Analysis on used cars - vol 2

Zsombor Hegedus

2021 february 4

Executive summary

This document is to build multiple linear regression models with increasing complexity and predict the prices of used cars. This analysis is a continuation of an earlier project of mine in which I uncovered a pattern of association between the age and a price of Renault Megane models. In this paper I will take this analysis to another level and try to enhance my apprach.

The analysis is submitted as data exercise from Chapter 13 exercise 3. Workfiles are available in my github repo.

Feature engineering

My task is to build more and more complex models and to see if by doing so, I can improve their predictive power. Since I will use OLS only, the increasing complexity will be achieved with my selection of predictors and their functional form. In my earlier exercise, I discovered that a log-level model using the age of a car as the only rght hand side variable - performed the best. Learning from that, I will use ln transformation for my target variable, which is the price of a used car. Another lesson from that exercise was that I noticed a set of observations which formed a sort of cluster, and had prices that did not fit into the general pattern at all - I labelled them as potentially luxurious models, and created a dummy variables for them for this exercise.

Since I scraped quite a few information on used cars, I will use multiple regression models where I include not only age, but other variables such as: `Displacement`, `Horsepower`, `Km`, `PictureCount`. The first two speak for themsleves, the third one, `Km` is the number of kilometers the car ran, and `PictureCount` is the number of photos someone uploaded in the advertisement. Other than these variables I also included interaction terms and second and third order polynomials for the more complex cases. So overall, my models correspond to the below variable sets:

- 1st OLS: `Age`
- 2nd OLS: 1st OLS + `Displacement` + `Km`
- 3rd OLS: 2nd OLS + `Horsepower` + `PictureCount` + `luxury_dummy`
- 4th OLS: 3rd OLS + second order polynomials of continuous variables
- 5th OLS: 4th OLS + all interactions between `Engine` and continuous variables
- 6th OLS: 5th OLS + third order polynomials of continuous variables
- 7th OLS: 6th OLS + all interactions between `Age` and continuous variables and `luxury_dummy` and continuous variables

The 7th case of course is ridiculously complex and probably overfits the data by a large margin, but let's see how they fair against each other in the next chapter

Model selection

I ran 7 OLS models where the target varible was log transformed price, while the predictors were defined as described in the previous chapter. I first set aside a holdout set, that was a 15% randomly picked portion of all the data, and used the rest to train my models. I used 5-fold cross validation and averaged the RMSE -

my loss function - of each 5 folds to get the final RMSE. I will evaluate the models based on RMSE only - the lower this is, the better the model.

Table 1: Model comparison for OLS models for CV RMSE

	CV RMSE
OLS1	0.3507
OLS2	0.2952
OLS3	0.2831
OLS4	0.2839
OLS5	0.2827
OLS6	0.2827
OLS7	0.2837

Based on the RMSEs I'd say that it makes sense to increase model complexity, but only until a point. Even though we see lower RMSEs until the third case, I couldn't really achieve any improvement. Based on this I will select the third model, and used that for prediction as that is not very complex but achieves quite a low RMSE compared to the others. Anyway I will also show how the models performed on the holdout set in the below table:

Table 2: Model comparison for OLS models on the holdout set

	Holdout RMSE
OLS1	0.3489
OLS2	0.3137
OLS3	0.2989
OLS4	0.2945
OLS5	0.2996
OLS6	0.2996
OLS7	0.2991

Prediction

I chose model 3 for prediction, and used all the data to predict a price given the predictors for each observation. When predicting I also converted the log price back to level by not only using the exponential of the predicted log price, but also used a correction term. The first plot shows a quasi prediction interval for cars of given age - I used lowess on the actual price, the predicted price, and the lower and upper bound of an 80% prediction interval. It's visible that the interval is narrower, the older the car which can be explained by the low variation of price for older cars. This is an interesting finding, but probably not a surprising one as older cars are generally less expensive. The second figure is a scatter plot between the predicted and actual prices from which it can be seen that the predictions are actually quite close.

Conclusion

I set out to analyse data on Renault Megane used cars that I collected from Hasznaltauto.hu. In an earlier exercise it turned out that using a simple log-level OLS with age of car as the only right hand side variable was fairly good when uncovering the association between price and age of used cars. In this exercise I increased model complexity by adding more variables and played around with their functional forms. It turns out that when it comes to prediction more complex models did outperform the simple log-level case, hence if the task at hand is price prediction it is better to go with more variables. However too complex models might overfit the data so one should keep that in mind when choosing the final model used for prediction.

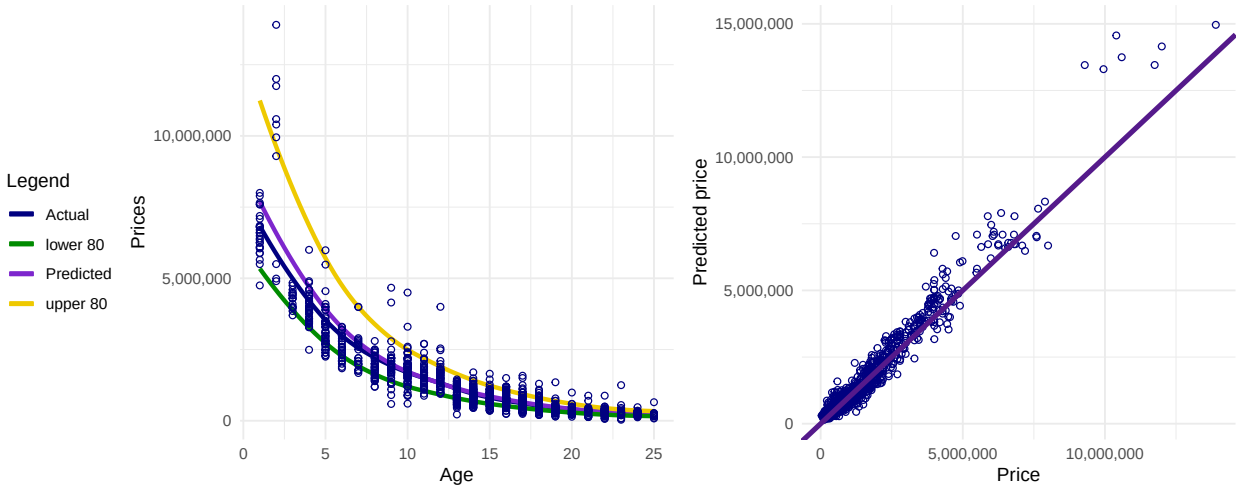

Figure 1: Price predictions