

Online Shoppers' Purchasing Intention

Group 11

Jubin Joung, Satvik Mahendra, Wonjun Ryhu

0. Introduction

0.1 Introduction

In this project, we will analyze the Online Shoppers Purchasing Intention dataset from the UCI Machine Learning Repository. This dataset captures user browsing behavior from an e-commerce platform during a 1-year period, including page interactions, engagement metrics (BounceRates, ExitRates, PageValues), visit timing (Weekend, SpecialDay, Month), and user profile indicators (VisitorType, Region, OS, Browser). The goal is to understand how these behavioral and contextual factors influence whether a session ends with a purchase.

Predicting purchasing intention in real time is critical in digital marketing and recommendation systems. By identifying which users are most likely to convert, companies can optimize targeted promotions and reduce advertising costs. Additionally, analyzing influencing factors may provide actionable insights into user preferences and seasonal shopping trends.

0.2 Data Description

The dataset used in this analysis is the *Online Shoppers Purchasing Intention* dataset, which contains detailed behavioral logs from 12,330 individual browsing sessions on an online retail website. Each row represents a single session rather than a unique user, and the primary objective is to determine whether the session resulted in a purchase, as indicated by the binary variable *Revenue*. The dataset includes a rich set of features that capture multiple dimensions of user behavior.

Several variables describe page-level interaction patterns, such as *Administrative*, *Informational*, and *ProductRelated*, along with the corresponding time spent in each category. These features reflect the depth and type of content explored by the user. Engagement and abandonment indicators, including *BounceRates* and *ExitRates*, provide insights into early exit behavior and browsing efficiency. A particularly important feature, *PageValues*, represents the estimated value of a page based on historical conversion data and therefore serves as an indirect measure of how close a user is to finalizing a transaction.

The dataset also contains contextual and demographic attributes such as *Month*, *OperatingSystems*, *Browser*, *Region*, *TrafficType*, and *VisitorType*. These variables capture the temporal environment and visitor identity, both of which may influence purchase behavior. Additionally, the *Weekend* variable indicates whether the session occurred on a weekend, and can help detect patterns in user activity across different days of the week. Overall, the dataset provides a comprehensive view of user behavior, enabling a robust analysis of the factors that drive purchase decisions.

0.3 Data Preprocessing

Before conducting any modeling, we performed a series of preprocessing steps to prepare the dataset for analysis. We first checked for duplicate entries and confirmed that no duplicates were present, leaving the dataset at 12,205 valid sessions after initial cleaning. The *Weekend* and *Revenue* columns, which were originally represented as string-like boolean values, were converted to consistent binary integers (0 or 1) to ensure compatibility with machine learning algorithms.

Next, the features were organized into meaningful groups based on their type. Numerical features which included page visit counts, durations, abandonment rates, and *PageValues*. These features were handled using median imputation to replace any missing values, followed by standardization so that all numeric variables were placed on a comparable scale. Categorical features such as *Month* and *VisitorType* were imputed

using the most frequent category and subsequently transformed through one-hot encoding. The *Weekend* variable, already expressed as a binary indicator, was passed through without modification.

A notable part of our preprocessing pipeline involved converting the *Month* feature into cyclical representations using sine and cosine transformations. Traditional one-hot encoding treats months as unrelated categories, but seasonality is inherently cyclical. The sin/cos encoding preserves this continuity by mapping December and January closer together and aligning similar seasonal months. This representation is especially useful for models sensitive to numerical relationships, allowing them to capture cyclical shopping patterns more effectively.

All preprocessing steps were implemented using a unified scikit-learn pipeline, ensuring reproducibility and preventing data leakage. The final processed dataset consisted of 20 engineered features and 12,205 observations, which served as the standardized input for all subsequent modeling procedures.

1. Question 1

What attributes are the most relevant for classifying whether a shopper makes a purchase?

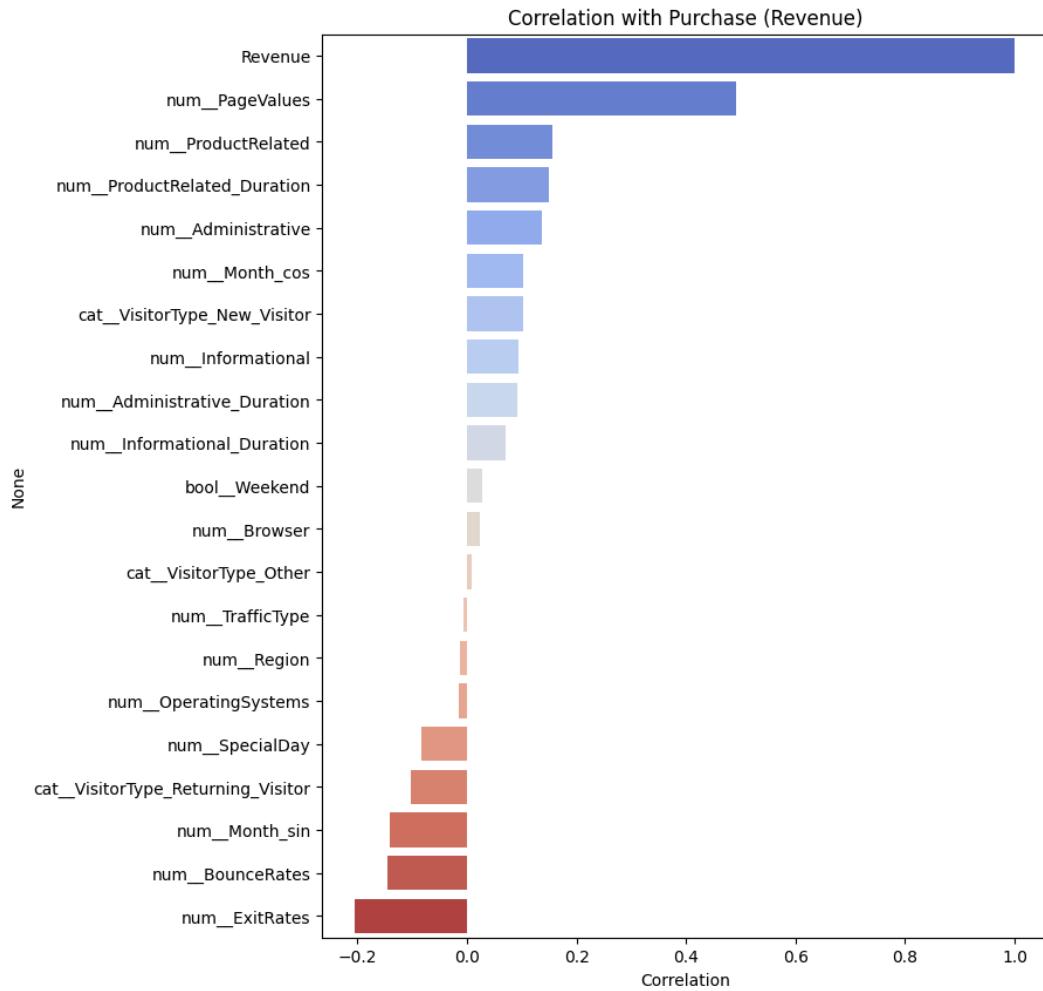
1.1 Introduction

The first research question aims to identify which attributes of a browsing session are most relevant for predicting whether a shopper ultimately makes a purchase.

Understanding the behavioral and contextual factors that drive conversion is essential not only for model performance but also for gaining actionable insights into user engagement, website design, and marketing strategy. Because the dataset contains a diverse set of behavioral, economic, temporal, and demographic features, it is important to evaluate their predictive importance using multiple complementary approaches.

To accomplish this, we conducted a series of analyses including correlation assessment, logistic regression, Random Forest, XGBoost, and SHAP-based interpretability. This combination of linear, non-linear, and model-agnostic methods allows us to examine both the direction and magnitude of each feature's influence while ensuring that our findings remain robust across different modeling frameworks. The following sections summarize the results from each method and highlight the attributes that consistently emerge as the strongest predictors of purchase behavior.

