Μηχανική και Στατιστική Μάθηση

Εργασία 2

Γεώργιος Παυλής

Τμήμα Στατιστικής
Οικονομικό Πανεπιστήμιο Αθηνών
Νοέμβριος 2023

Εισαγωγή

Μία εταιρία τηλεμάρκετινκ στο εγχείρημα της να προωθήσει ένα προϊόν πραγματοποιεί τηλεφωνικές κλήσεις, με σκοπό την πώληση του, για κάθε κλήση γνωρίζουμε διάφορα στοιχεία του πελάτη και αν τελικά αγόρασε το προϊόν. Τα δεδομένα που διαθέτουμε είναι 39883 παρατηρήσεις με τα εξής χαρακτηριστικά των πελατών :

Στοιχεία Πελάτη

1. age : Η ηλικία σε χρόνια.
2. job : Τύπος εργασίας (Διοικητικός, Χειρωνακτικός, Επιχειρηματίας, Οικιακός βοηθός, Management, Συνταξιοδοτούμενος, Υπηρεσιακός, Μαθητής, Τεχνικός, Άνεργος, Άγνωστο).
3. marital : Συζυγική κατάσταση (Διαζευγμένος, Παντρεμένος, Ανύπαντρος/Χήρος, Άγνωστο).
4. education : Εκπαίδευση (Βασικό Τετραετές, Βασικό Εξαετές, Βασικό Ενναετές, Δευτεροβάθμια, Αναλφάβητος, Επαγγελματίας, Πτυχιούχος, Άγνωστο).
5. default : Αν έχει αθετήσει πληρωμή πίστωσης (Ναι, Όχι, Άγνωστο).
6. housing : Αν έχει στεγαστικό δάνειο (Ναι, Όχι, Άγνωστο).
7. loan : Αν έχει προσωπικό δάνειο (Ναι, Όχι, Άγνωστο).

Σχετικά με προηγούμενες κλήσεις για την προώθηση προϊόντος.

8. contact : Μέσο Επικοινωνιακής Επαφής (Κινητό, Σταθερό).
9. month : Ο μήνας προηγούμενης επαφής
10. day of week : Η μέρα προηγούμενης επαφής.

11. duration : Η διάρκεια της κλήσης της προηγούμενης επαφής.

Άλλα στοιχεία

12. campaign : Ο συνολικός αριθμός επαφών στην διάρκεια της καμπάνιας.
13. pdays : Αριθμός των ημερών που πέρασαν από την τελευταία επαφή.
14. previous : Ο συνολικός αριθμός επαφών πριν αυτή την καμπάνια.
15. poutcome : Αποτέλεσμα προηγούμενης καμπάνιας (Αποτυχία, Ανύπαρκτη, Επιτυχία) Κοινωνικά και Οικονομικά Στοιχεία
16. emp.var.rate : Διακύμανση ενασχόλησης - τετραμηνιαίος δείκτης.
17. cons.price.idx : Δείκτης κατανάλωσης - μηνιαίος.
18. cons.conf.idx : Δείκτης οικονομικής αυτοπεποίθησης - μηνιαίος.
19. euribor3m : euribor τρεις μήνες.
20. nr.employed : Αριθμός υπαλλήλων - τετραμηνιαίος δείκτης.

Απαντητική Μεταβλητή που θέλουμε να μοντελοποιήσουμε.

21. SUBSCRIBED : Αν ο πελάτης αγόρασε το προϊόν (Ναι, Όχι).

Σκοπός μας είναι να αξιοποιήσουμε τα διαθέσιμα δεδομένα και να μοντελοποιήσουμε την μεταβλητή που μας ενδιαφέρει το SUBSCRIBED, ώστε να μάθουμε ποιες μεταβλητές επηρεάζουν σημαντικά το αποτέλεσμα και να μπορούμε δοθείσης νέων δεδομένων να προβλέψουμε αν ο πελάτης θα εγγράφει στο προϊόν. Για να το επιτύχουμε αυτό θα υποθέσουμε ότι υπάρχουν δύο ομάδες οι εγγεγραμμένοι και μη. Θα χρησιμοποιήσουμε διάφορα μοντέλα στην προσπάθεια μας να κατατάσσουμε στην σωστή ομάδα με γνωρίζοντας μονάχα τα χαρακτηριστικά και στην τελική θα κρίνουμε ποιο υπόδειγμα λειτουργεί καλύτερα.

LDA

Το LDA ή στα ελληνικά η Γραμμική Διακριτική Ανάλυση χωρίζει τους πληθυσμούς των δύο ομάδων μου με μία υπερδιάστατη γραμμή στον χώρο των δεδομένων, δηλαδή αναλόγως που βρίσκεται μία παρατήρηση στον χώρο αυτόν θα κατατάσσεται σε μία ομάδα. Το επιτυγχάνει υποθέτοντας ότι για κάθε ομάδα υπάρχει μία πολυμεταβλητή κανονική κατανομή που παίρνει ως όρισμα τα χαρακτηριστικά και επιστρέφει την πιθανότητα να ανήκει σε αυτή την ομάδα.

Συνεπώς για να ακολουθούν τα δεδομένα μου κανονική κατανομή πρέπει να είναι συνεχής και ιδανικά κανονικά κατανεμημένα. Για αυτό επιλέγω να τρέξω το μοντέλο μόνο με τις μεταβλητές age,duration,euribor3m. Από τις οποίες όλες απορρίπτεται η κανονικότητα με το Shapiro-Wilk test for normality. Παρόλαυτα αν το δοκιμάσουμε βγάζει ικανοποιητικά αποτελέσματα.

