Final Reporting

Executive Summary

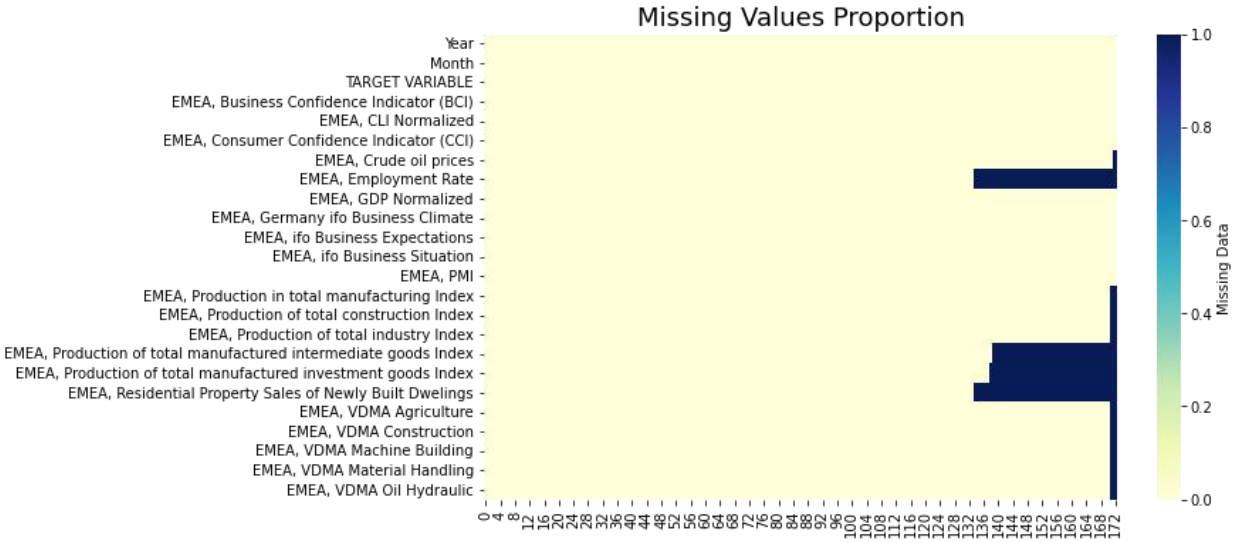
Danfoss is a Danish multinational company, based in Denmark, with more than 40,043 employees globally. It specializes in engineering solutions that help the globe use resources more wisely. The company in the past years has seen changes in its operations. A gap in their forecasting accuracy emerged as one of the major business challenges since the inception of COVID. In the effort to improve customer experience, more precise sales and inventory projections need to be made. Danfoss currently uses a simple linear regression with very little amount of macroeconomic data to forecast sales. Our proposed model looks to implement feature selection and regression models to address Danfoss's forecasting inaccuracy which will go a long way to help their Sales and inventory team properly plan inventory and enhance the customer experience. For any firm, exploratory data analysis is crucial. It allows us to examine the data before making any assumptions. We performed an EDA which guarantees that the findings are reliable and pertinent to the objectives and consequences of business. This report contains the results of our EDA carried out on the dataset and details of our proposed model to Danfoss.

Business Problem

A key business challenge that developed since the beginning of COVID is a gap in Danfoss's forecasting accuracy. Their inability to make accurate sales and inventory forecasts has influenced their customer experience. The use of a simple linear regression, a statistical technique that enables analysis and exploration of connections between continuous (quantitative) variables. proved inefficient for Danfoss.

Exploratory Data Analysis

To maximize our insight into a data set and into the underlying structure of the data set, we performed an EDA. The dataset was prepared for analysis using Python. 137 rows and 24 columns made up the dataset's total number of rows. These columns were made up of only numerical and integer variables. A total of 136 N/A values were found as depicted in the plot below. We chose to replace the N/A values with the respective columns' median as the histogram shows that most of them are either left-skewed or right-skewed. Also, there were no duplicate values observed.



