Predicting Online Shoppers’ Purchasing Intention Using Behavioral and Contextual Features

Group 11 : Satvik, Jubin, Wonjun

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.

Data

- Dataset Name: Online Shoppers Purchasing Intention
- Source: University of California, Irvine Machine Learning Repository
- Link: https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset

Dataset Description

- # of instances: 12,330 sessions
- # of features: 17 variables (10 numerical, 8 categorical)
- Target variable: `Revenue` (1 = purchase, 0 = no purchase)
- Class distribution:
  - 84.5% no purchase
  - 15.5% purchase → strong class imbalance
- Variables reflect 3 key categories:
  1. Page interaction metrics (Administrative, Informational, ProductRelated & Duration)
  2. Engagement metrics (BounceRates, ExitRates, PageValues)
  3. Time/context variables (Month, Weekend, SpecialDay, VisitorType, TrafficType)

Data Pre-processing:

We performed basic data preprocessing before starting the analysis. The original dataset was already quite clean, so only minimal steps were required. Specifically, we removed 125 duplicate rows, converted boolean variables (*Weekend* and *Revenue*) into numeric values (0/1), handled missing values (none were found), and standardized all numeric features. Categorical variables such as *Month* and *VisitorType* were encoded using one-hot and cyclical (sin/cos) transformations to preserve their structure. As a result, the final preprocessed dataset consists of 12,205 records and 21 fully numeric features, ready for modeling. The class distribution for the target variable (*Revenue*) is slightly imbalanced (around 15.6% purchases), which will be addressed later in the modeling stage.

Questions

Question 1: What attributes are the most relevant for classifying whether a shopper makes a purchase?

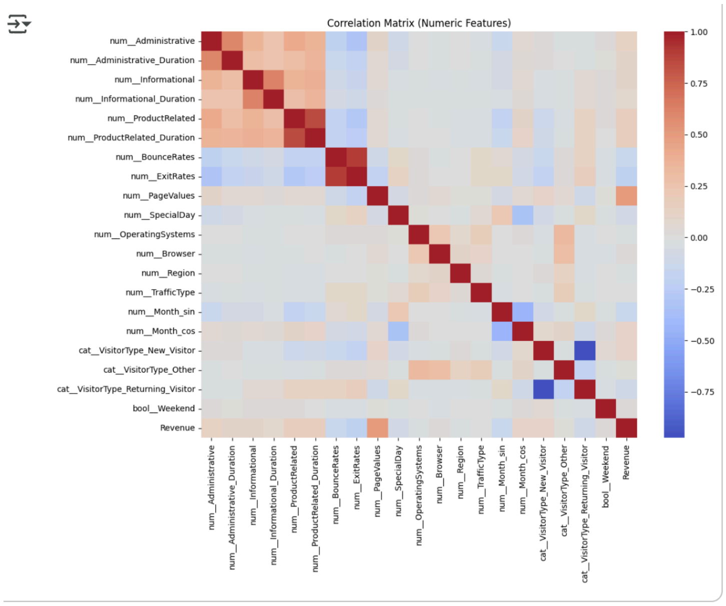


Figure 1. Correlation matrix of numeric features (red = positive, blue = negative correlations).

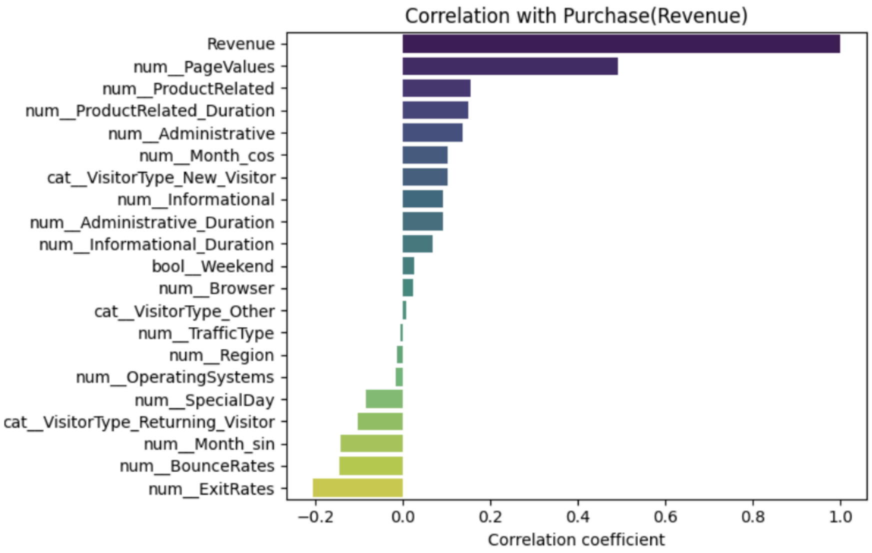


Figure 2. Correlation coefficients of all features with the target variable (*Revenue*).

