USED CAR VALUE PREDICTION PROJECT MILESTONE REPORT

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 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. 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 the subject website 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, and thus, was not utilized.

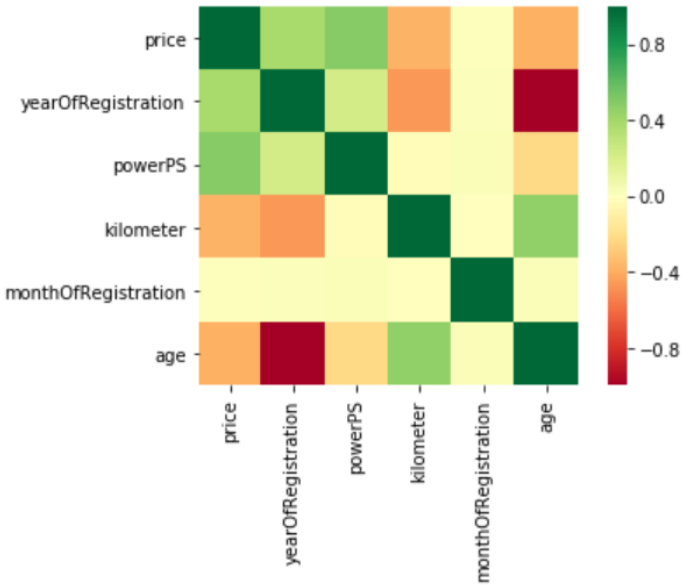
After the code written to retrieve data from the name column to fill missing values in brand and model columns is executed, there will be still some missing values. These are in transmission, vehicle type, fuel type. There are no intermediate values for these columns, so interpolation would not work.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

Also, 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 i.e. 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.

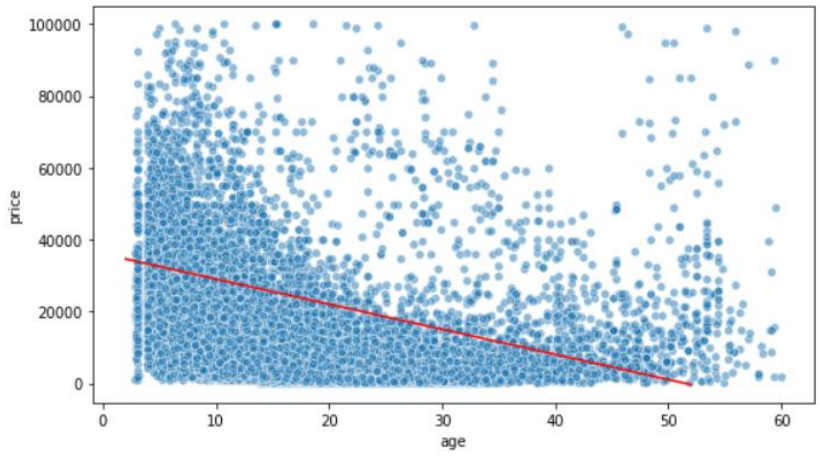
After running coded algorithms to fill missing values, another EDA was carried out, and remaining outliers have been eliminated.

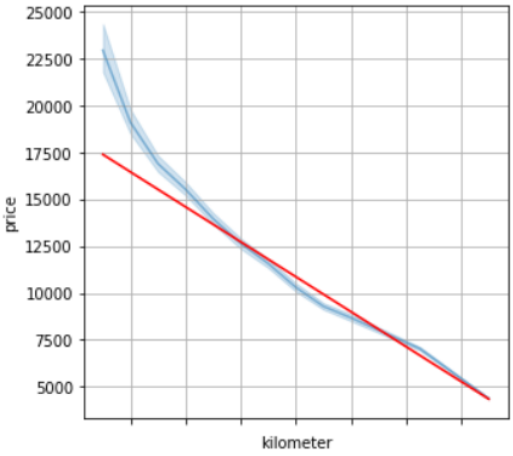
Initial findings from exploratory analysis:

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.

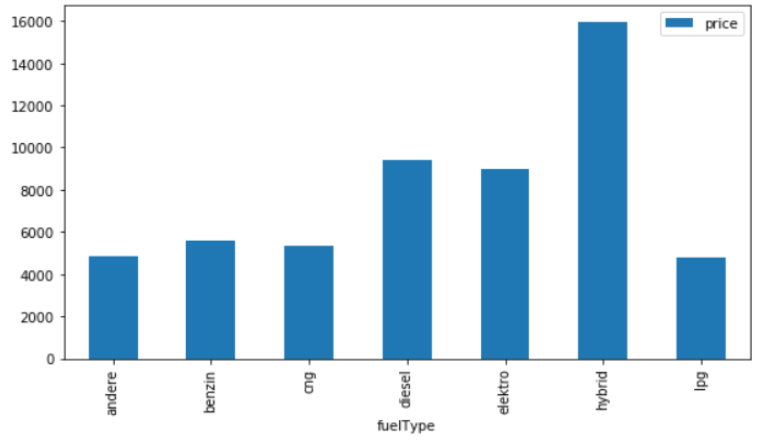


To further investigate the parameters affecting the price, more graphical analyses were made.
The red line represents the regression, in this (age, price) graph below. The regression line was shifted upwards in order to overlay. The trend is clearly visible, the age and price are negatively correlated.



