30 August 2025

# Interpreting Customers' Purchase on Online Shopping Website

## Final Project Presentation

MMDT MLAI101

Instructor : Dr.Myo Thida

Mentor: Ma Nuwai Thet

Team members: Ma May Mon Thant

Ko Myint Myat Aung Zaw

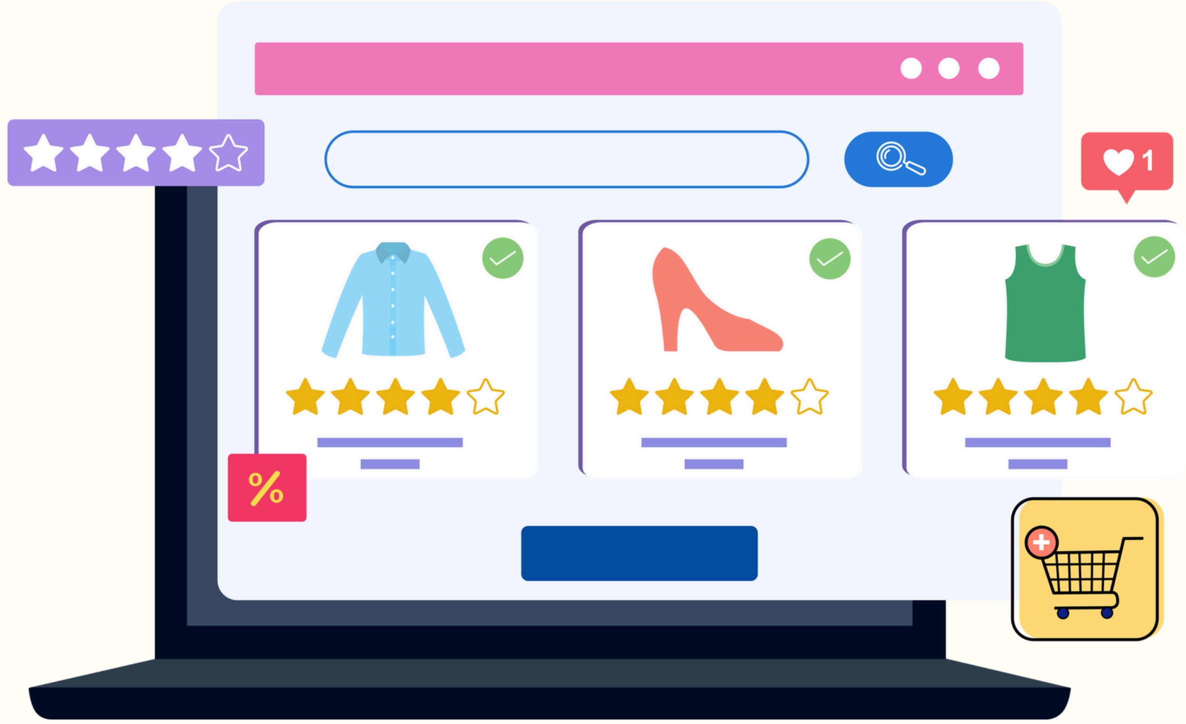
Ma Nilar Win

# Contents


- Introduction to the Project
- Dataset Description
- Data Preprocessing
- Model Selection
- Implementation Process
- Evaluation Metrics
- Results and Findings
- Discussion of Results
- Conclusion and Future Work


# Project overview

The project focuses on binary classification to predict whether an online shopper will make a purchase or not during their visit to an e-commerce website.




# Scope of the Project

- Identify our target audience with a high degree of confidence.
- Focus marketing efforts on potential buyers.
- Increase the conversion rate of our marketing campaigns.