Κατώφλι ταξινόμησης

Το LDA ανήκει σε μία κατηγορία μοντέλων ταξινόμησης που ονομάζονται soft classification models. Τα οποία μας δίνουν ως αποτέλεσμα πιθανότητες για κάθε παρατήρηση να ανήκει σε μία ομάδα και από εκεί σύμφωνα με κάποιο κατώφλι που θέτουμε εμείς αν έχουμε πιθανότητες που υπερβαίνουν αυτό το κατώφλι κατατάσσουμε εκεί.

Για να βρούμε το βέλτιστο κατώφλι πρέπει πρώτα να θέσουμε ποια στατιστικά μέτρα περιγράφουν αυτό που θέλουμε να επιτύχουμε και να το μεγιστοποιήσουμε. Αυτό που στοχεύουμε είναι να προβλέψουμε αν η επαφή είναι επιτυχημένη δηλαδή δοθείσης τα χαρακτηριστικά ενός εγγεγραμμένου πελάτη να προβλέψω ότι όντως θα αγοράσει το προϊόν δηλαδή να μπορώ να προμηνύω τις επιτυχημένες επαφές. Αυτό το μέτρο ονομάζεται sensitivity. Όμως αν προσπαθώ να μεγιστοποιήσω 'εθελοτυφλώντας' μονάχα αυτό το μέτρο αγνοώντας τα άλλα μέτρα καλής προσαρμογής τότε αυτό που θα συμβεί είναι να κατατάσσω διαρκώς στις επιτυχίες χωρίς να έχω αρκετά σημαντικές ενδείξεις για να το κάνω. Έτσι συχνά θα προβλέπω εσφαλμένα ότι θα εγγράφουν χωρίς να ισχύει στην πραγματικότητα.

Τελικά θα προσπαθήσω να μεγιστοποιήσω το accuracy, το ποσοστό ορθών κατατάξ-

εων και στις δύο ομάδες και ταυτόχρονα να μειώσω το 1-specificity, το ποσοστό των εγγεγραμμένων που εσφαλμένα ταξινόμησα στους μη εγγεγραμμένους. Με το να έχω μεγάλο accuracy και μικρό 1-specificity συμβάλω εν μέρη στην βελτιστοποίηση και του sensitivity. Το εργαλείο που θα χρησιμοποιήσω για να βρω το κατώφλι είναι το ROC curve.

Το ROC είναι ένα διάγραμμα που έχει στον κάθετο άξονα το True Positive, τις σωστές προβλέψεις της ομάδας που έχω ονομάσει ως επιτυχίες και στον κάθετο το False Negative τις λανθασμένες κατατάξεις της ομάδας που έχω ονομάσει ως αποτυχίες. Δοκιμάζωντας μία σειρά από διαφορετικές τιμές για το κατώφλι ταξινόμησης παίρνω και διαφορετικά True Positive και False Negative, στη συνέχεια επιλέγω εκείνη την τιμή που μου δίνει το μεγαλύτερο True Positive και μικρότερο False Negative, ή ισοδύναμα την τιμή στον διάγραμμα που αντιστοιχεί στο σημείο που είναι πιο κοντά στο (0,1).

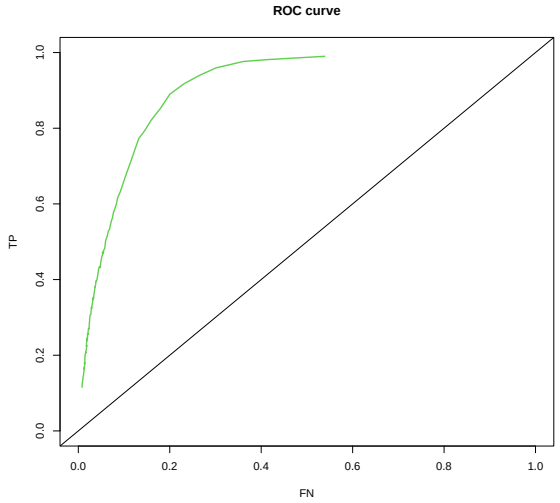

Figure 1: Το ROC curve για το LDA

Επέλεξα για πιθανότητες κατατάξεις στους εγγεγραμμένους, που τους θεώρησα ως την ομάδα επιτυχίας, πάνω του 0.06 να τους εντάσσω εκεί. Δηλαδή αν για μία παρατήρηση έχω πιθανότητα 0.07 να αγοράσει το προϊόν τότε θεωρώ ότι θα το αγοράσει. Τα μέτρα υπολογισμένα με 10 επαναλήψεις bootstrap που καταλήγω για το συγκεκριμένο είναι τα εξής :

Threshold	Sensitivity	1-Specificity	Accuracy
0.06	0.9182	0.2319	0.7829

K-fold Cross Validation

Λόγω του ότι ο κύριος σκοπός μας είναι να μπορούμε να προβλέψουμε σωστά τους πελάτες στις δύο ομάδες πρέπει με κάποιον τρόπο να μετρήσουμε αντικειμενικά την επιτυχία του μοντέλου μας. Προηγουμένως αναφέραμε κάποια μέτρα όμως, αν δοκιμάσουμε να τα υπολογίσουμε προβλέποντας στα δεδομένα από τα οποία έχουμε 'εκπαιδεύσει' το μοντέλο, τότε ελλοχεύει ο κίνδυνος τα μέτρα να είναι προσαυξημένα από τα πραγματικά. Η διαφορά των μέτρων στα δεδομένα εκπαίδευσης με καινούργια δεδομένα ονομάζεται overfitting. Η επίπτωση αυτού είναι να χρησιμοποιούμε διάφορες μεθόδους για να εκτιμήσουμε τα πραγματικά μέτρα δύο από αυτές είναι το K-fold cross Validation και το Bootstrap.

