Στέφανος Κανελλόπουλος 1115201200050

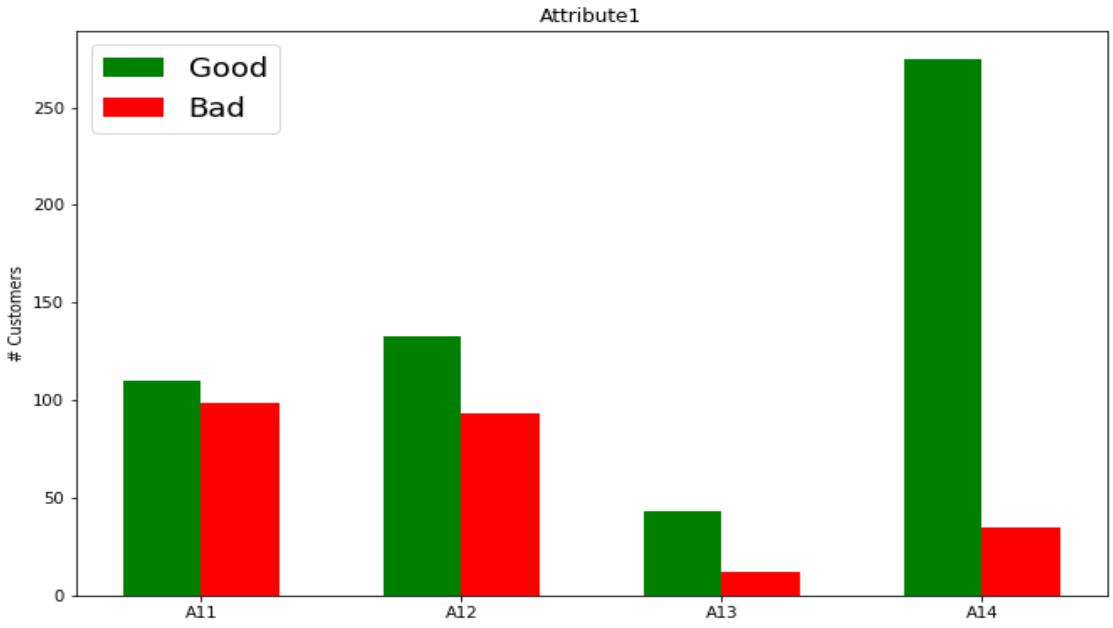
Χανιωτάκης – Ψύχος Χαρίδημος 1115201200194

2η Άσκηση Τεχνικές Εξόρυξης Δεδομένων

Η άσκηση ολοκληρώθηκε με επιτυχία, απατώντας σε όλα τα ζητούμενα ερωτήματα. Η υλοποίηση έγινε με Python 3.6 χρησιμοποιώντας την πλατφόρμα Anaconda (Spyder). Ακολουθεί περιγραφή σχετικά με τις τεχνικές υλοποίηση που ακολουθήσαμε.

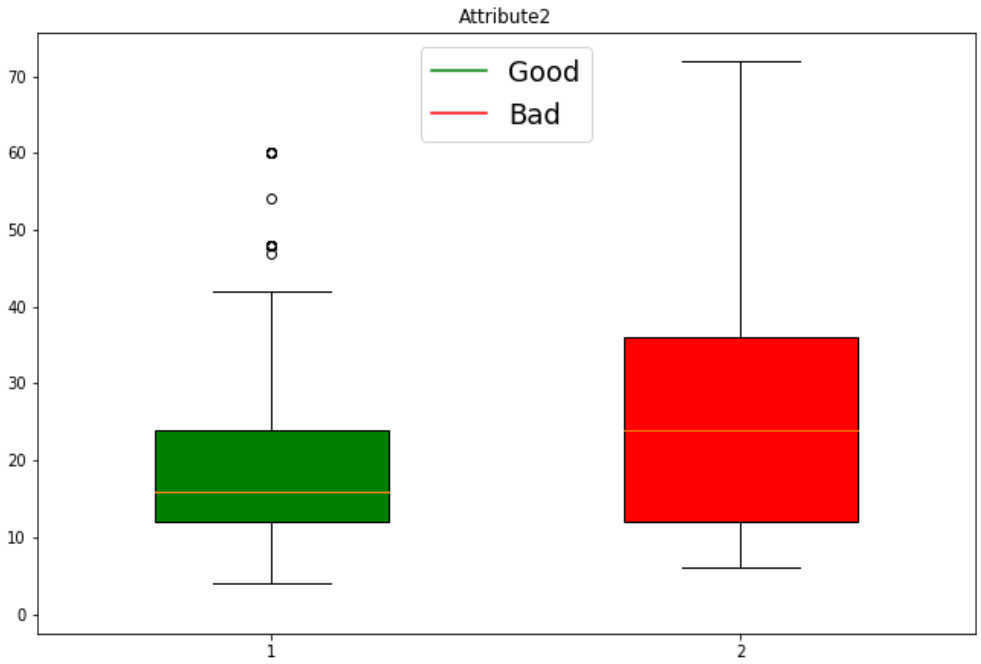
Οπτικοποίηση των Δεδομένων

- *Status of existing checking account (Categorical) :*



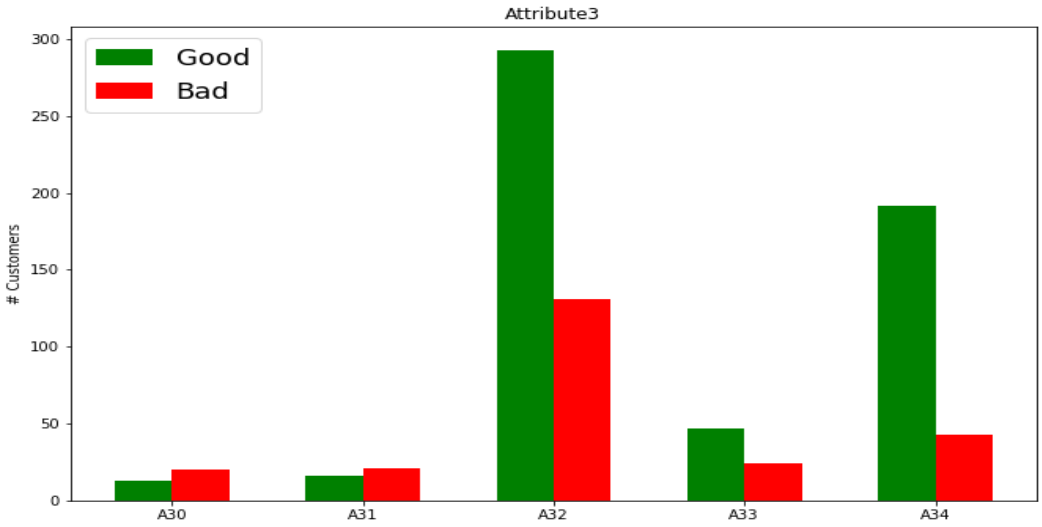
Στο παραπάνω ιστόγραμμα παρατηρούμε ότι το σημαντικότερο feature είναι το A14 (no checking account) μιας και υπάρχει αρκετά μεγάλη διαφορά στο πλήθος των καλών δανειοληπτών σε σχέση με τους κακούς. Τα υπόλοιπα features θα λέγαμε πως δεν αποτελούν ιδιαίτερα καλή πληροφορία για την αξιολόγηση ,του αν κάποιος είναι καλός ή κακός δανειολήπτης.

