Data Science: Capstone
(HarvardX PH125.9x)

## Capstone Project II

# Classic Trader

**Predict prices of classic and exclusive cars**

February 2022

submitted by Bela Koch

# Contents

# 1 Introduction

The goal of this exploratory machine learning project is, to understand the key drivers of the prices of classic and exclusive cars and to predict the prices using machine learning techniques. In a first step, the data used in the project is described and analysed. Then, the relevant metrics are described, the process of training the models is addressed and the results of the algorithms with the test data and the validation data are shown. Finally, possible weaknesses and limitations are discussed and suggestions for improvement are made.

# 2 The data

## 2.1 Data Source

The data used in this project was scraped from classic-trader.com - a website which describes itself as the international marketplace for classic vehicles. All advertisements currently being placed were downloaded. If you intend to run the scraping process, please note that this will take some time. Alternatively (and as default in the corresponding script), the already scraped data is downloaded via the GitHub repository associated with this project.

During the processing of this project, about 11 thousand observations could be obtained by scraping classic-trader.com. After data cleansing, about 8 thousand observations were available for the analysis.

## 2.2 Overview of the data

### 2.2.1 Variables considered

| Variable | Type | Description |
|---|---|---|
| ID | integer | Identification number of vehicle from website |
| manufacturer | Factor | Manufacturer of vehicle (factorized) |
| model | Factor | Model of vehicle (general, e.g. 3 Series; factorized) |
| model_name | Factor | Model name of vehicle (more specific, e.g. 318is; factorized) |
| year | Factor | Year of manufacture (factorized) |
| decade | Factor | Decade of manufacture (factorized) |
| mileage | integer | Mileage, all converted to kilometers |
| body_style | Factor | Body style of vehicle, e.g. sedan (factorized) |
| color | Factor | Color of exterior body (factorized) |
| power | integer | Metric to indicate power produced by engine, measured in kilowatts |
| ccm | integer | Measure of engine displacement, measured in cubic centimeter |
| cylinders | integer | Number of cylinders |
| steering | Factor | Placement of steering position, e.g. lefthand drive (factorized) |
| transmission | Factor | Type of transmission (factorized) |
| drive | Factor | Type of drivetrain, e.g. rear-wheel drive (factorized) |
| fuel | Factor | Fuel type (factorized) |
| condition | Factor | Condition of vehicle (predefined classes, factorized) |
| price | numeric | Advertised price measured in euro |

### 2.2.2 Missing data

The following table shows the missing data. As we can see, the variables mileage, color, condition and price contain many missing values.

| | NAs absolute | NAs relative |
|---|---|---|
| ID | 0 | 0% |
| manufacturer | 0 | 0% |
| model | 0 | 0% |
| model_name | 0 | 0% |
| year | 0 | 0% |
| mileage | 3219 | 28.46% |
| body_style | 0 | 0% |
| power | 145 | 1.28% |
| ccm | 0 | 0% |
| cylinders | 240 | 2.12% |
| steering | 0 | 0% |

| | NAs absolute | NAs relative |
|---|---|---|
| transmission | 0 | 0% |
| drive | 701 | 6.2% |
| fuel | 0 | 0% |
| color | 2545 | 22.5% |
| condition | 6308 | 55.77% |
| price | 1347 | 11.91% |
| decade | 0 | 0% |

The missing values for the variable price are due to the fact that the vehicles can be advertised with "price on request". For the variable "condition", the category "original" is assumed in case of missing values, which describes the condition of the car as "unchanged, little signs of usage". After filling the missing values of "condition", the remaining observations which contain missing values were dropped.

## 2.3 Exploratory Data Analysis

### 2.3.1 Price

The distribution of the variable to be estimated - the price - is analysed below. By means of visualisation of the distribution, a right skewed distribution of prices can be detected. Further checks are waived, since the right skewness is clearly recognisable and is typical for variables with a natural lower limit (i.e. zero) and a theoretical infinite upper limit.

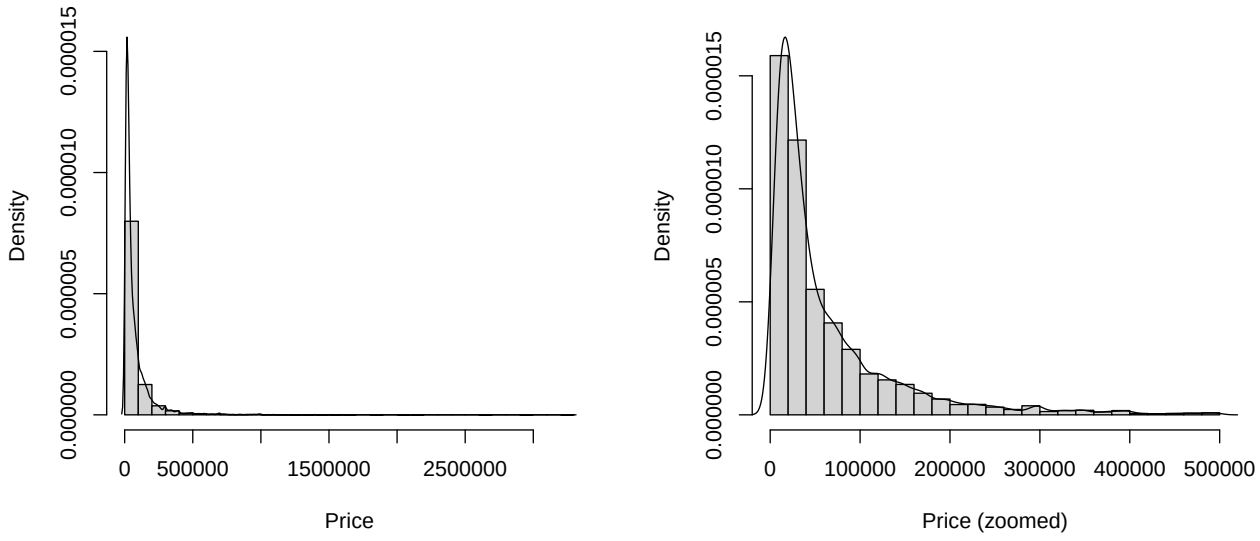


Figure 1: Distribution of prices

While this is, for example, not a problem for decision trees, linear models (among others) are sensitive to outliers and skewed data. For this reason, a log transformation was applied to approximate the distribution of prices to the normal distribution. The graph below shows that the distribution of prices after applying the transformation indeed is more similar - although not perfect - to the normal distribution.

