Predicting Defaulters for Lending Club Dataset

Ins: Prof. Daniel Acuna

Team Members: Anjali Nair, Rajesh Ayyalasomayajula, Shambhavi Godbole, Vishwanath Hegde

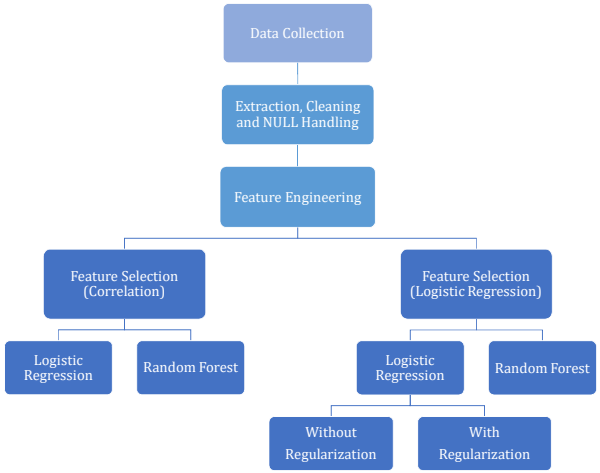
Problem Statement

Lending Club is the largest peer-to-peer lending company. It enables borrowers to create unsecured personal loans between $1,000 and $40,000. The purpose of this study is to identify the various factors affecting borrowers capability to repay the loan. The expected outcome is to predict the whether the loan will be defaulted or not.

Data Description

The dataset consists of 682215 records of customer loan data spread over 145 different attributes. The data was collected from Jan 2017 to July 2018. The data was discovered from Lending Club's website. The factor at test is the loan status which would be 1 if repaid and 0 if defaulted.

Our Approach



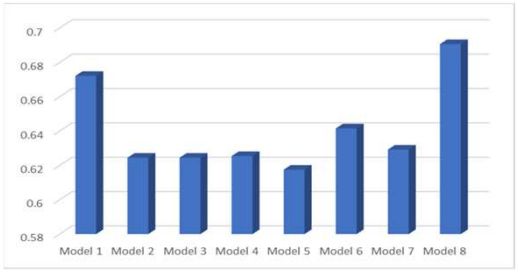
References

https://www.lendingclub.com/info/download-data.action

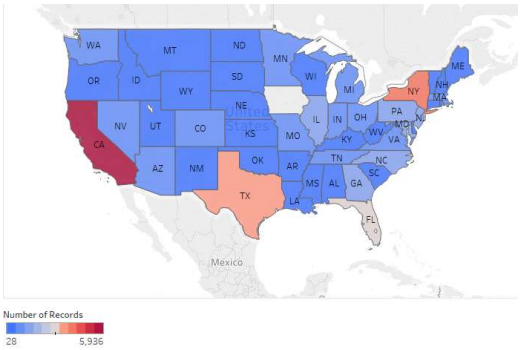
Model Description

Model	Features	Techniques	Evaluation
Logistic Regression	loan_amnt, term, mort_acc, int_rate, installment, total_rec_int, tot_hi_cred_lim, sub_grade_encoded, home_ownership_encoded, verification_status_encoded, purpose_encoded, revol_util_encoded	Spark Pipeline, String Indexer, One Hot Coding, Cross Validation, Regularization	AUC
Logistic Regression	num_tl_30dpd, mort_acc, num_tl_90g_dpd_24m, chargeoff_within_12_mths, pub_rec_bankruptcies, delinq_2yrs, collections_12_mths_ex_med, num_tl_120dpd_2m, acc_now_delinq, sub_grade_encoded, home_ownership_encoded, verification_status_encoded, purpose_encoded, revol_util_encoded	Spark Pipeline, String Indexer, One Hot Coding, Cross Validation, Regularization	AUC
Random Forest	loan_amnt, term, mort_acc, int_rate, installment, total_rec_int, tot_hi_cred_lim, sub_grade_encoded, home_ownership_encoded, verification_status_encoded, purpose_encoded, revol_util_encoded	Spark Pipeline, String Indexer, One Hot Coding, Cross Validation	AUC
Random Forest	num_tl_30dpd, mort_acc, num_tl_90g_dpd_24m, chargeoff_within_12_mths, pub_rec_bankruptcies, delinq_2yrs, collections_12_mths_ex_med, num_tl_120dpd_2m, acc_now_delinq, sub_grade_encoded, home_ownership_encoded, verification_status_encoded, purpose_encoded, revol_util_encoded	Spark Pipeline, String Indexer, One Hot Coding, Cross Validation	AUC

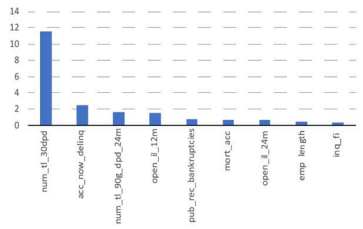
Model Performance



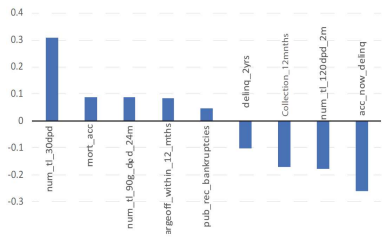
Model 1: LR; **Model 2:** LR (λ = 0, α = 0.4) ; **Model 3:** LR (λ = 0, α = 0.4); **Model 4:** LR (λ = 0, α = 0.4); **Model 5:** LR (λ = 0, α = 0.4); **Model 6:** RF; **Model 7:** RF (Correlation); **Model 8:** LR (Correlation)



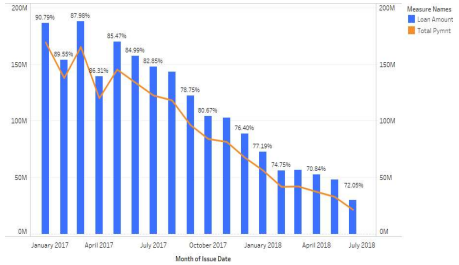
Regression Coefficients



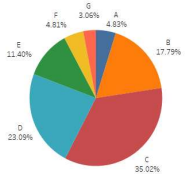
Feature Importance



Yearly Payment Decrease

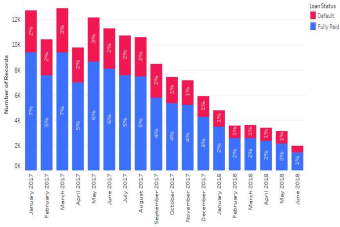


Defaulters per Grade



Mortgage Accounts Per Year

Default Vs Fully Paid



Proposed Solution

With a little information related to the loan, our model can predict whether the borrower will repay the loan or default. The model provides the best accuracy of 67.54% to classify the loans. This will help in determining the right borrowers for the loan be lent to. In future, we plan to implement more powerful big data analytics techniques to improve the performance of this model.