Μέθοδοι Μηχανικής Μάθησης στη Χρηματοοικονομική

Οικονομικό Πανεπιστήμιο Αθηνών

Τμήμα Λογιστικής και Χρηματοοικονομικής

2022-2023

Εργασία 3

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

# Άσκηση 1

Για το σκοπό της άσκησης, χωρίσαμε τα δεδομένα σε training και test sets, από 1929/09 έως 2015/12 και από 2016/01 έως 2022/12 αντίστοιχα, και στη συνέχεια τα κανονικοποιήσαμε στα υπολογιστικά φύλλα *Train Data* και *Test Data*.

Στο φύλλο *Ridge Train* δημιουργήθηκαν οι κατάλληλες μεταβλητές για την κατασκευή ενός γραμμικού μοντέλου Ridge, το οποίο εκπαιδεύτηκε για τρια διαφορετικά λ, 0.01, 0.05 και 0.1. Οι συντελεστές, τα MSEs και οι βέλτιστες τιμές φαίνονται στον πίνακα 1. Φαίνεται πως το μοντέλο Ridge απέδωσε καλύτερα για λ=0.05, ενώ οι διαφορές μεταξύ των μοντέλων είναι αριθμιτικά μικρές. Λέμε αριθμιτικά διότι η κλίμακα των δεδομένων είναι κανονικοποιημένη, οπότε μια διαφορά τάξης χιλιοστού ενδέχεται να είναι τεράστια στην κανονική κλίμακα των δεδομένων.

Table 1 Συγκεντρωτικός πίνακας τιμών συντελεστών, MSEs και βέλτιστων τιμών για διάφορα λ από το μοντέλο Ridge.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **RESULTS** | | | | | | | | | | | | | |
| **λ** | **b0** | **b1** | **b2** | **b3** | **b4** | **b5** | **b6** | **b7** | **b8** | **b9** | **b10** | **b11** | **MSE** | **Obj** |
| **0.01** | 5.01E-05 | 0 | 0.018728 | 0.093681 | 0 | 4.88E-05 | 0 | 0.000179 | 0 | 0 | 0.002304 | 0 | 0.987114 | 0.987205 |
| **0.05** | 9.1E-05 | 0 | 0.034895 | 0.089082 | 0 | 0.000914 | 0 | 0.065394 | 0 | 0 | 0.051132 | 0 | 0.980214 | 0.981016 |
| **0.10** | 9.16E-05 | 0 | 0.035115 | 0.08482 | 0 | 0 | 0 | 0.062303 | 0 | 0 | 0.048817 | 0 | 0.980289 | 0.981759 |

Εικόνα 1 Ραβδόγραμμα συντελεστών για διάφορα λ για το μοντέλο Ridge.

Αξίζει να σημειωθεί πως στις παραπάνω περιπτώσεις το μοντέλο «επέλεξε» να μηδενίσει τους συντελεστές b1, b4, b6, b8, b9 και b11 άρα να εξαλείψεί τη σημαντικότητα των αντίστοιχων μεταβλητών, δηλαδή των DY, tbl, lty, ntis, infl και svar. Επιπλέον, μπορούμε να πούμε πως οι τιμές των συντελεστών δεν έχουν έχουν τεράστιες διαφορές για τις διάφορες τιμές του λ.

Όμοια, ο πίνακας 2 παρουσιάζει τα αποτελέσματα του γραμμικού μοντέλου με LASSO regularization.

