The Effect of Application Criteria on Loan Approval

Uladzimir Kasacheuski

Abstract

Lending Club is a peer-to-peer lending company which processes thousands of loan applications each day. The company provides datasets on approved and rejected loans online for free. After properly engineering the given data for analysis, this data can be analyzed in order to find correlations between application criteria, such as Total Loan Sum requested and Employment Length, with loan approval probability. This research assesses the correlations between loan total, DTI, employment length, submission year, submission month, and applicant state of residency upon approval chances. The research found 3 strong correlations and even a surprising correlation including residency state.

Research Question

What properties affect loan approval probability the most? In other words, what factors have the greatest correlation to loan approval rate?

Background and Significance

Lending Club is a San Francisco based peer-to-peer lending company. Every day they receive thousands of applications for loans, pass the applications to their underwriters, and either accept or reject the loans based on information provided. For the attraction of new investors, Lending Club displays analyzed statistics about the loans that they accept. Particularly, they display statistics about how much profit an investor could expect to earn based on different classifications of the data. Lending Club also shares the datasets used to generate these results, the data about approved loan applications, as well as data about rejected loan applications.

With the data that Lending Club shares about its approved and rejected loan applications, it is possible to determine the probability of loan applications being approved or rejected based on certain criteria found in the applications. By analyzing the data of approved and rejected loan applications we will be able to determine if any of the data given by Lending Club correlate to a change in an application's probability of being approved or rejected. Upon finding correlations between application criteria and approval probability, it would then be possible to predict the chances of a loan being approved or denied before going through the underwriting process. It would also, then, be possible to determine what properties one should optimize as well as how they should be optimized in order to maximize a loans chances of being approved.

Methods - Obtaining and Analyzing the Data

Obtaining the Data

The data used in this research was obtained through the company Lending Club on their website at the link *https://www.lendingclub.com/info/download-data.action*. At this “online repository” Lending Club offers for download CSV files of approved loan data and rejected loan data, as well as a “data dictionary” which defines what the columns of the approved loan data values represent. The datasets 2007-2011 and 2007-2012 for Approved and Rejected loans, respectively, were chosen and extracted.

The data extracted from the Lending Club repository was heavily engineered for optimal use in analyzing the correlation between loan application criteria, such as the total loan amount requested, and loan approval. The data engineering required : assessing equivalent criteria, or columns, given between the two datasets and extracting this data, matching the data types data is represented in under equivalent columns between the two datasets, limiting the datasets to an equivalent time frame, extracting separate month and year data columns from date data described uniquely between accepted and rejected loans, and finally concatenating the approved and rejected loan data with an approved classifier. Details of how the data was engineered can be found in the appendix.

Analysing the Data

The final compiled and optimized dataset was then analysed, with a focus on determining what data criteria, or columns, affect or correlate with approval the most. The data analysis

consisted of, for each data criteria, analysis of the distribution of the data and an analysis of the acceptance ratio for a categorical distribution of the data. Each data criteria was first plotted in either a box plot, for continuous data, or a bar graph, for categorical data. Then, for the analysis of the acceptance ratio, all data was, by category, tabularized to calculate the frequency of occurrence for each category. The occurrence of accepted loans per category was divided by the occurrence of all loans per category to generate a ratio, the approval ratio, for each category. This ratio is the general probability function, describing the probability that an application will be approved given a certain value. An example of the significance of analyzing the ratio as opposed to absolute magnitude can be seen when comparing Figure 8 and Figure 9, both found in the appendix.

The following data criteria were analysed in this research: Loan Total, the total loan amount requested in the application; DTI, the debt to income ratio of the applicant; Employment Length, how long the applicant had been employed in years including '<1 year' and '10+ years' as initial and terminal categories; Year, the year at which the application was submitted; Month, the month at which the application was submitted; and, State, the state where the applicant resides.

Results of Analysis

The analysis of the data distributions and the approval ratio distributions of the data provided by Lending Club relevant to approval and rejection yields some intuitive and not intuitive results. It, additionally, yields qualitative results correlating the effect of certain data criteria values and the probability of a loan being approved.