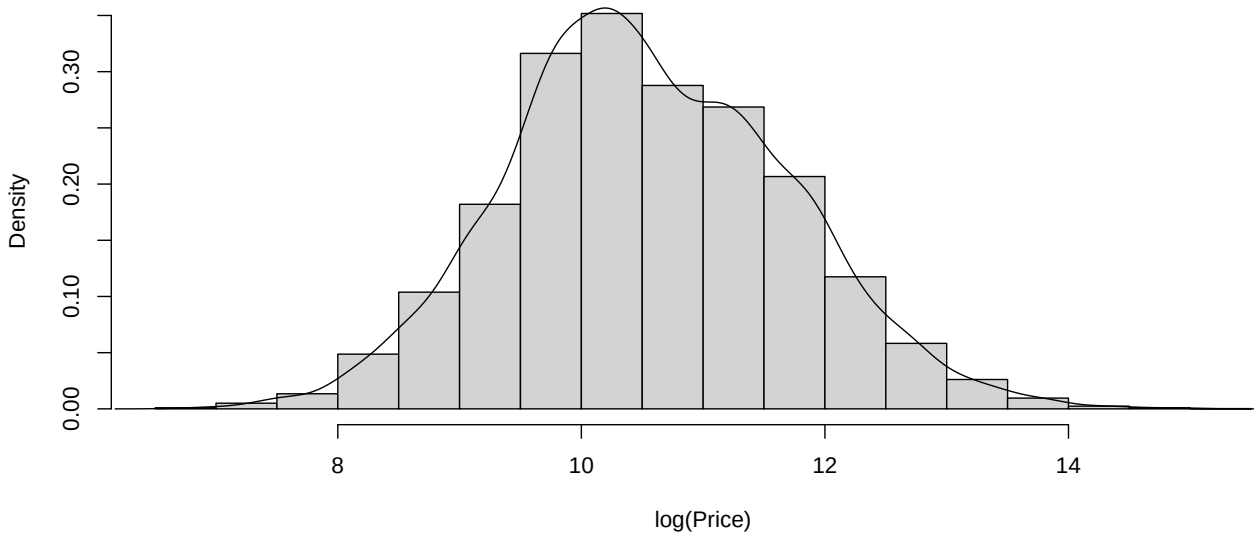


Figure 2: Distribution of prices (logarithmically transfomated)

### 2.3.2 Manufacturer

The following graph shows the 15 manufacturers from which the most vehicles are advertised on classic-trader.com.

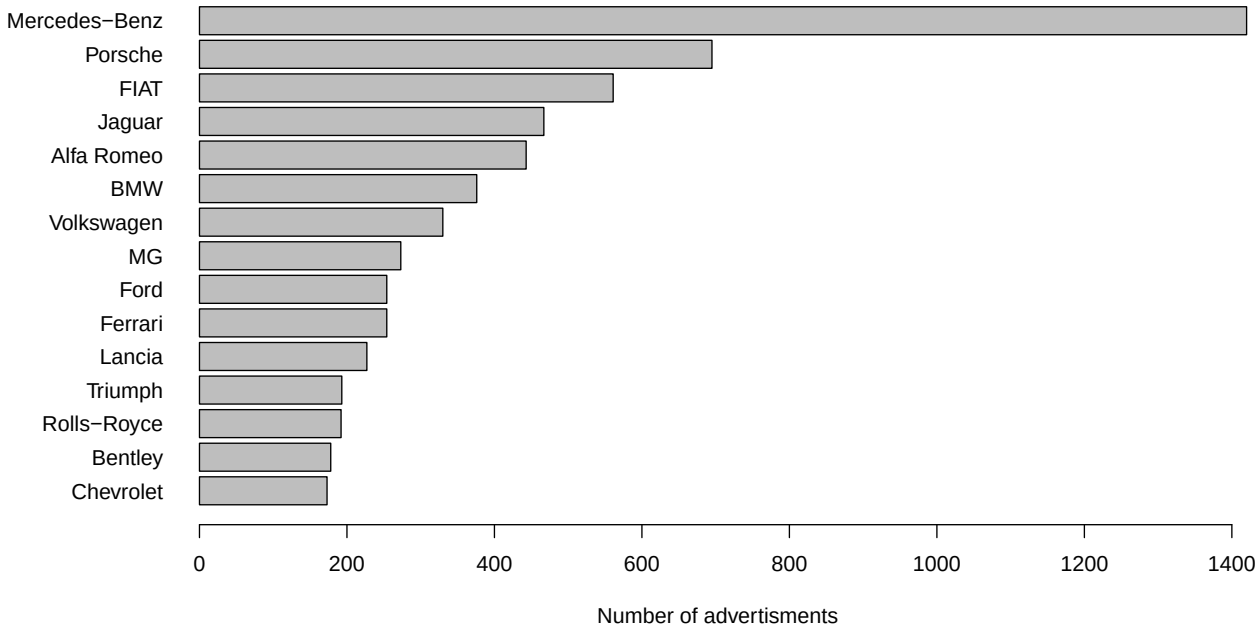


Figure 3: Top 15 manufacturers in dataset

As can already be expected, the manufacturer of a vehicle has an influence on the price of the corresponding vehicle. Thus, different manufacturers produce vehicles of different classes and thus also of different price ranges. The distribution of prices thus differs between the different manufacturers, which can be seen in the following graph of the ten most represented manufacturers in the data set.

