Πηνελόπη Μπουζιάνη-p3200115

# Εργασία 2 – Μέθοδοι Στατιστικής και Μηχανικής Μάθησης

Τα 3 μοντέλα πρόβλεψης που επιλέχθηκαν είναι : Δέντρα Αποφάσεων, Λογιστική Παλινδρόμηση και Μέθοδος Κοντινότερων Γειτόνων.

## Μεταβλητές

- Σε κάθε μέθοδο μετατρέπουμε τις κατηγορικές μεταβλητές ("job", "marital", "education", "default", "housing", "loan", "contact", "month", "day_of_week", "poutcome") σε δυαδικές ώστε να μπορούν να χρησιμοποιηθούν μαζί με τις αριθμητικές και έτσι δημιουργούνται νέες μεταβλητές .πχ από job σε jobadmin, jobblue_collar, jobunemployed etc.

- Δημιουργούμε ένα train και ένα data set με 80% και 20% των δεδομένων αντίστοιχα

- Μέσω του κώδικα σε όλα τα δεδομένα μας εξάγουμε τις 10 πιο σημαντικές μεταβλητές.

```
importance <- model$variable.importance

selected_features <- names(importance)[order(importance, decreasing = TRUE)][1:10]
```

" "duration" "nr.employed" "euribor3m" "emp.var.rate" "cons.conf.idx" "cons.price.idx" "pdays" "poutcomesuccess" "previous" "poutcomenonexistent"

## Δέντρα Αποφάσεων

**Πλεονεκτήματα**: Είναι εύκολα ερμηνεύσιμα και μπορούν να χειριστούν καλά δεδομένα με αρκετές διακλαδώσεις.

Η εφαρμογή αυτής της μεθόδου έχει τα εξής χαρακτηριστικά:

Accuracy: 0.9152564

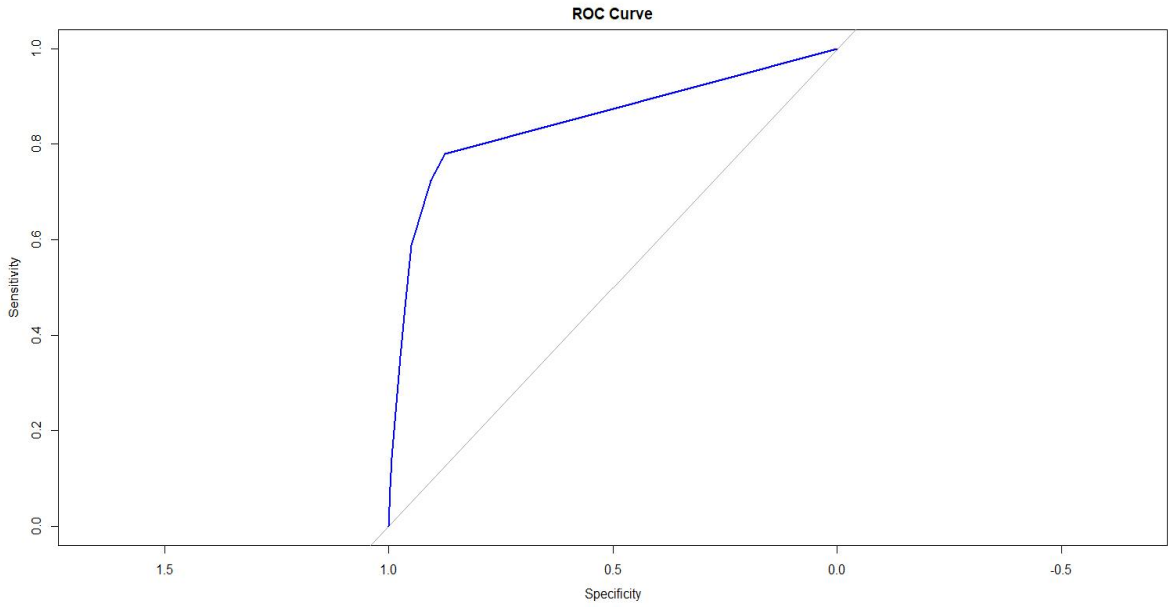
Precision: 0.5735294

Recall: 0.4582245

F1-Score: 0.509434

ROC-AUC: 0.8473335

Επειδή έχει μεγάλη ακρίβεια το δέντρο μας δεν προχωράμε σε pruning και βλέπουμε από την καμπύλη ROC ότι και το AUC είναι υψηλό




## Λογιστική Παλινδρόμηση

**Πλεονεκτήματα**: Κατάλληλη για προβλήματα δυαδικής ταξινόμησης

Εκπαιδεύοντας το μοντέλο μας με τον αρχικό τρόπο βλέπουμε ότι Accuracy: 0.0001253604 και άρα πρέπει να βελτιστοποιήσουμε το μοντέλο.

Έτσι, βελτιστοποιούμε τις υπερπαραμέτρους (hyperparameter tuning) μέσω διασταυρούμενης επικύρωσης (cross-validation). Αυτό βοηθάει στην επιλογή βέλτιστων τιμών υπερπαραμέτρων για το μοντέλο.

Επιπλέον, χρησιμοποιούμε τον αλγόριθμο glmnet για την εκπαίδευση του μοντέλου λογιστικής παλινδρόμησης. Αυτός ο αλγόριθμος υποστηρίζει λύσεις με σπαρσότητα (sparse solutions), οι οποίες είναι χρήσιμες όταν έχουμε πολλές μηδενικές τιμές στα χαρακτηριστικά.