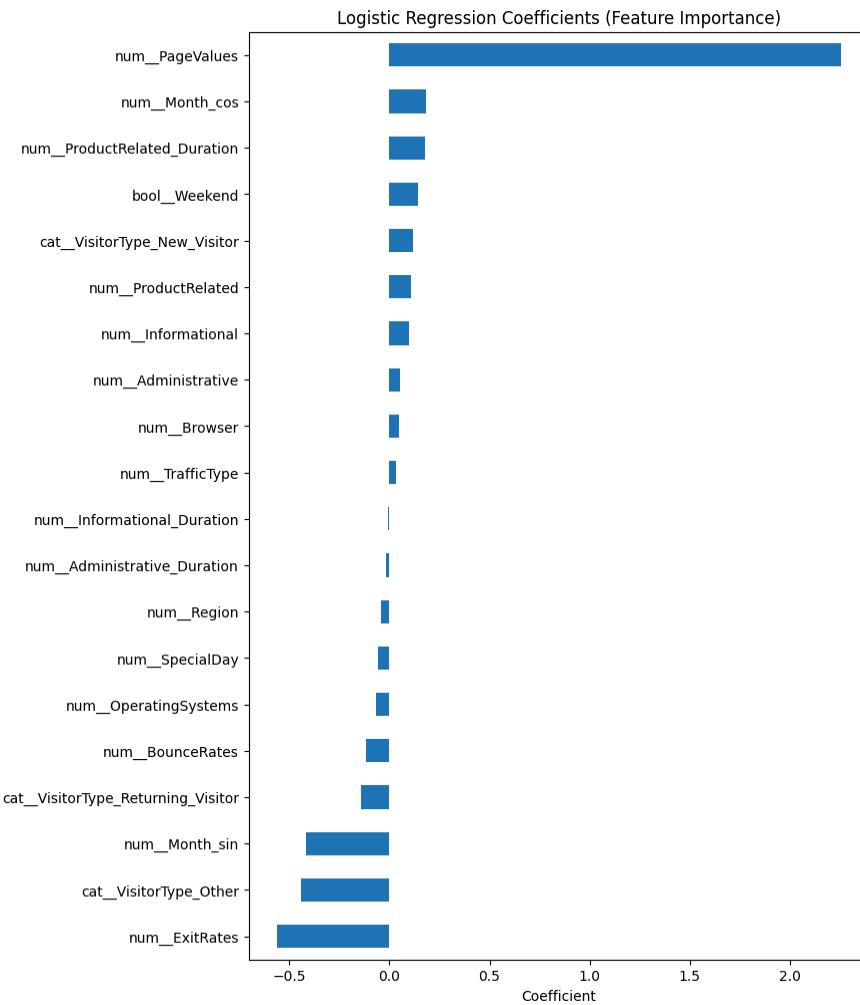
1.2 Correlation Analysis



We first examined the linear relationships between each feature and the purchase outcome using a correlation matrix. The results showed that PageValues had by far the strongest positive correlation with Revenue, indicating that sessions involving pages with historically high conversion value were substantially more likely to result in a purchase. Product engagement variables, particularly *ProductRelated* and *ProductRelated_Duration* also showed moderate positive correlations, suggesting that users who spend more time exploring product pages generally demonstrate higher purchase intent. In contrast, abandonment-related indicators such as ExitRates and BounceRates were strongly negatively correlated with purchases, reflecting the intuitive interpretation that users who leave pages quickly or abandon browsing flows early are

much less likely to convert. These initial trends provided a foundation for understanding which behavioral factors mattered most before applying more complex models.

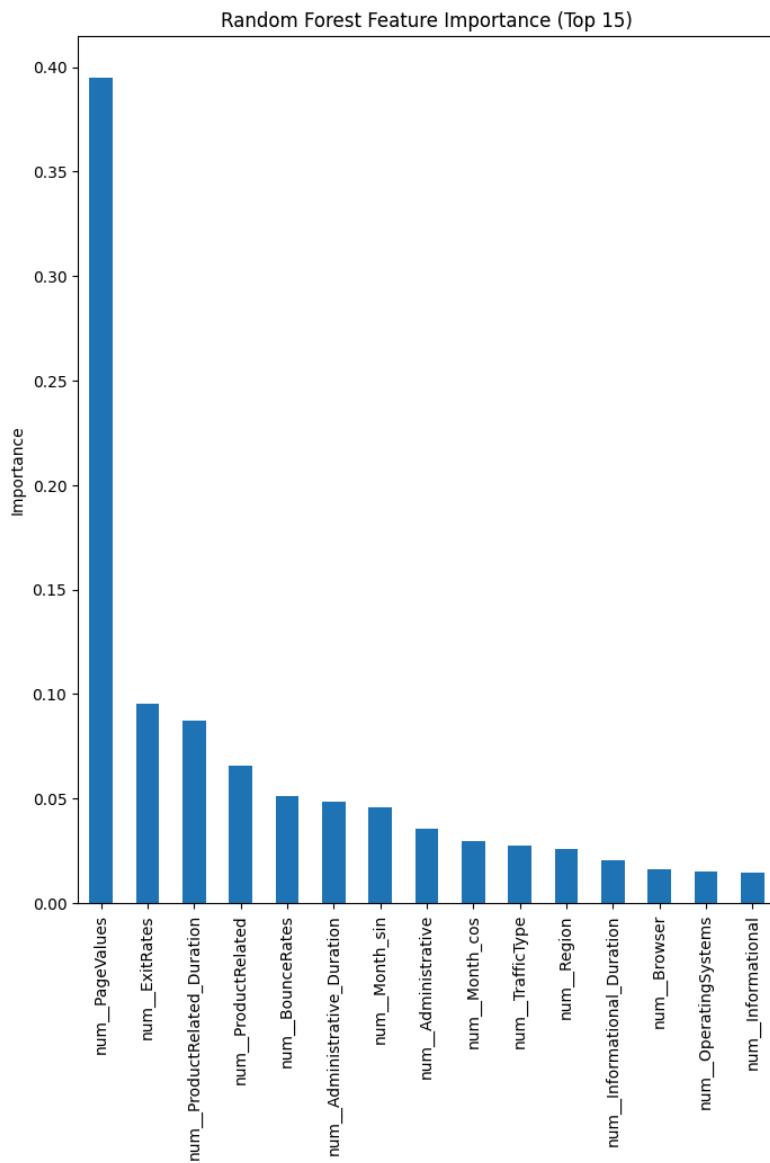
1.3 Logistic Regression



To capture directional effects while controlling for other variables, we fitted a logistic regression model. The coefficients once again identified PageValues as the strongest positive predictor, reinforcing its central importance across statistical methods. High PageValues essentially signal that the user has navigated closer to high-intent pages, which increases the likelihood that the session ends in a purchase. Meanwhile, features such as ExitRates, BounceRates, and Month_sin displayed strongly negative

