Lead Scoring case study

Submitted By:
Ashish Abraham George
Arya Balu
Abisha Jotheesh Bell

Problem Statement

About an educational company:

An education company named X Education sells online courses to industry professionals.

The company markets its courses on several websites and search engines like Google.

Although X Education gets a lot of leads, its lead conversion rate is very poor.

So the company needs to build a model wherein 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Business Goals of the Case Study

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.

Objectives

- To help the company to choose the most potential leads.
- Using the logistic regression, predict the lead conversion probabilities of a lead.
- Build a logistic regression model to assign a lead score (0-100) to each lead, indicating their likelihood to convert. Higher scores suggest "hot" leads, while lower scores indicate "cold" leads.

Methodology

Data understanding and preparation

Import Data

Data cleaning and preparation

EDA Analysis

Model Building

Logistic regression model

Assign a lead score for each leads

Test the mode on train set

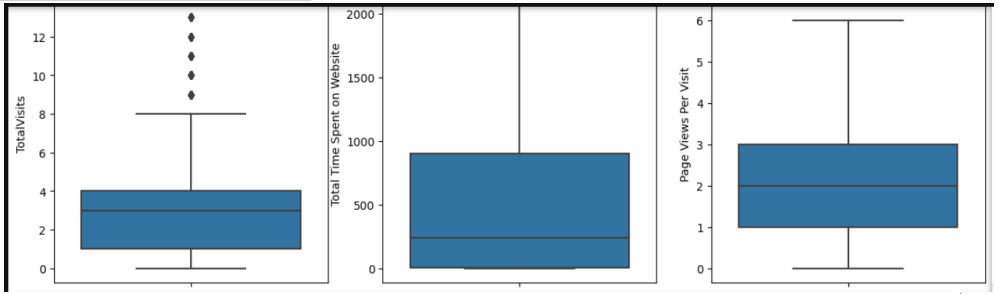
Evaluate

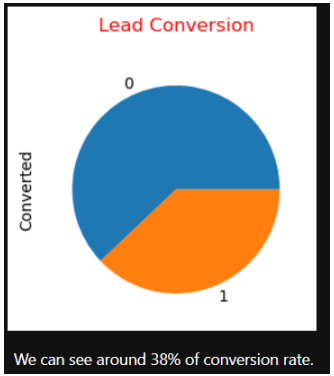
Test the model on test

Accuracy of the model

EDA

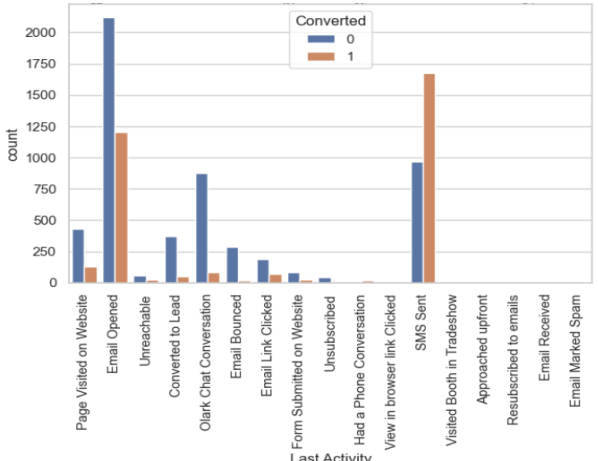
Plot the numerical columns using boxplots after the outlier treatment

