Lead Scoring Case Study

Ritika Mehra

Manit Malhotra

Varun Biyyala

Batch C47

Problem Statement

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. The typical lead conversion rate at X education is around 30%.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor.
- X Education has appointed you to help them select the most promising leads. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Strategy

1. Data understanding

    1.1 Loading dataset and describing them

    1.2 Data Formatting

    1.3 Dropping columns having higher percentage of missing values

2. Data cleaning

    2.1 Handling missing data

    2.2 Handling outliers

    2.3 Checking for data bias

    2.4 Data Imbalance Ratio (Conversion Ratio)

3. Data Analysis

    3.1 Univariate Analysis

    3.2 Bivariate Analysis

Strategy

4. Data Preparation

   4.1 Reducing Categories for categorical variables

   4.2 Creating dummy variables

   4.3 Multicollinearity check using correlation matrix

5. Model Building and Evaluation

   5.1 Train and Test Data Split

   5.2 Standardization

   5.3 Model Building and Prediction

   5.4 Model Evaluation

   5.5 Model Prediction on Test data

   5.6 Model Interpretation

Assumptions

- 40% cut-off selected for dropping columns with missing values.
- Pie chart has cut-offs mentioned for better visualization.
- Other category is not considered while mentioning EDA observations.
- Train Size = 0.7, Test Size = 0.3, Random State = 100

Approach

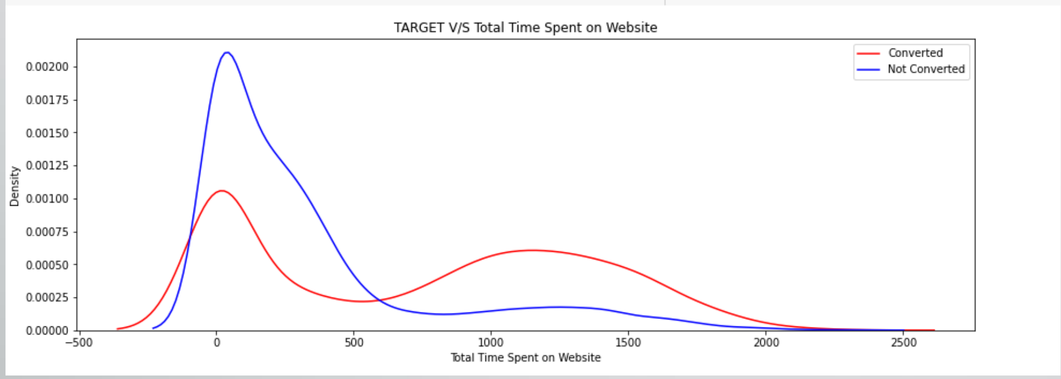
- Select category was replaced with NaN.
- Columns having more than 40% missing values were deleted.
- Other category was created for categorical variables with > 5% missing values, for the rest missing value replaced with mode.
- Missing values in numeric variables replaced with median due to presence of outliers
- Outliers were replaced using upper and lower limit of IQR.
- Columns with high data bias were removed.
- Since conversion rate of data is 38%, separate data frames created for EDA.
- Categories were reduced for dummy variables creation.
- Columns having high multi-correlation based on correlation matrix were removed.

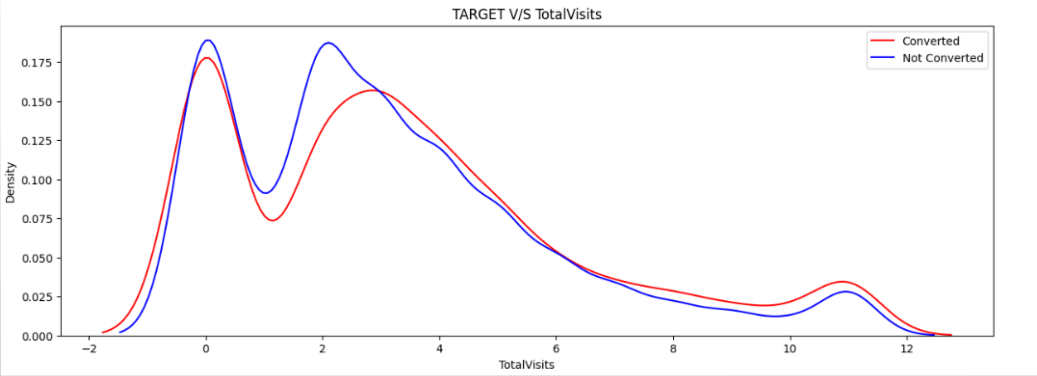
Approach

- After splitting in test and train data frames, numeric variables were standardized using standard scaler.
- RFE was used to select top 15 significant variables.
- Columns having high p-value were reduced one by one.
- ViF was checked for variables to ensure least multi-collinearity.
- Cut-off point was decided as 0.3 by plotting accuracy, sensitivity and specificity.
- Accuracy, confusion matrix, precision score and recall score were used as metrics for model evaluation.
- Lead score values were derived on the test based on the conversion probability.

