Contents

583 Project Report

Zhijia Ju, Lili Tang

March 24, 2023

Contents

1 Dataset Introduction and Hypotheses

The dataset we are using examines how the price of Mono-ethylene glycol (MEG) is related to upstream and downstream variables. Originally, there are 29 explanatory variables with a lot of missing data. With the help of the domain knowledge, We transformed them into 8 variables using the weighted method from the aspects of supply,demand,profit and inventory. We will use the new transformed variables in our analysis, including the price of crude oil and coal, upstream MEG profit, MEG profits, MEG operating rate, downstream profits, downstream operating rate, downstream inventory, and MEG inventory. The response variable is the price of MEG. The purpose of this report is to examine how the variables affect the price of MEG, and to predict the price trend using appropriate models.

The hypotheses are as follows: We expect that the price of MEG will be positively related to the price of crude oil and coal, the upstream MEG profit, MEG profit, downstream profits, and downstream operating rate. As the price of crude oil, coal and upstream MEG profit are measures of the cost of MEG, the price will be higher when the cost is higher. Besides, downstream profits and downstream operating rate are all measures of the demand of MEG, and the price will be higher when the demand is higher.As for the MEG profit, which is the overall reflection of supply and demand performance.The better performance(strong demand or short supply),the higher profit, therefore, we expect the price will be higher when MEG profit is higher

We also expect that the price of MEG will be negatively related to the inventory and MEG operating rate, as both of them are measures of the supply of MEG, and the price will be lower when the supply is higher.

2 Description of the Dataset

Variable Name	Unit of Measurement	Continuous vs Discrete
date	YYYY/M/D	Discrete
Price_MEG	CNY(¥)/ton	Continuous
Price_crude_oil_coal	USD($)/barrel	Continuous
Upstream_MEG_Profit	CNY(¥)/ton	Continuous
MEG_Profit	CNY(¥)/ton	Continuous
MEG_operating_rate	Percentage	Continuous
Downstream_Profits	CNY(¥)/ton	Continuous
Downstream_Operating_Rate	Percentage	Continuous
Downstream_Inventory	10,000 Tons	Continuous

Variable Name	Unit of Measurement	Continuous vs Discrete
MEG_Inventory	10,000 Tons	Continuous

Table 1: Description of the variables

As stated from the Table 1 above, the dataset contains 10 variables. The first variable is the date. Price of MEG is the response variable. All the other variables are explanatory variables. The date is a discrete variable, and the rest of the variables are continuous variables. The dataset collected weekly from 2018/1/5 to 2022/11/4, and there are 239 observations in this dataset. The dataset is collected from the Lili's previous employer, and Lili is allowed to use the dataset for this project.

3 Regression Analysis

3.1 Normality check for the response variable and residuals

We conduct a kernel density estimate for Price of MEG using an Epanechnikov kernel. The Epanechnikov kernel is a nonparametric kernel that is used to estimate the probability density function of a random variable.

