

DEMAND FORECASTING USING MACHINE LEARNING

Demand forecasting is the process of estimation of probable demand for a product or service in the future. It is used to predict what customer demand will be for a product or service, with varying levels of specificity. Accurate, timely forecasts are invaluable for both businesses and their customers.

The objective of this project is to predict daily sales of 50 distinctive items in 10 stores for a 2-month period (Nov 1st to Dec 31st, 2017) using Machine Learning.

Dataset:

Link: <https://www.kaggle.com/competitions/demand-forecasting-kernels-only/overview>

The dataset contains 5 years of sales data per 50 different items being sold in 10 different stores. It covers the period between 2013-01-01 and 2017-12-31. It contains features like

- date - Date of the sale data. There are no holiday effects or store closures.
- store - Store ID
- item - Item ID
- sales - Number of items sold at a particular store on a particular date.

The dataset comprises a substantial volume of information, totalling 913,000 entries.

Data Preparation:

- Checked for missing values and duplicates.
- Explored basic statistics and information about the dataset.
- Visualized the weekly and monthly average sales over time using line plots and a calendar heatmap.

Feature Engineering:

- Extracted various features from date column, such as day, month, year, day of the year, weekday, and indicators for weekends, month starts, and month ends.
- Lag features are created based on past sales values, capturing seasonality and trends.
- Rolling mean features are calculated to smooth out fluctuations.

Train - Test Split:

Split the dataset into training and testing sets, where the training set contains data up to Oct 31, 2017 and the testing set contains data from Nov 1st to Dec 31st, 2017.

Feature Scaling and Model Training:

- Categorical features (day, month etc.) are one-hot encoded to be suitable for the XGBoost model.
- An XGBoost regressor is chosen for its efficiency and accuracy in the time series forecasting.
- The model is trained on the training set, learning the relationships between features and sales.

Models Prediction and Evaluation:

- The trained model is used to make prediction on the testing set using the trained model.
- Mean Absolute Percentage Error (MAPE) is calculated to measure the model's performance.
- MAPE of 13.54% indicates a relatively good fit with average percentage difference between the predicted and actual sales values.
- Visualized the prediction for Item 1 sales for different stores using a line plot.