



# UNIVERSITY OF CAPE TOWN

DEPARTMENT OF STATISTICAL SCIENCES

STA5076Z - Supervised Learning Assignment 2

MSHTSA009

May 8, 2024

# Contents

Introduction	1
Data description and pre-processing	1
Modelling	5
Model evaluation	13
Optimising for MCC	14
References	15

# Analysis of Online Shoppers' Purchase Intentions

## Introduction

In the rapidly evolving landscape of e-commerce, understanding and predicting online shoppers' behaviors and purchase intentions has become a focal point of both academic research and practical applications in digital marketing. The emergence of sophisticated machine learning techniques has significantly enhanced the accuracy and efficiency of these predictions, leading to a more personalized shopping experience for consumers.

The dataset analyzed in this study, first introduced by Sakar et al. (2018), contains a comprehensive set of 17 features capturing various aspects of an online shopping session, such as page views, bounce rates, and time spent on the website. These features collectively provide a deep insight into user engagement and the likelihood of transaction completion. Following the foundational work by Sakar et al., subsequent studies, including that of Abdullah-All-Tanvir et al. (2023), have extended the analysis, leveraging advanced predictive models to enhance the understanding of consumer behavior in digital environments.

This report aims to replicate some of the key findings of Abdullah-All-Tanvir et al., focusing on the predictive accuracy of different machine learning models in determining whether online shoppers will finalize their purchases. By adhering to rigorous data handling protocols, the analysis ensures the integrity and reproducibility of results. The primary goal here is not just to achieve high accuracy but also to maintain transparency in methodology and clarity in the presentation of the findings, thus contributing valuable insights to the field of e-commerce analytics. This introduction sets the stage for a detailed exploration of supervised learning applications in predicting online shopping behaviors, ensuring that the analysis is both scientifically rigorous and practically relevant to the field of digital marketing and consumer behavior analysis.

## 1. Data description and pre-processing

### Data Description

The dataset contains 12,330 observations and 18 variables, it comprises a mixture of numerical and categorical features, reflecting various aspects of an online browsing session aimed at understanding and predicting online shopping behaviors. These features include:

- Administrative, Informational, ProductRelated: Counts of different types of pages visited.
- Administrative.Duration, Informational.Duration, ProductRelated.Duration: Total time spent on different types of pages.
- BounceRates, ExitRates, PageValues: Metrics that describe the engagement and value of the pages visited.
- SpecialDay: Closeness of the site visiting time to a special day (e.g., Mother's Day).
- Month, OperatingSystems, Browser, Region, TrafficType: Categorical variables representing the month, operating system used, browser type, visitor's region, and traffic type.
- VisitorType: Type of visitor (e.g., Returning Visitor, New Visitor).
- Weekend: Whether the visit occurred on a weekend.

- Revenue: Whether the session ended in a transaction (binary outcome).

## Data Preprocessing

1. **Reading the Data:** The dataset 'online\_shoppers\_intention.csv' is loaded into R, maintaining all features as indicated in the Sakar et al. (2018) study without prior editing.
2. **Initial Data Inspection:** Utilize `str(data)` and `summary(data)` to gain an initial understanding of data structure and summary statistics. The absence of missing values (**NAs**) across all features was confirmed, which aligns with the dataset's description in Sakar et al. (2018).
3. **Feature Exploration:** Feature exploration was a critical initial step aimed at understanding the influences and dynamics within the dataset. Key insights include:

- **Page Values and Durations:** Higher values and longer durations on specific pages correlate strongly with increased likelihood of purchase, reflecting deeper engagement.
- **Categorical Variables:** Factors such as the month of visit and type of visitor were transformed to enhance the model's interpretive power.

