LEAD SCORE CASE STUDY

Submitted By –

Tanmay Thapliyal

Mathangi S

Shahul Hameed

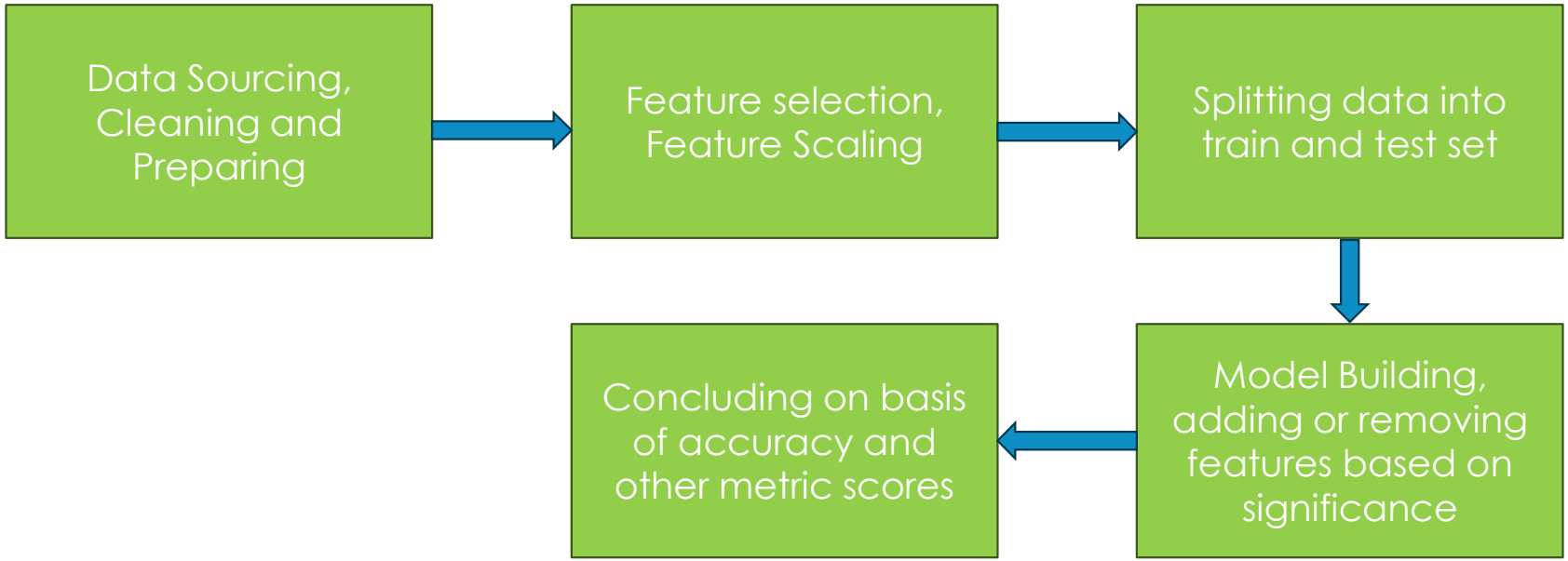
PROBLEM STATEMENT

- 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%.