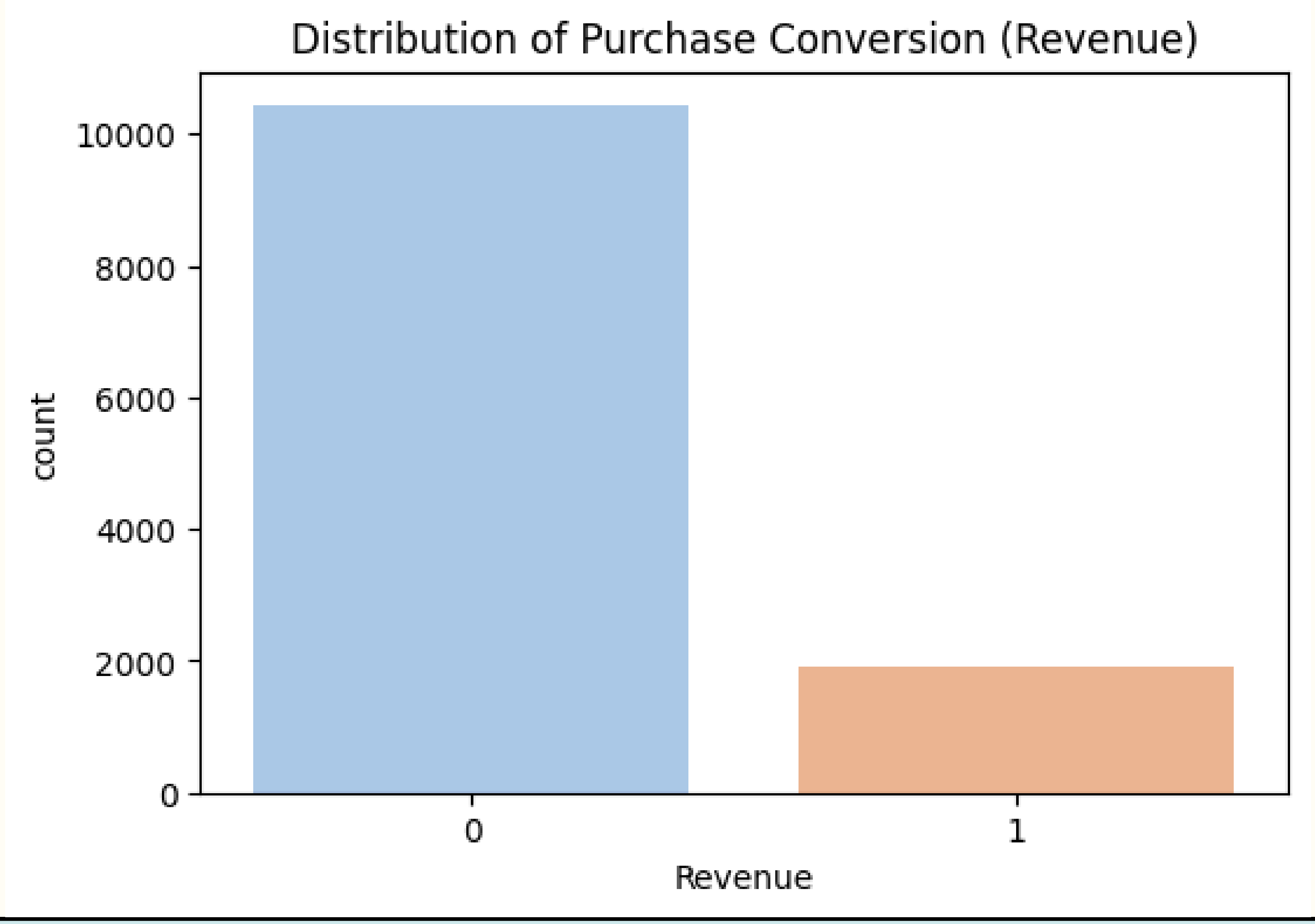
# DATASET DESCRIPTION

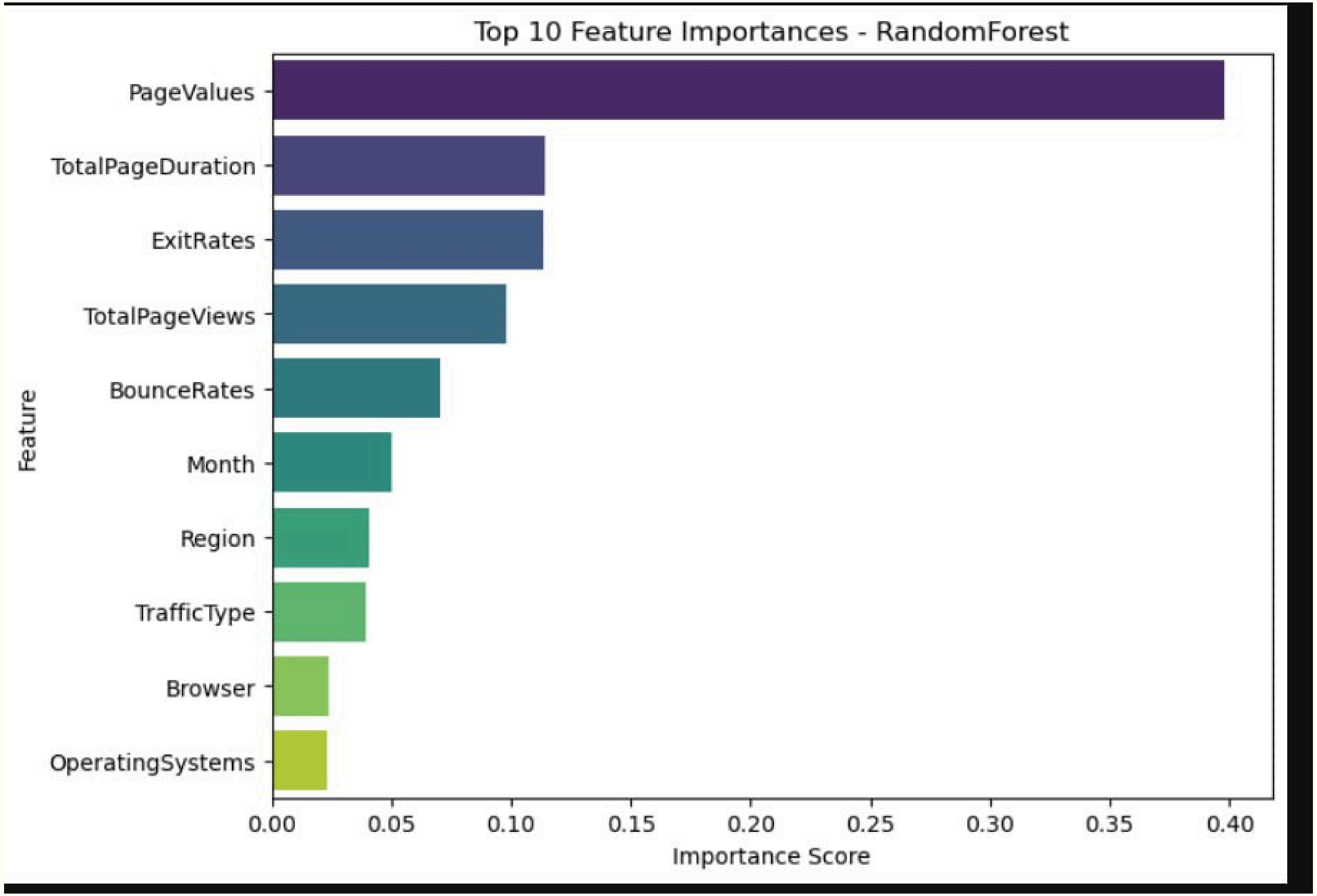
- Source: UCI Machine Learning Repository
- Size: 12,330 shopping sessions from different users over 1 year
- Features: 18 columns (10 numerical, 8 categorical)
- Important Features: PageValues, ExitRates, TotalPageDuration, TotalPageViews, Month, BounceRates
- Target Variable:  Revenue (Purchase, No Purchase)

