Lead Scoring Case Study

(Logistic Regression)

1. Jeyashree Kothai

2. Akash Agarwalla

Background

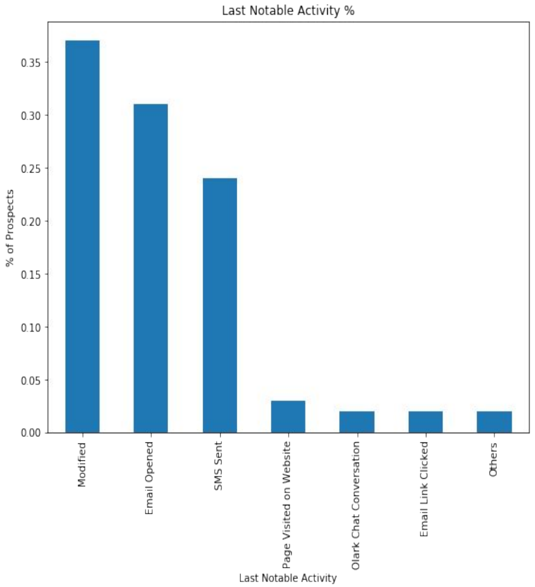
- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these 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.
- X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein we 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%.

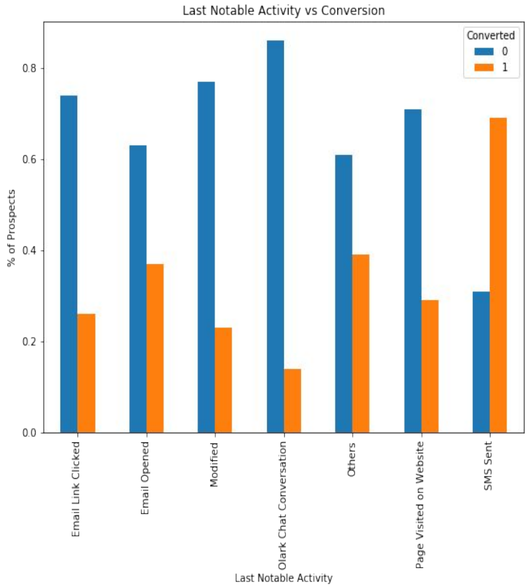
EDA VISUALISATIONS

1. Exploratory Data Analysis – Categorical Variables
2. Exploratory Data Analysis – Numerical Variables

Exploratory Data Analysis – Categorical Variables

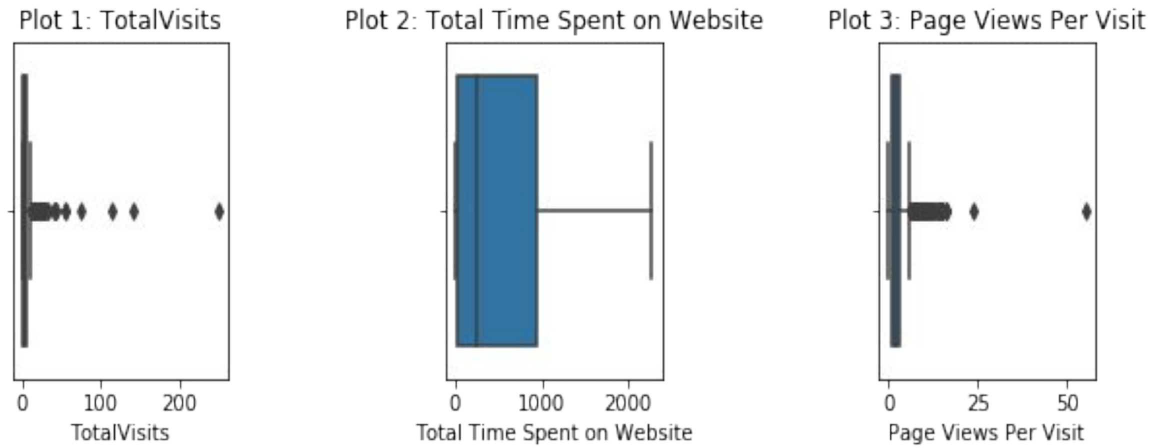
- ‘Select’ value is replaced by null values.
- Missing values check performed. If less than 40% then data imputation is done by mode of that categorical
- For some categorical variables having low frequency in the column , values are clubbed and categorized as other category.
- Univariate and bivariate analysis is performed to understand impact on lead conversion.
- Example – Lead source categorical variable shown in the graph.





Exploratory Data Analysis – Numerical Variables

- Missing values above 40 % in columns have been dropped.
- Corelation Map is checked for corelation and high corelation values are handled later.
- Outliers checked as shown below and handled by capping at IQR*3 range where IQR stands for Inter Quartile range. this reduces data loss crucial for analysis.