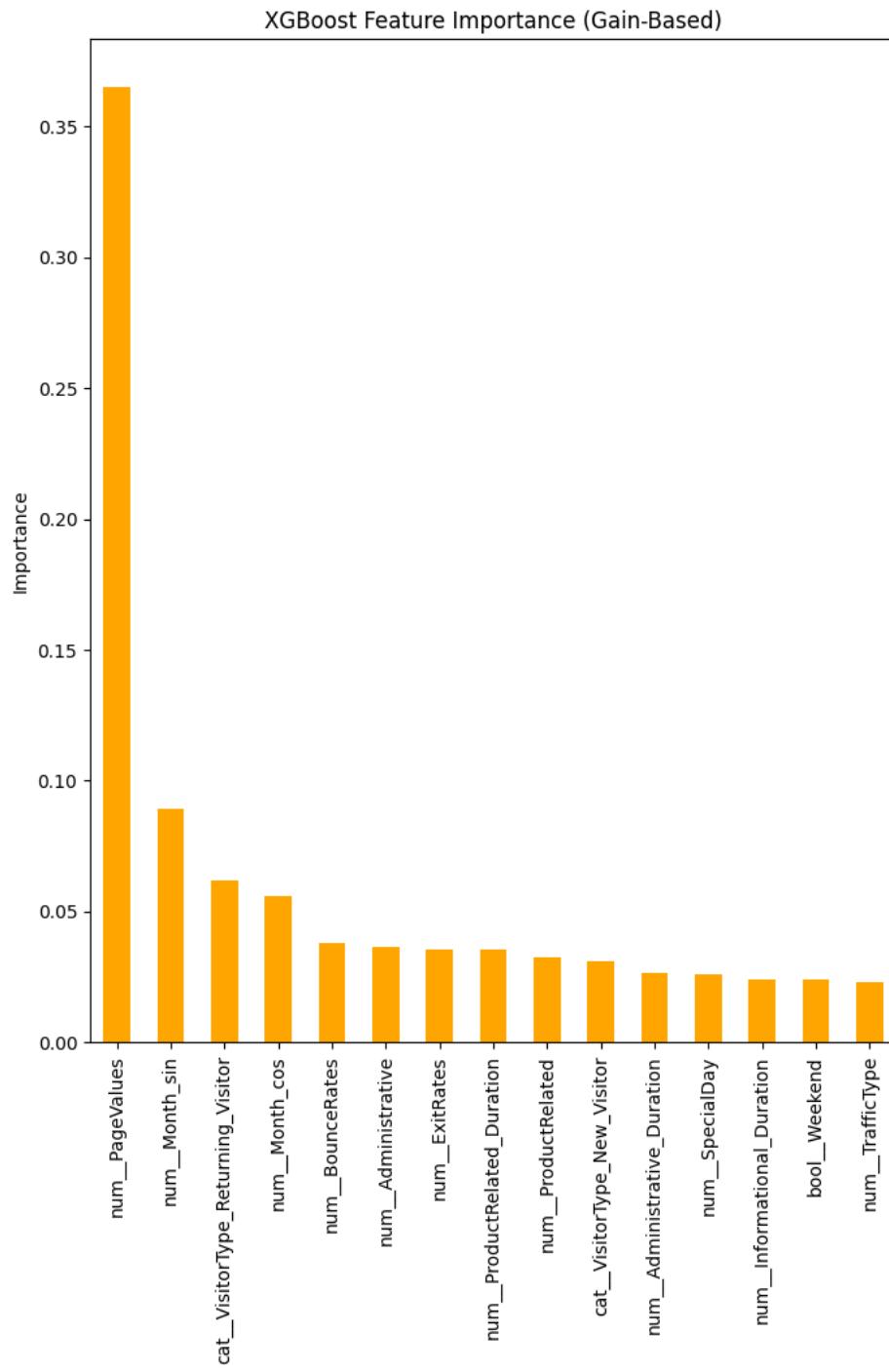
coefficients, indicating that higher values in these features significantly reduce conversion probability. The model achieved an ROC-AUC of approximately 0.90, demonstrating strong predictive power even under a linear specification. Together with the correlation results, the logistic regression supported a consistent interpretation of the key behavioral drivers behind purchases.

1.4 Random Forest Feature Importance



Because purchase decisions often reflect complex patterns, we next applied a Random Forest model to capture non-linear relationships. The importance rankings again placed PageValues overwhelmingly above all other features, showing that its predictive strength persists even when interactions are considered. Additional influential features included ExitRates, ProductRelated_Duration, ProductRelated, and BounceRates, all of which align with intuitive expectations: deeper engagement with product content increases conversion likelihood, whereas abandonment signals sharply reduce it. The consistency between Random Forest results and those from the logistic regression demonstrated that the strongest predictors remain stable across different modeling assumptions.

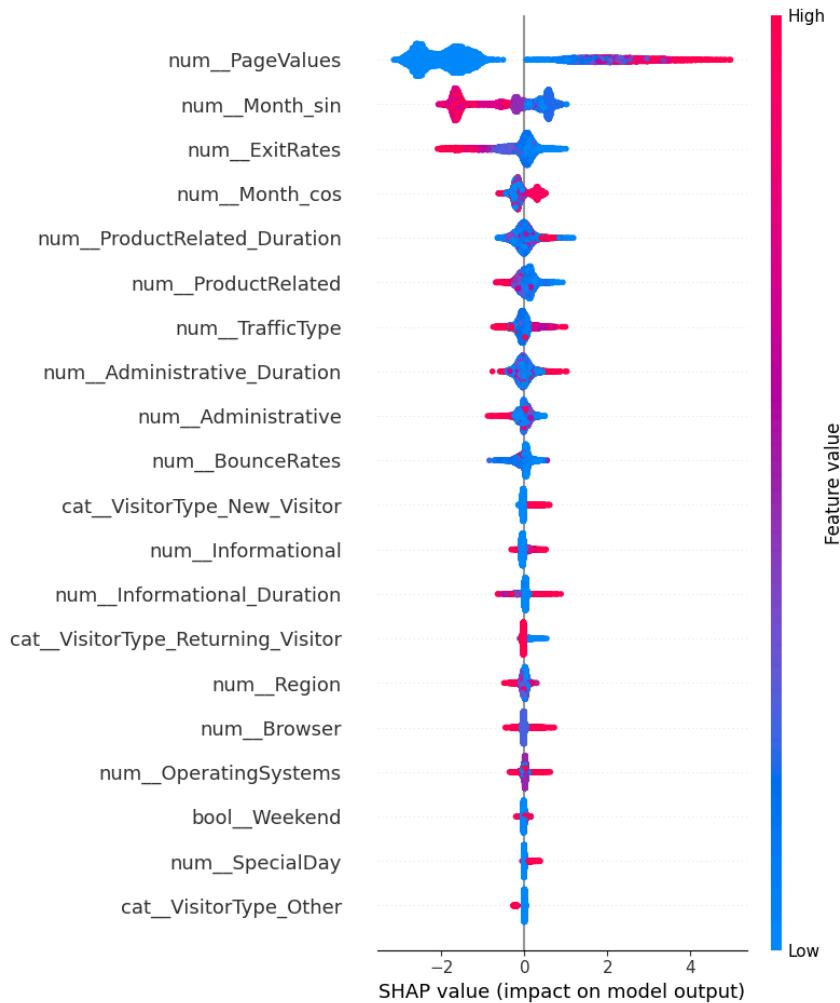
1.5 XGBoost Feature Importance

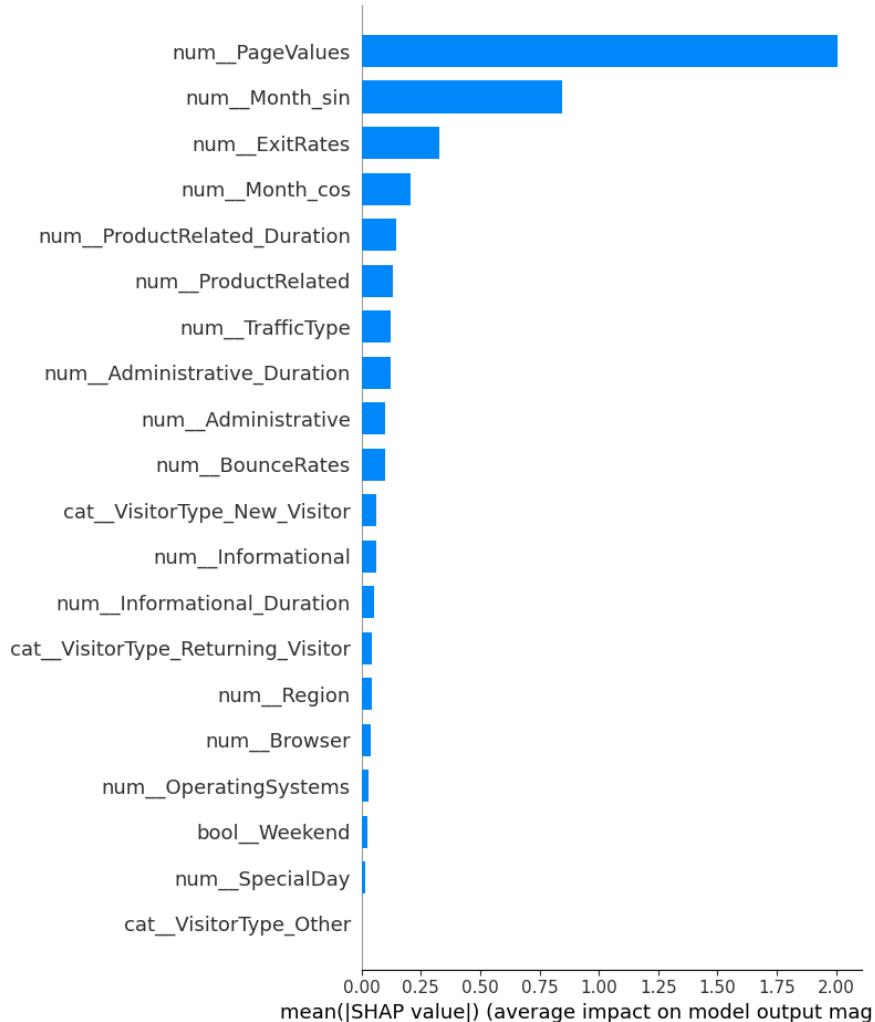


To further capture non-linear interactions, we used XGBoost, which is often more sensitive to subtle patterns. XGBoost reinforced previous findings by again identifying PageValues as the most critical feature, far surpassing others in importance.

