Lead Scoring Case Study

Summary Report

Contents

Business Understanding

X Education faces a challenge with a 30% lead conversion rate. To address this issue, the company is considers implementing a supervised classification model trained on historical data stored in the `"leads.csv"` dataset. The primary objective is to identify "Hot Leads" exhibiting a conversion likelihood exceeding 80%. By leveraging machine learning insights, this approach enables a more targeted strategy for the sales team, allowing them to concentrate efforts on leads with a higher probability of converting into paying customers. This data-driven solution aims to streamline the lead conversion process, improving overall efficiency and potentially boosting the company's revenue.

Logistic Regression – Theory

Logistic regression, a supervised learning algorithm for binary classification, predicts the likelihood of a record being labelled 0 or 1 using the logistic (sigmoid) function. Weights for important features are determined by minimizing the cost-function. The model passes the weighted sum of chosen features into the logistic function to obtain likelihood values. A cut-off is set based on business objectives, utilizing sensitivity-specificity or precision-recall trade-offs. Mathematically, it is represented as follows:

$$P(y = 1) = \frac{1}{1 + \mathrm{e}^{-(b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n)}}$$

Here,

$P(y = 1)$ is the probability that an observation belongs to Class 1;

$b_0, b_1, b_2, \ldots, b_n$ are coefficients obtained by minimizing the cost function;

$X_1, X_2, \ldots, X_n$ are the chosen input features for the model.

Data Analysis

Data Cleaning

- Dropped columns with over 45% null values to enhance data modelling.
- Dropped categorical columns with over 45% "Select" values and converted the rest of the “Select” values to "Unknown" category.
- Identified data imbalance, but opted not to address it because the final model effectively identifies the minority class.
- Removed categorical columns with low variability to prevent overfitting.

Visualizations

Visualizations of categorical and numerical columns were created with and without the target variable to identify patterns. Univariate categorical plots revealed that despite numerous categories, most had low number of records.

We also identified trends indicating higher conversion rates for specific categories and numerical variables. This understanding is used to check the validity of the model coefficients in the end.

Data Pre-processing

One-hot encoding is used to create dummy variables from categorical variables. There were no other data quality issues that needed addressing at this step. Data was prepared for modelling by completing the train-test split and scaling the data using min-max scaler.

Modelling – Logistic Regression

Feature Selection

Identifying the most important features for the final model was done in two steps:

1. Coarse selection of features using Recursive Feature Elimination (RFE) algorithm (automated selection of 15 features from 107 features)
2. Refined manual selection of features to get 11 features for the final model.
    a. Using p-value which gives the significance level of a coefficient ($\alpha = 0.05$ chosen)
    b. Using Variance Inflation Factor to identify and avoid multicollinearity ($VIF \leq 5$ tolerable)

Final Model

Finally chosen factors are:

- 'Total Time'
- 'Source_Welingak Website'
- 'Last Activity_Email Bounced'
- 'Last Activity_SMS Sent'
- 'Current Occupation_Unknown'
- 'Tags_Closed by Horizzon'
- 'Tags_Lost to EINS'
- 'Tags_Ringing'
- 'Tags_Will revert after reading the email'
- 'Tags_switched off'
- 'Last Notable Activity_Modified'

Following feature selection, we retained 11 key features for the logistic regression model, achieving an impressive ROC curve area of 0.97. Using Sensitivity-Specificity and Precision-Recall thresholds, we established a probability cut-off for classifying leads as "hot" or not.

Model Evaluation

Using the final model, we assigned Lead Scores ranging from 0 to 100 to evaluate the conversion likelihood for each record. The classification metrics computed are summarized in the table below:

Metric	Train Data	Test Data
Accuracy	0.9167	0.9108
Sensitivity	0.9086	0.9161
Specificity	0.9215	0.9073
Precision	0.8736	0.8632
Recall	0.9086	0.9161
F1 – Score	0.8908	0.8889

The model's performance on training and test data was assessed through key classification metrics, all surpassing 90%. While precision slightly dipped but remained above 85%, noteworthy was the model's superior performance on test data compared to training data, particularly in sensitivity/recall. This suggests the model accurately identifies underlying trends, showcasing its effectiveness in predicting lead conversion.

Conclusion

Identified Trends:

High Sensitivity/Recall: The model consistently achieved over 90% sensitivity/recall, showcasing its robust ability to identify leads with a high conversion likelihood.

Key Feature Selection: Features like 'Total Time,' 'Last Activity_SMS Sent,' and specific 'Tags' categories offer actionable insights, enhancing the interpretability of the model.

Lead Score Utility: The generated lead scores provide a practical tool for the sales team to prioritize leads, optimizing resource allocation for more efficient conversion efforts.

Recommendation:

Utilize High-Performing Features: Leverage insights from key features like 'Total Time' for informed sales strategies.

Implement Lead Scoring: Actively integrate lead scores to prioritize high-conversion potential leads, streamlining efforts.

Regular Model Monitoring and Updates: Establish a system for periodic model updates to adapt to evolving lead behaviour trends for sustained effectiveness.