UNIVERSITY OF CENTRAL OKLAHOMA

COLLEGE OF BUSINESS, GRADUATE UNIT

MSBA 5404: PREDICTIVE ANALYTICS AND ARTIFICIAL INTELLIGENCE (CAPSTONE)

FINAL RESEARCH REPORT

Predicting Online Shopping Purchases

Prepared by

Saugat Raut

Dr. Ho-Chang Chae

5/7/2025

TABLE OF CONTENTS

1. ABSTRACT

In the modern e-commerce environment, converting browsing users into paying customers remains a core business challenge. Research shows that nearly 70% of online shopping carts are abandoned, resulting in significant lost revenue (Rajamma, Paswan, & Hossain, 2009). While businesses often invest in strategies like email follow-ups and online ads to re-engage customers, these efforts are not always effective. This study addresses this by developing a predictive modeling framework for identifying user conversions using session-level e-commerce data. By integrating machine learning algorithms—such as Random Forest and Adaptive LASSO—with behavioral theory, this research aims to enhance predictive performance while producing interpretable insights for understanding user behavior. Novel features like Session Efficiency were engineered to reflect user engagement quality and purchasing likelihood. The strongest performing model, Random Forest, achieved a recall score of 0.9067 with strong generalization across both training and validation sets. Logistic Regression and Adaptive LASSO provided additional interpretability through odds ratios and standardized coefficients. To conceptualize the results, key features were linked to the Technology Acceptance Model (TAM), aligning session behavior with constructs such as Perceived Usefulness and Perceived Ease of Use. Findings indicate that users with efficient navigation paths and deeper engagement with product related content are significantly more likely to convert to purchasers. This study contributes a robust and theoretically backed approach for predicting purchases and offers insights into optimizing user experience, marketing strategies, and digital engagement in the e-commerce industry.

2. INTRODUCTION

One of the biggest challenges in online retail is shopping cart abandonment, where users interact with an e-commerce website, browse products, and sometimes even initiate checkout, but leave before completing the transaction. Research shows that nearly 70% of online shopping carts are abandoned, causing businesses to lose significant potential revenue (Rajamma et al., 2009). While some customers leave due to unexpected costs, security concerns, or a complicated checkout process, others simply never intended to make a purchase in the first place. The key challenge is identifying which customers are likely to buy and which ones are merely browsing. Addressing this issue requires businesses to implement data-driven decision-making approaches that enable them to anticipate customer behavior and take proactive actions to drive conversions (Sakar et al., 2019).

The existing literature on UCI's Online Shopper Purchasing Intention dataset primarily focuses on experimenting and evaluating the performance of different machine learning algorithms, preprocessing methods, and model tuning. The initial research implemented a Multilayer Perceptron and an LSTM recurrent neural network to perform real-time predictions on the purchasing intention of online shoppers (Sakar et al., 2019). It demonstrated the viability of LSTM recurrent neural networks and Multilayer Perceptron models for real-time prediction of online consumer behavior data.

Although many studies have modeled online purchasing behavior using the UCI dataset, few studies have attempted to connect consumer behavioral features to theoretical constructs derived from the Technology Acceptance Model (TAM). TAM emphasizes factors like perceived usefulness and ease of use (Davis, 1989; Andrina, D., et al., 2022). This creates an opportunity for new research that combines predictive power with behavioral interpretability.

3. MOTIVATION

3.1 PROBLEM BACKGROUND

Many businesses use broad, untargeted marketing campaigns that do not differentiate between high-intent buyers and casual visitors, resulting in wasted advertising budgets and missed revenue opportunities. Shopping cart abandonment can stem from unexpected costs, lack of payment options, and general indecision (Rajamma et al., 2009). While efforts have attempted to address this problem through retargeting ads and abandoned cart email reminders, these methods are often ineffective because they lack personalization and fail to account for real-time customer behavior. A more effective approach involves predictive modeling, where businesses can identify at-risk customers before they leave the site and take immediate action to influence their decision. This study applies machine learning techniques in SAS Enterprise Miner to classify customers based on their likelihood to convert.

3.2 PROJECT SIGNIFICANCE

The rise of smart devices and social media has transformed the way people shop online. The Internet has evolved into a powerful marketing tool that promotes domestic and international commerce because of its expansion into a global interconnected network for information sharing and distribution (Khan et al., 2023). This has reshaped international retail by giving consumers the ability to shop from national and international markets without the need for physical travel. In fact, more than 82% of international buyers shop at least once from foreign websites annually (Ramkumar & Jin, 2019).

By uncovering meaningful patterns in session-level behavioral data, machine learning algorithms help businesses dynamically adjust their marketing strategies to influence purchasing decisions at the most crucial moments. Beyond its practical impact, this work establishes an adaptive framework that can serve as a foundation for future studies in predictive analytics and consumer behavior modeling. Additionally, this methodology can be further enhanced to improve predictive performance and applicability in competitive digital markets.

4. LITERATURE REVIEW

4.1 DATA PREPARATION

A major obstacle in the dataset is the class imbalance of the target variable. Approximately 85% of sessions did not result in a purchase, and this is a pattern that can be comparable to the norm in real-world e-commerce behavior. Several studies highlight the importance of addressing the class imbalance to avoid misleading performance and improve generalization. Rana et al. (2023) and Frazier et al. (2022) employed techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and under-sampling to balance the classes. Other papers chose a different route, such as Torres Trevino and Cepeda (2024), which implemented cost-sensitive learning and custom loss functions to penalize false negatives more heavily, especially in recall-optimized models.

Feature engineering and preprocessing for the dataset have been done in several ways. Nearly every paper on the topic applies encoding for categorical features with one-hot-encoding or label encoding. Furthermore, the existing literature applies transformations and standardization of continuous variables to improve model convergence for those that are sensitive to having a difference in scales among features. Some common transformation and standardization techniques included: Min-Max Scaling and Z-Score Standardization. Wen et al (2023), showed that applying a log transformation or binning to highly skewed variables like Page Value significantly improved model performance. For data partitioning, Rana et al. (2023) and Frazier et al. (2022) explicitly mentioned that they divided the data into training and test sets using stratified sampling. However, other papers fail to reveal their specific method for splitting the data.

4.2 MODELING

A variety of machine learning algorithms have been applied to the task of predicting online shopping purchases, each study tests different model types to address tradeoffs in performance and sensitivity to class imbalance. After examining the existing literature, researchers commonly utilize tree-based ensemble methods, linear classifiers, probabilistic models, and deep learning systems. Random Forest has emerged as one of the most frequently used algorithms due to its robustness in predicting complex relationships and resistance to overfitting. Studies by Adhikari (2023) and Satu & Islam (2023) consistently reported that Random Forest was one of the top performing models, and this is due to its versatility and resilience to noisy data. XGBoost and Gradient Boosting have also gained significant popularity for their high predictive accuracy and ability to model complex non-linear relationships. Frazier et al. (2022) identified XGBoost as the best performing model in terms of accuracy, recall, and F1-score. AdaBoost was leveraged by Sunarto et al. (2023) and Kurniawan et al. (2020) to improve classification performance, especially when combined with SMOTE and feature selection techniques. Logistic Regression is commonly used as a baseline model in multiple studies due to its simplicity and interpretability. Additionally, its used as a benchmark for comparison with more complex modeling techniques. Islam et al. (2023) used support Vector Machines (SVM) and Naive Bayes models to assess classification effectiveness, and both models were a part of their recommended techniques for achieving the most accurate results. Additionally, deep learning techniques have been employed on the UCI dataset to model session or sequential behavior. The original study published on the dataset, Sakar et al. (2019) applied both Multilayer Perceptron and LSTM (Long Short-Term Memory) networks for real-time prediction. Their findings showed that LSTM performed the best, indicating the advantage of modeling temporal dependencies in consumer navigation behavior. More recently, CatBoost—a gradient boosting technique that naturally handles categorical variables and applies ordered boosting—was applied alongside other popular algorithms and achieved the highest recall and overall accuracy in their survey-based dataset (Tayal & Daniel, 2024).

4.3 ASSESSMENT

The evaluation of assessment methods in the literature regarding the predictive modeling of online shopping purchases has been approached with various methodologies. Each study contributes its distinct insights into the topics of predictive modeling and consumer behavior. However, one must go beyond using accuracy as the primary measure to effectively assess model performance on the subject. The reason behind this is due to the limited information it provides when applied to highly imbalanced data such as UCI's Online Shoppers Purchasing Intention dataset, where non-purchase sessions dominate by a wide margin. In situations like this, a model's accuracy results may achieve a deceptively high score by predicting non-purchase behavior while failing to capture the minority class (actual purchases), which is the focus of marketing analytics (Rana et al., 2023; Sakar et al., 2019).

To address this issue, this study and the existing literature emphasize a broader set of assessment measures. Precision, which quantifies the proportion of predicted purchases that are actual purchases, is important for minimizing wasted targeting efforts and unnecessary marketing expenses. Recall, on the other hand, displays the model's ability to correctly identify all true purchase events, which is a vital measure for maximizing the conversion rate of purchases. The F1-score, which is a measure that combines precision and recall, is useful for its ability to balance the cost of false positives and false negatives (Frazier et al., 2022; Islam et al., 2023). Furthermore, multiple studies also evaluate the use of AUC-ROC scores and confusion matrices to better understand the predictive power of classifiers in imbalanced data.

Despite the improvements in assessment methods, the literature reveals inconsistent handling of class imbalance. Several papers apply oversampling techniques such as SMOTE (Frazier et al., 2022; Kurniawan et al., 2020) or cost-sensitive learning (Torres Treviño & Cepeda, 2024), many studies still evaluate models on raw data, which potentially causes the performance results to be flawed. This error may lead to inflated accuracy and precision scores while failing to predict true purchases in real-world applications.

A noticeable gap in the literature is the limited integration of interpretable methods. Very few papers incorporate explainability into their model outputs, which is critical for business contexts. One exception is Wen et al. (2023), where Shapley Additive Explanations (SHAP) is integrated into their modeling pipeline. However, other studies are still treating their models as black boxes, and without interpretable results, adoption into real-world applications will remain limited. Additionally, the existing literature on the topic fails to incorporate any marketing or behavioral theories to support their research findings. A potential opportunity to apply this would be linking a consumer behavioral framework like the Technology Acceptance Model, which identifies perceived usefulness and perceived ease of use as primary drivers of system adoption. These theories closely align with consumer behavior features in the UCI dataset, such as Page Value, Bounce Rates, and Exit Rates. Page Value represents the perceived value and relevance of a webpage, while Exit Rates and Bounce Rates can be a representation of navigation ease and user satisfaction. Recent TAM-based studies (e.g., Andrina et al., 2022; Purwianti et al., 2024) have shown these behavioral dimensions to be significant predictors of e-commerce engagement. Incorporating TAM as a theoretical foundation provides a more interpretable understanding of shopper behavior beyond model performance alone.

4.4 CONCLUSIONS

The literature on the UCI Online Shoppers Purchasing Intention dataset reveals a crossroads of machine learning, consumer behavior analytics, and practical challenges seen in online retail. A recurring theme throughout the studies is the need to understand and reduce shopping cart abandonment, which is responsible for substantial revenue loss in e-commerce. Numerous modeling efforts have demonstrated the application of advanced machine learning algorithms and proved the effectiveness of its use in predicting online consumer behavior. However, much of the success hinges heavily on proper data preprocessing, handling of class imbalance, and evaluation with the appropriate metrics.

Most of the existing works have emphasized model experimentation but have not sufficiently addressed interpretability, this gap limits their applicability in real-world business contexts. Furthermore, behavioral constructs from theoretical frameworks like the Technology

Acceptance Model (TAM) are underutilized, despite their strong association with the consumer features in the UCI dataset. Incorporating a theoretical framework could deepen the understanding of user behavior and enhance the relevance of model outputs for decision-making.

Addressing these gaps presents an opportunity for further research; by combining robust machine learning models, interpretable outputs and insights from consumer behavioral theory, future studies can advance the development of deep, intelligent systems capable of identifying high-intent online shoppers and guiding marketing strategies. Pursuing these efforts will not only show promise to improve conversion rates but also contribute to creating a more efficient and personalized shopping environment.

4.5 TABLE OF LITERATURE

SI. NO.	SOURCE	HOW ARE THE KEY VARIABLES DEFINED AND MEASURED?	METHODOLOGY	ASSESSMENT MEASURES	IMPORTANT FINDINGS
1.	Adhikari (2023)	Target Variable: Revenue Independent Variable: PageValues, Month, OperatingSystems, Browser, Region, TrafficType, VisitorType, Weekend	Applied ensemble models (XGBoost, LightGBM, Random Forest) to the UCI dataset. SMOTE was used to balance classes, followed by grid search tuning.	Accuracy, precision, recall, and F1-score were used for model evaluation.	XGBoost achieved 93.54% accuracy, outperforming other models. PageValue and ProductRelatedDuration ranked highest in feature importance.
2.	Ahsain & Ait Kbir (2022)	Target**:** Revenue Independent**:** Administrative_Duration, Informational_Duration, BounceRates, ExitRates, Page value, product-related page views, exit rate, bounce rate, special day indicators	Used the UCI Online Shoppers Purchasing Intention Dataset; applied Decision Tree, Random Forest, and Gradient Boosting models via PyCaret.	Accuracy, precision, recall, F1-score, and area under the curve (AUC).	Random Forest achieved 91.3% accuracy and Gradient Boosting reached 90.4%. Product-related duration and page value were the strongest predictors.
3.	(Andrina, D., et al., 2022)	Independent variables: Perceived usefulness, perceived ease of use, trust. Dependent variable: Purchase intention. Variables measured using Likert scale survey items.	Online survey of consumers; Structural Equation Modeling (SEM).	Path coefficients, R², and p-values.	Perceived usefulness and trust strongly influence purchase intention; TAM constructs are valid in e-commerce.
4.	Baati and Mohsil (2020)	Target Variable: “Revenue” Class label indicating whether the visit has been finalized with a transaction Important Variable Numerical Features: Day	The study investigates Naïve Bayes classifier, C4.5 decision tree, and Random Forest. Evaluation metrics	The dataset from the UCI Machine Learning Repository used for the research	The final model presented in the research paper identifies Random Forest with a significantly higher accuracy and F1 Score

		Categorical Features: Operating Systems, Browser, Region, Traffic, Visitor, Month, Weekend	used include F1 score, accuracy	has 12,330 records and 18 attributes. Additionally, 30% of the data set consisting of 12330 samples is first excluded for testing and the oversampling SMOTE method is applied to the remaining 70% of the samples.	than Naïve Bayes and C4.5 decision tree classifiers. The study uses “Synthetic Minority Oversampling Technique” (SMOTE) methodology to improve the performance and the scalability of each classifier. Random Forest has the highest accuracy of 86.78% and F1 Score of 0.60 is obtained with the random forest classifier.
5.	Chatterjee & Kumar Kar (2020)	Dependent: Cart abandonment; Independent: Page load time, payment steps, product type, user intent.	Survey-based behavioral analysis and logistic regression modeling; e-retail platform case study.	Logistic regression coefficients, p-values, and behavioral outcome modeling.	Complex checkout process and slow load time increased abandonment; trust and UX were key mediators.
6.	Davis (1989)	Independent: Perceived usefulness, perceived ease of use. Dependent: Actual system use. Measured via self-report questionnaires.	Survey-based quantitative analysis; regression and factor analysis.	Regression weights, variance explained.	TAM introduced; perceived usefulness had greater influence on technology adoption than ease of use.
7.	Frazier et al. (2022)	Dependent: Purchase; Independent: Navigation patterns, session duration, visitor type.	Used UCI dataset; XGBoost, SVM, and Decision Trees; SMOTE for class imbalance.	Accuracy, precision, recall, F1-score, AUC-ROC	XGBoost had top performance; class balancing improved precision/recall significantly.

8.	Islam et al. (2023)	Dependent: Buying intention; Independent: User engagement metrics.	Behavioral modeling with SVM, Naive Bayes; evaluated accuracy and F1-score.	Accuracy and F1-score.	Recommended SVM and Naive Bayes for performance; emphasized minimizing false positives.
9.	Jiang Yuwei (2022)	Target Variable: Repurchase (1 = repurchase, 0 = not). Independent Variables: 50 Variables from user, merchant, and user-merchant interaction (click count, age, buy/click ratio etc.)	Real-world Tmall data. Used random under sampling for class balance. Machine Learning Models: Logistic Regression, KNN, Random Forest, and XGBoost. Integrated Soft-Voting and Stacking fusion methods	AUC (Area Under Curve) is used for model evaluation.	XGBoost was Best model. Combining models using weighted Soft-Voting and Stacking made the predictions more accurate, increasing the AUC score by 0.2% to 4% compared to using single models like Logistic Regression, KNN, Random Forest, and XGBoost.
10.	Ketipov et al. (2023)	Target: Purchase behavior Independent: Personality traits, bounce rate, session count, dwell time, exit rate	Survey data collected from 226 respondents in 10+ countries. Applied Random Forest and Decision Tree models; TPOT used for automated optimization.	Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE).	Random Forest reduced MAPE from 72.48% to 72.79%. Product reviews, delivery options, and payment security were leading factors in decision behavior.
11.	Kurniawan et al. (2020)	Target Variable: Online shoppers purchasing intention prediction. Important Variable: The document focuses on improving the classification of imbalanced data to predict the target	The dataset used for evaluation was obtained from the UCI repository. Analytical Methods:	The study proposes a model that integrates feature selection	The proposed method, which combines data level approaches and feature selection techniques, can

		variable. It explores the use of data level approaches and feature selection techniques to achieve this.	10-fold cross-validation Confusion Matrix AUC (Area Under the Curve), f-measure Accuracy, Sensitivity, Precision, Specificity	using Particle Swarm Optimization (PSO) with data level approaches like Random Under-Sampling (RUS) and SMOTE, and the AdaBoost algorithm.	effectively address the challenges of imbalanced data in predicting online shoppers' purchasing intentions. The combination of SMOTE and AdaBoost with classification algorithms showed improved performance compared to using single classifiers. The Random Forest classification algorithm outperformed other algorithms. The SMOTE + AdaBoost + Classification Algorithm model was identified as the final model which
12.	Purwianti et al., (2024)	Independent: Perceived usefulness, perceived ease of use, risk factors. Dependent: Online purchase intention. Variables based on TAM and measured through structured survey responses.	Survey data collected from online shoppers; Confirmatory Factor Analysis and SEM used.	Goodness-of-fit metrics, path coefficients.	TAM variables (usefulness, ease) and perceived risk significantly impact online shopping intention.
13.	Rajamma et al. (2009)	Key constructs: Cart abandonment behavior influenced by perceived waiting time, transaction inconvenience, and risk.	Empirical study using survey responses analyzed	Regression coefficients, ANOVA,	Transaction inconvenience and waiting time significantly influence

		Measured via shopper surveys and Likert scales.	via regression modeling.	significance levels.	cart abandonment. Trust and ease of checkout are key to reducing drop-off.
14.	Rana et al. (2023)	Dependent: Purchase intent; Independent: Admin/product/revenue variables from UCI dataset.	Applied transformations; used LR, SVM, RF; tested impact on model performance.	Accuracy, precision, recall, F1-score (transformation impact discussed in assessment section).	Data transformation greatly improved accuracy: RF gave the best performance overall.
15.	Sakar et al. (2019)	Dependent: Purchase (Binary); Independent: Bounce Rates, Exit Rates, Page Value, Duration, etc.	Used UCI dataset; MLP and LSTM models; evaluation via accuracy, precision, recall, and F1-score.	Accuracy, precision, recall, F1-score.	LSTM outperformed MLP; behavioral metrics like Page Values and Exit Rates were critical predictors.
16.	Satu and Islam (2023)	Target Variable: “Revenue” variable indicating whether a visit to an e-commerce site resulted in a transaction Important Variable: product quality, price, availability of products, special discounts, visitor types, reliable online services, and engagement on social media Other important variables: Administrative, Administrative duration, Informational, Informational duration, Product related, Product related duration, Bounce rate, Exit rate, Page value, Special day, Month, Operating systems, Browser, Region, Traffic type, Visitor type, Weekend	Data Collection: The Online Shoppers' Purchasing Intention Dataset was collected from the UCI Machine Learning Repository. Analytical Methods: The study uses various machine learning techniques, including data transformation, data balancing, outlier detection, feature	The study proposes a machine-learning model that uses multiple data analytics and machine learning techniques to predict customer buying intentions. The methodology involves data	Random Forest (RF) was found to be the most stable classifier for predicting customer purchase intention. The Random Forest classifier provided the best accuracy of 92.39% and an f-score of 0.924 for the Z-Score and Gain Ratio transformed subset. Z-Score transformation and Information Gain were identified as reliable methods for

			selection, and classification algorithms	collection, feature transformation, balancing, outlier detection, feature selection, and classification.	processing the online shoppers' customer intention dataset
17.	Shi, Xiang (2021)	Target Variables: Revenue (Binary:1 = purchase, 0 = no purchase) Key Independent Variables: Time spent on site, Page Value, Bounce Rate, Exit Rate.	Used Online Shoppers Purchasing Intention Dataset from UCI Repository. Applied descriptive statistics to identify correlations, followed by Machine Learning Models: Logistic Regression, Decision Tree, And Random Forest.	Accuracy was an evaluation metric. The study compares the performance of different models to see which one predicts purchases most accurately.	Random Forest was the best model. Time and Page values had positive correlation and bounce and exit rates had negative correlation with purchasing intent.
18.	Song & Liu, (2020)	Target Variable: Revenue (Binary:1 = purchase, 0 = no purchase) Key Independent Variables: Administrative, Product Related, Product Related Duration, Bounce Rates, Exit Rates, Page Values, Month_Category_ Nov,	Dataset was collected from the UCI Machine Learning Repository. Data was cleaned and modeled using	Evaluation Metrics: Accuracy, Precision, recall, F1 score and AUC	XGBoost was best Model with accuracy = 90.15%, AUC = 0.7744. ExitRates, BounceRate, and VisitorType were key predictors.

