Wiseanalytics

Data Science Internship Assignment: Sales Forecasting

Introduction

Welcome to the Data Science Internship Assignment. In this assignment, you will work with real-world retai

to develop a forecasting model that predicts future sales for thousands of product families across different s

This project will help you understand how external factors like promotions, holidays, economic conditions, a

impact sales, and how machine learning models can be used to improve demand forecasting.

This assignment is structured into two parts:

1. Data Processing and Feature Engineering (Day 1) - Cleaning, transforming, and exploring the dataset.
2. Model Selection, Forecasting, and Evaluation (Day 2) - Training different forecasting models, comparing
   and presenting insights.

Dataset Overview

The dataset consists of multiple files providing sales data and additional influencing factors:

- train.csv - Historical sales data.
- test.csv - The test set for which sales need to be predicted.
- stores.csv - Metadata about store locations and clusters.
- oil.csv - Daily oil prices (affecting Ecuador's economy).
- holidays_events.csv - Information about holidays and special events.

Your task is to forecast daily sales for each product family at each store for the next 15 days after the last t

Part 1: Data Processing and Feature Engineering (Day 1)

1. Data Cleaning

- Load the dataset using Pandas.
- Handle missing values in oil prices by filling gaps with interpolation.
- Convert date columns to proper datetime formats.
- Merge data from stores.csv, oil.csv, and holidays_events.csv into the main dataset.

2. Feature Engineering

- Time-based Features:
  - Extract day, week, month, year, and day of the week.
  - Identify seasonal trends (e.g., are sales higher in December?).
- Event-based Features:
  - Create binary flags for holidays, promotions, and economic events.
  - Identify if a day is a government payday (15th and last day of the month).
  - Consider earthquake impact (April 16, 2016) as a separate feature.
- Rolling Statistics:
  - Compute moving averages and rolling standard deviations for past sales.
  - Include lagged features (e.g., sales from the previous week, previous month).
- Store-Specific Aggregations:
  - Compute average sales per store type.
  - Identify top-selling product families per cluster.

3. Exploratory Data Analysis (EDA)

- Visualize sales trends over time.
- Analyze sales before and after holidays and promotions.
- Check correlations between oil prices and sales trends.

- Identify anomalies in the data.

4. Documentation

- Clearly document each preprocessing step in a Jupyter Notebook.

- Explain why each feature was created and how it helps in forecasting.

Part 2: Model Selection, Forecasting, and Evaluation (Day 2)

1. Model Training

Train at least five different time series forecasting models:

- Baseline Model (Naïve Forecasting) - Assume future sales = previous sales.

- ARIMA (AutoRegressive Integrated Moving Average) - A traditional time series model.

- Random Forest Regressor - Tree-based model to capture non-linear relationships.

- XGBoost or LightGBM - Gradient boosting models to improve accuracy.

- LSTM (Long Short-Term Memory Neural Network) - A deep learning-based forecasting model.

Bonus Challenge: If comfortable, implement a Prophet model for handling seasonality.

2. Model Evaluation

Compare models based on:

- Root Mean Squared Error (RMSE)

- Mean Absolute Percentage Error (MAPE)

- R-Squared Score

- Visual Inspection (Plot actual vs. predicted sales)

3. Visualization

- Plot historical sales and predicted sales.

- Compare model performances using error metrics.

- Visualize feature importance (for Random Forest/XGBoost).

4. Interpretation and Business Insights

- Summarize which model performed best and why.

- Discuss how external factors (holidays, oil prices, promotions) influenced predictions.

- Suggest business strategies to improve sales forecasting (e.g., inventory planning, targeted promotions).

Submission Guidelines

- Submit a GitHub repository or a Google Drive link containing:

  - A Jupyter Notebook with code and explanations.

  - A README.md file explaining how to run the scripts.

  - A final model comparison summary with key insights.

Why This Assignment?

- Real-world business impact: Sales forecasting is a critical business function in retail.

- Data engineering skills: Handling multiple data sources and creating meaningful features.

- Machine learning practice: Comparing traditional ARIMA models with advanced ML techniques like XGBo

- Business insights: Beyond model performance, understanding what drives sales.

This assignment tests your ability to process real-world data, build forecasting models, and provide busines