Lead Score Case Study

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.
- 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.

Business Goal:

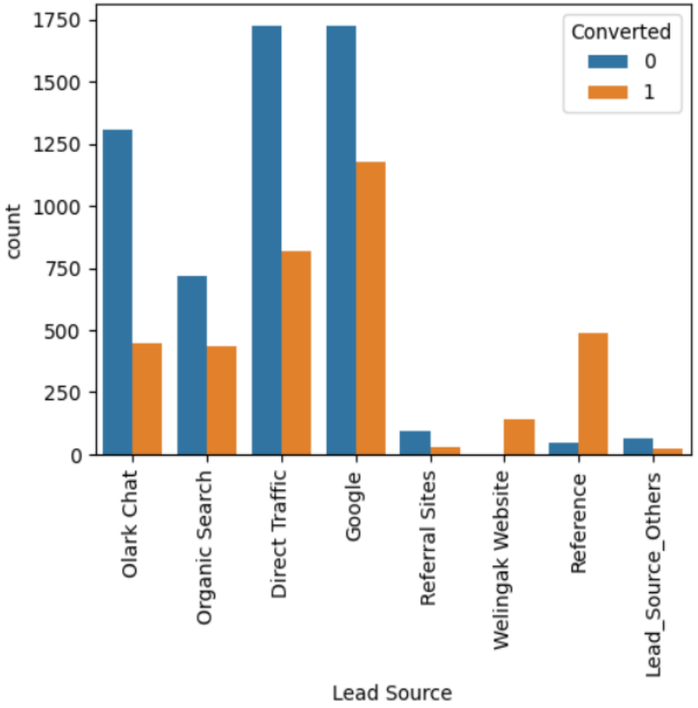
- X Education needs help is selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires a model wherein it needs to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance 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%.

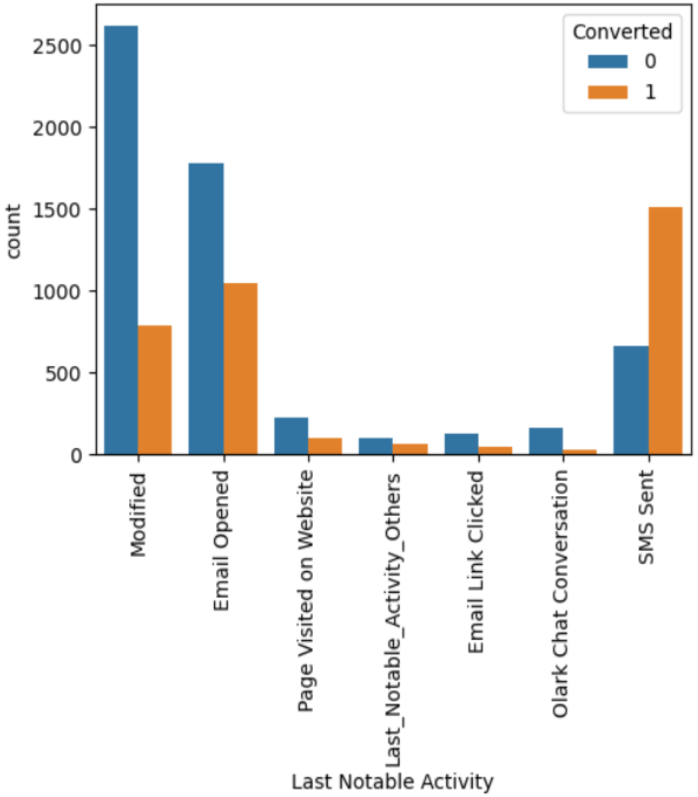
Strategy

- Data cleaning – removing and imputing the missing values wherever necessary.
- Data preparation – grouping the data and creating dummy variables on categorical data .
- EDA – using visualization tools to get insights on data.
- Model Building – used both automatic and manual feature selection based on P and VIF values.
- Prediction on train data – using the model to predict the outcome.
- Calculating metrics – used to determine the predictive power of the model.
- Prediction on test data – once satisfied with metrics, used the model to predict the test data.

