

**How to Get a Loan: An Exploration with Logistic Regression**

*by Tianhong Ding & William Raikes (Group 30)*

*Advanced Data Analysis – 12/12/16*

**Introduction / Background**

The Lending Club is a peer-to-peer lending company, founded in 2006 in San Francisco, and is currently the world's largest online credit marketplace. The company specializes in personal loans, business loans, and other types of financing needs, and to date, has funded over $22 billion dollars in loans. Former US Treasury Secretary, Larry Summers, has stated that the “Lending Club's platform has the potential to profoundly transform traditional banking over the next decade.”

The Lending Club has a fairly simple structure. Potential customers complete brief applications online, in which the Lending Club quickly and digitally assesses the risk, credit rating and interest rates from these applications, and then sends qualified applicants an offer. Individual and institutional investors are then able to invest in these loans. Therefore, the Lending Club acts similarly to a bank, and cuts out the ‘middleman’. The New York Times has reported that “the math pitch involves the increased efficiency of cutting out the banks, allowing relatively lower rates for borrowers and good returns for lenders.”

Overall, the Lending Club has made loans more accessible to common individuals, both in terms of the ease of applying and receiving loans, and through its affordable structure. In a sense, it is making the loan industry more affordable and accessible to the masses. The Lending Club is also fairly transparent, as it has made public all of its relevant data in loan applications and loan servicing since the time of its inception. This report will make use of this data to help prospective applicants increase their chances of being offered a loan, and will help the Lending Club serve some potentially under-served markets.

**Objectives**

The primary objective of this analysis is: how to get a loan from the Lending Club. This report will highlight the type of information that the Lending Club utilizes in granting loans, and will enable potential customer to understand what makes their applications more attractive and how this impacts their chances of receiving a loan.

This information will also be contrasted against whether an individual will default on their loan, and can serve as guidance to the Lending Club in understanding how certain information may or may not relate to the potential default status of an applicant – a crucial concern in reviewing loan applications.

**Material & Methods**

The data is publicly available at: https://www.lendingclub.com/info/download-data.action. After downloading, merging and processing the data, the final data frame contains over 9 million records –containing information from an individual loan application – which contain several descriptive variables for analysis. A logistic model was created by using these variables to understand the relationship between the information provided on an application and the outcome of receiving an offer for a loan. Also, the model will be compared to other high performing predictive models, to assess whether this inferential model can adequately perform at a reasonable level as a predictive model.

A second logistic model has also been created by using the same predictors as the first model, and will be used to ascertain whether these same predictors are useful in understanding if an individual will default on their loan. The only other main distinction from the first model will be the replacement of the outcome variable from rejection status to default status. This model will provide evidence that the information used to assess the riskiness of an applicant can be transferred over to whether the applicant will actually default on their loan.

The relevant data is as follows:

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Rejection Status | Binary variable recording whether an application was accepted. |
| Loan Amount | The loan amount requested by the applicant. |
| Debt-to-Income Ratio (DTI) | The monthly debt divided by the monthly income of the applicant. |
| Employment Length | A factor describing in years how long an individual has been employed. |
| Date | Date of loan application. |
| State | State of residence. |
| Default Status | Binary variable recording whether an individual defaulted on their loan. |

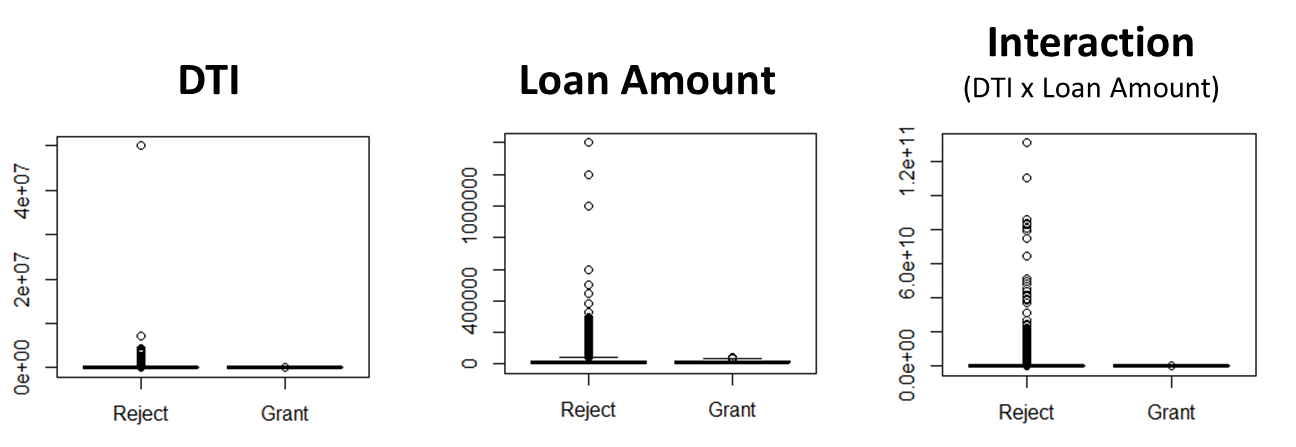
Each model will be assessed by using multiple diagnostic tools, such as: predictive accuracy, goodness of fit metrics, influential observations, multi-collinearity and etc. If the model fit is adequate, the model's estimated parameters can be used to guide individuals in increasing their likelihood to receive a loan from the Lending Club, and will help the Lending Club understand if these variables can be used to predict whether someone will default on a loan.