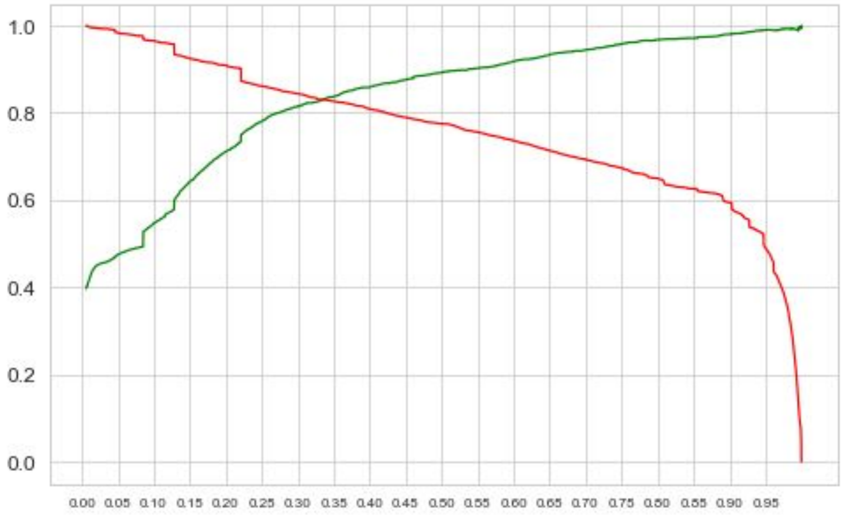


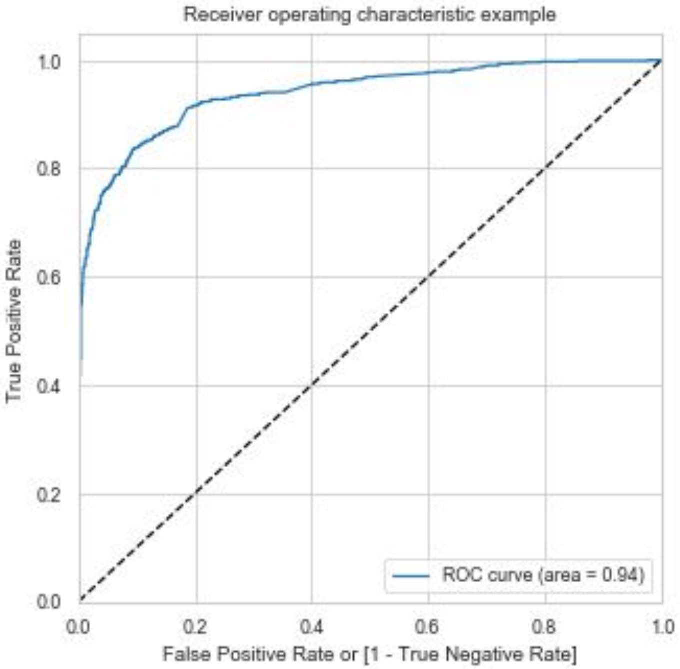
RESULTS IN BUSINESS TERMS

Model Summary

- Our final model gave us the following performance metrics

	Overall Model Accuracy	Precision	Recall/Sensitivity	Specificity	AUC
Train	87.2	83.4	82.9	89.9	0.93
test	87.8	84.6	83.6	90.4	0.94





- We used Recall/Precision trade-off graph to derive the optimal threshold value.
- 84% of recall value indicates that our model is able to predict 84% od actual conversion cases correctly.
- 85% of precision value indicates that 85% of the conversions that our model predicted is actually converted.

Lift and Gain chart

- Lead Scoring Decile: we divided the lead scores into 10 Deciles,after sorting them in descending order of their probability scores. This allowed us to have the top lead score targets in decile1 and the next best probable ones in decile 2 and so on.

- This helped us to maximise the lead conversion by choosing the decile with maximum no. of lead score.

- This can be used as focused conversion techniques by sales team.

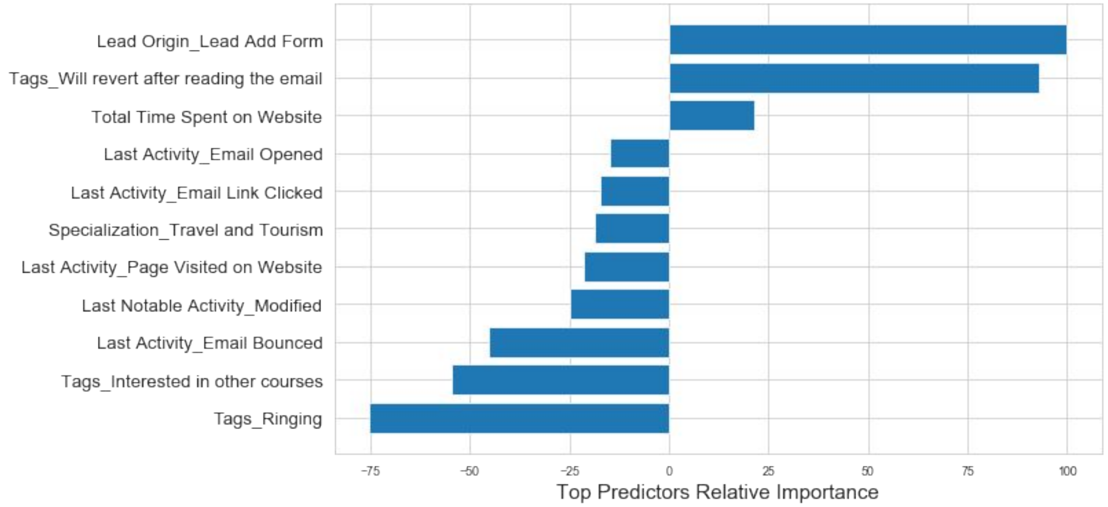
decile	gain	gain_percentage	lift
1	180	25.60	2.56
2	180	51.20	2.56
3	149	72.40	2.41
4	88	84.92	2.12
5	55	92.74	1.85
6	12	94.45	1.57
7	20	97.29	1.38
8	9	98.57	1.23
9	8	99.71	1.10
10	2	100	1.00

CONCLUSION

Top 3 features contributing to Lead conversions

Top 3 variables which contribute most towards the probability of a lead getting converted.

- Lead origin_Lead Add Form
- Tags_will revert after reading the email
- Tags_Ringing


Lead Origin_Lead Add Form
Tags_Will revert after reading the email
Total Time Spent on Website
Last Activity_Email Opened
Last Activity_Email Link Clicked
Specialization_Travel and Tourism
Last Activity_Page Visited on Website
Last Notable Activity_Modified
Last Activity_Email Bounced
Tags_Interested in other courses
Tags_Ringing
Top Predictors Relative Importance