GOOGLE MERCHANDISE STORE

ABOUT GOOGLE ANALYTICS

Google Analytics is a web analytics service offered by Google that tracks and reports traffic, currently as a platform inside the Google Marketing Platform brand. Google launched the service in November 2005 after acquiring Urchin.

EVOLUTION OF GOOGLE ANALYTICS TRACKING

2005

Google uses acquisition of Urchin Sofware to power Google Analytics

1

2007

Custom event tracking

2

2009

Faster and more accurate

3

2014

UID
Enhanced eComm

4

And still counting!

01
PROBLEM STATEMENT
What are we trying to predict

02
EXPLORATORY DATA ANALYSIS
What are the sources of traffic

03
MODELLING

04
SCORING
F1, Accuracy and AUC

05
CONCLUSION

Problem Statement

The 80/20 rule has proven true for many businesses–only a small percentage of customers produce most of the revenue. As such, marketing teams are challenged to make appropriate investments in promotional strategies. For this project,we are challenged to analyze a Google Merchandise Store (also known as GStore, where Google swag is sold) customer dataset and identify the best model to predict the probability of a session being revenue-generating.

The f1-score, accuracy and auc score from different machine-learning models will be compared and used to evaluate the best model for prediction.

DataSet

Data Shape
903,653 rows, 12 columns

Data Type Issue
includes JSON columns which required flattening

Final Data Shape
903,653 rows, 55 columns

GASP!!
That is a huge dataset!

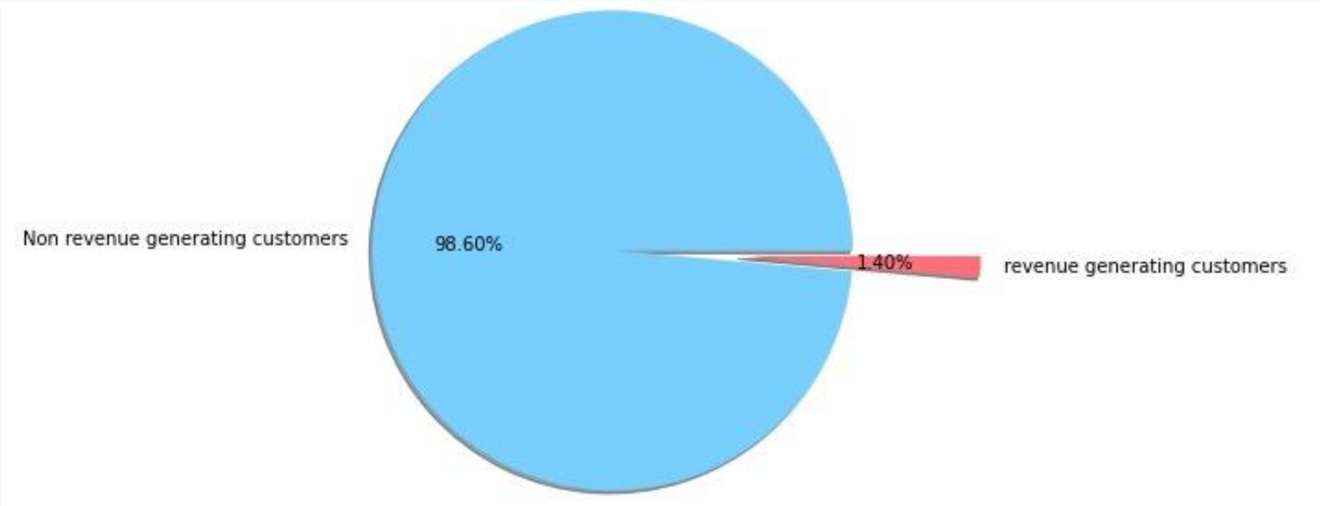
Revenue Generating Customers

Customers

Non-revenue customers:
704,171
Revenue-generating customers:
9996

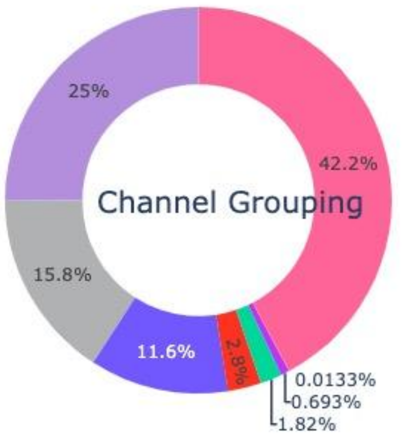
Percentage

1.3997%



Channel Grouping

3) Direct
Refers to traffic by typing the URL into the browser or through bookmarks



