Predicting Customer Purchase using Decision Tree Classifier

Introduction

The objective of this project was to build a decision tree classifier to predict whether a customer will purchase a product or service based on their demographic and behavioral data. The dataset used for this analysis is the Bank Marketing dataset from the UCI Machine Learning Repository, which is widely used for customer behavior prediction tasks.

Dataset Description

The dataset contains 41,188 observations and 21 attributes. The target variable y indicates whether the client subscribed to a term deposit ('yes' or 'no'). The attributes include:

- **Demographic Information**:
  - age: Age of the customer.
  - job: Type of job.
  - marital: Marital status.
  - education: Education level.
- **Financial Information**:
  - default: Whether the customer has credit in default.
  - housing: Whether the customer has a housing loan.
  - loan: Whether the customer has a personal loan.
- **Communication Information**:
  - contact: Type of communication contact (cellular, telephone).
  - month: Last contact month of the year.
  - day_of_week: Last contact day of the week.
  - duration: Last contact duration in seconds.
- **Campaign Information**:
  - campaign: Number of contacts performed during this campaign.

    - pdays: Number of days since the client was last contacted from a previous campaign.
    - previous: Number of contacts performed before this campaign.
    - poutcome: Outcome of the previous marketing campaign.
- **Economic Information**:
    - emp.var.rate: Employment variation rate.
    - cons.price.idx: Consumer price index.
    - cons.conf.idx: Consumer confidence index.
    - euribor3m: Euribor 3-month rate.
    - nr.employed: Number of employees.

Data Preprocessing

1. **Data Cleaning**:
    - Checked for missing values and handled them by dropping rows with missing data.
2. **Encoding Categorical Variables**:
    - Converted categorical variables to numeric using one-hot encoding to make them suitable for model training.
3. **Splitting Data**:
    - Split the data into training and testing sets using an 80-20 split (80% training, 20% testing).

Model Building

A Decision Tree Classifier was used to build the prediction model:

1. **Model Training**:
    - The model was trained using the training data.
2. **Model Testing**:
    - Predictions were made on the test data to evaluate the model's performance.

Evaluation Metrics

- **Accuracy**:

- The model achieved an accuracy of 0.87, meaning it correctly predicted the subscription status for 87% of the test instances.

Confusion Matrix:

```
[[677  55]
[ 49  43]]
```

The confusion matrix shows:

- True Positives (TP): 677
- True Negatives (TN): 43
- False Positives (FP): 55
- False Negatives (FN): 49

Classification Report:

	precision	recall	f1-score	support
no	0.93	0.92	0.93	732
yes	0.44	0.47	0.45	92
accuracy		0.87		824
macro avg	0.69	0.70	0.69	824
weighted avg	0.88	0.87	0.88	824

Visualizations

1 . **Decision Tree Visualization**:

    - This plot depicts the structure of the decision tree, showcasing the decision-making process of the model.
2. **Confusion Matrix**:
    - A heatmap representing the confusion matrix to illustrate true positives, true negatives, false positives, and false negatives.
3. **Feature Importance**:
    - A bar plot highlighting the top 10 most influential features in predicting customer subscriptions.
4. **Distribution of Subscription**:
    - A count plot displaying the distribution of the target variable, indicating the number of customers who subscribed versus those who did not.
5. **Pairplot of Numerical Features**:
    - Visualizes relationships between different numerical features to identify correlations and distributions.
6. **Age Distribution by Subscription**:
    - A histogram showing the age distribution for subscribed and non-subscribed customers to identify any age-related trends.
7. **Boxplot of Age by Subscription**:
    - Compares age distributions for subscribed and non-subscribed customers to highlight significant differences.
8. **Correlation Heatmap**:
    - A heatmap illustrating correlations between numerical features to identify strong relationships.
9. **Violin Plot of Categorical Variables**:
    - Violin plots for categorical variables by subscription status, showing distribution within each category.
10. **Feature Distribution**:
    - Histograms of all features to visualize their distribution across the dataset.
11. **Bar Plot of Categorical Features**:
    - Bar plots with count annotations for categorical features to display the frequency of each category.

12. **Boxplots of Numerical Features**:
    - Boxplots to visualize the spread and potential outliers of numerical features.

Results

The decision tree classifier achieved an accuracy of 0.87 in predicting whether customers would subscribe to a term deposit based on their demographic and behavioral data. The precision and recall metrics for both classes ('yes' and 'no') suggest that the decision tree model effectively captures patterns related to customer subscriptions.

Conclusion

The decision tree model provided a good balance between complexity and interpretability. The visualizations aided in understanding the data distribution and the significance of various features in predicting customer behavior. This analysis can be extended further by experimenting with other machine learning models and techniques to improve prediction accuracy and model performance. Additionally, hyperparameter tuning and cross-validation can further enhance the model's robustness and generalizability.