PREDICTING CONSUMER PURCHASE IN THE INITIAL JOURNEY OF ECOMMERCE PURCHASE

Applied Machine Learning - Project Report

Group 4
Rahul Jain (RXJ230006)
Abhinav Anand (AXA220249)
Abhishek Dubey (AXD220145)
Manmohan Dash (MXD220045)
Sai Sandeep Sreevaschava Surabhi (SXS220442)
Rucha Hrishikeshbhai Jadav (RHJ220001)

Contents

1. Executive Summary

The objective of this project is to predict the likelihood of a customer making a purchase online based on various input parameters, including time spent on various webpages, geographical information, bounce rate, exit rate, if it was a weekday or a weekend etc. Additionally, this project aims to provide insights into the data by using classification techniques with the goal of improving business outcomes. With the onset of e-commerce, it is important to identify the initial signs to determine the intent to purchase and take preemptive measures to nudge the platform's users into customers.

2. Data Overview

The dataset comprises features from 12,330 user sessions collected over a year. Each session represents a unique visit, ensuring diversity in the data and minimizing biases related to specific campaigns, special days, or user demographics.

The features include:

User Behavior Features: Metrics like Administrative, Informational, and Product-Related activities, along with their durations.

Website Interaction Metrics: Bounce Rates, Exit Rates, and Page Values.

Temporal Features: Month, Operating Systems, Browser, Region, Traffic Type, Visitor Type, Weekend.

Target Variable: Revenue, a Boolean indicating whether the session resulted in a purchase.

With the goal of the analysis in mind, we need to prepare the variable set such that they are available at the initial journey of the users. We wish to predict their intent to purchase when they begin their session. Thus, we have considered product page-related activity, exit rate, month, OS, browser, Region, Traffic Type, Visitor Type, and weekend as the measures to predict the target variable.

Statistical summary of data:

	Administrative	Administrative_Duration	Informational	\
count	12330.000000	12330.000000	12330.000000	
mean	2.315166	80.818611	0.503569	
std	3.321784	176.779107	1.270156	
min	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	
50%	1.000000	7.500000	0.000000	
75%	4.000000	93.256250	0.000000	
max	27.000000	3398.750000	24.000000	

	Informational_Duration	ProductRelated	ProductRelated_Duration	\
count	12330.000000	12330.000000	12330.000000	
mean	34.472398	31.731468	1194.746220	
std	140.749294	44.475503	1913.669288	
min	0.000000	0.000000	0.000000	
25%	0.000000	7.000000	184.137500	
50%	0.000000	18.000000	598.936905	
75%	0.000000	38.000000	1464.157214	
max	2549.375000	705.000000	63973.522230	

	BounceRates	ExitRates	PageValues	SpecialDay	\
count	12330.000000	12330.000000	12330.000000	12330.000000	
mean	0.022191	0.043073	5.889258	0.061427	
std	0.048488	0.048597	18.568437	0.198917	
min	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.014286	0.000000	0.000000	
50%	0.003112	0.025156	0.000000	0.000000	
75%	0.016813	0.050000	0.000000	0.000000	
max	0.200000	0.200000	361.763742	1.000000	

	OperatingSystems	Browser	Region	TrafficType
count	12330.000000	12330.000000	12330.000000	12330.000000
mean	2.124006	2.357097	3.147364	4.069586
std	0.911325	1.717277	2.401591	4.025169
min	1.000000	1.000000	1.000000	1.000000
25%	2.000000	2.000000	1.000000	2.000000
50%	2.000000	2.000000	3.000000	2.000000
75%	3.000000	2.000000	4.000000	4.000000
max	8.000000	13.000000	9.000000	20.000000

3. Data Preprocessing

The following preprocessing steps were designed to optimize the dataset for analysis:

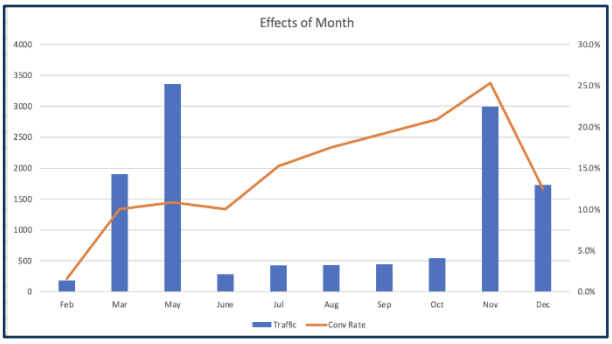
Data Cleaning: Checked for and confirmed no missing values. Columns incorrectly formatted as numerical (e.g., Operating Systems, Browser, Region, Traffic Type) were converted to categorical to represent the data accurately.

