Predicting Borrower Risk using Lending Club's loan data Augmented with Alternative Data

Neville Ed Jos, Subhash Talluri, Hamdan Siddiqui
Master of Science
University of Illinois Urbana Champaign
Champaign, IL, USA

The current US system of lending is not very good at assessing the risk of borrowers. It relies on a singular value known as a credit score which is biased against people who have had one late payment or people with no credit history. To cover the uncertainty of borrower risk, lenders set a relatively high interest rate. This leads to people having to pay for high interest loans even though they aren't risky enough. The goal of this paper is to challenge Lending Club's approach by assessing borrower risk without taking a look at credit score. Outside datasets and APIs are used to augment existing dataset. Through the process of trial and error, a model with new attributes was found that show a better probability of predicting risk than Lending Club's models.

Keywords—lending club; loan; risk; borrower; default

I. INTRODUCTION

The current US system of lending has some flaws. Mainly, it is not very good at assessing the risk of borrowers (look at the 2008 recession for example). Most major banks in the US rely regularly on a singular value to assess the risk of a borrower, known as a credit score. A credit score is a number assigned to a person based on their ability to make payments on previous credit cards, loans, mortgages etc. Though that may seem like a decent indicator of risk, it has been notoriously erroneous in actually predicting the chances of default for a borrower, since it is extremely biased against poor people who have had one late payment or people with no credit history at all [1]. Banks, knowing that credit scores are bad indicators (while still using them), will then set a relatively high interest rate to cover the uncertainty of borrower risk. This leads to people having to pay for high interest loans even though they aren't risky enough to deserve that high interest. Lending club [2] provides public access to data on their loans with attributes such as customer income, credit score, state, credit card utilization, home ownership etc. as well as information about the loan (loan amount, term, and whether paid back or defaulted). The dataset categorizes the risk of each borrower by their credit score and sets interest rates for them appropriately.

Our goal is to challenge Lending Club's approach by assessing borrower risk without taking a look at credit score. We will be using outside datasets and APIs to augment our existing dataset. For example, we will be pulling in Zillow to analyze average house prices for a zip code, consumer price index to measure cost of goods at the time of borrowing, unemployment rates from the US bureau of Labor Statistics, population density by county from US Census bureau and any other datasets we can get our hands on. Through this process of trial and error, we hope to find attributes about a borrower that predict better borrower risk, which can be used to give out better, low-interest loans to those who actually deserve it the most.

II. RELATED WORK

There has been lot of work done on Lending club loan data such as "Predicting borrowers chance of defaulting on credit loans" [3], "Predicting Probability of Loan Default" [4] and "Peer Lending Risk Predictor" [5]. They have applied machine learning to improve loan default prediction. The authors have compared the performance of various models such as Random Forest, Naïve Bayes and SVM's. The more recent "Predicting default risk of Lending Club Loans" [6] has added in census data, with info like regional median income, to train models on a more holistic set of features. We are taking this approach one step further by augmenting our current data set with outside data such as Zillow's average home price index, consumer price index at the time of borrowing and unemployment rates from the US bureau of Labor Statistics. Through this approach, we hope to show the true risk of a borrower without using credit score as an indicator of risk.

III. INTELECTUAL MERIT

In the age of information, there are troves of data being collected about the United States and its people every single day. The difficulty in this project is to be able to discern the data that have the ability to predict borrower risk, which means that our group had to spend as much time on reading about the causes for loan default as we did reading on analyzing the data. A purely computer science approach would not have factored in the nuances of the lending industry and the factors that impact it.

IV. BROADER IMPACT

The United States is one of the best countries to borrow money from, even if the methodology of using credit score is no the optimal method to do so. That is because, in comparison, other countries have no method at all for calculating borrower risk. Countries like Brazil can charge

upwards of 70% (Brazil Bank Lending Rates, 2017) as interest rates for personal and business loans due to the inability to properly access credit risk. Our model is extensible to work in any country since it does not rely on a specifically calculated credit score but rather data available to the public in almost every country such as home prices, treasury bill rates, unemployment rates etc. Since this model can use this public data to access borrower risk, banks can dramatically reduce uncertainty premium of loan interest rates and provide more accessible funds to people.

