Report

Khasay Mirzali

May 2021

Abstract

In the report, I will demonstrate classification,clustering and frequent pattern mining results in customer credit risk dataset.

Keywords: Classification Cluster Apriori.

1 Introduction

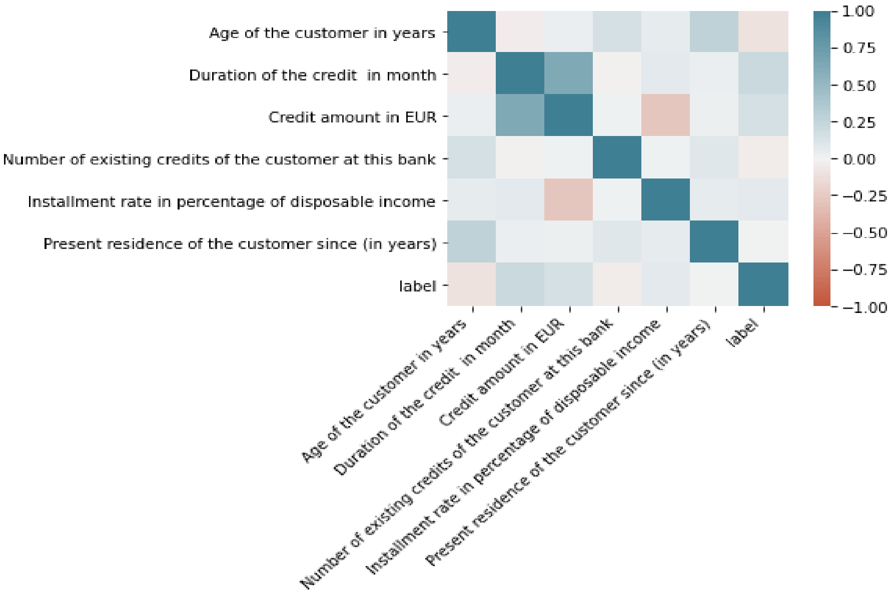
Data is all around us. Most companies nowadays use their data to make future decisions. One of the fields is the banking industry. Banks use data analysis for better customer segmentation and their classification. Because with classification they can learn which customer is potentially bad or good one. In this paper, we discuss the important algorithms for clustering,classification and pattern mining.

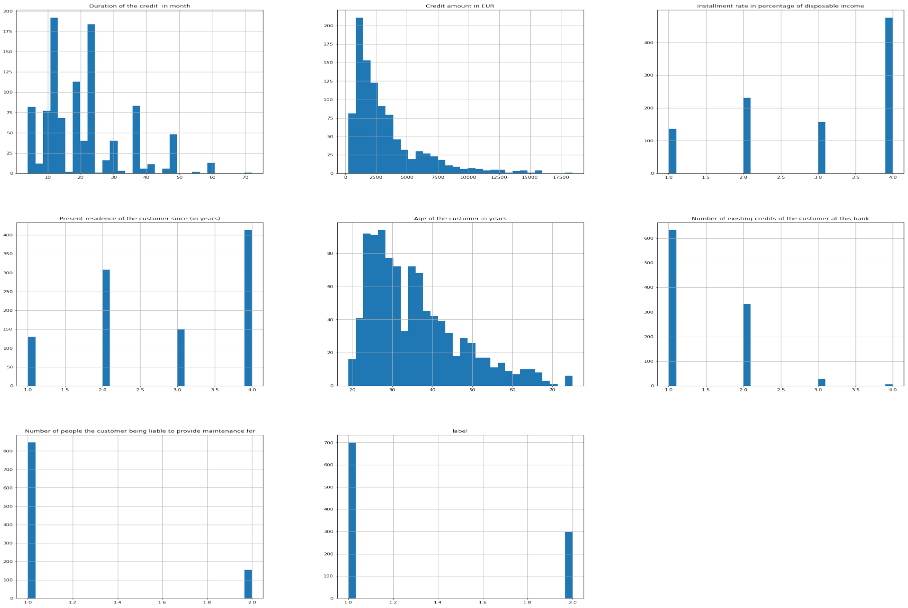
2 Data set information

Given data set is about Customer Credit risk. Data set contains 20 columns and 1000 rows. There are 7 numeric and 14 categorical variables.. 700 of them are classified as good and remaining 300 are bad. The data collection contains no missing values.

