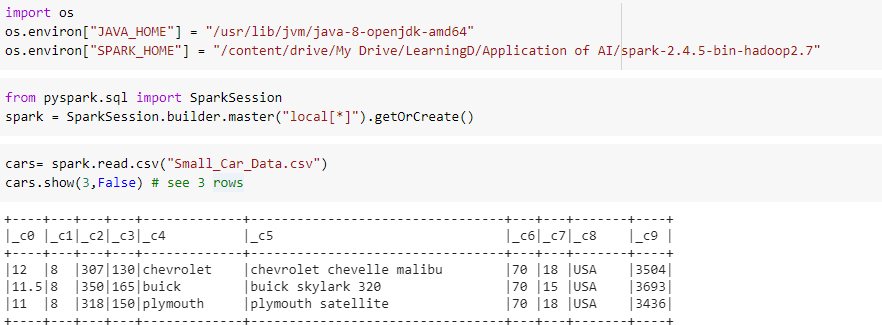
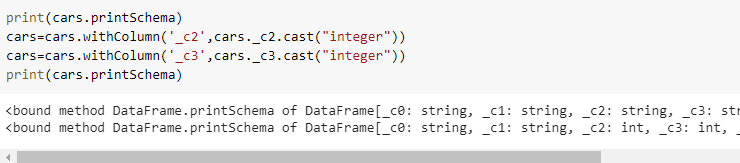
**Problem 1:**

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.

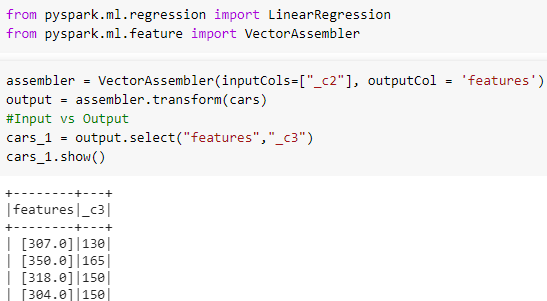
Read the csv file using into spark using read.csv function



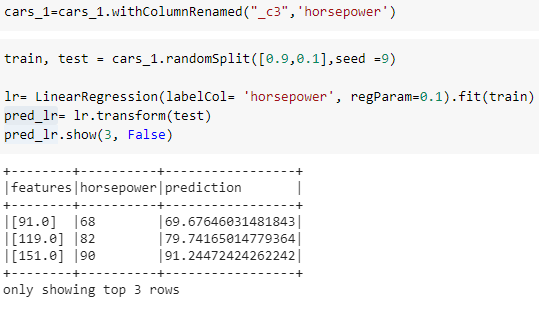
As spark only used numerical values, type casted the required variables to integer.

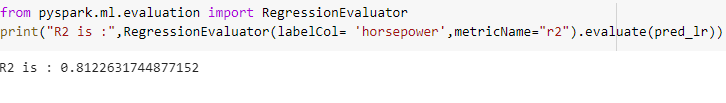


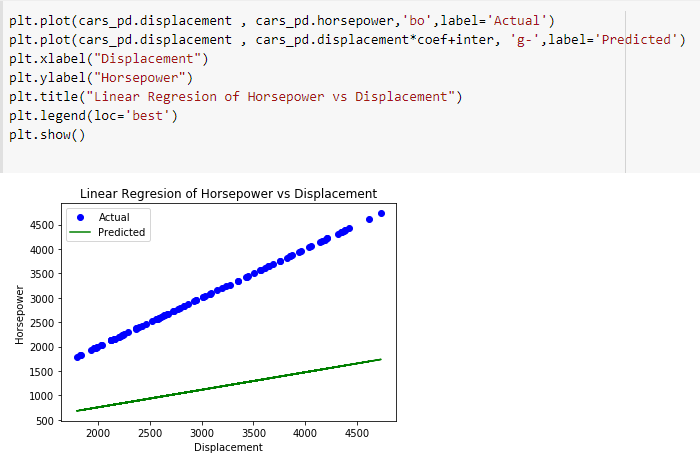
Created the feature vector using VectorAssembler



At 9:1 ratio train and test sets are randomly selected. Then the “LinearRegression” is applied. The R2 accuracy of linear model is 81.23%.



