Under the Microscope: An Investigation of Predictive Factors of Loan Acceptance

Hampton Leonard, hll4ce, Andrew Pomykalski, ajp5sb,
Tianye Song, ts7fx, Tyler Worthington, tjw4ry

Executive Summary:

A thought of many Americans today is that if they want to purchase a car or buy a house, they can just go to a loan provider, request a loan, and get that amount right away. But sadly, several hundreds of thousands of people in America cannot and do not get the loan amount they requested. The goal of our project is to investigate what factors are significant in whether or not a person receives a loan for the amount they requested. We will use a data set of with millions of observations outlining several factors collected on people applying for a loan. Through this analysis, we will attempt to determine factors that do and do not contribute to the success or failure of receiving a loan of a certain amount and to what degree each predictor does. We found that there are several factors that are significant in predicting whether the person received a loan or not, such as loan intent, amount requested, debt history, and employment. We also found that the longer an applicant was employed, the higher the chance at receiving a loan became. We believe that our results were limited because there were few predictors we were working with. There seems to be less information collected on applicants who get rejected leading to fewer predictors in our combined model of accepted and rejected.

1. Problem:

When a person applies for a loan, they are required to provide the loan provider with several pieces of information. These pieces are, but are not limited to, amount of loan requested, length of employment, credit history, and loan purpose. All of this data is stored and supposedly reviewed before a person is either accepted or rejected for a loan. However, many loan applications are rejected, and there can often be little understanding of the metrics behind these rejections and their influence on the final decision. There are very few, if any, loan providers that provide a clear outline or criteria in which a person needs to meet in order to receive a loan of any amount. This is where the problem lies. *How do I know if I am going to be seriously considered for the loan amount that I am requesting?* It would be beneficial to all loan applicants to have a robust understanding of what factors increase their chance of loan application acceptance, as well as have a method to predict the likelihood of their acceptance.

With this information more available to applicants, it could also benefit loan providers by hopefully providing them with more qualified applicants for loans. Loan providers get hundreds of thousands of loan requests per year and it takes them time and money to figure out who gets money. If people knew right away that they really did not have a high chance of getting a loan, they may not apply or may go elsewhere. This would cut down on the time a loan providing firm would have to allocate to go through all the underqualified people. By providing a more clear picture of the essential criteria for a loan, we can help speed up the loan process by cutting down on the number of rejections.

2. Hypothesis:

Looking through the loan data, there are several initial assumptions that have been made about loan applications over the years. Generally, success of loan application is associated with factors such as good credit and steady employment. Our hypothesis is that employment, debt history, location, and loan request information is related to whether or not a person's loan application is accepted. We believe that the longer the applicant has been employed, the better the chance he or she has at being accepted for the loan. This will be shown by comparing their odds ratio and checking which odds ratio is the highest.

3. Data:

A company known as Lending Club Corporation provides a thorough data set of loan applicants for the full year of 2015 and quarters one, two, and three for 2016. They break down their data sets into people who were accepted and people who were rejected. Each of the datasets has hundreds of thousands of observations for applicants but vary in the number of predictors. The data set that encompasses the applicants who were rejected has only 9 predictors where the data set of applicants who were accepted has 111 predictors. There are also several columns of missing data that will have to be dealt with and a large amount of cleaning in regards to column name difference and missing columns. We assert that these two data sets should give us an accurate and representative depiction of the loan environment over the past two years for our analysis.

4. Methodology:

4a. Data Cleaning:

Our first step in the data cleaning process was to look at similar columns in both the declined and accepted data sets to determine which factors we could use when we concatenate these two datasets together for use in our logistic model. We noticed that many of the columns/factors in the accepted dataset were not present in the declined dataset. For our logistic regression model to work we would have to concatenate the two datasets, and we can only concatenate them if they contain identical information column-wise. As a result, we had to filter down the features in the accepted dataset with comparison to the declined dataset. Once we identified these columns, "Amount Requested", "Loan Title", "Debt-To-Income Ratio", "State", "Employment Length", "Policy Code", "Application Date", we extracted these columns from both of the data sets and kept only those features. Labeling the accepted loans with a 1 and the declined as a 0 we next had to row bind the two data sets together. Before doing that however, we had to rename similar columns in the different data sets with the same label to merge the sets together.

