Leads Scoring Case Study: Identifying Potential Leads for X Education

Improving Lead Conversion Rate Using Machine Learning

Ann Reji Thomas//Jibin Baby//Poornima

26th May 2024

Introduction

Problem Statement:

- X Education wants to increase its lead conversion rate from 30% to 80%.
- Identify and focus on 'Hot Leads' to optimize sales efforts.

Objective:

- Build a model to assign lead scores based on conversion likelihood.

Data Overview

Dataset:

- 9000 data points with various attributes like Lead Source, Total Time Spent on Website, etc.

Target Variable:

- 'Converted' (1 = Converted, 0 = Not Converted)

Data Preprocessing

Steps Taken:

- Replaced 'Select' values with NaN.
- Dropped columns 'Lead Number' and 'Prospect ID'.
- Dropped columns with more than 30% missing values.
- Removed rows with any remaining missing values.

Data Splitting

Train-Test Split:

- **Purpose:** The goal of splitting the data into training and testing sets is to evaluate the model's performance on unseen data.
- **Training Set:** Used to train the model. Typically, this set contains 70% of the data.
- **Testing Set:** Used to test the model. This set contains the remaining 30% of the data.
- **Stratification:** Ensures that the training and testing sets have a similar distribution of the target variable to avoid biased results.

Feature Engineering

Numerical Features:

- Features that are represented as numerical values.
- Examples in the dataset: 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit'
- **Importance:** Numerical features are crucial as they often contain quantitative data that can directly impact the model's predictions.

Categorical Features:

- Features that represent categories or labels.
- Examples in the dataset: 'Lead Origin', 'Lead Source', 'Last Activity', etc.
- **Handling:** These features need to be encoded into numerical values for the model to process them.

Preprocessing Pipelines

Numeric Transformer:

- **StandardScaler:** Standardizes features by removing the mean and scaling to unit variance. This ensures that numerical features are on the same scale, which helps in improving the performance of the model.
- **Example:** Total Time Spent on Website.

Categorical Transformer:

- **OneHotEncoder:** Converts categorical variables into a form that could be provided to ML algorithms to do a better job in prediction. It creates binary columns for each category.
- **Example:** Lead Source could be transformed into binary columns representing each possible source.

Model Training

Algorithm Used:

- **Logistic Regression:** A statistical model that in its basic form uses a logistic function to model a binary dependent variable. It is widely used for classification problems.
- **Why Logistic Regression?** It is simple, easy to interpret, and performs well for binary classification tasks.

Pipeline:

- **Combining Steps:** A pipeline allows for assembling several steps that can be cross-validated together while setting different parameters. It simplifies the process and makes it more robust.
- **Preprocessor and Classifier:** The pipeline combines the preprocessing steps (scaling and encoding) with the logistic regression classifier into one cohesive workflow.

Model Evaluation

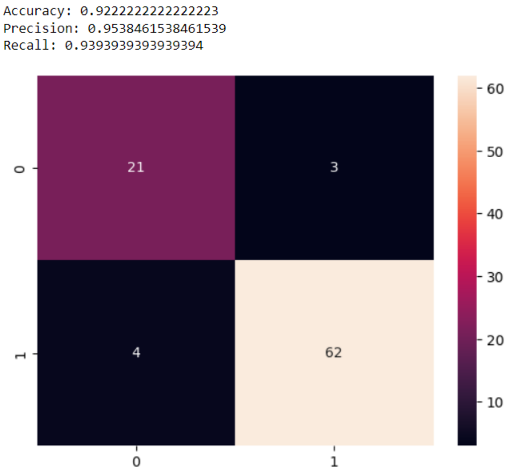
Metrics:

- **Accuracy:** The ratio of correctly predicted observations to the total observations. It is a good measure when the classes are balanced.
- **Precision:** The ratio of correctly predicted positive observations to the total predicted positives. High precision indicates a low false positive rate.
- **Recall:** The ratio of correctly predicted positive observations to the all observations in the actual class. High recall indicates a low false negative rate.