**Results**

Exploratory Data Analysis

Boxplots for the three numeric variables are shown below. “Loan Amount” and “DTI” has been described in the data descriptions, but a new term, “Interaction”, was created variable by multiplying the variable “Loan Amount” by “DTI”. This variable was created to assess the varying relationship between “Loan Amount” and “DTI” and their combined impact on if an applicant receives a loan.

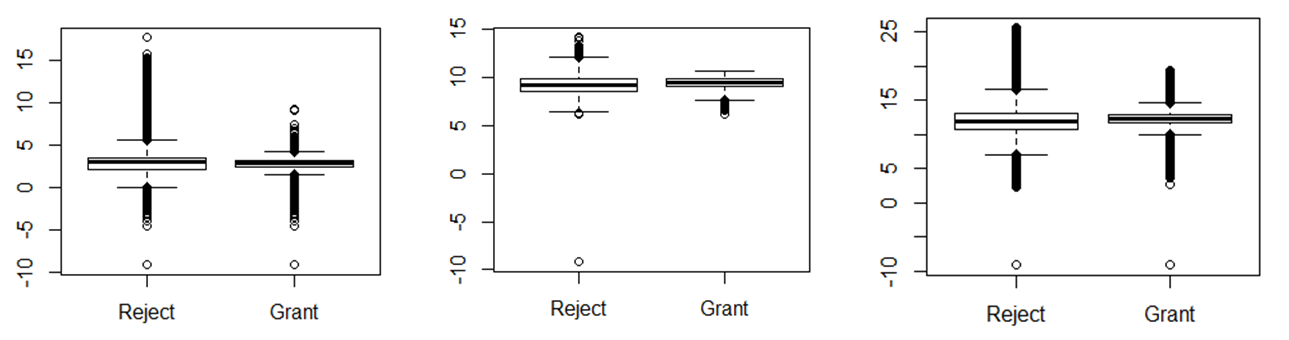
*Figure 1: Boxplots*



The boxplots are hard to analyze due to the skewness of the data, therefore, it’s difficult to determine the distributional properties of the data, and whether being rejected or offered a loan changes the distribution of the numeric predictors. It's clear that the outliers are distorting the figures, which brings the first piece of analysis into recognition: the distribution among those who were rejected from loans is more heavily skewed than those who were accepted. Yet, it should be noted that less than 1% of the data is actually within these elongated tails.

To get a better sense of the distributional properties of numeric data, the boxplots are shown again, after transforming the numeric variables into their logarithmic counterparts. This decreases the skewness of the variables and makes it easy to compare the different distributions.