Out of remaining 19 columns,7 are numeric data and 12 are categorical data. According to heatmap, it is understandable that there is high and positive correlation between "Duration of credit" and "Credit amount in EUR" variables. Additionally, 3 variables has left side skewness which they are, Duration of credit in month, Age of customer in the years and Credit amount in EUR.

Correlation heatmap result is:





Pairplot among the variables

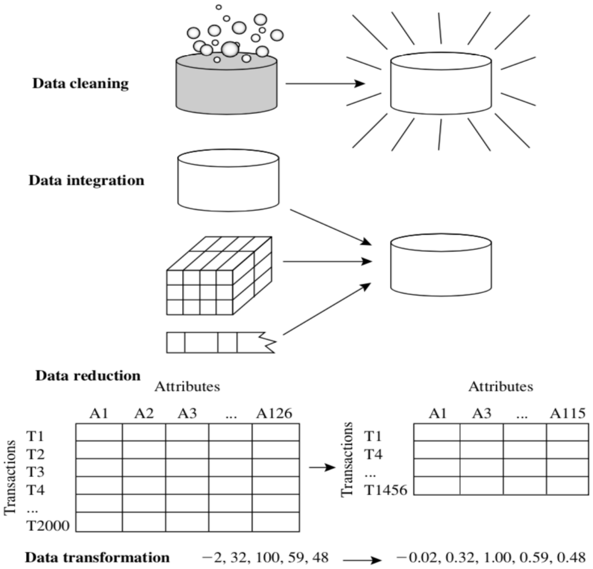
We can learn many things about data and some relationships. First of all, around 60 percent of customers choose credit option which duration of credit is between 10 and 25 month. It is clear proof that amount of credit will be small if duration of credit is less than 25 month.
On the other hand, we can see high customer numbers in age of 20 and 30 years. In that range, most people need money for making start-up or company.

Sometimes it can be education reason. Because between 22 and 25 age, people want to study master degree or MBA. It require money that is why they take credit.

The data in four categorical columns is mainly monotonic, meaning that more than 80 percent of the data falls into the same category: other payment plans, is consumer a foreign worker, and other parties. These columns will be used for classification and clustering, but they will be omitted from pattern mining analysis because they are nearly identical.

3 Data Pre-processing

First off all, we should understand the importance of data pre-processing. What is the data pre-processing? Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogenous sources. Low-quality data will lead to low-quality mining results. "How can the data be preprocessed in order to help improve the quality of the data and, consequently, of the mining results? How can the data be preprocessed so as to improve the efficiency and ease of the mining process?"
There are several data preprocessing techniques.
Data cleaning can be applied to remove noise and correct inconsistencies in data. There are some methods for data cleaning:
1. Ignore the tuple
2. Fill in the missing value manually
3. Use a global constant to fill in the missing value
4. Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value
Data integration merges data from multiple sources into a coherent data store such as a data warehouse.
Data reduction can reduce data size by, for instance, aggregating, eliminating redundant features, or clustering.



Data transformations (e.g., normalization) may be applied, where data are scaled to fall within a smaller range like 0.0 to 1.0. This can improve the accuracy and efficiency of mining algorithms involving distance measurements. Data cleaning routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. If users believe the data are dirty, they are unlikely to trust the results of any data mining that has been applied. Furthermore, dirty data can cause confusion for the mining procedure, resulting in unreliable output. We can move to our dataset. At first 13 columns which are categorical is to converted to numerical data by process called one hot encoding.One hot encoding add new column binary column for each categorical value on data set. The category which represent example is encoded as 1 and all other values as 0. When it comes to normalizing data, different normalization technique such as zero-mean and min max scaling algorithms applied to data first. Then both normalized and non-normalized data set fitted to models which will be discussed later. For classification, there was no difference between normalized and non-normalized data. For clustering normalization resulted in considerable bad silhouette score for each clustering algorithm which is clearly something not desirable. So no normalization technique was applied to data. We check also missing value in the data. We saw that there is no any missing value but if we observe some such values we can apply data imputation methods.