The boxplots of the continuous data criteria `Loan Total` (Appendix, Figure 1) and `DTI` (Appx, Figure 4) show us that the variance of rejected loans is much greater than for approved loans. It is also easy to see that the approved loans distributions are very symmetric (Appx, Figure 5). To gain a better understanding of the total distribution of these data, we plot a histogram including the distribution of all applications and approved applications for each data criteria (Appx, Figure 2, Figure 6). Clearly that the number of approved applications is much less than the number of total applications, however it does not tell us anything significant about relating the two criteria to approval chances. Interestingly, we do see a large peak at `DTI` of [0,2)% in applications, without a similar peak in approvals. It is possible that this peak is caused by credit seekers who have no history of credit, and thus no current debt.

The correlation between `Loan Total` and `DTI` with the approval of the loan is much more clearly described when analysing the approval ratio bar graph of the categorized data. Figure 3 demonstrates the approval ratio for Loan Totals categories with an interval of $4,000. We can clearly see a smooth gaussian curve, skewed to the left with a mean between 8 and 12 thousand with a 13% acceptance ratio - yet with a sudden overwhelming peak around the 20k amount with a 18.7% acceptance ratio. Figure 7 demonstrates the approval ratio for DTI categories with an interval of 5%. The result is a clear gaussian distribution, demonstrating that the largest acceptance rates (~17.5%) occur between [10, 25)% DTI. [0-5)% and [25-30)% DTI have slim, but real chances. [30,inf) has zero chance of being accepted. We can see that there is a clear relationship, and correlation, between Loan Total and approval as well as DTI and approval.

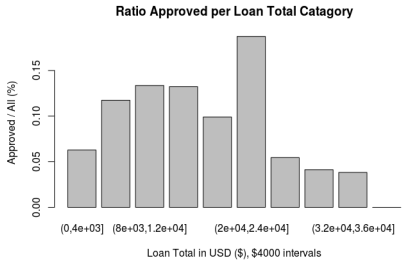

Figure 3 - Ratio Approved per Loan Total Category

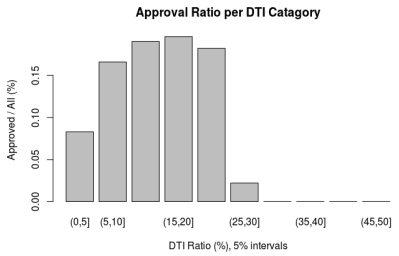

Figure 7 - Approval Ratio per DTI Category

The approval ratio graphs for the categorical data criteria `Employment Length`, `Year`, `Month`, and `State` show us clear relationships between the data criteria and their effect on approval. While, as we could have expected, Year (Appx, Figure 10) and Month (Appx, Figure 11) did not have any correlation with approval chances, Employment Length and State did. From Figure 9, describing the ratio approved per Employment Length, we instantly see an informative relationship: applicants with less than one year of employment have slim chances of getting approved, and that this chance rises until about 4 years of employment, after which point the acceptance ratio levels off at around 25%. Figure 12, describing the ratio approved per State, shows that there is a definitely discrepancies between the acceptance ratios of different states. For example, the district of columbia (DC) has an approval ratio of 18.93%, while IN only has an acceptance ratio of 3.69%.

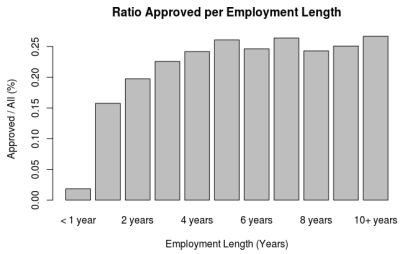

Figure 9 - Ratio Approved per Employment Length

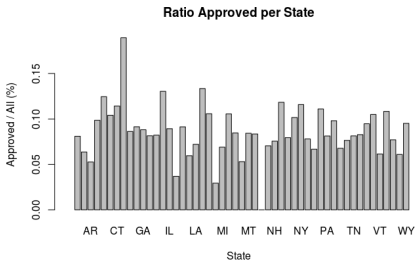

Figure 12 - Ratio Approved per State

Discussion

This research has shown that there is a definite correlation between Loan Total, DTI, Employment Length, and even State with the chance of a loan application being approved. While the first 3 are logical and likely a direct criteria in the loan approval process, the correlation between State and approval is more likely a secondary result of other factors. The current research is limited by the data that is provided by rejected applications. It would be interesting to determine how credit scores, credit history, cosigners, annual income, and other factors affect the approval ratio.

