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:

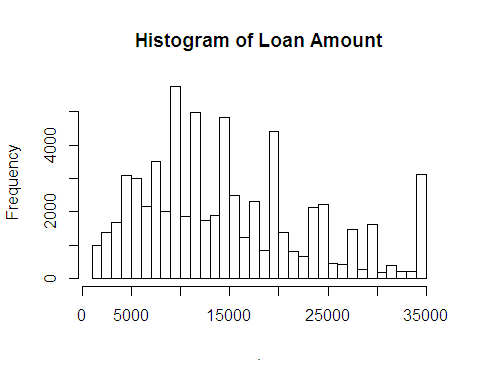
* Average current account balance
* Bankcard open-to-buy
* Bankcard utilization
* Months since most recent bankcard was opened

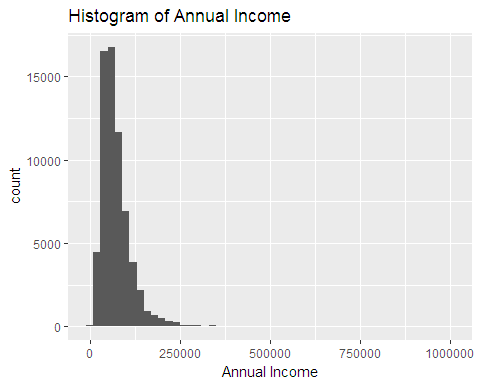
For the percentage of bankcard accounts where spending surpassed 75% of the card's limit, a missing value most likely indicated that the borrower does not have a bankcard. Missing values were therefore imputed with zero's. The same logic was applied to the number of revolving accounts. Employment length was imported as a factor consisting initially of 12 levels ranging from less than one year to more than 10 years, but one 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

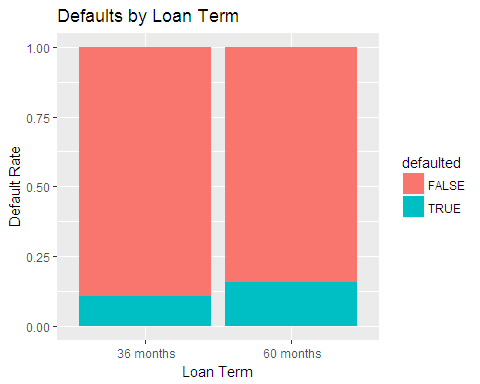
Due to computing constraints, the data was downsampled from 650,000 loans to 65,000. The analysis primarily consisted of sample mean comparisons across different segments of the dataset as well as hypothesis testing. Additionally, visualizations are provided to make the findings more impactful.

First, a look at loan amounts and annual income:



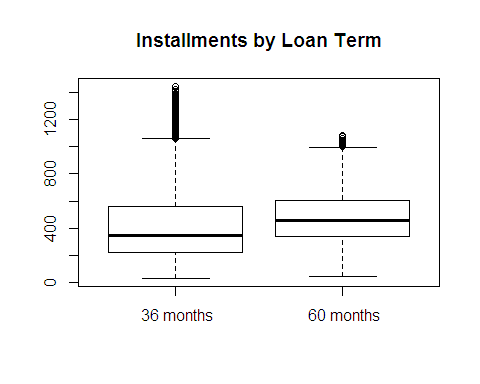


Loan amount is a slightly right tailed distribution with a large number of borrowers taking out the max loan amount of $35,000. Borrowers typically apply for loans in multiples of $5,000. Annual income is also a right-tailed distribution with a very small number of borrowers earning between $250,000 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.

