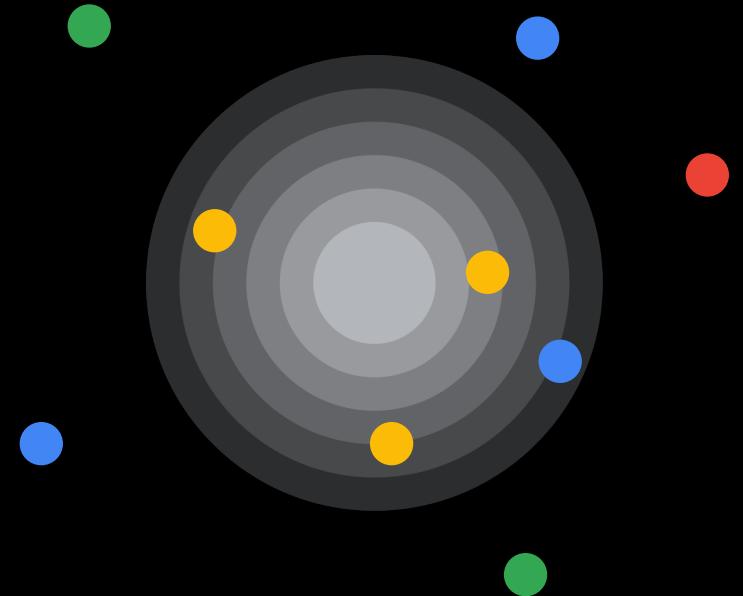


Revenue Radar



INSY 695 TERM PROJECT
APRIL 25TH 2024



Abdul-AA
aoluwolerotimi
TashfeenAhmed12
tiger7789
ykamm

Meet the Revenue Radar Team



ABDULRAHMAN AROWARAMIMO
DATA SCIENTIST



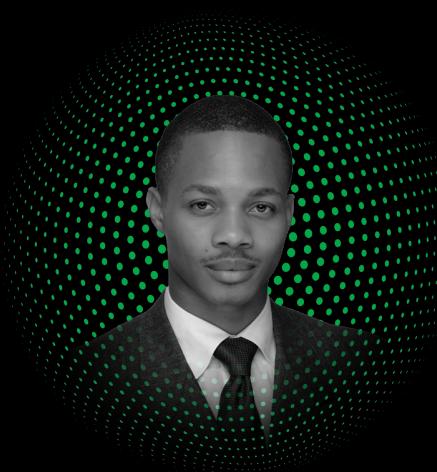
TASHFEEN AHMED
DATA SCIENTIST



ANGEL OLUWOLE-ROTINI
BUSINESS ANALYST



XINGHEN LUO
DATA SCIENTIST



YVAN KAMMELU
BUSINESS ANALYST

Dataset



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

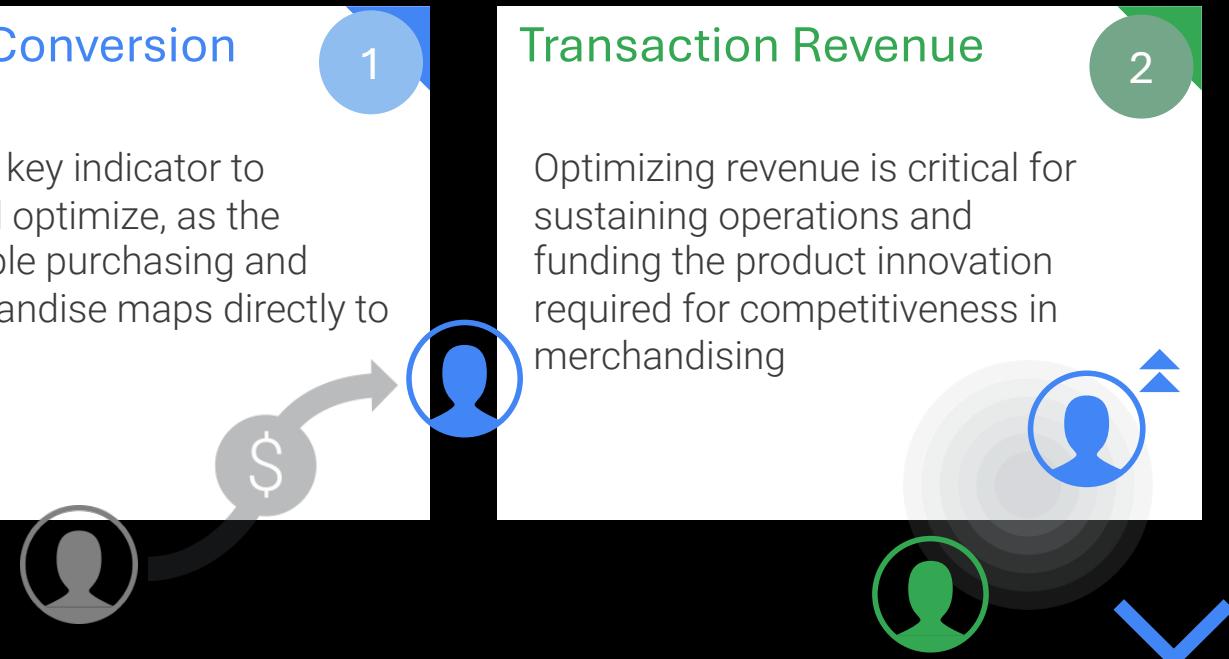
Business Context



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

1



Transaction Revenue

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

2

Exploring Our Dataset

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

#	Column	Non-Null Count	Dtype	
0	channelGrouping	903653	non-null	object
1	date	903653	non-null	int64
2	fullVisitorId	903653	non-null	object
3	sessionId	903653	non-null	object
4	socialEngagementType	903653	non-null	object
5	visitId	903653	non-null	int64
6	visitNumber	903653	non-null	int64
7	visitStartTime	903653	non-null	int64
8	continent	903653	non-null	object
9	subContinent	903653	non-null	object
10	country	903653	non-null	object
11	region	903653	non-null	object
12	metro	903653	non-null	object
13	city	903653	non-null	object
14	cityId	903653	non-null	object
15	networkDomain	903653	non-null	object
16	latitude	903653	non-null	object
17	longitude	903653	non-null	object
18	networkLocation	903653	non-null	object
19	browser	903653	non-null	object
20	browserVersion	903653	non-null	object
21	browserSize	903653	non-null	object
22	operatingSystem	903653	non-null	object
23	operatingSystemVersion	903653	non-null	object
24	isMobile	903653	non-null	bool
25	mobileDeviceBranding	903653	non-null	object
26	mobileDeviceModel	903653	non-null	object
27	mobileInputSelector	903653	non-null	object
28	mobileDeviceInfo	903653	non-null	object
29	mobileDeviceMarketingName	903653	non-null	object
30	flashVersion	903653	non-null	object
31	language	903653	non-null	object
32	screenColors	903653	non-null	object
33	screenResolution	903653	non-null	object
34	deviceCategory	903653	non-null	object
35	visits	903653	non-null	int64
36	hits	903653	non-null	int64
37	pageviews	903553	non-null	float64
38	bounces	450630	non-null	float64
39	newVisits	703060	non-null	float64
40	transactionRevenue	11515	non-null	float64
41	campaign	903653	non-null	object
42	source	903653	non-null	object
43	medium	903653	non-null	object
44	keyword	400724	non-null	object
45	adwordsClickInfo.criteriaParameters	903653	non-null	object
46	isTrueDirect	274005	non-null	object
47	referralPath	330941	non-null	object
48	adwordsClickInfo.page	21460	non-null	float64
49	adwordsClickInfo.slot	21460	non-null	object
50	adwordsClickInfo.gclId	21561	non-null	object
51	adwordsClickInfo.adNetworkType	21460	non-null	object
52	adwordsClickInfo.isVideoAd	21460	non-null	object
53	adContent	10946	non-null	object
54	campaignCode	1	non-null	object

Process Improvements

Data Cleaning & Preprocessing	Feature Engineering	Modelling	Model Evaluation
<ol style="list-style-type: none">1. EDA, handling missing or uninformative data2. Creation of user level dataset3. Correlation checks, outlier removal with isolation forest	<ol style="list-style-type: none">1. Creation of user-level features2. Binning of categorical Features3. Scaling target variables4. Feature selection with Random Forest	<ol style="list-style-type: none">1. Sampling methods appropriate for imbalance dataset for train, validation, test split2. Class weight parameters3. Varying model experimentations, including ensembles	<ol style="list-style-type: none">1. Best conversion model: Fine-tuned Logistic Regression (F1:0.44)2. Best transaction revenue model: Support Vector Regression (MAE:0.832)

Process Improvements

Data Cleaning & Preprocessing	Feature Engineering	Modelling	Model Evaluation	Productionization
<ol style="list-style-type: none">1. EDA, handling missing or uninformative data2. Creation of user level dataset3. Correlation checks, outlier removal with isolation forest4. Refined cleaning of geographical data5. New features included from additional domain knowledge6. Information Leakage prevention through cut-off date in train/test split	<ol style="list-style-type: none">1. Creation of user-level features2. Binning of categorical Features3. Scaling target variables4. Feature selection with Random Forest5. Dimensionality reduction with PCA and Clustering6. Creation of interaction variables	<ol style="list-style-type: none">1. Sampling methods appropriate for imbalance dataset for train, validation, test split2. Class weight parameters3. Varying model experimentations, including ensembles4. Additional models tested5. Experimentation logging with MLFlow6. Hyperparameter optimization with Optuna7. AutoML model and pipeline optimization with TPOT	<ol style="list-style-type: none">1. Best conversion model: Fine-tuned Logistic Regression (F1:0.44)2. Best transaction revenue model: Support Vector Regression (MAE:0.832)3. Best conversion model: XGBoost (F1:0.53)4. Best transaction revenue model: Support Vector Regression (MAE: 0.8195)	<ol style="list-style-type: none">1. Model registry via MLFlow2. Endpoint creation and MVP FastAPI webapp3. Databricks batch streaming to Confluent Cloud4. Dockerization

