MICROSOFT ENGAGE 2022

DATA ANALYSIS TRACK

A data analytics project based on the data from the automotive industry

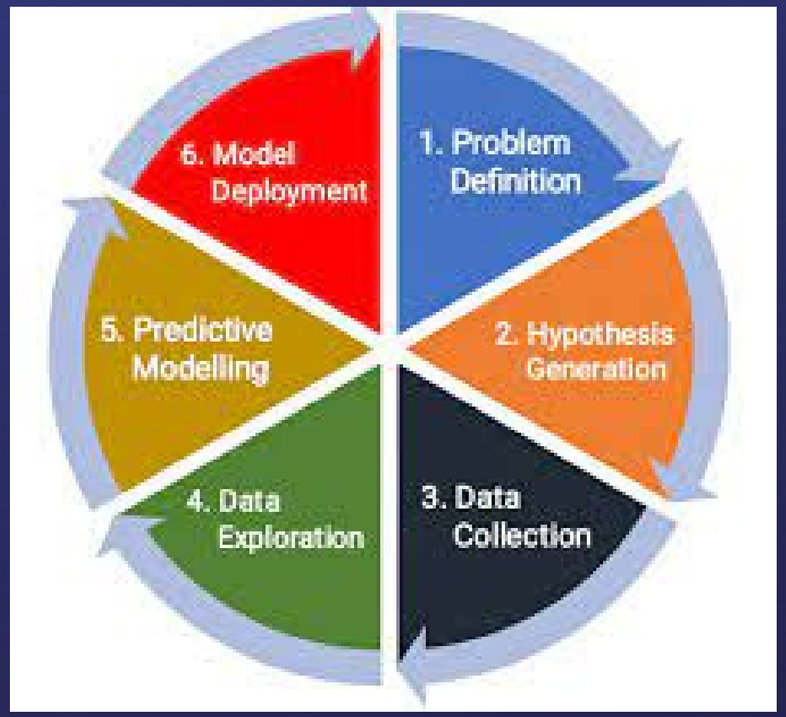
Gayatri Gyanaprava Bahali
National Institute of Technology, Rourkela
B.Tech, Civil Engineering (2024)
Mentor- Vaibhav Agarwal

INTRODUCTION

This project is solely targeted on the dataset provided to us by MS Engage which has been uploaded in the GitHub repo (cars_engage_2022.csv). However, the methodology followed is of general trend and can be effectively applied to other datasets for visualization as well as predictive analysis.

The entire project has been subdivided into majorly three sections:

- Data Preprocessing/cleaning
- Data Visualization
- Modelling and predictive analysis



TECH STACK

The project is built on a web-based interactive computing platform JUPYTER NOTEBOOK.

LIBRARIES AND PACKAGES USED (PYTHON):

- Numpy
- Pandas
- Statsmodels
- Matplotlib
- Sklearn
- Seaborn

DATA PREPROCESSING

In this process, the data is cleaned and organized so as to make it suitable for visualization and building and training a machine learning model.

DATA PREPROCESSING

- Columns having more than 50 percent of missing values have been dropped as they have insufficient data and any attempt to fill those missing values will only lead to misleading information.
- Columns having only one unique value have been dropped because they can no way contribute towards creating variations in the customer segments. (e.g.- Door_Ajar_Warning has 1192/1276 rows filled with 'Yes' while rest missing)
- Columns containing invalid or wrong data has been dropped because these can lead to misleading and skewed interpretations. (e.g.- the column of "Average_Fuel_Consumption" is filled with "Yes" which is invalid)
- Multiple columns which correlate to the same feature has been dropped as this will lead to multicollinearity. (e.g- USB_Compatibility comes under the column Audiosystem)

- Unique values in the same column which convey the same meaning has been treated so that they are not referred to as different features. (e.g.- "Front, Rear" & "Rear, Front" convey the same meaning)
- Multiple columns which depict some essential / compulsory features in modern day cars have been dropped because it is not a choice variable which customers can choose. (For e.g- Engine_Immobilizer is present in all modern day cars and hence cannot be a predictive parameter)
- Redundant parameters have been dropped as they are not that much important in predicting or creating variations. Similarly a few technical parameters which a customer never goes ahead and essentially looks for while buying a car and hence redundant considering the business perspective have been treated.
- Finally, all the missing values in numerical variables have been filled with mean of that column and those in categorical columns have been filled with the most frequently occurring value.

THE CLEANED DATA HAS BEEN EXPORTED AS 'processed_data.csv' FOR FUTHER PROCEDURES

DATA VISUALIZATION

The core part of the project which aims to reveal meaningful insights which is very useful considering the business perspective.

In-depth data analysis has been done using pie charts, bar graphs, histograms and joinplots to reveal and analyse the relationships existing between various predictive parameters. Using visual techniques, existing patterns in the customer segments and market trends directions have been identified.

Various queries have been addressed through this section, like which is the most popular brand or the specifications in major demand in the current market scenario. Along with this relationship and dependencies existing between various technical parameters have been identified which can be helpful future designing and planning.

MODELLING AND PREDICTIVE ANALYSIS

This section is devoted towards predicting prices of cars based on certain predictive parameters.

The data is further processed to be made fit to be fed into a model and for this the categorical variables have to be transformed into numerical values.

- There are many ways to do this - dummy variables, one- hot encoding, target encoding, etc. But the first two will add on to the dimensionality of the dataframe which can be especially problematic considering categorical variables with too many unique features.
- To counter this, simple variables with few unique values have been encoded with sensible numbers while avoiding ordered relationship formation. The other categorical variables have been encoded using TARGET ENCODING.
- On the basis of this LINEAR REGRESSION MODEL has been made to predict prices with appropiate training and testing which has resulted into an excellent performance of 0.86.

THANK YOU