Why do People Borrow and Run?

## Introduction

What causes people to default on their loans? Using Extreme Gradient Boosting (XG Boost) we created a model to predict the current status of loans. We consider “good” loans to be loans that are fully paid off, in their grace periods, or that are currently being paid back, while bad loans are loans that are in default, are late, or have been charged off.

Furthermore, we discovered that borrowers are more likely to be in the “bad” group if they have high debt-to-interest ratios, poor subgrades (and thus worse interest rates), and if they have had the loan for a long period of time.

## Feature Engineering

We first explored our training data and discovered that only 10% of our observations are “bad” loans. We then worked on adjusting our data by converting categorical variables into numerical variables. For the variable sub-grade, we appointed the best sub-grade (A1) as 1, and the worst sub-grade (G5) as 35. We also ordered home ownership. We then converted date values into numbers, and created the variable “issue datestamp”. This variable measures the distance from February 2017 to the month and year that a loan was issued. Finally, we made the response variable, loan status, a binary variable with “good” loans = 0 and “bad” loans = 1. We then utilized a label encoder to convert the other categorical variables into numbers.

We noticed that there are outliers in the debt-to-income (DTI) ratios, annual incomes, and the number of derogatory public records. DTI ratios over 100 are rare, so we assigned the value of 100 to any DTI ratio 100 and above. There are many DTI ratios of 9999, but these are a result of a borrower’s income being 0; debt/0 is infinity, and 9999 was assigned to all values of infinity in the data we were given. Thus, all DTI values where the income was 0 were assigned to have DTI ratios of 100. There were very few observations with more than 20 derogatory public records, although the maximum number of records is 86; thus, we assigned the value 20 to all observations with 20 or more derogatory public records. We also split the annual incomes into bins (Model 1).

Using ANOVA testing we determined predictor variable interactions were mostly insignificant and did not include them in the model.

## Modeling

We split the training data into 3 parts, training with ⅔ of our data and validating with the last ⅓. We attempted various models, utilizing the brier score and AUC score on the training and validation data to evaluate the model’s effectiveness.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| [Brier Score] | numeric only(9) | + categorical (26) | + new feature (29) | + outlier clipping |
| Benchmark [All Zeros] | 0.1014 |  |  |  |
| Random Forest | 0.0898 | 0.0857 | 0.0855 | 0.0854 |
| Logistic Regression | - | - | - | 0.0861 |
| Extra Trees | - | - | - | 0.0854 |
| XGBoost | - | - | - | 0.0844 |

We checked to ensure that our validation and training data scores’ both beat our other models and were close to each other, to ensure the models were effective and not overfitting. After testing a logistic regression, random forest model, and

We then predicted the probability

## Results

The variables that are most closely aligned with the status of the loan are \_\_\_\_\_ (logistic regression results and weight). Of these, we determined that people with higher DTI’s, worse subgrades, longer issued datestamps and a higher number of inquiries are more likely to have bad loans.

We examined the distributions of the training, testing and validation data sets. Similar distributions imply we’re running our tests on the correct data. A fault we found was that the distribution of delinquency in the past two years did not match the distribution we are predicting, looking back at our data we determined that this is because delinquency in the past two year has practically no weight in predicting the status of loans.

RESULTS

We determined

Key Result:

No interaction variables because they were not found to be significant through anova testing

XXX variables most important in determining loan\_status

Summary

* DTI, Subgrade, Issue Datestamp, and inquiry

## 

## 

## Feature engineering

* Benchmark validation: brier score
* Use logistic regression, rf, xg
* 3-fold cross validation (⅔ training, ⅓ valid)
* Tried manual parameter search
* USE BRIEr & AUC scores for valuation
  + Looked at training & valid
  + Values must not be too far apart so dont overfit
  + Try to avoid overfitting with still good validation score
* Did single obs. --- look at to see everything works
* Feature Importance tree
  + Did for tree model, random forest, logistic regression
    - Logistic regression gives wieight to how important
      * Utilize to see bad
    - Tree model
      * Not weights, just relative importance
    - Mostly overlapping (slight diff. - nature of models)
* ML model should be generalizable
* VIOLIN PLOT
  + Train data
  + Validation data
  + Test data
    - See distribution for each feature
      * If distributions differ, then training on bad/wrong data - must train so distribution on data is almost identical to test data
  + Delinquency 2 years ---> bad correlation
    - This is because weight is meaningless in regression
* DTI, Subgrade, Issue Datestamp, and inquiry
* Subgrade → as subgrade gets worst, the worst effect it has on being bad
  + Interest rate
* Inquiry Increase in Last 6 months, the higher the inquiry the less likely they are to pay
* Issue datestamp (distance between now and issued date)
* DTI
* Home ownership →
  + Owning house -> more likely to default
  + Renting
* a 2-3 page report containing at least two graphics or tables, a detailed description of the methods used to analyzing the data, and any key results that were obtained (submitted as a .pdf file)
* Same distribution of percentages (trai, test, validation)