Boxplot

There are a few outliers in many of the features taken into consideration, as can be seen from the box plot shown below. We have removed the outliers in extreme cases where they were influencing the best-fit-line. In other cases, we have limited them to 10th and 90th percentile.

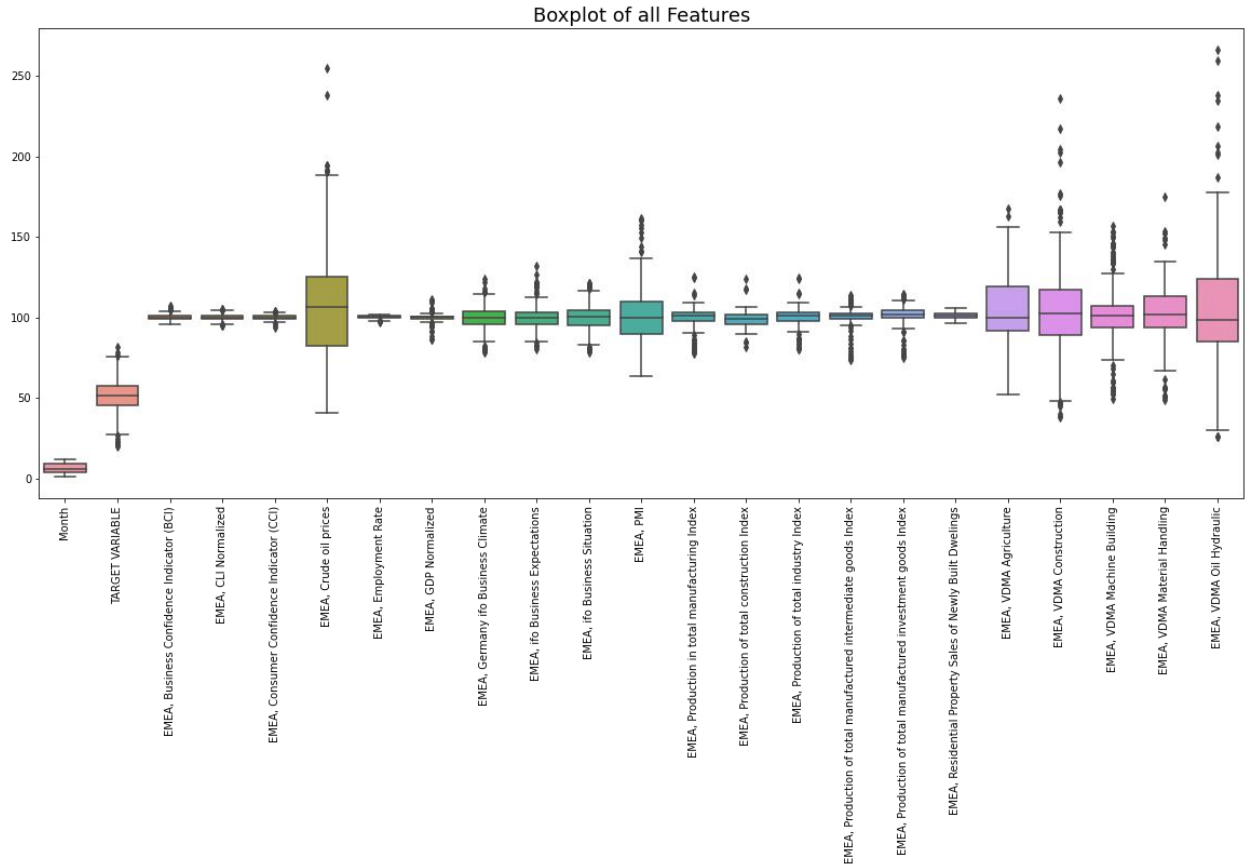


Figure: Finding Outliers

Correlation Plot

To comprehend how the variables in the dataset are related to one another, a correlation plot is considered. The total industrial production and manufacturing indexes were positively correlated (0.99) whereas VDMA agriculture and consumer confidence indicators showed a negative correlation (-0.16).

