Google Analytics Transaction Prediction

1. Introduction to Data

The Google Merchandise Store, or GStore, is where Google swag is sold. This data set involves customer data information for previous users and customers. The data includes device information, geographical information, trafficking, timing, and more.

2. Problem to Solve

A very small percentage of users make purchases and produce the revenue for a business. The problem is how to appropriately budget and allocate money for marketing purposes. Through the course of this project, the data will be analyzed to predicted whether a user will complete a transaction. If predicted accurately, marketing can focus budgets on the features of importance in hopes of increasing customers and revenue.

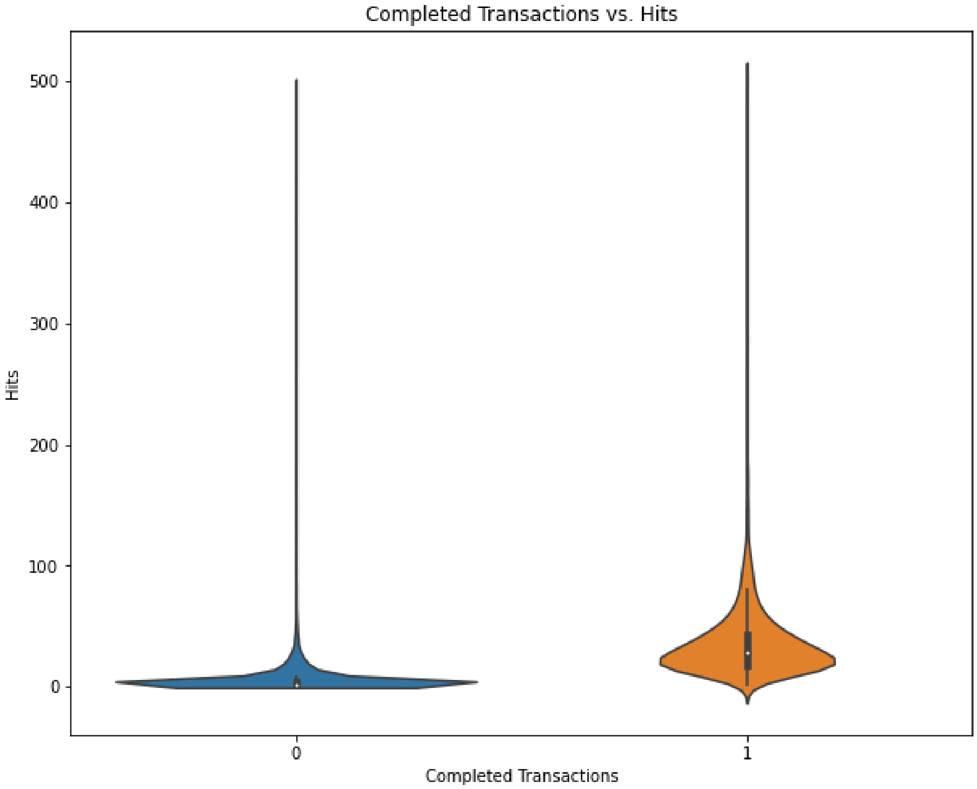
3. Data Cleaning

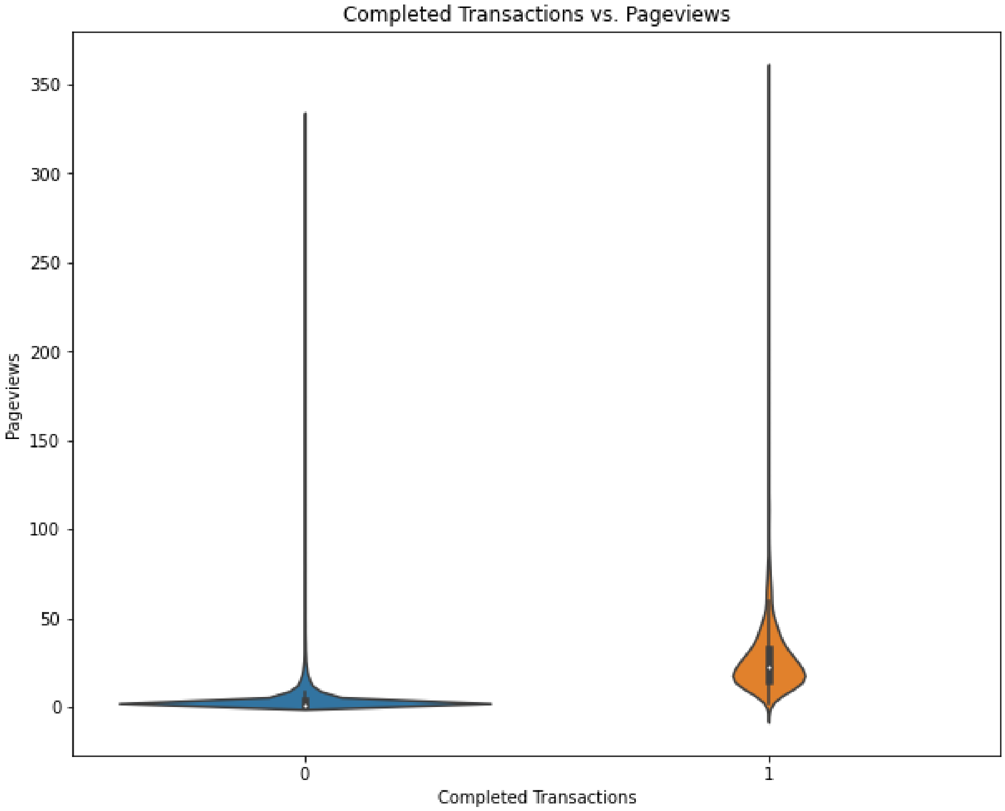
In the data cleaning process, I followed the following steps:

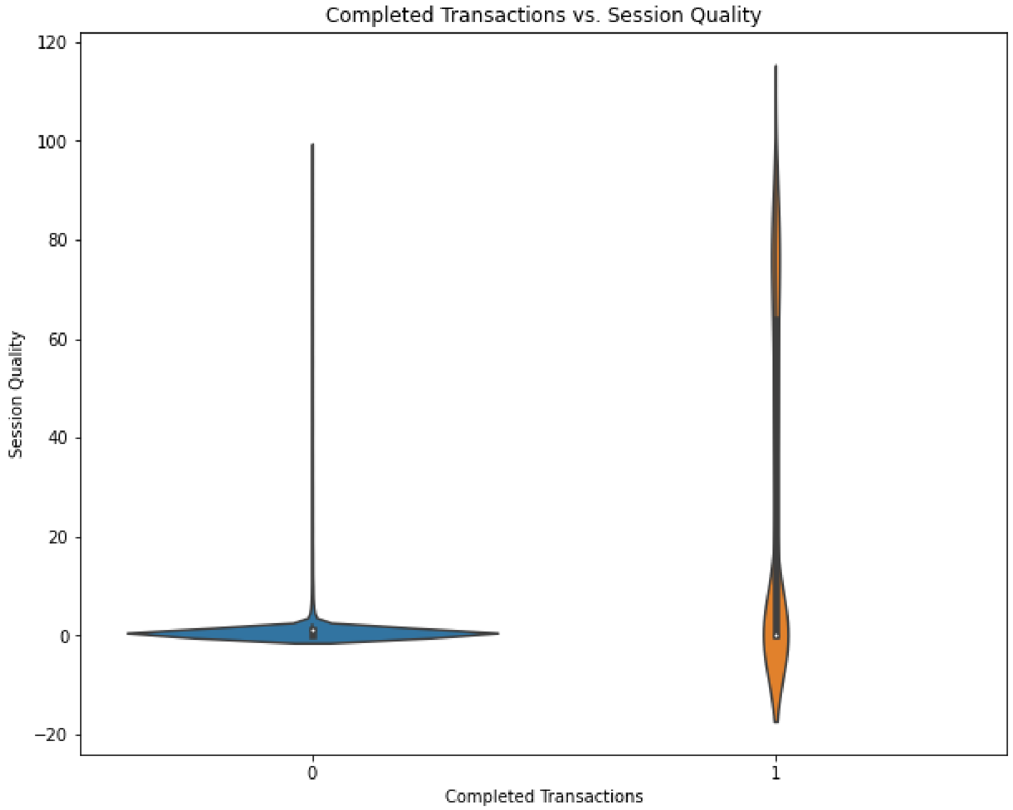
1. There were 4 json columns. The first step I took was to use json normalize to expanding those columns.
2. Some columns given did not have any information in this data set. Those columns were dropped.
3. Formatting issues were address, such as multiple notations for 'Unknown'.
4. Day, month, year, and weekday columns were created.
5. A previous customer column was created based on a customer's previous transactions.
6. Our target feature was created. We are predicting if a user will complete a transaction.