Interestingly, XGBoost also assigned substantial weight to the cyclical features Month_sin and Month_cos, indicating that purchasing behavior follows a seasonal rhythm. Features tied to abandonment behavior, such as ExitRates and BounceRates, also remained influential. This model strengthened the insight that conversion likelihood depends heavily on both user engagement depth and the broader temporal context.

1.6 SHAP Interpretability Analysis





To obtain interpretable, model-agnostic explanations, we applied SHAP values to the XGBoost model. The SHAP summary plot revealed that high PageValues consistently generated large positive contributions, increasing the predicted probability of purchase. Conversely, high values of ExitRates and BounceRates produced strong negative SHAP impacts, showing that early abandonment sharply lowers the likelihood of conversion. SHAP also highlighted meaningful contributions from *ProductRelated* engagement and seasonal components. Importantly, the SHAP rankings mirrored the importance patterns seen in Random Forest and XGBoost, confirming the robustness of these findings across multiple modeling frameworks.

1.7 Why PageValues Matters

PageValues is the most important feature because it reflects the estimated economic value of each page based on historical behavior. When a user reaches pages associated with high conversion probability such as checkout, pricing, or review pages so the PageValues metric increases. Unlike raw page counts or durations, PageValues directly encodes how close a user is to completing a transaction. Therefore, when PageValues is high, it meaningfully signals that the user has progressed into stages of the browsing journey that historically correlate strongly with purchases. This makes it a uniquely powerful predictor across all models.

1.8 Conclusion

Across all linear, non-linear, and interpretability-focused methods which include correlation, logistic regression, Random Forest, XGBoost, and SHAP, PageValues consistently emerges as the strongest and most reliable predictor of purchase behavior. Additionally, abandonment indicators such as ExitRates and BounceRates significantly reduce purchase likelihood, while deeper engagement with product-related content increases it. Seasonal patterns further contribute to predictive accuracy. The remarkable consistency of these results across multiple methods demonstrates the robustness of these drivers and provides actionable insights for improving website design, user flow, and conversion strategies.

2. Question 2

Based on the features in the dataset, will a user make a purchase?

2.1 Introduction

The second question we sought to answer was: Based on the features in the dataset, will a user make a purchase? Being able to confidently answer this question would provide valuable insight for online stores as they would be able to manage inventory more effectively based on user behavior and optimize the online store to improve conversion rate, thus driving higher revenue for the store owners.

2.2 Approach

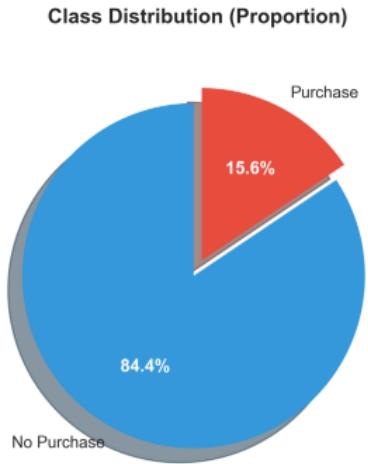
The first step in identifying how to answer this question was to understand which techniques would be the best to use. Since this question requires us to make a prediction about a binary outcome, we can conclude that this is a classification problem. Knowing this, there are a few approaches we sought to experiment with to answer the classification question.

We experimented with six distinct classification methods to try to identify the optimal predictive model.

1. Decision Tree
2. Support Vector Machine
3. K-Nearest Neighbors
4. Gradient Boosting
5. Random Forest
6. Logistic Regression

2.3 Class Imbalance

Our initial exploratory analysis of our dataset revealed a striking class imbalance.



Pie Chart Visualizing Class Imbalance

The visualization confirms that approximately 85% of user sessions do not result in a purchase. A base model predicting “False” for every interaction would thus achieve 85% accuracy, rendering “Accuracy” a potentially misleading metric. There were various approaches we considered to help address this class imbalance.

First, we chose appropriate metrics for evaluating our models. Given the class imbalance in our dataset, we evaluated these models using a comprehensive set of metrics: Accuracy, Precision, Recall, F1-Score, and Area Under the Receiver Operating Characteristic Curve (ROC-AUC). We specifically focused on F1-Score and ROC-AUC since other metrics may miss evaluating certain bias in the models. These metrics are especially good for handling the Precision-Recall tradeoff and ensuring that poor minority class performance is visible in our results.

Second, to mitigate the class imbalance, we also implemented Synthetic Minority Oversampling Technique, or SMOTE. This sampling method creates synthetic data that uses the k-nearest neighbors of minority class instances to create new data points. We applied SMOTE to specific models in order to experiment with how using SMOTE compares to the baseline version of models to see what difference it creates.

Third, we made use of certain parameters exposed by the APIs of the libraries we used to train our models. For example, when building a decision tree, we can set the ‘class_weight’ parameter to be ‘balanced’. Setting this parameter automatically adjusts class weights to be inversely proportional to class frequencies, which results in the model more heavily penalizing errors on the minority class. Thus, we can augment our model training by encouraging better predictions on the minority class, helping to counteract the potential bias caused by the class imbalance in our data.

2.4 Model Training

Prior to training our models, we had to implement a preprocessing pipeline to ensure our data was suitable for the models we were experimenting with and to ensure we maintained consistency between these parameters for our various models. As mentioned earlier in the report, this involved steps such as encoding categorical variables, imputing numeric variables, and scaling numeric variables. We then split the data into a train and test set, with 80% of data used for training and 20% used for testing. SMOTE was applied before model training but after preprocessing to prevent data leakage between our training and testing data splits.

2.5 Hyperparameter Tuning

To ensure our models were performing well, we also experimented with hyperparameter tuning to find the optimal configuration of hyperparameters for the metrics we were interested in examining. We specifically focused on tuning the hyperparameter for our top three performing models: Gradient Boosting, Random Forest, and Logistic Regression. The tuning process was designed to maximize the ROC-AUC rather than just accuracy to ensure the models were optimized to handle the class imbalance in our dataset.

We used a mixed strategy for hyperparameter tuning based on the scale of options for the model's parameters. For the tree-based ensemble models (Random Forest and Gradient Boosting) which have a larger number of potential hyperparameter combinations, we used `RandomizedSearchCV`. This samples a fixed number of parameter settings from specified distributions/ranges, which is a more efficient alternative to exhaustively searching all combinations. The Logistic Regression model has a smaller number of possible parameter combinations and thus we were able to perform an exhaustive search using `GridSearchCV`. For all of our models, we used 5-Fold Cross Validation to ensure that our selected hyperparameters would generalize well to unseen data.

2.6 Model Results

After training all of our models, we aggregated the results to compare the models' capabilities. The table below details the performance of our models.

