Loan Default Prediction using XGBoost with explainability

Author: Tareq Galala

Abstract: In this paper, we will look into a loan dataset and analyse the data by looking for patterns characterizing financial distress. Next, we look at feature importance, feature selection and explainability in finance and try to understand what leads to defaults. Several methods and models will be used to gain trust in the results.

Index Terms: Feature importance, feature selection, explainability, machine learning, loan defaults

1- ANALYSIS DOMAIN, QUESTIONS, PLAN

Introduction

One of the key challenges in machine learning today is called a black box which is when data scientists don't know exactly how the ML model works. Explainability decreases as the complexity of the model increases. Therefore, there is more demand today on how these models work in order to provide transparency on how a decision was made. This is important especially in finance where decision makers and central banks demand explainability and governance. [1]

In this paper, we will use a data driven approach in addressing this challenge by exploring the features that could lead to financial distress and apply different methods to understand the features importance and explain the characteristics of the loans related the probability of default then compare results to gain trust.

Data Source

The data used in our analysis is sourced from Kaggle competition and was originally used to improve credit scoring, by predicting the probability that somebody will experience financial distress in the next two years.

Research Questions

Financial Services are interested in adopting Machine Learning, but these models are not always easily explainable to regulators and other stakeholders. [1] From this understanding we ask a set of questions we aim to address in this paper

1- Can feature engineering help representing the underlying structure of the data?
2- Which features did the predicting model think are the most important?
3- What Feature selection techniques can we use to enhance our predictions for default?
4- How can we measure and explain the contribution of our features in a given model?

Analysis Strategy

Our objective in this paper is to perform explanatory data analysis to understand our data better and dive deeper in analysis, we will train a parametric and non-parametric model to measure feature importance, then we will select our top features that will increase the scoring and explore feature explainability methods.

The following are the steps needed for our analysis plan:

1- Data Pre-processing
Importing, cleaning, wrangling & filling missing data,

2- Exploratory Data Analysis
Feature exploration, distributions, transformation & outliers.

3- Feature engineering

4- Modelling
We will use modelling to capture feature importance & select our features that are most important. Then will tune our model and evaluate for feature explainability.

5- Discussion
Summarizing our findings and evaluating our results.

2- FINDINGS AND DISCUSSION

Findings

1- Features Distribution

Our first finding was feature distributions. It’s clear from figure 1 that most of our features are highly skewed. Also, by calculating the skewness in pandas, we found out that all features have a very high positive skewness which indicates an asymmetry in

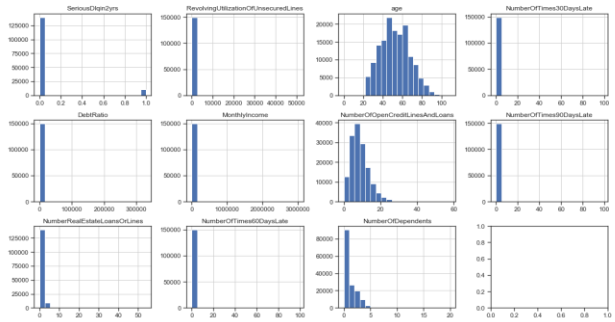


Figure 1: Features Distributions

the distribution towards the right-hand. Therefore, all the features were log transformed and many outliers got removed. Age_log feature skewness is around zero which indicates a near normal distribution. See figure 2.

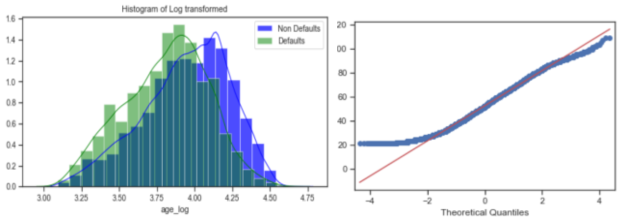


Figure 2: Left: Age log Histogram, Right: Q-Q plot for Age

2- Target Variable

Our second finding was our target variable which clearly shows that our data has high imbalanced classes, see figure 3. The default class represents 6.7% of the dataset. Solutions to this could be under sampling the majority class or oversampling the minority class example SMOTE. However, XGBoost model can deal with imbalanced data with 'scale_pos_weight' parameter. The parameter penalizes negative cases (0) and assign a higher weight to positive cases (1). The best parameter value was 7 which increased our AUC. Throughout this paper will be using AUC as our metric of choice.