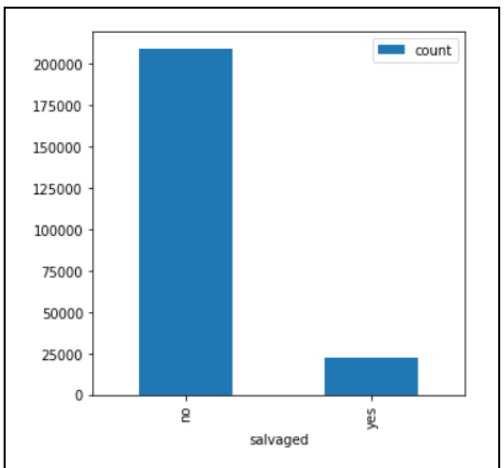


The (kilometer, price) graph above right reveals that the price is inversely proportional to the mileage. This information is also backed by the regression line, depicted in red, and the negative sign of the Pearson coefficient, which has been calculated as -0.37.



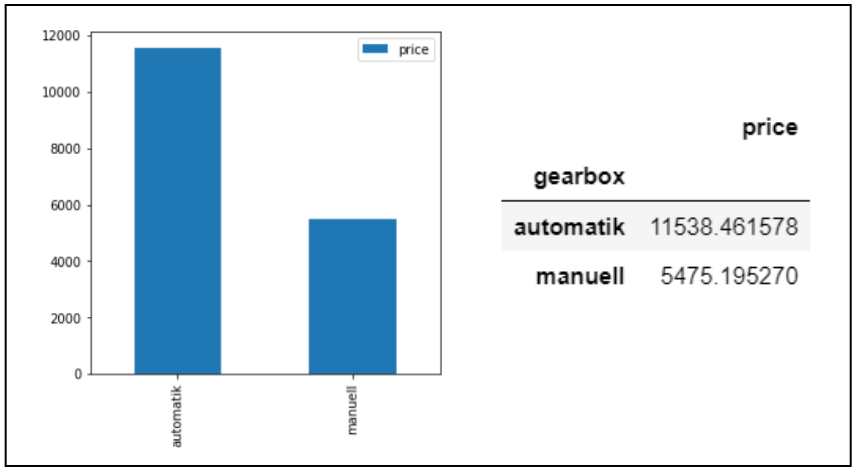
fuelType	price
andere	4886.744681
benzin	5583.599862
cng	5340.674757
diesel	9406.093050
elektro	8980.950000
hybrid	15973.527397
lpg	4773.563023