After row binding the two data sets together, our next biggest problem was the wide variety of “Loan Title” responses. Since not every person had the same reason for requesting a loan nor wrote it the same way, we decided to bin all the responses into one of 12 titles listed in the table below.

Once we changed all these, we had to fix our “Application Date” column. The two data sets, Accepted and Declined, both wrote their dates in different ways. First, we made all the dates into the same form. Next, since we only had the first three quarters of 2016 data, we decided to label each loan request by quarter of which it occurred in and the year. This way we could do analysis over quarters to get a more representative picture of a block of time. Another reason why we chose to group our dates by quarter was due to the fact that we are using 206 data that has not been recorded for the fourth quarter yet.

Our last data cleaning step before we were able to begin modeling was to make our categorical predictors into factors to use them in our logistic model later.

Amount Requested	Amount of loan $ requested
Loan Title	Title/purpose/category of the loan
Debt-To-Income Ratio	(Total debt/total income) * 100%
State	State where the request took place
Employment Length	How long the individual had been employed
Policy Code	Possible value of 1 or 2, indicating type of loan
Application Date	Date the application for loan was filed

Figure 1: Feature Chosen

Vacation	Renewable Energy	Small Business	Debt Consolidation
Credit Card	Car	Moving	Home Improvement
House	Major Purchase	Medical	Other

Figure 2: Loan Titles

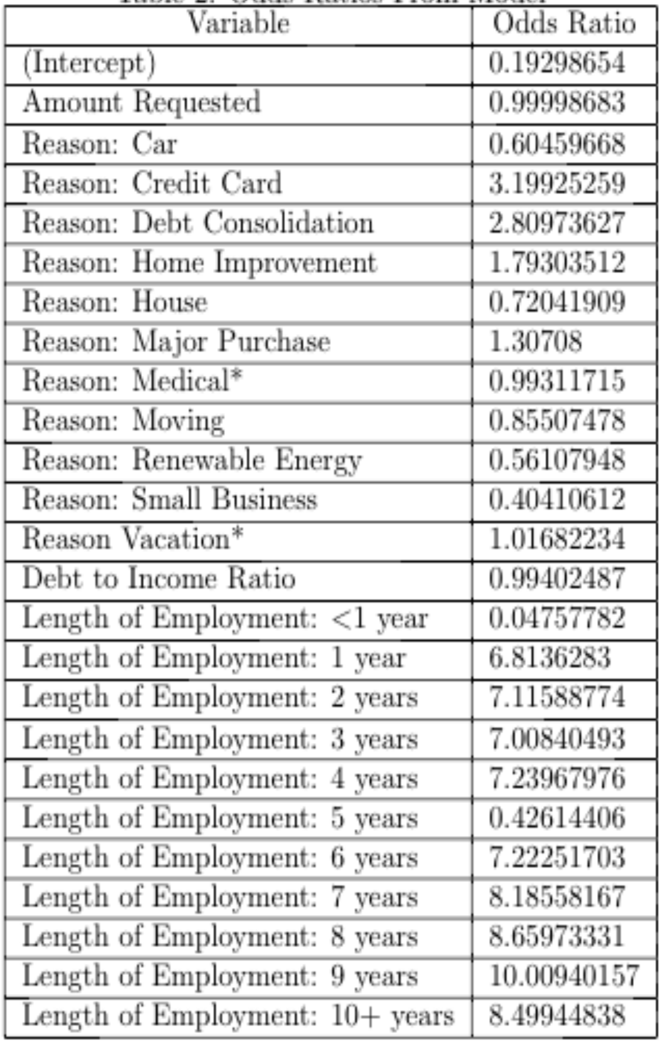
4b. Modeling:

Since our response variable was either accepted or declined for a loan, we decided that a logistic model would be the best for our problem. We began by throwing all the predictors into a model and checking our results. We then looked for insignificant terms based on p-values and removed those before running a new model. Once we settled on a model, we checked for multicollinearity because we are trying to build a model to interpret loan applications so multicollinearity could really throw off our model. Once we established there was no multicollinearity among the variables in our model, we ran it through 5 and 10-fold cross validation, and received an accuracy score of 0.936. Being satisfied with the cross-validated estimate of accuracy of our model, we formulated odds ratios to be able to interpret our results to possible suitors. Lastly, we ran ROC curves to check the ratio of our true positive and true negative rates.

