

Final Project: Loan Prediction

Group 4

ABSTRACT

The overall objective is to produce loan predictions by using clustering for insight generation and error analysis and to study the similar clusters and classes and to compare them; analyze pros and cons of algorithms, generate and communicate the insights.

AUTHORS:

Ifiok Charles (0300041771)

Kalada Reagan Tuwotamuno (8489637)

Gayatri (300095631)

Stevy Kuimi (8845324)

SUPERVISOR:

Prof. Arya Rahgozar

GNG 5125 X

Contents

| | | |
|------|-----------------------|----|
| I. | Introduction | 5 |
| II. | Data Preparation..... | 5 |
| III. | Clustering..... | 10 |
| IV. | Classification..... | 11 |
| V. | Error Analysis | 14 |
| | References..... | 15 |

| | | |
|----------------|---|-----------|
| <i>Table 1</i> | <i>Loan probabilities.....</i> | <i>7</i> |
| <i>Table 2</i> | <i>Histograms.....</i> | <i>8</i> |
| <i>Table 3</i> | <i>Accuracy table.....</i> | <i>11</i> |
| <i>Table 4</i> | <i>Confusion Matrices</i> | <i>11</i> |
| <i>Table 5</i> | <i>Metric records.....</i> | <i>12</i> |
| <i>Table 6</i> | <i>10-Fold Cross validation.....</i> | <i>12</i> |
| <i>Table 7</i> | <i>Cross-Validation of the champion and Challenger.....</i> | <i>13</i> |

| | |
|---|----|
| Figure 1: loan probability 1 | 5 |
| Figure 2 Loan probability 3 | 6 |
| Figure 3 Loan probability 2 | 6 |
| Figure 4 Loan probability 4 | 7 |
| Figure 5 Loan probability 5 | 7 |
| Figure 6 Loan probability 6 | 7 |
| Figure 7 Loan probability 7 | 7 |
| Figure 8 Loan amount-vs-Applicant Income | 7 |
| Figure 9 Histogram of loan amount | 8 |
| Figure 10 Histogram of income | 8 |
| Figure 11 Boxplot 1 | 8 |
| Figure 12 Boxplot 2 | 8 |
| Figure 13 Boxplot 3 | 8 |
| Figure 14 Correlation map | 9 |
| Figure 15 Random Forest features importance | 9 |
| Figure 16 Elbow curve | 10 |
| Figure 17 PCA plot with all feature | 10 |
| Figure 18 Random Forest confusion Matrix | 11 |
| Figure 19 SVM confusion matrix | 11 |
| Figure 20 Decision Tree confusion matrix | 11 |
| Figure 21 KNN Confusion Matrix | 11 |
| Figure 22 Metrics report Random Forest for important features | 12 |
| Figure 23 Metrics report SVM for important features | 12 |
| Figure 24 Metrics report Decision Tree for important features | 12 |
| Figure 25 Metrics report KNN for important features | 12 |
| Figure 26 Cross validation using the important features | 12 |
| Figure 27 Cross validation using all features | 12 |
| Figure 28 Random Forest cross validation | 13 |
| Figure 29 SVM cross validation | 13 |
| Figure 30 K-means with important features | 14 |

I. Introduction

Machine learning is a set of techniques that have proven to be very useful in almost all industries, be it media, medicine, engineering and finance. Banking institute, just like many other institutions are becoming more and more dependent on machine learning techniques to generate customers insight and future customer forecast (Li, et al., Nanjing). These predictions help the banks to analyze their clients and approve loans applications accordingly (Noy & McGuiness, 2001). To reduce risks, they are not recourse to data analyst, to help the using machine learning techniques. As such the aim of this project is to perform loan approval prediction based on several attributes. The data set used was taken from (Dubey, 2017).

II. Data Preparation

We started by performing the exploratory data analysis. We started by comparing the 11 attributes against each other to see if there was any correlation between them. Firstly, it was observed that the credit history was strongly correlated with the loan status. Indeed, it was observed that customers with credit history had higher chances of getting the loan than those with no credit history independently of other factors. For example, on Figure 1, it is noted that the probability of getting a loan with a credit history is way higher than that of no credit history, approximately 80% against 10%. It was also noticed on the same figure that, being married and/or graduate also slightly increase the probability of loan approval.

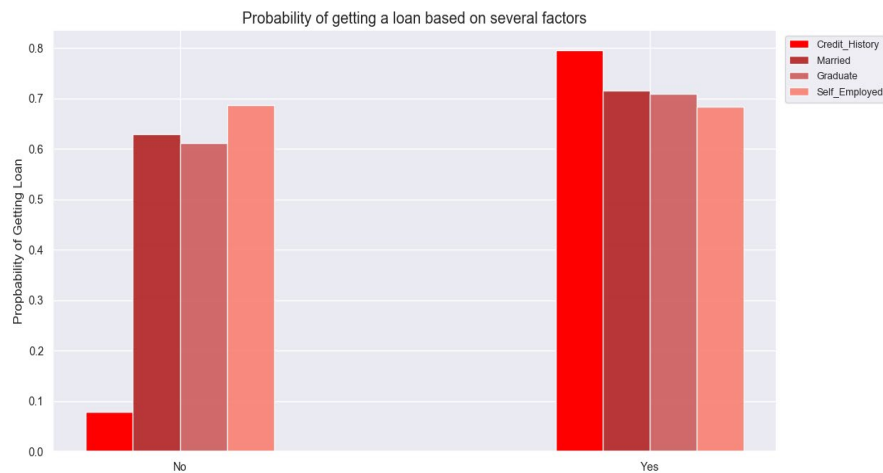


Figure 1: loan probability 1

Secondly, it was also found out that, being married and having a credit history, had more weight than being graduated or self-employed with a credit history as displayed on Figure 2 below.

