Μέθοδοι Μηχανικής Μάθησης στη Χρηματοοικονομική

Οικονομικό Πανεπιστήμιο Αθηνών

Τμήμα Λογιστικής και Χρηματοοικονομικής

2022-2023

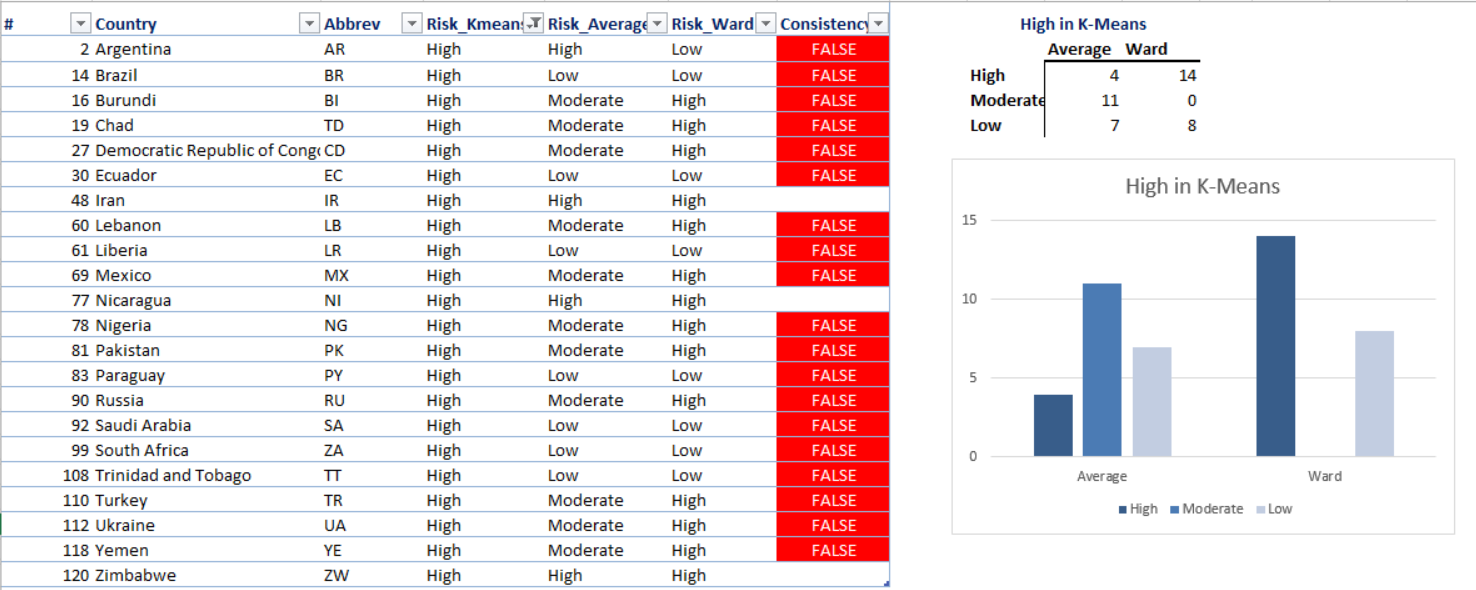
Εργασία 2

|  |  |
| --- | --- |
| Ονοματεπώνυμο Φοιτητή |  |
| Αριθμός Μητρώου |  |
| Ημερομηνία Παράδοσης |  |

# Άσκηση 1

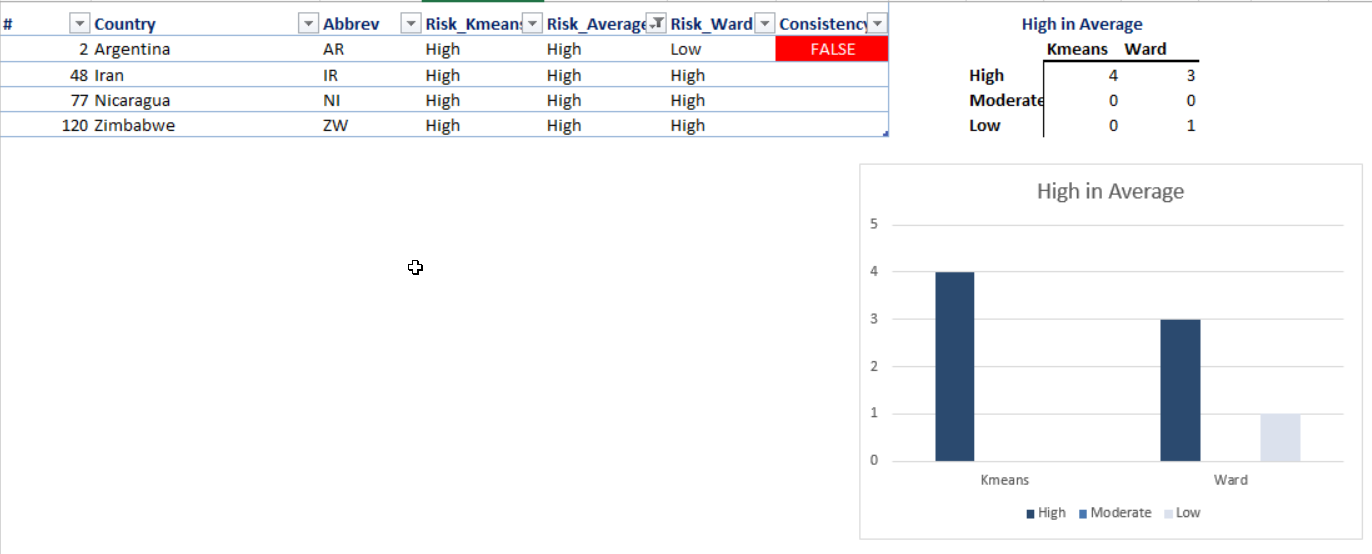
Μιας και τα βασικά βήματα της ανάλυσης παρουσιάστηκαν στην προηγούμενη εργασία, σε αυτήν την άσκηση θα τα παραλείψουμε. Θα αναφέρουμε μόνο ότι χρησιμοποιήθηκαν τα notebooks *4.3 hierarchical\_clustering\_averagemethod.ipynb* και *4.3 hierarchical\_clustering\_wardmethod.ipynb* τα οποία έχουν την ίδια δομή με το *4.1 K-means\_elbowmethod.ipynb* με μόνη διαφορά ότι γίνεται χρήση του αλγορίθμου *AgglomerativeClustering,* Ύστερα από ανάλυση των διαγραμμάτων διασποράς των κλάσεων με τις ετικέτες των χωρών, ανατέθηκαν οι ετικέτες Low, Moderate, High στις αντίσοιχες συστάδες και τελικά εξήχθηκαν τα δεδομένα σε csv αρχεία. Αφού φορτώθηκαν στο *Hierarchical vs KMeans.xlsx* δημιουργήσαμε τρια υπολογιστικά φύλλα στα οποία τα δεδομένα φιλτράρονται με την τιμή High για κάθε μια στήλη-ετικέτα που δημιουργήθηκαν από τους αντίστοιχους αλγορίθμους, π.χ. στο φύλλο *High KMeans* μελετάμε τις χώρες που είχαν *Risk\_Kmeans=High,* ενώ στο *High Ward* τις χώρες που είχαν *Risk\_Ward=High.* Με αυτόν τον τρόπο μπορούμε να μελετήσουμε τη συμπεριφορά των αλγορίθμων όταν ο ένας έχει ταξινομίσει την αντίστοιχη χώρα *High Risk.*

