Predicting Online Shoppers Purchasing Intention

Faisal Al-Shammari

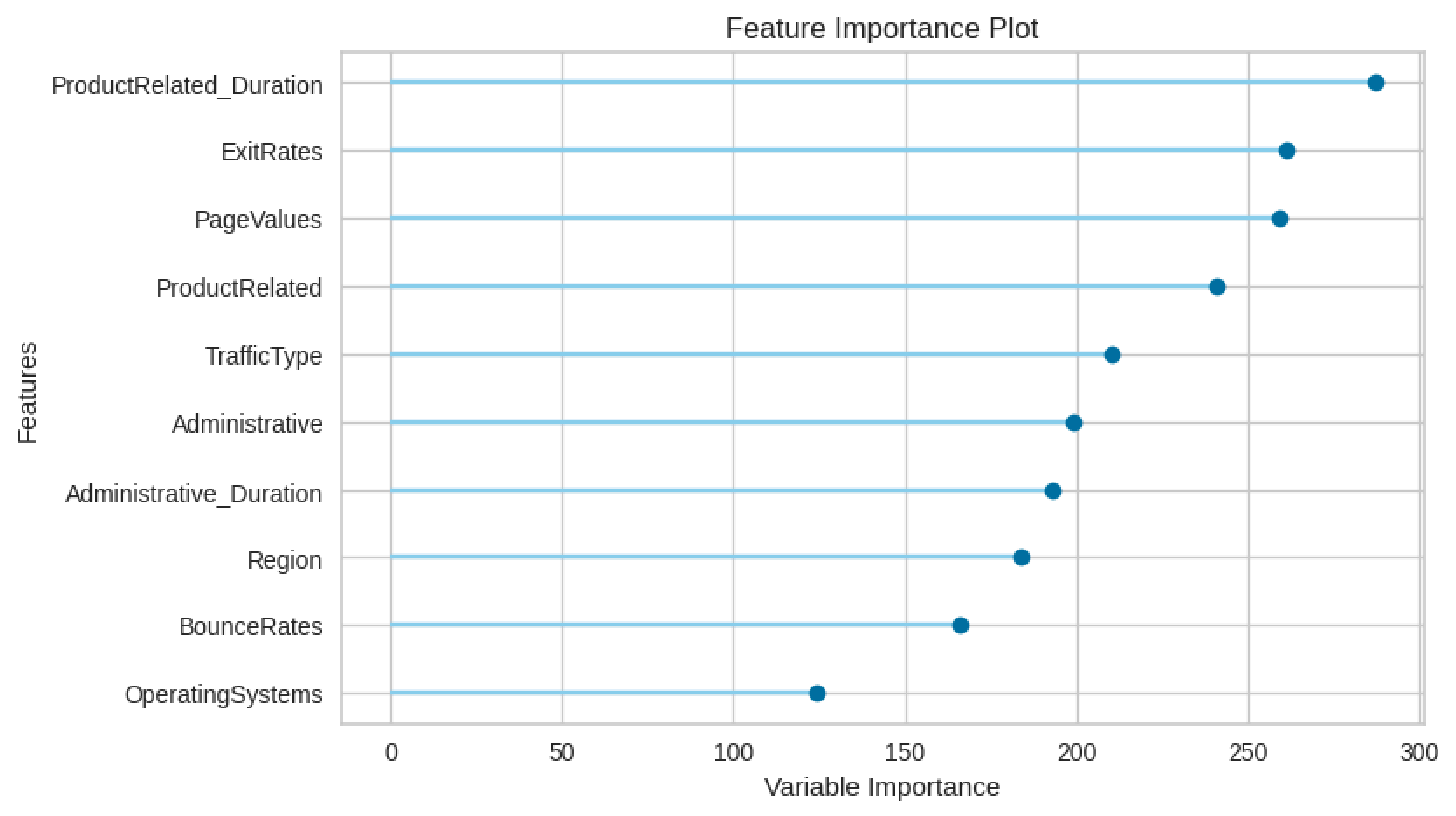
Sulaiman Alluhaib

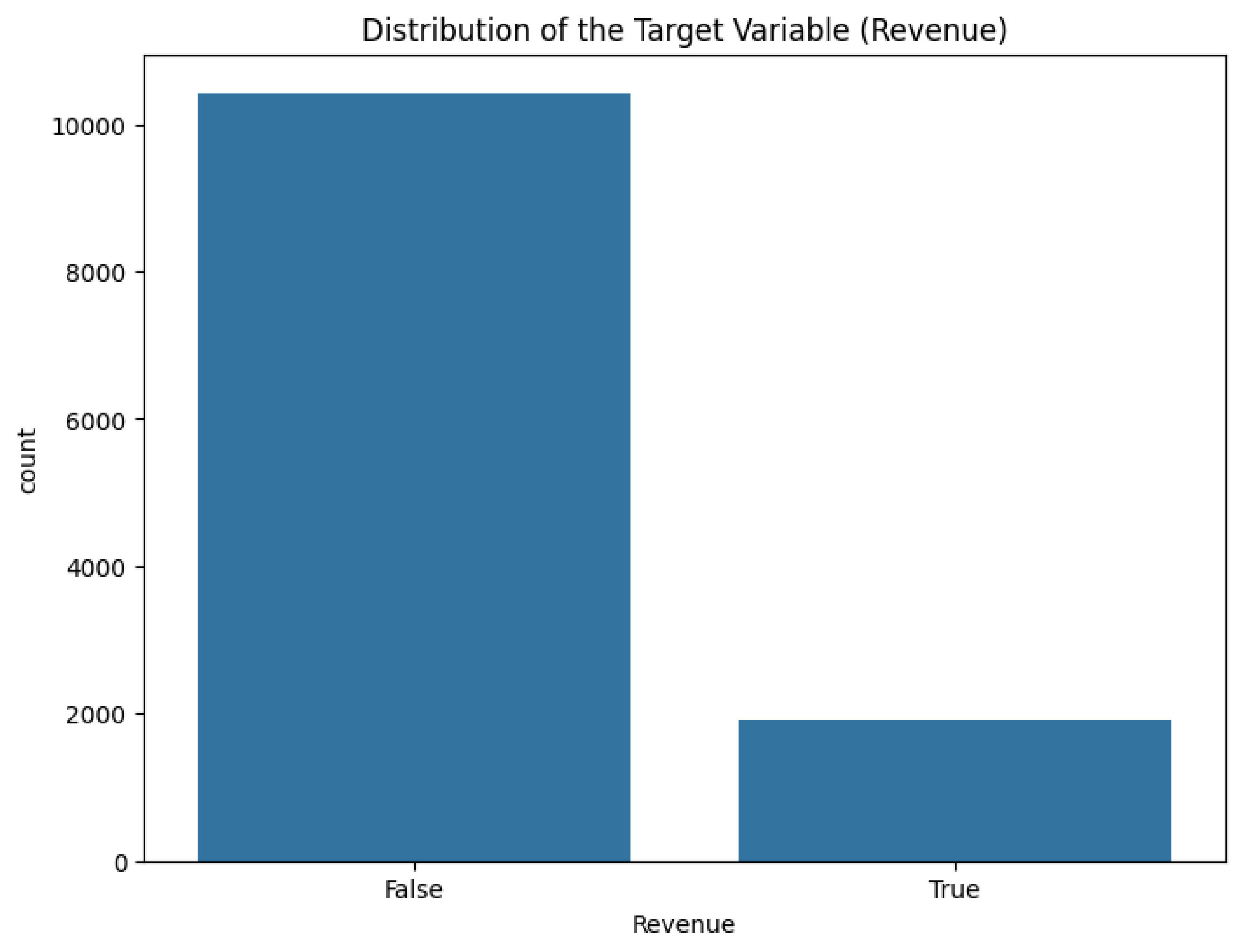
Project Overview

The primary objective of this project is to provide a model that accurately predicts whether a visitor will make a purchase on an e-commerce website based on their browsing behavior and session information.

By understanding these patterns, businesses can enhance user engagement and optimize conversion rates.

Data Summary





Exploratory Data Analysis (EDA)

1. ProductRelated and ProductRelated_Duration:
   - Strong positive correlation indicates more product page views lead to more time spent.

