VIX ID/X Partners

Credit Scoring Using Logistic Regression

Presented by Erlando Regita Febrian

Table of Contents

Background

As a data science intern at ID/X Partners, I was assigned to make a credit risk analysis from lending company. My objective here is to make a good prediction model that can clasify whether borrowers are good or not.

I used logistic regression as a model to predict credit risk and implemented weight evidence & information values to perfrom feature selection

Metrics used in this project are ROC-AUC score and KS-Statistic

My goals here are:

- build a prediction model with good ROC-AUC score (>0.8) and good KS-Statistic score (>0.45)
- build credit score (score card) each borrower and treshold recommendation list

Dataset Overview

Independent Feature

- Dataset contains of 75 Features and 466K records
- There are 16 features that contains 100% missing values (we dropped it)
- Some feature have missing values and no duplicated values
- We will perform feature engineering and drop unused features (e.g: id, member_id, url, zip, etc)

Target Feature

- loan_status is our target feature, it contains of 9 unique values
- we will convert it into binary feature with condition : Current, Fully Paid, and In Grace Period will be categorized as good loan and encoded as 1, the rest will be bad loan and it will be encoded as 0

Feature Engineering

- We dropped unused features, features that have 100% missing values, and also feature that has similar proportion each values by our target feature. So, total of dropped features are 30 features, so our dataset now only have 44 features left
- We converted target feature into binary feature
- we perform engineering on datetime feature (such as last_pymnt_d, next_pymnt, etc)

Data Preprocessing

01 Handling Missing Values

We imputed missing values with median and min (0 values), imputation method based on the feature condition

02 Drop Highly Correlated Features

We dropped 14 correlated independent features. We only drop 1 of each correlated features

03 Drop Quasi Constant Feature

We also dropped feature that only dominated by single values (pymnt_plan)

04 Cleaned Data

Summary, before feature selection, we have 27 features left that will be selected using weight of evidence (woe) and information value

Feature Selection

We will use Weight of Evidence and Information Value to select our features. The weight of evidence tells the predictive power of a every bin/category of a feature concerning its independent feature.

Information Value is a good measure of the predictive power of a feature and it also helps point out the suspicious feature. We perform feature selection with these following steps:

- Bin each categorical and numeric features
- We have to make sure that information value each feature follow the rule of thumb.
- If not, we will drop those features

Feature Selection

Before implementing weight of evidence and information value method, we have 27 features consist of 4 categorical features and 23 numeric features.

We combined features that have similar weight of evidence and dropped those who didn't follow rule of thumb.

Selected Features :

After implementing selection method, now we have dataset that consist of 466k records and 85 features

Modeling

We used logistic regression to predict whether the borrower will be a good borrower or not. We also used AUC score and KS-Statistic as our evaluation metric. After performing hyperparameter tuning and feature selection, the results are :