Οι χώρες οι οποίες ταξινομήθηκαν ως *High Risk* από τον KMeans φαίνονται στην εικόνα 1. Παρατηρούμε πως οι αλγόριθμοι δεν «συμφωνούν». Για αυτές τις χώρες, ο Agglomerative με τη μετρική Average ταξινόμησε 4 χώρες ως High, 11 ως Moderate και 7 ως Low. Ενδεικτικά μπορούμε να δούμε πως για τις 7 χώρες που ο Average είπε “Low” μάλλον έχει άδικο, ενώ ο ΚMeans φαίνεται να συμπεριφέρεται πιο σωστά. Από την άλλη ο Ward διακρίνει 14 χώρες ως *High* και 8 ως *Low*. Ομοίως, ο Ward φαίνεται να κάνει λάθος για αυτές τις 8 *Low Risk* χώρες. Παρατητώντας τις συμφωνίες-ασυμφωνίες των αλγορίθων, παρατηρούμε πως όταν ο KMeans ταξινομήσει μια χώρα *High* και ο Average *Low*, τότε και ο Ward θα την ταξινομήσει *Low*, ενώ αν ο Average την ταξινομήσει την χώρα ως *High* ή *Moderate*, ο Ward θα την ταξινομήσει ως *High*, με εξαίρεση την Αργεντινή.



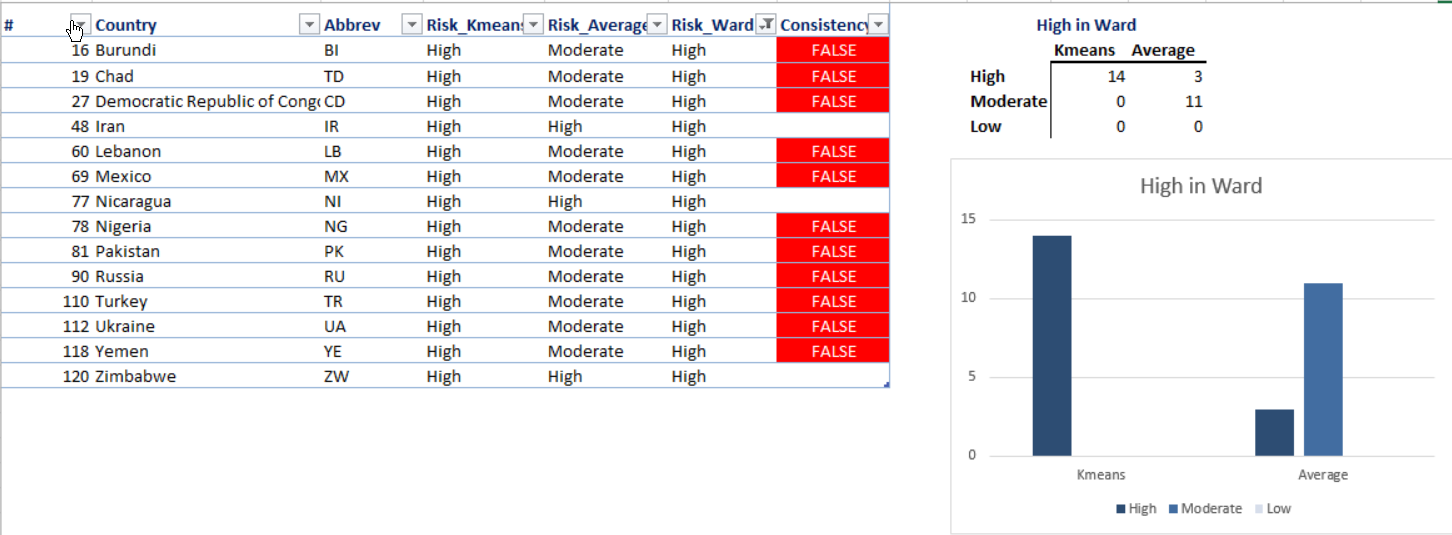
Εικόνα 1 Χώρες ταξινομημένες ως High Risk από τον KMeans και οι κατανομές των ετικετών από τις υπόλοιπες μεθόδους.

Στην περίπτωση όπου *Risk\_Average=High* τα πράγματα είναι πιο αυστηρά για τις ακραίες περιπτώσεις. Παρατηρούμε (Εικ. 2) ότι η μέθοδος Agglomerative Clustering με Average μετρική έχει ταξινομήσει μόνο 4 χώρες ως *High Risk* και σε συνέπεια με τις άλλες μεθόδους, εξαιρουμένης την Αργεντικής. Το μόνο που μπορούμε να σχολιάσουμε είναι πως η μέθοδος αυτή φαίνεται να έχει απομονώσει τις ακραίες χώρες σε μια συστάδα (Εικ 5.).



