**Probability of Default in peer-to-peer lending**

Gregory Stephen Happ

**Introduction**

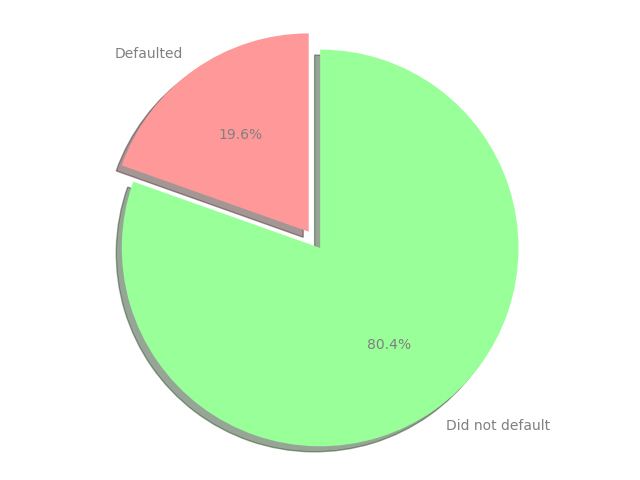
When assessing both the risk level of a loan, and the fair interest rate charged to a borrower, the probability of default is an extremely important variable. The probability of default is the probability that a borrower will not pay back in full. Because of its importance, it is used in the expected loss equation for a loan by credit analysts.

Expected Loss = Probability of Default \* Loss given Default \* Exposure at Default

Despite how vital the probability of default is in credit analysis, it is not a known variable when performing credit analysis, and it must be estimated. For individual borrowers, the two most important variables when estimating the probability of default are the borrower’s credit score, and the borrower’s debt-to-income ratio.

The researcher analyzed over 1.3 million peer-to-peer loans from Lending Club, and focused on analyzing their average default rates, their borrower’s FICO credit score grades, and their borrower’s debt-to-income ratios. Also, the researcher made a probability of default calculator by training a support vector machine algorithm over 1300 loans that were randomly selected from the population of over 1.3 million loans. When tested on over 1.3 million loans, the model was 80% accurate.

**Probability of Default**

19.57% of all loans defaulted. Thus, the probability of default for a given peer-to-peer loan from Lending Club is estimated to be about 20%.

**Debt-to-income ratios**

Debt-to-Income Ratio Summary Statistics:

