Predictive analysis for online shoppers behavior

| Virendra Rajpurohit, Reshma Vangala | January 2026 |

1. Abstract

This project predicts online shopper purchasing behavior using the Online Shoppers Purchasing Intention Dataset from UCI, which contains 12,330 user sessions with behavioral, temporal, and technical features. After cleaning and preprocessing, EDA showed strong class imbalance, high skewness in behavioral features, and strong links between purchases and PageValues, BounceRates, ExitRates, and product interactions. Statistical tests confirmed that weekend visits and user engagement significantly affect purchases. SMOTE was used to handle imbalance. Logistic Regression, Random Forest, and Gradient Boosting were trained, with Gradient Boosting performing best (89% accuracy, AUC 0.93). Feature importance showed PageValues as the strongest predictor, proving that user behavior is more important than technical factors in predicting purchases.

2. Introduction

This project focuses on analyzing and predicting **online shopper purchasing behavior** using machine learning techniques. Understanding user behavior during online sessions helps e-commerce businesses improve customer experience, optimize marketing strategies, and increase conversion rates. The analysis uses a real-world dataset that captures detailed user interaction metrics collected during website visits.

3. Data Source and Scope

The online shopper's dataset used in this project originates from an open data repository containing behavioral metrics from an e-commerce platform. The data quality was generally high with no major missing values, but several features required transformation for effective modeling. Categorical variables such as Month, Visitor Type, and Weekend indicators were converted using encoding techniques. Numerical fields including durations and rates were analyzed for skewness and scaled appropriately.

The dataset contains 12,330 records and 18 attributes. Each record represents a single online user session, capturing browsing behavior, session characteristics, and contextual information. The Revenue attribute is used as the target variable, indicating whether a user made a purchase during their online session.

- The dataset includes features representing the number of administrative, informational, and product-related pages visited, along with the time spent on each of these page types.
- Behavioral metrics such as BounceRates and ExitRates describe how quickly users leave the website and how frequently they exit from specific pages.
- PageValues represents the estimated value of the pages visited.
- SpecialDay, Month, OperatingSystems, Browser, Region, TrafficType, VisitorType, and Weekend capture temporal, technical, and user-related factors that influence purchasing behavior.

Overall, the dataset provides a comprehensive view of online customer interactions and is well-suited for user behavior analysis and machine learning classification models to predict purchasing intention.

4. Data Acquisition and Cleaning/Preprocessing:

4.1 Data Acquisition and Initial Load:

The dataset was loaded into a Pandas DataFrame using the read_csv()

4.2 Data Cleaning

- The dataset contained no missing values, so no imputation was required.
- Categorical features were converted into numerical format, suitable for machine learning models.
- Binary variables like Weekend and Revenue were encoded as:
  - False - 0, True - 1
- The VisitorType feature was encoded as:
  - Returning_Visitor - 0
  - New_Visitor - 1
  - Other - 2
- The Month feature was numerically encoded based on calendar order as follows: Feb = 2, Mar = 3, May = 5, June = 6, Jul = 7, Aug = 8, Sep = 9, Oct = 10, Nov = 11, and Dec = 12.

These cleaning steps ensured the dataset was consistent, machine-readable, and ready for preprocessing and model training.

5. Exploratory Data Analysis (EDA)

Analysis of browsing behavior through histograms and correlation matrices revealed that PageValues is by far the strongest predictor of Revenue. Sessions with high PageValues and long ProductRelated_Duration exhibit significantly higher purchase probability. ExitRates and BounceRates show negative correlation with Revenue, confirming that users leaving quickly are unlikely to convert. Weekend traffic tends to be slightly lower in conversion compared with weekday sessions.

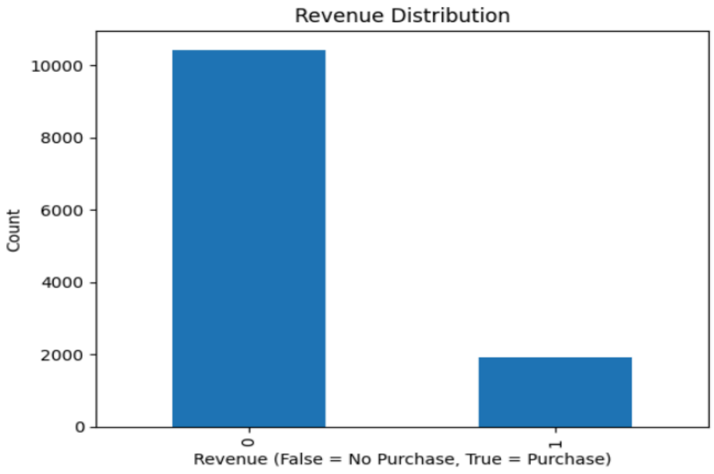
5.1 Feature Selection

- **Target:** Revenue (Binary variable indicating whether a purchase was made)
- **Features:** Administrative, Administrative_Duration, Informational, Informational_Duration, ProductRelated, ProductRelated_Duration, BounceRates, ExitRates, PageValues, SpecialDay, Month, OperatingSystems, Browser, Region, TrafficType, VisitorType, Weekend.

