# 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) |

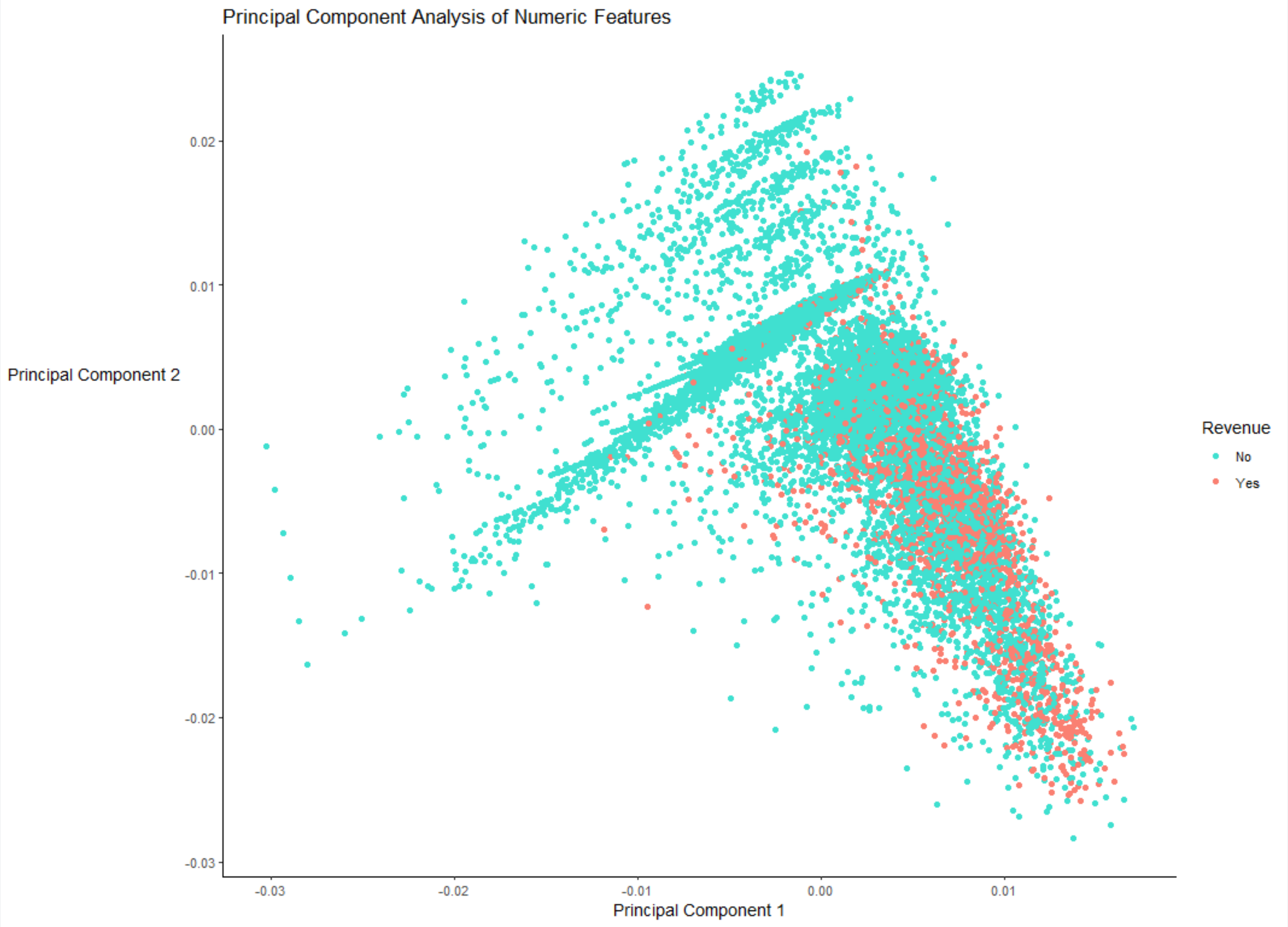
# APPENDIX A: VARIABLE INFLATION FACTOR (VIF) RESULTS