4 Clustering

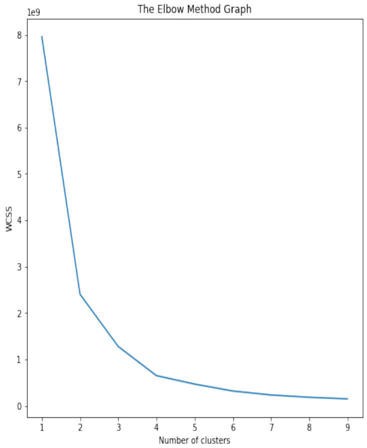
Clustering is an unsupervised Machine Learning method which is used to find similiraties between dataset's objects. Cluster analysis or simply clustering is the process of partitioning a set of data objects (or observations) into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. The set of clusters resulting from a cluster analysis can be referred to as a clustering. Cluster analysis has been widely used in many applications such as business intel- ligence, image pattern recognition, Web search, biology, and security. There three main algorithms are used for Clustering: Kmeans, Agglomerative and DBSCAN. For this dataset 2 different algorithms, namely, KMeans and Agglomerative clustering have been used.

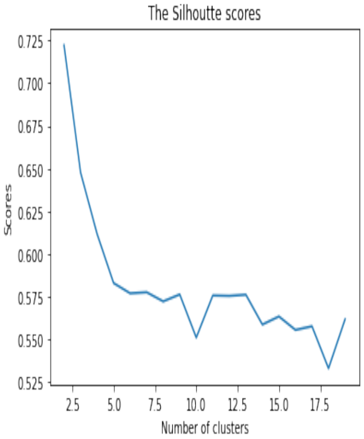
4.1 Distance methods

Most efforts to produce a rather simple group structure from a complex data set require a measure of "closeness" or "similarity". There is often a great deal of subjectivity involved in the choice of a similarity measure. Important considerations include the nature of the variables (discrete, continuous, binary), scales of measurement (nominal, ordinal, interval, ratio), and subject matter knowledge. There are Euclidiean, Minkowoski,Coisine,Chord, Jaccard distance used in clustering.
However we can other methods in Hierarchical Agglomerative Clustering:
1. Single linkage (minimum distance): groups are formed from the individual entities by merging nearest neighbors, where the term nearest neighbor connotes the smallest distance or largest similarity.
2.Complete linkage (maximum distance): Complete linkage clustering proceeds in much the same manner as single linkage clustering, with one important exception: At each stage, the distance between clusters is determined by the distance between the two elements, one from each cluster, that are most distant. Thus, complete linkage ensures, that all items in a cluster are within some maximum distance of each other.
3.Average linkage (average distance between all pairs of items): Average linkage treats the distance between two clusters as the average distance between all pairs of items where one member of a pair belongs to each cluster.

4.2 KMeans

K-means is one of the most popular unsupervised machine learning algorithm thanks to its simplicity and effetivness. Partitioning methods: Given a set of n objects, a partitioning method constructs k partitions of the data, where each partition represents a cluster. That is, it divides the data into k groups such that each group must contain at least one object. In other words, partitioning methods conduct one-level partitioning on data sets. The basic partitioning methods typically adopt exclusive cluster separation. That is, each object must belong to exactly one group. This requirement may be relaxed, for example, in fuzzy partitioning techniques. References to such techniques are given in the bibliographic notes. Most partitioning methods are distance-based. Given k, the number of partitions to construct, a partitioning method creates an initial partitioning. The main task is determining the K, number of clusters before starting process. For determining number of clusters, Elbow method and silhouette scores have been used.