		VisitorType_Category_Returning_Visitor, VisitorType_Category_New_Visitor	XGBoost and Random Forest. Python libraries were used.		
19.	Sunarto et al. (2023)	Target Variable: “Revenue” attribute to indicate whether a visit ended with a transaction or not Important Variable Numerical Features: Administrative, Administrative duration, Informational, Informational duration, Product related duration, Bounce rate, Exit rate, Page value, Special day Categorical Features: Operating Systems, Browser, Region, Traffic Type, Visitor Type, Weekend, Month	C4.5 algorithm for classification. Boosting technique (Adaboost) to reduce classification error and handle class imbalance. Evaluation metrics: Accuracy, Error, and Area Under the ROC Curve (AUC)	The research used the boosting technique with the C4.5 algorithm. The dataset from the UCI Machine Learning Repository used for the research has 12,330 records and 18 attributes. The dataset was split into training and testing sets with ratios of 90:10 and 80:20 The Adaboost method was used as a boosting technique. The Online Shoppers Purchasing	The final model presented in the research paper is the C4.5 algorithm enhanced with the Adaboost boosting technique. The boosting technique with the C4.5 algorithm outperforms all other models in terms of improved accuracy and reduction classification errors. The average increase in accuracy when using the boosting technique with C4.5 is 5.76 percent. The boosting technique also improves the AUC (Area Under the Curve) value.

				Intention dataset from the UCI Machine Learning Repository was used.	
20	Tayal & Daniel (2024)	Target Variable: Purchase Intention categorized in three classes (Positive, Neutral and Negative) based on Likert scale. Independent Variables: Perceived risks (Delivery, Financial, Health, Time-Loss, Cultural) and Demographic Factor (Age, Gender, Education Level, Monthly income and Occupation)	Survey Collected from 308 online shoppers using a Google Form. Machine Learning Model: CatBoost, Random Forest, Gradient Boost, Logistic Regression, SVM, XGBoost, Decision Tree, KNN, AdaBoost, Naive Baves	Evaluation Metrics: Accuracy, Precision, Recall, F1-Score. They managed Class imbalance during evaluation. CatBoost achieved highest accuracy and recall (74.19%).	Delivery risk had the biggest negative impact on purchase intent while Occupation was the most important factor to purchase intent. CatBoost was Best Model.
21.	Torres Trevino & Cepeda (2024)	Dependent: Purchase (binary); Independent: TimeOnPage, PageValue, ExitRates, BounceRates.	UCI dataset: Logistic Regression and Decision Trees used; metrics included recall and accuracy.	Accuracy, recall, cost-sensitive learning metrics.	PageValue and ExitRates were key features; cost-sensitive adjustments improved recall.
22.	Wen, Lin, & Liu (2023)	Target: Purchase intention Independent: Clickstream sequence, bounce rate, time on site, entry flow, page depth	Analyzed anonymized e-commerce session data using Logistic Regression, SVM, Random Forest, and	Accuracy, precision, recall, F1-score, and feature importance rankings.	XGBoost achieved 88.2% accuracy; Random Forest followed with 85.7%. Time-on-page and visit sequence

			XGBoost. Focused on behavioral signals.		were the most influential predictors.

Table 4.5.1: *Table of relevant literature*

4.6 CONCEPT MATRIX

The following tables indicate whether the key components of the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework were addressed in the reviewed literature. A checkmark denotes the inclusion of specific aspects, such as summary statistics or use of survey data, while a blank space indicates their absence. This refined approach ensures clarity when comparing studies that utilize the Online Shoppers' Purchasing Intention Dataset for predicting consumer behavior.

SI NO	AUTHORS	YEAR OF DATA COLLECTION	USED UCI DATASET	USED OTHER WEBSITE BEHAVIORAL DATA	SURVEY DATA	LONGITUDINAL DATA	SUMMARY STATS
1.	Adhikari	2023	✓				✓
2.	Ahsain & Ait Kbir	2022	✓				
3.	Andrina, D., et al.	2022			✓		
4.	Baati and Mohsil	2020	✓				✓
5.	Chatterjee & Kumar Kar	2020		✓			
6.	Davis	1989			✓		
7.	Frazier et al.	2022	✓				
8.	Islam et al.	2023	✓				
9.	Jiang Yuwei	2022		✓			✓
10.	Ketipov et al.	2023			✓		
11.	Kurniawan et al.	2020	✓				

12.	Purwianti et al.	2024			✓		
13.	Rajamma et al.	2009			✓		
14.	Rana et al.	2023	✓				
15.	Sakar et al.	2019	✓				
16.	Satu and Islam	2023	✓				✓
17.	Shi, Xiang	2021	✓				✓
18.	Song & Liu	2020		✓			✓
19.	Sunarto et al.	2023	✓				✓
20.	Tayal & Daniel	2024			✓		✓
21.	Torres Trevino & Cepeda	2024	✓				
22.	Wen, Lin, & Liu	2023		✓			

Table 4.6.1: *Concept Matrix for relevant literature*

4.7 RELEVANT WORKS – DATA PREPROCESSING

SI NO	AUTHORS	BALANCED DATASET	IMBALANCED DATASET	VARIABLE SELECTION	DATA CLEANING	COMBINING DATASETS	EXPLORATORY ANALYSIS	BINNING
1.	Adhikari		✓	✓	✓		✓	
2.	Ahsain & Ait Kbir		✓	✓	✓		✓	

3.	Andrina, D., et al.							
4.	Baati and Mohsil		✓	✓	✓		✓	✓
5.	Chatterjee & Kumar Kar							
6.	Davis							
7.	Frazier et al.		✓	✓				
8.	Islam et al.		✓	✓	✓		✓	
9.	Jiang Yuwei	✓	✓	✓	✓	✓	✓	
10.	Keptipov							
11.	Kurniawan et al.	✓	✓	✓	✓			
12.	Purwianti et al.							
13.	Rajamma et al.							
14.	Rana et al.		✓	✓	✓		✓	
15.	Sakar et al.		✓	✓	✓		✓	
16.	Satu and Islam		✓	✓	✓		✓	✓
17.	Shi, Xiang		✓				✓	
18.	Song & Liu		✓	✓	✓		✓	
19.	Sunarto et al.		✓	✓	✓		✓	
20.	Tayal & Daniel		✓		✓			✓

21.	Torres Trevino & Cepeda		✓	✓	✓		✓	
22.	Wen, Lin, & Liu		✓	✓	✓		✓	

Table 4.7.1: *Data preprocessing methods among relevant literature*

4.8 RELEVANT WORKS – MODELING

SI NO	AUTHORS	FINAL MODEL	LOGISTIC REGRESSION	DECISION TREE	RANDOM FOREST	SVM	LASSO	NN	BAYESIAN NETWORK
1.	Adhikari	Random Forest		✓	✓			✓	✓
2.	Ahsain & Ait Kbir	XGBoost			✓				
3.	Andrina, D., et al.								
4.	Baati and Mohsil	Random Forest		✓	✓				✓
5.	Chatterjee & Kumar Kar	Logistic Regression	✓						
6.	Davis								
7.	Frazier et al.	XGBoost		✓		✓			
8.	Islam et al.	Naive Bayes				✓			✓
9.	Jiang Yuwei		✓		✓				
10.	Kurniawan et al.	SMOTE + AdaBoost + Classification Algorithm		✓		✓		✓	
11.	Ketipov et al.	Random Forest		✓	✓				

12.	Purwianti et al.								
13.	Rajamma et al.								
14.	Rana et al.		✓		✓	✓			
15.	Sakar et al.	LSTM						✓	
16.	Satu and Islam	Random Forest	✓	✓	✓	✓		✓	✓
17.	Shi, Xiang	Random Forest	✓	✓	✓				
18.	Song & Liu	XGBoost		✓	✓				
19.	Sunarto et al.	C4.5 algorithm Boosting technique (Adaboost)		✓					
20.	Tayal & Daniel	CatBoost	✓	✓	✓	✓			
21.	Torres Trevino & Cepeda	Cost-Sensitive RF	✓	✓	✓	✓		✓	
22.	Wen, Lin, & Liu	XGBoost	✓	✓	✓	✓			

Table 4.8.1: *Algorithm & model selection among relevant literature*

4.9 RELEVANT WORKS – ASSESSMENT METHODS

SI NO	AUTHORS	RMSE	ROC/AUC	ODDS RATIO	F1 SCORE	ACCURACY	RECALL	PRECISION	SPECIFICITY
1.	Adhikari		✓		✓	✓	✓	✓	
2.	Ahsain & Ait Kbir				✓	✓			
3.	Andrina, D., et al.								
4.	Baati and Mohsil		✓		✓	✓	✓	✓	
5.	Chatterjee & Kumar Kar			✓					
6.	Davis								
7.	Frazier et al.		✓		✓	✓	✓	✓	
8.	Islam et al.		✓		✓	✓	✓	✓	
9.	Jiang Yuwei				✓	✓	✓	✓	
10.	Ketipov et al.	✓							
11.	Kurniawan et al.		✓		✓	✓	✓	✓	✓
12.	Purwianti et al.								
13.	Rajamma et al.	✓							
14.	Rana et al.				✓	✓	✓	✓	
15.	Sakar et al.				✓	✓	✓	✓	

16.	Satu and Islam		✓		✓	✓	✓	✓	
17.	Shi, Xiang					✓			
18.	Song & Liu		✓		✓	✓	✓	✓	
19.	Sunarto et al.		✓			✓			
20.	Tayal & Daniel				✓	✓	✓	✓	
21.	Torres Trevino & Cepeda					✓	✓		
22.	Wen, Lin, & Liu				✓	✓	✓	✓	

Table 4.9.1: *Assessment measures used among relevant literature.*

5. DATA UNDERSTANDING

Effectively predicting online shopping purchases requires a strong understanding of user behavior, proper handling of class imbalance, and a thorough approach to data preprocessing to ensure model performance and interpretability. The UCI Online Shoppers Purchasing Intention dataset contains user browsing sessions collected over a one-year period. Online behavioral and technical features describe each session and are accompanied by a binary target variable indicating whether a purchase was made. The dataset includes three broad feature categories: Administrative & Informational (Administrative, Information, Product Related, and their corresponding durations), Behavioral Metrics (Bounce Rates, Exit Rates, Page Values), Technical/User Attributes (Operating Systems, Browser, Visitor Type). These features have been shown to capture important behavioral patterns predictive of purchases. For example, Sakar et al. (2019) and Frazier et al. (2022) highlight Page values, Exit Rates, and the duration of time on a page as important drivers of purchase decisions. In studies that used datasets like the UCI dataset, such as Chatterjee & Kumar Kar (2020), it was documented that cart abandonment and session outcomes can often be inferred through the examination of behavioral feature paths such as repeated product views or elevated bounce rates.

5.1 DATA SOURCE

The dataset from the UCI Machine Learning Repository is originally derived from a real e-commerce website. Specifically, it is from Columbia Sportswear's Turkish website, representing e-commerce activity in Turkey and nearby regions. The dataset was donated to UCI in 2018 by its creators, C. Okan Sakar and Yomi Kastro, who obtained the raw data from collaborating with an e-commerce analytics firm.

All user interactions were tracked from the website's analytics system, which relied on the Google Analytics tracking code embedded in it. It was composed of actual web traffic logs and analytics data, where each user browsing the Columbia Sportswear site was recorded in real-time (Sakar et al., 2019).

6. DATA DESCRIPTION

The dataset includes 12,330 observations and 18 features, providing detailed insights into user behavior, such as browsing activity, time spent on different pages, traffic sources, and purchasing intent. The data dictionary defined below provides information on the distinctive features and the role they play in our research.

VARIABLE	DESCRIPTION	RATIONALE	ROLE	MEASUREMENT	SELECT	TAM CONSTRUCT	FEATURE TYPE
ADMINISTRATIVE	Number of administrative pages visited.	Indicates engagement with non-product pages.	Input	Interval	yes	Ease of Use	Behavioral Feature
ADMINISTRATIVE_ DURATION	Time (in seconds) spent on administrative pages.	Reflects depth of user interaction with admin pages.	Input	Interval	yes	Ease of Use	Behavioral Feature
INFORMATIONAL	Number of informational pages visited.	Measures user interest in learning more about the site.	Input	Interval	yes	Usefulness	Behavioral Feature
INFORMATIONAL_ DURATION	Time (in seconds) spent on informational pages.	Captures the extent of engagement with informational content.	Input	Interval	yes	Usefulness	Behavioral Feature
PRODUCTRELATED	Number of product-related pages visited.	A key metric for user intent towards purchasing.	Input	Interval	yes	Usefulness	Behavioral Feature
PRODUCTRELATED_ DURATION	Time (in seconds) spent on product-related pages.	Measures potential purchase intent through time spent.	Input	Interval	yes	Usefulness	Behavioral Feature
BOUNCERATES	Metric measures how often a user visits a page and exits quickly.	Helps assess user engagement level.	Input	Interval	yes	Ease of Use	Behavioral/ Technical Feature

EXITRATES	Percentage of site visits that end on the page exited from.	Indicates the likelihood of user drop-off.	Input	Interval	yes	Ease of Use	Behavioral/ Technical Feature
PAGEVALUES	Average value of a webpage based on conversion probability.	Measures importance of pages in conversion paths.	Input	Interval	yes	Usefulness	Behavioral/ Technical Feature
SPECIALDAY	Proximity of visit to a special day (e.g., holidays).	Could influence purchase behavior.	Input	Nominal	yes	N/A	Temporal Feature
MONTH	Month in which the user visited the website.	Captures seasonal effects on purchasing behavior.	Input	Nominal	yes	N/A	Temporal Feature
OPERATINGSYSTEMS	OS of the user visiting the website.	May impact user experience and conversion rates.	Input	Nominal	yes	Ease of Use	Technical Feature
BROWSER	Browser used by the visitor.	Helps in understanding platform compatibility issues.	Input	Nominal	yes	Ease of Use	Technical Feature
REGION	Geographic region of the user.	Useful for location-based marketing and analysis.	Input	Nominal	yes	N/A	Demographic Feature
TRAFFICTYPE	Source of visitor traffic to the website.	Helps analyze the impact of different traffic sources.	Input	Nominal	yes	Usefulness	Technical/ Behavioral Feature
VISITORTYPE	Type of visitor: "New Visitor,"	Important for identifying user retention trends.	Input	Nominal	yes	Usefulness	Behavioral Feature

	"Returning Visitor," etc.						
WEEKEND	Indicate if the visit was on a weekend.	Could impact shopping behavior.	Input	Binary	yes	N/A	Temporal Feature
REVENUE	Indicates if the visit resulted in a transaction.	Target variable for conversion prediction.	**Target**	**Binary**	**yes**	**N/A**	**Outcome Feature**

Table 6.1.1*: Online Shoppers Purchasing Intention Data Dictionary*

7. EXPLORATORY DATA ANALYSIS

7.1 SUMMARY STATISTICS

Based on the Online Shoppers Purchasing Intention Dataset, the data was collected through an E-commerce website over a year period. To begin our exploratory analysis, we will focus on generating and interpreting summary statistics based on the dataset. Undoubtedly, summary statistics provide a foundational understanding of the data by highlighting key measures such as the mean, median, standard deviation, minimum, and maximum values for each variable. By examining these descriptive metrics, we can gain valuable insights into the distribution, central tendencies, and variability of the data, which will help inform our subsequent analysis and modeling efforts.

VARIABLE	MEAN	STD DEV	MIN	MAX	N	SKEWNESS	KURTOSIS
ADMINISTRATIVE	2.315	3.322	0	27.000	12333	1.960	4.701
ADMINISTRATIVE_ DURATION	80.819	176.779	0	3398.750	12333	5.616	50.557
INFORMATIONAL	0.504	1.270	0	24.000	12333	4.036	26.932
INFORMATIONAL_ DURATION	34.472	140.749	0	2549.380	12333	7.579	76.317
PRODUCTRELATED	31.731	44.476	0	705.000	12333	4.342	31.212
PRODUCTRELATED_ DURATION	1194.750	1913.670	0	63973.520	12333	7.263	137.174
BOUNCERATES	0.022	0.048	0	0.200	12333	2.948	7.723
EXITRATES	0.043	0.049	0	0.200	12333	2.149	4.017
PAGEVALUES	5.889	18.568	0	361.764	12333	6.383	65.636

Table 7.1.1: *Dataset Overall Summary Statistics*

The summary statistics reveal that most interval variables in the dataset exhibit positive skewness and high kurtosis, indicating non-normal distributions. Variables such as Administrative, Informational, ProductRelated, and their durations show that most sessions lie within lower values, with outliers occurring where users spend significantly more time. For instance, Informational_Duration and ProductRelated_Duration have extreme right-skewness and kurtosis, reflecting unusually long sessions for a small portion of users. Similarly, PageValues shows that while most visits had a minor impact, a small portion of users contributed significantly to transaction value. BounceRates and ExitRates are also positively skewed, though less extreme, due to their ranges lying between 0 and 0.2. Likewise, Informational and ProductRelated metrics reveal significant variability in user behavior. While the average number of informational pages viewed is just 0.5, some users visited up to 24. Similarly, product-related page views average 31.7 but reach as high as 705. Bounce and exit rates remain low and consistent, with means around 0.02–0.04, suggesting most users have limited engagement with the site before leaving. The PageValues variable is particularly notable, with an average of 5.89 and a maximum of 361.76, and a large standard deviation of 18.57 indicating the impact of a few high-value sessions. These insights emphasize the importance of data transformation and robust modeling to manage the broad range of user behavior.

Furthermore, with Revenue defined as our target variable, both values of the variable are used to provide an in-depth view of those interval variables as displayed in those two tables below.

REVENUE	N	VARIABLE	MEAN	STD DEV	MIN	MAX	SKEWNESS	KURTOSIS
FALSE	10422	**Administrative**	2.118	3.202	0	27.000	2.095	5.484
		Administrative_Duration	73.740	171.018	0	3398.750	6.104	60.067
		Informational	0.452	1.212	0	24.000	4.450	33.945
		Informational_Duration	30.236	133.909	0	2549.380	8.433	94.382
		ProductRelated	28.715	40.745	0	705.000	4.694	38.839
		ProductRelated_Duration	1069.990	1803.800	0	63973.520	8.833	197.103
		BounceRates	0.025	0.052	0	0.200	2.664	5.973
		ExitRates	0.047	0.051	0	0.200	1.940	2.977
		PageValues	1.976	9.072	0	246.759	9.559	148.084

Table 7.1.2: *Dataset Overall Summary Statistics without Revenue*

REVENUE	N	VARIABLE	MEAN	STD DEV	MIN	MAX	SKEWNESS	KURTOSIS
TRUE	1908	**Administrative**	3.394	3.731	0	26.000	1.455	2.411
		Administrative_ Duration	119.483	201.115	0	2086.750	3.985	23.154
		Informational	0.786	1.521	0	12.000	2.699	9.290
		Informational_ Duration	57.611	171.619	0	1767.670	4.983	31.195
		ProductRelated	48.210	58.267	0	534.000	3.259	15.065
		ProductRelated _Duration	1876.21	2312.21	0	27009.86	3.338	17.725
		BounceRates	0.005	0.012	0	0.2	8.277	110.231
		ExitRates	0.020	0.016	0	0.2	3.478	25.938
		PageValues	27.265	35.192	0	361.764	3.252	17.921

Table 7.1.3: *Dataset Overall Summary Statistics with Revenue*

In the summary statistics tables above, in terms of user sessions, we are observing differences in the outcomes of the target variable. Indeed, user sessions that resulted in a purchase showed significantly higher engagement, with users visiting more pages and spending more time overall. This is reflected in the segmented summary statistics, where higher means were identified among purchasers for Administrative (3.39 vs. 2.12), ProductRelated (48.21 vs. 28.72), and PageValues (27.27 vs. 1.98). Notably, ProductRelated_Duration was much greater among purchasing users (1876.21 seconds) compared to non-purchasing ones (1069.99 seconds), indicating deeper browsing behavior. Besides, both purchasing and non-purchasing sessions show strong positive skewness and high kurtosis, indicating non-normal distributions with extreme outliers. For instance, PageValues is highly skewed in non-purchasing sessions (9.56) with extreme kurtosis (148.08) and remains notably skewed in purchasing sessions (3.25, kurtosis 17.92). Bounce rates are also much lower for purchasing users (0.005 vs. 0.025), supporting the link between engagement and conversions. These patterns highlight the need for data transformation and robust modeling techniques to manage the skewed and heavy-tailed distributions.

