LENDING CLUB LOAN DATA ANALYSIS

STA 141A FINAL PROJECT

Abstract

We examined information on over 800,000 loans provided by SF-based peer-to-peer lending service, Lending Club, to look at trends in types of loans, interest rates, and credit risk modeling.

Niveditha Achanta, Paul Kim, Sophia Lee

Lending Club Loan Data Analysis

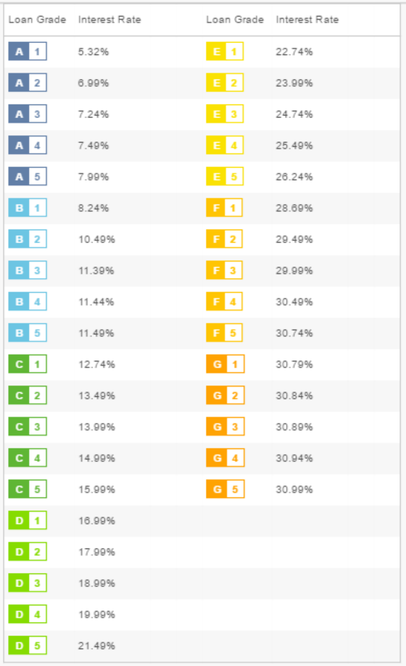
Members: Paul Kim (999140819), Sophia Lee (999720705), Niveditha Achanta (998857938)

Roles: Paul Kim and Niveditha Achanta focused on gathering appropriate datasets[1], cleaning the data, and determining how to graphically represent variables. Sophia Lee used trends and information in the data to model credit risk and build a regression to predict loan default rates. Each group member took care of roughly one part of the analysis, although we collaborated throughout the process, since some parts were more difficult and time-consuming than others.

Introduction

For our project, we will be starting with the analysis of a dataset containing information for over 800,000 loans from Lending Club, a peer-to-peer lending service based out of San Francisco[2]. As described in Bloomberg magazine, Lending Club sets itself apart by being a hallmark of transparency: “Lending club “matches users who need money with investors willing to lend…Loans of up to $40,000 at a time and are divided into $25 securities that anyone can buy, and Lending Club publishes detailed information about the loans in daily filings with [the SEC]. It maintains publicly available spreadsheets…about [prospective] borrowers.[3]”

Lending Club issues grades (and subgrades) to each loan. Per their website, “Based on each loan application and credit report, every loan is assigned a grade ranging from A1 to G5 with a corresponding interest rate.” Below is a table that visualizes the information, as given by Lending Club:



Loan Grade	Interest Rate	Loan Grade	Interest Rate
A 1	5.32%	E 1	22.74%
A 2	6.99%	E 2	23.99%
A 3	7.24%	E 3	24.74%
A 4	7.49%	E 4	25.49%
A 5	7.99%	E 5	26.24%
B 1	8.24%	F 1	28.69%
B 2	10.49%	F 2	29.49%
B 3	11.39%	F 3	29.99%
B 4	11.44%	F 4	30.49%
B 5	11.49%	F 5	30.74%
C 1	12.74%	G 1	30.79%
C 2	13.49%	G 2	30.84%
C 3	13.99%	G 3	30.89%
C 4	14.99%	G 4	30.94%
C 5	15.99%	G 5	30.99%
D 1	16.99%		
D 2	17.99%		
D 3	18.99%		
D 4	19.99%		
D 5	21.49%		

[1] http://financerecipes.com/cleaning_data.html
[2] Data will later be segmented into subcategories. We use the broadest dataset provided for the graphical representation, and a smaller one (roughly 400,000 rows) for the regression, since it is cleaner.
[3] http://www.bloomberg.com/news/features/2016-08-18/how-lending-club-s-biggest-fanboy-uncovered-shady-loans

Lending Club[4] also collects data about the purpose of loans, status, various dates for payments, and more. There are 75 variables for each loan, each recording a different piece of information.

