Title: Online Shoppers Purchasing Intention Dataset

Student name: Saad Baig

Enrolment no: 00919011922

Signature
Email Id: saadaventador@gmail.com
Contact no: 9873062020

TABLE OF CONTENTS

1. Introduction

This report presents the evaluation of various classification models on a dataset consisting of feature vectors belonging to 12,330 sessions. The dataset was carefully formed to ensure that each session belongs to a different user within a 1-year period, avoiding any bias towards specific campaigns, special days, user profiles, or periods. The objective is to predict whether a session will result in a revenue-generating transaction or not.

2. Data Loading and Preprocessing

The dataset is loaded using the Pandas library in Python. The features are stored in a DataFrame called 'data', and the target variable is labelled as 'Revenue'. To prepare the data for modeling, categorical values in the features are converted into integers using the LabelEncoder from scikit-learn. Specifically, the 'Month', 'VisitorType', and 'Weekend' columns are encoded.

3. Data Splitting

The data is split into training and testing sets using the train_test_split function from scikit-learn. The testing set comprises 20% of the total data, while the remaining 80% is used for training the models. This split allows for model evaluation on unseen data.

4. Classification Models

Several classification models are evaluated on the dataset to determine their performance in predicting revenue generation. The following models are applied:

- Decision Tree
- Support Vector Machine (SVM)
- Gaussian Naive Bayes (GNB)
- k-Nearest Neighbors (k-NN)
- Perceptron

For each model, the fitting process involves training the model using the training data and then making predictions on the test data. The

accuracy of each model is computed using the accuracy_score metric, which compares the predicted labels with the actual labels from the test data.

5. Results and Discussion

The accuracy scores obtained for each model are as follows:

Decision Tree: Accuracy = 0.877

Support Vector Machine: Accuracy = 0.901

Gaussian Naive Bayes: Accuracy = 0.846

k-Nearest Neighbors (k=5): Accuracy = 0.884

Perceptron: Accuracy = 0.871

Based on the accuracy scores, the Support Vector Machine model achieves the highest accuracy of 0.901, indicating its effectiveness in predicting online shopper intention. It outperforms other models such as the Decision Tree, Gaussian Naive Bayes, k-Nearest Neighbors, and Perceptron.

6. Conclusion

In conclusion, this report presents the evaluation of classification models on a dataset containing 12,330 sessions to predict online shopper intention. The Support Vector Machine model achieved the highest accuracy among the evaluated models, with an accuracy score of 0.901. The findings suggest that the Support Vector Machine model can be a valuable tool for predicting revenue generation in online shopping scenarios. Further analysis and optimization can be performed to improve the model's performance or explore other algorithms for comparison.