Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας, Πληροφορικής &
Υπολογιστών

Τεχνικές Μηχανικής Μάθησης σε Προβλήματα Πρόβλεψης Τιμών Ακινήτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΗΝΑΪΔΗ ΜΑΡΙΑ ΝΕΚΤΑΡΙΑ

Επιβλέπων : **Δημήτριος Φωτάκης**
Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2020

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας, Πληροφορικής &
Υπολογιστών

Τεχνικές Μηχανικής Μάθησης σε Προβλήματα Πρόβλεψης Τιμών Ακινήτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΗΝΑΪΔΗ ΜΑΡΙΑ ΝΕΚΤΑΡΙΑ

Επιβλέπων : Δημήτριος Φωτάκης
Αν. Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 22η Οκτωβρίου 2020.

…….	…….	…….
Δημήτριος Φωτάκης	Παπασπύρου Νικόλαος	Παγουρτζής Αριστείδης
Αν. Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2020

……
Μηναΐδη Μαρία Νεκταρία

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η αγορά των ακινήτων στην σημερινή εποχή, χαρακτηρίζεται από ιδιαίτερη οικονομική αστάθεια και επηρεάζεται από πλήθος, επίσης ασταθών και ποικίλης φύσεως, παραγόντων. Ως σημαντική πτυχή του ανθρώπινου πολιτισμού, είναι άρρηκτα συνδεδεμένη με την οικονομία της κάθε περιοχής και αποτελεί δείκτη οικονομικής ευμάρειας και ευημερίας, καθώς επίσης και καθοριστικό παράγοντα του επιπέδου διαβίωσης μιας κοινωνίας. Η κατανόηση της συμπεριφοράς και των τάσεων της αγοράς των ακινήτων μπορεί να παρέχει προτάσεις όχι μόνο για τους ενδιαφερόμενους αγοραστές, αλλά και για τους ερευνητές της αγοράς και τους αρμόδιους για την λήψη των αποφάσεων όσον αφορά την αγορά των ακινήτων, τον σχεδιασμό της πόλης και την αστική ανάπτυξη.

Οι παράγοντες που επηρεάζουν τις τιμές της γης και των ακινήτων είναι πολυάριθμοι και εξαρτώνται από πλήθος μεταβλητών, σχετίζονται έμμεσα ή άμεσα μεταξύ τους και αλληλεπιδρούν συνέχεια. Γίνεται, έτσι, κατανοητό ότι η μοντελοποίησή τους συνιστά μία αρκετά σύνθετη διαδικασία, η οποία απαιτεί την λεπτομερή μελέτη τους. Μεγάλες εταιρείες και οργανισμοί έχουν κάνει βήματα προς την ανάπτυξη και εδραίωση εφαρμογών, οι οποίες προβλέπουν τις τιμές των κατοικιών, βρίσκονται ωστόσο σε πρώιμα στάδια.

Στο πλαίσιο της παρούσας διπλωματικής εργασίας, μελετάται η πρόβλεψη των τιμών των ακινήτων. Παρόλο που πολλές μελέτες έχουν διεξαχθεί με σκοπό την ανάλυση του συγκεκριμένου ζητήματος, η εφαρμογή των συμπερασμάτων τους στην πράξη έχει γίνει σε ελάχιστες περιπτώσεις. Ιδιαίτερα στην χώρα μας, παρά τον ρόλο που κατέχει η αγορά των ακινήτων στην οικονομία μας, δεν υπάρχουν, τουλάχιστον δημοσιευμένες, έρευνες ή εφαρμογές που να απαντούν σε αυτό το πρόβλημα.

Στην εργασία αυτή, επιχειρούμε να δώσουμε μία λύση στο ζήτημα που παρουσιάστηκε, επικεντρώνοντας το ενδιαφέρον της μελέτης στην περιοχή της Αττικής, λαμβάνοντας υπόψιν σημαντικούς παράγοντες από τους οποίους καθορίζονται οι τιμές των ακινήτων και χρησιμοποιώντας μεθόδους και τεχνικές της μηχανικής μάθησης. Αναλύονται, στη συνέχεια, τα αποτελέσματα που λάβαμε από τις μεθόδους που εφαρμόσαμε και αξιολογούνται η ακρίβεια και η καταλληλότητά τους για το παρόν πρόβλημα.

Πιο συγκεκριμένα, γίνεται ιστορική αναδρομή και περιγράφεται το πλαίσιο της μελέτης μας, ενώ κατόπιν παρουσιάζεται αναλυτικά το πρόβλημα με το οποίο καταπιάνεται η εργασία. Συνεχίζουμε με την παρουσίαση του συνόλου δεδομένων που χρησιμοποιήθηκε και του τρόπου συλλογής του. Ακολουθεί η λεπτομερής ανάλυση και διερευνητική προεπεξεργασία του, προκειμένου να είναι δυνατή η πλήρης κατανόησή του και η βέλτιστη αξιοποίησή του.

Στη συνέχεια, υλοποιούνται έξι διαφορετικά μοντέλα μηχανικής μάθησης, τα οποία εκπαιδεύονται με βάση τα δεδομένα. Η πλειοψηφία των μοντέλων αυτών στηρίζεται σε τεχνικές και αλγορίθμους παλινδρόμησης, όπως είναι η πολλαπλή γραμμική παλινδρόμηση, η οποία αποτελεί τη βάση για την ανάπτυξη πιο σύνθετων και αποδοτικών τεχνικών παλινδρόμησης, όπως είναι η παλινδρόμηση Ridge, η Lasso και η παλινδρόμηση με χρήση της τεχνικής Gradient Boosting. Επίσης, υλοποιήθηκαν μοντέλα που βασίζονται στα δέντρα απόφασης, όπως είναι τα Τυχαία Δάση. Ακολούθως, παρουσιάζονται και αξιολογούνται τα αποτελέσματα της εφαρμογής τους στα δεδομένα εκπαίδευσης και τα δεδομένα ελέγχου.

Τέλος, παρουσιάζονται κάποιες συγκρίσεις των αποτελεσμάτων μας με αποτελέσματα αντίστοιχων ερευνών, εντοπίζονται τα σημεία που επιδέχονται βελτίωση στην μεθοδολογία

που ακολουθήθηκε και προτείνονται με βάση αυτά κάποιες μελλοντικές προοπτικές έρευνας για τις επόμενες μελέτες.

Λέξεις Κλειδιά

Πρόβλεψη τιμών ακινήτων, πολλαπλή γραμμική παλινδρόμηση, ενίσχυση κλίσης, lgbm, τυχαία δάση, μηχανική μάθηση, διερευνητική ανάλυση δεδομένων.

Abstract

The real estate market today is characterized by particular economic instability and is influenced by a number of factors, also unstable and of various natures. As an important aspect of human culture, it is inextricably linked to the economy of each region and is an indicator of economic prosperity and well-being, as well as a determining factor in the standard of living of a society. Understanding the behavior and trends of the real estate market can provide suggestions not only for interested buyers, but also for market researchers and those responsible for making decisions regarding the real estate market, city planning and urban development.

The factors that affect land and real estate prices are numerous and depend on a number of variables, directly or indirectly related to each other and constantly interact. It is thus understood, that their modeling is a rather complex process, which requires their detailed study. Large companies and organizations have taken steps to develop and consolidate applications that predict housing prices, but are in the early stages.

In the context of this dissertation, the forecast of real estate prices is studied. Although many studies have been conducted to analyze this issue, the application of their conclusions in practice has been done in a few cases. Especially in our country, despite the role of the real estate market in our economy, there are no, at least published, surveys or applications that answer this problem.

In this work, we try to provide a solution to the problem that arose, focusing the interest of the study in the region of Attica, taking into account important factors that determine property prices and using methods and techniques of machine learning. The results obtained from the methods we applied are then analyzed and their accuracy and appropriateness for the present problem are evaluated.

More specifically, a historical review is made and the context of our study is described, while then the problem that the work is dealing with is presented in detail. We continue with the presentation of the data set used and how it was collected. Following is the detailed analysis and exploratory preprocessing, in order to be able to fully understand and use it optimally.

Next, six different machine learning models are implemented, which are trained based on the data. The majority of these models rely on regression techniques and algorithms, such as multiple linear regression, which is the basis for the development of more complex and efficient regression techniques, such as Ridge regression, Lasso and Gradient Boosting regression. Models based on decision trees, such as Random Forests, were also implemented. Then, the results of their application to the training data and the control data are presented and evaluated.

Finally, some comparisons of our results with the results of relevant studies are presented, the points that can be improved in our research are identified and based on these, some future research perspectives are proposed.

Key Words

House price prediction, multiple linear regression, gradient boosting, lgbm, random forests, machine learning, exploratory data analysis.

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου, κύριο Δημήτριο Φωτάκη, για την συνεχή καθοδήγησή του και την άμεση ανταπόκρισή του σε κάθε μου ερώτημα, καθ' όλη τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας. Οι σωστές παρατηρήσεις του με βοήθησαν να ξεπεράσω τις δυσκολίες που παρουσιάστηκαν και να ολοκληρώσω με επιτυχία την παρούσα εργασία. Η αγάπη του για την μάθηση και την επιστήμη μεταδίδεται στους μαθητές του και αυτό είναι που τον ξεχωρίζει ως καθηγητή.

Επίσης, θα ήθελα να ευχαριστήσω ιδιαίτερα τους γονείς μου και τον αδερφό μου, οι οποίοι, χωρίς δεύτερη σκέψη, με στήριξαν σε όλα τα χρόνια των σπουδών μου. Χωρίς την ανιδιοτελή αγάπη τους και την καθοδήγησή τους δεν θα μπορούσα να τα είχα καταφέρει.

Ακόμα, θα ήθελα να πω ένα μεγάλο ευχαριστώ στους θείους μου και τη νονά μου, οι οποίοι προσφέροντας τις γνώσεις τους, με συμβουλεύουν, με στηρίζουν και είναι πάντα παρόντες. Στους παππούδες μου και τις γιαγιάδες μου οφείλω ένα ακόμα μεγάλο ευχαριστώ για την αμέριστη αγάπη τους.

Τέλος, θα ήθελα να ευχαριστήσω όλους τους φίλους μου και τους συμφοιτητές μου, οι οποίοι στάθηκαν δίπλα μου κατά τη διάρκεια των φοιτητικών μου χρόνων, βοηθώντας με ο καθένας με τον δικό του τρόπο και αποτελώντας ένα πολύ ευχάριστο κομμάτι των χρόνων αυτών.

Μηναΐδη Μαρία Νεκταρία,
Αθήνα, 22η Οκτωβρίου 2020.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κεφάλαιο 1

Εισαγωγή

Σε αυτήν την Ενότητα θα παρουσιαστεί το πρόβλημα με το οποίο καταπιάνεται η παρούσα διπλωματική εργασία, η υπάρχουσα βιβλιογραφία, καθώς επίσης και η δομή της εργασίας, όπως αυτή διαμορφώνεται στη συνέχεια.

1.1 Τιμολόγηση Ακινήτων

Από τα πρώτα χρόνια της ύπαρξης του ανθρώπου στην γη, μία από τις βασικότερες ανάγκες του, αποτέλεσε η ανάγκη για στέγαση. Η κατοικία, ή οικία του ανθρώπου, εμφανίζεται στην πιο απλή μορφή της από την παλαιολιθική εποχή, με τον άνθρωπο να καταφεύγει σε σπήλαια κατά τις μετακινήσεις του, όπου και διέμενε, προκειμένου να προστατευθεί από τις καιρικές συνθήκες ή άλλες απειλές. Οι πρώτες μόνιμες ανθρώπινες κατοικίες παρατηρούνται την νεολιθική εποχή, οπότε και έχουμε την ανάπτυξη των πρώτων μόνιμων οικισμών με πέτρινα ή ξύλινα σπίτια. Στην αρχαία Αθήνα, έχουμε απλές και λιτές κατοικίες από πέτρα, πηλό, ξύλο και κεραμίδια. Πλησιάζοντας το σήμερα, εδραιώνεται η έννοια της ιδιοκτησίας και σε συνδυασμό με την μεγάλη τεχνολογική πρόοδο που βιώνει η ανθρωπότητα, αλλάζει ριζικά η έννοια της κατοικίας.

Η αξία της γης και η εκτίμηση αυτής, αλλά και των ακίνητων κατοικιών, αποτελεί ζήτημα, το οποίο έχει απασχολήσει τον άνθρωπο από τον καιρό της εμφάνισης των πρώτων οικισμών. Σήμερα, η ανάγκη αυτή προκύπτει από την δυνατότητα διάθεσης του ακινήτου και συνεπώς κρίνεται καίριο να εκφραστεί σε χρήματα η αξία του. Οι παράγοντες από τους οποίους αυτή εξαρτάται, είναι τόσο πολυάριθμοι, όσο και περίπλοκοι, εφόσον ενδέχεται να είναι πολεοδομικής, πολιτικής, κοινωνικής, οικονομικής, αλλά και προσωπικής φύσεως. Οι συχνές αλλαγές των παραπάνω παραγόντων, αλλά και η πολυάριθμες έννοιες της αξίας, πολυπλέκουν ακόμα περισσότερο τα δεδομένα μας και έχουν ως συνέπεια την γένεση της ανάγκης του σχηματισμού και της θεμελίωσης ενός «συστήματος» κανόνων εκτίμησης της αξίας των κατοικιών και της γης. Την λύση στο παραπάνω ζήτημα έρχεται να δώσει το Σύνταγμα κάθε χώρας, με την συμβολή του οποίου κατοχυρώνεται η αξία του ακινήτου και προσδιορίζεται επακριβώς μέσω της επιστήμης και της συστηματικής σκέψης. [1] Η γνώση της ακριβούς αυτής αξίας συμβάλλει σε πολλούς σκοπούς και στόχους οι οποίοι είτε προβλέπονται από την νομοθεσία, όπως φορολογικής φύσεως, είτε αποτελούν απόρροια των αναγκών της καθημερινής ζωής.

Στην χώρα μας, η αξία των ακινήτων δεν προσδιορίζεται ενιαία και συνολικά από κάποια νομοθετική ρύθμιση, γεγονός που οδηγεί στην εκτίμηση αυτής με τρόπο διαφορετικό ανά την περίπτωση. Αποτέλεσμα αυτού αποτελεί, η πιθανότητα για ένα ακίνητο να υπάρχει ένα πλήθος υπερτιμημένων ή υποτιμημένων τιμών, οι οποίες ενδέχεται να πλησιάζουν την πραγματική αξία, οι αποκλίσεις ωστόσο μπορεί να είναι πολύ μεγάλες. Η αγορά των ακινήτων και η οικονομία μίας χώρας συνδέονται άρρηκτα μεταξύ τους, αλληλεπιδρούν και αλληλοεπηρεάζονται. Σε προσωπικό επίπεδο, η αγορά ενός ακινήτου αποτελεί, σημαντική απόφαση στην ζωή ενός πολίτη, καθώς η σωστή επιλογή μπορεί να

αποδειχθεί σημαντική επένδυση για το μέλλον του. Γεννιέται, λοιπόν, η αναγκαιότητα δημιουργίας και εδραίωσης ενός συστηματικού και καλά ορισμένου τρόπου προσδιορισμού της αξίας ενός ακινήτου, προκειμένου να καλυφθούν οι σημερινές και μελλοντικές, ολοένα πολλαπλασιαζόμενες ανάγκες της κοινωνίας μας. Το σύστημα αυτό κρίνεται απαραίτητο να είναι δυναμικό, να μεταβάλλεται, δηλαδή, η τιμή των ακινήτων ανάλογα με τις μεταβολές των παραγόντων που την επηρεάζουν.

Σημαντικό ρόλο στην σημερινή αγορά των ακινήτων στην Ελλάδα αλλά και το εξωτερικό, κατέχουν οι μεσίτες, οι οποίοι προσφέρουν κτηματομεσιτικές υπηρεσίες στους πολίτες που ενδιαφέρονται για την αγορά, πώληση ή ενοικίαση κάποιου ακινήτου ή γης. Έχουν την δυνατότητα να εκτιμήσουν την εμπορικότητα, εκμεταλλευσιμότητα, οικοδομησιμότητα, καταλληλότητα και την αξία των ακινήτων, αλλά και να ρυθμίσουν υπό όρους την αγοραπωλησία τους. Η ραγδαία εξέλιξη της τεχνολογίας και η ανάπτυξη της επιστήμης της Πληροφορικής και της Μηχανικής Μάθησης έχουν πλέον ανοίξει νέους δρόμους σε πολλές πτυχές της ζωής μας, αλλάζοντας σημαντικά τον τρόπο ζωής του σύγχρονου ανθρώπου.

Η Τεχνητή Νοημοσύνη και η Μηχανική Μάθηση έχουν κάνει τα πρώτα τους βήματα, εισχωρώντας εκτός των άλλων και στον τομέα της αγοράς των ακινήτων. Μεγάλες εταιρείες στην περιοχή του Real Estate, όπως είναι η Airbnb και η Zillow, έχουν υλοποιήσει εργαλεία που βασίζονται στην Τεχνητή Νοημοσύνη, βασικές λειτουργίες των οποίων είναι το ταίριασμα ενός αγοραστή ή ενοικιαστή ακινήτου με έναν ιδιοκτήτη, αλλά και η εκτίμηση της τιμής ακινήτων. Για παράδειγμα το πρόγραμμα Zestimate της εταιρείας Zillow αποτελεί ακριβώς αυτό: μία πλατφόρμα μέσω της οποίας κάθε κάτοικος της Αμερικής έχει την δυνατότητα να λάβει μία εκτίμηση της αξίας του ακινήτου του. Οι εκτιμήσεις αυτές βελτιώνονται καθημερινά, εφόσον ο όγκος των δεδομένων τους ολοένα και αυξάνεται.

Στην χώρα μας, δυστυχώς, δεν υπάρχει μέχρι στιγμής κάποια υπηρεσία που να προσφέρει το προαναφερθέν. Παρά την αυξημένη μελέτη γύρω από την περιοχή της αγοράς ακινήτων και της μηχανικής μάθησης, μικρή εφαρμογή των τεχνολογιών αυτών σε προβλήματα κοστολόγησης κατοικιών, φαίνεται να υπάρχει. Το τελευταίο, σε συνδυασμό και με το πόσο σημαντική κρίνεται η ακριβής αποτίμηση της αξίας των ακινήτων, γεννάει την επιθυμία για περαιτέρω έρευνα και πρακτική δοκιμή των μεθόδων μηχανικής μάθησης, με σκοπό την αυτοματοποίηση της τιμολόγησης των κατοικιών. Η παρούσα διπλωματική εργασία επικεντρώνεται σε αυτήν ακριβώς την προσπάθεια.

1.2 Υπάρχουσα Βιβλιογραφία

Η θεματική που θίγεται σε αυτήν την εργασία, αφορά το υποπεδίο της επιστήμης των υπολογιστών που ονομάζεται μηχανική μάθηση, η οποία στηριζόμενη σε μοντέλα της οικονομετρίας, κλάδου της επιστήμης των οικονομικών, χρησιμοποιείται για την μελέτη και υλοποίηση αλγορίθμων, που έχουν την δυνατότητα να «μαθαίνουν» από τα δεδομένα και με βάση αυτά να κάνουν τις εκάστοτε προβλέψεις. Το πρόβλημα που μελετάται στην παρούσα εργασία, αποτελεί ένα πρόβλημα πολλαπλής γραμμικής παλινδρόμησης επιβλεπόμενης μάθησης. Η μοντελοποίηση της σχέσης μεταξύ της εξαρτώμενης μεταβλητής, στην προκειμένη περίπτωση της τιμής των ακινήτων, και των ανεξάρτητων μεταβλητών, που αποτελούν τα χαρακτηριστικά του κάθε ακινήτου, συνιστά τον απώτερο σκοπό του. Εδώ και αρκετά χρόνια, πραγματοποιούνται προσπάθειες και γίνονται μελέτες με θέμα το παραπάνω πρόβλημα, επιστρατεύοντας τεχνικές μηχανικής μάθησης για την επίλυση αυτού, από πολλούς μελετητές και επιστήμονες, των οποίων το έργο υπάρχει δημοσιευμένο. Οι μελέτες στις οποίες στηρίχθηκε η παρούσα έρευνα παρουσιάζονται στη συνέχεια.

Το 2015 οι Yingyu Feng και Kelvyn Jones χρησιμοποίησαν δύο προηγμένες μεθόδους μοντελοποίησης, τα πολύ-επίπεδα μοντέλα και τα τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks), προκειμένου να προβλέψουν τις τιμές των σπιτιών της περιοχής του Μπρίστολ.[2] Τις μεθόδους αυτές τις σύγκριναν με την τυπική μέθοδο της ηδονικής τιμολόγησης, όσον αφορά την ακρίβεια της πρόγνωσης, την ικανότητα να συλλάβουν πληροφορίες τοποθεσίας και την επεξηγηματική τους δύναμη. Η έρευνα αυτή χρησιμοποίησε δεδομένα από το Μητρώο Ακινήτων του Μπρίστολ. Τα αποτελέσματα υποδεικνύουν ότι τα πολύ-επίπεδα μοντέλα δίνουν μεγαλύτερη ακρίβεια προβλέψεων και υψηλότερη επεξηγηματική ισχύ σε σχέση με τα τεχνητά νευρωνικά δίκτυα και την μέθοδο ηδονικής τιμολόγησης. Η συμπερίληψη της γειτονιάς ως μεταβλητή φαίνεται να βελτιώνει κατά σημαντικό ποσοστό τα αποτελέσματα, ακόμα και αν η περιοχή έχει ήδη περιληφθεί. Παρόλο που τα τεχνητά νευρωνικά μπορούν να εφαρμοστούν σε μεγάλα σύνολα δεδομένων, η ακρίβεια των αποτελεσμάτων που μας δίνουν δεν είναι απαραίτητα υψηλότερη από τις απλές μεθόδους μηχανικής μάθησης, όπως την παλινδρόμηση, η οποία κατά κύριο λόγο χρησιμοποιείται σε εφαρμογές τέτοιου είδους.

Το 2016 οι Wan Teng Lim, Lipo Wang, Yaoli Wang, και Qing Chang εφάρμοσαν δύο αλγορίθμους στην προσπάθειά τους να προβλέψουν τις τιμές των ακινήτων της Σιγκαπούρης και να συγκρίνουν την απόδοση αυτών. [3]Ο πρώτος αλγόριθμος στηρίχθηκε σε ένα μοντέλο τεχνητού νευρωνικού δικτύου (Artificial Neural Network), με χρήση της εξειδικευμένης τεχνικής autoregressive integrated moving average (ARIMA) και ο δεύτερος σε τεχνητό νευρωνικό δίκτυο κατασκευασμένο με βάση την κλασική μέθοδο πολλαπλής γραμμικής παλινδρόμησης. Στόχος των παραπάνω, αποτέλεσε η σύγκριση της απόδοσης της ικανότητας των μεθόδων αυτών να προβλέψουν τις τιμές των ακινήτων της Σιγκαπούρης. Το βέλτιστο από αυτά τα μοντέλα θα χρησιμοποιηθεί για την πρόγνωση των μελλοντικών αυτών τιμών. Τα αποτελέσματα της μελέτης, δείχνουν την υπεροχή του μοντέλου της απλής γραμμικής παλινδρόμησης, έναντι της μεθόδου ARIMA στην πρόβλεψη του δείκτη των τιμών. Παρόλα αυτά, περιορισμοί των δεδομένων δεν επιτρέπουν την εξαγωγή ενός στερεού συμπεράσματος και στο σημείο αυτό επισημαίνεται η αναγκαιότητα μελλοντικών ερευνών.

Το 2019 οι Tiancheng Cai, Kevin Han και Han Wu, με τη χρήση των δεδομένων που υπάρχουν στην πλατφόρμα Inside Airbnb, έχουν στόχο την μελέτη και έρευνα ενός μοντέλου πρόβλεψης τιμών των καταλυμάτων Airbnb στην Μελβούρνη της Αυστραλίας.[4] Η έρευνά τους περιλαμβάνει την μελέτη και σύγκριση διάφορων μοντέλων πρόβλεψης τιμής, με βάση τα νευρωνικά δίκτυα, καθώς επίσης και παραδοσιακών μεθόδων μηχανικής μάθησης, όπως είναι οι τεχνικές παλινδρόμησης, τα τυχαία δάση και η ενίσχυση κλίσης. Πιο συγκεκριμένα, οι μέθοδοι που χρησιμοποίησαν ήταν η γραμμική παλινδρόμηση, η παλινδρόμηση Ridge, τα τυχαία δάση, η ενίσχυση κλίσης, το Support Vector Machine και τέσσερα διαφορετικά νευρωνικά δίκτυα. Υπολογίζοντας το Μέσο Τετραγωνικό Σφάλμα και τον συντελεστή προσδιορισμού R^2, αξιολογούν τα παραπάνω μοντέλα και διαπιστώνουν ότι καλύτερη απόδοση έχει η μέθοδος παλινδρόμησης με ενίσχυση κλίσης, ενώ αμέσως μετά έρχεται η μέθοδος των τυχαίων δασών, η οποία ενδεχομένως να είχε βελτιωμένη απόδοση με αυστηρότερη επιλογή χαρακτηριστικών.

1.3 Πεδίο Έρευνας και Στόχοι

Όπως αναλύθηκε παραπάνω, η μεταβλητότητα της τιμής των ακινήτων, η οποία διαμορφώνεται από την επίσης μεταβλητή αξίας τους, έχει ως απόρροια την ανάγκη της δημιουργίας ενός μοντέλου πρόγνωσης αυτής. Το μοντέλο αυτό, επιθυμούμε, λαμβάνοντας υπόψιν τα χαρακτηριστικά του κάθε ακινήτου να προβλέπει, όσο το δυνατόν ακριβέστερα,

την τιμή του. Στο πεδίο έρευνας που έχουμε εστιάσει, ο κλάδος της επιστήμης των οικονομικών, ο οποίος ονομάζεται οικονομετρία, σε συνδυασμό με τις μεθοδολογίες προβλέψεων της μηχανικής μάθησης, χρησιμοποιούνται με σκοπό την κατασκευή ενός τέτοιου μοντέλου. Συλλέγοντας δεδομένα που αφορούν την αγορά των ακινήτων στην περιοχή της Αττικής, στόχος μας είναι η κατασκευή ενός μοντέλου, το οποίο βασιζόμενο στην οικονομετρία και με την χρήση τεχνικών μηχανικής μάθησης, θα προβλέπει με όσο το δυνατόν μεγαλύτερη ακρίβεια τις τιμές των ακινήτων τα οποία παρατίθενται προς πώληση. Στην προσέγγισή μας, εστιάσαμε στη χρήση διαφόρων τεχνικών παλινδρόμησης και της μεθόδου των τυχαίων δασών, δοκιμάστηκε ωστόσο και η απόδοση ενός νευρωνικού δικτύου και μίας μηχανής διανυσμάτων υποστήριξης. Τα αποτελέσματα που μας έδωσαν τα τελευταία δεν ήταν καθόλου ικανοποιητικά, εξαιτίας του μικρού αριθμού των δεδομένων μας και της απλότητας των μεταξύ τους σχέσεων, και για το λόγο αυτό στην παρούσα εργασία θα επικεντρωθούμε στην ανάλυση και παρουσίαση των αποτελεσμάτων των υπόλοιπων μεθόδων, οι οποίες αναλύονται στο επόμενο κεφάλαιο.

Τα δεδομένα μας αφορούν ακίνητα, τα οποία έχουν δημοσιευτεί στην ιστοσελίδα της Χρυσής Ευκαιρίας, την περίοδο 2019-2020 στον νομό της Αττικής. Ως επιμέρους στόχος που προέκυψε μέσα από την πορεία της εργασίας, τοποθετείται η συλλογή, ο καθαρισμός και η ανάλυση των δεδομένων των ακινήτων, καθώς το παραπάνω αποτελεί αναγκαίο βήμα στην προσπάθεια μας. Επομένως, μετά την διερευνητική ανάλυση των συλλεγμένων δεδομένων, κατασκευάζεται το σύνολο δεδομένων, που ως είσοδος στα μοντέλα μηχανικής μάθησης θα μας δώσει, ως αποτέλεσμα, την πρόβλεψη της τιμής κάθε ακινήτου.

1.4 Δομή Διπλωματικής Εργασίας

Στη συνέχεια περιγράφεται κάθε βήμα της πορείας προς την επίτευξη του παραπάνω σκοπού και πώς αυτό σκιαγραφείται σε κάθε ενότητα του παρόντος κειμένου.

Στο Κεφάλαιο 2, περιγράφεται το θεωρητικό υπόβαθρο, το οποίο αποτέλεσε τη βάση της μελέτης μας, μαζί με τις αντίστοιχες αναφορές στα πρωτότυπα κείμενα. Χωρίς αυτό, η κατανόηση της εργασίας δεν θα μπορούσε να είναι πλήρης. Γίνεται εδώ, εκτενής αναφορά στις μεθόδους τις οποίες χρησιμοποιήσαμε στην προσέγγιση της λύσης του προβλήματος. Αρχικά, παρουσιάζονται οι βασικές αρχές της οικονομετρίας και η σημασία της στην επίλυση ζητημάτων παρόμοιας φύσεως με αυτό που καταπιάνεται η παρούσα διατριβή. Στη συνέχεια, αναλύονται οι κύριοι πυλώνες της επιστήμης των δεδομένων στους οποίους στηριχθήκαμε. Έπειτα, ακολουθεί η εισαγωγή στην επιστήμη της μηχανικής μάθησης και η αναλυτικότερη περιγραφή των μεθόδων που χρησιμοποιήθηκαν στα πειράματά μας.

Στο Κεφάλαιο 3, γίνεται ανάλυση της μεθοδολογίας που ακολουθήθηκε για τον σχεδιασμό και την υλοποίηση του στόχου αυτής της εργασίας. Αποτυπώνεται λεπτομερώς, η διαδικασία της συλλογής, διερευνητικής ανάλυσης και επεξεργασίας των δεδομένων, καίριου βήματος προκειμένου να επιτευχθεί εμπεριστατωμένη έρευνα. Αναλύονται, κατ' αυτόν τον τρόπο τα επιμέρους βήματα που ακολουθήθηκαν και παρουσιάζονται τα προβλήματα που αντιμετωπίσαμε στην προσπάθεια της διαμόρφωσης του τελικού συνόλου δεδομένων, και σκιαγραφείται ο τρόπος επίλυσης καθενός εξ αυτών.

Στο Κεφάλαιο 4, παρουσιάζονται τα τελικά αποτελέσματα των αλγορίθμων μηχανικής μάθησης που εφαρμόσθηκαν και αξιολογείται η επίδοσή τους. Ακόμα, αναλύεται η επίδοση κάθε μία από τις μεθόδους που υλοποιήθηκαν και ερμηνεύονται τα αποτελέσματα κάθε διαφορετικού υποσυνόλου των δεδομένων, το οποίο χρησιμοποιήθηκε ως είσοδος σε κάθε αλγόριθμο, καθώς επίσης και η σημασία κάθε χαρακτηριστικού. Εξηγούνται σε βάθος οι λόγοι για τους οποίους οι αλγόριθμοι είχαν τις συγκεκριμένες επιδόσεις. Τέλος, ακολουθεί

μία σύνοψη όσων διαπιστώθηκαν, με στόχο την διευκόλυνση της μελλοντικής επιστημονικής έρευνας.

Στο Κεφάλαιο 5, πραγματοποιείται μία σύνοψη των συμπερασμάτων στα οποία καταλήξαμε κατά την πορεία προς την επίτευξη του στόχου μας και παρουσιάζονται κάποιες προτάσεις για περαιτέρω εξέλιξη της παρούσας μελέτης.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Επισκόπηση επιστημονικών πεδίων

Στο παρόν κεφάλαιο θα αναλυθεί το θεωρητικό υπόβαθρο, στο οποίο στηριχθήκαμε κατά τη συγγραφή της διπλωματικής εργασίας. Πιο συγκεκριμένα θα αναπτυχθούν οι βασικές αρχές της επιστήμης της οικονομετρίας, της επιστήμης των δεδομένων και της μηχανικής μάθησης. Ο συνδυασμός των τριών παραπάνω επιστημονικών πεδίων μας επέτρεψε να αναπτύξουμε τις μεθοδολογίες που αναλύονται στο επόμενο κεφάλαιο.