Future work should consist of analysing multiple criteria at the same time and their effect on approval of a loan. For example, we see from figure 3 that the ratio of approved loans decreases as you get higher. It would be interesting to determine whether the chance of getting a loan approved for a higher loan total is affected by a higher interest rate, lower credit score, or higher annual income.

Appendix

Details of Data Engineering

Between the approved and rejected datasets, not all of the data present in one was present in the other. In other words, the rejected loan application dataset was a strict subset of the approved loan dataset. Furthermore, the column names and data types stored under logically equivalent columns varied; for example, debt to income ratio was reported in both datasets: in the approved dataset it was reported under column `dti` and was represented as `XX.XX`, where x is any real positive number. In the rejected dataset, debt to income was reported under column `Debt.To.Income.Ratio` and was represented as `%XX.XX`. This pattern continued through equivalent data columns. Thus, equivalent data columns had to be found, such as `Amount.Requested` = `loan_amnt`, `Employment.Length` = `emp_length`, etc…, and renamed. Additionally, the data had to be cleaned in order to create matching datatypes under the new, matching, data columns. Other criteria, or columns, of data found in the approved loans but not found in the rejected loans was not analyzed in this research as there is no method by which we can determine whether the distribution of those criteria is affected by the distribution of applications, naturally, or whether the distribution is affected instead by approval criteria.

Data was further engineered by removing all data for loan applications after the year 2011, extracting the month and year from the date data, and finally concatenating the data with classification. Data beyond the year 2011 was removed as there is no accepted loan data beyond year 2011, while there is for the year 2012. The year and month were extracted from the date data in order to optimize analysis of time period and loan approval. Finally, after concatenation of the two, up to this point, separately manipulated datasets we have a final compiled, relevant dataset with optimized data structure that we can begin analysing. This dataset is further cleaned in order to remove certain outliers from rejected application values including DTI ratios greater than 2000 and loan total amounts greater than 40000, the companies stated maximum loan funding amount. Our dataset is reduced by 4.35%. This dataset has been saved in a new CSV file.

Figures

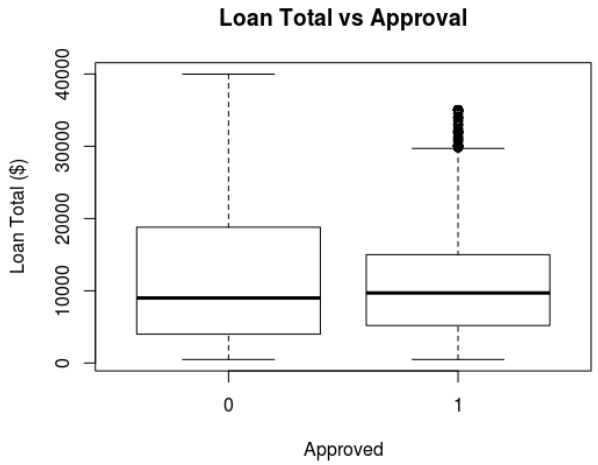

Figure 1 - Loan Total vs Approval

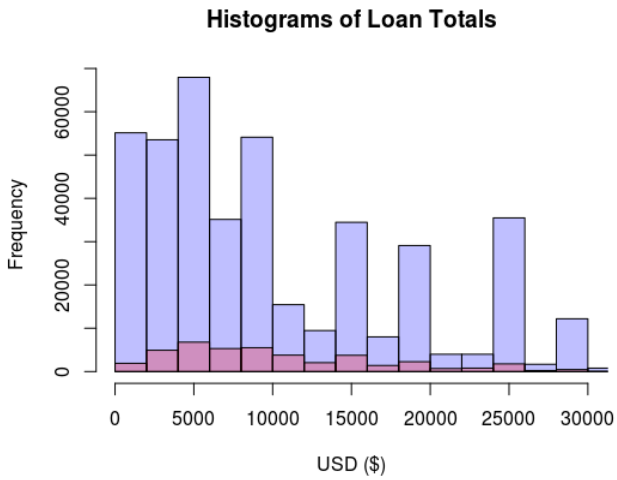

Figure 2 - Histograms of Loan Totals - Red is accepted applications, Blue is all applications

