Customer Credit Risk Analysis

Mhyar Kousa 1 and Dr.Tomas Horvath 2
1 MSc. Data Science, ELTE University, mhyarov.k@gmail.com,
2 Head of Data Science and Energeeing Department,tomas.horvath@inf.elte.hu
ELTE University - Faculty of Informatics.
1117 Budapest, Pázmány Péter str.

May 30, 2021

Abstract: One of the main areas of research for each banking institution is the assessment of the client's credit risk [1]. The historical data of the customer can be used to assess the paying capacity of new customers and the risks of credit defaults. In addition, customers with similar behaviour can be grouped by clustering algorithms and patterns of their behavior can be identified. In this research paper, logistic regression was compared with other well-known algorithms to classify customers as good or bad. Moreover, k-means clustering method was applied to form the groups of clients with relatively similar features. Finally, it was demonstrated that frequent pattern mining can be applied to the customer credit risk data and patterns can be found.

Keywords: customer credit risk · credit risk prediction · machine learning · classification · clustering · features engineering · correlation of the data · frequent pattern mining . association rules · appriori algorithm.

1 Introduction

Nowadays credit risk analysis is becoming a significant part of banking industry. Before deciding to approve a loan to a particular customer, the bank assesses the customer according to the available parameters and historical data. Due to the fact that banks process a large amount of information while assessing credit risk, machine learning techniques can be used to simplify this process and to get more accurate evaluations. In this paper, we are going to investigate several models to identify clients with good and bad credit risk. Moreover, clustering will be performed on client data. Finally, we will give insides on how to apply frequent pattern mining to the mentioned problem.

2 Data Exploration

2.1 Data Description

The given data provides information about customers who take a bank loan. The data set itself contains 1000 observations with 20 independent features, where 7 of them are numerical variables and other 13 are categorical variables.

2.2 Exploring the data type

To Begin,with info() function in pandas library to know the type of data for every column and if the null values exist or not.There are number of categorical features as the Table 1. shown.

Categorical Features	**Description**
Status account of the customer	Status of existing checking account of the customer.
Duration of the credit	Duration of the credit (requested by the customer from the bank) in mont
Credit History	Credit history of the customer
Purpose	Purpose of the credit
Savings account	Savings account/bonds of the customer
Present employment	Present employment of teh customer since
Status	Personal status and sex of the customer
Guarantors for the credit	Other debtors or guarantors for the credit
Property	Property owned by the customer
Installment plans	Other installment plans of the customer
Housing situation	Housing situation of the customer
Job situation	Job situation of the customer
Telephone	Telephone of the customer
Foreign Worker	if the customer is a foreign worker

Table 1: Categorical Features

There were no missing values in data as for each column the info() function returns 'Non-Null Count' value equals to 1000. Moreover, by checking the Series with value counts() we proved that there were no nulls in data as well. By calling the describe function it will return information about numerical features of the current data set.

From the Table 2 shown number of numerical features and a short description about each feature.

	Duration of the credit	Credit amount in EUR	Installment rate	Present residence	Age	Number of existing credits	Number of people the customer being liable
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	20.903000	3271.258000	2.973000	2.845000	35.546000	1.407000	1.155000
std	12.058814	2822.736876	1.118715	1.103718	11.375469	0.577654	0.362086
min	4.000000	250.000000	1.000000	1.000000	19.000000	1.000000	1.000000
25%	12.000000	1365.500000	2.000000	2.000000	27.000000	1.000000	1.000000
50%	18.000000	2319.500000	3.000000	3.000000	33.000000	1.000000	1.000000
75%	24.000000	3972.250000	4.000000	4.000000	42.000000	2.000000	1.000000
max	72.000000	18424.000000	4.000000	4.000000	75.000000	4.000000	2.000000

Figure 1: Description of numerical features.

Numerical Features	**Description**
Duration of the credit	Duration of the credit (requested by the customer from the bank) in month.
Credit amount	Credit amount in EUR.
Installment rate	Installment rate in percentage of disposable income.
Present residence	Present residence of the customer since (in years).
Age	Age of the customer in years.
Existing credits	Number of existing credits of the customer at this bank.
Number of people	Number of people the customer being liable to provide maintenance for.

Table 2: Numerical Features.

