**#Explain the Project background and Project Objective**

The data is collected about the features of different cars such cylinder, displacement, weight, Engine Alignment, Gear type etc. with the target variable as miles per gallon which needs to be predicted in future if it is not known.

The project is to create a predictive model for cars performance regarding miles per gallon.

**#Explain the Data that you have**

The data is having 12 columns -11 features and 1 target variable as mpg(miles per gallon) . Since the data is with the labels so supervised learning will be used for modelling.

Mpg will be our target variable to be used for creating the machine learning model with rest of the variables as features containing both categorical and continuous values.

mpg, car names , cyl , disp ,hp ,drat ,wt , qsec ,vs ,am ,gear ,carb

Mpg - Miles / Gallon

Cyl - number of cylinders

Disp - Displacement (cu. in.)

Hp - Gross Horsepower

Drat - Rear axle Ratio

Wt - Weight

Qsec - 1/4 mile time

Vs - Engine (0 - v-shaped , 1-straight)

Am - Transmission (0=automatic , 1 = manual)

Gear - Number of forward gears

Carb - Number of carburettors

**#Explain Data Exploration process and the result**

Data exploration is always the first step in data analysis which helps us in exploring and visualising the data insights and get the information out data.

In data exploration we figured out the columns with categorical values and number of respective categories. Identification of numerical columns is also a part of data exploration to calculate the correlation metrics. We plotted the explored correlation metrics to figure out highly created variables and drop them to address speed and storage issues

**#Explain Data Correlation Test Result**

Data correlation tells about multi collinearity between the features and gives us highly correlated variables. In this case “hp”, “disc” and “wt” comes out to be highly correlated feature variables and can be dropped of the data for further predictive modelling.

**#Explain Data Visualisation Result**

Data visualisation is used to visualise and get hidden insights of the data. Here visualisation let us know about number of different categories present in categorical column and magnitude of correlation between the features given in the data.

**#Explain Data Preparation/Data Cleansing**

Data Preparation means preparing the data for predictive modelling after cleaning or removing less impacting features and highly correlated variables.

Its a process of cleaning and transforming a raw data before analysis to enrich the data for creating a better predictive model which includes reformatting, correcting and filtering.

**#Explain Feature selection, feature engineering**

Feature engineering is process to build a better and complex model that we can build compared to model with raw data.

Feature engineering also needs domain knowledge to select and transform most relevant variables from raw data when creating a predictive model.

Feature selection will help us to limit the features into a manageable numbers.

Feature selection is a process where we manually or automatically select the features that contribute the most to our prediction or output variables

We need features which are more conductive in creating a accurate ML model that needs a better feature selection and feature engineering makes this process easier.

For example :-

1. We removed the car name column as it was non numerical and non categorical variable which makes it completely irrelevant for predictive modelling.
2. We explored highly correlated variables as a part of feature engineering and drop those variables as a part of feature selection.

**#Explain Algorithm Selection**

Selection of algorithm depends upon type of data as target variable.

If Target variables is a categorical column then we used algorithm for classification such as KNN, Random Forest, SVM etc.

If target variable is continuous then we use algorithm for regression based prediction such as linear regression, Non Linear Regression etc.

In our problem statement mpg (miles per gallon) is the target variable whose values are continuous on a scale of 0 to 35 so we prefer using linear regression for predictive modelling. It will give us a best fit linear line within the data using given features and labels with trained values of coefficients and intercepts for each given feature to make prediction about mpg for unknown future inputs.

**#Explain Residual and Evaluate the model (RMSE)**

Residual gives the measure of how far the data points are from regression line I.e difference between the predicted and given sets of values.

RMSE (Root Mean Squared Error) explains the spread of residuals or we can say it tells us about how concentrated the data is around best fit line.

**#Explain the Model created**

Linear Regression creates a model by deriving a relationship between two variables by fitting a linear equation to observed data. One of the variable is explanatory or independent variable known as features and other is considered as dependent or target variable.

In our model we are considering mpg (miles per gallon) as dependent entity on given features of cars. Which means our model is designed to predict mpg value on a linear scale using the trained value of coefficient derived as a relationship between given features such as number of cylinders, engine alignment , types of transmission etc. with respect to given mpg values. Since the mpg feature is continuous in nature so type of algorithm should be regression based and that’s why we used linear regression for our predictive modelling.

**#Why are we using linear model?**

We need to predict a variable which is continuous in nature that’s why we are using a regression based algorithm.

After deselecting highly correlated variables we are left out with 8 variables to be used as feature for predictive modelling. A linear model gives us R-squared value above 0.85 representing a good model for training data.

If we use any non linear model the feature size will increase which may improve the model performance but can also lead us to more overfit model and processing and storage requirement will also get increased without improving the model performance on test data.

**#Explain the use of Azure machine learning in training the model**

Azure Machine Learning is a cloud based service for managing and creating a machine learning model and projects.

Using Azure machine learning studio we can do analysis and visualisation on a given data variables. It has inbuilt statistical tools to give statistics of any given features for feature engineering. The built in algorithms and evaluation techniques helps us in creating a predictive model and evaluate it in much less interval of time. The web deployment feature used to deploy the model created for real time use case purpose.

Our system specifications doesn’t play much role in predictive model performance as all sorts of calculations , processing and data storage will get managed by Microsoft Azure cloud service.

**Activity 1: -**

Define data management structures to align and streamline processes of data ownership,

retrieval, combination and usage

Explain the process of the data flow , retrieval , combination of both R and AML and its usage

Data Management can be considered as a process to be encapsulated at the very beginning to solve a business problem in better strutted and streamlined way.

It involves accessing data from trusted resources, Create a frame for better understanding of data, Data analysis and quality assessment.

Process of Data Flow for R Script:-

1. Import The Data
2. Identified Categorical Columns and usr Factor
3. Separate dependent and Independent Variables
4. Identified numerical columns and calculate multicollinearity
5. Removes highly correlated variables as data filtering
6. Use of Linear Regression algorithm on preprocessed data to create predictive model
7. Evaluate the performance of predictive model

Process of Data Flow for AML:-