LEAD SCORING CASE STUDY

By,

Madhusoodhan H V

Sharath Chandra

Shivam Singh

(DS C70 Batch)

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.

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'.

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 a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

DATA UNDERSTANDING

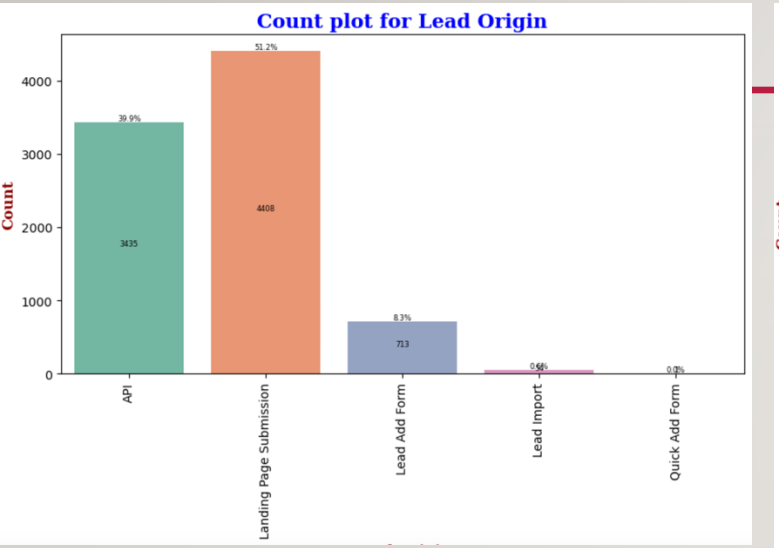
Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Country	Specialization	How did you hear about X Education	What is your current occupation	What matters most to you in choosing a course	Search
660737	API	Olark Chat	No	No	0	0.0	0	0.0	Page Visited on Website	NaN	Select	Select	Unemployed	Better Career Prospects	No
660728	API	Organic Search	No	No	0	5.0	674	2.5	Email Opened	India	Select	Select	Unemployed	Better Career Prospects	No
660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	Email Opened	India	Business Administration	Select	Student	Better Career Prospects	No

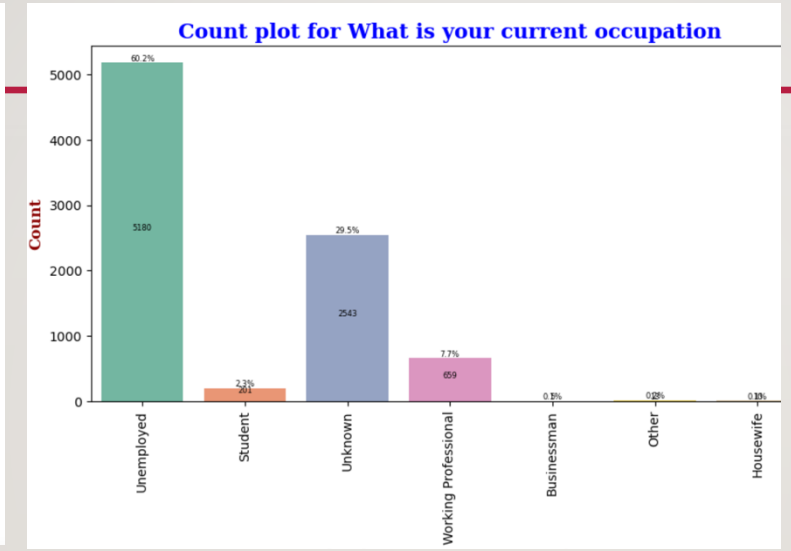
No of Rows	9240
No of Columns	37
Primary Key	Lead Number
Target Column	Converted