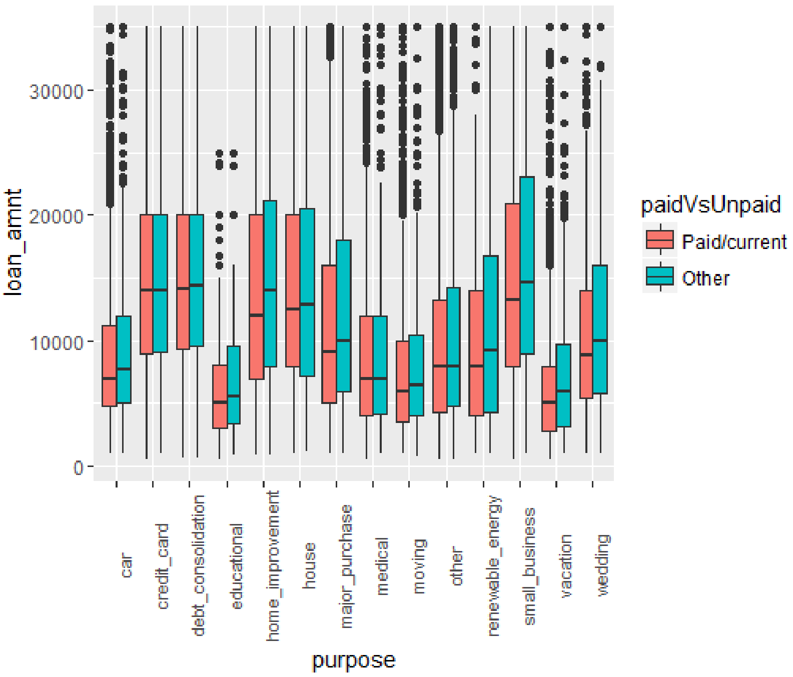
Objectives: Our data analysis will focus on:

1. The purpose of a loan and its payment status.
2. The grade of a loan and its effects on the payment plan and other factors.
3. Lending Club's current method of grading loans to better predict default rates or "bad" loans.

Part I: Purpose and Payment Status

We can use the aesthetic features of ggplot in R to look at distributions of loans across grades consolidated for subgroups, as shown in the following graphs[5]. The first graph will show the distribution of the loan by different purposes, and the second will look at the paid vs. unpaid loan amount across purposes.

This graph shows the distribution of the loan by different purposes. We note that the spreads and outliers of these loans vary greatly for the purposes they serve. For example, the "major purchase" purpose has a cluster of paid/current loans above $30,000 (and luckily for Lending Club, that cluster does not exist for the blue, unpaid section).

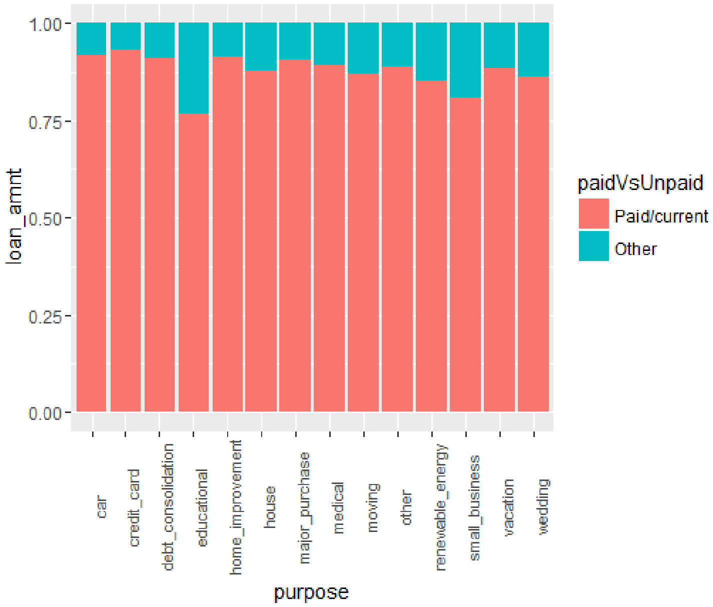


One thing to note is that a lot of this is definitional. These categories try to be as specific as possible, but "major purchase", for example, is broadly defined, which may be one reason the loan amounts are so skewed. Some of the loans without great skew or outliers include credit card, debt consolidation, and small business.