Το K-fold cross Validation λειτουργεί χωρίζοντας όλα τα δεδομένα σε κ ισομερή κομμάτια και στη συνέχεια τρέχουμε το μοντέλο αγνοώντας το 1 κομμάτι το οποίο θα χρησιμοποιηθεί για αντικειμενικό υπολογισμό μέτρων σε άγνωστα προς το μοντέλο δεδομένα.

Συγκεκριμένα χρησιμοποίησα διαφορετικό αριθμό folds για να δω αν διαφέρουν τα αποτελέσματα. Τελικά όμως είναι περίπου σταθερά οπότε έχουμε 0.916 Sensitivity και 0.756 Accuracy με το LDA.

Folds	Sensitivity	Accuracy
2	0.916	0.785
3	0.916	0.786
4	0.917	0.785
5	0.916	0.785
7	0.916	0.785
10	0.915	0.785
12	0.916	0.785
15	0.915	0.785
20	0.915	0.785

Bootstrap

Το Bootstrap είναι άλλος ένας τρόπος να βρούμε αντικειμενικά μέτρα/scores. Η βασική ιδέα είναι ότι κάνω δειγματοληψία με επανατοποθέτηση από όλα τα δεδομένα μέχρι να συλλέξω σε αριθμό το 100 % των αρχικών παρατηρήσεων. Πιο συγκεκριμένα επειδή έκανα επανατοποθέτηση μερικοί πελάτες επιλέχτηκαν πάνω από μία φορά και κάποιοι άλλοι καθόλου. Τελικά μοντελοποιώ πάνω σε αυτά που σύλλεξα και δοκιμάζω το υπόδειγμα σε αυτούς που δεν επέλεξα αρχικά.

Sensitivity		Accuracy	
Min.	0.8855	Min.	0.7681
1st Qu.	0.9093	1st Qu.	0.7805
Median	0.9144	Median	0.7853
Mean	0.9135	Mean	0.7856
3rd Qu.	0.9193	3rd Qu.	0.7921
Max.	0.9391	Max.	0.8027

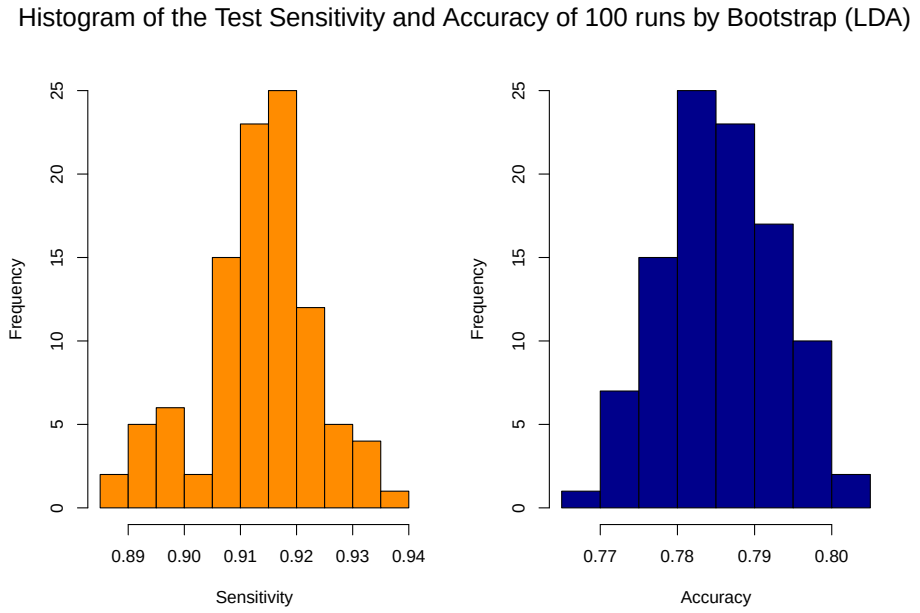

Figure 2: Ιστογράμματα των Sensitivity και Accuracy με Bootstrap για το LDA

Όπως φαίνεται παρόμοια αποτελέσματα του K-folds έχω και με το Bootrap.