7.2 FREQUENCY DISTRIBUTIONS FOR CATEGORICAL FEATURES

The dataset shows a strong class imbalance in the target variable "Revenue". The number of online shopping sessions that did not result in purchases is much higher than the number of visits that resulted in a purchase. This suggests that the conversion rate (the rate at which visitors make a purchase) is low, referring to the 10,422 sessions where the Revenue variable is "False" (No Purchase). However, there was 1,908 sessions that did result in a transaction that were labeled as “True”.

This imbalance is an important characteristic of the dataset and will need to be considered throughout data preprocessing, model building, and assessment. This is especially important since we will be trying to maximize as many true purchases as possible. Weighing the options for handling imbalanced datasets will be discussed and evaluated in a later section of the report.

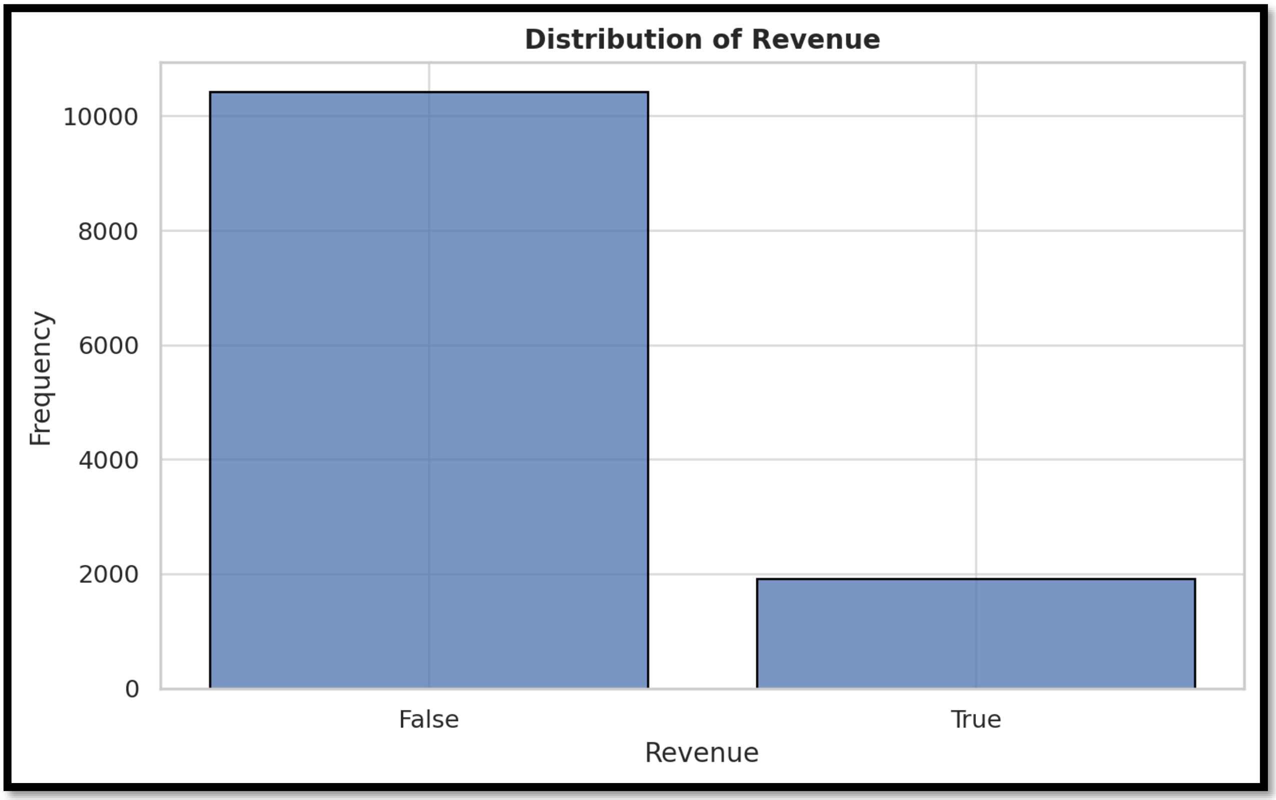

Figure 7.2.1: *Distribution of the target variable Revenue*

The number of returning visitors is the most frequent type of visitor and contributes the most to both revenue and non-revenue generating sessions, accounting for 10,551 out of the total observations. This group of customers have the highest number of purchases, and non-purchases, suggesting a large base with varying levels of purchase intent. New Visitors have approximately 1,694 visits with a relatively smaller number of visits resulting in either purchase or non-purchase.

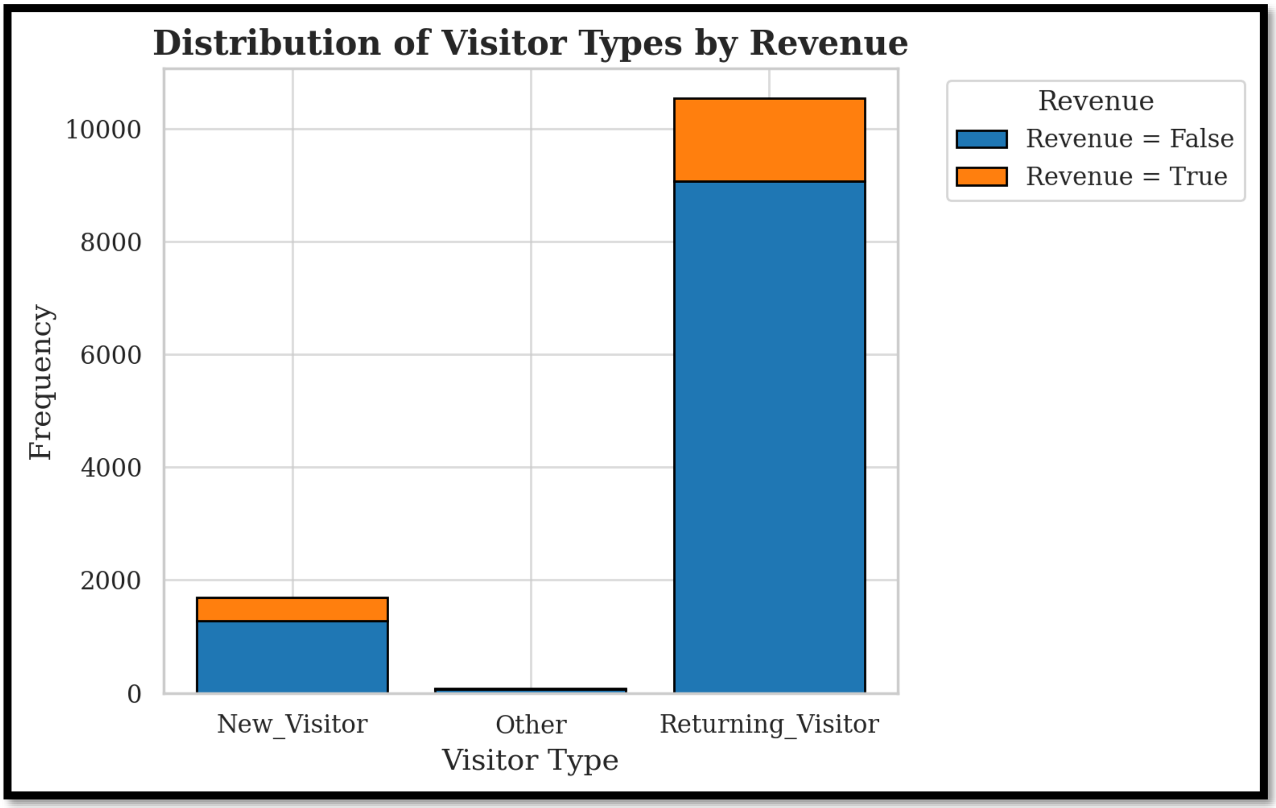

Figure 7.2.2: *Distribution of the target variable Revenue by Visitor type*

The dataset indicates that there is more website traffic and online shopping activity on weekdays, non-weekend visits (False), having a total number of 9,462 visits as compared to weekends which recorded a much lower number of visits. The number of weekday sessions is more than three times the number of weekend sessions. In comparing the number of sessions that generate revenue, though there are more sessions on weekdays, however the proportion of sessions that generate revenue on weekends may be slightly higher relative to total weekend sessions.

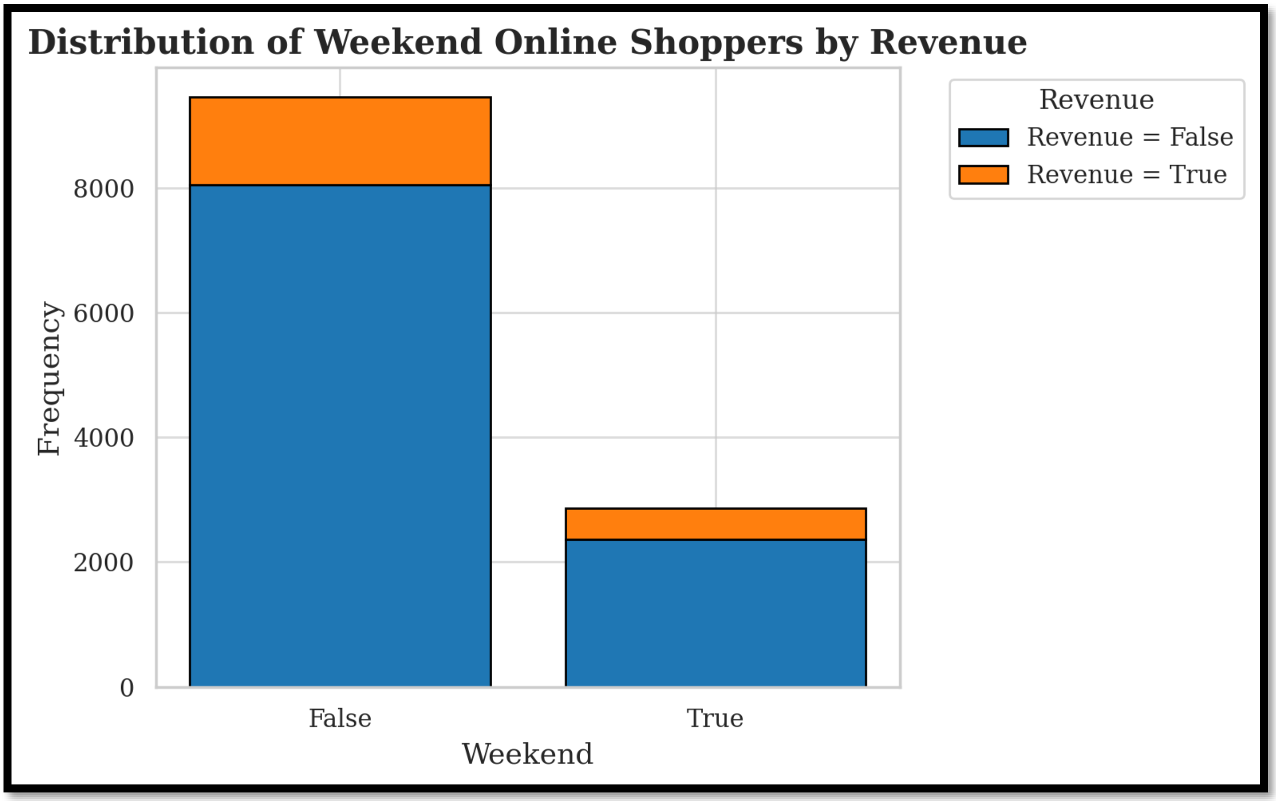

Figure 7.2.3: *Distribution of the target variable Revenue by Weekend*

In comparing month by month visits, May and November have the highest number of online shopper sessions, with 3364 and 2998 sessions, respectively. It is also important to note that these two months have had a high number of non-purchasing sessions and a noticeable level of purchasing sessions. The proportion of purchases is higher in November than May even though it has less sessions. Additionally, February has the least number of visits and a corresponding sparse number of both non-purchasing and purchasing sessions.

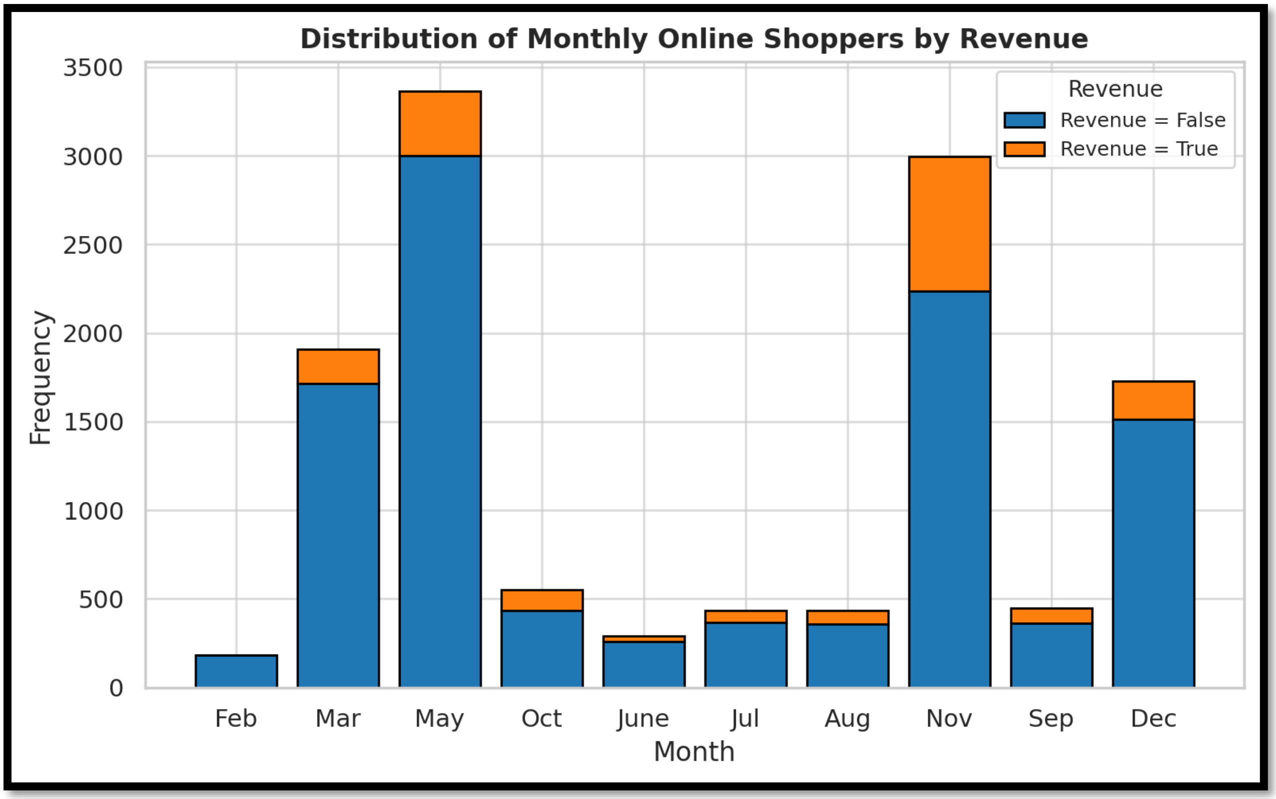

Figure 7.2.4: *Distribution of the target variable Revenue by Month*

7.3 INTERVAL FEATURE DISTRIBUTIONS

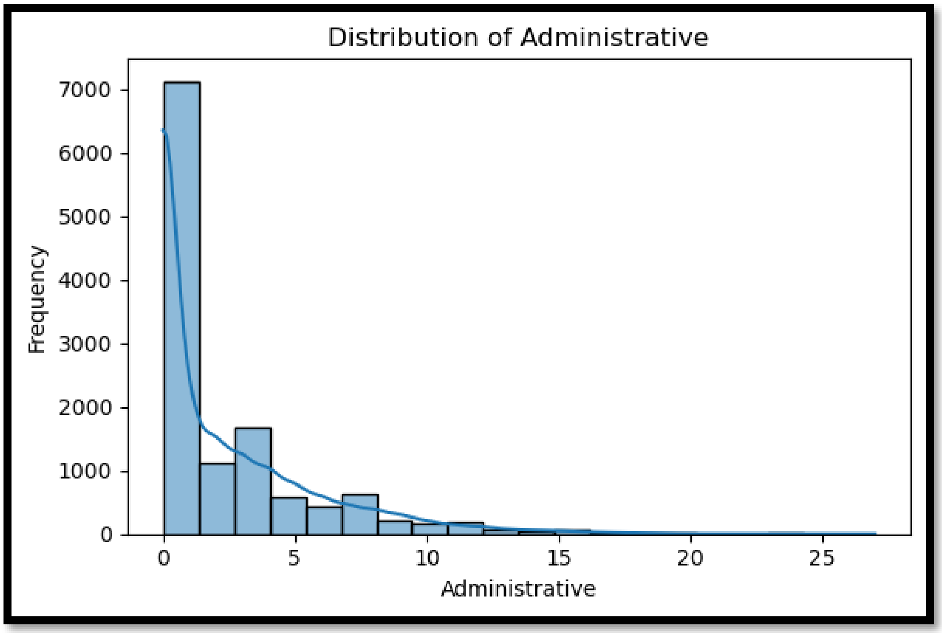

Figure 7.3.1: *Distribution of Administrative Page Visits*

The distribution of the Administrative variable, which reflects the number of administrative pages visited during a session, is highly skewed. As shown in Figure 5.3.1, most users viewed between 0 and 2 administrative pages, with the frequency sharply declining as the count increases.

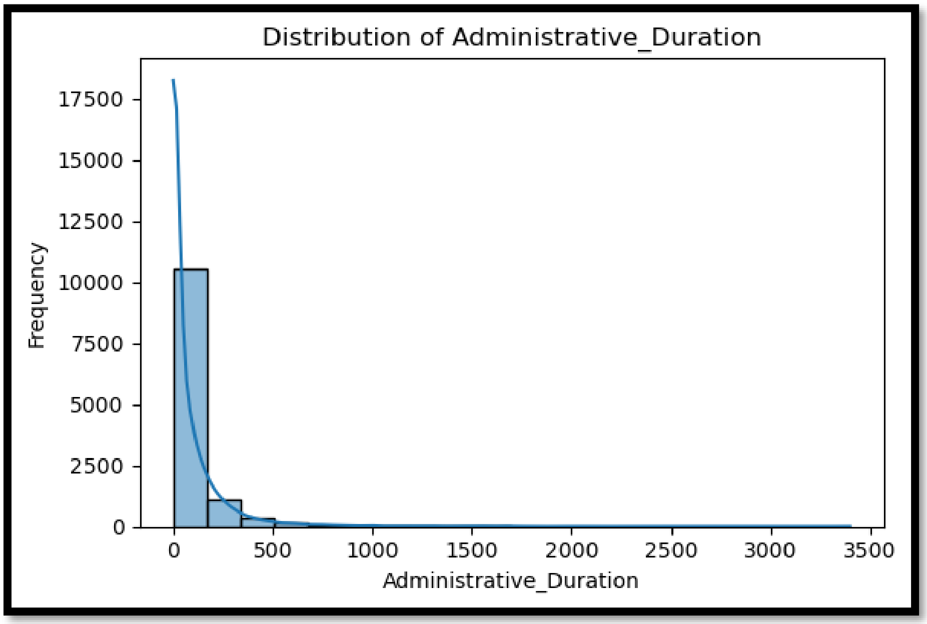

Figure 7.3.2: *Distribution of Administrative Page durations*

The Administrative_Duration variable, which captures the total time a user spent on administrative pages during a session, displays a highly right-skewed distribution. As shown in Figure 5.3.2, most visitors spent between 0 and 200 seconds on administrative pages and as time increases the number of visitors drops sharply.

