Summary Report on Lead Score Case Study

Problem Statement

X Education aims to improve its lead conversion process by focusing on high-potential leads, termed 'Hot Leads'. Specializing in online courses for professionals, the company aims to elevate its 30% conversion rate to around 80%. A sophisticated lead scoring model will be employed, prioritizing leads with higher conversion likelihood. This data-driven approach empowers the sales team for targeted communication, fostering engagement and personalized interactions. This initiative streamlines the lead conversion funnel, nurturing leads effectively and driving substantial growth.

About Dataset Provided

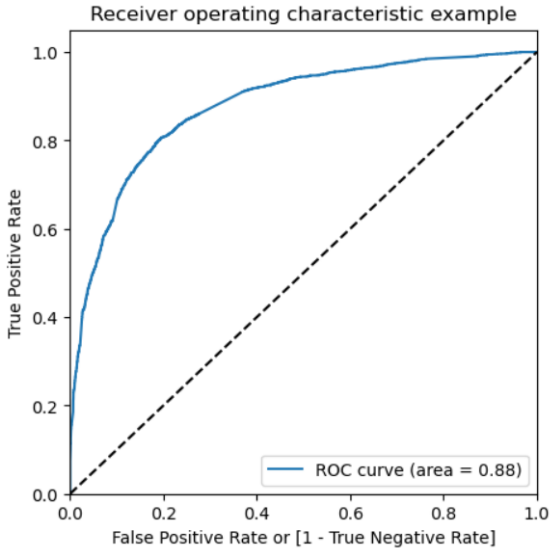
Provided is a dataset of around 9000 data points aimed at enhancing lead conversion strategies. Attributes like Lead Source, Total Time Spent on Website, Total Visits, and Last Activity contribute to the focus on the 'Converted' column (1 for success, 0 for failure). Notably, categorical variables include a 'Select' level, similar to null, necessitating special treatment. A data dictionary in the zip folder clarifies attribute roles. Through data-driven insights, this initiative seeks to elevate lead conversion tactics, improving business performance and customer engagement.

Model Building

Steps to be Followed

- **Dataset Reading**
- **Data Understanding**
  This Includes checking the shape of the data-frame, data types of all the columns, null value count for each column, null value percentage, unique values present in each column.
- **Exploratory data analysis**
  This includes dropping of the columns which has null value percentage higher then 40, replacing Select values with Null, null value imputation and Univariate and Bi-variate Analysis.
- **Data Preparation**
  Dummy variable Creation, mapping of columns which has Yes and No with 1 and 0.
- **Splitting the data in Test and Train**
  Splitting the dataset in test train in the ratio of 30 and 70
- **Feature Scaling**
  For numerical columns checked for any outliers if there's any, no major outliers were found therefore used MinMax Scalar for scaling the numerical features.

- **Feature Selection using RFE**
  RFE will identify the least significant features in each iteration and eliminate them from the model. This step helps in retaining only the most relevant features that significantly impact the conversion outcome.
- **Model Building**
  Build a Logistic regression model with total of 15 features, all the features are having p-significance value less than 0.05 and Variance Inflation factor less than 4.
- **Confusion Matrix**
  Train Data
  Accuracy - 79.78
  Specificity - 81.43
  Sensitivity - 78.77
- **AUR-ROC Curve**



- **Predication on Test Set**
  Test Data
  Accuracy - 79.82
  Specificity - 82.62
  Sensitivity - 78.09

Conclusion

The model showcases strong predictive performance with an accuracy of 79.78% on the training data and 79.82% on the testing data. The metrics of specificity (81.43% and 82.62%) and sensitivity (78.77% and 78.09%) underscore the model's proficiency in accurately identifying both non-converted and converted leads across datasets.

Feature Importance Analysis

An analysis of feature coefficients reveals influential factors. Notably, variables like Total Visits, Total Time Spent on Website, and the status of being a working professional exhibit

positive coefficient, indicating their positive impact on lead conversion likelihood. Conversely, certain lead sources, lead searches, and specific last activities display negative coefficients, suggesting their potential hindrance to conversion.

Hot Leads Selection

The stratification of leads based on lead scores is indicative of their potential conversion propensity. Leads with scores exceeding certain thresholds (90, 80, 70, and 60) form substantial clusters, highlighting varying degrees of conversion potential. This insight empowers targeted engagement strategies tailored to lead scores, optimizing conversion outcomes through resourceful interactions.

- Lead Scores > 90: The dataset encompasses a robust count of 283 leads whose lead scores exceed the threshold of 90.
- Lead Scores > 80: The dataset further includes a substantial cohort of 477 leads that possess lead scores surpassing 80.
- Lead Scores > 70: Within the dataset, a significant assembly of 602 leads showcases lead scores surpassing 70.
- Lead Scores > 60: Additionally, the analysis identifies a substantial grouping of 698 leads with lead scores exceeding 60.