Product Demand Forecasting: An Analysis Using Visualization and Machine Learning Techniques

COMP4449 Capstone

Yixuan Huang

# Introduction

## Purpose

This project aims to develop machine learning models for predicting product demand using various input variables. Accurate demand forecasting is essential for efficient supply chain management and inventory control. It ensures that products are available in the right quantities when needed, benefiting organizations like manufacturers, retailers, and distributors. This analysis caters to the specific needs of clients in the supply chain, optimizing resource allocation and reducing costs through strategic planning. The paper covers dataset analysis and the implementation of machine learning models to achieve accurate predictions.

## Significant

The significance of this project lies in its potential to provide valuable insights through visualizations and the ability to forecast product demand. By understanding the relationship between input variables and order demands, businesses can optimize their operations and improve financial planning.

## Research Question

My research questions for this analysis are: What practical strategies/solutions can we offer our clients to address/improve their concerns/experiences from the past? Which ML model provides the most accurate and reliable product demand forecasts, mitigating the risks of overselling or under-selling in the future?

## Dataset

The dataset, sourced from Kaggle, comprises historical order records for a manufacturing company. It consists of 1048575 rows and 5 columns, including information such as product info, warehouse info, and dates. The target variable is the Order\_Demands, while the input variables include the Product\_Code, Warehouse, Product\_Category, and Date. They are all in object data type.

# Data Preprocessing

## Data Preparation

Data preparation involved handling missing values, data type conversion, cleaning the target variable, and removing orders with zero demand, as the focus of the project was on predicting order demands.

## Exploratory data analysis and visualization techniques

A graph of different colored bars

Description automatically generated A chart with numbers and a bar

Description automatically generated

Here are some visualizations between Order\_Demand and different variables. Whse\_J processes more orders overall, when we consider the size of the orders, Whse\_C and Whse\_S are handling mid-size orders on average while Whse\_J appears to handle small orders primarily. Whse\_A has low demands over the years.

A graph with different colored bars

Description automatically generated

Category\_019 takes the lead overall, with strong sales in both Whse\_C and Whse\_S. This plot allows organizations to draw up inventory allocation strategies to ensure that products are stocked closer to where they are most likely to be ordered, which reduces shipping costs. Prior to modeling, various preprocessing steps were undertaken, including encoding, making time series stationary, outlier removal, and data scope reduction from January 2012 to December 2016. The distribution of the target variable revealed right-skewness and varying scales, suggesting the need for scaling before modeling.

## Data Splitting

To manage the large dataset, it was sampled to 1% of the original data and then split into 20% training and 80% testing sets for training and evaluating the XGBoost regression models. For time series models, data was resampled monthly and split into training(Jan 01, 2012 – Jan 01, 2016) and testing sets(Jan 01, 2016 – Jan 01, 2017).

# Model Building and Evaluation

A graph showing a number of blue and red bars

Description automatically generated

A graph showing a number of blue and red bars

Description automatically generated

**XGBoost Regression Models**: XGBoost regression models were initially implemented to predict order demands. Hyperparameter tuning was performed using Optuna to optimize model performance. The evaluation was based on Mean Absolute Error (MAE) and Mean Squared Error (MSE) scores, revealing low errors but inadequate fitting on the testing set, indicating overfitting.A graph with blue and orange lines

Description automatically generated

A graph showing the different types of graphs

Description automatically generated with medium confidence

A graph showing a line

Description automatically generated

**Time Series Models**: ARIMA and SARIMA models were implemented, with SARIMA performing well on the testing data and achieving a better MAE score. However, there remained an issue with aligning the prediction line with the original scale. Would this mean that the model performs badly?

# Conclusion

In conclusion, want to say SARIMA outperformed other models but had challenges aligning predictions with the original scale, although it generally exhibited the correct patterns.

## Lessons Learned

Throughout the project, dealing with irregular data and deciding on the necessity of resampling was challenging, as we may lose some data and it may affect the model when it comes to the high-frequency orders in Whse\_J. The possibility of overfitting in XGBoost models due to irregular data was a concern. However, Exploring regularization techniques like Lasso or Ridge regression to prevent overfitting in the future may be worthwhile.

By leveraging the insights gained from this analysis and implementing the recommended improvements, organizations can benefit from more accurate order demand predictions, leading to improved decision-making, resource allocation, and overall business success.

Data Resource: https://www.kaggle.com/datasets/felixzhao/productdemandforecasting/data