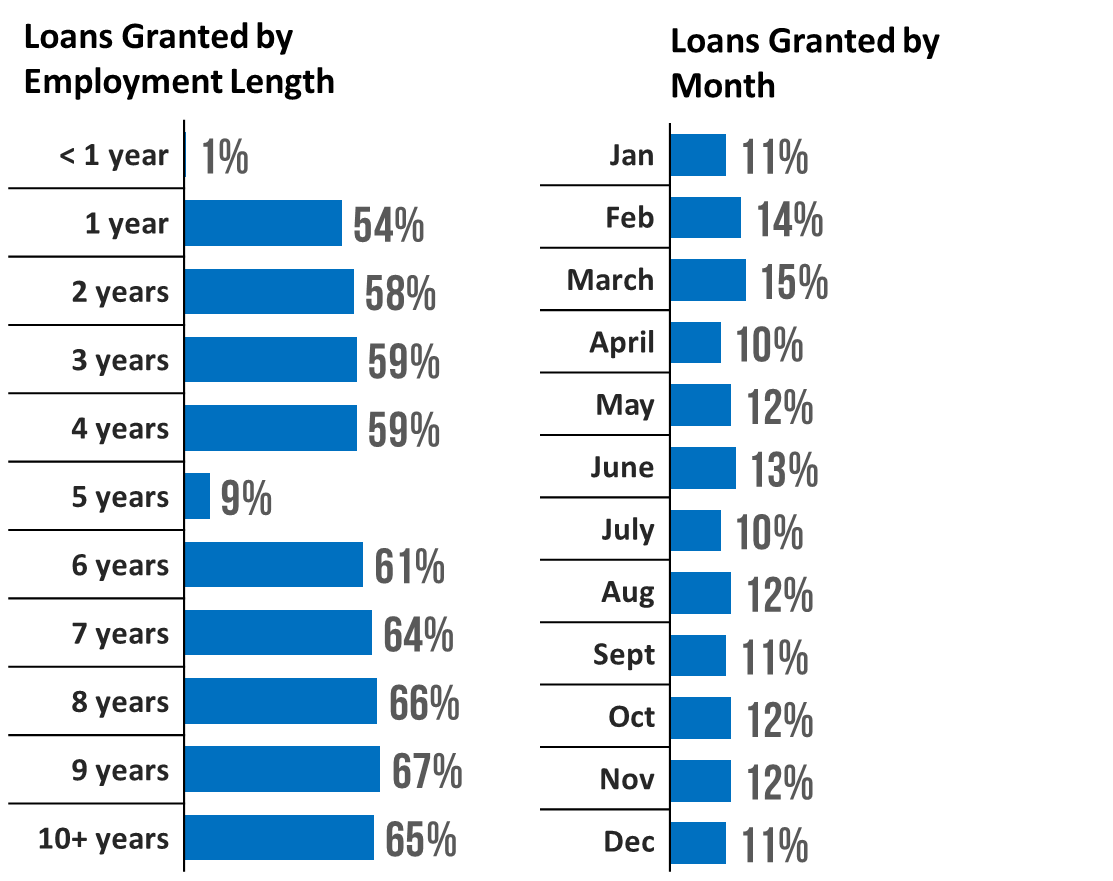
*Figure 2: Boxplots on Logarithm Scale*



The boxed quartiles are very small, reflecting the fact that this is a large dataset with more than 9 million records. Yet it is also apparent that the boxed quartiles are, somewhat, similar in size and location. This indicates that the structure of the data do not differ drastically when comparing the two populations of rejected applicants versus accepted applicants.

Bar charts are also provided below to explore whether the categorical variables are related to the ability to get a loan. The charts below show the proportion of those who received a loan by “Employment Length” and by month.

*Figure 3: Categorical Barcharts*

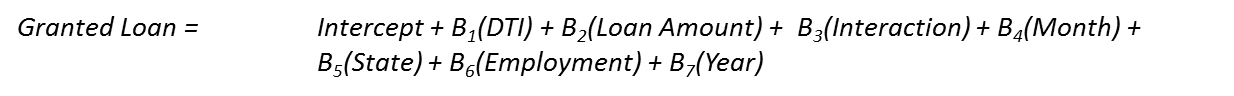


The chart detailing “Employment Length” is evidence that this variable is instrumental in being offered a loan. Less than 1% of applicants who have less than 1 years’ worth of employment were offered a loan. This number drastically increases to 54% if the applicant states that they have been employed for at least 1 full year, and will continue to increase until reaching the highest designation of 10+ years of employment, with a rate of offers at 65%. The only striking anomaly is the fact that only 9% of applicants received an offer if they stated they have been working for a full 5 years of employment.

The month variable, however, does not appear to impact the chances of someone obtaining a loan; meaning, seasonality is not a contributing factor for the model.