Thirdly, it was also strangely noticed that couples with 2 children had more chances of getting

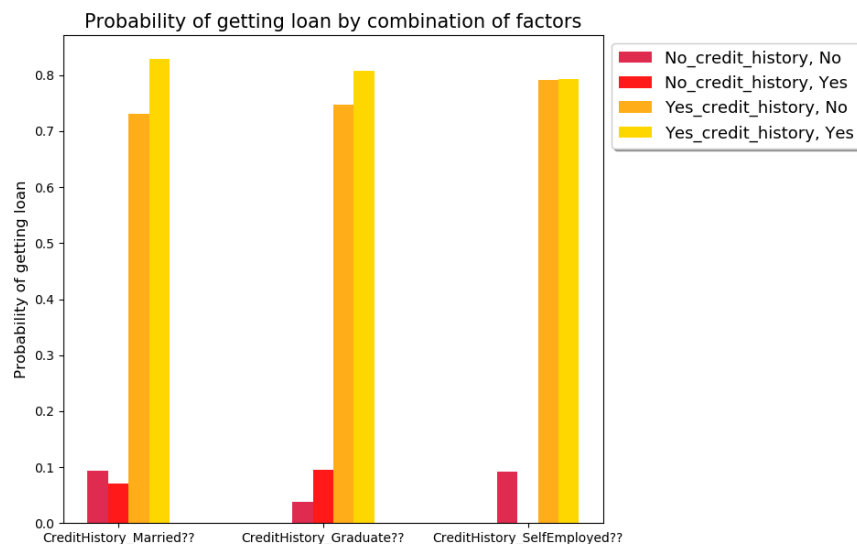


Figure 3 Loan probability 2

It was taken from Figure 8 that applicants with income between \$20000-70000 are more likely to get a positive response

independently of other factors, than those with an income between \$0-20000. This means that depending on your income, some attributes are taken in consideration or not. The loan term was proven not to be a reliable attribute to use alone in predicting loans because its data is not abnormally distributed. After more analysis, it figured out that this is the result of uneven sample. For example, from Figure 6, it I shown that loan terms of 12, 60 and 120 months have the highest loan probability, after checking the sample numbers, it was noticed that, there was only one applicant with a loan term of 12 months, 2 for 60 and 3 for 120. These numbers are small to base any decision on. On the other hand, we can see that, the longer the loan term, the shorter the probability to get a loan. This is logical since, financial situations can easily change with time. It was noted that the loan amount and income data can be reliable source of prediction since, the distribution is normalized as shown on Figure 9 and Figure 10.

the loan than those with less or more children (Figure 3) and people living in semi-urban areas had more chances of getting approved than those living in rural or urban areas as shown on Figure 4, while those living in urban areas will more chance of being successful than those in rural areas. Gender do not matter in loan prediction (Figure 5) since the data shows that male and female applicant have equal chances.

