Lead Scoring Case Study

Presented by :
Shubham Gupta
Sukanya
Sristi

Agenda

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 company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. 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, i.e. the leads that are most likely to convert into paying customers. 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%.

Approach

1 Extract the data for analysis

2 Reading and understanding the data

3 Data Cleaning & Preparation

4 EDA

5 Splitting the data into train-test dataset

6 Feature Scaling

7 Model Building

8 Model Evaluation : Specificity, Sensitivity, Precision & Recall

9 Making Predictions on the test data

Data Extraction Cleaning and Preparation

1. Read the data from CSV file
2. Checking for numerical columns- Univariate Analysis
3. Checking for outliers
4. Checking for Categorical columns
5. Converting column with binary values to 0/1
6. Creating Dummy variables for more than 2 categories
7. Train-Test Split
8. Feature Scaling

Outliers - Numerical Variables

Box plot for below variables:

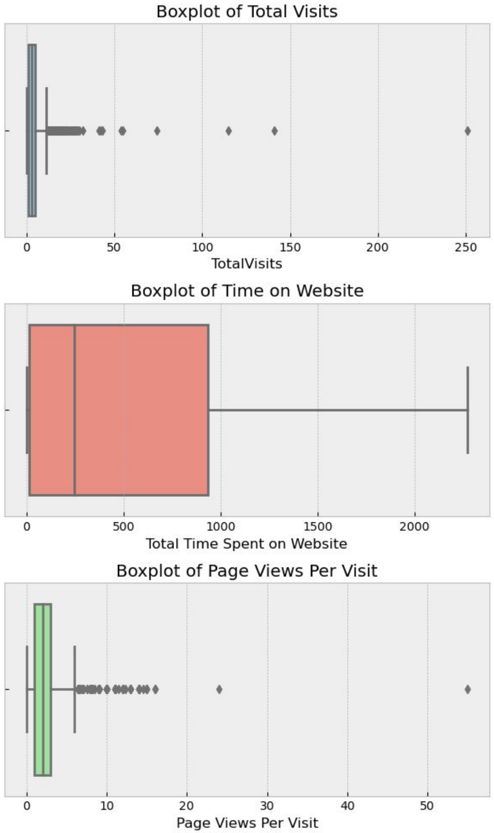
- Total Visits
- Time on Websites
- Page Views per Visits

Observations from Analysis:-

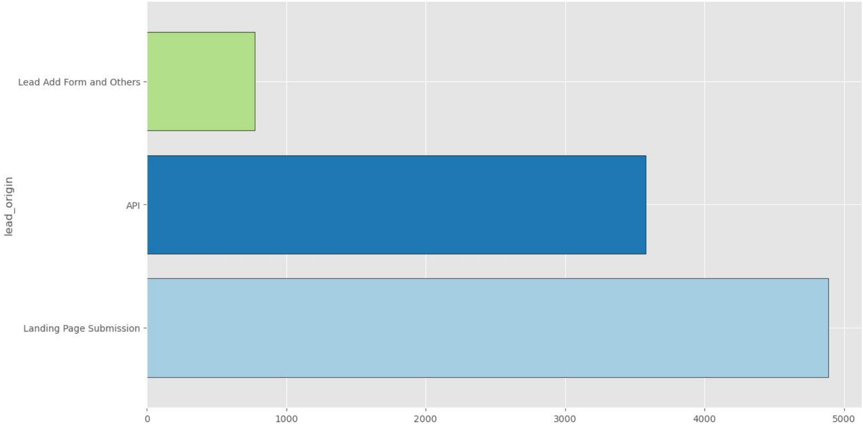
1. There are upper bound outliers in both “total_visits” and “page_views_per_visit”
2. It is clear that the data can be capped at the 99th percentile and that there are upper bound outliers in both variables