Customer Conversion

Additional Features Considered

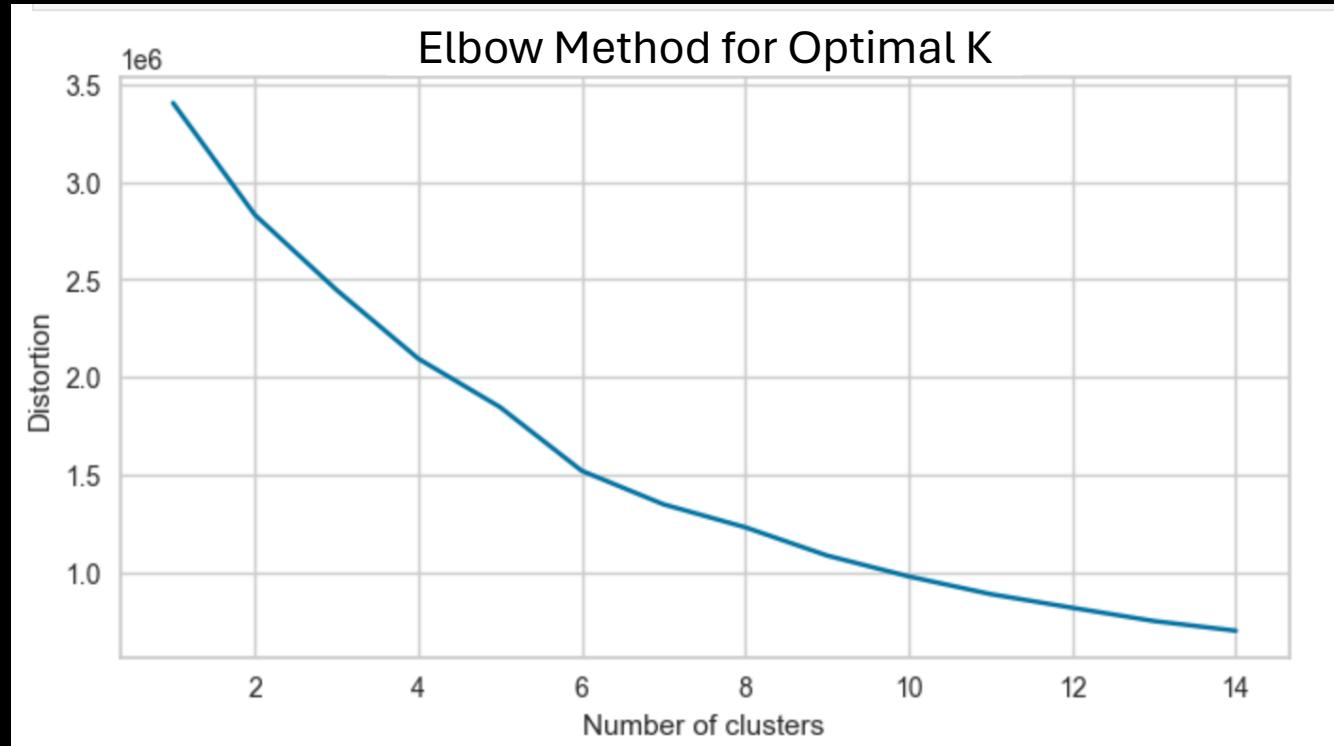
- Meaningful null values in isTrueDirect
- Multiple columns with high dimensionality
- Logging ad interactions transformed into single flag

Dimensionality Reduction

- Source column with 380 unique values transformed into 3 PCA components
- Country column with 50 unique values transformed into 3 PCA components
- Cluster label categorical features derived from frequency of site visits, types of devices used, depth of engagement on visits

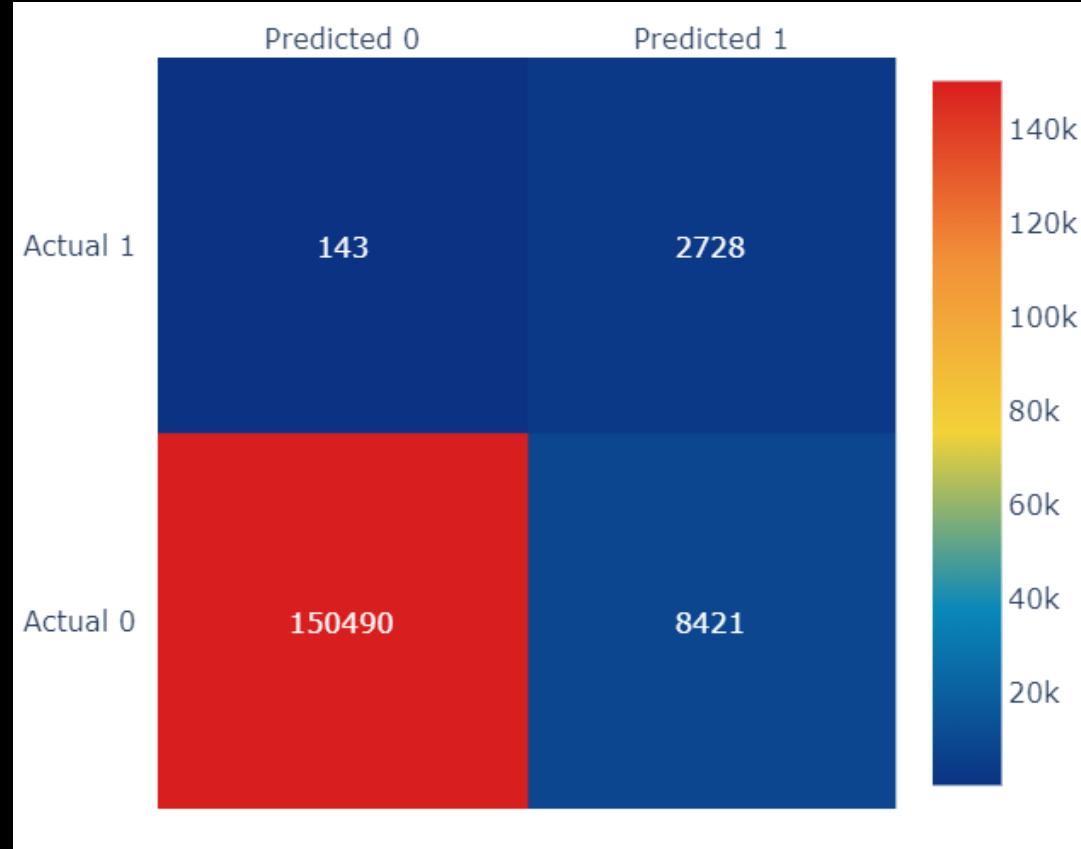
Interaction Variable

- User Engagement Depth



Customer Conversion

Approach	Model	Precision	Recall	F1-score
Hyperparameter F1 score	Random Forest	0.31	0.84	0.45
Stacking Ensemble	(Meta learner) Log. Regression	0.26	0.92	0.40
	Random Forest			
	XGBoost			
	Final Estimator XGBoost			
Optimal F1 Score/Recall Focused	Log. Regression	0.19	0.96	0.32
Optimal F1 Score Focused	Log. Regression	0.27	0.86	0.41
Focal Loss Function	XGBoost	0.6	0.34	0.43
Hyperparameter F1 score	XGBoost	0.39	0.79	0.53



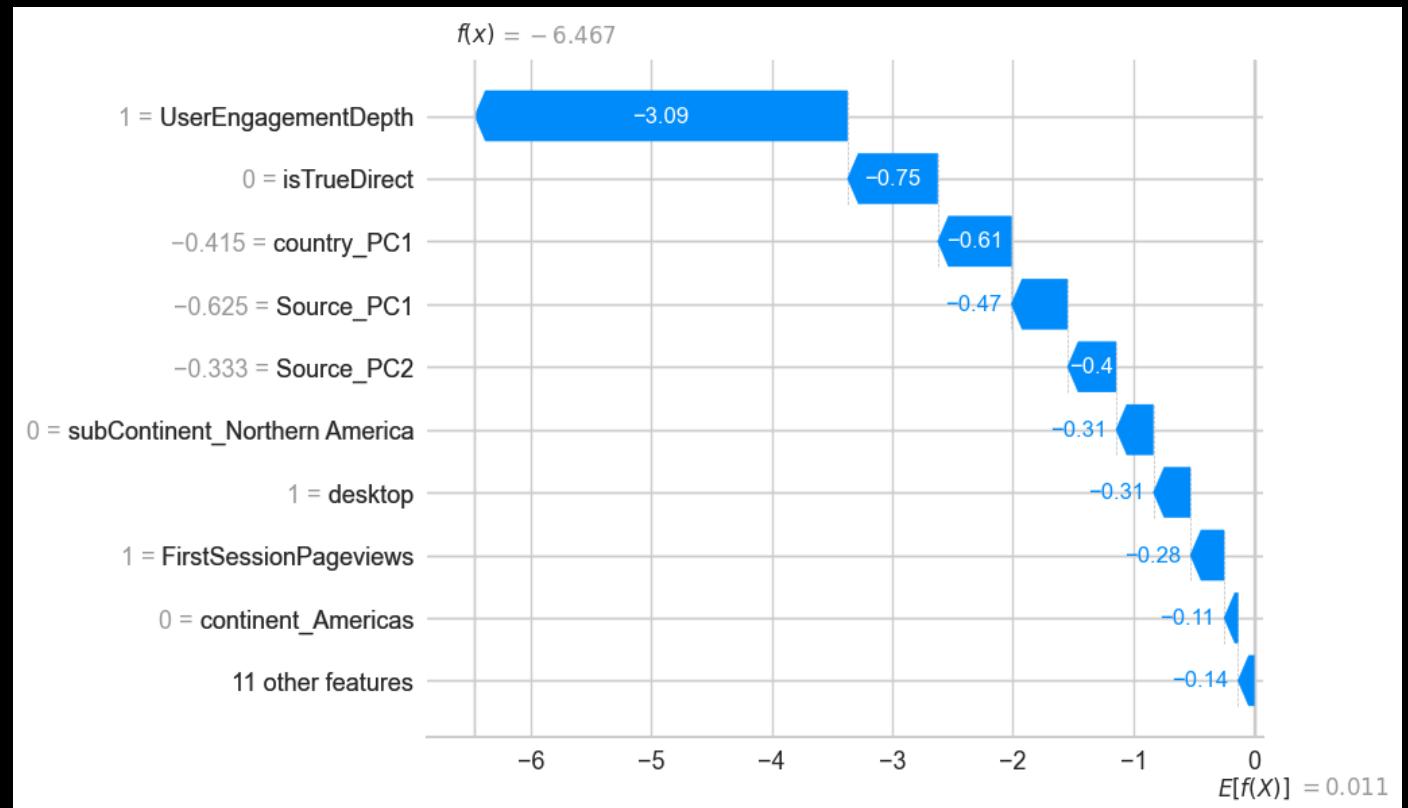
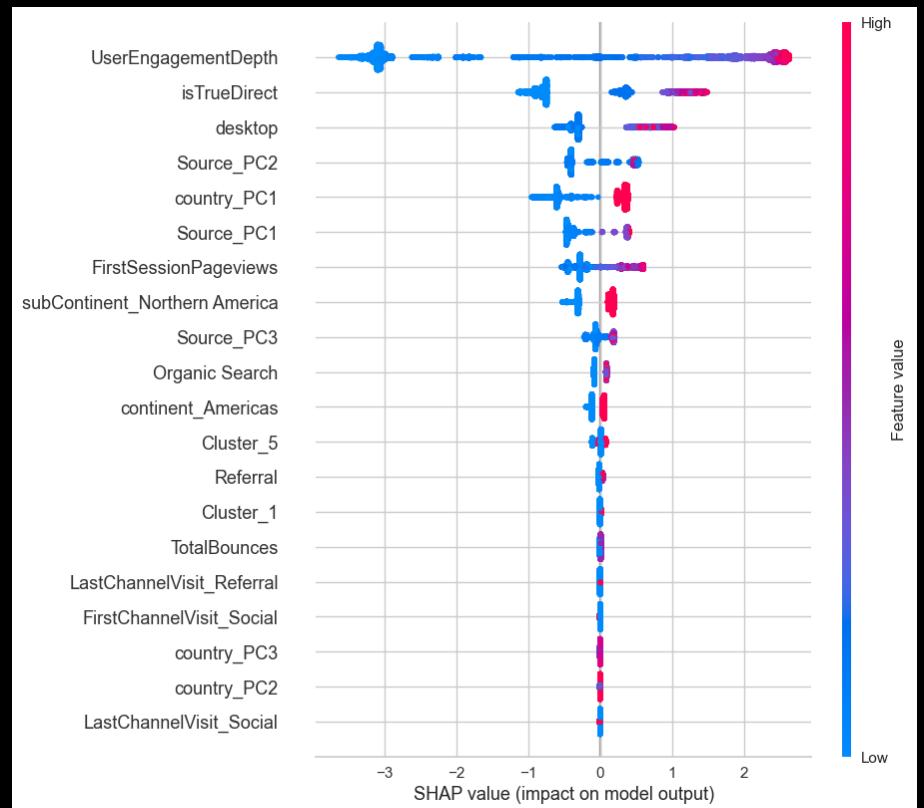
Threshold: 0.89