EDA – analysis of the categorical columns

Lead Score and Last Notable Activity



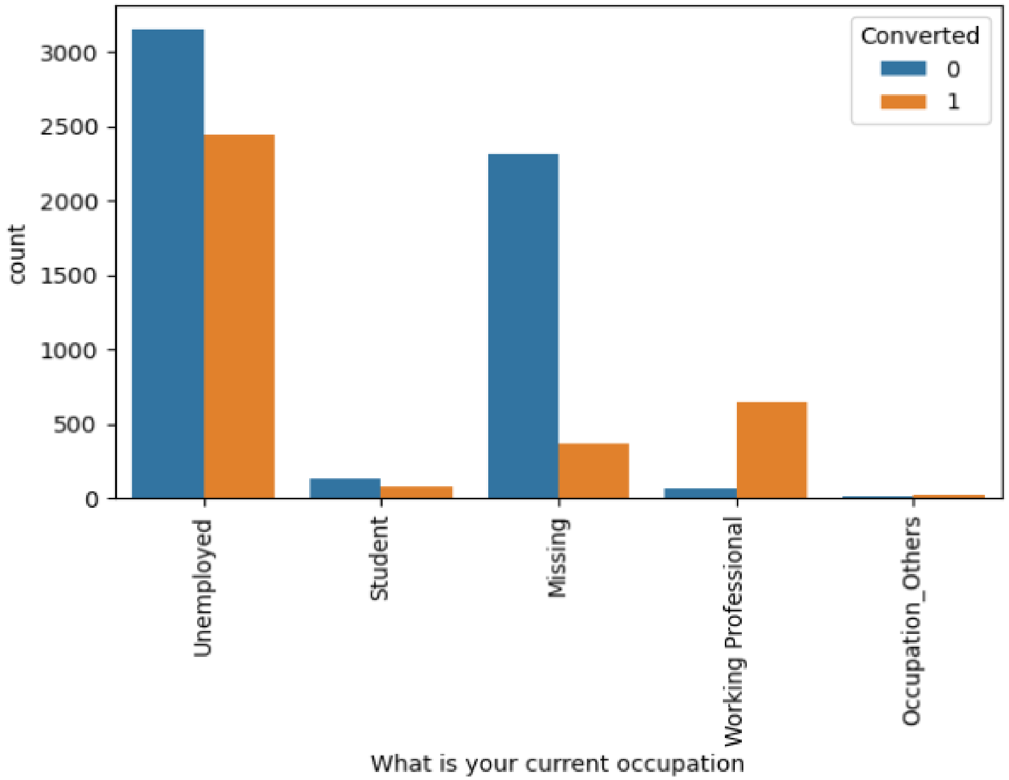


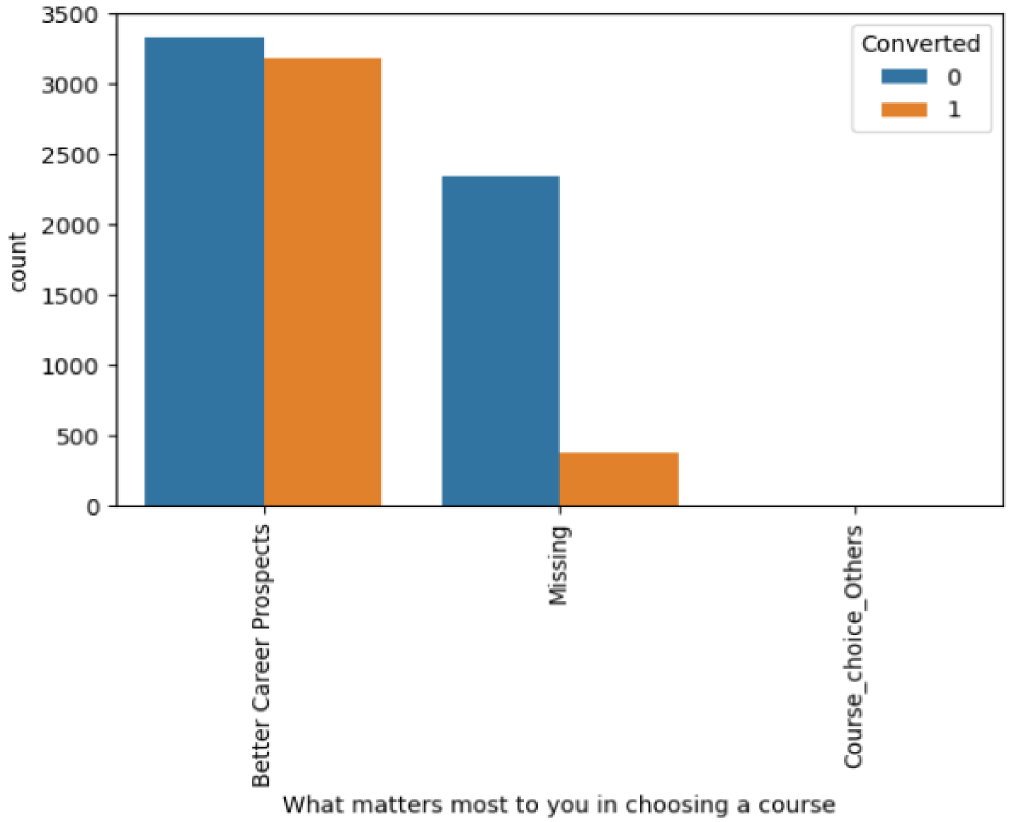
Observation:

1.In Lead Source, Direct traffic and Google has highest weightage, But reference has the highest conversions rate
2.SMS Sent has the good conversion rate, while modified and email opened has higher count value
3.Lead Import has very less count as well as conversion rate and hence can be ignored

EDA – analysis of the categorical columns

What is your current occupation



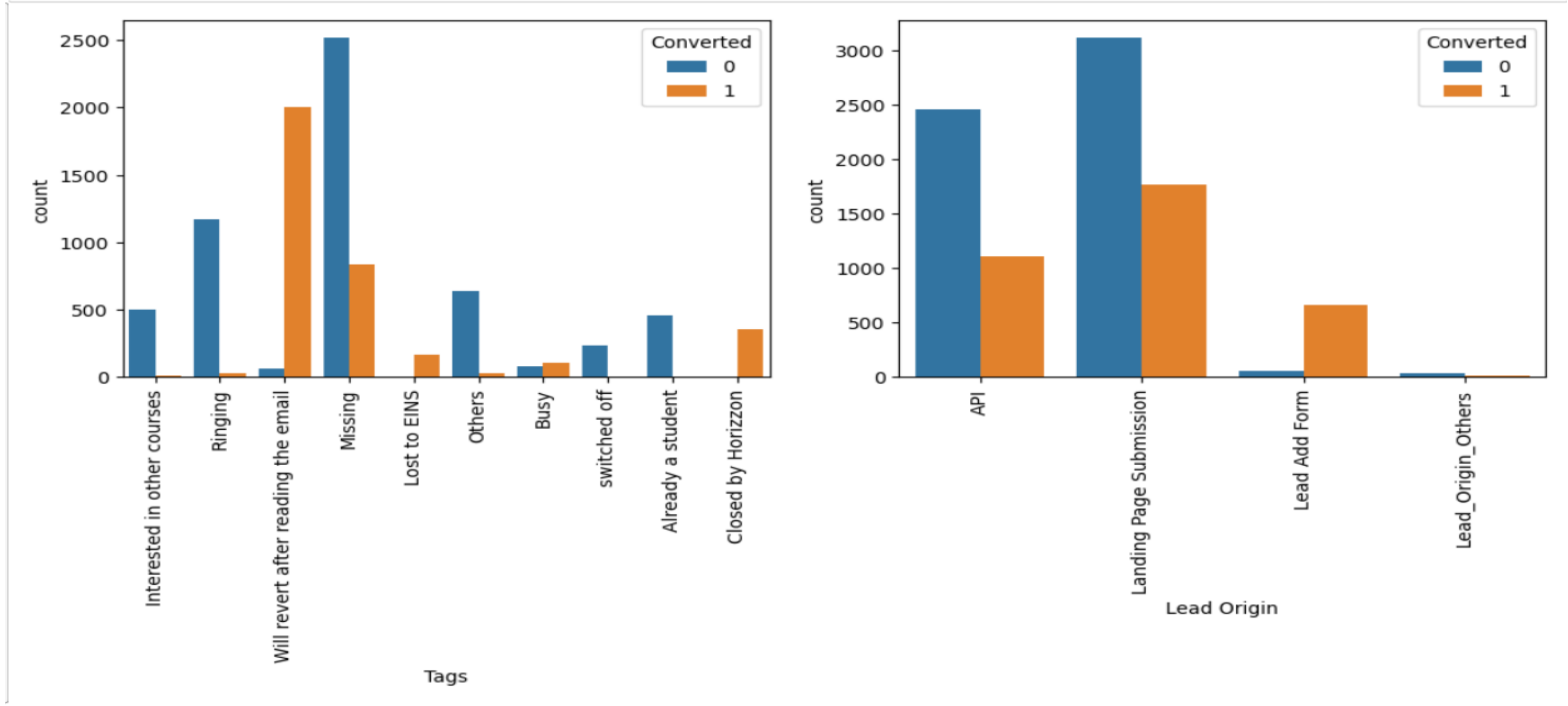


Observations:

1.Unemployed candidates have highest count, while working professionals have the highest conversion rate
2.Better career prospects have both higher weightage ~50% conversion rate
3.Mumbai has higher count and Thane and outskirts has better conversion rate

