Case Study Data Science

William R. Astley

May 2022

1 Introduction

Financial credit and lending services are a significant underlying mechanism of the global, interwoven contemporary financial system. Within this field, one of the most important problems financial service industries seek to solve is finding a particular clients Credit Default Risk - The risk, mathematically the probability, of loss arising from a debtor being unlikely to pay its loan obligations in full. Financial institutions all across the world have credit analysts - people's whose job is to analyze and asses whether or not a borrower can repay a loan. Although machine learning algorithms aren't correct 100% of the time, nor are these analysts; hence one naturally ponders whether or not machine learning's unparalleled predictive power, in conjunction with its computational speed, can be utilized in a model so that either its efficiency rivals that of a credit analyst, or it can be used as a supplement for credit analysts.

2 Data

2.1 Data Set

For the duration of the study, we will use the term "data" to refer to the particular dataset mentioned in the references section; the dataset is described in further detail as **Figure 1**.

link: https://www.kaggle.com/datasets/laotse/credit-risk-dataset

Feature Name	Description
person_age	Age
person_income	Annual Income
person*home*ownership	Home ownership
person*emp*length	Employment length (in years)
loan_intent	Loan intent
loan_grade	Loan grade
loan_amnt	Loan amount
loan*int*rate	Interest rate
loan_status	Loan status (0 is non default 1 is default)
loan*percent*income	Percent income
cb*person*default*on*file	Historical default
cb*preson*cred*hist*length	Credit history length

Figure 1: Data Overview

2.2 Finding Outliers

2.2.1 Null Entries

We first find all of the missing values in our dataset. Since the number of missing values is a relatively small percentage of the total entries, we can remove each of these rows entirely.

2.2.2 Descriptive statistics

We now explore the dataset to look for any outliers. The descriptive statistics for our data is listed in **Figure 2**, located on the next page. Obviously, there isn't someone who is 144 years old, and there isn't someone who has been employed for 123 years. These outliers in age and employment length could negatively influence our model, but before removing these, we first try to find other outliers through bivariate analysis.

2.2.3 Bivariate Analysis

Figure 3 depicts the scatter plot matrix between the variables listed in the axis. Visually, it is not hard to see that income also has an outlier. Thus, it follows that there is also an outlier in income that we need to remove.

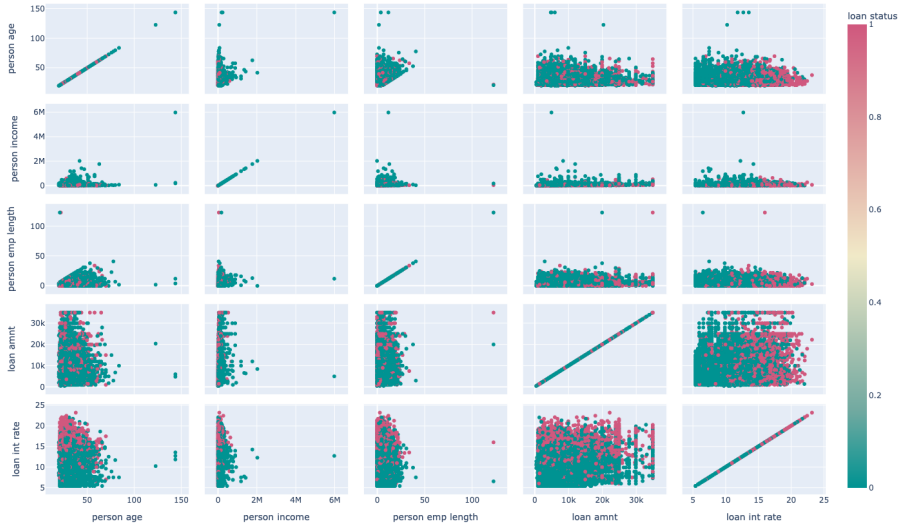

Figure 2: Scatter Matrix

2.3 Cleaning the Data

We now clean the data by

- Removing the rows containing a null entry.
- Removing the rows where Age is over 100.
- Removing the rows where income is over than 4000000

In addition, since we seek to use models that may not operate on categorical data, we use the one-hot encoding method to quantify these values.

3 Analyzing the Data

The nature of the problem in conjunction with the nature of the dataset implies that we are most likely dealing with some form of classification problem; Moreover, this is an imbalanced classification program as the data consists of 78.4% non-default cases. Since we are interested in the credit default rate, which in this data set is described by "*loan_status*", we see how this variable is related to other variables.

