**DSC 530 Final Project Summary**

Ganesh Kale

**Abstract**

The objective of this project was to perform the different statistical techniques as part of exploratory data analysis using python as programming language. The author of textbook ‘Think Stats’ has explained different techniques with real world examples. This document provides the highlight on exploratory data analysis steps performed, outcome of the project, challenges faced, assumptions made and any data missing etc. on completing the final project for this course.

**Statistical/Hypothetical Question –** The dataset consists of credit card customers information along with customers who closed their card and who did not. The main business problem was to predict customers who are going to churn, so that appropriate offers are given to such customers, or their issues will be handled if any.

The objective of this data analysis project is to better understand the existing customers information based on different data points and identify the attributes that are making impact on customer's decision to leave the credit card company and after finalizing the attributes another goal is to predict such customers who are going to leave.

* Identify the key features/factors that driving customer to drop off.
* Identify the customers having high likelihood of attrition

The customers information such as age, income level, credit card utilization, how often they contact customer service, education etc are driving them to leave or any other factors.

**Outcome of EDA –** The above stated business problems were tackled using exploratory data analysis and regression analysis. As part of exploratory data analysis, different statistical techniques and are used to explore the data and different charts are used for visualization.

The data validated for any missing or duplicate values, the data distribution is checked for any outliers and removed them appropriately. Different distributions are checked for few key variables such as PMF, CDF, analytical distribution such as logarithmic distribution, checked for the correlation, covariance between the variables, plotted scatter charts to see correlation visually. This exploratory analysis helped to identify features that are impacting customers decision to leave or stay with bank.

Finally, Regression analysis was performed using the key features to build the model to predict the customers who will drop-off and who will not.

This exploratory analysis helped understand the different features, their relationship and impact and regression analysis helped to build prediction model more accurately.

Overall, with this analysis we were able to solve the business problem that was stated to identify features impacting customer to leave or stay and predicting churn.

**What was missed During the Analysis**? In final analysis wherein logistic regression model was used to predict churning customers received accuracy of 89% but there could be other models that might give more accuracy than logistic regression, that has been not validated in analysis.

Comparison of cdf of all variables rather just one, could have got better understanding about the impactful features.

**Were there any variables could have helped in analysis?** Since problem statement was to predict churning customers, so data such as number payments, missed payments, defaults, employment status could have helped more to handle the business problem.

**Were there any assumptions made incorrect?** One of the assumptions made about the feature Income category that lower income category customers tends to leave but this assumption was wrong when cdf of attrited and not-attrited customers are compared and did not see much difference.

**What challenges faced and what did not understand?** To perform exploratory data analysis and regression analysis, challenges faced to compute cdf and plot cdf, Calculated the cdf is easy but plotting cdf faced challenge.

Another challenge faced in regression analysis in interpreting the results and needed to take help of book and online to understand each value.

Overall, the project was quite helpful to understand the exploratory data analysis methods and how to perform them.