German Car Insights

Stat-412

Yusuf Kağan ÖZKAN

2429215

1. Aim OF The Project

The aim of this project is to examine the factors affecting the prices of vehicles in the German car market, and to examine how the features of the vehicles affect each other.

2. Data Description

The data is taken from Kaggle, and it has 14 variables. It has 7 numerical variables which are Price (dependent variable), Power kW, Power PS, Fuel Consumption (L/100km), Fuel Consumption (g/km), and Milage, also Age variable is generated by looking at its manufacturing year of the car. It has 6 categorical variable which are Brand, Model, Color, Fuel Type, Transmission Type, and Offer Description which did not used for whole study. Also, it has Registration Date and Year column as date variable.

3. Data Cleaning and Tidying

For cleaning and tidying steps first, offer description variable is removed from the dataset, then column names of the variables change appropriately. After that we start to change miswritten observations for each variable to their appropriate format. Almost all the variables had miswritten observations, also some of them were in completely irrelevant format, so the ones can be change has been changed by appropriate techniques and the ones in irrelevant form or misplaced observations turned as NA. For example, for the variable Fuel Consumption (L/100km) there were observations ends with l/100 km, kg/100 km, and kWh/100 km, to put them in same scale and make the calculations possible, we removed the miswritten metrics and put all of them in same metric. Another example in Fuel variable, there were observations like Automatic, Manual etc. since these variables are miswritten in here, we removed them. Another example for Registration Date, there were too many observations in wrong format and there were different separators for different dates, so we put all the observations in a same format. Data Cleaning process has made for all the variables and at the end all the variables' classes changed accordingly where they belong.

3.1 Descriptive Statistics

brand	model	color	registration_date	year
audi :20738	Ford Focus : 3797	black :23657	Min. :1995-01-01	2019 :11795
bmw :19435	Audi A3 : 3493	grey :19304	1st Qu.:2014-04-01	2018 : 9786
ford :18442	Audi A4 : 3323	white :16916	Median :2018-03-01	2023 : 8500
hyundai: 6852	Audi A6 : 2828	blue :12439	Mean :2016-11-08	2017 : 7704
kia : 5615	Ford Fiesta: 2735	silver :11469	3rd Qu.:2020-07-01	2022 : 7405
(Other):26994	(Other) :81781	(Other):14225	Max. :2023-11-01	(Other):52723
NA's : 1924	NA's : 2043	NA's : 1990	NA's :2089	NA's : 2087

price	power_kw	power_ps	transmission	fuel
Min. : 150	Min. : 1.0	Min. : 1	Automatic :52186	Petrol :53213
1st Qu.: 12900	1st Qu.:110.0	1st Qu.:120	Manual :45195	Diesel :36828
Median : 20490	Median :135.0	Median :159	Semi-automatic: 128	Hybrid : 4203
Mean : 29445	Mean :155.3	Mean :191	Unknown : 496	Electric: 2424
3rd Qu.: 32480	3rd Qu.:185.0	3rd Qu.:218	NA's : 1995	LPG : 911
Max. :5890500	Max. :735.0	Max. :999		(Other) : 326
NA's :2034	NA's :23095	NA's :2060		NA's : 2095

fuel_consumption_l_100km	fuel_consumption_g_km	milage
Min. : 0.000	Min. : 0.0	Min. : 0
1st Qu.: 4.900	1st Qu.:117.0	1st Qu.: 25528
Median : 5.800	Median :137.0	Median : 68000
Mean : 6.051	Mean :138.6	Mean : 85628
3rd Qu.: 6.900	3rd Qu.:165.0	3rd Qu.: 127000
Max. :22.900	Max. :300.0	Max. :3800000
NA's :12702	NA's :20844	NA's :2080

Table 1: Descriptive Statistics

From above table we can see five number summary statistics of our data, for example for price variable we can see that minimum and maximum values are 150 euros and 5890500 euros respectively, the median is 20490 euros and mean value is 29445 euros, interquartile range is 19580. We can see that in price variable, the range is large and mean and median values are not close to each other and due to the extreme values, we can say that price variable is not symmetric. Also, it has 2034 NA values which will be remove in next part. For transmission variable we can see that we have 3 types of transmission and unknown transmissions, most frequent transmission type is Automatic one. Also, it has 1995 NA values which will be remove later.