*Count* 1,340,619

*Mean* 18.29820

*Std* 11.16129

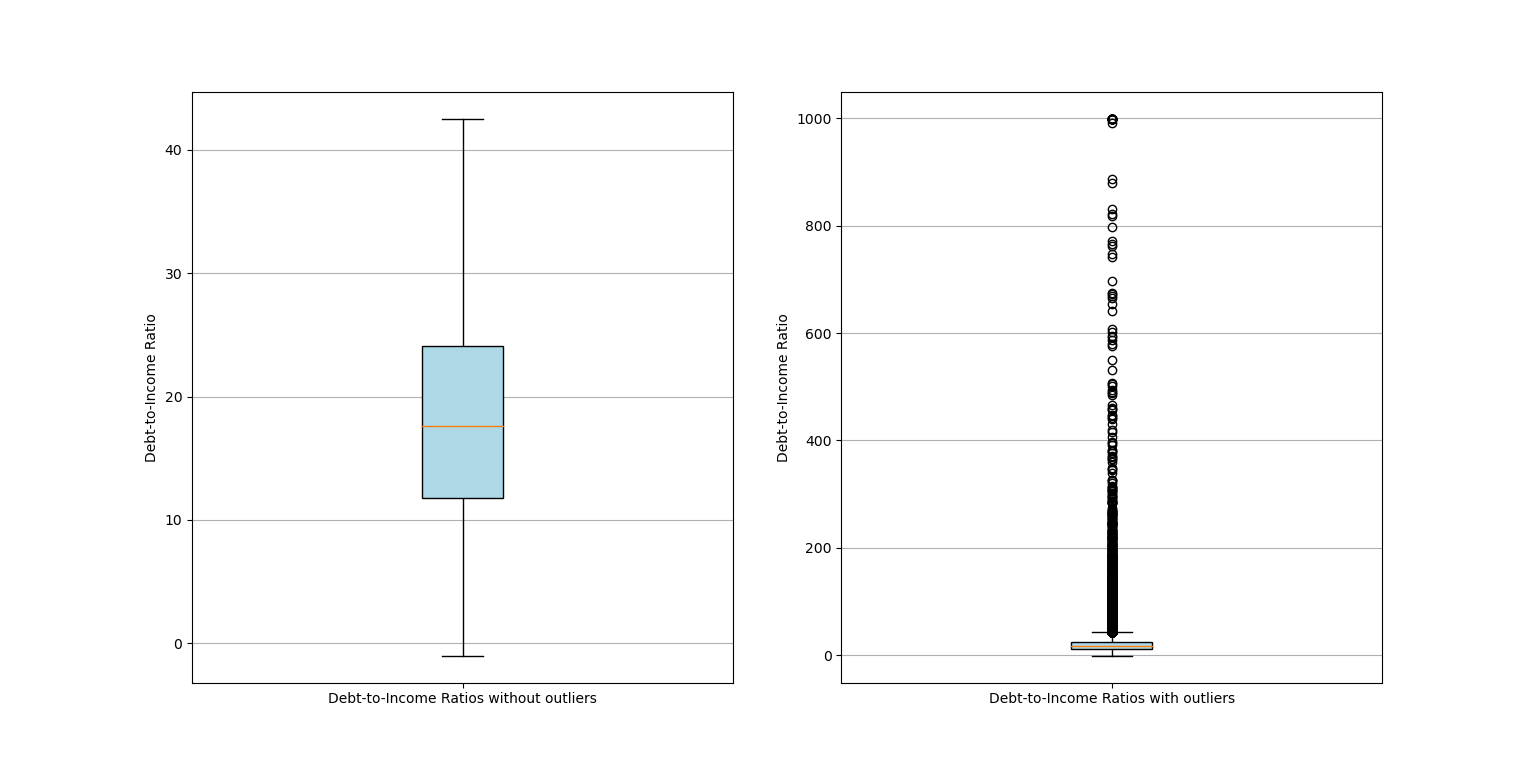
*Min* -1.000000

*25%* 11.80000

*50%* 17.63000

*75%* 24.08000

*Max* 999.0000



As seen in the box plots, there are clearly lots of outliers for debt-to-income ratios. Those outliers themselves are worrying regarding the overall riskiness of peer-to-peer lending for Lending Club. However, when outliers are removed, debt-to-income ratios appear a lot more normal and less risky, as most borrowers have a debt-to-income ratio below 30. The median debt-to-income ratio is 17.63. Half of all loans have debt-to-income rations between 11.80 and 24.08. A quarter of all borrowers have debt-to-income ratios below 11.80.