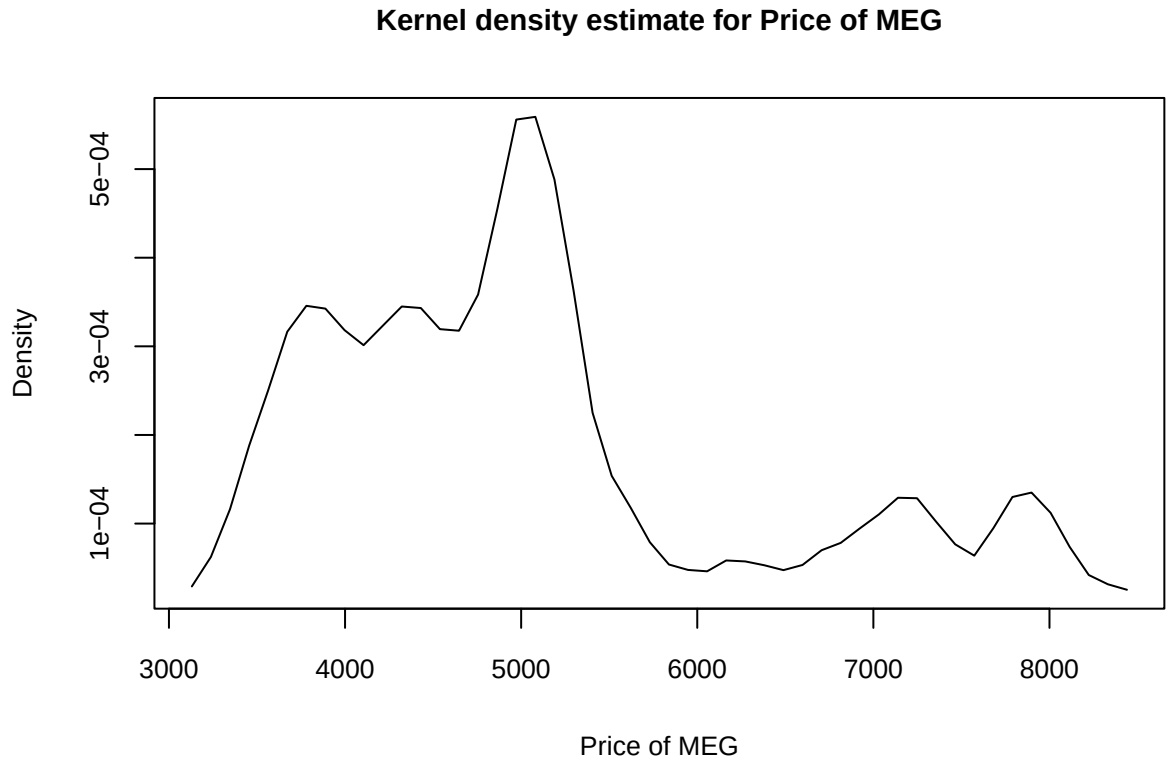

Figure 1: Distribution of the response variable

Figure 1 clearly shows the distribution of the response variable is not normal. We also conduct a Pearson test for normality. With an extremely small p-value, we reject the null hypothesis that the response variable is normally distributed. Therefore, we tried to transform the response variable to make it more normal first.

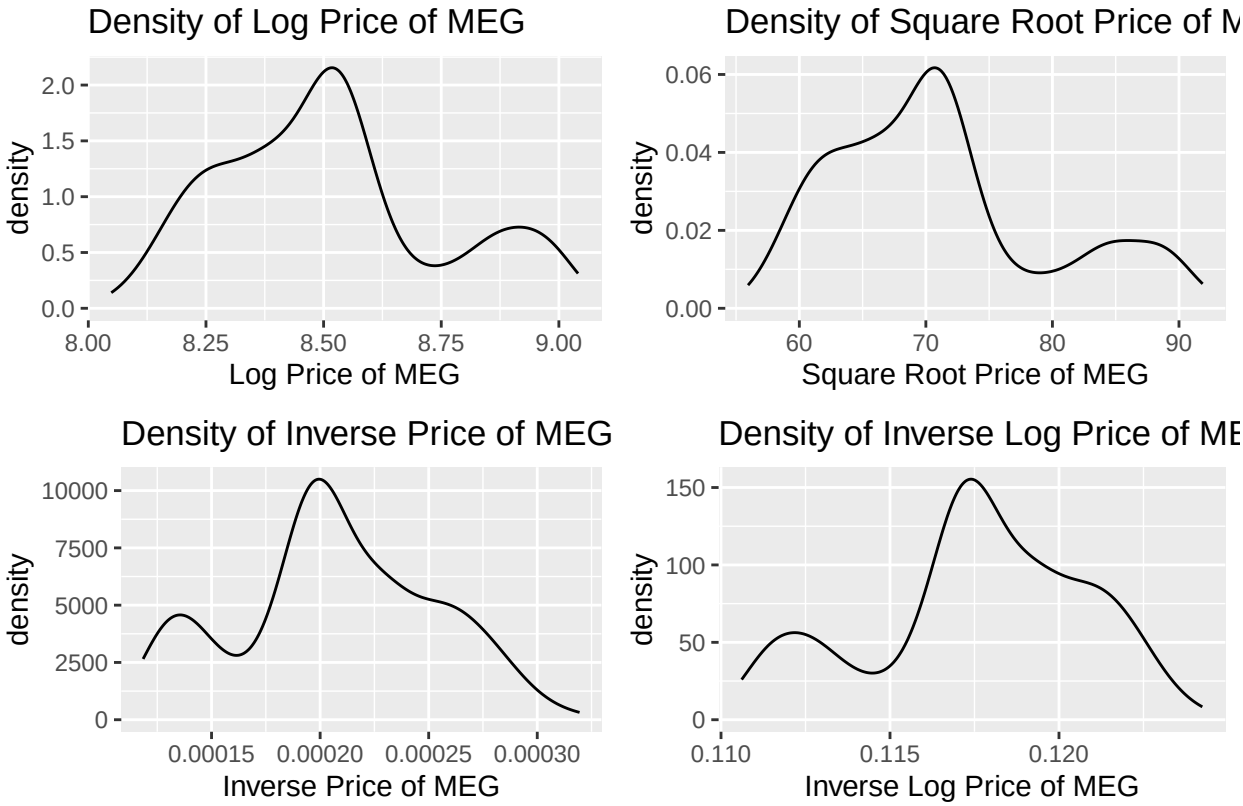

Figure 2: Density plots for different transformations of Price of MEG.