- 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 |

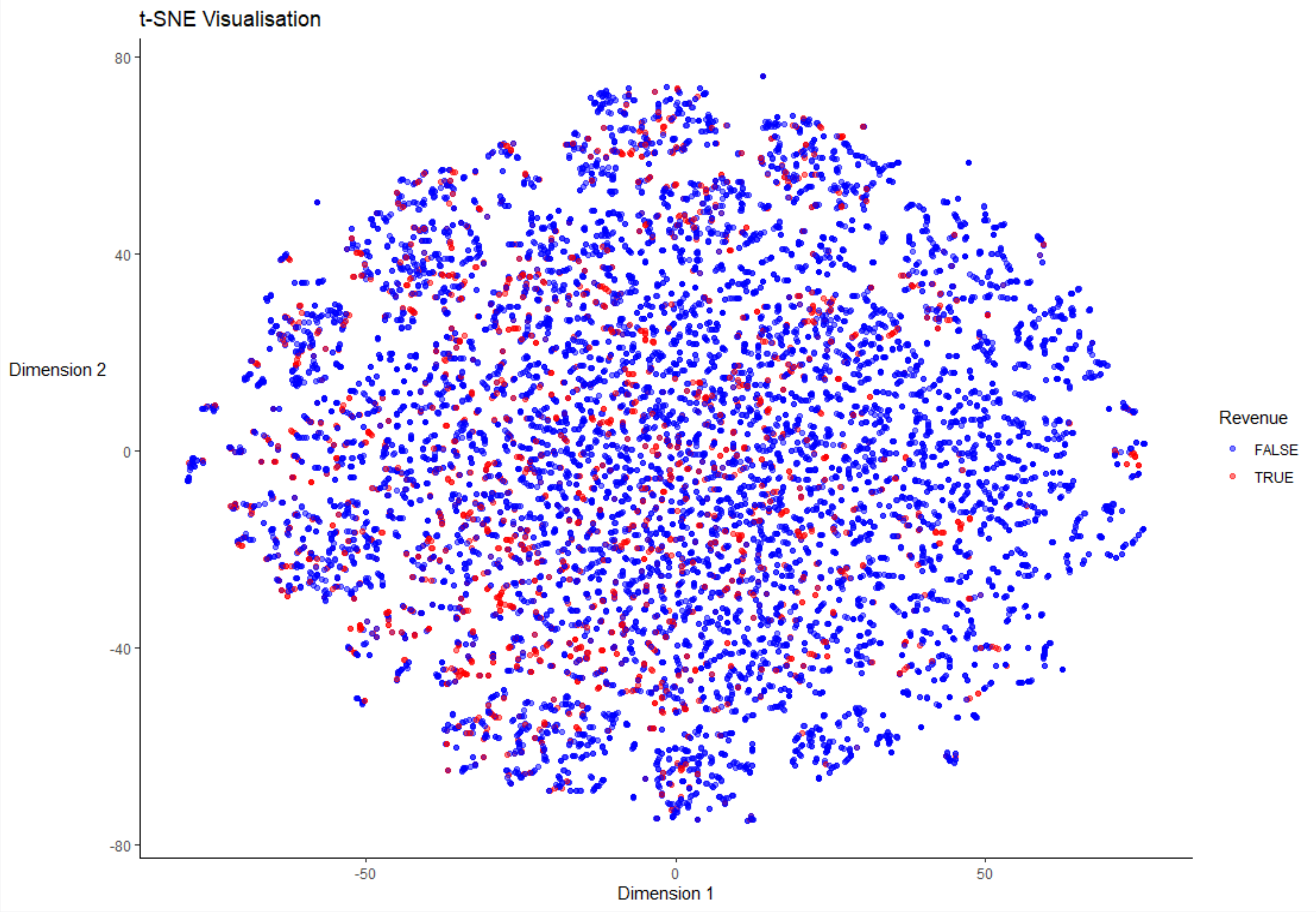
# APPENDIX B:PRINCIPAL COMPONENT ANALYSIS (2D PLOT)

- 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

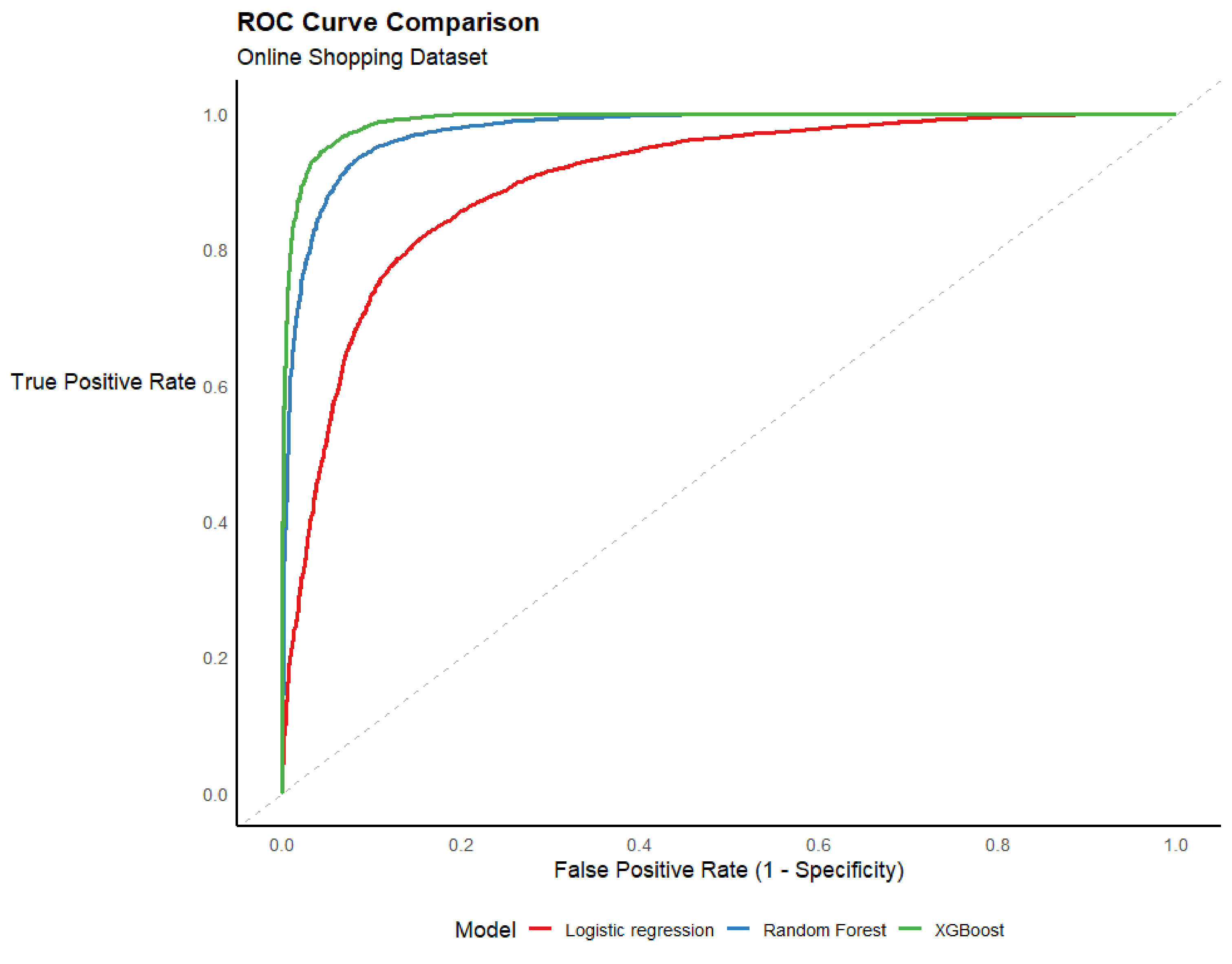
| Confusion Matrix | | |
|---|---|---|
| Predicted Class | Actual Class | Number of Observations |
| No | No | 5,282 |
| Yes | No | 939 |
| No | Yes | 1,142 |
| Yes | Yes | 4,967 |