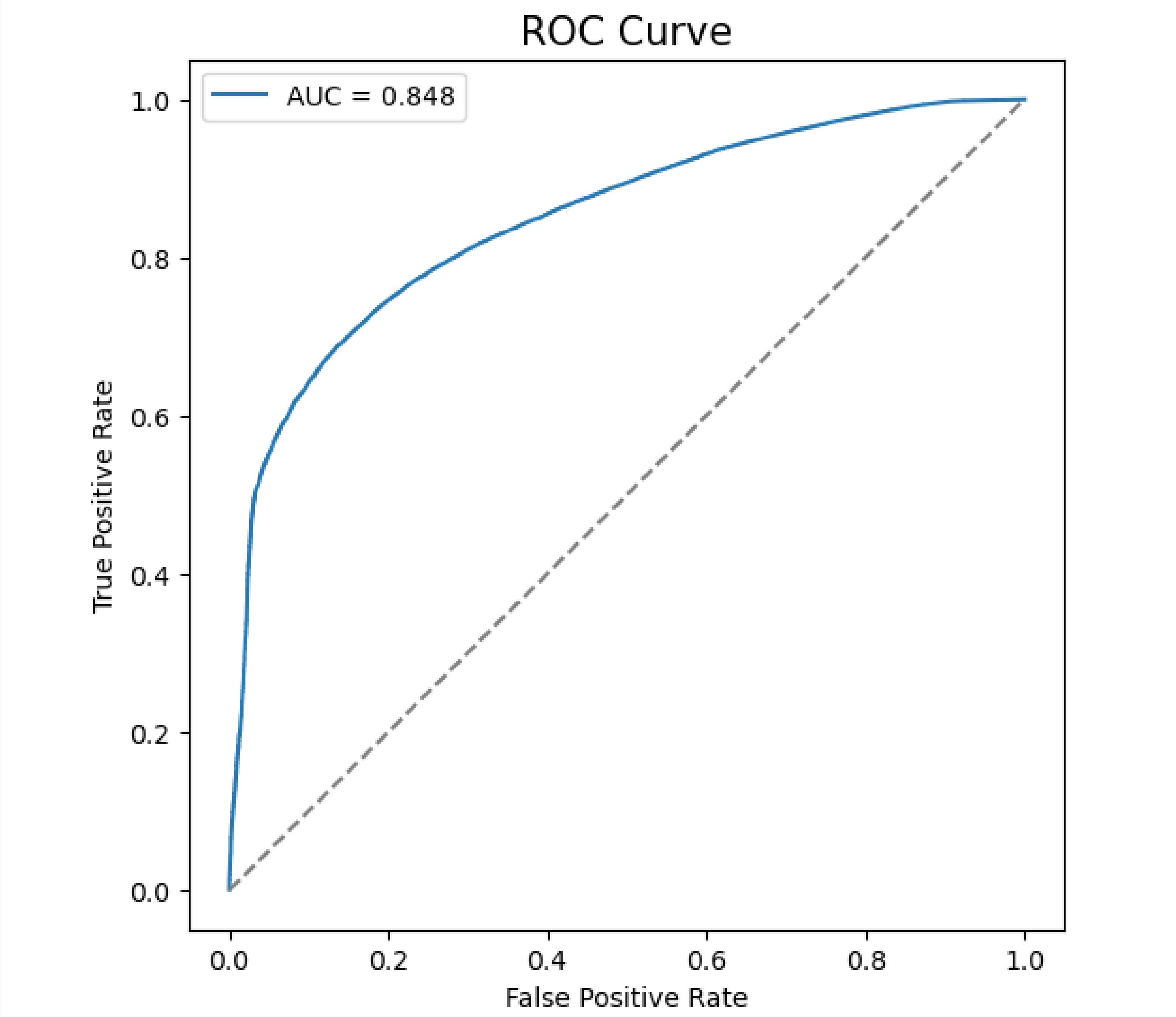
Evaluation Score

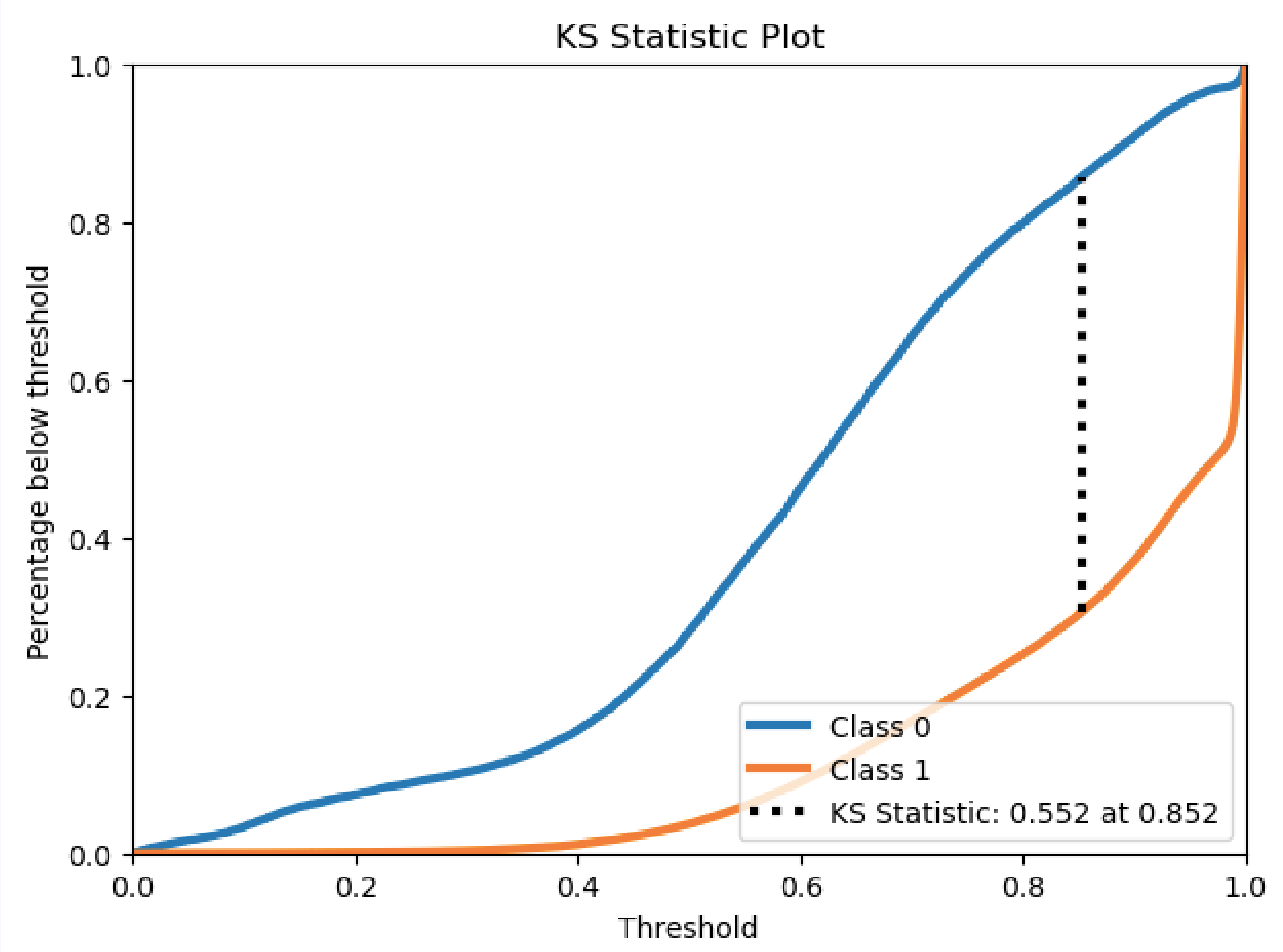
AUC Training Score : 0.847

AUC Testing Score : 0.848

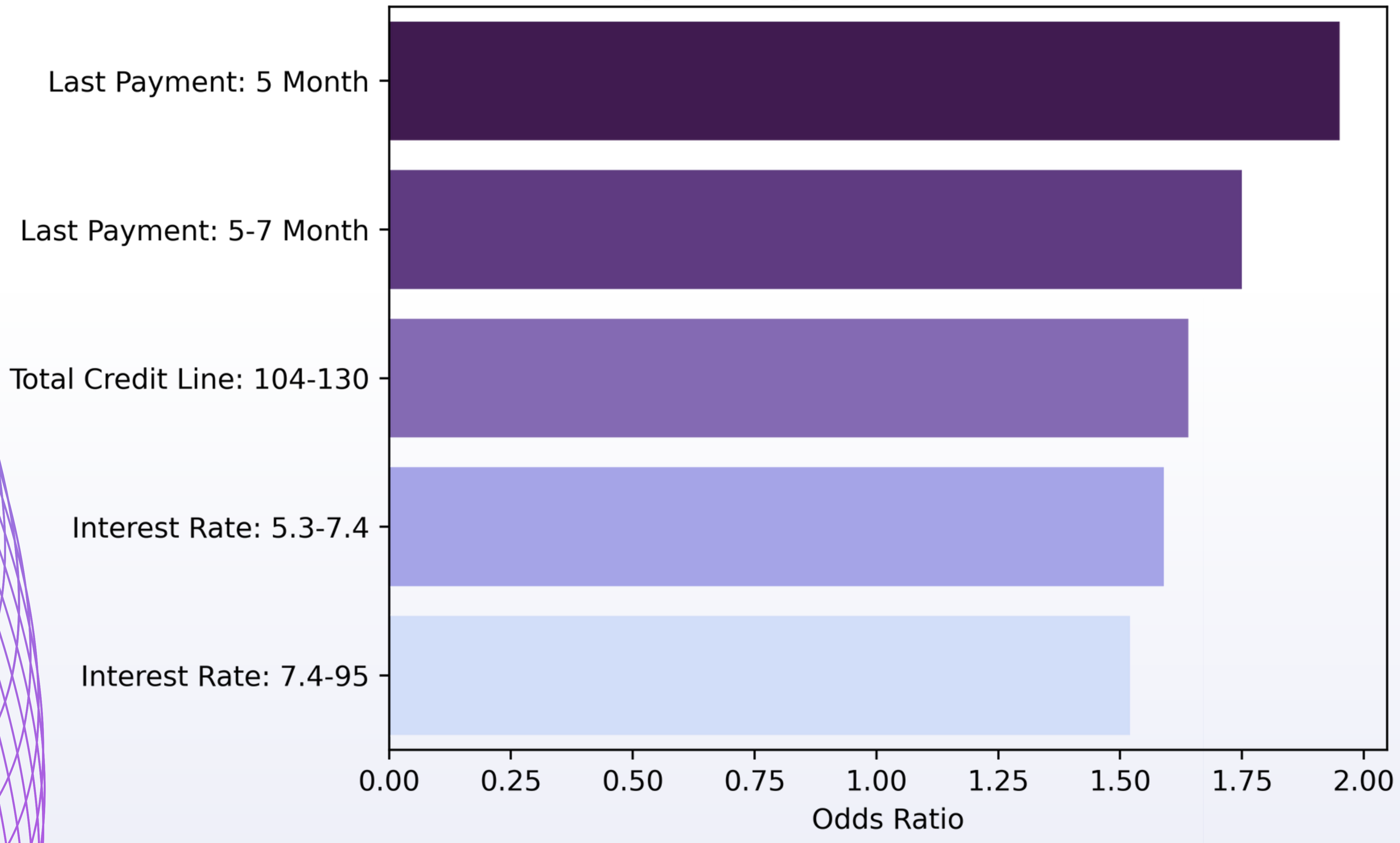
KS Statistic :0.552

Model Evaluation





Feature Importance



Last Payment:5 Month
Borrowers who made the last payment in the **past 5 months**, their odds of being a good loan borrowers will **increase by 1.9 times**

Last Payment:5-7 Month
Borrowers who made the last payment in the **past 5-7 months**, their odds of being a good loan borrowers will **increase by 1.7 times**

Interest Rate: 5.3%-7.4%
Borrowers with an **interest rate of 5.3%-7.4%**, their odds of being a good loan borrowers will **increase by 1.5 times**

Etc

Credit Score FICO Scale

*Glimpse of Credit Score

ID	Member ID	Credit Score
762003	962497	618
7342864	9004729	684
12385411	14397553	606
1076863	1277178	565

full credit score file: Download

Credit Score 0-100 Scale

*Glimpse of Credit Score

ID	Member ID	Credit Score
15510351	17602771	70.42
544737	702661	66.64
1457416	1711938	64.86
33290364	35933595	72.21

full credit score file: Download

Cut Off Treshold

This list helps business person to take proper decision making on treshold regarding to business condition