2.3 Numerical Variables

First of all, let us have a look at the distribution of 'Duration of Credit' variable (Figure 2). It is noticeable that credit records of good customers range from 1 year to 2 years, whereas records for bad customers varies from 1 to 3 years. It does make sense since long-term credit posses more risk for bank. Furthermore,observations for customers with bad credit risk has the larger median than compared to those of reliable customers.To point out, several outliers take place for good credit records with the duration of more than 40 months. As regards credit amount, the riskiest group of clients borrow a larger amount of money from the bank (2500 EUR up to 5000 EUR on the average), while good customers take a loan for a smaller amount (generally less than 4000 EUR). The age of borrowers is approximately the same for both groups and varies from 25 to over 43 years,however the median age of good clients is higher by approximately 2 years. ˚

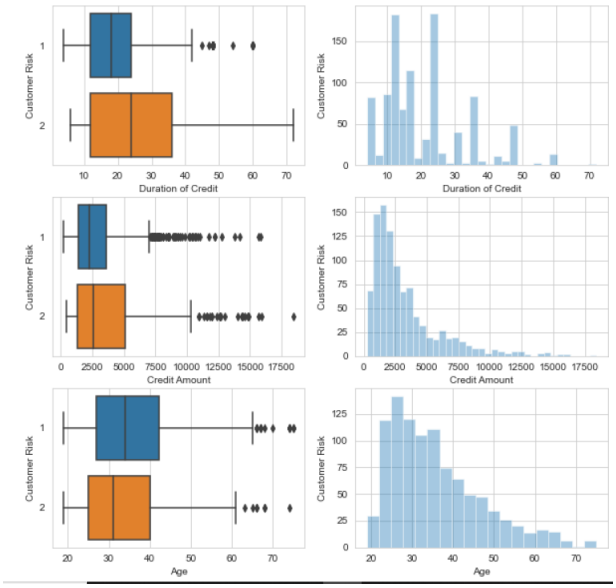

Figure 2: Distribution of numerical variables

2.4 Categorical Variables

2.4.1 Purpose of credit

First of all, let us have a look at how the amount of credit depends on the purpose of credit (Figure 3). It is noticeable that a larger loans are taken for buying a car. Moreover, loans for home related purposes are the smallest for both groups. However, bad customers borrow large sums of money for uncertain purposes.

Figure 3: Purpose of the credit

Some of the purpose categories have very few observations, so that I decided

to merge them under more general classes. As car(used) and car(new) have both sufficient number of observations, I left them as it was. Finally I have got these new categories: car(new), car(used), home equipment (radio/television, domestic appliances, repairs, furniture/equipment),education,business, others. Now the distribution looks as follows (Figure 4):

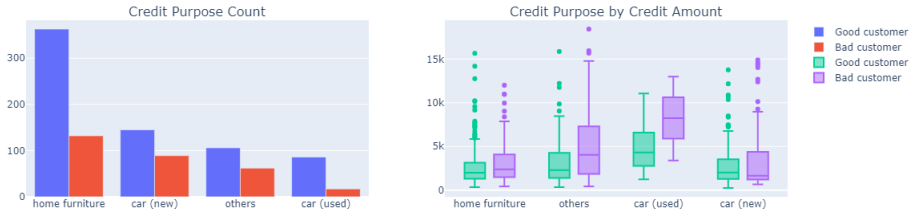

Figure 4: Purpose of the credit

2.4.2 Job situation of the customer

The share of good and bad customers within 4 job categories is approximately the same. However from figure 7 it can be noted that job category 'management/self-employed/highly qualified employee/officer' tends to take larger credits.

Figure 5: Job Situation Distribution

I decided to merge these 4 various job categories into two group: employed and unemployed.

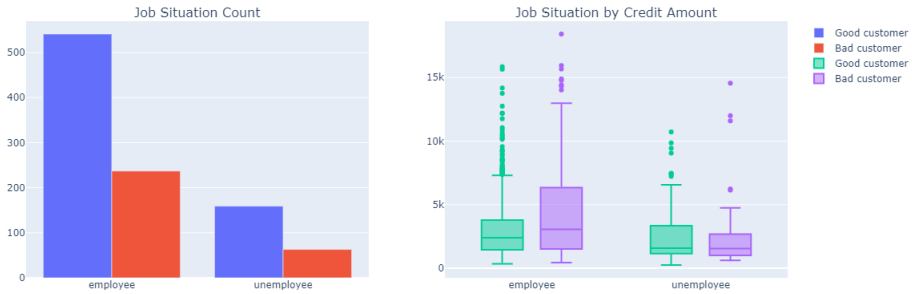

Figure 6: Job Situation Distribution

2.4.3 Credit History

The (Figure 7) shows that customers with bad credit risk have 4 times less critical accounts in comparison with good customers.The majority of customers paid back duly in both groups.