V. Challenges

The main challenge that we faced is in finding datasets that are detailed and free. Several of the datasets that we originally thought we could use were in fact either premium datasets or not as detailed as we would have liked (for e.g. annual data vs. country level data). Since we couldn't find several of the datasets we hypothesized would exponentially increase the quality of predicting borrower risk (for e.g. county crime data), we had to rework our premise from being that we can predict "true risk" to the premise that we can improve the current risk prediction models.

Furthermore, due to the Equal Credit Opportunity Act of 1974 (Fair Lending n.d.), all underwriting models in the United States have to be able to dictate the top 3 reasons why a loan was rejected/approved. This allows for the government to validate that a loan was not rejected on the premises of race, age, ethnicity, sex, sexual orientation etc. Though this law is great for consumers, it does inhibit the ability for modern underwriting systems based on Neural Networks, Random Forests, Clustering Algorithms etc to be used as underwriting systems since it is extremely difficult to be able to dictate the features that led to a loan being labeled as being charged-off in the future vs. not. This limitation dramatically decreases the types of machine-learning techniques we can use to build underwriting models.

VI. Methods

Lending Club is an online marketplace for peer-to-peer lending that connects borrowers and investors. Their platform enables borrowers to access loans through a fast and easy online or mobile interface. Investors provide the capital to enable many of the loans in exchange for earning interest.

Data provided by Lending Club consists of attributes such as borrower annual income, loan interest rate, and borrowing habits. Data is broken down into sets by period. For each period, the data describes a list of all loans issued, including current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The data set contains Lending Club's loan data from 2007-2015. It has a total of 887,379 observations of 74 variables.

A. Data Creation

Loan data is retrieved from Lending Club for each of the four financial quarters from year 2007-2015. These files were combined together with the variables that were influential and relevant to the project at hand (Table I).

The following datasets were used to augment the Lending Club data:

(1) Average Housing Prices for a zip code (from Zillow API)

(2) Consumer Price Index (from St. Louis Federal Reserve Bank API). CPI Index tracks the average price of all goods in the US annually. It is also used to calculate inflation in the US. The hypothesis is that higher CPI means that goods are more expensive to buy and a person will have less money to pay for their loans and therefore have a higher chance of defaulting.

(3) Unemployment Rate by state (from Bureau of Labor Statistics)

(4) Historical 3-Month Treasury Bill Rates (from St Louis Federal Reserve Bank API). Treasury Bill Rates are cheapest interest rates in the US. They determine the interest rates of almost every loan a person gets. The hypothesis is that higher interest rates lead to higher default rates. We hope that these data sources can better predict borrower risk than Lending Club's current modeling system.

After reviewing the data, we decided to remove the variables 'sub grade", 'int rate', 'fico range low' and 'fico ra nge high' since they are both values created by Lending Club after their underwriting process, which is what this project is trying to replicate.

TABLE I. Description of Loan Data Variables

Variable Name	Description
loan_amnt	Amount borrowed for loan
term	Length of the loan
emp_length	Length of the borrower's currently employment
home_ownership	Length of the borrower's home ownership
annual_inc	Annual Income of the borrower
verification_status	Boolean dictating whether the income of a borrower has been verified
loan_status	Boolean dictating whether the loan defaulted or not
zip_code	First three digits of the borrowers zipcode
addr_state	State of residence of the borrower
delinq_2yrs	Number of delinquencies the borrower has over the last two years
earliest_cr_line	Date of the earliest credit line of the borrower
inq_last_6mths	Number of inquiries into the borrower's credit score in the last six month
mths_since_last_delinq	Number of months since the borrower's last delinquency
mths_since_last_record	Number of months since the last activity on the borrower's credit score
open_acc	Number of borrower's open accounts
pub_rec	Number of borrowers' derogatory public records
revol_bal	Number of borrower's revolving accounts
revol_util	Amount of revolving account credit being used
total_acc	Total number of accounts held by the borrower
z_index	Home Price Index for the US from Zillow