Εικόνα 2 Χώρες ταξινομημένες ως High Risk από τον AgglomerativeClustering με Average μετρική και οι κατανομές των ετικετών από τις υπόλοιπες μεθόδους.

Στην τελευταία περίπτωση όπου ο φιλτράρουμε βάσει της *High* ετικέτας του Ward, ο Ward συμφωνεί με τον KMeans, ενώ με τον Average μονο στις ακραίες περιπτώσεις.



Εικόνα 3 Χώρες ταξινομημένες ως High Risk από τον AgglomerativeClustering με Ward μετρική και οι κατανομές των ετικετών από τις υπόλοιπες μεθόδους.

Αφού μελετήσαμε τις κατανομές ταξινόμησεις μπορούμε να συμπαιράνουμε από τα σύνολα των ετικετών πως ο KMeans έχει δημιουργήσει ένα υπερσύνολο *High Risk* χωρών σε σχέση με τα High Risk clusters των Average και Ward και ισχύει . Καταλαβαίνουμε πως ο Average ταξινομεί πιο «αυστήρα» μια χώρα ως *High Risk* σε σχέση με τον KMeans. Είναι καλύτερο αυτό; Όχι απαραίτητα. Εξαρτάται από τον σκοπό του αναλυτή αλλά και από την σημασία των υπολοίπων ετικετών. Να σημειωθεί πως εξετάσαμε μόνο το πως συμπεριφέρονται οι αλγόριθμοι για *High Risk* χώρες, και όχι για τις χώρες που ταξινόμησαν ως *Low* ή *Moderate.* Για παράδειγμα, είδαμε πως ο KMeans ταξινόμησε την Βραζιλία ως χώρα υψηλού ρίσκου, ενώ οι άλλοι μέθοδοι την χαρακτήρισαν λανθασμένα ως χαμηλού ρίσκου. Από την άλλη, ο Average μας βρήκε τις ιδιαίτερα υψηλού ρίσκου χώρες και τις ενσωμάτωσε σε μια συστάδα, ενώ τις υπόλοιπες θεωρητικά υψηλού ρίσκου τις ταξινόμησε στη μέτρια βαθμίδα.

Άξιο μελέτες αποτελούν τα διαγράμματα διασποράς των χωρών στις ανά-δυο διαστάσεις. Με κόκκινο χρώμα είναι οι χώρες υψηλού ρίσκου, ενώ με μπλε και πράσινο είναι οι μέτριου και χαμηλού, αντίστοιχα.

|  |  |  |
| --- | --- | --- |
|  | Εικόνα 4 Διαγράμματα διασποράς χωρών κατα KMeans. |  |
|  | Εικόνα 5 Διαγράμματα διασποράς χωρών κατα Ward |  |
|  | Εικόνα 6 Διαγράμματα διασποράς χωρών κατα Average. |  |

Στα παραπάνω διαγράμματα φαίνεται και η «αυστηρότητα» των αλγορίθμων, από τον πιο «χαλαρό» αλλά «ζυγισμένο» KMeans, στον πιο «αυστηρό» αλλά «ανισόρροπο» ως προς το πλήθος των χωρών ανά συστάδα Average. Κάποιος θα μπορούσε να πει πως ο KMeans μοιάζει περισσότερο με τον Agglomerative Clustering με Ward, παρά ο τελευταίος με τον Agglomerative Clustering με Average παρόλο που αποτελούν το ίδιο ουσιαστικά αλγόριθμο. Το γεγονός αυτό οφείλεται στην φύση των μετρικών. Στην περίπτωση του Ward, η μέθοδος χρησιμοποιεί την ελαχιστοποίηση της συνολικής εσωτερικής διακύμανσης των συστάδων ώστε να συγχωνεύσει δύο συστάδες. Με άλλα λόγια, θα πραγματοποιήσει συγχώνευση δύο συστάδων η οποία θα οδηγήσει στην μικρότερη αύξηση της διακύμανσης. Αυτό έχει ως αποτέλεσμα να δημιουργεί συμπαγείς συστάδες πολύ όμοιων, άρα κοντινών δεδομένων και συνήθως ισορροπιμένες ως προς το πλήθος των σημείων. Από την άλλη, ο Average δημιουργεί συστάδες με βάση την μέση απόσταση των σημείων μεταξύ δύο συστάδων. Αυτό καθιστά τη μέθοδο ευάλωτη σε ακραία σημεία και συστάδες αφού η μέση απόσταση τους με άλλες συστάδες θα είναι μεγάλη σε σχέση κάποια άλλη συγνχώνευση, με αποτέλεσμα να τα απομονώνει. Χρήσιμα είναι και τα τρισδιάστατα διαγράμματα διασποράς μιας και δίνουν την αίσθηση του χώρου.

|  |  |
| --- | --- |
| Εικόνα 7 Τρισδιάστατο διάγραμμμα διασποράς κατά Ward | Εικόνα 8 Τρισδιάστατο διάγραμμμα διασποράς κατά Average |

Εύλογα δημιουργείται η απορία γιατί ο Average δημιουργεί μια συμπαγή πράσινη συστάδα, ενώ ο Ward την σπάει σε δυο έχοντας υπόψιν πως ο Ward θα έπρεπε να είναι δημιουργεί συμπαγείς συστάδες. Αυτό οφείλεται στο γεγονός ότι η πράσινη συστάδα (αριστερά) του Average φαίνεται να είναι συμπαγής, όμως αποτελείται από μεγάλο αριθμό σημείων τα οποία τελικα αυξάνουν τη διακύμανση, με άλλα λόγια εξαπλόνονται στο χώρο. Αυτό έχει ως αποτέλεσμα να μην προτιμάει ο Ward μια τέτοια δομή, και να καταληγεί στην δεξία, η οποία τελικά, αριθμητικά και όχι απλά φαινομενικά, έχει μικρότερη διακύμανση.

