**Sales Prediction with Random Forests**

**1. Introduction**

This document outlines the process of predicting sales using a Random Forest classifier. The goal is to classify sales into categories (Low, Moderate, High) based on various features.

**2. Objective**

**Problem Statement**: Predict sales performance based on features such as competitor price, income, advertising, population, and others.

**Goal**: Classify sales into categories (Low, Moderate, High) to understand patterns and inform strategic decisions.

**3. Solution Architecture**

**1. Data Collection**:

* **Dataset**: Company\_Data.csv
* **Features**:
  + Numerical: Sales, CompPrice, Income, Advertising, Population, Price, Age, Education
  + Categorical: ShelveLoc, Urban, US

**2. Data Preparation**:

* **Handling Missing Values**: Verified and confirmed no missing values.
* **Outlier Detection**: Used box plots to visualize outliers in numerical features.
* **Encoding**:
  + **Ordinal Encoding**: For ShelveLoc (categories: Bad, Medium, Good).
  + **One-Hot Encoding**: For categorical features (Urban, US).
  + **Label Encoding**: For target variable (Sales).

**3. Exploratory Data Analysis (EDA)**:

* **Distribution Analysis**:
  + KDE plots for CompPrice and Price.
  + Bar plots for sales by location, shelves location, and advertising impact.
* **Visual Insights**:
  + Analyzed sales distribution and its relationship with other features.

**4. Machine Learning Modeling**:

* **Model Used**: Random Forest Classifier
* **Parameters**:
  + Number of Trees: 500
  + Maximum Depth: 10
* **Performance Metrics**:
  + **Accuracy**: 73.33%
  + **Classification Report**: Precision, recall, and F1-scores for each category.
  + **Confusion Matrix**: Visual representation of model performance.

**4. Methodology**

**1. Data Exploration**:

* Load and preview the dataset to understand its structure and features.
* Perform preliminary data exploration to check data quality and distribution.

**2. Data Cleaning**:

* Address missing values (if any).
* Identify and handle outliers.

**3. Feature Engineering**:

* Convert categorical features into numerical format using encoding techniques.
* Prepare the target variable by classifying sales into categories.

**4. Model Building**:

* Split the dataset into training and testing sets.
* Train the Random Forest classifier with the training data.
* Evaluate the model using the testing data and compute performance metrics.

**5. Model Evaluation**:

* Assess model accuracy and performance using classification metrics and confusion matrix.
* Visualize results to interpret model performance and insights.

**6. Visualization**:

* Generate plots and charts to visualize data distribution and model performance.

**5. Execution Time**

* **Total Execution Time**: 5.31 seconds

**6. Conclusion**

**Summary**:

* Achieved a classification accuracy of 73.33%.
* Provided valuable insights into sales patterns and influential factors.

**Business Impact**:

* **Strategic Insights**: Informed marketing strategies and resource allocation.
* **Decision Making**: Improved understanding of customer segments and sales drivers.