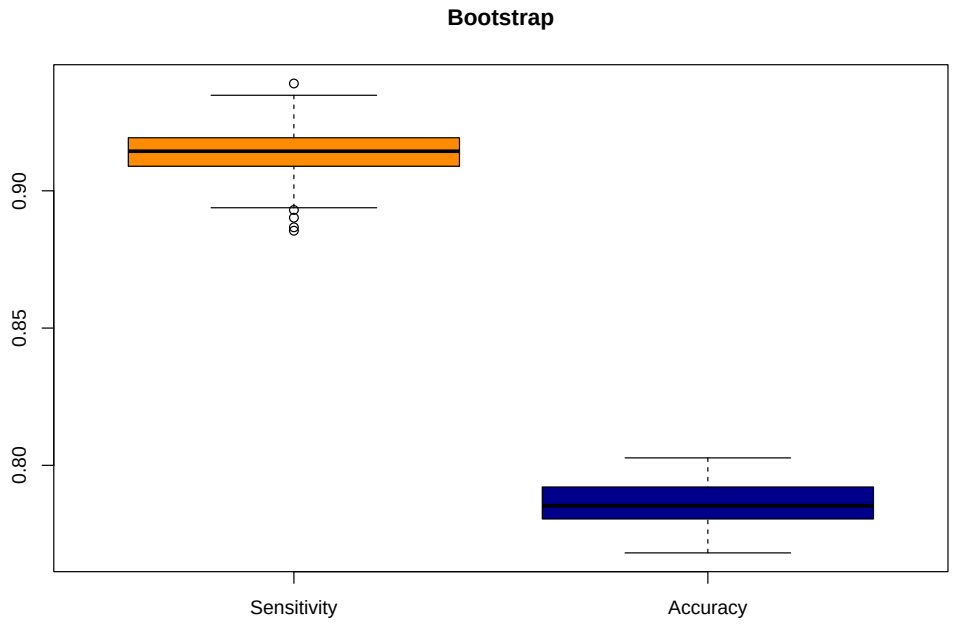

Figure 3: Boxplot των Sensitivity και Accuracy με Bootstrap για το LDA

QDA

Το QDA ή Τετραγωνική Διακριτική Ανάλυση είναι παρόμοιο του LDA μόνο που υποθέτει διαφορετικό πίνακα διακύμανσης κανονική κατανομής σε κάθε ομάδα, πιο απλά θεωρεί, ότι η ομάδες έχουν διαφορετικό σχήμα στον χώρο των δεδομένων, ενώ πριν κάναμε την παραδοχή ότι ήταν κοινός. Αυτό επιτρέπει στο μοντέλο πλέον να μην χωρίζει με ευθείες αλλά με καμπύλες.

Κατώφλι ταξινόμησης

Όπως το LDA και αυτό είναι soft classification με πιθανότητες κατάταξης αυτό συνεπάγεται ότι πρέπει να βρούμε το βέλτιστο κατώφλι με το ROC curve.

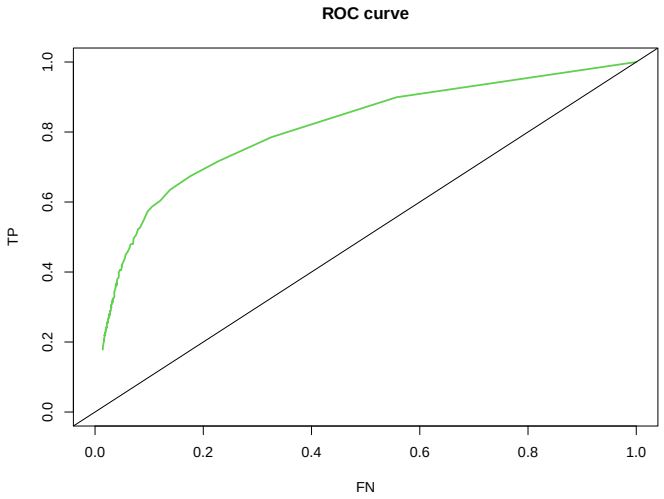

Figure 4: Το ROC curve για το QDA

Επιλέγω το σημείο κοντινότερο στο (0,1), το οποίο είναι αυτό με πιθανότητα 0.07 .

Τα μέτρα υπολογισμένα με 10 επαναλήψεις bootstrap που καταλήγω για το συγκεκριμένο είναι τα εξής :

Threshold	Sensitivity	1-Specificity	Accuracy
0.07	0.634	0.138	0.838

Διαισθητικά θα έπρεπε να είναι καλύτερα του LDA, καθώς δεν δίνω τον περιορισμό του ίδου πίνακα διακύμανσης, όμως βλέπουμε ότι είναι αρκετά χειρότερα.

K-fold Cross Validation

Με το Cross Validation ξαναβλέπουμε χειρότερα αποτελέσματα του LDA με Sensitivity 0.637 και Accuracy 0.839.

Folds	Sensitivity	Accuracy
2	0.636	0.838
3	0.637	0.838
4	0.637	0.839
5	0.637	0.838
7	0.636	0.838
10	0.636	0.838
12	0.635	0.838
15	0.636	0.838
20	0.637	0.838

Bootstrap

Συνολικά συμπεραίνω ότι έχω λίγο καλύτερο accuracy όμως πολύ χειρότερο sensitivity από το LDA.

Sensitivity		Accuracy	
Min.	0.604	Min.	0.824
1st Qu.	0.630	1st Qu.	0.833
Median	0.637	Median	0.838
Mean	0.637	Mean	0.837
3rd Qu.	0.645	3rd Qu.	0.841
Max.	0.667	Max.	0.848