**Credit Scores**

Credit Risk Grade Relative Frequency Table

*Grade Percent*

A 0.170495

B 0.289430

C 0.285385

D 0.151939

E 0.071190

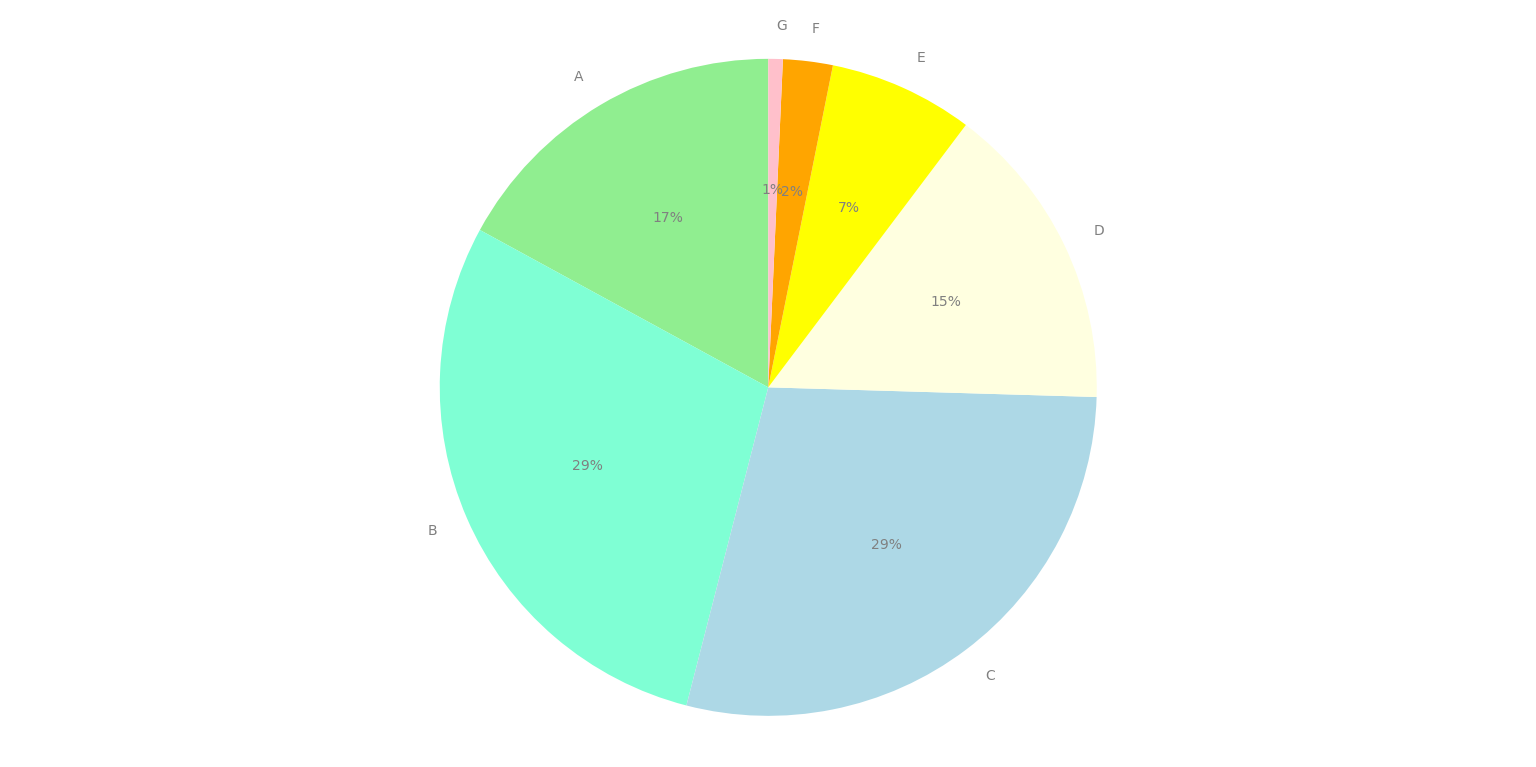
F 0.024454

G 0.007108

The FICO credit score letter of borrowers, in order of frequency, are:

B (29%), C(29%), A(17%), D(15%), E(7%), F(2%), G(1%)

FICO credit score letter grades usually range between A through D. Lending Club, however, has some borrowers with such poor FICO credit ratings, that there are borrowers with E, F and G credit score letter grades. However, E, F, and G borrowers only compromise about 10% of all borrowers, so they do not represent most of Lending Club’s peer-to-peer loans available to lenders.