# Άσκηση 2

Για τον σκοπό της άσκησης δημιουργήθηκε το υπολογιστικό φύλλο *5. Salary vs Age – Regression.xlsx.* Το φύλλο *Data* περιέχει τα σύνολα δεδομένων training, validation, test σε κανονική μορφή. Τα φύλλα *Linear Regression, Ridge Regression* και *Lasso Regression* περιέχουν την πολυωνυμική μορφή και την κανονικοποιημένη πολυωνυμική μορφή των training και validation data, καθώς επίσης και τα αντίστοιχα μοντέλα και τις αποδόσεις τους. Τέλος, στο φύλλο Test Analysis συγκεντρώνονται οι αποδόσεις των μοντέλων, επιλέγετε το καλύτερο και υπολογίζεται η απόδοσή του στο σύνολο δεδομένων test.

Πιο αναλυτικά, στο φύλλο *Linear Regression* ύστερα από προεπεξεργασία των δεδομένων και εκπαίδευση του γραμμικού μοντέλου με το εργαλείο Toolpak Analysis, βρέθηκαν οι παρακάτω συντελεστές και τα MSE για τα training και validation sets αντίστοιχα.

Table 1 Γραμμικοί Συντελεστές και MSEs για Linear Regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ***Coefficients*** |  |  | ***MSE*** |
| **Intercept** | 225555.5556 |  | **Training** | 166465778 |
| **Age** | -30788087.53 |  | **Validation** | 1346930838 |
| **Age^2** | 131623207.6 |  |  |  |
| **Age^3** | -213973307.4 |  |  |  |
| **Age^4** | 156515448.8 |  |  |  |
| **Age^5** | -43327496.62 |  |  |  |

Παρόλο που υπάρχει μεγάλη διαφορά μεταξύ των MSEs, το παρακάτω διάγραμμα δίχνει να «περνά» τα δεδομένα σε καλό βαθμό.

Εικόνα 9 Training και Validation data, και η συνάρτηση εκτίμησης.

Όμοια διαδικασία ακολουθήθηκε και στο *Ridge Regression* φύλλο για το αντίστοιχο μοντέλο, με τη διαφορά ότι δεν χρησιμοποιήθηκε το εργαλείο ToolPak Analysis, αλλά χρειάστηκε να δημιουργηθεί το γραμμικό μοντέλο και η συνάρτηση ελαχιστοποίησης και να αναζητηθούν οι βέλτιστες παράμετροι με το εργαλείο Solver. Η διαδικασία αυτή επαναλήφθηκε τρεις φορές για διαφορετικά λ. Συγκεντρωτικά, έχουμε τους παρακάτω συντελεστές και MSEs.

Table 2 Γραμμικοί Συντελεστές και MSEs για Ridge Regression

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Coefficients** | | |
|  | **λ=0.02** | **λ=0.05** | **λ=0.1** |
| **Intercept** | 225555.52 | 225555.45 | 225555.45 |
| **Age** | 106947.22 | 78284.75 | 57576.85 |
| **Age^2** | 33618.66 | 32193.65 | 26661.84 |
| **Age^3** | -15009.51 | -2990.40 | 1614.51 |
| **Age^4** | -39739.48 | -26769.99 | -16911.93 |
| **Age^5** | -44938.25 | -40587.11 | -29476.26 |
| **Training MSE** | 1253686320 | 1476254273 | 1785304925 |
| **Validation MSE** | 1130849787 | 1598702018 | 2197875499 |

Το παραπάνω πινακάκι δεν φανερώνει πολλά, πέρα από ότι παρατηρούμε το μικρότερο MSE για λ=0.02. Όμως στο δριάγραμμα της εικόνα 10 φαίνεται πως όσο μεγαλώνει το λ, χαλαρώνουν οι συντελεστές το οποίο οδηγεί σε μεγαλύτερη γενίκευση της συνάρτησης εκτίμησης. Επίσης, υπάρχει φθήνουσα βαρύτητα όσο αυξάνεται ο πολυωνυμικός βαθμός. Παρόλο που η κλίμακα των δεδομένων *Age, Age^1, …, Age^5* είναι η ίδια (κανονικοποιημένη), φαίνεται οι μεγαλύτεροι πολυωνυμικοί βαθμοί να λαμβάνουν μικρότερη βαρύτητα.[[1]](#footnote-1)

Εικόνα 10 Συντελεστές γραμμικής συνάρτησης Ridge για διαφορετικά λ.

Οι γραφικές παραστάσεις των εκτιμήσεων διαφέρουν ελάχιστα όπως φαίνεται στην εικόνα 11.

Εικόνα 11 Τraining και Validation data, και η συνάρτηση εκτίμησης κατά Ridge για διάφορα λ.

Η ίδια διαδικασία αλλά με διαφορετική συνάρτηση ελαχιστοποίησης εκτελέσθηκε στο φύλλο *Lasso Regression* για την αντίστοιχη μέθοδο λαμβάνοντας τα παρακάτω αποτελέσματα.