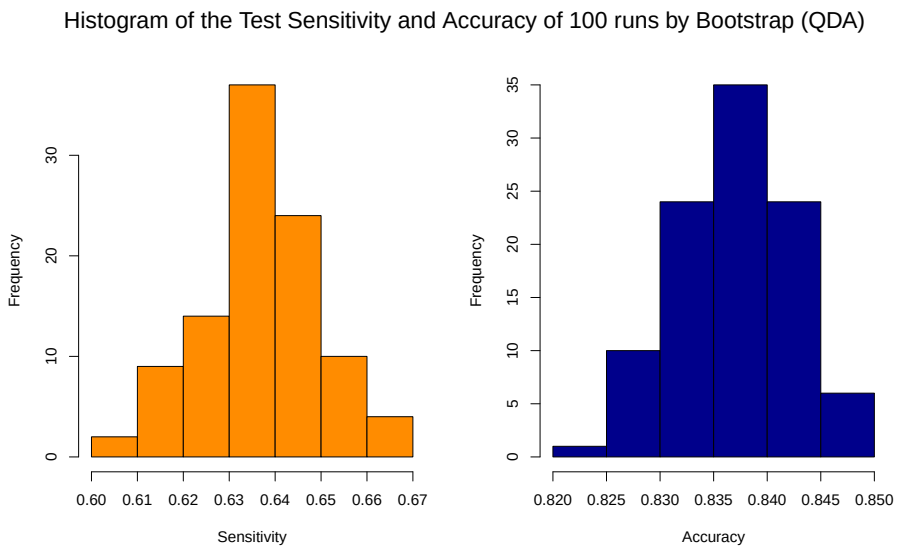

Figure 5: Ιστογράμματα των Sensitivity και Accuracy με Bootstrap για το QDA

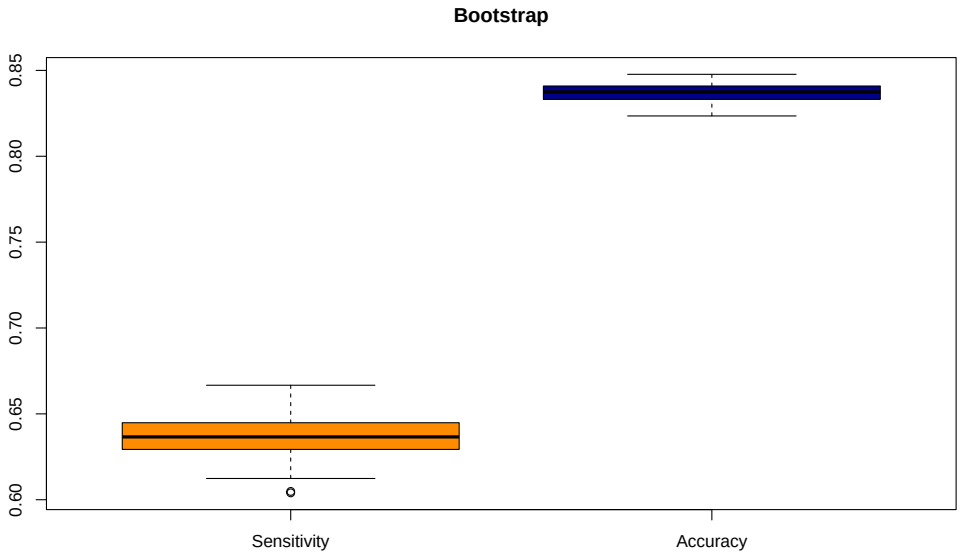

Figure 6: Boxplot των Sensitivity και Accuracy με Bootstrap για το QDA

LPM

Το LPM συμβολίζει το απλό γραμμικό μοντέλο, όπου ψάχνουμε μία γραμμική σχέση μεταξύ των χαρακτηριστικών των πελατών και της μεταβλητής που μας ενδιαφέρει ή αλλιώς μεταβλητή απόκρισης. Λόγω του ότι λειτουργεί για συνεχής μεταβλητή απόκρισης θα μετατρέψουμε τις κατηγορίες σε αριθμούς, συγκεκριμένα τις επιτυχίες θα τις όρισουμε με 1 και τις αποτυχίες με 0. Όπως και το LDA και το LPM χωρίζει γραμμικά τις ομάδες.

Επιλογή μεταβλητών

Ένα χρήσιμο χαρακτηριστικό του είναι ότι ενδεχομένως να μην χρειάζομαι όλες τις μεταβλητές, μπορώ να αξιολογήσω μοντέλα με διαφορετικές μεταβλητές στην προσπάθεια να βρω το βέλτιστο και πιο φειδωλό. Ειδικότερα θα χρησιμοποιήσω την διαδικασία forward step. Ξεκινάει με 1 μεταβλητή υπολογίζει για κάθε υποψήφια μεταβλητή ένα score, στην προκειμένη περίπτωση το Accuracy και επιλέχει εκείνη με το μεγαλύτερο score. Έχοντας επιλέξει μία εκτιμά το score με δύο μεταβλητές για κάθε υποψήφια δεύτερη και παίρνει το μεγαλύτερο. Συνεχίζει μέχρι να φτάσει το μεγαλύτερο Accuracy και αν προσθέσει ακόμα ένα παράγοντα το score μειώνεται.

Η διαδικασία φαίνεται παρακάτω με την σείρα που εισάγονται το Accuracy που σημειώνουν.

Αριθμ. Μεταβλ.	1	2	3	4	5	6	7
Μεταβλητές	pdays	duration	nr.employed	month	education	default	housing
Accuracy	0.900	0.909	0.911	0.912	0.912	0.913	0.913

Μόλις 7 μεταβλητές από τις 20 που έχω συνολικά.

Κατώφλι ταξινόμησης

Το LPM ανήκει και αυτό στα soft classification οπότε πρέπει να βρούμε το Κατώφλι ταξινόμησης

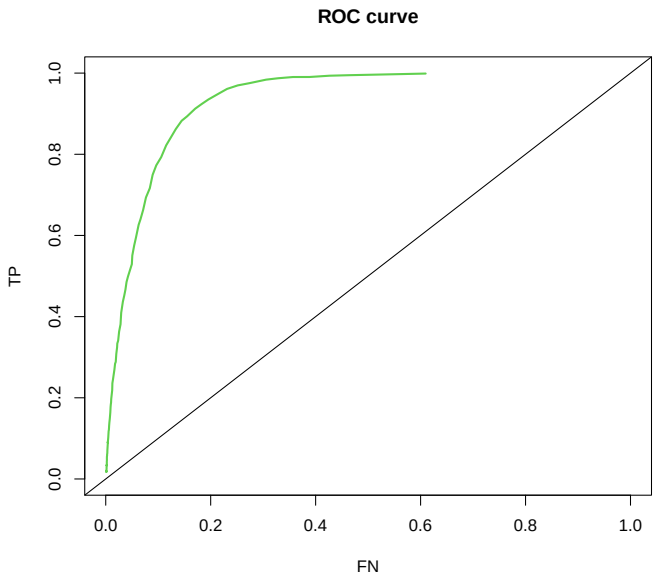

Figure 7: Το ROC curve για το LPM

Επιλέγω το σημείο κοντινότερο στο (0,1), το οποίο είναι αυτό με πιθανότητα 0.14 .