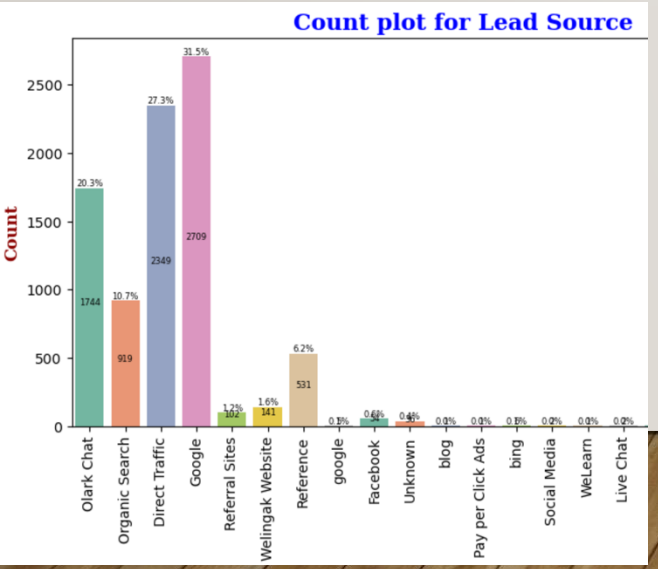
DATA CLEANING

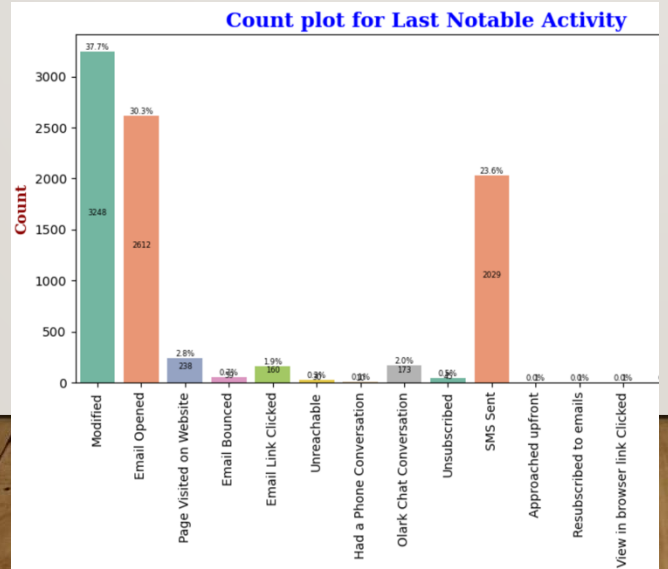
- Replaced columns having ‘Select’ value with Null. Practically student or unemployed people would not have the Industry Specialization.
- Dropped columns having 40% of null column values.
- Post analysis, computed categorical missing values with value ‘Unknown’.
- Post analysis, computed missing numerical value with median of the column value.
- Handled data outliers for numerical columns.
- Dropping high dominance columns which are having same value for all records.
- Renaming column names to meaningful one.

UNIVARIATE ANALYSIS









1. 50% of the leads arrived from the Landing page submission followed by API

2. 31% of the leads sourced via Google followed by Direct Traffic way

3. 38.5% of the leads have converted to hot leads who is contributing to the revenue of X company.

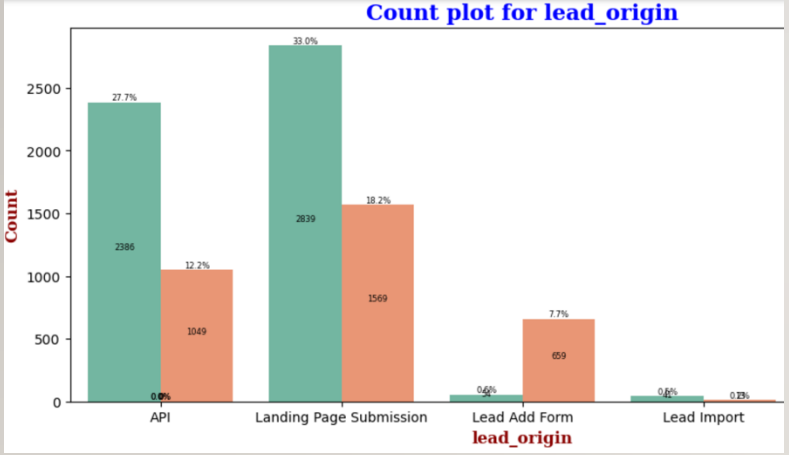
4. For most of the leads, marketing company has contacted via Email and SMS which contributes 60% of Last Activity happened

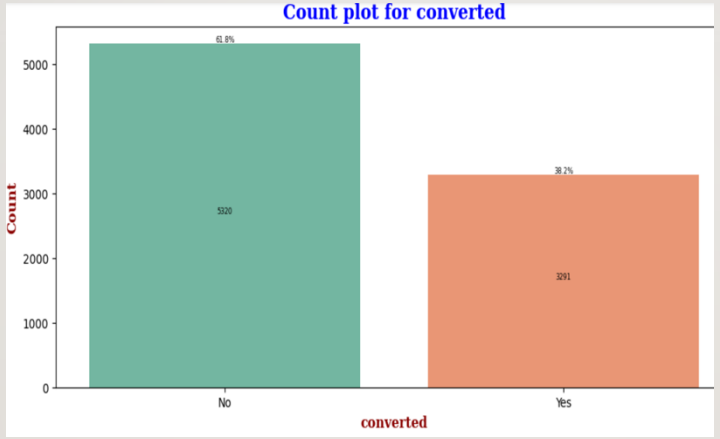
5. 70% of the leads belongs to India while 25% of leads are Unknown

