Τεχνικες Εξορυξης Δεδομενων

2η Ασκηση

Φοιτητες:

Ονομα: **ΓΚΑΡΑΓΚΑΝΗΣ ΕΥΑΓΓΕΛΟΣ**

Α.Μ. : **1115201400033**

Ονομα: **ΚΩΣΤΑΣ ΚΟΤΡΩΝΗΣ**

Α.Μ. : **1115201400074**

Το καθε ερωτημα και ζητουμενο της ασκησης υλοποιειται σε 4 αρχεια πηγαιου κωδικα σε python 3.5.

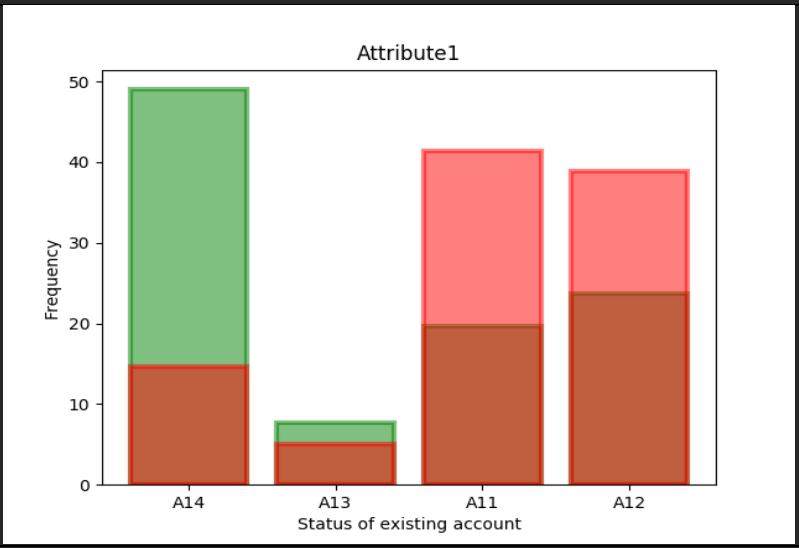
1. Οπτικοποιηση Δεδομενων:
   DataVisualization.py
2. Υλοποιηση Κατηγοριοποιησης:
   Classification.py
3. Επιλογη features (Χαρακτηριστικων):
   FeatureSelection.py
4. TestSet_Predictions:
   Predict.py

Στα αρχεια αυτα υπαρχουν καθοδηγητικα σχολια που εξηγουν την δομη και τη λειτουργια καθε βηματος για την υλοποιηση των ερωτηματων.

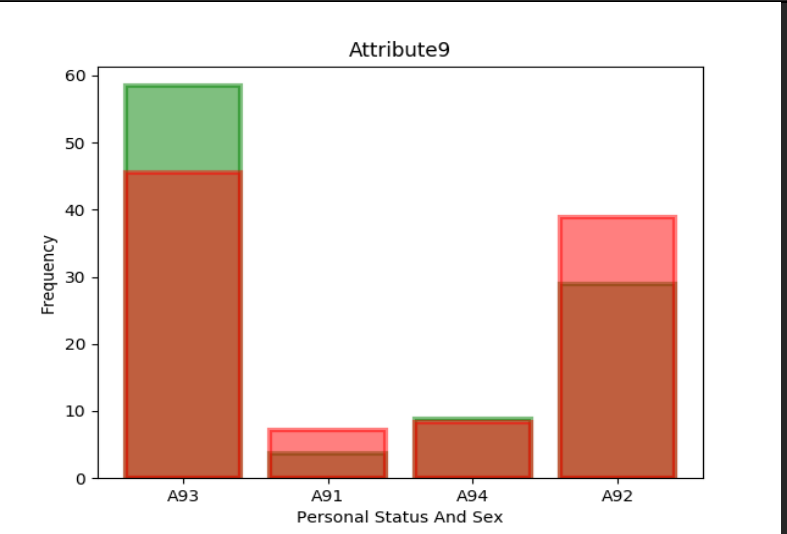
Οπτικοποιηση Δεδομενων (Data Visualization)

Σε αυτο το ερωτημα χρησιμοποιηθηκαν bar plots για την οπτικοποιηση των **categorical** δεδομενων και box plots για **numerical** . Πιο συγκεκριμενα , χρησιμοποηθηκε η βιβλιοθηκη matplotlib , με την οποια κατασκευαστηκαν τα εξης ειδους σχεδιαγραμματα :

Παραδειγματα bar blots



Παραδειγμα histogram , που απεικονιζει το πρωτο feature. Καθε κατηγορια , για την κατασταση του τρεχοντος λογαριασμου , κωδικοποιειται οπως μας δινεται απο την ασκηση , σε μια κωδικη ονομασια [A14 , A13 , A11 , A12] , το καθενα με την δικη του σημασια.

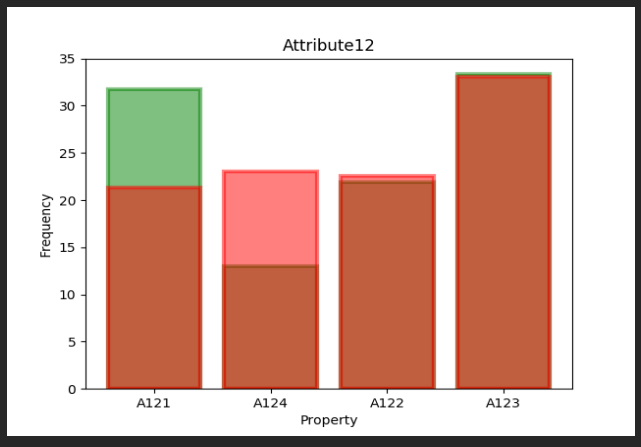


Αντιστοιχα και εδω.
Εξηγουμε:
Attribute 9: (qualitative)
Personal status and sex
A91 : male : divorced/separated
A92 : female : divorced/separated/married
A93 : male : single
A94 : male : married/widowed
A95 : female : single

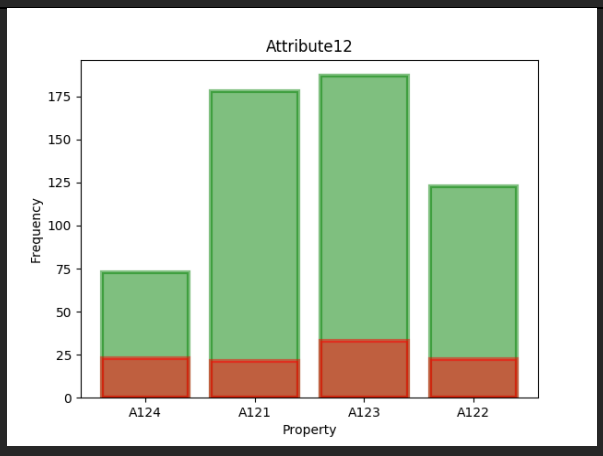
Επεξηγηση σχετικα με την σχεδιαστικη επιλογη των bar plot και του τροπου με τον οποιο αντλουμε πληροφοριες απο αυτα :

Στο επανω μερος των plot μας , αναγραφεται ο τιτλος του features που οπτικοποιειται . Αριστερα , στον αξονα y , εχουμε την συχνοτητα την οποια παρουσιαζει η καθε κατηγορια για καθε ειδος πελατη. Στο αξονα x, εχουμε την κατηγορια καθε κατηγορηματικου feature , ενω απο κατω επεξηγηται τι ακριβως συμβολιζει αυτο feature. Ως προς τα plots , με πρασινο χρωμα παριστανονται οι καλοι πελατες , ενω με κοκκινο οι κακοι. Το καφε χρωμα δημιουργειται λογω του οτι τα bars , πεφτουν το ενα πανω στο αλλο , το οποια επελεξαμε για λογους ευδιακριτης συγκρισης.

