**DATA SCIENCE PROJECT REPORT ON**

**SALES EFFECTIVENESS**

**PROJECT TEAM ID:** PTID-CDS-AUG-23-1607B

**PROJECT ID:** PM-PR-0019

**INTRODUCTION OF PROJECT**

FicZon Inc, a prominent IT solution provider offering a spectrum of products from on-premises solutions to SAAS-based offerings, relies heavily on its digital channels, especially the website, for lead generation. However, the company faces a pivotal challenge in sustaining sales effectiveness amidst a maturing market landscape and increased market entrants, leading to a noticeable downturn in sales performance. The pivotal factor influencing effective sales lies in the quality of leads, predominantly determined by manual categorization heavily reliant on the sales team's judgment.

While FicZon maintains a quality process for ongoing lead categorization, its primary value lies post-analysis, offering insights rather than immediate conversation strategies. Recognizing the critical role of lead quality in sales efficacy, FicZon seeks to delve into Machine Learning methodologies. The objective is to proactively categorize lead quality using predictive models, aiming to significantly augment sales effectiveness by transitioning from subjective, human-dependent categorization to data-driven, predictive lead quality assessment.

**ABSTRACT:**

This project revolves around the integration of Machine Learning into FicZon Inc's lead management system to pre-categorize lead quality and elevate sales performance. Currently facing a decline in sales due to market maturation and intensified competition, FicZon aims to alleviate this challenge by harnessing ML's predictive capabilities. The objective is to transition from manual lead categorization, heavily reliant on sales staff, to a proactive ML-driven approach that anticipates lead quality. By pre-determining lead quality, FicZon anticipates a considerable boost in sales effectiveness, optimizing resource allocation and refining marketing strategies for improved outcomes.

Methodology involves data preprocessing, exploratory data analysis (EDA), feature engineering, and the application of machine learning algorithms Through this methodology, FicZon aims to revolutionize its lead management process, transitioning from subjective manual categorization to a data-driven ML approach. This transformation is anticipated to drive substantial improvements in sales effectiveness, facilitating more informed decision-making and propelling FicZon towards sustained growth and success.

**DEVICE PROJECT IN-TO MULTIPLE STEPS:**

1. Data Collection
2. 2.Loading data
3. 3.Domain Analysis
4. 4.Basic Checks of data
5. EDA (Univariate Data Analysis)
6. Data Pre-processing
7. Feature Selection
8. Building ML Model
9. Training & Model Evaluation
10. Model Saving

**DATA COLLECTION:**

The dataset received from the Datamites educational platform and provided in a spreadsheet format, specifically as a PDF File.In PDF file contains the necessary data related to project, including DataBase Detail and relevant specifications.

**LOADING DATA:**

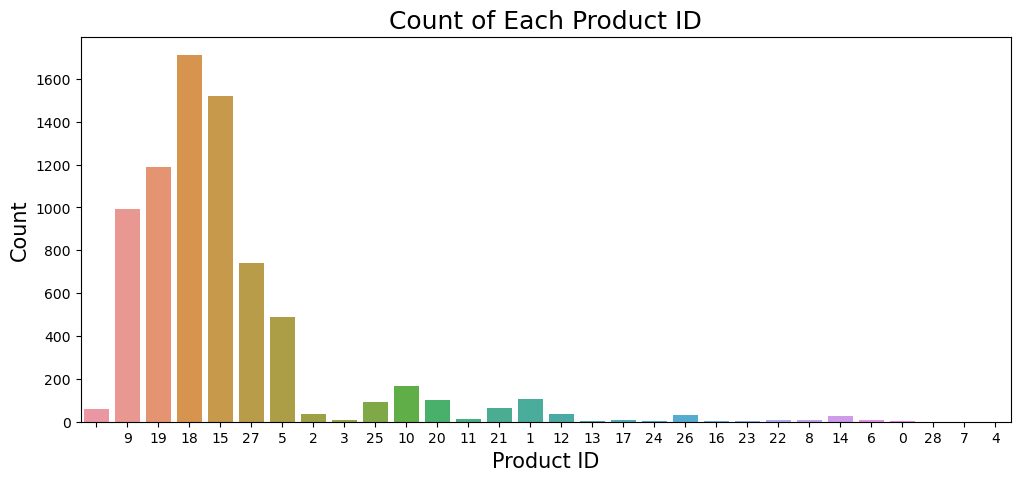
load data in python using panda’s library

