**Team Name: Aysmptote**

**Team Members:**

**Neel Kudu:** [**https://www.linkedin.com/in/neel-kudu-566219150/**](https://www.linkedin.com/in/neel-kudu-566219150/)

**Gaurav Mishra:** [**https://www.linkedin.com/in/gaurav-a-mishra/**](https://www.linkedin.com/in/gaurav-a-mishra/)

**Ajaykumar Panchal:** [**https://www.linkedin.com/in/ajaykumar-panchal-a9733389/**](https://www.linkedin.com/in/ajaykumar-panchal-a9733389/)

**Tilak Bire:** [**https://www.linkedin.com/in/tilak-bire-2450a438/**](https://www.linkedin.com/in/tilak-bire-2450a438/)

**Approach to for Mahindra First Choice Capstone Project**

**Exploratory Data Analysis for Capstone Project**

Productivity of a Plant:

1. Calculate Number of unique Technicians present at the particular Plant over a year
2. Calculate the total amount of service time in hours over a year
3. Calculate the total amount of revenue generated at the respective plant over a year
4. Productivity = Total Revenue(in year)/Total Service Time (in year) \* Number of Technicians worked (in year)

Business of a Plant:

    Total number of invoices generated by the plant per year or month.

**Classification of Customer based on Customer Lifetime value**

* Loading “Final Invoice” file
* Shape of the Invoice Dataframe (492314, 59)
* Checking for missing values in each column
* Dropping the column containing more than 60% missing values. Columns dropped in the process are 'Cash /Cashless Type',  'Claim No.', 'Expiry Date', 'Gate Pass Date', 'Insurance Company', 'Policy no.', 'Service Advisor Name'
* New shape of the Invoice Dataframe (492314, 52)
* Checking for unique entries in each 52 columns
* Dropping columns 'Amt Rcvd From Custom', 'Amt Rcvd From Ins Co', 'CGST(14%)', 'CGST(2.5%)', 'CGST(6%)', 'CGST(9%)', 'IGST(12%)', 'IGST(18%)', 'IGST(28%)', 'IGST(5%)', 'SGST/UGST(14%)', 'SGST/UGST(2.5%)', 'SGST/UGST(6%)', 'SGST/UGST(9%)', 'TDS amount', 'Total CGST', 'Total GST', 'Total IGST', 'Total SGST/UGST', 'Total Value' as all these columns have unique value as either “0” or “NaN”
* New shape of the Invoice Dataframe (492314, 32)
* Convert "Invoice Date" and "Invoice Time" to datetime format, similarly for "JobCard Time" and "JobCard Date"
* Drop columns "Invoice Date","Invoice Time","JobCard Date","JobCard Time" as they are already combined together for datetime format in the previous step.
* Calculate “Service\_Time” by subtracting “JobCard\_Datime” from “Invoice\_Datetime”
* Drop columns "Labour Total","Misc Total","OSL Total","Parts Total","Total Amt Wtd Tax." as these columns are already included in “Total Value” as Total Value is the summation of values of all of these columns.
* New shape of Invoice Dataframe (492314, 26)
* Cleaning “Area/Locality”, “City” and “District” using “pgeocode” library of Python with the help of “Pin Code” column of the DataFrame
* Dropping old “Area/Locality”, “City” and “District” columns as new clean columns are created for respective columns.
* Calculation of LifeTime Value for each customer over a period of 2013-2015 (3 years) based on “Make”, “Model” and “Location/Area”

What is 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.**

How to calculate Customer Life Time Value?

**To calculate customer lifetime value you need to calculate average purchase value, and then multiply that number by the average purchase frequency rate to determine customer value. Then, once you calculate average customer lifespan, you can multiply that by customer value to determine customer lifetime value.**

* Make a seperate DataFrame for the “Running Type” cars over a period 2013-2015
* On above DataFrame perform 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 for Modeling.
* 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.

**Clustering/Segmentation of different Customers based on different features.**

* Take a cleaned Invoice DataFrame from previous task with shape (492314, 26)
* Drop columns "Regn No","Technician Name" as they are not required for Customer Segmentation purpose
* New Invoice DataFrame shape (492314, 24)
* Load the Customer Data
* Drop columns 'Marital Status', 'Occupation', 'Date of Birth', 'Death date', 'Business Partner' as these columns have 90%+ missing values
* Fill the missing values “Partner Type”, “Data Origin” and “Title” with mode value of the respective columns.
* Remove trailing zeros from Customer No. of the Customer DataFrame
* Merge Customer DataFrame with Invoice DataFrame based on “Customer No.” as a common column.
* New invoice\_customer with shape (492312, 28)
* Load JTD data
* Drop column "Labor Value Number" as these columns are not required in Customer Segmentation/Clustering
* Group by JTD Data based on “DBM Order” and “Order Type” taking average of “Net Value” and “Order Quantity”
* Merge two DataFrames invoice\_customer and jtd based on the column “Job Card No” from invoice\_customer and “DBM Order” from jtd DataFrame
* The new DataFrame is formed by combining all DataFrames called invoice\_customer\_jtd of shape (523161,30)
* Group by invoice\_customer\_jtd on columns "Customer No.","Make","Model","Clean\_Location\_Area","Cust Type" and perform averaging on “Total Value”, “Service\_Time\_Hours”
* Label Encoding the Categorical Columns for Model building
* Drop “Customer No.” as it’s a unique identification number for customers not required in the clustering.
* Standard scaling all input features as K-Means Clustering is Distance based algorithm.
* Right number of Clusters are found out using the Elbow Method and Silhouette Score.