Zach Timmons
https://www.linkedin.com/in/ztimmons/
https://github.com/ztimmons

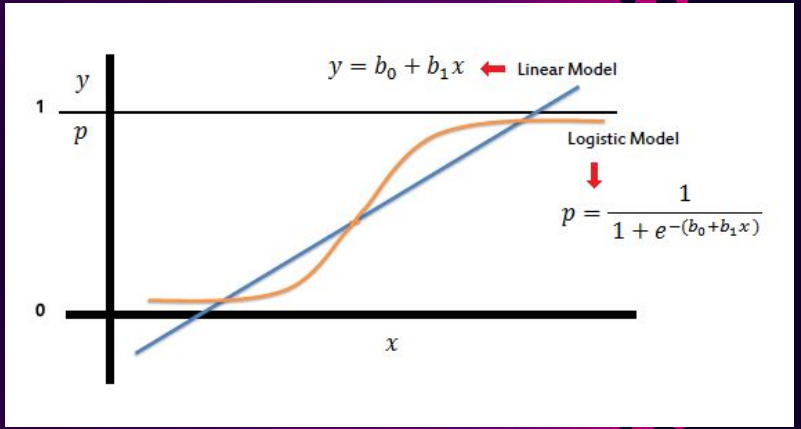
Analytics Customer Revenue Classification

Summary

Project Overview

Binomial Classification - Will a customer spend money or not?

Binomial Logistic Regression



Encoding and Transformation Methods for Data

- Min Max Scaling
- Target Encoding
- Weight of Evidence Encoding

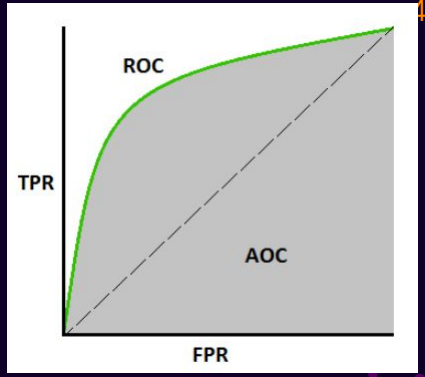
$$WOE = \ln\left(\frac{\text{Distribution of Goods}}{\text{Distribution of Bads}}\right)$$

WOE Calculation

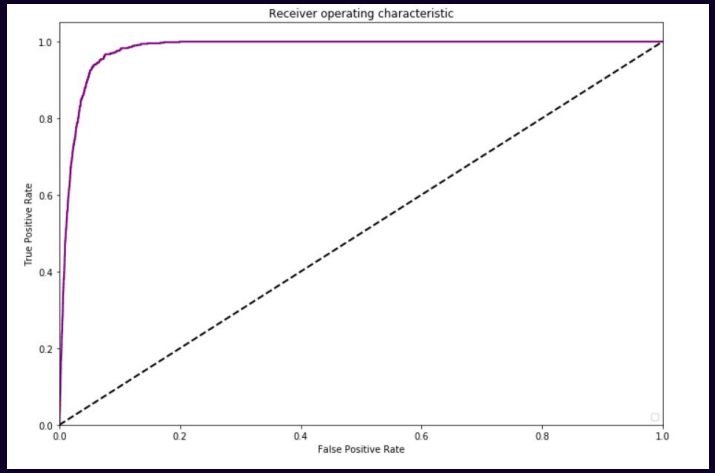
Metrics

Accuracy

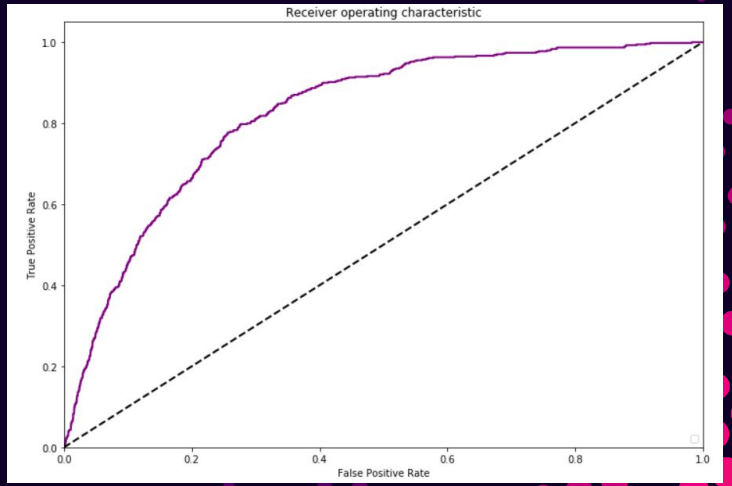
ROC-AUC Score



(Initial Logistic Regression ROC-AUC Curve)



