Μάθημα: Μηχανική Μάθηση

Άσκηση 1 : Πρόβλημα Παλινδρόμησης

Ονοματεπώνυμο: Κωνσταντίνος Βαρδάκας

AM: 522

Email: pcs0522@uoi.gr

Περιγραφή Dataset

Το παρόν project βασίζεται στο γνωστό διαγωνισμό "House Prices - Advanced Regression Techniques" που φιλοξενείται στην πλατφόρμα Kaggle. Το dataset αποτελείται από δεδομένα που αφορούν κατοικίες στην πόλη Ames, Iowa, και περιλαμβάνει πλήθος χαρακτηριστικών που μπορούν να χρησιμοποιηθούν για την πρόβλεψη της τελικής τιμής πώλησης ενός ακινήτου. Για το συγκεκριμένο project, χρησιμοποιήθηκε το train.csv, το οποίο περιέχει 79 χαρακτηριστικά (80 εάν συμπεριληφθεί και το id της εγγραφής), και την τιμή πώλησης (SalePrice), η οποία είναι και το χαρακτηριστικό στόχος.

Οι μεταβλητές του dataset καλύπτουν διάφορες κατηγορίες πληροφορίας, όπως την τοποθεσία και το οικόπεδο, την κατάσταση και την ποιότητα, χαρακτηριστικά σχετικά με τους χώρους του σπιτιού, υλικά που χρησιμοποιήθηκαν για την κατασκευή, και χρονικές πληροφορίες σχετικά με την κατασκευή του σπιτιού. Αυτές οι μεταβλητές είναι είτε αριθμητικές (ακέραιες ή δεκαδικές), είτε κατηγορικές (ονομαστικές ή διατεταγμένες). Αναφορικά με το χαρακτηριστικό στόχο, οι τιμές κυμαίνονται από 34900 έως 755000 δολάρια, και παρουσιάζει μία δεξιά ασυμμετρία, όπως φαίνεται στο ιστόγραμμα της εικόνας 1.

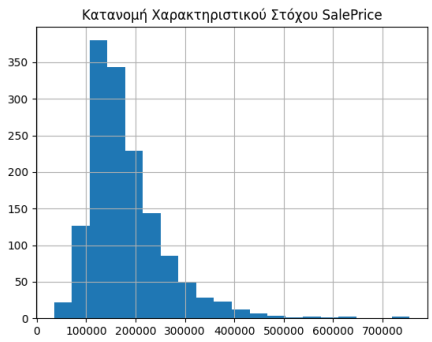

Figure 1: Κατανομή του χαρτηριστικού SalePrice

Προεπεξεργασία Δεδομένων

Η προεπεξεργασία των δεδομένων αποτελεί κρίσιμο βήμα για την ανάπτυξη μοντέλων μηχανικής μάθησης, καθώς επηρεάζει άμεσα όχι μόνο την απόδοση και τη γενίκευση των μοντέλων, αλλά και την ταχύτητα και την ικανότητα σύγκλισής τους κατά την εκπαίδευση. Ιδιαίτερα στην περίπτωση γραμμικών μοντέλων, είναι απαραίτητη η προσεκτική αντιμετώπιση των μεταβλητών, ώστε να ικανοποιούνται βασικές υποθέσεις όπως η γραμμικότητα, η ομοσκεδαστικότητα και η ανεξαρτησία των σφαλμάτων. Επιπλέον, τα γραμμικά μοντέλα είναι πιο ευαίσθητα σε ακραίες τιμές (outliers), γεγονός που καθιστά τον καθαρισμό, τη μετατροπή και την κωδικοποίηση των χαρακτηριστικών αναγκαία στάδια. Στο παρόν κεφάλαιο περιγράφονται ορισμένα σημαντικά βήματα προεπεξεργασίας που πραγματοποιήθηκαν. Για λόγους συντομίας, δεν παρουσιάζονται αναλυτικά όλα τα επιμέρους στάδια.

Αρχικά, εντοπίστηκαν και απομακρύνθηκαν χαρακτηριστικά με εξαιρετικά χαμηλή διακύμανση (quasi-constant features), δηλαδή μεταβλητές που παρουσιάζουν σχεδόν την ίδια τιμή σε ποσοστό άνω του 99% των παρατηρήσεων. Τέτοια χαρακτηριστικά έχουν ελάχιστη προβλεπτική αξία και ενδέχεται να εισάγουν θόρυβο ή να προκαλούν υπερπροσαρμογή (overfitting). Μερικά από αυτά τα χαρακτηριστικά είναι τα PoolQC και PoolArea (απουσία πισίνας στο 99.5% των περιπτώσεων), το Street (99.6% λιθόστρωτο) και το Utilities (99.9% AllPub).

Ένα από τα χαρακτηριστικά του dataset, το MSSubClass, αναπαριστά τον τύπο κατοικίας με βάση το στυλ και τη χρήση (π.χ. μονοκατοικία, διπλοκατοικία, split-level κτλ.), αλλά σε μορφή αριθμητικών κωδικών, χωρίς όμως να υποδηλώνεται κάποια ποσοτική ή ιεραρχική σχέση μεταξύ των τιμών. Για λόγους απλοποίησης και καλύτερης ερμηνείας, το χαρακτηριστικό μετασχηματίστηκε έτσι ώστε να εκφράζει τον αριθμό ορόφων της κατοικίας. Το mapping βασίστηκε σε κατηγοριοποίηση των τύπων κατοικιών σύμφωνα με την εξής αντιστοίχιση:

1-Story	1.5-Story	2-Story	2.5-Story	Split/Multi-Level	Multi-Family
0	1	2	3	4	5

Table 1: Mapping του χαρακτηριστικου MSSubclass, βάσει ορόφων

Αν και το νέο χαρακτηριστικό φαινομενικά αντιπροσωπεύει διατεταγμένες κατηγορίες (ordinal), η γραμμική συσχέτισή του με την τιμή πώλησης αποδείχθηκε αμελητέα (συντελεστής Pearson = -0.034). Συνεπώς, επιλέχθηκε η one-hot κωδικοποίηση για την ενσωμάτωσή του στο μοντέλο, ώστε να αποφευχθεί η επιβολή τεχνητής ιεραρχίας η οποία δεν αντανακλά τα πραγματικά μοτίβα των δεδομένων. Η απουσία γραμμικής συσχέτισης μπορεί να γίνει αντιληπτή και από τα παρακάτω boxplots.

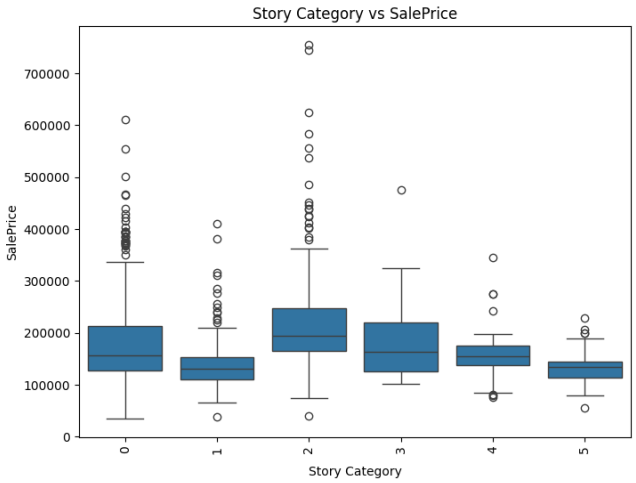

Figure 2: Box Plots που αναδεικνύουν απουσία γραμμικής σχέσης μεταξύ του χαρακτηριστικού Story Level με το χαρακτηριστικό στόχο

Το χαρακτηριστικό LotArea, το οποίο αντιπροσωπεύει το εμβαδόν του οικοπέδου, παρουσίαζε σημαντικές ακραίες τιμές (outliers) που μπορούσαν να επηρεάσουν δυσανάλογα την εκπαίδευση του μοντέλου, ειδικά στην περίπτωση χρήσης συναρτήσεων κόστους βασισμένων στο τετραγωνισμένο σφάλμα (squared error). Για τον λόγο αυτό, εφαρμόστηκε λογαριθμικός μετασχηματισμός, προκειμένου να μειωθεί η ασυμμετρία της κατανομής και να περιοριστεί η επίδραση των outliers, καθιστώντας τη μεταβλητή πιο κατάλληλη για μοντελοποίηση.

Στη συνέχεια, όπως και για κάθε άλλο αριθμητικό χαρακτηριστικό, εφαρμόστηκε κανονικοποίηση (μετασχηματισμός σε κατανομή με mean ± std = 0 ± 1). Η διαδικασία αυτή διευκολύνει τη σύγκλιση των αλγορίθμων εκπαίδευσης, ειδικά σε μοντέλα που βασίζονται σε βελτιστοποίηση μέσω παραγώγων (όπως και τα γραμμικά), και μειώνει τον κίνδυνο παγίδευσης σε τοπικά ελάχιστα κατά την εύρεση των παραμέτρων του μοντέλου. Επιπρόσθετα, η κλιμάκωση των χαρακτηριστικών στην

ίδιο εύρος τιμών, επιτρέπει την εξαγωγή συμπερασμάτων από τους συντελεστές των χαρακτηριστικών, σχετικά με τη σημασία τους στην τελική πρόβλεψη.

Στο ακόλουθο διάγραμμα, το οποίο αναδεικνύει τη σχέση μεταξύ LotArea και SalePrice, παρατηρείται ότι το confidence interval διευρύνεται καθώς αυξάνεται το LotArea. Αυτές οι υψηλές και ακραίες τιμές μπορούν να επηρεάσουν σημαντικά τη συμπεριφορά του μοντέλου, δεδομένου ότι το MSE δίνει μεγαλύτερο βάρος σε μεγάλες αποκλίσεις.

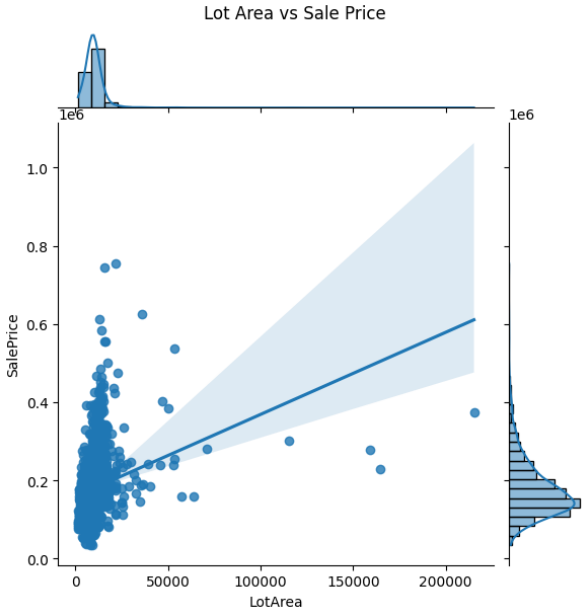

Figure 3: Συσχέτιση του χαρακτηριστικού LotArea με το χαρακτηριστικό στόχο SalePrice

Ανάμεσα στα χαρακτηριστικά OverallQual (συνολική ποιότητα) και OverallCond (συνολική κατάσταση), επιλέχθηκε να χρησιμοποιηθεί μόνο το OverallQual, καθώς εμφανίζει σαφέστερη γραμμική συσχέτιση με το SalePrice, όπως φαίνεται στο ακόλουθο σχήμα.

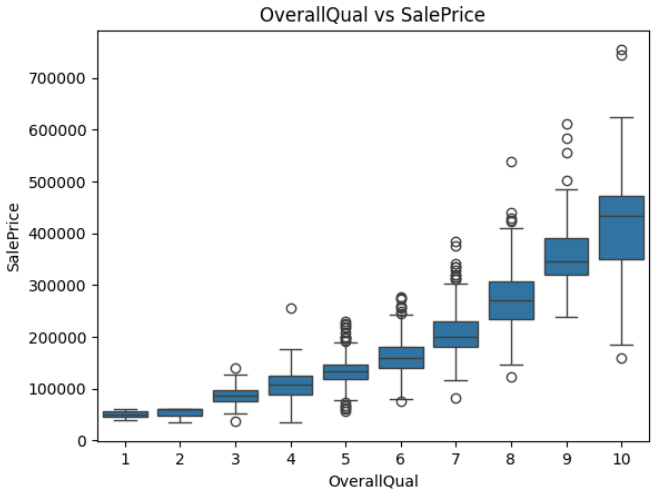

Figure 4: Συσχέτιση του χαρακτηριστικού OverallQual με το χαρακτηριστικό στόχο SalePrice

Επιπλέον, το OverallCond συγκεντρώνει το μεγαλύτερο μέρος των τιμών στην κατηγορία 5, η οποία εμφανίζει και υψηλό std (Saleprice με OverallCond = 5: std = 85,117 έναντι std του dataset = 79,442). Το χαρακτηριστικό OverallQual που κρατήθηκε, κανονικοποιήθηκε με min-max scaling, δεδομένου ότι είναι γνωστές και σταθερές οι ελάχιστες και μέγιστες επιτρεπτές τιμές της μεταβλητής (1 έως 10).