**Probability of default by FICO credit risk score grade**

Probability of Default by Credit Risk Grade:

*Grade Probability of Default*

A 0.060266

B 0.131853

C 0.217966

D 0.292751

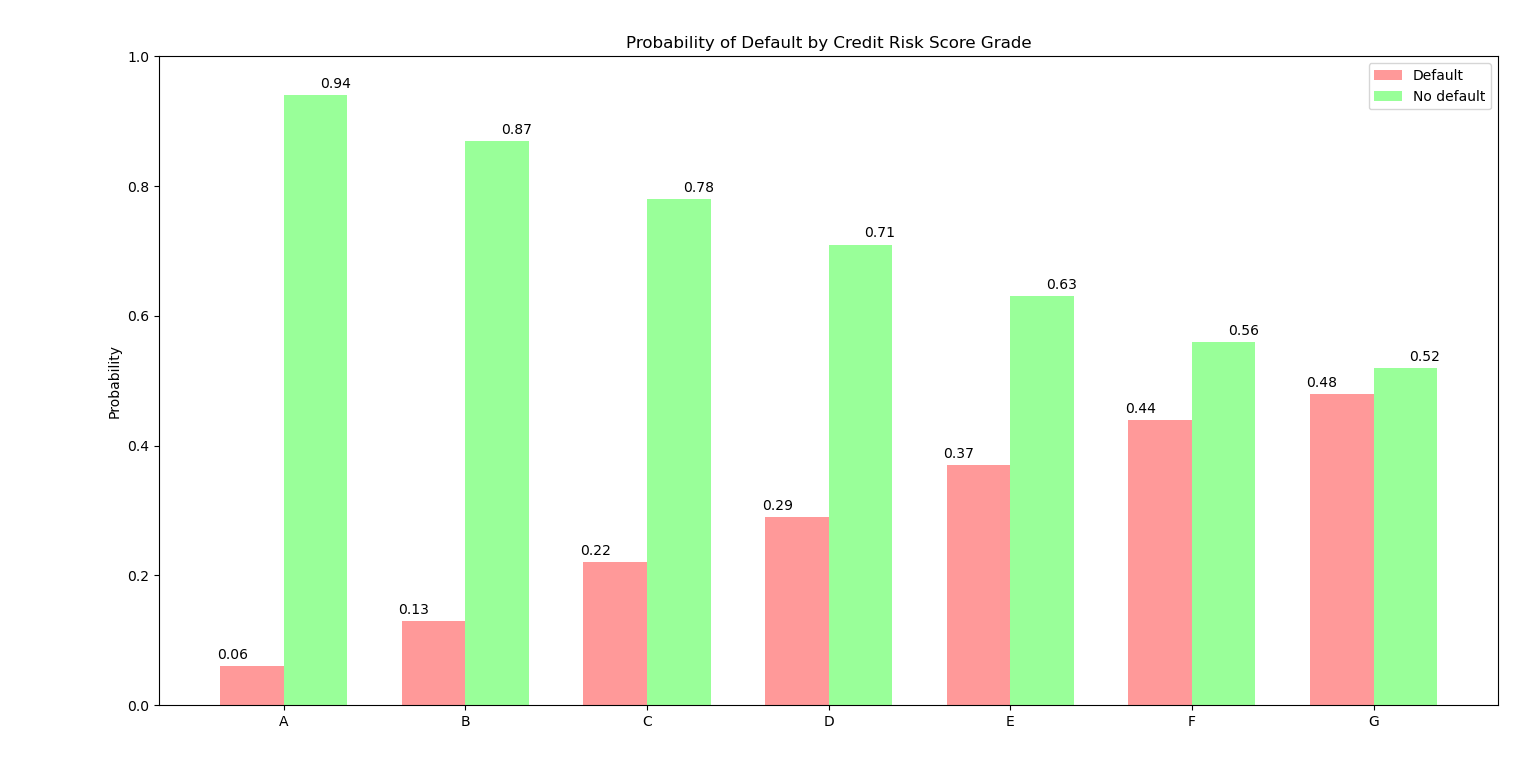
E 0.372200

F 0.437849

G 0.477490

There is clearly a very strong and positive relationship between better FICO score letter grades and lower probabilities of default. For example, “A” rated borrowers had an average probability of default of 6%, which is far lower than “G” rated borrowers’ 48% average.

