Lead Categorization for FicZon Inc.

Project Title:

Lead Categorization Using Machine Learning to Improve Sales Effectiveness at FicZon Inc.

Business Case Overview:

FicZon Inc, an IT solution provider with a product portfolio ranging from on-premise to SaaS-based offerings, primarily generates leads through digital channels and their website. As the competition grows and market saturation increases, FicZon is experiencing a decline in sales.
The effectiveness of the sales team is strongly dependent on the quality of incoming leads. Currently, lead categorization is done manually and heavily relies on the sales staff, which introduces bias and delays. Existing quality control processes are reactive, offering limited value for real-time sales conversions.

Objective:

FicZon aims to implement a Machine Learning (ML) model that can predict the lead quality (High Potential or Low Potential) at the time of lead generation to improve conversion rates and overall sales effectiveness.

Project Goals:

1. Perform data exploration and analysis to understand patterns in sales effectiveness.
2. Build a classification model to predict lead quality:
    - Categories: High Potential, Low Potential.

Dataset Summary:

Total Records: 7422
Features : 9 Columns

Data Collection:

Data Source:

The data was collected from the company's internal SQL database using SQL Workbench. The relevant table was queried and exported to a .csv file for local processing and machine learning workflows.

Steps Followed:

1. Accessed SQL Workbench using the following connection parameters:
    - Database Name: project_sales
    - Table Name : data
    - Host: 18.136.157.135
    - Port: 3306
    - Username: dm_team2
    - Password: DM!$Team&27@9!20!

2. Executed SQL Query to extract relevant data: < SELECT * FROM your_table_name; >
3. Exported the Result as a .CSV file from SQL Workbench.
4. Imported the CSV into Python using pandas.

Initial Observations:

- Total records: 7422
- Columns: Created, Product_ID, Source, Mobile, EMAIL, Sales_Agent, Location, Delivery_Mode, Status
- Missing values:

  i. Product_ID, Location → have some missing (NaN)
  ii. EMAIL → contains invalid entries like #VALUE!
  iii. Mobile → some data masked or invalid (e.g., XXXXXXX)
  iv. Created → datetime column

- Data Types:
  i. Created should be converted to datetime for time-based analysis
  ii. Most other fields are categorical or string

Domain Analysis:

Business Context:

FicZon Inc's sales success is highly dependent on the quality of leads. Currently, this lead classification (e.g., Potential, Open, Not Responding) is done manually. Delayed insights impact conversion rates.

Objective:

Use machine learning to pre-categorize leads as:

- High Potential
- Low Potential

This will:

- Improve sales agent productivity
- Enable early prioritization of quality leads
- Automate the lead scoring pipeline

Relevant Domain Insights:

- Channels (Source like Website, Call, Live Chat) indicate engagement intent
- Sales Agent and Location may reflect geographic or agent-specific performance trends
- Status is our target variable
  - Will be reclassified into High vs. Low Potential
  - Mapping Example :

    High Potential: CONVERTED, converted, Potential, In Progress Positive, Long Term

    Low Potential: Junk Lead, Not Responding, Just Enquiry, In Progress Negative, LOST, Open

Problem Definition:

Type of Problem;

- Machine Learning Task: Supervised Learning
- Problem Type: Classification
- Target Variable: Status

Objective:

To build a binary classification model to predict lead quality as either:

- High Potential: Likely to convert or progress positively.
- Low Potential: Unlikely to convert (junk, not responding, etc.

Final Target Variable :

- 1 : High Potential
- 0 : Low Potential

Goal:

To predict whether a new lead falls into the High Potential or Low Potential category based on features like:

- Source of lead
- Sales Agent
- Location
- Delivery Mode
- Contact information (possibly anonymized or dropped)

Exploratory Data Analysis:

1. **Checked Data Structure :**Reviewed shape, data types, and null values.
2. **Target Variable Analysis :**Analyzed class distribution of Lead_Potential (High vs. Low).
3. **Categorical Feature Insights**:

    - Explored Delivery_Mode vs. Lead_Potential using countplots.
    - Found that Mode-1 had balanced leads, while Mode-5 mostly had Low Potential leads.

