Data science - test task

Tomas Dzedulionis

4/16/24

1. Task Description

Develop a model for each of the four items that predicts the variable ZG (column A) for the period 01.08.21 - 23.02.22.

2. Modelling

Bayesian Additive Regression Trees (BART) has been chosen as the modeling technique for this task. BART offers a flexible approach to fitting regression models without imposing strict parametric assumptions. In BART, an ensemble of trees is constructed using a back-fitting algorithm. Initially, a small tree is fitted to the data, and then subsequent trees are built to capture the residuals iteratively.

One of the strengths of BART lies in its use of intelligent priors, which allow the model to systematically learn the appropriate amount of shrinkage and the depth of the trees. This feature helps prevent overfitting and ensures that the true relationships between variables are accurately captured. Additionally, BART has the capability to handle missing data without requiring imputation.

The optimal parameters for each model are determined through grid search, and model performance is assessed using 10-fold cross-validation. This approach ensures robustness and generalizability of the models.

2.1. Data pre - processing

The dataset was partitioned based on the article_no variable, resulting in four distinct models. To enhance the predictive power and account for calendar effects, additional regressors were incorporated by mapping columns representing month, day of the week, and holidays onto the dataframe.

In preparation for modeling, the values of the zg column for each model were transformed using a square-root transformation. This transformation helps in stabilizing variance and normalizing the distribution, thereby rendering the data more suitable to statistical analysis and modeling.

Moreover, potential outliers were identified using the Interquartile Range (IQR) method. Outliers were determined by calculating the difference between the 75th and 25th percentiles (the IQR), and any values falling below the 25th percentile minus 1.5 times the IQR, or above the 75th percentile plus 1.5 times the IQR, were flagged as outliers.

$$\text{IQR} = Q_3 - Q_1$$

2. Exploratory Data Analysis (EDA)

Second step was to Exploratory Data Analysis (EDA).

2.1. *Descriptive statistics*

Model nr.1 - Article no. 1294:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      0    2484    3190    3156    3926   16296     207
```

The sales data for Article No. 1294 exhibits a mean of 3156.32 and a median of 3189.70, with values ranging from 0 to 16296.38. This item also contains 207 missing values in the dataset.

Model nr.2 - Article no. 1782:

```
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
   44.86  14291.82  18937.87  28137.29  43650.64 152838.17       207
```

For Article No. 1782, the sales data demonstrates a notably higher mean of 28137.29 and a median of 18937.87. The range extends from 44.86 to 152838.17, and similarly, there are 207 missing values.

Model nr.3 - Article no. 89450:

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
   0.00     0.00     3.63  3159.42  2442.58 48316.92      207
```

Article No. 89450 showcases distinct sales characteristics, with a mean of 3159.42 and a median of only 3.63. The range spans from 0 to 48316.92, and like the others, it contains 207 missing values.

Model nr.4 - Article no. 89479:

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
   35.21  3266.78  9106.55 13774.05 22768.21 58284.57      207
```

Lastly, Article No. 89479 indicates a mean sales of 13774.05 and a median of 9106.55, with values ranging from 35.21 to 58284.57. Similarly, it contains 207 missing values.

2.2. Histograms

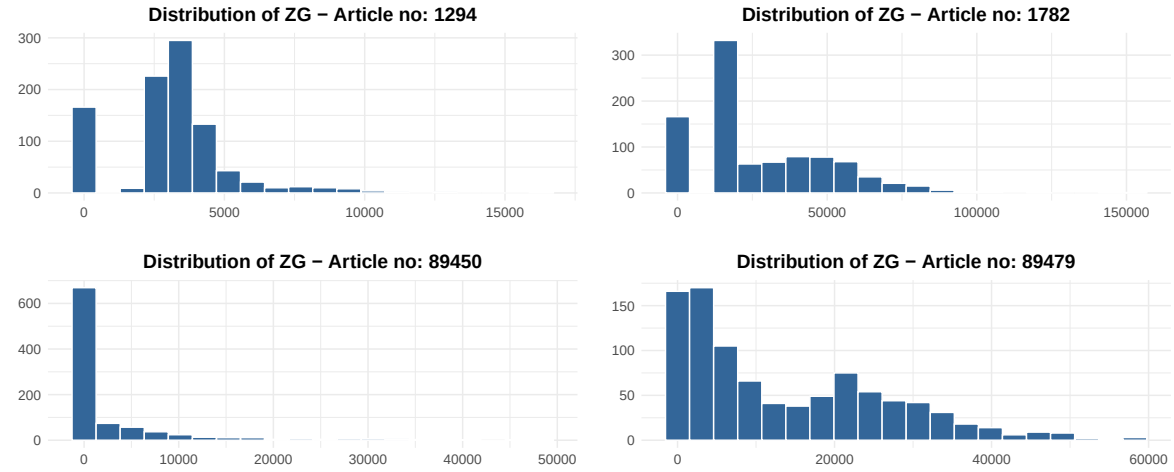


Figure 1: Distribution plots

The distributions of ZG values across the different article numbers exhibit a range of patterns, from right-skewed to bimodal, suggesting varying characteristics in the underlying data for each article.

2.3. Trend plots

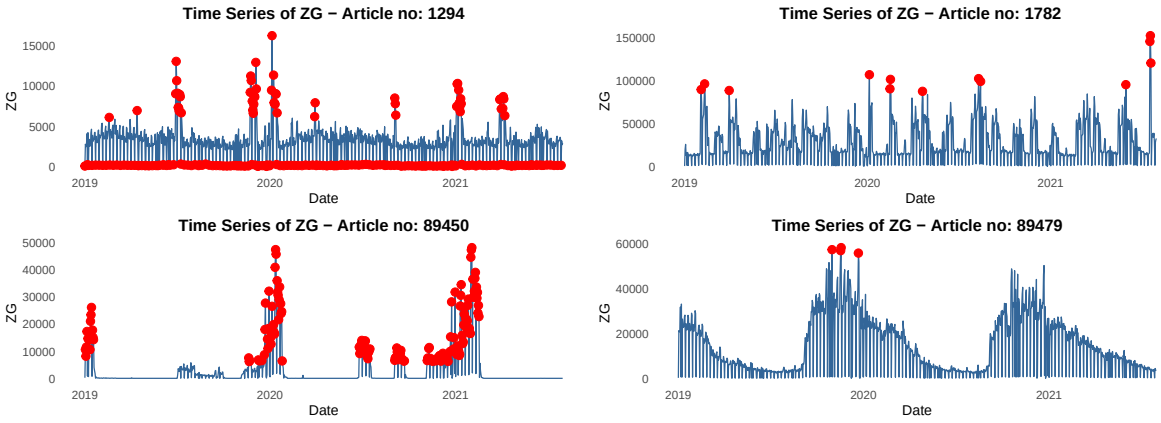


Figure 2: Trend plots

Red dots indicate previously detected outliers. For article no. 1294, the time series exhibits a volatile and fluctuating pattern, with several spikes in the ZG values, particularly in 2020. The red dots represent the outliers identified using the IQR method, indicating the presence of unusual or extreme ZG values throughout the time series.

