Bank data:

Problem statement:

The dataset includes the following columns:

**age**: age of a person

**job**: type of job

**marital**: marital status

**education**

**default**: has credit in default? ('no','yes','unknown')

**balance**: average yearly balance

**housing:** has housing loan? ('no','yes','unknown')

**loan**: has personal loanduration ? ('no','yes','unknown')

**contact**: contact communication type ('cellular','telephone')

**day:** The day of the month when the contact was last made.

**month:** The month of the year when the last contact was made.

**duration:** The duration of the last contact in seconds.

**campaign:** The number of contacts performed during the campaign for the individual.

**pdays:** The number of days that passed after the last contact from a previous campaign. A value of -1 indicates that the individual was not previously contacted.

**previous:** The number of contacts performed before this campaign for the individual.

**poutcome:** The outcome of the previous marketting campaign (e.g. success, failure, others)

The aim of the project is to predict if the client will subscribe (yes/no) to a term deposit. This is a classification model, because the target variable is categorical.

**Data Analysis Journey**

Step 1 : after loading the data we do the initial inspection to understand the structure of the dataset

Step 2 : then we are checking the count of missing values

Step 3:#handling categorical data after segregating the categorical and continuous columns

We are listing out the categorical columns and applying encoding to the relevant string columns

Step 4 : feature engineering

We are modifying and creating new features as necessary. We are also combining distinct categories of similar type into a parent category in order to make the data concised

Step 5 : feature scaling we are standardizing the numerical features such that the model can capture the patterns accurately

Step 6 : dividing into train test split to split the data for building the model

Plot :

using seaborn's countplot() function to create a countplot for categorical variables

using seaborn's boxtplot() function to create a boxtplot for continuous variables

**Splitting the Dataset**

* Splitted the dataset into training and test sets to make predictions using the test data breaking dataset to train and test to 75% train data and 25% test data

**Applying Machine Learning Models**

* Applied Logistic Regression model, decision tree

**Comparing Model Accuracy and Tuning**

* Cross-Validation Without Hyperparameter Tuning
* A DecisionTreeClassifier is initialized and cross-validation is performed using StratifiedKFold to ensure balanced splits of the dataset.
* Cross-validation scores are printed along with their mean
* Cross-Validation With Hyperparameter Tuning:
* A parameter grid is defined for hyperparameter tuning.
* GridSearchCV is used to perform an exhaustive search over specified parameter values for the DecisionTreeClassifier using cross-validation.
* The best parameters and cross-validation score are printed.
* Cross-validation scores are printed for the best model along with their mean.
* Visualization:
* The decision tree is visualized for the best model obtained from hyperparameter tuning using plot\_tree.

**Conclusion**

* While applying decision tree we got overfitting scenario, so we refitted the decision tree model with customization of parameters
* Considering the accuracy scores of Logistic Regression (with cross-validation and hyperparameter tuning along with their default scores), we concluded that the model gave the best accuracy score: 0.8004

**Best Model Metrics : decision tree**

**Accuracy**: 0.8004

**Precision**: 0.805

**Recall**: 0.802

**F1 Score**: 0.80