Audi Car price Prediction

-by P.Charith

1) How cleaning/EDA was performed:

The chosen Dataset file ‘audi.csv’ had no null values. The only form of cleaning performed was, removing column ‘model’ as it does not contribute in predicting the price of an Audi car.

Exploratory Data Analysis (EDA) was done using seaborn and matplotlib libraries. The types of data visualizations done here, include countplot, plot, bar chart and heatmap.

Countplot : visualizing count of the number of models released throughout the years till 2020, visualizing count of model types based on ‘transmission’ feature, visualizing count of model types based on ‘FuelType’ feature.

Heatmap: visualizing a correlation between all numerical columns from the dataset.

Plot: Trend of the number of cars released by Audi company.

Bar chart: count of the different types of models, count of the number of cars manufactured by the company with respect to the year (horizontal bar chart).

2) Your independent and dependent feature:

Independent features: 'year', 'price', 'transmission', 'mileage', 'fuelType', 'tax', 'mpg', 'engineSize'.

Dependent features: ‘price’.

3) Why and how selection/engineering/scaling were performed:

Feature engineering was required in the audi.csv dataset as two categorical columns had to be encoded/converted into numerical columns to make price prediction possible. Columns: transmission and fuelType had to be converted. Column transmission has 3 unique values [‘Manual’, ‘Automatic’, ‘Semi-Auto’] and Column fuelType has 3 unique values ['Petrol', 'Diesel', 'Hybrid'].

pandas.get_dummies(...) was used for the required conversion. After conversion, the numerical columns replaced the categorical columns and the dependent column was stored as targets (Y) and removed from the main dataframe . Hence, finally, the available columns for prediction model were: ['year', 'mileage', 'tax', 'mpg', 'engineSize', 'Manual', 'Semi-Auto', 'Hybrid', 'Petrol'].

train_test_split() method was used to separate train and test data. Following this was, Standardizing the training data using **StandardScaler.**

4) Which activation function was chosen and why?

The activation function chosen was relu . When relu was used as activation function for the layers it showed promising results i.e. less loss compared to tanh, sigmoid etc.

5) Which optimizer was chosen and why?

The optimizer used is adam. When other optimizers such as RMSprop, SGD the loss was higher compared to when adam optimizer was used.

6) Which neural network and why? Describe how your neural structuring?

For the audi car price prediction two model were done:

1. **Model from sklearn**: An **MLP** (multilayer perceptron) with a single hidden layer of size (512,), activation: relu and solver: adam. This neural network was chosen because it provided good result while training (low loss). The Neural structure consisted of 3 layers: input, hidden and output layer. Input layer size was the number of columns after feature engineering, hidden layer size was taken as (512,) and output layer had size 1.

2.**Model from tensorflow.keras**: A keras Sequential model with 5 Dense layers and each layer except last layer have activation function ‘relu’. The optimizer used is adam. This model was chosen as the loss and time taken to fit was comparatively low. The Neural structure consists of first/input layer having 512 neurons with input dimension as 9(number of columns after feature engineering), second layer having 256 neurons, third

layer having 128 neurons, fourth layer having 64 neurons, fifth/output layer having 1 neuron.