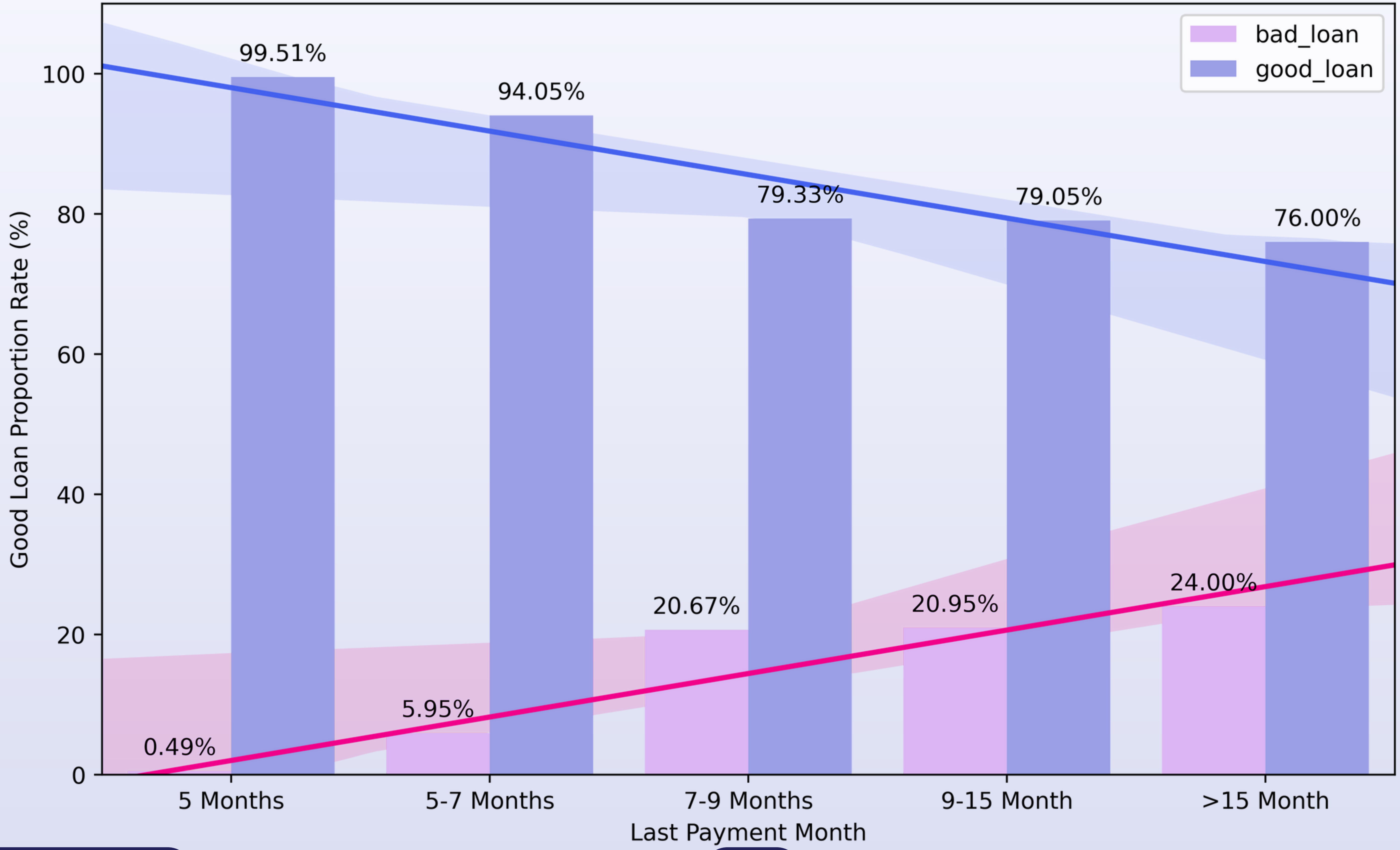
*Glimpse of Treshold

Treshold	Credit Score	Num Approved
76.00%	698	10,886
75.51%	692	17,285
73.68%	670	66726
70.05%	629	229,385