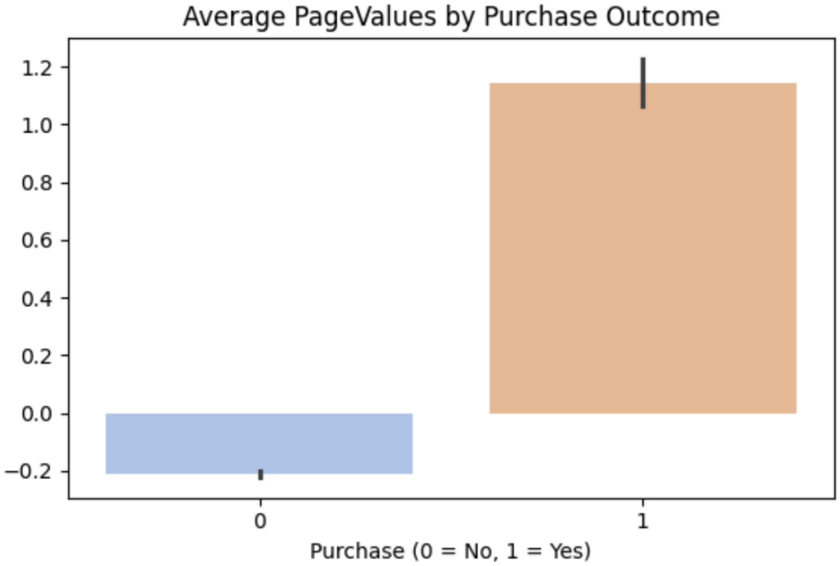


Figure 3. Average *PageValues* by purchase outcome, showing higher values for sessions that resulted in a purchase.

To gain an initial understanding of which attributes are most relevant for classifying whether a shopper makes a purchase, a brief exploratory analysis was conducted on the preprocessed dataset.

Correlation Analysis.

Figure 1 presents the correlation matrix for all numeric features. Overall, *PageValues*, *ProductRelated*, and *ProductRelated_Duration* show positive relationships with *Revenue*, whereas *ExitRates* and *BounceRates* are negatively correlated. This pattern suggests that sessions involving more product-related engagement and higher page value are more likely to lead to a purchase, while sessions that end quickly are less likely to convert.

Figure 2 illustrates each feature's correlation with the purchase variable (*Revenue*). The results confirm that **PageValues (r = 0.49)** is the most influential feature, followed by **ProductRelated (r = 0.16)** and **ProductRelated_Duration (r = 0.15)**. On the other hand, **ExitRates (r = –0.20)** and **BounceRates (r = –0.15)** are the strongest negative predictors, implying that high abandonment rates reduce purchase likelihood. Weak negative correlations for **SpecialDay (r = –0.08)** and **Month_sin (r = –0.14)** indicate mild seasonal effects.

Feature-Level Visualization.
To visualize these relationships more intuitively, Figure 3 compares the average *PageValues* between sessions that resulted in a purchase and those that did not. The difference is clear, sessions associated with purchases have substantially higher average *PageValues*. This aligns with the correlation results and highlights *PageValues* as a strong indicator of purchase intention.

Overall, these preliminary findings provide a clear direction for further modeling. Subsequent steps will include more rigorous feature selection and model-based importance evaluation to verify these early observations.

To further analyze the factors that influence online purchase behavior, we plan to apply a combination of supervised and unsupervised learning techniques.

Dimensionality Reduction (PCA).
Principal Component Analysis will be used to reduce dimensionality and identify the key components that explain the majority of variance in the dataset. This step will help detect underlying relationships between features and simplify subsequent modeling.

