UT AUSTIN

Online Shoppers' Purchasing Intention

by Group 11

ELEMENTS OF STAT MACH LEARNING
FINAL PROJECT

Presented by
Jubin Joung, Satvik Mahendra, Wonjun Ryhu

Get Started

AGENDA

1. Data Overview

2. Data Processing

3. Research Questions

Q1. What attributes are the most relevant for classifying whether a shopper makes a purchase?

Q2. Based on the features in the dataset, will a user make a purchase?

Q3. How Do Seasonal and Behavioral Contexts Influence Purchasing Intention?

Get Started

Data Overview

Dataset: Online Shoppers Purchasing Intention

- 12,330 sessions of online retail website behavior
- Each row = one user session, not a specific user
- Goal = Predict whether the session leads to a purchase (Revenue)

Feature Composition

- **Behavioral Metrics**
  - Administrative, Administrative_Duration
  - Informational, Informational_Duration
  - ProductRelated, ProductRelated_Duration
  - These represent what types of pages the user visited and how long they stayed.
- **Engagement / Drop-off Indicators**
  - BounceRates
  - ExitRates
  - Reflect abandonment behavior, often associated with lower purchase likelihood.

Feature Composition

- **Economic Indicator**
  - PageValues
  - Represents the estimated value of a page based on historical conversion — key metric.
- **Seasonality / Timing**
  - Month
  - Weekend
  - Capture when the session occurred, allowing insight into seasonal patterns.
- **Visitor Information**
  - VisitorType (Returning vs New)
  - Region, TrafficType, Browser, OperatingSystems

Data Processing

1. Data Cleaning

- Removed duplicate rows
  → (No duplicates found: 12,205 rows → 12,205 rows)
- Converted Weekend, Revenue into binary (0/1) format
  → Ensures consistent numeric representation for ML models

2. Feature Grouping

- **Numeric Features**
  - Administrative, Informational, ProductRelated counts and durations, BounceRates, ExitRates, PageValues, SpecialDay, OS/Browser/Region/TrafficType
- **Categorical Features**
  - Month, VisitorType
- **Boolean Features**
  - Weekend (already 0/1)

Data Processing

3.Cyclical Encoding for Seasonality

- Replaced "Month" with sin/cos encoding

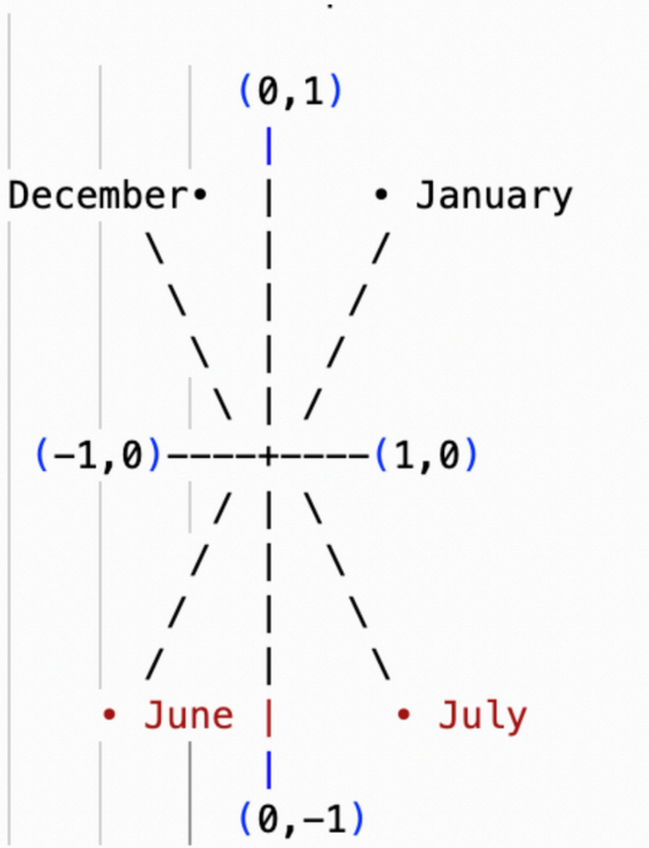
→ Captures seasonal pattern

→ Prevents artificial distance between December (12) and January (1)

Month	sin(2πm/12)	cos(2πm/12)
1 (Jan)	0.5	0.866
3 (Mar)	1.0	0.0
6 (Jun)	0.0	-1.0
9 (Sep)	-1.0	0.0
12 (Dec)	-0.5	0.866

Summer(June ~ August)
→ similar sin/cos patter

Winter(December ~ Feburary)
→ similar sin/cos pattern



Data Processing

4. Column-wise Transformations

- **Numeric**: Median imputation + StandardScaler
- **Categorical**: Mode imputation + One-Hot Encoding
- **Boolean**: Pass-through
- **Final Feature Count: 20 engineered features**

5. Output

- Built the final modeling dataset:
- **12,205 rows × 20 features**
- Saved as: online_shoppers_preprocessed.csv

Q1. What attributes are the most relevant for classifying whether a shopper makes a purchase?

