ABSTRACT

The Federal Reserve of the United States reported a drastic increase in consumer debt over the past few years, reaching $3.5 trillion in May 2015. Credit card debt accounts for only 26% of total consumer debt, however, the rest of the 74% is derived from student loans, automobile loans, mortgage etc. Lending loans has become an integral part of US consumers’ everyday life.

Have you ever wondered how lenders use various factors such as FICO score, annual income, the loan amount approved, tenure, debt-to-income ratio etc. and select your interest rates? The process, defined as ‘risk-based pricing’, uses a sophisticated algorithm that leverages different determining factors of a loan applicant. Selection of significant factors will help develop a prediction algorithm which can estimate loan interest rates based on clients’ information. On one hand, knowing the factors will help consumers and borrowers to increase their credit worthiness and place themselves in a better position to negotiate for getting a lower interest rate. On the other hand, this will help lending companies to get an immediate fixed interest rate estimation based on clients’ information.

By building various predictive models on diverse factors that might influence the interest rate set we take an attempt to answer the following problem statement:

1. Estimate if a borrower will receive a low or high interest rate.
2. The range of the interest rate based on various significant factors.

We took a two way approach in determining the interest rates, first we determine if the interest rate is going to be low or high and then in each category their range.

DATA PREPARATION & EXPLANATORY ANALYSIS

The downloaded dataset was in the form of a CSV file which was converted into SAS Datasets. This data was now taken in to SAS EM and was put under cleaning phase like: Removing the variables with large missing values, correcting the skewness in the data by data transformation etc., few variables had some missing values which were imputed using tree imputation.

Fig. 2. Shows the distribution of all the 2015 loan borrowers across all the states. We can see that majority of borrowers are from the states: California, Texas, New York and Florida respectively. Of all the borrowers there was no one from the states Idaho and Iowa.

Fig. 3. and Fig. 4. tells about the most common reasons for loan application are: Debt Consolidation i.e. About 59%, Credit Card, Home Improvement or a Major Purchase.
**METHODOLOGY (To Predict the Interest Rate Category)**

- **STEP 1:** From the data we found out that the median interest rate is around 15%, so any predicted interest rate below 15% was categorized as low interest rate and predicted interest rates above 15% as high interest rate.

- **STEP 2:** After data cleaning was done, we build several models like Logistic regression, neural network, decision tree, an ensemble model of the previous three and gradient boosting model, to best predict the category of the interest rate for all the borrowers.

- **STEP 3:** The variable selection neural network model was done via stepwise logistic regression.

- **STEP 4:** From Model Comparison we found that, Neural network best predicts the category of the interest rate, which is explained by back tracking the results into a decision tree.

**METHODOLOGY (To Predict the Interest Rate Range)**

- **STEP 1:** After the data was prepared, we built three models i.e. Linear regression, Neural Network and an ensemble of the previous two to predict the interest rate.

- **STEP 2:** The variable selection neural network model was done via stepwise linear regression.

- **STEP 3:** The Neural Network model best predicts the interest rate, as it has the lowest average squared error. From the fit statistics of the regression model we can speak about the performance of the model, it has an adjusted R squared value of 0.6217.
RESULTS (To Predict the Interest Rate Category)

From the Neural Network Model, we see that the most important variables in determining whether a person will get a low or high interest are: Term of Loan, Interest amount received till date, Principal loan amount received till date, current lower fico score followed by previous upper fico score.

We also see that the neural net model, measures the proportion of positives that are correctly identified as such, 92.48% times and measures the proportion of negatives that are correctly identified by 47.43%.

Fig. 13. Shows one of the English rules for the Prediction of low or high Interest rate. As per node 14 if a borrower has already paid interest amount up to 2485$, has a loan tenure of 36 months and as per last credit check has an upper fico score of 676.5, chances are the borrower will get a lower interest rate i.e. less than 15%.

• We set the number of branches as two and the number of levels as four, as a part of the pruning process. From the above tree we can see that the model gives four major splits. The first split occurs for the variable loan tenure, if it is for 36 months or 60 months.
• The second split occurs for previous lower fico score and current upper fico score.
• Next the model splits on the basis of total interest amount received till date followed by the total principal loan amount received till date.
• The Nodes in blue depicts that the decisions as per these nodes have high significance and the results show higher confidence.
CONCLUSION

• It is advisable for any borrower to maintain a good lower fico credit score on the last update credit history and have a high upper credit score in the current credit check in order to receive lower interest rates.

• The amount of loan amount doesn’t play a significant role while deciding the interest rate, chances are a person with high loan amount might receive a lower interest rate compared to a person who has minimal loan amount.

• Borrowers with employment more than 7 years tend to pay lower interest rates, compared to other people.

• The states from where a loan is being applied also plays a pivotal role in determining the interest rates, states like Washington, Virginia, Tennessee, Ohio, Minnesota, New York and Illinois fall in this category.

• If a borrower owns a house chances are he will pay lower interest rates compared to people who stay in a rented house.

• The Income and debt-to-income ratio of a borrower plays a significant role in determining the interest rate.

FUTURE WORK

• The scope of this model can be extended by bringing in macro economic variables like the inflation rate and the average annual income across all states, in determining the interest rates.

• Also it would be worth studying the effect of the stock exchange index like S&P 500, NASDAQ Composite, DOW Jones Industrial Average etc. effect on fluctuation in the interest rates.

ACKNOWLEDGEMENT

• I heartily thank Dr. Goutam Chakraborty, Professor, Department of Marketing and founder of SAS and OSU Data Mining Certificate program, Director MS in Business Analytics – Oklahoma State University for his constant inputs and guidance without which I couldn’t have done this research project.

REFERENCES