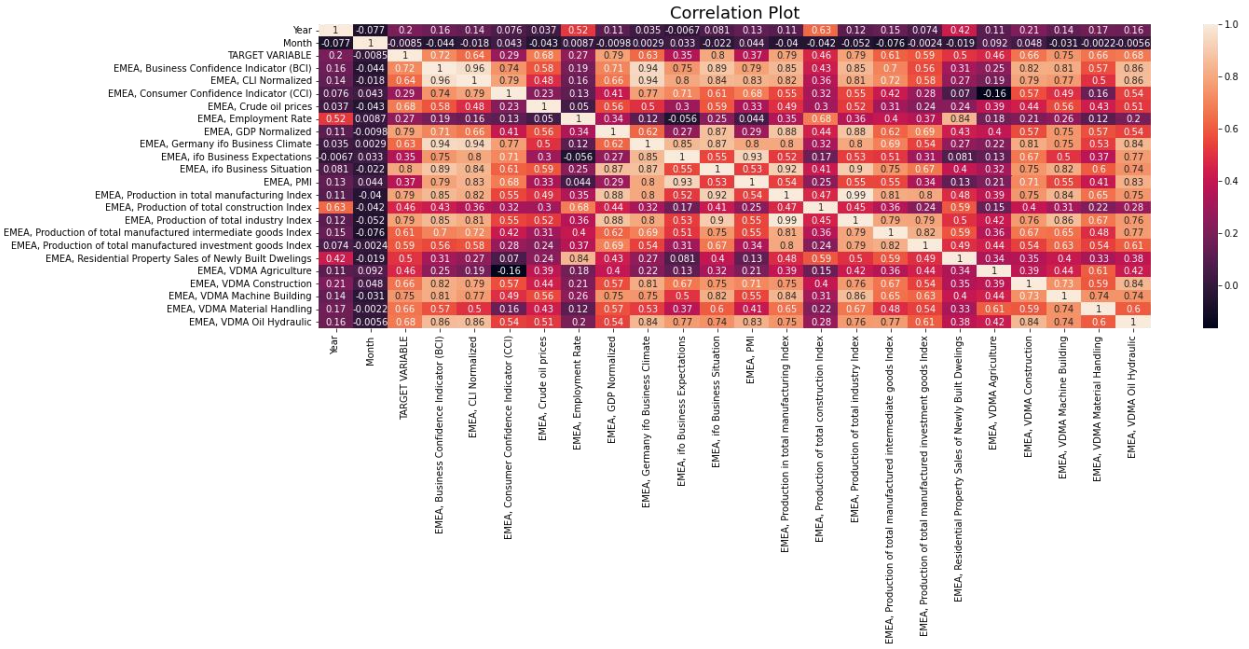


Figure: Correlation among parameters

Features with the highest correlation to the target feature are,

1. EMEA, Business Confidence Indicator (BCI) 0.7157223579365043
2. EMEA, GDP Normalized 0.7851050613959051
3. EMEA, ifo Business Situation 0.8009952209001024
4. EMEA, Production in total manufacturing Index 0.7900927257031622
5. EMEA, Production of total industry Index 0.7856863076869299

Line graph

Recognizable key performance indicators for Danfoss: total manufacturing index, total construction index, total industry index, total manufactured intermediate goods index and total manufactured investment goods index was analysed below is a line graph displaying the results.

