Business Intelligence
Final Submission Report

Nadia Verdha (11739391)
Person A
University of Technology of Vienna
Vienna, Austria

Elton Mpofu Tinashe (12333475)
Person B
University of Technology of Vienna
Vienna, Austria

1 Business Understanding

1.1 Data Source and Business Scenario

For this assignment a dataset about used cars [1] was chosen and downloaded from Kaggle. This data was sourced from AutoScout24, one of Germany's largest car sales websites. It includes information about car offers, covering cars manufactured from 1995 to 2023. This dataset is a perfect fit for everyone who wants to perform a market research study of the automotive industry in Germany, as it can be used to understand trends in car prices, or to explore popularity of car models over the years. The code used for the Data Understanding and Preparation can be found in our GitHub repository [2].

1.2 Business Objectives

Possible business objectives may include:

1. Market research and understanding of trends in the sale of used cars
2. Identify factors that affect the price of used cars
3. Indentify if a listed offer is placed well among other similar ones
4. As a result, optimize pricing strategy to increase revenues and sales

1.3 Business Success Criteria

After finishing this project, the user should be able to get an idea of the current trends in the automobile industry and know the market better. Also, the seller can get a better idea about at which price point to list his/her used car.

1.4 Data Mining Goals

Data Mining Goals might include:

1. Find out correlations and relationships among the features
2. Build a Price Prediction Model

1.5 Data Mining Success Criteria

Since the main topic of this project revolves around used car prices, I would say that the best success criteria would be a Predictive Model that predicts car prices as accurately as possible. In this case, we can use metrics such as Root Mean Squared Error(RMSE), Mean Squared Error (MSE), and R-squared to evaluate our model. We would want the RMSE and MSE to be as low and R-Squared to be as close to 1 as possible.

1.6 AI Risk Aspects

Some possible AI Risk Aspects one might recognize:

1. Bias in the Data - the possibility of inaccurate, incomplete information
2. Data Privacy Issues - data might raise privacy concerns if it contains personal/sensitive information
3. Interpretability of the model - complex models can be hard to explain. In our case, we should choose a model that can be easily interpreted (Linear Regression, Random Forest Regressor, etc.)

2 Data Understanding

In this part of the project, exploratory data analysis (EDA) will be performed in order to understand the data provided by the chosen dataset.

2.1 Attribute Types

The dataset has 251079 instances and 15 columns, which are listed below:

1. Index (integer) - This column can be used as an ID
2. Brand (string) - Manufacturer of the car eg. (Volkswagen)
3. Model (string) - Model of the car eg. (Volkswagen Golf)
4. Color (string) - Color of the model
5. Registration date (string) - Month/Year of registration date
6. Year (string) - Production Year
7. Price in Euro (string) - Price
8. Power KW (string) - Power of engine in KW
9. Power PS (string) - Metric Horsepower used commonly in Europe
10. Transmission Type (string) - Type of transmission
11. Fuel Type (string) - Type of fuel
12. Fuel Consumption l_100km (string) - Amount of fuel that car uses to travel for 100 km in liters
13. Fuel Consumption g_100km (string) - Amount of CO2 emitted in grams for every km drive
14. Mileage in KM (float) - Mileage of car in km

[1] https://www.kaggle.com/datasets/wspirat/germany-used-cars-dataset-2023
[2] https://github.com/nadiaverdha/BusinessIntelligence/tree/main

15. Offer description (string) - Description of scraped offer

As seen above, most of the columns are not in the correct type format.

2.2 Data quality

This step was listed as the third one in the exercise description,but I had to put it second because there were a lot of issues with our data.

2.2.1 Variables Type Format.

First of all, as mentioned before most of the data was not in the correct format and variables that should have been numeric were of string type. Since statistical properties cannot really be computed with non-numeric variables, this is a step we already performed in this stage of the project. While doing so we also noticed that some of the values in some columns for example "Manual", "04/2017" in column Year or "ROSTHREI" in column Price did not really make sense. So in this case, we decided to keep only those rows that had numeric data. Other problematic columns like Fuel Consumption l_100km and Fuel Consumption g_100km did not only contain numeric values but the unit of measure as well. In this case we decided to keep only the data without the unit of measure.

