The first part of this analysis (as seen in the appendix) was to explore the data to get a better idea of what kind of data that I have available for this analysis and identify where some cleaning and/or formatting might be needed and get familiar with the dataset in general. After this section and the identification of the areas that might need improvement, I started working on cleaning and formatting the data. During this step I decided to normalize the numerical columns, change the format from the boolean columns to numeric (0 & 1), drop missing values as they were around 15% of the training dataset (another alternative if could have been to replace the missing values for the median in case of numerical columns and the mode for categorical ones, which is what a did for the test data set as the missing values where around 5%), get one hot encodings for some of the categorical columns and to extract as much information as possible from the 2 columns that were formatted as a list of lists (VehFeats & VehHistory).

Once the data was clean and with the needed formatting, I decided to work on 2 different models: A regression for the prediction of the Dealer Listing Price and a classifier for the Vehicle Trim.

The first part of the regression model was to do a feature selection for the features that would be part of the model, for this section I went for an embedded method (Lasso) as it provides a way to minimize to zero those coefficients that contribute the least to the regression. For the Regression model itself I used the ElasticNet model from the sklearn – linear model package. I split the Training dataset into “Training” & “Validation” and selected multiple hyperparameters to select the best combination of values which I check by means of cross validation and looking at the r squared (it is important to notice that other measures of performance could have been used, such as mean squared error. This relates to the end use of the model). After this process the training and validation datasets were combined to train the selected model again and use this last model as the final one for the test data.

For the second model (classifier) I followed pretty much the same structure as before: feature selection (used Lasso as well), did a comparison of 3 models looking for the best hyperparameters for each of the classifiers (3 legacy machine learning models) using the “RandomizedSearchCV” function from the “sklearn.model\_selection” library. After looking at the performance of the validation dataset and verifying that no overfitting was present, I re-trained the model using the entire “Training dataset”.

I did no used deep learning because the dataset was too small for the noted application, and I believe interpretability was important for this model.­­