Methods Used

Correlation Analysis
: linear association

Logistic Regression
: directional effect on purchase probability

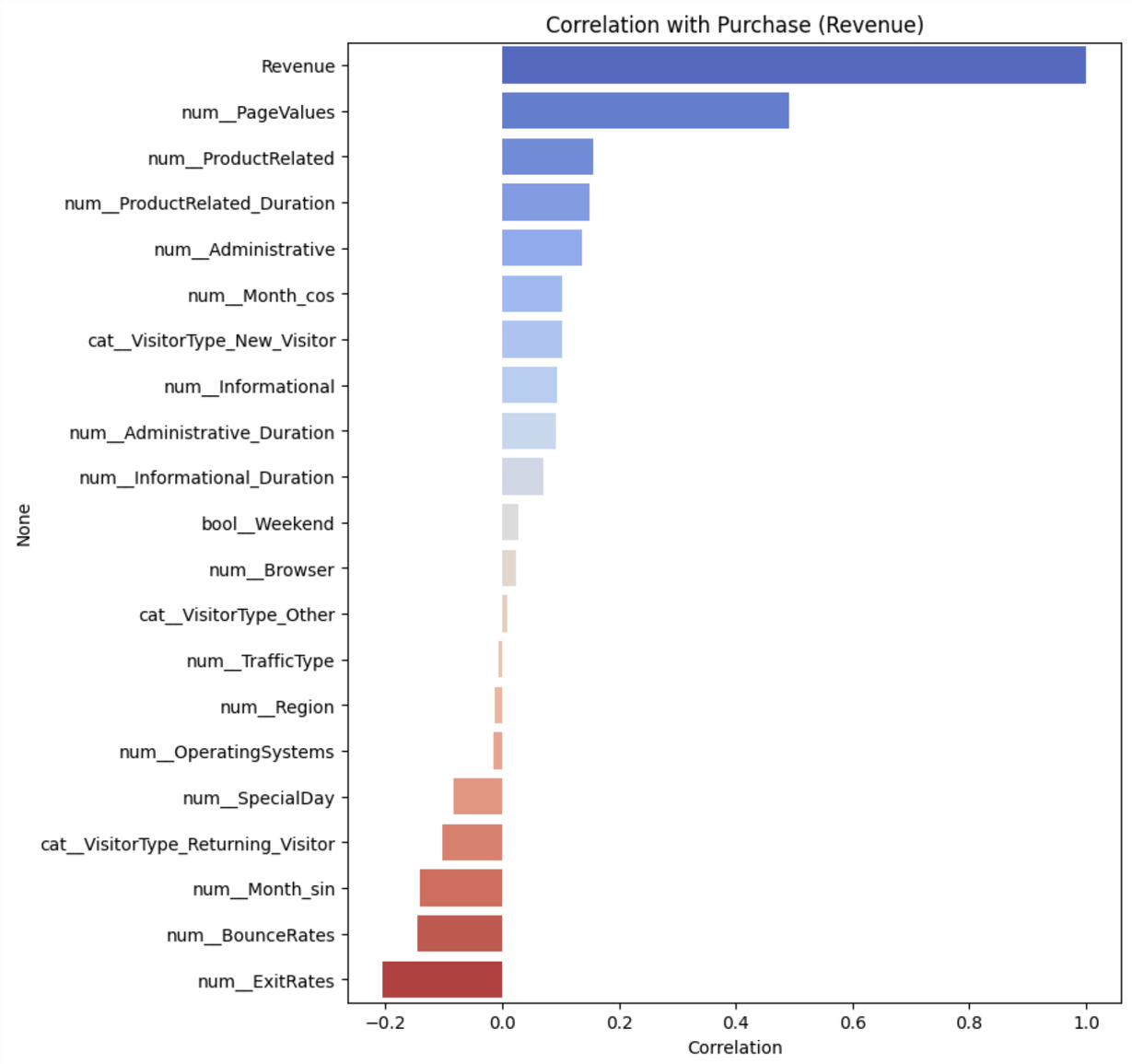
Random Forest
: non-linear feature importance

