Credit Loan Prediction

Rakamin x ID/X Partners VIX End-to-end

Dhoifullah Luth Majied

Introduction

This is my project data science in notebook. Let me tell u a bit about this project. This project is the final project for the Data Scientist intern at ID/X Partners, this project data come from a Lending club. In this project, I have to build a model that can credit loan predict using a dataset provided by the company which consists of data on received and rejected loans.

In this project, I need to understand what is credit risk, and the process of how loans are being issued. Not all data in this dataset can be used to make a prediction.

In general, this kernel will follow the way of Dr Ronen Meiri (Founder and CTO of DMWay Analytics)

you can read the detail of the document below : https://dmway.com/wp-content/uploads/2017/11/2017.06-WhitePaper_FINAL-Design-DMWAY.pdf

Problem

We are investors in P2P platforms. There are loans that get charged off in the end. If a loan get charged off or defaulted, we will lost our money. We want to prevent that, and minimize our loss.

Objective

To build a model that can predict credit risk for loan applicants, based on Loan dataset.

Outcome

- Improved accuracy of credit risk predictions: By building a model to predict credit risk, the company can improve the accuracy of their lending decisions.
- Faster lending decisions: With a model to predict credit risk, the company can automate part of the lending decision-making process.
- Increased efficiency: Automating part of the lending decision-making process can also increase efficiency, as it reduces the need for manual review of every loan application.

Goal

The goal of this project is to build a machine learning model to predict the probability that a loan will charge off. We will attempt to only use data available to customers via the data loan, including information about the borrower and the loan listing (the loan amount, loan purpose, loan grade, interest rate, installment, etc.). Such a predictive model could help customers make better-informed investment decisions. We will only consider loans that LendingClub accepted under its credit underwriting policy.

Business Understanding

- Credit risk is a concern for the company: The fact that the company wants to build a model to predict credit risk suggests that credit risk is a concern for the company. This is likely because the company wants to avoid lending money to individuals who are unlikely to pay it back, as this could result in financial losses.
- The model will help the company make lending decisions: The model being built will likely be used to inform lending decisions. By predicting credit risk, the model will help the company determine whether to approve or reject loan applications, and may also inform the terms of the loan (such as interest rate).

Table of Contents

Data Exploration

The Data contains complete loan data for all loans issued through the 2007-2014, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the “present” contains complete loan data for all loans issued through the previous completed calendar quarter. Additional features include credit scores, number of finance inquiries, address including zip codes, and state, and collections among others. The dataset consists of 466,285 rows and 74 columns consisting of categorical and numerical data.

Data Preparation

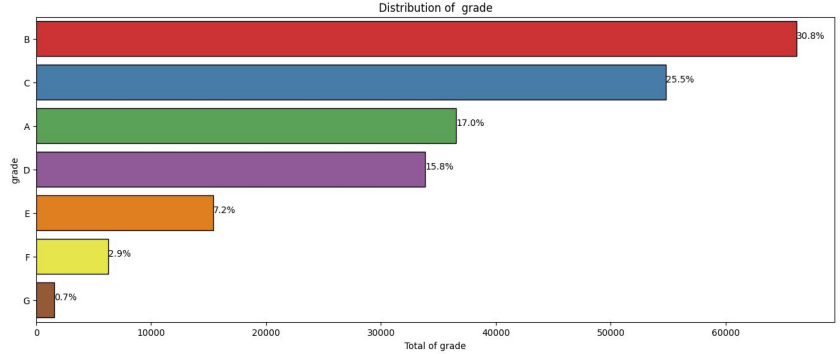
Cleansing

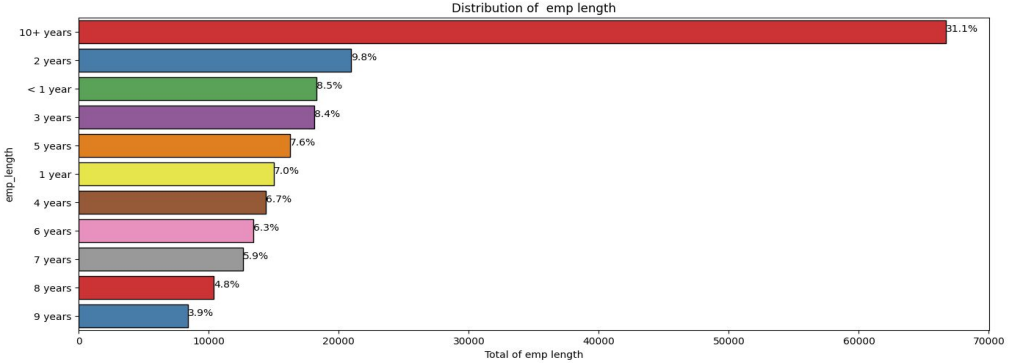
- Drop features with null values > 15%
- Define target predict labels from the data
- Impute missing value from numerical features with a media
- Processing outliers from numerical features
- Drop unnecessary feature