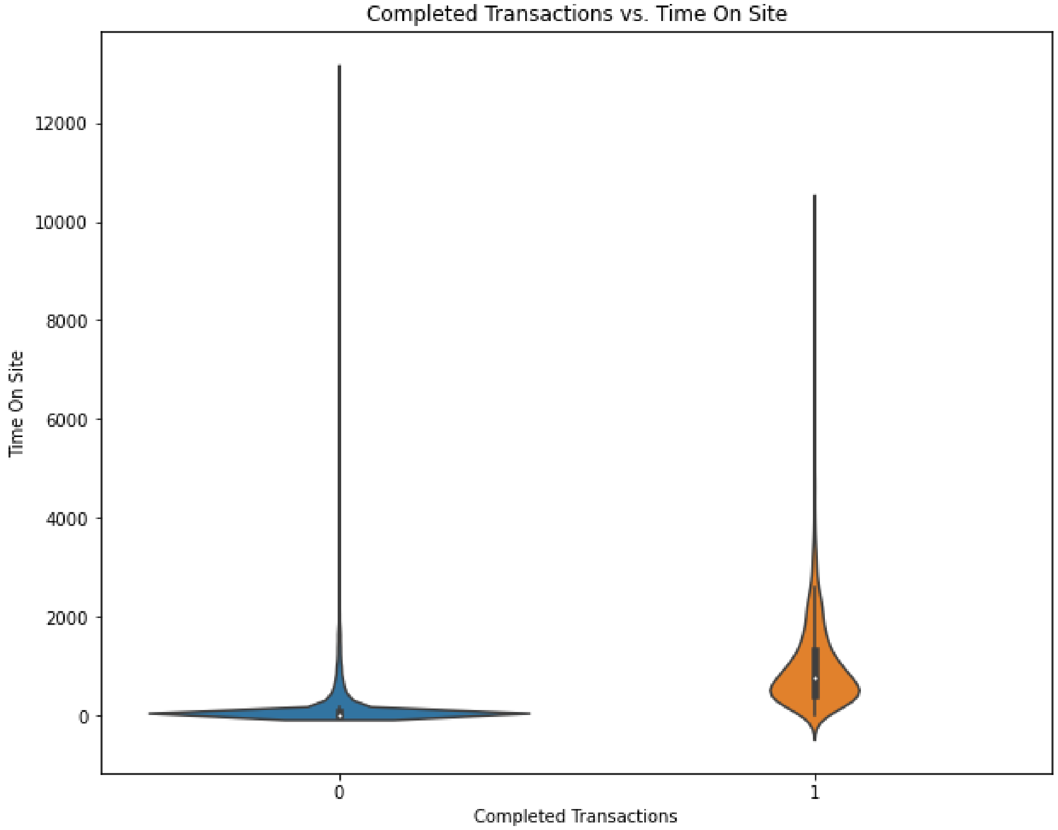
4. Exploratory Data Analysis

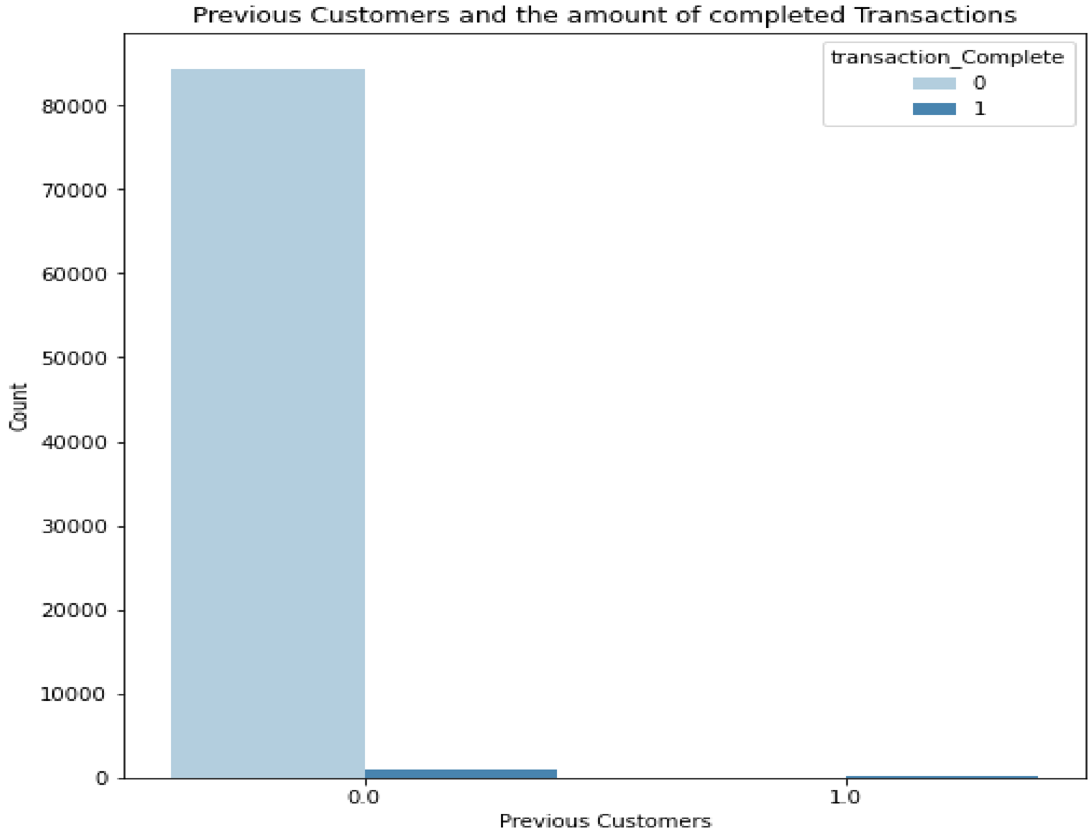
In the process of exploring the data, some of the features stood out as being possibly important to whether a user completes a transaction. The visuals below explore those features compared to our target variable of transactions complete.


Completed Transactions vs. Hits
Hits
500
400
300
200
100
0
0
1
Completed Transactions