- Experiments logging with Mlflow
- Hyperparameter tuning with Optuna

Customer Conversion

XGBoost Explainability

- Newly incorporated features, especially User Engagement Depth, display significant importance in determining the model's output



Transaction Revenue

Enterprise 1
Data

Data Cleaning
Expanded the model's geographic scope to include the United States, Canada, and Mexico

Feature Correlation

Feature Engineering
More deep analysis on feature engineering

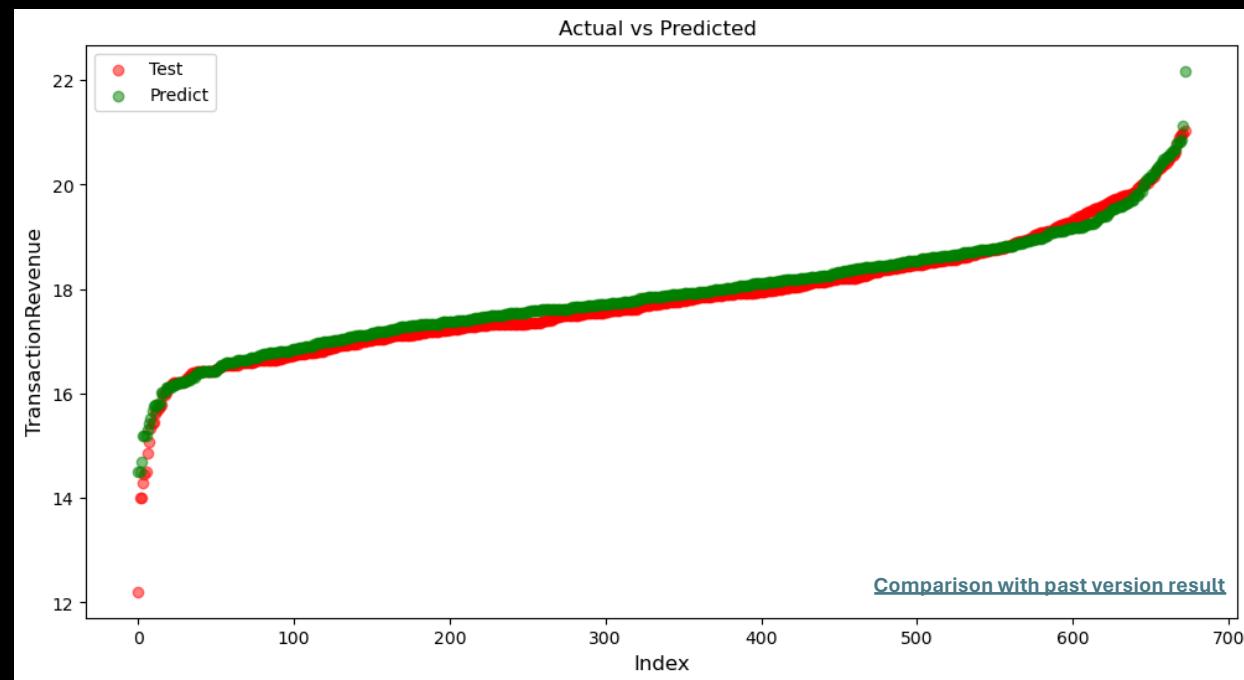
Advanced Model Tuning with AutoML
Integrated AutoML into our tuning processes

T-Pot
ElasticNetCV
Polynomial Regression

Enterprise 2 Model Comparison

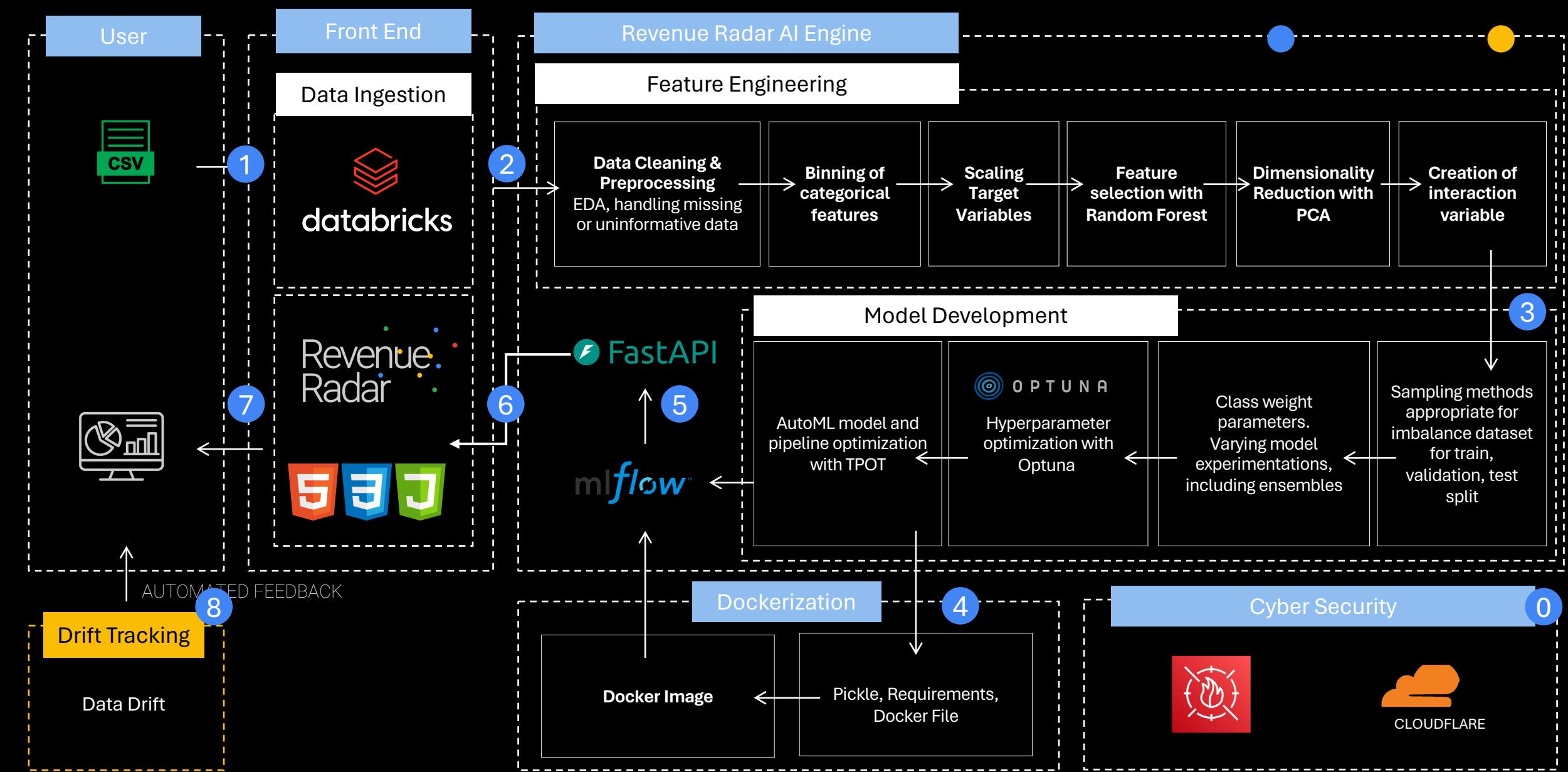
	Enterprise 1 Post Hypermeter Tuning Results	Enterprise 2
Type of the Model	Mean Absolute Error	Mean Absolute Error
Linear Regression	0.863	0.8255
Ridge Regression	0.848	0.8269
Lasso Regression	0.848	0.8259
Random Forest Regressor	0.838	0.8230
XGB Regressor	0.835	0.8217
Gradient Boosting Regressor	0.835	0.8215
AdaBoost Regressor	0.857	0.8322
Decision Tree Regressor	0.854	0.8360
Support Vector Regression	0.832	0.8195 (-1.6%)
AutoML (ElasticNetCV - Polynomial)	-	0.8201 (+0.06%)

TPOT Model Results
AutoML F.L
SVR. F.I.