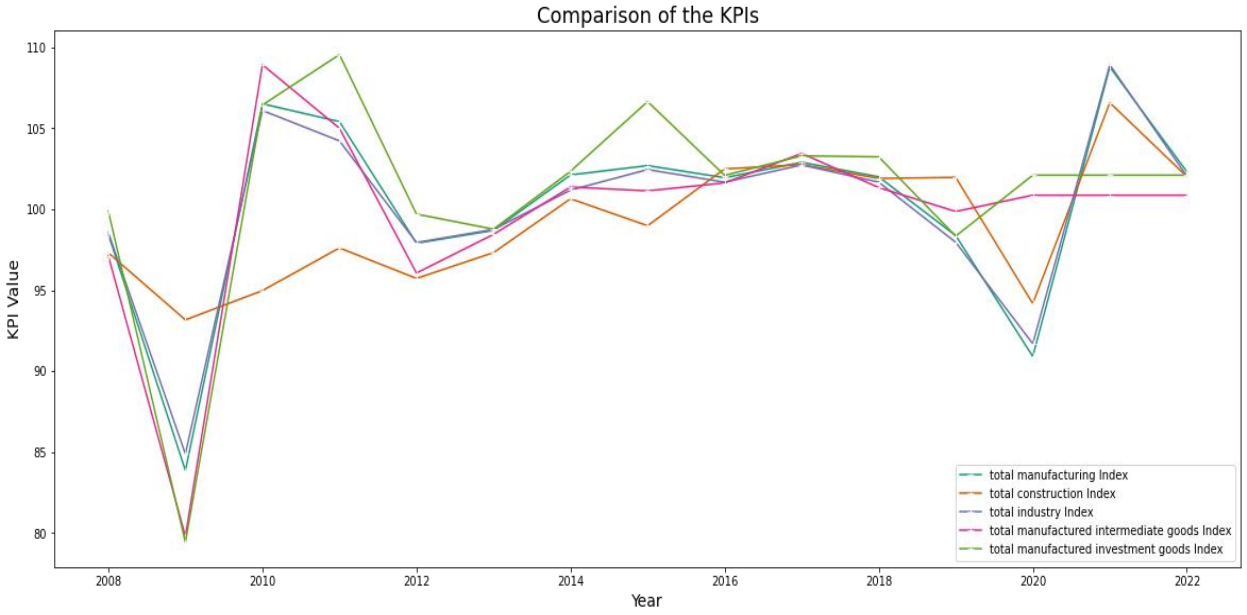


Figure: Comparison of KPI

- The overall construction index was the KPI with the lowest trends over time among the KPIs. It is evident that 2009 saw a significant downward trend.
- The trend in each KPI has shown enormous volatility between the years 2008 and 2017.
- The total manufactured investment goods index and the total manufactured intermediate goods index both show a consistent flow from 2022 to the present.
- For the remaining KPIs, a declining trend was seen in the year 2020. All KPIs saw significant growth between 2009 and 2010, except for the construction index KPI.

Feature Engineering and Transformation

Performance can be improved and the risk of overfitting can be reduced by selecting an acceptable selection of features. In this stage, we've used three main strategies:

- **Normalization Technique**

  The wide range of values for each attribute might make modeling techniques difficult. The normalizing procedure, which modifies all values between 0 and 1, has thus been employed. It is implemented for all the features as all of them are in numerical format.

  X = (X - X_Min) / (X_Max - X_Min)

- **Step-Wise Forward Feature Selection**

  This technique is employed to extract a subset of features that can give the best performance. There is a total of 24 features and we got rid of 2 features through this technique where they were performing well on the dataset with 0.036 RMSE.

- **Cross-Validation**

  Cross-validation helps to tackle issues such as bias and variance. It divides the data into K folds of the same size and utilizes K-1 folds for training and the last fold for testing. It repeats the process for K times with different combinations of the folds. For our dataset, the cross-validation was performed with the baseline model and the result's variance was very small which means that the data and model are not suffering from variance or bias.

Model Building

Danfoss is currently utilizing linear regression to estimate sales, but it is utterly failing. The main cause is that linear models perform worse when dealing with complicated data. In situations like this, we will propose Danfoss employ sophisticated models, like the ensemble technique, which combines a number of weak models into one strong model. Additionally, some

neural networks such as LSTM and RNN can predict future sales by taking past records into consideration. Also, we have implemented time series forecasting techniques such as ARIMA.