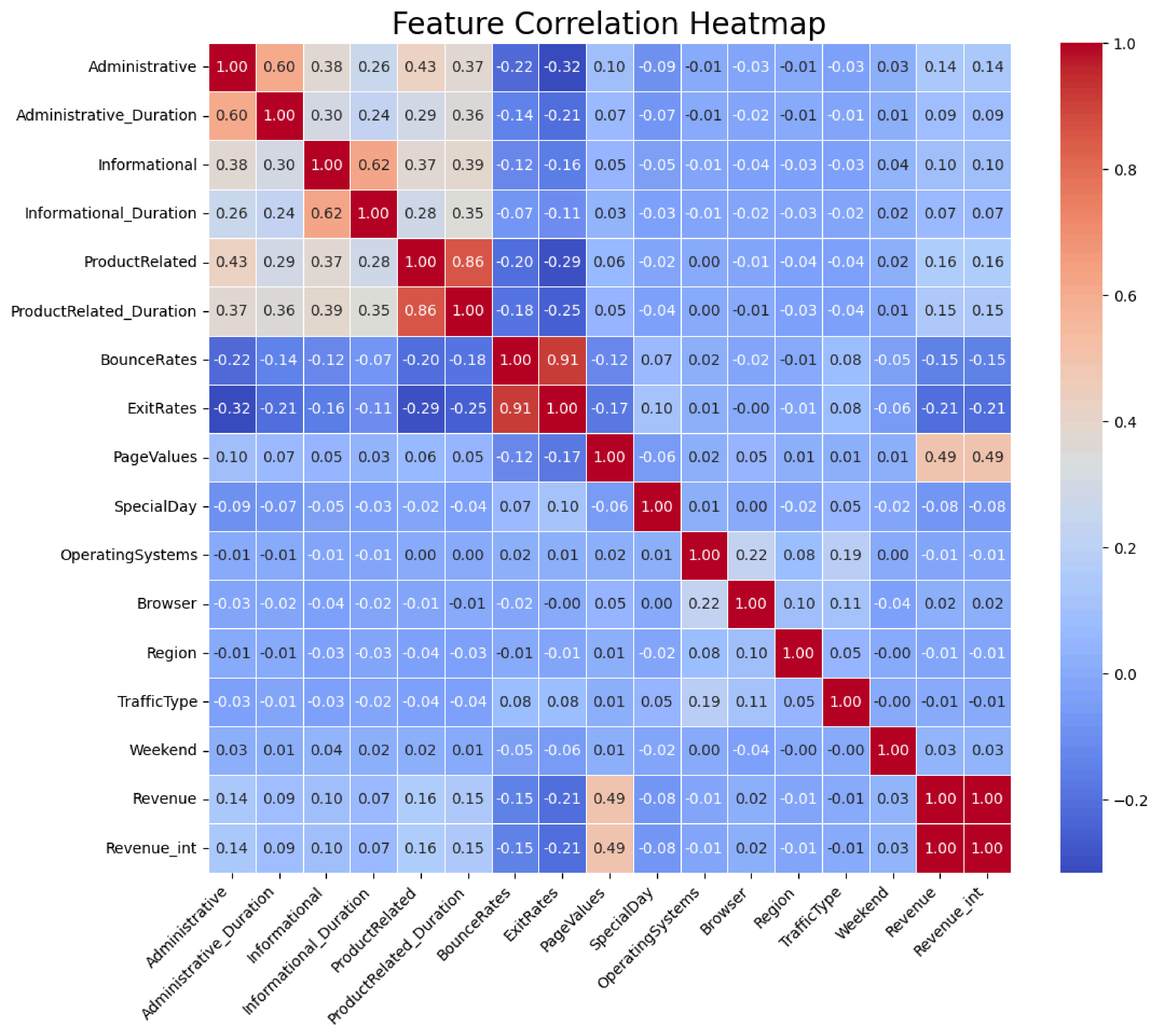
2. BounceRates and ExitRates:
   - High positive correlation suggests pages with high bounce rates also have high exit rates.

3. BounceRates and PageValues:
   - Moderate negative correlation shows valuable pages have lower bounce rates.

4. ExitRates and PageValues:
   - Moderate negative correlation indicates valuable pages have lower exit rates.

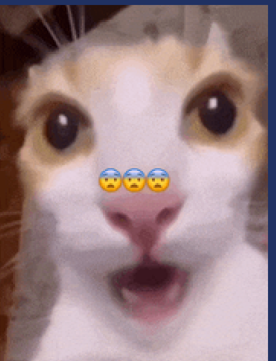
5. PageValues and Revenue:
   - Moderate positive correlation reveals higher page values lead to more purchases.

6. Other Features:
   - Significant correlations among Administrative, Informational views, and durations indicate navigation patterns.