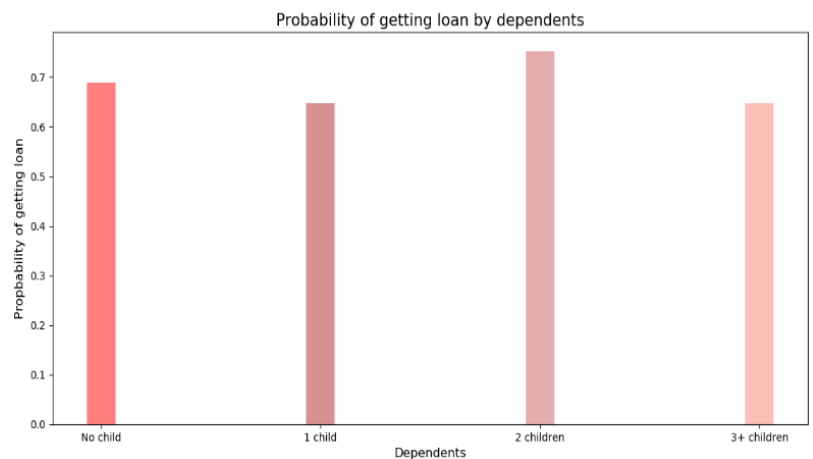


Figure 2 Loan probability 3

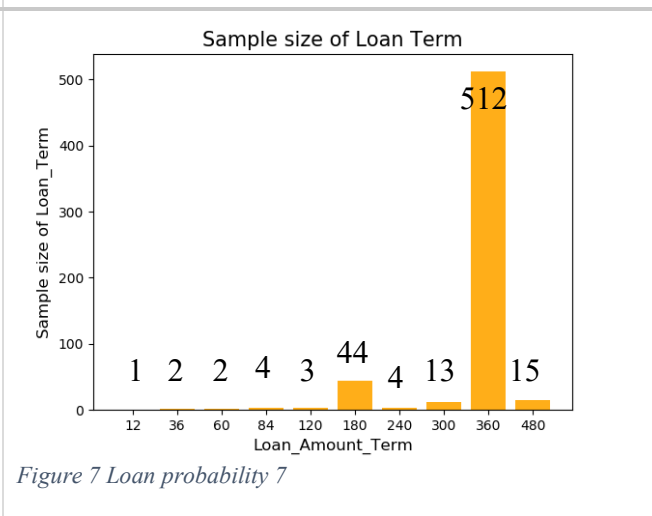
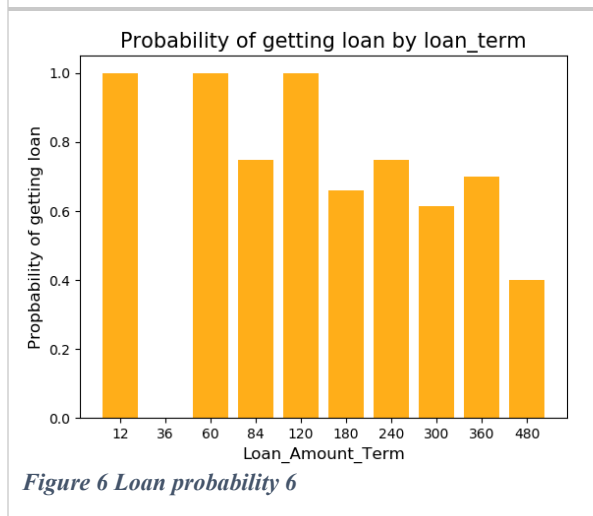
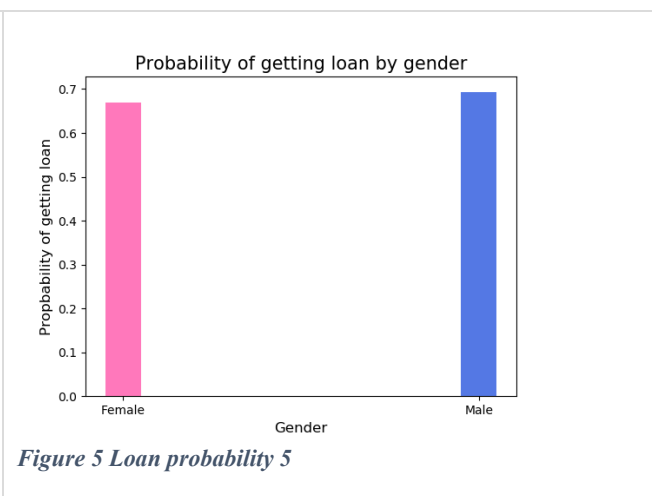
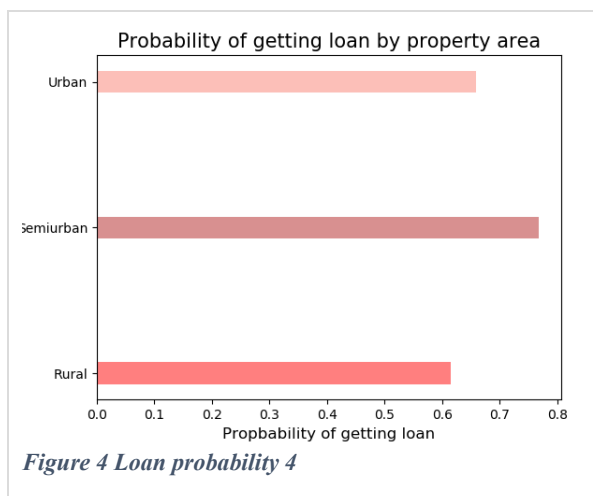