Τα bars δημιουργηκαν με βαση την παρακατω λογικη. Χωρισαμε το train data set σε καλους και κακους πελατες. (~561 good , 239 bad). Στην συνεχεια για το καθε ειδος συνολου πελατων , υπολογισαμε την συχνοτητα της καθε κατηγοριας του feature σε αυτην. Στην αρχη , επιλεξαμε να προβαλουμε αυτα τα αποτελεσματα αυτουσια στα γραφημαρα μας , δηλαδη το υψος καθε κατηγοριας να αγγιζει τον ακριβη αριθμο ατομων που ανηκουν σε αυτη. Αυτο ομως , δεν μας παρεδιδε ξεκαθαρα καποια εμφανης πληροφορια , αφου οι πλειοψηφια των πελατων μας ειναι good . Ετσι επιλεξαμε , να υπολογιζουμε το ποσοστο τις 100 , των πελατων που ανηκουν σε καθε κατηγορια , για το καθε συνολο πελατων , προκειμενου να εχουμε πιο αντιπροσωπευτικα αποτελεσματα. Συγκριση των bar plot , πριν και μετα την μετατροπη %.

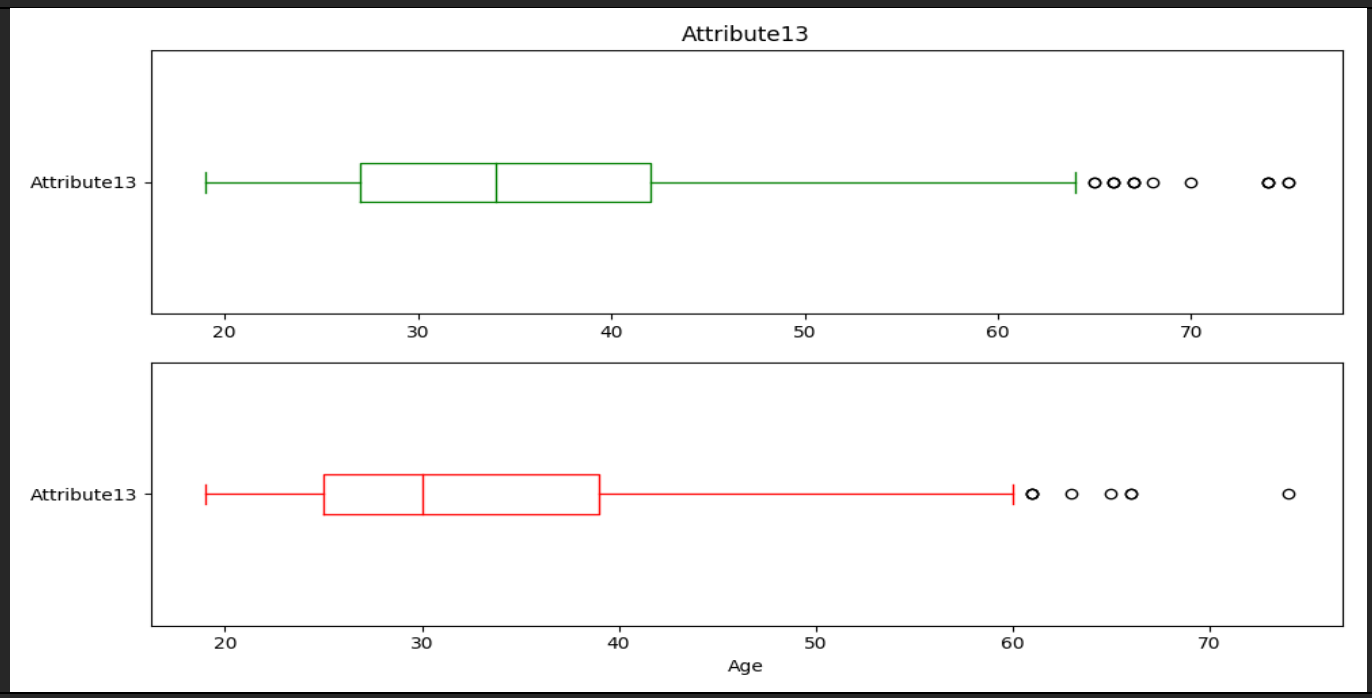


Μετα την μετατροπη

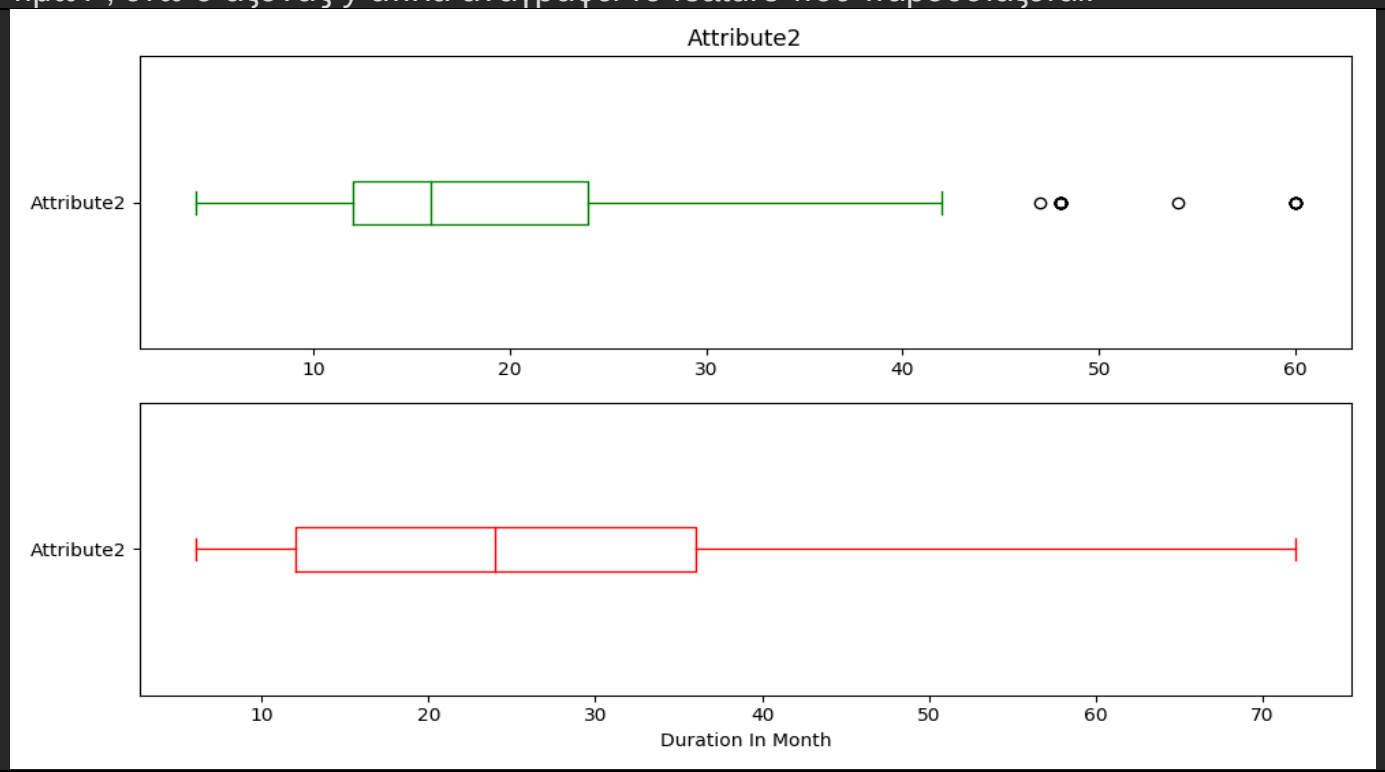


Πριν την μετατροπη

Αντιστοιχη επεξηγηση των box plots



Τα box plots παρουσιαζουν αντιστοιχα , ακολουθωντας τον ιδιο διαχωρισμο των δεδομενων , την αντιστοιχη συγκριση των χαρακτηριστικων μεταξυ των καλων και των κακων πελατων , αυτη την φορα για τα νουμερικα χαρακτηριστικα των πελατων.
Εδω , δεν εχουμε % αναπαρασταση των δεδομενων , αφου τα box bots , μας δειχνουν που υπαρχει η μεγαλυτερη συγκεντρωση ατομων . Ο αξονας x περιεχει το ευρος των τιμων , ενω ο αξονας y απλα αναγραφει το feature που παρουσιαζεται.



