Two step recommendation system for loan capital recovery and loss prevention

Matthew Gudorf

Introduction

Lots of capital involved with loans

Lots of capital lost from defaulted loans

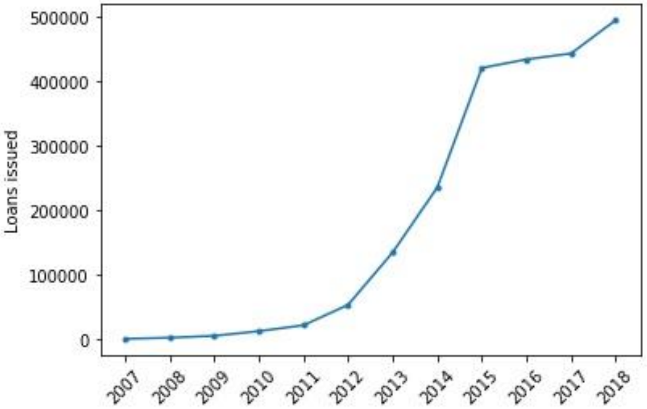
Create a two-step recommendation system

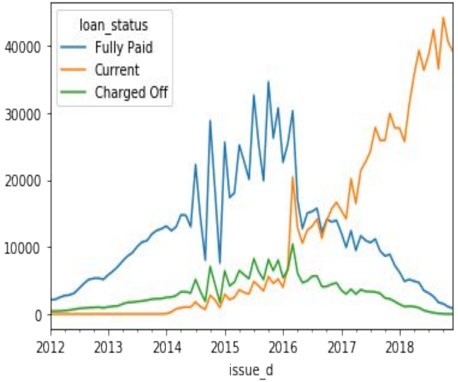
1. Create a predictive model for the outcome (“good” / “bad”) of a loan
2. Create a predictive model for the amount of money recovered from bad loans.

The combination of these two models enable for proactive and reactive actions.

Motivation

1. Number of annual loans issued has an upwards trend.
2. Defaults on loans is a time delayed process; cannot predict future growth of charge offs





Lending Club dataset

Millions of data points

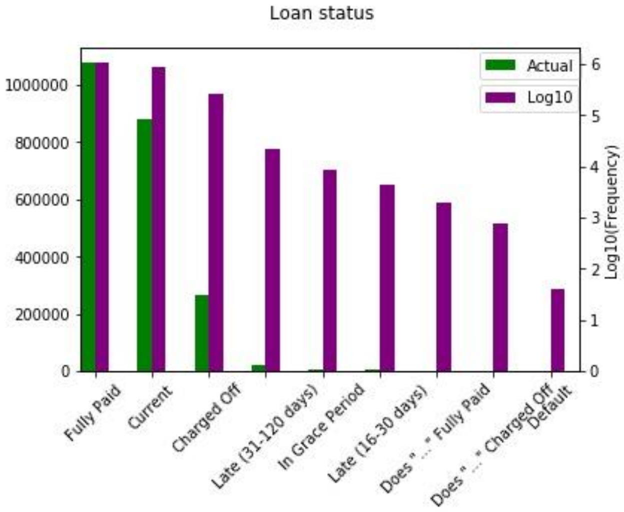
Data features split between loan descriptors (interest rate, principal amount, issuance date) and borrower descriptors (geographical location, income, employment title, etc.).

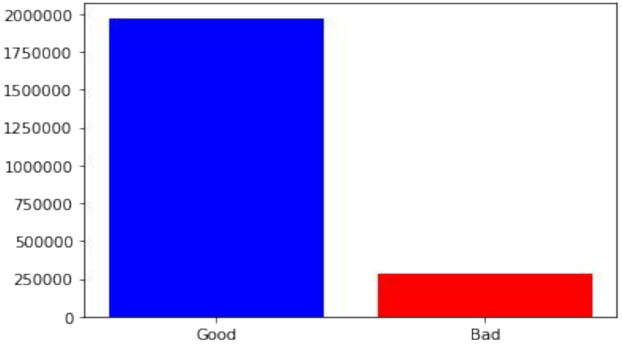
Dataset has issues:

1. Missing features
2. Missing values
3. Imbalanced features
4. Some features are essentially subcomponents of others (highly correlated)

Data story : Classification

1. Conversion of multilabel loan statuses to binary “good” and “bad”.
2. Compare logistic regression and random forest classifiers for this binary classification problem.