[See Model Evolution](#)

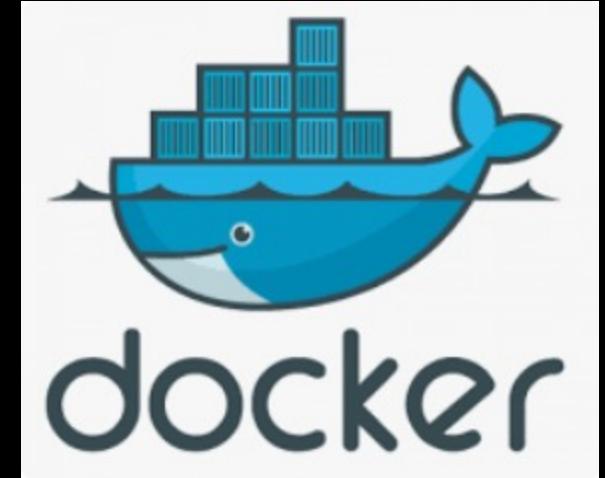
Revenue Radar Architecture



Appendix

Dockerization of ML and Web Applications

- Creating two separate Docker files: one for the MLflow model serving and another for the FastAPI web application
- Use of Docker Compose: Docker Compose orchestrates the deployment of multiple containers, handling the network setup that allows containers to communicate and work together

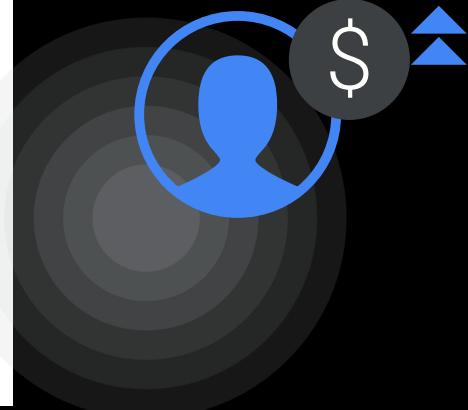


How Revenue Radar impacts the GStore?

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

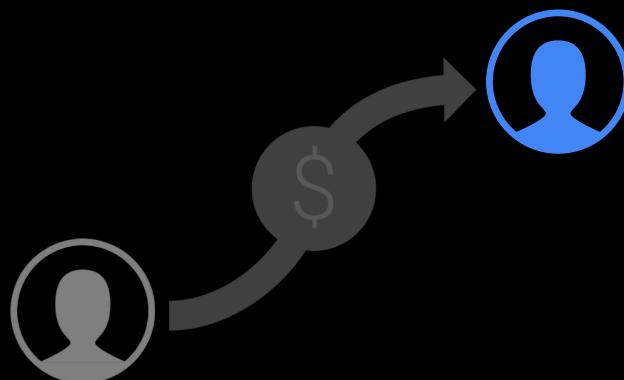
Training and maintenance via offline batch learning over pilot period

Boost number of conversions for a given ad spend budget

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

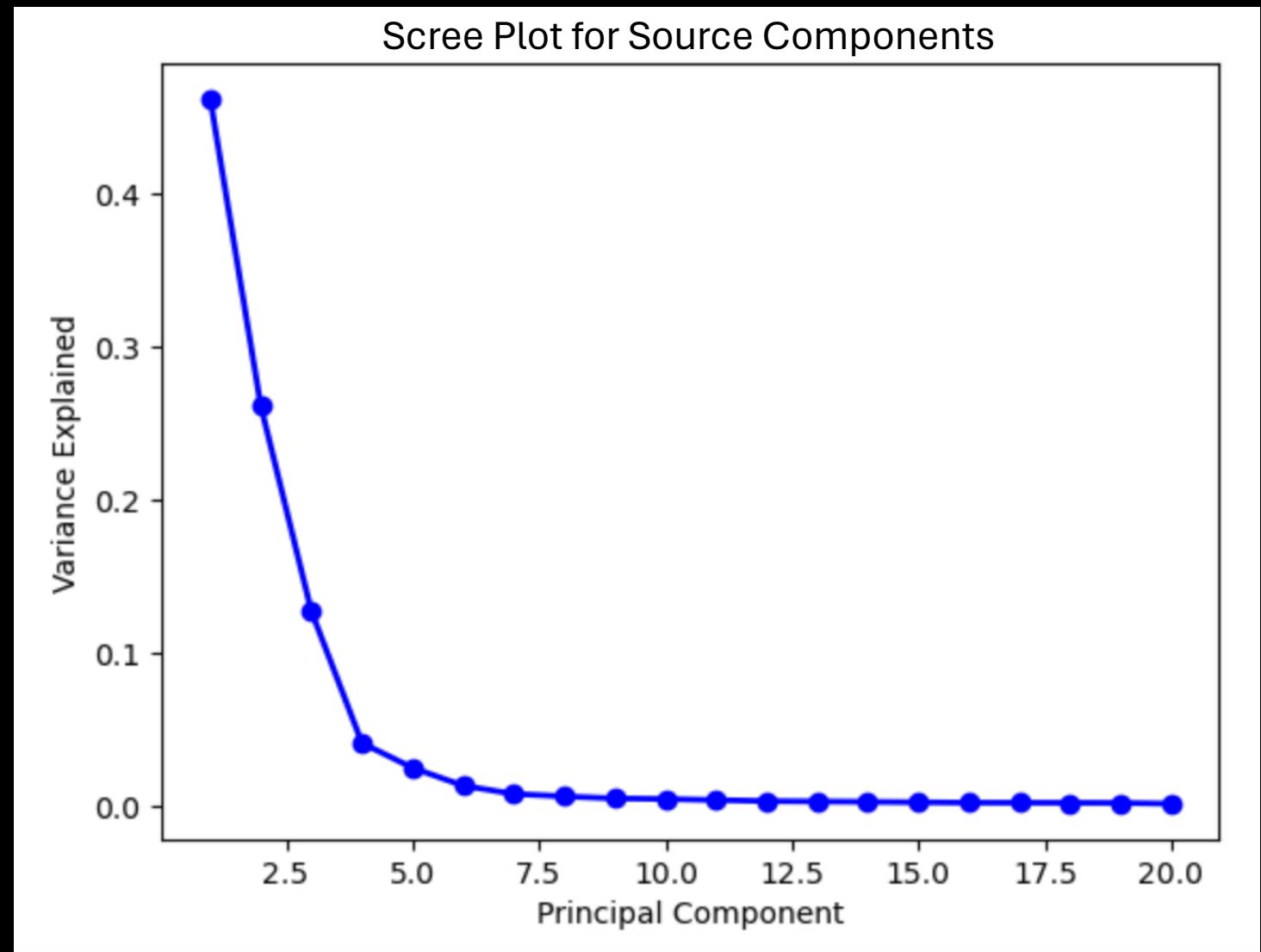
To facilitate user-centric analysis, the dataset was transformed from session level to user level

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

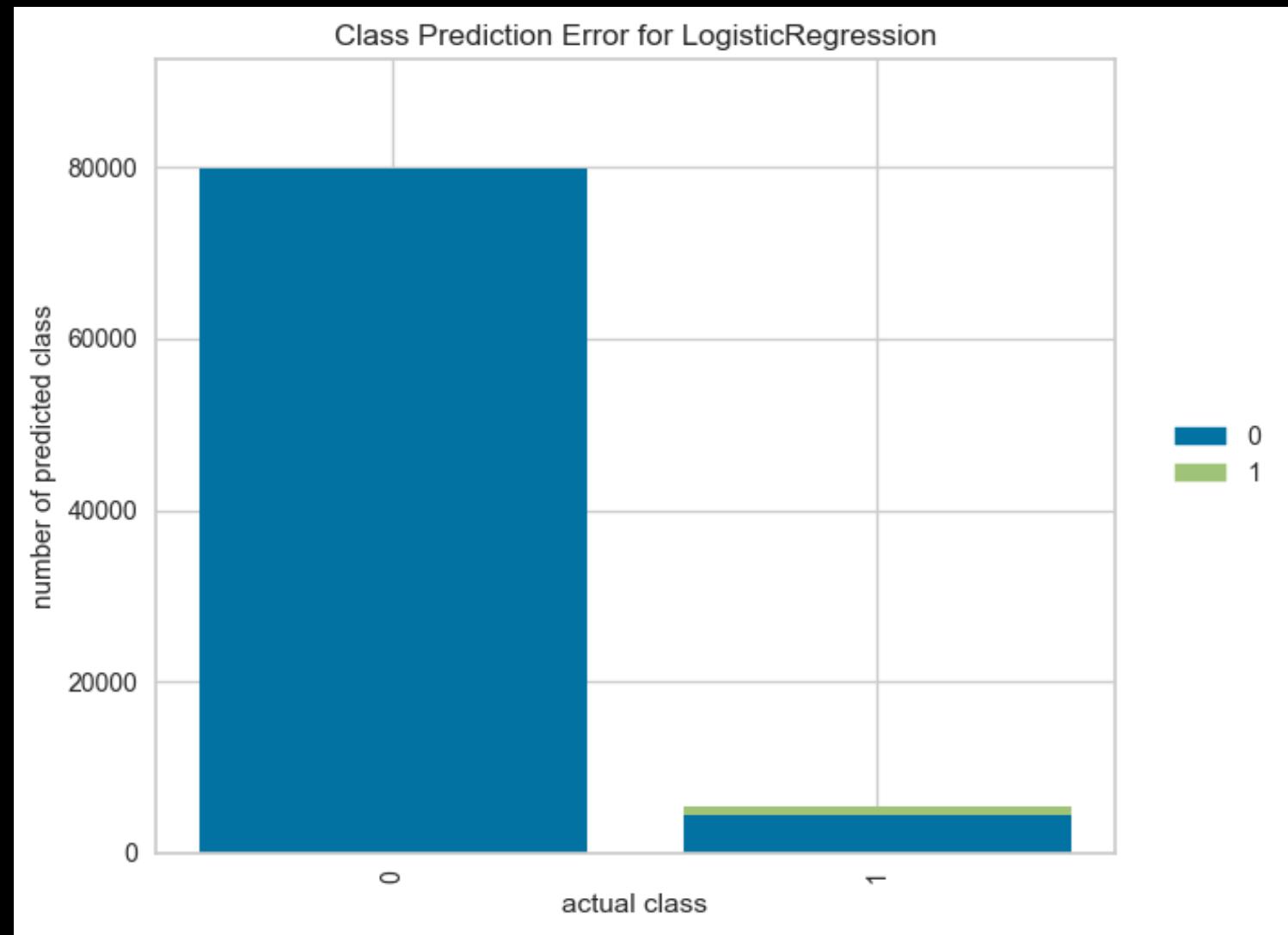
Device Category → Number of visits by desktop, mobile, tablet

