TECHNISCHE UNIVERSITÄT DRESDEN

Big Data Analytics in Transportation @TU Dresden

"Friedrich List" Faculty of Transport and Traffic Sciences Institute of Transport and Economics

Chair of Big Data Analytics in Transportation

Project thesis on subject
Application of Data Analytics

Modeling Gas Prices

Jiujiu Liao

Matriculation number: 5195367

Kirill Salnikov

Matriculation number: 5179909

Tim Stolz

Matriculation number: 5179877

27th August 2024

Supervisor

M.Sc. Lennart Schäpermeier

Supervising professor

Prof. Dr. rer. pol. Pascal Kerschke

Contents

List of Figures

1 Introduction

Fuel retail is a very interesting example of a market with high dynamics of changes. Price shifts on it occur several times during an hour for a majority of gas stations, which means its frequency is more comparable to the stock exchange, than to ordinary consumer or business oriented markets. This high elasticity stems from huge and relatively persistent demand combined with ease of price modifications due to a small number of products. Usually only E5, E10, and Diesel fuels are sold on regular gas stations in Germany.

A competitive market with a large number of brands but a small number of similar items leads to the need to outmaneuver opponents by very carefully adjusting prices to the market conditions of the moment. The interesting thing about this ecosystem is that this adjustment is made relatively independently at different geographic locations.

All together this creates a situation where data analysis methods can be used not only to understand or explain some patterns, but also to predict future behaviour of brands on various unique stations, since a relatively consistent, standardized data generating process is available. This can be vital information for competitors to estimate future price trends of their rivals, as a tiny difference in price can make a huge difference in the bottom-line profitability, due to the high volume of sales. For costumers, especially large logistics companies, information about expected price levels of fuel can also be valuable for route planning and profitability estimations.

Speaking more generally, finding the laws that make up the foundation of fuel market behaviour is significant, because gasoline prices impact the majority of all industries and therefore also most consumption goods. This could allow for better efficiency in the placement of new gas stations, as well as improving the planning of better refinery logistics, possibly benefiting the welfare of entire regions. These tasks are not directly linked to price determination on a given gas station, but investigating them lays a foundation for further research in the topic.

2 Background and literature review

During the last couple of years a lot of research has been done in order to explain the price setting behaviour of fuel brands within their networks of gas stations. Recognizing the various aspects of the importance of the topic, in this paper we focus on pricing strategies at individual gas stations, an understanding of which can highlight valuable details of fuel brands competition.

One of the key factors influencing the strategy of any economic agent is the real or expected competition with other agents. But the influence of competition on prices on the gasoline market, as can be discovered from literature, is not trivial, in fact it is complex to determine, because it tends to be heavily correlated with other factors.

A new study, exploring price patterns in five Italian cities (Fronzetti Colladon et al. (2024)), found that in rich areas the density of gas stations and therefore competition between brands can be higher, but the price is above average too, because the demand is relatively high and customers are wealthy. Besides that spatial zoning and regulation can have a big impact on where stations can be located (usually it is hard to place a gas station in the city center) and therefore substantially distort the balance between supply and demand in a particular area. On the other hand the same paper suggests that on areas with low competition and only few separate gas stations the prices are also higher than average, which fits good basic economic reasoning. Below average prices could be found on areas with average density of stations or poorer than average neighborhoods.

Modelling done by Pennerstorfer and Weiss (2013) found out that it is hard to determine effect of spatial competition on prices, since there are different patterns between individual gas stations and "clusters" of them, which in turn also needs to be determined. Calculations made by authors show that geographical "clustering" of gasoline stations leads to fall of competition between firms and therefore increases the average price. Thus, stations do not compete for the same customers with their immediate neighbors as much as they do with stations located further away. Following this reasoning, lower average prices can stem from competition of different clusters. "Merger simulations show that ignoring merger-induced changes in spatial characteristics will lead to a significant bias in the evaluation of merger effects." (Pennerstorfer and Weiss (2013))

These two pieces of research point to the application of machine learning to make predictions in this field being quite feasible. The expectation is that automated algorithms will be able to find patterns, which are too complex to distinguish and describe by people, without deep div-

ing into hundreds of different local situations.

The approach taken by Haucap, Heimeshoff, and Siekmann (2015) is more focused on brand competition for customers globally within a region. They noted that larger and well-known brands charge more for their fuel. This behaviour, of course, occur in a lot of consumer goods markets, however in this case the commodity (fuel) is uniform and its quality is relatively independent of brand. Locations of gas stations and their advertising were found to be related to socio-demographic characteristics. Thus, a brand can develop its marketing strategy and set different prices in different neighborhoods, although it sells exactly the same fuel. Besides that the study introduces two other major factors influencing final prices: distance to the closest high level road and refinery price.

The study by Kucher, Burnett, and Lacombe (2018) explains the interdependence of gasoline prices among neighbouring large areas in one country. That justifies the influence of wholesale gasoline prices, refinery prices on local retail. For our research (in which we want to explore local market tendencies), since we do not have data on refinery prices, it is important to try to substitute it with other parameters (we chose to include crude oil price from the stock market) and chose a period when no major shifts on the oil market happened. This also means that temporal factors play a significant role, because the market can be volatile and change trends quickly.

