ONLINE SHOPPING PURCHASE INTENTION

GROUP NUMBER 2

- Sophia Nelson - 6428678
- Sarah Daniel - 6909062
- Reethender Vedira - 6910677

UNIVERSITY OF SURREY

PROJECT AIM AND OBJECTIVES

Problem Statement

This project aims to identify the key behavioral and technical factors that influence whether a customer completes an online purchase, to improve the conversion rate for an e-commerce platform.

Hypotheses

- Higher user engagement (e.g., more time on product pages, higher page value scores) increases the likelihood of making a purchase.
- Lower bounce and exit rates are associated with a higher probability of conversion.
- The month of the visit and the type of content viewed (product vs. informational pages) significantly influence purchase behavior.

Aim

To apply machine learning techniques to predict online purchase behavior, identify key drivers of conversion, and provide actionable insights to optimise e-commerce strategies.

Objectives

- **Data Understanding and Preparation**: Explore the dataset, clean, and transform it for machine learning use.
- **Exploratory Data Analysis**: Analyse patterns and visualize relationships between features and purchase intent.
- **Modeling**: Train and evaluate multiple machine learning models (Logistic Regression, Random Forest, XGBoost) to predict purchase intent.
- **Evaluation**: Assess model performance using metrics such as Precision-Recall AUC, Recall, ROC AUC and F1 score among others.
- **Insights and Recommendations**: Provide actionable recommendations based on model findings to improve conversion rates.
- **Reporting**: Present findings and recommendations clearly and concisely to ecommerce/business managers, marketers, data analysts.

CRISP-DM FRAMEWORK APPLICATION

Core Business Objectives:
Identify the key behavioural and technical factors that most influence whether a customer completes a purchase online.
Use insights from shopper behaviour to recommend data-driven strategies that can improve the sales conversion rate.

Success Criteria:
Insights from the analysis enable targeted marketing strategies that seek to increase the online conversion rate for the long term.

Source of data:
The dataset ‘*Online Shoppers Purchasing Intention’* was collected from the external source University of California, Irvine (UCI) repository.
The dataset captures anonymised session-level behavioral data of visitors to an e-commerce website.

Initial observations:
The dataset contains 12,330 instances and 17 features. The data contains a mix of quantitative and categorical/symbolic variables.

Key variables:
The target variable is ‘Revenue’ and the values in the column are stored as a Boolean (True/False).

Some of the important predictor variables include page value score, exit rate, bounce rate and visit month.

Modelling approach:
The project focuses on a binary classification problem to predict the likelihood that a user session will result in a conversion.

Firstly, a baseline model for Logistic regression will be used, followed by a parsimonious model for feature reduction and interpretability.

To complement this, two tree-based models (Random Forest and XGBoost) will be used to maximise prediction accuracy and to verify whether those same predictors are still the most influential and to identify potential non-linear relationships that the logistic regression model may have missed.

Tools & Techniques:
R, PowerBI, tidyverse, caret, preprocessing (scaling, encoding), Variable Inflation Factor, ROSE for balancing, K-fold (n=10).

Evaluation:
confusion matrix, Roc Curve (overall ability separate classes), PR Curve (better for class imbalance), AUC, precision, recall, F1 score, Accuracy), feature importance, SHAP values.

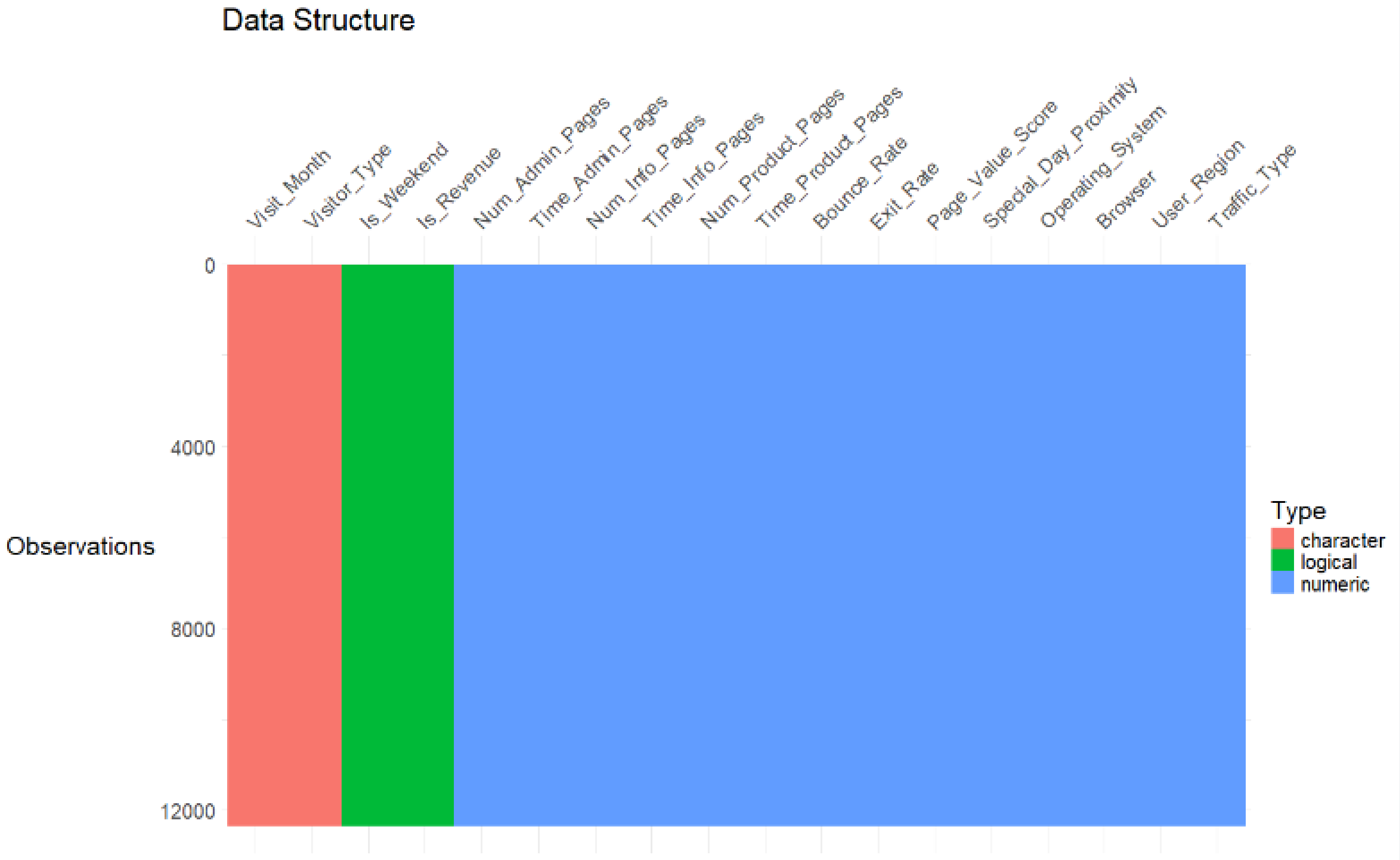
Reporting:
Provide actionable insights for business decision-makers to optimise marketing strategies that result in enhanced online conversion.

BENEFITS OF CRISP - DM

- **Structured Approach**
  - CRISP-DM provided a clear, systematic process, helping us stay focused on the business objectives and ensuring a thorough analysis. It guided us through each stage of the project in a logical and organized manner.
- **Improved Data Quality**
  - By following the Data Understanding and Data Preparation steps rigorously, we ensured that the data was properly cleaned, balanced, and ready for modeling. This led to better model performance and more reliable insights.
- **Effective Model Selection and Evaluation**
  - The iterative nature of CRISP-DM allowed us to experiment with different machine learning models, tuning them based on performance metrics. This led to the identification of the most accurate model for predicting purchase intent.
- **Actionable Business Insights**
  - By following CRISP-DM’s Evaluation and Deployment steps, we could link model findings directly to business strategies, such as improving website engagement and reducing exit rates, ensuring the insights were practical and actionable.
- **Flexibility and Iteration**
  - The iterative nature of CRISP-DM allowed us to revisit and refine earlier steps as needed, improving the overall quality of the analysis and ensuring that the final model was the best fit for the problem.

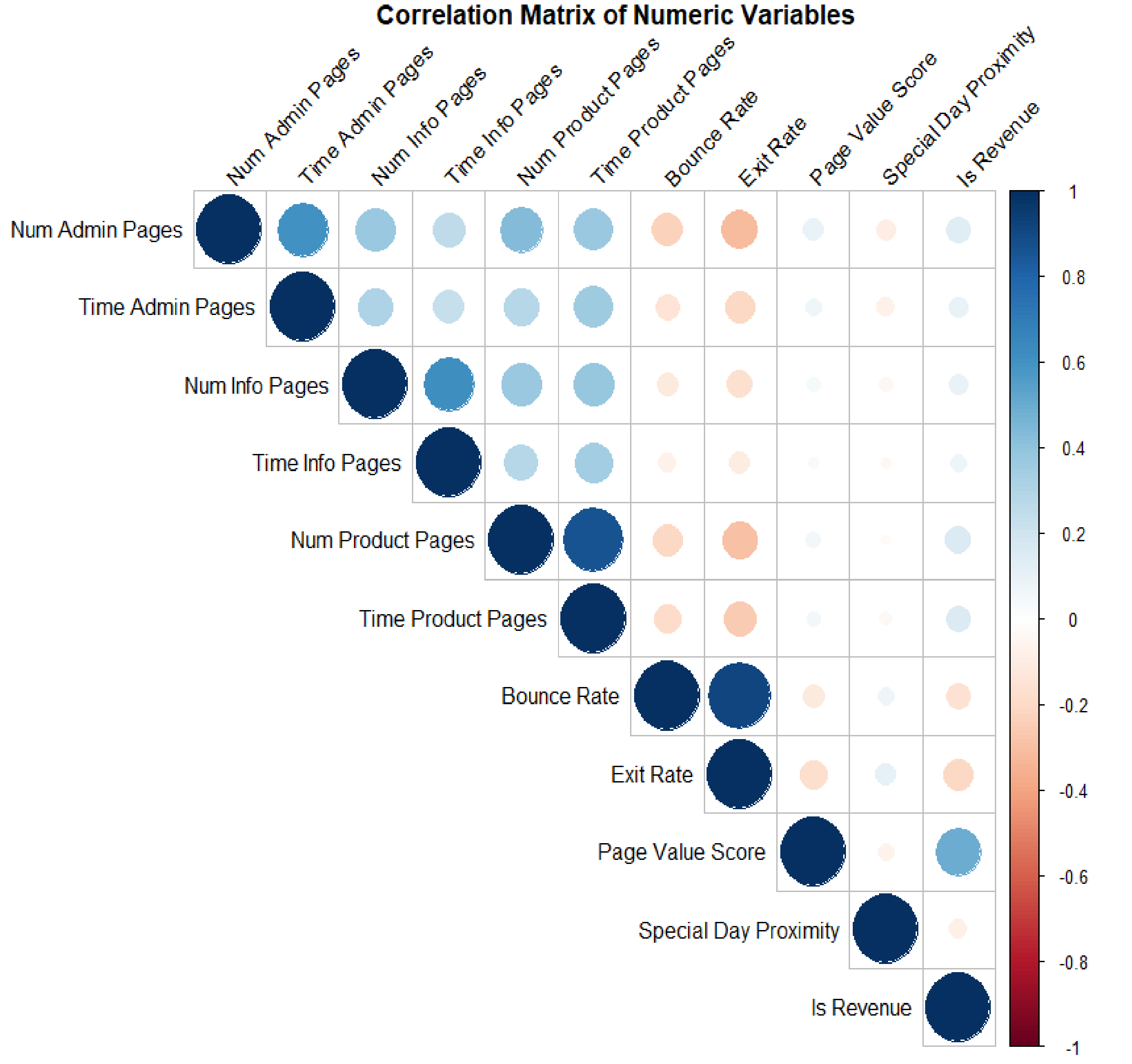
DATA QUALITY ASSESSMENT

- Data types are a mix of categorical/symbolic and numeric variables.
- Target variable is binary and originally denoted as Boolean TRUE/FALSE.
- No missing values identified in the dataset.
- Duplicate values were identified, and after reviewing them, they were retained in the main dataset for modelling, since they are not errors and reflect repeated behaviour of different users. The only exception was during t-SNE visualisation, where duplicates were removed to avoid overplotting and to ensure better visual clarity.
- Minority class (TRUE) is imbalanced compared to the FALSE class. Thus, to prevent biased predictions, this variable should be balanced.



DATA CLEANING & FEATURE ENGINEERING

- Column renaming for interpretability.
- Log transformation of some strongly right skewed variables (e.g., Page_Value_Score, Time_Admin_Pages, Time_Info_Pages, Time_Product_Pages) and discretization of (e.g., Num_Info_Pages, Num_Admin_Pages and Num_Product_Pages).
- High cardinality was identified among some categorical variables e.g., Browser, which had 13 levels, and Traffic_Type, which had 20 levels. To reduce the number of unique levels, rare categories (representing less than 5 percent of the data) were grouped into “Other.”
- Standardisation (Z-score) applied to numeric features to ensure that they have a mean of 0 and a standard deviation of 1, to improve comparability across features and better model performance.
- Target variable encoding from TRUE/FALSE to a factor (YES/NO) for logistic regression & random forest, although XGBoost was encoded numerically (1/0) to ensure compatibility.
- Converted categorical variables into numeric format using feature encoding techniques (e.g., one-hot encoding), ensuring compatiability with machine learning algorithms and improving model performance.
- Correlation matrix revealed weak to moderate correlation between majority of the features and the target variable.
- In addition, there were signs of moderate multicollinearity between some variables (e.g., Num_Product_Pages & Time_Product_Pages, Bounce_Rate & Exit_Rate).
- The Variable Inflation Factor (VIF) technique, helped to determine that none of the features were redundant (VIF <5) and therefore no variables were dropped from the dataset.


Correlation Matrix of Numeric Variables

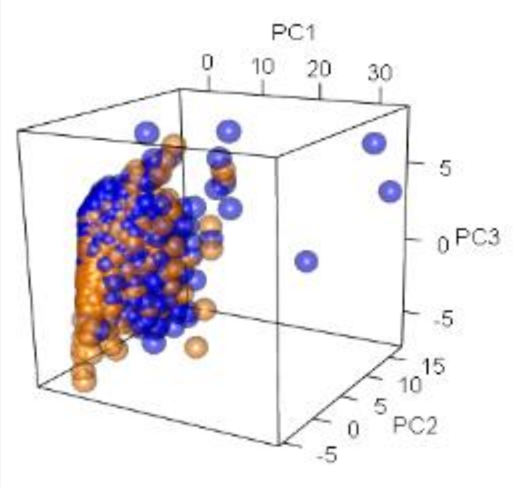
EXPLORATORY DATA ANALYSIS- DATA STRUCTURE

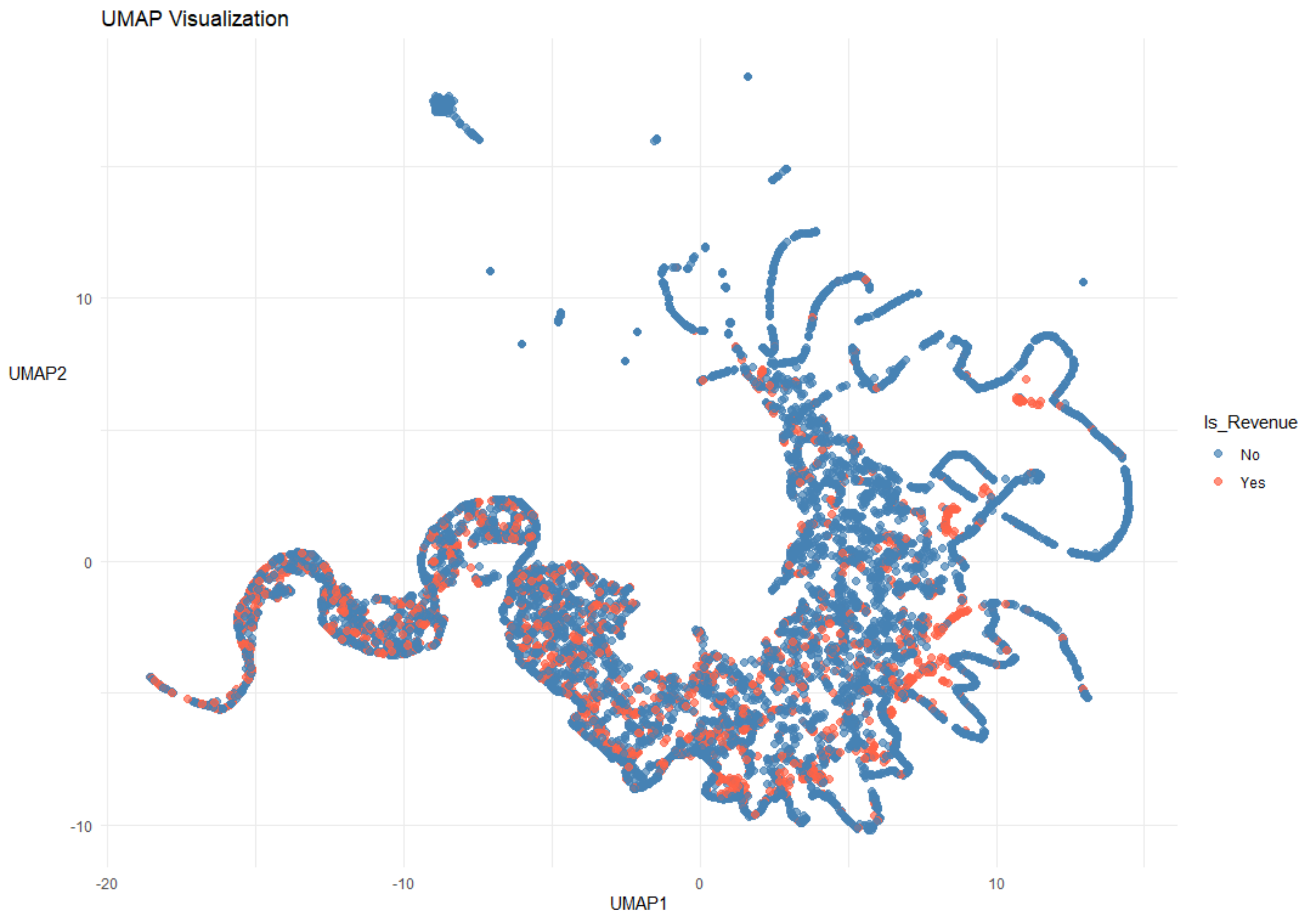
PCA 3D Plot

- Demonstrates the underlying structure of the high-dimensional dataset.
- Shows how observations are distributed across the first three principal components.
- Clear clustering patterns are limited → classes are not easily separable with linear methods.

