LEAD SCORING CASE STUDY

SUBMITTED BY:

ROSHANI SRIVASTAVA

AKANSHA SINGH

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 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. For example, 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. A typical lead conversion process can be represented using the following funnel:

- As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.
- 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 higher lead score have a higher conversion chance and the customers with 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%.

GOALS AND OBJECTIVES

There are quite a few goals for this case study.

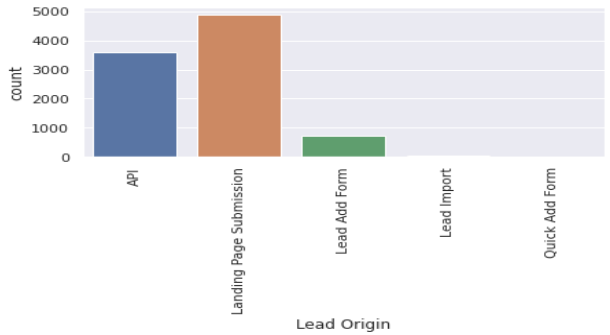
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

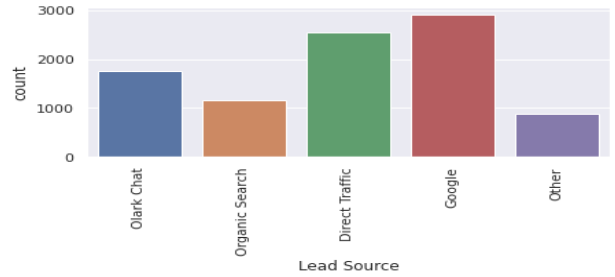
RESULTS EXPECTED

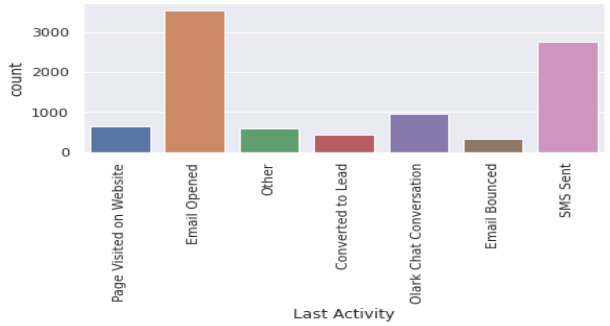
- A well-commented Jupyter note with at least the logistic regression model, the conversion predictions and evaluation metrics.
- The word document filled with solutions to all the problems.
- The overall approach of the analysis in a presentation
- Mention the problem statement and the analysis approach briefly
- Explain the results in business terms
- Include visualisations and summarise the most important results in the presentation
- A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

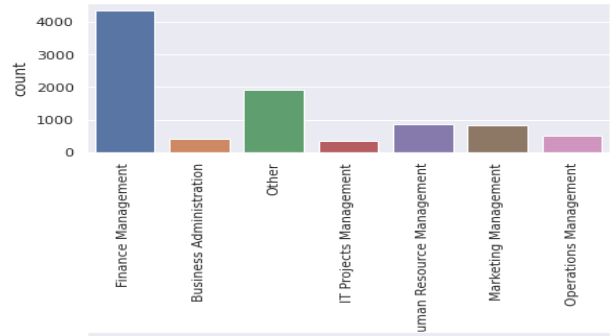
EDA

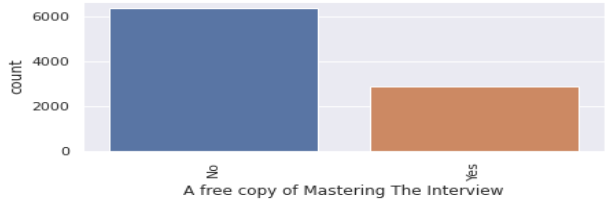
- In Lead Source we can see that Direct Traffic and Google are the two main source for Leads
- Email Opened and SMS Sent in Last Activity is high we can see
- Finance Management Specialization is the most chosen one in the specialization category

