Lead Score Case Study

- Student Name: Tran Hong Duc
- Email: ductran2306@gmail.com
- Class: DS C58 IIITB EPGDS July 2023

Introduction

- 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.
- Although X Education gets a lot of leads, its lead conversion rate is very poor. The company requires building a model to help them select the most promising leads.

Project Overview & Goal

- Current situation: The typical lead conversion rate at X education is around 30%.
- Goal:
  - The CEO want the target lead conversion rate to be around 80%.
  - Build a logistic regression model to assign a lead score between 0 and 100. A higher score would mean that the lead is hot.
  - The company's requirement can change in the future. The model should be able to adjust this this case.

Project Dataset

The dataset for this project include 01 file:

- 'Leads.csv' contains all the information of the past with around 9000 data points. The data is about whether **a lead has converted or not.**
- The target variable is "Converted" which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

Model Building Steps

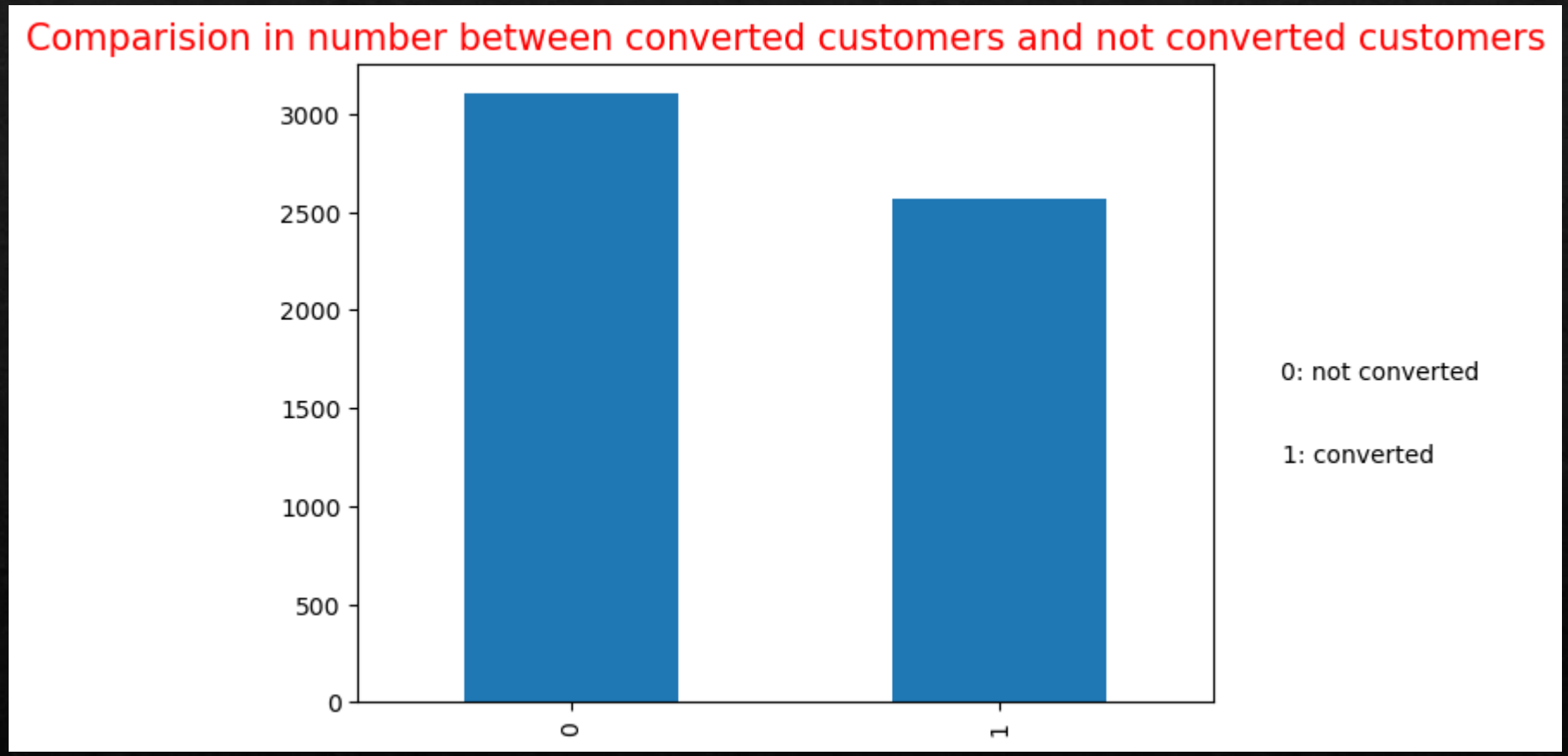
- Step 1: Import data from “Leads.csv”
- *Step 2: Clean data including: drop columns, handle missing values, outliers, duplicated rows.*
- *Step 3: EDA*
- *Step 4: Data Preparation*
- *Step 5: Building model*

