Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Ηλεκτρικών Βιομηχανικών Διαταξεων &
Συστηματων Αποφασεων

Ανάπτυξη και σύγκριση αλγορίθμων δυναμικής ανάθεσης τιμών σε προϊόντα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Φωτεινή Λαμπρογεώργου

Επιβλέπων : Ασημακόπουλος Βασίλειος
Καθηγητής Ε.Μ.Π.

Υπεύθυνος : Ευάγγελος Σπηλιώτης
Διδάκτωρ Ε.Μ.Π

Αθήνα, Οκτώβριος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Ηλεκτρικών Βιομηχανικών Διαταξεων & Συστηματων Αποφασεων

Ανάπτυξη και σύγκριση αλγορίθμων δυναμικής ανάθεσης τιμών σε προϊόντα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Φωτεινή Λαμπρογεώργου

Επιβλέπων : Ασημακόπουλος Βασίλειος
Καθηγητής Ε.Μ.Π.

Υπεύθυνος : Ευάγγελος Σπηλιώτης
Διδάκτωρ Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την

(Υπογραφή)	(Υπογραφή)	(Υπογραφή)
……	……	……
Βασίλειος Ασημακόπουλος	Ιωάννης Ψαρράς	Δημήτριος Ασκούνης

Αθήνα, Οκτώβριος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Ηλεκτρικών Βιομηχανικών Διαταξεων & Συστηματων Αποφασεων

(Υπογραφή)

.....

Λαμπρογεώργου Φωτεινή, Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Ηλεκτρονικών Υπολογιστών

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια των ερευνητικών δραστηριοτήτων της Μονάδας Προβλέψεων και Στρατηγικής του Τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου Αθηνών κατά το ακαδημαϊκό έτος 2020-2021. Η μονάδα υπάγεται στον Τομέα Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων της σχολής.

Η εργασία αυτή αποτελεί την κατακλείδα σε μια ακαδημαϊκή πορεία και το ξεκίνημα νέων δρόμων που ακολουθούν. Θα ήθελα να ευχαριστήσω θερμα:

Τον επιβλέποντα καθηγητή κ. Βασίλειο Ασημακόπουλο για την ανάθεση και επίβλεψη της διπλωματικής αυτής σαν μια ευκαιρία να ερευνήσω παραπάνω τον ενδιαφέροντα τομέα των προβλέψεων και της μηχανικής μάθησης.

Τον Καθηγητή κ. Ι. Ψαρρά και τον Καθηγητή κ. Δημήτρη Ασκούνη για τη συμμετοχή τους στην εξεταστική επιτροπή της εργασίας.

Τον Ευάγγελο Σπηλιώτη, ερευνητικό συνεργάτη της Μονάδας Προβλέψεων και Στρατηγικής, για την αποδοτική συνεργασία και τη διαθεσιμότητα του να συμβάλει στην εκπόνηση της διπλωματικής καθόλη τη διάρκεια της διαδικασίας αυτής.

Τέλος, τους δικούς μου ανθρώπους για τη στήριξη και τη βοήθεια σε όλο αυτό το ακαδημαϊκό ταξίδι.

Περίληψη

Η παρούσα διπλωματική εξετάζει το θέμα της τιμολόγησης των προϊόντων, ένα ζήτημα που έχει απασχολήσει ιδιαίτερα κάθε μορφής πάροχο που εμπλέκεται σε εμπορικές δραστηριότητες και αποτελεί ακόμη πιο πολύπλοκη διαδικασία στις μέρες μας, όπου ο ανταγωνισμός πληθαίνει και οι επιλογές που έχουν οι καταναλωτές αυξάνονται. Στόχος της εργασίας είναι η ανάπτυξη και η σύγκριση αλγορίθμων που θα προσφέρουν τη δυνατότητα σχεδιασμού μιας δυναμικής στρατηγικής τιμολόγησης αλλά και ορισμού των επιμέρους παραμέτρων πώλησης (όπως η αποστολή, ή η διαθεσιμότητα των προϊόντων) που θα συμπεριλαμβάνεται σε μια γενικότερη μεθοδολογία με τέτοια ευελιξία ώστε να μπορεί να μεταβάλλεται προσφέροντας κάθε φορά στο προϊόν περισσότερες πιθανότητες να προτιμηθεί έναντι των ανταγωνιστικών και υποκατάστατων προϊόντων.

Ειδικότερα, η μελέτη γίνεται σε ένα σύνολο δεδομένων ηλεκτρονικών προϊόντων μεγάλων παρόχων του ηλεκτρονικού εμπορίου. Πάνω σε αυτή τη βάση εκπαιδεύονται και αξιολογούνται αλγόριθμοι επιβλεπόμενης μηχανικής εκμάθησης με στόχο αρχικά την πρόβλεψη των τιμών των προϊόντων με τον πιο ακριβή και αποδοτικό τρόπο και σε επόμενο στάδιο τον καθορισμό και των επιμέρους χαρακτηριστικών τους. Συνολικά, επιδιώκεται ο σχεδιασμός μιας στρατηγικής τιμολόγησης και πώλησης η οποία θα συμβάλει στο να γίνουν τα προϊόντα ενός εκάστοτε παρόχου δημοφιλή και ανταγωνιστικά μια συγκεκριμένη χρονική στιγμή την οποία εξετάζουμε.

Μέσα από την εργασία αυτή, παρουσιάζονται τα βασικά χαρακτηριστικά των προβλέψεων με μεγαλύτερη ανάλυση να γίνεται στις καινοτόμες μεθόδους πρόβλεψης με χρήση νευρωνικών δικτύων. Αναλύονται οι δύο βασικές μέθοδοι της παλινδρόμησης και της ταξινόμησης και τα στοιχεία των αλγορίθμων αυτών. Έπειτα, περιγράφεται η μεθοδολογία που χρησιμοποιείται και έχει ως στόχο τόσο την αποτελεσματική ανάλυση και διαχείριση των δεδομένων μέχρι να παράγουμε ένα εύχρηστο μοντέλο όσο και την εφαρμογή και αξιολόγηση των αλγορίθμων είτε αυτό αφορά ακρίβεια (regression), είτε επιτυχία πρότασης (classification). Τέλος, λόγω της προσπάθειας καθορισμού μιας δυναμικής στρατηγικής, εξετάζονται συγκεκριμένες περιπτώσεις οι οποίες συμβάλλουν στην εξαγωγή συμπερασμάτων για τους παράγοντες που καθορίζουν τη δημοφιλία των προϊόντων.

Η εργασία ολοκληρώνεται με την αξιοποίηση αυτών των συμπερασμάτων με στόχο τη δημιουργία μιας ευρύτερης στρατηγικής με βάση την οποία ένας πάροχος μπορεί όχι μόνο να καθορίσει δυναμικά τις τιμές των προϊόντων αλλά και τα υπόλοιπα στοιχεία που αφορούν την πώληση και διάθεση του προϊόντος στα μέσα του ηλεκτρονικού εμπορίου.

Λέξεις Κλειδιά: Δυναμική Τιμολόγηση, Μοντέλα Πρόβλεψης, Μηχανική Μάθηση, Ανάλυση Δεδομένων, Παλινδρόμηση, Ταξινόμηση

Abstract

The present thesis addresses the issue of pricing of products, a matter that has been particularly concerning for each provider involved in commercial activities and is an even more complex process nowadays when competition is rising and options which consumers have are increasing. The goal of the research is to develop and compare algorithms that will will offer the possibility of designing a dynamic pricing strategy and defining the individual sales parameters (such as shipping or availability of products) that will be included in a general methodology with such flexibility that it can be easily changed in order to ensure each time that the product has a better chance of being preferred over competing and substituting products.

In particular, the study is done on a data set of electronic products of major providers of e-commerce. Using this dataset, supervised machine learning algorithms are trained and evaluated with the aim of forecasting the products' prices in the most accurate and efficient way and also determining the individual characteristics of the products. All in all, we are trying to form a pricing and sales strategy in order to make products from one provider popular and competitive at a specific point in time that we are considering.

Through this work, the main characteristics of forecasts are presented, with greater analysis being made into innovative forecasting methods using neural networks. The two basic methods of regression and classification and the elements of these algorithms are analyzed. Then, we present the methodology being used, which aims at the effective analysis and management of data until we produce a workable model as well as the application and evaluation of the algorithms for either regression or classification. Finally, so as to define a dynamic strategy we are using specific case studies that contribute to drawing conclusions about factors that determine products' popularity.

The project is being completed by taking use of these conclusions with the aim of creating a broader strategy which will be used by a provider not only dynamically to determine the prices of products but also the other features related to the sale and disposal of the product in the means of electronic commerce.

Keywords: Dynamic Pricing, Forecasting Models, Machine Learning, Data Analysis, Regression, Classification

Πίνακας Περιεχομένων

Περιεχόμενα Σχημάτων

Περιεχόμενα Πινάκων

Κεφάλαιο 1. Εισαγωγή

1.1 Αντικείμενο Εργασίας

Η επιλογή μεθόδων τιμολόγησης και ο καθορισμός των τιμών των προϊόντων ενός παρόχου ήταν ανέκαθεν ένα ζήτημα που απασχολούσε τους πωλητές και που γίνεται ολοένα και πιο σημαντικό όσο ο ανταγωνισμός από διαφορετικούς παρόχους μεγαλώνει, τα υποκατάστατα προϊόντα πολλαπλασιάζονται και οι τρόποι πώλησης ποικίλλουν. Είναι αδιαμφισβήτητα μια σημαντική πρόκληση για τους παρόχους καθώς ορίζει τη ζήτηση αλλά και τη διαφοροποίηση των προϊόντων. Τα τελευταία χρόνια οι διάφοροι πωλητές και ιδιαίτερα οι ηλεκτρονικοί πάροχοι και οι μεγάλες εταιρείες ηλεκτρονικού εμπορίου χρησιμοποιούν νέες μεθόδους για να καθορίσουν πιο αποτελεσματικά τις τιμές των προϊόντων τους. Οι μέθοδοι αυτοί βασίζονται κυρίως σε τεχνικές προβλέψεων με χρήση μηχανικής εκμάθησης και υπόσχονται μια πιο δυναμική αλλά και αποτελεσματική στρατηγική τιμολόγησης των προϊόντων.

Μέχρι τώρα η πρόβλεψη τιμών ενός αγαθού/προϊόντος ή μιας υπηρεσίας γινόταν με την αξιολόγηση διαφόρων παραγόντων όπως τα χαρακτηριστικά του, η ζήτηση, οι εποχιακές τάσεις, οι τιμές των άλλων εμπορευμάτων, οι προσφορές από τους διάφορους προμηθευτές κ.λπ. Με βάση αυτές τις προβλέψεις πέρα από την πώληση των προϊόντων οι πάροχοι οργάνωναν την εφοδιαστική αλυσίδα τους, προγραμμάτιζαν τις πωλήσεις τους και σχεδίαζαν τον προϋπολογισμό τους. Οι αναλύσεις αυτές γίνονταν με στατικό τρόπο ή με βάση αναλύσεις των εμπειρογνωμόνων. Δυστυχώς, οι τεχνικές αυτές δεν είναι πλέον σε θέση να αντανακλούν την πολυπλοκότητα των παραγόντων που επηρεάζουν την εξέλιξη των τιμών ειδικά στο πλαίσιο του ηλεκτρονικού εμπορίου.

Ακόμη, πέρα από την τιμολόγηση, οι πάροχοι καλούνται πλέον να ορίσουν πολλαπλούς παράγοντες που θα τους διαφοροποιήσουν από τους ανταγωνιστές και αναφέρονται στο ηλεκτρονικό εμπόριο, με σημαντικότερους τις μεθόδους κοστολόγησης της αποστολής των προϊόντων αλλά και τη διαχείριση της διαθεσιμότητας και της κατάστασης των προϊόντων. Με τις πολυάριθμες παραγγελίες που γίνονται καθημερινά σε διεθνές επίπεδο στα πλαίσια του ηλεκτρονικού εμπορίου, είναι τεράστια πρόκληση για έναν πάροχο να οργανώσει το απόθεμα με τρόπο τέτοιο ώστε να μην ζημιώνεται οικονομικά αλλά και παράλληλα να μην "χάνει" πελάτες και άρα δημοφιλία και επισκέψεις στην ιστοσελίδα του. Παράλληλα, η εισαγωγή και πώληση τόσο μεταχειρισμένων, όσο και επισκευασμένων προϊόντων έχει αυξήσει ακόμη περισσότερο την πολυπλοκότητα αυτή καθώς τα υποκατάστατα προϊόντα πολλαπλασιάζονται και οι στρατηγικές μεταβάλλονται.

Για την πρόβλεψη των τιμών, λοιπόν, η μηχανική μάθηση παρέχει έναν μοναδικό τρόπο ενσωμάτωσης όλων των διαφορετικών προσεγγίσεων αλλά και την δυνατότητα του αυτοματισμού της διαδικασίας κάτι το οποίο παρέχει ταχύτητα, ακρίβεια αλλά και ευελιξία.

Πιο συγκεκριμένα, οι διαδικτυακές πλατφόρμες λιανικής πώλησης λειτουργούν μέσω της εφαρμογής αλγορίθμων και εφαρμογών που βασίζονται στην τεχνολογία της τεχνητής νοημοσύνης, με χρήση των στοιχείων πωλήσεων, προϊόντων και ανταγωνιστών για έναν πιο δυναμικό καθορισμό των τιμών.

Η παρούσα διπλωματική εργασία επικεντρώνεται στο συνδυασμό της ανάλυσης δεδομένων και της μηχανικής μάθησης για την πρόβλεψη των τιμών και τον καθορισμό παραγόντων των προϊόντων που θα τα κάνει πιο ανταγωνιστικά στον χώρο του ηλεκτρονικού εμπορίου. Χρησιμοποιώντας ως εργαλείο τα νευρωνικά δίκτυα και αναλύοντας τα δεδομένα με τον πιο αποδοτικό τρόπο επιδιώκουμε την πιο δυναμική αλλά και αποδοτική στρατηγική τιμολόγησης. Πιο συγκεκριμένα, αντικείμενο της εργασίας αποτελεί η ανάπτυξη μιας δυναμικής μεθοδολογίας τιμολόγησης προϊόντων ενός ηλεκτρονικού παρόχου. Αρχικά, στοχεύουμε στην εύρεση των προϊόντων εκείνων που στερούνται δημοφιλίας και δεν προτιμώνται από τους πελάτες. Στόχος της εργασίας είναι να βρεθεί μια μέθοδος που θα εξασφαλίζει ότι αυτά τα λιγότερο δημοφιλή προϊόντα ενός παρόχου θα τιμολογηθούν με τέτοιο τρόπο και θα ορίσουμε τις υπόλοιπες μεταβλητές και χαρακτηριστικά τους έτσι ώστε να γίνουν πιο δημοφιλή έναντι των υπολοίπων.

1.2 Μέθοδοι Τιμολόγησης και Μηχανική Μάθηση

Η τιμολόγηση του προϊόντος ήταν μία από τις μεγαλύτερες προκλήσεις για τις εταιρείες σε όλη την πορεία ύπαρξης τους από το παρελθόν μέχρι σήμερα. Οι οικονομολογοι είχαν από νωρίς διατυπώσει την άποψη "ότι οι τιμές καθορίζονται από την προσφορά και τη ζήτηση. Αν η σχετική ζήτηση για ένα προϊόν αυξηθεί, οι καταναλωτές θα είναι πρόθυμοι να πληρώσουν περισσότερα για αυτό. Οι ανταγωνιστικές προσφορές των προϊόντων από τη μία θα υποχρεώνουν τους καταναλωτές να πληρώνουν περισσότερα αλλά και θα επιτρέψουν στους παραγωγούς να κερδίζουν περισσότερα."

Σε γενικές γραμμές, η πρόβλεψη των τιμών γίνεται με περιγραφικά και προγνωστικά εργαλεία.

Η περιγραφική ανάλυση (descriptive analytics) βασίζεται σε στατιστικές μεθόδους που περιλαμβάνουν τη συλλογή δεδομένων, την ανάλυση, την ερμηνεία, και την παρουσίαση των ευρημάτων. Η περιγραφική ανάλυση επιτρέπει τη μετατροπή παρατηρήσεων (raw observations) σε γνώση που μπορεί να κατανοήσει κανείς και να τη μοιραστεί. Με λίγα λόγια, αυτός ο τύπος ανάλυσης βοηθάει να απαντηθεί το ερώτημα σχετικά με το τι συνέβη με τα δεδομένα.

Η προγνωστική ανάλυση (predictive analytics) αφορά την ανάλυση τρεχόντων και ιστορικών δεδομένων για την πρόβλεψη της πιθανότητας μελλοντικών συμβάντων και αποτελεσμάτων στο πλαίσιο των προβλέψεων τιμών. Η προγνωστική ανάλυση απαιτεί

πολυάριθμες στατιστικές τεχνικές, όπως η εξόρυξη δεδομένων (data mining - προσδιορισμός μοτίβων σε δεδομένα) και η μηχανική μάθηση.

Η τάση λοιπόν τα τελευταία χρόνια είναι οι παραπάνω μέθοδοι να συνδυάζονται και να βασίζονται στους αλγορίθμους μηχανικής μάθησης και τη χρήση νευρωνικών δικτύων. Τα τεχνητά νευρωνικά δίκτυα (artificial neural network) έχουν επιστήσει ιδιαίτερη προσοχή σε πολλούς κλάδους που αφορούν την αναγνώριση και την πρόβλεψη προτύπων και έχουν βρεθεί χρήσιμα σε αρκετές μελέτες μάρκετινγκ και καταναλωτικής συμπεριφοράς. Για παράδειγμα, οι Bentz και Merunka (2000) τα χρησιμοποίησαν ως διαγνωστικό εργαλείο με βάση τη πολυωνυμική λογική (MNL) και στόχο τη μοντελοποίηση αποφάσεων επιλογής μάρκας από τους καταναλωτές. Σε μια μελέτη προσομοίωσης και μια μελέτη σχετικά με τη συμπεριφορά των καταναλωτών, οι West, Brocket, και Golden (1997) διαπιστώνουν ότι η μηχανική μάθηση μπορεί να προσφέρει σημαντική βελτίωση σε σχέση με τα παραδοσιακά γραμμικά μοντέλα όπως η λογιστική παλινδρόμηση, επειδή τα νευρωνικά δίκτυα μπορεί να συλλάβουν τις μη γραμμικές σχέσεις των μη γραμμικών κανόνων απόφασης που σχετίζονται με τη συμπεριφορική ανάλυση για τους αγοραστές. Οι Agrawal και Schorling (1996) εφάρμοσαν τέτοιες τεχνικές για να προβλέπουν μετοχές από μάρκες που αναφέρονταν σε κατηγορίες προϊόντων παντοπωλείου και διαπίστωσαν ότι είναι καλύτερες από το συνήθως χρησιμοποιούμενο πολυωνυμικό μοντέλο λογικής. Οι Kumar, Rao, και Soni (1995) συνέκριναν τα νευρωνικά δίκτυα με τη λογιστική παλινδρόμηση στη μοντελοποίηση της απόφασης μιας αλυσίδας σούπερ μάρκετ για το αν θα πρέπει να φέρνει νέα προϊόντα, και συμπέραναν ότι τα νευρωνικά παράγουν καλύτερη ταξινόμηση, διαχειρίζονται καλύτερα πολύπλοκες υποκείμενες σχέσεις και είναι ισχυρότερα στην παρεμβολή. Αυτές και άλλες μελέτες δείχνουν ότι τα νευρωνικά δίκτυα είναι πολλά υποσχόμενη τεχνική για την ταξινόμηση και τη λήψη αποφάσεων πολλαπλών κριτηρίων αλλά και την πρόβλεψη τάσεων από την άποψη της ακρίβειας, της προσαρμοστικότητας, και της ευελιξίας που προσφέρουν.

Στα πλαίσια αυτά, τέτοιοι αλγόριθμοι εφαρμόζονται και για την τιμολόγηση προϊόντων προφέροντας νέες επιλογές για πιο ακριβή αποτελέσματα για την πρόβλεψη των τιμών των προϊόντων. Ο στόχος της μηχανικής μάθησης είναι να δημιουργήσει συστήματα ικανά να βρίσκουν μοτίβα στα δεδομένα, μαθαίνοντας από αυτά χωρίς ανθρώπινη παρέμβαση και ρητό επαναπρογραμματισμό. Για να λύσουν το πρόβλημα της πρόβλεψης τιμών, οι αναλυτές δεδομένων (data analyst - data scientist) πρέπει πρώτα να κατανοήσουν ποια δεδομένα πρέπει να χρησιμοποιήσουν για να εκπαιδεύσουν μοντέλα μηχανικής μάθησης. Στη συνέχεια, συλλέγουν, επιλέγουν, προετοιμάζουν, προ-επεξεργάζονται και μετασχηματίζουν αυτά τα δεδομένα. Μόλις ολοκληρωθεί αυτό το στάδιο, οι ειδικοί αρχίζουν να δημιουργούν προγνωστικά μοντέλα. Για την τροφοδοσία ενός συστήματος ή μιας εφαρμογής, επιλέγεται ένα μοντέλο που προβλέπει τιμές με τον υψηλότερο ρυθμό ακρίβειας. Γι' αυτό ακριβώς χρειάζονται τόσο η περιγραφική όσο και η προγνωστική ανάλυση, με την πρώτη να είναι σημαντική στα πρώτα βήματα ενώ η άλλη στη συνέχεια.

Πιο συγκεκριμένα, η πρόβλεψη τιμής μπορεί να διατυπωθεί ως εργασία παλινδρόμησης (regression task). Η ανάλυση παλινδρόμησης (regression analysis) είναι μια στατιστική τεχνική που χρησιμοποιείται για την εκτίμηση της σχέσης μεταξύ μιας εξαρτημένης/στοχευόμενης μεταβλητής (τιμή ενός προϊόντος για παράδειγμα) και μίας ή πολλών ανεξάρτητων (αλληλοεξαρτώμενων) μεταβλητών πρόβλεψης, δηλαδή εκείνων που επηρεάζουν τη μεταβλητή-στόχο. Η ανάλυση παλινδρόμησης επιτρέπει επίσης τον προσδιορισμό του βαθμού επιρροής αυτών των προγνωστικών παραγόντων για την αντίστοιχη μεταβλητή-στόχο. Στην παλινδρόμηση, μια μεταβλητή προορισμού είναι πάντα αριθμητική.

Η τιμολόγηση που βασίζεται στη μηχανική εκμάθηση γίνεται σταδιακά, λοιπόν, βασικό χαρακτηριστικό του λιανικού εμπορίου. Σύμφωνα με μια μελέτη της IBM, 73 τοις εκατό των εταιρειών που ερωτήθηκαν σχεδιάζουν να βελτιστοποιήσουν τις τιμές και τις προωθητικές ενέργειές τους μέσω έξυπνων μεθόδων αυτοματισμού πριν από το τέλος του 2021. Για να παραμείνουν ανταγωνιστικές, οι λιανοπωλητές επιβάλλεται να εξετάσουν το ενδεχόμενο μεταβολής της υπάρχουσας μεθόδου τιμολόγησης σε τεχνολογία που βασίζεται στην μηχανική μάθηση.

Παρόλα αυτά, έχει παρατηρηθεί πως η ακριβής τιμή είναι δύσκολο να προσδιοριστεί, καθώς οι παράγοντες είναι πολλοί, μεταβάλλονται συνεχώς και καμία εταιρεία δεν μπορεί να διαθέτει όλα τα δεδομένα για να κάνει τις πιο αποδοτικές προβλέψεις. Για το λόγο αυτό, η προσέγγιση που αναφέρθηκε σταδιακά μεταβάλλεται και γίνεται μια πολυπαραγοντική διαδικασία καθορισμού της τιμής σύμφωνα με τη δημοφιλία των προϊόντων και τον καθορισμό και των υπόλοιπων χαρακτηριστικών τους πέρα από την τιμή με στόχο την ανταγωνιστική παρουσία τους στην αγορά. Έτσι πλέον μιλάμε για εναλλακτικές συνολικές προσφορές που θα προτιμηθούν έναντι άλλων και όχι μόνο καθορισμό τιμών αλλά και καθορισμό των επιμέρους παραγόντων.

Πολλαπλές μελέτες έχουν επικεντρωθεί στον καθορισμό των παραγόντων αυτών γύρω από τον όρο που ονομάζουμε δυναμική ανάθεση τιμών (dynamic pricing). Η βασική ιδέα της δυναμικής τιμολόγησης είναι η μοντελοποίηση της επίδρασης της αλληλεπίδρασης μεταξύ διαφόρων παραγόντων, όπως η τιμή του προϊόντος σε διαφορετικούς χρόνους, η ζήτηση του προϊόντος και η χρήση του μοντέλου αυτού για τη βελτιστοποίηση της τιμολογιακής πολιτικής. Αυτά τα μοντέλα ζήτησης προέρχονται κυρίως από τα ιστορικά δεδομένα για τις τιμές και τη ζήτηση των προϊόντων στο παρελθόν. Στην πράξη, γίνονται διάφορες παραδοχές σχετικά με τον τρόπο αλληλεπίδρασης αυτών των παραγόντων, με αποτέλεσμα να προκύπτουν διαφορετικοί τύποι μοντέλων ζήτησης. Για τη δυναμική τιμολόγηση σύμφωνα με τις μελέτες που έχουν γίνει είναι σημαντική και η ανάλυση και πρόβλεψη της συμπεριφοράς των καταναλωτών. Στο πλαίσιο αυτό, η προσέγγιση του νευρωνικού δικτύου έγκειται στην ικανότητά του να μιμείται τη λειτουργία του ανθρώπινου εγκεφάλου και να εξασφαλίζει ακριβή πρόβλεψη της συμπεριφοράς που βασίζεται σε προϊόντα ιδιότητες ή/και

εικόνα, ακόμη και όταν η υποκείμενη διαδικασία επιλογής είναι μη γραμμική από τη φύση της. Η αρχιτεκτονική αυτών των μοντέλων βασίζεται στο ευρέως αποδεκτό μοντέλο Spreading Activation Model of human memory (Anderson 1976, Collins και Loftus 1975, Quillan 1968) και μοντέλα αναγνώρισης προτύπων και μάθησης (Simon 1996). Σε μελέτη των S.Shakyaa, M.Kerna, G. Owusua, C.M. Chinb το 2012, οι τελευταίοι δημιουργούν ένα μοντέλο ζήτησης που βασίζεται σε νευρωνικά δίκτυα και χρησιμοποιούν εξελικτικούς αλγόριθμους για να βελτιστοποιήσουν την πολιτική στο μοντέλο δόμησης. Αυτή η προσέγγιση έχει δύο βασικά οφέλη. Η χρήση του νευρωνικού δικτύου το καθιστά αρκετά ευέλικτο ώστε να μοντελοποιήσει μια σειρά από διαφορετικά σενάρια ζήτησης που συμβαίνουν μέσα σε διαφορετικά προϊόντα και υπηρεσίες, και η χρήση του εξελικτικού αλγορίθμου το καθιστά αρκετά ευέλικτο ώστε να λύνει πολύ σύνθετα προβλήματα. Επίσης, αξιολογούνται οι τιμολογιακές πολιτικές που βρέθηκαν από αυτό το νευρωνικό δικτυακό μοντέλο σε σχέση με αυτές που βρέθηκαν από άλλα ευρέως χρησιμοποιούμενα μοντέλα ζήτησης. Τα αποτελέσματά τους δείχνουν ότι το προτεινόμενο μοντέλο είναι πιο συνεπές, προσαρμόζεται καλά σε μια σειρά από διαφορετικά σενάρια, και σε γενικές γραμμές, βρίσκει πιο ακριβή πολιτική τιμολόγησης από άλλα τρία συγκριτικά μοντέλα. Σε αντίστοιχη μελέτη το 2017, πάνω σε στοχαστικά δυναμική τιμολόγηση και διαφήμιση έχουν τιμολογηθεί προϊόντα σε ειδικές ολιγοπωλιακές αγορές με σταθερή ελαστικότητα τιμής και διαφήμισης σημαντικά αποδοτικά σε σχέση με τις παραδοσιακές μεθόδους.

Τέλος, η δυναμική τιμολόγηση συνδέθηκε από πολύ νωρίς με την πρόβλεψη ζήτησης, κάτι το οποίο θα μελετηθεί και στην παρούσα εργασία. Ειδικότερα, για να μπορεί να εφαρμοστεί η θεωρία της βέλτιστης τιμολόγησης σε πρακτικά προβλήματα, πρέπει να υπάρχει μια εκτίμηση της λειτουργίας ζήτησης. Το πρώτο γνωστό εμπειρικό έργο σχετικά με τις καμπύλες ζήτησης ήταν ο νόμος King-Davenant, ο οποίος αφορούσε την προσφορά και την τιμή του καλαμποκιού. Πιο προχωρημένες έρευνες για την εκτίμηση καμπυλών ζήτησης, μέσω στατιστικών τεχνικών όπως η συσχέτιση και η γραμμική παλινδρόμηση, ξεκίνησαν στις αρχές του 20ου αιώνα. Οι Benini, Gini και Lehfeldt εκτίμησαν τις καμπύλες ζήτησης για διάφορα αγαθά όπως ο καφές, το τσάι, το αλάτι και το σιτάρι, χρησιμοποιώντας διάφορες μεθόδους προσαρμογής καμπυλών. Περαιτέρω πρόοδος στη μεθοδολογία σημειώθηκε, μεταξύ άλλων, από τον Moore, τον Wright και τον Tinbergen με σταθμό το μνημειώδες έργο του Schultz που παρέχει μια εμπεριστατωμένη επισκόπηση της σύγχρονης εκτίμησης της ζήτησης στην εποχή του, συνοδευόμενη από πολλά παραδείγματα. Οι επόμενες μελέτες επικεντρώθηκαν σε ψηφιακά περιβάλλοντα πωλήσεων που γενικά παρέχουν στις εταιρείες μια πληθώρα δεδομένων πωλήσεων που δεν ήταν διαθέσιμη στο παρελθόν. Τα δεδομένα αυτά μπορεί να περιέχουν σημαντικές πληροφορίες σχετικά με τη συμπεριφορά των καταναλωτών, ιδίως σχετικά με τον τρόπο με τον οποίο οι καταναλωτές ανταποκρίνονται στις διαφορετικές τιμές πώλησης. Η αξιοποίηση των γνώσεων που περιέχονται στα δεδομένα και η εφαρμογή τους σε πολιτικές δυναμικής τιμολόγησης μπορεί να παρέχει βασικά ανταγωνιστικά πλεονεκτήματα, και η γνώση σχετικά με τον τρόπο με τον οποίο θα πρέπει να γίνει αυτό έχει μεγάλη πρακτική σημασία και θεωρητικό ενδιαφέρον. Το ζήτημα αυτό αποτελεί βασική κινητήρια δύναμη της έρευνας για τη δυναμική τιμολόγηση και μάθηση: η μελέτη της βέλτιστης δυναμικής τιμολόγησης σε ένα

αβέβαιο περιβάλλον, όπου τα χαρακτηριστικά της συμπεριφοράς των καταναλωτών μπορούν να διδαχθούν από τη συσσώρευση δεδομένων πωλήσεων.
Συμπερασματικά, καταλήγουμε στο ότι ενώ η τιμολόγηση και η πρόβλεψη ζήτησης των προϊόντων αποτέλεσε από νωρίς ένα ζήτημα που απασχολούσε οποιοδήποτε έμπορο, μόνο τα τελευταία χρόνια έχει επιτευχθεί μια πιο δυναμική προσέγγιση στο ζήτημα. Η προσέγγιση αυτή, βασισμένη στη μηχανική μάθηση και τα νευρωνικά δίκτυα, επιτρέπει την αξιοποίηση δεδομένων των καταναλωτών και των προϊόντων τόσο για την πρόβλεψη της ζήτησης όσο και για τη δυναμική τιμολόγηση σύμφωνα με τις εκάστοτε τάσης της αγοράς. Όλα αυτά προσφέρουν ένα νέο πεδίο στο χώρο των ηλεκτρονικών πωλήσεων και μπορεί να εξασφαλίσει την εξατομικευμένη στρατηγική τιμολόγησης με ακρίβεια και δυναμικότητα.

1.3 Δομή Εργασίας

Στο πρώτο κεφάλαιο της παρούσας εργασίας γίνεται μια εισαγωγή στο θέμα της διπλωματικής, μια ιστορική αναδρομή σχετικά με μελέτες πάνω σε αυτό αλλά και μια περιγραφή της βασικής δομής που θα ακολουθηθεί για την παρουσίαση του θέματος.

Στο δεύτερο κεφάλαιο, περιγράφονται οι στατιστικές μέθοδοι προβλέψεων που χρησιμοποιούνται και αφού γίνει μια ιστορική αναδρομή της εξέλιξης τους και περιγραφούν αναλυτικά, τις αξιολογούμε αναφερόμενοι στις περιπτώσεις χρήσης τους και στα πλεονεκτήματα και μειονεκτήματα τους. Έπειτα αναφερόμαστε στις μεθόδους προβλέψης με χρήση νευρωνικών δικτύων και επιβλεπόμενης μηχανικής μάθησης, αναφερόμενοι και πάλι στις ιστορικές εξελίξεις αλλά και αξιολογώντας τα οφέλη που επιφέρουν.

Στο τρίτο κεφάλαιο, παρουσιάζονται αναλυτικά όλα τα βήματα της μεθοδολογικής προσέγγισης του θέματος που θέλουμε να επιλύσουμε στην παρούσα διπλωματική αλλά και οι επιμέρους μέθοδοι που έχουμε χρησιμοποιήσει σε αυτά τα βήματα. Ακόμη αναλύονται τα μοντέλα και οι μετρικές αξιολόγησης τους.

Στο τέταρτο κεφάλαιο, αναλύεται εκτενώς η διαδικασία που ακολουθήθηκε πάνω στα πραγματικά δεδομένα, παρουσιάζονται τα αποτελέσματα και αξιολογούνται. Ακόμη στο κεφάλαιο αυτό παρουσιάζεται με την ίδια ανάλυση και η νέα προσέγγιση που ακολουθείται η οποία με τη σειρά της αξιολογείται έπειτα από την παρουσίαση των αποτελεσμάτων.

Στο πέμπτο κεφάλαιο, περιγράφονται τα case studies που θέσαμε σε εφαρμογή και αναλύονται τα αποτελέσματά τους με στόχο την εξαγωγή συμπερασμάτων από κάθε μια από τις περιπτώσεις που εξετάζουμε.

Στο τελευταίο και έκτο κεφάλαιο αναφερόμαστε στα γενικότερα συμπεράσματα που εξάγουμε από όλη τη μελέτη της παρούσας διπλωματικής εργασίας, αναφέρονται οι προεκτάσεις που αυτά μπορεί να έχουν σχετικά με την μελέτη αυτή και τέλος τα πιθανά θέματα που προκύπτουν για περαιτέρω ανάλυση.

Κεφάλαιο 2. Μέθοδοι Πρόβλεψης

2.1 Εισαγωγή

Οι άνθρωποι ήταν από πάντα ένα είδος που είχε ανάγκη να προβλέπει, να σχεδιάζει και να λαμβάνει αποφάσεις έπειτα από μια διαδικασία ανάλυσης και σκέψης. Ακόμη και αν έλειπαν οι εντονότερες αισθήσεις και το πιο δυνατό ένστικτο, το ανθρώπινο είδος αποτελούσε πάντα έναν "διανοούμενο σχεδιαστή" που προσπαθούσε να προβλέψει τον καιρό, την διαδρομή ενός θηράματος, την ύπαρξη τροφής, την καρποφορία του εδάφους και τα λοιπά. Αυτές οι προβλέψεις ήταν που έδωσαν και ένα μεγάλο πλεονέκτημα στην επιβίωση του ατόμου και άνοιξαν το δρόμο της ανάπτυξης του τομέα των προβλέψεων τόσο ατομικά όσο και συλλογικά σε επιστημονικό και ακαδημαϊκό επίπεδο.

Παρόλα αυτά όσο η πρόοδος συνεχιζόταν τόσο αυξανόταν η ανάγκη επίτευξης του στόχου της ακριβής πρόβλεψης, κάτι το οποίο άνοιξε με τη σειρά του το δρόμο της ανάπτυξης κάποιων πιο συστηματικών μεθόδων πρόβλεψης. Οι τελευταίες στόχευαν στην αποτελεσματικότερη παραγωγή προβλέψεων ακόμα και σε ένα απρόβλεπτο και περίπλοκο περιβάλλον με μεγάλη ακρίβεια. Στο πλαίσιο αυτής της σύγχρονης ανάγκης αναπτύχθηκε η παρούσα εργασία, για αυτό είναι σημαντικό αφού αναλύσουμε την εξέλιξη του τομέα αυτού να διατυπώσουμε κάποιες βασικές στατιστικές μεθόδους πρόβλεψης και τους τρόπους αξιολόγησης αλλά και την τελική αποτίμηση τους.

2.2 Κλασικές Στατιστικές Μέθοδοι Πρόβλεψης

2.2.1. Ιστορική Εξέλιξη

Αν και ορισμένες στατιστικές μέθοδοι έχουν ηλικία πάνω από δύο χιλιετίες, οι διάφορες μέθοδοι προβλέψεων εμφανίστηκαν περίπου τον 17ο αιώνα. Ο Karl Pearson παρουσίασε τον ευρέως χρησιμοποιούμενο συντελεστή συσχέτισης στιγμή-προϊόντος Pearson και ο John Galton ανέπτυξε βασικές έννοιες όπως η τυπική απόκλιση, η συσχέτιση, ακόμα και η ανάλυση παλινδρόμησης. Ο Ronald Fisher ανέπτυξε τη μηδενική υπόθεση και πολλές άλλες βασικές έννοιες. Τέτοιες ιδέες παραμένουν αναπόσπαστο μέρος της σύγχρονης επιστήμης δεδομένων, της μηχανικής μάθησης και της προγνωστικής ανάλυσης. Το ανθρώπινο είδος είχε τελικά έναν συστηματικό τρόπο συλλογής, ανάλυσης και παρουσίασης αριθμητικών δεδομένων που σχετίζονται με πιθανότητες. Η στατιστική επέτρεψε στους ερευνητές και τους αναλυτές να μετρούν, να επικοινωνούν και μερικές φορές ακόμη και να ελέγχουν την αβεβαιότητα.

Η προέλευση των στατιστικών μεθόδων πρόβλεψης χρονολογείται τουλάχιστον την ίδια περίοδο με την σύγχρονη στατιστική, αλλά δεν ήταν μέχρι τη δεκαετία του 1950 που διάφοροι οργανισμοί ξεκίνησαν να χρησιμοποιούν μοντέλα με υπολογιστές για να προβλέπουν τα πάντα, από τα καιρικά φαινόμενα μέχρι τους πιστωτικούς κινδύνους. Τη δεκαετία του 1970, αναπτύχθηκε το διάσημο μοντέλο Black Scholes για την πρόβλεψη των καλύτερων τιμών για τις μετοχές, και τη δεκαετία του 1990 έγινε η ευρεία χρήση της ανάλυσης για ένα ευρύ φάσμα δεδομένων, από αναζητήσεις στο διαδίκτυο μέχρι και μπέιζμπολ line-ups. Μόλις πρόσφατα έρευνα που διεξήχθη από τον Φίλιπ Τετλοκ και άλλους απέδειξε - μέσω προβλέψεων τουρνουά που πραγματοποιήθηκε από την Υπηρεσία Προηγμένων Ερευνητικών Έργων Πληροφοριών (Intelligence Advanced Research Projects Agency, IARPA) - ότι οι άνθρωποι μπορούν να γίνουν εμπειρικά καλύτεροι στις προβλέψεις. Η έρευνα δείχνει ότι οι καλοί ειδικοί στις προβλέψεις μπορούν ακόμη και να διδάξουν σε άλλους πώς να κάνουν πιο αποτελεσματικές τις προβλέψεις.

Έπειτα και καθώς τα τέλη του 19ου και οι αρχές του 20ου αιώνα επλήγησαν από μια σειρά κρίσεων που οδήγησαν σε σοβαρούς πανικούς παγκοσμίως καθώς επίσης και σημαντικές δημογραφικές αλλαγές, αφού οι χώρες πέρασαν από το να είναι κυρίως γεωργικές στο να είναι βιομηχανικές και αστικές, οι άνθρωποι αγωνίζονταν να βρουν σταθερότητα στο ευμετάβλητο περιβάλλον. Την περίοδο εκείνη, αναπτύχθηκαν οι προβλέψεις που βασίζονται σε στατιστικές και θα αναλυθούν παρακάτω.

Μελλοντικά είναι σίγουρο ότι θα υπάρξουν περαιτέρω βελτιώσεις στην τέχνη και την επιστήμη των προβλέψεων, δεδομένων όλων των οικονομικών και στρατηγικών κινήτρων για κάτι τέτοιο. Προβλέπεται ότι η αγορά του τομέα αυτού θα φτάσει τα 6,5 δισεκατομμύρια δολάρια μέχρι το 2019. Αν υπολογίσουμε μάλιστα τις πολυάριθμες χρήσεις των προβλέψεων σε πολιτικό, κοινωνικό, οικονομικό επίπεδο, όπως προβλέψεις τιμών, αποθεμάτων, πολιτικές δημοσκοπήσεις, πρόγνωση του καιρού και άλλα, η «βιομηχανία της πρόβλεψης» θα είναι πράγματι τεράστια. και θα αυξάνεται με ταχύ ρυθμό κάθε χρόνο.

2.2.2. Επισκόπηση Στατιστικών Μεθόδων Πρόβλεψης

Οι μέθοδοι πρόβλεψης είναι πολλές και διαφοροποιούνται ανάλογα με το είδος των προβλέψεων που θέλουμε να κάνουμε. Μια ενδεικτική κατηγοριοποίηση είναι η εξής:

1. Ποιοτικές Μέθοδοι: Αυτές οι μέθοδοι βασίζονται σε συναισθήματα, διαισθήσεις, κρίσεις, προσωπικές εμπειρίες και απόψεις. Αυτό σημαίνει ότι δεν χρησιμοποιούνται μαθηματικά στις μεθόδους ποιοτικής πρόβλεψης. Κάποια παραδείγματα τέτοιων μεθόδων είναι η Μέθοδος των Δελφών (Delphi Method), Έρευνα αγοράς (Market Survey), Εκτελεστική Γνώμη (Executive Opinion) και τα λοιπά.

2. Ποσοτικές Μέθοδοι: Αυτές οι μέθοδοι εξαρτώνται εξ ολοκλήρου από μαθηματικά ή ποσοτικά μοντέλα και υπολογισμούς. Αυτού του είδους οι προβλέψεις αποτελούν τα μοντέλα χρονοσειρών (Time Series Models) και συσχέτισης (Associative Models).

Παρόλα αυτά, παρακάτω θα παρουσιαστούν οι πιο ευρέως χρησιμοποιημένες και αυτές που αφορούν κατά κύριο λόγο προβλέψεις οικονομικών στοιχείων όπως μελεταμε στην παρούσα εργασία.

Αφελής Πρόβλεψη (Naïve Forecasting)

Η αφελής πρόβλεψη είναι μια τεχνική εκτίμησης σύμφωνα με την οποία οι τιμές της τελευταίας περιόδου χρησιμοποιούνται ως προγνώσεις αυτής της περιόδου, χωρίς να μεταβάλλονται ή να γίνεται προσπάθεια καθορισμού αιτιωδών παραγόντων (casual factors). Με άλλα λόγια, μια αφελής πρόβλεψη είναι απλώς η πιο πρόσφατα παρατηρούμενη τιμή. Υπολογίζεται με τον τύπο F(t) + k = y(t), όπου, κατά τον χρόνο t, η καθαρή πρόγνωση k-step-ahead (Ft+k) ισούται με την παρατηρούμενη τιμή κατά τον χρόνο t.

Στη βιομηχανία και το εμπόριο, η μέθοδος αυτή χρησιμοποιείται κυρίως για τη σύγκριση με τις προβλέψεις που παράγονται από τις πιο εξελιγμένες τεχνικές ως σημείο αναφοράς. Ωστόσο, μερικές φορές τα αποτελέσματα της μεθόδου αυτής είναι το καλύτερο που μπορεί να γίνει για πολλές χρονολογικές σειρές, συμπεριλαμβανομένων των περισσότερων δεδομένων τιμών μετοχών. Ακόμη η παρακολούθηση της αφελούς πρόβλεψης με την πάροδο του χρόνου και η εκτίμηση της προβλεπόμενης αξίας που προστίθεται στη διαδικασία προγραμματισμού συμβάλει στο να τεθεί μια βάση για τις προβλέψεις. Έτσι, ακόμα και αν δεν είναι η πιο ακριβής μέθοδος πρόβλεψης, παρέχει ένα χρήσιμο μέτρο αναφοράς για άλλες προσεγγίσεις.

Στατιστική πρόβλεψη (Statistical Forecasting)

Η στατιστική πρόβλεψη είναι μια μέθοδος που βασίζεται σε μια συστηματική στατιστική εξέταση δεδομένων που αντιπροσωπεύουν παρελθούσα παρατηρούμενη συμπεριφορά του συστήματος προς πρόβλεψη, η οποία περιλαμβάνει παρατηρήσεις χρήσιμων δεικτών πρόβλεψης εκτός του συστήματος. Με απλά λόγια, χρησιμοποιεί στατιστικές που βασίζονται σε ιστορικά δεδομένα για να προβάλει τι θα μπορούσε να συμβεί στο μέλλον. Οι δύο κύριες μέθοδοι στατιστικής πρόβλεψης είναι η πρόβλεψη βάσει χρονοσειρών και η πρόβλεψη βάσει μοντέλων.

Η πρόβλεψη βάσει χρονοσειρών (Time Series forecasting) είναι μια βραχυπρόθεσμη αμιγώς στατιστική μέθοδος πρόβλεψης που προβλέπει βραχυπρόθεσμες αλλαγές με βάση ιστορικά δεδομένα. Επεξεργάζεται δεδομένα με βάση τον χρόνο (έτη, ημέρες, ώρες και λεπτά), για να βρει κρυφές πληροφορίες. Η απλούστερη τεχνική πρόβλεψης χρονοσειρών είναι ένας απλός κινητός μέσος (simple moving average - SMA). Υπολογίζεται με την πρόσθεση των τιμών της τελευταίας περιόδου n και, στη συνέχεια, τη διαίρεση του αριθμού αυτού με το

n. Έτσι, η τιμή του κινούμενου μέσου χρησιμοποιείται στη συνέχεια ως πρόβλεψη για την επόμενη περίοδο.

