XN Project: Individual Draft Presentation

Esha Mulki
Mohammad Mohavedhi
Ajoy Kumar Nandakumar
Taiye Murtala
College of Professional Studies,
Northeastern University
ALY6080: Integrated Experiential Learning
Instructor: Dr. Matthew Goodwin

Executive Summary

- Improve Danfoss client's sales forecasting.
- Study of the sponsor business sector to discuss the dependent and independent variables with the sponsor.
- Data cleaning, reformatting, and exploratory analysis, model building and model evaluation.
- Used programming language: R
- Model Evaluation- computed RMSE and MAPE values

Approach in dealing with the missing values

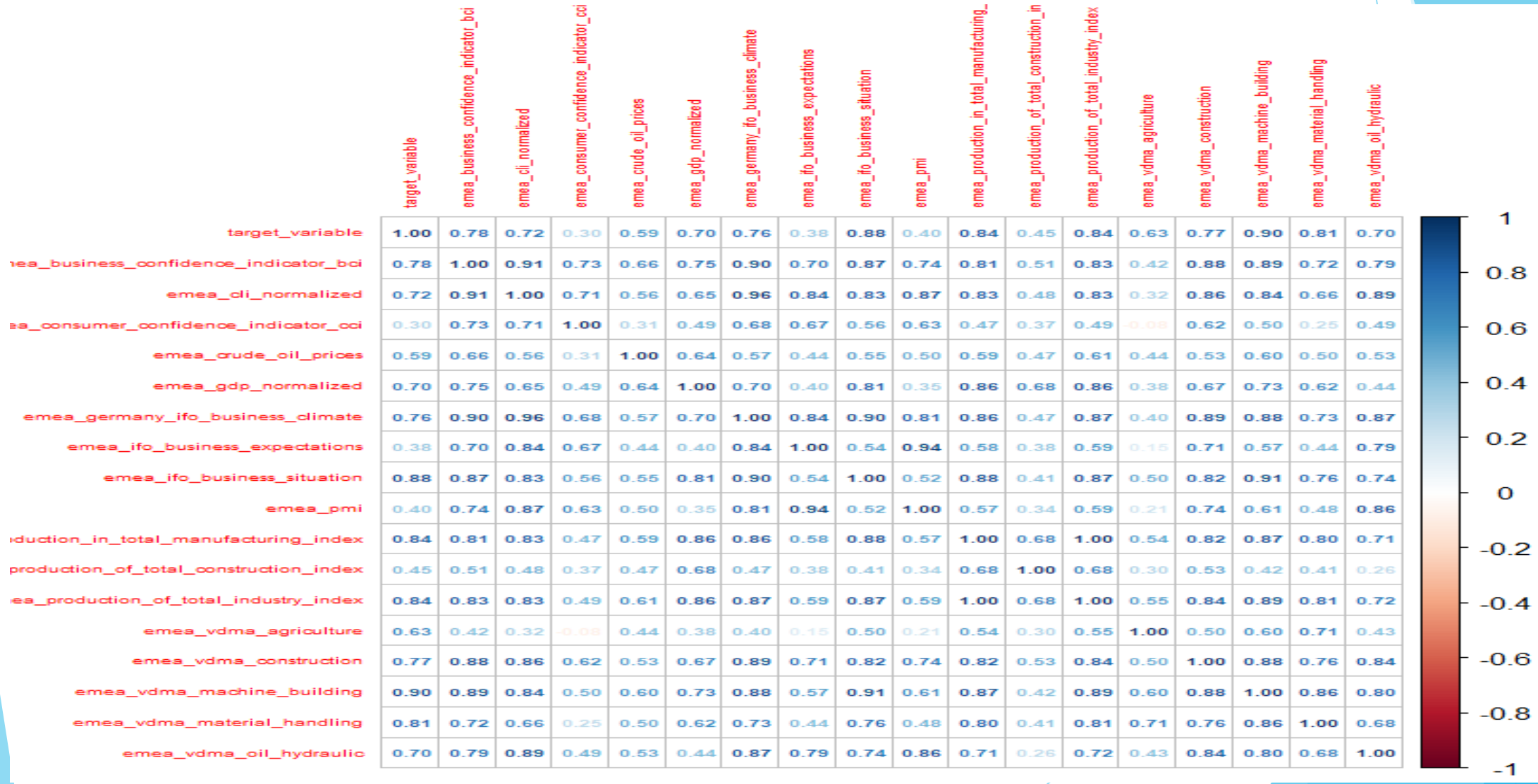
- There were missing values for some variables like the employment_rate. total_manufactured_intermediate_goods_index, total_manufactured_investment_goods_index.
- The missing values for these variables were imputed with the median value of the variable.
- This approach provides the tolerance to the outliers in the variables.

```
df1$x2[is.na(df1$x2)]<-median(df1$x2,na.rm=TRUE)
df1
```

EDA Visualization

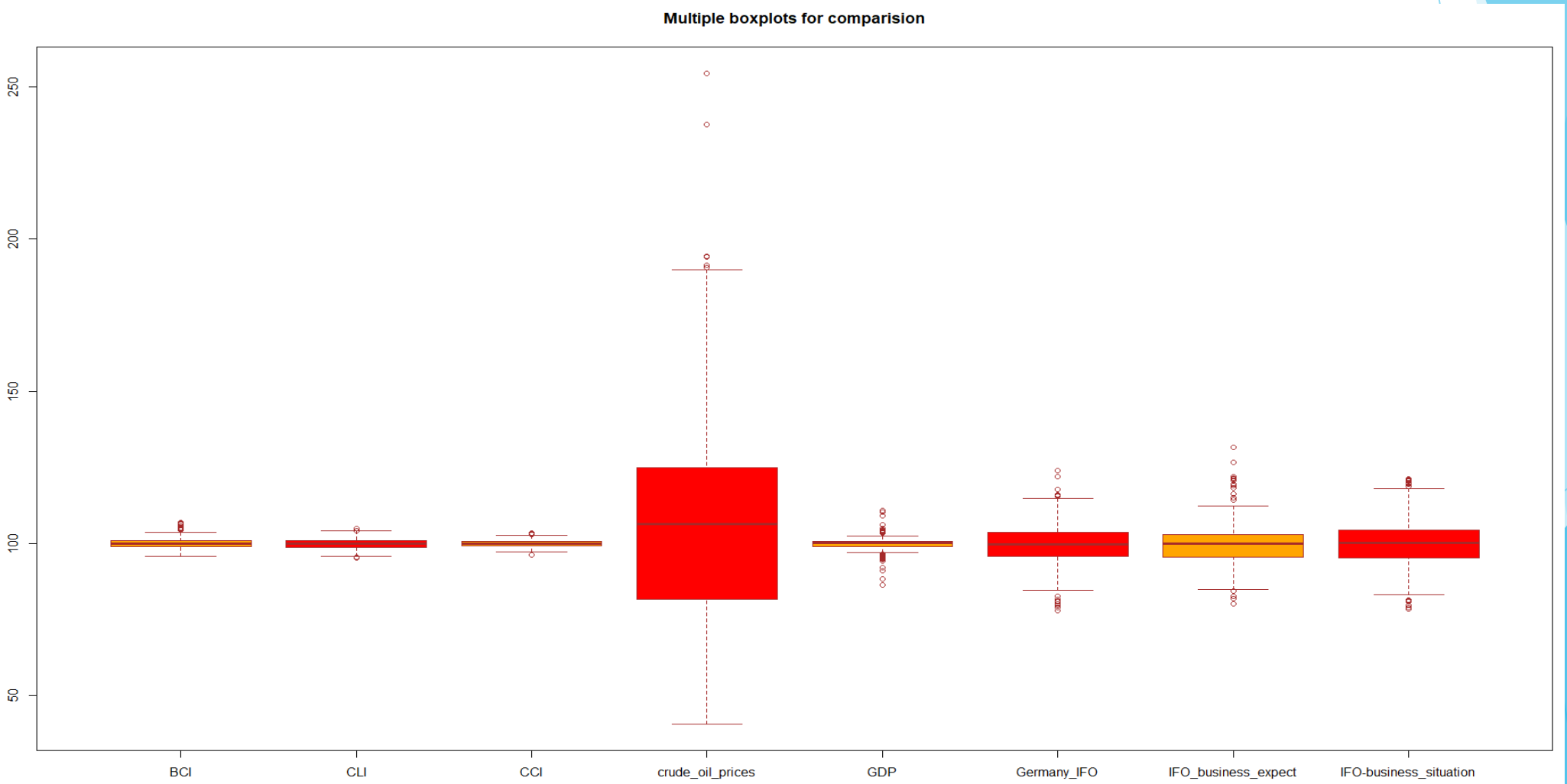
Correlation Plot

The below correlation matrix shows the correlation between the variables and the target variable. This helps us in proceeding further with the variable selection



