Analysis Report: Online Shopping Dataset

Introduction

This report investigates the factors influencing purchase decisions during online shopping sessions, using the "OnlineShopping.csv" dataset. The main goal is to predict whether a user will complete a transaction (Revenue: TRUE) based on various session attributes, including page interactions, Google Analytics metrics, special day proximity, visitor details, and temporal factors.

Data Summary and Preprocessing

The dataset was first cleaned by removing duplicates and null values, resulting in a complete set for analysis. A key issue identified was class imbalance in the Revenue variable, with far fewer positive (conversion) cases. This was intentionally addressed during modelling using class weighting and SMOTE. User engagement varied across page types: administrative and informational pages saw minimal interaction (median visits near zero), though a small group showed intense engagement—suggesting research intent. In contrast, product-related pages had high activity (median of 18 visits, max over 700), reflecting strong purchase behaviour in a select segment.

Bounce and exit rates were generally low (medians ~0.3% and ~2.5%), though some pages reached 20%, indicating possible UX friction. The PageValues metric was highly right-skewed, with most sessions generating no revenue but a small number contributing significantly (max ≈ 362), highlighting the importance of identifying high-value users. The SpecialDay feature showed minimal variation, implying weak promotional impact.

Technical features showed that most users accessed the site via common platforms (Windows/macOS, Chrome/Firefox), with moderately diverse traffic sources. Preprocessing steps included log transformations to correct skew in duration and rate metrics and targeted outlier treatment: Z-score filtering for near-normal distributions (e.g., Administrative_Duration_log), IQR for moderately skewed ones (e.g., ExitRates_log), and percentile clipping for extreme cases like PageValues_log. One-hot encoding was applied to categorical features like Month to retain seasonal signals. These transformations normalised feature distributions and ensured model-ready data.

Exploratory Data Analysis

The exploratory data analysis (EDA) aimed to uncover the behavioural and contextual factors most strongly associated with user conversion. This was achieved using a multi-pronged approach that included binned variable analysis, funnel path exploration, seasonal trend inspection, and traffic source evaluation—each contributing unique insights into what drives revenue-generating sessions.

Binned Analysis: To manage the large proportion of zero-values in engagement metrics such as Administrative and Informational durations and PageValues, segmentation was applied through binning to better capture behavioural differences across user sessions. This approach enabled clearer comparisons between levels of user interaction. Visualisations (Appendices 9, 13, and 14) revealed that even minimal engagement contributed predictive value: users who briefly interacted with administrative pages converted at approximately 22%, double the rate (11%) of those with no engagement. Similarly, longer durations on informational pages were associated with a 24% conversion rate versus 15% for non-engaged users—highlighting that non-transactional content can still influence purchasing intent. The most pronounced effect appeared in the PageValues variable: users in the highest value category converted at 62.57%, compared to just 4.28% in the zero-value group. These findings reinforce PageValues as both a powerful predictor and a behavioural signal of purchase readiness.

Funnel Analysis: Funnel diagrams were constructed to visualise key behavioural transitions in the user journey. While nearly all users transitioned from informational to product pages (2,395 of 2,399), only 23.6% of these sessions resulted in a purchase, exposing a substantial drop-off late in the funnel. However, alternative flows like "Administrative → Product → Conversion" (1,385 conversions) and "High PageValue → Conversion" (1,322 conversions) exhibited significantly higher conversion success, suggesting that certain engagement patterns (particularly those tied to high-value content or deeper site navigation) are more effective at pushing users toward checkout. In contrast, time-based flows such as "Product on Special Day" yielded only 50 conversions, highlighting that calendar proximity alone doesn't guarantee action.

Seasonal Trends: Monthly trend analyses revealed distinct seasonality in conversion behaviour **(Appendices 1, 2, and 10)**. November exhibited the highest conversion rate (27.13%), likely linked to events like Black Friday, and was supported by elevated average PageValues (0.79 log-scaled) and the longest session durations. Despite a lower volume of sessions, October also showed strong performance (21.31%) due to even higher engagement quality—particularly an average PageValue of 0.92. Conversely, SpecialDay proximity—meant to reflect periods around major promotions—did not correlate positively with conversions. For instance, May and February had high SpecialDay scores but relatively poor conversion rates, suggesting potential misalignment between campaign timing and user readiness or interest.

Traffic Source Insights: Analysis of traffic types (Appendix 15) highlighted three channels—Types 7, 20, and 2—as particularly effective at driving high-value traffic. Traffic Type 7, although representing a small number of sessions, showed the highest average PageValue (1.46) and a 30% conversion rate, likely indicative of targeted campaigns or returning loyal users. Type 20 performed similarly, with strong session-level engagement and a 28% conversion rate. Traffic Type 2, while lower in individual session value (0.84), delivered large volumes and consistent conversion rates—making it valuable at scale. These results suggest strategic investment in these sources may yield significant return, especially when combined with tailored user experiences and retargeting.

This multi-layered EDA revealed that successful conversions are strongly influenced by the depth of user interaction, timing of sessions, and source of traffic. High PageValues, longer sessions, and specific traffic types emerged as consistent predictors of revenue. These insights lay a strong foundation for model training and provide business teams with actionable strategies for campaign targeting, content optimisation, and customer journey refinement.

Model Analysis and Insights

Logistic Regression:

Logistic regression was chosen for its interpretability, efficiency, and suitability for binary classification, modelling the log-odds of conversion as a linear function of input features. Assumptions of log-odds linearity, independent observations, and absence of multicollinearity were addressed—VIF analysis led to the removal of highly correlated variables (e.g., Total_Session_Duration), which stabilised the model without reducing performance (ROC-AUC remained 0.9157). Class imbalance was mitigated using class weighting, yielding strong initial metrics (ROC-AUC = 0.9157, recall = 81%) and ensuring that conversion sessions were not overlooked. Recursive Feature Elimination refined the model to 13 key predictors (e.g., PageValues_log, ExitRates_log, Month_Nov), maintaining AUC (0.9148) and recall (79%). Applying SMOTE to oversample the minority class improved precision from 57% to 58%, with overall accuracy remaining at 87%. Regularisation tuning (C = 0.01) helped prevent overfitting while preserving predictive power. **(Appendices 3 and 4)**

To evaluate dimensionality reduction, PCA was used to reduce the input space to two components: PC1 (driven by exit behaviour) and PC2 (linked to page value engagement). Logistic regression trained on these retained solid performance (ROC-AUC = 0.8773, recall = 78%, precision = 58%), suggesting that core behavioural signals could be effectively summarised with minimal loss (Appendix 14). Strengths of this approach include clear interpretability, low overfitting risk, and consistently strong recall. Its main limitations are moderate precision and limited ability to model complex non-linear interactions compared to more advanced models. Nevertheless, logistic regression proved to be a reliable baseline, offering valuable insights into behavioural patterns driving conversion.

Random Forest:

Random Forest, a non-parametric ensemble method, was selected for its robustness to multicollinearity and capacity for capturing nonlinear interactions without extensive preprocessing. Class imbalance was managed via balanced class weighting, yielding an initial accuracy of 89% and ROC-AUC of 0.9229. The feature importance analysis highlighted PageValues_log as critical, alongside exit and session duration metrics, reinforcing their role in predicting conversion (Appendix 7). Hyperparameter tuning identified optimal parameters (200 estimators, unlimited depth), resulting in cross-validation accuracy of 89.4%, high precision (0.74), and recall (0.56) for revenue sessions. Calibration analysis revealed minor probability overestimation **(Appendix 5)**, indicating areas for further refinement. Decision confidence analysis showed clear predictions at extremes but ambiguity around mid-range probabilities, suggesting strategic intervention for uncertain predictions **(Appendix 6)**.

XGBoost:

XGBoost was employed due to its scalability, sensitivity to class imbalance, and ability to capture complex data structures more efficiently than Random Forest, particularly through its gradient boosting mechanism, which sequentially corrects errors made by previous trees. Initial modelling revealed strong overall accuracy (89.6%) but moderate recall (0.59). Adjustments via scale_pos_weight significantly boosted minority recall to 0.86 but compromised accuracy (85.2%). SMOTE achieved optimal balance (89% accuracy, recall = 0.76), confirming effective generalisation. Hyperparameter tuning (50 estimators, max depth = 3, learning rate = 0.1) improved ROC-AUC to 0.9279 and minority class recall to 0.60. Visual diagnostics demonstrated confident class separation, especially clear predictions, but uncertainty persisted near the threshold. Thus, XGBoost balances predictive robustness with sensitivity to data nuances, outperforming simpler models like logistic regression **(Appendix 7)**.