Η πρόβλεψη βάσει μοντέλου (Model-Based Forecasting) είναι πιο στρατηγική και μακροπρόθεσμη μέθοδος και προβλέπει αλλαγές στο επιχειρηματικό περιβάλλον και συμβάντα με λίγα δεδομένα. Αποτελεί μέθοδο παρόμοια με τα συμβατικά προγνωστικά μοντέλα που έχουν ανεξάρτητες και εξαρτημένες μεταβλητές, αλλά η ανεξάρτητη μεταβλητή είναι τώρα ο χρόνος. Η απλούστερη από αυτές τις μεθόδους είναι η γραμμική παλινδρόμηση (linear regression). Με δεδομένο ένα σύνολο εκπαίδευσης, εκτιμούμε τις τιμές των συντελεστών παλινδρόμησης για να προβλέψουμε μελλοντικές τιμές της μεταβλητής στόχου.

Με τον καιρό, οι βασικές στατιστικές μέθοδοι πρόβλεψης έχουν δει σημαντικές βελτιώσεις στις μεθόδους, σχηματίζοντας ένα φάσμα μεθόδων που βασίζονται σε δεδομένα και διαφορετικές τεχνικές μοντελοποίησης.

Μέθοδοι Εκθετικής Εξομάλυνσης και Holt-Winters Filtering

Η εκθετική εξομάλυνση (exponential smoothing) προτάθηκε για πρώτη φορά στη στατιστική βιβλιογραφία χωρίς αναφορά σε προηγούμενη εργασία του Ρόμπερτ Γκούντελ Μπράουν το 1956. Η εκθετική εξομάλυνση είναι ένας τρόπος εξομάλυνσης των δεδομένων αφαιρώντας μεγάλο μέρος του «θορύβου» με στόχο μια καλύτερη πρόβλεψη. Αποδίδει εκθετικά μειούμενα βάρη καθώς η παρατήρηση μεγαλώνει ως εξής: yx=α*yx + (1-a)*yx-1, όπου έχουμε σταθμισμένο κινητό μέσο με δύο συντελεστές α και 1-α. Αυτή η απλούστερη μορφή εκθετικής εξομάλυνσης μπορεί να χρησιμοποιηθεί για βραχυπρόθεσμη πρόβλεψη με μια χρονοσειρά που μπορεί να περιγραφεί με τη χρήση ενός προσθετικού μοντέλου με σταθερό επίπεδο και χωρίς εποχικότητα.

Η Holt-Winters Filtering είναι μια μέθοδος που προτάθηκε από τον Τσαρλς Χολτ το 1957 ως μια παραλλαγή της εκθετικής εξομάλυνσης για μια χρονοσειρά που μπορεί να περιγραφεί με τη χρήση ενός προσθετικού μοντέλου με αυξανόμενη ή φθίνουσα τάση και χωρίς εποχικότητα. Η ιδέα πίσω από αυτόν τον αλγόριθμο είναι η εφαρμογή εκθετικής εξομάλυνσης στα εποχιακά συστατικά εκτός από το επίπεδο και την τάση. Η εξομάλυνση εφαρμόζεται σε όλες τις εποχές, π.χ. η εποχική συνιστώσα του 3ου σημείου στην εποχή θα εξομαλυνθεί εκθετικά με αυτή από το 3ο σημείο της περασμένης σεζόν, 3ο σημείο δύο εποχές πριν κλπ.

Ολοκληρωμένο Αυτοπαλινδρομικό Μοντέλο Κινητού Μέσου Όρου (Autoregressive Integrated Moving Average - ARIMA)

Το ARIMA είναι μια στατιστική τεχνική που χρησιμοποιεί δεδομένα χρονοσειρών για να προβλέψει το μέλλον. Είναι παρόμοια με την εκθετική εξομάλυνση στο ότι είναι

προσαρμοστική, μπορεί να μοντελοποιήσει τάσεις και εποχιακά μοτίβα, και μπορεί να αυτοματοποιηθεί. Ωστόσο, τα μοντέλα ARIMA βασίζονται σε αυτοσυσχετίσεις (μοτίβα στον χρόνο) και όχι σε μια δομημένη και αυστηρή άποψη του επιπέδου, της τάσης και της εποχικότητας. Συνολικά, τα μοντέλα ARIMA λαμβάνουν υπόψη τις τάσεις, την εποχικότητα, τους κύκλους, τα σφάλματα και τις μη στατικές πτυχές ενός συνόλου δεδομένων όταν κάνουν προβλέψεις. Τα ARIMA ελέγχουν τη στατικότητα των δεδομένων και αν τα δεδομένα δείχνουν σταθερή διακύμανση με την αλλαγή του χρόνου. Η ιδέα πίσω από αυτά είναι ότι το τελικό υπόλειμμα θα πρέπει να μοιάζει με λευκό θόρυβο, διαφορετικά, υπάρχουν πληροφορίες διαθέσιμες στα δεδομένα προς εξαγωγή.

Τα μοντέλα ARIMA έχουν καλύτερες επιδόσεις από τα μοντέλα εκθετικής εξομάλυνσης για μεγαλύτερα, σταθερότερα σύνολα δεδομένων και όχι τόσο για πιο θορυβώδη και ευμετάβλητα δεδομένα. Αν και πολλά από τα μοντέλα χρονοσειρών μπορούν να δημιουργηθούν σε υπολογιστικά φύλλα, το γεγονός ότι βασίζονται σε ιστορικά δεδομένα τα καθιστά εύκολα αυτοματοποιημένα. Επομένως, τα πακέτα λογισμικού μπορούν να παράγουν αυτόματα μεγάλες ποσότητες από αυτά τα μοντέλα σε μεγάλα σύνολα δεδομένων. Ειδικότερα, τα δεδομένα μπορεί να διαφέρουν σημαντικά, και η υλοποίηση αυτών των μοντέλων ποικίλλει επίσης, οπότε το αυτοματοποιημένο στατιστικό λογισμικό μπορεί να βοηθήσει στον προσδιορισμό της βέλτιστης προσαρμογής σε κάθε περίπτωση ξεχωριστά.

Μοντέλα Παλινδρόμησης (Regression Models)

Τα μοντέλα δυναμικής παλινδρόμησης επιτρέπουν την ενσωμάτωση αιτιωδών παραγόντων, όπως οι τιμές, οι προαγωγές και οι οικονομικοί δείκτες στις προβλέψεις. Τα μοντέλα συνδυάζουν την Μέθοδο των Ελαχίστων Τετραγώνων (Ordinary Least Squares - OLS) με τη δυνατότητα να χρησιμοποιούν δυναμικούς όρους για να συλλάβουν την τάση, την εποχικότητα και τις χρονικά κατανεμημένες σχέσεις μεταξύ των μεταβλητών.

Ένα μοντέλο δυναμικής παλινδρόμησης παρέχει πληροφορίες για τις σχέσεις μεταξύ των μεταβλητών και επιτρέπει σενάρια «what if». Για παράδειγμα, αν μελετήσουμε τη σχέση μεταξύ πωλήσεων και τιμής, το μοντέλο μας επιτρέπει να δημιουργούμε προβλέψεις κάτω από διάφορα σενάρια τιμών, όπως «Τι θα συμβεί αν αυξήσουμε την τιμή;» ή «Τι θα συμβεί αν το χαμηλώσουμε;». Η παραγωγή αυτών των εναλλακτικών προβλέψεων μπορεί να βοηθήσει να καθοριστεί μια αποτελεσματική στρατηγική τιμολόγησης, όπως θα γίνει και στην παρούσα μελέτη. Ένα καλά καθορισμένο μοντέλο δυναμικής παλινδρόμησης συλλαμβάνει τη σχέση μεταξύ της εξαρτημένης μεταβλητής (αυτή που αποτελεί το αντικείμενο πρόβλεψης) και μία ή περισσότερες (σε περιπτώσεις γραμμικών ή πολλαπλών παλινδρομήσεων, αντίστοιχα) ανεξάρτητες μεταβλητές. Για να δημιουργηθεί μια πρόβλεψη, πρέπει να υπάρχουν προβλέψεις για τις ανεξάρτητες μεταβλητές.

2.2.3. Αξιολόγηση Στατιστικών Μεθόδων Πρόβλεψης

Γενικά, οι παραδοσιακοί αλγόριθμοι που αναφέρθηκαν παραπάνω έχουν την τάση να χρησιμοποιούν προκαθορισμένες τεχνικές και στατιστικά μοντέλα. Ο στόχος τους είναι σε μεγάλο βαθμό περιγραφικής φύσης, με βασικό γνώμονα την ανάλυση ενός μοναδικού συνόλου δεδομένων ή ενός πολυπαραγοντικού συνόλου δεδομένων με πεπερασμένα, μετρήσιμα και εξηγήσιμα κατηγορήματα. Έπειτα, χρησιμοποιούνται για την εκτίμηση της μελλοντικής αξίας, συνήθως από ιστορικά αρχεία μετρήσεων της επιχειρηματικής απόδοσης.

Για αυτό λοιπόν οι παραδοσιακές στατιστικές μέθοδοι μπορεί να παρέχουν μια εύλογη ακρίβεια πρόβλεψης με βάση ιστορικά δεδομένα όταν αυτά είναι σταθερά στο χρόνο και λιγότερο περίπλοκα. Λιγότερο περίπλοκα είναι τα δεδομένα όταν ο αριθμός των διαστάσεων που θα μπορούσαν να επηρεάσουν τις τιμές τους είναι πεπερασμένος και μετρήσιμος. Έτσι θα έχουμε ακρίβεια αλλά και παράλληλα έναν καλό βαθμό εξηγησιμότητας και κέρδος σε υπολογιστική ισχύ.

Παρόλα αυτά εμφανίζονται εμπόδια πέρα από τα οποία αυτές οι μέθοδοι πρόβλεψης δεν μπορούν να προβλέπουν με αξιοπιστία. Υπάρχουν πολλά συμβάντα και τιμές που δεν μπορούν να προβλεφθούν με ακρίβεια, επειδή είναι τυχαία γεγονότα και δεν υπάρχει σημαντική σχέση στα δεδομένα. Όταν οι παράγοντες που οδηγούν σε αυτό που προβλέπεται δεν είναι γνωστοί ή καλά κατανοητοί, όπως στις προβλέψεις των χρηματιστηριακών αγορών και των αγορών συναλλάγματος, οι προβλέψεις είναι επίσης συχνά ανακριβείς ή λανθασμένες, καθώς δεν υπάρχουν αρκετά στοιχεία για όλα όσα επηρεάζουν τις μεταβλητές. Ακόμη, σε επιχειρηματικές εφαρμογές με τεράστιες ποσότητες δεδομένων, οι παραδοσιακές τεχνικές προβλέψεων είναι λιγότερο ακριβείς και αποτελεσματικές λόγω του μεγάλου αριθμού των δεδομένων που εμπλέκονται και του γεγονότος ότι ο αλγόριθμος που χρησιμοποιείται μπορεί να μην είναι πολύ γραμμικός ή απλός. Ακόμη, απαιτείται κάτι με σημαντική διαφορά στην ταχύτητα και στην αυτοματοποίηση καθώς και περαιτέρω δυνατότητα εξέλιξης και ανάπτυξης όσο η γνώση πάνω στο αντικείμενο αναπτύσσεται. Για το λόγο αυτό, τα τελευταία χρόνια οι προβλέψεις σε διάφορους τομείς αλλά και στον επιχειρηματικό κόσμο πραγματοποιούνται με τη χρήση νευρωνικών δικτύων.

2.3 Νευρωνικά Δίκτυα στις Προβλέψεις

Η τεχνητή νοημοσύνη και η μηχανική μάθηση μέσω της χρήσης των νευρωνικών δικτύων θεωρούνται τα εργαλεία που μπορούν να φέρουν επανάσταση στις προβλέψεις. Ένας αλγόριθμος τεχνητής νοημοσύνης ο οποίος μπορεί να λαμβάνει υπόψη όλους τους πιθανούς παράγοντες που θα μπορούσαν να επηρεάσουν την πρόβλεψη, προσφέρει σημαντικές δυνατότητες για την εξαγωγή γνώσεων από μαζικά σύνολα δεδομένων που συγκεντρώνονται από οποιονδήποτε αριθμό εσωτερικών και εξωτερικών πηγών. Η εφαρμογή των αλγορίθμων μηχανικής μάθησης στα λεγόμενα προγνωστικά μοντέλα διερευνά τις γνώσεις και προσδιορίζει τις τάσεις που χάνονται από τις παραδοσιακές προβλέψεις ανθρώπινης διαμόρφωσης. Εκτός αυτού, οι αλγόριθμοι μηχανικής μάθησης

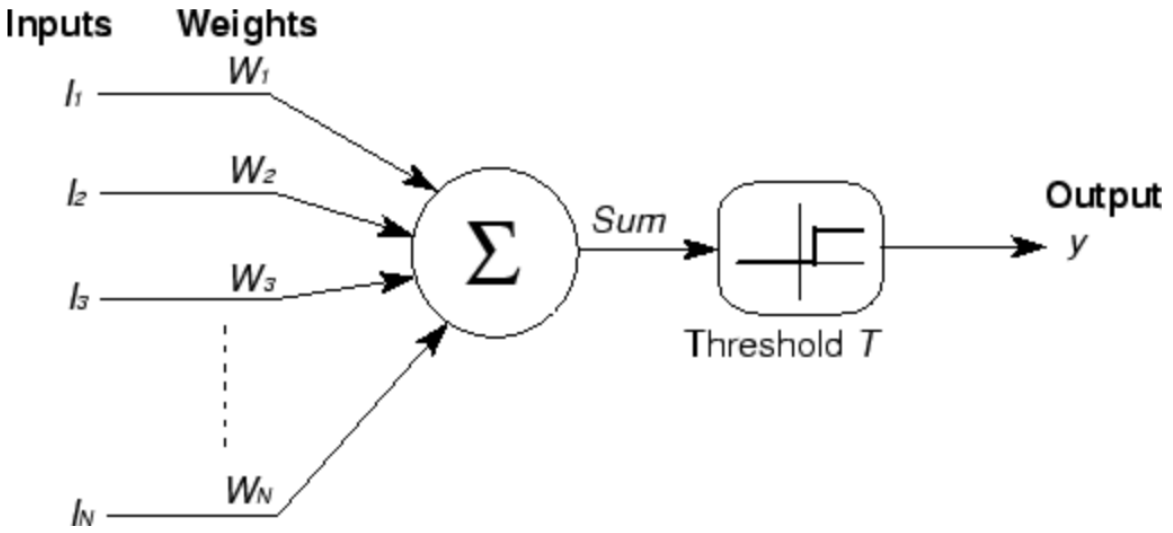
μπορούν ταυτόχρονα να δοκιμάσουν και να μάθουν, βελτιώνοντας συνεχώς εκατοντάδες προηγμένα μοντέλα. Το βέλτιστο μοντέλο μπορεί στη συνέχεια να εφαρμοστεί με μεγάλη ακρίβεια και αποτελεσματικότητα. Για το λόγο αυτό παρακάτω θα αναλυθεί η χρήση των νευρωνικών δικτύων στις προβλέψεις και τα πλεονεκτήματα που έχουν.

2.3.1 Ιστορική Εξέλιξη

Η ιδέα των νευρωνικών δικτύων ξεκίνησε όπως ήταν αναμενόμενο ως ένα μοντέλο του τρόπου αναπαράστασης της λειτουργίας των νευρώνων του εγκεφάλου, που ονομάζεται «συνδετισμός» (connectionism) και βασίζεται στη χρήση συνδεδεμένων κυκλωμάτων για την προσομοίωση ευφυούς συμπεριφοράς. Το 1943, το τελευταίο απεικονίζεται με ένα απλό ηλεκτρικό κύκλωμα από το νευροφυσιολόγο Warren McCulloch και μαθηματικό Walter Pitts. Ο Donald Hebb προχώρησε την ιδέα περαιτέρω στο βιβλίο του, The Organization of Behavior (1949), προτείνοντας ότι οι νευρικές οδοί ενισχύονται σε κάθε διαδοχική χρήση, ειδικά μεταξύ των νευρώνων που τείνουν να πυροδοτούνται ταυτόχρονα, ξεκινώντας έτσι το μακρύ ταξίδι προς την ποσοτικοποίηση των πολύπλοκων διαδικασιών του εγκεφάλου.

Δύο βασικές έννοιες που είναι πρόδρομοι των νευρωνικών δικτύων είναι η «Λογική ορίου» (Threshold Logic), δηλαδή η μετατροπή συνεχούς εισόδου σε διακριτή έξοδο και το «Hebbian Learning», που αποτελεί ένα μοντέλο μάθησης που βασίζεται στη νευρωνική πλαστικότητα, και προτείνεται από τον Donald Hebb στο προαναφερθέν βιβλίο. Οι δύο προτάσεις αυτές διατυπώθηκαν στη δεκαετία του 1940, ενώ κατά τη διάρκεια της δεκαετίας του 1950, και καθώς οι ερευνητές άρχισαν να προσπαθούν να μεταφράσουν αυτά τα δίκτυα σε υπολογιστικά συστήματα, το πρώτο Hebbian δίκτυο υλοποιήθηκε με επιτυχία στο MIT, συγκεκριμένα το 1954.

Εκείνη την εποχή, ο Frank Rosenblatt, ψυχολόγος στο Κορνέλ, εργαζόταν στην κατανόηση των συγκριτικά απλούστερων συστημάτων λήψης αποφάσεων που υπάρχουν στο μάτι μιας μύγας, τα οποία καθορίζουν την απόκρισή της σε φυγή. Σε μια προσπάθεια να κατανοήσει και να ποσοτικοποιήσει αυτήν τη διαδικασία, πρότεινε την ιδέα ενός Perceptron το 1958, το οποίο ονομάζεται Mark I Perceptron. Ήταν ένα σύστημα με μια απλή σχέση εισόδου-εξόδου, που διαμορφώθηκε από έναν νευρώνα McCulloch-Pitts, που προτάθηκε το 1943 από τον Warren S. McCulloch και τον Walter Pitts, για να εξηγήσει τις πολύπλοκες διαδικασίες απόφασης στον εγκέφαλο χρησιμοποιώντας μια γραμμική πύλη ορίου. Ένας νευρώνας McCulloch-Pitts δέχεται εισόδους, παίρνει ένα σταθμισμένο άθροισμα και επιστρέφει «0» αν το αποτέλεσμα είναι κάτω από το όριο και «1» διαφορετικά, όπως φαίνεται και στο σχήμα.



Σχήμα 2.1. Ένας νευρώνας McCulloch-Pitts με Ν εισόδους.

Το μεγάλο πλεονέκτημα του Mark I Perceptron έγκειται στο γεγονός ότι τα βάρη του «μαθαίνονται» με τη διαδοχική μεταβίβαση εισόδων, μειώνοντας παράλληλα τη διαφορά μεταξύ της επιθυμητής και της πραγματικής εξόδου. Μεγάλο μειονέκτημα όμως, είναι ότι αυτός ο τεχνητός νευρώνας μπορούσε να μάθει να διαχωρίζει μόνο γραμμικά διαχωρίσιμες κλάσεις, κάνοντας το απλό αλλά μη γραμμικό exclusive-or κύκλωμα ανυπέρβλητο εμπόδιο. Μετά το χρονικό διάστημα αυτό, και περίπου το 1959 στο Στάνφορντ, ο Bernard Widrow και ο Marcian Hoff ανέπτυξαν το πρώτο νευρωνικό δίκτυο το οποίο εφαρμόστηκε με επιτυχία σε ένα πραγματικό πρόβλημα. Τα συστήματα αυτά ονομάστηκαν ADALINE και MADALINE από τη χρήση των πολλαπλών ADAptive LINear Elements, το τελευταίο εκ των οποίων σχεδιάστηκε ειδικά για την εξάλειψη του θορύβου στις τηλεφωνικές γραμμές και εξακολουθεί να χρησιμοποιείται και σήμερα.

Όλα αυτά τελείωσαν το 1969 με την έκδοση ενός βιβλίου «Perceptrons» από τον Marvin Minsky, ιδρυτή του MIT AI Lab, και τον Seymour Papert, διευθυντή του εργαστηρίου. Το βιβλίο υποστήριξε οριστικά ότι η προσέγγιση απλής αντίληψης (single perception) του Rosenblatt στα νευρωνικά δίκτυα δεν μπορούσε να μεταφραστεί αποτελεσματικά σε πολυεπίπεδα νευρωνικά δίκτυα. Για να αξιολογηθούν οι σωστές σχετικές τιμές των βαρών των νευρώνων που απλώνονται σε στρώματα με βάση την τελική έξοδο θα χρειαζόταν αρκετές, αν όχι άπειρες, επαναλήψεις και θα χρειαζόταν πολύ χρόνο για να υπολογιστεί. Το αποτέλεσμα αυτών των θεωριών ήταν η διακοπή της έρευνας για τα επόμενα 10-12 χρόνια, και η αρχή μιας εποχής που αναφέρεται πλέον ως «ο χειμώνας της τεχνητής νοημοσύνης».

Έτσι οδηγηθήκαμε στην εκ νέου ανακάλυψη μιας έννοιας που υπήρχε ήδη από τη δεκαετία του '60 και η οποία βοήθησε τα νευρωνικά δίκτυα να ξεφύγουν από το παραπάνω αδιέξοδο. Η οπισθοδιάδοση (backpropagation), μια μέθοδος που επινοήθηκε από τους ερευνητές από τη δεκαετία του '60 και αναπτύχθηκε συνεχώς μέχρι και τον "χειμώνα της τεχνητής

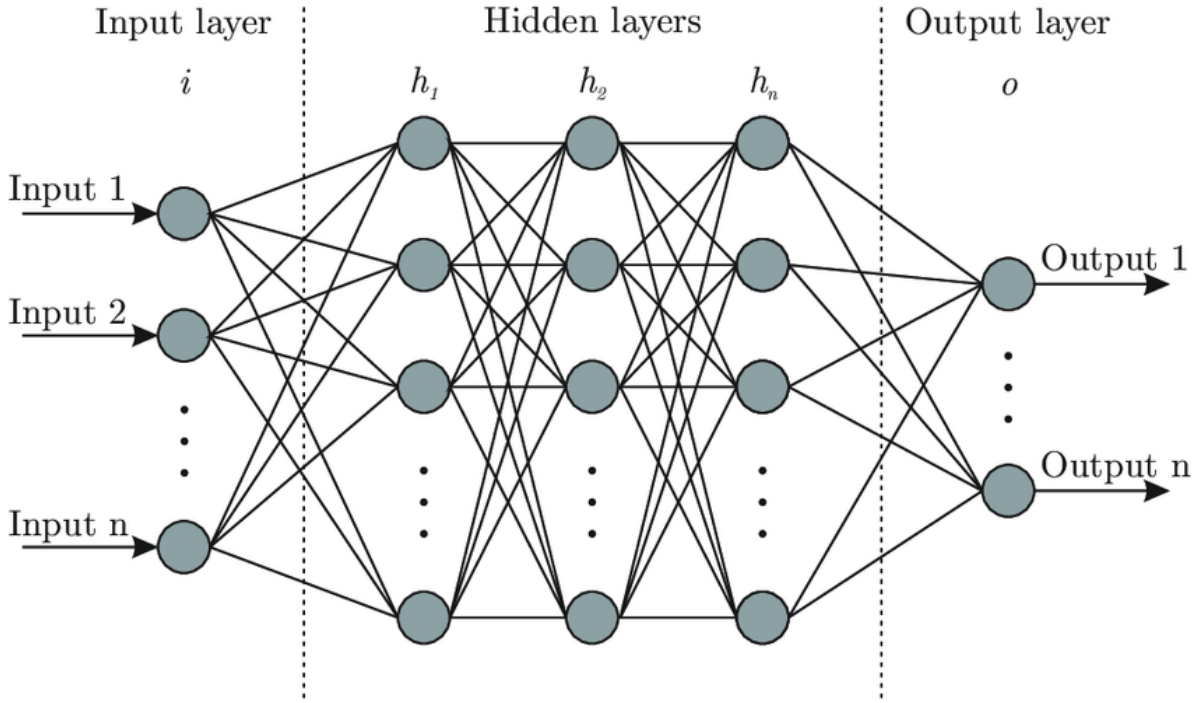
νοημοσύνης, ήταν μια διαισθητική μέθοδος που αποδίδει τη μείωση της σημασίας σε κάθε γεγονός καθώς πηγαίνουμε πιο μακριά στην αλυσίδα των γεγονότων. Η μέθοδος αυτή έγινε δεκτή χρόνια μετά και μετά από αρκετά ερευνητικά έργα που γράφτηκαν μέχρι και το 1985.

Τέλος από τη δεκαετία του 1990, τα νευρωνικά δίκτυα επέστρεψαν οριστικά, αυτήν τη φορά πραγματικά αιχμαλωτίζοντας τη φαντασία του κόσμου και τελικά κατακτώντας, αν όχι προσπερνώντας, τις προσδοκίες του και προσφέροντας πολυάριθμα οφέλη σε ποικίλους επιστημονικούς τομείς. Ταυτόχρονα με την αύξηση των υπολογιστικών πόρων, η σημερινή εποχή χαρακτηρίζεται από μια εκρηκτική ανάπτυξη των τεχνητών νευρωνικών δικτύων με χρήση των οποίων μοντελοποιούμε πλέον σχέσεις που δεν μπορούν να περιγραφούν από τα συμβατικά αναλυτικά μοντέλα λόγω ελλιπών γνώσεων για το εξεταζόμενο σύστημα και φυσικά πολυπλοκότητας των σχέσεων.

2.3.2. Τρόπος Λειτουργίας Νευρωνικών Δικτύων

Τα νευρωνικά δίκτυα, που συνθέτουν τους αλγορίθμους βαθιάς μάθησης (deep learning algorithms), στοχεύουν στην επεξεργασία δεδομένων εκπαίδευσης (training data) μεσω της μίμηση της διασύνδεσης του ανθρώπινου εγκεφάλου που αποτελείται από επίπεδα κόμβων. Κάθε κόμβος (node) αποτελείται από εισόδους (inputs), βάρη (weights), ένα κατώφλι (bias or threshold), και μια έξοδο (output). Αν αυτή η τιμή εξόδου υπερβαίνει ένα δεδομένο όριο, «πυροδοτεί» ή ενεργοποιεί τον κόμβο, μεταβιβάζοντας δεδομένα στο επόμενο επίπεδο του δικτύου. Τα νευρωνικά δίκτυα μαθαίνουν αυτήν τη συνάρτηση χαρτογράφησης μέσω της επιβλεπόμενης μάθησης (supervised learning), προσαρμόζοντας με βάση τη λειτουργία απώλειας (loss function) και μέσω της διαδικασίας gradient descent. Όταν η συνάρτηση κόστους είναι ακριβώς ή κοντά στο μηδέν, μπορούμε να είμαστε βέβαιοι για την ακρίβεια του μοντέλου να δώσουμε τη έξοδο.

Το απλούστερο νευρωνικό δίκτυο είναι ένα πλήρως συνδεδεμένο μοντέλο που αποτελείται από μια σειρά από πλήρως συνδεδεμένα επίπεδα. Σε ένα πλήρως συνδεδεμένο επίπεδο κάθε νευρώνας συνδέεται με κάθε νευρώνα στο προηγούμενο επίπεδο, και κάθε σύνδεση έχει το δικό της βάρος, όπως φαίνεται και στο σχήμα 2.2. Ένα τέτοιο μοντέλο μοιάζει με ένα απλό μοντέλο παλινδρόμησης που δέχεται μία είσοδο και βγάζει μία έξοδο. Τέτοια μοντέλα επαναλαμβάνουν τις προηγούμενες τιμές με ελαφρά μετατόπιση. Ωστόσο, τα πλήρως συνδεδεμένα μοντέλα δεν είναι σε θέση να προβλέψουν το μέλλον από τη μοναδική προηγούμενη τιμή.



Σχήμα 2.2. Μια βασική αναπαράσταση νευρωνικού δικτύου.

Τα πρόσφατα εισαχθέντα επαναλαμβανόμενα νευρωνικά δίκτυα αντιμετωπίζουν προβλήματα ακολουθίας. Μπορούν να διατηρούν μια κατάσταση από μια επανάληψη στην επόμενη, χρησιμοποιώντας τη δική τους έξοδο ως είσοδο για το επόμενο βήμα. Όσον αφορά τον προγραμματισμό, αυτό μοιάζει με την εκτέλεση ενός σταθερού προγράμματος με ορισμένες εισόδους και μερικές εσωτερικές μεταβλητές. Τέτοια μοντέλα μπορούν να μάθουν να αναπαράγουν το ετήσιο σχήμα των δεδομένων και δεν έχουν την καθυστέρηση που σχετίζεται με ένα απλό πλήρως συνδεδεμένο τροφοδότη νευρωνικό δίκτυο.

2.3.3. Πλεονεκτήματα Νευρωνικών Δικτύων

Με την πρόσφατη εμφάνιση των τεχνικών της Μηχανικής Μάθησης, τα νευρωνικά δίκτυα έχουν δει σημαντικές βελτιώσεις όσον αφορά την ακρίβεια και την ικανότητα αντιμετώπισης των πιο εξελιγμένων και πολύπλοκων εργασιών. Τα τελευταία μπορούν να παράγουν αποτελέσματα εξόδου από πολύπλοκα ή ανακριβή δεδομένα και χρησιμοποιούνται για τον εντοπισμό μοτίβων και τάσεων στα δεδομένα, τα οποία δεν είναι εύκολα ανιχνεύσιμα ούτε από τον άνθρωπο ούτε από τις μηχανές. Μπορούμε να κάνουμε χρήση των νευρωνικών σε οποιοδήποτε επιστημονικό κλάδο, καθώς είναι πολύ ευέλικτα και επίσης δεν απαιτούν πολύπλοκους αλγορίθμους.

Η πιο σημαντική διαφορά μεταξύ των παραδοσιακών μεθόδων και των μεθόδων μηχανικής μάθησης έγκειται στον τρόπο ελαχιστοποίησης. Αν και οι περισσότερες παραδοσιακές μέθοδοι χρησιμοποιούν εξηγήσιμες γραμμικές διεργασίες, οι περισσότερες μέθοδοι μηχανικής μάθησης χρησιμοποιούν μη γραμμικές τεχνικές για την ελαχιστοποίηση των συναρτήσεων απώλειας. Και ενώ οι μέθοδοι μηχανικής μάθησης είναι υπολογιστικά πιο απαιτητικές από τις στατιστικές, οι δυνατότητες που υπάρχουν σχετικά με την διαχείριση χιλιάδων παραγόντων για την εξαγωγή προβλέψεων κάνει τους αλγορίθμους μηχανικής μάθησης να ξεπερνούν τις στατιστικές μεθόδους. Ένα σύνολο διαφορετικών τεχνικών πρόβλεψης -τόσο γραμμικών όσο και μη γραμμικών- μπορεί να συνδυαστεί για να επιτευχθεί η μεγαλύτερη ακρίβεια.

Καθώς η τεχνολογία εξελίσσεται, οι εφαρμογές αυτές θα συνεχίσουν να φέρνουν επανάσταση στον επιχειρηματικό κόσμο. Για το λόγο αυτό, η χρήση της τεχνητής νοημοσύνης στην πρόβλεψη παρουσιάζει τεράστιο ενδιαφέρον για τις περισσότερες επιχειρήσεις λόγω της χρηστικότητας τους σε όλες τις λειτουργίες καθώς κερδίζουν σε χρόνο, προσπάθεια, αλλά και κόστος και οι προβλέψεις που βασίζονται στην μηχανική μάθηση έχουν αντικαταστήσει τις παραδοσιακές μεθόδους σε πολλές διαδικασίες ανάλυσης δεδομένων σε πολλούς διαφορετικούς κλάδους και τομείς.

2.3.4. Αρχιτεκτονικές Νευρωνικών Δικτύων

Για να γίνει πιο σαφής η χρήση των νευρωνικών δικτύων και για να καθορίσουμε μετέπειτα και τον αλγόριθμο μάθησης για το κάθε είδος νευρωνικού, είναι σημαντικό να παρατηρήσουμε τον τρόπο δόμησης των νευρώνων του. Σύμφωνα με αυτόν μπορούμε να κατηγοριοποιήσουμε τα νευρωνικά και να διατυπώσουμε συγκεκριμένους κανόνες μάθησης. Ενδεικτικά θα αναφερθούμε στις πιο σημαντικές κατηγορίες αρχιτεκτονικών νευρωνικών δικτύων παρακάτω:

Τεχνητός Νευρώνας (Perceptron)

Ο τεχνητός νευρώνας είναι το πιο βασικό από όλα τα νευρωνικά δίκτυα, καθώς είναι ένα θεμελιώδες δομικό στοιχείο για πιο σύνθετα νευρωνικά δίκτυα. Απλώς συνδέει ένα κελί εισόδου και ένα κελί εξόδου.

Δίκτυα Πρόσθιας Τροφοδότησης (Feed-Forward Networks)

Το δίκτυο πρόσθιας τροφοδοσίας είναι μια συλλογή από perceptrons, στην οποία υπάρχουν τρεις βασικοί τύποι στρώσεων, το επίπεδο εισόδου, το κρυφό επίπεδο, και το επίπεδο εξόδου. Κατά τη διάρκεια κάθε σύνδεσης, το σήμα από το προηγούμενο επίπεδο πολλαπλασιάζεται με ένα βάρος, προστίθεται σε μια πόλωση, και περνά μέσω μιας λειτουργίας ενεργοποίησης. Τα δίκτυα προώθησης χρησιμοποιούν την οπισθόζευξη για την επαναλαμβανόμενη ενημέρωση των παραμέτρων μέχρι να επιτευχθεί η επιθυμητή απόδοση.

Υπολειπόμενα Νευρωνικά Δίκτυα (Residual Networks - ResNet)

Ένα πρόβλημα με τα νευρωνικά δίκτυα τροφοδοσίας σε βάθος ονομάζεται vanishing gradient problem, σύμφωνα με το οποίο τα δίκτυα είναι πολύ μεγάλα για να μπορούν να διαδίδονται χρήσιμες πληροφορίες σε ολόκληρο το δίκτυο. Καθώς το σήμα που ενημερώνει τις παραμέτρους ταξιδεύει μέσω του δικτύου, μειώνεται σταδιακά μέχρι τα βάρη στο μπροστινό μέρος του δικτύου να μην αλλάζουν ή να μην χρησιμοποιούνται καθόλου.
Για την αντιμετώπιση αυτού του προβλήματος, ένα δίκτυο με υπολειπόμενη υποδομή χρησιμοποιεί υπερπηδούμενες συνδέσεις (skip connections), οι οποίες διαδίδουν τα σήματα σε ένα επίπεδο που ονομάζεται jumped layer. Αυτό μειώνει το πρόβλημα χρησιμοποιώντας συνδέσεις που είναι λιγότερο ευάλωτες σε αυτό. Με την πάροδο του χρόνου, το δίκτυο μαθαίνει να αποκαθιστά τα επίπεδα που παραλείπονται καθώς μαθαίνει τον χώρο των χαρακτηριστικών, αλλά και είναι πιο αποδοτικό στην εκπαίδευση αφού είναι λιγότερο ευάλωτο στο vanishing gradient problem και πρέπει να εξερευνά λιγότερο χώρο χαρακτηριστικών.

Επαναλαμβανόμενα νευρωνικά δίκτυα (Recurrent Neural Networks - RNN)

Ένα επαναλαμβανόμενο νευρωνικό δίκτυο είναι ένας εξειδικευμένος τύπος δικτύου που περιέχει βρόγχους, και "επαναλαμβάνει τον εαυτό του". Επιτρέποντας την αποθήκευση πληροφοριών στο δίκτυο, το RNN χρησιμοποιεί τη λογική από την προηγούμενη εκπαίδευση για να κάνει καλύτερες και πιο ενημερωμένες αποφάσεις για τα επερχόμενα γεγονότα. Λόγω της φύσης τους, τα RNNs χρησιμοποιούνται συνήθως για να χειρίζονται διαδοχικές εργασίες, όπως η πρόβλεψη δεδομένων χρονοσειρών. Μπορούν επίσης να χειρίζονται δεδομένα εισόδου οποιουδήποτε μεγέθους.

2.4 Επιβλεπόμενη Μηχανική Μάθηση

Με βάση τα παραπάνω, λοιπόν, ορίζουμε τη διαδικασία της μηχανικής μάθησης. Η Μηχανική μάθηση (Machine Learning - ML) είναι η μελέτη αλγορίθμων υπολογιστών που μπορούν να βελτιωθούν αυτόματα μέσω της εμπειρίας και με τη χρήση δεδομένων. Θεωρείται μέρος της τεχνητής νοημοσύνης. Οι αλγόριθμοι μηχανικής μάθησης (machine learning algorithms) κατασκευάζουν ένα μοντέλο το οποίο βασίζεται σε δείγματα δεδομένων, γνωστά ως «δεδομένα εκπαίδευσης» (training data), προκειμένου να κάνουν προβλέψεις ή αποφάσεις χωρίς να προγραμματίζονται ρητά για να το κάνουν.

Μπορούμε να κατηγοριοποιήσουμε τις διαδικασίες μάθησης μέσω των οποίων λειτουργούν τα νευρωνικά δίκτυα ως εξής: επιβλεπόμενη μάθηση ή εναλλακτικά μάθηση με εκπαιδευτή (supervised learning) και μη επιβλεπόμενη μάθηση ή μάθηση χωρίς εκπαιδευτή (unsupervised learning). Στην παρούσα εργασία θα επικεντρωθούμε στην επιβλεπόμενη μάθηση την οποία θα αναλύσουμε στα επόμενα υποκεφάλαια.

2.4.1. Τρόπος Λειτουργίας

Όπως αναφέρθηκε παραπάνω, η επιβλεπόμενη μηχανική μάθηση, είναι μια υποκατηγορία της μηχανικής μάθησης και της τεχνητής νοημοσύνης. Ορίζεται από τη χρήση συνόλων δεδομένων με ετικέτες για την εκπαίδευση αλγορίθμων που ταξινομούν αυτά τα δεδομένα ή προβλέπουν αποτελέσματα με ακρίβεια. Καθώς τα δεδομένα εισόδου εισάγονται στο μοντέλο, το τελευταίο προσαρμόζει τα βάρη του μέχρι να γίνει το κατάλληλο fitting, κάτι που συμβαίνει ως μέρος της διαδικασίας διασταυρούμενης επικύρωσης (cross validation). Πιο συγκεκριμένα, το σύνολο δεδομένων εκπαίδευσης περιλαμβάνει δεδομένα - εισόδους και τις αντίστοιχες σωστές εξόδους, οι οποίες επιτρέπουν στο μοντέλο να μαθαίνει με την πάροδο του χρόνου. Ο αλγόριθμος μετρά την ακρίβειά του μέσω της συνάρτησης απώλειας, επιδεχόμενος προσαρμογές και αλλαγές μέχρι το σφάλμα να ελαχιστοποιηθεί επαρκώς.
Η επιβλεπόμενη μάθηση βοηθά στην επίλυση διαφόρων προβλημάτων του εμπορικού και όχι μόνο κόσμου καθώς χρησιμοποιεί ένα σύνολο εκπαίδευσης για να διδάσκει μοντέλα τα οποία επιτυγχάνουν τα επιθυμητά αποτελέσματα με μεγάλη ακρίβεια τις περισσότερες φορές.

Ωστόσο, όπως είδαμε εμφανίζεται και η μη επιβλεπόμενη μηχανική μάθηση που συναντάται συχνά μαζί με την επιβλεπόμενη. Σε αντίθεση με την επιβλεπόμενη μάθηση, η μη επιβλεπόμενη μάθηση χρησιμοποιεί δεδομένα χωρίς ετικέτες. Από αυτά τα δεδομένα ανακαλύπτει μοτίβα που βοηθούν στην επίλυση προβλημάτων συσταδοποίησης ή συσχέτισης (clustering or association problems). Αυτό είναι ιδιαίτερα χρήσιμο όταν δεν υπάρχουν επαρκή στοιχεία και βεβαιότητα για τις κοινές ιδιότητες ενός συνόλου δεδομένων. Οι συνηθισμένοι αλγόριθμοι συσταδοποίησης είναι ιεραρχικά μοντέλα (hierarchical), μοντέλα k-means, και μοντέλα Gaussian mixture. Τέλος, η μάθηση με ημιεπίβλεψη πραγματοποιείται όταν έχει ετικέτες μόνο το ένα μέρος των δεδομένων εισόδου. Η μη επιβλεπόμενη και η ημιεπιβλεπόμενη μάθηση μπορεί να είναι πιο ελκυστικές εναλλακτικές λύσεις, καθώς μπορεί να είναι χρονοβόρα και δαπανηρή η διαδικασία αντιστοίχισης ετικετών στα δεδομένα με σκοπό την επιβλεπόμενη μάθηση.

Εφόσον θα επικεντρωθούμε στην επιβλεπόμενη μάθηση είναι σημαντικό να την χωρίσουμε σε δύο τύπους προβλημάτων, την ταξινόμηση (classification) και την παλινδρόμηση (regression).

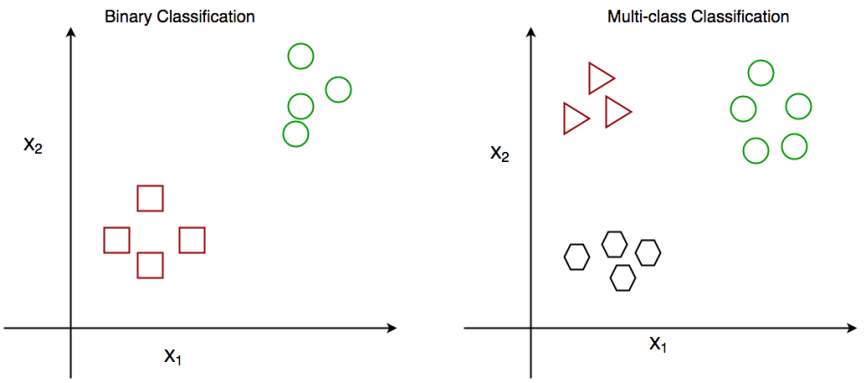
Ταξινόμηση (Classification)

Στην ταξινόμηση χρησιμοποιείται ένας αλγόριθμος που αντιστοιχίζει με ακρίβεια τα πειραματικά δεδομένα (test data) με συγκεκριμένες κατηγορίες. Αναγνωρίζει συγκεκριμένες οντότητες μέσα στο σύνολο δεδομένων και προσπαθεί να βγάλει κάποια συμπεράσματα για το πώς αυτές οι οντότητες θα πρέπει να επισημαίνονται ή να ορίζονται. Οι κοινοί αλγόριθμοι ταξινόμησης είναι γραμμικοί ταξινομητές (linear classifiers), διανυσματικές μηχανές υποστήριξης (support vector machines - SVM), δέντρα αποφάσεων (decision

trees), k-πλησιέστερος γείτονας (k-nearest neighbor), και τυχαίο δάσος (random forest), τα οποία περιγράφονται με περισσότερες λεπτομέρειες παρακάτω.

Στην ταξινόμηση, υπάρχει πάντα ο κατηγοριοποιητής (classifier), δηλαδή ο αλγόριθμος που χρησιμοποιείται για την αντιστοίχιση των δεδομένων εισόδου σε μια συγκεκριμένη κατηγορία. Ακόμη, εμφανίζεται το μοντέλο ταξινόμησης (classification model) δηλαδή το μοντέλο που προβλέπει ή εξάγει το συμπέρασμα στα δεδομένα εισόδου που δίνονται για εκπαίδευση, άρα πιο συγκεκριμένα προβλέπει την τάξη ή την κατηγορία για τα δεδομένα και το χαρακτηριστικό (feature), που αποτελεί μια ατομική μετρήσιμη ιδιότητα του φαινομένου που παρατηρείται.

Ακόμη, η ταξινόμηση χωρίζεται σε διαφορετικά είδη. Υπάρχει η δυαδική ταξινόμηση (binary classification) που περιλαμβάνει δύο αποτελέσματα (π.χ. αληθές ή ψευδές), η ταξινόμηση πολλαπλών κατηγοριών (Multi-class Classification), δηλαδή εκείνη με περισσότερες από δύο κλάσεις όπου κάθε δείγμα αποδίδεται σε μία και μόνο μία ετικέτα ή στόχο, όπως φαίνεται στο σχήμα 2.3. Αντίθετα στην ταξινόμηση πολλαπλών ετικετών κάθε δείγμα αποδίδεται σε ένα σύνολο ετικετών ή στόχων.



Σχήμα 2.3. Αναπαράσταση δυαδικής ταξινόμησης και ταξινόμησης πολλαπλών κατηγοριών.

Παλινδρόμηση (Regression)

Η παλινδρόμηση (regression) χρησιμοποιείται για να κατανοηθεί η σχέση μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Χρησιμοποιείται συνήθως για να κάνει προβλέψεις ή να βρει σχέσεις αιτίας - αποτελέσματος. Ο αλγόριθμος παλινδρόμησης μπορεί να διαιρεθεί περαιτέρω σε γραμμική (linear regression) και μη γραμμική παλινδρόμηση (non linear regression). Δημοφιλείς αλγόριθμοι παλινδρόμησης είναι λοιπόν η γραμμική παλινδρόμηση (linear regression), η λογιστική παλινδρόμηση (logistic regression) και η πολυωνυμική παλινδρόμηση (polynomial regression).

Τα μοντέλα γραμμικής παλινδρόμησης μπορούν να ενσωματώνουν τόσο συνεχείς όσο και κατηγορικές μεταβλητές πρόβλεψης και μπορούν να έχουν μία ή πολλές μεταβλητές πρόβλεψης, ενώ κάποιες να υπάρχουν στο μοντέλο αλλά να μην επηρεάζουν τη μεταβλητή στόχο. Ακόμη, σε ορισμένες περιπτώσεις, η σχέση μεταξύ του αποτελέσματος και των μεταβλητών πρόβλεψης μπορεί να είναι γραμμική, πολυωνυμική ή ακόμη και να εμφανίζεται σε μορφή καμπύλης.