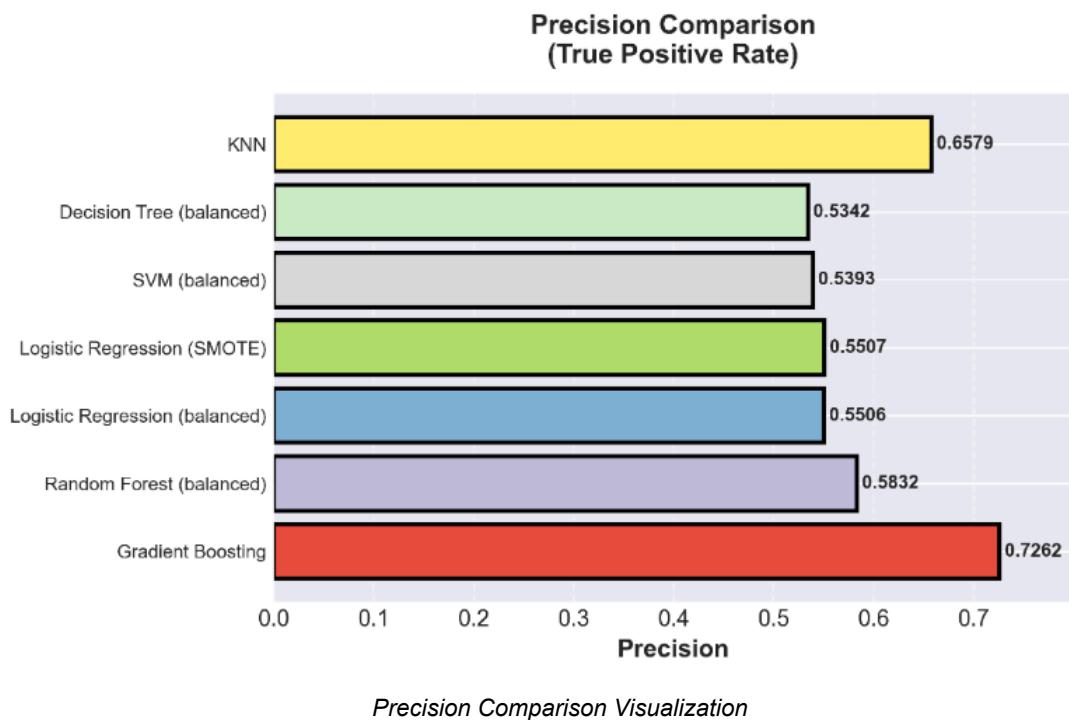
Model	ROC-AUC	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	0.9328	0.9037	0.7262	0.6178	0.6676
Random Forest (balanced)	0.9292	0.8783	0.5832	0.7801	0.6674
Logistic Regression (balanced)	0.9074	0.8656	0.5506	0.7696	0.6419
Logistic Regression	0.9074	0.8652	0.5507	0.7539	0.6365

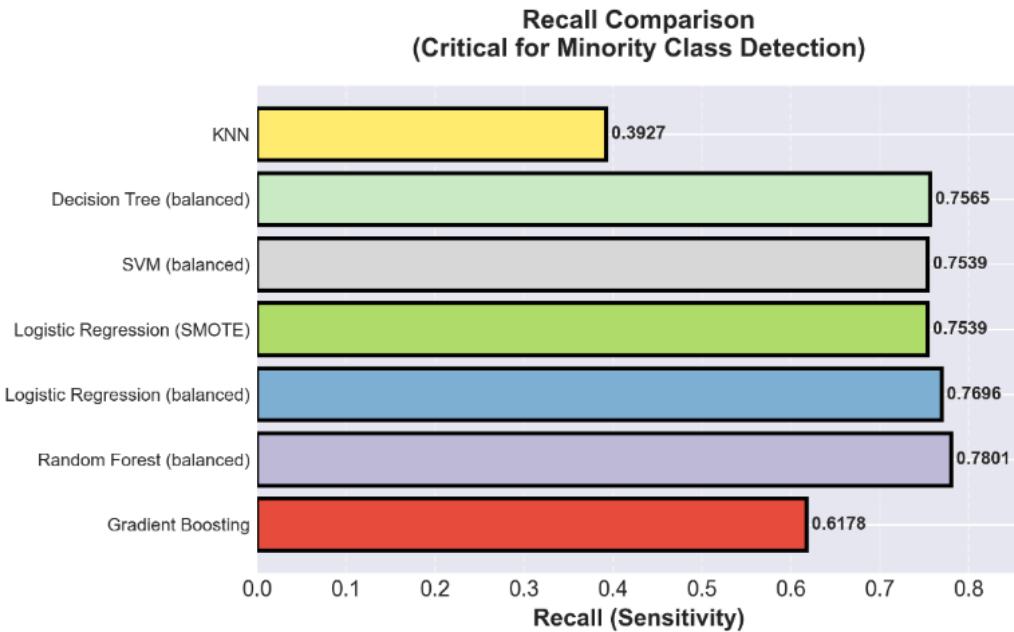
(SMOTE)					
Support Vector Machine (balanced)	0.8994	0.8607	0.5393	0.7539	0.6288
Decision Tree (balanced)	0.8633	0.8587	0.5342	0.7565	0.6262
K-Nearest Neighbors	0.8009	0.8730	0.6579	0.3927	0.4918

We made several key observations from our aggregated results. First, while the baseline Logistic Regression model performed adequately (ROC-AUC of 0.907), it was consistently outperformed by tree-based ensembles. This likely means that the relationship between session behaviors and whether or not a user makes a purchase is non linear and better captured by tree-based methods that can model these complex interactions. Overall, the Gradient Boosting Classifier was the best performing with an ROC-AUC of 0.933 and an F1-Score of 0.668. While the Random Forest model offered a higher Recall of 78%, the Gradient Boosting model provided the best balance. It had a significantly higher precision (72.6%), which means that when it predicted a purchase, it was correct almost 73% of the time. This demonstrates how the Gradient Boosting model balanced the Precision-Recall tradeoff the best. However, as we will discuss later, there is an argument to be made for Random Forest being the best model if one values Recall more than Precision.

Our worst performing models were the KNN and SVM models, which tended to underperform compared to the ensemble models. This is likely because of noise that is present in the sessions of users that didn't make purchases, which interfered with the distance calculations in higher-dimensional spaces.

Additionally, it is important to note that certain models have different strengths and weaknesses. For example, a model may have great precision but a poor recall. This information is important in practical applications since there may be situations where we seek to minimize certain types of errors as they may be more costly. The bar charts below demonstrate a visual representation of the models' performance in the context of their precision and recall.





Recall Comparison Visualization

From these plots, we can see that Gradient Boosting has the highest precision but actually has a relatively poor recall. Thus, even though Gradient Boosting was our best performing model overall, if we are more concerned with predictions of the minority class, it may be better to use a Random Forest model as this model performs better in that situation. On the other hand, although the KNN model was considered to be a poor performing model, we can see from the visualization that it is actually relatively powerful at identifying true positives, which once again may be a valuable insight when determining which model to use in production environments.

The plot below helps reinforce some of our findings from before by visualizing the ROC Curve and Precision-Recall Curve for the models we experimented with. There is not much to note from these plots as they primarily confirm the analysis from earlier.

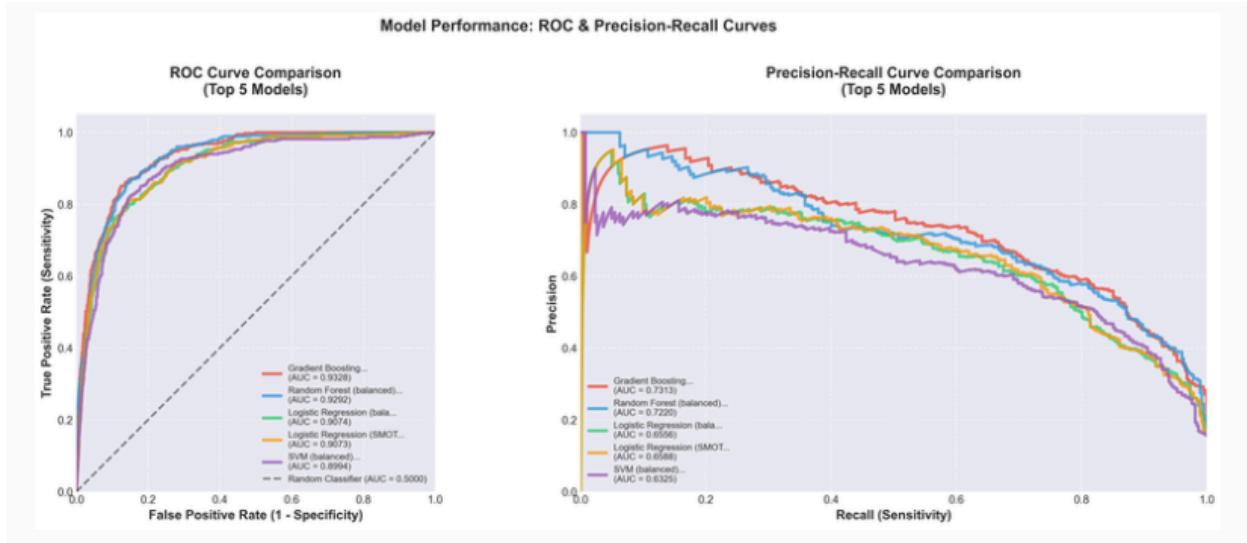


Figure [X]: ROC & Precision-Recall Curves for Different Models

Lastly, we can see a plot demonstrating the overall results from the Gradient Boosting model. The model excels at predicting true negatives and is pretty good at predicting true positives. However, this disparity between true negative prediction rate and true positive prediction rate further reinforces the challenges we anticipated from our imbalance dataset.