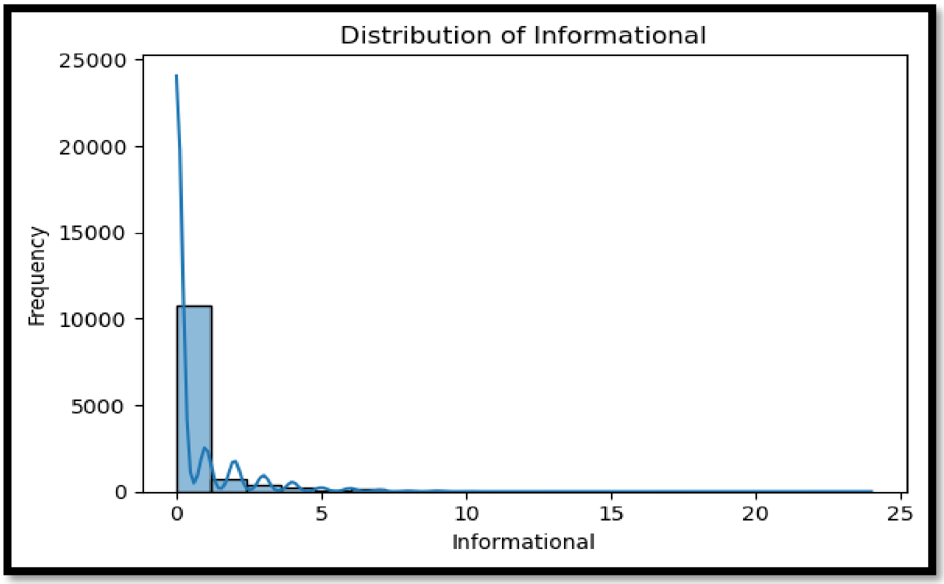

Figure 7.3.3: *Distribution of Informational page visits*

The Informational variable, representing the number of informational pages visited during a session, also displays a highly right-skewed distribution where most visitors viewed between 0 to 2 informational pages.

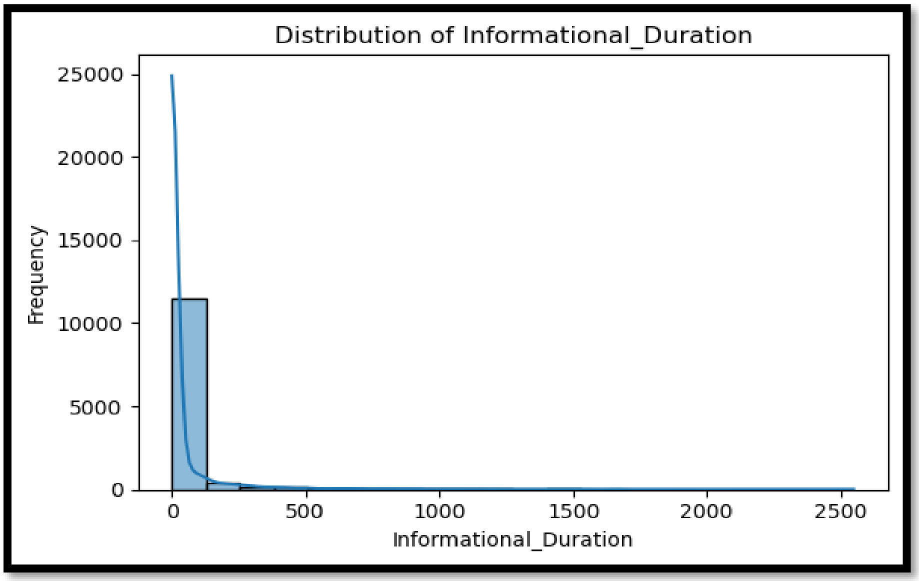

Figure 7.3.4: *Distribution of Informational page durations*

As seen in previous duration metrics, the Informational_Duration variable displays a highly right-skewed distribution, where most visitors did not spend an extended amount of time on informational pages. Interestingly, the average time spent on administrative pages is much higher than the time spent on informational pages.

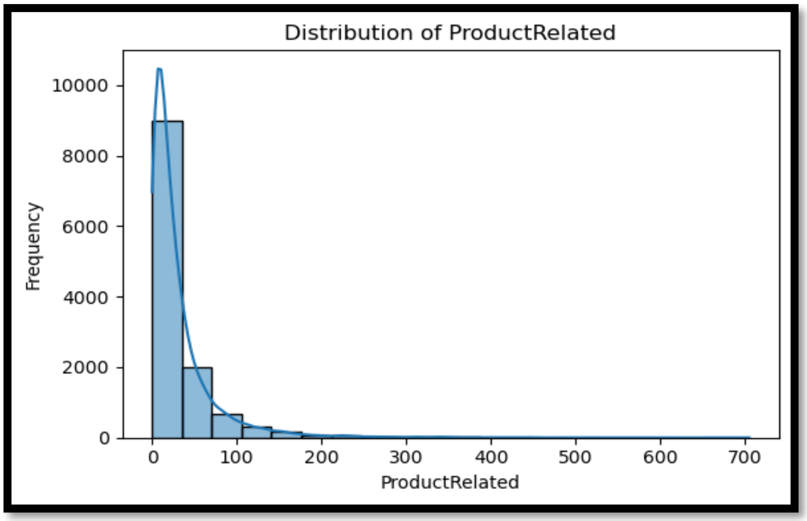

***Figure** 7.3.5: Distribution of Product Related page visits*

The ProductRelated variable, which measures the number of product related pages viewed during a session, displays a highly right-skewed distribution. Despite most users visiting under one hundred pages, there are numerous users with extremely high page visit values.

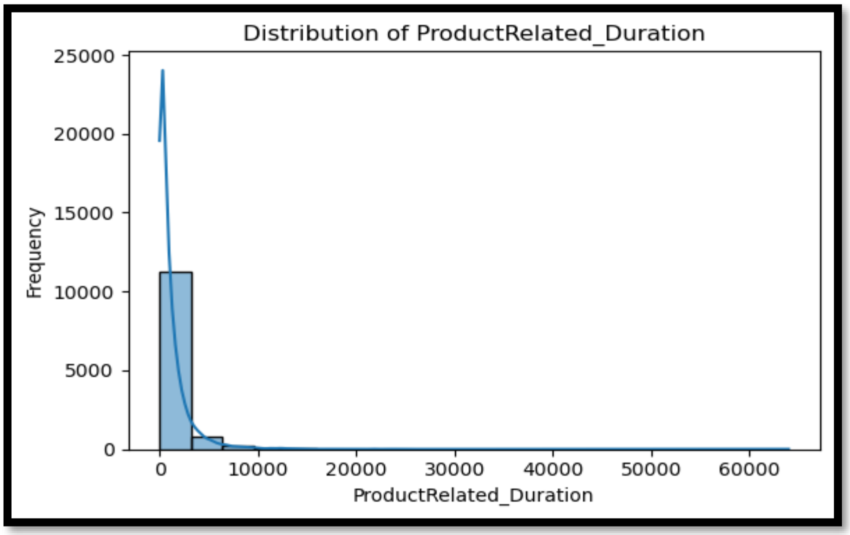

***Figure** 7.3.6: Distribution of Product Related page durations*

As seen in previous duration metrics, ProductRelated_Duration behaves in a very similar way, which is having a highly right skewed distribution with a long tail. Intuitively, Product related page durations are much higher on average than those of informational or administrative. This would align with existing beliefs about user traffic.

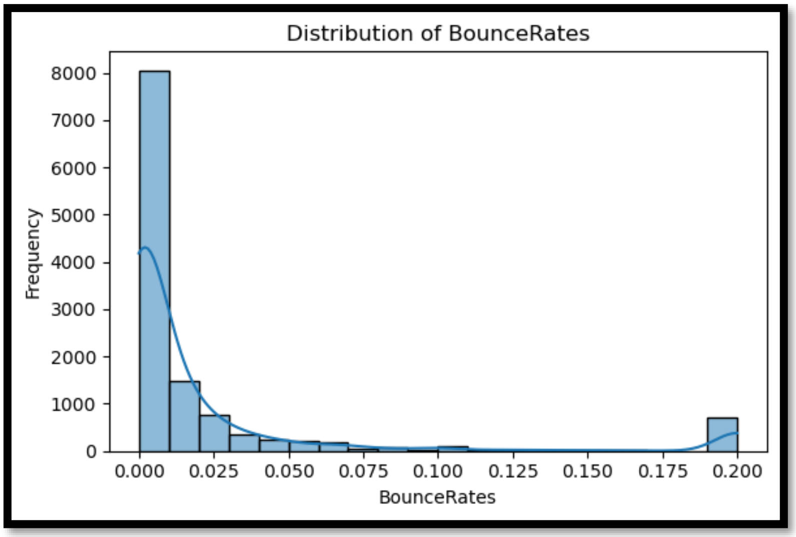

Figure 7.3.7: *Distribution of user Bounce Rates*

The BounceRates variable, which measures the proportion of single-page sessions within a browsing session, exhibits a strongly right-skewed distribution. Furthermore, most users recorded low bounce rates, indicating that they were engaged with multiple pages instead of exiting the site immediately.

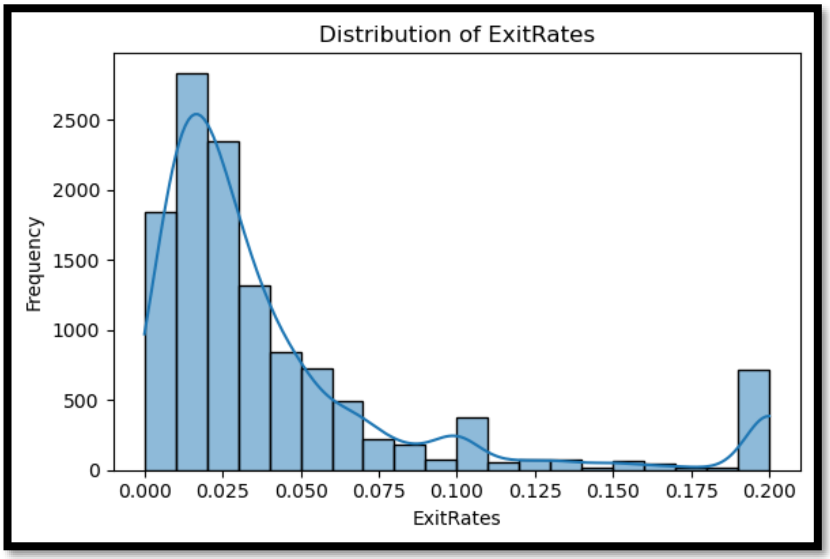

Figure 7.3.8: *Distribution of user Exit Rates*

The ExitRates variable, which measures the proportion of times a particular page was the last one viewed during a session, displayed a right-skewed distribution with a notable concentration of users at 0.20. In contrast to bounce rates, which reflect single-page sessions, exit rates display a broader spread in their distribution. This suggests that there is variability in which types of pages users tend to exit from.

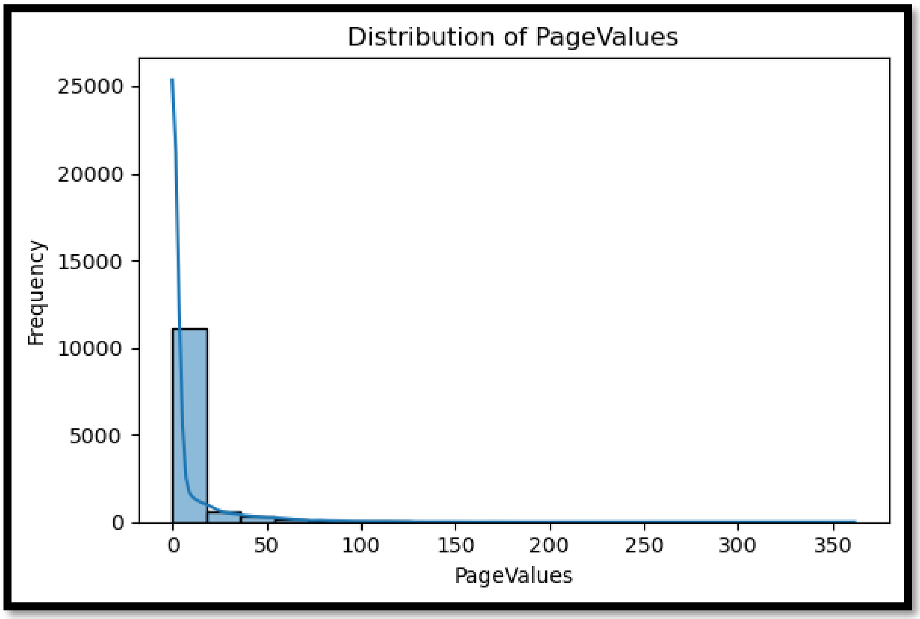

Figure 7.3.9: *Distribution of Page Values*

The PageValue variable represents the average historical page value that each user has visited in their session. Most sessions were associated with low or near zero-page values, indicating that most users interacted with pages that have poor conversion rates for purchases.

7.4 BOX PLOTS FOR INTERVAL FEATURES

To better understand the distribution and potential outliers in the dataset, box plots were generated for each of the key interval (continuous) features in the figures below.

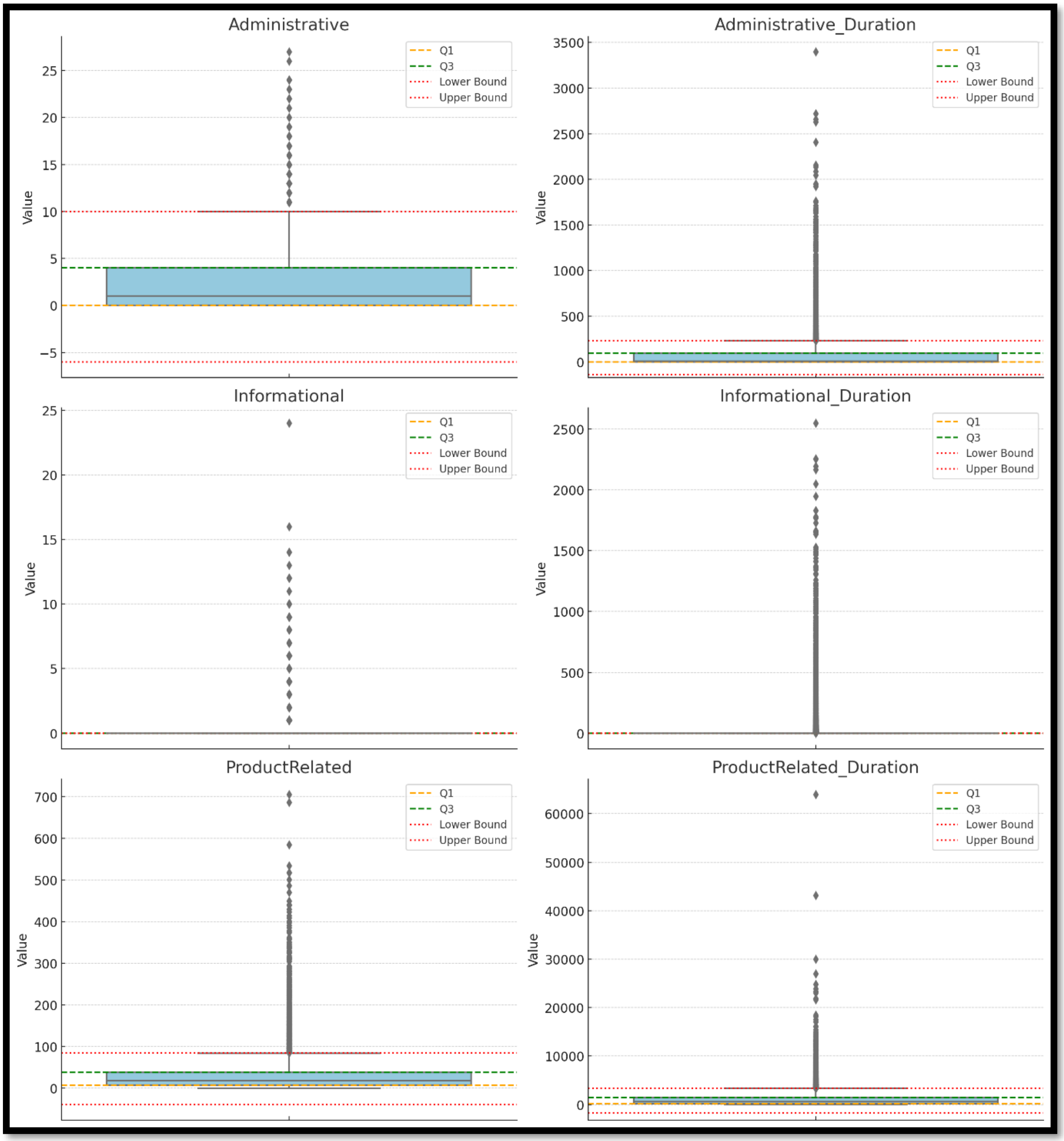

Figure 7.4.1: *Box Plots for Session Activity Features*

In Figure 7.4.1, there are box plots for six session activity features that help visualize the extreme values and potential outliers. All six features exhibit long right-tailed distributions and appear to have numerous extreme values. A method to account for these outliers will be applied for algorithms such as Logistic Regression, Support Vector Machines (SVM), and Neural Networks due to their sensitivity.

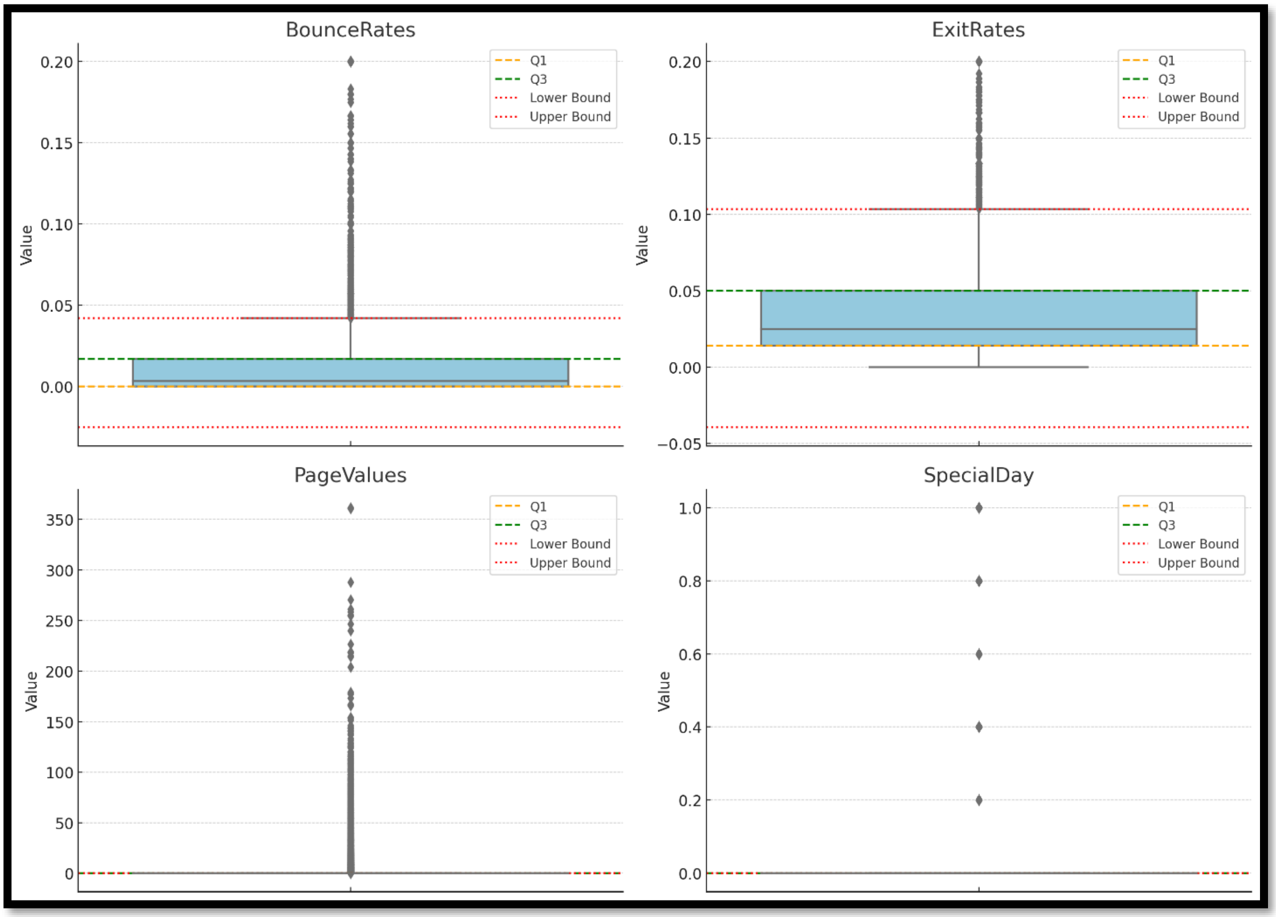

Figure 7.4.2: *Box Plots for Bounce Rates, Page Values, Exit Rates, and proximity to special days.*

In Figure 7.4.2, the SpecialDay feature is dominated by values near zero, indicating that the sparsity in the data may reduce its predictive power. Furthermore, it could benefit from being converted into a binary variable representing whether a session occurred in proximity to a holiday or other unique days of the year.