3.0.1 Box Plot

Observing **Figure 3**, there are two things that catch the eye:

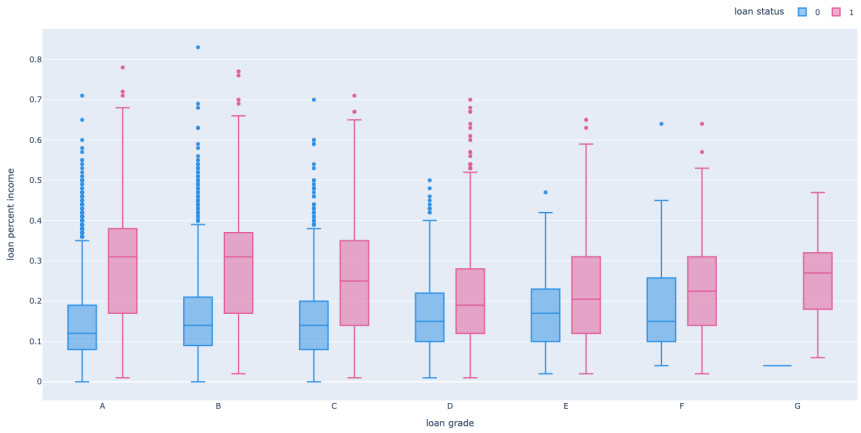

Figure 3: Box Plot

- No borrowers with loan grade G were able to repay their loan
- Those who don't default have a lower loan to income ratio mean value across all loan grades

3.0.2 Parallel Category Diagram

In order to understand how the categorical labels are related to one and other on the basis of loan status; we use a Parallel Category Diagram, depicted in **Figure 4**(on the next page), from which the following is deduced:

- Our dataset consists primarily of those who have not defaulted before - an observation initially made in *Section 3*.
- The most common loan grades are A and B; The least common load grades are F and G.
- Borrowers took out a loan for education the most; Borrowers took out a loan for home improvement the least.
- Defaults appear to be the most common when loans were incurred for covering medical expenses and debt consolidation.
- Home renters defaulted more than those with a mortgage;Homeowners defaulted the least.

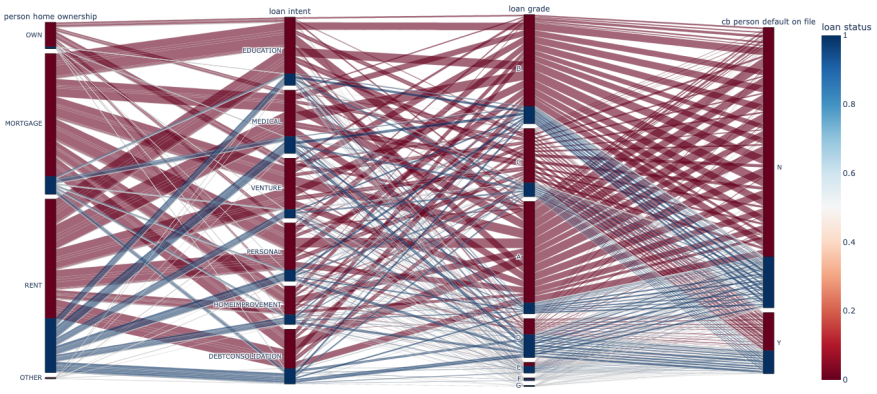

Figure 4: Parallel Category Diagram

4 Model Making & Training

Since this is an imbalanced classification problem, we use the following three models: K-nearest neighborhood models, logistic regression, and XGBoost. After splitting our data into a train and test split, we apply these various models to the data, and compare their ability to predict whether a borrower will default or not using a classification report (**Figure 5**).

```
              KNN
              precision    recall  f1-score   support

           0       0.86      0.96      0.90      4525
           1       0.72      0.39      0.51      1202

    accuracy                           0.84      5727
   macro avg       0.79      0.68      0.71      5727
weighted avg       0.83      0.84      0.82      5727

              LG
              precision    recall  f1-score   support

           0       0.82      0.98      0.89      4525
           1       0.72      0.17      0.27      1202

    accuracy                           0.81      5727
   macro avg       0.77      0.58      0.58      5727
weighted avg       0.80      0.81      0.76      5727

              XGBoost
              precision    recall  f1-score   support

           0       0.94      0.99      0.96      4525
           1       0.96      0.75      0.84      1202

    accuracy                           0.94      5727
   macro avg       0.95      0.87      0.90      5727
weighted avg       0.94      0.94      0.94      5727
```

