USED CAR VALUE PREDICTION PROJECT

Problem Statement:

Car Value Prediction Is Essential for Individuals and Companies

Each day, thousands of pre-owned cars are sold worldwide. Prediction of the second-hand vehicle price provides an important benchmark to both private buyer and the seller as well as business professionals such as car dealers, lenders and insurance companies.

Banks need to know the exact value of second-hand vehicles as they are mostly lienholders or they are transferring the loan from one person or another. Insurance companies alike need to be able to assess the value of the pre-owned vehicles, since they will be calculating premiums when they are making their risk assessment.

The used car market is also a large and strategically important market for car manufacturers since it is closely connected to the new car business. Trading-in used cars in new car retail sales and handling lease returns, repossessions and fleet returns from car rental companies necessitate car manufacturers to engage in the used car market. Therefore, car makers require sophisticated decision support systems to sustain the profitability of the used car business.

The necessity of prediction paved the way for now well-established companies like Edmunds, Kelley Blue Book, NADA Blue Book. These companies utilize statistical models on massive databases and they use machine learning algorithms to effectively predict the value of innumerable car brands and models, answering the market demand.

Dataset:
The dataset used for this project was retrieved from Kaggle.com. The original dataset is a German used car sales website and each data point is an advertisement placed by an individual. The dataset is therefore quite arbitrary, with great percentage of missing data, most of which cannot be completed with interpolation or other conventional fill methods. Once all erroneous data points and non-fillable missing values have been deleted, approximately 232 thousand data points out of 377 thousand data points remained, which is 62% of the original data. Data fidelity was preferred over the number of data points remaining, since data precision heavily affects model quality.

Initially, all strings were converted to uppercase and insignificant columns have been removed. Some variable names and elements had to be translated from German to English to provide ease of use.

A multi stage filtering was introduced to remove erroneous data. Since the data frame was prepared by non-professional individuals, wrong entries such as 20000 horsepower engines or 0€ car values were spotted. Some cars were significantly undervalued, and some were unreasonably high priced. Some online research of normal prices together with Exploratory Data Analysis by drawing graphs helped spot those outliers immediately. Special attention had to be paid to vehicles below 1000€, since it is common practice for dealers to rent cars rather than selling them at this price range, and illegally use this platform to place such ads.

Registration year values as low as 1800, and higher than 2016, had to be removed by a two-stage filter, since it is impossible to have registration date after the data crawl date.

After first stage of cleaning a function to fill missing model information was coded. The algorithm created populates a brands-models dictionary and finds all entered model values for the corresponding brand by searching the ad title for keywords. While proving useful, this

method pointed out another weakness of the dataset: Non-German vehicles were generally grouped under “sonstige autos” name, which means “other cars” and many of their models were labelled “andere”, meaning “others”. In order to reduce the deleted number of rows, the brands-models dictionary was augmented by adding foreign brands and their corresponding models. The retriever function was then run again and it significantly reduced the number of missing values.

Retrieving the trim data, engine capacity, and other strings of importance would be valuable. This, however, would exceed the limits and scope of this project and require data mining algorithms and NLP, thus was not utilized.

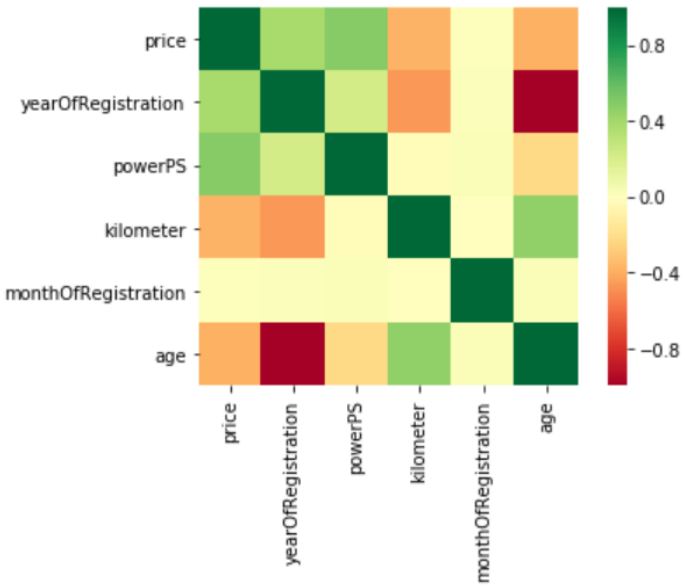
After the code written to retrieve data from the name column to fill missing values in brand and model columns is executed, there will be still some missing values. These are in transmission, vehicle type, fuel type. There are no intermediate values for these columns, so interpolation would not work. Forward or backward filling would create erroneous data. Omitting those rows was the best decision for the sake of the model.