4. Exploration of Missingness

The Dataset containing almost 6% of missing values, we observe this missingness from row data, the missing values came from after cleaning and 2% artificial missing values.

For Brand and Model variables we made imputation like this:

If Brand observation is missing, we took the first word of model observation because model observations were like that: if Brand is BMW the model was written as BMW X5, so by using that we impute Brand variable like this. For Model variable like previous one we impute the Brand name directly to Model observations. If both were NA, we leave them as NA. For Color variable we made imputation based on the mode of Color variable and it was Black.

For Color variable we made imputation based on the mode of Color variable and it was Black.

For Registration Date and Year variables we applied same strategy for Brand and Model Variables.

For Price variable, we impute them by their mean brand values. For example, if price value is missing for an Audi Car, the price of the car imputed as average prices of all Audi Cars.

For Power kW and Power PS variables same strategy applied, we impute them as their brands average Powers. For example, if a Ford Cars Power observation is missing, we impute them as average Powers of all Ford Cars.

Transmission Type observations imputed as randomly.

Fuel Type observation imputed as Electric if the cars' fuel consumptions are 0, and rest imputed as randomly.

Fuel Consumption (L/100km) and Fuel Consumption (g/km) again imputed as their average brand fuel consumptions. For example, if a Hyundai cars fuel consumption is missing, it is imputed by average Hyundai cars fuel consumption.

For Milage observation we used same technique above.

For Age variable it updated according to Year variable.

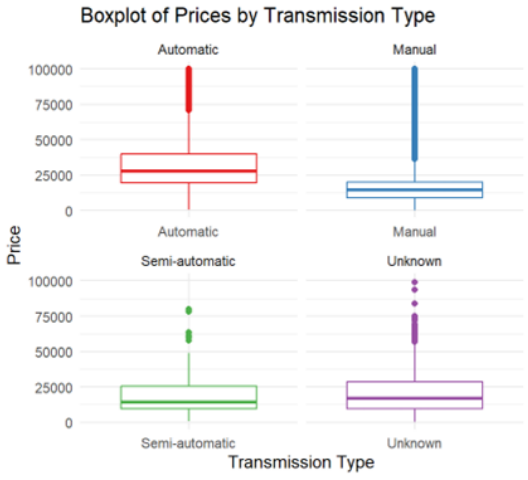
After all those steps we just had 125 missing value which we could not impute in a logical way, so we decide to remove those observations. If we calculate the amount of data that we lose it is accordingly 0.00125 this much which is not bad at all.

5. EDA (With Missing Values) and CDA

We are doing EDA to understand the behavior of the data visually and get ideas for data to what can be investigated. For that reason, we conduct some Exploratory Data Analysis and check them with Confirmatory Analysis. Let's start Our questions and their visual interpretations and their confirmatory analysis.

5.1 Is there any difference between Price of the cars by transmission types?

To investigate this question, we construct boxplots for different transmission types. To test this, we conduct Kruskal-Wallis test, since our normality and variance assumptions are not satisfied for ANOVA. All the groups have more than 10 observations and they are independent.