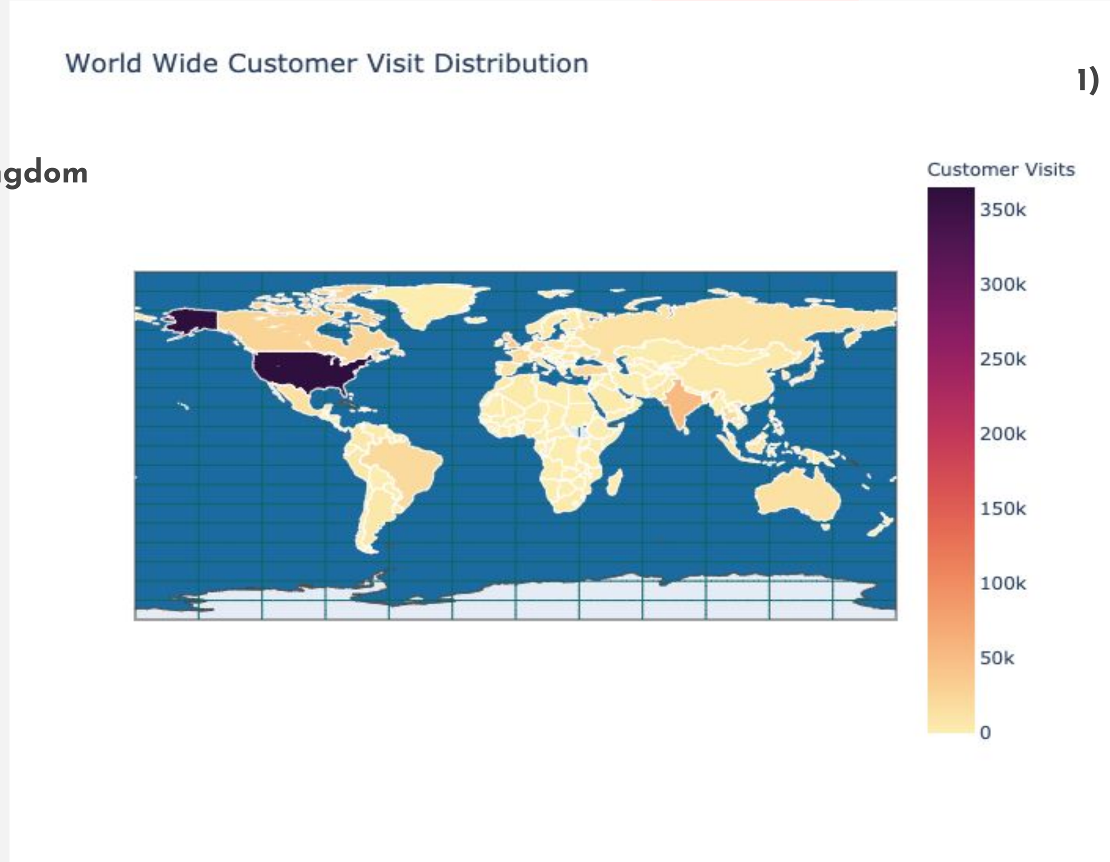
1) Organic Search
Based on unpaid ranking. Accounts for the highest visits to the website