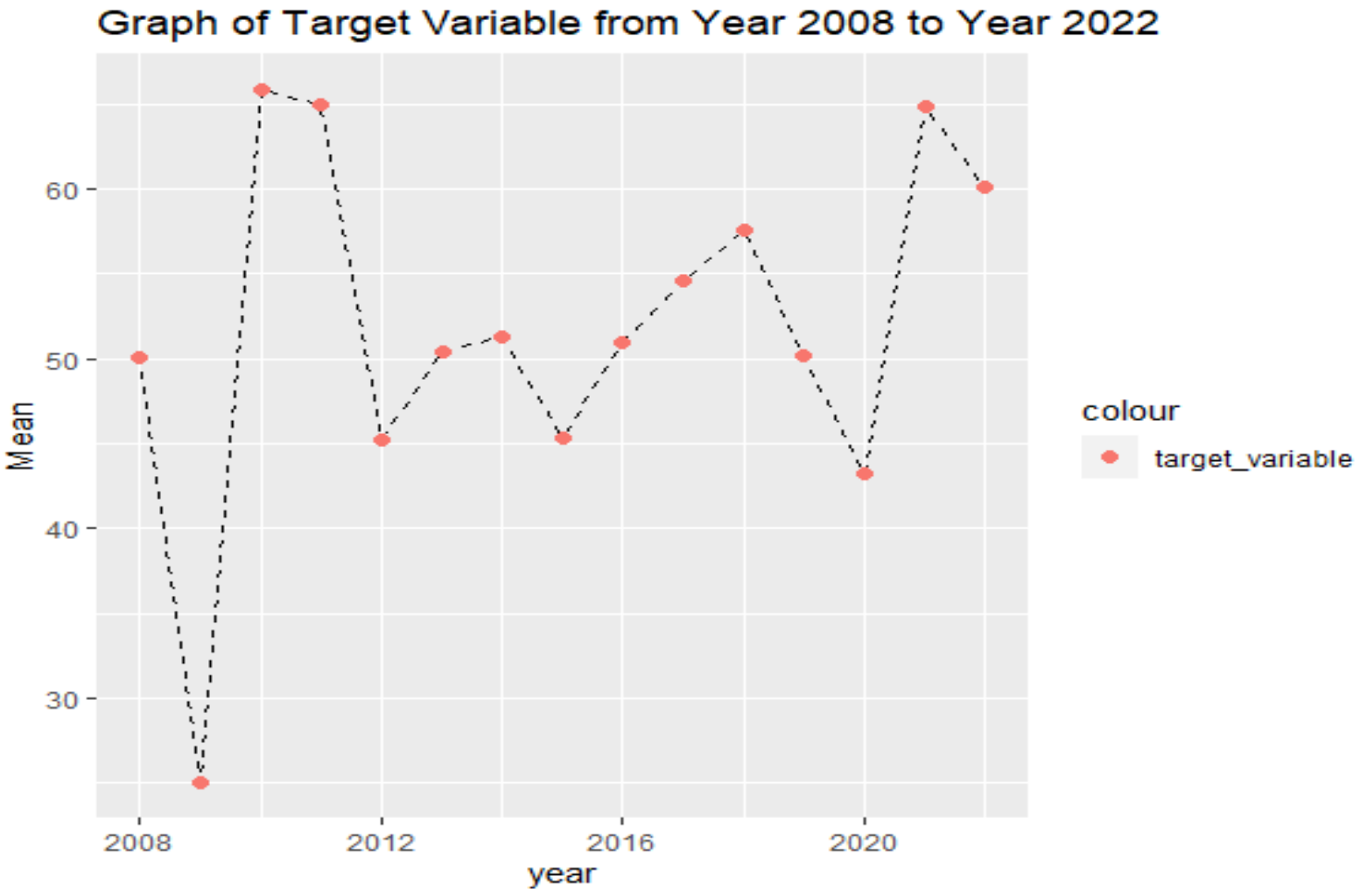
Boxplot

Below boxplot helps to understand the variables with outliers. Crude oil prices has huge outliers



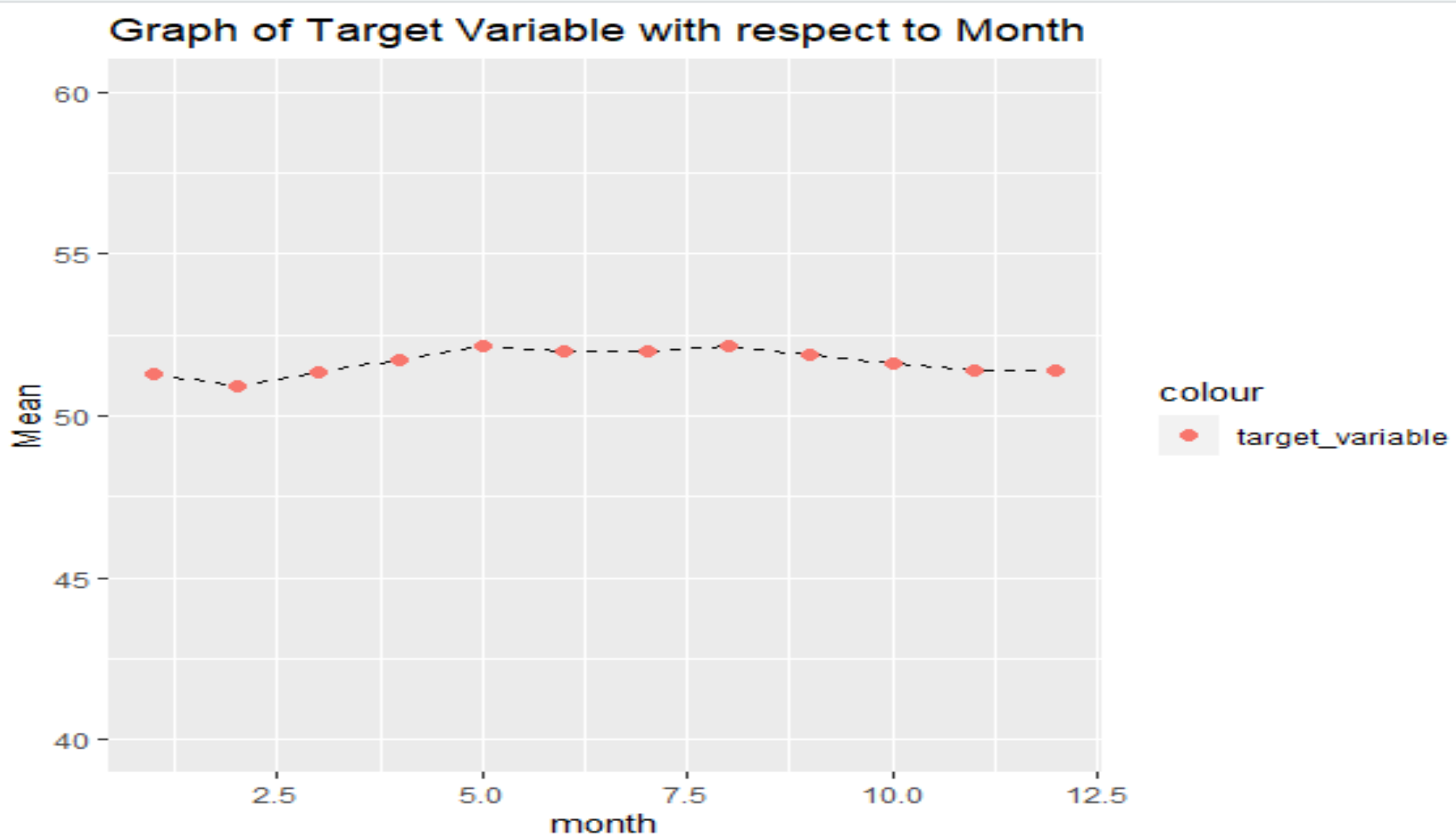
Mean Target Variable vs Year

The mean target variable values fluctuate from 2008 to 2012, but after that point, they become more or less stable until 2018, after which they start to fluctuate till 2022.