5.2 Handling Data Imbalance

To address this imbalance, SMOTE (Synthetic Minority Over-sampling Technique) was applied. SMOTE generates synthetic samples for the minority class (Revenue = 1), helping to balance the dataset. This improves model learning, reduces bias toward the majority class, and enhances the performance of classification models.

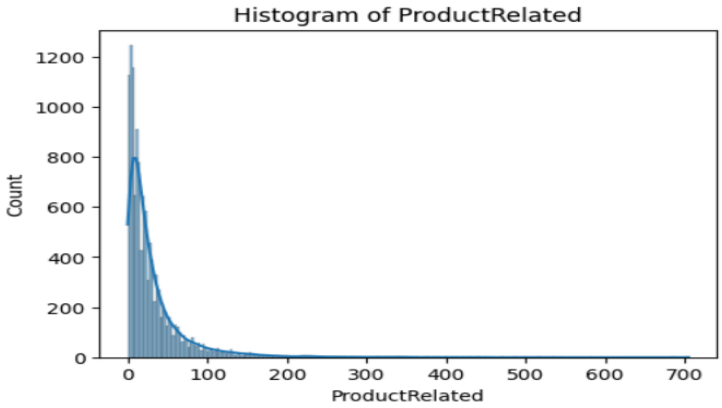
5.3 Target Variable Distribution



Observation

- The majority of users (84.53%) did not make a purchase.
- Only 15.47% of users completed a purchase.
- This indicates a class imbalance, where the number of non-purchase cases is significantly higher than purchase cases.

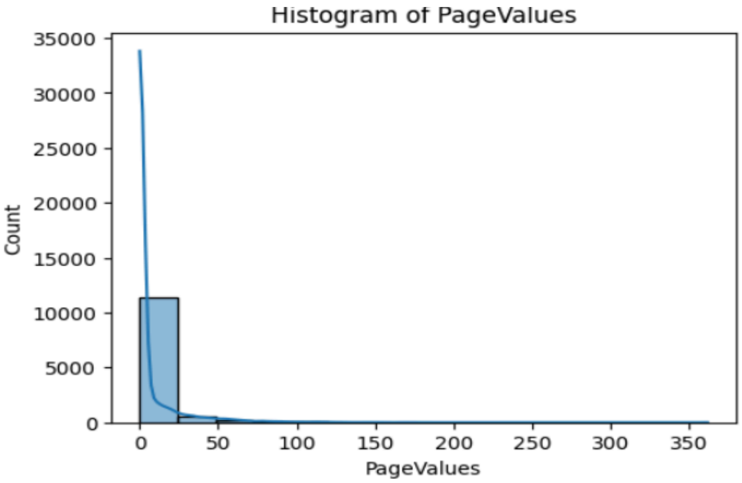
5.4 Distribution of Product-Related Page Visits



Observation:

- Most users visited very few product-related pages, indicating strong concentration near 0.
- The distribution is highly right-skewed, with a long tail extending up to around 700 pages.
- A small fraction of users visited a very large number of product-related pages.
- High variability in product page interactions could influence purchase behavior and conversion likelihood.

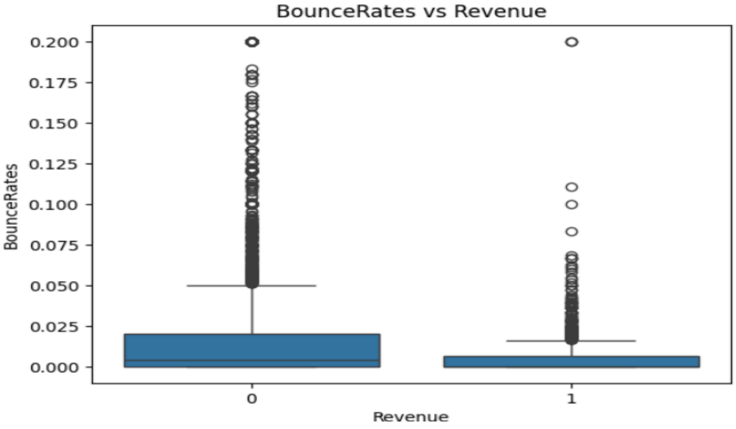
5.5 Distribution of Page Values



Observation:

- Most users have very low page values, concentrated near 0.
- The distribution is highly right-skewed, with a long tail extending up to around 350.
- A small number of users have very high page values, which could indicate high engagement or high likelihood of purchase.
- PageValues captures monetary intent, users with PageValues > 5 are very likely buyers.