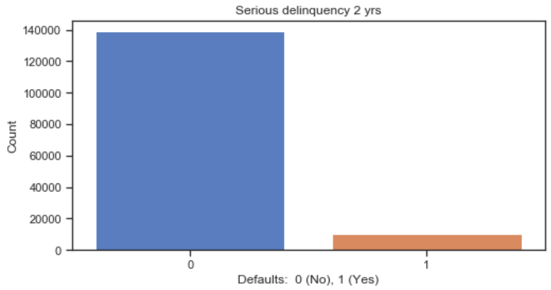


Figure 3: Target Variable SeriousDlqin2yrs

3- Relationships

As this is a binary classification problem, initially we couldn't find any patterns or relationships between our target variable and the rest of the features by plotting scatterplots. The heatmap in figure 4 shows us a pairwise correlation between features. We noticed that monthly debt is very high correlated with Monthly Balance, DebtRatio_log with MonthlyBalance and NumberofDependents_log with IncomeEachMember.

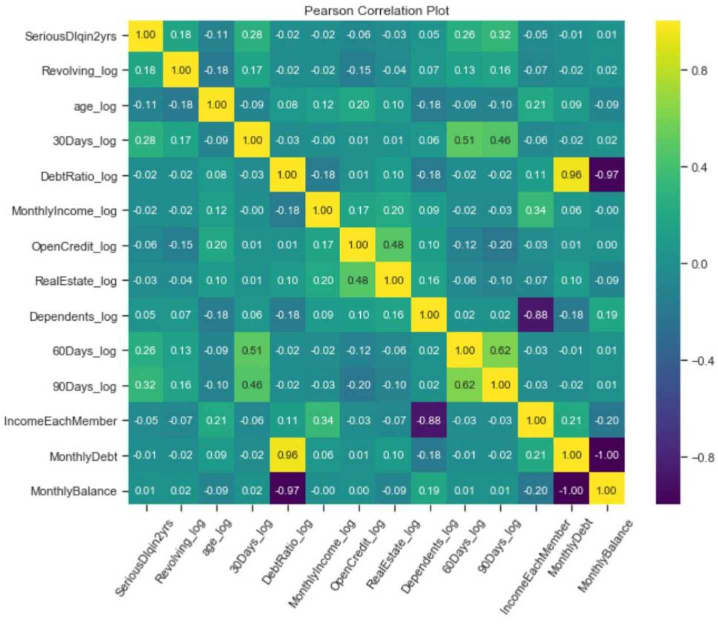


Figure 4: Correlation Heatmap

Another unexpected finding was age. Figure 5 shows a linear regression between default counts mean at each year of age. It is generally a negative relationship where the counts of defaults decrease by age. According to US studies younger people may have poor borrowing decisions and are more likely to default than the older adults. For these reasons, loan rates are much higher for younger people as they are a riskier profile, which can lead to financial distress. [2]

