LEAD SCORING CASE STUDY

SUBMITTED BY :

1. **Prashant Pal**
2. **Pallavi Singh**
3. **Parth Kanetkar**

CONTENTS

- Problem statement
- Business Objective
- Problem approach
- EDA and Visualisation
- Correlations
- Model Evaluation
- Observations
- Conclusion

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. They have process of form filling on their website after which the company that individual as a lead.
- 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, this means if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as Hot Leads.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone

BUSINESS OBJECTIVE

- Lead X Education wants us to build a model to give every lead a lead score between 0 -100 . So that they can identify the Hot leads and increase their conversion rate as well.
- The CEO want to achieve a lead conversion rate of 80%.
- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

Problem Approach

- Importing the data and inspecting the data frame.
- Data preparation
- EDA
- Dummy variable creation
- Test-Train split
- Feature scaling
- Correlations
- Model Building (RFE R-squared VIF and p-values)
- Model Evaluation
- Making predictions on test set

EDA – DATA CLEANING

- There are 17 columns that has missing values and out of these 17 columns, 5 columns has more than 4000 missing values and 1 has more than 3000 missing values. As the total number of data values is more than 9000, so dropped Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score, Asymmetrique Profile Score and Lead Quality.

- There are now 4 categories, Lead Source, TotalVisits, Page Views Per Visit and Last Activity, which have missing values between 30-140, so we'll just fill in the missing value with mode value of the data. 'What is your current occupation' column cannot be ignored as the data is skewed to Unemployed people and these set of people are looking for career transition or Upgrading the skills and may look for the career opportunity.

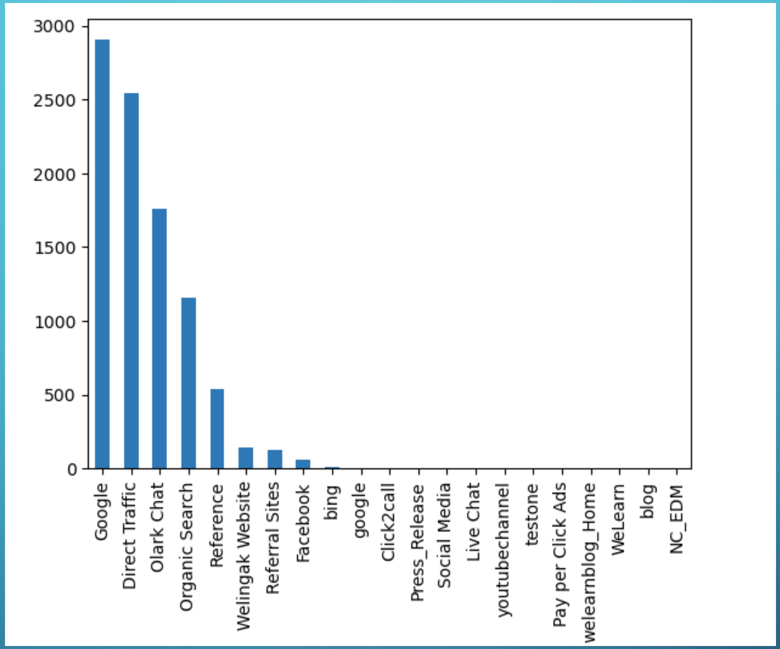
- We'll also fill in the 'Specialization' column with the mode value where ever there is a missing value and also we will combine the management related Specialization under one named as Management Specialization column.

- Looking at the Country column, we understood that 95% of the values in the data is coming from India and remaining is from different country. Along with that, City column shows that 41% data is collected from Mumbai region and remaining from other Indian regions and 28% is not selected as well. So combined the Mumbai and Thane & Outskirts together and will consider as 1 and anything outside from this region we'll consider as 0.

- 'What matters most to you in choosing a course' column can be ignored for now as there are only 3 categories and Better Career Prospects is highly skewed. So if 99.95% of the people has the same need so, no need to consider it in the model building.