Παραδειγματα box plots

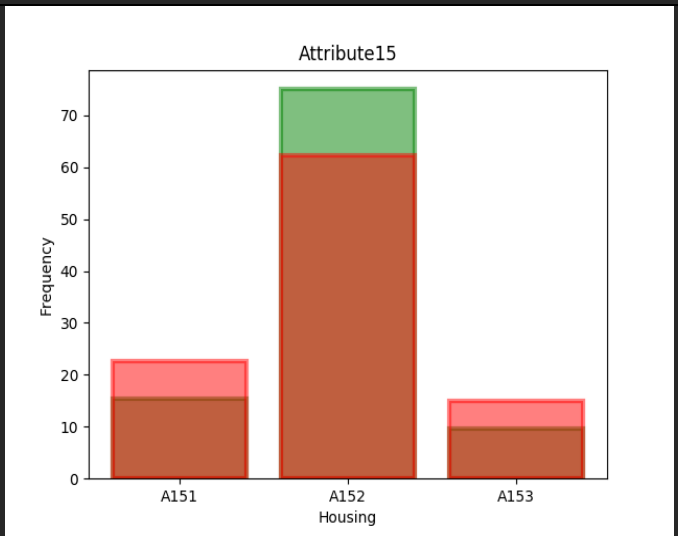
Παρατηρησεις / Προβλεψεις πανω στα plots

Τα δεδομενα που οπτικοποιηθηκαν επαληθευσαν , για το μεγαλυτερο ποσοστο , τις προβλεψεις μας , καθως οι κατηγοριες που περιμεναμε να ειναι πιο σημαντικες και στις οποιες ηταν αναμενομενο να δουμε διαφορες αναμεσα σετους καλους και κακους πελατες , πραγματι ειχαν αντιστοιχα γραφηματα. Τα features ανταποκρινονται στην πραγματικοτητα , εξαιρουμενου καποιον εκπληξεων , και παρουσιαζουν μια πολυ καλη εικονα για το ποια χαρακτηριστικα διακρινουν εναν καλο και ενα κακο πελατη.

Πιο συγκεκριμενα , μελετηθηκαν ολα τα features ξεχωριστα , τοσο ως προς τις δικες μας προβλεψεις και εκτιμησεις , οσο και ως προς τις παρατησεις και τα συμπερασματα που απορρεουν απο τα γραφηματα. Ετσι διακρινουμε τα features αναλογα με την σημασια τους για το ποιον ενος πελατη και για το ποσο αναμενομενη ηταν η προβλεψη μας. Τα παρουσιαζουμε παρακατω :

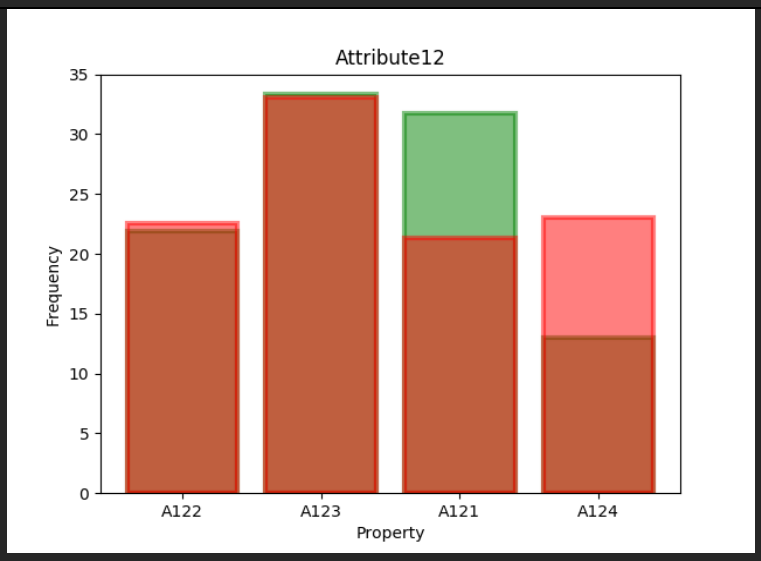
- **Σημαντικα features :**

Attribute 15 , Housing :



Αναμενομενο . Ειναι ακρως λογικο μια τραπεζα να θεωρει την κατοχη σπιτιου καλο στοιχειο για ενα πελατη , λογω της σταθεροτητας ενεπνεει για την ζωη του και για την παροχη εγγυησεων που ισως τις παρεχει , εναντι ενος πελατη που δεν εχει δικο του σπιτι ή μενει καπου δωρεαν. Αυτο το αποδυκνυει και το αντιστοιχο ιστογραμμα για το συγκεκριμενο feature.

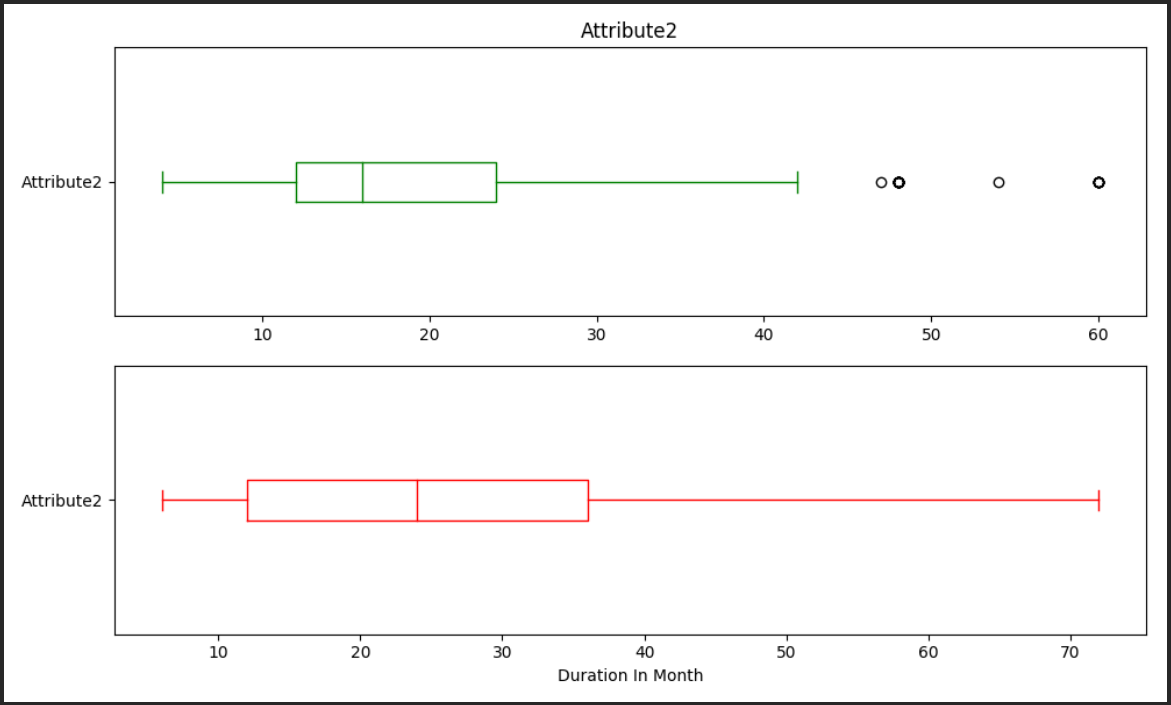
Attribute 12 , Property :

