Springboard

Capstone Project 1

Automotive Price Prediction - Indian Market

Data Wrangling

1. The data set:

The dataset 'cars_ds_final.csv' is obtained from kaggle uploaded by Abhinav Medharkar (https://www.kaggle.com/medhekarabhinav5/indian-cars-dataset) .

The dataset contains information about cars available for sale in the Indian Market and is updated 4th June 2020. It contains over 1200+ models and 140 features of these cars to study.

2. Data Wrangling Performed:

The following steps were taken to wrangle the data:

1. Shortlist features based on usability and common sense:

Due to the sparsity of the dataset, 24 features are preselected based on the kind of questions one could expect a prospective buyer to answer so that a reasonable prediction vector can be established. The choice of features were also based on the amount of information available for the features and a common sense understanding of correlation between the variables. Variables that had very little data (<50% of the dataframe size) or would not matter to a prospective buyer were removed.

Out of the 24 features selected, 11 were numerical, 10 were categorical and a combination of 3 would be used as the target variable for the clustering phase.

2. Addressing the target variables.

There were a few issues in the 'Make' and 'Model' entries, these were addressed by string search, splitting and replacing using the pandas module.

3. Cleaning and extracting the numerical features:

- To extract the numerical entries, the numerical columns were checked for string pattern consistency (since the units were a part of the numerical string entries). This was done using string comparisons with regex.
- Those with inconsistent entities were replaced by NaN entries.
- All the pattern consistent entries were then extracted using str accessors and regex.
- The numerical features were converted from object dtype to numerical dtype (float or int as required).

4. Filling the missing values in numeric features:

- The first attempt was to fill in the missing values by averaging by 'Model'. This would have ensured that accurate values are filled in. But this had the drawback of not filling up any if all the entries for a particular model and feature were all missing.
- Second attempt was to groupby 'Body_Type' and fill in the missing values by the median of 'Body_Type' for features that are functions of the Body style of a car. These are: Kerb_Weight, Ground_Clearance, Seating_Capacity, Wheelbase and Boot_Space
- The missing values for 'Mileage' were tackled by filling on an average by Body_Type and Fuel_Type. It was then converted into a more Robust feature: 'Cost_per_Km'.
- And finally 'Displacement' was tackled by sorting the dataframe using 'Power' and then interpolating.

5. Cleaning up the Categorical Variables:

- First, the unique values of the categorical variables were examined and the replaced for either duplication or relative usability to the customer.

  For example, for the category - 'Drivetrain', the categories of All_Wheel_Drive was replaced Four_Wheel_Drive since for practical purposes, both serve the same purpose.

- The missing values in the categorical variables were then filled out by taking the mode of the feature grouped by 'Make'. Since these variables would roughly remain the same within the bounds of a manufacturer.

These are: 'Drivetrain', 'Emission_Norm', 'Engine_Location', 'Fuel_System', 'Seats_Material', 'Transmission_Type'.

6. Removing Outliers:

The major outlier category that we are concerned with is ‘Price’. Primarily because we are predicting the price is our main motive. Also because multiple other features like power, torque etc are linked to price.

When we remove the outliers, we see that the price still reaches Rs. 4 Cr. mark. We cut this down further and limit our database to cars that fall within the Rs. 2 Cr. range.

We remove outliers from the numerical predictor variables using scipy.stats package for observations with a z-score > 3.

We will later also remove some outliers during our linear regression stage to obtain linearity with the dataset. We will also transform the target variable ‘Price’.

With this the dataset is clean and ready for use.