**DOMAIN ANALYSIS:**

* Understanding the meaning of each feature and grasping the significance of every attribute in defining sales effectiveness is essential. Status is the target Variable Represents the current status of the transaction or order.

**EDA (UNIVARIATE, BIVARIATE, MULTIVARIATE DATA ANALYSIS)**

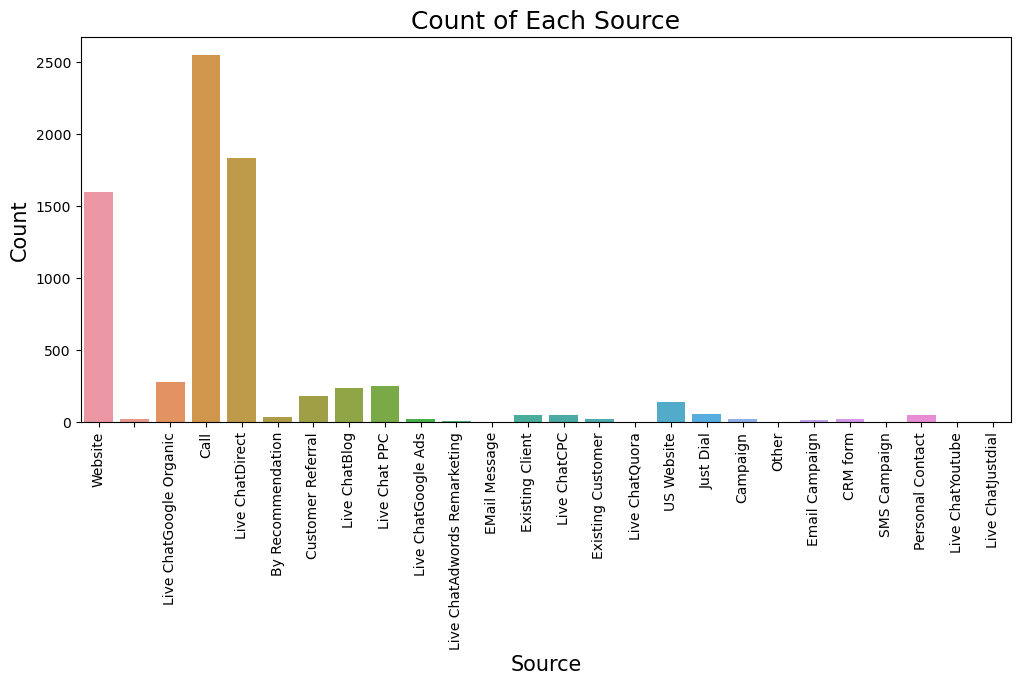
1. **UNIVARIATE DATA ANALYSIS**

**1. PRODUCT\_ID**

#### **OBSERVATION/INSIGHTS:-**

* In the sales dataset, product IDs 9, 19, 18, 15, 27, and 5 have recorded the highest sales volumes, while the remaining product IDs have lower sales figures by comparison.

#### **2. SOURCE**



#### **OBSERVATION/INSIGHTS:-**

* The sales dataset trends where the sources—website, call, live chat direct exhibit significantly higher sales figures compared to the remaining sources which register specifically lower sales.