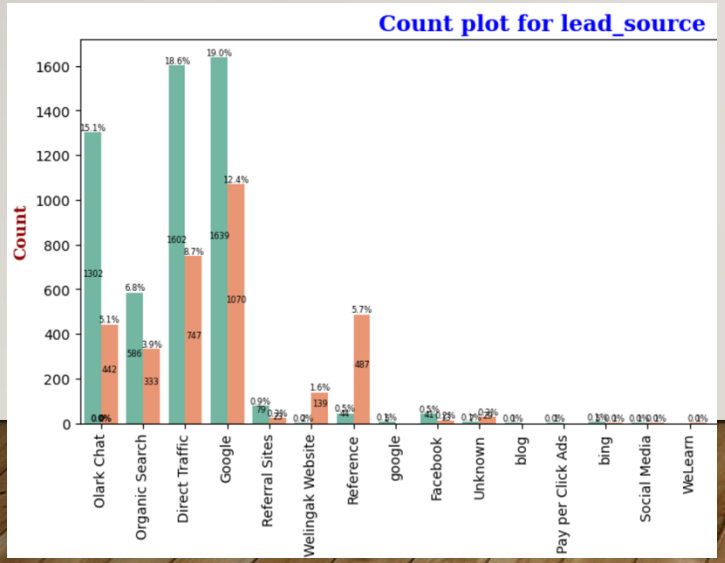
6. 60% of the leads are unemployed and we do not know the current occupation of 29% of leads. So we do not know 36% of the leads' specialization.

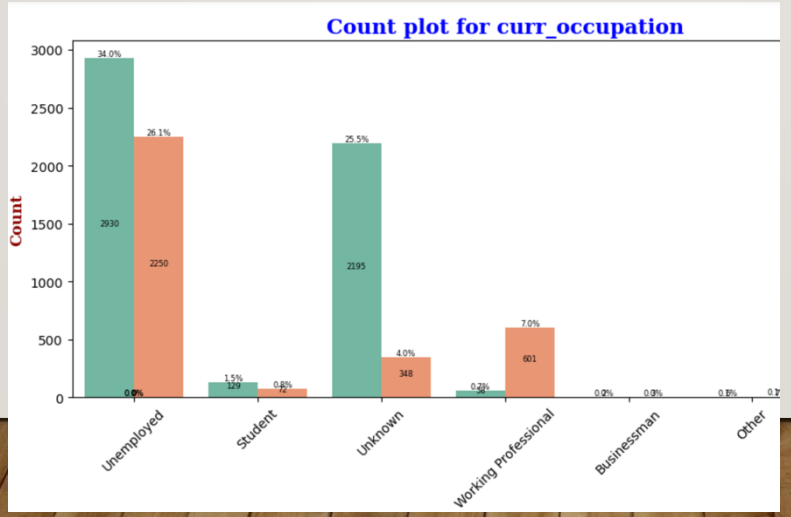
7. 70% of the leads opting for the course to have Better Career Prospects. This shows people are upgrading their skills.

BIVARIATE ANALYSIS









1. The lead conversion rate is highest for Lead Originating from Lead Add Form folowed by Landing Page Submission of the online education company.

2. Leads arriving via Welingak Website and through reference have the highest lead conversion rate.

3. The data contains 62:38 ratio of conversion data which is having clear balanced data.

4. Most of the leads have Email Opened and SMS Sent as their last_activity. Hence engaging the customers via contact details works.

5. Under the Specialization column, we could see most leads are from Unknown category, this population is clearly are students or unemployed people. We could also see good conversion rate among Working Professionals.

6. Working Professional has highest conversion rate as you could see in the curr_occupation plot. Unemployed and Students category has lesser conversion rate compared to working professional.

7. Most Working Professional opt the course for Better Career Prospects, hence there is a good conversion rate.

DATA PREPARATION

- Created dummy variables for all Categorical columns.
- Applied feature scaling to numerical columns.
- Split train and test dataset.

