An Investor's Guide to Lending Club

### Introduction

Lending Club is an online P2P lending platform that offers unsecured personal loans by connecting borrowers with investors who wish to back their loans. From the borrower's side, P2P lending services such as Lending Club are an increasingly popular new way to take out personal loans since the application process can be completed in under 30 minutes from the comfort of one's home, the approval process typically takes less than a day and the funds are wired directly to the borrower's bank account in 2-3 business days. Additionally, the size of these online marketplaces ensures that a borrower can find interest rates that are reasonable given their credit rating. From the investor's perspective, online lending platforms represent an investment opportunity that frequently offers higher returns than traditional investments such as stocks, bonds or mutual funds.

### About the Data

Data for personal loans issued through Lending Club is available going back to 2011 on the following website: <https://www.lendingclub.com/info/download-data.action>.

For this project, I chose to analyze loans issued in calendar years 2014 and 2015 for two reasons:

* Loans are issued with 36-month or 60-month payment terms, so data from several years ago will offer a more clear picture on each borrower's ability to repay.
* Going any further back in time may cause data to be of lesser quality. Furthermore, there may be macroeconomic trends in recent years that would cause older data to produce different results during analysis.

### Objective

This analysis will attempt to accomplish multiple tasks:

* Identify specific characteristics of loans that cause the probability of defaulting to be higher or lower.
* Create and refine a model that can be used to separate successful loans from potential defaults.
* Develop an investment strategy that will yield the highest overall ROI.

### An Overview of Data Wrangling

The original dataset consisted of 650,000 rows each representing a loan and 137 columns describing various details of the loans. In order to be of use, the data had to be thoroughly cleaned, which in this case primarily meant removing or imputing missing values. Columns with missing data were addressed first and in order ensure that each column had enough data to analyze, any column that was less than 15% populated was removed. For many of the remaining columns with missing data, new variables were created as factors indicating the presence of a value in the original variable. In the below example, the number of months since a borrower's last delinquency is converted to a factor describing whether or not the borrower had a delinquency in their credit history:

loans$ever\_delinq <- ifelse(is.na(loans$mths\_since\_last\_delinq), 0, 1)

This approach worked well for many of the variables that would indicate potentially problematic items in a borrower's credit history, such as recent bankcard delinquencies, payments that were over 120 days past due, or recent inquiries. There were a number of columns where missing values were imputed with medians:

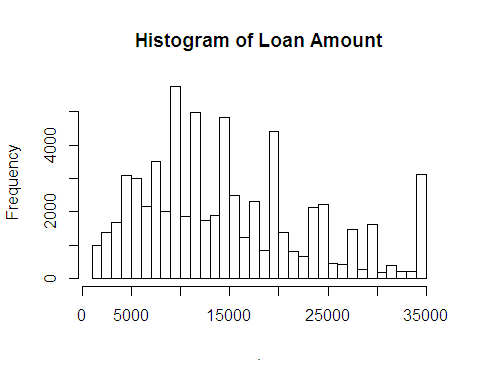
* Average current account balance
* Bankcard open-to-buy
* Bankcard utilization
* Months since most recent bankcard was opened

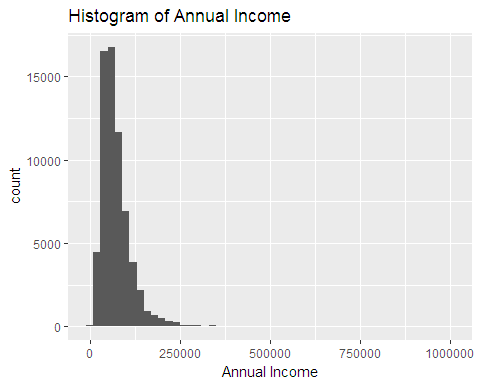
For the percentage of bankcard accounts where spending surpassed 75% of the card's limit, a missing value most likely indicated that the borrower does not have a bankcard. Missing values were therefore imputed with zero's. The same logic was applied to the number of revolving accounts. Employment length was imported as a factor consisting initially of 12 levels ranging from less than one year to more than 10 years, but one of level was the character string "n/a". The variable was collapsed to 11 levels so that "n/a" would actually read as a missing value. These were kept in order to analyze the connection between missing employment data and defaults. The issue date of each loan was listed as a character string containing an abbreviated month and two digit year. These were split into separate columns containing the full name of the issue month in one column and the 4-digit issue year in the other. Finally, the loan status column was used to created a new column indicating whether or not the loan had gone into default (charge-offs and defaults were included).