2) Social
Refers to traffic from social networks and social media platforms

Customer Visits Distribution

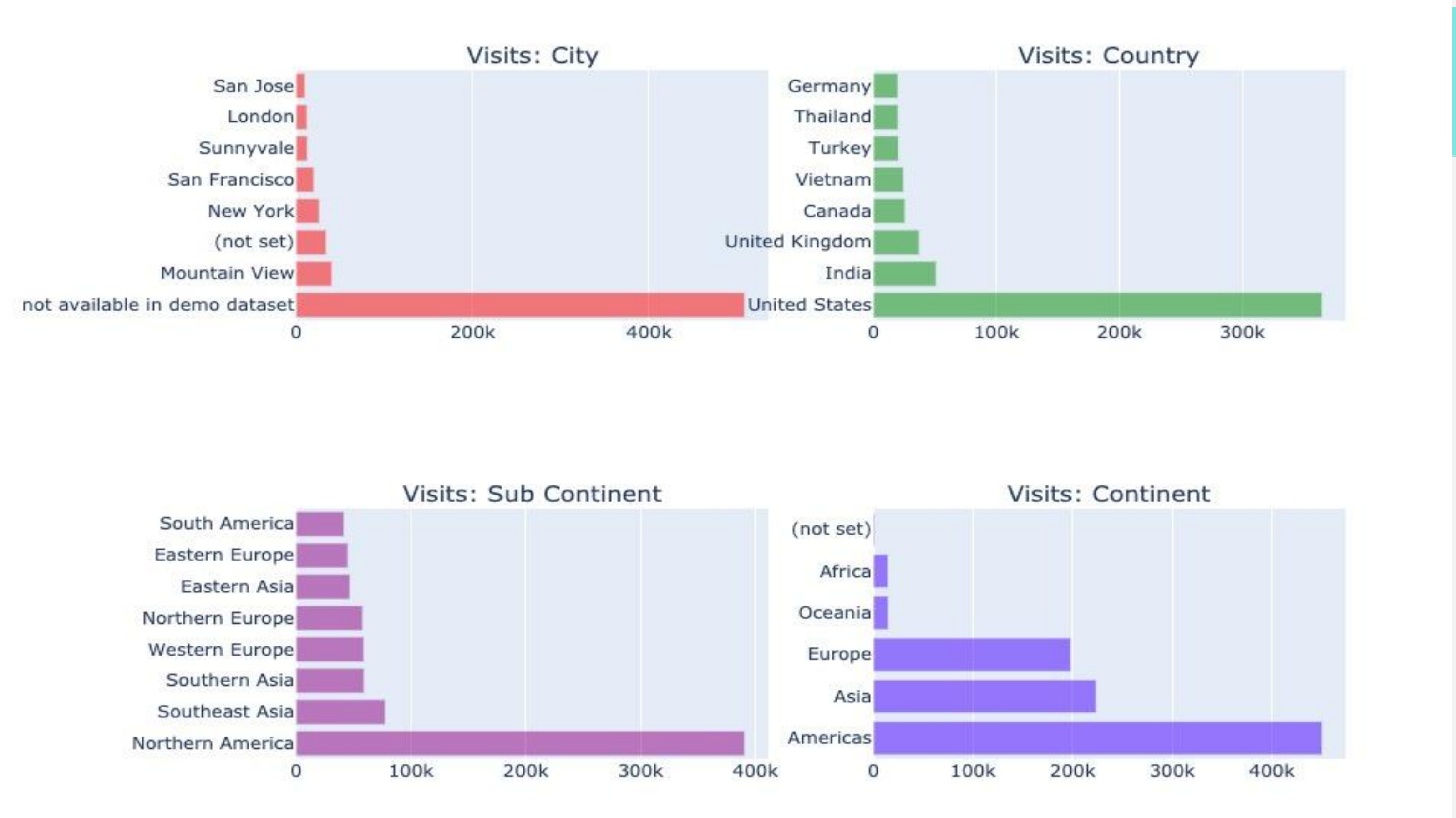
1) United States

3) United Kingdom



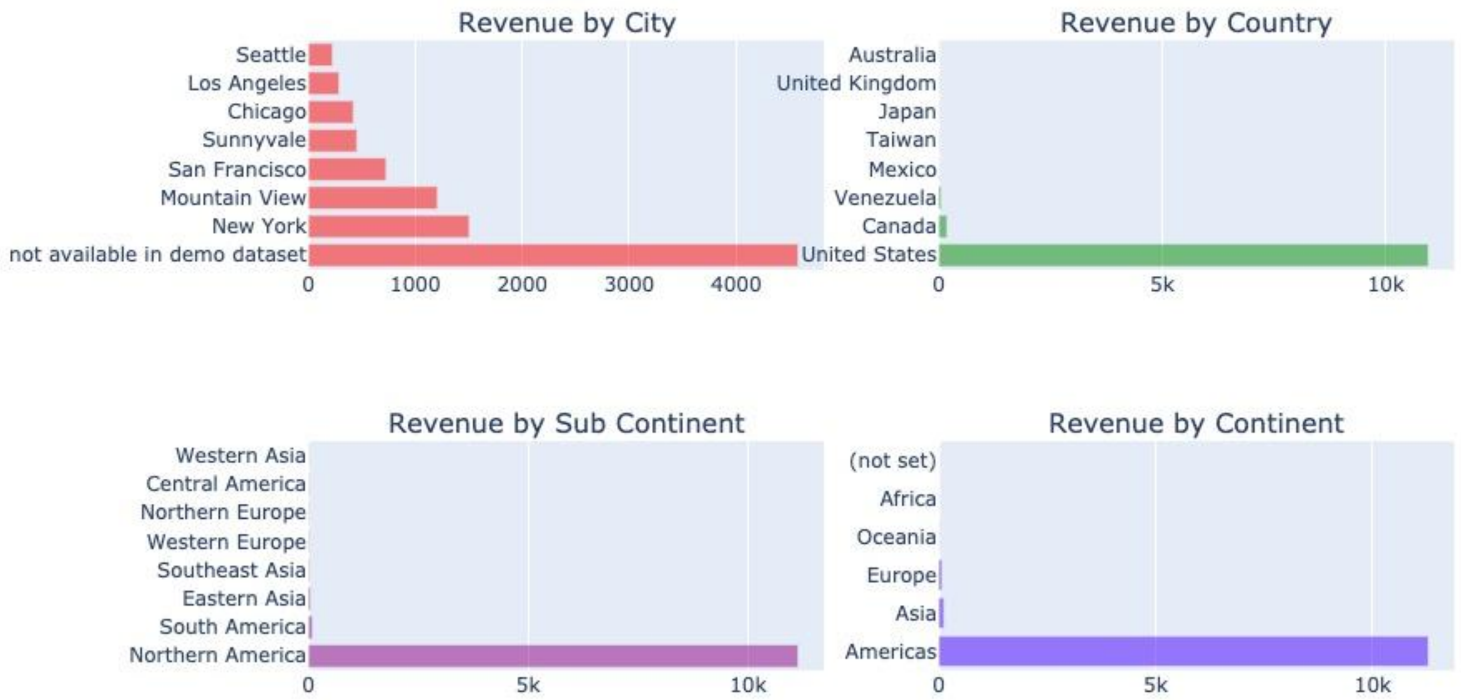
2) India

Customer Visits Distribution