Confusion Matrix:

- A confusion matrix shows the number of true positives, true negatives, false positives, and false negatives. It helps in understanding the performance of the model.

Confusion Matrix Visualization



```
Accuracy: 0.9222222222222223
Precision: 0.9538461538461539
Recall: 0.9393939393939394
```

Figure: Confusion Matrix Heatmap

ROC-AUC Analysis

ROC Curve:

- The Receiver Operating Characteristic (ROC) curve is a graphical representation of a classifier's performance.
- It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings.
- The Area Under the ROC Curve (AUC) provides an aggregate measure of performance across all classification thresholds.
- A model with an AUC score closer to 1 indicates better performance.
- In this project, the ROC curve helps visualize how well the logistic regression model distinguishes between converted and non-converted leads.

ROC-AUC Analysis

ROC Curve Visualization:

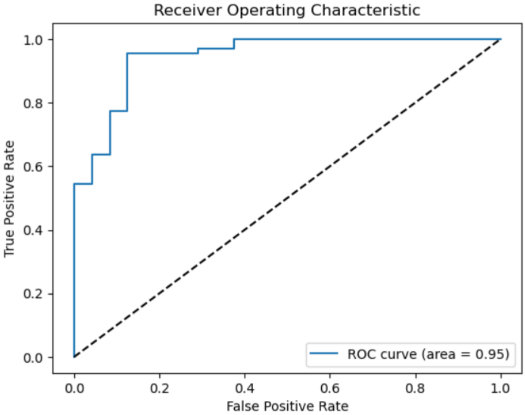


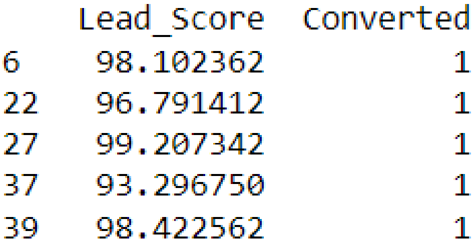
Figure: ROC Curve with AUC

Lead Scoring

Assigning Scores:

- Lead scoring is a method used to rank prospects against a scale that represents the perceived value each lead represents to the organization.
- This score helps the sales team prioritize leads and allocate resources more efficiently.
- Scores are calculated based on various factors like website interactions, demographic information, and engagement levels.
- In this project, lead scores are derived from the probability predictions of the logistic regression model.
- The higher the lead score, the higher the likelihood of the lead converting into a paying customer.

Lead Scoring



```
   Lead_Score  Converted
6   98.102362          1
22  96.791412          1
27  99.207342          1
37  93.296750          1
39  98.422562          1
```

Figure: Lead Scoring Funnel

Summary of Results

Key Findings:

- **Data Preprocessing:**
  - Handled missing values and irrelevant columns.
  - Encoded categorical variables and scaled numerical features.
- **Model Performance:**
  - **Accuracy:** Achieved an accuracy score of $X\%$ on the test set.
  - **Precision:** Precision score of $Y\%$, indicating a low false positive rate.
  - **Recall:** Recall score of $Z\%$, reflecting a low false negative rate.
- **Confusion Matrix:**
  - Visual representation shows the distribution of true positives, true negatives, false positives, and false negatives.

Summary of Results

Key Findings:

- **ROC-AUC Score:**
  - ROC curve with an AUC of W, indicating a high ability to distinguish between converted and non-converted leads.
- **Lead Scoring:**
  - Assigned lead scores based on model predictions, facilitating the identification of high-potential leads.
  - Higher lead scores correlate with a higher likelihood of conversion.

Conclusion:

- The logistic regression model effectively prioritizes leads, potentially increasing the conversion rate.
- Future steps include refining the model and incorporating more data to further improve accuracy.

Conclusion

Summary:

- Successfully built a model to predict lead conversion.
- Improved focus on potential leads can increase conversion rate.

Future Work:

- Further tuning of the model.
- Incorporate more features for better predictions.

Thank You!