Report

Samir Aliyev

December 2020

1 Abstract

In this report results of bank data set will be presented. Differences between the results of different classification and clustering is to be shown.

2 Introduction

Data is everywhere. Nowadays most businesses uses their data for making decision for future. One of the field is bank sector. For example, like what this report is about, banks uses their data to make decision about to which client they can give credit and to which not. They obtain it by classify them as good or bad customer. This paper will shows the results obtained using classification algorithm to predict new customers as good or bad. The other main aim is to maximize the accuracy of classifying good customer as much as possible. Because when "bad" customer is classified as "good", it may be lost for bank whereas classifying "good" as "bad" don't effect bank in considerable way.

3 EDA

This data set is consist of 1000 rows with 20 feature. 700 of them are classified as good and remaining 300 are bad. There's no missing values on data set.One of the feature is either the customer has a telephone or not. But if we consider that this data set is belong to late 20th century and having a telephone was not common then, we will not use this column for classification due that having a telephone didn't effect the if customer is good or bad.

Out of remaining 19 columns,7 are numeric data and 12 are categorical data. Correlation matrix result is:

	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
Duration of the credit in month	1	0.624984	0.0747488	0.0340672	-0.0361364	-0.0112836	-0.0238345
Credit amount in EUR	0.624984	1	-0.271316	0.0289263	0.0327164	0.0207946	0.0171422
Installment rate in percentage of disposable income	0.0747488	-0.271316	1	0.0493024	0.0582657	0.0216687	-0.0712069
Present residence of the customer since (in years)	0.0340672	0.0289263	0.0493024	1	0.266419	0.0896252	0.0426434
Age of the customer in years	-0.0361364	0.0327164	0.0582657	0.266419	1	0.149254	0.118201
Number of existing credits of the customer at this bank	-0.0112836	0.0207946	0.0216687	0.0896252	0.149254	1	0.109667
Number of people the customer being liable to provide maintenance for	-0.0238345	0.0171422	-0.0712069	0.0426434	0.118201	0.109667	1

From correlation heat map, we can see there's little to no correlation between variables except between credit amount and duration of credit. Which was expected that the higher credit amount takes longer time to be paid back.
4 categorical columns are mostly monotonic that is more than 80% of data fall in same category namely, other payment plans, is customer foreign worker and other parties. These columns will we used for classification and clustering but will be removed in frequent pattern mining analysis since they are mostly same.

4 Preprocessing

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. Data was discretized already. So no more discretization technique applied.

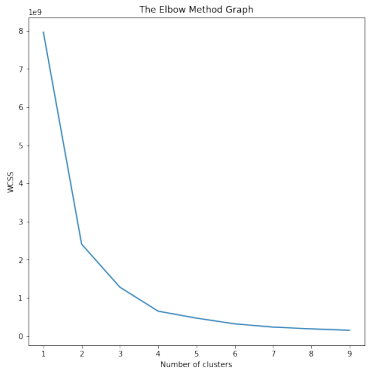
5 Clustering

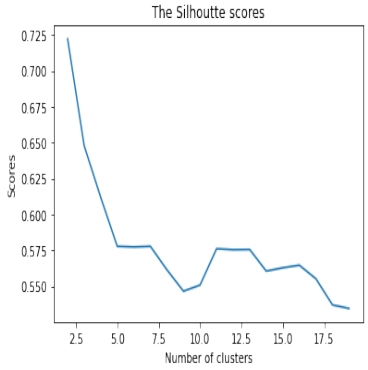
Clustering is an unsupervised Machine Learning method which is used to find similirities between dataset's objects. For this dataset 2 different algorithms, namely, KMeans and Agglomerative clustering have been used.

5.1 KMeans

K-means is one of the most popular unsupervised machine learning algorithm thanks to its simplicity and effetviness. K-means is an iterative algorithm that

tries to split the dataset into K pre-defined distinct non-overlapping clusters where each data point belongs to only one group.
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.

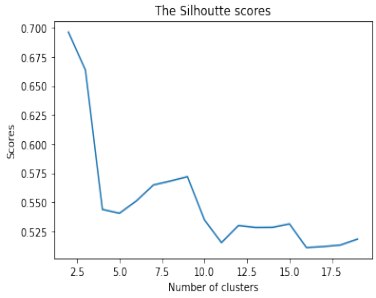




It's hard to tell what is the number of clusters from elbow method. It can be 2 or 4. It is because the clusters on dataset are not well defined.
Silhouette score takes its maximum value when number of clusters is equal to 2.

5.2 Agglomerative clustering

Agglomerative clustering, unlike KMeans, is hierarchical clustering technique. It starts treating each data point as a cluster. Then it merges all "clusters" into big cluster until the number of clusters is one. The result is tree-like object called dendogram
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.



labels of cluster	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
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

Anaylisys show that the main difference between clusters are the Credit amounts and duration of credit. Both credit amount and duration of credit are condiderably higher for second cluster than first one.

6 Classification

Classification is supervised Machine Learning method which is used to predict the class of new data based on labeled dataset.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 , DecisionTrees and Random Forest algorithms have been used for classification.

6.1 Logistic regression

Logistic regression is probability based classification algorithm. It assigns each data object value between 0 and 1. Based in these values classes are chosen. Cross Validation is used for getting average accuracy.