## Xgboost

| Confusion Matrix | | |
|---|---|---|
| Predicted Class | Actual Class | Number of Observations |
| No | No | 5,863 |
| Yes | No | 358 |
| No | Yes | 266 |
| Yes | Yes | 5,843 |

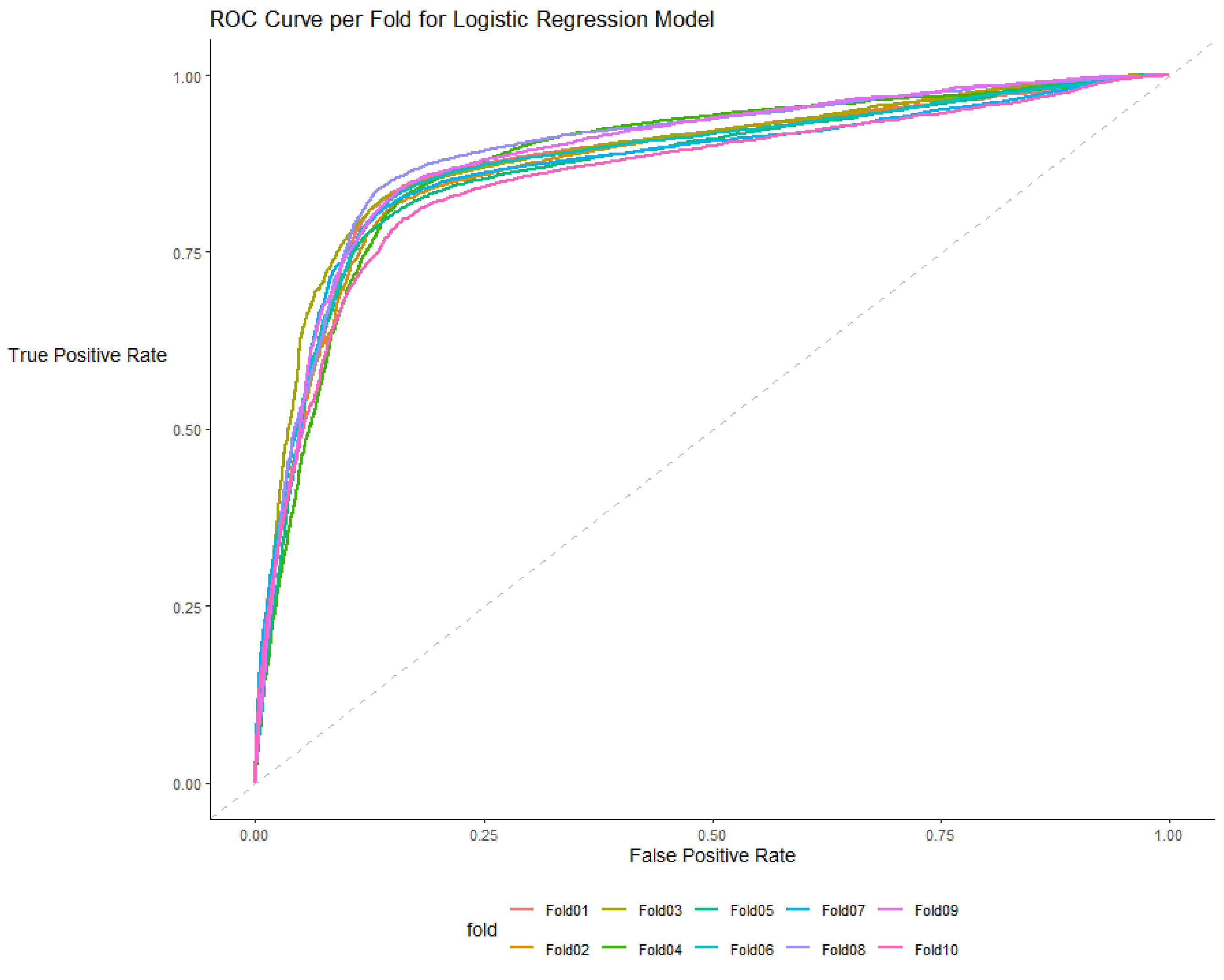
# APPENDIX E: RECEIVER OPERATING CHARACTERISTIC (ROC CURVE) FOR ALL THREE MODELS

- 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.