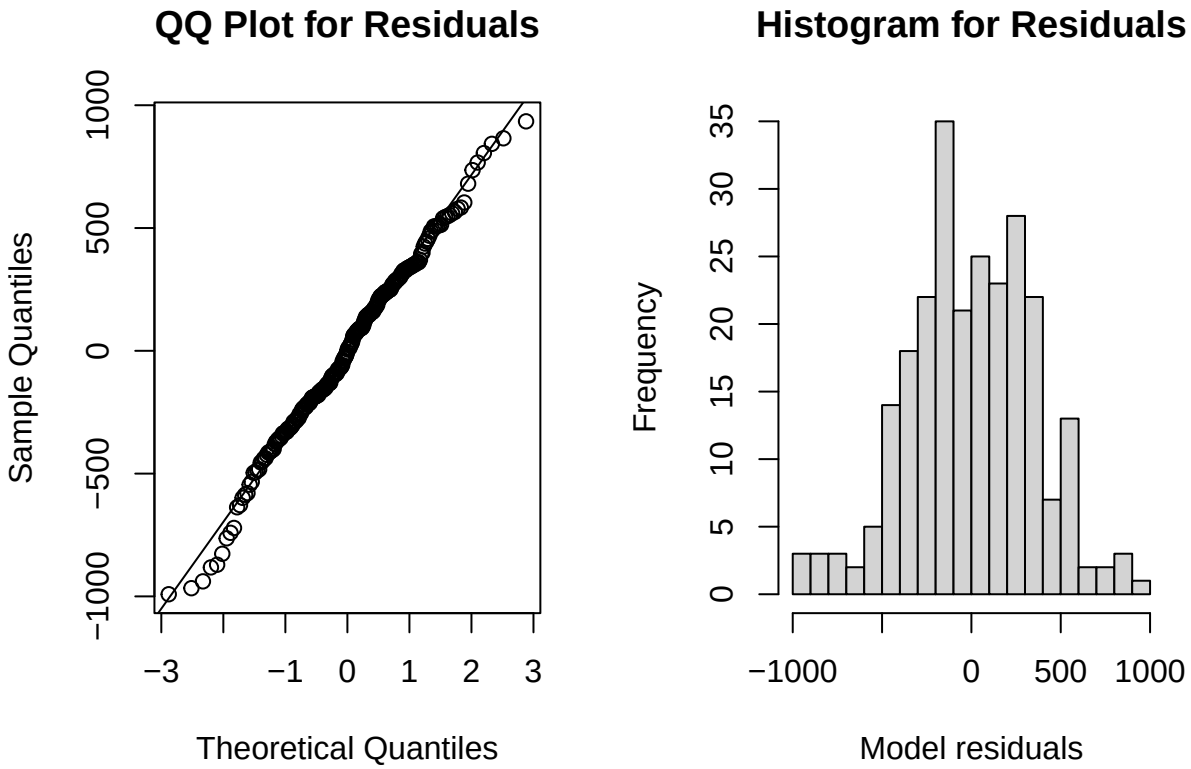

Figure 3: Diagnostics plots

As shown in Figure 2, after using log, square root, inverse transformation, and inverse log transformation, we found that all of them can not handle the skewness and multi-peaks of the response variable. And figure 3 also shows residuals are not normally distributed, so it is not appropriate to use linear regression model and we decided to use nonparametric regression for this dataset first.

3.2 Check the correlation between variables

Before fitting a nonparametric regression model, we need to check the correlation between the response variable and the explanatory variables.