Regression Model

We divided our dataset into the test and train sets with a ratio of 90:10. All the regressor models were implemented with a hyperparameter tuning technique named GridSearchCV which gives us the best p arameters as result. The metric that is chosen for evaluating the performance of the model is Root Mean Squared Error(RMSE). We are looking for a model which satisfies two conditions,

1. Has the lowest Test RMSE
2. There are no signs of overfitting i.e. major difference in test RMSE and train RMSE

Here are the results for all the methods used for the training and the testing data.

	model	metrics	value
0	Linear Regression	Training RMSE	0.036016
1	Linear Regression	Testing RMSE	0.031066
2	Gradient Boosting	Training RMSE	0.008722
3	Gradient Boosting	Testing RMSE	0.039890
4	Random Forest	Training RMSE	0.026215
5	Random Forest	Testing RMSE	0.035287
6	AdaBoost	Training RMSE	0.035506
7	AdaBoost	Testing RMSE	0.045287
8	XGBoost	Training RMSE	0.013466
9	XGBoost	Testing RMSE	0.038996
10	Decision Tree	Training RMSE	0.032830
11	Decision Tree	Testing RMSE	0.043571
12	SVM	Training RMSE	0.057447
13	SVM	Testing RMSE	0.058543

If we compare the testing RMSE, the Gradient boosting regressor has performed the best with an RMSE of 0.0087. There may exist chances of overfitting as there is a major difference between the test and train RMSE in gradient boosting regressor. To compare further, we have implemented LSTM and ARIMA.

Long short-term memory (LSTM)

Using LSTM, time series forecasting models can predict future values based on previous, sequential data. This provides greater performance for demand forecasters which results in better decision-making for the business. A trend in sales over a year can be observed in the below graph. From a cursory glance at the graph, it can be noticed that the highest sales were achieved in the year 2010. Sales dipped during 2020 when the COVID pandemic struck the world. Thereafter, it increase during 2021. The lowest sales were achieved during the year 2009 when it was clocked at about 22. The green shade around the line plot shows the maximum and minimum range of the specific year.