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

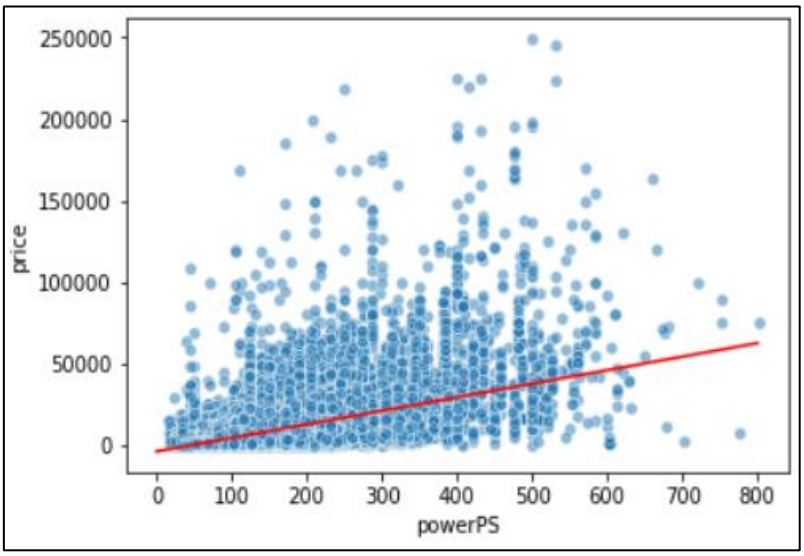


gearbox	price
automatik	11538.461578
manuell	5475.195270

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

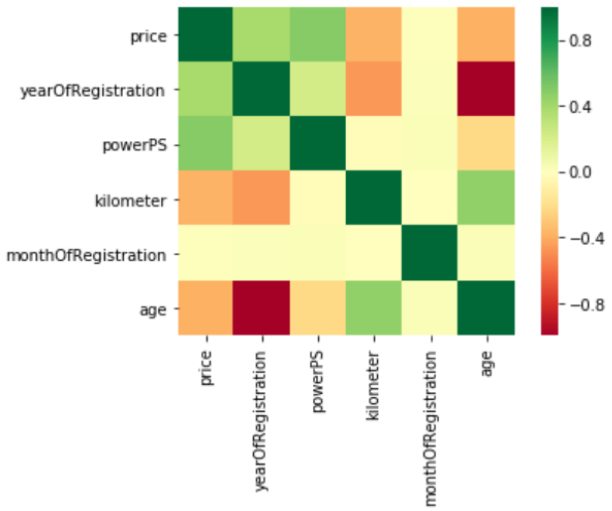
Automatic transmission vehicles are twice as expensive as manual transmission vehicles. Automatic transmission is more sophisticated, therefore more expensive. Moreover, some luxurious models only come in automatic transmission.

Exploratory Data Analysis:



After cleaning the large dataset and discarding all unwanted 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. 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”, 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 methods 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

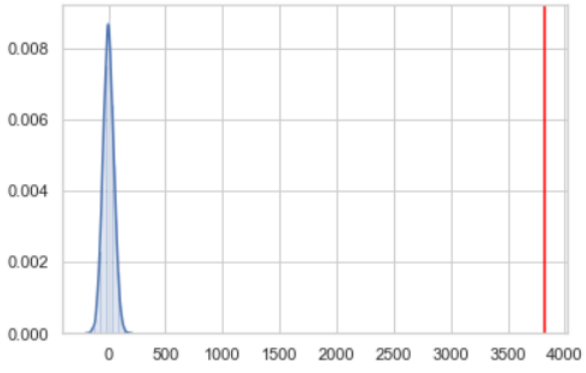
Please note that Skoda, a sister company of Volkswagen has similar values regarding the horsepower, since most vehicles use common engines between these two brands. This was reflected to t-test, and we can clearly see that a random vehicle selected from a pool of VW and Skoda have comparable average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

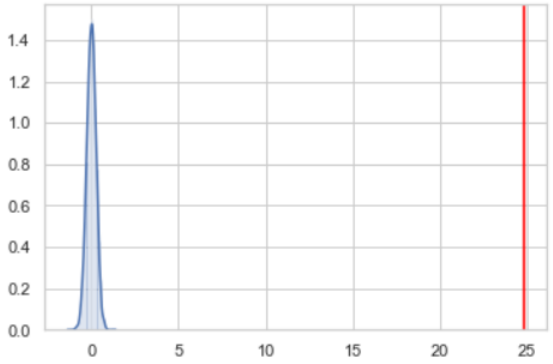
```
two_sample = stats.ttest_ind(df1[df1['brand']=='VOLKSWAGEN'].powerPS,
                             df1[df1['brand']=='SKODA'].powerPS)
print('The t-statistic is %.3f and the p-value is %.8f.' % two_sample)
```

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

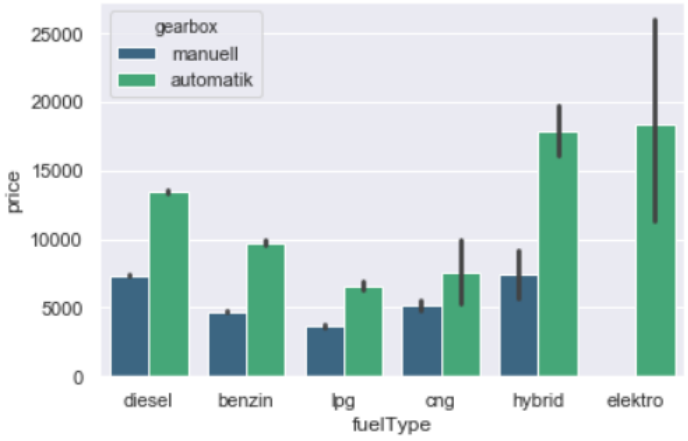
Another approach was using the bootstrap method as discussed above. Looking at the engine horsepower values of both Skoda and Volkswagen, we can see that the empirical difference of means of horsepower of each brand is 0.27965, and as we carry out the bootstrap method we can witness the normal distribution 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 an average 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.

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.