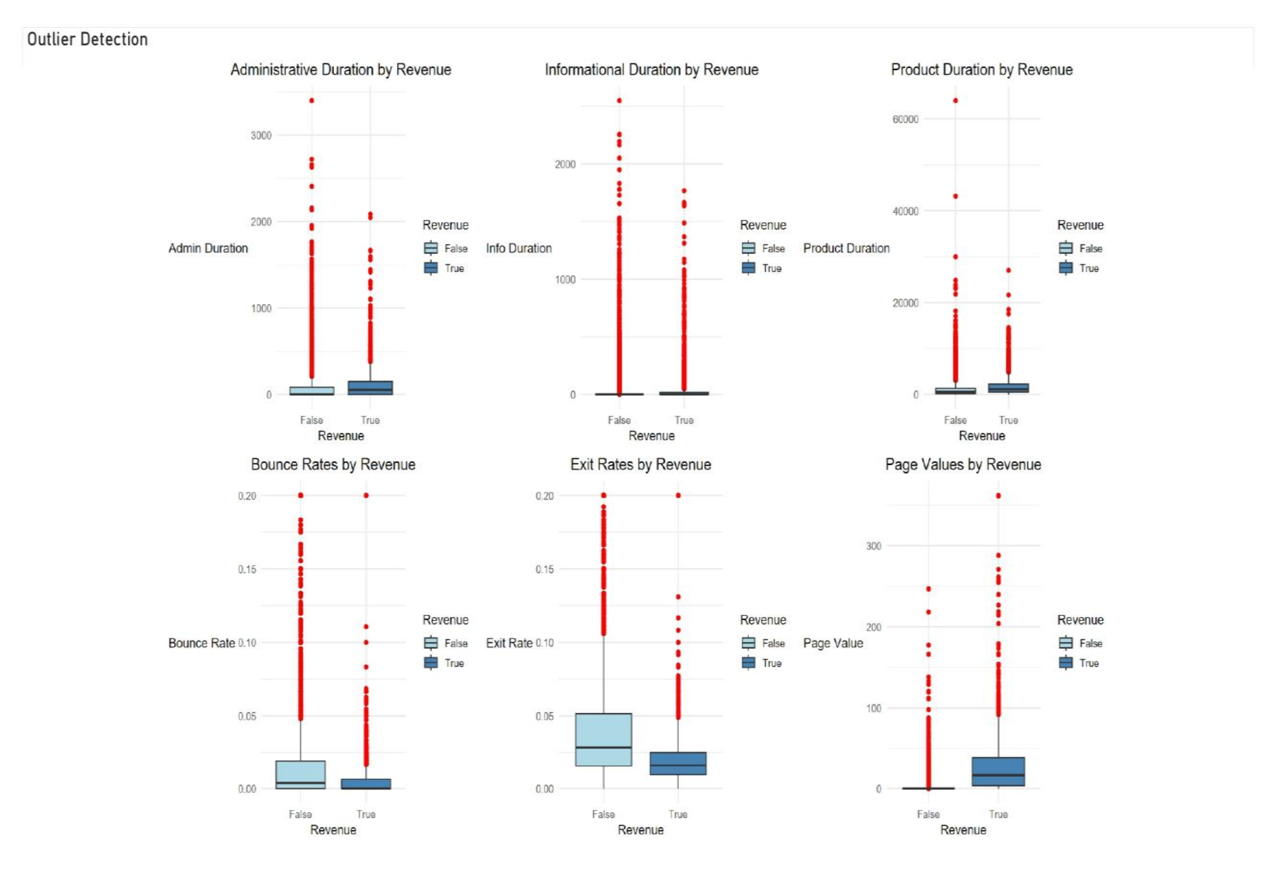
UMAP Plot

- Example run with fixed parameters.
- Captures **non-linear relationships** and complex data structure.
- Reveals overlapping clusters → classification is challenging.
- Indicates that advanced, non-linear models may be more effective.





EXPLORATORY DATA ANALYSIS- DISTRIBUTION ANALYSIS

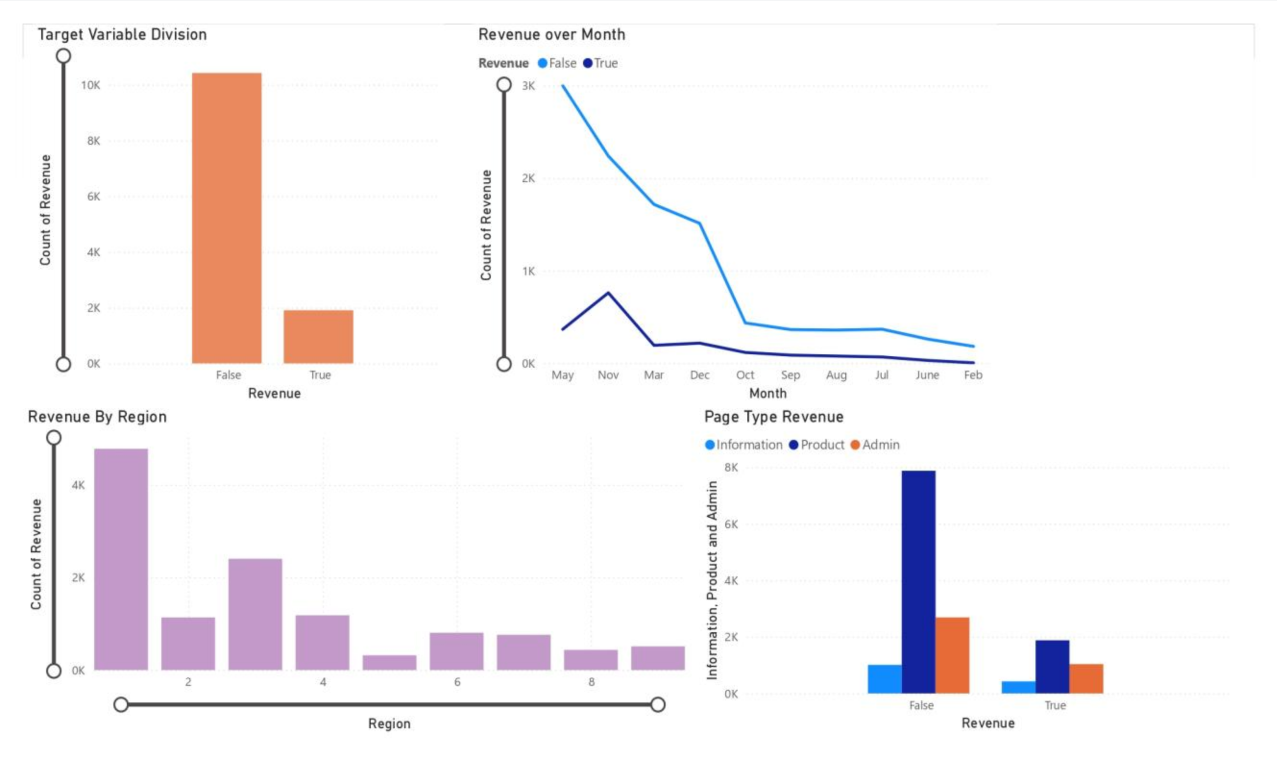


Revenue division has more of false that shows class imbalance

November and May has a peak in revenue suggesting seasonal hikes

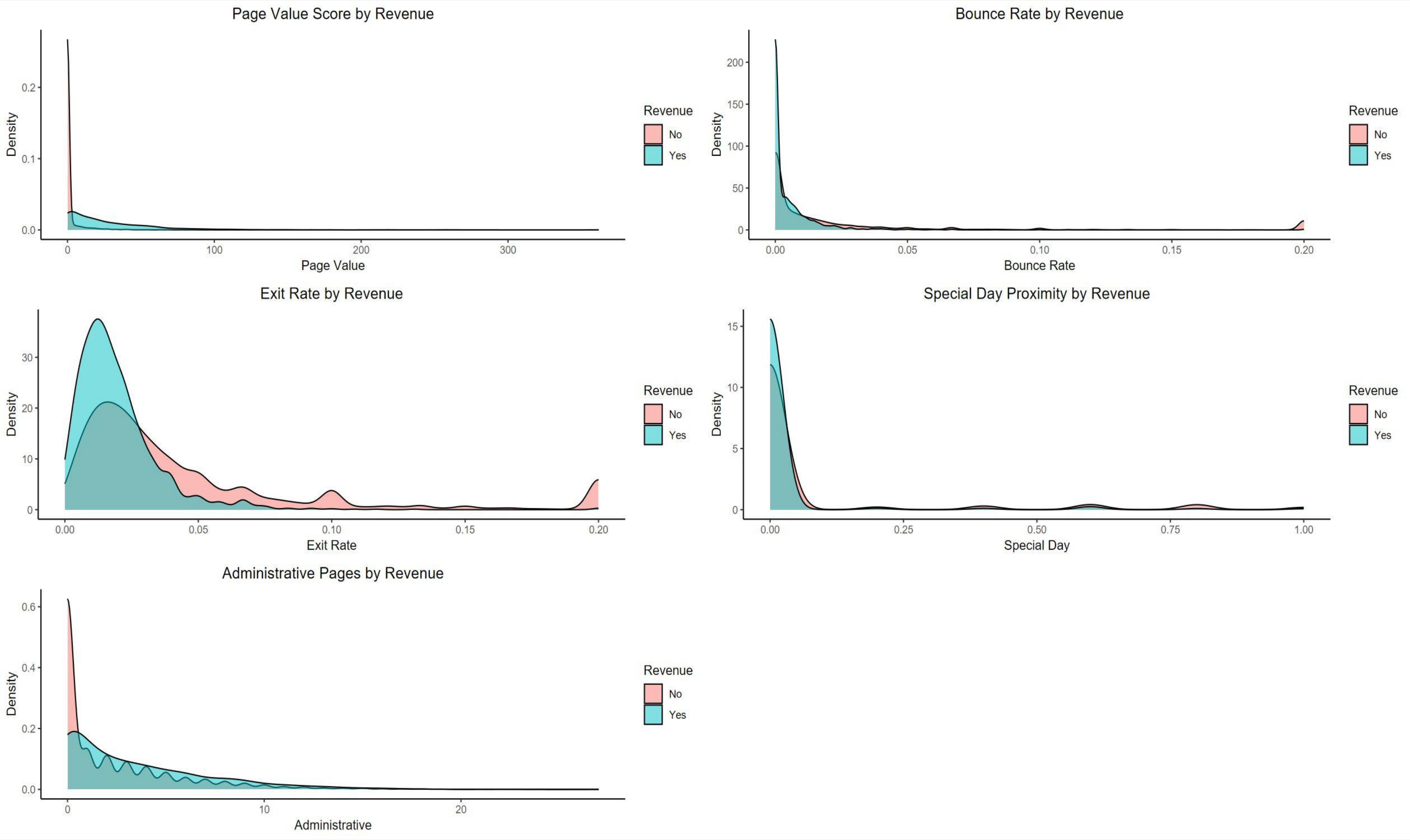
Region 1 has highest revenue followed by 3

Product related duration is higher than Informational and Administration as expected due to the website being a retail page



This set of boxplots visualises the distribution of key numeric features—such as page duration, bounce rate, exit rate, and page value—grouped by session revenue outcome (`Yes` vs. `No`). Notably, several variables show significant differences in spread and central tendency between converting and non-converting sessions. For instance, sessions that generated revenue tend to have higher values across engagement-related features, suggesting stronger user intent and interaction.

EXPLORATORY DATA ANALYSIS- DISTRIBUTION ANALYSIS



These plots compare distributions **between sessions that did and did not result in revenue generation**

- **Page Values**:

Revenue-generating sessions have a **notably higher density** at elevated Page Values. This confirms that **higher page value correlates with revenue**.

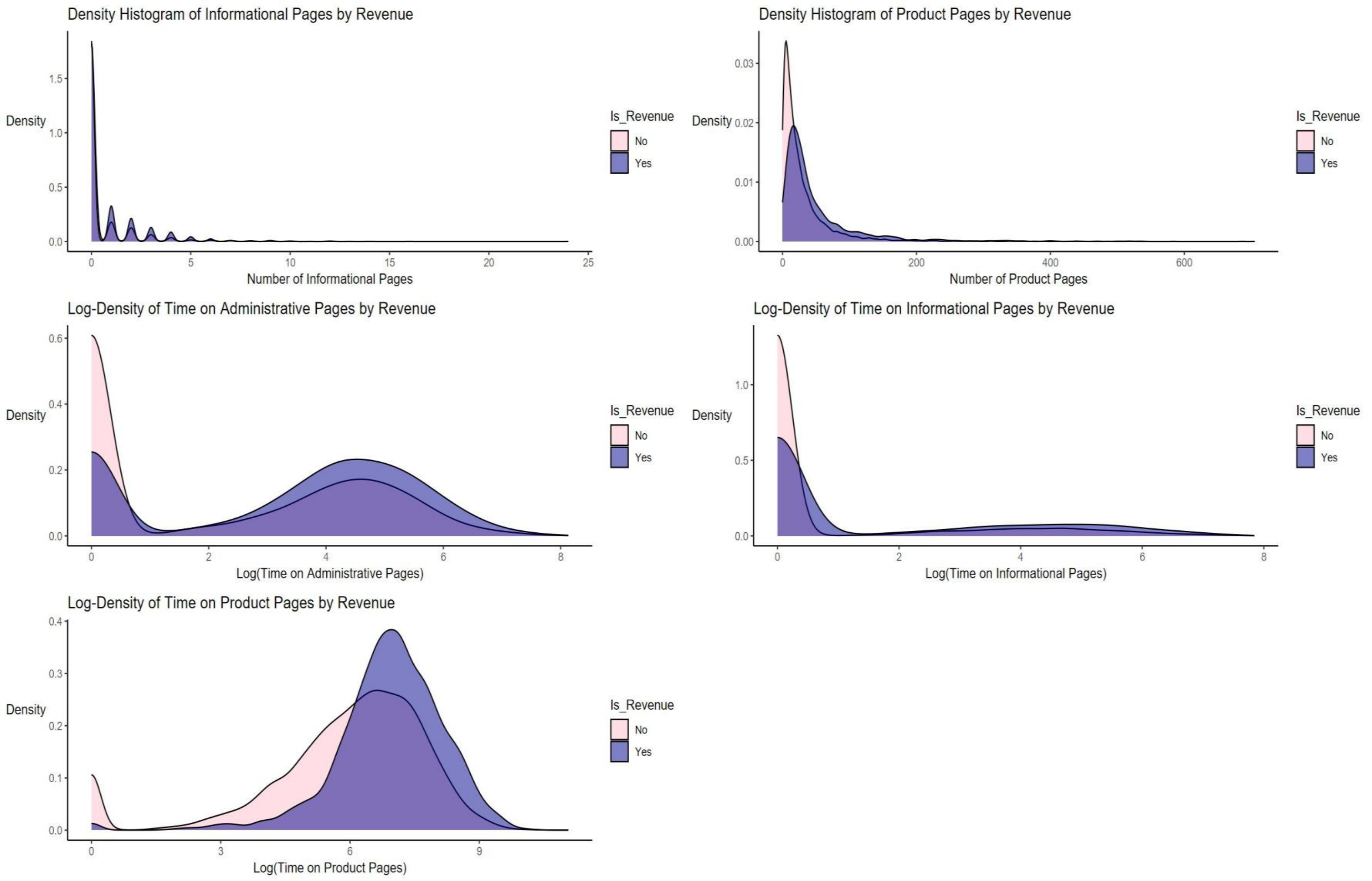
- **Exit Rate & Bounce Rate**:

Sessions without revenue tend to have **higher bounce and exit rates**.
Revenue sessions cluster more at **lower bounce and exit rates**, suggesting better engagement.