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

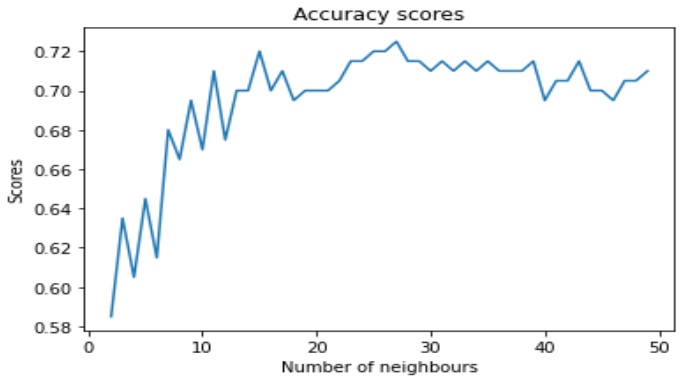
Which is about 74.5% accuracy. Accuracy score didn't change when normalizing applied to data.

	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.

6.2 KNN

KNN is vote-based supervised ML algorithm which can be used for both regression and classification. In classification, the class is defined by the how close is the data point to its neighbors. Number of neighbors that is taken into consideration k is the parameter.
Cross Validation scored for different number of k was used.




```
27
0.725
```

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

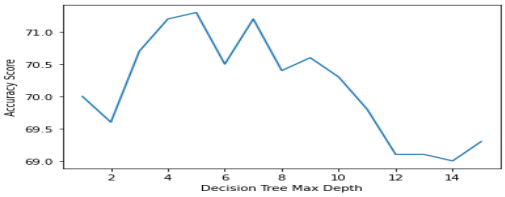
	precision	recall	f1-score	support
0	0.55	0.11	0.18	56
1	0.74	0.97	0.83	144
accuracy			0.73	200
macro avg	0.64	0.54	0.51	200
weighted avg	0.68	0.72	0.65	200

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

6.3 Decision Tree Classifier

Decision tree classifier is another classification algorithm which uses decision tree to observe the item using branches to conclude item's target class. Main parameter is the depth of decision tree. It represents how many layers our decision tree will have.
Cross Validation scored for different number of depth was used.

```
Best accuracy score |71.3| achieved at max depth |5|
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=5,
                       max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort=False,
                       random_state=None, splitter='best')
```



Accuracy achieved was 71.3 which is less than what is achieved by both KNN and Logistic regression.

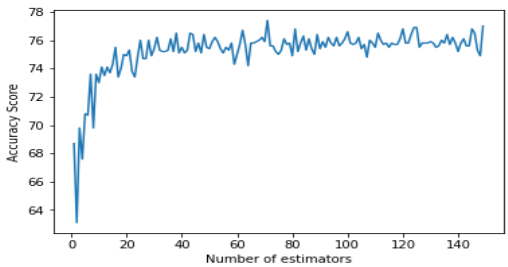
	precision	recall	f1-score	support
0	0.45	0.34	0.39	56
1	0.77	0.84	0.80	144
accuracy			0.70	200
macro avg	0.61	0.59	0.59	200
weighted avg	0.68	0.70	0.69	200

By looking at classification report, we can clearly state that Decision Tree didn't work well in this data set. Precision score was worse than previous two as well.

6.4 Random Forest Classifier

Random Forest is ensemble ML algorithm. Random forest is consist of large amount decision trees. Each decision tree return a class label (for classification problem) and the class with the most vote become the prediction of the model. This algorithm, most of the time, performs much better than decision trees. The main parameter in this algorithm is decision trees. Different numbers of decision trees(number of estimators) in range between 1 to 150 was tried.

```
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.

7 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. It is very frequently applied technique especially in classification, clustering, market basket analysis. There are several algorithms which are used to get "association rules". In this data set Apriori algorithm was used. Apriori algorithm starts with individual patterns and then extend them into larger sets where these patterns appear in data set sufficiently frequent. One of the main advantage of apriori algorithm is that it is unsupervised algorithm, thus we don't need labels. And also it is exhaustive algorithm . It finds all rules with certain confidence.

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
3705	(Property owned by the customer_car or other, ...	(Housing situation of the customer_owning)	0.064	0.713	0.062	0.968750	1.358696	0.016368	9.184000
3465	(Personal status and sex of the customer_male ...	(Housing situation of the customer_owning)	0.063	0.713	0.061	0.968254	1.358000	0.016081	9.040500
254	(Housing situation of the customer_accommodati...	(Property owned by the customer_unknown/no pro...	0.108	0.154	0.104	0.962963	6.253006	0.087368	22.842000
1409	(Savings account/bonds of the customer_X06 < 1...	(Property owned by the customer_unknown/no pro...	0.067	0.154	0.064	0.955224	6.202752	0.053682	18.894000
1612	(Personal status and sex of the customer_male ...	(Property owned by the customer_unknown/no pro...	0.085	0.154	0.081	0.952941	6.187930	0.067910	17.977500

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.

8 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 , Decision tree classifier and Random Forest Classifier algorithms applied to data. Hyperparameter optimization used for all except Logistic Regression. Each method evaluated based on their precision score, recall score and cross validation score. Random Forest Classifier gave best results (77.4%) 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.