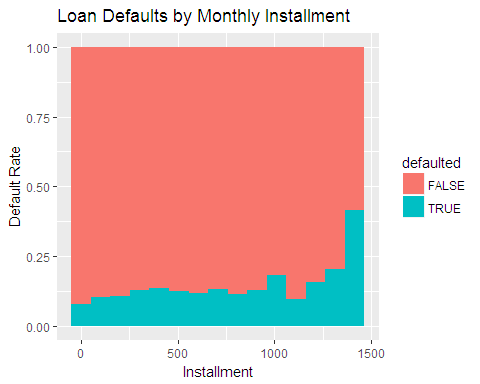


##   
## 2-sample test for equality of proportions with continuity  
## correction  
##   
## data: c(sum(loans60), sum(loans36)) out of c(length(loans60), length(loans36))  
## X-squared = 301.25, df = 1, p-value < 2.2e-16  
## alternative hypothesis: two.sided  
## 95 percent confidence interval: 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

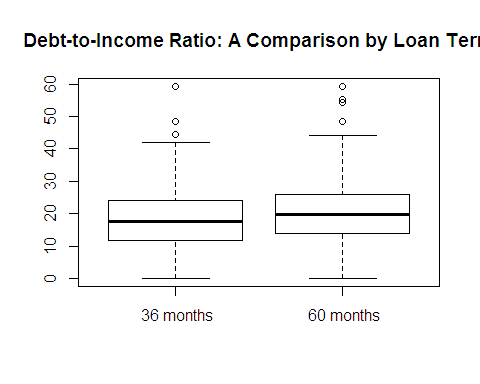
Borrowers on a 60 month loan term defaulted at a rate of 15.9% while borrowers on a 36 month loan defaulted 10.8% of the time. Using a 2-sample test of equal proportions, we find that the null hypothesis of equal proportions should be rejected at a 95% confidence interval and that 60-month loans indeed 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.



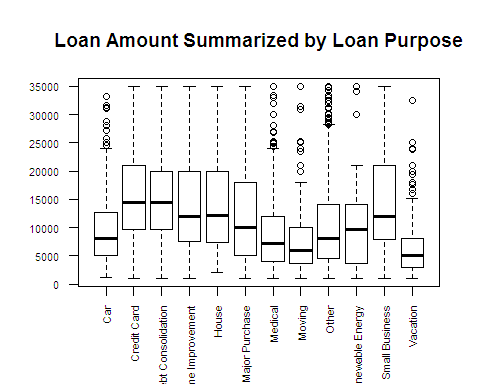
The connection between 60-month loans and higher installments is clear, but are higher installments also associated with higher default rates? The following density plot is binned by installment and filled with default rates.



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 | 0.8967495 | 0.1032505 |
| Credit Card | 0.9050633 | 0.0949367 |
| Debt Consolidation | 0.8659576 | 0.1340424 |
| Home Improvement | 0.8926192 | 0.1073808 |
| House | 0.8301887 | 0.1698113 |
| Major Purchase | 0.8693380 | 0.1306620 |
| Medical | 0.8587302 | 0.1412698 |
| Moving | 0.7957560 | 0.2042440 |
| Other | 0.8618578 | 0.1381422 |
| Renewable Energy | 0.8378378 | 0.1621622 |
| Small Business | 0.8170940 | 0.1829060 |
| Vacation | 0.8567493 | 0.1432507 |

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.

Default Rates by Homeownership Type

|  |  |  |
| --- | --- | --- |
|  | 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:

Average Income Grouped by Loan Purpose

|  |  |
| --- | --- |
| Purpose | Income |
| Car | 65320 |
| Credit Card | 76164 |
| Debt Consolidation | 74444 |
| Home Improvement | 88311 |
| House | 80387 |
| Major Purchase | 78010 |
| Medical | 71158 |
| Moving | 65842 |
| Other | 69270 |
| Renewable Energy | 65741 |
| Small Business | 97547 |
| Vacation | 67804 |