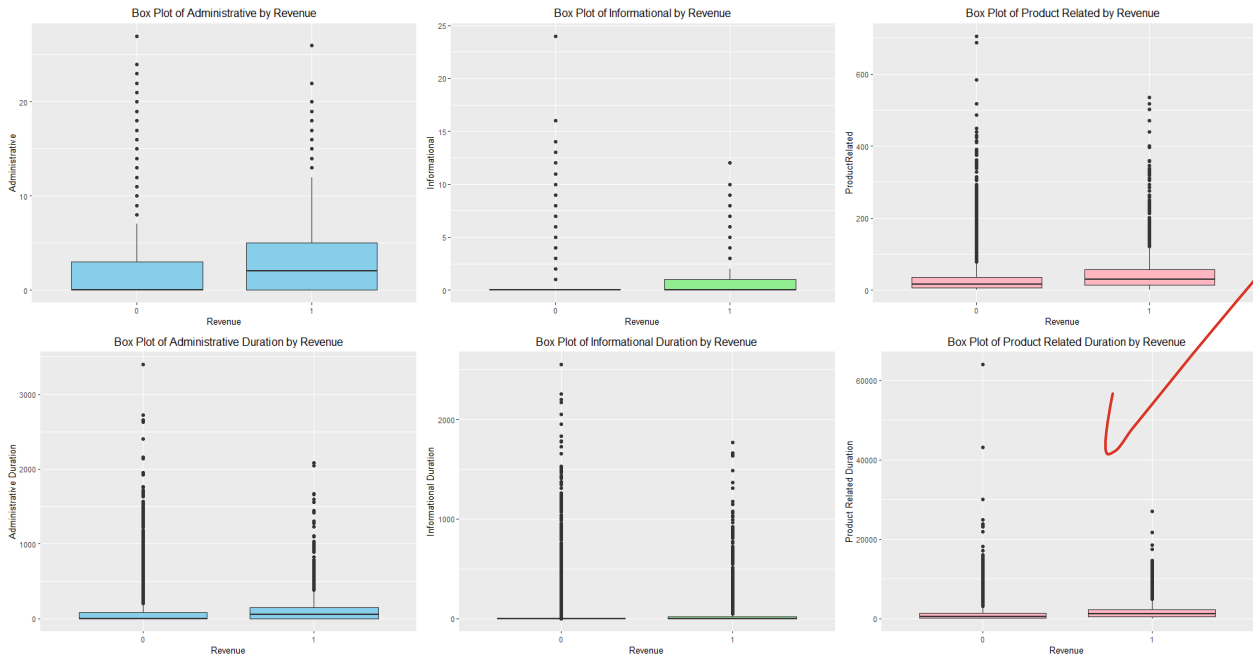


Figure 1: **Box Plots of Page Interaction Types by Revenue Outcome.** This figure presents a series of box plots illustrating the distribution of page interactions and their durations in relation to revenue generation. Each box plot contrasts metrics for sessions that resulted in purchases (Revenue = 1) versus those that did not (Revenue = 0). The metrics analyzed include the count and duration of administrative, informational, and product-related pages.

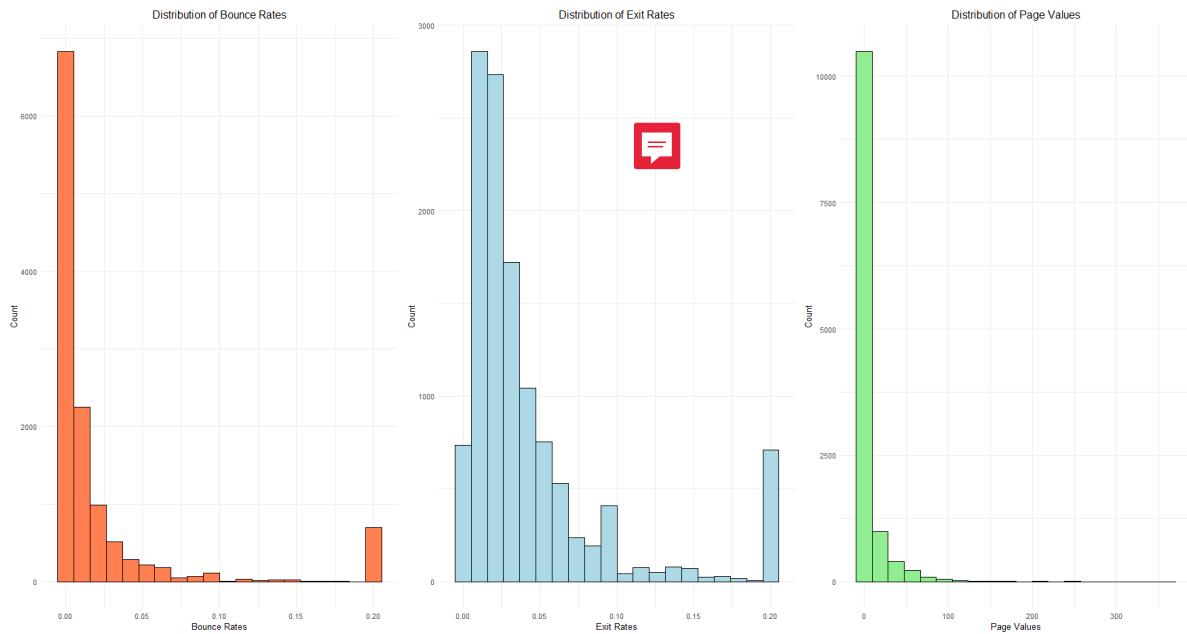


Figure 2: **Distribution of Key Website Metrics Across Sessions.** This figure illustrates the distribution of three critical metrics observed during online shopping sessions: Bounce Rates, Exit Rates, and Page Values. Each histogram represents the frequency of sessions corresponding to different ranges of these metrics, showcasing the variability in user engagement and page value during their browsing experience.