(Second Tier Logistic Regression with WOE)



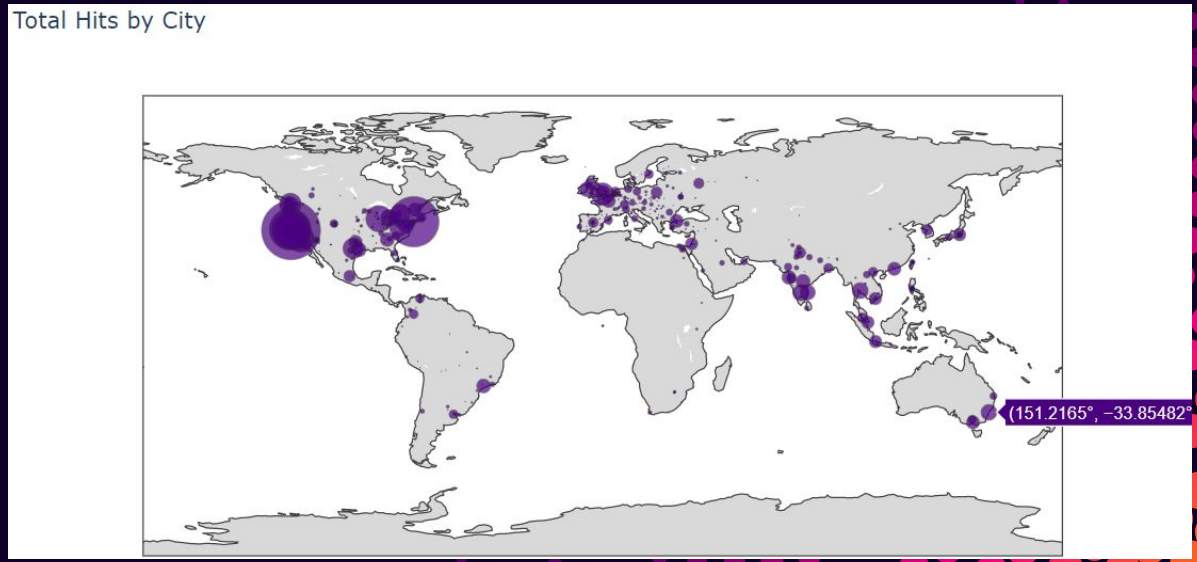
Exploratory Data Analysis

Data Cleaning

Downsampling Data - data consisted of 1.7 million rows and 30 columns for a 24GB CSV file. It was loaded in chunks with a chunk loader function in Jupyter Notebook and then downsampled to 300,000 rows.

Data cleaning and feature creation - week, day, month, year, binary 'has_revenue' features were created for the dataset from datetime and total revenue columns, as well as latitude and longitude values from cities being pulled using PyGeo API.

Visualizing Feature Relationships



Feature Permutation Importance

Permutation Importance defines a decrease in a model score when a single feature value is randomly shuffled. If a random shuffling of a feature is associated with a significant decrease in the model's value, then that feature is determined as more important than a feature that shows a less significant decrease in value upon shuffling.

The features most associated with model performance decrease in a logistic regression model are

-Total Page Views

-Total Hits

-Total New Visits

-Location (This makes sense as the vast majority of online purchases from the Google Merchandise Store came from the Americas).

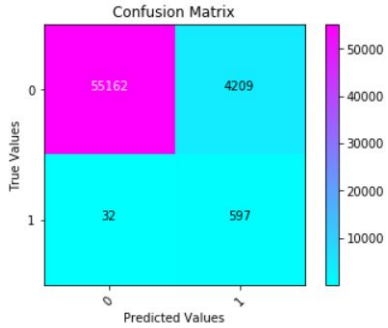
Charts based on Total Page Views and Total Hits illustrate significant relationships between not only our binary classification of whether or not customers will purchase a product from the Google Merchandise Store, but also how much they will spend.

