Assessing Credit Risk in Buy Now, Pay Later (BNPL): A Case Study on Rising Defaults and Consumer Exposure

Ryan Leddy

June 13, 2025

1 Introduction & Motivation

Buy now, pay later, or BNPL as what it is informally called, is a form of short-term financing that allows consumers to purchase goods or services and pay for it in structured installments over time. This type of financing has increased in popularity over the last few years, allowing consumers to finance everything from a car to a burrito. Recently, financial services that offer BNPL such as Klarna or Affirm have seen an increase in defaults from users. This spike in defaults poses a risk for not just the company and associated bank, but the overall economy. This paper aims to determine if any characteristics can predict the likelihood of default, ultimately helping lenders understand the riskiness of the borrower.

By using data from Lending Club Accepted Loan data, this paper aims to determine what characteristics are most indicative of a borrower's likelihood of default on a BNPL type loan. Variables such as past delinquency's and number of credit inquiries will be analyzed to identify if they are statistically significant of default. Furthermore, to detect patterns in borrowers' profiles, K-means and hierarchal clustering will segment the data, so regressions can be ran within each cluster to examine how predictive factors can differ by borrower group. A multiple regression framework will be used to determine the relationship between borrower characteristics and default risk. Moreover, logistic regression and discrete models will be used as well as machine learning tools, such as random forest models and neural networks, to assess predicative

power. This mixture of descriptive and predicative analytics provides practical insight into comer credit risk in the growing BNPL market.

2 Data & Empirical Methodology

2.1 Data and Summary Statistics

Since BNPL firms like Klarna and Affirm do not have information about their user's default rates available, this paper uses data from the "Lending Club Platform," a financial services company that connects borrowers with investors for personal loans among other banking products. Lending Club provides data for peer-to-peer lending, which acts as a proxy for BNPL data since they both have comparable credit profiles. Both Lending Club borrowers and BNPL users typically have subprime or near-prime credit, lower incomes, a higher likelihood of limited credit history, and take on small to medium-sized personal loans. Using data from 2007 to 2018 and analyzing the main variables of annual income, interest rate, debt-to-income, FICO score range, delinquencies in the past two years, credit card usage, number of credit inquiries in the past six months, months since last default, and loan amount, this study aims to identify key predictors of default among borrowers.

Since the data set given has approximately seventy-five variables to choose from, this paper mainly focuses on nine main variables to see if this can help predict the likelihood of default. Self-reported annual income, interest rate of the loan, debt-to-

income ratio, lower bound of the borrowers FICO credit score range, number of thirty or more day past-due incidents in the borrowers credit file in the past two years, credit card usage, number of credit inquires in the last six months, months since the borrowers last default, and the amount of the loan are the main variables this study will dissect.

Looking at Table 1, we can see the summary stats given for the variables of discussion. When looking at a sample size of 50,000 lending club applicants, we can see that the average self-reported annual income is $78,376, although we can see that it is very variable to change with a standard deviation of $92,346, ranging from $0 to $9.5 Million. The average interest rates on loans were 13.08%, with the average debt-to-income ratio of 18.78 and a maximum outlier of 999. The mean FICO store for the lower bound is around 698, showing a general near prime borrower user base for the data. The past defaults in the last two years are low relative to the mean of 0.31, and the number of credit card inquires in the past six months averaged 0.57, with a maximum of 7. When looking at credit card usage and average loan amount, we can see it stands at 50.39% and $15,051, with a minimum of $1,000 and maximum of $40,000. An important statistic that stands out from the data is that among the 24,406 borrowers with past defaults, the average time since their last default is 34.45 months. Lastly, looking at the p-values we can see that they all show that the variables are statistically significant regarding the data.