Τα μέτρα υπολογισμένα με 10 επαναλήψεις bootstrap που καταλήγω για το συγκεκριμένο είναι τα εξής :

Threshold	Sensitivity	1-Specificity	Accuracy
0.14	0.935	0.196	0.818

Αρκετά καλά αποτελέσματα.

K-fold Cross Validation

Με το Cross Validation βλέπω καλύτερα αποτελέσματα από το QDA και λίγο χειρότερα από το LDA με Sensitivity 0.882 και Accuracy 0.853.

Folds	Sensitivity	Accuracy
2	0.880	0.840
3	0.882	0.842
4	0.881	0.844
5	0.882	0.846
7	0.882	0.848
10	0.881	0.850
12	0.883	0.851
15	0.881	0.852
20	0.882	0.853

Bootstrap

Επαληθεύονται τα αποτελέσματα του Cross Validation.

Sensitivity		Accuracy	
Min.	0.856	Min.	0.846
1st Qu.	0.875	1st Qu.	0.854
Median	0.880	Median	0.858
Mean	0.881	Mean	0.857
3rd Qu.	0.888	3rd Qu.	0.861
Max.	0.898	Max.	0.866

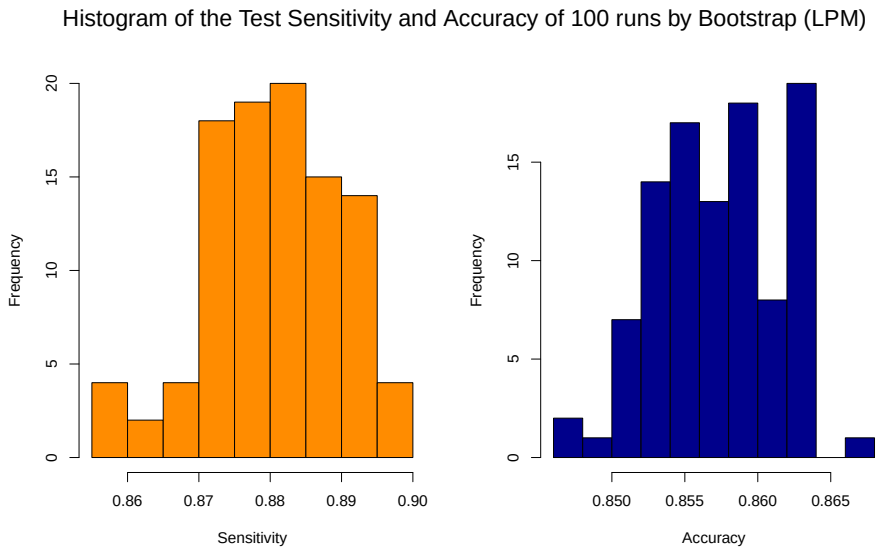

Figure 8: Ιστογράμματα των Sensitivity και Accuracy με Bootstrap για το LPM

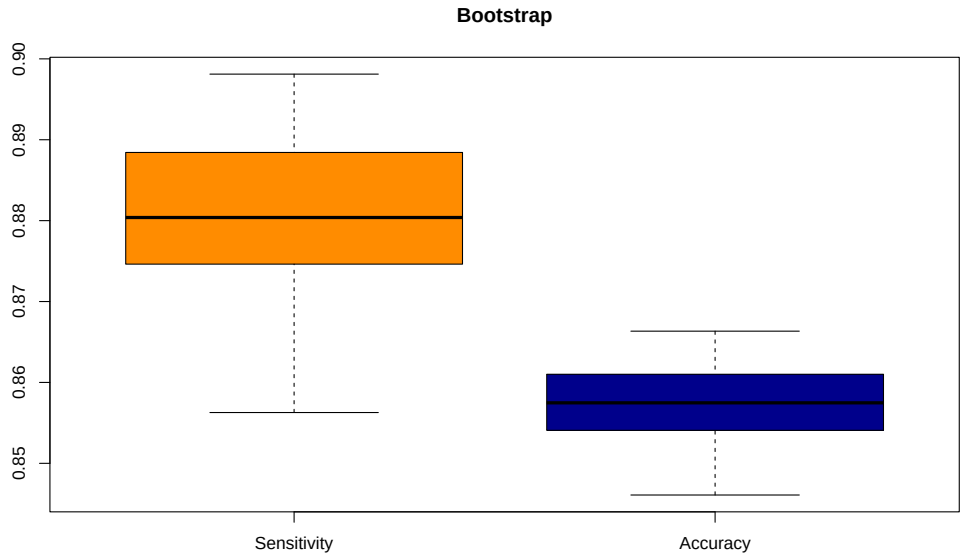

Figure 9: Boxplot των Sensitivity και Accuracy με Bootstrap για το LMP

Logistic Regression

Το Logistic Regression ή Λογιστική Παλινδρόμηση είναι το πρέπον γραμμικό μοντέλο για κατηγορικά δεδομένα υποθέτοντας ότι τα δεδομένα ακολουθούν Διωνυμική κατανομή, όποτε διαισθητικά πρέπει να τα πάει καλύτερα απο το LPM.

Επιλογή μεταβλητών

Ξανατρέχουμε forward step διαδικασία για επιλογή μεταβλητών με τον ίδιο ακριβώς τρόπο.