Table 1 Loan probabilities

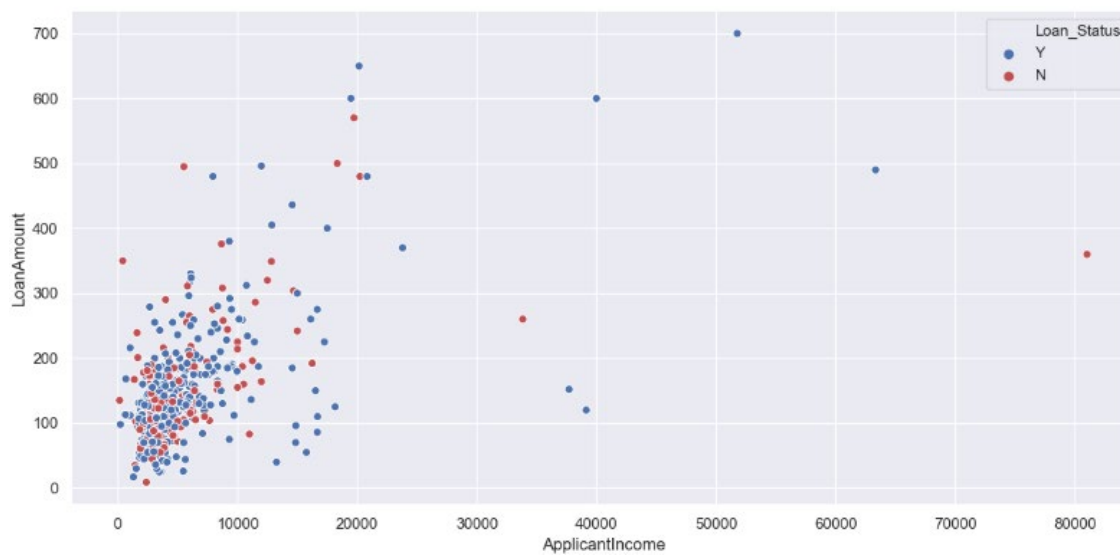


Figure 8 Loan amount-vs-Applicant Income

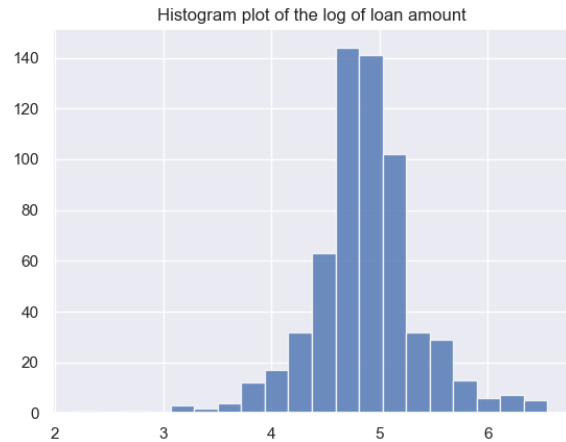


Figure 9 Histogram of loan amount

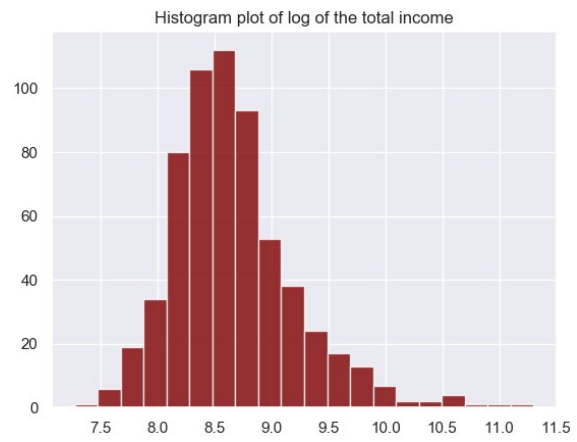


Figure 10 Histogram of income

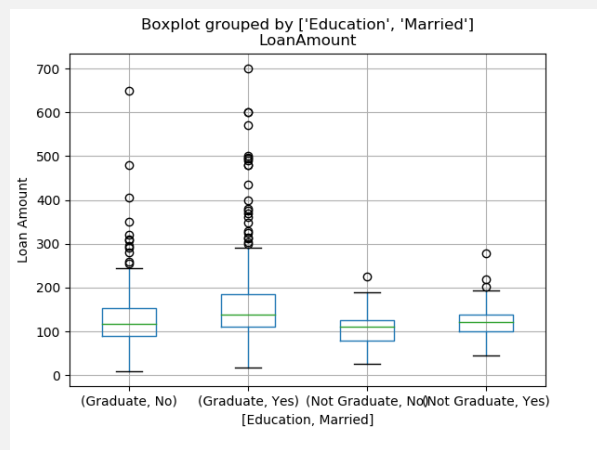


Figure 11 Boxplot 1

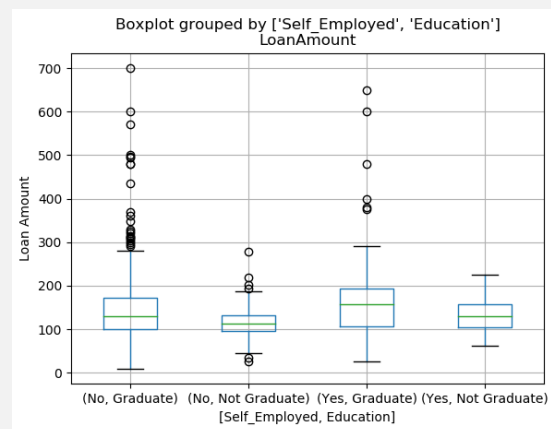


Figure 12 Boxplot 2

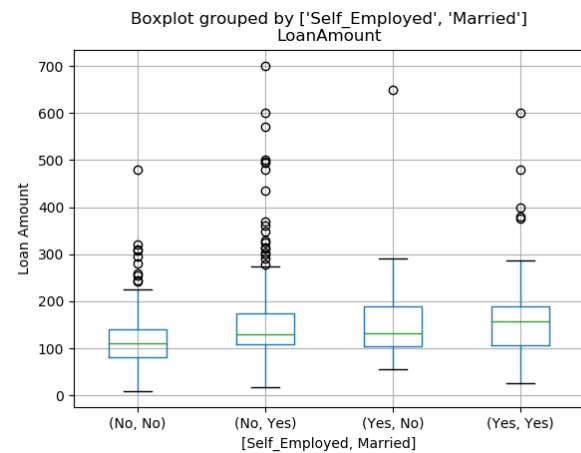


Figure 13 Boxplot 3

Table 2 Histograms

We used boxplots of three attributes to observe the data in general and it was found out that the data was nor normally distributed, so we used the log function to normalize the data. the class of