3.2 Methodology

To pull further insight from the data, this paper used estimating models such as a linear regression, logistic regression for binary outcomes, and clustered regression models to account for grouped variation. To break down the data further, variables where created to indicate if a borrower defaulted on their loan by setting default equal to one. Using this variable, four linear regression models where created, the first following default and annual income, the second containing default, annual income, interest rate, and debt-to-income ratio, the third containing default and all nine variables, and the fourth using stepwise selection on the third model to make the model more precise by adding or removing predictors. Once all models where created, they were ran through a root mean square errors test and an Area Under the ROC curve (AUC) model as well to see which model was best fitted. The linear regression model serves as a linear probability model, where the dependent variable represents defaults in a binary form. It is used within this paper to show the effect of various borrower characteristics and their probability of default with each coefficient representing a one-unit change in the model, showing us the potential predictability probabilities of default for each variable given.

$$\text{default}_i = \beta_0 + \beta_1 \cdot \text{annual\_inc}_i + \beta_2 \cdot \text{int\_rate}_i + \beta_3 \cdot \text{dti}_i + \cdots + \beta_9 \cdot \text{loan\_amount} + \varepsilon_i$$

Moreover, alongside linear regression models, logistic regression models were also used within this study to estimate the probability of a borrower defaulting given a binary environment such as this one. Two logistic regression models where used, the first containing all nine variables, and the second containing a select few variables such as annual income, interest rates, and debt-to-income. After the regression was ran,

thresholding was incorporated to return a probability between 0 and 1. The threshold was set at 0.2, which entails that the model is more cautious and will flag more loans as potentially risky. The model shown represents the log of the odds of default as a function of a borrower with loan characteristics. By estimating the relationship between the binary variable, also known as default in our case, and the variables that would be our predicators, we are able to see which borrower attributes are associated with a higher likelihood of default.

$$\log\left(\frac{P(\text{default}_i = 1)}{1 - P(\text{default}_i = 1)}\right) = \beta_0 + \beta_1 \cdot \text{annual\_inc}_i + \beta_2 \cdot \text{int\_rate}_i + \beta_3 \cdot \text{dti}_i + \cdots + \beta_9 \cdot \text{loan\_amount} + \varepsilon_i$$

Furthermore, another way to find the best fitted model is by clustering. To better tailor our regression models, this paper uses clustering to group loan applicants into certain clusters based on similar financial profiles. By using a K-Mean Clustering model and a Hierarchical Clustering model we can better identify risk in subpopulations within the data. The K-Mean model helps us achieve a generated flat cluster to classify the borrowers into specific clusters. The Hierarchical clustering shows how the borrowers relate across several levels of similarity, with both clustering ultimately leading for a better classification of clusters to better pull insight of default risk from specific subgroups.

K-Mean Clustering:

$$\text{Objective: } \min_{\{C_k\}} \sum_{k=1}^{K} \sum_{x_i \in C_k} |x_i - \mu_k|^2$$

Hierarchical Clustering:

$$\Delta E_{(A,B)} = \frac{n_A n_B}{n_A + n_B} |\mu_A - \mu_B|^2$$

Once borrowers are grouped into three main clusters using a K-Mean model, statistics where gathered on the demographics of each cluster to gain insight on the characteristics of each. Then three separate logistic regression models where ran on all three clusters to separate predictive models for defaults based on key characteristics. To compare the models, a Chi-Square Test was ran to test whether default rates differ across clusters. A frequency test of defaults by each cluster was also ran to find the actual default rates within each cluster. Penultimately, a random forest model was implemented to enhance default prediction performance. The random forest model was built to combine many decision trees to reduce overfitting and increase overall predictive performance. Moreover, the model also models nonlinear relationships and interactions between variables. The model used ROC curves and AUC to measure the performance of the model in predicting defaults. Lastly, a neural network model was used to benchmark predictive accuracy. The model used ten borrower variables, five hidden layers, and one node to predict the probability of default. This model was implemented to capture very high nonlinear relationships that the logistic regressions could have missed.