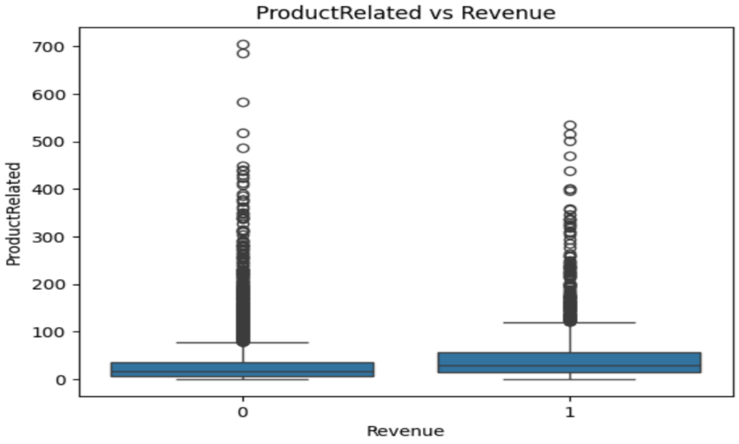
5.6 Bounce Rate Analysis by Purchase Behavior



Observation:

- Revenue = 0 (No Purchase): Mean bounce rate 0.025; most sessions 0–0.02 with outliers up to 0.20.
- Revenue = 1 (Purchase Made): Mean bounce rate 0.005; sessions tightly clustered near 0.
- Insight: Non-buyers leave faster, while buyers stay longer; lower bounce rates strongly correlate with purchases.
- Exit rate shows the same pattern.

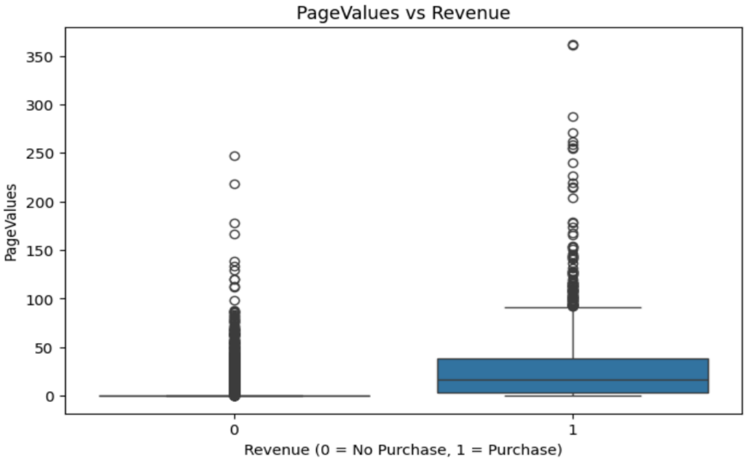
5.7 Product Interactions vs. Revenue



Observation:

- Buyers view more products (median ~25–30) than non-buyers (median ~10–15).
- Most users interact with only a few products.
- A few users browse a lot but don’t purchase.
- Higher product interaction generally increases the chance of buying.