Figure 7: Credit History Distribution

According to the given distribution,we can see that credit history data can be split into three generalized categories: critical account/delays (A33, A34), paid duly (A31, A32), no credits(A30).

Figure 8: Credit History Distribution

2.4.4 Personal status and sex of the customer

I decided to split this attribute into two new attributes: sex (male/ female) and personal status (single, married at least once). Further, I got the following distribution of observations. In general, more men than women are tend to take loans in banks. According to the statistics, single men posses more risks to bank (146 out of 1000 customers).

Customer Type	Female	Male	Single	Married at least one
Good Customers	201	499	402	298
Good Customers,%	28.7	71.2	57.4	42.5
Bad Customers	109	191	146	154
Bad Customers,%	36.3	63.6	48.6	51.3

Table 3: Table captions

2.4.5 Savings account of the customer

For the status of existing checking account of the customer I decide using four general classes(A14:no checking account, A11:poor, A12::modurate, A13:rich) depending on customer account.

The (Figure 9) illustrate that large ratio of bad customer going for poor and modurate customer.Wherear,the majority of good customers going for no checking account .

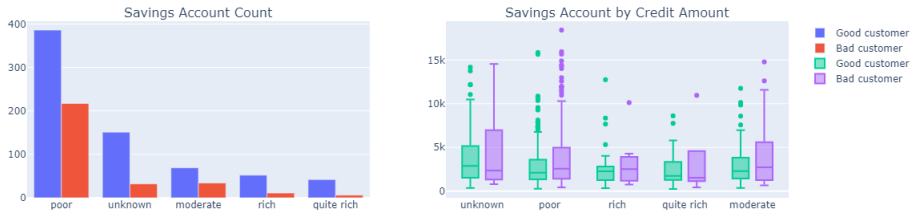

Figure 9: Savings Account Distribution

2.4.6 Status of existing checking account of the customer

For the status of existing checking account of the customer I decide using four general classes(A14:no checking account, A11:: poor, A12::modurate, A13:rich) depending on customer account .The Figure 8 illustrate that large ratio of bad customer going for poor and modurate customer .Wherear,the majority of good customers going for no checking account .

Figure 10: Customer Status Account Distribution

2.4.7 Housing situation of the customer

First,lets have a look for own house has the highest percentage that most of customer prefer to own house rather than renting that has ess than its predecessor and the lowest percentage go for accommodation.

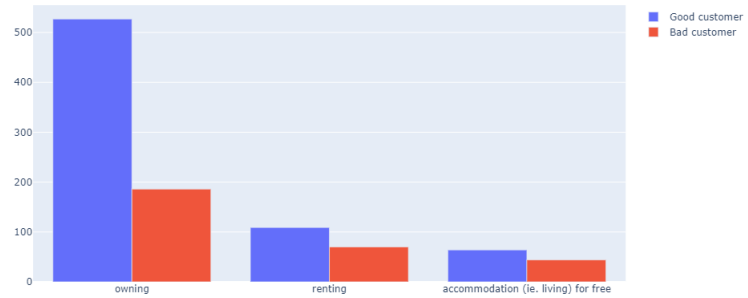

Figure 11: Housing Distribution

2.4.8 Present employment of the customer

For the present employment status I decide using four general classes(fresh graduated, A72: junior, A74:senior, A75:supervisor, A71: unemployed) depending on number of years .The Figure 8 illustrate that large ratio of bad customer going for junior,fresh graduated and supervisor employee .Wherear,the majority of good customers going for junior,supervisor and senior employee .

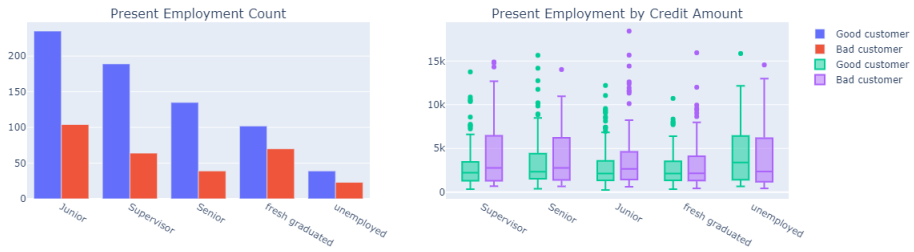

Figure 12: Present Employment Distribution

2.4.9 Other installment plans of the customer

It is noticeable, that customers per stores and bank have small ratio of representatives, so that we can merge stores and bank categories under a new one. Finally, we have two levels under category "Other installment plans of the customer": 1) none 2) bank and stores.

