SEAT CODE

Victor Farre, Tomas Montoya

Tackling the Problem

The first key decision we made for this assignment was choosing the variables for the model. Beyond the value these variables could bring to improving the accuracy of the model, we also assessed the viability of including it into the final product. For instance, the variable *Title* can explain variability within models, as there are different versions. However, there are over 3,537 different model sub-types, on the other hand there are only 111 model types, consequently we opted to remove the variable *Title*. Our rhetoric for choosing variables was to either measure its correlation for numerical variables with price, as is the case for horsepower or mileage. For categorical variables, we visualize the average price difference across different categories. For example, we knew Manufacturer was an imperative parameter as a Dacia on average costs 11,578€, whereas a Cupra costs on average 33,400 €, and this discrepancy must be present in our model. After sufficient data exploration we decided to pick: *Manufacturer, Model, Transmission, HP, Year & Mileage* as our variables. On the basis of relevancy of these variables and their impact on accuracy, but also how easy it will be for the user to input these into our final product.

The model

We decided to adopt a hierarchical approach for our prediction model. We use a Linear Mixed Model (LMM) which we use to group the vehicles by *Manufacturer*, we then used the variables *Model* & *Transmission* as our two categorical variables. The other variables used are *Horsepower*, *Year* and *Mileage.* Through this we can segregate Manufacturer and Model, furthermore, as we include HP, we can capture the price difference within the model subtypes.

$$Price = \beta_0 + \beta_1 YEAR_i + \beta_2 HP_i + \beta_3 C(MODEL_i) + \beta_4 C(TRANSMISSION_i) + \beta_5 YEAR_i \cdot MILEAGE_i$$

An excellent characteristic of the LMM is its simplicity. As a statistical model this is extremely advantageous. For one, we can reduce the computing time, furthermore, it's extremely easy to add, remove and alter the variables used within the formula. Moreover, we can add interaction terms, in our case β_5 explores the interaction between Year and Mileage. An important detail to point out is that, we added a function in our model to include all unique car models into the train data, this way we can ensure that the model can be inputed any of the models in the dataset. For further info n LMM and how the work reference (*Statsmodels, 2023*).

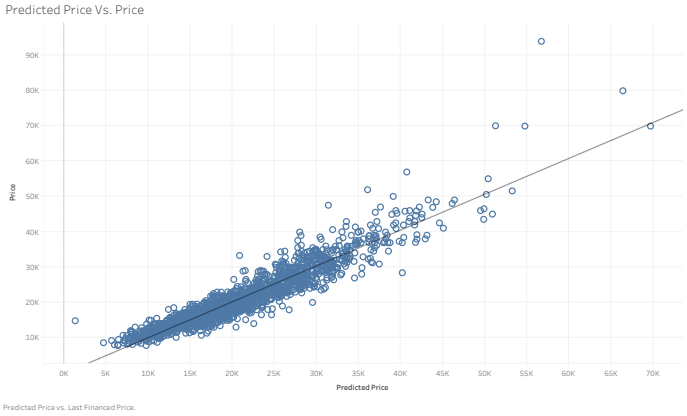
Results

Through this model we achieved a R^2 value of 0.89 on an 80-20 split. In addition, we calculated another measure of accuracy (*Equation 1,2*). This measure captures the variation between Predicted Price and the Real Price. Using this alternative method, we obtain a mean accuracy of 0.90. *Figure 2* reveals the variation of accuracy depending on car manufacturer, which shows a lot of consistency amongst all car manufacturers. In other words, the model does a great job at accurately predicting all vehicles regardless of which brand they're from. As of now, we have a working prototype of a website that runs locally, as seen in *Figure 3* and 4. In the current version, we manually input the Year of the car, Manufacturer…Mileage, and after all the data has been written and selected, we either click on the button or press the 'enter' button and the website will send the information to a python script which runs the model and then returns a prediction value back onto the website.

We want to continue working on this and provide even better insights for users and an overall better user experience. For one, we want to make a better interface, such as *Figure 5*. Another improvement is to make adjustments to the model, so in instances where the prediction model fails to properly predict the price for certain types of models, we can adjust certain parameters to properly capture the variables and their impact on the price.

