Lead Scoring Case Study

By Kanika Khattar Ahuja
Divya Namani

UpGrad iiit-b ज्ञानमुत्तमम्

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.
- 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 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 higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

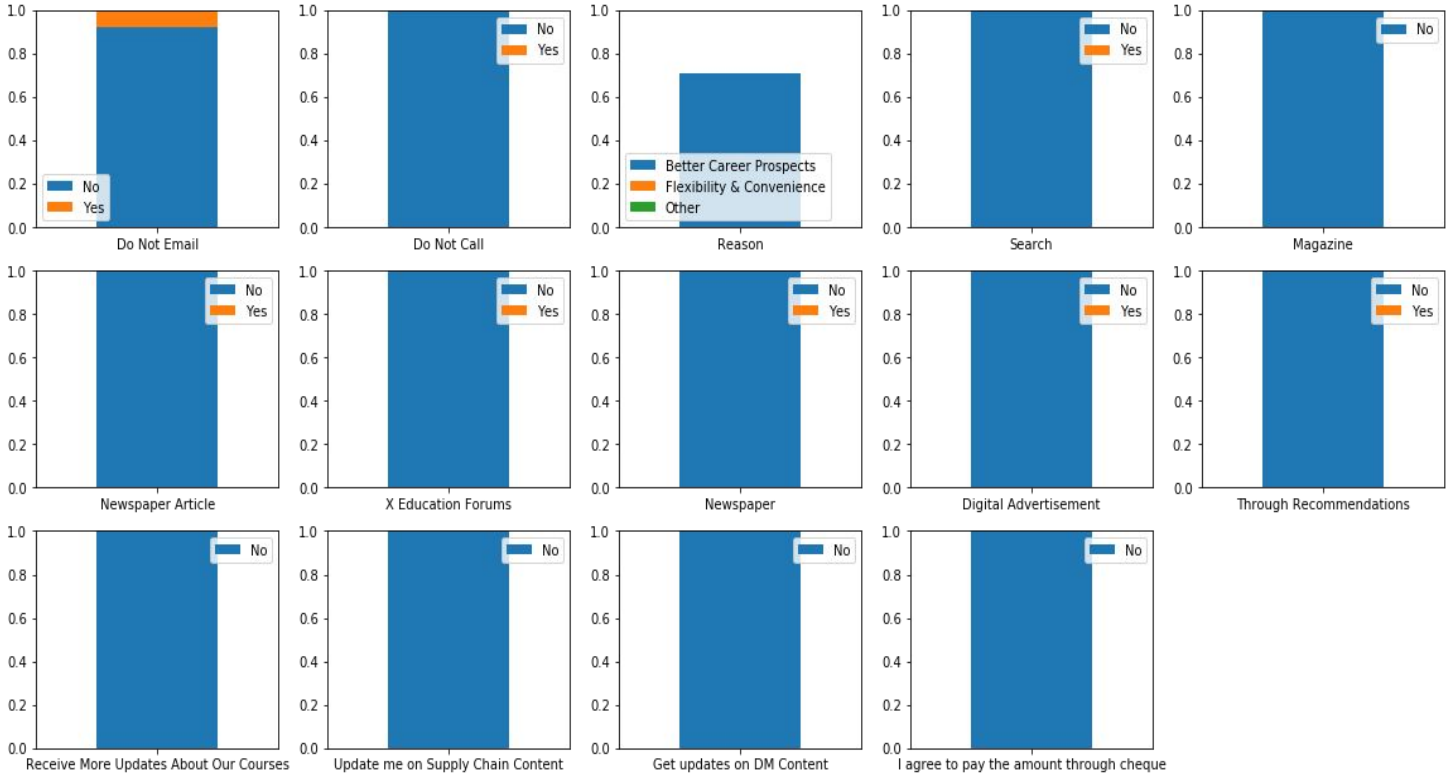
Analysis Approach [1/2]