Η οικονομετρία και η μηχανική μάθηση έχουν χρησιμοποιηθεί στο παρελθόν για να δώσουν απάντηση σε προβλήματα παρόμοιας φύσεως, κάθε μία από τις επιστήμες αυτές, ωστόσο, προσεγγίζει τα ζητούμενα από διαφορετική σκοπιά. Οι παραδοσιακές τεχνικές της οικονομετρίας επικεντρώνονται γύρω από τον προσδιορισμό των παραμέτρων, προκειμένου να μοντελοποιηθεί βέλτιστα η σχέση μεταξύ των ανεξάρτητων και εξαρτημένων μεταβλητών. Από την άλλη, η μηχανική μάθηση, εστιάζει στο πρόβλημα της πρόβλεψης, στην παραγωγή, δηλαδή, προβλέψεων για τις εξαρτημένες μεταβλητές από τις ανεξάρτητες. Πιο επεξηγηματικά, ενώ οι τεχνικές της μηχανικής μάθησης στοχεύουν στην πρόβλεψη της ενδιαφερόμενης ποσότητας, η οικονομετρία μελετάει και εντοπίζει τα αίτια και τους παράγοντες, που οδηγούν σε κάποιο αποτέλεσμα, καθώς επίσης και τις σχέσεις μεταξύ τους. [5] Ουσιαστικά, και οι δύο επιστήμες στοχεύουν στην γνώση των πραγματικών υποκείμενων σχέσεων, αλλά όπως ήδη αναφέρθηκε, η προγνωστική απόδοση αποτελεί κύριο δείκτη, ο οποίος χρησιμοποιείται στην αξιολόγηση εφαρμογών μηχανικής μάθησης, ενώ ο όσο το δυνατόν ακριβέστερος προσδιορισμός και η επεξήγηση των αιτωδών σχέσεων, αποτελεί κύριο στόχο της οικονομετρίας. Οι δύο αυτές επιστήμες αλληλοενισχύονται, καθώς η μία ευνοείται από την άλλη.

Αρχικά, η ανάπτυξη και η πρόσβαση σε μεγάλους όγκους δεδομένων, προσφέρει νέες μεθόδους και πηγές πληροφοριών στην οικονομετρία, όπως ο χωρισμός των δεδομένων σε σύνολα εκπαίδευσης και ελέγχου, οι τεχνικές bagging και boosting, η επιλογή μεταβλητών, τα εργαλεία διαχείρισης των δεδομένων κ.α. Στην μηχανική μάθηση η οικονομετρία προσφέρει την ανάπτυξη της αιτιώδους μοντελοποίησης, λαμβάνοντας υπόψιν τον ανθρώπινο παράγοντα και δίνοντας, έτσι, την δυνατότητα για την βαθύτερη κατανόηση των σχέσεων μεταξύ των μεταβλητών. [6] Φυσικά, χωρίς την συμβολή της επιστήμης των δεδομένων, δεν θα είχαμε στην διάθεσή μας τα δεδομένα με βάση τα οποία ορίζονται, ουσιαστικά, τα διάφορα ζητήματα. Διερευνώντας και αναλύοντας αυτά, μας δίνεται στη συνέχεια η δυνατότητα της βέλτιστης ερμηνείας τους. Με τον τρόπο αυτό, οι μελετητές δύνανται να μοντελοποιήσουν τις σχέσεις των δεδομένων, ευνοούμενοι από την ανάπτυξη της οικονομετρίας, της μηχανικής μάθησης, της επιστήμης των δεδομένων και της διαρκούς εξέλιξης αυτών. Στη συνέχεια της παρούσας διπλωματικής θα αναπτυχθούν λεπτομερέστερα οι τρεις αυτοί επιστημονικοί κλάδοι και τα πεδία αυτών, στα οποία βασιστήκαμε.

2.2 Οικονομετρία

2.2.1 Βασικές Αρχές της Οικονομετρίας

Η επιστήμη της Οικονομετρίας, βασίζεται στην ανάπτυξη και εφαρμογή στατιστικών μεθόδων, προκειμένου να εκτιμηθούν οικονομικές σχέσεις, να ελεγχθούν οικονομικές θεωρίες και να υλοποιηθούν κυβερνητικές και επιχειρηματικές πολιτικές. Με την βοήθεια της οικονομετρίας υπολογίζονται σημαντικές μακροοικονομικές μεταβλητές, όπως τα επιτόκια, τα ποσοστά πληθωρισμού και το ακαθάριστο εγχώριο προϊόν. Παρά την ευρέως χρήση τους στην πρόβλεψη οικονομικών δεικτών, οι μέθοδοι της οικονομετρίας είναι ιδιαίτερα χρήσιμες και σε οικονομικές προβλέψεις, οι οποίες δεν έχουν να κάνουν με την μακροοικονομία. Η εξέλιξη της οικονομετρίας συνδέεται σε μεγάλο βαθμό με την εξέλιξη της μαθηματικής στατιστικής, εφόσον η πρώτη στηρίζεται στην συλλογή και ανάλυση μη πειραματικών οικονομικών δεδομένων, τα οποία με βάση τη στατιστική αναλύονται και εξάγονται συμπεράσματα. Στις κοινωνικές επιστήμες είναι πολύ δύσκολο και μερικές φορές μη ηθικό, να πραγματοποιηθούν ελεγχόμενα πειράματα, με στόχο την συλλογή δεδομένων, για αυτό και το μεγαλύτερο μέρος της οικονομικής ανάλυσης στηρίζεται σε πραγματικά δεδομένα, τα οποία όπως είναι φυσικό, χρειάζονται περισσότερη προετοιμασία και ανάλυση προκειμένου να είναι έτοιμα για μοντελοποίηση[7]. Για την μοντελοποίηση των δεδομένων χρησιμοποιείται ευρέως η μέθοδος της πολλαπλής παλινδρόμησης (Multiple Regression) και στην επιστήμη της οικονομετρίας αλλά και στην στατιστική, παρόλο που ο τρόπος με τον οποίο εφαρμόζεται σε κάθε περίπτωση μπορεί να διαφέρει σημαντικά.

Στην εφαρμοσμένη οικονομική, οι οικονομετρικές διαδικασίες κατέχουν ρόλο εξέχουσας σημασίας, καθώς με τη χρήση τους ελέγχεται η ορθότητα των οικονομικών σχέσεων και θεωριών. Η εμπειρική ανάλυση χρησιμοποιεί δεδομένα για να εξετάσει θεωρίες ή να εκτιμήσει οικονομικές σχέσεις, μέσω μαθηματικών μοντέλων. Για την κατασκευή του οικονομικού μοντέλου ακολουθούνται κάποια βασικά βήματα. Αρχικά, σύμφωνα με τον Jeffrey M Wooldridge[8], το πρώτο βήμα είναι η ακριβής και σαφής διατύπωση της θεωρίας την οποία θέλουμε να στηρίξει το μοντέλο μας. Στη συνέχεια κατασκευάζεται το οικονομικό μοντέλο, το οποίο αποτελείται από μαθηματικές εξισώσεις οι οποίες περιγράφουν τις σχέσεις μεταξύ των μεταβλητών, και με βάση αυτό δημιουργείται εν τέλει το οικονομετρικό μοντέλο. Ουσιαστικά η οικονομετρία, αφορά την εξαγωγή νέων μεταβλητών και σχέσεων μεταξύ αυτών, προκειμένου να υπολογιστεί μία νέα επιθυμητή ποσότητα. Η σχέση που εκφράζει ένα τέτοιο μοντέλο είναι συνήθως της μορφής: $y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n + \varepsilon$, όπου y είναι η εξαρτημένη μεταβλητή που εκφράζει την ποσότητα που θέλουμε να υπολογίσουμε, τα x_i ισοδυναμούν με τις ανεξάρτητες μεταβλητές από τις οποίες εξαρτάται το y, οι σταθερές b_i αποτελούν τις παραμέτρους του οικονομετρικού μοντέλου και περιγράφουν το είδος και την ισχύ της εξάρτησης της μεταβλητής y από την αντίστοιχη μεταβλητή x_i. Ακόμα, ο όρος ε περιέχει παράγοντες τους οποίους ενδέχεται να μην έχουμε παρατηρήσει και αποτελεί το σφάλμα της παραπάνω σχέσης. Στη συνέχεια, είναι απαραίτητο να γίνει εξόρυξη και συλλογή των δεδομένων που αφορούν τις ανεξάρτητες μεταβλητές του μοντέλου, έτσι ώστε μετά την εφαρμογή των οικονομετρικών μεθόδων να υπολογιστούν οι παράμετροι και να εξετασθούν οι αρχικές υποθέσεις. Αυτό είναι το μοντέλο που χρησιμοποιείται σαν βάση, στις περισσότερες εκ των περιπτώσεων της εφαρμοσμένης οικονομετρικής ανάλυσης.

Το προαναφερθέν μοντέλο αποτελείται από ντετερμινιστικές σχέσεις μεταξύ των ανεξάρτητων και εξαρτημένων μεταβλητών, από τις οποίες καθορίζονται και προσδιορίζονται τις περισσότερες φορές οι τρόποι με βάση τους οποίους μεταβάλλονται οι

εξαρτημένες μεταβλητές, όταν μεταβληθούν οι ανεξάρτητες. Φυσικά, το μοντέλο αυτό αποτελεί μια απλοποίηση της πραγματικότητας και των ουσιαστικών παραγόντων που συμμετέχουν στον προσδιορισμό της ποσότητας της οποίας την τιμή προσπαθούμε να προσδιορίσουμε, καθώς κανένα μοντέλο δεν θα μπορούσε να περιλαμβάνει όλες τις τυχαίες πτυχές της πραγματικής ζωής. Η τυχαία ανθρώπινη συμπεριφορά, όπως αναφέρει και ο William H Green [9], είναι κάτι που δύσκολα μοντελοποιείται και κατά συνέπεια οι παρατηρήσεις μιας εξαρτημένης μεταβλητής θα παρουσιάζουν αλλαγές που δεν είναι δυνατόν να προσδιοριστούν λεπτομερώς. Συνεπώς, η εισαγωγή στο μοντέλο μιας μεταβλητής, που εκφράζει την τυχαία διαταραχή, είναι απαραίτητη, καθώς επίσης απαραίτητος είναι και ο έλεγχος των αποτελεσμάτων της μελέτης, προκειμένου να διαπιστωθεί, ότι ο υποτιθέμενος τυχαίος παράγοντας είναι πραγματικά αυτό και όχι κάποιο δικό μας λάθος στην διαμόρφωση του εκάστοτε μοντέλου.

2.2.2 Γραμμική Παλινδρόμηση

Όπως αναφέρθηκε παραπάνω, η απλή γραμμική παλινδρόμηση είναι η πιο ευρέως χρησιμοποιούμενη τεχνική της εφαρμοσμένης οικονομετρίας. Το μοντέλο της χρησιμοποιείται με σκοπό να επιτευχθούν ποσοτικές εκτιμήσεις οικονομικών σχέσεων, οι οποίες είχαν διατυπωθεί έως τότε μόνο θεωρητικά. Η παλινδρόμηση είναι μια στατιστική τεχνική η οποία προσπαθεί να εξηγήσει τον τρόπο με τον οποίο μεταβάλλεται Μία απλή εξίσωση, που ορίζει την απλή γραμμική παλινδρόμηση και συνδέει δύο μεταβλητές είναι: $y = b_0 + b_1 x + u$ (2.1). Το y ονομάζεται εξαρτημένη μεταβλητή, το x ανεξάρτητη, ο όρος u αποτελεί το σφάλμα ή τη διαταραχή. Αν αυτό είναι σταθερό, δηλαδή $\Delta u = 0$, τότε η μεταβολή της μεταβλητής y εξαρτάται γραμμικά από εκείνη της μεταβλητής x:

$$\Delta y = b_1 \Delta x \; if \; \Delta u = 0.$$

Αυτό σημαίνει ότι η παράμετρος b_1 είναι η παράμετρος της κλίσης της σχέσης μεταξύ y και x και έχει τραβήξει το ενδιαφέρον στην εφαρμοσμένη οικονομετρία. Η παράμετρος b_0, είναι ο σταθερός όρος. Η γραμμικότητα της σχέσης υποδεικνύει ότι αλλαγή μίας μονάδας του x θα έχει τα ίδια αποτελέσματα στην μεταβλητή y. Αυτή η σχέση αποτελεί μία βάση, αλλά προφανώς δεν είναι ρεαλιστική η εφαρμογή της σε πιο σύνθετες εφαρμογές. Η εξίσωση 2.1 μπορεί να θεωρηθεί ότι αποτελείται από δύο συνιστώσες, την ντετερμινιστική συνιστώσα $b_0 + b_1 x$ και την στοχαστική u. Υποθέτουμε ότι το σφάλμα u έχει μηδενική μέση τιμή, δηλαδή $E(u) = 0$, προκειμένου να μπορούμε να εξάγουμε αξιόπιστα συμπεράσματα για τη σχέση μεταξύ ανεξάρτητης και εξαρτημένης μεταβλητής.[8] Μία επίσης βασική υπόθεση είναι ότι η μέση τιμή του u δεν εξαρτάται από την μεταβλητή x: $E(u|x) = E(u)$ ή $E(u|x) = 0$ (2.2). Έτσι με βάση την εξίσωση 2.2, το ντετερμινιστικό κομμάτι της εξίσωσης 2.1 μπορεί να γραφτεί: $E(y|x) = b_0 + b_1 X$ (2.3), που δηλώνει ότι η μέση τιμή του y δεδομένου του x, είναι μία γραμμική συνάρτηση της ανεξάρτητης μεταβλητής. Βέβαια η τιμή του y η οποία παρατηρείται στον πραγματικό κόσμο, είναι μάλλον απίθανο να ισούται με την 2.3 και κατά συνέπεια η εισαγωγή ενός όρου σφάλματος κρίνεται απαραίτητη [3]. Άρα:

$$y = E(y|x) + \varepsilon = b_0 + b_1 X + \varepsilon. \quad (2.4)$$

2.2.3 Μέθοδος Ελαχίστων Τετραγώνων

Οι παράμετροι b της εξίσωσης 2.1 είναι αυτοί που επιθυμούμε να υπολογίσουμε. Η πιο διαδεδομένη μέθοδος που χρησιμοποιείται για αυτόν τον υπολογισμό είναι η μέθοδος των ελαχίστων τετραγώνων, η οποία θα αναλυθεί στην συνέχεια [8]. Έστω ένα δείγμα του πληθυσμού $\{(x_i, y_i), i = 1, \dots, n\}$, όπου n τυχαίο μέγεθος πληθυσμού. Εφόσον αυτά τα δείγματα υποθέτουμε ότι υπακούν στην 2.1 μπορούμε να γράψουμε $y_i = b_0 + b_1 x_i + u_i$, για κάθε i. Η μέθοδος των ελαχίστων τετραγώνων υπολογίζει τις παραμέτρους b_i, έτσι ώστε να ελαχιστοποιείται το άθροισμα των τετραγωνισμένων υπολοίπων, $\sum_{i=1}^{n} u_i^2$ $(i = 1, \dots, n)$. Γνωρίζουμε ότι $E(u) = 0$. Η σχέση αυτή γράφεται:

$$E(y - b_0 + b_1 x) = 0. \text{ (2.5)}$$

Εφόσον $E(u) = 0$, συμπεραίνεται ότι $Cov(x, u) = E(xu) = 0$ (2.6) και από τις (2.5) και (2.6) καταλήγουμε στην:

$$E(x(y - b_0 + b_1 x)) = 0. \text{ (2.7)}$$

Δοσμένου του δείγματος δεδομένων επιλέγονται εκτιμήτριες παράμετροι $\widehat{b_0}$ και $\widehat{b_1}$ και οι παραπάνω εξισώσεις μπορούν να μετασχηματιστούν στις:

$$n^{-1} \sum_{i=1}^{n} \left(y_i - \widehat{b_0} + \widehat{b_1} x_{\,i}\right) = 0 \text{ (2.8)}$$

και

$$n^{-1} \sum_{i=1}^{n} \left(y_i - \widehat{b_0} + \widehat{b_1} x_{\,i}\right) = 0. \text{ (2.9)}$$

Με τη χρήση των ιδιοτήτων του αθροίσματος από τη σχέση (2.8) καταλήγουμε στην σχέση:

$$\bar{y} = \widehat{b_0} + \widehat{b_1} \bar{x}, \qquad (2.10)$$

όπου $\bar{y} = \sum_{i=1}^{n} y_i$ είναι ο δειγματικός μέσος του y_i και αντίστοιχα $\bar{x} = \sum_{i=1}^{n} x_i$. Ο δειγματικός μέσος του, x. Η εξίσωση (2.10) γράφεται:

$$\widehat{b_0} = \bar{y} - \widehat{b_1} \bar{x}. \text{ (2.11)}$$

Από την (2.9) και την (2.11) έχω:

$$\sum_{i=1}^{n} x_i \left[y_i - \left(\bar{y} - \widehat{b_1} \bar{x}\right) - \widehat{b_1} x_i\right] = 0 \Rightarrow \sum_{i=1}^{n} x_i (y_i - \bar{y}) = \widehat{b_1} \sum_{i=1}^{n} x_i (x_i - \bar{x}).$$

Βάσει βασικών ιδιοτήτων του αθροιστή έχω:

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \ \kappa\alpha\iota \ \sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Συνεπώς:

$$\widehat{b_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \ \varepsilon\varphi\acute{o}\sigma o\nu \sum_{i=1}^{n} (x_i - \bar{x})^2 > 0. \ (2.12)$$

Έχοντας τελικά υπολογίσει το $\widehat{b_1}$ μπορούμε μέσω της (2.11) να υπολογίσουμε και το $\widehat{b_0}$. Η εξίσωση (2.12) αντιστοιχεί στην δειγματική συνδιακύμανση των x και y διαιρεμένη με την δειγματική διακύμανση του x. Τα $\widehat{b_1}$ και $\widehat{b_0}$ ονομάζονται εκτιμήτριες ελαχίστων τετραγώνων. Η τιμή $\widehat{y_\iota} = \widehat{b_0} + \widehat{b_1} x_\iota$, είναι η τιμή του y την οποία θα προβλέψουμε για $x = x_i$, και με τη δοσμένη κλίση. Το υπόλοιπο για την παρατήρηση i είναι η διαφορά μεταξύ του πραγματικού y_i και της τιμής που προβλέψαμε:

$$\widehat{u_\iota} = y_i - \widehat{y_\iota} = y_i - \widehat{b_0} - \widehat{b_1} x_i. (2.13)$$

Τα $\widehat{b_1}$ και $\widehat{b_0}$ τελικά υπολογίζονται έτσι ώστε το άθροισμα των τετραγώνων των ποσοτήτων της εξίσωσης (2.13) να είναι ελάχιστο.

2.3 Μοντέλο Ηδονικής Τιμολόγησης

Η θεωρία ηδονικής τιμολόγησης υποθέτει ότι ένα εμπόρευμα ή αγαθό, όπως για παράδειγμα ένα σπίτι, μπορεί να θεωρηθεί ότι συνίσταται από το άθροισμα επιμέρους μεμονωμένων στοιχείων ή χαρακτηριστικών. Οι καταναλωτές θεωρείται ότι αγοράζουν αγαθά τα οποία εκπροσωπούν δομές από χαρακτηριστικά, τα οποία μεγιστοποιούν την αξία τους [10]. Σύμφωνα με τον Triplett (1986), μέθοδοι που πλησίαζαν κατά πολύ το μοντέλο ηδονικής τιμολόγησης, χρησιμοποιούνταν σε δείκτες τιμών αρκετό καιρό πριν γίνει πλήρως κατανοητό το εννοιολογικό τους πλαίσιο. Ο όρος «ηδονικός» στο οικονομικό πλαίσιο χρησιμοποιείται για να περιγράψει την ευχαρίστηση που προκαλείται στον άνθρωπο από την κατανάλωση αγαθών.

Το μοντέλο ηδονικής τιμολόγησης, όπως γίνεται αντιληπτό από τους Tung-Leong Chin και K W Chau [11], δηλώνει πως η τα αγαθά αποτελούνται από ποικίλα γνωρίσματα τα οποία συνδυάζονται για να παράγουν δομές χαρακτηριστικών, που υπολήπτεται ο καταναλωτής. Για παράδειγμα, όπως και στην περίπτωση που θα μελετηθεί, η αξία ενός ακινήτου εξαρτάται από πολλά χαρακτηριστικά, όπως το μέγεθός του, η ηλικία του και το είδος του. Η εκτίμηση της τιμής του ακινήτου αφορά τον προσδιορισμό, ποσοτικό και ποιοτικό, και την ανάλυση αυτών των γνωρισμάτων. Συνήθως αυτή η εκτίμηση γίνεται από κάποιον ειδικό. Η δυσκολία, ωστόσο, έγκειται στο γεγονός ότι η αξία του κάθε γνωρίσματος ενδέχεται να είναι υποκειμενική, σε ατομικό επίπεδο αλλά και το κάθε γνώρισμα να έχει διαφορετική σημασία από ακίνητο σε ακίνητο. Σύμφωνα με τους Neil Dunse και Colin Jones [12], με τη χρήση τεχνικών παλινδρόμησης, καθίσταται ευκολότερη η εφαρμογή του μοντέλου ηδονικής τιμολόγησης και κατά συνέπεια ο υπολογισμός της αξίας του αγαθού, καθώς μέσω αυτής προσδιορίζονται και ποσοτικοποιούνται οι πιο καθοριστικοί στην αξία του παράγοντες.

Η θεωρία ηδονικής τιμολόγησης υλοποιείται με την συνάρτηση ηδονικής τιμολόγησης, η οποία είναι ουσιαστικά μία παλινδρόμηση, όπως αναφέρουν οι B W Brorsen, W R Grant

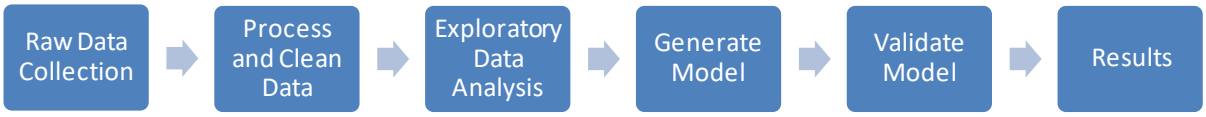
και M E Rister [13], της παρατηρούμενης τιμής ενός αγαθού συναρτήσει των γνωρισμάτων του. Η ηδονική τιμολόγηση αφορά, συνεπώς, την αξία του καθενός από τα επιμέρους χαρακτηριστικά χωριστά και άρα η τελική τιμή του προϊόντος αποτελεί σύνθεση των τιμών των χαρακτηριστικών του. Το μοντέλο αυτό έχει εφαρμοσθεί για την αποτίμηση της αξίας γεωργικών εμπορευμάτων, οικημάτων αλλά και αγαθών των οποίων η αξία καθορίζεται από διάφορους και πιο σύνθετους παράγοντες. Παρόλο που η χρήση του μοντέλου έχει ευρέως αναγνωριστεί, υπάρχουν και κάποιες δυσκολίες όσον αφορά το γεγονός ότι υποθέτει, πως τα χαρακτηριστικά των αγαθών είναι γραμμικώς ανεξάρτητα μεταξύ τους. Το τελευταίο είναι αρκετά ιδεαλιστικό, καθώς είναι πιο πιθανό στην αγορά να υπάρχουν ανισορροπίες και συσχετίσεις μεταξύ των δεδομένων. Στην παρούσα μελέτη, έχουμε στηριχθεί στο μοντέλο ηδονικής τιμολόγησης, προκειμένου να υλοποιήσουμε διαφορετικά είδη παλινδρόμησης και να καταλήξουμε σε συμπεράσματα αναφορικά με την αξία των ακινήτων αλλά και την σπουδαιότητα των επιμέρους χαρακτηριστικών τους.

2.4 Επιστήμη Ανάλυσης Δεδομένων

2.4.1 Ανάλυση Δεδομένων

Η Ανάλυση Δεδομένων ορίζεται ως η διαδικασία επιθεώρησης, καθαρισμού, μετατροπής και μοντελοποίησης των δεδομένων, που στόχο έχει την εξόρυξη πληροφοριών και συμπερασμάτων, προκειμένου να κατανοηθούν σε βάθος τα δεδομένα και η σημασία τους [14]. Στην εποχή που διανύουμε, η εύκολη πρόσβαση σε τεράστιο όγκο δεδομένων καθιστά πιο αναγκαία από ποτέ την βελτιστοποίηση της εξόρυξης, της αποθήκευσης, της ανάλυσης και της ερμηνείας τους, με την χρήση σύγχρονων τεχνολογικών εργαλείων και τεχνικών, που ολοένα και βελτιώνονται. Η επιστήμη των δεδομένων έχει πλέον εισχωρήσει σε κάθε πτυχή της ζωής του σύγχρονου ανθρώπου, με ποικίλες προσεγγίσεις και τεχνικές, από τους τομείς των επιχειρήσεων, των θετικών επιστημών, των κοινωνικών επιστημών, αλλά και τους τομείς της υγείας, διευκολύνοντας την ορθότερη λήψη σημαντικών αποφάσεων και την αποτελεσματικότερη λειτουργία των επιχειρήσεων και υπηρεσιών [15].

Τα βήματα που ακολουθούνται κατά την διαδικασία ανάλυσης των δεδομένων είναι, αρχικά, η συλλογή των δεδομένων τα οποία δύναται να προέρχονται από διάφορες πηγές. Τέτοιες πηγές δειγματοληψίας ενδέχεται να είναι η επικοινωνία ατόμων με οργανισμούς και εταιρείες που ειδικεύονται σε αυτούς τους τομείς, διαφόρων ειδών αισθητήρες που μπορεί να είναι τοποθετημένοι στο περιβάλλον, όπως για παράδειγμα κάμερες κυκλοφορίας, δορυφόροι, ή αισθητήρες άλλου είδους, συλλογές δεδομένων ιδιωτικών και μη επιχειρήσεων που σκοπό έχουν την βελτιστοποίηση της εξυπηρέτησης και ικανοποίησης του καταναλωτή. Στη συνέχεια τα δεδομένα αυτά υφίστανται επεξεργασία προκειμένου να προετοιμαστούν για την μοντελοποίηση τους, καθώς η ύπαρξη «θορυβώδους» πληροφορίας, όπως δεδομένα με σφάλματα, ακραίες ή κενές τιμές, αποτελούν εμπόδιο για την εξαγωγή συμπερασμάτων, καθώς οδηγούν σε εσφαλμένα συμπεράσματα και αποπροσανατολίζουν τους αλγορίθμους. Ακολούθως, εφαρμόζεται σε αυτά η διαδικασία της Διερευνητικής Ανάλυσης Δεδομένων η οποία αναλύεται στη συνέχεια. Τελικά, μέσω αλγορίθμων οι οποίοι θα εφαρμοστούν στα δεδομένα, λαμβάνουμε πληροφορίες, οι οποίες επιδρούν καθοριστικά στην εξαγωγή συμπερασμάτων και την λήψη αποφάσεων [16].



Εικόνα 1 Σχηματική αναπαράσταση της διαδικασίας Ανάλυσης Δεδομένων.

Υπάρχουν διάφορες τεχνικές ανάλυσης δεδομένων, ποσοτικές και ποιοτικές ανάλογα με τον τύπο τους. Η αριθμητική πληροφορία επιδέχεται μεθόδους ερμηνείας που προέρχονται από την επιστήμη της στατιστικής και η ανάλυση ονομάζεται στατιστική ανάλυση δεδομένων, ενώ η κατηγορηματική πληροφορία αναλύεται με μεθόδους αφηγηματικές και περιγραφικές [17]. Στην συνέχεια της παρούσας διπλωματικής εργασίας θα επικεντρωθούμε στην πρώτη μέθοδο, η οποία αφορά την εφαρμογή στατιστικών τεχνικών ως ερευνητική μεθοδολογία. Οι τεχνικές αυτές μας παρέχουν πολλά εργαλεία για την επίτευξη μιας πιο ενδελεχούς και εμπεριστατωμένης μελέτης και επεξεργασίας των δεδομένων. Μέσω των στατιστικών μοντέλων και τον προσδιορισμό των παραμέτρων αυτών, καταλήγουμε σε συμπεράσματα που συμβάλλουν στην βαθύτερη κατανόηση των δεδομένων που διαθέτουμε [18].

2.4.2 Διερευνητική Ανάλυση Δεδομένων

Τον όρο Διερευνητική Ανάλυση Δεδομένων (ΔΑΔ), όρισε ο John Tukey το 1961, ως οι «Διαδικασίες για την ανάλυση δεδομένων, τεχνικές ερμηνείας των αποτελεσμάτων τέτοιων διαδικασιών, τρόποι σχεδιασμού της συλλογής δεδομένων για να γίνει η ανάλυσή τους ευκολότερη ή ακριβέστερη και όλα τα μηχανήματα και τα μαθηματικά αποτελέσματα των στατιστικών στοιχείων που ισχύουν για την ανάλυσή τους». Ο Tukey (Exploratory Data Analysis, 1977) [14] τονίζει επίσης, πως η επιστήμη της ανάλυσης δεδομένων δεν υπάρχει επειδή είναι χρήσιμη, αλλά στόχος της είναι να επισημάνει στον ερευνητή σημαντικό αριθμό ποικίλων τεχνικών, προκειμένου να εξετάζονται και να κατανοούνται πιο αποτελεσματικά τα δεδομένα. Κατά τον ίδιο η ΔΑΔ είναι μια φιλοσοφία, μια κατάσταση στην οποία ο ερευνητής οφείλει να είναι ευέλικτος και διατεθειμένος να ψάξει για όσα πιστεύει ότι δεν είναι εκεί αλλά και για αυτά που πιστεύει ότι υπάρχουν. Η ΔΑΔ αποτελεί ένα από τα πρώτα βήματα που θα γίνουν στην προσπάθεια δημιουργίας του μαθηματικού μοντέλου και αποσκοπεί στην προσφορά γνώσης και στην ουσιαστική κατανόηση του προς επίλυση προβλήματος [19].

Με βάση τους Rachel Schutt & Cathy O'Neil [5], τα βασικά εργαλεία της ΔΑΔ είναι οι γραφικές παραστάσεις, τα γραφήματα και τα στοιχεία στατιστικής. Συνιστά μία μέθοδο κατά την οποία καλούμαστε να διερευνούμε συστηματικά τα δεδομένα σχεδιάζοντας τις κατανομές των μεταβλητών, μετασχηματίζοντας αυτές και εξάγοντας συμπεράσματα για τις σχέσεις μεταξύ τους. Ο μελετητής αναλύει μεγάλους όγκους δεδομένων αναζητώντας μοτίβα, πρότυπα, σύνθετες σχέσεις μεταξύ αυτών, αδύνατον να διαπιστωθούν με άλλον τρόπο. Η σημερινή εξέλιξη και πορεία της τεχνολογίας με τις δυνατότητες που μας παρέχει, αποτελεί καθοριστικό παράγοντα για την ανάλυση δεδομένων [20]. Η ΔΑΔ εφαρμόζεται αμέσως μετά την συλλογή των δεδομένων, και οι μέθοδοί της ποικίλουν ανάλογα με τον τύπο τους και τον στόχο της μελέτης. Πιο συγκεκριμένα, τα κατηγορηματικά δεδομένα αναλύονται με βάση τις τεχνικές της περιγραφικής στατιστικής, ενώ για την τα αριθμητικά δεδομένα, χρησιμοποιούνται γραφήματα γραμμής, ιστογράμματα, δισδιάστατα αλλά και

τρισδιάστατα γραφήματα διασποράς, χάρτες θερμότητας και γραφήματα ράβδων [21]. Τα αναφερθέντα συνεισφέρουν καθοριστικά στην απεικόνιση και την προετοιμασία των δεδομένων για μοντελοποίηση, καθώς μέσω αυτών, σε συνδυασμό και με την επιστήμη της στατιστικής, υπολογίζονται πρωτίστης σημασίας παράμετροι, χαρακτηριστικά και τιμές αριθμητικών περιγραφικών μέτρων, όπως είναι η λοξότητα, η διασπορά και η κύρτωση. Αφού τα παραπάνω εφαρμοστούν, θα έχουν προετοιμάσει κατάλληλα τα δεδομένα προκειμένου να μοντελοποιηθούν στη συνέχεια, με μεθόδους και αλγορίθμους μηχανικής μάθησης.

2.5 Μηχανική Μάθηση

2.5.1 Ορισμός και Βασικές Αρχές Μηχανικής Μάθησης

Η έννοια της Μάθησης, όπως και αυτή της νοημοσύνης, είναι εκ φύσεως δύσκολο να οριστεί, καθώς καλύπτει ένα ευρύ φάσμα διαδικασιών, πολλοί ορισμοί ωστόσο αναφέρουν φράσεις όπως «διαδικασία εμπέδωσης νέων δεδομένων, η απόκτηση γνώσεων ή δεξιοτήτων μέσω εμπειριών, οδηγιών ή μελέτης, η μετατροπή της συμπεριφοράς ενός υποκειμένου μέσω της εμπειρίας του». Η Μηχανική Μάθηση αφορά την διαδικασία της μάθησης, όπως αυτή λαμβάνει χώρα στον κόσμο των μηχανών. Έχοντας υπόψη την έννοια, την σημασία και τον τρόπο με τον οποίο πραγματοποιείται η μάθηση στο περιβάλλον των έμβιων οργανισμών, οι επιστήμονες χρησιμοποιούν τις γνώσεις αυτές, προκειμένου να αναπτύξουν μοντέλα και τεχνικές, τα οποία καθιστούν δυνατή και την μάθηση των μηχανών. Η επίτευξη αυτού συνιστά έναν μεγάλο και συναρπαστικό στόχο στον τομέα της Τεχνητής Νοημοσύνης [22].