2.4.10 Other debtors or guarantors

for the credit I combined the categories into two larger ones. Now there are 2 levels: 1) none 2) has guarantor, which unites 'co-applicant' and 'guarantor' levels.

2.5 Target Variable

Let's start looking through target variable and their distribution there is two type of customer good and bad,the highest percentage is for good customers while the lowest percentage is for bad customers.

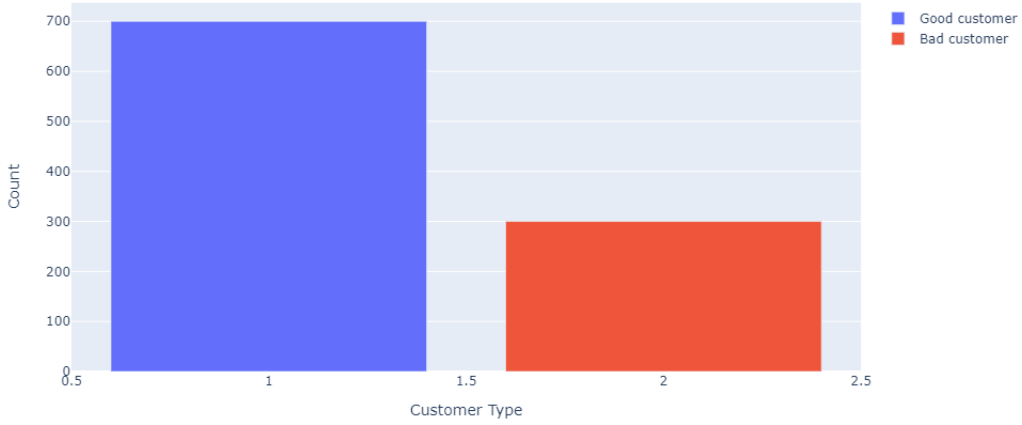

Figure 13: Distribution of target variable

3 Data Processing

3.1 Binarization and Dummy Variables

After modified levels of categorical variables,we are going further with creation of dummy and binary variables for categorical features.First of all,features with two level are transformed into binary representation.

1. **Customer Type:** 'good customer': 1 , 'bad customer': 0.
2. **Foreign Worker:** if the customer is a 'foreign worker': 1 or 'not foreign': 0 .
3. **Sex:** 'male': 1, 'female': 0 .

4. **Personal status:** 'married at least one': 1, 'single': 0 .

5. **Other debtors or guarantors for the credit:** 'has guarantors':1 , 'none': 0 .

6. **Job situation of the customer:** 'employed': 1, 'unemployed': 0 .

7. **Other installment plans of the customer:** 'bank and stores': 1 , 'none': 0 .

8. **Telephone of the customer:** 'no phone': 0, 'has phone': 1

The pandas get dummies function was used to convert other categorical variable into dummy ones. For instance, column 'X15' (Housing) was expanded to three new binary variables 'X15 A151' (Renting), 'X15 A152' (Owning) and 'X15 A153' (Accommodating for free). In case customer owns any accommodation (X15 is equal to A151), the corresponding dummy variable (X15 A151) will be 1 and other dummies will be 0.

...	Customer Status Account_no checking account	Customer Status Account_poor	Customer Status Account_rich	Credit History_no credits taken or all credits paid back duly	Credit History_paid duly	Property_if not A121/A122 - car or other, not in attribute X06_y	Property_real estate_y	Property_unknown/no property_y	Housing Situation_owning_y	Housing Situation_renting_y
..	0	1	0	0	0	0	1	0	1	0
..	0	0	0	0	1	0	1	0	1	0
..	1	0	0	0	0	0	1	0	1	0
..	0	1	0	0	1	0	0	0	0	0
..	0	1	0	0	0	0	0	1	0	0

Figure 14: Dummy Variables

3.2 Correlation of the data

Because there are a huge number of features in the data set i prefer to use a subset of features for measuring correlation coefficients between them using correlation matrix,from the Figure 9 It seems noticeable there is a strong relation between Credit amount and Duration of the credit and also there is a weak and negative relation between age and duration of the credit.