Regression Models.
We will begin with Multiple Linear Regression to examine linear relationships between predictor variables and purchase behavior. Since the target variable (*Revenue*) is binary, Logistic Regression will also be applied as a more appropriate classification model. The results from both models will help interpret which features most strongly predict purchasing intent.

Tree-based Models.
To capture nonlinear relationships and feature interactions that linear models may miss, we will explore ensemble methods such as Random Forest or Gradient Boosting. Feature importance from these models will be compared with correlation and regression results for validation.

Unsupervised Learning (Clustering).
Finally, we will conduct clustering analysis (e.g., K-Means) to segment shoppers into groups based on behavioral similarities. This step will provide additional insight into different types of user journeys and their likelihood of conversion.

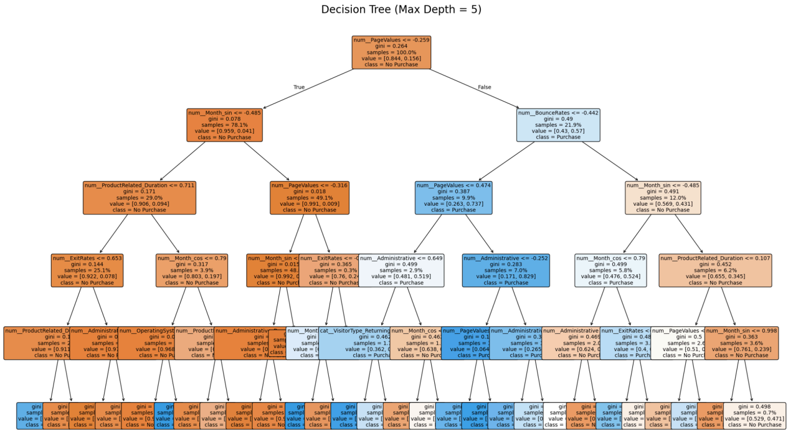
Evaluation.
Model performance will be assessed using metrics such as ROC-AUC, F1-score, and precision/recall to account for the slight class imbalance. Stratified cross-validation will be applied to ensure robustness and reliability of results.

<u>Question 2: Based on the features in the dataset, will a user make a purchase?</u>

To determine whether a user will make a purchase based on the features in the dataset, we conducted preliminary analysis by fitting a decision tree and SVM as a proof of concept that these classification models could effectively classify whether a user will make a purchase.

After fitting a decision tree on the preprocessed dataset, we were able to achieve an accuracy of 90%. This is promising and suggests that we will be able to classify whether a user makes a purchase based on the features in the dataset. However, one big concern is that since there is massive class imbalance in the data, we have a poor recall of 0.58.

We tried to resolve this by setting the class_weight parameter to 'balanced' to more heavily penalize mistakes on the minority class. This resulted in a drop in accuracy to 85% but improved recall to 0.87. However, we witnessed the precision-recall tradeoff and saw precision drop as a result. This model is far superior for the business goal of identifying the maximum number of potential buyers, while the unbalanced model is only useful for finding a smaller, more reliable subset of those buyers.

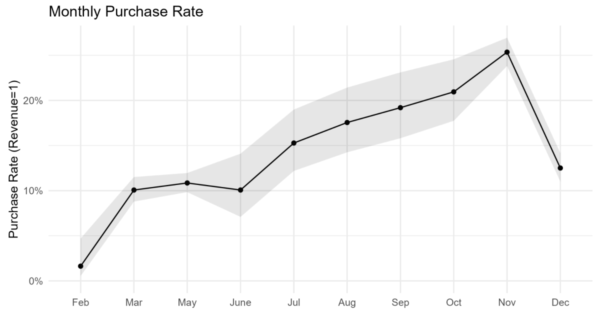


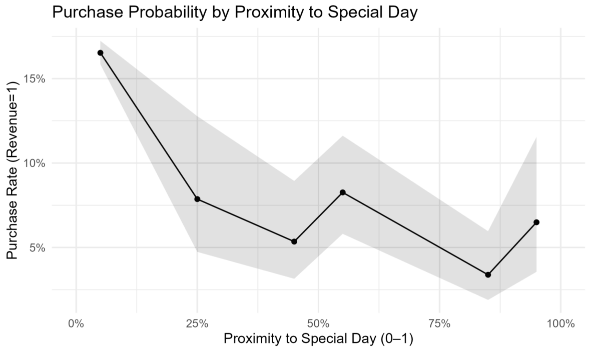
The visualization above shows the nodes and splits in the decision tree after completing training.