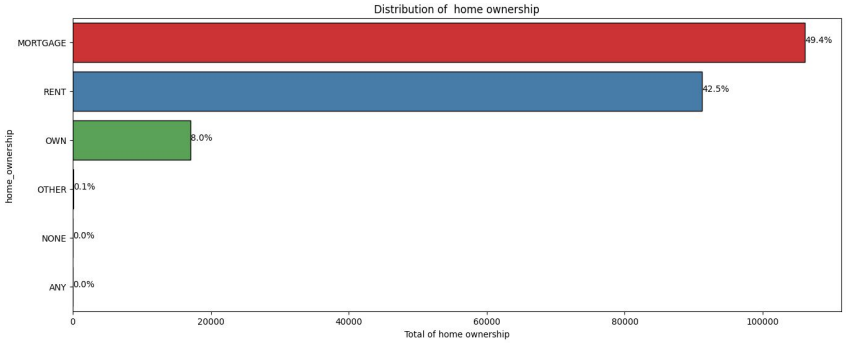
Feature Enginering

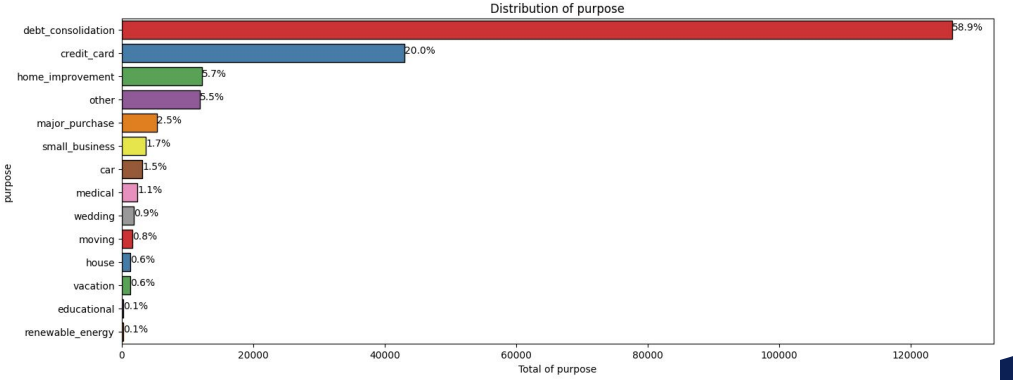
- Encode all categorical feature with one hot encoding
- Split data into train and test data
- Perform feature scaling on numerical data
- Normalization of data with min-max scaler
- Process the target data imbalance that will be predicted using the supervised learning

