Term project - Ford Focus analysis

Attila Serfozo

2020.12.30

Contents

1 Executive summary

We address the question, whether we can predict car prices based on their key variables age, kilometers, performance, fuel type and transmission type. Our results are based on the data from 2020 October of Ford Focus hatchback cars suggest that by using age, kilometers and performance we can catch 90% of the overall variation in price of Ford Focuses. Another key outcome was according to our selected model, until a Ford Focus is less than 3 years old, it loses in average 1.5 million from it's value in every year on the used car market and after 3 years, these vehicles tend to worth in average 0.3 million HUF less in every year.

2 Introduction

In this project our goal is to build a price prediction model to predict the price of Ford Focus Hatchback cars based on the data of available cars on the largest Hungarian used car website (hasznaltauto.hu). The dataset of cars was scraped from the used car website on 2020.10.13 15:32. In the exercise to provide an accurate prediction for prices we will take into account important variables like price, age of cars, kilometers ran, performance in horsepower, fuel type and transmission. In the end we would like to finish with a multiple regression model which can provide a reliable benchmark price for a Ford Focus car and hopefully can be applied to other brands and cars as well.

3 Data

3.1 Data collection

The data to be analyzed is collected from the largest Hungarian used car website (https://www.hasznaltauto.hu/). To conduct a transparent price comparison through the different variables we restrict our attention to a specific car type, the Ford Focus. The data includes used cars only, therefore advertised new cars or demo cars from the website are excluded. We also narrow down the variety of Ford Focuses to hatchback cars only. Sedans, minivans and estate cars can distort the model as they tend to have different pricing, since they are sometimes different market segments (sedans, minivans) or they are often old taxi or police cars (estate) decreasing the average price.

In the analysis we examine the price of the cars registered from 2010, so maximum 10 years old. The motivation behind is that I also have an approximately 8 years old Ford Focus, thus my main area of interest is in the price of cars less than 10 years old. Based on these criteria we finish with data on 244 Ford Focus cars.

For the data collection I used a publicly available point and click interface web scraper which can be added as an extension to Google Chrome at the following website (https://webscraper.io/).

3.2 Data cleaning

During the cleaning process we need to decide how to handle missing values. As price is the main variable of the analysis it was an obvious choice to exclude the 2 observations which lacked price values. We scale prices to million HUF and kilometers to thousand km to ease understanding of visualizations. The main variable of interest is age of cars which can be created from Registration Date. In case of missing month values we substitute January to calculate age. Empty values were also common in cylinder capacity variable, but this can be acceptable as we intend to use horsepower as the main performance attribute. Also we exclude the 3 electric cars from the analysis as there are not enough observations on this type of cars to predict their effect on price. As a result of the cleaning process we finish with 239 observations. The variables of the cleaned dataset with their description can be found in the variables.xlsx.

3.3 Exploration of variables

Our aim is to predict the prices of cars. To achieve this goal our main variable of interest is age in years and to create a better fit and estimation we include performance in horsepower, kilometers run, fuel type and transmission in the analysis.

Table 1 below shows the descriptive statistics of these variables. Fuel type is transformed into a dummy variable, where 0 means petrol and 1 means diesel. Transmission is also transformed into a dummy, where 0 means manual cars 1 means automatics.

Table 1: Descriptive statistics of key variables

Statistics	Price mHUF	Age	Performance in HP	Kilometers in thousand	Fuel type	Transmission
mean	3.22	6.57	125.41	122.98	0.37	0.08
median	2.65	7.17	116.00	124.00	0.00	0.00
min	0.93	0.33	80.00	0.00	0.00	0.00
max	12.69	10.75	402.00	391.20	1.00	1.00
sd	2.04	2.88	51.47	68.08	0.48	0.27

We also look at the distribution of the variables with histograms, this can ease understanding our variables. The histograms can be found in Appendix 1. As prices usually log-normally distributed we could use logarithmic transformation, but for this model we keep it simple as in the regression part it will not improve significantly the fit of regression, but it will help interpretation at predicting the price of a car. Logarithmic transformations on other variables did not look reasonable.

Finally I checked the correlation between variables and the results support our intuition that we should choose age as the main variable, it has a strong negative correlation with price (-83%). In addition kilometers has a relatively strong negative correlation with price and performance shows a positive relationship with price as well. Also it is important to highlight that there is a correlation (61%) between age and kilometers, but we keep both variables in the model as the correlation is not strong and they are both important variables.