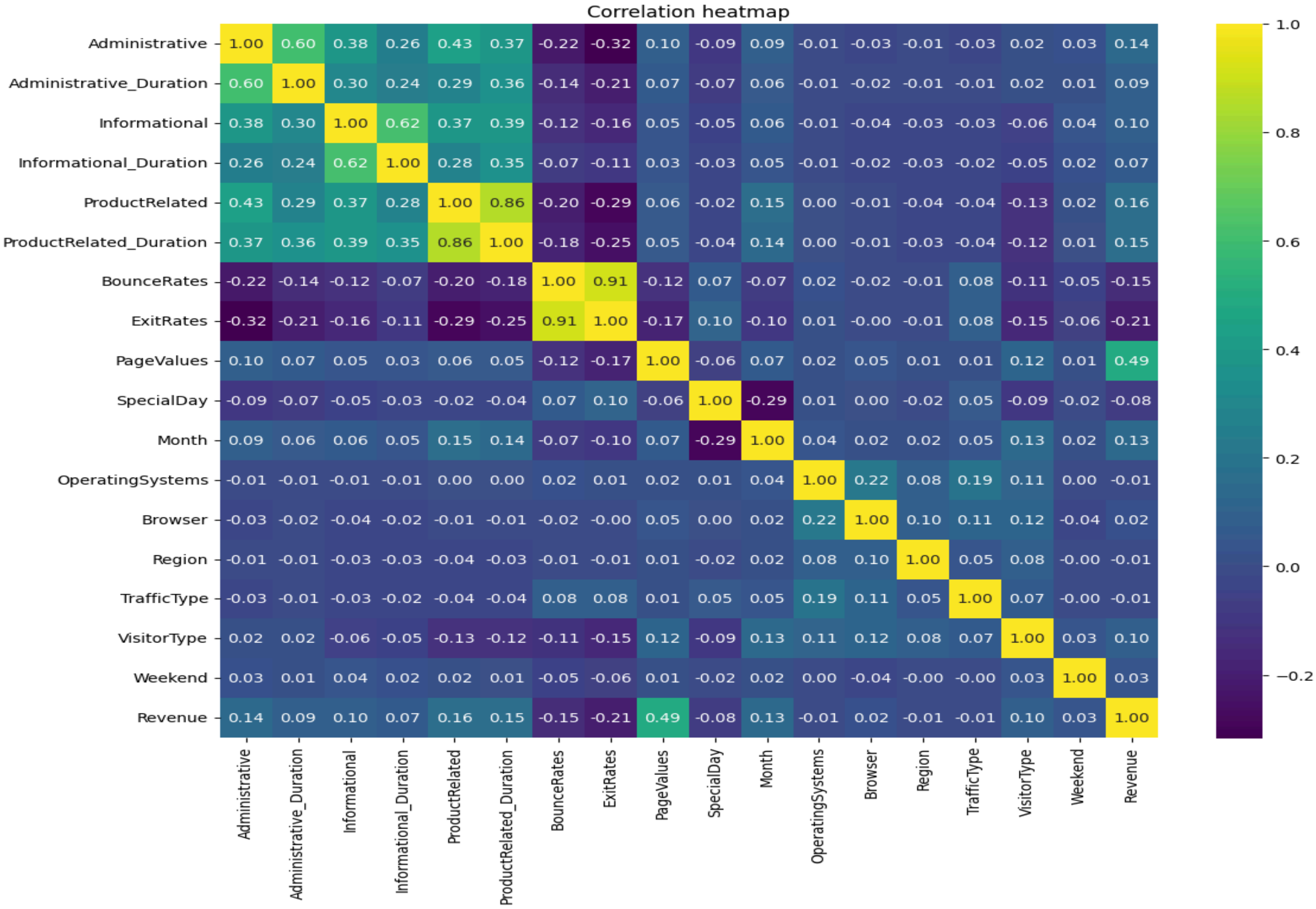
5.8 PageValues vs. Revenue



Observation

- Purchasers have significantly higher PageValues than non-purchasers.
- Most non-buyers show PageValues near zero.
- Higher PageValues strongly increase purchase likelihood.
- PageValues is a key predictor of conversion.

5.9 Correlation Heatmap:



Strong Predictors of Revenue:

Positive correlation:

- PageValues (0.49): When users visit more pages like checkout, they are much more likely to buy.
- ProductRelated (0.16): Viewing more product pages slightly increases the chance of purchase.
- Informational (0.15): Information pages can slightly increase purchase likelihood, but the effect is small.

Negative correlation:

- ExitRates (-0.21): Users who leave pages quickly are less likely to buy.
- BounceRates (-0.15): If users leave after viewing only one page, the chance of purchase is lower.

Other Key Observations (Multicollinearity): Below features are very strongly related.

- BounceRates & ExitRates (0.91)
- ProductRelated & ProductRelated_Duration (0.86)

6. Hypothesis Test

6.1 Chi-Square Test: Weekend vs Revenue

A Chi-square test of independence was conducted to examine the relationship between weekday vs weekend and Revenue (purchase vs no purchase).

Hypotheses:

- **H_0 (Null Hypothesis):** Weekend does not affect the likelihood of making a purchase.
- **H_1 (Alternative Hypothesis):** Users are more likely to make a purchase on weekends.

Contingency Table:

	No Purchase (0)	Purchase (1)
Weekday (0)	8053	1409
Weekend (1)	2369	499

Test Results: Since the p-value is **less than 0.05**, the null hypothesis is rejected.