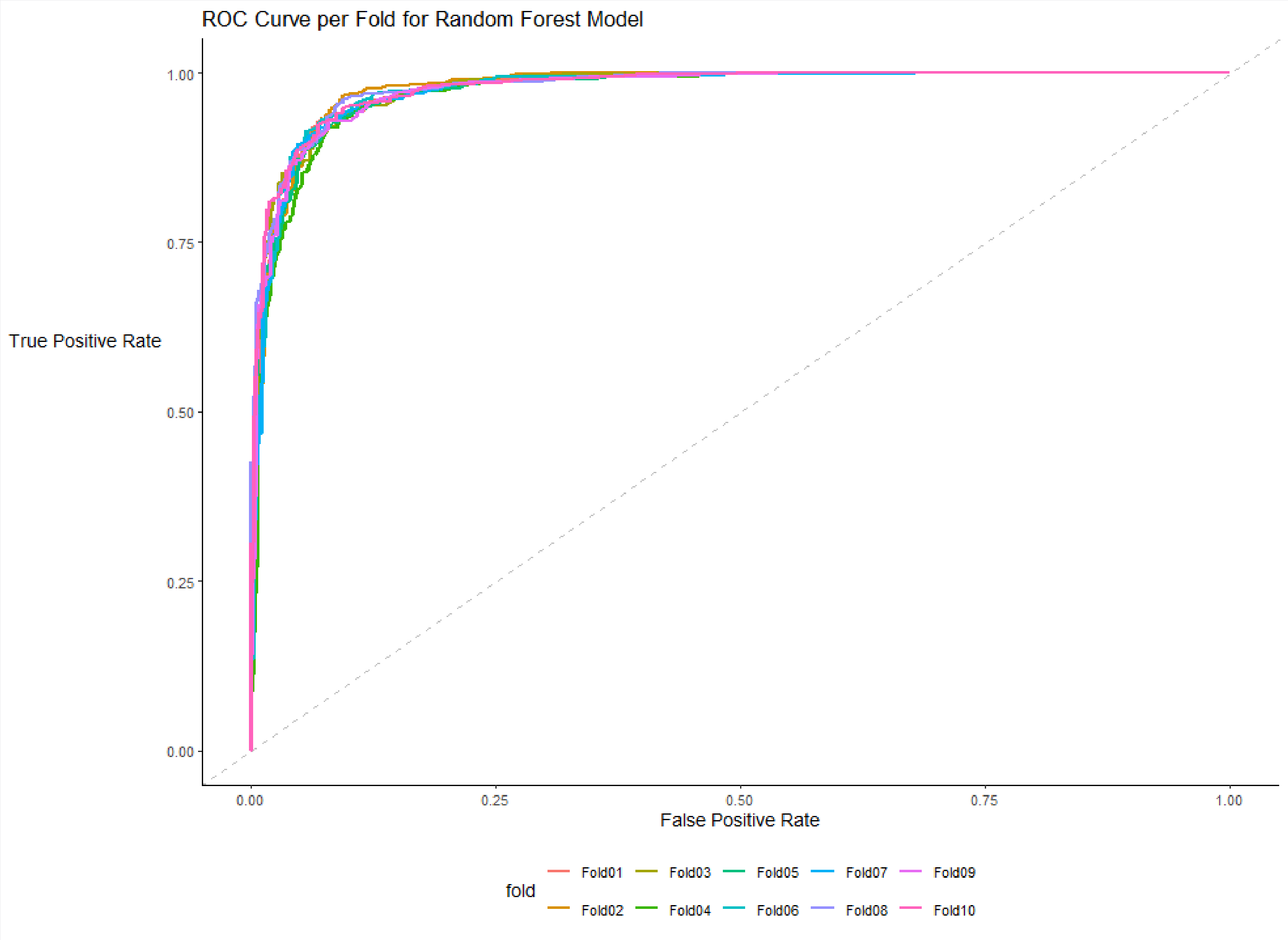
# APPENDIX F:LOGISTIC REGRESSION MODEL ROC CURVE PER FOLD

- 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.



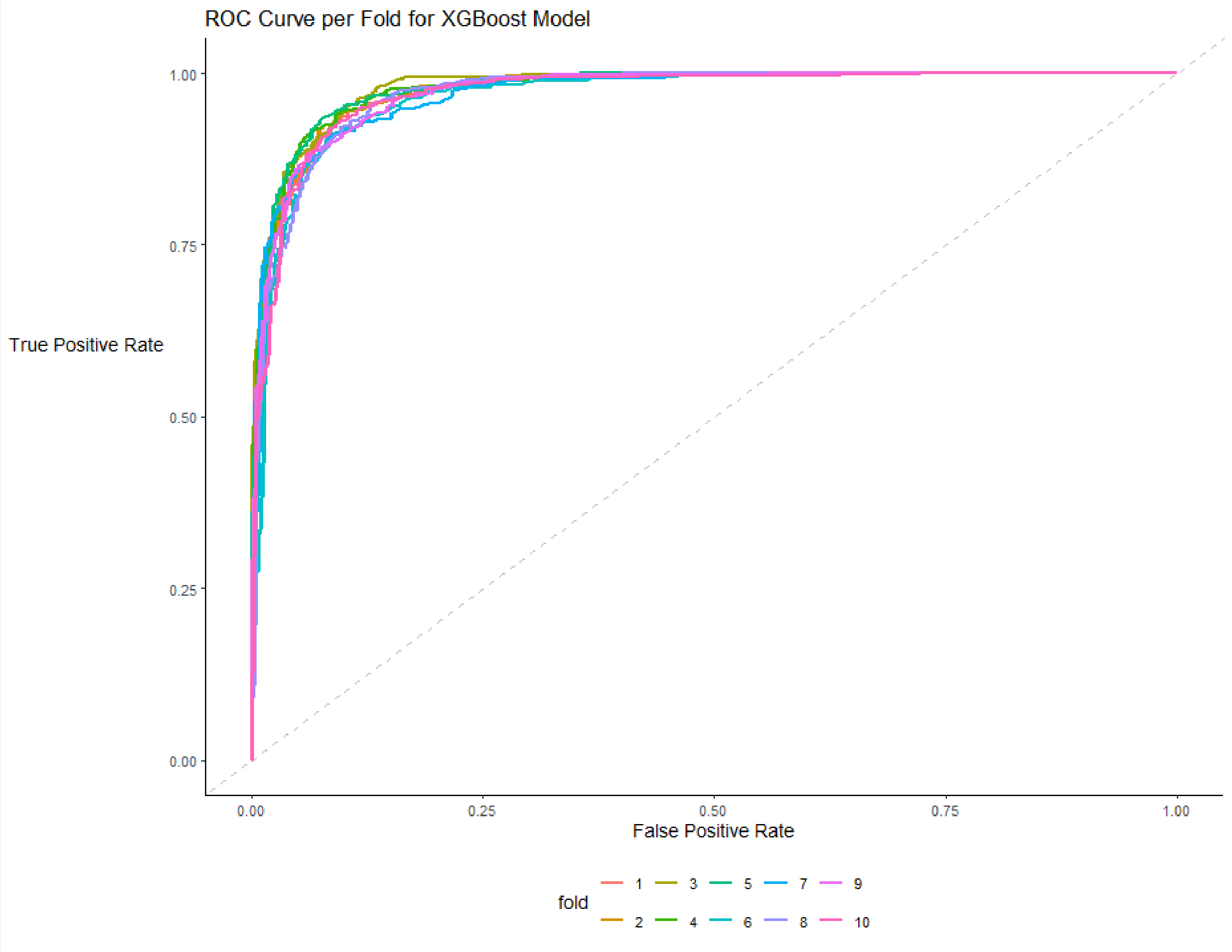
# APPENDIX G:RANDOM FOREST MODEL ROC CURVE PER FOLD

- 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.



