Revenue Radar

INSY 695 TERM PROJECT

FEBRUARY 23RD 2024

Meet the Revenue Radar Team

- 3 Data Scientists
- 2 Business Analysts
- 1 Data Analyst

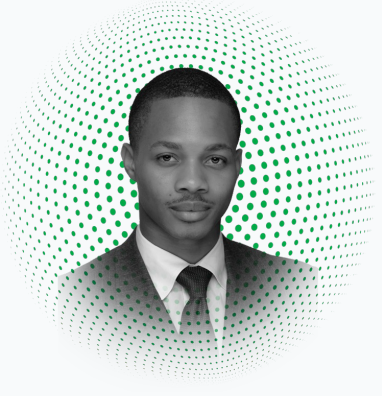
ABDULRAHMAN AROWORAMIMO
DATA SCIENTIST

AMMAD SOHAIL
DATA SCIENTIST

ANGEL OLUWOLE-ROTIMI
BUSINESS ANALYST

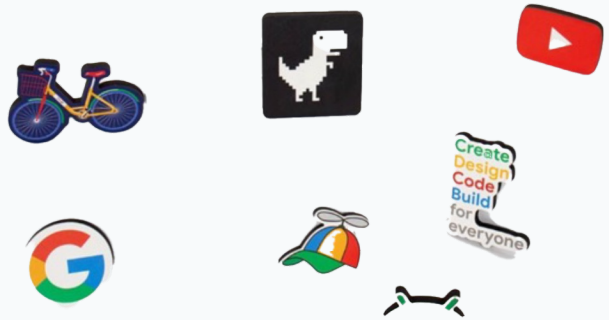
RODRIGO CASTRO
DATA ANALYST

XINGHEN LUO
DATA SCIENTIST



YVAN KAMMELU
BUSINESS ANALYST

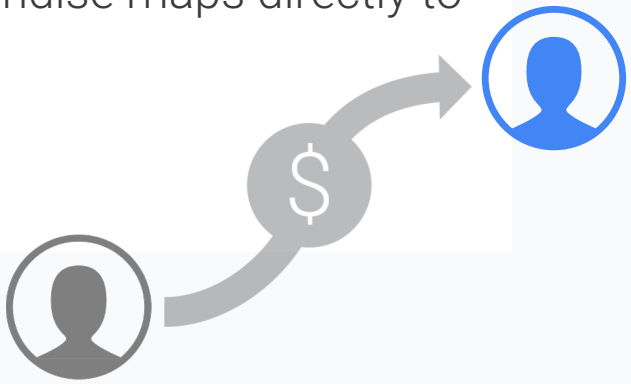
Business Context



Our Business Use Cases

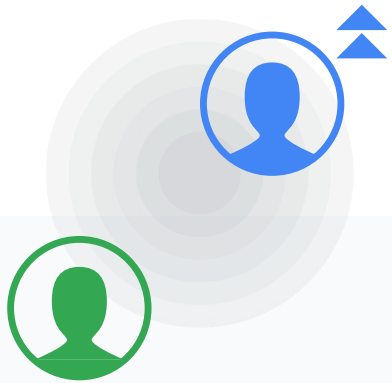
Customer Conversion

Conversion is a key indicator to understand and optimize, as the number of people purchasing and sporting merchandise maps directly to branding goals



Transaction Revenue

Optimizing revenue is critical for sustaining operations and funding the product innovation required for competitiveness in merchandising



Exploring Our Dataset

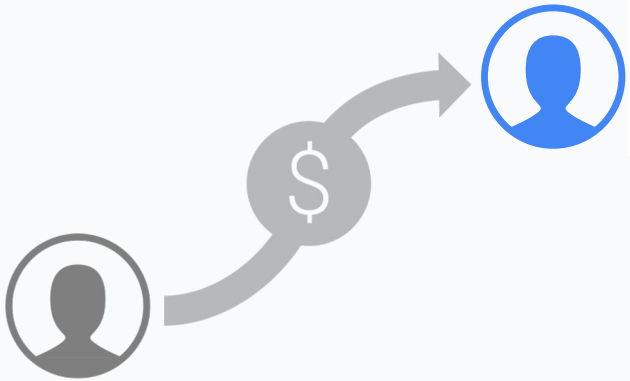
- 900K+ records
- 55 columns
- 4 ID
- 2 Temporal
- 10 Geographic
- 17 Operating System
- 22 Behavioural

Temporal	Geographic	Operating System	Behavioural
Date	Subcontinent	Browser	Traffic source
Visit Start Time	Country	Device Category	Page Views
	City		Transaction Revenue

Customer Conversion

Objectives

1. Determine which users to target for conversion nudges
2. Increase understanding of the user journey and profile of converting customers



Measuring Performance

Model Performance

Model performance should be measured with F1 for appropriate balance of precision and recall

Initiative Performance

Model must outperform current rule-based process to be piloted

Explainable models must be explored over the data science process

Data Transformations

To facilitate user-centric analysis, the dataset was transformed from session level to user level

Page Views → First Session Page Views, Last Session Page Views

Device Category → Number of visits by desktop, mobile, tablet

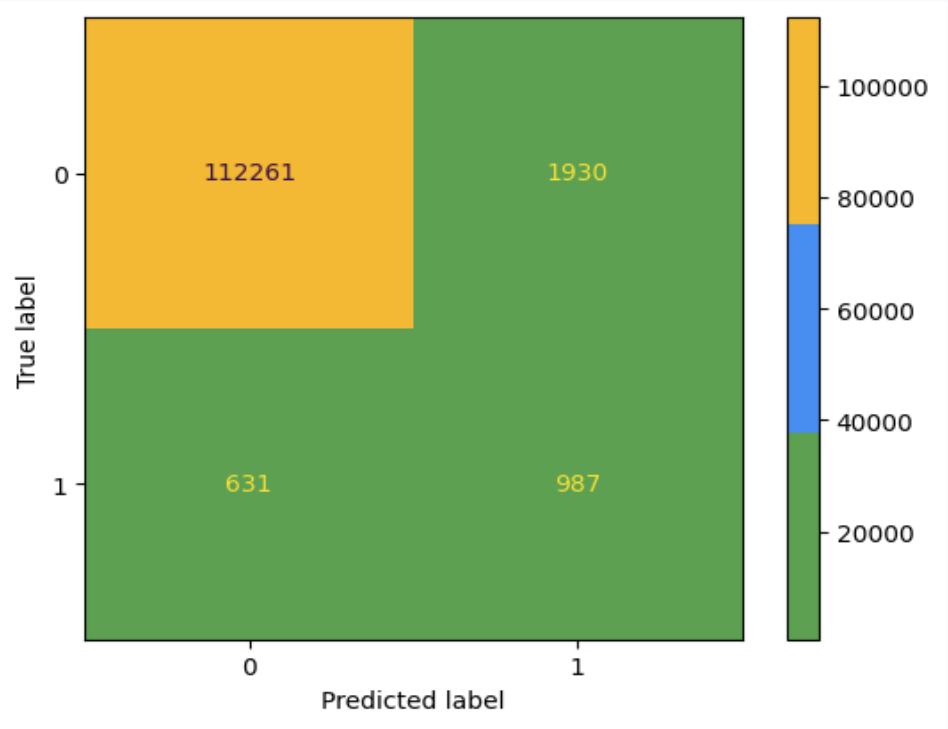
Columns excluded based on proportions of missing data and SME consultation

