Credict Risk Analysis - LendingClub Loan data

Sreejaya[1], Vuthy[1], & Suman[1]

[1] City University of New York (CUNY)

Author Note

test note......

Abstract

LendingClub is an online lending platform for loans.Borrowers apply for a loan online, and if accepted, the loan gets listed in the marketplace. As an investor/lender, you can browse the loans in the marketplace, and choose to invest in individual loans at your discretion - basically purchase *notes* backed by payments based on loans. In this project, we will attempt to analyse and predict the if a loan is risky loan or not so that we can avoid investment in high-risk notes.

Keywords: delinquent, dti (debt-to-income ratio), credit utilization ratio, mortgage accounts, open credit lines

Credict Risk Analysis - LendingClub Loan data

Introduction

LendingClub's peer to peer model can not only yeild better interest rates over the traditional banking counterparts, treasury bonds and other such financial instruments, it also has advantages like lower overhead costs, lower cost of capital etc. Based on each loan application and credit report, every loan is assigned a grade ranging from A to G, and sub-grade 1 to 5, with a corresponding interest rate. The higher the interest rate, the riskier the grade.

The LendingClub's datasets contains comprehensive list of features (about 115), we will shortlist few features to employ to train our model for predictions. We will build different models, and test those against the validation dataset and select a better performing model.

Literature Review

We have reviewed few of the previous projects/analysis/kaggle competitions done in this area ("GiveMeSomeCredit," 2011) (Jitendra Nath Pandey, 2011) (Liang, n.d.) (Shunpo Chang, 2015). Majority of these were applying machine learning to improve the loan default prediction. Some of these determined that Random Forest appeared to be better performing, and others logistic regression would be better. However, real-world records often behave differently from curated data in competetions like kaggle, so, we will try to apply different regression techniques including the logistic regression, naive bayes, and random forest on the real loan data during the years 2012-13 to continue search for a better predictive model. We will also include additional features like dti, credit utilization etc. from the LendingClub dataset, to see if they influence our target (credit risk) variable.

Methodology

Data Exploration

The dataset we used for this analysis is from publicly available data from LendingClub.com, which is basically an online market place that connects borrowers and investors. As part of the data exploration, we will perform *Exploratory data analysis* to better understand the relationships in the given data including correlations, feature distributions and basic summary statistics. We will also identify the outliers, missing data and any look for invalid data.

Data Preparation

As part of the data preparation, we will try to fix the data issues we noticed in our exploratory analysis, which involves treating the outliers, missing data, invalid data etc. We will also identify the data classifications, and create dummy variables wherver required, and will convert the categorical data to numeric data which would help us later in model building.

Model Development

Our primary objective here is to *identify risky loans*, which is binary outcome, so we will be using *logistics regression, naive bayes, and random forest* modeling techniques to examine and predict the credit risk using a training subset (80%) of the original full data.

Model Validation

A validation subset (20%) was used to test how well our candidate models predit the target variable. AUC and the model Accuracy will be used to select a better predicting model.

Experimentation and Results

Data Exploration and Preparation

Feature Selection: There are 111 fields and 188K observations. However, not all fields are intuitively useful for our model building, such as the loan ID, member ID, last payment month etc., so we will be removing such fields. We will also be removing features with majority of NAs (80% NAs). In order to label the dataset, we will classify the loans that defaulted, charged off, or were late on payments as negative cases, and those that were fully paid or current was classified as positive loans.

Outliers: Look for outliers, and remove outliers that might affect the model.