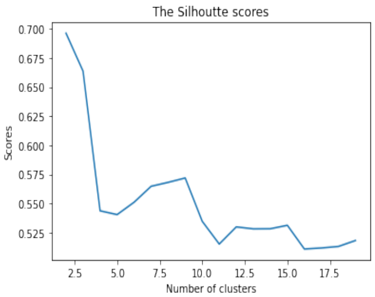


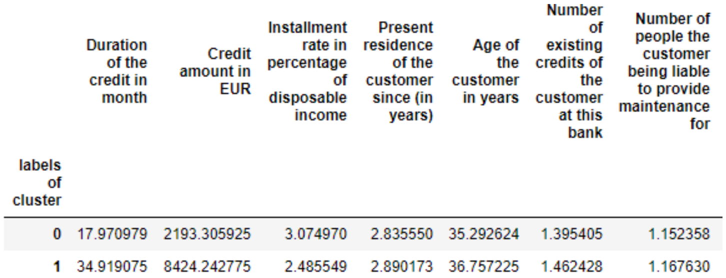


It's hard to tell what is the number of clusters from elbow method. It can be 2 or 4. It is because the clusters on dataset are not well defined.
Silhouette score takes its maximum value when number of clusters is equal to 2.

4.3 Agglomerative clustering

Agglomerative clustering, unlike KMeans, is hierarchical clustering technique. Hierarchical methods: A hierarchical method creates a hierarchical decomposition of the given set of data objects. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. A hierarchical clustering method works by grouping data objects into a hierarchy or "tree" of clusters. It successively merges the objects or groups close to one another, until all the groups are merged into one (the topmost level of the hierarchy), or a termination condition holds. The divisive approach, also called the top-down approach, starts with all the objects in the same cluster. In each successive iteration, a cluster is split into smaller clusters, until eventually each object is in one cluster, or a termination condition holds.
For agglomerative clustering we can't use Elbow method since there's no cluster centers in this algorithm. But again Silhoutte score take its maximum values at 2. Based on this analysis we can tell that there are 2 clusters in this dataset.





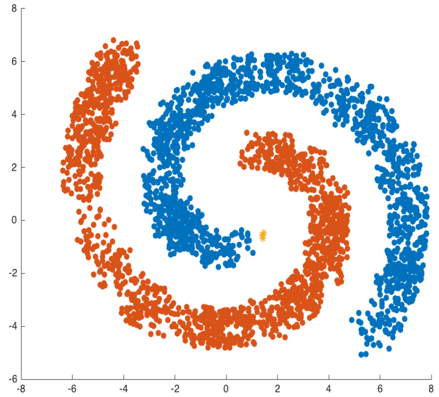
	Duration of the credit in month	Credit amount in EUR	Installment rate in percentage of disposable income	Present residence of the customer since (in years)	Age of the customer in years	Number of existing credits of the customer at this bank	Number of people the customer being liable to provide maintenance for
labels of cluster							
0	17.970979	2193.305925	3.074970	2.835550	35.292624	1.395405	1.152358
1	34.919075	8424.242775	2.485549	2.890173	36.757225	1.462428	1.167630

The credit amounts and length of credit are the key differences between clusters, according to Anaylisys. Both the credit sum and the credit period are significantly higher in the second cluster than in the first.

4.4 DBSCAN clustering

Density-based spatial clustering of applications with noise (DBSCAN) is a well-known data clustering algorithm that is commonly used in data mining and machine learning. Based on a set of points (let's think in a bidimensional space as exemplified in the figure), DBSCAN groups together points that are close to each other based on a distance measurement (usually Euclidean distance) and a minimum number of points. It also marks as outliers the points that are in low-density regions.

Parameters: The DBSCAN algorithm basically requires 2 parameters:
eps: specifies how close points should be to each other to be considered a part of a cluster. It means that if the distance between two points is lower or equal to this value (eps), these points are considered neighbors.
minPoints: the minimum number of points to form a dense region. For example, if we set the minPoints parameter as 5, then we need at least 5 points to form a dense region.