# Data Preprocessing

- Load the data from .csv file
- Check the null value and data type of features
- Encoded the target variable and Weekend feature
- Encoded VisitorType feature using OneHotEncoder
- Mapped Month feature
- Combined number of pages views as TotalPageViews
- Combined Time spent on each pages in seconds as TotalPageDuration
- Cleaned the column names by removing underscores
- Split the dataset into 70% train and 30% test data using stratify=y due to imbalanced dataset
- Scaled the numerical features using StandardScaler

# EDA


Distribution of Purchase Conversion (Revenue)
count
10000
8000
6000
4000
2000
0
0
1
Revenue


Top 10 Feature Importances - RandomForest
PageValues
TotalPageDuration
ExitRates
TotalPageViews
BounceRates
Month
Region
TrafficType
Browser
OperatingSystems
Feature
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Importance Score

# BUSINESS OBJECTIVES

Optimize Marketing: By minimizing wasted spend on irrelevant customer outreach.

Maximize Revenue: By increasing the purchase conversion rate.

# Aligning Model Metrics with Our Strategy

- False Positives (The Problem): A model prediction of **"will buy"** for a customer who ultimately does not, leads to wasted marketing costs. These are customers we mistakenly target with promotions.

- Precision (The Solution): This metric tells us how many of our positive predictions were actually correct. A high precision score ensures our marketing efforts are focused on the right people, leading to a higher conversion rate for every dollar spent.

# Implementation Process

## Library Usage

- pandas, numpy, matplotlib
- scikit-learn
- time ( to compute model running time)
- accuracy_score, precision_score, recall_score, f1_score, roc_auc_score, confusion_matrix (evaluation metrics)

# Implementation Process

## Data Implementation Process

- Cleaned dataset (removed missing values, standardized column names)
- Encoded categorical variables (Weekend , VisitorType, etc.)
- Split dataset into training (70%) and testing (30%)
- Performed feature scaling (StandardScaler for numeric features)
- Trained multiple models: Logistic Regression, Random Forest, Naive Bayes, SVM
- Evaluated using Accuracy, Precision, Recall, F1-score, ROC-AUC

Data Loading
↓
Data Preprocessing
↓
Feature Engineering
↓
Train/Test Split
↓
Model Training
↓
Model Evaluation
↓
Prediction