3 Results

3.1 Descriptive Analytics

Looking at these nine variables in comparison to default rates, this paper worked to pull valuable descriptive analytics to find meaning from the data. From a sample of 50,000 and using information from Table 1's summary statistics, Table 2's Correlation Matrix, Graph 3 and 4's box plot graphs, and ANOVA tests, we are able to get a better understanding of the data at hand. The results researched have been pulled using data from the variables annual income, interest rate, debt-to-income ratio (DTI), FICO score, delinquencies, credit inquiries, and loan amounts.

Looking at Table 1, we are able to see the summary statistics of the variables at hand in comparison to default rates. Firstly, annual income shows us a mean of $78,376 and a very high standard deviation of $94,346 suggesting a very highly skewed distribution of the data. Furthermore, when looking at interest rates we can see a range of 5.31% to 30.99%, with those that default, with a mean of 15.79%, paying much more than those of non-defaulters, with a mean of 12.7%. When looking at FICO scores, we can see that they are lower among defaulters with a mean of 687, than non-defaulters with a mean of 700, showing that high credit is a sign for borrowers to default less. When looking at the debt-to-income ratio, defaulters show a mean ratio of 20.15%, compared to that of non-defaulters with a mean of 18.6%, implying that a higher financial burden indicates higher risk. Looking to Table 2, the correlation matrix, we can see for interest

rates it is positively correlated with default at a rate of 0.2045, indicating that higher risk borrowers pay higher rates. For FICO scores, we can see that it is negatively corrected with default rates with a correlation of -0.1217, showing that credit worthiness indicates the rate of default. For Debt-to-Income and credit card usage, we can see a moderate positive correlation with default, showing that it is correlated but not to the degree of interest rates. For variables like annual income and loan amount, we see a very low correlation with default, suggesting an overall weaker relationship between those variables and default. When looking at the data overall, we can see that delinquency rates, interest rates, and credit inquiries are consistently higher among defaulters, as well as credit card usage is higher in defaulters at a rate of 54.36%, compared to that of non-defaulters for 49.85%.

Looking to Graph 3 and 4, we can see that interest rates and FICO scores show separation between defaulters and non-defaulters, with defaulters clustering around higher interest rates and lower credit scores. Moreover, when looking at the ANOVA results, we can see that for interest rates by default status it shows a very high F-Statistic and low P-value, confirming that it is statistically significant and has a difference in means.

3.2 Predictive Analytics

To better understand how the variables relate to the default rates, several test where ran such as K-Mean and Hierarchical clustering, linear regression, logistic regression models,

random forest model, and neural network models. Using these models, this paper works to understand which variables lead to a higher rate of default.

Looking to the four main linear regression models, Model 3 that used all variables, and Model 4 that used stepwise selection performed the best over Model 1 and 2. The Root Mean Squared Error (RMSE) for Model 1 was 0.3336, Model 2 was 0.3271, Model 3 was 0.3265, and Model 4 was 0.3265 as well. Since Model 3 and 4 have the lowest RMSE, they have the strongest fit. When looking at AUC scores Model 4 had the highest with a score of 0.66979, indicating moderate to high classification power. Overall, Model 4 performed the best out of the models, but still lacked in performance from being limited by linear assumptions.

For logistic regression models, Model 1 which contained all variables had the highest AUC of 0.67139 next to Model 2 that had selected variables with an AUC of 0.66252. For Model 1, key variables that predicted default was interest rates, annual income, FICO score, loan amount, and number of credit inquires in the last six months. For Model 2, the variables that showed high signs of default where annual income, interest rate, and debt-to-income. Overall, Model 1 performed the best out of the two and performed better than the linear regression models. Moreover, the models showed that interest rates, number of credit inquires in the last six months, and loan amount where consistently predicative.

When looking at the random forest models, Model 1 had an AUC of 0.65796 and Model 2 had a AUC of 0.65858, which is close but still under that of the logistic

regression. After one hundred iterations of the model, the best AUC was 0.6626, which entails that the moderate AUC score was consistent. The random forest model shows a strong nonlinear model but still yields a lower AUC. While the random forest model is not the best AUC, it does capture non-linearities missed in other models such as linear regression models or logistic.

