Sales Forecasting Using Machine Learning

1. Project Overview

The goal of this project is to develop a predictive model that forecasts future sales based on historical data, customer demographics, and order details. By leveraging data science techniques, the project aims to identify patterns and trends that can help optimize inventory management, enhance marketing strategies, and improve overall business performance.

2. Problem Statement

The organization seeks to improve its ability to predict sales accurately, reduce inventory waste, and anticipate customer demand. The objective is to build a model that can forecast sales at various levels, including by customer, product category, and geographic region.

3. Data Description

The dataset used for this project contains detailed information about transactions, shipping, customer demographics, and product details. The key fields include:

- **Type**: Type of transaction made.
- **Days for shipping (real)**: Actual shipping days.
- **Days for shipment (scheduled)**: Scheduled delivery days.
- **Benefit per order**: Earnings per order.
- **Sales per customer**: Total sales per customer.
- **Late_delivery_risk**: Indicates if the delivery is late.
- **Category Name**: Product category description.
- **Customer Segment**: Type of customer (Consumer, Corporate, etc.).
- **Order Date**: Date of order placement.
- **Sales**: Total sales amount.

Data Source: Kaggle - DataCo Smart Supply Chain

4. Project Objectives

- Develop a machine learning model to forecast sales based on historical data.
- Analyze and visualize sales trends.
- Identify key factors influencing sales performance.
- Deploy the model to provide real-time predictions.

5. Data Preparation

Steps Involved:

1. **Data Collection**: Import and merge data from multiple sources.
2. **Data Cleaning**:
    - Handle missing values by imputing or removing rows.
    - Remove duplicates.
    - Correct erroneous data (e.g., negative sales values).
3. **Feature Engineering**:

    - Calculate shipping delays (`Days for shipping (real) - Days for shipment (scheduled)`).
    - Create profit margin features (`Profit/Sales`).
    - Aggregate data by customer, product category, and region.
4. **Data Exploration**:
    - Visualize sales trends over time.
    - Analyze correlations between shipping time, discounts, and sales.

6. Model Development

Modeling Approach:

1. **Algorithm Selection**:
    - Linear Regression
    - Random Forest Regressor
    - XGBoost Regressor
2. **Training and Validation**:
    - Split the data into training (80%) and testing (20%) sets.
    - Perform cross-validation.
3. **Hyperparameter Tuning**:
    - Use GridSearchCV to optimize model parameters.
4. **Evaluation Metrics**:
    - RMSE (Root Mean Square Error)
    - MAE (Mean Absolute Error)
    - R-squared

7. Results and Analysis

- Visualize actual vs. predicted sales.
- Identify the best-performing model based on evaluation metrics.
- Highlight key insights, such as seasonal trends and customer segments driving high sales.

8. Deployment

- Develop a web application using Flask or Streamlit.
- Deploy the model to provide real-time sales forecasts based on user inputs.
- Share the web app link for stakeholder access.

9. Project Deliverables

- Jupyter Notebook with all code and analysis.
- Web application for model deployment.
- Dataset (cleaned and raw versions).
- Video presentation explaining the project.
- GitHub repository with project files.

10. Conclusion

This project leverages data science to address a real-world business challenge, offering valuable insights into sales forecasting. The predictive model and associated web application provide stakeholders with tools to make data-driven decisions, optimize inventory, and

improve customer satisfaction.