3 Data gathering and aggregation

For German gas prices a very large data-set is available, although it contains a modest amount of variables. It covers all the country with approximately 17000 gas stations and stores all gas price changes that occurred during the last 10 years with many of them just a few seconds apart. It is divided into 2 parts, each of them split up into daily csv files: station information and price level information for every time the price was changed. Thus, there are 2 fundamental "dimensions" of our data: one spatial, the other temporal. The general workflow to prepare this data for machine learning looked as follows:

1. Choose the subset of data to work with
   Data selection decisions were made by testing the machine learning procedure on different subsets of data. The limitations described here are dictated by constraints in computational resources, but the procedure in section "4 Predicting Prices" can be scaled up for larger amounts of data. The selected data set is restricted by:
   - State of Saxony. The region is neither uniform nor monocentric, allowing for analysis of various combinations of spatial aspects together.
   - February 13th to 15th (Tuesday to Thursday). First two of which are used for training and the third for the final test of the predictions on "unobserved" data.
   - Fuel type E5. Since observed patterns and data structure for other fuel types is very similar it makes it easy to extend the applied methods to other types of fuel if desired.
   - Four largest brands: ARAL, TotalEnergies, Shell, STAR. Which was applied only on the stage of final data selection for model training and predictions, meaning spatial interactions with other brands are estimated and preserved.
2. Extract the coordinates of stations and map them.
3. Gather (create) spatial variables which could have predictive power on price, based on background research and literature review (e.g. distance to the city, road or another competitor gas station).
4. Preprocess non-spatial data on the station dataset (e.g. brand name matching).
5. Preprocess price dataset. Converting the arbitrary set of price points into a regular series of the average price during every hour based on the amount of time they were active.
6. Merge the two datasets, creating a layer containing all the information about which station has which price in a given time interval.
7. Clean the data, remove lines containing controversial and implausible information. Manually check the data adequacy using cartographic services.

Details, concerning spatial and temporal “dimensions” of data are discussed below.

Spatial dimension
In terms of proximity and fuel retail service coverage the state of Saxony is non-homogeneous. Denser clusters of gas stations are located within three cities: Dresden, Leipzig, Chemnitz. In Chemnitz the highest density is observed in city center, while in Dresden and Leipzig there are more stations in peripheral neighbourhoods. There are some other areas which possess surprisingly tight clusters, such as the western counties of Zwickau and the Vogtlandkreis. Therefore, to distinguish between different locations, we introduced variables of county, population density, settlement type, and distance to the nearest city.

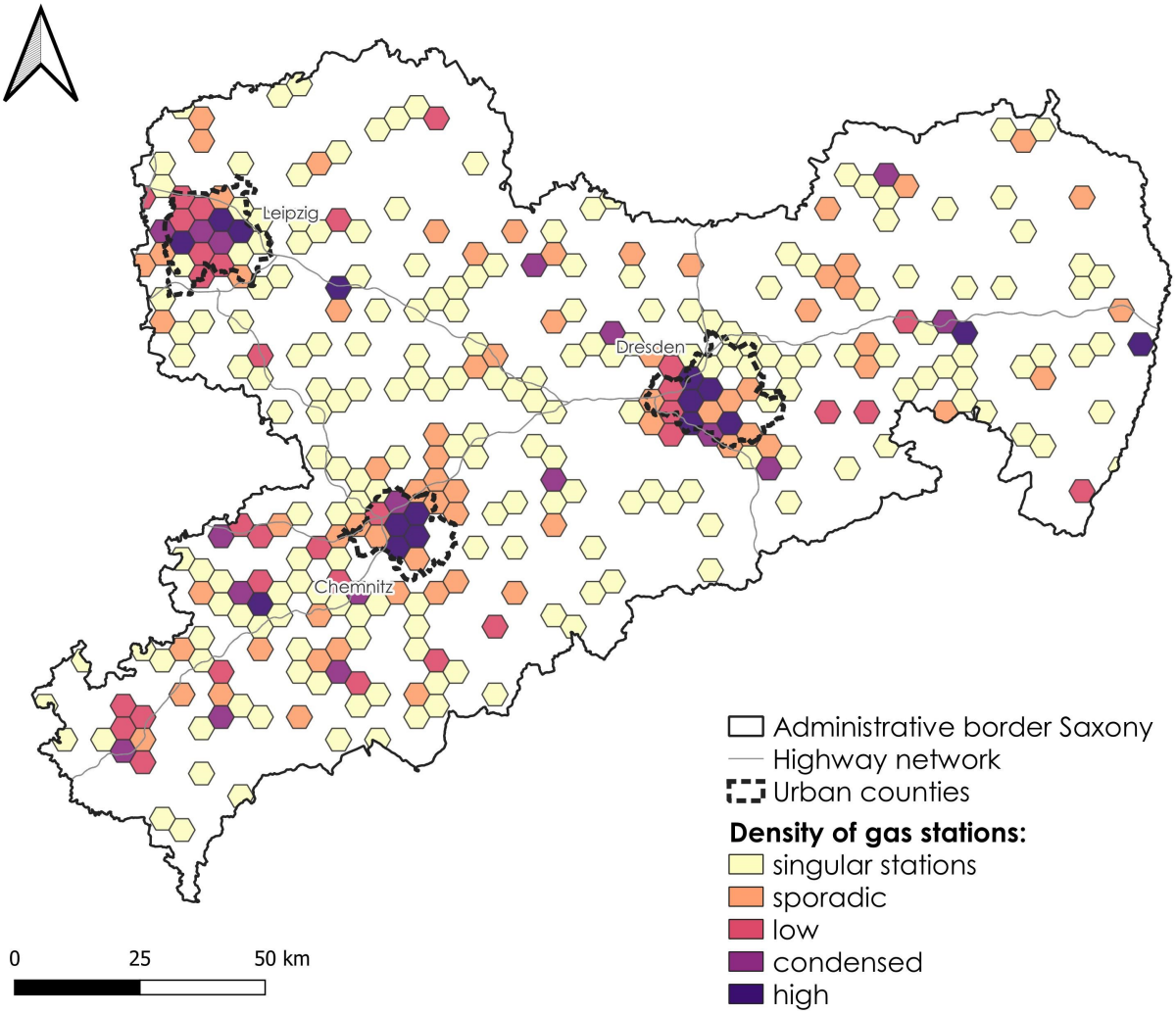

Figure 3.1: Density of gas stations across Saxony

Another factor of price impact that is considered important, according to the literature, is some measure of competition. The issue of creating a suitable measure or perhaps a combination of several measures (variables based on available data) is a complex topic and can be seen as an area of further possible research development.
For this project various approaches to determine spatial competition were explored and tested: Distance to the closest competitor, average distance of the 3 closest competitors, number of stations located within a 1 km radius, thiessen polygons, distance to another closest cluster of stations. For our final model "distance to the closest competitor" has been chosen (figure 4.1), which is simple enough and yet has noticeable and interpretable influence on the result.