The neural network model seen in Graph 2, used ten variables with five hidden layers and one node. The error function showed a value of around 3,074 which is large, and an AUC comparable to that of the random forest model. The neural network model adds more flexibility to a model but does not outperform that of the logistic regression model.

Once the K-Means and Hierarchical Clustering models have been ran, the K-Means cluster was divided into three main sectors. Sector 1 had few observations and very high income and loan amount, likely showing outliers. Cluster 2 showed 49,967 observations, near-prime consumers, and a moderate default rate. Cluster 3 showed low income and high debt-to-income in a small sample. Once all three clusters were regressed, Cluster 2 showed similar significant variables to that of other significant models, such as annual income, interest rate, FICO score, inquiry in the last six months, and loan amount all had a p-value less than 0.01, signally to its statistical significance.

Overall, the highest performing model was Model 1 with all variables present for the logistic regression model and the logistic regression model ran on Cluster 2. Key

variables that showed a high likeliness in default rates was interest rates, annual income, FICO score, loan amount, and number of credits inquires in the last six months.

4 Summary and Conclusion

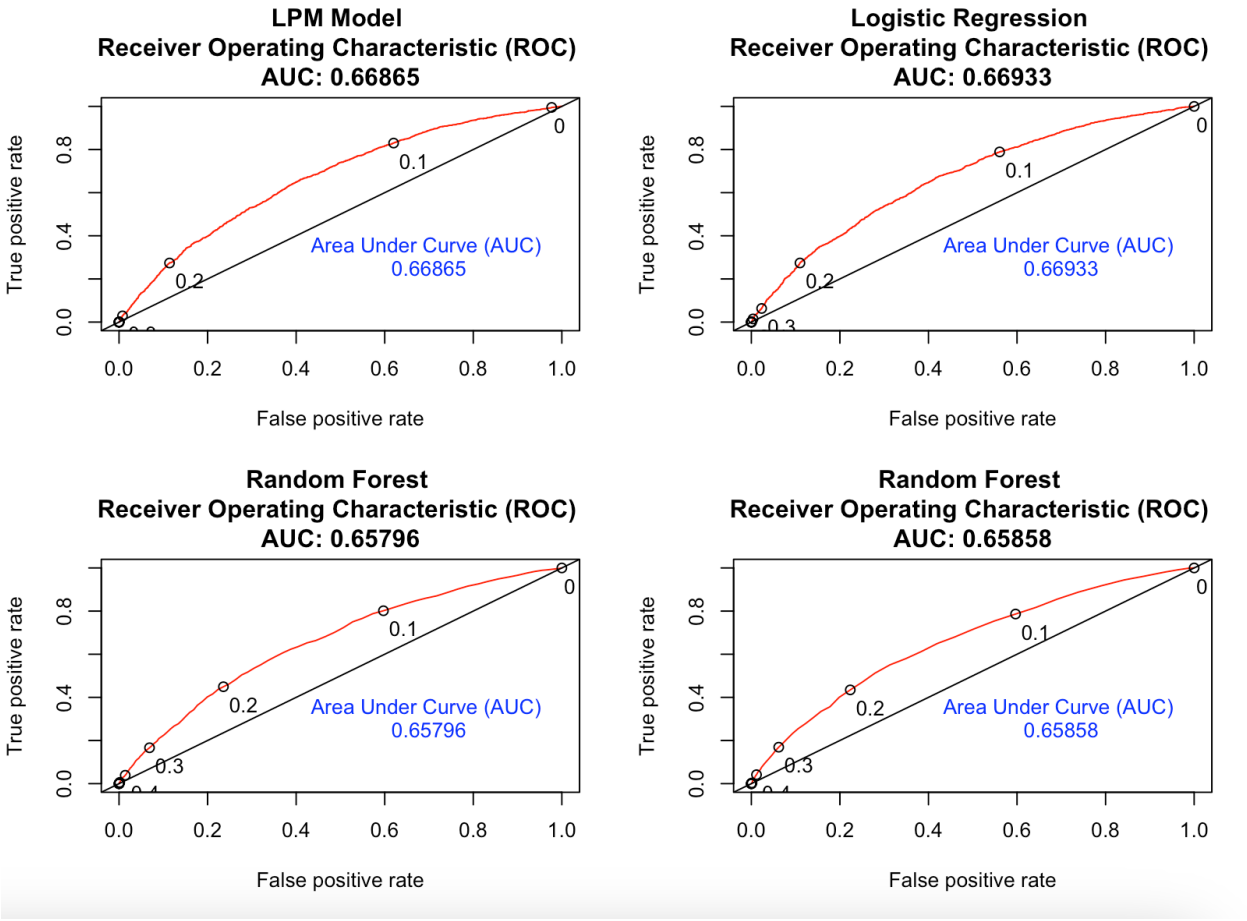
With the noticeable increase in popularity of Buy Now, Pay Later financing, a more noticeable increase in defaults has been seen as well. To help lenders better understand the riskiness of the borrower, this paper uses a myriad of models and to better understand what factors are indicative of a borrower defaulting on their debt. Using data from Lending Club loan data, and running linear regression, logistic regression, K-Mean and Hierarchical clustering, random forest, and neural network models, the findings concluded that the logistical regression model yielded the best fitted model, showing that the variables interest rates, annual income, FICO score, loan amount, and number of credits inquires in the last six months, are shown to have the highest likelihood of predicting defaults in borrowers. This entails that borrowers with high interest rates on their loans, low annual income, low FICO scores, a high loan amount, and a high number of credits inquires in the last six months will have a higher rate on defaulting on their debt. Going forward, firm such as Affirm and Klarna should keep these variables in mind when offering BNPL loans to borrowers. Moreover, to achieve a better model that is able to predict variables that lead to higher rates of default, this study should be ran again incorporating all seventy-five variables as well as the full sample size of the data. With