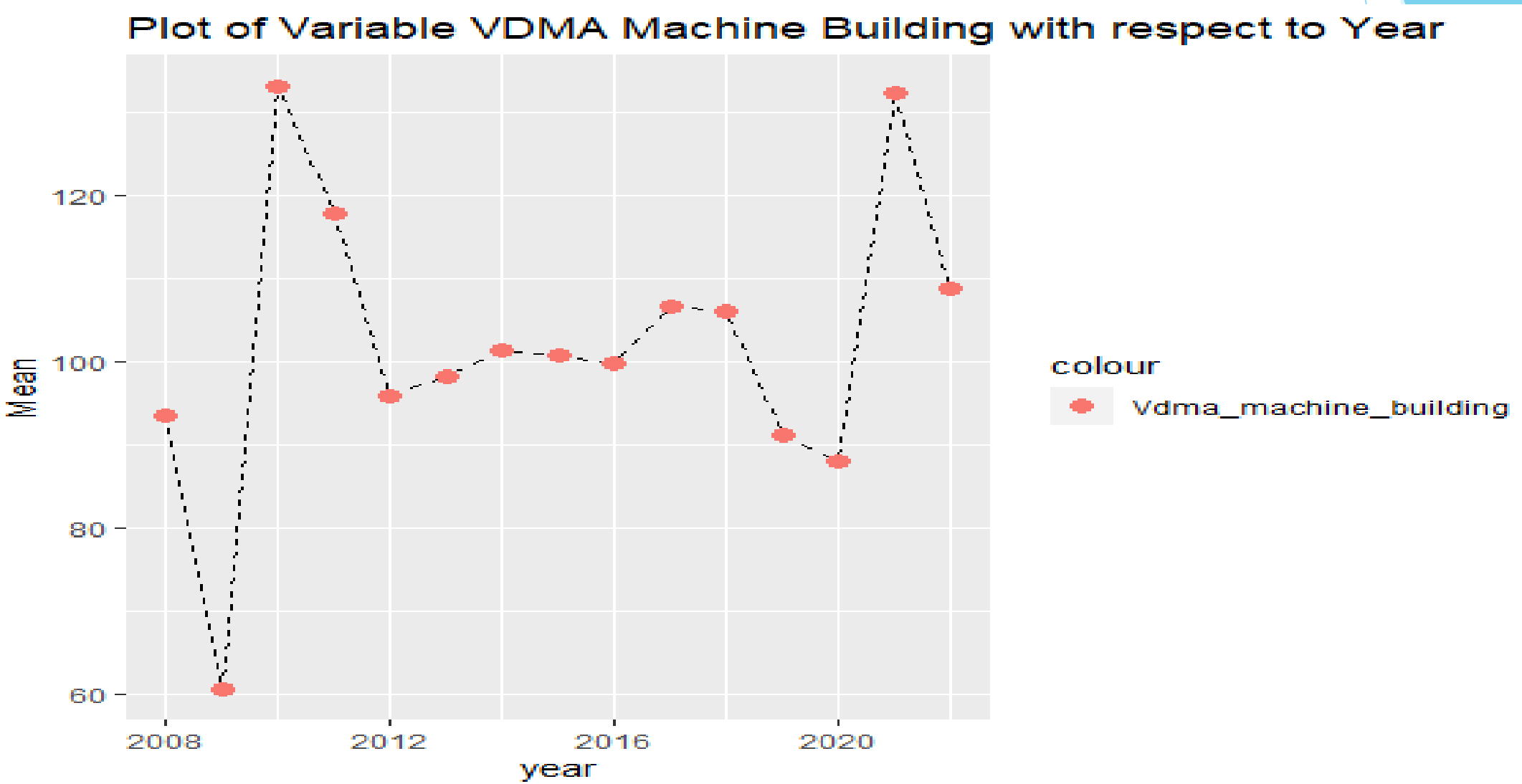
Mean Target Variable vs Month

There is not much variation in the target variable's graph with regard to the months.



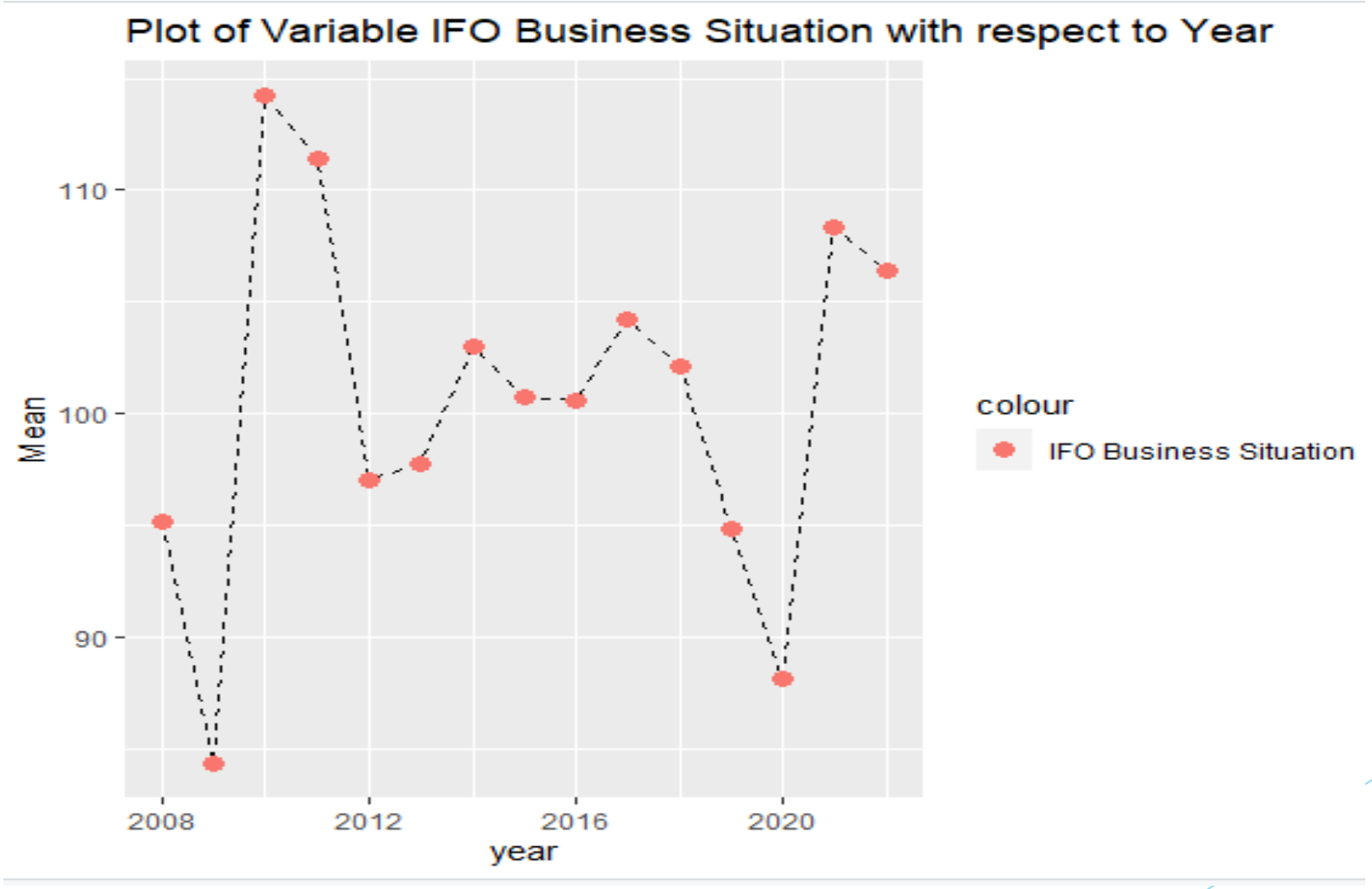
Mean value VDMA Machine Building vs Years

The mean values of the VDMA machine building variables vary from 2008 to 2012 and from 2019 to 2022. The values are more or less stable from 2012 to 2018.



Mean value of IFO Business Situation vs Years

High fluctuation in the value in the period from 2008 to 2012 and between 2020 and 2022