Linear Discriminant Analysis (LDA):

LDA was used for dimensionality reduction and interpretability, maximising between-class variance under linear decision boundaries. The single discriminant axis (LD1) effectively separated revenue sessions, as confirmed by KDE and strip plots. LDA coefficients emphasised PageValues_log, exit rates, and bounce rates as influential, aligning with prior analyses. Logistic regression on LD1 alone yielded impressive metrics (accuracy = 89%, recall = 0.60, precision = 0.72), confirming LD1's predictive strength. Despite its linear constraint, LDA provided valuable interpretative insights into behavioural drivers of conversions.

Support Vector Machines (Linear, RBF, Polynomial):

SVMs were tested with linear, RBF, and polynomial kernels to assess their ability to predict revenue. The Linear SVM offered strong interpretability and solid performance (ROC-AUC = 0.8815, recall = 0.64), identifying PageValues_log and session duration as key predictors. The RBF kernel slightly outperformed it in accuracy (89.31%) and ROC-AUC (0.8927) but had lower recall (0.56), indicating better capture of non-linear patterns but weaker conversion detection. The Polynomial kernel underperformed (ROC-AUC = 0.8638) due to overly complex boundaries.

The small performance gap between Linear and RBF suggests most relationships are linear. PCA and manual 2-feature plots showed compact, overlapping clusters, limiting separability (Appendix 13). Permutation importance from the RBF model confirmed PageValues_log and session duration as dominant features. Overall, the Linear SVM provided the best balance between performance and interpretability, with minimal gain from non-linear kernels.

Neural Network:

A feedforward neural network was implemented in PyTorch with three fully connected layers and ReLU activations, trained using BCEWithLogitsLoss and pos_weight to counter class imbalance. Log-transformed inputs and one-hot encoded categorical features were standardised to ensure stable learning across diverse scales. SMOTE was applied to the training data, which, when combined with class weighting, substantially improved minority class detection—addressing the key business need of identifying potential converters. Post-SMOTE, the model achieved excellent performance (ROC-AUC = 0.9599, accuracy ≈ 92%, precision/recall/F1 ≈ 0.92), with minimal overfitting as indicated by similar training/test ROC-AUC values (0.9950 / 0.957). Predicted probabilities were well-calibrated, making the model's outputs reliable for ranking users by conversion likelihood **(Appendices 10 and 11)**. Its strength lies in capturing complex, nonlinear patterns that tree-based models may miss, particularly in high-dimensional interactions. However, interpretability remains limited due to the black-box nature of neural networks, making them less transparent than linear or tree-based models. Despite this, the strong generalisation and probability confidence make it an ideal candidate for deployment or integration into ensemble frameworks.

Soft-Voting Ensemble:

A soft-voting ensemble combining Linear Discriminant Analysis (LDA), Random Forest (RF), and XGBoost was implemented to leverage their complementary strengths—LDA for linear interpretability, RF for non-linear pattern capture and robustness to outliers, and XGBoost for fine-grained performance in noisy, imbalanced data. This diverse trio (linear, bagged, and boosted) was trained on SMOTE-rebalanced data to handle class imbalance, with hyperparameters tuned via GridSearchCV (200 estimators for RF/XGBoost, max_depth=None for RF, and max_depth=5 and learning_rate=0.1 for XGBoost).

The ensemble achieved a cross-validated ROC-AUC of 0.9783 and generalised well (test accuracy = 88.7%, ROC-AUC = 0.9266, recall = 0.73, F1 = 0.69 for the revenue class). Its high CV AUC suggests consistent discrimination across folds (Appendix 10). Partial Dependence Plots (PDPs) confirmed PageValues_log as the top driver, followed by ExitRates_log (negative) and Month_Nov, aligning with seasonal insights (Appendix 6). While less interpretable than individual models, the ensemble delivered strong calibration, high recall, and adaptability—making it a practical candidate for real-world deployment.

Linear Regression: To explore what drives high-value sessions, linear regression was run on page_values_log—the strongest revenue predictor across models. With an R^2 of 0.14 and RMSE of 1.20, the model revealed that longer sessions, high interaction with product and administrative pages, and specific traffic sources (Types 7 and 20) were key drivers. Covariance analysis supported this, showing strong associations with session duration and product-related activity. These insights confirm that deeper engagement leads to higher page value, making these features strong proxies for conversion likelihood.

Non-Technical Interpretation and Conclusion

This analysis aimed to understand which types of user behaviour are most strongly linked to making an online purchase. By using a wide range of machine learning models—from simple, interpretable ones to more complex, high-performing ones—we were able to predict buying behaviour with high accuracy while also gaining actionable insights into what drives users to convert.

Simpler models like Logistic Regression and Linear Discriminant Analysis (LDA) provided clear explanations: users who spent time on high-value pages (like product or checkout pages), browsed during November (a peak sales month), or had longer, more engaged sessions were far more likely to make a purchase. These findings are immediately useful for marketing teams, such as targeting engaged users or tailoring promotions around seasonal trends.

More advanced models like Random Forest and XGBoost captured more complex behaviour patterns. They revealed that low exit rates, extended product page interactions, and returning visits were strong predictors of revenue. Though these models are harder to interpret directly, they excel at detecting subtle patterns and interactions that simple models might miss.

The neural network, when enhanced with class balancing using SMOTE, achieved the highest predictive performance, correctly identifying around 92% of buyers. While it doesn't offer clear explanations for its predictions, it is highly effective at detecting nuanced behavioural signals. Similarly, the soft-voting ensemble, which blends several model types, offered a strong compromise—balancing accuracy (89%) and minority class recall (73%)—making it well-suited for real-world deployment where both precision and consistency matter.

These models offer practical value for lead scoring, targeted marketing, and conversion optimisation by identifying users with high purchase intent based on their browsing behaviour—such as engaging with product pages, visiting during peak months like November, or spending longer on site. Such users can be prioritised for tailored incentives or follow-up strategies. While clustering was explored to group users into behaviour types, the results showed that user actions exist on a gradual scale rather than in clear-cut categories. This highlights the benefit of using continuous behavioural scores over fixed segments to deliver more personalised and timely marketing actions.

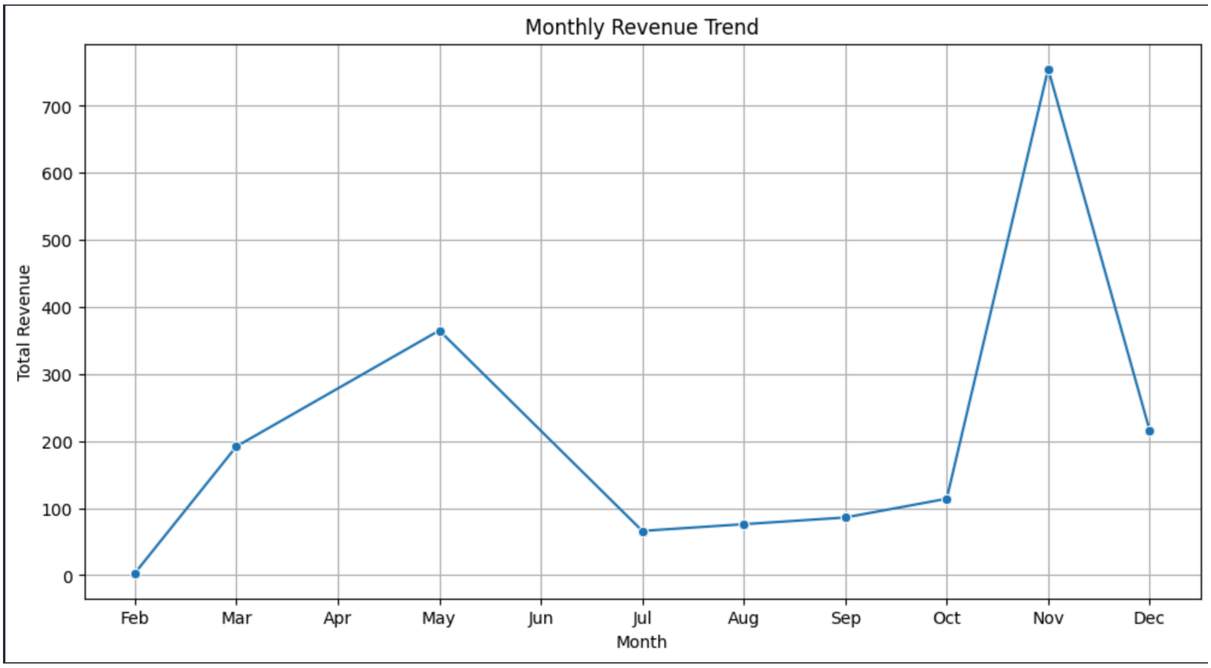
Conclusion

This project applied a suite of machine learning methods to predict revenue-generating sessions, balancing performance and interpretability. Across all models, PageValue_log consistently emerged as the most predictive feature, alongside exit rates, session duration, and November seasonality—validated through Partial Dependence Plots and binned visualisations.