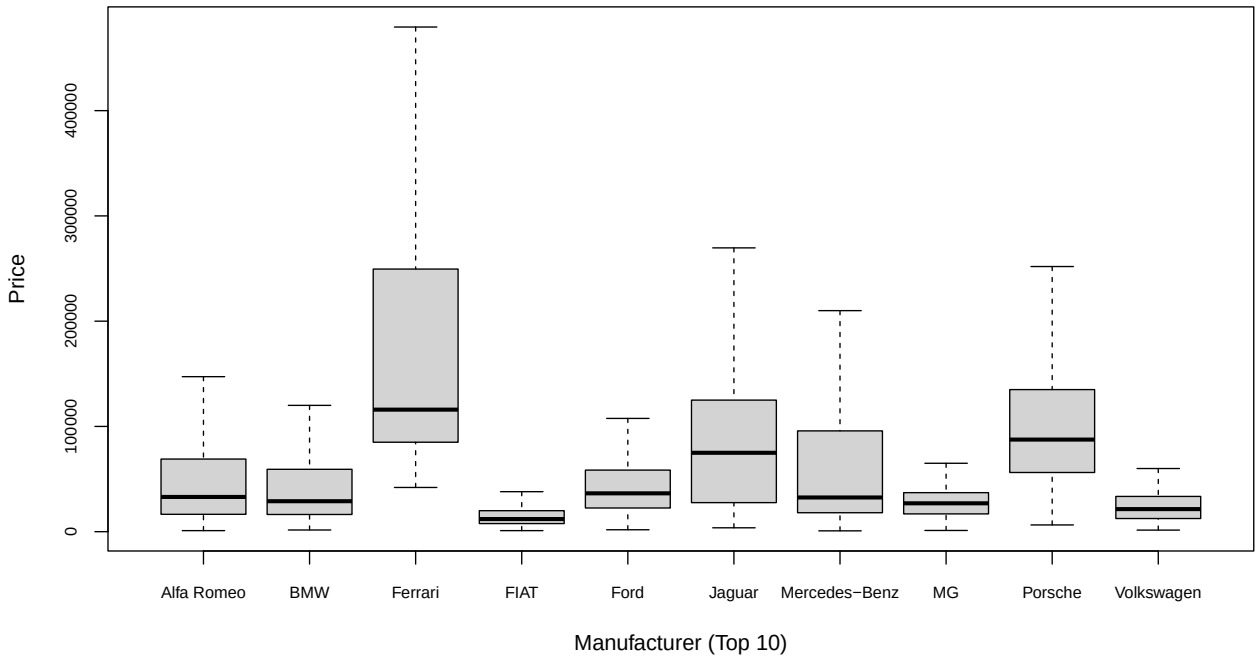


Figure 4: Prices by manufacturers

### 2.3.3 Mileage

In the used car market, the mileage of a car is an important indicator to determine the value and the condition of a vehicle. Although it is not possible to draw direct conclusions about the condition and value of the vehicle, it nevertheless provides a tendency in this respect. In formal terms, correlation but not necessarily causality is expected between condition/value and mileage of a vehicle.

The graph below shows the relationship between the price and the mileage of the cars advertised on classic-trader.com. Note that the data for creating the following plot was filtered to be under 1 million Euros so that the relationship is more clearly visible.

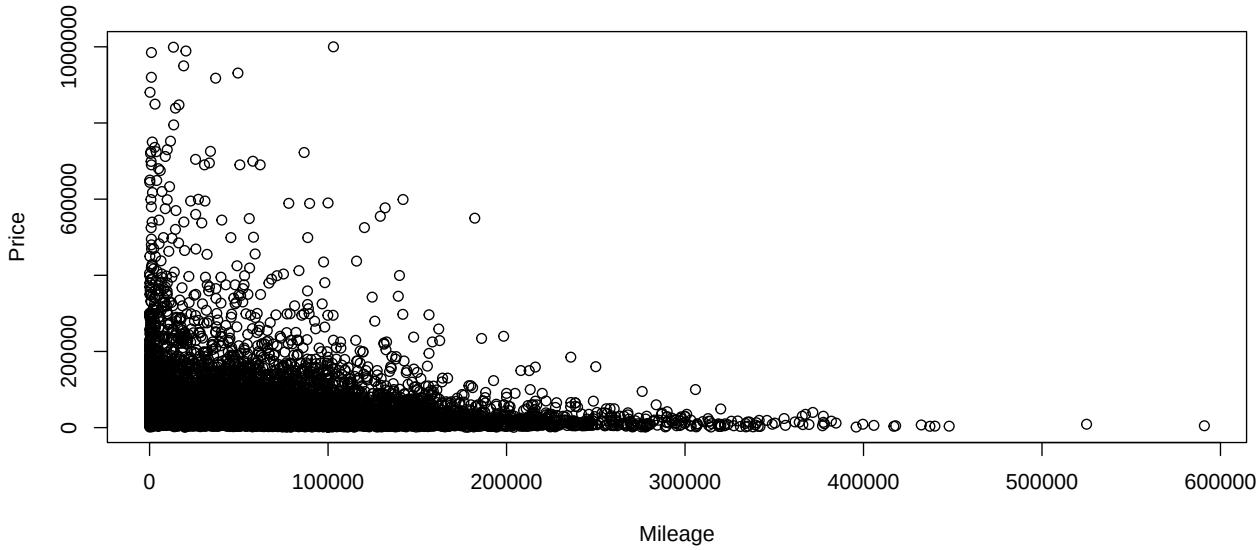


Figure 5: Mileage vs. Price (Price filtered $< 10^6$)

As we can see, there is a relationship between price and mileage which could be used as an indicator when estimating prices. Unfortunately, there are a lot of missing values for mileage in the data set, which is why this indicator is not considered in the analysis. For further studies however, it is recommended to include this indicator in the models.

#### 2.3.4 Year

In order to better understand the structure of the data, the distribution of the production years of the advertised vehicles are shown below. It is interesting to note that major historical events are discernible in the distribution of production years. For example, one sees a major slump in the period of World War 2 (1939-1945) and during the oil crisis of 1973. This can be relevant for this study insofar as car manufacturers are affected by what is happening in the world and in the global economy and therefore are factoring this into their business decisions when producing and developing new cars.