this amount of data and the number of variables given, different variables with higher AUC scores can be calculated, leading to better predictability going forward.

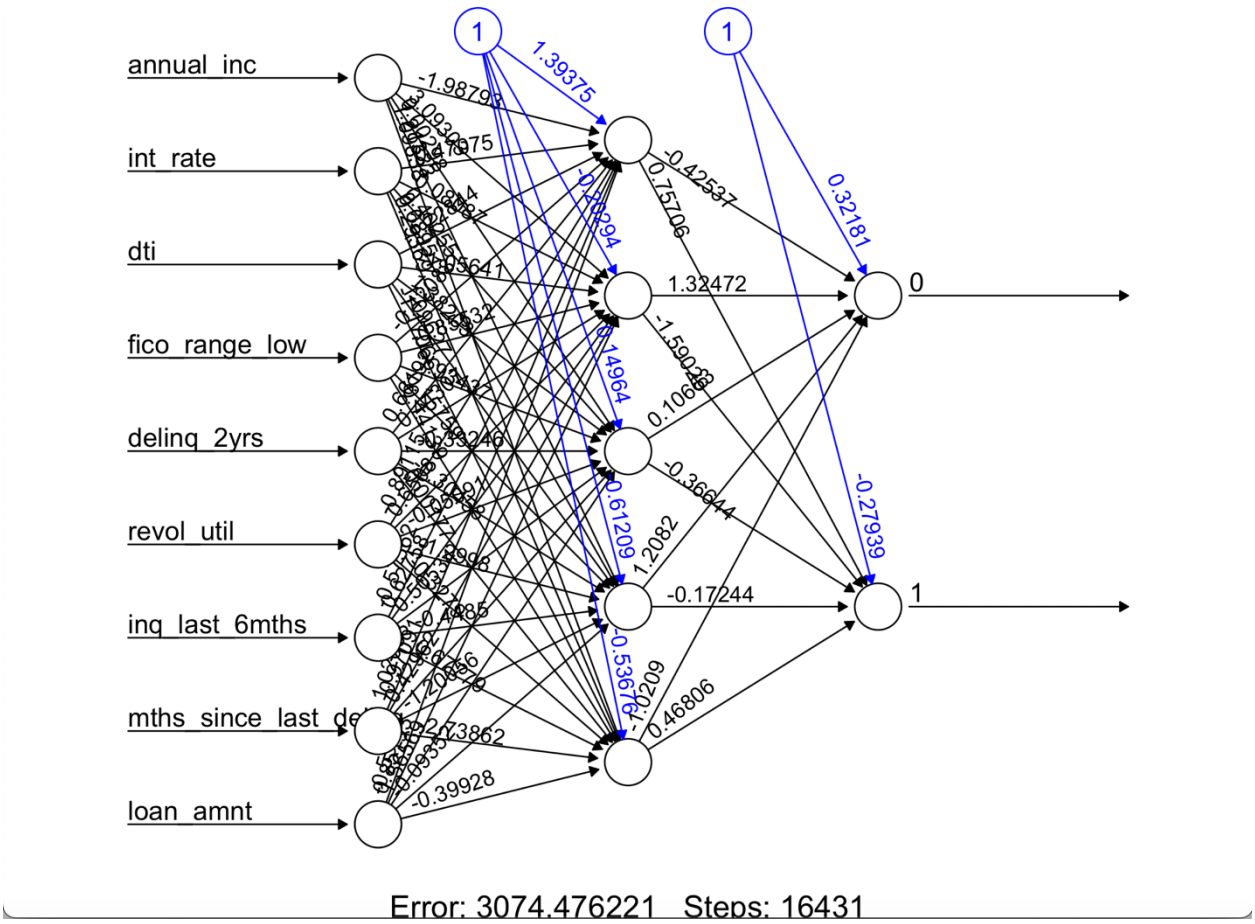
5 Bibliography

Variable	Sample	Mean	Standard Deviation	Minimum	Maximum	P-Value
Annual Income	**50000**	**78375.92**	**92345.85**	**0**	**9550000**	**0.01**
Interest Rate	**50000**	**13.078**	**4.81**	**5.31**	**30.99**	**0.01**
Debt-to-Income	**49971**	**18.78**	**13.91**	**0**	**999.00**	**0.01**
FICO Range Low	**50000**	**698.78**	**33.10**	**660.00**	**845.00**	**0.01**
Delinquency 2 Years	**50000**	**0.31**	**0.89**	**0**	**29.00**	
Credit Card Usage	**49968**	**50.39**	**24.71**	**0**	**134.40**	**0.01**
Number of Credit Inquiry in last 6 Months	**50000**	**0.57**	**0.87**	**0**	**7.0**	**0.01**
Months since last Default	**24406**	**34.45**	**21.94**	**0**	**145.0**	**0.01**
Loan Amount	**50000**	**15050.87**	**9220.81**	**1000.00**	**4000.0**	**0.01**

Table 1: Summary Stats for Key Variables.