Customer Conversion

Raw Data → **Target variable** df['at_least_one_conversion'] dealt with null values / duplicates → **Remove Outliers** Isolation Forest → **Feature Selection** Top 20 features – Random Forest → **Class Imbalance** Class Weight parameter → **Train Data**

Approach	Model	Accuracy	Precision	Recall	F1-score
Baseline	Rule based	0.98	0.31	0.44	0.36
Class Weight	Random Forest	0.98	0.29	0.53	0.38
Stacking Ensemble	Log. Regression (meta learner) Random Forest XGBoost Ada Boost D_tree	0.98	0.35	0.56	0.43
Optimal Threshold	Log. Regression	0.98	0.37	0.55	0.44
Fine Tuning	Log. Regression	0.98	0.34	0.61	0.44

Threshold: 0.59



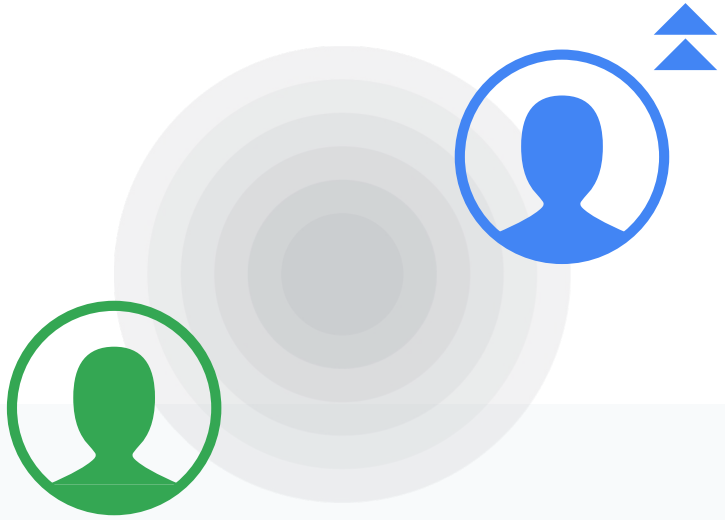
Transaction Revenue

Objective

1. Identify which customers are likely to spend more at GStore, focusing on understanding their spending behaviors and patterns.
2. Determine the key factors contributing to higher customer spending, enabling targeted marketing strategies and product innovation.

Measuring Performance

Adopted Mean Absolute Error (MAE) as our primary metric to evaluate the accuracy of the regression model in predicting customer spending.



Data Transformations

To facilitate transaction level analysis, the dataset was filtered for session with transaction revenue == 0

Columns excluded based on proportions of missing data and use case relevance.

Transaction Revenue

Raw Data

Data Cleaning
Remove Unary Columns
Group and Sort Columns
Drop Columns with Excessive NaNs

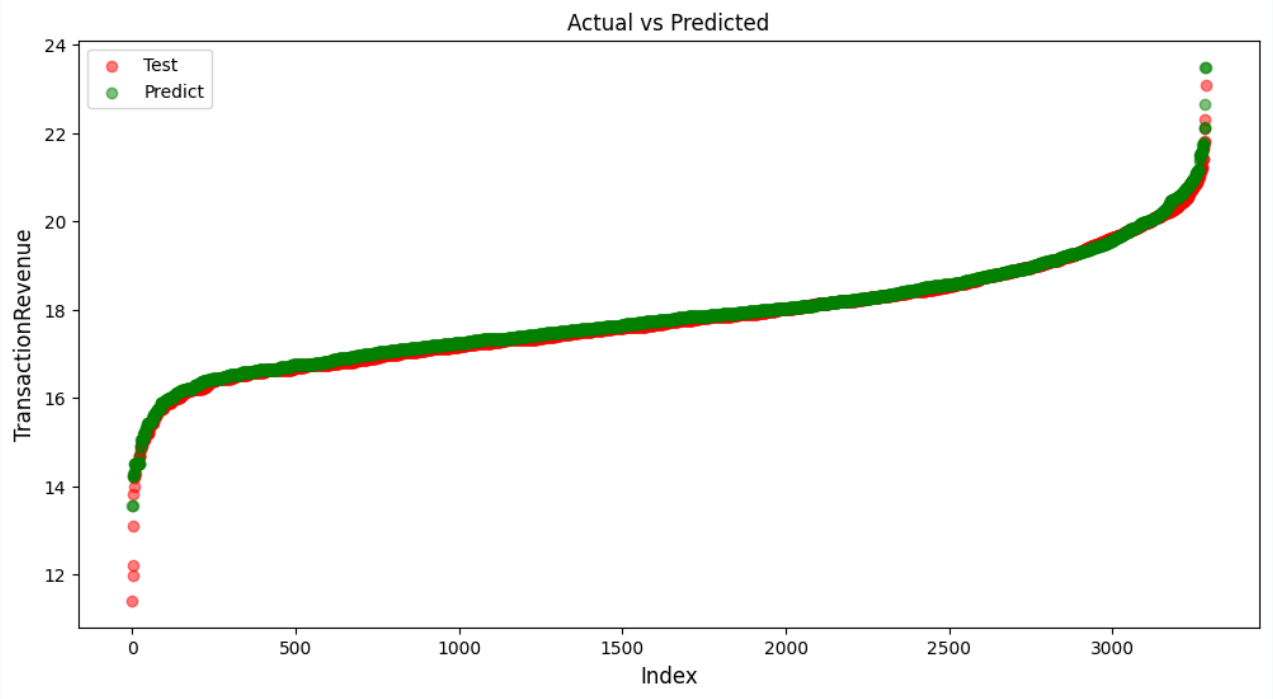
Feature Engineering
Transform and Consolidate Features

Encoding and Multicolineairty Analysis
Create Dummy Variables
Drop Highly Correlated Features

Data Scaling
Scaling Target Variable (TotalTransactionRevenue)

Split into Train and Test Dataset
(11000 rows)

	Pre-Hyperparameter Tuning			Post-Hyperparameter Tuning
Model	**Mean Absolute Error**	**Mean Squared Error**	**Median Absolute Error**	**Mean Absolute Error**
Linear Regression	0.863	1.255	0.687	-
Ridge Regression	0.863	1.255	0.687	0.848
Lasso Regression	0.882	1.313	0.711	0.848
Random Forest Regressor	0.942	1.477	0.779	0.838
XGB Regressor	0.902	1.360	0.738	0.835
Gradient Boosting Regressor	0.852	1.220	0.681	0.835
AdaBoost Regressor	0.920	1.359	0.773	0.857
Decision Tree Regressor	1.176	2.322	0.943	0.854
Support Vector Regression	0.856	1.264	0.671	0.832