[4] https://www.lendingclub.com/

[5] Using our own analysis as well as visual representations from https://www.kaggle.com/ashokn30/d/wendykan/lending-club-loan-data/lending-club-data-some-insights

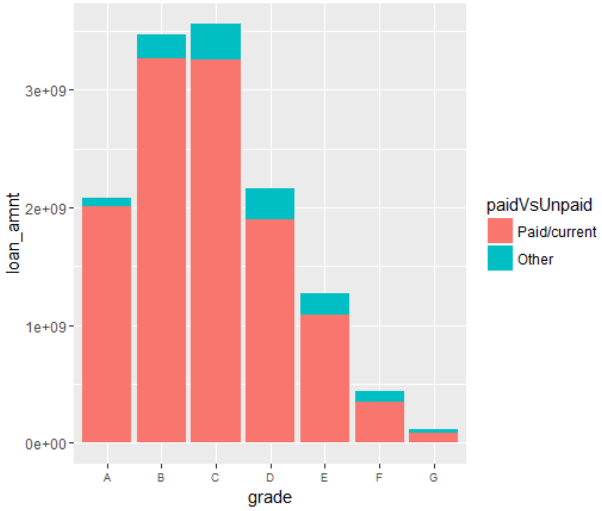
After looking at the distribution of loan amounts by purpose, we wanted to glean more information about the proportion of loans that were paid. Loans taken out for educational purposes seem to have the highest unpaid percentage, whereas credit card loans have the lowest. We found it interesting that the education unpaid percentage was so high, especially since, from the last picture, that category had one of the lowest mean loan amounts. The blue section represents unpaid loans, and the salmon represents paid or current, which means that this discrepancy cannot be attributed to the length of the loan.



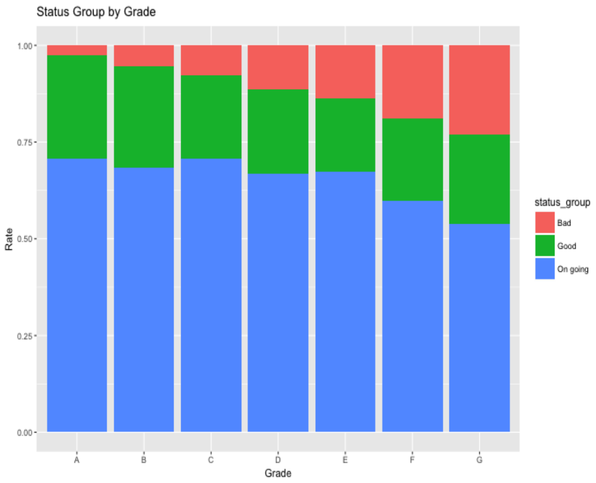
Part II: Grades and Interest Rates

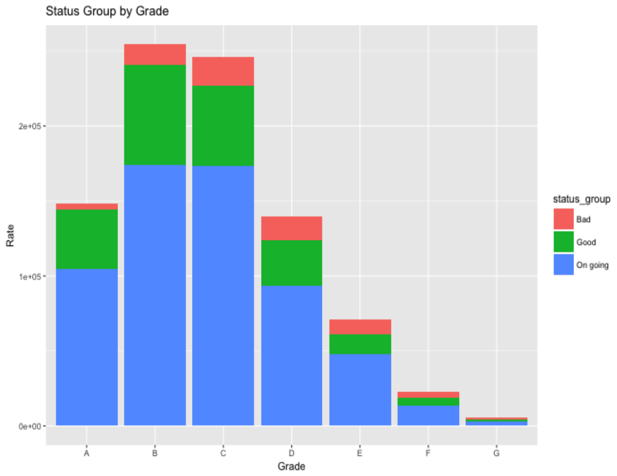
Although there is a lot more examination to be done about the purpose of these loans, we wanted to focus on the grades that Lending Club assigns, since it is unique to their business model.

An “A” represents the best possible grade (based on many factors such as credit score), and a “G” represents the worst. From the graph below, we see that C- and D- rated consumers have the highest unpaid proportion.