Weight	Feature
0.0890 ± 0.0037	totals_pageviews
0.0696 ± 0.0073	totals_hits
0.0131 ± 0.0032	totals_newVisits
0.0099 ± 0.0039	geoNetwork_subContinent
0.0080 ± 0.0008	geoNetwork_continent
0.0065 ± 0.0051	year
0.0046 ± 0.0054	geoNetwork_networkDomain
0.0042 ± 0.0026	visitStartTime
0.0037 ± 0.0016	device_isMobile
0.0033 ± 0.0053	geoNetwork_city
0.0022 ± 0.0048	trafficSource_source
0.0017 ± 0.0037	device_deviceCategory
0.0010 ± 0.0021	trafficSource_keyword
0.0009 ± 0.0019	trafficSource_medium
0.0002 ± 0.0004	device_browser
0.0001 ± 0.0013	visitNumber
0.0000 ± 0.0002	trafficSource_campaign
0 ± 0.0000	totals_bounces
-0.0001 ± 0.0002	weekday
-0.0001 ± 0.0015	trafficSource_referralPath
… 6 more …	

Other Classifiers

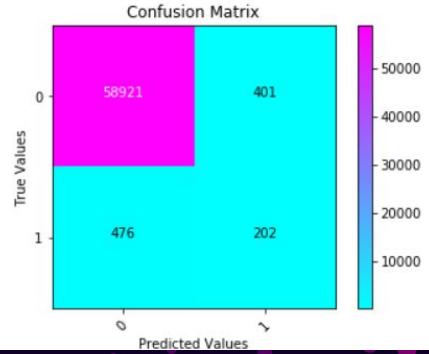
V-Treated Logistic Regression

```
Accuracy score: 0.929
ROC_AUC score: 0.939
Cross Validation Scores are: [0.97564938 0.97810696 0.97677371 0.97877244 0.97829593]
```



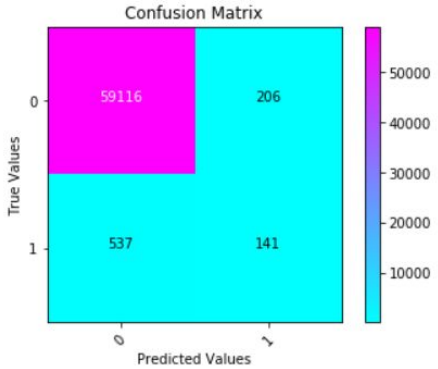
Decision Trees

```
model score: 0.985
ROC_AUC score: 0.646
Cross Validation Scores are: [0.63635043 0.63274357 0.65606028 0.64349554 0.65901188]
```



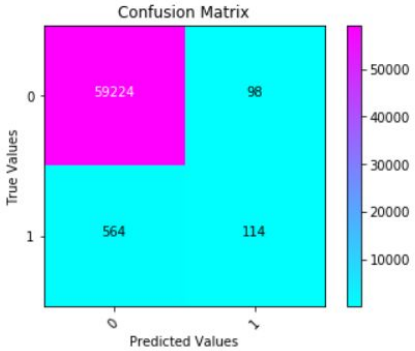
K Nearest Neighbors

```
model score: 0.988
ROC_AUC score: 0.602
Cross Validation Scores are: [0.73923959 0.78003591 0.74532605 0.74822874 0.75606372]
```



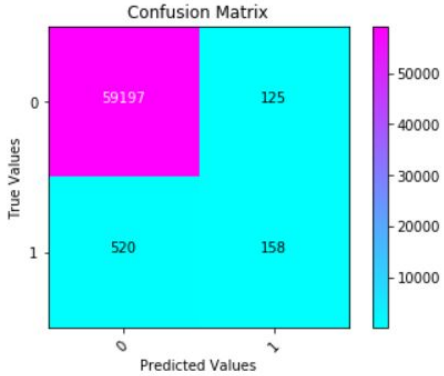
Random Forests

```
model score: 0.989
ROC_AUC score: 0.583
Cross Validation Scores are: [0.87497105 0.88239128 0.87739264 0.87950099 0.88413175]
```



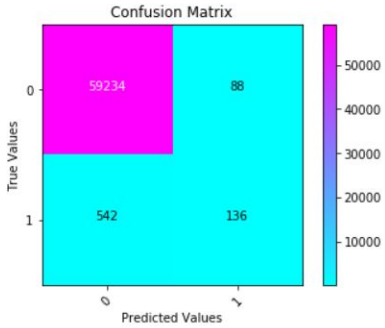
Adaptive Boosting

```
model score: 0.989
ROC_AUC score: 0.615
Cross Validation Scores are: [0.97109313 0.96869093 0.97358831 0.9694778  0.97075183]
```



Gradient Boosting

```
model score: 0.990
ROC_AUC score: 0.600
Cross Validation Scores are: [0.97959428 0.98341227 0.98363414 0.98605713 0.98626119]
```



Two Tiered Model

Round 1 Models

-Categorical Variables - Target Encoding
-Numerical Variables - Min Max Scaler
-Logistic Regression, V-Treated Logistic Regression, K Nearest Neighbors, Decision Trees, Random Forests, Adaptive Boosting, Gradient Boosting
-Logistic Regression - Best Model