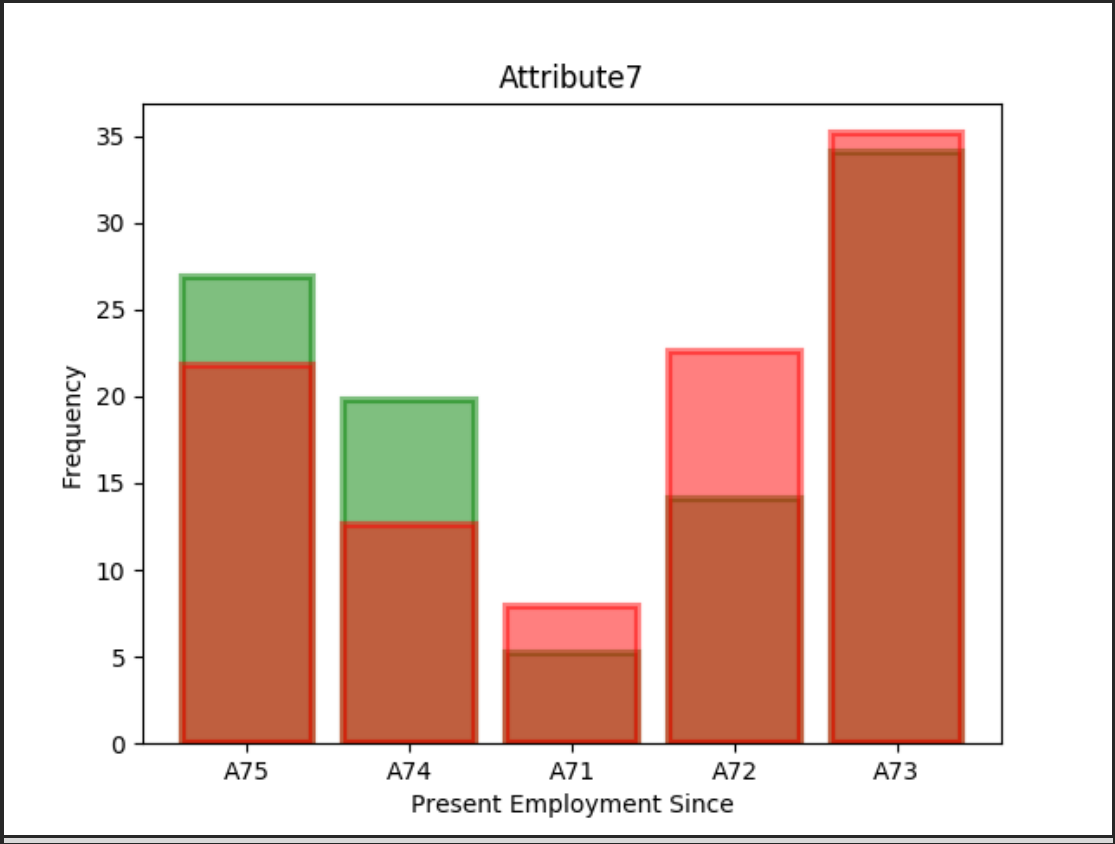


Ανεμενομενο και επισης λογικο. Πελατες οι οποιοι φροντιζουν για το μελλον τους με ασφαλειες ζωης και οικονομιες για κοινωνικες αναγκες , ειναι πιο ωριμοι ως προς την διαχειριση των χρηματων του , οποτε οι τραπεζα τους θεωρει καλους. Αντιθετως , με τους κακους πελατες , που χρησιμοποιουν τις αναγκες για υλιστικες αγορες

Attribute 2:Duration in month, Attribute 7: Employment Since:

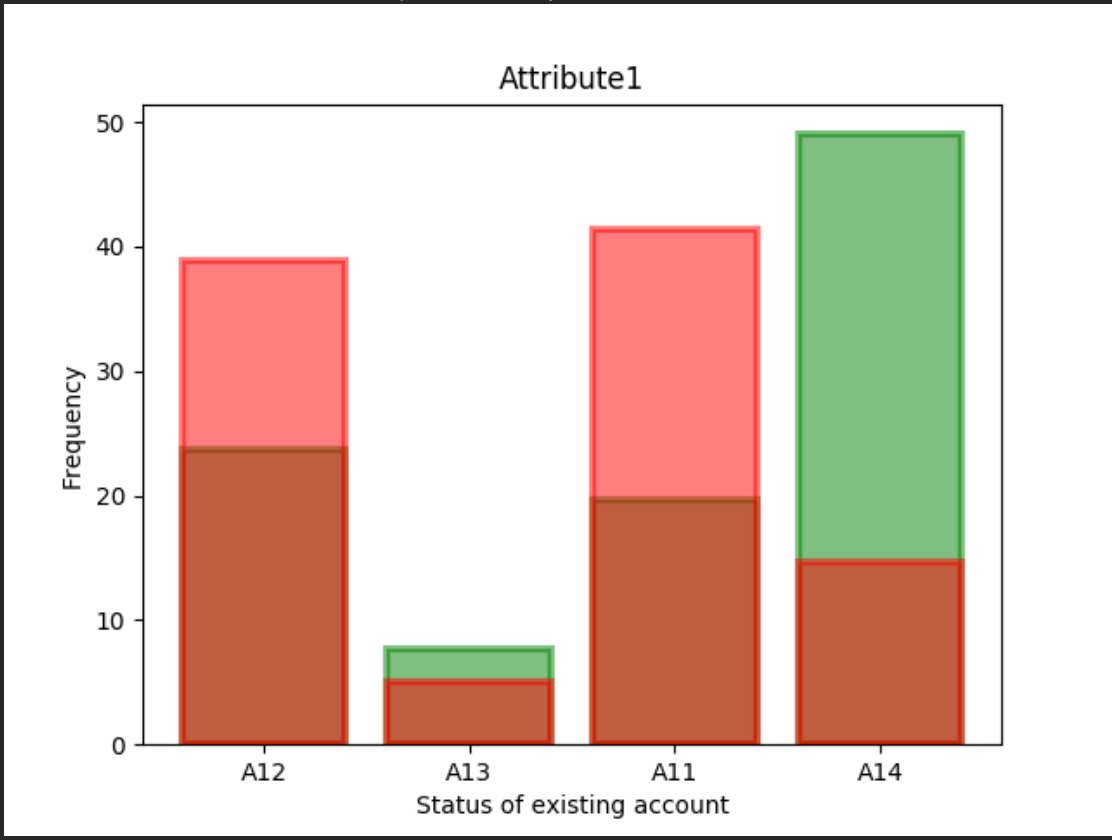
Τα χαρακτηριστικα σταθεροτητα και ωριμοτητα συνεχιζουν να διακρινουν τους καλους πελατες . Οπως βλεπουμε απο τα παρακατω γραφηματα :





Συνεπεις πελατες με μικρα δανεια , σταθεροι ως προς την ζωη τους αφου δουλεουν πανω απο 4 χρονια , ειναι λογικο να ειναι σημαντικοι.