The time series for article no. 1782 also shows a volatile trend, with multiple peaks and troughs in the ZG values. The red dots highlighting the outliers suggest that the dataset contains several extreme or atypical data points that may warrant further investigation.

Similarly, the time series for article no. 89450 demonstrates a fluctuating pattern, with the presence of multiple outliers, as indicated by the red dots, suggesting the potential existence of unusual ZG values within the data.

Lastly, the time series for article no. 89479 exhibits a more pronounced and sustained peak in the ZG values, followed by a gradual decline. The red dots again highlight the outliers, which appear to be concentrated in the higher value range of the time series. It seems that the highest peaks are around the start of each new year, thus the inclusion of months variable should help.

After further inspecting the plots, it seems that outliers may not be removed as the volatile and spiky nature of the ZG values, with frequent occurrences of the red dot outliers, suggests that these high or low values are not necessarily incorrect, but rather reflect the underlying dynamics of the data.

2.4. Weekdays effects plots

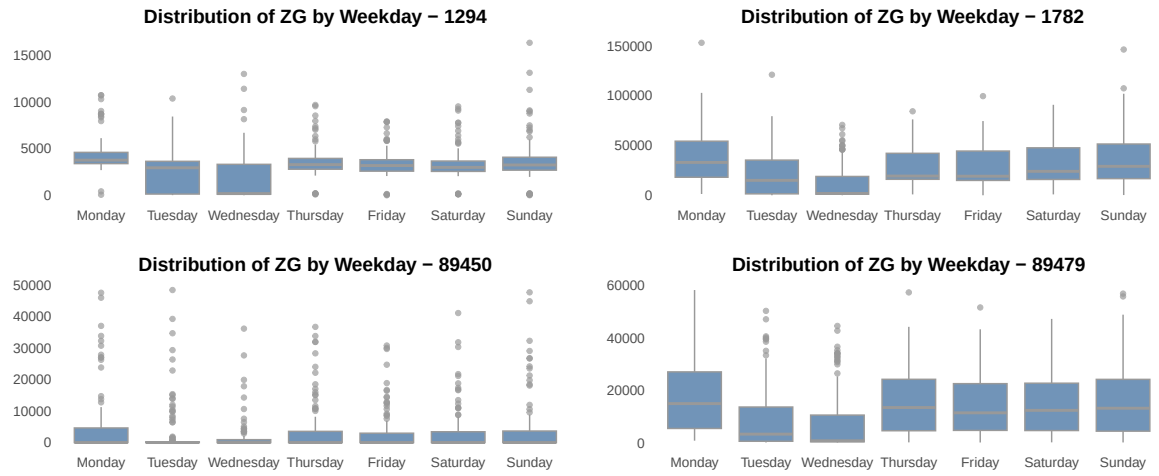


Figure 3: Weekday effects

The effects of weekdays were visualized using boxplots.

The distribution of ZG values across different weekdays shows some interesting patterns:

For article no. 1294, the ZG values tend to be higher on Wednesdays and Thursdays, with a few outliers on other days. This suggests that there may be specific factors or processes driving higher ZG values during the middle of the week for this particular article.

The pattern is similar for article no. 1782, where the ZG values are generally higher on Wednesdays and Thursdays, with some elevated values on other weekdays as well.