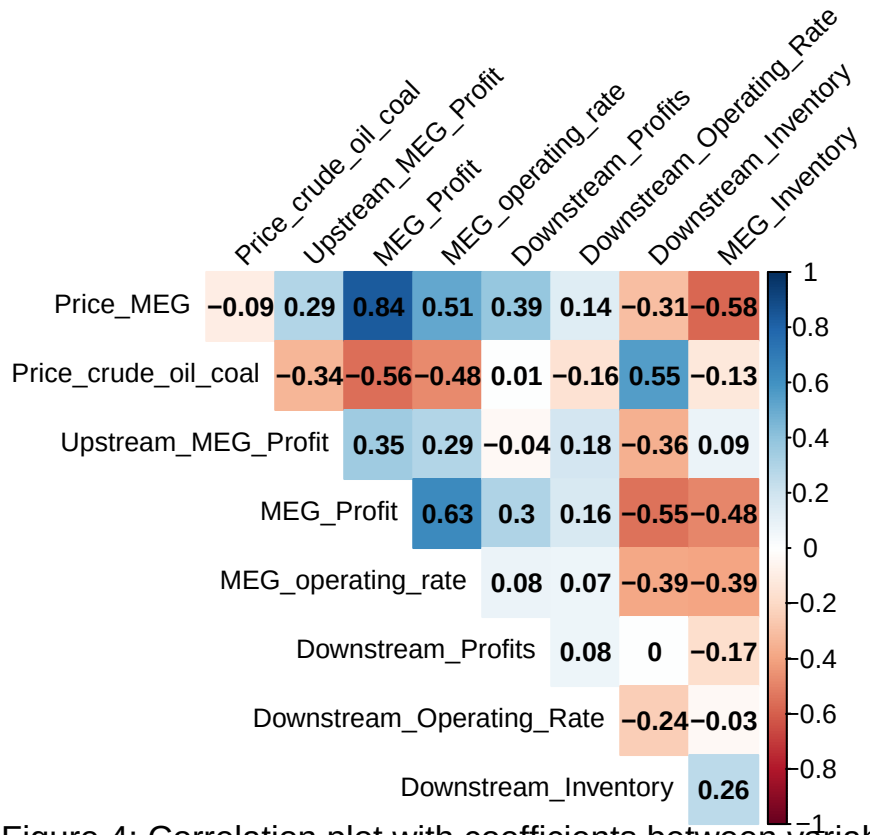

Figure 4: Correlation plot with coefficients between variables

Figure 4 shows that the price of MEG is positively highly correlated with MEG profit and negatively related to MEG inventory, and has almost no correlation with the price of crude oil and coal. The result for MEG profit and MEG inventory aligns with our previous hypotheses. But for price of crude oil and coal is surpising, as we expected the price of MEG will be correlated with it.

3.3 Check multicollinearity

```
##      Price_crude_oil_coal       Upstream_MEG_Profit                MEG_Profit
##                  2.970558                  1.434466                  3.647038
##        MEG_operating_rate        Downstream_Profits Downstream_Operating_Rate
##                  1.968274                  1.231785                  1.086147
##      Downstream_Inventory            MEG_Inventory
##                  1.952275                  2.517280
```

The results show that the VIF values are all smaller than 5. It indicates that the multicollinearity is not a potential big issue in this dataset, but still we need to be careful when interpreting the results.

3.4 Nonparametric local linear regression

After fitting the nonparametric local linear regression model, the summary of the model shows the R-squared is 0.991, which raises the concern of over-fitting. We then performed the consistent nonparametric test of significance, we found that the p-value of the price of crude oil and coal, upstream MEG profit, MEG profit, MEG operating rate, Downstream Profits and Downstream Operating Rate are smaller than 0.05, which means they are significant. All the other variables are not significant according to the test.