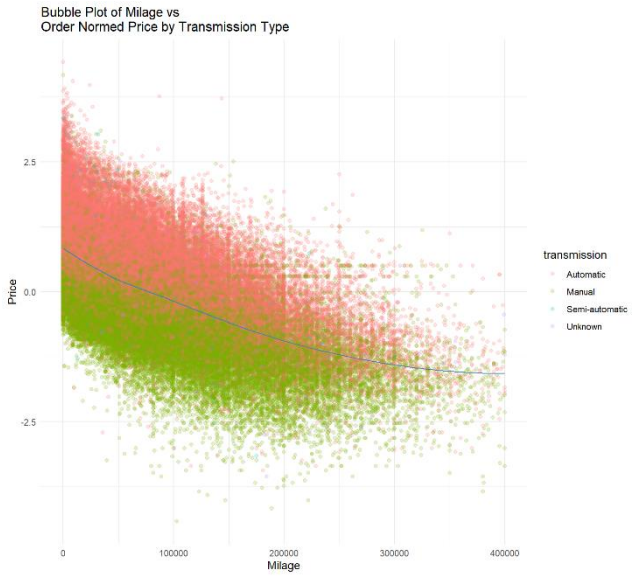
Plot 1: Boxplot of Prices by Transmission Type

The p-value is less than 0.05, so we can say that there is enough evidence to say that there are significant differences in the median values of price across different transmission types. Post-Hoc test has been made and results suggest that except Semi-automatic and Unknown transmissions, all the other transmissions differ significantly.

Col Mean- Row Mean	Automati	Manual	Semi-aut
Manual	172.5908 0.0000*		
Semi-aut	5.299757 0.0000*	-7.352679 0.0000*	
Unknown	14.61134 0.0000*	-10.04905 0.0000*	1.957293 0.1509

Table 2: Post-Hoc Analysis

5.2 How Price and Milage changes by Transmission Type?



Plot 2: Bubble Plot

From above plot, we can see that for Automatic transmission when milage is getting higher, price seems higher than other transmissions.

When we conduct multiple regression with dependent variable order norm transformed price variable and independent variables as milage, transmission types and interaction of price and milage variables we can see that transmission types have significant effect on price. We checked that after making order norm transformation to price variable it become normally distributed. Also, by Durbin Watson test we verify test our residuals are independent, and constant variance assumption is slightly violated because of outliers in milage variable.

6. Cross Validation

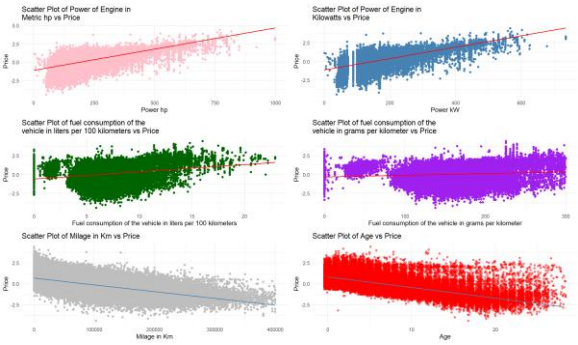
We are separating our data as Train and Test to measure our predictions accuracy after we made predictions. For that reason, we divide our data into 80% Train set and 20% Test set randomly before constructing model. Therefore, we used validation set approach.

7. Modeling

7.1 Multiple Regression Model

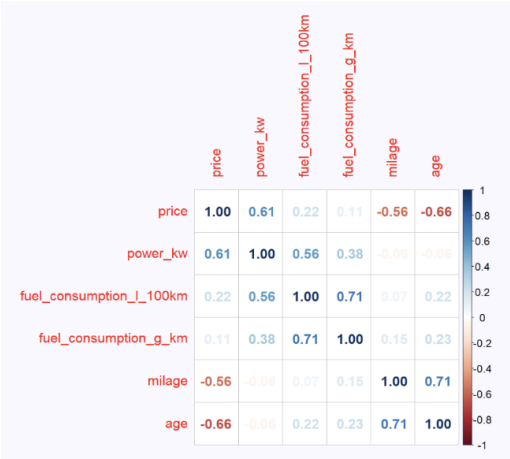
To understand which factors effects our response variable price, how our dependent variable has associated with other variables, and to make predictions for car prices we will construct multiple linear regression model.

