

Lead Scoring Model - Logistic Regression

Submitted by: Tanisha Grover, Tushar Rajput, Vansh Tondon

Project Overview

This project aims to build a **Logistic Regression-based Lead Scoring Model** to predict the likelihood of lead conversion. The model helps businesses identify and prioritize high-potential leads, improving marketing efficiency.

Data Overview

- **Dataset Size:** 9,240 records, 37 features
- **Target Variable:** **Converted** (1 = Converted, 0 = Not Converted)
- **Feature Types:**
 - **Categorical** (e.g., **Lead Source**, **Last Activity**, **Tags**)
 - **Numerical** (e.g., **TotalVisits**, **Total Time Spent on Website**)

Preprocessing & Feature Selection

- **Handling Missing Data:** Imputation/removal of missing values based on data distribution.
- **Categorical Encoding:** One-hot encoding for categorical variables.
- **Feature Selection:** Used **Recursive Feature Elimination (RFE)** to select the most relevant variables.

Model Building

Generalized Linear Model (GLM) - Logistic Regression

- Used **Logit Link Function** for binary classification.
- Evaluated features based on **p-values** and **Variance Inflation Factor (VIF)**:
 - Dropped features with **high p-values** (insignificant predictors).
 - Ensured **VIF < 5** to prevent multicollinearity.

Final Model Features:

- **Numerical:** **TotalVisits**, **Total Time Spent on Website**, **Page Views Per Visit**
- **Categorical (dummy variables):**
 - **City_Other Cities of Maharashtra**

- What is your current occupation_Working Professional
- Last Activity_Email Opened
- Tags_Other Tags, Tags_Ringing, Tags_Will revert after reading the email

Model Evaluation & Performance Metrics

Train Dataset

Metric	Value
Accuracy	96.65%
Sensitivity (Recall)	98.4%
Specificity	92.6%
Precision	96.85%
False Positive Rate (FPR)	7.4%
Negative Predictive Value (NPV)	96.15%

Confusion Matrix Analysis

- True Positives (TP): 246
- True Negatives (TN): 100
- False Positives (FP): 8
- False Negatives (FN): 4

Test Dataset

Metric	Value
Accuracy	94.97%
Sensitivity (Recall)	94.8%
Specificity	95.37%
Precision	97.93%
False Positive Rate (FPR)	4.6%
Negative Predictive Value (NPV)	88.79%

Confusion Matrix Analysis

- **True Positives (TP): 237**
- **True Negatives (TN): 103**
- **False Positives (FP): 5**
- **False Negatives (FN): 13**

Insights:

- **High Sensitivity (Train: 98.4%, Test: 94.8%):** The model captures converted leads accurately.
- **Good Precision (Train: 96.85%, Test: 97.93%):** Predicts lead conversion with minimal false positives.
- **Low False Positive Rate (Train: 7.4%, Test: 4.6%):** Ensures marketing efforts are not wasted on unlikely conversions.

Model Balance & Graphical Analysis

- **ROC Curve & AUC Score:** High AUC confirms strong model discrimination.
- **Precision-Recall Curve:** Validates the effectiveness of predictions.

Conclusion

- The **Logistic Regression model** provides a **robust and balanced approach** to lead scoring.
- **High accuracy and low error rates** make it **suitable for real-world deployment**.
- Businesses can **prioritize high-quality leads**, increasing efficiency in lead conversion strategies.