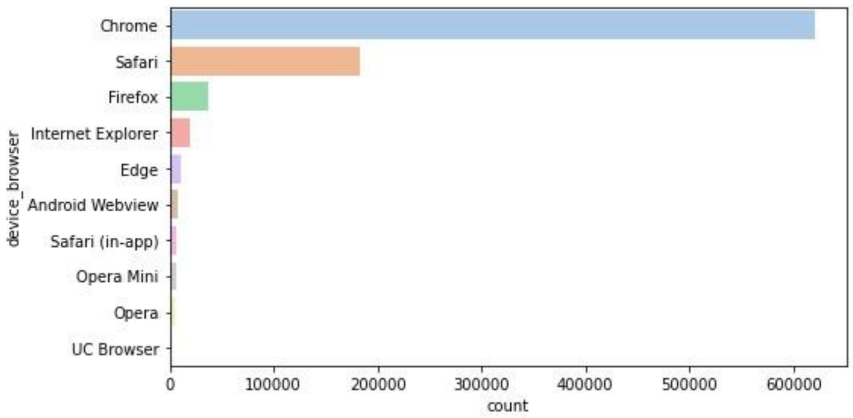


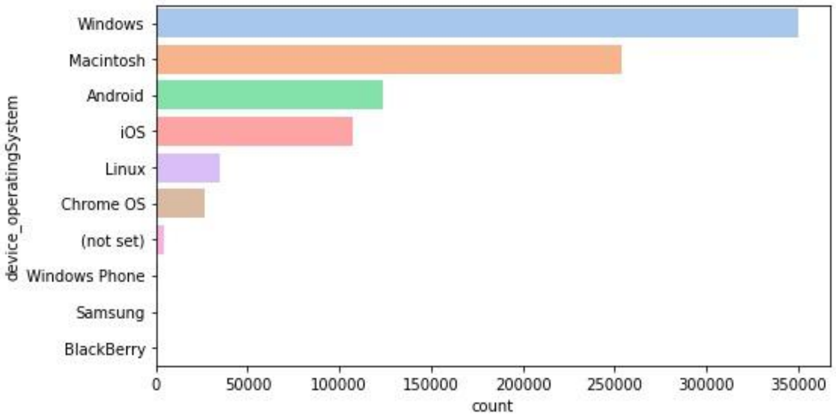
Visits and Revenue by Date

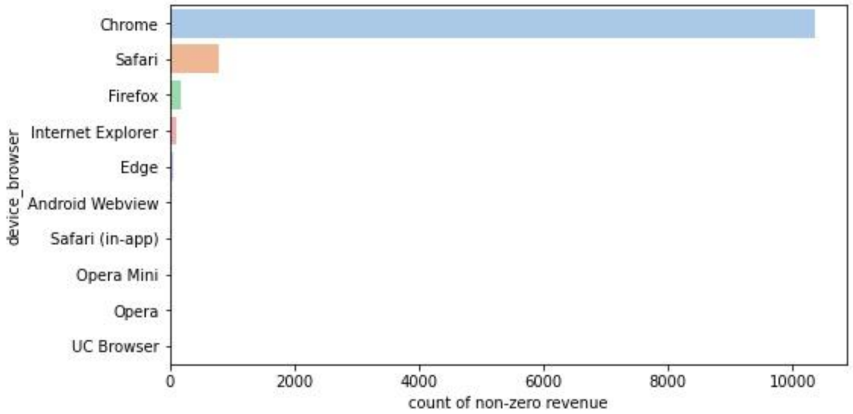
Revenue Sources

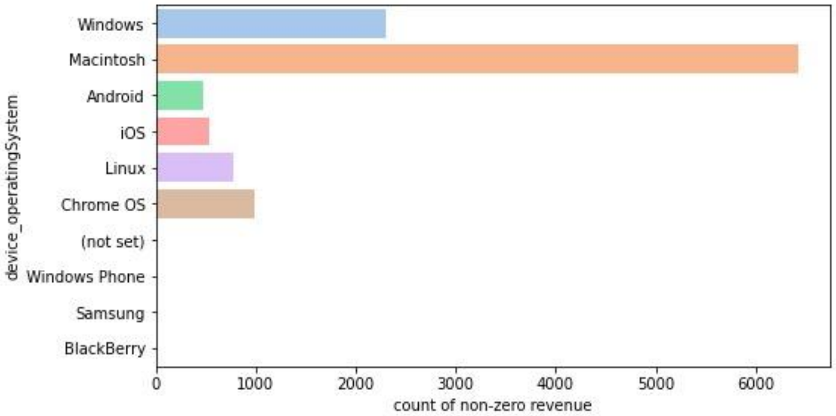


Browser and Operating System

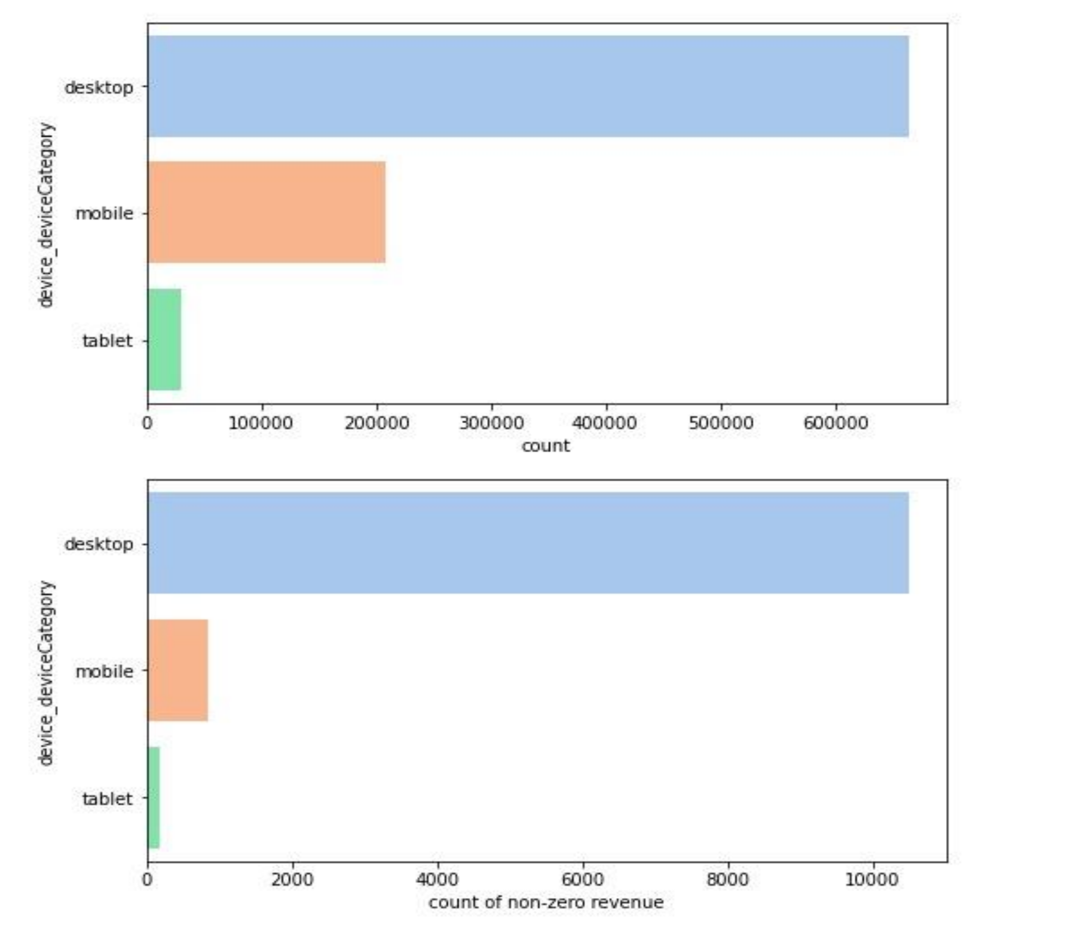