- **Administrative Pages & Special Day Proximity**:

Revenue sessions generally have a **slightly higher density** at low Administrative values.
Special Day shows almost **no clear distinction** between revenue and non-revenue sessions — this might not be a strong predictor.

EXPLORATORY DATA ANALYSIS- DISTRIBUTIONANALYSIS



- **Product Page Views**:

Revenue users view **20–100+ product pages**.
Non-revenue users mostly <10 pages

- **Time on Product Pages (log seconds)**:

Revenue users: peak at **log(6–7)**
Non-revenue: peak at **log(4–5)**

Informational Pages:

- Slightly higher count and time for revenue users.
- Modest impact.

Administrative Pages:

- Revenue users show **second peak at log(4–5)** suggesting deeper engagement.

Summary of findings:

- Most continuous variables are **skewed** → log-transform required for modelling.
- **High PageValue, low BounceRate/ExitRate** correlate with revenue generation.
- **Product engagement** (page views & time) is the **strongest revenue signal**.
- Informational/Admin content plays a **supporting role**, but less predictive.

DATA BALANCING APPROACH

Why Balancing was needed

- Approximately 15% of user sessions resulted in revenue generation, while 85% did not result in revenue generation.
- Class imbalance can cause models to favour the majority class, which weakens their ability to predict outcomes for the minority class.

ROSE Method (Random Over-Sampling Examples)

- Generates synthetic samples for the minority class.
- Helps create a more balanced dataset without simply duplicating existing rows.
- Preserves underlying feature distribution for better generalisation.

Random Stratified Over Sampling Advantages

- Ensures proportional representation of both classes in the resampled dataset.
- Reduces sampling and representation bias in classification models.
- Chosen to improve the model's ability to detect *Revenue* = *Yes* cases.

Before and After balance check (see table below for balancing results)

- Result: A more balanced dataset, leading to fairer representation of both classes and better model performance.

Yes	No
1,908 (15%)	10,422 (85%)

Yes	No
6,056 (49.6%)	6,149 (50%)

SUMMARY OF CLEAN DATASET

Dataset Balance (Post-Cleaning)

- Records: 12,205
- Revenue = Yes: ~49.6%
- Revenue = No: ~50%

Feature Transformation

- Applied log transformations to reduce skewness in time and page variables.
- Discretized selected continuous features to capture non-linear patterns.
- Standardised variables for consistency.

Feature Encoding

- Categorical variables such as *Operating System*, *Browser*, *Traffic Type*, *Visitor Type*, and *Month* were transformed into factors for analysis.
- High-cardinality categorical variables such as *Traffic Type* and *Browser* were grouped into an “Other” category to reduce sparsity.
- For logistic regression and random forest models, categorical factors were handled internally through dummy coding.
- For models such as logistic regression and XGBoost, categorical variables were explicitly one-hot encoded using model.matrix().

Target Variable Encoding

- The target variable Is_Revenue was originally stored as booleans (TRUE/FALSE).
- It was converted to a factor variable with “No” set as the reference category to ensure consistent interpretation in classification models.
- For some analyses and correlation checks, Is_Revenue was temporarily converted to a numeric variable (0 = No, 1 = Yes) which enabled ROC curve generation, probability predictions, and logistic regression analysis.
- Consistent factor encoding ensured that cross-validation and resampling methods (e.g., K-fold CV) worked correctly across all models.

Prepared for Modelling

- Cleaned, balanced, transformed, and validated for stable model training.

LOGISTIC REGRESSION MODEL

Overview: Logistic regression is a classic and interpretable statistical method commonly used for binary classification. It was selected for its simplicity, transparency and ability to provide clear insights regarding the influence of individual features on the target outcome.

Training Method: The model was trained using 10-fold cross-validation. In each iteration, the model was trained on nine folds (~11,097 rows) and tested on the remaining (~1,233 rows). This approach helps minimise bias and variance, ensuring that all observations are used both in training and evaluation.

Tuning Method: No complex tuning was required beyond standard preprocessing, as Logistic regression has relatively few hyperparameters. However, regularization was applied as it has in-built options such as L1 (Lasso) which was useful for penalising features that did not improve model accuracy.

Step 1: Baseline Model

- Includes all features from the dataset
- Ideal for benchmarking performance
- This model was refined using various techniques to create a parsimonious model that included meaningful predictors and lowered AIC.
- **AIC: 9,940**

Step 2:Tuning Parameters and Feature Selection using Lasso Regression (via glmnet)

- Purpose: penalise complexity and automatically selects features by shrinking coefficients to zero.
- Best tuning Parameters: Lambda (regularization strength) = **Seq(0.0001, 0.1, length.out = 10) and Alpha = 1 (lasso penalty).**
- Benefits: prevents overfitting and handles multicollinearity.

Step 3:Feature Selection using STEPWISE SELECTION (via AIC)

- Direction: Both forward and backward
- Criteria: Akaike Information Criterion (AIC)
- balanced models fit and complexity
- Lower AIC indicates better trade-off
- Benefits: selects parsimonious model with reasonable explanatory power.
- **AIC: 9,915**

Step 4: P-value Based Elimination

- Custom iterative function repeatedly removes variables with P-values > 0.05 (statistically insignificant)
- Stops when all the remaining predictors are statistically significant
- Benefit: ensures only meaningful predictors are kept for interpretability.
- Final model performance was compared with random forest and XGBoost model.
- **AIC: 9,921**
- Although there is a slight increase in AIC compared to stepwise model, this is outweighed by the model's reduced complexity and interpretability making it more practical for business use/decision-making.

Benefits of using all three approaches (tuning & feature selection):

- Robustness (Lasso)
- Simplicity and interpretability (StepAIC)
- Statistical validity (P-value filtering)

RANDOM FOREST MODEL

Overview: Random Forest is an ensemble learning algorithm that forms multiple decision trees and combines outputs (aggregating their results) to improve predictive accuracy and stability. It was selected for its robustness to overfitting, ability to model non-linear relationships and effectiveness in capturing complex interactions.

Training Method: Model training was performed using 10-fold cross-validation to ensure reliable performance across different data partitions.

Tuning Method:

- A grid search procedure was used to identify optimal hyperparameters, such as number of trees, tree depth, and minimum samples per split to maximise performance while reducing overfitting.

Table 1: Tuning Grid

Parameters	Value
mytry (number of variables tried at each split)	2, 4, 6
nodesize (minimum size of terminal nodes)	5, 10
Splitrule	gini (held constant)

Table 2: Best Model Parameters

Parameters	Best Tune Value
mytry (number of variables tried at each split)	6
nodesize (minimum size of terminal nodes)	5
Splitrule	gini (held constant)

XGBOOST MODEL

Overview: The XGBoost model is a powerful and efficient gradient boosting algorithm designed for speed and performance. Selected due to its reputation for delivering strong results in many classification tasks by combining gradient boosting with advanced regularization techniques to help overfitting.

Training Method: Model training was performed using 10-folds cross validation to ensure robust evaluation and generalisation across different subsets of data.

Tuning Method:

- The grid search approach was employed to systematically explore combinations of hyperparameters such as learning rate, max depth and subsample ratio. This process identified optimal settings that maximise model performance, balancing bias and variance.

Table 3: Tuning Grid

Parameters	Tune
nrounds	100
max_depth	4 to 6
eta (learning rate)	0.1
gamma	0
colsample_bytree	0.8
subsample	0.8
min_child_weight	1

Table 4: Best Model Parameters

Parameters	Best Tune Value
nrounds	96
max_depth	6
eta (learning rate)	0.1
gamma	0
colsample_bytree	0.8
subsample	0.8
min_child_weight	1

PR CURVE COMPARISON

Precision-Recall AUC as a key metric:

- Although the dataset used for modelling has been balanced, the original data was highly imbalanced, with far fewer positive cases (buyers) compared to non-buyers.
- Because of this initial imbalance, traditional accuracy metrics could still be misleading, as models might struggle to correctly identify the minority class.
- The Precision- Recall AUC provides a comprehensive summary of this balance, making it the most meaningful metric for assessing model performance in this context.
- To ensure a thorough evaluation additional performance metrics such as ROC AUC, F1 score and confusion matrix among others were used to analyse prediction errors.

Logistic Regression:

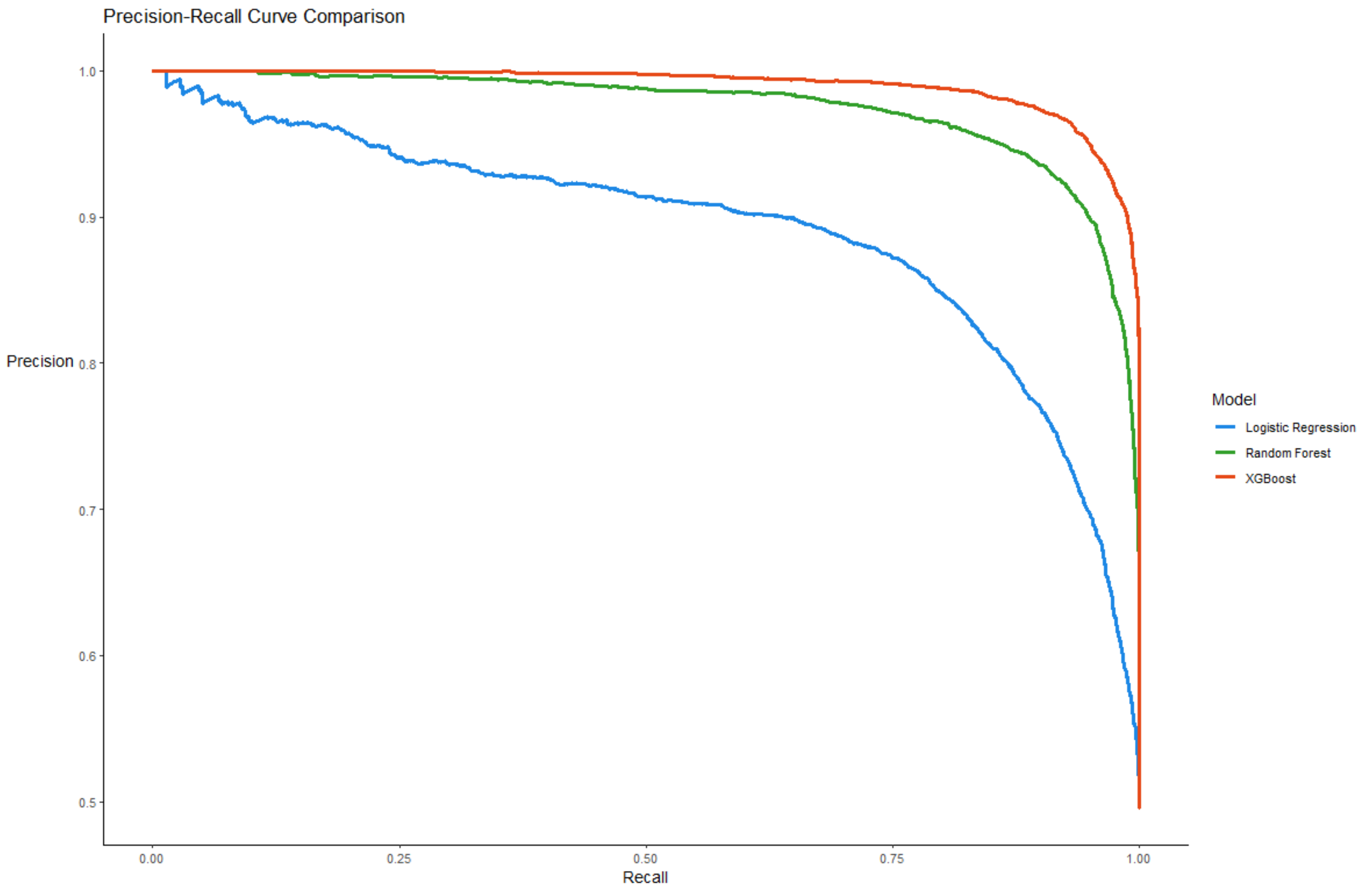
- Achieved a PR AUC of **0.89**, which indicates that the model is the least effective at distinguishing positive instances compared to the tree-based models. This suggests that it may struggle with capturing complex patterns in the data.

Random Forest:

- Improved performance with a PR AUC of **0.97**, demonstrating strong ability to identify positives while maintaining precision.

XGBoost:

- Best performing model with a PR AUC of **0.99**, showing excellent balance between precision and recall and capturing intricate relationships in the dataset.



PERFORMANCE METRICS

Logistic Regression Results:

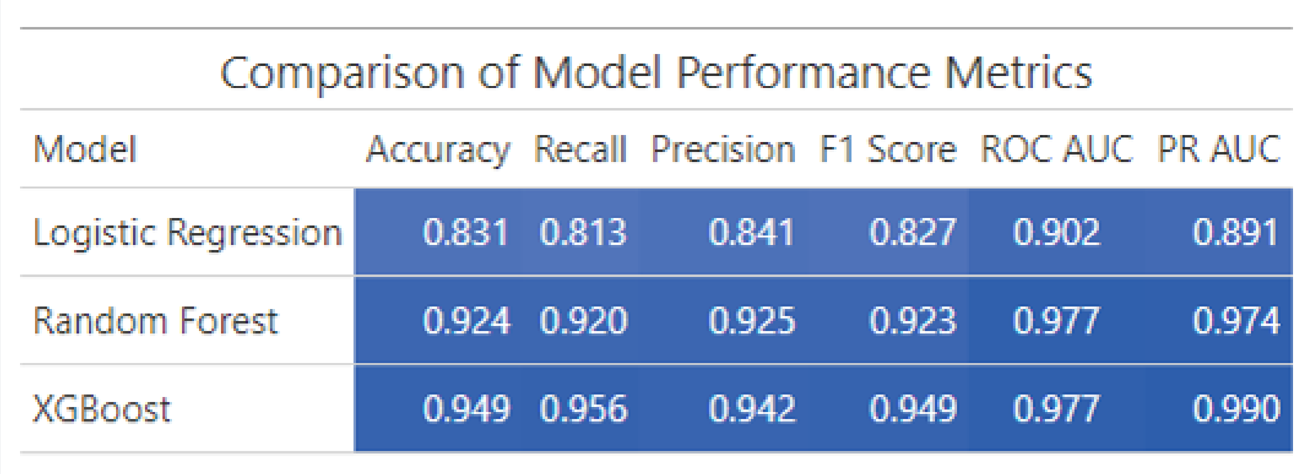
- The logistic regression model was the lowest performing model across all classification metrics.
- The model produced an accuracy score of **0.83**, indicating that 83% of predictions were correct overall.
- In addition, the model achieved a recall of **0.81**, indicating that it correctly identified 82% of actual buyers.
- The precision score was **0.84**, suggesting that when the model predicted that users would buy, it was correct 84% of the time.
- The F1 score, which balances precision and recall was **0.83**.
- The ROC AUC was **0.91**, demonstrating strong ability to distinguish buyers and non-buyers, though this is not as high as the result for the tree-based models.

Random Forest Results:

- The random forest model showed improved performance by obtaining an accuracy score of **0.93**, indicating that 93% of predictions were correct overall.
- It had a recall of **0.92**, highlighting its strong ability to correctly identify actual buyers and a precision score of **0.93**, showing increased reliability in positive predictions.
- The F1 score of **0.93** demonstrates a well-balanced performance.
- The ROC AUC score of **0.98** indicates excellent separation between classes, reflecting the model's robustness in classifying buyer intent.

XGBoost Results:

- XGBoost delivered the strongest results by outperforming the other models as indicated by the key performance metrics.
- Accuracy was **0.95**, the highest coming all models.
- Recall was the highest at **0.96** which indicates that the model correctly identified actual buyers (revenue-generating sessions) 96% of the time. This suggests that the model is highly effective at capturing the positive class and minimising the false negatives (**4%**) which is critical when the goal is to not miss potential buyers. Thus, this model is suitable as the priority is to maximise detection of actual buyers.
- The precision score was **0.94** which indicates that the model correctly predicted buyers as real buyers 94% of the time, which is an excellent score considering that it mostly avoids false positives.
- The F1 score was **0.95** which indicates a strong balance between precision and recall. This means it was both accurate in identifying actual buyers and effective at minimising false positives.
- The model achieved the same ROC AUC as the random forest model, indicating near-perfect ability to distinguish classes.
- These results are in line with the findings from Abdullah-All-Tanvir et al. (2023) and Deniz & Bülbül (2024). This consistency reinforces the effectiveness of ensemble models in handling behavioural prediction tasks in e-commerce contexts.



