**Assignment 2**

*Group 5*

**Q1. Discuss a conceptual model to predict the MSRP including significant variables, correlation, your plan for training and testing.**

Before analysis, we will do EDA to visualize variables, do statistical analysis, check missing values and outliers, conduct correlations between independent variables, checking multicollinearity, and correlation between MSRP and predictors, check importance of predictors. We will also do feature scaling, like one-hot encoding. After doing EDA, we will select models. We decided to use multiple linear regression and random forest regressor to predict MSRP. Then we split the training and validation set, 60% for training and 40% for validation. Finally, we will do model evaluation, such as comparing accuracy rates between two models and comparing the performance of using different significant variables.

**Q2. Build both Multi-Linear Regression model and Random Forest model to predict the MSRP**

**of a car.**

Multi-Linear Regression model

A screen shot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

Random Forest model

A screenshot of a computer code

Description automatically generated

A white screen with black text

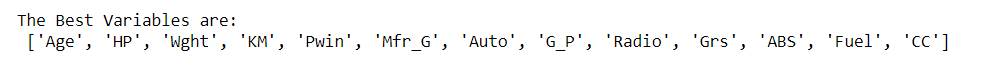
Description automatically generated

**Q3. Validate and interpret the model using a different number of significant variables.**

We used forward and backward selections to choose the best variables, and run model again to test their validation performances.

Multilinear Regression Model Validation

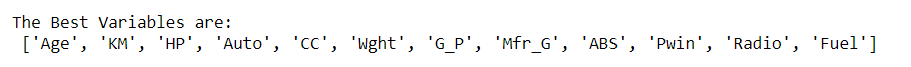
For forward selection



A white background with black text

Description automatically generated

For backward selection

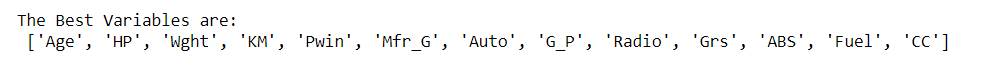


A screenshot of a computer error

Description automatically generated

Random Forest Model Validation

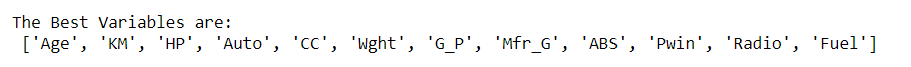
For forward selection



A white background with black text

Description automatically generated

For backward selection



A screenshot of a computer error

Description automatically generated

The variables reduced by backward selection are more significant than those selected by forward selection method, for example, with MPE value of -0.001 to 0.0303 for multilinear regression, and MPE value of -1.2989 compared to -1.1776 for random forest model.

**Q4. Compare results of random forest and linear regression, support each model by validating the results.**

Multilinear regression validation



A screenshot of a computer error

Description automatically generated

Random Forest validation

A close-up of a text

Description automatically generated

A white background with black text

Description automatically generated

Based on the validation analysis, we conclude that random forest model has better accuracy rate than that of linear regression, with MPE value of -1.1561 compared to 0.0589.

**Q5. Make a decision and offer your recommendations to the professor. Justify your decision based on your analytics results.**

After building multilinear regression and random forest model, splitting training and validation set, conducting forward and backward variable selection and evaluating the model performances, we found that:

1) Both linear regression and random forest models perform well on the training set, with low values for RMSE, MAE, and MAPE. Random forest generally provides more accurate predictions on the training set and validation set than linear regression.

2) On the validation set, random forest consistently outperforms linear regression in terms of RMSE, MAE, and MAPE. This suggests that the random forest model is more robust and better generalizes new data.

3) When we assess the predictors chosen through forward and backward selections and observe reduced RMSE, MAE, and MAPE on both the training and validation sets, it leads us to conclude that eliminating specific variables enhances the model's overall generalization performance.

4) Backward variable selection is more effective than forward selection based on their training and validation regression statistics summary.

The following chart shows the performance of validation set compared to training set, indicting the high accuracy using random forest model to predict MSRP:

A close-up of a computer screen

Description automatically generated

**A chart of data points

Description automatically generated with medium confidence**

Based on the finding above, we decided to use the significant variables (['Age', 'KM', 'HP', 'Auto', 'CC', 'Wght', 'G\_P', 'Mfr\_G', 'ABS', 'Pwin', 'Radio', 'Fuel\_Petrol','Fuel\_Diesel']) selected by backward reduction and Random Forest model to predict MSRP, which is more reliable and accurate for supporting business decisions and gaining insights into predictions of targeted variable.