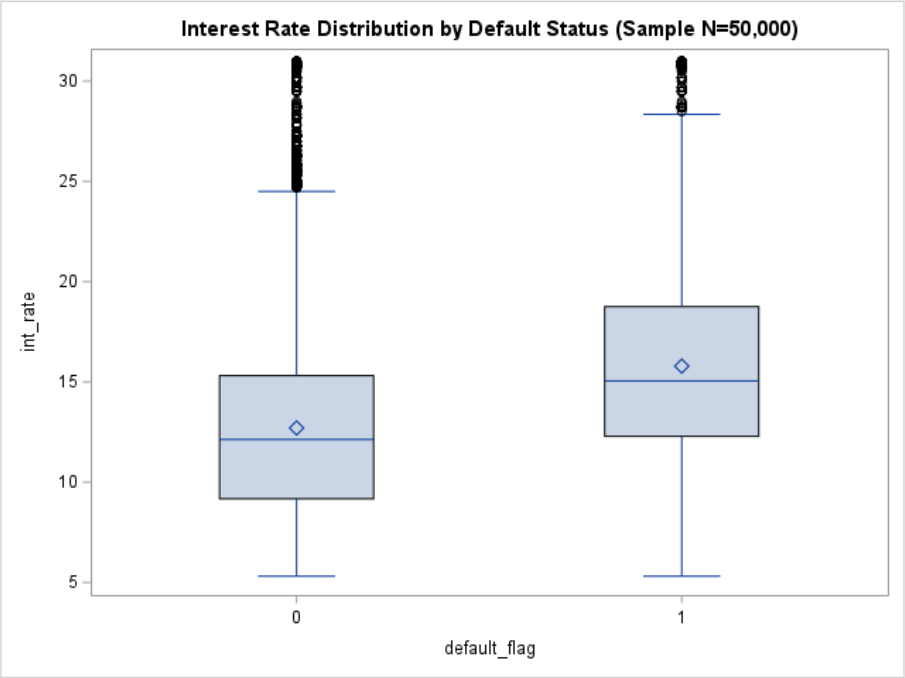
Graph 1: AUC Plots For Linear Regression, Logistic Regression, and two Random Forest Models.