7.5 BIVARIATE ANALYSIS

The purpose of this section is to explore the relationship between individual features and the **Revenue** target variable. Indeed, by examining how each feature varies concerning purchase outcomes, we can identify potential patterns, trends, or associations that may influence user conversion. This analysis provides valuable insights into which variables are most relevant for predicting purchases and supports the development of more accurate and targeted models.

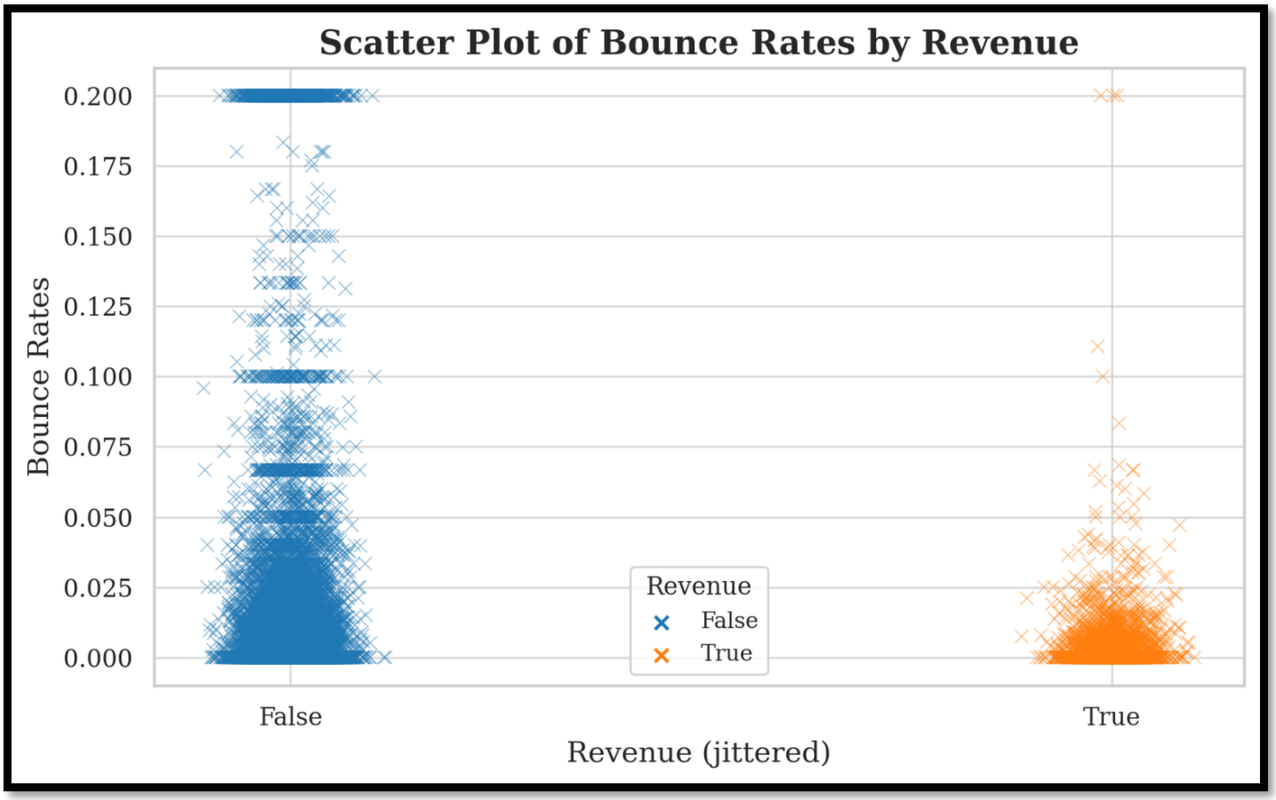

Figure 7.5.1: Bounce Rates by Revenue Scatter Plot

This segmented scatter plot shows the distribution of BounceRate observations that resulted in a purchase and ones that did not. The chart shows that users converted to purchasers tend to have much lower bounce rates than non-purchasers. These results align with previous studies done on the UCI dataset and support the theoretical constructs of TAM.

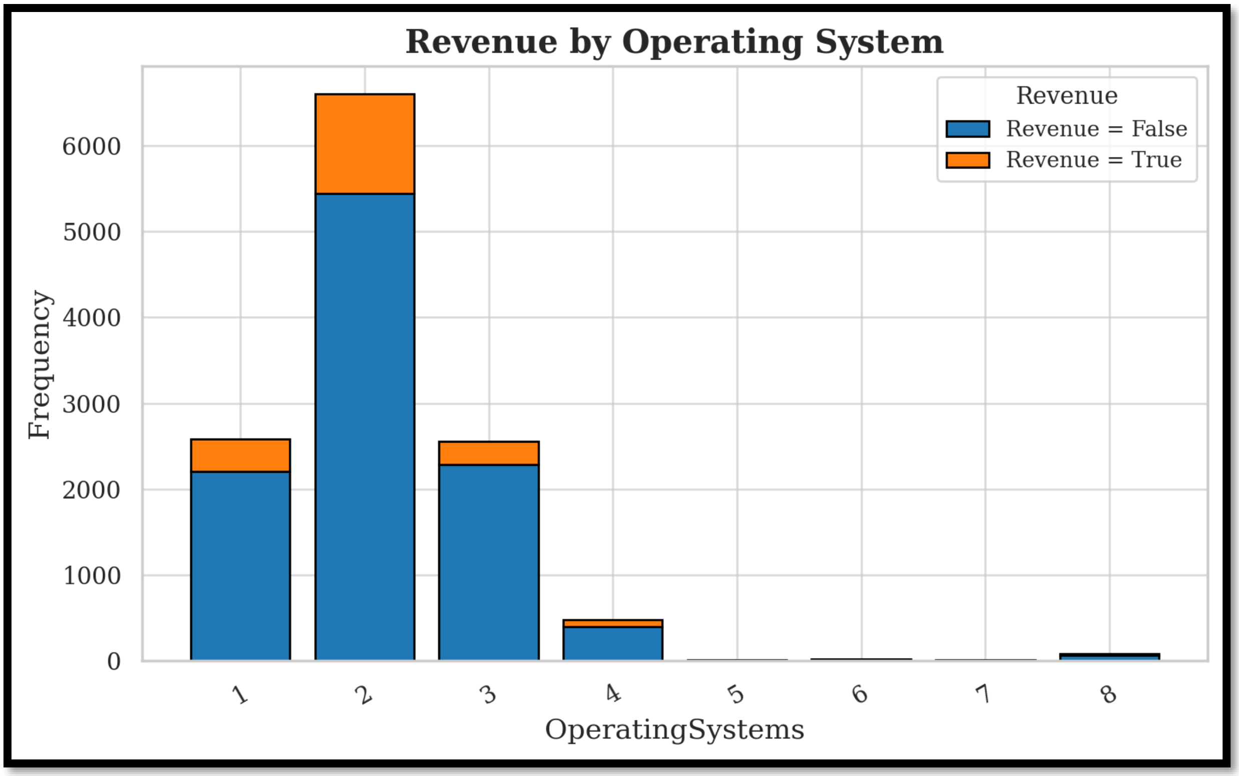

Figure 7.5.2: *Frequency of Operating System by Revenue*

The stacked bar chart above shows the distribution of OperatingSystems across purchasing (Revenue = True) and non-purchasing (Revenue = False) sessions. The top three most used operating systems make up the majority of all sessions, which include 94.4% of purchasing sessions. This could indicate that the feature could benefit from grouping class levels that do not have enough observations into an "Other" variable when encoding. This aligns with the preconceived beliefs about the distribution of operating systems among users and their conversion rates.

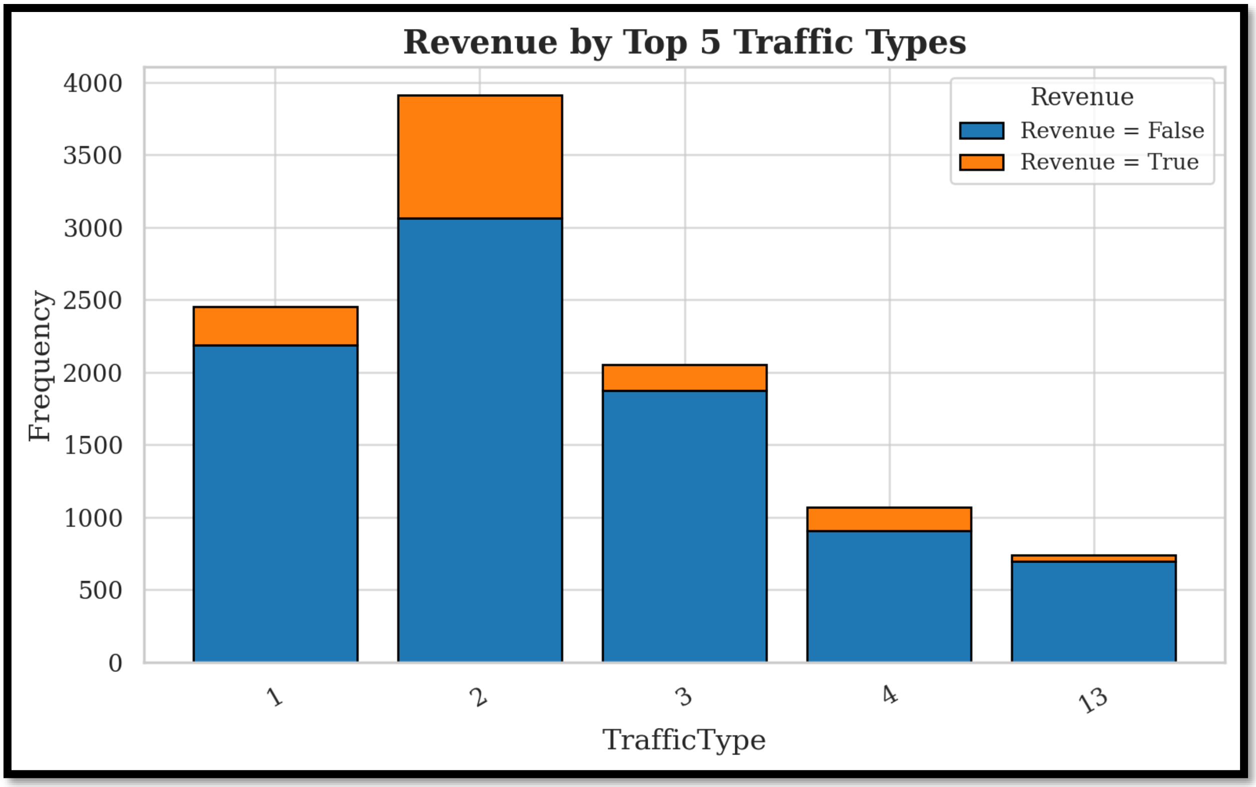

Figure 7.5.3: *TrafficType frequency by Revenue among the top five most common types*

The stacked bard chart above visualizes the frequency of the top five most common traffic types by the target variable (Revenue). The most frequent traffic types account for a majority of total sessions in the dataset and 78.4% of purchase sessions. In Google Analytics, traffic types are categorized based on the origin or method that a user accessed the site. For example, TrafficType 1 is the source associated with users arriving by Google's search engine or other organic routes. TrafficType 2 contains the largest number of conversions and represents users that reached the site directly without a specific source or by the URL. This information is valuable when trying to isolate conversion sessions and could align with the returning visitor class of the VisitorType feature due to reaching the site directly.

7.6 TWO-SAMPLE T-TESTS

To evaluate whether the mean values of continuous variables significantly differ between users who made a purchase (Revenue = 1) and those who did not (Revenue = 0), independent two-sample T-tests were performed. Furthermore, the tests do not assume equal variance between the two groups and the results are outlined in the table below.

FEATURE	REVENUE (1) MEAN	REVENUE (0) MEAN	T-STATISTIC	P-VALUE
ADMINISTRATIVE	3.3906	2.1177	-14.02	< 0.001
ADMINISTRATIVE_ DURATION	119.5	73.7401	-9.34	< 0.001
INFORMATIONAL	0.7862	0.4518	-9.09	< 0.001
INFORMATIONAL_ DURATION	57.6114	30.2362	-6.61	< 0.001
PRODUCTRELATED	48.2102	28.7146	-14.00	< 0.001
PRODUCTRELATED_ DURATION	1876.2	1070.0	-14.45	< 0.001
BOUNCERATES	0.00512	0.0253	34.85	< 0.001
EXITRATES	0.0196	0.0474	44.33	< 0.001
PAGEVALUE	27.2645	1.9760	-31.20	< 0.001
SPECIALDAY	0.0232	0.0684	12.97	< 0.001

Table 7.6.1: *Two-Sample T-Tests for all continuous features*

The results of the two sample T-tests revealed that all continuous variables are statistically significant between users who completed a purchase and users who did not.

7.7 CHI-SQUARE TESTS FOR CATEGORICAL FEATURES

Chi-Square Tests of Independence were conducted to examine whether the distribution of users across distinct levels of a categorical variable differs significantly based on whether a purchase occurred. The results of the tests are located in the table below:

FEATURE	DEGREES OF FREEDOM (DF)	CHI-SQ. STATISTIC	P-VALUE
MONTH	9	384.9348	< 0.001
OPERATINGSYSTEM	7	75.0271	< 0.001
BROWSER	12	27.7153	0.0061
REGION	8	9.2528	0.3214
TRAFFICTYPE	19	373.1456	< 0.001
VISITORTYPE	2	135.2519	< 0.001
WEEKEND	1	10.5818	0.0011

Table 7.7.1: *Chi-Square Tests for all Categorical Features*

7.8 CORRELATION MATRIX

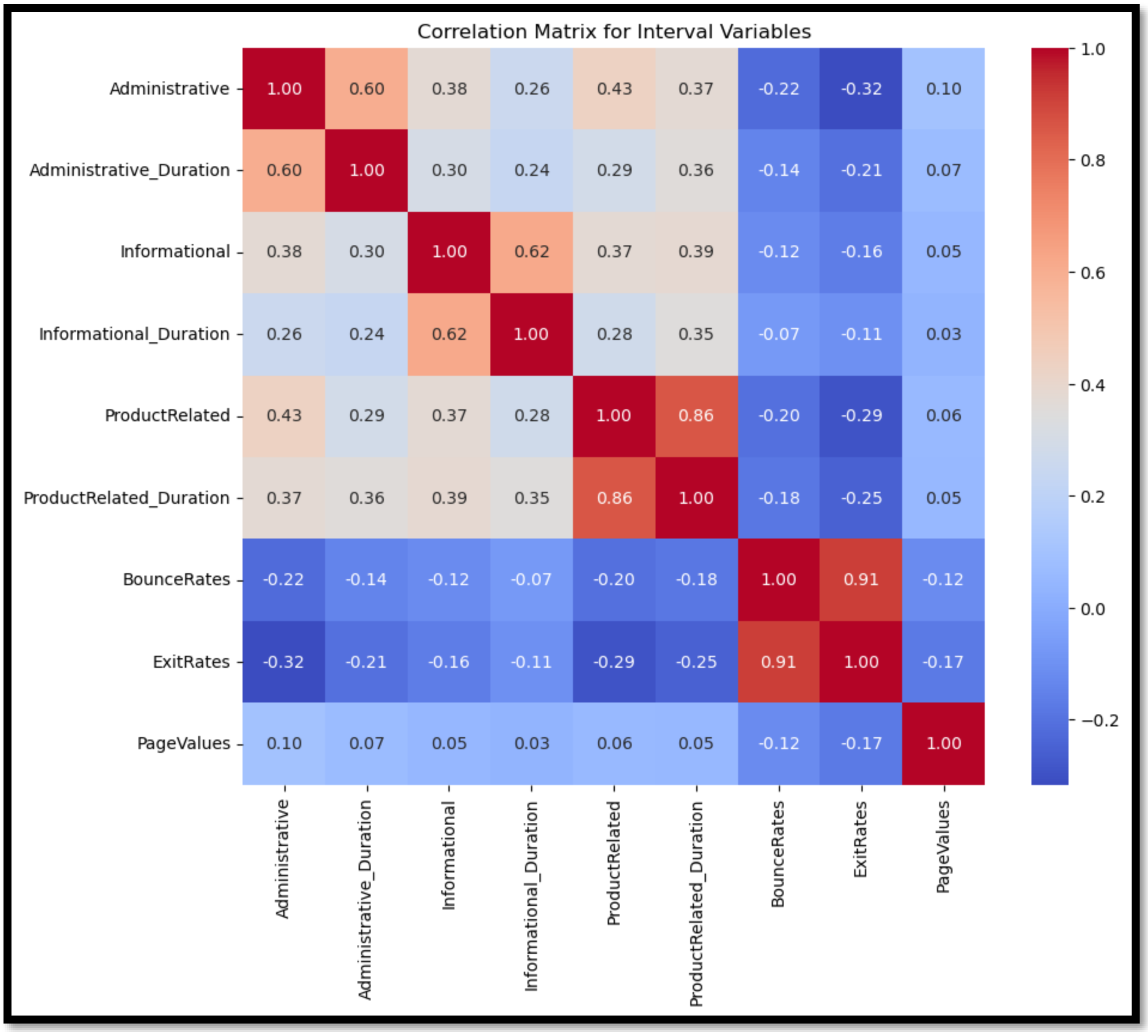

Figure 7.8.1: *Correlation Matrix of Interval Variables*

The correlation matrix illustrates how key user behaviors are related to each other. The matrix visualizes diagonal values reflecting perfect correlation of 1 between a variable and itself. The off-diagonal values show how two variables are related where value close to 1 means strong positive relationship and value near to –1 means strong negative relationship. A correlation of 0.86 shows a strong relationship between ProductRelated and ProductRelated_Duration. This means when visitors look at more product pages, they also spend more time on them. BounceRate and ExitRates have an exceedingly high correlation of 0.91, indicating that when a visitor leaves the site quickly that page is often the last page they visit. Informational and Informational_Duration shows a positive relationship of 0.62, means visitors look at more informational pages and they also spend more time reading them. Administrative and Administrative_Duration shows a positive correlation of 0.60, suggesting that visitors who visit more admin type pages also spend more time there. Negative correlations like -0.32 between Administrative and ExitRates suggest that those visitors who explore admin pages are less likely to leave the site right away.

7.9 EDA FINDINGS & INSIGHTS

The exploratory data analysis (EDA) uncovered several important behavioral patterns from UCI's Online Shopper Purchasing Intention Dataset that have direct implications for predicting conversions and align with the Technology Acceptance Model.

Sessions resulting in a purchase consistently displayed higher engagement: users visited more pages and spent significantly more time on the site across all page types, especially product related. For instance, ProductRelated_Duration was nearly 75% higher among converting users, suggesting that sustained interaction with product pages is a strong indicator of a future purchase. This aligns with the Perceived Usefulness construct in TAM: users are more likely to complete a purchase when they spend more time engaging with content that is directly related to their goals.

BounceRates & ExitRates were notably lower in sessions, which resulted in purchases. On average, BounceRates for non-purchasing sessions (0.025) were approximately five times greater than for purchasing sessions (0.005). This finding shows that highly engaged users tend to navigate deeper into the site without abandoning early. Furthermore, it supports the Perceived Ease of Use construct in TAM: users who experience smoother, more intuitive navigation paths are more likely to convert to purchases.

The PageValue feature, which measures the conversion contribution of each page visited, was much higher for buyers (27.27 vs. 1.98). This reinforces the idea of Perceived Usefulness, confirming that buyers tend to follow more valuable navigation paths. The high skewness and kurtosis of PageValue further show that only a minority of users engage with these high value behaviors, which makes identifying them critical for predictive modeling.

One unexpected result is that the average Administrative_Duration (80.82 sec) is much higher than that of Informational_Duration (34.47), considering that informational pages tend to receive more attention. This finding may indicate the presence of complicated or unintuitive website design among administrative pages. While administrative engagement may indicate deeper paths that lead to conversion, the relatively high duration compared to informational pages may be an indicator of friction for the user. Further investigation should be conducted to potentially improve the classification of users or the user interface design.

TAM Construct	Feature	Interpretation
Perceived Usefulness	PageValue, ProductRelated_Duration	Users who find content valuable or relevant to their goals engage more deeply and convert more.
Perceived Ease of Use	BounceRates, ExitRates, Administrative_Duration	Users that experience friction or poor navigation are more likely to abandon sessions.