3.4 Extreme values

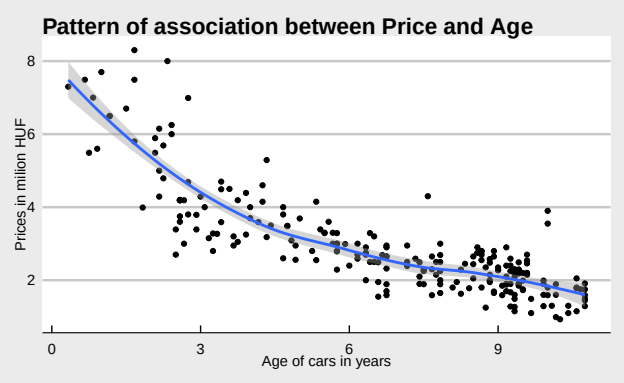
As we can see on the price histogram in Appendix 1, price is skewed with a long right tail. The reason of the large skewness are seven extreme values with cars above 10 million HUF. These observations are not errors, the cars are Ford Focus RS cars, which are very expensive due to their high performance above 350 HP, which results the long right tail on the performance histograms as well. Thus we exclude the 8 RS cars as their pricing is special and would make the fit of the regression worse for traditional cars in the analysis. The list of the excluded RS cars can be found in the Appendix 2.

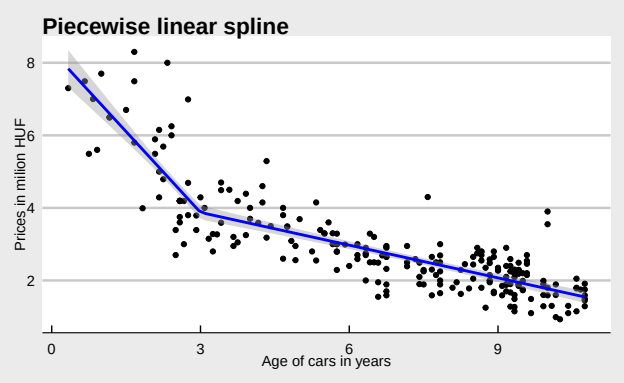
After the exclusion we finish with 231 Ford Focus cars, which is our final dataset for analysis. The quality of the data is rather good, measurement errors are treated in the data cleaning process and Ford Focus RS cars were identified as extreme values and were excluded as well.

4 Model

4.1 Model choice

The starting point of our model is to regress price on age. First we check the pattern of association with a non-parametric estimator to have an idea about the functional form between these two variables. The first scatterplot chart below shows the results.





The connection can be captured better with a piecewise linear splines. In case of age a piecewise linear spline a knot at 3 years provides the best fit, it can be seen above on the second scatterplot. To provide a better fit we can extend the model with the key variables. Surprisingly regression 4 was the best model in the analysis. It regresses price on age, performance in horsepower and kilometers. Adding fuel type or transmission type to the model did not increase its fit, therefore we do not include them in the final model in order not to overfit our data. The results of the regressions can be found in the out folder of the project in the ford_focus_models.html file.

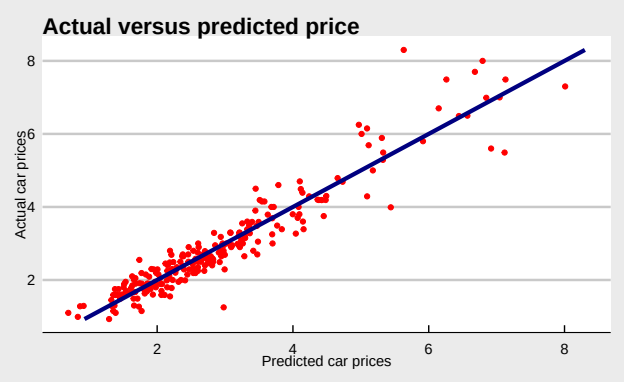
The equation of the model:

$$Price^E = \beta_0 + \beta_1 \times Age_{(Age<3)} + \beta_2 \times Age_{(Age>=3)} + \beta_3 \times Performance_{(HP)} + \beta_4 \times Kilometers_{(thousand)}$$