Milestones of Project

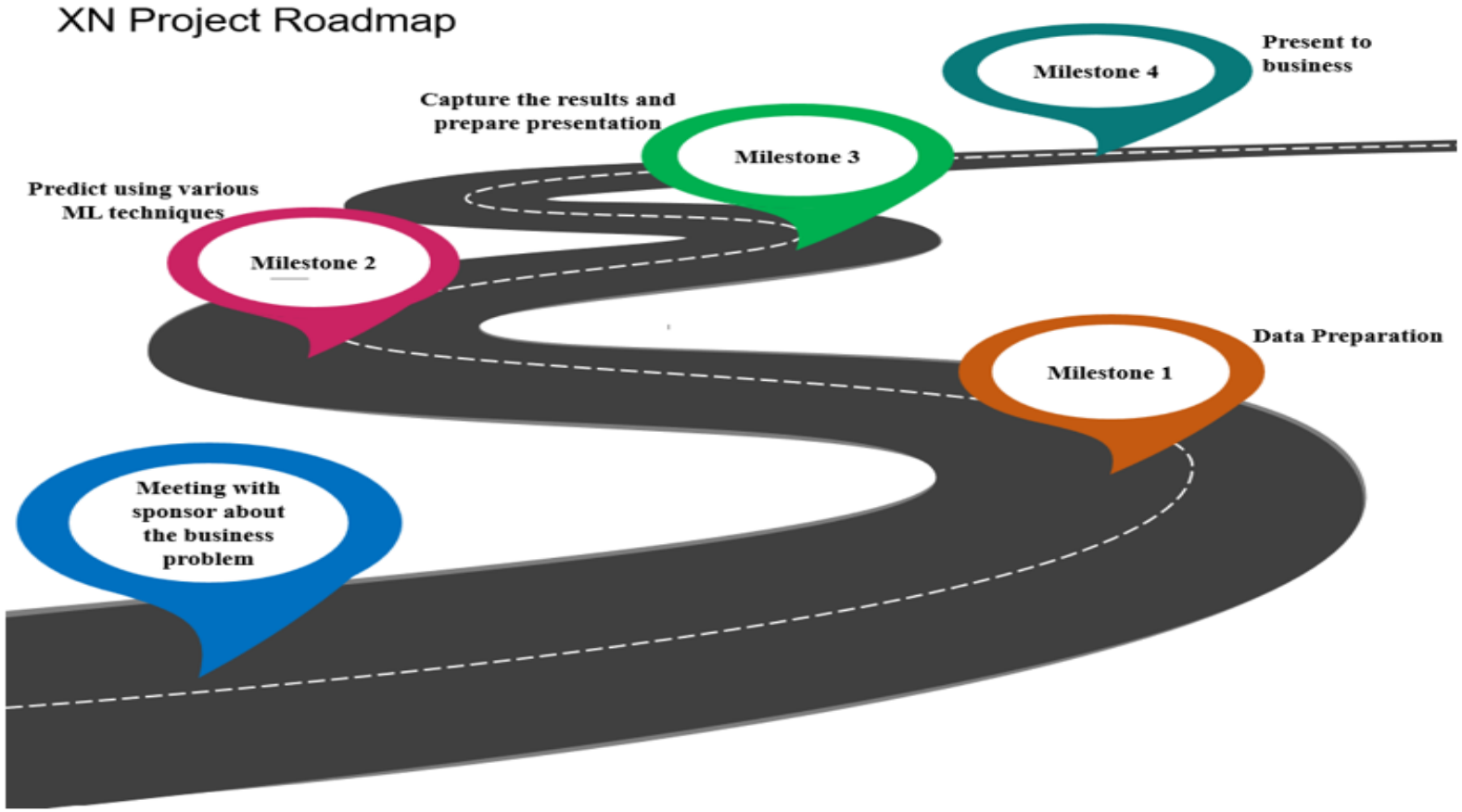
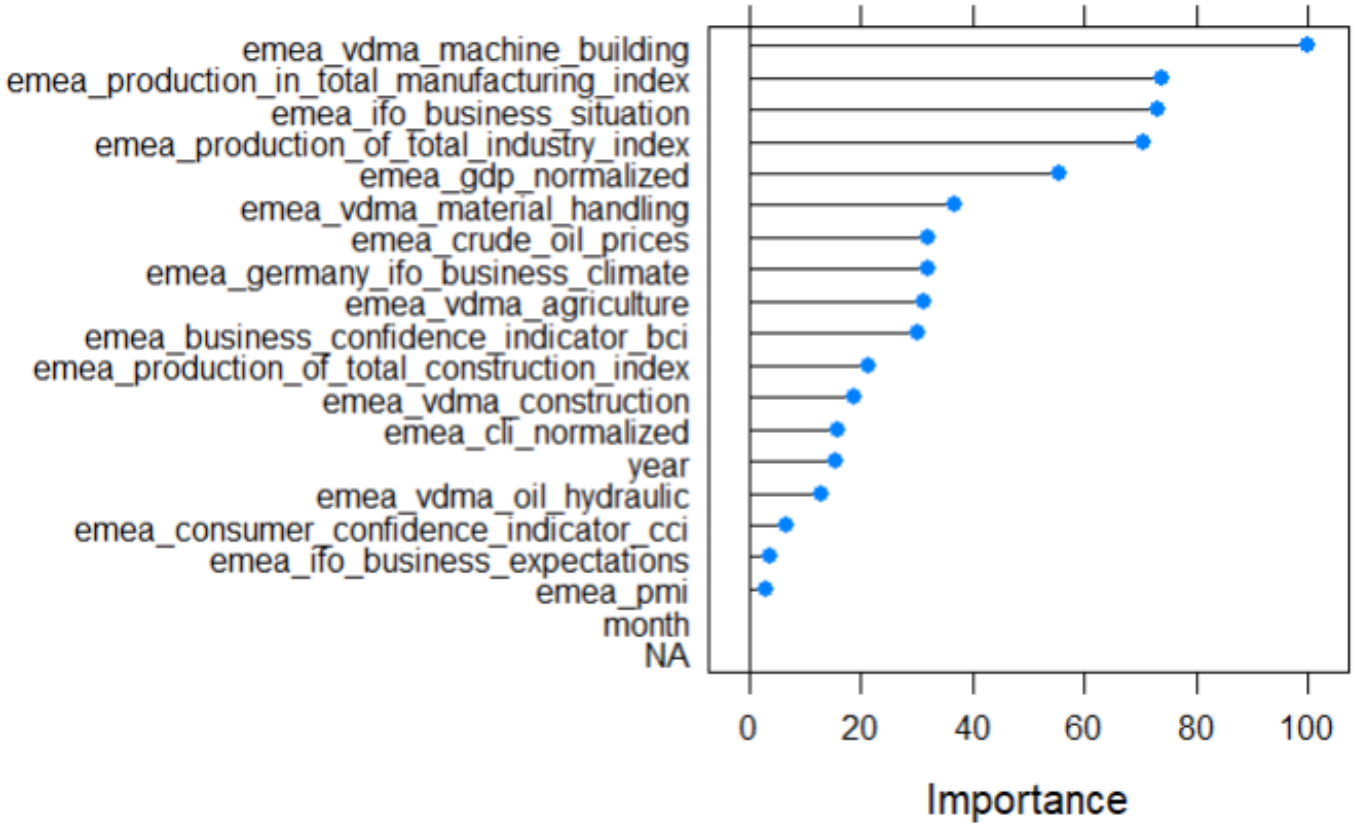


Fig 1: Milestones for the project

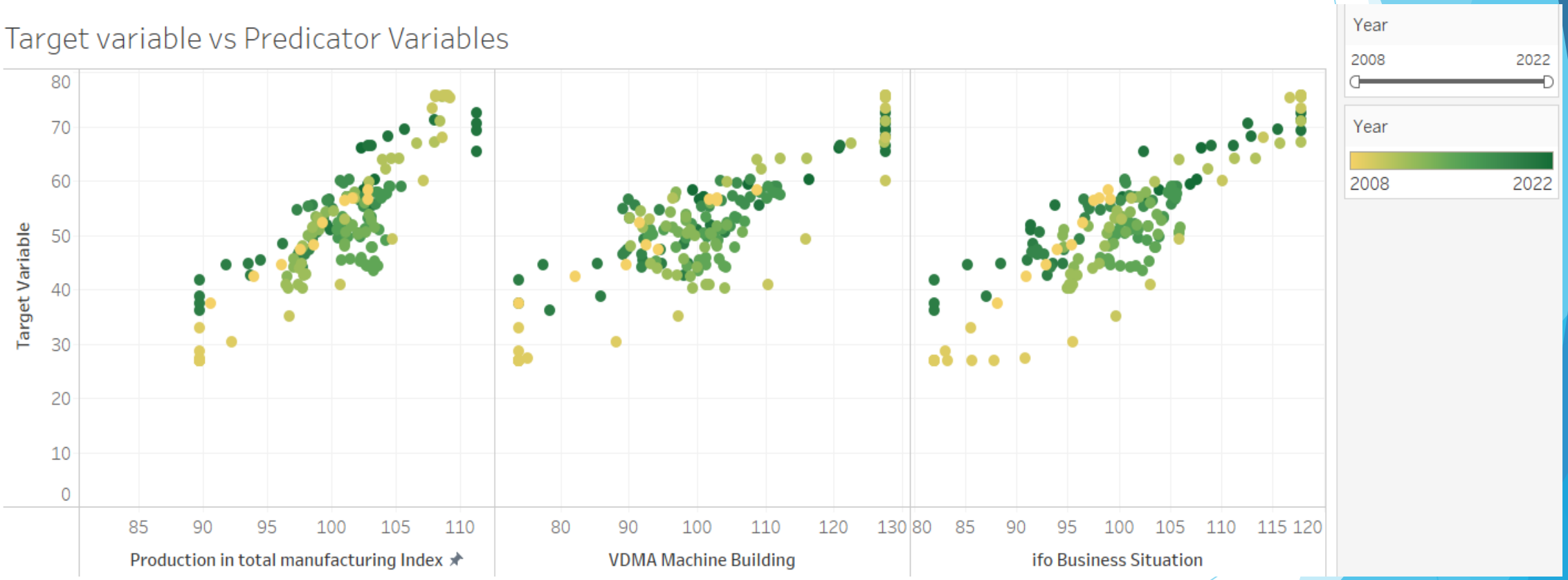
Some of the significant variables for our model