Η Μηχανική Μάθηση αποτελεί πεδίο της Επιστήμης των Υπολογιστών και σκοπός της, όπως αναφέρει ο Phil Simon [23], είναι «να καταφέρουν οι υπολογιστές να αποκτήσουν την ικανότητα της μάθησης, χωρίς να έχουν ρητά προγραμματιστεί». Η έννοια μάθηση για τους υπολογιστές, αφορά την τροποποίηση ή προσαρμογή των ενεργειών ή αποκρίσεών τους, προκειμένου αυτές να γίνουν περισσότερο ακριβείς, με την έννοια του πόσο καλά προσεγγίζουν τα ορθά αποτελέσματα, αυτά των ενεργειών τους. Επιστήμες όπως η νευροεπιστήμη, η βιολογία, η στατιστική, τα μαθηματικά και η φυσική έχουν συμβάλλει στην πρόοδο της Μηχανικής Μάθησης και την εξέλιξη των μεθόδων της. Τα νευρωνικά δίκτυα και η εξόρυξη δεδομένων αποτελούν απόρροιες αυτής ακριβώς της εξέλιξης. Ένα άλλο στοιχείο των μεθόδων της μηχανικής μάθησης που αξίζει να αναφερθεί, είναι η υπολογιστική πολυπλοκότητά των αλγορίθμων της. Εφόσον αυτοί είναι που εφαρμόζονται σε πολύ μεγάλα σύνολα δεδομένων, χρειάζεται προσοχή προκειμένου να μην έχουν μεγάλο βαθμό πολυπλοκότητας πολυωνυμικού χρόνου στον αριθμό των δειγμάτων, καθώς κάτι τέτοιο αποτελεί πρόβλημα και δεν είναι επιθυμητό [24].

Υπάρχουν διάφορα είδη Μηχανικής Μάθησης στα οποία ταξινομούνται και οι αλγόριθμοι που χρησιμοποιεί. Τα βασικά από αυτά είναι η *επιβλεπόμενη μάθηση*, η *μη επιβλεπόμενη μάθηση* και η *ενισχυτική μάθηση*. Στην επιβλεπόμενη μάθηση έχουμε ένα σύνολο δεδομένων εισόδου και ένα σύνολο επιθυμητών τιμών εξόδου, οι οποίες αντιστοιχούν στα δεδομένα εισόδου. Αυτό που καθιστά αποτελεσματικούς τους αλγόριθμους μηχανικής μάθησης είναι η ικανότητα της γενίκευσης που διαθέτουν. Πιο συγκεκριμένα, ο αλγόριθμος έχει την δυνατότητα να παράγει λογικά αποτελέσματα για εισόδους στις οποίες δεν έχει εκπαιδευτεί κατά τη διάρκεια της μάθησης. Κάποια παραδείγματα προβλημάτων στα οποία χρησιμοποιείται είναι η ταξινόμηση και η πρόβλεψη. Στην μη επιβλεπόμενη μάθηση, διαθέτουμε ένα σύνολο δεδομένων εισόδου,

χωρίς το αντίστοιχο σύνολο δεδομένων εξόδου. Σκοπός τους αλγορίθμου είναι να καταφέρει να εντοπίσει την δομή των δεδομένων που του δόθηκαν ως είσοδο, χωρίς να διαθέτει την επιθυμητή έξοδο. Χρησιμοποιείται κυρίως για προβλήματα ομαδοποίησης και ανάλυσης συσχετισμών. Τέλος, στην ενισχυτική μάθηση ο αλγόριθμος διδάσκεται μία ακολουθία ενεργειών σε ένα περιβάλλον, συνήθως το περιβάλλον αυτό είναι διατυπωμένο σε μορφή στοχαστικής Διαδικασίας Απόφασης Markov (ΔΑΜ), καθώς πολλοί αλγόριθμοι χρησιμοποιούν δυναμικό προγραμματισμό και το περιβάλλον ΔΑΜ διευκολύνει τέτοιες διαδικασίες. Κύρια προβλήματα που επιδέχονται το συγκεκριμένο είδος μάθησης, είναι προβλήματα σχεδιασμού, σχετιζόμενα με την επιστήμη της ρομποτικής και του αυτομάτου ελέγχου, αλλά και προβλήματα βελτιστοποίησης, όπως η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρους [25]. Στη συνέχεια, για την μοντελοποίηση των δεδομένων και την εκπαίδευση των αλγορίθμων που εφαρμόσθηκαν, χρησιμοποιήθηκαν τεχνικές επιβλεπόμενης μάθησης και σε αυτήν θα επικεντρωθούμε.

Ένας αλγόριθμος επιβλεπόμενης μάθησης υλοποιεί μία συνάρτηση πρόγνωσης $f: X \rightarrow Y$, όπου X είναι ο χώρος εισόδου και Y ο επιθυμητός χώρος εξόδου, με $P(\boldsymbol{x}, y)$ μία (άγνωστη) από κοινού συνάρτηση κατανομής πιθανότητας στον χώρο $X \times Y$. Δεδομένου ενός δείγματος εκπαίδευσης $\{(x_i, y_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} P(\boldsymbol{x}, y)$, η επιβλεπόμενη μάθηση εκπαιδεύει την συνάρτηση f, σε μια οικογένεια συναρτήσεων F, έτσι ώστε η συνάρτηση $f(x)$, να προβλέπει την πραγματική έξοδο y για την αντίστοιχη είσοδο x, όπου $(\boldsymbol{x}, y) \overset{i.i.d.}{\sim} P(\boldsymbol{x}, y)$. Αν ο χώρος εξόδου είναι διακριτός, τότε μιλάμε για πρόβλημα ταξινόμησης, ενώ αν είναι συνεχής για πρόβλημα πρόβλεψης. Η συνάρτηση f επιλέγεται ως η συνάρτηση που ελαχιστοποιεί την ποσότητα $\mathrm{E}_{(x,y)} \sim P[c(\boldsymbol{x}, y, f^*(\boldsymbol{x}))]$, όπου $c(\cdot)$ μια συνάρτηση απώλειας, η οποία καθορίζει το κόστος μίας λανθασμένης πρόβλεψης, και είναι γνωστή ως σφάλμα Bayes (Bayes error). Ο στόχος του αλγορίθμου είναι να βρει μία f η οποία θα φτάσει όσο πιο κοντά στην f^* γίνεται [26].

Οι πιο διαδεδομένοι αλγόριθμοι επιβλεπόμενης μάθησης είναι οι γραμμικοί ταξινομητές, όπως είναι η γραμμική παλινδρόμηση, η λογιστική παλινδρόμηση, ο ταξινομητής Naïve Bayes, τα perceptrons και οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines), τα Δέντρα Αποφάσεων (Decision Trees), ο αλγόριθμος ομαδοποίησης k-Means, τα νευρωνικά δίκτυα, τα δίκτυα Bayes και οι αλγόριθμοι ενδυνάμωσης (boosting). Στη συνέχεια θα αναλυθούν οι αλγόριθμοι που χρησιμοποιήθηκαν για την κατασκευή του μοντέλου και την επίτευξη του στόχου της διπλωματικής.

2.5.2 Πολλαπλή Γραμμική Παλινδρόμηση

Στην ενότητα 2.1 είδαμε το μοντέλο της απλής γραμμικής παλινδρόμησης $y = b_0 + b_1 x + u$ και υπολογίσθηκαν οι παράμετροι b_0 και b_1 μέσω των εκτιμητριών των ελαχίστων τετραγώνων. Σε αυτήν την ενότητα θα επεκτείνουμε το παραπάνω για περισσότερες εκ της μία εξόδου, και επίσης περισσότερες εισόδους. Έστω τώρα ότι το μοντέλο δίνεται από την σχέση: $Y_i = X_i \beta + \varepsilon_i$, όπου το Y αποτελεί την μεταβλητή που επιθυμούμε να προβλέψουμε ως έξοδο και Y_ι είναι η τιμή της μεταβλητής στην i παρατήρηση. Το Y καλείται εξαρτημένη ή μεταβλητή απόκρισης και αποτελεί την έξοδο του μοντέλου μας. Υποθέτουμε ότι το X είναι ένα $1 \times n$ διάνυσμα, που αποτελείται από τις παρατηρούμενες, τυχαίες μεταβλητές εισόδου και X_i αποτελεί το διάνυσμα των εισόδων κατά την i παρατήρηση. Ακόμα, το β είναι ένα διάνυσμα παραμέτρων μεγέθους $n \times 1$, οι οποίες αρχικά είναι άγνωστες και στη συνέχεια εκτιμούνται με την εφαρμογή της μεθόδου των ελαχίστων τετραγώνων, με βάση τα δεδομένα. Κάθε μία από τις παραμέτρους εκφράζει την μεταβολή του Y, όταν μεταβάλλεται μόνο η συγκεκριμένη μεταβλητή X, στην οποία αναφέρεται το αντίστοιχο β.

Οι όροι ε_i είναι ανεξάρτητοι και όμοια κατανεμημένοι, με μηδενική μέση τιμή και διακύμανση σ^2, δηλαδή $\varepsilon_i \sim N(0, \sigma^2)$. Για παράδειγμα για την πρώτη παρατήρηση η εξίσωση της πολλαπλής γραμμικής παλινδρόμησης γράφεται: $Y_1 = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon_1$ [27].

Η εξίσωση της γραμμικής παλινδρόμησης ορίζει ευθεία γραμμή όταν σε αυτήν υπάρχει μόνο μία ανεξάρτητη μεταβλητή X, επίπεδο όταν συμμετέχουν δύο μεταβλητές X και υπερεπίπεδο όταν έχουμε περισσότερες εκ των δύο μεταβλητές X. Με βάση την μέθοδο των ελαχίστων τετραγώνων η οποία θα εφαρμοστεί για την εκτίμηση των παραμέτρων β, ελαχιστοποιείται το σφάλμα του αθροίσματος των σφαλμάτων, τα οποία αφορούν τις αποστάσεις των τιμών Y που υπολογίσθηκαν από το υπερεπίπεδο προσαρμογής. Η πολλαπλή γραμμική παλινδρόμηση υποθέτει γραμμική σχέση μεταξύ των ανεξάρτητων και εξαρτημένων μεταβλητών, όχι μεγάλη συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών, σταθερή διακύμανση των υπολειμμάτων, όπως αναφέρθηκε και παραπάνω, ανεξαρτησία των παρατηρήσεων, δηλαδή ανεξαρτησία των τιμών των σφαλμάτων και κανονική κατανομή των σφαλμάτων [28].

2.5.3 Παλινδρόμηση Κορυφογραμμής – Ridge Regression

Η τεχνικής της παλινδρόμησης Ridge χρησιμοποιείται για την ανάλυση δεδομένων τα οποία πάσχουν από πολυσυγγραμμικότητα. Η μέθοδος των ελαχίστων τετραγώνων, όταν υπάρχει πολυσυγγραμμικότητα, υπολογίζει αμερόληπτες εκτιμήτριες $\hat{\beta}$, αλλά με μεγάλη διακύμανση, γεγονός που μπορεί να προκαλέσει, εύκολα, σφάλματα στις τιμές τους. Η παλινδρόμηση Ridge μειώνει αυτά τα σφάλματα, περιορίζοντας τις παραμέτρους και προσθέτοντας έναν βαθμό κλίσης στον υπολογισμό τους. Αυτό έχει ως συνέπεια οι εκτιμήτριες να μην είναι πλέον αμερόληπτες, πετυχαίνεται ωστόσο μείωση της διακύμανσης και μετριασμό των σφαλμάτων που οφείλονταν σε αυτήν.

Έστω το μοντέλο γραμμικής παλινδρόμησης $Y = X\beta + \varepsilon$. Σε αυτό, η μέθοδος των ελαχίστων τετραγώνων εκτιμά τις παραμέτρους β ως $\hat{\beta}$, ελαχιστοποιώντας το άθροισμα των τετραγώνων των υπολειμμάτων και εν τέλει μας δίνει $\hat{\beta} = (X^T X)^{-1} X^T Y$. Αν η μέση τιμή των σφαλμάτων είναι μηδενική και η διακύμανση ίση με σ^2, τότε η διακύμανση των εκτιμητών είναι $Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$. Με την υπόθεση κανονικοποίησης για τα Y, λαμβάνουμε $\sigma^2 = 1$ και συνεπώς βρίσκουμε: $Var(\hat{\beta}) = \frac{1}{1 - X^T X^2}$. Θέτοντας $R = X^T X$, έχω $Var(\hat{\beta}) = \frac{1}{1 - R^2}$, όπου R είναι ο πίνακας συσχέτισης των ανεξάρτητων μεταβλητών [29].

Η Ridge παλινδρόμηση, προσθέτει έναν μικρό παράγοντα μεροληψίας k, όπου $0 \geq k \geq 1$ και οδηγεί στον επόμενο τύπο για τις εκτιμήτριες των συντελεστών της: $\widehat{\beta_k} = (R + kI)^{-1} X^T y$. Η εκτιμήτρια $\widehat{\beta_k}$, όπως είπαμε δεν είναι αμερόληπτη, ωστόσο με την προσθήκη της φαίνεται ότι αυξάνοντάς την, μειώνεται η διακύμανση, αυξάνεται ωστόσο και το bias: $E(\hat{\beta} - \beta) = [(X^T X + kI)^{-1} X^T X - I]\beta$ [30]. Υπάρχει γενικά μία βέλτιστη τιμή του $k > 0$ για κάθε πρόβλημα, όπως απέδειξε ο Hoerl (1970), παρόλα αυτά δεν έχει ακόμα αναπτυχθεί αναλυτική μέθοδος για την εύρεσή του. Ο παράγοντας αυτός καθορίζει πόσο έντονο θα είναι το κόστος που θα προστεθεί. Δοκιμάζονται συνήθως διάφορες τιμές του, με στόχο να οδηγήσουν σε λύση με μικρότερο μέσο τετραγωνικό σφάλμα των παραμέτρων από ότι η μέθοδος των ελαχίστων τετραγώνων [31]. Ουσιαστικά, για ένα μοντέλο το οποίο είχε υπερεκπαιδευτεί σε συγκεκριμένα δεδομένα, η πρόσθεση της μικρής κλίσης στην γραμμή μπορεί να μην ταιριάζει στα δεδομένα εκπαίδευσης τόσο καλά, αλλά η μείωση της διακύμανσης, που έχει ως αποτέλεσμα η προσθήκη αυτή, το καθιστά πιο αξιόπιστο σε βάθος χρόνου.

2.5.4 Lasso Παλινδρόμηση – Lasso Regression

Η Lasso (Least Absolute Shrinkage and Selection Operator) Παλινδρόμηση είναι μία μέθοδος ανάλυσης παλινδρόμησης η οποία ελαχιστοποιεί το άθροισμα των τετραγώνων των υπολειμμάτων, κάτω από τον περιορισμό της l^1 – νόρμας του διανύσματος των παραμέτρων. Επιλύει δηλαδή το παρακάτω πρόβλημα ελαχιστοποίησης:

$$\min_{\beta_1,\dots,\beta_m} \tfrac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} x_{ij}\,\beta_j)^2 \text{ , με } \sum_{j=1}^{m}|\beta_j| \le t, \text{ για κάποιο } t,$$

όπου $(x_{i,1},\dots,x_{i,m},y_i), i = 1,\dots,n$ αποτελούν τα δεδομένα μας, με x_{ij} και y_i τις ανεξάρτητες και την εξαρτημένες μεταβλητές αντίστοιχα, κατά την παρατήρηση i. Η παράμετρος t ονομάζεται ρυθμιστική παράμετρος και αν είναι μεγαλύτερη ή ίση της l^1-νόρμας των εκτιμητών των ελαχίστων τετραγώνων, τότε η εκτιμήτρια δεν αλλάζει με την εφαρμογή της μεθόδου Lasso. Για μικρότερες, ωστόσο, τιμές, η Lasso συρρικνώνει το διάνυσμα των συντελεστών της παλινδρόμησης που εκτιμώνται, προς το αρχικό. Συνήθως, θέτει κάποιον συντελεστή ίσο με το μηδέν, κάτι που έχει ως αποτέλεσμα και την μείωση των χαρακτηριστικών. Συνεπώς, η μέθοδος Lasso εκτελεί και επιλογή χαρακτηριστικών κατά την εκτίμηση παραμέτρων, σημείο στο οποίο πλεονεκτεί σχετικά με την παλινδρόμηση κορυφογραμμής (Ridge Regression) [32].

Ο υπολογισμός, ωστόσο και η επίλυση του παραπάνω ζητήματος, αποτελεί ένα απαιτητικό, υπολογιστικά, πρόβλημα τετραγωνικού προγραμματισμού. Ένα πρόβλημα ισοδύναμο στο προηγούμενο είναι: $\min_{\beta_1,\dots,\beta_m} \tfrac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} x_{ij}\,\beta_j)^2 + \lambda\sum_{j=1}^{m}|\beta_j|$, εφόσον για δοσμένο λ, με $0 \le \lambda < \infty$, υπάρχει $t \ge 0$, τέτοιο ώστε τα δύο αυτά προβλήματα να έχουν την ίδια λύση. [1] Όσο πιο μεγάλες είναι οι τιμές του λ, τόσοι περισσότεροι συντελεστές μηδενίζονται κατά την εφαρμογή του αλγορίθμου. Το λ υπολογίζεται με τη μέθοδο της Διασταυρωμένης Επικύρωσης (Cross Validation). Οι εκτιμήσεις, τελικά, των παραμέτρων της μεθόδου Lasso, υπολογίζονται βάσει της σχέσης:

$$\hat{\beta}^{Lasso} = \min_{\beta}\{||Y - X\beta||_2^2 + \lambda||\beta||_1\}\,.$$

Όμοια με την Παλινδρόμηση Κορυφογραμμής, έτσι και η Lasso, ανταλλάσει την εισαγωγή μιας μικρής κλίσης στην εκπαίδευση του μοντέλου, προκειμένου να επιτύχει μικρότερη διακύμανση της γραμμής παλινδρόμησης. Παρά τα πλεονεκτήματα της Lasso, και καθώς η επιλογή των μεταβλητών κρίνεται ολοένα και περισσότερο σημαντική για την ανάλυση των δεδομένων, υπάρχουν αρκετές περιπτώσεις στις οποίες η τεχνική αυτή υστερεί. Οι περιορισμοί της μεθόδου έχουν να κάνουν με τις περιπτώσεις στις οποίες το πλήθος των ανεξάρτητων μεταβλητών είναι μεγαλύτερο του αριθμού των δειγμάτων, οπότε και ο αλγόριθμος επιλέγει το πολύ τόσες μεταβλητές όσος και ο αριθμός των παρατηρήσεων. Ακόμα στις περιπτώσεις που έχουμε πολύ μεγάλη συσχέτιση μεταξύ δύο ή περισσότερων μεταβλητών, η μέθοδος Lasso θα επιλέξει μόνο τη μία από αυτές, χωρίς κάποιο άλλο κριτήριο, γεγονός που οδηγεί στην καλύτερη επίδοση της τεχνικής Ridge σε σχέση με την Lasso, για τις συγκεκριμένες περιπτώσεις.

2.5.5 Elastic Net Παλινδρόμηση

Το 2005 οι Hui Zou και Trevor Hastie [33], ανέπτυξαν μία νέα μέθοδο εφαρμογής των μοντέλων της γραμμικής παλινδρόμησης, η οποία υλοποιεί κανονικοποίηση και επιλογή μεταβλητών, την Elastic Net. Κίνητρο για την ανάπτυξη της μεθόδου, αποτέλεσε η ανάγκη για ένα μοντέλο το οποίο να έχει την δυνατότητα να δώσει ικανοποιητικά αποτελέσματα στις περιπτώσεις όπου η τεχνική Lasso δεν ήταν ικανή να ανταποκριθεί με επάρκεια. Παρόμοια με την Lasso, η Elastic Net εφαρμόζει επιλογή χαρακτηριστικών, ενώ ταυτόχρονα, όπως η Παλινδρόμηση Κορυφογραμμής, υλοποιεί και συρρίκνωση του διανύσματος των παραμέτρων, έχοντας την δυνατότητα να επιλέγει σύνολα μεταβλητών οι οποίες έχουν υψηλή συσχέτιση μεταξύ τους. Σε πειράματα δεδομένων, η μέθοδος Elastic Net μας δίνει καλύτερα αποτελέσματα από ότι η Lasso και η Παλινδρόμηση Κορυφογραμμής, όσον αφορά την ακρίβεια των προβλέψεων. Το τελευταίο είναι λογικό επακόλουθο του γεγονότος ότι η Elastic Net αποτελεί συνδυασμό των δύο προαναφερθέντων μεθόδων.

Η τεχνική Elastic Net εισάγει δύο σταθερές κόστους λ_1 και λ_2, καθώς επίσης και το τετράγωνο του όρου $||\beta||$ και επιλύει το πρόβλημα:

$$\hat{\beta} = \underset{\beta}{\arg\min} \quad \left(\left(||y - X\beta|| \right)^2 + \lambda_2 ||\beta||^2 + \lambda_1 ||\beta||_1 \right).$$

Το L1 μέρος της ποινής μας δίνει ένα αραιό μοντέλο, ενώ η προσθήκη του τετραγώνου της παραμέτρου β, καθιστά την παραπάνω εξίσωση κυρτή και συνεπώς έχει μοναδικό ελάχιστο. Ως αποτέλεσμα, αφαιρείται ο περιορισμός ως προς τον αριθμό των επιλεχθέντων χαρακτηριστικών και παράλληλα ενθαρρύνεται η ομαδοποίηση και η συστηματοποίηση των αποτελεσμάτων.

2.5.6 Μάθηση Συνόλου

Η μάθηση συνόλου είναι μία διαδικασία, κατά την οποία συνδυάζονται πολλαπλοί αλγόριθμοι και μοντέλα εκπαίδευσης, προκειμένου να καταλήξουμε σε ένα βελτιστοποιημένο αποτέλεσμα, το οποίο θα επιλύει κάποιο πρόβλημα υπολογιστικής νοημοσύνης. Είναι μέθοδος επιβλεπόμενης μάθησης, παρόλο που παρόμοιες μέθοδοι έχουν αναπτυχθεί και για τις περιπτώσεις της επιβλεπόμενης μάθησης, και χρησιμοποιείται για να βελτιωθεί η απόδοση ενός μοντέλου ή για να μειωθεί η πιθανότητα της λανθασμένης επιλογής ενός μη ακριβούς. Τα επιμέρους μοντέλα τα οποία χρησιμοποιούνται για την δημιουργία του συνολικού μοντέλου, μπορεί είτε να είναι ανεξάρτητα είτε να έχουν κάποιον βαθμό εξάρτησης μεταξύ τους. Στην πρώτη περίπτωση κατασκευάζονται ανεξάρτητα και στο τέλος συνδυάζονται, για να αποκτηθεί το τελικό μοντέλο, ενώ στην δεύτερη περίπτωση για την κατασκευή του ενός μοντέλου λαμβάνονται υπόψιν τα αποτελέσματα του αμέσως προηγούμενου.

Ο Schapire (1990) [34] συνδύασε πολλαπλούς ταξινομητές με μία τεχνική που αργότερα αποκάλεσε ενίσχυση (boosting). Η ενίσχυση αποτέλεσε πρόγονο πολλών μεθόδων, όπως η οικογένεια αλγορίθμων AdaBoost. Από τότε έχουν αναπτυχθεί μεγάλος αριθμός αλγορίθμων και τεχνικών που αφορούν μοντέλα τα οποία στηρίζονται σε συστήματα συνόλου, όπως για παράδειγμα τα Τυχαία Δάση, η Παλινδρόμηση Σωρού και ο συνδυασμός πολλαπλών μοντέλων, τα οποία εξηγούνται στη συνέχεια και χρησιμοποιούνται για την εκπαίδευση των μοντέλων στην συνέχεια της εργασίας.

2.5.7 Τυχαία Δάση

Τα Τυχαία Δάση (Random Forest) αποτελούν μία τεχνική Δένδρων Απόφασης και Μάθησης Συνόλου, που χρησιμοποιείται για την επίτευξη της ταξινόμησης, της παλινδρόμησης και άλλων μεθόδων. Είναι μία ελκυστική μέθοδος μοντελοποίησης των δεδομένων, χάρης στην ευκολία κατανόησης τους και την ταχύτητα με την οποία μας δίνουν αποτελέσματα. Για το λόγο αυτό, έχουν μελετηθεί σε βάθος τις τελευταίες δύο δεκαετίες και χρησιμοποιούνται σε ένα μεγάλο εύρος εφαρμογών. Ο Breiman απέδειξε ότι η χρήση ενός συνόλου δέντρων δίνει ακριβέστερα αποτελέσματα ταξινόμησης και παλινδρόμησης. Κάθε ένα από τα δέντρα μεγαλώνει ανάλογα με την τιμή μιας τυχαίας παραμέτρου και οι τελικές προβλέψεις λαμβάνονται από το συνολικό μοντέλο, το οποίο, εφόσον αποτελείται από επιμέρους μικρότερα δέντρα, ονομάζεται δάσος και εξαιτίας της τυχαιότητας του τρόπου με τον οποίο αυτά κατασκευάζονται, καλείται Τυχαίο Δάσος.

Τα τυχαία δάση είναι γρήγορα και εύκολα στην υλοποίηση τους, παράγουν εξαιρετικά ακριβή αποτελέσματα και έχουν την δυνατότητα να διαχειριστούν μεγάλο όγκο δεδομένων. Θεωρούνται, άλλωστε, ως μία από τις πιο αξιόπιστες γενικού σκοπού, τεχνική μάθησης. Το θεμελιώδες στοιχείο, από το οποίο κατασκευάζονται τα τυχαία δάση, είναι το δυαδικό δέντρο. Κατά με τον Gerard Biau [35], κάθε δέντρο στο σύνολο, σχηματίζεται επιλέγοντας τυχαία, για κάθε κόμβο, ένα σύνολο μεταβλητών εισόδου, ή χαρακτηριστικών, με βάση τα οποία θα γίνει ο διαχωρισμός, στη συνέχεια, στους κόμβους του αμέσως κατώτερου επιπέδου, υπολογίζοντας ταυτόχρονα τον καλύτερο τρόπο διαχωρισμού, ανάλογα με τα δεδομένα του δείγματος εκπαίδευσης. Το δέντρο μεγαλώνει με την εφαρμογή της τεχνικής των Ιεραρχικών Δέντρων Ταξινόμησης (CART), με βάση την οποία χρησιμοποιείται ένας αλγόριθμος, ο οποίος εισήχθηκε από τον Breimen το 1984, που υπολογίζει συγχρόνως και την τιμή μιας αριθμητικής εξαρτημένης μεταβλητής, γεγονός που αντιστοιχεί σε παλινδρόμηση, αλλά και την τιμή μίας κατηγορικής μεταβλητής και άρα υλοποιεί και ταξινόμηση (classification). Με βάση την CART ένας κόμβος διασπάται δυαδικά σε ομοιογενείς ή σχεδόν ομοιογενείς κόμβους. Ο στόχος είναι τα παιδιά του αρχικού κόμβου να έχουν υψηλότερο βαθμό ομοιογένειας από ότι αυτός [36]. Στην παραπάνω τεχνική εφαρμόζεται ακόμα και η μέθοδος bagging, προκειμένου να ανανεώνονται τυχαία το δείγμα εκπαίδευσης αλλά και τα χαρακτηριστικά που επιλέγονται από τους κόμβους γονείς, κάθε φορά που δημιουργείται ένα νέο δέντρο, προσθέτοντας, έτσι, ακόμα έναν βαθμό τυχαιότητας στην κατασκευή του δάσους, ο οποίος έχει ως αποτέλεσμα και την μείωση της συσχέτισης του συνόλου. Εκτός από την τεχνική bagging, πολύ καλά αποτελέσματα δίνουν οι μέθοδοι Random Split Selection, η οποία προτάθηκε από τον Dieterrich το 1998, όπως επίσης και η Adaboost, από τους Freund και Schapire το 1996. Να σημειωθεί ότι τα δέντρα που σχηματίζονται κατά την παραπάνω διαδικασία δεν κλαδεύονται [37].

Κάθε φορά που εφαρμόζεται η μέθοδος bagging, στην εκπαίδευση των δέντρων του συνόλου, λαμβάνει τυχαία ένα δείγμα από το δείγμα εκπαίδευσης, μεγέθους όσο αυτό. Αν εφαρμοστεί, δηλαδή Μ φορές bagging, θα έχω εν τέλει Μ σύνολα δειγμάτων, τα οποία θα περιέχουν το καθένα n διανύσματα, όσα και το πλήθος των δεδομένων εκπαίδευσης. Επίσης επιλέγονται τα χαρακτηριστικά με βάση τα οποία θα γίνει η εκπαίδευση του κάθε δέντρου, ως ένα υποσύνολο των συνολικών χαρακτηριστικών των δειγμάτων μας. Τα δέντρα εκπαιδεύονται με βάση τα Μ σύνολα και στη συνέχεια από τα αποτελέσματα που μας δίνουν, κρατάμε τα κατά μέσο όρο καλύτερα ή αυτό με το μεγαλύτερο πλήθος περιπτώσεων στην έξοδο, σε περίπτωση που έχουμε πρόβλημα ταξινόμησης.

Τα Τυχαία Δάση είναι μια πολύ καλή λύση στο πρόβλημα που δημιουργείται όταν ένα δέντρο μεγαλώνει πολύ σε βάθος, γεγονός που πολλές φορές προκαλεί υπερεκπαίδευση. Η τεχνική αυτή έχει ως αποτέλεσμα σημαντική μείωση της διακύμανσης, όπως αναφέρει ο Reiman. Το τελευταίο κρίνεται σημαντικό, καθώς πιο ακριβή τυχαία δάση έχουν χαμηλή διακύμανση, όπως και χαμηλή συσχέτιση μεταξύ των επιμέρους δέντρων, αλλά υψηλότερη ισχύ, δηλαδή ευστοχία πρόβλεψης. Οι παραπάνω συντελεστές κατευθύνονται σημαντικά από τον αριθμό των χαρακτηριστικών τα οποία επιλέγονται για τον σχηματισμό του εκάστοτε δέντρου.

2.5.8 Ενίσχυση Κλίσης - Gradient Boosting

Η Ενίσχυση αποτελεί μία μέθοδο Μηχανικής Μάθησης και μάλιστα Μάθησης Συνόλου, η οποία χρησιμοποιείται σε προβλήματα ταξινόμησης και παλινδρόμησης. Η προσέγγιση της αφορά τον συνδυασμό μεγάλου αριθμού αδύναμων, σχετικά, μοντέλων, προκειμένου να αποκτήσει μία ισχυρότερη συνολική πρόβλεψη. Συνήθως τα μοντέλα μηχανικής μάθησης, τα οποία συνδυάζονται, είναι τα Τυχαία Δάση και τα Νευρωνικά Δίκτυα και κοινή τεχνική τους αποτελεί η απλή εύρεση του μέσου όρου τους, για να χρησιμοποιηθεί στο συνολικό μοντέλο. Όπως αναφέρουν και οι Alexey Natekin και Alois Knoll [38], κύριο χαρακτηριστικό της ενίσχυσης αποτελεί η προσθήκη νέων μοντέλων στο σύνολο, διαδοχικά. Σε κάθε επανάληψη ένα νέο αδύναμο μοντέλο εκπαιδεύεται, λαμβάνοντας υπόψιν το σφάλμα του μέχρι τώρα συνόλου.

Η Ενίσχυση Κλίσης βασίζεται, σε αντιστοιχία με τα παραπάνω, στην διαδοχική εκπαίδευση νέων ασθενών μοντέλων, που κατά κανόνα είναι πιο αδύναμα από το τελικό σύνολο των μοντέλων, με σκοπό την επίτευξη μίας πιο ακριβούς προσέγγισης της μεταβλητής απόκρισης. Βασική αρχή αυτού του αλγορίθμου αποτελεί η κατασκευή των νέων εκπαιδευόμενων μοντέλων, έτσι ώστε να έχουν μέγιστη συσχέτιση με την αρνητική κλίση της συνάρτησης κόστους του συνόλου, την οποία επιλέγει ο ερευνητής. Συνεπώς η Παλινδρόμηση Ενίσχυσης Κλίσης, η οποία θα χρησιμοποιηθεί στην παρούσα εργασία, είναι μία γενίκευση της Ενίσχυσης Κλίσης και αποτελείται από τρία βασικά στοιχεία: την συνάρτηση κόστους που θέλουμε να βελτιστοποιήσουμε, το αδύναμο προς εκπαίδευση μοντέλο και ένα μοντέλο που προσθέτει τα αδύναμα προς εκπαίδευση μοντέλα, στο σύνολο. Τα ασθενή μοντέλα που χρησιμοποιούνται ως «μαθητές» είναι τα δένδρα απόφασης, το μέγεθος των οποίων παραμένει σταθερό κατά τη διάρκεια της εκπαίδευσης [39].

