LEAD SCORING CASE STUDY

Enhancing Lead Conversion for X Education

Koteshwar Reddy Pullaiahgari
Chaitanya Prabhune
Pradip Tamgale

Introduction

- X Education, an education company, aims to boost its lead conversion rate by targeting "Hot Leads," or those with a higher probability of converting into customers.

- The goal is to build a model that assigns a lead score to each prospect, enabling the sales team to prioritize their efforts and achieve a conversion rate of approximately 80%.

Data Overview

- Data set has a total of 37 columns and 9240 rows.
- Out of all the 37 columns , 4 column values are of float , 3 column values are of Int and 30 column are of object type.
- Drop columns with over 30% missing values if they don’t provide important information. This keeps the data cleaner and helps the analysis or model perform better .
- Fill missing values with averages or predictions.This keeps the data clean and ensures accurate analysis.

Exploratory Data Analysis (EDA)

Based on univariate analysis, we found columns with too many missing values, constant values, and low variability that should be removed to improve the dataset's quality.

Columns which has removed

['Magazine', 'Search', 'Newspaper Article', 'X Education Forums','Country' 'Newspaper', 'Digital Advertisement', 'Lead Number', 'Through Recommendations','Receive More Updates About Our Courses', 'Update me on Supply Chain Content','Prospect ID', 'Get updates on DM Content','I agree to pay the amount through cheque']

Data creation

Created dummy variables for categorical data variables.

Columns used for dummy creation:

'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'What matters most to you in choosing a course', 'A free copy of Mastering The Interview', 'Last Notable Activity'

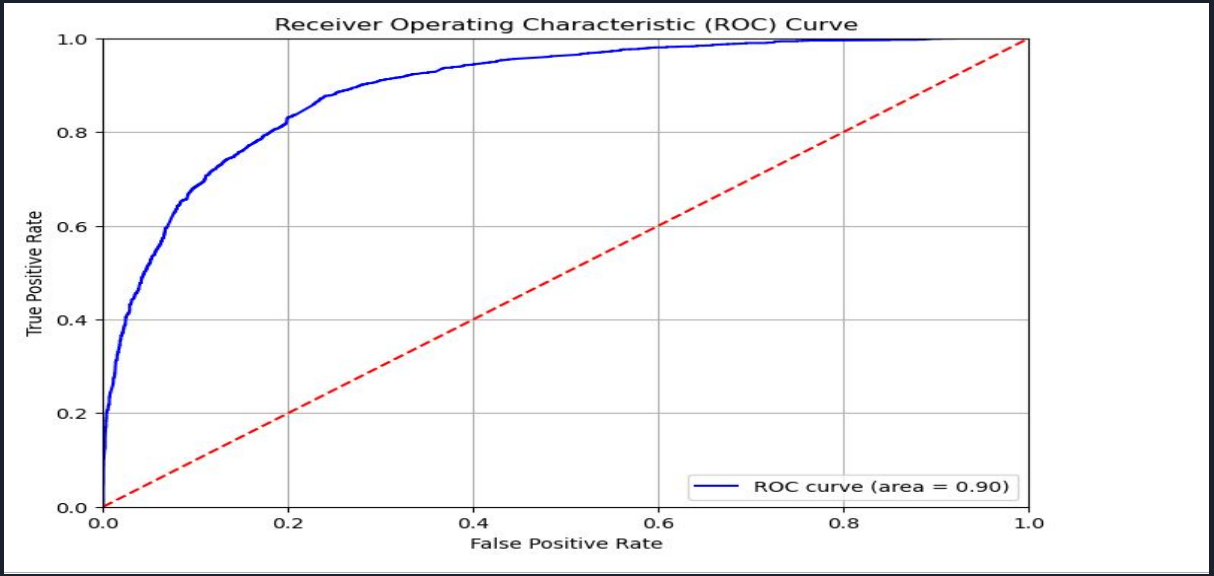
Model creation

Build Logistic regression model with RFE.