Some missing values in type of vehicle and type of fuel columns cannot be filled, since some vehicle/models have various body types and engines, and it would lack precision to assert if a missing value for a specific brand/model body type is sedan or hatchback, since there is no way to find out the actual value. For example, a Mercedes Benz E Klasse has both a sedan and station wagon body type, and both diesel and gasoline engine, so if these values were missing the whole row was discarded, since it is no more than a blind guess trying to find those values.

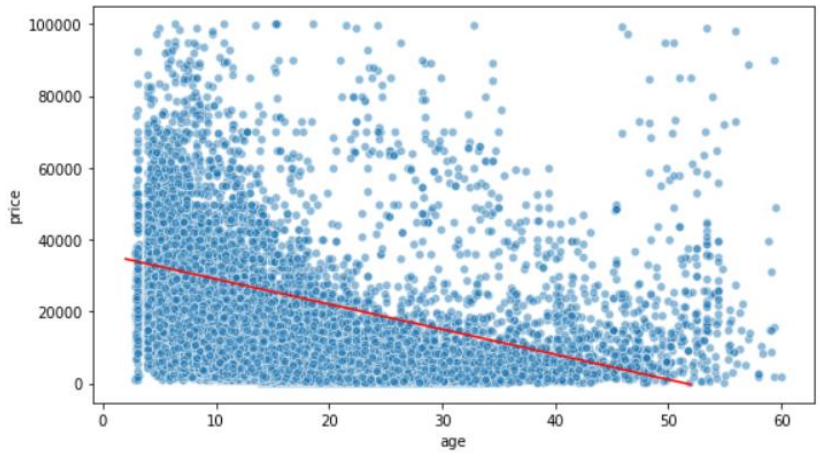
After running coded algorithms to fill missing values, another EDA was carried out, and remaining outliers have been eliminated.

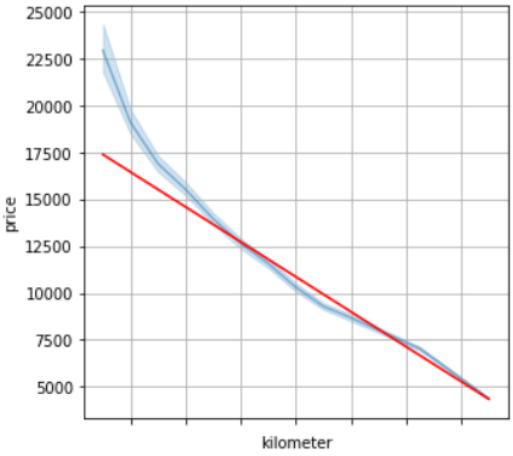
Initial findings from exploratory analysis:

After cleaning the data, it was time to analyze the prominent factors determining the value of the used vehicle. The heat map generated to display the correlation coefficients can be seen below. An internal algorithm coded into Pandas automatically blocks out columns that are not continuous numeric variables, leaving out transmission type, fuel type etc. These will be handled as a part of the classification problem in the machine learning phase.

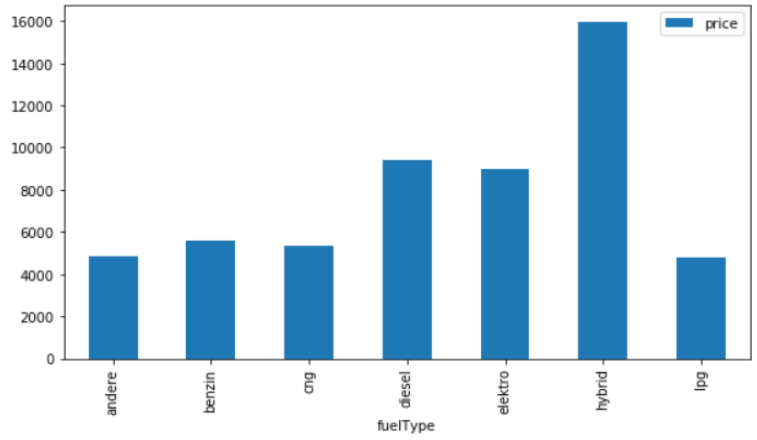


To further investigate the parameters affecting the price, more graphical analyses were made.
The red line represents the regression, in this (age, price) graph below. The regression line was shifted upwards in order to overlay. The trend is clearly visible, the age and price are negatively correlated.