Below graph shows the important of variables obtained from the output from the decision tree



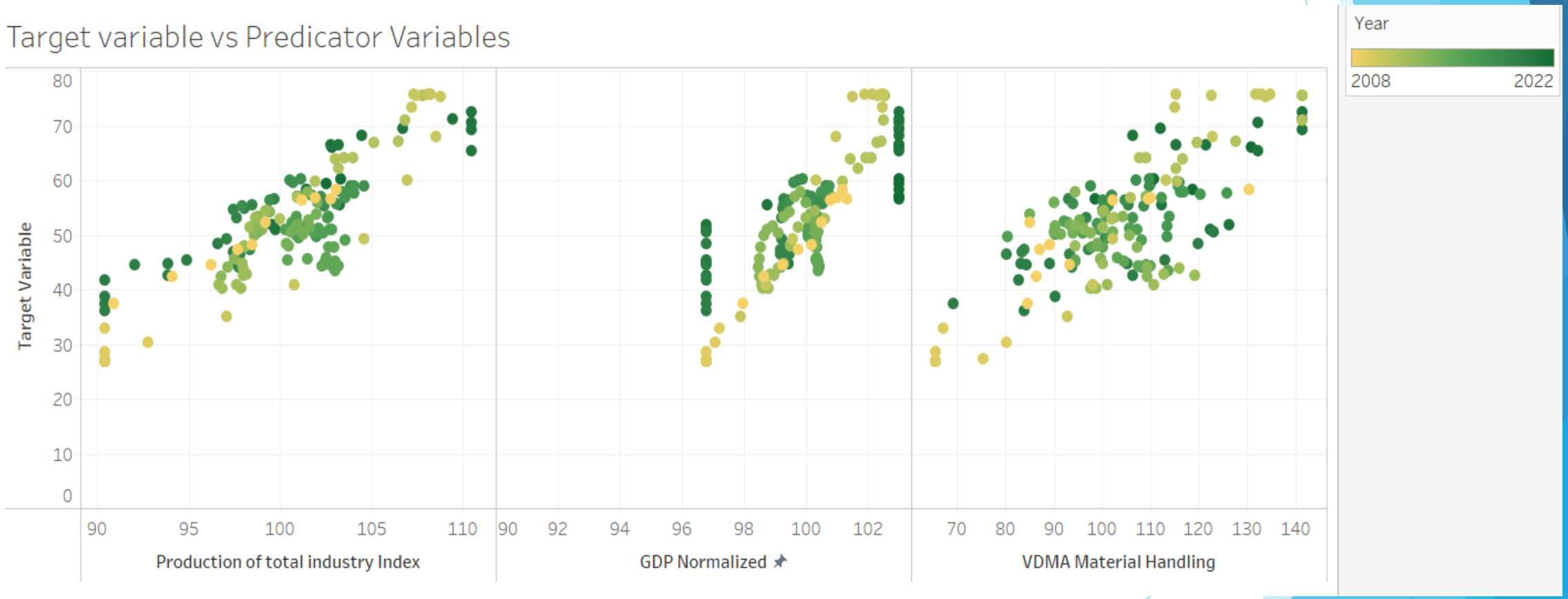
Relationship of Target variable with the significant variables

The below graph shows that the target variable has a somewhat linear relationship with the below variables



Relationship of Target variable with the significant variables

The below graph shows that the target variable has a somewhat linear relationship with the below variables



Summary: EDA Findings

- The mean target variable values varied between the years of 2008 and 2012, but after that, they stayed largely stable until 2018, then they started to vary once more.
- When compared to the months, it was observed that the target variable does not vary significantly.
- The target variable and some of the variables such as the IFO business scenario, VDMA machine building, Production in Total Manufacturing Index, and VDMA Material Building have a linear relationship.

- The high collinearities of some of the variables also contribute to multicollinearity.

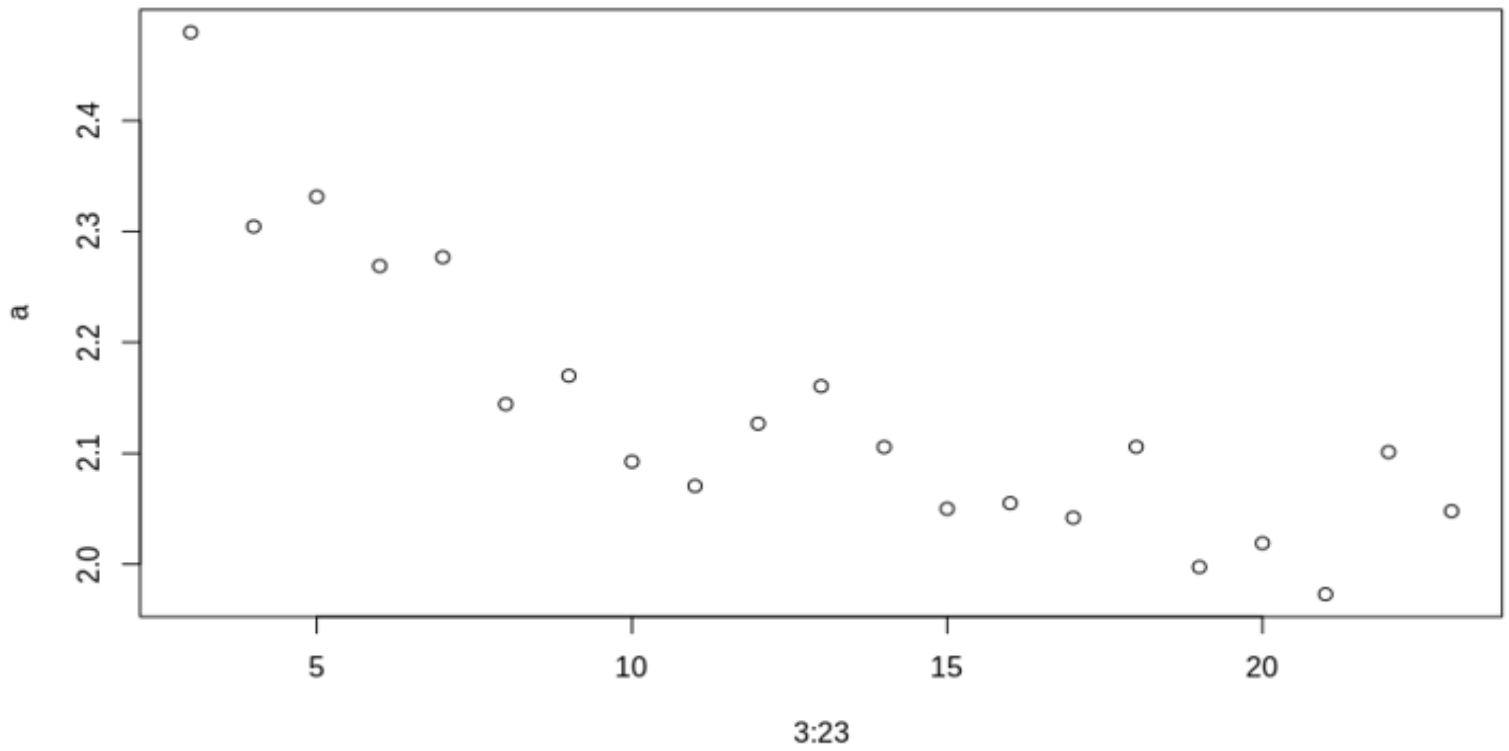
Summary: EDA Findings(contd)

- The price of crude oil has changed significantly throughout the years, as can be seen from the value of crude oil prices over the years.
- Between the years of 2008 and 2012, it was discovered that the mean IFO business scenario value varied; however, after that, the value remained essentially constant until 2018, at which time it started to vary once more.
- In general, it can be seen that Danfoss's revenue was significantly impacted by both the pandemic that occurred between 2020 and 2022 and the world recession that occurred between 2008 and 2009.

