Capstone 2: Project Proposal

Problem Statement

- Whether customers will buy or not based on previous searches on google ecommerce platform.

Context

- The input dataset contains website users data (Google Analytics data) from August 2016 to August 2017. The sample dataset contains obfuscated Google Analytics 360 data from the Google Merchandise Store, a real ecommerce store. The Google Merchandise Store sells Google branded merchandise. The data is typical of what one would see for an ecommerce website. It includes the following kinds of information.

Criteria For Success

- Build a predictive model to calculate whether customers will transc with 80% > accuracy rate.

Scope of solution space

- Model will be reviewed and approve by Springboard Mentor

Constraints

- We have a large dataset consisting of 903656 rows spanning over 1 year, if a customer has transc before or after the data range that will not be captured and might create a less accurate model.
- Data is exported as a JSON file and will need to be parsed before cleaning.
- BigQuery only allows export of 1GB data limit, we would have to divide the dataset under the given constraints and export. Below Query was written for partition of dataset by each month and exported to google drive.

```
SELECT * ,
FROM `bigquery-public-data.google_analytics_sample.ga_sessions*`
AS BQ
WHERE BQ.date > '20170701' AND BQ.date < '20170731'
ORDER BY BQ.date DESC
```

Stakeholders

- Springboard Mentor

Data sources

- Big Query: Google Analytics 360 Ecommerce data