Article no. 89450 exhibits a more uniform distribution of ZG values across the weekdays, without any clear spikes or patterns on specific days.

Finally, for article no. 89479, the ZG values appear to be consistently higher on Mondays, with a more dispersed distribution throughout the rest of the week.

2.5. Promo plots

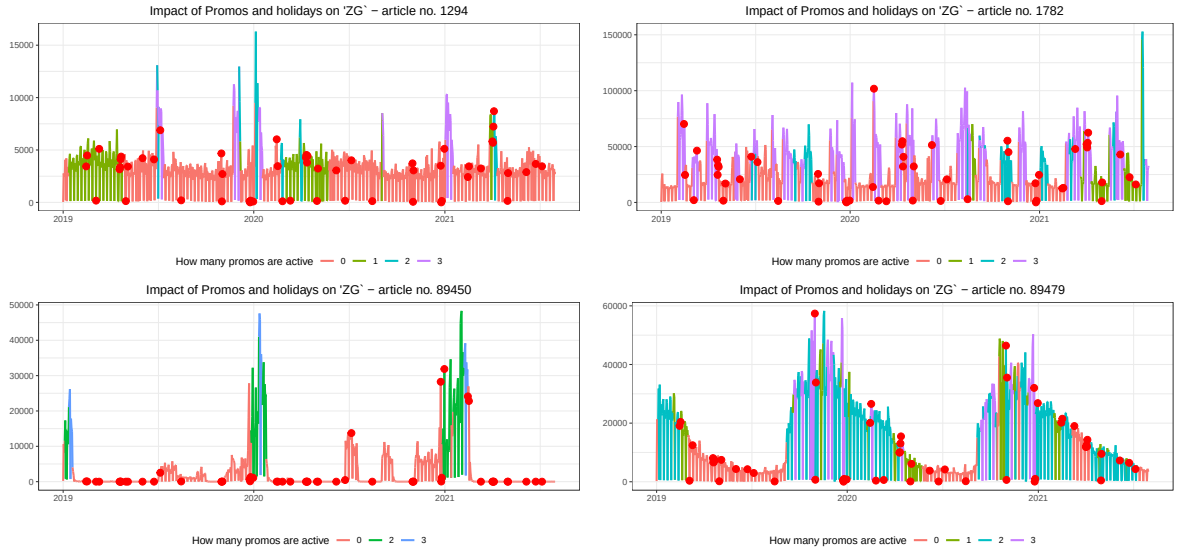


Figure 4: Promos and Holidays effects

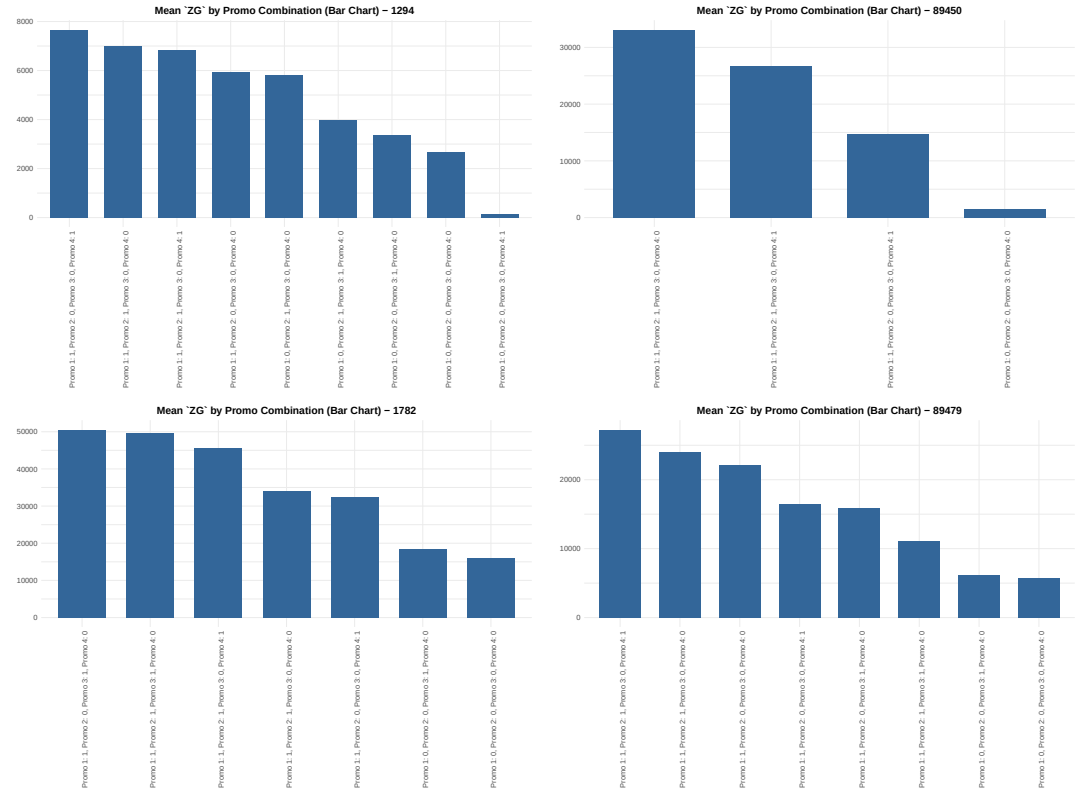
The histograms shows the impact of various combinations of promos, while the time series plots provide a comprehensive view of how promotional activities and holidays impact the 'ZG' metric across different article numbers.

For article no. 1294, the 'ZG' values exhibit pronounced fluctuations, often coinciding with spikes in the number of active promotions. The influence of holidays is evident, as some of the major 'ZG' peaks appear to align with holiday periods.