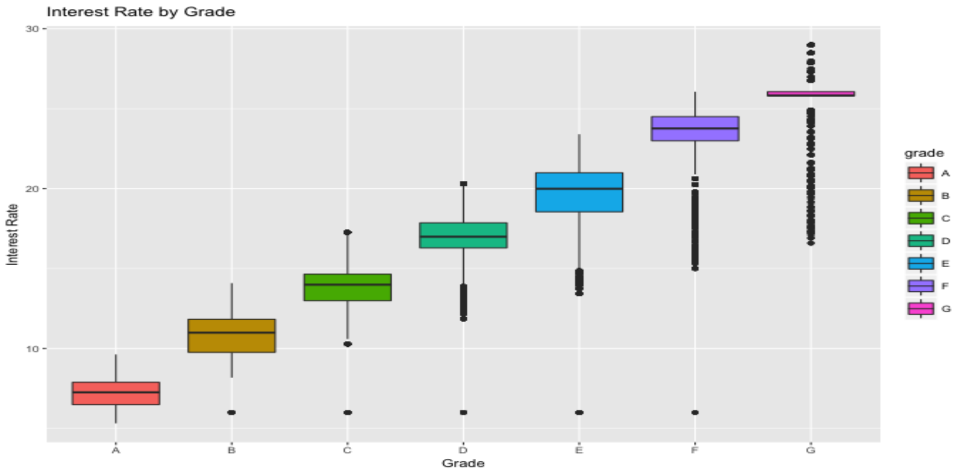
We can also separate this loan status further into 3 different categories: Ongoing, Good, and Bad, to examine more trends:





We note that the "bad" loans increase as we go from A through G, and that the rates all seem to have relatively proportional lengths to the different categories Bad, Good, and Ongoing.

Additionally, we can examine the interest rate by grade to see how these grades determine the kind of interest rate Lending Club gives them:



In accordance with our graph, we see that there are a lot of outliers for the grades D, E, F, and G against interest rate. This goes to show that the grades D, E, F, and G don't correlate with interest rate, while grades A, B, and C do correlate with interest rate because they have few to no outliers. We will take this into account while considering the information used to build effective models.

Part III: Modelling

Lastly, to reach our proposal's goal of creating an accurate predictor model for default loan rates, we downloaded another dataset from the Lending Club Website[6]: Due to the large size of the csv file, and several missing variables, we first had to do some data cleaning.

[6] https://www.lendingclub.com/info/download-data.action.

Loan Status After Standardization

loan_status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Charged Off	5658	14.30	5658	14.30
Fully Paid	33902	85.70	39560	100.00

For instance, for the predictor variable "loan_status" we removed Lending Club's records where the status of the loan was not specific to "charge off" or "fully paid". We did this to remove records where the loan status was a partial payout— we wanted our model to be built off of data with a definite outcome. Moreover, these partial payouts would interfere with the model being able to distinguish a "charge off" from a "fully paid" record. Overall, the purpose of the exercise was to predict the status of the loan (fully paid or charged off) using the variables in the data set. We defined a positive outcome as a fully paid account and a negative outcome as an account the charged off (closed with a loss).

Loan Status Before Standardization

loan_status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	3	0.01	3	0.01
Charged Off	5658	13.30	5661	13.31
Current	201	0.47	5862	13.78
Default	1	0.00	5863	13.78
Does not meet the credit policy. Status:Charged Off	761	1.79	6624	15.57
Does not meet the credit policy. Status:Fully Paid	1988	4.67	8612	20.25
Fully Paid	33902	79.70	42514	99.94
In Grace Period	9	0.02	42523	99.96
Late (16-30 days)	5	0.01	42528	99.98
Late (31-120 days)	10	0.02	42538	100.00

We also "cleaned" Lending Club's data by separating it into two randomly-sampled data sets— a development dataset and a validation data set. The original Lending Club csv file contained 39,560 records; therefore, 26,459 records (about 70%) were randomly assigned to the development dataset and the remaining 13,101 records (about 30%) were randomly assigned to the validation data set. This was done to ensure that during our research and investigation that we did not incorporate bias assumptions that would result in over fitting the final regression model. Moreover, we performed all our research investigation on the 70% training (development) sample. Then when our final regression was complete, the same regression was applied onto the 30% validation dataset. If the summary results were similar between the two regressions, we would have reason to believe that our model is statistically robust. Otherwise, a difference between the two regressions would mean that we made invalid assumptions, our model is over fitted, and we need to revisit our methodology.