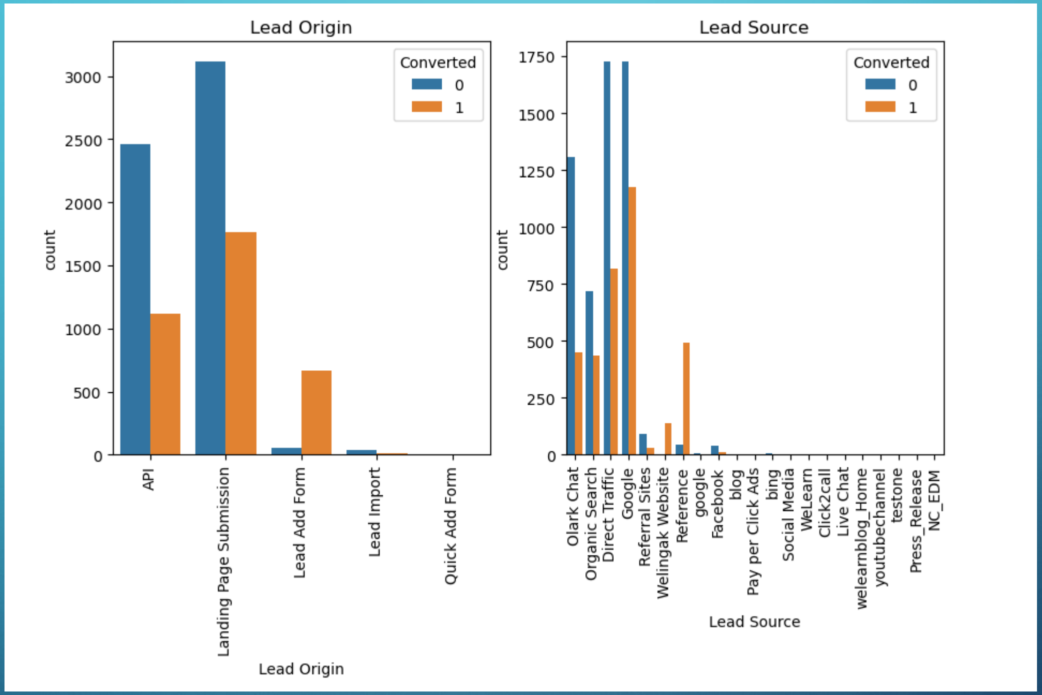
DATA VISUALIZATION

- **Lead Score:** This column displays that the maximum number of possible leads can be converted. Google, Direct Traffic and Olark Chat are some of them.



DATA VISUALIZATION

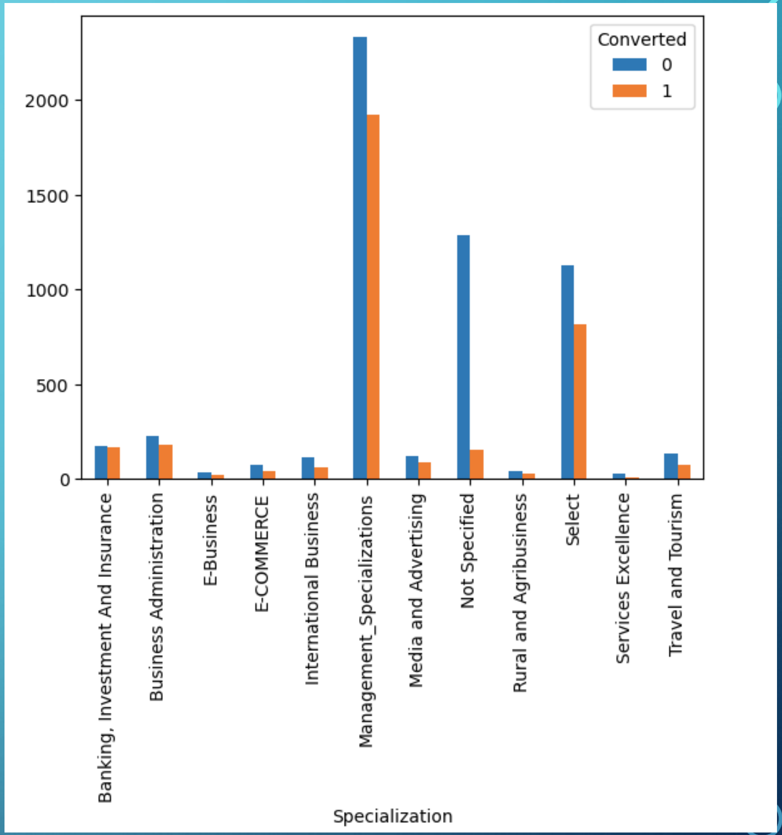
- **Lead Source vs Lead Origin:** If we compare these two columns with Converted column, we come to know that there are few values which displays difference between failed and successful leads converted.



