1. Is this a classification or regression (prediction) problem?

Ans. This is a regression (prediction) problem

1. Which 3 machine learning models did you select? Are there any particular reasons why?

Ans. We chose multiple linear regression, k-NN (KNeighborsRegressor) and Gradient Boosting (GradientBoostingRegressor) models for this particular problem –

1. The KNN algorithm uses 'feature similarity' to predict the values of any new data points, which seems like an appropriate way to predict.
2. Linear regression is less of a Blackbox and easier to communicate and understand
3. Gradient boosting because it often provides predictive accuracy that cannot be trumped. It provides several hyperparameter tuning options that make the function fit very flexible.
4. Outline how you preprocessed the data. Be sure to discuss (i) predictor selection, (ii) dealing with NA values, (iii) dealing with categorical predictors, and (iv) scaling.

Ans.

1. Predictor Selection: We have removed color based on practicality, cylinders column had the same value throughout hence we removed it and Fuel\_Type\_Diesel since it was highly correlated with fuel\_type\_petrol, other predictors exists.
2. Dealing with NA values:
   * + - “Mistlamps” has very large number of NA values, the entire column should be removed.
       - “ID” column is just an identifier
       - “Cylinders” because it has only 1 value (4.0)
       - “Removed” the rows having NA values for Price, Color, CC, Mfr\_Guarantee, Airco since the number of rows having these NA values are not that large.
3. Dealing with categorical predictors:
   * + - There is only one categorical variable in our set of predictors namely Fuel\_Type.
       - After flagging such categorical variable, we dropped fuel\_type\_diesel since it was highly correlated with fuel\_type\_petrol.
4. Scaling: We have used the standard scaler and converted all values into their z value to normalize the values in the dataset.
5. For each model that you selected, which hyperparameter(s) did you change?

Ans.

1. In Multiple linear regression, we don’t have any hyperparameters. Hence no changes made.
2. In k-NN algorithm, we changed the n-neighbor’s (k) value to the optimal value of n-neighbor. We found the optimal value of n-neighbor i.e k as 19.
3. For the tree-based models, we checked the performance of two models, Random Forest, and Gradient Boosting.

* For Random Forest we have changes the n-estimators and max-depth. The values are 600 and 10 respectively.
* For Gradient Boosting we have changed the n-estimators and learning-rate. The values are 60 and 0.2 respectively.

1. What are the training and testing performances for each model?

Ans.

1. For linear regression:

* For training set Root Mean Squared Error (RMSE): 1326
* For testing set Root Mean Squared Error (RMSE): 1445

1. For k-NN regression:

* For training set Root Mean Squared Error (RMSE): 1309
* For testing set Root Mean Squared Error (RMSE): 1517

1. For Tree Based Models:
2. For Random Forest Regression

* For training set Root Mean Squared Error (RMSE): 669
* For testing set Root Mean Squared Error (RMSE): 1084

1. For Gradient Boosting Regression

* For training set Root Mean Squared Error (RMSE): 803
* For testing set Root Mean Squared Error (RMSE): 1008

1. How would you judge the fit of each of the 3 models? Why?

Ans.

1. For linear regression, there is very small increase (119) in the RMSE from training to test. Hence, we can say there is no overfitting.
2. For k-NN, there is some increase (208) in the RMSE from train to test. Hence, we can say there is little overfitting.
3. For tree-based models:

* For Random Forest Regression There is large increase (415) in the RMSE from train to test. Hence, we can say there is overfitting.
* For Gradient Boosting There is some increase (205) in the RMSE from train to test. Hence, we can say there is little overfitting.

1. Which model would you recommend to ACB Auctions? Why?

Ans.

* I would recommend ACB solutions to use Gradient boosting because it has the best performance when compared to the other 2 models.
* The model which has less RMSE is a better model.
* As we compare the RMSE values of the models, Gradient Boosting has a better performance.