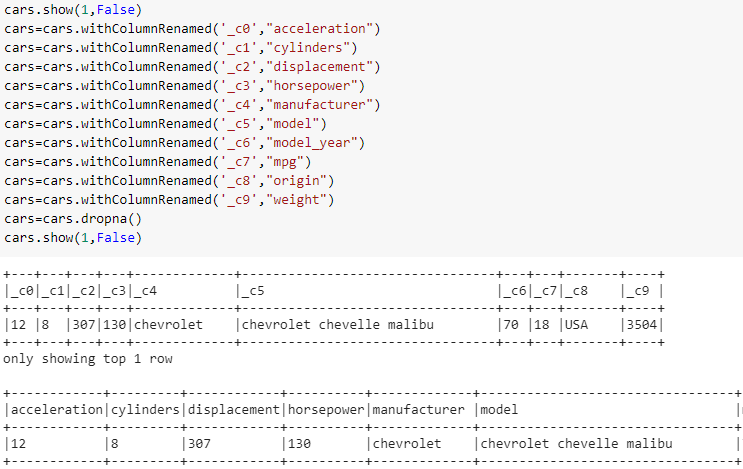


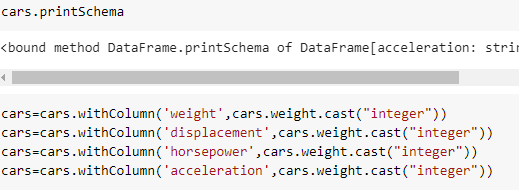
I couldn’t figure out the D3, so used matplotlib to plot the actual vs predicted.

**Problem 2:**

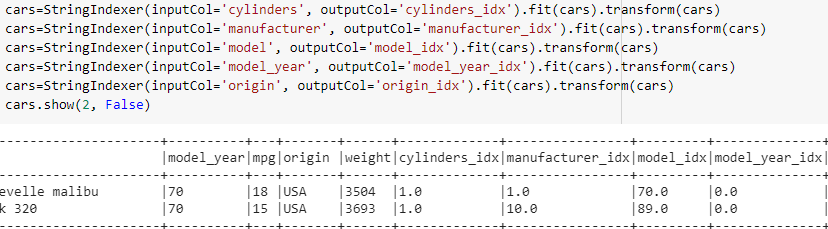
Treat: cylinders, displacement, manufacturer, model year, origin and weight as features and use linear regression to predict two target variable: horsepower and acceleration. Please note that some of those are categorical variables. Use test data to assess quality of prediction for both target variables.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

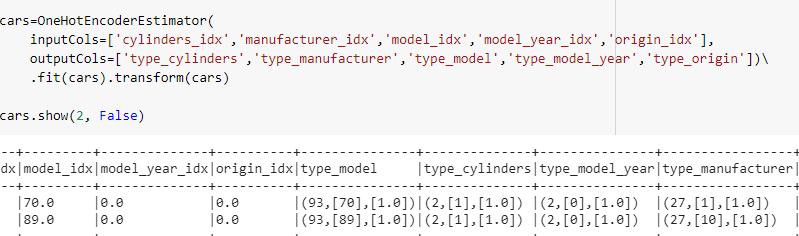
Renamed the default columns to respective names and dropped the missing rows

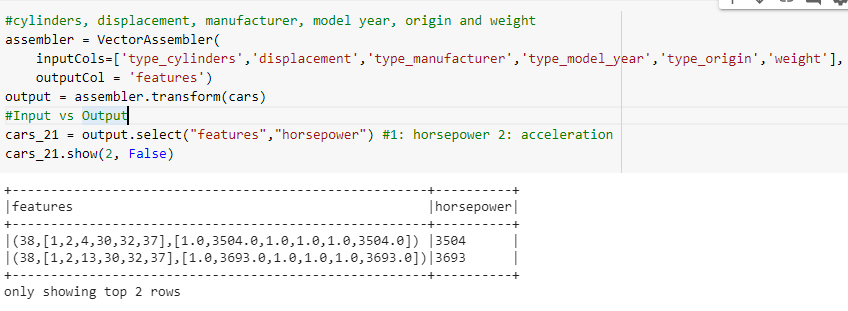


Checked the schema of spark data frame, and type casted to integer

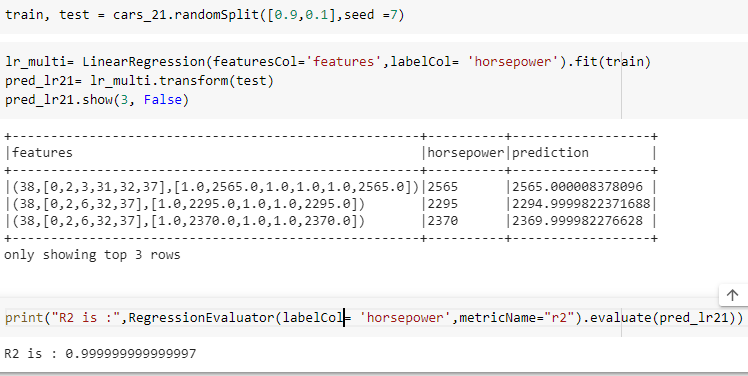
The cars data has many categorical variables. I have converted them to numerical indexes and then to a space vector with one hot encoding.