XGBoost
: boosted trees with high predictive power

SHAP Values
: model-agnostic feature attribution

We use multiple approaches to ensure consistent and reliable results

Correlation with Purchase

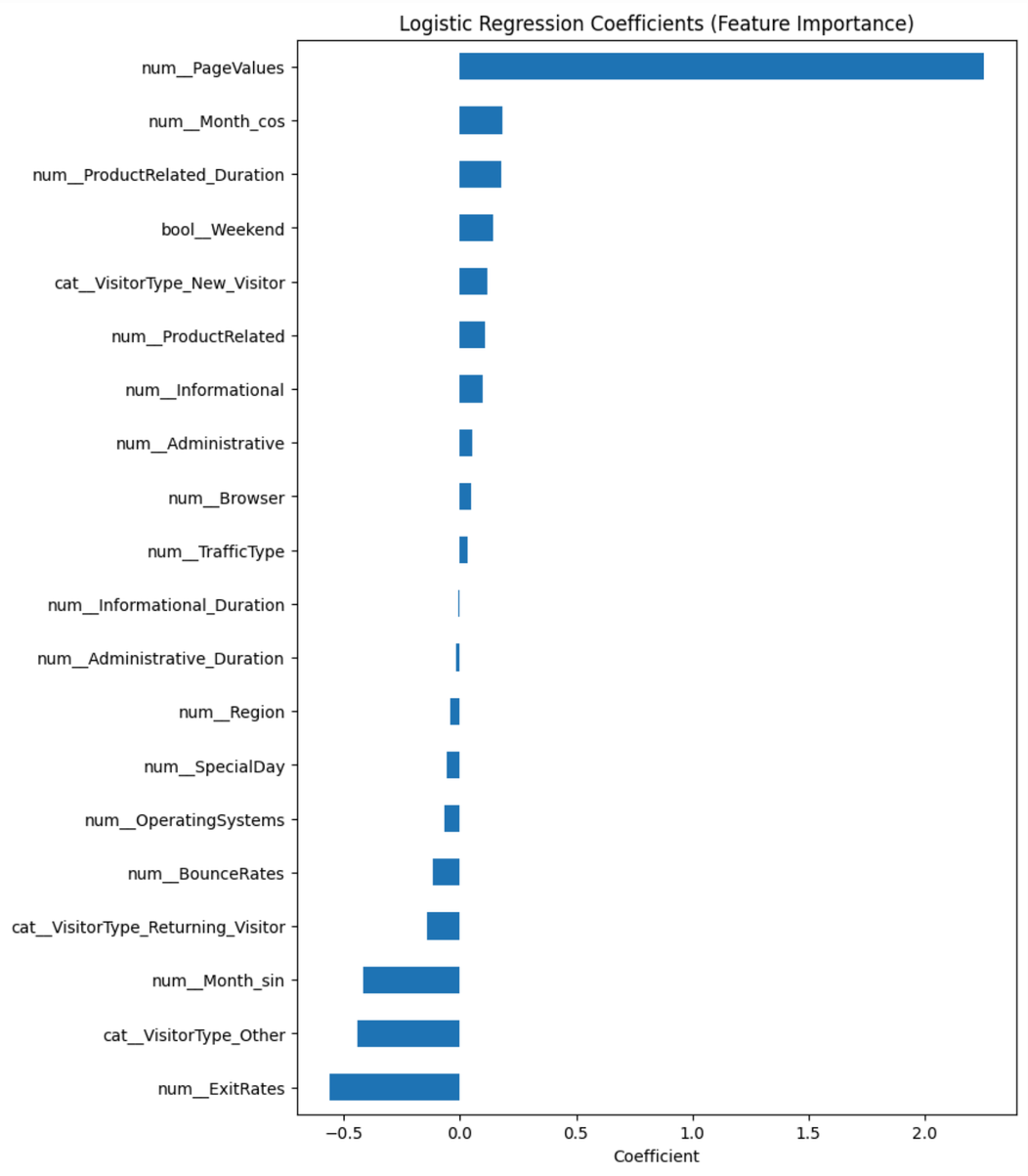


Key Findings

PageValues shows the highest positive correlation with purchase

ExitRates and **BounceRates** show substantial negative correlation

Time spent on product-related pages increases likelihood of purchase



Logistic Regression Coefficients

Interpretation

PageValues is the strongest positive predictor of purchase

ExitRates, BounceRates, Month_sin → strongly negative impact

Logistic regression confirms the correlation findings

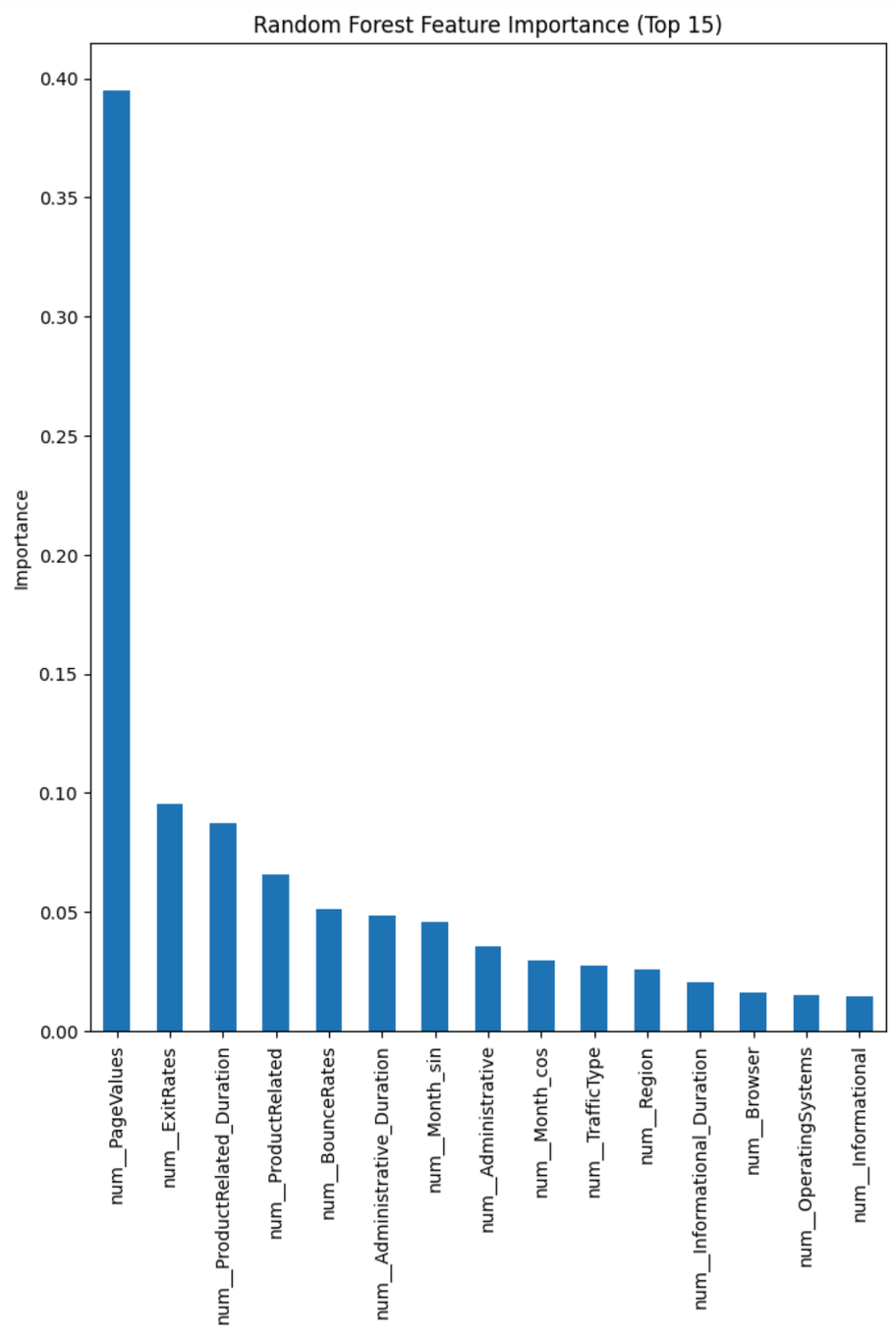
Model Performance: ROC-AUC = 0.9021401381078409