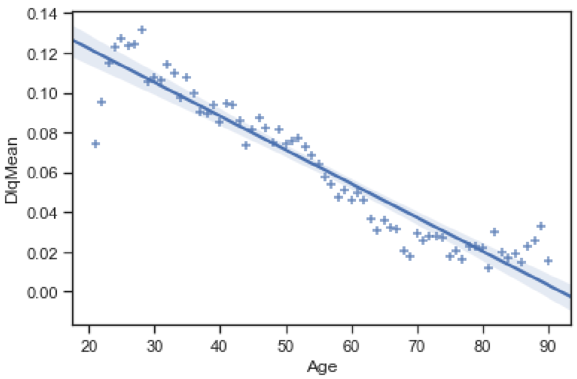


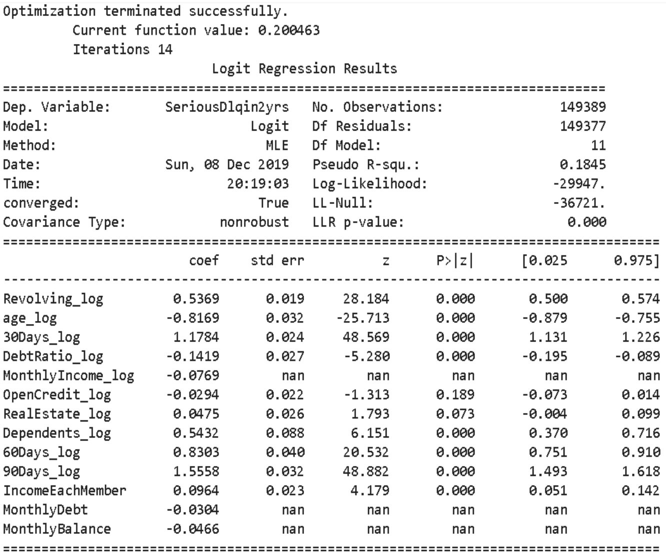
Figure 5: Linear Regression between default counts mean and age

4- Modelling

In the modelling section we will train two models to capture feature importance and try to select the top features that leads to financial distress. One model will be a parametric (Logistic Regression) and the other will be non-parametric (XGBoost) and will try to look at the difference between the findings.

Logistic Regression model

We ran the stats model version of logistic regression model and Sklearn version to find the coefficient values, Figure 6 table shows the coefficients values where the higher the coefficient, the higher the importance of a feature.



```
Optimization terminated successfully.
         Current function value: 0.200463
         Iterations 14
                           Logit Regression Results
==============================================================================
Dep. Variable:       SeriousDlqin2yrs   No. Observations:               149389
Model:                          Logit   Df Residuals:                   149377
Method:                           MLE   Df Model:                           11
Date:                Sun, 08 Dec 2019   Pseudo R-squ.:                  0.1845
Time:                        20:19:03   Log-Likelihood:                -29947.
converged:                       True   LL-Null:                       -36721.
Covariance Type:            nonrobust   LLR p-value:                     0.000
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
Revolving_log         0.5369      0.019     28.184      0.000       0.500       0.574
age_log              -0.8169      0.032    -25.713      0.000      -0.879      -0.755
30Days_log            1.1784      0.024     48.569      0.000       1.131       1.226
DebtRatio_log        -0.1419      0.027     -5.280      0.000      -0.195      -0.089
MonthlyIncome_log    -0.0769        nan        nan        nan         nan         nan
OpenCredit_log       -0.0294      0.022     -1.313      0.189      -0.073       0.014
RealEstate_log        0.0475      0.026      1.793      0.073      -0.004       0.099
Dependents_log        0.5432      0.088      6.151      0.000       0.370       0.716
60Days_log            0.8303      0.040     20.532      0.000       0.751       0.910
90Days_log            1.5558      0.032     48.882      0.000       1.493       1.618
IncomeEachMember      0.0964      0.023      4.179      0.000       0.051       0.142
MonthlyDebt          -0.0304        nan        nan        nan         nan         nan
MonthlyBalance       -0.0466        nan        nan        nan         nan         nan
=====================================================================================
```

Figure 6: Stats model logistic regression results

The plot below is a coefficient values plot which suggests the important features and matches the stats model report.

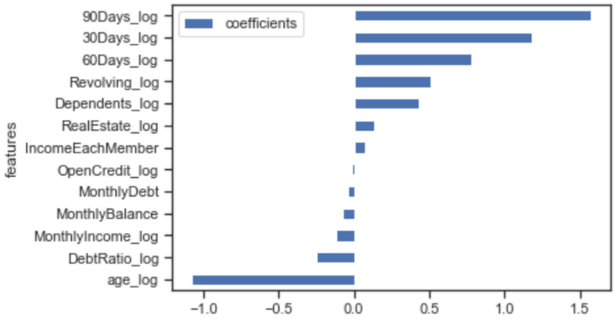


Figure 7: Sklearn logistic regression coefficient values plot

Not only the results nearly matched in both methods, but also our score of AUC was high at 0.835 which indicates trust in the result. Logistic regression is a very good baseline for binary cl assification problems, especially in credit risk.