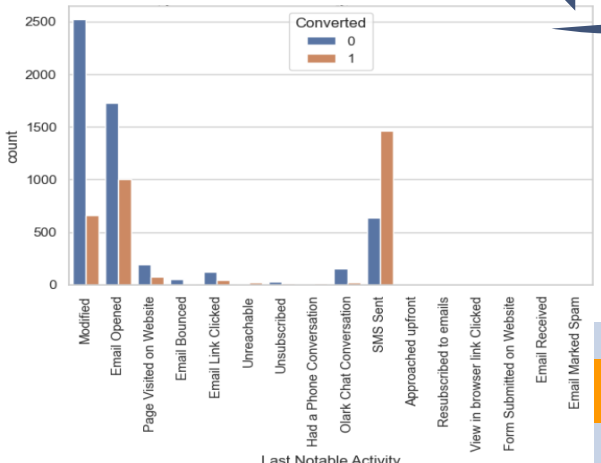




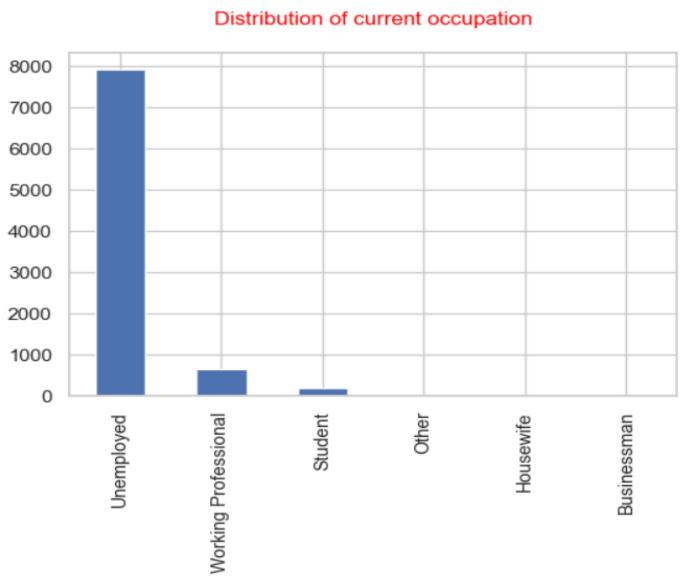
Plotting the conversion rate (Target)

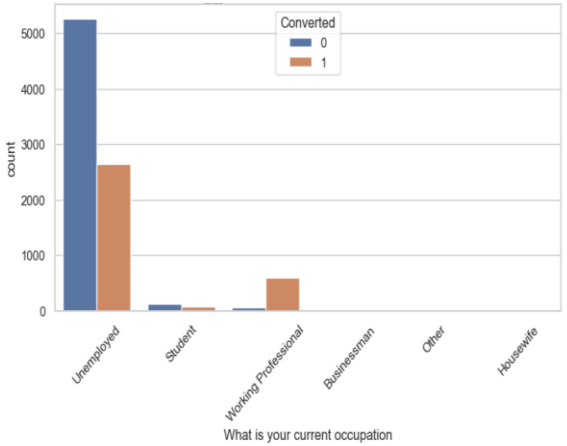
Plotting count plot for categorical variables (Last Activities) against the Target column

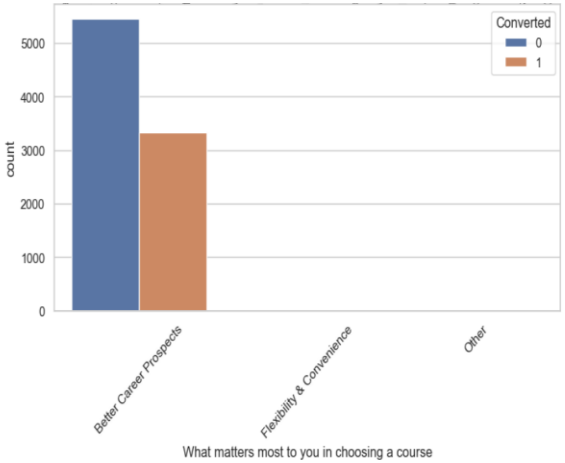




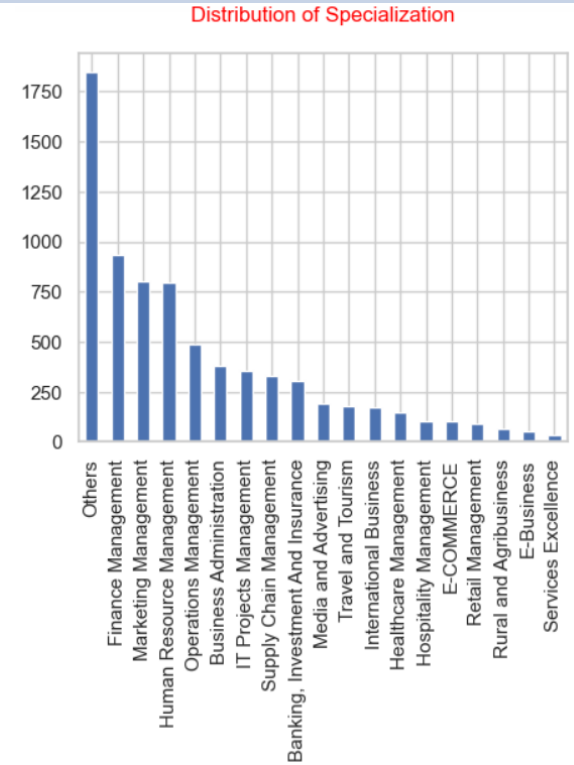
Plotting the distribution of their current occupation among the leads





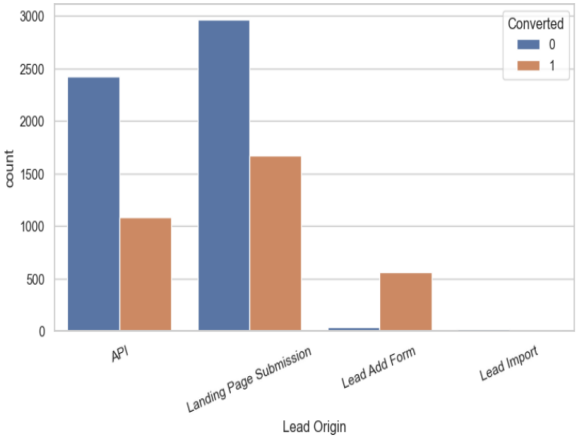


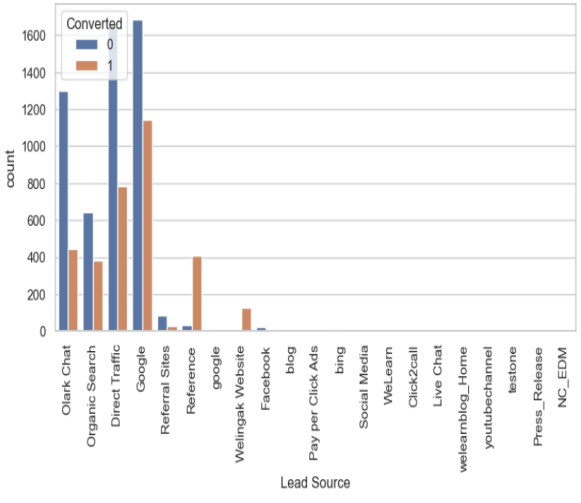
Plotting count plot for the categorical variables (Current Occupation & Course choosing factor) against the Target column