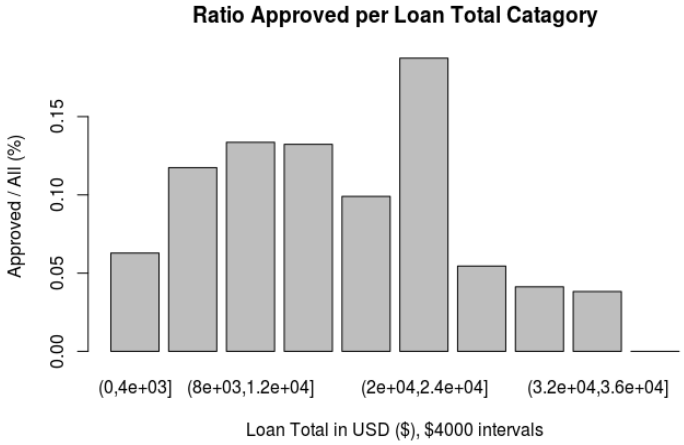

Figure 3 - Ratio Approved per Loan Total Category

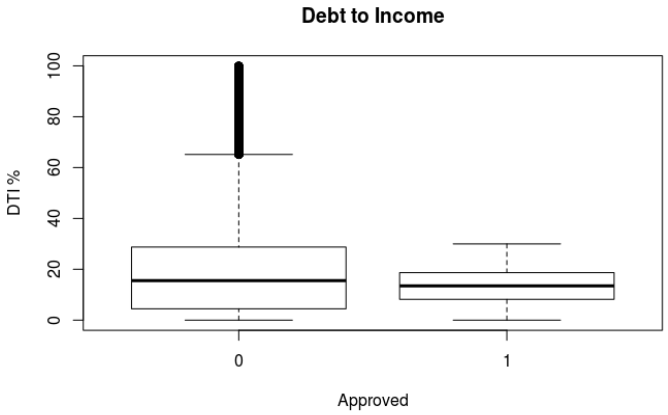

Figure 4 - Debt To Income

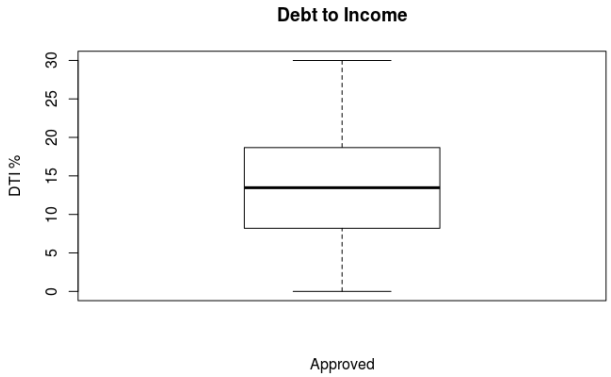

Figure 5 - (Approved) Debt to Income

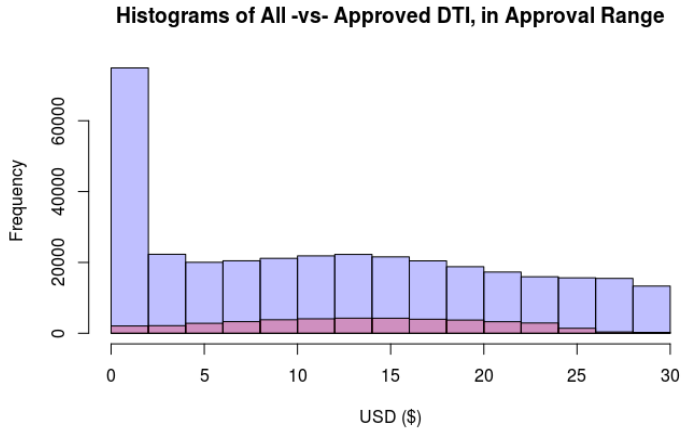

