LOAN REPAYMENT PREDICTION

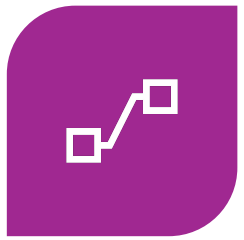
USING MACHINE LEARNING

TABLE OF CONTENTS

INTRODUCTIO
N

DATA
ANALYSIS

MODELLING



CONCLUSION

REFERENCES

Introduction

The loan is one of the most important products of the financial institutes. All the institutes are trying to figure out effective business strategies to persuade more customers to applytheir loans. However, there are some customers are not able to pay off the loan after their application are approved

The dataset that used in this paper is from Lending Club, a website that connects borrowers and investors over the Internet. It includes 9,578 observations that were fundedthrough the Lending Club.com platform between May 2007 and February 2010

Data Analysis

Standardization of datasets is a common requirement for many machine learning models. Itis possible that the machine learning algorithm behave badly if the individual features do not more or less look like standard normally distributed data. Therefore, in order to improve the result of the prediction model. We will standardize the data first

The main advantage of standardization scaling is to avoid attributes in greater numericranges dominating those in smaller numeric ranges. Another advantage is to avoid numericaldifficulties during the calculation

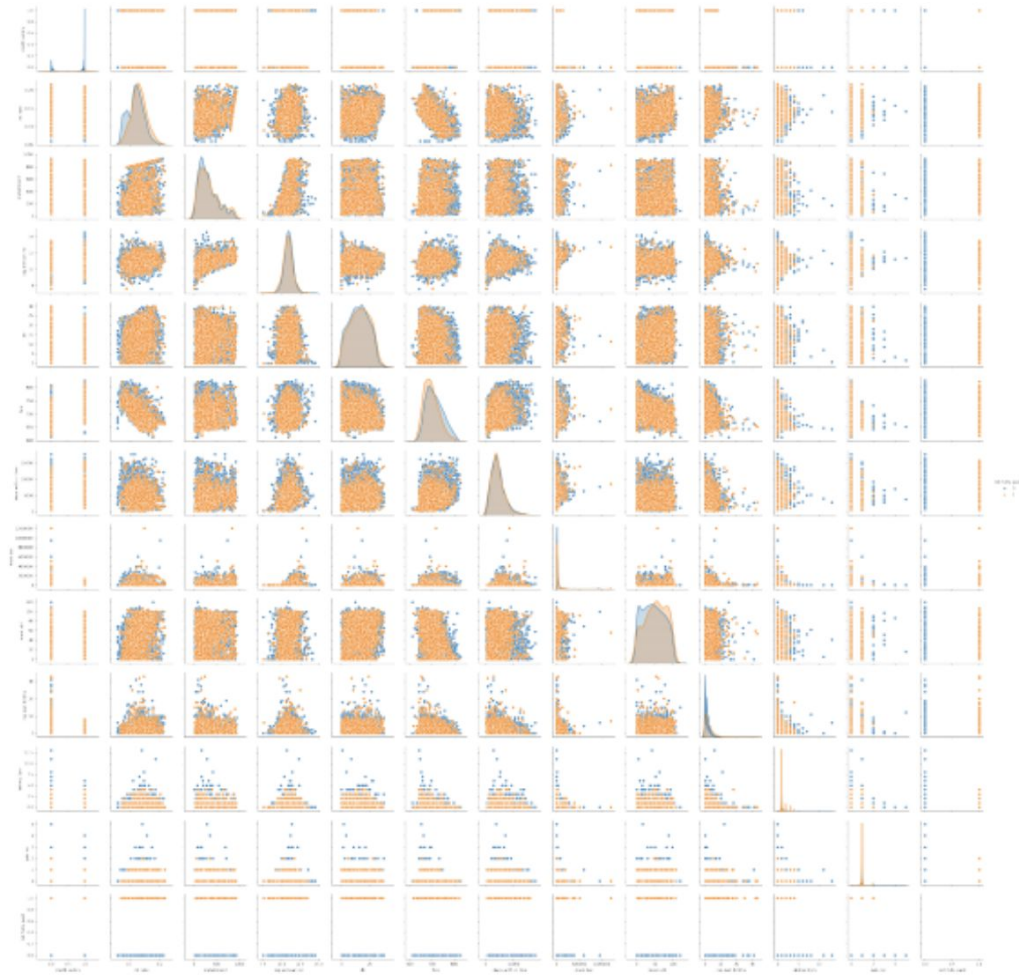


Figure 2.1: Density plot for numerical variables

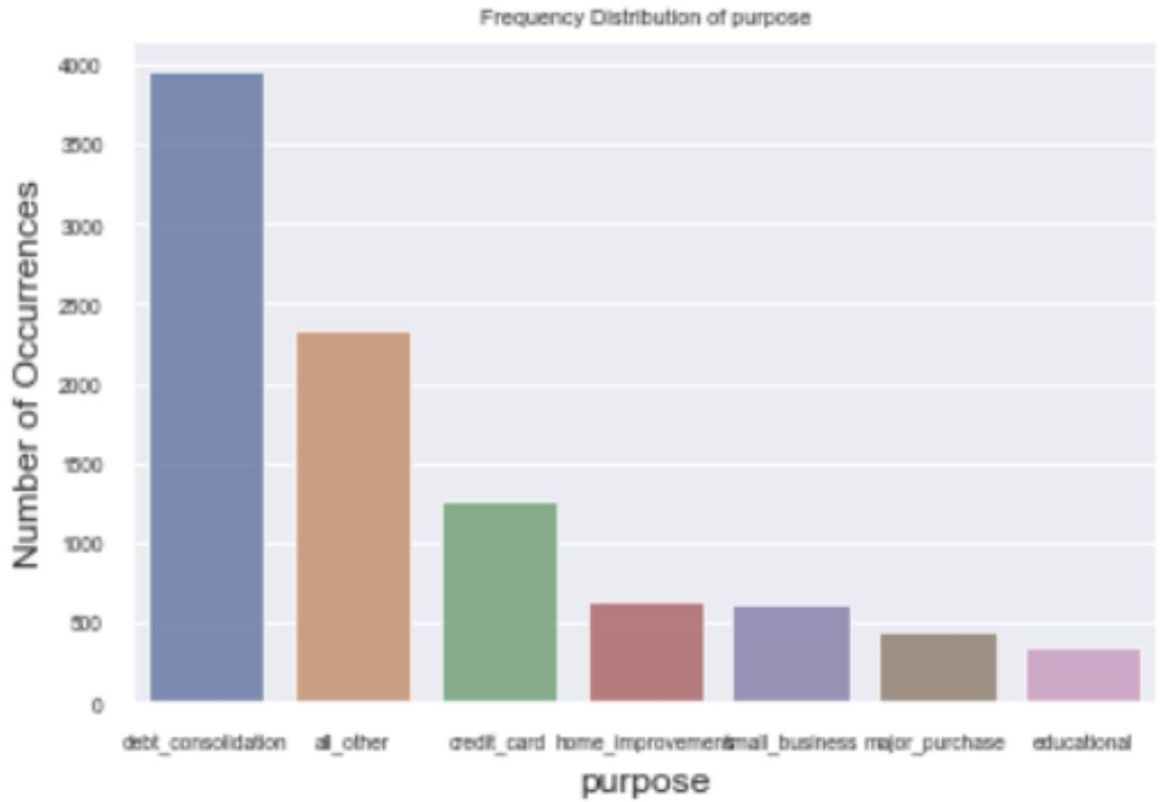


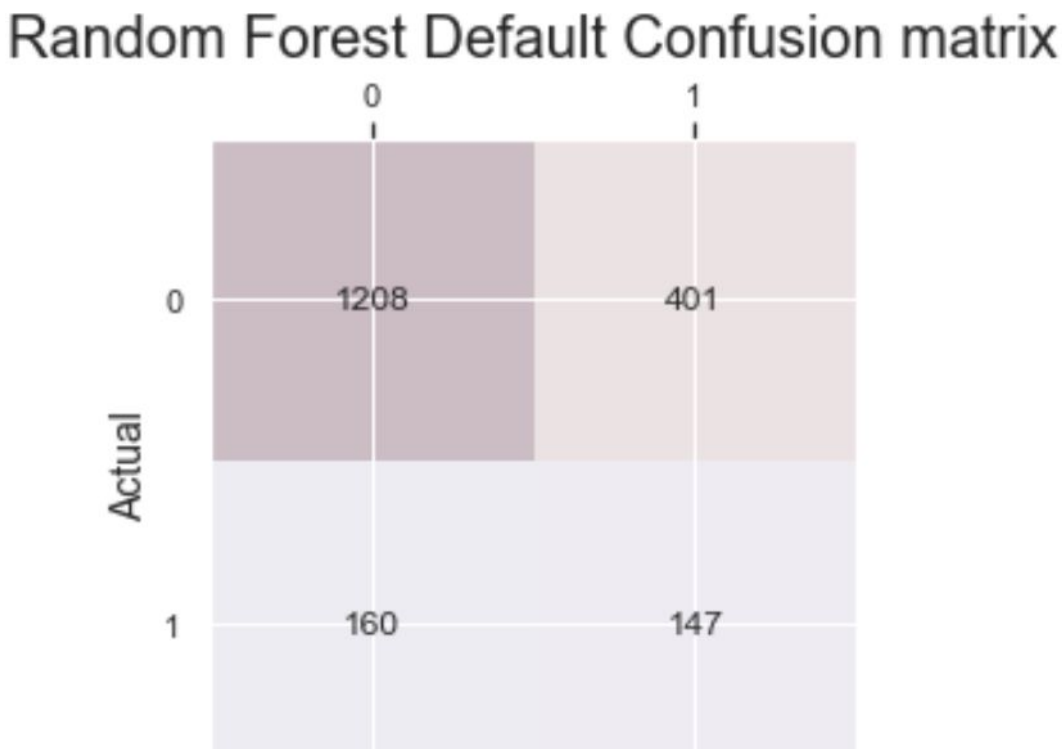
Figure 2.2: Count plot for categorical variables