A similar dynamic is observed for article no. 1782, where the 'ZG' values demonstrate substantial variations, frequently corresponding to changes in promotional intensity. The impact of holidays is also visible, though not as pronounced as in the previous case.

The pattern for article no. 89450 is somewhat different, with a less volatile 'ZG' trend. While some of the 'ZG' increases correlate with promotional activities, there are also periods where the 'ZG' values rise without a direct link to promotions, particularly in the latter half of the year.

For article no. 89479, the 'ZG' time series exhibits a complex and dynamic behavior, with sharp spikes and troughs. These fluctuations appear closely tied to the variations in the number of active promotions, which tend to be concentrated in the latter part of the year and carry over into the following year. The influence of holidays is also evident in this case.



3. BART - Modelling

Once the optimal parameters for the Bayesian Additive Regression Trees model are determined through grid search and the models are constructed, they are saved into the 'logs' folder. This practice helps save time when running the code in the future.

Following the construction of the models, convergence diagnostics tests and error assumption tests are conducted to ensure the models are reliable and accurate. Additionally, the model fit is assessed by comparing modeled versus actual values, and variables importance tests are generated to identify the most influential predictors in the model.

Overall, the models perform well in capturing the relationship between the predictors and the target variable. However, it's noted that three of the models (1st, 2nd, and 4th) occasionally tend to overestimate lower values.

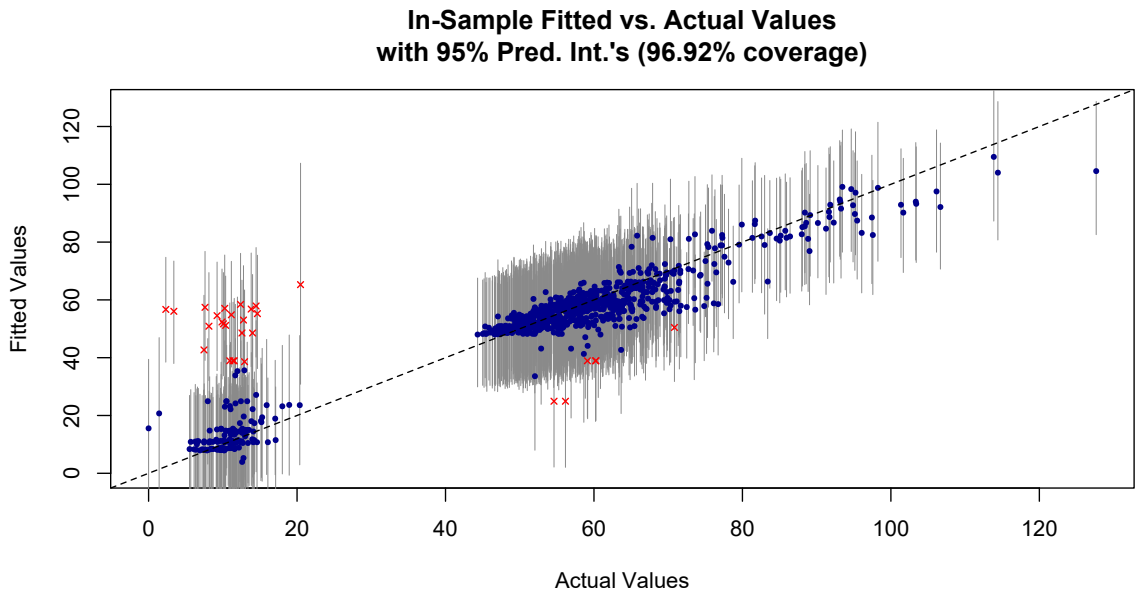


Figure 5: 1st model

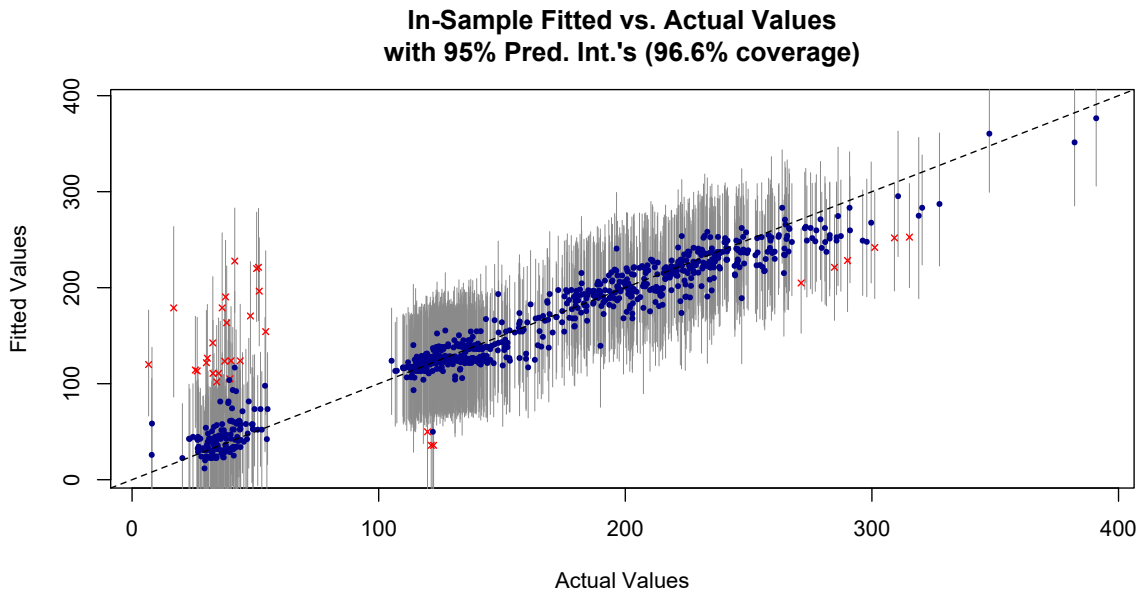


Figure 6: 2nd model

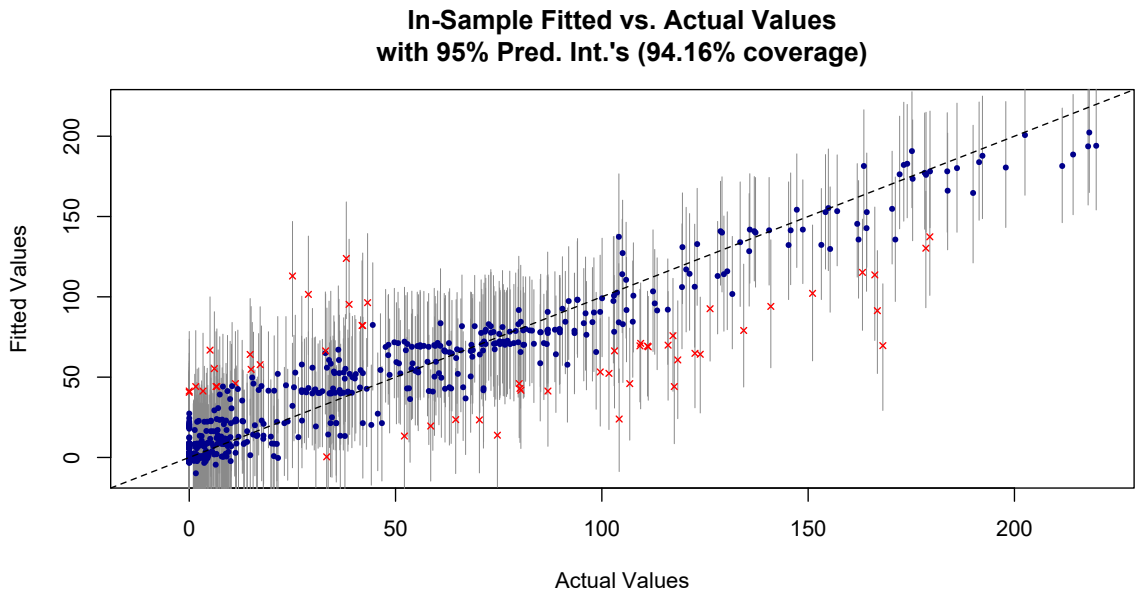


Figure 7: 3rd model