2.2.2 Missing Data.

When converting the data in the correct type format, we also noticed that the columns Consumption l_100km and Fuel Consumption g_100km also contained data points with values such as "-" , which we replaced with NaN values that will be handled in the next part of the project.
The columns that contain missing values are summarized below:

1. Color - 166 entries (0.06%)
2. Power KW - 128 entries (0.05%)
3. Power PS - 128 entries (0.05%)
4. Fuel Consumption l/100km - 27965 entries (11%)
5. Fuel Consumption g/km - 36707 entries (14%)
6. Mileage in KM - 62 entries (0.02%)

I think that the missing data in the column color is completely missing at random as it seems to be unrelated to other attributes. This type of missing values can be easily dropped, however, I would suggest to impute it using the mode. Power KW and Power PS are both Missing at Random and connected to each-other because if one is missing so is the other one. In this case I would suggest to group rows by model and then impute the values using mean. As shown above there are clearly more missing values in case of Fuel Consumption in l/km and g/km which are not exactly missing at random as they are related to the fuel type feature of the dataset. For example, if the car is electric then the amount of fuel consumed by it should be 0. I think for imputing the other missing entries again mean could be used.

2.2.3 Outliers.

In terms of outliers our dataset proved itself to be very problematic, and the main reason for that is that many of the information found in the data did not make sense. For example a fuel consumption value of 789 l/100km when the car was of electric type was quite illogical.Also, there were many instances where this feature had values higher than 15 l/km which is usually the maximum a car can have. In this case, such values can be substituted by the median, since this statistic is not really affected by outliers.

Figure 1 shows the values before quality issues of the data were resolved.

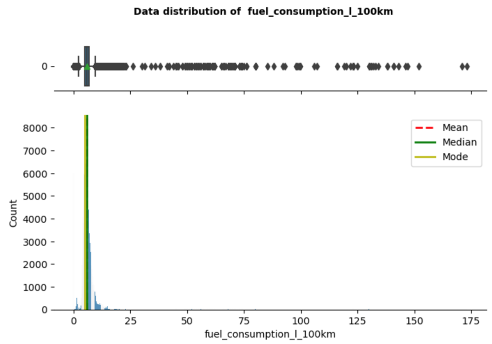

Figure 1. Before dealing with data quality issues

2.2.4 Previous Data Cleaning.

The data was quite messy and the quality of it was really questionable.Therefore, I don't think that any cleaning steps were performed by the authors previously.

2.3 Statistical Properties

2.3.1 Correlation.

Figure 2 shows how our numerical columns correlate with each-other. Some insights based on the figure:

- Expected strong negative correlation between year and mileage in km
- Expected negative correlation between year and fuel consumption l/100 km
- Expected negative correlation between year and amount of CO2 emitted
- Strong positive correlation between power in KW and power PS (represent same thing)
- Positive correlation between price in euro and power PS/power KW
- Low positive correlation between prince in euro and fuel consumption l/100km

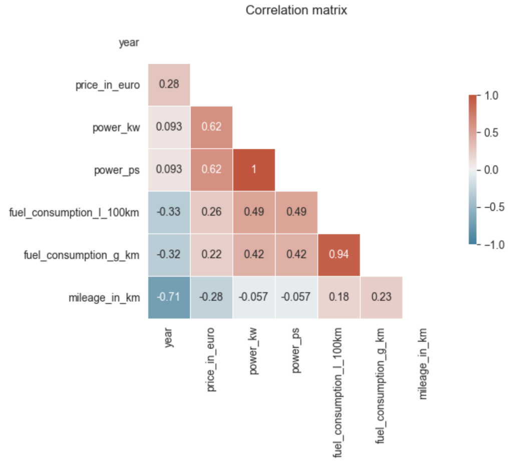

Figure 2. Correlation of numeric features

2.3.2 Other statistics.

We also had a look into the distribution of all our numeric variables just to get a better idea about our data and check for presence of outliers in it. For example for column Year we can recognize a left-skewed distribution with median being at the value of 2019 and mean at the value of 2016. There are some outliers present, mostly related to the fact that listed offers of cars manufactured before 2000 are fewer in number when compared to cars manufactured in later years.