Αριθμ. Μεταβλ.	pdays	duration	euribor3m	month	contact
Μεταβλητές	1	2	3	4	5
Accuracy	0.810	0.820	0.822	0.824	0.824

Μόλις 5 μεταβλητές από τις 20 που έχω συνολικά ακόμα λιγότερες από τις 7 του LPM.

Κατώφλι ταξινόμησης

Το Logistic Regression ανήκει και αυτό στα soft classification οπότε πρέπει να βρούμε το Κατώφλι ταξινόμησης.

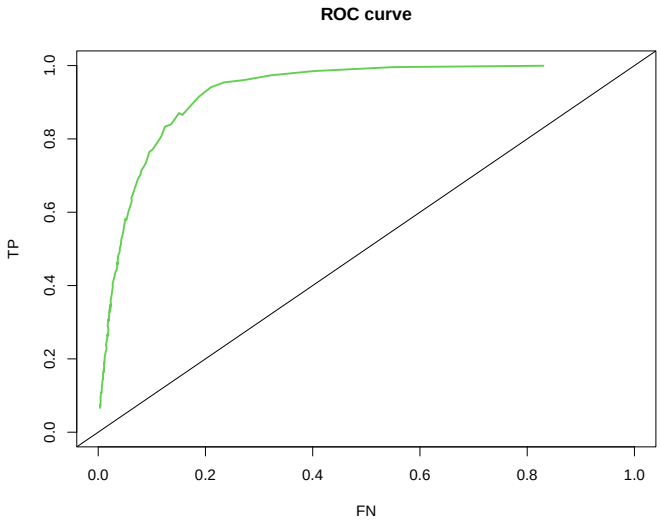

Figure 10: Το ROC curve για το Logistic Regression

Επιλέγω το σημείο κοντινότερο στο (0,1), το οποίο είναι αυτό με πιθανότητα 0.07 .

Τα μέτρα υπολογισμένα με 10 επαναλήψεις bootstrap που καταλήγω για το συγκεκριμένο είναι τα εξής :

Threshold	Sensitivity	1-Specificity	Accuracy
0.07	0.919	0.190	0.821

Λίγο καλύτερα από αυτά του LDA που έχει επικρατήσει μέχρι τώρα.

K-fold Cross Validation

Με το Cross Validation βλέπω τα καλύτερα μέχρι τώρα αποτελέσματα με Sensitivity 0.917 και Accuracy 0.785.

Folds	Sensitivity	Accuracy
2	0.931	0.806
3	0.934	0.806
4	0.934	0.806
5	0.935	0.806
7	0.935	0.806
10	0.934	0.807
12	0.935	0.807
15	0.935	0.807
20	0.935	0.807

Bootstrap

Επαληθεύονται τα αποτελέσματα του Cross Validation.

Sensitivity		Accuracy	
Min.	0.913	Min.	0.793
1st Qu.	0.928	1st Qu.	0.803
Median	0.934	Median	0.806
Mean	0.933	Mean	0.806
3rd Qu.	0.939	3rd Qu.	0.809
Max.	0.950	Max.	0.817

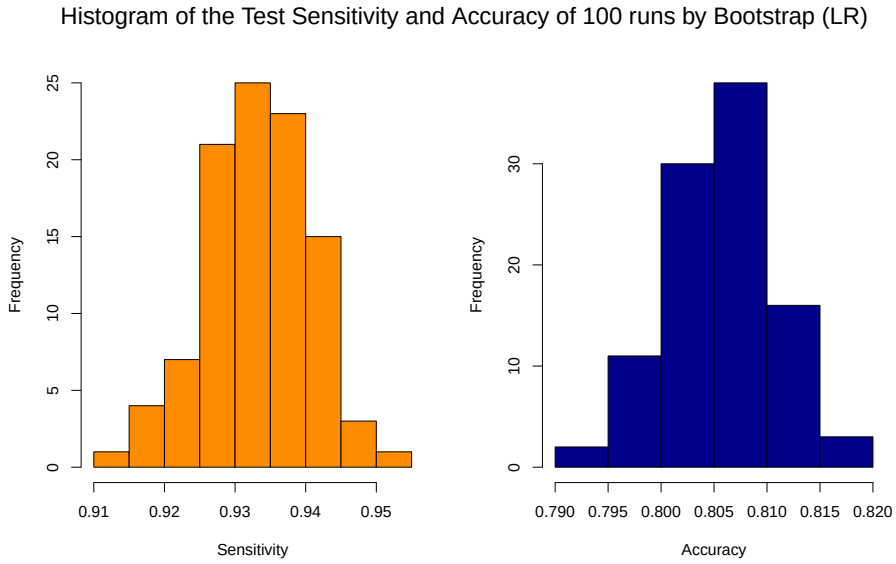

Figure 11: Ιστογράμματα των Sensitivity και Accuracy με Bootstrap για το Logistic