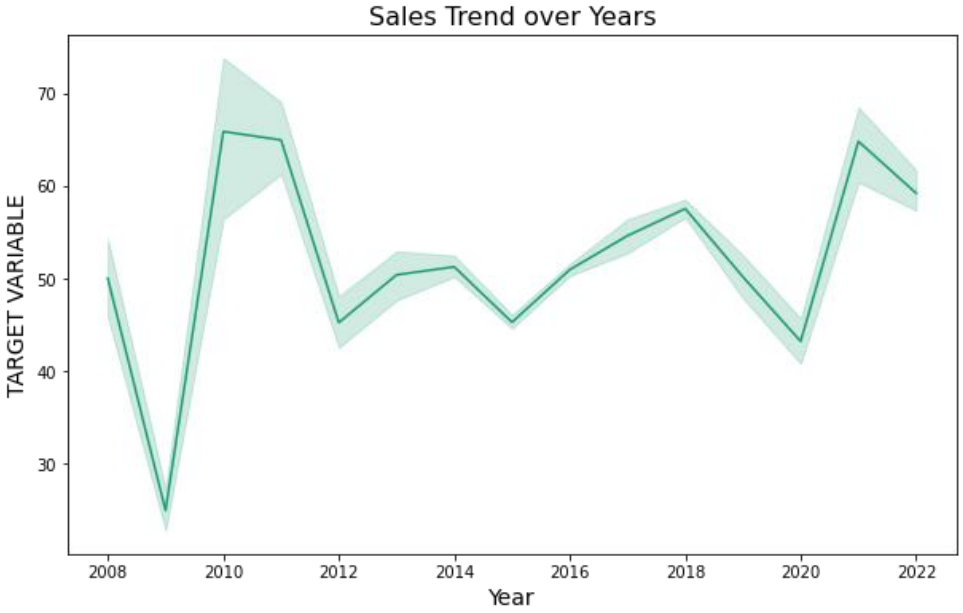


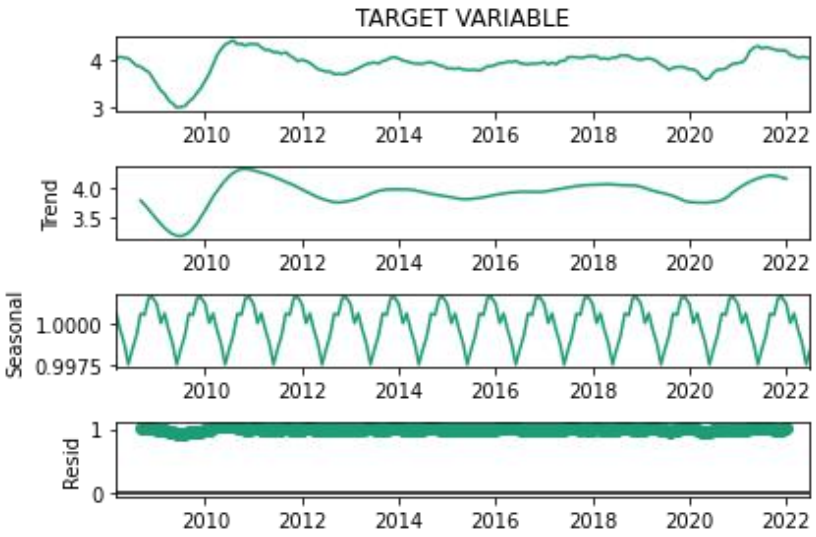
Figure: Trend of sales over years

Long short-term memory networks, or LSTMs, are employed in deep learning. Many recurrent neural networks (RNNs) are able to learn long-term dependencies, particularly in tasks involving sequence prediction. While considering this, the dataset has been divided into train and test and a model has been applied to it. The result while applying LSTM for training is 0.036 RMSE and for the test would be 0.0465 RMSE.

Autoregressive integrated moving average (ARIMA)

A statistical analysis model called ARIMA uses time series data to either forecast future trends or provide a better understanding of the current data set. Lagged moving averages are used by ARIMA to smooth time series data.

Step 1: To understand different components of the time series



The time series decomposition shows that trend is fluctuating and is almost negligible. The time series has a minor influence of seasonality as well. The residual resides at 1 which can be a matter of concern for better prediction.

Step 2: Check the stationarity of the time series

In order to implement time series models, the time series must be stationary. A stationary time series does not have trend and seasonality. From above image, it is unclear that we have both components or not so we will apply Ad Fuller test to confirm the stationarity.

H0: The time series is non-stationary. In other words, it has a time-dependent structure and does not have constant variance over time.

HA: The time series is stationary.

```
ADF Statistic: -4.127697
p-value: 0.000871
Critical Values:
        1%: -3.471
        5%: -2.880
        10%: -2.576
```

The above image shows the results of the statistical test. Here, we can conclude that we have stationarity as the p-value is less than 0.05

Step 3: Plotting Autocorrelation(ACF) and Partial Autocorrelation (PACF)

A set of current data is compared to a set of previous values using autocorrelation to see whether they are correlated. It is widely utilized in forecasting and time series analysis. We can determine the connection between current time-series observations and lag observations, or observations from earlier time steps.