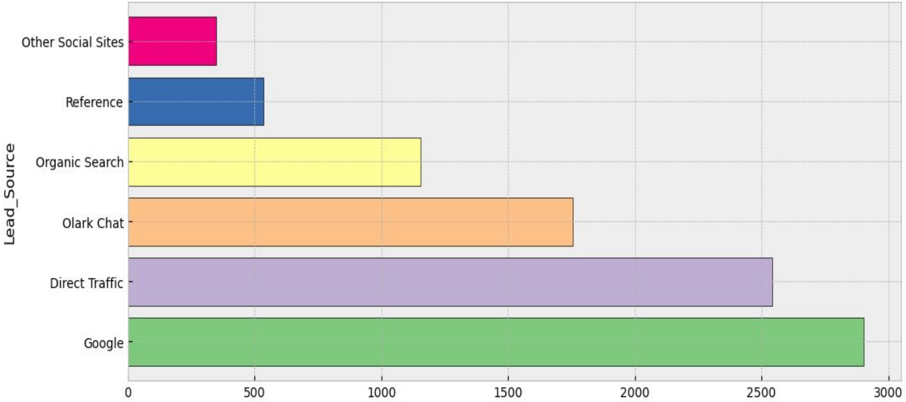
Recommendation

1. Outliers needs to be treated for both the variables



Categorical Variables Analysis

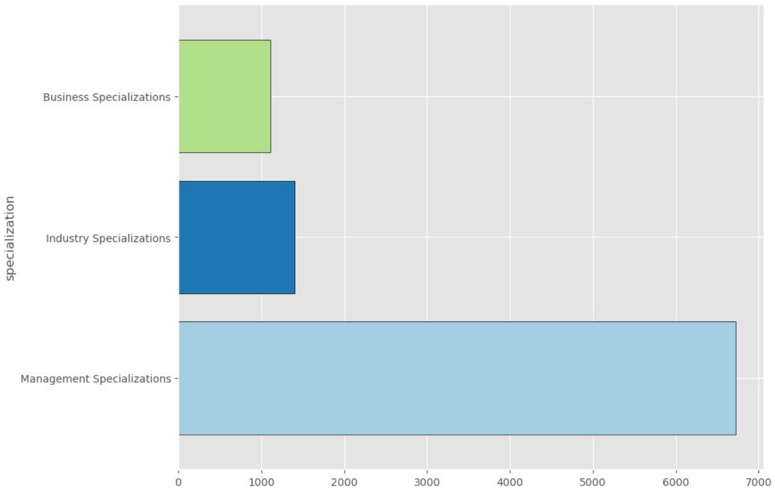


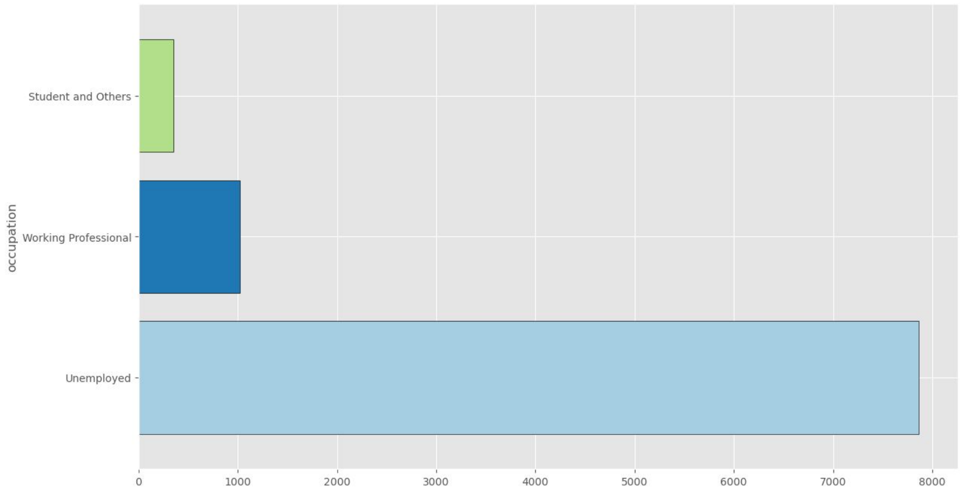


Leads_Origin:
Observation : “Landing page Submission” are having high conversion rate

