A DEEP LEARNING MACHINE LEARNING PROJECT TO PREDICT THE PRICES OF HOUSES IN DOLLARS

A machine learning project for predicting the prices of various cars based on their features. Features include; **'Symbolling', 'Fuel Type', 'Aspiration', 'Door Number', 'Drive Wheel', 'Engine Location', 'Wheelbase', 'Car Length', 'Car Width', 'Car Height', 'Curb Weight', 'Engine Size', 'Bore Ratio', 'Stroke', 'Compression Ratio', 'Horsepower', 'Peak rpm', 'City mpg', 'Highway mpg'.**

Below is a table which shows features in a better presentation;

Car ID	symbolling	Car Name	Fuel type	aspiration	Door number	Car Body	Drive Wheel
1	3	alfa-romero giulia	gas	std	two	convertible	rwd
2	3	alfa-romero stelvio	gas	std	two	convertible	rwd
3	1	alfa-romero Quadrifoglio	gas	std	two	hatchback	rwd
4	2	Audi 100 ls	gas	std	four	sedan	fwd
5	2	Audi 100ls	gas	std	four	sedan	4wd
6	2	Audi fox	gas	std	two	sedan	fwd
7	1	Audi 100ls	gas	std	four	sedan	fwd

Engine location	Wheel base	Car length	Car width	Car Height	Curb weight	Engine type	Cylinder number
front	88.6	168.8	64.1	48.8	2548	dohc	four
front	88.6	168.8	64.1	48.8	2548	dohc	four
front	94.5	171.2	65.5	52.4	2823	ohcv	six
front	99.8	176.6	66.2	54.3	2337	ohc	four
front	99.4	176.6	66.4	54.3	2824	ohc	five
front	99.8	177.3	66.3	53.1	2507	ohc	five
front	105.8	192.7	71.4	55.7	2844	ohc	five
front	105.8	192.7	71.4	55.7	2954	ohc	five

Based on the above features, I built models to predict the prices of cars in dollars.

The total process includes;

Data Collection, Data Cleaning, Exploratory Data Analysis and Statistical Analysis, Feature Engineering, Model Building with Scikit-Learn and Keras (Comparing accuracy of both models), Model Testing, Model Deployment.

- **DATA COLLECTION**

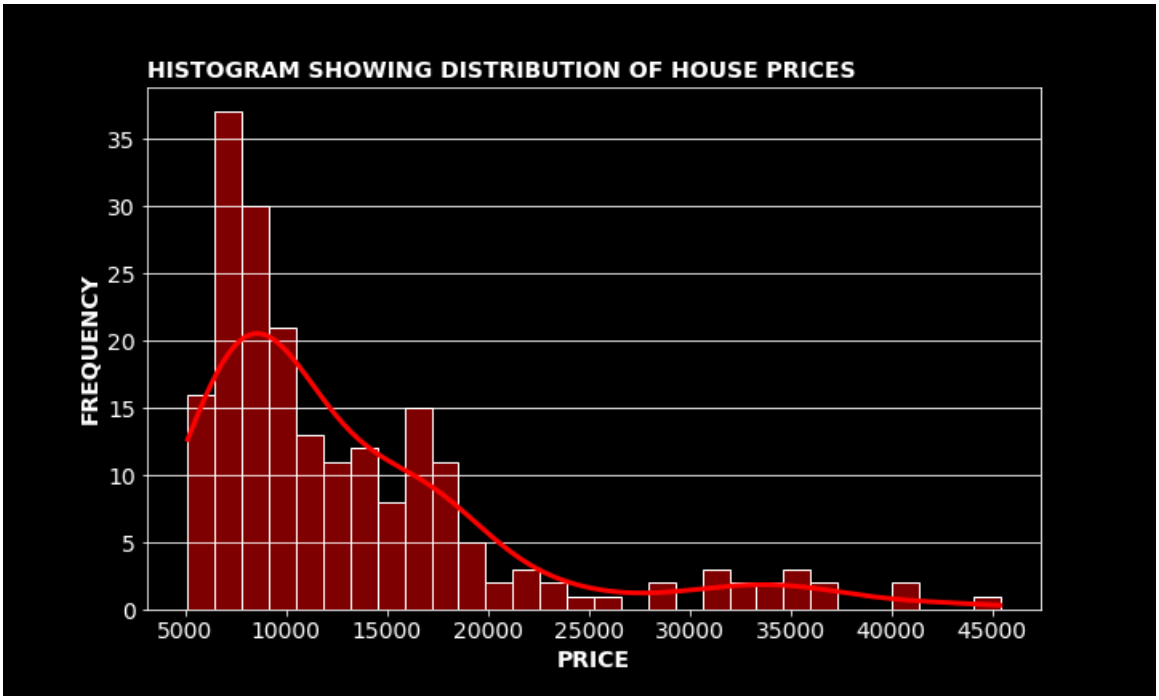
The given data was collected, consists of about 200 rows and 15+ columns as seen in the introduction.