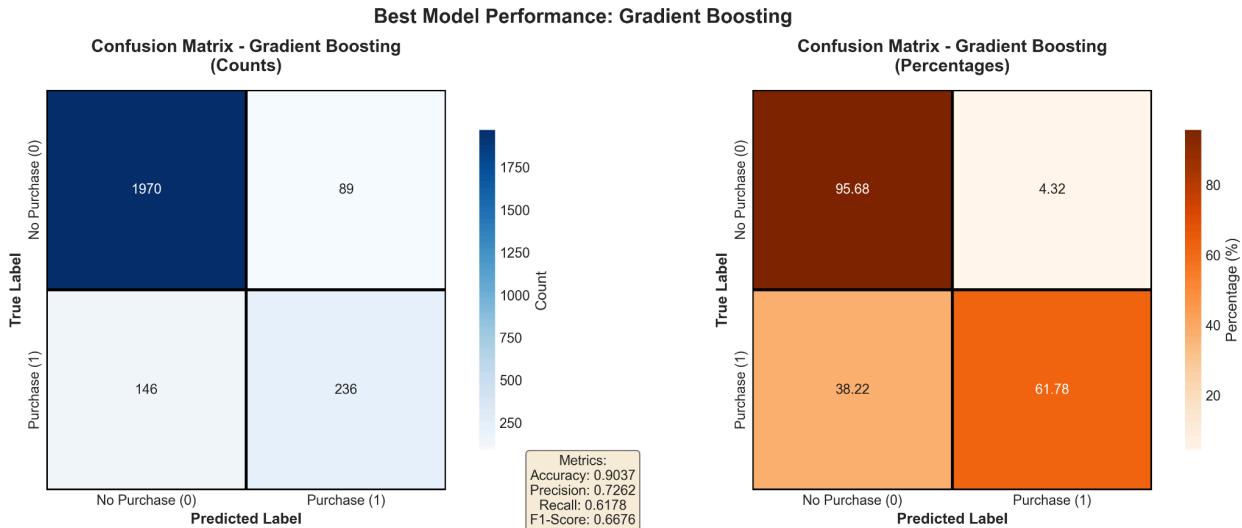


Figure [X]: Confusion Matrix Heatmap for Gradient Boosting Model

2.7 Conclusion

Overall, we were confidently able to answer the question of: Based on the features in the dataset, will a user make a purchase? By training various classification models, appropriately handling a class imbalance in our dataset, and rigorously comparing our models using more robust metrics, we were able to converge on a Gradient Boosting model being the best for making predictions about user purchasing based on our dataset. This information can be especially valuable for those running e-commerce platforms as it can allow store owners to modify their sites to help drive conversions. For instance, if the user's session activity suggests they may be a user who is more likely to make a purchase, relevant discounts or cross-sells can be dynamically displayed to help ensure the user does indeed make a purchase. Additionally, this information can be useful for allowing store owners to better manage inventory based on the types of users they expect to be visiting their site and whether or not they will make a purchase.

3. Question 3

How Do Seasonal and Behavioral Contexts Influence Purchasing Intention?

To deepen our understanding of what drives online purchasing behavior, we expanded our analysis beyond static session-level features and examined how **seasonality** and **behavioral patterns** jointly shape purchase intention. Online shopping is inherently time-sensitive: consumer motivation fluctuates across seasons, holidays, and promotional cycles. Our goal in Question 3 was to quantify these temporal dynamics and uncover latent shopper profiles that help explain why certain users convert while others do not.

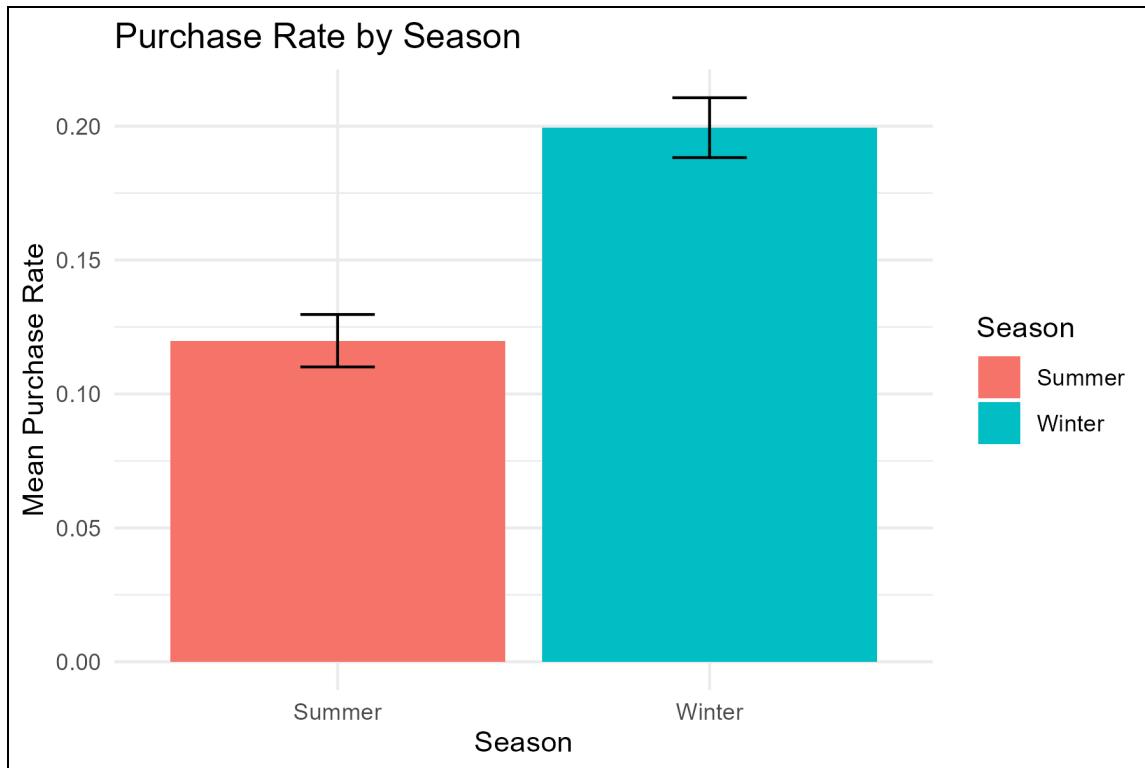
Our methodological pipeline integrates three complementary approaches:

1. **Seasonal purchase-rate comparisons**
2. **Season × VisitorType interaction analysis**
3. **Season-specific logistic regression models**
4. **Unsupervised clustering to identify behavioral shopper segments**

Together, these allow us to capture both macro-level seasonal shifts and micro-level behavioral heterogeneity.

3.1 Seasonal Effects on Purchasing Behavior

We first compared purchase probabilities between Summer (May–August) and Winter (November–February). To preserve raw temporal variation, we used the original dataset without smoothing or resampling.



Key Finding:

Winter shoppers convert at nearly twice the rate of Summer shoppers.

The confidence intervals barely overlap, indicating that the seasonal gap is statistically meaningful.