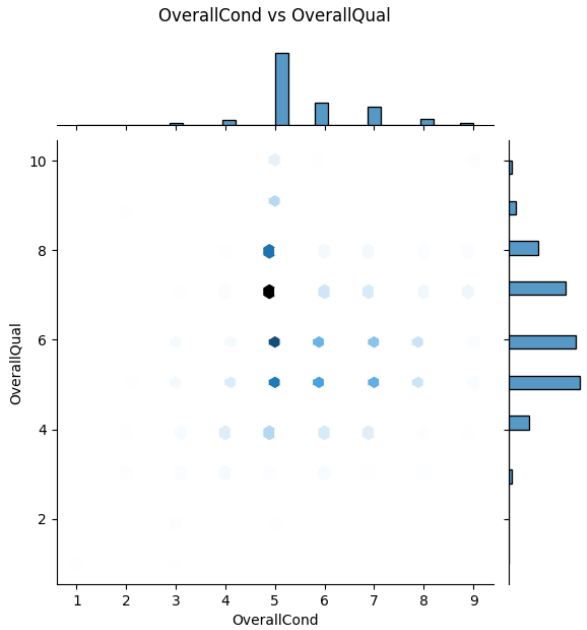

Figure 5: Διαφορές κατανομών των χαρακτηριστικών OverallQual και OverallCond.

Αναφορικά με τα χρονολογικά χαρακτηριστικά (όπως YearBuilt, YearRemodAdd, GarageYrBlt), εφαρμόστηκε μετασχηματισμός σε σχετικές χρονικές διαφορές ως προς το YrSold, ώστε τα δεδομένα να αποτυπώνουν την ηλικία του κάθε

χαρακτηριστικού κατά τον χρόνο πώλησης. Για παράδειγμα, το YearBuilt μετασχηματίστηκε σε YrSold - YearBuilt.

Παρόλο που τα αρχικά χαρακτηριστικά αντικαταστάθηκαν από τις σχετικές χρονικές τιμές, το YrSold διατηρήθηκε, καθώς η χρονιά πώλησης ενδέχεται να περιέχει πληροφορία σχετική trends της αγοράς κάθε χρονιάς. Τα τρία χρονολογικά χαρακτηριστικά που αποτυπώνουν την χρονολογική διαφορά (YearBuiltDiff, YearRemodDiff, GarageYrBltDiff) κανονικοποιήθηκαν στη συνέχεια μέσω standard scaling. Αντίθετα, το YrSold, κωδικοποιήθηκε με one-hot encoding, προκειμένου να αποτυπωθούν ενδεχόμενες μη γραμμικές τάσεις ανά έτος στην αγορά ακινήτων.

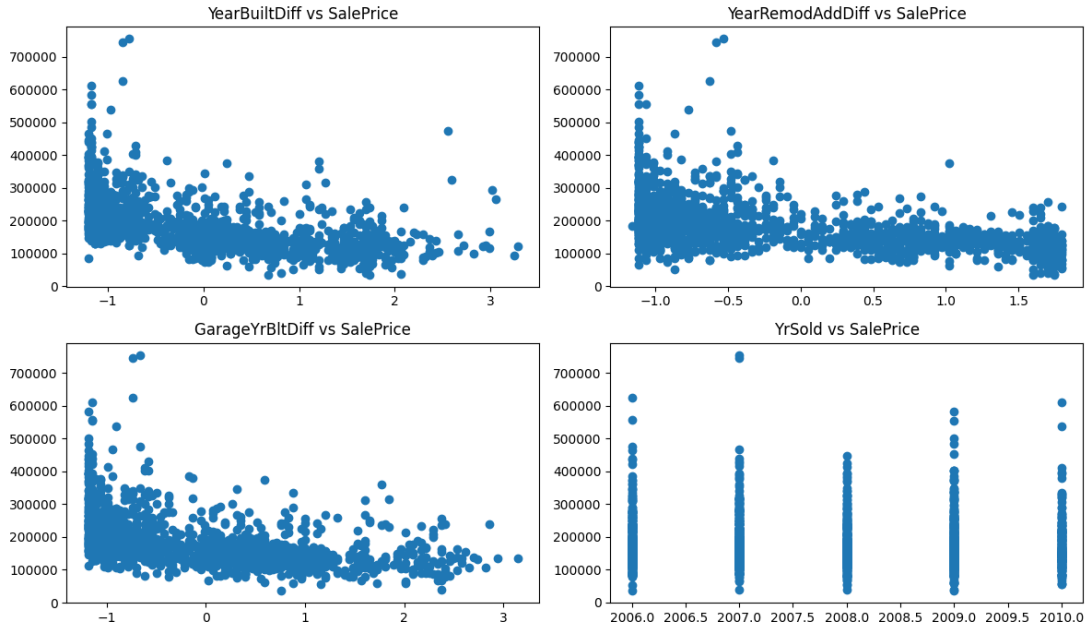

Figure 6: Συσχέτιση των Χρονικών χαρακτηριστικών με το χαρακτηριστικό στόχο

Το χαρακτηριστικό GarageYrBltDiff, το οποίο εκφράζει τη διαφορά ανάμεσα στο έτος πώλησης και το έτος κατασκευής του γκαράζ, παρουσιάζει ελλείπουσες τιμές στις περιπτώσεις όπου δεν υπάρχει γκαράζ. Οι τιμές αυτές αντικαταστάθηκαν με το 0, καθώς η απουσία γκαράζ αναπαρίσταται από το χαρακτηριστικό GarageCars, στο οποίο προστέθηκε και η κατηγορία 0 για τα ακίνητα χωρίς γκαράζ.

Μεταξύ των χαρακτηριστικών GarageCars (αριθμός οχημάτων που χωράει το γκαράζ) και GarageArea (εμβαδόν γκαράζ), παρατηρήθηκε ισχυρή συσχέτιση με συντελεστή Pearson correlation = 0.88, γεγονός που τα καθιστά πλεονασματικά χαρακτηριστικά. Ωστόσο, προτιμήθηκε η χρήση του GarageCars, καθώς παρουσίασε

υψηλότερη προβλεπτική ισχύ σε απλή γραμμική παλινδρόμηση με το SalePrice (R^2 = 0.419 έναντι 0.385 του GarageArea).

Το τελευταίο χαρακτηριστικό που σχετίζεται με τον χρόνο είναι ο μήνας πώλησης (MoSold). Για να μειωθεί η πολυπλοκότητα του χαρακτηριστικού και να αποτυπωθεί καλύτερα η εποχικότητα, επιλέχθηκε να ομαδοποιηθούν οι μήνες σε εποχές. Αντί της απλής κατηγοριοποίησης ανά τρίμηνο ή τετράμηνο, ακολουθήθηκε μια πιο ευέλικτη προσέγγιση μέσω κυκλικού μετασχηματισμού και μη εποπτευόμενης μάθησης.

Οι μήνες μετασχηματίστηκαν σε κυκλική μορφή, προκειμένου να διατηρηθεί η φυσική περιοδικότητα του χρόνου. Συγκεκριμένα, κάθε μήνας μετατράπηκε αρχικά σε γωνία μέσω του τύπου:

$$\vartheta = \frac{2 \cdot \pi \cdot \mu\acute{\eta}\nu\alpha\varsigma}{12}$$

Στη συνέχεια εφαρμόστηκαν οι συναρτήσεις ημίτονο και συνημίτονο, δημιουργώντας δύο νέα χαρακτηριστικά. Με αυτόν τον μετασχηματισμό, μήνες όπως ο Δεκέμβριος και ο Ιανουάριος απεικονίζονται ως χρονικά κοντινοί, κάτι που δεν ισχύει στον αρχικό γραμμικό χώρο.

Αξιοποιώντας αυτή τη μετατροπή, εφαρμόστηκε clustering στον τρισδιάστατο χώρο που περιλαμβάνει τις τιμές cos(x), sin(x) και SalePrice, με στόχο την ομαδοποίηση μηνών που είναι ταυτόχρονα χρονικά κοντινοί και παρουσιάζουν παρόμοιες μέσες τιμές πώλησης. Για να αποφευχθεί η κυριαρχία του SalePrice στο αποτέλεσμα, πραγματοποιήθηκε κανονικοποίησή του πριν την εφαρμογή του clustering αλγορίθμου. Τέλος, ο αριθμός των clusters ορίστηκε σε τέσσερις, και οι τελικές εποχιακές κατηγορίες κωδικοποιήθηκαν μέσω one-hot encoding. Η ομαδοποίηση που προέκυψε φαίνεται στο παρακάτω διάγραμμα.

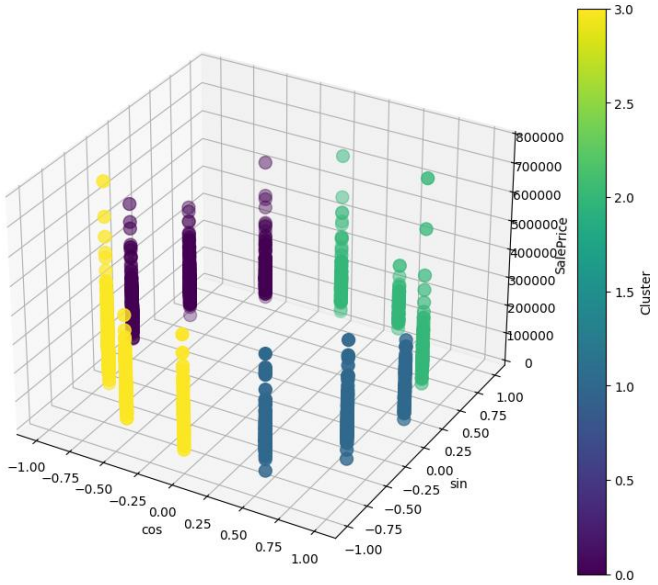

Figure 7: Ομαδικοποίηση των εποχών, με βάση τη τιμή πώλησης ακινήτων.

Επίσης, στο παρακάτω διάγραμμα φαίνεται η ομαδοποίηση στον αρχικό χώρο, χρησιμοποιώντας τον μέσο όρο του Standarized Sale Price κάθε μήνα.

Figure 8: Ομαδικοποίηση των εποχών, με βάση τη τιμή πώλησης ακινήτων στον αρχικό χώρο

Αναφορικά με τα υπόλοιπα χαρακτηριστικά, εφαρμόστηκαν επιπλέον τεχνικές επιλογής και επεξεργασίας, με στόχο τη μείωση της διαστατικότητας και τη βελτίωση της ποιότητας των εισόδων προς το μοντέλο. Συγκεκριμένα, αριθμητικά χαρακτηριστικά που παρουσίαζαν χαμηλή γραμμική συσχέτιση με την τιμή πώλησης (Pearson correlation μικρότερο του 0.4) αφαιρέθηκαν, καθώς κρίθηκε ότι η συμβολή τους στην πρόβλεψη είναι περιορισμένη και ενδέχεται να εισάγουν θόρυβο ή να ενισχύσουν το overfitting.

Παράλληλα, τα κατηγορικά χαρακτηριστικά υπέστησαν ομαδοποιήσεις με βάση είτε εννοιολογική συνάφεια (π.χ. παρόμοια υλικά κατασκευής ταξινομήθηκαν μαζί), είτε οικονομική προσέγγιση (π.χ. περιοχές ομαδοποιήθηκαν σε "υψηλού", "μεσαίου" και "χαμηλού" κόστους, ανάλογα με τις μέσες τιμές των ακινήτων).

Τέλος, κατηγορικά χαρακτηριστικά που διέθεταν ταξινομημένη κλίμακα τιμών (π.χ. βαθμός ποιότητας ή κατάστασης από το 1 έως το 5) και παρουσίαζαν σχετικά σταθερή γραμμική συσχέτιση με την τιμή πώλησης, κωδικοποιήθηκαν ως ordinal μεταβλητές, ώστε να διατηρηθεί η ιεραρχία των κατηγοριών.

Επιλογή Χαρακτηριστικών

Σε αυτό το στάδιο, τα χαρακτηριστικά έχουν προεπεξεργαστεί πλήρως και είναι πλέον κατάλληλα για συγκριτική αξιολόγηση και επιλογή. Οι ελλείπουσες τιμές δεδομένα έχουν συμπληρωθεί, τα κατηγορικά χαρακτηριστικά έχουν ομαδοποιηθεί με βάση την πληροφορία που μεταφέρουν, είτε ως προς τη σημασία τους, είτε σε σχέση με το στόχο (SalePrice), και τα αριθμητικά χαρακτηριστικά έχουν μετασχηματιστεί ώστε να παρουσιάζουν, όσο είναι εφικτό, γραμμική συσχέτιση με την τιμή πώλησης.