BUSINESS GOAL

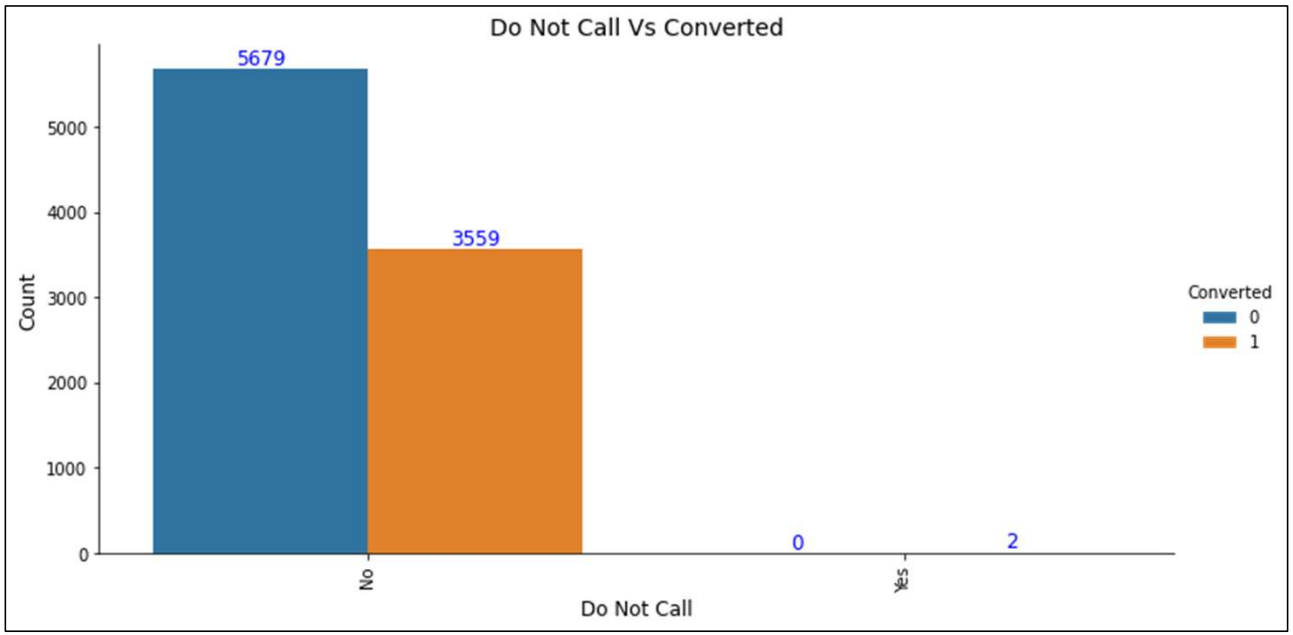
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will 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.

FLOW ON HOW TO SOLVE THE PROBLEM



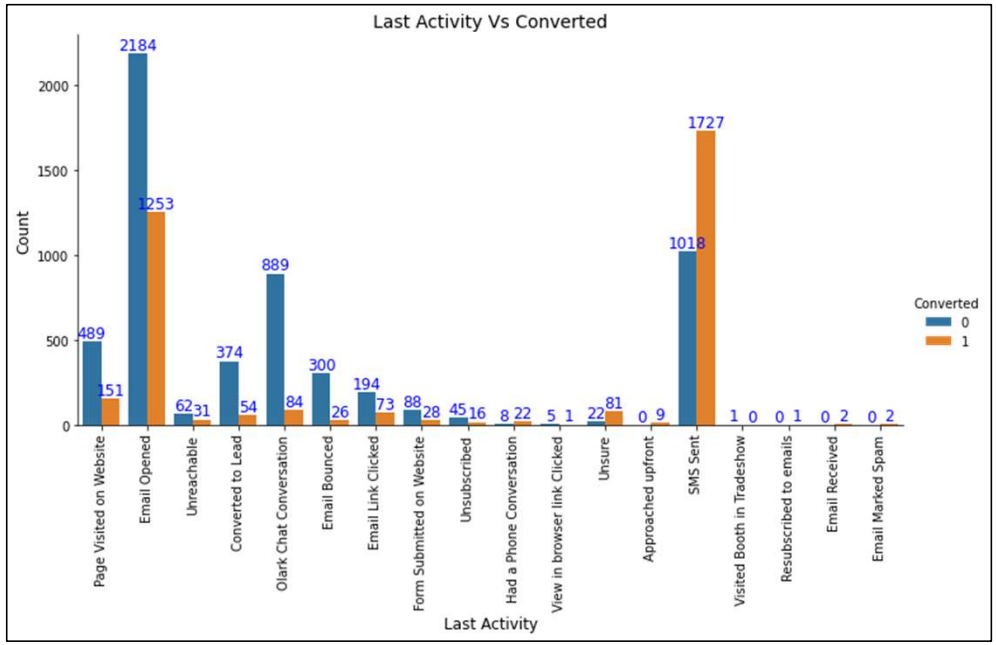
EXPLORING DATA

- For Do Not Call, major conversion happened when calls were made, but 2 leads got converted when do not call was opted too.