2.4.2. Αλγόριθμοι Επιβλεπόμενης Μάθησης

Παρακάτω παρατίθενται κάποιοι από τους βασικότερους αλγορίθμους επιβλεπόμενης μηχανικής μάθησης, οι οποίοι χρησιμοποιούνται είτε για ταξινόμηση είτε για παλινδρόμηση είτε και για τα δύο καθώς και ο τρόπος λειτουργίας τους:

Naive Bayes

Ο Naive Bayes είναι μια προσέγγιση ταξινόμησης που υιοθετεί την αρχή της ταξικής εξαρτώμενης ανεξαρτησίας από το θεώρημα Bayes. Αυτό σημαίνει ότι η παρουσία ενός χαρακτηριστικού δεν επηρεάζει την παρουσία ενός άλλου στην πιθανότητα ενός δεδομένου αποτελέσματος, και κάθε προβλεπτικός παράγοντας έχει ίση επίδραση σε αυτό το αποτέλεσμα. Υπάρχουν τρεις τύποι ταξινομητών Naïve Bayes: Multinomial Naïve Bayes, Bernoulli Naïve Bayes και Gaussian Naïve Bayes. Αυτή η τεχνική χρησιμοποιείται κυρίως σε συστήματα ταξινόμησης κειμένου (text classification), αναγνώρισης ανεπιθύμητων μηνυμάτων (spam identification) και συστήματα συστάσεων (recommendation systems).

Ο Naive Bayes χρησιμοποιείται κυρίως για ταξινόμηση που απαιτεί μια μικρή ποσότητα δεδομένων εκπαίδευσης για να εκτιμήσει τις απαραίτητες παραμέτρους και να εξάγει τα αποτελέσματα. Είναι εξαιρετικά γρήγορος σε σύγκριση με άλλους ταξινομητές. Το μόνο μειονέκτημα είναι ότι δεν έχει ευελιξία σχετικά με την επιλογή παραγόντων αφού τους χρησιμοποιεί όλους ανεξαρτήτως κριτηρίων και με προκαθορισμένο τρόπο.

Ιδιαίτερα δημοφιλής είναι ο Gaussian Naive Bayes που αποτελεί μια παραλλαγή του Naive Bayes αλλά ακολουθεί Gaussian κανονική κατανομή και υποστηρίζει συνεχή δεδομένα.

Γραμμική παλινδρόμηση (Linear Regression)

Η γραμμική παλινδρόμηση χρησιμοποιείται για να προσδιοριστεί η σχέση μεταξύ μιας εξαρτημένης μεταβλητής και μίας ή περισσότερων ανεξάρτητων μεταβλητών και συνήθως χρησιμοποιείται για να κάνει προβλέψεις για μελλοντικά αποτελέσματα. Όταν υπάρχει μόνο μία ανεξάρτητη μεταβλητή και μία εξαρτημένη μεταβλητή, είναι γνωστή ως απλή γραμμική παλινδρόμηση (simple linear regression). Καθώς ο αριθμός των ανεξάρτητων μεταβλητών αυξάνεται, αναφέρεται ως πολλαπλή γραμμική παλινδρόμηση (multiple linear regression). Για κάθε τύπο γραμμικής παλινδρόμησης, ο αλγόριθμος επιδιώκει να απεικονίσει μια

γραμμή βέλτιστης προσαρμογής, η οποία υπολογίζεται με τη μέθοδο των ελαχίστων τετραγώνων.

Η γραμμική παλινδρόμηση χρησιμοποιείται πιο αποδοτικά για παλινδρόμηση. Η μέθοδος αυτή είναι απλή στην υλοποίηση και ευκολότερη στην ερμηνεία των συντελεστών εξόδου. Από την άλλη πλευρά, στην τεχνική της γραμμικής παλινδρόμησης οι ακραίες τιμές μπορούν να έχουν τεράστιες επιπτώσεις στην παλινδρόμηση και τα όρια είναι γραμμικά σε αυτήν την τεχνική. Όταν είναι γνωστό ότι η σχέση μεταξύ ανεξάρτητης και εξαρτημένης μεταβλητής είναι γραμμική, ο αλγόριθμος αυτός είναι πολύ αποδοτικός και λιγότερο πολύπλοκος σε σύγκριση με άλλους αλγόριθμους. Από την άλλη, στις περιπτώσεις που αυτό δεν ισχύει εμφανίζει σημαντικό μειονέκτημα.

Λογιστική παλινδρόμηση (Logistic Regression)

Ενώ η γραμμική παλινδρόμηση προτιμάται όταν οι εξαρτημένες μεταβλητές είναι συνεχείς, η λογιστική παλινδρόμηση επιλέγεται όταν η εξαρτημένη μεταβλητή είναι κατηγορηματική, που σημαίνει ότι έχει δυαδικές εξόδους, όπως "αληθές" και "ψευδές" ή "ναι" και "όχι". Ενώ και τα δύο μοντέλα παλινδρόμησης επιδιώκουν να κατανοήσουν τις σχέσεις μεταξύ των δεδομένων εισόδου, η λογιστική παλινδρόμηση χρησιμοποιείται κυρίως για την επίλυση προβλημάτων δυαδικής ταξινόμησης.

Η λογιστική παλινδρόμηση είναι χρήσιμη για την κατανόηση του τρόπου με τον οποίο ένα σύνολο ανεξάρτητων μεταβλητών επηρεάζουν το αποτέλεσμα της εξαρτημένης μεταβλητής για αυτό είναι παραπάνω χρήσιμη για την ταξινόμηση. Το κύριο μειονέκτημα του αλγορίθμου λογιστικής παλινδρόμησης είναι ότι υποθέτει ότι τα δεδομένα είναι απαλλαγμένα από ελλείπουσες τιμές και ότι τα κατηγορήματα είναι ανεξάρτητα μεταξύ τους.

Παλινδρόμηση Lasso (Lasso Regression)

Η παλινδρόμηση Lasso είναι ένας τύπος γραμμικής παλινδρόμησης που χρησιμοποιεί συρρίκνωση (shrinkage). Συρρίκνωση είναι το σημείο όπου οι τιμές δεδομένων συρρικνώνονται προς ένα κεντρικό σημείο, όπως ο μέσος.

Η διαδικασία αυτή ενθαρρύνει απλά, λιγότερο πολύπλοκα μοντέλα (δηλαδή μοντέλα με λιγότερες παραμέτρους). Αυτός ο συγκεκριμένος τύπος παλινδρόμησης είναι κατάλληλος για μοντέλα που εμφανίζουν υψηλά επίπεδα πολυπολιτισμικότητας ή όταν είναι επιθυμητή η αυτοματοποίηση της επιλογής των μεταβλητών ενός μοντέλου.

Μηχανή διανυσματικής υποστήριξης (SVM)

Μια μηχανή διανυσματικής υποστήριξης είναι ένα δημοφιλές μοντέλο επιβλεπόμενης μάθησης που αναπτύχθηκε από τον Vladimir Vapnik, που χρησιμοποιείται τόσο για την

ταξινόμηση δεδομένων όσο και για παλινδρόμηση. Συνήθως, βέβαια, χρησιμοποιείται για προβλήματα ταξινόμησης, κατασκευάζοντας ένα υπερεπίπεδο όπου η απόσταση μεταξύ δύο κλάσεων με σημεία δεδομένων είναι η μέγιστη. Αυτό το υπερεπίπεδο είναι γνωστό ως όριο απόφασης (decision boundary) και χωρίζει τις κλάσεις των σημείων δεδομένων σε κάθε πλευρά του επιπέδου.

Ο αλγόριθμος SVM λειτουργεί σχετικά καλά όταν υπάρχει σαφές περιθώριο διαχωρισμού μεταξύ κλάσεων, σε χώρους υψηλής διάστασης και σε περιπτώσεις όπου ο αριθμός των διαστάσεων είναι μεγαλύτερος από τον αριθμό των δειγμάτων. Παρόλα αυτά δεν είναι κατάλληλος για μεγάλα σύνολα δεδομένων, στις περιπτώσεις που οι κλάσεις προορισμού επικαλύπτονται, καθώς και σε περιπτώσεις όπου ο αριθμός των χαρακτηριστικών για κάθε σημείο δεδομένων υπερβαίνει τον αριθμό των δειγμάτων δεδομένων εκπαίδευσης. Καθώς ο ταξινομητής διανυσμάτων υποστήριξης λειτουργεί τοποθετώντας σημεία δεδομένων πάνω και κάτω από το ταξινομητικό υπερεπίπεδο, δεν υπάρχει πιθανολογική εξήγηση για την ταξινόμηση.

Κ πλησιέστερος γείτονας (K-nearest neighbor)

Ο Κ-πλησιέστερος γείτονας, επίσης γνωστός ως αλγόριθμος KNN, είναι ένας μη παραμετρικός αλγόριθμος που ταξινομεί τα σημεία δεδομένων με βάση την εγγύτητά τους και τη συσχέτισή τους με άλλα διαθέσιμα δεδομένα. Ο αλγόριθμος αυτός υποθέτει ότι παρόμοια σημεία δεδομένων μπορούν να βρεθούν το ένα κοντά στο άλλο. Ως αποτέλεσμα, επιδιώκει να υπολογίσει την απόσταση μεταξύ των σημείων δεδομένων, συνήθως μέσω της Ευκλείδειας απόστασης, και στη συνέχεια τα αναθέτει σε μια κατηγορία με βάση την πιο συχνή κατηγορία ή το μέσο όρο.

Η ευκολία χρήσης και ο χαμηλός χρόνος υπολογισμού του τον καθιστούν προτιμώμενο αλγόριθμο από τους επιστήμονες δεδομένων, αλλά όσο το σύνολο δεδομένων δοκιμής (test dataset) μεγαλώνει, ο χρόνος επεξεργασίας παρατείνεται, καθιστώντας το λιγότερο ελκυστικό για εργασίες ταξινόμησης. Το KNN χρησιμοποιείται συνήθως για μηχανές συστάσεων (recommendation engines) και αναγνώριση εικόνων (image recognition).

Δέντρα Αποφάσεων (Decision Tress)

Ένα δέντρο αποφάσεων παράγει μια ακολουθία κανόνων που μπορούν να χρησιμοποιηθούν για την ταξινόμηση των δεδομένων, όταν δοθούν δεδομένα για ιδιότητες μαζί με τις κλάσεις τους.

Το δέντρο αποφάσεων είναι απλό να κατανοηθεί και να οπτικοποιηθεί, απαιτεί λίγη προετοιμασία δεδομένων και μπορεί να χειριστεί τόσο αριθμητικά όσο και κατηγορηματικά δεδομένα. Παρόλα αυτά, μπορεί να δημιουργήσει πολύπλοκα δέντρα που δεν γενικεύονται καλά, και τα δέντρα αποφάσεων μπορεί να είναι ασταθή επειδή μικρές παραλλαγές στα

δεδομένα μπορεί να έχουν ως αποτέλεσμα τη δημιουργία ενός εντελώς διαφορετικού δέντρου.

Τυχαίο δάσος (Random Forest)

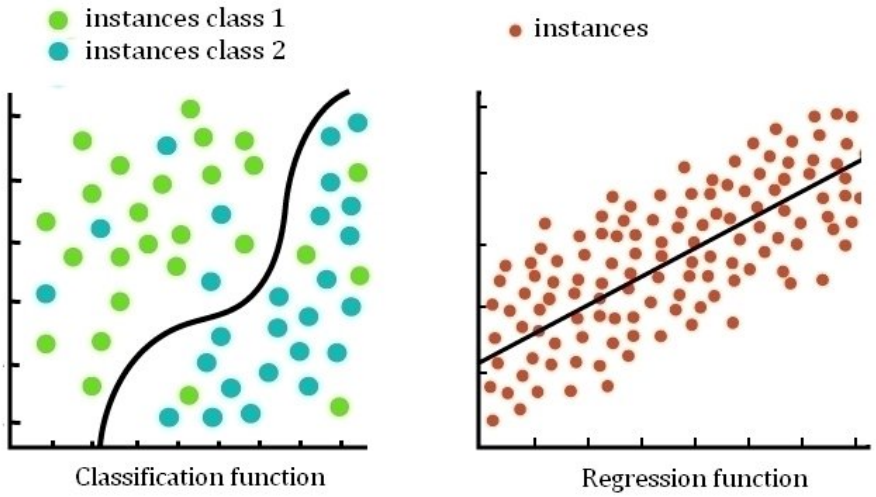
Το τυχαίο δάσος είναι ένας άλλος αλγόριθμος μηχανικής επιβλεπόμενης μάθησης που χρησιμοποιείται τόσο για σκοπούς ταξινόμησης όσο και παλινδρόμησης. Το «δάσος» αναφέρεται σε μια συλλογή από ασύνδετα δέντρα αποφάσεων, τα οποία στη συνέχεια συγχωνεύονται για να μειώσουν τη διακύμανση και να δημιουργήσουν πιο ακριβείς προβλέψεις δεδομένων.

Το πλεονέκτημα του τυχαίου δάσους είναι ότι είναι πιο ακριβές από τα δέντρα απόφασης λόγω της μείωσης της υπερπροσαρμογής. Το μόνο μειονέκτημα με τους τυχαίους ταξινομητές δασών είναι ότι κάθε δέντρο μπορεί να έχει πρόσβαση σε συγκεκριμένα από τα χαρακτηριστικά (features). Οπότε, αν τα χαρακτηριστικά είναι λίγα, κάποια δέντρα δεν θα έχουν πρόσβαση σε πιθανώς χρήσιμη πληροφορία. Επίσης τα δέντρα δεν μαθαίνουν σειριακά, αφού το καθένα είναι μόνο του, άρα ο αλγόριθμος παρουσιάζει κάποια πιο περιορισμένη δυνατότητα εκπαίδευσης.

2.4.3. Συγκριτική Ανάλυση Αλγορίθμων Επιβλεπόμενης Μάθησης

Οι αλγόριθμοι επιβλεπόμενης μάθησης όπως παρουσιάστηκαν και παραπάνω είναι πολλοί και εξετάστηκαν στην παρούσα εργασία από την αποδοτικότητα τους ή όχι στις 2 βασικές λειτουργίες, την ταξινόμηση και την παλινδρόμηση. Είναι σημαντικό όμως να διαφοροποιήσουμε τους αλγορίθμους αυτούς και να ξεχωρίσουμε την ιδιαίτερη τους χρήση σε κάθε διαφορετική περίπτωση.

Αρχικά, η πιο σημαντική διαφορά μεταξύ των προβλημάτων ταξινόμησης και παλινδρόμησης είναι ότι η ταξινόμηση αφορά την πρόβλεψη μιας διακριτής ετικέτας κλάσης ενώ η παλινδρόμηση αφορά την πρόβλεψη μιας συνεχούς ποσότητας. Οπότε, ο αλγόριθμος παλινδρόμησης αντιστοιχίζει την τιμή εισόδου (x) με τη μεταβλητή συνεχούς εξόδου (y), ενώ ο αλγόριθμος ταξινόμησης με τη διακριτή μεταβλητή εξόδου (y). Άρα τα προβλήματα της ταξινόμησης είναι διαφορετικά από τα προβλήματα της παλινδρόμησης χωρίς αυτό να αποκλείει την ενδεχόμενη ύπαρξη επικάλυψης. Πιο συγκεκριμένα, ένας αλγόριθμος ταξινόμησης μπορεί να προβλέπει μια συνεχή τιμή, αλλά η συνεχής τιμή να έχει τη μορφή πιθανότητας για μια ετικέτα κλάσης. Αντίστοιχα, ένας αλγόριθμος παλινδρόμησης μπορεί να προβλέπει μια διακριτή τιμή, αλλά η διακριτή τιμή να έχει τη μορφή ακέραιας ποσότητας.



Σχήμα 2.4. Ένα διάγραμμα ταξινόμησης στα αριστερά και ένα παλινδρόμησης στα δεξιά.

Ακόμη, οι αλγόριθμοι εμφανίζουν διαφορές και στο λειτουργικό τους κομμάτι. Στην παλινδρόμηση, στοχεύεται η εύρεση της καλύτερης γραμμής προσαρμογής, η οποία μπορεί να προβλέψει το αποτέλεσμα με μεγαλύτερη ακρίβεια. Στην ταξινόμηση, στοχεύεται η εύρεση του ορίου απόφασης, το οποίο μπορεί να χωρίσει το σύνολο δεδομένων σε διαφορετικές κλάσεις.

Όπως παρατηρήθηκε παραπάνω, μερικοί αλγόριθμοι μπορούν να χρησιμοποιηθούν τόσο για την ταξινόμηση όσο και για παλινδρόμηση με μικρές τροποποιήσεις, όπως τα δέντρα αποφάσεων, ενώ ορισμένοι δεν μπορούν καθόλου ή δεν μπορούν εύκολα να χρησιμοποιηθούν και για τους δύο τύπους προβλημάτων, όπως η γραμμική παλινδρόμηση για παλινδρόμηση και λογιστική παλινδρόμηση για ταξινόμηση.

Μια άλλη σημαντική διαφορά είναι ο τρόπος με τον οποίο αξιολογούμε τις προβλέψεις ταξινόμησης και παλινδρόμησης που ποικίλλει και δεν επικαλύπτεται. Ειδικότερα, οι προβλέψεις ταξινόμησης μπορούν να αξιολογηθούν χρησιμοποιώντας τη μετρική accuracy, η οποία μας δείχνει την ακρίβεια μιας πρόβλεψης και θα αναλυθεί σε επόμενο κεφάλαιο. Αντιθέτως, οι προβλέψεις παλινδρόμησης μπορούν να υπολογιστούν χρησιμοποιώντας το μέσο τετραγωνικό σφάλμα ή τη ρίζα αυτού, μετρικές που θα αναλυθούν ομοίως σε επόμενα κεφάλαια αλλά δεν χρησιμοποιούνται για τις προβλέψεις ταξινόμησης.

Συμπερασματικά, στην εργασία αυτή χρησιμοποιήθηκαν τόσο η παλινδρόμηση όσο και η ταξινόμηση στα πλαίσια της επιβλεπόμενης μάθησης νευρωνικών δικτύων και για την παραγωγή επιθυμητών αποτελεσμάτων. Οι δύο αυτοί τύποι αλγορίθμων έδωσαν αποτελέσματα για διαφορετικά προβλήματα και ερωτήματα που τέθηκαν τα οποία θα αναλυθούν περεταίρω στη συνέχεια.

Κεφάλαιο 3. Μεθοδολογική Προσέγγιση

3.1 Εισαγωγή

Το συγκεκριμένο πρόβλημα αποτελεί ουσιαστικά ένα πρόβλημα τιμολόγησης τυχαίων δεδομένων που αποτελούν προϊόντα από μια τυχαία βάση, το οποίο όμως πρέπει να γίνει με έναν αποδοτικό τρόπο και λαμβάνοντας υπόψη διαφορετικούς παράγοντες. Στόχος της διπλωματικής, όπως αναφέρθηκε, είναι ουσιαστικά να αναπτυχθεί μια μεθοδολογία σύμφωνα με την οποία ένας πάροχος ηλεκτρονικού εμπορίου θα μπορεί όχι απλά να ορίσει τις τιμές των προϊόντων του για μια συγκεκριμένη χρονική στιγμή αλλά και να καθορίσει τους επιμέρους παράγοντες που ορίζουν τα προϊόντα και ουσιαστικά καθιστούν μια ολοκληρωμένη στρατηγική τιμολόγησης και πώλησης τους ηλεκτρονικά. Ας σημειωθεί ότι αυτή η μεθοδολογία ουσιαστικά είναι ένας δυναμικός τρόπος στον οποίο οδηγηθήκαμε αξιολογώντας τον πιο στατικό τρόπο πρόβλεψης τιμών και βλέποντας πως δεν εξάγει ικανοποιητικά αποτελέσματα. Για το λόγο αυτό οδηγηθήκαμε σε δοκιμές εναλλακτικών μεθόδων και στη διατύπωση μιας πιο δυναμικής μεθοδολογίας που θα παρουσιαστεί παρακάτω.

Ειδικότερα, η εκκίνηση της μεθοδολογίας αποτελείται από ένα βασικό κομμάτι προεπεξεργασίας και ανάλυσης δεδομένων.

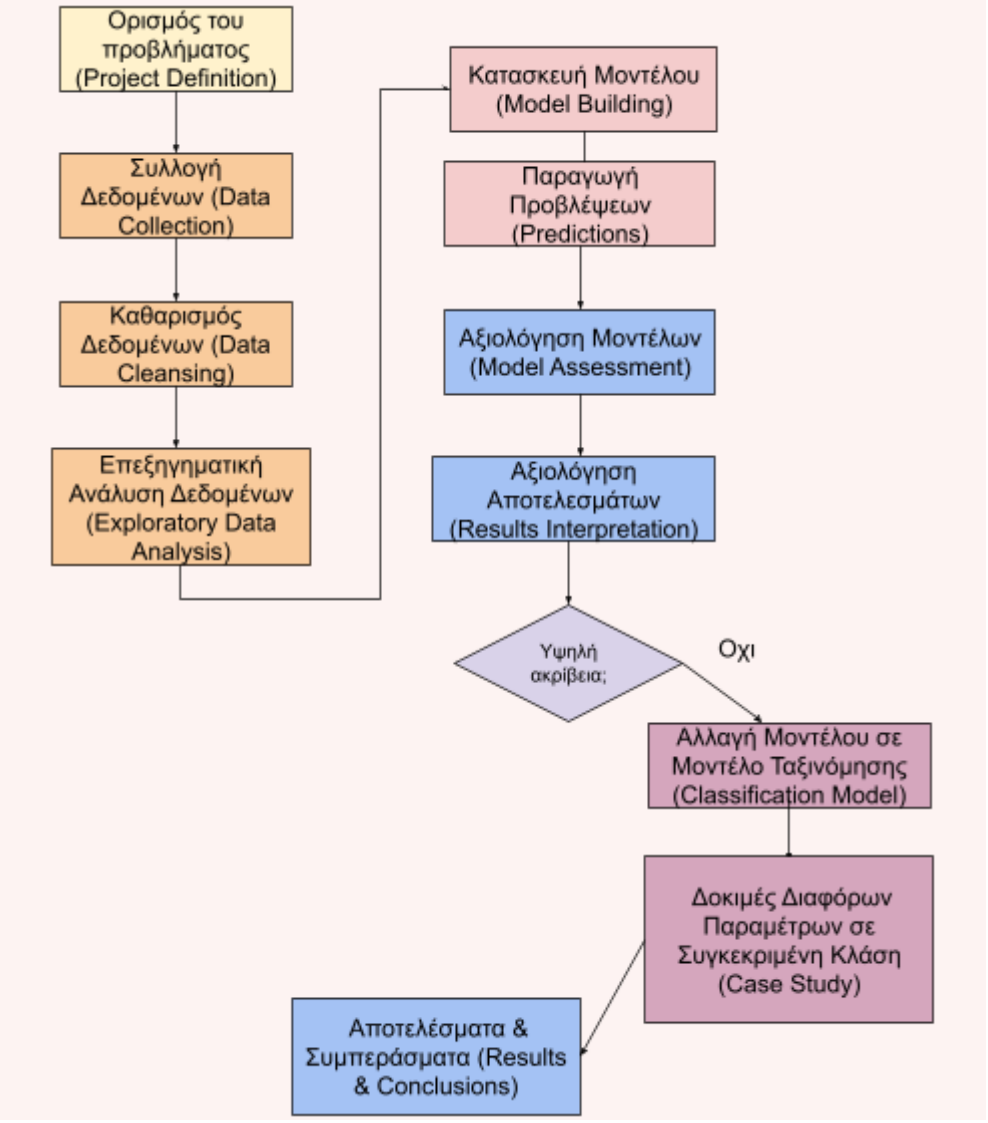
Η προεπεξεργασία δεδομένων (data preprocessing) είναι η διαδικασία μετατροπής μη επεξεργασμένων δεδομένων σε κατανοητή μορφή. Είναι ένα σημαντικό βήμα πριν την ανάλυση δεδομένων, καθώς δεν μπορούμε να εργαστούμε με ανεπεξέργαστα δεδομένα. Η ποιότητα των δεδομένων πρέπει να ελέγχεται πριν από την εφαρμογή αλγορίθμων μηχανικής μάθησης και συγκεκριμένα εξαγωγής προβλέψεων στην περίπτωση της εργασίας αυτής. Ο έλεγχος της ποιότητας των δεδομένων που γίνεται κατά την προεπεξεργασία βασίζεται σε 5 κυρίως άξονες. Αρχικά ελέγχεται η ακρίβεια τους, δηλαδή αν τα δεδομένα εισόδου είναι τα σωστά και επιθυμητά. Επειτα εξετάζονται ως προς την πληρότητα, δηλαδή αν είναι διαθέσιμα και έχουν καταγραφεί ολοκληρωμένα και επαρκώς και ως προς τη συνέπεια, δηλαδή αν τα ίδια δεδομένα διατηρούνται σε όλα τα μέρη που εμφανίζονται. Τέλος, εξασφαλίζεται η επικαιρότητα, η αξιοπιστία και η ερμηνευτικότητα τους ώστε να είναι δεδομένα εύκολα κατανοητά από τους χρήστες και σε μορφή τέτοια που θα εξάγουν αξιόπιστα αποτελέσματα εξόδου.

Το επόμενο βήμα, το οποίο με τη σειρά του αποτελείται από πολλά επιμέρους, είναι η ανάλυση των δεδομένων (data analysis). Η τελευταία είναι η διαδικασία της συστηματικής εφαρμογής στατιστικών ή/και λογικών τεχνικών για την περιγραφή και την επεξήγηση, τη σύμπτυξη, την ανακεφαλαίωση, και την αξιολόγηση των δεδομένων. Στόχος της ανάλυσης δεδομένων είναι ο καθαρισμός, η μετατροπή και τελικά η μοντελοποίηση των δεδομένων για την ανακάλυψη χρήσιμων πληροφοριών, για τη λήψη αποφάσεων και στην περίπτωση

μας για την εξαγωγή συμπερασμάτων για την τιμολόγηση και την γενικότερη στρατηγική πώλησης των προϊόντων.

Η διαδικασία αυτή λοιπόν ολοκληρώνεται με την εξαγωγή χρήσιμων πληροφοριών από τα δεδομένα και τη λήψη της απόφασης που αποτελεί το σκοπό μας, με βάση την ανάλυση αυτή.

Μια συνοπτική παρουσίαση των βημάτων βρίσκεται παρακάτω:



Σχήμα 3.1. Σχηματική Αναπαράσταση της Μεθοδολογίας.

3.2 Επιμέρους Βήματα Μεθοδολογίας

3.2.1. Ορισμός προβλήματος (Project Definition)

Το πρώτο βήμα που εμφανίζεται σε κάθε παρόμοια τέτοια μελέτη είναι ο ορισμός του προβλήματος. Είναι ένα βήμα το οποίο είναι σημαντικό να μην παραλείπεται καθώς, πριν ξεκινήσει η συλλογή των δεδομένων είναι απαραίτητο να καθοριστεί ποιο είναι το επιθυμητό αποτέλεσμα στο οποίο πρέπει να οδηγήσουν τα δεδομένα αυτά. Το επιχειρηματικό πρόβλημα αυτό πρέπει να είναι ξεκάθαρο, ακριβές και καλά ορισμένο αλλά και να απαντά σε ένα σύνολο μετρήσιμων, σαφών, και περιεκτικών ερωτήσεων. Οι απαντήσεις στις ερωτήσεις αυτές θα ορισίσουν μια κατάλληλη υπόθεση που θα είναι η αρχή της διαδικασίας.

Το ξεκίνημα με έναν σαφή στόχο είναι ένα ουσιαστικό βήμα στη διαδικασία ανάλυσης δεδομένων. Για αυτό στη συγκεκριμένη εργασία σαν επιθυμητό αποτέλεσμα και ορισμός του επιχειρηματικού προβλήματος τέθηκε η τιμολόγηση προϊόντων ενός ηλεκτρονικού παρόχου με αποδοτικό και δυναμικό τρόπο, τέτοιο ώστε να τα κάνει προτιμητέα και επιθυμητά έναντι άλλων εναλλακτικών προϊόντων σε μια συγκεκριμένη χρονική στιγμή. Συμπληρωματικά, επιθυμητό αποτέλεσμα είναι η εξαγωγή γενικών συμπερασμάτων σχετικά με τους παράγοντες εκείνους που καθορίζουν τη δημοφιλία των προϊόντων και ορίζουν πιο αποδοτικά αυτές τις στρατηγικές τιμολόγησης.

3.2.2. Συλλογή Δεδομένων (Data Collection)

Αφού οριστεί το πρόβλημα, είναι απαραίτητη η συλλογή των δεδομένων που θα δώσουν τις γνώσεις που απαιτούνται για την επιθυμητή λύση. Αυτό το μέρος της διαδικασίας περιλαμβάνει τη μελέτη των δεδομένων και την εύρεση τρόπων για τη λήψη αυτών των δεδομένων, είτε πρόκειται για ερωτήματα σε εσωτερικές βάσεις δεδομένων, είτε για την εύρεση εξωτερικών συνόλων δεδομένων. Η ανάμιξη και η συγχώνευση δεδομένων από όσο το δυνατόν περισσότερες προελεύσεις είναι κάτι που καθιστά τη μελέτη ακόμη πιο ολοκληρωμένη. Εφόσον λοιπόν έχουμε τις δομημένες ερωτήσεις και απαντήσεις από τον ορισμό του προβλήματος, η εύρεση δεδομένων γίνεται με βάση αυτές. Ο τελικός στόχος αυτού του βήματος είναι μια ολοκληρωμένη βάση δεδομένων που αντιστοιχεί στα απαραίτητα δεδομένα του προβλήματος προς λύση.

Για τη συλλογή δεδομένων στη παρούσα εργασία έγινε έρευνα στις διάφορες ιστοσελίδες με δωρεάν συλλογές δεδομένων. Η αναλυτική έρευνα η οποία έγινε βασιζόταν στα κριτήρια που τέθηκαν από τον ορισμό του προβλήματος δηλαδή την εύρεση μιας βάσης με δεδομένα προϊόντων ηλεκτρονικού εμπορίου από διάφορους παρόχους και τα στοιχεία τιμολόγησης και διάθεσής τους για μια χρονική περίοδο. Τελικά, επιλέχθηκε ένα αρχείο δεδομένων από την ιστοσελίδα data.word το οποίο περιλαμβάνει δεδομένα σχετικά με τα ηλεκτρονικά

προϊόντα και τις πληροφορίες τιμολόγησής τους από τη βάση δεδομένων προϊόντων Datafiniti για ηλεκτρονικούς παρόχους. Τα δεδομένα ενημερώθηκαν μεταξύ Ιανουαρίου και Ιουλίου 2018 από πηγές που αποτελούν κάποιους από τους μεγαλύτερους παρόχους ηλεκτρονικού εμπορίου που κυριαρχούν αυτή την στιγμή στην αγορά. Δεν έγινε ανάμιξη με άλλα δεδομένα καθώς η βάση ήταν πλήρης και κρίθηκε κατάλληλη για τη λύση του επιθυμητού προβλήματος.

3.3.3. Καθαρισμός Δεδομένων (Data Cleansing)

Η συλλογή των κατάλληλων δεδομένων δεν εξασφαλίζει ότι τα δεδομένα είναι έτοιμα για χρήση. Αυτό συμβαίνει επειδή τα ανεπεξέργαστα δεδομένα σπανίως μπορούν να χρησιμοποιηθούν στην τρέχουσα μορφή τους. Συχνά υπάρχουν ελαττώματα μέσα σε αυτά, όπως ανακρίβειες, λανθασμένες τιμές, τιμές που λείπουν και τα λοιπά. Ακόμη, ορισμένες φορές τα δεδομένα μπορεί να είναι αρκετά «μπερδεμένα», ιδιαίτερα αν δεν έχουν συλλεχθεί σωστά και με μεθοδικό και προσεκτικό τρόπο. Εμφανίζονται τιμές ως κενές (null), διπλότυπες τιμές, και τιμές που λείπουν. Είναι σημαντικό όχι μόνο να γίνει έλεγχος για όλα αυτά αλλά και να γίνουν οι κατάλληλες αλλαγές, προσθήκες και τροποποιήσεις. Αν και φαινομενικά είναι ήσσονος σημασίας, αυτά τα σφάλματα μπορεί να είναι αρκετά επιζήμια κατά την ανάλυση και τη μοντελοποίηση καθώς μπορεί να παραμορφώσουν τα αποτελέσματα.

Η διαδικασία, που ονομάζεται καθαρισμός δεδομένων είναι λοιπόν αυτό που κάνουμε με στόχο την τροποποίηση ή την αφαίρεση λανθασμένων ή περιττών δεδομένων, καθώς και τον έλεγχο για ατέλειες ή ασυνέπειες. Είναι ένα σημαντικό βήμα που θα καθορίσει την ποιότητα των δεδομένων. Στη συγκεκριμένη βάση χρειάζεται να εξεταστούν συγκεντρωτικά στοιχεία των γραμμών και των στηλών, να αφαιρεθούν μη έγκυρα δεδομένα, να αντικατασταθούν λανθασμένα δεδομένα ενώ και να συμπληρωθούν ελλιπή δεδομένα. Για το σκοπό αυτό, χρησιμοποιήθηκαν τόσο δομημένες μέθοδοι που θα αναλυθούν παρακάτω όσο και στοιχεία από έρευνα των προϊόντων στο διαδίκτυο αλλά και μια διαισθητική τροποποίηση.

3.3.4. Επεξηγηματική Ανάλυση Δεδομένων (Explanatory Data Analysis)

Η επεξηγηματική ανάλυση δεδομένων είναι το επόμενο βήμα της διαδικασίας το οποίο χρησιμοποιείται για την ανάλυση και διερεύνηση των συνόλων δεδομένων και τη σύνοψη των κύριων χαρακτηριστικών τους, συχνά χρησιμοποιώντας μεθόδους οπτικοποίησης δεδομένων όπως γραφικές παραστάσεις. Βοηθά στον προσδιορισμό του καλύτερου τρόπου χειρισμού των δεδομένων για τη λήψη των απαντήσεων αφού βασίζεται σε μια μεθοδολογία ανακάλυψης μοτίβων, παρατήρησης ανωμαλιών και ελέγχου υποθέσεων και παραδοχών.

Η διαδικασία αυτή συμβάλλει επίσης στο να γίνει σαφές ποια δεδομένα μπορούν να προσφέρουν πληροφορίες χρήσιμες πέρα από την τυπική εργασία μοντελοποίησης. Ακόμη, εξασφαλίζεται μια καλύτερη κατανόηση των μεταβλητών του συνόλου δεδομένων και των σχέσεων μεταξύ τους κάτι το οποίο με τη σειρά του συμβάλει στο να προσδιοριστεί το αν είναι κατάλληλες οι στατιστικές τεχνικές που χρησιμοποιούνται για την ανάλυση δεδομένων.

Σε αυτό το βήμα, λοιπόν, αρχίζουμε να αναλύουμε εις βάθος τα δεδομένα, να τα εμπλουτίζουμε με νέες μεταβλητές και να αντλούμε σημαντικές πληροφορίες από αυτά, να εξάγουμε μεταδεδομένα και να ανακαλύπτουμε τις συσχετίσεις που τα διέπουν. Φυσικά, στο βήμα αυτό είναι σημαντικό να δοθεί η απαραίτητη προσοχή για να αποφευχθεί η ακούσια προκατάληψη που μπορεί να εισάγει ανεπιθύμητα μοτίβα στην διαδικασία. Ακόμη, επειδή η βάση περιλαμβάνει έναν σχετικά μεγάλο όγκο δεδομένων, η οπτικοποίηση είναι ο καλύτερος τρόπος για να τη διερεύνηση αλλά και την κοινοποίηση των ευρήματων. Η γραφική απεικόνιση, βοηθάει ώστε να καταλήγουμε σε συμπεράσματα, μοτίβα και σχέσεις μεταξύ των δεδομένων αλλά και να τις αποδείξουμε. Μερικοί τρόποι απεικόνισης που επιλέγουμε είναι το διάγραμμα “κουτί” (boxplot), το διαγράμματα διασποράς (scatter plot), το διάγραμμα μπαρών (barplot) και οι πίνακες συσχετίσεων (correlation matrix).

3.3.5. Κατασκευή Μοντέλου Πρόβλεψης (Model Building)

Το τελευταίο βήμα ώστε τα δεδομένα να είναι έτοιμα να εισαχθούν σε μοντέλα για να παράγουν τις επιθυμητές προβλέψεις περιλαμβάνει την κατασκευή του μοντέλου, μια διαδικασία η οποία αρχικά απαιτεί την κανονικοποίηση των δεδομένων (data normalization) και την κωδικοποίηση των δεδομένων.

Ο σκοπός της κανονικοποίησης είναι η μετατροπή των δεδομένων με τέτοιο τρόπο ώστε να είναι είτε αδιάστατα είτε να περιγράφονται με παρόμοιες κατανομές. Η κανονικοποίηση είναι ένα βασικό βήμα στην προεπεξεργασία δεδομένων σε οποιαδήποτε εφαρμογή μηχανικής μάθησης και προσαρμογής μοντέλων, καθώς βελτιώνει εντυπωσιακά την ακρίβεια του μοντέλου. Ο τρόπος με τον οποίο λειτουργεί είναι να δίνει ίσα βάρη σε κάθε μεταβλητή έτσι ώστε καμία μεταβλητή να μην κατευθύνει την απόδοση του μοντέλου προς μία κατεύθυνση μόνο και μόνο επειδή είναι ένας μεγαλύτερος αριθμός. Οι τρόποι κανονικοποίησης που δοκιμάστηκαν στην παρούσα βάση δεδομένων αλλά και η τελική επιλογή θα παρουσιαστεί αναλυτικά στο επόμενο κεφάλαιο.

Μετά την κανονικοποίηση, απαιτείται η κατάλληλη κωδικοποίηση των δεδομένων, δηλαδή η μετατροπή τους σε μορφή τέτοια που θα είναι κατανοητά για αλγόριθμους και για την αποτελεσματική παραγωγή προβλέψεων. Συνήθως η κωδικοποίηση αποτελεί τρόπο μετατροπής των μεταβλητών σε δυαδική μορφή με διάφορες τεχνικές. Μια αποτελεσματική κωδικοποίηση κάνει τα δεδομένα πιο χρήσιμα και εκφραστικά, και μπορεί να αναπροσαρμόζεται εύκολα. Ο τρόπος που αυτό έγινε στην παρούσα βάση θα αναλυθεί παρακάτω.

Η αποδοτική κωδικοποίηση και κανονικοποίηση είναι χρήσιμες καθώς οι αλγόριθμοι μηχανικής μάθησης αντιμετωπίζουν τη σειρά των αριθμών ως ένα χαρακτηριστικό σημαντικότητας. Με άλλα λόγια, εκλαμβάνουν έναν μεγαλύτερο αριθμό ως καλύτερο ή πιο σημαντικό από έναν μικρότερο αριθμό. Αν και αυτό είναι χρήσιμο για ορισμένες συνηθισμένες καταστάσεις, κάποια δεδομένα εισόδου δεν έχουν κάποια κατάταξη για τιμές κατηγοριών, και αυτό μπορεί να οδηγήσει σε προβλήματα με προβλέψεις και κακή απόδοση. Ετσι με την κανονικοποίηση και την κωδικοποίηση μετατρέπουμε τα δεδομένα στην κατάλληλη μορφή για την εισαγωγή τους σε μοντέλα πρόβλεψης.

Τέλος, τα κανονικοποιημένα και κωδικοποιημένα δεδομένα πρέπει τώρα να διαμορφώσουν την τελική βάση δεδομένων. Εκεί εντάσσεται ένα τελευταίο βήμα που αποτελεί την επιλογή χαρακτηριστικών (feature selection) αν αυτό είναι απαραίτητο. Όπως υποδηλώνει το όνομα, η επιλογή χαρακτηριστικών είναι κυριολεκτικά η διαδικασία επιλογής ενός υποσυνόλου χαρακτηριστικών από έναν αρχικά μεγάλο όγκο χαρακτηριστικών. Εφόσον μια από τις πιο σημαντικές πτυχές της μηχανικής μάθησης είναι το μοντέλο να μπορεί να αποκτήσει και να παράγει αξιοποιήσιμες γνώσεις, είναι σημαντικό να επιλεγεί ένα υποσύνολο των σημαντικών χαρακτηριστικών από το μεγαλύτερο σύνολο που έχουν έντονη συσχέτιση με την επιθυμητή μεταβλητής πρόβλεψης.

Τέλος, αυτά τα επιλεγμένα δεδομένα είναι πλέον έτοιμα να εισαχθούν στα μοντέλα προβλέψεων. Κατά την ανάπτυξη μοντέλων μηχανικής μάθησης, είναι επιθυμητό το εκπαιδευμένο μοντέλο να αποδίδει καλά σε νέα, άγνωστα δεδομένα. Για την προσομοίωση των νέων αυτών δεδομένων, τα διαθέσιμα δεδομένα υποβάλλονται σε διαχωρισμό δεδομένων, όπου χωρίζονται σε 2 τμήματα (train - test split). Ειδικότερα, το πρώτο τμήμα είναι το μεγαλύτερο υποσύνολο δεδομένων που χρησιμοποιείται ως σύνολο εκπαίδευσης (train dataset) και αντιστοιχεί περίπου στο 70-90% των αρχικών δεδομένων. Το δεύτερο είναι συνήθως ένα μικρότερο υποσύνολο και χρησιμοποιείται ως σύνολο δοκιμών (test dataset) και αντιστοιχεί στο υπόλοιπο 30-10% των δεδομένων.

3.3.6 Παραγωγή Προβλέψεων (Predictions)

Το επόμενο βήμα είναι αφού έχουμε ήδη τα 2 διαφορετικά σύνολα δεδομένων (εκπαίδευσης και δοκιμών) και επιθυμούμε την παραγωγή προβλέψεων. Ειδικότερα, το σύνολο εκπαίδευσης χρησιμοποιείται για τη δημιουργία ενός προγνωστικού μοντέλου και αυτό το εκπαιδευμένο μοντέλο στη συνέχεια εφαρμόζεται στο σύνολο δοκιμών για να κάνει προβλέψεις. Πιο αναλυτικά, θέτουμε σαν μεταβλητή στόχο αυτή την οποία θέλουμε να προβλέψουμε η οποία στη συνέχεια διαφοροποιείται σε μια ξεχωριστή βάση δεδομένων. Εκπαιδεύουμε το εκάστοτε μοντέλο έτσι ώστε να παράγει επιθυμητά αποτελέσματα της μεταβλητής στόχου με είσοδο τα υπόλοιπα δεδομένα με όσο μεγαλύτερη ακρίβεια γίνεται. Δοκιμάζουμε διαφορετικά μοντέλα, όπως αυτά που αναφέρθηκαν παραπάνω για να μπορούμε έπειτα να τα αξιολογήσουμε και να επιλέξουμε το κατάλληλο. Η επιλογή του βέλτιστου μοντέλου γίνεται με βάση την απόδοση του μοντέλου στο σύνολο δοκιμών. Η

υλοποίηση αυτή ουσιαστικά αποτελεί αυτό το επιθυμητό βήμα προς την απόκτηση γνώσεων και την πρόβλεψη μελλοντικών τάσεων.

Παράλληλα, για να γίνει η πιο οικονομική χρήση των διαθέσιμων δεδομένων, χρησιμοποιείται συνήθως μια διαδικασία που ονομάζεται διασταυρούμενη επικύρωση N-fold (Cross Validation - CV) όπου το σύνολο δεδομένων χωρίζεται σε N πτυχές (folds), συνήθως πέντε ή δέκα (5-fold ή 10-fold CV). Σε ένα τέτοιο N-fold CV, μία από τις πτυχές αφήνεται έξω ως τα στοιχεία δοκιμής, ενώ οι υπόλοιπες πτυχώσεις χρησιμοποιούνται ως τα στοιχεία κατάρτισης για την κατασκευή μοντέλου.

Μπορεί επίσης να πραγματοποιηθεί βελτιστοποίηση υπερπαραμέτρων (hyperparameter optimization). Οι υπερπαράμετροι είναι ουσιαστικά παράμετροι του αλγορίθμου μηχανικής μάθησης που επηρεάζουν άμεσα τη διαδικασία μάθησης και την απόδοση πρόβλεψης. Δεδομένου ότι δεν υπάρχουν ρυθμίσεις υπερ-παραμέτρων "one-size fits all" που θα λειτουργούν καθολικά για όλα τα σύνολα δεδομένων, θα χρειαστεί να πραγματοποιηθεί βελτιστοποίηση των υπερ-παραμέτρων γνωστή και ως ρύθμιση υπερ-παραμέτρων ή ρύθμιση μοντέλου (hyperparameter tuning or model tuning). Η διαδικασία αυτή σκοπεύει στην εύρεση των υπερπαραμέτρων ενός συγκεκριμένου αλγορίθμου μηχανικής μάθησης που παρέχουν την καλύτερη απόδοση. Οι υπερπαράμετροι αυτοί, σε αντίθεση με τις παραμέτρους μοντέλου, καθορίζονται από το χρήστη του μοντέλου πριν από την εκπαίδευση. Αποτελούν ουσιαστικά τις ρυθμίσεις του μοντέλου ώστε να μπορεί να λύσει βέλτιστα το πρόβλημα της μηχανικής μάθησης.

Στην παρούσα εργασία, θέτουμε σαν μεταβλητή στόχο την τιμή των προϊόντων και παράγουμε τις προβλέψεις σύμφωνα με διαφορετικά μοντέλα. Ουσιαστικά αποτελεί μια διαδικασία παλινδρόμησης (regression), καθώς προβλέπουμε μια συνεχή μεταβλητή, την τιμή ενός προϊόντος σύμφωνα με κάποια χαρακτηριστικά που αυτό έχει και αποτελούν την είσοδο στο μοντέλο. Στη συνέχεια, πραγματοποιούμε cross validation αλλά και hyperparameter optimization ώστε να στοχεύουμε σε υψηλή απόδοση.