Before building model let's check some visualizations for price variable to make our findings more interpretable and logical. We construct multiple Scatter plots to see different variables relationship between price variable, and by looking at the correlation plot we can see that which variables are more correlated with each other. For the model we removed Power PS variable because it was explaining almost same relationship with Power kW variable, and it was causing multicollinearity problem. Also, we add transmission variable to our model how transmission types of effects price of a car. Before starting our model, we scaled the whole data and to make our response variable we applied OrderNorm transformation which is suggested by BestNormalize package.



Plot 3: Multiple Scatter Plots of variables vs Price

We can see that most of the observations has linear relationship with order norm transformed price.



Plot 4: Correlation plot

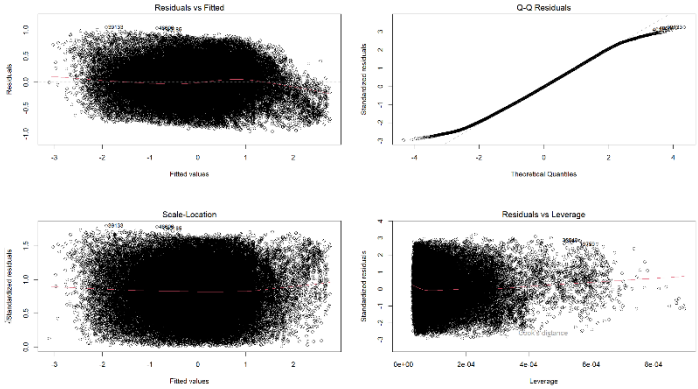
We can see that the correlations between variables from above plot.

```
Call:
lm(formula = price ~ ., data = cleaned_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95567 -0.22841 -0.00638  0.22119  1.05768

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.069858   0.005240  -13.33   <2e-16 ***
power_kw                 4.486661   0.019736  227.34   <2e-16 ***
fuel_consumption_l_100km -0.612166   0.035730  -17.13   <2e-16 ***
fuel_consumption_g_km     0.781371   0.019126   40.85   <2e-16 ***
milage                   -9.747007   0.111530  -87.39   <2e-16 ***
age                      -2.763065   0.012573 -219.76   <2e-16 ***
transmissionManual       -0.404703   0.002922 -138.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3263 on 70273 degrees of freedom
Multiple R-squared:  0.8676,    Adjusted R-squared:  0.8675
F-statistic: 7.672e+04 on 6 and 70273 DF,  p-value: < 2.2e-16
```

Let’s See the output of our regression table.

Table 3: Regression Output

The model that we build contains the variables that order norm transformed price and independent variables. Since our normality and constant variance assumptions are violated before, we removed influential points and outliers. Also, several transformations have been made to catch the assumptions. Finally, removing influential points gives the best model in terms of both assumptions and model metrics. Let’s see the assumptions related plots.



Plot 5: Diagnostic Checking

From above plots we can see that residual vs fitted plot shows horizontal band and observations deviates around 0, also for qq plot we can see that there are some outliers at the tails, but generally the pattern is almost at 45 degrees line, again for squared residuals plot we can see that they are in horizontal band around 0. Statistical tests suggest that residuals are independent, but there is slight violation of normality of residuals and constant variance assumptions. We know that with a large sample size, the central limit theorem can mitigate some of these issues, particularly with the normality of residuals. For variance assumption weighted least squares regression is also used but it did not change anything like other approaches so, at the end we assume that we have constant variance. Now, lets interpret our model output.

We can see that the minimum and maximum values are very small and less than 3, which indicates there is no outliers in the model, also median value is very close to 0 which indicates it is symmetric. Also, we had 4 type oof transmissions but while we are removing influential points the ones semi-automatic and unknown transmission types are removed also. It is just manual and automatic transmissions now.

When we checked model significance the model is significant, and our adj_Rsquared value is almost

0.87 which is pretty good. It indicates that 87 percent of the variability in price can be explained by independent variables. All the independent variables are significant in our model and if we want to interpret some of them, we can say follows.

For one unit increase in mileage, our scaled response decreases 9.74 unit.

For one unit increase in age of the car, our scaled response decreases 2.76 unit.