The results of the regressions confirms our instinct that the PLS model has a better fit than the simple regression. The R^2 increased from the simple linear's 69% to 78% due to the PLS transformation, by adding the extended model to it improves the fit further to 90%. The results suggests that without any other variable included, a year older car is in average 0.42 million HUF less expensive. However in reality we know that new cars tend to have a more dramatic price decrease as soon as they are leaving the saloon. To capture this let's see the results of PLS with a knot at 3 years. The model is significant below 3 and above 3 years at any significant level. According to the coefficients in case of cars less than 3 years old, cars 1 year older are in average 1.5 million HUF less expensive. On the other, between cars more than 3 years old, vehicles 1 year older worth in average 0.3 million HUF less. Based on the results of the extended model it seems that age, performance and kilometers are valuable part of variation in prices. Age and kilometers have negative and performance has positive association with prices, all are significant at 1%.

4.2 Prediction

The aim of this analysis was prediction. As we can see on the chart below the model captures prices well as the prices lay narrow close to the line. One of my main goal with this project was to be able to predict the price of my own Ford Focus car. So let's check whether we are getting a sensible result. I have a Ford Focus Hatchback car registered on 2011 February (8.67 years old now) with 125 HP performance and 96,800 km in it. By substituting the properties into their relevant variables in the model, we get that the predicted price of the car is 2.62 mHUF, which is quite realistic nowadays on the market.



4.3 Residuals

After checking the prediction power of the model let's have a look at the best and worst deals on the market. We can see that there are good opportunities to get cars below price for 1.5 mHUF less than the model predicted. Also there are overpriced cars, which are mostly younger vehicles.

Table 2: Best Ford Focus cars deals based on model prediction

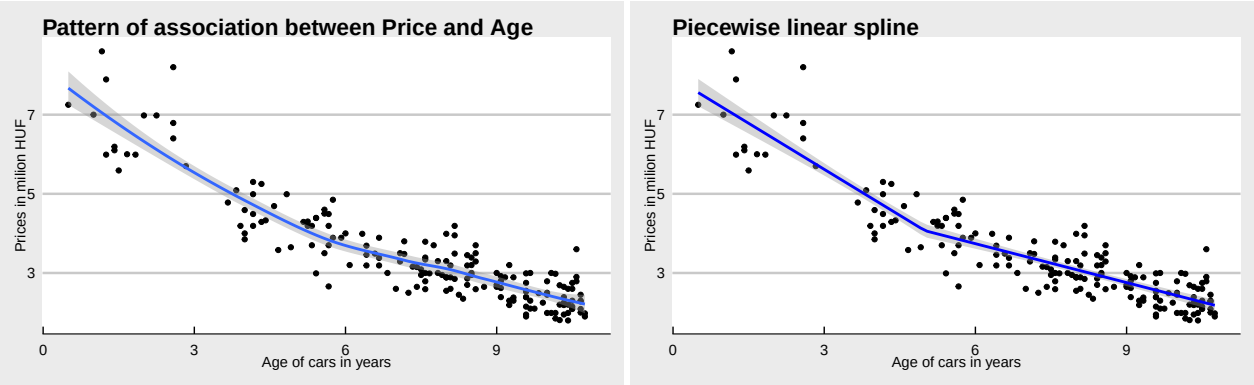
ID	Price mHUF	Predicted price	Residual	Registration date	Kilometers thousand	Performance in HP
318	1.25	2.980	-1.730	2012	0.4	116
325	5.49	7.118	-1.628	2020	6.1	125
327	3.99	5.441	-1.451	2018	53.7	120
370	5.60	6.919	-1.319	2019	1.5	124
385	4.29	5.093	-0.803	2018	53.0	125

Table 3: Worst Ford Focus cars deals based on model prediction

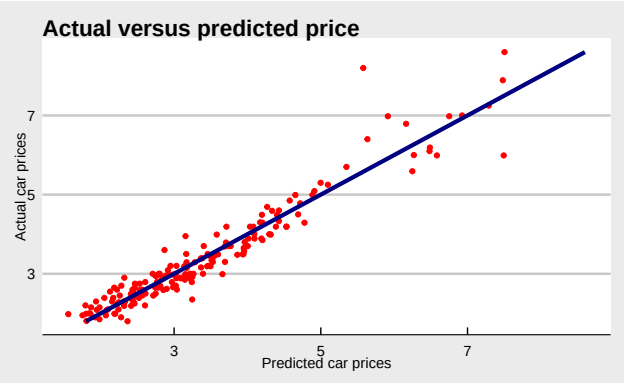
ID	Price mHUF	Predicted price	Residual	Registration date	Kilometers thousand	Performance in HP
352	8.30	5.632	2.668	2019	57.5	120
353	8.00	6.794	1.206	2018	8.5	250
356	7.49	6.261	1.229	2019	13.8	150
364	6.25	4.971	1.279	2018	15.5	125
365	6.15	5.092	1.058	2018	40.1	120