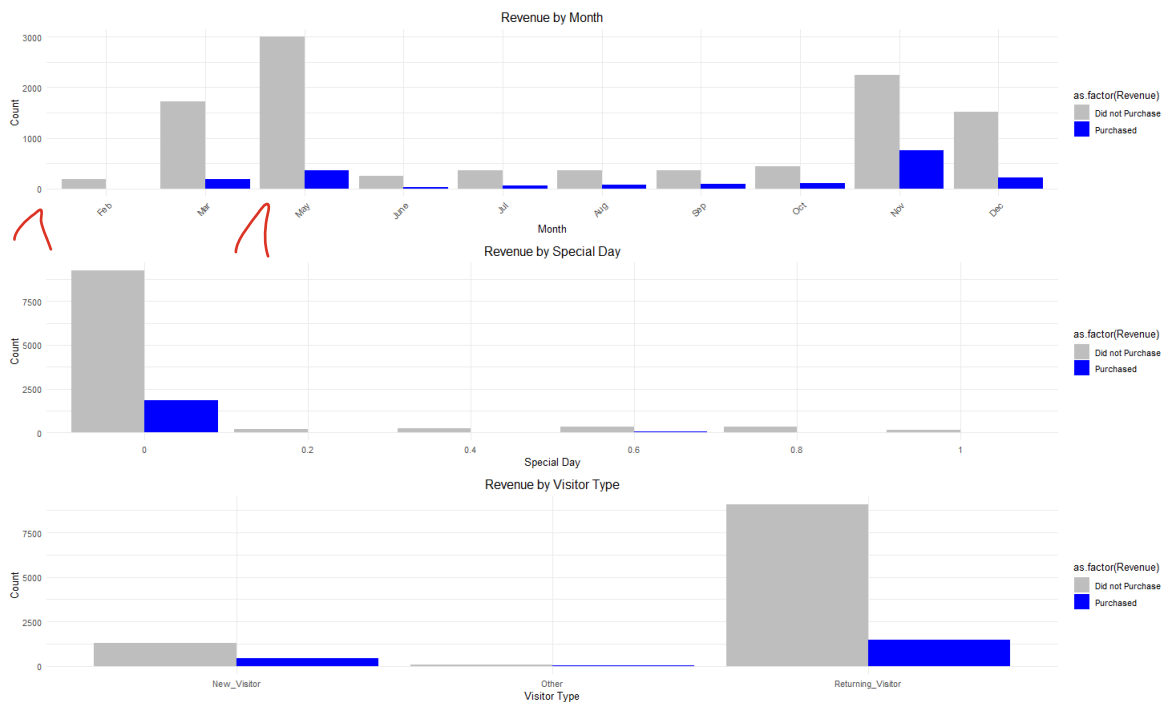


Figure 3: **Seasonal and Demographic Influences on Revenue Generation.** This figure displays bar graphs analyzing the impact of various factors on revenue generation, including the month of visit, proximity to special days, and visitor types. The graphs depict the count of shopping sessions, distinguishing between those that led to purchases and those that did not, across different time frames and visitor demographics.

#### 4. Handling Categorical Variables:

- Convert the 'Month' and 'VisitorType' to factors to treat these variables as categorical in the model
- Most of the dataset is numerical, either integers or floats; Revenue and Weekend are boolean type, and they can easily be transformed into binary type (0 & 1).

5. **Data Split:** Data is split into training and testing sets using a commonly recommended ratio of 70:30 for training and testing respectively. This split ensures adequate data for learning while also providing enough data for a robust evaluation.

6. **Data Scaling:** Numeric features are like 'Administrative Duration', 'Informational Duration', and 'Product Related Duration' are scaled in the training data to have zero mean and unit variance. This step is essential for models that are sensitive to the scale of input features, such as logistic regression.

There are no discrepancies observed in the feature descriptions or the data structure between the paper's description and the actual dataset provided. The data characteristics and features used for the analysis in the paper match those in the dataset you uploaded, confirming the fidelity of the dataset to the study's reported structure.

Excellent intro!

## 2. Modelling

### 2.1 Logistic Regression

#### Logistic Regression with All Features:

In this logistic regression model, we aimed to predict the likelihood of a purchase (Revenue) based on various features derived from online browsing sessions. Here's a discussion of the significant coefficients:

1. **Intercept:** The intercept represents the log odds of the baseline scenario where all predictor variables are zero. In this case, it's significantly negative (-1.727), indicating a lower probability of purchase when all other predictors are zero.
2. **PageValues:** This coefficient (1.605) is highly significant with a very low p-value ( $p < 0.001$ ), suggesting that PageValues have a substantial positive effect on the likelihood of purchase. This aligns with the intuitive understanding that pages with higher values are more likely to lead to purchases.
3. **ExitRates:** The coefficient (-0.817) for ExitRates is significant ( $p < 0.001$ ) and negative, indicating that higher ExitRates decrease the likelihood of a purchase. This implies that users who exit the website more frequently without making a purchase are less likely to convert.
4. **ProductRelated\_Duration:** This coefficient (0.138) is significant ( $p = 0.035$ ) and positive, suggesting that longer durations spent on product-related pages increase the likelihood of a purchase. It indicates that users who spend more time engaging with product-related content are more likely to make a purchase.
5. **MonthDec, MonthMar, MonthMay, MonthNov:** These coefficients represent the effect of different months on purchase likelihood compared to the reference month (January). MonthDec, MonthMar, MonthMay, and MonthNov have negative coefficients, indicating that these months are associated with a decreased likelihood of purchase compared to January.
6. **Weekend:** The coefficient (0.178) for Weekend is significant ( $p = 0.034$ ) and positive, suggesting that visits occurring on weekends are associated with a slightly higher likelihood of purchase compared to weekdays.
7. **VisitorTypeReturning\_Visitor:** The coefficient (-0.287) is significant ( $p = 0.006$ ) and negative, suggesting that returning visitors have a slightly lower likelihood of purchase compared to new visitors.

Overall, the model's null deviance (7431.9) compared to its residual deviance (4952.4) and the associated AIC (Akaike Information Criterion) value (5006.4) indicate that the model provides a significant improvement in explaining the variance in the response variable compared to the null model. However, the model's performance can be further evaluated using additional metrics and compared to other models for better insights into its predictive capability.