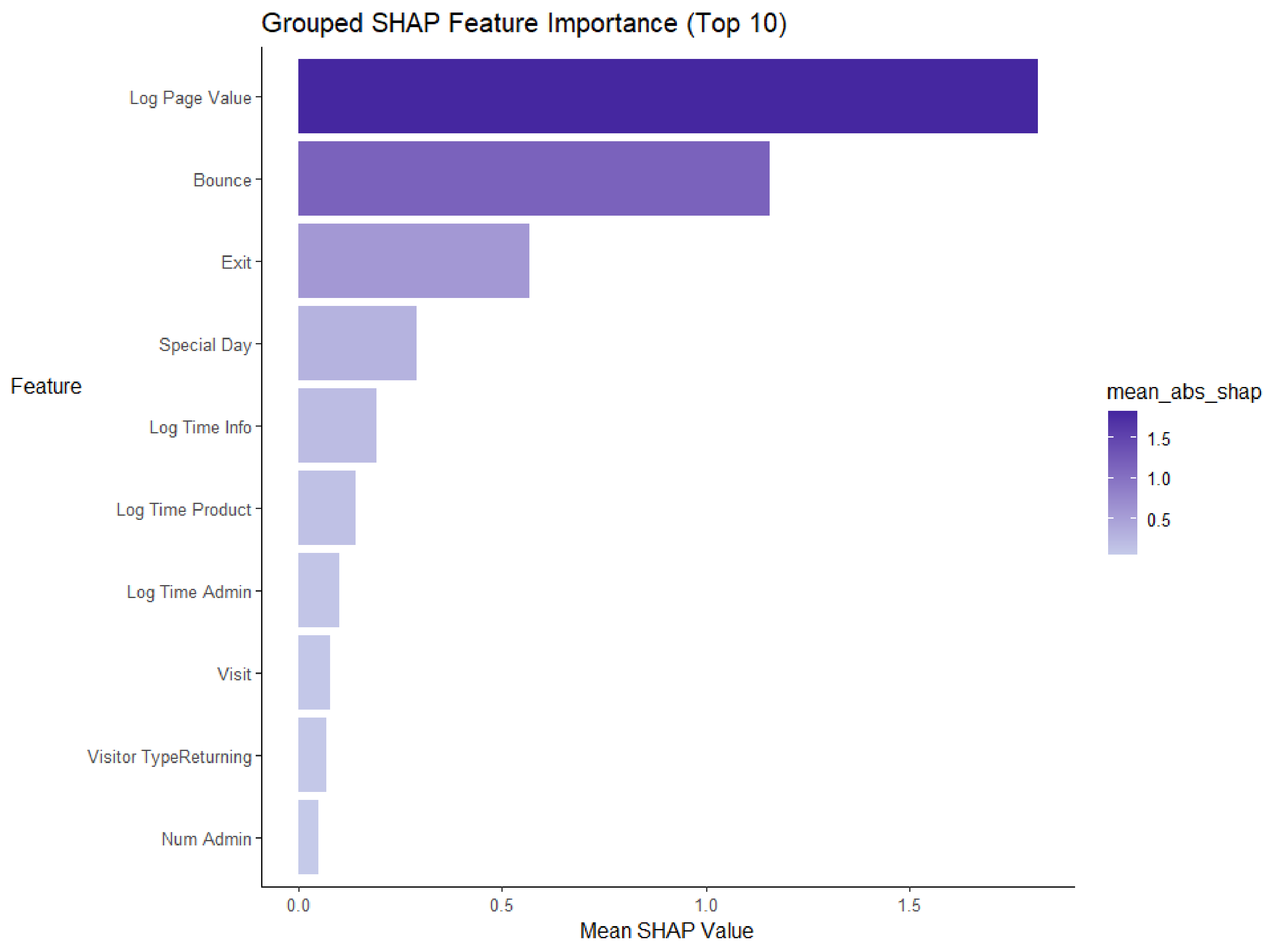
# APPENDIX H:XGBOOST MODEL ROC CURVE PER FOLD

- 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.