- **Chi-square statistic:** 10.391
- **p-value:** 0.001

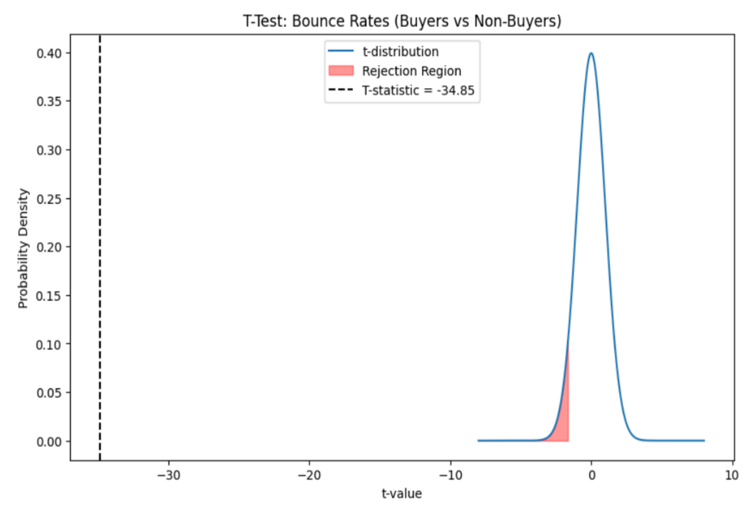
Conclusion: There is a statistically significant association between weekend visits and purchase behavior.

6.2 T-Test: Bounce Rate vs Purchase Behavior

An independent two-sample t-test was conducted to examine whether there is a significant difference in **bounce rates** between users who made a purchase and those who did not.

Hypotheses:

- **H_0 (Null Hypothesis):** There is no difference in bounce rates between buyers and non-buyers.
- **H_1 (Alternative Hypothesis):** Bounce rates differ between buyers and non-buyers.



Test Results:

- **T-statistic:** −34.85
- **Degrees of freedom:** 11,795.09
- **Critical t-value ($\alpha = 0.05$):** −1.645
- **P-value:** 1.29×10^{-253}

p-value is effectively zero, so reject the null hypothesis.

Conclusion: There is a statistically significant difference in bounce rates between buyers and non-buyers. The negative t-statistic indicates that less bounce rate implies more purchase likelihood.

7. Model Building

7.1 Data Preparation for Modeling

Before training the machine learning models, the dataset was preprocessed to ensure that all features were in a suitable format and scale for modeling.

- **Numerical features** were standardized using **StandardScaler** to normalize their ranges and prevent features with larger magnitudes from dominating the model.
- **Categorical features** were encoded using **OneHotEncoder** with handle_unknown='ignore' to safely manage unseen categories in the test data.

The dataset was then split into training and testing sets using an **80:20 ratio**, ensuring that the class distribution of the target variable (**Revenue**) was preserved through **stratified sampling**.

- **Training set shape:** (9,864, 17)
- **Test set shape:** (2,466, 17)

7.2 Logistic Regression Model

Model Description

A Logistic Regression model was implemented as a baseline classification algorithm to predict user purchase behavior (Revenue). To ensure proper feature scaling and stable convergence, the model was built using a pipeline that integrates StandardScaler with Logistic Regression. The maximum number of iterations was increased to 2000 to avoid convergence issues. The model was trained on the SMOTE-balanced training dataset to address class imbalance and improve the detection of purchasing sessions.

Model Evaluation

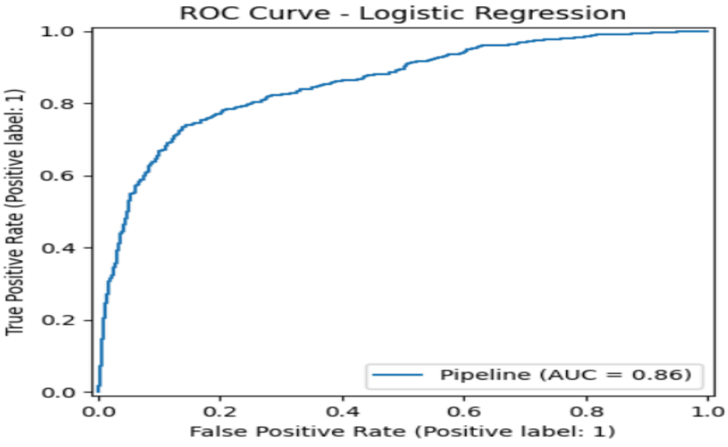
The model was evaluated on the unseen test dataset using classification metrics and ROC-AUC score.
ROC-AUC stands for Receiver Operating Characteristic Area Under the Curve.