5. Evaluation and Results:

By testing for significant variables in multiple logistic models, the final model included the variables: amount of money requested for the loan, the intent of use for the loan, the debt history of the applicant, and the employment length of the applicant. All levels of the categorical variables were significant, with the exception of loan titles relating to medical or vacation bills. More specifically, using a loan title related to paying for a car, paying for a house, moving, covering renewable energy, or paying for a small business will decrease your chances of receiving a loan, with small business loan titles decreasing the odds of loan acceptance by 59.59% relative to the reason being other, holding all else constant. Conversely, intents relating to credit card requests, debt consolidation, home improvement, and major purchases all improve chances. Employment length of less than one year decreases chances, as well as employment length of five years, all other employment lengths increased chances, relative to the length being unknown and holding all else constant. Amount requested also decreases loan acceptance chances with increase in amount, for every $1000 requested the odds decrease by 1.3%. Chances of receiving a loan decrease 0.06% per each 10% increase in debt to income ratio. Variables that were not found to be significant were the residing state of the applicant and the date of application. All resulting odds ratios can be seen in the table below.

Table 2: Odds Ratios From Model



Variable	Odds Ratio
(Intercept)	0.19298654
Amount Requested	0.99998683
Reason: Car	0.60459668
Reason: Credit Card	3.19925259
Reason: Debt Consolidation	2.80973627
Reason: Home Improvement	1.79303512
Reason: House	0.72041909
Reason: Major Purchase	1.30708
Reason: Medical*	0.99311715
Reason: Moving	0.85507478
Reason: Renewable Energy	0.56107948
Reason: Small Business	0.40410612
Reason Vacation*	1.01682234
Debt to Income Ratio	0.99402487
Length of Employment: <1 year	0.04757782
Length of Employment: 1 year	6.8136283
Length of Employment: 2 years	7.11588774
Length of Employment: 3 years	7.00840493
Length of Employment: 4 years	7.23967976
Length of Employment: 5 years	0.42614406
Length of Employment: 6 years	7.22251703
Length of Employment: 7 years	8.18558167
Length of Employment: 8 years	8.65973331
Length of Employment: 9 years	10.00940157
Length of Employment: 10+ years	8.49944838

Our first evaluation was to use Cross Validation to compare the cross validation estimation of accuracy of a couple models. We used 5-fold cross validation on our final model and found that our estimation of accuracy of our final model to be 0.936 which is pretty good. This accuracy estimation can range from zero to one. This is our first evaluation technique used to check our model.

The implementation of an ROC curve revealed the predictive power of our model to correctly classify loan acceptance or denial. We created predictions by using a logistic model trained on loan data from 2015 to predict loan response on quarter one, quarter two, and quarter three data from 2016. By comparing the predicted values to the true responses found in the 2016 dataset, the ROC curve below was generated.

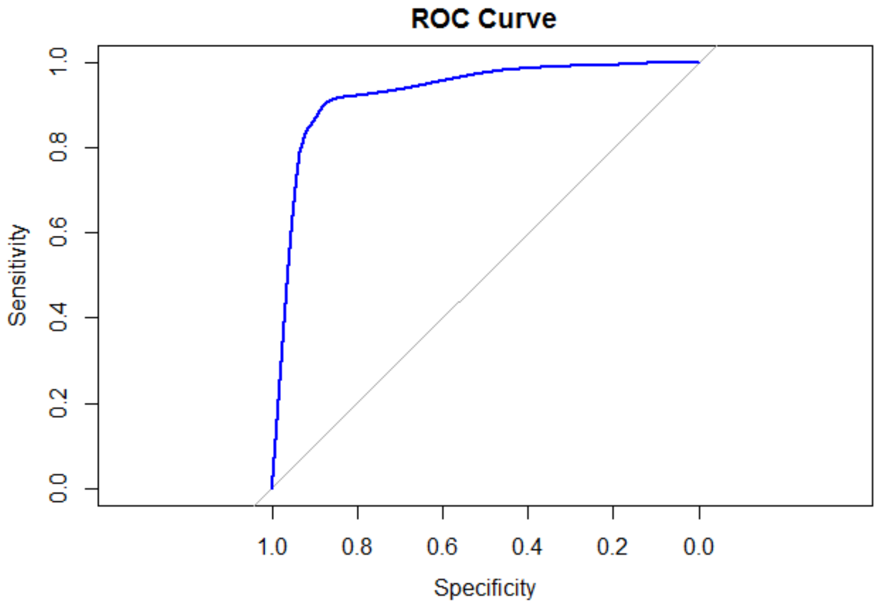


Figure 4: ROC Curve

This graph shows the sensitivity rates, or true positive rate, and the specificity, or true negative rate, of the model. The area under the curve was 0.9275. This shows that the model is very successful at correctly identifying loan acceptances when they were actually accepted as well as correctly identifying loan denials when they were actually denials. This supports our hypothesis that employment length, debt history, and debt request information are significant predictors for whether a loan is accepted or denied.