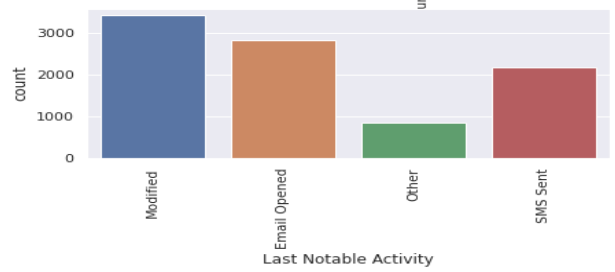




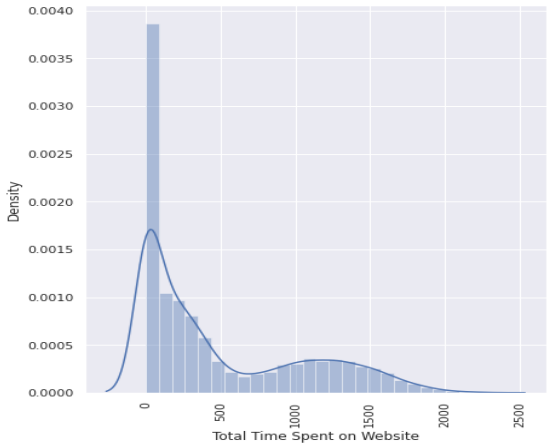


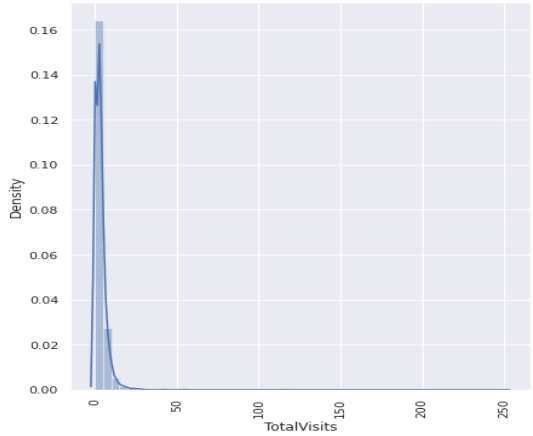


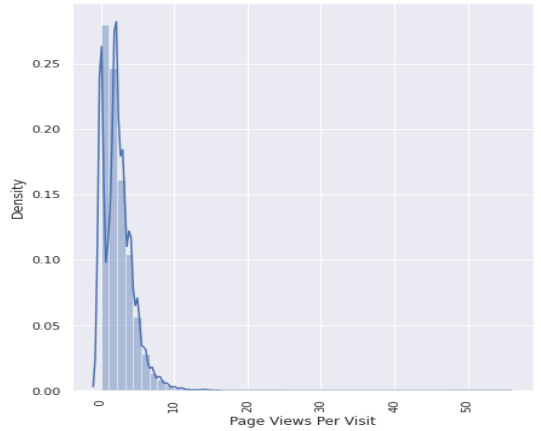


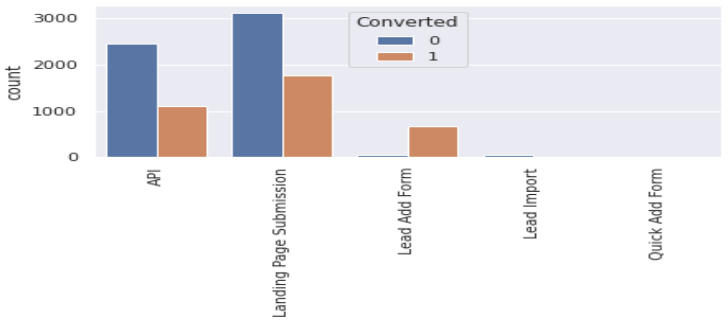


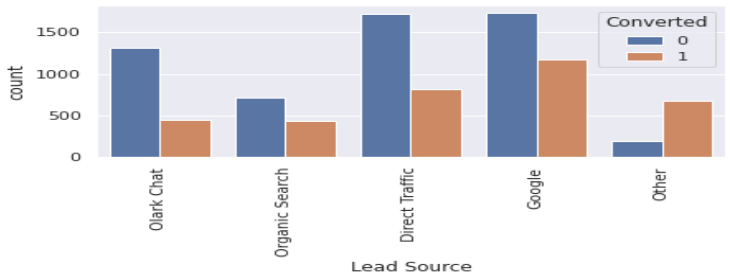
- Continuous Variables are not in Normal distribution
- Outliers are there in Total Visits and Page Views Per Visit
- We can see that the total visits have more values is between 0-50 and page views per visits 0-20



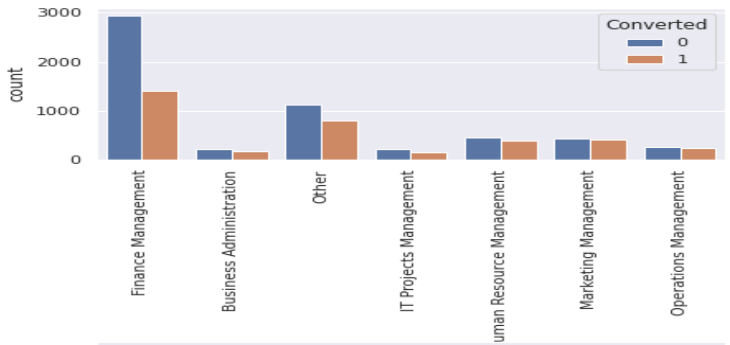


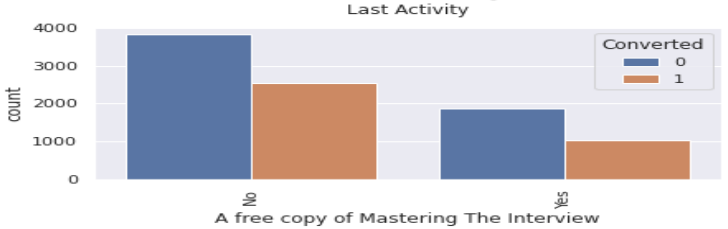


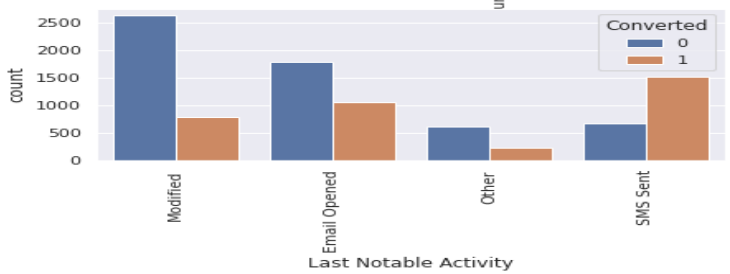

Converted
0
1
count
3000
2000
1000
0
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin


Converted
0
1
count
1500
1000
500
0
Olark Chat
Organic Search
Direct Traffic
Google
Other
Lead Source


Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Other
Converted to Lead
Olark Chat Conversation
Email Bounced
SMS Sent
Last Activity


Converted
0
1
count
3000
2000
1000
0
Finance Management
Business Administration
Other
IT Projects Management
uman Resource Management
Marketing Management
Operations Management


Converted
0
1
count
4000
3000
2000
1000
0
No
Yes
A free copy of Mastering The Interview


Converted
0
1
count
2500
2000
1500
1000
500
0
Modified
Email Opened
Other
SMS Sent
Last Notable Activity