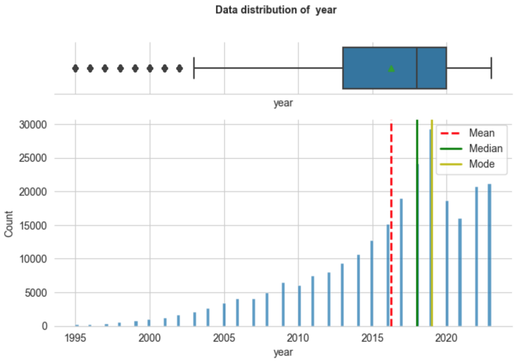

Figure 3. Distribution of column Year (Year of Manufacturing)

The following plot displays the relationship between Manufacturing Year and Price in euro. I put the Price column in logarithmic scale as it makes the plot more visible when you deal with very big/ very small numbers. As seen below, there is some kind of positive relationship between these two variables, as with the increase in years also the prices tend to be higher.

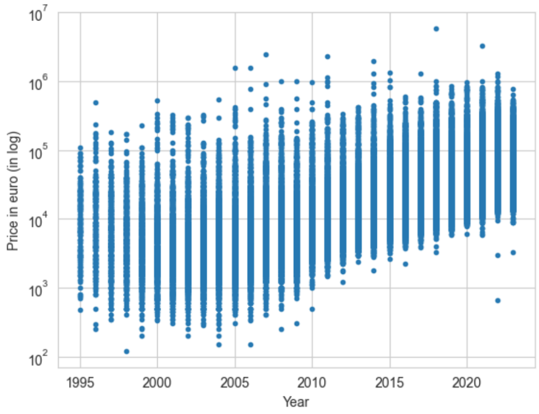

Figure 4. Relationship between year and price in euro

Figure 5 aims to show the relationship between Mileage in Km and Price in euro. As show in the correlation plot, figure 2, these two variables had a negative correlation between each-other. Again, I have put their values in the log scale. The relationship between these two variables is not exactly linear, however, one can notice a descrease in price as the mileage in km values increase.

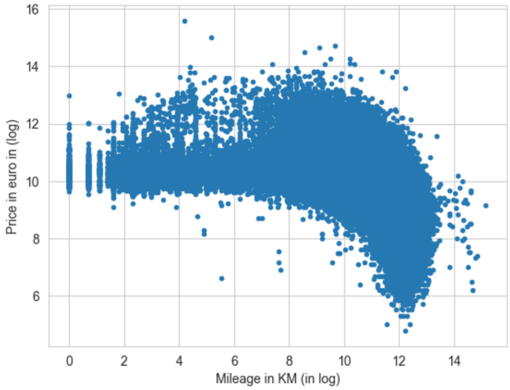

Figure 5. Relationship between mileage in km and price in euro (both in log scale)

2.4 Visual Exploration

Figure 4 shows the number of listed offers per brand. As seen below, the majority of used cars per sale are Volkswagen, Mercedes-Benz and Audi. The other plot, figure 5 shows the average price of used cars based on the brand. As expected,Lamborghini leads the group, followed second by Ferrari. The most listed brands like Mercedes and Audi lie somewhere in the middle.

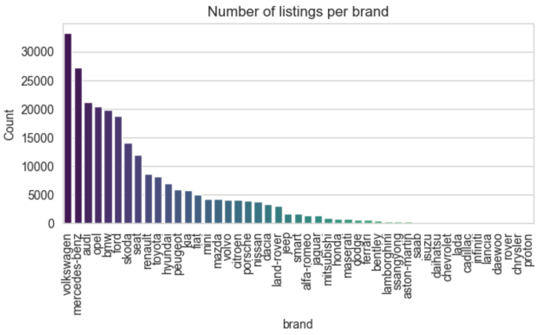

Figure 6. Number of listings per brand

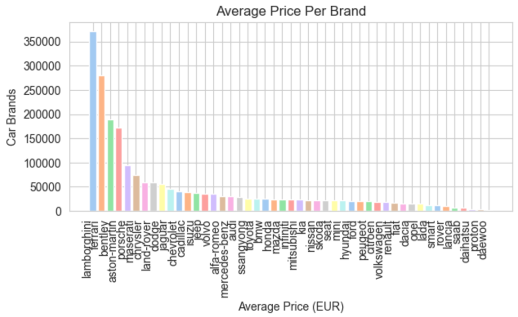

Figure 7. Average Price per Brand

The following figures, figure 8 and figure 9, give us insights on the price for different types of fuel and transmissions. As shown below, based on the transmission types, the median of the price for automatic cars is higher then the rest. Same can be observed for Diesel Hybrid cars when comparing them based on fuel types.