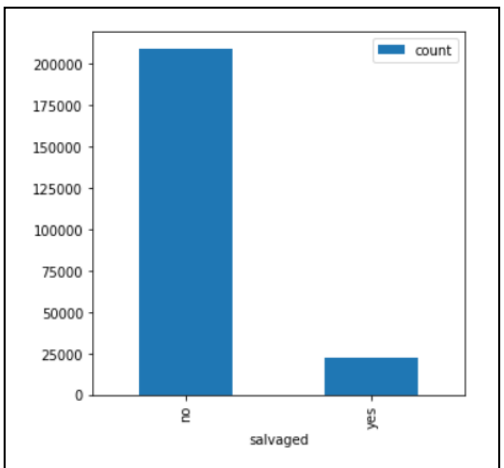


The (kilometer, price) graph above right reveals that the price is inversely proportional to the mileage. This information is also backed by the regression line, depicted in red, and the negative sign of the Pearson coefficient, which has been calculated as -0.37.



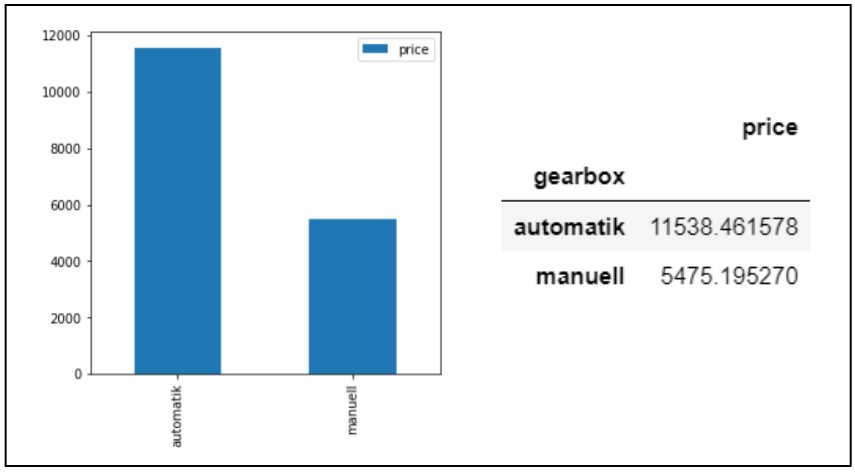
fuelType	price
andere	4886.744681
benzin	5583.599862
cng	5340.674757
diesel	9406.093050
elektro	8980.950000
hybrid	15973.527397
lpg	4773.563023

The price according to fuel types have been depicted above. The fact that hybrid vehicles are the most expensive proves to be because of this reason: the only electric cars available at second hand market -thus this database- were ultra-compact city cars that has front row of seats only. Those vehicles are very small, therefore not so expensive. Many brands have hybrid models, that have cutting age technology to curb down fuel consumption, making them considerably higher priced. Diesel vehicles burn less amount of fuel and diesel is considerably cheaper than gasoline in Europe, therefore these cars are preferred over gasoline ones and they are more expensive.



salvaged	price
no	7345.329905
yes	2584.351481

The chart on the left shows the relation between the salvaged status of vehicles and price. The average price of a salvaged vehicle is almost three times lower than an accident free vehicle, which is