5 Classification

Classification is supervised Machine Learning method which is used to predict the class of new data based on labeled dataset. Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels. For example, we can build a classification model to categorize bank loan applications as either safe or risky. Such analysis can help provide us with a better understanding of the data at large. Many classification methods have been proposed by researchers in machine learn- ing, pattern recognition, and statistics. In this dataset there are two classes: Good customer and Bad customer.Objective is getting accuracy score as much as possible. For classification, 80% of data choosen for training and 20% for testing, which is about 800 to 200 examples. Logistic Regression, KNN ,Support Vector Machine, Neurol Network and Random Forest algorithms have been used for classification.

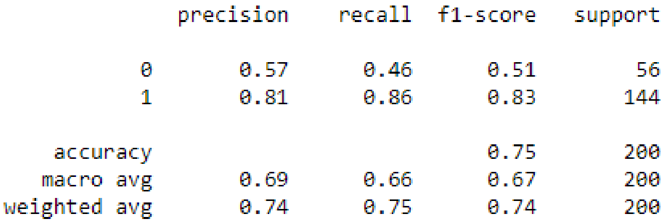
5.1 Logistic regression

Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Consider a scenario where we need to classify whether an email is spam or not. If we use linear regression to solve this problem, we'll need to set a threshold by which we can classify the data. If the actual class is malignant, the expected continuous value is 0.4, and the threshold value is 0.5, the data point would be labeled as not malignant, potentially resulting in severe consequences in real time. There are 3 types of Logistic Regression 1. Binary Logistic Regression The categorical response has only two 2 possible outcomes. Example: Spam or Not 2. Multinomial Logistic Regression Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg,

Vegan) 3. Ordinal Logistic Regression Three or more categories with ordering. Example: Movie rating from 1 to 5
Cross Validation is used for getting average accuracy.

array([0.745, 0.745, 0.76 , 0.74 , 0.735])

Which is about 74.5% accuracy. Accuracy score didn't change when normalizing applied to data.

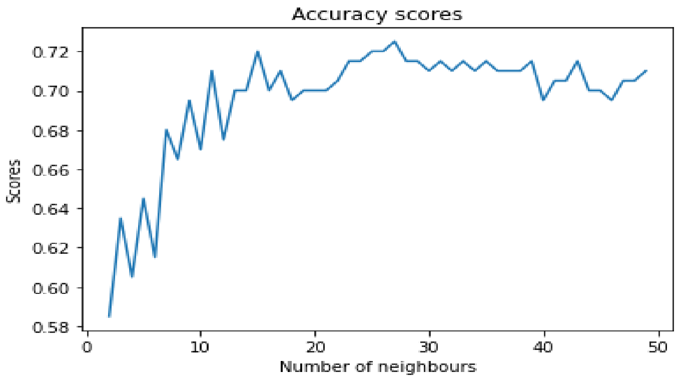


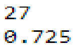
```
              precision    recall  f1-score   support

           0       0.57      0.46      0.51        56
           1       0.81      0.86      0.83       144

    accuracy                           0.75       200
   macro avg       0.69      0.66      0.67       200
weighted avg       0.74      0.75      0.74       200
```

From Classification report we can see that Logistic regression in this data set didn't work quite well. It could find about 75% of labels correctly. It wasn't that succesfull to classify bad customers. Only 57% precision we could get for classifying "bad" labels.But it is pretty succesul in classifying "good" labels.

