Lead Scoring Case Study

Vishnu Kulathunkal
Tahir Khan
Sai

X Education System Problem Statement

30% Current Lead Conversion Rate

Objective Sales Team Struggling to priotrize the genunie leads and missed opportunities

Goal To build and train a model to properly identify “Hot Lead” and increase the conversion rates to near 80%

Goals For Analysis

- Develop a Lead Scoring Model
- Business Insights: Key driving factors
- Optimize Sales Efforts for Hot Leads

Data Overview

Data Set

- Total Records:- Approx 9000
- Key Features:- Lead Source,Tags,Total time spent on website, Occupation.
- Target Variable:- Converted

Challenges

- Missing Values in Key columns.
- Sparse Categories like “select”

Data Cleaning & Preprocessing

- Handled Missing Values : Columns with >40% missing data were Dropped
- Encoded Binary Variables to 0/1
- Combined Sparse Categories into others to reduce noise
- Addressed Multicollinearity using VIF

Feature Selection

Chi-Square Test

Identified Statistically significant Categorical features based on their assoiciation with target

RFE

selected Top 20 Features based on their predictive power.

Final Features

Selected the feature set which was given by RFE.

Model Building

Algorithm Used:

Logistic Regression.
Chosen for its interpretability and ability to output probabilities that can be converted into lead scores.

Training and Testing Split:

Dataset split into 70% training and 30% testing to evaluate performance on unseen data.
Used stratified splitting to maintain the class balance in training and testing datasets.

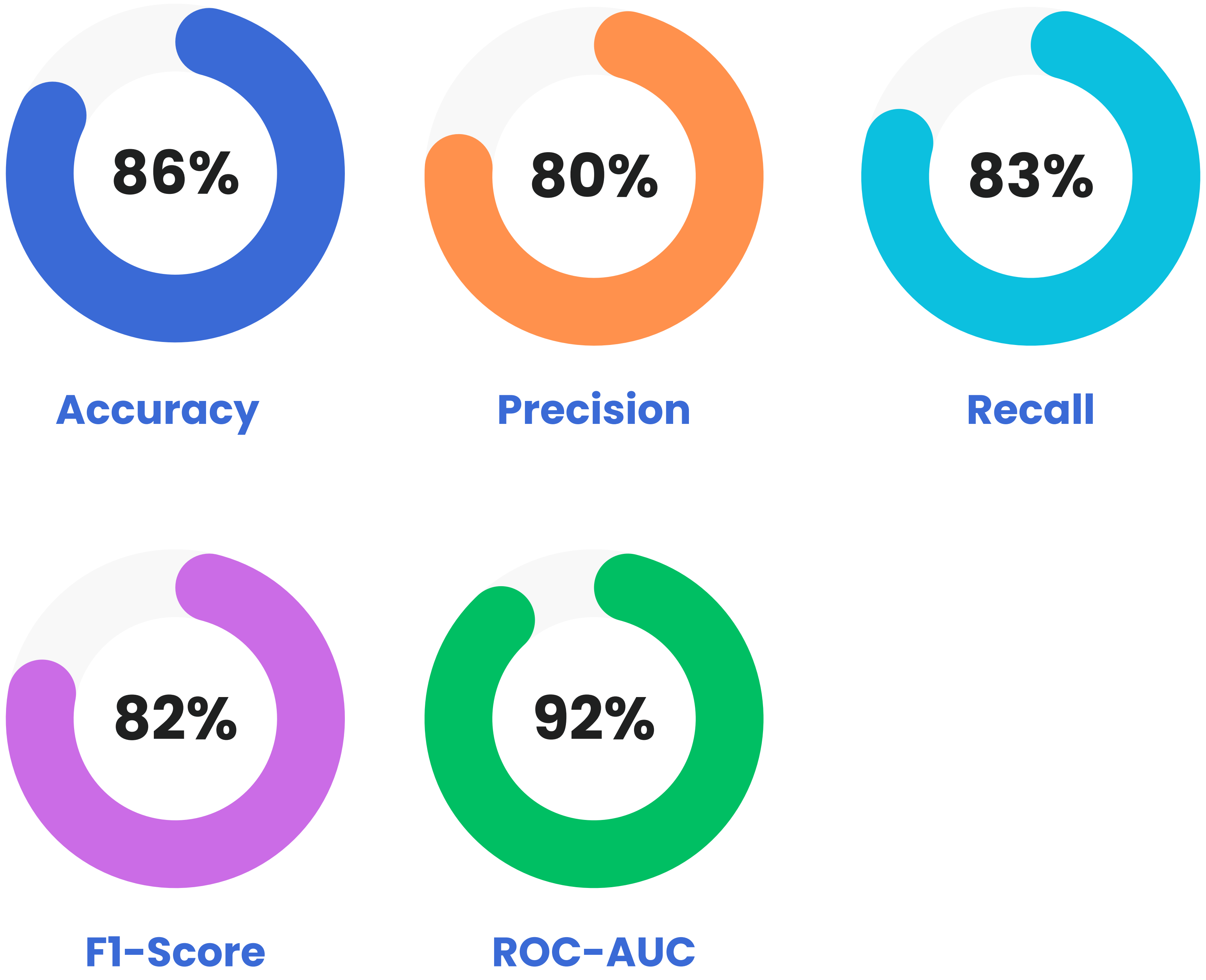
Feature Engineering:

Standardized numerical features (Total Time Spent on Website, Total Visits) for better model performance.
Included only statistically significant features identified (RFE).

Threshold Optimization:

The model's default threshold of 0.5 was tuned to optimize recall and precision
Final Threshold Selected: 0.41 (balances recall and precision).

Metrics



Accuracy: 92%
Measures overall correctness of predictions.

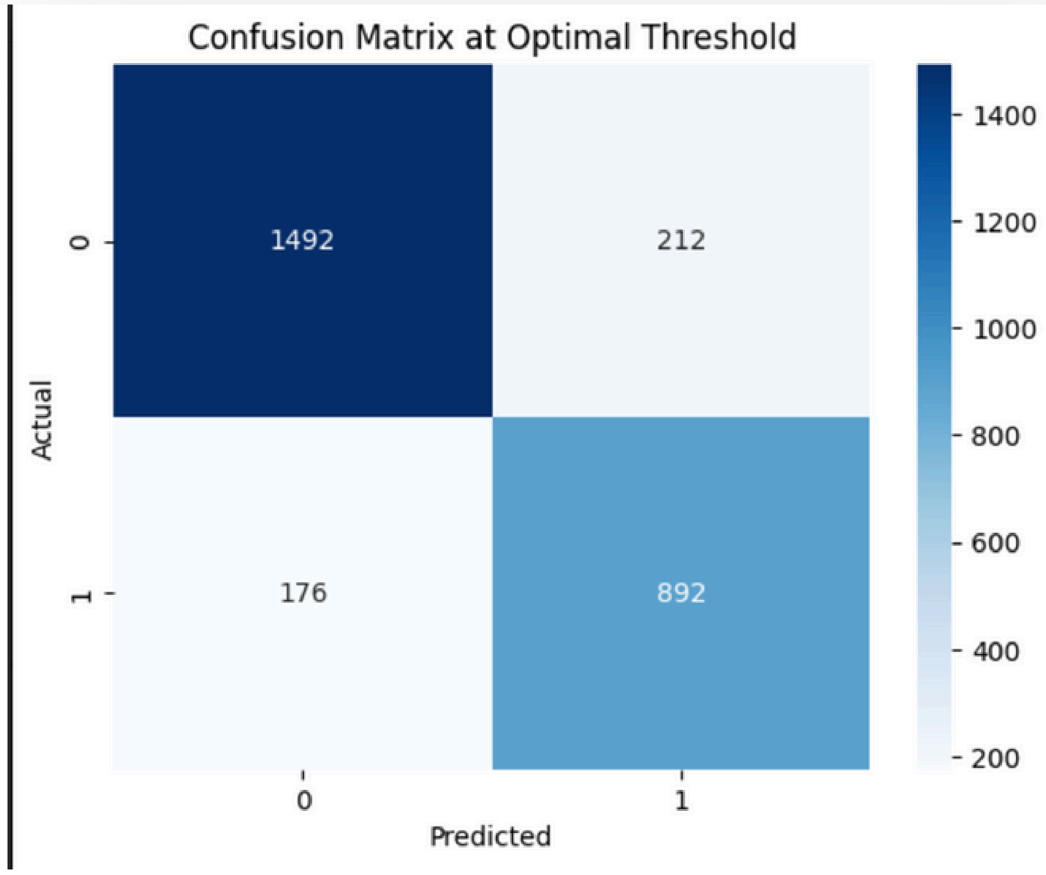
Precision: 88%
Measures the proportion of predicted converters (1s) that are actually correct.