Leads_Source:
Observation : Google sources are having high conversion rate

Categorical Variables Analysis

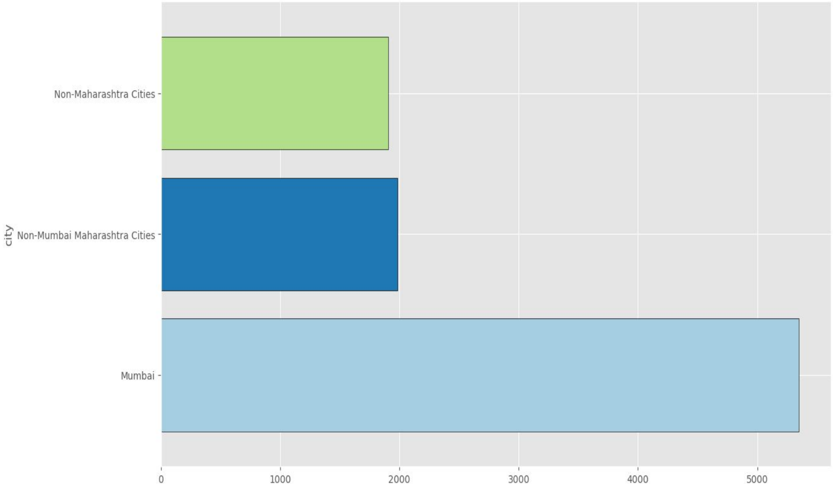


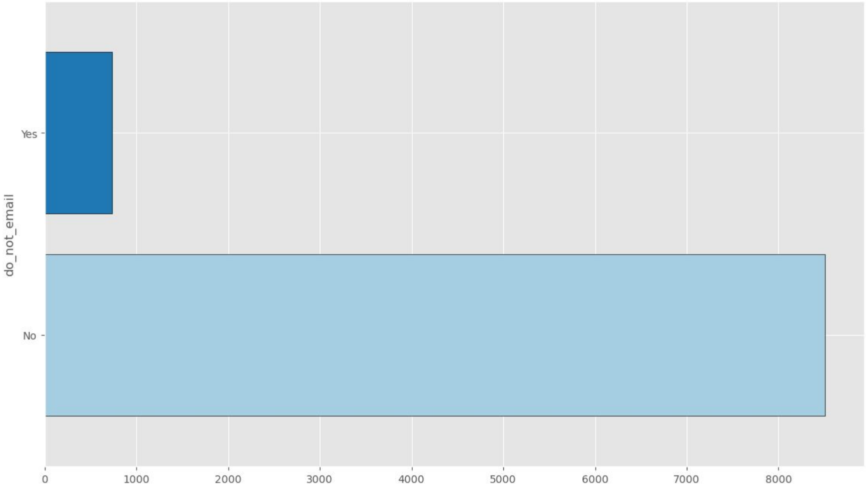


Specialization:
Observation : “Management” are having maximum conversion rate

Occupation:
Observation : “Unemployed” are considered for leads

Categorical Variables Analysis





City:

Created 3 variables by combining them:

- Non-Maharashtra Cities
- Non-Mumbai Maharashtra Cities
- Mumbai

Observation : “Mumbai” are significant leads

Do_not_email:

Observation : “No” will have higher chance to become a lead

Data Preparation

- **Converting columns with binary variables to 0/1**
  - Do_Not_Email
  - A_free_copy_of_Mastering_The_Interview
- **Created Dummy Variable for Categorical variables**
  - Lead_Origin
  - Lead_Source
  - Specialization
  - What_is_your_current_occupation
  - City

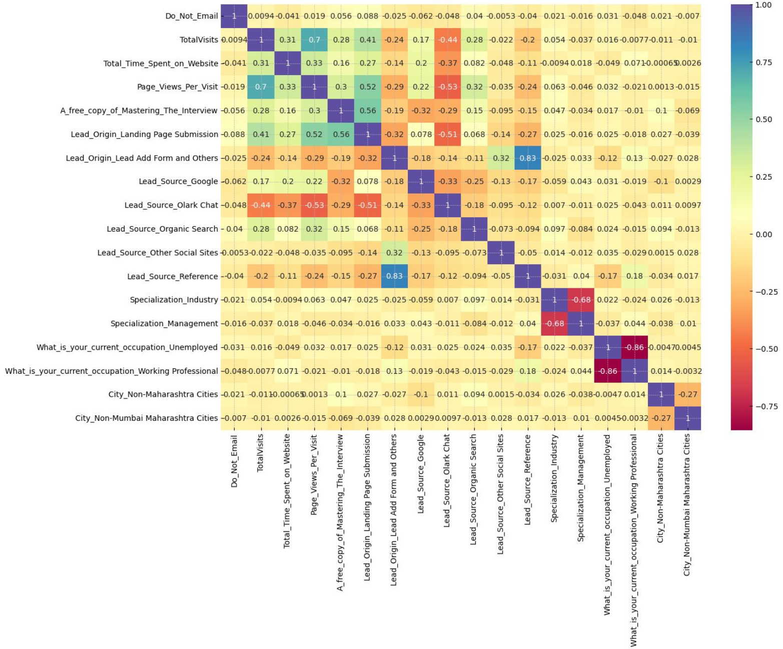
Feature Scaling & Splitting test & train Sets

