ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

2η Εργασία για το μάθημα

Τεχνικές Εξόρυξης Δεδομένων

Χρήστος Κιτσανέλης

Α.Μ. : 1115201200065

Μαρία-Ισαβέλλα Μανωλάκη-Σεμπάγιος

Α.Μ. : 1115201300093

Επιβλέποντες: Γουνόπουλος Δημήτριος, Καθηγητής

ΑΘΗΝΑ
ΜΑΙΟΣ 2017

Οπτικοποίηση

Διαβάζουμε το αρχείο train, διατρέχουμε μία μία τις στήλες και για κάθε attribute αποθηκεύουμε σε ένα set τις δυνατές απαντήσεις και σε μία λίστα τις απαντήσεις όσων είναι καλοί και σε μια άλλη τις απαντήσεις όσων είναι κακοί. Στη συνέχεια, εξετάζουμε την πρώτη απάντηση κάθε attribute για να δούμε εαν είναι numerical ή όχι, ώστε να αποφασίσουμε αν θα οπτικοποιήουμε το αποτέλεσμα χρησιμοποιώντας histogram ή boxplot.

Η οπτικοποίηση πραγματοποιείται με τις συναρτήσεις hist και box στις οποίες περνάμε τις λίστες με τις απαντήσεις των καλών και των κακών, τις ταξινομημένες δυνατές απαντήσεις και το όνομα με το οποίο θέλουμε να αποθηκευτεί το αποτέλεσμα.

Hist: Υπολογίζουμε το πλήθος της κάθε απάντησης σε κάθε λίστα και το φέρνουμε στην ταξινομημένη μορφή που θέλουμε βάση του set. Δημιουργούνται τα tuples και αναλόγως το πλήθος των tuples διαλέγουμε τις κατάλληλες ιδιότητες για το histogram. Χρησιμοποιήσαμε subplots για να απεικονίσουμε και τους καλούς και τους κακούς στο ίδιο Plot.

Box: Ορίζουμε ότι θέλουμε να γίνουν plot και τα δεδομένα των καλών και των κακών και θέτουμε τις ιδιότητες που θέλουμε να υπάρχουν στα κουτιά, στα μέσα και τα fliers.

Όλα τα plots που προκύπτουν υπάρχουν στο φάκελο Plots. Από τα αποτελέσματα μπορούμε να συμπεράνουμε τα εξής:

- Attribute1 : Περίπου το 80% όσων είναι κακοί δανειολήπτες ανήκουν στις κατηγορίες A11 και A12, ενώ περίπου το 50% των καλών ανήκουν στη κατηγορία A14. Η αναλογία καλοί:κακοί στην κατηγορία A14 είναι περίπου 6 προς 1, επομένως ένας νέος πελάτης που ανήκει σε αυτή την κατηγορία, είναι αρκετά πιθανό να είναι καλός.

- Attribute2 : Οι τιμές των κακών έχουν και μεγαλύτερο εύρος και μεγαλύτερο median.

- Attribute3 : Μπορούν να μας βοηθήσουν στην αξιολόγηση οι κατηγορίες

A32 και A34 στις οποίες ανήκουν η πλειοψηφία των δανειοληπτών.

- Attribute4 : Σχεδόν όλοι οι δανειολήπτες πήραν δάνειο για τις κατηγορίες αμάξι, εξοπλισμός, τηλεόραση και επιχείρηση

- Attribute5 : Οι καλοί έχουν πιο “μαζεμένο” εύρος τιμών σε μικρότερες τιμές.

- Attribute6 : Σχεδόν όλοι οι κακοί έχουν μικρό savings account, ενώ στις κατηγορίες A63,A64,A65 υπάρχει μεγαλύτερη αναλογία καλών:κακών.

- Attribute7 : Όσο αυξάνονται τα χρόνια πρόσληψης τόσο καλύτερη αναλογία καλών:κακών παρατηρούμε.

- Attribute8 : Παρόλο που έχουν ίδιο εύρος τιμών, στους κακούς το median είναι ίσο με 4, άρα η πλειοψηφία των κακών έχουν installment rate 4%. Άρα κάποιος νέος πελάτης με installment rate κάτω από 4% μπορεί να θεωρηθεί καλός.

- Attribute9 : Η μεγάλη πλειοψηφία των αντρών είναι καλοί, ενώ στις γυναίκες δεν υπάρχει μεγάλη διαφοροποίηση μεταξύ καλών και κακών.

- Attribute10 : Πολυ λίγα άτομα ανήκουν στις κατηγορίες A102 και A103 για να μας βοηθήσει να βγάλουμε συμπέρασμα για αυτό το feature.