Παράλληλα, έχουν εντοπιστεί και μειωθεί οι επιδράσεις των ακραίων τιμών, ενώ όλες οι μεταβλητές έχουν κανονικοποιηθεί, με αποτέλεσμα να κινούνται σε παρόμοια κλίμακα μεγέθους. Αυτό είναι ιδιαίτερα σημαντικό, καθώς καθιστά δυνατή τη συνεπή σύγκριση μεταξύ αριθμητικών και κατηγορικών χαρακτηριστικών και επιτρέπει την ορθή εφαρμογή τεχνικών επιλογής χαρακτηριστικών, χωρίς να επηρεάζονται δυσανάλογα από διαφορές στην κλίμακα.

Το πρώτο βήμα στην επιλογή χαρακτηριστικών βασίστηκε στη χρήση του συντελεστή συσχέτισης Pearson. Συγκεκριμένα, εξετάστηκαν όλα τα πιθανά ζεύγη χαρακτηριστικών μεταξύ τους, και επιλέχθηκαν εκείνα που παρουσίαζαν απόλυτη συσχέτιση μεγαλύτερη του 0.8. Σε κάθε ζεύγος με υψηλή συσχέτιση, διατηρήθηκε μόνο το χαρακτηριστικό που εμφάνιζε τον υψηλότερο συντελεστή Pearson ως προς την τιμή πώλησης (SalePrice), ενώ το άλλο αφαιρέθηκε ως πλεονάζον.

Η επιλογή αυτής της μεθόδου στηρίχθηκε στο γεγονός ότι ο συντελεστής Pearson μετρά τη γραμμική εξάρτηση μεταξύ μεταβλητών, κάτι που είναι ιδιαίτερα

χρήσιμο για γραμμικά μοντέλα όπως η γραμμική παλινδρόμηση. Ωστόσο, είναι σημαντικό να σημειωθεί πως αυτή η προσέγγιση δεν εντοπίζει μη γραμμικές σχέσεις (π.χ. λογαριθμικές ή εκθετικές), γεγονός που την καθιστά περιορισμένη όταν πρόκειται για την επιλογή χαρακτηριστικών σε πιο πολύπλοκα μοντέλα, όπως νευρωνικά δίκτυα ή Gaussian Processes με RBF kernels.

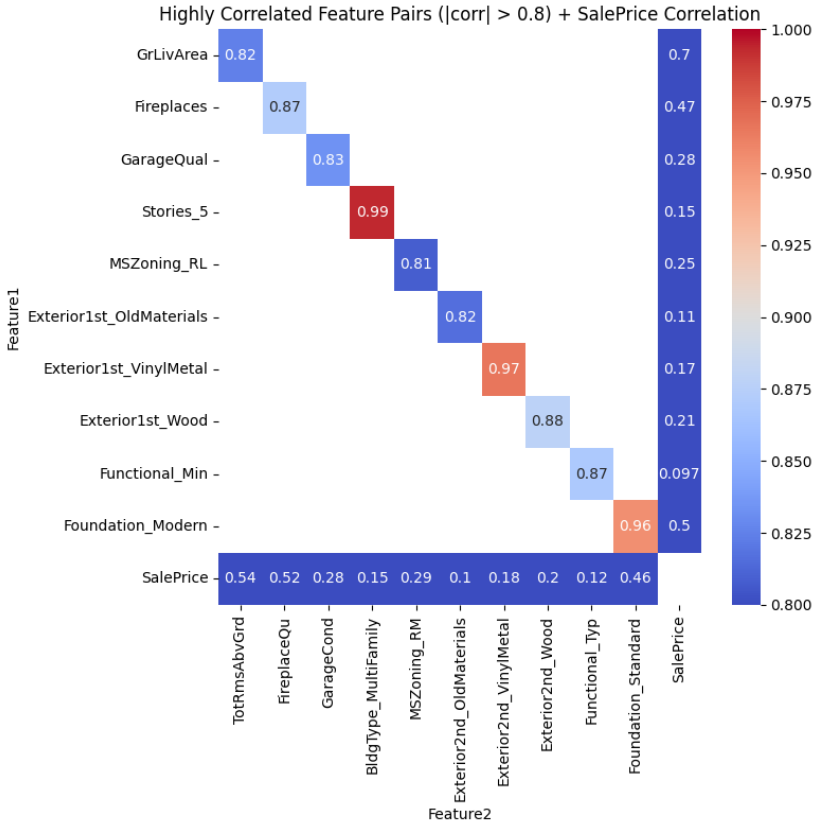

Figure 9: Ζεύγη χαρακτηριστικών με υψηλή απόλυτη τιμή συντελεστή Pearson Correlation και συσχέτιση αυτών με το χαρακτηριστικό στόχο

Για την περαιτέρω μείωση της διαστασιμότητας και την ανάδειξη των χαρακτηριστικών με τη μεγαλύτερη προβλεπτική ισχύ, εφαρμόστηκε στη συνέχεια η μέθοδος Sequential Feature Selection (SFS) με κατεύθυνση forward, χρησιμοποιώντας γραμμική παλινδρόμηση ως estimator. Η επιλογή έγινε από το πλήρες σύνολο των χαρακτηριστικών, και αξιολογήθηκε μέσω 5-fold cross-validation, με γνώμονα την επίδοση ως προς τον συντελεστή R^2.

Στο παρακάτω διάγραμμα αποτυπώνεται η μεταβολή του cross-validation score σε σχέση με τον αριθμό των επιλεγμένων χαρακτηριστικών. Παρατηρείται ότι η μέγιστη επίδοση επιτυγχάνεται με 34 χαρακτηριστικά, ενώ από το σημείο αυτό και

μετά, η προσθήκη νέων χαρακτηριστικών οδηγεί σε μείωση της ακρίβειας, πιθανόν λόγω εισαγωγής θορύβου ή πλεονάζουσας πληροφορίας.

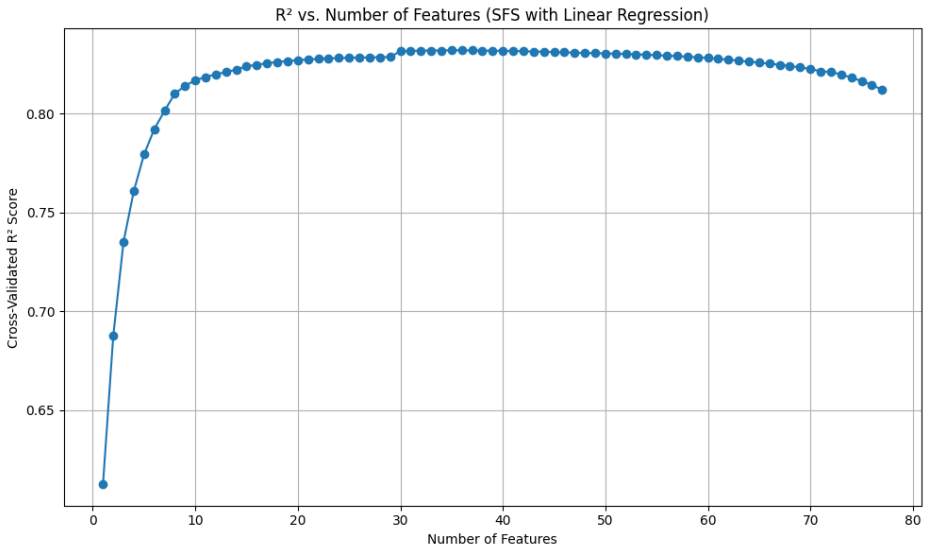

Figure 10: Διάγραμμα R^2 συναρτήσει του αριθμού των χαρακτηριστικών

Αξίζει να σημειωθεί ότι, όπως και στην περίπτωση της Pearson correlation, το ιδανικό υποσύνολο χαρακτηριστικών ενδέχεται να διαφέρει ανάλογα με τη φύση του μοντέλου (γραμμικό ή μη γραμμικό).

Μια ενδιαφέρουσα παρατήρηση αφορά την προσθήκη των one-hot encoded τιμών του χαρακτηριστικού SaleCondition. Συγκεκριμένα, με την προσθήκη του χαρακτηριστικού SaleCondition_Normal (29ο κατά σειρά), η βελτίωση στο score είναι οριακή (~$5*10^{-5}$). Ωστόσο, όταν ενσωματώνονται και η πληροφορία για τις υπόλοιπες δύο κατηγορίες, SaleCondition_NonNormal και SaleCondition_AllocaLand, ολοκληρώνεται η αναπαράσταση της κατηγορικής μεταβλητής και παρατηρείται σημαντική αύξηση του R^2 (~$2*10^{-3}$, 100 φορές μεγαλύτερη αύξηση).

Εκπαίδευση και Ανάλυση Προβλεπτικών Μοντέλων

Για κάθε ένα από τα μοντέλα που ακολουθούν, χρησιμοποιήθηκε κοινό training split και testing split, προκειμένου να διασφαλιστεί η συγκρισιμότητα μεταξύ των αποτελεσμάτων. Επίσης, στο training split εφαρμόστηκε το ίδιο αντικείμενο της κλάσης KFold cross-validation (με 10 folds), ώστε το αρχικό training set να χωρίζεται

κάθε φορά στα ίδια δέκα υποσύνολα και έτσι να εκπαιδεύονται και να επικυρώνονται στα ίδια δεδομένα.

Linear Regression

Το πρώτο μοντέλο που εξετάστηκε είναι το μοντέλο Linear Regression. Αυτό το μοντέλο επιχειρεί να προσαρμόσει μία υπερεπίπεδη επιφάνεια στα δεδομένα, ελαχιστοποιώντας το μέσο τετραγωνικό σφάλμα μεταξύ των προβλεπόμενων και πραγματικών τιμών.

Μετά από εκπαίδευση του γραμμικού μοντέλου στο training set με τα 34 επιλεγμένα χαρακτηριστικά, προκύπτει μέσο cross-validation $R^2 = 0.836 \pm 0.044$ (10-fold CV), υποδεικνύοντας μια σταθερή απόδοση στο εκπαιδευτικό σύνολο. Στη συνέχεια, πραγματοποιήθηκε fitting του μοντέλου σε ολόκληρο το training set και αξιολόγηση στο testing set, όπου προέκυψε $R^2 = 0.87$, αναδεικνύοντας καλή ικανότητα γενίκευσης.

Στο ιστόγραμμα των residuals (διαφορά μεταξύ πραγματικών και προβλεπόμενων τιμών), παρατηρείται ότι ακολουθούν σε γενικές γραμμές κανονική κατανομή, με μερικά outliers και στις δύο κατευθύνσεις. Αυτά τα ακραία σφάλματα φτάνουν τιμές της τάξης των ±100.000, γεγονός που υποδηλώνει πιθανή ευαισθησία του μοντέλου σε μεμονωμένες περιπτώσεις με εξαιρετικά υψηλή ή χαμηλή τιμή πώλησης.

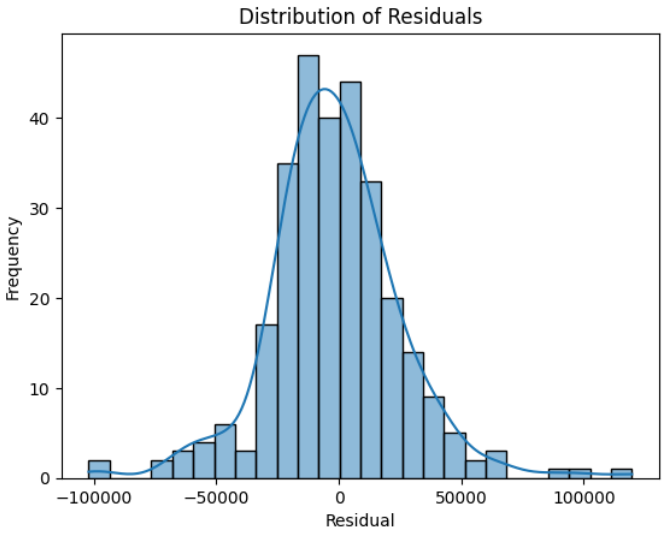

Figure 11: Ιστόγραμμα των residuals

Το QQ-plot ενισχύει αυτή την παρατήρηση, καθώς εμφανίζονται αποκλίσεις από την ευθεία της κανονικότητας κυρίως στα άκρα της κατανομής, επιβεβαιώνοντας την παρουσία outliers. Η ύπαρξη αυτών των τιμών μπορεί να έχει σημαντική επίδραση στους συντελεστές του μοντέλου, όπως έχει ήδη αναφερθεί.