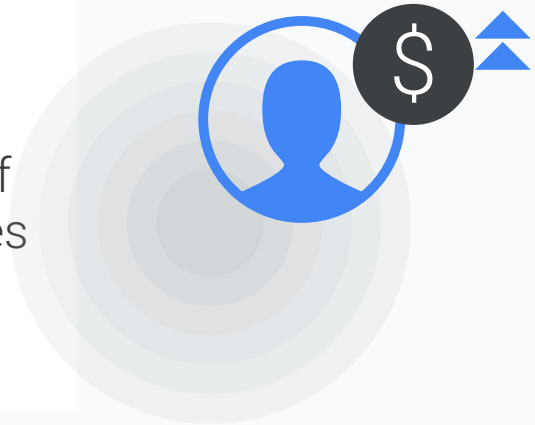


How Revenue Radar impacts the GStore?

Transaction Revenue

The higher profitability associated with returning visitors signals a strategic pivot – from a broad-brush marketing approach to a more nuanced, targeted strategy.

Allocating a more significant portion of our budget towards retention strategies and targeted marketing for previous visitors



Customer Conversion

Automated prioritization of ad-space bidding should be transitioned from rules-based model to predictive model

Training and maintenance via offline batch learning over pilot period

Boost number of conversions for a given ad spend budget

Future State

Areas of Improvement

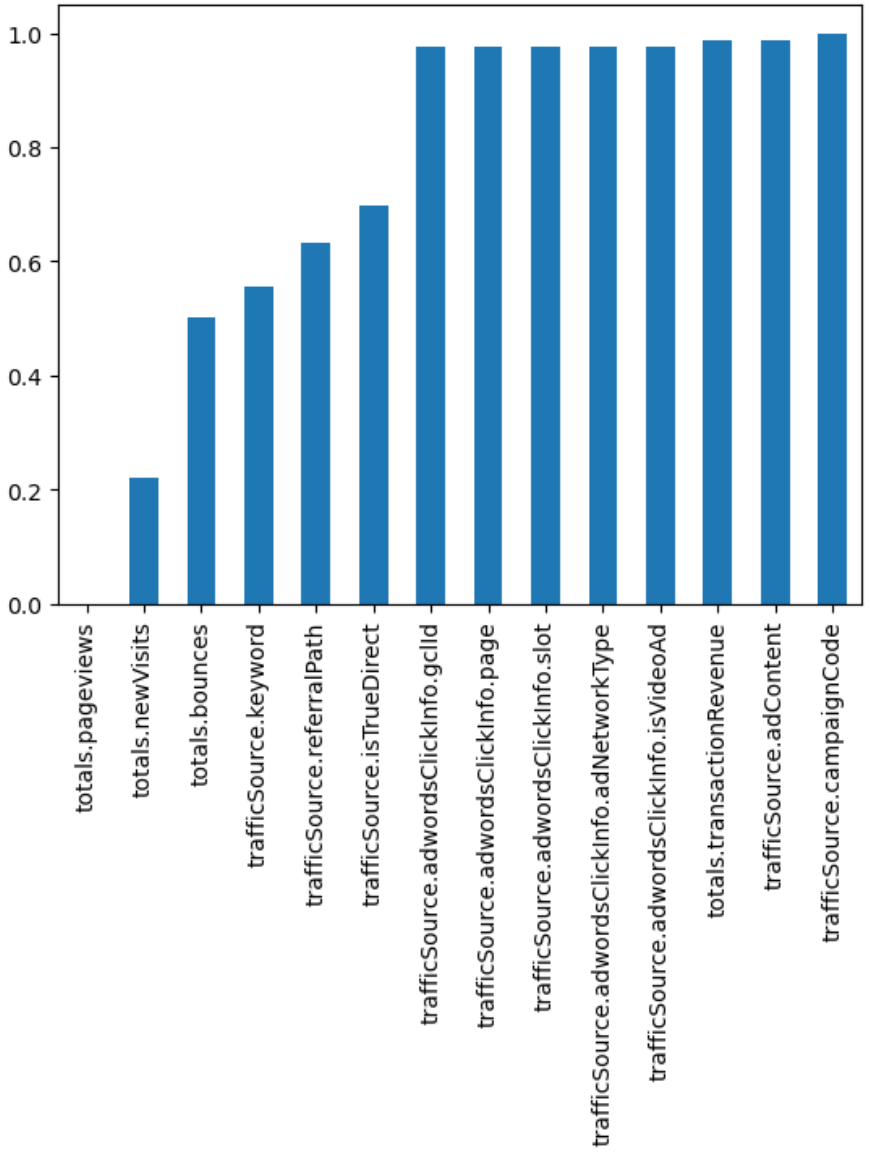
- Uplift modeling (Causal ML)
- Further Customer Segmentation
- High impact marketing efforts: Further optimize marketing cost by zeroing in on the ‘persuadable’ and not spending on ‘sure things’

Other Applications

- Subscription based services
- Financial services

Appendix

Missing Data (%)



