Lending Club Loan Default Prediction

by

Radhika Vijayaraghavan
netID# zg4894

Instructor: Dr.Eric Suess

STAT 652, California State University East Bay

Spring 2023

Contents

Abstract

The objective of this project is to apply the various Machine Learning modeling techniques taught in STAT 652 course. Effort has been made to incorporate the 5-step process (collect-explore-train-evaluate-improve) for each model. Model improvements have been done using Cross Validation, Tuning. The algorithms used for this predicting `loan default` are Null model, Elastic Net Regression, Boosted C5.0, Random Forest. Since we're interested in being able to predict which of 'Fully Paid' or 'Charged Off/Default' a loan will fall under, so we can treat the problem as *binary classification*.

As part of data cleaning, the below were performed:- - The columns that had *greater than 10% of missing values* were removed. - Converted variables to its correct data type such as characters to factors - Removed redundant variables. - Removed variables that *leak data* from the future, Eg:- `funded_amnt`, `recoveries`, `total_pymnt`, `collection_recovery_fee` etc) - I also remove all the loans that don't contain either 'Fully Paid' or 'Charged Off'/'Default' as the loan's status and then transform the 'Fully Paid' values to 0 for the positive case and the 'Charged Off/default' values to 1 for the negative case

As part of feature selection, the below were performed:- - *Boruta algorithm* was used for feature selection, and *recipe* for creating dummy variables for categorical variables - Feature Importance Plot and combines ROC curves have been plotted for the best model(C5.0).

Data Description

a. Data Source

The source of this data set is Kaggle(Lending Club data from 2012-2014).

b. Data Description

The cleaned data set consist of 366603 observations on the following 25 variables. 2 new variables have been added. Below are the description of some of the variables I used in my modeling.

- **loan_amnt** = The listed amount of the loan applied for by the borrower.
- **term** = The number of payments on the loan. Values are in months and can be either 36 or 60.
- **installment** = The monthly payment owed by the borrower if the loan originates.
- **home_ownership** = The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
- **purpose** = A category provided by the borrower for the loan request.
- **dti** = A ratio calculated using the borrower's total monthly debt payments on the total debt obligations
- **open_acc** = The number of open credit lines in the borrower's credit file.
- **revol_bal** = Total credit revolving balance

Additionally below variables were added:-

- **loan_default** - Binary values 0(Fully Paid) or 1(Default/Charged Off), extracted from initial `loan_status` variable
- **fico_average** - Average of `last_fico_range_high` and `last_fico_range_low`

Executive Summary of model accuracy

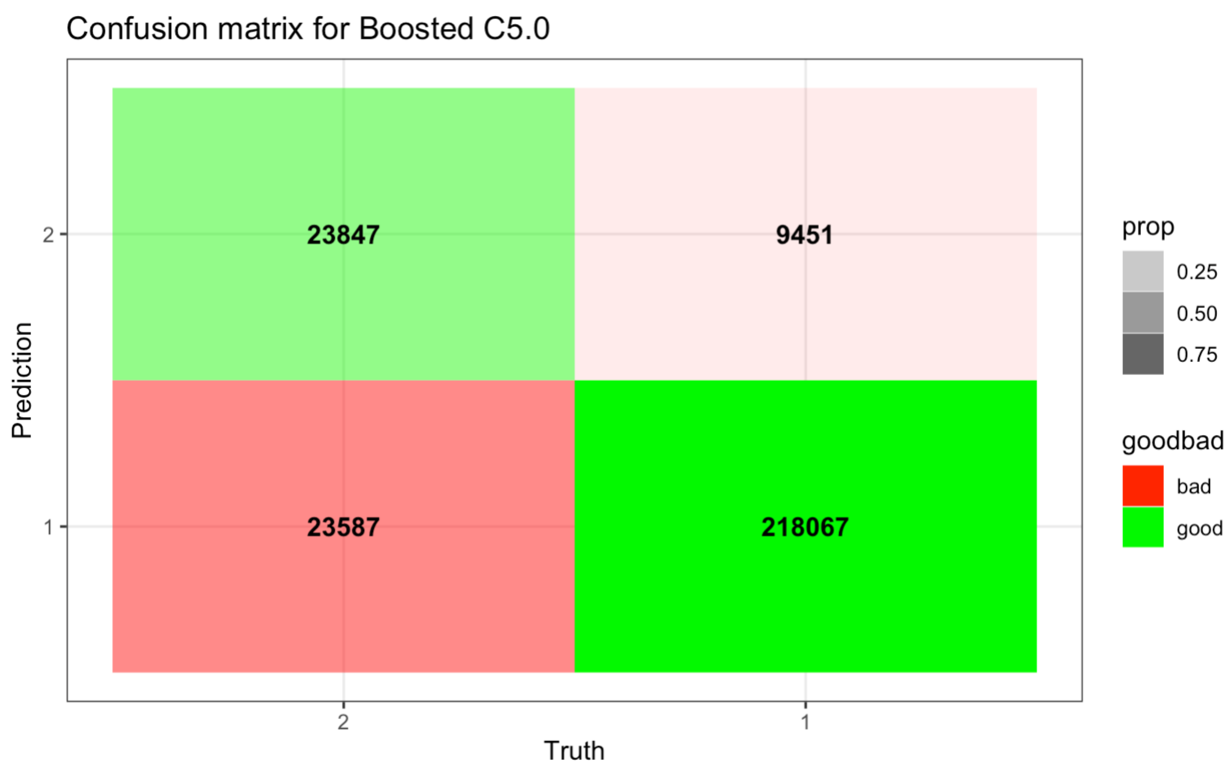
	Null model	Logistic regression using GLMNET	Random Forest	Boosted C5.0
Before Cross Validation	0.8290	0.8714	0.8642	0.8833
After Cross Validation	-	0.8699	0.8665	0.8798