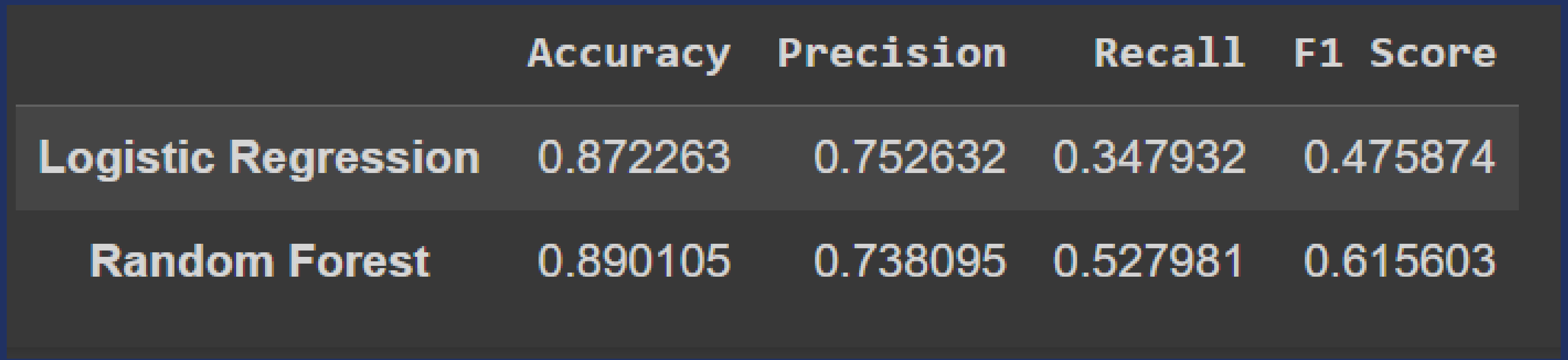


Data Preprocessing

- Handling Missing Values: Ensured data completeness by verifying no missing values were present.
- Encoding Categorical Variables: Transformed categorical variables using one-hot and ordinal encoding.
- Normalizing Features: Standardized numerical features to have a mean of 0 and standard deviation of 1.
- Transformation: Applied transformations to reduce skewness in feature distributions.
- Handling Class Imbalance: Used SMOTE to generate synthetic samples for the minority class.
- Removing Multicollinearity: Excluded highly correlated features to improve model performance.



Initial Model Performance



	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.872263	0.752632	0.347932	0.475874
Random Forest	0.890105	0.738095	0.527981	0.615603

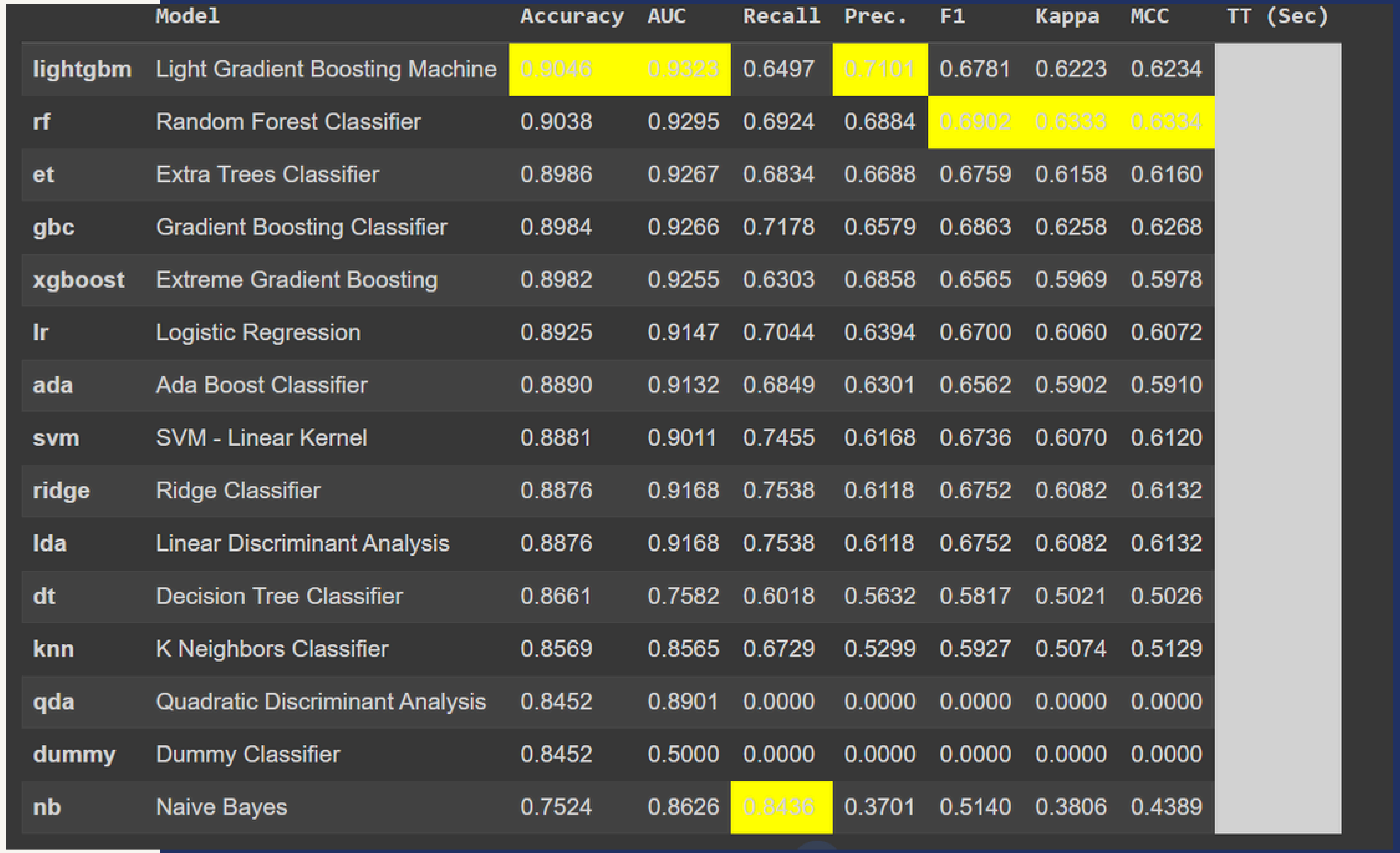
- Logistic Regression: A simple yet effective model for binary classification problems.

- Random Forest:
  - An ensemble method that builds multiple decision trees and merges them to get a more accurate and stable prediction.
  - In the initial testing, Random Forest performed better than Logistic Regression.

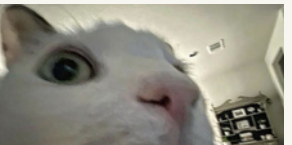
- Comparison of Results:
  - Accuracy: Random Forest outperformed Logistic Regression in terms of accuracy, indicating it was better at correctly predicting both classes.
  - Precision:Logistic Regression had a slightly higher precision, meaning it was better at predicting true positives (purchases) among the predicted positives.
  - Recall:Random Forest had a significantly higher recall, indicating it was better at capturing the actual positives (purchases) from the dataset.
  - F1 Score:The F1 score, which balances precision and recall, was higher for Random Forest, making it the more robust model overall.