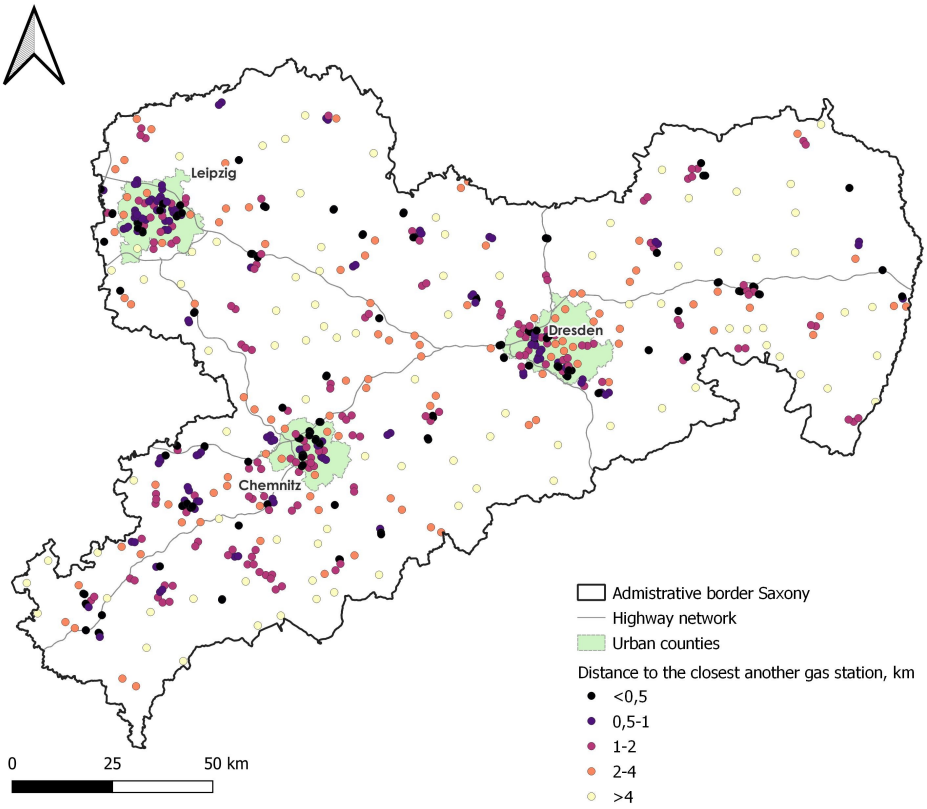

Figure 3.2: Distance to the closest competitor gas station in Saxony

Finally, a gas station being located on or near a high level road has proven to be an important factor by the literature we analysed, while smaller, rural roads are not expected to have a clear and strong influence. Thus, we have taken the backbone roads of the German road network to analyse stations proximity to them (figure 3.3). Smaller links or streets in the cities were not included.

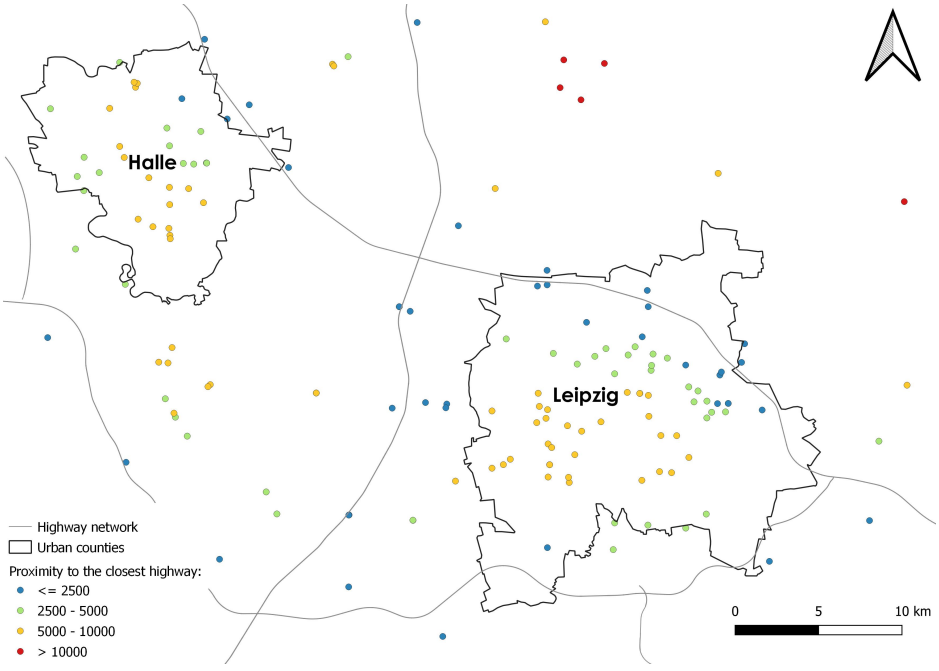

Figure 3.3: Distance to the closest highway (high level road)

Temporal dimension and other variables
As far as the time dimension is concerned, the possible areas of consideration are: daily, weekly, seasonal, and yearly trends. In this study, the focus was placed on changes over the course of a day and the frequency of alterations, which results in two other variables to predict prices. No significant difference between patterns on different working days during the week was measured (figure 3.4). The approach of analyzing prices over several days in a row gives greater insight into local price determinants and helps you understand the differences between nearby located gas stations and brands.

