Lead Conversion Prediction for X Education

Data Analysis and Predictive Modeling to Optimize Lead Management

Ditipriya Dutta

Objective of the Analysis

- **Business Goal:**

  Improve lead conversion rates by identifying "hot leads" and optimizing sales efforts.
  Focus sales efforts on the most promising leads to optimize resources.
  Minimize unnecessary calls during periods of low activity.

- **Technical Goal:**

  Build a predictive model to classify leads into "hot" (likely to convert) and "cold" (unlikely to convert).
  Provide interpretable results for actionable insights using logistic regression.

Dataset Overview

- Size: 9,240 rows and 37 attributes.
- Target Variable: Converted (1 = Converted, 0 = Not Converted).
- Key Features :-
  - Demographics (e.g., Country, City).
  - Behavioral metrics (e.g., Total Time Spent on Website, Last Activity).
  - Behavioral metrics (e.g., Total Time Spent on Website, Last Activity).
  - Lead source (e.g., Google, Reference).

Data Cleaning and Preprocessing

- Dropped columns with >30% missing data (e.g., Lead Quality).
- Imputed missing values using the mode (categorical) or median (numerical).
- Treated "Select" values in categorical variables as missing data and imputed accordingly.
- Capped numerical features like TotalVisits and Page Views Per Visit at the 95th percentile to ensure stability.
- Converted categorical variables into dummy variables for analysis.
- Converted categorical variables into dummy variables for analysis.

Exploratory Data Analysis (EDA)

- **Target Distribution**:

  38.5% of leads converted, indicating class imbalance.

- **Behavioral Insights**:

  Leads spending more time on the website have a higher likelihood of conversion.

  Tags like "Will revert after reading the email" and "Lost to EINS" correlate positively with conversion.

- **Categorical Insights**:

  Sources like "Reference" and "Welingak Website" have higher conversion rates than channels like "Olark Chat."

- **Visualization Examples:**

- Target distribution plot.

  Conversion rates by lead source and tags.

Feature Selection

Key Features for the Model -

- **Methods:**

  Correlation Analysis: Identified numerical features with a strong correlation to the target variable.

  Logistic Regression Coefficients: Ranked dummy variables by predictive importance.

- **Top Features:**

  Tags_Lost to EINS (+6.64): Indicates a renewed focus on lost leads.

  Tags_Closed by Horizzon (+5.51): Represents engaged leads.

  Tags_Will revert after reading the email (+2.55): Reflects explicit interest.

  Total Time Spent on Website (+0.36): Strong engagement indicator.

- **Business Insight:**

  Behavioral and engagement-driven features are the strongest predictors.

Predictive Model Development

- **Model Used:**

  Logistic Regression

- **Why Logistic Regression?**

  Provides interpretative coefficients for actionable insights.

  Balances simplicity with strong performance in binary classification tasks.

- **Steps Taken:**

  Data split into 70% training and 30% testing sets.

  Standardized numerical features for consistent scaling.

  No hyperparameter tuning required—default settings provided excellent results.

- **Output:**

  No hyperparameter tuning required—default settings provided excellent results.

Model Performance Metrics

- **Results:**

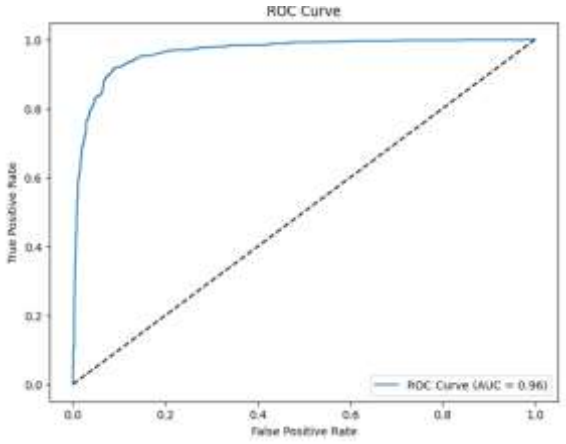
  Accuracy: 90.8%

  Precision: 89.2%

  Recall: 86.7%

  F1 Score: 87.9%

  ROC-AUC: 96.4%



- **Key Visual:**

  ROC Curve demonstrating excellent discriminatory power.

- **Business Impact:**

  High accuracy ensures reliable identification of "hot leads."

  Balanced precision and recall minimize missed opportunities and wasted efforts.

Business Applications of the Model

- **Aggressive Lead Conversion:**

  Lower threshold (e.g., 0.3) during high-resource periods to maximize coverage.

- **Efficient Resource Allocation:**

  Raise threshold (e.g., 0.8) to focus on high-confidence leads during low-activity periods.

- **Strategic Insights:**

  Use tags and behavioral features to guide targeted sales strategies.

  Prioritize high-performing channels like "Reference" and "Welingak Website."

Key Learnings and Takeaways

- **Data Cleaning:**

  Addressing missing values and outliers significantly improved model performance.

- **Behavioral Features Are Key:**

  Tags like "Lost to EINS" and "Will revert after reading the email" were critical predictors.

- **Threshold Adjustment:**

  Dynamic thresholds align sales strategies with business priorities.

- **Model Insights Drive Action:**

  Logistic regression provided actionable, interpretable results.

	Feature	Coefficient
9	Tags_Lost to EINS	6.381391
7	Tags_Closed by Horizon	5.396103
2	Tags_Will revert after reading the email	2.666118
4	Lead Origin_Lead Add Form	2.043807
11	Lead Source_Welingak Website	1.824333
1	Last Notable Activity_SMS Sent	1.408312
16	Tags_Busy	1.228531
3	Last Activity_SMS Sent	0.949238
5	What is your current occupation_Working Profes...	0.819383
13	Asymmetrique Activity Index_02.Medium	0.729897

	Lead Score	Converted
8305	32.46	0
1591	78.45	1
8604	2.48	0
1333	6.99	0
4260	0.11	0
2357	98.41	1
1900	74.82	1
9077	1.35	0
6302	97.16	1
8158	91.00	1