### Analysis

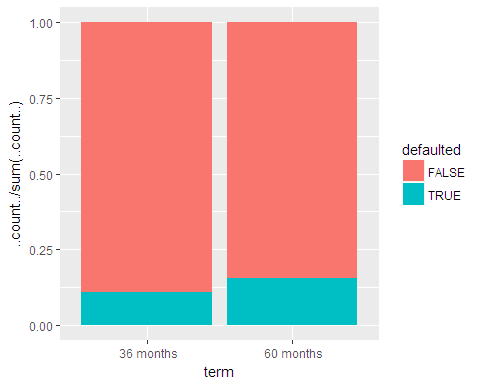
Due to computing constraints, the data was downsampled from 650,000 loans to 65,000. The analysis primarily consisted of sample mean comparisons across different segments of the dataset as well as hypothesis testing. Additionally, visualizations are provided to make the findings more impactful.

First, a look at loan amounts and annual income:



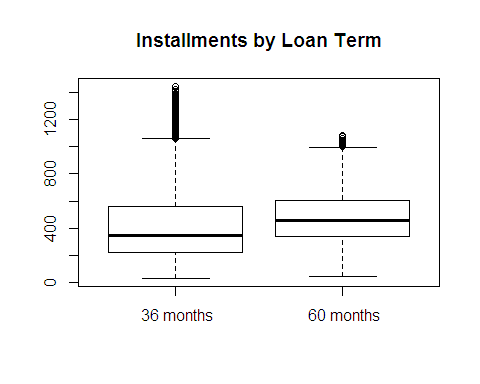


Loan amount is a slightly right tailed distribution with a large number of borrowers taking out the max loan amount of $35,000. Borrowers typically apply for loans in multiples of $5,000. Annual income is also a right-tailed distribution with a very small number of borrowers earning between $250,000 and $1,000,000 annually and 75% of borrowers earning less than $90,000/year. A quick calculation of the sample proportion of defaults revealed that 12.4% of all Lending Club loans go into default. This provides the benchmark for comparing the default rate among smaller segments of this sample. A review of the most important explanatory variables follows. Generally, a longer term on a loan reflects a desire to spread out repayment of the loan over a longer period of time so as to lower the monthly installments. A longer loan term therefore may be indicative of certain budget constraints for the borrower which may eventually cause difficulty making payments.

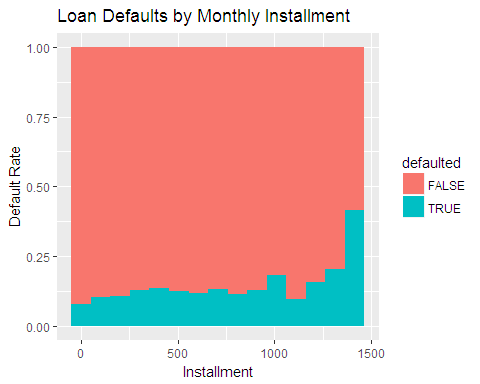


##   
## 2-sample test for equality of proportions with continuity  
## correction  
##   
## data: c(sum(loans60), sum(loans36)) out of c(length(loans60), length(loans36))  
## X-squared = 301.25, df = 1, p-value < 2.2e-16  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## 0.04216404 0.05363156  
## sample estimates:  
## prop 1 prop 2   
## 0.1566660 0.1087682