3.3.7. Αξιολόγηση Μοντέλων (Model Assessment)

Αφού πραγματοποιήσουμε την εκπαίδευση και τις προβλέψεις, είναι σημαντικό να αξιολογήσουμε την απόδοση των μοντέλων που χρησιμοποιούμε. Με αυτόν τον τρόπο θα εκτιμηθεί ο βαθμός στον οποίο ένα εκπαιδευμένο μοντέλο μπορεί να προβλέψει με ακρίβεια τις τιμές των δεδομένων εισόδου. Για το λόγο αυτό, χρησιμοποιούμε τις μετρικές αξιολόγησης μοντέλων. Οι τελευταίες απαιτούνται για την ποσοτικοποίηση της απόδοσης των μοντέλων. Η επιλογή των μετρήσεων αξιολόγησης εξαρτάται από τη συγκεκριμένη εργασία μηχανικής μάθησης που πραγματοποιούμε κάθε φορά (όπως ταξινόμηση, παλινδρόμηση).

Η αξιολόγηση είναι απαραίτητη καθώς ολόκληρη η ιδέα της κατασκευής μοντέλων μηχανικής μάθησης λειτουργεί με βάση μια εποικοδομητική αρχή ανατροφοδότησης. Ένα

μοντέλο κατασκευάζεται, λαμβάνεται ανατροφοδότηση από μετρήσεις, γίνονται βελτιώσεις, και η διαδικασία συνεχίζεται μέχρι να επιτευχθεί μια επιθυμητή ακρίβεια. Οι μετρικές αξιολόγησης εξηγούν την απόδοση ενός μοντέλου με σαφή και ποσοτικό τρόπο, ώστε να μπορούμε να διακρίνουμε τα αποτελέσματα των μοντέλων μεταξύ τους. Σύμφωνα με αυτές γίνεται και η επιλογή του μοντέλου με την υψηλότερη ακρίβεια.

Στην συγκεκριμένη περίπτωση και επειδή αναφερόμαστε σε προβλήματα παλινδρόμησης θα χρησιμοποιηθούν μετρικές αυτών των προβλημάτων, δηλαδή, η Ρίζα Μέσου Τετραγωνικού Σφάλματος (Root Mean Squared Error - RMSE), το Μέσο Απόλυτο Σφάλμα (Mean Absolute Error - MAE), το Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error - MSE) και και το R2 Score. Περαιτέρω ανάλυση των μετρικών θα γίνει σε παρακάτω κεφάλαιο.

3.3.8. Αξιολόγηση Αποτελεσμάτων (Results Interpretation)

Αφού έχουμε πλέον μετρικές που αξιολογούν τα μοντέλα μας είναι σημαντικό να αξιολογήσουμε συνολικά το αποτέλεσμα της όλης διαδικασίας που χρησιμοποιήθηκε και να εξασφαλίσουμε ότι οδηγεί στην επίλυση του αρχικού προβλήματος.

Αρχικά, σχετικά με τα μοντέλα είναι σημαντικό να γνωρίζουμε ότι το RMSE θα είναι πάντα μεγαλύτερο ή ίσο με το MAE και όσο μεγαλύτερη είναι η διαφορά μεταξύ τους, τόσο μεγαλύτερη είναι η διακύμανση στα μεμονωμένα σφάλματα του δείγματος. Όσο ο δείκτης MAE τόσο και ο δείκτης RMSE (και MSE) μπορούν να κυμανθούν από 0 έως ∞. Αυτοί οι βαθμοί είναι αρνητικού προσανατολισμού. Έτσι οι χαμηλότερες τιμές είναι καλύτερες άρα τα μοντέλα με χαμηλότερα σφάλματα είναι πιο αποδοτικά και ακριβή και τείνουν να επιλέγονται. Ακόμη σχετικά με το r2 score γνωρίζουμε ότι όσο πλησιάζει στην τιμή 1, τόσο πιο τέλεια εκπαιδεύεται το μοντέλο. Άρα επιλέγονται τα μοντέλα με το υψηλότερο r2 score.

Γενικότερα, μια καλή μετρική σχετίζεται με το συγκεκριμένο σύνολο δεδομένων. Είναι καλή ιδέα να δημιουργηθεί εξ αρχής μια τιμή βάσης των μετρικών για το σύνολο δεδομένων και να αξιολογήσουμε συνολικά όλα τα μοντέλα σύμφωνα με αυτήν. Στην παρούσα εργασία αποφασίστηκε ότι αν οι τιμές σφαλμάτων δεν είναι ικανοποιητικές σύμφωνα με αυτό που τέθηκε σαν βάση τότε τα συνολικά αποτελέσματα της διαδικασίας θα αξιολογηθούν ως μη ικανοποιητικά άρα δεν θα θεωρήσουμε μια αποδοτική επίλυση του αρχικού προβλήματος. Πιο συγκεκριμένα, θα έχουμε προβλέψεις τιμών που είναι ανακριβείς άρα δεν εξασφαλίζουν ότι έχουμε μια σωστή στρατηγική τιμολόγησης αλλά και ότι τα προϊόντα θα προτιμηθούν έναντι άλλων. Απαιτείται, λοιπόν σε αυτήν την περίπτωση, η αλλαγή της προσέγγισης και η δοκιμή ενός άλλου εναλλακτικού τρόπου επίλυσης του προβλήματος.

3.3.9. Μοντέλο Ταξινόμησης (Classification Model)

Ο εναλλακτικός αυτός τρόπος επίλυσης του προβλήματος μετατρέπεται σε ένα πρόβλημα ταξινόμησης. Ειδικότερα, διατυπώνεται η ιδέα τα προϊόντα να ταξινομηθούν σε δημοφιλή και μη δημοφιλή βάσει των προβολών που έχουν από διάφορους χρήστες στο διαδίκτυο. Στόχος είναι να γίνει εύρεση αυτών των μη δημοφιλών προϊόντων ώστε να διατυπωθεί η κατάλληλη στρατηγική τιμολόγησης και πώλησης για αυτά και εν τέλει να γίνουν δημοφιλή και να προτιμηθούν από τους χρήστες.

Σε τεχνικό επίπεδο, η κατασκευή του μοντέλου και η εκπαίδευση του είναι μια διαδικασία αντίστοιχη με αυτή που αναφέρθηκε παραπάνω για την παλινδρόμηση. Παρόλα αυτά εδώ έχουμε τη δημιουργία μιας στήλης που περιλαμβάνει την κλάση με τα "μη δημοφιλή προϊόντα". Η κλάση αυτή παίρνει τιμές 0 και 1 αν ένα προϊόν είναι δημοφιλές ή όχι αντίστοιχα. Αυτή η κλάση αποτελεί επίσης και τη μεταβλητή στόχο (target) σε αυτή τη διαδικασία ταξινόμησης. Οι αλγόριθμοι, λοιπόν, που δοκιμάζουμε επιθυμούμε να παράγουν δυαδικά αποτελέσματα σχετικά με το αν ένα προϊόν ανήκει στην κλάση των μη δημοφιλών (άρα είναι μη δημοφιλές) ή όχι.

Για την αξιολόγηση των μοντέλων ταξινόμησης χρησιμοποιούμε τη μετρική της ακρίβειας. Η ακρίβεια (accuracy) είναι ο αριθμός των σωστών προβλέψεων ως αναλογία όλων των προβλέψεων που έγιναν. Όσο μεγαλύτερη είναι η μετρική αυτή τόσο καλύτερο είναι το μοντέλο ταξινόμησης και τόσο πιο ακριβή είναι τα αποτελέσματα μας. Ακόμη, για την αξιολόγηση της ταξινόμησης χρησιμοποιείται και το recall και precision. Όλες αυτές οι μετρικές θα παρουσιαστούν αναλυτικά στη συνέχεια.

3.3.10. Δοκιμές Σεναρίων (Case Study)

Όπως αναφέραμε και παραπάνω, χρησιμοποιούμε τη μετρική της ακρίβειας για να επιλέξουμε τον κατάλληλο και άρα πιο ακριβή ταξινομητή. Αυτό συμβαίνει ώστε στο επόμενο βήμα να χρησιμοποιηθεί στην πραγματικότητα για να παράξει τις ταξινομήσεις των προϊόντων στα πλαίσια κάποιων δοκιμών που έχουν ως στόχο τον ορισμό συγκεκριμένων παραμέτρων.

Ειδικότερα, θεωρούμε 3 ευρύτερες περιπτώσεις - case studies, οι οποίες ορίζονται θέτοντας σαν αντικείμενο μελέτης κάθε έναν από τους 3 κύριους παρόχους ηλεκτρονικού εμπορίου. Για κάθε έναν από αυτούς δοκιμάζουμε 6 υποπεριπτώσεις αλλαγών στα χαρακτηριστικά των προϊόντων, με στόχο την εύρεση εκείνων που θα επιφέρουν αλλαγή στην δημοφιλία των προϊόντων. Αποτελεί δηλαδή μια επαναληπτική διαδικασία ταξινόμησης έπειτα από μεταβολή παραγόντων των δεδομένων. Για κάθε διαφορετική υποπερίπτωση του κάθε case

study λαμβάνουμε τον νέο αριθμό των μη δημοφιλών προϊόντων, ώστε να εξάγουμε συμπεράσματα σχετικά με την συνολική μεταβολή από μη δημοφιλή σε δημοφιλή προϊόντα.

3.3.11. Αποτελέσματα και Συμπεράσματα

Το τελευταίο βήμα της διαδικασίας αυτής είναι η παραγωγή αποτελεσμάτων και συμπερασμάτων σχετικά με το αρχικό μας πρόβλημα. Αφού έχουμε παράξει τα αποτελέσματα κάθε υποπερίπτωσης είναι σημαντικό να παράγουμε τα αντίστοιχα συμπεράσματα σχετικά με τους παράγοντες που είναι σημαντικοί. Συγκεντρώνοντας όλα αυτά τα συμπεράσματα, μπορούμε να καταλήξουμε και σε ακόμη πιο γενικά πορίσματα σχετικά με τις στρατηγικές τιμολόγησης και πώλησης των προϊόντων ηλεκτρονικού εμπορίου, κάτι που ήταν και ο αρχικός μας στόχος. Στο σημείο αυτό η διαδικασία ολοκληρώνεται και διαθέτουμε μια συνολική λύση του προβλήματός μας.

3.3 Επιλογή Μεθόδων

Πέρα από την γενική μεθοδολογία που αναλύθηκε παραπάνω, είναι σημαντικό να αναλυθούν περαιτέρω και οι μέθοδοι και τα μεγέθη που αποτέλεσαν σημεία αναφοράς στα διάφορα αυτά βήματα που παρουσιάστηκαν. Οι μέθοδοι αυτές εμφανίστηκαν σε όλο το κομμάτι της προεπεξεργασίας, ανάλυσης, παραγωγής και αξιολόγησης των προβλέψεων καθώς και στην εξαγωγή συμπερασμάτων.

3.3.1 Διαχείριση Κενών ή Ελλιπών Δεδομένων (Dealing with Null or Missing Values)

Η μέθοδος της διαχείρισης των κενών ή ελλιπών δεδομένων εμφανίστηκε στο κομμάτι του καθαρισμού δεδομένων που αναφέρθηκε πρωτύτερα. Τα κενά ή ελλιπή δεδομένα (missing or null values) είναι δεδομένα τα οποία έχουν για κάποιο λόγο αφαιρεθεί η τροποποιηθεί από τη βάση. Ο λόγος για τον οποίο λείπουν αυτές οι τιμές μπορεί να είναι από ανθρώπινα λάθη, διακοπές στη ροή των δεδομένων, ανησυχίες για το προσωπικό απόρρητο και άλλοι παράγοντες. Το φαινόμενο αυτό είναι ένα από τα πιο συνηθισμένα προβλήματα των βάσεων δεδομένων και επηρεάζει άμεσα την απόδοση των μοντέλων μηχανικής μάθησης.

Ορισμένες πλατφόρμες μηχανικής μάθησης απορρίπτουν αυτόματα τις σειρές που περιλαμβάνουν τις κενές τιμές στη φάση εκπαίδευσης του μοντέλου (training phase) κάτι που όμως μειώνει την απόδοση μοντέλου λόγω του μειωμένου μεγέθους της βάσης εκπαίδευσης (training dataset). Από την άλλη, οι περισσότεροι αλγόριθμοι δεν δέχονται σύνολα δεδομένων τα οποία έχουν τιμές που λείπουν και για αυτό επιστρέφουν την ειδοποίηση για κάποιο σφάλμα. Αρα κρίνεται αναγκαίο να διαχειριστούμε από τα πρώιμα στάδια τις γραμμές και τις στήλες με τις τιμές αυτές.

Διαγραφή κενών ή ελλιπών δεδομένων

Η πιο απλή λύση για τις τιμές που λείπουν είναι η διαγραφή (drop) των γραμμών ή ολόκληρης της στήλης. Δεν υπάρχει βέλτιστο όριο για κατάργηση, αλλά ενδείκνυται η χρήση του 70% ως βάση σύμφωνα με την οποία θα απορροφηθούν οι γραμμές και οι στήλες που έχουν τιμές που λείπουν σε ποσοστό μεγαλύτερο από αυτό το όριο. Παρόλα αυτά αυτή η λύση δεν αποτελεί ιδανική εκδοχή καθώς χάνουμε σε μέγεθος βάσης δεδομένων, άρα χρειαζόμαστε λύσεις που γεμίζουν αυτές τις στήλες και γραμμές και όχι τις διαγράφουν και οι οποίες φαίνονται παρακάτω.

Αντικατάσταση με την πιο συχνά εμφανιζόμενη τιμή (mode)

Η μέθοδος αυτή αφορά αριθμητικές αλλά και κατηγορικές κατά κύριο λόγο μεταβλητές. Όταν τα δεδομένα που λείπουν, λοιπόν, είναι σε τυχαίες θέσεις και στήλες (Missing At Random - MAR) και οι τιμές που λείπουν είναι η πλειοψηφία των τιμών μιας στήλης, τότε μπορούμε να τις αντικαταστήσουμε με την πιο συχνή τιμή της αντίστοιχης μεταβλητής/στήλης.

Για να γίνει αυτό, αρχικά βρίσκουμε την τιμή που εμφανίζεται περισσότερο σε κάθε κατηγορία χρησιμοποιώντας το mode() ή αλλιώς τη μέθοδο value_counts. Αντικαθιστούμε όλες τις τιμές NAN σε αυτήν τη στήλη με αυτήν την τιμή. Βασικό πλεονέκτημα της επιλογής αυτής είναι η απλή και εύκολη υλοποίηση για κατηγορικές μεταβλητές/στήλες. Τα μειονεκτήματα που εμφανίζονται είναι ότι η αντικατάσταση στηλών με πολύ μεγάλο αριθμό κοινών τιμών μπορεί να προδιαθέτει την πρόβλεψη και παράλληλα το γεγονός ότι τροποιοείται αρκετά το ποστοστό της πιο συχνής τιμής στη βάση στην αντίστοιχη γραμμή και στήλη, κάτι το οποίο αλλοιώνει τα δεδομένα.

Αντικατάσταση με το μέσο όρο ή τη διάμεσο (mean/median)

Οι ελλιπείς και κενές τιμές που αφορούν αριθμητικές μεταβλητές, αντικαθίστανται με τις τιμές μέσου όρου/ διαμέσου (mean/median) της εκάστοτε μεταβλητής. Αυτή είναι η στατιστική μέθοδος χειρισμού των τιμών αυτών και αποδίδει και αυτή καλύτερα αποτελέσματα όταν συγκρίνεται με την κατάργηση των στηλών/γραμμών με κενές τιμές. Πιο συγκεκριμένα, ο μέσος όρος των αριθμητικών δεδομένων της στήλης (mean) χρησιμοποιείται για την αντικατάσταση τιμών null όταν τα δεδομένα κατανέμονται κανονικά (normally distributed). Η διάμεσος των αριθμητικών δεδομένων (median) χρησιμοποιείται όταν τα δεδομένα περιλαμβάνουν αρκετές ακραίες τιμές (outliers).

Για να υλοποιήσουμε τη μέθοδο αυτή, μπορούμε αρχικά να σχεδιάσουμε το ιστόγραμμα των δεδομένων μας. Ένα ιστόγραμμα (histogram) είναι μια γραφική αναπαράσταση που παρουσιάζει μια ομάδα σημείων δεδομένων σε εύρη που καθορίζονται από τον χρήστη. Έχει εμφάνιση παρόμοια με ένα ραβδόγραμμα, καθώς συμπυκνώνει μια σειρά δεδομένων σε μια οπτική μορφή που ερμηνεύεται εύκολα, λαμβάνοντας πολλά σημεία δεδομένων και

ομαδοποιώντας τα σε λογικές περιοχές ή κλάσεις. Έτσι μπορούμε να συγκρίνουμε το προκύπτον ιστόγραμμα με αυτά της κανονικής κατανομής και να επιλέξουμε τη χρήση του μέσου όρου αν ταυτίζεται ή τη χρήση της διαμέσου αν παρατηρούμε ακραίες διακυμάνσεις. Η μέθοδος αυτή έχει το πλεονέκτημα ότι είναι έγκυρη στατιστικά αλλά το μειονέκτημα ότι δεν λειτουργεί για κατηγορικές μεταβλητές ή για κατανομές με πολύ μεγάλη διακύμανση και ακραίες τιμές για τις οποίες ακόμη και η διάμεσος αποτελεί σχεδόν τυχαία επιλογή.

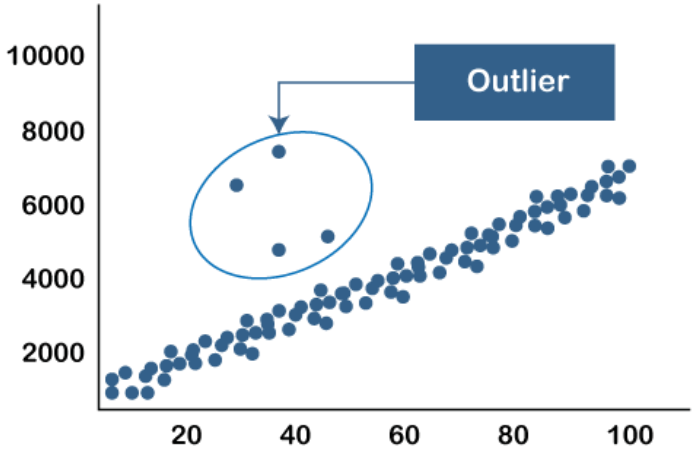
Αντικατάσταση με τυχαίες τιμές

Τέλος, μια εναλλακτική μέθοδος είναι ο ορισμός μιας νέας τιμής για τις τιμές NAN, δηλαδή τυχαία τιμή όπως "Other", "Unknown", "Not Defined" κτλ. Η μέθοδος αυτή μπορεί να χρησιμοποιηθεί τόσο για αριθμητικές όσο και για κατηγορικές μεταβλητές και χρησιμοποιείται κατά κύριο λόγο όταν οι κενές τιμές είναι πολλές αριθμητικά, περισσότερες από τις μη κενές. Αυτό συμβαίνει καθώς σε αυτήν την περίπτωση δεν έχουμε αρκετές πληροφορίες και η αντικατάσταση τους με κάποιον από τους παραπάνω τρόπους θα ήταν αρκετά αυθαίρετη.

Η υλοποίηση της μεθόδου είναι απλή και περιλαμβάνει απλά την αντικατάσταση των κενών τιμών με ένα συγκεκριμένο όνομα - τιμή. Το βασικό της πλεονέκτημα είναι ότι αποτελεί μια απλή και εύκολη στην υλοποίηση διαδικασία για κατηγορικές μεταβλητές/στήλες η οποία διατηρεί τη διακύμανση των μεταβλητών και δεν προκαλεί μεταβολή στις πρωταρχικές στατιστικές. Ως μειονέκτημα της μεθόδου θεωρείται η δημιουργία τυχαίων δεδομένων στην περίπτωση που η κατηγορία που λείπει είναι ένα μεγάλο ποσοστό της βάσης, κάτι που αλλοιώνει τα επιθυμητά αποτελέσματα.

3.3.2 Διαχείριση Ακραίων Τιμών (Outliers)

Μια ακόμη μέθοδος που εμφανίστηκε συγκεκριμένα στο κομμάτι της Επεξηγηματικής Ανάλυσης των Δεδομένων για την προετοιμασία τους για το μοντέλο είναι η διαχείριση των ακραίων τιμών. Οι ακραίες τιμές μιας βάσης δεδομένων (outliers) είναι ακραίες τιμές που διαφέρουν από άλλες παρατηρήσεις πάνω σε δεδομένα, οι οποίες μπορεί να υποδεικνύουν μεταβλητότητα σε μια μέτρηση, πειραματικά σφάλματα ή ανεπιθύμητες τροποποιήσεις στα δεδομένα. Με άλλα λόγια, μια ακραία τιμή (outlier) είναι μια παρατήρηση η οποία αποκλίνει από ένα συνολικό μοτίβο ενός δείγματος και δεν εμπίπτει στο γενικό πεδίο εφαρμογής των άλλων παρατηρήσεων. Μια γραφική αναπαράσταση των ακραίων τιμών παρουσιάζεται και στο σχήμα 3.2. που βρίσκεται παρακάτω.



Σχήμα 3.2. Γραφική Αναπαράσταση των Ακραίων Τιμών.

Οι πιο συνηθισμένες αιτίες ύπαρξης ακραίων τιμών σε ένα σύνολο δεδομένων είναι:

- Σφάλματα καταχώρησης δεδομένων (ανθρώπινα σφάλματα)
- Σφάλματα μετρήσεων (σφάλματα οργάνων)
- Πειραματικά σφάλματα (σφάλματα εξαγωγής δεδομένων ή σφάλματα προγραμματισμού/εκτέλεσης πειραμάτων)
- Εκ προθέσεως (φτιαγμένα για να εξεταστεί η ακρίβεια των μεθόδων εύρεσης των τιμών αυτών)
- Σφάλματα επεξεργασίας δεδομένων (μη επιθυμητές τροποποιήσεις κατά τη διάρκεια του χειρισμού ή της συλλογής των δεδομένων)
- Σφάλματα δειγματοληψίας (εξαγωγή ή ανάμειξη δεδομένων από λάθος ή διαφορετικές πηγές)
- Φυσικά σφάλματα (όχι λάθη αλλά καινοτομίες στα δεδομένα)

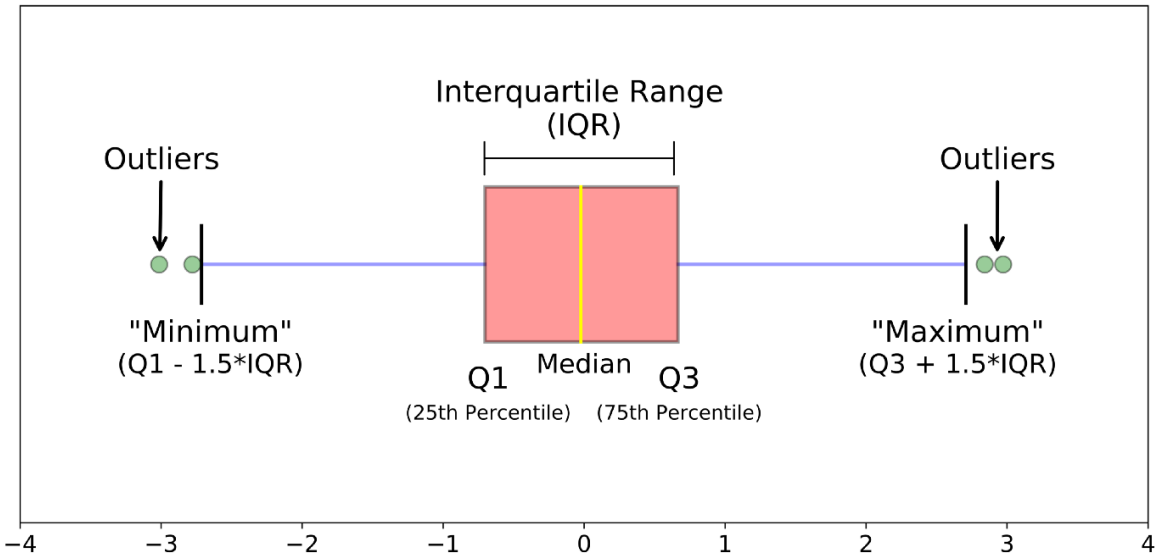
Είναι προφανές, λοιπόν, ότι κατά τη διαδικασία παραγωγής, συλλογής, επεξεργασίας και ανάλυσης δεδομένων, οι ακραίες τιμές μπορούν να προέρχονται από πολλές πηγές και να κρύβονται σε πολλές διαστάσεις. Αυτές που δεν είναι προϊόν σφάλματος ονομάζονται καινοτομίες (novelties). Η ανίχνευση ακραίων τιμών είναι μεγάλης σημασίας για σχεδόν οποιοδήποτε ποσοτικό κριτήριο. Στη μηχανική μάθηση συγκεκριμένα είναι τόσο σημαντική όσο και η ποιότητα ενός μοντέλου πρόβλεψης ή ταξινόμησης καθώς καθορίζει την ποιότητα των δεδομένων.

Για την εύρεση των τιμών αυτών υπάρχουν αρκετοί μέθοδοι όπως οι εξής:

- Z-Score or Extreme Value Analysis
- Probabilistic and Statistical Modeling
- Linear Regression Models

Η μέθοδος που επιλέχθηκε να χρησιμοποιηθεί είναι η μέθοδος IQR, σύμφωνα με την οποία για τον εντοπισμό ακραίων τιμών θα στήσουμε ένα «φράχτη» έξω από τα Q1 και Q3. Για να οικοδομήσουμε αυτό το φράχτη παίρνουμε 1,5 φορές το IQR (Διατεταρτημοριακό Εύρος) και στη συνέχεια αφαιρούμε αυτήν την τιμή από το Q1 και προσθέτουμε αυτήν την τιμή στο Q3. Οποιεσδήποτε παρατηρήσεις που είναι περισσότερο από 1,5 IQR κάτω από Q1 ή περισσότερο από 1,5 IQR πάνω από Q3 θεωρούνται ακραίες τιμές και αποκλείονται από την βάση σε επόμενο βήμα.

Ένα τέτοιο παράδειγμα φαίνεται στην παρακάτω εικόνα:



Σχήμα 3.3. Αναπαράσταση της μεθόδου IQR για τη διαχείριση ακραίων τιμών.

3.3.3 Μέθοδοι Κανονικοποίησης (Normalization)

Στο βήμα της κατασκευής του μοντέλου αναφερθήκαμε στην κανονικοποίηση δεδομένων, η οποία είναι καθοριστική για να φέρουμε τα δεδομένα στην επιθυμητή μορφή πριν εισέλθουν στο μοντέλο πρόβλεψης. Παρακάτω θα αναλύσουμε τις μεθόδους κανονικοποίησης που δοκιμάσαμε στα δεδομένα.

- **MaxAbsScaler**

Γίνεται κανονικοποίηση κάθε χαρακτηριστικού βάσει της μέγιστης απόλυτης τιμής του, έτσι ώστε η μέγιστη απόλυτη τιμή κάθε χαρακτηριστικού στο σύνολο εκπαίδευσης να είναι 1.0. Δεν μετατοπίζει ή κεντράρει τα δεδομένα, και έτσι δεν καταστρέφει την ποικιλομορφία των δεδομένων. Πρακτικά κατά την κανονικοποίηση αυτή διαιρείται κάθε παρατήρηση με τη μέγιστη τιμή της μεταβλητής. Το αποτέλεσμα είναι μια κατανομή στην οποία οι τιμές ποικίλλουν περίπου εντός του εύρους -1 έως 1. Ισχύει ο τύπος:

$$xscaled = \frac{x}{max(x)}$$

- **MinMaxScaler**

Η μέθοδος κανονικοποίησης αυτή είναι η απλούστερη μέθοδος και συνίσταται στην εκ νέου κλιμάκωση του εύρους των χαρακτηριστικών για την κλιμάκωση του εύρους στο [0, 1] ή [-1, 1]. Η επιλογή του εύρους-στόχου εξαρτάται από τη φύση των δεδομένων. Πρακτικά η κανονικοποίηση αυτή κλιμακώνει και μεταφράζει κάθε χαρακτηριστικό ξεχωριστά, έτσι ώστε να είναι σε ένα δεδομένο εύρος στο σύνολο εκπαίδευσης. Ισχύει ο τύπος:

$$xscaled = \frac{x - min(x)}{max(x) - min(x)}$$

- **StandardScaler**

Ο StandardScaler μετασχηματίζει τα δεδομένα έτσι ώστε η κατανομή τους να έχει μέση τιμή 0 και τυπική απόκλιση 1. Στην περίπτωση πολυμεταβλητών δεδομένων, αυτό γίνεται σε επίπεδο χαρακτηριστικών (με άλλα λόγια ανεξάρτητα για κάθε στήλη των δεδομένων). Με δεδομένη την κατανομή των δεδομένων, κάθε τιμή στο σύνολο δεδομένων θα κανονικοποιείται αρχικά αφαιρώντας τη μέση τιμή και στη συνέχεια διαιρώντας με την τυπική απόκλιση ολόκληρου του συνόλου δεδομένων (ή χαρακτηριστικού στην περίπτωση πολυμεταβλητής). Για μ η μέση τιμή και σ η τυπική απόκλιση ισχύει ο τύπος:

$$xscaled = \frac{x - \mu}{\sigma}$$

- **RobustScaler**

Αυτή η μορφή κανονικοποίησης αφαιρεί το διάμεσο (median) και κανονικοποιεί τα δεδομένα σύμφωνα με το εύρος των ποσοστών IQR. Το IQR είναι το εύρος μεταξύ του 1ου τεταρτημορίου και του 3ου τεταρτημορίου. Η κλιμάκωση γίνεται ανεξάρτητα σε κάθε χαρακτηριστικό, υπολογίζοντας τα σχετικά στατιστικά στοιχεία των δειγμάτων στο σύνολο εκπαίδευσης. Στη συνέχεια αποθηκεύονται η διάμεση τιμή και η περιοχή των τεταρτημορίων για να χρησιμοποιηθούν σε μεταγενέστερα δεδομένα με τη μέθοδο μετασχηματισμού (transform method). Ισχύει ο τύπος:

$$xscaled = \frac{Xi - Q1(x)}{Q3(x) - Q1(x)}$$

3.3.4 Κωδικοποίηση one-hot (One-hot Encoding)

Ένα πρόβλημα που προέκυψε στη συνέχεια και αναφέρθηκε στο ίδιο στάδιο κατασκευής του μοντέλου, αφορούσε τα κατηγορικά δεδομένα που περιείχε η βάση δεδομένων. Πιο συγκεκριμένα, μερικοί αλγόριθμοι μπορούν να λειτουργούν με κατηγορικά δεδομένα άμεσα. Πολλοί αλγόριθμοι μηχανικής μάθησης, όμως, δεν μπορούν να λειτουργήσουν άμεσα με αυτά και απαιτούν όλες οι μεταβλητές εισόδου και εξόδου να είναι αριθμητικές.

Αυτό ήταν το πρόβλημα και στη δική μας περίπτωση όπου χρειάστηκε κωδικοποίηση των κατηγορικών δεδομένων δηλαδή μετατροπή τους σε αριθμητική μορφή. Για το σκοπό αυτό χρησιμοποιήθηκε η κωδικοποίηση one-hot, δηλαδή είναι η χρήση των 2 bits δηλαδή 0 και 1 και η δημιουργία "ψευδομεταβλητών" για την αναπαράσταση των κατηγορικών δεδομένων.

Η συνάρτηση που χρησιμοποιούμε για αυτό το σκοπό ήταν η συνάρτηση get_dummies, η οποία μετέτρεψε σε dummy όλες τις κατηγορικές μεταβλητές. Οι dummy μεταβλητές είναι αυτές που παίρνουν μόνο την τιμή 0 ή 1 για να δείξουν ουσιαστικά με αυτόν τον τρόπο, την απουσία ή την παρουσία κάποιας κατηγορηματικής επίδρασης που μπορεί να αναμένεται ότι θα μετατοπίσει το αποτέλεσμα. Μπορούν να θεωρηθούν ως αριθμητικές βάσεις για ποιοτικά γεγονότα σε ένα μοντέλο παλινδρόμησης, ταξινομώντας τα δεδομένα σε αμοιβαία αποκλειόμενες κατηγορίες. Στην περίπτωση μας για παράδειγμα χρησιμοποιήθηκε για να ορίσει ένα προϊόν το οποίο ήταν σε κατάσταση καινούρια, ή χρησιμοποιημένη κτλ.

3.4 Επιλογή Μοντέλων Πρόβλεψης

Πέρα από τις μεθόδους αυτές, έγινε και η επιλογή μοντέλων επιβλεπόμενης μηχανικής εκμάθησης καθώς και μεθόδων αξιολόγησης και βελτιστοποίησης τους, τα οποία θα παρουσιαστούν αναλυτικά.

3.4.1. Μοντέλα Πρόβλεψης και Τεχνικές Βελτιστοποίησης

Όπως αναφέραμε το πρώτο στάδιο αποτελούσε μια διαδικασία προβλέψεων των τιμών των προϊόντων η οποία βασίζεται στη διαδικασία παλινδρόμησης, υποκατηγορία της επιβλεπόμενης μάθησης. Οι αλγόριθμοι οι οποίοι χρησιμοποιήθηκαν ήταν οι εξής:

- Linear Regression
- Decision Tree Regression
- Random Forest Regression
- Lasso Regression
- K-Neighbors Regression

Οι παραπάνω αλγόριθμοι έχουν αναλυθεί στο παραπάνω κεφάλαιο. Για καθέναν από αυτούς έγιναν προβλέψεις αλλά και δημιουργήθηκαν γραφικές παραστάσεις και πίνακες με τις αναμενόμενες και προβλεπόμενες τιμές.

Έπειτα, εφαρμόστηκε η διαδικασία του cross validation που περιγράψαμε παραπάνω για όλους αυτούς τους αλγορίθμους με στόχο να βρεθεί το κατάλληλο K που δίνει την καλύτερη απόδοση και το μέγεθος της βελτίωσης της απόδοσης του καλύτερου μοντέλου μετά τη διαδικασία αυτή.

Το επόμενο στάδιο, που όπως αναφέραμε είναι η ρύθμιση υπερπαραμέτρων, έγινε με δύο τρόπους.

Αρχικά, δοκιμάστηκε η αναζήτηση πλέγματος (Grid Search) η οποία είναι αναμφισβήτητα η πιο βασική μέθοδος ρύθμισης υπερπαραμέτρων. Με αυτήν την τεχνική, απλά χτίζουμε ένα μοντέλο για κάθε πιθανό συνδυασμό όλων των παρεχόμενων τιμών υπερπαραμέτρων, αξιολογώντας κάθε μοντέλο, και επιλέγοντας την αρχιτεκτονική που παράγει τα καλύτερα αποτελέσματα. Η αναζήτηση πλέγματος δεν εφαρμόζεται μόνο σε έναν τύπο μοντέλου, αλλά μπορεί να εφαρμοστεί σε όλη την εκμάθηση μηχανής για να υπολογίσει τις καλύτερες παραμέτρους για χρήση για οποιοδήποτε μοντέλο. Στην συγκεκριμένη περίπτωση την εφαρμόσαμε για τους καλύτερους σύμφωνα με τις μετρικές αλγορίθμους παλινδρόμησης δηλαδή τα decision trees, το lasso regression και το k-neighbors regression.

Έπειτα, επιλέγοντας το μοντέλο που είχε την καλύτερη απόδοση, εφαρμόσαμε αλγορίθμους για να ελέγξουμε τις υπερπαραμέτρους του. Με τις διάφορες δοκιμές καταλήγουμε στις καλύτερες τιμές των παραμέτρων και στην νέα καλύτερη απόδοση του μοντέλου.

Τέλος, δοκιμάσαμε και έναν αυτοματοποιημένο αλγόριθμο επιλογής χαρακτηριστικών (Feature Selection). Όπως έχει αναφερθεί, η απόδοση του μοντέλου μηχανικής μάθησης που χρησιμοποιούμε είναι ευθέως ανάλογη με τα χαρακτηριστικά των δεδομένων που χρησιμοποιούνται για την εκπαίδευσή του. Οι επιδόσεις του μοντέλου δηλαδή θα επηρεαστούν αρνητικά αν τα χαρακτηριστικά των δεδομένων που παρέχονται σε αυτό δεν έχουν συσχέτιση με την έξοδο του μοντέλου. Από την άλλη πλευρά, η χρήση των σχετικών χαρακτηριστικών στοιχείων μπορεί να αυξήσει την ακρίβεια του μοντέλου σημαντικά. Η διαδικασία επιλογής των χαρακτηριστικών, λοιπόν, μπορεί να οριστεί ως η διαδικασία με τη βοήθεια της οποίας επιλέγουμε από τα δεδομένα μας τα χαρακτηριστικά εκείνα που είναι πιο σχετικά με τη μεταβλητή εξόδου ή πρόβλεψης στην οποία ενδιαφερόμαστε. Ονομάζεται επίσης διαδικασία επιλογής ιδιοτήτων (attribute selection).

Η αυτόματη επιλογή χαρακτηριστικών γίνεται στην περίπτωση αυτή με τη μέθοδο Αναδρομικής Απαλοιφής Χαρακτηριστικών (Recursive Feature Elimination - RFE), ουσιαστικά μια μέθοδο κατάταξης χαρακτηριστικών μιας βάσης δεδομένων. Σύμφωνα με τη μέθοδο αυτή ενας εξωτερικός εκτιμητής αναθέτει βάρη - συντελεστές (coefficients) στα χαρακτηριστικά ενός μοντέλου και έχει ως στόχο την επιλογή χαρακτηριστικών με

αναδρομική εξέταση συνεχώς μικρότερων συνόλων χαρακτηριστικών. Πρώτον, ο εκτιμητής εκπαιδεύεται στο αρχικό σύνολο χαρακτηριστικών και η σημασία του κάθε χαρακτηριστικού λαμβάνεται μέσω οποιασδήποτε συγκεκριμένης ιδιότητας. Στη συνέχεια, τα λιγότερο σημαντικά χαρακτηριστικά περικόπτονται από το τρέχον σύνολο χαρακτηριστικών. Αυτή η διαδικασία επαναλαμβάνεται αναδρομικά στο κάθε νέο μικρότερο σύνολο μέχρι να επιτευχθεί τελικά ο επιθυμητός αριθμός χαρακτηριστικών που θα επιλεχθούν. Η διαδικασία αυτή στοχεύει στη μείωση της υπερπροσαρμογής, του χρόνου εκπαίδευσης και της αύξησης της ακρίβειας. Παρόλα αυτά στη συγκεκριμένη βάση και λόγω του business case αλλά και του αποτελέσματος του αλγορίθμου αποφασίστηκε η αποφυγή αφαίρεσης άλλων χαρακτηριστικών πέρα από αυτά που είχαν αρχικά αφαιρεθεί από εμάς πριν την εισαγωγή των μοντέλων.

3.4.2 Μετρικές Αξιολόγησης Μοντέλων Πρόβλεψης

Για την αξιολόγηση των μοντέλων πρόβλεψης χρησιμοποιήθηκαν οι παρακάτω μετρικές (Metrics for Model Assessment) :

- **Mean Squared Error (Μέσο Τετραγωνικό Σφάλμα)**

Το μέσο τετραγωνικό σφάλμα, ή MSE, είναι ένα δημοφιλές μετρικό σφάλμα για τα προβλήματα παλινδρόμησης. Το MSE υπολογίζεται ως ο μέσος όρος των τετραγωνισμένων διαφορών μεταξύ προβλεπόμενων και αναμενόμενων τιμών σε ένα σύνολο δεδομένων. Ο τύπος είναι ο εξής:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_{hat_i})^2$$, όπου:

- y_i είναι η i'οστή αναμενόμενη τιμή στο σύνολο δεδομένων,
- y_{hat_i} είναι η αντίστοιχη i'οστή προβλεπόμενη τιμή και
- η διαφορά μεταξύ αυτών των δύο τιμών υψώνεται στο τετράγωνο, με αποτέλεσμα την αφαίρεση του προσήμου και άρα τη δημιουργία θετικής τιμής σφάλματος.

Το γεγονός ότι είναι στο τετράγωνο έχει επίσης ως αποτέλεσμα τη "διόγκωση" ή "μεγέθυνση" μεγάλων σφαλμάτων. Δηλαδή, όσο μεγαλύτερη είναι η διαφορά μεταξύ των προβλεπόμενων και των αναμενόμενων τιμών, τόσο μεγαλύτερο είναι το θετικό σφάλμα που προκύπτει στο τετράγωνο. Αυτό έχει ως αποτέλεσμα να «τιμωρούνται» τα μοντέλα περισσότερο για μεγαλύτερα λάθη όταν το MSE χρησιμοποιείται ως δείκτης για την αποδοτικότητα του μοντέλου.

- **Mean Absolute Error (Μέσο απόλυτο σφάλμα)**

Το Μέσο Απόλυτο Σφάλμα, ή MAE, είναι μια δημοφιλής μετρική επειδή οι μονάδες του βαθμού σφάλματος αντιστοιχούν στις μονάδες της τιμής στόχου που προβλέπεται. Για αυτό, οι αλλαγές στο MAE είναι γραμμικές και επομένως διαισθητικές. Το MAE δεν δίνει περισσότερο ή λιγότερο βάρος σε διαφορετικούς τύπους σφαλμάτων και αντ 'αυτού οι βαθμολογίες αυξάνονται γραμμικά με αυξήσεις σε σφάλματα.

Όπως υποδηλώνει το όνομά του, το MAE υπολογίζεται ως ο μέσος όρος των απόλυτων τιμών σφάλματος. Η απόλυτη συνάρτηση ή η abs() είναι μια μαθηματική συνάρτηση που απλώς κάνει έναν αριθμό θετικό. Επομένως, η διαφορά μεταξύ μιας αναμενόμενης και μιας προβλεπόμενης τιμής μπορεί να είναι θετική ή αρνητική και αναγκάζεται να είναι θετική κατά τον υπολογισμό του MAE. Το MAE μπορεί να υπολογιστεί ως εξής:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} abs(y_i - y_{hat_i}),$$ όπου:

- y_i είναι η i'οστή αναμενόμενη τιμή στο σύνολο δεδομένων,
- το y_{hat_i} είναι η αντίστοιχη i'οστή προβλεπόμενη τιμή και
- η abs() είναι η απόλυτη συνάρτηση.

- **Root Mean Squared Error**

Όπως δηλώνει το όνομα του το RMSE είναι η ρίζα του μέσου τετραγωνικού σφάλματος. Το χρησιμοποιούμε εναλλακτικά γιατί είναι άμεσα ερμηνεύσιμο όσον αφορά τις μονάδες μέτρησης, και έτσι είναι ένα καλύτερο μέτρο της καλής προσαρμογής από έναν συντελεστή συσχέτισης.

- **R2 Score**

Η μετρική r2 είναι μια μετρική που κυμαίνεται μεταξύ 0 και 100%. Είναι στενά συνδεδεμένη με την MSE που αναλύθηκε παραπάνω αλλά πολλές φορές είναι πιο ευρέως χρησιμοποιούμενη. Ορίζεται ως το ποσοστό της διακύμανσης στην εξαρτημένη μεταβλητή που είναι προβλέψιμο από την ανεξάρτητη μεταβλητή (ή μεταβλητές). Εναλλακτικά μπορούμε να πούμε ότι αντιστοιχεί στο κλάσμα με αριθμητή την συνολική διακύμανση από το μοντέλο προς τη συνολική διακύμανση γενικά. Έτσι, αν είναι 100%, οι δύο μεταβλητές συσχετίζονται τέλεια, δηλαδή, χωρίς καμία διακύμανση. Μια χαμηλή τιμή θα έδειχνε ένα χαμηλό επίπεδο συσχέτισης, που σημαίνει ότι το μοντέλο δεν είναι τόσο έγκυρο και ακριβές. Το R2 μπορεί να υπολογιστεί ως εξής:

$$R^2 = 1 - \frac{\Sigma (y_i - y_{hat_i})^2}{\Sigma (y_i - y_{mean_i})^2}$$, όπου:

- y_i είναι η i'οστή αναμενόμενη τιμή στο σύνολο δεδομένων,
- το y_{hat_i} είναι η αντίστοιχη i'οστή προβλεπόμενη τιμή και
- η y_{mean_i} είναι η μέση τιμή.

3.4.3. Μοντέλα Ταξινόμησης

Αφού αποφασίστηκε να χρησιμοποιηθεί η εναλλακτική μέθοδος επίλυσης του προβλήματος εφαρμόστηκαν όπως είπαμε μοντέλα ταξινόμησης. Τα μοντέλα αυτά ήταν:

- Gaussian Naive Bayes Classifier
- K Neighbors Classifier
- Support Vector Machines Classifier
- Decision Tree Classifier
- Logistic Regression Classifier

Οι παραπάνω αλγόριθμοι έχουν αναλυθεί στο παραπάνω κεφάλαιο. Για καθέναν από αυτούς έγιναν ταξινομήσεις της επιθυμητής μεταβλητής.

Σε αυτό το σημείο δεν χρησιμοποιήθηκε cross validation όπως παραπάνω καθώς οι τιμές αξιολόγησης των μοντέλων ήταν ικανοποιητικές αλλά και γιατί η προσέγγιση θέλαμε να είναι σε επίπεδο δοκιμών συγκεκριμένων περιπτώσεων και όχι αυτόματης διαδικασίας βελτιστοποίησης.

3.4.4. Μετρικές Αξιολόγησης Μοντέλων Ταξινόμησης

- **Ακρίβεια (Accuracy)**

Για την αξιολόγηση των μοντέλων ταξινόμησης χρησιμοποιήθηκε η μετρική της ακρίβειας (accuracy). Η τελευταία είναι μια μετρική που συνοψίζει την απόδοση ενός μοντέλου ταξινόμησης ως τον αριθμό των σωστών προβλέψεων διαιρεμένο με το συνολικό αριθμό των προβλέψεων. Είναι εύκολη να υπολογιστεί και όσο πιο υψηλή είναι η τιμή της τόσο καλύτερο είναι το μοντέλο.