3.3 Features Selection

Univariate feature selection For each customer is represented by 58 features.It should be noticed that not an every feature can contribute to the model or have a strong descriptive capacity.Therefore there is a need to drop less important attributes. First of all, univariate feature selection can be applied.This method examines each feature independently to determine the strength of the relationship of the feature with the target variable.Scikit-learn provides f classif method to calculate the p-values. It computes the ANOVA F-value for the provided sample. I chose the common threshold for p-values of 0.15, meaning

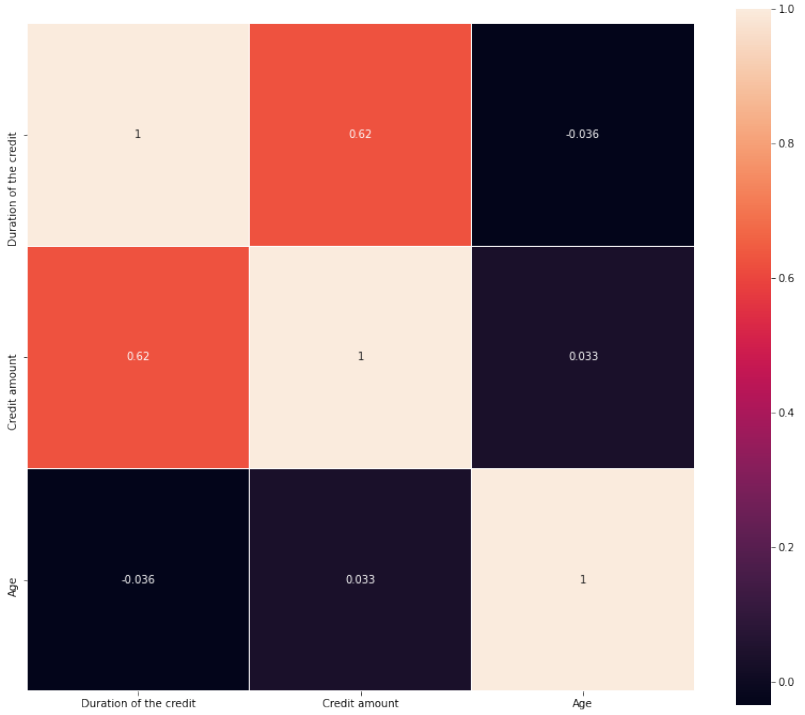

Figure 15: Correlation between Features

that anything less than 0.15 may be considered significant. According to this we can drop features with p-value more than or equals to 0.15.

Features Contribution ' Credit Duration', 'Credit Amount', 'Installment Rate','Age', 'Foreign Worker', 'Customer Type', 'Sex','Personal Status','Present Employment_Supervisor','Present Employment$_u nemployed'$,'$SavingsAccount\_poor$','$SavingsAccount\_quiteri$
$carorother, notinattributeX06'$,'$Property\_realestate'$,'$Property\_unknownnoproperty'$,'$HousingSituation\_o$

3.4 Data Scaling

Before using different kind of models to predict customer risk ,numerical variables were scaled.Sklearn.preprocessing package provide for us scale() method which standardize data. Standardization is a popular scaling technique where the values are centered around the mean with a unit standard deviation. So that the variable has a mean of zero and the obtained distribution has a unit standard deviation.

	Credit Duration	Credit Amount	Installment Rate	Credit Guarantors	Present Residence	Age	Existing Credits Number	Job Situation	People Number	Telephone	...	Customer Status Account_no checking account	Customer Status Account_poor	Customer Status Account_rich
1	-1.236478	-0.745131	4	1	4	2.766456	2	0	1	0	...	0	1	0
2	2.248194	0.949817	2	1	2	-1.191404	1	0	1	1	...	0	0	0
3	-0.738668	-0.416562	2	1	3	1.183312	1	0	2	1	...	1	0	0
4	1.750384	1.634247	2	0	4	0.831502	1	0	2	1	...	0	1	0
5	0.256953	0.566664	3	1	4	1.535122	2	0	2	1	...	0	1	0

5 rows × 32 columns

Figure 16: Scaling Numerical Features

4 Classification Model

4.1 Training and testing data

Initially there is X set of selected features and Y set of the target values which demonstrates whether the type of customer is good or bad. The data is split by using 80% of the whole data to train models and the rest 20% of the data to evaluate them. Moreover, by setting random seed for results reproduction.

4.2 Classifiers

Following models were considered:

1. Random Forest (RF)
2. K-Nearest Neighbors (KNN)
3. Gaussian Naive Bayes (NB)
4. Logistic Regression (LR)
5. Decision Tree (DT)

Tuning parameters As regard LR,I set class weight="balanced", because it basically replicates the smaller class until we have as many samples as in the larger one, but in an implicit way.I used grid search to tune criterion,max depth and criterion parameters for DT. Similarly, this approach was used to find out best max depth, max features and n estimators for RF classifier.KNN model's parameters like number of neighbors, distance metric were tuned with grid search as well.