Model Built

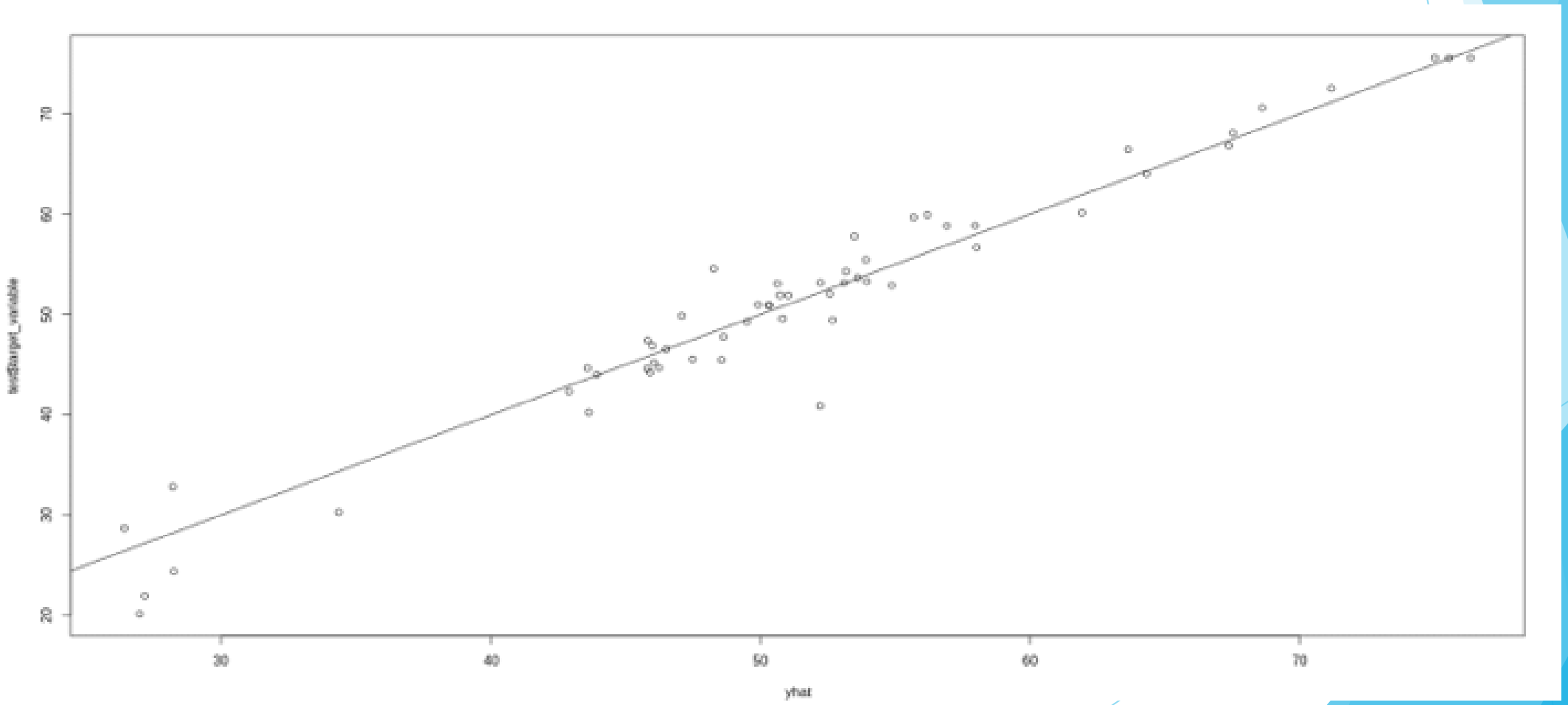
Prediction using Random Forest

- Random Forests - Used in a wide range of industries. Its foundation is the creation of several decision trees, each of which is built using a different subset of your training data.
- The plot below shows the mean of error for each forest based on different parameter or number of variables picked up in each tree.



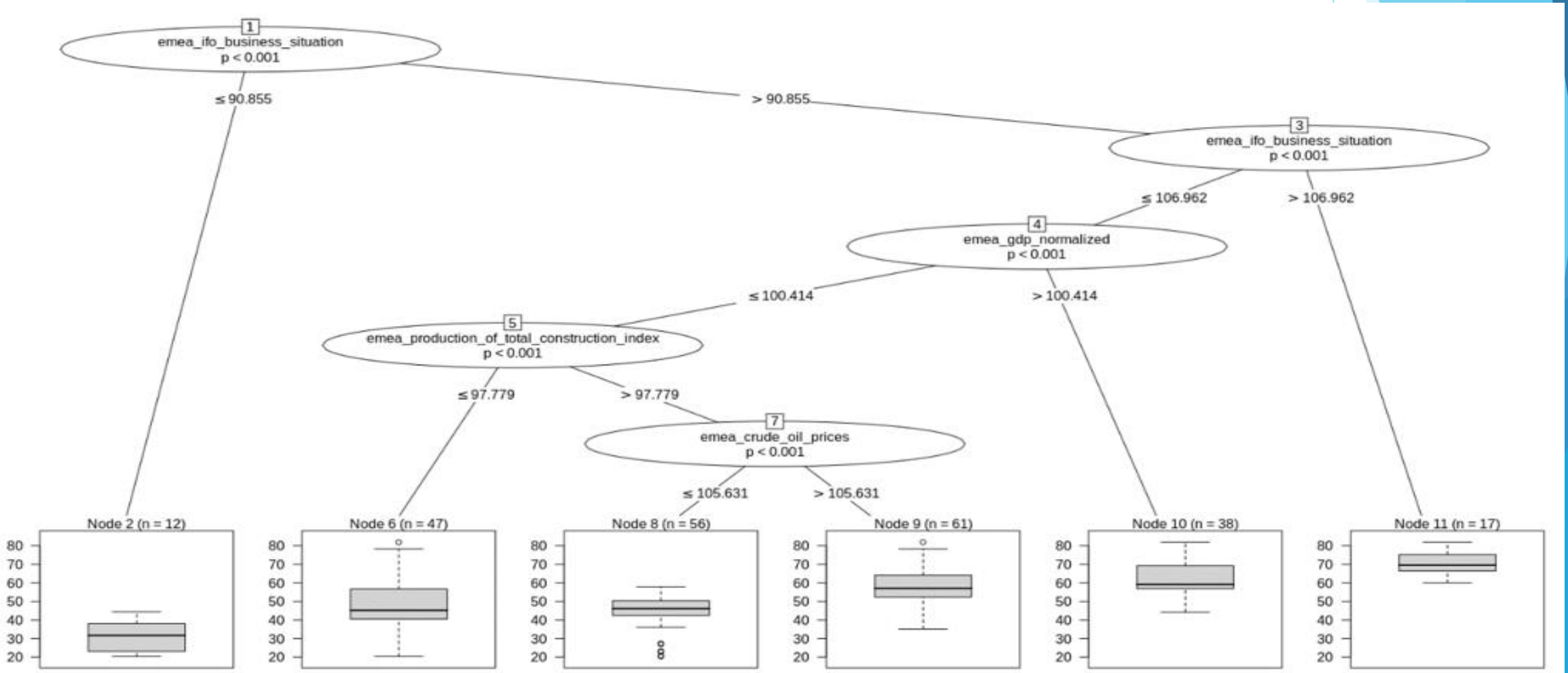
Prediction using Random Forest

- Most of the points are close to y=x line indicating that the residuals are almost 0.
- The RMSE value for final random forest is 8.056074



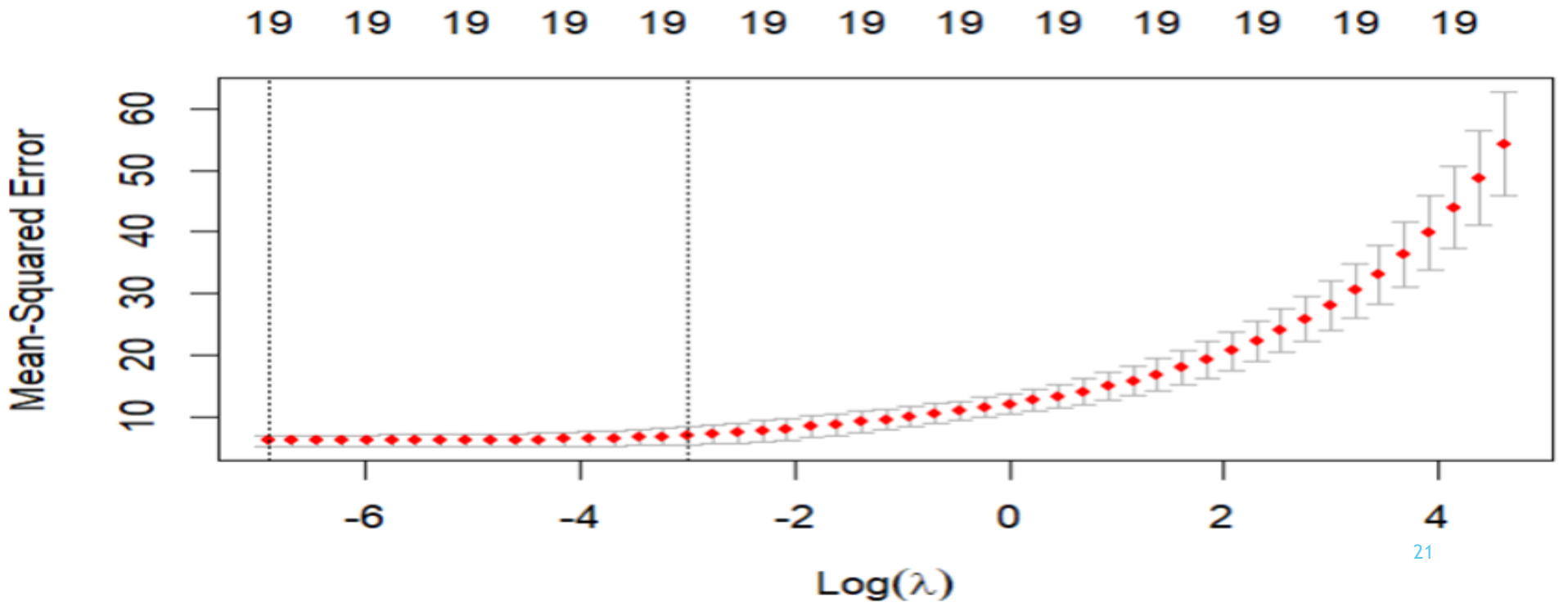
Prediction using Random Forest

- Decision tree is generated by the random forest algorithms.
- It's not the most successful tree but it's impressive how it has done a great job in some nodes.