Variable Name	Description
cpi	Consumer Price Index of the US
3_month_treasury	3-Month Treasury Bill Rates in the US
unemployment_rate	Unemployment rate in the state of residence of the borrower

B. *Data Cleaning*

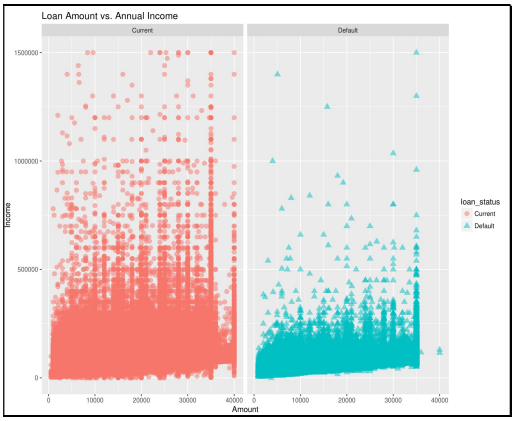
The data has been preprocessed to verify for any invalid characters, null or missing values. We examined the dataset closely as not all the variables were useful. Any presence of anomalies in the variables that were not significant have been removed. A subset of the data was defined with variables that were most interesting (Table I). Some data needed transformations of a few string variables into numerics, removing or recasting na's, and the creation of a few binary variables.

C. *Data Exploration*

The preprocessed dataset that is free from anomalies is used for exploratory analysis. A binary variable "default_risk" was created based on whether the "loan_status" was "Paid Fully" or not.

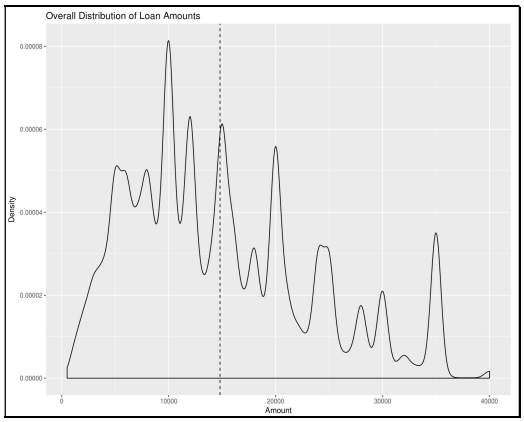
There is quite a bit of variation in the annual incomes of people applying for all different loan amounts (Fig 1). There's also quite a few observations that appear to be outside the plausible range. How many people with an annual income of $9.5mil are applying for $24k loans on a peer-lending platform? There seems to be some patterns in specific income verticals.

Fig. 1. Loan Amount vs Annual Income



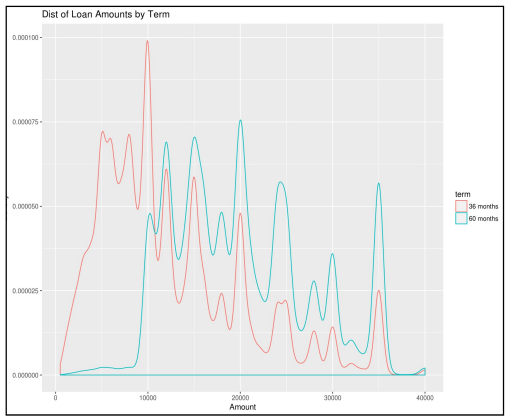
It appears that highest number of loans are centered around $10k range, with a mean around $15k (Fig 2).

Fig. 2. Overall Distribution of Loan Amounts



As might be expected, we can see that 36-month terms are more common for smaller loan amounts, and almost non-existent for loans over roughly $27.5k (Table 3).

Fig. 3. Distribution of Loan Amounts by Term



D. *Data Modeling*

Due to the limitations described in the "Challenges" section, we are left with mainly using Logistic Regression and Naïve Bayes as our primary methodologies of modeling, since they are able to dictate how much each factor affects the final result.

