Nakul Pacheriwala (np2455)

ML Assignment 1 (EDA)

I have performed EDA on a Dataset containing details about different kind of cars which was scraped from a car enthusiast website.

On this Data we can do supervised learning of two kinds-

Regression if dependent variable in Continuous

Classification if dependent variable is not continuous (or is discrete/categorical)

I will evaluate the dataset to choose which variables can be a suitable target variable for this dataset.

Summarizing the EDA.

- Most of the work done here is pre-processing of data so that it can be used for prediction
- Extracted hidden information from columns and plot their densities to check if they are normally distributed or not
- The distribution is not normal for most cases, mini datasets with a few random (with different seeds) normally distributed values were generated to check if normal distribution makes a difference or not, my inference was that it makes a small difference but not significant enough to discard the remaining data points.
- Sparse data points with more than 70 percent NA values were removed
- Features with collinearity more than 0.85 were removed
- Features with multiple information in one column were split into 2 or more and original one was removed.
- Lots of features were converted to integers/ double by extracting numbers from the column
- Found all continuous variables and took a subset of data for numerical analysis
- Using Variable importance wasn't very useful as the variables with the highest values were ones that are unique for most cars in the given dataset like (name of the car) so while that would have given a great training accuracy it would be unable to help in prediction of price/mileage/country when new never seen before data comes up.
- Visualized variables for subjective analysis of important features which helps decide if any further processing is required or not.
- There are multiple dependent variables that can be chosen in this dataset, while most are regression problems like MSRP, Fuel economy etc, we can also use country or brand if we want to do categorical analysis
- Both Country and Brand have more than 2 classes, thus we cannot use methods like logistic regression or SVM directly without serious modifications.

After the EDA I understood that ideally we can predict the country of the car by training a model using this data or predict MSRP if using regression models and choose appropriate features depending on which model we generate.

Statistical methods for picking important variables seem like a hit and miss.

They are not very useful as they can be used to only eliminate a few features, however it cannot be used to rank the features also statistically every feature will have some effect even if it is very little. So, the statistical tests will not reject those on such a large dataset as some rows might have some very minor effect.

If not using any subjective methods, I feel in this case it would be better if we directly run the models and use regularization methods like Lasso to eliminate features.

I had tried various tests like t-tests and chi and others which were excluded from code as they were of no significance. I have left a few just to give an example.

When generating models, we would pick variables that are continuous but not correlated along with categorical features which aren't unique for each row.

Then we can use regularization to remove variables that are not important.

Below is the pdf file generated using latex on an ipynb file.