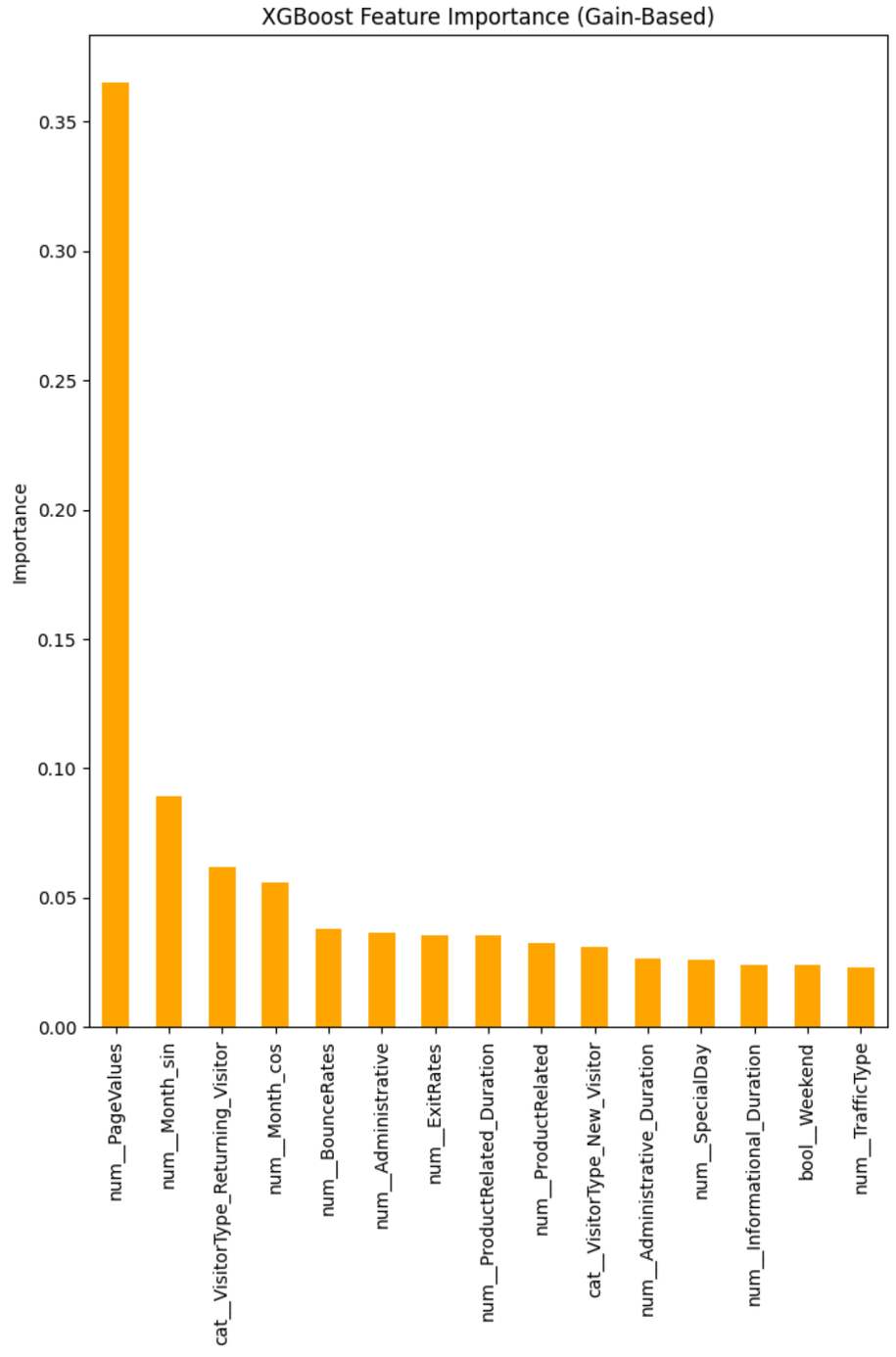
Random Forest Feature Importance

Top Features Identified

1. PageValues
2. Exit Rates
3. ProductRelated_Duration
4. Product Related
5. Bounce Rates

→ Non-linear model again identifies PageValues as overwhelmingly important





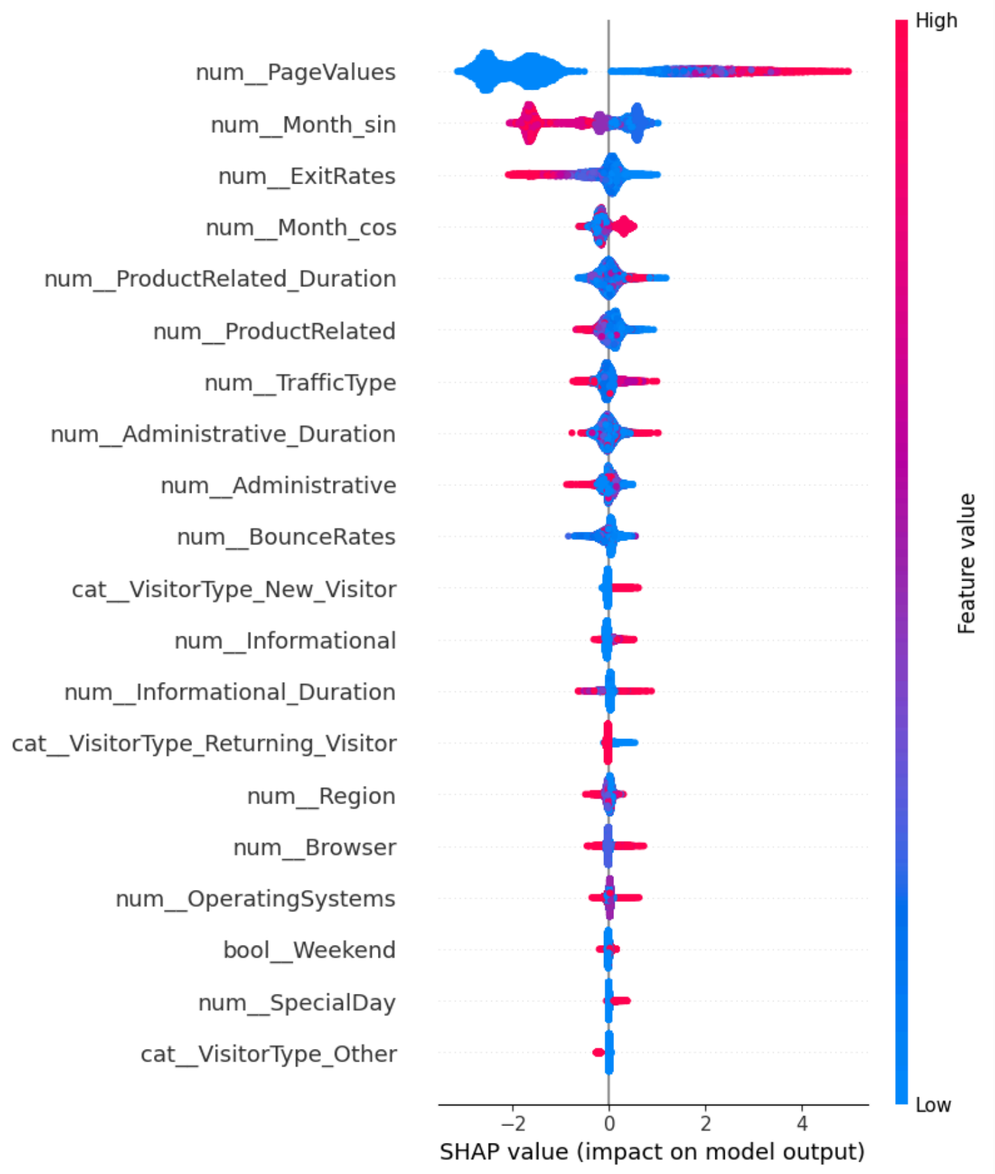
XGBoost Feature Importance

Insights

PageValues remains the dominant feature

Seasonal pattern (Month_sin, Month_cos) appear meaningful

ExitRates and BounceRates consistently lower purchase probability



SHAP Summary Plot

Interpretation

High PageValues → consistently large positive SHAP values

High ExitRates/ BounceRates → consistently negative SHAP values

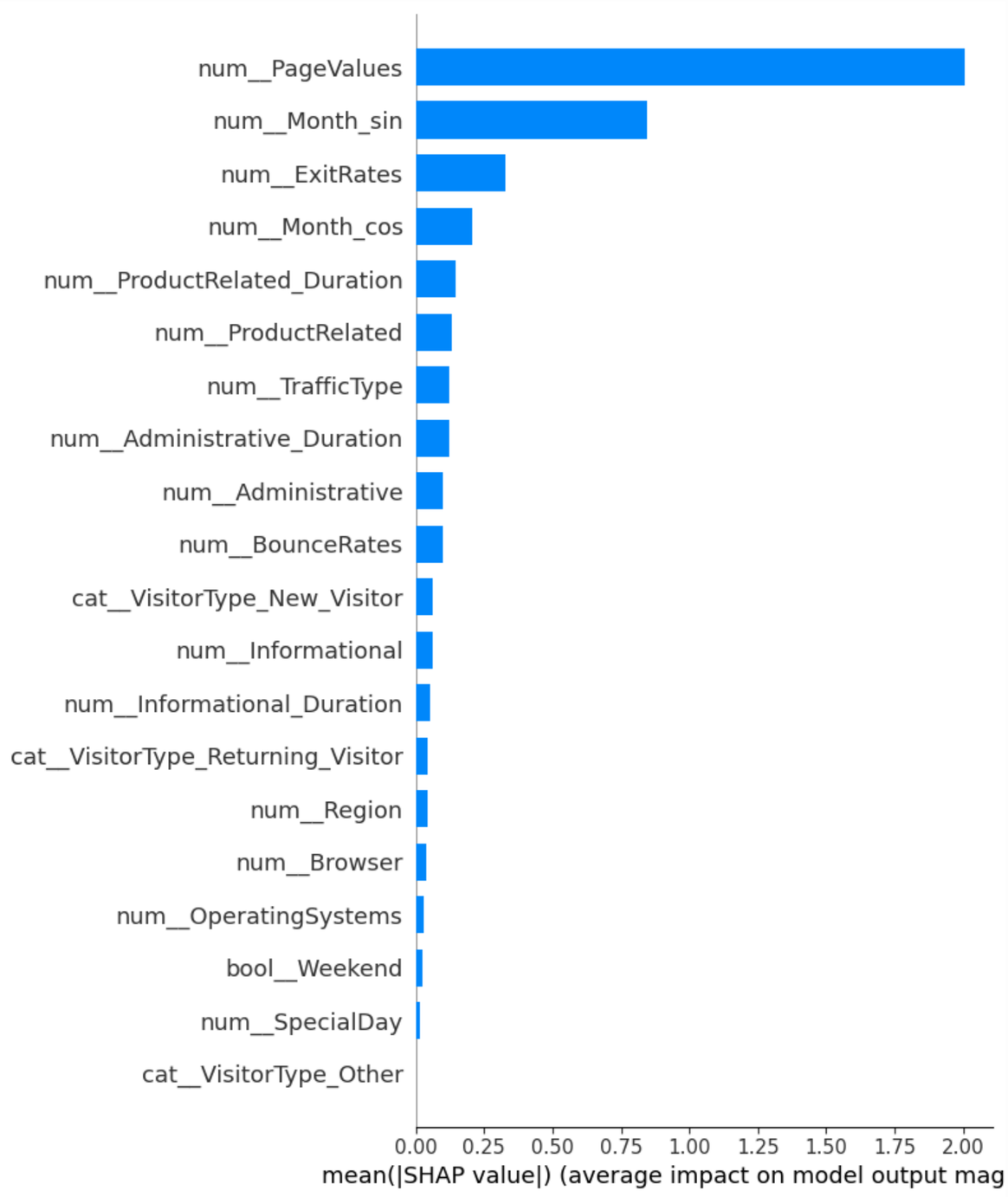
Model explanation aligns with tree-based importance