A confusion matrix was also created to further evaluate the predictive power of the model. In order to generate this, a cut-off value of 50% was chosen for the prediction probabilities provided by the model. All values greater than 50% were assigned a 1 for loan received and all values less than 50% were assigned a 0 for loan denied.

	Reference		
Predicted		**0**	**1**
	0	2,760,738	73,177
	1	188,875	257,684

Figure 5: Confusion Matrix

This shows the number of correct and incorrect cases predicted by the model. The majority of the predictions are correct predictions of cases where the loan was not received and correct predictions of where the loan was received. The confusion matrix revealed a prediction accuracy of 92% and a prediction accuracy confidence interval of (0.9198, 0.9204). The sensitivity rate of the model was 0.9360 and the specificity rate was 0.7788. This shows that our model performs with high accuracy, but has a slightly better true positive rate than true negative rate.

Lastly, we did explore to see if there was a difference in the loan variables and coefficients if we looked at larger loans versus smaller loans. We broke the data set into two groups; loans greater than the median which was about $10,000 and loans less than the median. This left us with sets of 1.4 million and 1.7 million respectively. After doing all the above fitting and testing on both, we looked to see if there was any major change in coefficients from the whole data set, to the large loans to the small loans.

We found that the signs of the coefficients were pretty similar signifying that the increase or decrease in your probability of receiving a large verses a smaller loan are the small. However, the coefficients on some the predictors did change significantly. We found that the coefficients for smaller loans were just about 1 unit larger meaning they increased the probability by about 1%. To us, that makes sense; for a smaller loan, one has a higher chance of receiving the loan because they are asking for a smaller, more manageable amount for the loan provider to come up with. The coefficients for these two data sets can be viewed in the appendix.

6. Conclusion:

One thing we learned about this problem is that there exists very few predictors that go into the process of loan applicant selection. With less than 10 predictors in both the accepted and declined datasets, it is difficult to get a good sense of all the factors that go into the loan selection process. We built and interpreted the logistic model with the data that we had available.

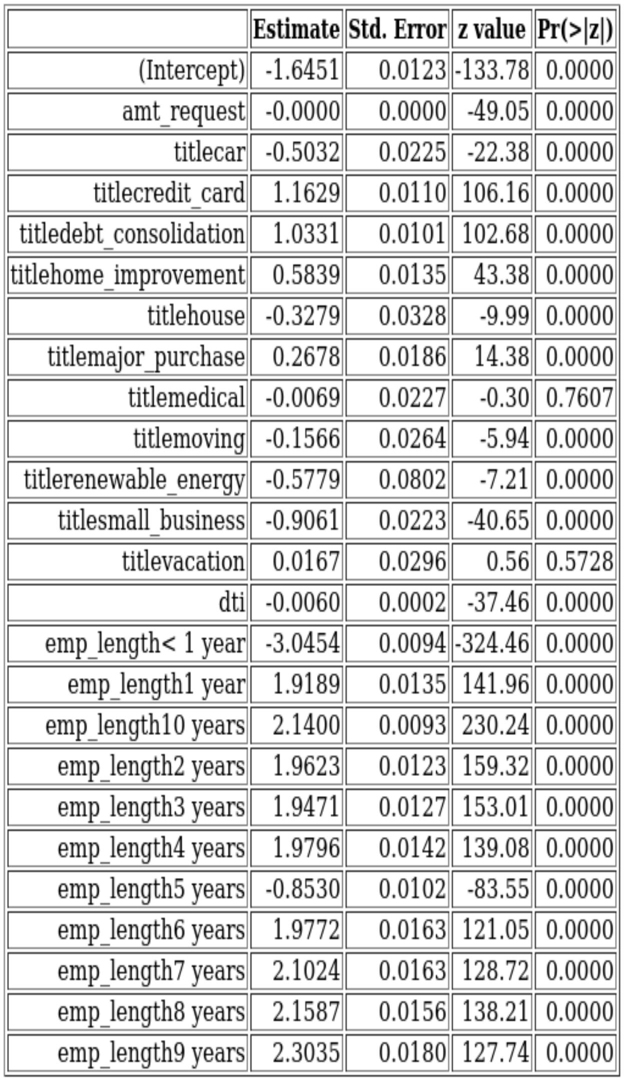
Our results have proven our hypotheses that employment, debt history, loan request information, and amount requested have an effect on loan acceptance. However, we also found that we could not prove our hypotheses that date or location of application had any effects on loan acceptance. Our other hypothesis that applicants' probability increased as the years of employment increased was mostly proven true as well. The odds ratio was largely positive for all the years greater than 1 however, at five years of employment the ratio decreases slightly.