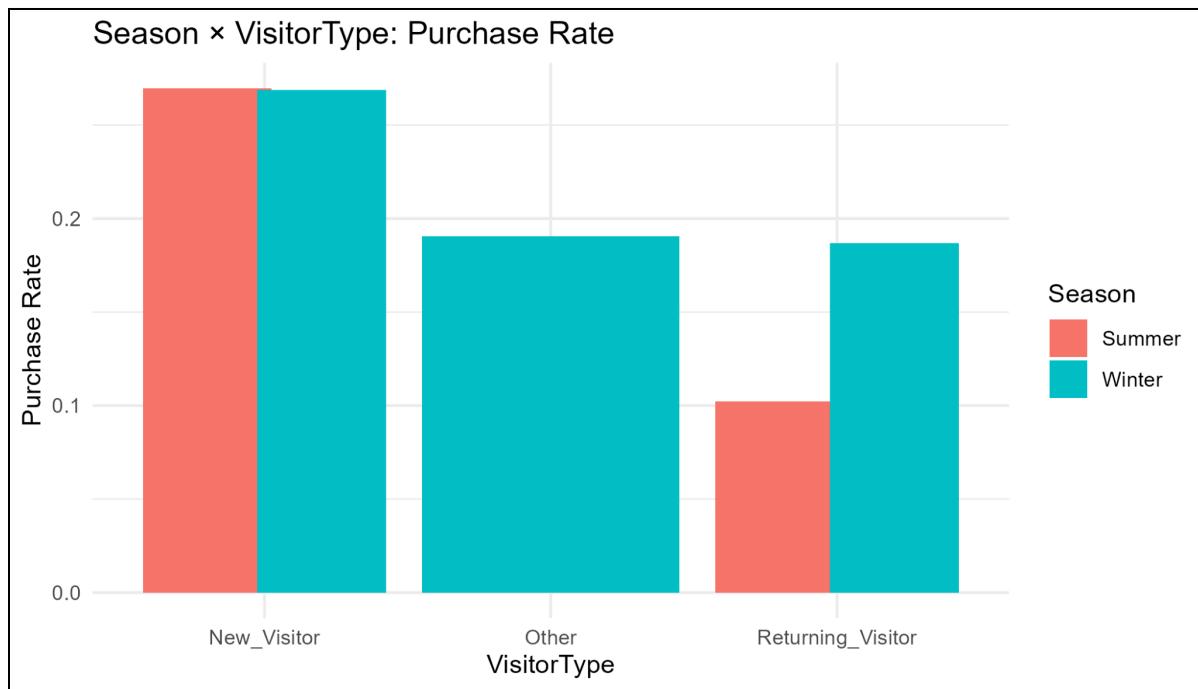
This aligns with well-known retail dynamics:

- Black Friday and Cyber Monday promotions
- Holiday gift shopping
- Year-end discount cycles
- Higher consumer urgency and motivation in Winter

In contrast, Summer tends to show more casual browsing and lower intent.

3.2 Season × Visitor Type Interaction

Next, we explored whether these seasonal patterns apply uniformly across shopper types. The dataset includes three groups: New Visitors, Returning Visitors, and Other.



Interpretation of Interaction Results:

- Returning Visitors show the most dramatic seasonal increase. Winter boosts their purchase rate by roughly 8–10 percentage points, suggesting that loyal users become especially goal-oriented during the holiday period.
- New Visitors display relatively similar conversion rates across seasons. This implies that first-time engagement is less sensitive to temporal incentives.
- Other Visitors (neither clearly new nor returning) show a moderate Winter uplift, though smaller than Returning Visitors.

Insight:

Seasonal effects *amplify existing behavioral tendencies*. Users who already have stronger intent (Returning Visitors) respond more to seasonal cues.

3.3 Seasonal Logistic Regression

To quantify seasonal behavioral sensitivity more precisely, we fit separate logistic regression models for Summer and Winter.

Key predictors included:

- ProductRelated (number of product-focused pages visited)
- ProductRelated_Duration
- BounceRates

Instead of showing raw regression output, we summarize the essential findings below.

<SummerRegression>

```
Call:  
glm(formula = Revenue_num ~ ProductRelated + ProductRelated_Duration +  
  BounceRates, family = "binomial", data = df_season[df_season$Season ==  
  "Summer", ])  
  
Coefficients:  
              Estimate Std. Error z value Pr(>|z|)  
(Intercept) -1.591e+00 6.486e-02 -24.530 <2e-16 ***  
ProductRelated -9.393e-04 2.091e-03 -0.449 0.653  
ProductRelated_Duration 6.687e-05 4.700e-05 1.423 0.155  
BounceRates    -4.482e+01 5.021e+00 -8.927 <2e-16 ***  
---  
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 3101.5 on 4228 degrees of freedom  
Residual deviance: 2878.8 on 4225 degrees of freedom  
AIC: 2886.8  
  
Number of Fisher Scoring iterations: 7
```

<WinterRegression>

```
Call:  
glm(formula = Revenue_num ~ ProductRelated + ProductRelated_Duration +  
    BounceRates, family = "binomial", data = df_season[df_season$Season ==  
    "Winter", ])  
  
Coefficients:  
              Estimate Std. Error z value Pr(>|z|)  
(Intercept) -1.491e+00 5.298e-02 -28.142 < 2e-16 ***  
ProductRelated 6.202e-03 1.475e-03  4.206 2.60e-05 ***  
ProductRelated_Duration 4.034e-05 3.646e-05  1.106  0.269  
BounceRates    -2.431e+01 3.227e+00 -7.533 4.96e-14 ***  
---  
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 4905.2 on 4908 degrees of freedom  
Residual deviance: 4536.2 on 4905 degrees of freedom  
AIC: 4544.2  
  
Number of Fisher Scoring iterations: 7
```

Summary of Key Coefficients

Predictor	Summer	Winter	Interpretation
ProductRelated	Not significant	Positive & highly significant	Winter shoppers convert more when engaging with more product pages
ProductRelated_Duration	Not significant	Not significant	Time spent alone does not explain seasonal differences
BounceRates	Negative & significant	Negative & significant	High bounce behavior strongly reduces conversion in all seasons

Interpretation: Seasonal Behavioral Sensitivity

The contrast in ProductRelated significance is particularly revealing:

- In Summer, browsing more products does *not* meaningfully increase the chance of conversion — users may be casually exploring without strong purchase motivation.
- In Winter, the same behavior sharply increases the likelihood of purchase — indicating that Winter shoppers browsing multiple products are in a more deliberate, goal-driven decision mode.

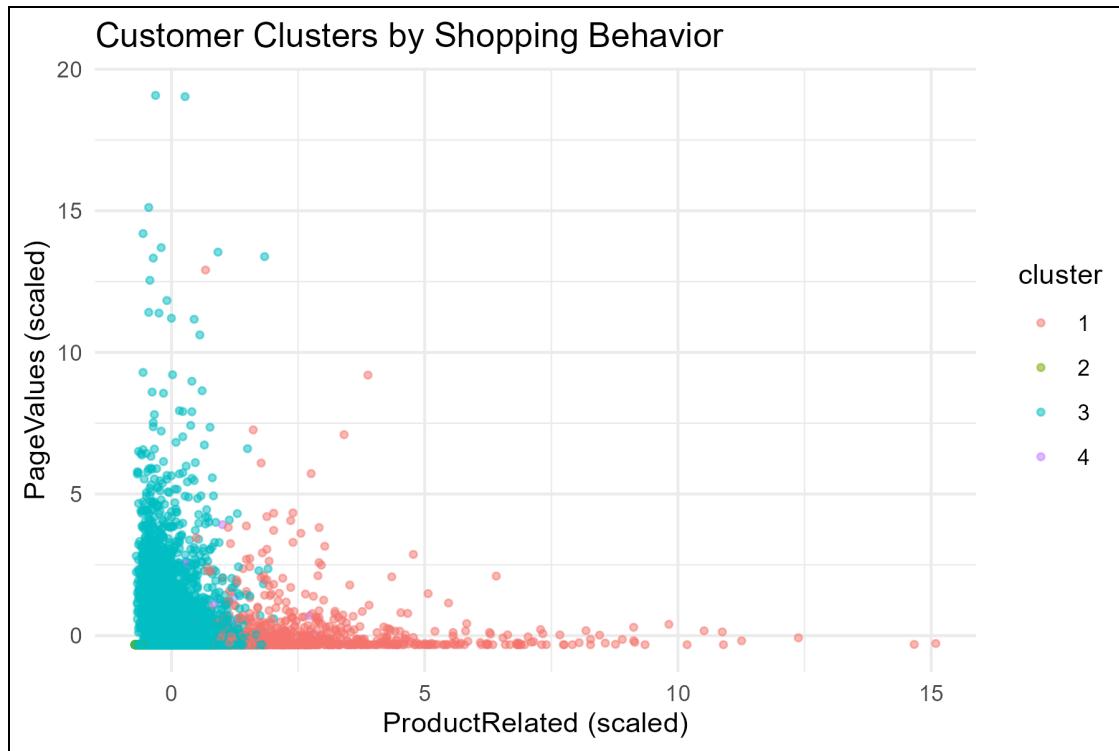
BounceRates remain a universally negative signal across both seasons, confirming that early disengagement consistently predicts non-purchase.

