

claudiatsai / Predict\_loan\_default Public


0 stars 0 forks

Star

Unwatch

- <> Code
- Issues
- Pull requests
- Actions
- Projects
- Wiki
- Security
- 


main

 claudiatsai 1123commit ... 4 minutes ago 5

[View code](#)

☰ README.md

# Loan Default Prediction



# Business Problem

---

Banks have a huge volume of applicants applying for loans. Some of the applicants do not have credit history or some might have very light credit score. It doesn't mean that lending money to them is highly risky so we should reject all of these applicants. This project uses data about personal loan. Our goal was to develop a model that could step by step explain the results of the model we built and what impact on the likelihood of the case falling into one of the binary categories (loans paid-off and charged). This model predict the 68 % of loan default and would be useful for the banks to make the best decision.

## Data

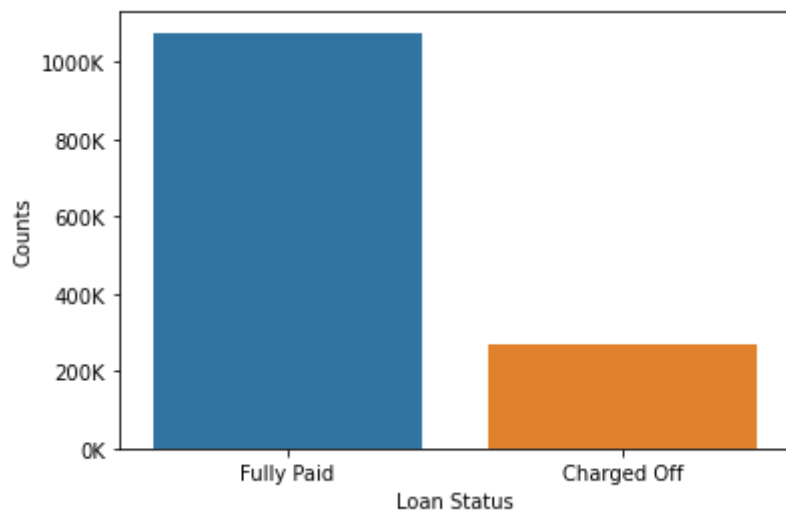
---

- LendingClub loan dataset
- Data range from 2007 to 2018
- 2.3M observations
- 151 features
- Loan status, interest rate, grade/subgrade, home ownership, annual income, etc
- <https://www.kaggle.com/datasets/wordsforthewise/lending-club>

## Methods

---

- Clean null values and outliers in the dataset
- Convert categorical variables into dummy variable for fitting machine learning models
- Select the features for modeling
- Due to the fact that the existing dataset is not balanced, which means that there are many more customers with clear loan status than customers who default, I used the sampling method to address this issue
- The datasize is large, so the major class was undersampled

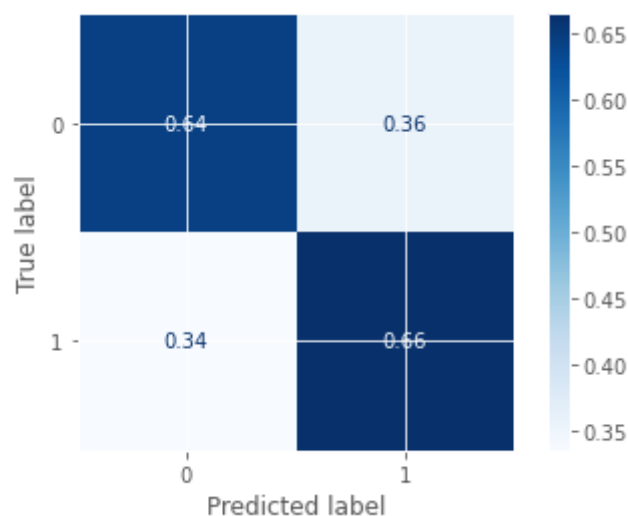


## Model

- Logistics Regression Model, Decision Tree Model, Random Forest Model and XGBoost Model were used to train the data
- XGBoost Model has recall score 68% and accuracy score 66%
- Tuning parameter in XGBoost Model to improve performance