Comparison of Model Performance Metrics

Model	Accuracy	Recall	Precision	F1 Score	ROC AUC	PR AUC
Logistic Regression	0.831	0.813	0.841	0.827	0.902	0.891
Random Forest	0.924	0.920	0.925	0.923	0.977	0.974
XGBoost	0.949	0.956	0.942	0.949	0.977	0.990

STRENGTHS AND LIMITATIONS

Model	Strengths	Limitations
Logistic Regression	• Highly interpretable (coefficients show feature effects). • Fast to train. • Works well with linearly separatable data.	• Lower predictive performance. • Assumes linear relationships. • May underperform with complex patterns.
Random Forest	• Handles non-linear relationships. • Robust to overfitting. • Good with imbalanced datasets (if tuned).	• Less interpretable than Logistic Regression. • Slower to train with many trees. • Performance not as high as XGBoost.
XGBoost	• High predictive accuracy. • Handles complex feature interactions and non-linear relationships. • Includes regularization.	• Complex to tune. • Less interpretable (“black-box”). • Can overfit if not managed.

FEATURE IMPORTANCE

Importance Measurement by Model Type

Logistic Regression: Absolute Coefficient

- Magnitude of coefficient.
- Larger the absolute coefficient means that a feature has more influence on the likelihood of the outcome.

Random Forest: Permutation

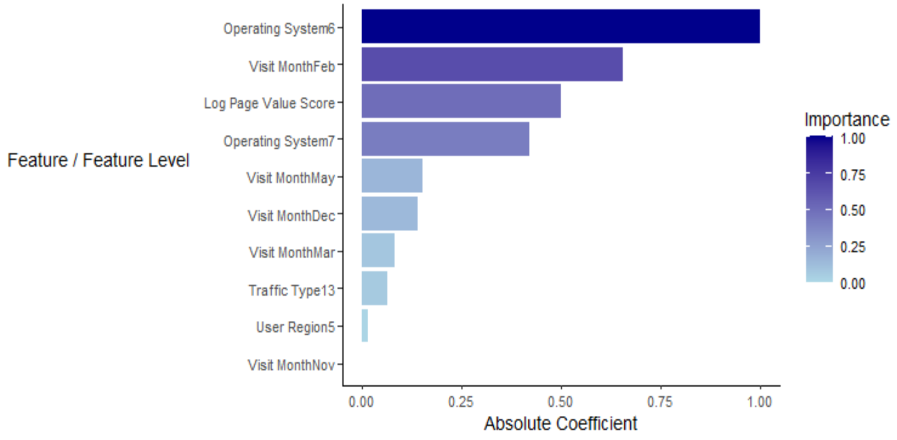
- Randomly shuffles one feature at a time and sees how much the model performance drops
- If the performance drops significantly, the feature was important.

XGBoost: Gain

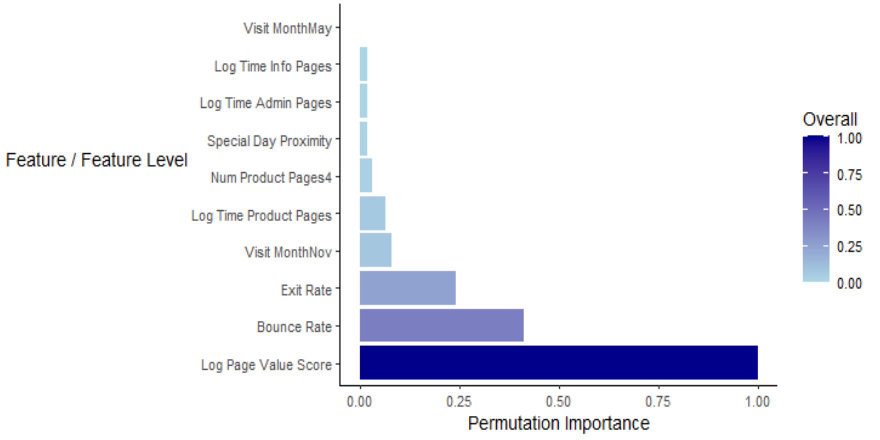
- It represents the improvement in performance accuracy (reducing loss) brought by a feature each time it's used to split.
- Higher gain implies that a feature is informative.

Results:

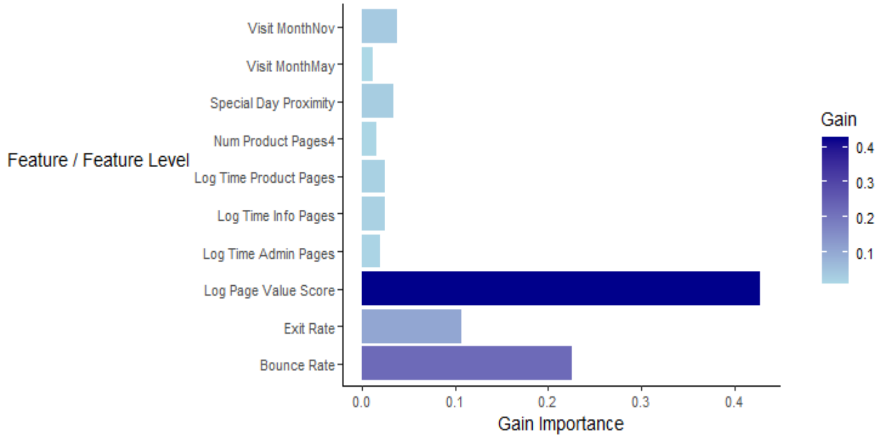
- Among all features, the XGBoost model determined that page value score was the most important predictor of purchase likelihood, followed by bounce rate, exit rate and visit month November.
- The results for both tree-models (XGBoost and Random Forest) were very similar. Reflecting their ability to effectively capture the complex and non-linear relationships.
- However, the logistic regression model had very different results, with operating system 6 the most important predictor followed by visit Month Feb and page value score.



Feature Importance for Logistic Regression Model



Feature Importance for Random Forest Model



Feature Importance for XGBoost Model

CONTEXTUALISING MODEL EXPLANATIONS

Feature	Model Statistical Insight (SHAP Value)	Comparison
Page Value (log)	• The strongest predictor of purchasing intent obtained a mean Shap value of approximately 0.2, indicating a moderate positive contribution to the model's prediction. In this case, higher values of the feature increase the likelihood of conversion. • This suggests that greater user engagement with high-value pages is strongly associated with purchasing behaviour, reinforcing the importance of content-rich interactions in driving conversions.	• The results align with prior research that emphasises the importance of user engagement and content-rich page interactions in driving purchasing intent. For example, Close and Kukar-Kinney (2009) identified motivations such as information search, entertainment and organisation as key drivers of online cart use and purchase behaviour – suggesting that engagement extends beyond immediate transactional intent. • Similarly, Markov chain-based attribution models recognise the role of product and category pages in shaping the user's journey, even if they are not the final touchpoint before conversion. This finding that higher page value scores yield positive SHAP values (~0.2) supports this view, indicating that value-rich, informative pages contribute meaningfully to purchase intent, even if they are not directly associated with the final purchase click. • Thus, the SHAP analysis reinforces the idea that user interactions across multiple high-value pages cumulatively influence conversion, highlighting the importance of moving beyond last-click attribution to understand purchasing behaviour holistically.
Bounce Rate	• Interestingly, the obtained a mean SHAP value indicates a low to moderate influence (approximately -0.1 or less). The negative sign suggests that this feature level is associated with a lower probability of purchase. While the effect is not strong, it does imply that users represented by this feature level are less likely to convert. • This may reflect browsing/exploratory behaviour without strong purchasing intent – users are engaging with the site but not exhibiting actions that significantly drive purchase predictions.	• The model insights resonate with findings from Cialdini (2001), who noted that users initially motivated to browse can be persuaded to buy later through attractive incentives. This suggest that while the model predicts low conversion likelihood for these browser users, they represent a segment with potential for conversion if targeted effectively with offers or promotions. Thus, the combination of SHAP-based model interpretation and behavioural research highlights an opportunity to nurture browsers into buyers with well-timed incentives.
Exit Rate	• Similarly, the mean SHAP value indicates a low to moderate influence (approximately -0.1 or less). The negative sign indicates a small downward contribution to the model's purchase likelihood. This suggests that users with this this exit rate are less likely to purchase, although the effects are minimal. • These results may imply that users are not exiting the site immediately, but also not taking actions associated with conversion – reflecting low purchase intent despite some level of engagement.	• The results for the mean SHAP value for exit rate align with prior findings. For example, Close and Kukar-Kinney (2010) observed that many users use online carts as wish lists or for comparison, not for immediate purchases. Similarly, hedonic browsers (those engaging with for entertainment) often stay longer without goal-directed actions. As such, simply remaining on the site does not necessarily reflect high purchase intent, echoing what the SHAP value suggests. • Moreover, Shafir et al. (1993) found that visible, time-limited discounts can successfully convert general browsers into buyers. McConnell et al. (2000) also noted price guarantees can discourage comparison shopping and drive users to complete purchases. • Together, these findings support the interpretation that prolonged engagement without conversion signals low intent, but with the right incentives (e.g. limited time offers or price reassurances) this segment can still be influenced toward conversion.

RECOMMENDATIONS AND NEXT STEPS

Page Value:

Major driver of purchasing intention.

Action:

- Optimise high-value pages (e.g. product information and category) to better support purchase journeys.
- Use recommendation engines to increase time spent on valuable content.
- Implement multi-touch attribution (e.g. Markov chains) to capture the true impact of page sequences.
- Incorporate AI/ML tools (predictive search and personalised recommendations) to align content with user intent in real time.

Bounce Rate:

low purchase probability- even if users stay, they may not convert (these users are likely in the consideration phase).

Action:

- Enhance site functionality to help users quickly find what they need.
- Highlight key benefits (e.g. free returns, fast delivery) throughout the journey.
- Introduce nudges and reassurance, such as simplified checkout or chatbot support.
- Use social proof (e.g. best sellers, reviews, live activity) to reduce hesitation and build trust.

Exit Rate:

Neither exiting or converting – likely browsing aimlessly.

Action:

- Use exit-intent popups with time limited discounts to prompt action.
- Introduce goal-driven navigation (e.g. *“Complete the look”, “Customers also bought”*) to guide purchasing.
- Segment cart users who bookmark items and target them with timed offers and low stock alerts.
- Offer price guarantees to reduce comparison shopping and encourage conversion.

UNIVERSITY OF
SURREY

THANK YOU !!

REFERENCES

Andriy Burkov (2019). *THE HUNDRED-PAGE MACHINE LEARNING BOOK*. Andriy Burkov.

Bell, L., McCloy, R., Butler, L. and Vogt, J. (2020). Motivational and Affective Factors Underlying Consumer Dropout and Transactional Success in eCommerce: An Overview. *Frontiers in Psychology*, 11. doi:https://doi.org/10.3389/fpsyg.2020.01546.

Borges, J. and Levene, M. (2007). Evaluating Variable-Length Markov Chain Models for Analysis of User Web Navigation Sessions. *IEEE Transactions on Knowledge and Data Engineering*, 19(4), pp.441–452. doi:https://doi.org/10.1109/tkde.2007.1012.

Close, A.G. and Kukar-Kinney, M. (2010). Beyond buying: Motivations behind consumers' online shopping cart use. *Journal of Business Research*, 63(9-10), pp.986–992. doi:https://doi.org/10.1016/j.jbusres.2009.01.022.

Deniz, E. and Semanur Çökekoğlu Bülbül (2024). Predicting Customer Purchase Behavior Using Machine Learning Models. *Information Technology in Economics and Business* , [online] 1(1). doi:https://doi.org/10.69882/adba.iteb.2024071.

Gao, B., Liu, T.-Y., Liu, Y., Wang, T., Ma, Z.-M. and Li, H. (2011). Page importance computation based on Markov processes. *Information Retrieval*, 14(5), pp.488–514. doi:https://doi.org/10.1007/s10791-011-9164-x.

Gkikas, D.C. and Theodoridis, P.K. (2024). Predicting Online Shopping Behavior: Using Machine Learning and Google Analytics to Classify User Engagement. *Applied Sciences*, [online] 14(23), p.11403. doi:https://doi.org/10.3390/app142311403.

Huang, J.Z. (2014). An Introduction to Statistical Learning: With Applications in R By Gareth James, Trevor Hastie, Robert Tibshirani, Daniela Witten. *Journal of Agricultural, Biological, and Environmental Statistics*, 19(4), pp.556–557. doi:https://doi.org/10.1007/s13253-014-0179-9.

Mir, I.A. (2021). Self-Escapism Motivated Online Shopping Engagement: a Determinant of Users' Online Shopping Cart Use and Buying Behavior. *Journal of Internet Commerce*, 22(1), pp.1–34. doi:https://doi.org/10.1080/15332861.2021.2021582.

REFERENCES

Office for National Statistics (2025). *Internet Sales as a Percentage of Total Retail Sales (ratio) (%) - Office for National Statistics*. [online] Ons.gov.uk. Available at: https://www.ons.gov.uk/businessindustryandtrade/retailindustry/timeseries/j4mc/drsi.

Statista (2025). *Online shopping cart abandonment rate worldwide between 2006 to 2025*. [online] Statista. Available at: https://www-statista-com.surrey.idm.oclc.org/statistics/477804/online-shopping-cart-abandonment-rate-worldwide/ [Accessed 3 Aug. 2025].

Tanvir, A.-A., Ali Khandokar, I., Muzahidul Islam, A.K.M., Islam, S. and Shatabda, S. (2023). A gradient boosting classifier for purchase intention prediction of online shoppers. *Heliyon*, 9(4), p.e15163. doi:https://doi.org/10.1016/j.heliyon.2023.e15163.

Team, A.C. (2022). *Ecommerce bounce rate — what it is and how to improve it*. [online] Adobe.com. Available at: https://business.adobe.com/blog/basics/ecommerce-bounce-rate#what-is-the-average-ecommerce-bounce-rate [Accessed 3 Aug. 2025].

Thiyagarajan, G. and Swathi, Y. (2025). Temporal Dynamics of Consumer Engagement in E-Commerce. In: *In Proceedings of the 2025 International Conference on Computing for Sustainability and Innovation (COMP-SIF)*. [online] Available at: https://www.researchgate.net/publication/391257130_Temporal_Dynamics_of_Consumer_Engagement_in_E-Commerce [Accessed 1 Aug. 2025].

Wolfgang Jank (2011). *Business analytics for managers*. New York: Springer.

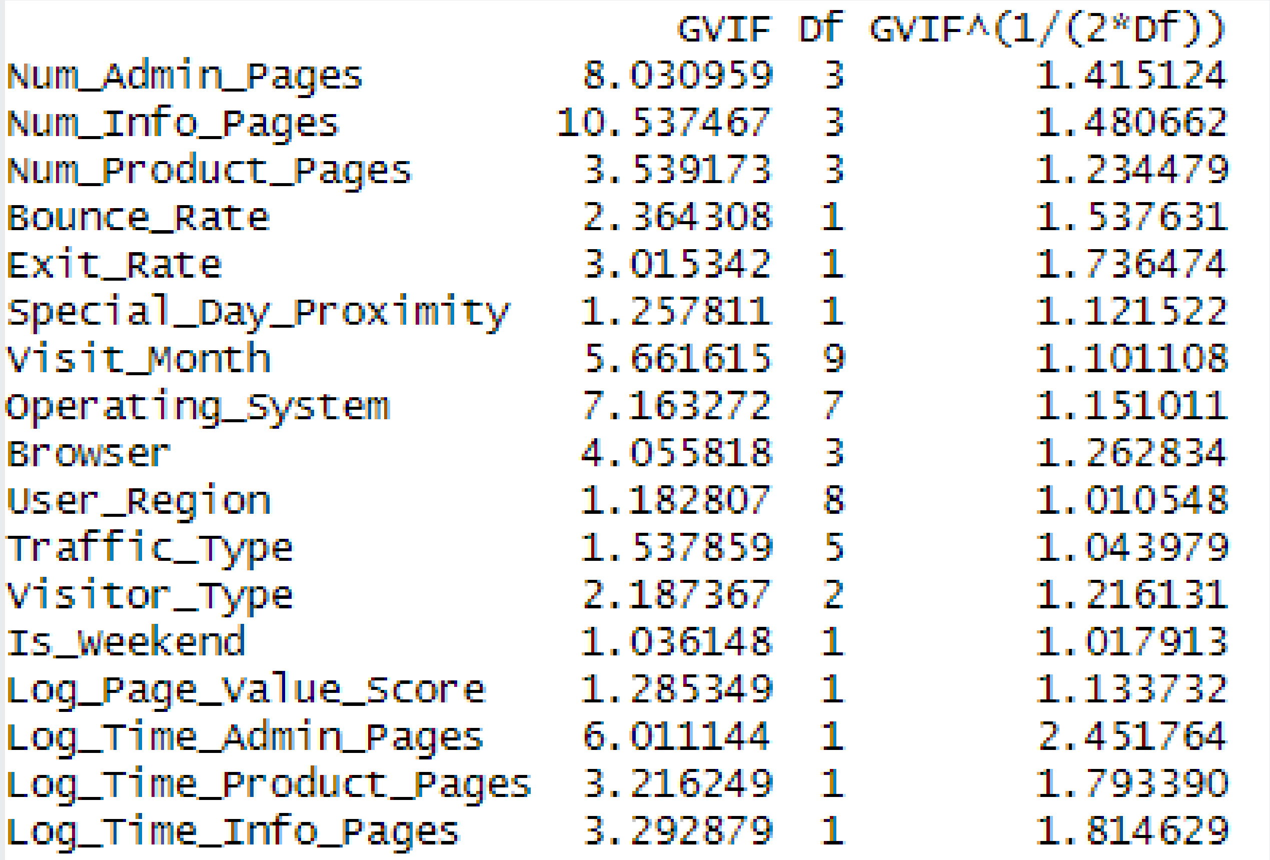
APPENDICES

The table lists all variables from the original UCI Online Shopping Purchase Intention dataset with their original column names. Each variable's description is provided to clarify its meaning and help interpret the dataset's features and target variable.