3.4 Clustering Analysis: Identifying Shopper Segments

To uncover deeper behavioral structures, we performed K-Means clustering ($k = 4$) on six standardized features:

- ProductRelated
- Duration
- BounceRates
- ExitRates
- PageValues
- SpecialDay

Cluster assignments were merged with purchase outcomes to compute conversion rates.



Cluster Summary

Cluster	Size (n)	Purchase	Description
		Rate	
1	842	33.4%	High-value shoppers; strong PageValues, moderate product engagement
2	847	0.7%	Bouncers; high ExitRates/BounceRates
3	9,560	16.3%	Browsers; general-purpose, moderate engagement
4	956	6.4%	Early holiday explorers; high SpecialDay scores and low engagement

Cluster Insights

- Cluster 1 users exhibit high intent, consistent with shoppers who compare products carefully or arrive with clear purchase goals.
- Cluster 2 represents almost complete non-buyers; their engagement patterns indicate strong misalignment between user expectations and site content.
- Cluster 3 forms the backbone of typical browsing behavior with moderate conversion.
- Cluster 4 captures seasonal deal-seekers who may be surveying promotions without immediate intent.

3.5 Integrating Seasonal and Behavioral Findings

Across the seasonal and clustering analyses, several themes emerge:

1. Winter amplifies purchase motivation, especially among Returning Visitors and high-engagement sessions.
2. Product engagement becomes a more powerful predictor in Winter, reinforcing the idea that seasonal shoppers are more intentional.
3. Clusters with higher PageValues and ProductRelated activity (especially Cluster 1 and parts of Cluster 3) align closely with the behaviors that the Winter regression model identifies as significant.
4. Seasonal uplift is not uniform — loyal users and highly engaged browsers are the most sensitive to seasonal incentives, while low-intent users remain largely unaffected.

3.6 Conclusion

Seasonality and behavioral context jointly shape purchasing intention in meaningful ways. Winter increases overall conversion probability, strengthens the predictive power of product-engagement features, and disproportionately boosts conversion among loyal and deliberate shoppers. At the same time, clustering reveals sharp heterogeneity in

behavioral profiles, reaffirming that seasonal effects interact strongly with underlying user intent.

These insights suggest that season-aware and segment-specific targeting strategies—such as prioritizing Returning Visitors in Winter and tailoring engagement prompts to high-value clusters—could significantly improve digital marketing effectiveness.

4. Conclusion and Future Work

4.1 Conclusion

Across the three research questions, our project provides a comprehensive understanding of the behavioral, contextual, and seasonal factors that drive online purchasing intention. By combining statistical analysis, machine-learning-based prediction models, and behavioral segmentation techniques, we uncover a consistent narrative about how users browse, engage, and ultimately decide whether to make a purchase.

First, in **Question 1**, we found overwhelming evidence that *PageValues* is the single strongest predictor of conversion across all analytic approaches—including correlation analysis, logistic regression, Random Forest, XGBoost, and SHAP. *PageValues* effectively captures a user’s proximity to high-intent pages, making it a uniquely powerful indicator of imminent purchase behavior. Additional features such as *ExitRates* and *BounceRates* also consistently emerged as strong negative predictors, while deeper engagement with product-related content positively influenced purchase likelihood. These results highlight that both the *quality* and *depth* of user engagement are essential in understanding purchasing intent.

In **Question 2**, we evaluated a diverse set of classification models to predict whether a session ends in a purchase. After careful handling of class imbalance, extensive preprocessing, and hyperparameter tuning, Gradient Boosting achieved the strongest

overall performance with an ROC-AUC of 0.933 and the best precision among all models. Although other models—such as Random Forest—excelled in recall, the Gradient Boosting model provided the most balanced tradeoff, making it the most reliable classifier for this dataset. These predictive insights offer practical value for e-commerce platforms, enabling targeted interventions such as personalized promotions and improved inventory planning based on real-time user behavior.

In **Question 3**, we extended our analysis into temporal and behavioral contexts, revealing distinct seasonal dynamics. Winter shoppers exhibited nearly double the purchase probability of Summer shoppers, with Returning Visitors showing the most dramatic seasonal increase. Seasonal logistic regression further demonstrated that ProductRelated engagement becomes significantly more predictive in Winter, suggesting that shoppers' intent intensifies during the holiday period. Complementing this, our clustering analysis identified four latent shopper segments that differ sharply in engagement style and purchase rates. High-value shoppers (Cluster 1) and moderate-engagement browsers (Cluster 3) aligned closely with the behavioral patterns associated with Winter conversions, while bouncers (Cluster 2) remained non-responsive across seasons.

Taken together, these findings reveal that purchasing intention is shaped by the interaction of behavioral depth, contextual engagement, and seasonal motivation. Shoppers do not behave uniformly; the same behavioral signals can carry different meanings depending on the time of year and the shopper's underlying intent. For practitioners, this suggests that effective e-commerce strategy requires both *feature-aware* and *season-aware* decision-making. Tailoring website design, promotional strategies, and targeting algorithms to specific user segments—and adjusting them seasonally—can substantially improve conversion outcomes.

Overall, our project demonstrates that integrating machine learning with behavioral insight provides a powerful toolkit for understanding and predicting online shopper behavior. The multi-method approach ensures robust conclusions, while the actionable

implications offer clear pathways for enhancing digital marketing effectiveness and user experience design.

4.2 Future Work

In the future, there are a few steps we would like to take to make our findings more conclusive and insightful. First, it could be interesting to experiment with different approaches to feature engineering and see how this affects our results. For example, we could add a feature for `Avg_Time_Per_Product_Page` that uses `ProductRelated_Duration` and `ProductRelated` to give us more information into whether a user is rushing through pages or reading carefully and how this affects purchasing behavior. Next, we may attempt to analyze how `OperatingSystem` and `Browser` affects user purchasing behavior. There are certain observations in e-commerce that users from specific operating systems such as iOS may be more able to make purchases so seeing the effect that these variables have on purchasing outcomes may provide some insights. Additionally, it may be useful to experiment with a neural network based approach to see if using these classes of models can provide more robust predictions of user purchasing behavior from our dataset. Lastly, we could also experiment with using an ensemble of our Random Forest, Logistic Regression, and Gradient Boosting to create a voting system where we classify a record based on the majority vote from these three models. This may help to smooth out errors and improve our predictive power.

5. Contribution Statement

All group members contributed meaningfully to the development of the final project. The primary responsibilities are summarized below:

- Jubin Joung
Led the data preprocessing pipeline, including cleaning, feature engineering, and transformation procedures. Conducted the full set of analyses for Question 1, including correlation analysis, logistic regression, Random Forest, XGBoost, and SHAP interpretability, and wrote the corresponding sections of the report.
- Satvik Mahendra
Led the modeling workflow for Question 2, including handling class imbalance, training and tuning classification models, and generating performance comparisons. Also contributed to writing the project Introduction and refining the problem framing.
- Wonjun Ryhu
Conducted all analyses for Question 3, including seasonal modeling, interaction effects, and clustering. Produced the visualizations for seasonal and behavioral analyses and coordinated the overall structure, editing, and integration of the full report. Also prepared materials for the presentation.

All members participated in discussions of methodology, interpretation of results, and final report review. The project reflects a collaborative and balanced team effort.