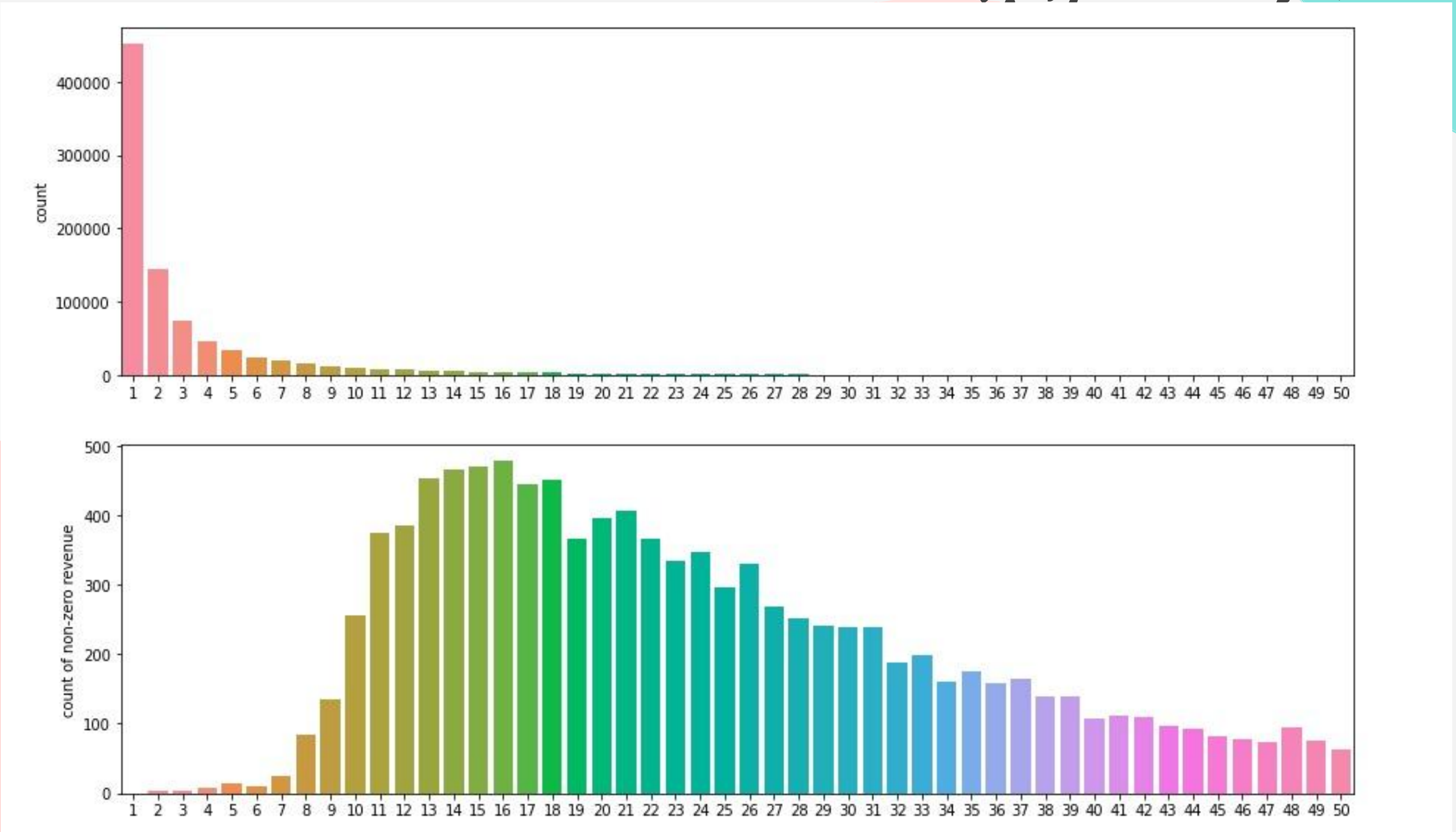




Device Categories



Totals PageViews



Data Cleaning and PreProcessing

Columns with 1 value and too many missing values are dropped

Identify Outliers

One Hot Encoding on categorical features

Modelling

Logistic Regression
Target Variable is Binary

