**1. Introduction**

**Customer purchase prediction is a significant application of machine learning in marketing. It allows companies to identify potential buyers and optimize their sales strategies. This project focuses on building a decision tree classifier to predict whether a customer will purchase a product or service based on demographic and behavioral data. The dataset used is the Bank Marketing dataset from the UCI Machine Learning Repository, which contains information about customer interactions with a banking campaign.**

**2. Objective**

The primary objectives of this analysis are:

1. Build a decision tree classifier to predict customer purchase decisions based on available data.
2. Analyze the importance of different features in predicting customer behavior.
3. Evaluate the performance of the classifier and visualize key insights.

**3. Dataset Overview**

The dataset used in this analysis includes data collected from a direct marketing campaign of a Portuguese bank. It contains several features, including:

* **Age**: The age of the customer.
* **Job**: The type of job the customer has.
* **Marital Status**: Whether the customer is single, married, or divorced.
* **Education**: The education level of the customer.
* **Default**: Whether the customer has credit in default.
* **Housing Loan**: Whether the customer has a housing loan.
* **Loan**: Whether the customer has a personal loan.
* **Campaign Data**: Data about previous marketing campaigns and customer interactions.
* **Target Variable (y)**: Whether the customer subscribed to a term deposit (yes/no).

The dataset was split into a **training set** and a **test set** to evaluate the model's performance.

**4. Data Preprocessing**

**4.1 Handling Missing Data**

* The dataset was checked for missing values. Missing data was imputed or dropped as necessary to ensure data quality.

**4.2 Encoding Categorical Variables**

* Categorical variables like **job**, **marital status**, and **education** were encoded using **one-hot encoding** to convert them into numerical format for the decision tree classifier.

**4.3 Feature Scaling**

* Although decision trees are not affected by feature scaling, continuous features were examined to ensure their suitability for model training.

**4.4 Splitting the Dataset**

* The dataset was split into training (80%) and testing (20%) sets to evaluate model performance.

**5. Exploratory Data Analysis (EDA)**

**5.1 Distribution of Features**

* **Histograms** and **boxplots** were used to visualize the distribution of numeric features like **age**, **balance**, and **duration of previous contacts**.

**5.2 Correlation Matrix**

* A **correlation matrix** was plotted to identify any strong correlations between numeric variables. However, decision trees can handle uncorrelated data well.

**5.3 Purchase Patterns Based on Demographics**

* Visualizations were created to explore the relationship between demographic factors (age, education, job) and purchase behavior.
* **Bar charts** were used to examine how the likelihood of purchasing changes based on marital status and housing loan status.

**6. Model Training and Evaluation**

**6.1 Decision Tree Classifier**

* A **decision tree classifier** was built using the **scikit-learn** library. The **Gini impurity** criterion was used to split nodes in the tree.
* Hyperparameters such as **maximum depth** and **minimum samples per split** were tuned to optimize the performance of the model.

**6.2 Model Performance**

* The model was evaluated using the following metrics:
  + **Accuracy**: The overall correctness of the classifier.
  + **Precision**: The proportion of positive predictions that were correct.
  + **Recall**: The proportion of actual positives that were correctly identified.
  + **F1-score**: The harmonic mean of precision and recall.
* A **confusion matrix** was plotted to visualize the number of true positives, true negatives, false positives, and false negatives.

**6.3 Feature Importance**

* The importance of each feature in predicting customer purchases was calculated and plotted. **Age**, **job**, and **previous campaign data** emerged as important factors influencing purchase decisions.

**7. Model Visualization**

**7.1 Decision Tree Plot**

* The trained decision tree was visualized using **matplotlib** and **graphviz** to show how different features split the data to make predictions.

**7.2 Confusion Matrix**

* A **confusion matrix** heatmap was plotted to visualize the model's performance in predicting customer purchases.

**8. Conclusion**

This project successfully built a decision tree classifier to predict whether a customer will purchase a product based on their demographic and behavioral data. The model revealed several insights:

* **Age**, **job**, and **previous campaign data** were the most important predictors of purchase behavior.
* The model achieved good accuracy and balanced performance in classifying customer purchase behavior.

**Key Recommendations:**

1. **Targeted Marketing**: Use demographic features like age and job to better target marketing efforts towards potential buyers.
2. **Campaign Optimization**: Focus on improving customer interactions in the campaign, as past interactions play a critical role in determining purchase decisions.
3. **Further Model Improvement**: Explore ensemble methods like **Random Forest** to potentially improve prediction accuracy.

**9. Future Work**

Future improvements to this project could include:

* **Hyperparameter Tuning**: Further optimize the decision tree by using **Grid Search** or **Random Search** for hyperparameter tuning.
* **Use of Other Models**: Compare the performance of the decision tree with other classifiers such as **Random Forest** or **Gradient Boosting**.
* **Cross-Validation**: Implement **k-fold cross-validation** to ensure that the model is not overfitting to the training data.

**BOBBY NARANG**  
MCA Graduate, Data Science Intern  
Date: 20.09.2024