Important feature list

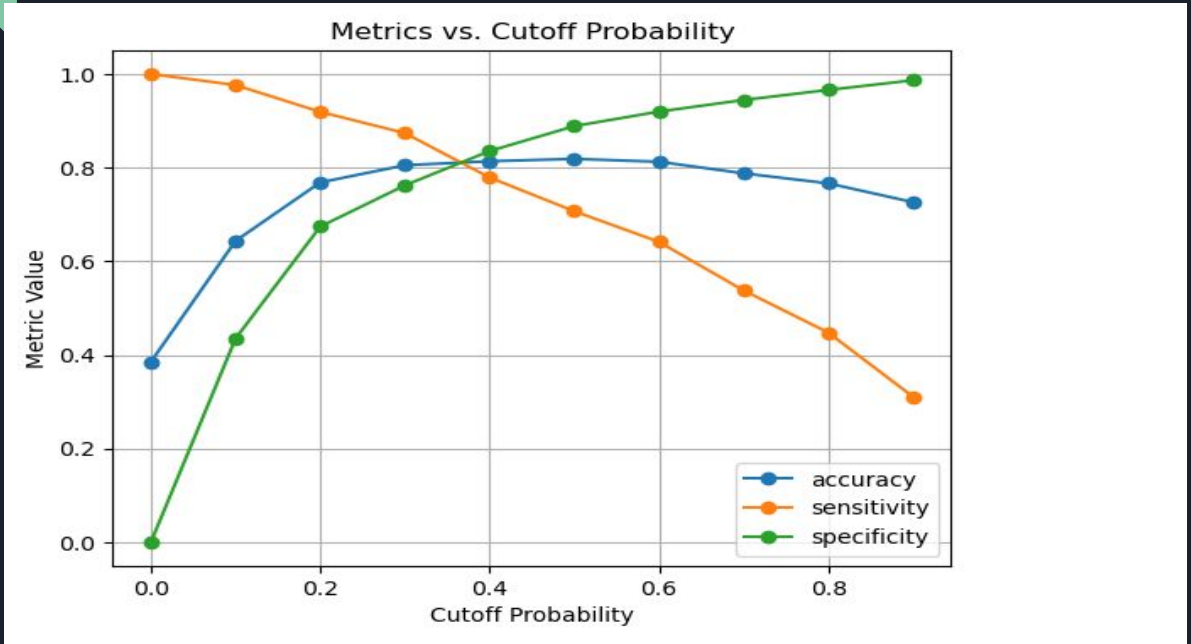
Lead Source_Welingak Website , Lead Source_Reference, What is your current occupation_Working Professional, Last Activity_Other_Activity, Last Notable Activity_Unreachable, Last Activity_Unsubscribed, Last Activity_SMS Sent, Total Time Spent on Website, Lead Source_Olark Chat, Last Notable Activity_Modified, Last Activity_Olark Chat Conversation, Lead Origin_Landing Page Submission, Specialization_Others

ROC CURVE



An AUC of 0.90 indicates excellent model performance, with a 90% chance of correctly classifying positive and negative cases.

Metrics vs. Cutoff Probability



The optimal cutoff probability was found to be approximately 0.35,

Model Performance Metrics

Train dataset

The model is performing well with an accuracy of 81.80%, correctly identifying 88.86% of positive cases and having a positive predictive value of 82.97%. However, its specificity is 70.97%, and the negative predictive value is 79.56%, indicating there's some room for improvement in classifying negative cases.

Test dataset

The model is performing decently with an accuracy of 79.69%, correctly identifying 76.82% of positive cases and achieving a high positive predictive value of 89.82%. However, its specificity is 84.73%, and the negative predictive value is 67.58%, suggesting there's room for improvement in accurately classifying negative cases.

Recommendation

- The company should make calls to leads from "Welingak Websites" and "Reference," as these are more likely to convert.
- Also target working professionals.
- Include those whose last activity was sending an SMS or other activities.
- Engage with those who spend more time on the website.
- Contact leads coming from "Olark Chat."
- The company should avoid calling leads whose last activity was "Olark Chat Conversation," as they are less likely to convert.
- Avoid leads coming from "Landing Page Submission."
- Skip those with the "Do Not Email" option active.

Thanks !!