**Project Approach**

**Mahindra First Choice**

**ML Saints**

**Team Members:**

Abhijit Katkar : <https://www.linkedin.com/in/abhijit-katkar-022b63bb>

Aarti Palande : <https://www.linkedin.com/in/aarti-palande-6aa249100/>

Manisha Upendra : [https://linkedin.com/in/manisha-rai-170082164](http://linkedin.com/in/manisha-rai-170082164)

Praveen Shekar : <https://www.linkedin.com/in/aarti-palande-6aa249100/>

**A. Exploratory Data Analysis:**

A.1. Geolocation Based Analysis:

* Statewise car services.
* Statewise revenue earned throughout the time period.
* Plant Locations

A.2. Market Segmentation -

* Revenue earned against each type of car models.
* Time Series for Revenue earned
* Time Series for Monthly growth
* Time Series for active customers each month
* Time Series for Monthly visits
* New customer to Old customer ratio
* Time Series graph depicting New customers and old customers in the data
* Average Efficiency of Plants
* Statewise customers from each state
* Gender Ratio
* Count of Origination of data
* Type of partner
* Car brands used by Retail type partner
* Orgination of data for different partner types
* Count of car models arriving at the shop
* Ratio of Luxury to Economy class car
* Time series for car models throughout the time period

A.3. Efficiency of Plants:

The following points were extracted for calculating Efficiency -

* + Service Time (Hrs)
  + Count of services
  + Total Revenue
* Subtracting the features ‘Invoice datetime’ and ‘Jobcard’ from the data and returning values in hours using pd.datetime() function.
* Grouping the datasets on Customer no. and getting the unique count of the services amd the sum all the Total Amt Wtd Tax. Feature.
* Efficiency = [Total Revenue / (Total service time in hrs. \* Count of invoices) ] \* 100

**B. Data Cleaning and Merging**

* **Invoice Data Cleaning**
  + Dropped Columns: -
* Amt Rcvd From Custom, Amt Rcvd From Ins Co,
* CGST (14%), CGST (2.5%), CGST (6%), IGST (12%), CGST (9%),
* IGST (18%), IGST (28%), IGST (5%), SGST/UGST (14%),
* SGST/UGST (2.5%), SGST/UGST (6%), SGST/UGST (9%),
* TDS amount, Service Advisor Name, Outstanding Amt,
* Total CGST, Total GST, Total IGST, Total SGST/UGST,
* Policy no., Cash /Cashless Type, Expiry Date, Gate Pass Date,
* Insurance Company, Claim No., ODN No, Technician Name',
* Regn No',Total Value

* **Customer Data Cleaning**
  + Dropped columns: -
    - Title
    - Marital Status
    - Occupation
    - Date of birth
    - Death date
* **Jtd Data Cleaning**
  + Dropped columns: -
    - Labor Value Number
* **Data Merging**
  + Joined Customer & Invoice Data on Customer No.
  + Grouped the Jtd data on dbm order to reduce duplicate order entries
  + Left Joined the invoice with grouped Jtd data based on job card no & dbm order.
  + Final shape of merged dataframe (merged\_data)
  + Shape of joined dataframe (936275, 33)

**C. Feature Cleaning**

Following columns were cleaned using pgeocode library of python using pin codes given in Invoice data.

* **City**
  + Using pin code correct city was extracted
* **District**
  + Using pin code correct district was extracted
* **Area/Locality**
  + Using pin code correct location area was extracted

**D. Feature Engineering**

* **Invoice Date Time**
  + Using Invoice date & invoice time **Invoice\_DateTime** was created
* **Job Card Date Time**
  + Using Jobcard date & Jobcard time **Jobcard\_DateTime** was created
* **Service Time**
  + By Subtracting invoice date time with Jobcard date time
  + Invoice date time – Jobcard date time

**E. Customer Segmentation -**

After all cleaning & data pre-processing tasks, A new dataframe was created with the help of cleaned merged dataframe (merged\_data.csv).

Grouping the merged data on Customer No, Cust Type, Make, Model, City, Order Type calculated average servicing time & average revenue & Count of invoices.

**Approach**

* Separating customers with zero spends
  + Separated customers having total revenue <= 0
  + 26327 rows were filtered as zero spenders
* Separating customers with nonzero spends
  + Separated customers having total revenue > 0
  + 284656 rows were filtered as non-zero spenders
* Rule Based Clustering on non-zero spenders using Average revenue
  + 4 clusters were identified based on quantile-based analysis on Average revenue.
  + Low Revenue, Medium Revenue, Average Revenue, High Revenue
* Rule Based Clustering on non-zero spenders using Average servicing time
  + 4 clusters were identified based on quantile-based analysis on Average servicing time in Hrs.
  + Super-Fast, Fast, Super Slow, Slow
* K-means Clustering on non-zero spenders
  + Label Encoded below columns
    - Make, Model, City, Cust Type, Order Type
  + Scaled all the features
    - Using standard scaler scaled all features
  + Used Elbow method to determine optimal no of clusters
    - We tried fitting range of 2 to 12 clusters on the scaled data.
    - Using elbow method, we found 6 clusters as optimal clusters
  + Run K-means clustering algorithm
    - 6 segments were generated by applying K-means algorithm with 6 clusters on scaled data.

**Customer Lifetime Value Prediction**

**Customer Lifetime Value**

Customer lifetime value is a metric that indicates the total revenue a business can reasonably expect from a single customer account. It considers a customer's revenue value, and compares that number to the company's predicted customer lifespan. Businesses use this metric to identify significant customer segments that are the most valuable to the company.

**Customer Lifetime Value Calculation**

CLTV = Customer frequency \* Average spend

**Approach For Classification**

* Created a Separate Data Frame for cars over a period of year 2016
* On above Data Frame performed the group by on “Make”, “Model” “Location/Area” and “Customer No.” and perform the count on “Invoice No” (For the frequency of the particular Customer No.) , mean on “Service Time” (For average Service Time), mean on “Total Value” (For average purchase value).
* Multiply Count of Invoice No of individual customer to Average value of the respective customer to get the Customer Lifetime Value.
* Binned the Customer Lifetime value as “Low”, ‘Medium” and “High” to convert it into categorical format.
* Label Encode the Categorical columns.
* Drop columns "Customer No.","LTV (for Running Repairs)","Invoice No" as these columns are not required for Machine Learning Model building.
* Build a Machine Learning model to predict the classification of the customer in to categories of Customer Lifetime as “Low”, “Medium” and “High”. Use different Machine Learning and different features to try and improve the accuracy of the model.