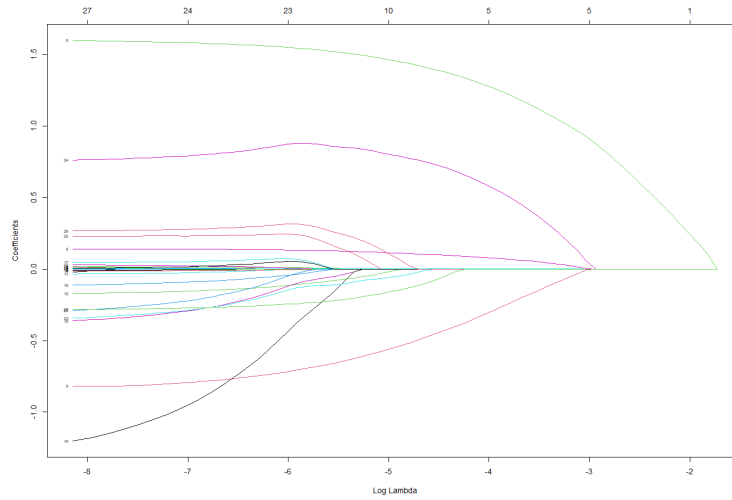
#### Logistic Regression with L1-Regularization (Lasso):

The lasso logistic regression model was fitted to perform variable selection and regularization, potentially improving model interpretability and generalization.

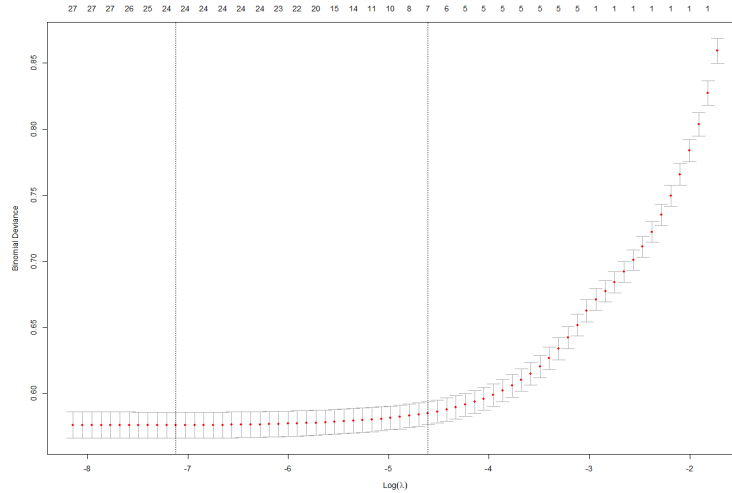
- **Lasso Coefficients:**

- The lasso model shrinks some coefficients towards zero, effectively performing variable selection by reducing the impact of less important predictors.

- Coefficients that are shrunk to zero are effectively eliminated from the model, leading to a sparse model with fewer predictors.
- **Interpretation:**
  - The significant predictors identified in the lasso model largely align with those from the logistic regression model with all features. Notably, **PageValues**, **ExitRates**, and various months (e.g., **MonthDec**, **MonthMar**, **MonthMay**, and **MonthNov**) retain their significance in the lasso model.
  - Some predictors are eliminated from the model due to regularization, such as **MonthOct** and **VisitorTypeOther**, which have coefficients of zero.



(a) Lasso Path Plot for Model Coefficients



(b) Binomial Deviance across Different Lambda Values

Figure 4: **(a)** The upper plot shows the Lasso path for model coefficients, illustrating the effect of increasing regularization on the coefficients of the logistic regression model. **(b)** The lower plot presents the binomial deviance across different lambda values, with error bars showing variability across cross-validation folds. The optimal lambda, minimizing the deviance, is indicated by a vertical dotted line.



Table 1: Comparison of Coefficients between Logistic Regression Models



Predictor	Logistic Regression (All Features)	Lasso (L1-Regularization)
Administrative	-0.006	-0.002
Administrative_Duration	0.018	0.011
Informational	0.019	0.016
Informational_Duration	0.008	0.007
ProductRelated	0.001	0.001
ProductRelated_Duration	0.138	0.139
BounceRates	-0.060	-0.018
ExitRates	-0.817	-0.816
PageValues	1.605	1.595
SpecialDay	-0.122	-0.110
OperatingSystems	-0.039	-0.034
Browser	0.031	0.028
Region	-0.011	-0.010
TrafficType	-0.014	-0.013
Weekend	0.178	0.169
MonthAug	-0.220	0.033
MonthDec	-0.659	-0.368
MonthFeb	-1.597	-1.189
MonthJul	-0.035	0.216
MonthJune	-0.232	0.002
MonthMar	-0.587	-0.296
MonthMay	-0.632	-0.350
MonthNov	0.484	0.752
MonthOct	-0.261	0
MonthSep	NA	0.256
VisitorTypeOther	-0.002	0
VisitorTypeReturning_Visitor	-0.287	-0.282

Factors!