From previous examination of Lending Club's original dataset, we chose to *not* use "grade", "subgrade", or "interest rate" as variables for our regression model because these variables were already defined by Lending Club. Additionally, we wanted our statistical model to be independent of Lending Club's grade assignment practices, and rather, use it as a benchmark comparison for model. Therefore, for regression model we used the following variables:

i) Annual_inc – higher annual incomes should be positively correlated with paying back the loan

ii) Revol_util - Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. Individuals with higher utilization rates would have higher risk, since they could be using credit beyond their financial means.

iii) Loan_Purpose – loan purpose describes the borrower's reason for taking out the loan

iv) Loan_term – this variable is 36 or 60 month loan period. We hypothesize that an individual seeking a longer-term loan will exhibit higher risk since they are seeking a lower monthly payment, indicating that they might be financially stressed.

v) Inq_last_6mths – this is the number of credit bureau inquires in the last 6 months. Individuals with more recent inquiries would indicate higher risk, since the individual is actively looking for credit and most likely have been turned down.

vi) Funded_amnt – the funded amount of the loan that the borrower is requesting.

vii) Pub_rec – public records consists of bankruptcies and foreclosures.

In order to begin creating our model to predict loan default rates (based on the variable "Bad"), we first had to fit our development dataset using previous payment and income information (the variables listed above). Next, as a comparison, we fitted a regression model based on Lending club's benchmark predictor variable "loan grade". After we observed that our model's adjusted r-squared value was indeed higher than Lending Club's grade-based model, we fitted our model without loan purpose to see how it affected the overall accuracy of our default estimator. After taking a summary of this new fit model, we saw that our r-squared value only lowered by .00002, which is not proportionately significant enough to assume that loan purpose has a significant effect on our model's prediction accuracy. In other words, the loan purpose information by itself may be useful for Lending Club, but not in combination with the other variables in our model. In addition, models with the fewest number of variables are generally said to be the preferred model because each variable provides a higher contribution towards the predictor variable, in this case, the loan status. Removing variables also reduced multicollinearity within our model, which could have artificially inflated our regression's adjusted r-squared value.

In relation, for our validation dataset we fitted our regression model without loan purpose since this variable did not provide additional predictiveness in determining a charge-off vs. fully paid account. We then validated the findings of our development model by using the command "compfit" to make an anova comparison of our larger-variable model and our model without loan purpose. As concluded in our development dataset analysis, the p-value of our comfit was greater than .05, therefore our predictor model is not significantly less accurate without the information provided in the variable category loan purpose. Lastly, we calculated a final summary of our regression model (without loan purpose) and our r-squared value comfortingly was larger than our development regression model r-squared value; the closer this value is to 1 means that our validation model is even more accurate than Lending Club's grade-method predicting model!

Closing Remarks

Due to the sheer size of the datasets and the various types that we could clean and use, we suppose there are some discrepancies between our data visualization and our modeling. However, it is all pulled from Lending Club’s data, so we have a centralized source. Given that our validation model exceeded our expectations, given more time, we would have tested this to see if this always holds true, and to try to apply it to other major financial services companies. Lending Club has historically opened its data up to major companies who use this information to do things like what we did for our final project, but on a smaller scale. If we had a longer time frame for this project, we would be able to delve more deeply into each of our objectives.

One that is most relevant to us which we only briefly touched upon is education. It may be possible, through grades and other consumer information, to examine why education has the highest amount of unpaid loans. Models built from Lending Club data could potentially give us insight into the relationship between education and credit risk, as well as other variables that we would examine in the future.

Finally, if we had a greater allowance for time for this project, we would be able to use more sources. Our primary source for already-analyzed data has come from Kaggle, where we found a link to this dataset, but since Lending Club data is one of the most analyzed pieces of information in the financial sector, it would be interesting to try to incorporate and compare as much analysis as we could, to determine which companies do a better job of credit risk modeling and prediction.

In summary, this dataset can be manipulated to provide an incredible amount of insight into companies’ decision making. In recent years, Lending Club has been recovering from the effects of subprime lending, which is another point of interest we could have tried to find. The data is easy to visualize and glean trends from, and it shows that even students with basic programming abilities and statistical knowledge can build models that are implementable in the real world.