4.3 Results obtained

In the (Table 4) the obtained results are presented. It is noticeable that logistic regression with C equals to 0.01 and balanced class weights performs better results. Even though we have not attained the state-of-art accuracy, the AUC is 0.78. In other words, when AUC is more than 0.5 there is a high probability that our model will have a capacity to distinguish bad clients from good clients.

Model	Accuracy	Precision	Recall	F1_Score	ROC_AUC
GNB	0.71	0.84	0.72	0.78	0.70
RF	0.77	0.80	0.90	0.85	0.68
KNN	0.74	0.78	0.89	0.83	0.64
DT	0.74	0.84	0.78	0.82	0.71
LR	0.78	0.80	0.91	0.85	0.68

Table 4: Comparison between results of several classifiers

The (Figure 17) shows that more numbers of TP and TN are detected than FP and FN. Moreover, it can be noted that around 73% of bad clients on test data are distinguished.

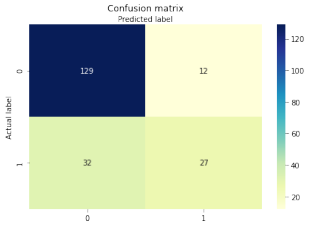

Figure 17: Confusion Matrix

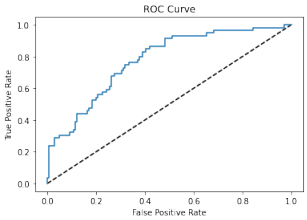

Figure 18: Roc curve for LR.

4.4 Metrics for evaluation

I evaluated models with accuracy,precision,recall, f1-score and roc-auc score. To point out, it is five times worse to classify customers as good when they are bad than it is to classify customers bad when they are good.Therefore, it is important to pay attention to recall, as it shows what proportion of actual positives (bad customers) was identified correctly. However, good customers should not be underestimated, so that other metrics are considered as well.

5 Clustering

5.1 K-Means Algorithm

The algorithm has several steps that are repeated until meeting stop condition: centroids initialization, datapoints assignment to closest centroids and centroids update.Traditional K-Means with euclidean distance metric does not work with categorical variables.

Steps for working K-Means algorithm:

Step 1-: Choose the number of clusters k. ...
Step 2-: Select k random points from the data as centroids. ...
Step 3-: Assign all the points to the closest cluster centroid. ...
Step 4-: Recompute the centroids of newly formed clusters. ...
Step 5-: Repeat steps 3 and 4.

5.2 The number of clusters

To choose the number of clusters I tried to calculate inertia values for various numbers of clusters and apply an elbow method.The idea was to find the point with small number of clusters and small inertia. From Figure 11 it can be noticed that there is a definite point (no of clusters=3) where inertia stops decreasing significantly.

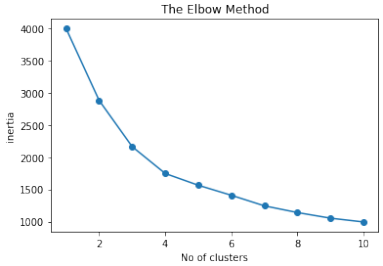

Figure 19: No of clusters

5.3 Principal components

To visualised the clusters into 2D there is a technique called,The data was represented with 2 principal components which explain 87% of data variance.From figure 13 it can be noticed clusters are located close to each other, however, the dividing borders between clusters can be found.

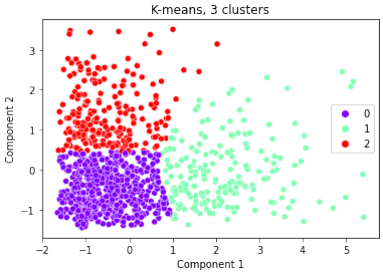

Figure 20: No of clusters

5.4 Hierarchical clustering

In hierarchical clustering, you categorize the objects into a hierarchy similar to a tree-like diagram which is called a dendogram. The distance of split or merge (called height) is shown on the y-axis of the dendogram below.