This comparison highlights the consistency in significant predictors between the two models. Both models offer valuable insights into the predictors influencing online shoppers' purchase intentions. The logistic regression model with all features provides a comprehensive understanding of the predictors' effects, while the Lasso model selects a subset of predictors, enhancing model interpretability and potentially reducing overfitting. By effectively shrinking some coefficients to zero, **the Lasso model** achieves a more concise representation with fewer predictors, aiding in model interpretability and potentially improving generalization performance.

## 2.2 Classification Tree Model

The classification tree model was built to predict revenue generation based on online shopping behavior features. After cross-validation, the optimal complexity parameter (CP) for pruning was found to be approximately **0.003998**. Pruning the tree with this CP value simplified its structure, enhancing interpretability without sacrificing predictive power.

The pruned tree revealed key predictors of revenue generation:

- **PageValues:** Pages with higher page values significantly contribute to revenue generation.
- **ExitRates and Administrative:** These features further refine the prediction, with different thresholds

indicating whether revenue is likely to be generated or not.

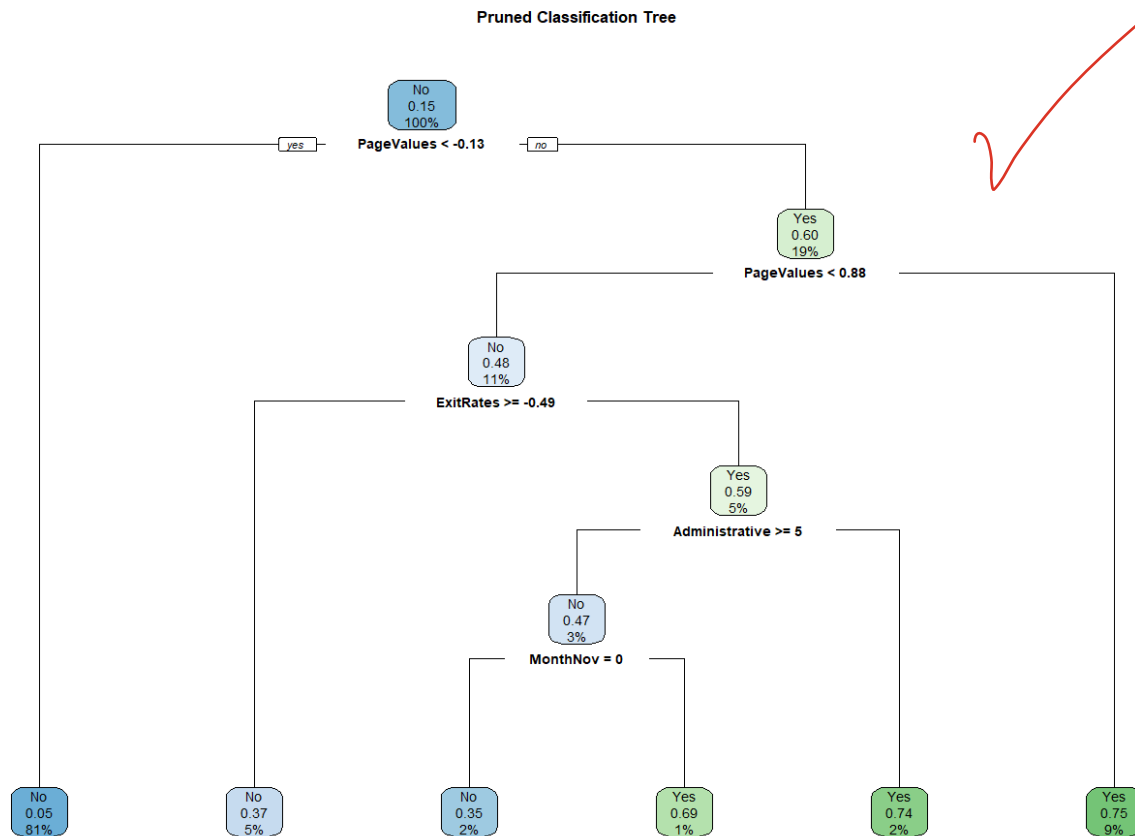


Figure 5: Decision Tree Visualization for Revenue Prediction


The tree structure shows the decision nodes (splitting criteria) and terminal nodes (final classification outcomes).

- Root Node: The initial node of the tree with 8631 observations. The majority class is "No" (revenue not generated), indicating that most observations in the dataset do not result in revenue generation.
- Internal Nodes: Decision nodes based on feature thresholds. For example, PageValues less than -0.1269001 is a splitting criterion in one branch.
- Terminal Nodes: Leaf nodes where classification decisions are made. Each terminal node represents a final classification outcome (Yes or No) along with the probability distribution of classes.

## 2.3 Random Forest Model

The random forest model was fitted to the training dataset with the following hyperparameters:

- Number of trees (**ntree**): 500 *Need to show it's enough.*
  - Setting ntree to 500 provides a robust ensemble of trees, reducing the risk of overfitting while capturing complex relationships in the data.

- Number of variables tried at each split (~~mtry~~): 5 (square root of the total number of features)
  - Choosing the square root of the total number of features as mtry ensures diversity in feature selection at each split, leading to improved model generalization. 