EDA Observations

- Total Time Spent on Website > 600 seconds indicates higher chance of conversion.
- Total Visits > 3 indicates higher chance of conversion.
- Page Views Per Visits < 1 indicates slightly higher chance of conversion.

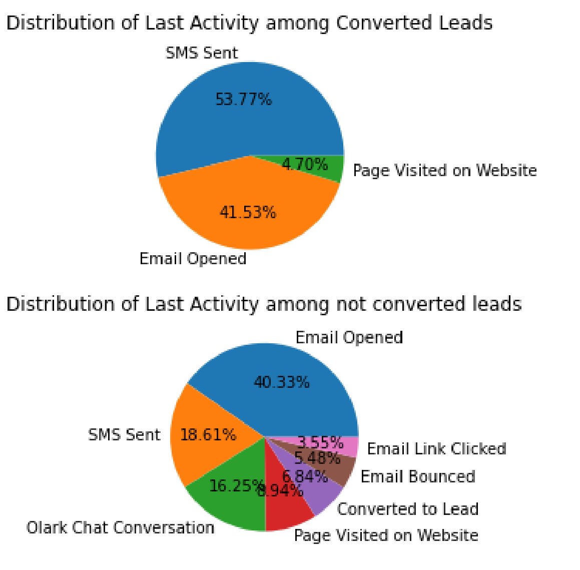


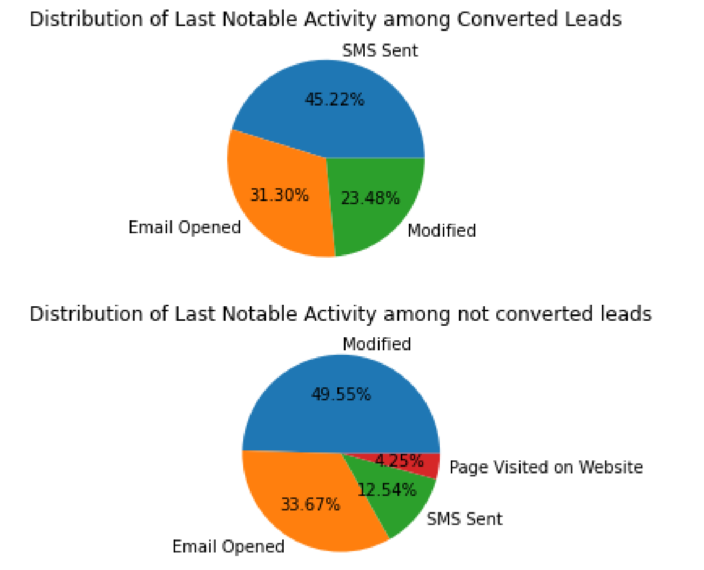


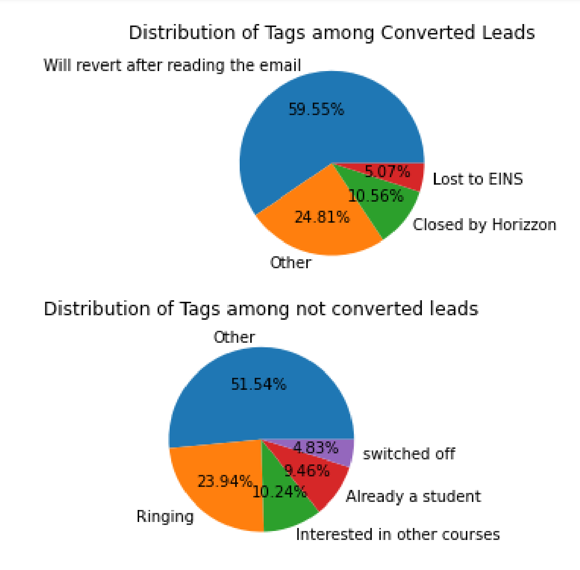
EDA Observations

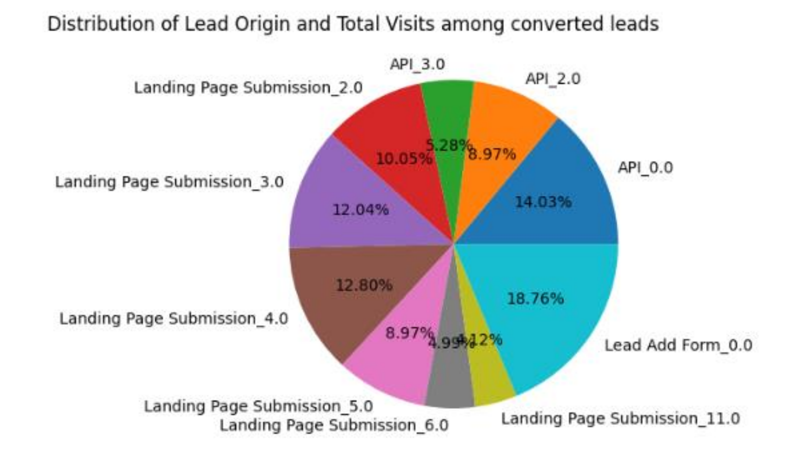
- Will revert after reading the email and Closed by Horizzon are common tags for converted leads.
- SMS Sent Activity has resulted in good conversion rate.
- Last Activity, Last Notable Activity, Tags and City have different categories contributing to converted and not converted leads, thus, they can be good factors.
- Landing page submission with total visits 4 has highest converted leads where as Landing page submission with total visits 2 has highest non converted leads.

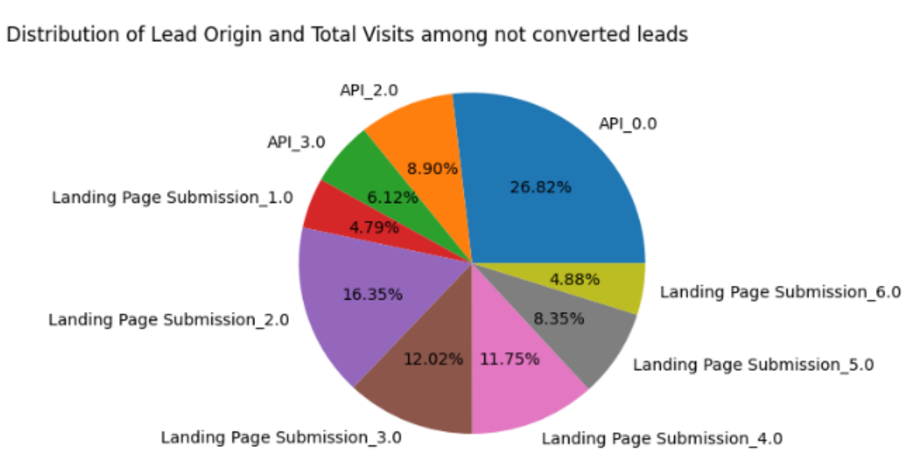
EDA Observations











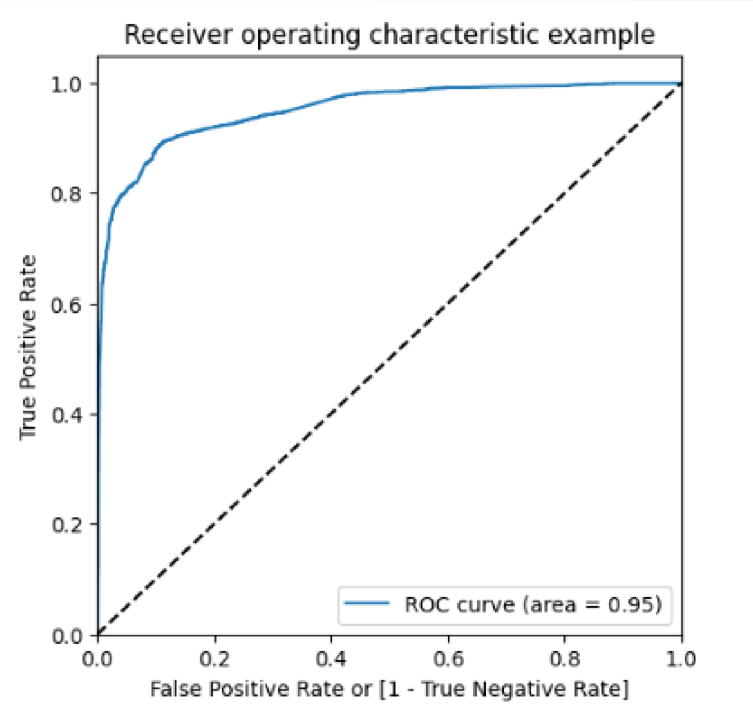
Model Evaluation

Train Data Frame:

- Accuracy: 0.895
- Specificity: 0.894
- Sensitivity: 0.887
- Precision Score: 0.837
- Recall Score: 0.887

Test Data Frame:

- Accuracy: 0.896
- Specificity: 0.90
- Sensitivity: 0.889
- Precision Score: 0.854
- Recall Score: 0.889



Model Interpretation

The following variables contribute the most to lead conversion,
Tags – Lost to Competitors, Reply Awaited
Lead Origin – Lead Import
Last Notable Activity – SMS Sent

Note – Categories from actual data were merged to form Lost to Competitors and Reply Awaited which are as follows:
Lost to Competitors – Closed by Horizzon, Lost to Others, in touch with EINS, Lost to EINS
Reply Awaited - In confusion whether part time or DLP, Still Thinking, Will revert after reading the email

Recommendations

- Even if a lead is lost to competitors, there is still a possibility to convert them. Providing them with offers may be useful in converting them.
- Following up with leads that are yet to confirm back by organizing doubt resolution or alumni sessions can help improve the chances of conversion.
- Not to invest much time on leads that are imported from other sources.
- Working with development team in making landing page more user friendly.
- Allocating lesser budget on channels like Magazines, Newspapers, Digital marketing etc, and increasing the budget on Referrals.
- Focusing on leads that came through reference or are working professionals may also be useful.
- Filtering out leads and excluding where the email had bounced, interested in other courses or preferring leads who spent higher amount of time on website.