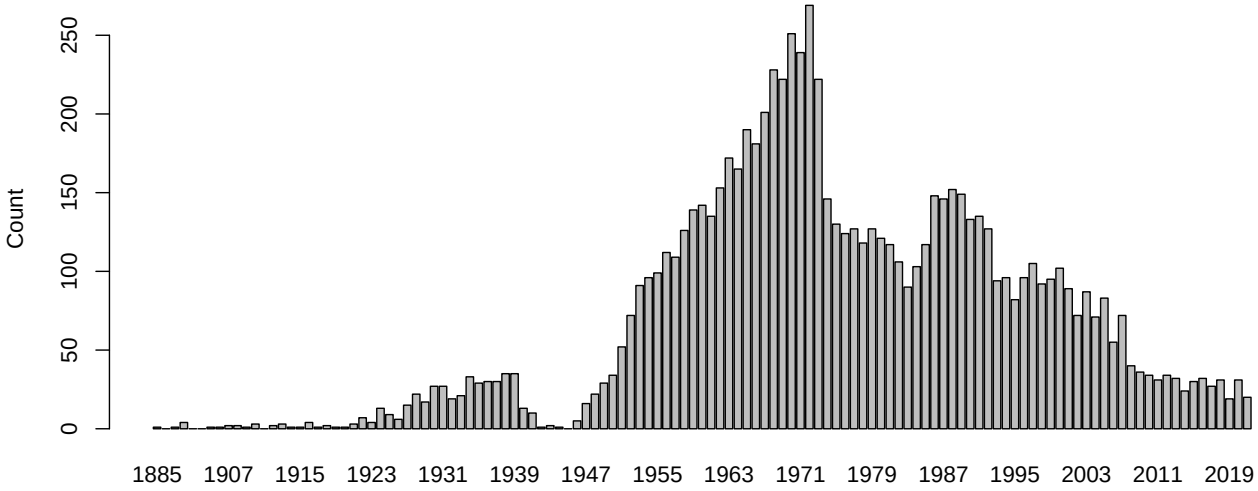


Figure 6: Year of manufacture

#### 2.3.5 Power and engine displacement

Since both variables are integer, the distributions are analysed. In both cases, a tendency towards right skewness can be observed, although not as clearly as in the distribution of prices as seen before. The best result when trying to make the distribution of the variables conform to the normal distribution was obtained with the box cox transformation. Although this transformation does not provide perfect results, an improvement in the performance of the models was nevertheless observed.

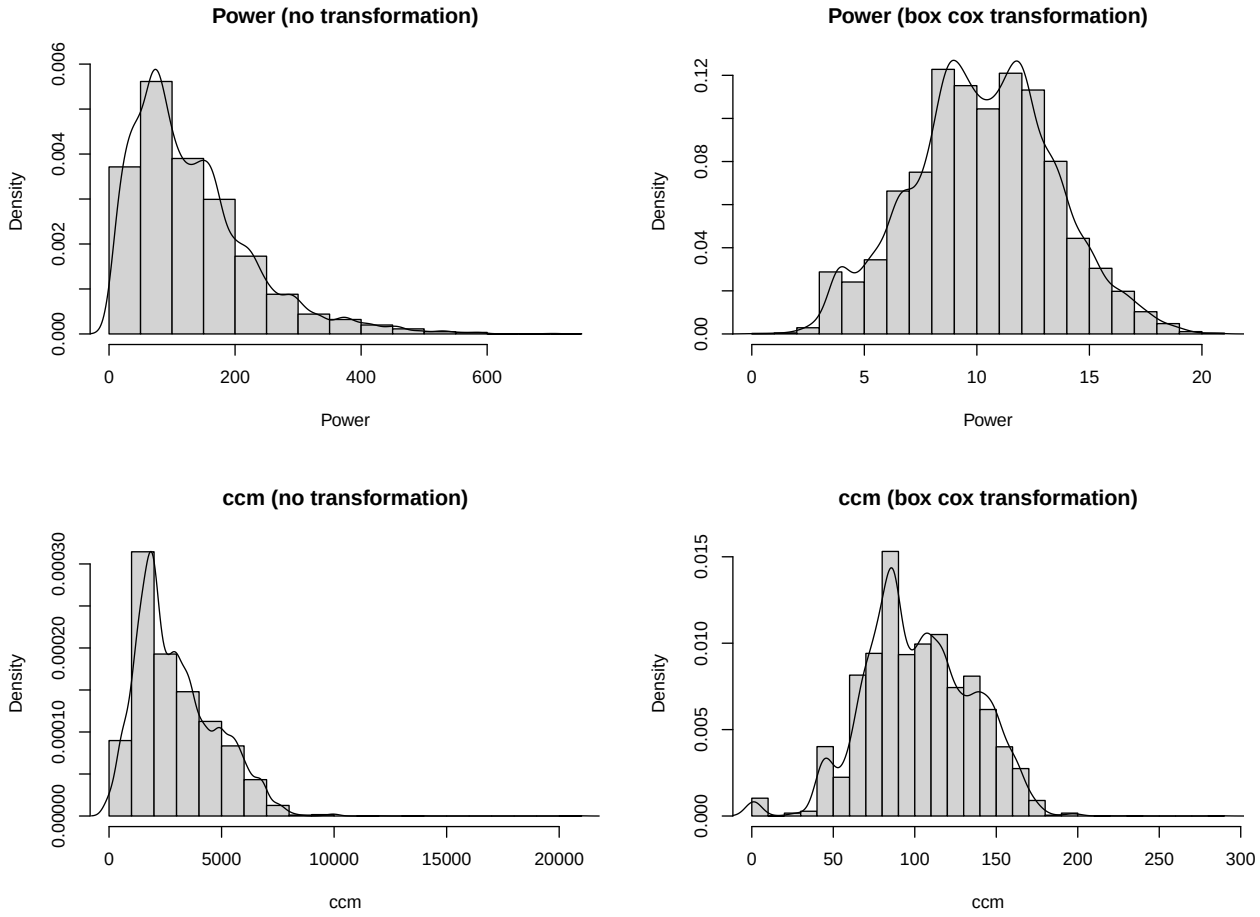


Figure 7: Distribution of Power and ccm

### 2.3.6 One-hot encoding

A large part of the variables considered in this project correspond to categorical variables. Since some algorithms cannot work with categorical data directly, the data was also encoded using the one-hot encoding method, since the categories have no ordinal relationship.