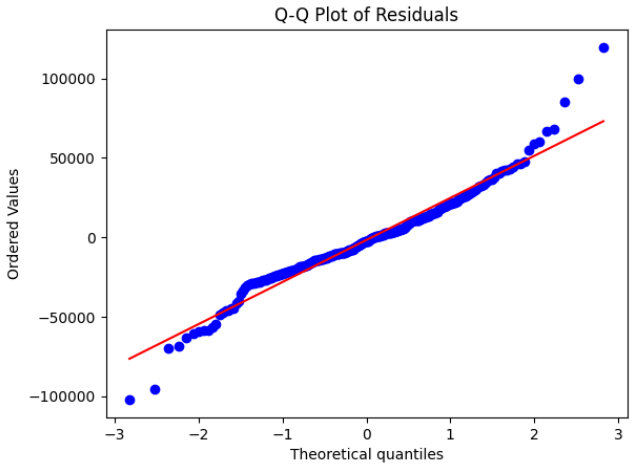

Figure 12: QQ-plot των residuals

Με σκοπό τον περιορισμό της επιρροής των outliers στους συντελεστές του μοντέλου, πραγματοποιήθηκε λογαριθμικός μετασχηματισμός στο target. Ο μετασχηματισμός αυτός εξομαλύνει την κατανομή της μεταβλητής-στόχου, η οποία πλέον πλησιάζει περισσότερο την κανονική κατανομή. Ως αποτέλεσμα, τα outliers συμμορφώνονται καλύτερα με τη συνολική τάση των δεδομένων και η επίδρασή τους

στην εκπαίδευση του μοντέλου μειώνεται σημαντικά, οδηγώντας σε πιο σταθερούς και αντιπροσωπευτικούς συντελεστές.

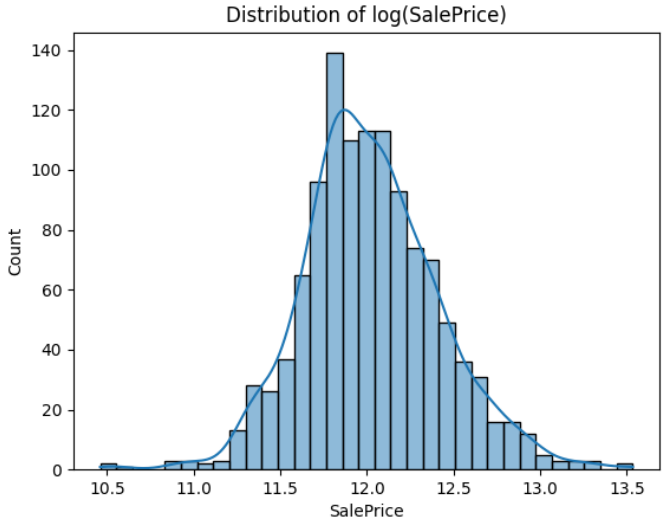

Figure 13: Κατανομή του χαρακτηριστικού στόχου μετά τον λογαριθμικό μετασχηματισμό

Μετά την εφαρμογή του λογαριθμικού μετασχηματισμού και την εκπαίδευση του μοντέλου, το cross-validation R^2 βελτιώθηκε προς 0.889 ± 0.026, δηλαδή σημειώθηκε αύξηση κατά 0.047, ενώ η τυπική απόκλιση μειώθηκε κατά 0.018, υποδεικνύοντας πιο σταθερά αποτελέσματα. Το test R^2 βελτιώθηκε επίσης, φτάνοντας το 0.897, καταγράφοντας μια αύξηση 0.02 σε σχέση με το R^2 πριν τον μετασχηματισμό. Αυτή η βελτίωση αποδεικνύει την αποτελεσματικότητα του λογαριθμικού μετασχηματισμού στην ενίσχυση της ικανότητας γενίκευσης του μοντέλου.

Παρακάτω παρουσιάζεται το ιστόγραμμα των residuals μετά τον λογαριθμικό μετασχηματισμό του target. Παρατηρείται ότι τα outliers στην αριστερή πλευρά της κατανομής έχουν περιοριστεί αισθητά, ενώ στην δεξιά πλευρά εξακολουθούν να υπάρχουν αποκλίσεις.

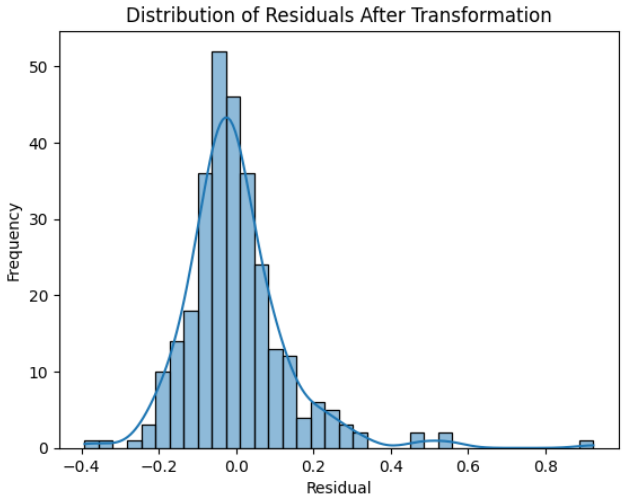

Figure 14: Ιστόγραμμα των residuals μετά τον λογαριθμικό μετασχηματισμό

Αξίζει να σημειωθεί ότι, έπειτα από τον λογαριθμικό μετασχηματισμό του SalePrice, το μοντέλο πλέον προβλέπει τη λογαριθμισμένη τιμή πώλησης.

Polynomial Regression

Η εφαρμογή πολυωνυμικής παλινδρόμησης συνεπάγεται εκθετική αύξηση του αριθμού των χαρακτηριστικών, ειδικά όταν εξετάζονται πολυώνυμα υψηλού βαθμού (π.χ. έως και 10ου). Για τον λόγο αυτό, αντί να χρησιμοποιηθούν τα 34 αρχικά χαρακτηριστικά, εφαρμόστηκε Principal Component Analysis (PCA), με στόχο τη μείωση της διαστασιμότητας και τη συγκράτηση της υπολογιστικής πολυπλοκότητας.

Συγκεκριμένα, επιλέχθηκαν οι 5 πρώτες κύριες συνιστώσες, οι οποίες διατηρούν το μεγαλύτερο ποσοστό της πληροφορίας του αρχικού dataset. Στη συνέχεια, τα δεδομένα μετασχηματίστηκαν σε πολυωνυμικά χαρακτηριστικά μέχρι και βαθμού 10, και εκπαιδεύτηκε γραμμικό μοντέλο σε αυτόν τον εμπλουτισμένο χώρο χαρακτηριστικών.

Έτσι, ο αριθμός των χαρακτηριστικών που προκύπτουν από την εφαρμογή πολυωνυμικού μετασχηματισμού βαθμού d=10, υπολογίζεται με βάση τον τύπο:

$$\binom{n+d}{d} = \frac{(n+d)!}{d!}$$

Όταν n = 34, προκύπτουν 2481256778, ενώ όταν n = 5 προκύπτουν μόνο 3003 χαρακτηριστικά. Μετά την παραγωγή των πολυωνυμικών χαρακτηριστικών για κάθε

βαθμό πολυωνύμου, είναι κρίσιμο να εφαρμόζεται κανονικοποίηση ώστε όλα τα χαρακτηριστικά να βρίσκονται στην ίδια κλίμακα, και έτσι να μπορεί να συγκλίνει σε κάποιο ελάχιστο του training set.

Κατά την εφαρμογή του PCA, οι αρχικές διαστάσεις των δεδομένων μετασχηματίζονται σε έναν νέο ορθογώνιο χώρο, όπου κάθε νέα διάσταση είναι ένας γραμμικός συνδυασμός των αρχικών χαρακτηριστικών. Οι κύριες συνιστώσες ταξινομούνται κατά φθίνουσα σειρά ως προς την ποσότητα διασποράς που εξηγούν στα δεδομένα. Συγκεκριμένα, η πρώτη κύρια συνιστώσα είναι εκείνη που ευθυγραμμίζεται με τη μέγιστη διακύμανση των δεδομένων και επομένως καταγράφει το σημαντικότερο μέρος της συνολικής πληροφορίας (χωρίς αυτή αναγκαστικά να σχετίζεται με το χαρακτηριστικό στόχο). Αυτό σημαίνει πως η κατεύθυνση αυτή περιέχει το κυρίαρχο μοτίβο των δεδομένων, ενώ οι επόμενες συνιστώσες καταγράφουν δευτερεύουσες διαφοροποιήσεις.

Στο ακόλουθο σχήμα απεικονίζεται το χαρακτηριστικό – στόχος ως συναρτήσει της κύριας συνιστώσας (PCA component 1), με τις προβλέψεις που προκύπτουν για κάθε βαθμό του πολυωνύμου. Αρχικά, παρατηρείται ότι υπάρχει γραμμικότητα μεταξύ της κύριας συνιστώσας και του στόχου, συνεπώς φαίνεται να έχει διατηρηθεί σημαντική πληροφορία για την πρόβλεψη του χαρακτηριστικού στόχου.

Παρατηρείται ότι όταν ο βαθμός του πολυωνύμου είναι μικρός, το μοντέλο φαίνεται να είναι ικανό να γενικεύει και να πλησιάζει τις πραγματικές τιμές του test split. Με την αύξηση του βαθμού του πολυωνύμου, δίνεται περισσότερη ευελιξία στο μοντέλο, και έτσι μαθαίνει τον θόρυβο του dataset και όχι την πληροφορία, δηλαδή κάνει overfitting.

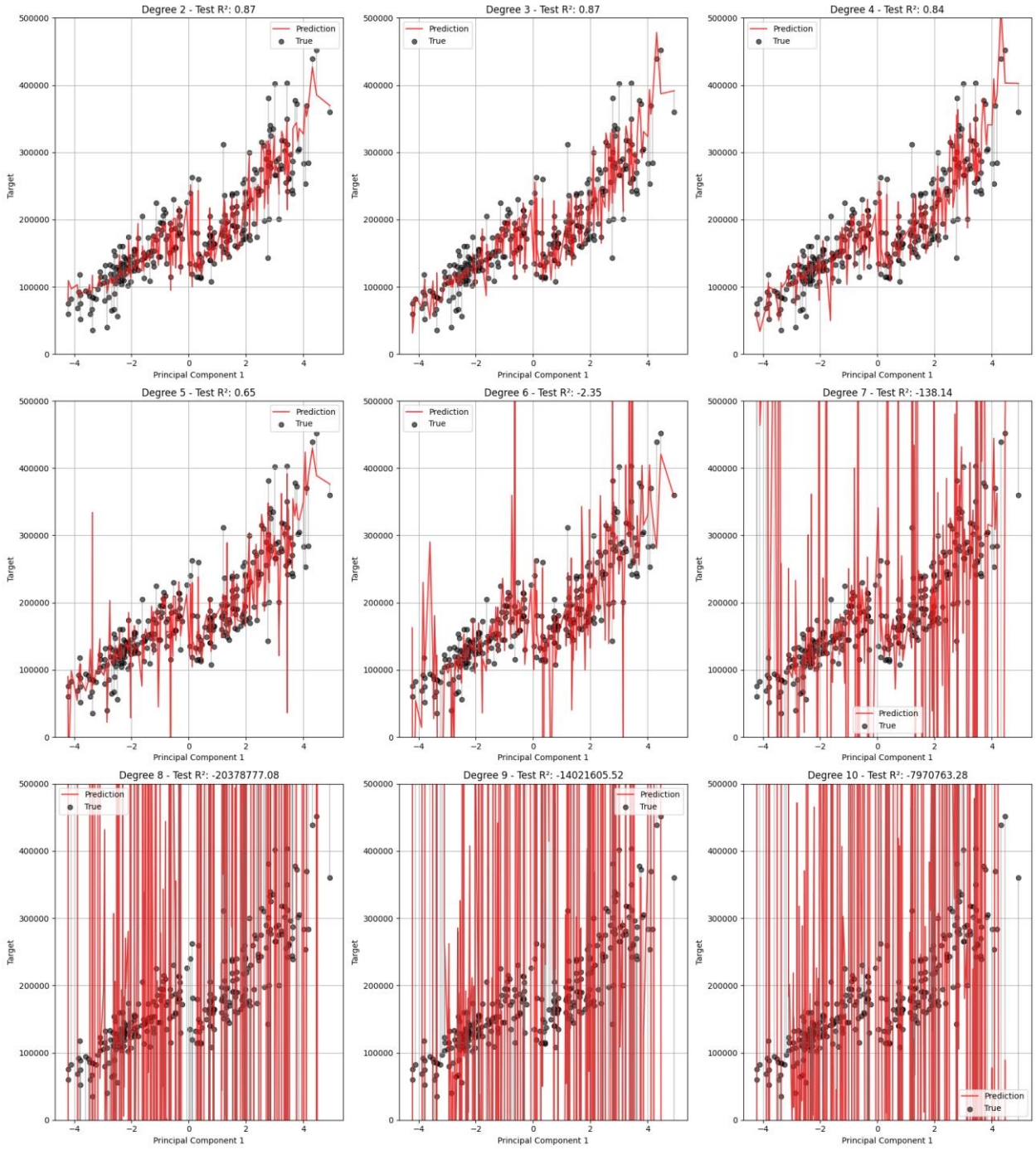

Figure 15: Γραμμή της συνάρτησης f(x) για κάθε σύνολο χαρακτηριστικών x, σε σύγκριση με το διάγραμμα PCA component 1 vs Saleprice

Αυτή η ευελιξία του μοντέλου με την αύξηση του πολυωνυμικού βαθμού, αποτυπώνεται και στον παρακάτω πίνακα. Συγκεκριμένα, το μοντέλο μηδενίζει το σφάλμα στο training set, αυξάνει το σφάλμα του test σετ, και αυξάνει το variance στα cross validation sets και έτσι το μοντέλο αδυνατεί να γενικεύσει σε νέα δεδομένα.