- *Duration in month (Numerical) :*



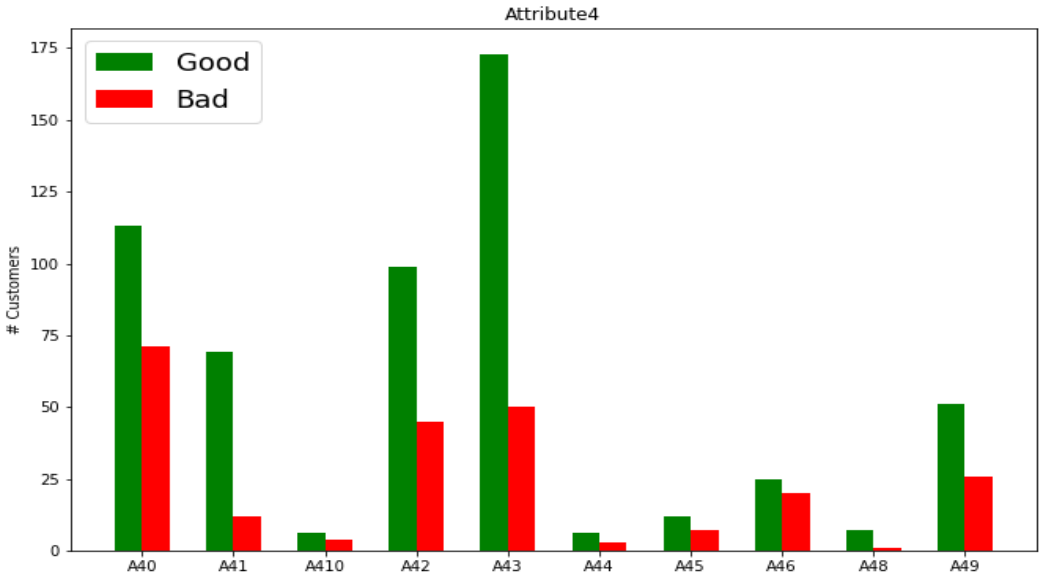
Αυτό το attribute αναφέρεται στο χρονικό διάστημα αποπληρωμής του δανείου σε μήνες. Παρατηρούμε ότι οι καλοί αποπληρωτές ζητάνε μικρότερο χρονικό διάστημα, κατά μέσο όρο 15 μήνες. Σε αντίθεση οι κακοί αποπληρώτες ζητάνε πιο μεγάλο χρονικό διάστημα αποπληρωμής, κατά μέσο όρο 25 μήνες.

- *Credit History (Categorical) :*



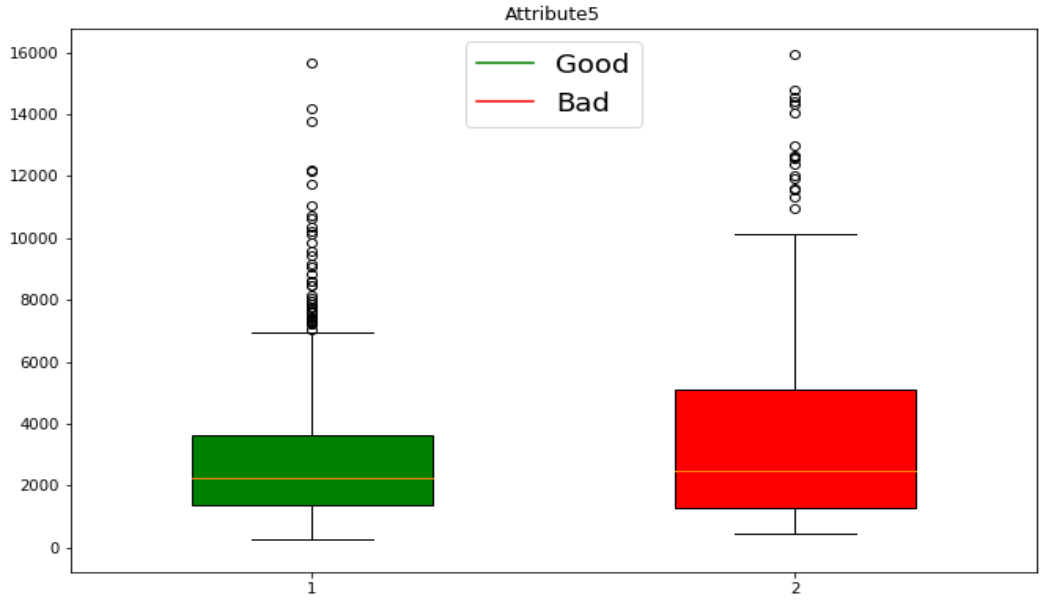
Στο παραπάνω ιστόγραμμα παρατηρούμε ότι τα σημαντικότερα features είναι τα A32 (existing credits paid back duly till now), A34 (other credits existing not at this bank) μιας και υπάρχει αρκετά μεγάλη διαφορά στο πλήθος των καλών υποψήφιων δανειοληπτών σε σχέση με τους κακούς. Τα υπόλοιπα features θα λέγαμε πως δεν αποτελούν ιδιαίτερα καλή πληροφορία για την αξιολόγηση ,του αν κάποιος είναι καλός ή κακός δανειολήπτης.

- *Purpose (Categorical) :*



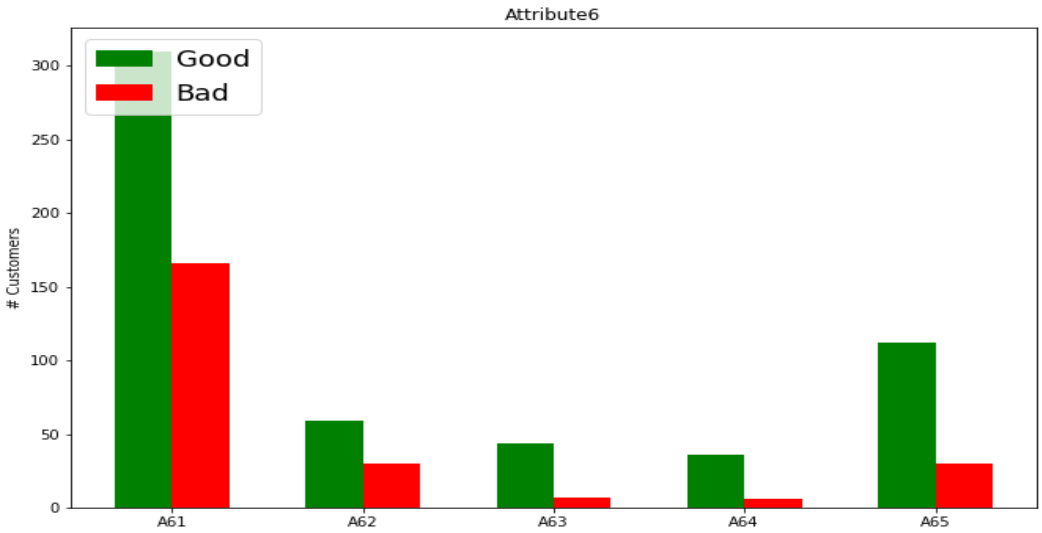
Στο παραπάνω ιστόγραμμα παρατηρούμε ότι τα σημαντικότερα με διαφορά features είναι τα A41 (used car), A42 (furniture/equipment), A43 (radio/television). Θεωρούνται βασικά αγαθά για την καθημερινότητα και οι πελάτες που παίρνουν δάνεια για αυτό τον σκοπό, είναι περισσότερο πιθανό να το αποπληρώσουν. Τα υπόλοιπα features θα λέγαμε πως δεν αποτελούν ιδιαίτερα καλή πληροφορία για την αξιολόγηση ,του αν κάποιος είναι καλός ή κακός δανειολήπτης.