Table 2 Συγκεντρωτικός πίνακας τιμών συντελεστών, MSEs και βέλτιστων τιμών για διάφορα λ από το μοντέλο LASSO.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **RESULTS** | | | | | | | | | | | | | |
| **λ** | **b0** | **b1** | **b2** | **b3** | **b4** | **b5** | **b6** | **b7** | **b8** | **b9** | **b10** | **b11** | **MSE** | **Obj** |
| **0.01** | 0.000134 | 0 | 0.027367 | 0.093784 | 0 | 0 | 0 | 0.062422 | 0 | 0 | 0.04875 | 0 | 0.980234 | 0.982558 |
| **0.05** | 0.000132 | 0 | 0.000629 | 0.088187 | 0 | 0 | 0 | 0.032816 | 0 | 0 | 0.028045 | 0 | 0.982732 | 0.990223 |
| **0.10** | 2.43E-05 | 0 | 0 | 0.058169 | 0 | 0 | 0 | 0.002804 | 0 | 0 | 0.002606 | 0 | 0.989169 | 0.995529 |

Εικόνα 2 Ραβδόγραμμα συντελεστών για διάφορα λ για το μοντέλο LASSO.

Στην περίπτωση του LASSO, το μοντέλο «μηδενίζει» τις μεταβλητές DY, tbl, dfr, lty, ntis, infl και svar, το οποίο είναι σχεδόν το ίδιο σύνολο «απορριθφέντων» μεταβλητών του Ridge. Όμως εδώ παρατηρούμε μεγαλύτερες αποκλίσεις στις τιμές των συντελεστών για τις διάφορες τιμές του λ. Παρόλα αυτά, την καλύτερη απόδοση φαίνεται να την έχει το μοντέλο με λ=0.01

Το φύλλο *Test* περιέχει τους συντελεστές των αποδοτικότερων μοντέλων, το test dataset, τις αντίστοιχες προβλέψεις του δείκτη και τις πραγματικές τιμές. Ως πρώτο σχόλιο, αξίζει να σημειωθεί πως οι τιμές των συντελεστών των μοντέλων Rigde και Lasso δεν έχουν αισθητές διαφορές μεταξύ τους, όπως φαίνεται στην εικόνα 3. Βεβαία, όπως είπαμε, λόγω τις κλίμακας ενδέχεται αυτές οι μικρές διαφορές να έχουν μεγάλο αντίκτυπο στην τελική απόδοση.

Εικόνα 3 Ραβδόγραμμα σύγκρισεις συντελεστών των αποδοτικότερων μοντέλων Ridge και Lasso.

Όμως η τελική απόδοση φαίνεται να είναι όμοια και εξίσου κακή. Δεν υπάρχει «αποδοτικό» μοντέλο, αλλά σίγουρα το Ridge φαίνεται να αποδίδει ελάχιστα καλύτερα από το Lasso.

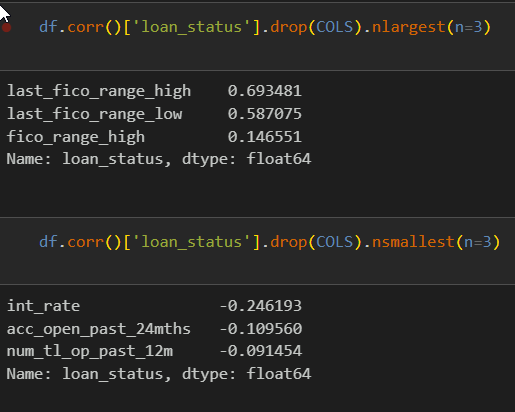
Table 3 Πίνακας μετρικών στα test δεδομένα.

|  |  |  |
| --- | --- | --- |
|  | **MSE** | **R Squared** |
| **Ridge** | **0.763438** | **0.008612** |
| **Lasso** | **0.760732** | **0.009568** |