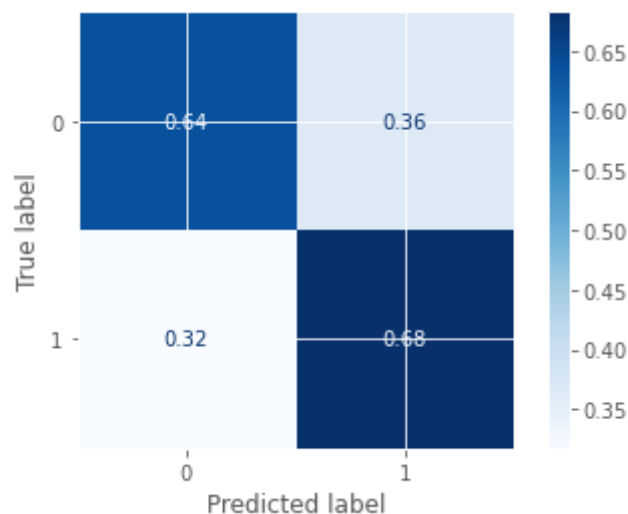
## Baseline Model

- Recall score 66%
- Accuracy score 65%

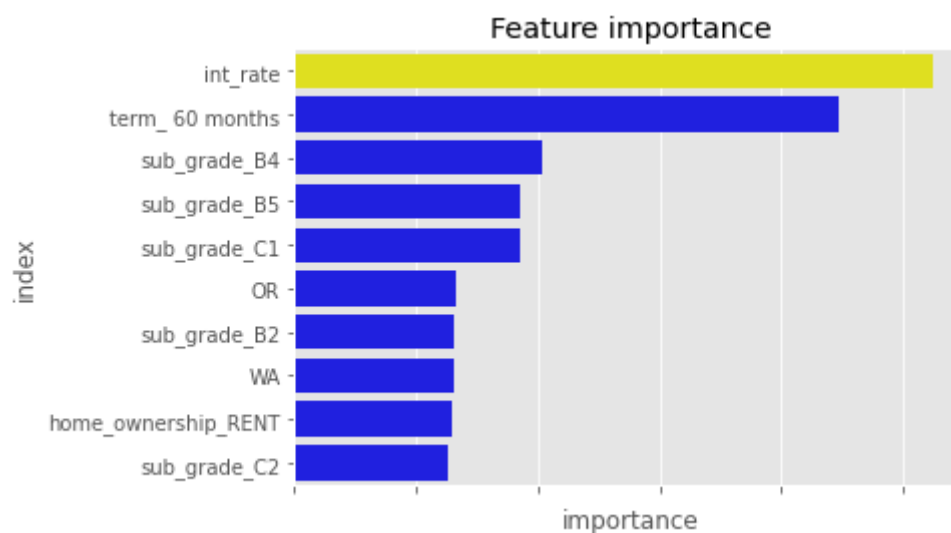


## Final Model

- Recall score 68%
- Accuracy score 66%



## Features Importance



## Conclusions

Interest Rate, term, subgrade, and home ownership affect the model prediction most.

Our model achieved 68% prediction on the test set.

From the confusion matrix, we can see our classifier has high recall. This means the proportion of borrowers predicted to default the loan is high.

# Future Improvements

---

- More classification models should be tried out
- Analyze the data by region or state to help banks to assess credit risk, provide accurate credit scores and make decisions on their loans in minutes after receiving each new incoming loan application
- Set up different threshold to improve recall score by business goal. It's because the binary classification models usually give the prediction of probability first and then assign the probabilities to 1 or 0 based on the default threshold of 0.5

# Repository Structure

---

```
├── README.ipynb
├── images
├── Lending Club_Presentation.pdf
├── .gitignore
├── Prediction_original.ipynb
├── README.md
├── Project_Proposal.pdf
├── notebook_project_loan_1.pdf
└── selected_features.csv
```

## Releases

No releases published

[Create a new release](#)

---

## Packages

No packages published

[Publish your first package](#)

---

## Languages

● Jupyter Notebook 100.0%