4. **Time Series Analysis:**

    - Analyzed Source and Created_Year vs. Lead_Potential.
    - Identified high-performing sources in specific years (e.g., Calls, Website).

5. **Identified Patterns :** Detected possible class imbalance and source/channel effectiveness.

Feature Engineering:

- Drop/Impute nulls (Mobile, Email,etc)
- Encode categorical features (LabelEncoder)
- Extract features from Created (e.g., Month, Weekday)
- Drop identifiers like EMAIL, Mobile if uninformative
- Scaling (MinMaxScaler for numerical features)

Splitting Data:

- Use train_test_split (e.g., 80% training, 20% testing)

Machine Learning Models Applied:

- **Linear Regression**
- **Decision Tree (base + tuned)**
- **Random Forest (base + tuned)**
- **K-Nearest Neighbors (KNN)**
- **Support Vector Machine (SVM)**
- **Artificial Neural Network (ANN using MLPClassifier)**
- **Gradient Boosting**
- **XGBoost**

Model Performance :

We experimented with multiple classification models to predict lead potential. Here's the overview:

1. **Logistic Regression:** Moderate performance with equal training and testing accuracy (66%).
2. **Decision Tree:** High overfitting observed (Training: 91%, Testing: 69%).
3. **Tuned Decision Tree:** Reduced overfitting but still moderate generalization (Training: 77%, Testing: 67%).
4. **Random Forest**: High training accuracy but slight overfitting (91% vs 71%).
5. **Tuned Random Forest:** Improved generalization compared to untuned version (83% vs 71%).
6. **KNN:** Moderate scores with slight overfitting (81% vs 67%).
7. **ANN (MLP):** Performed fairly with a good balance (71% vs 68%).
8. **SVC:** Consistent but lower performance (69% vs 67%).
9. **Gradient Boosting:** Chosen Model — slightly lower training accuracy (76%) but best generalization with highest testing accuracy (73%) and smallest gap.
10. **XGBoosting**: Good performance (85% vs 71%) but higher overfitting than Gradient Boosting.

Final Model: Gradient Boosting:

We selected Gradient Boosting as the final model due to:

- Best generalization ability.
- Highest testing accuracy (73%).
- Smallest train-test accuracy gap (just 3%).

Conclusion:

Through extensive data exploration, visualization, and machine learning modeling, we identified key patterns and built a predictive system to classify leads as High Potential or Low Potential. Among multiple models tested, Gradient Boosting emerged as the most effective, balancing accuracy and generalization (Training: 76%, Testing: 73%).

This model enables sales teams to focus on high-quality leads, improving conversion rates and optimizing effort allocation.

Recommendations to Improve Sales Effectiveness:

1. **Prioritize High-Potential Leads:** Use the model's predictions to assign experienced sales agents to high-potential leads to maximize conversions.

2. **Optimize Underperforming Sources:** Focus marketing efforts on top-performing sources like Live Chat, Calls, and Website, while re-evaluating or improving campaigns and referrals that generate mostly low-potential leads.

3. **Agent Performance Monitoring:** Identify top-performing agents (like Sales_Agent4 and Sales_Agent11) and replicate their strategies across the team.

4. **Geo-Focus Strategy:** Concentrate resources on high-performing locations such as Bangalore, Other Locations, and Trivandrum to leverage regional strengths.

5. **Time-Based Campaigning:** Plan campaigns during months like May to July, when you usually get more leads.

6. **Improve Delivery Modes:** Promote effective delivery modes like Mode_1 and Mode_5 that are most associated with successful leads.

7. **Target Lead Statuses Smartly:** Focus nurturing efforts on statuses like 'Potential', 'Just Enquiry', and 'In Progress Positive' to push them toward conversion.

8. **Feedback Loop:** Regularly retrain the model with updated data to adapt to changing market dynamics and customer behavior.

9. **Sales Dashboard:** Deploy a dashboard for real-time monitoring of lead quality, agent performance, and conversion trends.