full cut off treshold file: Download

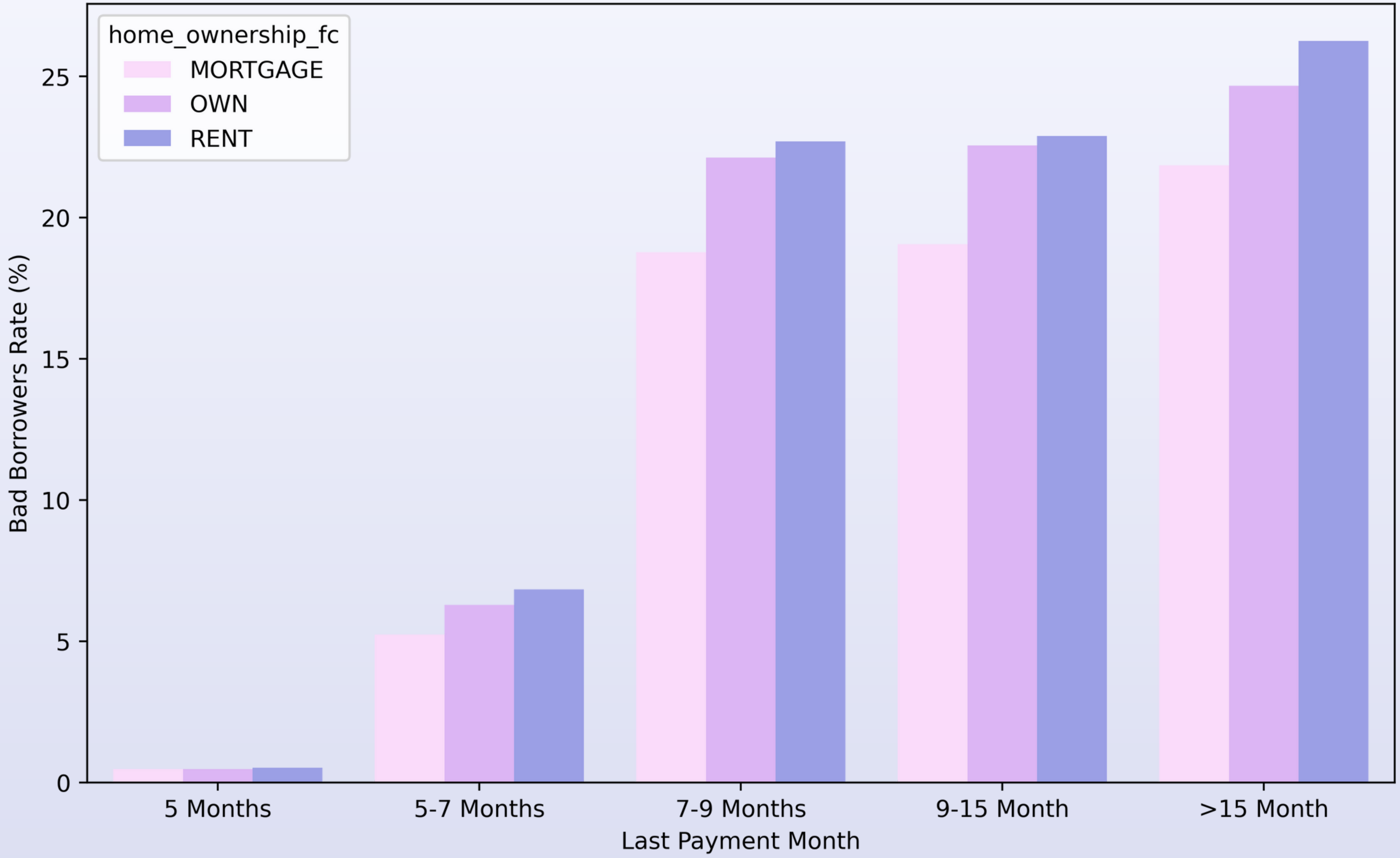
Negative Trends on Good Loan Rate of Loan Application Per Last Payment Months

The longer last payment month received tend to be more into bad loan borrowers, it means that borrowers with recent last payment are more likely to be a good loan borrowers



Proportion of Bad Loan Borrowers Per Last Payment Months Based on Home Ownership Type