All of the step can be reproduce in notebook.

Warning in Data Cleaning

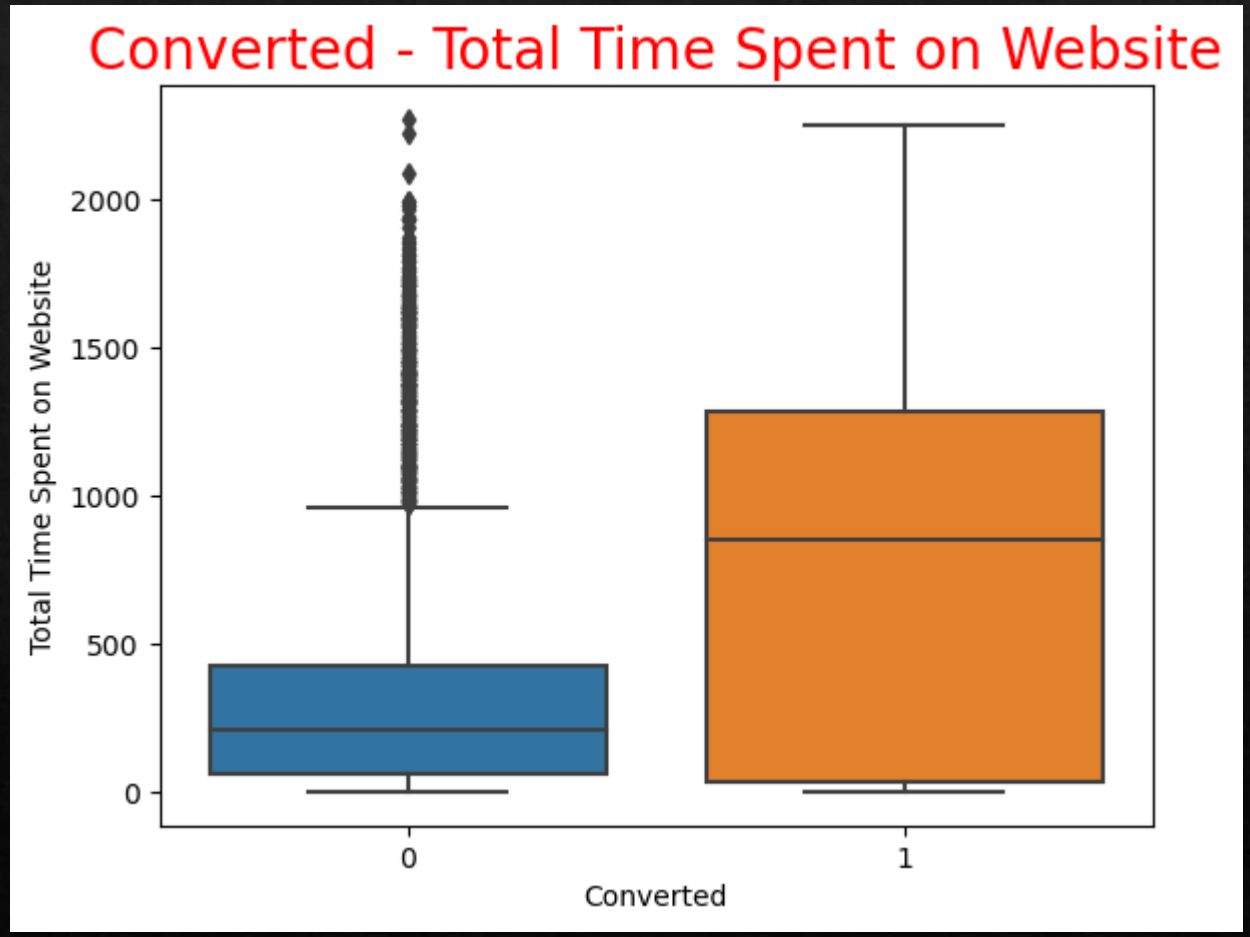
- There are many variables that have null data.
- *Most of these variables are handled by **removing whole variable or just null data points**. This is because we don't know the cause of these null value. Filling these null values makes our model unreliable.*
- *After cleaning data, we have total 5666 records.*