Other methods that are used for logistic regression feature selection are recursive feature elimination (RFE) and Select from Models (SFM), but for the purpose of this paper, we restrict this analysis to coefficient values in order to keep our work manageable.

XGBoost (Extreme Gradient Boosting) model

XGB feature importance counts the number of times each feature is split on across all trees and visualizes the result as a bar graph, with the features ordered according to how many times they appeared. We did a random and grid search and tuned our model and reached an AUC score of 0.87, but we noticed that the feature importance changed, See figure 8.

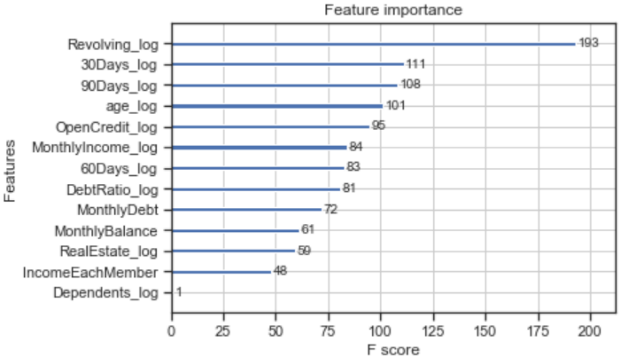


Figure 8: XGBoost best model feature importance chart

We found that Revolving_log was the most important in XGBoost model, but in logistic regression the 90Day_ Log ranked first, which both can lead to financial distress.

5- Feature Selection

Our feature selection was based on the feature importance by XGBoost as it scored better. The two new features MonthlyDebt and MonthlyBalance performed good and were selected in the top 10 features. The new feature IncomeEachMember unexpectedly didn't perform well and wasn't important enough.

6- Explainability

We can make an interpretation on the model easily with the plot tree in XGBoost, see figure 10, and have a direct understanding on the results but things become messy if we had more trees, therefore, we looked into SHAP, which is probably the current state-of-the-art explainability method. SHAP's main advantages are local explanation and consistency in a global model structure.