SHAP Bar Plot

Key Variables (Based on Model-Agnostic SHAP)

1. PageValues
2. Month_sin
3. ExitRates
4. Month_cos
5. Product-related variables

SHAP confirms the robustness of findings across methods



Most important Features Influencing Purchase

⭐ **1. PageValues — strongest, most consistent predictor**

💡 Represents estimated value of page interactions; highest driver of conversion.

⭐ **2. ExitRates**

💡 High exit rates strongly decrease purchase likelihood.

⭐ **3. ProductRelated / ProductRelated_Duration**

💡 Interest level and time spent on product pages increase purchase likelihood.

⭐ **4. BounceRates**

💡 Quick abandonment correlates with failed purchases.

⭐ **5. Seasonality (Month_sin, Month_cos)**

💡 XGBoost/SHAP show seasonal shopping patterns affecting purchases.

Q1 Conclusion

Answer

- Across all linear, non-linear, and model-agnostic analyses, PageValues emerges as the strongest predictor of purchase because it captures how close a user is to completing a transaction which means that pages with high value are historically visited right before conversion.

- Additional important attributes include ExitRates, BounceRates, ProductRelated behavior, and seasonality patterns.

Why this matters

- Helps improve website design
- Supports targeted marketing
- Optimizes user flow
- Improves conversion strategies

Q2. Based on the features in the dataset, will a user make a purchase?

Approaches

Experimenting with Various Models

Comparing Model Strengths

Defining Relevant Metrics

Handling Class Imbalances

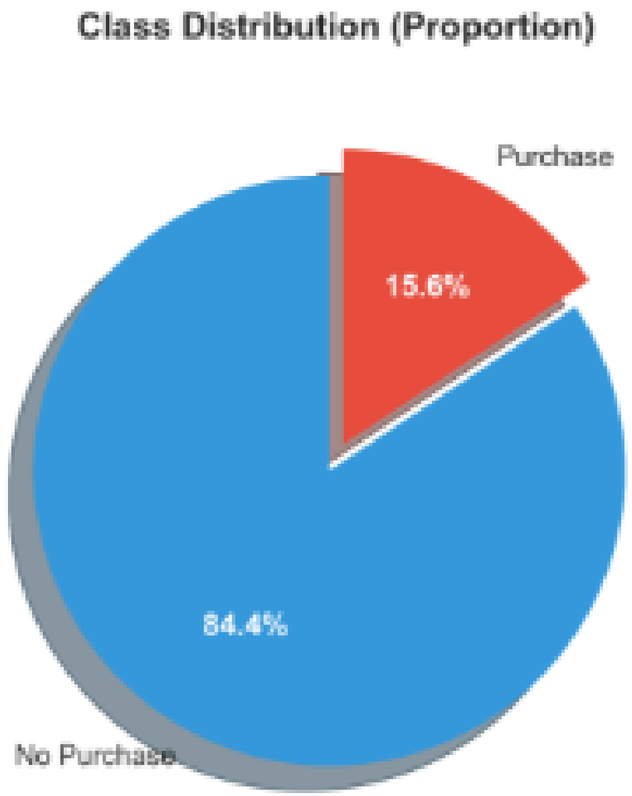
We use multiple models to determine the best for making predictions

Evaluation Metrics

To confidently answer the question, we need to evaluate our models on relevant metrics. Accuracy alone is not a robust enough metric because it misses potential bias in our models

- ROC-AUC
- Precision
- Recall
- F1-Score

Class Imbalances



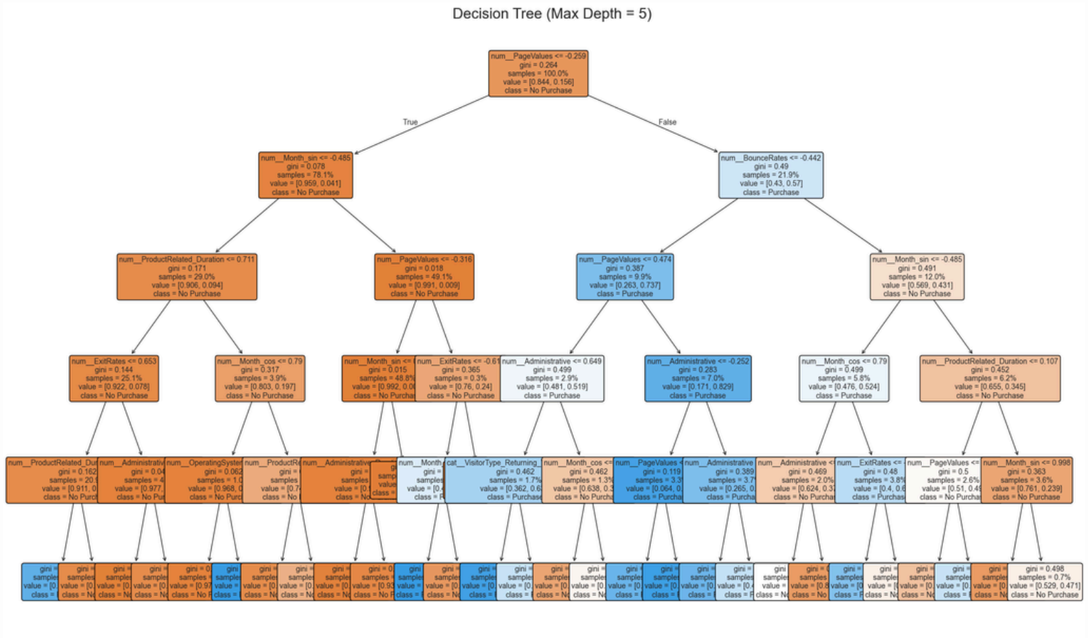
Since there is a large **class imbalance** in the data, we need to address this properly to avoid bias in our model

- ‘class_weight = balanced’ parameter
- Hyperparameter tuning
- SMOTE (Synthetic Minority Oversampling Technique)