Ridge Regression(L2 Norm)

- The 19 numbers at the top of the graph display the number of variables (non-zero coefficients) that would be kept in the model for the specified value of.
- The dashed line on the right represents the greatest value of within one standard error (1se) of the minimum, which is -6.907755, the dashed line on the far left represents the least value of lambda that minimises out-of-sample loss, which is -2.993361.



Ridge Regression(L2 Norm)

```
> predictions_train <- predict(ridge_reg, s = optimal_lambda, newx = train_x)
> eval_results(train_y, predictions_train, train_x)
      RMSE   Rsquare
1 1.976778 0.9728505
>
> # Prediction and evaluation on test data
> predictions_test <- predict(ridge_reg, s = optimal_lambda, newx = test_x)
> eval_results(test_y, predictions_test, test_x)
      RMSE   Rsquare
1 2.657362 0.9449558
```

- The RMSE and R square values for the train and test dataset is shown above.
- It can be seen that the RMSE for the test data is higher than the train and the R squared value is lower for the test than the train data.

Lasso Regression(L1 Norm)

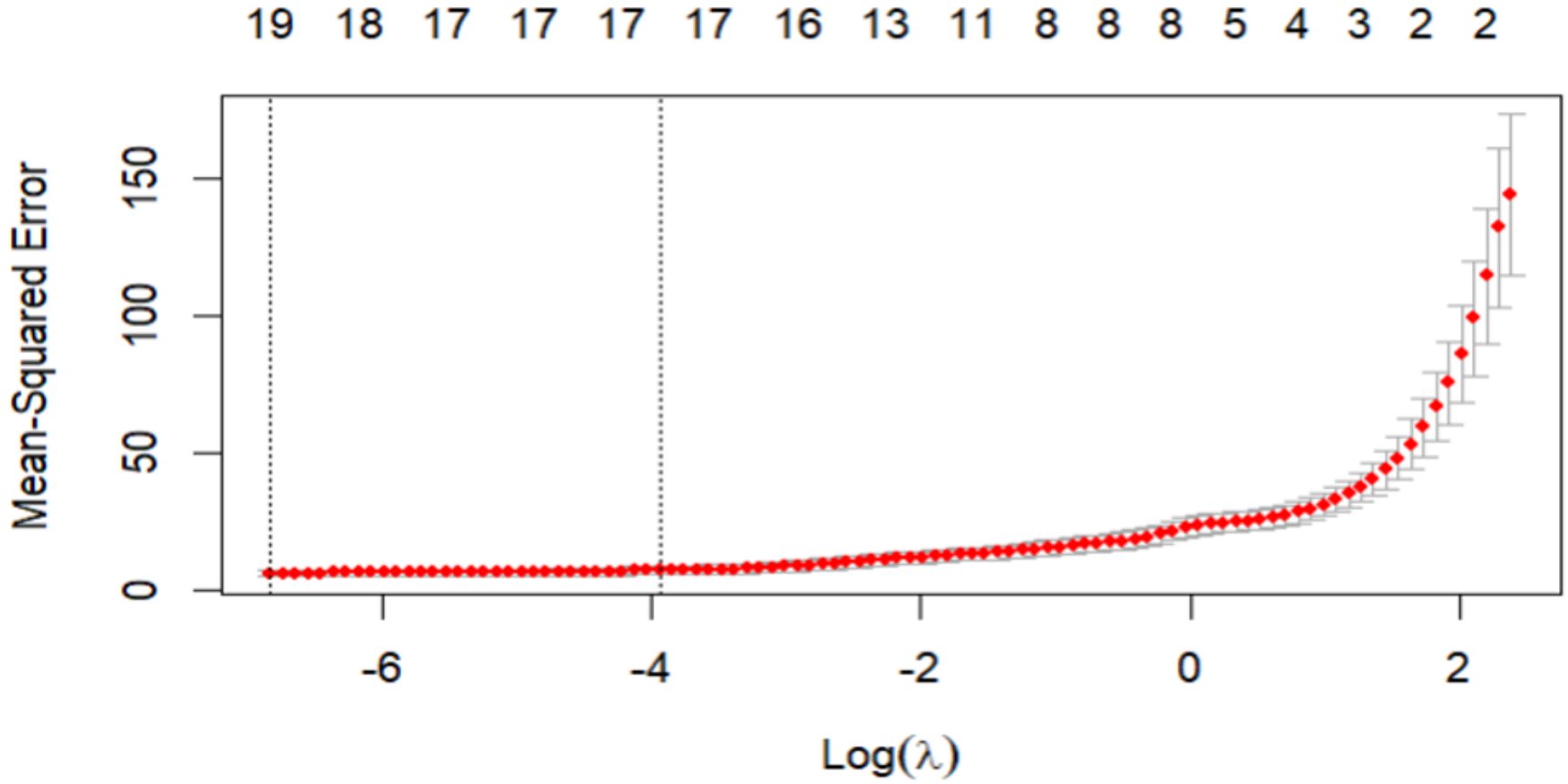
The regression coefficients are brought closer to zero by punishing the regression model with a penalty term called L1 norm, which is the sum of the absolute coefficients.

With regard to lasso regression, the penalty has the result of setting some coefficient estimates that barely affect the model exactly equal to zero.

To reduce the complexity of the model, variable selection using lasso might be viewed as an alternative to subset selection methods.

Lasso Regression(L1 Norm)

- Cross validation's output shows that there are 19 nonzero coefficients for the Lamda.min model and 17 for the Lamda.1se model.
- The value of log lamda min is -6.826579, and log lamda.1se is equal to -3.942533.
- Furthermore, when comparing LASSO regression to Ridge regression, we can observe a decline in the actual lambda values.