Table 3 Γραμμικοί Συντελεστές και MSEs για Lasso Regression

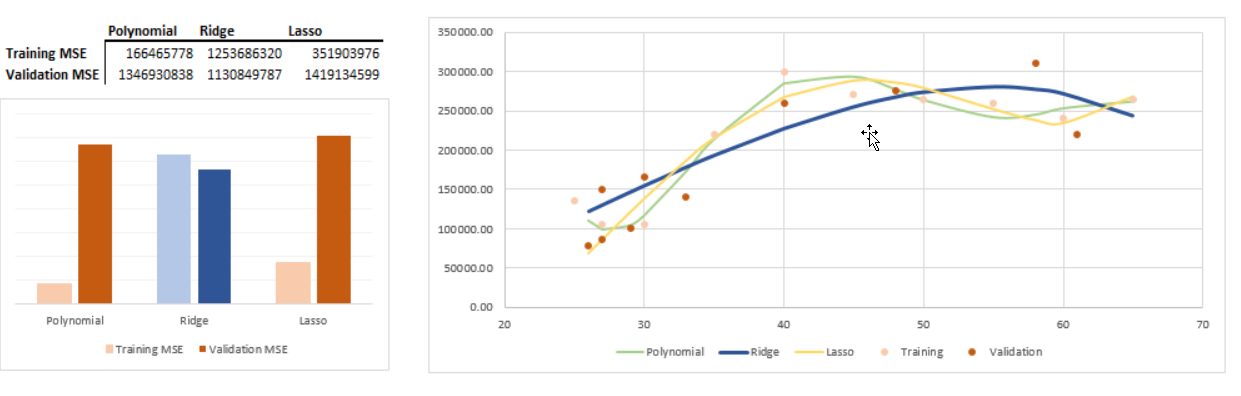
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Coefficients** | | |
|  | **λ=0.02** | **λ=0.05** | **λ=0.1** |
| **Intercept** | 225553.73 | 225555.46 | 225555.46 |
| **Age** | -519103.69 | -480108.04 | -498178.80 |
| **Age^2** | 2171097.60 | 2017093.82 | 2052754.22 |
| **Age^3** | -267414.00 | -34735.05 | -24365.68 |
| **Age^4** | -3748218.03 | -3906599.50 | -3965422.02 |
| **Age^5** | 2423069.86 | 2463791.92 | 2494636.34 |
| **Training MSE** | 351903976 | 352298627 | 352353403 |
| **Validation MSE** | 1419134598 | 1419452847 | 1419465801 |

Το MSE φαίνεται μειώθηκε περίπου κατά 70% σε σχέση με το μοντέλο Ridge στο training set και αυξήθηκε κατά 25% στο validation για λ=0.02, ενώ για τα υπόλοιπα λ έχουμε αντίστοιχη μείωση στο training, αλλά και μείωση στο validation. Επιπλέον, υπάρχει αξιοσημείωτη διαφορά στη βαρύτητα των συντελεστών. Σε αυτήν την περίπτωση, το μοντέλο δεν αποδίδει φθήνουσα βαρύτητα στους συντελεστές, ούτε υπάρχει κάποιο μονοτονία ως των συντελεστών ως προς το λ. Αυτό οφείλεται στη «φύση» του Lasso αφού προσπαθεί να δημιουργήσει ένα αραιό μοντέλο, δηλαδή να μηδενίσει τους συντελεστές των πιο αδύναμων μεταβλητών πράγμα το οποίο είναι καλό στις περιπτώσεις με πολυσυγγραμμικότητα. Γενικά φαίνεται να μην υπάρχουν έντονες διαφορές στους συντελεστές για διαφορετικά λ, με αποτέλεσμα να παράγονται συγκλινόμενες συναρτήσεις εκτίμησης, όπως φαίνεται στην εικόνα 13.

Εικόνα 12 Συντελεστές γραμμικής συνάρτησης Lasso για διαφορετικά λ.

Εικόνα 13 Τraining και Validation data, και η συνάρτηση εκτίμησης κατά Lasso για διάφορα λ.

Για την επιλογή του αποδοτικότερου μοντέλου συγκεντρώθηκαν οι μετρικές στο φύλλο Test Analysis. Για τα μοντέλα Ridge και Lasso, επιλέχθηκαν τα μοντέλα για λ=0.02 αφού είχαν το μικρότερο validation MSE. Από όλα τα μοντέλα, την καλύτερη επίδοση στο Validation MSE την έχει το Ridge μοντέλο, το οποίο όμως έχει παράλληλα την κατά πολύ χειρότερη επίδοση στο Training MSE. Μπορούμε να συμπεράνουμε πως τα μοντέλα Polynomial και Lasso έγιναν overfit στα training δεδομένα.



Εικόνα 14 Επιδόσεις MSE των διαφόρων μοντέλων και οι γραφικές παραστάσεις των εκτιμήσεων.

Από τις γραφικές παραστάσεις των συναρτήσεων εκτίμησης μπορούμε να δούμε πως όντως τα μοντέλα Polynomial και Lasso «ακολουθούν» σε μεγάλο βαθμό τα training δεδομένα σε αντίθεση με το Ridge, το οποίο έχει μια πιο γενικευμένη μορφή, εξώ και το μεγαλύτερο training MSE. Αυτή είναι και μια κύρια διαφορά των μοντέλων, bias vs variance. Το Ridge προσπαθεί να μειώσει το variance δηλαδή τη δυνατότητα του μοντέλου να είναι ευαίσθητο σε μικρές αλλαγές των δεδομένων, πράγμα το οποίο κάνουν τα άλλα μοντέλα με το overfitting στα training data, και παράλληλα αυξάνει το bias δηλαδή το σφάλμα που έχει το μοντέλο όταν είναι πολύ απλό και αδυνατεί να προσεγγίσει τα πραγματικά δεδομένα. Αυτό οδηγεί σε regularization, δηλαδή γενίκευση του μοντέλου.

Επιλέγοντας, λοιπόν, το Ridge ως το αποδοτικότερο μοντέλο, το χρησιμοποιούμε στα test data και έχουμε συνοπικτά τα παρακάτω MSEs.

Table 4 MSEs στα training, validation και test δεδομένα, καθώς και το test RMSE από το Ridge μοντέλο.

|  |  |
| --- | --- |
|  | **Ridge, λ=0.02** |
| **Training MSE** | 1,253,686,320 |
| **Validation MSE** | 1,130,849,787 |
| **Test MSE** | 1,472,434,923 |
| **Test RMSE** | 38,372 |

