Project Report

Predicting Online Shoppers' Purchase Intention

Vaishnavi Vishwas Mane
Campus Id - IT75930
vmane1@umbc.edu
MPS Data Science

Business Problem:

The business problem that this project aims to solve is to predict whether an online shopper will make a purchase or not. This is a common challenge faced by e-commerce websites, as they want to understand their customers' behavior and optimize their website's design and marketing strategies to increase the likelihood of conversion. By using machine learning to predict which sessions are likely to end with shopping, e-commerce websites can personalize their recommendations and marketing efforts, ultimately leading to increased sales and reven. The dataset used for this project is the "Online Shoppers Purchasing Intention Dataset" available at

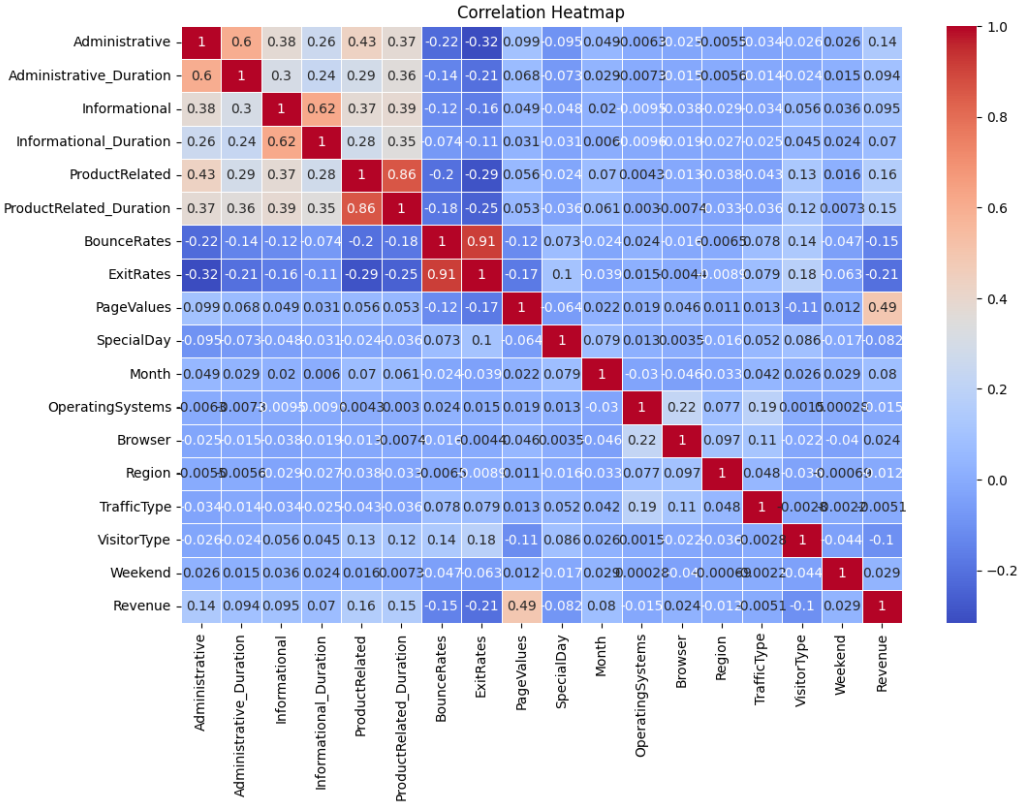
https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset.

Project Approach:

The Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology will be used for this project. The CRISP-DM methodology is a widely used data mining process that consists of six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

1. **Data Preprocessing and Feature Selection:**The datasets categorical variables were converted to integers using scikit-learn's LabelEncoder. The 'Month', 'VisitorType', 'Weekend', and 'Revenue' fields were encoded to allow for model training and

evaluation. The dataset was then divided using a 70:30 ratio into training and testing sets. For standardizing the features and to bring them on a similar scale, I used the StandardScaler from scikit-learn.For reducing the dimensionality of the dataset, I used Principal Component Analysis (PCA) using the PCA function from scikit-learn.