Models Used

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

Ex: Decision Tree



Findings

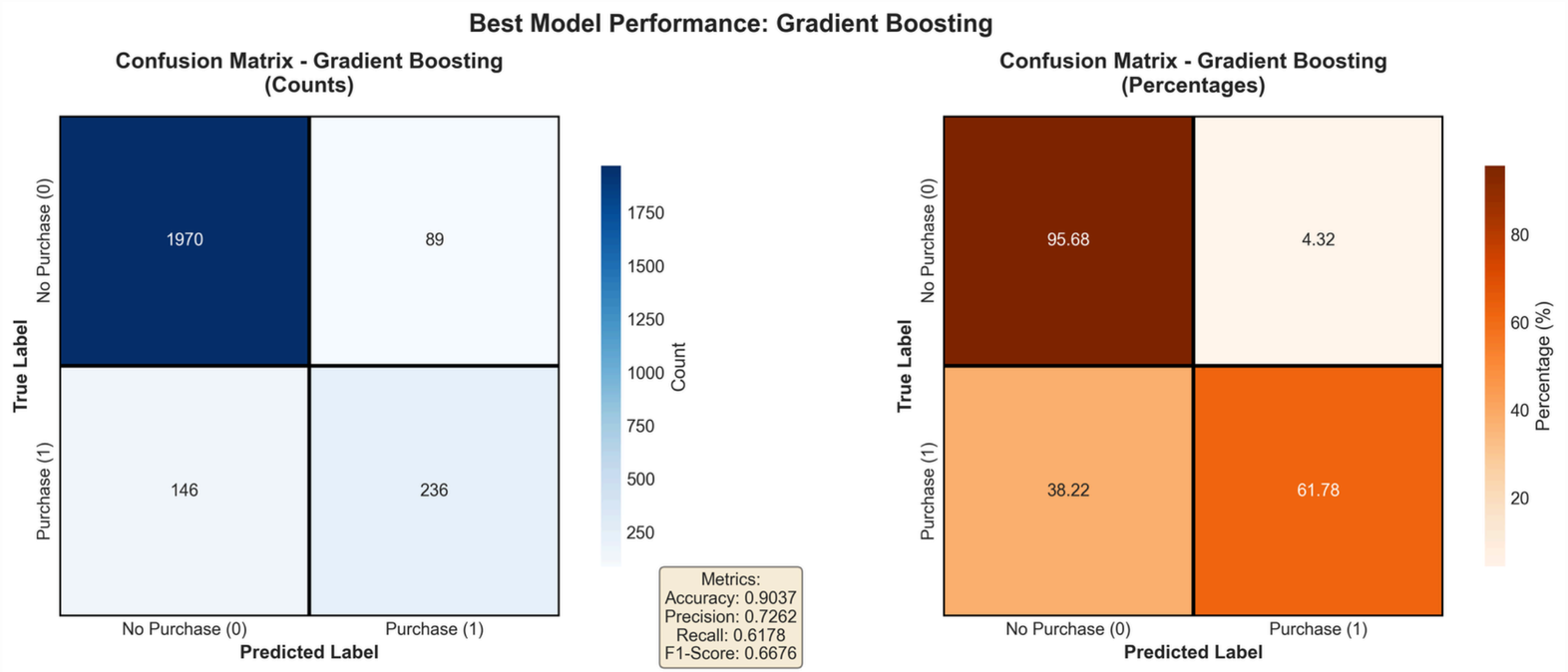
Accuracy was initially 90% but recall was high due to class imbalance

Correcting for class imbalance resulted in 85% accuracy

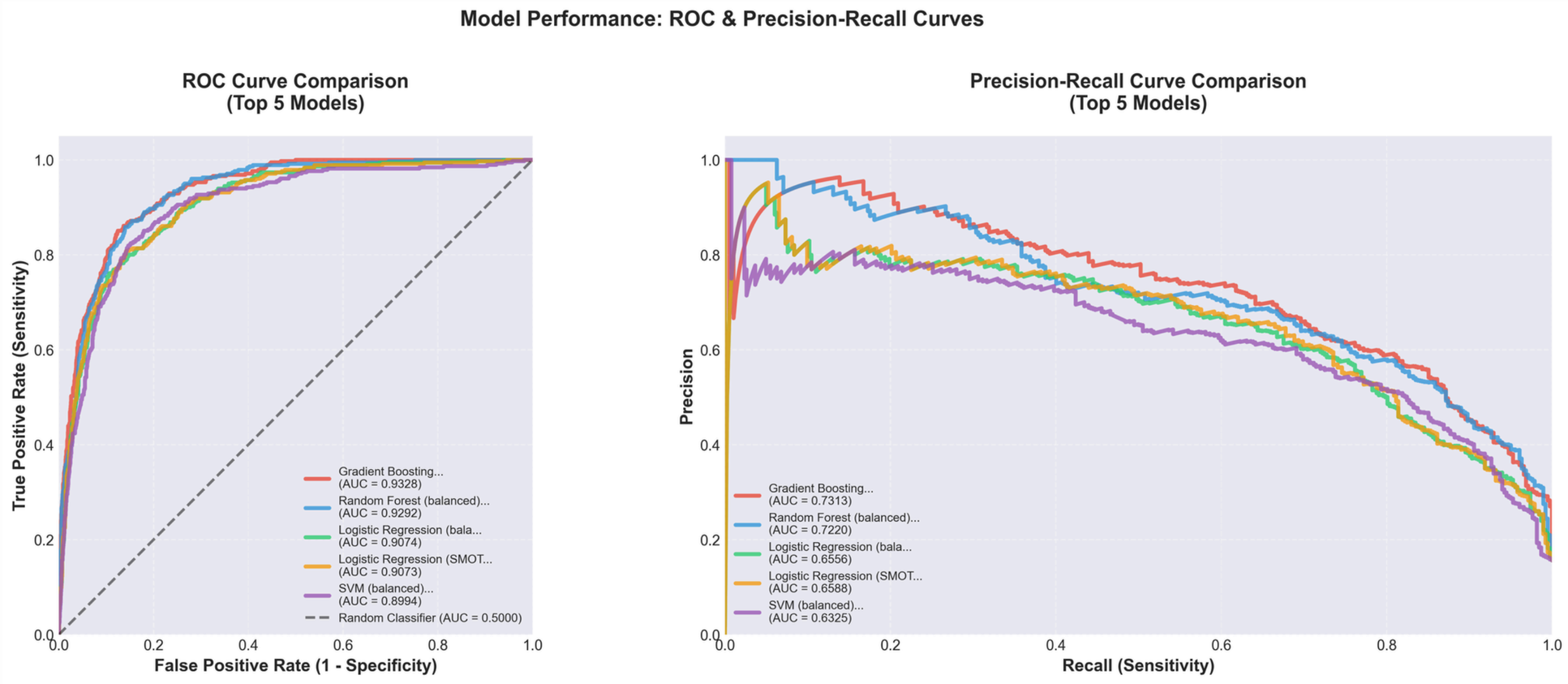
Recall improved from 0.58 to 0.87 after resolving class imbalance