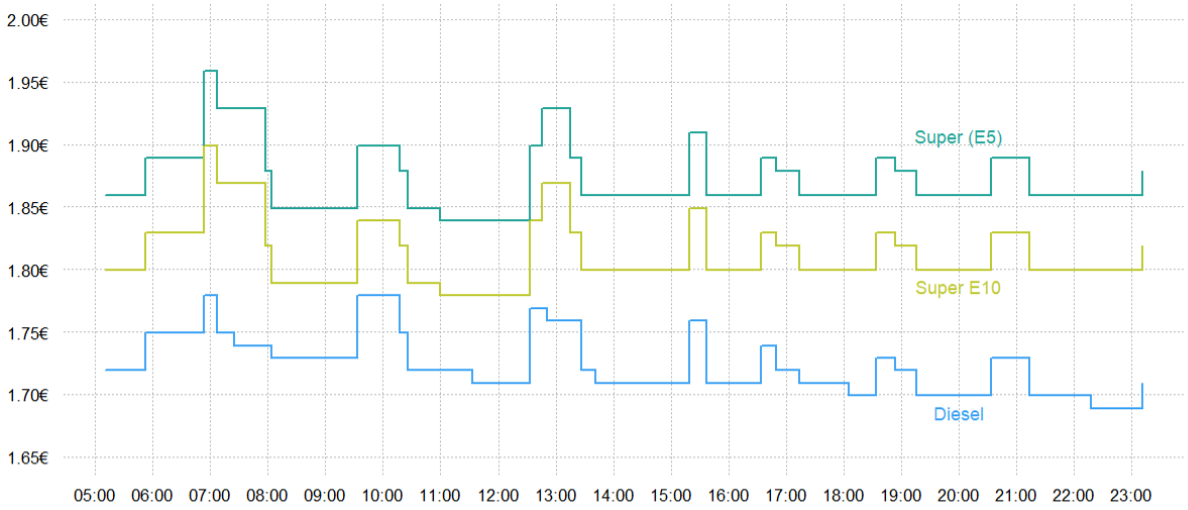

Figure 3.4: Daily changes of fuel prices on one gas station (common example in data)

Weekly changes might bring new inputs to the study, but they were found more computationally expensive and less important compared to spatial variables. Analyzing seasonal and yearly trends requires data on refinery price changes as well as multiple times more computational resources. In this case the general economic conditions, inflation and changes in oil prices will play a noticeable role and therefore need to be specified individually.
Although the period was chosen to try to mitigate changes in resource prices (refinery prices), the variable containing the price of crude oil on the exchange was taken into account as a proxy for missing refinery price data.

4 Predicting Prices

Our data setup is to use Tuesday and Wednesday as training and Thursday as the holdout to test if our predictions are actually successful. To select the best learner for this task we followed a two step process:

1. Compare a range of learners with default settings side by side to get an initial estimation for which learners may be best suited
2. Take the best performing learners through a tuning process to get even better predictions

The initially selected learners and their performance can be seen in figure 4.1. The initial learners we opted for are a decision tree learner, a k-nearest neighbour learner, a random forest learner, a generalized linear model learner, a one layer neural network learner and a support vector machine learner. The best performances were from the decision tree, k-nearest neighbour algorithm and the random forest, so we chose them to be tuned in the next step. A disadvantage with this approach is, that it is possible for one of the other learners to outperform our chosen ones after tuning. Their initial bad performance could be entirely due to very unfit default hyperparameters. We still opted for this approach, as it yielded a more certain outcome and seemed generally more fitting to our constrained computational resources.