- **Feature Scaling of Numerical variables**
  - Total_Visits
  - Total_Time_Spent_on_Website
  - Page_Views_Per_Visit
- **Splitting data into Test & Train Set**
  - Train : 70%
  - Test : 30%

Correlation

- **Looking for correlation with heatmap**
  - Total_Visits
  - Total_Time_Spent_on_Website
  - Page_Views_Per_Visit

- **Splitting data into test & Train Set**
  - **Analyse highly correlated variables**
    - lead_origin_Lead Add Form and Others
    - specialization_Industry
    - What_is_your_current_occupation_umemployed
    - What _is_your_occupation_Working_Professional



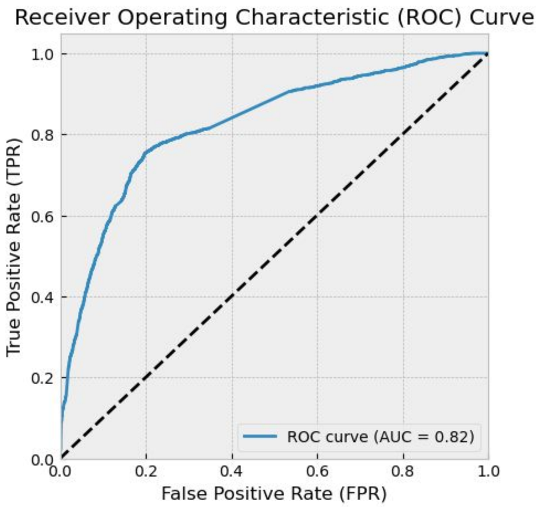
Model Building

- **Feature Selection using RFE with 15 variables**
    - Do_Not_Email
    - Total_Visits
    - Total_Time_Spent_on_Website
    - Page_Views_Per_Visit
    - Lead_Origin_Lead Add Form and Others
    - Lead_Source_Google
    - Lead_Source_Olark Chat
    - Lead_Source_Organic Search
    - Lead_Source_Other Social Site
    - Lead_Source_Reference
    - Specialization_Industry
    - Specialization_Management
    - What_is_your_current_occupation_Working Professional
    - City_Non-Maharashtra Cities
    - City_Non-Mumbai Maharashtra Cities

Model Building

- **Determine Optimal model using Logistic Regression Model**
  - Selecting variables with p-values less than 0.05 and VIFs less than 5

- **Analysed:**
  - Confusion matrix
  - Accuracy
  - Sensitivity
  - Specificity
  - Precision
  - Recall
  - ROC Curve
  - Evaluate Model
  - True Positive rate
  - False Positive rate

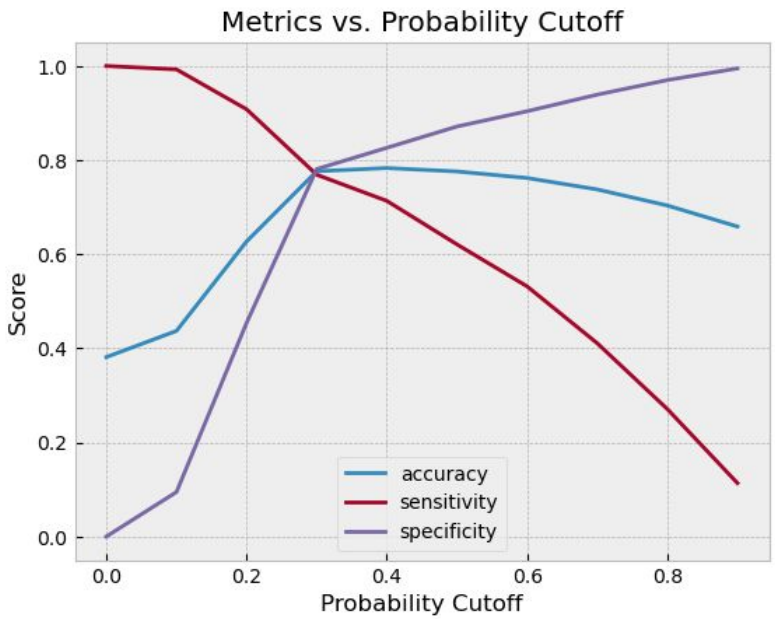


Variable Impacting Conversion Rate

- Do_Not_Email
- Total_Visits
- Total_Time_Spent_on_Website
- Lead_Origin_Lead Add Form and Others
- Lead_Source_Google
- Lead_Source_Olark Chat
- Lead_Source_Other Social Sites
- What_is_your_current_occupation_Working Professional