Gradient Boosting

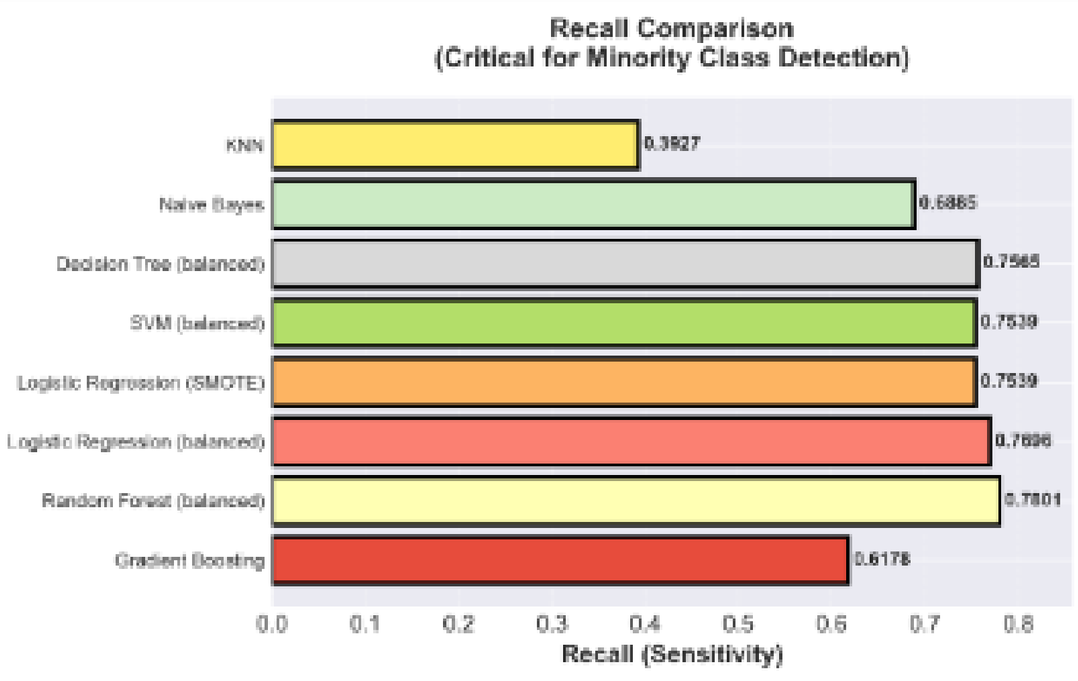
The best overall model was gradient boosting, with ROC-AUC of 0.9328 and an F1-Score of 66.8%. The model highly accurate overall and excels at identifying non-purchasers. Some purchasers are missed (moderate recall) but this is the best relative to other models.

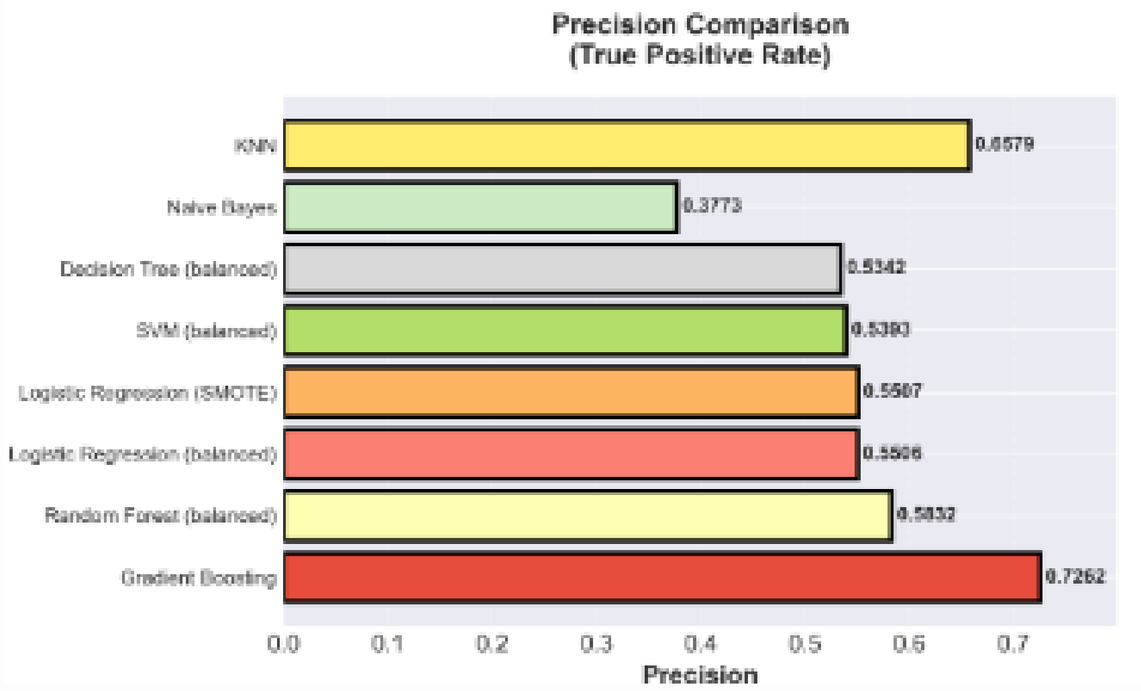


Model Comparison



Model Comparison





Different models excel at minimizing certain types of errors when predicting purchasers vs. non-purchasers

Q2 Conclusion

Answer

- We can answer the question of will users make a purchase with high certainty as proven by metrics from various classification models.
- Adjusting for class imbalances, tuning hyperparameters, and experimenting with different techniques helps yield a more confident answer.

Why this matters

- Helps predict purchasing patterns based on relevant data about websites, temporal patterns, etc. to manage inventory
- Leveraging strengths of models allows for adjustments based on costs of false positive/false negatives