Columns excluded based on proportions of missing data and SME consultation

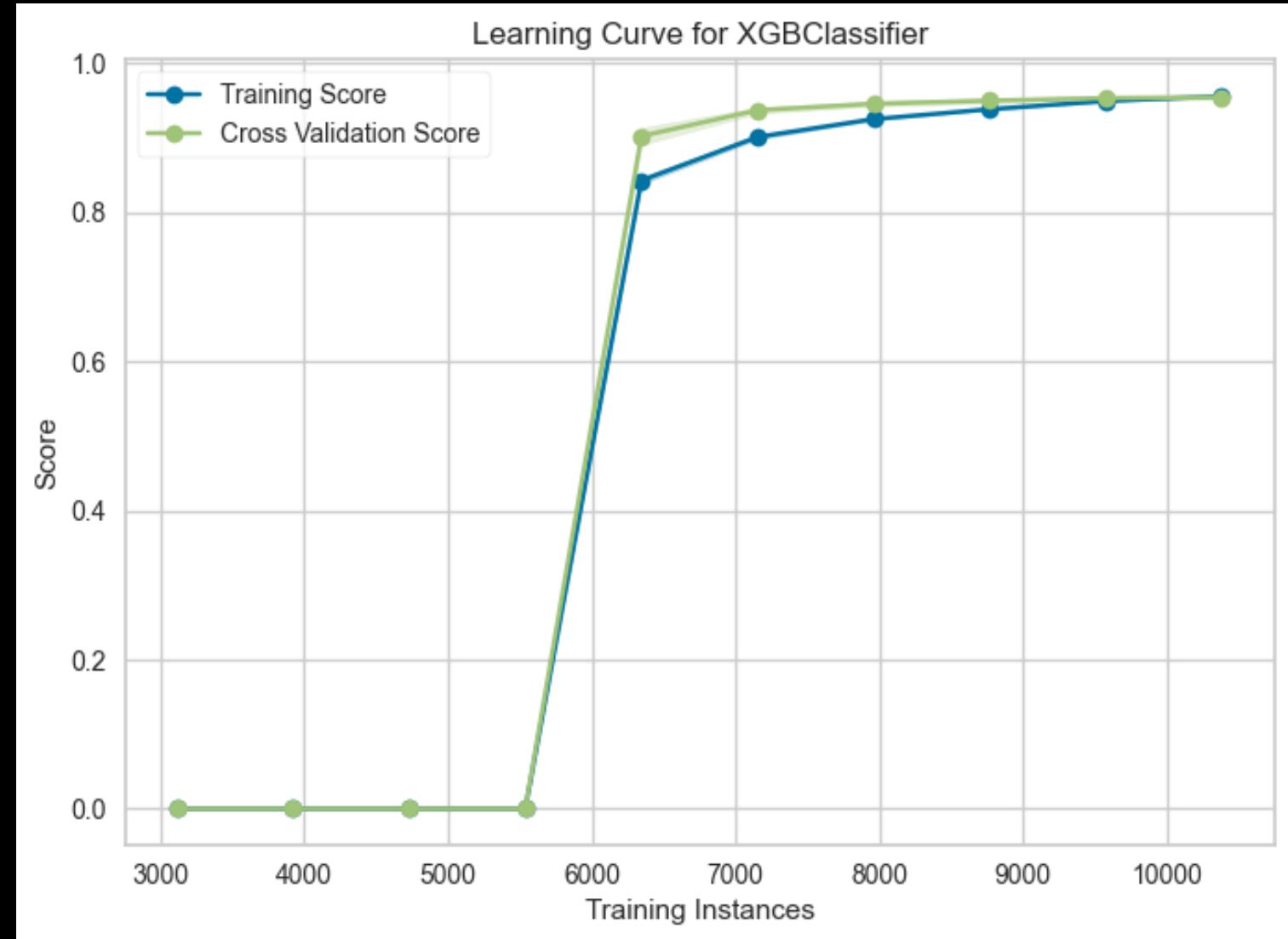
Source Component Number Selection



Logistic Regression Class Prediction Error



XGBClassifier Learning Curve



Estimated Costs & Benefits of Performance

Confusion Matrix Element	Direct Impact Estimate	Description
True Positive (TP)	\$133	Revenue less ad spend
False Positive (FP)	- \$0.75	Ad spend
True Negative (TN)	\$0	No profit / loss impact
False Negative (FN)	\$0	No profit / loss impact

Confusion Matrix Element	Expanded Impact Estimate	Description
FP + FN	-\$134.50	Ad spend + opportunity cost

Assumptions:

- \$0.75 for Social Media Advertisement (CPC) [Source: K6 Media Agency](#)
- Average order value of \$133.75 derived from data
- True Positives convert when nudged
- Fixed advertising budget which must be dispensed when lead identified (use it or lose it organizational policy) but has opportunity costs when dispensed incorrectly

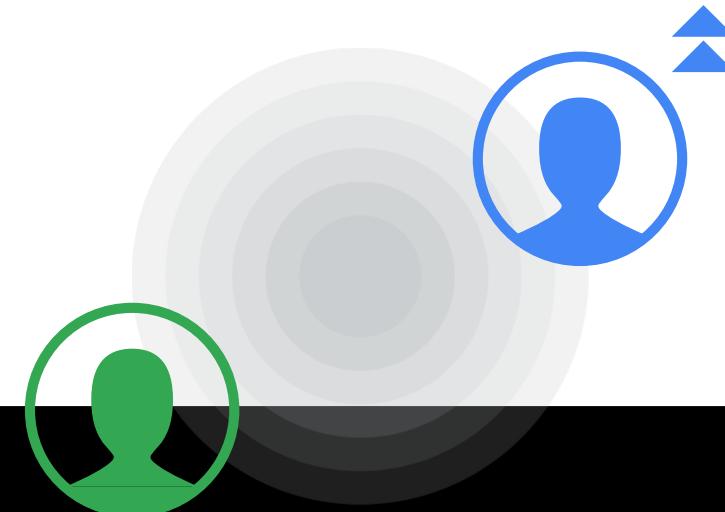
Transaction Revenue

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

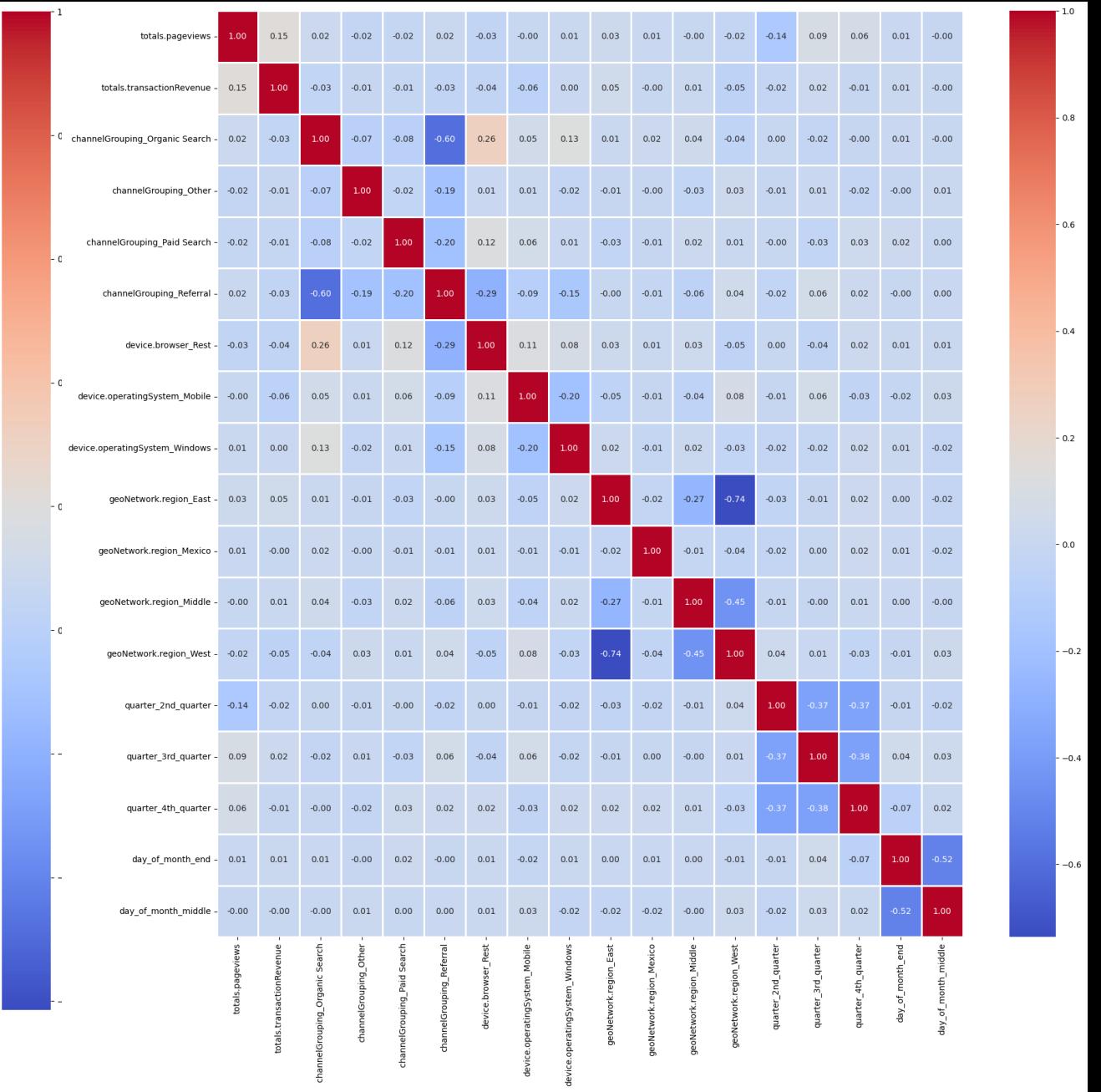
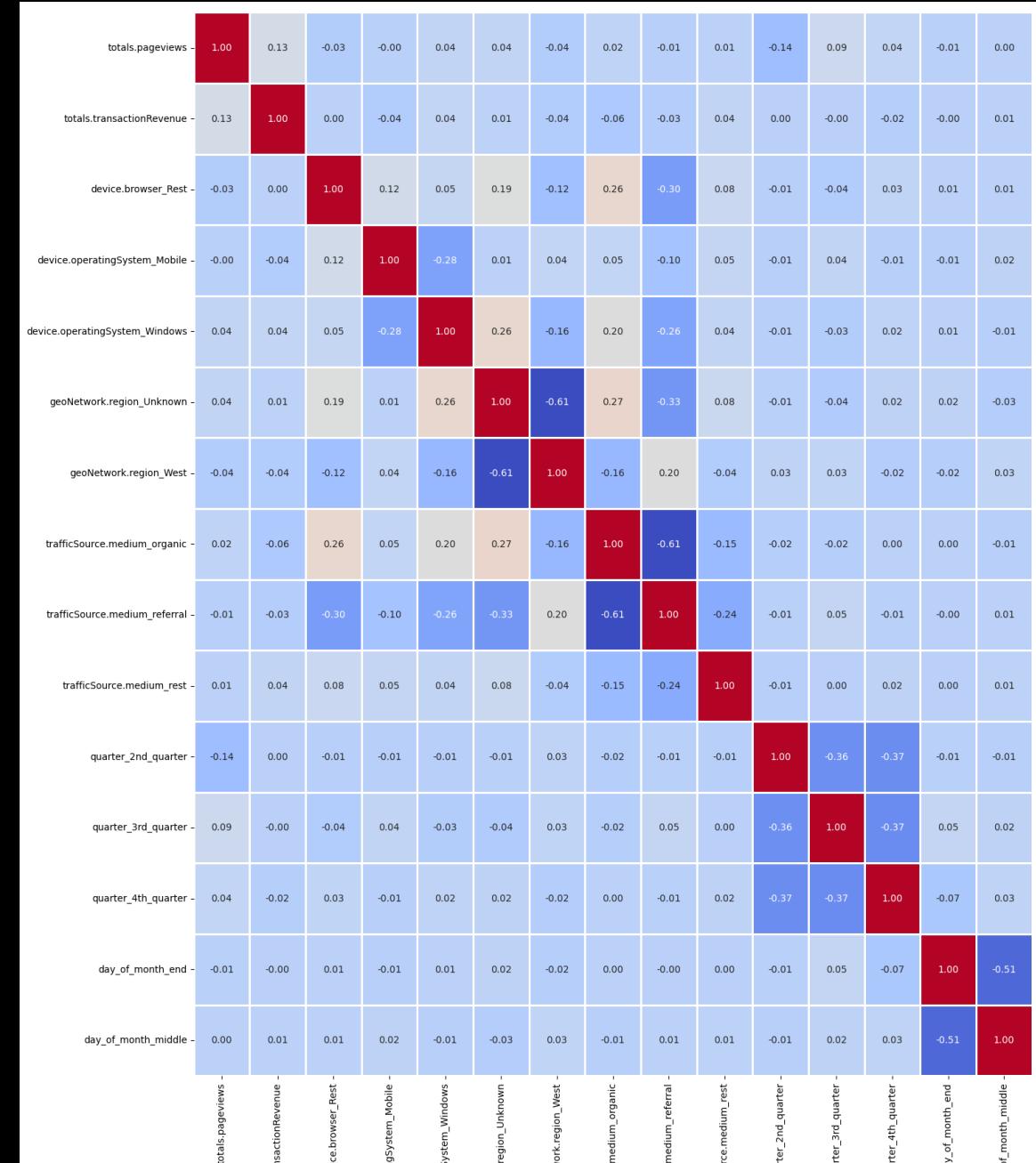
To facilitate transaction level analysis, the dataset was filtered for session with transaction revenue == 0