Appendix

(*Figure 1*)



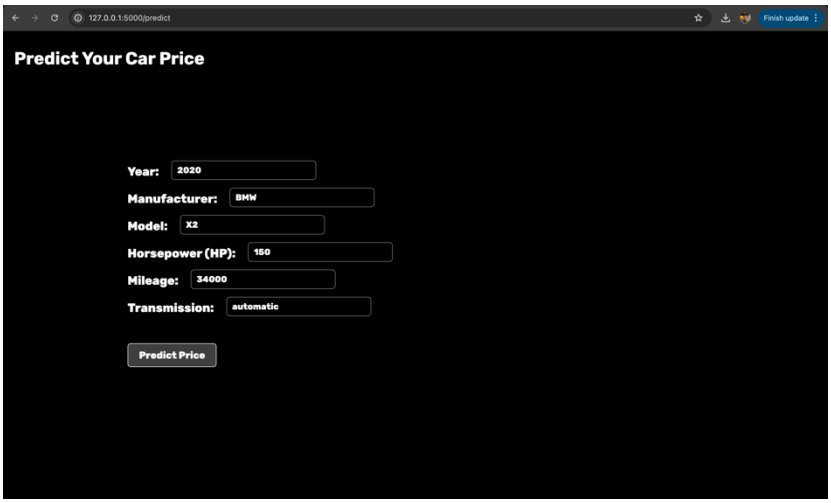
Off Amount = |Price - Predicted Price| (*Equation 1*)

Accuracy = 1 − (Off Amount/Price) (*Equation 2*)

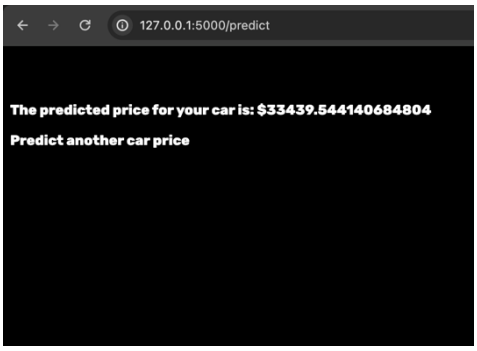
Accuracy by Manufacturer

Manufacturer	
AUDI	0.91513
BMW	0.89889
CITROEN	0.88962
CUPRA	0.92361
DACIA	0.83772
FORD	0.91946
HONDA	0.92695
HYUNDAI	0.89771
KIA	0.91180
MAZDA	0.93361
MERCEDES	0.91267
MINI	0.93285
NISSAN	0.92457
OPEL	0.90571
PEUGEOT	0.90835
RENAULT	0.89959
SEAT	0.89989
SKODA	0.91107
TOYOTA	0.89279
VOLKSWAGEN	0.88973
VOLVO	0.89389

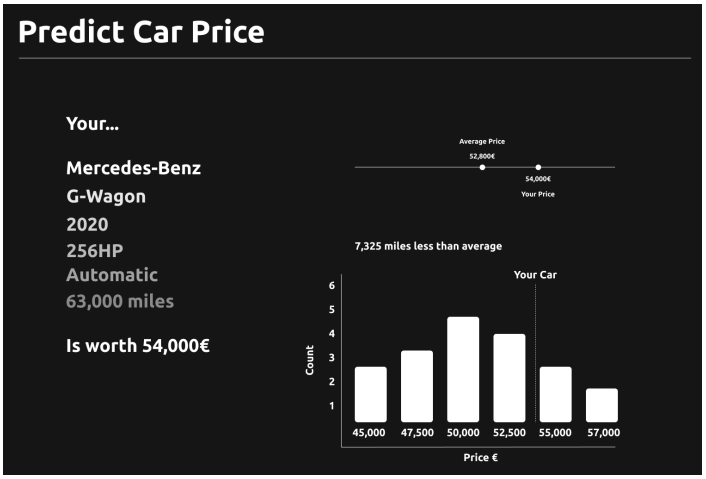
(*Figure 2*)



(*Figure 3*)



(*Figure 4*)



(*Figure 5*)

References

Linear Mixed Effects Models - statsmodels 0.14.1. (n.d.).

https://www.statsmodels.org/stable/mixed_linear.html