Variable Name	Description
Administrative	Number of administrative pages visited during a session
Administrative_Duration	Total time spent on administrative pages (seconds)
Informational	Number of informational pages visited
Informational_Duration	Time spent on informational pages (seconds)
ProductRelated	Number of product-related pages visited
ProductRelated_Duration	Time spent on product-related pages
BounceRates	Percentage of users who leave the site after viewing one page
ExitRates	Percentage of users who exited the website from a specific page
PageValues	Average value of a webpage based on transaction completion and navigation data
SpecialDay	Closeness of the session date to a special day (e.g. Valentine's day) (0 to 1)
Month	The month when the visit happened during the year
OperatingSystems	Operating system used by the visitor
Browser	Browser used by the visitor
Region	Visitor's geographic region
TrafficType	Source of the website traffic (e.g. direct, referral)
VisitorType	Type of visitor: Returning, New, or Other
Weekend	Whether the session took place on a weekend (Boolean TRUE/FALSE)
Revenue	Target variable – whether the visit resulted in a purchase (Boolean TRUE/FALSE)

APPENDIX A: VARIABLE INFLATION FACTOR (VIF) RESULTS

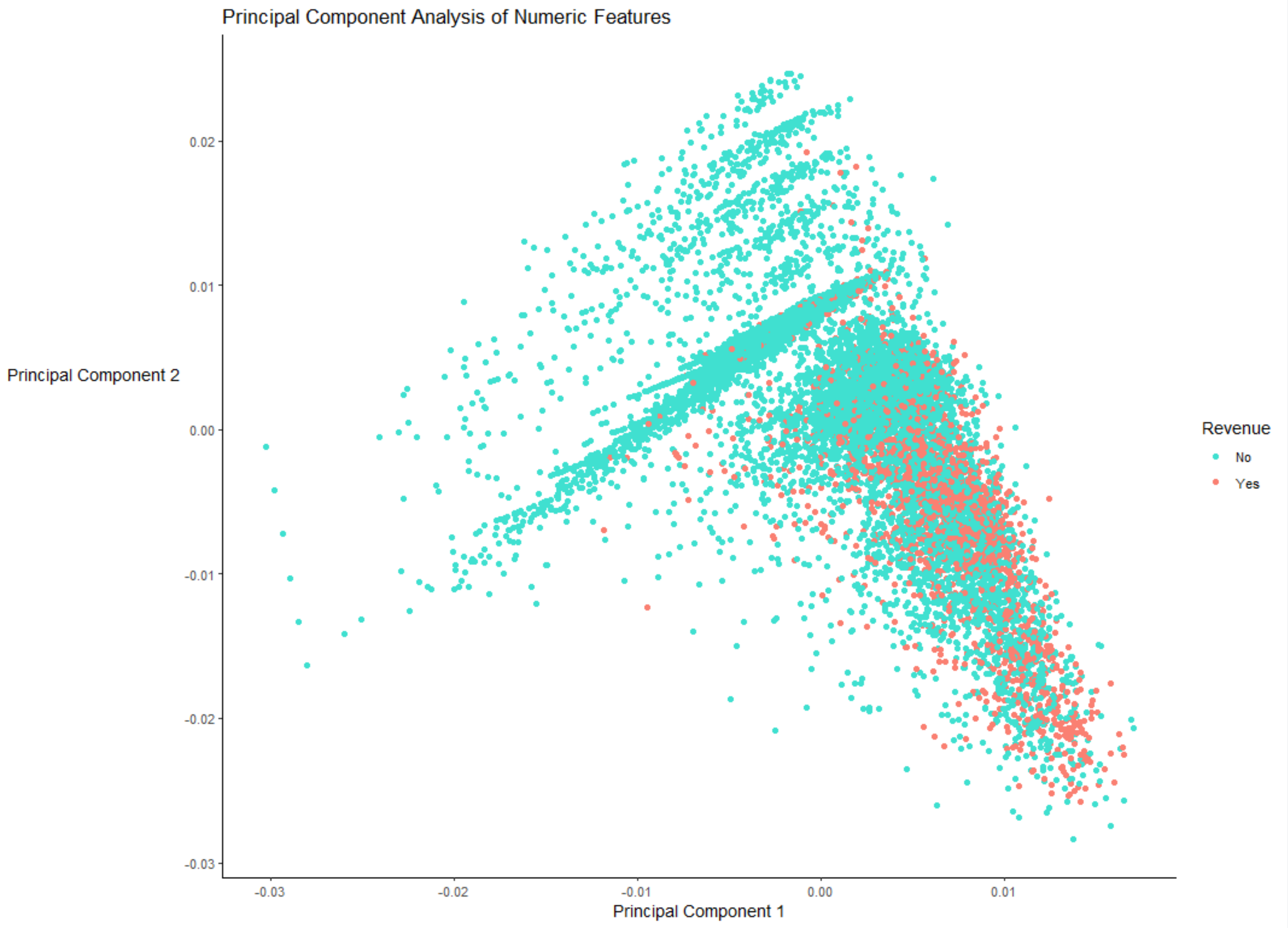
- To ensure that there was no harmful multicollinearity among predictors, a VIF analysis was conducted after variable transformation, but before balancing and scaling to diagnose multicollinearity in the original predictor space.
- The adjusted generalised VIF values (GVIF^(1/(2*Df))) were used for interpretation, as several categorical variables had more than two levels.
- All adjusted VIF values were below the common threshold of 5, indicating that none of the predictors were highly correlated with one another. This indicates that multicollinearity is not a concern in the dataset, and all predictors were retained for modelling.



```
                           GVIF Df GVIF^(1/(2*Df))
Num_Admin_Pages        8.030959  3        1.415124
Num_Info_Pages        10.537467  3        1.480662
Num_Product_Pages      3.539173  3        1.234479
Bounce_Rate            2.364308  1        1.537631
Exit_Rate              3.015342  1        1.736474
Special_Day_Proximity  1.257811  1        1.121522
Visit_Month            5.661615  9        1.101108
Operating_System       7.163272  7        1.151011
Browser                4.055818  3        1.262834
User_Region            1.182807  8        1.010548
Traffic_Type           1.537859  5        1.043979
Visitor_Type           2.187367  2        1.216131
Is_Weekend             1.036148  1        1.017913
Log_Page_Value_Score   1.285349  1        1.133732
Log_Time_Admin_Pages   6.011144  1        2.451764
Log_Time_Product_Pages 3.216249  1        1.793390
Log_Time_Info_Pages    3.292879  1        1.814629
```

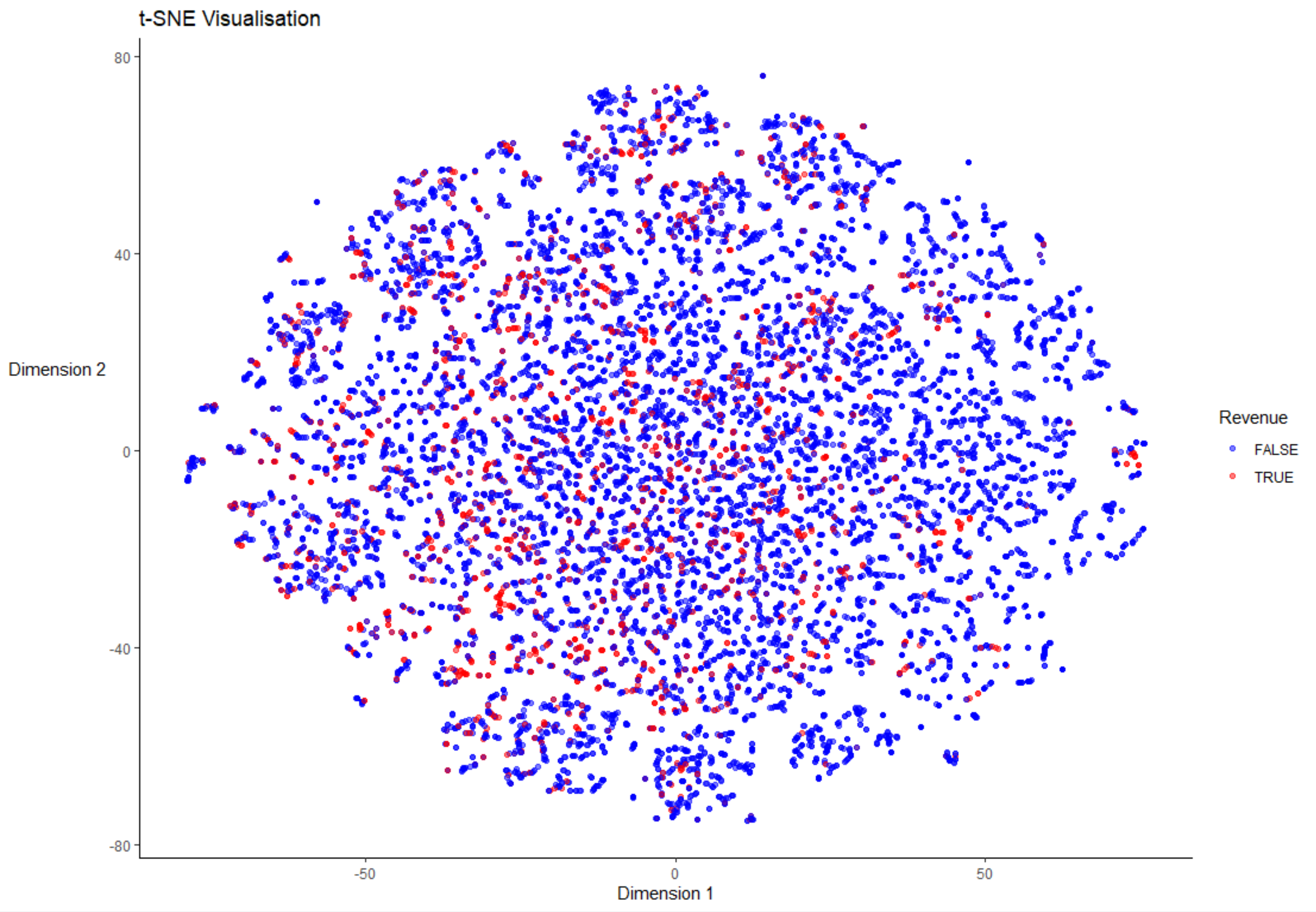
APPENDIX B:PRINCIPAL COMPONENT ANALYSIS (2D PLOT)

- Data forms a dense cluster of observations with limited spread.
- Significant overlap between “Yes” and “No” cases indicates weak separability.
- Some “Yes” cases appear at edges, but not enough for clear distinction.
- Confirms lack of strong linear separability
- Indicates the need for non-linear methods (e.g., UMAP, t-SNE, advanced classifiers).



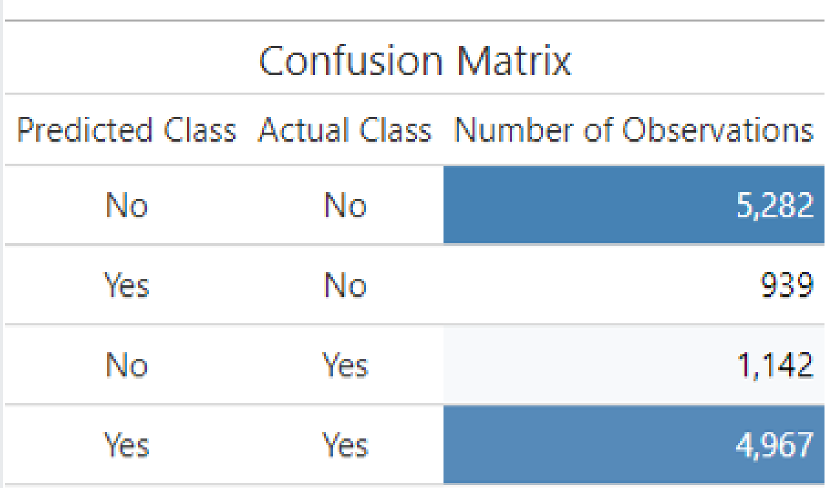
APPENDIX C: T-SNE VISUALIZATION

- Non-linear dimensionality reduction highlighting local data structure.
- Data points form dense clusters, but *Revenue vs. No Revenue* overlap significantly
- Confirms complexity and non-linearity of the classification problem.
- The data for this visualization was primarily intended for exploratory analysis and not for training the models.



APPENDIX D: CONFUSION MATRIX FOR ALL THREE MODELS

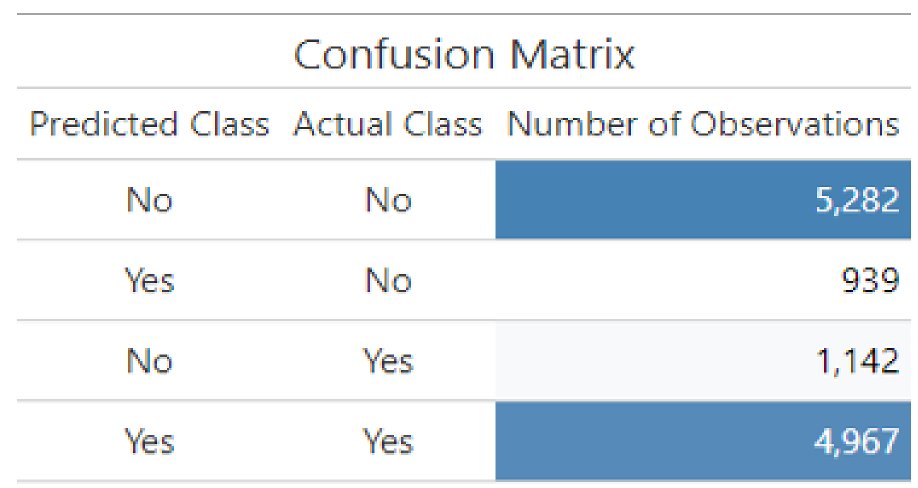
Logistic Regression



Confusion Matrix

Predicted Class	Actual Class	Number of Observations
No	No	5,282
Yes	No	939
No	Yes	1,142
Yes	Yes	4,967

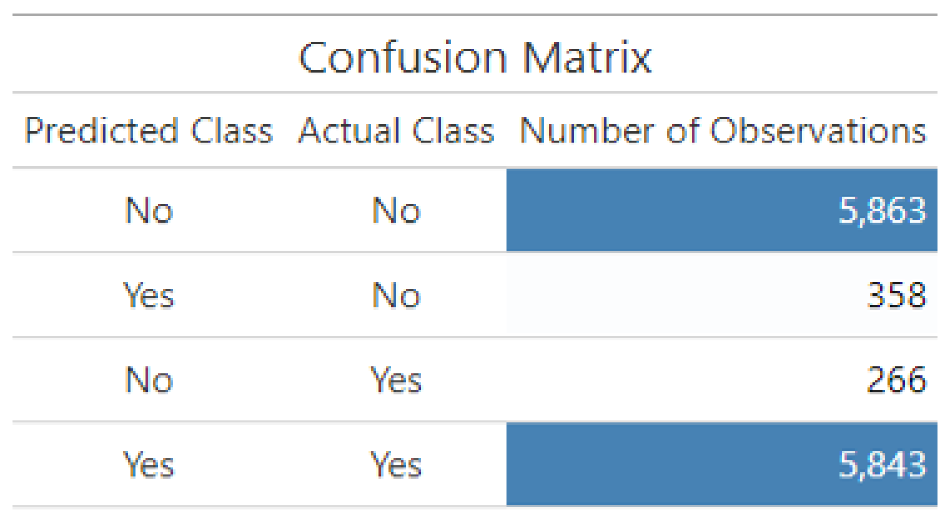
Random Forest



Confusion Matrix

Predicted Class	Actual Class	Number of Observations
No	No	5,282
Yes	No	939
No	Yes	1,142
Yes	Yes	4,967