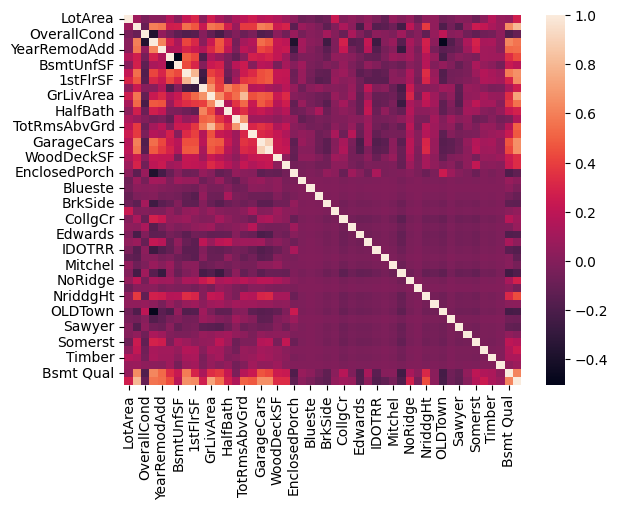
Το test MSE αυξήθηκε αλλά θα θέλαμε να εξετάσουμε πόσο πολύ επηρεάζει τις προβλέψεις μας. Για αυτό υπολογίσαμε το RMSE –πλέον μέσο σφάλμα- το οποίο είναι μια καλή εκτίμηση του μέσου σφάλματος στην κλίμακα των δεδομένων, και όχι του τετραγώνου των δεδομένων. Το μέσο σφάλμα είναι περίπου 38 χιλιάδες ευρώ ετήσιο μισθό. Τα δεδομένα μας, training, validation και test, βρίσκονται μεταξύ 78 και 314 χιλιάδες, δηλαδή έχουν εύρος 236 χιλιάδες. Σε γενικές γραμμές, σε μια τέτοια κλίμακα δεδομένων, ένα σφλάμα 38 χιλιάδων θα λέγαμε πως είναι «χαμηλό» προς «μέτριο». Σίγουρα καλύτερο από το να χρησιμοποιούσαμε τη μέση τιμή 216 χιλιάδες, ως πρόβλεψη, με τυπική απόκλιση 78 χιλιάδες. Βέβαια, η ελαστικότητα αυτή είναι στο χέρι του αναλυτή και το είδος της ανάλυσης.

Εικόνα 15 Γραφική παράσταση της συνάρτησης εκτίμησης του μοντέλου Ridge και όλα τα σύνολα δεδομένων.

# Άσκηση 3

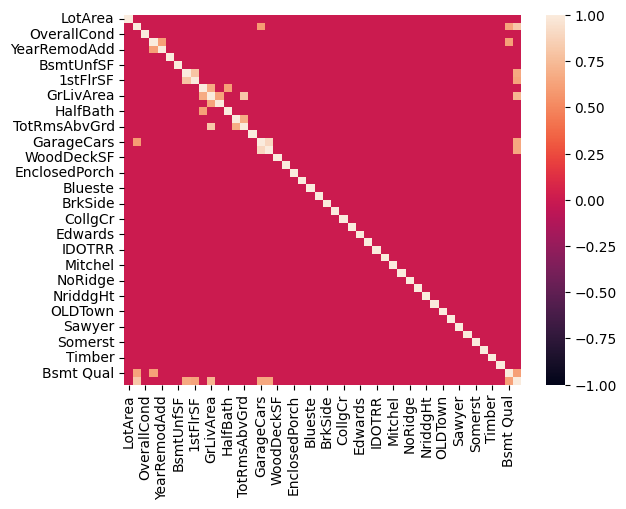
Η άσκηση αυτή «θυμίζει» την προηγούμενη, όμως σε αυτή την περίπτωση έχουμε περισσότερα δεδομένα και περισσότερα αλληλοσυσχετιζόμενα χαρακτηριστικά. Στο notebook *Regression Comparison.ipynb* έχει υλοποιηθεί η ανάλυση των δεδομένων και η εκπαίδευση και αξιολόγηση των μοντέλων.

Αφού ανακατεύθηκαν τα δεδομένα, υπολογίσθηκε ο βαθμός γραμμικής συσχέτισης μεταξύ των χαρακτηριστικών. Στο heatmap της εικόνας 16 εντοπίζουμε κάποια μοτίβα, όμως λόγω του πλήθους των χαρακτηριστικών και τη «κομμένης» κλίμακας δεν μπορούμε να λάβουμε συμπεράσματα.



Εικόνα 16 Feature correlation heatmap.

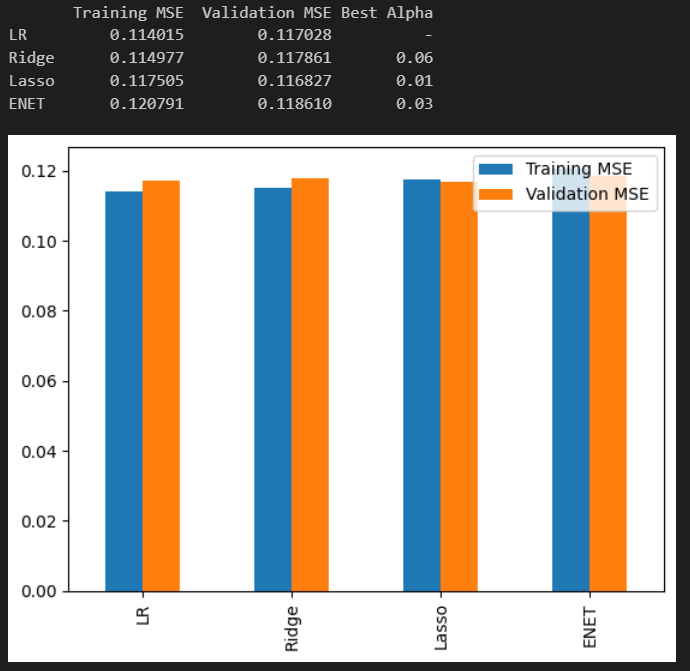
Επειδή μας ενδιαφέρει να εξετάσουμε τα χαρακτηριστικά για πολυσυγγραμικότητα, θα θέλαμε να δούμε ποια ζευγάρια έχουν βαθμό γραμμικής συσχέτισης μεγαλύτερο στο σύνολο . Έτσι, μηδενίζοντας τα υπόλοιπα και θέτοντας σταθερό εύρος χρώματος λαμβάνουμε το heatmap της εικόνας 17.



Εικόνα 17 Feature correlation heatmap for

Παρατηρώντας το δεν θα λέγαμε πως υπάρχει έντονη πολυσυγγραμμικότητα η οποία θα επηρέαζε τα μοντέλα παλινδρόμησης. Βέβαια, το μεγάλο πλήθος χαρακτηριστικών ίσως αποτελεί πρόβλημα.