- **DATA CLEANING**

After collection of data, the data had to be read through and cleaned (getting rid of null values, getting rid of outliers). This can be done by assigning values to entries where there were null values or even complete removal of rows that had such entries as it'd make it easier to get useful information without them. But fortunately, this data of cars given had zero outliers and even null values

- **EXPLORATORY DATA ANALYSIS**

This aspect entails visualization of our data, mainly graphically using libraries such as ***MATPLOTLIB, SEABORN AND CUFFLINKS***. Some of this exploratory data analysis include; Graphs showing correlation between attributes, Histograms showing frequency of prices of cars and so on.



The above graph shows a good type of distribution for the model; thus, no outliers are found and no null-valued rows are found.

- **FEATURE ENGINEERING**

Involves making the attributes (columns) in a dataset fit to be used for training and testing. For example, non-numerical values can't be used to predict a numerical figure, so columns like '**Car Name**' would have to be dropped. Other features like '**Fuel Type**', '**Door Number**' should also be dropped as they are non-numerical but they very much determine the price of the car so getting their Dummy Variables will be very much useful.

- **MODEL BUILDING**

In this aspect, I made use of Keras and TensorFlow to build a qualitative analysis model and even Scikit-Learn to evaluate the performance of the Model, the model Building entails Splitting the dataset into two parts, one for training and the other for testing, Normalizing the values to be used to train and test and adding Layers to the Model since it was built with Keras. \

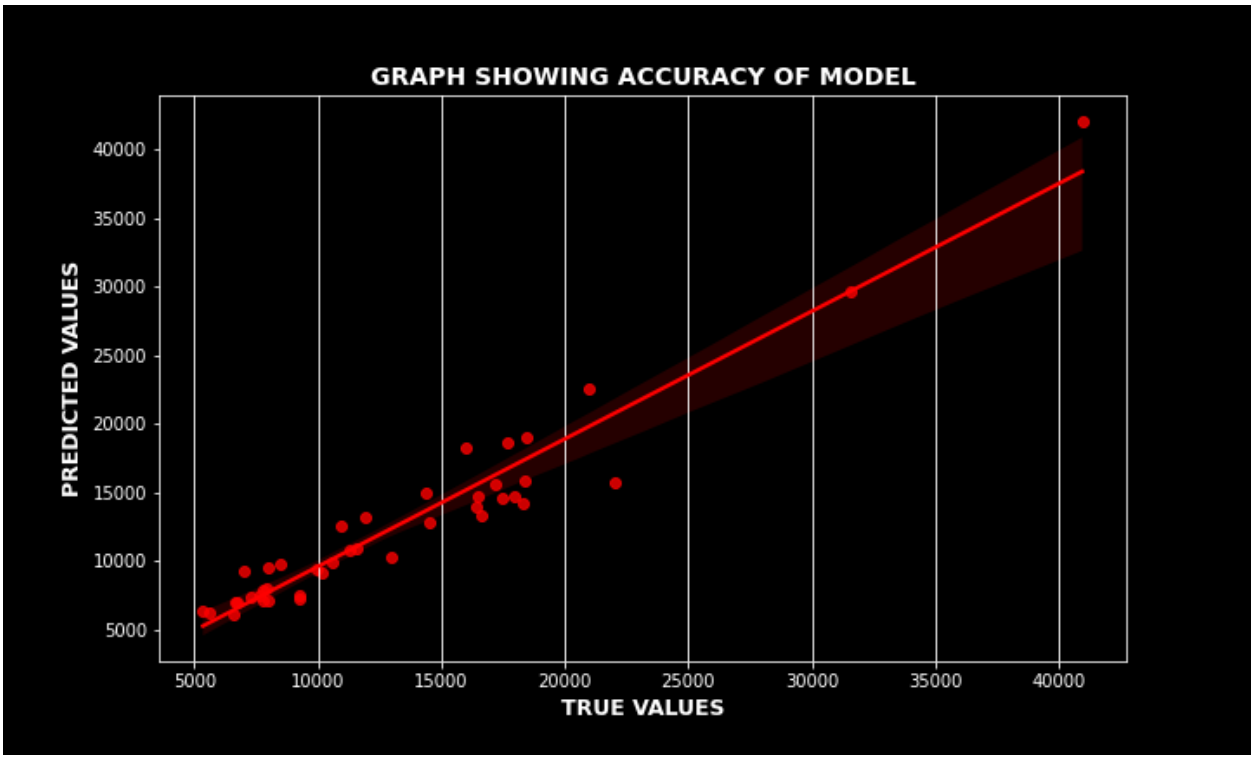
- **MODEL TESTING**

Quick Overview of the model Accuracy with different API'S and algorithm's

API USED (Application Programming Interface)	MACHINE LEARNING ALGORITHM	ACCURACY OF MODEL
SCIKIT-LEARN	Linear Regression	91%
	DecisionTreeRegressor	93%
	RandomForestRegressor	93%
TENSORFLOW-KERAS	Sequential Model	93%

The above table simply shows different machine learning algorithms used to train the models and that after several training and Testing, we arrived at 93% which occurs to be very much impressive.

Below is a regression plot which shows the predicted values against true values;



- **MODEL DEPLOYMENT**

Integrating a model into an existing production environment to make practical business decision based on data. This model hasn't been deployed yet, just tested.

This project is a very, much successful one due to the fact that the goal of the project was achieved and the Model is a very much accurate one.