Coefficients from Conversion Final Model

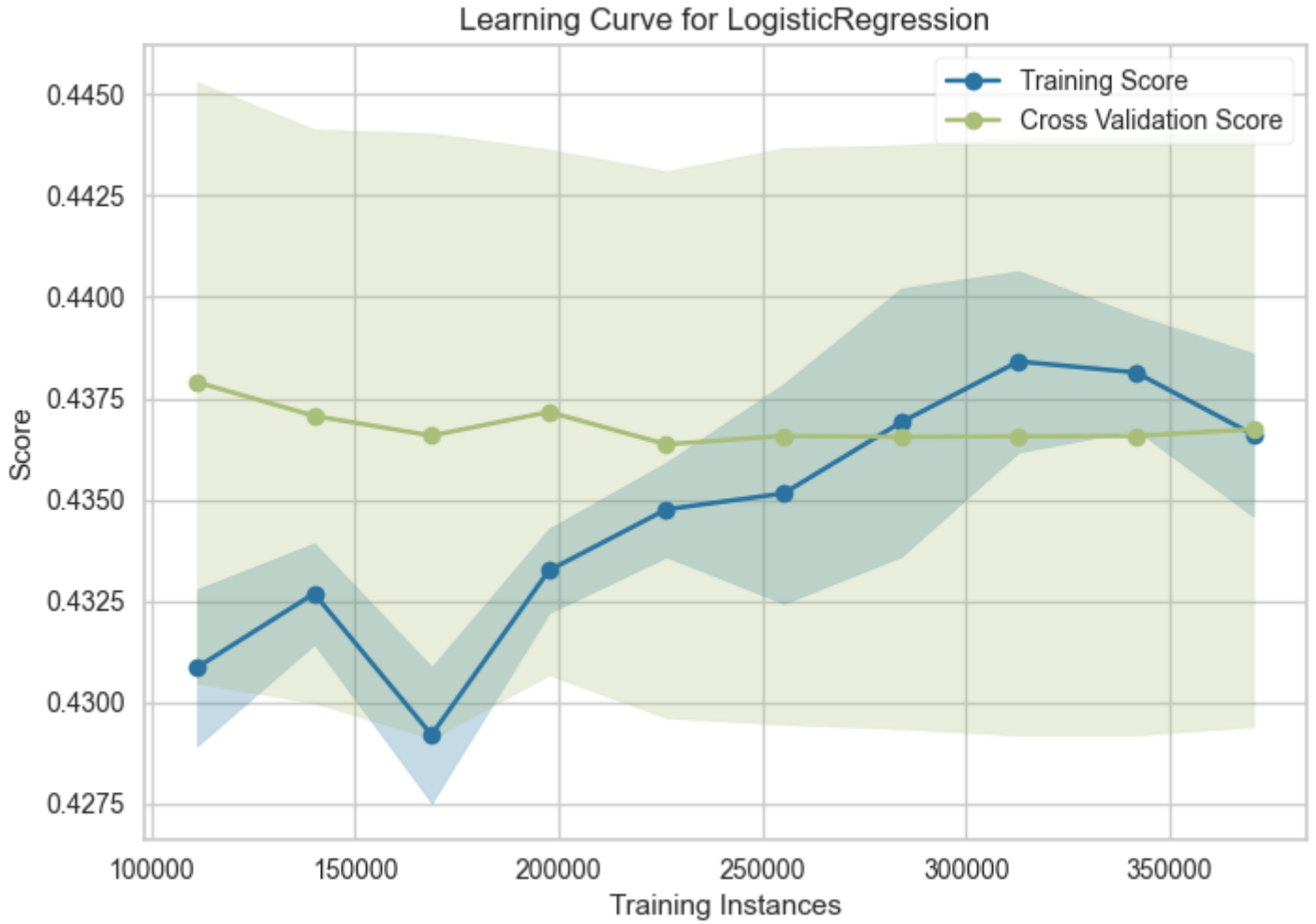
	Coefficient
subContinent_Northern America	2.094998
continent_Americas	1.495318
subContinent_Southern Asia	-1.300933
FirstChannelVisit_Social	-1.102428
TotalBounces	-0.949972
LastChannelVisit_Referral	0.915346
LastChannelVisit_Social	-0.757257
FirstChannelVisit_Organic Search	-0.729678
continent_Asia	0.630254
Organic Search	0.626897
Social	0.611898
Paid Search	0.567424
Referral	0.536795
Direct	0.483182
FirstChannelVisit_Referral	-0.463616
desktop	0.360982
mobile	0.180981
FirstSessionPageviews	0.131693
continent_Europe	-0.116171
LastChannelVisit_Organic Search	0.062603