Η μαθηματική μορφή που λαμβάνει τελικά το μοντέλο είναι η εξής:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x), \qquad (2.4.1)$$

Με

$$F_m(x) = \sum_{m=1}^{M} \gamma_m h_m(x), \qquad (2.4.2)$$

Όπου $h_m(x)$ είναι οι ασθενείς συναρτήσεις βάσης των δέντρων απόφασης. Σε κάθε βήμα το δέντρο απόφασης $h_m(x)$ επιλέγεται για την ελαχιστοποίηση της συνάρτησης κόστους, δοσμένου του τρέχοντος μοντέλου και του $F_{m-1}(x_i)$.

$$F_m = F_{m-1}(x) + arg\,{}^{min}_{h} \quad \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) - h(x)) \qquad (2.4.3)$$

Σύμφωνα με τον Friedman (1999) [40], η επαναληπτική διαδικασία χρησιμοποιεί την συνάρτηση κόστους και τις ασθενείς συναρτήσεις $h_m(x)$ και στοχεύει στον υπολογισμό – προσέγγιση της αρνητικής κλίσης της συνάρτησης σφάλματος, προκειμένου να ενημερωθεί η συνάρτηση – εκτίμηση του επόμενου βήματος. Η αρνητική κλίση της συνάρτησης σφάλματος δίνεται παρακάτω:

$$-g(x_i) = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]$$

Με βάση την παραπάνω εξίσωση, υπολογίζεται το βέλτιστο βήμα της ενίσχυσης κλίσης γ_m, και προσαρμόζεται η συνάρτηση $h_m(x)$ σύμφωνα με αυτό, όπως φαίνεται και στην εξίσωση (2.4.1) και όπου γ_m:

$$\gamma_m = argmin_\gamma \sum_{i=1}^{n} L(y_i, \hat{F}_{m-1}(x_i) + \gamma h(x_i)).$$

Έτσι, δημιουργείται ένα σύνολο από δέντρα απόφασης F_m, καθένα από τα οποία για την εκπαίδευσή του χρησιμοποιεί στοιχεία από τα προηγούμενα δέντρα, προκειμένου να βελτιστοποιείται η ακρίβεια του τελικού συνόλου των δέντρων απόφαση.

2.5.9 Παλινδρόμηση Σωρού

Ας υποθέσουμε ότι έχουμε Κ μοντέλα παλινδρόμησης ή ταξινόμησης, $v_1(x), \dots, v_K(x)$, μίας αριθμητικής μεταβλητής απόκρισης y, που έχουν κατασκευαστεί χρησιμοποιώντας το ίδιο σύνολο εκπαίδευσης $L = \{(\boldsymbol{x_i}, y_i), i = 1, \dots, N\}$, όπου κάθε $\boldsymbol{x_i}$ αποτελεί το διάνυσμα των εισόδων του δείγματος. Κοινώς, τα μοντέλα αυτά είναι ίδιου τύπου, με τυχόν διαφορές στην πολυπλοκότητα. Το ζητούμενο είναι η κατασκευή ενός μοντέλου, λαμβάνοντας υπόψιν όλα τα μοντέλα $v_1(x), \dots, v_K(x)$. Σύμφωνα με τον Breiman (1996) [41], αυτό μπορεί να επιτευχθεί βρίσκοντας το βέλτιστο μοντέλο παλινδρόμησης ή ταξινόμησης από τα v_K μοντέλα, είτε μέσω ενός συνόλου ελέγχου, είτε με τη χρήση της τεχνικής διασταυρωμένης επικύρωσης (cross validation), για να υπολογιστεί και να συγκριθεί το σφάλμα πρόβλεψης κάθε v_K σε μελλοντικά δεδομένα.

Ο Wolpert (1992) [42] προτείνει μία νέα τεχνική, ως γενίκευση σωρού, κατά την οποία αντί να επιλέγεται το βέλτιστο μοντέλο από το σύνολο $\{v_1(x), \dots, v_K(x)\}$, ένα πιο ακριβές μοντέλο παλινδρόμησης ή ταξινόμησης μπορεί να κατασκευαστεί συνδυάζοντας όλα τα μοντέλα του παραπάνω συνόλου. Αυτή η μέθοδος βασίζεται στον κατάλληλο γραμμικό συνδυασμό των εξόδων των μοντέλων, όπως δηλώνεται από τη σχέση:

$$\hat{\boldsymbol{y}} = [\hat{y}_1, \dots, \hat{y}_K]\boldsymbol{w} \qquad (2.4.4)$$

Προκειμένου να ελαχιστοποιηθεί η ποσότητα

$$\sum_n (y_n - \sum_k w_k v_k(x_n))^2 \qquad (2.4.5)$$

Στην σχέση (2.4.4) το διάνυσμα y αποτελεί όπως και πριν το διάνυσμα των αποκρίσεων των μοντέλων μας και το διάνυσμα w αφορά τις παραμέτρους του γραμμικού συνδυασμού αυτών. Η ουσία της μεθόδου βρίσκεται στον τρόπο υπολογισμού του διανύσματος w. Για

να αποφευχθεί η υπέρ-εκπαίδευση, η οποία είναι πολύ πιθανόν να υπάρξει με τη χρήση της σχέσης (2.4.5), αντί για αυτήν βρίσκουμε το διάνυσμα w, έτσι ώστε να ελαχιστοποιείται η ποσότητα:

$$\sum_n (y_n - \sum_k a_k z_{kn})^2 \qquad (2.4.6)$$

Επίσης, επειδή οι παράμετροι w έχουν μεγάλη συσχέτιση, καθώς προέρχονται από μοντέλα εκπαιδευμένα στο ίδιο σύνολο δεδομένων, μικρές αλλαγές στα δεδομένα προκαλούν σημαντικές αλλαγές στα αποτελέσματα του παραπάνω και μειώνεται η γενίκευση. Για το λόγο αυτό, χρήσιμη εδώ αποτελεί η εφαρμογή της παλινδρόμησης κορυφογραμμής, η οποία ελαχιστοποιεί την ποσότητα της εξίσωσης (2.4.6), υπό τον περιορισμό $\alpha_\kappa \geq 0, k = 1, \dots, K$. Η παλινδρόμηση σωρού αποτελεί τεχνική μάθησης συνόλου και χρησιμοποιείται στην παρούσα μελέτη, για τον συνδυασμό των μοντέλων των προηγούμενων μεθόδων, οι οποίες εφαρμόσθηκαν.

Κεφάλαιο 3

Σχεδιασμός και Υλοποίηση

Στο παρόν κεφάλαιο θα αναπτυχθεί αναλυτικά η μεθοδολογία που ακολουθήθηκε για την εκτέλεση της παρούσας διπλωματικής εργασίας και την εξαγωγή των επιθυμητών αποτελεσμάτων. Τα βήματα που εκτελέστηκαν είναι: α) η συλλογή των δεδομένων, β) η ανάλυση και επεξεργασία των δεδομένων και η δημιουργία διαφορετικών συνόλων δεδομένων προς μοντελοποίηση και γ) η ανάπτυξη των μοντέλων μάθησης που αναλύθηκαν στο προηγούμενο κεφάλαιο. Στη συνέχεια περιγράφεται αναλυτικά η ανωτέρω διαδικασία.

3.1 Συλλογή Δεδομένων

Για την συλλογή των δεδομένων υλοποιήθηκε ένα πρόγραμμα σε γλώσσα προγραμματισμού python, όπου χρησιμοποιήθηκε η βιβλιοθήκη BeautifulSoup και το εργαλείο Selenium, με την χρήση του οποίου καταφέραμε να συγκεντρώσουμε δεδομένα ακινήτων από τον ιστότοπο xe.gr. Αυτός ο ιστότοπος αποτελεί μία πλατφόρμα, στην οποία δημοσιεύονται αγγελίες κάθε είδους, για τις ανάγκες, βέβαια, της εργασίας μας, ενδιαφερόμαστε για το τμήμα xe property, όπου καθένας μπορεί να βρει αγγελίες για την αγορά ή ενοικίαση ακινήτων κάθε είδους, στις πόλεις της Ελλάδας, και όχι μόνο. Δυστυχώς, εξαιτίας της φύσης των δεδομένων και επειδή δεν υπάρχουν, στα πλαίσια των Ελληνικών ιδιοκτησιών, συγκεντρωμένα και προσιτά δεδομένα ακινήτων, η διαδικασία συλλογής τους αποδείχθηκε ιδιαίτερα χρονοβόρα. Για το λόγο αυτό συνολικά συλλέχθηκαν οι περιγραφές 15.098 κατοικιών, οι οποίες ήταν δημοσιευμένες προς πώληση την περίοδο του Δεκεμβρίου 2019, έως τον Απρίλιο 2020 στον νομό Αττικής. Συγκεκριμένα, τα δεδομένα αφορούσαν τις περιοχές των Βορείων, Νοτίων, Ανατολικών και Δυτικών Προαστίων της Αττικής, το κέντρο Αθήνας και την περιοχή του Πειραιά. Τα δεδομένα αυτά αποθηκεύτηκαν σε μία βάση δεδομένων mySQL.

3.2 Εξαγωγή Χαρακτηριστικών

Αρχικός στόχος μας αποτέλεσε η εξαγωγή των χαρακτηριστικών από την περιγραφή του κάθε ακινήτου. Αποσκοπούμε στην δημιουργία ενός συνόλου δεδομένων το οποίο θα αποτελείται, όχι από τις περιγραφές των ακινήτων, αλλά από μεταβλητές που θα περιγράφουν εξίσου αποτελεσματικά κάθε μία από τις καταχωρήσεις που έχουν συλλεχθεί στην βάση δεδομένων. Αυτό επιτεύχθηκε με την υλοποίηση ενός προγράμματος επεξεργασίας φυσικής γλώσσας σε γλώσσα προγραμματισμού Python. Αρχικά μελετήθηκαν οι εγγραφές της βάσης δεδομένων, προκειμένου να αποφασιστεί το σύνολο χαρακτηριστικών τα οποία θα εξάγουμε. Στη συνέχεια με την χρήση του εργαλείου NLTK,

καταφέραμε να εξάγουμε, για κάθε ένα από τα ακίνητα, τα χαρακτηριστικά που φαίνονται στον ακόλουθο πίνακα.

Όνομα Πεδίου	**Τύπος**	**Περιγραφή**
Price	Int	Τιμή ακινήτου
Area	Int	Εμβαδόν ακινήτου
Year	Int	Έτος κτίσης
Housetype	Varchar(45)	Είδος ακινήτου
Neighbourhood	Varchar(45)	Περιοχή ακινήτου
Bedrooms	Int	Αριθμός υπνοδωματίων
Bathrooms	Int	Αριθμός μπάνιων
WC	Int	Αριθμός WC
Floor	Varchar(45)	Όροφος ακινήτου
Renovation	Int	Έτος ανακαίνισης
Heating	Varchar(45)	Είδος θέρμανσης
Penthouse	Int	Ρετιρέ
Fireplace	Int	Τζάκι
Pool	Int	Πισίνα
Safetydoor	Int	Πόρτα ασφαλείας
Transport	Int	Πλησίον ΜΜΜ
Market	Int	Πλησίον αγοράς
School	Int	Πλησίον σχολείου
Park	Int	Πλησίον Πάρκου
Hospital	Int	Πλησίον νοσοκομείου
Parking	Int	Ύπαρξη πάρκινγκ
Furniture	Int	Επιπλωμένο ακίνητο
Yard	Int	Ύπαρξη αυλής/κήπου
Elevator	Int	Ύπαρξη ασανσέρ
Orientation	Varchar(45)	Προσανατολισμός ακινήτου

Πίνακας 1 Χαρακτηριστικά που εξήχθησαν από την περιγραφή κάθε ακινήτου.

Στη συνέχεια θα εξηγήσουμε τις τιμές που λαμβάνουν κάποια από τα δεδομένα και που θεωρούμε ότι χρειάζεται κάποια επεξήγηση. Το πεδίο Renovation λαμβάνει την τιμή 0 σε περίπτωση που δεν έχει συμβεί ανακαίνιση στο παρελθόν, την τιμή 1 σε περίπτωση που έχει γίνει ανακαίνιση χωρίς όμως να αναφέρεται το έτος που αυτή έγινε και σε περίπτωση που και έχει γίνει ανακαίνιση και αναφέρεται η χρονολογία της, τότε η μεταβλητή παίρνει την τιμή του έτους που αυτή έγινε. Το πεδίο Housetype λαμβάνει τις τιμές διαμέρισμα, οροφοδιαμέρισμα, μονοκατοικία, μεζονέτα ή κτίριο ανάλογα με το είδος του ακινήτου. Το πεδίο Heating λαμβάνει τις τιμές αυτόνομη, κεντρική, αέριο, ενδοδαπέδια, αναλόγως με το είδος θέρμανσης. Τα πεδία Penthouse, Fireplace, Pool, Safetydoor, Parking, Furniture, Elevator αφορούν την ύπαρξη ή μη, ρετιρέ ακινήτου, τζακιού, πισίνας, πόρτας ασφαλείας, πάρκινγκ, επίπλων και ασανσέρ σε κάθε ακίνητο και παίρνουν τις τιμές 0 ή 1. Τα πεδία Transport, Market, School, Park, Hospital αποτελούν δυαδικές μεταβλητές και λαμβάνουν, επίσης, τις τιμές 0 ή 1 σε περίπτωση που το ακίνητο βρίσκεται ή όχι κοντά σε ΜΜΜ, αγορά, σχολείο, πάρκο ή νοσοκομείο. Τέλος, το πεδίο Yard παίρνει την τιμή 0 σε περίπτωση που η ιδιοκτησία δεν διαθέτει αυλή, την τιμή 1 σε περίπτωση που διαθέτει και την τιμή του εμβαδού της αυλής, αν αυτή υπάρχει και αναφέρεται το μέγεθός της. Αυτά ήταν τα πεδία με τα οποία ξεκινήσαμε και στα οποία στηρίξαμε την έρευνά μας, τα οποία όμως, τροποποιήσαμε και αναλύσαμε, όπως διαφαίνεται στην συνέχεια.

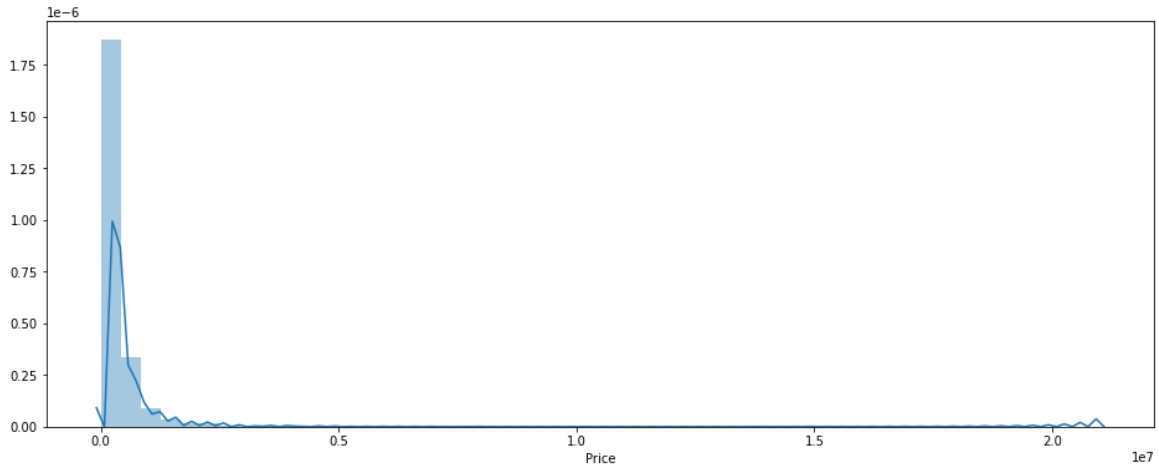
3.3 Διερευνητική Ανάλυση Δεδομένων

Στόχος του παρόντος κεφαλαίου αποτελεί η διερευνητική ανάλυση του συνόλου δεδομένων, προκειμένου να ληφθούν όλες οι δυνατές πληροφορίες που μπορούν να εξαχθούν από αυτό. Το συγκεκριμένο κρίνεται καίριο βήμα στην διαδικασία της έρευνας, και προτού μοντελοποιηθούν τα δεδομένα, καθώς, όπως αναφέρει και ο Tukey (1977) [14], μεγάλης σημασίας αποτελεί η κατανόηση των όσων έχουμε πραγματικά την δυνατότητα να καταφέρουμε, πριν αξιολογήσουμε το πόσο καλά τα καταφέραμε. Η ανάλυση του συγκεκριμένου συνόλου δεδομένων έγινε σε Python Notebook με την χρήση της πλατφόρμας Google Colab. Στη συνέχεια ακολουθούνται τεχνικές και μέθοδοι μονομεταβλητής και πολυμεταβλητής ανάλυσης, απεικόνισης και επεξεργασίας των δεδομένων και παρουσιάζονται τα αποτελέσματα.

3.3.1 Διερεύνηση και Απεικόνιση Μεταβλητής Στόχου

Η μεταβλητή που μας ενδιαφέρει και την οποία θέλουμε αρχικά να διερευνήσουμε, είναι η μεταβλητή Price, η οποία μας δίνει την τιμή του κάθε ακινήτου και είναι αυτή που θα μελετηθεί αρχικά. Παρακάτω φαίνονται τα χαρακτηριστικά και το ιστόγραμμα της μεταβλητής.

```
count    1.509700e+04
mean     3.414729e+05
std      5.329412e+05
min      0.000000e+00
25%      1.150000e+05
50%      2.100000e+05
75%      3.750000e+05
max      2.100000e+07
Name: Price, dtype: float64
```



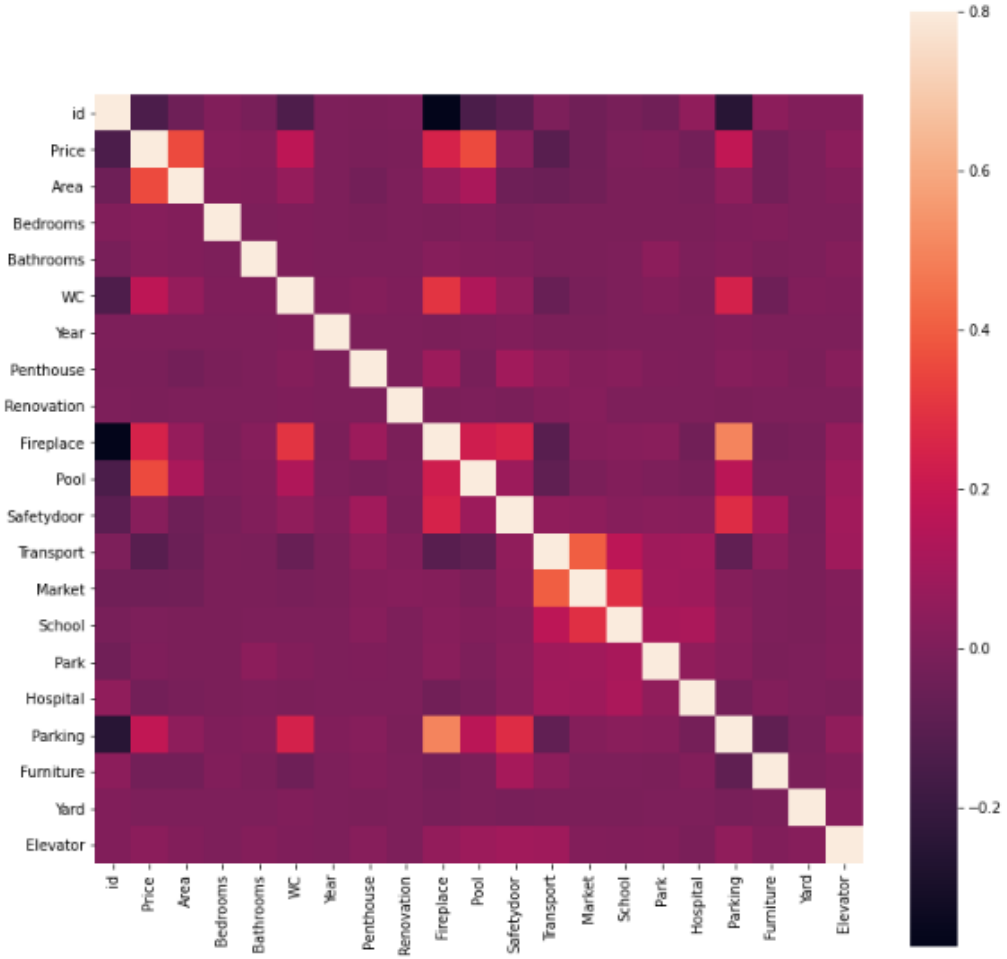
Εικόνα 2 Χαρακτηριστικά και Ιστόγραμμα της Τιμής Ακινήτων.

Όπως φαίνεται η τιμή έχει κάποιες μηδενικές τιμές, το οποίο φυσικά δεν είναι επιθυμητό και θα διορθωθεί στην συνέχεια. Επίσης, η κατανομή της διαφέρει από τη συμμετρική κανονική κατανομή και έχει θετική λοξότητα που υπολογίστηκε ίση με 12,499. Η λοξότητα

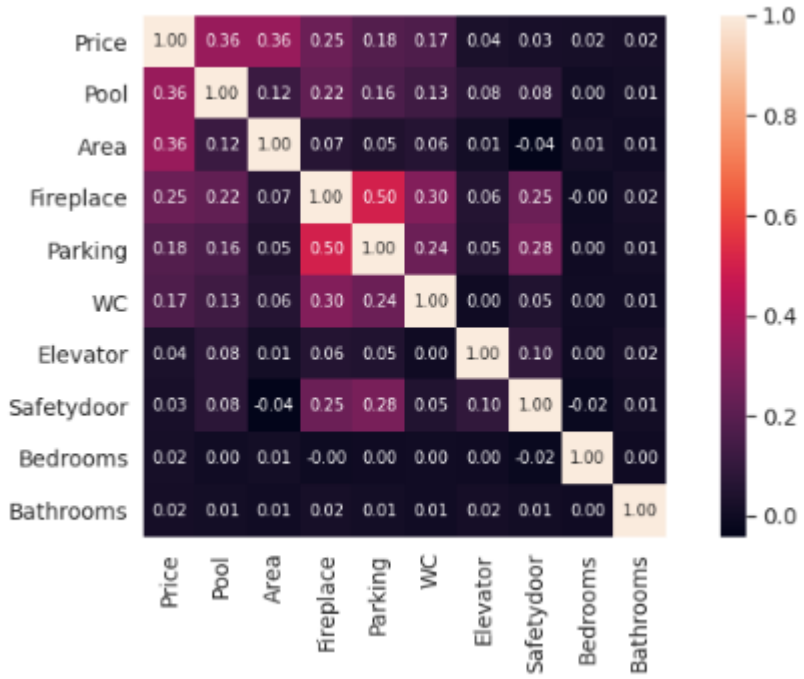
είναι αρκετά μεγάλη και όπως φαίνεται και από το ιστόγραμμα, παρατηρείται μεγάλη συσσώρευση τιμών σε μικρότερες τιμές παρά σε μεγαλύτερες και αυτό έχει ως αποτέλεσμα μία αρκετά ασύμμετρη κατανομή. Κάτι τέτοιο είναι, φυσικά, αναμενόμενο, εφόσον τα περισσότερα ακίνητα στον νομό Αττικής αποτελούν κατοικίες της μεσαίας τάξης και καθώς αυξάνεται η αξία τους, λιγοστεύει και ο αριθμός τους. Παρατηρείται, ακόμα, έντονη κορυφή και ουρά στην κατανομή, με την κύρτωση να μετριέται 325,274, τιμή αρκετά υψηλή. Η μέση τιμή της τιμής των ακινήτων πριν την επεξεργασία είναι 341.472,9€ με μέγιστη τα 21000000€ και ελάχιστη, όπως αναφέρθηκε τα 0€. Στη συνέχεια θα μελετήσουμε αριθμητικές μεταβλητές με μεγάλη συσχέτιση με την μεταβλητή Price.

3.3.2 Σχέση Τιμής με Αριθμητικές Μεταβλητές

Υπάρχουν συνολικά 24 μεταβλητές εκτός της τιμής, 19 από τις οποίες είναι αριθμητικές και 5 είναι κατηγορηματικές. Στην συγκεκριμένη παράγραφο θα μελετηθεί η σχέση των αριθμητικών μεταβλητών με την τιμή των ακινήτων. Αρχικά απεικονίζεται παρακάτω ο πίνακας συσχετίσεων όλων των αριθμητικών μεταβλητών, καθώς επίσης και ο πίνακας με τις 10 μεταβλητές οι οποίες έχουν το μεγαλύτερο επίπεδο συσχέτισης με την μεταβλητή της τιμής, μαζί με τον συντελεστή συσχέτισης.

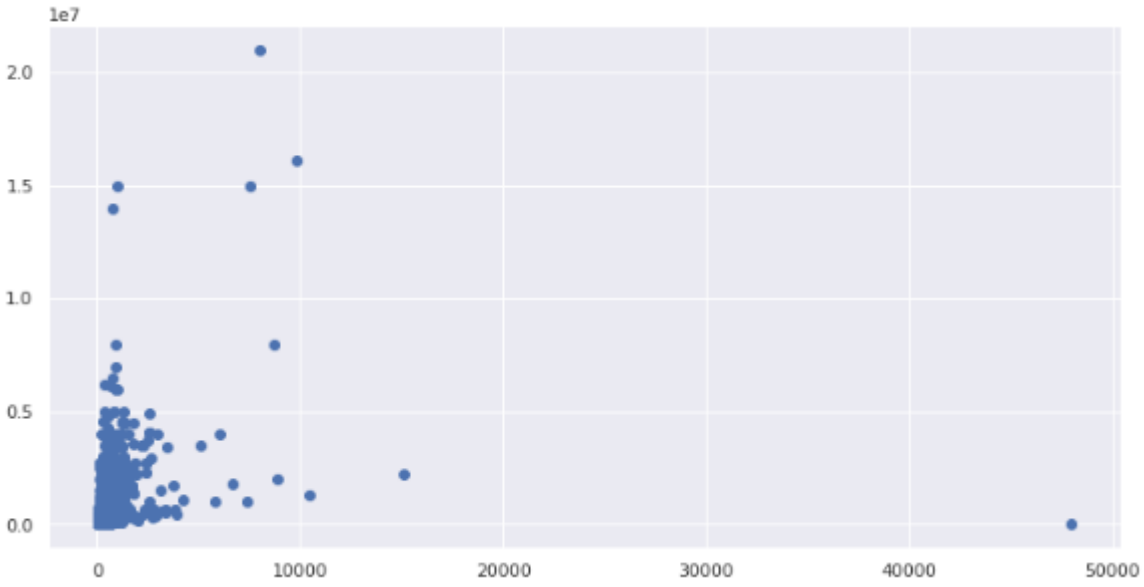


Εικόνα 3 Πίνακας Συσχετίσεων Μεταβλητών.

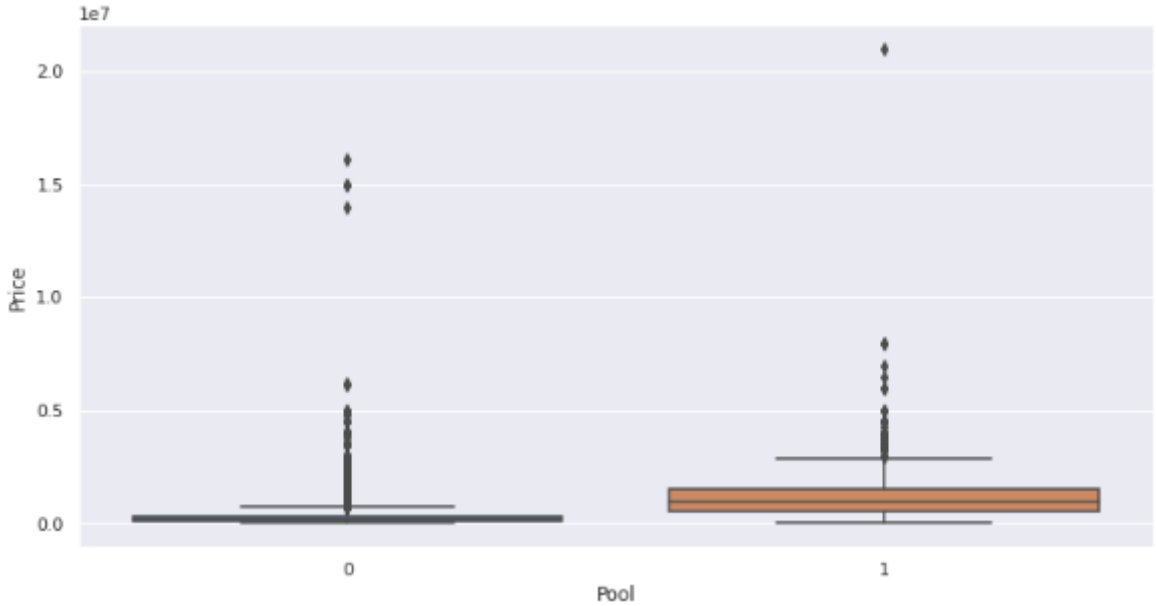


Εικόνα 4 Πίνακας των 10 περισσότερο Συσχετισμένων με την Τιμή Μεταβλητών.

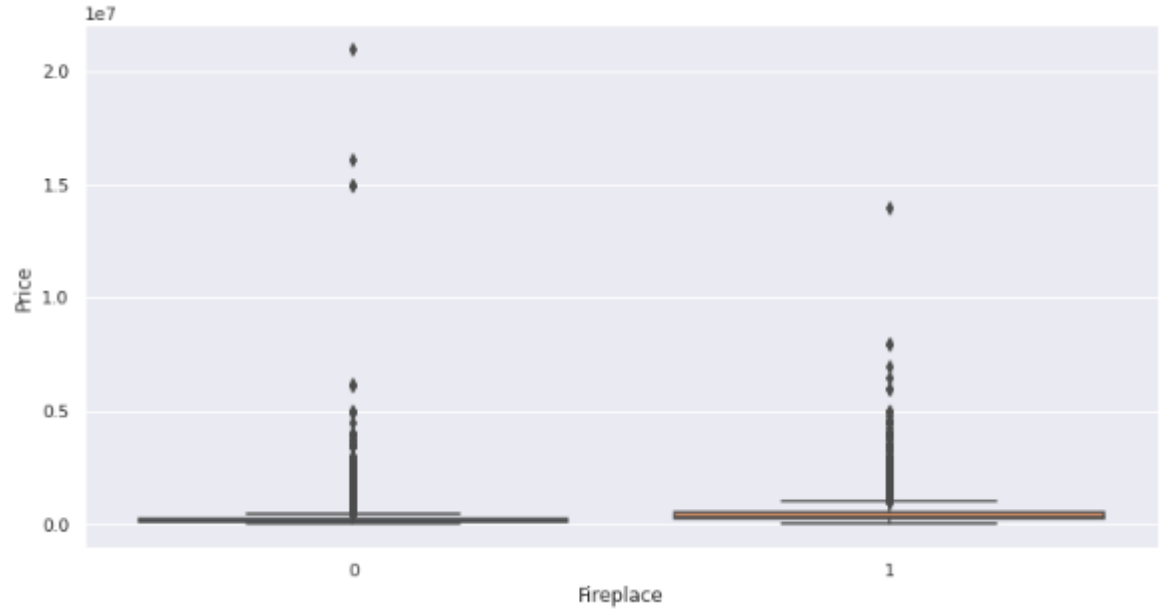
Γίνεται αντιληπτό ότι τα δεδομένα απαιτούν μία προεπεξεργασία, προκειμένου να αυξηθούν οι συσχετίσεις μεταξύ των μεταβλητών και να μπορέσουν να μας δώσουν ικανοποιητικά αποτελέσματα. Όπως μπορούμε να παρατηρήσουμε, μεγαλύτερη συσχέτιση με την μεταβλητή Price, έχουν οι μεταβλητές Pool, Area και Fireplace. Ας δούμε ακολούθως τη σχέση της κάθε μίας με την τιμή. Θα ξεκινήσουμε με το διάγραμμα διασποράς της μεταβλητής Area με την μεταβλητή Price και θα συνεχίσουμε με τα θηκογράμματα των μεταβλητών Pool και Fireplace με την τιμή.



Εικόνα 5 Διάγραμμα Διασποράς Εμβαδόν – Τιμής.