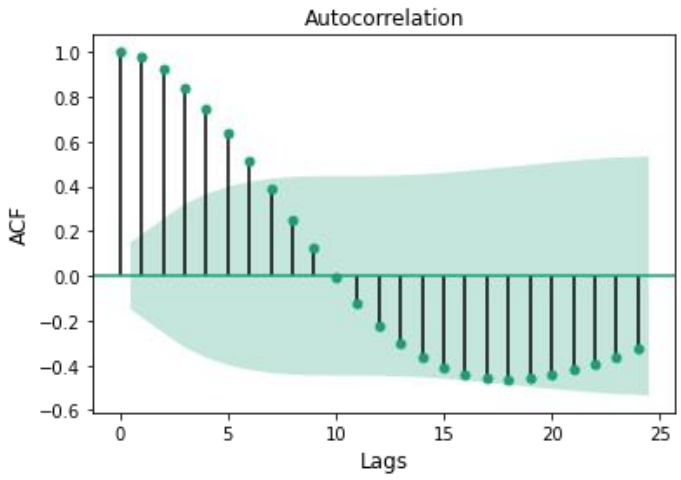
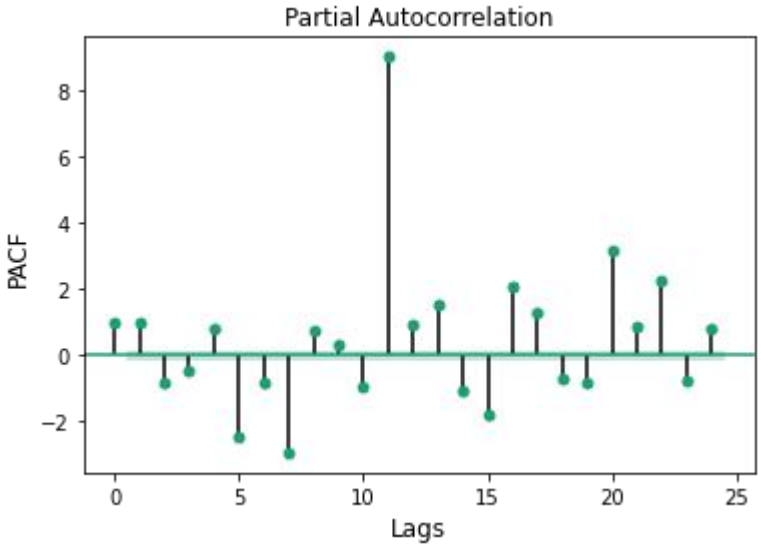


Figure: Correlation of ACF and Lags

From the above ACF plot, we can conclude that the previous time series of lag 1 to 6 are correlated with the current time series as they are outside the error band(green shaded area) whereas others are within the error band. The previous time series from lag 10 is negatively correlated with the current time series.



Partial Autocorrelation computes the direct correlation of the previous time series of lag X to the current time series. It can specify if any previous lag time series can be utilized to predict the current time series. Surprisingly, the time series with lag 11 has a correlation of 0.9 with the current time series and it can be included as a feature for better prediction.

Step 4: Implementing the ARIMA model

ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a generalization of the simpler AutoRegressive Moving Average and adds the notion of integration.

- **Autoregression, or AR:** A model that takes into account the dependency between an observation and a certain number of lag observations.
- **Integrated, or I:** The process of differencing raw observations to render a time series stationary, for as by removing an observation from an observation made at a previous time step.
- **Moving Average, or MA:** A model that makes use of the relationship between a lagged observation and a residual error from a moving average model.

For ARIMA, the values of the hyperparameters p,d, and q are taken as 5,1,3. Through this model, we have achieved an RMSE of 0.9 on the test dataset.

