Capstone Project

A car dealership sells different vehicle models manufactured in one of three regions (US, Europe, and Japan). After collecting a sample of data on cars in their inventory, they have hired your data science and analytics consulting team to provide them insights into their dataset. Specifically, they would like to know (1) How do American, European, and Japanese cars differ? (2) Can we build a model to predict which cars are American based on the reported features of the vehicle? (3) Additionally, the firm would like to understand how the variables recorded in the dataset affect mpg.

Ultimately, the dealership would like to understand how different countries of origin and different mpg performance affects the features of cars. This will help their sales team tailor their strategy.

Use the scripts from previous lessons to help you build classification and regression models to complete this assignment. Use the cars dataset for this project.

Using the ***cars dataset*** set, conduct a full analysis of the data set and build a *logistic regression* to predict which vehicles are American. Complete the capstone project by completing the following steps:

1. Import your data, inspect and clean it, remove any missing values

2. Conduct exploratory and statistical data analysis. This analysis should be geared toward answering question (1) that the dealership asked of your team.

2.1 Produce descriptive statistics on the vehicles overall and by country of origin

2.2 Produce relevant plots for visualizing features in the dataset, both overall and by country of origin. Further the dealership has asked for a series of data visualization in tableau for executives ***and*** a series of data visualization using ggplot/R studio ***or*** plotnine/Python for the dealership's in-house data scientists.

2.3 Conduct a series of statistical tests (at least three different statistical tests) to determine if American, Japanese, and European cars differ in statistically significant ways on different data features.

3. Build a logistic regression, this will answer question (2) for the client. As the '*brand*' column has three values ('US.','Europe.','Japan.'), you will have to construct a new binary variable to tell you which observations are American ('1') and which are not ('0'). After constructing a suitable logistic regression model, produce the following:

3.1 a confusion matrix

3.2 a cumulative probability chart

3.3 a lift chart.

4. Now the dealership would like you to build a multiclass classifier to also distinguish between European and Japanese cars. Use the TensorFlow and/or XGBoost package to build an appropriate classifier for the *'brand'* variable and report the confusion matrix.

5. Build a linear regression model to predict '*mpg*' from the other variables. Try building several models and choose the most effective one. Produce a plot of the true values (x-axis) vs your model's predicted values (y-axis).

6. Build a nonlinear regression model to predict '*mpg*' from the other variables. You may use TensorFlow or XGBoost to do this. Try building several models and choose the most effective one. Produce a plot of the true values (x-axis) vs your model's predicted values (y-axis).

7. Take your conclusions from the EDA in 2., the logistic regression in 3., the classifier in 4., and your regression models in 5. And 6., and produce a final writeup for the dealership. Support your finding directly with output and plots. Emphasize the business value and significance of your findings for the executives.