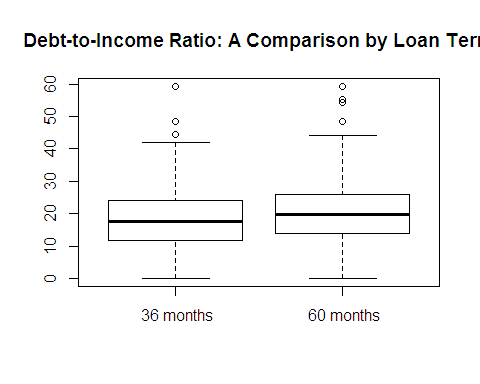
Borrowers on a 60 month loan term defaulted at a rate of 15.9% while borrowers on a 36 month loan defaulted 10.8% of the time. Using a 2-sample test of equal proportions, we find that the null hypothesis of equal proportions should be rejected at a 95% confidence interval and that 60-month loans indeed have a higher default rate than 36 month loans. To shed more light on why 60 month loans default more often, the median annual income and loan amount was compared between 60 month loans and 36 month loans. Somewhat unexpectedly, the 60-month loans had a median annual income of $70,000, which is 14.8% higher than the median of $61,000 for 36-month loans. However, comparing loan amount reveals a major finding: 60-month loans were issued for a median amount of $19,200, a staggering 92% higher than the median for 36-month loans, which was $10,000. This may suggest that borrowers of 60-month loans are struggling to keep up with their higher monthly payments. The boxplot below shows a comparison of installments between the two loan terms.



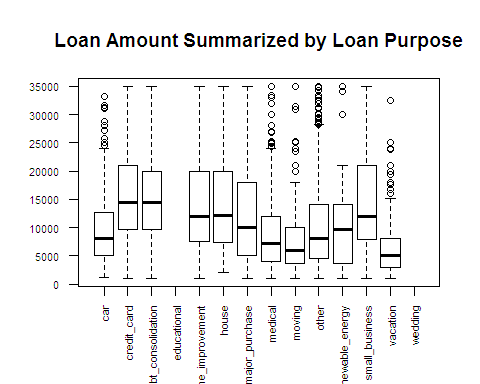
The connection between 60-month loans and higher installments is clear, but are higher installments also associated with higher default rates? The following density plot is binned by installment and filled with default rates.



Loans with higher installments appear to default more often than those with lower installments, which intuitively makes sense. A person's ability to successfully pay off debt hinges on their expenses being less than or equal to their income over the same time period. Another variable that may negatively impact a borrower's ability to pay off their loan is their total debt as a percentage of their annual income. Other things equal, a greater amount of total debt translates to a greater amount of interest paid on that debt, in addition to having to pay off the larger principal amount.



The boxplot above shows only a slight difference in the average debt-to-income ratio among 60-month loans versus 36-month loans, the former being 19.6% and the latter being 17.6%. It is worth noting, however, that the difference in the total amount of debt is greater since borrowers of 60-month loans have a higher income. Examining the variety of loan purposes reveals a significant variation in loan amounts by purpose.



Small business loans tend to be the largest amount, followed by credit card loans and debt consolidation loans. A table of default rate by loan purpose shows a re-shuffling of those ranks:

##   
## FALSE TRUE  
## car 0.89674952 0.10325048  
## credit\_card 0.90506329 0.09493671  
## debt\_consolidation 0.86595761 0.13404239  
## educational   
## home\_improvement 0.89261917 0.10738083  
## house 0.83018868 0.16981132  
## major\_purchase 0.86933798 0.13066202  
## medical 0.85873016 0.14126984  
## moving 0.79575597 0.20424403  
## other 0.86185777 0.13814223  
## renewable\_energy 0.83783784 0.16216216  
## small\_business 0.81709402 0.18290598  
## vacation 0.85674931 0.14325069  
## wedding

Moving loans have the highest chance of defaulting at 20.4%, followed by small business loans at 18.3% and house loans at 17%. This is an interesting finding as moving loans are among the lowest average loan amounts. Somewhat unexpectedly, loans for credit card debt have the lowest default rate at 9.5% despite having one of the highest average loan amounts. Comparing homeownership types by loan purpose, we find that 79% of moving loans were taken out by renters while only 40.6% of credit card loans went to renters. The question that arises is whether or not renters default at a higher rate than the overall population.