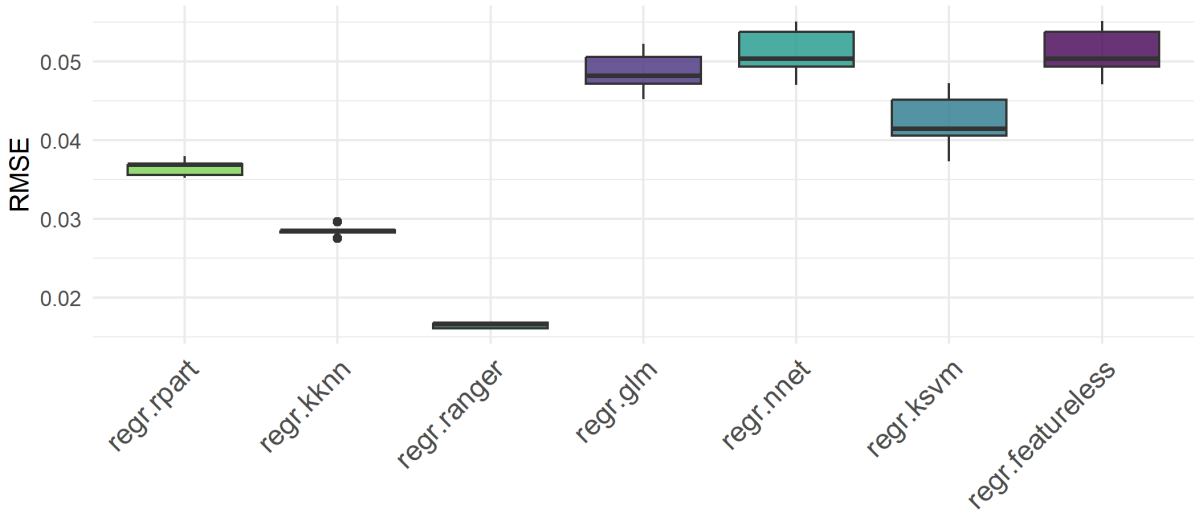

Figure 4.1: Learner performance with default settings

Having chosen our three learners to be tuned we set the tuning parameters for each of them. The decision tree is using the default mlr3 search space with varying degrees of cost parameters for their tree size, as well as varying levels of minimum split and bucket sizes. For the k-nearest neighbour learner we are considering different k parameters and our random forest is set to use 1000 trees, while exploring different values for the fraction of features used for splitting and the sample fraction for each of the trees. The Autotuners optimize within these spaces using 3 fold cross validation and getting 60 seconds of computational time do so (20 sec for the decision tree) and this process is iterated 5 times, due to the 5 fold cross validation outer resampling. We applied flexible Bayesian optimization as the tuner process for finding the best parameters. Following the guide of the mlr3 book by Bischl et al. (2024). They recommend Bayesian optimization as it can yield very good results even on smaller computational budgets. This is due to the process being sophisticated enough to efficiently move toward good performing parameter settings. The results of this process can be seen in 4.2. The ranger, which performed best in its untuned state improved slightly, similarly the k-nearest neighbour learner, which also had a slight improvement, but most interestingly the decision tree was able to cut its root mean square error down below the k-nearest neighbour learner almost reaching the same level as the random forest. Of course with a result like this, immediately the suspicion arises that what we are observing here may be due to overfitting the data and does not actually translate into better general prediction results. Naturally the next step is to make sure that isn't the case.

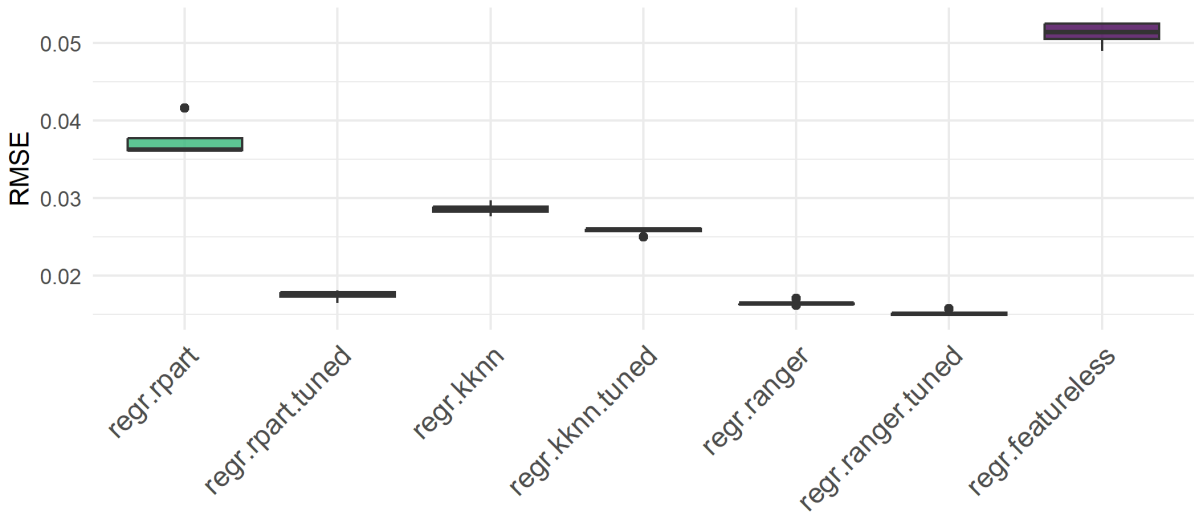

Figure 4.2: Learner performance after tuning

How this validation on the test set turned out can be seen in figure 4.3. Thankfully only the k-nearest neighbour learner seems to overfit significantly, while both the decision tree and the random forest get results very close to their training performance on the testing set. The reason that the k-nearest neighbour learner is overfitting likely has to do with the way our dataset is structured. Every gas station ends up 48 times in our training dataset (24 in testing). As such, if the internal random sampling uses two-thirds of those 48 hour averages as training and tests on the other third and the tuning returns a value for k lower than 32, then it is likely for the learner to simply choose all the sampled prices of the same gas station and average them for a prediction for the missing data. Here the learner suffers from the fact that the internal cross validation of the tuner does not sample in sequence but randomly, whereas the testing set is in sequence after the training set. Additionally, as we have multiple spatial variables, but these stay the same over time, it is certain that the first couple nearest neighbours will always be the same gas station at different points in time. We do not observe an overfit of the same degree for either the random forest or the decision tree learners. As such we choose the best performing learner on the testing data, the random forest learner, as the model to investigate further in the interpretable machine learning section. It would also be the learner recommended by us for a real world application based on our work in this project.

Figure 4.3: Learner performance on training and testing data

5 Interpretation of results

Our tuned and trained random forest learner predicts E5 prices much better than the featureless learner, but the question is: how does it achieve these results? For this purpose methods of interpretable machine learning were applied. Among these methods are: PDP (Partial Dependence Plot), ALE (Accumulated Local Effects), ICE (Individual Condition Expectation) plots. The first figure 5.1 shows us feature importance. From it we can distinguish between features that had high and low influence on the predictions.