Μπορούμε να αποδώσουμε την κακή αυτή απόδοση στο γεγονός ότι η μεταβολή του δείκτη S&P500 εξαρτάται από τη μεταβολή των τιμών χρηματηστηρίου πεντακοσίων εταιριών, οι οποίες με τη σειρά του επηρεάζονται από ένα τεράστιο σύνολο κοινών ή και όχι, γνωστών και άγνωστων παραγόντων, για παράδειγμα ο πληθωρισμός, η οικονομική ευημερία, μια απεργία, ένα λανσάρισμα ενός προϊόντος, ένα κακό βίραλ βίντεο που εκθέτει μια εταιρία ή έναν κλάδο. Σίγουρα, οι έντεκα μεταβλητές που δόθηκαν αποτελούν κύριους παράγοντες της μεταβολής του δείκτη, αλλά όχι ικανές ώστε να προβλέψουν τη μεταβολή. Μερικές ακόμα διαστάεις που πρέπει να ενσωματωθούν στο πρόβλημα, άρα και να μοντελοποιηθούν είναι τα βασικά συστατικά των χρονωσειρών, δηλαδή η εποχικότητα, η τάση, η κυκληκότητα και φυσικά η αυτοσυσχέτιση. Τα μοντέλα που φτιάξαμε δεν περιλαμβάνουν αυτά τα συστατικά, παραμόνο ένα στιγμιότυπο της προηγούμενης «κατάστασης». Μοντέλα που έχουν αναπτυχθεί ειδικά για χροσοσειρές ενδέχεται να δώσουν καλύτερα αποτελέσματα του πρόβλημα που μελετάμε.

# Άσκηση 2

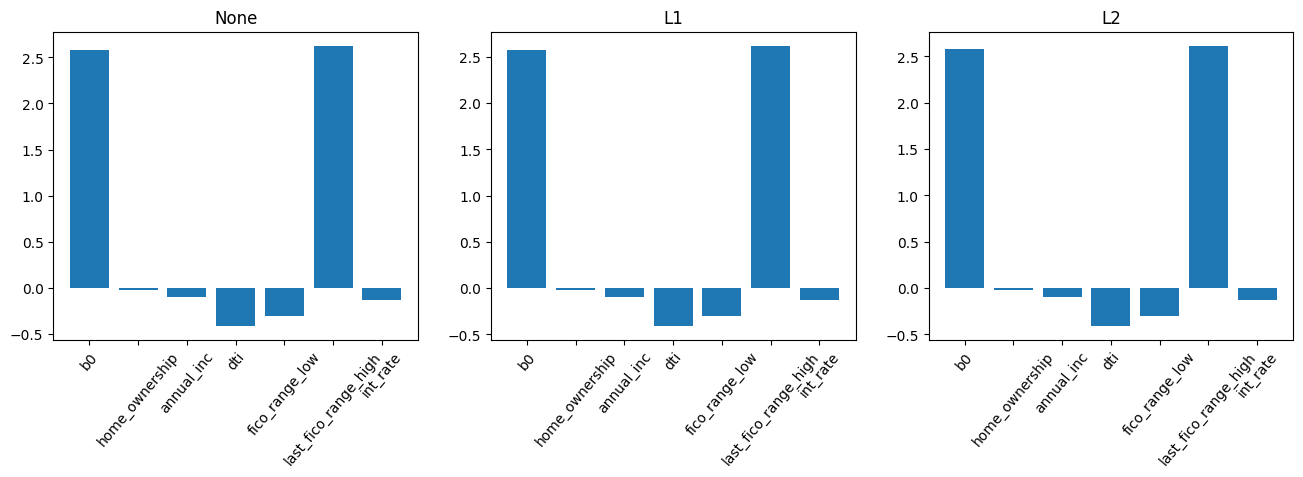
Ξεκινώντας την δημιουργία του μοντέλου πρέπει να επιλέξουμε τις δυο επιπλέον μεταβλητές, από τις 135, που θα προσθέσουμε. Για το σκοπό αυτό υπάρχουν πολλοί μέθοδοι, για παράδειγμα να συμβουλευτούμε έναν ειδικό τομέα, επιλογή χαρακτηριστικών με διαδοχική αφαίρεση/προσθήκη, χρήση PCA στα υπόλοιπα χαρακτηριστικά για την παραγωγή δυο νέων. Εμείς θα ακολουθήσουμε έναν απλό, γραμμικό τρόπο, χρησιμοποιώντας την γραμμική συσχέτιση μεταξύ των χαρακτηριστικών και της κατάστασης του δανείου. Μελετώντας τα τρια μεγαλύτερα και τα τρια μικρότερα αναζητώντας υψηλή θετική και αρνητική συσχέτιση, επιλέξαμε τελικά τα last\_fico\_range\_high και int\_rate. Θα μπορούσαμε να επιλέξουμε τα last\_fico\_range\_low ή fico\_range\_high αντί του int\_rate αφού έχουν πιο ισχυρή συσχέτιση, όμως τα εν λόγω δύο ενδέχεται να έχουν υψηλή συσχέτιση και με την πρώτη επίλογη, last\_fico\_range\_high με αποτέλεσμα να εμφανιστεί πολυσυγγραμικότητα στο μοντέλο μας.