Degree	2	3	4	5	6	7	8	9	10
train R^2	0.844	0.862	0.897	0.928	0.955	0.983	1.0	1.0	1.0
test R^2	0.867	0.865	0.838	0.653	-2.34	$-1e^2$	$-2e^7$	$-1e^7$	$-8e^6$
cv mean	0.819	0.771	0.651	0.052	-32.3	$-2e^4$	$-1e^7$	$-5e^{18}$	$-9e^{18}$
cv std	0.135	0.230	0.446	1.168	59.3	$5.6e^5$	$2.9e^7$	$1.5e^{19}$	$2.8e^{19}$

Table 2: Μετρικές του μοντέλου για κάθε βαθμό πολυωνύμου που χρησιμοποιήθηκε

Χρησιμοποιώντας την λογαριθμημένη τιμή του χαρακτηριστικού στόχου, δεν παρατηρήθηκε κάποια διαφορά.

Lasso Regression

Η Lasso Regression είναι μία μέθοδος που χρησιμοποιείται για να βελτιώσει την ακρίβεια του μοντέλου σε άγνωστα δεδομένα, προσθέτοντας στον όρο του σφάλματος έναν όρο κανονικοποίησης. Η συνάρτηση απώλειας για το Lasso έχει τη μορφή:

$$Loss = \sum_{i=1}^{M} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=0}^{p} |w_j|$$

Ο πρώτος όρος είναι ο ίδιος με τον όρο του μοντέλου Linear Regression, ο οποίος αναπαριστά το τετραγωνικό σφάλμα μεταξύ των παρατηρήσεων και των προβλέψεων. Ο δεύτερος όρος είναι ο όρος κανονικοποίησης, ο οποίος προσθέτει μια ποινή για την απόλυτη τιμή των βαρών w_j. Η υπερπαράμετρος λ καθορίζει την ένταση της κανονικοποίησης και κατά συνέπεια το βάρος του όρου της κανονικοποίησης ως προς το ολικό σφάλμα.

Η Lasso Regression είναι ιδιαίτερα χρήσιμη όταν έχουμε μεγάλο αριθμό χαρακτηριστικών, καθώς μέσω της κανονικοποίησης μπορεί να επιλέξει αυτόματα τα πιο σημαντικά χαρακτηριστικά, εξαλείφοντας τα λιγότερο σημαντικά και συνεπώς αποφεύγοντας το overfitting. Με αυτόν τον τρόπο, ενισχύει τη γενίκευση του μοντέλου σε άγνωστα δεδομένα.

Όπως και στην περίπτωση του μοντέλου Linear Regression, παρατηρήθηκε ότι οι προβλέψεις βελτιώνονται όταν χρησιμοποιούνται οι λογαριθμισμένες τιμές του SalePrice. Συνεπώς, για την παρουσίαση και αξιολόγηση των αποτελεσμάτων, θα

χρησιμοποιηθούν οι λογαριθμισμένες τιμές. Επίσης, καθώς η Lasso τείνει να μηδενίζει τα βάρη χαρακτηριστικών που είτε εισάγουν θόρυβο είτε παρέχουν πλεονάζουσα πληροφορία, θα χρησιμοποιηθεί το πλήρες σύνολο χαρακτηριστικών. Η επιλογή των πιο σημαντικών χαρακτηριστικών θα προκύψει μέσα από τη διαδικασία εκπαίδευσης, μέσω του μηδενισμού των αντίστοιχων συντελεστών βαρών από το ίδιο το μοντέλο.

Υπό αυτές τις συνθήκες, το εύρος των τιμών που μελετήθηκε είναι από $1e^{-5}$ μέχρι $1e^{2}$ και τα αποτελέσματα φαίνονται στο παρακάτω σχήμα:

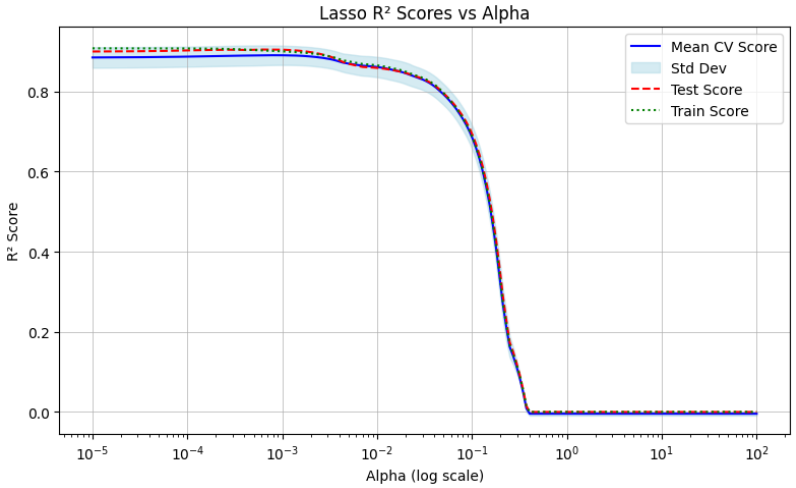

Figure 16: Διάγραμμα της μετρικής R^2 συναρτήσει της τιμής penalty

Το test score φαίνεται να βελτιώνεται μέχρι ένα συγκεκριμένο σημείο (περίπου για τιμή $\lambda \approx 10^{-3}$) πετυχαίνοντας και την καλύτερη ικανότητα γενίκευσης (test score = 0.905, CV score = 0.891 ± 0.024). Από εκεί και πέρα, ο όρος της κανονικοποίησης αρχίζει να κυριαρχεί, μειώνοντας ακόμα και συντελεστές που συμβάλλουν ουσιαστικά στην πρόβλεψη της μεταβλητής-στόχου. Αυτό έχει ως αποτέλεσμα τη μείωση τόσο του train score όσο και του test score. Τελικά, για πολύ μεγάλες τιμές του λ, οι συντελεστές μηδενίζονται και η πρόβλεψη οφείλεται στον όρο του bias του μοντέλου, ο οποίος παίρνει την τιμή του μέσου όρου του SalePrice, οδηγώντας σε $R^2 = 0$.

Στο ακόλουθο σχήμα, φαίνονται οι συντελεστές του καλύτερου μοντέλου που προέκυψε. Η πολυπλοκότητα του μοντέλου φαίνεται να είναι μικρή, καθώς οι περισσότεροι συντελεστές έχουν ήδη μηδενιστεί, χρησιμοποιώντας ένα αρκετά μικρό λ (0.0008).

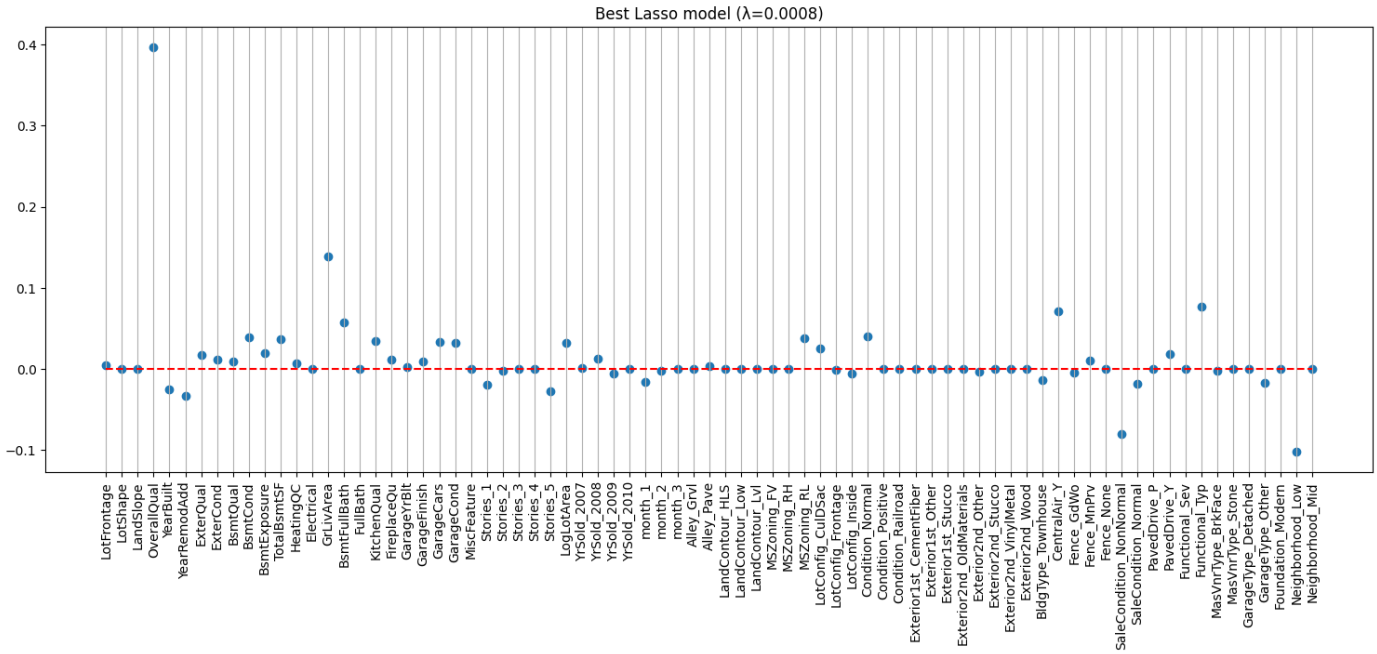

Figure 17: Συντελεστές των χαρακτηριστικών για τη τιμή λ που δίνει το καλύτερο μοντέλο, βάσει του test score

Το χαρακτηριστικό Overall Quality φαίνεται να έχει τη μεγαλύτερη επίδραση στο μοντέλο, καθώς προσδίδει περίπου 0.4 λογαριθμικές μονάδες στο log(SalePrice) όταν μεταβάλλεται από το ελάχιστο στο μέγιστο (δηλαδή από 0 σε 1 στη min-max κλίμακα). Για να μετατρέψουμε τη μεταβολή από λογαριθμική τιμή σε κανονική μονάδα τιμής, εφαρμόζουμε την αντίστροφη του λογαρίθμου δηλαδή $\exp(0.4) \approx 1.5$. Αυτό σημαίνει ότι ένα ακίνητο με Overall Quality στο μέγιστο της κλίμακας (1 μετά την κανονικοποίηση) προβλέπεται να έχει περίπου 49% υψηλότερη τιμή σε σχέση με ένα ακίνητο με Overall Quality στο ελάχιστο της κλίμακας (0), κρατώντας όλα τα υπόλοιπα χαρακτηριστικά σταθερά.

Τα χρονικά χαρακτηριστικά φαίνεται να έχουν αρνητική συσχέτιση με την τιμή του ακινήτου, καθώς όσο πιο παλιά είναι η ημερομηνία κατασκευής ή ανακαίνισης, τόσο χαμηλότερη είναι η προβλεπόμενη τιμή πώλησης. Αυτό υποδηλώνει ότι τα νεότερα σπίτια τείνουν να αποτιμώνται υψηλότερα. Επιπλέον, το χαρακτηριστικό Neighborhood_Low εμφανίζει έντονα αρνητική επίδραση, υποδηλώνοντας ότι η τοποθεσία αυτή σχετίζεται με σημαντικά χαμηλότερες τιμές σε σχέση με άλλες περιοχές.

Στα παρακάτω διαγράμματα αποτυπώνεται η εξέλιξη των συντελεστών καθώς αυξάνεται η τιμή της παραμέτρου κανονικοποίησης. Παρατηρείται ότι με την αύξηση

του penalty, οι συντελεστές συρρικνώνονται σταδιακά και τελικά μηδενίζονται, φανερώνοντας τη διαδικασία επιλογής χαρακτηριστικών που εφαρμόζει η Lasso.