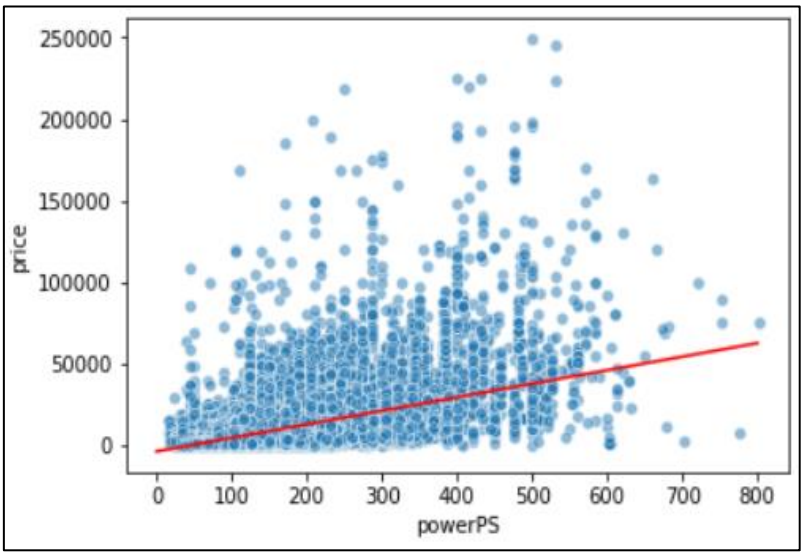


gearbox	price
automatik	11538.461578
manuell	5475.195270

quite obvious, since salvaged vehicles are less preferred. The bars represent the number of salvaged cars

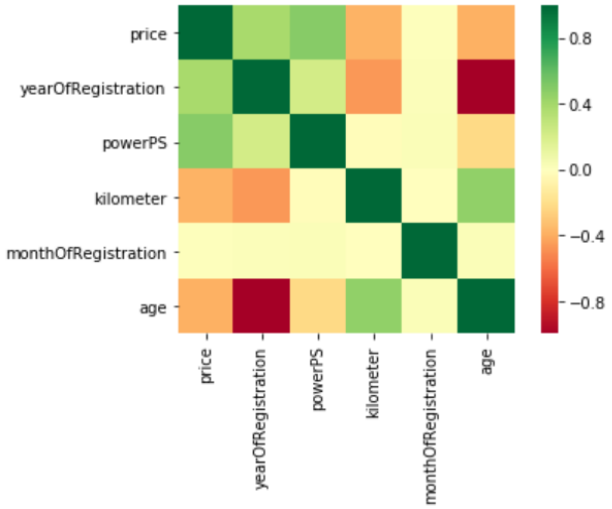
Automatic transmission vehicles are twice as expensive as manual transmission vehicles. Automatic transmission is more sophisticated, therefore more expensive. Moreover, some luxurious models only come in automatic transmission.

Exploratory Data Analysis:



After cleaning the large dataset and discarding all unwanted and values, the data frame is ready for analysis. Since the dataset variables are mostly binary such as the “salvaged” value or mostly strings, such as “brand” and “model”, but not continuous numeric variables, the size of the covariance matrix where we see the variable correlations will be limited. The heat map generated to display the covariance can be seen on top, next page. Please note that the algorithm processes only continuous numeric variables, leaving out transmission type, fuel type etc. Binary or discrete variables will be handled as a part of the classification problem in the machine learning phase. In this project, correlation was measured by Pearson coefficient r, which shows strong correlation between continuous numeric variables if r = 1, no correlation if r=0, and strong negative correlation if r=-1.

Once we look at the price row of the heatmap, we can see that there is a significant positive correlation between a car’s value and year of registration and power of the engine, with the latter being more prominent. The Pearson correlation coefficient between price and year of the vehicle is 0.39 while the same coefficient for power and price is as high as 0.5, which signifies a very strong correlation, whereas the correlation



coefficient between kilometer and price is -0.37. The variable “age”, which was created to include the month data to calculate how old the vehicle is, since using only the year of registration would simply ignore that data.

Two-sided T tests were carried out to check if the null hypothesis that claims 2 independent samples have identical average values are true. Similar tests were also carried out by bootstrap method to analyze relationships between prices of discrete variables, i.e. prices of diesel vs manual vehicles.

The tables below depict the t-test executed to check if different engine and transmission types were similar in changing the price. The **null hypothesis of equal averages** is rejected for a p-value less than 0.05.

PRICE	T Statistic	P Value	Result
DIESEL vs GAS	83.050	0.00000000	Reject H_o
MANUAL VS AUTO	120.509	0.00000000	Reject H_o
SALVAGED VS CLEAN	-64.047	0.00000000	Reject H_o

The table below depicts the t-test executed to check if different engine and transmission types were similar in changing the horsepower of the vehicle.