EDA – analysis of the categorical columns

Tags and Lead Origin



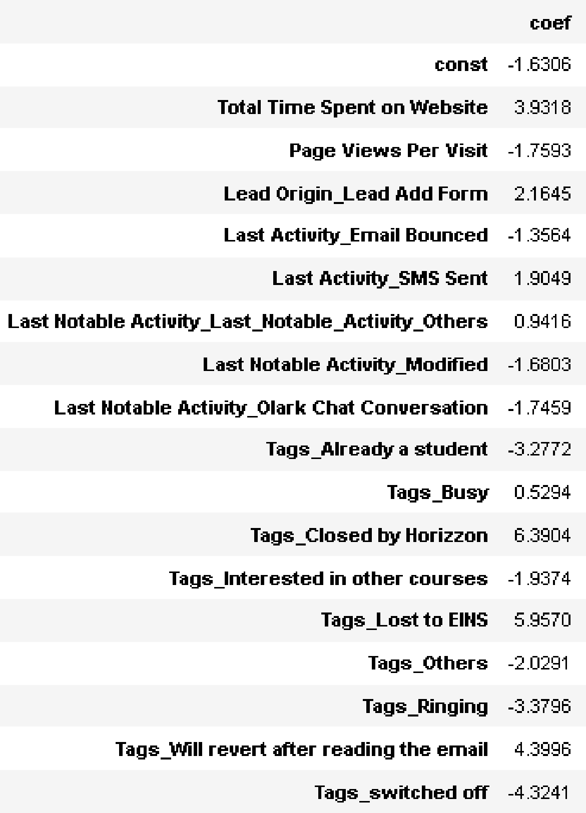
Observation:

1) revert after reading the email has extremely high conversion rate
2) In Lead Origin, API and Landing Page Submission has highest weightage in terms of counts, Lead Add Form has the highest conversion rate

- Feature scaling :

- **Using the Recursive Feature Elimination we went ahead and selected the 20 top important features. Using the statistics generated, we recursively tried looking at the P-value and VIF value in order to select the most significant values that should be present and dropped the insignificant variables.**
- **we choose these values as cutoff and removed the variables/features one by one, Cutoff Value for P=0.05 and Cutoff Value for VIF = 5.**

Final variables list :