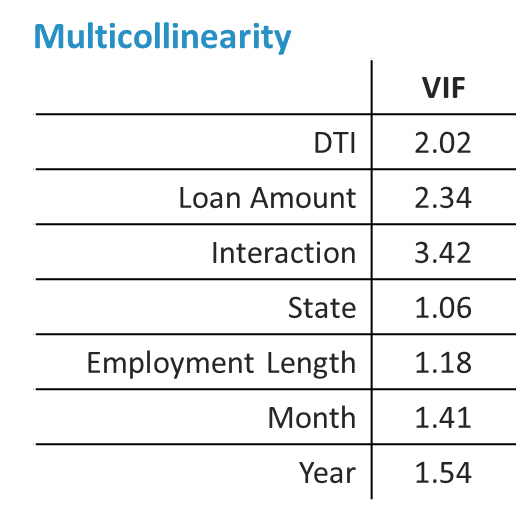
Rejection Status: Model Fitting

A logistic (logit) mode was fit to the data as follows:



Several diagnostics were used to determine if the model fit was appropriate. The Variance Inflaction Factors (VIF), as shown in the chart below, indicate that there is a moderate amount of collinearity between “DTI”, “Loan Amount” and the interaction term; the VIF values are 2.02, 2.34, and 3.42, respectively. This is to be expected, and the collinearity is not high enough to suggest that a different model is more appropriate.

*Figure 4: Multicollinearity*



The Likelihood Ratio test confirms that the inclusion of the interaction term is suitable for this model, as evident by the p-value of < 0.001. The “Goodness of Fit” test also confirms the fit of the model, with McFadden's (pseudo) R2 value of 0.503 - a fairly high score.

Influential observations were calculated using Cook's Distance. Zero data points were considered influential when using a cutoff value of 1. However, about 7% of the data points were considered influential when considering a cutoff value of 4/n. Even though removing these influential points led to a 6% drop in the AIC score, the coefficients and statistical significance of the predictors were left relatively unchanged. Therefore, it seems prudent to leave the influential observations in the analysis.

Rejection Status: Interpretation

Below is a chart detailing the regression coefficients of the rejection status model for the three numeric variables: “DTI”, “Loan Amount” and the Interaction term.

*Figure 5: Coefficients*



Debt-to-Income (DTI) is a statistically significant predictor for rejection status, and for each one unit increase in DTI, the odds-ratio for being granted a loan decreases by 0.4%. This may seem trivial at first, but the values for DTI range from 0 to 10,000 – so a substantial change in this variable will see a large change in probability of receiving an offer.

The “Loan Amount” variable, however, does not influence the likelihood of receiving an offer, since the variable has a low coefficient (< 0.001) with a p-value of 0.704. Therefore, an applicant should not worry that the size of their requested loan would be used as a catalyst for an offer. The interaction term, however, is statistically significant, with a p-value of < 0.001. But the practical significance is only relevant when there are large shifts in in either “DTI”, “Loan Amount” or both, as the coefficient registers at less than 0.0001.

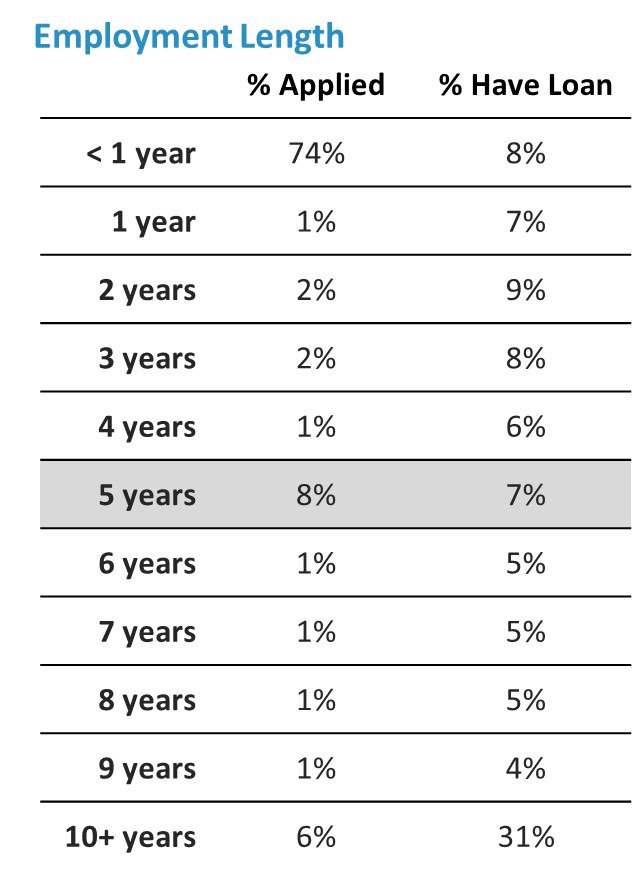
The coefficients for the factored categorical variables are also shown in the charts below. Similar to what was observed in the exploratory data analysis, the “Employment Length” variable is crucial in obtaining a loan. As less than 1 years’ worth of employment is set to the baseline, reaching at least 1 full years’ worth of employment increases the odds-ratio of obtaining a loan by a factor of at least 118, and ranges to a peak of 170 (with the exception of 5 years’ worth of employment – which is discussed below). This variable is crucial in determining if the applicant will receive a loan or not.