- Step 1. Reading and Understanding the Data
- Step 2.Data Cleaning and EDA
  - 2.1 Missing value check
  - 2.2 Cleaning and Visualizing categorical variables
  - 2.3 Cleaning and Visualizing numerical variables
  - 2.4 Outlier Treatment
  - 2.5 Check for data type conversion
- Step 3. Preprocessing and Data Preparation
  - 3.1 Categorizing variables
  - 3.2 Creating dummy variables
  - 3.3 Train test split
  - 3.4 Scaling data

Analysis Approach [2/2]

- Step 4. Model Building
  - Step 4.1 Build Logistic Model
  - Step 4.2 Prediction and evaluation on Training Set
  - Step 4.3 Prediction and evaluation on Testing Set
- Final Analysis

Cleaning and Visualizing Categorical Variables



- Skewness in the categorical variables with more than 90% of value being same.
- All these features were removed as they did not provide any insights.

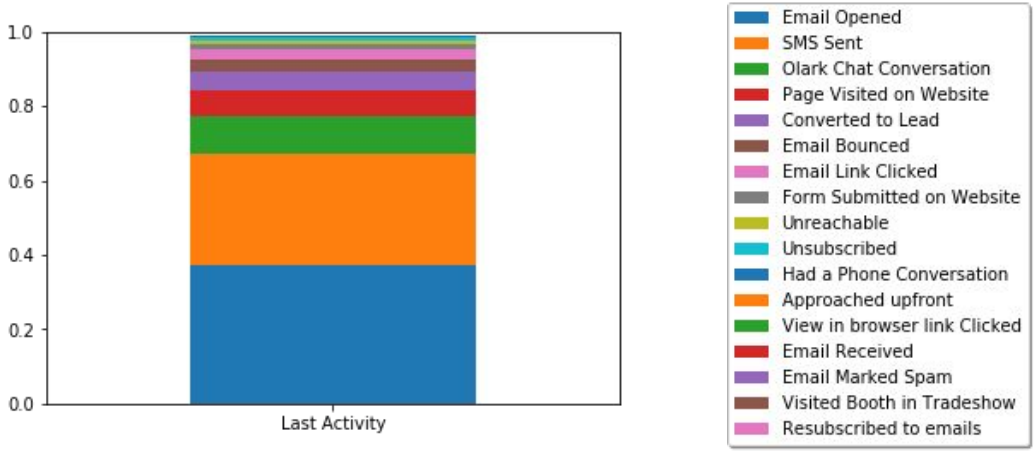
Cleaning and Visualizing Categorical Variables - Lead Source

- **Welingak Website** has the maximum conversion rate followed by **Reference**
- **Google** and **Direct Traffic** have top counts and a good conversion rate of close to 30%
- We noticed that Lead Source is highly affected by Lead Origin. Therefore, replaced the missing values of Lead Source by the mode of Lead Source depending on the Lead Origin

Lead Source	Counts	Total%	Converted
Google	2873	0.310931	0.399234
Direct Traffic	2543	0.275216	0.321667
Olark Chat	1755	0.189935	0.255271
Organic Search	1154	0.124892	0.377816
Reference	534	0.057792	0.917603
Welingak Website	142	0.015368	0.985915
Referral Sites	125	0.013528	0.248000
Facebook	55	0.005952	0.236364

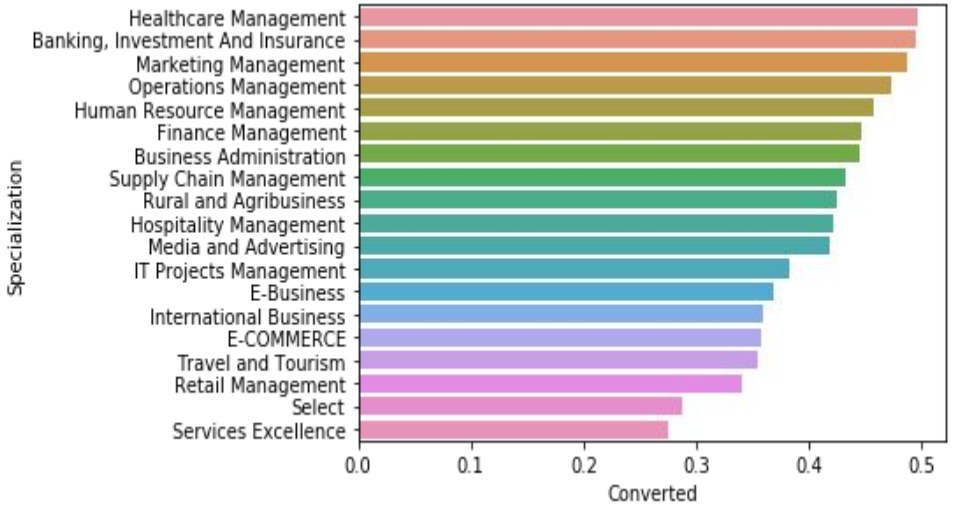
Cleaning and Visualizing Categorical Variables - Last Activity