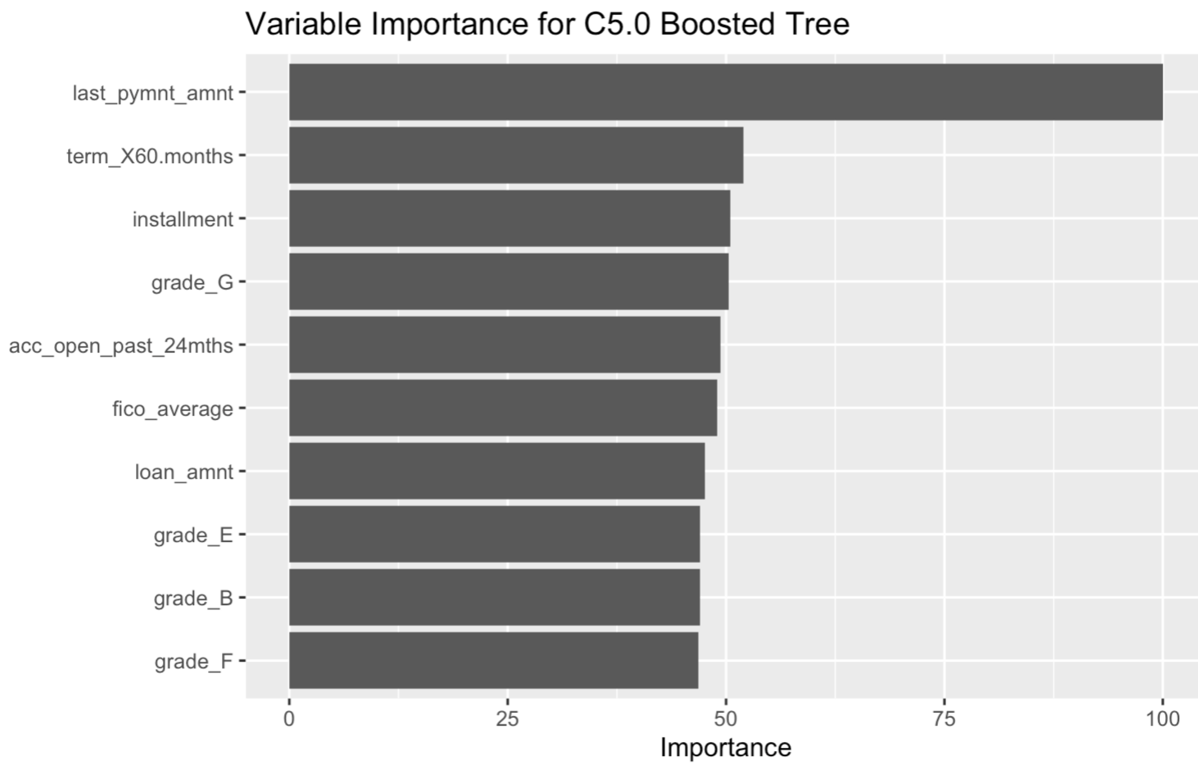
Accuracy score and ROC_AUC for Boosted C5.0 after 5-fold Cross Validation