We do not have an explanation for this but we see that the odds do increase and are largely in favor of the applicant the more years he or she is employed.

Some ideas for future improvements to this research could be to investigate more variables. Our model approach was found to be very accurate, but our data gave us access to a limited number of variables to analyze. It would be interesting to investigate other possible related variables, such as employment title and payment history. Our research has revealed important factors that may be used by loan companies as well as loan applicants to better predict the chances of giving or receiving a loan. This could allow companies to create streamlined processes for giving loan acceptances or denials, as well as help applicants determine what variables need to be improved or changed before applying.

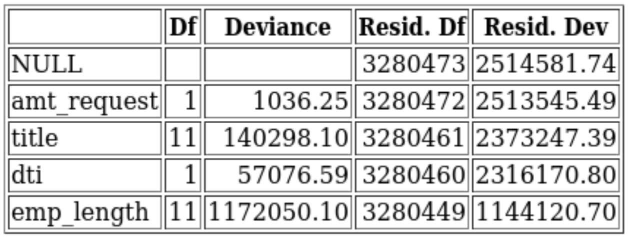
7. Appendix:

Summary of Final Model



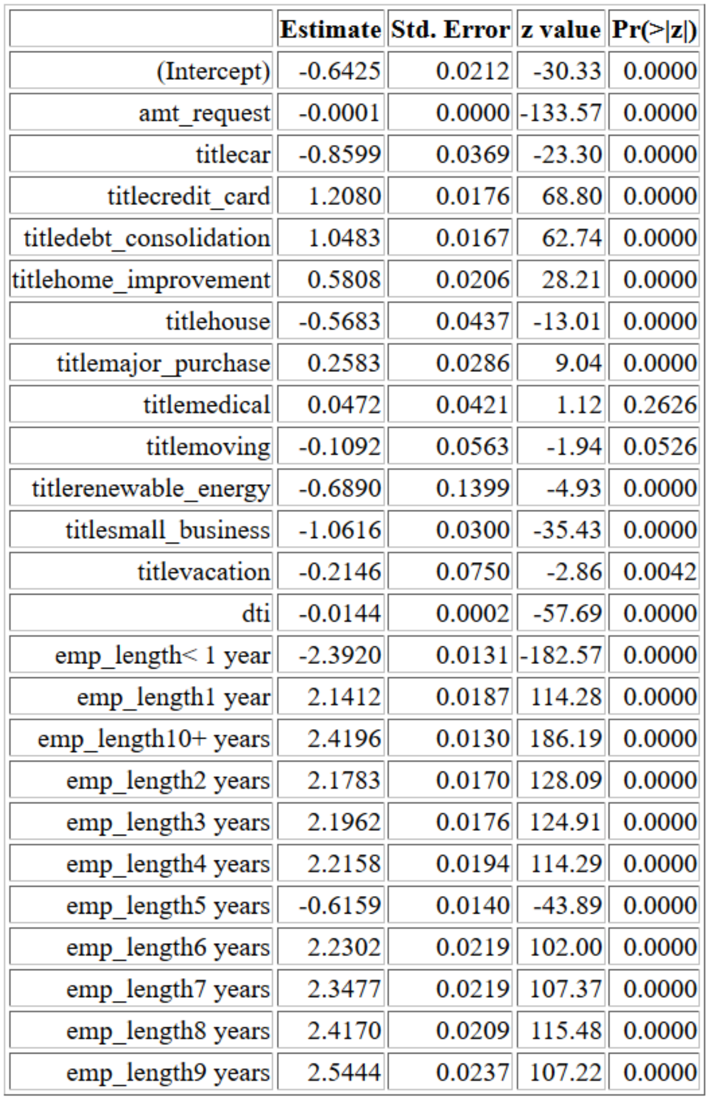
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.6451	0.0123	-133.78	0.0000
amt_request	-0.0000	0.0000	-49.05	0.0000
titlecar	-0.5032	0.0225	-22.38	0.0000
titlecredit_card	1.1629	0.0110	106.16	0.0000
titledebt_consolidation	1.0331	0.0101	102.68	0.0000
titlehome_improvement	0.5839	0.0135	43.38	0.0000
titlehouse	-0.3279	0.0328	-9.99	0.0000
titlemajor_purchase	0.2678	0.0186	14.38	0.0000
titlemedical	-0.0069	0.0227	-0.30	0.7607
titlemoving	-0.1566	0.0264	-5.94	0.0000
titlerenewable_energy	-0.5779	0.0802	-7.21	0.0000
titlesmall_business	-0.9061	0.0223	-40.65	0.0000
titlevacation	0.0167	0.0296	0.56	0.5728
dti	-0.0060	0.0002	-37.46	0.0000
emp_length< 1 year	-3.0454	0.0094	-324.46	0.0000
emp_length1 year	1.9189	0.0135	141.96	0.0000
emp_length10 years	2.1400	0.0093	230.24	0.0000
emp_length2 years	1.9623	0.0123	159.32	0.0000
emp_length3 years	1.9471	0.0127	153.01	0.0000
emp_length4 years	1.9796	0.0142	139.08	0.0000
emp_length5 years	-0.8530	0.0102	-83.55	0.0000
emp_length6 years	1.9772	0.0163	121.05	0.0000
emp_length7 years	2.1024	0.0163	128.72	0.0000
emp_length8 years	2.1587	0.0156	138.21	0.0000
emp_length9 years	2.3035	0.0180	127.74	0.0000