Xgboost

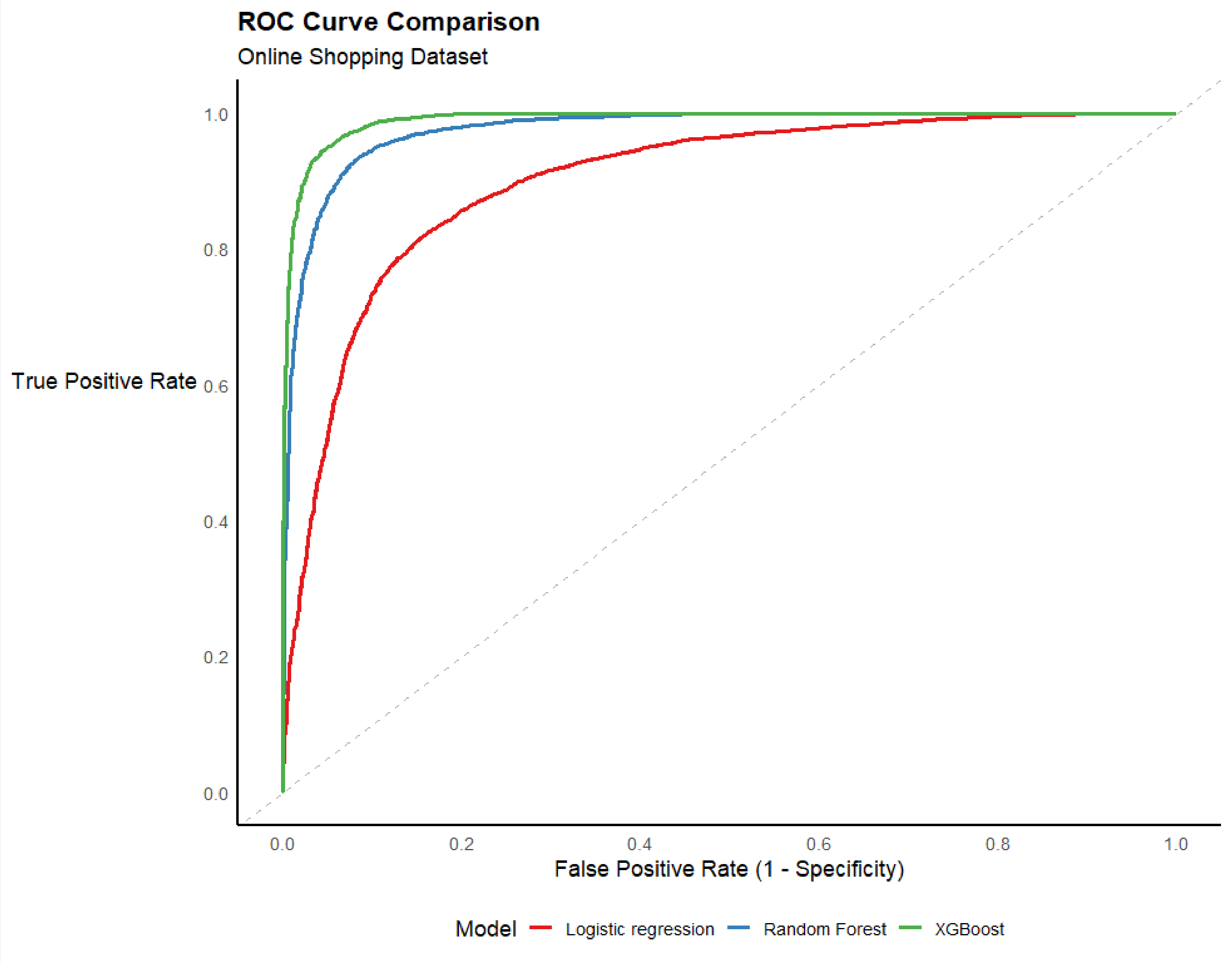


Confusion Matrix

Predicted Class	Actual Class	Number of Observations
No	No	5,863
Yes	No	358
No	Yes	266
Yes	Yes	5,843

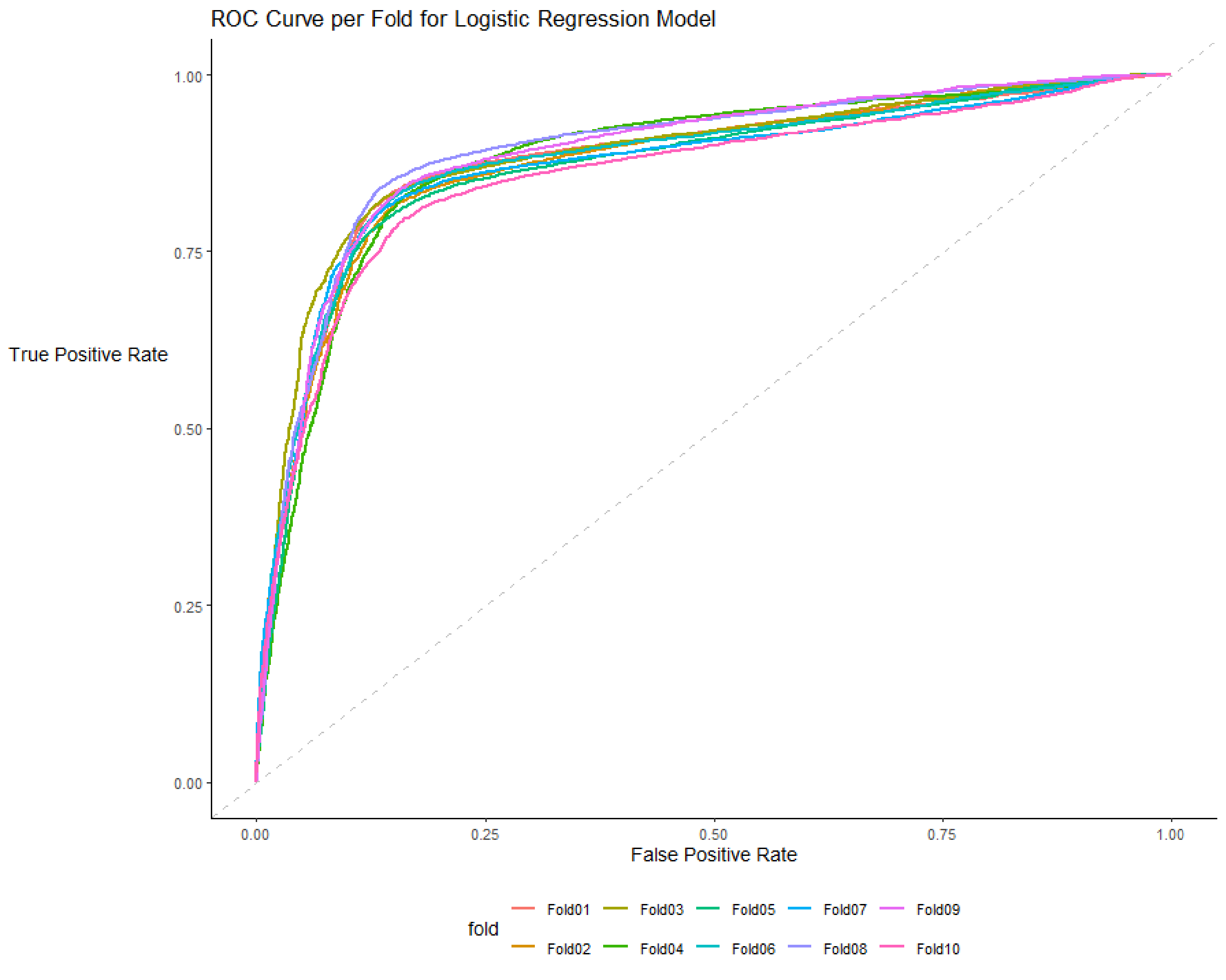
APPENDIX E: RECEIVER OPERATING CHARACTERISTIC (ROC CURVE)FOR ALL THREE MODELS

- The ROC Curve shows the true positive rate (recall/sensitivity) against the false positive rate at various threshold settings.
- The visualization shows that the XGBoost model performed the best at distinguishing between positive and negative classes, followed by random forest and lastly logistic regression.



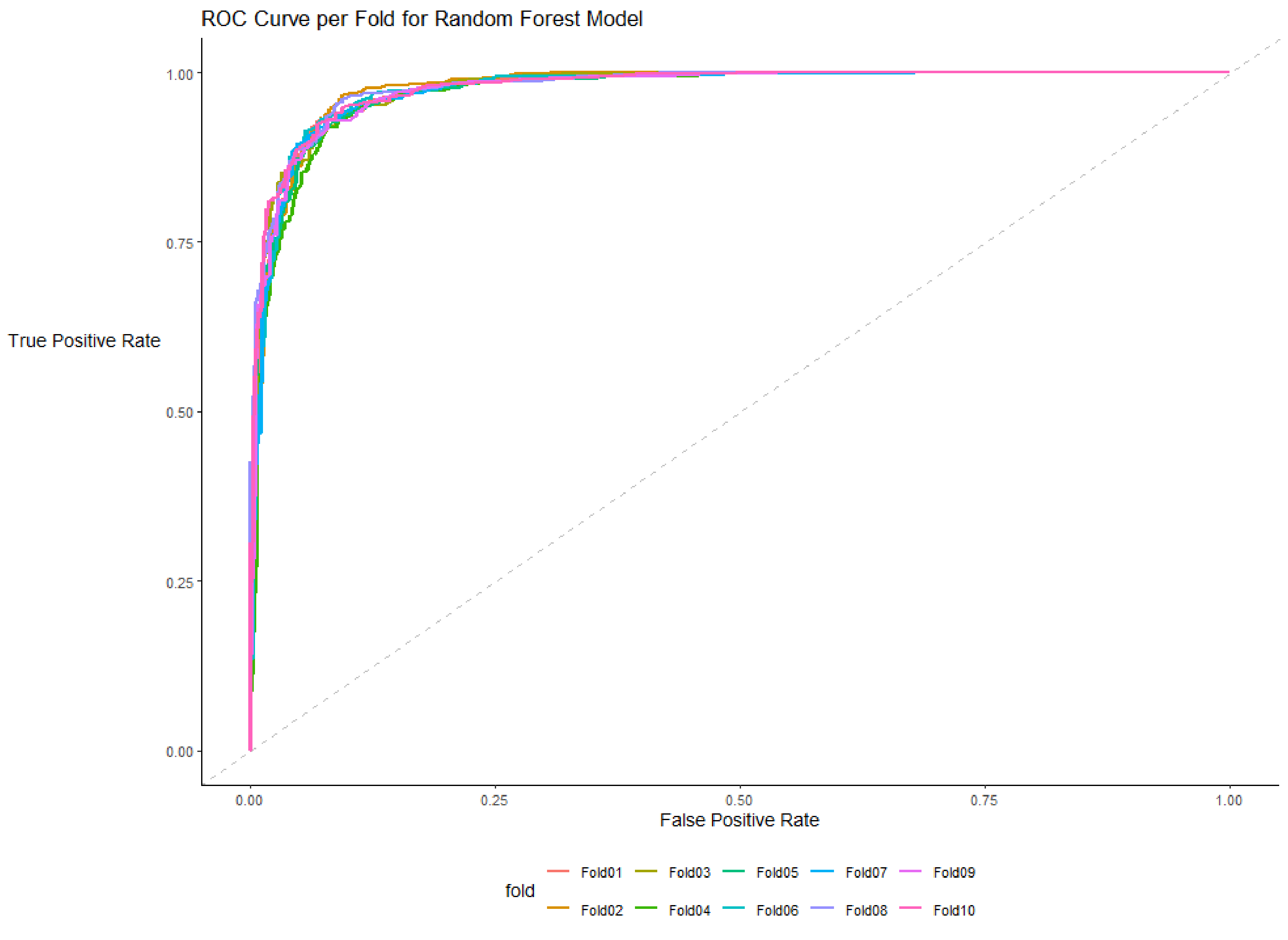
APPENDIX F:LOGISTIC REGRESSION MODEL ROC CURVE PER FOLD

- The ROC curve per fold visualises the performance of each of the 10 folds during cross-validation.
- Each line represents a fold’s true positive rate (recall/sensitivity) against the false-positive rate.
- The graph confirms that all folds demonstrate relatively consistent performance, suggesting that model generalises well across all subsets of the data.
- The area under the curve values across the folds are tightly grouped , indicating low variance and good model stability.



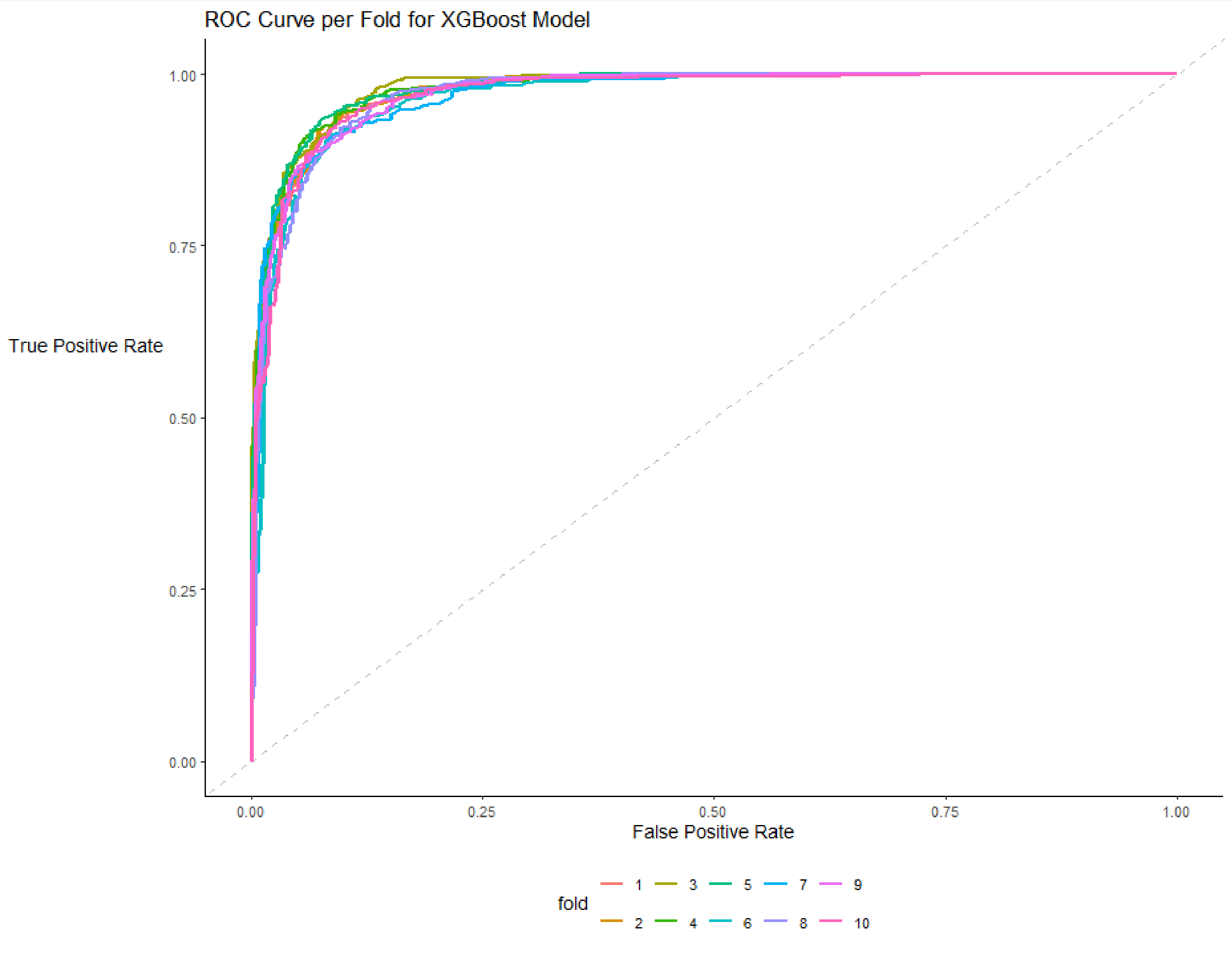
APPENDIX G:RANDOM FOREST MODEL ROC CURVE PER FOLD

- The ROC curve per fold visualises the performance of each of the 10 folds during cross-validation.
- Each line represents a fold’s true positive rate (recall/sensitivity) against the false-positive rate.
- The graph confirms that all folds demonstrate relatively consistent performance, suggesting that model generalises well across all subsets of the data.
- The area under the curve values across the folds are tightly grouped , indicating low variance and good model stability.