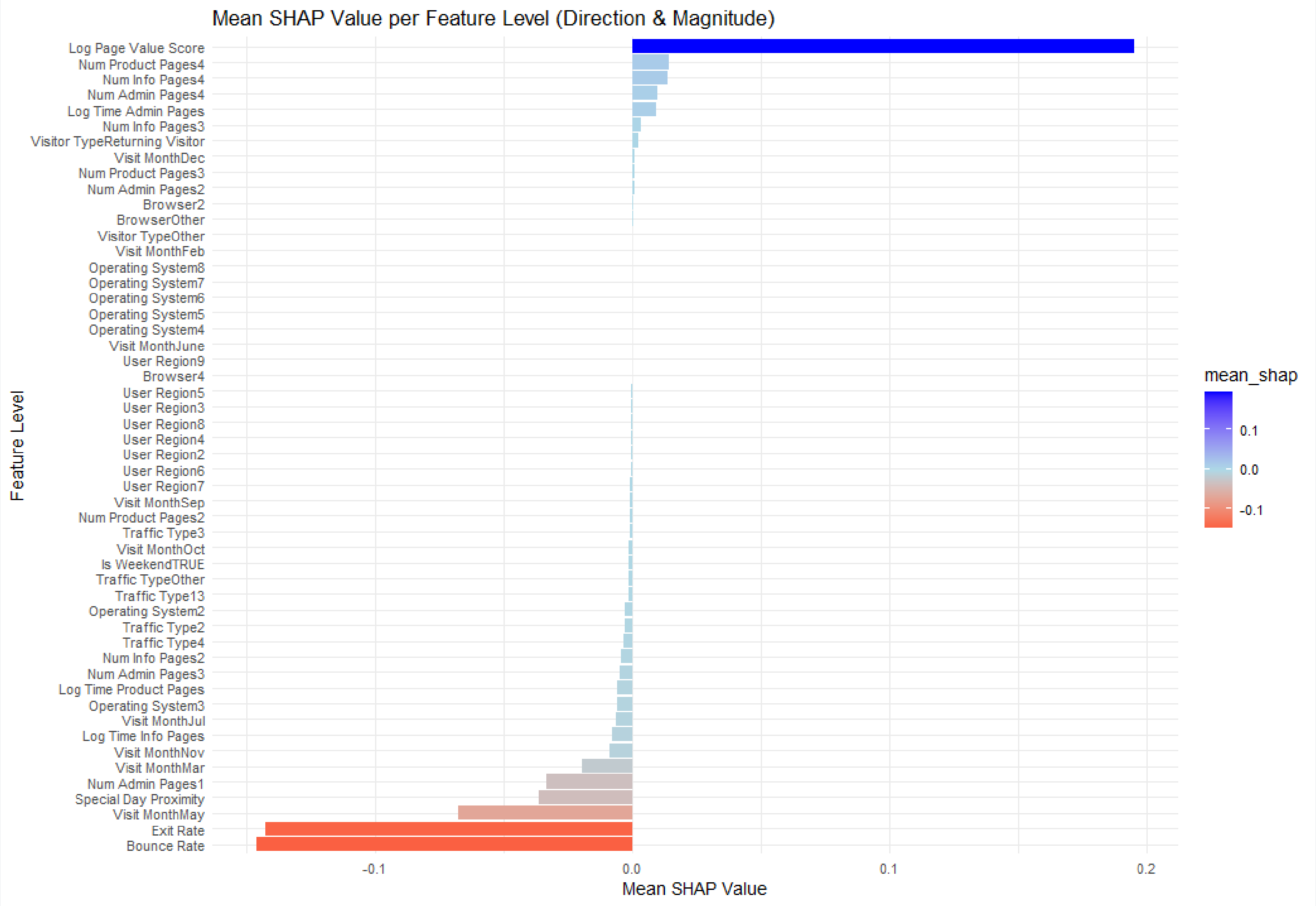
# APPENDIX I:GROUPED SHAP FEATURE IMPORTANCE PLOT (TOP 10)

- 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.