POWER PS	T Statistic	P Value	Result
DIESEL vs GAS	95.365	0.00000000	Reject H_o
MANUAL VS AUTO	249.735	0.00000000	Reject H_o
VW vs SKODA*	-0.385	0.70049115	Accept H_o

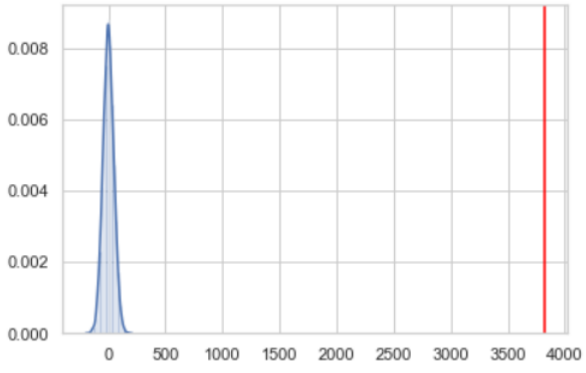
Please note that Skoda, a sister company of Volkswagen has similar values regarding the horsepower, since most vehicles use common engines between these two brands. This was reflected to t-test, and we can clearly see that a random vehicle selected from a pool of VW and Skoda have comparable average horsepowers.

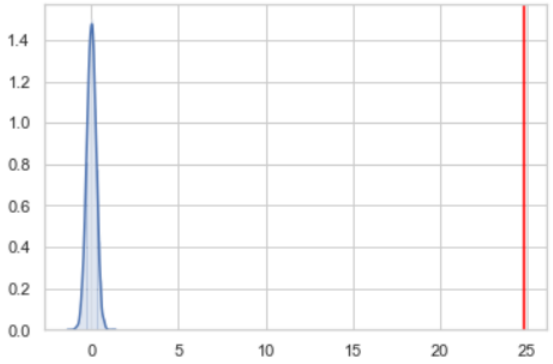
```
two_sample = stats.ttest_ind(df1[df1['brand']=='VOLKSWAGEN'].powerPS,
                             df1[df1['brand']=='SKODA'].powerPS)
print('The t-statistic is %.3f and the p-value is %.8f.' % two_sample)
```

```
The t-statistic is -0.385 and the p-value is 0.70049115.
```

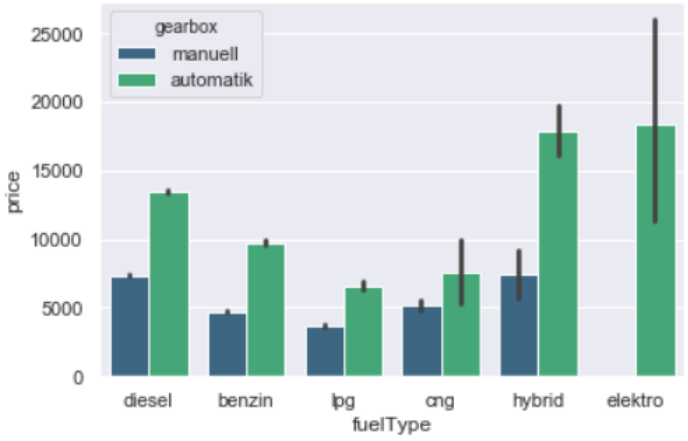
Another approach was using the bootstrap method as discussed above. Looking at the engine horsepower values of both Skoda and Volkswagen, we can see that the empirical difference of means of horsepower of each brand is 0.27965, and as we carry out the bootstrap method we can witness the normal distribution of permuted bootstrap replicates and the red line represents the empirical difference of means, with a p value of 0.347.

The graph on the left below shows permuted bootstrap samples drawn, between diesel and gasoline powered vehicles and their mean prices. The blue shaded histogram is





the distribution curve of these values. The empirical difference of 3822$ is depicted by the red vertical line and can be observed well right of the distribution curve shaded in blue. This result suggests us that the mean price of a diesel vehicle and a gasoline vehicle is not similar. We can also carry out the same test for analyzing the power of engines according to their fuel types. The graph is very similar, with the red line at 24.8 Hp, indicating the mean difference of a diesel vs. a gasoline engine. Diesel engines are simply more powerful and more expensive.



Another possible depiction of how car prices are affected by the engine type is the bar plots as seen in the graph below. It is clearly visible that diesel automatic vehicles are more expensive. Diesel engine is very sophisticated and diesel automatic vehicles are much preferred in Europe, since their cost per mile is very low.

The non-numeric discrete variables are a part of the classification problem in machine learning, thus prevalent features were identified.

Data Analysis Conclusion:

A large dataset of automobile sale website was cleaned, filtered and analyzed. The most prevalent factors are power, year of registration and kilometer, significance of which have been pointed out in this project so far. These initial findings are to be applied together with the classification data to the prediction model. A revisit to cleaning phase may be required to get better results as required.