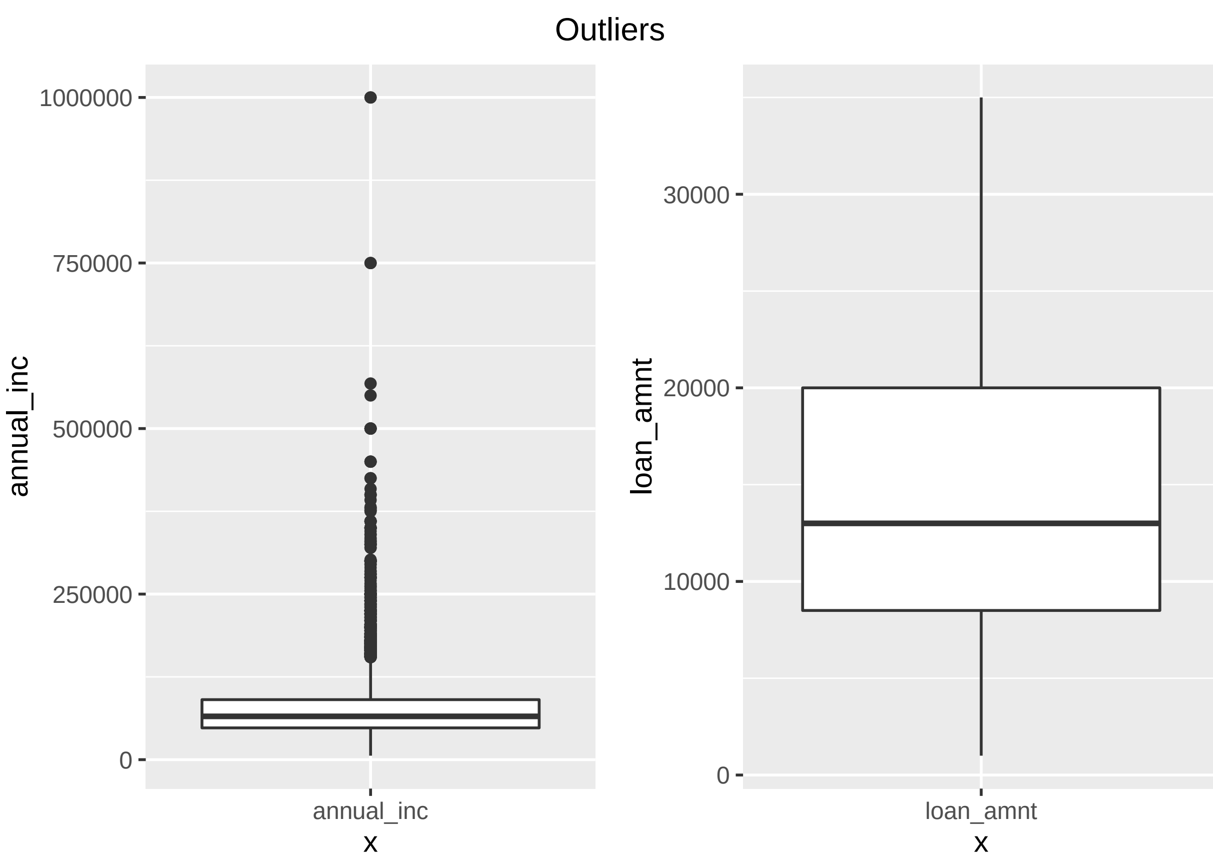

Figure 1

Visualize:

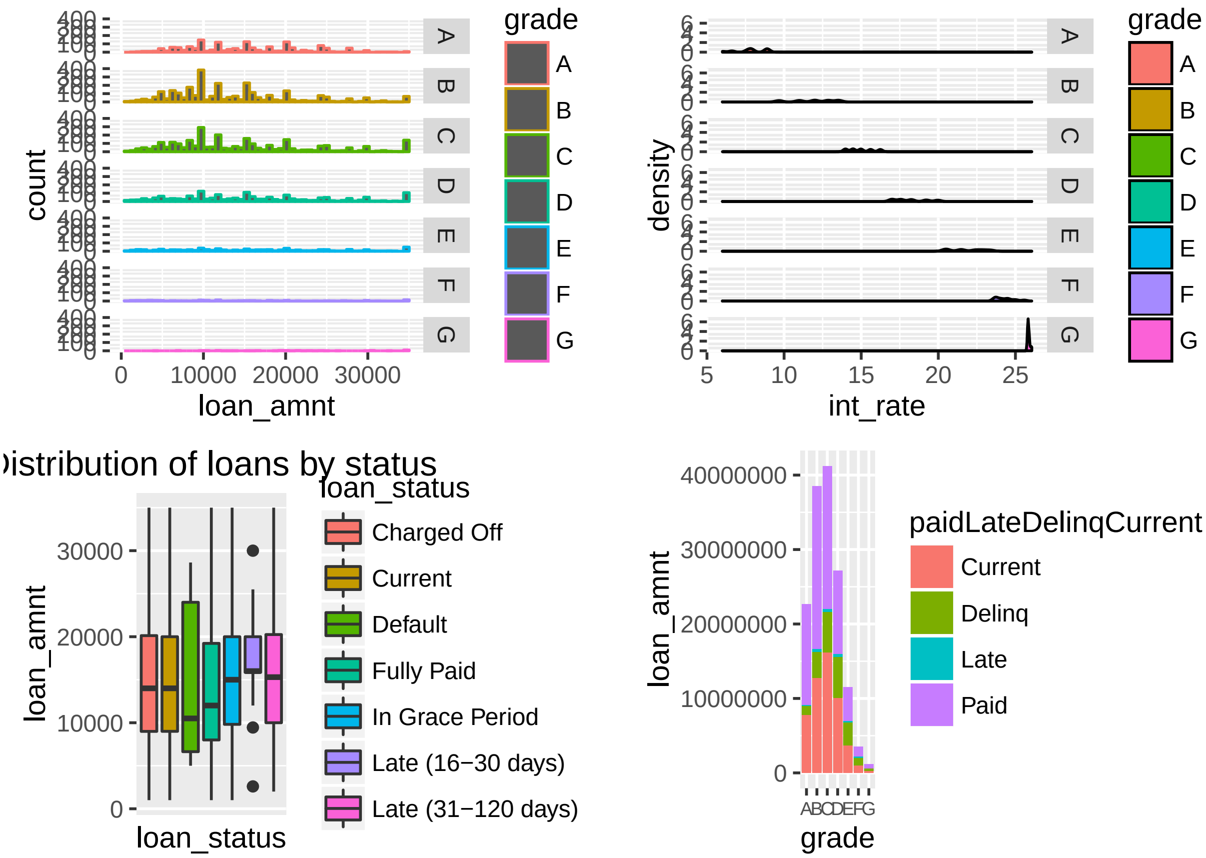

Figure 2

Tidy Data: Convert the date fields like issue date to proper date type, and remove % sign for interest rates, dti and convert those into numeric values.And consider matured loans only. [issue date + term months < today], and remove variables where more than 20% of the observations are missing values.