Figure 6 - Histograms of All -vs- Approved DTI, in Approval Range

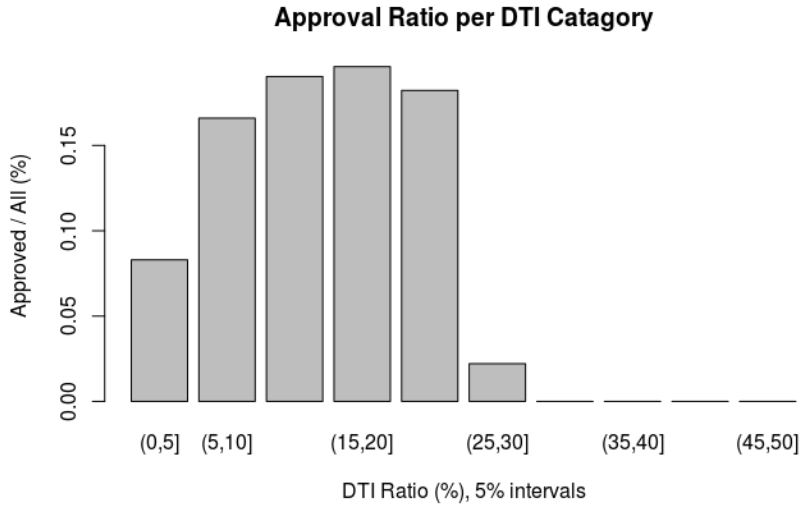

Figure 7 - Approval Ratio per DTI Category

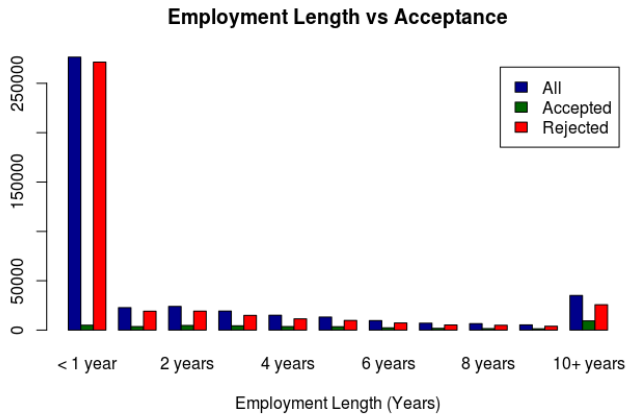

Figure 8 - Employment Length vs Acceptance

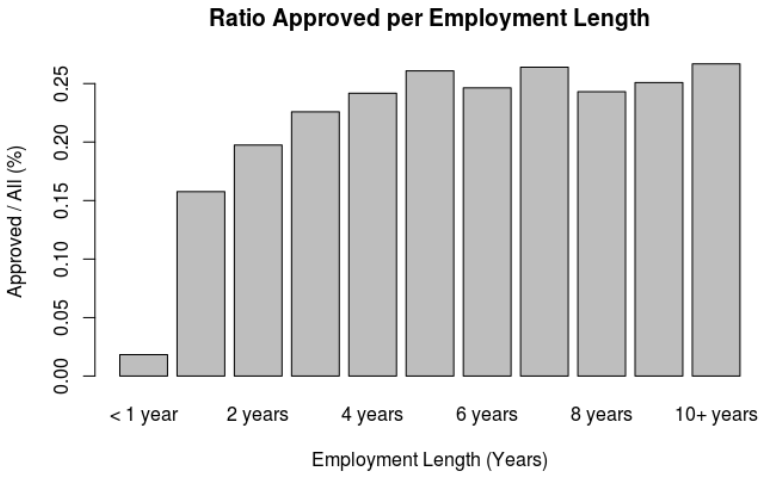