Regressor Models:

Each machine learning problem is closely related to the dataset and the problem-solving phase starts with a clean dataset and good understanding of the weight of those parameters in building up the model.

```
1  np.exp(y).describe()

count    2.311570e+05
mean     6.894628e+03
std      1.055829e+04
min      2.000000e+02
25%      1.700000e+03
50%      4.000000e+03
75%      8.900000e+03
max      1.250000e+06
Name: price, dtype: float64
```

The dataset was not prepared by professionals and a rigorous clean-up process had to be carried out to minimize the loss of data and to zero out faulty data feed to the prediction models. The observations being quite arbitrary, coupled with great percentage of missing data posed great difficulty. Missing values, most of which could be normally completed with interpolation or other conventional fill methods had to be picked out deleted, since all were non-fillable missing values. This resulted in loss of approximately %38 percent of the original data. However, despite the spread and diversity of the data, adequate data fidelity was achieved and the extensive data wrangling efforts paid off with satisfactory model outputs.

Dealing with used car values ranging from 200€ to 1.25M € built from 1950's to 2016 was not going to be easy for any machine learning model. The model had to process data that has a price standard deviation of 10558€ and a mean price of 6894€. There are many brands that have similar features but a wide range of price. Initial thought was to implement a linear regression model based on Scikit learn, however the results were not satisfactory, since data consists of both continuous variables and classifiers and linear models cannot have classifiers as input and simply have to ignore them. In order to better model this diverse dataset, next approach was to utilize regression tree models, again from Scikit learn.

The decision tree models proved to be far more successful from the start since they were able to capture non-linear relationships between features and the target: prices, following these steps:

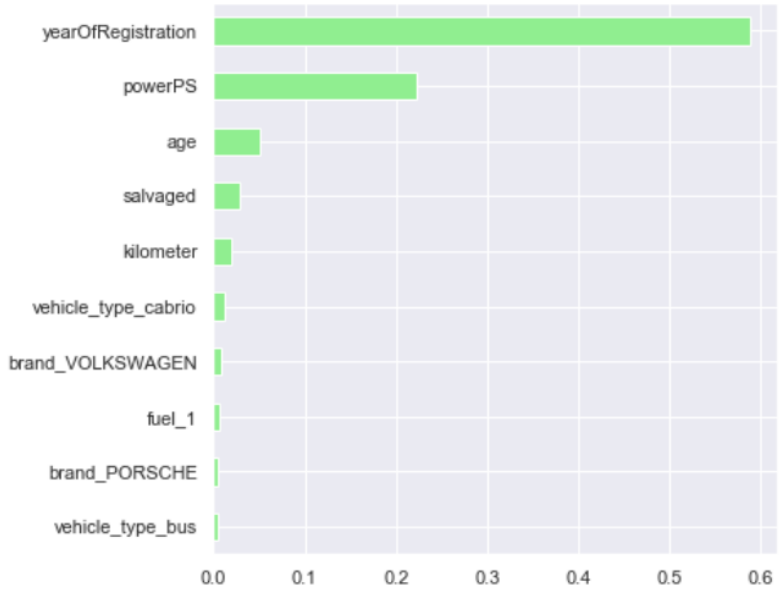
- Dummy variables were created so that categorical features could be inserted to the model.
- Data was split to various test fractions from %20 to %40 by train test split.
- Model was instantiated and then fit on training data.
- Cross-validation of 3-10 folds were applied, depending on processing time.
- Fitted model predicted test set values, the unseen part of the dataset.

- Grid-search and Randomized search and hyper parameter tuning was carried out, processing power and time permitting.

- Sklearn metrics R^2 and RMSE were used for scoring.

Based on the general guidelines above, numerous models including DecisionTreeRegressor, RandomForestRegressor, BaggingRegressor, GradientBoost, Adaboost and Ridge were evaluated. In all of those models, one drawback was immediately identified: the test size was dominant on the quality of the model, which was a strong indicator of some observations being too outlying for the model to explain. Basically, the validation performance or the test accuracy of the model was heavily dependent on where outlying data was. Test score turned out to be higher than train score, which clearly indicated a problem.