Round 2 Models

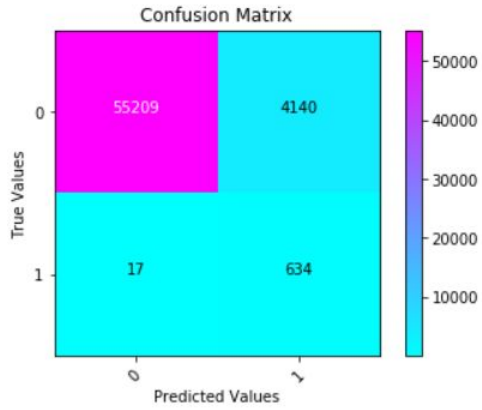
-DataFrames sliced from Predicted Customers
-Logistic Regression - Target Encoding & Weight of Evidence Encoding
-Second Model - K Nearest Neighbors, Decision Trees, Random Forests, Adaptive Boosting, and Gradient Boosting.
-Logistic Regression with Weight of Evidence Encoding was still the best performing model with the stacked model as well, although Logistic Regression with Target was almost identical.

These Models may represent a two-tiered approach to marking with customers. You could apply one advertising budget to customers who met the initial binary prediction requirement, and filter a second advertising budget to customers who met the second requirement.

Results

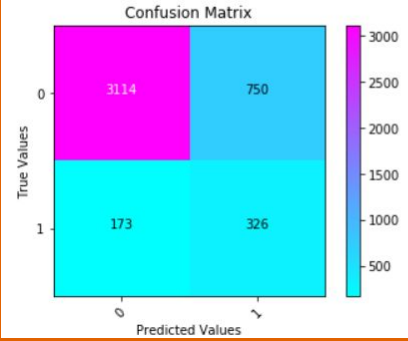
Tier 1 - Simple Logistic Regression - Target Encoding

```
Accuracy score: 0.931
ROC_AUC score: 0.952
Cross Validation Scores are: [0.98097657 0.98152717 0.98180791 0.98297627 0.98265345]
```



Tier 2 - Simple Logistic Regression - Weight of Evidence Encoding

```
Accuracy score: 0.788
ROC_AUC score: 0.730
Cross Validation Scores are: [0.79396032 0.79048946 0.77974493 0.79619704 0.79511703]
```

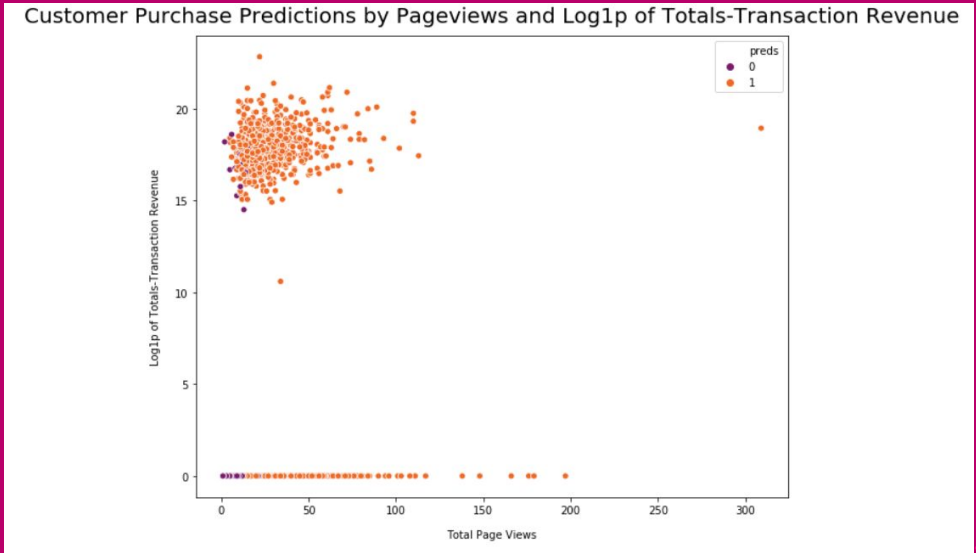


Conclusions

Logistic Regression is an effective model for our predictions that leaves few paying customers unadvertised in our marketing strategy

A multi-tiered approach to our marketing strategy may be the most beneficial to generating revenue.

There is a strong correlation between the number of times someone visits the Google Merchandise Store Website and their likelihood to spend money.



Future Improvements

V-Treat on Gradient and Ada Boosting

SHAP Modeling - These are only to explain Ensemble tree methods but they could prove Useful in explaining feature importance for Future models.

Feature Engineering - New models based on user's pageviews and hits

Applying more time-series features to my model to adjust predictions based on yearly or seasonal trends