4. Exploratory Data Analysis (EDA)

Outlier Analysis: Conducted a brief analysis using statistical summaries through IQR and visualization techniques like scatter plots to identify and mitigate the impact of outliers.

Insights from the EDA highlighted several key factors:

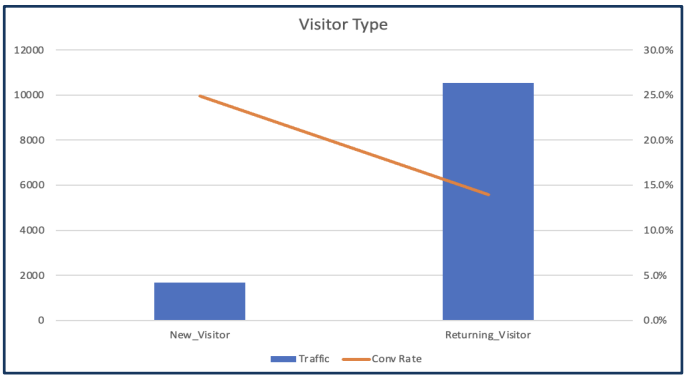
Seasonal Trends: Seasonal trends in the dataset highlight strategic months, like November and December, which exhibit notably higher success rates for converting website visits into revenue, potentially aligned with holiday shopping seasons.



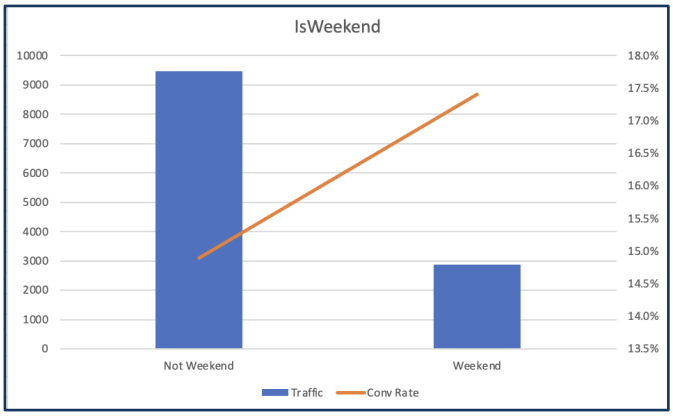
Traffic Type Variability: Different traffic types show substantial variation in conversion efficiency, with some demonstrating high revenue generation potential despite lower traffic volumes.

Regional Analysis: Region-based traffic analysis reveals that higher visitor counts do not necessarily correspond to higher revenue conversion rates, indicating opportunities for region-specific strategies to optimize conversion efficiency.

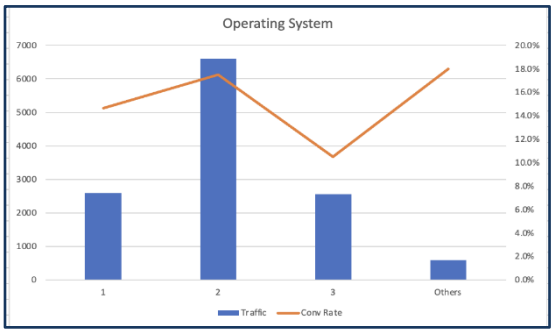
Visitor Type Impact: This shows that returning visitors not only contribute to most of the traffic but also have a success rate in making purchases close to the overall average, highlighting their importance in driving revenue.

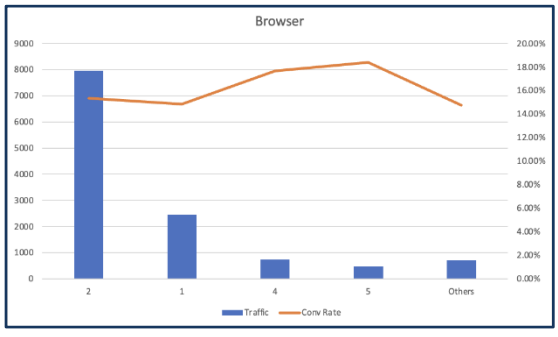


Weekend Effectiveness: Weekend visits, while fewer in number, show a marginally higher success rate in converting browsing into revenue, suggesting a quality-over-quantity dynamic that could inform targeted weekend marketing strategies.