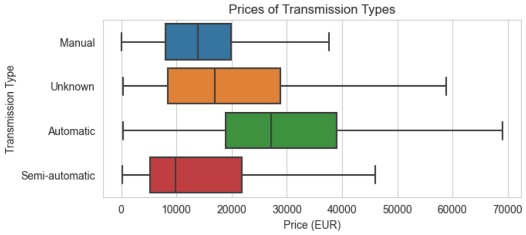

Figure 8. Price of Transmission Types

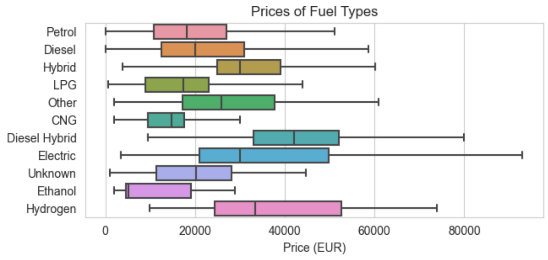

Figure 9. Price of Fuel Types

2.5 Sensitive Attributes/Representation of Groups

This dataset contains information about used cars and their features, and since it does not contain any personal or demographical information about buyers or sellers, I don't recognize any ethically sensitive attributes. In regards to representation of the groups, there are for sure some types of cars that are more present than the others. For example, in terms of fuel type Petrol and Diesel cars take up to 91% of the listed offers, leading to other types of cars such as Hybrid or Electric to not be very well-presented in this dataset.

2.6 Risks and Biases

One of the main research questions of this project is market research of automobile industry in Germany. However, this dataset comes from one car website only and ,therefore, it may not be representative of the whole used car market, as certain brands may be underrepresented and others may be overrepresented. Moreover,everything is manually entered on the website and there were a lot of innacuracies in the data, which might as an effect introduce bias in our analysis. For certain attributes such as Fuel Consumption l/100km we had to do research on the web, and in this case an expert's knowledge could have helped.

2.7 Next Steps

All in all, there were a lot of issues in our dataset. Consequently, in the next steps we should focus on dealing with outliers and innacuracies, filling in the missing values and preparing our data for fitting the model.

3 Data Preparation Report

In this part of the project we will prepare the data for modeling. To do so ,we will perform necessary actions to solve issues that were identified on the data understanding part.

3.1 Actions based on analysis

3.1.1 Outliers.

The dataset contained outliers that required careful handling. Initially, I focused on addressing instances where the fuel type was electric, yet incongruent values for fuel consumption were present. I specifically had to target those

entries, replacing their values with zeros. Subsequently, to address instances where the fuel consumption exceeded 15 l/100km, I replaced these outliers with the median fuel consumption value for each respective car brand. This strategy is good because it accounts for brand-specific variations in fuel efficiency. The same method was also applied to the fuel consumption in g/km, considering the inherent connection between these two variables.
Figure 10 below shows how the data looked after dealing with some of the outliers that did not make sense.

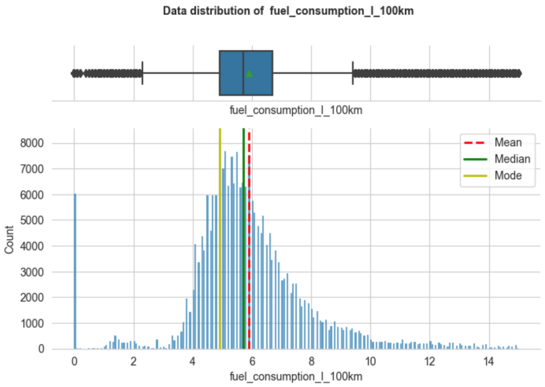

Figure 10. After dealing with data quality issues

3.1.2 Missing Data.

In this part of the project, I had to impute the values with missing values. For the categorical attribute Color, the mode was employed to impute missing values, ensuring consistency in the imputed values while preserving the categorical nature of the attribute. Furthermore, to enhance the dataset's completeness, cars were grouped by brand, then mean values for Power(Kw), Power(Ps), and Fuel Consumption (l/100km) were calculated and used to impute missing values. This approach aims to maintain data integrity and avoid bias in subsequent analyses.

3.1.3 Unwanted Columns.

In this phase of the project, the decision was made to eliminate the Registration column, considering that the year column provided more utility and relevance. The year column, offering a concise representation of the registration date, emerged as a more practical alternative. Additionally, the ID column was removed because it served the same purposes as the index. This step streamlined the dataset by eliminating redundancy and optimizing its structure for further analysis.