Variables	VIF
power_kw	1.719206
fuel_consumption_l_100km	3.795005
fuel_consumption_g_km	3.557869
milage	2.839714
age	3.027305
transmission	1.406914

Table 4: VIF Values

Also, when we checked VIF values all of them less than 10 so we can say that there is no multicollinearity problem in the model.

7.2 Neural Network Model

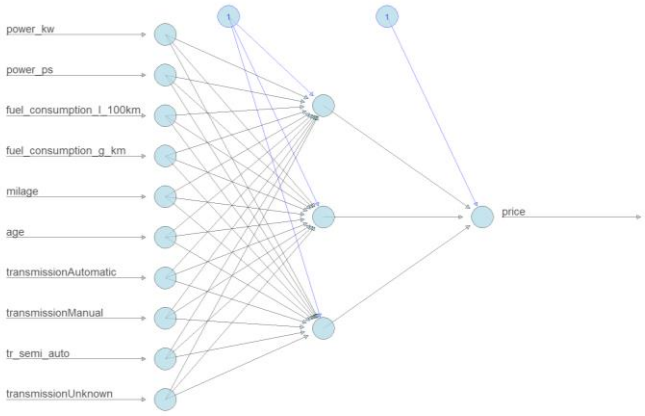
Before Starting our machine learning models, all data is scaled between 0 and 1. Also, the categorical variable transmission, has been encoded with one hot encoding. Also, we used same train and test set for all the models.

size	decay	RMSE	Rsquared	MAE
1	0.001	0.01854491	0.7565192	0.008523067
1	0.010	0.01872799	0.7552001	0.008224479
1	0.100	0.02352295	0.6195690	0.011315189
3	0.001	0.01734662	0.7869762	0.007640797
3	0.010	0.01854977	0.7592206	0.008232266
3	0.100	0.02352641	0.6201058	0.011316203
5	0.001	0.01798146	0.7772934	0.007738854
5	0.010	0.01841332	0.7625835	0.007972779
5	0.100	0.02361281	0.6168393	0.011374501

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were size = 3 and decay = 0.001.

Table 5: NN Model tuning

For this model we applied 10-fold cross validation with different parameters to be tuned. As we can see the best parameters selected as 3 hidden neurons and 0.001 learning rate. After we get these parameters, we build our model according to these parameters. Here we can see plot of our neural networks.



Plot 6: Neural Network Scheme

Also here is the table of some of the weights.

From 1 to 1	-4.095224
From 1 to 2	2.231241
From 1 to 3	1.247896
From 2 to 1	1.283295
From 2 to 2	0.07893972
From 2 to 3	-4. 995761
From 3 to 1	-1. 853806
From 3 to 2	-1. 170969
From 3 to 3	-0. 9048998

Table 6: Some weights from NN model

7.3 SVM Model

SVM regression has been built. After that we try to tune the parameters with 10-fold cross validation with different lists of parameters. Then we get following results for our tuning.

sigma	C	RMSE	Rsquared	MAE
0.01	0.5	0.02267716	0.7060509	0.007841048
0.01	1.0	0.02150512	0.7325220	0.007616222
0.01	2.0	0.02032362	0.7562311	0.007399856
0.01	5.0	0.01865485	0.7843045	0.007109672
0.01	10.0	0.01774196	0.7979505	0.006926171
0.05	0.5	0.01939805	0.7735024	0.007070697
0.05	1.0	0.01795336	0.7948200	0.006845940
0.05	2.0	0.01707982	0.8088783	0.006710081
0.05	5.0	0.01612271	0.8228407	0.006521127
0.05	10.0	0.01570763	0.8273103	0.006474857
0.10	0.5	0.01901576	0.7715749	0.006886562
0.10	1.0	0.01801710	0.7900935	0.006703436
0.10	2.0	0.01730660	0.8037175	0.006609048
0.10	5.0	0.01630582	0.8191388	0.006501358
0.10	10.0	0.01577380	0.8278225	0.006452405
0.50	0.5	0.02128131	0.7067505	0.007125699
0.50	1.0	0.02021000	0.7248187	0.006945110
0.50	2.0	0.01959227	0.7360636	0.006946977
0.50	5.0	0.01919783	0.7428328	0.007121633
0.50	10.0	0.01889414	0.7489146	0.007255492

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were sigma = 0.05 and C = 10.