Using Pycarot

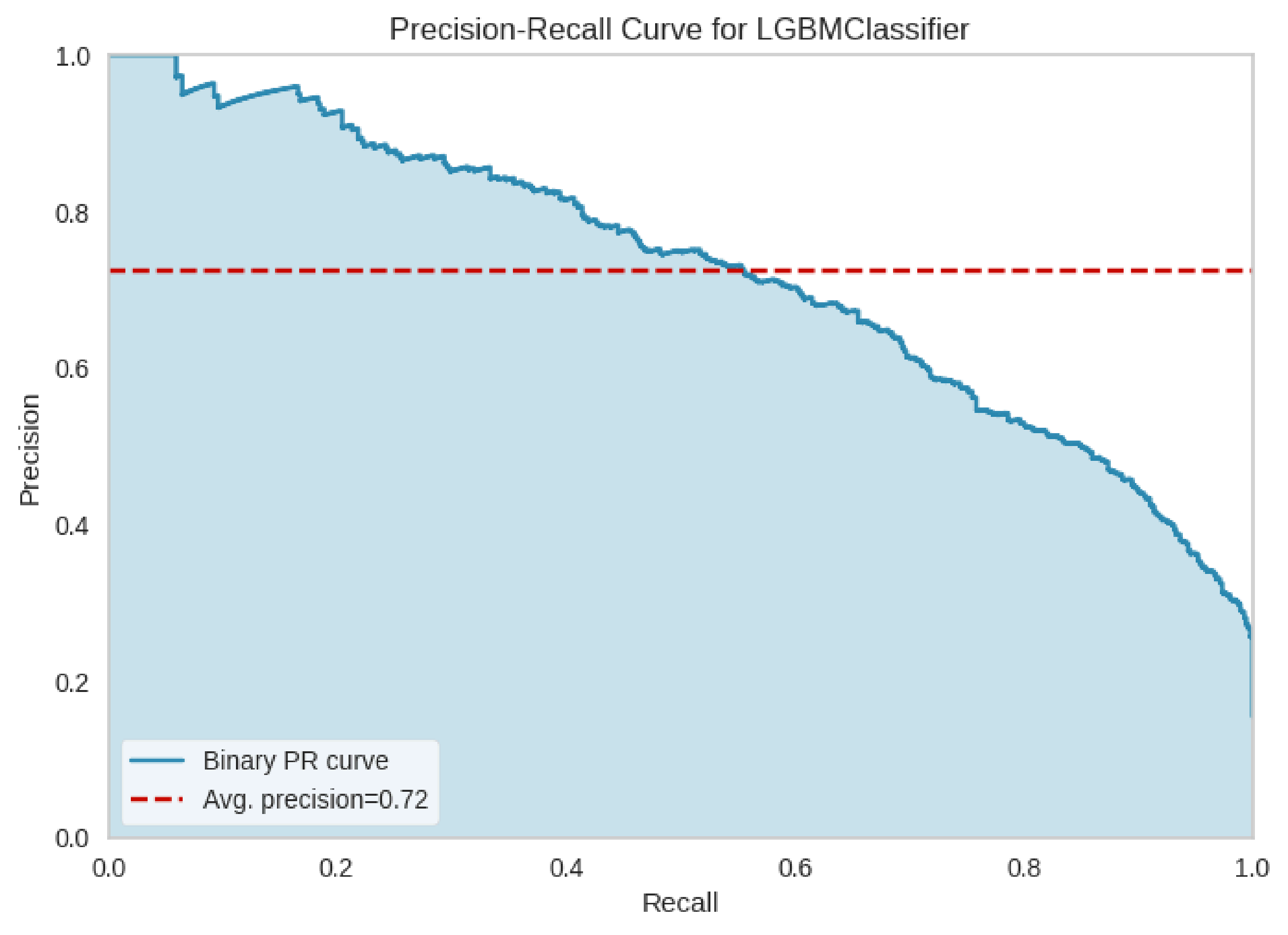
PyCaret is a machine learning library in Python that simplifies the process of preparing data, training models, and evaluating results.