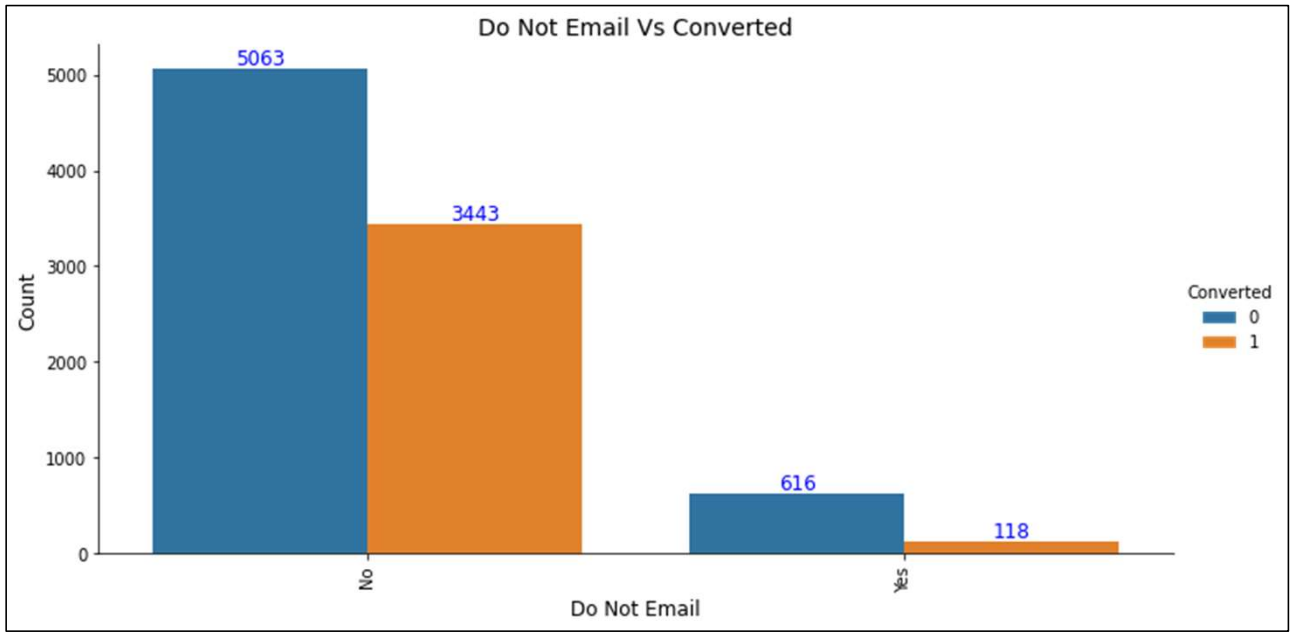
EXPLORING DATA

- For Last Activity, most conversions happened when sms was sent.



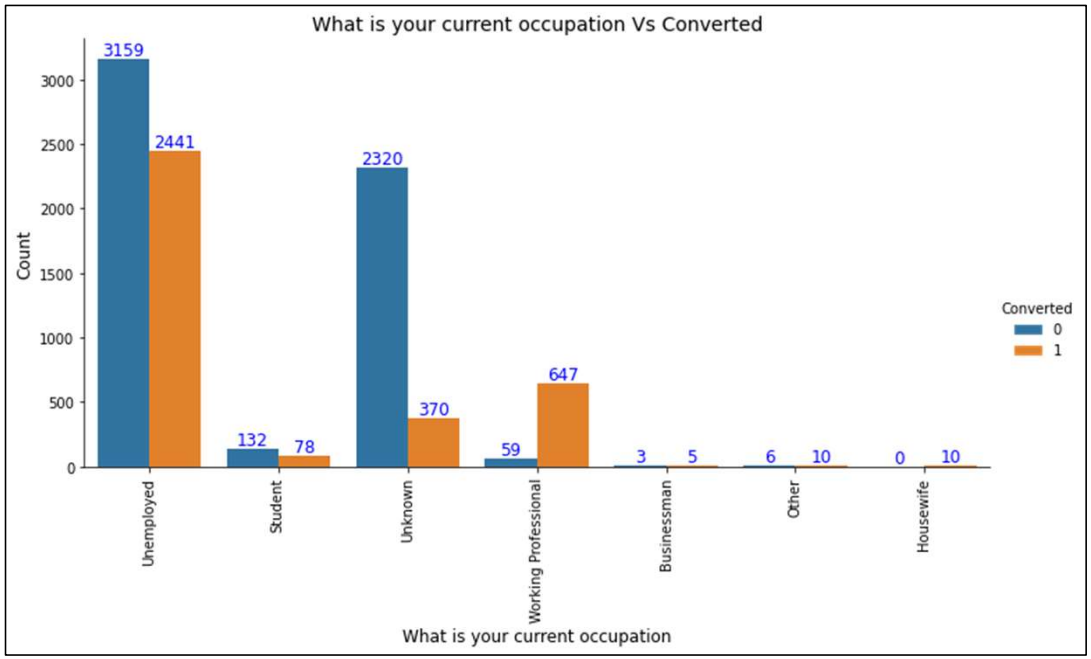
EXPLORING DATA

- For Do Not Email, most conversions seen when email is sent.



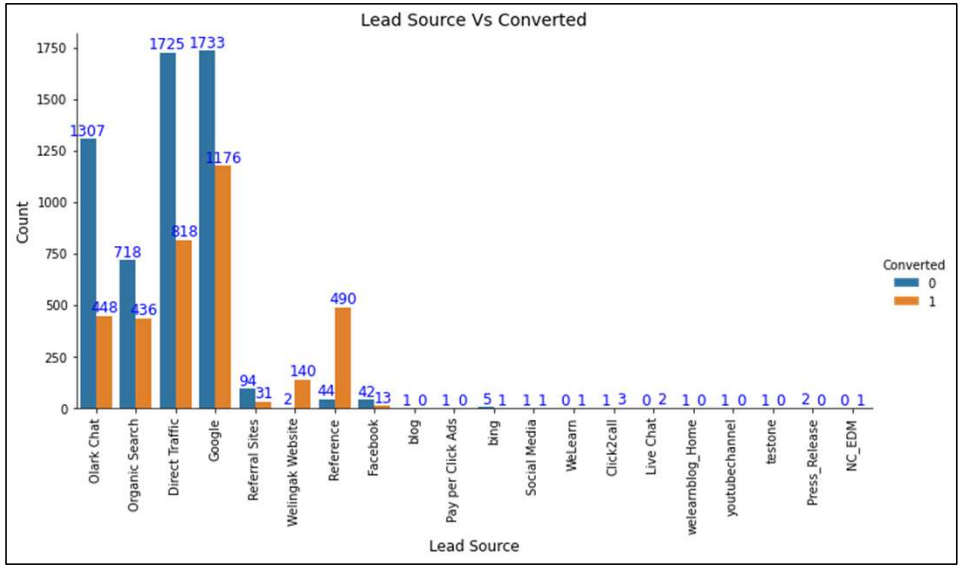
EXPLORING DATA

- Most Unemployeed had positive conversion rate.



EXPLORING DATA

- From Google as Lead Source, most conversions happened.



FEATURE SELECTION

The final features selected for model are below -

- 'Total Time Spent on Website'
- 'LastNotableActivity - Modified'
- 'LeadOrigin - API'
- 'LeadOrigin - Lead Add Form'
- 'LastActivity - Olark Chat Conversation'
- 'LastActivity - SMS Sent'
- 'CurrentOccupation - Unknown'
- 'CurrentOccupation - Working Professional'

FEATURE SELECTION

Top 3 features selected for model are below -

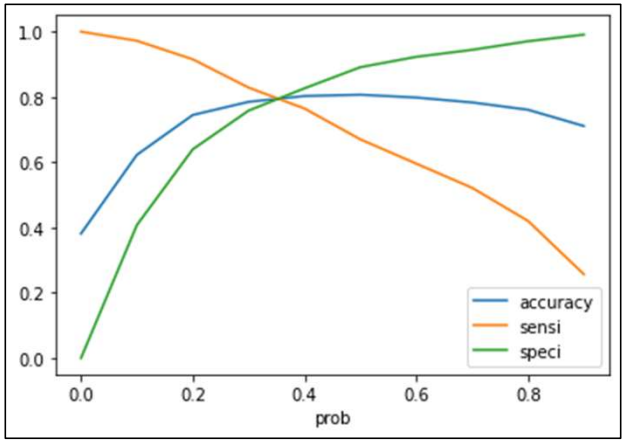
- Total Time Spent on Website
- Lead Add Form (Lead Origin)
- Working Professional (Current Occupation)

