Correlation Analysis to Identify the Effective Data for Prognostic Model Development using Regression Technique.

[1]Mujtaba Ashraf Qureshi
Mewar university, Chittorgarh, Rajasthan
Dept. of Information Technology.

[2]Dr. Azad Kumar Shrivastava
Assistant Professor, Mewar university,
Chittorgarh, Rajasthan.

Abstract: Correlation analysis is a widely used statistical method that recognizes suitable and interesting relationships in data. The conclusions and inferences drawn from such relationships of data are proved very much effective for machine learning, data mining and artificial intelligence like techniques to develop various prediction models. Machine learning, data mining and artificial intelligence methodologies are playing their efficient character to modernize every domain of the world. Machine learning is the practice employed to train the machines by the applications of specified datasets. To collect the required datasets from different locations and databases is done by data mining technology. Datasets acts a baseline for the development of the successful model. If there would be any noise, outlier or any unrelated data in datasets, then there is a good chance of failure of the model. Thus it becomes an imperative for data scientists to study and analyze the datasets to be used for model development. This experimental study is employed to study the existing relationships between various variables of car data set. Finally car dataset is used to devise the prediction model using regression technique.

Keywords: Correlation Technique, Regression method, Data Mining, Machine learning, Features, Prediction Model.

I. INTRODUCTION

Data mining, machine learning and artificial intelligence are interrelated technologies existing in modern digital world. In other words better to say that every field of modern world is influenced by use and applications of these advanced technologies. Machine learning is the practice employed to train the machines by the applications of specified datasets. To collect the required datasets and acquire useful knowledge from diverse databases is done by data mining technology. To acquire the hidden patterns of data, user oriented outline is made accessible by data mining process [4]. Datasets acts as a baseline for the development of the successful models by means of machine learning, data mining and artificial intelligence algorithms. If there would be any noise, outlier or any unrelated data in datasets, then there is a good chance of failure of the model. Thus it becomes an imperative for data scientists to study and analyze the datasets to be used for model development. In this experimental work 'car data' set (renamed) is imported into the platform -anaconda (Jupyter). A classification regression (simple) technique is used to devise the model based on the dataset taken into consideration. Before importing of this dataset, all the required libraries are also imported into the working environment for their implementation at various places in the program. Feature engineering also plays a vital role to process messy and bulky datasets and make them fit to be used for upcoming stages in the model development. This defines how to guide different data types of datasets like how to handle missing data, categorical data, numerical data, text data etc. Another algorithm is principal component analysis (PCA) reduces dimensions of the datasets as per need and requirement. In experimental environment a study of processed datasets is incorporated to study and investigate the variables of the dataset. *Correlation*, a statistical method employed to find correlation between different variables of a dataset with respect to our experiment. This is proved very efficient method to made research work successful. It depicts the need and essence of any variable for the project taken into consideration. Thus helps to avoid any kind of deviation from the goal of the research. To analyze the effect of some occurrences over other occurrences is also studied in various research fields. Some different fields where correlation technique is studied to find relationship between variables or occurrences such as agriculture, economy, image data analysis, big data etc. are prominent. Correlation is used also employed for financial market analysis to find associations

among stock returns in European markets and to find market trends in the world [2, 3]. Sometimes the rules defined by association techniques may be irrelevant or misleading. However correlation in such type of situations came as rescuer by defining the strength of association between items sets[1].In other words we can say that correlation is called a best statistical measure which is employed to find out the degree of association or connection between two or more variables. There are many questions taken into consideration such as;

1.1. What are the dimensions of the dataset?

1.2. What is the shape of dataset?

1.3. What is the number of input variables?

1.4. Which variable is selected as output variable?

1.5. What is the existing relationship between the employed variables of dataset?

1.6. Whether there is any missing or null value in the dataset.

1.7. What is the effect of input variables on the output of the model?

1.8. Whether there are any variable which decreases the efficiency of the model.

All the questions get resolved taken above mentioned dataset for the experimental purpose using statistical correlation technique. In this experimental work more focus is given to study the relationships existing between the variables which have very good impact upon the efficiency of the model development.

II. DATASET

An experimental approach is performed using anaconda (Jupyter) platform to study and investigate the dataset. Feature extraction is prerequisite for attaining and investigating eloquent knowledge from raw data [5].Following lines of code are used to import dataset into Jupyter notebook;

```
cardata=pd.read_csv("Desktop\car\carmpg.csv")
```

An overview of the imported dataset is depicted in table 1.

Table 1: Dataset

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino
...	...	...	...	...	...	...	...	...	...
393	27.0	4	140.0	86	2790	15.6	82	1	ford mustang gl

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
394	44.0	4	97.0	52	2130	24.6	82	2	vw pickup
395	32.0	4	135.0	84	2295	11.6	82	1	dodge rampage
396	28.0	4	120.0	79	2625	18.6	82	1	ford ranger
397	31.0	4	119.0	82	2720	19.4	82	1	chevy s-10

398 rows × 9 columns

The imported dataset now called *"cardata"* consists of 398 rows and 9 columns. These variables are used to train the algorithms and produce results when test data is supplied to them. To perform an experimental approach datasets are divided into training data (approximately 70 or 80 %) and testing data sets (approximately 30 or 20 %) to train and testify the developed model. Our data sets contain, mpg (milispergallon), cylinders, displacement, horsepower, weight, acceleration, model year, origin and car name as variables. An out variable is designated as 'mpg' column i.e. used as predicted class. Let's precede our experimental approach to know about dataset like missing values, outliers, data types, scaling factors etc. and thus its achievement on the model efficiency. The manageable and existing unprocessed data needs to assemble and stock in an organized format [6].

III. CORRELATION AND REGRESSION TECHNIQUES

Correlation is a statistical method employed to find correlation between different variables of a dataset with respect to our experiment. This is proved very efficient method to made research work successful. It depicts the need and essence of any variable for the project taken into consideration. Thus helps to avoid any kind of deviation from the goal of the research. In other words we can say that correlation is called a best statistical measure which is employed to find out the degree of association or connection between two or more variables. Coefficient of correlation designated by 'ρ' or r is employed to determine expanse of correlation. When there appears any change in one variable there also appears a change in corresponding variables also, however the change may be either positive or negative.

The value of Correlation coefficient exists between -1 and +1. If r=-1, then Correlation coefficient is perfect negative. Similarly, if r=+1, then Correlation coefficient is perfect positive. Also if r=0, then no Correlation exists between variables. Figure 1 shows various types of relationships existing between variables of different datasets. *In regression techniques*, an input of features is supplied. Response variable is also continuous like input features. In this manner the model is developed and presents an output when supplied by test data. Some of the important applications related to prediction such as to predict about sales performed in a particular month, stock price prediction etc. Linear regression and multiple regressions are two main basic types of regression technique. Linear regression is better to implement when there is one input feature and one out. A model learns based on input and output of features. Mathematics behind the linear regression is as follows;

Lt's take two data points size of a car and price of car. Here size of acts as input for algorithm and price of a car acts as predicted variable or output. Thus we can say that our data in this example lies in x-axis as size of a car and in y-axis as price of a car. In simple regression a line needs to fit and following equation is used to fit the line as.

$$f(x) = \beta_1 x + \beta_0$$

Output Shown by,

$$y_i = \beta_i x_i + \beta_0 + \epsilon_i$$

ϵ_i *=error term*
β_1 *and* β_0*=coefficient (Slope and intercept).*

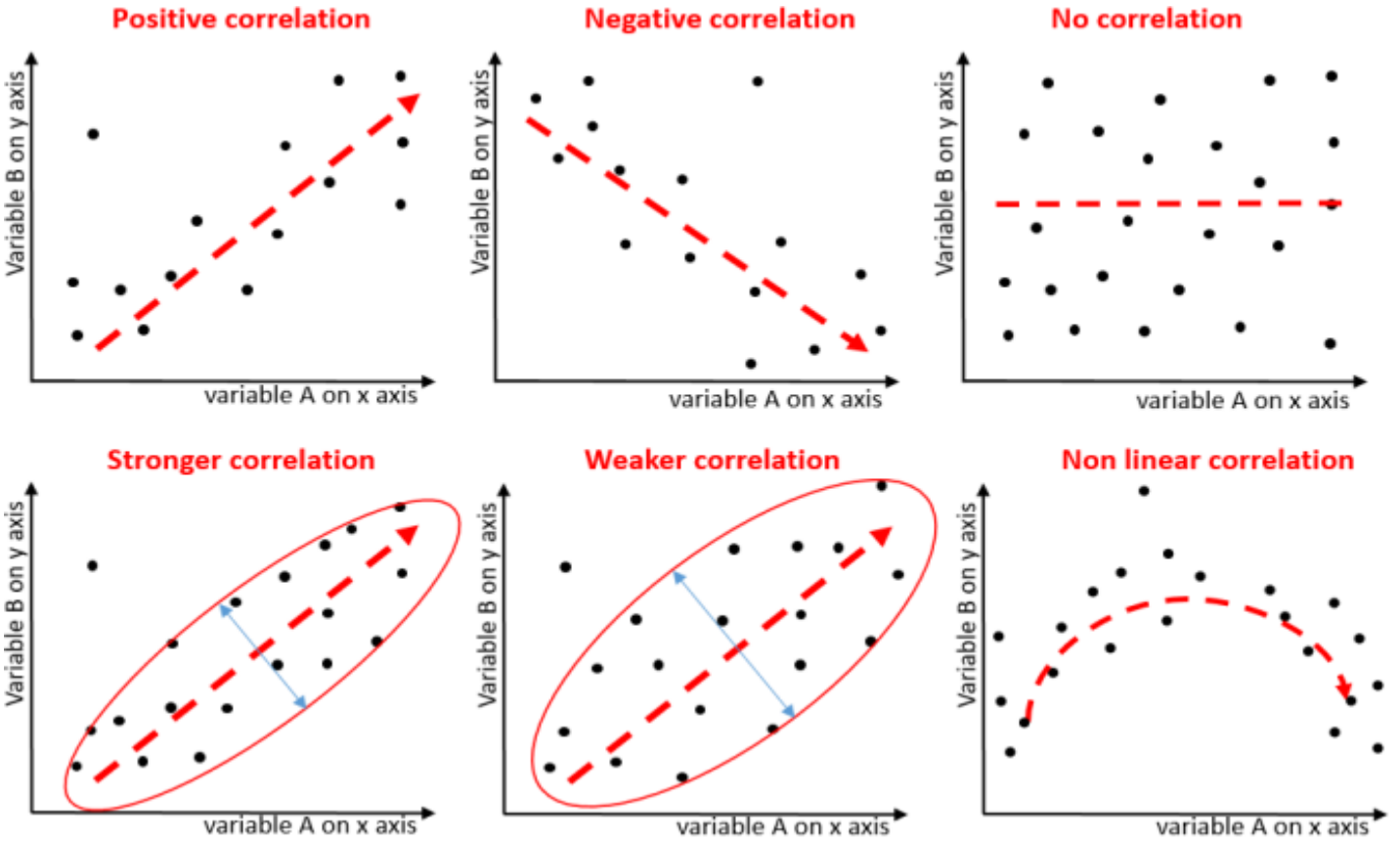

Figure 1: Types of Correlation.

Multiple regressions are also a linear regression having multiple features. In real world problems most of the datasets are composed of multiple features to present an output result. When output is not based on single input variable instead dependent upon higher degree of inputs, which is called polynomial regression. Model of polynomial regression is given below.

$$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots\ldots\ldots + \beta_p x^p + + \epsilon_i$$

Parametrs$= (\beta_0, \beta_1, \beta_2, \beta_3 + \ldots\ldots\ldots + \beta_p)$

Features $= (x, x^2, x^3 + \ldots + x^p)$

IV. EXPERIMENTAL APPROACH

To study and investigate our dataset, an empirical approach is employed using anaconda platform. Different kinds of methods are implemented to verify and validate the variables in the dataset. Consequently feature extraction is prerequisite for attaining and investigating eloquent knowledge from raw data [5].Dimensions regarding the employed dataset are already known to us as mentioned above To know the data types of different variables, a function .info() is employed. We get the data types of every variable and found that input 'hp' is object type that is not recommended and accept. So needs to convert into numeric form using related function. So here to the conclusion as regression deals with continuous values only. We removed null and missing values also to purify our data set further.

In this experimental approach, the fundamental goal is to study the correlation existing between the variables of a car dataset. Following lines of code is/are employed to get correlation information.

```
tele_data.corr()
```

Table 2: Statistical Measures

	mpg	cylinders	displacement	weight	acceleration	model year	origin
mpg	1.000000	-0.775396	-0.804203	-0.831741	0.420289	0.579267	0.563450
Cylinders	-0.775396	1.000000	0.950721	0.896017	-0.505419	-0.348746	-0.562543
displacement	-0.804203	0.950721	1.000000	0.932824	-0.543684	-0.370164	-0.609409
weight	-0.831741	0.896017	0.932824	1.000000	-0.417457	-0.306564	-0.581024
Acceleration	0.420289	-0.505419	-0.543684	-0.417457	1.000000	0.288137	0.205873
Model year	0.579267	-0.348746	-0.370164	-0.306564	0.288137	1.000000	0.180662
origin	0.563450	-0.562543	-0.609409	-0.581024	0.205873	0.180662	1.000000

A pair plot is function of sea born library. This function has best features related to visualization effect of data. It helps to understand the hidden information from data very precisely. A clear and lucid view of relationship existing between car data variables is shown. Clear visuals are seen from figure 2. A researcher or a data scientist could attain a very fruitful and helpful knowledge form pair plot graph, so that it can be used for the development of successful model further. Medical databases contain huge, unstructured and distributed datasets. The manageable and existing unprocessed data needs to assemble and stock in an organized format [6].

V. RESULTS

An investigation of experimental study reveals very deep knowledge that could be very much beneficial to devise various diagnosis models or to employ for any other related work. A number of questions could be understood easily and professionally by getting insights. There are various conclusions drawn when deep study is made to investigate about the relationship of variables. Some of the deductions made are;

- Acceleration is following normal distribution
- Displacement and weight have linear relationship.
- Displacement and mpg are negatively correlated.
- Horse power, displacement and acceleration are related linearly.

Finally processed dataset is divided into train (75%) and test (25%) part using train_test splitting method. A training set is given to linear regression to develop prediction model and test set is to know the performance results of this model. Performance results achieved from the devised model shows an acceptable level of accuracy i.e. 85.0% accuracy is achieved on test data.

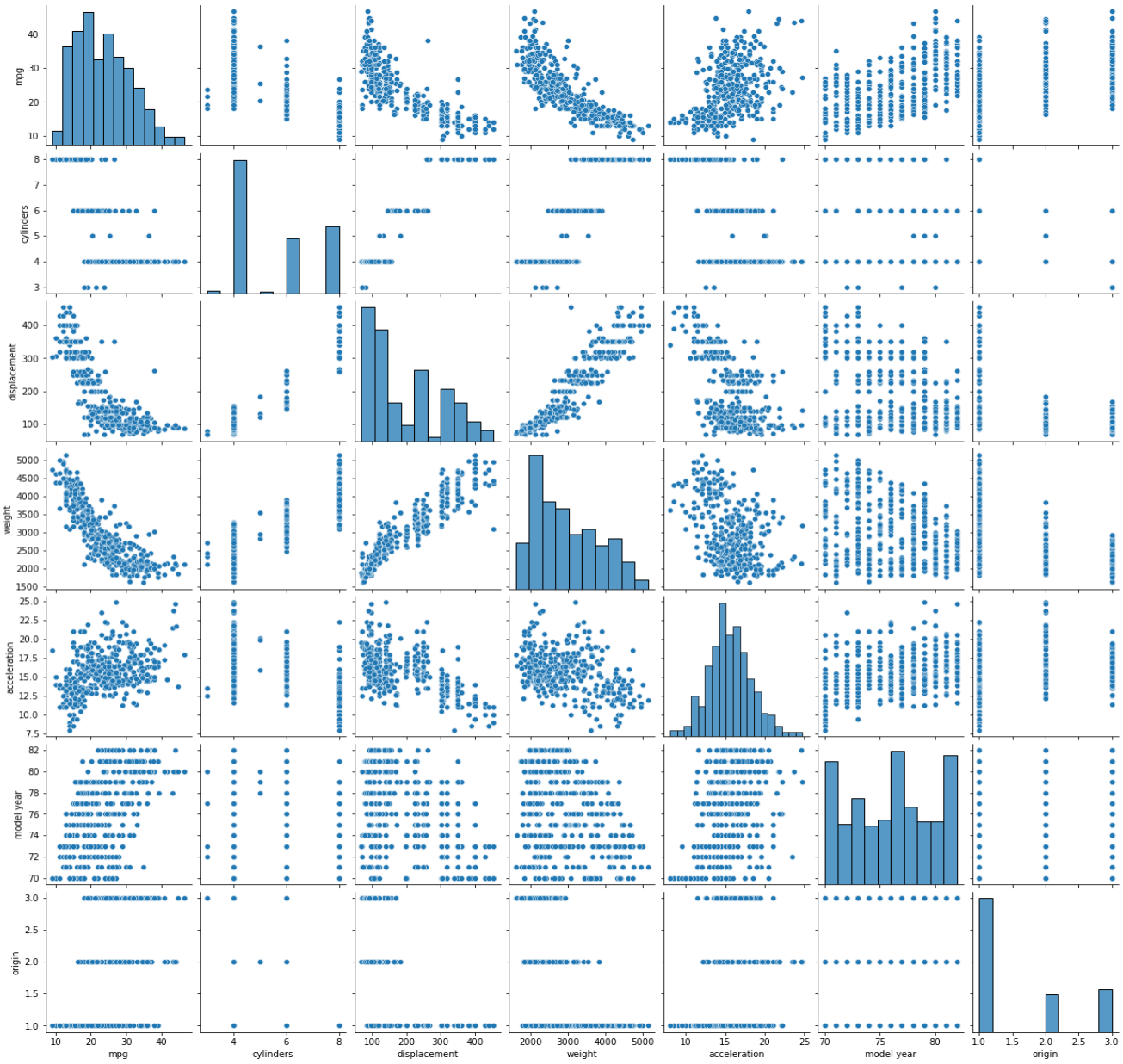

Figure 2: Pair Plot of variables

VI. CONCLUSION

To study and get perfect insights about the employed datasets is much need to strengthen data mining, machine learning and artificial intelligence further. This study makes us aware, what is the need of any variable? What is the strength of a variable? How much problematic a variable could be. Such studies acts as open solutions to questions like that. There were also various questions raised in introduction part and a number of questions are responded very well using this experimental study. An insight of pair plot and correlation table depicts relationship among variables like linear relationship, negative relationship or any other. To study such existing relationships are proved very much beneficial for a scientist or scholar.

This research work is open to future work as more insights and conclusions could be drawn from statistical results. Also negatively correlated variables could be removed from the employed dataset to enhance model accuracy and efficiency further.

VII. REFERENCES

[1]. Han, J.; Kamber, M. Data Mining: Concepts and Techniques, 2nd ed.; University of Illinois at Urbana-Champaign: Champaign, IL, USA, 2006; ISBN 9781558609013.

[2]. Syllignakis, M.N.; Kouretas, G.P. Dynamic correlation analysis of financial contagion: Evidence from the Central and Eastern European markets. Int. Rev. Econ. Financ. 2011, 20, 717–732. [CrossRef]

[3]. Sandoval Junior, L. Correlation of financial markets in times of crisis. Phys. A Stat. Mech. Appl. 2012,391, 187–208. [CrossRef]

[4]. Ralf Mikut and Markus Reischl Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 1, Issue 5, pages 431–443, September/October 2011.

[5]. Wagner, J.; Kim, J.; Andre, E. From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6–8 July 2005; pp. 940–943.

[6]. Wilson, P. e. (1998). Prediction of Coronary Heart Disease using Rusk Factor Categories. American Heart Association Journal.