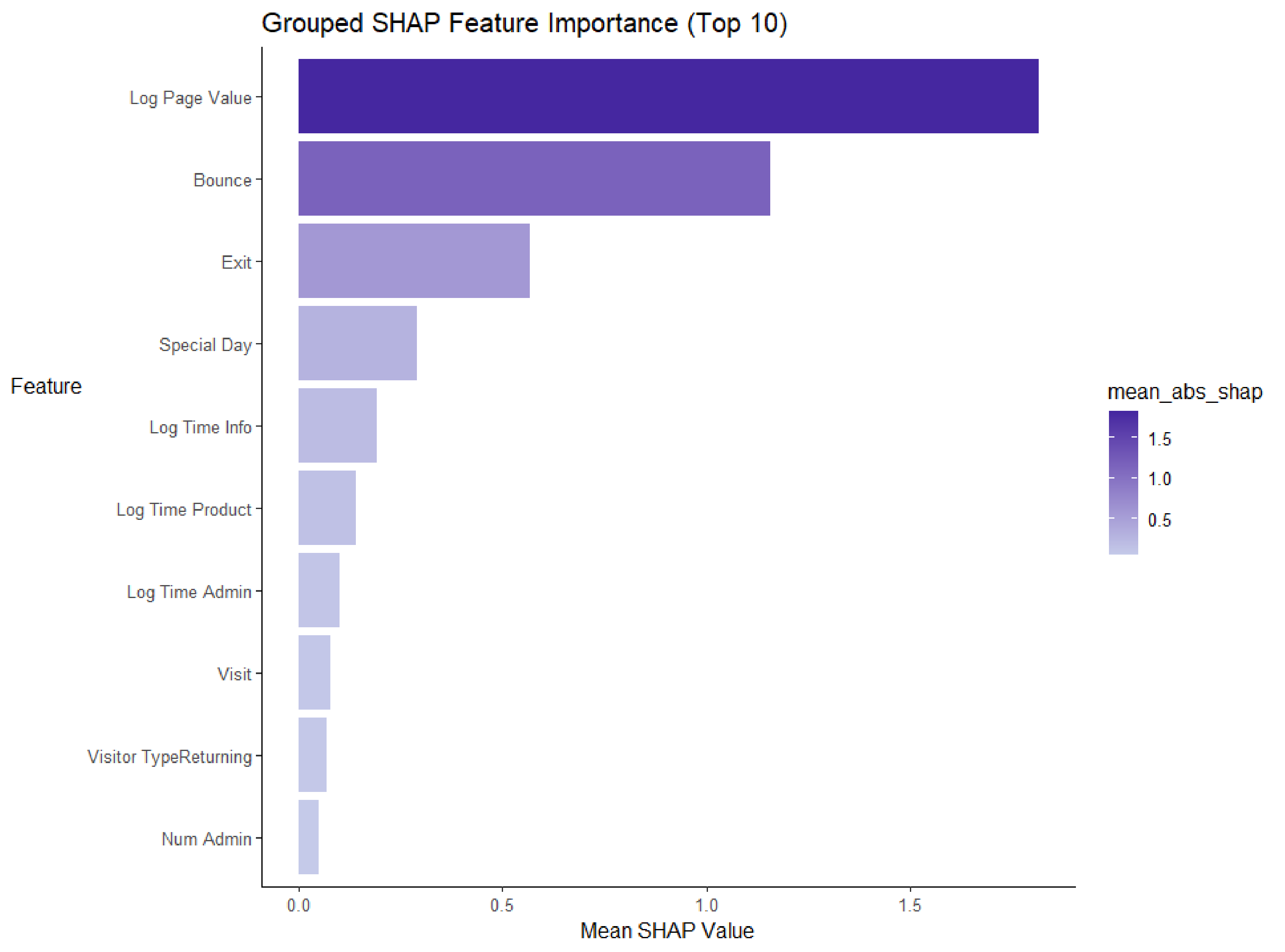
APPENDIX H:XGBOOST MODEL ROC CURVE PER FOLD

- The ROC curve per fold visualises the performance of each of the 10 folds during cross-validation.
- Each line represents a fold's true positive rate (recall/sensitivity) against the false-positive rate.
- The graph confirms that all folds demonstrate relatively consistent performance, suggesting that model generalises well across all subsets of the data.
- The area under the curve values across the folds are tightly grouped , indicating low variance and good model stability.



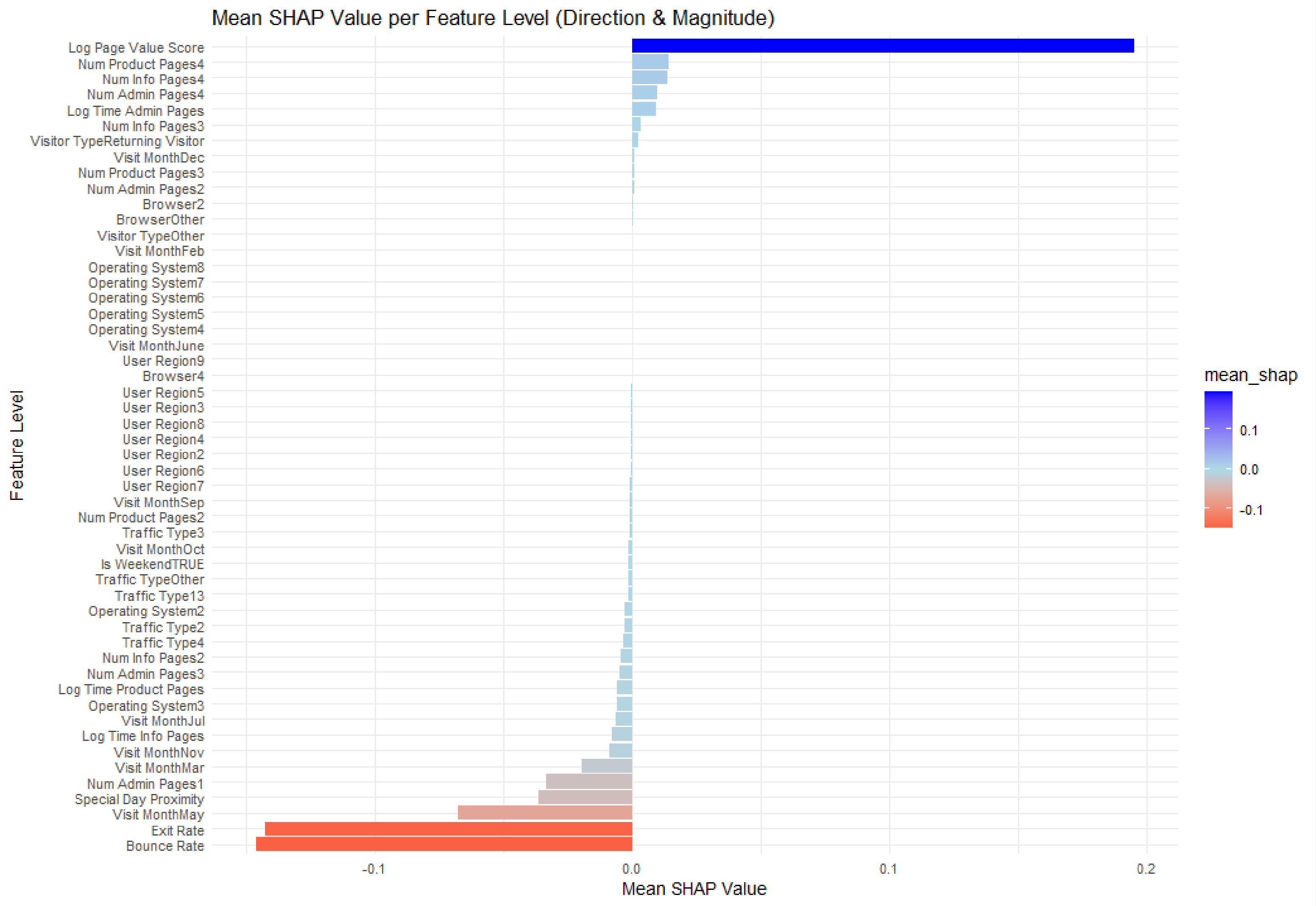
APPENDIX I:GROUPED SHAP FEATURE IMPORTANCE PLOT (TOP 10)

- Visualisation shows the aggregate mean absolute SHAP values from the output of the XGBoost model.
- It shows that page value (log) is the most important value followed by bounce rate, exit rate and special day.
- Notably, this is a global measure and doesn’t capture local or level-specific effects.

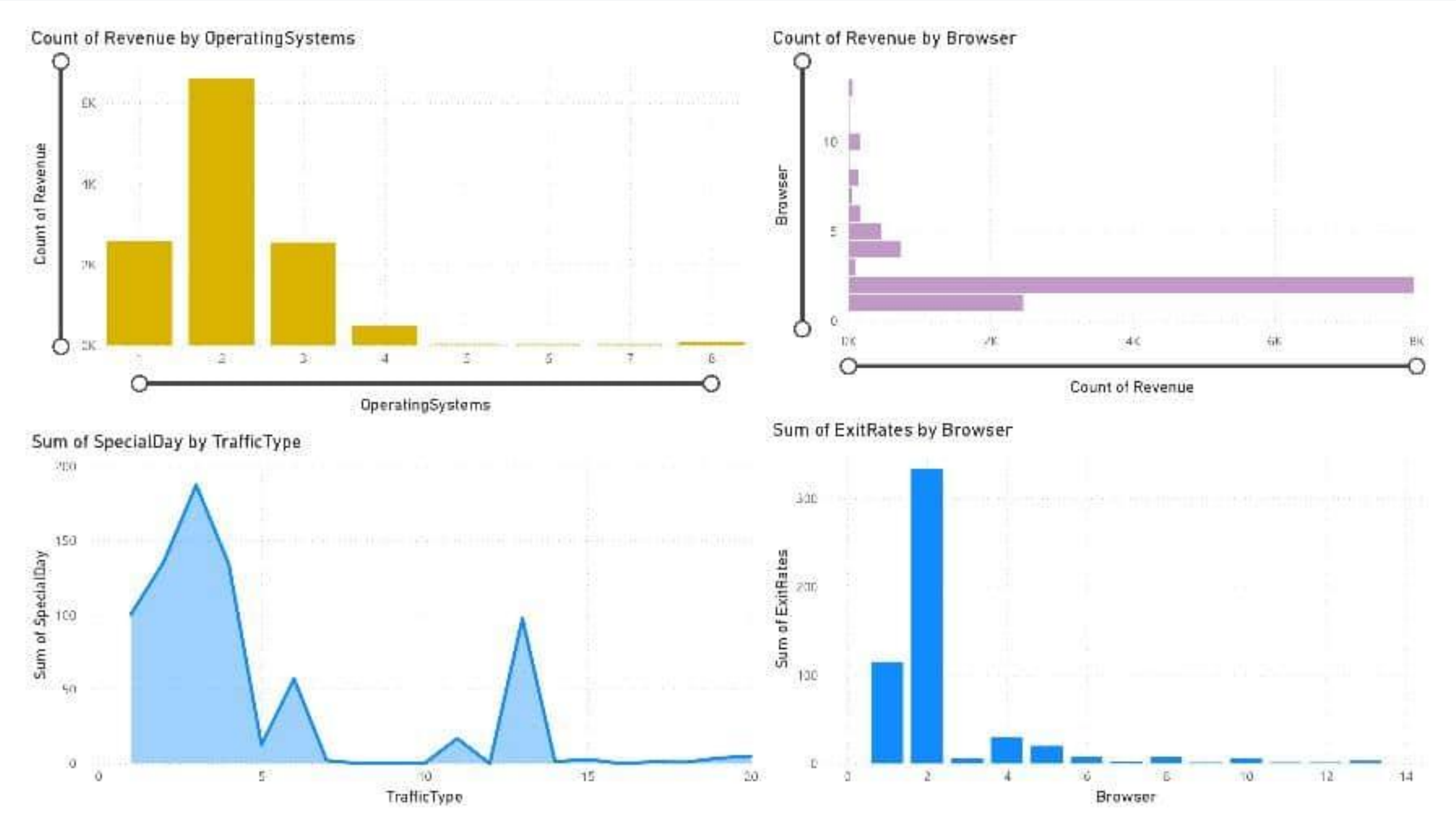


APPENDIX J:SHAP DEPENDENCE PLOT

- Shows the relationship between features and the target variable (Revenue) for the XGBoost model.
- The visualisation shows the magnitude and direction of how a feature impacts the model's prediction for individual data points.



APPENDIX K: EDA VISUALISATIONS



Count of Revenue by OperatingSystems

Count of Revenue by Browser

Sum of SpecialDay by TrafficType

Sum of ExitRates by Browser

Browser 2 shows high exit rates but also shows high revenue suggests there is overlapping

OS 2 and 3 have high revenue

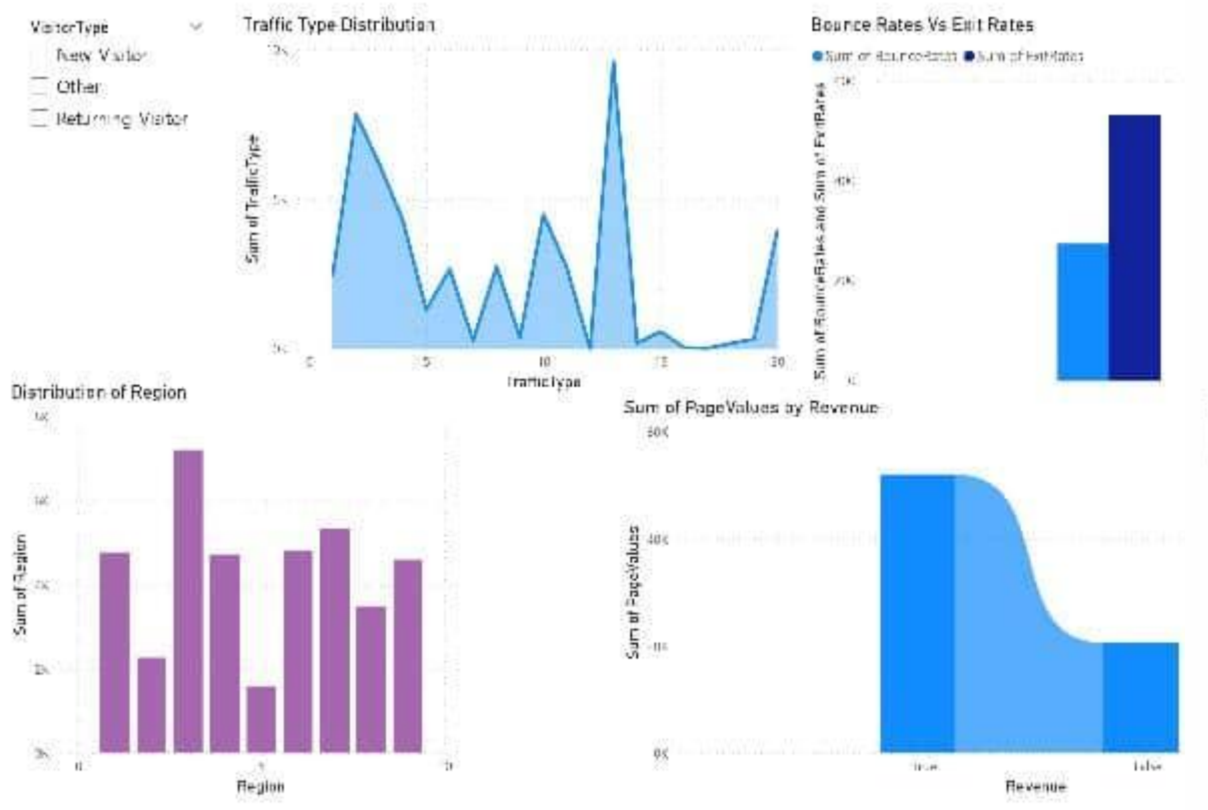
The plot shows traffic type on a special day