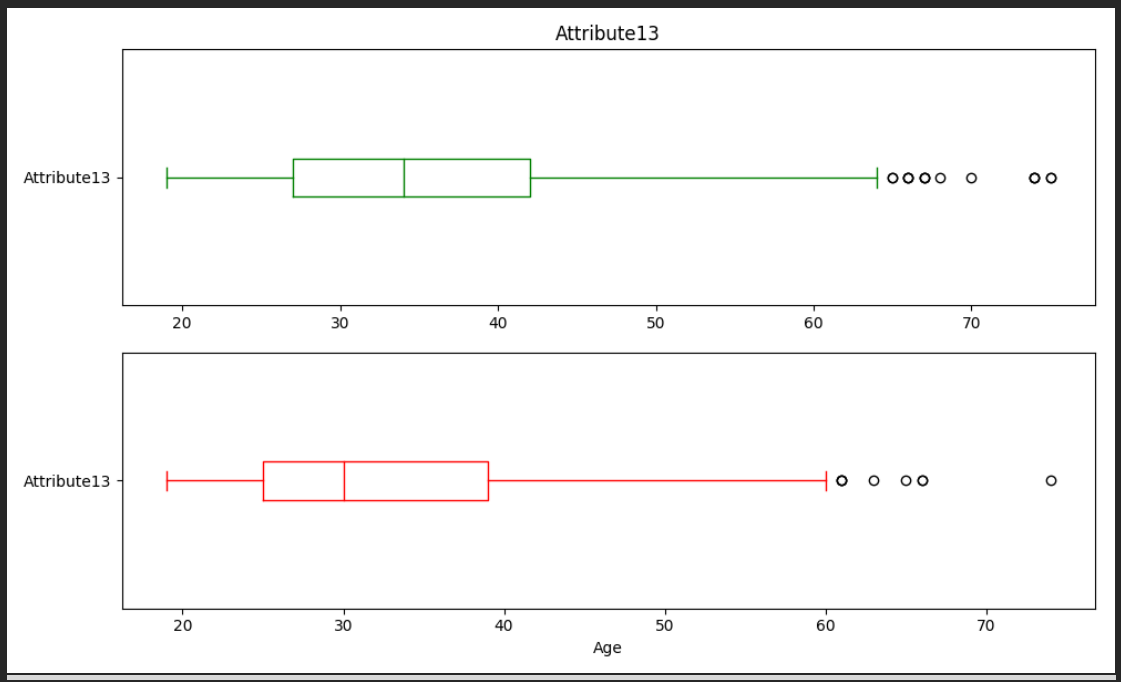
Αντιστοιχη λογικη ακολουθουν και αλλα features οπως ειναι η κατασταση του τρεχοντος λογαριασμου (Att 6), εαν χρωστουν στην τραπεζα ή οχι (**Attribute 1 :**

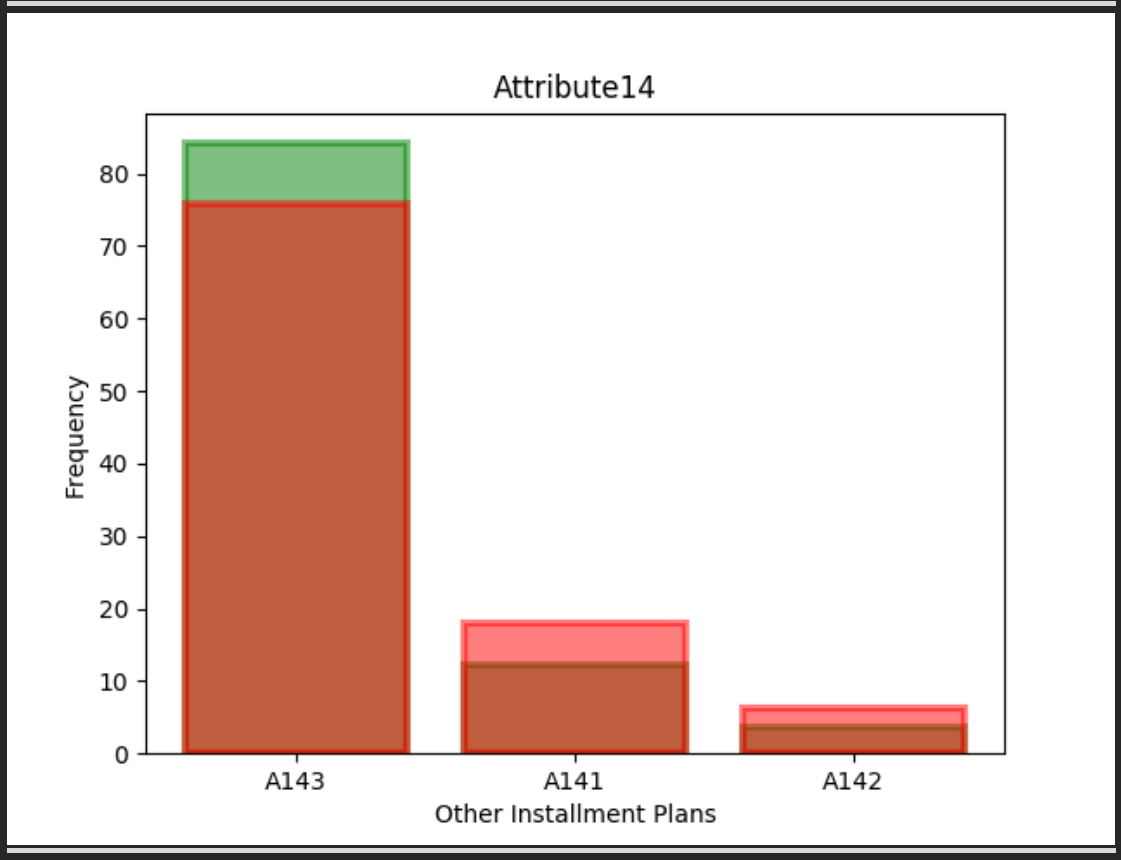


) , των οποιων τα γραφηματα μιλουν απο μονα τους.

- **Μετρια features :**

Θα αναφερουμε να χαρακτηριστικα που εχουν μετρια σημασια. Attr 14 , τροπος εξοφλησης , Attr 13 η ηλικια , Attr 15 τα χρηματα στην τραπεζα , Attr 4 , ο σκοπος . Ολα αυτα ειναι πολυ σχετικες ιδιοτητες που με την λογικη και μονο δεν εχουν καποια αμεση σχεση με το αν εισαι καλος πελατης ή οχι. Παρατηρουμε οτι πελατες με μεγαλυτερη ηλικια , με μεσαια εισοδηματα στην τραπεζα , η δουλεια τους (Atrr 17) , δεν επηρεαζει αμεσα την κατασταση. Αυτο φαινεται και στα διαγραμματα , με την μετρια διαφορα τους .