Στη συνέχεια, χωρίσαμε τα δεδομένα σε σύνολα ανεξάρτητων μεταβλητών και εξαρτημένης μεταβλητής (*Sale* Price) και έπειτα σε σύνολα training, validation και test. Για την εκπαίδευση των μοντέλων και την παραγωγή των μετρικών δημιουργήθηκαν τρεις συναρτήσεις. Η *train\_and\_validate* δέχεται ένα μοντέλο, και αφού το εκπαιδεύσει, υπολογίζει τα Training MSE και Validation MSE και τα επιστρέφει μαζί με τους αντίστοιχους συντελεστές. Η *CV* χρησιμοποιεί cross validation για να βρει την παράμετρο alpha –αντίστοιχο λ– σε ένα μοντέλο και στη συνέχεια χρησιμοποιεί την *train\_and\_validate* με είσοδο το «καλύτερο» μοντέλο. Τέλος, η *compare\_models* χρησιμοποιεί τις παραπάνω συναρτήσεις για να εκπαιδεύσει και να παράξει σύνολα μετρικών αξιολόγησεις για τα μοντέλα *Linear Regression, Ridge, Lasso* και *Elastic Net.* Κατ’ αυτόν τον τρόπο, λαμβάνουμε τα παρακάτω αποτελέσματα.

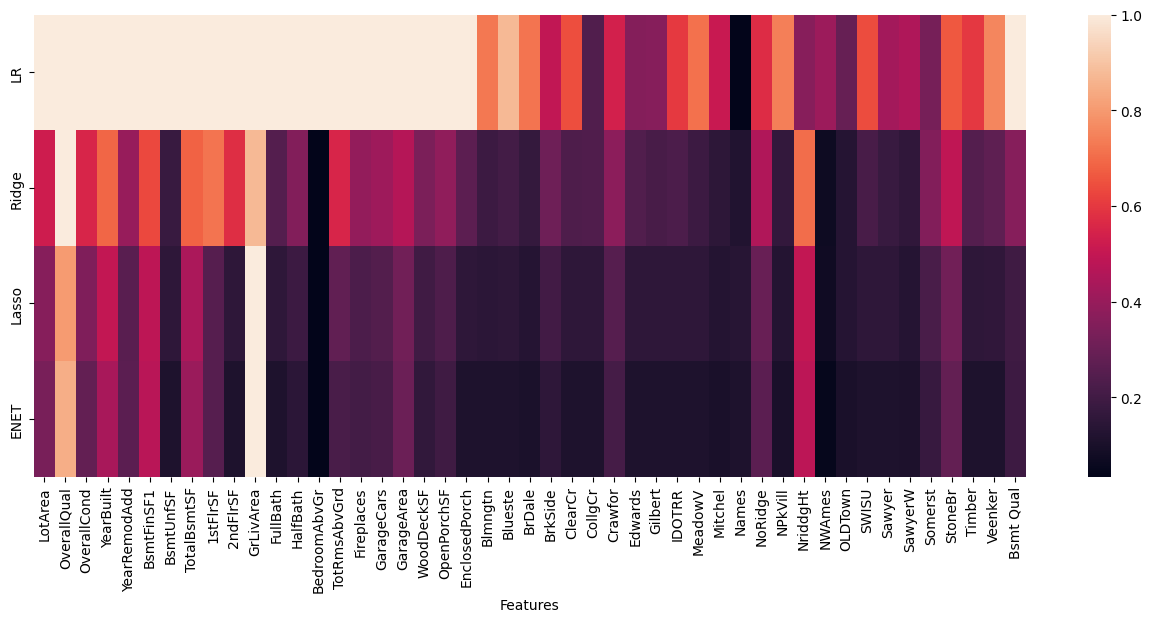


Εικόνα 18 Αποτελέσματα εκπαίδευσης και αξιολόγησης των μοντέλων.

Κάποιος ίσως βιαστεί και συμπαιράνει πως όλα τα μοντέλα είναι καλά αφού το MSE είναι κοντά στο μηδέν και όλα τα μοντέλα έχουν διαφορές στο τρίτο δεκαδικό. Όμως θα πρέπει να λάβει υπόψιν την κλίμακα, την κατανομή της εξαρτημένης μεταβλητής και την τετραγωνική κλίμακα του MSE. Για να λάβουμε παραπάνω πληροφορία ως προς τη συμπεριφορά των μοντέλων αξίζει να δούμε τους συντελεστές τους με τη βοήθεια ενός heatmap (Εικ. 19). Σημειώστε πως οι συντελεστές έχουν κανονικοποιηθεί στην κλίμακα και καλύτερη οπτικοποίηση και σύγκριση.

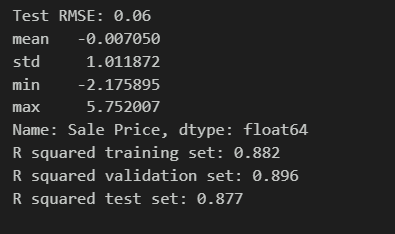
Φαίνεται πως το πολυμετάβλητο γραμμικό μοντέλο (LR) έδωσε μεγάλη έμφαση στα πρώτα 22 και το τελευταίο χαρακτηριστικό, ενώ μικρότερη αλλά εξίσου σημαντική στα υπόλοιπα, πλην του *Names.* Το μοτίβο αυτό εμφανίζεται και στο heatmap γραμμικής συσχέτισης και μπορούμε να πούμε ότι είναι λογικό. Από τη στιγμή που τα χαρακτηστικά αυτά εμφανίζουν γραμμική συσχέτιση μεταξύ τους αλλά και μεταξύ της εξαρτημένης μεταβλητής (τελευταία σειρά/στήλη στις εικόνες 16, 17) είναι φυσικό επόμενο το γραμμικό μοντέλο να εντοπίσει και να αξιοποιήσει αυτές τις σχέσεις. Βέβαια, όπως γνωρίζουμε η πολυσυγγραμμικότητα δεν είναι πάντα καλή και τα υπόλοιπα μοντέλα υλοποιούν τρόπους για να την αντιμετοπίσουν.