Data Visualization

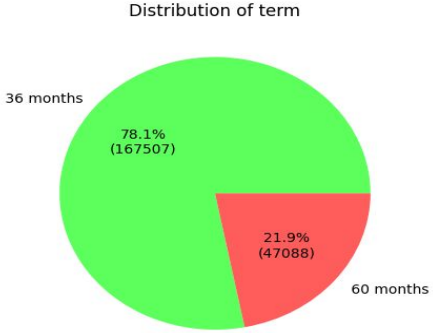


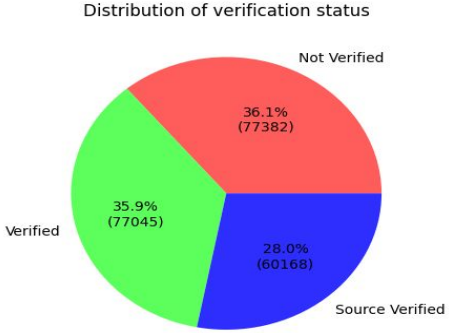


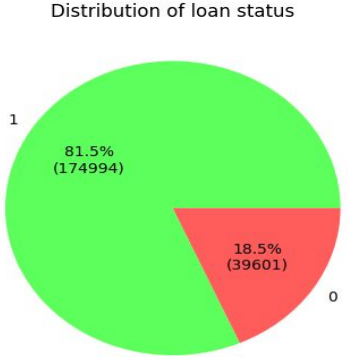


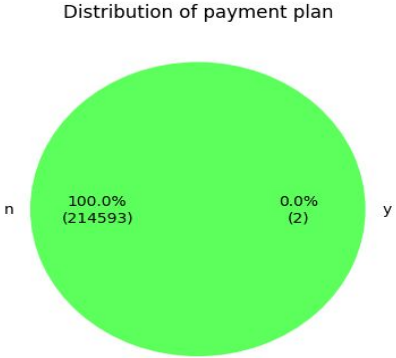


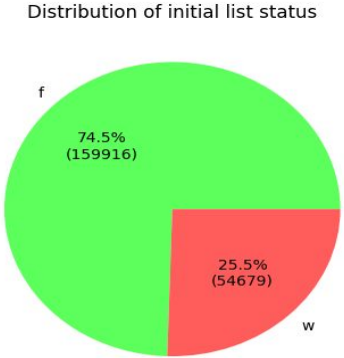
Data Visualization



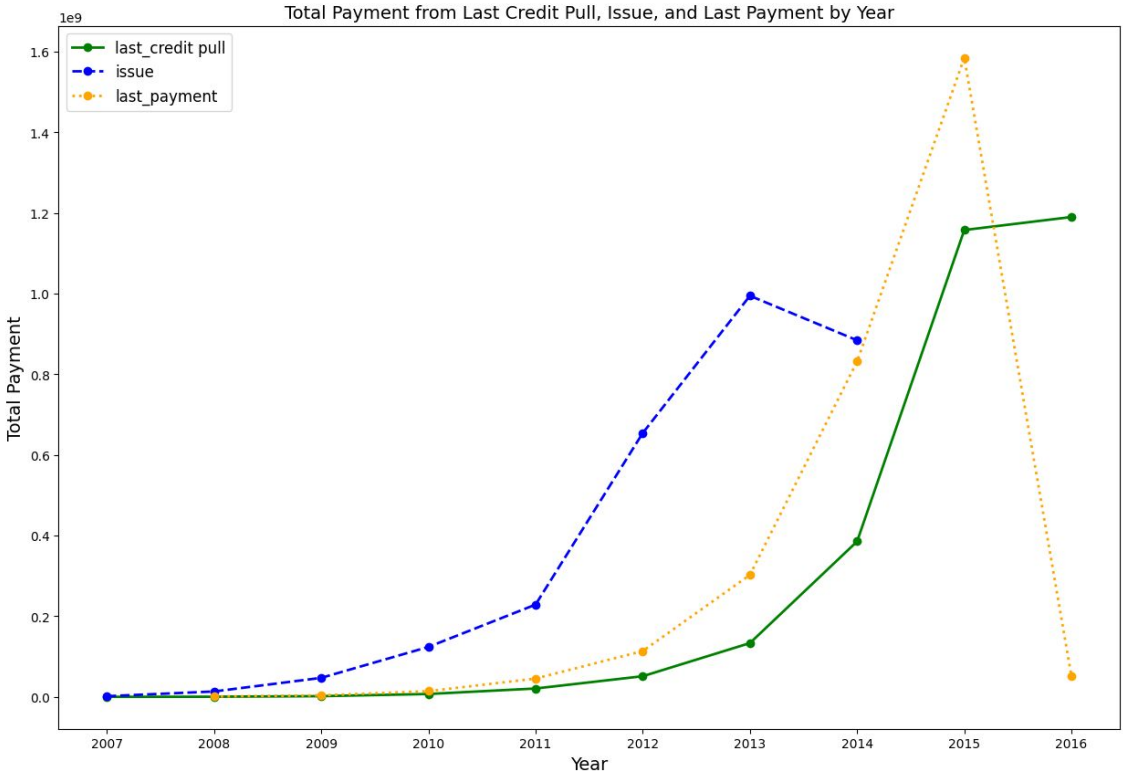






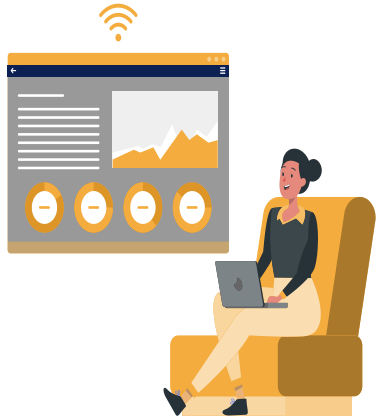


Data Visualization



Data Modeling

Model	Percent %
KNeighborsClassifier	96.5%
Logistic Regression	99.6%
Decision Tree	99.8%
Random Forest	99.8%
Naive Bayes MultinomialNB	81.5%
Naive Bayes GaussianNB	78.3%