Optimal Class Weight

0:1

1:10

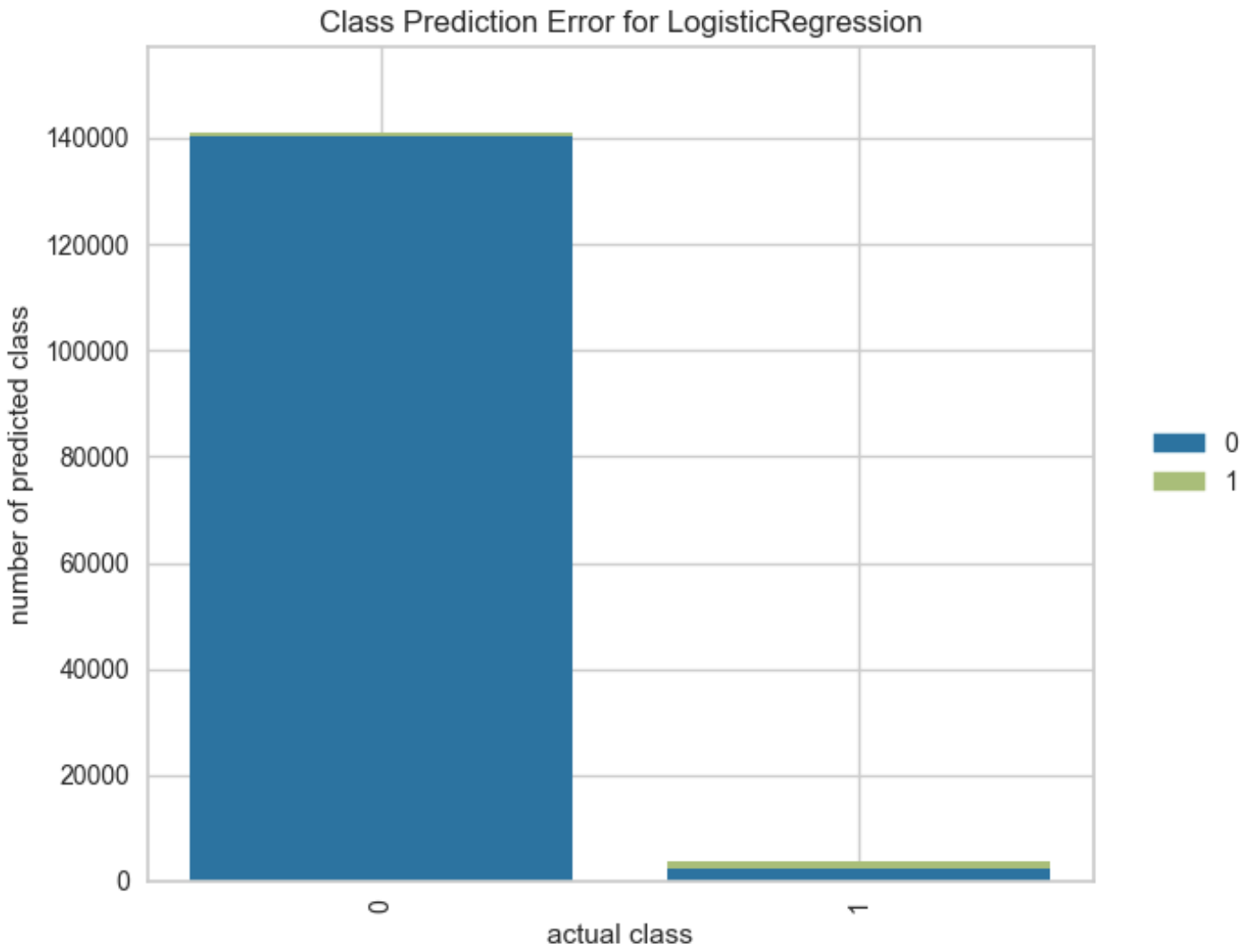
Learning Curve



Performance on Test (Holdout) Set

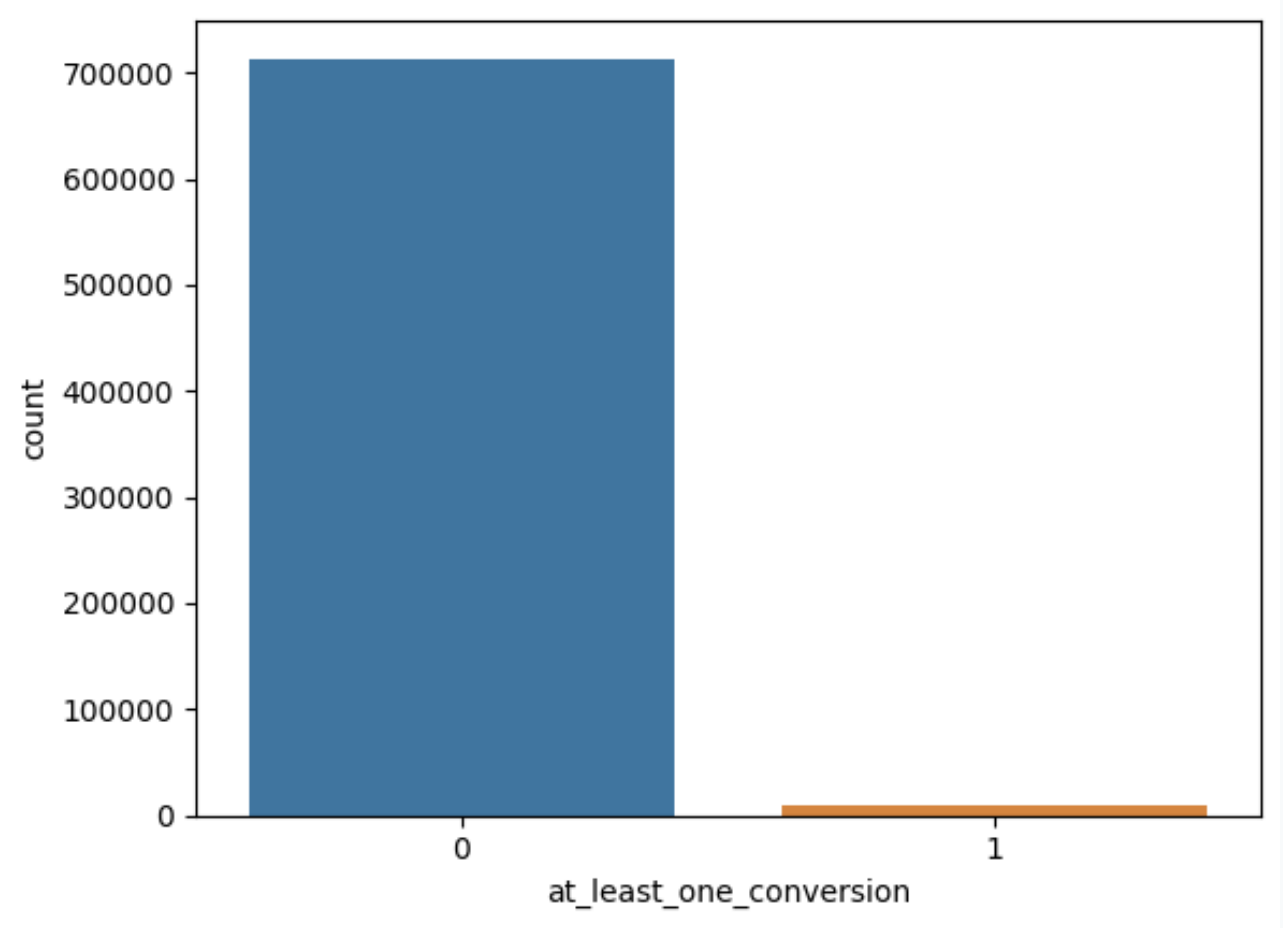
	precision	recall	f1-score	support
0	0.99	0.98	0.99	142738
1	0.33	0.63	0.43	2023
accuracy			0.98	144761
macro avg	0.66	0.80	0.71	144761
weighted avg	0.99	0.98	0.98	144761