- *Credit Amount (Numerical) :*



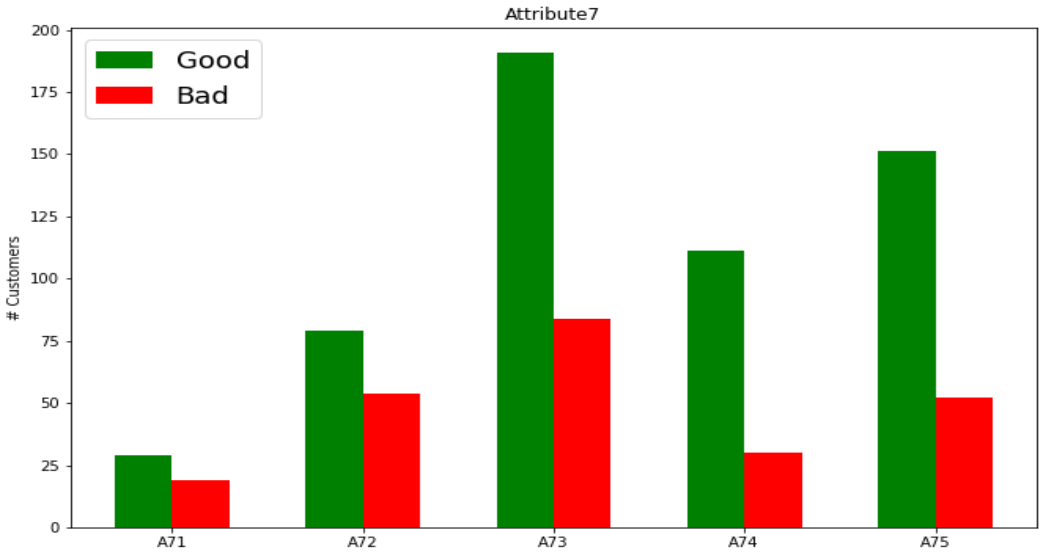
Από το παραπάνω box plot βγάζουμε το συμπέρασμα ότι το συγκεκριμένο attribute δεν μας προσφέρει σημαντική πληροφορία, τέτοια ώστε να διαχωρίσουμε τους πελάτες, με βάση το ποσό δανείου που ζητάνε, σε κάλους και κακούς αποπληρωτές.

- *Savings account/bonds (Categorical) :*



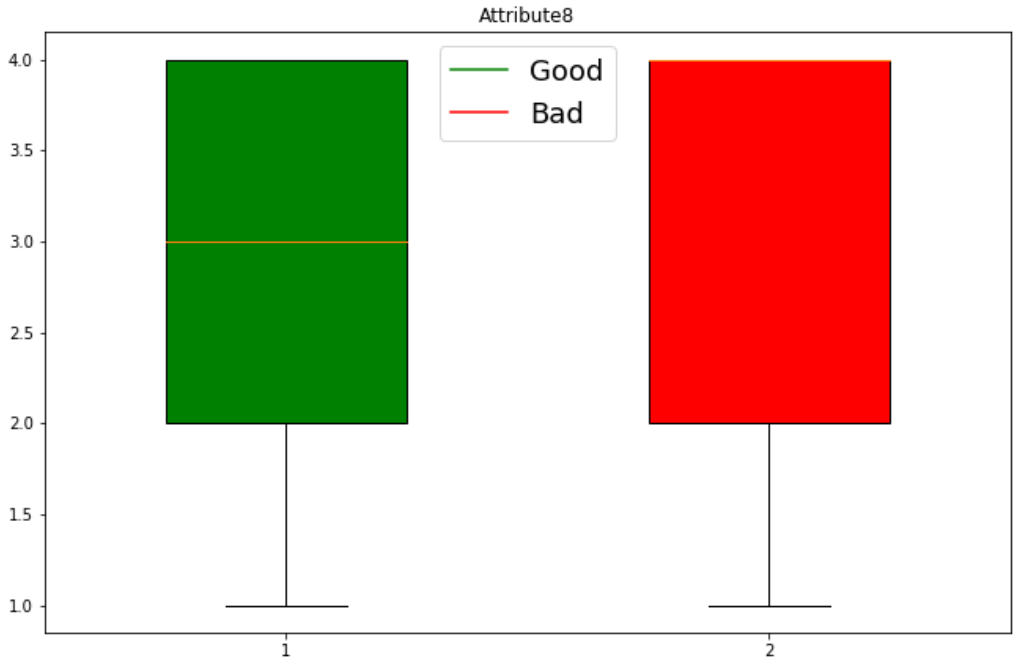
Στο παραπάνω ιστόγραμμα παρατηρούμε ότι τα σημαντικότερα features είναι τα A63 (500 DM < savings < 1000 DM), A64 (1000 DM < savings), A65 (unknown savings) μιας και υπάρχει αρκετά μεγάλη διαφορά στο πλήθος των καλών υποψήφιων δανειοληπτών σε σχέση με τους κακούς. Το συμπέρασμα που βγάζουμε είναι ότι αν κάποιος πελάτης έχει αποταμιεύσεις, το πιο πιθανό είναι να αποπληρώσει το δάνειο του.

- *Present employment since (Categorical) :*



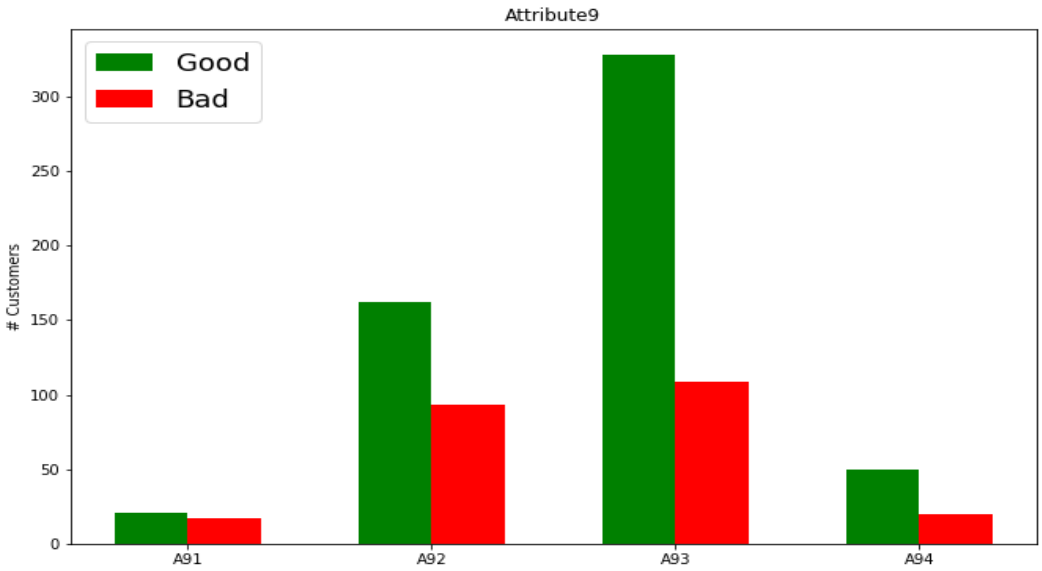
Στο παραπάνω ιστόγραμμα παρατηρούμε ότι τα σημαντικότερα features είναι τα A73 (1 έως 4 χρόνια), A74 (4 έως 7 χρόνια), A75 (περισσότερο από 7 χρόνια) μιας και υπάρχει αρκετά μεγάλη διαφορά στο πλήθος των καλών υποψήφιων δανειοληπτών σε σχέση με τους κακούς. Αυτό μας δείχνει ότι αν κάποιος έχει σταθερή δουλειά, δηλαδή ένα εγγυημένο μηνιαίο εισόδημα, είναι πιο βεβαιότερο ότι το δάνειο του θα αποπληρωθεί.