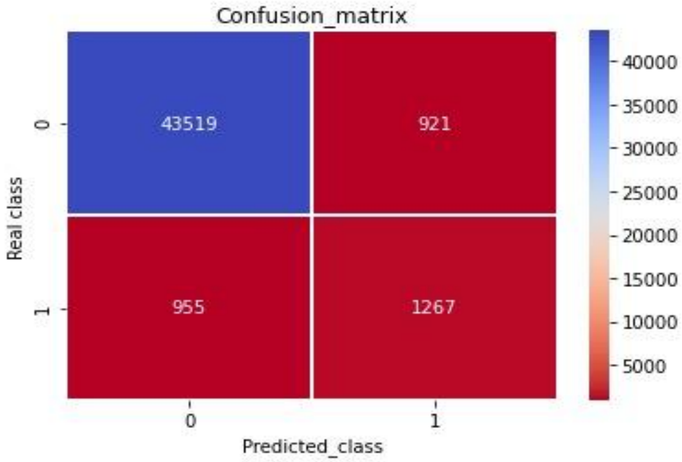
Extra Trees
Ensemble learning technique and predictions are made by using majority voting

Gradient Boosting
Trains many models in a gradual, additive and sequential method

Model Scores

	F1 Scores	Test Accuracy	Auc score
LogReg	0.623775	0.969547	0.760835
LogReg Filtered	0.566162	0.966204	0.727228
LogReg GridSearch	0.567107	0.966268	0.727689
ExtraTrees	0.444305	0.962046	0.656424
Extra Trees Filtered	0.558448	0.962925	0.739401
ExtraTrees GridSearch	0.00000	0.952381	0.50000
Gradient Boosting GridSearch	0.634515	0.966868	0.794487