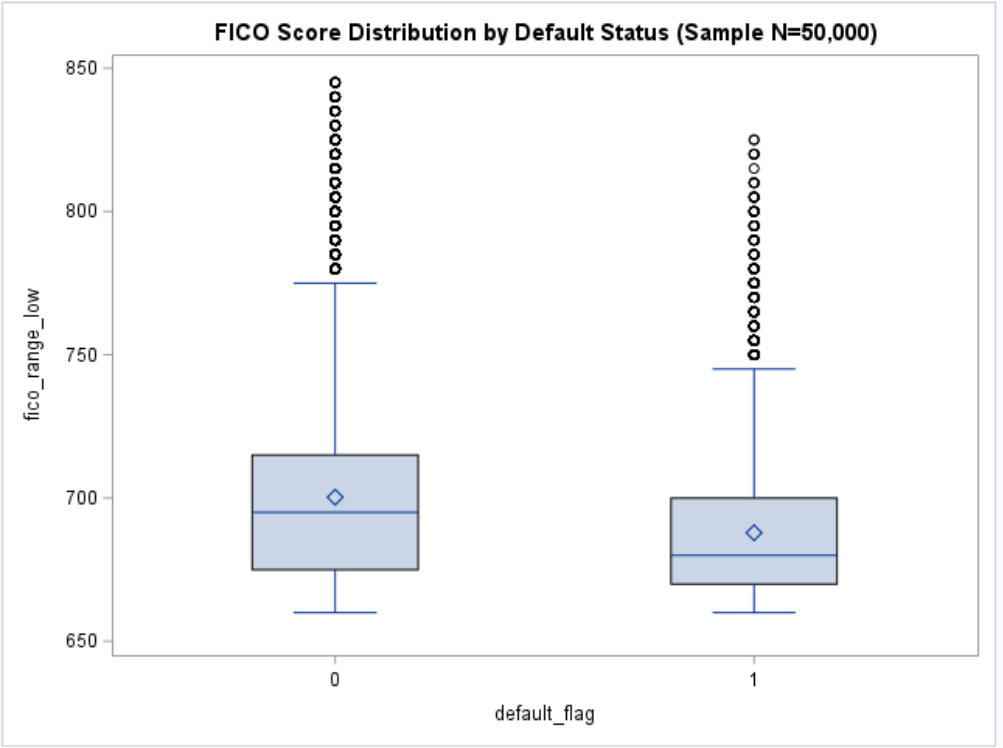
Graph 2: Neural Network Model

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations										
	annual_inc	int_rate	dti	fico_range_low	delinq_2yrs	revol_util	inq_last_6mths	mths_since_last_delinq	loan_amnt	default_flag
annual_inc	1.00000 2223667	-0.05071 <.0001 2223667	-0.08225 <.0001 2221994	0.03712 <.0001 2223667	0.02578 <.0001 2223667	0.02794 <.0001 2221943	0.02040 <.0001 2223666	-0.03051 <.0001 1082266	0.19670 <.0001 2223667	-0.02488 <.0001 2223667
int_rate	-0.05071 <.0001 2223667	1.00000 2223667	0.12442 <.0001 2221994	-0.41735 <.0001 2223667	0.05929 <.0001 2223667	0.26464 <.0001 2221943	0.19127 <.0001 2223666	-0.04359 <.0001 1082266	0.09679 <.0001 2223667	0.20450 <.0001 2223667
dti	-0.08225 <.0001 2221994	0.12442 <.0001 2221994	1.00000 2221994	-0.02823 <.0001 2221994	-0.01202 <.0001 2221994	0.11513 <.0001 2220274	-0.01062 <.0001 2221993	0.01347 <.0001 1081672	0.04363 <.0001 2221994	0.03571 <.0001 2221994
fico_range_low	0.03712 <.0001 2223667	-0.41735 <.0001 2223667	-0.02823 <.0001 2221994	1.00000 2223667	-0.17926 <.0001 2223667	-0.47788 <.0001 2221943	-0.09339 <.0001 2223666	0.10222 <.0001 1082266	0.11057 <.0001 2223667	-0.12166 <.0001 2223667
delinq_2yrs	0.02578 <.0001 2223667	0.05929 <.0001 2223667	-0.01202 <.0001 2221994	-0.17926 <.0001 2223667	1.00000 2223667	-0.00001 0.9866 2221943	0.02537 <.0001 2223666	-0.55289 <.0001 1082266	-0.01004 <.0001 2223667	0.01968 <.0001 2223667