Prediction Errors



Class Distribution

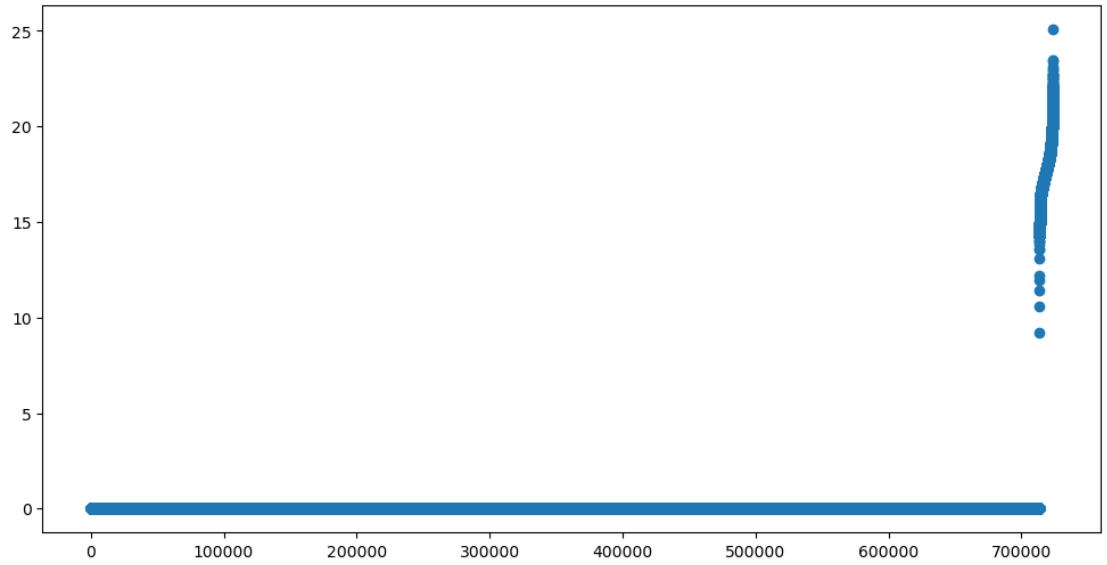
99:1%



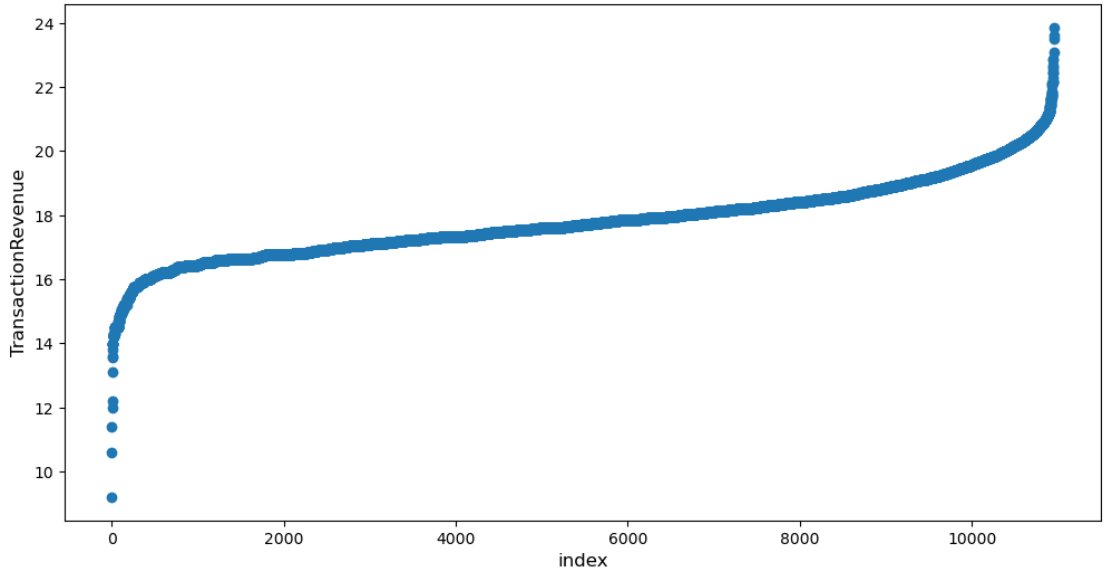
Revenue Distribution

```
import numpy as np
import pandas as pd

gdf = train_df.groupby("fullVisitorId")["totals.transactionRevenue"].sum().reset_index()

gdf["totals.transactionRevenue"] = pd.to_numeric(gdf["totals.transactionRevenue"], errors="coerce")

gdf = gdf.dropna(subset=["totals.transactionRevenue"])

gdf["totals.transactionRevenue"] = np.sort(gdf["totals.transactionRevenue"])

total_customers = gdf.shape[0]
revenue_customers = gdf[gdf["totals.transactionRevenue"] > 0].shape[0]
percentage_revenue_customers = (revenue_customers / total_customers) * 100

print("Percentage of customers producing revenue: {:.2f}%".format(percentage_revenue_customers))
```

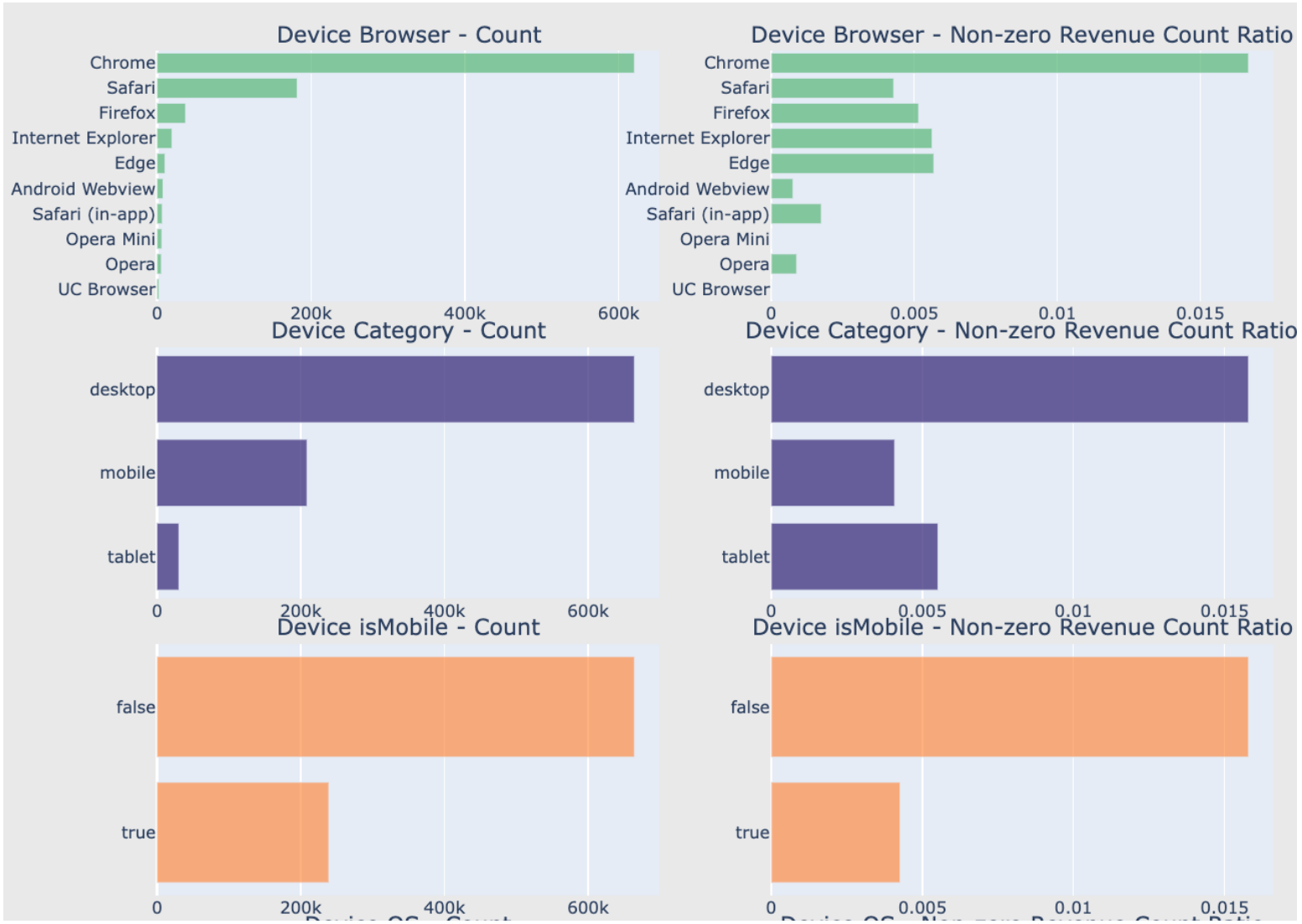
```
Percentage of customers producing revenue: 1.40%
```