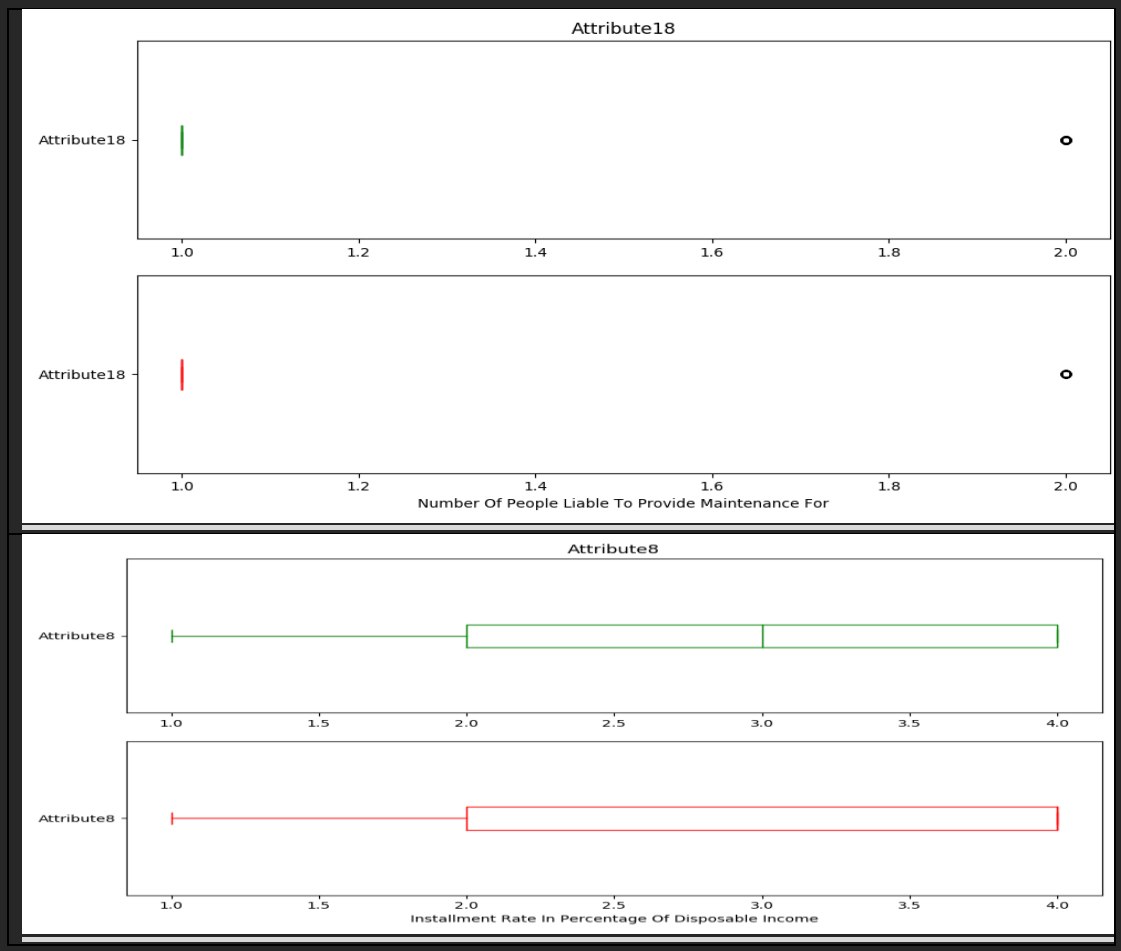




- **Ασημαντα features :**

Ασημαντα ειναι τα features τα οποια εχουν απειροελαχιστη διαφορα στα plot τους , πραγμα που δειχνει

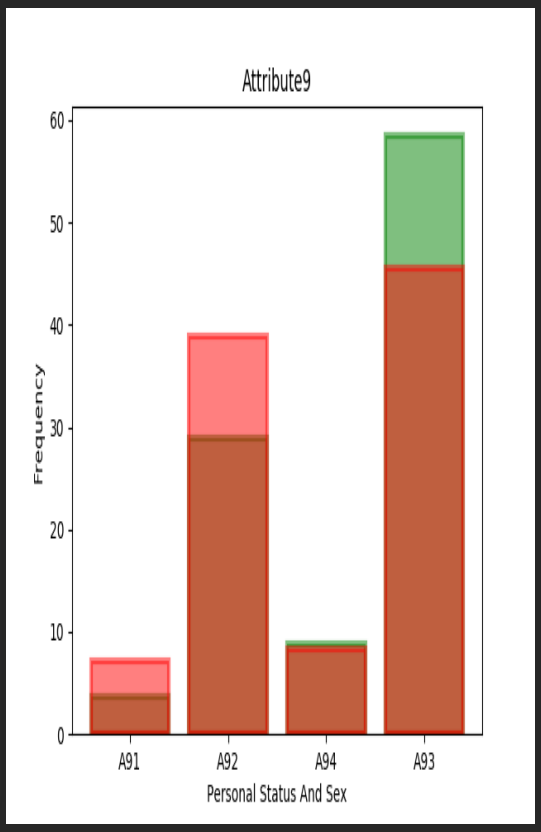
οτι οι πελατες των συγκεκριμενων feature δεν παρουσιαζουν καποια διαφορα που να παιζει ρολο για το αν ειναι καλοι πελατες ή οχι. Τετοια features ειναι Attr 19 , Attr 18 , Attr 8 , Attr 20 κτλπ. Δηλαδη η καταγωγη του πελατη , το αν εχει τηλεφωνο ή οχι , το ποσο λογριασμους εχει στην τραπεζα , δηλαδη αυτα τα feature που απλα ειναι θεμα επιλογης / τυχαιοτητας για τον πελατη, δεν εχουν τοσο σημασια. Χαρακτηριστικα ειναι τα γραφηματα των Attributes 18 και 8 , με τα γραφηματα τους να μην παρεχουν καποι προφανη πληροφορια ή καποιο πορισμα.



Συμπεραινουμε λοιπον , πως τα στοιχεια που ειναι σημαντικα για την διακριση ενος πελατη σε καλο ή κακο , ειναι αυτα που μας λεει και η λογικη. Αυτα ειναι η ωριμοτητα , η

συνεπεια του , το ποσο φροντιζει για το μελλον του , η σταθεροτηα της ζωης του , η επιλογες του κτλπ. Στοιχεια οπως η τρεχον κατασταση του πελατη, οπως ειναι το αν εχει τηλεφωνο ή οχι , τα λεφτα που εχει στην τραπεζα , το ποσο χρονων ειναι , δεν παιζουν τοσο σημαντικο ρολο. Ασημαντα χαρακτηριστικα θεωρουντα αυτα για τα οποια δεν ειναι θεμα επιλογης του πελατη, οπως η καταγογη , και αυτα που δεν εχουν τοσο σημαντικο ρολο.

Fun Fact:



Οπου A93 = Single

Υλοποιηση Κατηγοριοποιησης (Classification)

Για την υλοποιηση της κατηγοριοποιησης χρησιμοποιηθηκε η scikit με ηδη γνωστους αλγοριθμους , που χρησιμοποιηθηκαν και σχολιαστηκαν εκτενως στην προηγουμενη ασκηση . Σε αυτο το ερωτημα , χρειαστηκε αναλογη προεπεξεργασια των δεδομενων , προκειμενου να μετατραπουν ολα τα κατηγορηματικα feature σε νουμερικα .

(Η προεπεξεργασια και μετατροπη αυτων των δεδομενων εγινε με την βοηθεια του dummy coding . το οποιο χρησιμοποιείται συνήθως για τη μετατροπή μιας μεταβλητής κατηγορικής εισόδου σε συνεχή μεταβλητή. Το "Dummy", όπως υποδηλώνει το όνομα, είναι μια διπλή μεταβλητή που αντιπροσωπεύει ένα επίπεδο μιας κατηγορικής μεταβλητής. Η παρουσία ενός επιπέδου είναι αντιπροσωπευτική κατά 1 και η απουσία αντιπροσωπεύεται από το 0. Για κάθε υπάρχον επίπεδο, θα δημιουργηθεί μια ψευδομεταβλητή.)

. Η μεθοδοι classification που προτιμηθηκαν ειναι οι Random Forest , Naïve Bayes και SVM , πιο συγκεκριμενα ο linear svm λογο της καλυτερης αποδοσης του. Οι αλγοριθμοι επαληθευτηκαν με 10 fold cross validation , αμε αντιστοιχες συναρτησεις που παρεχονται απο την sci-kit.