Classification Results: Accuracy: 85% , **ROC-AUC:** 0.856.

Class	Precision	Recall	F1-Score
0 (No Purchase)	**0.94**	**0.88**	**0.91**
1 (Purchase)	**0.52**	**0.69**	**0.59**

Confusion Matrix Interpretation

- **True Negatives (1,842):** Non-buyers correctly predicted.
- **True Positives (264):** Buyers correctly predicted.
- **False Positives (242):** Non-buyers incorrectly classified as buyers.
- **False Negatives (118):** Buyers incorrectly classified as non-buyers.



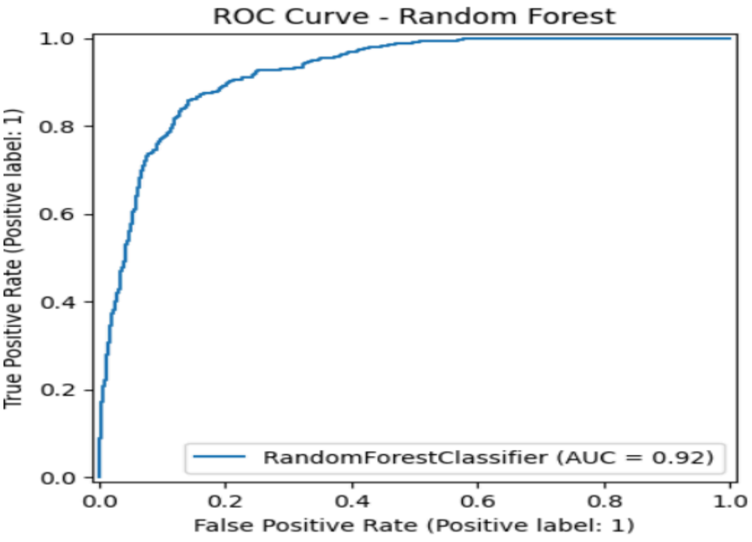
Precision: Out of all the sessions the model predicted as buyers, how many were actual.
Recall: Out of all the actual buyers, recall tells us how many the model was correctly found.
F1-Score: A single score that balances both above.
Observation
The model performs well in distinguishing between buyers and non-buyers, with an AUC of 0.86. It is particularly strong at identifying non-buyers and achieves a recall of 0.69 for buyers, meaning most purchasing sessions are correctly detected.

7.3 Random Forest Model

Model Description:

A Random Forest classifier was built using 200 decision trees with a maximum depth of 10. The model was trained on SMOTE-balanced training data to handle class imbalance between buyers (1) and non-buyers (0). Random state was fixed to ensure reproducible results. Random Forest combines multiple decision trees to improve prediction accuracy and reduce overfitting.

Model Evaluation: Accuracy: 0.88, **ROC–AUC:** 0.92



- The model shows strong overall performance. High ROC–AUC indicates excellent ability to separate buyers from non-buyers. Accuracy of 88% means most predictions are correct even with imbalanced data.
- The model is very strong at identifying non-buyers.
- For buyers, recall is high (0.76), meaning most actual buyers are detected.
- Precision for buyers is moderate (0.59), meaning some predicted buyers are actually non-buyers.
- This trade-off is acceptable when the goal is to not miss potential buyers.

Observation: The model successfully captures most buyers while keeping false alarms at a reasonable level. It slightly favors identifying buyers, which is useful for business tasks like marketing or customer targeting where missing a buyer is costlier than contacting a non-buyer.

Classification Report:

Class	Precision	Recall	F1-Score
0 (No Purchase)	0.95	0.90	0.93
1 (Purchase)	0.59	0.76	0.67

7.4 Gradient Boosting Tree Model

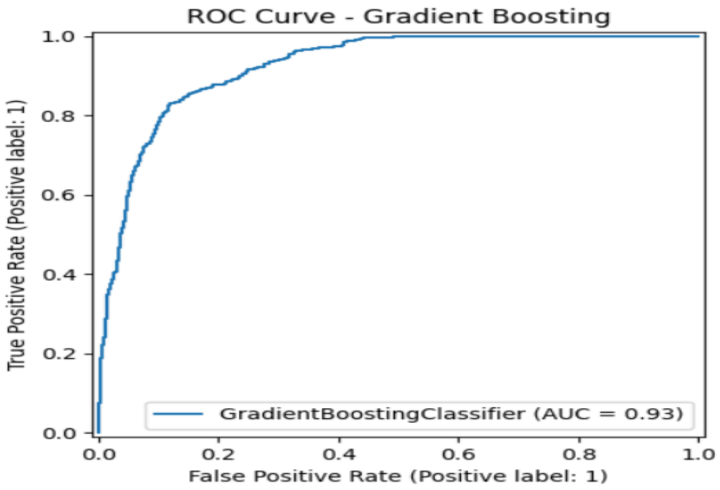
Model Description

A Gradient Boosting Classifier was trained using 150 boosting stages, a learning rate of 0.1, and tree depth of 3. The model was trained on SMOTE-balanced training data to handle class imbalance and evaluated on the original test set.