DATA VISUALIZATION

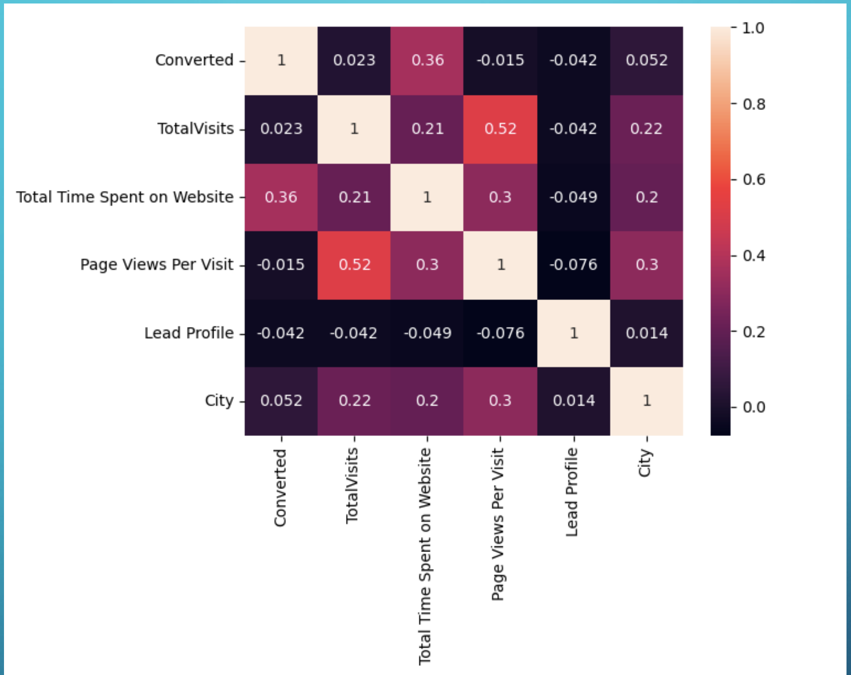
- **Specialization:** If we compare the specialization column with Converted column, we get to know that majority of Management specialization personnel leads are either getting converted or else failed.

- But there are few others who have not selected any specialization but also has a good number of conversion rate.



CORRELATIONS

- Converted column is highly corelated with Total Time Spend on



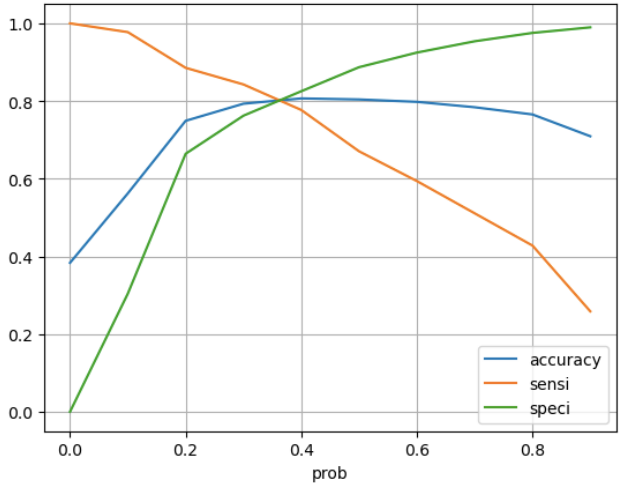
	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Profile	City
Converted	1	0.023	0.36	-0.015	-0.042	0.052
TotalVisits	0.023	1	0.21	0.52	-0.042	0.22
Total Time Spent on Website	0.36	0.21	1	0.3	-0.049	0.2
Page Views Per Visit	-0.015	0.52	0.3	1	-0.076	0.3
Lead Profile	-0.042	-0.042	-0.049	-0.076	1	0.014
City	0.052	0.22	0.2	0.3	0.014	1