5 Generalization and external validity

So we finished with a model with a strong prediction power, but we should check as well whether it is only true for Ford Focus cars only. To test our results we will do the analysis on another car brand, which is in our case the Honda Civic. As we tried to avoid overfitting our model and the final multiple regression explains 90% of the overall variation in price I would expect the results to be valid. The data we are using was also scraped from hasznaltauto on the date of 2021.01.02 23:12. The scatterplot results can be seen below for Honda Civic.



After applying the same regression models for the Honda Civic the model gave similarly strong results. The regression explains 92% of overall variation in Honda Civic prices and all variables are significant at 1%. The prediction accuracy can be seen below on the Actual versus predicted price chart. The results of the regressions can be found in the out folder of the project in the honda_civic_models.html file.



6 Summary

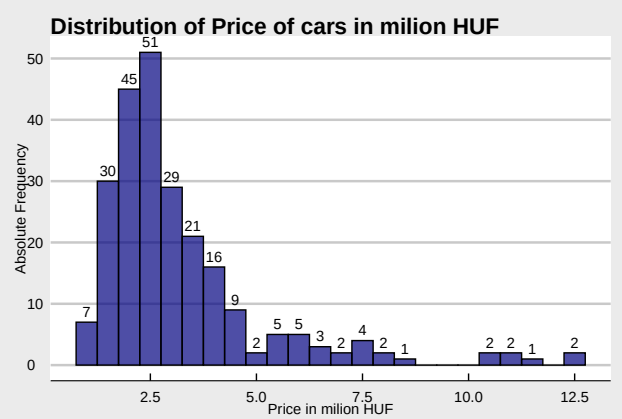
We have analyzed the relationship between Ford Focus Hatchback age and prices. We have used a piecewise linear splines for age of cars with a break at 3 years with extended variables of performance in horsepower and kilometers to catch the pattern better. Using a model we arrived to the conclusion that with 1% level of significant we can conclude that price correlates negatively with age and kilometers and positively with performance in horsepower. According to the model until a Ford Focus is less than 3 years old, it loses in average 1.5 million from it's value in every year on the used car market. On the other, after 3 years, these vehicles tend to worth in average 0.3 million HUF less year by year. Interesting outcome was that including fuel type or transmission type in the model do not seem to have a significant effect on the results.

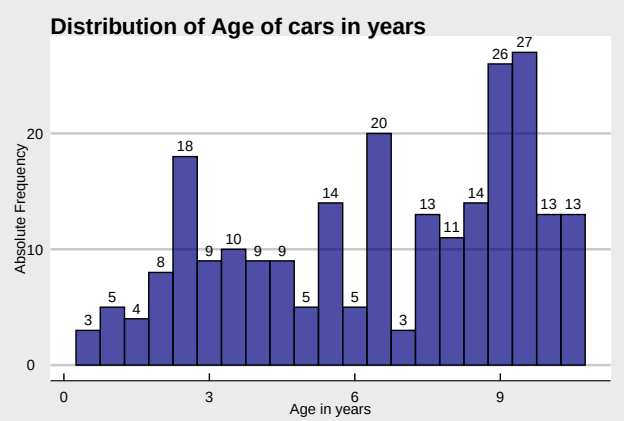
In my view the results can be very interesting for people trying to price their cars based on available cars on the market. I would also suggest as further research areas to test the model for other car segments as

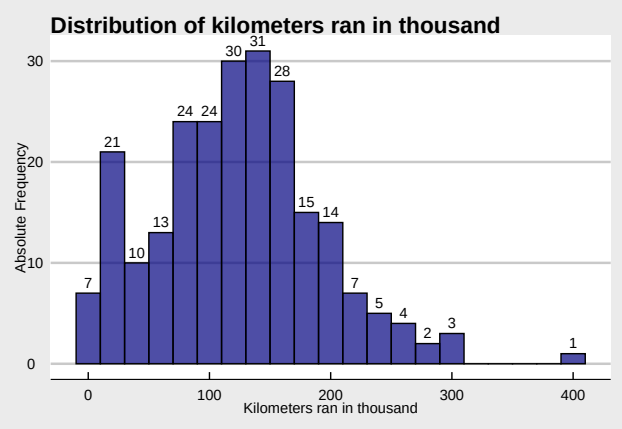
well such as estate cars, minivans or luxury cars, whether the prediction power stands there. Also it would be very exciting to compare these results to international car markets as well, such as the German used car website mobile.de.