Logistic regression delivered ~86% accuracy and high interpretability, making it useful for understanding behavioural drivers. Random Forest and XGBoost effectively captured non-linear interactions, achieving ROC-AUC scores above 0.92. The neural network, enhanced with SMOTE, provided the best overall performance (ROC-AUC = 0.96, precision/recall ≈ 0.92). The soft-voting ensemble offered a strong balance, reaching 88.7% accuracy and 0.9266 ROC-AUC with improved recall (0.73) for conversions.

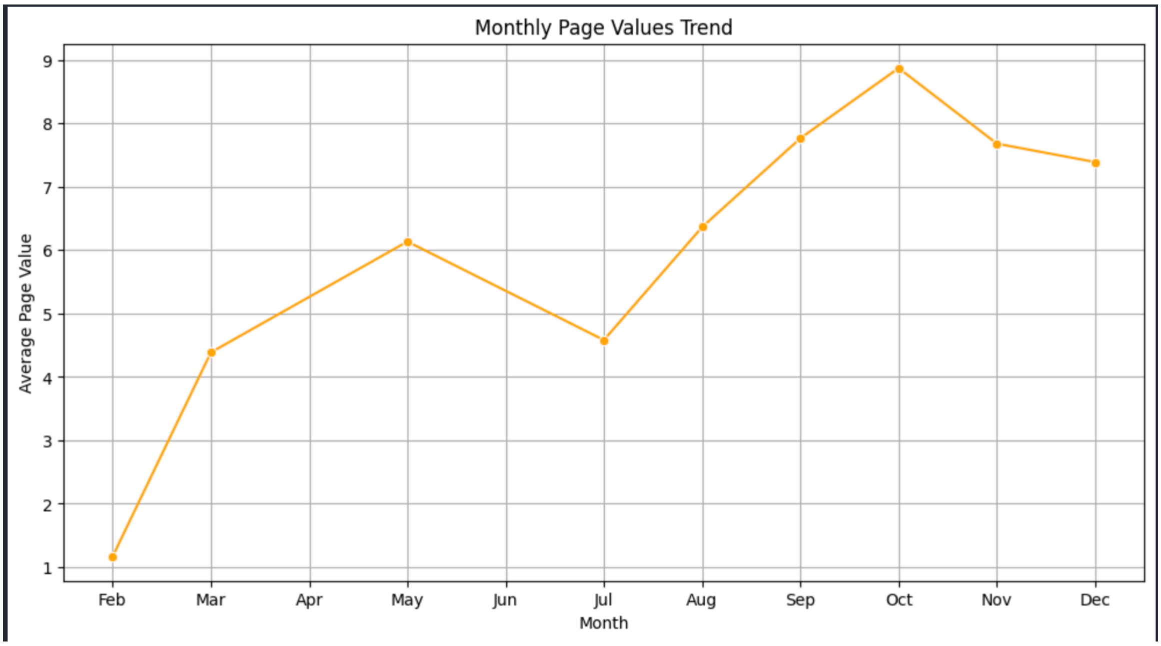
While complex models improved accuracy, they required more tuning and offered less transparency. Key limitations included residual multicollinearity, dependence on SMOTE, and static session-level features. Future work could explore sequential modelsOverall, this multi-method approach delivered strong predictive performance and actionable insights, providing businesses with effective tools to understand and influence online user conversions.

Appendix



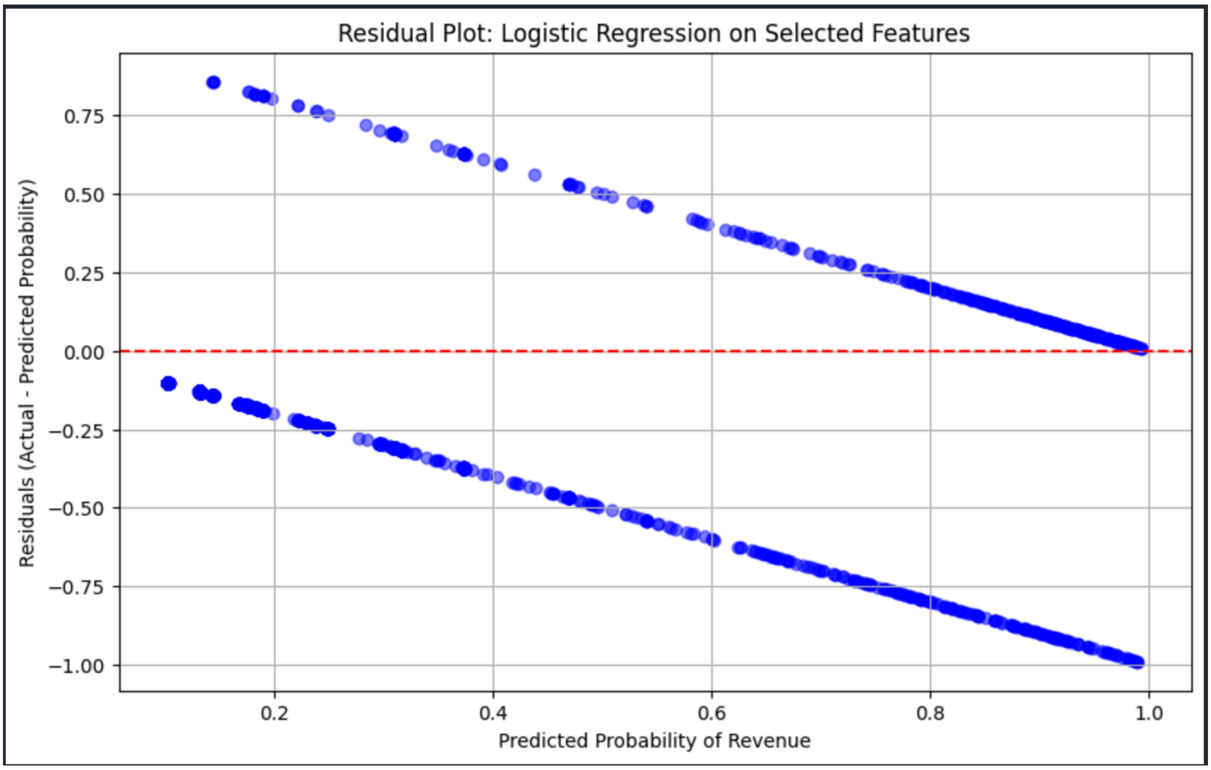
Appendix 1: Monthly Revenue Trend Analysis

The monthly revenue line chart reveals clear seasonal dynamics. Revenue shows a steady rise from February to May, with a modest peak in May, followed by a sharp decline during June and July—likely due to reduced engagement over the summer period. The most prominent increase occurs in November, marking the highest revenue point, likely driven by major seasonal events such as Black Friday and Cyber Monday. Although revenue remains relatively high in December, it declines from the November peak, suggesting a typical post-promotion drop-off. These patterns underscore the need to synchronise marketing efforts with seasonal highs—particularly in November—to maximise conversions and revenue impact.



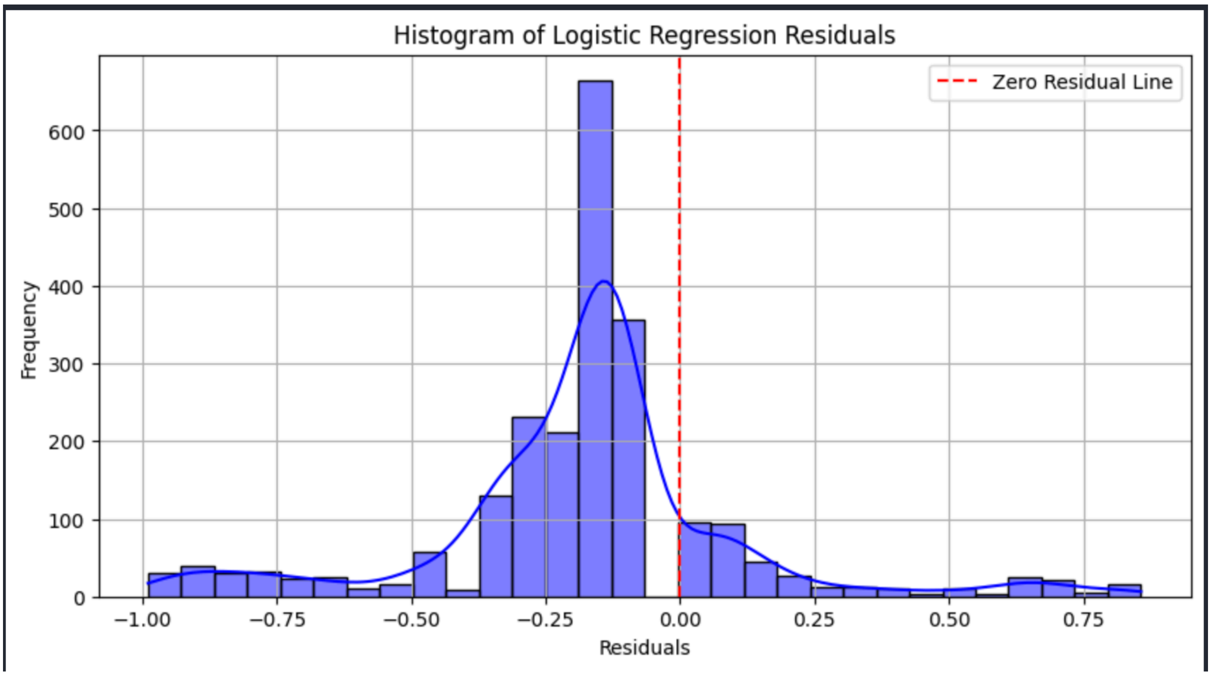
Appendix 2: Monthly Page Value Trend Analysis