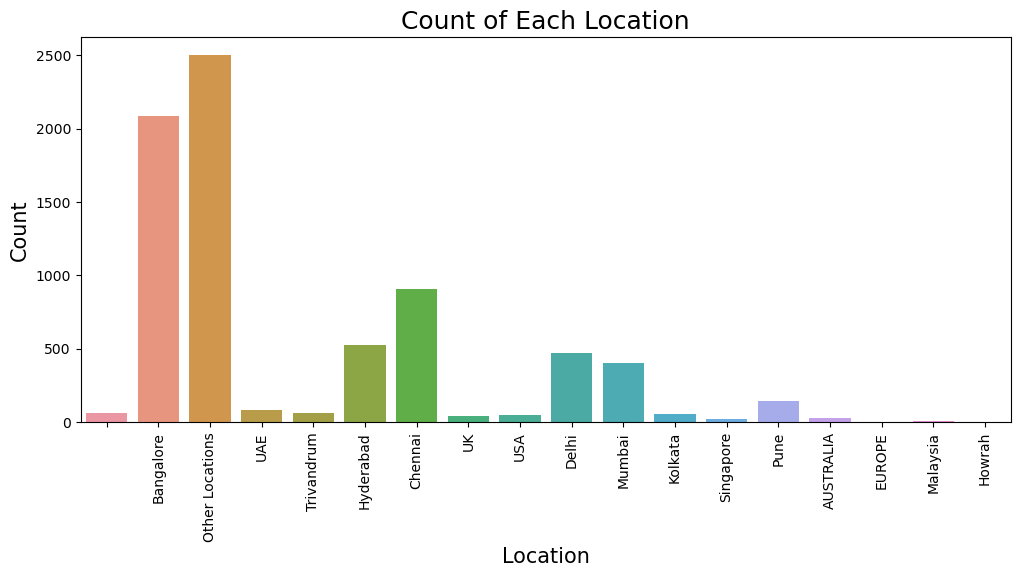
#### **3. SALES\_AGENT**

### 

#### **OBSERVATION/INSIGHTS:-**[**¶**](http://localhost:8890/notebooks/Downloads/Internship%20Projects/Client%20Project/PM-PR-0019/Sales-Effectiveness.ipynb#OBSERVATION/INSIGHTS:-)

* In the provided sales dataset the sales agents identified as sales\_agent11, sales\_agent3, sales\_agent4, sales\_agent9, sales\_agent7, and sales\_agent5 stand out for determine specifically high sales performance as compared to other sales agents.

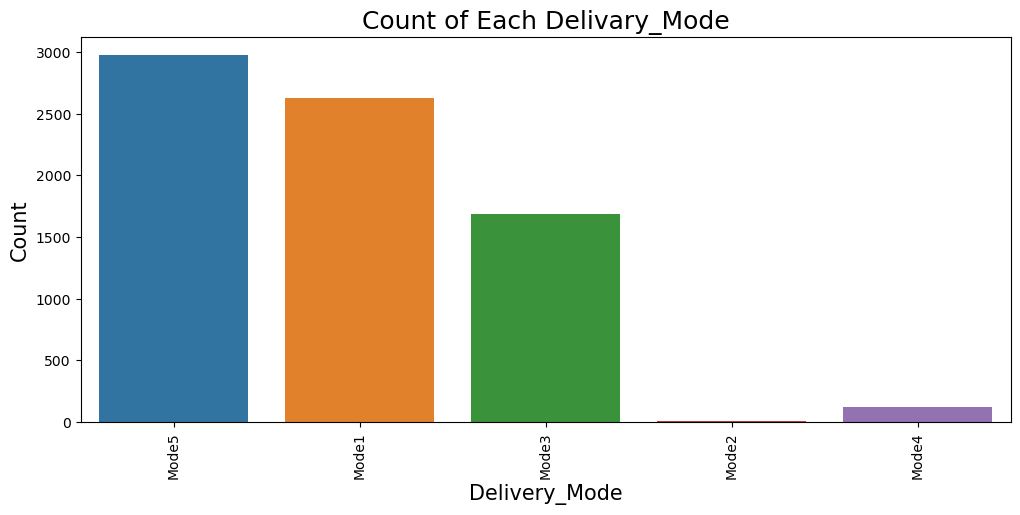
#### **4. LOCATION**



#### **OBSERVATION/INSIGHTS:-**[**¶**](http://localhost:8890/notebooks/Downloads/Internship%20Projects/Client%20Project/PM-PR-0019/Sales-Effectiveness.ipynb#OBSERVATION/INSIGHTS:-)