### Model Summary:

- Type: Classification
- Number of trees: 500
- Out-of-bag (OOB) estimate of error rate: 9.54%

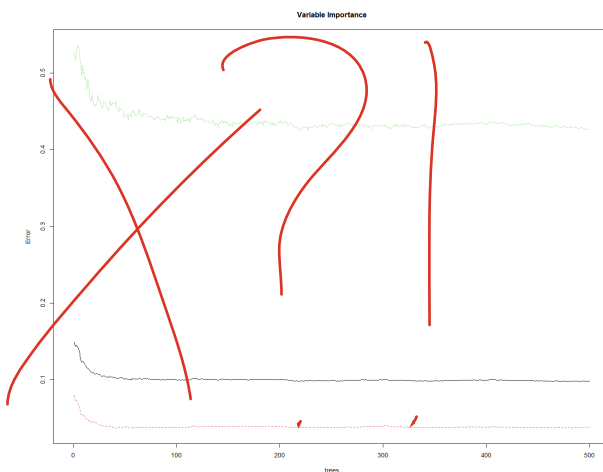
The confusion matrix summarizes the model's performance:

	Predicted No	Predicted Yes
Actual No	7044	253
Actual Yes	570	764

### Class Error:

- Class error for "No" class: 3.47%
- Class error for "Yes" class: 42.73%

The random forest model achieved an out-of-bag error rate of 9.54%, indicating its robust performance in predicting revenue generation based on online shopping behavior features. The confusion matrix provides insights into the model's classification accuracy for both revenue-generating and non-revenue-generating instances.



(a) Variable Importance Plot



(b) Partial Dependence Plot for PageValues

Figure 6: **(a)** ranks the features based on their importance in predicting the outcome (purchase or no purchase). High importance suggests that changes in these features are strongly associated with changes in the likelihood of a purchase. **(b)** illustrates the effect of the most influential variables on the probability of purchase. By showing how changes in a feature like 'PageValues' affect the likelihood of generating revenue, the plot provides insights into how different levels of this feature correlate with purchasing behavior.

## 2.4 Gradient Boosted Trees

### a. Gradient Boosted Trees using GBM

- **Model Training:**

- The Gradient Boosted Trees model using the GBM (Generalized Boosted Regression Models) algorithm was trained on the provided dataset.
- The model was trained with a total of 500 trees, an interaction depth of 4, and a shrinkage value of 0.01.
- The distribution was set to "bernoulli" as it is suitable for binary classification tasks.

- **Variable Importance:**

- Variable importance analysis provides insights into the contribution of each feature to the model's predictive performance.

Table 2: Variable Importance Analysis Results for Gradient Boosted Trees using GBM

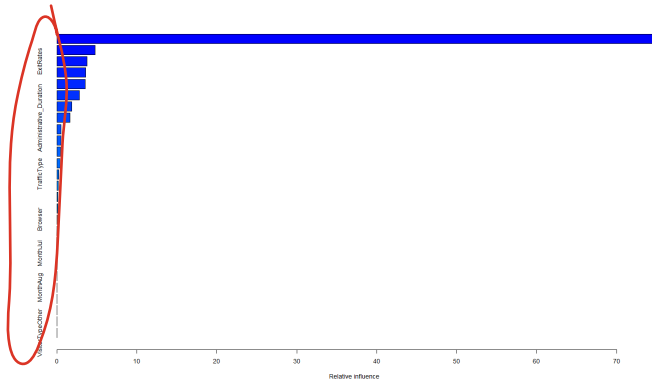
Variable	Relative Importance (%)
PageValues	75.19
MonthNov	4.76
ExitRates	3.74
Administrative	3.62
ProductRelated	3.57
ProductRelated_Duration	2.80
BounceRates	1.87
Administrative_Duration	1.66
VisitorTypeReturning_Visitor	0.51
MonthMar	0.49
MonthMay	0.45
Informational_Duration	0.40
TrafficType	0.22
Informational	0.16
Region	0.15
OperatingSystems	0.13
Browser	0.09
Weekend	0.07
MonthSep	0.07
MonthJul	0.03
MonthDec	0.02
SpecialDay	0.01
MonthAug	0.00
MonthFeb	0.00
MonthJune	0.00
MonthOct	0.00
VisitorTypeOther	0.00

*Rather plot*

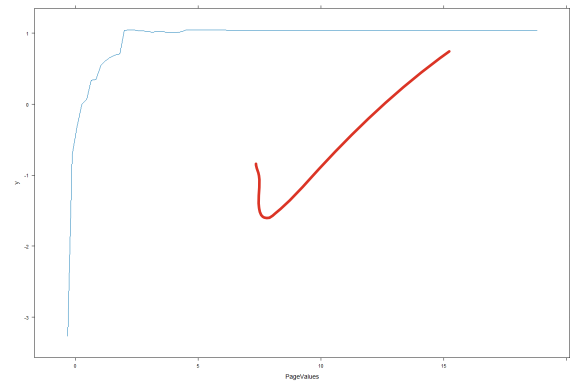
- The "Relative Importance" column represents the relative influence of each feature on the model's predictions.
- Key findings from the variable importance analysis:

Unnecessary

- \* **PageValues:** The most important feature, with a relative influence of approximately 75%. This indicates that the average value of pages visited by the user significantly impacts their likelihood of making a purchase.
- \* **MonthNov:** Indicates that the month of November has a relatively high importance in predicting purchase intentions, with around 4.76% relative influence.
- \* **ExitRates, Administrative, ProductRelated:** These features also show significant importance, contributing to the model's predictive power.
- \* **VisitorTypeReturning\_Visitor, MonthMar, MonthMay, Informational\_Duration, TrafficType:** While less influential compared to the top features, they still contribute to the model's performance.