The line chart of average monthly PageValues shows a consistent upward trend from February onwards, suggesting increasing user engagement with revenue-driving content. PageValues start just above 1 in February and peak in October at nearly 9, indicating highly valuable sessions likely tied to pre-holiday research activity. In contrast, November—despite generating the highest total revenue—sees a slight dip in average PageValue. This suggests a surge in lower-value sessions, possibly driven by increased but less engaged traffic during major sales events. The findings highlight a trade-off between session volume and quality, reinforcing the importance of targeted engagement during high-traffic periods.



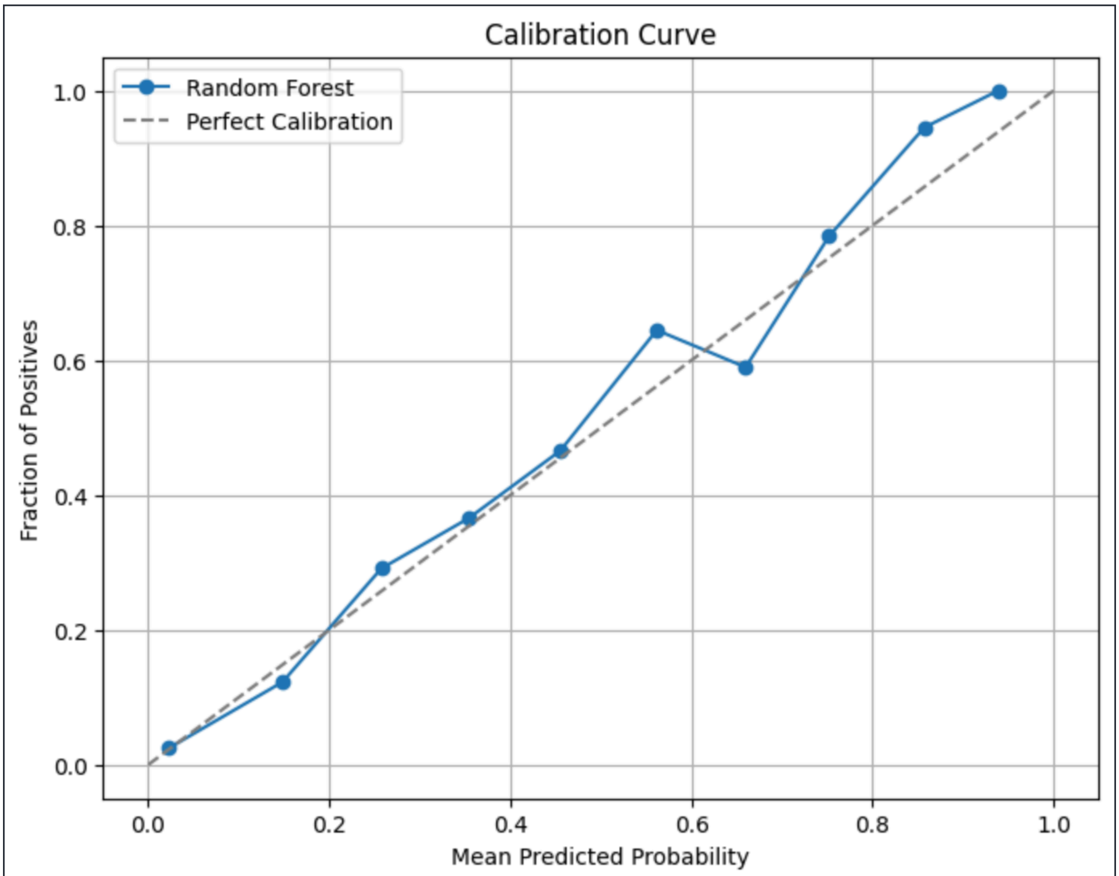
Appendix 3: Logistic Regression Residual Plot Analysis

The residual scatter plot illustrates the difference between actual and predicted probabilities for the logistic regression model. Residuals are symmetrically distributed around the zero line, forming two linear bands corresponding to correct and incorrect classifications. This structure indicates good model calibration and confidence in its predictions. Slight widening at the extremes of predicted probabilities suggests minor uncertainty in highly confident predictions, but overall the plot supports a well-fitted and reliable model.



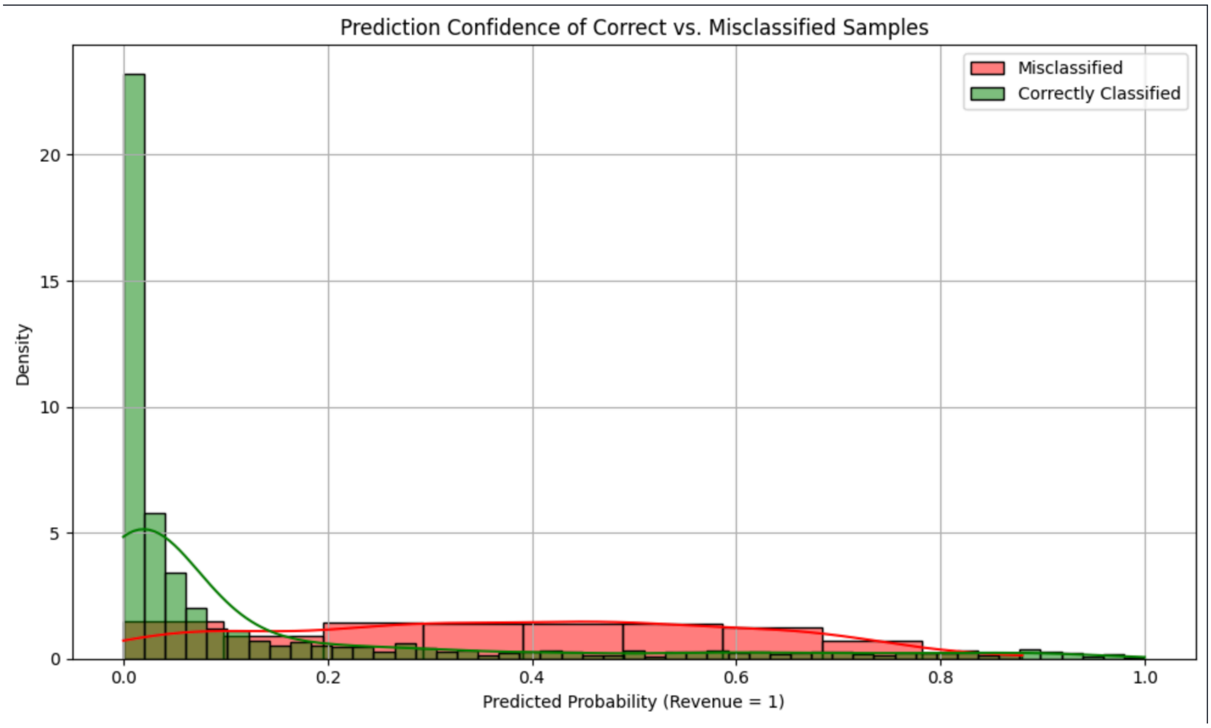
Appendix 4: Histogram of Logistic Regression Residuals

The histogram of residuals shows a sharp peak slightly left of zero, with most values concentrated in the negative range. This slight left skew indicates the model tends to mildly underestimate the probability of revenue generation. Despite this, the distribution remains tight with limited extreme values, supporting a good overall model fit. The modest tails suggest natural variability in user behaviour, but no strong systematic bias is present.