3.5 Variable importance

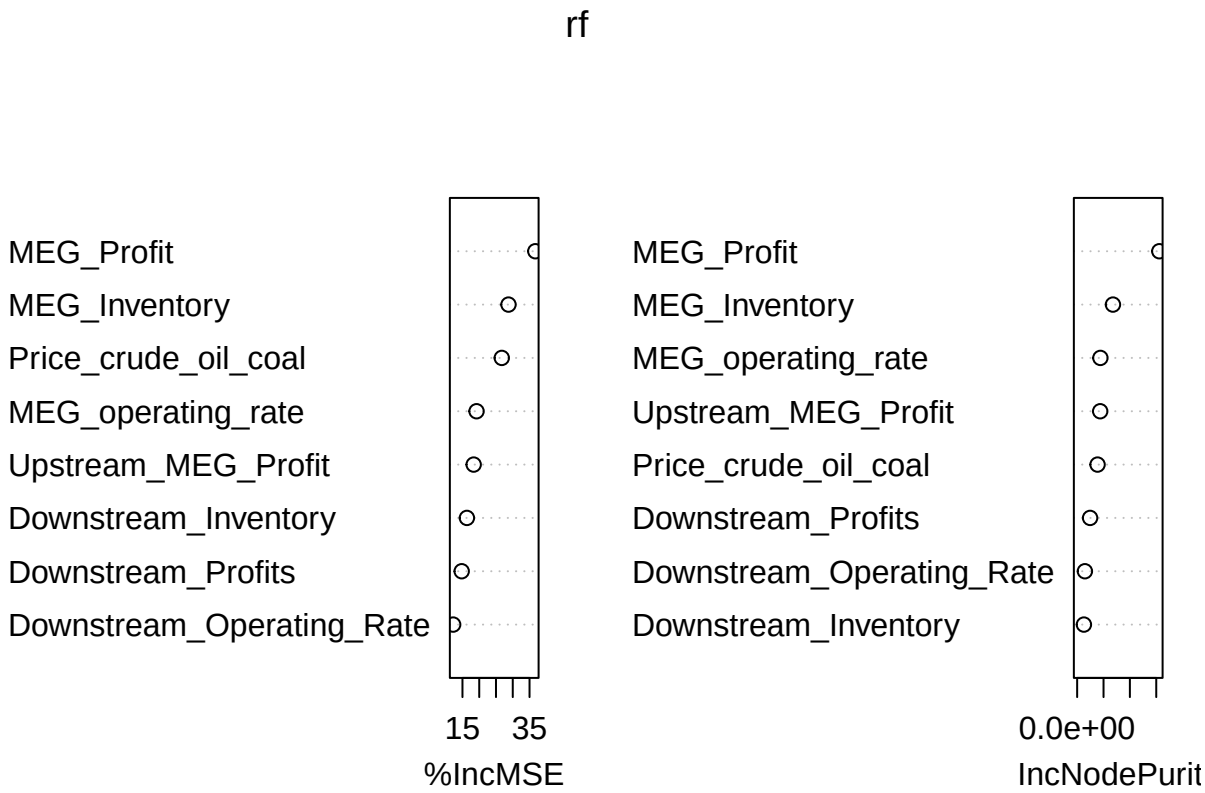

Figure 5: Variable importance plot for random forest

We check the variable importance using random forest and boosting. Both results show that inventory and MEG profit are the 2 most important variables. This results agree with the nonparametric local linear regression model that MEG profit is important variables, but inventory is not significant in the nonparametric local linear regression model.

Overall, for the purpose of variable selection, if we would like to keep 2 variables, we would choose MEG profit and inventory. If we would like to keep 3 variables, we would choose MEG profit, inventory, and the price of crude oil and coal.