5.2 KNN

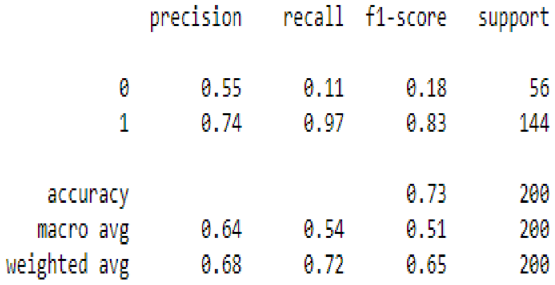
The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The important factor in KNN is choosing right number for K. There are some rules:
1. As we decrease the value of K to 1, our predictions become less stable. Just think for a minute, imagine K=1 and we have a query point surrounded by several reds and one green , but the green is the single nearest neighbor. Reasonably, we would think the query point is most likely red, but because K=1, KNN incorrectly predicts that the query point is green.
2.Inversely, as we increase the value of K, our predictions become more stable due to majority voting / averaging, and thus, more likely to make more accurate predictions. Eventually, we begin to witness an increasing number of errors. It is at this point we know we have pushed the value of K too far.
3. In cases where we are taking a majority vote among labels, we usually make K an odd number to have a tiebreaker.
Advantages: The algorithm is simple and easy to implement. There's no need to build a model, tune several parameters, or make additional assumptions. The algorithm is versatile. It can be used for classification, regression, and search (as we will see in the next section).
Disadvantages: The algorithm gets significantly slower as the number of examples and/or predictors/independent variables increase.





```
27
0.725
```

Best number of k is reached at 27 in our dataset. Which is not something normal. Also CVS (72.5) is less than Logistic regression (74.5).Classification report as follows:



```
              precision    recall  f1-score   support

           0       0.55      0.11      0.18        56
           1       0.74      0.97      0.83       144

    accuracy                           0.73       200
   macro avg       0.64      0.54      0.51       200
weighted avg       0.68      0.72      0.65       200
```

Precision is also less than what is achieved in Logistic Regression.

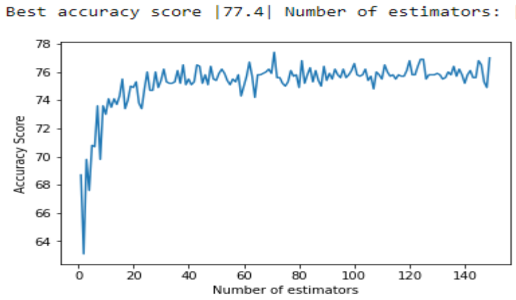
5.3 Support Vector Machine

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of

features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds 3.
Accuracy achieved was 72 which is less than what is achieved by Logistic Regression but higher than KNN.

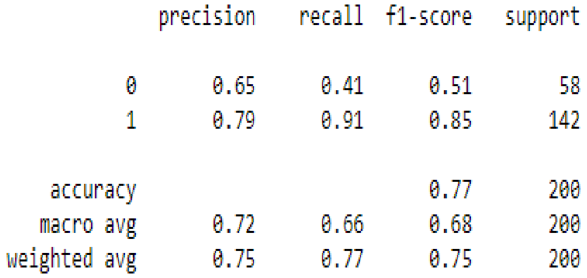
5.4 Random Forest Classifier

Random Forest is ensemble ML algorithm.Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well is: A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models. So the prerequisites for random forest to perform well are: There needs to be some actual signal in our features so that models built using those features do better than random guessing. The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.



```
Best accuracy score |77.4| Number of estimators: |71|
```

Best accuracy score is reached when the number of decision trees are 71. This algorithm with 71 decision trees outperformed all algorithms which mentioned above with 77.4 average accuracy. When we look at graph we can tell that starting from 30 it takes average values between 74 and 76.

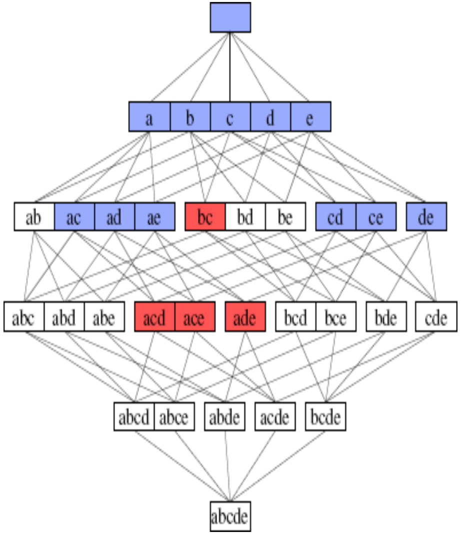


```
              precision    recall  f1-score   support

           0       0.65      0.41      0.51        58
           1       0.79      0.91      0.85       142

    accuracy                           0.77       200
   macro avg       0.72      0.66      0.68       200
weighted avg       0.75      0.77      0.75       200
```