MODEL BUILDING

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.8595	0.161	-11.523	0.000	-2.176	-1.543
email_unsubscribe	-1.7328	0.198	-8.751	0.000	-2.121	-1.345
total_visits	0.2710	0.056	4.833	0.000	0.161	0.381
time_spend_website	1.0557	0.042	25.000	0.000	0.973	1.139
page_views_per_visit	-0.2611	0.063	-4.115	0.000	-0.385	-0.137
subscribe_master_interview	-0.3054	0.096	-3.176	0.001	-0.494	-0.117
lead_origin_Landing Page Submission	-0.6914	0.138	-4.995	0.000	-0.963	-0.420
lead_origin_Lead Add Form	3.5587	0.216	16.465	0.000	3.135	3.982
lead_source_Olark Chat	1.1336	0.146	7.739	0.000	0.847	1.421
country_Others	-0.5278	0.222	-2.375	0.018	-0.963	-0.092
specialization_Unknown	-0.9851	0.129	-7.613	0.000	-1.239	-0.731
curr_occupation_Student	1.2175	0.247	4.934	0.000	0.734	1.701
curr_occupation_Unemployed	1.0865	0.092	11.749	0.000	0.905	1.268
curr_occupation_Working Professional	3.3222	0.208	15.964	0.000	2.914	3.730
last_notable_activity_Email Opened	0.6435	0.088	7.279	0.000	0.470	0.817
last_notable_activity_Olark Chat Conversation	-1.0235	0.382	-2.681	0.007	-1.772	-0.275
last_notable_activity_Others	1.9660	0.295	6.664	0.000	1.388	2.544
last_notable_activity_SMS Sent	2.1783	0.097	22.566	0.000	1.989	2.368

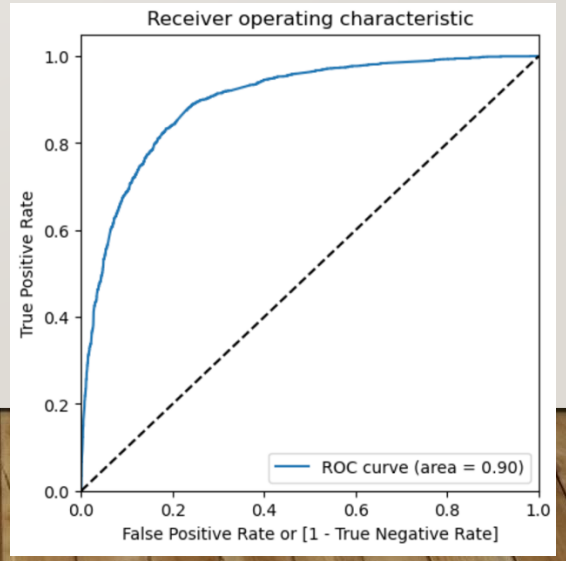
Features	VIF
lead_origin_Landing Page Submission	3.77
page_views_per_visit	3.21
curr_occupation_Unemployed	2.77
total_visits	2.64
lead_source_Olark Chat	2.58
specialization_Unknown	2.19
subscribe_master_interview	2.15
last_notable_activity_Email Opened	1.69
last_notable_activity_SMS Sent	1.66
lead_origin_Lead Add Form	1.63
curr_occupation_Working Professional	1.33
time_spend_website	1.30
email_unsubscribe	1.28
last_notable_activity_Others	1.15
last_notable_activity_Olark Chat Conversation	1.08
curr_occupation_Student	1.07
country_Others	1.05

- RFE is applied to select the significant features.
- Using the Generalized Linear Model, model is built on the train dataset.
- Final model is built by recursively checking the p-value and VIF.