##   
## FALSE TRUE  
## MORTGAGE 0.8940371 0.1059629  
## OWN 0.8724490 0.1275510  
## RENT 0.8541139 0.1458861

The proportion table above shows that renters have a default rate of 14.6% which is higher than our total sample proportion of 12.4%. Additionally, the t-test confirms that our null hypothesis of equal proportions should be rejected. A comparison of incomes across these segments shows that renters earn an average annual income of $64,920 compared to the "Own" and "Mortgage" groups at $70,150 and $85,194, respectively. Similarly, comparing the annual income among borrowers grouped by their loan purpose:

## # A tibble: 12 x 2  
## purpose income  
## <fctr> <dbl>  
## 1 car 65319.65  
## 2 credit\_card 76164.04  
## 3 debt\_consolidation 74444.34  
## 4 home\_improvement 88311.22  
## 5 house 80387.31  
## 6 major\_purchase 78010.40  
## 7 medical 71158.25  
## 8 moving 65842.29  
## 9 other 69269.72  
## 10 renewable\_energy 65740.70  
## 11 small\_business 97547.28  
## 12 vacation 67803.73

It turns out that borrowers of moving loans had the second lowest mean annual income, which makes sense as so many of them are renters. It's likely that someone who takes out a loan in order to pay moving expenses has very little savings in addition to zero home equity (as they are usually renters), meaning the additional monthly expense of a personal loan is more likely to cause difficulty for them. From a lender's standpoint, it is common to view a borrower's employment length as a measure of financial stability. A borrower who frequently leaves jobs may not have the consistent income required to make monthly payments on a loan, while a borrower who has been employed longer is not only likely to have higher earnings, but may have an easier time budgeting as they have been living on a consistent income for longer. Below is a table of default rates and incomes by employment length:

## # A tibble: 12 x 4  
## emp\_length default income med\_income  
## <fctr> <dbl> <dbl> <dbl>  
## 1 < 1 year 0.1395571 70934.52 60000  
## 2 1 year 0.1334600 71496.46 60000  
## 3 10+ years 0.1088417 82327.80 72000  
## 4 2 years 0.1271868 73167.03 60450  
## 5 3 years 0.1283994 73955.46 61250  
## 6 4 years 0.1165837 74721.25 62400  
## 7 5 years 0.1158650 75600.91 64500  
## 8 6 years 0.1328622 75068.99 63100  
## 9 7 years 0.1255479 75719.93 65000  
## 10 8 years 0.1342105 77116.76 65000  
## 11 9 years 0.1354246 76504.77 65000  
## 12 <NA> 0.1652026 49463.53 45000

As expected, income gradually increases as employment length increases, from a mean of $70,935 for borrowers with less than 1 year employed to $82,328 for those with 10 or more years of employment. Additionally, the default rate gradually decreases as employment length increases, with less than 1 year of employment defaulting 14% of the time and ten or more years defaulting 10.9% of the time. What is even more convincing is the borrower's who listed "N/A" for their employment length: a mean income of just $49,464 and a default rate of 16.5%.

### Conclusion

Analyzing the Lending Club data set has revealed the key elements of good and bad loans. The dataset consisted of roughly 12.4% bad loans, but we were able to dive down and find certain characteristics that are associated with riskier loans. Some were pretty intuitive, such as annual income or debt-to-income ratio, while others may come as a surprise to someone with less domain knowledge, such as more defaults among longer loan terms or less defaults for loans that are used to pay off credit card debt. These variables will be used to build models that can predict the chances of a loan being successfully paid off as well as the potential payout, i.e. interest collected.