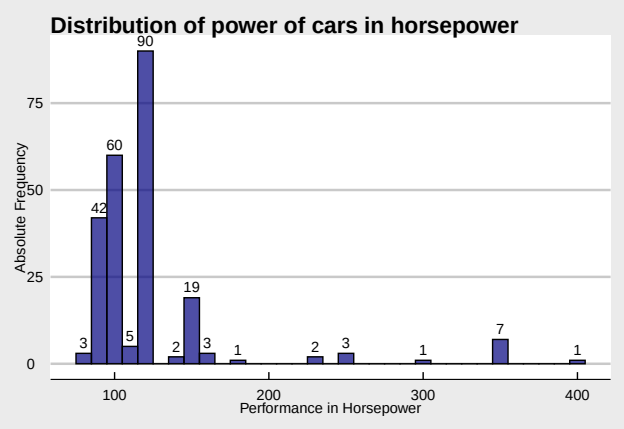
7 Appendix

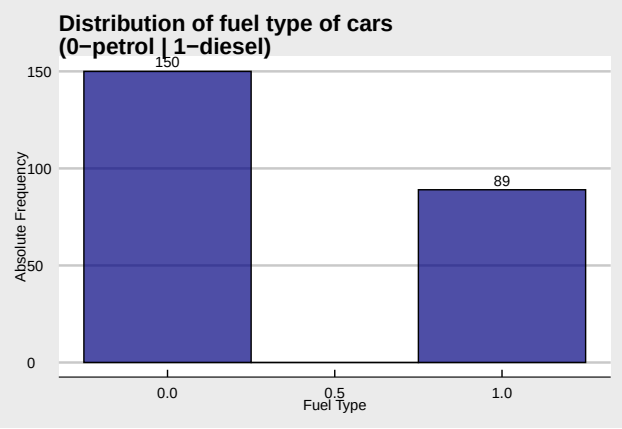
7.1 Appendix 1 - All histograms

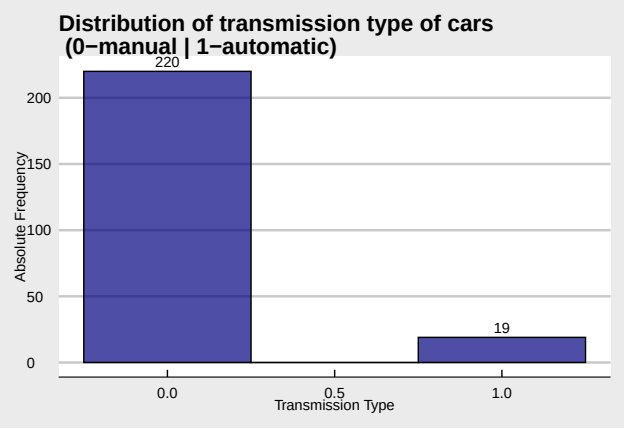












7.2 Appendix 2 - Observations with extreme values

Table 4: Extreme values in Ford Focus Price distribution

ID	Name	Price mHUF	Registration date	Kilometers thousand	Performance in HP
345	FORD FOCUS 2.3 EcoBoost RS	12.69	2017	61.4	349
346	FORD FOCUS 2.3 EcoBoost RS	12.49	2018	20.5	349
347	FORD FOCUS 2.3 EcoBoost RS	11.39	2018	23.9	349
348	FORD FOCUS 2.3 EcoBoost RS	10.99	2016	34.0	402
349	FORD FOCUS 2.3 EcoBoost RS	10.90	2017	20.3	349
350	FORD FOCUS 2.3 EcoBoost RS	10.59	2017	13.7	349
351	FORD FOCUS 2.3 EcoBoost RS	10.45	2016	18.9	349
354	FORD FOCUS 2.3 EcoBoost RS	7.90	2016	145.0	349