Τα νέα αποτελέσματα του αλγορίθμου είναι:

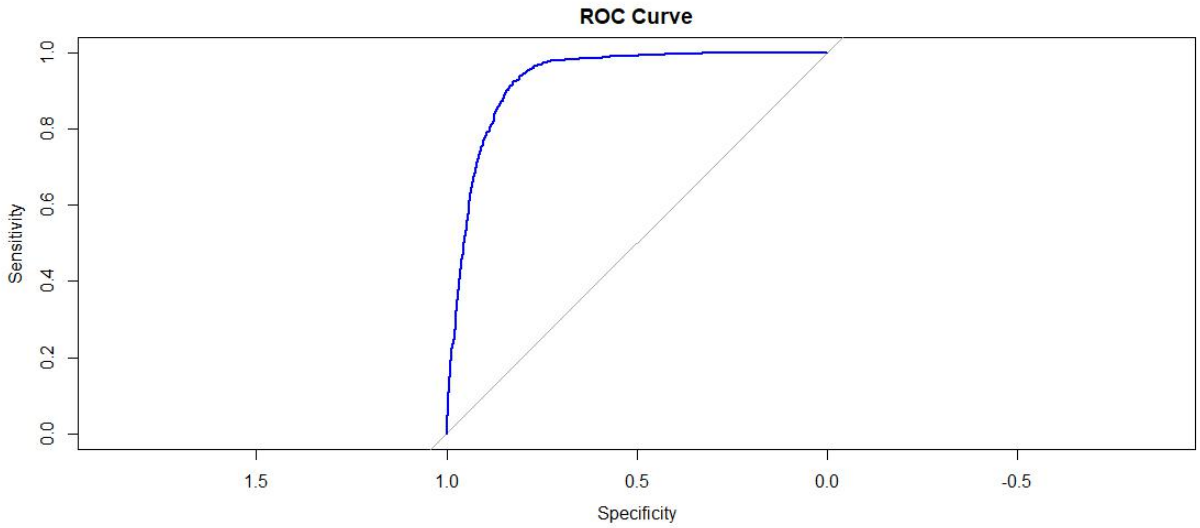
Accuracy: 0.9140028

Precision: 0.583682

Recall: 0.3642298

F1-Score: 0.4485531

ROC-AUC: 0.9289083




## Μέθοδος Κοντινότερων Γειτόνων

**Πλεονεκτήματα**: Κατάλληλη για περιπτώσεις όπου η δομή των δεδομένων είναι σημαντική και όπου μπορεί να υπάρχει περιορισμένος αριθμός χαρακτηριστικών.

Η εφαρμογή αυτής της μεθόδου με k = 5 έχει τα εξής χαρακτηριστικά:

Accuracy: 0.91187163093895

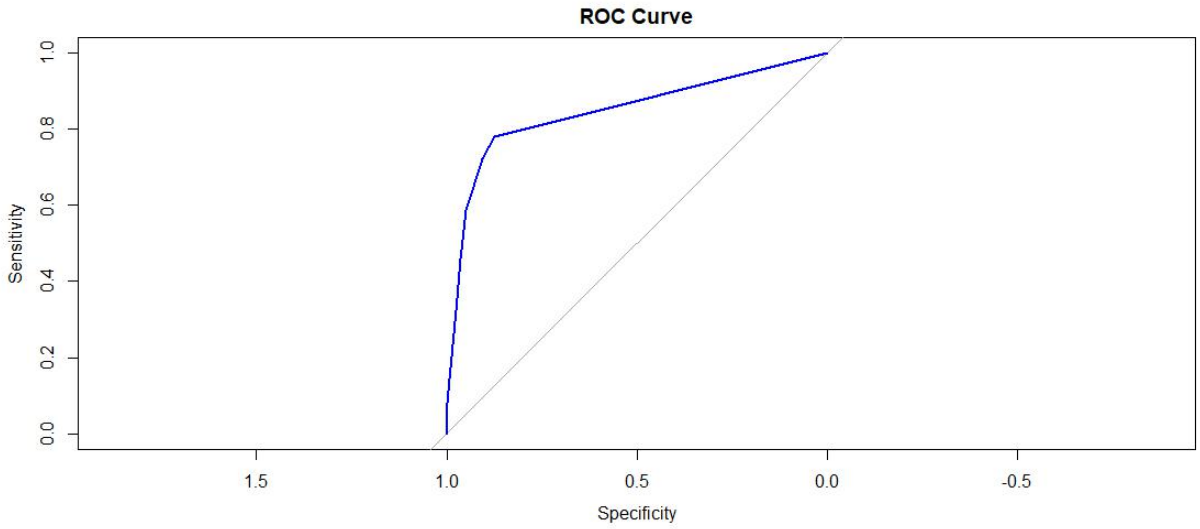
Precision: 0.5735294

Recall: 0.4582245

F1-Score: 0.509434

ROC-AUC: 0.8473335

Το k έχει επιλεχθεί δοκιμαστικά αλλά εφόσον έχουμε πολύ καλά αποτελέσματα ακρίβειας χρησιμοποιείται για τον τελικό υπολογισμό.




## Τελικό Μοντέλο

Τα αποτελέσματα των μεθόδων είναι πολύ καλά και στις 3 μεθόδους αλλά καλύτερο μοντέλο θεωρείται το βελτιστοποιημένο μοντέλο της Λογιστικής Παλινδρόμησης αφού έχει και καλύτερη καμπύλη ROC και όπως αναφέρθηκε είναι καλύτερο για δυαδικά δεδομένα όπως τα δικά μας.