Για την εύρεση της ακρίβειας ταξινόμησης γίνεται πρώτα η χρήση ενός μοντέλου ταξινόμησης για τη δημιουργία μιας πρόβλεψης για κάθε παράδειγμα σε ένα σύνολο δεδομένων ελέγχου. Στη συνέχεια, οι προβλέψεις συγκρίνονται με τις γνωστές ετικέτες εκείνων των παραδειγμάτων στο σύνολο δοκιμών. Η ακρίβεια υπολογίζεται στη συνέχεια ως το ποσοστό των παραδειγμάτων στο σύνολο δοκιμών που προβλέφθηκαν σωστά, διαιρούμενο με όλες τις προβλέψεις που έγιναν στο σύνολο δοκιμών. Ο τύπος της είναι ο εξής:

$$Accuracy = \frac{\text{Σωστές Προβλέψεις } (Correct\ Predictions)}{\text{Συνολικές Προβλέψεις } (Total\ Predictions)}$$

Η ακρίβεια μπορεί θεωρητικά να οδηγήσει και στην εύρεση του ποσοστού σφαλμάτων εφόσον διαιρέσουμε τις ανακριβείς προβλέψεις με τις ολικές ή απλώς με τον τύπο:

$$Accuracy = 1 - \frac{\text{Ανακριβείς Προβλέψεις } (Incorrect\ Predictions)}{\text{Συνολικές Προβλέψεις } (Total\ Predictions)} = 1 - Error\ Rate$$

- **Precision και Recall**

Εναλλακτικές μετρικές οι οποίες δεν χρησιμοποιηθήκαν αλλά θα παρουσιαστούν αποτελούν η μετρική precision και η μετρική recall, δύο διαδεδομένες μετρικές για τη διαδικασία της ταξινόμησης.

Για να ορίσουμε τις μετρικές αυτές είναι σημαντικό αρχικά να ορίσουμε τέσσερις βασικές έννοιες, τις έννοιες της σωστής θετικής πρόβλεψης (true positive) και σωστής αρνητικής πρόβλεψης (true negative) και τις έννοιες της λανθασμένης θετικής πρόβλεψης (false positive) και λανθασμένης αρνητικής πρόβλεψη (false negative). Ένα αποτέλεσμα που ορίζεται ως true positive είναι ένα αποτέλεσμα όπου το μοντέλο προβλέπει σωστά τη θετική κλάση ενώ true negative σωστά την αρνητική. Αντιθέτως, ένα false positive είναι ένα αποτέλεσμα όπου το μοντέλο προβλέπει λανθασμένα την θετική κλάση ενώ false negative λανθασμένα την αρνητική.

Με βάση αυτά μπορούμε να αναλύσουμε τις δύο μετρικές που αναφέρθηκαν παραπάνω. Ειδικότερα, η μετρική precision αποτελεί ουσιαστικά μια μετρική που ποσοτικοποιεί τον αριθμό των true positive προβλέψεων που έγιναν. Επομένως, υπολογίζει την ακρίβεια της τάξης μειονότητας. Πιο συγκεκριμένα, υπολογίζεται ως ο λόγος των σωστά προβλεπόμενων θετικών παραδειγμάτων (true positive) διαιρεμένος με το συνολικό πλήθος των θετικών παραδειγμάτων που είχαν προβλεφθεί είτε σωστά είτε λανθασμένα από το μοντέλο. Ο τύπος της είναι ο εξής:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Το αποτέλεσμα είναι μια τιμή μεταξύ 0,0 για μηδενική ακρίβεια και 1,0 για πλήρη ή τέλεια ακρίβεια. Όπως είναι προφανές η μετρική precision είναι χρήσιμη αλλά δεν περιγράφει πλήρως το μοντέλο, καθώς δεν σχολιάζει πόσα πραγματικά θετικά παραδείγματα κλάσεων προβλέφθηκαν ότι ανήκουν στην αρνητική κλάση, τα αποκαλούμενα δηλαδή false negatives.

Αντιθέτως, η μετρική recall είναι ένας δείκτης που ποσοτικοποιεί τον αριθμό των σωστών θετικών προβλέψεων (true positives) από όλες τις θετικές προβλέψεις που θα μπορούσαν να είχαν γίνει, δηλαδή αυτές που είναι είτε σωστά προβλεπόμενες θετικές είτε λανθασμένα ως αρνητικές. Σε αντίθεση με την precision, λοιπόν, παρέχει μια ένδειξη χαμένων θετικών προβλέψεων. Με αυτόν τον τρόπο, προσφέρει κάποια έννοια κάλυψης της θετικής τάξης. Πιο συγκεκριμένα, υπολογίζεται ως ο λόγος των σωστά προβλεπόμενων θετικών παραδειγμάτων (true positive) διαιρεμένος με το συνολικό πλήθος των true positives αλλά και false negatives. Ο τύπος της είναι ο εξής:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Το αποτέλεσμα είναι μια τιμή μεταξύ 0.0 και 1.0 για πλήρη ή τέλεια ανάκληση.

Και οι δύο μετρικές μπορούν να χρησιμοποιηθούν για προβλήματα ταξινόμησης και μέτρησης της ακρίβειας τους μαζί με τη μετρική accuracy. Η μεγιστοποίηση της μετρικής precision ελαχιστοποιεί τον αριθμό των ψευδώς θετικών αποτελεσμάτων, ενώ η μεγιστοποίηση της recall ελαχιστοποιεί το πλήθος των ψευδώς αρνητικών. Άρα στην πρώτη περίπτωση η εστίαση είναι στα false positive αποτελέσματα ενώ στη δεύτερη στα false negative. Μερικές φορές, θέλουμε εξαιρετικές προβλέψεις της θετικής τάξης γενικότερα άρα υψηλή τιμή και στις δύο μετρικές. Κάτι τέτοιο μπορεί να είναι δύσκολο, καθώς συχνά οι αυξήσεις στην μία μετρική συχνά έρχονται σε βάρος των μειώσεων στην άλλη. Παρόλα αυτά, αντί να επιλέγουμε το ένα ή το άλλο μέτρο, μπορούμε να επιλέξουμε μια πιο γενική μετρική δηλαδή τη μετρική accuracy, η οποία είναι και αυτή που έχουμε επιλέξει για την αξιολόγηση των μοντέλων μας στην παρούσα διπλωματική.

Κεφάλαιο 4. Αναλυτική Περιγραφή Εργασίας

4.1 Εισαγωγή

Όπως αναφέρθηκε, για το πειραματικό κομμάτι της εργασίας χρησιμοποιήθηκε ένα αρχείο δεδομένων από την ιστοσελίδα data.word το οποίο ήταν ουσιαστικά μια βάση δεδομένων προϊόντων ηλεκτρονικών παροχών και των στοιχείων πώλησης και τιμολόγησης τους. Η μορφή της βάσης αυτής ήταν αρχικά 7249 σειρές και 31 διαφορετικές στήλες. Οι στήλες είχαν δεδομένα σε μορφή ακεραίων αριθμ΄βν, γραμματοσειρών αλλά και κειμένου ανάλογα με το χαρακτηριστικό του προϊόντος που όριζαν.

Παρακάτω φαίνεται ένα λεξικό των δεδομένων αυτών:

Όνομα Μεταβλητής	Τύπος Δεδομένων	Μορφή Δεδομένων	Περιγραφή	Σχόλια
asins	string	Numbers & Letters - Keyword	Το αναγνωριστικό ASIN (Amazon identifier) για το προϊόν.	Δεν διατηρήθηκε στα τελικά δεδομένα.
brand	string	Text	Το όνομα της μάρκας του προϊόντωος.	Δεν διατηρήθηκε στα τελικά δεδομένα.
categories	string	Text	Μια λίστα από λέξεις - κλειδιά κατηγοριών στις οποίες ανήκει το προϊόν.	Δημιουργία 2 νέων κατηγοριών με βάση αυτό.
dataAdded	date	20XX-XX-XXTXX:XX:XXZ	Η ημερομηνία που εισάχθηκε στη βάση δεδομένων.	Χρήση για τον υπολογισμό του Number of Years.
dataUpdated	date	20XX-XX-XXTXX:XX:XXZ	Η πιο πρόσφατη ημερομηνία που το προϊόν ανανεώθηκε στο σύστυμα.	Δεν διατηρήθηκε στα τελικά δεδομένα.
ean	string	Numbers - Keyword	Οι κωδικοί EAN για αυτό το προϊόν. Μπορεί να υπάρχουν περισσότεροι από ένας EAN στη λίστα όταν ένα προϊόν έχει πολλές παραλλαγές, με τον καθένα να χρησιμοποιεί διαφορετικό EAN.	Δεν διατηρήθηκε στα τελικά δεδομένα.
manufacturer	string	Text	Ο κατασκευαστής του ποριόντος.	Χρήση για τους primary manufacturers
merchants	string	Text	Μια λίστα εμπόρων που πωλούν αυτό το προϊόν. Αυτοί είναι συνήθως third - party έμποροι που βρίσκονται σε ιστότοπους ηλεκτρονικού εμπορίου.	Χρήση για τους primary merchants
primatyCategories	string	Text	Λίστα τυποποιημένων κατηγοριών στις οποίες ανήκει το προϊόν.	Δεν διατηρήθηκε στα τελικά δεδομένα.
sourceURLs	string	Keyword	Μια λίστα διευθύνσεων URL που χρησιμοποιούνται για τη δημιουργία δεδομένων για αυτό το προϊόν.	Δεν διατηρήθηκε στα τελικά δεδομένα.
prices.amountMax	float	Number &USD or CAD	Η ελάχιστη τιμή που αναφέρεται.	Χρησιμοποιήθηκε ο μέσος όρος της τιμής min και max
prices.amountMin	float	Number &USD or CAD	Η μέγιστη τιμή που αναφέρεται.	Χρησιμοποιήθηκε ο μέσος όρος της τιμής min και max
prices.availability	string	Keyword	Μια αληθής ή ψευδής μεταβλητή σχετική με αν αυτό το προϊόν είναι διαθέσιμο σε αυτήν την τιμή.	Απλοποίηση τιμών του.
prices.currency	string	Text	Η νομισματική μονάδα.	Μεταβολή όλων σε USD.
prices.isSale	bool	Text	Μια αληθής ή ψευδής μεταβλητή σχετική με αν αυτό το προϊόν είναι σε έκπτωση/ προσφορά.	Χρήση ως true/false.
prices.dataSeen	date	20XX-XX-XXTXX:XX:XXZ	Ένας κατάλογος των ημερομηνιών όταν χρήστες είδαν το προϊόν στη συγκεκριμένη τιμή.	Χρήση για δημιουργίας νέας στήλης και για καθορισμό μεταβλητής στόχου στην ταξινόμηση με τη δημιουργία κλάσεων.
id	string	Keyword	Το id του προϊόντος.	Δεν διατηρήθηκε στα τελικά δεδομένα.
prices.condition	string	Keyword	Η κατάσταση του προϊόντος όταν πωλείται σε αυτήν την τιμή.	Χρήση απλοποιημένης μορφής.
prices.merchant	string	Keyword	Ο έμπορος ή / και ιστοσελίδα πώλησης σε αυτήν την τιμή.	Χρήση απλοποιημένης μορφής.
prices.shipping	string	Text	Οι όροι αποστολής που σχετίζονται με αυτήν την τιμή.	Χρήση απλοποιημένης μορφής.
prices.sourceURLs	string	Keyword	Ένας κατάλογος διευθύνσεων URL στις οποίες εμφανίστηκε αυτή η τιμή.	Δεν διατηρήθηκε στα τελικά δεδομένα.
imageURLs	string	Keyword	Μια λίστα με διευθύνσεις URL εικόνων για αυτό το προϊόν.	Δεν διατηρήθηκε στα τελικά δεδομένα.
keys	string	Keyword	Μια λίστα με εσωτερικά αναγνωριστικά Datafiniti για αυτό το προϊόν. Το πεδίο χρησιμοποιείται για τη συγχώνευση ανεπεξέργαστων δεδομένων από μεμονωμένες πηγές στην κύρια εγγραφή του αρχείου δεδομένων.	Δεν διατηρήθηκε στα τελικά δεδομένα.
manufacturerNumber	string	Keyword	Ο αριθμός κατασκευαστή ή μοντέλου του προϊόντος αυτού.	Δεν διατηρήθηκε στα τελικά δεδομένα.
name	string	Keyword	Το όνομα του εμπόρου.	Δεν διατηρήθηκε στα τελικά δεδομένα.
upc	string	Keyword	Ο κωδικός UPC για αυτό το προϊόν. Μπορεί να υπάρχουν περισσότερα από ένα UPC στον κατάλογο όταν ένα προϊόν έχει πολλαπλές παραλλαγές, κάθε μια με διαφορετικό UPC.	Δεν διατηρήθηκε στα τελικά δεδομένα.
weight	float	Text (number + string of unit)	Το βάρος του προϊόντος. Περιλαμβάνονται μονάδες.	Χρήση απλοποιημένης μορφής σε pounds.

Σχήμα 4.1. Λεξικό της βάσης δεδομένων.

4.2 Ανάλυση και Επεξεργασία Δεδομένων

Το πρώτο βήμα πριν από την υλοποίηση των μοντέλων όπως είδαμε και στο προηγούμενο κεφάλαιο είναι η ανάλυση των δεδομένων και έπειτα η κατάλληλη επεξεργασία τους με στόχο να είναι σε μορφή κατάλληλη για την παραγωγή αποτελεσμάτων.

4.2.1 Κατανόηση Δεδομένων (Data Understanding)

Αρχικά, εμφανίσαμε τα στοιχεία της βάσης δεδομένων ώστε να καταλάβουμε τις διαφορετικές στήλες. Από αυτές απευθείας παρατηρείται πως χρειάζεται να απαλλαγούμε από την στήλη Unamed που δεν περιέχει καμία χρήσιμη πληροφορία για εμάς, καταλήγοντας σε μια βάση με 26 στήλες. Έπειτα, ελέγχοντας για διπλές εισαγωγές δεν έχουμε κάποια μείωση στον αριθμό των γραμμών.

4.2.2 Καθαρισμός Δεδομένων (Data Cleansing)

Ο καθαρισμός δεδομένων, όπως είδαμε και παραπάνω, είναι μια απαραίτητη διαδικασία εντοπισμού και διόρθωσης ή αφαίρεσης κατεστραμμένων, ασήμαντων, ελλιπών ή ανακριβών δεδομένων από τη βάση δεδομένων μας. Στη συγκεκριμένη βάση, στο βήμα αυτό πραγματοποιείται μια διαδικασία προσδιορισμού ημιτελών, εσφαλμένων, ανακριβών ή μη σχετικών τιμών των δεδομένων και στη συνέχεια αντικατάστασης με αντίστοιχες έγκυρες τιμές μετά από έρευνα και χρήση κατάλληλων μεθόδων αντικατάστασης ή τροποποίησης αλλά και διαγραφής αυτών όπου κρίθηκε απαραίτητο και κατάλληλο.

Μη απαραίτητα δεδομένα

Σε πρώτο στάδιο ελέγχουμε τις στήλες εκείνες που περιέχουν δεδομένα τα οποία δεν μας δίνουν κάποιες χρήσιμες για την μετέπειτα μεθοδολογία μας πληροφορίες.

Αρχικά, μια τέτοια στήλη είναι η στήλη του ID, δηλαδή ενός κωδικού που ορίζει τα διαφορετικά δεδομένα. Από αυτό και εμφανίζοντας τις μοναδικές τιμές ID παίρνουμε την πληροφορία για την ύπαρξη 835 διαφορετικά προϊόντα από τα 7249 τα οποία έχει η βάση άρα υπάρχουν πολλαπλά ίδια προϊόντα καταχωρημένα με διαφορετικά στοιχεία στα πεδία του παρόχου, των τιμών, των μεθόδων αποστολής κλπ.

Παρομοίως, ελέγχοντας την στήλη “Primary Category” η οποία όπως φαίνεται περιλαμβάνει μόνο την τιμή “Electronics” αντιλαμβανόμαστε πως δεν χρειάζεται να κρατηθεί καθώς όλα μας τα προϊόντα ανήκουν στην ευρύτερη κατηγορία των Ηλεκτρονικών Ειδών.

Τέλος, αντίστοιχη στήλη είναι η στήλη “Ean” που περιέχει στοιχεία δεδομένων που δεν είναι χρήσιμα για τα επόμενα βήματα.

Σημαντικό να σημειωθεί σε αυτό το στάδιο είναι ότι σε πρώτη όψη και οι στήλες με τα URLs θα αποτελούσαν μη σημαντικές στήλες για την μεθοδολογία αλλά αποτελούν σημαντικές στήλες για την ανάλυση και επεξεργασία δεδομένων καθώς μέσω των υπερσυνδέσμων αυτών μπορούμε να βρούμε τυχόν ελλιπή στοιχεία για τα δεδομένα.

Μη σχετικά δεδομένα - Διαφορά στις μονάδες μέτρησης

Παρατηρούμε ότι υπάρχουν κάποια αριθμητικά δεδομένα τα οποία έχουν διαφορετικές μονάδες μέτρησης οπότε σε επόμενο στάδιο θα προσπαθήσουμε να τα μετατρέψουμε σε μια κοινή μονάδα ώστε να χρησιμοποιηθούν αποδοτικά στη συνέχεια.

Αρχικά, αυτό γίνεται για τη μεταβλητή της τιμής που εμφανίζεται σε USD αλλά και σε CAD. Αποφασίζεται να γίνει η μετατροπή όλων των δεδομένων σε USD καθώς χρησιμοποιείται ευρέως αλλά και εμφανίζεται και στην μεγάλη πλειοψηφία των δεδομένων της βάσης μας.

Έπειτα, το ίδιο γίνεται για την μεταβλητή του βάρους που εμφανίζεται σε διαφορετικές μονάδες μέτρησης όπως pounds, ounces, lb κλπ. Επιλέγεται να μετατραπεί σε lb που αποτελεί το πιο σύνηθες στη βάση με τις κατάλληλες μετατροπές των μονάδων. Στη μεταβλητή αυτή, βέβαια, απαιτείται και μια επιπλέον επεξεργασία που θα αναλυθεί στο βήμα των μη έγκυρων δεδομένων παρακάτω.

Μη έγκυρα/Εσφαλμένα Δεδομένα

Εμφανίζοντας τα δεδομένα των επιμέρους στηλών αντιλαμβανόμαστε πως υπάρχουν διάφορες στήλες στις οποίες τα δεδομένα χρειάζονται τροποποίηση και διόρθωση είτε γιατί είναι εσφαλμένα είτε γιατί είναι “κατεστραμμένα” άρα μη έγκυρα.

Αρχικά, στη στήλη weight βλέπουμε ότι σε πολλές γραμμές έχουμε αποτελέσματα σε μορφή URL αντί για μια τιμή κιλών. Σε άλλες περιπτώσεις έχουμε παραπάνω από μια τιμή ενώ σε άλλες έχουμε την ίδια τιμή σε όλες τις μονάδες. Για να αντιμετωπίσουμε αυτό το πρόβλημα και να αποκτήσουμε έγκυρα δεδομένα, χρησιμοποιήσαμε το URL του προϊόντος ώστε να το αναζητήσουμε και να βρούμε τις τιμές του βάρους που εμφανίζεται. Στις περιπτώσεις που υπήρχαν διαφορετικές τιμές χρησιμοποιήσαμε το μέσο όρων αυτών.

Ακόμη, μια παρόμοια διαδικασία χρησιμοποιήθηκε για την μεταβλητή “Category” όπου και πάλι μέσω έρευνας του προϊόντος ορίστηκε η έγκυρη κατηγορία του προϊόντος στις περιπτώσεις που ήταν πολύ γενική ή όχι ακριβής.

Κενά ή Ελλιπή Δεδομένα (Null or Missing Values)

Από την αρχή έγινε έλεγχος όλων των επιμέρους μεταβλητών/στηλών για null values. Οι στήλες στις οποίες παρατηρήθηκαν null values ήταν τρεις και εξετάστηκαν διαφορετικά ώστε να γίνει ο κατάλληλος χειρισμός δηλαδή είτε αντικατάσταση των τιμών αυτών είτε τροποποίηση τους με κατάλληλες μεθόδους.

Αρχικά, για την στήλη του shipping βρέθηκαν 2972 null values, τα οποία αποφασίστηκε να αντικατασταθούν από το mode, σύμφωνα με αντίστοιχες αναφορές από πειραματικές μεθόδους. Το mode μιας κατηγορικής μεταβλητής είναι ουσιαστικά αυτή η τιμή που εμφανίζεται πιο συχνά άρα στη περίπτωση αυτή η τιμή "FREE"

Επειτα, βρέθηκαν 4014 null values στη στήλη manufacturer η οποίες επειδή αποτελούν πάνω από το 50% των δεδομένων αποφασίστηκε να αντικατασταθούν από την τιμή "Unknown", αφού δεν θα ήταν έγκυρο να αντικατασταθούν από άλλη υπάρχουσα τιμή και να οριστούν σωστά.

Η τρίτη στήλη με null values ήταν η στήλη ean που ήδη είπαμε ότι δεν έμεινε στη βάση μας.

4.2.3 Επεξηγηματική Ανάλυση Δεδομένων (Explanatory Data Analysis)

Η Επεξηγηματική Ανάλυση Δεδομένων αποτελεί το επόμενο βήμα το οποίο γίνεται τώρα σε κάποια πιο "καθαρά" δεδομένα που όμως χρειάζονται ακόμη προεπεξεργασία. Μέσω αυτής της μεθόδου, όπως προαναφέρθηκε, οδηγούμαστε σε μια διαδικασία έρευνας και περαιτέρω επεξήγησης των δεδομένων, ώστε να ανακαλυφθούν μοτίβα, να εντοπιστούν ανακρίβειες που δεν ήταν τόσο ξεκάθαρες μέχρι τώρα, να γίνει έλεγχος και εξαγωγή υποθέσεων και να παρουσιαστούν συνοπτικές στατιστικές αλλά και γραφικές αναπαραστάσεις.

Δημιουργία Μετα-Δεδομένων (Creation of Meta-Data)

Στο στάδιο της δημιουργίας μετα-δεδομένων ήταν σημαντικό να δημιουργηθούν νέες στήλες - μεταβλητές οι οποίες αποτελούν ουσιαστικά μεταδεδομένα, καθώς είναι δεδομένα που παρέχουν πληροφορίες για άλλα δεδομένα ή εναλλακτικά «δεδομένα σχετικά με δεδομένα». Πιο συγκεκριμένα, για τις μεταβλητές τις οποίες θεωρήθηκε ότι απαιτούνται πιο απλοποιημένα συμπεράσματα ή μεταβλητές που παρέχουν καλύτερες πληροφορίες για το μοντέλο, φτιάχτηκαν νέες στήλες σχετικές με αυτές. Οι στήλες αυτές μας έδωσαν σημαντικές πληροφορίες και βοήθησαν μετέπειτα στην εξαγωγή συμπερασμάτων. Οι νέες στήλες που δημιουργήθηκαν θα αναλυθούν παρακάτω συγκεκριμένα για την κάθε μεταβλητή.

Απόρριψη Δεδομένων

Έπειτα από το πρώτο στάδιο που είχαμε διαγράψει κάποιες στήλες που είχαν μη χρήσιμες πληροφορίες, μπορούμε πλέον και αφού έχουμε συμπληρώσει όλα τα ελλιπή στοιχεία να διαγράψουμε και τις υπόλοιπες.

Έτσι, οι στήλες manufacturer number, upc, sourceURLs, imageURLS, keys and asins εφόσον δεν μας δίνουν κάποιες πληροφορίες που συσχετίζουν τα προϊόντα με τις επιθυμητές μεταβλητές είναι πλέον ασήμαντες και μπορούν να διαγραφούν.

Σε αργότερο βήμα, μπορεί να διαγραφεί και η στήλη name η οποία προς το παρόν μπορεί να χρειαστεί για την επεξήγηση των δεδομένων.

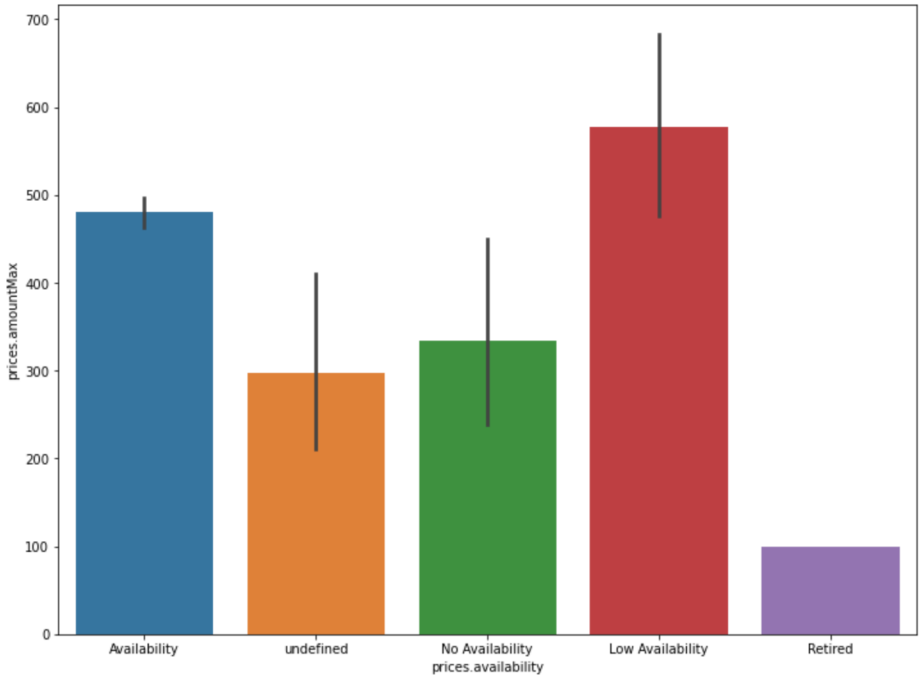
Τέλος, επιλέγουμε να μην κρατήσουμε και την στήλη manufacturer καθώς παραπάνω από τις μισές τιμές είναι "Unknown" όπως εξηγήθηκε και παραπάνω ενώ οι άλλες επικαλύπτονται και εξηγούνται καλύτερα από την μεταβλητή brand.

Επεξήγηση Δεδομένων

Για την επεξήγηση των δεδομένων και την εξαγωγή συμπερασμάτων ασχοληθήκαμε με κάθε μεταβλητή ξεχωριστά. Τα επιμέρους στοιχεία θα παρουσιαστούν παρακάτω για όλες αυτές τις μεταβλητές που αναλύθηκαν.

Διαθεσιμότητα (Availability)

Για να απλοποιηθεί η μελέτη της μεταβλητής σχετικά με τη διαθεσιμότητα χρειάστηκε να κάνουμε κάποια ομαδοποίηση των τιμών που περιείχε η βάση των δεδομένων, καθώς υπήρξαν πολλές παρόμοιες τιμές. Πιο συγκεκριμένα, λοιπόν, ομαδοποιούμε όλες τις τιμές που μας έδωσαν αληθή διαθεσιμότητα ως "Availability". Επειτα, ως "Low Availability" ορίστηκαν όλα τα προϊόντα με συγκεκριμένο αριθμό stock (π.χ. υπήρχαν τιμές όπως 7, 9 κλπ) αλλά και εκείνα που εμφανίζονταν ως "special order" και "more on the way" καθώς μετά από έρευνα τα συγκεκριμένα αποτελούν προϊόντα που οι ηλεκτρονικοί έμποροι παρουσιάζουν ως εκείνα με περιορισμένη διαθεσιμότητα. Τέλος, όσα προϊόντα ηταν μη διαθέσιμα ανατέθηκαν ως "No availability". Τέλος για εκείνα με ελλιπή στοιχεία κρατήσαμε τον ορισμό "Undefined" καθώς και το "Retired" που υπήρχε στη βάση.

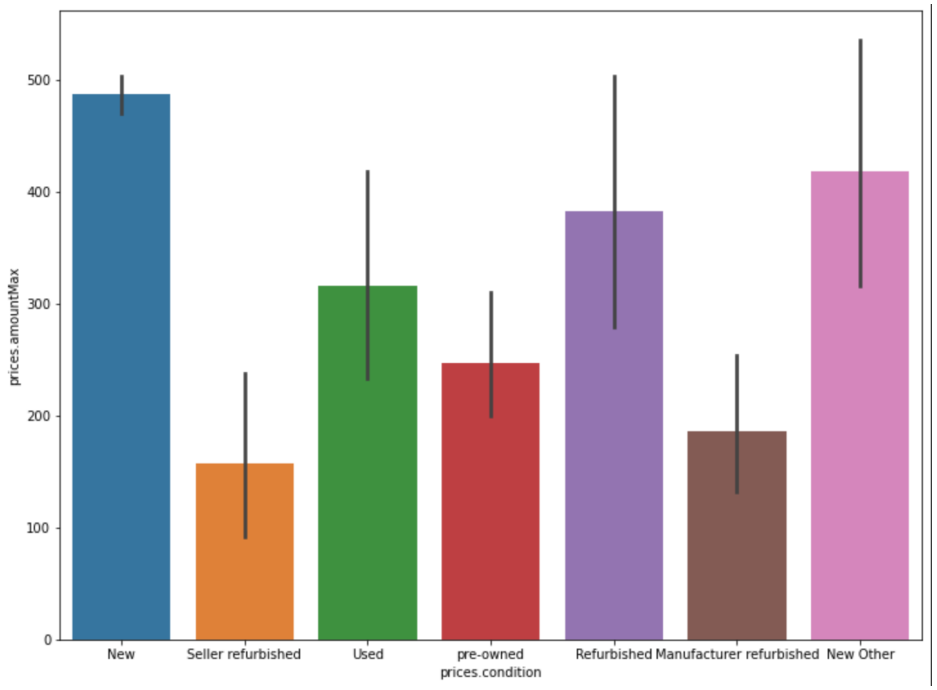


Σχήμα 4.2. Οι διαφορετικές μεταβλητές διαθεσιμότητας συγκριτικά με την τιμή των προϊόντων.

Δημιουργώντας λοιπόν το αντίστοιχο διάγραμμα καταλήγουμε στο συμπέρασμα πως οι τιμές των προϊόντων εκείνων που βρίσκονται σε περιορισμένη διαθεσιμότητα είναι πιο υψηλές σε σχέση με τα υπόλοιπα. Παρεμφερές συμπέρασμα είναι και το ότι τα προϊόντα με υψηλότερες τιμές είναι συνήθως σε περιορισμένη διαθεσιμότητα. Η μαύρη γραμμή στο διάγραμμα αντιστοιχεί στην τυπική απόκλιση σ της τιμής.

Κατάσταση (Condition)

Αντίστοιχος διαχωρισμός χρειάστηκε και σχετικά με τη στήλη της κατάστασης στην οποία βρίσκονται τα προϊόντα που πωλούνται. Χωρίσαμε τα προιοντα σε εκείνα που είναι καινούργια για όποια έχουν την τιμή “new”, καθώς και τιμές όπως “brand new” και παρεμφερείς τιμές. Αφήσαμε ξεχωριστή κατηγορία εκείνα με τιμή “New (other)” καθώς από έρευνα καταλήξαμε στο συμπέρασμα πως περιλαμβάνει προϊόντα διαφορετικά από τα καινούργια καθώς είναι αχρησιμοποίητα αλλά μπορεί να έχουν κάποιες ελλείψεις όπως να μην διαθέτουν την αρχική τους συσκευασία, να έχουν ανοιχτεί από προηγούμενο ιδιοκτήτη, ή να μην πωλούνται πλέον στην αγορά. Ακόμη, κρατήσαμε τις διαφορετικές τιμές των ανακατασκευασμένων (“Refurbished”) προϊόντων καθώς άλλα έχουν ανακατασκευαστεί από τους πωλητές και άλλα από τους κατασκευαστές. Αφήνουμε και τη γενική κατηγορία “Refurbished”, καθώς θα ήταν αλλοίωση των αποτελεσμάτων αν την προσθέταμε αυθαίρετα σε κάποια άλλη πιο ειδική κατηγορία. Τελικά, καταλήξαμε σε επτά διαφορετικές κατηγορίες σχετικά με την κατάσταση ενός προϊόντος.

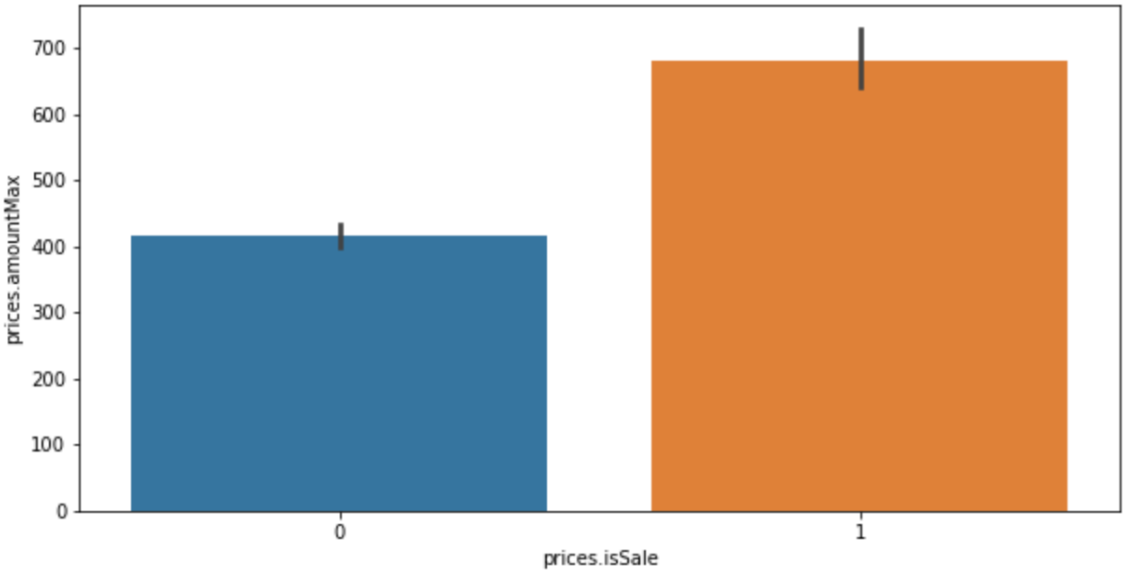


Σχήμα 4.3. Οι διαφορετικές μεταβλητές κατάστασης συγκριτικά με την τιμή των προϊόντων.

Κάνοντας το παραπάνω διάγραμμα συμπεραίνουμε ότι υψηλότερες τιμές παρατηρούνται στα καινούρια προϊόντα καθώς και στα "New other" ενώ χαμηλότερες στα ανακατασκευασμένα από τους πωλητές και κατασκευαστές. Η μαύρη γραμμή στο διάγραμμα αντιστοιχεί στην τυπική απόκλιση σ της τιμής.

Sale (Έκπτωση)

Τη στήλη που μας δίνει την πληροφορία σχετικά με το αν το προϊόν είναι σε έκπτωση ή όχι αρκεί να τη μετατρέψουμε σε δυαδική μορφή. Για το σκοπό αυτό, ομαδοποιούμε εκείνες τις τιμές που αποδεικνύουν ότι το προϊόν είναι σε έκπτωση και δίνουμε την τιμή "1" ενώ διαφορετικά την τιμή "0".



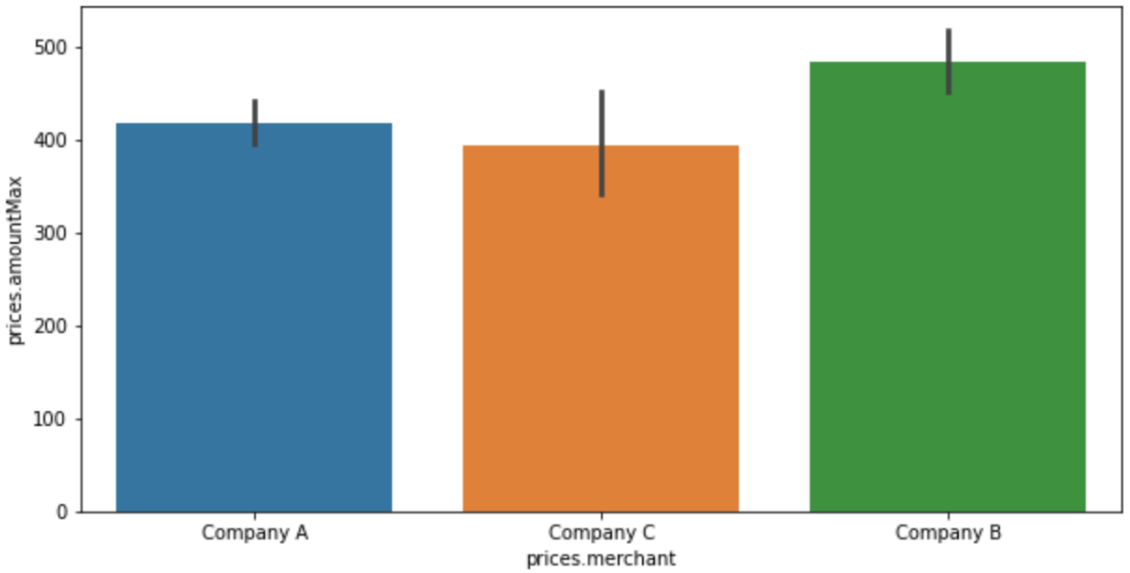
Σχήμα 4.4. Η συνθήκη σχετικά με το αν ένα προϊόν είναι σε έκπτωση ή όχι συγκριτικά με την τιμή των προϊόντων.

Από το διάγραμμα προκύπτει ότι υψηλότερες τιμές έχουν τα προϊόντα που είναι σε έκπτωση. Η μαύρη γραμμή στο διάγραμμα αντιστοιχεί στην τυπική απόκλιση σ της τιμής.

Merchant (Έμπορος)

Παρατηρούμε ότι στη βάση εμφανίζονται πολλές διαφορετικές τιμές εμπόρων κάτι το οποίο κάνει τη στήλη δύσκολη στο να τη διαχειριστούμε και να εξάγουμε επιθυμητά αποτελέσματα. Για αυτό, αποφασίζουμε αρχικά να επεξεργαστούμε τα δεδομένα ώστε να διορθώσουμε γραμματικά λαθη, διπλές τιμές με άλλη σύνταξη κτλ.

Επειτα, δημιουργούμε μια νέα στήλη την “Merchant Simplified” που περιλαμβάνει μόνο τους 3 εμπόρους με τα περισσότερα προϊόντα στη βάση και την τιμή “Other” για όλους τους άλλους. Αυτούς τους παρόχους τους ονομάζουμε Company A,B και C αντίστοιχα. Αυτός ο διαχωρισμός είναι σημαντικός, καθώς θα χρειαστεί και στη συνέχεια κατά τη μελέτη των περιπτώσεων. Για να συγκρίνουμε τις τιμές των εμπόρων κάνουμε το παρακάτω διάγραμμα.



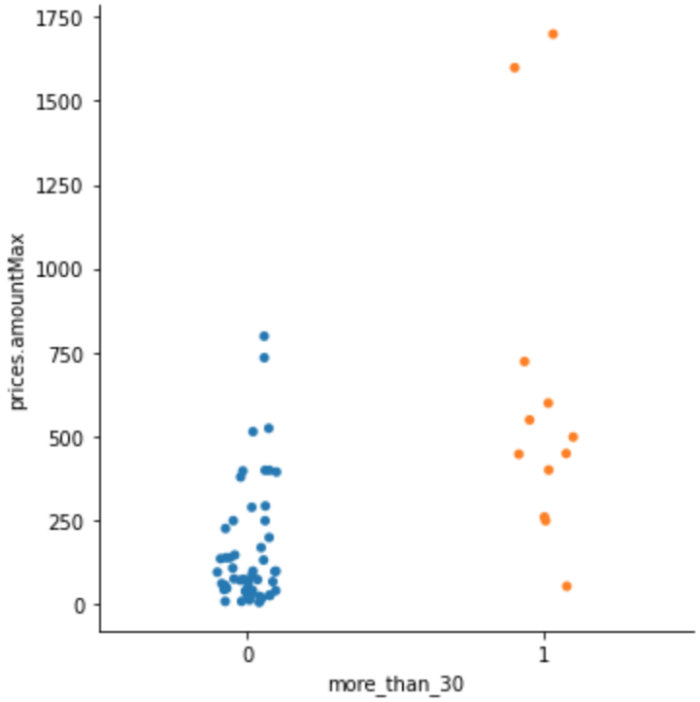
Σχήμα 4.5. Οι τρεις κύριες εταιρείες - πάροχοι συγκριτικά με την τιμή των προϊόντων τους.

Καταλήγουμε στο συμπέρασμα ότι πιο ακριβά προϊόντα με μικρή διαφορά έχει ο έμπορος που αντιστοιχεί στην μεταβλητή Company B. Η μαύρη γραμμή στο διάγραμμα αντιστοιχεί στην τυπική απόκλιση σ της τιμής.

Shipping (Αποστολή)

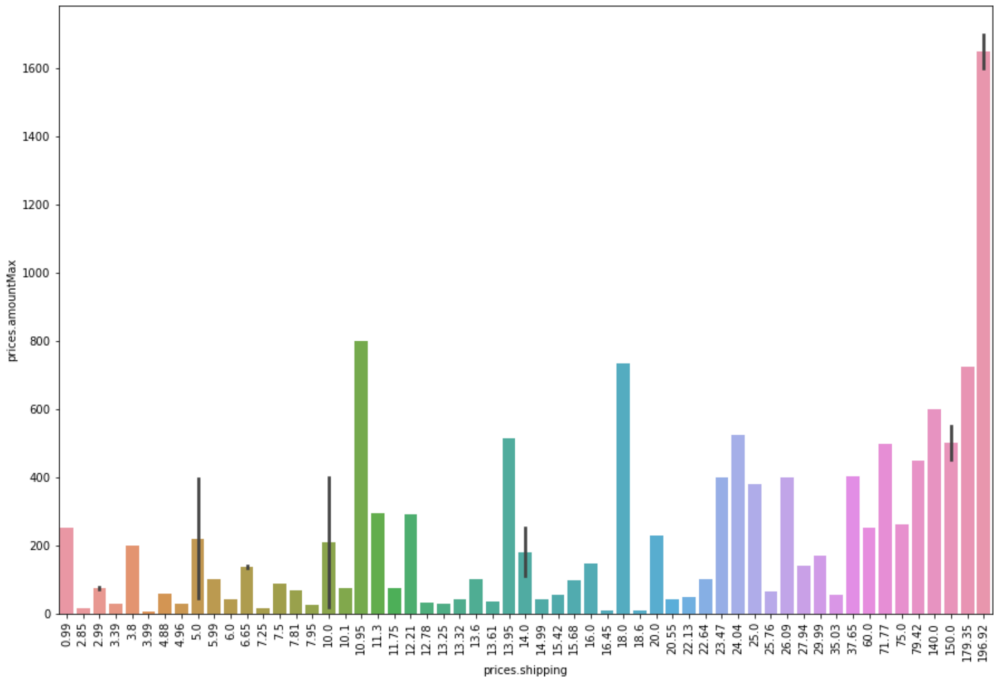
Η στήλη shipping περιλαμβάνει τις κοστολογικές πληροφορίες σχετικά με την αποστολή των προϊόντων και έχει κι αυτή περίπλοκες και διαφορετικές τιμές που χρειάζονται ομαδοποίηση και απλοποίηση για να μπορούμε να τις χρησιμοποιήσουμε και να εξάγουμε σημαντικά αποτελέσματα. Παρατηρείται επίσης ότι περιλαμβάνει και κατηγορικές τιμές (π.χ. “Free”) αλλά και αριθμητικές τιμές που αντιστοιχούν στα χρήματα σε USD που κοστίζει η αποστολή του εκάστοτε προϊόντος. Για αυτό είναι σημαντικό να διαχωρίσουμε αυτά τα δύο είδη τιμών ώστε να μπορούμε να διαχειριστούμε τη στήλη.

Αρχικά, λοιπόν, διαχωρίζουμε τα προϊόντα τα οποία έχουν τις τιμές αποστολής. Χρησιμοποιώντας αυτά τα προϊόντα μόνο, παρατηρούμε ότι είναι όλα σε USD οπότε το αφαιρούμε από την στήλη. Επειτα υπολογίζοντας την μέση τιμή των τιμών βλέπουμε ότι είναι το 30 άρα με τη βάση αυτή δημιουργούμε μια καινούρια στήλη, την “more_than_30”, που δίνουμε την τιμή 1 αν η τιμή αποστολής είναι πάνω από 30 ευρώ και 0 διαφορετικά. Ετσι, κατηγοριοποιούμε κατά κάποιο τρόπο τα προϊόντα με “ακριβό” κόστος αποστολής (δηλαδή παραπάνω από το μέσο όρο) και αντίστοιχα αυτά με το πιο “φθηνό”. Παρακάτω είναι το διάγραμμα που αντιστοιχεί σε αυτά τα προϊόντα:



Σχήμα 4.6. Η μεταβλητή του κόστους αποστολής συγκριτικά με την τιμή των προϊόντων.

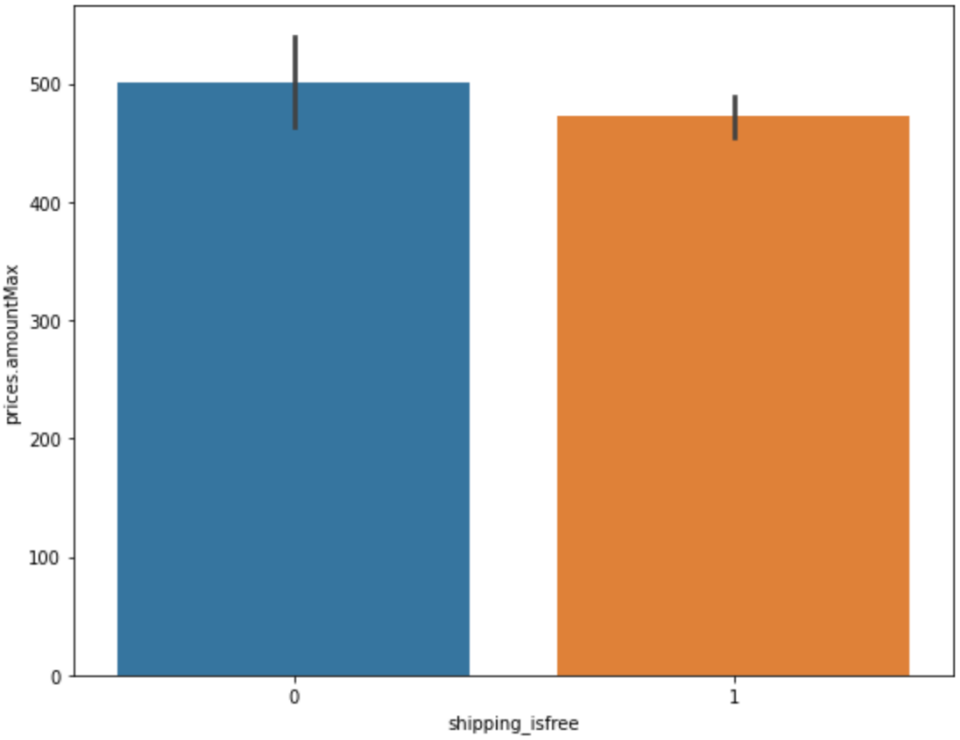
Από το διάγραμμα προκύπτει ότι δεν έχουμε κάποια σημαντική συσχέτιση με την τιμή της αποστολή και την τιμή των προϊόντων ειδικά σχετικά με το αν είναι πάνω από τα 30 ευρώ. Για να το επιβεβαιώσουμε κάνουμε και το διάγραμμα όλων των διαφορετικών προϊόντων με τιμών αποστολής καταλήγοντας στο εξής:



Σχήμα 4.7. Οι διαφορετικές τιμές αποστολής συγκριτικά με την τιμή των προϊόντων.