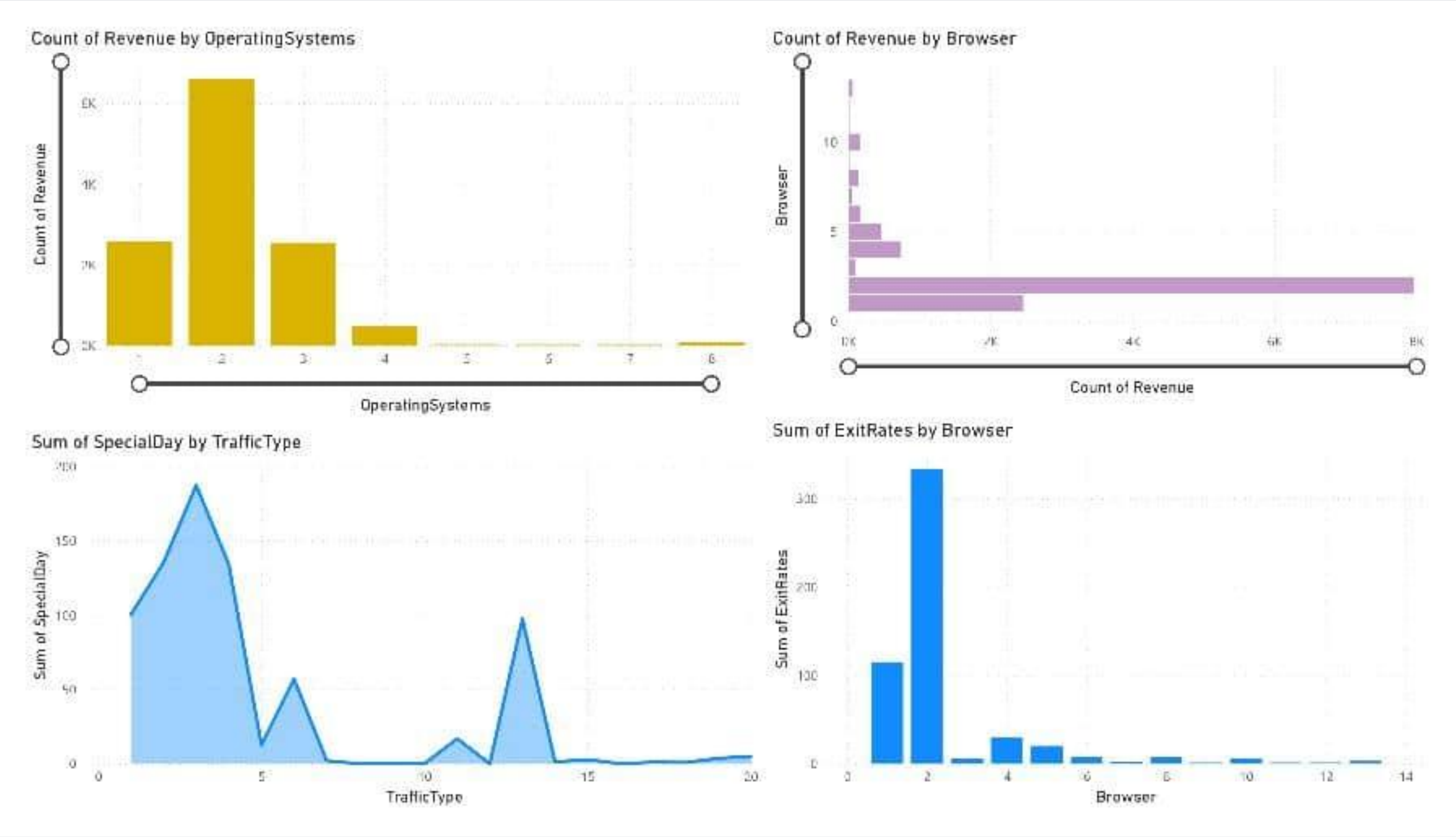


# APPENDIX J:SHAP DEPENDENCE PLOT

- 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




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

OS 2 and 3 have high revenue

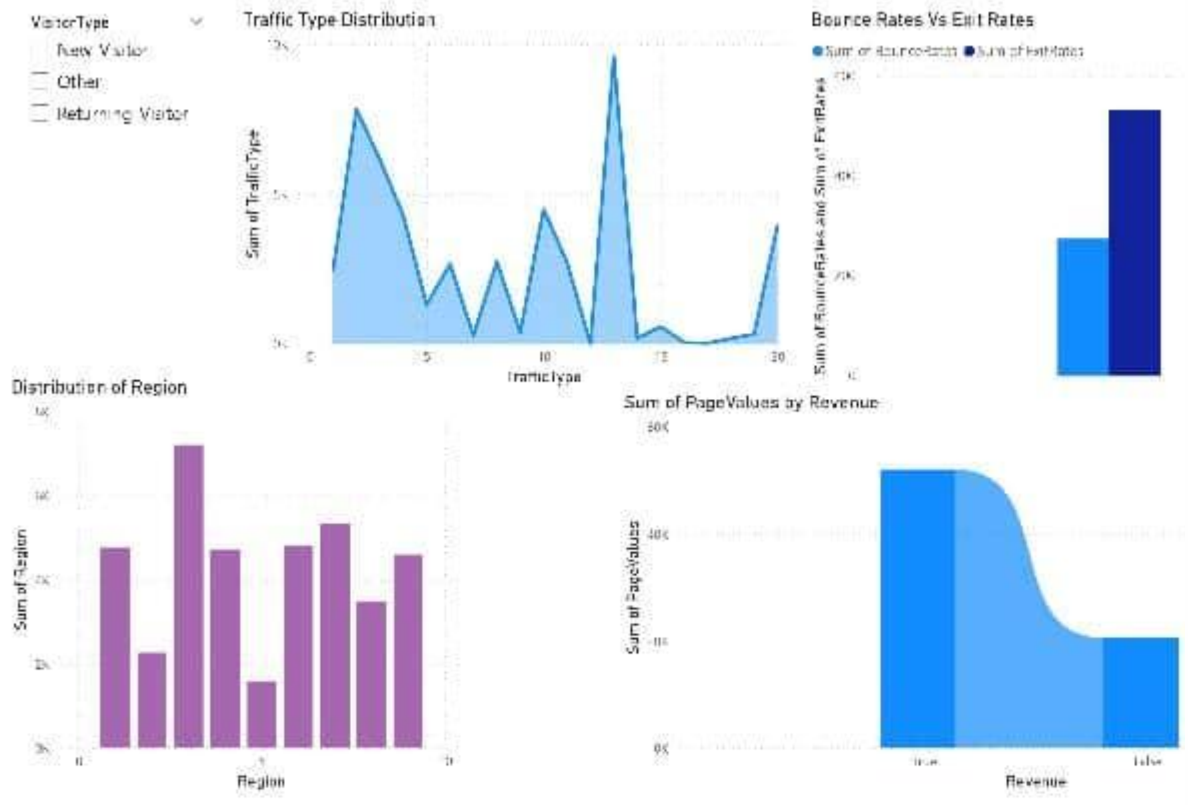
The plot shows traffic type on a special day

# APPENDIX L: EDA VISUALISATIONS

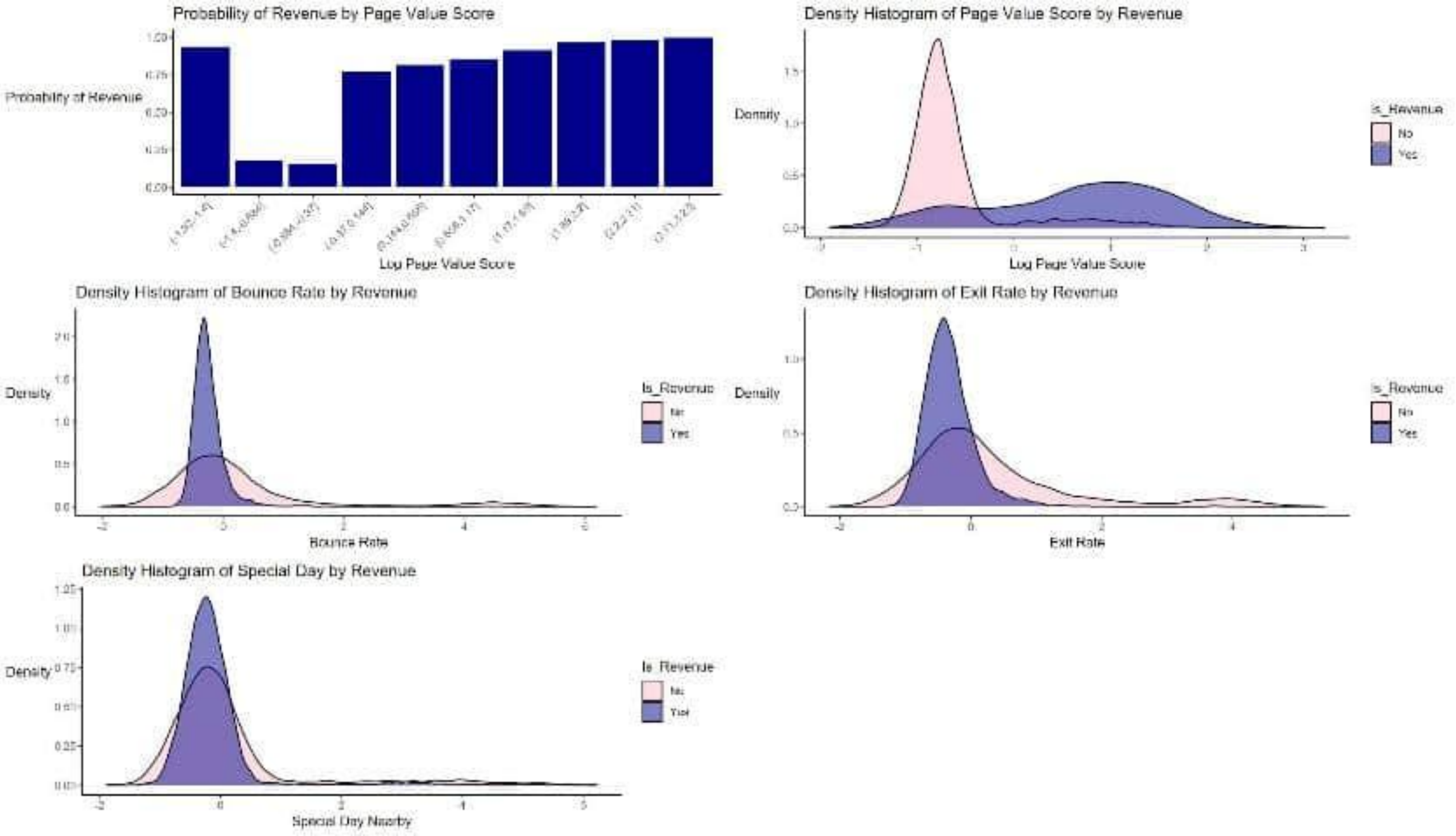
- While many users exit the site, fewer leave after viewing just one page (which would count as a bounce).
- It helps identify whether users are disengaging early (bouncing) or after viewing more content (exiting)