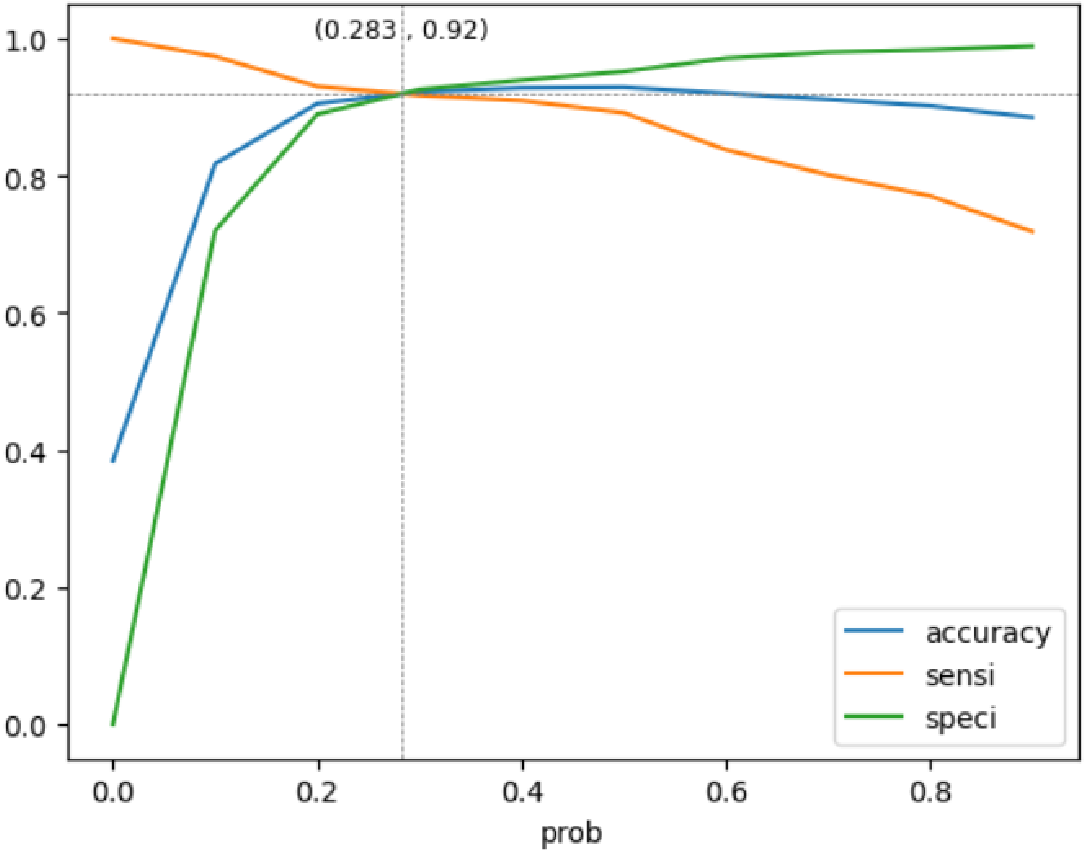


	coef
const	-1.6306
Total Time Spent on Website	3.9318
Page Views Per Visit	-1.7593
Lead Origin_Lead Add Form	2.1645
Last Activity_Email Bounced	-1.3564
Last Activity_SMS Sent	1.9049
Last Notable Activity_Last_Notable_Activity_Others	0.9416
Last Notable Activity_Modified	-1.6803
Last Notable Activity_Olark Chat Conversation	-1.7459
Tags_Already a student	-3.2772
Tags_Busy	0.5294
Tags_Closed by Horizzon	6.3904
Tags_Interested in other courses	-1.9374
Tags_Lost to EINS	5.9570
Tags_Others	-2.0291
Tags_Ringing	-3.3796
Tags_Will revert after reading the email	4.3996
Tags_switched off	-4.3241

```
[[3679 305] [ 204 2280
```

Model evaluation

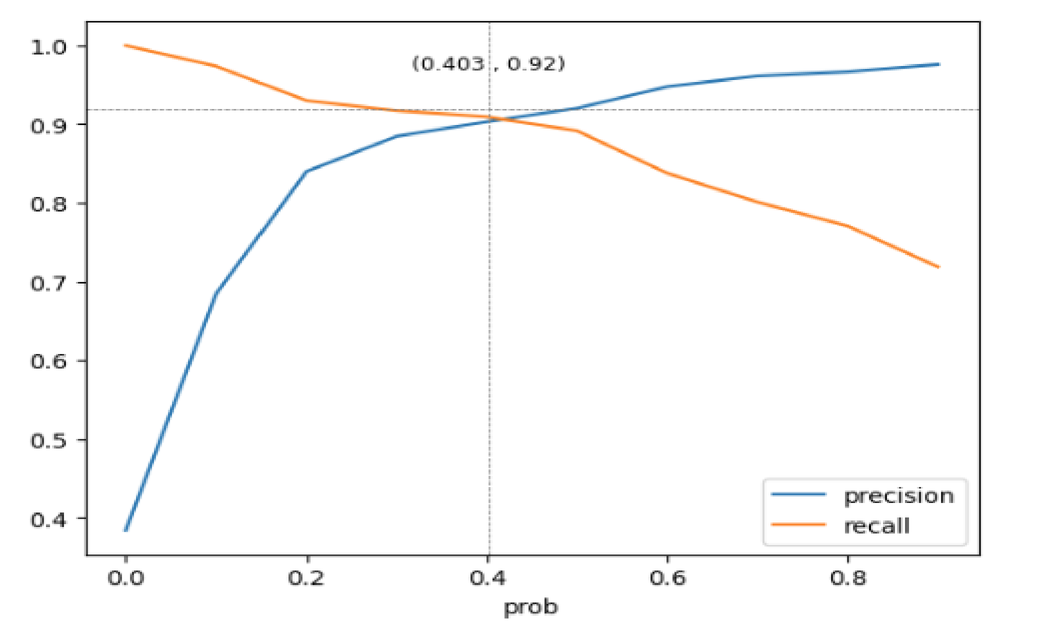
- Specificity and sensitivity:



- Accuracy - 92%
- Sensitivity - 91%
- Specificity - 92 %

Model evaluation

- Precision and Recall:



- Accuracy - 92%
- Precision - 90%
- Recall - 90 %

For the given problem statement sensitivity/Recall are important, hence we choose 0.283 cut off, since that's the maximum sensitivity/recall.

Test data Specificity and sensitivity

Following are the metrics obtained on test data :

Sensitivity/Recall : 0.9239

Specificity : 0.9192

precision : 0.879

False-Positive-Rate : 0.0808

We observed that there are no much deviations between train metrics and test metrics , this model can be accepted for prediction .

Conclusion

1. we have chosen Specificity and sensitivity over precision and precall

2. Also the lead score calculator shows the conversion rate on the final prediction model of test data is 92%.

3.We have top 3 variables for the X company to focus on:

- Tags
- Last notable activity
- Total time spent on the website