3.6 Generalized additive model

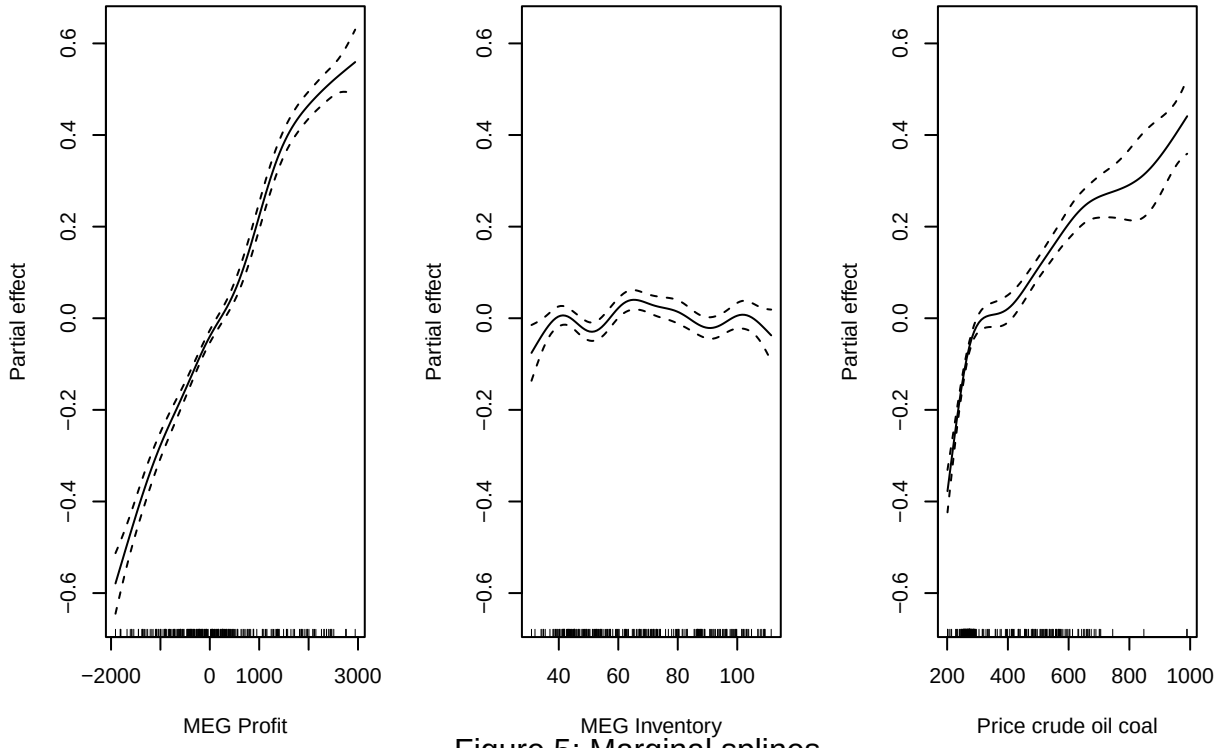

Figure 5: Marginal splines

We fitted a generalized additive model with the Price of MEG as the response variable, and MEG profit,MEG Inventory and price of crude oil and coal as the explanatory variables. The summary of the model shows that the adjusted R-squared is 0.954, which is similar to the nonparametric local linear regression model.

This partial effect plots in Figure 5 can show the relationships between the response variable and the explanatory variable, while accounting for the effect of the other explanatory variable.

The plots show that the Price of MEG increases as the MEG profit and Price crude oil coal increase. However, for MEG Inventory, the overall partial effect is low and the trend is flat. The plots show uncertainty in the fitted smooth function, which means that the relationship between the response variable and the explanatory variable is more complex than a linear relationship. And the estimated degree of freedom for the variables are larger than 3, which means that the relationship between the response variable and the explanatory variables is non-linear.

3.7 Logistic regression

In financial markets, predicting the exact price movement of a particular asset can be a challenging task as it influenced by a complex range of factors. As a result, it can be more practical to focus on predicting the trend of growth or decline, rather than the exact price movement. One approach that could be used to predict the trend of growth or decline in price trend is logistic regression.

The first thing we will need to convert our continuous response variable (MEG price) into a binary variable. This involves assigning a value of 1 to represent growth and 0 to represent decline.Then, after fitting the logistic model, we found that only three variables are significant. In order to improve the model's predictive

power and reduce the risk of overfitting, we use the Akaike Information Criterion (AIC) method to perform variable selection.

By applying the AIC method to the logistic regression model, we can identify that AIC increases from 323 to 317 and the selected five variables(Price_crude_oil_coal,MEG_Profit,MEG_operating_rate,Downstream_Profits and Downstream_Inventory) are all significant.

4 Model Evaluation

After fitting the model, we need to compare which model has the best performance in prediction or classification.

4.1 One-step Cross validation for time series data

When dealing with time series data, it is important to consider the temporal order of the data points. To evaluate the performance of a predictive model, we typically use cross-validation to divide the data into training and testing sets. However, in time series data, we cannot simply divide the data randomly as this would break the temporal order.

Instead, one approach is to use one-step cross-validation, which involves sequentially dividing the data into training and testing sets. To do this, we start by using the first 70% of the data as the initial training set. We then predict the next value in the time series using this training set. And we increase training set by adding one true value and then predict the next value. We repeat this process by using the updated training set to predict the next value and continue this process until used all of the data used.

By using this one-step cross-validation approach, we can evaluate the performance of our predictive model on data that is temporally similar to the data that it will encounter in the future. This can help us to determine how well our model will generalize to new, unseen data.