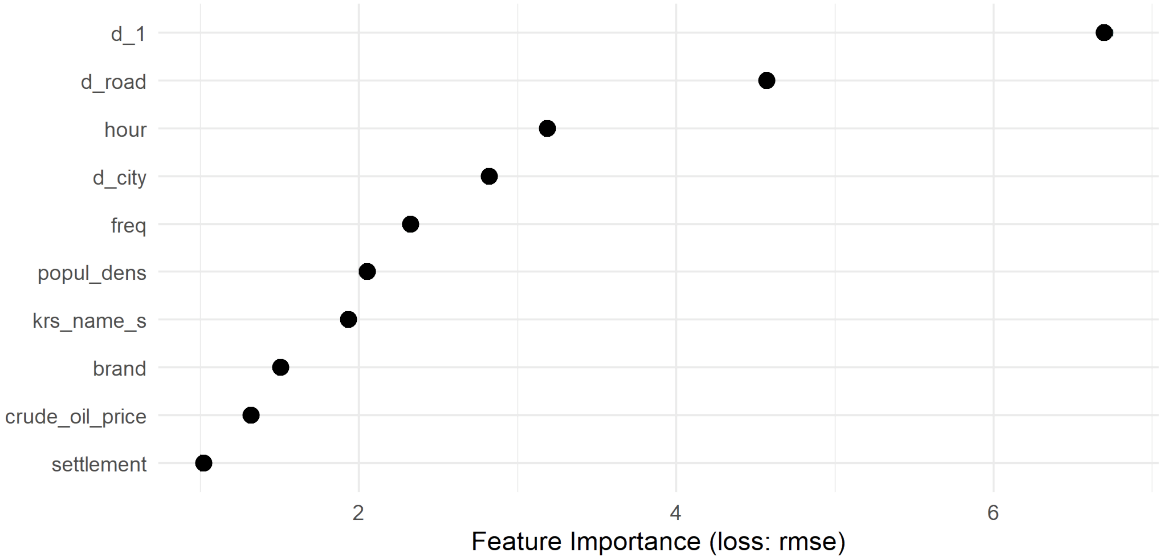

Figure 5.1: Feature importance

A set of spatial distance variables appears to have significant influence on gas price: d_1 (distance to the closest competitor), d_road (distance to the closest high level road), d_city (distance to the closest city). These results are aligned closely with the studies mentioned in the literature review.
The impact of spatial competition (d_1) should be measurable according to the literature but its influence is not always linear and not easily separable from other variables. A general upward but not uniform trend is expected (higher distances lead to higher prices).
From the ALE competitor distance plot (figure 5.2) it can be seen that there are several areas of d_1 values with different impact on price. Gas stations located closely to each other have lower than average prices, with possible exception of very dense clusters (d_1 < 200 meters). This refers us back to the research that nearby stations (clusters of stations) can work together

and compete with other clusters to a greater extent than between each other so their prices are not necessarily low. Stations located further away from others (200 meters < d_1 < 2000 meters) have lower prices due to location in a high competition zone. Following that, with distances to the closest competitor growing over 2000 meters, the prices start to increase and later are always above average. The slight decrease in prices for d_1 > 4000 meters can possibly occur because those stations are located in more rural areas with measurably lower effective demand. Since our population density data are not high-resolution, we cannot entirely offset this effect.

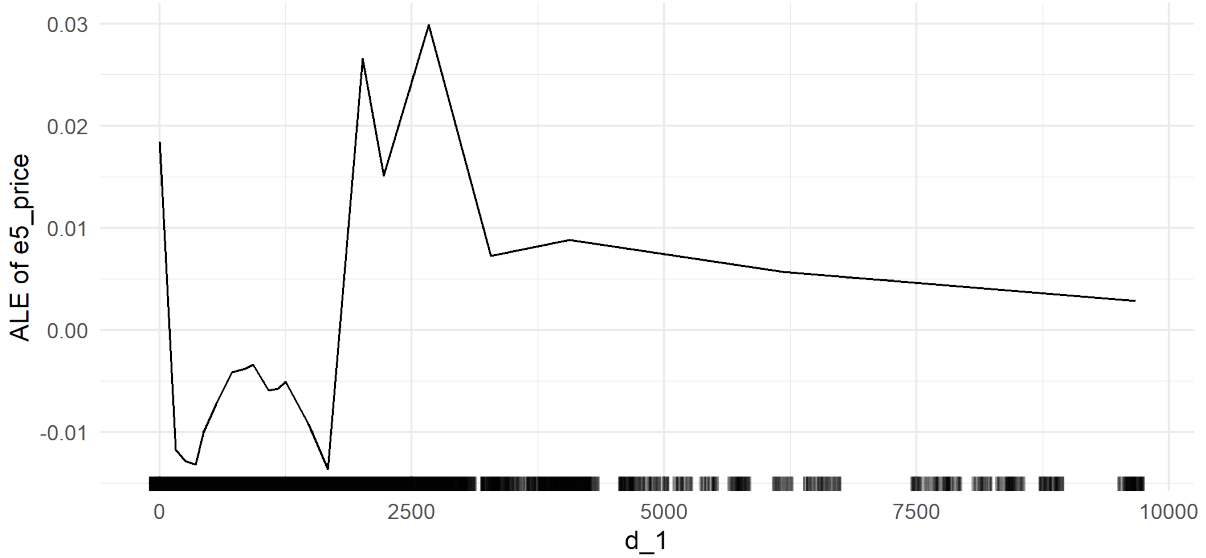

Figure 5.2: ALE distance to the closest competitor

As for the distance to the closest high-level road, we can see the pattern that is expected: on stations located in close proximity to these roads prices tend to be higher (5.3). Starting from values over 5000 meters the influence seems to be unpredictable. This is in agreement with the conclusion that there is no influence of highways on stations located far away from them, while the fluctuations of the graph line are due to correlation with other parameters on this data set.