Οι παρακατω πληροφοριες βασιστηκαν πανω στις εξης πηγες :

- https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling/
- https://www.quora.com/Which-algorithm-fits-best-for-categorical-and-continuous-independent-variables-with-categorical-response-in-Machine-Learning
- https://users.cs.duke.edu/~rvt/msthesis.pdf

Ακολουθουν τα αποτελεσματα του evaluation metric μας:

	A	B	C	D	E
1	Statistic Measure	Naive Bayes	Random Forest	SVM	
2	Accuracy	0.61375	0.7287500000000001	0.6637500000000001	
3					

Static Measure	Naïve Bayes	Random Forest	SVM
Accuracy	0.61375	0.72875	0.66375

Κατοπιν πολλων προσπαθειων και δοκιμων αλγοριθμων κατηγοριοποισης πανω στα δεδομενα , οπως παρατηρουμε και στο αρχειο εξοδου , με εμφανη διαφορα καλυτερος ειναι ο Random Forest Classifier. Η διαφορα αυτη , στην ικανοποιητικη αποδοση του αποδιδεται στα δεντρα αποφασης που δημιουργει ο random forest , πραγμα που καθοδηγει τον αλγοριθμο σε μονοπατια μεγαλυτερης αποκτησης πληροφοριας οπως θα δουμε και παρακατω. Παρα το γεγονος οτι ο Random Forest ειναι ισως 'κουραστικος' για την μηχανη στην οποια εκτελειται , αφου αυξανουν αρκετα την υπολογιστη πολυπλοκοτητα προκειμενου να δημιουργησουν τα δεντρα , οπως αναφερεται και σε αντιστοιχο αρθρο του Wikipedia:

"For data including categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels"

Οσον αναφορα τους υπολοιπους αλγοριθμους φαινεται επισης να ειναι σχετικα ικονοποιτικοι , και δουλεουν αρκετα καλα με τα categorical data , οσο και με τα numerical.

Τα αποτελεσματα των μετρησεων μας ανταποκρινονται στην πραγματικοτητα και στο κατα ποσο ευκολο ειναι να βγαλουμε συμπερασματα πανω στο συγκεκριμενο data set που μας δινεται . Οπως παρατηρησαμε στη παραπανω γενικη ιδεα που μας προσεφερε η οπτικοποιηση των δεδομενων , και οπως επισης θα καταλαβουμε στην συνεχεια απο το feature selection και το ποσο πληροφορια εκλαμβανουμε απο τα feature μας , καταλαβαινουμε οτι η κατηγοριοποιηση δεν ειναι αρκετα ευκολη. Αυτο συμβαινει λογω των χαμηλων συγκρισεων που μας παρεχει το data set , αφου οι κακοι και οι καλοι πελατες δεν εχουν παντα εμφανης διαφορες και δεν ακολουθουν καποια πολυ συγκεκριμενα μοτιβα και συμπεριφορες. Αυτο αποδεικνυεται και παρακατω , με τον υπολογισμο του information gain του καθε feature.

Τα παραπανω , περα απο προσωπικες μας λογικες εκτιμησεις και υπολογισμους , φαινεται να αποδεικνυονται και παρακατω:

What is Imbalanced Data?

Imbalanced data typically refers to a problem with classification problems where the classes are not represented equally. For example, you may have a 2-class (binary) classification problem with 100 instances (rows). A total of 80 instances are labeled with Class-1 and the remaining 20 instances are labeled with Class-2. This is an imbalanced dataset and the ratio of Class-1 to Class-2 instances is 80:20 or more concisely 4:1.

Το data set μας , με ratio περιπου 1:3 , εκφραζει μια σχετικη ασταθια , η οποια εκδηλωνεται τοσο στα plots μας , οσο και στις μεθοδους classification και το information gain που μας παρεχουν

Οι αλγοριθμοι που χρησιμοποιουμε και ειδικα ο random forest , θεωρειται πως εχει υψηλα επιπεδα ευστοχιας για το συγκεκριμενο data set , αφου μονιμα μας παρεχει ενα 70% ευστοχιας. Πηγη: http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

Επιλογη Feature (Feature Selection)

Για την επιλογη των feature , το ερωτημα αναλυθηκε σε 2 σταδια και συνδυαστηκε στην συνεχεια ,οποτε θα εξηγησουμε το καθενα τοσο ξεχωριστα , οσο και συνδυαστικα στην συνεχεια.

- Υπολογισμος Information Gain για καθε feature:

Επειδη το ερωτημα αυτο υλοποιηθηκε κατοπιν εκτενης ερευνας και κατασκευαστηκε απο εμας , με δικες μας υλοποιησεις , θα ξεκινησουμε με τον ορισμο του information gain.

This measure of *purity* is called the **information**. It represents the expected amount of information that would be needed to specify a prediction.

Entropy on the other hand is a measure of *impurity* (the opposite).

Η αποκτηση της πληροφοριας ειναι αλληλενδετη με την εντροπια , οπως εξηγηται και παραπανω. Ο μαθηματικος τυπος της εντροπιας ειναι ο παρακατω :

$$H(S) = -\sum_{x \in X} p(x) \log_2 p(x)$$

Where,

- S - The current (data) set for which entropy is being calculated (changes every iteration of the ID3 algorithm)
- X - Set of classes in S
- $p(x)$ - The proportion of the number of elements in class x to the number of elements in set S

Η συνδεση της εντροπιας με την πληροφορια εξηγηται στην παρακατω συλλογιστικη πορεια.

- is a measure of the amount of uncertainty in the (data) set S (i.e. entropy characterizes the (data) set S).
- in other words, it is the average amount of information contained in each message received (message here stands for an event, sample or character drawn from a distribution or data stream)
- it characterizes the uncertainty about our source of information (Entropy is best understood as a measure of uncertainty rather than certainty, as entropy is larger for more random sources)
- a data source is also characterized by the probability distribution of the samples drawn from it (the less likely an event is, the more information it provides when it occurs)
- it makes sense to define information as the negative of the logarithm of the probability distribution (the probability distribution of the events, coupled with the information amount of every event, forms a random variable whose average (expected) value is the average amount of information (entropy) generated by this distribution).

Με βαση αυτη την σχεση , και μετα απο αρκετες προσαρμογες , δημιουργησαμε την δικη μας συναρτηση σε γλωσσα python η οποια υπολογιζει την εντροπια ενος attribute σε ενα συγκεκριμενο dataframe. Ο κωδικας μαζι με την επεξηγηση του υπαρχει στα αρχεια παραδοσης (συγκεκριμενο στο InformationGain.py).

Συνεχιζοντας , το information gain:

$$IG(A,S) = H(S) - \sum_{t \in T} p(t)H(t)$$

Where,

- $H(S)$ - Entropy of set S
- T - The subsets created from splitting set S by attribute A such that $S = \bigcup_{t \in T} t$
- $p(t)$ - The proportion of the number of elements in t to the number of elements in set S
- $H(t)$ - Entropy of subset t

- is the measure of the difference in entropy from before to after the data set S is split on an attribute A
- in other words, how much uncertainty in S was reduced after splitting data set S on attribute A
- it is a synonym for Kullback–Leibler divergence (in the context of decision trees, the term is sometimes used synonymously with mutual information, which is the expectation value of the Kullback–Leibler divergence of a conditional probability distribution. The expected value of the information gain is the mutual information I(X; A) of X and A – i.e. the reduction in the entropy of X achieved by learning the state of the random variable A. In machine learning, this concept is used to define a preferred sequence of attributes to investigate to most rapidly narrow down the state of X. Such a sequence (which depends on the outcome of the investigation of previous attributes at each stage) is called a decision tree. Usually an attribute with high mutual information should be preferred to other attributes).

Παρομοια , δημιουργησαμε την αντιστοιχη συναρτηση information gain , βασιζομενοι σε αυτες τις πληφοριες.

(πηγες: https://stackoverflow.com/questions/1859554/what-is-entropy-and-information-gain

http://christianherta.de/lehre/dataScience/machineLearning/decision-trees.php)

Το information gain για καθε feature ειναι το παρακατω :

Το Information gain καθε Attribute αντιστοιχει τοσο στις προβλεψεις και εκτιμησεις που καναμε με βαση την λογικη μας , οσο και στα διαγραμματα τα οποια εξηγησαμε.