EVALUATION METRICS

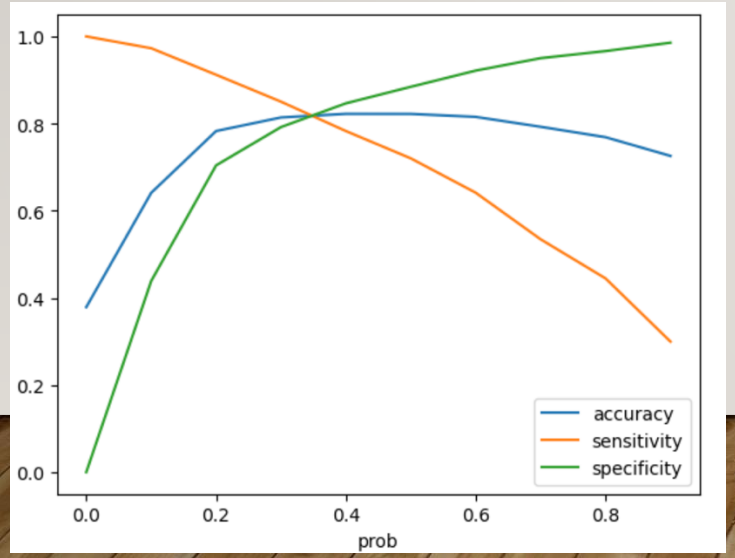
- Post applying model on testing dataset, the probability of conversion (Lead Score) is predicted.
- Selected 0.5 as Arbitrary cut-off value for predicted binary value.
- Model has accuracy of 82% but sensitivity is 72%.
- Plotted ROC curve to evaluate the model's performance. The AUC = 0.9 which signifies that the model classifies the Positives and Negatives.

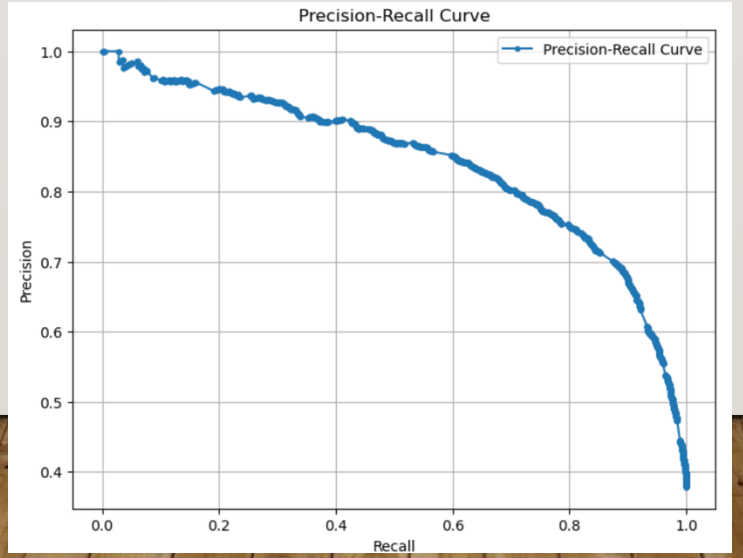
Metric Name	Value
Accuracy	82.21
Sensitivity	72.01
Specificity	88.43
Positive Predictive Value	79.15
Negative Predictive Value	83.82



OPTIMAL CUT-OFF VALUE

- To improve the sensitivity of the model.
- Plotted accuracy, sensitivity and specificity for different values of threshold between 0 to 1.
- Plotted Precision-Recall curve to check model's performance.
- 0.34 seems to be the optimal cut-off value.
- Resulted in balanced Accuracy, Specificity and Sensitivity of the model.





Metric Name	Value
Accuracy	81.98
Sensitivity/Recall	83.31
Specificity	81.17
Precision	72.96
Positive Predictive Value	72.96
Negative Predictive Value	88.86

TEST DATASET

- Applied the model on training dataset with 0.34 as cut-off value.
- Computed Lead score to indicate the chance of being converted.
- Not much deviation of Accuracy, Sensitivity and Specificity value for test dataset .

lead_index	converted	converted_probability	predicted	Lead_Score
376	0	0.051726	0	5
8914	0	0.107278	0	11
7331	0	0.238022	0	24
6344	1	0.249126	0	25
3783	0	0.265977	0	27

Metric Name	Value
Accuracy	79.61
Sensitivity/Recall	81.45
Specificity	78.43
Precision	70.71
Positive Predictive Value	70.71
Negative Predictive Value	86.86

RECOMMENDATION

Actionable recommendations to Sales team to increase the revenue.

- Engage with *Working Professionals* as they need to upskill time on time and there is a high conversion rate.
- Analyze lead behavior who originated from *Lead Add Form* and who *spends more time on website*.
- Understand the *Student's* or *Unemployed* people's background and recommend the course which adds value to them. Offer free counselling with mentors which convinces customer to opt for course.
- Automate *e-mail and SMS subscription* for the leads by sending the course brochure and curriculum. This interests customer to opt for the course.
- Target customers having *Lead Score above 80* by rolling out special discounts or encouraging them to opt for higher end products.

THANK YOU