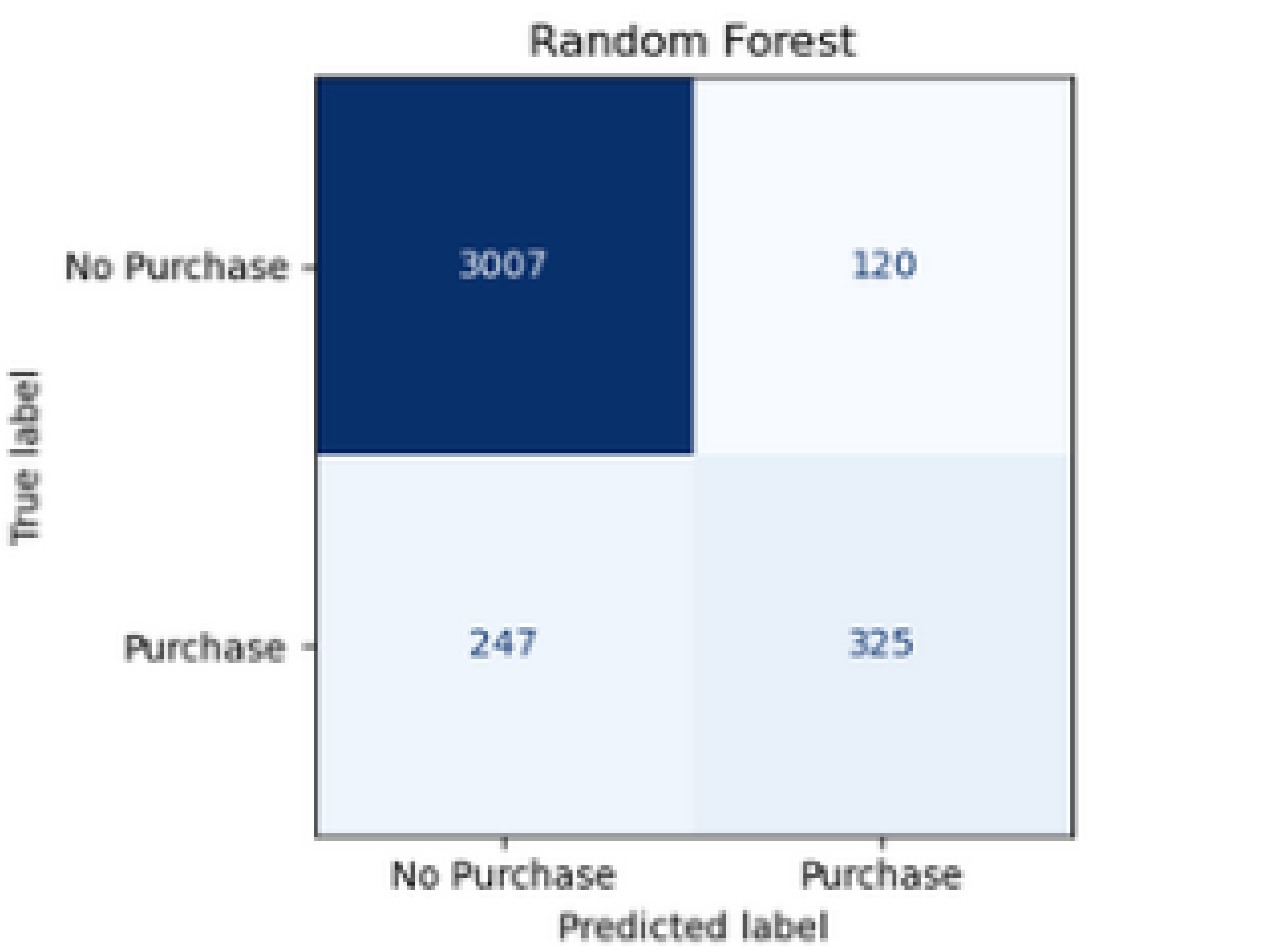
# Model Selection

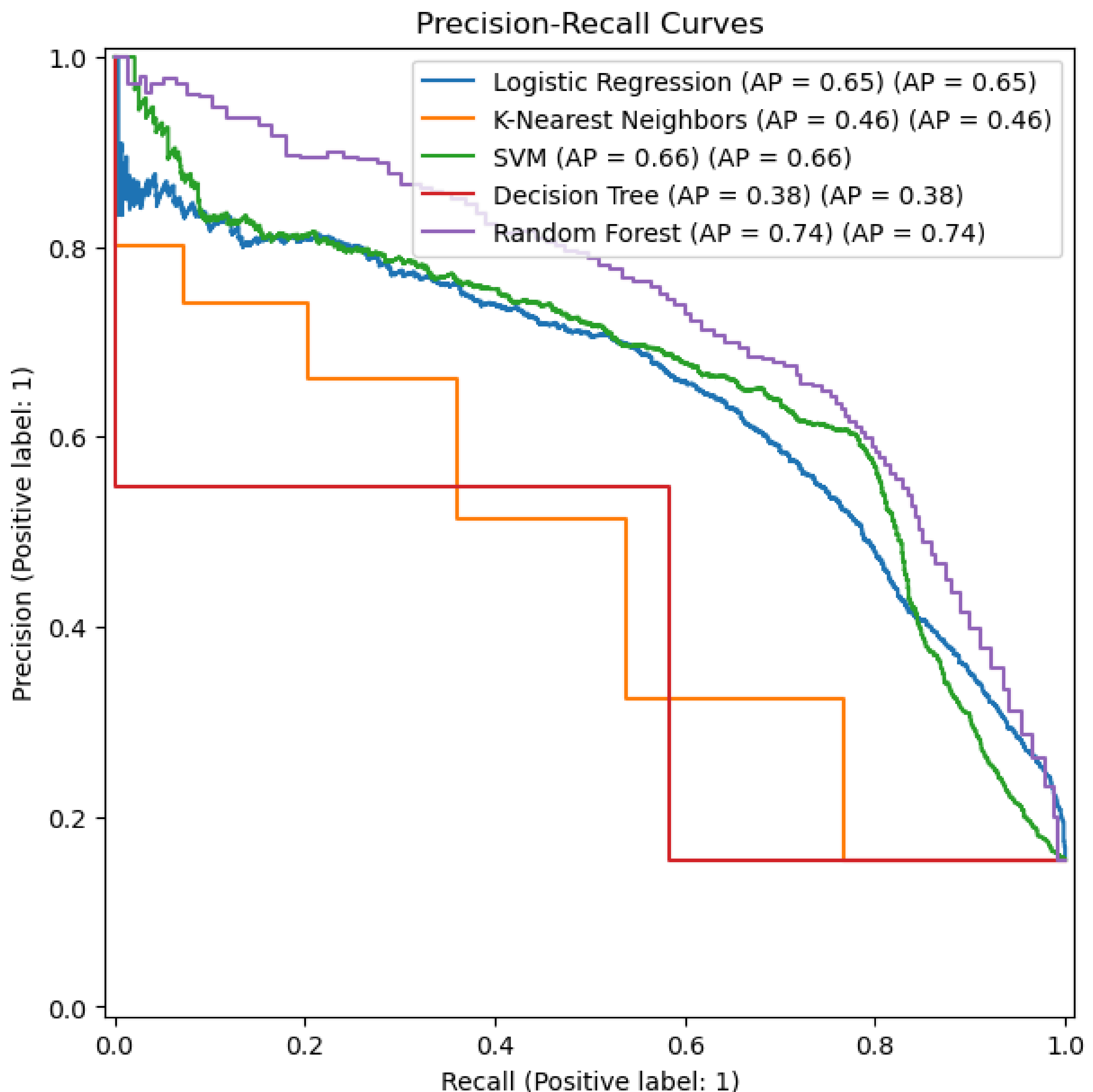
- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest
- K-Nearest Neighbors (KNN)

# Our Models

--- Final Model Performance Table Order By F1-Score ---

| Model | Precision | Recall | F1 Score | ROC AUC | Accuracy |
|---|---|---|---|---|---|
| Random Forest | 0.73 | 0.57 | 0.64 | 0.91 | 0.90 |
| Decision Tree | 0.51 | 0.53 | 0.52 | 0.72 | 0.85 |
| SVM | 0.72 | 0.39 | 0.50 | 0.88 | 0.88 |
| K-Nearest Neighbors | 0.71 | 0.38 | 0.50 | 0.79 | 0.88 |
| Logistic Regression | 0.71 | 0.34 | 0.46 | 0.89 | 0.88 |


Random Forest
No Purchase
Purchase
3007
120
247
325
True label
No Purchase
Purchase
Predicted label


Precision-Recall Curves
Logistic Regression (AP = 0.65) (AP = 0.65)
K-Nearest Neighbors (AP = 0.46) (AP = 0.46)
SVM (AP = 0.66) (AP = 0.66)
Decision Tree (AP = 0.38) (AP = 0.38)
Random Forest (AP = 0.74) (AP = 0.74)
Precision (Positive label: 1)
Recall (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0

# CONFUSION MATRIX

Type I and Type II errors for our problem:

- A type I (false positive) error would be predicting that a customer will make a purchase, when they in fact do not.
- A type II (false negative) error would be predicting that a customer will not make a purchase, when in fact they do.