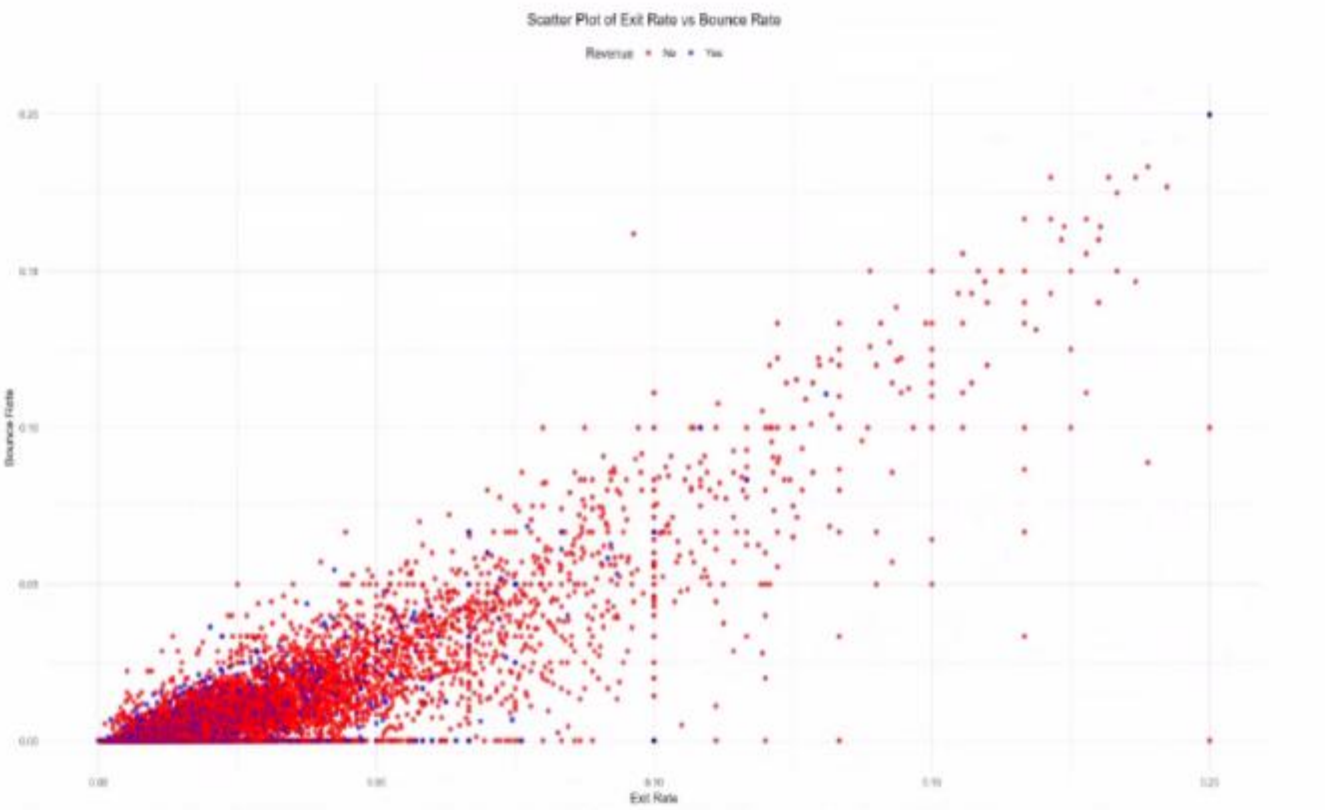
APPENDIX L: EDA VISUALISATIONS

- While many users exit the site, fewer leave after viewing just one page (which would count as a bounce).
- It helps identify whether users are disengaging early (bouncing) or after viewing more content (exiting)

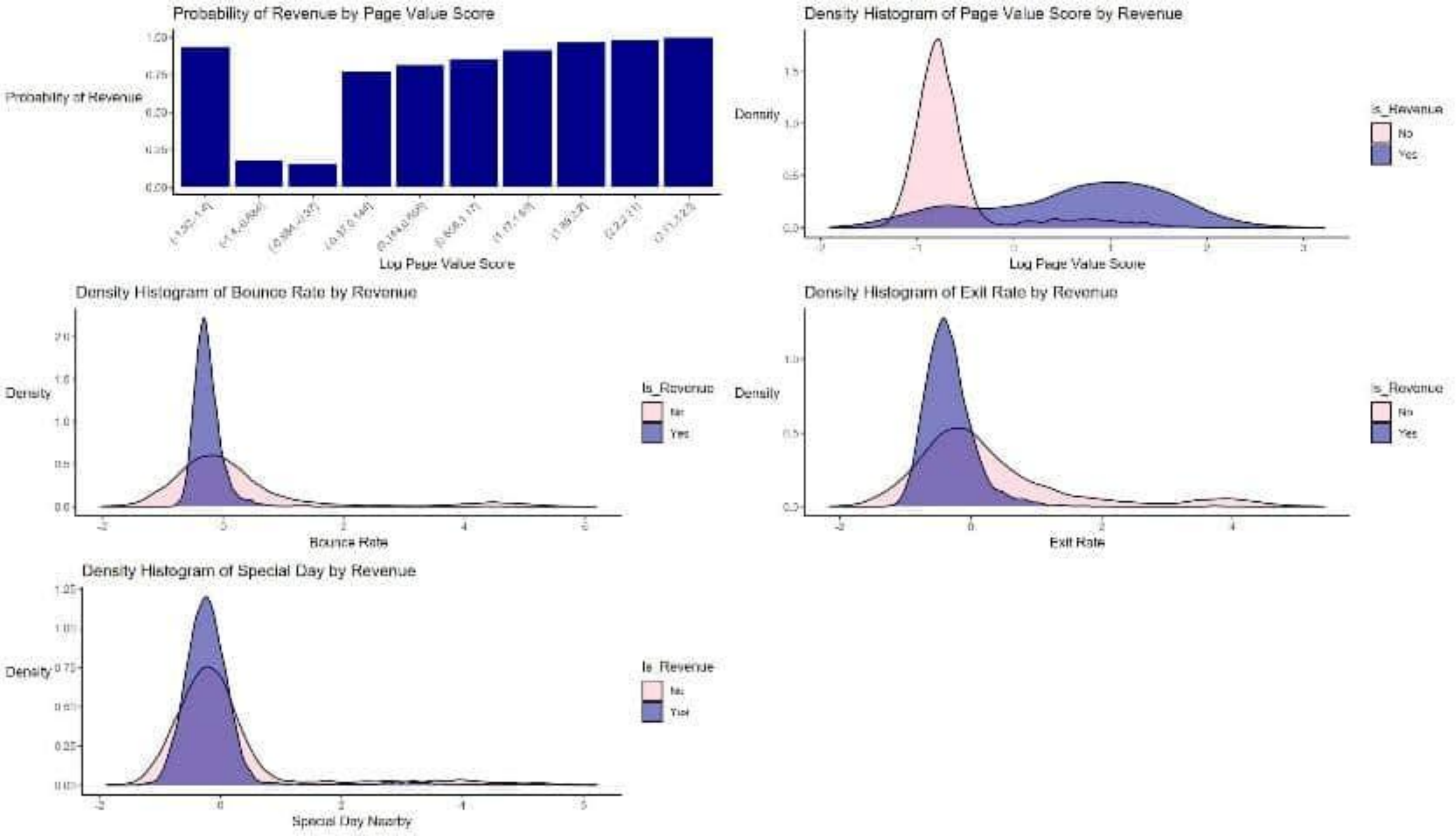
Traffic Type and Region differ greatly with regard to each visitor Page Values is highly Correlated and contributes more to revenue i.e has more True

- There is a visible **positive correlation**: as **Exit Rate increases, Bounce Rate also increases**.
- Most points cluster in the lower-left region, suggesting many sessions have both low exit and bounce rates.
- Revenue-generating sessions (blue) are more scattered and appear in areas with **lower bounce and exit rates**, hinting that users who stay longer (lower exit/bounce) are more likely to convert.





APPENDIX M: EDA VISUALISATIONS



1. **Probability of Revenue by Page Value Score**
- Binned log page value scores show rising revenue probability with higher scores.
- Users with high page value scores are significantly more likely to convert.
- Strong predictive feature for classification models.

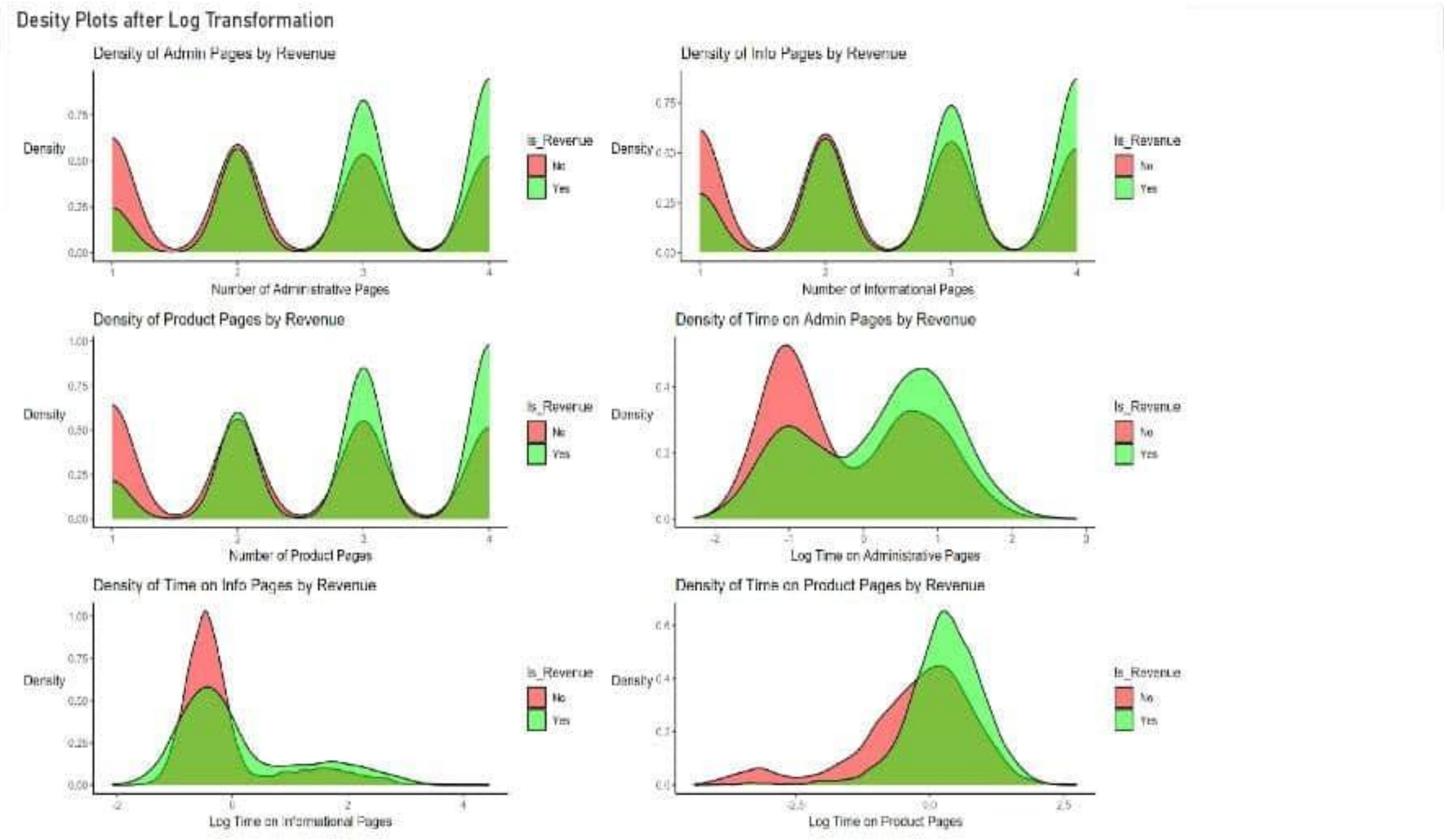
2. **Density of Page Value Score by Revenue**
- Revenue users (purple) are concentrated in the positive score range.
- Non-revenue users (pink) cluster around negative values.
- Reinforces that higher page value score correlates with revenue generation.

3. **Density of Bounce Rate by Revenue**
- Revenue users have a narrower, sharper peak close to 0.
- Suggests low bounce rate is common among converting users.

4. **Density of Exit Rate by Revenue**
- Exit rate is lower for revenue-generating sessions.
- Users making purchases are less likely to leave the site early.

5. **Density of Special Day Nearby by Revenue**
- Revenue and non-revenue groups are similar, but revenue users have slightly higher density close to special days.
- Minor but potential seasonal/holiday effect.

APPENDIX N: EDA VISUALISATIONS



Desity Plots after Log Transformation

1. **Density of Admin Pages by Revenue**

- Shows distribution of log-transformed number of administrative pages.
- Users who generated revenue (green) tend to view more admin pages than those who didn't (red).
- Suggests administrative interactions may correlate with conversions.

2. **Density of Info Pages by Revenue**

- Both groups (revenue/no revenue) show similar peaks, but revenue users lean slightly towards higher log counts.
- Indicates some interest in information pages might aid conversion.

3. **Density of Product Pages by Revenue**

- Clear separation: revenue users (green) tend to view more product pages.
- Strong indicator that product page interaction is tied to revenue generation.

4. **Density of Time on Admin Pages by Revenue**

- Revenue group has a more spread-out time distribution, peaking higher than the non-revenue group.
- Indicates longer or repeated admin interactions may be tied to purchases.

5. **Density of Time on Info Pages by Revenue**

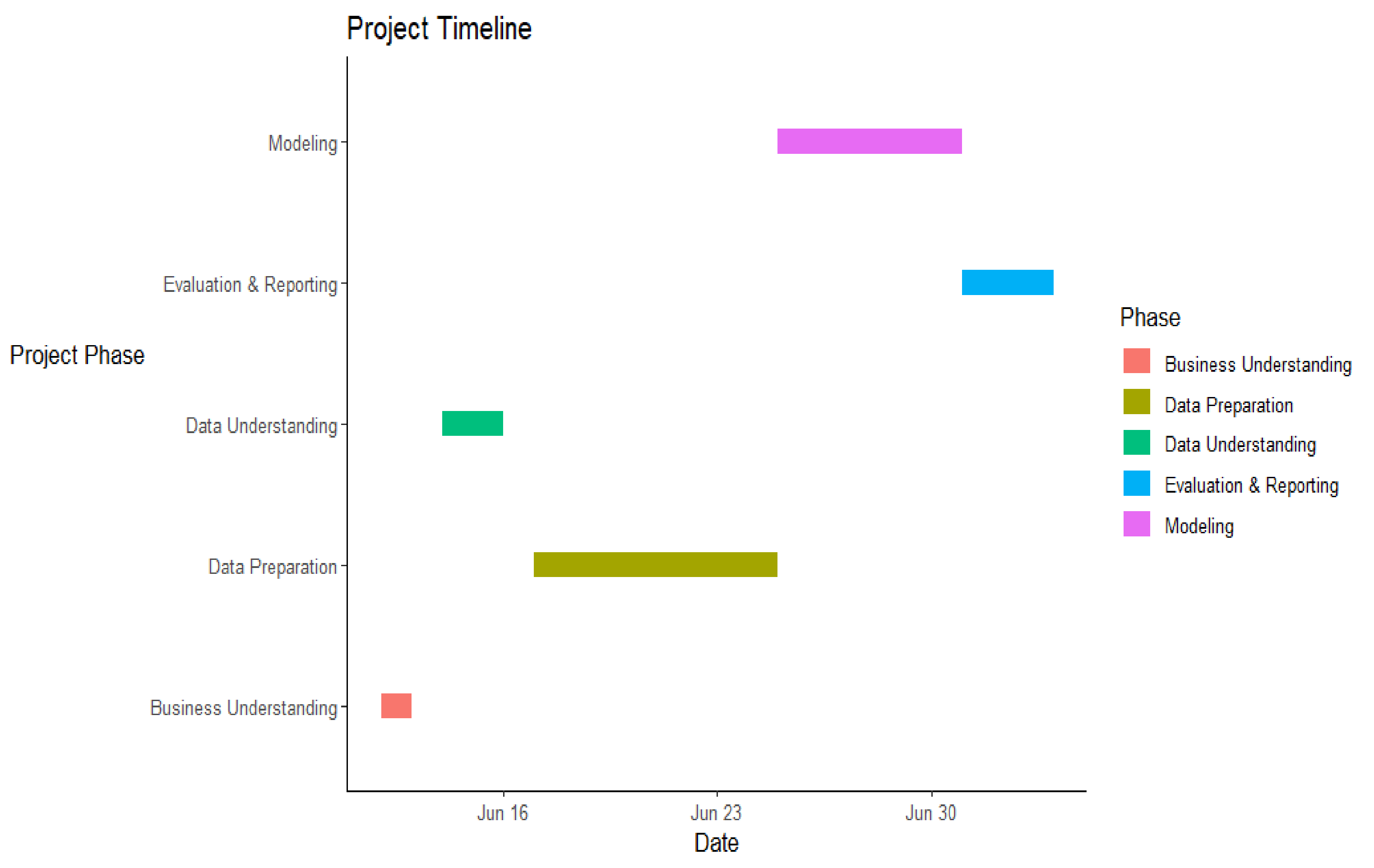
- Users who didn't generate revenue tend to have a higher density at lower log times.
- Revenue group shows longer time spent on info pages, though less sharply peaked.

6. **Density of Time on Product Pages by Revenue**

- Revenue group has a more right-skewed distribution.
- Indicates that longer time on product pages is positively associated with revenue.

APPENDIX O: GANTT CHART

- Shows the main activities completed by following the CRISP DM framework.
- Task durations were decided based on typical project workflows.
- The data preparation, modelling and evaluation & reporting phases include +1 contingency days whereas the business understanding, and data understanding phases contain 0 contingency days.


Project Timeline
Project Phase
Modeling
Evaluation & Reporting
Data Understanding
Data Preparation
Business Understanding
Jun 16
Jun 23
Jun 30
Date
Phase
Business Understanding
Data Preparation
Data Understanding
Evaluation & Reporting
Modeling