- *Installment rate in percentage of disposable income (Numerical) :*



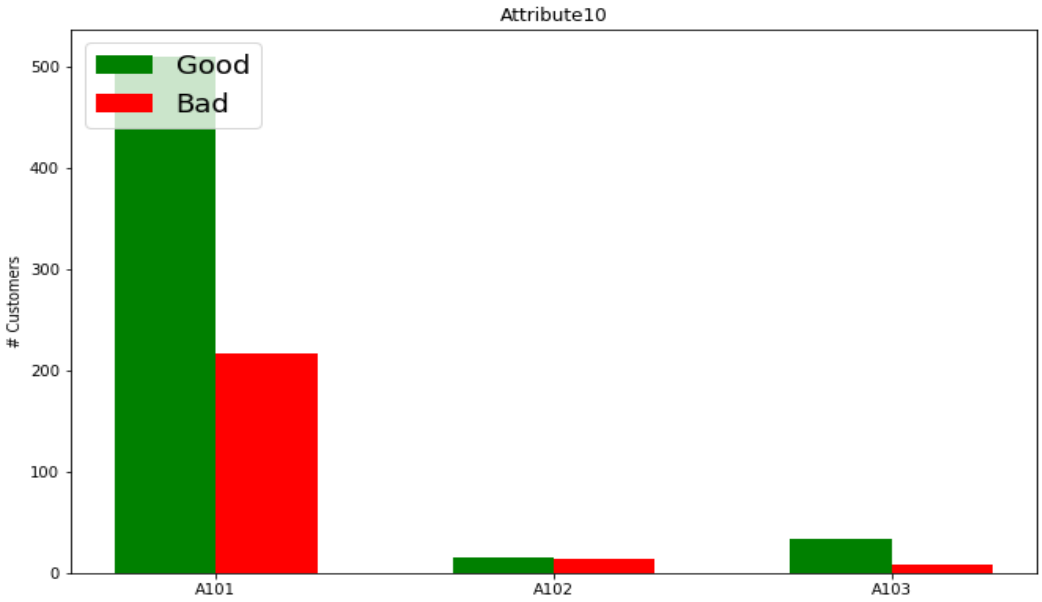
Από το παραπάνω box plot το συμπέρασμα που βγάζουμε είναι ότι οι καλοί αποπληρωτές διαθέτουν κατά μέσο όρο ένα 30% από το μηναίο μισθό τους, ενώ οι κακοί αποπληρωτές κατά μέσο όρο το 40%.

- *Personal status and sex (Categorical) :*



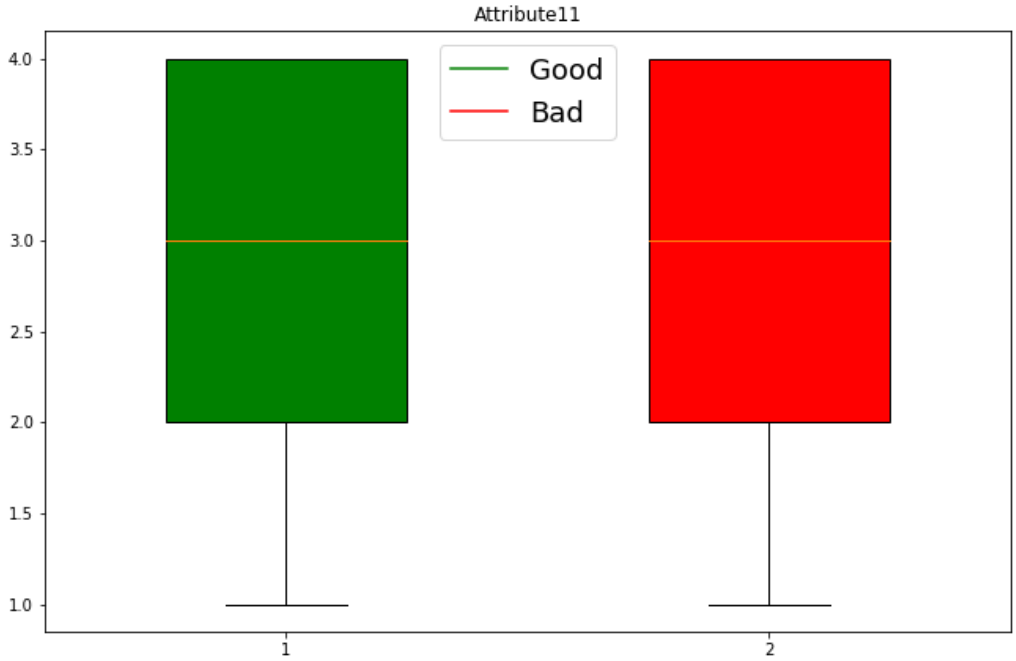
Στο παραπάνω ιστόγραμμα παρατηρούμε ότι το σημαντικότερο feature είναι το A93 (male single) μιας και υπάρχει αρκετά μεγάλη διαφορά στο πλήθος των καλών υποψήφιων δανειοληπτών σε σχέση με τους κακούς. Καταλήγουμε στο συμπέρασμα ότι οι ελεύθεροι άντρες είναι οι καλύτεροι υποψήφιοι για να πάρουν δάνειο.

- *Other debtors/guarantors (Categorical) :*



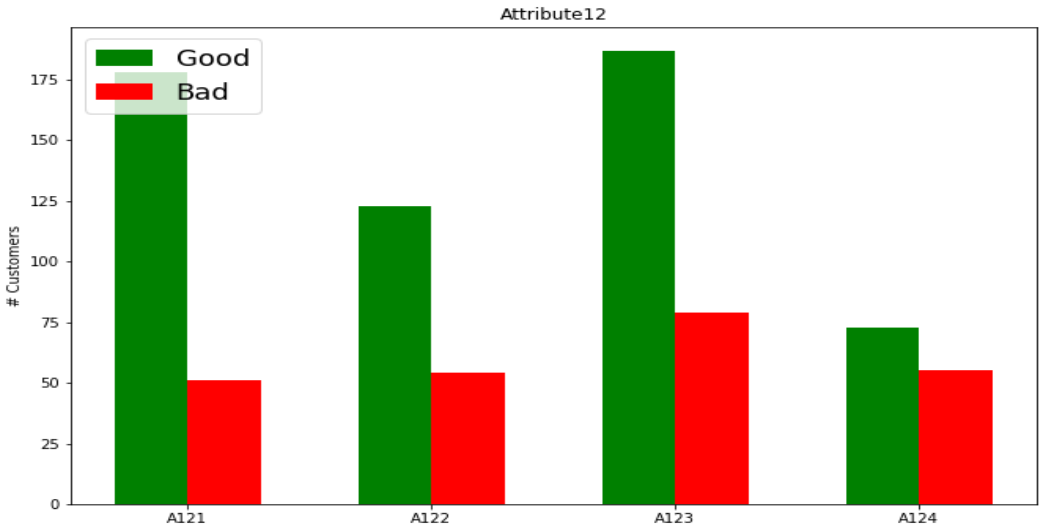
Στο παραπάνω ιστόγραμμα παρατηρούμε ότι τα σημαντικότερα features είναι τα A101 (none), A103 (guarantor) μιας και υπάρχει αρκετά μεγάλη διαφορά στο πλήθος των καλών υποψήφιων δανειοληπτών σε σχέση με τους κακούς. Από τα παραπάνω δεδομένα φαίνεται ότι οι πελάτες που έχουν κάποιον εγγυητή είναι πιο αξιόπιστοι για την αποπληρωμή του δανείου.