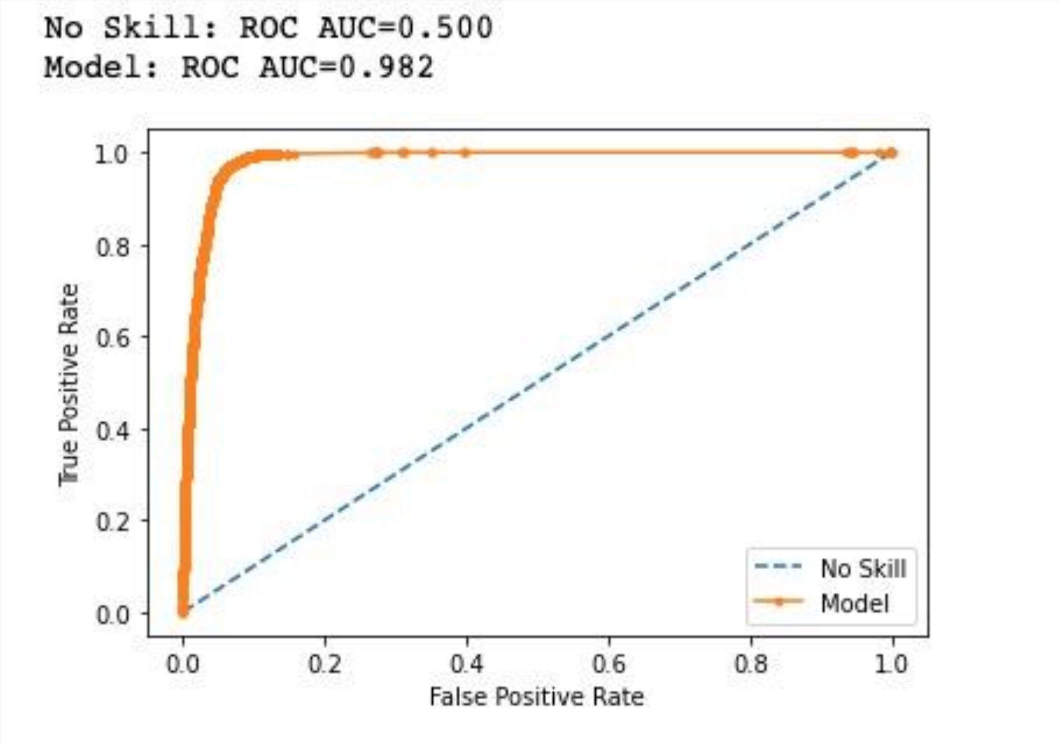
Best Model -Gradient Boosting Confusion Matrix



----------Classification Report----------------------------------

	precision	recall	f1-score	support
0	0.98	0.98	0.98	44440
1	0.58	0.57	0.57	2222
accuracy			0.96	46662
macro avg	0.78	0.77	0.78	46662
weighted avg	0.96	0.96	0.96	46662

Best Model Performance - ROC-AUC Curve



Conclusion

Logistic Regression, Extra Trees and Gradient Boosting models are used to model the data. There is a heavy class-imbalance in the data, at 100:1 ratio of non-revenue generating against revenue generating sessions. Random UnderSampler was used to adjust for the imbalance.

Gradient Boosting has shown to be the best performing model at an overall accuracy of 96.7%,with f1 score and auc scores of 63% and 79% respectively. ROC-AUC score achieved 98% using this method.

The test recall score works out to be 60% which is expectedly lower due to the highly imbalanced data and by using random undersampling, all of the training data points from the minority class (revenue generating) are used but instances are randomly removed from the majority training set till the desired balance is achieved which in this case the ratio applied was 1:20. One disadvantage of this approach is that some useful information might be lost from the majority class due to the undersampling.

Future Considerations

One future consideration is to use undersampling methods in conjunction with an oversampling technique for the minority class, which may result in better performance than using oversampling or undersampling alone on the training dataset

For the next iteration, the classifier model should be applied on unseen data to validate the scoring.

THANKS

Does anyone have any questions?

Presenter: Emily Kong