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:

|  |  |  |  |
| --- | --- | --- | --- |
| emp\_length | default | income | med\_income |
| 1. < 1year | 0.1395571 | 70934.52 | 60000 |
| 10. 9 years | 0.1354246 | 76504.77 | 65000 |
| 11. 10+ years | 0.1088417 | 82327.80 | 72000 |
| 2. 1 year | 0.1334600 | 71496.46 | 60000 |
| 3. 2 years | 0.1271868 | 73167.03 | 60450 |
| 4. 3 years | 0.1283994 | 73955.46 | 61250 |
| 5. 4 years | 0.1165837 | 74721.25 | 62400 |
| 6. 5 years | 0.1158650 | 75600.91 | 64500 |
| 7. 6 years | 0.1328622 | 75068.99 | 63100 |
| 8. 7 years | 0.1255479 | 75719.93 | 65000 |
| 9. 8 years | 0.1342105 | 77116.76 | 65000 |
| 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%. 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.

### Logistic Regression

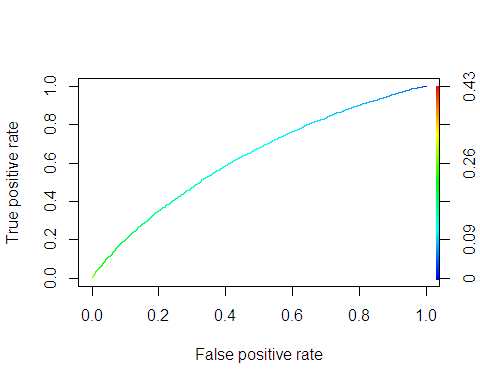
As previously stated, the objective is to use the findings from the statistical analysis to build a model that can accurately predict whether or not a loan will go into default. This model can then be used in conjunction with a basic formula to calculate each loan's total return on investment. Finally, the question of whether a low-risk loan portfolio will outperform high-risk loans will be answered. To begin the process of building a model, the following features were selected:

* Loan Amount
* Term
* Employment Length
* Homeownership Type
* Annual Income
* Loan Purpose
* Debt-to-income ratio

Prior to creating the model, the downsampled dataset of 65,000 rows was split into a training dataset (with 60% of the data) and a test dataset (with the other 40%). The training dataset was used to create the model using the features mentioned above, with the response variable being whether or not the loan went into default. Below is a summary of the model's coefficients and significance levels:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| (Intercept) | -2.4499835 | 0.2027021 | -12.0866189 | 0.0000000 |
| loan\_amnt | 0.0000116 | 0.0000025 | 4.6298759 | 0.0000037 |
| term 60 months | 0.3894153 | 0.0367981 | 10.5824836 | 0.0000000 |
| emp\_length10. 9 years | 0.0479374 | 0.0912583 | 0.5252934 | 0.5993793 |
| emp\_length11. 10+ years | -0.2084994 | 0.0606130 | -3.4398471 | 0.0005820 |
| emp\_length2. 1 year | -0.0127481 | 0.0793277 | -0.1607012 | 0.8723287 |
| emp\_length3. 2 years | -0.0769547 | 0.0738768 | -1.0416629 | 0.2975680 |
| emp\_length4. 3 years | 0.0174191 | 0.0747350 | 0.2330786 | 0.8157004 |
| emp\_length5. 4 years | -0.1884610 | 0.0851087 | -2.2143559 | 0.0268043 |
| emp\_length6. 5 years | -0.1384660 | 0.0849391 | -1.6301803 | 0.1030634 |
| emp\_length7. 6 years | -0.0316233 | 0.0897661 | -0.3522858 | 0.7246239 |
| emp\_length8. 7 years | -0.1003653 | 0.0878615 | -1.1423127 | 0.2533241 |
| emp\_length9. 8 years | 0.1058290 | 0.0831843 | 1.2722237 | 0.2032936 |
| home\_ownershipOWN | 0.1658368 | 0.0557635 | 2.9739292 | 0.0029401 |
| home\_ownershipRENT | 0.3310902 | 0.0361410 | 9.1610659 | 0.0000000 |
| annual\_inc | -0.0000048 | 0.0000005 | -8.9980813 | 0.0000000 |
| purposeCredit Card | -0.3185535 | 0.1927757 | -1.6524573 | 0.0984414 |
| purposeDebt Consolidation | 0.0997705 | 0.1901704 | 0.5246377 | 0.5998351 |
| purposeHome Improvement | 0.1588981 | 0.2015230 | 0.7884861 | 0.4304124 |
| purposeHouse | 0.3847710 | 0.3062789 | 1.2562764 | 0.2090158 |
| purposeMajor Purchase | 0.1033036 | 0.2238756 | 0.4614330 | 0.6444880 |
| purposeMedical | 0.1795631 | 0.2490644 | 0.7209508 | 0.4709398 |
| purposeMoving | 0.5125471 | 0.2667026 | 1.9217925 | 0.0546319 |
| purposeOther | 0.3223941 | 0.2012454 | 1.6019950 | 0.1091567 |
| purposeRenewable Energy | 0.3009089 | 0.6540112 | 0.4600974 | 0.6454463 |
| purposeSmall Business | 0.6133905 | 0.2382268 | 2.5748179 | 0.0100293 |
| purposeVacation | 0.2749775 | 0.2842403 | 0.9674122 | 0.3333380 |
| dti | 0.0206761 | 0.0019873 | 10.4039920 | 0.0000000 |

