Introduction to Machine Learning - Tasks

Task 1: Data Summary and Initial Observations

Objective: Conduct an initial analysis to understand the dataset's structure and patterns.

Instructions:

Load sample_data.csv and print the first 5 rows.

Display summary statistics for each feature, focusing on identifying any unusual values or patterns.

Calculate and display the distribution of churn (the percentage of customers who churn).

List 3 observations about the dataset, such as surprising values or any imbalance in churn.

Expected Outcome: Code for basic EDA and a short description of initial insights.

Task 2: Advanced EDA with Visualizations

Objective: Visualize relationships between features to discover trends.

Instructions:

Create a histogram for annual_income to observe income distribution.

Plot a boxplot comparing spending_score across different category values to see if there's a spending trend in specific categories.

Plot a scatter plot of annual_income vs. spending_score, using color to distinguish churn status.

Write a brief comment for each visualization, describing any trends or outliers observed.

Expected Outcome: Three visualizations with written interpretations.

Task 3: Creating Custom Metrics for Customer Value

Objective: Create a metric that combines customer activity and spending to estimate customer value.

Instructions:

Calculate a new feature customer_value_score as: spending_score * membership_years.

Sort and display the top 10 highest-value customers based on this metric.

Discuss briefly how this metric could be useful for a business in targeting high-value customers.

Expected Outcome: Code that generates customer_value_score, a sorted list of high-value customers, and a business interpretation.

Task 4: Implementing a Baseline Model for Churn Prediction

Objective: Build a simple baseline model to predict churn.

Instructions:

Use spending_score and annual_income as features to train a Decision Tree model for predicting churn.

Split data into training and test sets (80/20), train the model, and output test accuracy.

Comment on the accuracy and explain if it seems reasonable as a baseline.

Expected Outcome: Code for training the Decision Tree and test accuracy, with a short accuracy evaluation.

Task 5: Comparing Models with Cross-Validation

Objective: Compare the performance of different models on the same task.

Instructions:

Train three models (e.g., Logistic Regression, Decision Tree, and K-Nearest Neighbors) using cross-validation on age, annual_income, spending_score, and membership_years to predict churn.

Use 5-fold cross-validation to evaluate each model.

Compare the mean cross-validation accuracy and suggest which model seems best for this data.

Expected Outcome: Code comparing models and a recommendation based on cross-validation scores.