Συμπεραίνουμε, λοιπόν, ότι η πληροφορία για τα ακριβή κόστη αποστολής δεν είναι μια σημαντική πληροφορία για την τιμή.

Τέλος δημιουργήσαμε και την στήλη “shipping_isfree” η οποία έχει την τιμή 1 μόνο στην περίπτωση που έχουμε δωρεάν αποστολή. Παρακάτω είναι και το διάγραμμα που αντιστοιχεί σε αυτή τη στήλη:

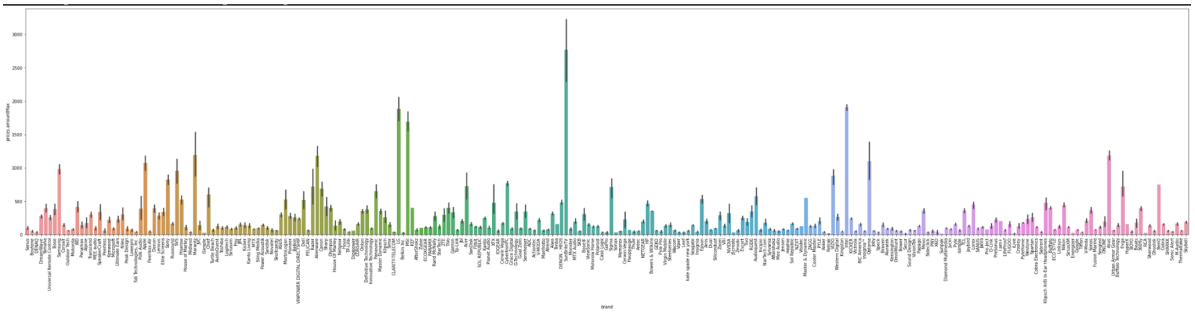


Σχήμα 4.8. Η μεταβλητή της δωρεάν αποστολής συγκριτικά με την τιμή των προϊόντων.

Παρατηρούμε ότι δεν υπάρχει μεγάλη διαφορά στις τιμές των προϊόντων με δωρεάν έξοδα μεταφοράς ή επί πληρωμή άρα η μεταβλητή αυτή δεν μας δίνει κάποια ιδιαίτερη πληροφορία. Η μαύρη γραμμή στο διάγραμμα αντιστοιχεί στην τυπική απόκλιση σ της τιμής.

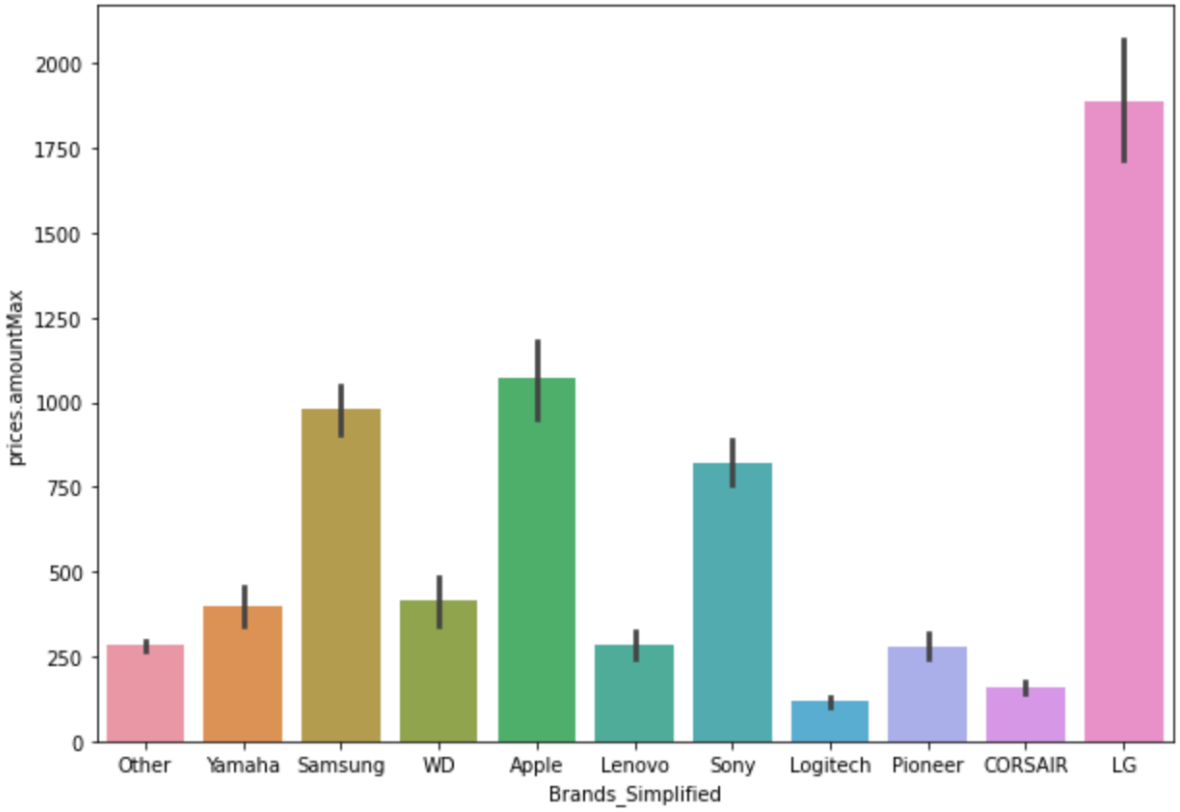
Brand (Μάρκα - Μοντέλο)

Στη στήλη με τις μάρκες - μοντέλα των προϊόντων αντιμετωπίσαμε το ίδιο πρόβλημα με αυτό της στήλης των εμπόρων, έχοντας δηλαδή πολλές διαφορετικές μάρκες για τα προϊόντα και καταλήγοντας στο εξής διάγραμμα:



Σχήμα 4.9. Οι διάφορες μάρκες συγκριτικά με την τιμή των προϊόντων.

Για το λόγο αυτό αποφασίσαμε πάλι να δημιουργήσουμε μια νέα στήλη την "Brand Simplified" στην οποία κρατήσαμε τις 10 μάρκες με τα περισσότερα προϊόντα και αποδώσαμε την τιμή "Other" στις υπόλοιπες μάρκες καταλήγοντας στο εξής διάγραμμα:



Σχήμα 4.10. Οι 10 κύριες μάρκες συγκριτικά με την τιμή των προϊόντων.

Παρατηρούμε ότι υψηλότερες τιμές έχουν τα προϊόντα LG και χαμηλότερες τα προϊόντα Logitech με τις περισσότερες όμως τιμές να κυμαίνονται σε χαμηλά επίπεδα.

Categories (Κατηγορίες)

Η στήλη των κατηγοριών όπως αναφέρθηκε παραπάνω είχε πολλές εσφαλμένες τιμές οι οποίες τροποποιήθηκαν. Παρόλα αυτά παρατηρήσαμε ότι η μορφή που η στήλη αυτή παρουσιάζει τις κατηγορίες είναι ξεκινώντας συνήθως από μια γενικότερη κατηγορία που ανήκει το προϊόν (π.χ. Computer Accessories) και καταλήγοντας σε μια πιο ειδική χωρίζοντας την κάθε υποκατηγορία με παύλες, ως εξής: Computer Accessories/Input Accessories/ Touch Pads/… Για το λόγο αυτό αποφασίσαμε να κρατήσουμε από την κάθε τιμή μόνο αυτό που βρίσκεται στην πρώτη αγκύλη και να εισάγουμε αυτή την τιμή σε μια νέα στήλη με όνομα "Primary Category".

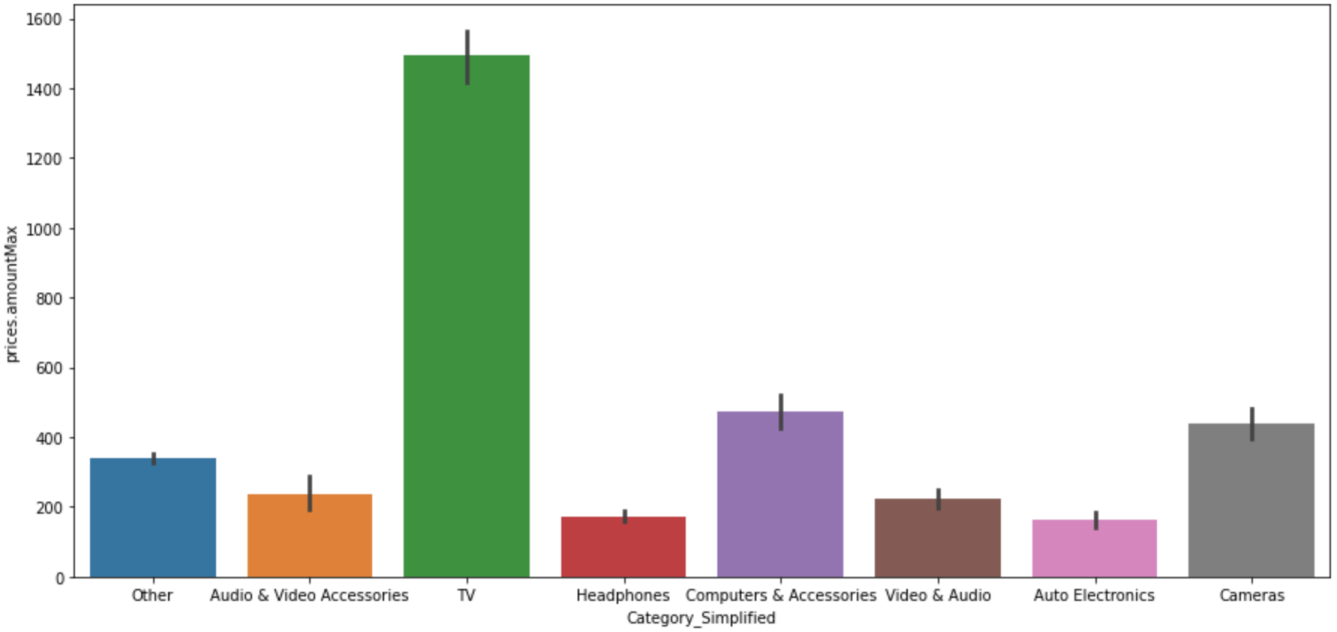
Η στήλη “Primary Category” λοιπόν περιλαμβάνει την κύρια κατηγορία στην οποία ανήκουν τα προϊόντα μας, δηλαδή τη γενικότερη από τις κατηγορίες που εντάσσονται στα ηλεκτρονικά προϊόντα. Παρόλα αυτά, παρατηρείται και πάλι μια μη ομοιομορφία στο πόσο γενική είναι αυτή η κατηγορία οπότε αποφασίζεται να γίνει ένας έλεγχος και εισαγωγή ή τροποποίηση τιμών πάλι ύστερα από έρευνα των προϊόντων.

Μετέπειτα, για να καταλήξουμε πάλι σε μια πιο συγκεκριμένη στήλη δημιουργούμε τη νέα στήλη “Primary Category Simplified” με τις 4 πιο συνήθεις κατηγορίες και την τιμή “Other” για τα προϊόντα που δεν ανήκουν σε κάποιες από αυτές. Ετσι έχουμε τις εξής κατηγορίες:

- Other
- Entertainment Electronics
- Computers & Accessories
- Entertainment Electronics Accessories
- Car Accessories

Τέλος, επειδή η κατηγορία των προϊόντων είναι καθοριστική για το πως θα τα διαχειριστούμε στη συνέχεια και επειδή οι κατηγορίες οι οποίες εμφανίστηκαν ως πιο συνήθεις ήταν επικαλυπτόμενες και δεν αντιστοιχούσαν επαρκώς σε όλες τις κατηγορίες των προϊόντων που υπάρχουν στη βάση, αποφασίστηκε η δημιουργία της νέα στήλης, της “Category Simplified”. Η στήλη αυτή περιλαμβάνει τις πιο γενικές κατηγορίες των προϊόντων που υπάρχουν και ορίστηκαν από εμάς ομαδοποιώντας κάποιες υπάρχουσες και προσπαθώντας να καλύψουμε όλες τις περιπτώσεις προϊόντων. Αυτό μπορεί να φανεί στο παρακάτω διάγραμμα με τις εξής κατηγορίες:

- Other
- TV
- Computers & Accessories
- Video & Audio
- Audio & Video Accessories
- Headphones
- Auto Electronics
- Cameras

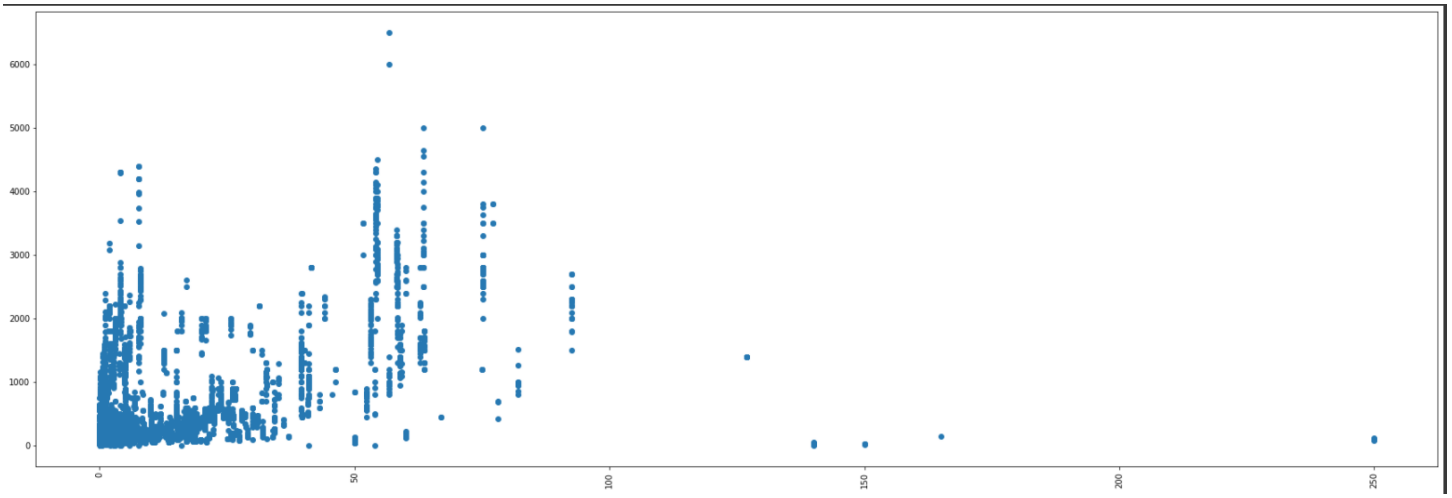


Σχήμα 4.11. Οι κύριες κατηγορίες συγκριτικά με την τιμή των προϊόντων.

Από το διάγραμμα παρατηρούμε ότι τα πιο ακριβά προϊόντα είναι αυτά που ανήκουν στην κατηγορία "TV" και έπειτα στην "Computers & Accessories" ενώ τα λιγότερο ακριβά τα "Auto Electronics" και "Headphones". Η μαύρη γραμμή στο διάγραμμα αντιστοιχεί στην τυπική απόκλιση σ της τιμής.

Weight (Βάρος)

Η μεταβλητή σχετικά με το βάρος του προϊόντος δεν είναι μια μεταβλητή η οποία μπορεί να χρησιμοποιηθεί με αποδοτικό τρόπο καθώς έχει πολλές διαφορετικές τιμές που δεν έχουν σχέση με την τιμή των προϊόντων όπως φαίνεται και στο διάγραμμα:



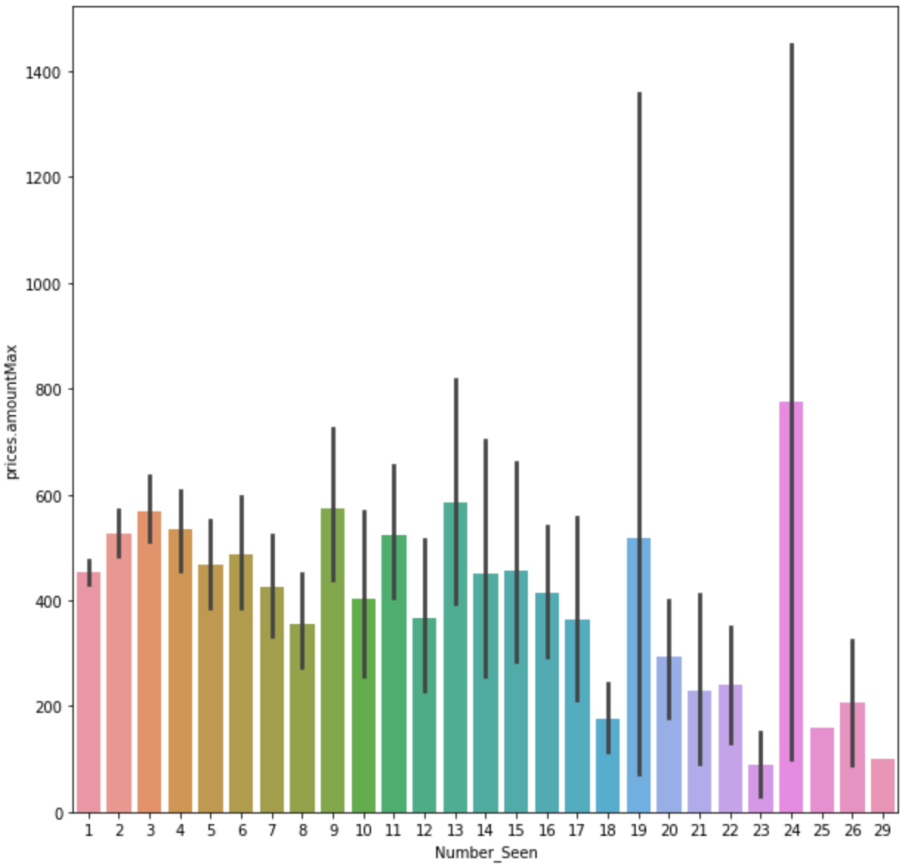
Σχήμα 4.12. Οι τιμές των προβολών των προϊόντων συγκριτικά με την τιμή των προϊόντων.

Είναι μια στήλη που είναι δύσκολο να απλοποιηθεί και να ομαδοποιηθεί οπότε δεν κάνουμε κάποιο επιπλέον βήμα στην ανάλυσή μας.

Dates (Ημερομηνίες)

Εχουμε 2 διαφορετικές στήλες με ημερομηνίες, η μια ως “dateSeen” με τις ημερομηνίες που ένα προϊόν προβλήθηκε από καταναλωτές και μια ως “dateAdded” με τις ημερομηνίες που το προϊόν προστέθηκε στην εκάστοτε ιστοσελίδα.

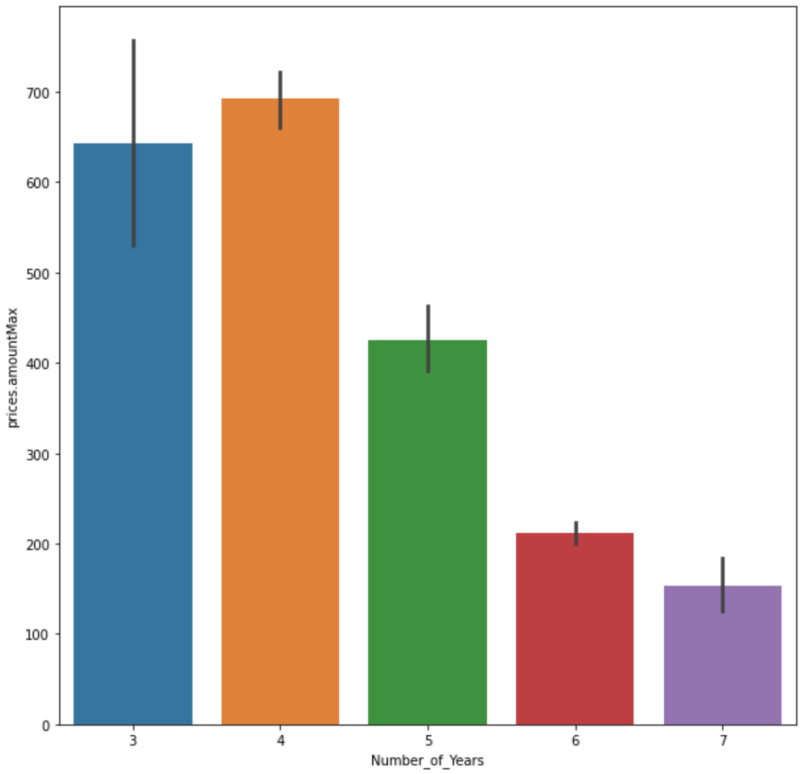
Χρησιμοποιώντας τη στήλη “dataSeen” δημιουργούμε μια νέα στήλη ως “Number_Seen” μετρώντας αυτές τις διαφορετικές ημερομηνίες για να υπολογίσουμε πόσες φορές έχει προβληθεί το κάθε προϊόν. Αυτό θα αποτελέσει στη συνέχεια και δείκτης της δημοφιλίας του προϊόντος. Αυτό θέλουμε να συσχετίσουμε και με την τιμή δημιουργώντας το παρακάτω διάγραμμα.



Σχήμα 4.13. Οι τιμές των προβολών των προϊόντων συγκριτικά με την τιμή των προϊόντων.

Από το παραπάνω διάγραμμα, λοιπόν, παρατηρούμε ότι μερικά από τα προϊόντα με τις περισσότερες προβολές έχουν υψηλότερες τιμές, χωρίς όμως αυτό να είναι απόλυτο καθώς δεν υπάρχει σχέση αναλογίας μεταξύ προβολών και τιμής. Η μεταβλητή αυτή, λοιπόν, δεν είναι αναλογικά συνδεδεμένη με την τιμή. Η μαύρη γραμμή στο διάγραμμα αντιστοιχεί στην τυπική απόκλιση σ της τιμής.

Χρησιμοποιώντας τη στήλη “dateAdded” δημιουργούμε μια νέα στήλη ως “Year_Added” ώστε να κρατήσουμε μόνο τη χρονιά που προστέθηκε το προϊόν σαν μεταβατική στήλη με σκοπό τη δημιουργία της στήλης “Number_of_Years”. Στη στήλη αυτή υπολογίζουμε και κρατάμε τον αριθμό των χρόνων που είναι στην ιστοσελίδα το κάθε προϊόν μέχρι τη χρονιά 2021 που είναι η χρονιά αναφοράς.



Σχήμα 4.14. Τα χρόνια που τα προϊόντα είναι στη βάση συγκριτικά με την τιμή των προϊόντων.

Παρατηρούμε ότι υψηλότερες τιμές έχουν τα πιο καινούρια προϊόντα στη βάση δηλαδή με τα μικρότερα χρόνια. Η μαύρη γραμμή στο διάγραμμα αντιστοιχεί στην τυπική απόκλιση σ της τιμής.

4.2.4 Έλεγχος Ακραίων Τιμών (Outliers Check)

Όπως αναφέρθηκε παραπάνω, ένα από τα βήματα της ανάλυσης δεδομένων είναι ο έλεγχος για ακραίες τιμές. Για το βήμα αυτό επιλέχθηκε η χρήση της τιμής IQR. Πιο συγκεκριμένα, τα δύο φράγματα Q1 και Q2 που τέθηκαν είναι το 0.3 και 0.7 αντίστοιχα έπειτα από δοκιμές οι οποίες ξεκίνησαν από την τιμή 0.05 και 0.95 αντίστοιχα και με βήμα 0.05. Οι τελικές τιμές που επιλέχθηκαν ήταν αυτές που εν τέλει μας έδωσαν κάποια τροποποίηση στις ακραίες τιμές της βάσης μας και άρα τα καλύτερα επιθυμητά αποτελέσματα.

4.2.5 Συσχετίσεις Μεταβλητών (Correlations)

Έπειτα χρησιμοποιώντας κατα βάση τις μεταβλητές εκείνες που έχουν μεγάλη σχετικά συσχέτιση αρχίσαμε να παράγουμε τους πίνακες συσχέτισης των μεταβλητών όπως φαίνεται παρακάτω.

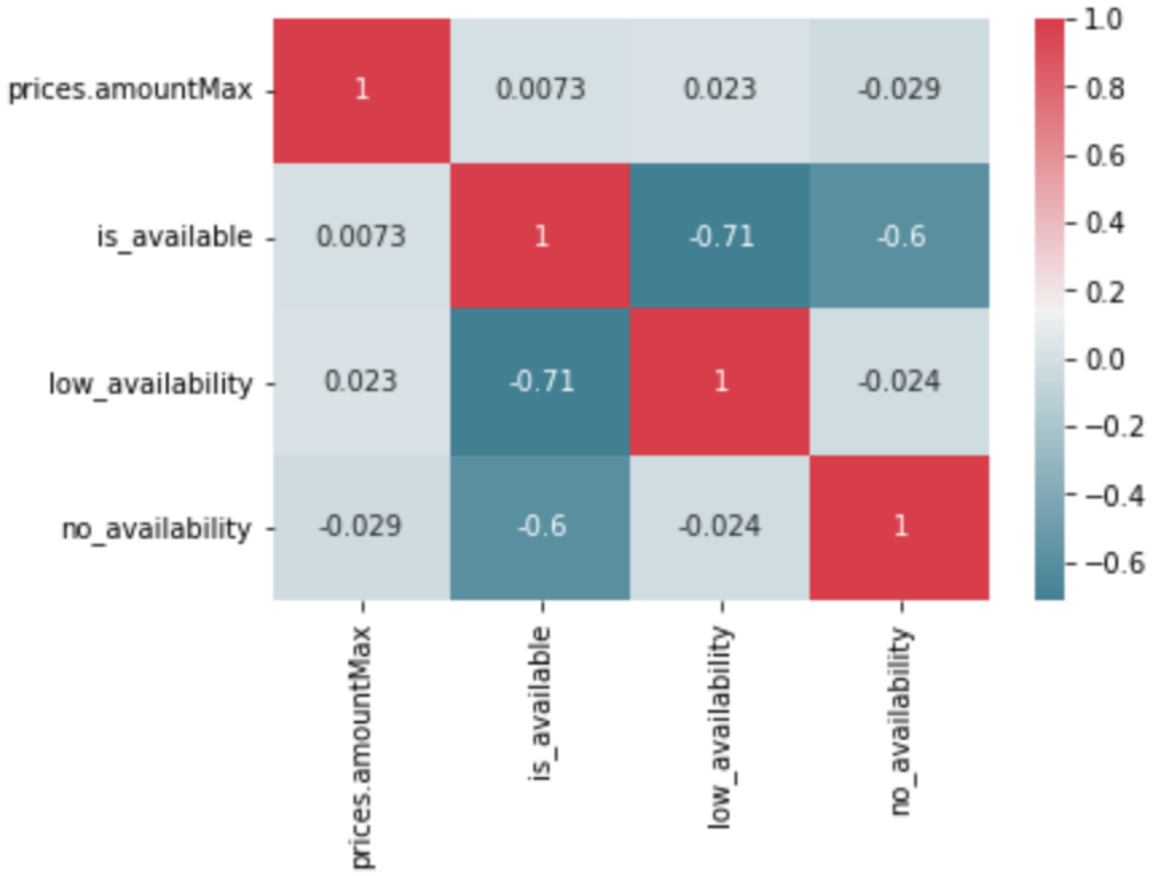
Συσχέτιση Τιμής με Βάρος και Εξοδα Αποστολής

Σχήμα 4.15. Πίνακας συσχέτισης βάρους με έξοδα αποστολής ακριβότερα από 30 και τιμή.

Σχήμα 4.16. Πίνακας συσχέτισης βάρους με δωρεάν έξοδα αποστολής και τιμή.

Όπως φαίνεται, τα προϊόντα με μεγαλύτερο βάρος έχουν αυξημένες τιμές ενώ δεν υπάρχει άμεση συσχέτιση με τα έξοδα αποστολής κάτι το οποίο φαίνεται και από τα δύο διαγράμματα όπου το ένα περιλαμβάνει τις στήλες με τις τιμές αποστολής ενώ το άλλο με την μεταβλητή σχετικά με το αν η αποστολή είναι δωρεάν ή όχι.

Συσχέτιση Τιμής με Διαθεσιμοτητα



Σχήμα 4.17. Πίνακας συσχέτισης των μεταβλητών διαθεσιμότητας και τιμή.

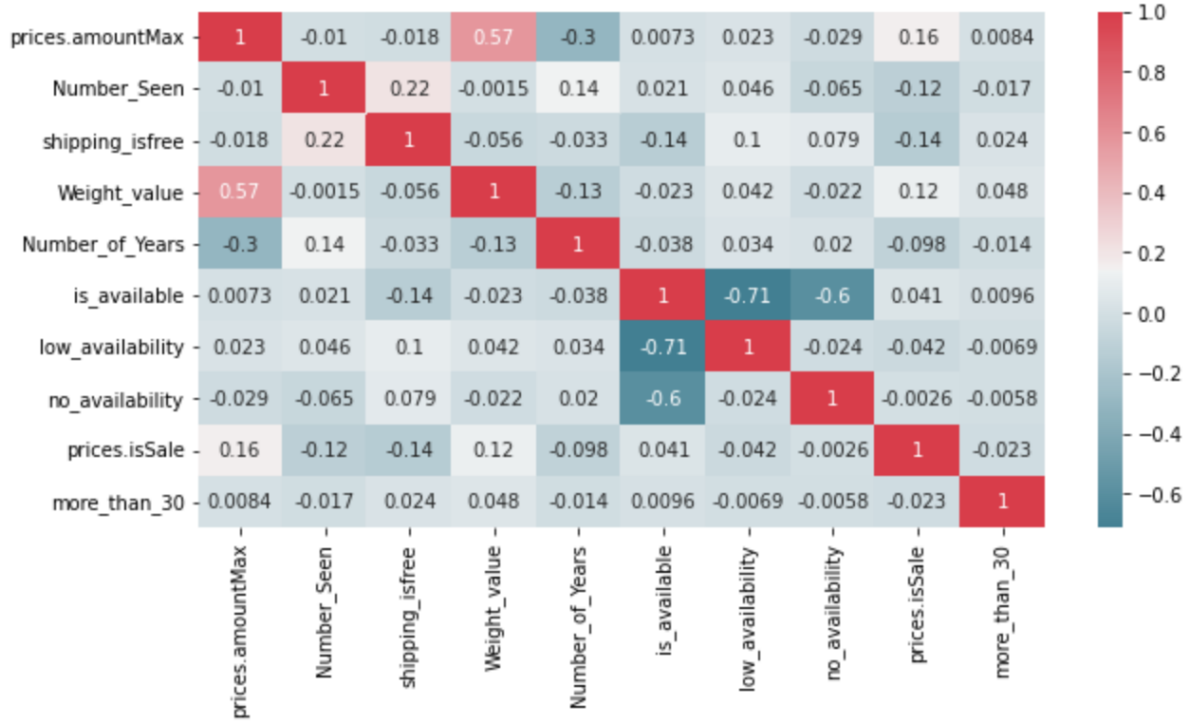
Παρατηρούμε ότι δεν έχουμε έντονη συσχέτιση τιμής με διαθεσιμότητα, ενώ μεγαλύτερη συσχέτιση παρατηρείται σε αυτά με περιορισμένη και καθόλου διαθεσιμότητα και υψηλότερες τιμές, όπως έχει διαπιστωθεί και παραπάνω.

Συσχέτιση Τιμής με Ημερομηνίες

Σχήμα 4.18. Πίνακας συσχέτισης προβολών με χρονιά που προστέθηκε και χρόνια που είναι στη βάση και τιμή.

Δημιουργούμε τον πίνακα συσχετίσεων μεταξύ της τιμής και του αριθμού των προβολών ενός προϊόντος, της χρονιάς που προστέθηκε στη βάση και του αριθμού των χρόνων που είναι μέσα στη βάση. Παρατηρούμε πως υπάρχει σχετικά μεγάλη συσχέτιση με τη χρονιά που έχει κάτι προστεθεί στη βάση και την τιμή αλλά και αρνητική συσχέτιση με το χρόνο που είναι στη βάση κάτι που είναι το αντίστοιχο. Άρα όσο πιο παλιό είναι ένα προϊόν στη βάση άρα πιο πολλά χρόνια τόσο πιο μειωμένη είναι η τιμή του και αντίστροφα για τα νέα προϊόντα. Δεν παρατηρείται έντονη συσχέτιση των προβολών με την τιμή, αλλά παρατηρούμε ότι περισσότερες προβολές έχουν τα προϊόντα που είναι και πιο πολύ καιρό στη βάση κάτι που επίσης είναι λογικό.

Συσχέτιση Όλων των Μεταβλητών



Σχήμα 4.19. Πίνακας συσχέτισης όλων των αριθμητικών και δυαδικών μεταβλητών.

Ο συνολικός πίνακας μας δίνει τα ίδια συμπεράσματα με παραπάνω αλλά και μας συσχετίζει κάποιες μεταβλητές μεταξύ τους. Συνοπτικά, λοιπόν, ισχύει ότι:

- Η τιμή του προϊόντος σχετίζεται πιο έντονα με το βάρος, αν είναι καινούργιο στη βάση και αν είναι σε έκπτωση.
- Ο αριθμός των προβολών σχετίζεται έντονα με το αν είναι δωρεάν η αποστολή αλλά και με το αν είναι νέο στη βάση.

Αξίζει να σημειωθεί ότι στο παραπάνω διάγραμμα δεν εντάχθηκαν όλες οι μεταβλητές καθώς λείπουν οι κατηγορικές μεταβλητές, τον οποίων οι κωδικοποίηση τους θα γίνει σε μεταγενέστερο στάδιο. Οι μεταβλητές αυτές δεν έχει νόημα να ενταχθούν καθώς δεν είναι σε μορφή τέτοια ώστε να παραχθεί αποτέλεσμα. Δεν θα είναι σε τέτοια μορφή ακόμη και μετά την κωδικοποίηση τους επειδή έχουν αρκετές διαφορετικές μεταβλητές που εντάσσονται άρα δημιουργούν πολλές στήλες και έτσι θα έκαναν τον πίνακα πολύ περίπλοκο και δεν θα ήταν δυνατή η εξαγωγή συμπερασμάτων. Παρόλα αυτά παρουσιάζονται εναλλακτικοί τρόποι εύρεσης της σημαντικότητας αυτών των μεταβλητών και άρα του βαθμού συσχέτισης παρακάτω.

4.2.6 Επιλογή Ανεξάρτητων Μεταβλητών (Feature Selection)

Με βάση τις παραπάνω συσχετίσεις αλλά και λόγω της δημιουργία καινούργιων στηλών αποφασίζεται ότι ορισμένες στήλες δεν προσφέρουν κάτι στη βάση μας και άρα θα τις κάνουμε drop.

Πιο συγκεκριμένα αυτές είναι:

- Η prices.availability, καθώς έχουμε δημιουργήσει τις 3 νέες διαφορετικές στήλες με τα "επίπεδα" διαθεσιμότητας (no,low, available) και η συγκεκριμένη έχει στοιχεία με μη καλά ορισμένη μορφή,
- Η prices.shipping, καθώς έχουμε δημιουργήσει τις 2 νέες διαφορετικές στήλες για τη δωρεάν και επι πληρωμή αποστολή και η συγκεκριμένη έχει στοιχεία με μη καλά ορισμένη μορφή,
- Η prices.dateSeen, καθώς έχει δημιουργηθεί η στήλη Number_Seen για τις προβολές,
- Η categories καθώς έχουμε τις 2 νέες στήλες με κύριες και απλοποιημένες κατηγορίες,
- Η στήλη brand καθώς έχουμε την απλοποιημένη μορφή,
- Η στήλη prices.merchant, καθώς έχουμε την απλοποιημένη μορφή,
- Η στήλη primary_category, καθώς έχουμε την απλοποιημένη μορφή,
- Η στήλη dateAdded, καθώς έχουμε τη στήλη με τα χρόνια που είναι στη βάση,
- Η στήλη Year_Added, καθώς έχουμε τη στήλη με τα χρόνια που είναι στη βάση,
- Η στήλη dateUpdated, καθώς δεν μας δίνει κάποια πληροφορία και τέλος ομοίως
- Η στήλη index, καθώς δεν μας δίνει κάποια πληροφορία.

Έτσι καταλήγουμε σε 10 στήλες τις οποίες κι όλας μετονομάζουμε ώστε να είναι εύκολα αναγνωρίσιμες ως εξής:

```
Int64Index: 7249 entries, 0 to 7248
Data columns (total 17 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Price                        7249 non-null   float64
 1   Condition                    7249 non-null   object
 2   Is_Sale                      7249 non-null   int64
 3   Difference of MAX-MIN        7249 non-null   object
 4   Weight_value                 7249 non-null   float64
 5   shipping_isfree              7249 non-null   int64
 6   is_available                 7249 non-null   int64
 7   low_availability             7249 non-null   int64
 8   no_availability              7249 non-null   int64
 9   Merchants_Simplified         7249 non-null   object
 10  more_than_30                 7249 non-null   float64
 11  Brands_Simplified            7249 non-null   object
 12  Primary_Category_Simplified  7249 non-null   object
 13  Category_Simplified          7249 non-null   object
 14  Number_Seen                  7249 non-null   int64
 15  Number_of_Years              7249 non-null   int64
 16  Year_Updated                 7249 non-null   object
dtypes: float64(3), int64(7), object(7)
```

Σχήμα 4.20. Τελική μορφή της βάσης δεδομένων μετά την επεξηγηματική ανάλυση.

4.3 Κανονικοποίηση και Κωδικοποίηση Δεδομένων

4.3.1. Μετατροπή κατηγορικών μεταβλητών

Το πρώτο βήμα πριν από την κανονικοποίηση των δεδομένων μας είναι να μετατρέψουμε τις κατηγορικές μεταβλητές (categorical variables) σε αριθμητική μορφή (dummy/indicator variables) μέσω της συνάρτησης get_dummies που εξηγήθηκε παραπάνω. Η τελική μορφή της βάσης δεδομένων φαίνεται παρακάτω:

```
Int64Index: 7249 entries, 0 to 7248
Data columns (total 48 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   Price                                                   7249 non-null   float6
 1   Is_Sale                                                 7249 non-null   int64
 2   Difference of MAX-MIN                                   7249 non-null   object
 3   Weight_value                                            7249 non-null   float6
 4   shipping_isfree                                         7249 non-null   int64
 5   is_available                                            7249 non-null   int64
 6   low_availability                                        7249 non-null   int64
 7   no_availability                                         7249 non-null   int64
 8   more_than_30                                            7249 non-null   float6
 9   Number_Seen                                             7249 non-null   int64
 10  Number_of_Years                                         7249 non-null   int64
 11  Year_Updated                                            7249 non-null   object
 12  Condition_Manufacturer refurbished                      7249 non-null   uint8
 13  Condition_New                                           7249 non-null   uint8
 14  Condition_New Other                                     7249 non-null   uint8
 15  Condition_Refurbished                                   7249 non-null   uint8
 16  Condition_Seller refurbished                            7249 non-null   uint8
 17  Condition_Used                                          7249 non-null   uint8
 18  Condition_pre-owned                                     7249 non-null   uint8
 19  Merchants_Simplified_Bestbuy.com                        7249 non-null   uint8
 20  Merchants_Simplified_Other                              7249 non-null   uint8
 21  Merchants_Simplified_Walmart.com                        7249 non-null   uint8
 22  Merchants_Simplified_bhphotovideo.com                   7249 non-null   uint8
 23  Brands_Simplified_Apple                                 7249 non-null   uint8
 24  Brands_Simplified_CORSAIR                               7249 non-null   uint8
 25  Brands_Simplified_LG                                    7249 non-null   uint8
 26  Brands_Simplified_Lenovo                                7249 non-null   uint8
 27  Brands_Simplified_Logitech                              7249 non-null   uint8
 28  Brands_Simplified_Other                                 7249 non-null   uint8
 29  Brands_Simplified_Pioneer                               7249 non-null   uint8
 30  Brands_Simplified_Samsung                               7249 non-null   uint8
 31  Brands_Simplified_Sony                                  7249 non-null   uint8
 32  Brands_Simplified_WD                                    7249 non-null   uint8
 33  Brands_Simplified_Yamaha                                7249 non-null   uint8
 34  Category_Simplified_Audio & Video Accessories           7249 non-null   uint8
 35  Category_Simplified_Auto Electronics                    7249 non-null   uint8
 36  Category_Simplified_Cameras                             7249 non-null   uint8
 37  Category_Simplified_Computers & Accessories             7249 non-null   uint8
 38  Category_Simplified_Headphones                          7249 non-null   uint8
 39  Category_Simplified_Other                               7249 non-null   uint8
 40  Category_Simplified_TV                                  7249 non-null   uint8
 41  Category_Simplified_Video & Audio                       7249 non-null   uint8
 42  Primary_Category_Simplified_Car Accessories             7249 non-null   uint8
 43  Primary_Category_Simplified_Computer & Accessories      7249 non-null   uint8
 44  Primary_Category_Simplified_Computers & Accessories     7249 non-null   uint8
 45  Primary_Category_Simplified_Entairtainment Electronics  7249 non-null   uint8
 46  Primary_Category_Simplified_Entairtainment Electronics Accessories  7249 non-null   uint8
 47  Primary_Category_Simplified_Other                       7249 non-null   uint8
dtypes: float64(3), int64(7), object(2), uint8(36)
memory usage: 1.3+ MB
```

Σχήμα 4.21. Τελική μορφή της βάσης δεδομένων μετά την κωδικοποίηση.

4.3.2 Δοκιμές Διαφορετικών Μεθόδων Κανονικοποίησης

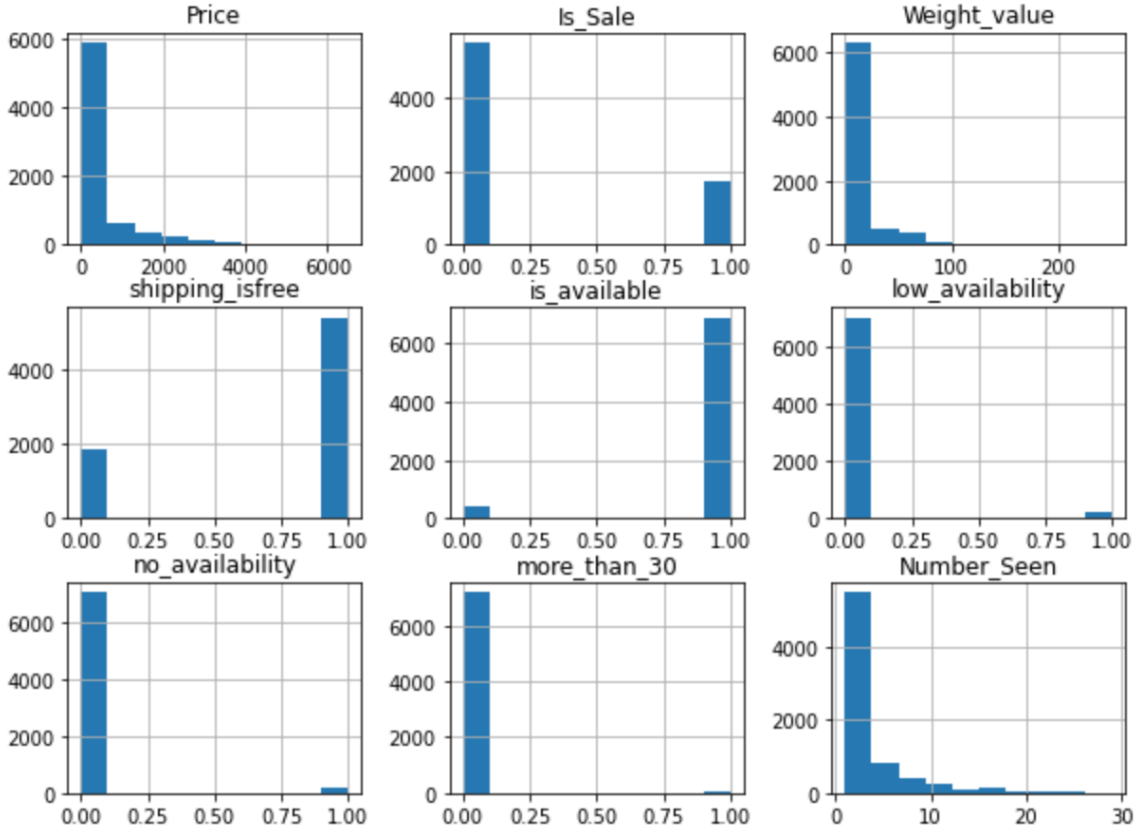
Το επόμενο βήμα αφορά την κανονικοποίηση των δεδομένων. Συγκεκριμένα θα κανονικοποιήσουμε τις αριθμητικές μεταβλητές άρα την στήλη Τιμη (Price) και Βάρος (Weight Value). Για να γίνει αυτό δοκιμάζονται οι τέσσερις διαφορετικοί μέθοδοι κανονικοποίησης που αναφέρονται παραπάνω με σκοπό να βρεθεί ο ιδανικότερος για τα δεδομένα μας. Πιο συγκεκριμένα, τα δεδομένα κανονικοποιούνται με MaxAbsScaler, MinMaxScaler, StandardScaler και RobustScaler. Παρατηρούμε ότι έχουμε παρόμοια αποτελέσματα για τους δύο πρώτους τρόπους κανονικοποίησης ενώ αρκετά μεγαλύτερες τιμές για τους άλλους δύο.