4.2 MSE comparison between Generalized additive model and random forest

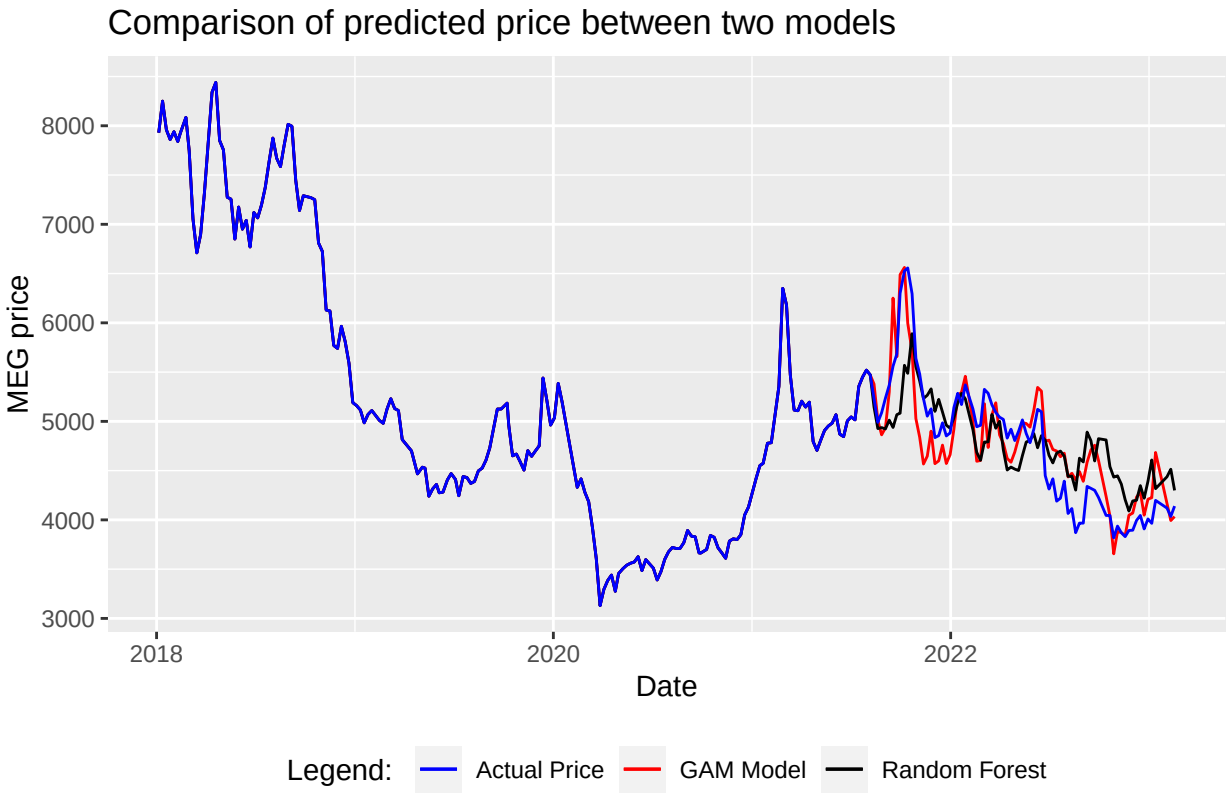

Figure 6: Comparison of GAM Model, Random Forest, and Actual Price

Based on the results of our analysis, we can see from Figure XX that both the generalized additive model and random forest model were able to fit the response variable (MEG price) to some extent. Although the trends in the predicted values are not exactly the same as the true values, they are relatively similar.

To better compare the performance of the two models, we calculated the mean squared error (MSE) between the predicted values and the true values. Our results showed that the generalized additive model had a smaller MSE than the random forest model. This suggests that the generalized additive model was better at accurately predicting the MEG price trend.

In summary, based on both the graph and the MSE results, we can conclude that the generalized additive model outperformed the random forest model in predicting the MEG price trend.

4.3 Logloss evaluation of logistic regression

The logistic regression model we used in our analysis was evaluated using two common metrics, namely the misclassification rate and the log loss. The log loss of the model was found to be 0.63, which falls in the range of 0 to infinity. A lower log loss value indicates better performance of the model. Although the log loss value of our model is not very close to 0, it is still small, which suggests that the model is performing reasonably well.

Furthermore, we found that the misclassification rate of our model is 0.25, meaning that 25% of the predicted outcomes are incorrect. This implies that the model is incorrectly predicting the class of one out of every four observations. However, it is important to note that in financial markets, accurately predicting the direction of the market movement is very challenging, and achieving an accuracy rate of 75% is considered

good. Hence, our logistic regression model's accuracy rate of 75% can be deemed satisfactory for predicting the trend growth or decline in the financial market.