As we can see from classification report, precision score is better than the other algorithms. Only little weaker for classifying good label than Logistic regression. Recall score is also outnumbered the other algorithms.
So for this data set, Random Forest Classifier performed best by taking 77.4 accuracy score and 80% precision to detect the good labels.

6 Frequent pattern mining

Frequent pattern mining is part of Data Mining techniques. The main aim of FPM is finding pattern in data set which are most frequent and relevant. Frequent patterns are patterns (e.g., itemsets, subsequences, or substructures) that appear frequently in a data set. For example, a set of items, such as milk and bread, that appear frequently together in a transaction data set is a frequent itemset. Frequent pattern mining searches for recurring relationships in a given data set. This section introduces the basic concepts of frequent pattern mining for the discovery of interesting associations and correlations between itemsets in transactional and relational databases.



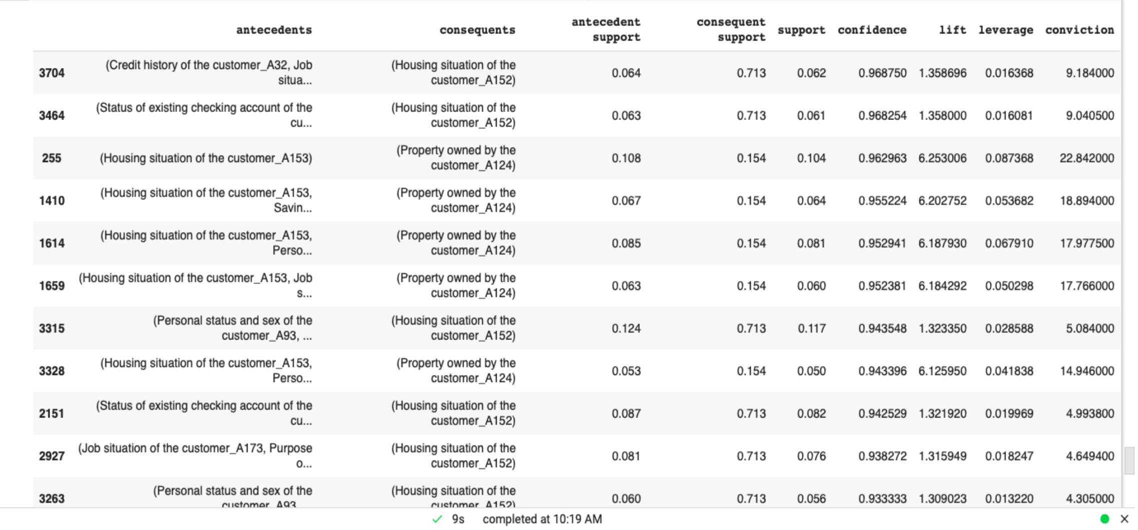
A frequent pattern represents a set of items co-occurring in comparatively more transactions, for instance in the horizontal layout example item1 and item2 appear frequently together. This frequency is quantified using the support metric. Itemset support is the number of transactions where the itemset elements appear together divided by the total number of transactions. Minimum support is a threshold used by the following algorithms in order to discard sets of items from the analysis which don't appear frequently enough. The strength of the association rule between 2 items, (for instance item1 and item2) or the association confidence represents the number of transactions containing item1 and item 2 divided by the number of transactions containing item1. In FPM, mainly two algorithms are used: Apriori , Eclat and FP tree.