Εικόνα 6 Θηκόγραμμα των μεταβλητών Pool – Price.



Εικόνα 7 Θηκόγραμμα των μεταβλητών Fireplace – Price.

Η μορφή του διαγράμματος διασποράς Εμβαδόν – Τιμής είναι αναμενόμενη, καθώς είναι λογικό ότι τα μεγαλύτερα σπίτια θα κοστίζουν και περισσότερο. Αυτό που δεν είναι λογικό είναι το σημείο κάτω δεξιά, το οποίο θα διορθωθεί σε επόμενο βήμα. Από τα θηκογράμματα των μεταβλητών Πισίνα – Τιμή και Τζάκι – Τιμή, μπορούμε να καταλήξουμε στο, επίσης, λογικό συμπέρασμα ότι ως επί το πλείστον σπίτια που διαθέτουν πισίνα ή τζάκι έχουν μεγαλύτερη τιμή πώλησης από αυτά που δεν διαθέτουν.

Κατόπιν θα προχωρήσουμε στην επεξεργασία του συνόλου δεδομένων, των χαρακτηριστικών δηλαδή, που έχουμε κατασκευάσει, προκειμένου να έρθει σε μορφή ικανή μετά την είσοδό του στα μοντέλα, να μας παρέχει χρήσιμα αποτελέσματα.

3.3.3 Κενές και Ακραίες Τιμές

3.3.3.1 Κενές Τιμές

Το σύνολο των δεδομένων που έχουμε κατασκευάσει περιέχει κάποιες κενές, αλλά και ακραίες τιμές, τις οποίες σε αυτήν την παράγραφο θα διαχειριστούμε προκειμένου να διορθώσουμε. Αρχικά παρουσιάζουμε τα πεδία που περιέχουν κενές τιμές και στη συνέχεια για κάθε ένα θα αποκαθιστάμε τις τιμές που λείπουν.

	Total	Percent
Orientation	10624	0.703856
Heating	3097	0.205181
Housetype	745	0.049357
Floor	146	0.009673
Elevator	0	0.000000
Neighborhood	0	0.000000

Εικόνα 8 Σύνολο πεδίων με κενές τιμές.

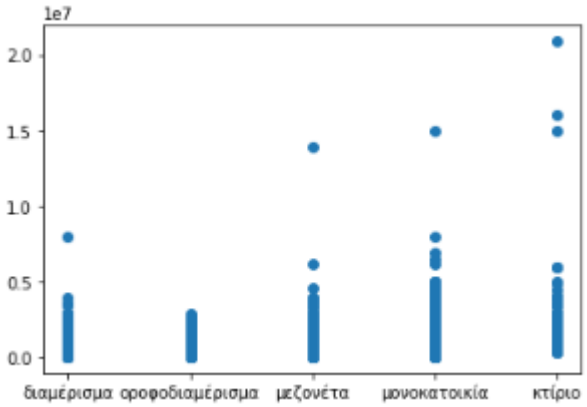
Παρατηρούμε ότι τον μέγιστο αριθμό κενών τιμών έχει το πεδίο Orientation, που περιγράφει τον προσανατολισμό του κάθε ακινήτου. Επιπροσθέτως, κενές τιμές έχουν τα πεδία Heating, Housetype και Floor, τα οποία περιγράφουν το είδος θέρμανσης, το είδος ακινήτου και τον όροφο του ακινήτου. Εκείνο που μπορούμε να συμπεράνουμε είναι ότι, εφόσον στο 70,39% των ακινήτων δεν καθορίζεται ο προσανατολισμός της κάθε ιδιοκτησίας, το χαρακτηριστικό αυτό δεν αποτελεί κάτι που συνήθως θα ληφθεί υπόψιν στην Ελλάδα, και συγκεκριμένα στον νομό της Αττικής τον οποίο αφορούν τα δεδομένα, κατά την διαδικασία αγοράς ενός ακινήτου. Μικρότερο είναι το ποσοστό των ακινήτων για τα οποία δεν αναφέρεται είδος θέρμανσης, είδος ακινήτου και ο όροφος, με κάθε ένα από τα ποσοστά να αγγίζουν τις τιμές 20,51%, 4,9% και 0.96% αντίστοιχα.

Ξεκινώντας από την μεταβλητή Orientation, αποφασίζουμε, εξαιτίας του μεγάλου αριθμού κενών τιμών και της μικρής συσχέτισής της με την τιμή του ακινήτου, να την διαγράψουμε από το σύνολο δεδομένων. Παρόλο που μπορούμε να υποθέσουμε ασφαλώς, ότι η αφαίρεση της μεταβλητής δεν θα επηρεάσει αρνητικά τα αποτελέσματά μας, πρέπει να σημειώσουμε ότι ο προσανατολισμός της οικίας είναι μια παράμετρος που λαμβάνεται υπόψιν σε σημαντικό βαθμό, όταν αγοράζεται ένα ακίνητο. Το γεγονός ότι τόσο μεγάλο ποσοστό ακινήτων δεν είχε αναφορά στον προσανατολισμό του, μας υποδεικνύει ότι η επιμελής οργάνωση και παρουσίαση των στοιχείων των ακινήτων είναι μία πτυχή της αγοράς, η οποία επιδέχεται βελτίωση.

Συνεχίζοντας, βλέπουμε ότι η μεταβλητή Heating διαθέτει επίσης μεγάλο ποσοστό κενών τιμών. Η αντικατάστασή τους, ειδικά για την συγκεκριμένη μεταβλητή, είναι δύσκολη και θεωρήθηκε ότι θα αλλοίωνε την πραγματικότητα η τυχαία αντικατάσταση με μία από τις τέσσερις δυνατές τιμές της μεταβλητής, καθώς δεν υπάρχει τρόπος να γνωρίζουμε με ικανοποιητική ακρίβεια το είδος θέρμανσης του κάθε ακινήτου. Για τους προαναφερθέντες λόγους, το πεδίο Heating αφαιρέθηκε από το σύνολο δεδομένων. Τα συμπεράσματα που βγάζουμε για την οργάνωση της αγοράς των ακινήτων, στον νομό τουλάχιστον της Αττικής, συμπίπτουν με τα προηγούμενα.

Ακολούθως, θα διαχειριστούμε τις κενές τιμές της μεταβλητής Housetype. Το ποσοστό των κενών τιμών του συγκεκριμένου πεδίου ήταν αποδεκτό, για το λόγο αυτό αποφασίσαμε να το χειριστούμε ως εξής: Ανάλογα με τον όροφο και το εμβαδόν του ακινήτου, σε

περίπτωση βέβαια που αναφέρεται ο όροφος, προσπαθήσαμε να προσδιορίσουμε με όσο το δυνατόν μεγαλύτερη ακρίβεια το είδος του ακινήτου. Λήφθηκε επίσης υπόψιν το γεγονός ότι στα μέχρι τώρα καταγεγραμμένα είδη ακινήτων, τα διαμερίσματα αποτελούν το 59,67%, οι μεζονέτες το 13,35%, οι μονοκατοικίες το 12,25% και τα οροφοδιαμερίσματα το 9,69%. Έτσι, για παράδειγμα, ακίνητα που αποτελούνταν μόνο από έναν όροφο και το εμβαδόν τους κυμαινόταν από 20 έως 150 τ.μ. κατατάσσονταν ως διαμερίσματα, ακίνητα των οποίων το εμβαδόν ξεπερνούσε τα 800 τ.μ. κατατάσσονταν ως κτίρια, κ.ο.κ. Με τον τρόπο αυτό διευθετήθηκαν όλες οι κενές τιμές για το πεδίο Housetype. Παρακάτω μπορούμε να δούμε πως τελικά σχηματίστηκε η μεταβλητή Housetype και ποια είναι η σχέση της με την τιμή. Όπως θα φανεί, μεγαλύτερη τιμή πώλησης έχουν οι μονοκατοικίες, οι μεζονέτες και τα κτίρια, ενώ ακολουθούν τα διαμερίσματα, αλλά και τα οροφοδιαμερίσματα με χαμηλότερες τιμές. Είναι λογικό και περιμέναμε, άλλωστε, μεγαλύτερη τιμή για τις μεζονέτες και τις μονοκατοικίες. Όσον αφορά τα οροφοδιαμερίσματα, ωστόσο, θα περιμέναμε ίσως υψηλότερες τιμές από ότι αυτές των διαμερισμάτων. Το γεγονός ότι κάτι τέτοιο δεν συμβαίνει, οφείλεται αρχικά, στον μικρό αριθμό των ακινήτων που καταχωρούνται ως οροφοδιαμερίσματα, σε σύγκριση με αυτά που καταχωρούνται ως διαμερίσματα, κάτι που ενδέχεται να αποτελεί συνέπεια της πιθανής καταχώρησης πολλών οροφοδιαμερισμάτων ως διαμερίσματα.



Εικόνα 9 Διάγραμμα Διασποράς Είδους Ακινήτου – Τιμής.

Τέλος, για το πεδίο Floor έχουμε 146 κενές τιμές και διορθώνονται ως εξής: Για τα ακίνητα τα οποία αποτελούν μονοκατοικίες τοποθετούμε στην μεταβλητή Floor την τιμή «ισόγειο». Αυτή η ενέργεια άλλαξε την τιμή μόνο τεσσάρων ακινήτων. Θα μπορούσαμε να καταχωρήσουμε τυχαία ορόφους για τα διαμερίσματα με κενές τιμές, και λαμβάνοντας υπόψιν την αναλογία των διαμερισμάτων σε κάθε όροφο, προτιμήσαμε ωστόσο να διατηρήσουμε την μορφή των δεδομένων ως έχει και να διαγράψουμε τα 142 ακίνητα για τα οποία δεν αναφέρεται κάποιος όροφος, εφόσον το ποσοστό αυτών είναι μικρό. Έτσι καταλήγουμε με 14.946 δεδομένα, για τα οποία συνεχίζουμε την ανάλυση.

Υπάρχουν κάποιες μεταβλητές, για τις οποίες η κενή τιμή δεν προσδιορίζεται ως NaN, αλλά ως 0. Αυτές οι μεταβλητές είναι οι εξής: Price, Area, Bedrooms, Bathrooms και Year. Συνεχίζοντας, θα διευθετήσουμε την ύπαρξη κενών τιμών στα παραπάνω πεδία.

Αρχικά διαγράφουμε από το σύνολο δεδομένων, όσες εγγραφές έχουν μηδενική τιμή πώλησης, δηλαδή μηδενική τιμή στην μεταβλητή Price. Στην συνέχεια θα αφαιρέσουμε και όλες τις εγγραφές με μηδενικό εμβαδόν. Ακίνητα με μηδενική καταχώρηση στο πεδίο Bedrooms δεν υπάρχουν, οπότε δεν προχωράμε σε κάποια ενέργεια. Για τις κενές τιμές στο πεδίο Bathrooms, υπολογίσαμε την μέση τιμή του αριθμού των μπάνιων ανά είδος ακινήτου και αντικαταστήσαμε αναλόγως τις τιμές που έλειπαν. Προχωρώντας στην μεταβλητή Year, παρατηρούμε πως υπάρχουν 602 ακίνητα για τα οποία δεν προσδιορίζεται το έτος κτίσης

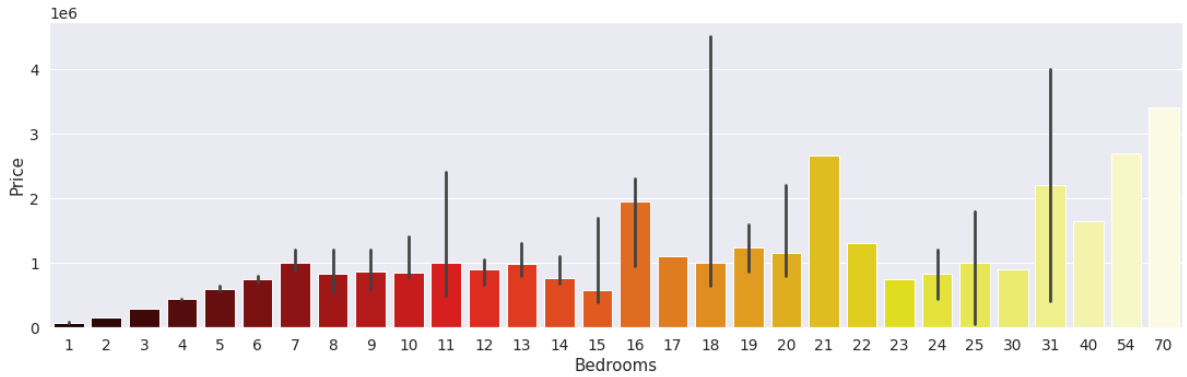
τους. Αυτά αντιστοιχούν σε ποσοστό 3,988% του συνολικού αριθμού των ακινήτων, το οποίο ποσοστό είναι σεβαστό, και συνεπώς η διαγραφή τους μπορεί να προκαλέσει δυσάρεστες αλλοιώσεις στο σύνολο. Αποφασίζουμε, καταληκτικά να αντικαταστήσουμε τις τιμές αυτές με τον μέσο όρο του έτος κτίσης των ακινήτων.

Πρέπει να σημειωθεί ότι όλα τα παραπάνω δεδομένα που διαγράφθηκαν από το σύνολο δεδομένων, δεν είχαν κάποιον βαθμό συσχέτισης είτε μεταξύ τους, είτε με άλλα πεδία του συνόλου δεδομένων και η διαγραφή τους δεν επηρέασε με κάποιον ανεπιθύμητο τρόπο το αποτέλεσμα, πχ εισάγοντας κάποια προκατάληψη στο σύνολο.

3.3.3.2 Ακραίες Τιμές

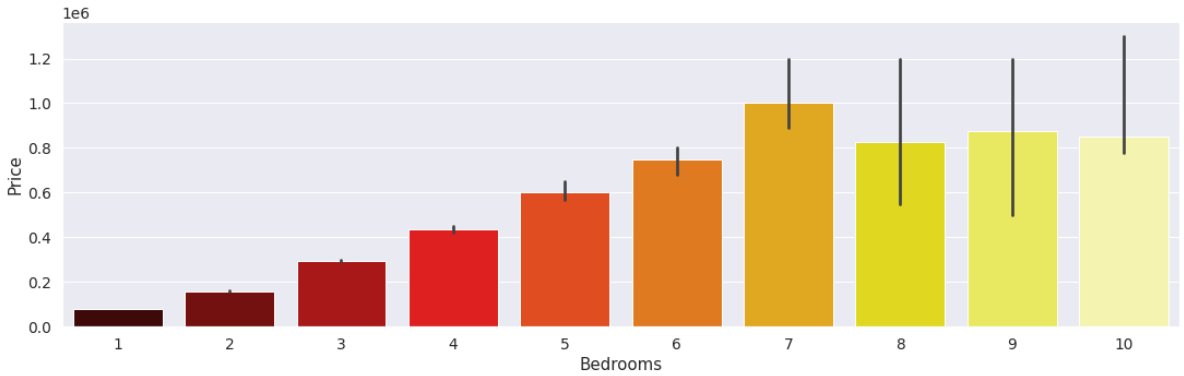
Σε αυτήν την ενότητα θα διευθετήσουμε τις ακραίες τιμές που υπάρχουν στο σύνολο δεδομένων. Οι ακραίες τιμές έχουν την δυνατότητα να μας προσφέρουν σημαντικές πληροφορίες για την συμπεριφορά του δείγματος και επηρεάζουν σημαντικά την εκπαίδευση του μοντέλου. Στην συνέχεια θα διορθώσουμε για κάθε ένα από τα πεδία τις ακραίες τιμές που ενδέχεται να υπάρχουν.

Αρχικά, ξεκινώντας με το πεδίο Area, θα διαγράψουμε το ακίνητο που φαίνεται στο Σχήμα 3.4 με το πολύ μεγάλο εμβαδόν και την χαμηλή τιμή. Επιπλέον, θα διαγράψουμε τα 6 ακίνητα που έχουν εμβαδόν μεγαλύτερο των 7000 τ.μ. Ακολούθως, συνεχίζουμε με την μεταβλητή Bedrooms. Η συγκεκριμένη μεταβλητή έχει αρκετές ακραίες τιμές, οι οποίες είμαστε σίγουροι ότι δεν συμπίπτουν με την πραγματικότητα. Η αντικατάσταση αυτών των τιμών έγινε, υπολογίζοντας τον μέσο όρο του αριθμού των υπνοδωματίων ανά είδος ακινήτου και αντικαθιστώντας αναλόγως. Η σχέση του αριθμού των υπνοδωματίων με την τιμή φαίνεται παρακάτω.



Εικόνα 10 Σχέση Αριθμού Υπνοδωματίων – Τιμής.

Παρατηρούμε ότι η τιμή του ακινήτου αυξάνεται με τον αριθμό των υπνοδωματίων μέχρι αυτός να πάρει την τιμή 7 και στη συνέχεια μειώνεται για να αυξηθεί πάλι μετά τα 15 υπνοδωμάτια. Οι τιμές μεγαλύτερες του 10, ωστόσο, αποτελούν μη χρήσιμη πληροφορία για το μοντέλο που επιθυμούμε να κατασκευάσουμε, ενώ αντιστοιχούν σε ακριβώς 95 ιδιοκτησίες, στο 0,6%, δηλαδή, των συνολικών ακινήτων. Για το λόγο αυτό η διαγραφή τους κρίνεται ασφαλής και εκτελείται. Η σχέση αριθμού υπνοδωματίων – τιμής ακινήτου διαμορφώνεται ως εξής:



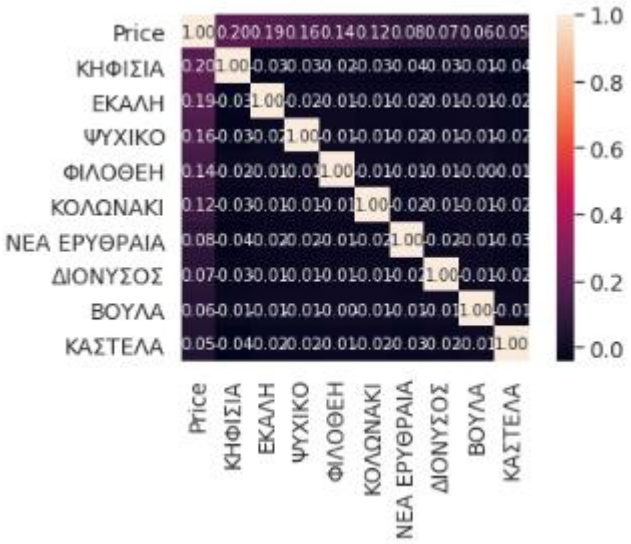
Εικόνα 11 Τελική Σχέση Αριθμού Υπνοδωματίων – Τιμής.

Συνεχίζουμε, αντικαθιστώντας τις μη λογικές τιμές των πεδίων Bathrooms και WC με τον μέσο όρο των πεδίων τους για κάθε είδος ακινήτου αντίστοιχα. Όσον αφορά το πεδίο Renovation, διατηρούμε τιμές μικρότερες του 2020 και μεγαλύτερες του 1950, αντικαθιστώντας με την μονάδα τις ακραίες τιμές που βρέθηκαν. Τέλος, για την μεταβλητή Yard, αντικαθιστούμε τις μη ρεαλιστικές τιμές εμβαδού με τον μέσο όρο των εμβαδών όσων ακινήτων διαθέτουν αυλή.

3.3.4 Επεξεργασία Κατηγορηματικών Χαρακτηριστικών

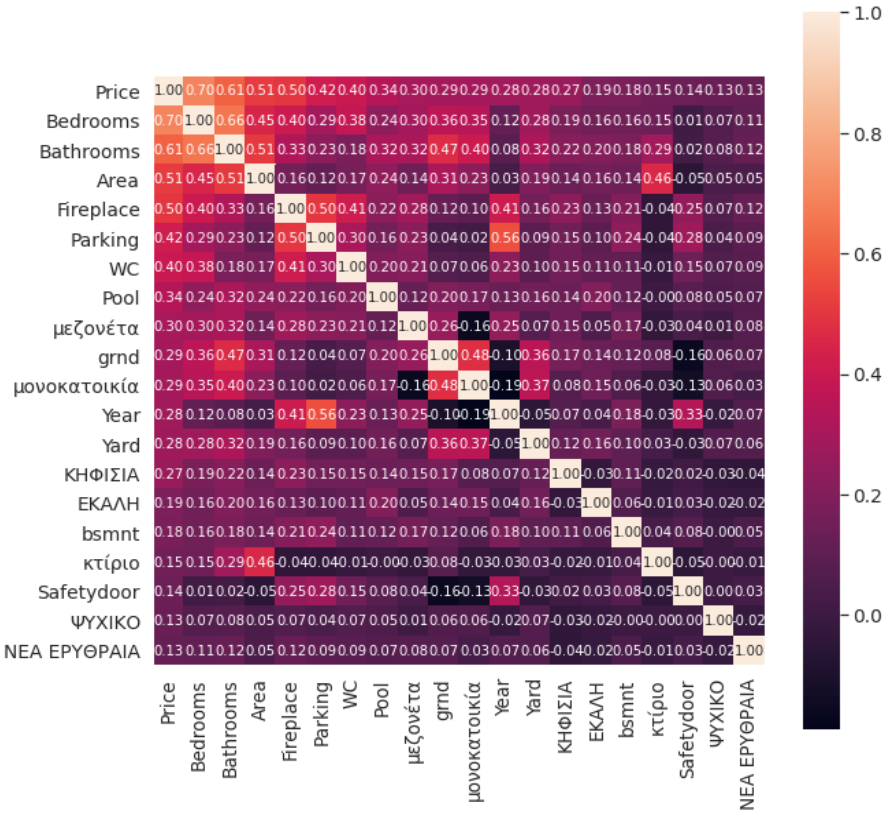
Προκειμένου να μπορούν να χρησιμοποιηθούν κατά το βέλτιστο δυνατό τρόπο όλα τα δεδομένα που έχουν συλλεχθεί, κρίνεται καίριο να μετατραπούν όλα τα κατηγορηματικά δεδομένα σε αριθμητικά. Η παρούσα ενότητα επιδιώκει να παρουσιάσει τους τρόπους με τους οποίος το παραπάνω υλοποιήθηκε στην συγκεκριμένη μελέτη.

Τα κατηγορηματικά χαρακτηριστικά που σκοπεύουμε να μετατρέψουμε σε αριθμητικά συνίστανται από τα πεδία Neighborhood, Housetype και Floor. Αρχίζοντας από τη μεταβλητή Neighborhood, η οποία περιγράφει την γειτονιά του κάθε ακινήτου, παρατηρούμε ότι διαθέτει 210 διαφορετικές τιμές, δηλαδή υποθετικά 210 διαφορετικές περιοχές. Μειώνουμε αυτόν τον αριθμό, καθώς διακρίνουμε ότι πολλές από αυτές τις τιμές αφορούν ίδιες περιοχές με διαφορετικό τρόπο γραφής και επιλέγουμε την ομαδοποίηση των τιμών του πεδίου, ανάλογα με την ευρύτερη περιοχή της Αττικής στην οποία ανήκει η κάθε ιδιοκτησία του συνόλου δεδομένων. Τελικά, καταλήγουμε σε 114 διαφορετικές τιμές για την μεταβλητή Neighborhood. Κάτι το οποίο παρατηρούμε και ήταν άλλωστε προβλεπόμενο, είναι το γεγονός ότι σε περιοχές όπως το Κολωνάκι, το Ψυχικό, η Κηφισιά και η Εκάλη οι τιμές των ακινήτων είναι αρκετά πιο υψηλές από ότι είναι σε περιοχές όπως τους Αμπελόκηπους, το Παγκράτι και τα Ιλίσια. Αυτό φαίνεται και από τον πίνακα συσχετίσεων των περιοχών των ακινήτων με το πεδίο της τιμής, ο οποίος παρατίθεται ακολούθως. Ο συγκεκριμένος πίνακας αφορά μόνο τις δέκα περιοχές με τον μεγαλύτερο συντελεστή συσχέτισης με την τιμή. Για το πεδίο Neighborhood χρησιμοποιήθηκε label encoding, με τη χρήση του οποίου κωδικοποιήσαμε την ετικέτα της γειτονιάς κάθε ακινήτου και τελικά οι διαφορετικές περιοχές αναπαριστάνονται από 114 διαφορετικούς ακεραίους. Μία ακόμα μέθοδος περιγραφής της κάθε περιοχής, η οποία δοκιμάστηκε στα πειράματά μας, ήταν η ομαδοποίηση των προαναφερθεισών 114 περιοχών σε 5 ομάδες: το κέντρο, τα βόρεια προάστια, τα νότια προάστια, τον Πειραιά και μία ομάδα που αφορούσε τις υπόλοιπες περιοχές, δηλαδή τα δυτικά και τα ανατολικά προάστια των Αθηνών.



Εικόνα 12 Πίνακας Συσχετίσεων Περιοχής Ακινήτου – Τιμής.

Συνεχίζοντας με την μεταβλητή Housetype, δημιουργήθηκαν πέντε διαφορετικά πεδία, κάθε ένα από τα οποία αντιστοιχεί σε ένα είδος ακινήτου. Εξήχθησαν, δηλαδή, οι ψευδομεταβλητές (dummy variables) του πεδίου. Η ίδια μέθοδος ακολουθήθηκε και για την μεταβλητή Floor, από την οποία προέκυψαν έντεκα διαφορετικές ψευδομεταβλητές. Ο πίνακας συσχετίσεων, όπως διαμορφώθηκε μετά από την επεξεργασία και την εκκαθάριση του συνόλου δεδομένων, των είκοσι πεδίων με τον μεγαλύτερο συντελεστή συσχέτισης με την τιμή πώλησης, παρατίθεται στην συνέχεια.



Εικόνα 13 Πίνακας συσχετίσεων των είκοσι πεδίων με την υψηλότερη συσχέτιση με την τιμή.

Επιπροσθέτως, σε αυτό το σημείο εξάγαμε 9 διαφορετικές ομάδες για τις τιμές του πεδίου Year. Έτσι, κάθε διάστημα χρονολογικό έτος από το 1940 έως το 2019 αντικαθίσταται με το χρονικό διάστημα στο οποίο ανήκει. Αυτό συμβάλλει στην βελτίωση των αποτελεσμάτων των αλγορίθμων που θα χρησιμοποιήσουμε στην συνέχεια, καθώς ο όγκος των δεδομένων δεν είναι αρκετός, προκειμένου να είμαστε σε θέση να κρατήσουμε κάθε μία ξεχωριστή χρονολογία και ταυτοχρόνως να έχουμε ικανοποιητικά αποτελέσματα.

3.3.5 Έλεγχος Στατιστικών Υποθέσεων

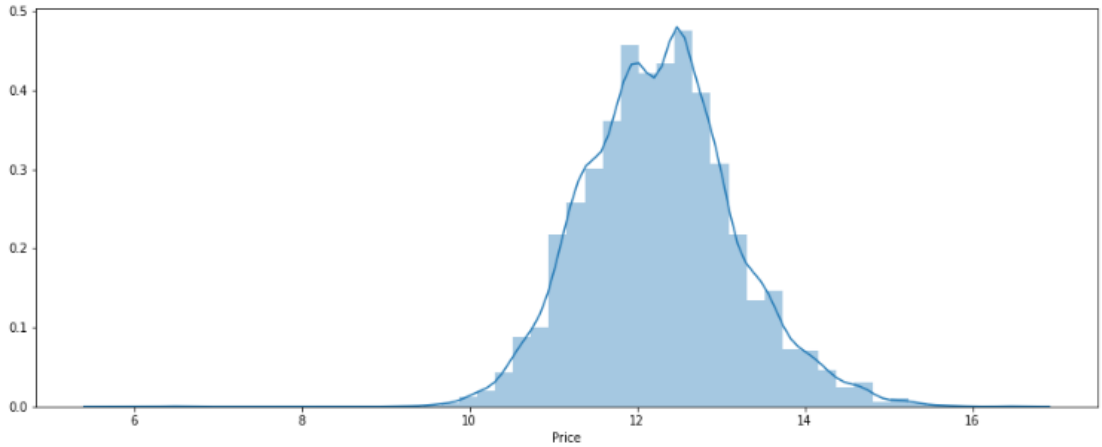
Τα προγενέστερα βήματα που ακολουθήθηκαν, της διαχείρισης των κενών και ακραίων τιμών, επιχείρησαν να «καθαρίσουν» τα δεδομένα που έχουμε στη διάθεσή μας και να τα μετατρέψουν σε μία μορφή περισσότερο κατάλληλη για πολυμεταβλητή ανάλυση. Ο έλεγχος, τώρα, των δεδομένων, προκειμένου να εξεταστούν οι στατιστικές υποθέσεις που αφορούν τις τεχνικές πολυμεταβλητής ανάλυσης, ασχολείται με τις βάσεις που πρέπει να υπάρχουν και στις οποίες θα στηριχτεί η στατιστική ανάλυση. Όπως αναφέρει και ο Hair (2009)[43], η ανάγκη του ελέγχου των στατιστικών υποθέσεων στις περιπτώσεις όπου στην συνέχεια θα εφαρμοστεί πολυμεταβλητή ανάλυση, όπως στην περίπτωση του πλαισίου της συγκεκριμένης μελέτης, αυξάνεται εξαιτίας δύο χαρακτηριστικών της. Πρώτον, η πολυπλοκότητα των σχέσεων μεταξύ των μεταβλητών, η οποία αποτελεί απόρροια της, τυπικά, χρήσης πολλών μεταβλητών, αυξάνει την πιθανότητα ύπαρξης στρεβλώσεων και προκαταλήψεων μεταξύ των δεδομένων. Δεύτερον, η πολυπλοκότητα της ανάλυσης και των αποτελεσμάτων, ενδέχεται να υποκρύψει τους δείκτες οι οποίοι προμηνύουν παραβιάσεις των υποθέσεων, οι οποίες είναι πιθανότατα πιο εμφανείς στην απλούστερη μονομεταβλητή ανάλυση. Εν πάση περιπτώσει, τα μοντέλα πολυμεταβλητής ανάλυσης θα υπολογίσουν αποτελέσματα ακόμα και όταν οι στατιστικές υποθέσεις οι οποίες αφορούν τα δεδομένα δεν ικανοποιούνται. Για το λόγο αυτό, ο ερευνητής είναι σημαντικό να λαμβάνει υπόψιν του αυτές τις παραβιάσεις και τις επιπτώσεις που μπορεί να έχουν στην διαδικασία εκτίμησης και την ερμηνεία των αποτελεσμάτων. Στην συνέχεια του κεφαλαίου θα εξεταστούν οι στατιστικές υποθέσεις της κανονικότητας, της ομοσκεδαστικότητας, της γραμμικότητας και της απουσίας συσχετισμένων σφαλμάτων.

3.3.5.1 Κανονικότητα

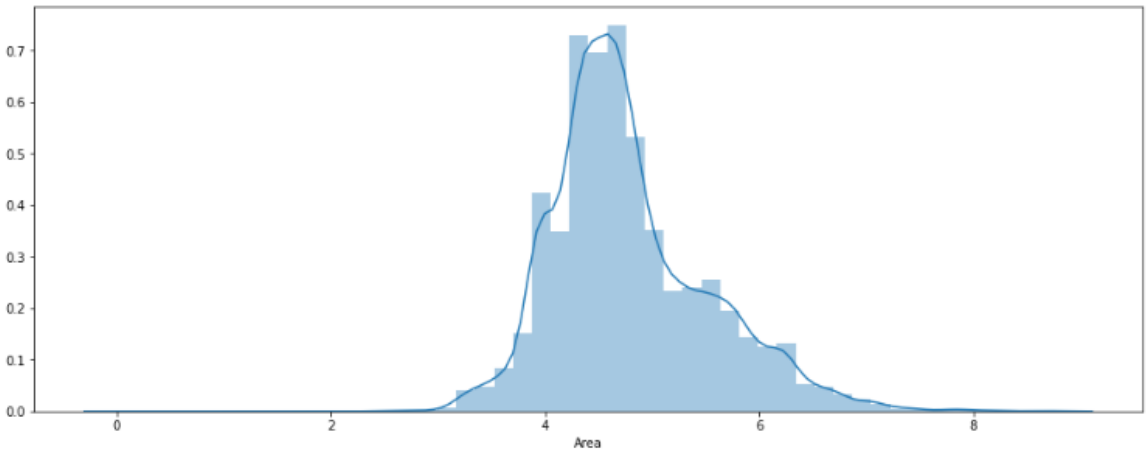
Με τον όρο κανονικότητα εννοούμε ότι η κατανομή του κάθε πεδίου ξεχωριστά επιθυμούμε να συγκλίνει στην κανονική κατανομή. Αποτελεί την σημαντικότερη εκ των υποθέσεων, καθώς αν η απόκλιση από την κανονική κατανομή είναι μεγάλη, όλα τα αποτελέσματα των στατιστικών ελέγχων είναι μη έγκυρα, καθώς η κανονικότητα απαιτείται προκειμένου να χρησιμοποιηθούν οι F και t στατιστικές υποθέσεις. Στην περίπτωση της μίας μεταβλητής, η κανονικότητα είναι εύκολο να ελεγχθεί με την βοήθεια του ιστογράμματος. Παρόλο που για τα μεγάλα σύνολα δεδομένων η ύπαρξη κανονικότητας είναι πολλές φορές δεδομένη, καθένας οφείλει να ελέγχει την κανονικότητα όλων των συνεχών αριθμητικών μεταβλητών που συμμετέχουν στην ανάλυση.