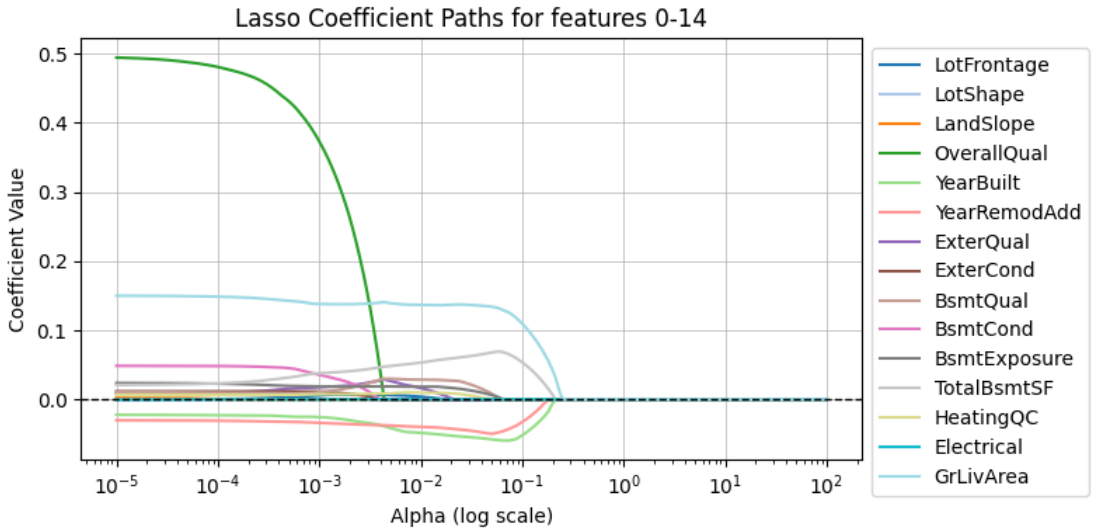

Figure 18: Συντελεστές των χαρακτηριστικών 0-14 συναρτήσει των τιμών του log(penalty)

Στο συγκεκριμένο διάγραμμα παρουσιάζεται και το χαρακτηριστικό που είχε το μεγαλύτερο βάρος στο καλύτερο μοντέλο, δηλαδή το Overall Quality. Παρατηρείται ότι, για μικρές τιμές της παραμέτρου κανονικοποίησης, ο συντελεστής του είναι σημαντικά μεγαλύτερος σε σχέση με τα υπόλοιπα χαρακτηριστικά. Ωστόσο, καθώς η ποινή αυξάνεται, ο συντελεστής του μηδενίζεται νωρίτερα από άλλους. Ο μηδενισμός αυτός φαίνεται να συνοδεύεται από την ενίσχυση των συντελεστών άλλων χαρακτηριστικών, τα οποία πιθανώς μεταφέρουν παρόμοια (αν όχι πλεονάζουσα) πληροφορία. Πολλά από αυτά τα χαρακτηριστικά εμφανίζονται στο επόμενο

διάγραμμα. Τα διαγράμματα των υπόλοιπων χαρακτηριστικών (30-76), για λόγους συντομίας, δεν παρουσιάζονται στην αναφορά.

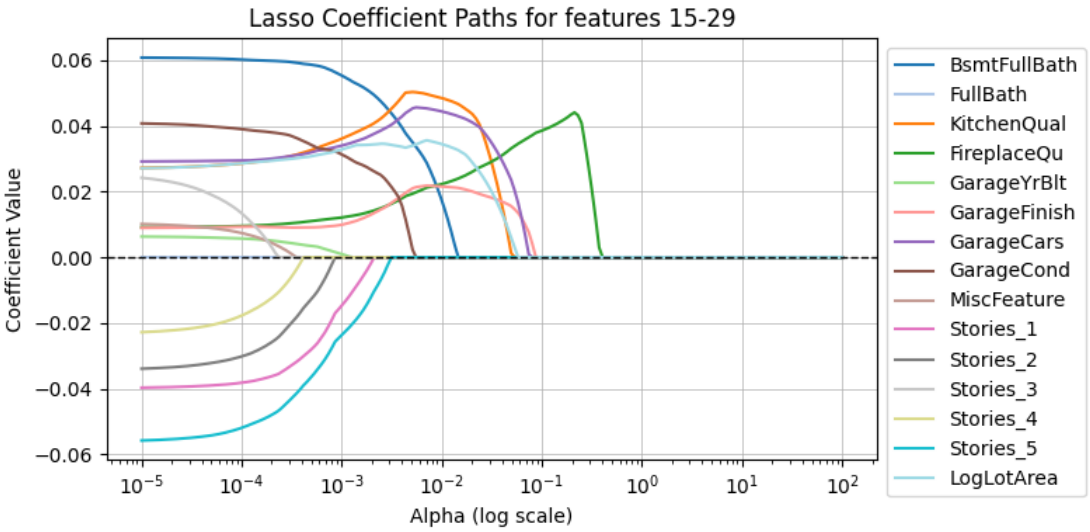

Figure 19: Συντελεστές των χαρακτηριστικών 15-29 συναρτήσει των τιμών του log(penalty)

Πολυεπίπεδα Νευρωνικά Δίκτυα

1 Hidden Layer

Τα πολυεπίπεδα νευρωνικά δίκτυα (multi-layer neural networks) αποτελούν ισχυρά μοντέλα μηχανικής μάθησης, ικανά να αποτυπώνουν πολύπλοκες μη γραμμικές σχέσεις στα δεδομένα. Σε αντίθεση με απλούστερα γραμμικά μοντέλα, τα νευρωνικά δίκτυα διαθέτουν κρυφά επίπεδα (hidden layers), στα οποία γίνεται αυτόματα εξαγωγή και συνδυασμός χαρακτηριστικών (feature engineering). Με αυτόν τον τρόπο, το δίκτυο είναι σε θέση να εντοπίζει πολύπλοκες, μη γραμμικές σχέσεις στα δεδομένα, χωρίς την ανάγκη αυτή η διεργασία να γίνει χειροκίνητα. Έτσι, τα χαρακτηριστικά του επιπέδου n προκύπτουν από μη γραμμικούς μετασχηματισμούς συνδυασμών των χαρακτηριστικών του επιπέδου n−1, επιτρέποντας στο δίκτυο να "χτίσει" σταδιακά πιο αφηρημένες και περιγραφικές αναπαραστάσεις των δεδομένων.

Για την εφαρμογή της μη γραμμικότητας, χρησιμοποιήθηκε η συνάρτηση ενεργοποίησης ReLU (Rectified Linear Unit). Η ReLU εισάγει μη γραμμικότητα με πολύ χαμηλό υπολογιστικό κόστος, καθώς η παράγωγός της είναι απλή: μηδέν για τιμές μικρότερες του μηδενός και ίση με την είσοδο για θετικές τιμές. Στο πλαίσιο αυτό,

εφαρμόστηκαν δύο διαφορετικές αρχιτεκτονικές: μία με ένα και μία με δύο hidden layers, ώστε να μελετηθεί η επίδραση του βάθους στην ακρίβεια και τη δυνατότητα γενίκευσης του μοντέλου.

Αρχικά, το πρώτο μοντέλο που δοκιμάστηκε ήταν αυτό με το ένα hidden layer το οποίο περιλάμβανε 10 νευρώνες, στις μη λογαριθμημένες τιμές του target. Το μοντέλο πέτυχε CV score ίσο με 0.8706 ± 0.0522 ενώ στο test set σημείωσε τελικό $R^2=0.8885$, επιδεικνύοντας και σε αυτήν την περίπτωση καλή ικανότητα γενίκευσης.

Στο παρακάτω διάγραμμα heatmap απεικονίζεται το feature engineering που πραγματοποιήθηκε στο κρυφό επίπεδο μέσω της βελτιστοποίησης των συντελεστών κατά την εκπαίδευση στο training set. Οι ενεργοποιήσεις των νευρώνων στο κρυφό επίπεδο προκύπτουν από μη γραμμικούς συνδυασμούς των αρχικών χαρακτηριστικών, μέσω των εκπαιδευμένων βαρών και της συνάρτησης ενεργοποίησης ReLU. Δεδομένου ότι έχει προηγηθεί κανονικοποίηση των χαρακτηριστικών, με αποτέλεσμα να βρίσκονται σε συγκρίσιμη κλίμακα, οι συντελεστές μπορούν να ερμηνευτούν ως δείκτες σημαντικότητας: όσο μεγαλύτερο η απόλυτη τιμή ενός βάρους, τόσο μεγαλύτερη η επίδραση του αντίστοιχου χαρακτηριστικού στην ενεργοποίηση του συγκεκριμένου νευρώνα.

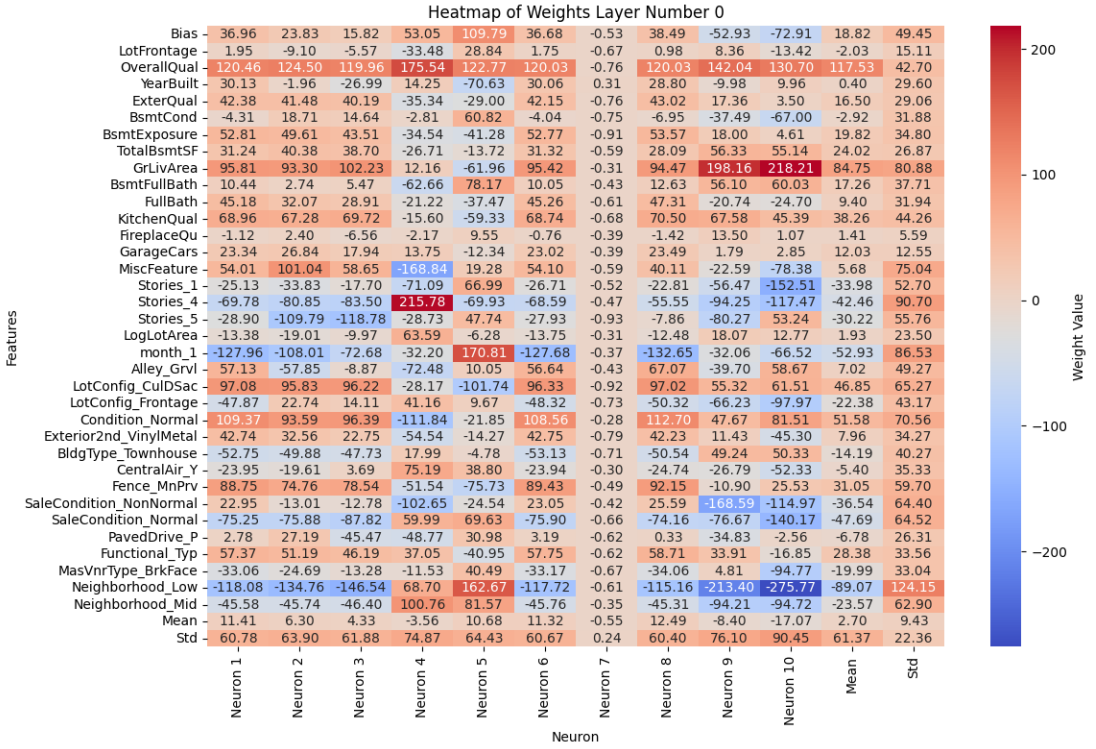

Figure 20: Heatmap με τους συντελεστές κάθε νευρώνα για κάθε χαρακτηριστικό.

Αναφορικά με τα χαρακτηριστικά, προκύπτει μια εικόνα παρόμοια με εκείνη της ανάλυσης των συντελεστών του Lasso Regressor. Το χαρακτηριστικό OverallQual αναδεικνύεται ως το πιο σημαντικό, καθώς συμβάλλει σημαντικά στην εκτίμηση της τιμής ενός ακινήτου. Συγκεκριμένα, εμφανίζει τον μεγαλύτερο μέσο συντελεστή σε όλα τα hidden units, με τιμή περίπου 117.53 ± 42.70. Το χαρακτηριστικό Neighborhood_Low πάλι φαίνεται να είναι αυτό που επηρεάζει περισσότερο αρνητικά την τιμή ενός ακίνητου, καθώς ο μέσος όρος του συντελεστή είναι 89.07. Παρουσιάζει ωστόσο μεγάλη διακύμανση (124.15), καθώς στον νευρώνα νούμερο 5 διαθέτει πολύ μεγάλη και θετική τιμή (162.67). Επίσης το χαρακτηριστικό που εξάχθηκε από το clustering με KMeans, η εποχή, φαίνεται να είναι σημαντική καθώς την Άνοιξη (χαρακτηριστικό month1), έχει πολύ μικρό συντελεστή (-52.93).

Αναφορικά με το feature engineering, ιδιαίτερο ενδιαφέρον παρουσιάζουν οι νευρώνες 4, 5 και 10, καθώς είναι αυτοί με τους μεγαλύτερους συντελεστές (κατά απόλυτη τιμή) προς την έξοδο του μοντέλου: 179.29, 192.11 και 122.90 αντίστοιχα, όπως φαίνεται και στο Σχήμα 21.