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

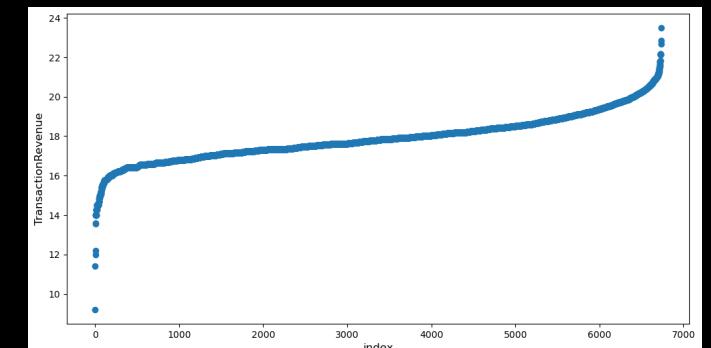
Before

After

Back 



Model Evolution



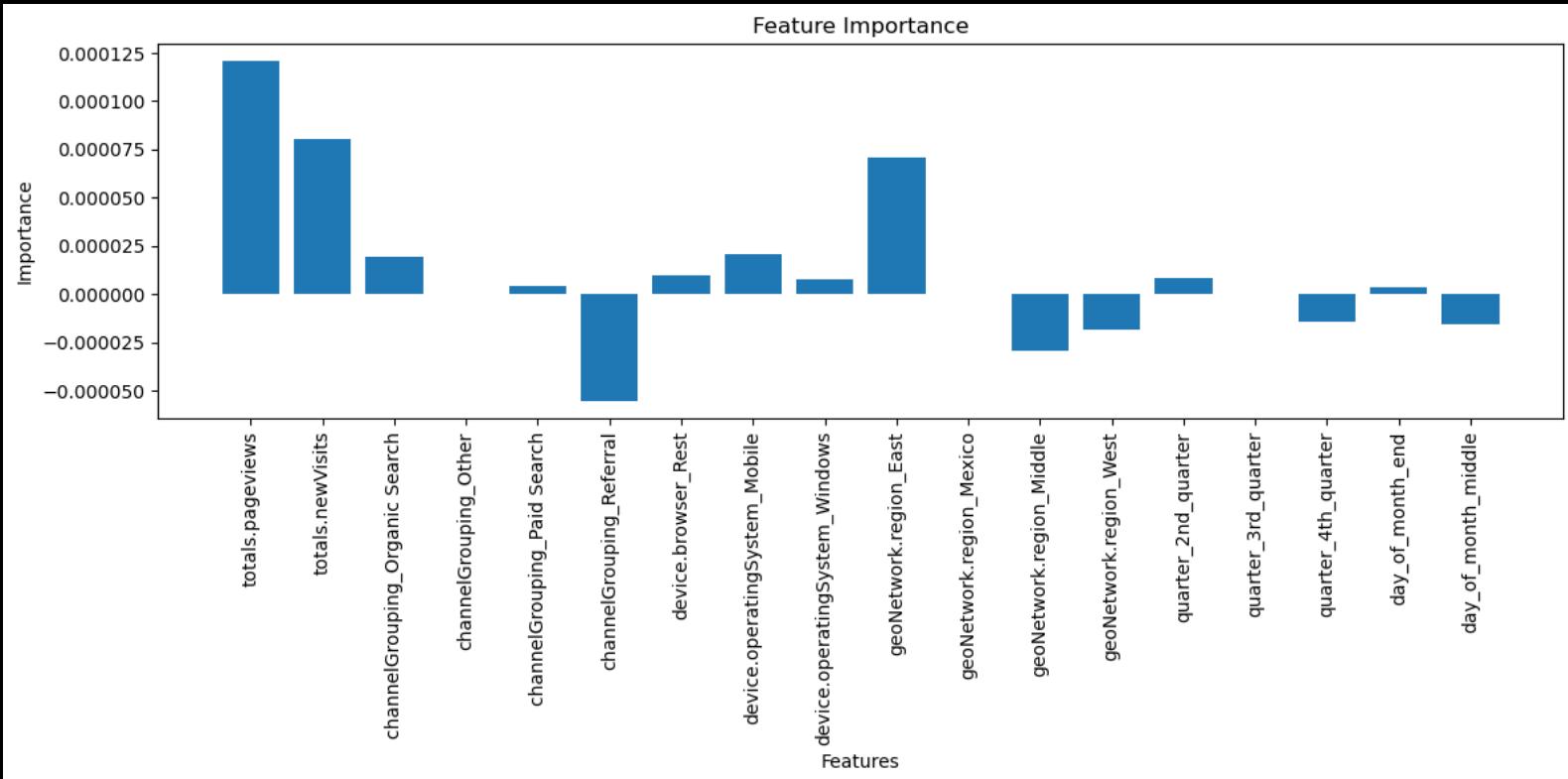
Initial Stage



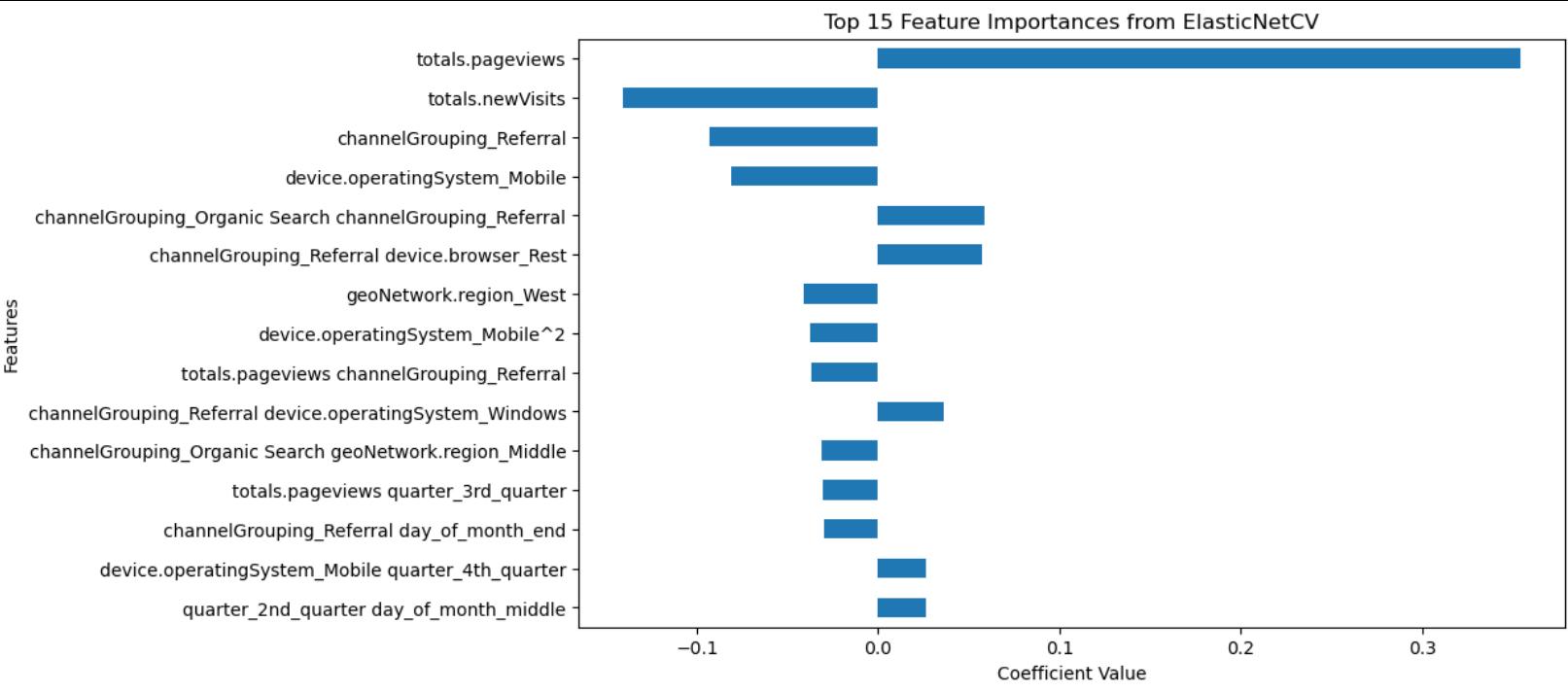
Final Model

Back

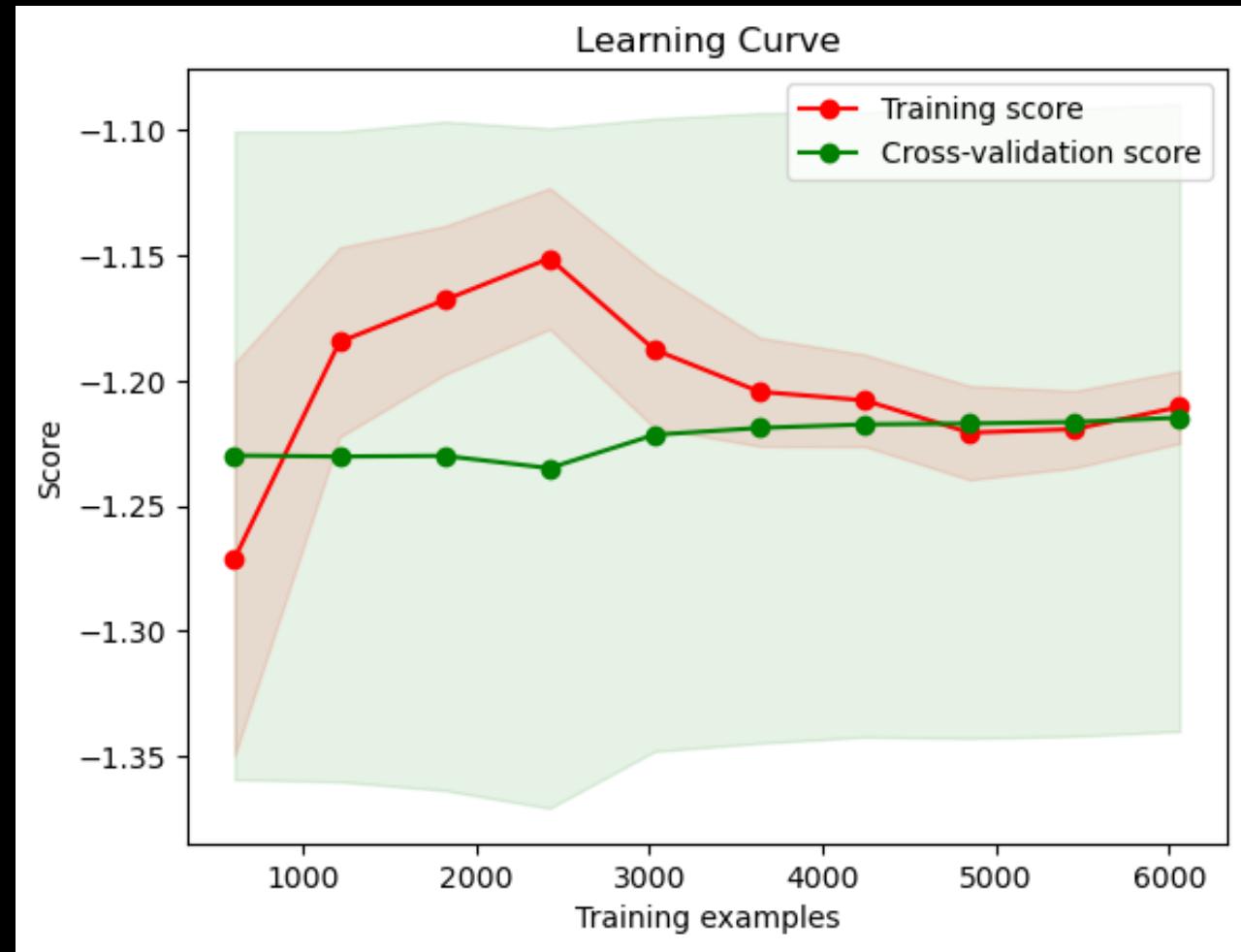
Support Vector Regression



AutoML Feature Importance

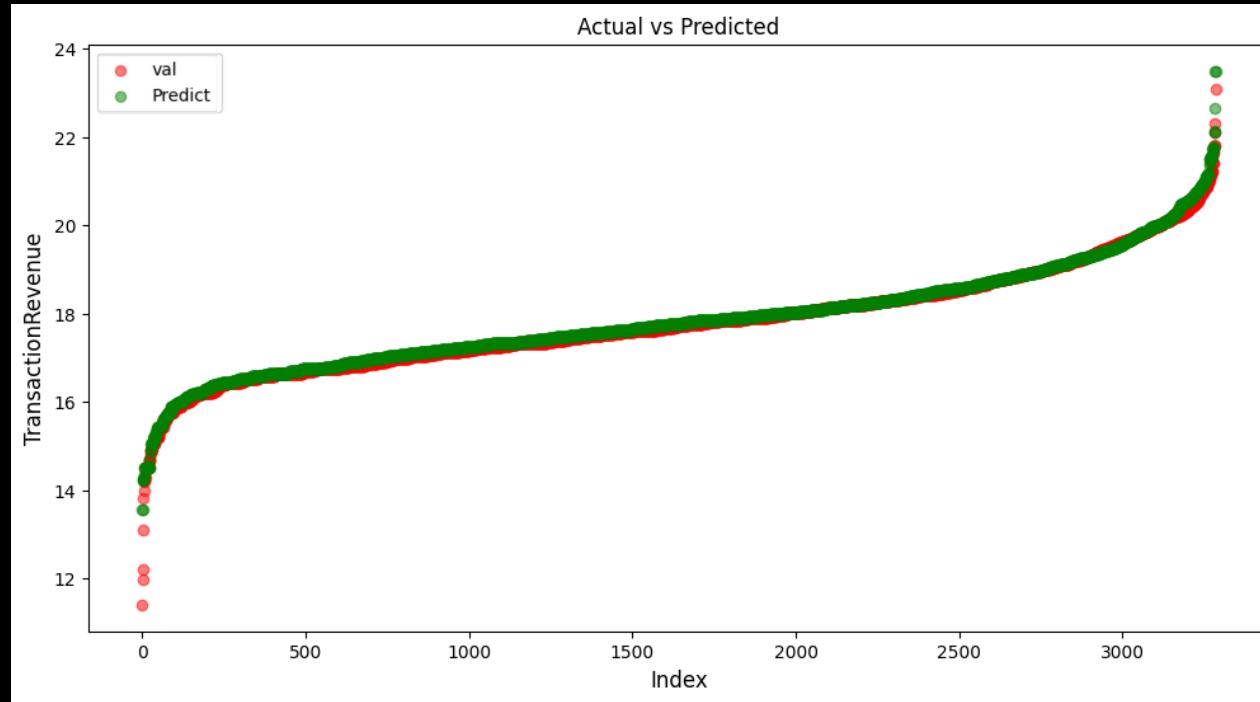


Learning Curve for AutoML



Back

Old Deviation Plot between Val and Predict



AutoML Pipeline

Pipeline ⓘ ⓘ

PolynomialFeatures ⓘ
PolynomialFeatures(include_bias=False)