3.2 Derived Attributes

From the existing attributes, I derived two additional features. The first, car age, was calculated by subtracting the registration year from the current year. This attribute holds significance as the age of a car tends to inversely correlate with its price, playing a crucial role in determining its market value. I also introduced the power-to-price ratio as another derived attribute. This ratio becomes valuable as a higher value indicates that the car delivers more power for the price paid. This metric proves beneficial for market researchers aiming to discern whether a car offers optimal value for money.

3.3 External data sources

For this project, challenges arise due to biases and limitations inherent in the data. The dataset, originating solely from one website, poses a potential problem. An effective solution would involve acquiring additional data from various independent websites. This strategic move aims to capture a broader range of attributes, ultimately enhancing the accuracy of our predictions. By steering clear of reliance on manually entered data, especially for critical attributes such as power (ps) and fuel consumption (l/100km), we position ourselves to achieve more precise and reliable results.

3.4 Pre-processing steps

3.4.1 Scaling.

I performed scaling on the columns representing power (kW), power (PS), fuel consumption (L/100km), and mileage (km). This process normalizes the values, enhancing the sensitivity of our model and ensuring a fair evaluation of its performance.

3.4.2 Highly correlated features.

To identify and mitigate potential challenges during our modeling phase, I employed a heatmap to visually inspect the degree of correlation of all the numerical data within the dataset. Figure 11 below shows the heatmap. Notably, I observed significant correlations among certain features, namely power(ps), and power(kw). To address the issue of multicollinearity, I opted to remove one feature from each pair that exhibited high correlation. This strategy is aimed at improving the reliability of our modeling process by eliminating redundant information and ensuring a more robust analysis.

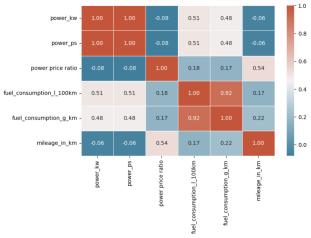

Figure 11. Heat map showing correlation of features

3.4.3 Encoding the categorical data.

I employed label encoding for categorical variables, including color, fuel type, brand, and model. This approach replaces categorical labels with numerical values, enabling the model to comprehend and extract meaningful information. I removed categorical distinctions and substituted them with numeric representations. This transformation enhances the model's ability to interpret and analyze the data effectively.

3.4.4 Data spliting.

First, I excluded the column representing the variable we aim to predict from the dataset, price in euros. After that, I split the remaining data into training and testing sets. This division is crucial for training the model on one subset and evaluating its performance on another. This will allow us to gauge how well the model generalizes to new, unseen data and aids in identifying potential issues such as overfitting or underfitting during the model development phase.

4 Modeling

4.1 Data Mining Algorithms

As previously mentioned, the focus of this assignment is to build a Price Prediction Model. Since we are dealing with numeric data, the reasonable choice is to go for a regression model. Suitable data mining algorithms that can be used for these types of problems can be Linear Regression, Ridge Regression, XGBoost, Random Forest Regressor, etc.. We decided to work with Random Forest since it can be robust to very noisy data.

4.2 Hyperparameters

When dealing with Random Forest Regressors, multiple hyperparameters can be considered for fine-tuning, such as:

1. n_estimators - number of trees in the forest
2. max_features - max number of features considered for splitting a node
3. max_depth - max number of levels in each decision tree
4. min_samples_split - min number of data points placed in a node before the node is split
5. min_samples_leaf - min number of data points allowed in a leaf node
6. bootstrap - method for sampling data points (with or without replacement)

I believe fine-tuning the parameter n_estimators can be considered more essential when compared to other parameters. This is because the number of trees used can play a huge role in underfitting (low number of trees) or overfitting (large number of trees) of the model. Moreover, increasing the number of trees can also increase the ability of the model to generalize well in unseen data.

4.3 Splitting data

We decided to split the data based on the following ratio: 70% train, 15% validation, and 15% test set. For that, we made use of the train_test_split of the sklearn library and also used a random state of 42 for reproducibility.

4.4 Model Training

Multiple models were trained on train sets for different values of the parameter n_estimators, such as 25,50,75,100,125, 150,175,200. Each of these models was evaluated on the validation set and their performance was stored in a list for later observation.

4.5 Performance

The metric used to evaluate the performance of the different models on the validation set was Root Mean Squared Error which is typically used for regression tasks. As seen in the figure 12 below, there is a clear decrease in the RMSE when the number of trees increases which is a common behavior for random forest models. Increasing the number of trees can improve the ability of the ensemble to generalize and provide accurate predictions. As a result, the best model is the random forest with 200 trees which achieves an RMSE of 1975.