The models were built on the cleaned data that we formed with "default_risk" as the response variable and the other variables set as the predictor variables. To ensure that our models were not over-fitting, we cross validated by creating training data using 80% of randomly chosen rows from the original data frame and the rest of the 20% of rows set as test data.

VII. Result

TABLE II. Model Accuracies

Model Type	Accuracy
LendingClub (Base)	*86.13%*
Logistic Regression	**86.16%**
Naïve Bayes (Gaussian)	85.44%
Naïve Bayes (Multinomial)	60.13%
Naïve Bayes (Bernoulli)	86.16%

TABLE III. Logistic Regression Features

Feature	Coefficient
z_index	**1.06628185161e-05**
loan_amnt	1.3251799826e-07
revol_bal	1.31596786693e-07
annual_inc	7.93945301355e-08
cpi	**3.63177034105e-08**
revol_util	1.51111414715e-08
unemployment_rate	**3.60462022742e-09**
mths_since_last_record	3.30176986866e-09
total_acc	2.10427712198e-09
mths_since_last_delinq	2.05043732339e-09
new_emp_length	1.02910445539e-09
inq_last_6mths	3.34900616867e-10
Verified	2.71306003743e-10
open_acc	1.59476488517e-10
Source Verified	1.28189020382e-10
RENT	1.0164943106e-10
MORTGAGE	9.92736794805e-11
pub_rec	9.26893127116e-11
delinq_2yrs	6.9798923711e-11
OWN	2.1936707264e-11
3_month_treasury	**1.45612245884e-11**
OTHER	6.93504761739e-13
NONE	5.00962255994e-13

VIII. Evalution

From the comparison of the models we could use, the best was the Logistic Regression model. It yielded an accuracy rating of **86.16%**. We assume that the Logistic Regression model fit best since around half of the data did not fit a Gaussian distribution (home_index, loan_amount etc) and therefore were harder to predict using those models. This accuracy of 86.16 is slightly higher than the default risk calculated by the LendingClub model of 86.13% (based on actual default rate that LendingClub experienced from the years of 2007 – 2015).

Since we chose Logistic Regression, we can find the most impactful variables that accounted for the prediction of loan defaults (Table III). As shown, the newly created variables of 'z_index', 'cpi', and 'unemployment_rate' rank amongst the top 10 factors that accounted for loan defaults, out of the total 25 factors, with 'z_index' (The Home Price Index of the US from Zillow) topping the list of most important features. '3_month_treasury' did not have much of an effect on the data as we had originally thought.

IX. Conclusion

Though we removed FICO credit scores, LendingClub loan grades and interest rates from the provided data, our dataset consisting of other variables plus newly created variables of 'z_index', 'cpi', 'unemployment_rate' and '3_month_treasury' from publicly available data sources could predict loan default rate as well as the LendingClub underwriting models could.

X. Works Cited

2017. *Brazil Bank Lending Rates.* http://www.tradingeconomics.com/brazil/bank-lending-rate.

n.d. *Fair Lending.* http://www.aba.com/Compliance/Pages/FairLending.aspx.

XI. References

[1] Gillian B. White (2017, January 10). Can the flaws in credit Scoring be fixed? Not easily. Retrieved from http://www.theatlantic.com/amp/article/512702

[2] Lending Club. (2006). Lending club statistics. Retrieved February 25, 2017, from https://www.lendingclub.com/info/download-data.action

[3] Liang, Junjie. "Predicting borrowers' chance of defaulting on credit loans."

[4] Pandey, Jitendra Nath. "Predicting Probability of Loan Default Stanford University, CS229 Project report Jitendra Nath Pandey, Maheshwaran Srinivasan. Stanford University.

[5] Tsai, Kevin, Sivagami Ramiah, and Sudhanshu Singh. "Peer Lending Risk Predictor."

[6] Shunpo Chang, Simon Dae-oong Kim and Genki Kondo. "Predicting default risk of Lending Club Loans". CS229, Machine Learning. Autumn 2015-2016. Stanford University.