Χαρακτηρικα βλεπουμε το πως τα features:

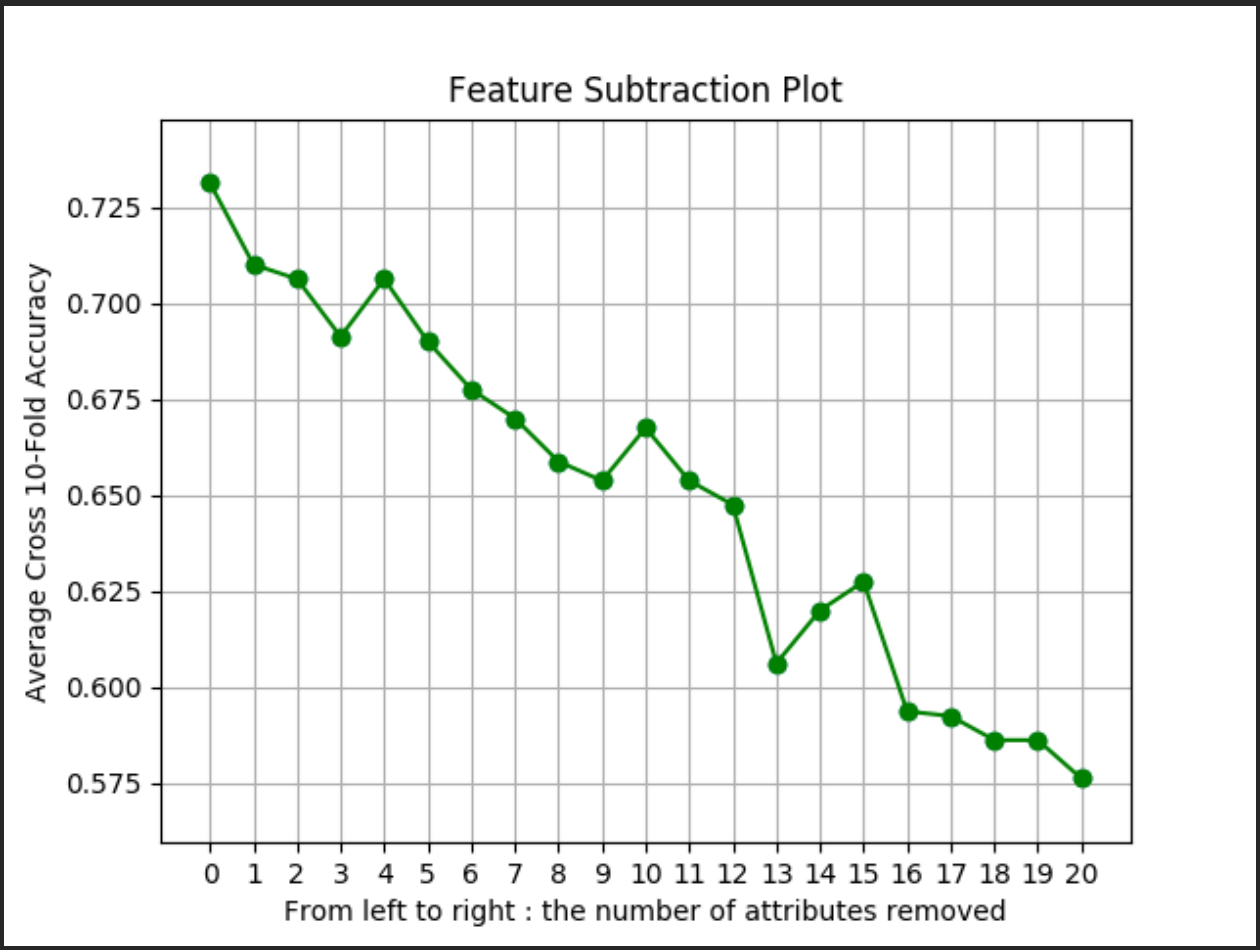
Attribute 1 , 7 , 2 , 12 ,15 μας παρεχουν υψηλα επιπεδα information gain. Αυτο δικαιολει και δικαιολογειται απο τις παρατηρησεις που καναμε . Αντιστοιχα , μεταβλητες που εκτιμησαμε να εχουν μικρη σημασια , οπως ειναι οι Attributes 8 και 18 , εχουν πολυ μικρο IG. Το ιδιο συμβαινει και για αυτες με μεσαια σημασια.

- Σταδιακη αφαιρεση features:

Διαλεξαμε τον random forest , ως καλυτερος classifier απο προηγουμενο ερωτημα , και εκτελεσαμε , ακριβως οπως εξηγησαμε παραπανω ,Classification καθε φορα αφαιρωντας ενα feature.

Παρουσιαση Plot / Πινακα σχεσης μεταξυ IG και Accuracy του Classifier:

Στο παρακατω plot παρουσιαζω το πως αλλαζει το μεσο accuracy για 10 fold cross validation , καθως αφαιρω feature απο τον classifier



Attribute 1	0.71	0.93
Attribute 2	0.70625	0.03179
Attribute3	0.69125	0.03789
Attribute4	0.70625	0.0269

Attribute5	0.69	0.0153
Attribute6	0.6775	0.0222
Attribute7	0.67	0.01455
Attribute8	0.65875	0.00734
Attribute9	0.65375	0.01275
Attribute10	0.6675	0.00568
Attribute11	0.65375	0.00023
Attribute12	0.65375	0.01491
Attribute13	0.6475	0.01175
Attribute14	0.60625	0.00705
Attribute15	0.62	0.01162
Attribute16	0.62751	0.0024
Attribute17	0.59375	0.00295
Attribute18	0.5925	0.00013
Attribute19	0.58625	0.00121
Attribute20	0.57626	0.00771

Παρουσιαζεται ο πινακας με το καθε feature

Που αφαιρουμε και το αντιστοιχο information gain του

Τελος , παρουσιαζουμε σε ενα plot , το πως το μεσο accuracy με cross fold 10 validation , του random forest classifier , φθηνει καθε φορα που αφαιρουμε και ενα attribute. Αυτο συμβαινει , γιατι καθε φορα χαινεται «υλικο» με το οποιο ο κατηγοριοποιητης μας κανεις τις προβλεψεις του. Παρατηρουμε οτι οι πτωσεις απο αφαιρεση σε αφαιρεση , ειναι αναλογες του information gain που παρεχουμε.

Τελος , παρουσιαζουμε το Prediction του random forest , πανω στο test set:

Client_ID	Predicted_Label
11100	1
11099	2
11098	1
11097	1
11096	1
11095	1
11094	2
11093	1
11092	1
11091	1
11090	1
11089	1
11088	1
11087	1
11086	2
11085	1
11084	1
11083	2
11082	2
11081	1
11080	1
11079	1
11078	1
11077	1
11076	1
11075	1
11074	2
11073	2

Συνοψιζοντας

Εχοντας ολοκληρησει την εργασια , εχουμε ασχοληθει με την εξερευνηση των δεδομενων μιας τραπεζας , την αξιολογηση των χαρακτηριστικων τους και πως αυτα μπορουν να εκμεταλευτουν για το δικος της κερδος.

Απο ολη αυτη την εξερευνηση και μελετη των δεδομενων(γραφικα , αλγοριθμικα , μαθηματικα) μας κανει εντυπωση η ανταποκριση των αποτελεσματων στην πραγματικοτητα και πηραμε μια εικονα εις βαθος για τις πτυχες και το βαθος στο οποιο υπαρχουν δεδομενα.

Στην προηγουμενη ασκηση ειδαμε το πως οι αλγοριθμοι βοηθουν εμας , για την εξορυξη των δεδομενων , σε αυτην , το πως εμεις μπορουμε να τους βοηθησουμε.