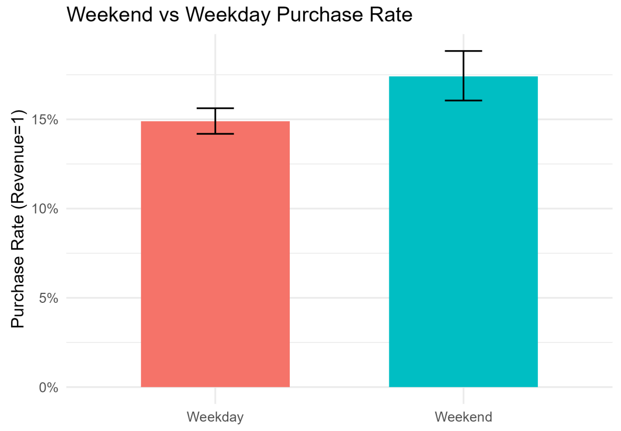
Next, we attempted to classify whether a user makes a purchase using a Support Vector Machine (SVM). This first involved scaling the data as doing so helps when working with SVMs. Then, after fitting the SVM, we were able to get an accuracy of 89%. However, due to there being a class imbalance in our data, our recall for users who actually made a purchase was 0.54 and our precision was 0.73. Thus, despite having high accuracy, the model struggled with the class that we had less data on. To attempt to address this, we set the class_weight parameter to be ‘balanced’ to penalize mistakes on the minority class more heavily. This resulted in a drop in accuracy to 86%, an improvement in recall from 0.54 to 0.74, but a drop in precision from 0.73 to 0.54. This makes the model a bit more balanced but still has tradeoffs that we will need to continue addressing. We could try tuning hyperparameters or use more advanced sampling techniques in the future to fix this.

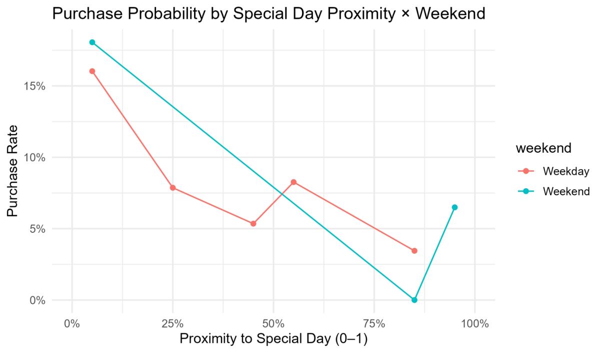
Overall, we are able to make predictions about whether a user will make a purchase based on the features in the dataset as indicated by the initial experimentation with the models described above. However, we will continue this analysis with further model experimentation to get more confident predictions about user purchase intent.

Question 3: How do temporal factors such as proximity to a ‘Special Day’, weekends, and months affect purchasing intention?









(Since the preprocessed dataset primarily serves for model consistency, we used the original dataset for temporal visualizations to preserve the full variability of special-day and time-related features.)

To explore how temporal patterns influence online purchasing behavior, we conducted a preliminary analysis using visualizations and a simple logistic regression model. As shown in the **Monthly Purchase Rate** plot, purchase probability appears to increase steadily throughout the year, reaching its highest levels in November before dropping slightly in December. This trend may reflect stronger consumer motivation during end-of-year holidays and promotional events.

From the **Special Day Proximity** plot, purchase likelihood does not linearly increase as a special day approaches. Instead, higher conversion rates are observed when users are far from such days, suggesting early preparation behaviors or that most browsing sessions naturally occur on regular, non-event days.

The **Weekend vs Weekday** plot shows that weekend sessions exhibit slightly higher purchase rates than weekday sessions, which may indicate that users are more relaxed or have greater time availability to complete transactions on weekends. However, as illustrated in the **Special Day × Weekend** interaction plot, the weekend effect weakens when special days are near, possibly because many users browse or compare products during event weekends but postpone actual purchases until later.

Overall, these preliminary findings highlight meaningful temporal dynamics in consumer behavior. In the full report, we plan to extend this analysis using more advanced models to better capture nonlinear seasonal effects and interaction terms.