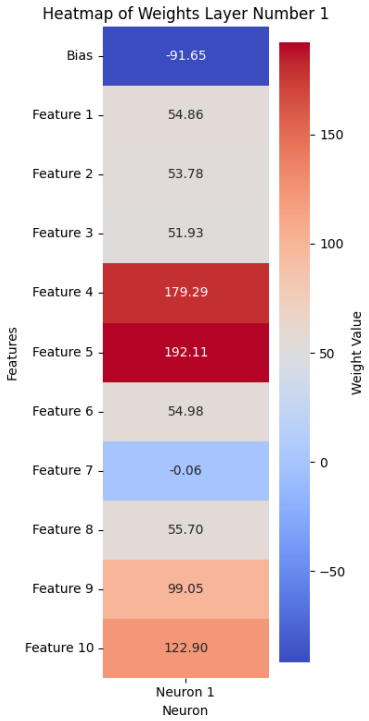

Figure 21: Βάρη και συντελεστής του output layer του μοντέλου

Ωστόσο, αν και ο νευρώνας 10 έχει μικρότερο συντελεστή συγκριτικά, η ενεργοποίησή του από τα χαρακτηριστικά εισόδου είναι σημαντικά υψηλότερη κατά απόλυτη τιμή (-17.07 ± 90.45), γεγονός που θα μπορούσε να ενισχύει τη συμβολή του στην τελική πρόβλεψη. Καθώς όμως χρησιμοποιείται ReLU ως συνάρτησης ενεργοποίησης, κάνει τη συνεισφορά αυτού του νευρώνα μηδενική, σε περιπτώσεις που εξάγεται αρνητικός αριθμός. Επίσης, ο νευρώνας 4, έχει μικρότερα βάρη κατά μέσο όρο από την είσοδο των χαρακτηριστικών (-3.56), συνεπώς οι ολική συνεισφορά προς τη τελική πρόβλεψη αναμένεται μικρότερη. Αξίζει να σημειωθεί ότι, ο νευρώνας 7 έχει αμελητέα επίδραση, καθώς τόσο οι ενεργοποιήσεις του από τα input χαρακτηριστικά όσο και ο συντελεστής εξόδου είναι κοντά στο μηδέν.

Ο νευρώνας νούμερο 5 φαίνεται να παίζει σημαντικό ρόλο στη συνολική πρόβλεψη, καθώς διαθέτει τον μεγαλύτερο συντελεστή στο τελικό output layer. Ωστόσο, η εστίασή του σε χαρακτηριστικά που σχετίζονται με ακίνητα χαμηλότερης αξίας, όπως το Neighborhood_Low και ο μήνας πώλησης (Month_1), είναι αντιδιαισθητική, καθώς αυτά συνδέονται συνήθως με χαμηλές τιμές. Παρόλα αυτά, η συνολική μέση επίδραση των συγκεκριμένων χαρακτηριστικών σε όλους τους νευρώνες, όπως έχει ήδη αναφερθεί, είναι μικρή, γεγονός που υποδεικνύει πως η

επιρροή τους περιορίζεται, ενδεχομένως λόγω χαμηλών ενεργοποιήσεων (ReLU) ή αντιστάθμισης από άλλα χαρακτηριστικά.

Αξίζει να σημειωθεί ότι η παραπάνω ανάλυση βασίζεται στην υπόθεση πως όλα τα χαρακτηριστικά βρίσκονται στην ίδια κλίμακα, πράγμα που ισχύει λόγω της κανονικοποίησης που εφαρμόστηκε. Ωστόσο, γίνεται επίσης μια έμμεση παραδοχή ότι οι κατανομές των χαρακτηριστικών είναι ομοιόμορφες (uniform), γεγονός που θα σήμαινε ότι κάθε χαρακτηριστικό ενεργοποιεί τους νευρώνες με παρόμοια συχνότητα. Για τον λόγο αυτό, ακολουθεί μια πιο εμπειρική ανάλυση, όπου παρουσιάζονται οι έξοδοι κάθε νευρώνα όταν το training set περνά από το δίκτυο.

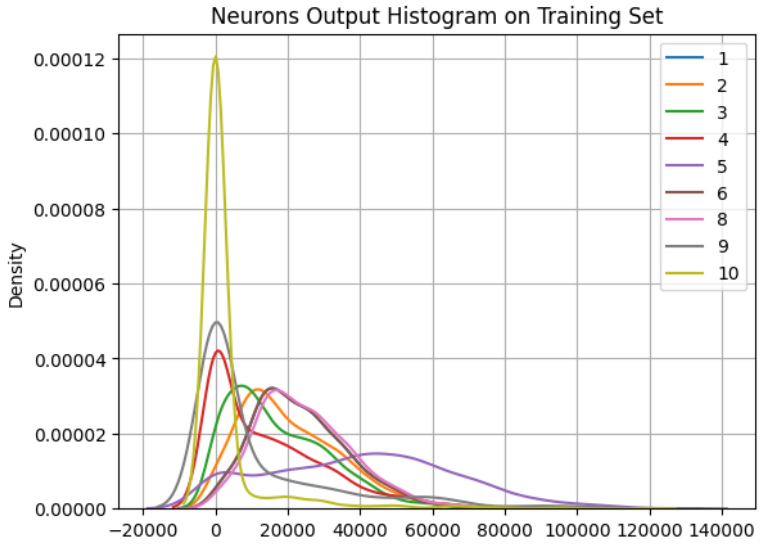

Figure 22: Κατανομές των output κάθε νευρώνα όταν εισάγεται το training set

no Neuron	1	2	3	4	5	6	7	8	9	10
out value (mean ± std)	23621 ± 13207	20666 ± 14019	17375 ± 13865	14299 ± 15933	40094 ± 25615	23584 ± 13201	-92 ± 0	24833 ± 13323	12082 ± 21641	3256 ± 11963

Table 3: Τιμές mean±std για τα ouput κάθε νευρώνα στο training set

Όπως φαίνεται από τα παραπάνω αποτελέσματα, ο νευρώνας με τη μεγαλύτερη επιρροή στην τελική πρόβλεψη είναι πράγματι ο νευρώνας νούμερο 5. Η υψηλή διακύμανση των εξόδων του πιθανόν οφείλεται τόσο στο μεγάλο τελικό βάρος που φέρει (192.11), το οποίο ενισχύει τις επιδράσεις των εισερχόμενων χαρακτηριστικών, όσο και στη σημαντική συμβολή του χαρακτηριστικού OverallQual, το οποίο

παρουσιάζει ευρύ φάσμα τιμών και υψηλό βάρος. Από την άλλη, ο νευρώνας νούμερο 7 παρουσιάζει μηδενική διακύμανση και μέση τιμή ίση με το τελικό bias του, γεγονός που υποδηλώνει πως η έξοδός του μετά τη ReLU είναι συνεχώς μηδενική, συνεπώς, η μοναδική συνεισφορά του είναι μέσω του bias, το οποίο προστίθεται σταθερά στο τελικό άθροισμα. Αναφορικά με την επίδραση του νευρώνα νούμερο 10, επιβεβαιώνεται η αρχική υπόθεση πως η συνεισφορά του στην τελική πρόβλεψη είναι πολύ περιορισμένη.

2 Hidden Layers

Το μοντέλο με δύο hidden layers έχει τη δυνατότητα να εξάγει πιο περίπλοκα και αφηρημένα χαρακτηριστικά, συνδυάζοντας τα ήδη μετασχηματισμένα χαρακτηριστικά του πρώτου layer σε ακόμα πιο σύνθετες μη γραμμικές σχέσεις. Με αυτόν τον τρόπο, μπορεί να «μάθει» πιο βαθιές δομές μέσα στα δεδομένα, προσφέροντας ενδεχομένως καλύτερη προσαρμογή σε πολύπλοκα patterns που δεν μπορούν να αποτυπωθούν με ένα μόνο hidden layer. Για λόγους συντομίας τα αποτελέσματα θα περιγραφούν πιο γενικά.

Η αρχιτεκτονική που χρησιμοποιήθηκε ήταν 2 hidden layers με 10 νευρώνες η καθεμία, και ReLU activation function. Το μοντέλο σημείωσε ελαφρώς καλύτερο test score $R^2 = 0.8897$ και ενώ εμφάνισε οριακά χαμηλότερο CV $R^2 = 0.8706 \pm 0.0522$. Η διαφορά στην απόδοση είναι πρακτικά αμελητέα, ιδιαίτερα αν ληφθεί υπόψη το επιπλέον υπολογιστικό κόστος: το δίκτυο περιλαμβάνει 10×10 περισσότερα βάρη και 10 επιπλέον biases. Αυτό υποδεικνύει ότι, για το συγκεκριμένο πρόβλημα, ένα μόνο hidden layer είναι επαρκές για να αποτυπώσει τα patterns του dataset.

Γενικότερα, παρατηρείται ότι οι τελικοί συντελεστές του μοντέλου με δύο hidden layers είναι μικρότεροι σε μέγεθος. Αυτό είναι αναμενόμενο, καθώς ξεκινώντας από τα ίδια αρχικά χαρακτηριστικά, το επιπλέον hidden layer «απλώνει» την επιρροή κάθε χαρακτηριστικού σε περισσότερα ενδιάμεσα βήματα. Οι ενεργοποιήσεις περνούν πλέον από ένα επιπλέον στάδιο μετασχηματισμού, με αποτέλεσμα οι τελικές τιμές να καταλήγουν στα ίδια επιθυμητά outputs, αλλά με πιο διασκορπισμένες συνεισφορές από κάθε feature.

Και στις δύο περιπτώσεις παρατηρείται το training R^2 να είναι ελαφρώς καλύτερο από το test set, υποδεικνύοντας έναν μικρό βαθμό overfitting (1 hidden layer: 0.027, 2 hidden layers: 0.034).

Gaussian Processes

Οι Gaussian Processes (GP) αποτελούν μια ισχυρή μη παραμετρική προσέγγιση παλινδρόμησης, η οποία αντί να μαθαίνει άμεσα τα βάρη w ενός μοντέλου, εκφράζει τη λύση ως γραμμικό συνδυασμό των παρατηρήσεων μέσω ενός διανύσματος α, δηλαδή $f(x) = \sum a_i k(x, x_i)$. Με αυτόν τον τρόπο, αποφεύγεται η άμεση προβολή σε υψηλές διαστάσεις (όπως έγινε στην περίπτωση του polynomial regression), και το πρόβλημα μετατρέπεται σε έναν υπολογισμό μεταξύ δειγμάτων μέσω του kernel trick, επιτυγχάνοντας έτσι έμμεση μείωση της διαστασιμότητας και καλύτερο χειρισμό της πολυπλοκότητας.

Ένα ακόμη σημαντικό πλεονέκτημα των Gaussian Processes είναι ότι, εκτός από τις προβλέψεις, παρέχουν και τον πίνακα συνδιακύμανσης (covariance matrix) των προβλέψεων. Λαμβάνοντας τα διαγώνια στοιχεία αυτού του πίνακα, δηλαδή τη συνδιακύμανση κάθε πρόβλεψης με τον εαυτό της, προκύπτει η διακύμανση της κάθε πρόβλεψης, από την οποία μπορεί εύκολα να εξαχθεί η αντίστοιχη τυπική απόκλιση ως μέτρο αβεβαιότητας.

Η συνάρτηση $k(x, x_i)$ που αναφέρθηκε προηγουμένως, είναι η συνάρτηση πυρήνα (kernel function) που χρησιμοποιείται για τον υπολογισμό της ομοιότητας μεταξύ των σημείων x και x_i. Στην περίπτωση αυτή, χρησιμοποιούνται οι γραμμικός πυρήνας (linear) και ο πυρήνας Gaussian (RBF).

Αναμενόμενα, ο γραμμικός πυρήνας παρήγαγε αποτελέσματα ισοδύναμα με το μοντέλο Linear Regression, καθώς και οι δύο μέθοδοι βασίζονται σε γραμμικές σχέσεις μεταξύ των μεταβλητών. Συγκεκριμένα, όταν χρησιμοποιήθηκε η λογαριθμημένη τιμή του SalePrice, το test R^2 ανήλθε σε 0.8966 και το cross-validation R^2 σε 0.8889 ± 0.0260, ενώ χωρίς τον λογαριθμικό μετασχηματισμό, οι αντίστοιχες τιμές ήταν 0.8707 για το test R^2 και 0.8363 ± 0.043 για το cross-validation R^2.

Στο παρακάτω διάγραμμα απεικονίζεται η πρόβλεψη του μοντέλου (μετασχηματισμένη τιμή SalePrice) σε συνάρτηση με το GrLivArea για 50 δεδομένα από το σύνολο των test δεδομένων. Το confidence interval (±2 τυπικές αποκλίσεις) έχει πολλαπλασιαστεί με 10^5 προκειμένου να γίνει πιο ευδιάκριτο στο διάγραμμα. Παρατηρείται ότι το εύρος του confidence interval είναι σημαντικά μικρότερο σε περιοχές με υψηλή πυκνότητα σημείων, ενώ σε περιοχές με λιγότερα σημεία είναι μεγαλύτερο.