- *Present residence since (Numerical) :*



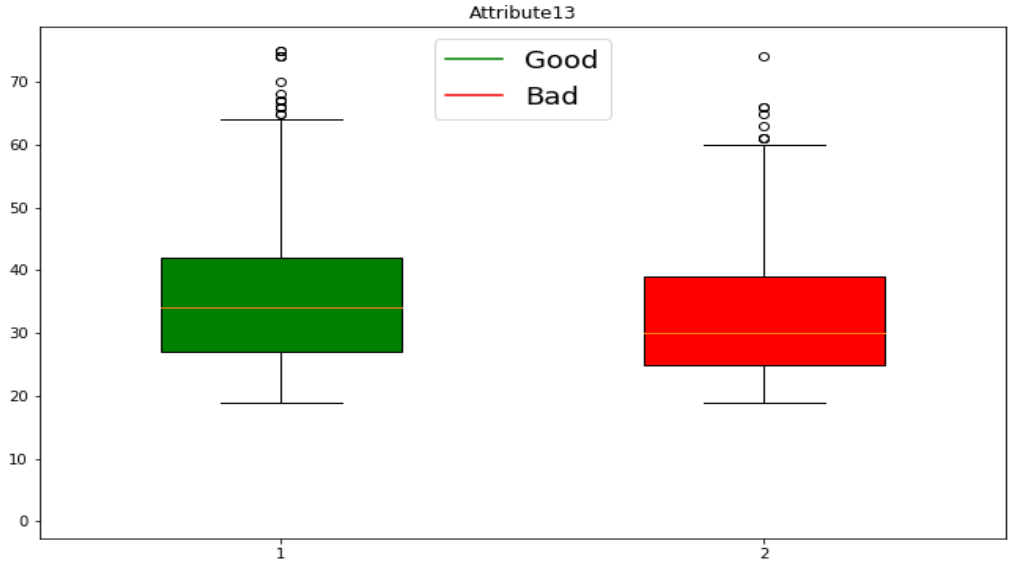
Από το παραπάνω box plot βγάζουμε το συμπέρασμα ότι το συγκεκριμένο attribute δεν μας προσφέρει σημαντική πληροφορία, τέτοια ώστε να διαχωρίσουμε τους πελάτες, με βάση την περιοχή κατοικίας τους, σε κάλους και κακούς αποπληρωτές.

- *Property (Categorical) :*



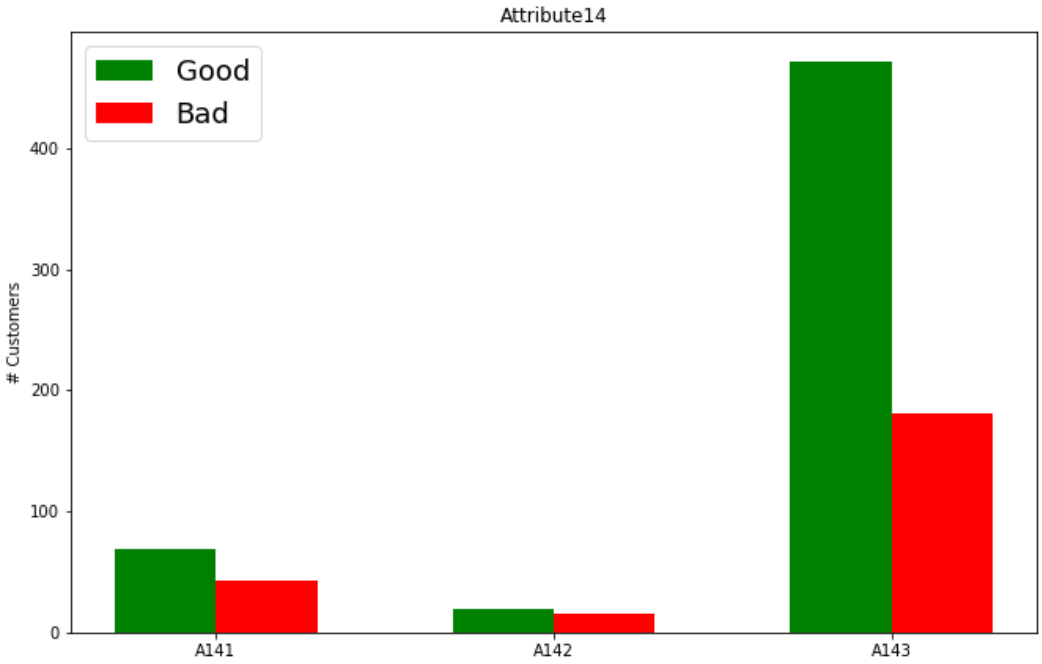
Στο παραπάνω ιστόγραμμα παρατηρούμε ότι τα σημαντικότερα features είναι τα A121 (real estate), A122 (life insurance), A123 (car) μιας και υπάρχει αρκετά μεγάλη διαφορά στο πλήθος των καλών υποψήφιων δανειοληπτών σε σχέση με τους κακούς. Το συμπέρασμα που βγάζουμε είναι ότι αν κάποιος πελάτης έχει ακίνητη περιούσια, ασφάλεια ζωής ή αμάξι, είναι πιθανότερο να αποδειχθεί καλός αποπληρωτής.

- *Age in years (Numerical) :*



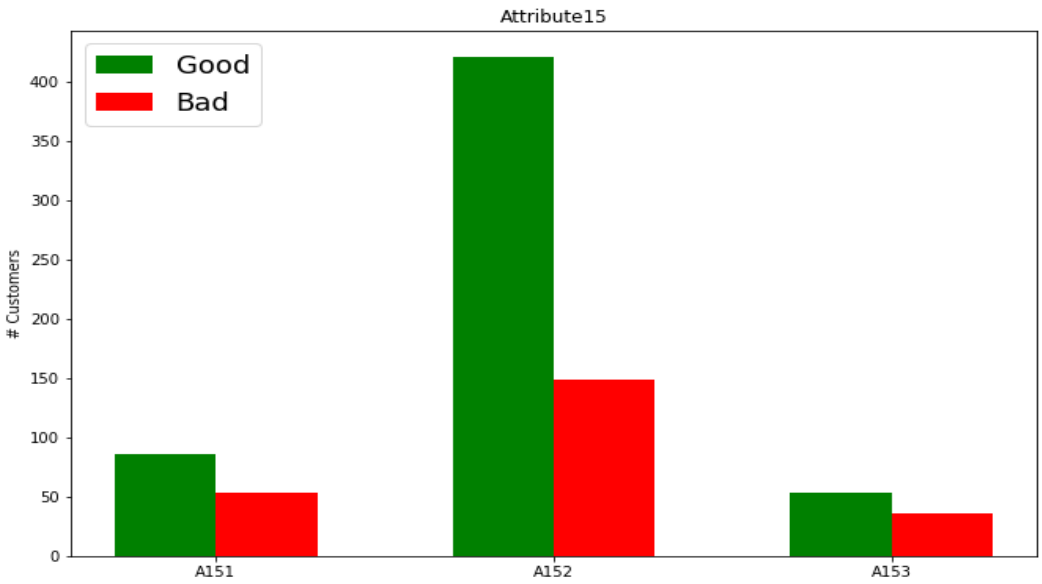
Από το παραπάνω box plot βγάζουμε το συμπέρασμα ότι οι πελάτες με μεγαλύτερη ηλικία (μέσος ορός 35 ετών) είναι πιθανότερο να αποπληρώσουν το δάνειο, σε αντίθεση με τους νεότερους (μέσος ορός 30 ετών).