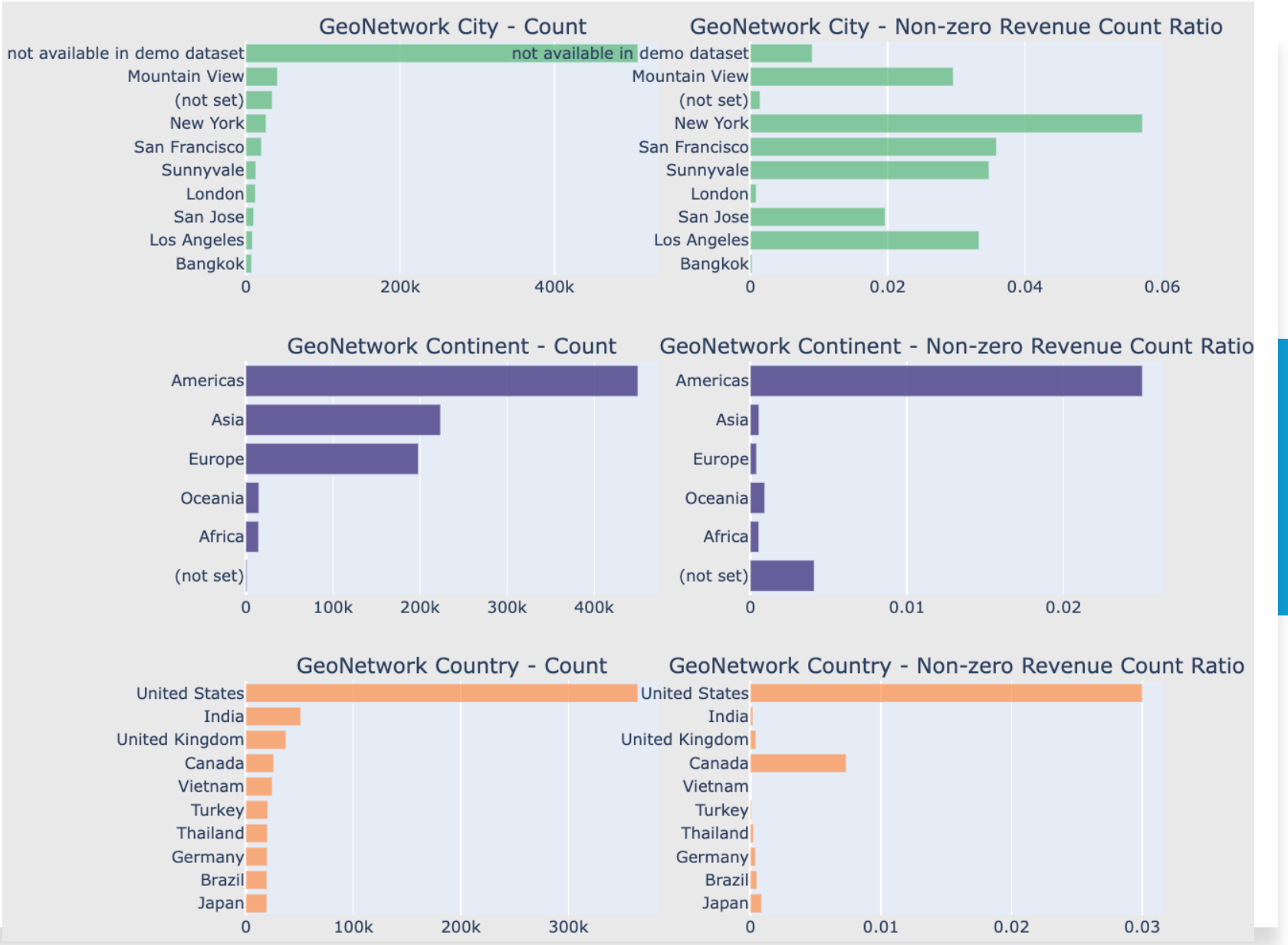
Before the cleaning



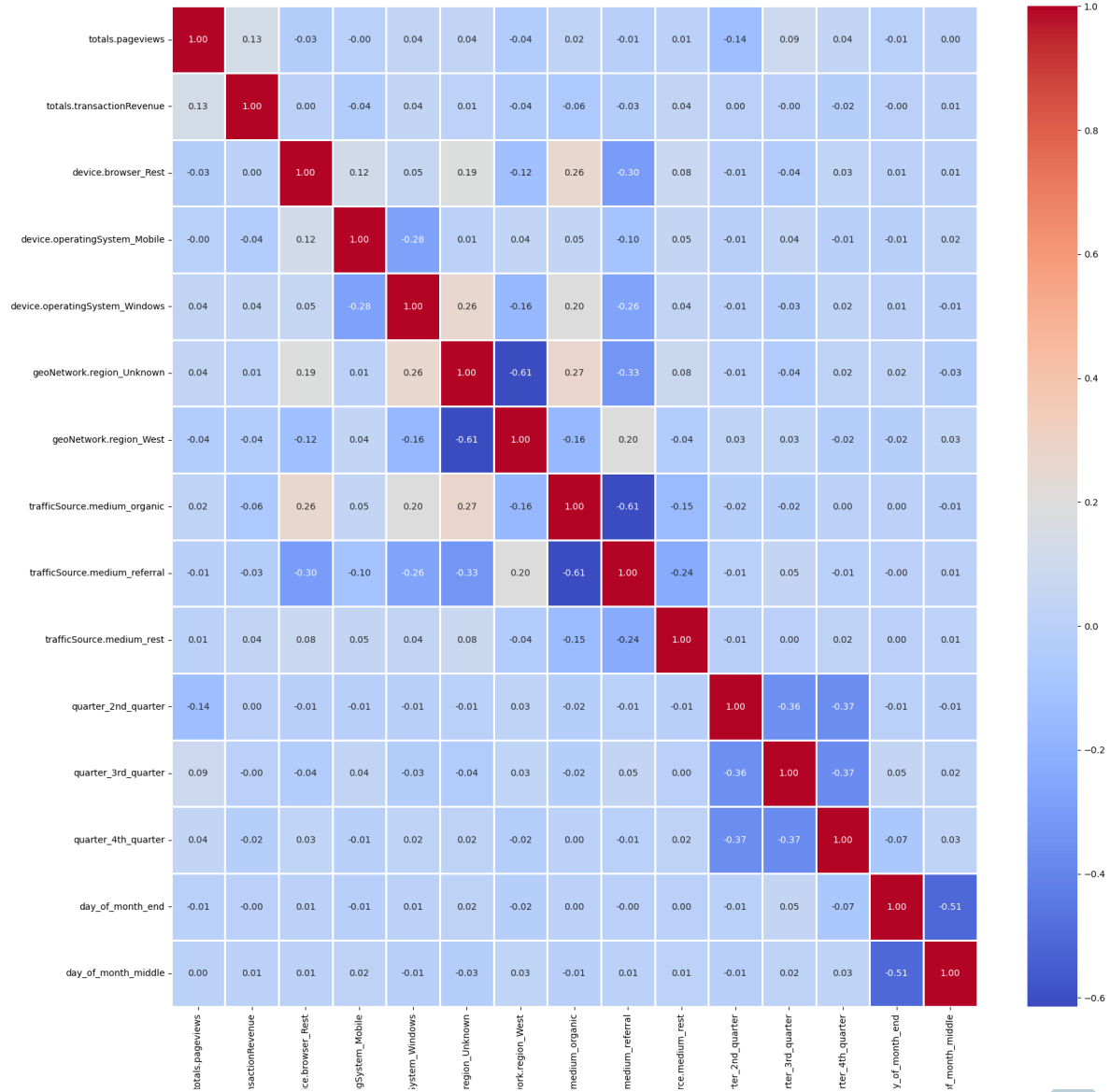
After Cleaning and Log Transformation


Device Browser - Count
Chrome
Safari
Firefox
Internet Explorer
Edge
Android Webview
Safari (in-app)
Opera Mini
Opera
UC Browser
0
200k
400k
600k
Device Browser - Non-zero Revenue Count Ratio
Chrome
Safari
Firefox
Internet Explorer
Edge
Android Webview
Safari (in-app)
Opera Mini
Opera
UC Browser
0
0.005
0.01
0.015
Device Category - Count
desktop
mobile
tablet
0
200k
400k
600k
Device Category - Non-zero Revenue Count Ratio
desktop
mobile
tablet
0
0.005
0.01
0.015
Device isMobile - Count
false
true
0
200k
400k
600k
Device isMobile - Non-zero Revenue Count Ratio
false
true
0
0.005
0.01
0.015

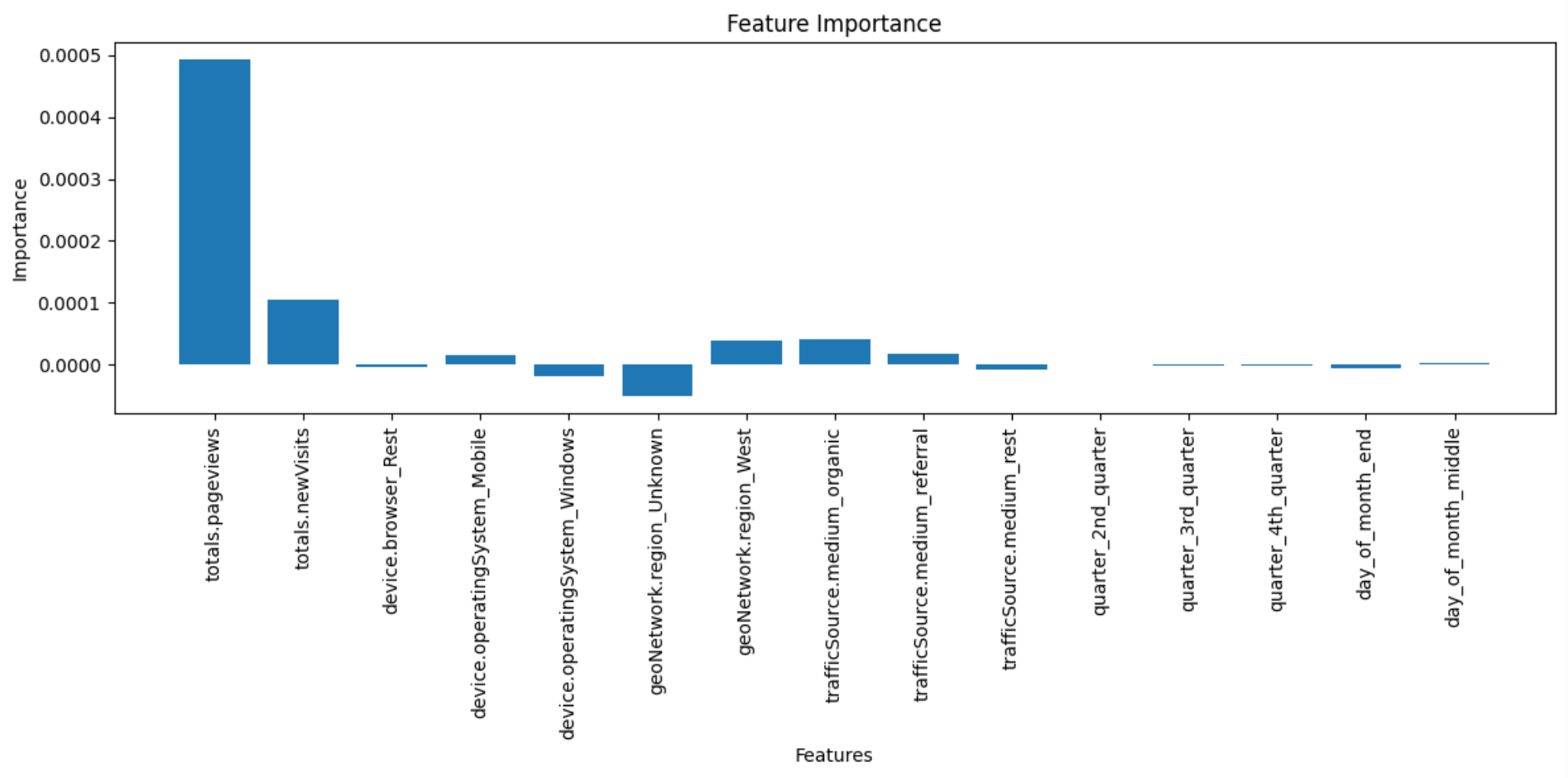
Device


GeoNetwork City - Count
not available in demo dataset
Mountain View
(not set)
New York
San Francisco
Sunnyvale
London
San Jose
Los Angeles
Bangkok
0
200k
400k
GeoNetwork City - Non-zero Revenue Count Ratio
not available in demo dataset
Mountain View
(not set)
