Lending Club Loan Classification Summary

Abstract

LendingClub is the world's largest peer-to-peer lending platform and was the first peer-to-peer leader to register its offerings as securities with the Securities and Exchange Commission(SEC). How can we forecast borrowers' payback ability before release loan.In this project, we apply many classification machine learning algorithms to Lending Club data from 2012-2014 by using five step learning process.The aim for this project is to find the best classification model for LedingClub.

Introduction

LendingClub enables borrowers to create loan listings on its website by supplying details about themselves and the loans that they would like to request. Investors make money from interest. Rates vary from 6.03% to 26.06%, depending on the credit grade assigned to the loan request.The challenge for LendingClub is how to decrease the borowers defaulting rate in order to protect investors' fund.
The data set for the years 2012 to 2014 consists of 423810 observations and 152 varibles.After dealing with NA data, the data set become 324353 observations.
The target varibable in this dataset is 'loan_status' which shows the status of the approved load.It includes three different unique values:Fully Paid(Loan has been fully repaid),Default(Loans has not been current for 121 days or more),Charged Off(Loan for which there is no longer a reasonable expectation of further payments).Here we only consider Fully paid and Charged Off.
For Predictor Variables, we choose 10 most related features for our model.

Predictor Variables	Description
loan_amnt	Total applied loan amount
funded_amnt	Total amount committed to that loan at that point time
term	Term of laon
int_rate	Interest rate for the loan
installment	how many installments of the loan
home_ownership	7 levels of the home ownership status
annual inc	The self-reported annual income provided by the borrower during registration
verification_status	4 levels of the verification
open_acc	open account time
total_pymnt	The total payment of the loan

In this project, we use Null model, logistic regression,KNN,Decision Tree Cart, C5.0,Naive Bayes, Randomforests, multiple liner regression algorithms to train our models. For evaluation part,we set up cross table of predicted and actual value to calculate accuracy for each model.By comparing each model's accuracy and ROC curve, we can pick the best model.

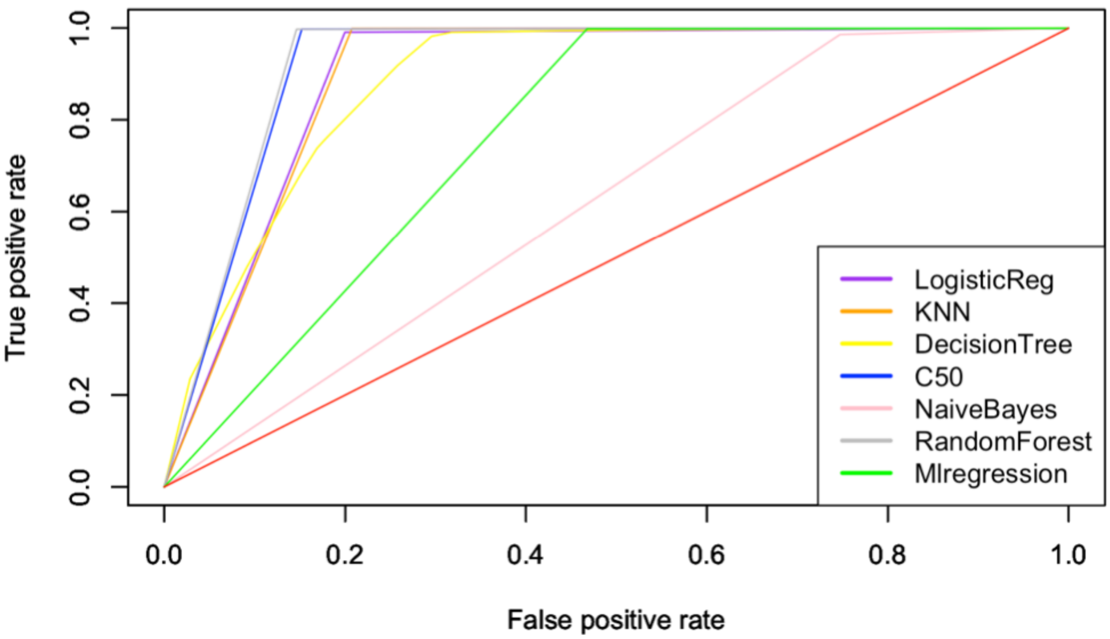
Conclusion

After evaluating each model, the best model is Random forest with accuracy 0.9719.
Model Accuracy Table

Model	Accuracy	AUC
Logistic Regression	0.9569	0.8956

Model	Accuracy	AUC
KNN(k = 11)	0.9622	0.8956
DecisionTree CART	0.8894	0.8848
C5.0(boost with trial = 10)	0.9702	0.9230
NaiveBayes	0.8553	0.6195
RandomForest	0.9719	0.9256
MultipleLinearRegression	0.9155	0.7655

ROC curve



Refit random Forest model with all 2012-2014 data and then use this model to classify all of the 2015 data.The accuracy for predicting 2015 data is 0.9806 and AUC is 0.9256.