Using the average probability of default by FICO score grade to estimate the probability of default for a given loan with a given grade, A’s have a probability of default of 6%, B: 13%, C: 22%, D 29%, E 37%, F 44%, and G 48%.



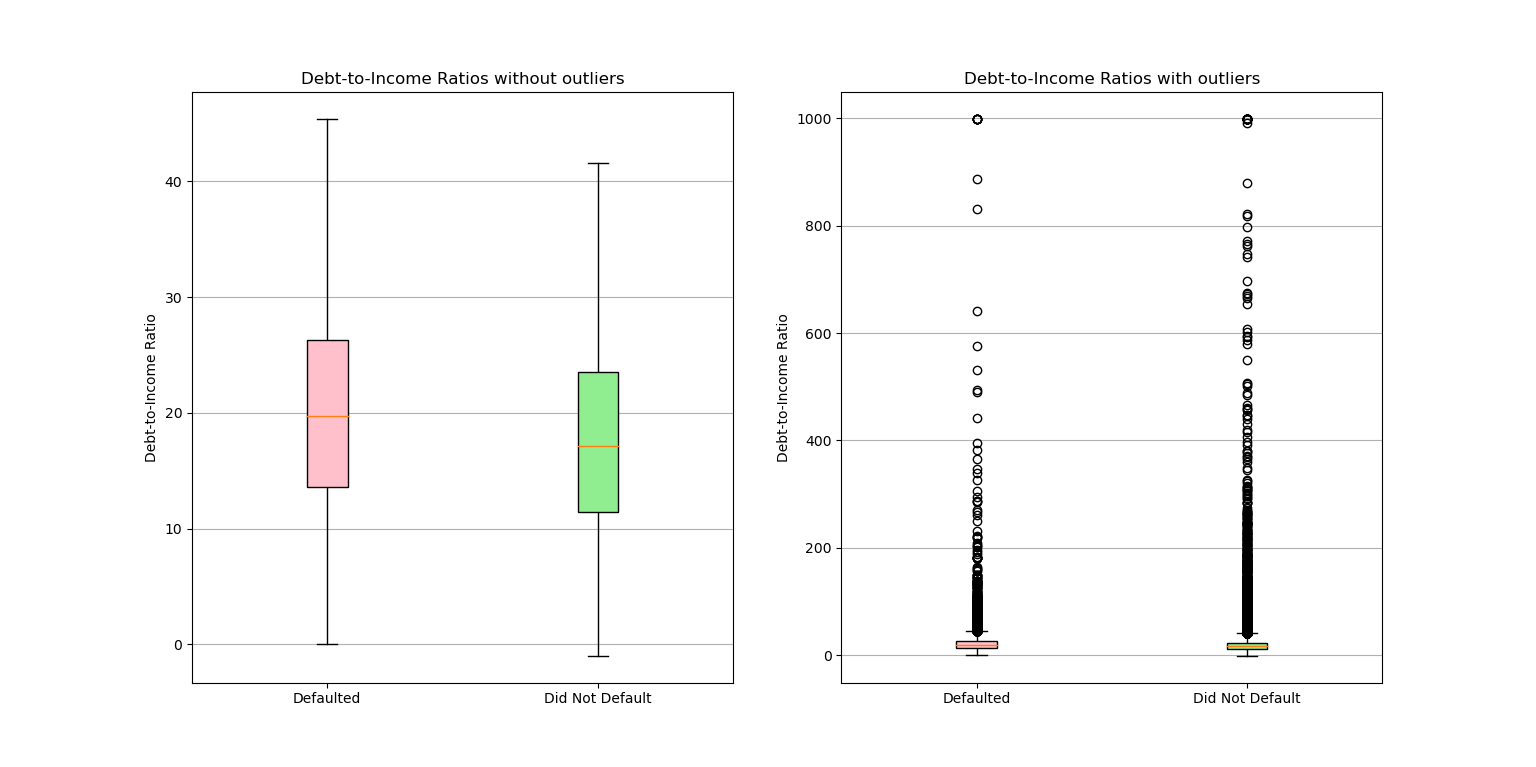
**Debt-to-income ratios by defaults**

Debt-to-Income Ratio Summary Statistics by Defaulted

*Defaulted count mean std min 25% 50% 75% max*

*No* 1078234.0 17.848802 11.084099 -1.0 11.44 17.14 23.49 999.0

*Yes* 262385.0 20.144915 11.286755 0.0 13.56 19.76 26.30 999.0



There is clearly a positive relationship between defaulting and higher debt-to-income ratios. Although the