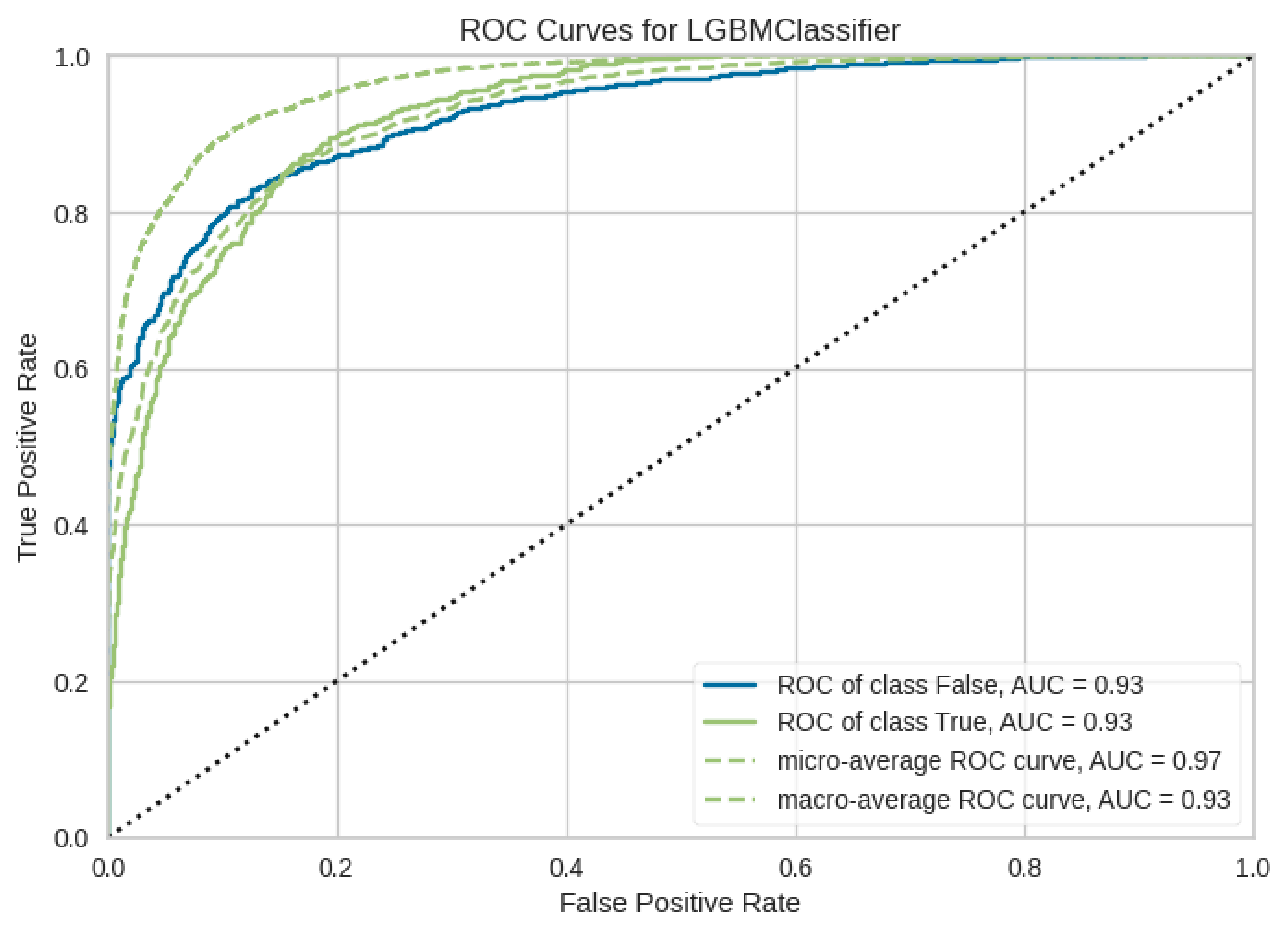


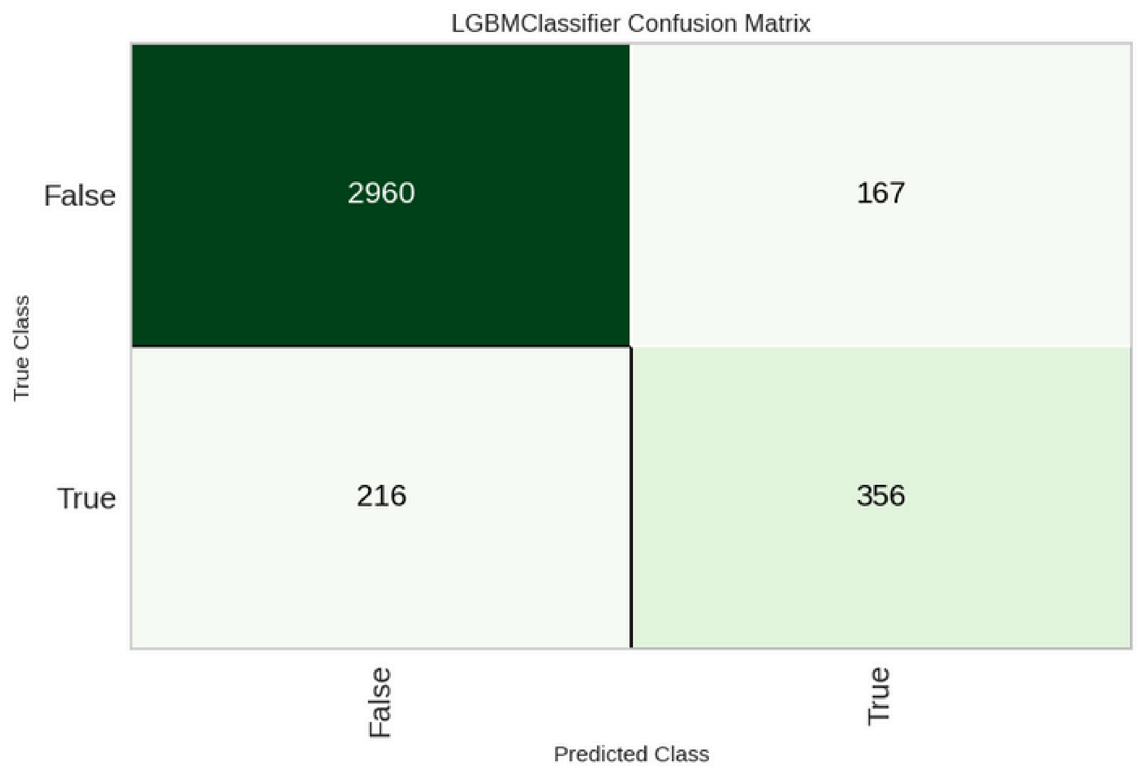
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.9046	0.9323	0.6497	0.7101	0.6781	0.6223	0.6234	
rf	Random Forest Classifier	0.9038	0.9295	0.6924	0.6884	0.6902	0.6333	0.6334	
et	Extra Trees Classifier	0.8986	0.9267	0.6834	0.6688	0.6759	0.6158	0.6160	
gbc	Gradient Boosting Classifier	0.8984	0.9266	0.7178	0.6579	0.6863	0.6258	0.6268	
xgboost	Extreme Gradient Boosting	0.8982	0.9255	0.6303	0.6858	0.6565	0.5969	0.5978	
lr	Logistic Regression	0.8925	0.9147	0.7044	0.6394	0.6700	0.6060	0.6072	
ada	Ada Boost Classifier	0.8890	0.9132	0.6849	0.6301	0.6562	0.5902	0.5910	
svm	SVM - Linear Kernel	0.8881	0.9011	0.7455	0.6168	0.6736	0.6070	0.6120	
ridge	Ridge Classifier	0.8876	0.9168	0.7538	0.6118	0.6752	0.6082	0.6132	
lda	Linear Discriminant Analysis	0.8876	0.9168	0.7538	0.6118	0.6752	0.6082	0.6132	
dt	Decision Tree Classifier	0.8661	0.7582	0.6018	0.5632	0.5817	0.5021	0.5026	
knn	K Neighbors Classifier	0.8569	0.8565	0.6729	0.5299	0.5927	0.5074	0.5129	
qda	Quadratic Discriminant Analysis	0.8452	0.8901	0.0000	0.0000	0.0000	0.0000	0.0000	
dummy	Dummy Classifier	0.8452	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	
nb	Naive Bayes	0.7524	0.8626	0.8436	0.3701	0.5140	0.3806	0.4389	

