Update to be done on the project notebook and Report:

i. load dataset adverts and show its general information and say what you saw – use .info()

ii. give a descriptive analysis of numerical and categorical features – use .describe()

iii. Analysis of Univariate Distribution of Mileage, Price and Year of Registration – Talk about percentage of cars that price are less than 1,500 between 1,500 and 50,000, above 50,000. Also talk about percentage of cars that Mileage is less 300 miles, less than 6 months old, between 6 months to 15 years, above 15years

1. Data Processing for Machine Learning

1.1 Detect and deal with noise (i.e., erroneous values): The dataset had wrong years; some of them were really out of line, so they need to be dropped like the value 1515 and 1063, while some have “1” missing at the start, for instance, the value “999” and others a misplaced value of “1” instead of “2” at the start like 1009, 1007, 1008, 1006, 1016, 1017, 1010, 1015, 1018. All of these cases mentioned should be handled by dropping the out of line date entries and adding “1” or replacing with “2” where applicable. Convert the crossover\_car\_and\_van feature from Boolean to object(string).

1.2 Dealing with Missing Values

Year of Registration & Reg Code: Dealing with missing values in the Year\_of\_registration and Reg\_Code columns by scraping the UK's car registration codes for age and year identifiers tables from the Wikipedia page UK Vehicle Registration Codes and mapping the reg\_code from the Wikipedia table to fill in the missing year of registration. There are also missing values in the reg\_code column and some other columns. The vehicle\_condition also seems to have majorly NEW cars. About 31,570 missing values in the reg\_code column also have their year\_of\_registration to be missing. For every NEW car, there's no year of registration and reg\_code attached to it. This is because it hasn't been purchased and registered by any user. So, fill the missing Year\_of\_registration for NEW cars with 2020 and Reg\_Code for NEW cars with 20. This is because the maximum year\_of\_registration in the dataset is 2020 and also the Public\_reference column indicates the year, month, and day the advert was made, this also aids in my decision to use 2020 as the year of registration for new cars. Only USED cars have their year\_of\_registration missing but have some values for reg\_code. After filling the missing Year of Registration by mapping the Reg\_code and Year of Registration gotten from the UK's Car Vehicle Registration Code. There are still missing values in the Year\_of\_registration column and their vehicle condition was USED. So fill the rows with the mode of the year\_of\_registration of USED cars or use KNN Imputer.

Categorical Features and Mileage: The missing values of the categorical features; colour, body type and fuel type are can be replaced by the mode of their respective models or use KNN Imputer, while the missing quantitative feature; millage can be replaced by the average or using KNN Imputer.

1.3 Dealing with outliers and Noise

- Show an initial Plot a (1:2) boxplot for mileage and price to illustrate the distribution or outliers

- Mileage processing for ML: To process mileage for ML create a new feature called age of car by subtracting the year car was register from 2021 (not current year) and select data that mileage is greater than or equal to 300miles and age is less than or equal to 15years. This selection is based on the business (i.e AutoTrader) strategy of excluding market analysis of cars outside this category.

- Price processing for ML: To process price for ML we take into consideration cars of prices greater than or equal to 1,500 to less than or equal to 50,000. This selection is based on the business (i.e AutoTrader) strategy of excluding market analysis of cars outside this category.

- Show a post plot a (1:2) boxplot for new selected mileage and price to show cleaned or processed dataset.

- Drop public\_reference feature as this is just a unique ID for each car advert and won’t contribute greatly for any better correlation between the itself and the price.

- Some of the features in the dataset have the wrong data type and should be transformed to the correct datatype. For instance, the year was in string, convert into an integer, and the cossover\_Car\_or\_van was boolean should be converted into string.

1.4 Categorically-encode, rescale data; split data into predictors and target; obtain train/validation/test folds

- Divide the dataset into numerical and categorical features

- Categorically-encode: use LabelEncoder

- Rescale data

- Split data into predictors and target

- Obtain train/validation/test fold

2. Feature Engineering:

2.1 Derive car\_age (if it was not done earlier) by subtracting year of registration from 2021. Also derive annual\_mileage by dividing the mileage by the car\_age.

2.2 Produce Polynomial:

2.2.1 Produce a polynomial without interaction with degrees or orders [1,2,3,4,5,6,7,8] or One polynomial regression with degree 2 on a fraction of the dataset and check Under/Overfiting using linear regressor

- Visualize each polynomial degree order against the predictor price to see the best degree

- check the rmse score

- Integrate the above with hyper-parameter search (gride search) on Ridge regressor and 10-fold cross validation scoring. Compare the result above and save the best model (degree and hyperparameter (alpha))

**Use Pipeline**

2.2.2 Produce a polynomial regression with interaction on the original cleaned dataset

- Integrate the above with hyper-parameter search (gride search) on Ridge regressor and 10-fold cross validation scoring. Compare the result above and save the best model (hyperparameter (alpha))

**Use Pipeline**

Discuss your findings

3. Feature Selection and Dimensionality Reduction.

3.1 Manual Feature Selection and Exploratory Data Analysis: perform manual selection guided by domain knowledge and exploratory data analysis by:

Firstly, present graphically a plot of each on the features (predictors) against the target and visually see the plots that is a good predictor with the target.

Secondly, perform a simple ols (statsmodel) regression model on the dataset and get the results and display the summary, see and comment on the P>|t| values of both the intercept and the feature parameter. (Note: Those with P>|t| less than 5% are likely important features).

- Compare and discuss the output of the P>|t| values with the graphical plots in first part above if there are correlations

3.2 Automated Feature Selection Algorithms:

- Dimensionality Reduction with PCA: Do a dimensionality reduction with PCA on a copy of the original dataset and do Firstly a PCA-based model that makes better predictions at predicting price. You should use that as trial phase (preliminary model). Then, secondly, build another model that will zoom into the features in the original dataset, that is more explainable and potentially more suitable for communication with stakeholders.

- Run at least 2 automated feature selection algorithms (SelectKBest and RFECV) for most useful predictors. Let the 3.1 guide in terms of number of features needed.

**Use Pipeline**

Discuss your findings

4. Model Building

- Choose suitable algorithms that is:

4.1 A Linear Model

4.2 A Random Forest

4.3 A Boosted Tree

- For each of this model you will fit and tune the models; perform grid-search, rank, and select best hyperparameter models of each of them based on evaluation metrics and under/overfit trade-off.

4.4 An Averager/Voter/Stacker Ensemble: Build an ensemble with best performing models/configurations) from above.

**Use Pipeline**

**Discuss your findings. This section very important, execute it properly**

5. Model Evaluation and Analysis

5.1 Overall Performance with Cross-Validation: Evaluate selected model(s) according to popular score and loss metrics with cross-validation and discuss findings

5.2 True vs Predicted Analysis: Analyse true vs predicted plot and discuss.

5.3 Global and Local Explanations with SHAP: Analyse true vs predicted plot, gain and discuss insights based on feature importance and model output space using SHAP

5.4 5.4 Partial Dependency Plots: Display a PDP plots of the most predictive features and discuss your findings

**This section very important, execute it properly**