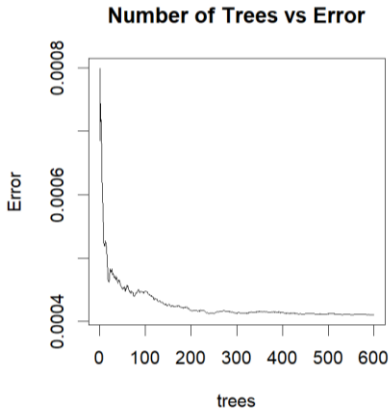
Table 7: SVM Tuning

The best parameters suggested from model is 0.05 Sigma and 10 as a c value which is cost in here. With this parameter we build our new, tuned model and

calculate its metrics. Performance metrics will be compared later for all models.

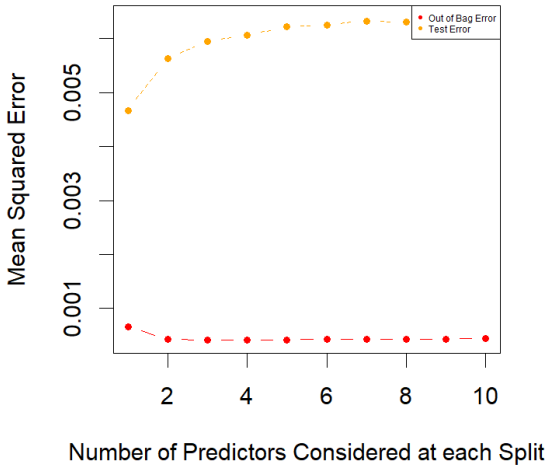
7.4 Random Forest Model

Random Forest model has been built with 500 trees initially, then it is decided to 450 trees will be better after checking graph below and according to variance explained results.



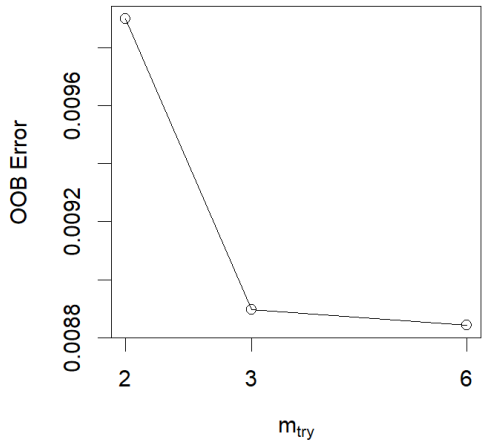
Plot 7: Number of trees vs Error

After Deciding tree number, we try to decide mtry number which is another hyperparameter to decide this we checked another plots and comparisons.



Plot 8: Out of Bag Error vs Test Error

We can see that after 3rd step test error becomes same for other levels.



Plot 9: Mtry Number vs Our of Bag Error

After checking above plots we observe that best mtry number would be around 3. Then according to the comparisons between different number of "mtry" values we observe that 3 is the optimum value for mtry. Then we build our random forest model with 450 trees and 3 for "mtry" parameters.

The variance explained by this random forest model is %80.48 which is reasonably good.

7.5 XGBoost Model

XGBoost model is the last model that we build for this project. First, we fit our XGBoost model, then we try to tune the parameters with 5-fold cross validation with different lists of hyperparameters. Then we observe that the best parameters are 150 for "nrounds" parameter, 3 for "max depth" parameter,0.1 for "eta" parameter, 0 for "gamma" parameter, 0.6 for "col sample by tree" parameter, 1 for "min child weight" parameter, and 0.75 for "subsample" parameter. After we done with tuning process, we build our final model and calculate performance measures of it.

