IEORE4523 Data Analytics Final Project

Project name : Revenue Analytics

Data sources : https://tinyurl.com/y7t4c8fg

Aries Li (jl5239) I Hung-Yu Chou (hc3035) I Jordan Hsieh (jch2211) I Wenyu Zhou (wz2445)

1. Introduction and Goal:

We analyzed a dataset from Google Merchandise Store (Google Strore). This dataset contained information about their online customers, including the type of device they were using, the sources that brought them into the website, their geographical location, the length of their browsing sessions and the amount of revenue they generated, etc.

We would like to **check if the 80/20 rule was vaild** in this dataset. The 80/20 rule generally exists in the business world, stating that 80% of the effects/revenue comes from 20% of the causes/customers. We were interested in whether the rule also applied to Google Store's dataset and what marketing strategies they could use based on the results.

From this dataset, we also wanted to generate recommendations after **analyzing the general customer type** and **figuring out what the most important features were** in terms of driving the revenue. We implemented a variety of data analytics techniques to process the data and visualized the results.

2. Techniques we used:

• **JSON normalize**:
After reading the dataset, we found there were four columns *('device', 'geoNetwork', 'totals', 'trafficSource')* with JSON, therefore we needed to flatten these columns first. For each of these four columns, we applied `json_normalize` from `pandas.io.json` package to the entire columns and created new columns according to keys in the original data. For example, the original column `‘geoNetwork’` had data `{"continent": "Asia", "subContinent": "Western…}`, we created new columns named `geoNetwork.continent` and `geoNetwork.subContinent` and dropped the original column for later calculation.
• **Numpy and Pandas**:
`Pandas` was mostly used for grouping when we were trying to do feature analysis on categorical features. After grouping features, we applied `DataFrameGroupBy.agg` function to aggregate by total number and mean.

• **Data visualization**:
`Matplotlib` and `Seaborn` are two libraries that we used to interpret our data before further technical analysis. For example, using `df.groupby()`, `sns.barplot()` and `plt.subplots()` with the `sharey` arguement, we could make a horizontal comparison on how different subcategories varied among users number, revenue contribution and the mean value given a specific category. We used this skill to analyze users' behavior on four columns *('device.browser',*

'device.deviceCategory', 'isMobile', 'OS').

- **Datetime conversion**:
In order to perform time series analysis and visualize datetime data clearly, we used `datetime.date()` and `df.apply()` to transform cells in our dataframe from the original type `yyyymmdd` to `yyyy-mm-dd` so that we could group by date in our analysis.

- **Data processing**:
We handled the missing data in the dataset and changed the data format to fit our use. We used the **mode** as the value to fit in the missing data for every column with numerical data and filled missing values in categorical columns with a keyword `unknown`. We hypothesized that most visitors would have the similar behavior/value. In addition, for usage in regression model, we converted all the numbers in `string` format to`float` and used `preprocessing` from `sklearn` to encode every categorical column.

- **Machine learning**:
`LightGBM`: A fast, distributed, high-performance gradient boosting framework based on decision tree algorithms. After importing `lightgbm`, we used the default regression in `LightGBM` to predict the revenue given some chosen features. We divided the training set into two parts: **training** and **validation**.
Since the time span for the data was one year, we set the last two months as the validation dataset. Finally, we applied our model to the test dataset. In order to get an overview of the feature importance, we used `plot_importance()` to get the importance plot.

3. Strategic Advices for Google Store:

- **The 80/20 rule is confirmed for Google Store.**
We advise that the store set **2,000 USD** as a boundary to separate their customers into two groups. For the "80 group," Google could consider lowering the price to an extent where a much larger amount of non-zero revenue could be attracted. As for the "20 group," Google could focus on developing higher quality products and setting higher prices in order to benefit from their spending power.

- **Top 3 feature importance: pageviews, visit start time, and hits.**
The feature importance analysis shows that the number of page views and the number of hits (i.e. mouse clicks made during browsing) had a higher association with generating revenue. Therefore, we suggest that Google focus on customizing their website in order to encourage customers to browse more. However, as shown in our visualization tool, an interval between **[10, 20]** for the number of page views and hits worked the best in terms of generating revenue; a higher number was associated with a decrease in revenue.
The visit start time indicates that Google could focus on these time sessions when customers were more likely to make purchases. For example, Google could send out advertisement emails during these time sessions to be more effective.

- **Potential for Windows users and Asia & Europe Markets.**
In terms of the user types of their customers, Google should try to increase the conversion rate of Windows users because they were a large group but contributed little in revenue. Google should also consider expanding to the Europe and Asia markets in an aggressive way.