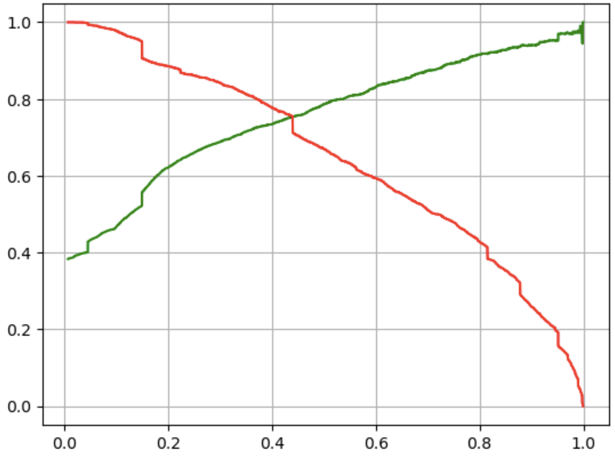
MODEL EVALUATION

ROC curve

0.42 is the tradeoff between Precision and Recall -

Thus we can safely choose to consider any Prospect Lead with Conversion **Probability higher than 42 % to be a hot Lead**





OBSERVATIONS

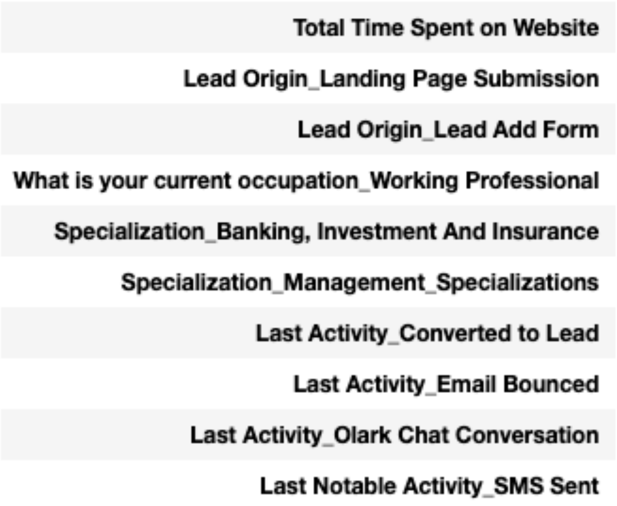
Train Data:

Accuracy : 79.31%
Sensitivity : 84.27%
Specificity : 76.22%

Test Data:

Accuracy : 79.07%
Sensitivity : 82,42%
Specificity : 76.94%

Final Features list:



Total Time Spent on Website
Lead Origin_Landing Page Submission
Lead Origin_Lead Add Form
What is your current occupation_Working Professional
Specialization_Banking, Investment And Insurance
Specialization_Management_Specializations
Last Activity_Converted to Lead
Last Activity_Email Bounced
Last Activity_Olark Chat Conversation
Last Notable Activity_SMS Sent

CONCLUSION

- We see that the conversion rate is 30-35% (close to average) for API and Landing page submission. But very low for Lead Add form and Lead import. Therefore we can intervene that we need to focus more on the leads originated from API and Landing page submission.
- We see max number of leads are generated by google / direct traffic. Max conversion ratio is by reference and welingak website.
- Leads who spent more time on website, more likely to convert.
- Max conversion with working professional and people with Management Specialization.