THE LOAN RANGERS

Predicting Loan Default in P2P Lending

Javier Recasens | Tanya Nadkarni | Suhas Tummalapalli | Skyler Shasky

Georgia Tech

Introduction

LendingClub is the world's largest online marketplace for peer to peer lending. Users can post their financial profiles in the hopes of getting a crowdsourced loan from a pool of investors. Currently, these loans are graded on a 35 point scale (A1-G5).

Loan Grade	Interest Rate
A	5.3% - 7.9%
B	8.2% - 11.5%
C	12.7% - 15.9%
D	16.9% - 21.4%
E	19.9% - 26.2%
F	24.2% - 30.7%
G	28.6% - 30.9%

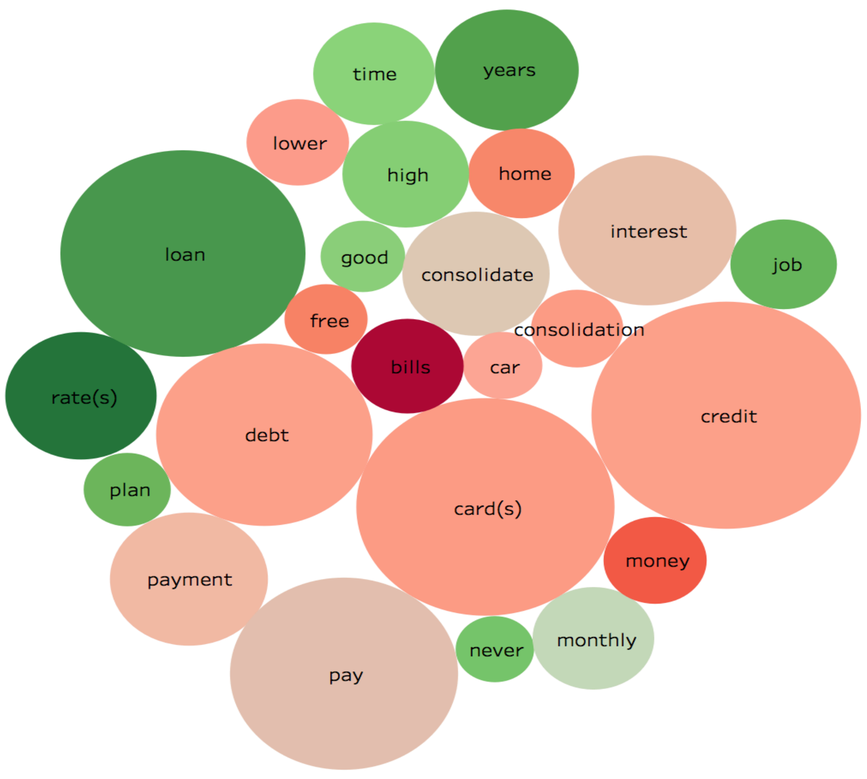
We seek to empower borrowers by providing an interactive tool that gives them the knowledge necessary to achieve the best chance at receiving a loan at the lowest interest rate.

Modeling Approach

We tried multiple classification models, but landed on binomial logistic regression as the best choice. The predictive power is strong and it leaves room for interpretation of coefficients to understand primary drivers for the loan default predictions.

Text Analysis of Loan Description

- Analyzed loan description text to find predictors of loan default rates.
- Identified key statistics such as description length, word count, etc. that are significant predictors of loan default rates and the slope of the relationship.
- Identified specific words whose presence in the description affects the expected loan default rate. Below is a chart representing the magnitude of the affect:



The size of each bubble represents the frequency of the word in the training data.

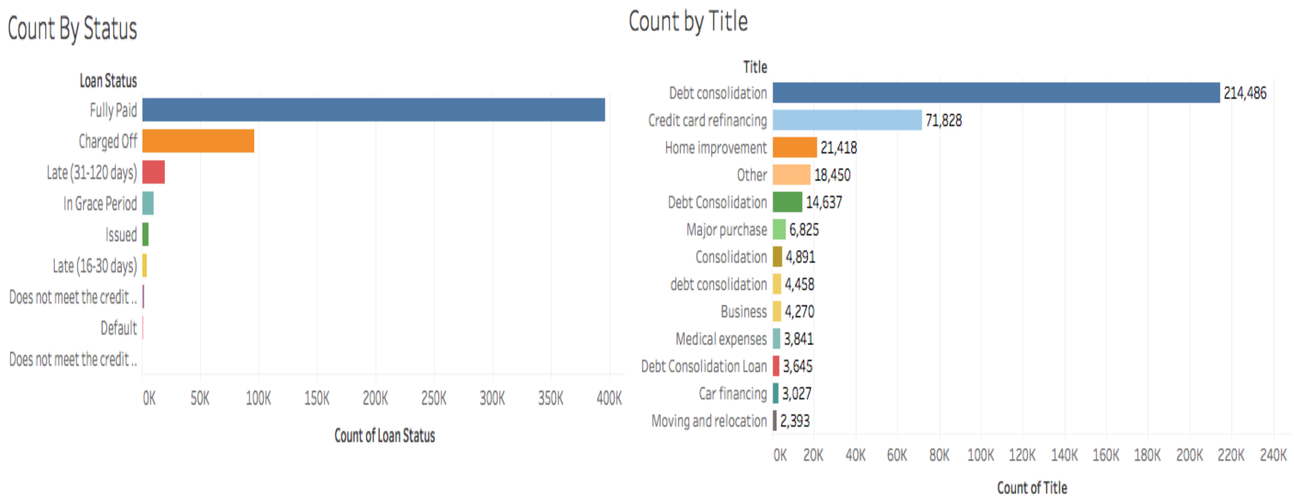
The color of each bubble measures the numeric value of the effect of the presence of the word in the description.

Evaluation

We split the data into a training and testing set. The training set is used to build a model which is then tested on the testing set.
Using a cutoff for the logistic regression of .5, we predict users who will default and compare to actual defaults. After plotting the ROC curve, we are able to validate our model has solid predictive power with an AUC of over .7.

Dataset (downloaded from lendingclub.com)

- 1GB dataset. Loans issued from 2007 to 2016.
- 1.3 million observations and 113 predictors.
- Joined IRS Tax Statistics data with Zip Codes.
- We narrowed this down to include only loans issued after 2014 due to lack of information in previous years.
- Each row contains information about the loan status and payment information, as well as financial demographics of the borrower such as credit score, income, and debt to income ratio.



Interactive Dashboard

We provide users with a customizable interface where they can input information about their personal financial situation. The app will provide key information such as their perceived likelihood of default, and the relative effects of each of the variables that are used in the predictive model