- *Other installment plans (Categorical) :*



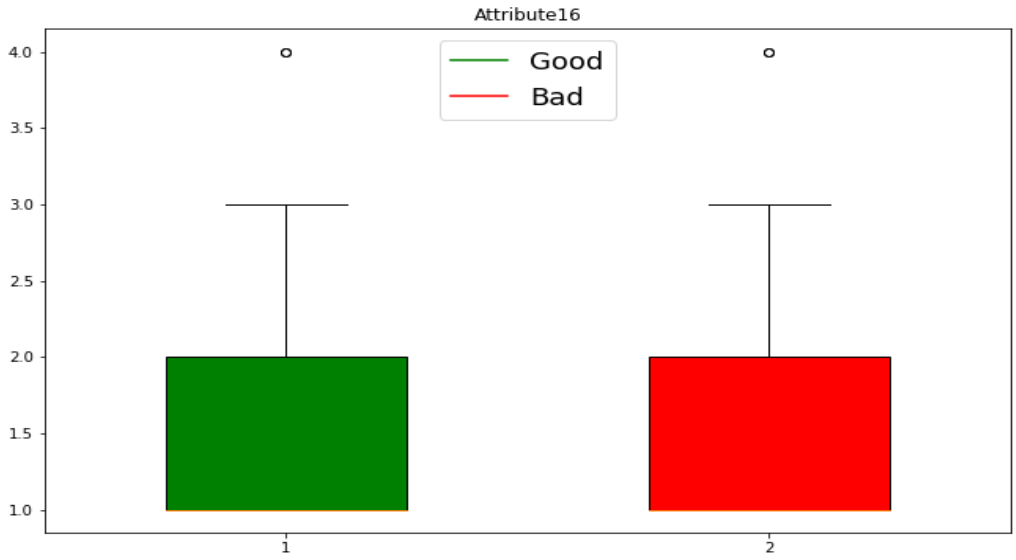
Στο παραπάνω ιστόγραμμα παρατηρούμε ότι το σημαντικότερο feature είναι το A143 (None) μιας και υπάρχει αρκετά μεγάλη διαφορά στο πλήθος των καλών υποψήφιων δανειοληπτών σε σχέση με τους κακούς. Καταλήγουμε στο συμπέρασμα ότι χωρίς installment plan είναι πιθανότερο το δάνειο να αποπληρωθεί.

- *Housing (Categorical) :*



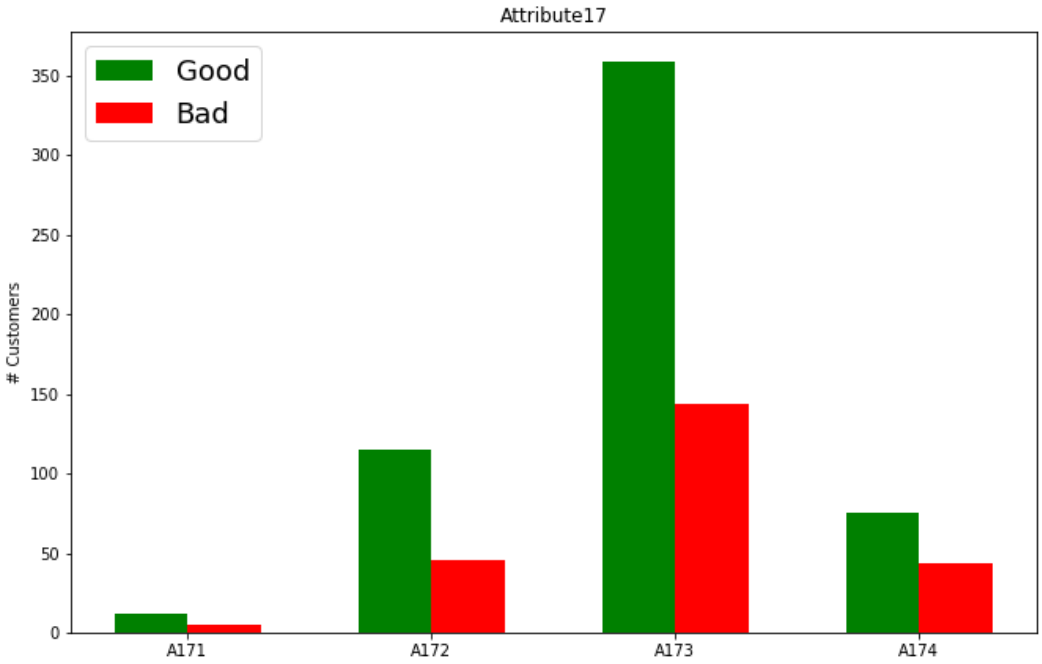
Στο παραπάνω ιστόγραμμα παρατηρούμε ότι το σημαντικότερο feature είναι το A152 (own house) μιας και υπάρχει αρκετά μεγάλη διαφορά στο πλήθος των καλών υποψήφιων δανειοληπτών σε σχέση με τους κακούς. Άρα, οι πελάτες που έχουν δικό τους σπίτι αποδεικνύονται καλοί αποπληρωμές.

- *Number of existing credits at this bank (Numerical) :*



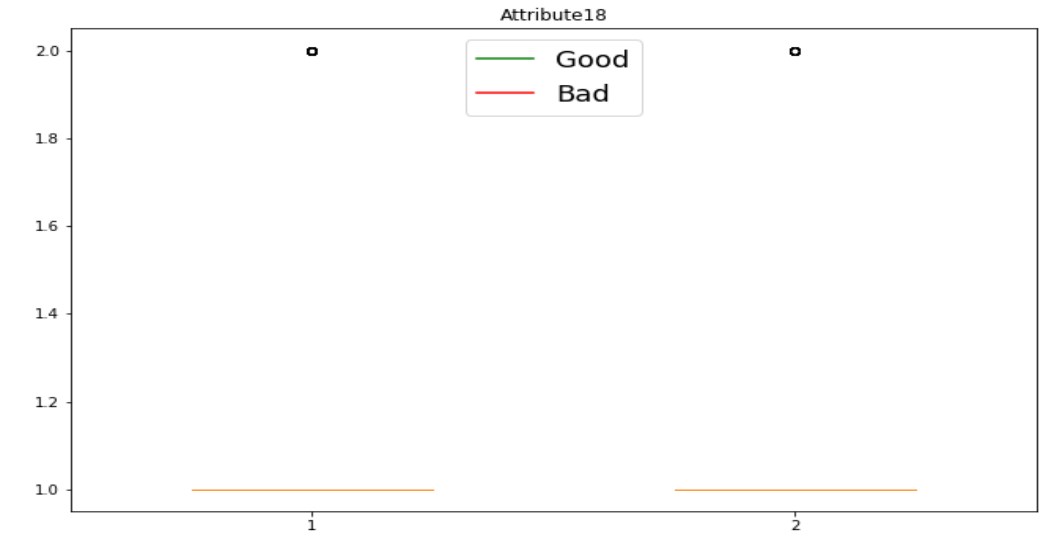
Από το παραπάνω box plot βγάζουμε το συμπέρασμα ότι το συγκεκριμένο attribute δεν μας προσφέρει σημαντική πληροφορία, τέτοια ώστε να διαχωρίσουμε τους πελάτες, με βάση το ποσό δανείου που ζητάνε, σε κάλους και κακούς αποπληρωτές.