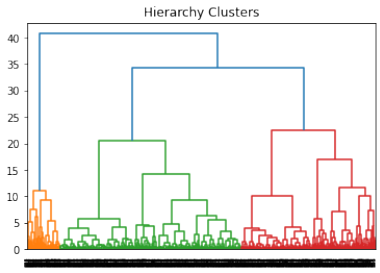

Figure 21: Dendogram

5.5 Statistics results obtained for clusters

Cluster	Credit Duration	Credit Amount	Age
0	16.65	1871.74	35.30
1	29.92	5640.57	36.25
2	39.16	11539.01	35.89

Table 5: Clustering Results

6 Frequent Pattern Mining

Frequent patterns are collections of items which appear in a data set at an important frequency (usually greater than a predefined threshold)and can thus reveal association rules and relations between variables.However, instead of the set of items like apples, beer etc., we have the set of customers' descriptive features.Therefore by using pattern mining we can find out having which features implies customer will be bad for us or vice versa. Alternatively, we can simply identify a set of features going together without reference to credit risk.

6.1 Association Rules

Association rule mining is a technique to identify frequent patterns and associations among a set of items and used when you want to find an association between different objects in a set, find frequent patterns in a transaction database, relational databases or any other information repository. The applications of

Association Rule Mining are found in Marketing, Basket Data Analysis (or Market Basket Analysis) in retailing, clustering and classification. It can tell you what items do customers frequently buy together by generating a set of rules called Association Rules. In simple words, it gives you output as rules in form if this then that.To implement association rule mining, many algorithms have been developed. Apriori algorithm is one of the most popular and arguably the most efficient algorithms among them.

- **Support** refers to the default popularity of an item and can be calculated by finding number of transactions containing a particular item divided by total number of transactions.

  Support(B) = (Transactions containing (B))/(Total Transactions)

- **Confidence** refers to the likelihood that an item B is also bought if item A is bought. It can be calculated by finding the number of transactions where A and B are bought together, divided by total number of transactions where A is bought.Mathematically, it can be represented as:

  Confidence(A→B) = (Transactions containing both (A and B))/(Transactions containing A)

- **Lift** refers to the increase in the ratio of sale of B when A is sold. Lift(A → B) can be calculated by dividing Confidence(A → B) divided by Support(B). Mathematically it can be represented as:

  Lift(A→B) = (Confidence (A→B))/(Support (B))

- **Conviction** a high conviction value means that the consequent is highly depending on the antecedent

6.2 Appriori Algorithm

The Apriori algorithm uses frequent itemsets to generate association rules, and it is designed to work on the databases that contain transactions. With the help of these association rule, it determines how strongly or how weakly two objects are connected. This algorithm uses a breadth-first search and Hash Tree to calculate the itemset associations efficiently. It is the iterative process for finding the frequent itemsets from the large dataset.

Steps for working Apriori algorithm:

Step-1: Determine the support of itemsets in the transactional database, and select the minimum support and confidence.
Step-2: Take all supports in the transaction with higher support value than the minimum or selected support value.

Step-3: Find all the rules of these subsets that have higher confidence value than the threshold or minimum confidence.
Step-4: Sort the rules as the decreasing order of lift.

Data and steps of working:

I worked with the same binarized data set that I used previously for classification.Let us suppose individual customer (row) is a transaction where items are our binary features. In case feature (so-called item) equals to 1, it will present in transaction.
First,I will start with all individual items called candidates and calculate their support counts.This is called candidate list generator then remove candidate that fail min sup count .The list now containing the frequent item sets.

	support	itemsets
0	0.907	(Credit Guarantors)
1	0.596	(Telephone)
2	0.690	(Sex)
3	0.690	(Personal Status)
4	0.603	(Savings Account_poor)
5	0.530	(Credit History_existing credits paid back dul...
6	0.713	(Housing Situation_owning)
7	0.814	(Installment Plans_none)
8	0.531	(Telephone, Credit Guarantors)
9	0.624	(Sex, Credit Guarantors)
10	0.624	(Personal Status, Credit Guarantors)
11	0.532	(Savings Account_poor, Credit Guarantors)
12	0.647	(Housing Situation_owning, Credit Guarantors)
13	0.742	(Installment Plans_none, Credit Guarantors)
14	0.690	(Personal Status, Sex)
15	0.517	(Housing Situation_owning, Sex)
16	0.554	(Installment Plans_none, Sex)
17	0.517	(Housing Situation_owning, Personal Status)
18	0.554	(Installment Plans_none, Personal Status)
19	0.576	(Installment Plans_none, Housing Situation_own...
20	0.624	(Sex, Personal Status, Credit Guarantors)
21	0.505	(Installment Plans_none, Sex, Credit Guarantors)
22	0.505	(Installment Plans_none, Personal Status, Cred...

Figure 22: Frequent item sets.