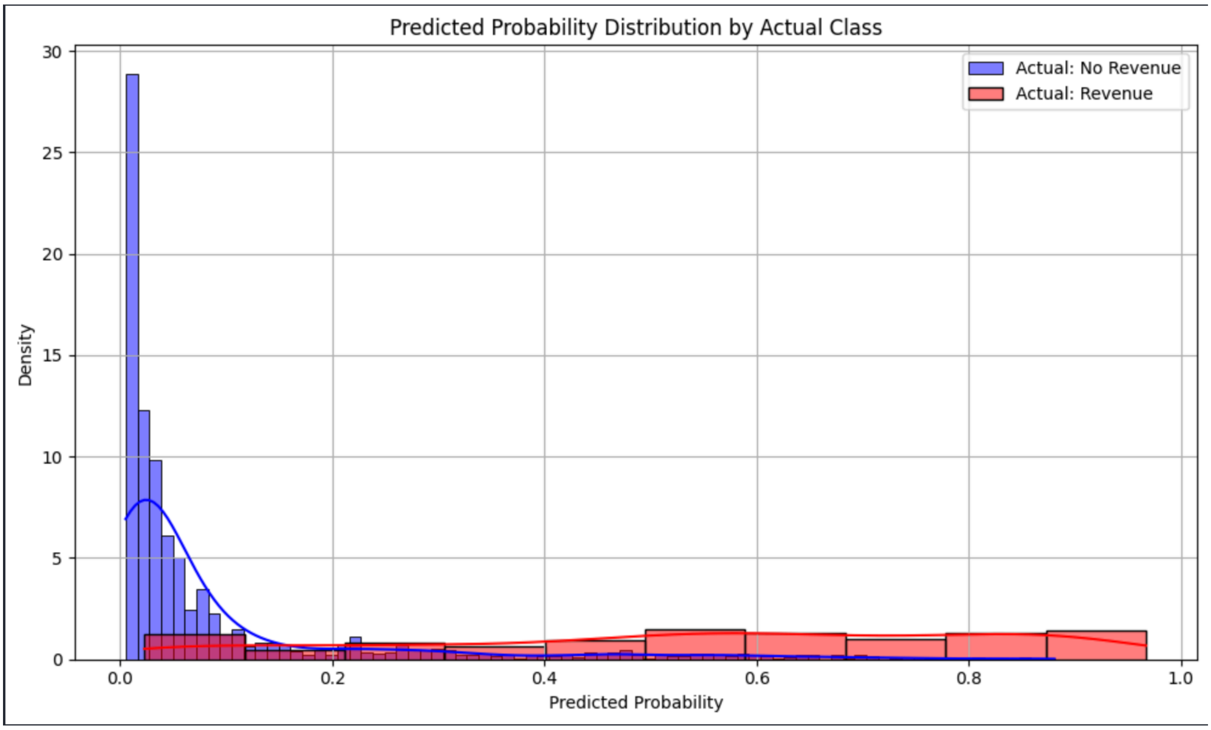
Appendix 5: Calibration Curve – Random Forest

The calibration curve for the Random Forest model closely follows the ideal diagonal, indicating good alignment between predicted probabilities and actual outcomes—particularly in mid-to-high probability ranges. However, a slight overestimation is observed around the 0.6–0.7 interval, suggesting mild miscalibration. While overall confidence is reliable, post-hoc calibration methods like Platt scaling or isotonic regression could enhance probability accuracy for deployment scenarios requiring precise thresholding.



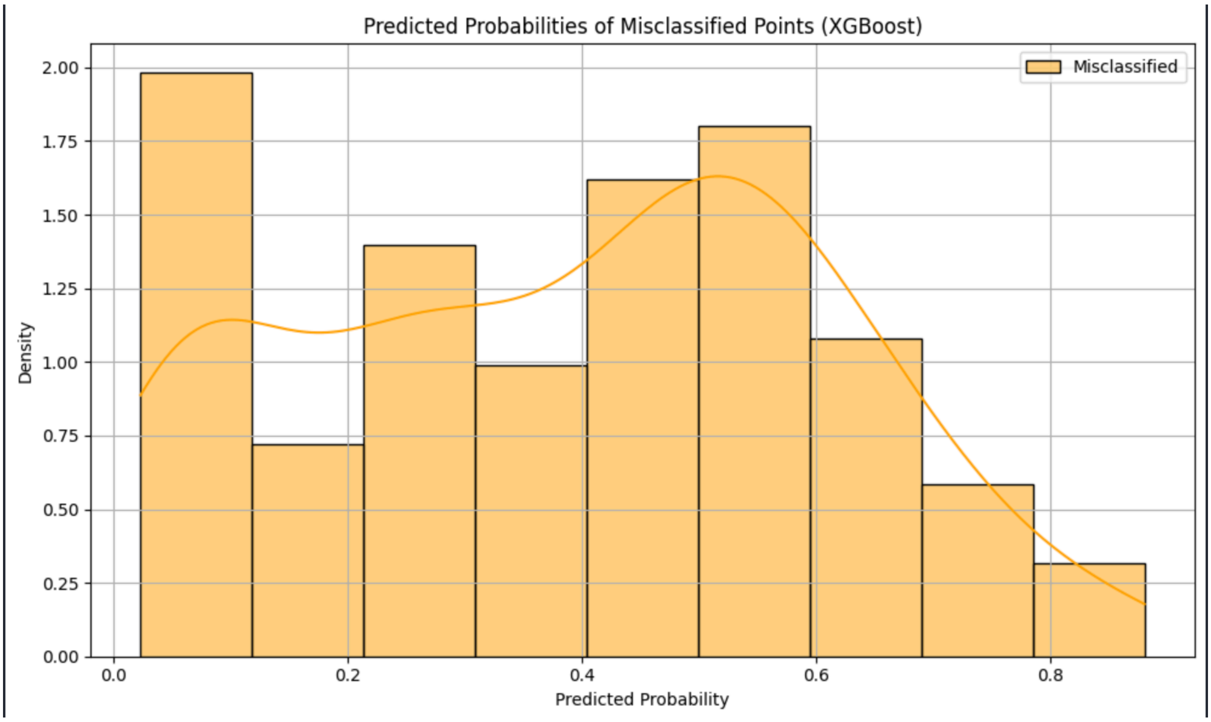
Appendix 6: Prediction Confidence Distribution – Revenue Class - random forest

The confidence distribution plot highlights a clear distinction between correctly and incorrectly classified revenue sessions. Correct predictions cluster at low probabilities, reflecting confident identification of non-converting users. Misclassifications peak between 0.2 and 0.6, suggesting model uncertainty near the decision boundary. Minimal overlap at the probability extremes indicates strong confidence when predictions are accurate. These insights support the use of confidence-based thresholding to refine decision-making, particularly in cases where false negatives or false positives carry different business implications.



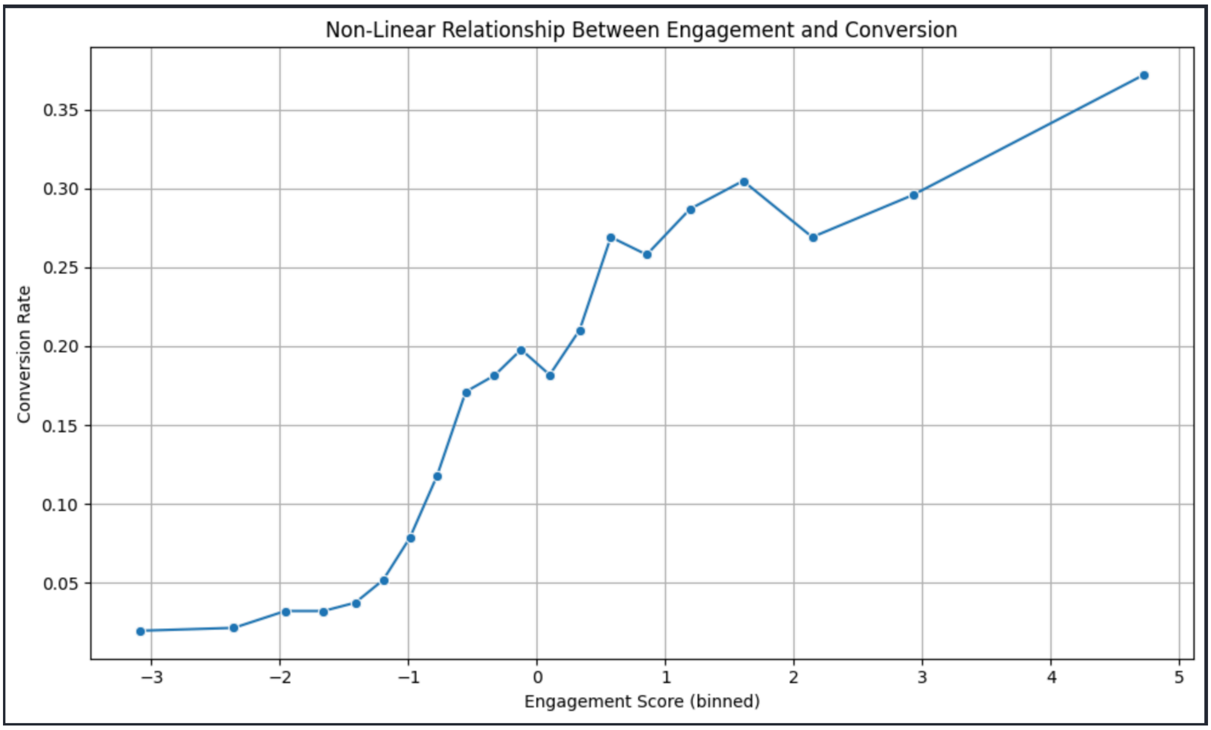
Appendix 7: XGBoost Predicted Probability Distribution

The class-wise probability distribution from XGBoost demonstrates clear separation between revenue and non-revenue sessions. Non-converting sessions (blue) are concentrated near 0, indicating high confidence in negative predictions. Converting sessions (red) are more spread, with increased density above 0.5—showing the model reasonably captures positive class likelihoods. This separation aligns with XGBoost's high ROC-AUC, reinforcing its discriminative power. Practically, it supports confident targeting of likely buyers while reducing false positives from uncertain sessions.



