MGSC 410

Group 10

**Action Plan**

1. Group members explored data and compiled main problems and findings on the dataset.
   1. Data are in different currencies
   2. There are a lot of values in “Purchase Amounts” that are extremely high
   3. There are a lot of NULL values in “Purchase Amounts”column
   4. All NULL values in “Purchase Amounts” are from App platform
   5. App platform is not released until January 2019
   6. Only 30% of limited consumers renewed their subscriptions
   7. Almost all App purchase data is inaccurate
   8. Purchase amounts are not in sync with the duration
   9. You would expect a high duration with a high purchase amount
2. Data cleaning and feature transformation to create a clean dataset that can be analyzed
3. Exploratory analysis through Power BI, Tableau, Python, and Excel
4. Created scoring system to determine high-value customers for Question 1. Used logistic regression to determine most predictive traits of a high-value customer
5. Clustering analysis (K-means) was used for Question 2 to determine different segments and define certain characteristics of each segment
6. A linear probability model was used for Question 3 to determine what were the variables that would increase/decrease the probability of the consumer renewing their subscription
7. For Question 4, we created a “Quitter” variable to define customers that stopped using the product, used logistic regression to determine which traits and indicative of a subscriber quitting Rosetta stone services.
8. Additional analysis and findings from questions 1 to 4 were left for question 5
9. Gathered business insights and findings and consolidated them for PowerPoint presentation

**Group Work**

**Ishan Supanekar**

* Involved with introductory data exploration in terms of basic analysis in Excel Pivot Tables
* Involved with initial data cleaning and featurization of variables: (ex: Duration) (ex: worked with Tim to merge App Session information to Subscriber information)
* Question 1: Collaborated with JJ Goh
  + Performed research on how to implement a cost function into possible CLV formula (Rosetta Stone Financial Statements, Earnings reports)
  + Worked to help implement a ranking system through suggestion of using percentiles opposed to Inter Quartile Range
  + Involved in decision of which variables to include for scoring metric to best showcase positive engagement of customer behavior
  + Implemented Logistic Regression through Python to predict Champion Customers
  + Helped create mock Rosetta Stone promotions for Champion Customers

**Timothy Tan**

* Created PowerBI dashboard for exploratory data analysis
  + Used dax functions/queries to find insights (relationships in data)
* Cleaned App Session data by converting from vertical to horizontal data with excel functions (App Start, App Onboarding, etc.)
  + Added cleaned app session data to final csv for analysis
* Question 3: Worked solely on this question
  + Used R to create linear probability model to determine renewal users
  + Used chi-squared test to validate results
* Attempted to use ENET to use as a variable selector in determining which variables were the most important
  + Ended up using backward stepwise model instead of forward stepwise model but found the same predictors as in linear probability model so I chose not to present it

**JJ Goh**

* Involved with data cleaning in Subscriber Info.
  + Standardizing all currencies to USD
* Attempted to predict purchase amounts for missing values, especially for app users, based on duration
* Question 1: Collaborated with Ishan
  + Attempted to calculate CLV for customers in dataset
  + Researched on Rosetta 10-K to calculate cost of acquiring customer
  + Developed scoring metric for “Champion customers”
  + Performed logistic regression to determine which variables were of most importance to predict a high value customer
  + Attempted Gradient Boosting (XGB) for higher predictive accuracy

**Brian Kim**

* Involved with data cleaning
  + Dropped outliers that were skewing the numerical variables
  + Factor lumped categorical variables that had too much variance within the variable
  + Created dummy variables for categorical variables
  + Created PCA variables to create visuals and use for models
* Segmented customer to answer question 2
  + Used K-means clustering to create 4 Different segments
    - Graphed out each segments through facet wrapping and ggplot to see defining characteristics of each segments
    - Aggregated the average using group by functions to see the averages of numeric functions

**Irene Huang**

* Involved with introductory data exploration with Python, Excel, and Tableau
* Attempted to use different models to examine the data and define the outstanding variables
  + PCA, LASSO, K-Fold, Gaussian Mixture
* Question 4 to identify subscriber barriers
  + Used Excel to filter data and created variable “Quitter”
  + Used Logistic regression to determine which variables play the main role in deciding whether a subscriber is a quitter or not
* Organized first draft of action/analytical plan

**Jackie Zhao**

* Involved with data exploration through Excel, R, and Tableau
* Attempted to identify most significant variables & predict dependent variables like Purchase Amounts (prior to exempting from analysis), duration, defaulting Quitter/NonQuitter,
  + Logistic, elastic net, LASSO, forward/backward stepwise
* Question 4
  + Involved with how to define Quitters using Duration anomalies
  + Visualized findings to support Irene’s logistic regression
* Question 5
  + Explore App data and evaluate iOS performance based on time period
  + Highlight key insights with Tableau