Figure 5: Classification Report

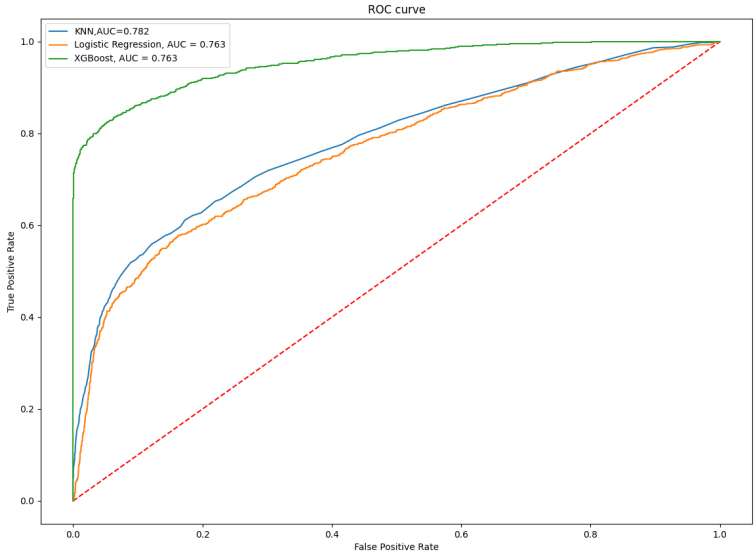

Figure 6: ROC Graph

4.1 Evaluation Metrics:

In this particular case, since our data is imbalanced, the weight for which we place on "Accuracy" should be low. Since Accuracy measures the ratio of total truly predicted values to the number of input samples, the model will get high accuracy when predicting the majority case, but won't be very accurate when predicting the minority case. Hence, we will use the F1 Score, Precision, and Recall as our three primary assessment metrics. Ostensibly, it seems as if the XGBoost model is probably the best model as it has the highest F1 Score, Precision, and Recall out of all the models.

4.2 Comparing Models

4.2.1 Predicting Class Labels

In order to compare the assess which model is the "best" at predicting class labels, we use an ROC curve(**Figure 6**, above) - a probability function where the x-axis is the rate of false positives, and the y-axis is the rate of true positives. It then follows that the closer the AUC is to 1, the more accurate it is(it's indicative of a high true positive rate and low false positive rate). Consistent with out initial belief, it appears that the XGBoost is our best performing model, as it has the largest AUC value.

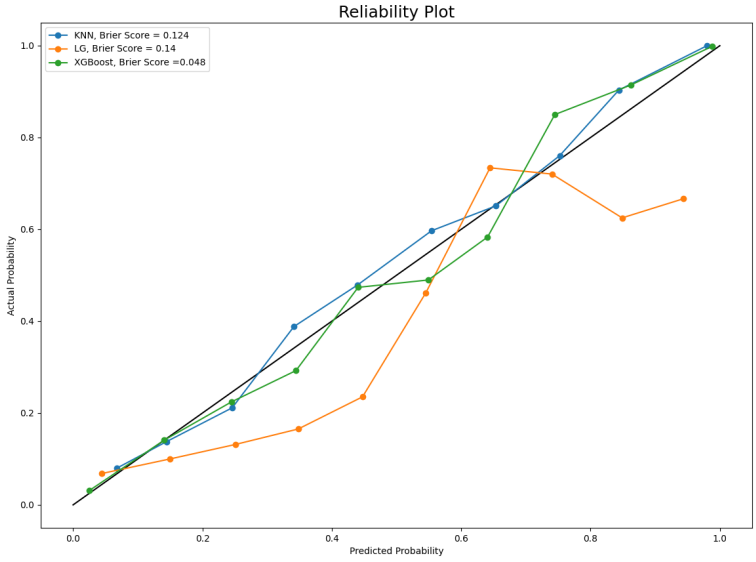

Figure 7: Reliability Plot & Briar Score

4.3 Model Calibration

we'll now evaluate their performance at predicting the probability of the sample belonging to the positive class. In order to compare these models with one and other, we use both a Reliability Plot and Brier Score(**Figure 7**, above). The reliability plot graphs actual probabilities versus the predicted probabilities on a test set; The Brier Score, a cost function, calculates the mean squared error between predicted probabilities and their respective positive class values. Thus it follows that a lower Brier Score is indicative of a more accurate prediction.

5 Conclusion

It follows that since the XGBoost model performs the best across all our metrics that is the best out of the 3 models. We can further analyze the structure of this model through information gain which measures each feature's contribution for each tree in XGBoost(**Figure 8**, next page). We can then deduce that the three most important features for predicting loan defaults and its probability are:

1. rent as home status
2. Loan to income ratio

3. loan grade C

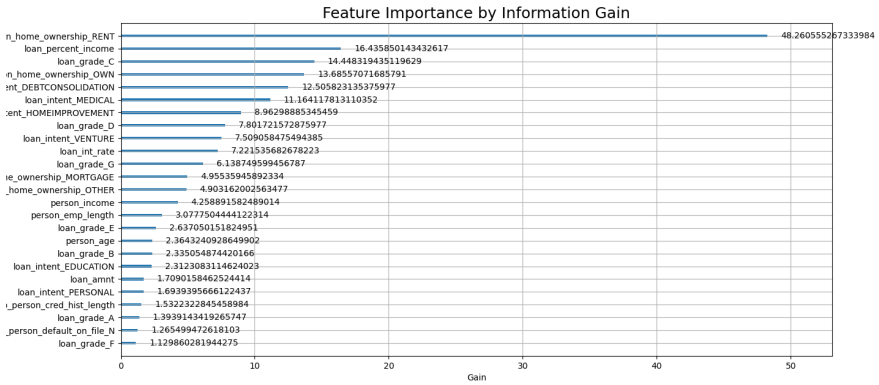

Figure 8: Feature Importance