New York
San Francisco
Sunnyvale
London
San Jose
Los Angeles
Bangkok
0
0.02
0.04
0.06
GeoNetwork Continent - Count
Americas
Asia
Europe
Oceania
Africa
(not set)
0
100k
200k
300k
400k
GeoNetwork Continent - Non-zero Revenue Count Ratio
Americas
Asia
Europe
Oceania
Africa
(not set)
0
0.01
0.02
GeoNetwork Country - Count
United States
India
United Kingdom
Canada
Vietnam
Turkey
Thailand
Germany
Brazil
Japan
0
100k
200k
300k
GeoNetwork Country - Non-zero Revenue Count Ratio
United States
India
United Kingdom
Canada
Vietnam
Turkey
Thailand
Germany
Brazil
Japan
0
0.01
0.02
0.03

Geo Data


totals.pageviews
totals.transactionRevenue
device.browser_Rest
device.operatingSystem_Mobile
device.operatingSystem_Windows
geoNetwork.region_Unknown
geoNetwork.region_West
trafficSource.medium_organic
trafficSource.medium_referral
trafficSource.medium_rest
quarter_2nd_quarter
quarter_3rd_quarter
quarter_4th_quarter
day_of_month_end
day_of_month_middle
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Feature Importance using SVR



Example of Underfitting