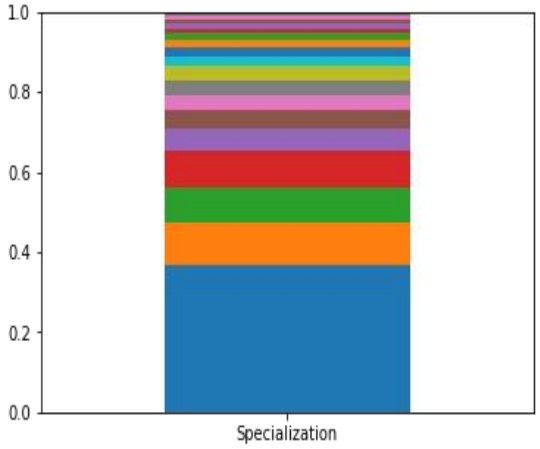
Last Activity	Counts	Total%	Converted
Email Opened	3540	0.383158	0.376836
SMS Sent	2745	0.297110	0.629144
Olark Chat Conversation	973	0.105314	0.086331
Page Visited on Website	640	0.069272	0.235937
Converted to Lead	428	0.046325	0.126168
Email Bounced	325	0.035177	0.076923
Email Link Clicked	267	0.028899	0.273408
Form Submitted on Website	116	0.012555	0.241379



- **SMS Sent** has the maximum conversion rate followed by **Email Opened.** They both also has the top counts.
- **Olark Chat Conversion** comprises of 10% of the leads and has the lowest conversion rate of 8%.
- Replaced 103 missing values of Last Activity with its mode Email Opened

Cleaning and Visualizing Categorical Variables - Specialization





- As the missing was large, replaced these value with others
- Leads with Finance Management as Specialization are maximum and has a good conversion rate.

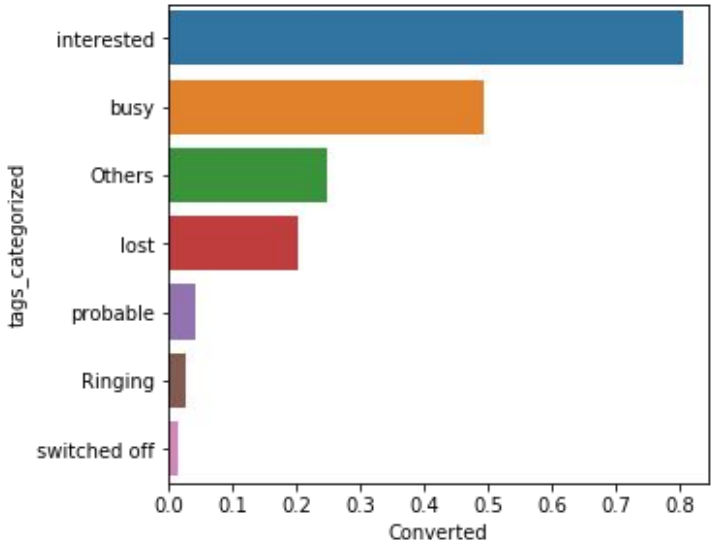
Cleaning and Visualizing Categorical Variables - Occupation

- 60% of leads are **Unemployed** and they have good conversion rate of **43%**
- Though number of **working professional** is less, but their conversion rate is higher at 91%
- The missing percentage is 29%. Replacing with mode might have skewed the data.
- Therefore, replaced null values in Occupation with value Other