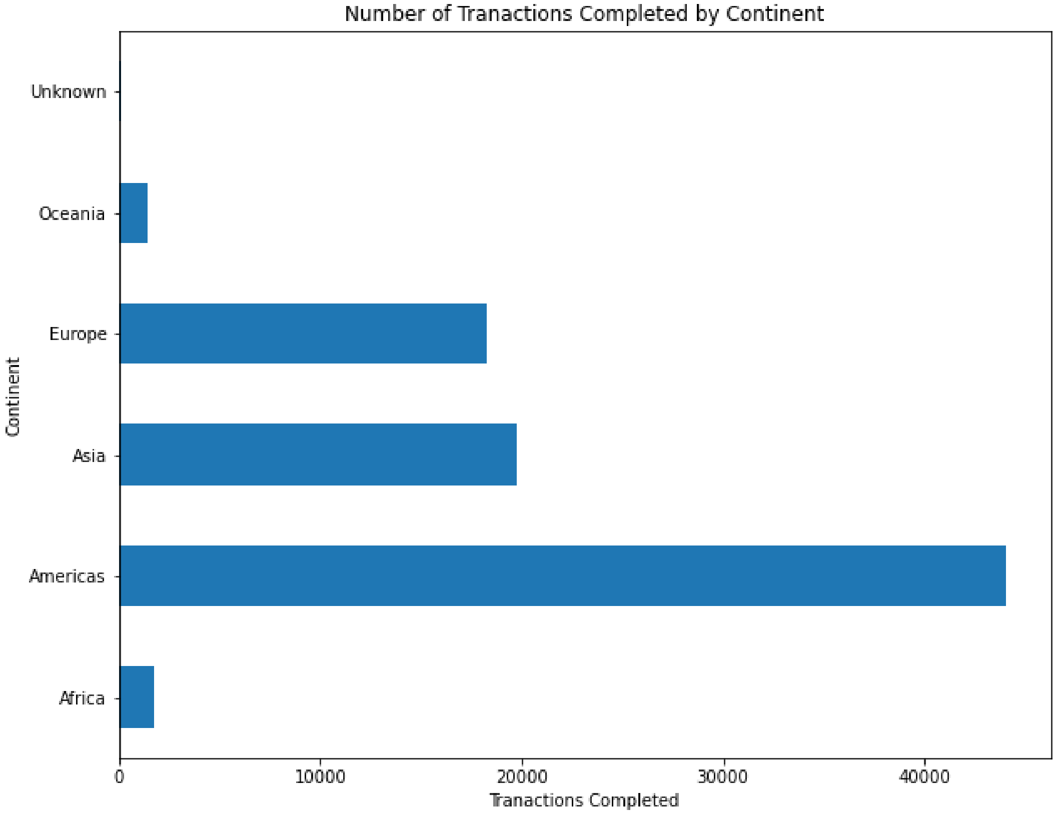

Completed Transactions vs. Pageviews
Pageviews
350
300
250
200
150
100
50
0
0
1
Completed Transactions


Completed Transactions vs. Session Quality
Session Quality
120
100
80
60
40
20
0
−20
0
1
Completed Transactions


Completed Transactions vs. Time On Site
Time On Site
12000
10000
8000
6000
4000
2000
0
0
1
Completed Transactions



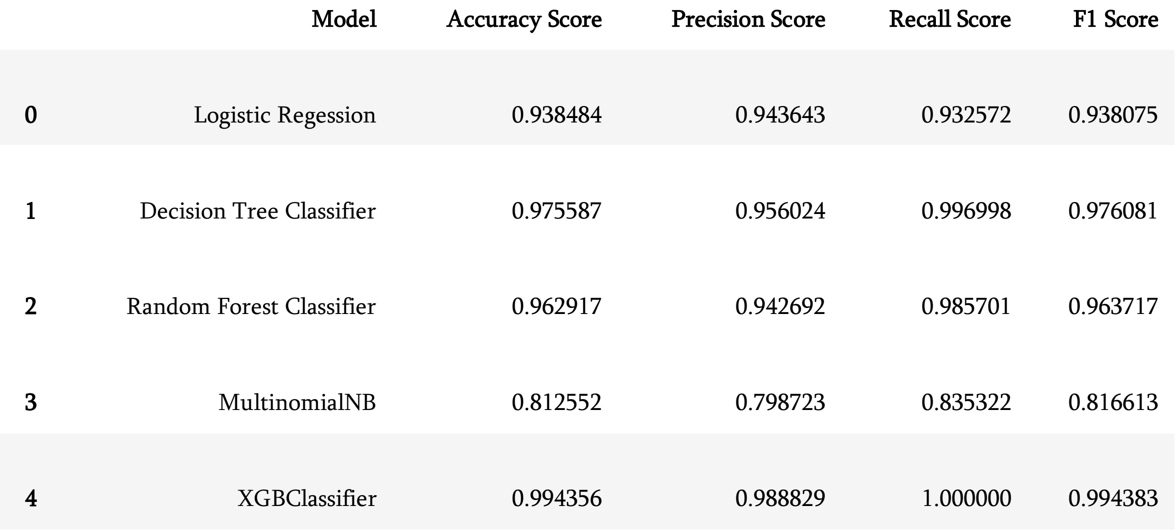
Previous Customers and the amount of completed Transactions
transaction_Complete
0
1
80000
70000
60000
50000
40000
30000
20000
10000
0
Count
0.0
1.0
Previous Customers