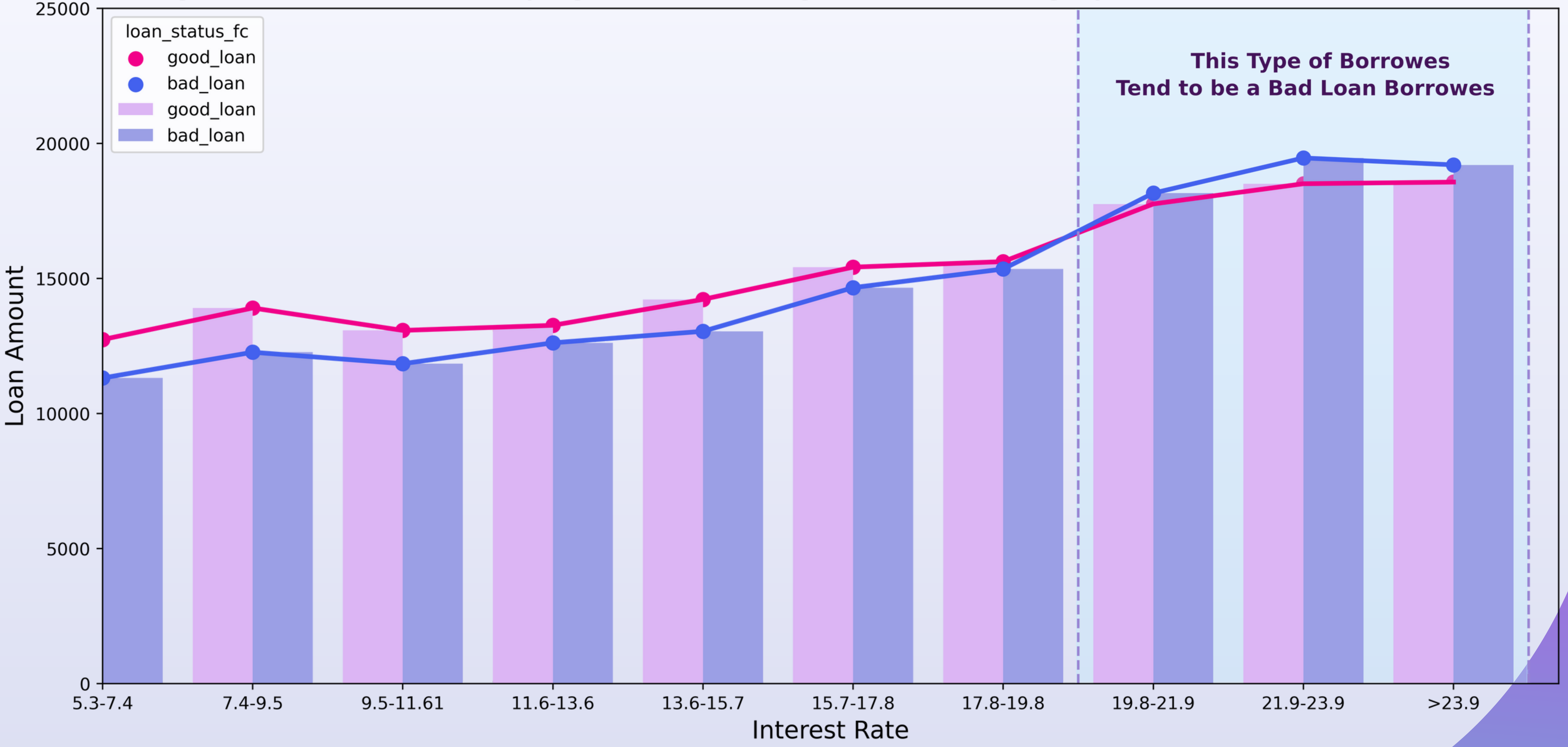
Borrowers who have rent home ownership type are more likely to be a bad loan borrowers, also borrowers who have long last payment month tend to be a bad loan borrowers, this is probably because borrowers who own a house or martgage have higher buying power or we can say they earn higher than someone who rent a house



Business Insights

Loan Amount Per Interest Rate Based On Borrowers Loan Status

The higher loan amount followed by high interest rate, higher interest rate slightly tend to be a bad loan borrowers



Business Insights

About Me

I graduated of bachelor's degree from Bandung Institute of Technology, School of Business and Management, Business degree. I also graduated from Rakamin Data Science bootcamp with outstanding grade, awarded as best final project team, and also my role as team leader.

I experienced on the following scope:

- Classification
- Regression
- Time Series Forecasting
- A/B Testing
- Deep learning using TensorFlow and Pytorch
- PySpark
- Recommender System
- Customer Lifetime Value
- Segmentation
- SQL
- Data Visualization (Tableau & Power BI)

Contact

github.com/erlndofebri

brian-insights.site/

erlandoregita99@gmail.com

0821-1000-4094

linkedin.com/in/erlandoregita/

THANK YOU