* In the sales dataset the sales distribution along with locations shows that Bangalore and other unspecified locations collectively exhibit higher sales compared to Hyderabad, Chennai, Delhi, and Mumbai.
* Hyderabad, Chennai, Delhi, and Mumbai indeed exhibit considerable sales figures. when directly compared to locations with very high sales, such as Bangalore and other unspecified locations.
* Additionally the remaining locations outside of these major cities indicate considerably lower sales in the dataset.

**5. DELIVERY\_MODE**



#### **OBSERVATION/INSIGHTS:-**[**¶**](http://localhost:8890/notebooks/Downloads/Internship%20Projects/Client%20Project/PM-PR-0019/Sales-Effectiveness.ipynb#OBSERVATION/INSIGHTS:-)

* In the sales dataset the sales associated with delivery modes specifically mode5, mode1, and mode3, stand out for determine higher sales figures in comparison to mode2 and mode4.

**6. STATUS**

### 

#### **OBSERVATION/INSIGHTS:-**[**¶**](http://localhost:8890/notebooks/Downloads/Internship%20Projects/Client%20Project/PM-PR-0019/Sales-Effectiveness.ipynb#OBSERVATION/INSIGHTS:-)

* In the sales dataset, the status labeled as "junk lead" appears to have the highest sales figures.
* Additionally, statuses such as potential, in progress positive, not responding, just enquiry, CONVERTED, in progress negative, and long term also high considerable sales compared to statuses such as open and converted.

**Note**:-Due to the absence of numerical features and also lots of categories are present in categorical feature, bivariate and multivariate analysis are not feasible.

**DATA PREPROCESSING**

* First, we check the missing values, In this data, there is few feature have missing values present.
* Product-ID, Source, Mobile, Sales\_Agent, Location have missing values present in these features.
* Handle Missing values in categorical data can be handled by imputing them with the mode, which represents the most frequently occurring value within the respective categorical feature.
* Second, features is contain lots of different label so we compresses and merged the label
* I Handle categorical data and use frequency encoding. Because features have contained lots of label.

**FEATURE ENGINEERING**

* First we drop unique and constant feature there are three unique feature such as Created, Mobile, EMAIL.
* Before check correlation changing data type after that Check the correlation with the help of heatmap and seen the from the above heatmap I analyzing the heatmap, it appears that there are no highly correlated features present. Therefore, we do not drop any feature based on correlation.
* I haven't addressed duplicates as I've compressed and merged labels, ensuring the dataset doesn't contain any duplicates due to label transformations
* After that save & load the preprocess data, Save the dataframe to a CSV file.

**MODEL CREATION & EVALUATION**

* Define Independent and dependant variable and split the data into training and testing.
* For Training 80% & Testing 20% data
* I check data is balanced or not after check the data has imbalanced so I utilized the SMOTE technique to balance the data.
* Use classification algorithm algorithms including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, KNN, Bagging and Artificial Neural Network (ANN).
* Afterward, I evaluated the models using accuracy, recall, precision, and F1 score metrics.

**BEST MODEL**

* Our best performed model with accuracy (0.72) metric is XG Boost. This classifier could achieve accuracy rate 0.72 that is average accuracy.

**MODEL SAVING**

* Save the model using pickle file.

**CONCLUSION**

* In a project focused on improving sales by identifying high-quality leads, various machine learning models were tested using a dataset. The goal was to find the most effective model based on different performance measures like Accuracy, Recall, Precision, and F1 scores.
* After evaluating these models, it was determined that the XG Boosting outperformed the others by achieving the highest accuracy score. This suggests that, compared to the other models tested, XG Boosting was most successful in accurately predicting lead quality, making it the top performer for this particular task.

----------------------------------Thank You -----------------------------------