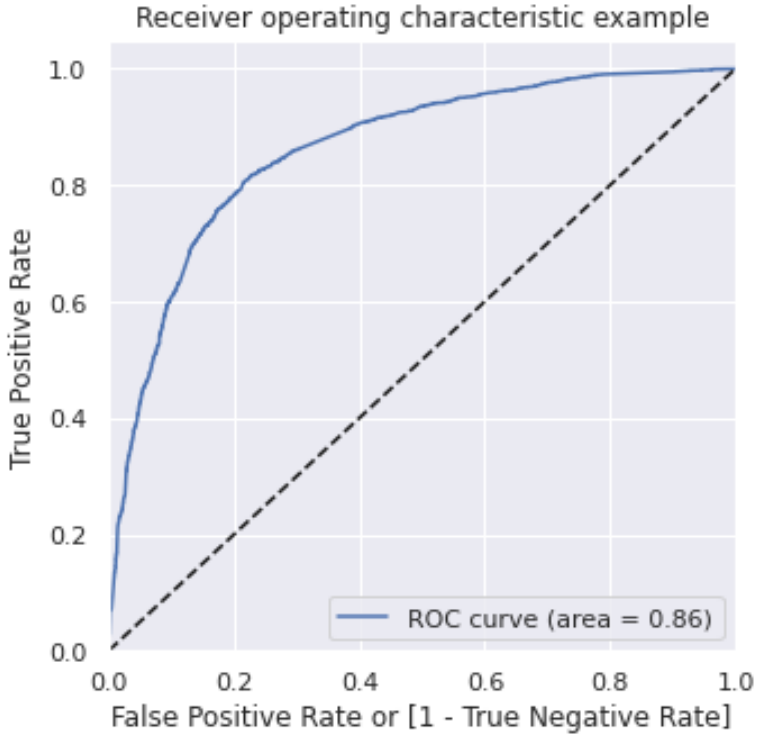
DATA CONVERSION

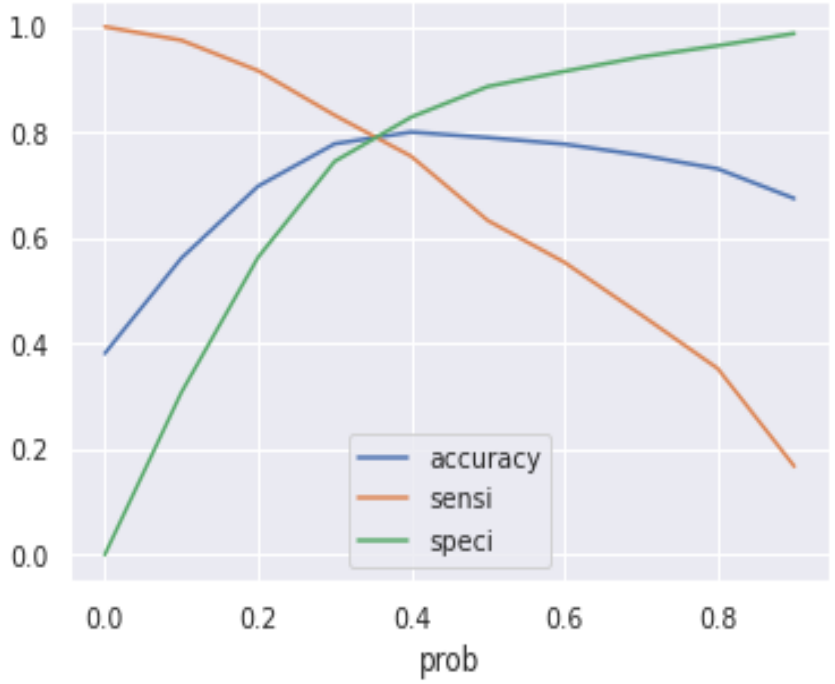
- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5
- Predictions on test data set
- Overall accuracy 81%

ROC CURVE

- Finding Optimal Cut off Point
- Optimal cut off probability is that
- Probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.





CONCLUSION

Below are the points via which we can a get a understanding between a hot lead and a cold lead

- The total time spend on the Website.
- Total number of visits.
- When the lead source was from the below sites:
  a. Google
  b. Direct traffic
  c. Organic search
  d. Welingak website
- When the lead origin is Lead add format
- When their current occupation is as a working professional.