Browser and Operating System: An analysis of traffic by operating system reveals that there isn't any significant variation in revenue across different operating systems and browsers, although some categories have a few samples. Moreover, because of the confidentiality of the data less information can be interpreted.





5. Data Processing

Feature Engineering - Reduction of Multicollinearity: Keeping only one of the highly correlated independent features and dropped features: Bounce Rates, Administrative, Administrative_Duration, Informational, PageValues, SpecialDay, and ProductRelated_Duration to prevent perfect correlation in model predictions.

Feature Encoding: Transformed categorical variables into dummy variables to better suit the modeling process.

Data Sampling: During the process of the model evaluations, we found that the model is not performing as expected due to very low positive samples. Thus, we oversampled the positive samples to build predictive power for the modeling exercise.

6. Model Building and Evaluation

Decision Tree

Objective: Build a model that offers intuitive decision rules and easy interpretability.

Model Configuration:

Criterion: Gini impurity is used to measure the quality of splits.

Max Depth: Limited to 5 to prevent the model from becoming overly complex and overfitting.

Min Samples Split: Minimum of 2 samples to split a node.

Max Leaf Nodes: Set to 9 to control the tree's growth.

Hyperparameters Tuned:

Max Depth: Tested from 1 to 10.

Min Samples Split: Range from 2 to 11.

Max Leaf Nodes: Tested from 2 to 10.

Performance (without preprocessing):

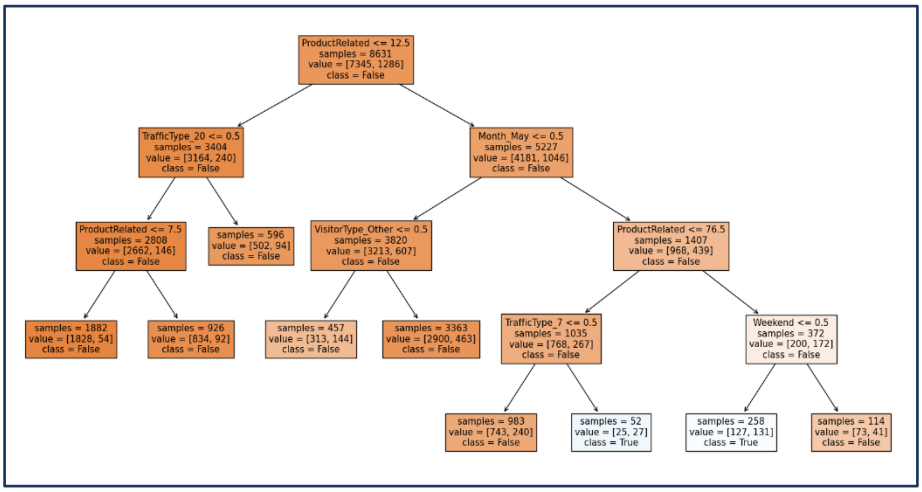
Training AUC: 0.708

Test AUC: 0. 694

F1 Score: 0.205

True Positive Rate: 12.9%

False Negative Rate: 87.1%



With this model, we can see the true positive rate is very low, and the false negative rate is very high. Thus, the model cannot detect the true nature due to the low proportion of positive cases. Hence, we would over-sample the positive cases in the dataset to increase the predictive power.

With oversampling of positive cases, the proportion of positive and negative cases is equal.

Performance (with preprocessing):

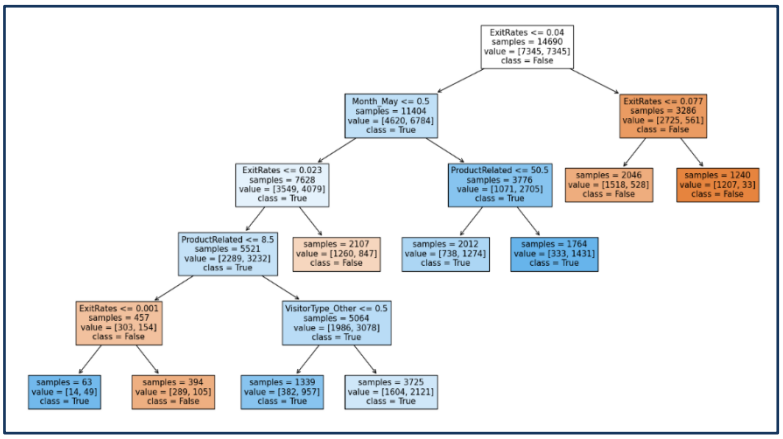
Training Accuracy: 0.719

Test Accuracy: 0.707

F1 Score: 0.386

True Positive Rate: 52.1%

False Negative Rate: 47.9%



<u>Logistic Regression</u>

Objective: Establish a baseline for model performance with a straightforward approach.