*Figure 6: Coefficients*



This analysis shows an interesting interaction between offers given and 5 years’ worth of employment. This anomaly is looks suspect at first. However, a deeper investigation into the data offers some clues as to the potential reasoning. The chart below shows the population of individuals grouped by their lengths of employment. Almost three-fourths of all applicants reported less than 1 full years’ worth of employment. This proportion ranges from 1% to 2% for all other levels of employment, except for two: 5 years and 10+ years.

*Figure 7: Frequencies*



When looking at the proportion of individuals who received a loan, the reasoning behind the weak relationship between 5 years’ worth of employment and receiving an offer is clear: a disproportionate amount of individuals who have worked for 5 years are receiving loans at a rate less their counterparts (with the exception of those who have worked for 10+ years).

For instance, 1% of all applicants have 1 full years’ worth of employment, and this group makes up 7% of all loan recipients. Likewise, 1% of all applicants have 6 years of full employment, and make up 5% of all loan recipients. Those who have reported employment lengths of 10+ years make up 6% of applicants, and 31% of all loan recipients. Those who were employed for 5 full years make up 8% of all applicants, yet only 7% of loan recipients.

Rejection Status: Predictive Performance

The model was fit primarily for inferential analysis. Yet, the intention of this analysis was to inform how to increase the probability of receiving a loan from the Lending Club. So a question naturally arises from this logic: how well does the model perform for prediction?

After building the model and validating its goodness of fit, ROC-AUC metrics were calculated using a cross-validated approach. The model was compared to other, more common predictive models from the machine-learning and classical statistics fields. An Extreme Gradient Boosting (XGBoost) model, a Random Forest model, a Linear Discriminant Analysis model and a Generalized Additive Model were fit using the same variables as the logistic regression model, and compared through their AUC scores. The comparison is shown in the graph below.

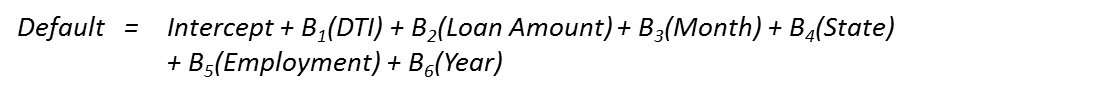
*Figure 8: Predictive Performance*



It's evident that the XGBoost model outperforms the rest, with an AUC score of 96%. The Random Forest model comes in a close second with 95%, and our logistic regression model results in an AUC score of 92%. The LDA and Generalized Additive Model follow slightly behind with scores of 92% and 89%, respectively. Although our model did not outperform the more modern, and powerful, machine-learning algorithms such as XGBoost and Random Forests, it still performs quite well, plus has the added benefit of offering inferential statistical analysis to show the intricate relationship between the outcome variables and the predictors (something that cannot be easily achieved for the other modes).

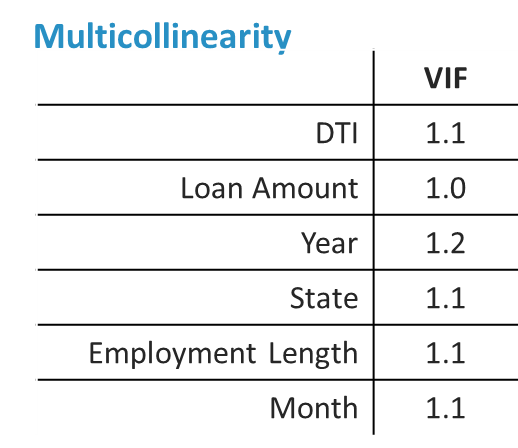
Default Status: Modeling Fitting

To determine the relationship among the predictor variables and the outcome of defaulting, the following logistic regression model was fit:



The same diagnostics used earlier were also employed on the “Default” status model, and are shown below. The Variance Inflation Factors revealed that including the interaction term based on the “Loan Amount” and “DTI” variables resulted in high multi-collinearity, and an unstable model. Therefore, this variable was omitted. The new VIF values, which are shown below, suggest that there is independence among the predictors.