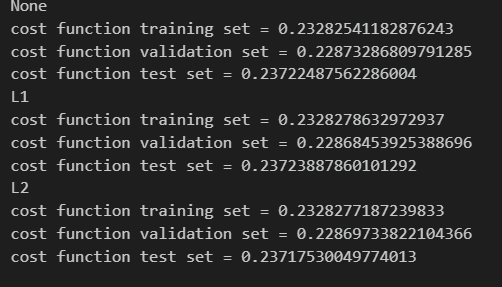
Εικόνα 4 Βαθμός γραμμικής συσχέτισης των υπολοίπων χαρακτηριστικών με τη κατάσταση του δανείου.

Η συνέχεια είναι πολύ απλή. Χωρίσαμε το σύνολο δεδομένων σε training, validation και test και εκπαιδεύσαμε τα μοντέλα. Γιατί μοντέλα και όχι ένα μοντέλο; Στην ερώτηση αν θα χρησιμοποιούσαμε regularization η εκ των προτέρων απάντηση είναι ανάμεικτη. Ναι μεν χρησιμοποιούμε λίγα χαρακτηριστικά στο μοντέλο άρα δεν είναι αρκετά περίπλοκο ώστε να κάνουμε χαλάρωση, αλλά δε, δεν υπάρχει απόλυτη και σωστή απάντηση σε τέτοιες ερωτήσεις στο χώρο της μηχανικής μάθησης, μόνο προεσεγγιστική. Έχοντας τη δύναμη της υπολογιστικής ισχύος και την ευκολία των εργαλείων μπορούμε να εκπαιδεύσουμε μοντέλα με regularization και να εξετάσουμε τη συμπερφιφορά τους.

Ύστερα, λοιπόν, από κανονικοποίηση των δεδομένων, εκπαιδεύσαμε τρια μοντέλα λογιστικής παλινδρόμησης, ένα χωρίς regularization, ένα με L1 και ένα με L2. Τα αποτελέσματα εφμανίζουν μηδαμινές διαφορές, τόσο στους συντελεστές, όσο και στην απόδοση όπως φαίνεται στις εικόνες 2 και 3.



Εικόνα 5 Συντελεστές των τριών μοντέλων.



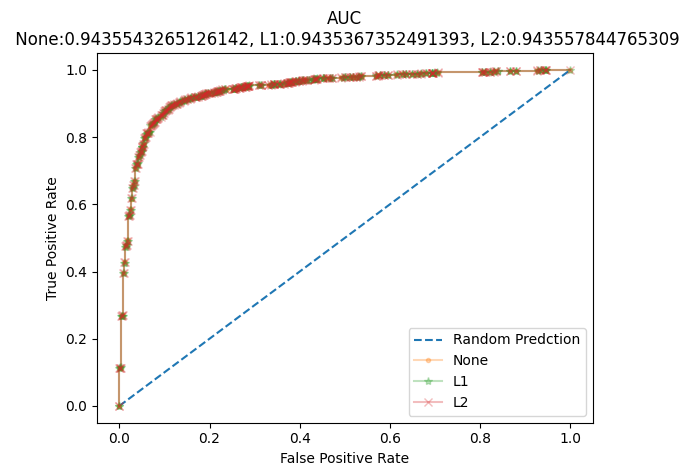
Εικόνα 6 Αποδόσεις των τριών μοντέλων.

