**BANA 6660: Project-3**

**Team Members: Avi Manawat, Sijia Li & Harshit Rai**

**Group Name: Head\_Bangers**

**Data Cleansing/Preparation and Exploratory Data Analysis:**

Modeling types were changed for different columns as per requirement. Binning was done in the columns deliquent where people with 3 or more were binned together, standing where paid and no due were binned. Data was standardized using Z-transformations for numeric data and skewness of variables were adressed using log or square root. 15 rows with missing data in Utilization column were excluded while all missing values in months were substitued with 0. An upward trend in the interest rates can be seen as it has gone up in general with increase in amount funded and requested. Average rate of interest for loans taken for a length of 60 months has been 15.07% while for 36 months it has been lower in comparison at 11.11%which majority of people prefer as their loan time. Most loans funded are in the range of $10,000 to $11,000. State of California has maximum number of loans sanctioned which almost double the number of second most popular state: New York. Mostly people with an income lower than $5000 have been funded loans. Most of the loan seekers have been employed for less than 5 years and also a major portion of loan seekers are people with 10+ years of employment. About 49% people have taken loans for debt consolidation. People with highest category of FICO score (781+) pay least interest rate on average of 8.08% while for scores between 751-780 average rates are 8.65%, for scores between 716-750 average rates are 10.41%, for scores between 681-715 average rates are 13.62% and highest for scores between 660-680 at an average of 16.43%. Majority of people taking loan have their FICO scores between 681-750, do not own a home and have 0 delinquencies for past 2 years.

**Briefly describe your final model. How did you arrive at model? What variables does it use? How good is it?**

**Final Model:**

We inititally started with different modelling techniques such as regression, boosted trees, decision tree, boosted forest, neural network, etc and got better r squares for neural network and boosted tree in intial phases of modelling. We tried nested neural networks by first trying to predict amount funded and monthly payment variables since they both were not available in the new dataset for which we had to predict so that we can use either of the two variables (or both variables) for ultimately predicting the interest rates which could have resulted in an r square value of 98.4 ideally (if predictions for amount funded and monthly payment were spot on) but the accuracy for prediction of those two variables was not good enough (about 92% only) and ultimately would predict rates with an accuracy of only about 70% r square. For verification of our nested neural network model we used the test set of the personal loans dataset to check for all the qualities by treating it as a new dataset. Then we started trying models with boosted trees using different combinations of predictor variables (based on multi coliniearity and trial and error). In our final chosen model which was a boosted tree, number of layers was taken as 195, splits per tree as 3 and the learning rate as 0.1. The variables used in the final model are: Amt Requested, Length, State, Income, Open LoC, Total LoC, Employ, Purpose, FICO, Home, Utilization, D/I, Balance ,Delinquent, Months 2. The final model created using a boosted tree resulted in a r-squared value of 0.847 for training data, r-squared value of 0.842 for validation data and r-squared value of 0.818 for test data.