(a) Variable Importance Plot



(b) Partial Dependence Plot

Figure 7: (a) The Variable Importance plot identified 'PageValues' as the most influential predictor, showcasing its substantial impact on purchase decisions. This suggests that the economic value of visited pages strongly indicates purchasing intent, likely due to their effectiveness in engaging users and leading them to transaction. (b) for 'PageValues' illustrates how changes in this feature affect the predicted probability of a purchase. It depicts a positive relationship between PageValues and purchase likelihood, emphasizing the importance of optimizing page content to maximize its value, as it directly correlates with higher conversion rates.

### • Hyperparameter Tuning:

- The hyperparameters such as the number of trees, interaction depth, shrinkage, and minimum observations in a node were tuned to optimize the model's performance.
- A cross-validation strategy with 5 folds was employed to ensure robustness and avoid overfitting.

### Conclusion:

- The Gradient Boosted Trees model using the GBM algorithm demonstrates promising results in predicting online shoppers' purchase intentions.
- Features such as PageValues, MonthNov, and ExitRates play significant roles in determining whether a shopper will finalize a transaction.
- Further fine-tuning of hyperparameters and exploration of additional features could potentially enhance the model's performance and provide deeper insights into online shopping behavior.

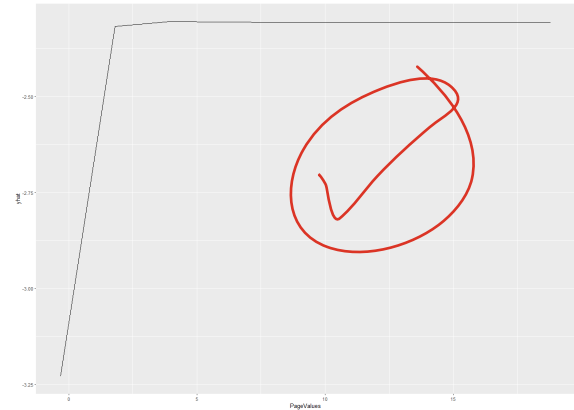
## b. XGBoost Model

The model was trained over 1000 iterations with the log loss decreasing to 0.0003886830 on the training set and 0.0003907794 on the evaluation set by the final iteration. These low values indicate a very good model fit, with no apparent signs of overfitting, as the evaluation and training log losses are almost identical.

### Variable Importance and Partial Dependence Plot



(a) Variable Importance Plot



(b) Partial Dependence Plot

Figure 8: **(a)** highlights 'PageValues,' 'ExitRates,' and 'MonthNov' as influential features in predicting purchase likelihood, guiding targeted optimizations for improved user engagement. **(b)** illustrates a positive correlation between 'PageValues' and purchase probability, emphasizing the need to enhance page content to boost its value and drive purchasing behavior.

### Variable Importance Analysis

#### • PageValues:

- Gain: The Gain is about 9.6%, showing a high impact on the model's decisions, which aligns with its role in influencing user purchase decisions.

#### • ExitRates:

- Gain: 0.86%, a moderate influence relative to other features.

#### • Administrative and Administrative Duration:

- Gain: Around 0.40% and 0.35% respectively, indicating a moderate impact on the model's decision process.

#### • MonthNov, ProductRelated Duration, and ProductRelated:

- Gain: Each shows smaller gain values (around 0.5% to 0.53%), suggesting moderate effects on the prediction outcomes.

### Partial Dependence Plot (PDP) for PageValues

- The PDP shows a sharp decrease in log loss as PageValues increase from 0 to 5, signifying a strong positive correlation between higher page values and the likelihood of a purchase.
- The log loss stabilizes after PageValues reach around 5, indicating that increases beyond this point do not significantly enhance the probability of making a purchase.

### 3. Model evaluation

This section presents a comparative analysis of model performance metrics between my experimental results and those reported in the study by Abdullah-All-Tanvir et al. (2023).

#### Models and Metrics Compared

The models evaluated include Logistic Regression, Classification Tree, Random Forest, and Gradient Boosting Machines (GBM). The key performance metrics analyzed are Accuracy, Precision, F1 Score, and ROC AUC.

Table 3: Model Performance Comparison

Model	Accuracy	Precision	F1	Recall	Specificity	ROC_AUC	PR_AUC	MCC
Logistic Regression	0.283	0.167	0.285	0.988	0.988	0.630	0.200	0.109
Classification Tree	0.876	0.573	0.664	0.790	0.790	0.841	0.508	0.602
Random Forest	0.874	0.869	0.391	0.220	0.220	0.851	0.610	0.399
GBM	0.880	0.634	0.577	0.530	0.530	0.863	0.631	0.511

XGBoost?



- **Accuracy:**

- My models generally underperform compared to Abdullah-All-Tanvir et al., with significant differences in accuracy. For example, the Classification Tree model reported an accuracy of 87.6% in my experiments compared to 90.54% in theirs.