Number of Tranactions Completed by Continent
Unknown
Oceania
Europe
Asia
Americas
Africa
Continent
0
10000
20000
30000
40000
Tranactions Completed

5. Modeling Approach

Before I began modeling, I created a label encoder for categorical data. Then, I had to address the issue of vastly different amounts of transactions completed or not completed. To correct this, I oversampled the minority, transactions completed. Next, I created 5 different models with hyperparameter tuning. The following chart is the resulting scores from each of the models.



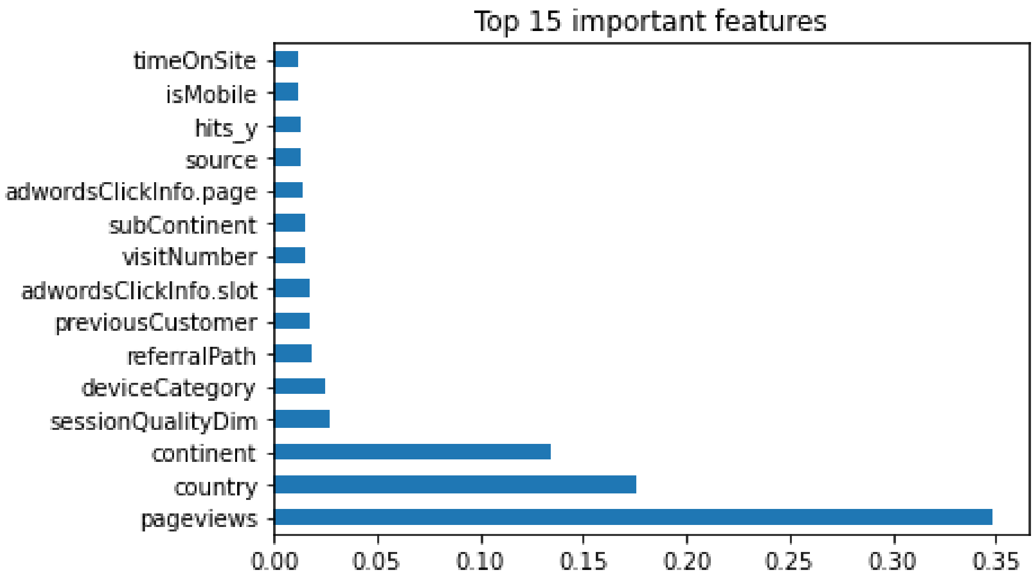
	Model	Accuracy Score	Precision Score	Recall Score	F1 Score
0	Logistic Regession	0.938484	0.943643	0.932572	0.938075
1	Decision Tree Classifier	0.975587	0.956024	0.996998	0.976081
2	Random Forest Classifier	0.962917	0.942692	0.985701	0.963717
3	MultinomialNB	0.812552	0.798723	0.835322	0.816613
4	XGBClassifier	0.994356	0.988829	1.000000	0.994383

6. Findings

The following model and hyperparameters were chosen to be the best model.

```
XGBClassifier(base_score=0.5, booster='gbtree',  class_weight=None,colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=0.8, gamma=0.5, gpu_id=-1,
              importance_type='gain', interaction_constraints='', learning_rate=0.1,
              max_delta_step=0, max_depth=20, min_child_weight=1, missing=nan,
              monotone_constraints='()', n_estimators=100, n_jobs=8,
              num_parallel_tree=1, random_state=0, reg_alpha=0, reg_lambda=1,
              scale_pos_weight=1, subsample=0.8, tree_method='exact',
              validate_parameters=1, verbosity=None)
```

The features of importance were investigated. Here is the result:



7. Use of Findings

Based on the finding, the top five features deviceCategory, sessionQualityDim, Continent, Country, and pageviews. Marketing teams should focus their budgets to increase customers based on these features.

8. Further Research

With further research, I would investigate to results of combining some of the columns. This could potentially create some new features of importance and help make more accurate models.