Occupation	Counts	Total%	Converted
Unemployed	5599	0.606018	0.435792
Other	2706	0.292889	0.140429
Working Professional	706	0.076415	0.916431
Student	210	0.022730	0.371429
Housewife	10	0.001082	1.000000
Businessman	8	0.000866	0.625000

Cleaning and Visualizing Categorical Variables - Tags

- There were 27 different values of Tags.
- Bucketing them into following categories based upon the business knowledge
  - Interested
  - Busy
  - Probable
  - Lost
- Interested and busy tags had the maximum conversion rate



Cleaning and Visualizing numerical Variables

There are three Numerical columns

1.TotalVisits.

2.Total Time Spent on Website(Website time)

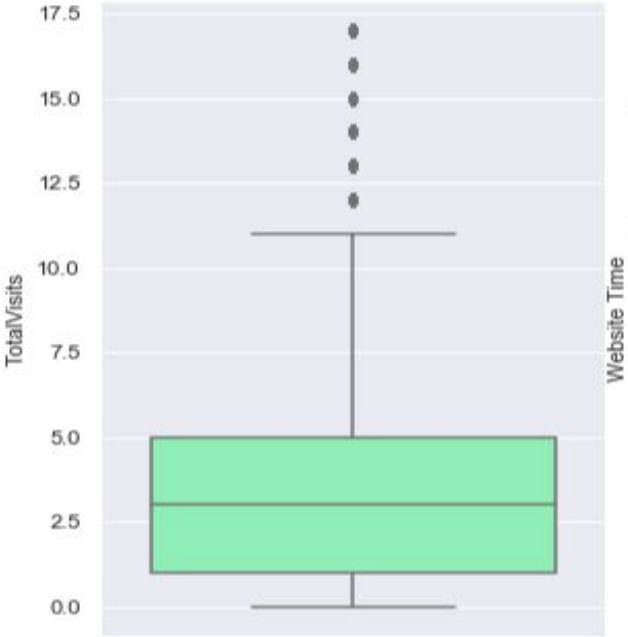
3.Page Views Per Visit

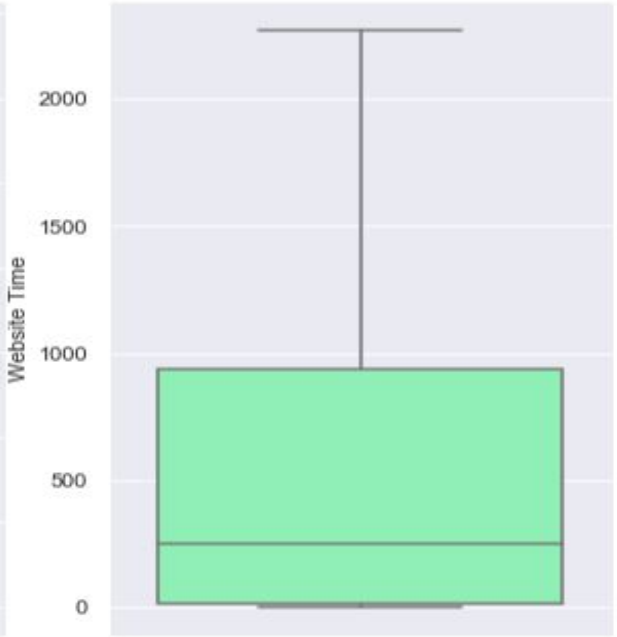
Outlier Treatment:

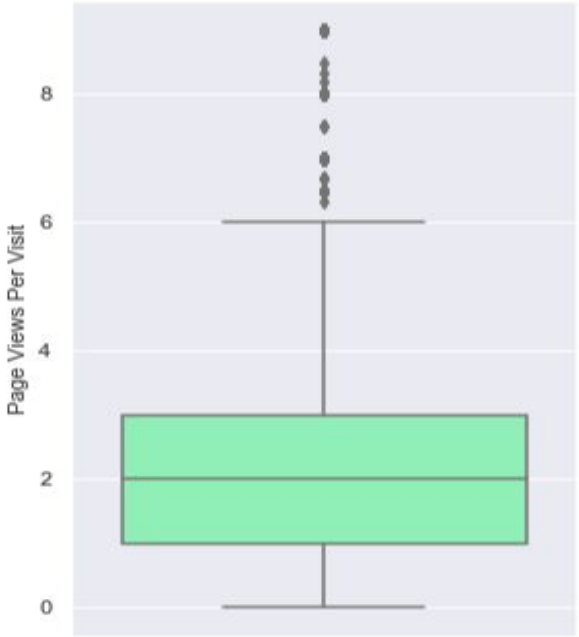
Columns TotalVisits and Page Views Per Visit have been capped at 99% values because of the presence of outliers.

Cleaning and Visualizing numerical Variables

Data spread of the numerical columns after treating the outliers:

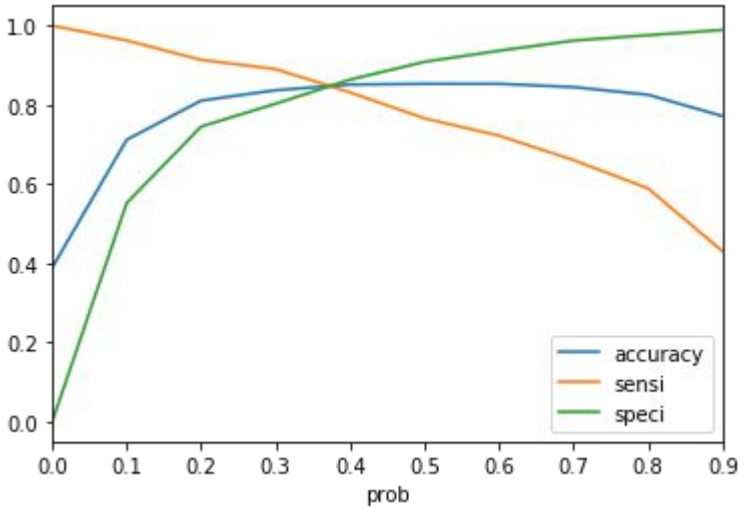






Model Building

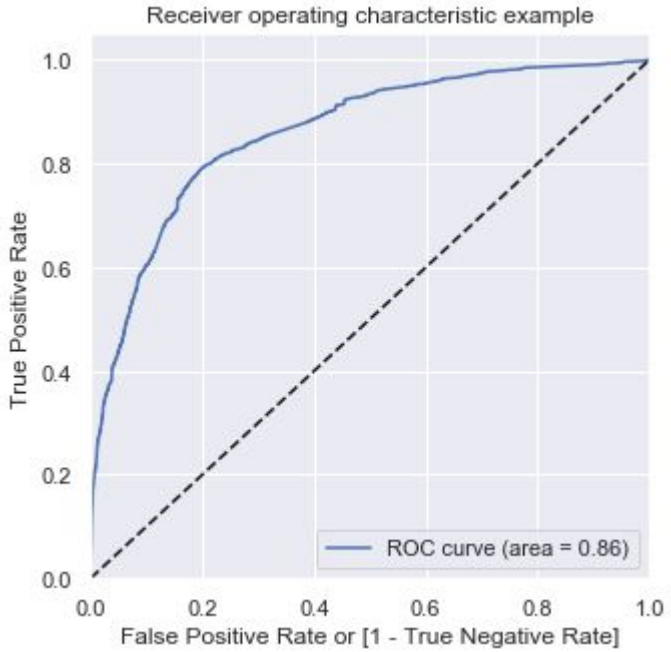
- **RFE** was used to reduce the variables
- Model was finalized with 12 variables.
- Optimal cutoff was decided based on **trade off of accuracy,sensitivity and specificity.**
- The model predicted the probabilities which was further used to calculate the **Lead Score**



The value 3.5 seems to be a good cutoff as all the metrics values are intersecting.