Model Configuration:

Penalty: L2 (Ridge) regularization was used to prevent overfitting and manage multicollinearity among features.

Max Iterations: Set to 10,000 to ensure convergence.

Hyperparameters:

Regularization Strength (C): Default value used (1.0), balancing accuracy and complexity.

Performance:

Training Accuracy: 0.659

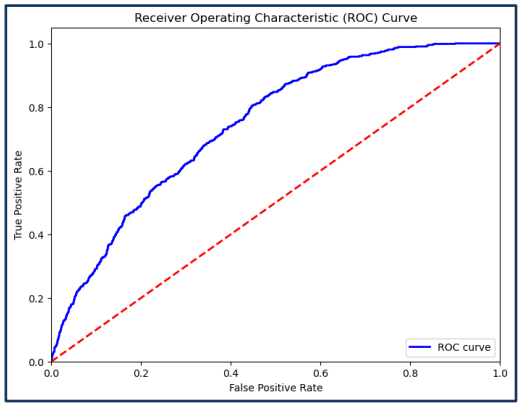
Test Accuracy: 0.654

AUC: 0.714

F1 Score: 0.384

Feature Importance:

The coefficients extracted indicate the relative influence of each feature on the prediction. Features like Month-Nov and browsers had higher coefficients, suggesting significant predictive power during the holiday shopping season and specific browser usage patterns.



Features	Odds ratio
Browser_13	2.5979
Browser_12	2.5133
TrafficType_16	2.468
TrafficType_7	2.1187
Month_Nov	1.8606
TrafficType_15	0.1415
TrafficType_8	1.8194
Month_Feb	0.2481
TrafficType_18	0.3835
TrafficType_13	0.4264
TrafficType_5	1.5696
Month_Oct	1.5088
OperatingSystems_5	0.5005
Browser_11	0.5005

Random Forest

Objective: Improve decision tree performance by ensemble learning, reducing variance without significantly increasing bias.

Model Configuration:

Number of Trees (n_estimators): Tested from 100 to 700 in increments of 100.

Bootstrap Samples: True, to allow sampling with replacement.

Max Samples: Tested various thresholds from 100 to 800 to determine the optimal sample size per tree.

Hyperparameters Tuned:

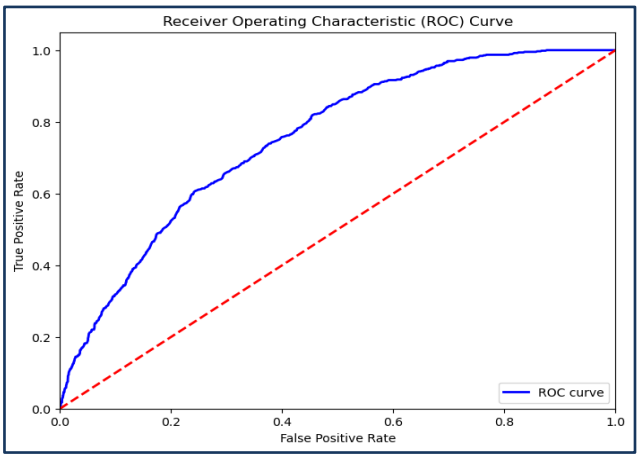
Max Features: Number of features to consider when looking for the best split.

Performance:

Training AUC: 0.898

Test AUC: 0.729

F1 Score: 0.404



AdaBoost

Objective: Boost the performance of decision trees by focusing on difficult cases and adjusting weights iteratively.

Model Configuration:

Base Estimator: Decision Tree with a max depth of 1.

Number of Estimators: 500, to progressively refine the decision boundaries.

Learning Rate: 1.0, to weigh the contribution of each subsequent classifier.

Hyperparameters Tuned:

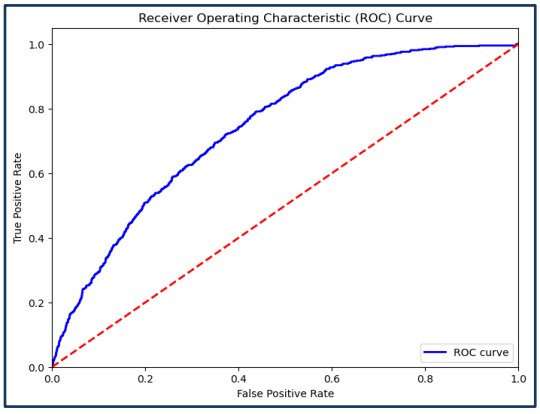
Learning Rate: Tested rates from 0.1 to 1.0 to optimize the contribution of each tree.