Table 7.9.1: *Theoretical link between features and TAM*

Another key insight found in the EDA was the presence of strong positive correlations between ProductRelated & ProductRelated_Duration (0.86). Additionally, the high correlation between BounceRates & ExitRates (0.91) suggests that they are explaining the same information and could produce redundancy among features. To address this, one feature from each highly correlated pair will be dropped when using linear algorithms to meet model assumptions and avoid issues involving multicollinearity.

In conclusion, these findings not only justify the inclusion of the key features in the EDA but also provide a theoretical basis for interpreting user behavior in a way that is valuable to marketing and website design strategies. Moving forward, the integration of TAM into the modeling and interpretation methodology improves both predictive and explanatory value of the analysis.

8. DATA PREPROCESSING

This section outlines the steps taken to clean, transform, and prepare the dataset to ensure quality and consistency for the modeling process. The goal was to clean the raw inputs by removing redundant variables, noise, and reshaping the dataset into a structure that is compatible with the different machine learning algorithms that will be applied.

8.1 DATA CLEANING SUMMARY

The dataset was assessed for integrity, completeness, and consistency. The initial analysis revealed that there are no missing values across any variables in the original dataset, which indicated that no imputation methods were required. Furthermore, it was checked for duplicate entries and any instances identified were removed to prevent redundancy. Boolean variables such as the target (Revenue) and Weekend were converted to a binary numerical form to obtain compatibility with classification algorithms.

8.2 OUTLIER DETECTION & TREATMENT

Outlier detection analysis was performed on all continuous variables using the Interquartile Range (IQR) method. Specifically, Administrative_Duration, Informational_Duration, ProductRelated_Duration, and PageValues displayed long tailed distributions and contained a considerable number of extreme values. For the variables containing outliers, a decision was made on each variable about whether to keep the observations to preserve information or to apply Winsorization or a Log transformation to reduce skewness and mitigate their influence on any sensitive models. Visual examinations were performed to validate these findings and determine the appropriate decisions to make.

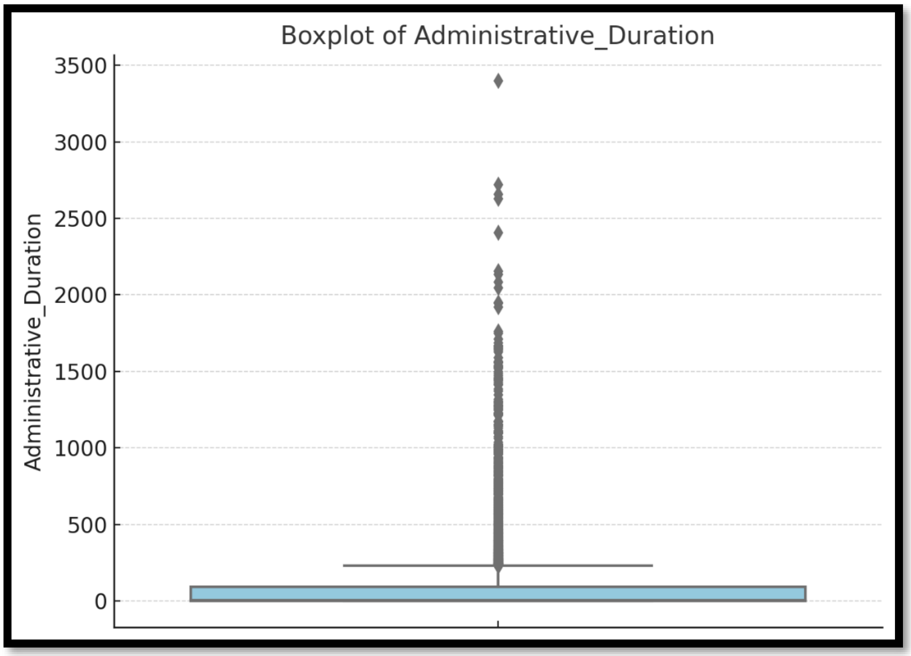

Figure 8.2.1: *Box Plot of Administrative page durations that highlights extreme values.*

8.3 ENCODING CATEGORICAL FEATURES

Categorical features were preprocessed using encoding & binning techniques that are suitable for both linear and non-linear algorithms. The Month feature was binned into quarters (Q1, Q2, etc.) and one-hot encoded to reduce dimensionality while maintaining information. OperatingSystem was binned into two one-hot encoded variables (IsMajorOS & OtherOS) because the exploratory analysis lacked any indication that it was meaningful when predicting purchases, but we wanted to maintain some information to verify. Browser, indicating which internet browser used, was also binned into two features (IsPopularBrowser & OtherBrowser) and one-hot encoded due to the same reasoning and high cardinality. Furthermore, the TrafficType variable, which exhibited high cardinality, was binned to retain only the four most frequently occurring traffic types, while all remaining categories were consolidated into a single class labeled TrafficTypeOther. The resulting categories were then one-hot encoded to enable the assessment of whether the most common traffic sources demonstrated any predictive relationship with purchase behavior. However, the original form of TrafficType was retained to experiment with applying other methods to address high cardinality such as Smoothed Weight of Evidence (SWOE). SWOE is an encoding technique that transforms categorical features into continuous numeric values that reflect their relationship with the target variable.

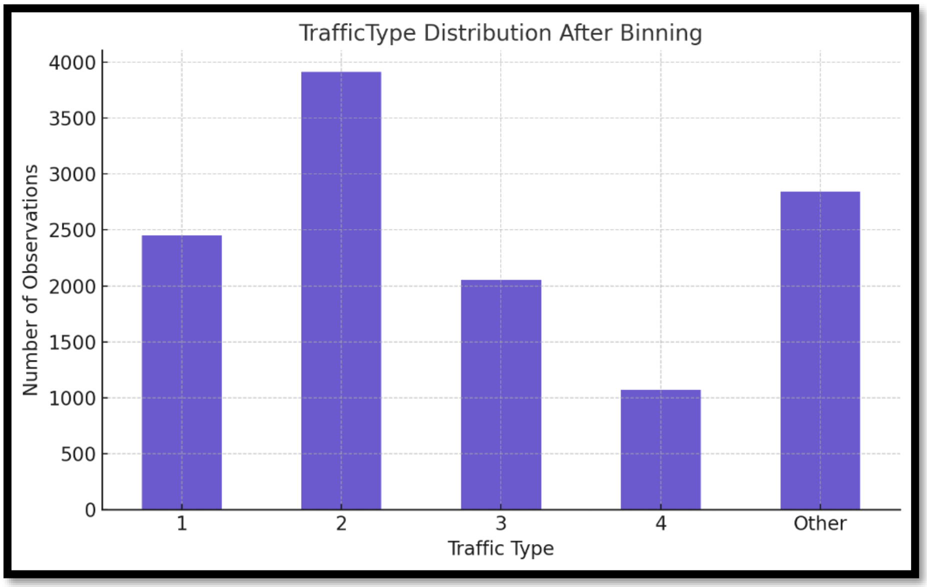

Figure 8.3.1: *Frequency of Observations in Traffic Type Bins*

8.4 FEATURE SCALING & TRANSFORMATION

Interval features were scaled using a few different techniques, we will conduct an experiment on how model performance differs between them. Multiple interval features exhibited heavily skewed distributions; to address this issue Log transformations will be applied to handle this and reduce the impact of outliers. Additionally, in a separate set, features like PageValue & ExitRates have been quantile binned in SAS to evaluate if it better captures any relationship with the target. For features that displayed relatively low skewness, we will apply standardization and examine whether applying Log transformations provide any difference in performance.

8.5 FEATURE ENGINEERING

To potentially enhance the predictive power of the dataset, multiple domain-relevant features were engineered. Total_PageViews and Total_TimeSpent are two features created by calculating the sum of all pages visited and time spent on webpages. AverageTime_PerPage was engineered by dividing the total time spent on webpages by the total page views of each user. Furthermore, SessionEfficiency has been generated by dividing each user's PageValue by total time spent on webpages and adding one to the denominator (PageValue / (Total_TimeSpent + 1). to capture user session efficiency. Adding one to the denominator prevents division by zero and stabilizes the metric for sessions with minimal duration. SessionEfficiency allows us to better distinguish between high-intent users who convert quickly and low-efficiency users who may browse extensively without making a purchase. Furthermore, this feature not only improves signal clarity for modeling but also aligns with the business goal of identifying sessions that make purchases with minimal resource consumption. The engineered features will be used to see if they enhance the representation of the dataset and better capture indicators that lead to a consumer completing an online purchase.

Engineered Feature	**Description**
Total_Page_View	Total number of pages visited in a session.
Total_TimeSpent	Total time spent across all types of pages in a session.
AvgTime_PerPage	Average time spent per page visited during the session.
SessionEfficiency	Ratio represents the efficiency of a session generating value. (PageValue / (Total_TimeSpent + 1))

Table 8.5.1: *Engineered features and their descriptions*

8.6 FINAL FEATURE SET SUMMARY

The final dataset consists of a combination of encoded, transformed, and engineered features optimized for both accurate predictions and interpretability. Additionally, redundant variables like ExitRates & ProductRelated were dropped from the final feature set to address any multicollinearity issues on models. Separate feature sets were created to conduct an experiment on which preprocessing techniques perform the best on the dataset. The following tables summarize the final feature set:

FEATURE	ORIGIN	TRANSFORMATION METHOD	DESCRIPTION
ADMINISTRATIVE	Original	Standardization	Number of administrative pages visited.
ADMINISTRATIVE_ DURATION	Original	Log Transformation	Time spent on administrative pages.
INFORMATIONAL	Original	Log Transformation	Number of informational pages visited.
INFORMATIONAL_ DURATION	Original	Log Transformation	Time spent on informational pages.
PRODUCTRELATED_ DURATION	Original	Quantile Binning	Time spent on product-related pages.
BOUNCERATES	Original	Standardization	Percentage of visitors who enter and exit quickly.
PAGEVALUES	Original	Log Transformation	The average value of a page is based on conversion probability.
AVGTIME_PERPAGE	Engineered	Log Transformation	Total_TimeSpent divided by Total_PageView
SESSION EFFICIENCY	Engineered	Quantile Binning	PageValue divided by (Total_TimeSpent + 1)

Table 8.6.1: *Final Feature Set for Interval Variables*

FEATURE	ORIGIN	ENCODING METHOD	DESCRIPTION
Q1	Engineered	One-Hot Encoding	Users visited e-commerce sites in the 1st quarter of the year. (Jan.-Mar.)
Q2	Engineered	One-Hot Encoding	Users visited e-commerce sites in the 2nd quarter of the year. (Apr.-Jun.)
Q3	Engineered	One-Hot Encoding	Users visited e-commerce sites in the 3rd quarter of the year. (Jul.-Sept.)
Q4	Engineered	One-Hot Encoding	Users visited e-commerce sites in the 4th quarter of the year. (Oct.-Dec.)
ISMAJOROS	Engineered	One-Hot Encoding	Consumer uses one of the three most popular operating systems.
ISPOPULARBROWSER	Engineered	One-Hot Encoding	Consumer uses one of the three most popular internet browsers.
TRAFFICTYPE	Original	SWOE	Source of visitor traffic to the website.
TRAFFIC_TYPE1	Engineered	One-Hot Encoding	Direct, the user typed the URL manually or used a bookmark
TRAFFIC_TYPE2	Engineered	One-Hot Encoding	Organic Search, the user came via search engine like Google/Bing.
TRAFFIC_TYPE3	Engineered	One-Hot Encoding	Paid Search, users arrived via paid ad campaigns.
TRAFFIC_TYPE4	Engineered	One-Hot Encoding	Referral, Clicked a link from another website or banner ad.
TRAFFIC_TYPEOTHER	Engineered	One-Hot Encoding	The remaining Google Analytics Traffic Types weren’t specified.
RETURNING_VISITOR	Engineered	One-Hot Encoding	Whether the user was a returning site visitor or not.
WEEKEND	Original	Label Encoding	Whether the user visited the site on a weekend or not.

Table 8.6.2: *Final Feature Set for Categorical Variables*

9. MODELING

Data partitioning, variable transformations, and algorithms were developed and implemented using SAS Enterprise Miner.

9.1 DATA PARTITIONING

The dataset was partitioned into two equal subsets using stratified sampling, with 50% of the data used for training the models and the remaining 50% reserved for the validation set to assess model performance. Furthermore, we addressed the class imbalance of the target variable using two distinct approaches: under-sampling the majority class and implementing cost-sensitive learning.

9.2 TREE-BASED ALGORITHMS

Tree-based algorithms are machine learning techniques that are suited well for classification modeling. Their design allows them to capture complex, non-linear relationships among features with minimal data preprocessing. To classify purchases, we apply multiple tree-based models: Decision Trees, Gradient Boosting, and Forest-based methods.

I. DECISION TREES

A Decision Tree was the first model implemented on the dataset to establish a baseline for tree-based model performance. Decision trees apply a split-search algorithm to segment the dataset based on splits that maximize class purity. It searches for potential splits across all input features and selects those that result in the largest reduction in impurity. These splits create a tree-like structure using the maximum amount of input features and are then pruned into features that maximize the correct prediction of cases using logworth. The final selected model represents the simplest iteration with the highest validation assessment measure that is specified. In this study, the assessment measure employed was based on the misclassification of predictions for the target variable (Revenue).

II. FOREST MODELS

To improve upon predictions and address the risk of overfitting when using single-tree models, we implemented a high-performance Forest model. Forest based algorithms use an ensemble approach that combines the outputs of multiple decision trees to produce a more robust prediction that will generalize better to new datasets. During the tree building process, each split is determined using a random subset of input variables to reduce the correlation between trees (Nyongesa, D., 2020). The main advantages of using a Forest-based model are that it reduces the occurrence of overfitting and does not require you to standardize continuous features. Furthermore, it also provides you with a feature-importance scoring method based on the marginal reduction in impurity.

III. GRADIENT BOOSTING

Gradient Boosting is a tree-based ensemble method that performs well in classifications tasks by combining a sequence of "weak learners," with each additional tree trained to correct the residual errors of the combined model (SAS Institute Inc., n.d.). This means that each additional tree added is focused on the observations that were previously misclassified, which allows it to capture non-linear relationships in the data. To prevent overfitting and improve the performance of the model, hyperparameter tuning can be applied by adjusting the learning rate, maximum number of trees, and tree depth. Like Forest models, Gradient Boosting also produces feature importance rankings that are based on the most influential predictors across all trees constructed in the model.

9.3 LINEAR MODELS

Linear models function as a valuable tool in predictive modeling due to their simplicity and interpretability. In this study, we applied several linear algorithms to predict whether a user will complete a purchase and those included are: Logistic regression, Partial Least Squares (PLS) regression, and both LASSO & Adaptive LASSO regression.

I. LOGISTIC REGRESSION

Logistic regression is a commonly used linear classification algorithm that estimates the probability of a binary outcome based on the linear combination of the input features. The algorithm is useful due its explainability and ability to provide meaningful coefficient estimates that quantify the predictive importance of each feature based on the log-odds of the target event. In this study, three feature selection methods were paired with the regression model to gain insights into which resulted in the best performance. The feature selections used were: Forward Selection, Backward Selection, and Stepwise Selection. Forward Selection starts with the most statistically significant input and at each step adds the feature that provides the greatest improvement in the specified assessment measure. Backward Selection starts with all inputs included in the model and at each step removes the least significant variable until all remaining inputs are statistically significant. Stepwise selection is a hybrid of forward and backward methods. It begins with the most significant input and, after each addition, reassesses all previously included variables to remove any that are no longer statistically significant.

In SAS Enterprise Miner, regression model outputs contain standardized estimates that provide the ability to rank variables in order of their importance. Another useful tool provided by logistic regression models are the odds ratio estimates, which tell you how much the log odds of an outcome increase or decrease when an input's value changes.

While logistic regression provides a good baseline and multiple interpretability tools, its assumption of linear relationships and limited capability of capturing complex relationships may lead to lower predictive performance than more advanced algorithms.

II. PARTIAL LEAST SQUARES REGRESSION

Partial Least Squares Regression is a linear modeling technique designed to handle datasets that contain correlated inputs by combining them into a smaller set of factors that are the most useful for predicting the target. The extracted factors maximize the covariance between inputs and the target, which enables effective predictions even when multicollinearity is present in the data. However, PLS lacks the interpretability seen in traditional logistic regressions for each input's coefficient estimate.

III. LASSO & ADAPTIVE LASSO REGRESSION

LASSO (Least Absolute Shrinkage and Selection Operator) is a linear regression technique that performs both feature selection and regularization by adding a penalty term that pushes the less important inputs towards zero. The penalty strength can be tuned to balance the complexity of the model, improve predictive performance, and reduce overfitting through the removal of the least important features that do not add value to the model. In this study, LASSO was applied to improve generalization in the presence of potentially redundant or correlated predictors and improve over the baseline logistic regression model.

Adaptive LASSO takes the original method further by assigning different weights to each feature's penalty, which allows the model to effectively distinguish between strong and weak predictors. This process aids in feature selection and importance because it retains the weaker, but meaningful predictors that standard LASSO might prematurely drop (SAS Institute Inc., n.d.). Furthermore, another valuable tool that the adaptive LASSO provides is the standardized estimates of both continuous and categorical features. This allows for the ranking of all features based on their magnitude even when applying preprocessing techniques like binning as we have done in this study.

9.4 NEURAL NETWORKS

To capture complex relationships among the features in user behavior, this study applied a Multilayer Perceptron (MLP) neural network. Neural networks are a type of machine learning algorithm that is modeled after the structure of the human brain. They are made up of layers of connected nodes called neurons that work together to learn patterns in the data. Neural networks are particularly good at identifying complex relationships that simpler models could miss which usually leads to strong predictions. However, they are also known as "black boxes" due to their lack of interpretability. This is the case because unlike other models, neural networks do not provide you with feature importance methods and their multiple layered construction makes it difficult to understand exactly how they produced their predictions. Additionally, in environments where explainability is crucial, this limitation can make it harder for businesses to trust the model's outputs.

In this study, we applied a specific type of neural network called a Multilayer Perceptron. This algorithm takes in the data through an input layer that processes it through one or more hidden layers that produce the predictions in the output layer. Each layer provides more information, and the model learns by adjusting the strength of connections based on how well it is performing on predictions. This process is repeated until it reaches the maximum number of iterations or lacks the ability to improve its assessment measures.

9.5 SUPPORT VECTOR MACHINES

Support Vector Machines (SVM) are advanced classification algorithms that generate predictions by finding the most optimal hyperplane that separates the outcome classes of the target. Its goal is to maximize the margin between classes and then choose the decision boundary that leaves the greatest distance between itself and the closest data points of each class. This allows SVM's to be effective when the data has clear but complex patterns that are not easily captured by more traditional models.

In SAS Enterprise Miner, the SVM supports both linear and non-linear classification using kernel functions, which allows the model to find separation in cases where the data is not linearly separable. Their strength is in its ability to manage high dimensional data and maintain levels of high performance. However, SVM's are often considered to be less interpretable and do not provide exact decision rules to explain why a prediction was made. In this study, multiple SVM models were implemented to gain insight into what kernel functions produce the best predictions of user behavior and how their results compare with other models.

10. RESULTS & ANALYSIS EVALUATION

The following section presents the results of the analysis and performance comparisons across all preprocessing techniques, algorithms, and as well as evaluating their interpretations. Furthermore, it highlights the comparison of feature importance techniques implemented to see if they are consistent and align with existing literature and the Technology Acceptance Model. Model performance was assessed using validation data, with a wide range of assessment measures that include Accuracy, Precision, Recall, F-1 Score, and AUC-ROC. A table highlighting the different algorithms and techniques applied is located below.

MODEL TYPE	ALGORITHM	DISTINCT TECHNIQUES/FEATURES
TREE-BASED MODELS	Decision Tree	Recursive splitting using ProbF for interval, ProbChisq for nominal, and Entropy for ordinal features. Final model selected on decision assessment.
	HP Forest	Ensemble bagged trees using majority voting, evaluated both Loss Reduction & Random Branch Assignment feature importance methods.
	Gradient Boosting	Ensemble that Boosts decision trees with shrinkage (Learning Rate).
LINEAR MODELS	Logistic Regression	Evaluated Backward, Forward, and Stepwise feature selection methods. Final model selected based on Misclassification and Decision assessment.
	Partial Least Square Regression	Projects data into latent components and manages multicollinearity.
	LASSO Regression	Feature selection using L1 Regularization and shrinkage.
	Adaptive LASSO Regression	Customizes penalty for each variable, helps preserve key features. Produces standardized estimates for both categorical and continuous features.
NONLINEAR MODELS	Neural Network (MLP)	Learns complex relationships through layers of connected nodes using weighted inputs and activation functions.
	Support Vector Machines	Supports kernel functions to separate non-linear classes. Evaluated and compared between kernel functions: Linear, Polynomial, Radial, and Sigmoid.