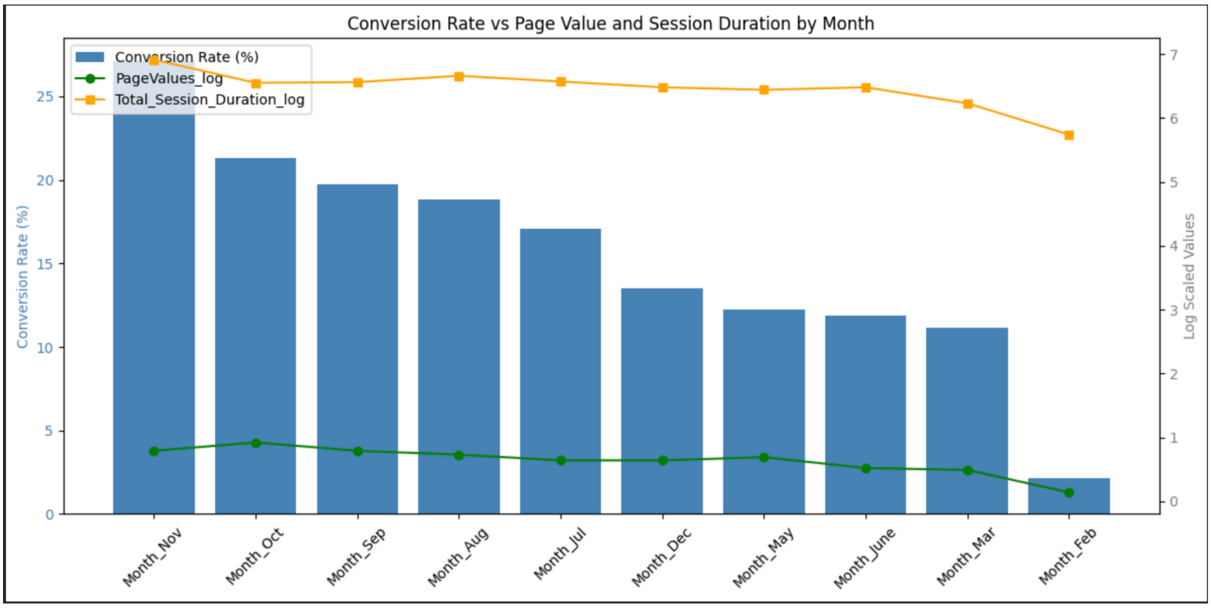
Appendix 8: XGBoost Misclassified Probability Distribution

This histogram focuses on predicted probabilities for misclassified samples. Most errors cluster between 0.2 and 0.6, peaking near the 0.5 threshold—highlighting uncertainty in borderline predictions. Very few misclassifications occur at the extremes, confirming that XGBoost is generally reliable when confident. For practical use, this suggests high-confidence outputs can be trusted, while medium-range scores may warrant additional scrutiny or intervention.



Appendix 9: Conversion Rate vs. Binned Engagement

The line plot of actual conversion rates across binned engagement scores shows a non-linear relationship. While conversions tend to rise with engagement, the curve includes sharp increases, plateaus, and dips—implying that high engagement doesn't always lead to purchase. This behavioural complexity supports the use of non-linear models (e.g. Random Forests), which are better suited to capture such irregular patterns compared to linear approaches.

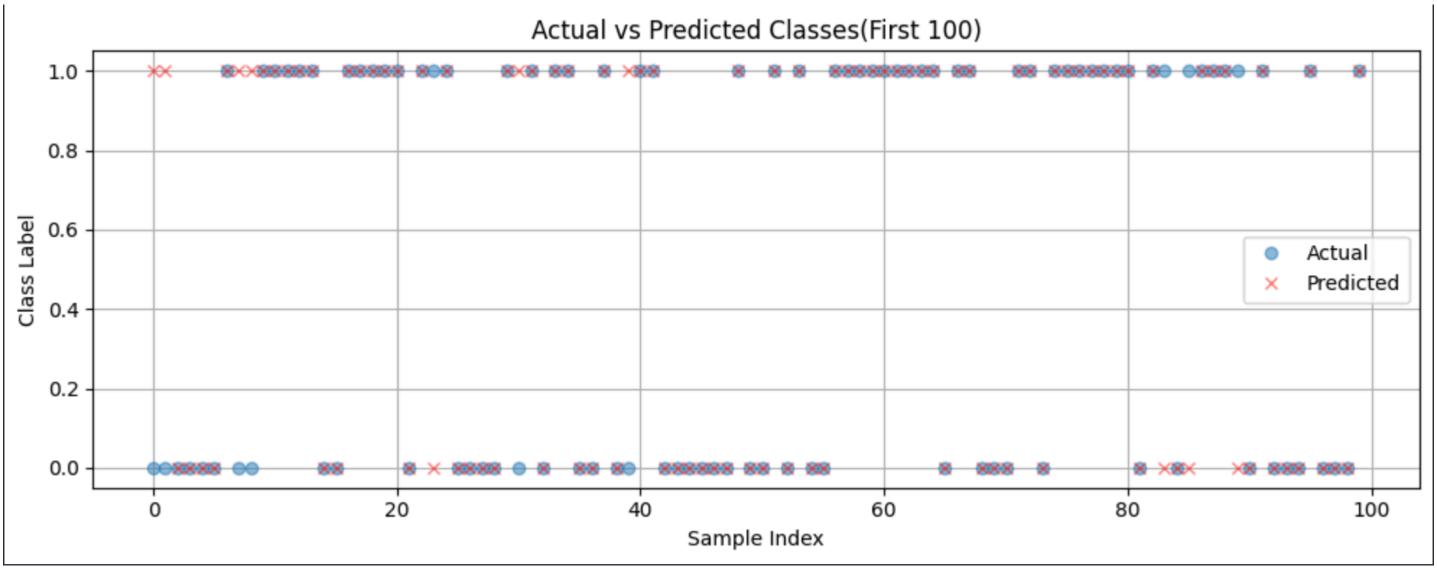


Appendix 10: Monthly Conversion Patterns and Behavioural Drivers

Monthly conversion rates show strong seasonal variation. November leads with the highest conversion rate (27.13%), followed by October (21.31%) and September (19.72%), while February performs poorest (2.14%). Notably, session volume alone doesn't explain these trends—May, with the most sessions, has a low conversion rate (12.26%), while October achieves high conversion with low traffic, suggesting that session quality outweighs quantity.