4.3.3 Επιλογή Μεθόδου Κανονικοποίησης

Από τις παραπάνω μεθόδους που δοκιμάστηκαν, εν τέλει επιλέχθηκε η μέθοδος κανονικοποίησης MinMaxScaler η οποία στην περίπτωση μας όπως αναφέρθηκε είχε τα ίδια αποτελέσματα με την MaxAbsScaler εφόσον κανονικοποιούμε μόνο τη μεταβλητή του βάρους και της τιμής.

Η επιλογή αυτή έγινε καθώς γνωρίζαμε περίπου τα άνω και κάτω όρια κατά προσέγγιση στα δεδομένα και τα τελευταία δεν είναι ιδιαίτερα ομοιόμορφα κατανεμημένα ενώ παράλληλα μας ενδιαφέρουν οι διάφορες ακραίες τιμές καθώς έχουν σημαντικό ρόλο στο κομμάτι της ανάλυσης. Έτσι η επιλογή του MinMaxScaler έγινε εφόσον επιθυμούσαμε μικρές μεταβολές και αποφυγή στρεβλώσεων. Πιο συγκεκριμένα, διατηρήθηκε το σχήμα της αρχικής κατανομής, χωρίς αλλαγή των πληροφοριών που υπήρχαν στα αρχικά δεδομένα και χωρίς μείωση της σημασίας των ακραίων τιμών.

Για την μορφή των δεδομένων και τον έλεγχο της κατανομής και των ακραίων τιμών παράγουμε τα παρακάτω ιστογράμματα:



Σχήμα 4.22. Ιστόγραμμα των μεταβλητών της βάσης δεδομένων.

Από αυτά ενδιαφερόμαστε για τα διαγράμματα της τιμής και του βάρους τα οποία παρατηρούμε προφανώς ότι δεν κατανέμονται ομοιόμορφα, επιβεβαιώνοντας τη μορφή κανονικοποίησης που επιλέχθηκε.

4.4 Παραγωγή Προβλέψεων

Όπως έχει ήδη εξηγηθεί στο κεφάλαιο 3, αρχικά τέθηκε σαν μεταβλητή - στόχος η τιμή του προϊόντος και η βάση δεδομένων χωρίστηκε σε βάση εκπαίδευσης (train set) και δοκιμών (test set) με αναλογία 85% και 15%. Το ποσοστό αυτό προέκυψε ως το βέλτιστο κάνοντας αρκετές δοκιμές στα μοντέλα. Στη συνέχεια έγιναν δοκιμές και παράχθηκαν προβλέψεις με χρήση πολλών διαφορετικών αλγορίθμων παλινδρόμησης στα πλαίσια της επιβλεπόμενης μηχανικής μάθησης. Αναλύθηκαν ήδη παραπάνω οι 5 διαφορετικοί αλγόριθμοι που τέθηκαν σε εφαρμογή.

4.4.1. Αποτελέσματα Προβλέψεων

Ακόμη, σε προηγούμενο υποκεφάλαιο αναλύθηκαν οι 4 μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση των αλγορίθμων που χρησιμοποιήθηκαν. Παρακάτω θα παρουσιαστούν τα αποτελέσματα που λάβαμε, όπου με πιο έντονη γραφή (bold) θα παρουσιάζεται το μοντέλο με την πιο αποδοτική τιμή ανά μετρική:

	Linear Regression	Decision Tree	Random Forest	Lasso Regression	KNeighbors Regression
MSE	0.01	**0.00**	**0.00**	0.01	**0.00**
MAE	0.05	0.04	0.04	0.07	**0.03**
RMSE	0.08	0.07	0.07	0.10	**0.06**
R2 Score	0.15	0.50	0.52	-3.66	**0.63**

Πίνακας 4.1. Τα αποτελέσματα των μετρικών των αλγορίθμων παλινδρόμησης.

Παρατηρούμε ότι η μετρική MSE επειδή είναι πολύ μικρή δεν μας δίνει επαρκείς πληροφορίες ενώ σε μικρές μονάδες κινούνται και οι μετρικές MAE και RMSE. Αυτό συμβαίνει γιατί τα δεδομένα μας είναι κανονικοποιημένα άρα είναι και αυτά σε μικρή κλίμακα. Ακόμη, μια τιμή που παρεκκλίνει πολύ είναι και η τιμή του Lasso για R2 Score. Παρ' όλα αυτά επιλέγεται σαν μετρική στην οποία θα δώσουμε παραπάνω βαρύτητα η μετρική R2 Score γιατί για τους υπόλοιπους αλγορίθμους μας δίνει μετρήσεις τις οποίες μπορούμε να συγκρίνουμε και να αξιολογήσουμε.

4.4.2. Επιλογή και Βελτιστοποίηση Αλγορίθμων Προβλέψεων

Όπως παρατηρούμε τα αποτελέσματα αυτά δεν είναι ικανοποιητικά άρα όπως αναφέρουμε και παραπάνω κάνουμε κάποια σειρά από βήματα για τη βελτιστοποίηση αυτών των μετρικών.

Αρχικά, ξεκινάμε τη διαδικασία βελτιστοποίησης με εφαρμογή cross validation μια μέθοδος που έχει περιγραφεί σε προηγούμενα κεφάλαια και γίνεται για όλους τους παραπάνω αλγορίθμους παλινδρόμησης με τιμές του N από 2 έως 10. Λαμβάνουμε τα εξής αποτελέσματα τα οποία παρουσιάζουμε με τρία δεκαδικα ψηφία ώστε να είναι ξεκάθαρες οι διαφοροποιήσεις και με έντονη γραφή (bold) για τον κάθε αλγόριθμο:

	Linear Regression	**Decision Tree**	**Random Forest**	**Lasso Regression**	**KNeighbors Regression**
2 Folds	0.549	0.675	0.711	0.170	0.598
3 Folds	0.556	**0.686**	0.711	0.173	0.618
4 Folds	0.553	0.683	0.707	0.172	0.615
5 Folds	0.555	0.664	**0.716**	0.173	0.623
6 Folds	0.553	0.674	0.707	0.169	**0.628**
7 Folds	**0.558**	0.683	0.715	**0.174**	**0.628**
8 Folds	0.557	0.679	0.713	0.172	**0.628**
9 Folds	0.556	0.676	0.711	0.170	**0.628**
10 Folds	0.555	0.675	0.713	0.170	**0.628**

Πίνακας 4.2. Τα αποτελέσματα του cross validation των αλγορίθμων παλινδρόμησης.

Παρατηρούμε ότι συνολικά η καλύτερη τιμή R2 Score είναι η τιμή 0.716 που προκύπτει για N = 5 Folds για τον αλγόριθμο Random Forest.

Εφαρμόζουμε αναζήτηση πλέγματος (grid search) με τους τρεις πιο αποδοτικούς αλγορίθμους σύμφωνα με τις τιμές των μετρικών που λάβαμε στο πίνακα 4.2 και λαμβάνουμε τα εξής αποτελέσματα:

	Decision Tree	Lasso Regression	KNeighbors Regression
Best R2 Score	**0.85**	0.17	0.65
Best Parameters	criterion: mse	alpha: 1	n_neighbors: 5
	splitter: best	selection: random	

Πίνακας 4.3. Τα αποτελέσματα του grid search των αλγορίθμων παλινδρόμησης.

Σύμφωνα με τα παραπάνω αποτελέσματα βλέπουμε ότι ο αλγόριθμος Decision Tree έχει τις πιο αποδοτικές μετρικές. Καταλήγουμε λοιπόν στην επιλογή του αλγορίθμου αυτού ως την καλύτερη επιλογή για τις προβλέψεις της τιμής των προϊόντων. Για τον αλγόριθμο αυτό, μέχρι τώρα η καλύτερη τιμή R2 Score που λάβαμε ήταν η τιμή 0.686 για N = 3 Folds.. Για να βελτιστοποιήσουμε ακόμη περισσότερο τις αποδόσεις του επιλέγουμε να κάνουμε τη διαδικασία ρύθμισης υπερπαραμέτρων (Hyperparameter Tuning) μέσω της αναζήτησης ενός συνόλου τιμών των υπερπαραμέτρων ενός μοντέλου που θα βελτιστοποιήσει την αρχιτεκτονική του.

Το συγκεκριμένο μοντέλο, δηλαδή το δέντρο αποφάσεων είναι ένας από τους πιο δημοφιλείς και πιο ευρέως χρησιμοποιούμενους Αλγόριθμους Μηχανικής Μάθησης λόγω της ανθεκτικότητας που παρουσιάζει στο θόρυβο, της ανοχής έναντι ελλειπουσών πληροφοριών, του χειρισμού μη σχετικών, περιττών προγνωστικών τιμών χαρακτηριστικών, του χαμηλού υπολογιστικού κόστους, της ερμηνευτικότητας, του γρήγορου χρόνου εκτέλεσης και των αποδοτικών προβλέψεων. Οι κύριοι παράμετροι του αποτελούν το μέγιστο βάθος (max depth), το ελάχιστο πλήθος δειγμάτων που απαιτούνται σε έναν κόμβο φύλλου (min samples leaf), ο μέγιστος αριθμός κόμβων φύλλου (max leaf nodes) και άλλοι.

Για να τους προσδιορίσουμε και αφού εκτελέσουμε τη διαδικασία αυτή, βρίσκουμε την καλύτερη επιλογή παραμέτρων για N = 3 Folds ως εξής:

splitter	max depth	min samples leaf	min weight fraction leaf	max features	max leaf nodes	R2 Score
best	5	4	0.1	auto	None	0.45

Πίνακας 4.4. Τα αποτελέσματα της ρύθμισης υπερπαραμέτρων για το δέντρο αποφάσεων.

Παρατηρούμε ότι για τις τιμές των παραμέτρων αυτών λαμβάνεται ως καλύτερη απόδοση για το R2 Score η τιμή 0.45. Παρατηρούμε ότι καλύτερη μέτρηση παραμένει η τιμή 0.85 που βρέθηκε κατά την εφαρμογή του grid search άρα ο αλγόριθμος δεν βελτιώθηκε.

4.4.3. Αποτίμηση Αποτελεσμάτων

Σύμφωνα με αυτά που έχουμε αναφέρει και στο κεφάλαιο 3 για τις μετρικές αξιολόγησης των αλγορίθμων παρατηρούμε ότι τα αποτελέσματα μας ακόμη και αν βελτιώθηκαν από τα αρχικά, παραμένουν ακόμη αρκετά χαμηλα σε απόδοση. Αυτό σημαίνει ότι έχουμε προβλέψεις τιμών οι οποίες είναι πολύ ανακριβείς και έτσι οι επιθυμητές προβλέψεις που θα παράξουμε δεν θα είναι σωστές και δεν θα οδηγηθούμε σε αποδοτικές στρατηγικές τιμολόγησης. Θα παράγουμε κάποιες πιθανές τιμές για τα προϊόντα που θέλουμε οι οποίες όμως θα είναι αρκετά διαφορετικές από τις θεωρητικά σωστές τιμές και έτσι δεν θα εξασφαλίσουν ότι είναι αποδοτικές για πώληση των προϊόντων.

Ακόμη, κάποια άλλα ερωτήματα που προκύπτουν έχουν να κάνουν με το πως θα εξασφαλίσουμε ότι αυτά τα προϊόντα θα προτιμηθούν από τους καταναλωτές. Εφόσον η βάση περιλαμβάνει έναν τεράστιο όγκο προϊοντων, πολλά ανταγωνιστικά μεταξύ τους, αλλά υποκατάστατα ενώ άλλα ίδια να πωλούνται όμως από διαφορετικούς παρόχους με διαφορετικές στρατηγικές πώλησης και τιμολόγησης δεν είναι ξεκάθαρο ότι ο καθορισμός τιμών με πρόβλεψη θα εξασφαλίσει ότι θα θέσει την τιμή έτσι ώστε να γίνει προτιμητέο και να πωληθεί έναντι όλων των άλλων προϊόντων. Τέλος, πέρα από την τιμή παρατηρούμε ότι υπάρχουν και άλλοι παράγοντες που παίζουν καθοριστικό ρόλο στην πώληση ενός προϊόντος, όπως για παράδειγμα τα έξοδα αποστολής, κάτι το οποίο δεν καταφέραμε να καθορίσουμε με την παραπάνω μεθοδολογία.

Αντιλαμβανόμαστε λοιπόν πως η παραπάνω προσέγγιση αποτελεί ουσιαστικά έναν πιο στατικό τρόπο καθορισμού των τιμών των προϊόντων όπου μας δίνει την τιμή που πρέπει να έχει ένα προϊόν θεωρητικά και σύμφωνα με τα χαρακτηριστικά του αφού το έχουμε συγκρίνει με τις τιμές παρόμοιων προϊόντων. Για το λόγο αυτό, η προσέγγιση μας μεταβάλλεται στοχεύοντας σε έναν εναλλακτικό τρόπο δυναμικής τιμολόγησης αλλά και ορισμού των επιμέρους χαρακτηριστικών που θα βασίζεται στο να γίνουν τα προϊόντα προτιμητέα και θα αναλυθεί περαιτέρω παρακάτω. Αυτός ο τρόπος αποτελεί μια πιο δυναμική ιδέα στρατηγικής η οποία θα εφαρμοστεί και σε συγκεκριμένες περιπτώσεις με στόχο την εξαγωγή συμπερασμάτων.

4.5 Αλλαγή Προσέγγισης

Η αλλαγή της προσέγγισης όπως αναφέρθηκε έχει ως βασικό στόχο τον καθορισμό της τιμής και των επιπρόσθετων χαρακτηριστικών των προϊόντων που δεν είναι προτιμητέα από τους πελάτες με στόχο να γίνουν πιο δημοφιλή. Για τη νέα αυτή προσέγγιση ακολουθήθηκαν κάποια νέα βήματα τα οποία θα παρουσιαστούν αναλυτικά παρακάτω.

4.5.1. Αναπροσαρμογή των Δεδομένων

Λόγω του μεγάλου όγκου της βάσης δεδομένων αλλά και για να μπορούμε να εφαρμόσουμε μετά τις προβλέψεις μας σε συγκεκριμένα παραδείγματα είναι σημαντικό να μεταβάλλουμε κατάλληλα τη βάση δεδομένων.

Για το σκοπό αυτό κρατάμε σαν μια υπο-βάση, εκείνη που προκύπτει από τα προϊόντα που πωλούν μόνο οι τρεις κύριοι πάροχοι που έχουμε ορίσει ως Company A, Company B και Company C. Η νέα αυτή βάση έχει τώρα 5017 προϊόντα έναντι των 7249 που είχε στην προηγούμενη μορφή της.

Έπειτα, δημιουργούμε μια νέα στήλη με το ονομα popularity (δημοφιλία) στην οποία καθορίζουμε τις τιμές ως "popular" και "not popular" για τα δημοφιλή και μη δημοφιλή προϊόντα αντίστοιχα. Πιο συγκεκριμένα, για να καθαρίσουμε το βαθμό δημοφιλίας των προϊόντων χρησιμοποιούμε τη στήλη "Number_Seen" που περιλαμβάνει τον αριθμό των προβολών κάθε προϊόντος. Θεωρούμε λοιπόν ως δημοφιλή τα προϊόντα εκείνα που έχουν περισσότερες από 5 προβολές και μη δημοφιλή εκείνα με 5 και λιγότερες. Το νούμερο αυτό τέθηκε σαν βάση αφού παρατηρήθηκαν οι τιμές των προβολών που κυμαίνονταν από 1 εως 29 με περισσότερα προϊόντα να έχουν λιγότερες από 10. Για το λόγο αυτό το 5 τέθηκε ως μια βάση που αναπαριστά τις λίγες προβολές και άρα τα μη δημοφιλή προϊόντα.

Στη συνέχεια, συγκρίνοντας τα id των προϊόντων παρατηρήθηκε ότι η βάση περιλαμβάνει πολλά ίδια προϊόντα τα οποία πωλούνται από διαφορετικούς ή τους ίδιους παρόχους, σε άλλη ή ίδια χρονική στιγμή και με άλλες συνθήκες πώλησης. Για το σκοπό αυτό, δημιουργήθηκε η ανάγκη κάτι τέτοιο να συγκεκριμενοποιηθεί. Ειδικότερα, εφόσον στοχεύουμε στο να καθορίσουμε την κατάλληλη στρατηγική για τα προϊόντα είναι σημαντικό να αναφερόμαστε σε προϊόντα που πωλούνται την ίδια χρονική στιγμή στις διάφορες ιστοσελίδες και από τους διάφορους παρόχους ηλεκτρονικού εμπορίου. Εφόσον τα προϊόντα μας αφορούν και ηλεκτρικά είδη, γνωρίζουμε ότι οι τιμές τους μεταβάλλονται πολύ σημαντικά με την πάροδο του χρόνου όταν εισέλθει στην αγορά το νέο πιο εξελιγμένο μοντέλο. Για το λόγο αυτό, κρατάμε πάλι σαν υπο-βάση τα προϊόντα που έχουν εισαχθεί το μήνα Ιούνιο στη βάση δεδομένων μας. Ο Ιούνιος επιλέχθηκε καθώς τότε έχουμε τα περισσότερα δεδομένα άρα θα έχουμε έναν αρκετά ικανοποιητικό αριθμό για να γίνει η εκπαίδευση των μοντέλων μας. Η πληροφορία αυτή επίσης βρέθηκε από την στήλη "Date Added" κρατώντας μόνο το μήνα που εισάχθηκε. Η νέα μας βάση έχει τώρα τη μορφή:

```
Index: 2843 entries, 1 to 7248
Data columns (total 21 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   id                            2843 non-null   object
 1   Price                         2843 non-null   float64
 2   Condition                     2843 non-null   object
 3   Is_Sale                       2843 non-null   int64
 4   Difference of MAX-MIN         2843 non-null   object
 5   Weight_value                  2843 non-null   float64
 6   shipping_isfree               2843 non-null   int64
 7   is_available                  2843 non-null   int64
 8   low_availability              2843 non-null   int64
 9   no_availability               2843 non-null   int64
 10  Merchants_Simplified          2843 non-null   object
 11  more_than_30                  2843 non-null   float64
 12  Brands_Simplified             2843 non-null   object
 13  Primary_Category_Simplified   2843 non-null   object
 14  Category_Simplified           2843 non-null   object
 15  Number_Seen                   2843 non-null   int64
 16  Number_of_Years               2843 non-null   int64
 17  Date                          2843 non-null   object
 18  Month                         2843 non-null   object
 19  Year_Updated                  2843 non-null   object
 20  popularity                    2843 non-null   object
dtypes: float64(3), int64(7), object(11)
```

Σχήμα 4.23. Η νέα βάση δεδομένων με τα προϊόντα των εταιρειών A,B, C τον Ιούνιο.

Παρατηρούμε, λοιπόν, ότι πλέον έχουμε 2843 σειρές και 21 στήλες (λόγω των δύο επιπλέον στηλών για το μήνα και τη δημοφιλία των προϊόντων).

Παρόλα αυτά παρατηρούμε πως η βάση μας ακόμη περιλαμβάνει δεδομένα με το ίδιο id που πολλές φορές αντιστοιχούν σε προϊόντα που πουλάει ο ίδιος πάροχος με διαφορετική τιμή και αλλαγή αυτής της τιμής εντός του μήνα Ιουνίου. Για την επίλυση του προβλήματος αυτού αποφασίζουμε για τα προϊόντα που συμβαίνει αυτό να κρατήσουμε ως τιμή το μέσο όρο των τιμών που εμφανίζονται. Έτσι, λοιπόν, κρατώντας αυτή τη νέα τιμή και διαγράφοντας τώρα τα διπλότυπα δεδομένα (duplicates), καταλήγουμε στην τελική μορφή της βάσης η οποία έχει την ίδια εικόνα μεταβλητών με την παραπάνω διαθέτοντας όμως 1433 μοναδικές γραμμές προϊόντων των εταιρειών A,B,C τον μήνα Ιούνιο.

Τα επόμενα βήματα της διαδικασίας είναι παρόμοια με την μοντελοποίηση που έγινε για τους αλγορίθμους παλινδρόμησης. Πιο συγκεκριμένα αυτή τη φορά κάνουμε μόνο κωδικοποίηση των δεδομένων με την ίδια μέθοδο get_dummies λαμβάνοντας εν τέλει την τελική βάση έτοιμη για την εφαρμογή σε αλγορίθμους ταξινόμησης.

4.5.2. Ταξινόμηση Δεδομένων

Το επόμενο βήμα, λοιπόν, αφορά σε πλήρη αντιστοιχία με πριν, την προετοιμασία του μοντέλου για την ταξινόμηση, την εκπαίδευση και αξιολόγηση των αλγορίθμων. Ομοίως με παραπάνω λοιπόν χωρίζουμε τη βάση σε δεδομένα εκπαίδευσης και δοκιμών αυτή τη φορά με αναλογία 70% και 30% που προέκυψε ως η καλύτερη εκδοχή έπειτα από αντίστοιχες δοκιμές. Ακόμη, θέτουμε σαν στήλη - στόχο τη στήλη "popularity_not_popular" η οποία έχει την τιμή 1 για τα προϊόντα εκείνα που είναι μη δημοφιλή και την τιμή 0 για εκείνα που είναι δημοφιλή. Άρα αυτές είναι και οι δύο κλάσεις οι οποίες θέλουμε να προβλέψουμε εξάγοντας συμπεράσματα για την δημοφιλία των προϊόντων.

Θέτοντας σε εφαρμογή τους πέντε αλγόριθμους και λαμβάνοντας τα αποτελέσματα της μετρικής της ακρίβειας τόσο για το σύνολο εκπαίδευσης όσο και για το σύνολο δοκιμής έχουμε τα παρακάτω νούμερα, όπου με πιο έντονη γραφή (bold) παρουσιάζεται το μοντέλο με την πιο αποδοτική τιμή ανά μετρική ως εξής:

	Gaussian Naive Bayes Classifier	**KNeighbors Classifier**	**SVM Classifier**	**Decision Tree Classifier**	**Logistic Regression Classifier**
Accuracy on training set	0.52	0.77	0.70	**0.94**	0.73
Accuracy on test set	0.46	0.57	**0.72**	0.57	0.71

Πίνακας 4.5. Τα αποτελέσματα των μετρικών των αλγορίθμων παλινδρόμησης.

Από αυτά μας ενδιαφέρει περισσότερο η ακρίβεια στο test set δηλαδή στη βάση δοκιμών καθώς έτσι αξιολογείται το μοντέλο πιο αποδοτικά.

4.5.3. Αποτίμηση Αποτελεσμάτων

Παρατηρούμε ότι τα αποτελέσματα της ακρίβειας που λαμβάνουμε είναι αρκετά ικανοποιητικά με τις τιμές της ακρίβειας να είναι σχετικά υψηλές. Από αυτές επιλέγεται η καλύτερη τιμή ακρίβειας στη βάση δοκιμών άρα η τιμή 0.72 του αλγορίθμου SVM Classifier. Άρα επιλέγεται αυτός ο αλγόριθμος ως ο επιθυμητός για τη συνέχεια της διαδικασίας. Πράγματι, γνωρίζουμε ότι η SVM είναι μια μέθοδος με καλύτερη απόδοση για πολλές εφαρμογές και ένας καλός ταξινομητής ειδικά αν υπάρχει ένα πρόβλημα δύο κλάσεις, με ισορροπημένα σύνολα δεδομένων και χωρίς θόρυβο ή με λίγο θόρυβο, όπως στην περίπτωσή μας.

Χρησιμοποιώντας, λοιπόν, αυτόν τον ταξινομητή παράγουμε τις προβλέψεις τις οποίες τις προσθέτουμε σε μια νέα βάση δεδομένων με στήλη τα αποτελέσματα των προβλέψεων άρα τις δυαδικές τιμές 0 και 1 για τα δημοφιλή και μη δημοφιλή προϊόντα αντίστοιχα. Τέλος, ενώνουμε αυτή τη βάση με την αρχική μας, έχοντας λοιπόν όλα τα στοιχεία των δεδομένων μας αλλά και το αν προβλέφθηκε από τον ταξινομητή ότι είναι δημοφιλή ή όχι.

Εδώ, ολοκληρώνεται και η διαδικασία που επιθυμούμε καθώς έχουμε ουσιαστικά μια νέα βάση δεδομένων όπου τα προϊόντα μας έχουν αξιολογηθεί σε σχέση με τη δημοφιλία τους. Πάνω σε αυτή τη νέα βάση είμαστε σε θέση να πραγματοποιήσουμε τις δοκιμές ώστε να ξέρουμε τις τιμές και τους παράγοντες που θα μετατρέψουν τα προιοντα σε δημοφιλή και προτιμητέα. Η δοκιμές αυτές γίνονται με τη δημιουργία κάποιων σεναρίων και την μελέτη περιπτώσεων πάνω σε αυτά τα σενάρια. Η διαδικασία αυτή, λοιπόν, θα παρουσιαστεί στη συνέχεια.

Κεφάλαιο 5. Αποτελέσματα και Μελέτη Περιπτώσεων

Στη συνέχεια της παραπάνω διαδικασίας και στοχεύοντας σε μια πιο δυναμική στρατηγική ανάθεσης τιμων αποφασίστηκε η μελέτη κάποιων ειδικών περιπτώσεων σύμφωνα με τις οποίες θα αξιολογείται κάθε φορά η δημοφιλία των προϊόντων και το αν αυτό μεταβλήθηκε ή όχι λόγω των αλλαγών που πραγματοποιήθηκαν. Οι αλλαγές αυτές αφορούσαν μεταβολές τόσο στις τιμές όσο και σε κάποια άλλα χαρακτηριστικά των προϊόντων που κρίθηκαν σημαντικά. Ο απώτερος σκοπός είναι η διατύπωση μίας γενικότερης στρατηγικής τιμολόγησης αλλά και πώλησης που θα προσφέρει σε ένα προϊόν περισσότερες προβολές και άρα δημοφιλία και μεγαλύτερης πιθανότητα αγοράς.

Για τη διαδικασία αυτή χρησιμοποιήθηκε ο ταξινομητής K-Neighbors, καθώς παρόλο που εμφάνισε χαμηλότερες τιμές ακρίβειας στη βάση δοκιμών, είχε υψηλές τιμές στη βάση εκπαίδευσης και παρουσίασε μεγαλύτερη ευαισθησία στα αποτελέσματα, δηλαδή περισσότερες τιμές 0 στην τελική ταξινόμηση έναντι των άλλων ταξινομητών που είχαν σχεδόν όλα τα αποτελέσματα ως 1 κάτι το οποίο δεν θα βοηθούσε στη συνέχεια με την μελέτη των διαφορετικών περιπτώσεων. Οι προβλέψεις δηλαδή που έγιναν εδώ ήταν με τον ταξινομητή knn και τη δημιουργία της νέας βάσης δεδομένων με τη διαδικασία που αναφέρθηκε παραπάνω αλλά με αυτές τις προβλέψεις.

5.1 Αξιολόγηση Σεναρίων

Για τις διάφορες περιπτώσεις που θέλαμε να δοκιμάσουμε, εφαρμόστηκαν κάποια case studies τα οποία μπορούμε να τα χωρίσουμε σε 3 μεγάλες κατηγορίες με 6 υποπεριπτώσεις για την κάθε μια. Το κάθε case study έγινε για μια άλλη εταιρία άρα έχουμε τα 3 case study για τις 3 εταιρίες αντίστοιχα και έπειτα τις 6 υποπεριπτώσεις μεταβολών ως εξής:

1. Μείωση τιμής προϊόντος κατά 5%
2. Μείωση τιμής προϊόντος κατά 10%
3. Μείωση τιμής προϊόντος κατά 20%
4. Αλλαγή εξόδων αποστολής σε δωρεάν έξοδα αποστολής για τα προϊόντα
5. Αλλαγή κατάσταση διαθεσιμότητας προϊόντος σε διαθέσιμο
6. Αλλαγή κατάστασης προϊόντος σε καινούριο

Παρακάτω θα παρουσιάσουμε τα αποτελέσματα για αυτές τις 6 περιπτώσεις και για κάθε εταιρεία ξεχωριστά.

5.1.1. Case Study I - Εταιρεία C

Αρχικά ξεκινάμε από την εταιρεία C καθώς έχει τη μικρότερη βάση δεδομένων άρα μπορούμε εύκολα να βγάλουμε σαφή αποτελέσματα. Αρχικά η βάση αυτή περιέχει 21 προϊόντα που είναι μη δημοφιλή και 8 που είναι δημοφιλή. Για την μελέτη της περίπτωσης αυτής παίρνουμε μόνο τα 21 προϊόντα που είναι μη δημοφιλή σε μια νέα βάση. Θα εξετάσουμε πως αλλάζει η κλάση τους, άρα η κατάσταση της δημοφιλίας τους, μετά τις μεταβολές που θα κάνουμε.

Πιο συγκεκριμένα τα αποτελέσματα που παίρνουμε είναι:

1. Με τη μείωση της τιμής κατά 5% έχουμε 19 μη δημοφιλή προϊόντα άρα 2 προϊόντα έγιναν δημοφιλή άρα μια μεταβολή κατά $\frac{2}{21}$= 0,095 άρα 9,5%.
2. Με τη μείωση της τιμής κατά 10% έχουμε την ίδια μεταβολή κατά 9,5%.
3. Με τη μείωση της τιμής κατά 20% έχουμε 20 μη δημοφιλή προϊόντα άρα 1 προϊόν έγινε δημοφιλές άρα μια μεταβολή κατά $\frac{1}{21}$= 0,048 άρα 4,8%.
4. Με την αλλαγή των εξόδων αποστολής δεν έχουμε καμία μεταβολή στη δημοφιλία των προϊόντων.
5. Η αλλαγή της διαθεσιμότητας δεν μπορούσε να εξεταστεί καθώς όλα τα προϊόντα στη βάση ήταν διαθέσιμα.
6. Για τον ίδιο λόγο ούτε η αλλαγή της κατάστασης μπορούσε να μετρηθεί.

5.1.2. Case Study II - Εταιρεία B

Έπειτα δοκιμάζουμε την εταιρεία B, η οποία έχει έναν μεγαλύτερο όγκο δεδομένων. Αρχικά η βάση αυτή περιέχει 72 προϊόντα που είναι μη δημοφιλή και 29 που είναι δημοφιλή. Για την μελέτη της περίπτωσης αυτή παίρνουμε μόνο τα 72 προϊόντα που είναι μη δημοφιλή σε μια νέα βάση. Θα εξετάσουμε πως αλλάζει η κλάση τους, άρα η κατάσταση της δημοφιλίας τους, μετά τις μεταβολές που θα κάνουμε.

Πιο συγκεκριμένα τα αποτελέσματα που παίρνουμε είναι:

1. Με τη μείωση της τιμής κατά 5% έχουμε 61 μη δημοφιλή προϊόντα άρα 11 προϊόντα έγιναν δημοφιλή άρα μια μεταβολή κατά $\frac{11}{72}$= 0,153 άρα 15,3%.
2. Με τη μείωση της τιμής κατά 10% έχουμε 62 μη δημοφιλή προϊόντα άρα 10 προϊόντα έγιναν δημοφιλή άρα μια μεταβολή κατά $\frac{10}{72}$= 0,139 άρα 13,9%.
3. Με τη μείωση της τιμής κατά 20% έχουμε 64 μη δημοφιλή προΪόντα άρα 8 προϊόντα έγιναν δημοφιλή άρα μια μεταβολή κατά $\frac{8}{72}$= 0,111 άρα 11,1%
4. Η αλλαγή των εξόδων αποστολής δεν μπορούσε να εξεταστεί καθώς όλα τα προϊόντα στη βάση είχαν δωρεάν αποστολή.

7. Με την αλλαγή της διαθεσιμότητας δεν έχουμε καμία μεταβολή στη δημοφιλία των προϊόντων.
8. Με την αλλαγή της κατάστασης δεν έχουμε καμία μεταβολή στη δημοφιλία των προϊόντων.

5.1.3. Case Study III - Εταιρεία Α

Έπειτα δοκιμάζουμε την εταιρεία Α, η οποία είναι εκείνη με το μεγαλύτερο όγκο δεδομένων. Αρχικά η βάση αυτή περιέχει 224 προϊόντα που είναι μη δημοφιλή και 76 που είναι δημοφιλή. Για την μελέτη της περίπτωσης αυτή παίρνουμε μόνο τα 224 προϊόντα που είναι μη δημοφιλή σε μια νέα βάση. Θα εξετάσουμε πως αλλάζει η κλάση τους, άρα η κατάσταση της δημοφιλίας τους, μετά τις μεταβολές που θα κάνουμε.

1. Με τη μείωση της τιμής κατά 5% έχουμε 192 μη δημοφιλή προϊόντα άρα 32 προϊόντα έγιναν δημοφιλή άρα μια μεταβολή κατά $\frac{32}{224}$= 0,143 άρα 14,3%.
2. Με τη μείωση της τιμής κατά 10% έχουμε 196 μη δημοφιλή προΪόντα άρα 28 προϊόντα έγιναν δημοφιλή άρα μια μεταβολή κατά $\frac{28}{224}$= 0,125 άρα 12,5%.
3. Με τη μείωση της τιμής κατά 20% έχουμε 198 μη δημοφιλή προΪόντα άρα 26 προϊόντα έγιναν δημοφιλή άρα μια μεταβολή κατά $\frac{8}{224}$= 0,116 άρα 11,6%
4. Η αλλαγή των εξόδων αποστολής δεν μπορούσε να εξεταστεί καθώς όλα τα προϊόντα στη βάση είχαν δωρεάν αποστολή.
5. Με την αλλαγή της διαθεσιμότητας δεν έχουμε καμία μεταβολή στη δημοφιλία των προϊόντων.
6. Με την αλλαγή της κατάστασης δεν έχουμε καμία μεταβολή στη δημοφιλία των προϊόντων.

5.2 Αποτίμηση Αποτελεσμάτων

Για την ευκολότερη αποτίμηση των αποτελεσμάτων χρησιμοποιείται ο παρακάτω πίνακας όπου παρουσιάζονται συνολικά όλες οι αλλαγές που έγιναν και όλες οι μεταβολές που είχαμε στη δημοφιλία των προϊόντων για κάθε μια από τις τρεις εταιρείες που εξετάστηκαν ξεχωριστά. Φυσικά σε κάθε ποσοστό που περιέχει ο πίνακας αναφερόμαστε σε αύξηση των δημοφιλών προϊόντων.

	Εταιρεία Α	Εταιρεία Β	Εταιρεία C
Μείωση Τιμής 5%	14,3%	15,3%	9,5%
Μείωση Τιμής 10%	12,5%	13,9%	9,5%
Μείωση Τιμής 20%	11,6%	11,1%	4,8%
Αλλαγή Εξόδων Αποστολής	-	-	0%
Αλλαγή Διαθεσιμότητας	0%	0%	-
Αλλαγή Κατάστασης	0%	0%	-

Πίνακας 5.1. Συνολικά αποτελέσματα των 3 Case Studies.

5.2.1. Μεταβολές στην τιμή

Αρχικά, λοιπόν, οι παρατηρήσεις μας θα επικεντρωθούν στη μεταβολή της τιμής και πως η μείωση της επηρεάζει την δημοφιλία των προϊόντων. Παρατηρείται ότι η μικρότερη μείωση που επιφέρουμε δηλαδή η μείωση κατά 5% αυξάνει τη δημοφιλία των προϊόντων όπως ήταν αναμενόμενο. Παρόλα αυτά η περαιτέρω μείωση των τιμών ενώ αυξάνει τη δημοφιλία των προϊόντων προκαλεί μικρότερη αύξηση όσο μεγαλώνει η μείωση της τιμής.

Για να αναλύσουμε αυτήν την παρατήρηση, αξίζει να σημειωθεί ότι ο καθορισμός μιας συγκεκριμένης τιμής για το προϊόν δεν έχει να κάνει μόνο με το να γίνει προσιτό για τους πελάτες. Παίζει επίσης έναν κρίσιμο ψυχολογικό ρόλο στην αξία του προϊόντος όπως εκτιμάται από τους καταναλωτές. Ο τρόπος που τιμολογούνται τα προϊόντα είναι μια άμεση επικοινωνία με τους πελάτες, καθώς συνδέονται με το πόσο αξίζουν τα προϊόντα και η μάρκα τους αλλά και η μάρκα της εταιρείας - εμπόρου που τα πουλάει. Αν αποφασιστεί η τιμολόγηση του προϊόντος χαμηλότερα από τον ανταγωνισμό, τότε στέλνεται ένα διαφορετικό είδος μηνύματος ανάλογα με τον παραλήπτη.

Πιο συγκεκριμένα, σε έναν αγοραστή αξίας (value shopper), δηλαδή κάποιον που αναζητά την καλύτερη δυνατή αξία για το κόστος του προϊόντος ή της υπηρεσίας η χαμηλή τιμή σημαίνει ότι το προϊόν είναι σε τιμή ευκαιρίας σε σύγκριση με τον ανταγωνισμό. Αυτό εξηγεί και την αύξηση της δημοφιλίας που βλέπουμε. Για έναν υψηλού επιπέδου αγοραστή (high-end shopper) που αναζητά ένα προϊόν ή μια υπηρεσία που θα του δώσει την αίσθηση ότι ανήκει σε μια ομάδα αποκλειστικότητας, η χαμηλή τιμή δηλώνει ότι το προϊόν είναι κατώτερης ποιότητας σε σχέση με τα υπόλοιπα και χάνει την αίσθηση της αποκλειστικότητας, καθώς μια μεγαλύτερη μερίδα αγοραστών δύνανται να αγοράσουν το

συγκεκριμένο προϊόν. Αντίθετα, μια υψηλότερη τιμή για ένα προϊόν μπορεί να κάνει έναν τέτοιο καταναλωτή πιο επιρρεπή στο να το αγοράσει άρα να αυξήσει τη δημοφιλία του απλά επειδή το μήνυμα που παίρνει είναι ότι του προσφέρεται ένα αποκλειστικό αγαθό πολυτελείας. Αυτό, συνδέεται με τη μείωση της αύξησης που βλέπουμε αν λάβουμε ως δεδομένο ότι περισσότεροι καταναλωτές ανήκουν στην ομάδα των value shoppers.

Ακόμη, ο καθορισμός της τιμής έχει άμεση σχέση με τον ανταγωνισμό και τον τρόπο χειρισμού του αλλά και με τη θέση της εταιρείας σε σχέση με το μερίδιο αγοράς που κατέχει. Έχει μεγάλο αντίκτυπο στο πόσο καλά ανταγωνίζεται η εταιρεία τις άλλες επιχειρήσεις του κλάδου αλλά παράλληλα μπορεί να επηρεάσει και τον τρόπο με τον οποίο άλλες επιχειρήσεις του κλάδου είναι σε θέση να την ανταγωνιστούν. Ειδικότερα, για μια κυρίαρχη εταιρεία στην αγορά που ελέγχει το μερίδιο της, οι όγκοι των πωλήσεών είναι τόσο μεγάλοι και τα κόστη τόσο χαμηλά που πιθανότατα μπορεί να επιβιώσει με χαμηλότερα περιθώρια κέρδους και χαμηλότερες τιμές, άρα μπορεί να μειώσει τις τιμές και να παραμείνει κυρίαρχη αναγκάζοντας σε μείωση τιμών και τις άλλες μικρότερες εταιρείες. Για μια εταιρεία η οποία δεν κατέχει μεγάλο μερίδιο αγοράς η μείωση τιμών δεν της προσφέρει απαραίτητα δημοφιλία στα προϊόντα καθώς όπως περιγράψαμε υποδηλώνει μείωση ποιότητας. Αυτές οι εταιρείες δεν είναι σε θέση να αυξήσουν το μερίδιο αγοράς ακόμη και με 20% μείωση των τιμών γιατί οι άλλες κυρίαρχες εταιρείες έχουν ήδη διαμορφώσει σχέσεις με τους καταναλωτές που την προτιμούν, έχουν κάνει γνωστό το όνομα τους στην αγορά και δεν είναι εύκολο να χάσουν πελάτες που θα προτιμήσουν τις άλλες.

Στη συγκεκριμένη βάση δεδομένων, λοιπόν, επειδή αναφερόμαστε σε ηλεκτρονικά προϊόντα μεταξύ των οποίων είναι και εκείνα τα οποία απευθύνονται σε high-end shoppers όπως τηλεοράσεις, κινητά τηλέφωνα και υπολογιστές πολυτελών μαρκών, η μείωση της τιμής συνδέεται με το μήνυμα της υποβάθμισης της ποιότητας για αυτό έχουμε μικρότερη αύξηση της δημοφιλίας. Από τη μία κερδίζουμε το μερίδιο των value shoppers, οι οποίοι προτιμούν τα προϊόντα σε εκπτώσεις και χαμηλές τιμές αλλά δεν προσφέρουμε την αποκλειστικότητα που έχουν ανάγκη οι high - end shoppers. Ακόμη, επειδή μέσα στη βάση περιλαμβάνονται και μάρκες που κυριαρχούν στην αγορά αλλά και άλλες μικρότερες που συνήθως έχουν τα μη δημοφιλή προϊόντα είναι δύσκολο η μείωση τιμής από τις τελευταίες να καθιστά τόσο εύκολα τα προϊόντα δημοφιλή.

Ακόμη μια σημαντική παρατήρηση είναι το γεγονός ότι εμφανίζεται μεγαλύτερη αύξηση της δημοφιλίας για την εταιρεία Β που έχει περισσότερα προϊόντα από την C στη βάση και λιγότερα από την Α. Κάτι τέτοιο μας οδηγεί στην παρατήρηση ότι ίσως η εταιρεία Β είναι μια εταιρεία η οποία έχει ήδη ένα ικανοποιητικό μερίδιο αγοράς αλλά ίσως να μην είναι η κυρίαρχη και έτσι η μείωση των τιμών φέρνει μεγαλύτερη αύξηση της δημοφιλίας και την καθιστά πιο ανταγωνιστική. Αντιθέτως, η εταιρεία Α είναι ήδη κυρίαρχη άρα δεν παρουσιάζεται τόσο σημαντικά μεγαλύτερη αύξηση της δημοφιλίας της καθώς έχει ήδη τους καταναλωτές που την προτιμούν. Τέλος για την εταιρεία C έχουμε τη μικρότερη αύξηση καθώς ούτως η άλλως φαίνεται να είναι ένας μικρότερος πάροχος στην αγορά άρα

δεν μπορεί με τη μείωση τιμών να έχουμε τόσο σημαντικές μεταβολές στη δημοφιλία της και των προϊόντων της. Ακόμη η εταιρεία C περιλαμβάνει τα περισσότερα μη δημοφιλή προϊόντα από τις άλλες δύο, άρα ήδη είναι μια εταιρεία με χαμηλά ποσοστά δημοφιλίας στα προϊόντα της, τα οποία ποσοστά διατηρεί και μετά τις μεταβολές καθώς όπως είπαμε δεν είναι αρκετές για να προκαλέσουν μεγάλες αλλαγές.

Παρόλα αυτά. επειδή σε κάθε περίπτωση η μείωση της τιμής επιφέρει αποτελέσματα είναι σημαντικό να αναλυθεί ο τρόπος με τον οποίο κάτι τέτοιο μπορεί να επιτευχθεί. Πιο συγκεκριμένα, η μείωση της τιμής όταν πρόκειται για μια στρατηγική μείωσης των τιμών αρκετών προϊόντων μιας ηλεκτρονικής πλατφόρμας, μπορεί να γίνει προσωρινά δηλαδή με κάποια έκπτωση σε μια συγκεκριμένη περίοδο εκπτώσεων ή σε κάποια εμπορική καμπάνια που διοργανώνει η εταιρεία. Αυτή είναι συνήθως μια αποτελεσματική βραχυπρόθεσμη τακτική για να βοηθήσει στην αύξηση των πωλήσεων και συνήθως χρησιμοποιείται παράλληλα με τη μακροπρόθεσμη τιμολόγηση και τις συνολικές στρατηγικές μάρκετινγκ. Εναλλακτικά μπορεί να γίνει μόνιμα, κάτι το οποίο συμβαίνει περισσότερο στα προϊόντα τα οποία έχουν εξελιχθεί σε νέα μοντέλα και δεν έχουν πια την πιο σύγχρονη τεχνολογία. Υποθέτοντας ότι τα κόστη παραμένουν τα ίδια, η μείωση των τιμών βραχυπρόθεσμα για την αύξηση των πωλήσεων μειώνει επίσης το περιθώριο κέρδους για κάθε μονάδα που πωλείται. Από την άλλη, οι χαμηλότερες τιμές για μια μεγαλύτερη περίοδο θα οδηγήσουν σε υψηλότερους όγκους πωλήσεων, οι οποίοι μπορεί να αντισταθμίσουν το χαμηλότερο περιθώριο κέρδους.

Ακόμη, η μείωση των τιμών γίνεται στα πλαίσια της στρατηγική απώλειας ηγεσίας (loss leader strategy) που περιλαμβάνει την πώληση ενός προϊόντος ή μιας υπηρεσίας σε τιμή που δεν είναι κερδοφόρα, αλλά πωλείται για να προσελκύσει νέους πελάτες ή για να πουλήσει πρόσθετα προϊόντα και υπηρεσίες σε αυτούς τους πελάτες. Αποτελεί μια κοινή πρακτική όταν μια επιχείρηση εισέρχεται για πρώτη φορά σε μια αγορά.