# 3 The model

## 3.1 Performance Metrics

The performance metrics used in this project are briefly described below. Thereby. $\hat{y}_n$ corresponds to the predicted value of the dependent variable $y_n$ for the $n$th observation of $N$ observations.

### 3.1.1 Root Mean Squared Error (RMSE)

Root Mean Squared Error calculates the average of the squared errors across all samples and takes the square root of the result. The RMSE provides an error measure in the same unit as the target variable, which makes RMSE as a performance metric directly interpretable. In general, a lower RMSE is better than a higher. The square avoids the errors of cancel each other out and penalizes larger errors more than smaller ones.

$$\text{RMSE} = \sqrt{\frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{N}}$$

### 3.1.2 Mean Absolute Error (MAE)

The Mean Absolute Error calculates the absolute value of the errors (since the direction of the error is not relevant for this study) and takes the average of these values. This also prevents the errors of cancelling each other out. In contrast to the RMSE, MAE does not penalize larger errors more severely. The error is also measured in the same unit as the target variable, which allows direct interpretation. Likewise, a lower value is better.

$$\text{MAE} = \frac{\sum_{n=1}^{N}|y_n - \hat{y}_n|}{N}$$

### 3.1.3 Mean Absolute Percentage Error (MAPE)

The Mean Absolute Percentage Error is calculated similarly to the Mean Absolute Error, but the error is shown as percentage, which allows to understand how off a prediction is in relative terms. Here, too, a lower value is better.

$$\text{MAPE} = \frac{1}{N}\sum_{n=1}^{N}\left|\frac{y_n - \hat{y}_n}{y_n}\right|$$

## 3.2 Developing the model

### 3.2.1 Linear Regression

First, a linear model is estimated using all predictor variables. To evaluate the model, the diagnostic plots are analysed.

1. Residual vs Fitted:
   Since equally spread residuals around the horizontal line without distinct patterns can be found, it is a good indication that there are no non-linear relationships in the data that could not be captured by the linear model.
2. Normal Q-Q:
   With the Normal Q-Q plot it can be determined that the distribution of the residuals corresponds to a distribution with “heavy tails”, which indicates that the data has more extreme values than expected if they truly came from a Normal distribution. The coefficient estimators should still be quite reasonable, however, the prediction intervals are likely to be too short (since they do not account for heavy tails), which is expected to worsen the performance of the model.

3. Scale-Location:
   This plot shows that the residuals are equally spread along the ranges of predictors, which indicates that the residuals have a constant variance. If this was not the case, less precise coefficients would be expected.

4. Residuals vs Leverage:
   Since no dashed red lines are visible on the graph, which would correspond to a Cook's distance of 0.5 and 1, none of the observations lay outside these boundaries and therefore no action is required.