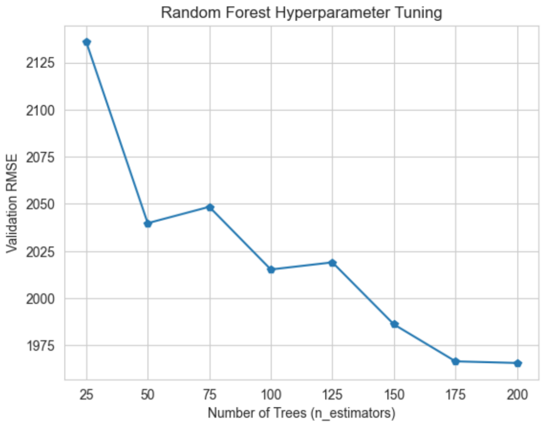

Figure 12. RMSE for different values of n_estimators

The figure 13 below shows that most of the predicted values are concentrated around the black line which also reflects the ground truth.

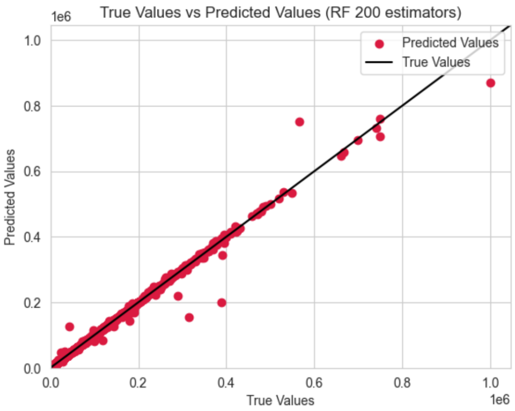

Figure 13. Predicted Values vs True Values for the best model

5 Evalution

5.1 Apply the final model

I use the final model to see how it performs on the test data it had not seen. The RMSE for the test data was 17647. The figure below shows the predicted value vs True values from the best model. From the graph, it shows that also the test data was around the black line and it shows that the model performed well.

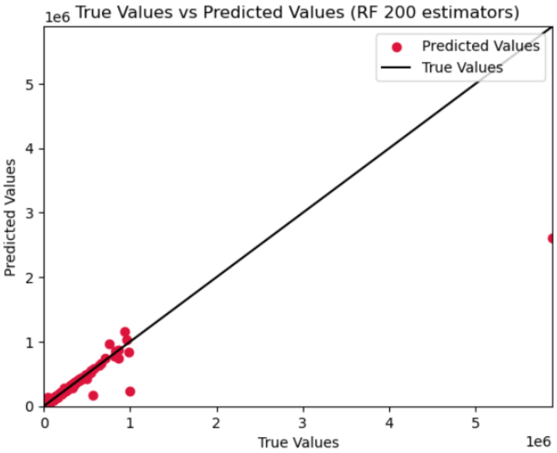

Figure 14. Predicted Values vs True Values for the best model for test data

5.2 Re-training the data

To retrain the data with both the validation data and train data together I concatenated both the validation data and train data. After that, I did the model fitting with the hyper-parameters that were used for the previous section. Next, test data was used to calculate the performance of the model.The retrained model showed a slight increase in RMSE (17739) compared to the previous section 17647.

5.3 State of the art and Base-line Performance

5.3.1 State of the Art Perfomance. During my research, I tried to find papers that used the same dataset as we did. I found notebooks that had performed analysis with our dataset but most of them focused their analysis on a specific car model and had performed different preprocessing steps. For this reason, we could not find a baseline that could also be comparable to our model.

5.3.2 Base-line Perfomance. Since we could not find a baseline, we decided to implement a dummy regressor. The dummy regressor results proved to be the worst,yielding a RMSE value of 44716. So from those results, it turned out that our model was performing quite better than our baseline.

5.4 Comparing performance with benchmark and baseline

Since we were not able to get the benchmark value we had to use our baseline using a dummy regressor. From the RMSE value, we were able to get some insights into how our model was performing. The model performed well compared to how the dummy regressor did. We also noticed that features like power ratio and power in kw had the largest influence on the model.