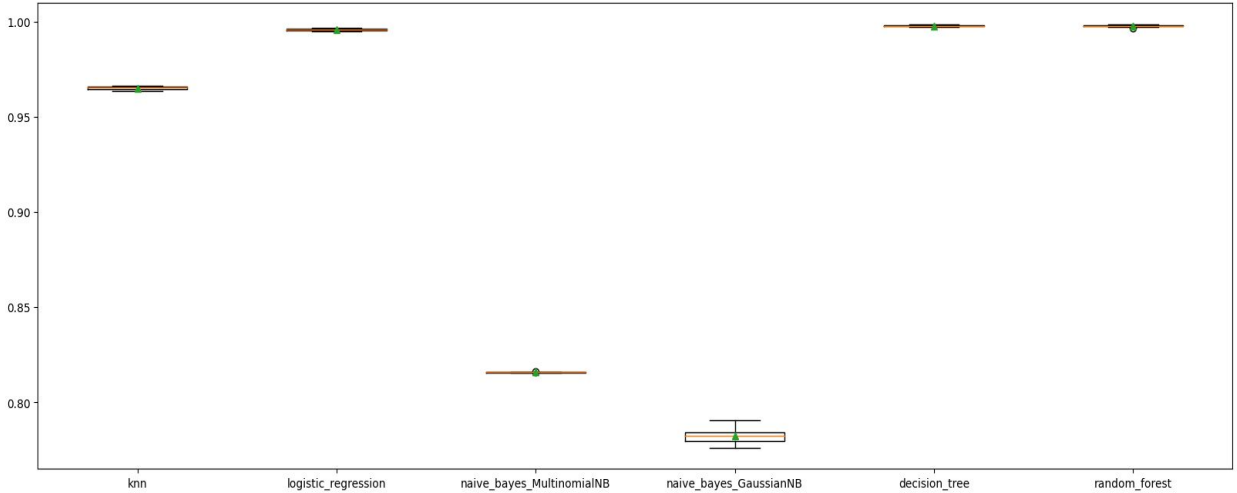
To build a system capable of predicting credit risk, we prefer to use 6 model (KNN, Logistic Regression, Naive Bayes MultinomialNB, Naive Bayes GaussianNB, Decision Tree, and Random Forest) by using 80 percent of the data as material for machine learning and the remaining 20 percent as test data to test the ability of the system to predict credit risk.

Result

The six modules that I use, I choose the three models that I apply, namely Logistic Regression, Decision Tree, and Random Forest algorithms to create credit risk, the following results are obtained. By this model it was decided to take the Decision Tree algorithm because the calculation model is more accurate and in the next process if you want to make an API from this model you can be sure that the search will be more precise.

Finally, it can be said that the developed method is quite accurate at predicting credit risk, which is of course very helpful for businesses to screen potential customers to reduce the danger of loss from clients who won't pay.

Result



Decision Tree	
Accuracy	99%
Precision	99%
Recall	99%
AUC	99%

Conclusion

Accuracy for each class (bad loans and good loans) is obtained with a very stable value after conducting a training model using oversampling data (the average accuracy for each class is > 70%). Therefore, it can be said that the model can distinguish bad loans from good loans quite effectively when oversampling is used during training.

Among all the models discussed above, the Decision Tree Classifier produces the best average accuracy results, with an average accuracy value of 99% (bad loan recall = 99% and good loan withdrawals = 100%) and the Random Forest Classifier produces the results the same best average accuracy, with an average accuracy value of 99% (bad loan recall = 99% and good loan withdrawals = 100%). The accuracy scores of the two models are the same, so I chose the Decison Tree Classifier model because it is computationally faster. Recall is calculated by dividing the "positive" number by the accurately anticipated "positive" number. This shows that the model accurately predicts 100% of total good loans and 99% of total bad loans.

Thank You

Contact Me :

dhoifullah.luthmajied95@gmail.com

https://www.linkedin.com/in/dhoifullahluthmajied/