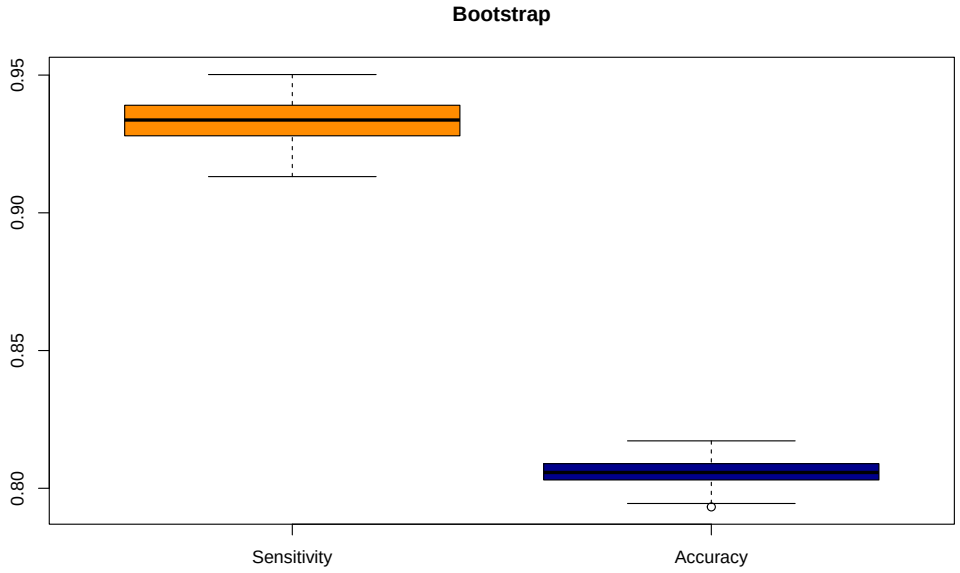

Figure 12: Boxplot των Sensitivity και Accuracy με Bootstrap για το Logistic

Validation

Για να γίνω ακόμα πιο αντικειμενικός θα τρέξω Bootstrap του καλύτερου μου μοντέλου (Logistic) στο 10% των αρχικών δεδομένων που είχα αφήσει στην άκρη εξαρχής για να έχω μία πιο ρεαλιστική άποψη στα αποτελέσματα που θα περιμένω αν το τρέξω στην πραγματικότητα. Ο λόγος που δεν αρκούμε στα test δεδομένα είναι διότι πάνω σε αυτά διάλεξα το καλύτερο μοντέλο όποτε αναγκαστικά θα ήταν προσαυξημένα από'τι θα έπρεπε.

Sensitivity		Accuracy	
Min.	0.841	Min.	0.785
1st Qu.	0.906	1st Qu.	0.807
Median	0.919	Median	0.813
Mean	0.920	Mean	0.815
3rd Qu.	0.937	3rd Qu.	0.823
Max.	0.980	Max.	0.853

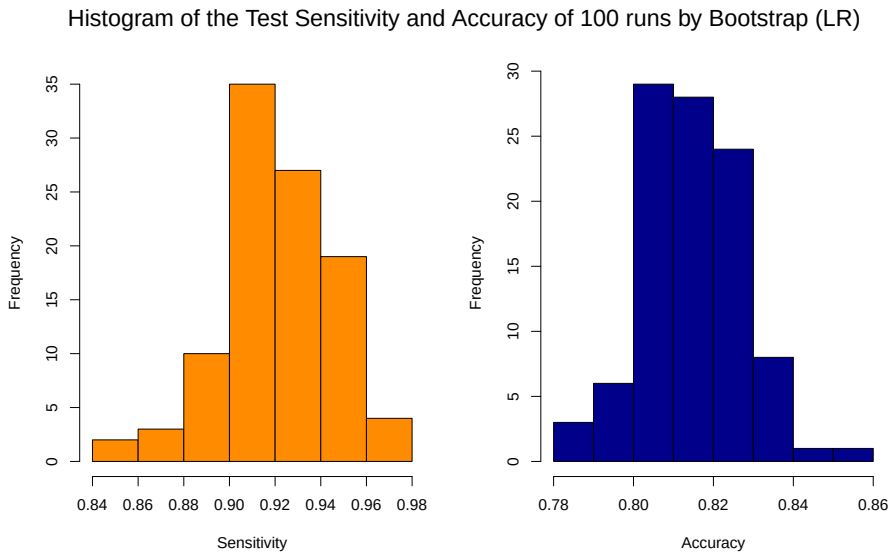

Figure 13: Ιστογράμματα των Sensitivity και Accuracy με Bootstrap για το Logistic

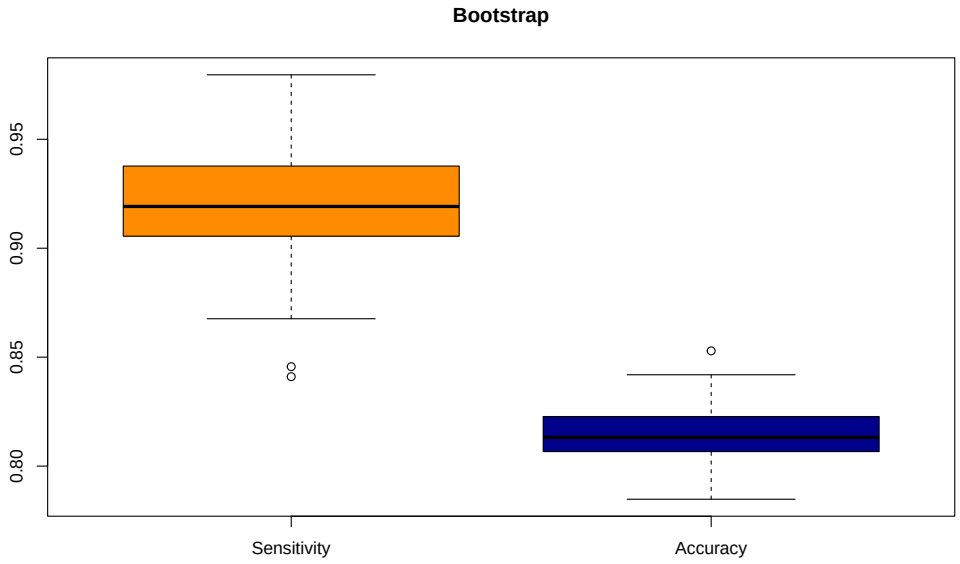

Figure 14: Boxplot των Sensitivity και Accuracy με Bootstrap για το Logistic