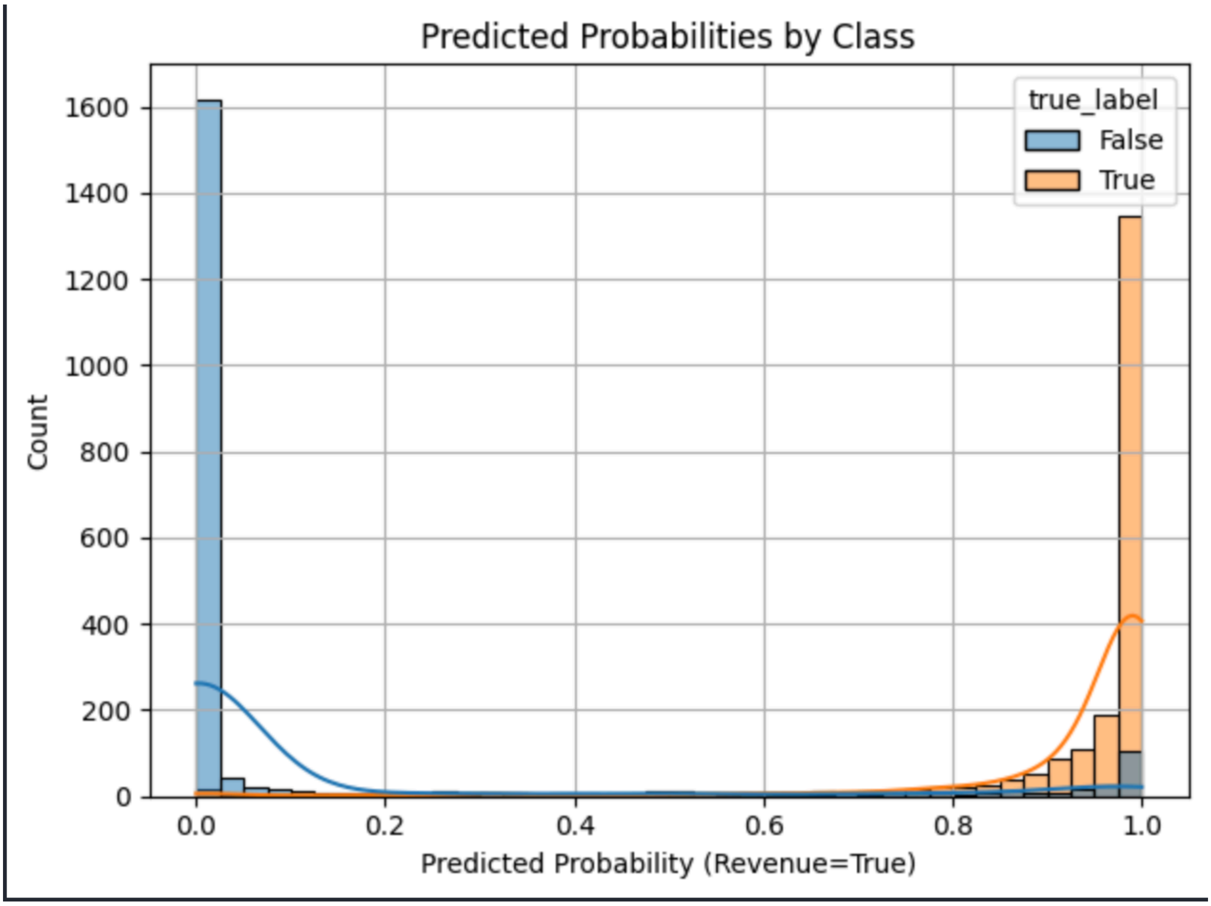
Despite assumptions, high-converting months did not align with high SpecialDay scores—November and October had zero values, whereas February and May, with the highest SpecialDay averages, saw poor conversion. This suggests promotional proximity isn't a reliable standalone driver.

Instead, engagement metrics were key: October had the highest average PageValues_log (0.92), while November had the longest session durations (6.91, log-scaled), indicating deeper, value-driven browsing behaviour during these peak months. These findings highlight that conversion success is more tightly linked to session depth and engagement than calendar events.



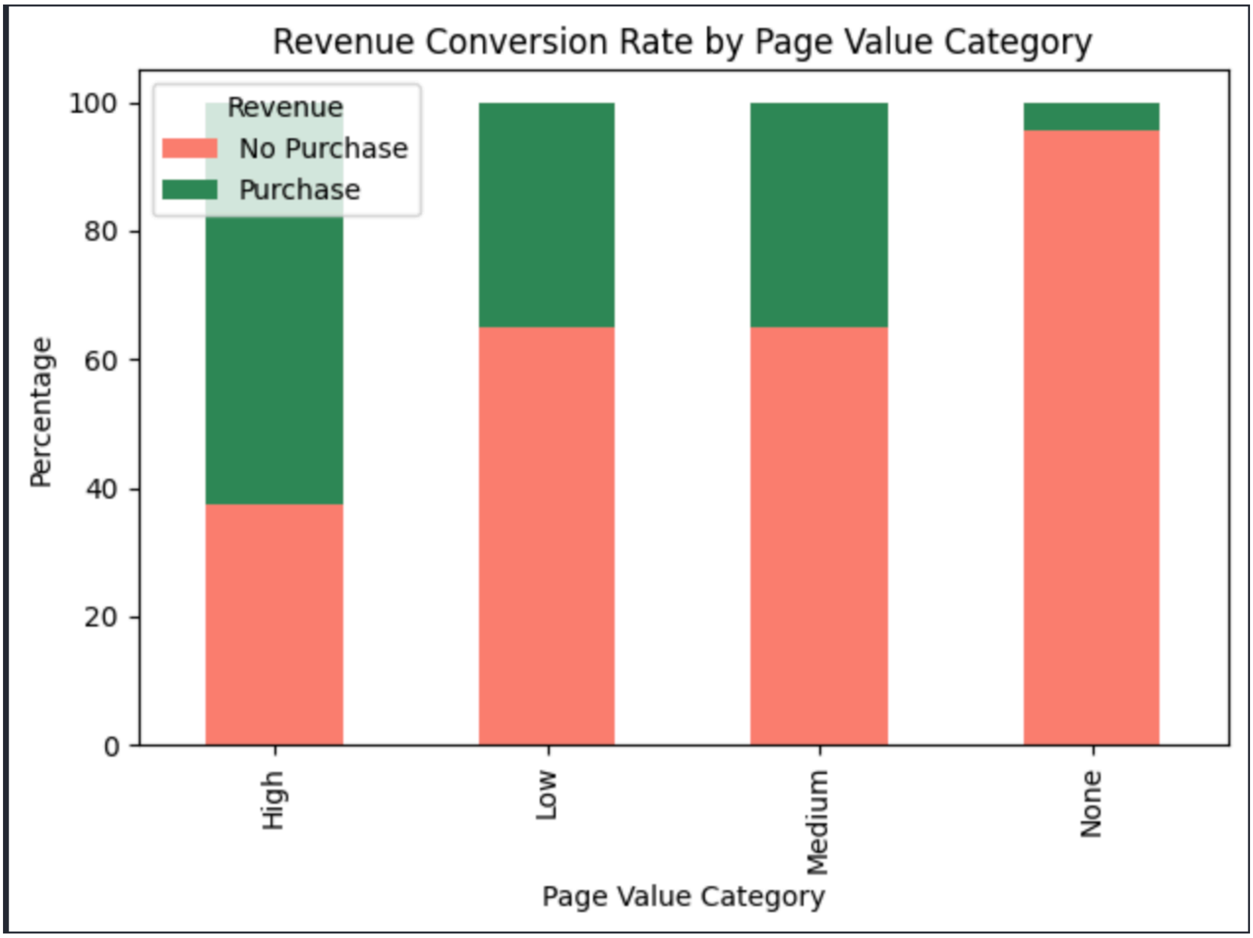
Appendix 11: Actual vs Predicted Class Scatter Plot (First 100 Samples) - Neural Network

The scatter plot comparing actual and predicted classes for the first 100 test samples demonstrates strong model accuracy. Blue circles (true labels) and red crosses (predictions) show high overlap, indicating consistent correct classifications across both classes. Misclassifications are minimal and evenly spread, supporting the model's reported **93% test accuracy** and **ROC-AUC of 0.9599**. The visual alignment confirms that the model generalises well, handles class imbalance effectively, and produces reliable predictions on unseen data.