Converted & not Converted Customers

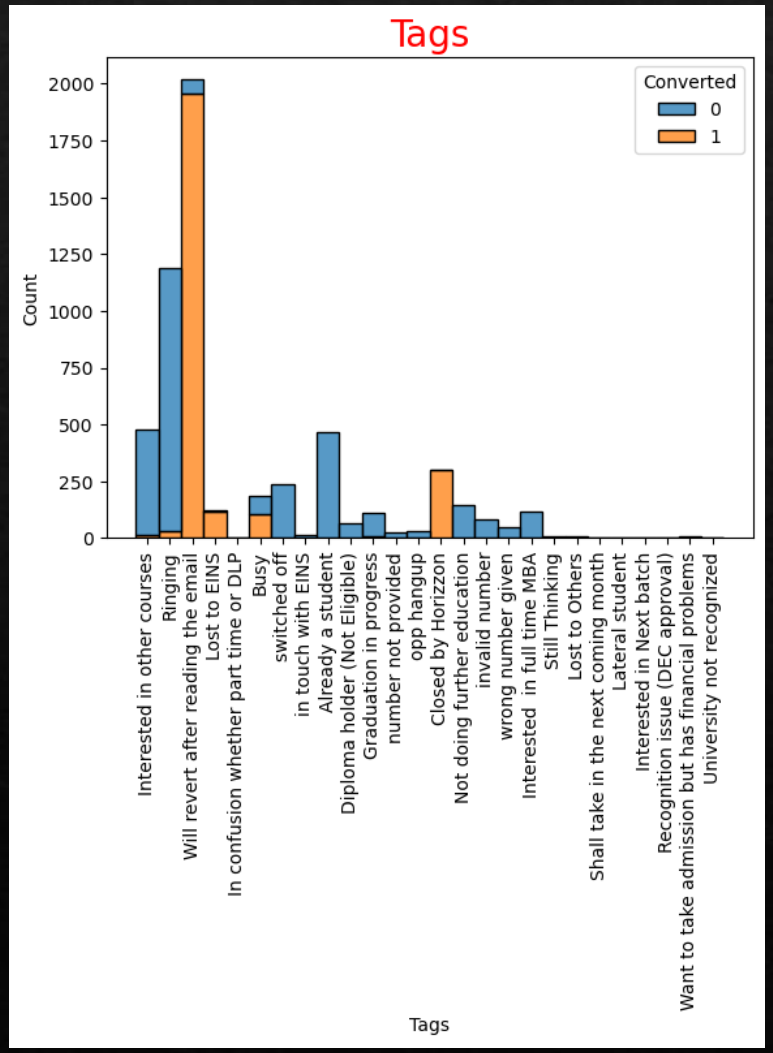


Number of converted customers is 3100 while Number of not converted customers is 2600