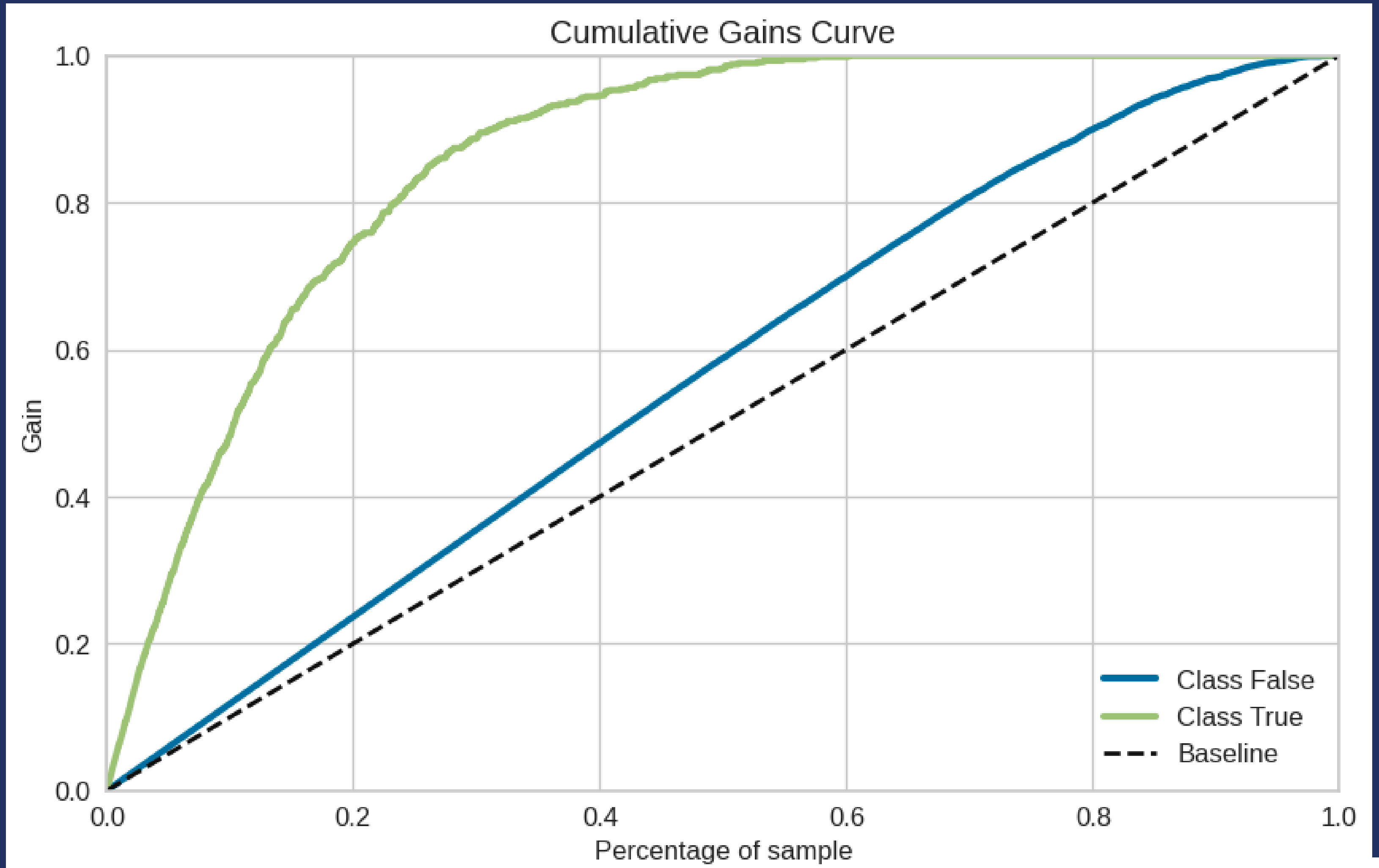


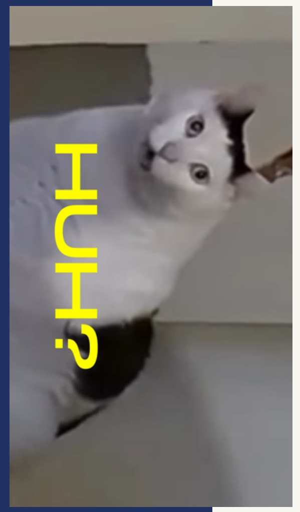
Analyzing results of LightGBM model

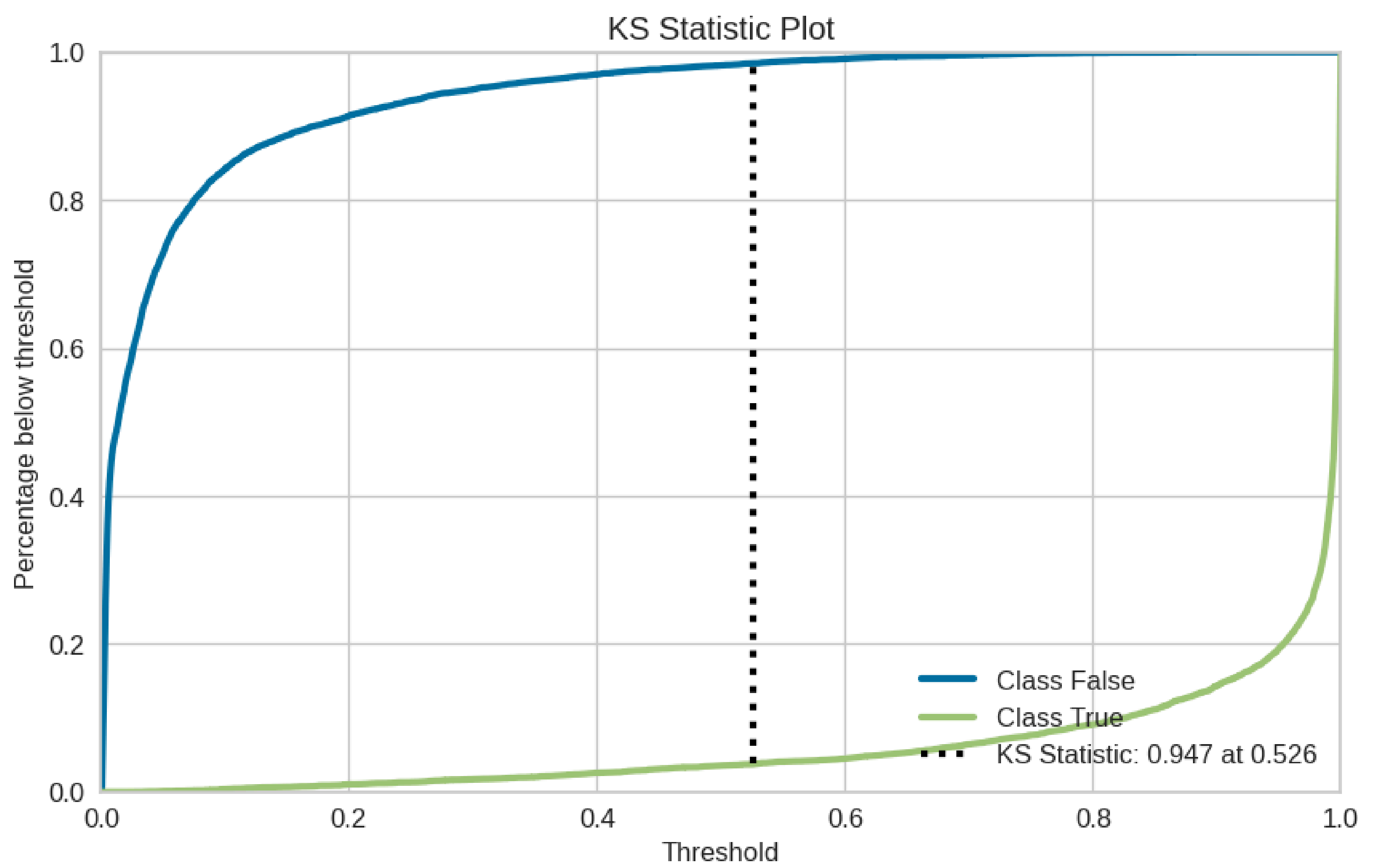


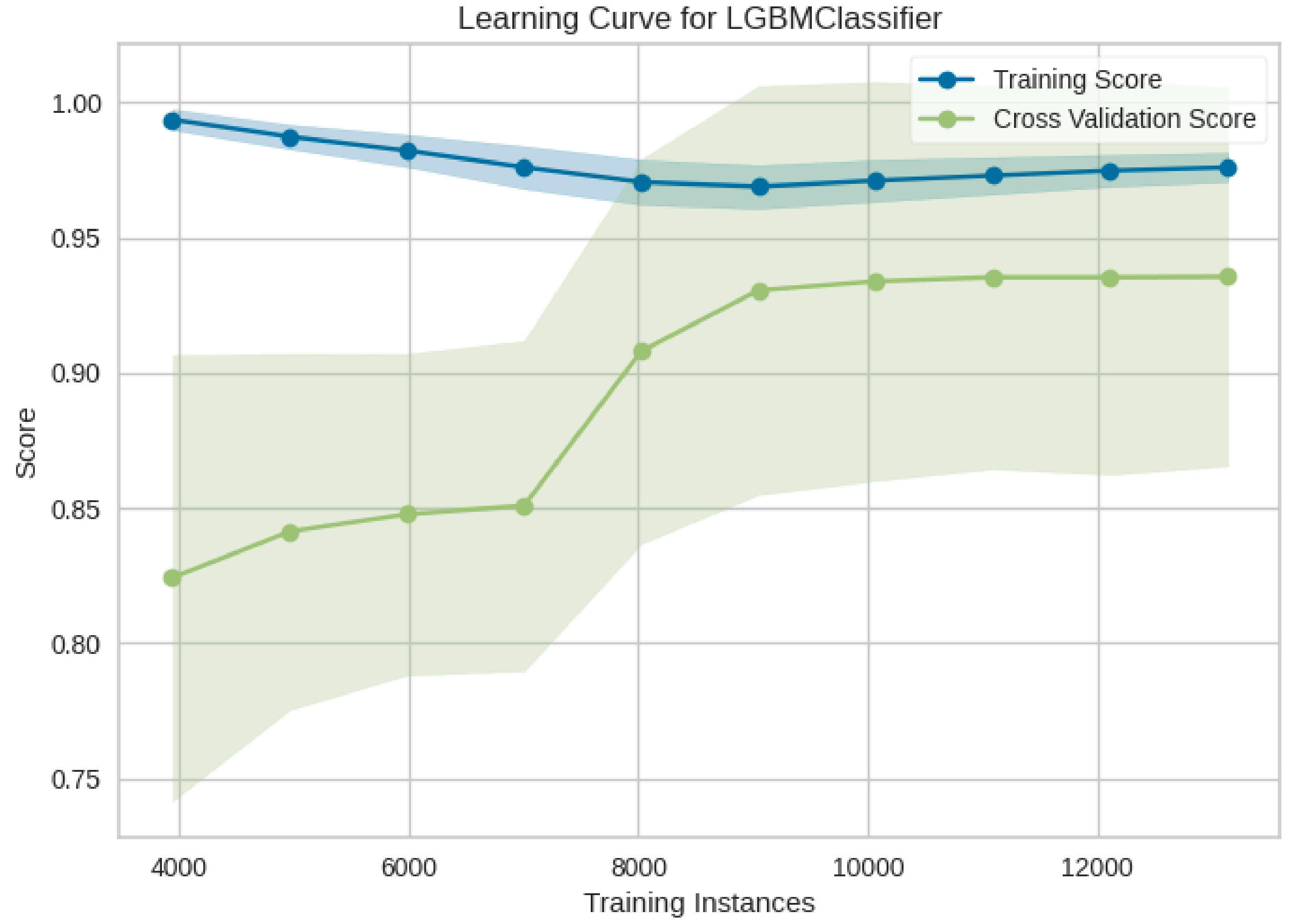

ROC Curves for LGBMClassifier
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC of class False, AUC = 0.93
ROC of class True, AUC = 0.93
micro-average ROC curve, AUC = 0.97
macro-average ROC curve, AUC = 0.93

