Instructions

a. This is a team exam. You can take help from anyone, but understanding is more important
b. Please refer to the “Batch37_team_allocation-For-Viva-CUTe7302c.csv” to know your team.
c. On the submission, clearly mention team names
d. You need to submit the R code and a presentation.
e. Each step in R code should be commented
f. Your team code and PPT should be zipped and uploaded in Grader tool before the due date
g. Due Date: **4th February 2017 12:00 noon**
h. Assignment submitted in Piazza will not be accepted (before or after the dead line)
i. You along with team should prepare for a 15 min viva session. Will announce the viva schedule by Feb 2nd, 2018, Friday

Find the Customer Life time values for Retail Company

Business problem

Hopmonk currently has a million customers on their Enterprise Data Warehouse (EDW) where the data is spread across various tables. Hopmonk has implemented both Oracle and Cognos solutions designed to enable their business users to extract and analyze their data. Hopmonk feels they have locked away valuable details on consumer behavior, segmentation, demographics, and more. They believe that their path to success is through more effective use of the data by enhancing the data analysis capability in order to provide increased customer satisfaction/retention, higher profitability, extended loyalty, increase upsell/ cross-sell opportunities, and increase customer acquisition. You are required to turn data into tangible business intelligence and insight – at high speed. Using machine learning techniques analyze the data and predict the **Customer Life Time Value CLV** that will enable Hopmonk to target and acquire customers based on the net potential as profit.

About the data

The raw data spread across 7 tables. The descriptions about each table and the column names are provided in an excel file – ‘Help’. Please refer the sheet: “Tables_Information”. As we may not use all the available data to get the life time value of a customer, let us first segment the data considering the nominations (registrations) for a period of 30 days and games related data over a period of 3 months after the nomination date. Please refer the help document (sheet: Tables_Information”) to understand the exact date filter conditions to be applied.

We will not use the data in its original format and need to pull out the variables specific to our project. We have extracted 108 variables, please refer the help document sheet: "Variable_list" to see the list of variables. At high level, these variables are categorized into Recency, Frequency and Monitory (RFM). Please refer the guide document for the detailed instructions (sheet name: 'Guide_to_get_the_data_108Vars') on how to extract the data.

Answer the following questions

1. Please refer the variable list in the ''help.xlsx", sheet: Variable_list. You need to get the data for at least **30** variables from the list. A column "levelOfDifficulty" indicates how easy or difficult to get the data for each of these variables. Of your 30 variables, you should include the data for at least 10 variables from levelOfDifficulty : "difficult".
2. Construct the basic visualizations (Box plots, Bar charts and histograms)

3. Your target variable is "TotalRevenueGenerated".
   a. Build a regression model
   b. Tune the regression model by implementing
      i. Lasso and Ridge regression model
      ii. AIC models
   c. Bin the target variable into 3/5 bins and build a logistic and naïve bayes algorithms to predict the revenue bin

   d. Compute the error metrics in each of the above scenarios and explain the reason to choose a specific error metric