customers who are not married and not graduates and customers who married but not graduated, have a constant loan amount, with few or none outliers, this means that their loan amount is constant within a range Figure 11. Similar observation was made for non-self-employed/not graduate and self-employed and graduate customers (Figure 12), as well as for self-employed single customers and self-employed and married customers (Figure 13). These could indicate that education has influences the loan amount in a sense that graduate customers ask for very little loan amount. Also, self-employed people tend to be more constant or more realistic with their loan application that non-self-employed customers. As conclusion of our data exploratory analysis, it was taken that, the loan could be predicted using only few features: Loan amount, loan amount term, credit history as displayed on the correlation map on Figure 14 and confirmed by the random forest methods of finding these important feature on Figure 15.

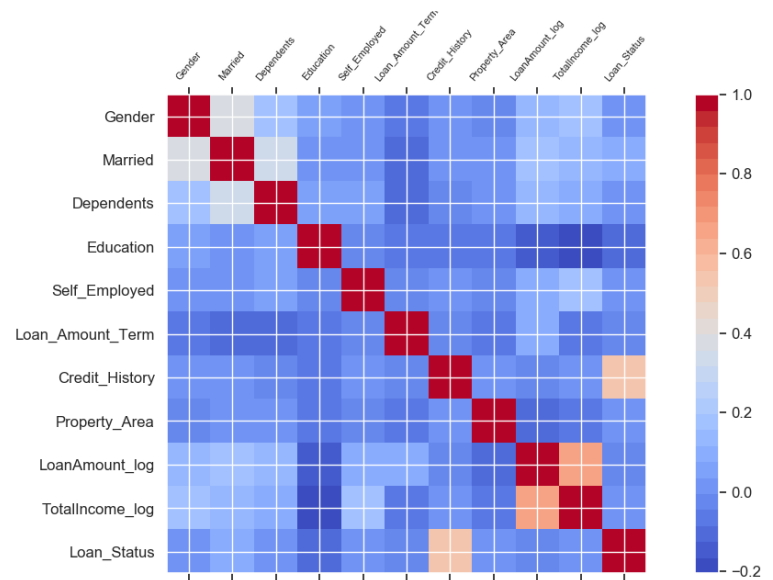


Figure 14 Correlation map

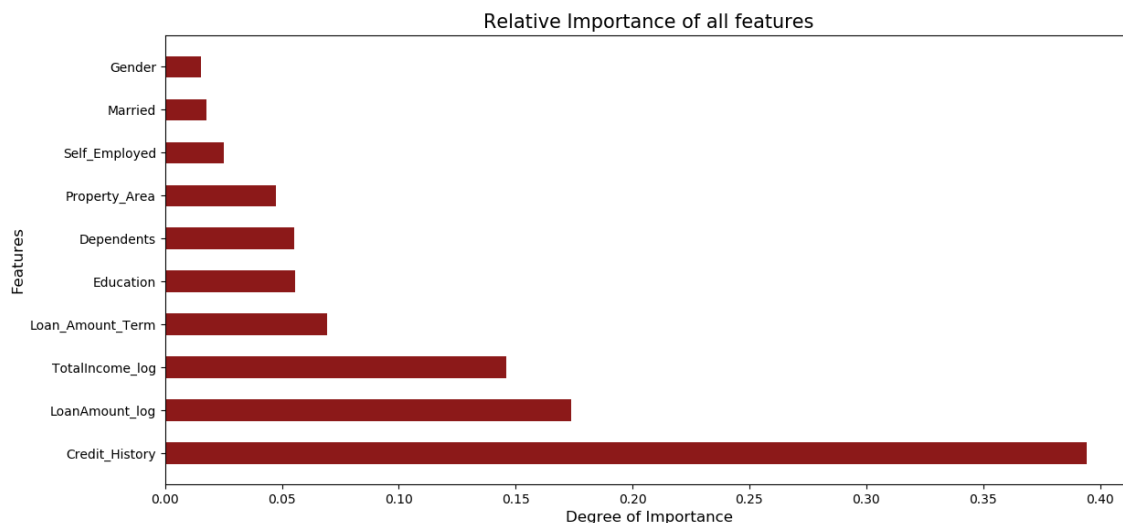


Figure 15 Random Forest features importance

III. Clustering

The algorithm used for clustering is K-means. This is the most used form of unsupervised machine learning technique for clustering. It has proven to be very fast and simple. K represent the number of clusters and these are the steps of the algorithm:

1. Select the number of clusters, K
2. Take each point and find the nearest centroid
3. Match each point to the closest centroid again
4. Repeat until the clusters cannot be improved anymore.