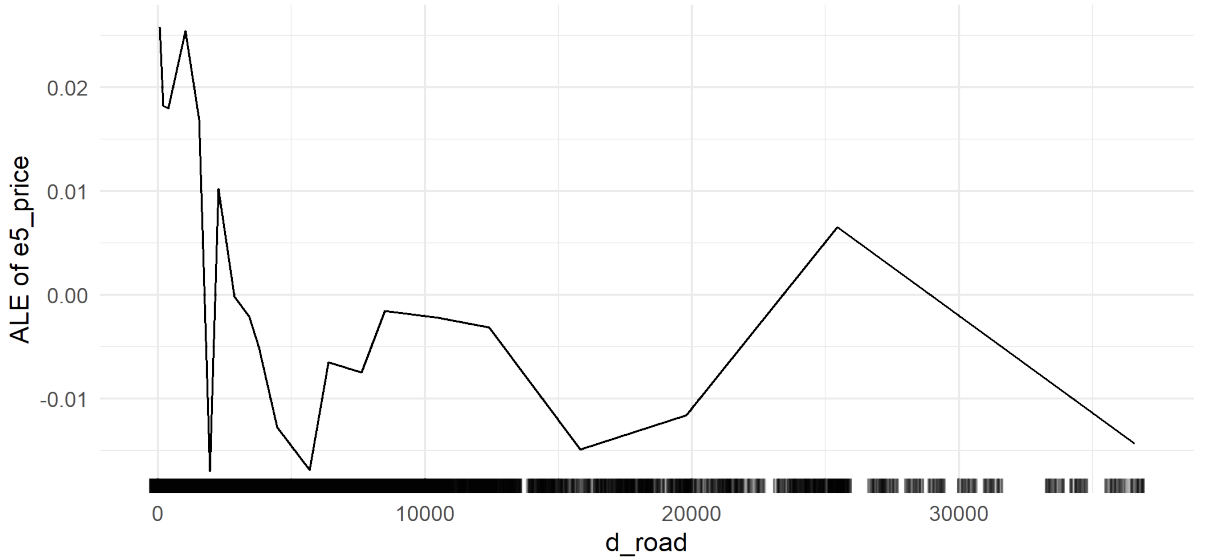

Figure 5.3: ALE distance to the closest high-level road

Along with spatial variables, hour of the day appears to be important. From ALE figure for "hour" (5.4) we can conclude that in the morning prices are higher than average, while in the evening lower. This finding is valid when speaking about the whole region, but it should be treated with caution when analyzing individual cities or agglomerations. Different patterns can occur in dense urban areas, since traffic flows and therefore demand are more concentrated in time.

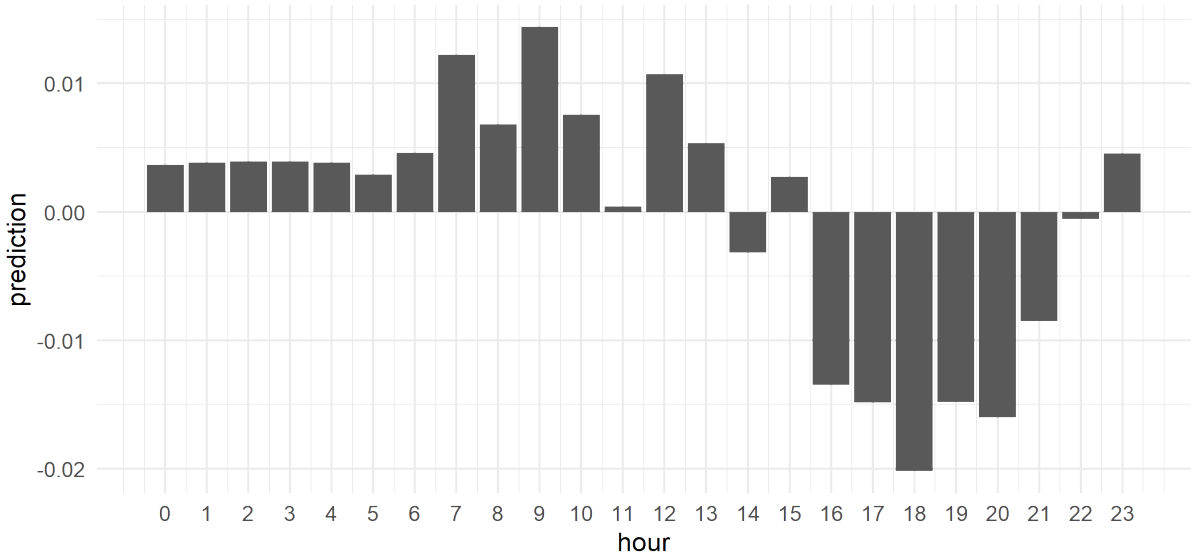

Figure 5.4: ALE hour during the day

Crude oil price has a low influence since we tried (successfully) to choose a period of time to minimize its impact. Effects of settlement type (variable "settlement") are mostly substituted by all counties considered with their own names (variable "krs_name_s") and therefore those additional variables are not much needed for model predictions (ALE plot can be seen on figure 5.5). This highlights that similarities among cities (urban counties) are not strong (which also can be seen from the plot below) and therefore cities should be considered individually.