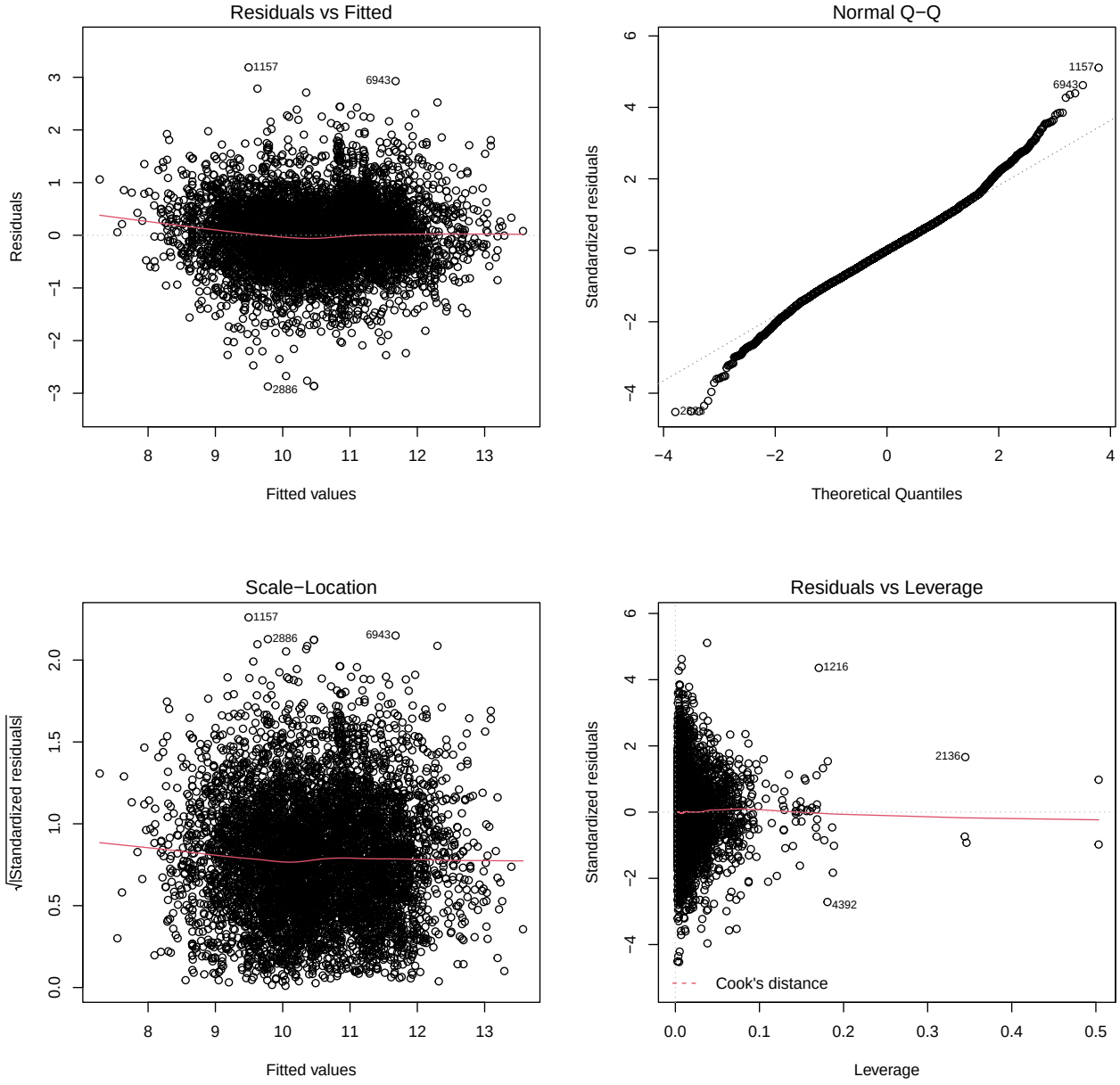


Figure 8: Linear model

When applying the linear model on the training set, following measures are reached:

| model | RMSE | MAE | MAPE |
|---|---|---|---|
| linear regression | 124722.8 | 39567.06 | 63.33423 |

### 3.2.2 Random Forest

Second, a random forest algorithm is trained. In a first step, the best value of mtry - the number of variables randomly sampled as candidates at each split - is sought. The best value of mtry minimized the minimum out of bag (OBB) error.

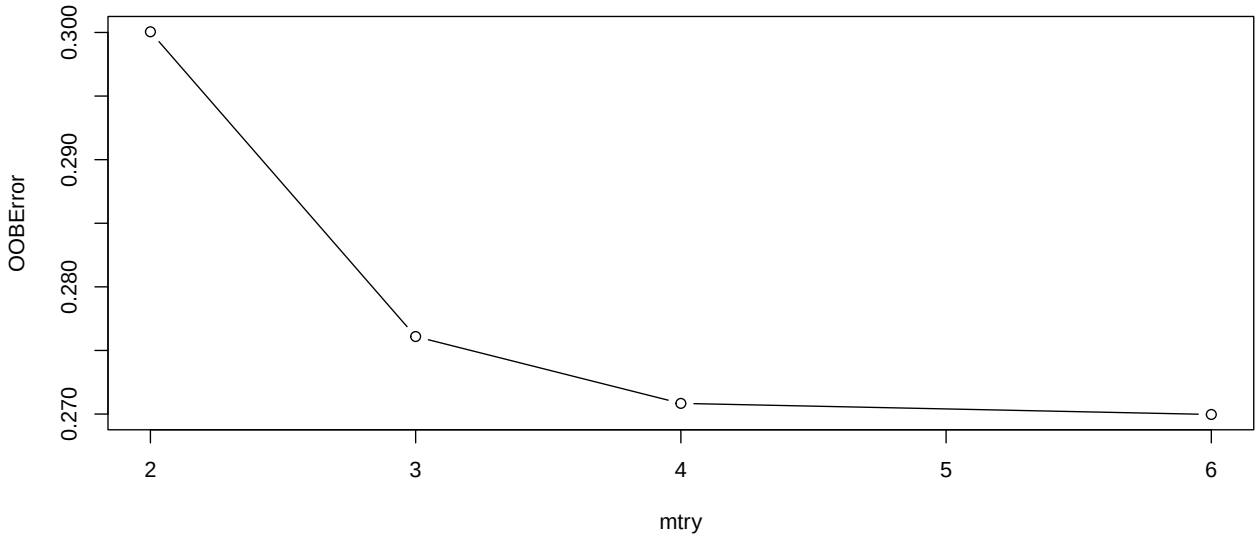


Figure 9: Random Forest Tuning

Next, the random forest algorithm is trained with the previously determined best value of mtry (here: 6). In the following graph, the variable importance as measured by a Random Forest can be seen.

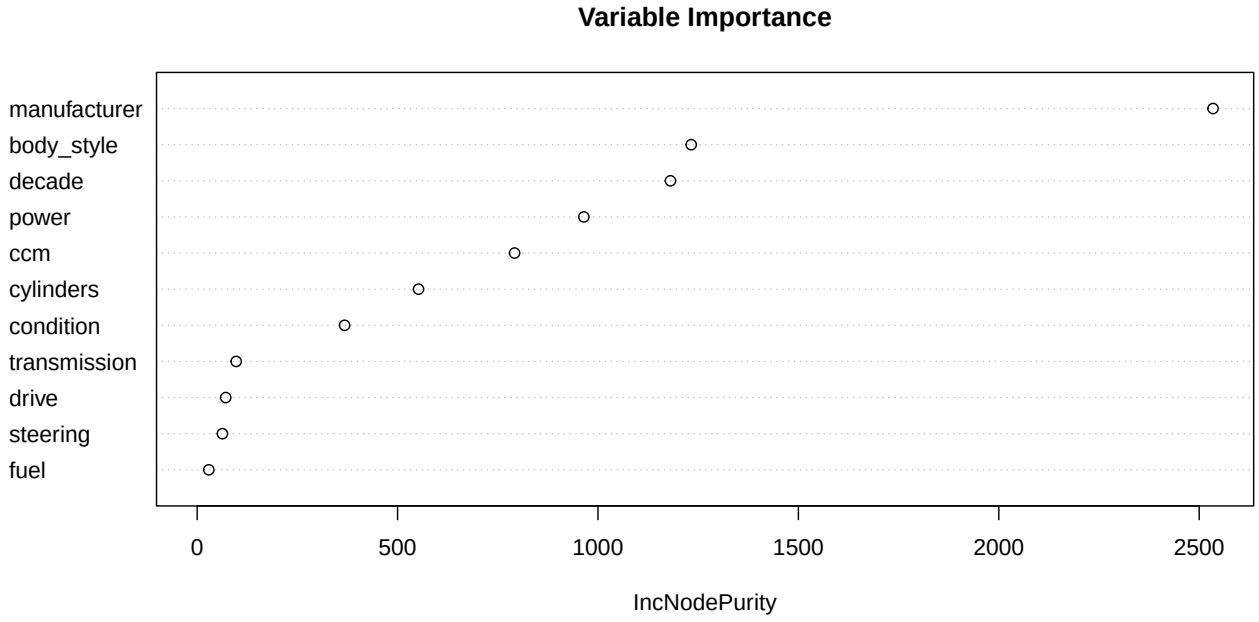


Figure 10: Variable Importance of Random Forest

When applying the random forest algorithm on the training set, following measures are reached:

| model | RMSE | MAE | MAPE |
|---|---|---|---|
| linear regression | 124722.8 | 39567.06 | 63.33423 |
| random forest | 94013.9 | 27573.68 | 46.10535 |

### 3.2.3 K-nearest Neighbors

RMSE was used to select the optimal model using the smallest value. The final value used for the model was k = 5.