Performance:

Training Accuracy: 0.683

Test Accuracy: 0.644

F1 Score: 0.394



Multi-Layer Perceptron (MLP)

Objective: Explore the capability of neural networks to capture complex patterns in the data.

Model Configuration:

Architecture: Three layers with 200, 100, and 50 neurons, respectively.

Activation Function: ReLU for non-linear transformations.

Solver: Adam, an efficient stochastic optimization method.

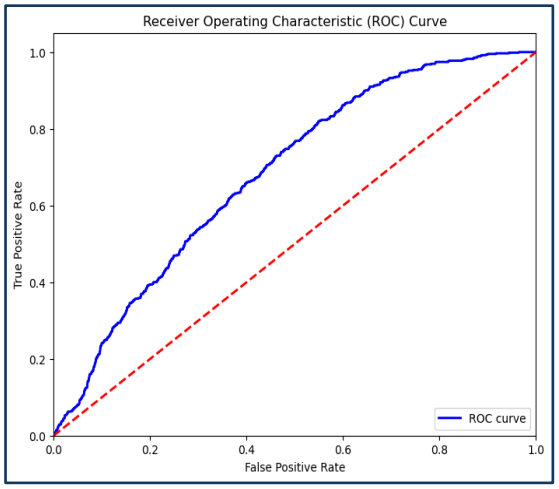
Max Iterations: 1000 to allow sufficient training epochs.

Performance:

Test Accuracy: 0.756

Test AUC: 0.674

F1 Score: 0.296



7. Model Comparison and Selection

Model Selection based on Evaluation Metrics: We used the ROC curve as an evaluation metric for model selection. Based on that, random forests outperformed with an excellent value above 0.729, making it the preferred choice, followed by decision trees.

Preference for Decision Tree: We opt for decision trees due to their simplicity and interpretability. We can leverage its predictive power to focus on improving classification for better segmentation of users.

8. Conclusions

With our predictability of the user's purchase decision in the initial part of the user journey, we can recommend two sets of users. One who is more likely to purchase and thus these users should not be intervened in their journey. Simultaneously, we also predict the users who will not purchase if not intervened

with a suitable nudge. Thus, we not only improve the conversion rate but also reduce marketing expenses on the overall set of users.

The model's predictive accuracy is approximately 72%, representing the combined proportion of accurately predicted true positive and true negative cases within the total test set. From a business standpoint, true positive cases signify individuals correctly identified as converting customers by the model. The true positive rate indicates the percentage of individuals accurately identified as potential buyers. This proportion is 9% for the model.

On the other hand, true negative cases consist of customers correctly identified as non-converting, resulting in no purchase. The true negative rate reflects the proportion of individuals correctly classified as non-buyers, highlighting the effectiveness of avoiding targeting these individuals to prevent unnecessary costs. Such a population comprises 63% in our model prediction.

False positive cases represent potential customers who were predicted to convert but did not, indicating a missed opportunity. Despite being expected to, these individuals didn't make a purchase, suggesting areas for improvement in conversion strategies. The false positive rate measures the percentage of individuals erroneously classified as potential buyers. This opportunity cost, which the business incurred due to some flaw with the user experience, constitutes 20% of the model predictions.

Conversely, false negative cases are individuals who weren't expected to purchase but did so while interacting with the website. It highlights missed conversion opportunities and suggests areas for enhancing targeting and engagement strategies. The false negative rate denotes the percentage of individuals mistakenly identified as non-buyers. These sets of customers comprise around 8% in the model prediction outcome.

In conclusion, we recommend prioritizing efforts to capture the false positive cases, which represent missed opportunities for conversion. Betterment of the user experience on the platform can enhance engagement and encourage these individuals to become actual buyers. Additionally, it's advisable for the business to save resources by refraining from targeting the true negative cases, as these individuals demonstrate minimal potential for conversion to actual buyers. By focusing on converting missed opportunities and optimizing resource allocation, the business can drive growth and efficiency in its marketing strategies.

The key takeaway with the exploratory analysis is that visitor type and time of the week gives us a good separation on conversion probability. We would like to effectively use the parameters to engage with the model.

9. References

Data source: Online shoppers purchasing intention

https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset