

EXPLORATORY DATA ANALYSIS ON AUTOMOBILE MANUFACTURING USING MACHINE LEARNING

Ashik Varghese*¹, Dr. Anjana S Chandran*²

*¹Department Of Computer Applications, SCMS School Of Technology And Management, Muttom, Aluva, 683 106

*²Associate Professor, Department Of Computer Applications, SCMS School Of Technology And Management, Muttom, Aluva, 683 106

ABSTRACT

Automobiles has become a virtual part of our daily life. Usage and demand of cars are getting increased day by day. Ten or twenty years back, a company used to produce various models at a rate of two or three per years. But now the things got changed. Companies are in a heavy contest to satisfy their customers and to increase their sales and profit. So now they release new models even on a monthly basis. So a review on their models can help the company to make improvement changes on their upcoming models. So in this project, a dataset which contain all details regarding various models and variants are used as input and then it is proceeds using various tools to obtain valuable results on various parameters output.

Keywords: Machine Learning, Automobile, Comparison, Data Analysis.

I. INTRODUCTION

At present, industry has spent Transfer amount has been increased subsequently over the years. So, transfers play an important role in vehicle development. in step with the manufacture analysis produce their products basically for supplies to the vehicle manufacturers. These products are manufactured as per the drawings and specifications of the vehicle producer. Since the Automotive manufacturer's aim is to manage the value, the Ancillary supplier is left with a meagre margin when he supplies his produce to the vehicle manufacturers. These supplies are referred to as supplies to Original Equipment Manufacturers.

Exploratory Data analysis on Automobile analytics allows companies to form decisions supported performance of their manufacture products. Also, after analysis car companies and peoples to induce relevant information about the vehicles. The accuracy is extremely questionable during this case. So, with the assistance of machine learning algorithms, we are able to determine the accurate information of the vehicles.

Jupyter notebook is running using your browser, it could run locally on your machine as an area server or remotely on a server. the rationale it's called notebook is because it can contain live code, rich text elements like equations, links, images, tables, and so on. Therefore, you may have a awfully nice notebook to explain your idea and therefore the live code beat one document. Thus Jupyter notebook becomes really popular thanks to test ideas, writing blogs, papers and even books, as an example, this book is written entirely within Jupyter notebook. Of course, it's many other advantages, and that we will only cover the fundamentals of the Jupyter notebook to induce you started.

Jupyter notebook is running using your browser, it could run locally on your machine as an area server or remotely on a server. the explanation it's called notebook is because it can contain live code, rich text elements like equations, links, images, tables, and so on. Therefore, you may have a awfully nice notebook to explain your idea and also the live code dead one document. Thus Jupyter notebook becomes really popular thanks to test ideas, writing blogs, papers and even books, as an example, this book is written entirely within Jupyter notebook. Of course, it's many other advantages, and that we will only cover the fundamentals of the Jupyter notebook to induce you started. The different library tools like Anaconda, Sciket Learn, Numpy, Pandas, Seaborn using data analysis for more flexibility.

II. METHODOLOGY

Here the project is based on the exploratory data analysis in automobile manufacturing using machine learning with python. We use different modules and library functions and also a regression methods are used for data analysis process. Based on the data analysis using these properties and find the accurate and detailed information about automobile vehicles.

Our project goes through 7 phases: -

1) Learning:

In this step, we take that data and learn from it. And here comes the supervised and unsupervised learning.

2) Data Cleaning:

The first step is we'd like to clean and format the info. (That is because computers don't seem to be smart when it involves working out the difference between an image or text once we send it in), that the very first thing we do is typically clean the info so all our datas are in one file and text is being processed separately.

3) Prediction:

In this phase, we're actually using it or putting it into commercial use which is to try and do a prediction and on there now we've got our train model and our new data move and output goes to be a prediction of what we are searching for. we will see that within the sort of predicted data.

4) Exploratory Data Analysis:

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so on discover patterns, to spot anomalies, to test hypothesis and to test assumptions with the assistance of summary statistics and graphical representations.

5) Building a Model:

A dataset is that the place to begin in your journey of building the machine learning model. Simply put, the dataset is basically an $M \times N$ matrix where M represents the columns (features) and N the rows (samples). In this paper regression model is used.

6) Evaluating Model's Result:

Python's Matplotlib package can be used for plotting the lead to graphical format. Attributes like vehicles body, model, engine specs, color, rating, height and weight are used for analyzing data.

7) Conclusion:

Finally after the six phases the accurate and detailed information's are fended using the data analysis technique in automobile manufacturing using machine learning.

III. MODELING AND ANALYSIS

The data analysis method are mainly used with regression technique. In statistics, exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model are often used or not, but primarily EDA is for seeing what the information can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the info, and possibly formulate hypotheses which will cause new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables PRN. EDA encompasses IDA. Analyzing make more helpful permanently findings about corresponding systems. Automakers are using Big Data already to manufacture better cars and to produce quality customer services. Vehicles interconnected with Big Data will change our lives indeed. The role of massive Data Team is crucial. For starters, it analyzes data regarding planning, sales, or marketing.

The main focus is to analyze the automobile manufacture details using the machine learning techniques. The information is fetched from the kaggle website. The automobile datasets are mainly contains of different types of detailed specifications related in automobile vehicles like engine side, body, ratings, manufacture, power, speed etc. Then Data cleansing is finished which helps to get rid of issues like duplicate entries, unwanted characters, split lines etc. Exploratory data analysis is conducted within the next innovate order to grasp and explore the structure of information. With the assistance of regression model, relationship between the target and predicted variable is discovered. In line with the experimental results, change was clearly visible between actual and predicted market values. Results are visualized in graphical form used with the python matplotlib functions.

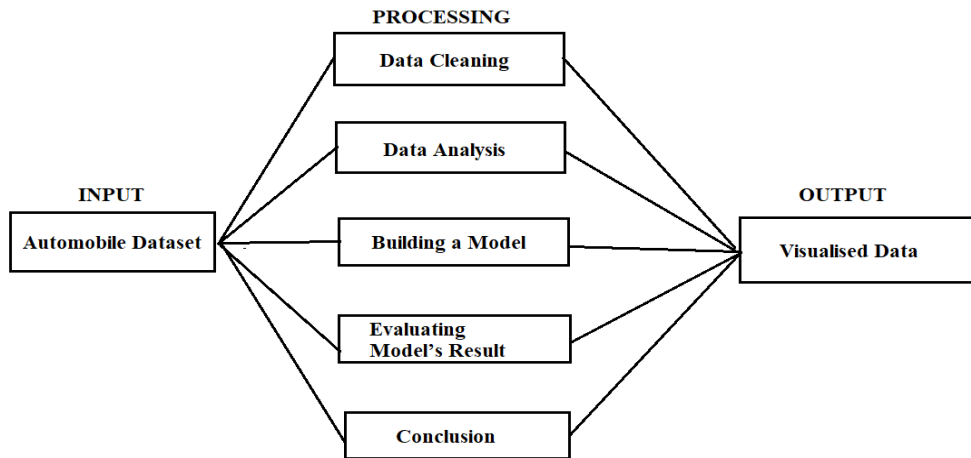


Figure 1: General Experiment Setup

At first the automobile dataset is taken as input .Then the data is cleaned as it removes unwanted or corrupted entries from the dataset .Cleaned data is then analyzed. After that a regression model is built were is explored required data. After building the model, the result of the model is evaluated .The conclusion is reached with the help of visualized data using matplotlib function. The detailed experiment is shown in figure 2.

| | symboling | normalized-losses | wheel-base | length | width | height | curb-weight | engine-size | bore | stroke | compression-ratio | horsepower |
|-------|------------|-------------------|------------|------------|------------|------------|-------------|-------------|------------|------------|-------------------|------------|
| count | 205.000000 | 205.000000 | 205.000000 | 205.000000 | 205.000000 | 205.000000 | 205.000000 | 205.000000 | 205.000000 | 205.000000 | 205.000000 | 205.000000 |
| mean | 0.834146 | 122.000000 | 98.756585 | 174.049268 | 65.907805 | 53.724878 | 2555.565854 | 126.907317 | 3.329751 | 3.255423 | 10.142537 | 104.253659 |
| std | 1.245307 | 31.681008 | 6.021776 | 12.337289 | 2.145204 | 2.443522 | 520.680204 | 41.642693 | 0.270844 | 0.313597 | 3.972040 | 39.519219 |
| min | -2.000000 | 65.000000 | 86.600000 | 141.100000 | 60.300000 | 47.800000 | 1488.000000 | 61.000000 | 2.540000 | 2.070000 | 7.000000 | 48.000000 |
| 25% | 0.000000 | 101.000000 | 94.500000 | 166.300000 | 64.100000 | 52.000000 | 2145.000000 | 97.000000 | 3.150000 | 3.110000 | 8.600000 | 70.000000 |
| 50% | 1.000000 | 122.000000 | 97.000000 | 173.200000 | 65.500000 | 54.100000 | 2414.000000 | 120.000000 | 3.310000 | 3.290000 | 9.000000 | 95.000000 |
| 75% | 2.000000 | 137.000000 | 102.400000 | 183.100000 | 66.900000 | 55.500000 | 2935.000000 | 141.000000 | 3.580000 | 3.410000 | 9.400000 | 116.000000 |
| max | 3.000000 | 256.000000 | 120.900000 | 208.100000 | 72.300000 | 59.800000 | 4066.000000 | 326.000000 | 3.940000 | 4.170000 | 23.000000 | 288.000000 |

Figure 2: Summary Statistics of Automobiles Specs

The Summary Statistics of Automobiles specification is shown in figure 2. Various specification parameters such as symboling, normalizes losses, wheel base, length, width, height, curb-weight, engine size, bore, stroke, compression ratio and horsepower were taken into consideration during experiments. There were 205 entries received after data cleaning process. Mean horse power was observed as 104.25.

SETTING UP THE SYSTEM

Jupyter:

The Jupyter Notebook is an open-source web application that permits you to form and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and far more.Jupyter Notebook is our surroundings used for computation. Jupyter can be installed by using ‘pip install notebook’ and command for server running is ‘jupyter notebbok’.

Numpy:

Numpy is a math library to figure with n-dimensional arrays in Python. It enables you to try and do computation efficiently and effectively. It's better than regular python due to its amazing capabilities.. NumPy is installed by using ‘pip install numpy’ command.

Matplotlib:

Matplotlib is a plotting package that provides 2D plotting as well as 3D plotting. Matplotlib is installed by using ‘pip install matplotlib’ command.

Pandas:

Pandas library could be a very high-level python library that has high-performance, easy to use data structures. it's many functions for data importing, manipulation and analysis. specifically, it offers data structures and operations for manipulating numerical tables and statistic. Pandas is installed by using 'pip install pandas' command.

Seaborn:

Seaborn could be a Python data visualization library supported matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. For a quick introduction to the ideas behind the library, you'll be able to read the introductory notes. Seaborn is installed by using 'pip install seaborn' command.

IV. RESULTS AND DISCUSSION

Data analysis in automobile data set are analyzed with help of machine learning technique. After analyzing the result, different types of automobile specification datas are will be fended. In this analyzing process the datas are will be represented accurate table value and graphical representation. It is clearly defined the main vehicle specifications like fuel type, doors pattern, engine, body etc are will be founded

Table 1: Data Analyzed in Table Format

| Sl.No | Brand | Fuel Type | Aspiration | Doors | Style | Drive Wheels | Engine location | Peak-RPM | Normalized losses |
|-------|-------------|-----------|------------|-------|----------|--------------|-----------------|----------|-------------------|
| 1 | Alfa-altraz | Gas | Std | Two | Xuv | 4wd | Front | 3000 | 122 |
| 2 | Setro | Gas | Std | four | Sedan | rwd | Front | 3000 | 123 |
| 3 | Xuv | Diesel | Std | four | Xuv | 4wd | Front | 5000 | 122 |
| 4 | Sande | Gas | Std | four | Hachback | rwd | Front | 3000 | 164 |
| 5 | Supra | Gas | Std | Two | Sedan | rwd | Front | 6000 | 165 |
| 6 | LC | Diesel | Std | four | Xuv | 4wd | Front | 5000 | 122 |
| 7 | Paggie | Gas | Std | four | Sedan | rwd | Front | 3000 | 125 |
| 8 | Samro | Gas | Std | four | Sedan | rwd | Front | 3000 | 121 |

In table 1, a detailed comparison of car models are shown. From the table most cars are having gas/petrol as fuel type and four door style appered to be more common among in the models. Brand Supra has marked with maximum RPM of 6000 rpm.

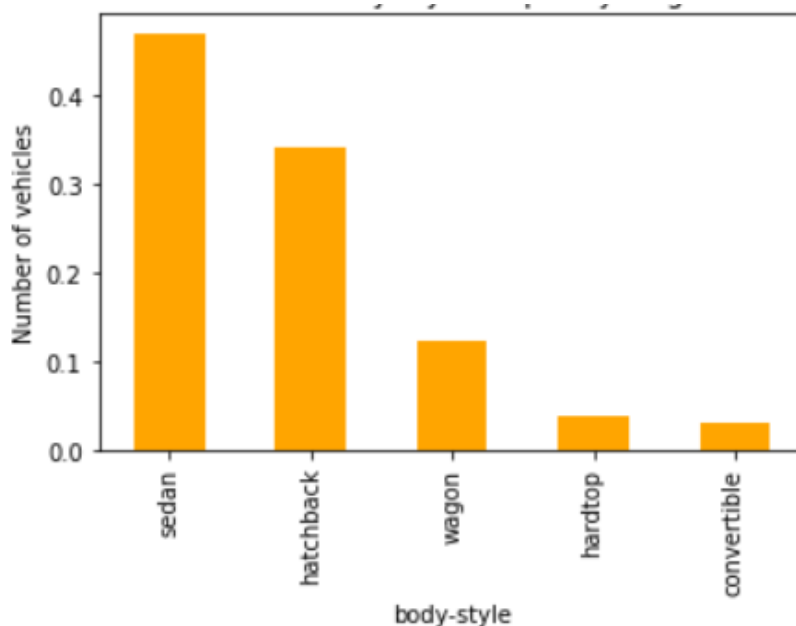


Figure 3: Number of Body style Frequency Diagram

In figure 3, its comparison between the body style and the number of vehicles. It was found that the car with body style 'sedan' has the most number of count whereas 'convertible' has the least. Hatchback was found to be the second favorite body style model.

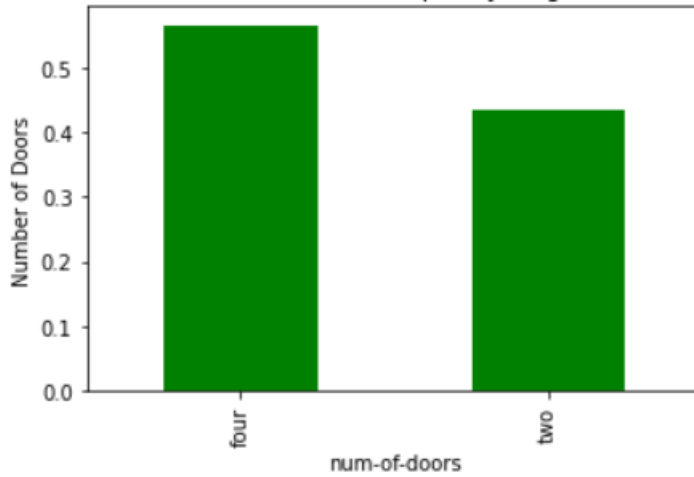


Figure 4: Number of Door Frequency Diagram

In figure 4, its comparison between the door style and the number of vehicles. It was found that the most car has the door style 'four'. Two door car models are comparatively low.

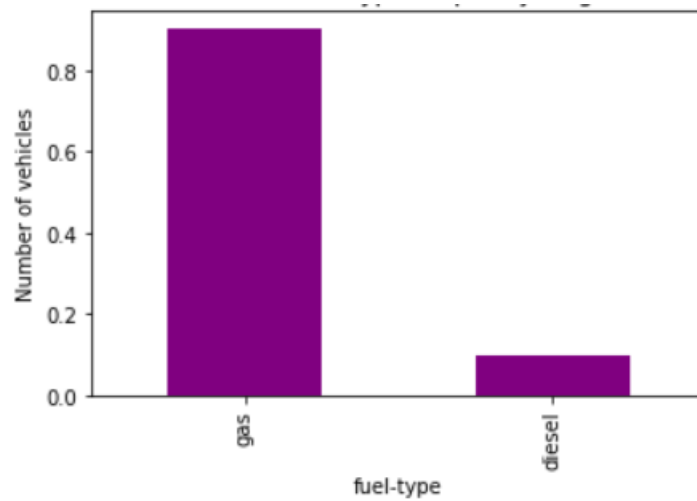


Figure 5: Number of Fuel Type Frequency Diagram

In figure 5, its comparison between the fuel type and the number of vehicles. It was found that the most cars have gas/petrol as its fuel. Surprisingly, count of cars with diesel as fuel is comparatively very negligible.

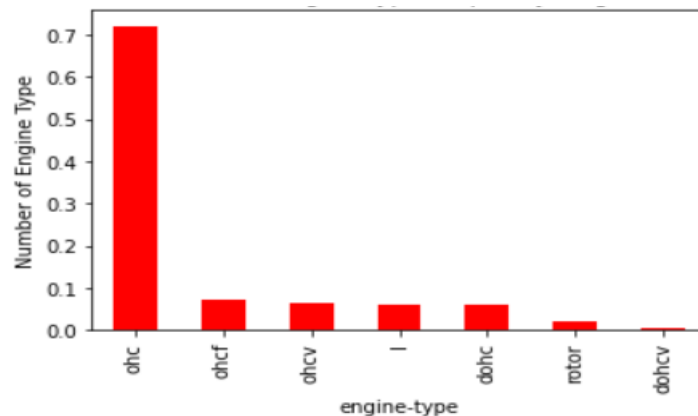


Figure 6: Number of Engine Type Frequency Diagram

In figure 6, its comparison between the engine and the number of vehicles. It was found that the most cars have engine style 'ohc'. Dohcv has the least number of count. Huge difference were seen with 1st and 2nd engines. Clear dominance of ohc engine was clearly visible.

V. CONCLUSION

From these results we can analyze the different types of automobile are included car has a Curb Weight is in range 1900 to 3100, the Engine Size is in range 60 to 190, most vehicle has horsepower 50 to 125, peak rpm is mostly distributed between 4600 to 5700, more than 70 % of the vehicle has Ohc type of Engine, 57% of the cars has 4 doors, gas(petrol) is preferred by 85 % of the vehicles, most produced vehicle are of body style sedan around 48% followed by hatchback 32%.

VI. REFERENCES

- [1] Chia-Nan Wang , Hector Tibo and Hong Anh Nguyen "Malmquist Productivity Analysis of Top Global Automobile Manufacturers" April 2020
- [2] Karl Koscher," Experimental Security Analysis of a Modern Automobile"
- [3] Julian Happian-Smith "Modern Vehicle Design analysis".IEEE Jan 2021
- [4] Razman Bin Mat Tahar & Ali Asghar J.Adham" Design and Analysis of Automobiles Manufacturing System Based on Simulation Model"(June 2010)
- [5] Ching-Shin Norman Shiau , Jeremy J. Michalek a,b, Chris T. Hendrickson "A structural analysis of vehicle design responses to Corporate Average Fuel Economy policy" May 2008
- [6] Do-Hyoung Kim , Hyun-Gyung Kim , Hak-Sung Kim "Design optimization and manufacture of hybrid glass/carbon fiber reinforced composite bumper beam for automobile vehicle" June 2015
- [7] Christopher D. Ittner & Marshall L. Fisher" The Impact of Product Variety on Automobile Assembly Operations: Empirical Evidence and Simulation Analysis"
- [8] Mehmet Cakmakci"Process improvement: performance analysis of the setup time reduction-SMED in the automobile industry"
- [9] Arunmozhi Manimuth" Modelling and analysis of artificial intelligence for commercial vehicle assembly process in VUCA world: a case study" 2020
- [10] Surender Kumar and Rabinder Singh Bharj" Experimental analysis of a light weight refrigerated electric vehicle" October 2020
- [11] Olha Sakno , Tetiana Kolesnikova ,Vasyl Ollo "Simulation of Vehicle Maintenance on the Basis of Functional Resonance Analysis Method using Function-Oriented Technology" Spetember 2020
- [12] Hui Wan , Jens Mathis Rieckmann , Qianqian Zhang and Qiao Ping" Research on New Energy Automobile Manufacturing Service Derivatization Based on TRIZ" August 2020
- [13] Anyi Zhang; Jianwen Wang" A method for analyzing abnormality of automobile sunroof manufacturing process by using Bayesian method" November 2020
- [14] S.Kaliappan" Mechanical Design and Analysis of AGV for Cost Reduction of Material Handling in Automobile Industries" January 2018
- [15] Mohammed N. Jamala" Predicting MPG for Automobile Using Artificial Neural Network Analysis". 2012
- [16] Dr. S. Smys" ROBOT ASSISTED SENSING, CONTROL AND MANUFACTURE IN AUTOMOBILE INDUSTRY" 2019
- [17] Ashok Deo" Fuel consumption of new passenger cars in India: Manufacturers' performance in fiscal year 2018-19" May 2020
- [18] Razman Bin Mat Tahar" Design and Analysis of Automobiles Manufacturing System Based on Simulation Model" June 2010
- [19] Sayanan Sivaraman" A Survey of Vision-Based Automobile Vehicle Detection, Tracking, and Behavior Analysis" December 2013
- [20] Soroush Shafiee" Investigating the Impacts of Plug-in Hybrid Automobile Electric Vehicles on Power Distribution Systems" September 2013
- [21] Alireza Khaligh" Comprehensive Topological Analysis of Conductive and Inductive Charging Solutions for Plug-In Automobile Electric Vehicles". October 2012
- [22] Chaozhong Wu" Visualization Analysis of Intelligent Vehicles Research Field Based on Mapping Knowledge Domain" May 2020

- [23] ZhaohuaWang” Fuel economy of Automobile car manufacturers: An efficiency analysis perspective” April 2021
- [24] Roland Örtengren “The impact of poor assembly ergonomics on product quality: A cost-benefit analysis in car manufacturing” November 2009
- [25] Parvathy Vinod Sheela & Fred Mannering” The effect of information on changing opinions toward autonomous vehicle adoption: An exploratory analysis” 2019.