MODEL SUMMARY

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6459
Model Family:	Binomial	**Df Model:**	8
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2702.2
Date:	Mon, 17 Jul 2023	**Deviance:**	5404.3
Time:	18:26:24	**Pearson chi2:**	6.63e+03
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.3897
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9324	0.076	-25.435	0.000	-2.081	-1.783
Total Time Spent on Website	4.0544	0.150	26.983	0.000	3.760	4.349
LastNotableActivityD_Modified	-0.8949	0.079	-11.261	0.000	-1.051	-0.739
LeadOriginD_API	0.7522	0.077	9.831	0.000	0.602	0.902
LeadOriginD_Lead Add Form	3.6810	0.180	20.404	0.000	3.327	4.035
LastActivityD_Olark Chat Conversation	-0.5951	0.171	-3.482	0.000	-0.930	-0.260
LastActivityD_SMS Sent	1.1790	0.073	16.096	0.000	1.035	1.323
CurrentOccupationD_Unknown	-1.0617	0.086	-12.335	0.000	-1.230	-0.893
CurrentOccupationD_Working Professional	2.5578	0.186	13.757	0.000	2.193	2.922

MODEL EVALUATION – TRAINING SET




```
print("Sensitivity - ",TP/(TP+FN))
```

```
Sensitivity -  0.7789943227899432
```

```
print("Specificity - ",TN/(TN+FP))
```

```
Specificity -  0.8133433283358321
```

```
print(metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.Predicted_Values))
```

```
0.8002473716759431
```

accuracy of ~ 80% is pretty good

MODEL EVALUATION – TEST SET

```
print('precision ',precision_score(y_predicted_final.Converted, y_predicted_final.predicted_final))

# recall
print('recall ',recall_score(y_predicted_final.Converted, y_predicted_final.predicted_final))
```

```
precision  0.7450302506482281
recall  0.7872146118721461
```

```
print("Specificity - ",TN/(TN+FP))
```

```
Specificity -  0.8240906380441264
```

```
print("Sensitivity - ",TP/(TP+FN))
```

```
Sensitivity -  0.7872146118721461
```

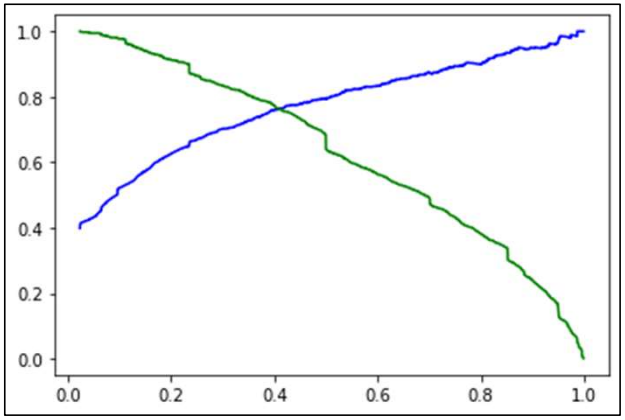
```
metrics.accuracy_score(y_predicted_final.Converted, y_predicted_final.predicted_final)
```

```
0.8095238095238095
```

accuracy of ~ 80% on test data is good.

MODEL EVALUATION – TEST SET

Precision Vs Recall



CONCLUSION

- we see accuracy of about 80% on test data and about 80% on train data.
- Specificity, Sensitivity of test data is 0.82 and 0.78 respectively.
- Specificity, Sensitivity of train data is 0.81 and 0.77 respectively.
- Precision score on test data is 0.74 and Recall score is 0.78.
- model seems to perform good on train and test set.
- Company should focus more on Total Time Spent on Website, Lead Add Form (Lead Origin), Working Professional (Current Occupation) features.
- The conversion probability cut off is 0.38.