Είναι σημαντικό να μεταβάλλουμε το όριο απόφασης και να μελετήσουμε την απόδοση των μοντέλων και αυτό γιατί ο αντίκτυπος λανθασμένης ταξινόμησης είναι διαφορετική για κάθε κλάση. Για παράδειγμα, θα ήταν πιο σοβαρό να ταξινομήσουμε ένα δάνειο ως ασφαλές ενώ στην πραγματικότητα είναι επισφαλές, από ότι να ταξινομήσουμε ένα δάνειο επισφαλές ενώ είναι ασφαλές. Λέμε δηλαδή πως θέλουμε να μειώσουμε τα False Positives, ή το Type 1 Error στην ταξινόμηση επισφαλών δανείων.

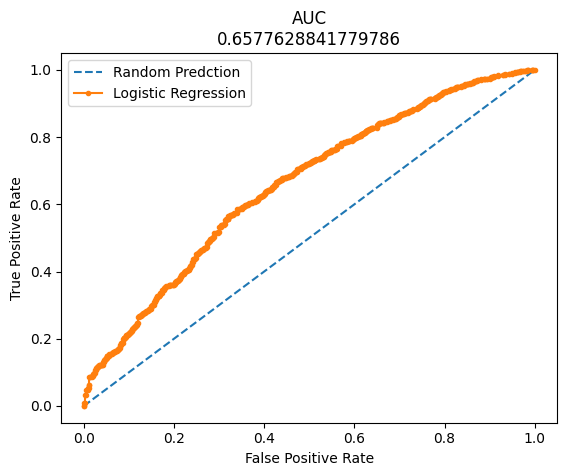
Μπορούμε να εξετάσουμε το Receiver Operating Curve το οποίο μας δείχνει την απόδοση των μοντέλων για διάφορα όρια απόφασης. Πιο συγκεκριμένα δείχνει ζευγάρια τιμών False Positive Rate και True Positive Rate όσο μεταβάλλεται το threshold από το 0 στο 1. Το False Positive Rate απαντάει στην ερώτηση «από το σύνολο των επισφαλών δανείων, πόσα δάνεια ταξινομήθηκαν ως ασφαλή;», ενώ το True Positive Rate απαντάει στην ερώτηση «από το σύνολο των ασφαλών δανείων, πόσα δάνεια ταξινομήθηκαν ως ασφαλή;» Στην εικόνα 4 φαίνεται πως και τα τρια μοντέλα αποδίδουν ισάξια σε κάθε τιμή ορίου απόφασης.

Καταλαβαίνουμε πως η προσθήκη regularization δεν προσέφερε κάποια βελτίωση στην απόδοση, πράγμα το οποίο ενισχύει, αλλα δεν επιβεβαιώνει, την υπόθεση μας πως λόγω του μικρού πλήθους χαρακτηριστικών και του απλού μοντέλου δεν χρειάζεται χαλάρωση.

Σίγουρα όμως βλέπουμε βελτίωση σε σχέση με το αρχικό μοντέλο, το μοντέλο με τα τέσσερα χαρακτηριστικά (Εικόνα 5). Παρατηρούμε πως έχει αυξηθεί δραματικά το AUC, από 0.6577 σε 0.9435, ενώ η καμπλύλη AUC που έτεινε προς την καμπύλη τυχαίας εκτίμησης στο αρχικό μοντέλο, πλέον τείνει στην πάνω αριστερή γωνία.



Εικόνα 7 ROC και AUC των τριων μοντέλων.



Εικόνα 8 Α ROC και AUC του αρχικού μοντέλου βάσης.