IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 6, June 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106052 |

Exploratory Data Analysis on Automobile Dataset

Tanvi Malpure

Student, Department of Electronics and Telecommunications, Cummins College of Engineering, Pune, India

ABSTRACT : The automotive market is characterized by cut-throat competition. Every vehicle OEM aims to capture the market by ensuring their vehicle offerings attract a wider customer segment compared to their competition. This would be possible only if careful and accurate data analysis supported by the modern methods borrowed from the emerging discipline of data science is carried out. Such an exercise would soundly support the further decision-making process towards ensuring launch of product equipped with specifications and feature mix and adds significant value to the end customer experience. This paper is one such attempt in this direction. It uses a publicly available data source containing relevant data. The paper is aimed to showcase an appropriate methodology towards attaining the aim of accurate market-entry decisions discussed above

KEYWORDS : Data Visualization, Exploratory Data Analysis, Automotive Market Analysis, Recommendation, Competition Analysis

I. INTRODUCTION

The automotive industry is made up of a diverse group of businesses and organisations engaged in the design, development, production, marketing, and sale of automobiles as well as their repairs and modifications. Fast-moving technical developments, shifting consumer behaviour, and disruptions brought on by the recent pandemic are all driving significant changes in the automotive business. Industry participants are developing fresh and solid business plans to be relevant throughout the transition as autonomous driving, cloud computing, electric vehicles, machine learning, blockchain, networking, etc. become more widely adopted. Additionally, the pandemic-related global technological revolution in supply chain and logistics has forced the automotive industry to adopt new approaches supported by technological advancements. Data scientists use exploratory data analysis (EDA), which frequently makes use of data visualisation techniques, to examine and study data sets and summarise their key properties. It makes it simpler for data scientists to find patterns, identify anomalies, test hypotheses, or verify assumptions by determining how to modify data sources to achieve the answers they need. EDA's major goal is to encourage data analysis before making any assumptions. It can assist in finding glaring errors, better understanding data patterns, spotting outliers or unusual occurrences, and discovering intriguing relationships between the variables. A cloud-based platform called Databricks offers tools for data engineering, data science, and analytics to assist businesses in managing and analysing massive amounts of data. The platform provides several integrated tools for data intake, processing, storage, and analysis and is built on top of Apache Spark, an open-source big data processing framework. In a team-based, cloud-based environment, Databricks users can carry out operations like ETL (extract, transform, load), machine learning, and stream processing. Additionally, the platform provides automated infrastructure provisioning, management, and security features that let users concentrate on tasks related to their data rather than the underlying infrastructure. For individuals and small teams, Databricks provides a Community Edition version of its platform that is free to use. A powerful tool for data exploration, experimentation, and learning, the Community Edition offers access to a constrained number of features and resources. Users of the Community Edition have access to a notebook environment for data exploration and analysis, may build data processing clusters, and can collaborate with other users via shared notebooks and dashboards.

II. RELATED WORK

[1] Explanatory research design was done due to which all the factors having impact on the customer satisfaction level is explored. Using the quantitative approach in this primary study, the data was collected from the survey-based responses. Different visualization techniques and correlation is used to get the result. This study thus suggests that along with providing the innovation-based facilities, connected technologies is helpful for the social and environmental welfare too, thus, it is important to make this process more cost-effective and in order to target the middle-income

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

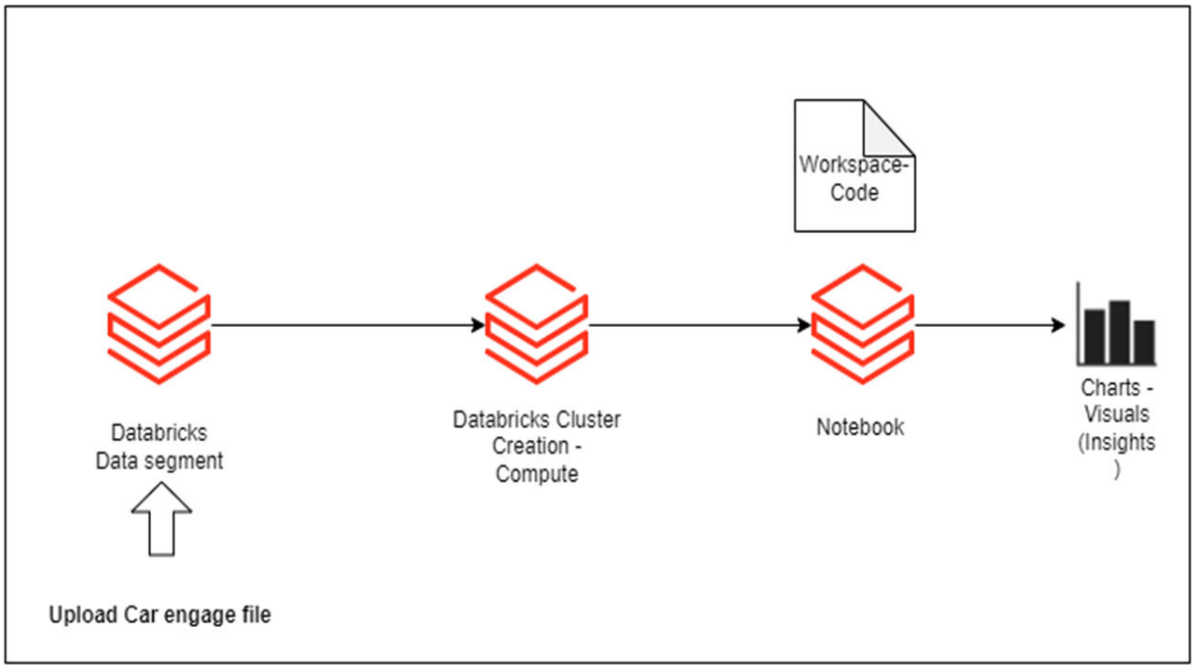
|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106052 |

group people too. [2] Based on the Service and Management center for EVs (SMC-EV), this paper conducts a statistical analysis of the travel data of more than 1,000 FCVs in operation. Different graphs like side by side bar charts, line charts and combination of line and bar charts are used for getting useful insights. [3] This paper uses exploratory data analysis in automobile manufacturing using machine learning with python. Data analysis in automobile data set is analyzed with help of machine learning technique. After analyzing the result, different types of automobile specification is fended. In this analyzing process the data is represented as 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. [11] This paper provided a summary of the most common data analysis techniques. It first describes data preparation methods which are an essential process in analyzing data. Then, common methods are reviewed, and the tools for the most important techniques are discussed. Qualitative data analysis and its strategies are also discussed more specifically in the final section

II. METHODOLOGY

The project is based on exploratory data analysis on car dataset with python. The Community edition of Databricks for visualisation was used for this work. This edition is available as a freeware. This facilitated the formulation of different plots and graphical representations for the data analysis process. From this result we can derive similar patterns in the data.



Architecture Diagram

1. Defining the question

Propose a methodology for estimating the possibly most attractive specifications, features at a price point that is likely to attract a wider market segment. It needs to be noted that this methodology is only a first step setting the course of further decision making process which may need jury trials, customer surveys and value engineering exercises to refine/modify the findings/proposals arrived at by this strategy

2. Data Cleaning

Link for dataset used: https://www.kaggle.com/datasets/medhekarabhinav5/indian-cars-dataset

This data is adequate to formulate the preliminary decision-making methodology this paper intends to showcase. For a real-life problem more authentic data-source with carefully tailored information packaged in rows and columns may be required. The raw data obtained from the above source had to undergo some data cleaning such as changing a column name from Ex-showroom price to price, changing data type to integer wherever applicable for ease of plotting etc.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106052 |

3. Exploratory Data Analysis

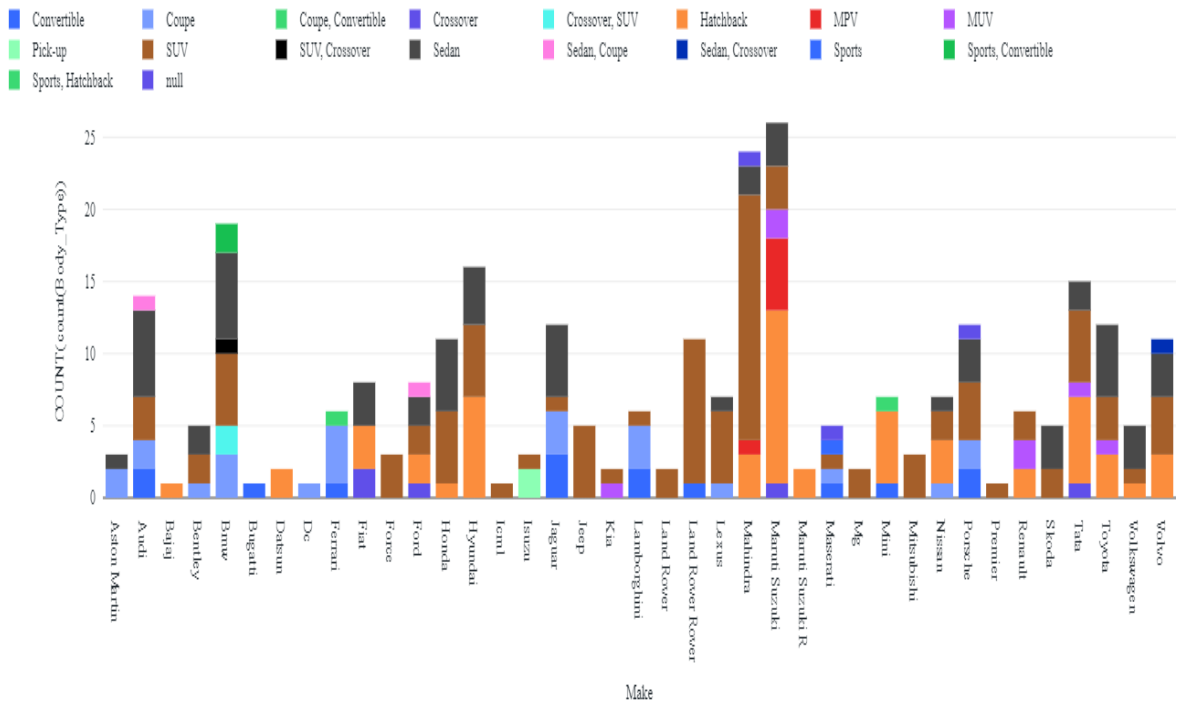
As a next step EDA was performed to find patterns, anomalies, to test hypothesis and to test assumptions from graphical representations.

III. RESULTS AND DISCUSSION

III.1

The first step of analysis focussed upon identification of those vehicle types which have less competition. This is more likely to identify the relatively untapped market opportunities. Appendix II discusses various body types available in the market for the vehicles.

Count of each body type based on Make of a vehicle analysis.



Graph is plotted for showing the correlation between Vehicle Make – Dimensions – Body Type

This analysis has been done to identify the vehicle segments where there is less competition.

From this graph it can be concluded that only Mahindra and Maruti Suzuki have MPV and it could be a good choice.

Power of the EDA tools that have been used for the above plot is evident here. A single bar can include multiple-coloured segments so that a "multi-dimensional" information can be accommodated in a single bar chart. This allows the decision makers to compare many aspects involved in a decision by studying a single plot and speedily arrive at the decision.

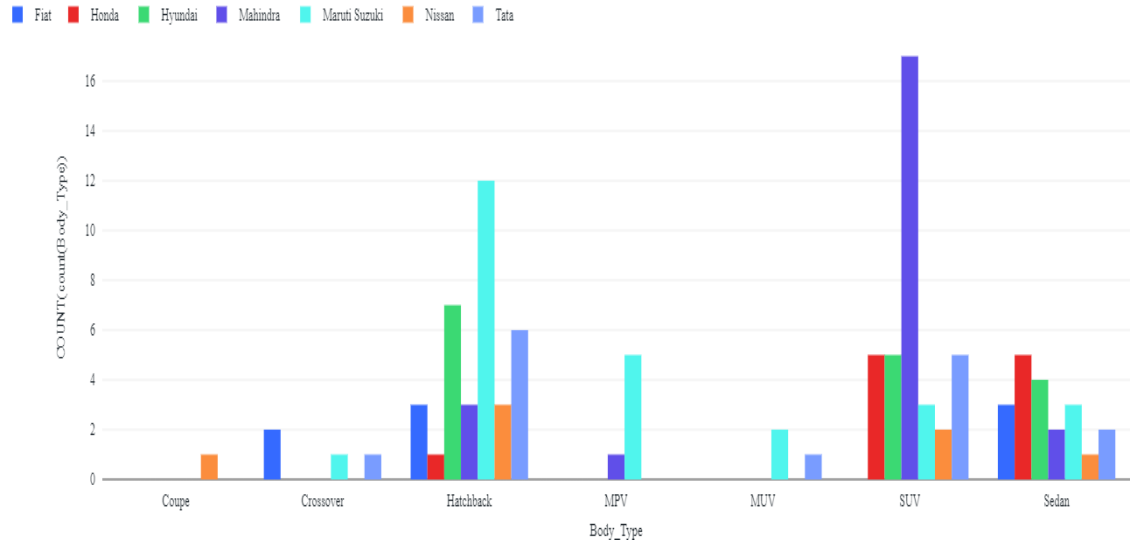
One needs to note that the inference drawn from this plot (i.e. to explore MPV segment) spells out only a rough direction for the further decision making and NOT the final decision as such. Such a preliminary inference can set the course of further customer surveys, dealer feedback and jury trials etc. The plot showcased above conserves time towards further steps by eliminating those market segments which already have been crowded by the competition

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

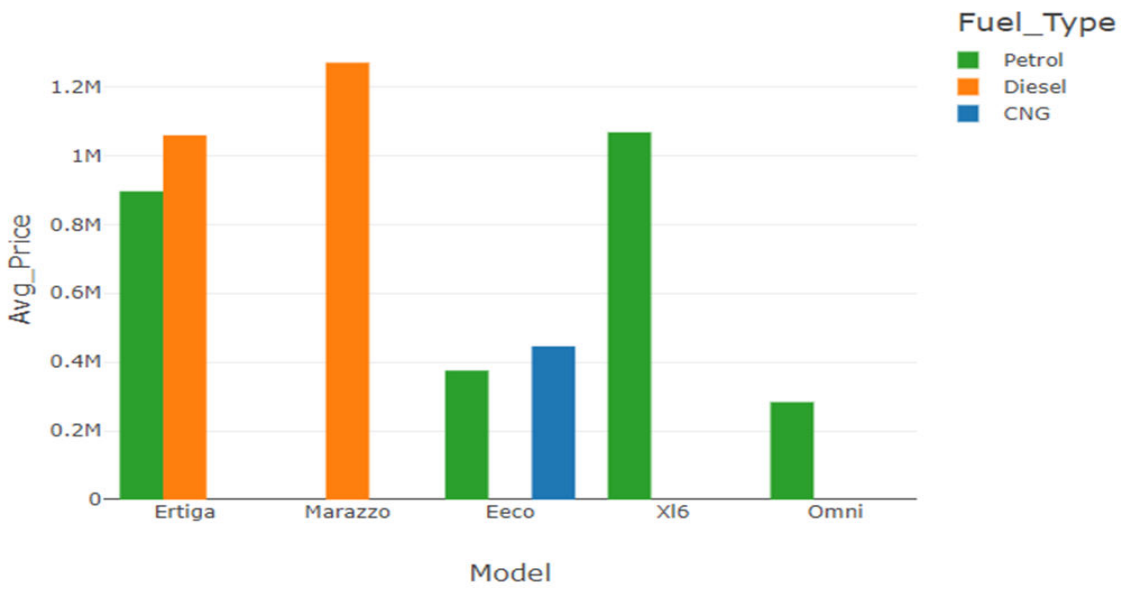
| DOI: 10.15680/IJIRCCE.2023.1106052 |

2. Most Preferred Vehicle Type for Indian Market



The next step focusses upon the most preferred vehicle type that is sought after by most of the Indian customers. This information coupled with the one obtained in the previous section will zero in on the vehicle type which is both sought after widely and yet not many players have exploited the same.
Graph has been plotted for showing the correlation between Body Type – and count of the make of vehicle for shortlisted Indian OEMs.
This analysis points to two distinct body types: Coupe and MPV as the most preferred vehicle types by the customers. However, as we analysed in the previous section, MPV also qualifies for "limited competition criterion". Hence MPV segment is recommended for exploring further analysis. It should be noted that "limited competition" may result from the factors such as formidable entry barriers etc. and hence further study/data analysis is required before finalizing the decision.

3. Fuel Type



| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

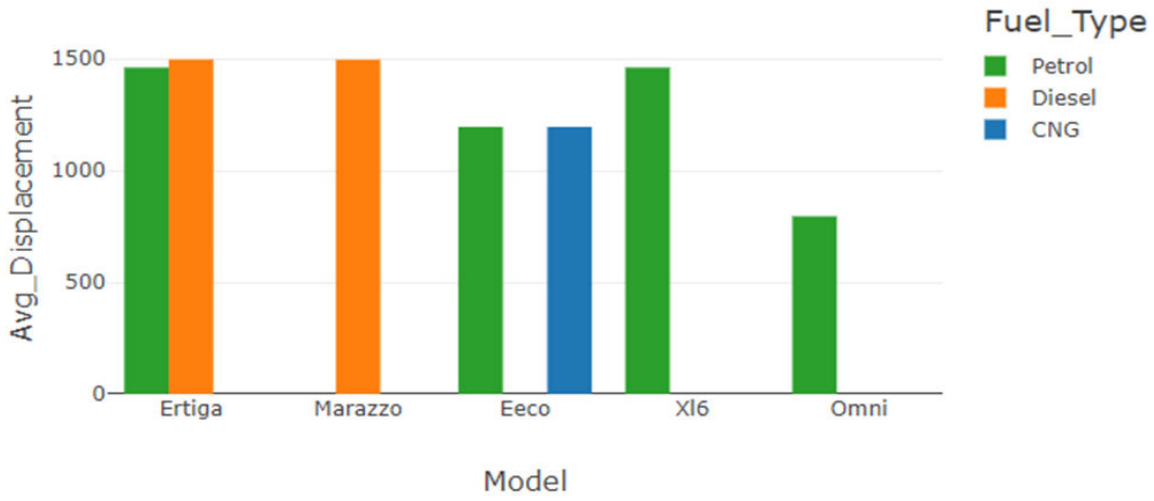
|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106052 |

From this analysis it is concluded that Fuel type - Diesel will be appropriate for vehicle to be launched. Because there is less competition as there are only two vehicles (Ertiga and Marazo) available in this MPV segment.

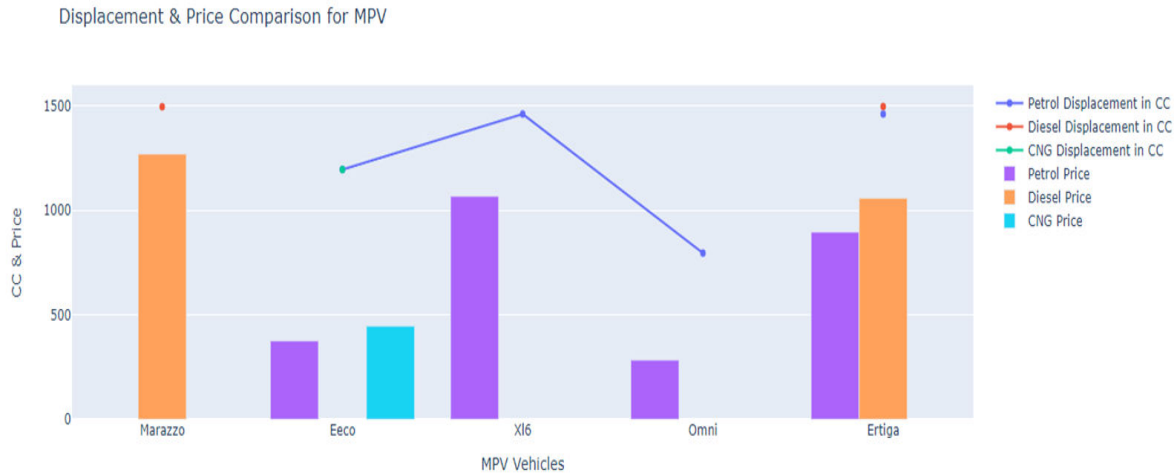
This step tries to further focus on the fuel-based vehicle subvariant. The diesel appears to be an apparent winner. However, further analysis regarding life-cycle cost other entry barriers such as maintenance infra structure, and emission road-map needs to be considered before a decision is finalized.

4. Engine Displacement



This analysis suggests that if we could decide to launch the vehicle with slightly less Engine Displacement e.g. 1250 then we will be able to launch the vehicle with slightly lower price than the existing fuel type - diesel models (Ertiga and Marazzo).

5. MPV Engine Displacement VS Price



This analysis is done to verify previous analysis and shortlisting of Engine displacement of 1250 CC and Price of INR 10 Lakhs for it.