- Attribute11 : Τα boxplots είναι πανομοιότυπα οπότε δε θα μας βοηθήσουν στην ομαδοποίηση.

- Attribute12 : Οι κακοί είναι σχεδόν ομοιόμορφα κατανεμημένοι σε όλες τις κατηγορίες, ενώ οι καλοί βρίσκονται κυρίως στις κατηγορίες A121 και A123.

- Attribute13 : Οι μέσες ηλικίες των καλών είναι μεγαλύτερες από αυτές των κακών, άρα ίσως μας βοηθήσει λίγο στην αξιολόγηση.

- Attribute14 : Περίπου το 85% όλου του train set δεν έχει άλλο installment plan, οπότε δε περιμένουμε να μας βοηθήσει πολύ αυτό το feature.

- Attribute15 : Περίπου το 70% των καλών έχει δικό του σπίτι, οπότε αυτή η απάντηση μπορεί να μας βοηθήσει στην αξιολόγηση.

- Attribute16 : Τα boxplots είναι πανομοιότυπα οπότε δε θα μας βοηθήσουν στην ομαδοποίηση.

- Attribute17 : Η αναλογία σε όλες τις απαντήσεις εκτός της A174 μοιάζουν οπότε δε περιμένουμε να παίξει κάποιο ρόλο στην αξιολόγηση.

- Attribute18 : Τα boxplots είναι πανομοιότυπα οπότε δε θα μας βοηθήσουν στην ομαδοποίηση.

- Attribute19 : Και στις δύο απαντήσεις, η αναλογία καλών:κακών είναι περίπου 2,5:1 , οπότε δε φαίνεται να σχετίζεται η ύπαρξη τηλεφώνου με την αξιολόγηση.

- Attribute20 : Οι foreign workers είναι ως επί το πλείστον καλοί αλλά είναι πολλοί λίγοι για να φτάσουμε σε λογικό συμπέρασμα.

Συμπερασματικά, περιμένουμε τα πιο χρήσιμα features για την αξιολόγηση να είναι τα Attribute1, Attribute2 και Attribute6. Από την άλλη, features που δε θα μας βοηθήσουν καθόλου είναι τα Attribute11, Attribute16 και Attribute18.

Κατηγοριοποίηση

Αρχικά διαβάζουμε το αρχείο και έπειτα διατρέχουμε τα columns του μέχρι και αυτό που αντιστοιχεί στο τελευταίο Attribute ώστε να κρατήσουμε μια λίστα με όσα από αυτά είναι categorical. Στην συνέχεια καλούμε την συνάρτηση get_dummies πάνω στο αρχικό dataframe δίνοντας την λίστα με τα categorical columns και κρατάμε μια μορφή του dataframe με τα categorical δεδομένα να είναι πλέον σε μια numerical αναπαράσταση. Κρατάμε το column με τα Labels και το διαγράφουμε από το νέο dataframe. Ακολουθούμε την αντίστοιχη διαδικασία για το test αρχείο, ορίζουμε τους κατηγοριοποιητές και κάνουμε normalization στα training δεδομένα (κάτι που αύξησε το accuracy). Έπειτα γίνεται το training κάθε κατηγοριοποιητή με χρήση 10-fold, και επιλέγεται αυτός με το καλύτερο accuracy. Τέλος, με τη χρήση του ακριβέστερου κατηγοριοποιήτη γίνεται το prediction του Label στο test αρχείο.

Σημείωση: Όσον αφορά την επιλογή των κατηγοριοποιητών, για την μέθοδο SVM χρησιμοποίηθηκε η LinearSVC λόγω του μικρότερου χρόνου εκτέλεσης και για την μέθοδο Naive Bayes χρησιμοποίηθηκε η BernoulliNB λόγω του υψηλότερου accuracy.