Model Comparison

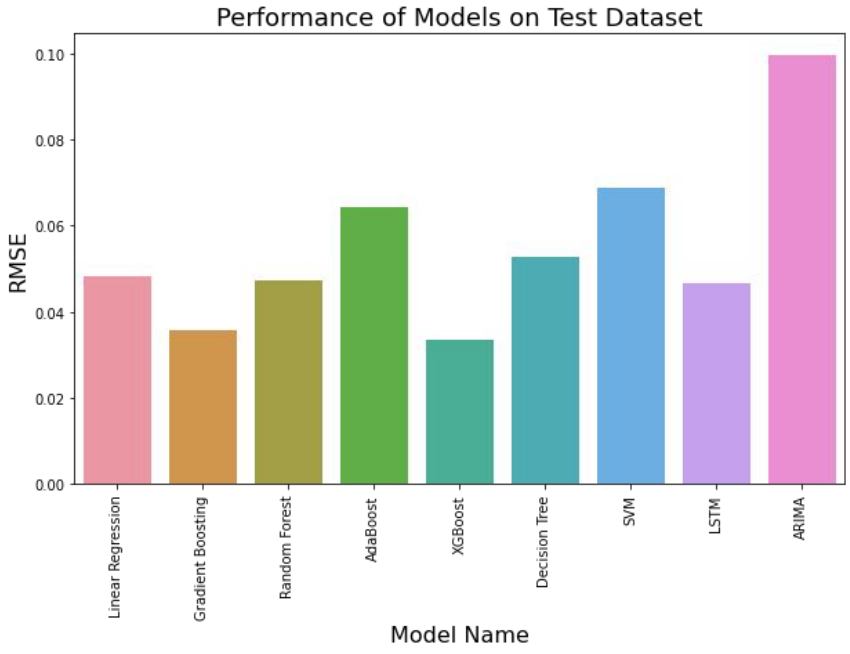


Figure: Performance of Models

Above is the graph where the performance for all the models on the test dataset has been compared. The model with the least loss is the XGBoost and Gradient boosting regressor with the RMSE of 0.0325 and 0.0375 respectively. ARIMA is the worst performer with maximum loss.

Findings

- Anomalies such as missing values and outliers were detected in several features of the dataset. Missing values were replaced by the feature's median value. Outliers were either trimmed or limited to certain values depending on their severity.
- Good predictors were yielded from the correlation plot. Also, There were traces of multicollinearity from the correlation plot where independent features were highly correlated.
- There was a steady flow in the total manufactured intermediate goods index and total manufactured investment goods index in 2020 while a downward slope was observed for the rest of the KPIs.
- Data is transformed using the normalization technique so that all the features fall in the same range. Step-wise feature selection techniques yielded with best 22 features which can help to boost the performance of the models. Additionally, cross-validation showed that there are no symptoms of bias or variance.
- We implemented several modeling techniques such as regression, neural networks such as LSTM, and time series forecasting such as ARIMA. The model with the least loss is the XGBoost and Gradient boosting regressor with the RMSE of 0.0325 and 0.0375 respectively. ARIMA is the worst performer with maximum loss.

Conclusion

Initially, data is cleaned to get rid of missing values and outliers. Secondly, key features are selected using correlation plots that are highly correlated with the target feature. The line plot compared different KPIs and factors such as BCI and CCI. The line plot of the target variable across the time period of the year showed the impact of the COVID pandemic where the sales decreased by a big margin in 2020. Feature transformation and engineering techniques such as normalization and step-wise forward feature selection helped to extract a subset of the best features in the same range that gives the best performance. Data modeling techniques such as regression, neural network, and ARIMA were employed where the models such as XGboost regressor and Gradient boosting regressor outperformed other models.

References

Danfoss- Engineering Tomorrow. (n.d.). Retrieved October 29, 2022, from https://www.danfoss.com/en-us/

Jordan, M. (2021, August 26). What is Regression Analysis and Why Should I Use It? Alchemer. https://www.alchemer.com/resources/blog/regression-analysis/

Wikipedia contributors. (2022b, April 11). Danfoss. Wikipedia. https://en.wikipedia.org/wiki/Danfoss

Pandian, S. (2021, November 12). *A comprehensive guide to time series analysis*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-to-time-series-analysis/

Karaman, B. (2022, September 29). *Predicting sales*. Medium. https://towardsdatascience.com/predicting-sales-611cb5a252de

Needle, F. (2020, December 21). *How to use regression analysis to forecast sales: A step-by-step guide*. HubSpot Blog | Marketing, Sales, Agency, and Customer Success Content. https://blog.hubspot.com/sales/regression-analysis-to-forecast-sales