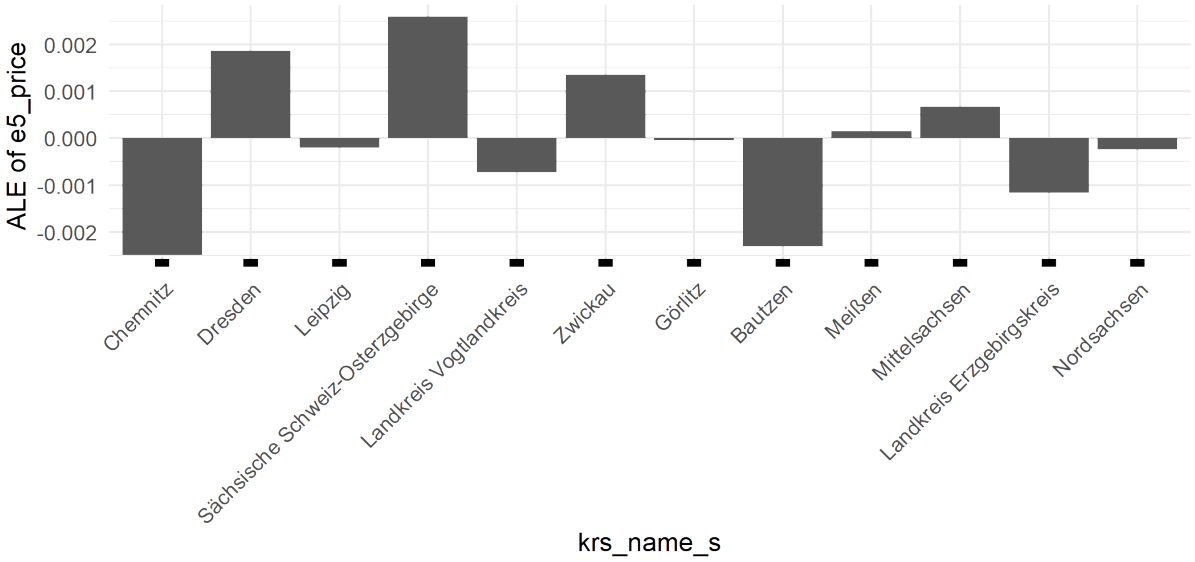

Figure 5.5: ALE name of the county

6 Conclusion

This project finally presents a machine learning procedure for predicting gas prices at local stations in Saxony. The prediction quality achieved by the trained model for this task is significantly higher than that of a featureless learner.

Limitations of the introduced approach are related to:

- Complexity of description of some parameters (e.g. spatial competition).
- Lack of data about some processes (e.g. refinery price changes, actual sale levels, presence of stores at gas stations, incomes of population in the areas, traffic conditions on nearby roads).
- Limited computational power.

Further possible research directions could include training and predicting in different regions or time periods. In case that data can be collected, several other price influencing factors described above could be introduced or modified.
The data frame created during described preprocessing procedure is highly structured and adaptable. Therefore different set of learners as well as different hyperparameter combinations are possible to test and compare. For this project we chose hourly averages over the course of three days, but both these parameters could be varied freely to adapt it to other possible research questions. For a high frequency price change interaction it would be possible to look at averages for every 5 minutes over the course of half a day, for example.

Finally, possible uses for this algorithm include:

- Creating an app for consumers and logistics operators, allowing not only to see, but to predict and compare future prices on gas stations of interest.
- Creating a tool for fuel brands to analyse and predict the market strategy of their rivals as well as compare their prices with close competitors in real-time.
- Helping local authorities, planners, consulting firms, or logistics companies to compare areas offering lower or higher price levels for fuel to make informed decisions about new infrastructure development or route planning.

Bibliography

Bischl, Bernd, Sonabend, Raphael, Kotthoff, Lars, and Lang, Michel, eds. (2024). *Applied Machine Learning Using mlr3 in R*. CRC Press. ISBN: 9781032507545. URL: https://mlr3book.mlr-org.com.

Fronzetti Colladon, Andrea, Verdoliva, Giulia, Segneri, Ludovica, and Vitali, Andrea G. (2024). "Analyzing gasoline prices in five Italian cities: Insights from social network analysis". In: *Cities* 150, p. 105075. DOI: https://doi.org/10.1016/j.cities.2024.105075.

Haucap, Justus, Heimeshoff, Ulrich, and Siekmann, Manuel (2015). "Price Dispersion and Station Heterogeneity on German Retail Gasoline Markets". In: *Forthcoming in: The Energy Journal*.

Kucher, Oleg, Burnett, J. Wesley, and Lacombe, Donald (2018). "Spatial spillovers in US wholesale gasoline markets". In: *Papers in Regional Science* 97.3, pp. 687–711. DOI: https://doi.org/10.1111/pirs.12270.

Pennerstorfer, Dieter and Weiss, Christoph (2013). "Spatial clustering and market power: Evidence from the retail gasoline market". In: *Regional Science and Urban Economics* 43.4, pp. 661–675. DOI: https://doi.org/10.1016/j.regsciurbeco.2013.04.002.

Statement of Individual Contribution

In the preparation of this document, the following contributions were made by each group member:

Kirill Salnikov: Explored background of the topic, conducted the literature review, collected and preprocessed spatial data for usage in machine learning, tested different sets of variables in the machine learning procedure.

Tim Stolz: Preprocessed and aggregated the data related to time variables and prices, chose and trained final model for predictions, interpreted the results of machine learning in plots.

Jiujiu Liao: Developed the machine learning procedure, including initial learners set, tuning procedure, benchmarking, made suggestions for future improvements.

Statement of authorship

I hereby certify that I have authored this document entitled *Modeling Gas Prices* independently and without undue assistance from third parties. No other than the resources and references indicated in this document have been used. I have marked both literal and accordingly adopted quotations as such. There were no additional persons involved in the intellectual preparation of the present document. I am aware that violations of this declaration may lead to subsequent withdrawal of the academic degree.

Dresden, 27th August 2024

Jiujiu Liao | Kirill Salnikov | Tim Stolz