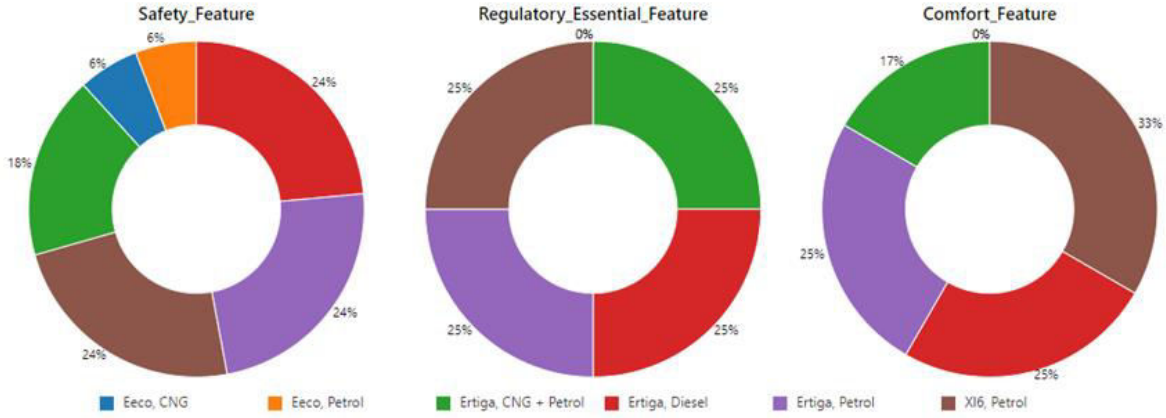
It should again be stressed that the preliminary inference supporting a lesser engine displacement engine needs to be further scrutinized based upon the repercussion on the end user experience and its impact on the saleability of the vehicle offered. For example, a limited engine displacement vehicle may not be able to offer adequate amount of payload and acceleration/speed performance needed by MPV users.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106052 |

6. Identifying the Most Preferred Features



Features are categorized into three types safety, Regulatory_Essential and Comfort features.This analysis shows that the new model should have features like Maruti Suzuki Xl6 which contains highest number of Regulatory_Essential Features, highest number of Comfort Features and highest number of Regulatory Features.
This information is useful once a preferred vehicle offering is finalized. A correct feature mix to further enhance saleability of the vehicle is required. The above plots significantly support the decision process for arriving at an optimum feature mix delivering maximal cost/benefit advantage to win the market.

IV. CONCLUSION

As highlighted at the outset, this paper only showcases a methodology to set the course for a "first time right decision making" towards arriving at the potentially most saleable vehicle for a given market. The steps followed direct us to identify areas with limited competition, vehicle types which are most preferred ones by the end customers, suggestions towards fuel type and engine displacement which are likely to attract a wider market segment as well as arriving at an optimum feature mix. As frequently stressed during the above analysis, these preliminary inference suggestions need to be further supported by customer/dealer surveys, entry barrier and emission/technology roadmap analysis etc. However, this methodology soundly supports arriving at the initial direction of the decision-making process

V. FUTURE SCOPE

With net sales of a make and model of car known for a particular year, a machine learning model can be used to recommend the make, price, fuel type, and features to be launched.

REFERENCES

[1] Sukhpreet Singh and Dr. Vijay Bhardwaj, "Implementation of Big Data Analytics in Automotive Industry", 2019 JETIR April 2019, Volume 6, Issue 4
[2] Shailendra Kumar Srivastava, Ravi Kant Sharma, Pramod Kumar Srivastava and Ruchira Srivastava, "Statistics Review of Indian Automobile Industry Using Correlation& Linear Regression Techniques ", 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)
[3] Hao Pang, Peng Liu, Shuo Wang, Zhenpo Wang and Zhaosheng Zhang, "Usage Pattern Analytics of Fuel Cell Vehicle Based on Big Data Analysis " , 2020 10th International Conference on Power and Energy Systems
[4] Matthieu Komorowski, Dominic C. Marshall, Justin D. Salciccioli and Yves Crutain. "Exploratory Data Analysis", The Author(s) 2016 MIT Critical Data, Secondary Analysis of Electronic Health Records, DOI 10.1007/978-3-319-43742-2_15
[5] Ashik Varghese and Dr. Anjana S Chandran, "EXPLORATORY DATA ANALYSIS ON AUTOMOBILE MANUFACTURING USING MACHINE LEARNING", International Research Journal of Modernization in Engineering Technology and Science
[6] Qi Zhang, Hongfei Zhan and Junhe Yua, "Car Sales Analysis Based On the Application of Big Data", International Congress of Information and Communication Technology (ICICT 2017)