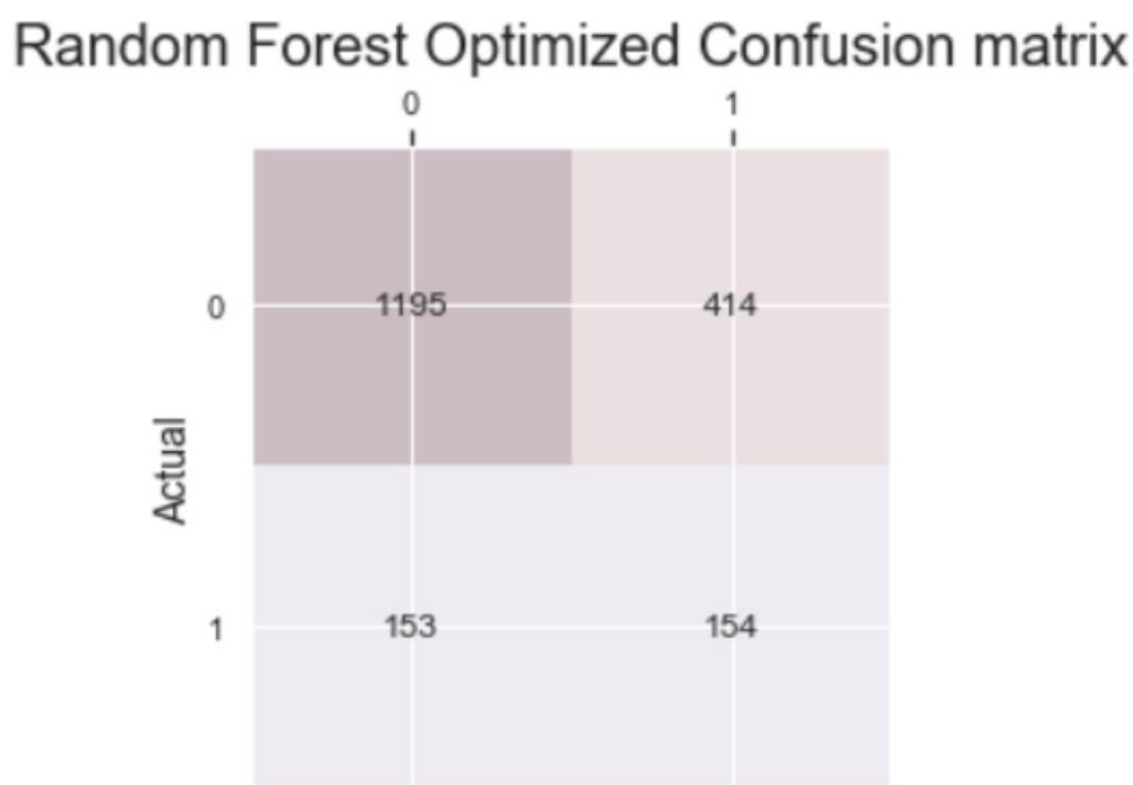
Modelling

- In machine learning, the classification task described above is commonly referred to as super-vised learning. In supervised learning there is a specified set of classes, and example objects are labeled with the appropriate class. In this case, the goal is to identify if the lender would be able to repay the loan

Precision=TruePositives/(TruePositives+FalsePositives)

Recall=TruePositives/(TruePositives+FalseNegatives)


Random Forest Default Confusion matrix
0
1
Actual
0
1
1208
401
160
147


Random Forest Optimized Confusion matrix
0
1
Actual
0
1
1195
414
153
154

Conclusion

Nowadays, the loan business becomes more and popular, and many people apply for loans for various reasons. However, there are cases where people do not repay the bulk of the loan amount to the bank which results in huge financial loss. Hence, if there is a way that can efficiently classify the loaners in advance, it would greatly prevent the financial loss .

In this study, the dataset was cleaned first, and the exploratory data analysis and feature engineering were performed. The strategies to deal with both missing values and imbalanced data sets were covered. Then we propose four machine learning models to predict if the applicant could repay the loan, which are Random Forest, Logistic Regression, SupportVector Machine, and K-Nearest Neighbours

REFERENCES

- Naomi S Altman. “An introduction to kernel and nearest-neighbour non parametric regression .
- Niklas Donges. “The Random Forest Algorithm.”Statistical Methods