Second,Take all the subset that have higher confidence value than the threshold or minimum confidence (Figure 23) represent the obtained results.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
24063	(Installment Plans_none, Existing Credits Numb...	(Installment Rate, Credit Amount, Present Resi...	0.205	0.252	0.205	1.000000	3.968254	0.153340	inf
24064	(Installment Plans_none, Existing Credits Numb...	(Installment Rate, Credit Amount, Personal Sta...	0.205	0.252	0.205	1.000000	3.968254	0.153340	inf
24065	(Personal Status, Installment Rate, Existing C...	(Installment Plans_none, Credit Amount, Presen...	0.252	0.205	0.205	0.813492	3.968254	0.153340	4.262553
24066	(Installment Rate, Existing Credits Number, Sex)	(Installment Plans_none, Credit Amount, Person...	0.252	0.205	0.205	0.813492	3.968254	0.153340	4.262553
24067	(Personal Status, Credit Amount, Existing Cred...	(Installment Plans_none, Installment Rate, Pre...	0.252	0.205	0.205	0.813492	3.968254	0.153340	4.262553
24068	(Credit Amount, Existing Credits Number, Sex)	(Installment Plans_none, Installment Rate, Per...	0.252	0.205	0.205	0.813492	3.968254	0.153340	4.262553
24069	(Personal Status, Existing Credits Number, Pre...	(Installment Plans_none, Installment Rate, Cre...	0.252	0.205	0.205	0.813492	3.968254	0.153340	4.262553

Figure 23: The subset by minimum confidence

Finally,the subset result will be sorting with left and the minimum value of threshold the obtained result shown in (Figure 24).

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
42553	(Present Residence, Sex)	(Existing Credits Number, Installment Plans_no...	0.252	0.205	0.205	0.813492	3.968254	0.153340	4.262553
42554	(Credit Duration_1.3870748387117167, Present R...	(Existing Credits Number, Installment Plans_no...	0.342	0.205	0.205	0.599415	2.923977	0.134890	1.984599
42555	(People Number, Present Residence)	(Existing Credits Number, Installment Plans_no...	0.342	0.205	0.205	0.599415	2.923977	0.134890	1.984599
42556	(Credit Duration_1.3870748387117167, Sex)	(Existing Credits Number, Installment Plans_no...	0.252	0.205	0.205	0.813492	3.968254	0.153340	4.262553
42557	(People Number, Sex)	(Existing Credits Number, Installment Plans_no...	0.252	0.205	0.205	0.813492	3.968254	0.153340	4.262553
42558	(Credit Duration_1.3870748387117167, People Nu...	(Existing Credits Number, Installment Plans_no...	0.342	0.205	0.205	0.599415	2.923977	0.134890	1.984599
42559	(Existing Credits Number)	(Installment Plans_none, Installment Rate, Cre...	0.342	0.205	0.205	0.599415	2.923977	0.134890	1.984599
42560	(Installment Plans_none)	(Existing Credits Number, Installment Rate, Cr...	0.814	0.252	0.205	0.251843	0.999376	-0.000128	0.999790
42561	(Installment Rate)	(Existing Credits Number, Installment Plans_no...	0.342	0.205	0.205	0.599415	2.923977	0.134890	1.984599
42562	(Credit Amount)	(Existing Credits Number, Installment Plans_no...	0.342	0.205	0.205	0.599415	2.923977	0.134890	1.984599
42563	(Personal Status)	(Existing Credits Number, Installment Plans_no...	0.690	0.205	0.205	0.297101	1.449275	0.063550	1.131031
42564	(Present Residence)	(Existing Credits Number, Installment Plans_no...	0.342	0.205	0.205	0.599415	2.923977	0.134890	1.984599

Figure 24: The subset by left

7 Conclusion

To help banking institutions predict the assessment of the client's credit risk by applied various machine learning techniques such as classification,clustering and frequent pattern mining. This paper is the starting point for further learning and research activity. There are dozens of ways and methods to improve obtained results like using balanced data for training models, another features preparation etc.However, with more practice and knowledge, better results will be performed.

References

[1] Maryam Zangeneh:Customer credit risk assessment using artificial neural networks.In: IJ Information Technology and Computer Science, pp.5866. (2016)

[2] HierarchicalClustering,https://pub.towardsai.net/fully-explained-hierarchical-clustering-with-python-ebb256317b50.Last access 20 May 2021.

[3] AprioriAlgorithm.Know How to Find Frequent Itemsets,https://pyshark.com/market-basket-analysis-using-association-rule-mining-in-python.Last accessed 20 May 2021.