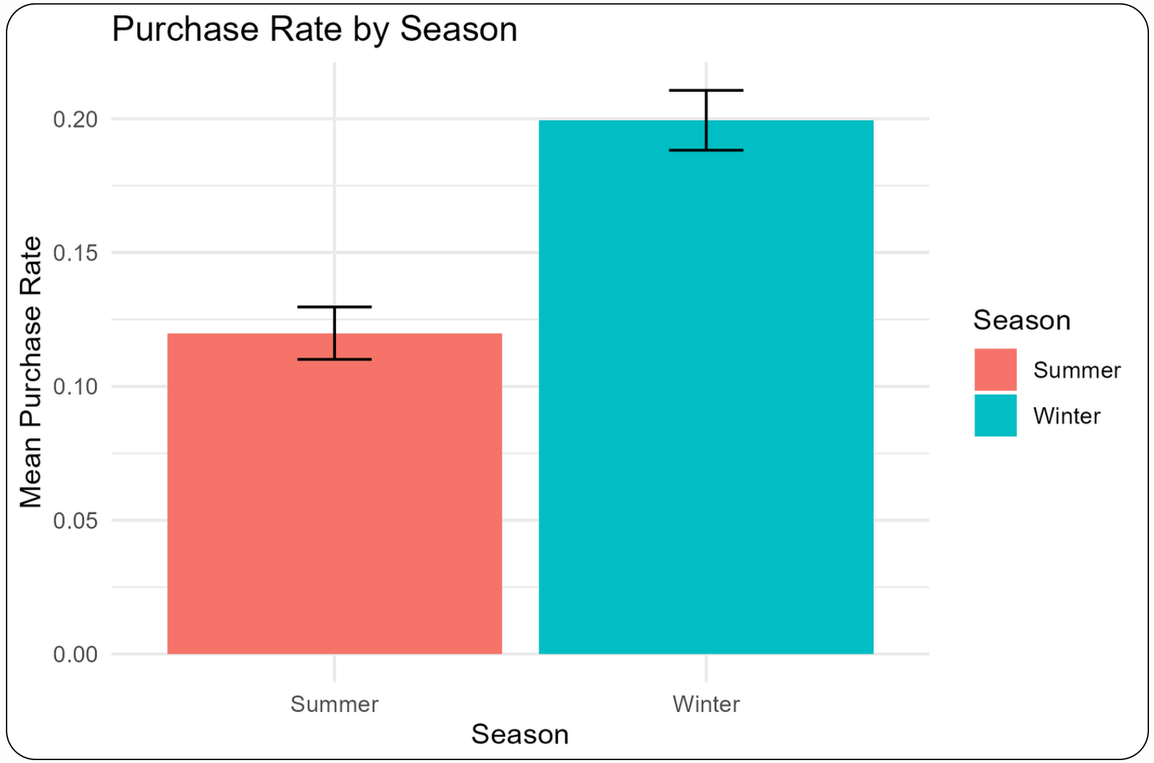
Q3. How Do Seasonal and Behavioral Contexts Influence Purchasing Intention?

3.1. Seasonal purchase gap

3.2. Visitor-type differences

3.3. Behavioral clustering

Seasonal Purchase Rate



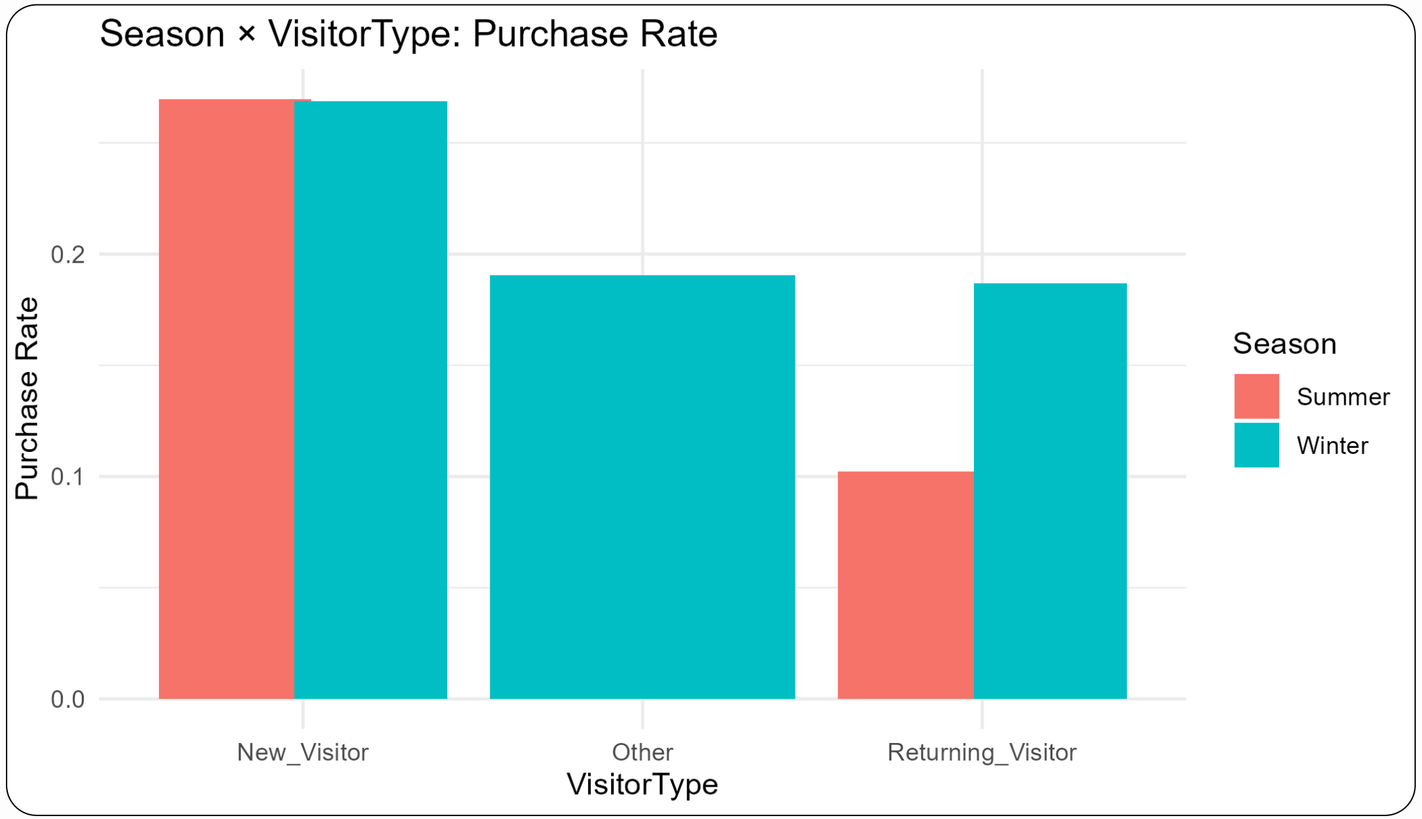
Winter

- Higher purchase intent during winter months
- Seasonal events naturally drive stronger motivation

Seasonal Patterns

- Shopping cycles elevate user engagement
- Winter browsing becomes more deliberate and goal-oriented

Visitor-type differences



Returning Visitors

- Winter significantly boosts conversions.
- Loyal users show stronger seasonal sensitivity.

New & Other Visitors

- New visitors convert consistently across seasons.
- “Other” visitors show moderate winter uplift.

Behavioral Clustering

Shoppers segment

Cluster	Size (n)	Purchase Rate
1	842	33.4%
2	847	0.7%
3	9,560	16.3%
4	956	6.4%

1: High-value buyers
2: low-intent bouncers
3: general browsers
4: early seasonal explorers

Q3 Conclusion

Answer

- Winter consistently strengthens purchasing intention across all user groups, especially Returning Visitors.

- Product-engagement signals strengthen in Winter, and clustering reveals distinct shopper types—from high-value buyers to low-intent bouncers—showing clear behavioral differences.

Why this matters

- Enables season-aware targeting strategies
- Boosts marketing efficiency for high-value and returning users
- Helps personalize user experience based on behavioral segments
- Supports strategic planning for peak-season demand and promotions

Thank You

Presented by Jubin Joung, Satvik Mahendra, Wonjun Ryhu