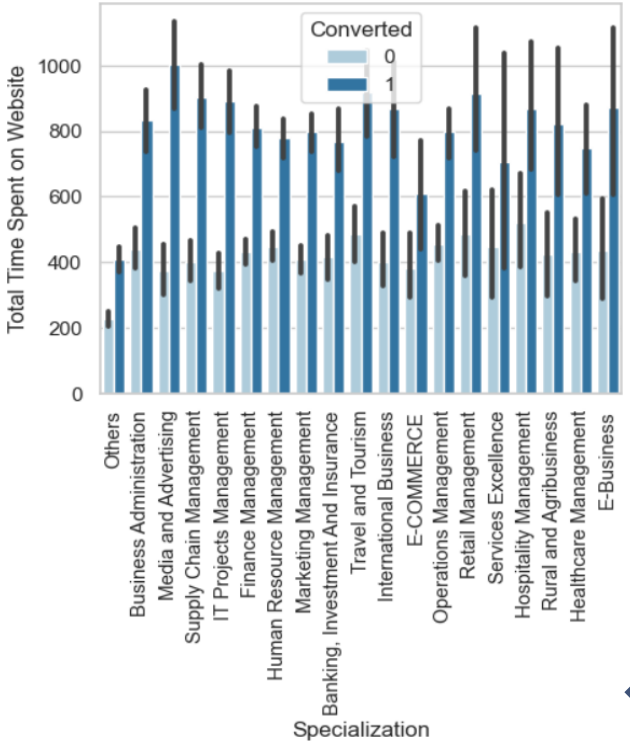


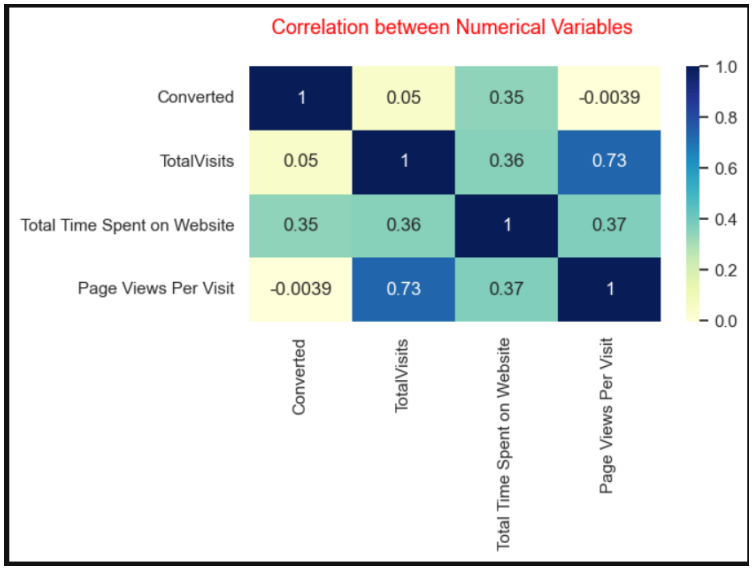
Plot the distribution of specialization of the leads

Plotting count plot for the categorical variables (Lead Origin & Lead Source) against the Target column









Heatmap to visualize the correlation of numerical variables

Correlation of Specialization against Total Time spent on Website & Conversaion rate

Observations from the above plots :

- Working Professionals are more likely to convert.
- Those who have the last activity 'SMS Sent' are more likely to convert.
- Those leads who are originated from 'Lead Add Form' are more likely to convert.
- Those leads with Google as source are more likely to convert.
- Apart from ‘Others’, most leads are from Finance, Marketing & HR management specializations.
- Those who have spent more time on the website are more likely to convert, and highly those who are in Media & advertising field.
- As per the heatmap, ‘Total Time spent on the Website’ has good positive correlation with the Target variable.

Model Approach

Logistic regression model

Model Prediction & Evaluation

Final Model Results & VIF using RFE approach

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.9154	0.103	-28.423	0.000	-3.116	-2.714
TotalVisits	1.4548	0.241	6.041	0.000	0.983	1.927
Total Time Spent on Website	4.7136	0.171	27.593	0.000	4.379	5.048
Lead Origin_Lead Add Form	4.9722	0.238	20.911	0.000	4.506	5.438
Lead Origin_Lead Import	1.8349	0.502	3.654	0.000	0.851	2.819
Lead Source_Olark Chat	1.6032	0.122	13.159	0.000	1.364	1.842
Last Activity_Had a Phone Conversation	2.0702	0.828	2.499	0.012	0.447	3.694
Last Activity_Olark Chat Conversation	-1.5159	0.170	-8.943	0.000	-1.848	-1.184
Last Activity_SMS Sent	1.3844	0.075	18.382	0.000	1.237	1.532
What is your current occupation_Working Professional	2.8791	0.195	14.738	0.000	2.496	3.262
Last Notable Activity_Unreachable	2.1581	0.513	4.205	0.000	1.152	3.164
Do Not Email	-1.4744	0.171	-8.630	0.000	-1.809	-1.140

	Features	VIF
1	Total Time Spent on Website	1.99
0	TotalVisits	1.96
7	Last Activity_SMS Sent	1.50
4	Lead Source_Olark Chat	1.42
6	Last Activity_Olark Chat Conversation	1.40
2	Lead Origin_Lead Add Form	1.15
8	What is your current occupation_Working Profes...	1.15
10	Do Not Email	1.05
5	Last Activity_Had a Phone Conversation	1.01
9	Last Notable Activity_Unreachable	1.01
3	Lead Origin_Lead Import	1.00

Model Prediction

The column Lead Score contains the score of each leads between 0 to 100 based on their probability of conversion. The leads with higher lead score shall be considered as the potential leads.