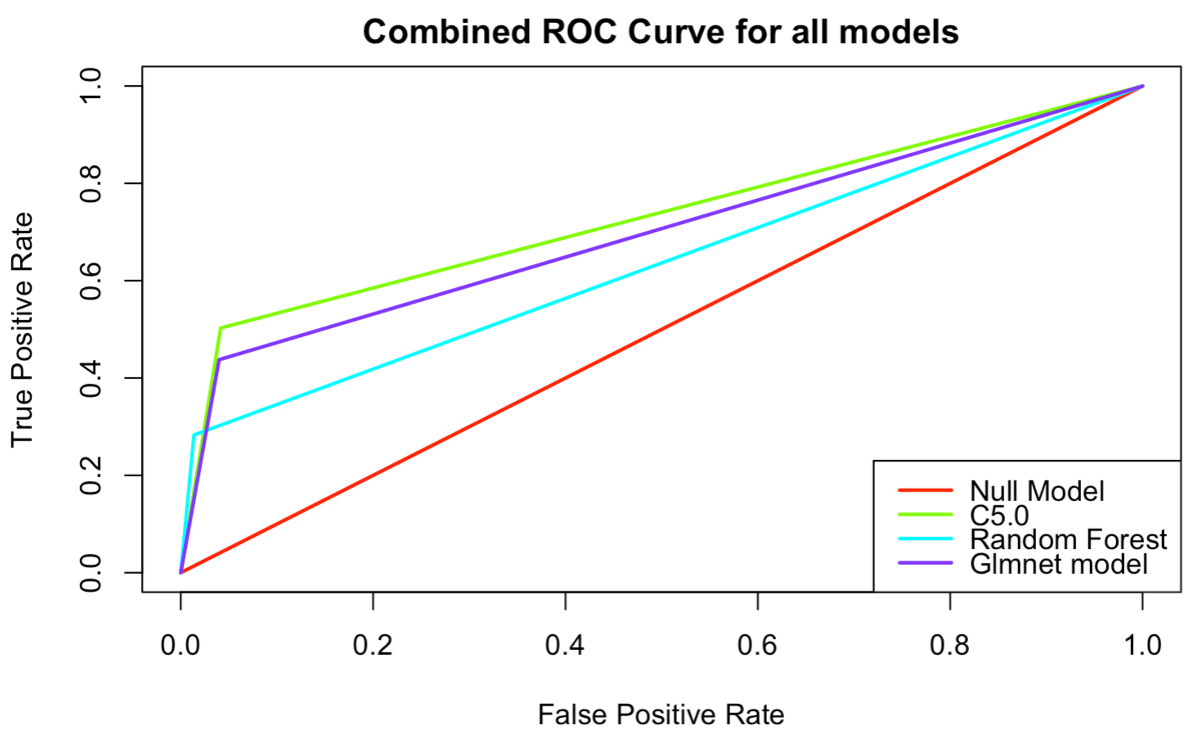
A tibble: 2 × 6

.metric <chr>	.estimator <chr>	mean <dbl>	n <int>	std_err <dbl>	.config <chr>
accuracy	binary	0.8798408	5	0.0002937985	Preprocessor1_Model1
roc_auc	binary	0.9103067	5	0.0005880179	Preprocessor1_Model1

2 rows




Variable Importance for C5.0 Boosted Tree
last_pymnt_amnt
term_X60.months
installment
grade_G
acc_open_past_24mths
fico_average
loan_amnt
grade_E
grade_B
grade_F
0
25
50
75
100
Importance


Combined ROC Curve for all models
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Null Model
C5.0
Random Forest
Glmnet model

Conclusion

From all the above models, Boosted C5.0 tree is the best model for this dataset with an accuracy of 0.8833 and ROC-AUC of 0.91. After 5-fold cross validation, the accuracy dropped to 0.8763 and ROC-AUC and 0.90. As per the feature importance plot, the features that seem to be highly important in predicting `loan_default` are `last_pymnt_amt`, `term_X60.months`, `installment`, `grade_G`, `acc_open_past_24mths`.

Findings so far

A lender must consider the following variables while deciding whether to Loan or not:-

- Grade: When a person is assigned Grade A, the risk of default is lowest and G grade shows the risk of default is highest. This is because interest rate increase from A-G

- Term: default rate is high on 60 months term

- High interest rate: The interest rate increases with increase in loan amount leading to higher chances of default

- inq_last_6mths: inquiries in last 6 months There is a increase in default when number of inquiries increases in last 6 months. Too many inquiries in 6 months may indicate that the borrower is not getting loan from anywhere and is desperate to find one, hence, the number of inquiries are high.

Acknowledgement

I would like to express my heartfelt gratitude to Dr.Eric Suess for his continued guidance, excellent teaching, extremely useful assignments and case studies. I would also like to thank my fellow classmates in Statistics department with whom I've had an opportunity to have many constructive conversations through the course this semester.