- **Precision:**

- Similar trends are observed in precision, where all models in the Abdullah-All-Tanvir et al. study consistently showed higher precision. The Random Forest model, for instance, demonstrated a precision of 86.9% in my experiments versus 89.93% in theirs.

- **F1 Score:**

- The F1 Scores are substantially higher in the study by Abdullah-All-Tanvir et al. This is particularly noticeable in the Random Forest and GBM models, indicating a better balance between precision and recall in their models.

- **ROC AUC:**

- The ROC AUC values also indicate better performance in Abdullah-All-Tanvir et al.'s study. My GBM model achieved an ROC AUC of 0.863, while theirs reported a higher value of 0.937.

x No visualisation

## 4. Optimising for MCC

This section of the report delves into the optimization of predictive modeling for online shopper purchase intentions, focusing specifically on enhancing the Matthews Correlation Coefficient (MCC). MCC is a more informative metric compared to traditional metrics like accuracy or F1 score, especially in imbalanced datasets where it provides a balanced measure of the true positive and negative rates.

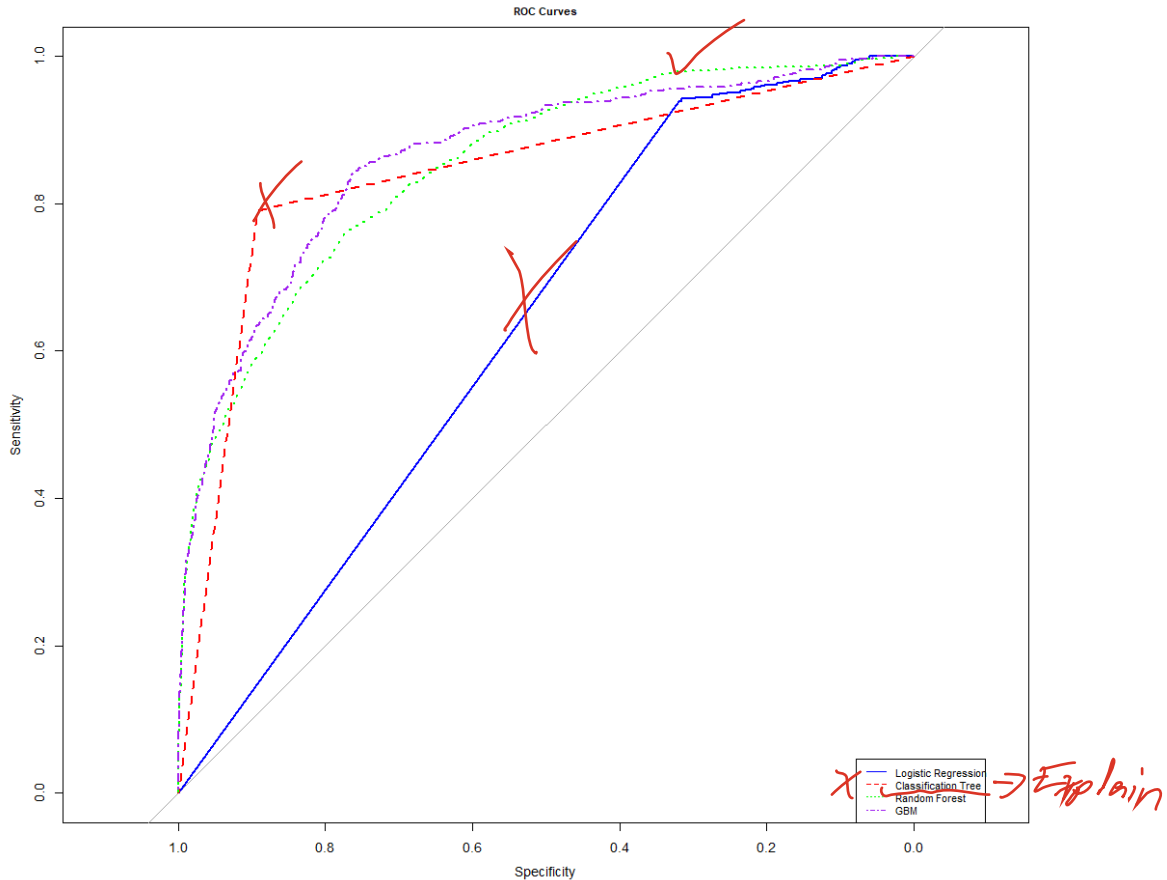


Figure 9: ROC Curves for Multiple Models.

The ROC curve for a classification tree is constructed by varying the decision threshold for the probability of the positive class (purchase). At each threshold:

- The model classifies an observation as positive if the predicted probability exceeds this threshold.
- Sensitivity (True Positive Rate) is calculated as the proportion of actual positives correctly identified by the model.
- $1 - \text{Specificity}$  (False Positive Rate) is the proportion of actual negatives incorrectly classified as positives.

nn



## References

1. Abdullah-All-Tanvir, Ali Khandokar et al. (2023). “A gradient boosting classifier for purchase intention prediction of online shoppers”. In: *Heliyon* 9.4. doi: <https://doi.org/10.1016%2Fj.heliyon.2023.e15163>.
2. Sakar, Cemal Okan et al. (2018). “Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks”. In: *Neural Computing and Applications* 31, pp. 6893–6908. doi: <https://doi.org/10.1007/s00521-018-3523-0>.