Στην συνέχεια, θα ελέγξουμε την ύπαρξη της κανονικότητας για τις μεταβλητές της τιμής, του εμβαδού και της αυλής. Με αφετηρία το πεδίο της τιμής των ακινήτων, όπως αναφέρθηκε και στο κεφάλαιο 3.3.1, βλέπουμε ότι η κατανομή της τιμή αποκλίνει αρκετά από την κανονική, με αρκετά μεγάλη θετική λοξότητα και κύρτωση. Το ίδιο παρατηρείται και για την μεταβλητή του εμβαδού και της αυλής, για τις τιμές μεγαλύτερες της μονάδας. Τη λύση στο πρόβλημα αυτό έρχονται να δώσουν οι μετασχηματισμοί δεδομένων. Οι

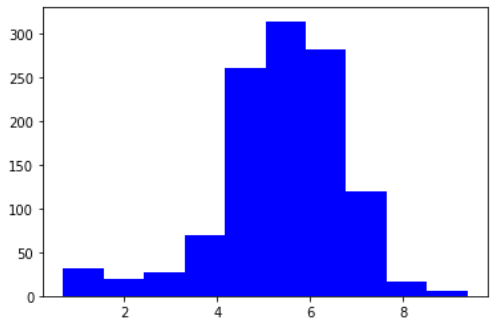
τελευταίοι παρέχουν ένα μέσο τροποποίησης των δεδομένων, με σκοπό την διόρθωση της παραβίασης στατιστικών υποθέσεων και την βελτίωση της συσχέτισης μεταξύ των εμπλεκόμενων μεταβλητών. Στην συγκεκριμένη εργασία και εφόσον τα δεδομένα μας είχαν θετικές λοξότητες, χρησιμοποιήθηκε για τις τρεις παραπάνω μεταβλητές ο λογαριθμικός μετασχηματισμός. Τα αποτελέσματα της εφαρμογής αυτού του μετασχηματισμού στις κατανομές των προαναφερθέντων πεδίων διακρίνονται στην συνέχεια. Όσον αφορά την μεταβλητή που περιγράφει την ύπαρξη και το μέγεθος της αυλής κάθε ακινήτου, δοκιμάστηκε αρχικά η εφαρμογή του λογαριθμικού μετασχηματισμού για τις τιμές μεγαλύτερες της μονάδας και εξετάστηκαν στην συνέχεια της μελέτης, τα αποτελέσματα των μοντέλων με την αυλή σε διάφορες μορφές (με μετασχηματισμό ή χωρίς, με συσταδοποίηση κτλ.), τα οποία θα παρουσιαστούν σε επόμενη ενότητα.



Εικόνα 14 Κατανομή της Τιμής Ακινήτων Μετά την Εφαρμογή του Λογαριθμικού Μετασχηματισμού.



Εικόνα 15 Κατανομή του Εμβαδού Ακινήτων Μετά την Εφαρμογή του Λογαριθμικού Μετασχηματισμού.

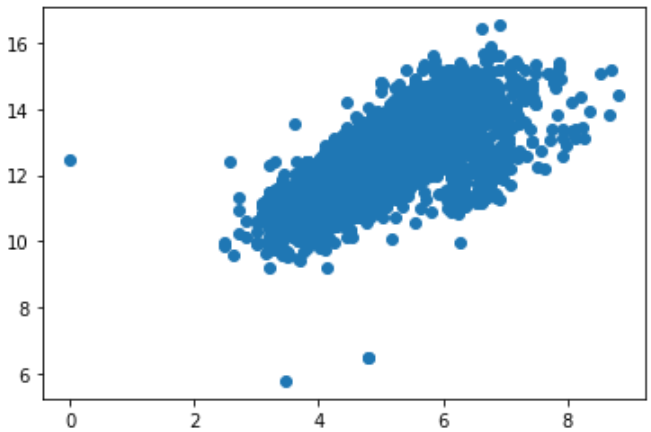


Εικόνα 16 Κατανομή του Εμβαδού της Αυλής Ακινήτων Μετά την Εφαρμογή του Λογαριθμικού Μετασχηματισμού.

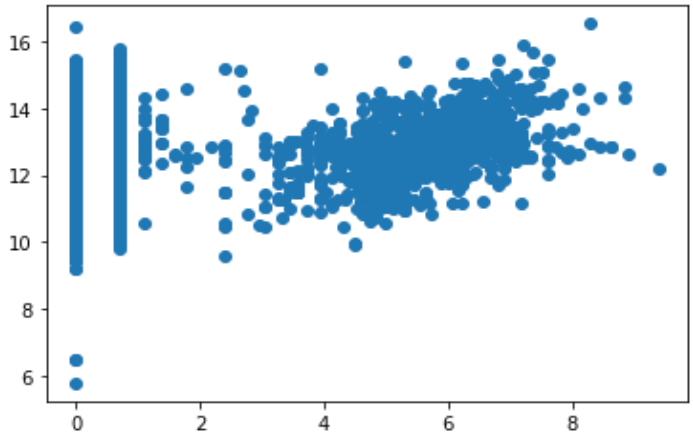
Η τιμή της λοξότητας για κάθε μία από τις παραπάνω μεταβλητές μετά την εφαρμογή του μετασχηματισμού είναι 0,211, 0,219 και -1,0269 για την τιμή, το εμβαδόν ακινήτου και το εμβαδόν αυλής αντίστοιχα. Μπορούμε να παρατηρήσουμε ότι ο λογαριθμικός μετασχηματισμός πράγματι ευνόησε την κανονική μορφή των κατανομών των μεταβλητών μας. Ακολουθεί η μελέτη της υπόθεσης της ομοσκεδαστικότητας για τις ίδιες μεταβλητές.

3.3.5.2 Ομοσκεδαστικότητα

Η υπόθεση της ομοσκεδαστικότητας σχετίζεται κυρίως με τις σχέσεις εξάρτησης μεταξύ των μεταβλητών. Στηρίζεται στην υπόθεση ότι οι εξαρτημένες μεταβλητές παρουσιάζουν ίδια επίπεδα διακύμανσης στο εύρος τιμών των ανεξάρτητων μεταβλητών. Η ομοσκεδαστικότητα είναι επιθυμητή, καθώς η διακύμανση της εξαρτημένης μεταβλητής που επεξηγείται από την σχέση εξάρτησης, δεν πρέπει να συγκεντρώνεται μόνο σε μία περιορισμένη περιοχή των τιμών των ανεξάρτητων μεταβλητών. Αν το παραπάνω δεν τηρείται, η σχέση μεταξύ των ανεξάρτητων και της εξαρτημένης μεταβλητής ονομάζεται ετεροσκεδαστική. Αυτό έχει ως αποτέλεσμα οι προβλέψεις του μοντέλου να είναι καλύτερες για συγκεκριμένες τιμές των ανεξάρτητων μεταβλητών, έναντι άλλων. Η ύπαρξη της ομοσκεδαστικότητας εξασφαλίζει ίδια τιμή σφάλματος σε όλο το εύρος των ανεξάρτητων μεταβλητών. Η υπόθεση αυτή, ελέγχεται με απεικονίσεις των σχέσεων της τιμής ακινήτου, Price, η οποία στην συγκεκριμένη περίπτωση αποτελεί την εξαρτημένη μεταβλητή, με τις συνεχείς αριθμητικές μεταβλητές Area και Yard, όπως παρουσιάζονται στην συνέχεια.



Εικόνα 17 Διάγραμμα Διασποράς Εμβαδού Ακινήτου – Τιμής Ακινήτου.



Εικόνα 18 Διάγραμμα Διασποράς Εμβαδού Αυλής Ακινήτου – Τιμής Ακινήτου.

Αυτό που παρατηρείται από τα παραπάνω διαγράμματα διασποράς είναι ότι ενώ προηγούμενα διαγράμματα διασποράς, τα οποία απεικόνιζαν την σχέση της τιμής με την μεταβλητή του εμβαδού ακινήτων και του εμβαδού της αυλή ακινήτου, είχαν κωνικό σχήμα (βλ. Σχήμα 3.4), τώρα τα διαγράμματα αυτά έχουν σχήματα με πολύ πιο ομοιόμορφη κατανομή της διακύμανσης της τιμής που εξηγείται από τις σχέσεις μας. Αυτό επιτεύχθηκε με την εφαρμογή του λογαριθμικού μετασχηματισμού και την εξασφάλιση της κανονικότητας των κατανομών. Βέβαια, παρόλο που η διακύμανση φαίνεται να διατηρείται περισσότερο σταθερή, συνεχίζουν να υπάρχουν κάποιες αποκλίσεις.

3.3.5.3 Γραμμικότητα

Η σιωπηρή υπόθεση όλων των πολυμεταβλητών μεθόδων, οι οποίες βασίζονται σε συσχετιστικά μέτρα σύνδεσης, συμπεραλαμβανομένης και της πολλαπλής παλινδρόμησης, είναι η γραμμικότητα. Οι συσχετίσεις των μεταβλητών αντικατοπτρίζουν μόνο τη γραμμική συσχέτιση μεταξύ των μεταβλητών και συνεπώς οποιαδήποτε μη γραμμική σχέση δεν εξηγείται από αυτές. Αυτή η παράλειψη οδηγεί σε μη – ακριβή εκτίμηση, και μάλιστα υποεκτίμηση, της πραγματικής ισχύος των μεταξύ τους σχέσεων. Είναι συνετό, να εξετασθούν όλες οι σχέσεις μεταξύ των μεταβλητών, προκειμένου να αναγνωριστούν αποκλίσεις από τη γραμμικότητα, που ενδεχομένως να επηρεάσουν το αποτέλεσμα.

Στην παρούσα εργασία, όπως φαίνεται και από τα διαγράμματα διασποράς που έχουν παρατεθεί παραπάνω, οι σχέσεις μεταξύ των μεταβλητών μας παρουσιάζουν γραμμική συμπεριφορά, και για το λόγο αυτό δεν θα χρειαστεί να προχωρήσουμε σε περαιτέρω μετασχηματισμούς των δεδομένων μας, προκειμένου να ικανοποιήσουμε την υπόθεση της γραμμικότητας. Συνεχίζουμε με την μελέτη της απουσίας συσχετισμένων σφαλμάτων.

3.3.5.4 Απουσία Συσχετισμένων Σφαλμάτων

Οι προβλέψεις που προκύπτουν από την εφαρμογή οποιασδήποτε μεθόδου, η οποία στηρίζεται στην εξάρτηση των δεδομένων, σε σπάνιες περιπτώσεις είναι αλάνθαστες, παρόλα αυτά οφείλουμε να προσπαθήσουμε για την εξασφάλιση της μη συσχέτισης των σφαλμάτων που θα προκύψουν. Εδώ, αναζητούμε μοτίβα που μπορεί να παρουσιάζουν τα σφάλματα, τα οποία μαρτυρούν συστηματικές σχέσεις μεταξύ των μεταβλητών, τις οποίες δεν έχουμε ακόμα εντοπίσει. Στην παρούσα διπλωματική εργασία τα δεδομένα συλλέχθηκαν με τέτοιον τρόπο, έτσι ώστε να μπορούμε με αυτοπεποίθηση να ισχυριστούμε

ότι αυτός δεν συντέλεσε στην προσθήκη κάποιου συστηματικού σφάλματος, σε κάποιο τμήμα των δεδομένων. Επιπλέον, το γεγονός ότι το σύνολο δεδομένων δεν αποτελείται από δεδομένα μορφής χρονοσειρών, έχει ως συνέπεια την ανεξαρτησία μεταξύ των δεδομένων και άρα την εξασφάλιση της ανεξαρτησίας των σφαλμάτων τους.

3.3.6 Τελικά Σύνολα Δεδομένων

Μετά την εφαρμογή των τεχνικών που αναλύθηκαν σε αυτήν την ενότητα, η μορφή του αρχικού συνόλου δεδομένων είναι η εξής: αποτελείται από 14.842 εγγραφές, 25 στήλες και συνίσταται από τα πεδία που φαίνονται στον παρακάτω πίνακα.

Όνομα Πεδίου	**Τύπος**	**Περιγραφή**
Price	Float	Τιμή ακινήτου
Area	Float	Εμβαδόν ακινήτου
Year	Int	Έτος κτίσης
Apartment	Int	Διαμέρισμα
FloorAp	Int	Οροφοδιαμέρισμα
House	Int	Μονοκατοικία
Maisonette	Int	Μεζονέτα
Building	Int	Κτίριο
Neighborhood	Int	Περιοχή Ακινήτου
Bedrooms	Int	Αριθμός υπνοδωματίων
Bathrooms	Int	Αριθμός μπάνιων
WC	Int	Αριθμός WC
Floor_-2-Floor_8	Int	Όροφος ακινήτου
Renovation	Int	Ανακαίνιση
Penthouse	Int	Ρετιρέ
Fireplace	Int	Τζάκι
Pool	Int	Πισίνα
Safetydoor	Int	Πόρτα ασφαλείας
Transport	Int	Πλησίον ΜΜΜ
Market	Int	Πλησίον αγοράς
School	Int	Πλησίον σχολείου
Park	Int	Πλησίον Πάρκου
Hospital	Int	Πλησίον νοσοκομείου
Parking	Int	Ύπαρξη πάρκινγκ
Furniture	Int	Επιπλωμένο ακίνητο
Yard	Float	Ύπαρξη αυλής/κήπου
Elevator	Int	Ύπαρξη ασανσέρ

Πίνακας 2 Σύνολο Χαρακτηριστικών μετά την επεξεργασία.

Για την μεταβλητή Neighborhood έχουμε εφαρμόσει label encoding, έτσι ώστε κάθε διαφορετική περιοχή να περιγράφεται από έναν μοναδικό ακέραιο αριθμό και για την μεταβλητή Floor έχουμε εξάγει τις ψευδομεταβλητές της, για τις οποίες το κάθε ακίνητο λαμβάνει την τιμή 1 σε μία μόνο από αυτές. Για παράδειγμα εάν το ακίνητο βρίσκεται στον

πρώτο όροφο, τότε η μεταβλητή Floor_1 θα λάβει την τιμή 1 και οι υπόλοιπες μεταβλητές Floor_-2-Floor_8 θα λάβουν μηδενική τιμή.

3.3.6.1 Επεξεργασία Συνόλου Δεδομένων

Το παραπάνω σύνολο χαρακτηριστικών υπέστη ορισμένες μετατροπές, και διάφορες μορφές αυτού δοκιμάστηκαν ως είσοδοι στα μοντέλα μηχανικής μάθησης που χρησιμοποιήσαμε, προκειμένου να ερευνηθούν τα διαφορετικά αποτελέσματα που μπορούν να ληφθούν.

Αρχικά, για να αντιμετωπίσουμε τον μη επαρκή αριθμό των δεδομένων που είχαμε στην διάθεση μας, σε σχέση με τον αριθμό των χαρακτηριστικών κινηθήκαμε ως εξής: Αφού μελετήθηκε η συσχέτιση των χαρακτηριστικών με την τιμή κάθε ακινήτου, δοκιμάσαμε σαν είσοδο στους αλγορίθμους μας διαφορετικούς συνδυασμούς των χαρακτηριστικών που διέθεταν υψηλά επίπεδα συσχέτισης με την τιμή, ξεκινώντας από έναν μικρό αριθμό χαρακτηριστικών και προσθέτοντας επιπλέον χαρακτηριστικά όσο με την προσθήκη τους παρατηρείται βελτίωση των αποτελεσμάτων. Επιπροσθέτως, για κάθε ένα από τα χαρακτηριστικά, το οποίο δοκιμάστηκε σαν μέλος του συνόλου εισόδου στα μοντέλα μας, μελετήθηκε και επιλέχθηκε η βέλτιστη μορφή του, η οποία εξυπηρετούσε τους στόχους μας, καθώς επίσης και ο βέλτιστος συνδυασμός αυτών. Αναλυτικότερα, θα εξηγήσουμε παρακάτω ποιες μορφές εξετάσθηκαν, πέραν της αρχικής, για κάθε ένα από τα γνωρίσματα του συνόλου δεδομένων.

Τα χαρακτηριστικά Bedrooms, Bathrooms, WC, Fireplace, Pool, καθώς και τα υπόλοιπα χαρακτηριστικά που αρχικά είχαν δυαδική μορφή κρατήθηκαν ως ήταν. Για τις μεταβλητή Area δοκιμάστηκε ομαδοποίηση και κανονικοποίηση. Για τη μεταβλητή Year, επίσης δοκιμάστηκαν ομαδοποιήσεις σε διάφορους αριθμούς ομάδων, με βέλτιστο αυτόν που αναφέρεται στο προηγούμενο υποκεφάλαιο. Όμοια και για τη μεταβλητή Neighborhood. Για τη μεταβλητή Price ελέγχθηκε η κανονικοποίηση. Οι τεχνικές κανονικοποίησης των αριθμητικών δεδομένων, που εξετάσθηκαν, είναι η κανονικοποίηση λογαρίθμου και η κανονικοποίηση ελαχίστου – μεγίστου (MinMax Scaler). Η μεταβλητή Yard ομαδοποιήθηκε, καθώς τα δεδομένα ήταν ανεπαρκή για την διατήρηση της αρχικής της μορφής. Τέλος, μεταβλητή Renovation εκφράστηκε ως δυαδική, λαμβάνοντας, δηλαδή, την τιμή 1 για τα ακίνητα στα οποία έγινε ανακαίνιση και την τιμή 0 για όσα δεν έχει γίνει ανακαίνιση. Οι παραπάνω περιπτώσεις μελετήθηκαν και εφαρμόσθηκαν ως είσοδοι στα μοντέλα που αναλύθηκαν στο Κεφάλαιο 2. Τα αποτελέσματα αυτών, καθώς επίσης και η ερμηνεία τους, παρατίθενται στο επόμενο Κεφάλαιο.

3.3.6.1 Δυσκολίες που αντιμετωπίσαμε

Στη συνέχεια, θα αναφερθούμε στα προβλήματα τα οποία αντιμετωπίσαμε, κατά την πραγματοποίηση της παρούσας εργασίας, και συγκεκριμένα σε προβλήματα που αφορούν την συλλογή των δεδομένων και την κατασκευή του συνόλου δεδομένων. Το κύριο πρόβλημα που συναντήσαμε, αφορούσε την «ποιότητα» των δεδομένων, τα οποία είχαμε την δυνατότητα να εντοπίσουμε και να συλλέξουμε. Αρχικά, τα ακίνητα και οι περιγραφές τους στον ιστότοπο της Χρυσής Ευκαιρίας, καταχωρούνται και ενημερώνονται από τους χρήστες, χωρίς αυστηρούς κανόνες και μορφοποιήσεις, αλλά και χωρίς να γίνεται κάποιος έλεγχος για την εγκυρότητα των στοιχείων των ακινήτων. Αυτό σημαίνει πως ο καθένας μπορεί να δημοσιεύσει κάποιο ακίνητο προς πώληση, με οποιαδήποτε στοιχεία αυτός επιθυμεί, ακόμα και αν δεν συμπίπτουν με την ιδιοκτησία. Τέτοιες περιπτώσεις ακινήτων

των οποίων η περιγραφή είναι σε ελλιπής, συναντήσαμε σε αρκετά μεγάλο αριθμό καταχωρήσεων. Τα χαρακτηριστικά των καταχωρήσεων αυτών, όπου ήταν δυνατόν διορθώσαμε, παρόλα αυτά υπήρξαν πολλές εγγραφές που διαγράφηκαν από το σύνολο δεδομένων εξαιτίας των ελλιπών τους στοιχείων. Ακόμα, κάθε χρήστης μπορεί να δημοσιεύσει το ακίνητο που επιθυμεί, περισσότερες εκ της μίας φορές, γεγονός που οδήγησε στο να συλλέξουμε μεγάλο ποσοστό διπλοτύπων, κάτι που καθυστέρησε ακόμα περισσότερο την δημιουργία του συνόλου δεδομένων. Επίσης, η μη αναφορά σημαντικών μεταβλητών στις αγγελίες των ακινήτων, οι οποίες έχουν υψηλό ποσοστό επιρροής στην τιμή τους, όπως είναι το επίπεδο πολυτέλειας του ακινήτου ή η κατάσταση των ηλεκτρικών συσκευών του επιδρούν αρνητικά στην ακρίβεια των προβλέψεων. Εδώ θα αναφερθούμε για ακόμα μία φορά στην δυσκολία συλλογής των δεδομένων που χρησιμοποιήθηκαν, διαδικασία η οποία κρίθηκε πολύ χρονοβόρα και για το λόγο αυτό ο όγκος των δεδομένων μας δεν είναι ο επιθυμητός. Αντί αυτού, είναι σημαντικά λιγότερες, από ότι αρχικά υπολογίζαμε, οι εισοδοι του συνόλου δεδομένων μας. Το γεγονός αυτό έχει σημαντική επίπτωση στην απόδοση των μοντέλων που χρησιμοποιήσαμε και στον αριθμό των χαρακτηριστικών που τελικά λήφθηκαν υπόψιν κατά την εκπαίδευση των μοντέλων.

Ορισμένα από τα αναφερθέντα προβλήματα θα μπορούσαν να βελτιωθούν με την καλύτερη οργάνωση των ακινήτων που δημοσιεύονται, την εισαγωγή πιο αυστηρών κανόνων για τους χρήστες της ιστοσελίδας και την απαίτηση ενός σεβαστού αριθμού χαρακτηριστικών τα οποία θα αποτελούν απαραίτητη προϋπόθεση, προκειμένου να δημοσιευθεί μία ιδιοκτησία προς πώληση.

Κεφάλαιο 4

Αποτελέσματα Αλγορίθμων και Συγκρίσεις

Το παρόν κεφάλαιο πραγματεύεται την παρουσίαση, ανάλυση και σύγκριση των αποτελεσμάτων που λήφθηκαν, μετά την εφαρμογή των μεθόδων και αλγορίθμων Μηχανικής Μάθησης, που παρουσιάστηκαν αναλυτικά στο Κεφάλαιο 2. Κάθε ένα από αυτά τα μοντέλα, εκπαιδεύεται στα σύνολα δεδομένων που είδαμε στην προηγούμενη ενότητα. Αφού παρουσιαστούν επεξηγηματικά οι μετρικές με βάση τις οποίες αξιολογήθηκε η επίδοση των αλγορίθμων, θα παραθέσουμε τα αποτελέσματα, την ερμηνεία τους και κατόπιν θα προχωρήσουμε στις συγκρίσεις μεταξύ των μοντέλων που υλοποιήθηκαν.

4.1 Μετρικές Αξιολόγησης

Οι μετρικές που αξιοποιήθηκαν για την αξιολόγηση των αποτελεσμάτων των μοντέλων Μηχανικής Μάθησης που υλοποιήθηκαν παρουσιάζονται αναλυτικά παρακάτω.

Πρώτον, μετρήθηκε η ποσότητα R^2, η οποία ονομάζεται *συντελεστής προσδιορισμού*, και εκφράζει το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής ή μεταβλητής απόκρισης, το οποίο μπορεί να εξηγήσει αποτελεσματικά και να προβλέψει το μοντέλο, δηλαδή οι ανεξάρτητες μεταβλητές ή μεταβλητές εισόδου. Με μαθηματικές σχέσεις ορίζεται ως:

$$R^2 = 1 - \frac{SSres}{SStot},$$

όπου το $SSres$ εκφράζει το άθροισμα των τετραγώνων των σφαλμάτων μεταξύ της πραγματικής μεταβλητής απόκρισης και της προβλεφθείσας τιμής της και το $SStot$ εκφράζει το άθροισμα της διαφοράς των πραγματικών τιμών της μεταβλητής απόκρισης από την δειγματική μέση τιμή της.

Εν συνεχεία μετρήθηκε η ποσότητα Huber Loss, η οποία αποτελεί μία συνάρτηση κόστους, λιγότερο ευαίσθητη στις ακραίες τιμές από ότι είναι η συνάρτηση κόστους του μέσου τετραγωνικού σφάλματος (Mean Square Error). Το Huber Loss εκφράζεται μαθηματικά ως:

$$L_\delta\big(y, f(x)\big) = \begin{cases} \frac{1}{2}(y - f(x))^2 & ,για\ |y - f(x)| < \delta \\ \delta|y - f(x)| - \frac{1}{2}\delta^2, & διαφορετικά. \end{cases}$$

Ακόμη ένας παράγοντας τον οποίο λάβαμε υπόψιν στην αξιολόγηση των αποτελεσμάτων, είναι ο στατιστικός δείκτης του μέσου απόλυτου σφάλματος (Mean Absolute Error). Το μέσο απόλυτο σφάλμα εκφράζει ένα μέτρο σφαλμάτων μεταξύ όμοιων παρατηρήσεων του ίδιου φαινομένου. Το μέσο απόλυτο σφάλμα υπολογίζεται μαθηματικά ως εξής:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Τέλος, μία ακόμη μετρική αξιολόγησης των μοντέλων, που θα χρησιμοποιήσουμε είναι το Area Under the Curve (AUC), με μία τροποποίηση για να μπορεί να εφαρμοστεί στο είδος των δεδομένων της παρούσας εφαρμογής. Η καμπύλη ROC αναπαριστά την διαγνωστική ικανότητα ενός δυαδικού ταξινομητή, όσο το διακριτικό όριο μεταβάλλεται. Δημιουργείται, απεικονίζοντας το True Positive Rate (TPR), ή αλλιώς ευαισθησία, έναντι του False Positive Rate (FPR), σε διάφορα όρια ταξινόμησης. Η μείωση αυτού του ορίου, έχει ως αποτέλεσμα την αύξηση των αντικειμένων που ταξινομούνται ως θετικά, ενώ η μείωσή του έχει το αντίθετο αποτέλεσμα. Η περιοχή κάτω από την καμπύλη ROC είναι αυτή που ορίζεται ως AUC και παρέχει ένα συνολικό μέτρο απόδοσης της ταξινόμησης. Ένας τρόπος να ερμηνευθεί είναι η πιθανότητα ότι το μοντέλο κατατάσσει ένα τυχαίο θετικό παράδειγμα υψηλότερα από ότι ένα τυχαίο αρνητικό παράδειγμα. Προφανώς η μετρική αυτή αφορά, όπως είπαμε δυαδικούς ταξινομητές και όχι μοντέλα πρόβλεψης μίας συνεχούς τιμής, όπως συμβαίνει στην περίπτωση του ζητουμένου της εργασίας μας. Για το λόγο αυτό, προκειμένου να μπορούμε να το χρησιμοποιήσουμε, εισάγουμε μία νέα ερμηνεία του AUC, η οποία ορίζεται ως εξής: έχοντας ταξινομήσει τις τιμές των ακινήτων των συνόλων εισόδου και εξόδου των μοντέλων μας, διατηρώντας τους δείκτες των εγγραφών, υπολογίζουμε το AUC score ως το ποσοστό των αντιστροφών που πρέπει να γίνουν στον πίνακα εξόδου προκειμένου να ταυτιστεί η ταξινόμηση των εγγραφών του με τον αντίστοιχο πίνακα εισόδου. Δηλαδή:

$$AUC = 1 - \frac{Number\ of\ Inversions}{Maximum\ Number\ of\ Inversions}$$

4.2 Αποτελέσματα Αλγορίθμων

Σε αυτήν την ενότητα θα παρουσιαστούν τα αποτελέσματα των μετρικών που αναφέρθηκαν, για τους αλγόριθμους Μηχανικής Μάθησης που χρησιμοποιήθηκαν για την κατασκευή των μοντέλων, που μας έδωσαν τα βέλτιστα αποτελέσματα. Αυτά τα αποτελέσματα προέκυψαν από την εκπαίδευση των μοντέλων, με το 80% των συνολικών δεδομένων να αποτελεί το σύνολο εκπαίδευσης και το 20% το σύνολο ελέγχου. Στα σύνολα αυτά εφαρμόστηκε μείωση των χαρακτηριστικών, προκειμένου να ληφθούν υπόψιν μόνο οι βέλτιστες παράμετροι και δεν εισήχθησαν ως είσοδοι των μοντέλων όλες οι ανεξάρτητες μεταβλητές. Τα σύνολα που προέκυψαν μετά την περαιτέρω επεξεργασία, όπως και ο τρόπος που οδηγηθήκαμε σε αυτά, παρουσιάζονται, επίσης, στη συνέχεια.

4.2.1 Light Gradient Boosting Machine (LGBM)

Το Light Gradient Boosting Machine (LGBM) είναι μία δομή, με τη βοήθεια της οποία υλοποιείται ενίσχυση κλίσης (gradient boosting), η οποία βασίζεται στην κατασκευή αλγορίθμων εκπαίδευσης που χρησιμοποιούν δέντρα αποφάσεων. Διαφέρει από άλλους αλγορίθμους, οι οποίοι κατασκευάζουν το μοντέλο τους με την χρήση δέντρων απόφασης, στο γεγονός ότι τα δέντρα μεγαλώνουν στην κατακόρυφη διεύθυνση αντί για την οριζόντια.

Παρακάτω παρατίθενται τα αποτελέσματα αυτού του αλγορίθμου, για τα δύο σύνολα δεδομένων, τα οποία μας έδωσαν την καλύτερη απόδοση. Ξεκινάμε με το dataset1, το οποίο αποτελεί το αρχικό και απλοποιημένο σύνολο δεδομένων, που περιέχει τα χαρακτηριστικά με τη μεγαλύτερη συσχέτιση με την τιμή.

Metric	**Train Set**	**Test Set**
R^2	0.7518	0.7143
MAE	59686.56	62826.77
Huber Loss	59686.06	62826.27
AUC	75,06%	74,1%

Πίνακας 3 Αποτελέσματα αρχικού συνόλου δεδομένων Lgbm.

Συνεχίζουμε, με το δεύτερο σύνολο δεδομένων, dataset2, στο οποίο έχουμε προσθέσαμε ένα – ένα χαρακτηριστικά, όσο βελτιωνόταν η απόδοση των μοντέλων και έχουμε ομαδοποιήσει την μεταβλητή Neighborhood, με βάση την ευρύτερη περιοχή της Αττικής, στην οποία βρίσκεται κάθε ακίνητο.

Metric	**Train Set**	**Test Set**
R^2	0.7787	0.7253
MAE	56274.31	61698.31
Huber Loss	56273.81	61697.81
AUC	76,73%	75,62%

Πίνακας 4 Αποτελέσματα τελικού συνόλου δεδομένων Lgbm.

4.2.2 Παλινδρόμηση Ενίσχυσης Κλίσης (Gradient Boosting Regressor)

Για την παλινδρόμηση ενίσχυσης κλίσης, λαμβάνουμε αντίστοιχα για το πρώτο σύνολο δεδομένων, dataset1:

Metric	**Train Set**	**Test Set**
R^2	0.7763	0.72205
MAE	55607.13	62099.56
Huber Loss	55606.63	62099.06
AUC	77,66%	74,7%

Πίνακας 5 Αποτελέσματα αρχικού συνόλου δεδομένων, Gradient Boosting.

Συνεχίζουμε με το δεύτερο σύνολο δεδομένων, dataset2:

Metric	Train Set	Test Set
R^2	0.7813	0.7259
MAE	55114.56	61285.89
Huber Loss	55114.06	61285.39
AUC	76,52%	75,96%

Πίνακας 6 Αποτελέσματα τελικού συνόλου δεδομένων, Gradient Boosting.

4.2.3 Τυχαία Δάση (Random Forest)

Τέλος, αποτυπώνονται στους επόμενους πίνακες τα αποτελέσματα που μας έδωσε ο αλγόριθμος των τυχαίων δασών για τα δύο σύνολα δεδομένων. Για το dataset1:

Metric	Train Set	Test Set
R^2	0.7511	0.7119
MAE	58768.01	63742.54
Huber Loss	58767.51	63742.04
AUC	76,41%	75,44%

Πίνακας 7 Αποτελέσματα αρχικού συνόλου δεδομένων, Τυχαία Δάση.

Για το dataset2:

Metric	Train Set	Test Set
R^2	0.7579	0.7256
MAE	57664.41	61266.2
Huber Loss	57663.91	61265.7
AUC	77,03%	77,45%

Πίνακας 8 Αποτελέσματα τελικού συνόλου δεδομένων, Τυχαία Δάση.

