Lead scoring case study

By Priyanka vaish

Problem Statement

- X Education, an online course provider, faces a challenge with its lead conversion process. Despite attracting numerous leads through website visits and referrals, only a small percentage actually convert into paying customers. The company aims to enhance this process by identifying the most promising leads, referred to as 'Hot Leads,' with the potential to convert at a higher rate.
- To achieve this, they seek to build a model that assigns a lead score to each prospect. The higher the lead score, the more likely a lead is to convert into a paying customer. By focusing efforts on leads with higher scores, X Education aims to increase its overall lead conversion rate, with a target set at around 80%.

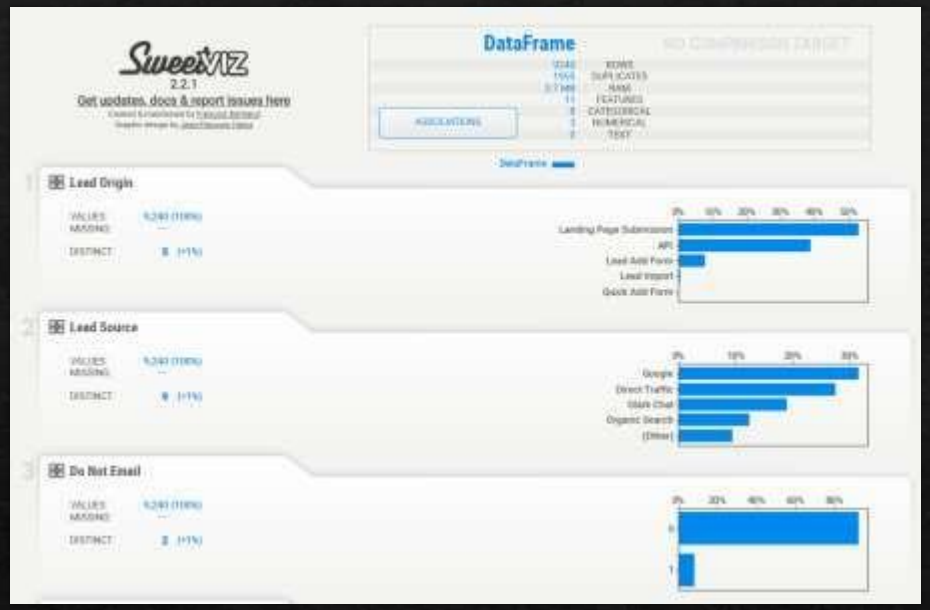
Business Objective

- **The primary objective of this project is to increase the lead conversion rate from the existing 30% to approximately 80%. This will be achieved by implementing a lead scoring system that evaluates and prioritizes leads based on their conversion potential.**

Data cleaning and data manipulation

- As mentioned in problem statement many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value, So ‘Select’ values were converted to null.
- 2. Checked and handle null values and missing values.
- 3. Dropped the columns containing missing values more than 40%.
- 4. Dropped the columns that are not useful for the analysis.
- 5. Visualized data and dropped highly skewed columns.
- 6.Imputation of the values as required done.
- 7. Dropped the columns highly correlated to each other.
- 8. Grouped low frequency values, or labeled as ‘Other’

Auto EDA


SweetViz
2.2.1
Get updates, docs & report issues here
DataFrame
Lead Origin
Lead Source
Do Not Email


Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
Last Activity


Last Activity
Specialization
What is your current occupation
A free copy of Mastering The Interview

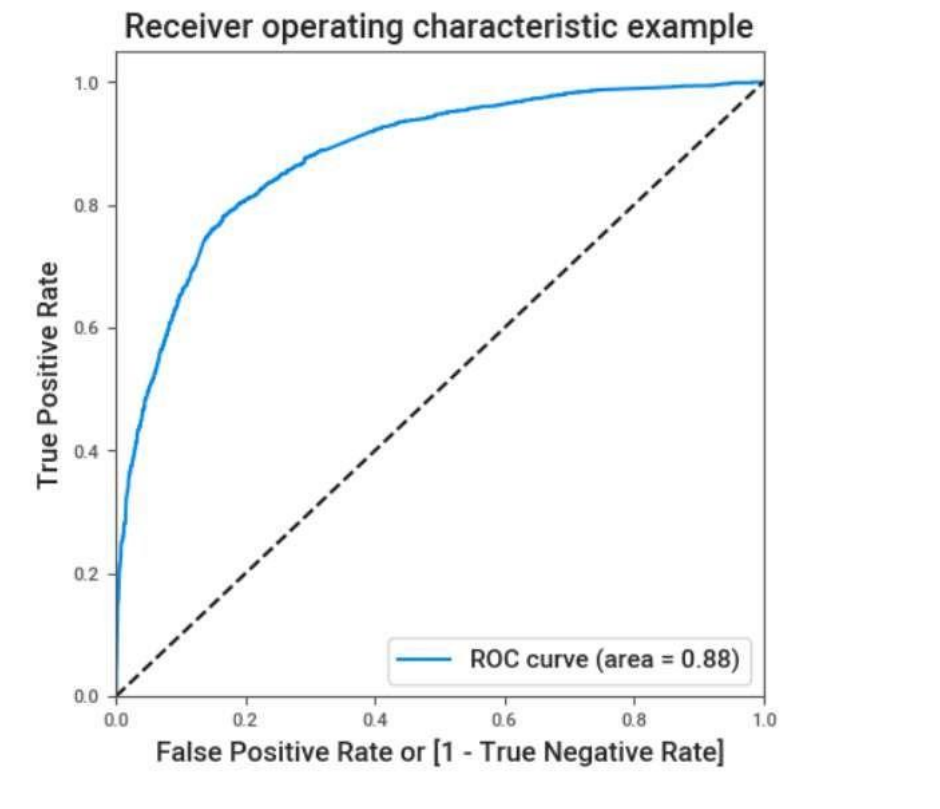
Dummy Variable Creation

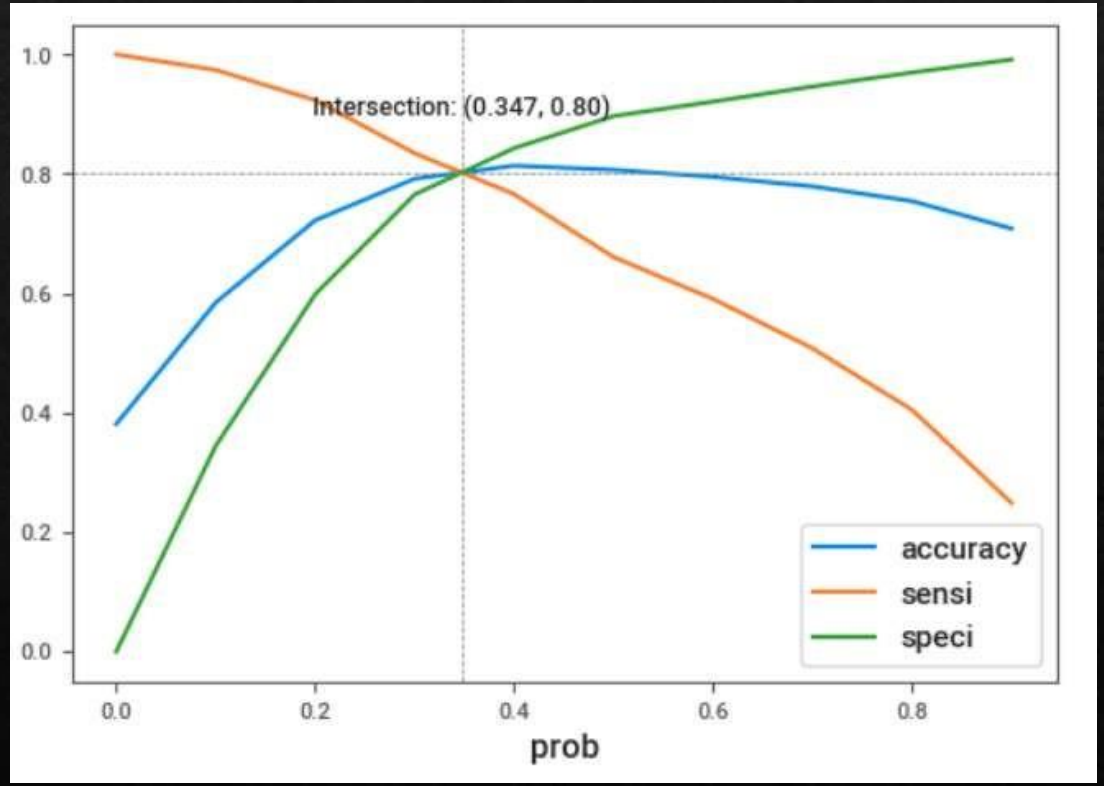
- Dummy variables are used to convert categorical data into a numerical format suitable for machine learning models. They help represent categories with binary values (0 or 1), enabling algorithms to work with non-numeric data. Dummy variables enhance model compatibility, prevent misinterpretation, and contribute to better performance by accurately representing categorical information.
- Created Dummy variables and dropped the categorical columns

Model Building

- Splitting Train & Test Sets: 70:30 ratio
- Feature Scaling using Standardization
- Used RFE to reduce variables from 48 to 15. To make dataframe more manageable. Dropped variables with p – value > 0.05.
- Total 2 models were built before reaching final Model 3 which was stable with (p-values < 0.05) and there was no sign of multicollinearity with VIF < 5.
- Logm3 was selected as final model with 13 variables, we used it for making prediction on train and test set.

ROC Curve





From the second graph it is visible that the optimal cut off is at 0.347

Model Evaluation

- Confusion matrix was made and cut off point of 0.347 was selected based on accuracy, sensitivity and specificity plot. This cut off gave accuracy, specificity and precision all around 80%. Whereas precision recall view gave less performance metrics around 75%.
- As to solve business problem where CEO asked to boost conversion rate to 80%, but metrics dropped when we took precision-recall view. So, we will choose sensitivity-specificity view for our optimal cut-off for final predictions
- Lead score was assigned to train data using 0.347 as cut off.

Making Predictions on Test Set

- Scaling and predicting using final model.
- Evaluation metrics for train & test are very close to around 80%.
- Lead score was assigned.
- Top 3 features are:

➢ Lead Source_Welingak Website

➢ Lead Source_Reference

➢ What is your current occupation_Working Professional

Observations

- Focus on features with positive coefficients for targeted marketing strategies.
- More budget can be spent on Welingak Website in terms of advertising ,marketing etc.
- Working professionals to be can be targeted as they have high conversion rate and will be able to pay high fees too as they are financially able.
- Areas with negative coefficients like 'Lead Origin_Landing Page Submission' should be analysed and put more work on.

THANK YOU