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.

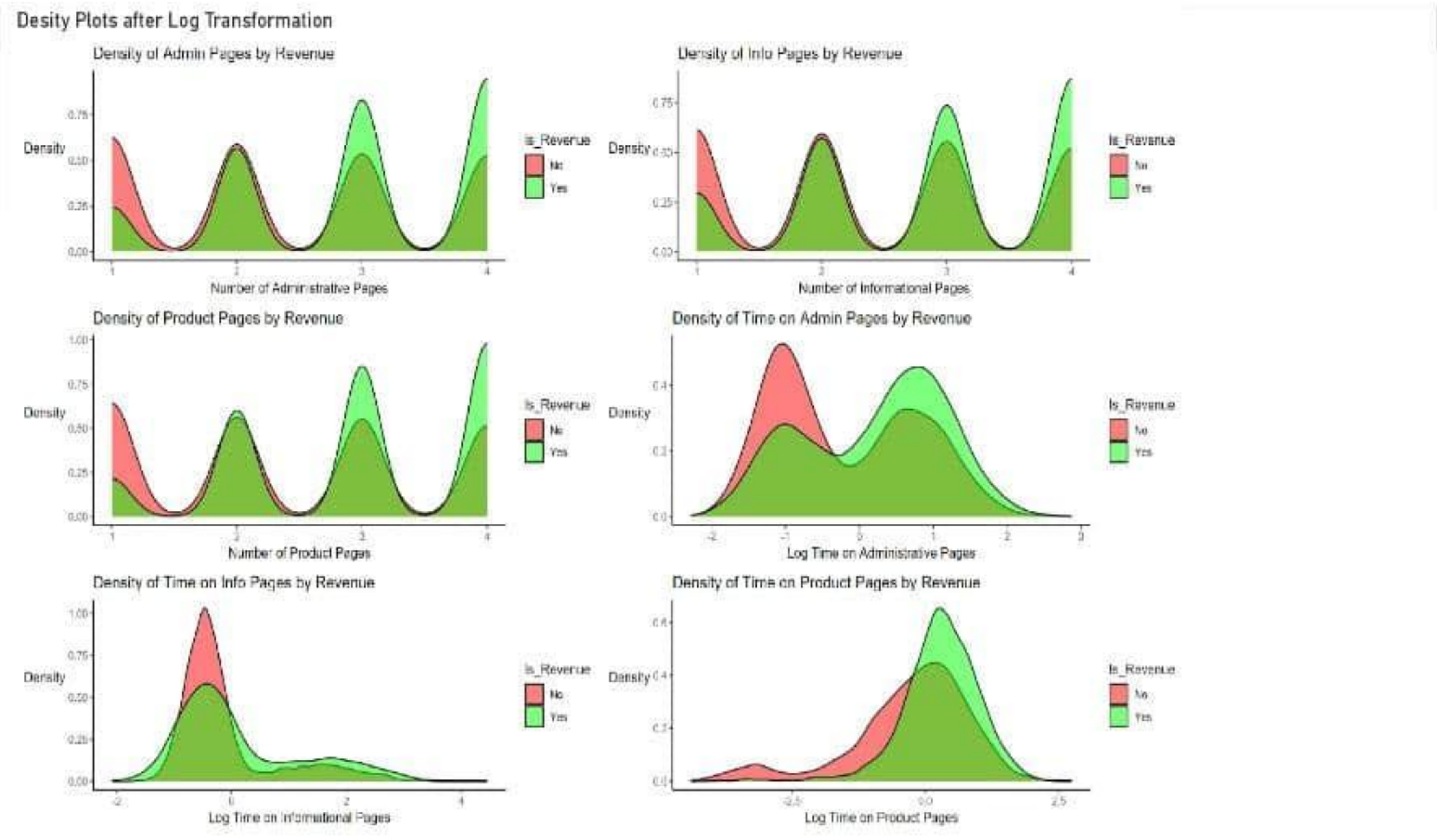
2. **Density of Page Value Score by Revenue**
- Revenue users (purple) are concentrated in the positive score range.
- Non-revenue users (pink) cluster around negative values.
- Reinforces that higher page value score correlates with revenue generation.

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




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.