Έτσι, παρατηρώντας τα υπόλοιπα μοντέλα (*Ridge, Lasso, ENET*) βλέπουμε πως προσπαθούν να επιλέξουν μόνο τα πιο σημαντικά χαρακτηριστικά και να μηδενίσουν τα υπόλοιπα. Πιο συγκεκριμένα, το Ridge προσπαθεί να γενικεύσει το γραμμικό μοντέλο «τιμορώντας» τις μεγάλες τιμές των συντελεστών, το Lasso προσπαθεί να γενικεύσει «πιέζοντας» τις τιμές συντελεστών των χαρακτηστικών που δεν δίνουν μεγάλη αξία στην πρόβλεψη είτε λόγω της φύσης τους είτε λόγω πολυσυγγραμικότητας. Το ENET συνδυάζει τις μεθόδους των Rigde και Lasso καταλήγοντας ορισμένες φορές σε ένα καλύτερο μοντέλο. Να σημειωθεί πως «καλύτερο» δεν είναι πάντα ένα μοντέλο με το μικρότερο MSE, αλλά ίσως κάποιο το οποίο επιλέγει τα λιγότερα χαρακτηριστικά ή είναι πιο ερμηνεύσιμο.



Εικόνα 19 Heatmap συντελεστών των μοντέλων.

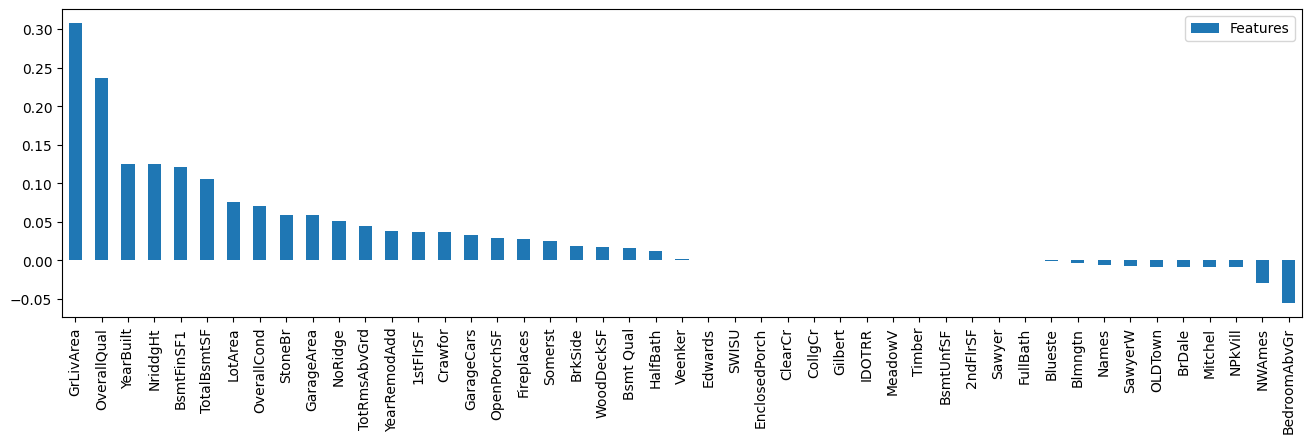
Στην περίπτωσή μας, ως το «καλύτερο» μοντέλο θα επιλέξουμε το Lasso, αφού ναι μεν έχει το μικρότερο MSE, αλλά έχει δε μειώσει, ή καλύτερα εξασθενίσει ένα μεγάλο πλήθος χαρακτηστικών. Έτσι, το Lasso παράγει ένα Test MSE ίσο με 0,119. Πάλι πριν βιαστούμε να βγάλουμε συμπεράσματα για την απόδοση, αξίζει να δούμε το RMSE και να το συγκρίνουμε τα περιγραφικά στατιστικά της εξαρτημένης μεταβλητής.



Εικόνα 20 Περιγραφικά στατιστικά της 'Sale Price', Test RMSE και R Squared για το Lasso.

Το Test RMSE=0.06 είναι ικανοποιητικό σε σχέση με τα στατιστικά της *Sale Price,* σε ένα εύρος δεδομένων περίπου 8 και τυπική απόκλιση 1, το σφάλμα 0.06 είναι συγκριτικά μικρό. Η απόδοση του μοντέλου είναι ικανοποιητική και επιβεβαιώνεται και από την υψηλή τιμή R Square στο 0,87.

Τέλος, αξίζει να μελετήσουμε τις τιμές των συντελεστών. Το παρακάτω γράφημα μας δίχνει πως το μοντέλο θεωρεί ιδιαίτερα σημαντικά 6 χαρακτηστικά με τιμή συνελεστή μεγαλύτερη του 10%. Αναλυτικά είναι GrLivArea, OverallQual, YearBuilt. NriddgHt, BsmtFinSF1, TotalBsmtSF. Αυτή είναι και η μεγαλύτερη αξία αυτών των μοντέλων, σε αντίθεση για παράδειγμα με μοντέλα Νευρωνικών Δικτύων, δηλαδή η ευκολία να εξηγήσει κανείς ποια χαρακτηριστικά έχουν σημαντική προβλεπτική ικανότητα. Ένας domain expert μπορεί να αξιοποιήσει αυτή την πληροφορία για περαιτέρω ανάλυση.



Εικόνα 21 Τιμές συντελεστών του μοντέλου Lasso.

1. Υπάρχει υποψία πως το φαινόμενο αυτό σχετίζεται με την συσχέτιση της κατανομής των πολυωνυμικών ανεξάρτητων μεταβλητών με την κατανομή της εξαρτημένης μεταβλητής. Με απλά λόγια, αν Age­1=40, Salary1=300,000 και Age­2=43, Salary2=380,000 έχουμε μεταβολή (Salary2-Salary1)/(Age­2-Age­1­)=26,666, όμως (Salary2-Salary1)/(Age­25-Age­15­)=80,000/44,608,443=0,001. [↑](#footnote-ref-1)