Anova of Final Model



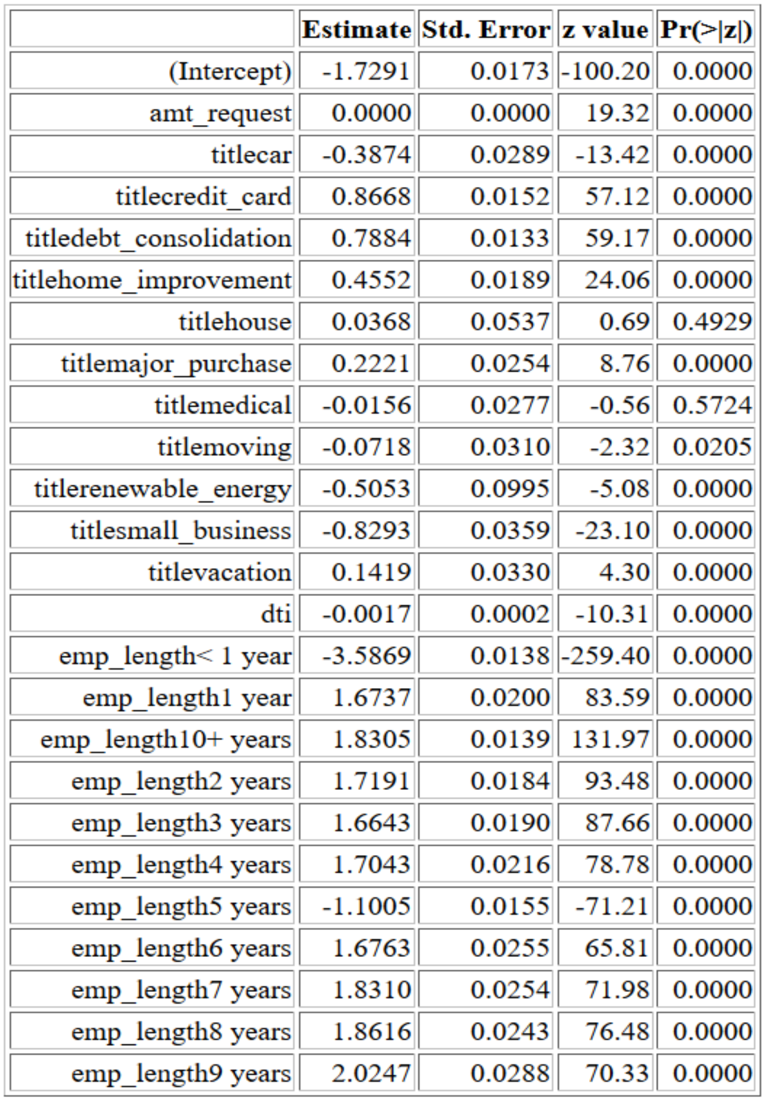
	Df	Deviance	Resid. Df	Resid. Dev
NULL			3280473	2514581.74
amt_request	1	1036.25	3280472	2513545.49
title	11	140298.10	3280461	2373247.39
dti	1	57076.59	3280460	2316170.80
emp_length	11	1172050.10	3280449	1144120.70