5.5 Does it meet our success Criteria

The model performed better than we expected. We could be able to predict the price of cars though it won't be the best price. However, the results seemed to be closer to the actual results which was a good thing for us. We were also able to analyze trend markets and see which car models were getting the most sales and were able to see which type of cars people like to use these days. So we can consider this a success in some way.

5.6 Bias on attributes

We decided to check if there was any bias in our model in regards to the car's fuel type.For that we grouped the test data based on fuel type so that we could evaluate whether the model exhibited bias towards different groups. The graph below shows the results of the RMSE values obtained for different fuel type's groups:

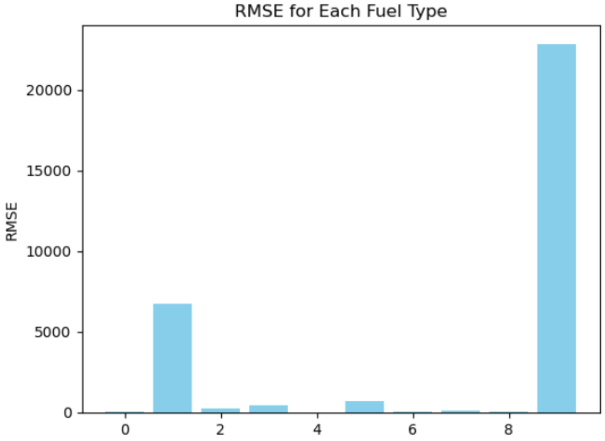

Figure 15. RMSE values of subgroups of fuel type

The graph showed that there is actually bias because some fuel types had better results as compared to some of them.

6 Deployment

6.1 Performance and business objectives

We believe that the data exploration performed in the first part of this project provides valuable insights into market trends while enabling analysis of such trends. As for the model part, the model successfully predicts prices of unknown data. In this way, it helps sellers make a more informed decision regarding their car price by optimizing their pricing strategy. Moreover, having an accurate price prediction model can also provide businesses with a competitive advantage as it enables them to adjust pricing in real time and offer competitive deals. However, due to the high RMSE on the test set, our model is far from perfect. Probably a future improvement would be to experiment with other regressor models as we might achieve better performance. Worth mentioning is that very often performance of the model depends a lot on the quality of the data fed into it, and our data was quite messy forcing us to make a lot of assumptions during our analysis. Therefore, probably finding a better dataset might offer us better and more accurate predictions. Since the final model still requires some improvement, we would suggest a hybrid solution. Integrating domain knowledge and expertise in certain cases might enhance decision-making.

6.2 Potential ethical aspects and risks in deployment

The dataset used in the project does not contain any explicit information related to sellers and buyers, therefore, we believe that no direct privacy concerns might arise. However, we were able to identify some other risks related to the deployment of this model. Since random forest models are not that easy to interpret and explain, for example when compared to decision trees, lack of interpretability might lead to distrust. It might hinder the broader acceptance of the model by the stakeholders. Moreover, the model's predictions might not always be accurate, and therefore, relying solely on the model without considering its limitations might pose a risk by causing financial losses or missed business opportunities.

6.3 Important aspects

Some important aspects to be monitored during deployment:

1. Model performance - monitor RMSE and intervene when there are consistent deviations between actual and predicted prices
2. Data Security - monitor any unusual unauthorized access to the model
3. User feedback - monitor feedback from users and intervene if the feedback is consistently negative

6.4 Reproducibility aspects

Being aware that reproducibility is a critical aspect of any data science project, we made sure to explain our analysis and implementation in as much detail as we could in this report. Moreover, by using a random seed when needed we ensured that everyone reproducing this project would get similar results as us. Last but not least, the link to our GitHub repository is also provided for anyone who would be interested in recreating this project.

7 Conclusion

The biggest issue we faced in course of this project was the messy nature of the dataset. Dealing with outliers and ensuring data cleanliness posed significant problems. The outcome, as reflected in the high RMSE values for all test data, indicates that the model's predictive performance was not as successful as anticipated. We would say that the dataset's inherent problems likely played a role in having success of the experiment. Moreover, the data cleaning process required careful consideration to avoid introducing bias or undesired effects. Despite the efforts invested in preprocessing, the experiment's overall outcome suggests limitations we had because of the dataset. What we learned from this experience is the importance of data quality and the impact it can have on the success of a machine-learning project. Moving forward, we would consider an alternative dataset that has better quality and less noise. In summary, while the experiment faced challenges and did not get the desired outcomes it provided insights into the complexities of working with real-world, because dealing with messy datasets is quite common.