Figure 9 - Ratio Approved per Employment Length

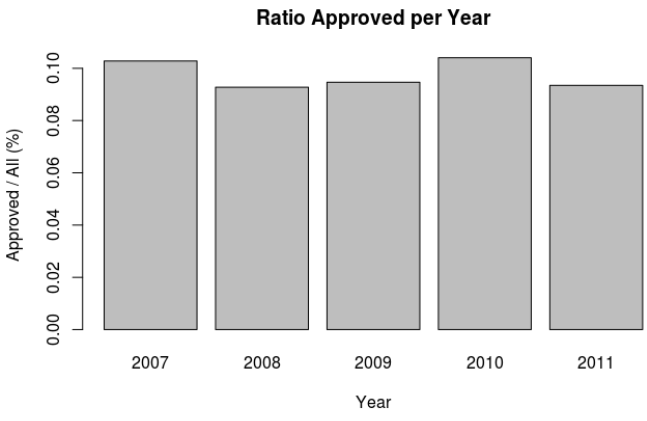

Figure 10 - Ratio Approved per Year

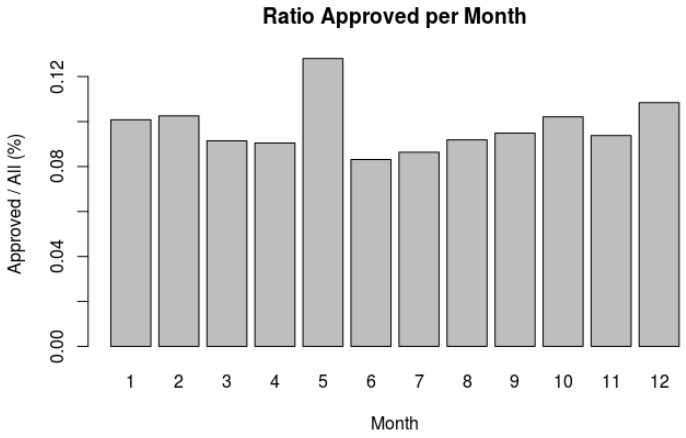

Figure 11 - Ratio Approved per Month

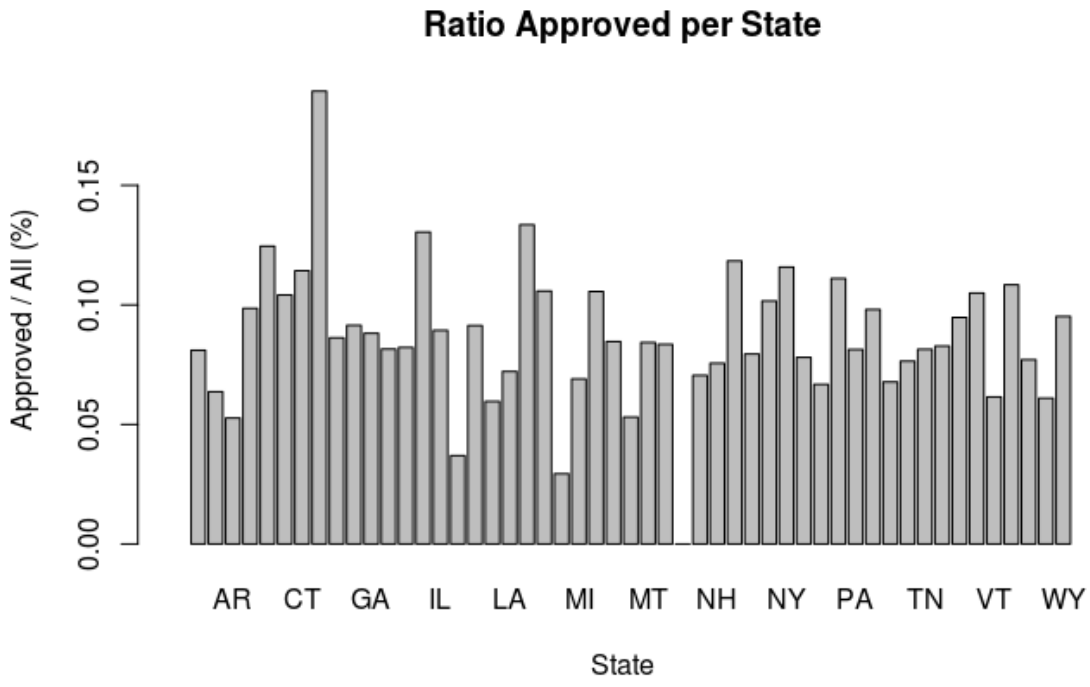

Figure 12 - Ratio Approved per State