In general, the coefficients in this model reaffirm the earlier findings. Higher loan amounts, 60-month loan terms, renters, lower annual income and debt-to-income ratio are all associated with a higher probability of defaulting and each of their respective coefficients has a high significance level. Employment length of greater than 10 years is shown to significantly reduce the default rate. However, loan purpose does not appear to be particularly effective in predicting defaults, with only small business loans showing significance. Next, the model was used on the test data to determine each loan's probability of defaulting. To find the optimal probability threshold for predicting a default, the true positive rate and false positive rate at each possible threshold value are plotted on a Receiver Operator Characteristic (ROC) curve.



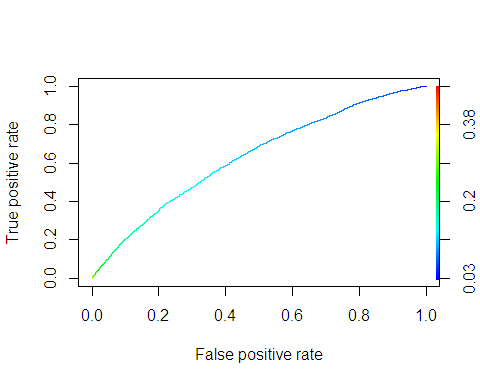
Based on a visual assessment of the ROC plot, a threshold of 0.13 was selected. In order to gain a better understanding of the model's accuracy at this threshold, a confusion matrix was created in the form of an un-weighted proportion table showing what percentage of defaults and non-defaults were accurately predicted. The true positive rate comes in at 54.2% while the true negative rate is 64.1%. Adding the weighting back to account for the relatively small proportion of defaults, the model's out-of-sample accuracy is calculated at 62.9%. This initial model can be improved upon by adding features as well as engineering a new feature. For the second model, loan amount and annual income were replaced with the ratio of loan amount to annual income:

loansTrain$loan\_income\_ratio <- loansTrain$loan\_amnt / loansTrain$annual\_inc  
loansTest$loan\_income\_ratio <- loansTest$loan\_amnt / loansTest$annual\_inc