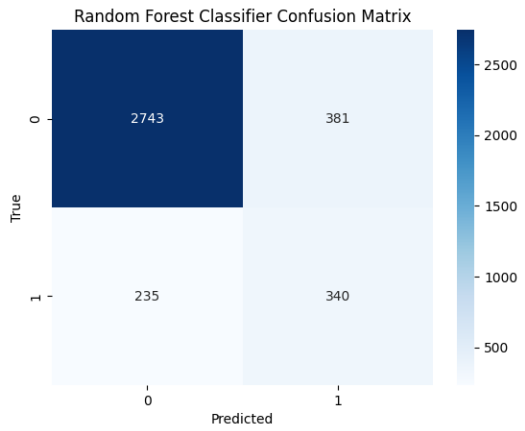


2. **Modeling and Evaluation:** Several classification models were implemented to predict the revenue generation potential of online shoppers.Each model was developed using training data and then tested using evaluation data. To evaluate each model's performance, the accuracy score, confusion matrix, and classification report were produced. The following models were utilized: Decision Tree Classifier, Random Forest Classifier, XGBoost Classifier, Logistic Regression, K-Nearest Neighbors (KNN) Classifier,Support Vector Machine (SVM) Classifier,Naive Bayes Classifier, Neural Network Classifier, Gradient Boosting Classifier.

For evaluating the performance of the classifiers, I used metrics such as accuracy, precision, recall, and F1-score. To check how well the classifiers can accurately classify instances of customers who generate revenue and those who do not. The classification_report function from scikit-learn generated a detailed report that includes these metrics.

Conclusion:

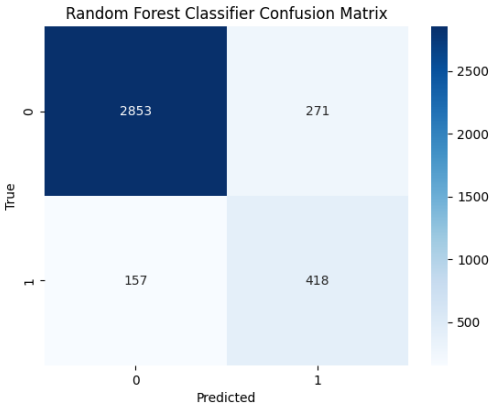
By predicting the likelihood of a shopper making a purchase, online retailers can improve their website design, marketing strategies, and customer service, leading to increased sales and revenue. This project will help online retailers make data-driven decisions and optimize their online shopping experience.



With the use of the classification report, confusion matrix, and accuracy score, each model's performance on the testing data was assessed. Based on the results, the Random Forest Classifier outperformed all other tested models with an accuracy of 88.43%. It fared better than other models including the Gradient Boosting Classifier, Decision Tree Classifier, XGBoost Classifier, Logistic Regression, KNN Classifier, SVM Classifier, Naive Bayes Classifier, and Neural Network Classifier.

The dataset utilized for training and evaluation was unbalanced, with more instances belonging to the negative class (shoppers who made no purchase) than the positive class (shoppers who made a purchase). I tried to balance the dataset using SMOTE from the imbalance-learn library.

The observed increase in accuracy after removing PCA and Standard Scaler can be due to two key considerations. To begin, the option to keep all attributes without dimensionality reduction (PCA) implies that each characteristic in the dataset significantly contributes to predicting the output variable. Second, it's possible that the dataset has intrinsic scaling qualities or has already been scaled adequately. This feature allows the models to operate well without the need for additional scaling using Standard Scaler.



References:

- Online Shoppers Behavior Prediction. (n.d.). Online Shoppers Behavior Prediction | Kaggle. https://www.kaggle.com/code/annettecatherinepaul/online-shoppers-behavior-prediction
- UCI Machine Learning Repository: Data Set. (n.d.). UCI Machine Learning Repository: Data Set. https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset.
- Baati, K., & Mohsil, M. (2020, May 6). Real-Time Prediction of Online Shoppers' Purchasing Intention Using Random Forest. PubMed Central (PMC). https://doi.org/10.1007/978-3-030-49161-1_4
- *Precision, Recall, and F1 Score: A Practical Guide Using Scikit-Learn*. (2022, November 8). Proclus Academy. https://proclusacademy.com/blog/practical/precision-recall-f1-score-sklearn/
- Zach. (2022). How to Interpret the Classification Report in sklearn (With Example). *Statology*. https://www.statology.org/sklearn-classification-report/