Table 10.1: *Outline of algorithms applied and their distinct features.*

10.1 CLASS IMBALANCE MITIGATION

The class imbalance of the target (Revenue) posed a significant challenge in modeling user conversion behavior, with the number of non-purchasing sessions outnumbering purchasing ones by a large margin. In this study, we applied two methods to counteract this issue, and the approaches used were Undersampling & Cost-Sensitive Learning.

When applying under-sampling, the majority class (non-purchasers) was reduced to match the size of the minority class (purchasers) to allow models to be trained on a more balanced dataset. The cost sensitive learning implementation involved using a decision matrix to apply the inverse prior probabilities of the target as decision weights. This allows the model to apply penalties to misclassifications for each decision outcome and prioritize the outcomes that are the most important.

In the case of predicting e-commerce purchases, recall and F1-score carry much more value when assessing the model. After applying both methods on the data, the results showed a significant improvement in all assessment metrics besides accuracy. The decrease in accuracy is understandable due to the loss of observations in the dataset when under-sampling was applied. A comparison of performance for the baseline logistic regression before and after the methods were applied are displayed in the table below.

MODEL	METHOD	ACCURACY	PRECISION	RECALL	F-1 SCORE
LOGISTIC REGRESSION	**None**	90.01%	72.2%	57.65%	64.11%
LOGISTIC REGRESSION	**Undersampling & Cost-Sensitive Learning**	83.49%	87.4%	73.48%	79.84%

Table 10.1.1: Class Imbalance method performance comparison

10.2 COMPARISON OF PREPROCESSING TECHNIQUES

To improve model performance and interpretability, multiple preprocessing strategies were applied and evaluated to compare their differences in model performance. These transformations were used to address scale differences and to verify that the input data is compatible with various algorithms. This section compares the predictive impact of two preprocessing comparisons: One-Hot encoding vs. SWOE for the categorical feature Traffic Type, and Log transformations vs. binning techniques for key continuous features.

For key continuous features such as PageValue, ProductRelated_Duration, and SessionEfficiency, both Log transformations and two different binning techniques (Quantile & Optimal) were applied to reduce skewness and create more stable distributions. Models like logistic regression and neural networks specifically benefit from applying transformations like these due to their model assumptions and smoother optimization processes.

After applying both methods to each of the features on our baseline logistic regression, the results displayed that Optimal binning improved the performance of the model and allowed it to identify relationships it previously missed. For example, before binning both ProductRelated_Duration and SessionEfficiency the model failed to include them in their final model. after both features were selected due to being statistically significant. Furthermore, after being applied to the model, both features were selected due to being statistically significant and resulted in the improvement of assessment measures. These results will also improve the interpretability of the model due to being able to segment users into the bins of both features. Despite the success seen by the other features, the results for PageValue showed that the model benefitted more from a Log transformation. The findings resulted in the decision to apply binning to both ProductRelated_Duration and SessionEfficiency but continue to use a Log transformation for PageValue. The table below shows the before and after effect of applying our techniques on the model.

MODEL	**FEATURE TRANSFORMATION**	**ACCURACY**	**PRECISION**	**RECALL**	**F-1 SCORE**
LOGISTIC REGRESSION	**PageValue:** **Log Transformation**	83.39%	87.78%	77.57%	82.36%
	ProductRelated_Duration: **Log Transformation**				
	SessionEfficiency: **Log Transformation**				
LOGISTIC REGRESSION	**PageValue:** **Log Transformation**	84.64%	86.38%	79.77%	82.94%
	ProductRelated_Duration: **Optimal Binning**				
	SessionEfficiency: **Optimal Binning**				

Table 10.2.1: *Baseline Regression performance before and after comparison*

Additionally, we applied the Smoothed Weight of Evidence (SWOE) on the TrafficType feature to evaluate if there was a relationship with the target that had previously been left uncaptured in literature. The results of our baseline logistic regression showed that TrafficType was included in the model, statistically significant (P-value = <0.0001), and had a standardized coefficient of 0.183. These results align with our EDA findings where we stated that some TrafficTypes contain substantially higher numbers of purchases than others, which could aid the process of identifying user conversions. Furthermore, the addition of the SWOE transformed feature led to a slight improvement in correct classifications for the model.

10.3 PREDICTIVE MODELING PERFORMANCE RESULTS

This section outlines the final performance results for each type of predictive model developed in this study. A list of abbreviations is provided below for reference.

- OPT: Optimal binning
- LOG: Log transformed
- STD: Standardization

I. DECISION TREE

The decision tree's subtree assessment method was configured to "Assessment" with an assessment measure of "Decision". The final model chose six important variables that were used to determine the splits, the most important of which was the SessionEfficiency feature. The resulting tree has a total of seven leaves and a validation F-1 score of 84.01%. The table and figure below contain the feature importance table and tree diagram of the model.

FEATURE	NUMBER OF SPLITTING RULES	IMPORTANCE	VALIDATION IMPORTANCE	RATIO OF VALIDATION TO TRAINING IMPORTANCE
SESSIONEFFICIENCY	1	1.000	1.000	1.000
Q4	1	0.3013	0.2134	0.7082
PRODUCTRELATED_ DURATION	1	0.1826	0.1419	0.7772
TRAFFICTYPE	1	0.1631	0.0652	0.3994
Q3	1	0.1608	0.1824	1.1341
ADMINISTRATIVE_ DURATION	1	0.1176	0.1310	1.1135

Table 10.4.1: *Decision tree feature importance rankings*

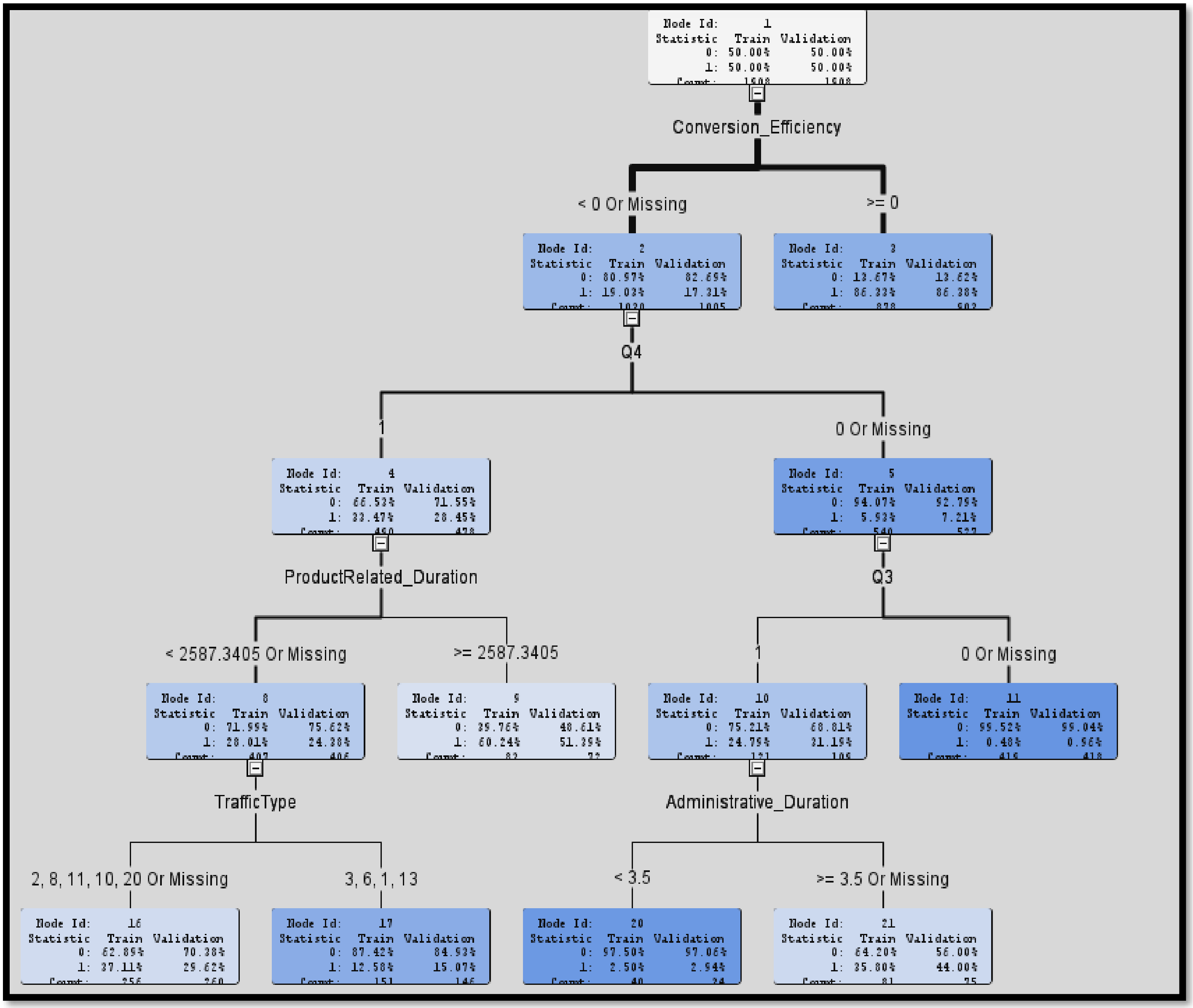

Figure 10.4.1: *Tree diagram for final decision tree model.*

II. RANDOM FOREST

The Random Forest model developed in our pipeline was configured to use Random Branch Assignment for feature selection, 200 trees, and the proportion of observations in each sample was set to 0.6. The validation F-1 score was 0.8466 and the top three variables identified by the model were PageValue, SessionEfficiency, and ProductRelated_Duration. The feature importance rankings are in the table below.

FEATURE	NUMBER OF SPLITTING RULES	VALIDATION GINI REDUCTION	VALIDATION MARGIN REDUCTION
PAGEVALUE	352	0.11923	0.23420
SESSIONEFFICIENCY	130	0.05373	0.10562
PRODUCTRELATED_ DURATION	202	0.00664	0.01809
Q4	300	0.00598	0.01691
BOUNCERATE	214	0.00220	0.00816
AVGTIME_PERPAGE	73	0.00226	0.00661
RETURNING_VISITOR	159	0.00312	0.00656
TRAFFICTYPE	134	0.00134	0.00544
ADMINISTRATIVE	63	0.00260	0.00494
PRODUCTRELATED	84	0.00146	0.00494
INFORMATIONAL	123	0.00011	0.00181
Q3	50	0.00079	0.00161
Q2	25	0.00055	0.00128
ADMINISTRATIVE DURATION	50	0.00020	0.00117
Q1	38	0.00020	0.00114
INFORMATIONAL_DURATION	35	-0.00023	0.00022
ISPOPULARBROWSER	28	-0.00011	0.00013

Table 10.4.2: *Random Forest feature importance rankings*

III. GRADIENT BOOSTING

The Gradient Boosting model implemented on the data was set to use subtree assessment measure “Decision” with the “Best Assessment Value” subseries. Furthermore, it was configured to have 100 trees, a maximum depth of 4, and a shrinkage rate of 0.01. The validation F-1 score was 0.8441 and the top three features identified in the feature importance table were SessionEfficiency, PageValue, and Q4. The feature importance rankings are in the table below.

FEATURE	NUMBER OF SPLITTING RULES	IMPORTANCE	VALIDATION IMPORTANCE	RATIO OF VALIDATION TO TRAINING IMPORTANCE
SESSIONEFFICIENCY	141	1	1	1
PAGEVALUE	91	0.969088	0.971673	1.002668
Q4	95	0.409244	0.221278	0.540701
TRAFFICTYPE	148	0.310228	0.112711	0.363318
BOUNCERATES	152	0.306484	0.075552	0.246512
PRODUCTRELATED_ DURATION	121	0.292679	0.13745	0.469628
PRODUCTRELATED	88	0.223316	0.092639	0.414835
Q3	74	0.209294	0.252422	1.206064
ADMINISTRATIVE_ DURATION	82	0.200099	0.105483	0.527151
AVGTIME_PERPAGE	83	0.197117	0.035582	0.180513
ADMINISTRATIVE	89	0.192353	0.047392	0.246381
REGION	117	0.185589	0.016186	0.087214
INFORMATIONAL_ DURATION	66	0.171824	0.016324	0.095006
INFORMATIONAL	40	0.118895	0.024006	0.201907
RETURNING_VISITOR	8	0.070872	0	0
Q2	9	0.057667	0.019041	0.3330181
WEEKEND	3	0.034837	0	0
ISMAJOROS	2	0.033944	0	0
ISPOPULARBROWSER	2	0.022442	0	0

Table 10.4.3: *Gradient Boosting feature importance rankings*

IV. LOGISTIC REGRESSION

Stepwise selection was chosen as the feature selection method because it had the best overall performance and a validation F1-score of 0.8428. The two tables below will contain the Analysis of Maximum Likelihood Estimates and the Odds Ratio Estimates.

PARAMETER	DF	ESTIMATE	STANDARD ERROR	WALD CHI-SQUARE	PR > CHISQ	STANDARDIZED ESTIMATE	EXP (EST)
INTERCEPT	1	1.9591	0.1808	117.43	<0.0001		7.093
OPT_SESSION EFFICIENCY 01: LOW	1	-2.5721	0.1262	415.36	<0.0001		0.076
OPT_SESSION EFFICIENCY 02: MID	1	0.2556	0.1301	3.86	0.0494		1.291
OPT_PRODUCT RELATED_ DURATION 01: LOW	1	-0.3986	0.1378	8.37	0.0038		0.671
OPT_PRODUCTRE LATED_ DURATION 01:MID	1	-0.2559	0.0978	6.84	0.0089		0.774
Q3 (0)	1	-0.4839	0.1123	18.58	<0.0001		0.616
Q4 (0)	1	-0.6310	0.0815	59.91	<0.0001		0.532
STD_ ADMINISTRATIVE	1	-1.0489	0.4499	5.43	0.0197	-0.0926	0.350
STD_ BOUNCERATE	1	-2.3136	0.7582	9.31	0.0023	-0.2558	0.099
R_TRAFFICTYPE	1	0.7908	0.1748	20.47	<0.0001	0.1826	2.205
RETURNING_ VISITOR	1	0.2283	0.0939	5.91	0.0151		1.256

Table 10.4.4: *Analysis of Maximum Likelihood Estimates for SW Regression*

EFFECT		POINT ESTIMATE
OPT_SESSIONEFFICIENCY 01: LOW	Low vs. High	0.008
OPT_SESSIONEFFICIENCY 01: MID	Mid vs. High	0.127
OPT_PRODUCTRELATED_ DURATION 01: LOW	Low vs. High	0.349
OPT_PRODUCTRELATED_ DURATION 01: MID	Mid vs. High	0.402
Q3	0 vs. 1	0.380
Q4	0 vs. 1	0.283
STD_ADMINISTRATIVE		0.350
STD_BOUNCERATE		0.099
R_TRAFFICTYPE		2.205
RETURNING_VISITOR	0 vs. 1	1.579

Table 10.4.5: *Odds Ratio Estimates for SW Regression.*

V. PARTIAL LEAST SQUARES REGRESSION

The validation F1-Score of the PLS regression was 0.8466 and the VIP cutoff for feature selection was 0.8. The tables below outline the Variable Selection, Analysis of Maximum Likelihood Estimates, and the Odds Ratio estimates.

FEATURE	STD. PARAMETER ESTIMATE	VIP	REJECTED BY STD. PARAMETER EST.	REJECTED BY VIP	ROLE
LOG_PAGEVALUE	0.09286	1.704943	Yes	No	Input
OPT_SESSION EFFICIENCY 01: LOW	-0.31940	1.743293	No	No	Input
OPT_SESSION EFFICIENCY 01: MID	0.12766	0.814768	No	No	Input
OPT_SESSION EFFICIENCY 01: HIGH	0.25464	1.291217	No	No	Input
OPT_PRODUCT RELATED_DURATION 01: LOW	-0.04071	0.804459	Yes	No	Input
R_TRAFFICTYPE	0.13902	0.662192	Yes	No	Input

Table 10.4.6: *PLS Variable Selection*

PARAMETER	DF	ESTIMATE	STANDARD ERROR	WALD CHI-SQUARE	PR > CHISQ	STANDARDIZED ESTIMATE	EXP (EST)
INTERCEPT	1	0.6000	0.2892	4.30	<0.0380		1.822
OPT_SESSION EFFICIENCY 01: LOW	1	-1.8350	0.2960	38.44	<0.0001		0.160
OPT_SESSION EFFICIENCY 02: MID	1	0.00214	0.1344	0	0.9873		1.002
OPT_PRODUCT RELATED_ DURATION 01: LOW	1	-0.5914	0.1253	22.26	<0.0001		0.554
OPT_PRODUCTRELATED_ DURATION 01: MID	1	-0.2470	0.0937	6.94	0.0084		0.781
LOG_PAGEVALUE	1	0.2672	0.1369	3.81	0.051	0.2249	1.306
R_TRAFFICTYPE	1	1.2001	0.1612	55.46	<0.0001	0.2771	3.321

Table 10.4.7: *PLS Regression Likelihood Estimates*

EFFECT		POINT ESTIMATE
OPT_SESSIONEFFICIENCY 01: LOW	Low vs. High	0.026
OPT_SESSIONEFFICIENCY 01: MID	Mid vs. High	0.160
OPT_PRODUCTRELATED_ DURATION 01: LOW	Low vs. High	0.239
OPT_PRODUCTRELATED_ DURATION 01: MID	Mid vs. High	0.338
LOG_PAGEVALUE		1.306
R_TRAFFICTYPE		3.321

Table 10.4.8: PLS Regression Odd Ratio Estimates

VI. LASSO REGRESSION

LASSO regression had a validation F1-score of 0.8394. The tables below contain both the Analysis of Maximum Likelihood Estimates, and the Odds Ratio Estimates for the model.

PARAMETER	DF	ESTIMATE	STANDARD ERROR	WALD CHI-SQUARE	PR > CHISQ	STANDARDIZED ESTIMATE	EXP (EST)
INTERCEPT	1	0.8191	0.3754	4.76	<0.0291		2.269
OPT_SESSION EFFICIENCY 01: LOW	1	-2.0102	0.3125	41.38	<0.0001		0.134
OPT_SESSION EFFICIENCY 02: MID	1	0.1512	0.1436	1.11	0.2924		1.163
OPT_PRODUCT RELATED_ DURATION 01: LOW	1	-0.4244	0.1375	9.53	0.0020		0.654
OPT_PRODUCTRELATED_ DURATION 01: MID	1	-0.2709	0.0980	7.64	0.0057		0.763
LOG_PAGEVALUE	1	0.2720	0.1443	3.55	0.0594	0.2499	1.313
R_TRAFFICTYPE	1	0.7178	0.1753	16.77	<0.0001	0.1657	2.050
ISMAJOROS	1	-0.2238	0.1600	1.96	0.1618		0.799
ISPOPULAR BROWSER	1	0.1515	0.0944	2.58	0.1084		1.164
Q2	1	0.2693	0.1013	7.07	0.0078		1.309
Q4	1	-0.3756	0.0864	18.88	<0.0001		0.687
STD_ ADMINISTRATIVE	1	-0.8939	0.4506	3.94	0.0473	-0.0789	0.409
STD_BOUNCE RATE	1	-2.2708	0.7385	9.45	0.0021	-0.2511	0.103
RETURNING_ VISITOR	1	0.2448	0.0951	6.63	0.0100		1.277

Table 10.4.9: *LASSO Regression Likelihood estimates*

EFFECT		POINT ESTIMATE
OPT_SESSIONEFFICIENCY 01: LOW	Low vs. High	0.021
OPT_SESSIONEFFICIENCY 01: MID	Mid vs. High	0.181
OPT_PRODUCTRELATED_ DURATION 01: LOW	Low vs. High	0.326
OPT_PRODUCTRELATED_ DURATION 01: MID	Mid vs. High	0.381
LOG_PAGEVALUE		1.313
R_TRAFFICTYPE		2.050
ISMAJOROS		0.639
ISPOPULAR BROWSER		1.354
Q2		1.713
Q4		0.472
STD_ ADMINISTRATIVE		0.409
STD_BOUNCE RATE		0.103
RETURNING_ VISITOR		1.632

Table 10.4.10: LASSO Regression Odds Ratio estimates

VII. ADAPTIVE LASSO REGRESSION

LASSO regression had a validation F1-score of 0.8394. The tables below contain both the Analysis of Maximum Likelihood Estimates, and the Odds Ratio Estimates for the model.