*Figure 9: Multicollinearity*



The Goodness of Fit metrics, on the other hand, suggest another key point of the analysis. The same variables that were used to adequately predict whether someone would be rejected from a loan are not viable predictors to determine whether an individual will default on that loan. The McFadden's (pseudo) R2 value is 0.033, which is almost 0.50 points below the McFadden's R2 value from the original model. In addition, the cross-validated ROC-AUC value of this model is 60.4%, which is over 30 percentage points lower than the original model. Although these models cannot be compared exactly, since they do include significant variations (i.e. especially different outcome variables and a missing interaction term), the result is still valid: the same variables that adequately predict rejection status, cannot adequately predict default status. This is an interesting contrast, since it is assumed that a major reason why someone will be rejected from a loan is simply because the lender believes the person is too risky and may default on their loan.

There is a fair amount of influential data points within this model fit, however. Using a cutoff point of 4/n, about 6% of the data were classified as influential after calculating their Cook's distance. Removing these records reduces the AIC value by 28%, and increases McFadden's R2 value to 0.133. This is a substantial improvement in fit, yet not enough to claim that this model fits the data well; and furthermore, the coefficients remained relatively the same. Therefore, removing the influential data points does not seem beneficial, and the conclusion remains the same: the predictive variables for rejection status are not predictive for default status.

Default Status: Interpretation

The coefficients for the default model are shown in the charts below. Although this model does not fit well, it’s interesting to look at the coefficients to understand the (albeit tenuous) relationship between default status and the predictors. There are some interesting trends that validate some of Lending Club's loan accepting practices, yet call into question some of the variables used to determine who should or should not be offered a loan.

*Figure 10: Coefficients*



The Debt-to-Income ratio seems on par with the first model constructed. An increase in DTI will weigh on an individual's ability to pay back their loan, at a level similar to the rate at which they're rejected for a loan. The “Loan Amount” also plays only a small role in predicting if someone will default on their loan, and as already shown, this variable has no weight in whether someone will be offered a loan in the first place.

The slightly incongruent results are for employment length. The variable employment was a keystone factor in deciding if an applicant were to receive an offer for a loan. This variable also predicts whether someone will default on a loan, however, the relationship is much weaker. This shows that there may be opportunity for the Lending Club to provide more loans to more individuals, especially in the sub-populations of those with employment length levels at less than 1 year or 5 years.

**Limitations**

One variable is missing in the main analysis, although it is collected during the time of application – FICO score. This variable may be a confounder, and other variables in the model could be a proxy – such as “DTI”. But since there is a high level of fit, and prediction accuracy, the absence of the FICO metric does not seem to impede on the analysis provided.

**Conclusion**

In conclusion, this analysis shows there is one main criteria in trying to get a loan from the lending club, which is an applicant should be employed for at least 1 full year. By doing so, the odds-ratio of the applicant will increase by a factor of 118 at the minimum. And each sequential year will increase the odds-ratio as well, with the maximum factor being 170. There is one condition however, an applicant should not report that they have been employed for exactly 5 years. The increase in the odds-ratio of the applicant is only a factor 8 – substantially lower than the other years given. This is due to the fact that an inordinate amount of individuals report having 5 years of employment, and a disproportionate amount of these individuals receive an offer – compared to the other groupings.

Furthermore, there are some restrictions in qualifying loans. An applicant should not request a loan greater than $40,000, should not have a Debt-to-Income ratio greater than 10,000, and the interaction term between the two should be restricted to 250 million. Lastly, and least significantly, if an applicant increases their income or decreases their monthly debt, and in effect improving their DTI ratio, then that applicant will have a better chance to be offered a loan – but not to the same effect as being employed for a full year. However, the loan amount requested by an applicant does not impact whether someone is offered a loan or not, and the time of the year also does not play into account.

Likewise, there are potential opportunities for the Lending Club to improve upon their lending practices. First, there may be untapped potential in the large cohort of individuals who have less than a full year's worth of employment. This group does default in loans more often than their longer employed counterparts, but yet not at the same rate by which they are rejected for loans – the same sentiment can be applied to those who have been employed for 5 years as well. Therefore, the Lending Club may want to rethink their loan rejection criteria, since these qualifiers are shown to not be predictive of actual default rates.