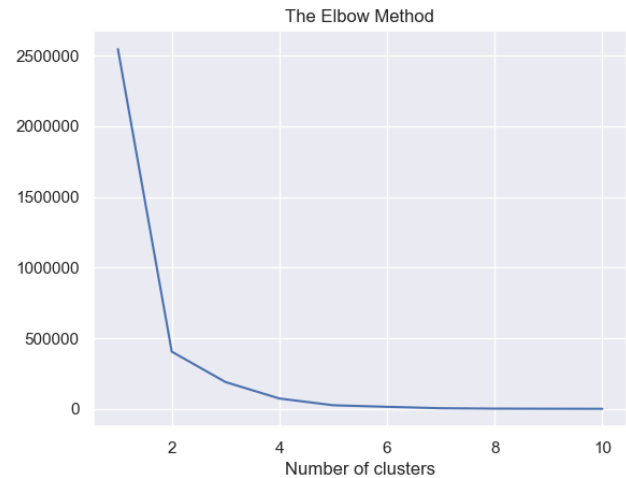


Figure 16 Elbow curve

We used the elbow curved to determine the number of clusters to use. On Figure 16, it was taken that the right number of clusters is 5. Following up to that, we started by drawing the PCA plot with all features as shown on Figure 17.

- The green cluster represent in majority the following features: gender, loan amount term, credit history, marriage
- The yellow cluster: gender, dependent, self-employed, credit history
- The grey cluster: gender, married, credit history, property area

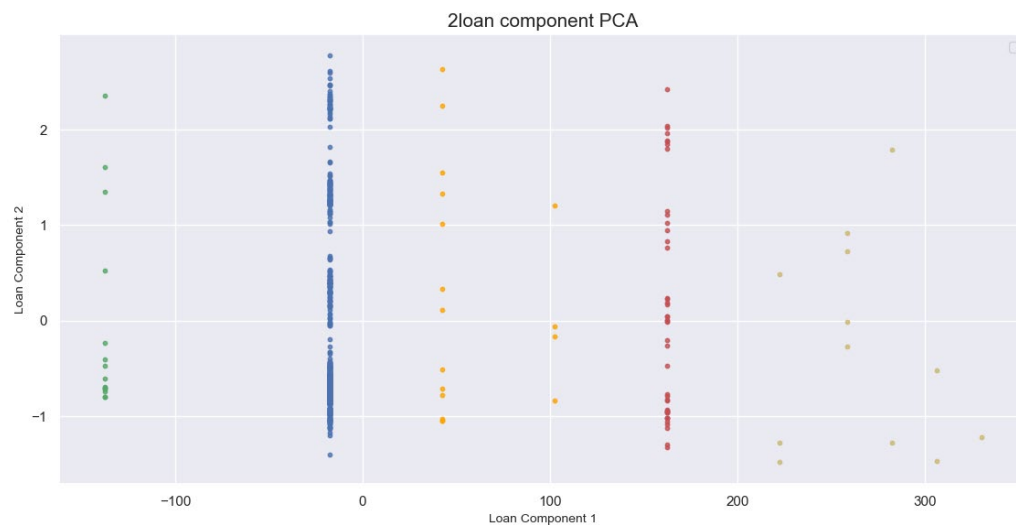


Figure 17 PCA plot with all feature

IV. Classification

For the classification exercise we implemented four algorithms:

- Random Forest,
- SVM,
- Decision tree,
- KNN

We first run the experiment for all the algorithms with all features. Then we performed 10-fold cross validation (Figure 27), we picked the champion and the challenger. We did it a second time, but only with the most important features, ran the 10-fold cross validation again (Figure 26) and picked the new champion and challenger. Table 4 displays the confusion matrices of the four algorithms while Table 5 displays the metric records of each algorithm. We can see that SVM has the highest average F1 score followed by Random Forest. Decision Tree is shown to have the lowest score. Table 3 concludes with the summary of accuracy of all experiences. It can be observed that Random Forest is the champion for the runs with all the features with SVM as challenger. When running with only the most important features, SVM become the champion and Random Forest the challenger.

| | All features | Most important features |
|---------------|--------------|-------------------------|
| Random Forest | 0.804 | 0.8013 |
| SVM | 0.793 | 0.809 |
| Decision Tree | 0.713 | 0.698 |
| KNN | 0.727 | 0.7655 |

Table 3 Accuracy table

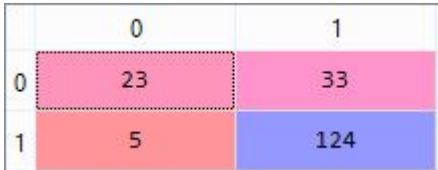
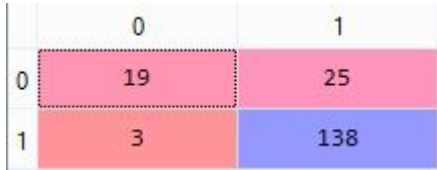
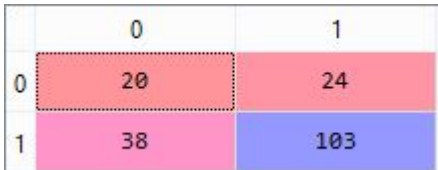
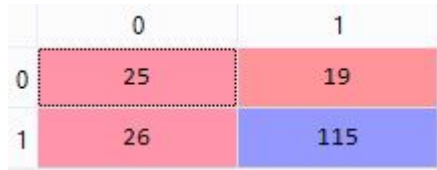
| | |
|---|--|
|  <p>Figure 18 Random Forest confusion Matrix</p> |  <p>Figure 19 SVM confusion matrix</p> |
|  <p>Figure 20 Decision Tree confusion matrix</p> |  <p>Figure 21 KNN Confusion Matrix</p> |