Τέλος, πολλές εταιρείας και ιδιαίτερα οι ηλεκτρονικοί πάροχοι επιλέγουν τη μείωση τιμών για να αποφύγουν τον εγκλωβισμό σε στατικά αποθέματα. Το απόθεμα αντιπροσωπεύει δεσμευμένα χρήματα, και δεν βοηθά στην ταμειακή ροή της επιχείρησης. Σε τέτοιες περιπτώσεις, συνήθως είναι καλή ιδέα η μείωση της τιμής προκειμένου να ξεκινήσει η πώληση των αποθεμάτων. Μερικές φορές είναι ακόμη και λογικό να το πωληθεί σε τιμή που δεν παράγει κέρδος, αφού η πώληση του μειώνει τα χρήματα που ξοδεύονται στην αποθήκευση του. Μπορεί επίσης να χρησιμοποιηθεί ακόμη και για σκοπούς μάρκετινγκ, και τα μελλοντικά κέρδη που θα προκύψουν να αντισταθμίσουν τις ζημίες που θα προκληθούν κατά την πώληση του αποθέματος.

5.2.2. Μεταβολές στα έξοδα αποστολής

Η επόμενη παρατήρηση αφορά τα έξοδα αποστολής τα οποία μπορούμε να αξιολογήσουμε μόνο με την περίπτωση της εταιρείας C, καθώς για τις άλλες δεν μπορούσαμε να

πραγματοποιήσουμε τις μεταβολές. Πιο συγκεκριμένα, βλέπουμε ότι η μεταβολή των εξόδων αποστολής σε δωρεάν αποστολή δεν επιφέρει κάποια αλλαγή στη δημοφιλία των προϊόντων.

Για να ερμηνεύσουμε αυτήν την παρατήρηση είναι σημαντικό αρχικά να διατυπωθεί ότι γνωρίζουμε ότι η επίδραση των εξόδων αποστολής στις αγοραστικές συνήθειες είναι αρκετά μεγάλη. Κατά τη διάρκεια μιας πρόσφατης μελέτης που διεξήχθη σε 2.500 καταναλωτές στις Η.Π.Α., 64,3% των ερωτηθέντων είπε ότι η τιμή είναι ο πιο σημαντικός παράγοντας που σχετίζονται με την αποστολή προϊόντων. Η δωρεάν αποστολή φαίνεται να προσφέρεται από την πλειοψηφία των λιανοπωλητών ηλεκτρονικού εμπορίου ειδικά όσο η ανάπτυξη του τελευταίου είναι ραγδαία και όλο και περισσότεροι καταναλωτές το προτιμούν έναντι των κλασσικών μεθόδων αγοράς. Η στρατηγική αυτή είναι ένας πολύ καλός τρόπος για να αυξηθούν οι πωλήσεις, καθώς είναι δελεαστικό για τους καταναλωτές και καθιστά επίσης τη διαδικασία πληρωμής λιγότερο περίπλοκη. Για να προσφέρουν δωρεάν αποστολή οι λιανοπωλητές έχουν δύο επιλογές. Είτε κάνουν ενσωμάτωση του κόστους αποστολής στην τιμή καταλόγου προϊόντων, ενδεχομένως καθιστώντας το προϊόν πιο ακριβό από τους ανταγωνιστές είτε απορροφούν το κόστος αποστολής και μειώνουν τα περιθώρια κέρδους τους.

Παρόλα αυτά, από τη συγκεκριμένη περίπτωση λαμβάνουμε το συμπέρασμα ότι τα δωρεάν έξοδα αποστολής ακόμη και αν είναι μεγάλης σημασίας παράγοντας δεν κάνουν ένα προϊόν που είναι μη δημοφιλές να γίνει δημοφιλές. Αυτό σημαίνει ότι μπορεί να επηρεάσει θετικά στο να επιλεγεί ο ηλεκτρονικός πάροχος από τον οποίο θα αγοραστεί ένα προϊόν που εμφανίζεται σε πολλαπλούς παρόχους αλλά όχι το συγκεκριμένο προϊόν έναντι κάποιου άλλου μόνο λόγω του γεγονότος ότι έχει δωρεάν έξοδα αποστολής. Κάτι τέτοιο είναι ακόμη πιο λογικό εφόσον μιλάμε για ηλεκτρονικά προϊόντα στα οποία η μάρκα παίζει πολύ σημαντικό ρόλο άρα ένας καταναλωτής σπάνια θα επιλέξει με βασικότερο κριτήριο τα έξοδα αποστολής έναντι της μάρκας που γνωρίζει και είναι ήδη δημοφιλής στην αγορά.

Ένας άλλος παράγοντας που είναι επίσης πολύ σημαντικός είναι ότι τα ηλεκτρονικά προϊόντα εμφανίζονται σε πολλούς διαφορετικούς ιστότοπους κάποιος από τους οποίους θα προσφέρει δωρεάν αποστολή. Εφόσον μάλιστα οι περισσότεροι πάροχοι ακολουθούν τη στρατηγική του να προσφέρουν δωρεάν αποστολή όταν η τιμή περάσει ένα συγκεκριμένο κατώφλι, είναι συχνό η αγορά ηλεκτρονικών ειδών που έχουν αυξημένες τιμές να γίνεται με δωρεάν έξοδα αποστολής. Τέλος, κάτι το οποίο είναι καθοριστικό είναι το γεγονός ότι πλέον με τη μεγάλη ανάπτυξη του ηλεκτρονικού εμπορίου εμφανίζονται πολλές εναλλακτικές για τους αγοραστές που δεν επιθυμούν να πληρώσουν έξοδα αποστολής. Πιο συγκεκριμένα, υπάρχει η επιλογή παραγγελίας ενός προϊόντος και παραλαβής του τόσο από φυσικά καταστήματα, όσο και από άλλα φυσικά σημεία παραλαβής κάτι το οποίο εξασφαλίζει δωρεάν κόστος αποστολής αφού το προϊόν συμπεριλαμβάνεται σε αυτά που ούτως η άλλως θα είχε αποστείλει η εταιρεία προς τα φυσικά της σημεία.

Αυτό που θα μπορούσε να αποτελέσει σημαντικό συμπέρασμα όμως είναι ότι είναι σημαντική η διαφάνεια σχετικά με τα έξοδα αποστολής. Όταν ένας ενδεχόμενος αγοραστής γνωρίζει τα έξοδα αποστολής, είναι πιο εύκολο να αισθανθεί ότι παίρνει τις αποφάσεις για τα έξοδα του και έχει μια εξατομικευμένη εμπειρία. Η διαδικασία αυτή θα δημιουργήσει εμπιστοσύνη και αφοσίωση των πελατών, και τελικά θα αυξήσει τις πωλήσεις. Η διαφάνεια σχετικά με τα έξοδα αποστολής είναι επίσης σημαντική για την οικοδόμηση της πίστης στη μάρκα των προϊόντων. Η προσφορά πολλαπλών επιλογών αποστολής σε διαφορετικές τιμές είναι ένας άλλος τρόπος που ενδεχομένων παίζει ρόλο στη δημοφιλία καθώς παράλληλα μειώνει τα ποσοστά εγκατάλειψης του καροτσιού, δηλαδή του φαινομένου όταν ένας αγοραστής αποφασίζει να μην αγοράσει το προϊόν λόγω των εξόδων αποστολής.

Όλα τα παραπάνω αποδεικνύουν ότι ο παράγοντας των εξόδων αποστολής είναι αδιαμφισβήτητης σημασίας αλλά στην περίπτωση των ηλεκτρονικών ειδών που εξετάζουμε όχι καθοριστικής ώστε να κάνει τα προϊόντα πιο δημοφιλή και προτιμητέα έναντι των υπολοίπων. Σε κάθε περίπτωση είναι ένα χαρακτηριστικό που είναι σημαντικό να εξεταστεί από έναν πάροχο αλλά δεν γίνεται σε επίπεδο προϊόντος που εξετάζουμε στην παρούσα εργασία αλλά σε επίπεδο ηλεκτρονικού εμπόρου έναντι άλλων κάτι που δεν εντάσσεται στην παρούσα έρευνα.

5.2.3. Μεταβολές στην διαθεσιμότητα

Στη συνέχεια, πραγματοποιήθηκαν μεταβολές στη διαθεσιμότητα οι οποίες δεν επέφεραν κάποια αλλαγή στη δημοφιλία των προϊόντων παρόλο που εξετάστηκαν τόσο για την εταιρεία Α όσο και για την Β.

Για να εξηγηθεί η παραπάνω παρατήρηση είναι σημαντικό να αναφερθούμε σε μια έρευνα που πραγματοποιήθηκε σχετικά με την επίδραση που έχουν τα μηνύματα των ηλεκτρονικών παρόχων σχετικά με την περιορισμένη ή όχι διαθεσιμότητα των προϊόντων στους καταναλωτές. Σύμφωνα με την έρευνα αυτή, οι καταναλωτές που έλαβαν μηνύματα σχετικά με τη διαθεσιμότητα των προϊόντων αλλά και εκείνοι που έλαβαν μηνύματα σχετικά με τη περιορισμένη διαθεσιμότητα δεν επηρεάστηκαν για την αγορά τους με κάποιον από τους δύο τρόπους. Μια πιθανή αιτία για αυτό είναι ότι οι ενημερώσεις αυτές δεν είναι αξιόπιστες στο εμπορικό περιβάλλον του διαδικτύου σύμφωνα με την κρίση των καταναλωτών. Η αξιοπιστία είναι απαραίτητη προϋπόθεση για την αποτελεσματικότητα οποιασδήποτε αξίωσης πειθούς ιδιαίτερα σε ηλεκτρονικά πλαίσια, όπου ο ανθρώπινος παράγοντας απουσιάζει εντελώς. Επειδή η αξιολόγηση της διαθεσιμότητας των προϊόντων είναι πιο περιορισμένη στην ηλεκτρονική αγορά, οι αγοραστές μπορεί να θεωρήσουν την περιορισμένη διαθεσιμότητα ως τη χειραγώγηση που ασκεί το τμήμα μάρκετινγκ για να παρακινεί τις πωλήσεις. Έτσι, υποθέτουμε ότι ο ισχυρισμός περί περιορισμένης διαθεσιμότητας δεν γίνεται αντιληπτός ως αξιόπιστος και δεν έχει επίδραση στη δημοφιλία του προϊόντος. Αντίστοιχα και τα μηνύματα σχετικά με τη διαθεσιμότητα των προϊόντων αποδείχθηκαν να μην έχουν μεγάλη ψυχολογική επίδραση στους καταναλωτές, οι οποίοι

επιθυμούσαν να έχουν την ελευθερία επιλογής χωρίς να επηρεάζονται από τέτοιους παράγοντες.

Παρόλα αυτά, το να είναι τα προϊόντα συνεχώς σε ελλείψεις μπορεί να έχει αρνητικές συνέπειες που δύνανται να κάνουν δημοφιλή προϊόντα να γίνουν μη δημοφιλή. Αρχικά, υπάρχουν πολλοί λόγοι για τους οποίους εμφανίζονται ελλείψεις στα αποθέματα των προϊόντων. Μερικοί από αυτούς είναι οι ανεπαρκείς προβλέψεις, η υποτίμηση της ζήτησης για ένα προϊόν, τα χαμηλά επίπεδα αποθεμάτων ασφαλείας και οι χαμηλές παραγγελίες. Όποιοι όμως και αν είναι οι λόγοι κάτι τέτοιο έχει αρνητικές συνέπειες για την εταιρεία. Η "Holiday Outlook Report 2015: To What Matters Most During Peak Shopping Season" δείχνει ότι από τους 500 διαδικτυακούς αγοραστές που συμμετείχαν στην έρευνα, το 79 τοις εκατό ανέφερε ότι αντί να περιμένουν ένα προϊόν να γίνει διαθέσιμο, είναι πιθανό να αλλάξουν μάρκες όταν τα είδη που αγοράζουν είναι εκτός αποθέματος.

Το φαινόμενο αυτό, λοιπόν, μπορεί να οδηγήσει σε αρνητικά αισθήματα στους πελάτες και εν τέλει να χαθούν πελάτες, σχέσεις μαζί τους και μεμονωμένες πωλήσεις. Η χαμηλή διαθεσιμότητα προϊόντων οδηγεί και αυτή σε απογοητευμένους και θυμωμένους πελάτες και δημιουργεί μια πολύ αρνητική εμπειρία αγορών. Αυτό μπορεί να έχει ως αποτέλεσμα οι πελάτες να μην ξαναπροτιμήσουν τον συγκεκριμένο ηλεκτρονικό πάροχο και το προϊόν αυτό αλλά και τη μάρκα του φοβούμενοι ότι θα έχουν άλλη μια παρόμοια αρνητική εμπειρία αγορών που θα αφήσει τις αγοραστικές τους ανάγκες ανικανοποίητες. Κάτι τέτοιο ακόμη, θα αναγκάσει τους πελάτες είτε να αγοράσουν ένα υποκατάστατο προϊόν είτε ένα προϊόν από άλλη μάρκα, επηρεάζοντας την δημοφιλία των προϊόντων. Το τελευταίο αναπόφευκτα θα οδηγήσει σε μια αρνητική αντίληψη η οποία θα μαθευτεί και θα επηρεάσει γενικά τη δημοφιλία του προϊόντος.

Τέλος, όλα τα παραπάνω μπορεί να επηρεάσουν και τη σχέση μεταξύ του πωλητή και του προμηθευτή, ειδικά όταν αναφερόμαστε για εταιρείες ηλεκτρονικού εμπορίου και μεγάλες μάρκες ηλεκτρονικών προϊόντων. Καθώς η κατάσταση αυτή παρακινεί τους πελάτες να δοκιμάσουν μια διαφορετική μάρκα, αυτό μπορεί να προκαλέσει απώλεια πωλήσεων για τον πελάτη που αποτελεί έναν μεγάλο κατασκευαστή ηλεκτρονικών ειδών και να τον παρακινήσει να στραφεί σε ανταγωνιστές ηλεκτρονικούς παρόχους εφόσον επιθυμεί να μην χάσει η μάρκα του την αξιοπιστία και δημοφιλία της.

Στην παρούσα περίπτωση, λοιπόν, αντιλαμβανόμαστε ότι δεν λάβαμε κάποια μεταβολή στα μη δημοφιλή προϊόντα. Παρόλα αυτά η χαμηλή και καθόλου διαθεσιμότητα επηρεάζει σίγουρα τα δημοφιλή προϊόντα, τα οποία μπορεί σταδιακά και με την πάροδο του χρόνου να χάσουν τη δημοφιλία τους λόγω ύπαρξης δυσαρεστημένων καταναλωτών κάτι που θα επηρεάσει και άλλους ενδεχόμενους μελλοντικούς όπως περιγράφηκε παραπάνω.

5.2.4. Μεταβολές στην κατάσταση

Η τελευταία περίπτωση που μελετήθηκε αφορούσε τη μεταβολή της κατάστασης των προϊόντων από χρησιμοποιημένα, ανακατασκευασμένα και τα λοιπά σε καινούρια. Προφανώς αυτό το σενάριο αποτελεί μια θεωρητική αλλαγή καθώς δεν είναι δυνατόν να μετατρέψουμε πρακτικά τα μη καινούρια προϊόντα σε καινούρια. Παρόλα αυτά αυτή η θεωρητική δοκιμή έγινε για να μελετηθεί η επιρροή που έχει σαν χαρακτηριστικό η κατάσταση που βρίσκεται ένα προϊόν σε σχέση με τη δημοφιλία του. Πιο συγκεκριμένα, ομοίως με προηγουμένως βλέπουμε ότι η μεταβολή αυτή δεν επιφέρει κάποια αλλαγή στη δημοφιλία των προϊόντων.

Σε γενικότερο πλαίσιο, η δημοτικότητα της αγοράς μεταχειρισμένων προϊόντων αυξάνεται συνεχώς σε όλα τα κοινωνικά επίπεδα. Η βιομηχανία ανακατασκευασμένων (refurbished) ηλεκτρονικών ειδών έχει αναπτυχθεί αρκετά και γίνεται ολοένα και πιο δημοφιλής. Στο παρελθόν, τα ανακατασκευασμένα ηλεκτρονικά προϊόντα, τα οποία αναφέρονται σε επαγγελματικά επισκευασμένα μεταχειρισμένα προϊόντα είχαν αμφιλεγόμενη φήμη καθώς πολλοί καταναλωτές αμφέβαλαν σχετικά με την ποιότητα τους αλλά και την αξιοπιστία τους. Οι πελάτες συχνά ανησυχούσαν ότι τα προϊόντα θα δυσλειτουργούσαν ή, σε ορισμένες περιπτώσεις, θα ήταν πλαστά. Αλλά όπως και για τα άλλα είδη, έτσι και για τα ηλεκτρονικά, οι πωλήσεις τους έχουν αυξηθεί σημαντικά τα τελευταία χρόνια. Για παράδειγμα, μια μάρκα χρησιμοποιημένων προϊόντων γνωστής εταιρείας έγινε η τρίτη πιο δημοφιλής μάρκα ηλεκτρονικών ειδών για τον Ιούνιο του 2020. Πλεον, αυτές οι μάρκες και οι αγορές που πωλούν μεγάλες ποσότητες ανακατασκευασμένων ηλεκτρονικών ειδών προσπαθούν να διατηρήσουν τους πελάτες τους, διευρύνοντας την επιλογή των αποθεμάτων τους και υποσχόμενες πιο αυστηρό έλεγχο ποιότητας. Παρόμοια τάση υπάρχει και για τα χρησιμοποιημένα ηλεκτρονικά είδη τα οποία πωλούνται σήμερα ακόμη και μέσω των σελίδων κοινωνικής δικτύωσης εύκολα και άμεσα, είτε μέσω ειδικών πλατφορμών και εφαρμογών που βοηθούν το χρήση αλλά και μέσω των μεγάλων παρόχων ηλεκτρονικού εμπορίου που μελετάμε και στην παρούσα εργασία.

Για το λόγο αυτό η παρατήρηση που προκύπτει σχετικά με το πόσο η κατάσταση του προϊόντος επηρεάζει τη δημοφιλία είναι λογική αν κρίνουμε από αυτές τις νέες τάσεις στην αγορά. Το να είναι ένα προϊόν καινούριο δεν ισοδυναμεί απαραίτητα με το να αυξηθεί η δημοφιλία του όπως βλέπουμε και από το πείραμα που κάναμε. Είναι σίγουρα ένα χαρακτηριστικό που σχετίζεται με τη δημοφιλία αλλά όχι καθοριστικό ώστε να την μεταβάλλει.

Κεφάλαιο 6. Συμπεράσματα και Προεκτάσεις

6.1 Συμπεράσματα

Είναι σαφές ότι η αρχική μέθοδος που χρησιμοποιήσαμε, δηλαδή οι αλγόριθμοι επιβλεπόμενης μηχανικής μάθησης που βασίζονταν σε μεθόδους παλινδρόμησης δεν είχε ικανοποιητικά αποτελέσματα προβλέψεων. Πιο συγκεκριμένα, η πρόβλεψη των τιμών των προϊόντων με τις διαφορετικές αυτές μεθόδους βρέθηκε να έχει αρκετά μεγάλο ποσοστό σφάλματος αλλά παράλληλα να χάνει και σε δυναμικότητα. Κάτι τέτοιο θα είχε συνέπεια τα προϊόντα να μην προτιμηθούν από τους καταναλωτές αλλά παράλληλα και η μεθοδολογία μας να μην είναι εύκολο να μεταβάλλεται σύμφωνα με τις τάσεις της αγοράς και να μην προσφέρει κανένα βαθμό ευελιξίας. Παράλληλα, δεν προσέφερε τη δυνατότητα να εξάγουμε κάποιο σενάριο σχετικά με μια γενικότερη στρατηγική τιμολόγησης αλλά ούτε να καθορίσουμε τα επιμέρους χαρακτηριστικά πώλησης των προϊόντων.

Η εναλλακτική μέθοδος που τέθηκε έπειτα σε εφαρμογή ήταν η ταξινόμηση των προϊόντων σε δημοφιλή και μη και οι δοκιμές μεταβολών των χαρακτηριστικών των προϊόντων με στόχο την μετατροπή των μη δημοφιλών προϊόντων σε δημοφιλή. Οι δοκιμές στηρίχτηκαν σε κάθε εταιρεία από τις 3 κυρίαρχες στην ηλεκτρονική αγορά και σε 6 διαφορετικά σενάρια, τρία από τα οποία συνδέθηκαν με μεταβολές στις τιμές των προϊόντων και 3 από τα οποία βασίστηκαν στα επιμέρους χαρακτηριστικά της πώλησης δηλαδή τα έξοδα αποστολής, την κατάσταση και την διαθεσιμότητα των προϊόντων. Σημαντικό στοιχείο για την εξαγωγή των αποτελεσμάτων αποτέλεσε το γεγονός ότι η βάση αποτελείται από δεδομένα ηλεκτρονικών προϊόντων ένα συγκεκριμένο μήνα και για εταιρείες ηλεκτρονικού εμπορίου.

Όσον αφορά τα αποτελέσματα, αυτά καθ΄ αυτά, μπορούμε να συνοψίσουμε τις παρατηρήσεις μας ως εξής:

- Η μείωση τιμής των προϊόντων προκαλεί αύξηση στη δημοφιλία των προϊόντων με τις εξής παραμέτρους:
    - Όσο η αύξηση της τιμής μεγαλώνει τόσο μικρότερη είναι η αύξηση στη δημοφιλία.
    - Μεγαλύτερη αύξηση της δημοφιλίας παρουσιάζεται για την εταιρεία Β που έχει περισσότερα προϊόντα από την C στη βάση και λιγότερα από την Α άρα είναι αρκετά σημαντικός πάροχος αλλά όχι ο κυρίαρχος.
- Ο καθορισμός δωρεάν αποστολής των προϊόντων που εξετάστηκε μόνο για μια εταιρεία δεν επιφέρει κάποια μεταβολή στη δημοφιλία.
- Ο ορισμός των προϊόντων ως διαθέσιμα που εξετάστηκε για δύο από τις τρεις εταιρείες δεν επιφέρει κάποια μεταβολή στη δημοφιλία.
- Ομοίως, ο ορισμός των προϊόντων ως καινούρια που εξετάστηκε για δύο από τις τρεις εταιρείες δεν επιφέρει κάποια μεταβολή στη δημοφιλία.

Σύμφωνα με τις παραπάνω παρατηρήσεις στα αποτελέσματα καταλήγουμε στα εξής συμπεράσματα λαμβάνοντας υπόψη τη μελέτη αυτή συνολικά:

- Η μείωση της τιμής αποτελεί μια στρατηγική που έχει αποτέλεσμα για αγοραστές που στοχεύουν σε χαμηλού κόστους αγορές και κάνει τις εταιρείες σίγουρα πιο δημοφιλείς και τα προϊόντα περισσότερο προτιμητέα λόγω αυτών των καταναλωτών.
- Η μείωση της τιμής όμως για μια συγκεκριμένη μερίδα καταναλωτών ισοδυναμεί με υποβάθμιση στη ποιότητα των προϊόντων κάτι το οποίο ζημιώνει την δημοφιλία τόσο των προϊόντων όσο και της ίδια της μάρκας.
- Η μείωση της τιμής συνδέεται επίσης με τη θέση της εταιρείας έναντι του ανταγωνισμού κάτι το οποίο υποδηλώνει πως οι πιο κυρίαρχες εταιρείες έχουν πιο μεγάλο περιθώριο μείωσης τιμών και καλύτερα αποτελέσματα λόγω της ήδη υπάρχουσας σχέσης με τους καταναλωτές και ονόματος στην αγορά που δεν αφήνει περιθώριο για μεγάλη υποβάθμιση της αξίας τους.
- Τα δωρεάν έξοδα αποστολής, μπορεί σαν παράγοντας να επηρεάσει θετικά στο να επιλεγεί ο ηλεκτρονικός πάροχος από τον οποίο θα αγοραστεί ένα προϊόν που εμφανίζεται σε πολλαπλούς παρόχους αλλά όχι το συγκεκριμένο προϊόν έναντι κάποιου άλλου για αυτό δεν παρατηρείται αύξηση της δημοφιλίας.
- Με τη μεταβολή των προϊόντων σε διαθέσιμα δεν λάβαμε κάποια μεταβολή στα μη δημοφιλή προϊόντα καθώς οι καταναλωτές δεν επηρεάζονται τόσο από αυτό το παράγοντα, εφόσον η αγορά ηλεκτρονικών ειδών αποτελεί μια αγορά που γίνεται έπειτα από μια συγκεκριμένη ανάγκη μια συγκεκριμένη περίοδο και η οποία εναλλακτικά καλύπτεται από μια σειρά από υποκατάστατα προϊόντα και σε επίπεδο είδους και σε επίπεδο μάρκας. Παρόλα αυτά η χαμηλή και καθόλου διαθεσιμότητα δημιουργεί δυσαρεστημένους καταναλωτές επηρεάζοντας τα δημοφιλή προϊόντα σε βάθος χρόνου.
- Η θεωρητική περίπτωση που ένα προϊόν είναι καινούριο αντί για χρησιμοποιημένο ή ανακατασκευασμένο δεν ισοδυναμεί με το να αυξηθεί η δημοφιλία λόγω της τάσης της αγοράς, η οποία υποδεικνύει μια μεγάλη αύξησης της δημοφιλίας των χρησιμοποιημένων ειδών και ειδικά ηλεκτρονικών που μπορούν να παρέχονται άμεσα και εύκολα μέσω πολύ διαδεδομενων και εύκολων τρόπων αγορών.

Όλα αυτά τα συμπεράσματα αναδεικνύουν ποιοι παράγοντες είναι περισσότερο ή λιγότερο σημαντικοί για την αύξηση της δημοφιλίας των προϊόντων. Σύμφωνα με αυτούς τους παράγοντες είναι σημαντικό να καθοριστεί μια γενικότερη στρατηγική τόσο τιμολόγησης όσο και πώλησης. Αυτή η στρατηγική είναι απαραίτητο όχι απλά να ορίζει σταθερά κάποιους παράγοντες αλλά και να μπορεί να μεταβάλλεται σημαντικά σύμφωνα με νέες τάσεις τιμολόγησης (π.χ. εκπτώσεις) αλλά και άλλων χαρακτηριστικών (π.χ. εβδομάδα με δωρεάν έξοδα αποστολής που γίνεται από πολλαπλούς παρόχους). Σύμφωνα με όλα τα παραπάνω προτείνεται μια στρατηγική η οποία να περιλαμβάνει τα εξής βήματα:

- Σχεδιασμός μιας στρατηγικής μείωσης των τιμών που θα εντάσσεται σε μια γενικότερη εμπορική καμπάνια είτε εκπτώσεων, είτε άλλου είδους προσφορών και θα στοχεύει τόσο στην πώληση των συγκεκριμένων προϊόντων όσο και στην προσέλκυση νέων πελατών με τους οποίους θα χτιστεί μετέπειτα μια σχέση εμπιστοσύνης και θα υπάρχουν μακροπρόθεσμα οικονομικά οφέλη.
- Μόνιμη μείωση τιμών μόνο για τα προϊόντα τα οποία έχουν γίνει πλέον παρωχημένα τεχνολογικά, σε συνδυασμό με μια αντίστοιχη καμπάνια η οποία θα έχει ως στόχο τη διατήρηση του ονόματος της εκάστοτε μάρκας και την ανάδειξη της αξίας της ακόμη και έπειτα από τη μείωση αυτή.
- Μελέτη των τακτικών εξόδων αποστολής των ανταγωνιστών ηλεκτρονικών παροχών και όχι μελέτη των εξόδων σε επίπεδο προϊόντων με στόχο την προσφορά στους καταναλωτές της επιλογής διαφορετικών μεθόδων παραλαβής των προϊόντων που ακόμη και αν δεν είναι δωρεάν αποτελούν ανταγωνιστικές στρατηγικές έναντι των υπολοίπων παρόχων.
- Προσπάθεια για ύπαρξη διαθέσιμων προϊόντων σε αντιστοιχία με την εκάστοτε ζήτηση ώστε να αποφευχθούν οι χαμένες πωλήσεις που θα προκύψουν και οι μακροπρόθεσμες χαμένες πωλήσεις από σχέσεις με δυσαρεστημένους καταναλωτές και ενδεχόμενους χαμένος καταναλωτές που θα βασιστούν στη φήμη σχετικά με τη μη διαθεσιμότητα.
- Συνεργασίες σε επίπεδο παρόχων τόσο με μάρκες καινούριων προϊόντων όσο και μεταχειρισμένων και ανακατασκευασμένων καθώς αποτελεί μια δημοφιλή κατηγορία προϊόντων με ενδεχόμενα κέρδη στην αγορά.

Είναι προφανές, λοιπόν, πως η διαδικασία καθορισμού των τιμών εμπλέκεται άμεσα με τη διαδικασία καθορισμού των υπόλοιπων χαρακτηριστικών της πώλησης όπως είναι τα έξοδα αποστολής από τους ηλεκτρονικούς παρόχους, η κατάσταση που θα βρίσκεται το προϊόν και τα επίπεδα διαθεσιμότητας τους στα αποθέματα των παρόχων. Για να σχεδιαστεί μια τέτοια στρατηγική είναι σημαντικό όχι απλά να προβλέψουμε τις τιμές αλλά να σχεδιάσουμε τη μείωση τους όποτε χρειάζεται και συγκριτικά με τον ανταγωνισμό και αφού ορίσουμε τη γενικότερη στρατηγική μαρκετινγκ και πωλήσεων να καθορίσουμε τα επιμέρους χαρακτηριστικά των προϊόντων.

Τέλος, είναι βασική προϋπόθεση αυτή η στρατηγική να αναπροσαρμόζεται συνεχώς ώστε να αποκρίνεται όσο πιο δυναμικά γίνεται στις νέες τάσεις και ανάγκες της αγοράς και έτσι ο πάροχος και το εκάστοτε προϊόν να είναι ανταγωνιστικό και προτιμητέο από τους καταναλωτές και τους νέους ενδεχόμενους αγοραστές που εντάσσονται στην αγορά.

6.2 Προεκτάσεις

Στην παρούσα διπλωματική εξετάστηκαν κάποιοι παράγοντες που καθορίζουν τις πωλήσεις των προϊόντων και παίζουν σημαντικό ρόλο στην εκάστοτε στρατηγική τιμολόγησης που θα

ακολουθηθεί. Παρόλα αυτά υπάρχουν και επιπλέον παράγοντες που δεν μελετήθηκαν και θα αναφερθούν παρακάτω με στόχο την επέκταση που θα μπορούσε να γίνει στη μελέτη αυτή, καθώς είναι πολλές οι παράμετροι που μπορούμε να συνδυάσουμε με τις τιμές για να κάνουμε το προϊόν ελκυστικό.

Αρχικά, μια σημαντική υπηρεσία που συνδέεται με τα χαρακτηριστικά αυτά και αποτελεί και βασικό στοιχείο του ηλεκτρονικού εμπορίου είναι η καλή εξυπηρέτηση πελατών. Η παροχή υπηρεσιών στους πελάτες είναι ένας από τους καλύτερους τρόπους για να θεμελιωθεί και να βελτιωθεί η εμπιστοσύνη των πελατών στην επιχείρησή. Είναι σημαντικό ένας ηλεκτρονικός πάροχος να διαθέτει τμήμα εξυπηρέτησης πελατών που είτε θα γίνει σε πραγματικό χρόνο είτε λίγο ετεροχρονισμένα να μπορεί να ικανοποιήσει όλα τα ερωτήματα πελατών εγκαίρως και να διατηρήσει μια φιλική προσέγγιση κατά την επικοινωνία με τους αγοραστές. Η αντιμετώπιση των προβλημάτων των πελατών τους βοηθά να αισθάνονται πιο άνετα προς την επιχείρησή ηλεκτρονικού εμπορίου και εν τέλει αυξάνει τη δημοφιλία των συγκεκριμένων προϊόντων.

Ακόμη, ένα άλλο χαρακτηριστικό που θα μπορούσε να μελετηθεί είναι η διαφήμιση και γενικά το πόσο επιτυχημένο είναι το μάρκετινγκ του προϊόντος. Η προώθηση στα κοινωνικά δίκτυα προσφέρει μια σημαντική ευκαιρία για να διαφημιστεί τόσο ο πάροχος όσο και το προϊόν και να ενισχύσει την παρουσία της μάρκας στο διαδίκτυο. Ακόμη, η δημιουργία περιεχομένου σχετικά με προϊόντα και ο τρόπος παρουσίασης, περιγραφής και γενικότερα δημοσίευσή τους στην ιστοσελίδα είναι ένα πρόσθετο κομμάτι που είναι σημαντικό να μελετηθεί και παίζει ρόλο για τη δημοφιλία τους.

Συμπληρωματικά, οι μέθοδοι πληρωμής επηρεάζουν σίγουρα τη δημοφιλία ενός προϊόντος και γενικά ενός παρόχου. Οι online συναλλαγές χρημάτων γίνονται δημοφιλείς ολοένα και περισσότερο και οι αγοραστές είναι εξοικειωμένοι με τη μέθοδο πληρωμής σε ηλεκτρονικά συστήματα. Ένα προϊόν μπορεί να αγοραστεί με διαφορετικές επιλογές όπως συναλλαγές με πιστωτικές και χρεωστικές κάρτες, καθαρές τραπεζικές συναλλαγές, ή ακόμη και εξατομικευμένο "πορτοφόλι" (wallet) του παρόχου ή ευρέως χρησιμοποιούμενα τέτοια ηλεκτρονικά πορτοφόλια. Ο τρόπος πληρωμής μπορεί να είναι και αυτός ένας παράγοντας που μπορεί να εξεταστεί και να επηρεάσει την προτίμηση ενός καταναλωτή.

Τέλος, πολλά άλλα χαρακτηριστικά των προϊόντων μπορούν να παίζουν άλλοτε μικρότερο και άλλοτε μεγαλύτερο ρόλο στον καθορισμό των τιμών. Μια επέκταση της παρούσας μελέτης, λοιπόν, θα ήταν η έρευνα για στοιχεία δεδομένων και για αυτά τα χαρακτηριστικά και η εφαρμογή τους στο μοντέλο ώστε να καθοριστεί ο βαθμός που επηρεάζουν τη δημοφιλία των προϊόντων και άρα τη στρατηγική που περιγράφηκε.

Ακόμα πέρα από τα επιπλέον χαρακτηριστικά, μια προέκταση της διπλωματικής αυτής θα αφορούσε και τη χρήση εναλλακτικών ή παραπάνω δεδομένων. Τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα εργασία ήταν μηναία και άρα ο ορίζοντας των

προβλέψεων περιορισμένος. Αλλάζοντας τον τύπο των δεδομένων θα μπορούσε να προκύψει μια νέα μελέτη για εβδομαδιαία ή ετήσια δεδομένα.

Συμπληρωματικά, ένας ακόμη παράγοντας που θα αποτελούσε παραλλαγή της εργασίας θα ήταν η μελέτη όχι μόνο της δημοφιλίας μέσω των προβολών αλλά και μέσω του αριθμού των πωλήσεων. Η εύρεση δεδομένων πωλήσεων σχετικά με τα προϊόντα αυτά θα βοηθούσε ώστε να καθοριστεί πιο εύκολα η ζήτηση των προϊόντων και έτσι να γίνει μια πιο εκτενής και ρεαλιστική μελέτη για τη δημοφιλία τους και τελικά τη στρατηγική τιμολόγησης και πώλησης τους.

Τέλος, εφόσον η διαδικασία βασίζεται σε αλγορίθμους μηχανικής μάθησης και ιδιαίτερα ταξινόμησης, μια παραλλαγή που προτείνεται είναι η δοκιμή και άλλων μεθόδων ταξινόμησης με στόχο την ακόμη μεγαλύτερη ακρίβεια ή ακόμη και προσπάθεια βελτιστοποίησης των υπαρχόντων ώστε να επιτευχθεί μέγιστη αποδοτικότητα.

6.3 Επίλογος

Ένας από τους πυλώνες της πώλησης προϊόντων είναι αδιαμφισβήτητα η τιμολόγηση. Είναι μια διαδικασία που συνδέεται άρρηκτα με την ανάπτυξη του προϊόντος, τη θέση του, και τον χώρο όπου πωλείται. Ειδικότερα στις ηλεκτρονικές πωλήσεις που αυτά τα στοιχεία είναι ακόμη λιγότερο σημαντικά, η τιμή αποτελεί έναν καθοριστικό παράγοντα για τη δημοφιλία του προϊόντος. Όποια τιμή και αν οριστεί, θα αποτελέσει σημαντικό παράγοντα σε πολλά στοιχεία όπως ο όγκος των πωλήσεων, τα κέρδη ακόμα και ο τρόπος με τον οποίο γίνεται αντιληπτή η μάρκα του προϊόντος αφού όλα αυτά καθορίζονται από την εκάστοτε στρατηγική τιμολόγησης.

Παρόλα αυτά λόγω της πολυπαραγοντικοτητας που εμφανίζει πλέον το ηλεκτρονικό εμπόριο, με μεγάλη ποικιλία παρόχων και διαφορετικές μεθόδους πωλήσεων, η τιμή δεν είναι το μοναδικό χαρακτηριστικό που πρέπει να μελετηθεί. Αντιθέτως είναι σημαντικό να σχεδιαστεί μια ολοκληρωμένη στρατηγική πώλησης των προϊόντων, η οποία θα είναι σε πλήρη παραλληλία με τις τάσης της αγοράς και θα στοχεύει στην αντιμετώπιση του ανταγωνισμού, ώστε το εκάστοτε προϊόν να γίνει δημοφιλές και προτιμητέο.

Μέσω της παρούσας διπλωματικής, αποδείχθηκε αρχικά η σημασία της χρήσης της μηχανικής μάθησης σε όλη αυτή τη διαδικασία. Οι αλγόριθμοι που αναπτύχθηκαν αποτέλεσαν τη βάση των προβλέψεων, οι οποίες έγιναν άμεσα, εύκολα και με σχετικά μεγάλη αποδοτικότητα και ακρίβεια. Για αυτό βασικό πόρισμα της διπλωματικής είναι η ευρεία χρήση των νευρωνικών δικτύων σε αντίστοιχα προβλήματα που δύναται να παρέχει επιθυμητά και αποδοτικά αποτελέσματα.

Ακόμη, μέσω της διαδικασίας που ακολουθήθηκε καταλήγουμε στην αδυναμία καθορισμού των τιμών με στατικό και μονόπλευρο τρόπο. Αντιθέτως προτείνεται μια διαδικασία που

βασίζεται στη μελέτη της δημοφιλίας των προϊόντων και στις μεταβολές παραγόντων σύμφωνα με τις νέες ανάγκες και με δυναμικό τρόπο έτσι ώστε να προϊόντα να γίνονται πιο δημοφιλή και πιο ανταγωνιστικά στην εκάστοτε αγορά και ιδιαίτερα στην ηλεκτρονική αγορά που αποτελεί και το αντικείμενο έρευνας στην παρούσα περίπτωση.

Συμπερασματικά η διπλωματική αυτή είναι μια μελέτη που μπορεί να σταθεί ως σημείο αναφοράς και βάση για περαιτέρω ανάπτυξη μιας στρατηγικής τιμολόγησης και πώλησης προϊόντων με έναν δυναμικό τρόπο και με στόχο τη συνεχή μεταβολή παραγόντων που θα καθιστούν ένα προϊόν πιο ανταγωνιστικό και δημοφιλές στην αγορά.

Βιβλιογραφία

1. Νευρωνικά Δίκτυα και Μηχανική Μάθηση - 3η έκδοση, Simon Haykin, Εκδόσεις Παπασωτηρίου, 2009
2. Επιχειρησιακές Προβλέψεις, Πετρόπουλος Φ., Ασημακόπουλος Β., (2011). εκδόσεις συμμετρία, Αθήνα.
3. John C. Chambers, Satinder K. Mullick, and Donald D. Smith, How to Choose the Right Forecasting Technique (1971), Magazine, Harvard Business Review
4. Persistent Forecasting of Disruptive Technologies. (2010) Washington, DC: The National Academies Press
5. Robert Nau, Review of basic statistics and the simplest forecasting model: the sample mean (2014), Fuqua School of Business, Duke University
6. Frischmann, Tanja & Hinz, Oliver & Skiera, Bernd. (2012). Retailers' Use of Shipping Cost Strategies: Free Shipping or Partitioned Prices?. International Journal of Electronic Commerce. 16. 65-87. 10.2307/23106403.
7. Hyun Ju Jeong & Kyoung-Nan Kwon (2012) The Effectiveness of Two Online Persuasion Claims: Limited Product Availability and Product Popularity, Journal of Promotion Management, 18:1, 83-99
8. S. Shakya, M. Kern, G. Owusu, C.M. Chin, Neural network demand models and evolutionary optimisers for dynamic pricing, Knowledge-Based Systems, Volume 29, 2012, Pages 44-53, ISSN 0950-7051
9. Min Qi, Sha Yang, Forecasting consumer credit card adoption: what can we learn about the utility function?, International Journal of Forecasting, Volume 19, Issue 1, 2003, Pages 71-85, ISSN 0169-2070
10. Patricia M. West, Patrick L. Brockett, Linda L. Golden (1997) A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice. Marketing Science 16(4):370-391.
11. Harald Hruschka, Werner Fettes, Markus Probst, An empirical comparison of the validity of a neural net based multinomial logit choice model to alternative model specifications, European Journal of Operational Research, Volume 159, Issue 1, 2004, Pages 166-180, ISSN 0377-2217
12. Rainer Schlosser, Stochastic dynamic pricing and advertising in isoelastic oligopoly models, European Journal of Operational Research, Volume 259, Issue 3, 2017, Pages 1144-1155, ISSN 0377-2217
13. Somayeh Najafi- Ghobadi, Jafar Bagherinejad, Ata Allah Taleizadeh, A two-generation new product model by considering forward-looking customers, Dynamic pricing and advertising optimization, Journal of Retailing and Consumer Services, Volume 63, 2021, 102387, ISSN 0969-6989
14. Dongfan Wang, Zhen He, Shuguang He, Zhaomin Zhang, Yiwen Zhang, Dynamic pricing of two-dimensional extended warranty considering the impacts of product price fluctuations and repair learning, Reliability Engineering & System Safety, Volume 210, 2021, 107516, ISSN 0951-8320

15. Nishika Bhatia, Nalan Gülpınar, Nurşen Aydın, Dynamic production-pricing strategies for multi-generation products under uncertainty, International Journal of Production Economics, Volume 230, 2020, 107851, ISSN 0925-5273
16. Arnoud V. den Boer, Dynamic pricing and learning: Historical origins, current research, and new directions, Surveys in Operations Research and Management Science, Volume 20, Issue 1, 2015, Pages 1-18, ISSN 1876-7354
17. Deksnyte, Indre & Lydeka, Prof. (2012). Dynamic Pricing and Its Forming Factors. International Journal of Business and Social Science. 3
18. Wen Guang Qu, Alain Pinsonneault, Daniel Tomiuk, Shaoqing Wang, Yuan Liu, The impacts of social trust on open and closed B2B e-commerce: A Europe-based study, Information & Management, Volume 52, Issue 2, 2015, Pages 151-159, ISSN 0378-7206
19. Jying-Nan Wang, Jiangze Du, Ya-Ling Chiu, Jin Li, Dynamic effects of customer experience levels on durable product satisfaction: Price and popularity moderation, Electronic Commerce Research and Applications, Volume 28, 2018, Pages 16-29, ISSN 1567-4223
20. He, Qiao-Chu & Chen, Ying-Ju. (2017). Dynamic Pricing of Electronic Products with Consumer Reviews. Omega. 80. 10.1016/j.omega.2017.08.014.
21. Pricing Electronic Products, Volume 1, Issue 1&2, January, 1995
22. Jaspreet, A Concise History of Neural Networks (2016), towardsdatascience.com
23. Chanin Nantasenamat, How to Build a Machine Learning Model: A Visual Guide to Learning Data Science (2020), towardsdatascience.com
24. Mahbubul Alam, Data normalization in machine learning (2020), towardsdatascience.com
25. Will Koehrsen, Hyperparameter Tuning the Random Forest in Python (2018), towardsdatascience.com
26. Sruthi Korlakunta, Leading a Data Science Project from Scratch (2021), towardsdatascience.com
27. Jeff Hale, Scale, Standardize, or Normalize with Scikit-Learn (2019), towardsdatascience.com
28. Sergio Santoyo, A Brief Overview of Outlier Detection Techniques (2017), towardsdatascience.com
29. Mukesh Mithrakumar, How to tune a Decision TreeMukesh Mithrakumar (2019), towardsdatascience.com
30. Gabriela Barkho, Refurbished electronics marketplaces are having a moment (2021), modernretail.co
31. Avery Walts, How to Determine the Best Shipping Cost For Your Business (2020), skuvault.com
32. Nicky LaMarco, Is Cutting Prices a Good Marketing Strategy? (2019), Small Business, Business Planning & Strategy, Market Pricing Strategies, smallbusiness.chron.com/
33. S won Lee, Predict the price of products using Machine Learning (2020), medium.com

34. Sciforce, Evolution of Forecasting from the Stone Age to Artificial Intelligence, medium.com
35. Leané Mulder, Consistent Product Availability Leads To Improved Customer Experiences (2021), dotactiv.com
36. Joey Blanco Joey Blanco , Shipping Rates 101: How to Calculate Shipping Costs (2021), bigcommerce.com
37. Jason Brownlee, Failure of Classification Accuracy for Imbalanced Class Distributions (2020), machinelearningmastery.com
38. Jason Brownlee, Difference Between Classification and Regression in Machine Learning (2017), machinelearningmastery.com
39. Jason Brownlee, Recursive Feature Elimination (RFE) for Feature Selection in Python (2020), machinelearningmastery.com
40. Jason Brownlee, How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification (2020), machinelearningmastery.com
41. Steve Mutuvi, Introduction to Machine Learning Model Evaluation (2019), heartbeat.comet.ml
42. Sadhvi Anunaya, Data Preprocessing in Data Mining -A Hands On Guide (2021), analyticsvidhya.com
43. Matt Payne, Dynamic Pricing: How Pricing Optimization And Revenue Management Benefit From Machine Learning (2021), scalr.ai
44. Leo Gimenez, 6 steps for data cleaning and why it matters (2020), Maintenance, geotab.com
45. Lauren Erdelyi, The Five Stages of Data Analysis (2021), lighthouselabs.ca
46. Jeff Bodenstab, The Evolution of Forecasting (2015), toolsgroup.com
47. Mohammad Waseem, How To Implement Classification In Machine Learning? (2021), edureka.co
48. Alivia Smith, 7 Fundamental Steps to Complete a Data Analytics Project (2019), dataiku.com