ElasticNetCV ⓘ
ElasticNetCV(l1_ratio=0.3500000000000003, random_state=42, tol=0.01)

Elastic Net model with iterative fitting along a regularization path.

L1_ratio: float or list of float. Float between 0 and 1 passed to ElasticNet (scaling between L1 and L2 penalties). For $0 < \text{l1_ratio} < 1$, the penalty is a combination of L1 and L2

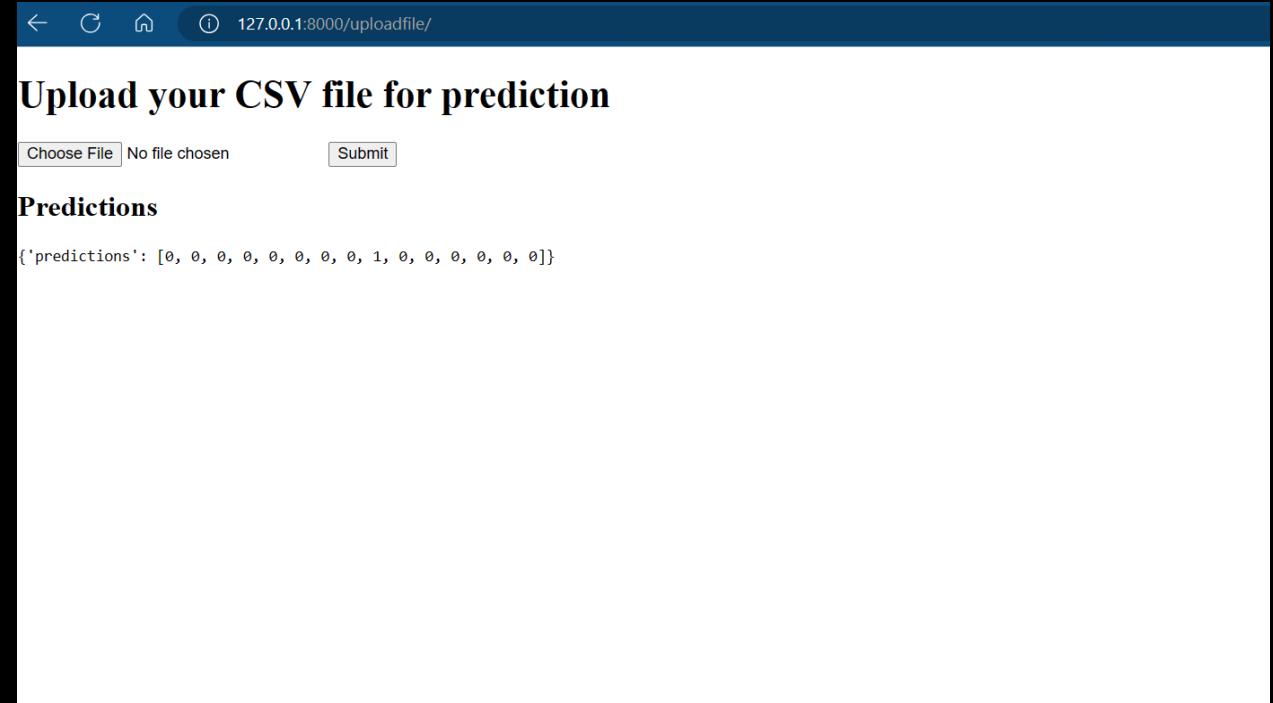
Tol: The tolerance for the optimization: if the updates are smaller than tol, the optimization code checks the dual gap for optimality and continues until it is smaller than tol.