Model Evaluation (Considered Threshold 0.5)

Checking the Accuracy & Evaluating the metrices :

The overall accuracy: 0.8146944083224967 (81.5%)

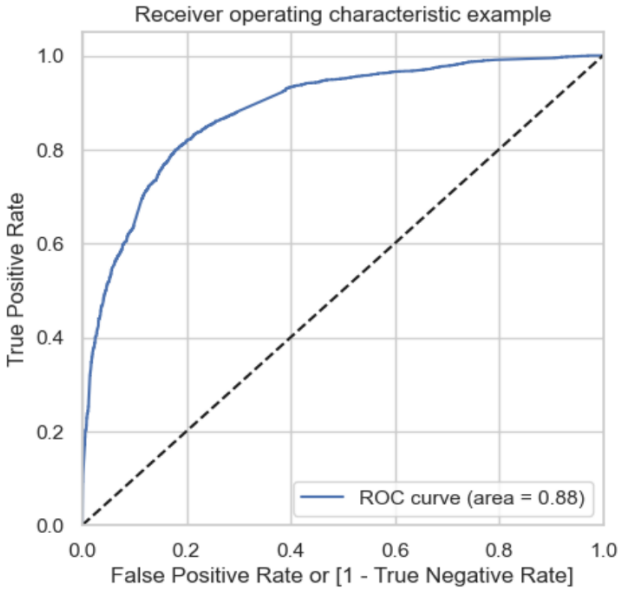
Sensitivity of the model: 0.705655526992288 (70.6%)

Specificity of the model: 0.8813514929282347 (88%)

	Converted	Conversion_Probabilty	Predicted	Lead Score
0	0	0.117472	0	12.0
1	0	0.212980	0	21.0
2	1	0.968965	1	97.0
3	0	0.055819	0	6.0
4	0	0.241388	0	24.0

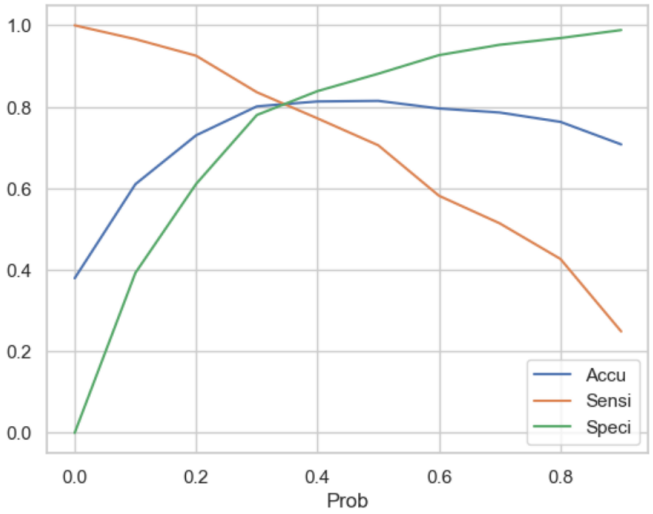
ROC Curve

The area under the ROC curve is 0.88.



Finding the Optimal Cutoff

Trade off plot: The cut-off point comes around 0.35.