PARAMETER	DF	ESTIMATE	STANDARD ERROR	WALD CHI-SQUARE	PR > CHISQ	STANDARDIZED ESTIMATE	EXP (EST)
INTERCEPT	1	1.7754	0.1600	123.19	<0.0001		2.269
OPT_SESSION EFFICIENCY 01: LOW	1	-2.5173	0.1227	420.87	<0.0001	-0.73108	0.081
OPT_SESSION EFFICIENCY 02: MID	1	0.1783	0.1251	2.03	0.1541	-0.17068	1.195
OPT_PRODUCT RELATED_ DURATION 01: LOW	1	-0.3564	0.1365	6.82	0.0090	-0.10458	0.700
OPT_PRODUCTRE LATED_ DURATION 01: MID	1	-0.2393	0.0974	6.04	0.0140	-0.10581	0.787
R_TRAFFICTYPE	1	0.7808	0.1748	19.95	<0.0001	0.1803	2.183
Q3	1	-0.4662	0.1110	17.64	<0.0001	-0.04806	0.627
Q4	1	-0.6131	0.0809	57.47	<0.0001	-0.13154	0.542
STD_BOUNCE RATE	1	-2.1901	0.7482	8.57	0.0034	-0.2422	0.112
RETURNING_ VISITOR	1	0.2265	0.0933	5.89	0.0152	0.023392	1.254

Table 10.4.11: *Adaptive LASSO Regression Likelihood estimates*

EFFECT		POINT ESTIMATE
OPT_SESSIONEFFICIENCY 01: LOW	Low vs. High	0.008
OPT_SESSIONEFFICIENCY 01: MID	Mid vs. High	0.115
OPT_PRODUCTRELATED_ DURATION 01: LOW	Low vs. High	0.386
OPT_PRODUCTRELATED_ DURATION 01: MID	Mid vs. High	0.434
R_TRAFFICTYPE		2.183
Q3		0.394
Q4		0.293
STD_BOUNCE RATE		0.112
RETURNING_ VISITOR		1.573

Table 10.4.12: *Adaptive LASSO Regression Odds Ratio estimates*

VIII. NEURAL NETWORK

The neural network model was set to a maximum of 100 iterations for training optimization and preliminary training was set to “No”. The variables and estimates used the same feature set as the stepwise logistic regression model to implement a feature selection technique into the neural network. The validation F-1 score was 0.8461.

IX. SUPPORT VECTOR MACHINE

Four different SVM models were assessed to identify which kernel function worked best on the dataset. After running the models, the best SVM model used the “Interior Point” as its optimization method and the polynomial kernel function to achieve a validation F1-score of 0.8450.

X. MODEL ASSESSMENT

In the table below, are the final models selected for each algorithm and their validation assessment scores for Accuracy, Precision, Recall, F1-score, and AUC-ROC. The models are ranked in a descending order from the best to worst in terms of validation recall score.

RANK	MODEL NAME	PARTITION SET	ACCURACY	PRECISION	RECALL	F1-SCORE
1	SVM Polynomial	Training	0.8443	0.8484	0.8386	0.8434
		Validation	0.8433	0.8359	0.8543	0.8450
2	Neural Network	Training	0.8433	0.8533	0.8291	0.8410
		Validation	0.8454	0.8422	0.8501	0.8461
3	Gradient Boosting	Training	0.8606	0.8667	0.8522	0.8594
		Validation	0.8438	0.8424	0.8459	0.8441
4	Random Forest	Training	0.8464	0.8573	0.8312	0.8441
		Validation	0.8470	0.8484	0.8449	0.8466
5	PLS Regression	Training	0.8349	0.8668	0.7914	0.8274
		Validation	0.8470	0.8484	0.8449	0.8466
6	Stepwise Regression	Training	0.8370	0.8645	0.7956	0.8286
		Validation	0.8459	0.8603	0.8260	0.8428
7	LASSO Regression	Training	0.8391	0.8664	0.8019	0.8329
		Validation	0.8428	0.8578	0.8218	0.8394
8	Adaptive LASSO Regression	Training	0.8365	0.8681	0.7935	0.8291
		Validation	0.8407	0.8548	0.8208	0.8374
9	Decision Tree	Training	0.8344	0.8633	0.7945	0.8275
		Validation	0.8443	0.8638	0.8176	0.8401

Table 10.4.13: *Final Model Assessment measure comparison & recall ranking.*

After finishing the assessment of each model's performance on the UCI dataset, the conclusions on which model is the best came down to multiple considerations such as their recall, interpretability, and the stability of their assessment scores from the training to validation set. The champion models that were decided upon are the Random Forest and the Adaptive LASSO regression. The Random Forest was chosen as the top model for predicting user conversions due to its stable, consistent scores across all assessment measures. Furthermore, when paired with the cutoff node and reducing the decision cutoff to 0.4, the Random Forest reaches a recall score of 0.9067 while keeping precision at 0.8049. Another factor that came into the decision was the ability to have feature importance rankings and interpretability integrated with the high performance. Additionally, we recommend supplementing the Random Forest with an Adaptive LASSO model to enhance the level of interpretability due to being able to evaluate the standardized estimates of both continuous and categorical features. This plays a vital role in understanding which variables are the most important for predicting e-commerce purchases.

10.4 FEATURE ENGINEERING EVALUATION

Two engineered features were included in the final feature set for modeling: SessionEfficiency and AvgTime_PerPage. After running the final feature set through our modeling pipeline, SessionEfficiency proved to be a significant predictor of conversions and valuable addition to the model. Specifically, the binned version of SessionEfficiency was identified as the most important feature for predicting purchases in nearly all tree-based and linear models. This discovery displayed that SessionEfficiency can be treated as a critical differentiator when identifying whether a user will complete a purchase. The evidence to support this finding will be displayed in the feature importance section.

10.5 FEATURE IMPORTANCE

To understand the most influential factors in predicting online shopping purchases, we compared the feature importance rankings across multiple models that included both tree-based and linear algorithms. The analysis incorporates standardized estimates, odds ratio estimates, and decision-based importance scores. In the table below, a final feature importance ranking has been compiled that considers all the outcomes of the models.

FEATURE	OVERALL RANK	TREE-BASED RANK	LINEAR MODEL RANK	INTERPRETATION
SESSIONEFFICIENCY	1	1	1	Users who have an efficient navigation path are highly likely to purchase.
PAGEVALUE	2	2	3	Higher page values lead to much higher likelihood of purchases.
PRODUCTRELATED_ DURATION	3	3	2	Higher durations of time spent on product pages leads to a higher likelihood of purchasing.
BOUNCERATE	4	6	4	High bounce rates greatly reduce the likelihood of conversion.
TRAFFIC TYPE	5	5	5	Certain routes taken to reach a site have much higher odds of a purchase occurring.
Q4	6	4	6	Seasonal Indicator of purchasing behavior.
ADMINISTRATIVE	7	7	7	More visits to administrative pages can indicate potential purchases.
RETURNING_VISITOR	8	8	8	If a user is returning to a site, it's associated with a higher likelihood of purchases.

Table 10.5.1: *Feature importance rankings developed from modeling assessment*

SessionEfficiency came out as the top predictor of conversions due to consistent top ranks across all models. SessionEfficiency (binned: “low”) had extremely low odds ratios across all linear models (ex: LASSO: 0.008, SW Reg: 0.021) indicating that users with low efficiency were extremely unlikely to convert to purchasers. Furthermore, it had the highest standardized estimate coefficient (-0.73108) further proving its ability to differentiate between users.

PageValue was a consistently strong predictor of conversions, ranking within the top three in each model it was selected as a feature in. Its odds ratio in linear models like LASSO (1.313) indicated that for each additional unit of PageValue, the odds of purchase are 31.3% higher. To support this further, it also had a standardized estimate of 0.2499 in the LASSO regression.

ProductRelated_duration demonstrated a clear effect across both linear and tree-based models. Users in the low engagement bin (ProductRelated_duration 01: Low) have significantly lower odds of conversion (Adaptive LASSO: 0.386) when comparing it to the high engagement bin. Furthermore, both low and mid engagement bins for the feature had standardized estimates (Low: -0.10458, Mid: -0.10581) that indicated it was influential to the target variable.

BounceRate is aligned with previous findings found in the literature that indicate it is negatively associated with the target. In linear models, it maintained low odds ratio estimates (SW Reg: 0.099, Adaptive LASSO: 0.112) which indicates that users with high bounce tendencies were far less likely to convert to purchasers. Furthermore, its standardized estimate confirmed this with -0.2558 in the Stepwise Regression and very similar values in the other models.

When encoded using Smoothed Weight of Evidence (SWOE), TrafficType proved to be a valuable predictor of conversion in linear models and the original variable further backed this up as a feature in all tree-based models. The odds ratio estimates for TrafficType were significantly above 1.0 (Adaptive LASSO: 2.183), indicating that certain traffic sources were far more likely to complete a purchase. Furthermore, the standardized estimates (Adaptive LASSO: 0.1803) were consistent with these findings and proved that traffic sources are a valuable addition to feature sets for predicting conversions.

Returning_Visitor demonstrated a positive relationship with conversion likelihood. In most linear models, the odds ratio ranged from 1.25 to 1.57, indicating that repeat visitors had higher chances of purchasing. While it did not rank among the top features for tree-based models, its presence in both logistic regression and LASSO models suggests it remains a useful segmentation tool when paired with session duration metrics.

10.6 TECHNOLOGY ACCEPTANCE MODEL ANALYSIS

To enhance the interpretation of the modeling results and give them theoretical backing, the findings of this study were evaluated through the lens of the Technology Acceptance Model (TAM). TAM suggests that a user's decision to engage with or adopt a digital system is primarily rooted in two constructs: Perceived Usefulness and Perceived Ease of Use. Both constructs influence user behavior and usage, and in our case conversions in e-commerce. The table below outlines the link between our model assessment and findings with TAM.

FEATURE	TAM CONSTRUCT	INTERPRETATION
SESSIONEFFICIENCY	Perceived Ease of Use	High efficiency reflects smooth, frictionless experiences that encourage conversion.
PAGEVALUE	Perceived Usefulness	A higher PageValue suggests the user is receiving content or offers of value.
PRODUCTRELATED_ DURATION	Perceived Usefulness	More time spent on product pages reflects deeper engagement and relevance (usefulness) of content.
BOUNCERATE	Perceived Ease of Use	High bounce rates may indicate difficulty or dissatisfaction, reducing the likelihood of adoption.
TRAFFICTYPE	Perceived Usefulness	Certain traffic sources like paid ads and direct search often deliver users to goal-relevant or high value pages. This can increase perception of perceived usefulness.
RETURNING_VISITOR	Perceived Usefulness	Repeat visits indicate prior positive experiences that align with systems giving users value.

Table 10.6.1: *Features & findings link with TAM*

11. RESEARCH DELIVERABLES AND CONTRIBUTION

This study aimed to create a link between data-driven modeling and behavioral theory to better understand the drivers of online shopping conversions. By integrating machine learning techniques, data preprocessing & engineering, and behavioral frameworks such as the Technology Acceptance Model, several key research deliverables and contributions were met. The deliverables are outlined in the table below.

Research Deliverables	Impact
A well-documented predictive modeling pipeline to classify purchases	Provides a high-performing, reproducible ML framework
Performance comparison of different data preprocessing & feature engineering techniques	Provides understanding for best preprocessing & engineering methods
Interpretable Algorithms & Results	Filling a research gap, due to the lack of interpretability tools in prior research
Use of TAM as a behavioral lens to interpret model findings	Adds a theoretical layer to support findings and explain why features matter
Business Insights for e-commerce behavior & engagement.	Connection between data science & conversion optimization

Table 11.1: *Research Deliverables & contribution*

I. DEVELOPMENT OF A PREDICTIVE MODELING PIPELINE

Using session-level data from the UCI Machine Learning Repository and advanced machine learning techniques, a series of predictive models were developed to identify user conversions in online shopping. The Random Forest model was able to attain stable, consistent scores in both training and validation sets, while achieving a strong recall score of 0.9067 and balancing the precision at 0.8049 when reducing the decision cutoff to 0.4. Additionally, supplementing that with an Adaptive LASSO model gives you a solid grounding for both predicting conversions and interpretating what influences them.

II. FEATURE ENGINEERING FOR BEHAVIORAL INSIGHT

New features such as Session Efficiency and optimally binning ProductRelated_Duration were engineered to better reflect the quality of user engagement. These features not only enhanced the performance of the models, but also provided interpretable insights into how efficiently users navigate and how much value they generate per session.

III. ASSESSMENT OF DATA PREPROCESSING TECHNIQUES

This study compared the approaches of applying Log transformations and binning techniques to highly skewed, long-tailed continuous variables common in the e-commerce industry. Additionally, it displayed the application of Smoothed Weight of Evidence (SWOE) to high cardinality variables such as TrafficType that have not been performed on the dataset in previous work. These preprocessing techniques improved recall and interpretability in linear models and proved that traffic sources are influential when predicting user conversions.

IV. BEHAVIORAL FRAMEWORK INTEGRATION USING TAM

By linking key predictive features (SessionEfficiency, PageValue, etc.) to Perceived Usefulness and Perceived Ease of Use, this research demonstrated how behavioral modeling can be interpreted through TAM constructs. This hybrid approach enhances both predictive performance and theoretical understanding of user behavior.

12. CONCLUSION

This study presented a comprehensive approach to predicting and interpreting online shopping user conversions by combining advanced machine learning techniques with behavioral theory. By leveraging session-level website traffic data, multiple predictive models were developed and assessed, including tree-based algorithms (Random Forest, Gradient Boosting), linear algorithms (Logistic Regression, LASSO, PLS), and nonlinear algorithms (SVM, Neural

Network). Among these, the Random Forest model achieved the highest overall performance with a 0.9067 recall score and strong generalization between training and validation sets.

The research introduced a novel engineered feature, SessionEfficiency, which consistently proved to be the top predictor across all models. Other key features such as PageValue, ProductRelated_Duration, and BounceRate aligned with previous studies and further proved the importance of user engagement metrics in predicting conversions. The traffic source feature TrafficType was also discovered to be significant when paired with the appropriate preprocessing techniques like SWOE.

To enhance the interpretation of results, the study linked key predictive features to the Technology Acceptance Model (TAM). This behavioral framework conceptualized model outputs with two key constructs: Perceived Usefulness and Perceived Ease of Use. The alignment between predictive variables and TAM validated the behavioral logic of the findings and also provides a behavioral lens for applying insights to business practices.

In summary, this research contributes robust, interpretable, and theoretically backed solutions for predicting online shopping purchase conversions. It demonstrates the value of combining data science with behavioral theory to improve predictive performance and decision-making for digital marketing and e-commerce optimization.

13. APPENDIX OF TABLES & FIGURES

13.1 APPENDIX A: TABLES

13.2 APPENDIX B: FIGURES

14. REFERENCES

Adhikari, S. (2023). Predicting customer purchase intention using ensemble models: XGBoost, LightGBM, and SMOTE on the UCI dataset. International Journal of Applied Machine Learning, 22(4), 150-165. https://doi.org/10.1145/3469921.3470410

Ahsain, S., & Ait Kbir, M. (2022). Predicting the client's purchasing intention using machine learning models. E3S Web of Conferences, 351, 01070. https://doi.org/10.1051/e3sconf/202235101070

Granic, Andrina & Davis, F. D.(2022). *The Technology Acceptance Model*. SpringerLink. https://link.springer.com/book/10.1007/978-3-030-45274-2

Baati, K., & Mohsil, M. (2020). Real-Time prediction of online shoppers' purchasing intention using Random Forest. In *IFIP advances in information and communication technology* (pp. 43–51). https://doi.org/10.1007/978-3-030-49161-1_4

Chatterjee, S., & Kumar Kar, A. (2020, March 19). Why do small and Medium Enterprises use social media marketing and what is the impact: Empirical insights from India. International Journal of Information Management. https://www.sciencedirect.com/science/article/abs/pii/S0268401219316676?via%3Dihub

Chatterjee, S., & Kumar Kar, A. (2020, March 19). Why do small and Medium Enterprises use social media marketing and what is the impact: Empirical insights from India. International Journal of Information Management. https://www.sciencedirect.com/science/article/abs/pii/S0268401219316676?via%3Dihub

Frazier, A., Li, X., Chen, Y., & Maloku, F. (2022). *(PDF) Data Analysis of Online Shopper's purchasing intention machine learning for prediction analytics*. Data Analysis of Online Shopper's Purchasing Intention Machine Learning for Prediction Analytics. https://www.researchgate.net/publication/364072790_Data_Analysis_of_Online_Shopper's_Purchasing_Intention_Machine_Learning_for_Prediction_Analytics

Islam, M. S., Naeem, J., Emon, A. S., Baten, A., Mamun, M. A., Waliullah, G. M., Rahman, M. S., & Mridha, M. F. (2023). Prediction of Buying Intention: Factors Affecting Online Shopping. *2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM*, *4*, 1–6. https://doi.org/10.1109/ncim59001.2023.10212766

Jiang, Y. (2022). *Research on prediction of e-commerce repurchase behavior based on multiple fusion models*. The 4th International Conference on Computing and Data Science (CONF-CDS). https://doi.org/10.54254/2755-2721/2/20220555

Ketipov, R., Vladimirov, M., & Nikolov, P. (2023). Predicting purchase intentions on website features: The role of Big Five personality traits and e-commerce usability factors. Journal of Behavioral Science and Technology, 12(2), 98-112. https://doi.org/10.1016/j.jbst.2023.05.002

Kurniawan, I., Abdussomad, N., Akbar, M. F., Saepudin, D. F., Azis, M. S., & Tabrani, M. (2020). Improving The Effectiveness of Classification Using The Data Level Approach and Feature Selection Techniques in Online Shoppers Purchasing Intention Prediction. Journal of Physics Conference Series, 1641(1), 012083. https://doi.org/10.1088/17426596/1641/1/012083

Nyongesa, D. (2020). Variable Selection Using Random Forests in SAS®. Proceedings of the SAS Global Forum 2020. https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2020/4826-2020.pdf

Purwianti, L., Yulianto, E. (2024). The Influence of Social Presence on Purchase Intention in Social Commerce: The Mediating Roles of Trust and Attitude. International Journal of Applied Research in Business and Management, 5(2). https://doi.org/10.51137/wrp.ijarbm.2024.lpts.45613

Rajamma, R. K., Paswan, A. K., & Hossain, M. M. (2009). Why do shoppers abandon shopping carts? Perceived waiting time, risk, and transaction inconvenience. *Journal of Product & Brand Management, 18*(3), 188–197.

Rana, N., Patel, A. K., Singh, A., Parayitam, S., Dwivedi, Y., & Dutot, V. (2023, April 26). Assessing customers' attitude towards online apparel shopping: A three-way interaction model. Journal of Business Research. https://www.sciencedirect.com/science/article/abs/pii/S0148296323002758?via%3Dihub

Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2018). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-018-3523-0

SAS Institute Inc. (n.d.). *Adaptive LASSO Selection - SAS Help Center*. Retrieved from https://documentation.sas.com/doc/en/statug/15.2/statug_glmselect_details11.htm

SAS Institute Inc. (n.d.). *Gradient Boosting Node - SAS Help Center*. Retrieved from https://documentation.sas.com/doc/en/emref/15.2/n0t6j7sk2xn3mon1e7ulvypjppew.htm

Satu, M. S., & Islam, S. F. (2023). Modeling online customer purchase intention behavior applying different feature engineering and classification techniques. *Discover Artificial Intelligence*, *3*(1). https://doi.org/10.1007/s44163-023-00086-0

Shi, X. (2021). *The Application of Machine Learning in Online Purchasing Intention Prediction*. *International Conference on Big Data*. https://doi.org/10.1145/3469968.3469972

Song, P., & Liu, Y. (2020). An XGBoost algorithm for predicting purchasing behaviour on e-commerce platforms. *Tehnički vjesnik – Technical Gazette*, *27*(5), 1467–1471. https://doi.org/10.17559/TV-20200808113807

Sunarto, A., Kencana, P. N., Munadjat, B., Dewi, I. K., Abidin, A. Z. A., & Rahim, R. R. (2023). Application of Boosting Technique with C4.5 Algorithm to Reduce the Classification Error Rate in Online Shoppers Purchasing Intention. *Journal of Wireless Mobile Networks Ubiquitous Computing and Dependable Applications*, *14*(2), 01–11. https://doi.org/10.58346/jowua.2023.i2.001

Tayal, K., & Daniel, S. (2024). Predicting the customer purchase intention based on perception of risk using machine learning techniques. *European Economics Letters, 14*(1), 1803–1813. https://doi.org/10.52783/eel.v14i1.1252

Torres, D., & Cepeda, L. K. (2024). (PDF) machine learning for predicting online shoppers' purchase intentions. Machine Learning for Predicting Online Shoppers' Purchase Intentions. https://www.researchgate.net/publication/381185713_Machine_Learning_for_Predicting_Online_Shoppers'_Purchase_Intentions

UCI Machine Learning Repository. (2024). Online Shoppers' Purchasing Intention Dataset. Retrieved from https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset

Wen, Z., Lin, W., & Liu, H. (2023). Machine-learning-based approach for anonymous online customer purchase intentions using clickstream data. Systems, 11(5), 255. https://doi.org/10.3390/systems11050255