Recall: 87%
Measures the ability to identify all actual converters.

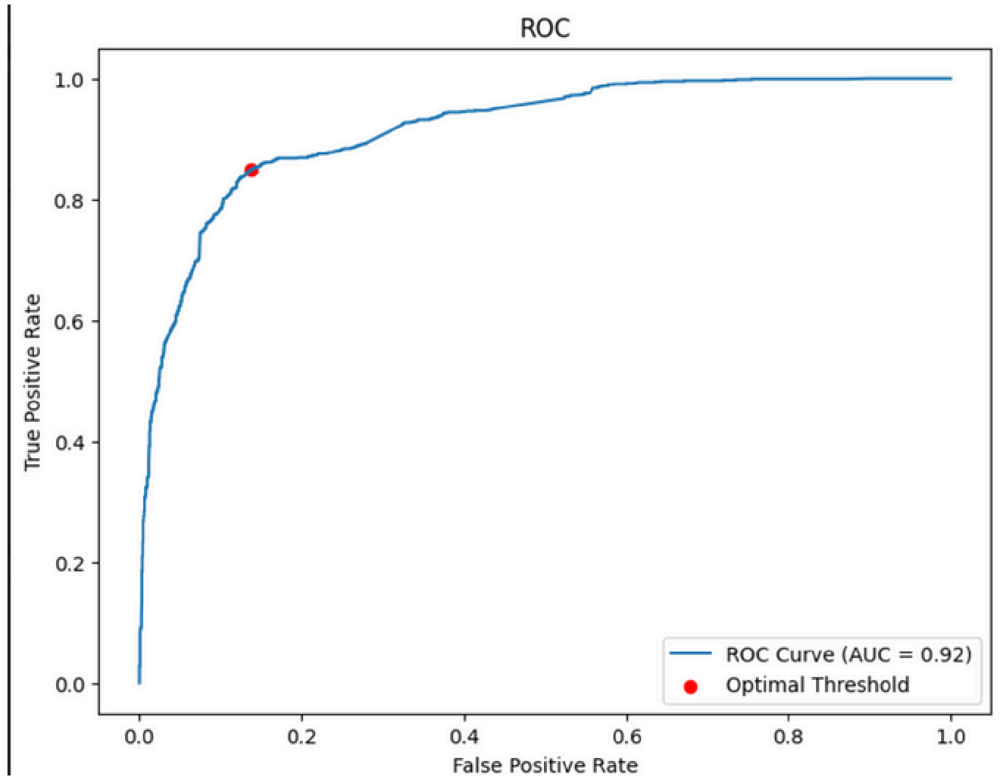
F1-Score: 87.5%
Harmonic mean of precision and recall, balancing both.

