Loan Default Prediction

Samuel Sears
Summer 2021
https://github.com/ssears219

Domain

The data comes from Lending Club which had traditionally been a "peer-to-peer" lending company.

1. Lending Club website for information about the data - https://www.lendingclub.com/company/about-us
2. Loan Default Prediction competition with several notebooks shared - https://www.kaggle.com/c/loan-default-prediction/code
3. Case Study article on Loan Default Prediction for Profit Maximization - https://towardsdatascience.com/loan-default-prediction-for-profit-maximization-45fcd461582b
4. Academic paper on Peer-to-peer loan acceptance and default prediction - https://royalsocietypublishing.org/doi/10.1098/rsos.191649
5. Case Study article on Loan Default Prediction - https://medium.datadriveninvestor.com/can-you-predict-customers-loan-default-using-machine-learning-be774489b8f5
6. Peer-to-Peer Lending information - https://www.investopedia.com/terms/p/peer-to-peer-lending.asp
7. News article concerning Lending Club's P2P lending model - https://www.fool.com/the-ascent/personal-loans/articles/lendingclub-ending-its-p2p-lending-platform-now-what/
8. Current P2P lending sites - https://www.investopedia.com/articles/investing/092315/7-best-peertopeer-lending-websites.asp
9. Loan Grades - https://www.lendingclub.com/foliofn/rateDetail.action
10. Lending Club review by NerdWallet.com - https://www.nerdwallet.com/reviews/loans/personal-loans/lendingclub-personal-loans?scrollTo=full-review-scroll-target

Data

Data set - https://www.kaggle.com/itssuru/loan-data

The data is composed of 13 predictor variables and 1 target variable. There are 9,579 records. The data set originated from LendingClub.com.

Variable	**Type**	**Description**
credit.policy	categorical (binary)	if the customer meets the credit underwriting criteria of LendingClub.com
purpose	categorical (str)	Purpose of loan
int.rate	continuous (float)	Interest Rate on the loan
installment	continuous (float)	The monthly payment owed by the borrower if the loan originates.
log.annual.inc	continuous (float)	The self-reported annual income provided by the borrower during registration.

dti	continuous (float)	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.
fico	continuous (int)	FICO score
days.with.cr.line	continuous (float)	The number of days the borrower has had a credit line.
revol.bal	continuous (int)	Total credit revolving balance
revol.util	continuous (float)	The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available)
inq.last.6mths	continuous (int)	The borrower's number of inquiries by creditors in the last 6 months
delinq.2yrs	continuous (int)	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years.
pub.rec	continuous (int)	The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments)
not.fully.paid	categorical (binary)	Whether the loan was fully paid or not

Research Questions and Benefits

A Peer-to-Peer lending company connects borrowers with investors. As an investor, you want to invest in in loans that have a high probability of being paid back. If a loan goes into default, the investor loses that portion of their investment. This data is composed of potential information available to an investor during the process of evaluating which loan(s) to invest in. I will analyze what associations exist between these variables and whether or not the loan was paid back in the end. I will then attempt to model to data as a means to predict whether or not the loan was paid back based on information known at the origination of the loan. Having knowledge of the association between these variables and default risk as well as being able to predict default would enable an investor to minimize their risk and attain more profit.

Methods

There are two main tasks for this project that I will be pursuing. First, I will explain the relationship between each variable and the target variable through exploratory data analysis. There can be a lot of data associated with a potential loan investment. Having clear understanding of the most important variables will provide value in the decision process. The second task will be predictive modeling. The dataset is composed of loans with data available at origination and an outcome (not fully paid or fully paid). In predictive modeling, I will determine whether a model could be created that predicts whether a loan will be paid back in full or not.

Potential Issues

This is an unbalanced data set with only 16% of the data represents borrowers who did not pay the loan back in full. Like in my Health Insurance Cross Sell project, I will utilize SMOTE, Synthetic Minority Over-sampling Technique, class weights, and potentially other strategies to ensure the minority class is learned by the model.

LendingClub has discontinued its Peer-to-Peer lending as of last fall. This means the data sources and data dictionaries that the various Kaggle data sets reference is not available anymore. This could potentially cause issues with interpretation and/or validation of the data. Beyond that, it could make the results slightly less valuable since it no longer would be applicable directly to LendingClub.

Conclusion

In the Peer-to-Peer lending domain, investors need to be able to predict which loans will be paid back in full. The better the investors are in choosing the loans they invest in, the more profitable they will become. Through exploratory data analysis and predictive modeling on a LendingClub loan dataset, I will provide valuable insight to investors attempting to minimize their risk. I will highlight the most important variables to focus on when evaluating loan risk and create a model for predicting whether a loan will be paid back in full or not.