Model Evaluation: **Accuracy:** 0.89, **ROC–AUC:** 0.93

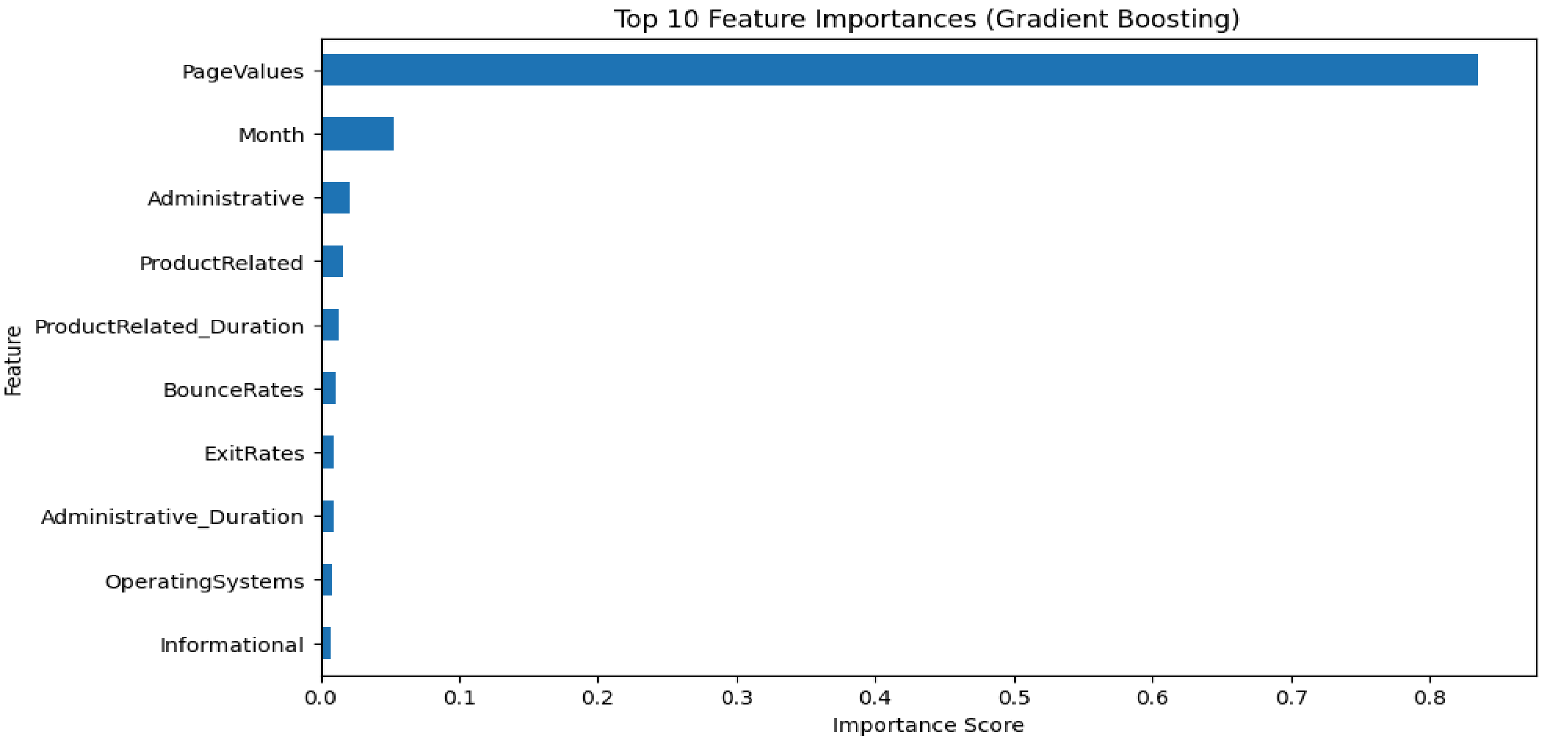
Classification Report:

Class	Precision	Recall	F1-Score	Support
0 (No Purchase)	0.95	0.92	0.93	2084
1 (Purchase)	0.61	0.73	0.67	382



- The model shows strong overall performance, with slightly better results than Random Forest. The high ROC–AUC indicates excellent ability to distinguish between buyers and non-buyers.
- The model performs very well for non-buyers.
- For buyers, recall is 0.73, meaning most buyers are correctly identified.
- Precision for buyers is 0.61, slightly better than Random Forest.
- This balance is useful when identifying likely purchasers.