Dataset

#	Column	Non-Null Count	Dtype
0	channelGrouping	903653	non-null object
1	date	903653	non-null int64
2	fullVisitorId	903653	non-null object
3	sessionId	903653	non-null object
4	socialEngagementType	903653	non-null object
5	visitId	903653	non-null int64
6	visitNumber	903653	non-null int64
7	visitStartTime	903653	non-null int64
8	continent	903653	non-null object
9	subContinent	903653	non-null object
10	country	903653	non-null object
11	region	903653	non-null object
12	metro	903653	non-null object
13	city	903653	non-null object
14	cityId	903653	non-null object
15	networkDomain	903653	non-null object
16	latitude	903653	non-null object
17	longitude	903653	non-null object
18	networkLocation	903653	non-null object
19	browser	903653	non-null object
20	browserVersion	903653	non-null object
21	browserSize	903653	non-null object
22	operatingSystem	903653	non-null object
23	operatingSystemVersion	903653	non-null object
24	isMobile	903653	non-null bool
25	mobileDeviceBranding	903653	non-null object
26	mobileDeviceModel	903653	non-null object
27	mobileInputSelector	903653	non-null object
28	mobileDeviceInfo	903653	non-null object
29	mobileDeviceMarketingName	903653	non-null object
30	flashVersion	903653	non-null object
31	language	903653	non-null object
32	screenColors	903653	non-null object
33	screenResolution	903653	non-null object
34	deviceCategory	903653	non-null object
35	visits	903653	non-null int64
36	hits	903653	non-null int64
37	pageviews	903553	non-null float64
38	bounces	450630	non-null float64
39	newVisits	703060	non-null float64
40	transactionRevenue	11515	non-null float64
41	campaign	903653	non-null object
42	source	903653	non-null object
43	medium	903653	non-null object
44	keyword	400724	non-null object
45	adwordsClickInfo.criteriaParameters	903653	non-null object
46	isTrueDirect	274005	non-null object
47	referralPath	330941	non-null object
48	adwordsClickInfo.page	21460	non-null float64
49	adwordsClickInfo.slot	21460	non-null object
50	adwordsClickInfo.gclId	21561	non-null object
51	adwordsClickInfo.adNetworkType	21460	non-null object
52	adwordsClickInfo.isVideoAd	21460	non-null object
53	adContent	10946	non-null object
54	campaignCode	1	non-null object

Back

MVP Web App



MLFlow Experimentation

The screenshot shows the MLflow Experimentation interface. On the left, there's a sidebar titled "Experiments" with a search bar and three entries: "Default", "classification" (which is selected and highlighted in blue), and "classification_experiment". On the right, the main area is titled "classification" and displays a table of experimental runs. The table has columns for "Run Name", "Created", "Dataset", "Duration", "Source", and "Models". The data shows various models like XGBoost and Random Forest being run multiple times with different configurations.

Run Name	Created	Dataset	Duration	Source	Models
XGBoost Model with thre...	1 day ago	-	12.7s	c:\Users\t...	XGBoost/1
XGBoost with Loss Function	1 hour ago	-	23.1s	c:\Users\t...	xgboost
XGBoost Model with thre...	1 day ago	-	12.7s	c:\Users\t...	xgboost
Random Forest Model wit...	1 day ago	-	30.6s	c:\Users\t...	sklearn
Stacking Classifier with Th...	1 day ago	-	1.5min	c:\Users\t...	-
Random Forest Model wit...	1 day ago	-	37.4s	c:\Users\t...	sklearn
Stacking Classifier Optimizi...	1 day ago	-	1.6min	c:\Users\t...	sklearn
XGBoost Model Optimizatio...	1 day ago	-	12.8s	c:\Users\t...	sklearn
Random Forest Model Opti...	1 day ago	-	1.6min	c:\Users\t...	sklearn

This screenshot shows the "Logistic Regression Model Recall" page within the MLflow interface. It includes sections for "Overview", "Model metrics", "System metrics", and "Artifacts". The "Artifacts" tab is active, displaying a tree view of files: "best_model_logistic_f1" (containing "metadata", "MLmodel", "conda.yaml", "model.pkl", "python_env.yaml", and "requirements.txt") and "dataset_info" (containing "dataset_info.json"). To the right, there's a detailed view of the "best_model_logistic_f1" artifact, including its path and a section titled "MLflow Model" with code snippets for making predictions using a Spark DataFrame.

Model schema
Input and output schema for your model. [Learn more](#)

Name	Type
No schema. See MLflow docs for how to include input and output schema with your model.	

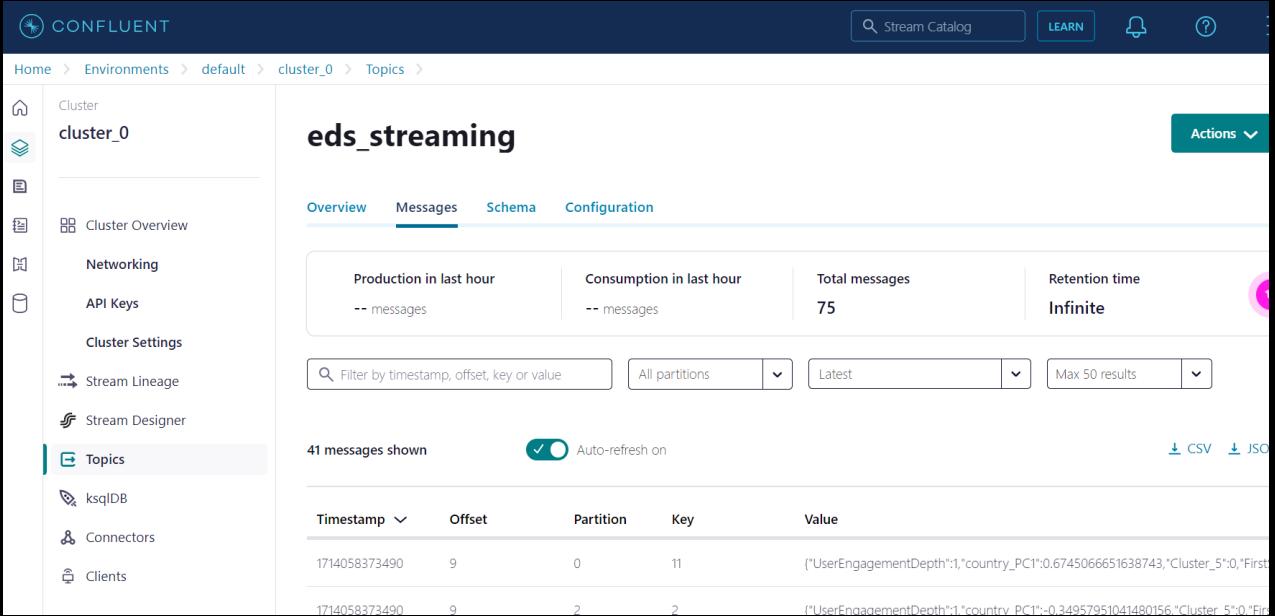
Make Predictions
Predict on a Spark DataFrame:

```
import mlflow
from pyspark.sql.functions import struct, col
logged_model = 'runs:/a036502ca14e46f5a720f9cb5cc68549/best_model_logistic_f1'

# Load model as a Spark UDF. Override result_type if the model does not return double values.
loaded_model = mlflow.pyfunc.spark_udf(spark, model_uri=logged_model, result_type='double')

# Predict on a Spark DataFrame.
```

Batch Streaming



```
1 from pyspark.sql import SparkSession
2 from pyspark.sql.functions import monotonically_increasing_id, to_json, struct, col
3
4 # Initialize Spark Session
5 spark = SparkSession.builder \
6     .appName("KafkaDataWriter") \
7     .getOrCreate()
8
9 # Assume X_test_spark is already loaded as a Spark DataFrame
10 X_test_spark = X_test_spark.withColumn("key", monotonically_increasing_id())
11 X_test_spark = X_test_spark.withColumn("value", to_json(struct([col(c) for c in X_test_spark.columns if c != 'key'])))
12
13 # Prepare data for Kafka
14 kafka_ready_df = X_test_spark.selectExpr("CAST(key AS STRING) AS key", "CAST(value AS STRING) AS value")
15
16 # Write to Kafka
17 api_key = "AO35DNAQ4YOSZUR"
18 api_secret = "Jtslhj091q1wes/lo6AN/J4sRorMMvEbAxTM7Oo37IkPOin3F55AHUwZhOE02693"
19 bootstrap_servers = "pkc-56d1g.eastus.azure.confluent.cloud:9092"
20 topic_name = "eds_streaming"
21
22 kafka_ready_df.write.format("kafka") \
23     .option("kafka.bootstrap.servers", bootstrap_servers) \
24     .option("topic", topic_name) \
25     .option("kafka.security.protocol", "SASL_SSL") \
26     .option("kafka.sasl.mechanism", "PLAIN") \
27     .option("kafka.sasl.jaas.config", f"kafkashaded.org.apache.kafka.common.security.plain.PlainLoginModule required username='{api_key}' password='{api_secret}';") \
28     .save()
29
```

Data Drift Investigation

Feature Drift Risks from Upstream Data Structure Changes	Data Drift Risks from Seasonality
<ul style="list-style-type: none">Columns becoming deprecated (e.g. networkDomain, networkLocation, isMobile identified during revised EDA)Transformations applied to numerical features being changed (e.g. currently Transaction Revenue is multiplied by 10^6 when passed to Google Analytics)<ul style="list-style-type: none">For the regression model this would also represent Target Drift	<ul style="list-style-type: none">Experiencing seasonality from peak shopping periods. (e.g. products as gifts during winter holidays), which could affect statistical properties of the input data.For the conversion use case which involves aggregating user data over a period, it could impact the relationship between the prediction data and the performance period data, constituting a form of concept drift
<i>Detection Options</i>	<i>Detection Options</i>
<ul style="list-style-type: none">Validation of expected data structure at point of ingestion into pipelinePeriodic reviews of upstream data practices	<ul style="list-style-type: none">Statistical techniques (e.g. KL Divergence), ML based tests (ability of model to classify input datasets)Overall model performance tracking compared to actuals
<i>Mitigation Options</i>	<i>Mitigation Options</i>
<ul style="list-style-type: none">Regularly update and re-engineer features to align with the current data environment and business contextImplement data versioning for tracking changesAdjust models to be robust against missing features, potentially using fallback rules for imputation until a more resilient fix can be implemented	<ul style="list-style-type: none">Incorporate time-based features that can help capture seasonalityRetrain the model with more historical examples of seasonality effects