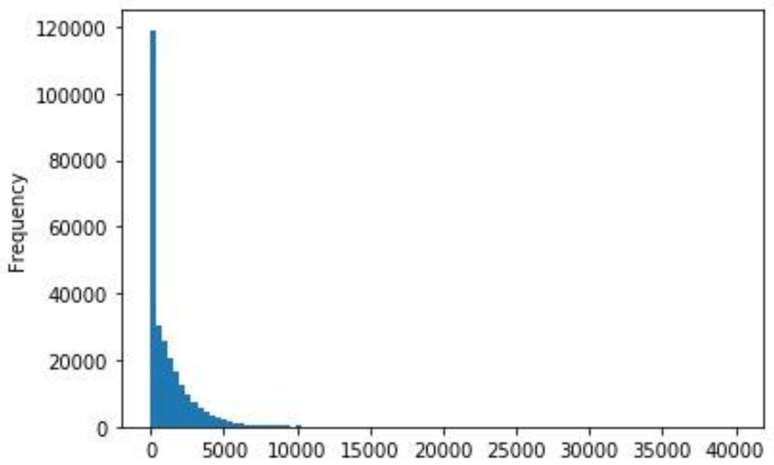


Recoveries on defaulted loans

Out of 268559 defaulted loans, only 184684 have had capital recovered from (68.8%).

There is still billions of dollars of capital that could possibly be recovered.

How to decide on which defaulted loans should be pursued? This is where the recovery regression model comes in.

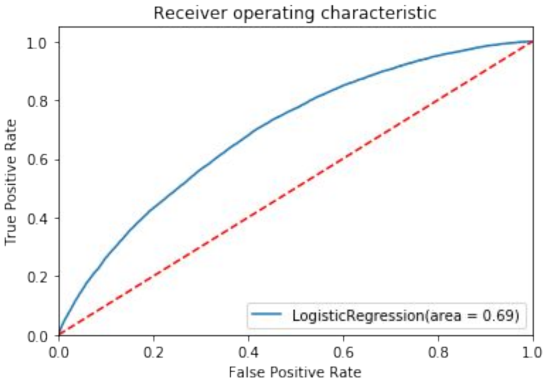


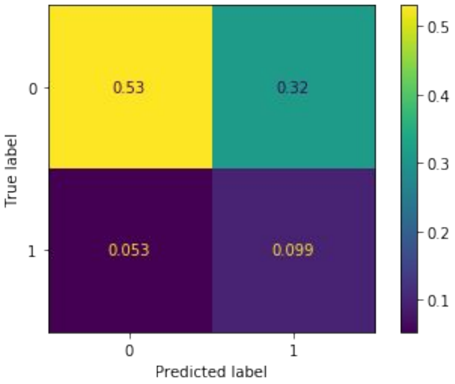
Classification analysis

Random forest and Logistic regression have similar performance. Figures displayed here show the performance of the Logistic Regression model.

Things to note: many false positives (many rejected loans which would have been fully paid) in exchange for reducing the number of potentially defaulting loans.

Recall = 0.65, Precision =

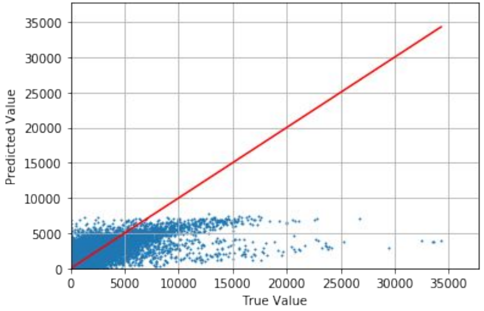


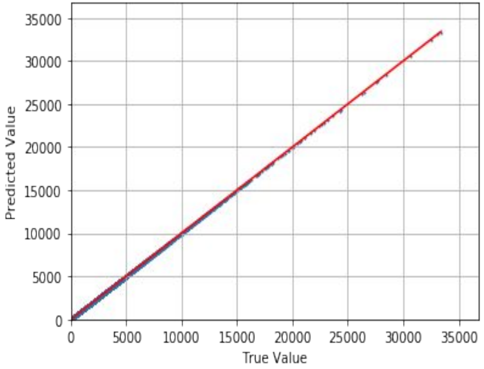


Regression analysis

Model is clearly missing something as it seems to be unable to predict larger recovery values. R^2 value is ~0.47.

The performance becomes nearly perfect upon inclusion of three features, which would seemingly indicate collinearity but the correlations with the dependent variable are not significant. More investigation required.





Conclusion

The system I propose to prevent and recover capital is as follows:

1. By predicting the outcome of loans, we can reduce the number of defaulted loans, but not to 0.
2. For the loans that still default, predict the recovery amount so that we know which loans to pursue and attempt to recover capital from.

Next step: Theoretically we want to only use the second step when necessary, therefore I would first work on improving the classification model.