Evaluations on training data

- Training Accuracy: 0.79
- Training Sensitivity: 0.80
- Training Specificity: 0.79
- The F1 score of training is 80%
- The ROC curve area is 86% which is good enough



Evaluations on testing data

- Testing Accuracy = 0.79
- Testing Sensitivity: 0.80
- Testing Specificity: 0.79
- F1 score of test data is 80%

Final Analysis - Features

Features	Coeff
Occupation_Working Professional	2.9662
Lead Source_Reference	2.8502
Last Activity_SMS Sent	1.6785
Website Time	0.9859
Last Activity_Email Opened	0.6105
Lead Source_Olark Chat	0.2311
TotalVisits	0.0683
Lead Source_Direct Traffic	-0.2514
Country_India	-0.2983
Lead Origin_Landing Page Submission	-0.5406
Last Activity_Converted to Lead	-0.8215
Last Activity_Olark Chat Conversation	-0.9478

Final Analysis - Conclusion

As sensitivity is 80%, therefore, all the leads with lead score > 35 (a potential lead) have 80% chance of getting converted.

Dummies that contributes towards increasing the probability are:

- **Occupation_Working Professional** (currently 92% conversion rate)
- **Lead Source_Reference** (currently 91% conversion rate)

Of all the leads who are predicted as Converted, the sales team should follow the below strategy to increase the conversion rate:

- If all the leads who are a **working professional** are targeted, 92% of them has a chance of getting converted. These leads should be targeted first.
- Leads who come by **reference** have a chance of 91% conversion. Therefore, sales team should next target these leads.

Final Analysis - Conclusion

Variables that need to be worked on by the team are

- **Country_India**(38% conversion rate)
- **Last Activity_Olark Chat Conversation**(currently 8% conversion rate)
- **Last Activity_Converted to Lead(**currently 12% conversion rate)
- **Lead Origin_Landing Page Submission** (currently 36% conversion rate)

To increase the probability of conversion, teams should working on the following:

- A whooping 73% of the leads are from the **Country-India** ,but only 38% of them are converted.Hence,this is an indicator of area of improvement.
- Even though a major share of the leads originate from **Landing page submission** – around 52% of them ,only 36% of them are converted. Team can focus more on this area as there are many leads who have not converted.
- **Olark Chat Conversion** comprises only 10% of the leads and out of which only 8% are converted. Chat Conversation is a good tool to know more about candidates and market the products. Therefore, team should work more on getting candidates from Olark chat and should improve the marketing skills on this platform.
- **Converted to lead** in Last Activity comprises only 4% of the leads and out of which only 12% are converted. Converted to lead seems to be the first step when a particular person is treated as a lead. Teams should work on reaching to these customers at the earliest through other mediums like email, SMS or chat.

Final Analysis - Recommendations

- Reaching out to Working Professionals:
  - Team should give high priority to working professionals.Hence,they can be approached over phone only on weekends(as most of the Organisations have weekends as holidays).
  - On weekdays,the prospective candidates can be contacted through email,SMS.
- Attracting more references:
  - After a detailed brainstorming of the options,various referral offers should be rolled out to the current students as past data shows that references have played a key role in more leads getting converted.
- Increasing visitor count to the website:
  - Effectively promote courses and the X educational company to attract more visitors to the website as Visits has a positive impact on the lead conversion.
- Improving website content:
  - As we see that the Time spent on the website and Total Visits are key drivers in a lead getting converted,the website content can be improved with the latest trends in the market,providing crisp and to-the-point information(keeping in mind the busyness of the working professionals who are highly converted).

Final Analysis - Recommendations

- Email content:
  - As Last Activity_Email Opened is positively correlated with the conversion,the email content can be improved to attract more enrollment. There have been successful usecases where simple changes in Email content like below can have a significant impact.
    - Addressing the receive by their Name(based on the details provided by them on the website).
    - Customizing content based on their profiles(for instance,different content for students and working professionals and so on).
- Business communications training:
  - As we see Olark Chat is a key source for attracting prospective students,but the conversations on Olark chat have less contribution to the lead conversion.Hence, a formal business communications training to the sales team especially using Olark Chat can result in more lead conversion rate.