Observation: Gradient Boosting shows strong and balanced performance, slightly outperforming Random Forest in overall accuracy and precision for buyers, making it a good choice for purchase prediction tasks.



PageValues is by far the most important feature in Gradient Boosting, dominating the model's decisions. Month has some influence, while all other features contribute very little. This shows the model mainly relies on page value and, to a smaller extent, seasonal effects to predict purchases.

8. Conclusion

This study successfully demonstrated how machine learning can be used to predict online shopper purchasing intention using real-world session data. Exploratory analysis showed that most users do not purchase, making class imbalance a major challenge, which was effectively handled using SMOTE. Behavioral features such as PageValues, product interactions, bounce rate, and exit rate were found to be the strongest indicators of purchase behavior. Statistical tests confirmed that both user engagement and timing (weekend vs weekday) significantly influence conversions.

Among the models tested, Gradient Boosting performed the best with 89% accuracy and an AUC of 0.93, slightly outperforming Random Forest and clearly improving over the baseline Logistic Regression model. Both tree-based models showed strong ability to detect buyers while maintaining high performance on non-buyers. Feature importance analysis revealed that PageValues dominates prediction, while technical attributes such as operating system or browser have minimal impact.

Overall, the results show that user behavior during a session is far more important than technical or demographic factors in predicting purchases. These findings can help e-commerce businesses focus on improving user engagement, optimizing high-value pages, and targeting users more effectively to increase conversion rates.

9. Future Work

The hypothesis that browsing behavior can accurately predict purchase intention has been validated. The project produced a useful model suitable for operational marketing deployment. Future improvements may include richer user-level history features, deeper calibration techniques, and online A/B testing integration. Additional ensemble approaches combining XGBoost and LightGBM probabilities may further enhance robustness.

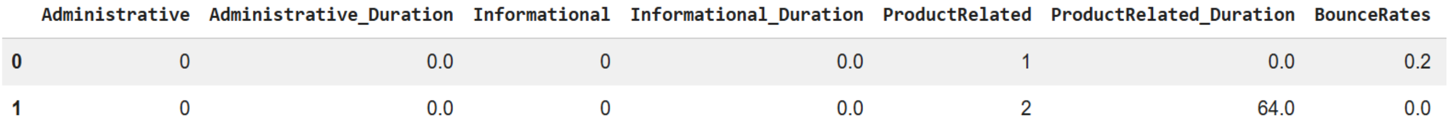
- Enhance targeting for high-value visitor categories.

- Optimize pages with high exit/bounce rates.

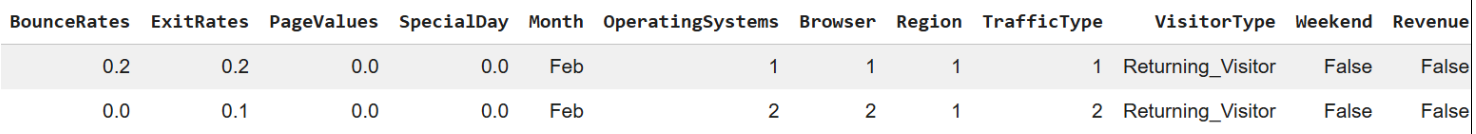
10. Appendix – Data Sample

- Dataset Link: https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset
- A small sample of the dataset is below:

```
df=pd.read_csv('online_shoppers_intention.csv')
 df.head(2)
```



	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates
0	0	0.0	0	0.0	1	0.0	0.2
1	0	0.0	0	0.0	2	64.0	0.0



BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0.2	0.2	0.0	0.0	Feb	1	1	1	1	Returning_Visitor	False	False
0.0	0.1	0.0	0.0	Feb	2	2	1	2	Returning_Visitor	False	False

- Features encoding and target, Code for model training, ROC curves, confusion matrices, and SHAP generation are fully documented in the Jupyter notebook.
- ROC-AUC stands for Receiver Operating Characteristic Area Under the Curve.