Factorize the loan status levels with proper ordering.

Lets create a label *bad.loan* which indicates a loan is bad if it is delinquent, or late consider as negative, otherwise positive.

Shortlist Numeric variables: Identify the numeric variable, and compare how those distributions plot against the good, bad label, and pick a few variables with differences in the bad and good populations.(yhat, 2013)

Lets visualize the correlation graph of these numeric variable:

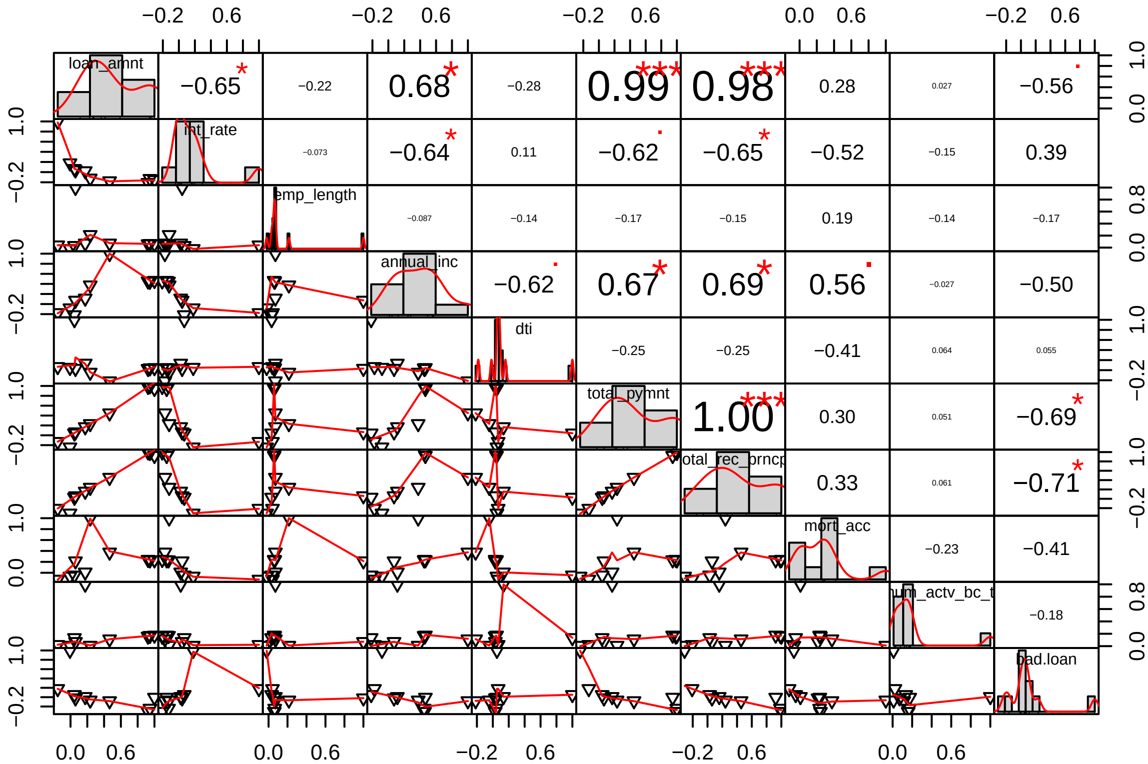

Figure 3

From our numerical feature list, it appears like the *total_pymnt*, and *total_rec_prncp* are having high correlation with the *bad_loan* classification.

Dummy variables: Since R's glm function will take care of the dummy variables (?), we just need to make sure that the categorical variables are factorized, and remove other unnecessary variables.

Split the dataset into training and test: We will randomly split our dataset into training (80%) and test (20%).

Model Development

Model Validation

Conclusion

conclude your findings, limitations, and suggest areas for future work.

References

GiveMeSomeCredit. (2011). Retrieved from https://www.kaggle.com/c/GiveMeSomeCredit

Jitendra Nath Pandey, M. S. (2011). Predicting probability of loan default. Project Report. Retrieved from http://cs229.stanford.edu/proj2011/PandeySrinivasan-PredictingProbabilityOfLoanDefault.pdf

Liang, J. (n.d.). Predicting borrowers chance of defaulting on credit loans. Kaggle Submission. Retrieved from http://cs229.stanford.edu/proj2011/JunjieLiang-PredictingBorrowersChanceOfDefaultingOnCreditLoans.pdf

Shunpo Chang, G., Simon. (2015). Predicting default risk of lending club loans. Project Report. Retrieved from http://cs229.stanford.edu/proj2015/199_report.pdf

yhat. (2013). Machine learning for predicting bad loans. Retrieved from http://blog.yhat.com/posts/machine-learning-for-predicting-bad-loans.html