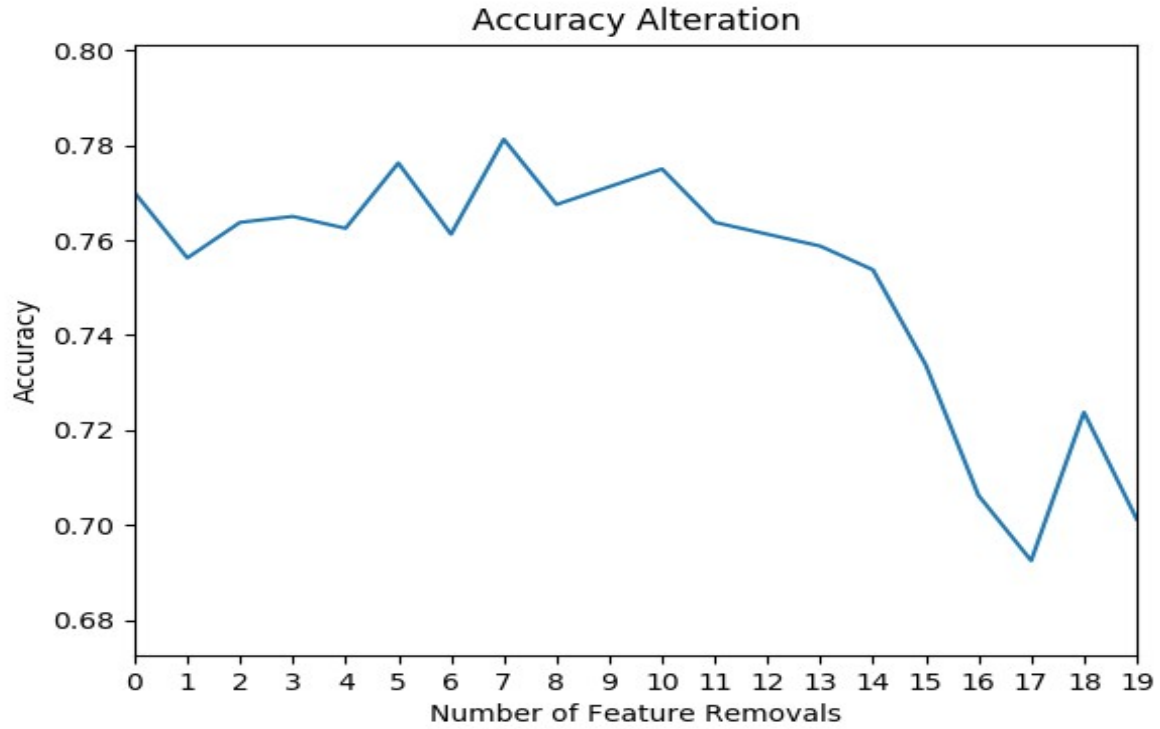
Η ακρίβεια κάθε αξιολογητή μετά τη χρήση 10-fold Cross Validation είναι η εξής:

	Random Forest	Naive Bayes	SVM
Accuracy	0.77	0.73	0.74

Επιλογή Features

Μετά το διάβασμα του αρχείου, διατρέχουμε τη στήλη Labels και κρατάμε το πλήθος των καλών και των κακών πελατών, το συνολικό πλήθος και την συνολική εντροπία. Στη συνέχεια διατρέχουμε το αρχικό dataframe, ελέγχουμε ποια columns είναι categorical και τα μετατρέπουμε σε numerical με την χρήση της pd.cut() (για την διαδικασία γίνεται χρήση 5 bins). Δημιουργούμε ένα λεξικό που περίεχει τις κατηγορίες και το πλήθος τους μέσα στο feature, και ένα λεξικό των κατηγοριών και το πλήθος των καλών πελατών που τους αντιστοιχούν, για κάθε column. Υπολογίζουμε την εντροπία κάθε κατηγορίας και το Information Gain κρατώντας σε μια ταξινομημένη λίστα κάθε ζευγάρι attribute-information gain. Οι πληροφορίες αυτές γράφονται σε αρχείο. Έπειτα από ένα αντίγραφο του αρχικού dataframe αφαιρούμε ένα ένα τα attributes με σειρά από αυτό με το μικρότερο information gain προς αυτό με το μεγαλύτερο, ενώ ακολουθούμε κάθε φορά την διαδικασία κατάλληλης μετατροπής των training δεδομένων που περιγράφηκε (get_dummies, normalization). Ακολουθεί η διαδικασία εκπαίδευσης του καλύτερου κατηγοριοποιήτη που βρέθηκε, για τα δεδομένα μετά την αφαίρεση κάθε feature ενώ οι τιμές του accuracy αποθηκεύονται σε λίστα η οποία χρησιμοποιείται για την δημιουργία του plot στην συνέχεια.

Το διάγραμμα που προκύπτει είναι το εξής:



Η σειρά με την οποία διαγράφονται τα features και τα αντίστοιχα Information Gain τους είναι η εξής:

Attribute	Information Gain
Attribute18	0.0001296657
Attribute11	0.0002205713
Attribute19	0.0012028626
Attribute16	0.0023957701
Attribute17	0.0029403166
Attribute10	0.0056743998
Attribute14	0.0070415063
Attribute8	0.0073305001
Attribute20	0.0077043865
Attribute15	0.0116188868
Attribute9	0.0127468412
Attribute13	0.0134129805
Attribute7	0.0145478652
Attribute12	0.0149055309
Attribute5	0.0184611461
Attribute6	0.0221989661
Attribute4	0.026897452
Attribute2	0.0329634294
Attribute3	0.0378894062
Attribute1	0.093827963