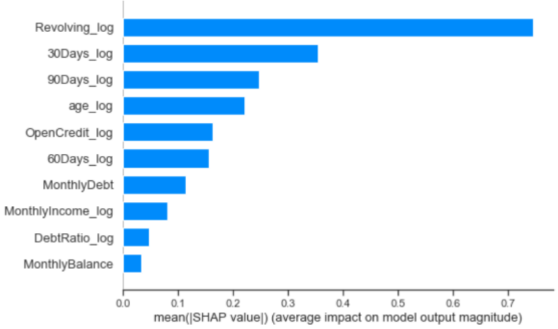


Figure 9: SHAP Feature importance

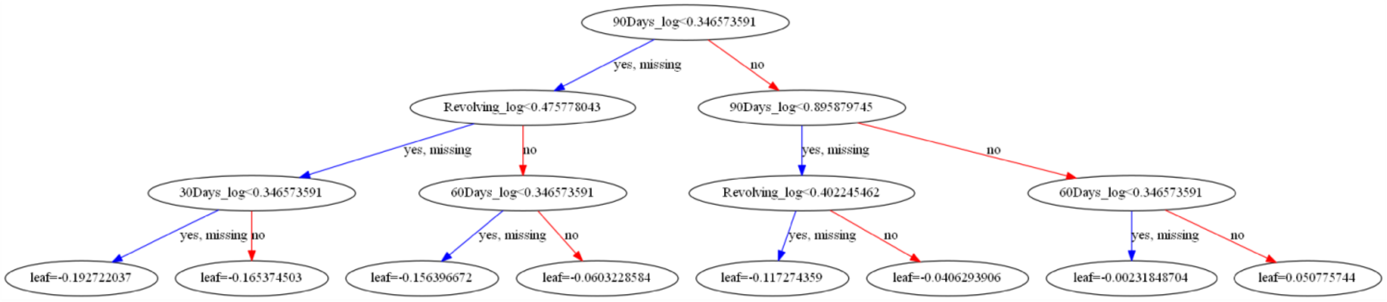


Figure 10: XGBoost tree plot

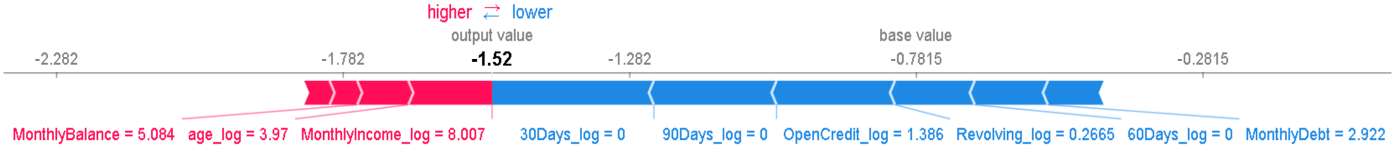


Figure 11: SHAP force plot

Figure 11 shows features contributing to push the prediction from the base value. Features pushing the prediction higher are shown in red. Features pushing it lower appear in blue. The record we are testing from the test set has a lower than average predicted value at -1.52 compared to -0.78 base value. The details here explain why an individual prediction was made on a local level. This is important to help us include recommendations with other decision factors.

Figure 12 is a summary plot that shows global feature importance which can lead to a better understanding of overall patterns.

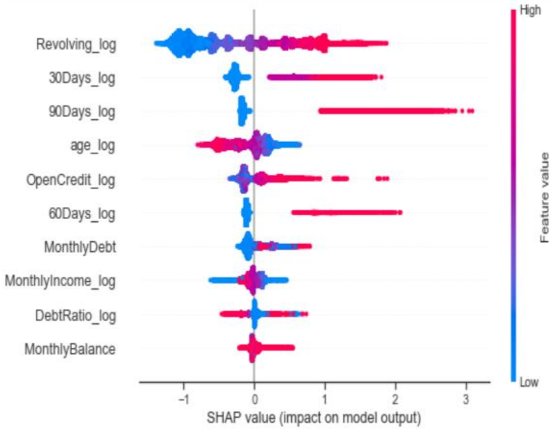


Figure 12: SHAP Summary plot

Discussion

To conclude, it is worth discussing our research questions and evaluate if our analysis answered the subject of this paper.

We engineered four new features which helped in our analysis and found hidden patterns. The new features were based on discussion with people from the finance domain, so domain knowledge is vital in analysis. We also found the most important features from our loan dataset that leads to financial distress. Two of our engineered features were selected in the top 10 features.

There are many techniques that can be used for feature selection, we used model coefficients values, correlation heatmaps and models feature importance.

Model explainability, is extremely important today. We used SHAP which helped us to measure the contribution of particular features in a given prediction on a local level and global level.

Finally, our objectives in this paper was met by understanding the data more, but the generalizability of the results is limited by the data in hand. More research should consider other factors that can lead to financial distress.

References

[1] Philippe Bracke, Anupam Datta, Carsten Jung and Shayak Sen: Machine learning explainability in finance, Bank of England working paper series. August 2019

[2] Sara Davies, Andrea Finney, Sharon Collard and Lorna Trend: BORROWING BEHAVIOUR. August 2019

[3] Mona J. Gardner and Dixie L. Mills: Evaluating the Likelihood of Default on Delinquent Loans. November 2019

[4] Benedict Guttman-Kenney and Stefan Hunt: Preventing financial distress by predicting unaffordable consumer credit agreements: An applied framework. The Financial Conduct Authority (FCA) Occasional Papers. July 2017

[5] Extreme Gradient Boosting with XGBoost. Data camp course