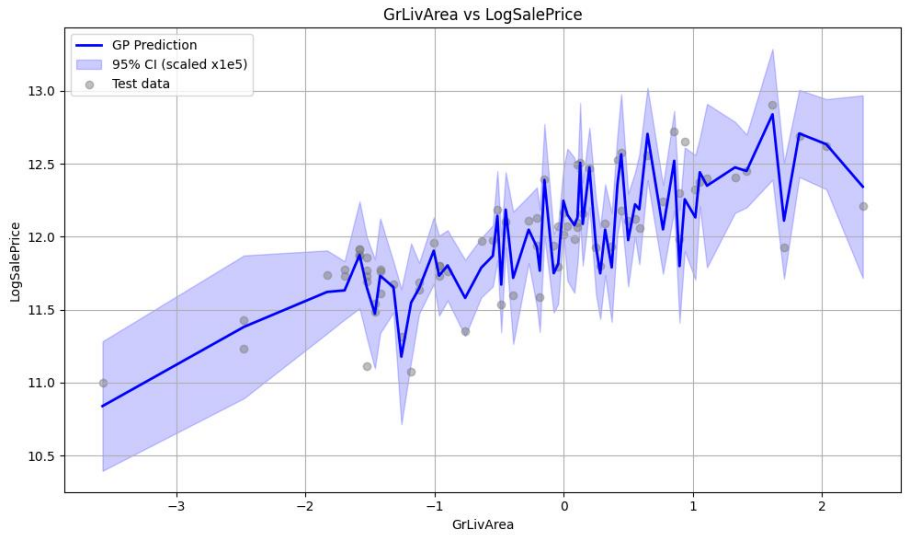

Figure 23: Προβλέψεις του μοντέλου συναρτήση ενός συνεχούς χαρακτηριστικού, του GrLivArea

Αναφορικά με τον RBF πυρήνα, παρατηρήθηκε ότι χωρίς την προσθήκη θορύβου το μοντέλο έκανε overfitting στον θόρυβο των δεδομένων, με αποτέλεσμα να μην μπορεί να γενικεύσει σωστά. Για να αντιμετωπιστεί αυτό, προστέθηκε θόρυβος που ακολουθεί κανονική κατανομή, και όταν ο θόρυβος ήταν 1, η απόδοση του μοντέλου ήταν βέλτιστη όσον αφορά το testing score. Επιπλέον, το training score ήταν παρόμοιο με το testing score, υποδεικνύοντας ότι το μοντέλο δεν έκανε πλέον από overfitting.

Αναφορικά με τη παράμετρο σ, δεν παρατηρήθηκε κάποια ευαισθησία όταν ο θόρυβος είναι επαρκής ώστε το μοντέλο να βγάλει αξιόλογες προβλέψεις. Αντιθέτως, όταν ο τεχνητός θόρυβος είναι πολύ μικρός, παρατηρήθηκε ότι μικρές τιμές του σ (που μεγιστοποιούν την επίδραση των κοντινών γειτονικών σημείων) οδηγούν σε καλύτερα αποτελέσματα. Στον παρακάτω πίνακα φαίνονται τα train και test score για κάθε σίγμα και Noise Level (train R^2| test R^2).

		Noise Level					
		0	1e-5	1e-3	1e-1	1	10
σίγμα	1e-2	1|0.5	1|0.53	1|0.68	0.98|0.85	0.89|0.89	0.76|0.84
	1e-1	1|-5.8	1|0.53	1|0.68	0.98|0.85	0.89|0.89	0.76|0.84
	1	1|-0.7	1|0.53	1|0.68	0.98|0.85	0.89|0.89	0.76|0.84
	1e1	1|-5.9	1|-5.9	1|-5.9	0.98|0.85	0.89|0.89	0.76|0.84
	1e2	1|-5.9	1|-5.9	1|-5.9	0.98|0.85	0.89|0.89	0.76|0.84

Table 4: Μετρικές του μοντέλου για κάθε τιμή σίγμα και θορύβου

Στο παρακάτω διάγραμμα φαίνονται και οι προβλέψεις στο test set σε 50 παραδείγματα, με το confidence interval να είναι πολλαπλασιασμένο με το 2e4, ώστε να είναι διακριτό στο διάγραμμα.

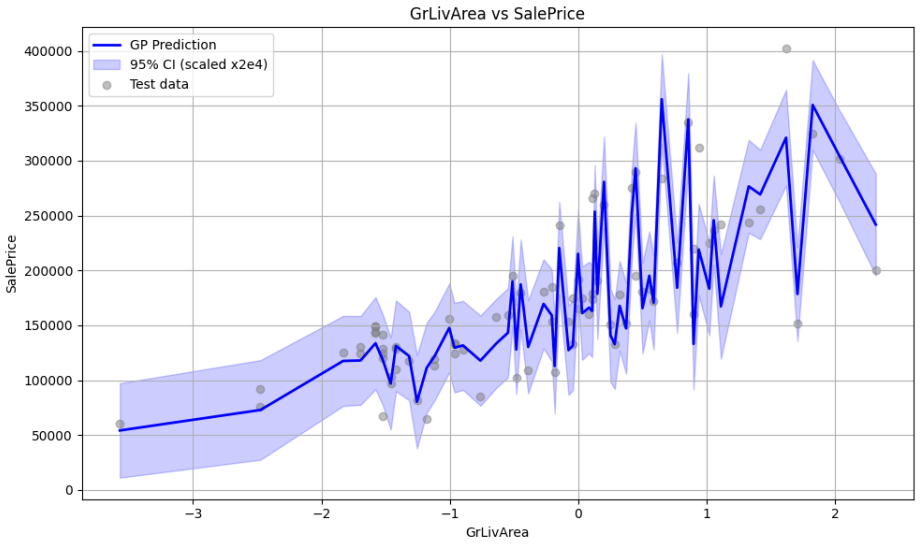

Figure 24: Προβλέψεις του μοντέλου συναρτήση του GrLivArea

Σύγκριση Μοντέλων και Συμπεράσματα

Στην παρούσα εργασία εφαρμόστηκαν και συγκρίθηκαν διαφορετικά μοντέλα παλινδρόμησης για την πρόβλεψη της τιμής πώλησης ακινήτων, ξεκινώντας από απλές γραμμικές μεθόδους και φτάνοντας έως πιο σύνθετες προσεγγίσεις, όπως νευρωνικά δίκτυα και Gaussian Processes.

Το απλούστερο μοντέλο που χρησιμοποιήθηκε αρχικά ήταν το Linear Regression, το οποίο παρουσίασε εξαιρετικά αποτελέσματα, ιδιαίτερα όταν το χαρακτηριστικό - στόχος είχε λογαριθμηθεί. Ο μετασχηματισμός αυτός μετέτρεψε την

αρχικά skewed κατανομή του target (εικόνα 1) σε κατανομή τύπου Gaussian (εικόνα 13), μειώνοντας την επίδραση ακραίων τιμών στα τελικά βάρη και στο συνολικό σφάλμα, δεδομένου ότι χρησιμοποιήθηκε ως συνάρτηση κόστους η MSE.

Το μοντέλο Polynomial Regression παρουσίασε καλή ικανότητα γενίκευσης όταν χρησιμοποιήθηκαν πολυώνυμα χαμηλού βαθμού. Τα χαρακτηριστικά που χρησιμοποιήθηκαν προήλθαν από την προβολή των αρχικών χαρακτηριστικών του Linear Regression σε έναν υποχώρο 5 διαστάσεων, με στόχο τη μείωση της διαστασιμότητας και τη δυνατότητα χρήσης έως και 10 βαθμών. Με την αύξηση του βαθμού του πολυωνύμου, παρατηρήθηκε αναμενόμενα overfitting, με το R^2 του training set να προσεγγίζει τη μονάδα, ενώ το test score και το μέσο cross-validation score να μειώνονται σημαντικά, ενώ το τελευταίο παρουσίαζε αυξανόμενη διακύμανση. Σε σύγκριση με το Linear Regression, δεν παρατηρήθηκε βελτίωση ως προς τη γενίκευση σε άγνωστα δεδομένα.

Αναφορικά με το Lasso Regressor, η βέλτιστη απόδοση επιτεύχθηκε χρησιμοποιώντας το πλήρες σύνολο χαρακτηριστικών, με την τιμή του penalty να οδηγεί σταδιακά σε μηδενισμό των περιττών χαρακτηριστικών. Όπως αναμενόταν, οι καλύτερες προβλέψεις επιτεύχθηκαν όταν το target είχε λογαριθμηθεί, όπως και στην περίπτωση του Linear Regression. Η απόδοση του μοντέλου στο test set ήταν καλύτερη από κάθε άλλο μοντέλο που χρησιμοποιήθηκε.

Τα δύο νευρωνικά δίκτυα που χρησιμοποιήθηκαν παρουσίασαν παρόμοια αποτελέσματα. Το δίκτυο με τα δύο hidden layers είχε συγκρίσιμη απόδοση με εκείνο που διέθετε ένα hidden layer, παρά το γεγονός ότι περιλάμβανε 10x10 περισσότερα βάρη και 10 επιπλέον biases μεταξύ του πρώτου και του δεύτερου κρυφού επιπέδου.

Τέλος, τα μοντέλα Gaussian Processes παρουσίασαν εξίσου καλή ικανότητα γενίκευσης. Με γραμμικό πυρήνα, οι επιδόσεις ήταν αναμενόμενα παρόμοιες με αυτές του Linear Regression, τόσο στη λογαριθμημένη όσο και στη μη μετασχηματισμένη μορφή του target. Αντίθετα, με τον RBF kernel τα καλύτερα αποτελέσματα επιτεύχθηκαν στο μη μετασχηματισμένο target, ωστόσο χωρίς προσθήκη τεχνητού θορύβου το μοντέλο εμφάνισε έντονο overfitting για κάθε τιμή του σίγμα. Με την προσθήκη θορύβου, μειώθηκε η υπερπροσαρμογή, αλλά το καλύτερο μοντέλο σημείωσε τελικά χαμηλότερο R^2 στο testing set σε σύγκριση με το αντίστοιχο με γραμμικό πυρήνα.

Στον παρακάτω πίνακα περιλαμβάνονται τα αποτελέσματα των μοντέλων καθώς και το Target που έδωσε αυτές τις μετρικές. Για το μοντέλο Polynomial Regression φαίνεται το μοντέλο με βαθμό πολυωνύμου 2.

Μοντέλο	Train R^2	CV R^2	Test R^2	Target
Linear Regression	0.897	0.889 ± 0.026	0.897	log(SalePrice)
Polynomial Regression	0.844	0.819 ± 0.135	0.867	SalePrice
Lasso Regression	0.902	0.891 ± 0.024	0.905	log(SalePrice)
NN 1 Hidden Layer	0.916	0.871 ± 0.052	0.889	SalePrice
NN 2 Hidden Layer	0.924	0.861 ± 0.087	0.890	SalePrice
GP Linear Kernel	0.897	0.889 ± 0.026	0.897	log(SalePrice)
GP RBF Kernel	0.892	0.855 ± 0.039	0.889	SalePrice

Table 5: Μετρικές κάθε μοντέλου και μορφή του target που χρησιμοποιήθηκε

Θα πρέπει να σημειωθεί ότι στο training set υπήρχαν μερικά outlier, ενώ το test set ήταν πιο καθαρό, γεγονός που εξηγεί γιατί το test score είναι καλύτερο από το cross-validation score, συγκεκριμένα στην περίπτωση των μη λογαριθμημένων τιμών του target, όπου τα outliers αφέθηκαν ως είχαν. Αυτό παρατηρήθηκε επίσης σε μερικά batches, τα οποία είχαν μεγάλη διαφορά στο R^2 από τα υπόλοιπα, κατά τη διάρκεια του cross-validation (φαίνεται και από τον παραπάνω πίνακα, όπου τα μοντέλα που χρησιμοποιούν SalePrice έχουν μεγαλύτερο std στο R^2 score).

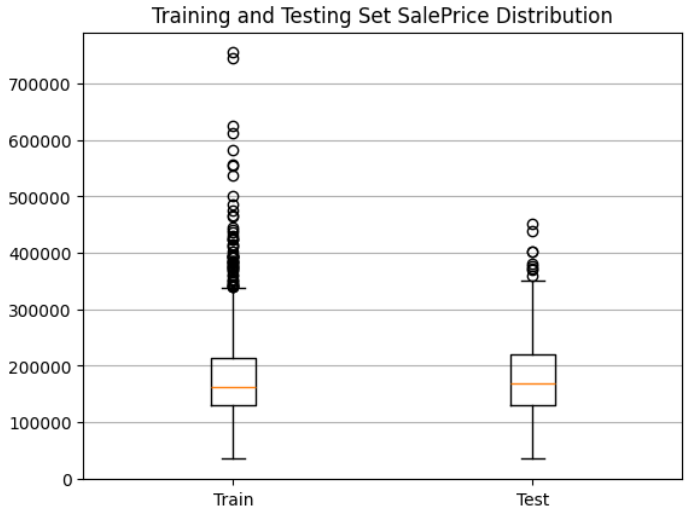

Figure 25: Boxplot του training και του test set

Σημειώσεις για τον κώδικα

Συμπεριλαμβάνονται δύο αρχεία κώδικα, ένα αρχείο preprocess.ipynb, όπου πραγματοποιείται η προεπεξεργασία των δεδομένων, και ένα αρχείο train.ipynb, όπου γίνεται αρχικά ένα βασικό feature selection και στη συνέχεια πραγματοποιούνται τα fit των μοντέλων. Επίσης περιλαμβάνεται και το train.csv από την ιστοσελίδα Kaggle, το οποίο διαβάζεται στο preprocess.ipynb και στο τέλος εξάγεται από το ίδιο αρχείο το αρχείο data_preprocessed.csv. Το τελευταίο διαβάζεται από το αρχείο train.ipynb και πραγματοποιούνται οι διεργασίες που αναφέρθηκαν.