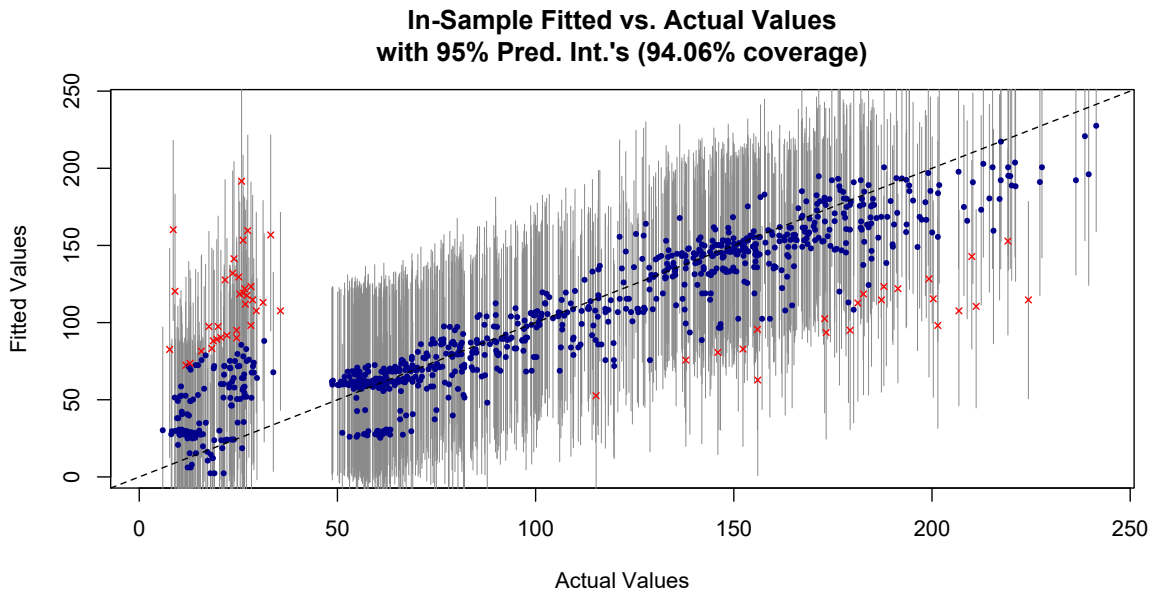
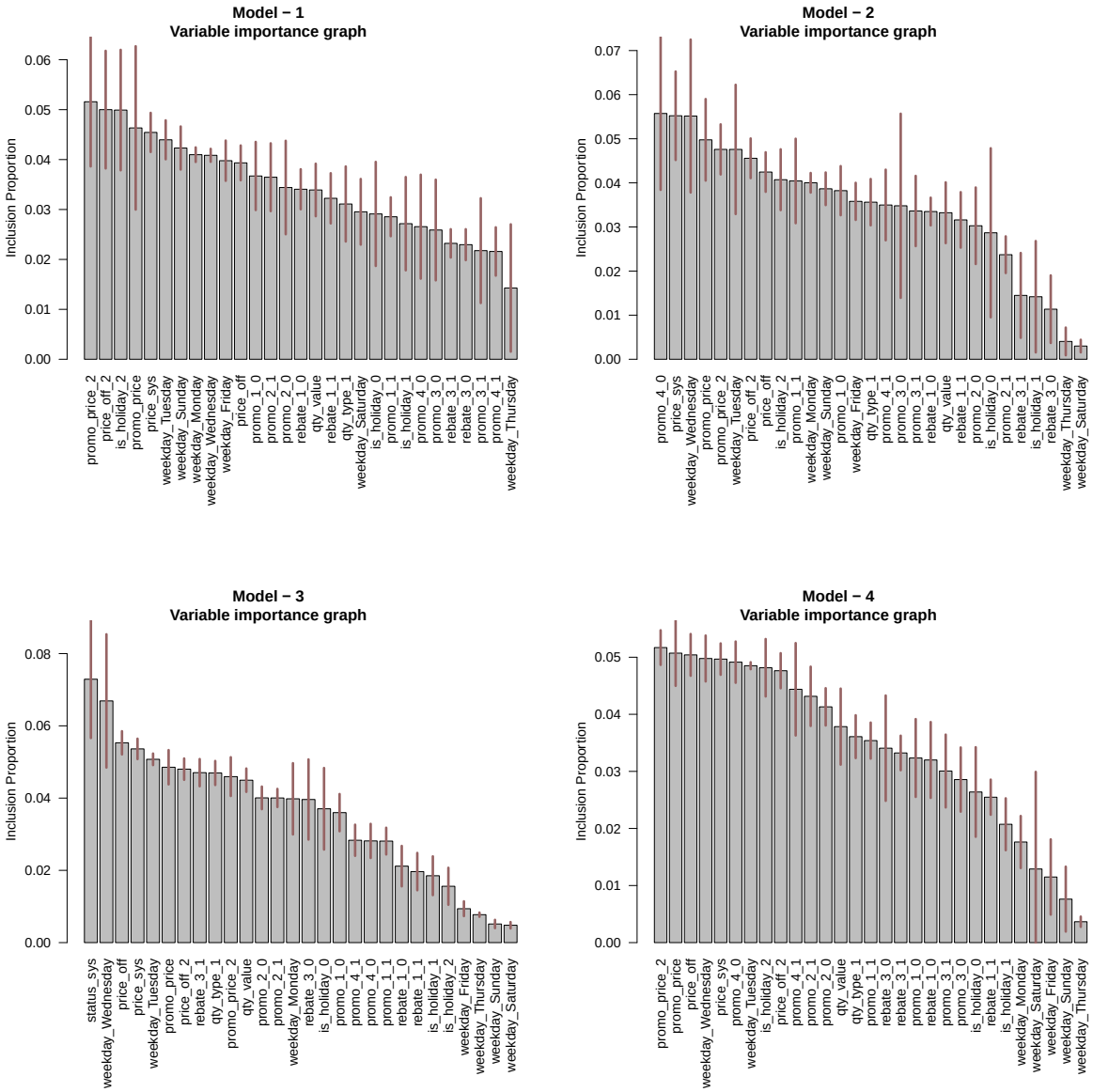
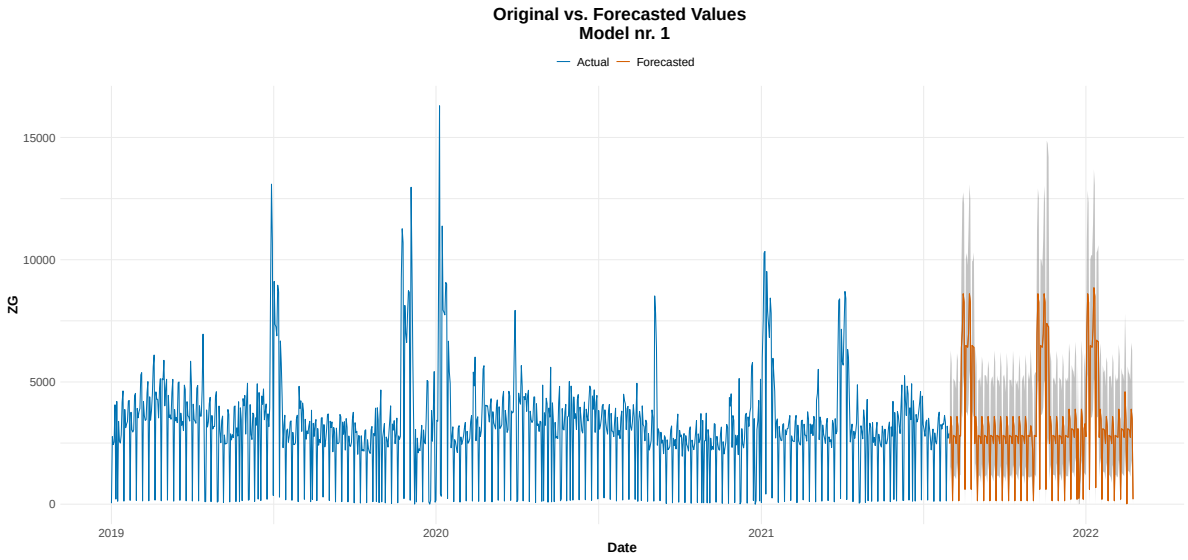


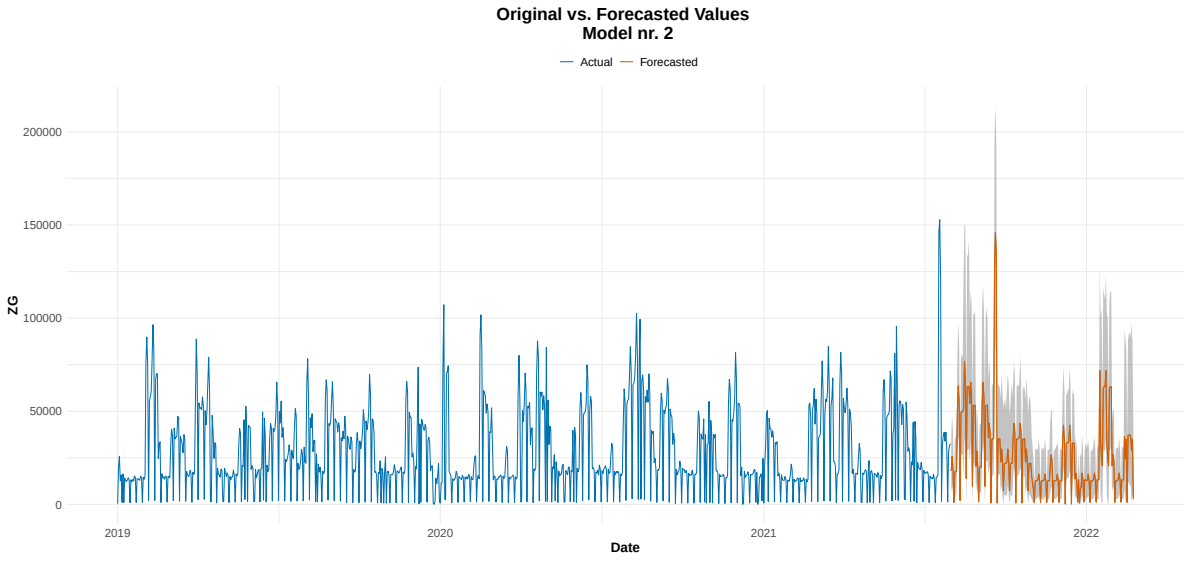
Figure 8: 4th model

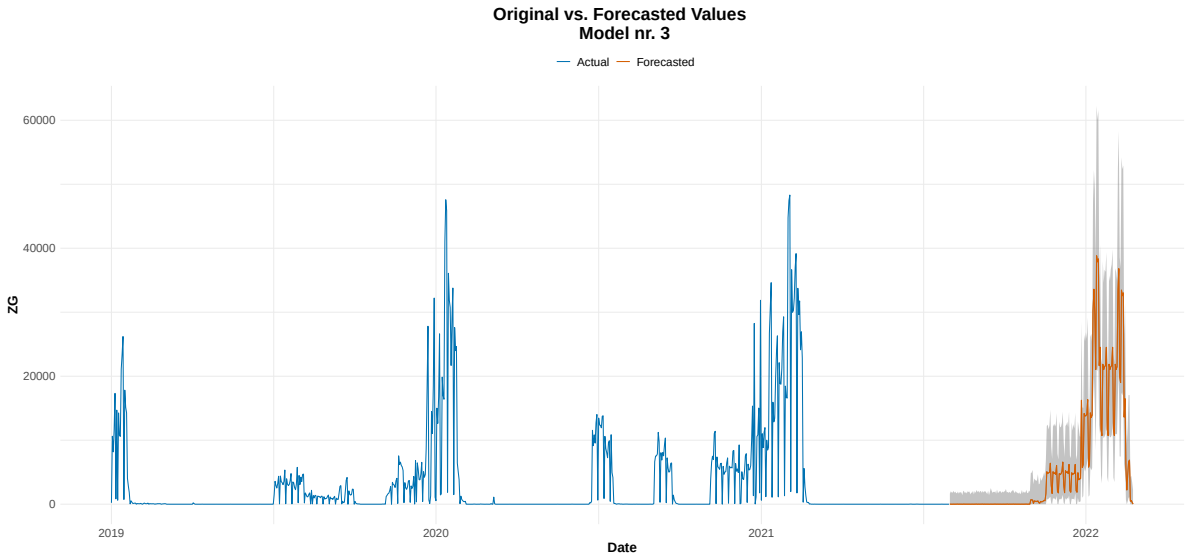


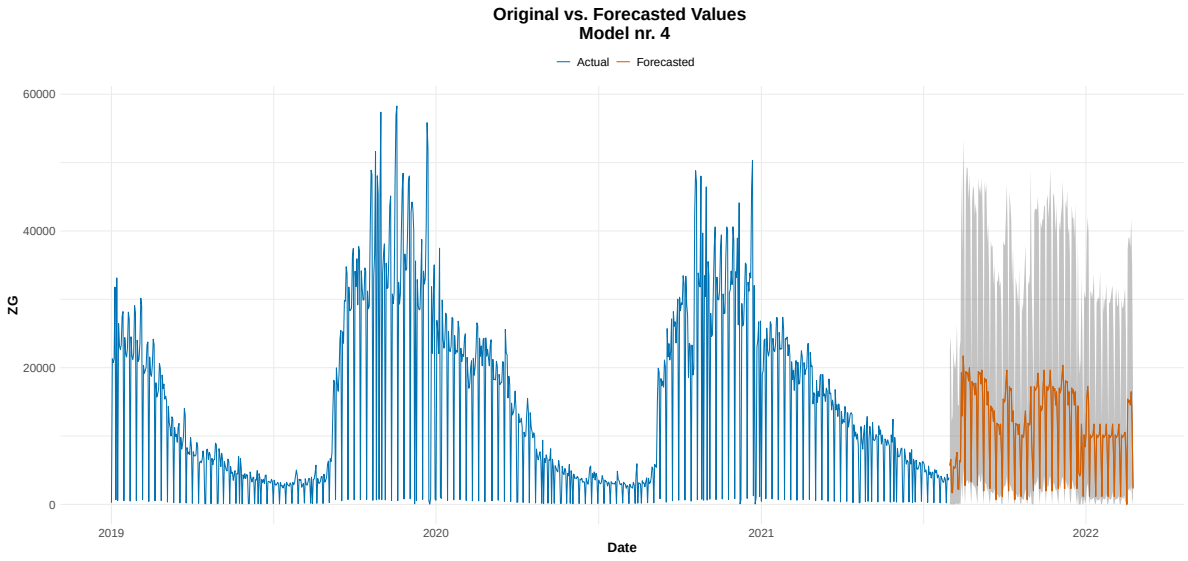
4. Predictions

The final step involves generating predictions for each of the models, which are visualized in plots. In these plots, the orange line represents the predictions, while the grey background illustrates the 95% confidence interval.


Original vs. Forecasted Values
Model nr. 1
Actual
Forecasted
15000
10000
5000
0
ZG
2019
2020
2021
2022
Date


Original vs. Forecasted Values
Model nr. 2
Actual
Forecasted
200000
150000
100000
50000
0
ZG
2019
2020
2021
2022
Date


Original vs. Forecasted Values
Model nr. 3
Actual
Forecasted
60000
40000
20000
0
ZG
2019
2020
2021
2022
Date


Original vs. Forecasted Values
Model nr. 4
Actual
Forecasted
60000
40000
20000
0
ZG
2019
2020
2021
2022
Date