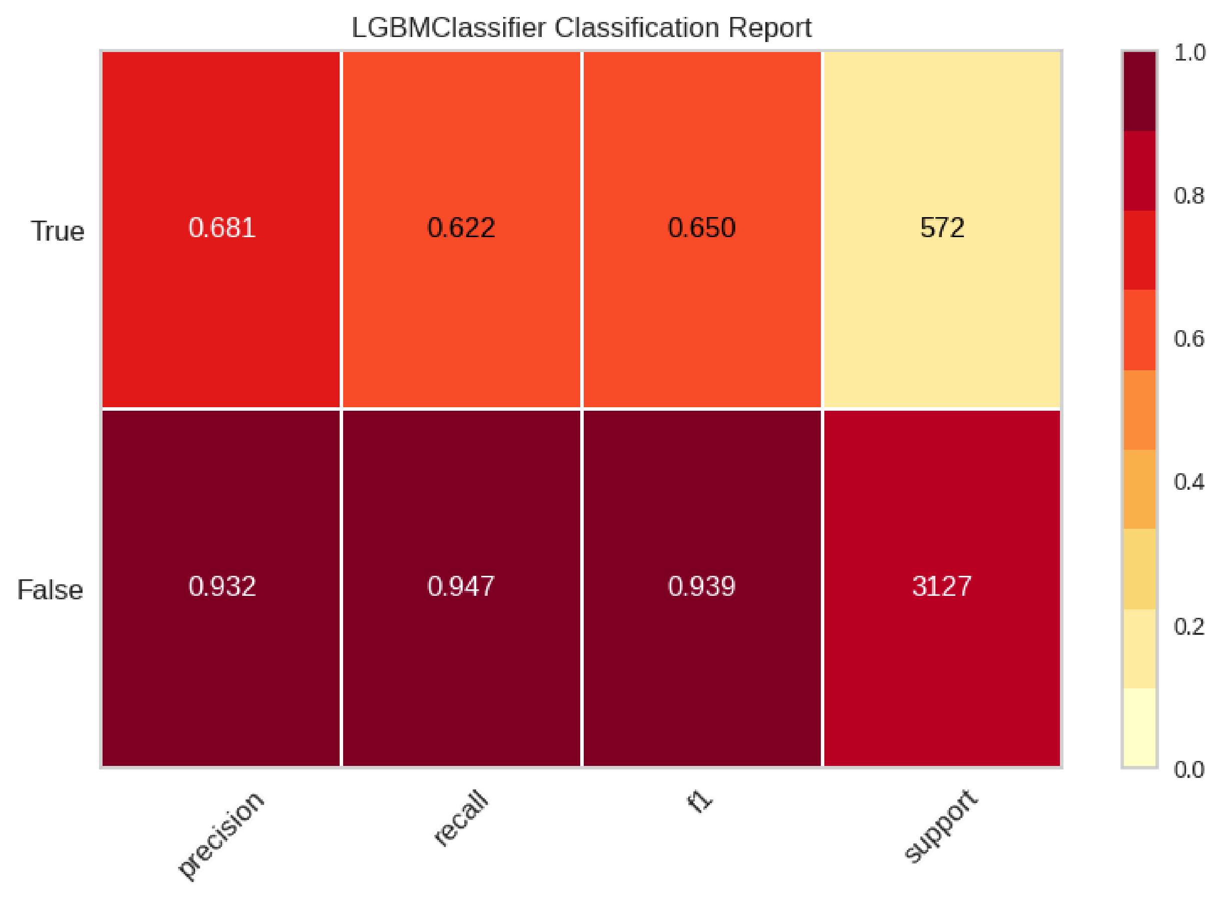



Cumulative Gains Curve
Gain
Percentage of sample
0.0
0.2
0.4
0.6
0.8
1.0
Class False
Class True
Baseline


HUH?

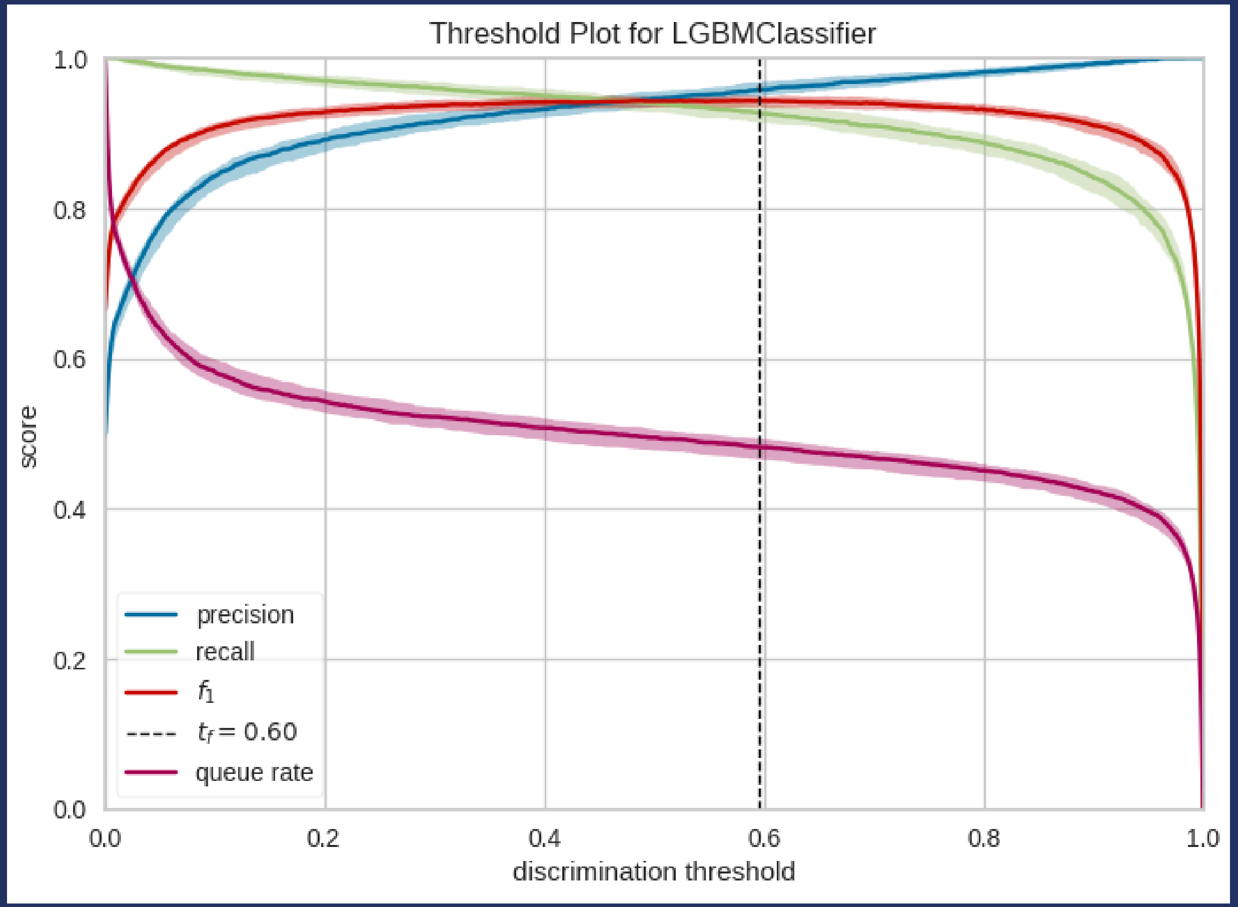

KS Statistic Plot
Percentage below threshold
Threshold
0.0
0.2
0.4
0.6
0.8
1.0
Class False
Class True
KS Statistic: 0.947 at 0.526


Learning Curve for LGBMClassifier
Score
Training Instances
1.00
0.95
0.90
0.85
0.80
0.75
4000
6000
8000
10000
12000
Training Score
Cross Validation Score



LGBMClassifier Classification Report

	precision	recall	f1	support
True	0.681	0.622	0.650	572
False	0.932	0.947	0.939	3127



In conclusion,

By leveraging data-driven insights and advanced machine learning techniques, the project provides actionable strategies to enhance e-commerce platforms' effectiveness in converting visitors into buyers. The findings highlight the importance of features such as page values, product-related duration, and exit rates in predicting purchasing behavior.

Thank You