4.3 Ερμηνεία Αποτελεσμάτων

Από τα αποτελέσματα που παρουσιάστηκαν στην ενότητα 4.2, γίνεται αντιληπτό ότι η συμπεριφορά των αλγορίθμων δεν διαφέρει σε σημαντικό βαθμό από το ένα σύνολο δεδομένων στο άλλο, αλλά και από αλγόριθμο σε αλγόριθμο. Παράλληλα, φαίνεται ότι οι τιμές των μετρικών που προκύπτουν, δεν είναι υψηλές και για το λόγο αυτό οι αποκλίσεις των πραγματικών από τις προβλεφθείσες τιμές είναι αρκετά μεγάλες. Παρά τις προσπάθειες που έγιναν για την βελτιστοποίηση των μοντέλων μας, δεν καταφέραμε να λάβουμε καλύτερα αποτελέσματα από τα παραπάνω. Οι λόγοι εξαιτίας των οποίων συμβαίνει αυτό, θα εξετασθούν στην συνέχεια, στην παρούσα ενότητα.

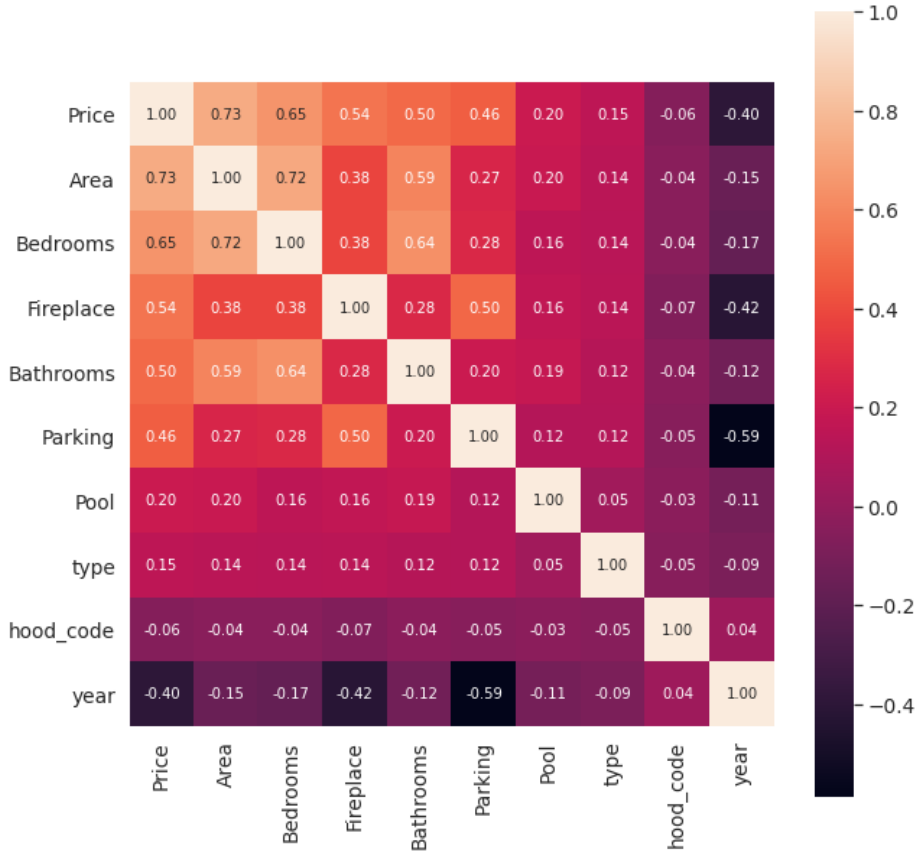
4.3.1 Σύνολα Δεδομένων και Χαρακτηριστικά

Αρχικά, φαίνεται από τα αποτελέσματα των αλγορίθμων ότι λίγο βελτιωμένα αποτελέσματα μας έδωσε το δεύτερο σύνολο δεδομένων. Αξίζει να αναφερθεί εδώ ο τρόπος που οδηγηθήκαμε σε αυτό. Όπως έχει ειπωθεί και σε προηγούμενη ενότητα, τα αποτελέσματα που μας έδωσε το αρχικό σύνολο δεδομένων, χωρίς κάποια τροποποίηση, ήταν αποτρεπτικά. Για το λόγο αυτό προχωρήσαμε σε feature engineering, και μετά από πολυάριθμες δοκιμές καταλήξαμε στα δύο αυτά σύνολα, dataset1 και dataset2, τα οποία διαφαίνονται στην συνέχεια.

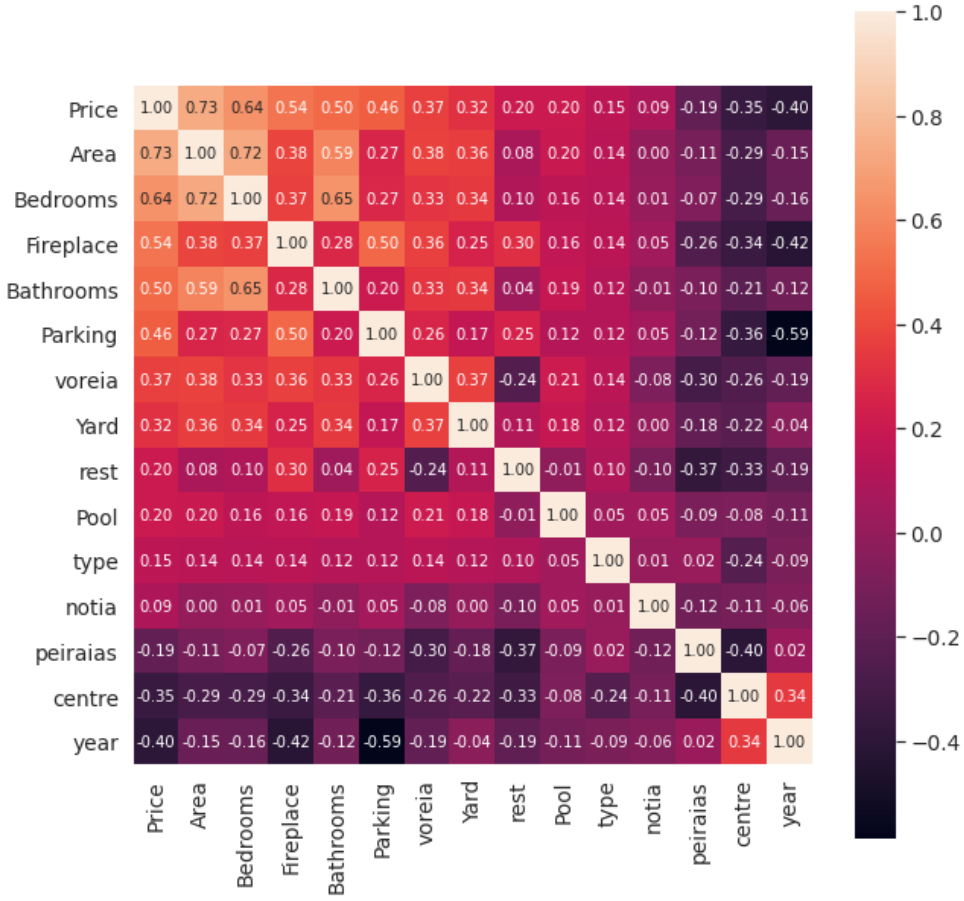
Dataset1	**Dataset2**
Area	Area
Bedrooms	Bedrooms
Bathrooms	Bathrooms
Neighborhood	Year
Year	
Type	Type
Fireplace	Fireplace
Pool	Pool
Parking	Parking
Price	Price
	Voreia_proastia
	Notia_proastia
	Centre
	Peireias
	Anatolika_ditika
	Yard

Πίνακας 9 Σύνολα δεδομένων που εφαρμόσθηκαν ως είσοδοι στα μοντέλα.

Ο λόγος για τον οποίο εφαρμόσθηκε η επιλογή των χαρακτηριστικών, ήταν το ελλιπές μέγεθος του αρχικού μας δείγματος. Όπως γνωρίζουμε και από την κατάρα της διαστατικότητας, μετά από έναν συγκεκριμένο αριθμό χαρακτηριστικών, για δεδομένο αριθμό δειγμάτων, η απόδοση του ταξινομητή μειώνεται. Έτσι λοιπόν, αυξάνοντας στην περίπτωσή μας τον αριθμό των χαρακτηριστικών, χωρίς αύξηση του δείγματος, αυξάνεται η πιθανότητα υπερεκπαίδευσης για το μοντέλο μας, γεγονός που κληθήκαμε να αντιμετωπίσουμε. Για το λόγο αυτό, επιλέξαμε τα χαρακτηριστικά που είχαν μεγαλύτερη πιθανότητα να καταφέρουν να εξηγήσουν όσο δυνατόν μεγαλύτερο ποσοστό της διασποράς του δείγματος, δηλαδή αυτά που διέθεταν μεγαλύτερο ποσοστό συσχέτισης με την τιμή. Ξεκινώντας από αυτά, είδαμε πραγματικά εξάλειψη της υπερεκπαίδευσης και συνεχίζοντας προσεκτικά την πρόσθεση επιπλέον χαρακτηριστικών καταφέραμε να φτάσουμε στο βέλτιστο δυνατό αποτέλεσμα για τα συγκεκριμένα δεδομένα. Παρακάτω φαίνονται οι πίνακες συσχετίσεων των χαρακτηριστικών των δύο συνόλων δεδομένων με την τιμή.



Εικόνα 19 Πίνακας συσχετίσεων πρώτου συνόλου δεδομένων.



Εικόνα 20 Πίνακας συσχετίσεων δεύτερου συνόλου δεδομένων.

Γίνεται αντιληπτό από τους πίνακες 4.1 και 4.2, ότι το εμβαδόν του ακινήτου και ο αριθμός των υπνοδωματίων έχουν τον υψηλότερο βαθμό συσχέτισης με την τιμή του ακινήτου. Σημαντική συσχέτιση έχει επίσης το τζάκι, τα μπάνια και το πάρκινγκ. Αυτό που παρατηρείται και θεωρούμε ενδιαφέρον, είναι ότι ενώ στο πρώτο σύνολο δεδομένων η περιοχή του κάθε ακινήτου δεν έχει καθόλου υψηλή συσχέτιση με την τιμή του, γεγονός το οποίο δεν περιμέναμε, αλλά που είναι λογικό αν λάβουμε υπόψιν μας το μεγάλο πλήθος των περιοχών σε σχέση με τον μικρό αριθμό δεδομένων, στο δεύτερο σύνολο δεδομένων βλέπουμε ότι, μετά την ομαδοποίηση των περιοχών των ακινήτων, αυξάνεται κατά πολύ η συσχέτισή τους με την τιμή. Για παράδειγμα, βλέπουμε ότι η μεταβλητή που αντιστοιχεί στα ακίνητα, τα οποία βρίσκονται στα βόρεια προάστια έχει αρκετά υψηλή συσχέτιση με την τιμή, λογικό επόμενο, καθώς κατά κανόνα η μεγάλη πλειοψηφία των ακινήτων που βρίσκονται στα Βόρεια Προάστια της Αττικής, έχουν αρκετά υψηλότερες, από τον μέσο όρο τιμές. Αντίστοιχα, βλέπουμε ότι η μεταβλητή που αντιστοιχεί στις κατοικίες του κέντρου της Αθήνας, έχει αρνητική συσχέτιση με την τιμή, γεγονός που επιβεβαιώνεται από την πραγματικότητα.

Φαίνεται ακόμα η θετική συσχέτιση που υπάρχει μεταξύ των μεταβλητών Area, Bathrooms, Bedrooms, Fireplace, Parking, Voreia, Yard. Φυσικά το μεγάλο εμβαδόν θα συνοδεύεται και από μεγαλύτερο αριθμό υπνοδωματίων και μπάνιων, μεγαλύτερη πιθανότητα ύπαρξης τζακιού, πάρκινγκ και αυλής, και δεδομένων των προηγουμένων φαίνεται ότι αυξάνεται και η πιθανότητα το ακίνητο να βρίσκεται στα Βόρεια Προάστια. Ενδιαφέρουσα είναι επίσης, η αρνητική συσχέτιση της μεταβλητής Year με την τιμή, αλλά και η θετική με τη μεταβλητή centre, η οποία αφορά ακίνητα του κέντρου της Αθήνας. Αυτό συμβαίνει, καθώς όσο αυξάνεται η μεταβλητή Year, τόσο αυξάνεται και η ηλικία του σπιτιού. Έτσι λοιπόν, φαίνεται ότι τα νεότερα ακίνητα έχουν αυξημένη τιμή και ότι τα ακίνητα του κέντρου της πόλης έχουν αυξημένη ηλικία, γεγονός που επιβεβαιώνεται και από την πραγματικότητα. Τέλος, θα πρέπει να αναφέρουμε, ότι η ανυπαρξία υπολογίσιμων συσχετίσεων μεταξύ των δεδομένων μας, αποτελεί ένα επιθυμητό χαρακτηριστικό των συνόλων δεδομένων, που χρησιμοποιούνται ως είσοδοι σε αλγορίθμους μηχανικής μάθησης. Τόσο στην περίπτωση της παλινδρόμησης, όσο και στην περίπτωση των δέντρων απόφασης, υψηλή συσχέτιση μεταξύ χαρακτηριστικών, είναι πιθανό να οδηγήσει σε μη έγκυρα αποτελέσματα και εσφαλμένα συμπεράσματα, εξαιτίας της πολυσυγγραμμικότητας των χαρακτηριστικών, ειδικά ως προς την αξιολόγηση των χαρακτηριστικών. Στην περίπτωση του πρώτου αλγόριθμου, επηρεάζεται ο υπολογισμός των συντελεστών που δημιουργούνται για κάθε χαρακτηριστικό, με αποτέλεσμα πολύ μικρές αλλαγές στα δεδομένα να δημιουργούν αδικαιολόγητα σοβαρές αλλαγές στους συντελεστές, ενώ στη δεύτερη, αν επιλεγούν και τα δύο συσχετιζόμενα χαρακτηριστικά στο σύνολο για την κατασκευή του δένδρου, είναι πιθανό να μην έχει την δυνατότητα ο αλγόριθμος να ξεχωρίσει τον πραγματικό βαθμό σημαντικότητας του καθενός, ενώ οι προβλέψεις για τα πιθανώς να μην επηρεαστούν σε σημαντικό ποσοστό.

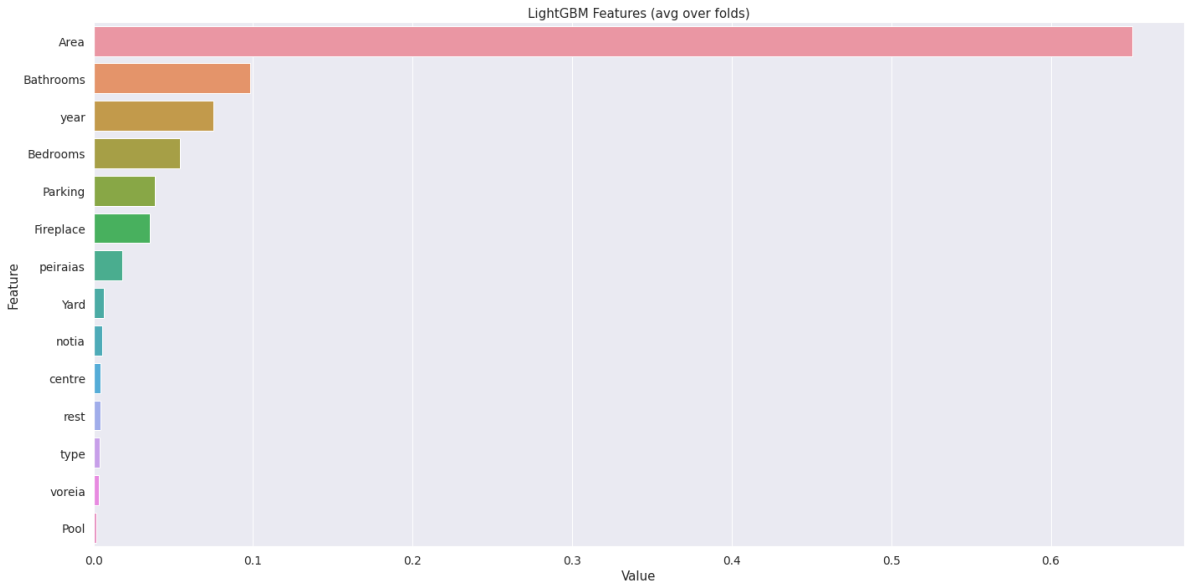
Προχωρώντας, κάτι που παρατηρήθηκε και θεωρούμε σημαντικό, είναι η μεγάλη διακύμανση των τιμών της μεταβλητής της τιμής ακινήτου, για τις υψηλότερες τιμές. Πιο συγκεκριμένα, έχουμε το 25-εκατοστιαίο σημείο των δεδομένων να έχει την τιμή 115.000, το 75-εκατοστιαίο σημείο έχει τιμή 365.000, ενώ το maximum έχει τιμή 15.000.000. Ακόμα, βλέπουμε μέσω υπολογισμών, ότι η μέση τιμή των ακινήτων, τα οποία βρίσκονται στα βόρεια προάστια, είναι 419.900. Ένα παράδειγμα από το οποίο διακρίνεται η έντονη διακύμανση των τιμών αποτελεί το γεγονός ότι για τα ακίνητα με εμβαδόν μεταξύ 500τ.μ. και 600τ.μ., το εύρος τιμών της τιμής πώλησής τους είναι [100.000, 4.800.000], το οποίο προφανώς είναι πολύ μεγάλο. Και ενώ ισχύουν τα ανωτέρω, αντιλαμβανόμαστε ότι δεν υπάρχει κάποια μεταβλητή ή μεταβλητές, που να περιγράφουν λεπτομερέστερα στοιχεία του ακινήτου, όπως είναι ο βαθμός πολυτέλειας του κτιρίου και των χαρακτηριστικών του,

η κατάσταση του κτιρίου, η κατάσταση των σημαντικών ηλεκτρικών συσκευών του, όπως για παράδειγμα ο θερμοσίφωνας, η κουζίνα και ο ηλιακός. Για τον λόγο αυτό, περιορίσαμε εκ νέου, ακόμα παραπάνω, το σύνολο τιμών των συνόλων δεδομένων στα οποία δουλεύαμε. Το νέο εύρος τιμών πώλησης είναι [20.000, 900.000] και τα σύνολα δεδομένων μειώθηκαν μόλις κατά 1.143 ακίνητα. Μετά από αυτήν την ενέργεια το σφάλμα μας μειώθηκε σχεδόν στο μισό, με τα αποτελέσματα να είναι αυτά που διαφαίνονται στην ενότητα 4.2.

Για την αξιολόγηση της σημασίας κάθε χαρακτηριστικού, σε ό,τι αφορά τους αλγορίθμους, χρησιμοποιήθηκαν οι συντελεστές που αντιστοιχούν σε κάθε χαρακτηριστικό και είναι ενδεικτικοί της βαρύτητάς τους στον υπολογισμό της τιμής των ακινήτων. Η διερεύνηση της σπουδαιότητας των χαρακτηριστικών δεν είναι μία καλά ορισμένη διαδικασία. Στην παρούσα περίπτωση, θα χρησιμοποιηθεί η ενσωματωμένη συνάρτηση της βιβλιοθήκης sklearn, feature_importance(). Η συνάρτηση αυτή, κατατάσσει τα χαρακτηριστικά, με βάση το κέρδος τους στον αλγόριθμο, το οποίο υπολογίζεται από την συνάρτηση:

$$Gain = \frac{1}{2}\left[\frac{{G_L}^2}{H_L+\lambda} + \frac{{G_R}^2}{H_R+\lambda} + \frac{(G_R+G_L)^2}{H_R+H_L+\lambda}\right] - \gamma$$

Τα αποτελέσματα στην περίπτωσή μας, για τον βέλτιστο αλγόριθμο Gradient Boosting, παραδίδονται στην Εικόνα 4.3.



Εικόνα 21 Σπουδαιότητα χαρακτηριστικών.

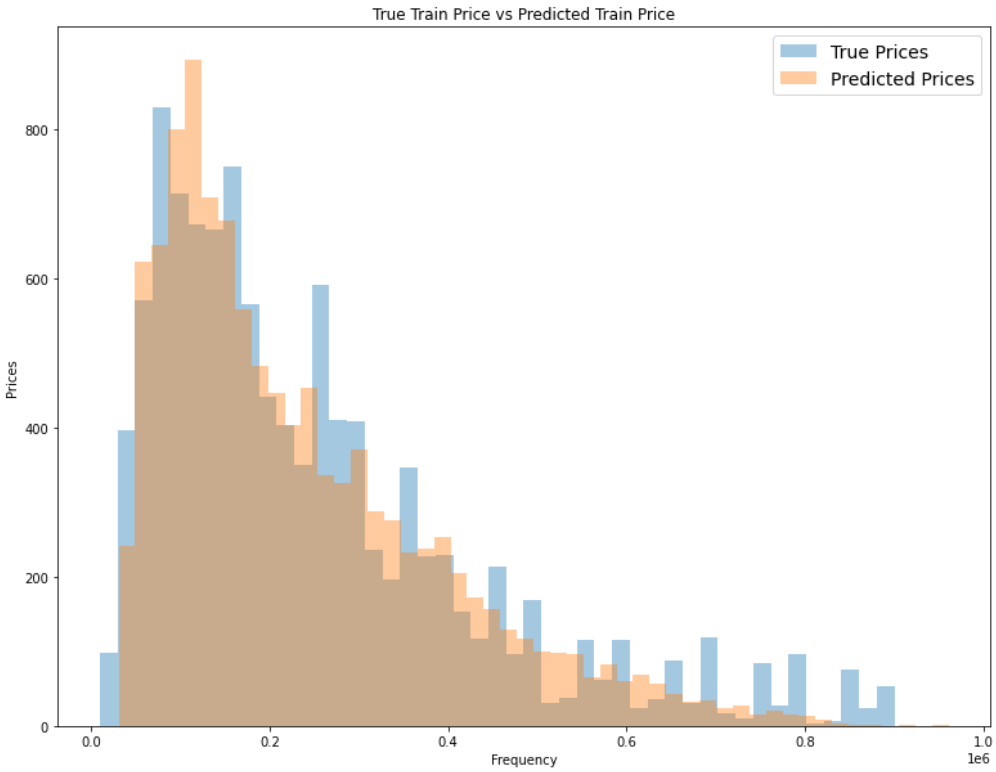
Όπως φαίνεται, τα χαρακτηριστικά που αποτελούν ισχυρότερους παράγοντες για το αποτέλεσμα του αλγορίθμου είναι, όπως περιμέναμε, το εμβαδόν, τα υπνοδωμάτια και τα μπάνια, καθώς επίσης και το έτος κτίσης. Βλέπουμε εδώ, πως παρά τον σχετικά υψηλό βαθμό συσχέτισης της περιοχής των βορείων προαστίων με την τιμή, μικρή επιρροή είχε στο αποτέλεσμα του αλγορίθμου, ενώ σημαντικό ρόλο έλαβε η μεταβλητή που αντιστοιχεί στα ακίνητα της περιοχής του Πειραιά. Πράγματι, το πεδίο voreia, όπως έχει ήδη αναφερθεί, είχε ένα πολύ διευρυμένο σύνολο τιμών, το οποίο είναι δύσκολο να διαχειριστεί ένα μοντέλο σαν το δικό μας. Τα μη δυαδικά χαρακτηριστικά, φαίνονται ιδιαίτερα χρήσιμα για τους αλγορίθμους μας, εφόσον βλέπουμε ότι η σπουδαιότητά τους είναι συνολικά μεγαλύτερη από ότι τα δυαδικά χαρακτηριστικά. Αυτό είναι λογικό, καθώς σκοπός του

αλγορίθμου είναι η πρόβλεψη μίας τιμής μέσα σε ένα φάσμα τιμών, το οποίο έχει και αρκετά μεγάλο εύρος στην προκειμένη περίπτωση.

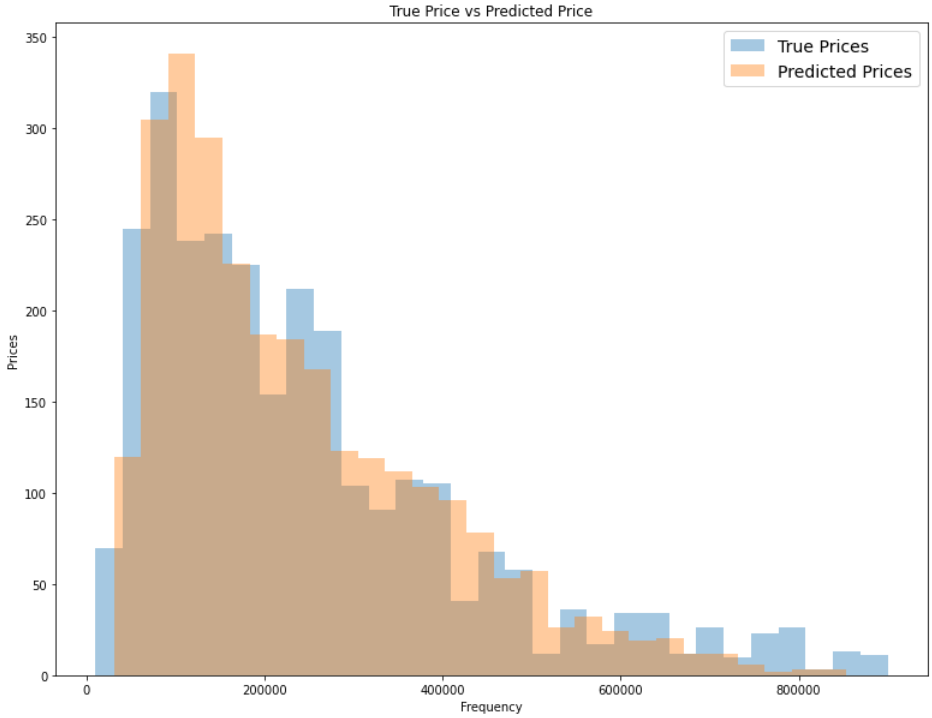
Φαίνεται, τέλος, ότι το σύνολο δεδομένων που διαθέταμε παρουσιάζει σημαντικές ελλείψεις και δεν είναι ικανό να μας δώσει έγκυρα αποτελέσματα, στον βαθμό που επιθυμούμε, για το εύρος τιμών των ακινήτων στις διάφορες περιοχές της Αττικής. Παρόλο που προσπαθήσαμε, μέσω ποικίλων τροποποιήσεων, δοκιμών και δημιουργίας νέων χαρακτηριστικών, όπως η τιμή ανά τετραγωνικό μέτρο κάθε περιοχής, μικρές βελτιώσεις είδαμε στα αποτελέσματα. Παρακάτω, αξιολογούνται τα αποτελέσματα των αλγορίθμων που εφαρμόσθηκαν, καθώς επίσης και επιχειρούμε να συγκεντρώσουμε τους λόγους για τους οποίους πιστεύουμε πως προέκυψαν αυτές οι επιδόσεις.

4.3.2 Αξιολόγηση Αποτελεσμάτων

Παρακάτω παρουσιάζονται τα διαγράμματα στα οποία φαίνονται τα ιστογράμματα της πραγματικής τιμής των ακινήτων, σε αντιπαράθεση με αυτά της προβλεφθείσας τιμής, για τα σύνολα εκπαίδευσης και ελέγχου, αντίστοιχα.



Εικόνα 22 Ιστόγραμμα πραγματικής τιμής – προβλεφθείσας τιμής για το σύνολο εκπαίδευσης.



Εικόνα 23 Ιστόγραμμα πραγματικής τιμής – προβλεφθείσας τιμής για το σύνολο ελέγχου.

Τα παραπάνω, αποτελούν αποτελέσματα που μας έδωσε αλγόριθμος Gradient Boosting, λαμβάνοντας ως είσοδο το δεύτερο σύνολο δεδομένων, dataset2. Όλοι οι αλγόριθμοι που δοκιμάστηκαν είχαν παραπλήσια αποτελέσματα, με τους τρεις αλγορίθμους που παρουσιάζονται στην ενότητα 4.2 να μας δίνουν τα βέλτιστα. Από αυτούς, οι διαφορές είναι πολύ μικρές, με τα Τυχαία Δάση και τον Gradient Boosting, να έχουν μικρότερη τιμή σφάλματος από ότι ο Light Gradient Boosting Regressor. Η απόδοση αυτή των αλγορίθμων δεν μας εξέπληξε, εφόσον αξιολογώντας το σύνολο δεδομένων, βλέπουμε ότι πρόκειται για ένα μικρού μεγέθους σύνολο, γεγονός το οποίο μας απέτρεψε από την χρήση αλγορίθμων βαθιάς μηχανικής μάθησης. Ακόμα, η φύση του συνόλου δεδομένων και οι σχέσεις μεταξύ τόσο των αριθμητικών, όσο και των κατηγορηματικών μεταβλητών, φαίνεται να περιγράφονται καλύτερα από πολλαπλές γραμμικές σχέσεις και αλγορίθμους παλινδρόμησης ή δενδρικών δομών.

Αυτό που προκύπτει από την μελέτη των αποτελεσμάτων, αφορά την αδυναμία πρόβλεψης των υψηλών τιμών. Τόσο στο σύνολο εκπαίδευσης, αλλά και στο σύνολο ελέγχου, φαίνεται να υπάρχει συστηματική πρόβλεψη μικρότερης τιμής για τα ακριβά ακίνητα, περισσότερο από όσο συμβαίνει το αντίθετο. Αυτό, συνδέεται με το ύψος της μετρικής MAE, της οποίας η τιμή κυμαίνεται γύρω στις 55.000-60.000, με την τιμή του R^2 να ισούται με 72-78%. Παρόλο, δηλαδή, που το μοντέλο μας είναι σε θέση να εξηγήσει το 75% περίπου, των δεδομένων, το μέσο απόλυτο σφάλμα παραμένει υψηλό. Προσπαθώντας να εξηγήσουμε το παραπάνω, καταλήγουμε στο γεγονός ότι οι αστοχίες του μοντέλου κυρίως για ακίνητα με πολύ υψηλή τιμή, αλλά και για ακίνητα με χαμηλότερη τιμή, έχουν ως αποτέλεσμα μια αρκετά μεγάλη τιμή του μέσου σφάλματος. Σε προσπάθεια κατανόησης της συμπεριφοράς του μοντέλου, αφαιρέσαμε παραπάνω υψηλές τιμές, προσπαθώντας να δημιουργήσουμε ένα σύνολο δεδομένων με μικρότερη διακύμανση για την μεταβλητή της τιμής. Έτσι, καταλήξαμε σε ένα σύνολο με 11234 εισόδους δεδομένων και με την τιμή πώλησης να κυμαίνεται στο διάστημα [20.500, 400.000]. Εφαρμόζοντας τους αλγορίθμους στο συγκεκριμένο σύνολο, είχαμε παρόμοια τιμή της μετρικής R^2, με ελάχιστη βελτίωση, ενώ το Huber Loss και το μέσο απόλυτο σφάλμα έπαιρναν τιμές στο διάστημα

[35.000,39.000], τιμές σχεδόν οι μισές από ότι στα προηγούμενα πειράματα. Βλέπουμε, λοιπόν, και πάλι, την αδυναμία του συνόλου δεδομένων να εκφράσει τις ιδιομορφίες στην κατανομή των τιμών των ακινήτων.

Συγκεντρωτικά, οι δύο κύριοι λόγοι για τους οποίος πιστεύουμε ότι είχαμε τα συγκεκριμένα αποτελέσματα και την μειωμένη αυτή απόδοση του μοντέλου, είναι οι εξής:

- Πρώτον, κύριο λόγο εξαιτίας του οποίου δεν έχουμε την δυνατότητα να βελτιώσουμε στον επιθυμητό βαθμό τα αποτελέσματά μας, είναι ο αριθμός και η ποιότητα των δεδομένων μας. Για το πρώτο είναι προφανές ότι με λιγότερα δεδομένα καταλήγουμε σε λιγότερο έγκυρα αποτελέσματα. Για το δεύτερο, όπως έχει αναφερθεί σε προηγούμενο κεφάλαιο, τα δεδομένα μας, τα οποία όντας από μία ιστοσελίδα στην οποία δεν συμβαίνουν συστηματικοί έλεγχοι για τα υπάρχοντα ακίνητα, αλλά και δεν ζητείται συγκεκριμένος αριθμός χαρακτηριστικών για να δημοσιευθεί μία αγγελία, είχαν σημαντικές διαφορές στον αριθμό των γνωρισμάτων που περιείχαν. Κάποια ακίνητα είχαν λεπτομερή περιγραφή, ενώ άλλα μπορεί να μην διέθεταν. Το γεγονός αυτό, δημιούργησε πολλές κενές τιμές στα χαρακτηριστικά του συνόλου δεδομένων, υποβιβάζοντας την ποιότητά του.
- Δεύτερον, η μεγάλη διακύμανση των τιμών της τιμής πώλησης των ακινήτων, σε συνδυασμό με την έλλειψη επάρκειας χαρακτηριστικών για να την εξηγήσουν, οδηγεί στα μεγαλύτερα αυτά σφάλματα και αποτελεί εξίσου σημαντική αιτία αυτών.