# METRICS

| Random Forest Model | Test Data | | | Train Data | | |
|---|---|---|---|---|---|---|
| Performance | Precision | Recall | F1-score | Precision | Recall | F1-score |
| Original | 0.71 | 0.54 | 0.61 | - | - | - |
| SMOTE | 0.59 | 0.69 | 0.64 | - | - | - |
| K-Fold CV | 0.72 | 0.50 | 0.59 | 0.74 | 0.57 | 0.64 |
| SMOTE & K-Fold CV | 0.61 | 0.70 | 0.65 | 0.62 | 0.74 | 0.67 |
| Best Parameters | 0.72 | 0.58 | 0.64 | - | - | - |

# Our Model's Performance

- The model's Precision Score is 72%.
- This means that when our model predicts a customer will buy, it is correct 72% of the time.

```
--- Step 3: Final Model with Best Parameters Test Data ---
 🔹 Final Evaluation on Test Set 🔹
F1 Score: 0.6409
Confusion Matrix:
 [[2995  132]
 [ 240  332]]
Classification Report:
               precision    recall  f1-score   support

           0       0.93      0.96      0.94      3127
           1       0.72      0.58      0.64       572

    accuracy                           0.90      3699
   macro avg       0.82      0.77      0.79      3699
weighted avg       0.89      0.90      0.90      3699

ROC AUC Score: 0.9199
```

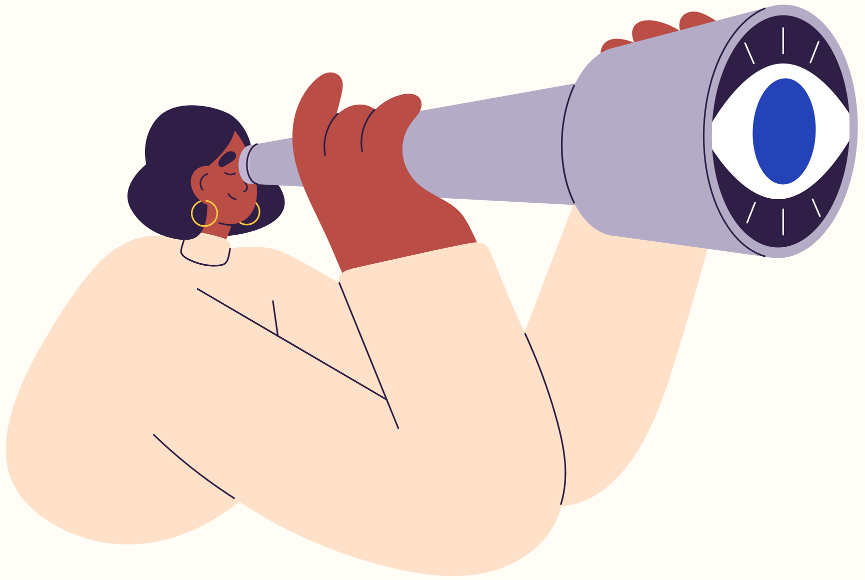
# CHALLENGES

**Highly Imbalanced Dataset:** The number of customers who make a purchase is very small compared to those who don't.

**Data Validation:** The data source for customer engagement may not be fully validated. We must be mindful of potential inaccuracies or biases in the data

# DISCUSSION

- The model might be very good at predicting purchases for frequent visitors but poor for first-time customers.
- Given our website's focus on high-volume, casual products (similar to Amazon), we prioritize maximizing the efficiency of our marketing campaigns to drive mass sales.



# FUTURE WORK

Acquire More Data in Future

Improve our data foundation

Implement and test for business implementation

# THANK YOU!