USE CASE STUDY REPORT

Group No.: Group 19

Student Names: Priyanka Donta and Anugraha Rose Varkey

Executive Summary: The goal of the study was to predict the customer revenue of the Gstore's dataset based on 42 predictor variables such as channel grouping, browser, pageviews, hits, country, adwordsClickInfo.page, etc. The dataset is downloaded from Kaggle.com and processed the data to find out missing values and perform other exploratory data analysis based on the variables. We have used the predictive model Light GBM, the gradient boosting algorithm which uses tree-based learning algorithm.

I. Background and Introduction

The Pareto principle (also known as the **80/20 rule**, the law of the vital few, or the principle of factor sparsity) states that, for many events, roughly **80**% of the effects come from **20**% of the causes and as per this rule it has proven that many businesses only a small percentage of customers produce most of the revenue and the marketing teams are challenged to make appropriate investments in business strategies.

- **The problem**

In this project, we have analyzed a Google Merchandise Store also known as Google Store, where Google swag is sold) customer dataset to predict the revenue per customer based on date, week-day, month, channel grouping, operating systems, devices, browsers, page views and countries.

- **The Goal of our study:**

Keeping the concept of 80/20 in mind, we have segmented, and categorized Google store's customers based on the revenue they bring to the company in order to help the company improve their promotional strategies. The conclusions and results we derived from this project can help the company take actionable operational changes and put together a better use of marketing budgets to improve their customer revenue.

- **Possible solution**

We have determined the percentage of customers that contribute to the revenue of the store using various algorithms and predictive analysis models in RStudio. Used different visualization charts and graphs to represent the data for different sections of the products.

II. Data Exploration and Visualization

#EDA after pre-processing the data-set

```
glimpse(gstoretrain)
## Observations: 300,000
## Variables: 44
## $ X                    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11…
## $ channelGrouping      <fct> Organic Search, Referral, Direct,…
## $ customDimensions     <fct> "[{'index': '4', 'value': 'EMEA'}…
```

```
## $ date                        <date> 2017-10-16, 2017-10-16, 2017-10-…
## $ fullVisitorId                <dbl> 3.162356e+18, 8.934117e+18, 7.992…
## $ hits                         <fct> "[{'hitNumber': '1', 'time': '0',…
## $ socialEngagementType             <fct> Not Socially Engaged, Not Sociall…
## $ visitId                   <int> 1508198450, 1508176307, 150820161…
## $ visitNumber                  <int> 1, 6, 1, 1, 1, 1, 1, 1, 1, 2, 1, …
## $ visitStartTime               <dttm> 2017-10-17 00:00:50, 2017-10-16 …
## $ browser                      <fct> Firefox, Chrome, Chrome, Chrome, …
## $ operatingSystem              <fct> Windows, Chrome OS, Android, Wind…
## $ isMobile                     <lgl> FALSE, FALSE, TRUE, FALSE, FALSE,…
## $ deviceCategory                 <fct> desktop, desktop, mobile, desktop…
## $ continent                    <fct> Europe, Americas, Americas, Asia,…
## $ subContinent                 <fct> Western Europe, Northern America,…
## $ country                      <fct> Germany, United States, United St…
## $ region                    <fct> NA, California, NA, NA, NA, Calif…
## $ metro                       <fct> NA, San Francisco-Oakland-San Jos…
## $ city                         <fct> NA, Cupertino, NA, NA, NA, San Fr…
## $ networkDomain                <fct> NA, NA, windjammercable.net, unkn…
## $ visits                       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
## $ hits1                        <int> 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, …
## $ pageviews                    <int> 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, …
## $ bounces                      <int> 1, NA, NA, NA, NA, NA, NA, NA, NA…
## $ newVisits                    <int> 1, NA, 1, 1, 1, 1, 1, 1, 1, NA, 1…
## $ sessionQualityDim                <int> 1, 2, 1, 1, 1, 2, 1, 1, 1, 1, 1, …
## $ timeOnSite                   <int> NA, 28, 38, 1, 52, 12, 9, 15, 34,…
## $ transactions                 <int> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ transactionRevenue               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ totalTransactionRevenue          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ campaign                     <fct> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ source                       <fct> google, sites.google.com, (direct…
## $ medium                       <fct> organic, referral, NA, organic, o…
## $ keyword                      <fct> water bottle, NA, NA, NA, NA, NA,…
## $ referralPath                 <fct> NA, /a/google.com/transportation/…
## $ isTrueDirect                 <lgl> NA, NA, TRUE, NA, NA, NA, NA, NA,…
## $ adContent                    <fct> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ campaignCode                   <fct> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ adwordsClickInfo.page            <int> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ adwordsClickInfo.slot            <fct> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ adwordsClickInfo.gclId           <fct> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ adwordsClickInfo.adNetworkType <fct> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ adwordsClickInfo.isVideoAd       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, N…
#Displaying the % of missing values in the variables
```

#Plot the count of transaction revenue between 0 and 1000

#Time series of sessions and revenues by date
#Time series of sessions and revenues by weekday
Time series of sessions and revenues by month
#Plotting graph for Channel grouping variable

#Plotting graph sessions and revenue based on device category and operating system, browser

#Plot for page views

#Plot for sessions and revenues by country

III. Data Preparation and Preprocessing

The data-set has 3000 observations 44 variables

- It contains JSON variables which are converted into character/numerical variables.
- The percentage of missing values in each variable was found out
- The variable 'StartTime' was converted into 'UTC Time zone' , parsing week month and day into one type (date) and replacing the NULL values in 'Transaction Revenue' to 0
- Data was split into training and test data

IV. Data Mining Techniques and Implementation

You are expected to explore multiple data mining techniques as appropriate to your problem. Clearly state the problem in data mining context (e.g., classification, prediction, supervised/unsupervised learning, etc.). It is desirable to have a flowchart for the entire process from data cleaning/manipulation/variable selection and transformation to specific techniques/algorithms implemented in R.

V. Performance Evaluation

The data of 3 million records was divided into training, test and validation based on. Where we set the transactionRevenue column value to be NULL so that we could predict the value after training the model based on train dataset. We then used the Light GBM model to predict the value and calculated the RMSE value which varied as follows.

```
## [1]: val's rmse:1.52945
## [101]:   val's rmse:0.560072
## [201]:   val's rmse:0.205592
## [301]:   val's rmse:0.0765531
## [401]:   val's rmse:0.0312312
## [501]:   val's rmse:0.0179661
## [601]:   val's rmse:0.0154999
## [701]:   val's rmse:0.0151778
## [801]:   val's rmse:0.0151286
```

VI. Discussion and Recommendation

We analyzed the dataset that consisted of various data formats such as JSON fields and as it has 42 variables there was a need to use the model which would boost the performance of the analysis. Hence, we used Light GBM where we had to download the package on the system from git and connect to the library. After applying the model on validation dataset, we received the RMSE value at 0.0151286 which shows there is minimum error. Hence, we recommend Light GBM to be the best model for this predictive analysis.

VII. Summary

After performing exploratory data analysis there were few things that will help to

Appendix: R Code for use case study

Please show the R code you generated for the use case study. Please do not show results here, only the code.