- *Job (Categorical) :*



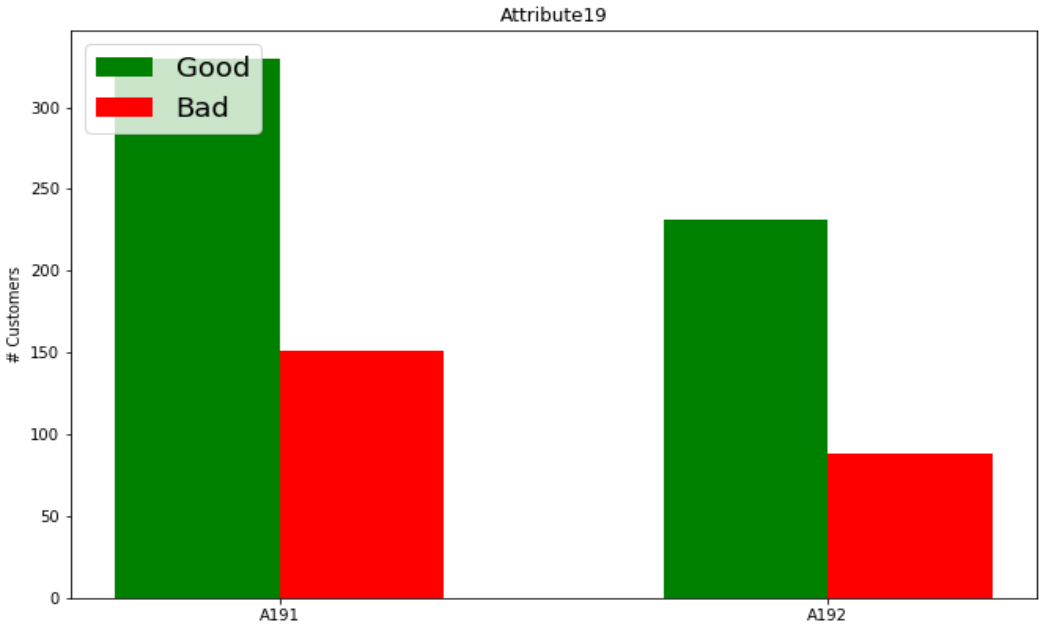
Στο παραπάνω ιστόγραμμα παρατηρούμε ότι τα σημαντικότερα features είναι τα A172 (unskilled resident), A173 (skilled employee) μιας και υπάρχει αρκετά μεγάλη διαφορά στο πλήθος των καλών υποψήφιων δανειοληπτών σε σχέση με τους κακούς. Το συμπέρασμα που βγάζουμε είναι ότι αν κάποιος πελάτης είναι ανειδίκευτος ή εξειδικευμένος εργαζόμενος, σε αντίθεση με τους ανέργους και τους ελεύθερους επαγγελματίες, είναι πιθανότερο να αποδειχθεί καλός αποπληρωτής.

- *Number of people being liable to provide maintenance for (Numerical) :*



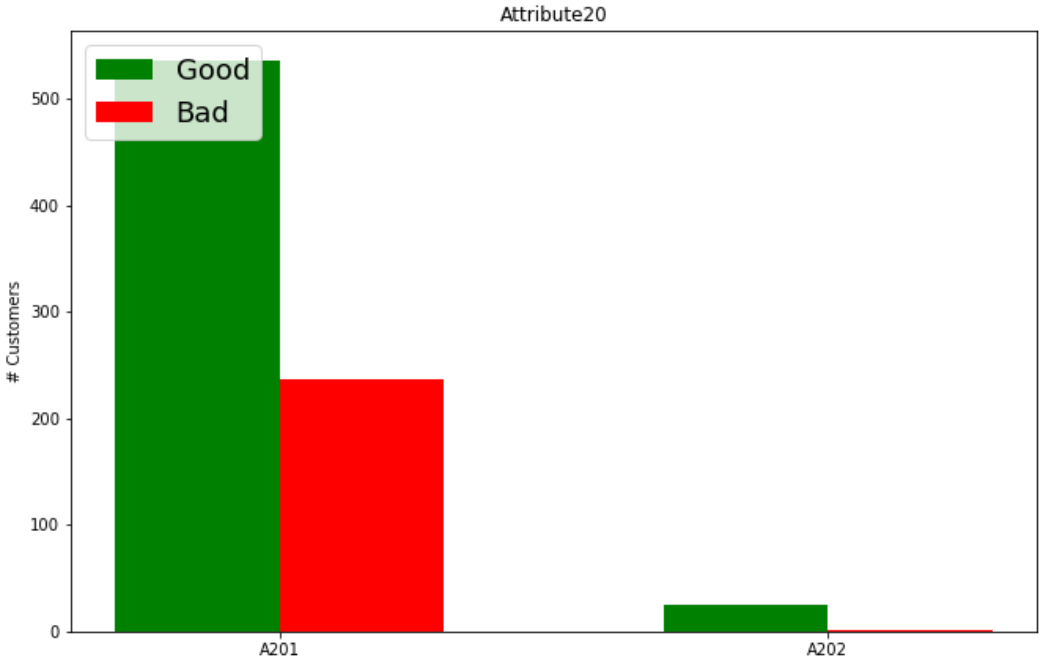
Από το παραπάνω box plot βγάζουμε το συμπέρασμα ότι το συγκεκριμένο attribute δεν μας προσφέρει σημαντική πληροφορία, τέτοια ώστε να διαχωρίσουμε τους πελάτες, με βάση το ποσό δανείου που ζητάνε, σε κάλους και κακούς αποπληρωτές.

- *Telephone (Categorical) :*



Από το παραπάνω ιστόγραμμα που αφορά την ύπαρξη κινητού τηλεφώνου βγάζουμε το συμπέρασμα ότι δεν μας προσφέρει σημαντική πληροφορία, τέτοια ώστε να διαχωρίσουμε τους πελάτες, με βάση το ποσό δανείου που ζητάνε, σε κάλους και κακούς αποπληρωτές.

- *Foreign worker (Categorical) :*



Στο παραπάνω ιστόγραμμα παρατηρούμε ότι το σημαντικότερο feature είναι το A202 (no) μιας και υπάρχει αρκετά μεγάλη διαφορά στο πλήθος των καλών υποψήφιων δανειοληπτών σε σχέση με τους κακούς. Φαίνεται, ότι οι πελάτες που δεν είναι από άλλη χώρα, είναι πιο αξιόπιστοι για την αποπληρωμή του δανείου.