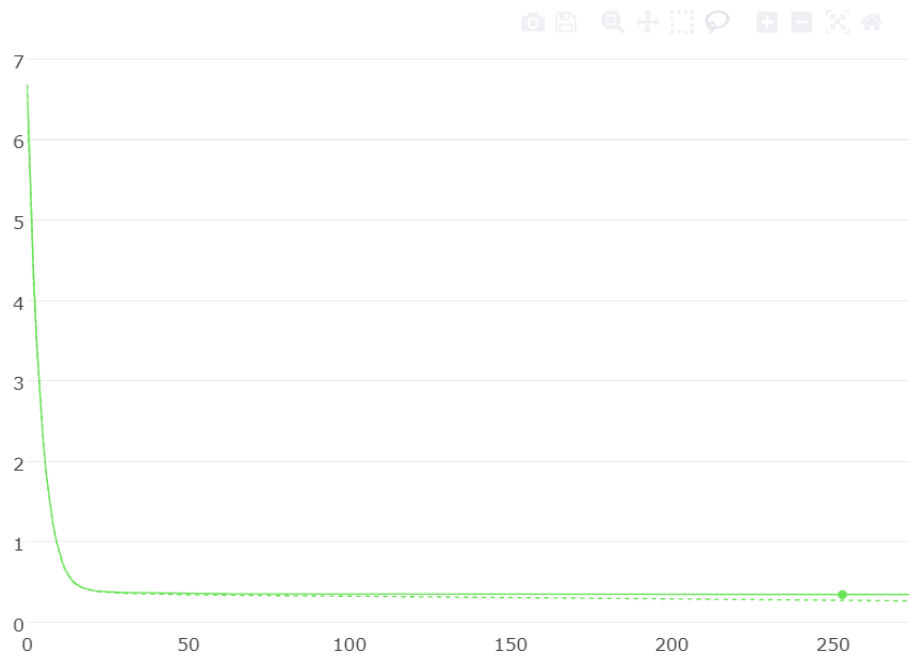
Initial reaction was to revisit the data frame and filter out vehicles older than year 1990 and more expensive than 400.000€. However, these surprisingly did not change the nature of error.



Still, data was sensitive to folding and train test split, only the amplitude was reduced. With further pruning and sacrificing the outlying data points, train test score anomaly was fixed, and after many hyperparameter tuning trials the results' value of 2281€ and test score of 0.917 was achieved with a RandomForestRegressor, which appeared to be the second most successful model so far, surpassed only by GradientBoostRegressor that produced slightly better results:2232€ and 0.920.

Even though the R^2 score of the test data was good, the RMSE value was not satisfactory since it was too large. In order to fix this problem, the price column (labels) of the dataset was transformed to logarithmic scale and the prediction models were re-run. Also, in addition to previous models, LightGBM(LGB) and CatBoostRegressor(CB) were added. Moreover, the last stage of the data filtering were reverted, including very old and very expensive cars, increasing the spread of the data. Surprisingly, after the log transformation, the RF results turned R2 score of 0.893 and an RMSE of 1.43€. Same results for GB were 0.901, 1.417€ . Newly introduced CB and LGB proved to be performing slightly better, with score values of 0.907, 1.41€ and 0.903, 1.41€, without sacrificing any data to filtering.

The other method for normalizing the price variable, box cox power transformation, provided similar results. CB model had an R^2 score of 0.91 and RMSE score of 1.6€



Linear models failed to provide an effective mechanism to predict the price, however decision tree based non-linear models and ensemble models together with boosting models achieved high success.

```
Fitting 5 folds for each of 1 candidates, totalling 5 fits

[Parallel(n_jobs=-1)]: Done   2 out of   5 | elapsed: 16.8min remaining: 25.2min
[Parallel(n_jobs=-1)]: Done   5 out of   5 | elapsed: 17.1min finished
```

Fine hyperparameter tuning, a step that was skipped due to lack of computational power, remains as an area of improvement.

Conclusion:

An arbitrary database retrieved from German online sale website has been retrieved from Kaggle.com. After an extensive data cleaning and manipulation process, data frame was price column has been predicted by numerous non-linear machine learning models. Following normalization via log or box cox transformation, all models were able to predict with root mean square error of less than 2€, with the best regressor error being as low as 1.40€. Error rates are interpreted to be compatible with market demand explained in the *"Problem Statement"* phase. Moreover, extensive hyper parameter tuning would further increase the results.

```
1  np.exp(rmse_test)
```

```
1.4060107218905358
```