4.4 Σύγκριση Αποτελεσμάτων

Στην παρούσα υποενότητα, θα αναφερθούμε στην σύγκριση που διεξήγαμε μεταξύ των αποτελεσμάτων των αλγορίθμων μας και των τιμών που προκύπτουν από τις αντικειμενικές αξίες των ακινήτων, όπως αυτές ορίζονται από το Υπουργείο των Οικονομικών. Ακόμα θα παρουσιάσουμε αποτελέσματα ορισμένων παρόμοιων ερευνών, με σκοπό τη σύγκριση της απόδοσής τους με αυτήν των μοντέλων μας.

Πιο συγκεκριμένα, λήφθηκαν υπόψιν οι δημοσιευμένοι πίνακες του Υπουργείου των Οικονομικών, οι οποίοι ορίζουν την τιμή ανά τετραγωνικό μέτρο κάθε περιοχής της Ελλάδος, ως την τιμή ζώνης. Έτσι, με βάση την τιμή ζώνης κάθε περιοχής της Αττικής, υπολογίστηκαν οι αντικειμενικές τιμές ανά τετραγωνικό μέτρο των συστάδων της μεταβλητής Neighborhood του συνόλου δεδομένων μας. Είχαμε δηλαδή τη βασική τιμή ανά τετραγωνικό μέτρο, συνολικά για τα Βόρεια Προάστια, τα Νότια Προάστια, τον Πειραιά και τα Δυτικά και Ανατολικά Προάστια. Στη συνέχεια, τις τιμές αυτές τις υπολογίσαμε και συνολικά με βάση το σύνολο δεδομένων μας. Συνεπώς, η πρώτη σύγκριση που κάναμε, ήταν μεταξύ των βασικών τιμών ανά τετραγωνικό μέτρο κάθε περιοχής, που προβλέπονται από το Υπουργείο Οικονομικών και αυτών που υπολογίσαμε από το σύνολο δεδομένων μας για κάθε ομάδα περιοχής. Τα αποτελέσματα που λάβαμε φαίνονται στον παρακάτω πίνακα.

	Αντικειμενική αξία ανά τ.μ.	**Αξία ανά τ.μ. με βάση το σύνολο δεδομένων.**
Βόρεια Προάστια	1587,66	2044,37
Νότια Προάστια	1869,99	2805,21
Πειραιάς	1195,2	1768,27
Ανατολικά και Δυτικά Προάστια	1430,79	2235,37

Πίνακας 10 Σύγκριση αξιών ανά τετραγωνικό μέτρο.

Η πρώτη παρατήρηση κοιτώντας τα παραπάνω αποτελέσματα, είναι το γεγονός ότι οι αντικειμενικές αξίες είναι πραγματικά χαμηλότερες από ότι οι πραγματικές τιμές των τιμών ανά τετραγωνικό μέτρο κάθε περιοχής. Το χάσμα ανάμεσα στις πραγματικές και τις αντικειμενικές τιμές των αξιών των ακινήτων στην χώρα μας είναι γνωστό αποδεικνύεται πρακτικά από τα παραπάνω στοιχεία. Η υψηλότερη εμπορική αξία των ακινήτων είναι κάτι αναμενόμενο, το οποίο επιβεβαιώνεται. Συνεχίζοντας, με βάση τις αξίες αυτές ανά τετραγωνικό μέτρο, υπολογίζουμε την τιμή κάθε ακινήτου του συνόλου δεδομένων μας. Καταλήγουμε, δηλαδή σε δύο σύνολα δεδομένων, τα οποία περιέχουν τις τιμές των ακινήτων μας, το ένα με βάση τις βασικές τιμές ζώνης κάθε περιοχής και το άλλο με βάση τον μέσο όρο της τιμής ανά τετραγωνικό μέτρο κάθε περιοχής. Με βάση τα δύο αυτά σύνολα δεδομένων, υπολογίζονται, για το κάθε ένα ξεχωριστά, οι μετρικές αξιολόγησης που χρησιμοποιήσαμε για το αρχικό μας μοντέλο προκειμένου να συγκριθούν με τα αποτελέσματα των μοντέλων μας. Παρακάτω φαίνονται οι πίνακες των αποτελεσμάτων αυτών.

Metric	**Train Set**	**Test Set**
R^2	0.5134	0.4994
MAE	299236.08	300145.08
Huber Loss	132934.87	136282.904
AUC	24,53%	23,95%

Πίνακας 11 Αποτελέσματα με βάση την αντικειμενική τιμή των ακινήτων.

Metric	**Train Set**	**Test Set**
R^2	0.4753	0.4406
MAE	310656.72	330807.38
Huber Loss	92345.89	95946.52
AUC	20,43%	19,16%

Πίνακας 12 Αποτελέσματα με βάση τον μέσο όρο ανά τ.μ. κάθε περιοχής.

Οι τιμές που λάβαμε από τα παραπάνω ήταν αρκετά χειρότερες από τις τιμές που λάβαμε με το αρχικό μας μοντέλο. Συγκεκριμένα, για το σύνολο δεδομένων που προέκυψε με βάση την βασική τιμή ζώνης, το μέσο απόλυτο σφάλμα στο σύνολο ελέγχου ισούται με 300.145,08. Για το σύνολο δεδομένων που προέκυψε με βάση τη μέση τιμή ανά τετραγωνικό μέτρο κάθε περιοχής, το μέσο απόλυτο σφάλμα ισούται με 330.607,38 στο σύνολο ελέγχου. Παρατηρούμε ότι τα σφάλματα είναι πολύ μεγαλύτερα με τη χρήση των νέων συνόλων, από ότι τα σφάλματα των προβλεφθείσων, από τα μοντέλα, τιμών. Το γεγονός αυτό, είναι κάτι που αναμέναμε, εφόσον όπως είπαμε πριν, οι αντικειμενικές τιμές των ακινήτων είναι πολύ χαμηλότερες από την εμπορική τιμή πώλησης του, και αυτό επιβεβαιώνεται από τα σύνολα δεδομένων. Τόσο οι τιμές που προβλέψαμε, όσο και οι πραγματικές τιμές των ακινήτων ήταν αρκετά υψηλότερες από τις τιμές με βάση την αντικειμενική τους αξία. Με αυτή τη σύγκριση, επιθυμούμε να επαληθεύσουμε ότι το μοντέλο μας, κρίνεται περισσότερο χρήσιμο από τις τιμές πώλησης που προκύπτουν σύμφωνα με τις βασικές τιμές ζώνης κάθε περιοχής.

Προχωρώντας, θα παρουσιάσουμε ακολούθως, τα αποτελέσματα κάποιων μελετών, παρόμοιων με την δική μας. Το σημαντικά μεγαλύτερο ποσοστό τέτοιων δημοσιευμένων ερευνών, παρουσιάζει και ερμηνεύει τα αποτελέσματα που προκύπτουν με βάση το λογάριθμο των δεομένων. Οι μετρικές αξιολόγησης, δηλαδή, αφορούν τον λογάριθμο των

τιμών που προβλέπουν τα μοντέλα και όχι τις πραγματικές τιμές. Στη συνέχεια θα αναφερθούμε σε μελέτες οι οποίες εμφανίζουν τα αποτελέσματα των σφαλμάτων χωρίς κάποια μετατροπή.

Το 2015 οι Yingyu Feng και Kelvyn Jones [2], μελέτησαν τις αποδόσεις του Ηδονικού Μοντέλου Τιμολόγησης και των τεχνητών νευρωνικών δικτύων στην πρόβλεψη των τιμών των ακινήτων του Μπρίστολ. Χρησιμοποίησαν δεδομένα από 65.302 ακίνητα προς πώληση στην ευρύτερη περιοχή του Μπρίστολ, από τα οποία τα ακίνητα δημοσιευμένα έως το 2012 αποτέλεσαν το σύνολο εκπαίδευσης, ενώ ακίνητα του 2013 αποτέλεσαν το σύνολο ελέγχου. Εξέτασαν τρία διαφορετικά σενάρια με τρία διαφορετικά σύνολα δεδομένων. Το πρώτο σύνολο περιείχε δεδομένα που αφορούσαν μόνο τα χαρακτηριστικά των ακινήτων, το δεύτερο περιέχει τα χαρακτηριστικά των κατοικιών αλλά και μεταβλητές που αφορούν την τοποθεσία τους και το τρίτο περιέχει τα χαρακτηριστικά των κατοικιών και τα χαρακτηριστικά κάθε γειτονιάς. Για την μέτρηση της απόδοσης χρησιμοποίησαν παράμετρο R^2, το μέσο απόλυτο σφάλμα και το μέσο απόλυτο ποσοστό σφάλματος. Τα αποτελέσματα των σφαλμάτων τους ήταν χαμηλότερα από τα δικά μας, με το χαμηλότερο μέσο απόλυτο σφάλμα να κυμαίνεται γύρω στις 48.000. Η εισαγωγή στο dataset των γεωγραφικών χαρακτηριστικών και των χαρακτηριστικών της γειτονιάς, έχει ως αποτέλεσμα την μείωση της διασποράς που δεν μπορεί να εξηγηθεί από το μοντέλο στο 34%. Παρόλο που τα αποτελέσματα δεν είναι εκπληκτικά καλύτερα από τα δικά μας, οι Yingyu Feng και Kelvyn Jones, χρησιμοποίησαν περισσότερες μεταβλητές στο σύνολο δεδομένων τους που αφορούσαν γεωχωρικά δεδομένα και είχαν επίσης στην διάθεσή τους τις ημερομηνίες δημοσίευσης της αγγελίας και πώλησης του σπιτιού. Τα τελευταία φαίνεται να συμβάλανε στην βελτίωση των αποτελεσμάτων τους.

Μία ακόμα ενδιαφέρουσα μελέτη είναι εκείνη των Tiancheng Cai, Kevin Han και Han Wu [4], οι οποίοι το 2019, όπως έχει αναφερθεί και σε προηγούμενο κεφάλαιο προσπάθησαν να προβλέψουν τις τιμές των καταλυμάτων του Airbnb στην Μελβούρνη. Τα τελικά σύνολα δεδομένων τους αποτελούνταν από 22.725 εισόδους και 80 γνωρίσματα, εκτός της τιμής για τα αριθμητικά δεδομένα και για τα κατηγορηματικά δεδομένα είχαν 485.416 εισόδους, που αντιστοιχούσαν στην περιγραφή των καταλυμάτων. Η τιμές των καταλυμάτων περιορίστηκαν στο εύρος [x, 1000], όπου x η τιμή του φθηνότερου καταλύματος. Οι μέθοδοι που χρησιμοποίησαν ήταν η γραμμική παλινδρόμηση, η παλινδρόμηση ridge, η παλινδρόμηση gradient boosting, τα τυχαία δάση, η μηχανή διανυσμάτων υποστήριξης και τέσσερα διαφορετικά νευρωνικά δίκτυα. Τα αποτελέσματά τους αξιολογήθηκαν με βάση τις μετρικές R^2 και το μέσο τετραγωνικό σφάλμα. Καλύτερη απόδοση από τα παραπάνω είχε το μοντέλο της παλινδρόμησης με την τεχνική gradient boosting και τα τυχαία δάση, με μέσο τετραγωνικό σφάλμα 4024.7052 και 4422.7124, αντίστοιχα. Οι τιμές του R^2 είναι 0,69 και 0,66 αντίστοιχα για τα δεδομένα ελέγχου. Το μοντέλο τους, αναφέρουν ότι τείνει να υποτιμά τις τιμές των ακριβότερων καταλυμάτων. Περιορίζοντας την μέγιστη τιμή καταλύματος στα 500 δολάρια, το μέσο τετραγωνικό σφάλμα μειωνόταν σχεδόν στα μισά. Η τιμές αυτές είναι εμφανώς χαμηλότερες από τις τιμές των αποτελεσμάτων της μελέτης μας, παρόλα αυτά πρέπει να σημειωθεί ότι χαμηλότερες είναι και οι τιμές των καταλυμάτων που προσπαθούν να ερμηνεύσουν τα μοντέλα. Βέβαια, υπολογίζοντας το μέσο ποσοστιαίο σφάλμα τους βλέπουμε ότι στην μελέτη τους αυτό ισούται με περίπου 15%, ενώ σε εμάς ισούται με περίπου 24%. Για μελλοντικές μελέτες προτείνεται η προσεκτική επιλογή χαρακτηριστικών και η εκπαίδευση των μοντέλων για διαφορετικά εύρη τιμών.

Κεφάλαιο 5

Επίλογος

5.1 Σύνοψη

Στην παρούσα διπλωματική εργασία μελετήθηκε το πρόβλημα της μοντελοποίησης των τιμών των ακινήτων. Ένα πρόβλημα εξέχουσας σημασίας για ένα σεβαστά μεγάλο τμήμα του Ελληνικού και βέβαια όχι μόνο, πληθυσμού, εφόσον η αγορά των ακινήτων επηρεάζει άμεσα την οικονομία μας. Στην συγκεκριμένη μελέτη, επικεντρωθήκαμε στην αγορά των ακινήτων του νομού της Αττικής. Πραγματοποιήθηκε μία προσπάθεια υλοποίησης ενός μοντέλου πρόβλεψης των τιμών των ακινήτων, ακολουθώντας τις μεθοδολογίες και τη λογική της μηχανικής μάθησης, που προβλέπονται μέσω της βιβλιογραφίας.

Αρχικά, συλλέξαμε δεδομένα από την ιστοσελίδα της Χρυσής Ευκαιρίας, τα οποία αφορούσαν ακίνητα προς πώληση, και προχωρήσαμε στην λεπτομερή επεξεργασία τους. Η αναλυτική διερεύνηση των δεδομένων, μας επέτρεψε να κατανοήσουμε σε βάθος την φύση των δεδομένων και τα χαρακτηριστικά του, γεγονός που μας οδήγησε στην βέλτιστη χρήση του και στην επιλογή του ορθότερου τρόπου επεξεργασίας του. Η επιλογή των καταλληλότερων μεθόδων μηχανικής μάθησης, οι οποίες χρησιμοποιήθηκαν για την μοντελοποίηση της μεταβλητής στόχου, αποτέλεσε απόρροια της βαθύτερης ανάλυσης των δεδομένων μας. Για να φτάσουμε στην βέλτιστη μορφή του συνόλου δεδομένων, κληθήκαμε να εντοπίσουμε έναν αποτελεσματικό τρόπο διαχείρισης της μέτριας, έως και κακής, ποιότητάς του, την υψηλή τυπική του απόκλισή, όσον αφορά την μεταβλητή στόχο και τις πολυάριθμες κενές τιμές του, λαμβάνοντας, επίσης, υπόψιν μας τις σχέσεις μεταξύ των γνωρισμάτων. Το μικρό μέγεθος του dataset, μας οδήγησε στην μείωση και απλοποίηση των χαρακτηριστικών εισόδου, προκειμένου να αντιμετωπιστεί υπερεκπαίδευση των μοντέλων μας.

Δοκιμάστηκαν, έτσι, ποικίλοι αλγόριθμοι, για την εκπαίδευση των μοντέλων με είσοδο το σύνολο δεδομένων, από τους οποίους βέλτιστα αποτελέσματα μας έδωσαν οι τεχνικές παλινδρόμησης LGBM, η παλινδρόμηση με χρήση της τεχνικής Gradient Boosting και τα Τυχαία Δάση. Με βάση αυτά, διακρίνουμε και αναλύουμε τη σπουδαιότητα των χαρακτηριστικών και κατανοούμε τον ρόλο που κατείχε το καθένα από αυτά στην όψη των αποτελεσμάτων. Όλοι οι αλγόριθμοι έκριναν ότι το εμβαδόν των ακινήτων, το έτος κτίσης και ο αριθμός των υπνοδωματίων αποτελούν τα κρισιμότερα γνωρίσματα, επηρεάζοντας σε σημαντικό βαθμό την τιμή πώλησης.

Τα αποτελέσματα των πειραμάτων μας, υποδεικνύουν την πολυπλοκότητα του προβλήματος και του πλήθος των παραγόντων από τους οποίους αυτό εξαρτάται. Εξετάζοντας τις τιμές των μετρικών που λάβαμε από τα μοντέλα μας, αναρωτιόμαστε ποιοι είναι οι λόγοι εξαιτίας των οποίων αυτές δεν ήταν δυνατόν να βελτιωθούν. Ως φυσικό επακόλουθο λοιπόν, έρχεται η ερμηνεία και αξιολόγηση των αποτελεσμάτων. Καταλήγουμε, έτσι, σε ένα σύνολο αιτιών, οι οποίες κρίνουμε ότι κατέχουν ρόλο εξέχουσας σημασίας στην διαμόρφωση της επίδοσης των μοντέλων μας. Εδώ πρέπει να σημειωθεί, ότι

σύμφωνα με τα πειράματα που διεξήχθησαν, αλλά και με τη μελέτη πολυάριθμων ερευνών ίδιου θέματος με αυτό που καταπιανόμαστε, μπορούμε να καταλήξουμε με ασφάλεια στην ορθότητα της επιλογής των συγκεκριμένων μεθόδων και μοντέλων πρόβλεψης και να αποκλείσουμε την πιθανότητα να ευθύνονται για τα επίπεδα των σφαλμάτων. Αντιθέτως, το περιορισμένο μέγεθος του dataset και η απουσία χαρακτηριστικών που εκφράζουν σημαντικές λεπτομέρειες, οι οποίες διαχωρίζουν τα υψηλής με τα χαμηλής τιμής ακίνητα, σε συνδυασμό με την μεγάλη διακύμανση των τιμών της τιμής πώλησής τους, ακόμα και για ακίνητα στην ίδια περιοχή και με παραπλήσιο εμβαδόν, αποτελούν από τους κυριότερους παράγοντες, οι οποίοι πιστεύουμε πως συνέβαλαν στην ποιότητα των αποτελεσμάτων.

Στη συνέχεια, προχωράμε σε κάποιες συγκρίσεις που διεξήγαμε, μεταξύ παρόμοιων δημοσιευμένων ερευνών, οι οποίες ακολούθησαν παραπλήσια σειρά βημάτων με αυτήν που ακολουθήθηκε και στην παρούσα μελέτη. Ακόμα, ενδιαφέρουσα κρίθηκε η σύγκριση των τιμών, οι οποίες προκύπτουν με βάση την αντικειμενική αξία των ακινήτων, αρχικά με τις πραγματικές τιμές πώλησης του dataset, αξιολογώντας τις τιμές των μετρικών της μεταξύ τους σύγκρισης και έπειτα με τις τιμές πώλησης που προέβλεψαν τα μοντέλα μας.

Συμπερασματικά, η πρόβλεψη των τιμών πώλησης των ακινήτων κρίνεται ιδιαίτερα απαιτητική και πολύπλοκη διαδικασία, καθώς η τιμή ενός ακινήτου επηρεάζεται από πλήθος παραγόντων, ποικίλων φύσεων, εφόσον συνδέεται άρρηκτα με την κατάσταση και την οικονομία κάθε κοινωνίας, τις αλλαγές που υφίστανται και τους παράγοντες που τις επηρεάζουν. Προτείνεται, όμως, εδώ, μία βασική μεθοδολογία αντιμετώπισης του συγκεκριμένου ζητήματος, κατά την οποία γίνεται προσπάθεια αξιοποίησης των υπαρχόντων πόρων, με σκοπό να ληφθούν όσο το δυνατόν ακριβέστερα αποτελέσματα.

5.2 Μελλοντικές Προτάσεις

Η παρούσα διπλωματική εργασία ανέδειξε τα προβλήματα που προκύπτουν στην προσπάθεια πρόβλεψης της τιμής πώλησης των ακινήτων με τεχνικές μηχανικής μάθησης. Στα πλαίσια της μελέτης αυτής διερευνήθηκαν διάφορες προσεγγίσεις και μεθοδολογίες του ζητήματος με το οποίο καταπιάνεται και παράλληλα διαπιστώθηκαν ορισμένα σημεία στα οποία θεωρούμε ότι διαφορετική αντιμετώπιση θα ωφελούσε. Ακόμα, κρίνοντας πως στο μέλλον οι μελέτες και εφαρμογές γύρω από το συγκεκριμένο ζήτημα θα πληθύνουν, παρουσιάζουμε κάποιες προτάσεις που θα μπορούσαν να συμβάλλουν το μελλοντικό έργο.

Αρχικά, για μελλοντικές και πιο ολοκληρωμένες μελέτες, περισσότερα και πιο έγκυρα δεδομένα απαιτούνται για την επιτυχία τους. Η μελέτη ενός μεγαλύτερου συνόλου δεδομένων θα έδινε μια πιο γενικευμένη και πραγματική εικόνα, η οποία θα μας έδινε εγκυρότερα αποτελέσματα. Ακόμα, απαιτούνται περισσότερες μεταβλητές, οι οποίες θα περιγράφουν πιο λεπτομερείς και σημαντικούς παράγοντες και χαρακτηριστικά, που επηρεάζουν άμεσα την τιμή πώλησης ενός ακινήτου. Χαρακτηριστικά τέτοιου είδους αποτελούν το επίπεδο πολυτέλειας του ακινήτου και των υλικών από τα οποία είναι δομημένο, η κατάσταση των βασικών ηλεκτρικών συσκευών και εγκαταστάσεων του, η λεπτομερέστερη περιγραφή του πιθανού κήπου του. Επιπλέον, εξίσου σημαντικό κρίνεται να ληφθούν υπόψιν περιβαλλοντικοί παράγοντες που επηρεάζουν ένα ακίνητο, όπως τα επίπεδα μόλυνσης της ατμόσφαιρας κάθε περιοχής. Τα δεδομένα μας συγκεντρώθηκαν κατά μία περίοδο διάρκειας αρκετών μηνών, γεγονός που σημαίνει ότι δεν περιέχουν σημαντικά στοιχεία εποχικότητας ή οικονομικά στοιχεία σεβαστής σημασίας. Σε μελλοντικές, λοιπόν, μελέτες προτείνεται να δοθεί προσοχή σε μακροοικονομικούς παράγοντες, αλλά και σε στοιχεία των πωλήσεων ακινήτων που έχουν ήδη ολοκληρωθεί στο παρελθόν. Γενικότερα, η καλύτερη ποιότητα των δεδομένων που θα μπορούσε να

εξασφαλιστεί εάν αυτά παρέχονται με περισσότερη ευκολία από τους αρμόδιους φορείς θα βελτίωνε σημαντικά τις μοντελοποιήσεις.

Ένα ακόμη ενδιαφέρον σημείο των μελλοντικών κατευθύνσεων, θα μπορούσε να αποτελέσει η αξιοποίηση των δεδομένων εικόνας των φωτογραφιών των ακινήτων, αλλά και δορυφορικών δεδομένων, καθώς επίσης και η αξιοποίηση των γεωχωρικών δεδομένων τους. Οι τεχνικές αυτές φαίνεται να συμβάλουν στην βελτίωση της ακρίβειας των αποτελεσμάτων και η ποιότητα τους στο μέλλον μόνο θα αναβαθμίζεται.

Επιπροσθέτως, θα μπορούσαμε να προτείνουμε την κατασκευή και εκπαίδευση διαφορετικών μοντέλων για διαφορετικές περιοχές ή σύνολα περιοχών με παρόμοια χαρακτηριστικά. Εάν τα δεδομένα είναι επαρκή, το παραπάνω μόνο να βελτίωνε θα μπορούσε την εγκυρότητα των προβλέψεων. Παράλληλα, ενθαρρύνεται η προσπάθεια υλοποίησης μοντέλων στηριζόμενα σε διαφορετικές τεχνικές, όπως είναι η βαθιά μάθηση και τα νευρωνικά δίκτυα, με τη βοήθεια των οποίων θα μοντελοποιούταν ευκολότερα και ακριβέστερα η σχέση της τιμής του ακινήτου με το πλήθος των μεταβλητών από τις οποίες αυτή εξαρτάται.

Επιπλέον, ενδιαφέρουσα κρίνεται η δημιουργία ενός δυναμικού συστήματος πρόβλεψης των τιμών των ακινήτων, προτείνεται ως πιθανός στόχος μελλοντικών ερευνών. Με τον όρο δυναμικό, εννοούμε ένα σύστημα το οποίο θα μεταβάλλεται ζωντανά, με βάση τους παράγοντες του παρόντος από τους οποίους επηρεάζεται. Έτσι, θα μπορούσε να καταγράφεται και να μελετάται σε βάθος χρόνου η συμπεριφορά και η πορεία της αγοράς των ακινήτων, αλλά και να προτείνονται σε αληθινό χρόνο οι τιμές των ακινήτων στους ενδιαφερόμενους.

Συνιστάται, ακόμα, η έρευνα της εφαρμογής παρόμοιων μεθόδων πρόβλεψης σε προβλήματα διαφόρων τομέων, όπως είναι η μελλοντική κατεύθυνση της αγοράς των ακινήτων, η πρόβλεψη των τιμών πετρελαίου και η πρόβλεψη των τιμών των μετοχών στο χρηματιστήριο, προς γενίκευση της παρούσας μελέτης.

Τέλος, η πιθανότητα πώλησης ενός ακινήτου, όπως αυτή εξελίσσεται σε μία συγκεκριμένη χρονική περίοδο από την ημέρα δημοσίευσής της αγγελίας του, ανάλογα με την τιμή του, αποτελεί μία δυνατότητα που θα μπορούσε να προστεθεί σε μελλοντικές εφαρμογές.

Βιβλιογραφία

[1] P. Zentelis, "Η Αξία Των Ακινήτων," pp. 85–102, 2015.

[2] Y. Feng and K. Jones, "Comparing multilevel modelling and artificial neural networks in house price prediction," *ICSDM 2015 - Proc. 2015 2nd IEEE Int. Conf. Spat. Data Min. Geogr. Knowl. Serv.*, pp. 108–114, 2015, doi: 10.1109/ICSDM.2015.7298035.

[3] W. T. Lim, L. Wang, Y. Wang, and Q. Chang, "Housing price prediction using neural networks," *2016 12th Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov. ICNC-FSKD 2016*, pp. 518–522, 2016, doi: 10.1109/FSKD.2016.7603227.

[4] T. Cai, K. Han, and H. Wu, "Melbourne Airbnb Price Prediction," pp. 1–6, [Online]. Available: https://drive.google.com/open?id=1D32jVpSfvEYCDCoVt6FYxS98KVFhp3Fm.

[5] S. Mullainathan and J. Spiess, "Machine learning: An applied econometric approach," *J. Econ. Perspect.*, vol. 31, no. 2, pp. 87–106, 2017, doi: 10.1257/jep.31.2.87.

[6] R. Basu and J. Ferreira, "ScienceDirect Understanding household vehicle ownership in Singapore through a comparison of econometric and machine learning models," *Transp. Res. Procedia*, vol. 00, no. May, pp. 1674–1693, 2019, doi: 10.1016/j.trpro.2020.08.207.

[7] Studenmund, *Using Econometrics: A Practical Guide*, 7th ed. Harlow, United Kingdom : Pearson, 2017.

[8] Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*. 2002.

[9] William H. Green, *Econometric Analysis*, vol. 97, no. 457. 2002.

[10] V. Limsombunc, C. Gan, and M. Lee, "House Price Prediction: Hedonic Price Model vs. Artificial Neural Network," *Am. J. Appl. Sci.*, vol. 1, no. 3, pp. 193–201, 2004, doi: 10.3844/ajassp.2004.193.201.

[11] K. W. Chau and T. L. Chin, "A Critical Review of Literature on the Hedonic Price Model," *Int. J. Hous. Sci. Its Appl.*, vol. 2, no. 27, pp. 145–165, 2003, [Online]. Available: http://papers.ssrn.com/abstract=2073594.

[12] N. Dunse and C. Jones, "A hedonic price model of office rents.," 2006.

[13] B. W. Brorsen, W. R. Grant, and M. E. Rister, "A Hedonic Price Model for Rough Rice Bid/Acceptance Markets," *Am. J. Agric. Econ.*, vol. 66, no. 2, pp. 156–163, 1984, doi: 10.2307/1241032.

[14] J. W. Tukey, *Eploratory Data Analysis*. 1977.

[15] B. S. Xia and P. Gong, "Review of business intelligence through data analysis," *Benchmarking*, vol. 21, no. 2, pp. 300–311, 2014, doi: 10.1108/BIJ-08-2012-0050.

[16] R. Schutt and Cathy O'Neil, *Doing Data Science*, vol. 53, no. 9. O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2014.

[17] J. Ashirwadam, "Communication Research Methods Methods of Data Analysis," *Tamilnadu Theol. Semin.*, no. August, pp. 1–6, 2014.

[18] Φ. Ίσαρη and Μ. Πούρκος, *Ποιοτική Μεθοδολογία Έρευνας*. 2015.

[19] J. W. Tukey, "The Future of Data Analysis," *Ann. Math. Stat.*, vol. 33, no. 1, pp. 1–67, 1962, doi: 10.1214/aoms/1177704711.

[20] Ευστάθιος Γ. Κύρκος, *Επιχειρηματική Ευφυΐα & Εξόρυξη Δεδομένων*, vol. 53, no. 9. 2015.

[21] M. I. T. C. Data, *MIT Critical Data Secondary Analysis of Electronic Health Records*. .
[22] N.J.Nilson, "Introduction machine learning," *Science (80-.).*, pp. 1996–1996, 1996.
[23] E. Hunt, *Machine learning: An artificial intelligence approach (vol. 2)*, vol. 31, no. 3. 1987.
[24] M. Stephen, *Machine Learning An Algorithmic Perspective Second Edition*. 2014.
[25] Γεωργούλη Κατερίνα, *Τεχνητή Νοημοσύνη*. 2015.
[26] X. Goldberg, *Introduction to semi-supervised learning*, vol. 6. 2009.
[27] Δ. Πετρίδης, *Ανάλυση Πολυμεταβλητών Τεχνικών - Εφαρμογές Περιπτώσεων*. 2015.
[28] D. A. Freedman, *Statistical Models*. Cambridge University Press, New York, 2009.
[29] H. Ishijima and A. Maeda, "Real Estate Pricing Models: Theory, Evidence, and Implementation," *Asia-Pacific Financ. Mark.*, vol. 22, no. 4, pp. 369–396, 2015, doi: 10.1007/s10690-013-9170-7.
[30] D. W. Marquardt and R. D. Snee, "Ridge regression in practice," *Am. Stat.*, vol. 29, no. 1, pp. 3–20, 1975, doi: 10.1080/00031305.1975.10479105.
[31] C. Huang and A. Mintz, "Ridge regression analysis of the defence-growth tradeoff in the United States," *Def. Econ.*, vol. 2, no. 1, pp. 29–37, 1990, doi: 10.1080/10430719008404676.
[32] M. R. Osborne, B. Presnell, and B. A. Turlach, "On the LASSO and its Dual," *J. Comput. Graph. Stat.*, vol. 9, no. 2, pp. 319–337, 2000, doi: 10.1080/10618600.2000.10474883.
[33] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005, doi: 10.1111/j.1467-9868.2005.00503.x.
[34] R. E. Schapire, "The Strength of Weak Learnability," *Mach. Learn.*, vol. 5, no. 2, pp. 197–227, 1990, doi: 10.1023/A:1022648800760.
[35] G. Biau, "Analysis of a Random Forests Model," *Anaesthesiol. Intensive Ther.*, vol. 49, no. 5, pp. 373–381, 2017, doi: 10.5603/AIT.a2017.0074.
[36] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, no. 6, pp. 323–329, 2012, doi: 10.1016/j.ygeno.2012.04.003.
[37] Y. L. Pavlov, "Random forests," *Random For.*, pp. 1–122, 2019, doi: 10.1201/9780367816377-11.
[38] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front. Neurorobot.*, vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.
[39] A. Keprate and R. M. C. Ratnayake, "Using gradient boosting regressor to predict stress intensity factor of a crack propagating in small bore piping," *IEEE Int. Conf. Ind. Eng. Eng. Manag.*, vol. 2017-Decem, pp. 1331–1336, 2018, doi: 10.1109/IEEM.2017.8290109.
[40] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.
[41] L. Breiman, "Stacked regressions," *Mach. Learn.*, vol. 24, no. 1, pp. 49–64, 1996, doi: 10.1023/A:1018046112532.
[42] D. Wolpert, "Stacked Generalization (Stacking)," *Neural Networks*, vol. 5, pp. 241–259, 1992.
[43] J. F. H. JR., W. C. Black, B. J. Babin, and R. E. Anderson, "Multivariate Data Analysis." pp. 135–144, doi: 10.1016/j.foodchem.2017.03.133.