magnitude of the relationship is not drastically large, the size of the relationship was at least a few points. For example, the average debt-to-income ratio for defaulters was 20, and the average for those who did not default was 18. This difference wasn’t large, but it was still there.

**Debt-to-income ratios by grade**

Debt-to-Income Ratio Summary Statistics by Credit Risk Grade

*Grade count mean std min 25% 50% 75% max*

*A* 228569.0 15.607868 9.005063 0.0 9.89 14.98 20.72 999.0

*B* 388016.0 17.369628 10.248553 -1.0 11.27 16.77 22.85 999.0

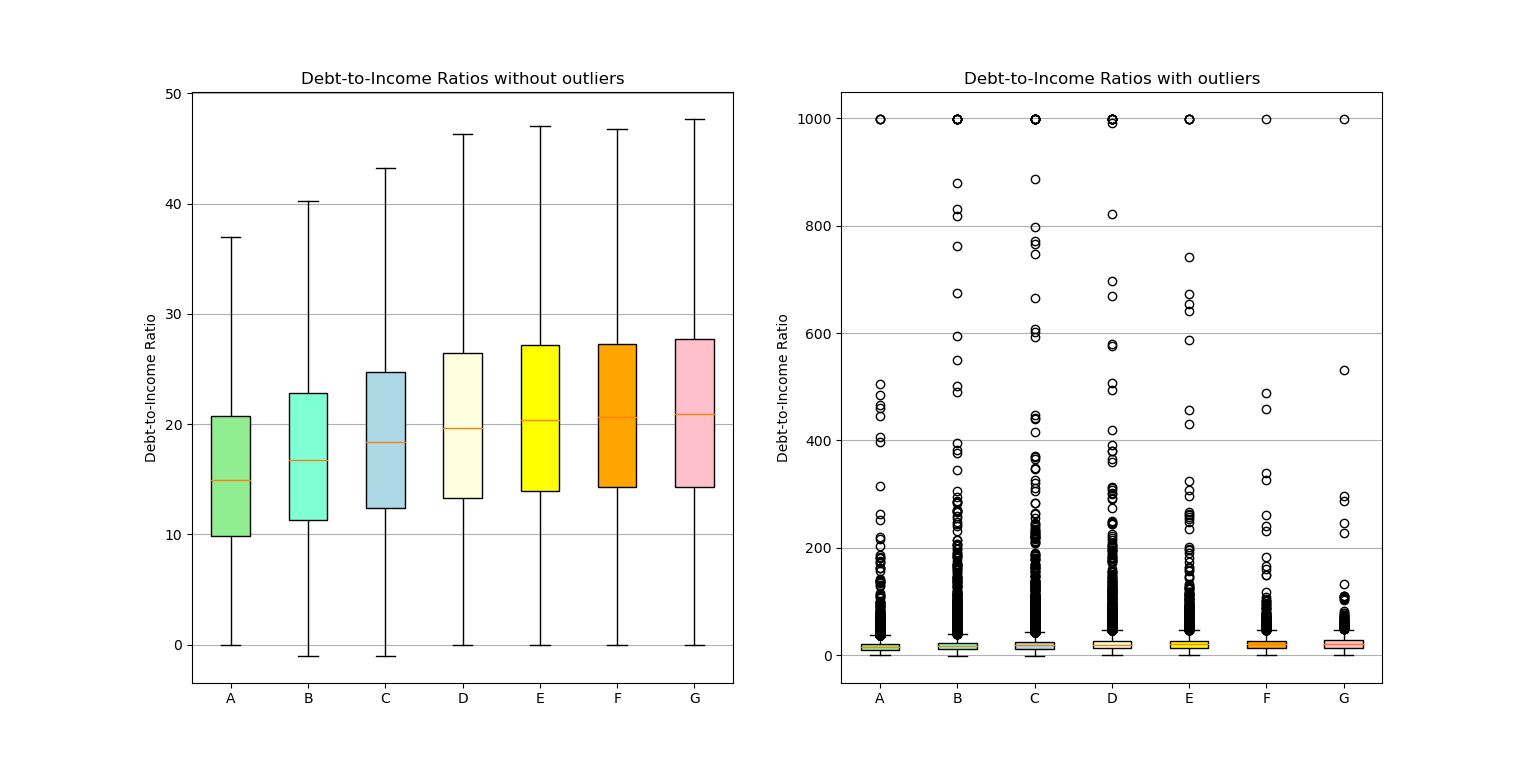
*C* 382592.0 18.912586 11.371688 -1.0 12.44 18.34 24.74 999.0

