Machine Learning Sales Prediction Project Report

Introduction

This project aims to predict sales based on various factors such as family, store number, date, holiday status, and oil prices. The dataset comprises historical sales data, holiday events, and oil prices. Machine learning techniques, particularly linear regression, are employed to develop a predictive model for sales.

Dataset and Tools

Dataset: Includes historical sales data, holiday events, and oil prices.

Tools: Python libraries including pandas, numpy, matplotlib, scikit-learn.

Data Preprocessing

Load and preprocess the training dataset, including handling missing values and converting categorical variables to numerical using label encoding.

Merge datasets based on the date column, incorporating holiday events and oil prices.

Interpolate missing values in the oil price column using linear interpolation.

Exploratory Data Analysis (EDA)

Visualize the correlation between oil prices and dates, family and sales, and dates and sales to understand relationships between variables.

Model Development

Prepare the data by assigning independent and dependent variables.

Split the data into training and testing sets.

Train a linear regression model on the training data to predict sales.

Model Evaluation

Evaluate the model's performance using mean squared error (MSE) on the test data.

Assess the accuracy of the predictions and the model's ability to generalize to unseen data.

Prediction

Preprocess the test dataset and convert it into a format suitable for prediction.

Use the trained linear regression model to predict sales for the test data.

Conclusion

This project demonstrates the application of machine learning techniques for sales prediction based on historical data and external factors such as holidays and oil prices. By preprocessing the data, training a linear regression model, and evaluating its performance, accurate predictions can be made to assist in business decision-making and planning.