revol_util	0.02794 <.0001 2221943	0.26464 <.0001 2221943	0.11513 <.0001 2220274	-0.47788 <.0001 2221943	-0.00001 0.9866 2221943	1.00000 2221943	-0.07907 <.0001 2221942	0.00452 <.0001 1081289	0.10004 <.0001 2221943	0.06640 <.0001 2221943
inq_last_6mths	0.02040 <.0001 2223666	0.19127 <.0001 2223666	-0.01062 <.0001 2221993	-0.09339 <.0001 2223666	0.02537 <.0001 2223666	-0.07907 <.0001 2221942	1.00000 2223666	0.01333 <.0001 1082266	-0.02548 <.0001 2223666	0.08766 <.0001 2223666
mths_since_last_delinq	-0.03051 <.0001 1082266	-0.04359 <.0001 1082266	0.01347 <.0001 1081672	0.10222 <.0001 1082266	-0.55289 <.0001 1082266	0.00452 <.0001 1081289	0.01333 <.0001 1082266	1.00000 1082266	-0.01113 <.0001 1082266	-0.01378 <.0001 1082266
loan_amnt	0.19670 <.0001 2223667	0.09679 <.0001 2223667	0.04363 <.0001 2221994	0.11057 <.0001 2223667	-0.01004 <.0001 2223667	0.10004 <.0001 2221943	-0.02548 <.0001 2223666	-0.01113 <.0001 1082266	1.00000 2223667	0.02194 <.0001 2223667
default_flag	-0.02488 <.0001 2223667	0.20450 <.0001 2223667	0.03571 <.0001 2221994	-0.12166 <.0001 2223667	0.01968 <.0001 2223667	0.06640 <.0001 2221943	0.08766 <.0001 2223666	-0.01378 <.0001 1082266	0.02194 <.0001 2223667	1.00000 2223667

Table 2: Correlation Matrix



Graph 3: Boxplot for Interest Rates



Graph 4: Boxplot for FICO Score