Udacity Capstone Project Proposal

Background

The 80/20 rule states that 80% of the effort of anything is usually generated by 20% of the effort. In business this can be translated as 80% of a company's revenue being created by 20% of its customers. Machine learning is broadly used in the marketing industry to help companies to better understand their customers segmentations in order to design marketing campaigns that result in high revenues.

In this project, I am going to analyze a Google Merchandise Store to predict revenue per customer. Hopefully, this will find the high-value customers for those companies in the Google Merchandise store and thus to help them to better target and maximize the revenue of their marketing expenses.

Problem Statement

1. What will be the tracking metric?
2. Customers have different needs and their own profiles. What will be the best features to describe them?
3. Companies invest in customers to generate revenues. Can I help these companies identify customers' behavior patterns, customers' profiles thus they can act accordingly to maximize their revenues?

Datasets and Inputs

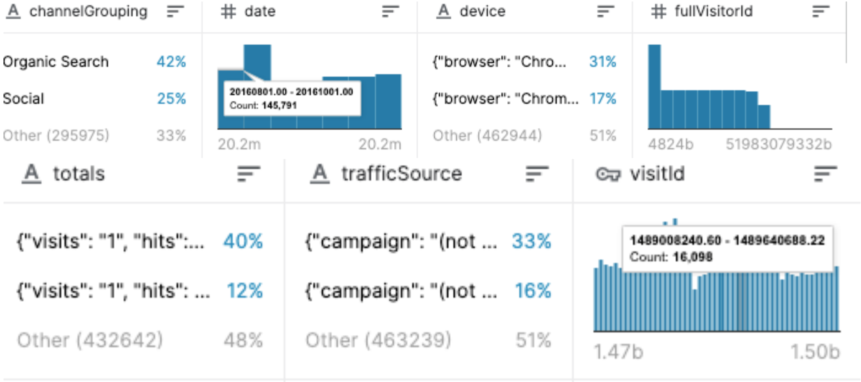
I will use the data set provided by Kaggle's Google Analytics Customer Revenue Prediction Competition. The training data contains user transactions from August 1st, 2016 to April 30th, 2018. The test data set contains user transactions from May 1st,2018 to Oct 15th, 2018.

Data fields include,

- fullVisitorId- A unique identifier for each user of the Google Merchandise Store.
- channelGrouping - The channel via which the user came to the Store.
- date - The date on which the user visited the Store.
- device - The specifications for the device used to access the Store.
- geoNetwork - This section contains information about the geography of the user.
- socialEngagementType - Engagement type, either "Socially Engaged" or "Not Socially Engaged".
- totals - This section contains aggregate values across the session.
- trafficSource - This section contains information about the Traffic Source from which the session originated.
- visitId - An identifier for this session. This is part of the value usually stored as the _utmb cookie. This is only unique to the user. For a completely unique ID, you should use a combination of fullVisitorId and visitId.
- visitNumber - The session number for this user. If this is the first session, then this is set to 1.
- visitStartTime - The timestamp (expressed as POSIX time).
- hits - This row and nested fields are populated for any and all types of hits. Provides a record of all page visits.
- customDimensions - This section contains any user-level or session-level custom dimensions that are set for a session. This is a repeated field and has an entry for each dimension that is set.
- totals - This set of columns mostly includes high-level aggregate data.

Data Inputs on training set include

- The data is quiet clean. There are no null values in the training data set.



channelGrouping
Organic Search 42%
Social 25%
Other (295975) 33%

date

device
{"browser": "Chro... 31%
{"browser": "Chrom... 17%
Other (462944) 51%

fullVisitorId

totals
{"visits": "1", "hits":... 40%
{"visits": "1", "hits": ... 12%
Other (432642) 48%

trafficSource
{"campaign": "(not ... 33%
{"campaign": "(not ... 16%
Other (463239) 51%

visitId

Solution Statement

I am going to predict the log of the sum of transactions per user from December 1st 2018 to January 21st, 2019 using XGBoost model by minimizing the root mean squared error (RMSE).

Benchmark Model

I am going to use a Linear Regression model or a past solution to the problem on Kaggle leaderboard as a benchmark.

Evaluation Metrics

I will use RMSE to evaluate the model.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where y hat is the natural log of the predicted revenue for a customer and y is the natural log of the actual summed revenue value plus one.

Project Design

1. Select Metrics – I am going to use transaction per customer
2. Customer Segmentation – I am going to use Recency- Frequency- Monetary (RFM) value to create customer segments. Theoretically, I will have segments including, low value, mid value and high value customers. I plan to apply K-means clustering for the calculation.
3. Customer Lifetime Value Prediction -I will select a time window to calculate the life time Value using the features derived in step 2.
4. Customer Transactions Prediction – I will apply EDA on customers by segmentation and life time value calculated in step 2 and step 3 to select the best features. Then I will use XGBoost to predict the transactions per customer.

Reference

https://www.kaggle.com/c/ga-customer-revenue-prediction/data?select=train_v2.csv