Υλοποίηση Κατηγοριοποίησης (Classification)

- *Εύρεση αποδοτικότερου classifier*

Αρχικά διαβάσαμε το αρχείο ‘train.tsv’. Στην συνέχεια για τα Categorical Atrributes χρησιμοποιήσαμε τον Label Encoder και για τα Numerical τα αφήσαμε ως έχουν. Επόμενο βήμα ήταν να κάνουμε 10-Fold Cross Validation, και για τους τρεις classifiers στα encoded δεδομένα, υπολογίζοντας κάθε φορά το accuracy της πρόβλεψης. Τέλος δημιουργήσαμε το αρχείο ‘EvaluationMetric_10fold.csv’ που περιέχει το μέσο accuracy για τον καθένα.

Statistic Measure	Naïve Bayes	Random Forest	SVM
Accuracy	0.7125	0.73875	0.70125

Λαμβάνοντας υπόψιν τα παραπάνω αποτελέσματα, αποφανθήκαμε ότι ο πιο αποδοτικός classifier είναι ο Random Forest.

- *Πρόβλεψη του test set*

Για το predict του test set χρησιμοποιήσαμε το Random Forest classifier. Επεξεργαστήκαμε τα δεδομένα όπως στην διαδικασία της εύρεσης του information gain για κάθε attribute. Από το παρακάτω διάγραμμα (Accuracy – Number of Attributes) παρατηρούμε ότι το μέγιστο accuracy το πετυχαίνουμε με 12 attributes. Συγκεκριμένα αφαιρέσαμε τα Attribute18, Attribute11, Attribute19, Attribute16, Attribute17, Attribute 10, Attribute14, Attribute8. Αφού τον κάναμε training με τα δεδομένα από το αρχείο ‘train.tsv’, στην συνέχεια τον βάλαμε να προβλέψει αν τα άτομα από το αρχείο ‘test.tsv’ είναι καλοί ή κακοί δανειολήπτες. Τα αποτελέσματα βρίσκονται στο αρχείο ‘testSet_Predictions.csv’.

Επιλογή Features

Έχουμε κατασκευάσει δυο συναρτήσεις οι οποίες υπολογίζουν την εντροπία ολόκληρου του dataset και την εντροπία για κάθε attribute, βασισμένοι στον τύπο που υπάρχει στην Wikipedia.

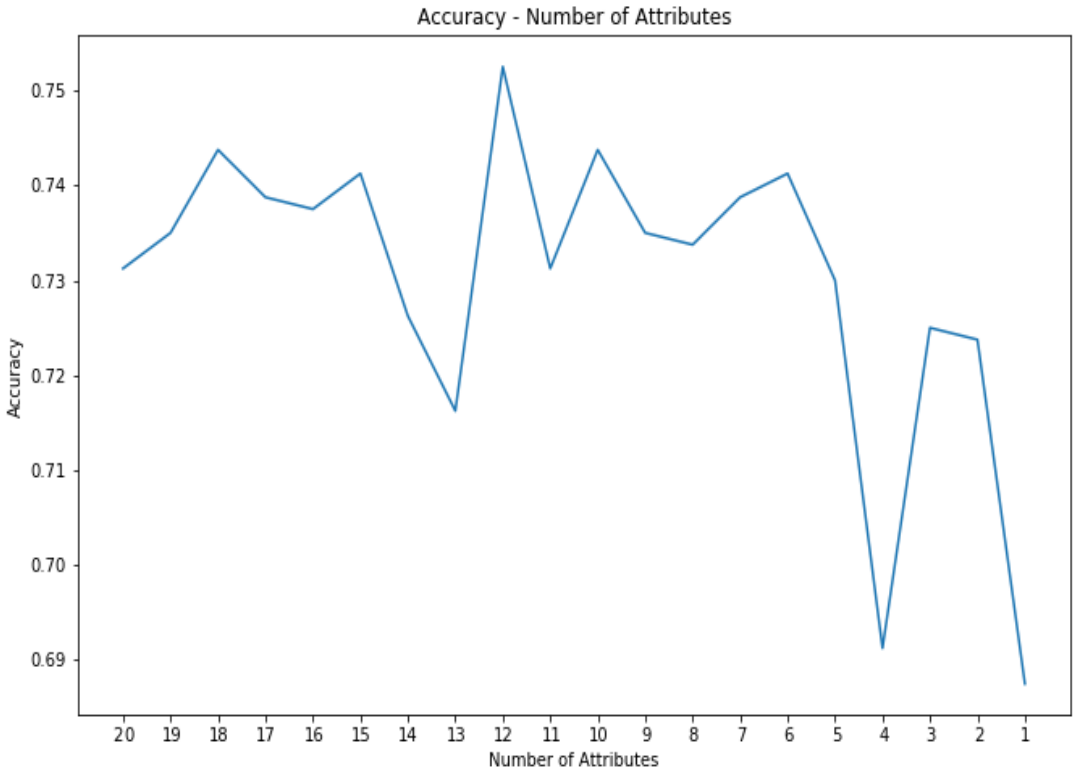
Στην προ επεξεργασία των δεδομένων για τα categorical χρησιμοποιήσαμε τον Label Encoder ενώ για τα numerical τα μετατρέψαμε σε categorical με την χρήση της συνάρτησης pandas cut, χωρισμένα σε 5 κατηγορίες (bins). Για τα numerical δεδομένα που έχουν λιγότερες από 5 διακριτές τιμές, τα αφήσαμε ως έχουν αφού δεν χρειάζονται κάποια άλλη επεξεργασία.

Στην συνέχεια υπολογίσαμε το information gain για κάθε attribute ξεχωριστά και τα αποθηκεύσαμε σε ένα dictionary (οπού keys είναι ο αριθμός του attribute και values το information gain), το οποίο ταξινομήσαμε με βάση τα values από το μικρότερο στο μεγαλύτερο.

Έπειτα εφαρμόσαμε 10-Fold Cross Validation αφαιρώντας κάθε φορά το attribute με το μικρότερο information gain.

Loop	Removed Attribute	Information Gain
1	Attribute 18	0.000129665701928
2	Attribute 11	0.000220571349274
3	Attribute 19	0.00120286259108
4	Attribute 16	0.00239577011259
5	Attribute 17	0.00294031663129
6	Attribute 10	0.00567439979016
7	Attribute 14	0.00704150632514
8	Attribute 8	0.00733050007683
9	Attribute 20	0.00770438654644
10	Attribute 15	0.0116188868237
11	Attribute 9	0.0127468411562
12	Attribute 13	0.0134129805330
13	Attribute 7	0.0145478652302
14	Attribute 12	0.0149055308773
15	Attribute 5	0.0184611461323
16	Attribute 6	0.0221989660524
17	Attribute 4	0.0268974520331
18	Attribute 2	0.0329634294231
19	Attribute 3	0.0378894062215
20	Attribute 1	0.0938279630235

Παρακάτω ακολουθεί το διάγραμμα μεταβολής του μέσου accuracy καθώς αφαιρούμε attributes από τον Random Forest classifier. Παρατηρούμε ότι το μέγιστο accuracy το έχουμε στα 12 attributes.



Στον παραδοτέο φάκελο περιέχονται επίσης τα παρακάτω πηγαία αρχεία που δίνουν όλα τα παραπάνω αποτελέσματα.

- info_gain.py :
- metrics.py :
- plots.py :
- predict_test_set.py :