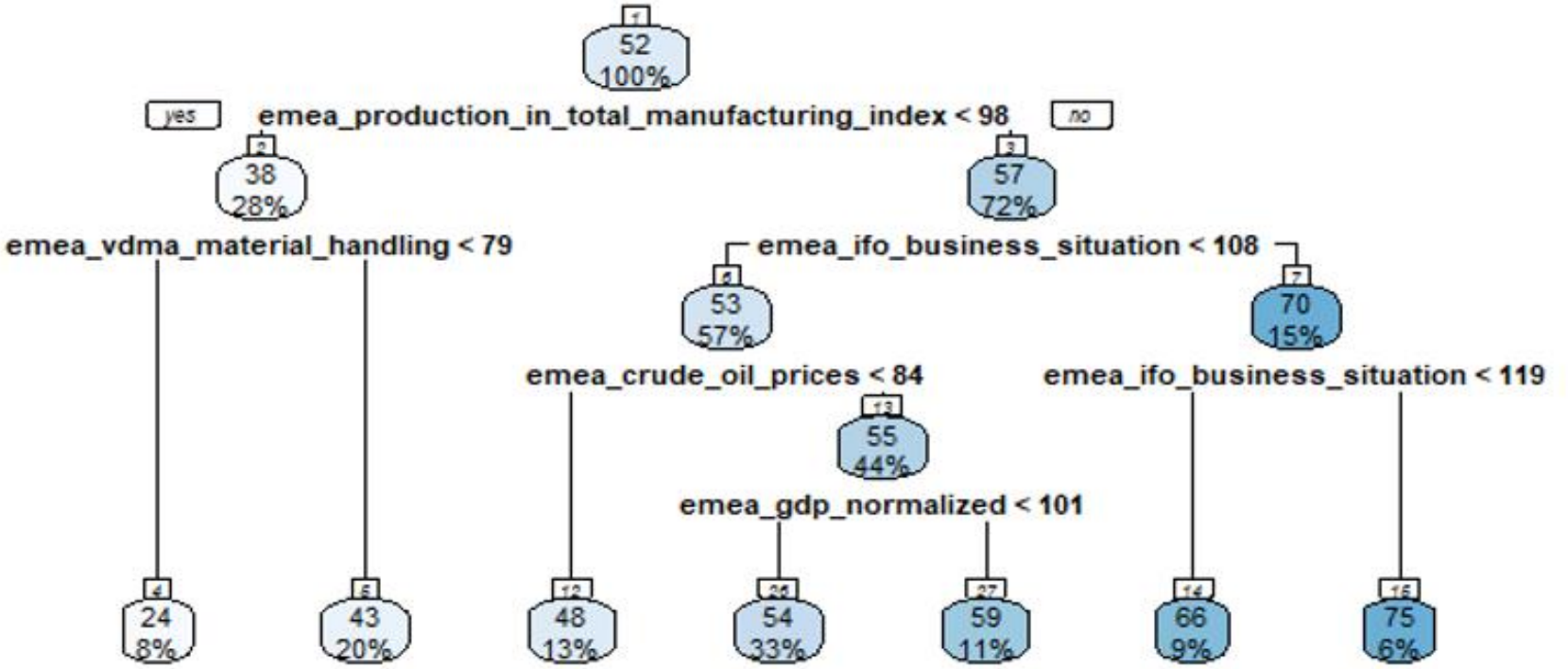


Lasso Regression(L1 Norm)

```
> predictions_train <- predict(lasso_model, s = lambda_best, newx = train_x)
> eval_results(train_y, predictions_train, train_x)
      RMSE   Rsquare
1 1.975359 0.9728894
>
> predictions_test <- predict(lasso_model, s = lambda_best, newx = test_x)
> eval_results(test_y, predictions_test, test_x)
      RMSE   Rsquare
1 2.681538 0.9439497
```

- The RMSE and R square values for the train and test dataset is shown above.
- It can be seen that the RMSE for the test data is higher than the train and the R squared value is lower for the test than the train data.

Decision Tree Regression



It is seen that the tree is first split on emea_production_in_total_manufacturing_index attribute for values < 98

Next split using emea_vdma_material_handling attribute with values < 79 and similarly for other nodes.

On the leaf node it can be seen that the leftmost leaf node contains 8% of data and the predicted value is 24. Similarly the rest of the tree can be interpretated.

Decision Tree Regression Results

```
No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 108, 107, 107, 107, 107, 107,
Resampling results:

  RMSE      Rsquared   MAE
  3.795476  0.8953531  3.022114
```

The above result of the decision tree regression is obtained by using bagging and 10 fold cross validated

ARIMA Model

An autoregressive integrated moving average ARIMA model was also used for predicting the future 3 months sales.

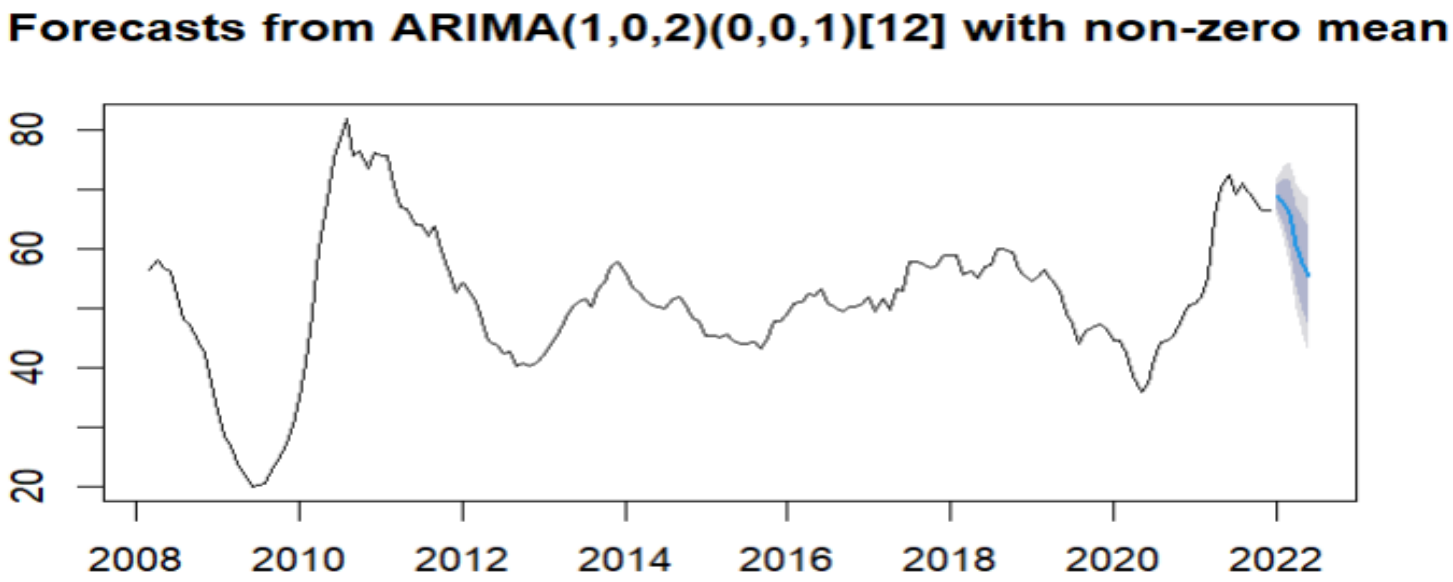
```
> arima_model_final
Series: ts_final_uni
ARIMA(1,0,2)(0,0,1)[12] with non-zero mean

Coefficients:
         ar1     ma1     ma2     sma1     mean
      0.9373  0.6262  0.5650  -0.8191  51.7298
s.e.  0.0300  0.0674  0.0629   0.0817   1.0356

sigma^2 = 2.767:  log likelihood = -335.58
AIC=683.16   AICc=683.67   BIC=702.01
```

```
> RMSE(pred_data$arima_pred, pred_data$target_variable)
[1] 5.071104
>
```

The RMSE value for the ARIMA model is 5.07, the AIC and BIC values should be as low as possible. Log Likelihoods determines how well the model fits the provided data

ARIMA Model

The time frame in question in this case is from March 2008 to October 2022. We make predictions for August, September, and October of 2022.The confidence band is depicted by the grey-shaded area. With a 95% confidence level, we project the sales figures for August, September, and October 2022.



Summary

- The random forest, lasso & ridge regression and decision tree regression and ARIMA models has been used predicting the sales.
- Below are the results for the same.
- The ridge and lasso helps in the better prediction by using the penalization method for large number of variables

Model				
Random Forest	**Ridge Regression**	**Lasso Regression**	**Decision Tree**	**ARIMA Model**
RMSE - 8.05	RMSE - 2.68	RMSE - 2.65	RMSE - 3.79	RMSE - 5.07

Recommendations

- Using the ARIMA model for predicting the sales for the next three months it is anticipated that the sales of the company will be decreasing.
- Hence necessary measures can be taken by Danfoss regarding this.
- Also it can focus on increasing the values for the most significant variables like VDMA machine building, production in total manufacturing index and ifo business situations to prevent the sales from decreasing.

Future Research

- Future research will be done to reduce the RMSE values below 1.8 by using hyperparameter tunning techniques and other tunning techniques.
- Other machine learning algorithms will also be tried to see which ML algorithm gives the best result and suits best for our sales prediction.

References

- Kelwig,D.(2022, June 24).The definitive guide to sales forecasting methodologies. Zendesk. Retrieved from https://www.zendesk.com/blog/5-essential-sales-forecasting-techniques/
- GeeksforGeeks. (2021). Design a Learning System in Machine Learning. GeeksforGeeks. Retrieved from https://www.geeksforgeeks.org/design-a-learning-system-in-machine-learning/
- Logallo, N. (2019, December). Data Science Methodology 101. Towards Data Science. https://towardsdatascience.com/data-science-methodology-101-ce9f0d660336

References

- Pires, S. (2017, April 10). *A very basic introduction to Random Forests using R | Oxford Protein Informatics Group*. Blopig.com. https://www.blopig.com/blog/2017/04/a-very-basic-introduction-to-random-forests-using-r/

- Singh. (2019, November 12). *Linear, Lasso, and Ridge Regression with R*. PluralSight. https://www.pluralsight.com/guides/linear-lasso-and-ridge-regression-with-r

- Kassambara. (2018, November 11). *Penalized Regression Essentials: Ridge, Lasso & Elastic Net*. STHDA. http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/153-penalized-regression-essentials-ridge-lasso-elastic-net/

- Maheshwari, S. (2020, April 21). *Predicting Sales using R programming*. Towards AI. https://pub.towardsai.net/predicting-sales-using-r-programming-84b66d11c35d

Thank You