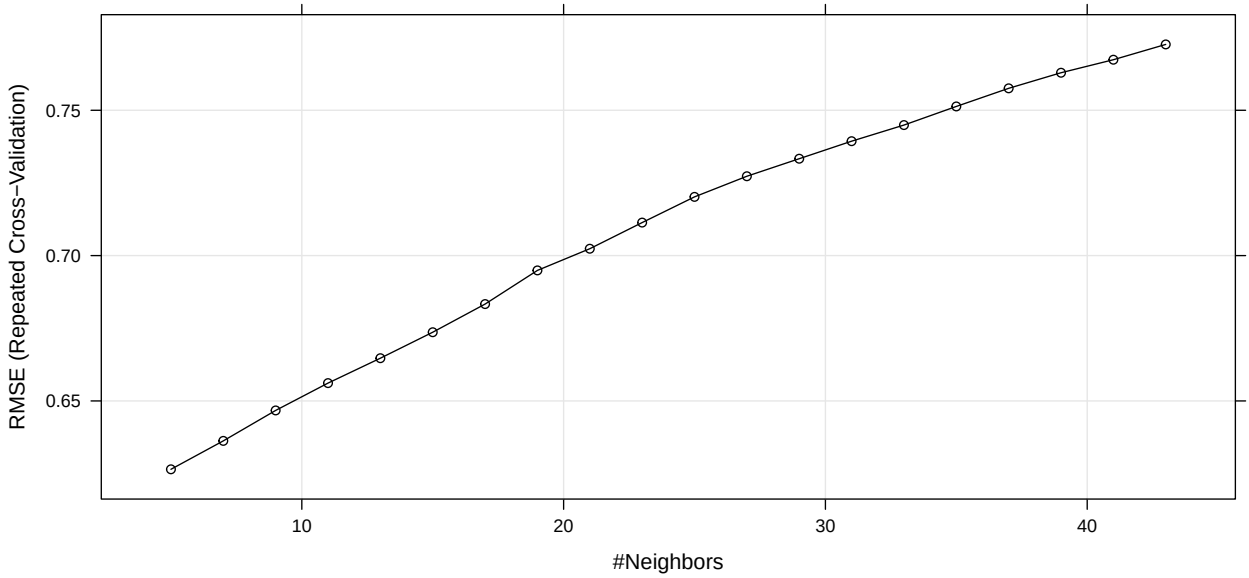


Figure 11: Tune K-nearest neighbors

The following table shows the variable importance of the K-nearest neighbors algorithm

| | Overall |
|---|---|
| power | 100.00000 |
| ccm | 69.16202 |
| body_style.Saloon | 66.65241 |
| cylinders.4 | 55.80344 |
| drive.Front | 45.69647 |
| body_style.Coupe | 39.32848 |
| manufacturer.FIAT | 34.06515 |
| cylinders.6 | 31.85660 |
| drive.Rear | 28.45789 |
| manufacturer.Porsche | 27.22387 |

When applying the K-nearest Neighbor algorithm on the training set, following measures are reached:

| model | RMSE | MAE | MAPE |
|---|---|---|---|
| linear regression | 124722.83 | 39567.06 | 63.33423 |

| model | RMSE | MAE | MAPE |
|---|---|---|---|
| random forest | 94013.90 | 27573.68 | 46.10535 |
| K-nearest neighbors | 95239.96 | 30456.70 | 54.79448 |

#### 3.2.4 Ensemble model (averaging)

Finally, the individual predictions of the different models are combined with the calculation of the arithmetic mean to form a single prediction.

When applying the ensemble model on the training set, following measures are reached:

| model | RMSE | MAE | MAPE |
|---|---|---|---|
| linear regression | 124722.83 | 39567.06 | 63.33423 |
| random forest | 94013.90 | 27573.68 | 46.10535 |
| K-nearest neighbors | 95239.96 | 30456.70 | 54.79448 |
| ensemble model (average) | 100680.40 | 30028.67 | 50.97492 |
| ensemble model (average, without linear model) | 93670.71 | 28034.08 | 48.84203 |

### 3.3 Applying the model

As must be noted with the development set, the models unfortunately do not provide reliable predictions of the prices. Nevertheless, the models are applied with the validation set in order to further analyse the results.

Since the random forest algorithm outperforms the other models in both MAE and MAPE and the ensemble model that does not include the linear model outperforms the other models in RMSE, these two models are used with the validation set. With the data from the validation set, following results are obtained:

| model | RMSE | MAE | MAPE |
|---|---|---|---|
| random forest (validation set) | 56968.03 | 23392.77 | 40.18039 |
| ensemble model without linear model (validation set) | 59365.98 | 25032.26 | 45.59154 |

The following plot shows the distribution of the relative errors made by the two models:

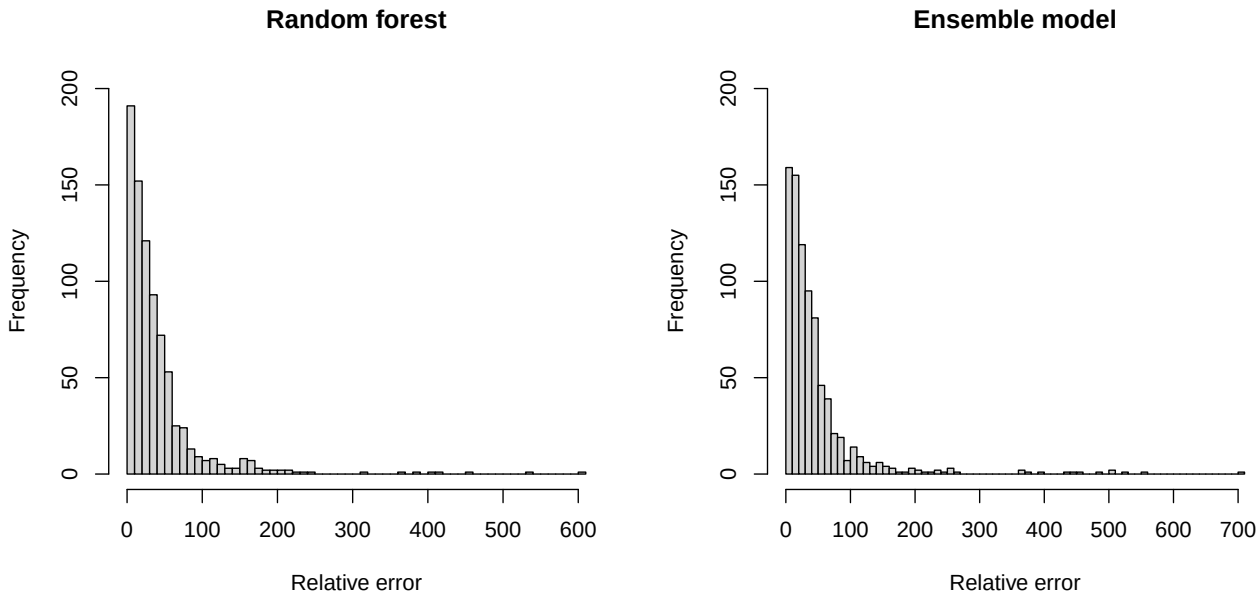


Figure 12: Distribution of relative errors

As can be seen in the table and the plot, random forest seems to make better predictions - even if they are not really of much use. What is much more interesting to note is that the distribution of the errors made by both models are heavily skewed to the right. Thus, there are few estimates with very strong deviations, which degrade the observed performance of the models very strongly.

# 4 Conclusion and Limitations

## 4.1 Data Quantity and Quality

With around eight thousand observations, the analysis does not have very much data available for developing the models. With an increasing number of observations, the performance of the models could be expected to improve.

As was noticed when applying the models with the validation set, very high relative errors are made in estimating prices for some observations, which is why some estimates with very high deviations were investigated. It was found that some of these (very) large errors are due to the assumption which was made when filling the NAs in the variable "condition" (i.e. filling NAs with "original"). For example, an Austin-Healey 100/4 (BN1) was estimated to have a price of 100'546.10 euros, whereas the vehicle was advertised with a price of 15'850 euros (this represents a relative discrepancy of 534.36 percent). At the time of writing, the average price of this vehicle was 78'266.61 euros (calculated across all conditions of vehicles), with some vehicles being advertised with a price as high as 155'383 euros. The error of the estimate in relation to the average price corresponds to 28.47 percent, i.e. significantly lower than the 534.36 percent. It was found that the observation with the very large difference was incorrectly attributed the condition "original", although the vehicle is merely a rusty bodyshell without an engine (remember: original condition describes the condition of the car as unchanged with little signs of usage)[1].

To improve the model, the assumption in filling the condition-variable has to be reconsidered. Using the data, which can also be obtained via the web-scraping process, one could also try, for example, to use text classification with the description of the vehicle (unsystematic text written by the seller) to infer the condition. With the available data, it would even be possible to attempt to classify the condition using image recognition with the pictures of the vehicles which are uploaded by the seller.

## 4.2 Omitted Variable Bias

The prices of vehicles, especially classic and exclusive vehicles, are often influenced by various other variables which are not taken into account in this project. Possible examples that could significantly influence prices of a vehicle and that were not taken into account in this project are successes in Motorsports of the vehicle model or the cult status of a vehicle, established for example through appearances in movies or through ownership by various celebrities.

For example, the BMW E30 M3 (built from 1986-1991) is traded at comparatively high prices, which is certainly partly due to the car's success in Motorsport and the reputation it has built up as a result: it is often entitled as the "most victorious racing car of all time" with 1'436 victories in just 1'628 days.

Another example is the VW Type 2 T1 (the "Hippie bus" built from 1950-1967), which is traded at comparatively high prices (in some special versions in top condition, prices reach six figures). One possible reason for the high prices is that this vehicle is a symbol of the counterculture movement of the 1960s and stands for cultural openness and diversity. Even today, this vehicle conveys a certain attitude of life and stands for freedom, which is why the vehicle is also used as a symbol in modern advertisements, for example, precisely because of this characteristics.

For the reasons mentioned, success in Motorsport, for example, could be introduced as a variable (e.g. number of wins as ratio of participation in races). Emotional factors, such as the cult status, are more difficult to capture.

## 4.3 Interaction Effects

In addition, some features of a vehicle may have positive effect on the price of a vehicle in some cases and a negative effect in others. For example, the impact of a right-hand drive is expected to be positive in the case of a Japanese vehicle, as this is (in many cases) due to the fact that the vehicle was built for the Japanese

[1]The advertisements can be found with the ID, for exmaple contained in the object "analyze_results_final_rf", using the search function on classic-trader.com

market, which may increase collector's value, whereas a right-hand drive might have a negative impact on the price of a German vehicle

As another example, it would also be expected that the number of cylinders in an engine tends to increase the value of the corresponding vehicle. However, it is possible that for some vehicle models a smaller engine is more popular, for example due to reliability, which means that the model with the smaller engine is priced higher.

A possible solution to this potential source of error could be the introduction of interaction-terms, for example between steering ($x_1$) and country of origin of the vehicle ($x_2$):

$$\hat{y} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$