Table 4 Confusion Matrices

| | precision | recall | f1-score | support |
|-------------|-----------|--------|----------|---------|
| 0 | 0.82 | 0.41 | 0.55 | 56 |
| 1 | 0.79 | 0.96 | 0.87 | 129 |
| avg / total | 0.80 | 0.79 | 0.77 | 185 |

Figure 22 Metrics report Random Forest for important features

| | precision | recall | f1-score | support |
|-------------|-----------|--------|----------|---------|
| 0 | 0.86 | 0.43 | 0.58 | 44 |
| 1 | 0.85 | 0.98 | 0.91 | 141 |
| avg / total | 0.85 | 0.85 | 0.83 | 185 |

Figure 23 Metrics report SVM for important features

| | precision | recall | f1-score | support |
|-------------|-----------|--------|----------|---------|
| 0 | 0.34 | 0.45 | 0.39 | 44 |
| 1 | 0.81 | 0.73 | 0.77 | 141 |
| avg / total | 0.70 | 0.66 | 0.68 | 185 |

Figure 24 Metrics report Decision Tree for important features

| | precision | recall | f1-score | support |
|-------------|-----------|--------|----------|---------|
| 0 | 0.49 | 0.57 | 0.53 | 44 |
| 1 | 0.86 | 0.82 | 0.84 | 141 |
| avg / total | 0.77 | 0.76 | 0.76 | 185 |