*D*  203692.0 20.171601 12.482689 0.0 13.28 19.63 26.48 999.0

*E* 95438.0 20.811170 12.804991 0.0 13.94 20.35 27.18 999.0

*F*  32783.0 21.004682 11.937965 0.0 14.27 20.65 27.28 999.0

*G*  9529.0 21.447148 15.864161 0.0 14.34 20.94 27.73 999.0



As with the relationship between debt-to-income ratios and defaulters, there is clearly a positive relationship between worse FICO score letter grades and debt-to-income ratios. As with the relationship between debt-to-income ratios and defaulters, the size is not drastically large, but still very strong.

**Chi Square test for default and FICO score grade**

Chi-Squared Test Results

*test statistic* 85980.42810158321

*p-value* 0.0

*degrees of freedom* 6

The results of the chi-squared test suggest that the relationship between defaults and FICO score grades is statistically significant.

**Z test for probability of default and debt-to-income ratio**

Z-Test Results

*test statistic* 94.8207592669868

*p-value*  0.0

The results of the z-test suggest that the relationship between defaults and debt-to-income ratios is statistically significant.

**Probability of default support vector machine model**

The model was trained on 0.1% of the dataset. As the dataset consist of over 1.3 million loans, the training data consisted of over 1,300 loans. The training data was randomly selected. After the model was trained on the random loans, it was tested on the rest of the population, which consists of over 1.3 million loans. The model used was a support vector machine algorithm.

This process was repeated 9 times, meaning that each of the nine support vector machine algorithms were trained over 9 different sets of randomly selected loans. After each algorithm was trained, new loans were randomly selected.

9 SVM models with 9 different random training samples

*SVM model iteration Accuracy*

1st 0.803966

2nd 0.801249

3rd 0.800916

4th 0.803946

5th 0.804295

6th 0.804284

7th 0.803959

8th 0.804269

9th 0.801239

Summary statistics of the 9 SVM models’ accuracy scores

*Count*  9.000000

*mean*  0.803125

*std*  0.001502

*min* 0.800916

*25%* 0.801249

*50%* 0.803959

*75%* 0.804269

*Max%*  0.804295

It is very clear that the model is 80% accurate.

The support vector machine model was from the python package scikit-learn:

sklearn.svm.SVC(kernel="linear", probability=True)

SVC(C=1.0, break\_ties=False, cache\_size=200, class\_weight=None, coef0=0.0,

decision\_function\_shape='ovr', degree=3, gamma='scale', kernel='linear',

max\_iter=-1, probability=True, random\_state=None, shrinking=True, tol=0.001,

verbose=False)