8. Model Evaluation

To compare and select the best model we calculate models' performance metrics such as RMSE, MAPE etc. Now, let's see and compare performance metrics for both train and test set.

	RMSE	MAPE	MAE	MPE
Neural Network	0.03885	0.3339	0.0090	**0.0679**
SVM	0.03821	**0.2768**	0.0074	0. 1345
Random Forest	0.05296	1.3507	0.02363	1.33541
XGBoost	**0.03576**	0.2985	**0.00715**	0.15885
MLR	0.08575	0.2962	0.03576	0.24441

Table 8: Test Set Performance Metrics

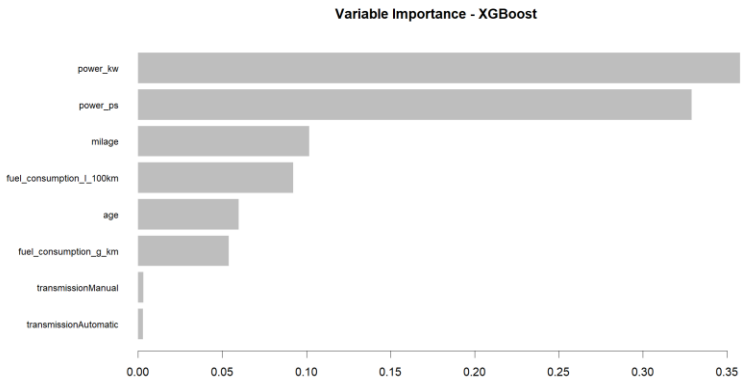
	RMSE	MAPE	MAE	MPE
Neural Network	0.01716	-	0.00766	-
SVM	0.01464	-	0.00586	-
Random Forest	0.04293	-	0.02226	-
XGBoost	**0.00924**	-	**0.00511**	-
MLR	0.09531	-	0.06366	-

Table 9: Train Set Performance Metrics

According to the test results we can see that for Test set performance, XGBoost model has better metrics and for training set XGBoost is again the best model among others. When we are selecting model, our final decision would be XGBoost model according to the test set, also train set suggest that this model is the best one.

8.1 Variable Importance

Since our best model is XGBoost, we will look at the variable importance plot of it.



Plot 10: Variable Importance of XGBoost model

In summary, the XGBoost model places the highest importance on the vehicle's power output (both in kW and PS), followed by mileage, fuel consumption (in liters per 100 km), and the vehicle's age. Fuel consumption measured in grams per km has moderate importance, while the type of transmission has minimal impact on the model's predictions. This analysis indicates that performance and usage metrics of the vehicle are the most influential factors in the model.

9. Discussion/ Conclusion

In this study, first we cleaned the data, then we made imputations with different techniques. For missingness mechanism we encounter with different type of missingness which are MCAR, MNAR. With appropriate imputation we get rid of this missingness situations. After these steps, we conduct EDA part with research questions, and we made appropriate statistical test to answer these questions. After that, we made cross validation to split the data to measure our model performance. Then we build our statistical model Multiple Linear Regression, while building this model several transformations, and different statistical modeling techniques have been used to reach regression assumptions. At the end constant variance assumption is slightly violated. Then neural network model has been built with 3 hidden layers after parameter tuning process. After that SVM method used to predict price values like others, in here we again tune the parameters and select the best model. Then Random Forest model conducted, here also we tune the parameters and select the best model. Finally, we build XGBoost model, tuning process was a bit longer in here, but we again build the best model with best parameters. After those steps we compared our model's performance metrics, and we observe that XGBoost is the best model among others.

10. References

Yozgatlıgil, C. (2023). Statistical Data Analysis [Boosting Algorithms]. METU, Ankara.

Güler, B. (2024). Statistical Data Analysis [Stat412 Recitation 13]. METU, Ankara.

Güler, B. (2024). Statistical Data Analysis [Stat412 Recitation 12]. METU, Ankara.

OpenAI. ChatGPT. 2024. https://openai.com/chatgpt