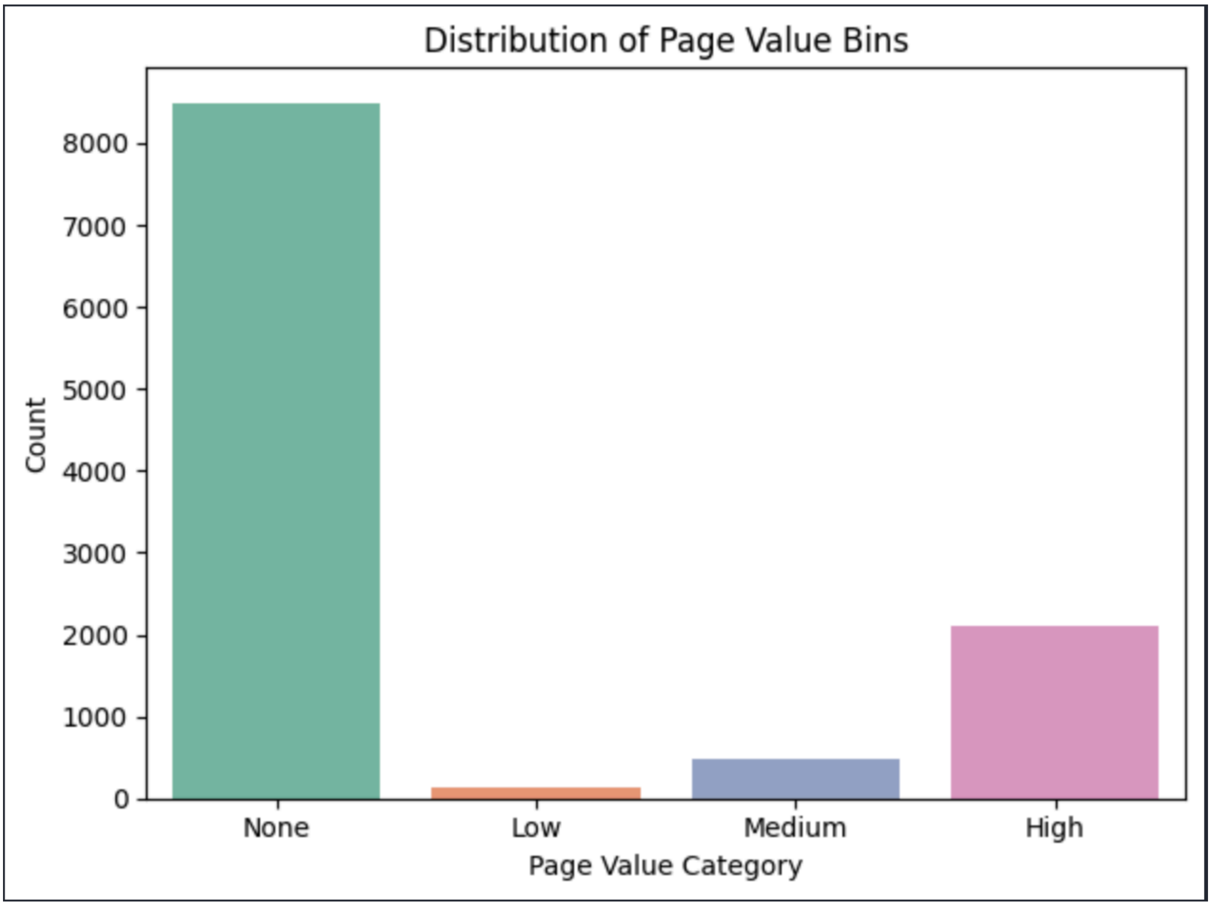
Appendix 12: Predicted Probabilities by Class Histogram - Neural Network

This histogram shows the distribution of predicted probabilities for each true class. Predictions for the negative class (False, blue) are concentrated near **0.0**, while those for the positive class (`True`, orange) peak sharply near **1.0**. The distinct separation indicates that the model assigns probabilities with high confidence and minimal overlap, aligning with its strong **ROC-AUC score**. This well-calibrated output enhances trust in the model's predictions—especially in applications where decisions depend on probability thresholds.



Appendix 13: Revenue Conversion Rate by Page Value Category

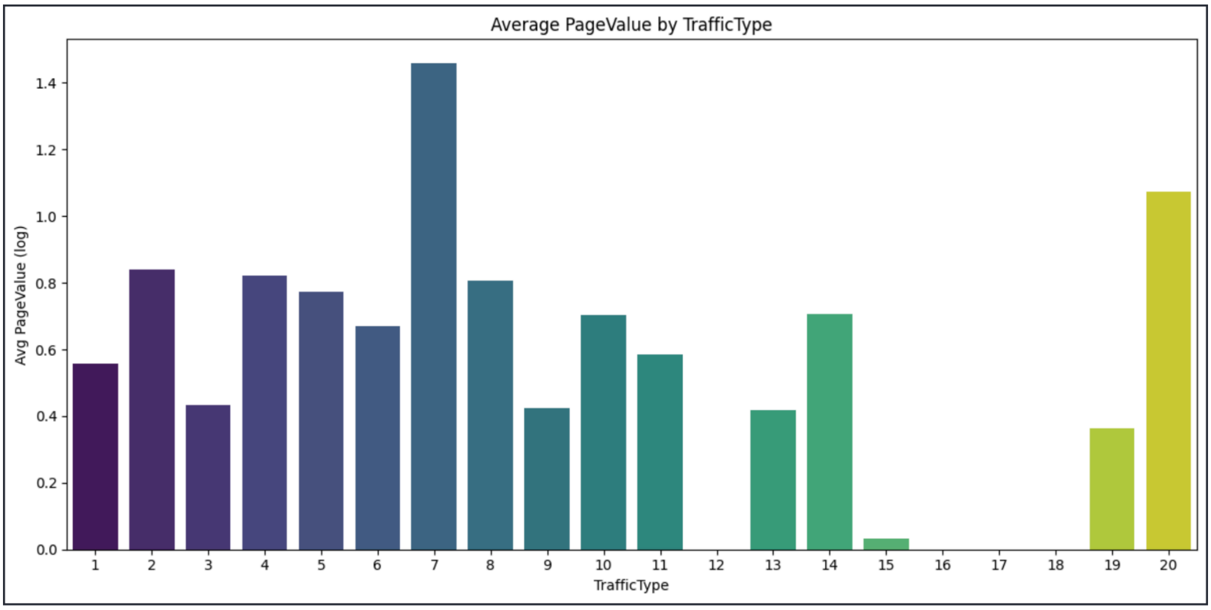
This stacked bar chart shows the percentage of purchase vs. non-purchase sessions across four Page Value categories: High, Medium, Low, and None. A clear upward trend is visible—conversion rates rise with page value, peaking in the High category where over 60% of sessions result in purchases. In contrast, the None group has the lowest conversion (~5%). Low and Medium groups display similar conversion rates (~40%), indicating that any engagement with value-generating content improves purchase likelihood, but the biggest gains occur at high page values. From a business standpoint, this confirms that Page Value is a strong proxy for purchase intent, and driving users toward high-value content—like product or checkout pages—can significantly increase conversions.



Appendix 14: Distribution of Page Value Categories

This bar chart illustrates the distribution of sessions across four Page Value categories: None, Low, Medium, and High. A striking imbalance is evident—the vast majority of sessions fall into the 'None' category, indicating little to no interaction with value-generating content. In contrast, High Page Value sessions, though far fewer, are strongly associated with conversions, as shown in prior analyses.

This imbalance underscores a key insight: while most users do not engage with transactional content, those who do are significantly more likely to convert. It also highlights the importance of ensuring models don't overfit to the dominant 'None' group, but instead focus on the behavioural significance of Page Value as a strong proxy for purchase intent.



Appendix 15: High PageValue Traffic Types and Revenue Performance

This bar chart and summary table identify top-performing traffic sources based on average PageValue_log—a key predictor of conversion. TrafficType 7 leads with the highest PageValue (1.46) and a 30% conversion rate, despite lower volume (40 sessions), indicating high purchase intent—likely from targeted campaigns such as loyalty or remarketing.

TrafficType 20 follows closely, with a 28% conversion rate and a PageValue of 1.07, supported by strong engagement across 174 sessions. TrafficType 2, while slightly lower in PageValue (0.84), combines scale (3832 sessions) with high product interaction, making it a consistent and valuable source.

These traffic types demonstrate the clearest link between channel and conversion intent. Focusing on these high-performing sources—via expanded targeting, campaign replication, or enhanced user experience—could improve overall performance. Lower-yielding types may require reevaluation of targeting or creative strategy.

References to Python Libraries and Use Cases

1. **Pandas**
   -Used for data loading, cleaning, wrangling, and exploratory analysis (EDA). Enabled efficient manipulation of tabular session data and feature engineering for modelling.

2. **Numpy**
   -Used for numerical operations, including array manipulations, log transformations, and vectorised computations to support feature processing and model preparation.

3. **Scikit-learn (sklearn)**
   -Used extensively for model training (Logistic Regression, SVM, Random Forest, PCA, etc.), hyperparameter tuning (GridSearchCV), dimensionality reduction (PCA), performance evaluation (ROC-AUC, precision, recall, classification report), preprocessing (e.g., StandardScaler, OneHotEncoder), and feature selection (RFE, VIF calculation).

4. **Imbalanced-learn (imblearn)**
   -Used to apply SMOTE (Synthetic Minority Oversampling Technique) for addressing class imbalance by generating synthetic samples for the underrepresented revenue class during training.

5. **Scipy**
   -Used for statistical analysis and outlier treatment, including the application of Z-score methods and IQR filtering during data preprocessing.

6. **Matplotlib**
   -Used to create custom visualisations such as bar plots, line charts, confusion matrices, and decision boundary plots to support technical and non-technical communication.

7. **seaborn**
   -Used to improve the visual quality of data distributions, KDE plots, heatmaps, and correlation matrices during exploratory data analysis.

8. **Xgboost**
   -Used for training gradient-boosted tree models (XGBoostClassifier) known for high predictive accuracy. Also employed for hyperparameter tuning and extracting feature importances.

9. **Pytorch**
   -Used to build and train a feedforward neural network for binary classification. Implemented custom layers with ReLU activations and trained using BCEWithLogitsLoss and pos_weight to manage class imbalance.

10. **Sklearn.metrics & sklearn.inspection**
    -Used for generating calibration curves and evaluating model reliability