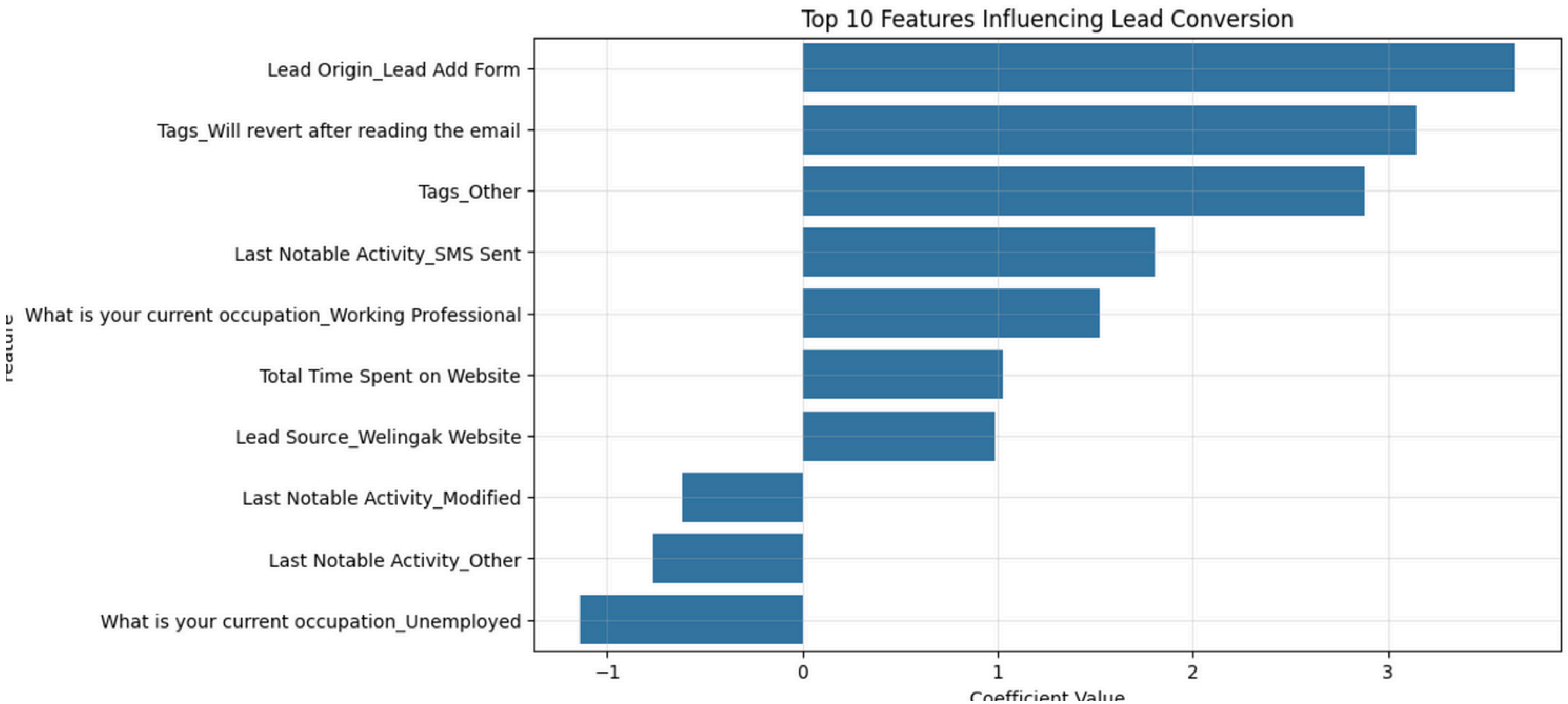
ROC-AUC: 0.92
Indicates excellent discrimination ability between converters and non-converters.

- **Confusion Matrix:**
  - True Positives: Successfully identified converters.
  - True Negatives: Non-converters correctly classified.
  - False Positives: Non-converters incorrectly identified as converters.
  - False Negatives: Missed opportunities (actual converters classified as non-converters).



- **Roc - Curve**
  - At this point, the model achieves a high true positive rate while keeping the false positive rate relatively low.
  - The selection aligns with business goals by ensuring a good balance between identifying true converters and minimizing unnecessary effort on false positives.
  - A high AUC value (0.92) confirms that the model is highly reliable and performs well in identifying potential leads effectively.





Top Positive Features

- Lead Origin_Lead Add Form: Strongest driver of conversion.
- Tags_Will revert after reading the email: High engagement leads.
- Total Time Spent on Website: Indicates strong interest.

Top Negative Features:

- Do Not Email: Leads opting out of emails convert less.
- Tags_Ringing: Indicates poor response.
- What is your current occupation_Unemployed: Lower conversion likelihood.

Business Impact

Increased Conversion

Projected Increase from 30% to 80 % for priortized leads

Efficient Resource Allocation

Sales Team focuses only on high potential Leads

Driver-Indication

Able to see top and bottom drivers that affects conversion

Conclusion

The lead-scoring model is a data-driven approach to improving sales efficiency.

Ensures flexibility for different business scenarios.

Results show strong potential for increased conversions and optimized resource use.

Thank You