Apriori algorithm uses data organized by horizontal layout. It is founded on the fact that if a subset S appears k times in a database, any other subset S1 which contains S will appear k times or less.

Eclat (Equivalence Class Clustering and bottom up Lattice traversal) algorithm uses data organized by vertical layout which associates each element with the list of underlying transactions. Eclat algorithm is generally faster than apriori and requires only one database scan which will find the support for all itemsets with 1 element.

FP tree algorithm uses data organized by horizontal layout. It is the most computationally efficient algorithm from the 3 presented in this post. It only performs 2 database scans and keeps the data in an easily exploitable tree structure.

At first, all columns of this data set were used for FPM. Since there were some columns with mostly same values, as mentioned at EDA part, result contained items with 100% confidence. These results were expected, because when columns are consist of same elements, these elements are most likely to be appear on every other categories. So The columns which are more than 80% same categories are removed. After these columns removed number of patterns with more than 90% confidence dropped down to 25. The top 5 patterns are as follows:



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
3704	(Credit history of the customer_A32, Job situa...	(Housing situation of the customer_A152)	0.064	0.713	0.062	0.968750	1.358696	0.016368	9.184000
3464	(Status of existing checking account of the cu...	(Housing situation of the customer_A152)	0.063	0.713	0.061	0.968254	1.358000	0.016081	9.040500
255	(Housing situation of the customer_A153)	(Property owned by the customer_A124)	0.108	0.154	0.104	0.962963	6.253006	0.087368	22.842000
1410	(Housing situation of the customer_A153, Savin...	(Property owned by the customer_A124)	0.067	0.154	0.064	0.955224	6.202752	0.053682	18.894000
1614	(Housing situation of the customer_A153, Perso...	(Property owned by the customer_A124)	0.085	0.154	0.081	0.952941	6.187930	0.067910	17.977500
1659	(Housing situation of the customer_A153, Job s...	(Property owned by the customer_A124)	0.063	0.154	0.060	0.952381	6.184292	0.050298	17.766000
3315	(Personal status and sex of the customer_A93, ...	(Housing situation of the customer_A152)	0.124	0.713	0.117	0.943548	1.323350	0.028588	5.084000
3328	(Housing situation of the customer_A153, Perso...	(Property owned by the customer_A124)	0.053	0.154	0.050	0.943396	6.125950	0.041838	14.946000
2151	(Status of existing checking account of the cu...	(Housing situation of the customer_A152)	0.087	0.713	0.082	0.942529	1.321920	0.019969	4.993800
2927	(Job situation of the customer_A173, Purpose o...	(Housing situation of the customer_A152)	0.081	0.713	0.076	0.938272	1.315949	0.018247	4.649400
3263	(Personal status and sex of the customer_A93	(Housing situation of the customer_A152)	0.060	0.713	0.056	0.933333	1.309023	0.013220	4.305000

✓ 9s completed at 10:19 AM

From association rules we see that, with confidence 0.96 we may say that customers single male customers who is official worker are owning car and house. And obviously whose housing situation are accommodation don’t posses a house.

7 Conclusion

1. During preprocessing all categorical data converted to numeric data using dummy variables. No normalization and discretization technique applied. During data cleaning process unnecessary column (telephone of customers) dropped.
2. Data set split into clusters. Both Elbow method and Silhouette score for both KMeans and Agglomerative clustering showed that 2 is the number of clusters. KMeans shows slightly higher performance as oppose to Agglomerative clustering. Two clusters were mainly differ in the duration of credit and amount of credit. Duration of credit for one cluster was twice and duration of credit was 4 time higher than the other clusters. Normalization dropped the performance of both algorithms.
3.Four different: Logistic regression, KNN , Support Vector Machine, Random Forest Classifier and NN algorithms applied to data. Each method evaluated based on their precision score, recall score and cross validation score. Random Forest Classifier gave best results (79%) with number of estimators.
4.Apriori algorithm was used for Frequent Pattern Mining. Some columns which consist of mainly same categories dropped. Items with more than 0.9 confidence analyzed.