Time series Demand Forecasting Challenge

1) Problem Statement

The challenge is to build a statistical forecasting model that predicts Industry demand (column E - Western European tractor industry >160 HP) for the next 12 months. It should also be invested whether exogenous factors (columns 'Wheat Price', 'Milk Price EU', 'CEMA Business Barometer Index (TR&HV)') have an impact on the forecast.

2) Approach

A thorough exploratory data analysis (EDA) revealed the following insights from the data:

- There's no clear increasing/decreasing trend in the demand (column E)
- There's a clear yearly cyclic/seasonal pattern in the demand
- Demand is higher for two different times in a year. This aligns with our common understanding of farming seasons, and we expect demand to be higher during planting and harvesting seasons.
- There's some correlation between the demand and exogenous variables

I used a multi-pronged approach to solve this challenge:

i. Univariate time series forecasting

The first question I asked myself was can the demand be predicted by just its past values? So, I implemented classical univariate approaches such as Exponential Smoothing and ARIMA. Since one of the objectives was to investigate the impact of exogenous variables on the demand, I also implemented SARIMAX. SARIMAX performs the best in this category.

ii. Multivariate time series forecasting

To further improve the performance of the model, I used VARMAX (Vector Auto Regressor Moving Average with Exogenous variables). This approach didn't work great since VARMAX doesn't support seasonality and in our use case, seasonality needs to be accounted for.

iii. Regression

Although regression is not a popular approach for time series data, I chose it because it can provide statistical significance for different features. I created lagged variables and fit a linear regression model. It didn't perform great since high multicollinearity was an issue. I used the variable selection method Lasso regression to counter that. Lasso didn't perform great either. I validated the assumptions of the linear model, and they didn't hold true. In such cases, predictions can't be trusted and hence, I moved on to fb prophet.

iv. Facebook Prophet library

Facebook Prophet is a very powerful library that accounts for trends, and seasonality and supports multivariate data. This approach works the best out of all the approaches I have tried.

A couple of additional points that I took care of:

- Used the same split for training and testing sets across approaches to ensure a fair comparison
- To get the best out of the limited data, used walk-forward cross-validation

3) Results

To evaluate the model performance, I chose the following two metrics:

i. RMSE (Root Mean Squared Error)
ii. MAPE (Mean Absolute Percentage Error)

I chose them because RMSE can give us the error on the original scale and MAPE can tell us the same on the % scale. Together they paint a complete picture of the error.

Results are summarized in the following table:

Method	RMSE	MAPE
Simple exponential smoothing	730.11	23.95%
Holt's method	727.95	23.57%
Holt Winter's method	370.52	11.72%
ARIMA	463.10	12.63%
SARIMAX	330.58	9.80%
VARMAX	719.49	25.75%
Linear Regression	485.09	13.08%
Lasso Regression	484.82	13.11%
FB Prophet	231.44	8.16%

4) Challenges

The biggest challenge for me was to identify which features to keep in the model. I tried different approaches such as univariate vs multivariate time series forecasting and Lasso Regression, but I strongly believe that by carefully reviewing them with a domain expert and engineering new features, I can further improve the model performance.

5) Conclusion

To solve this challenge, I explored numerous avenues. The open-source library Facebook Prophet performed the best in predicting the demand, but I believe that the model performance can further be improved by tuning hyperparameters and including exogenous variables. This library currently doesn't support exogenous variables. SARIMAX revealed that exogenous variables Wheat Price and Milk Price EU are statistically significant in predicting the demand.