Figure 25 Metrics report KNN for important features

Table 5 Metric records

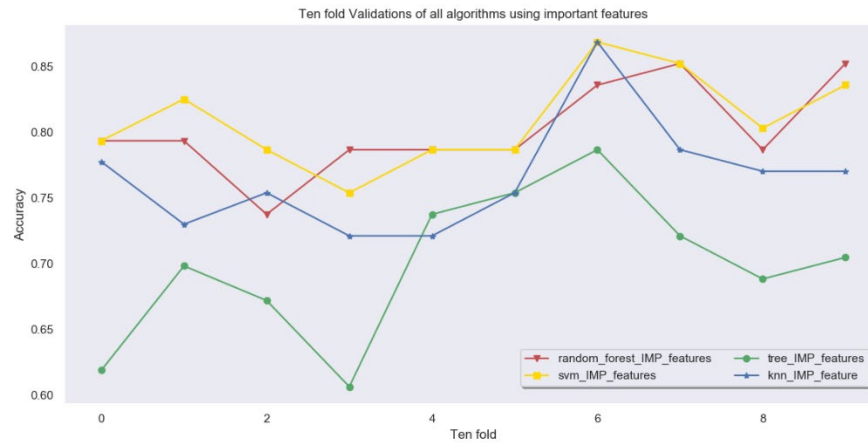


Figure 26 Cross validation using the important features

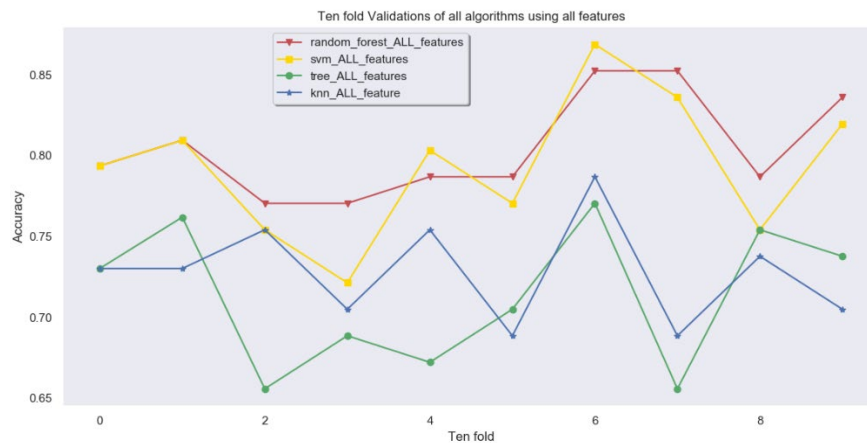


Figure 27 Cross validation using all features

Table 6 10-Fold Cross validation

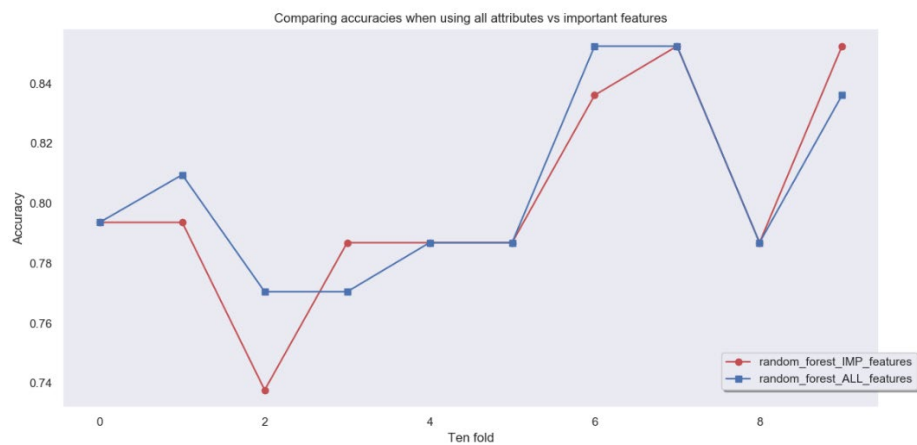


Figure 28 Random Forest cross validation

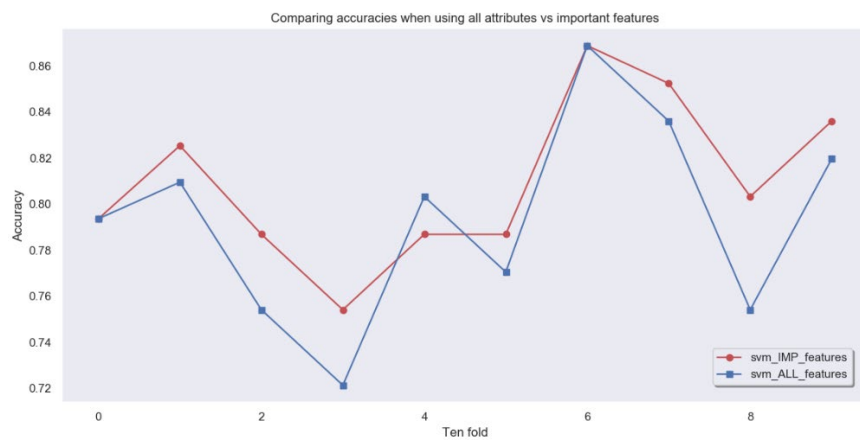


Figure 29 SVM cross validation

Table 7 Cross-Validation of the champion and Challenger

V. Error Analysis

Figure 30 shows the cluster plot when we used only the important features. This figure is meant to be a 2-cluster plot but as we can see, there are some anomalies that makes it difficult for the K-means algorithm to split the dataset into two clusters. This may be due to the lack of correlation between the important features.

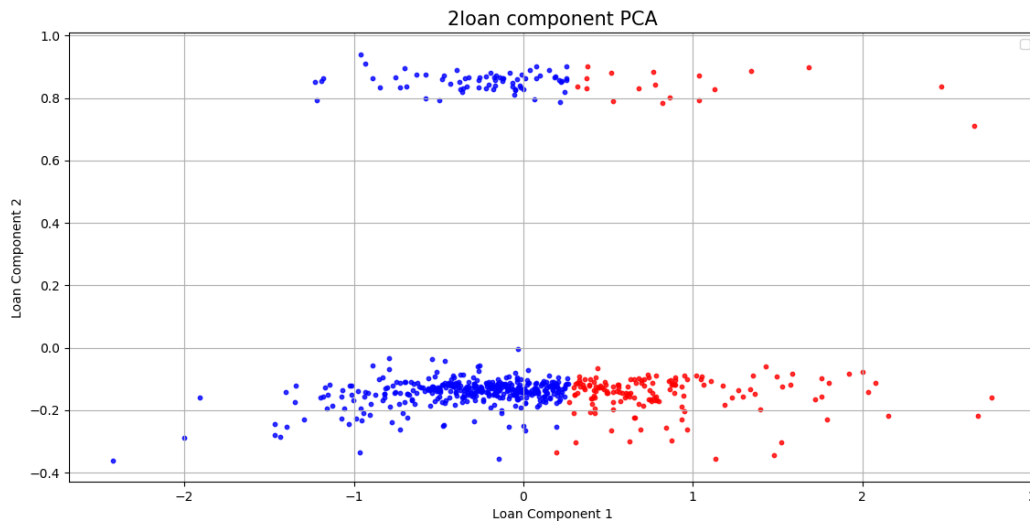


Figure 30 K-means with important features

References

- Dubey, V., 2017. *Kaggle*. [Online]
Available at: <https://www.kaggle.com/vikasdubey/loan-prediction/output>
- G., A. & C., S., 1959. *Prediction of Loan Status in commercial bank using machine learning classifier*. s.l., IEEE Xplore.
- Gruber, T. R., 1993. *Knowledge Acquisition*. Stanford, Elsevier, pp. 199-220.
- Kumar, A., Gary, I. & Kaye, S., 2016. Loan Approval Prediction based on machine learning approach. *IOSR-JCE*, pp. 18-21.
- Li, X. et al., Nanjing. *Overdue Prediction Of Bank Loan Based on LSTM-SVM*. 2018, IEEE.
- Noy, N. F. & McGuinness, D. L., 2001. *Ontology Development 101: A Guide to Creating Your First Ontology*, Stanford Knowledge Systems Laboratory Technical Report: Stanford.
- Turkson, R. E., Baagyere, E. Y. & Wenya, G. E., 2016. A Machine Learning Approach for Predicting Bank Credit Worthiness. *IEEE*.