Converted customers spent more time on website



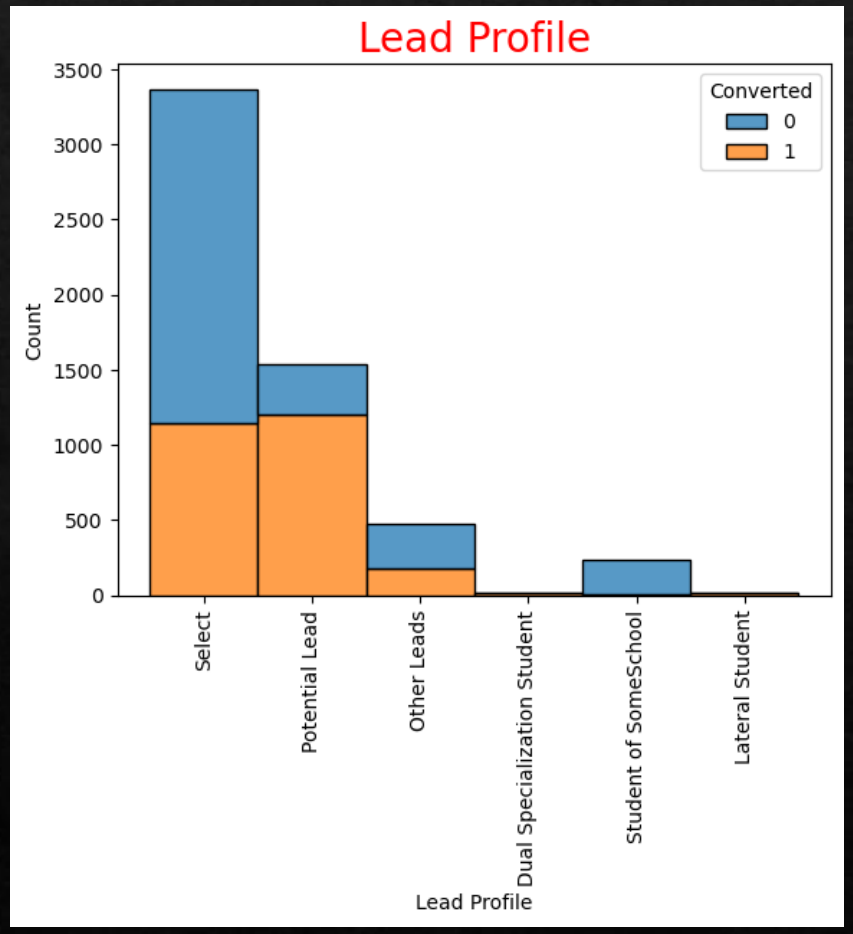
Which tags are most converted?



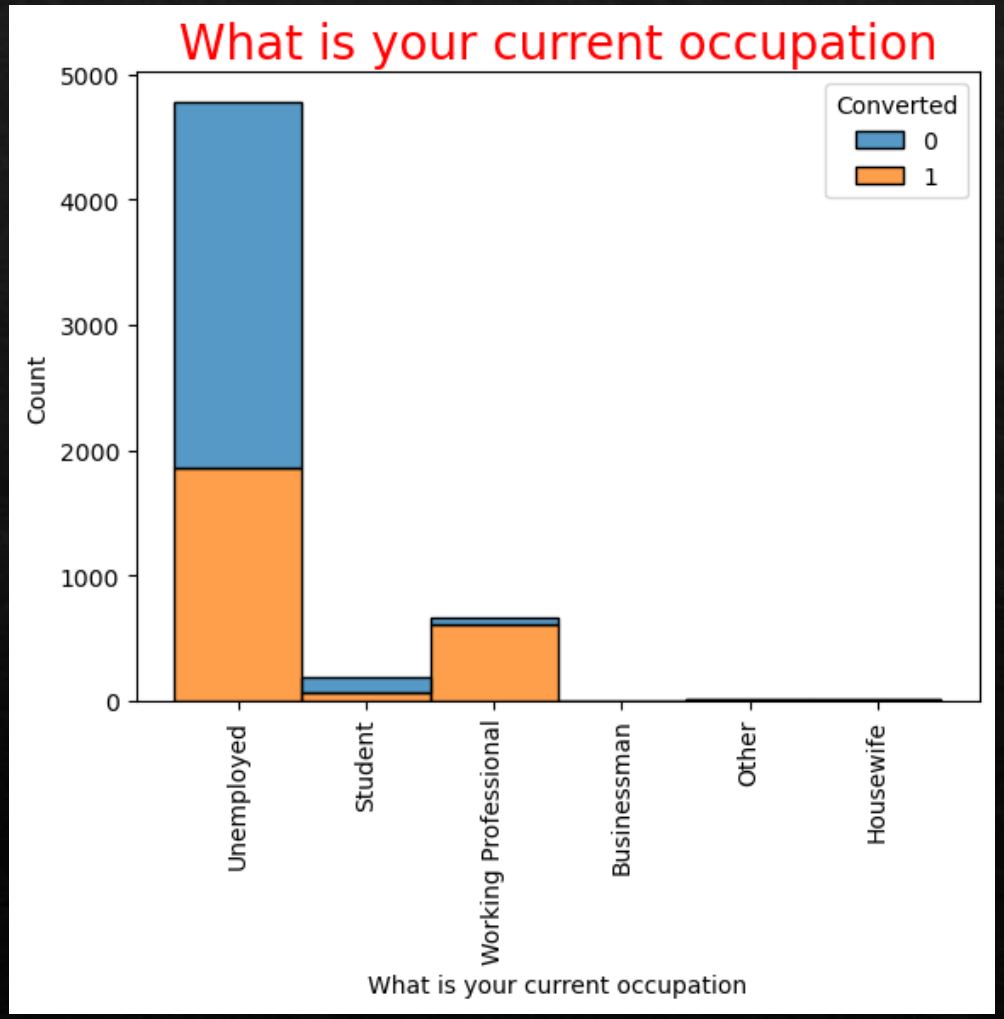
These tags have very high conversion rate:

- Will revert after reading the email
- Lost to EINS
- Closed by Horizzon

Potential Lead has the most conversion rate



Working Professional has the most conversion rate



Final Model

Dep. Variable:	Converted	**No. Observations:**	3966
Model:	GLM	**Df Residuals:**	3958
Model Family:	Binomial	**Df Model:**	7
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-545.57
Date:	Sun, 14 Jan 2024	**Deviance:**	1091.1
Time:	09:49:27	**Pearson chi2:**	3.66e+03
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.6690
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.9548	0.165	-23.981	0.000	-4.278	-3.632
Last Activity_SMS Sent	1.4007	0.187	7.489	0.000	1.034	1.767
Tags_Busy	3.2770	0.229	14.281	0.000	2.827	3.727
Tags_Closed by Horizzon	9.3447	1.019	9.169	0.000	7.347	11.342
Tags_Lost to EINS	7.4705	0.666	11.214	0.000	6.165	8.776
Tags_Will revert after reading the email	6.9221	0.212	32.680	0.000	6.507	7.337
How_did_you_hear_about_X_Education_Advertisements	1.9093	0.690	2.769	0.006	0.558	3.261
Lead_Profile_Student of SomeSchool	-2.2961	0.931	-2.466	0.014	-4.121	-0.471

Which evaluation metrics?

- Evaluation metrics are chosen based on business requirements.
- *Business language: The main target of the project is to increase the conversion rate to 80%. This means there are 80% are targets and 20% are not targets.*
- *Technical language: The positive class predicted result from the model must have true positive 80%.*

Which evaluation metrics?

Predicted converted

Actual converted

	No	Yes
No	True Negative	False Negative
Yes	False Positive	True Positive

Precision is a metric that measure how often a machine learning model correctly predicts the positive class.

Precision = True Positive / (True Positive + False Negative)

The model must have Precision >= 80% to meet business requirement.

Which evaluation metrics?

Predicted converted

Actual converted

	No	Yes
No	True Negative	False Negative
Yes	False Positive	True Positive

Recall is a metric that measures how often a machine learning model correctly identifies positive instances (true positives) from all the actual positive samples in the dataset.

Recall = True Positive / (True Positive + False Positive)

The model should have as high as possible to not miss potential customers.

Model Evaluation

The final model has very good evaluation metrics:

	Precision	Recall
Training Set	0.94	0.97
Test Set	0.93	0.97

Lead Score

Lead Score is calculated based on the converted probability that the model predicts.

Lead Score = Converted Prob * 100

Lead Score > 10 is considered hot lead and need to be taken care.

	Lead Number	Score
0	1392	1.880205
1	3123	95.107588
2	8181	7.215323
3	3791	1.880205
4	4318	98.748282

Example Lead Score from dataset

Model Adjustment

The model is able to adjust based on business requirement:

- *Consider to lower the lead score threshold (current is 10) if the business want to reach more customers and trade off for the decreasing conversion rate.*
- *Consider to increase the lead score threshold (current is 10) if the business have limited resources at the moment. The customers are less but the conversion rate is higher.*

Conclusion

- “Closed by Horizzon”, “Lost to EINS”, “Will revert after reading the email” are top 3 tags that we need to pay attention.
- With the predicted result from the model, customers who have lead score > 10 are our target.
- Consider to lower or raise lead score threshold (current recommendation is 10) to meet business strategy.