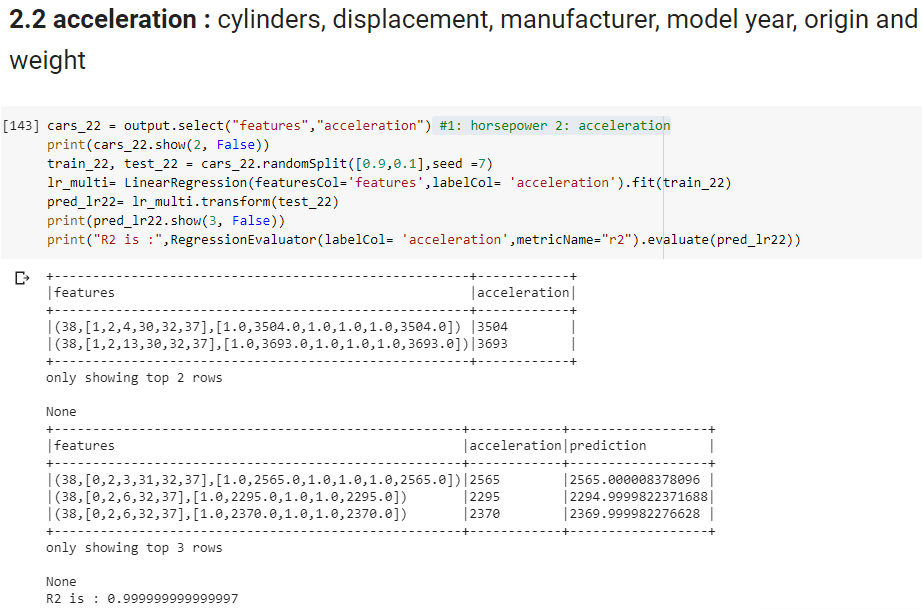


Assembled all the required columns as a dense vector named feature. This column will be used for all the remaining regressions. The master file “output” will be used as a master file and the required columns will be filtered from it for all the models. 

At 9:1 ratio train and test sets are randomly selected. Then the multiple logistic regression is applied. With addition of new features the R2 accuracy has increased to 99.99%.



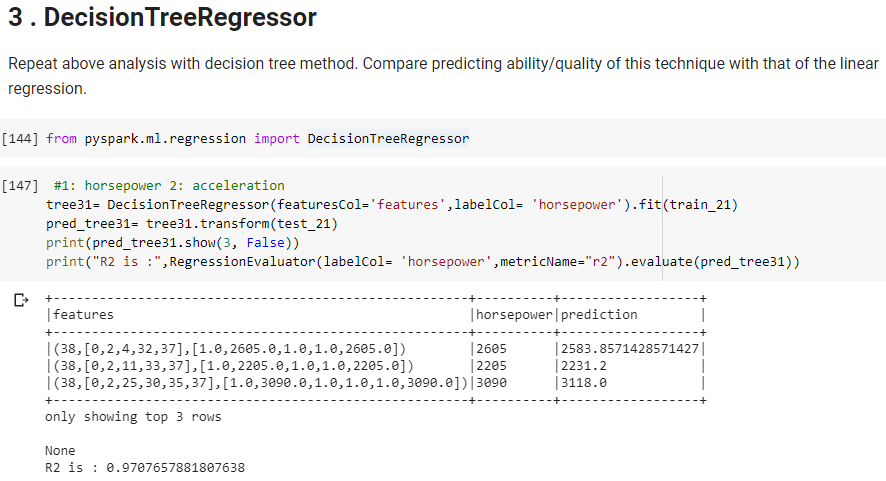
Similar steps are followed to find the linear relationship between the selected features and acceleration. At 9:1 ratio train and test sets are randomly selected. Then the multiple logistic regression is applied. With addition of new features the R2 accuracy has increased to 99.99%.



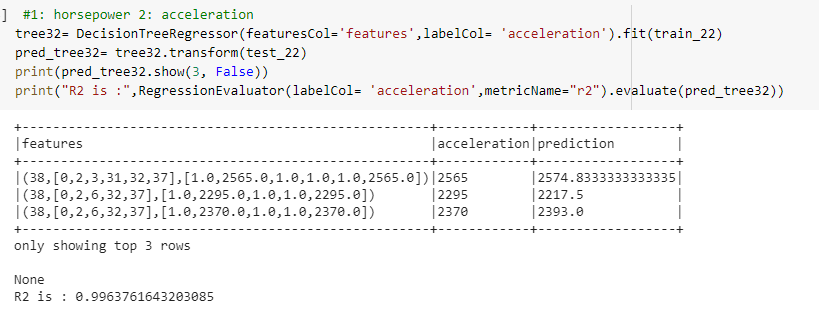
**Problem 3:**

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

Similar steps are followed to predict the horsepower of cars. At 9:1 ratio train and test sets are randomly selected. Then the “DecisionTreeRegressor” is applied. The R2 accuracy of tree model is 97.07%.



Similar steps are followed to predict the acceleration of cars. At 9:1 ratio train and test sets are randomly selected. Then the “DecisionTreeRegressor” is applied. The R2 accuracy of tree model is 99.64%.



------------------------------------------------------------------------------------------------------------