***Use this word document to describe all your steps after each problem. Show all commands & screen pages and describe your steps. Points are deducted for missing steps.***

Describe all steps for each problem set and capture screen shots for each step with a JPG image within this document. Do not forget to put your name on the top of this file for your submitted homework. Also, please replace \_ALL\_ text in red with your information! Please include in your MS Word document only relevant portions of the console output or output files. Sometime either console output or the result file is too long and including it into the MS Word document makes that document too hard to read. PLEASE DO NOT EMBED files into your MS Word document. You are not obliged to use Java or Eclipse. You are welcome to use any language and any IDE of your choice.

**Problem 1:**

Remove the header of the attached Samll\_Car\_Data.csv file and then import it into Spark. Randomly select 10% of you data for testing and use remaining data for training. Look initially at horsepower and displacement. Treat displacement as a feature and horsepower as the target variable. Use MLlib linear regression to identify the model for the relationship. Use test data to illustrate accuracy of your ability to predict the relationship. Create a diagram using D3 which presents the model (straight line), original test data and predictions of your analysis. Please label your axes and use different colors for original data and predicted data.

**Total points: 35**

The first step is to install pyspark in Jupyter:

A screenshot of a cell phone

Description automatically generated

And then import packgaes from pyspark ↓

A screenshot of a cell phone

Description automatically generated

For the problem 1, load the data first by using spark.read.format and define dataset as car\_df:

A screenshot of a cell phone

Description automatically generated

Then, we can take a look at this dataset by print out schema and show the descriptive data:

A screenshot of a social media post

Description automatically generated

To find out the correlation between “Displacement” and the target variable “Horsepower” by using ‘six’ package below. We can see the coefficient is 0.91 (a very strong positive correlation).

A screenshot of a cell phone

Description automatically generated

The next step is to prepare the simple linear regression with defining feature and label:

A screenshot of a social media post

Description automatically generated

Then split the training and testing dataset: randomly select 10% of you data for testing and use remaining data for training. Because there are 100 rows in this dataset, so there are 10 test test\_df and 90 train\_df.

A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

R square at 0.838 indicates that in our linear regression model, there are about 83.8% of the variability in “Horsepower” from training data can be explained using the model. While for the test data, the RMSE is 15.848 and the R² is 0.764 as shown below:

A screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generated

Let’s make the prediction using linear model:

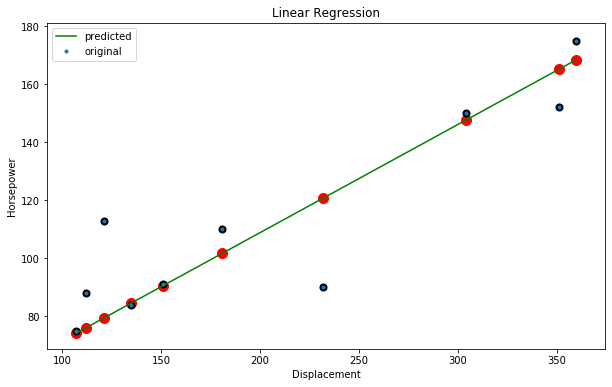
A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Plotting the linear straight line by displaying the original test data (red) and the predictions (blue).



**Problem 2:**

Treat: cylinders, displacement, manufacturer, model year, origin and weight as features and use linear regression to predict two target variable: horsepower and acceleration. Please note that some of those are categorical variables. Use test data to assess quality of prediction for both target variables. Which of two target variables is easier to predict, in the sense that predicted values differ less from the original values

**Total points: 35**

Put all your steps and screen captures here.

Firstly, if we want to include more features we have to convert categorical variables “Origin” and “Manufacturer” into numerical data by using StringIndexer

A screenshot of a social media post

Description automatically generated

A close up of text on a white background

Description automatically generated

Then to prepare machine learning with features & label by using VectorAssembler

A screenshot of a social media post

Description automatically generated A screenshot of a cell phone

Description automatically generated

**A screenshot of a cell phone

Description automatically generated** Split the dataset by 10:90 (test & train)

**A screenshot of a social media post

Description automatically generated**

As shown below: For predicting the “Horsepower”, the R^2 at 0.873 indicates that in our linear regression model, there are about 87.3% of the variability in “Horsepower” can be explained using the model. And the R square for “Acceleration” linear model is 0.5985

**A screenshot of a cell phone

Description automatically generated**

For the test data, we use similar prediction approach to compare the result:

**A screenshot of a social media post

Description automatically generated**

We can conclude the “Horsepower” is easier to predict as the target variable because it has much higher R square as the model predicting ‘Acceleration’ on the test data according to results shown below:

**A close up of text on a white background

Description automatically generated**

**A screenshot of a social media post

Description automatically generated**

Root Mean Squared Error (RMSE) measures the differences between predicted values by our model and the actual values.

A screenshot of a cell phone

Description automatically generated

But the second model to predict ‘Acceleration’ has lower RMSE, so it’s predicted values differ less from the original values.

A picture containing screenshot

Description automatically generated

**Problem 3:**

Repeat above analysis with decision tree method. Compare predicting ability/quality of this technique with that of the linear regression.

**Total points: 30**

Put all your steps and screen captures here.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

The R sqaure and the RMSE of the linear regression model are : 0.942 and 10.1. While for the decision tree model: R sqaure is 0.838, RMSE is 16.9

Obviously, the linear regression has a better performance than decision tree model.

A screenshot of a social media post

Description automatically generated