**A Predictive Journey into Oman Car Prices**

Author: Sri Charan Bodduna

**Purpose:**

This project deals about predicting the price of the Cars in country Oman with the help of statistical modeling and machine learning models using R language. Further, I will compare each model candidates and assess their results and performance to finalize best and appropriate model for this dataset.

**Dataset:**

Dataset used in this project is named as oman\_car\_prices\_2023.csv (comma separated file) downloaded from Kaggle at https://www.kaggle.com/datasets/amjadkhatabi/oman-car-prices-2023. Data consists of 16898 car details including their prices. It contains columns related to the car brand, model, manufacturing year, color, regional specifications, transmission types, type of fuel used, exterior colors, mileage, type of paint, condition, body condition, licensing, insurance, payment methods for purchase, adjusted price, cities, neighborhoods and exterior and interior options

**Description:**

As a preliminary processing, I have conducted data cleaning, preprocessing and feature extraction steps to make the data ready for modeling.

Data Cleaning:

From the overview of the data, I see there are inconsistent data in Year column, so I cleaned it. There are few columns which are out of scope for this project, so I removed them from the dataset. Furthermore, I saw there is an inconsistency between car condition and Kilometers traveled, I have corrected them by looking realistic assumptions. I have also removed outliers present in the price column to minimize errors.

Feature extraction:

As we know the price scales would differ between regular and luxury cars, so I have created a Boolean column which tells whether a car falls into luxury car or not.

**Exploratory Data Analysis:**

I have created multiple box plots between categorical columns and price to understand the relationship between each category and price distributions. I have also used these analyses for feature selection. I have also plotted histogram of price and correlation matrix between variables in the dataset.

**Train Test Split:**

I have split dataset in ratio of 6:2:2 for training, testing and validation datasets respectively.

**Modeling:**

I built multiple models using multiple linear regression, forward stepwise regression, regularization technique known as Ridge regression, decision tree regressor, Artificial Neural Network (ANN), K nearest neighbors (KNN). I have calculated Training Mean Squared Error(MSE), Mean Squared Prediction Error (MSPE) for each model and validated the model on the results generated.

**Evaluation:**

From the above analysis, I found out to be KNN is performing better and has provided least MSE on training data but on Test data MSE is comparatively higher than other models.

I saw that least test MSE is found to be on Decision Tree model with all variables which is 35239243.56. Which is little higher than linear models but linear models fails to validate assumptions.

To conclude based on the train and test data MSEs, I decided that decision tree as our final model.

To better understand the results, again I compared the Linear model, Decision Tree, and KNN models using validation data. Decision Tree model have resulted better results overall on the validation data. Which means Decision Tree can predict car price better on the new data points, which is reliable.

**Conclusion:**

I have filtered the data set by year by dropping the rows for the year > 2015. I converted the categorical variables with the dummy values for modeling. Created new features such as Luxury\_Car, Fleet\_Size, and mean\_kilometers.

From our modeling and analysis of the created model candidates, I have seen there performance using Mean Squared Error on different data sets. From this analysis, I can choose the better model performance characteristics as I tested on different data subsets.

From the analysis I found that decision tree out performed. I can be infer that the data contains many categorical variables which makes decision tree to better on this data.

Best MSPE achieved is 36318409.051 using Decision Tree Regressor model.

Nonetheless, This analysis not limited. I have more space to improve overall analysis. I can investigate data patterns on complete data set, use further data exploration techniques, hyper parameter tuning of models like variable transformation in linear models, increasing nodes and hidden layers for ANN, considering min-sample-split etc. for decision tree. These can be further implemented to see whether I can extract even more better model.