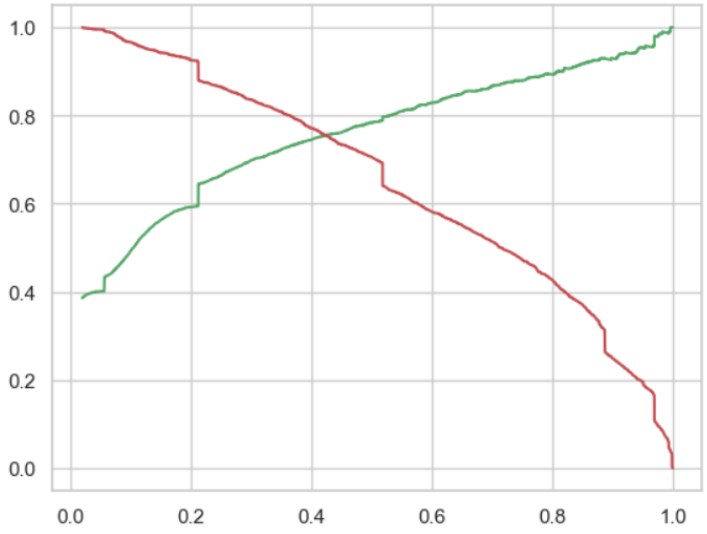


With new Threshold 0.35 :

Accuracy : 81%
Sensitivity : 80.8%
Specificity : 81.06%

Precision and recall tradeoff

Trade off plot: The cut-off point comes around 0.42.



Making Predictions on the Test Set

Precision : 74.15%

Recall : 71.63%

Accuracy : 80%

	Converted	Conversion_Probability	Final_Predicted	Lead Score
0	0	0.655112	1	66.0
1	0	0.339531	0	34.0
2	1	0.968965	1	97.0
3	1	0.937559	1	94.0
4	0	0.749393	1	75.0

Key Variables Influencing Lead Conversion

- *Total Visits:* *1.454840*
- *Total Time Spent on Website:* *4.713648*
- *Lead Origin Lead Add Form:* *4.972153*
- *Lead Origin Lead Import:* *1.834902*
- *Lead Source Olark Chat:* *1.603159*
- *Last Activity Had a Phone Conversation:* *2.070172*
- *Last Activity Olark Chat Conversation:* *-1.515929*
- *Last Activity SMS Sent:* *1.384397*
- *Current Occupation Working Professional:* *2.879075*
- *Last Notable Activity Unreachable:* *2.158080*
- *Do Not Email:* *-1.474431*

- More **frequent visits to the website** positively correlate with a higher likelihood of lead conversion.
- Spending **more time on the website** strongly indicates a higher chance of conversion
- Leads added through the "**Lead Add Form**" have a very strong positive influence on conversion likelihood.
- **Leads imported** into the system also show a positive, though moderate, correlation with conversion chances.
- Leads **sourced from Olark Chat** are moderately likely to convert.
- A recent **phone conversation** is a strong indicator of potential conversion.
- Leads whose **last activity was an Olark Chat Conversation** are less likely to convert, as indicated by the negative coefficient.
- **Sending an SMS** as the last activity shows a positive impact on conversion but is less influential compared to other factors like phone conversations.
- Leads who are **working professionals** are significantly more likely to convert, likely due to their financial stability and career aspirations.
- Leads marked as "**Unreachable**" in their last notable activity still show a strong likelihood of conversion, which may indicate persistence pays off.
- Leads marked as "**Do Not Email**" are less likely to convert, as indicated by the negative coefficient. This might reflect limited communication opportunities.

Recommendations and action plans

- The company should focus on leads who spend significant time on the X-Education website, as this indicates strong interest in the courses. These leads are more likely to convert.
- The company should prioritize leads who frequently visit the X-Education website, as their repeated visits suggest they are comparing courses and exploring options. These leads have a higher likelihood of conversion.
- Emphasize targeting leads who are currently working professionals. They are more likely to convert due to their career-oriented goals and financial stability.
- Leads whose most recent interaction with the company was a phone conversation demonstrate a higher probability of conversion.

Conclusion:

The final predictions are based on the optimal cut-offs derived from the Sensitivity-Specificity and Precision-Recall trade-offs.

The model effectively segments leads into "hot" and "cold" categories, enabling the sales team to focus their efforts more efficiently.

Total Time Spent on Website: This is a strong predictor with a high coefficient (4.71), indicating that spending more time on the website strongly correlates with a higher likelihood of conversion.

Lead Origin - Lead Add Form: With a high coefficient (4.97), leads who have submitted the Lead Add Form are significantly more likely to convert, making these leads a priority for the sales team.