

1. Load the dataset : We start by loading the dataset from a CSV file using Pandas “read_csv” function. This step is essential to access the data and begin our analysis.

2. Perform EDA (Exploratory Data Analysis) : EDA helps us understand the structure and characteristics of the dataset. We use descriptive statistics, count plots, pair plots, and correlation matrices to explore the data visually and identify patterns or relationships between variables.

3. Clean the data : Here we make sure the data is clean by deleting any repeated data and useless data such as non-used column then we fill missing values with the mean of the column for numeric data or with the mode function for non-numerical data.

4. Apply some pre-processing techniques : In this step, we apply pre-processing techniques such as one-hot encoding to convert categorical variables into numerical format (country to integer for example), which is required for modeling.

5. Do a train-test-split : We split the dataset into training and testing sets to train our models on a portion of the data and evaluate their performance on unseen data.

6. Train the dataset using various models : We train three classification models: Logistic Regression, Decision Tree, and Random Forest. These models will learn from the training data to make predictions.

7. Apply cross-validation and check accuracy of different models : Cross-validation helps estimate the performance of machine learning models and assess their generalization ability. We use cross-validation to compute the accuracy scores of the models.

8. Compare the results of all the different models : This code segment evaluates the performance of three trained classification models (Logistic Regression, Decision Tree, and Random Forest) on a separate test dataset. It predicts labels for the test data using each model and calculates the accuracy of these predictions. Finally, it prints the accuracy scores for each model, providing insights into their effectiveness in making correct predictions on unseen data.

9. Visualize the results with heatmap or some graphs : We compare the heatmaps of the confusion matrices and the classification reports. Then we visualize the performance metrics (precision, recall, and F1-score) of each model using bar plots. Additionally, we plot ROC curves and feature importances for Random Forest.

10. Conclusion : Based on the evaluation metrics and comparison of the models, we draw conclusions about their effectiveness in predicting customer churn. We quickly discuss the advantages and disadvantages of each model.