Additionally, bankcard utilization percentage and history of delinquency were included in the new model. A summary of this new model is shown below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| (Intercept) | -3.1761371 | 0.2037212 | -15.5906045 | 0.0000000 |
| loan\_income\_ratio | 1.4816745 | 0.1593865 | 9.2961116 | 0.0000000 |
| term 60 months | 0.3051654 | 0.0356363 | 8.5633315 | 0.0000000 |
| emp\_length10. 9 years | 0.0464246 | 0.0913848 | 0.5080125 | 0.6114446 |
| emp\_length11. 10+ years | -0.2194847 | 0.0606576 | -3.6184170 | 0.0002964 |
| emp\_length2. 1 year | -0.0060930 | 0.0794258 | -0.0767130 | 0.9388518 |
| emp\_length3. 2 years | -0.0710161 | 0.0740049 | -0.9596142 | 0.3372494 |
| emp\_length4. 3 years | 0.0238771 | 0.0748554 | 0.3189766 | 0.7497443 |
| emp\_length5. 4 years | -0.1875524 | 0.0852101 | -2.2010578 | 0.0277319 |
| emp\_length6. 5 years | -0.1356270 | 0.0850281 | -1.5950846 | 0.1106933 |
| emp\_length7. 6 years | -0.0226587 | 0.0898730 | -0.2521186 | 0.8009494 |
| emp\_length8. 7 years | -0.0968602 | 0.0879595 | -1.1011917 | 0.2708132 |
| emp\_length9. 8 years | 0.1110780 | 0.0832841 | 1.3337230 | 0.1822946 |
| home\_ownershipOWN | 0.1860260 | 0.0557629 | 3.3360187 | 0.0008499 |
| home\_ownershipRENT | 0.3683790 | 0.0354864 | 10.3808537 | 0.0000000 |
| purposeCredit Card | -0.4365210 | 0.1932024 | -2.2593976 | 0.0238587 |
| purposeDebt Consolidation | -0.0071678 | 0.1905478 | -0.0376171 | 0.9699930 |
| purposeHome Improvement | 0.0887213 | 0.2017856 | 0.4396813 | 0.6601679 |
| purposeHouse | 0.3108297 | 0.3068166 | 1.0130798 | 0.3110221 |
| purposeMajor Purchase | 0.0639103 | 0.2242407 | 0.2850075 | 0.7756384 |
| purposeMedical | 0.1368785 | 0.2495305 | 0.5485440 | 0.5833184 |
| purposeMoving | 0.4634556 | 0.2670138 | 1.7356989 | 0.0826171 |
| purposeOther | 0.2779772 | 0.2017071 | 1.3781229 | 0.1681654 |
| purposeRenewable Energy | 0.2537631 | 0.6551126 | 0.3873580 | 0.6984912 |
| purposeSmall Business | 0.5167944 | 0.2381410 | 2.1701190 | 0.0299978 |
| purposeVacation | 0.2643767 | 0.2845750 | 0.9290228 | 0.3528773 |
| dti | 0.0200991 | 0.0019970 | 10.0648600 | 0.0000000 |
| bc\_util | 0.0041123 | 0.0006255 | 6.5746239 | 0.0000000 |
| ever\_delinq | 0.1329718 | 0.0324294 | 4.1003517 | 0.0000413 |

The loan-to-income ratio appears to be a very strong indicator of a loan's likelihood of defaulting. Bankcard utilization is also highly significant with an extremely low standard error but the coefficient is much closer to zero. A delinquency in the borrower's credit history also seems to positively impact the probability of a default. Below is the ROC curve for this model:



Multiple thresholds were tested with the model and ultimately a threshold of 0.13 was chosen. The below confusion matrix shows a slight decrease in the true positive rate at 52.7%, however the true negative rate has improved to 65.4%. Overall, the out-of-sample accuracy of this model has improved by one basis point to 63.9%.

|  |  |  |
| --- | --- | --- |
|  | FALSE | TRUE |
| FALSE | 0.6545413 | 0.3454587 |
| TRUE | 0.4733202 | 0.5266798 |

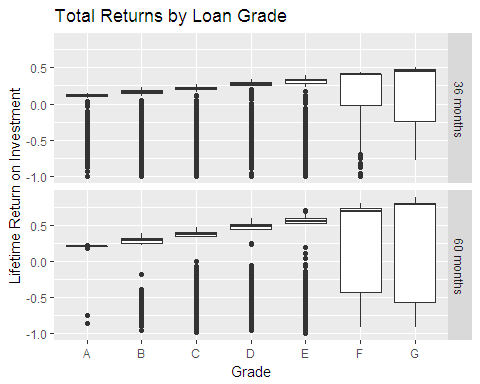
The logistic regression model can predict with considerable accuracy whether or not a loan will be successfully paid off. Another application for this model is to categorize loans into different risk levels and compare the return on investment. Remaining at the threshold probability of 0.13, the loans were separated into high-risk loans (greater than 0.13) and low-risk loans (less than or equal to 0.13). The following assumptions were made in order to calculate return on investment:

* All loans are repaid in their regular monthly installment for the entire duration of the loan term.
* Any loan that has not defaulted will be repaid in full.

Given these assumptions, the total expected payment on any loan that is not in default is the product of the loan's monthly installment and its term in months. If the loan defaulted, the total expected payment is equal to the total payments to date. The return on investment is the difference of total expected payment and loan amount, divided by loan amount. For the loans that fell into the low-risk category, the default rate improved from 12.4% to 8.4%. The return on investment for these loans is 16.4%. Loans that were categorized as high-risk had a default rate of 15.1%, while the return on investment was 22%. Of course, the high risk loans have a higher proportion of 60-month loans, which tend to have higher returns simply because they are longer. In order to account for this, each category was broken into sub-categories for 36- and 60-month loans.

## 36 Months 60 Months  
## Low-risk 0.1164248 0.2879165  
## High-risk 0.1006197 0.2802960

Although high-risk loans initially appeared to deliver better returns, accounting for loan terms gives a clearer picture. Low risk loans have a higher average return on investment due to the higher probability of loans being successfully paid off. The below plot visualizes the potential returns for each loan grade:



As the loan's grade gets worse, both the maximum and median returns increase. The downside of these loans is the risk of incurring losses becomes considerably higher, so much so that the *average* return for these loans actually becomes lower. The logistic regression model can be used in conjunction with loan grade to dive deeper and find the lowest risk loans in each grade and evaluate their average returns.

Lifetime Returns: Below Threshold

|  |  |  |
| --- | --- | --- |
| Grade | 36 Months | 60 Months |
| A | 0.0970 | 0.1887 |
| B | 0.1170 | 0.2468 |
| C | 0.1342 | 0.2874 |
| D | 0.1322 | 0.3330 |
| E | 0.1413 | 0.3273 |
| F | 0.2325 | 0.3312 |
| G | 0.1316 | 0.3637 |

Lifetime Returns: Above Threshold

|  |  |  |
| --- | --- | --- |
| Grade | 36 Months | 60 Months |
| A | 0.0818 | 0.2146 |
| B | 0.1050 | 0.2306 |
| C | 0.1004 | 0.2772 |
| D | 0.1202 | 0.3040 |
| E | 0.0765 | 0.2816 |
| F | 0.0098 | 0.2897 |
| G | 0.0209 | 0.2099 |

With the help of the logistic regression model, we were able to find loans that were rated poorly but still fell below the threshold probability of default. On average, these loans offer the highest lifetime return on investment as seen in the table above.

### Conclusion

When evaluating investment opportunities, the basic factors to be taken into consideration are the the investment's risk and rewards. Lending Club's investment platform provides this information and gives the investor the power to build a portfolio based on their preferences. A highly risk-averse investor can choose to invest in loans with a grade of A while an investor seeking high returns and willing to accept the risk may choose loans with lower grades and higher interest rates. During statistical analysis, the factors contributing to a loan's successful repayment or default were identified: loan terms, total debt as a percentage of income (before and after the loan was approved), employment and history of delinquencies were the most significant. Using predictive modeling, an investor can find loans that may be "underrated" -- in other words, poorly rated due to a bad credit rating, etc. but still considered low-risk by the logistic regression model. These loans offer the highest returns on average, meaning that a well-diversified portfolio of low-grade loans with a relatively low probability of defaulting will eventually result in the most profit.