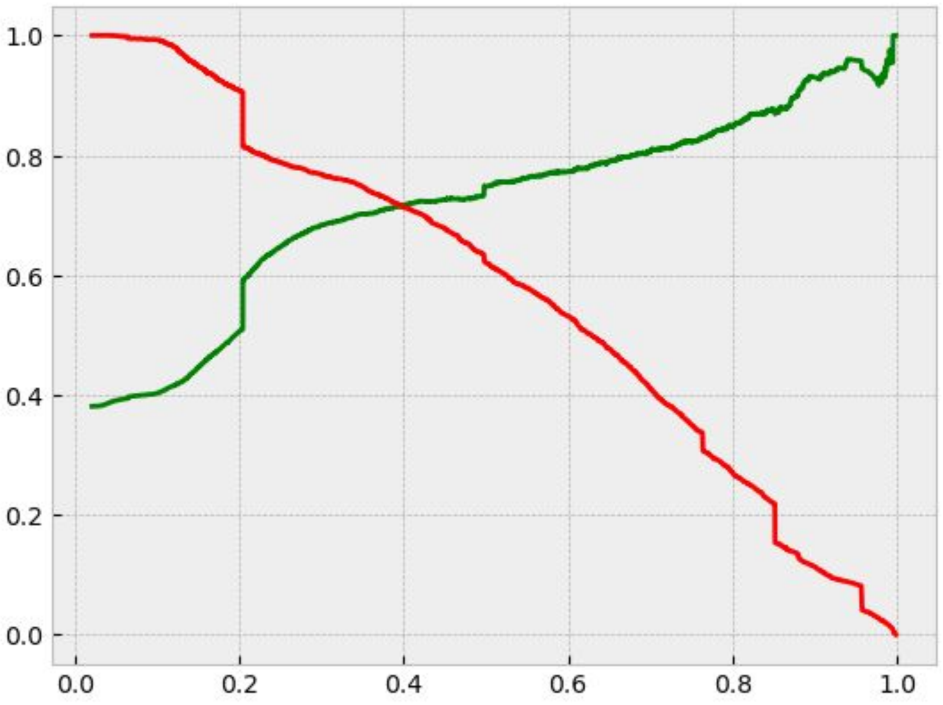
Model Evaluation : Sensitivity & Specificity on Train Data

- Graph depicts an optimal cutoff : 0.3
  - Accuracy : 77.63%
  - Sensitivity : 76.89%
  - Specificity : 78.09%



Model Evaluation : Precision & Recall on Test Data

- Graph depicts an optimal cutoff : 0.39
  - Accuracy : 77.45%
  - Precision : 71.68%
  - Recall : 70.96%



Conclusion

After trying several models, we have finally chosen model 9 to be our final model with the following characteristics:

All variables have p-value < 0.05. All the features have very low VIF values, meaning, there is hardly any multicollinearity among the features. This is also evident from the heat map.

The overall accuracy 77.63% at a probability threshold of 0.39 on the test dataset is also very acceptable. In the initial stage, a large number of leads are generated, but only a few of them convert into paying customers. During the middle stage, it is crucial to nurture potential leads effectively, which involves educating them about the product and maintaining consistent communication.

To identify the best prospects among the leads, focus on factors such aLead_Origin_Lead Add Form and Others, ,What_is_your_current_occupation_Working Professional,' anTotal_Time_Spent_on_Websiteit,' as these contribute significantly to the probability of lead conversion.

It's essential to maintain a list of leads to keep them informed about new courses, services, job offers, and future opportunities for higher studies. Monitor each lead carefully and tailor the information you send to them based on their interests. Providing job offerings, course information, or services that align with the leads' interests will enhance the likelihood of converting them into prospects.

Place particular emphasis on converted leads. Conduct question-answer sessions with leads to gather the necessary information about them. Schedule follow-up inquiries and appointments with the leads to understand their intentions and willingness to join online courses. A well-thought-out plan to address the needs of each lead will significantly improve lead capture and conversion rates