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106052 |

[7] Andreas Vogelsang, "An exploratory study on improving automotive function specifications", Andreas Vogelsang Technische Universität München, Germany
[8] Chaozhong Wu" Visualization Analysis of Intelligent Vehicles Research Field Based on Mapping Knowledge Domain" May 2020
[9] Deloitte (2015) Service Delivery Trend Outlook: The potential future of government customer service delivery, The Government Summit Thought Leadership Series
[10] Gagandeep Jagdev et al., "A Comparative study of conventional data mining algorithms against Map-Reduce algorithm",International Journal of Advanced Research in Science and Engineering. 2017 May. 6 (5), pp. 325-335
[11] Hamed Taherdoost, "Different Types of Data Analysis; Data Analysis Methods and Techniques in Research Projects", International Journal of Academic Research in Management (IJARM) Vol. 9.

Appendix 1

Databricks is an analytics service based on the Apache Spark open source project. Apache Spark is a batch processing and real time processing environment. Apache Spark is quite popular among data scientists because of its ability to analyze huge amounts of data, its streaming capabilities, graph computation, machine learning, and interactive queries engine. Spark provides in-memory cluster computing.

key aspects of Databricks:

1. Unified Workspace: Databricks offers a unified workspace that brings together the various tools and services necessary for data analysis and processing. The Databricks Notebook is an interactive notebook interface that supports multiple programming languages such as Python, Scala, R, and SQL. This notebook interface allows users to write, run, and collaborate on code, visualisations, and documentation. It provides an efficient and collaborative platform for users to get their work done quickly and easily.
2. Scalability and Performance: Databricks leverages the distributed computing capabilities of Apache Spark to process large-scale datasets efficiently. It allows users to process and analyze massive amounts of data by distributing the workload across multiple nodes in a cluster. Databricks also provides built-in optimizations to accelerate data processing tasks and improve overall performance.
3. Collaboration and Sharing: Databricks enables collaboration among team members by providing shared workspaces and notebooks. Multiple users can work on the same notebooks simultaneously, making it easier to collaborate and share insights.
4. Automated Data Engineering: Databricks' automated data engineering capabilities are a powerful tool for data professionals. It provides an intuitive visual interface that makes it easy to connect to various data sources, use Spark's DataFrame API to transform data, and create pipelines to schedule data processing jobs. This eliminates the need for manual coding and allows data engineers to quickly and accurately automate tasks.
5. Machine Learning and AI Capabilities: Databricks offers built-in tools and libraries for developing and deploying machine learning models at scale. It provides a managed MLflow service for tracking experiments, managing model versions, and deploying models as web services. Databricks also integrates with popular machine learning frameworks and libraries like TensorFlow, PyTorch, and scikit-learn.
6. Data Visualization and Exploration: Databricks provides built-in data visualization capabilities for exploratory data analysis. It supports interactive visualizations and charts to help users understand and communicate insights effectively. Databricks also integrates with popular visualization libraries like Matplotlib and Plotly.

Appendix 2

1. MPV
   MPV stands for Multi-Purpose Vehicle. MPVs are sometimes called 'people carriers', too, which is perhaps a more accurate name. They have tall, box-like bodies designed to create as much interior space as possible and often have more seats than a comparable hatchback or saloon. All MPVs have at least five seats. The biggest have as many as nine, which is the maximum a car can have. MPVs with more than five seats have three rows. A seven-seat MPV has a 2-3-2 layout. An eight-seat MPV has a 2-3-3 layout. A nine-seat MPV has a 3-3-3 layout. There are a few six-seat MPVs with a 2-2-2 layout, as well. They're designed to carry larger groups of people.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 6, June 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1106052 |

2. Coupe
These are three-box configuration, fixed roof, two-door cars but are sporty and compact. Signature characteristics of coupe cars are long hood and sloping roof. These are called “coupe” based on a type of vintage carriage capable of seating two. Coupe cars are mostly high-performance cars as they are capable of stuffing big engines in their long nose. Most luxury car manufacturers opt for a coupe configuration when developing a fast, high-performance car.

3. Convertible
A convertible car body type, also known as a cabriolet or simply a "convertible," is a type of car that features a retractable roof or a convertible top. The roof can be folded down or removed, allowing occupants to enjoy open-air driving and experience the sensation of wind and sunshine while on the road.Convertible cars are often considered a symbol of freedom, style, and luxury.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details