5 Conclusion

```
library(knitr)
library(kableExtra)

table_data <- data.frame(
  Model = c("Nonparametric regression", "", "", "",
            "Random Forest",
            "Boosting",
            "Generalized additive model", "",
            "Logistic regression", "", ""),
  Variables = c("Price_crude_oil_coal; MEG_Profit",
                "Upstream_MEG_Profit; MMEG_operating_rate",
                "Downstream_Profits; Downstream_Operating_Rate",
                "",
                "MEG_Profit; MEG Inventory",
                "MEG_Profit",
                "Price_crude_oil_coal; MEG_Profit; MEG Inventory",
                "",
                "Price_crude_oil_coal; MEG_Profit;",
                "MEG_operating_rate; Downstream_Profits",
                "Downstream_Inventory")
)
table_data
```

```
##                          Model                                          Variables
## 1    Nonparametric regression                Price_crude_oil_coal; MEG_Profit
## 2                                      Upstream_MEG_Profit; MMEG_operating_rate
## 3                                 Downstream_Profits; Downstream_Operating_Rate
## 4
## 5               Random Forest                          MEG_Profit; MEG Inventory
## 6                    Boosting                                           MEG_Profit
## 7  Generalized additive model Price_crude_oil_coal; MEG_Profit; MEG Inventory
## 8
## 9         Logistic regression                Price_crude_oil_coal; MEG_Profit;
## 10                                      MEG_operating_rate; Downstream_Profits
## 11                                                          Downstream_Inventory
```

5.1 Interpretation between predicators and response variables

Model	Significant variables or variable importance
Nonparametric regression	Price_crude_oil_coal; MEG_Profit Upstream_MEG_Profit; MMEG_operating_rate Downstream_Profits; Downstream_Operating_Rate
Random Forest	MEG_Profit;MEG Inventory

Model	Significant variables or variable importance
Boosting	MEG_Profit
Generalized additive model	Price_crude_oil_coal;MEG_Profit;MEG Inventory
Logistic regression	Price_crude_oil_coal; MEG_Profit; MEG_operating_rate; Downstream_Profits Downstream_Inventory

The figure indicates that across five different models, two variables, namely "Price_crude_oil_coal" and "MEG_Profit", consistently show significance or importance. Moreover, the variable "MEG Inventory" displays significance or importance in three models, namely "Random Forest", "Boosting", and "GAM". Notably, the "GAM" model with significant three variables, namely "Price_crude_oil_coal", "MEG_Profit", and "MEG Inventory", has a large R-square of 0.95. This indicates that these three variables have a substantial impact on the MEG price.

The statistical results also confirms our previous hypothesis from the fundamental perspective: 1.As the price of crude oil, coal is a measure of the cost of MEG, the price will be higher when the cost is higher. 2.we expect the price will be higher when MEG profit is higher. As for the MEG profit, which is the overall reflection of supply and demand performance.The better performance(strong demand or short supply),the higher profit, therefore, we expect the price will be higher when MEG profit is highers for the MEG profit, which is the overall reflection of supply and demand performance.The better performance(strong demand or short supply),the higher profit. 3.We also expect that the price of MEG will be negatively related to the inventory and MEG_operating_rate, as both of them are measures of the supply of MEG, and the price will be lower when the supply is higher.

The previous correlation plot shows that the price of MEG is positively highly correlated with MEG profit and negatively related to MEG inventory.

Therefore, based on the statistical analysis of the models and variables, we can conclude that the hypothesis based on the fundamental perspective has been supported.

5.2 Prediction of response variables(Model Selection)

When response variable is continuous When evaluating different models for our dataset, we found that linear regression was not an appropriate choice because the response variable and residuals were not normally distributed. Nonparametric local linear regression, while it has a high R-square value close to 0.99, raised concerns about overfitting and has some insignificant variables. Furthermore, it required a large dataset.

In comparison, the random forest model has a larger Mean Squared Error (MSE) than the Generalized Additive Model (GAM), which was our preferred choice. The GAM had a smaller MSE, all significant variables, and a large R-square value. Moreover, the GAM was more transparent and easier to interpret than the black box nature of the random forest model.

Therefore, we can conclude that the Generalized Additive Model is the best model for our dataset, as it had strong predictive power, all significant variables, and good interpretability.

When response variable is binary Predicting the exact movement of financial markets can be very difficult, so it may be more useful to focus on predicting the general trend of growth or decline.

Our logistic model has an misclassification rate of 0.25 and a log loss of 0.63, which is not considered to be a strong model. However, given the inherent unpredictability of financial markets, this model may still have some value. In summary, while our model may not be optimal, it can still be considered a reasonable approach for predicting market trends.