Summary of Model where Loan Amount Greater than $10,000



	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.6425	0.0212	-30.33	0.0000
amt_request	-0.0001	0.0000	-133.57	0.0000
titlecar	-0.8599	0.0369	-23.30	0.0000
titlecredit_card	1.2080	0.0176	68.80	0.0000
titledebt_consolidation	1.0483	0.0167	62.74	0.0000
titlehome_improvement	0.5808	0.0206	28.21	0.0000
titlehouse	-0.5683	0.0437	-13.01	0.0000
titlemajor_purchase	0.2583	0.0286	9.04	0.0000
titlemedical	0.0472	0.0421	1.12	0.2626
titlemoving	-0.1092	0.0563	-1.94	0.0526
titlerenewable_energy	-0.6890	0.1399	-4.93	0.0000
titlesmall_business	-1.0616	0.0300	-35.43	0.0000
titlevacation	-0.2146	0.0750	-2.86	0.0042
dti	-0.0144	0.0002	-57.69	0.0000
emp_length< 1 year	-2.3920	0.0131	-182.57	0.0000
emp_length1 year	2.1412	0.0187	114.28	0.0000
emp_length10+ years	2.4196	0.0130	186.19	0.0000
emp_length2 years	2.1783	0.0170	128.09	0.0000
emp_length3 years	2.1962	0.0176	124.91	0.0000
emp_length4 years	2.2158	0.0194	114.29	0.0000
emp_length5 years	-0.6159	0.0140	-43.89	0.0000
emp_length6 years	2.2302	0.0219	102.00	0.0000
emp_length7 years	2.3477	0.0219	107.37	0.0000
emp_length8 years	2.4170	0.0209	115.48	0.0000
emp_length9 years	2.5444	0.0237	107.22	0.0000

Summary of Model where Loan Amount Less than $10,000



	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.7291	0.0173	-100.20	0.0000
amt_request	0.0000	0.0000	19.32	0.0000
titlecar	-0.3874	0.0289	-13.42	0.0000
titlecredit_card	0.8668	0.0152	57.12	0.0000
titledebt_consolidation	0.7884	0.0133	59.17	0.0000
titlehome_improvement	0.4552	0.0189	24.06	0.0000
titlehouse	0.0368	0.0537	0.69	0.4929
titlemajor_purchase	0.2221	0.0254	8.76	0.0000
titlemedical	-0.0156	0.0277	-0.56	0.5724
titlemoving	-0.0718	0.0310	-2.32	0.0205
titlerenewable_energy	-0.5053	0.0995	-5.08	0.0000
titlesmall_business	-0.8293	0.0359	-23.10	0.0000
titlevacation	0.1419	0.0330	4.30	0.0000
dti	-0.0017	0.0002	-10.31	0.0000
emp_length< 1 year	-3.5869	0.0138	-259.40	0.0000
emp_length1 year	1.6737	0.0200	83.59	0.0000
emp_length10+ years	1.8305	0.0139	131.97	0.0000
emp_length2 years	1.7191	0.0184	93.48	0.0000
emp_length3 years	1.6643	0.0190	87.66	0.0000
emp_length4 years	1.7043	0.0216	78.78	0.0000
emp_length5 years	-1.1005	0.0155	-71.21	0.0000
emp_length6 years	1.6763	0.0255	65.81	0.0000
emp_length7 years	1.8310	0.0254	71.98	0.0000
emp_length8 years	1.8616	0.0243	76.48	0.0000
emp_length9 years	2.0247	0.0288	70.33	0.0000