Lead Scoring Case Study

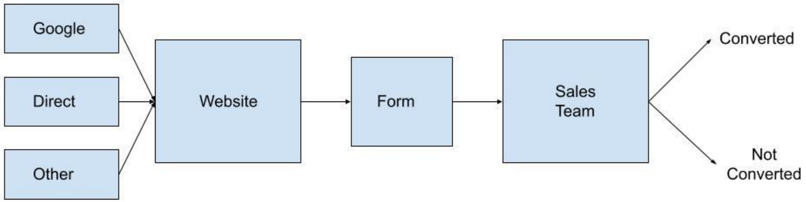
Problem Statement

A company called X Education, which sells online courses, generates a large number of leads for its sales team. The goals of the analytics project are:

- to build a model to assign a lead score to each of the leads (where a higher lead score indicates a more promising lead),
- to identify the most promising leads, called 'Hot Leads', for the sales team,
- to improve the lead conversion rate to 80%, and
- to develop solutions to some problems presented by the company.

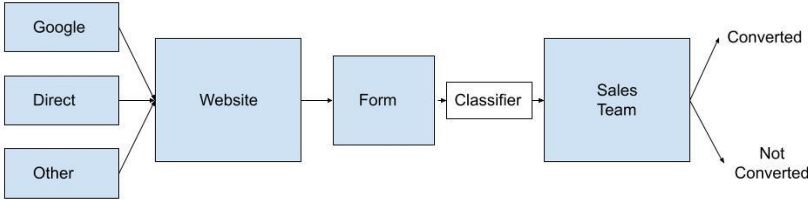
The model will be used to provide more qualified leads to the sales team and to improve their lead conversion rate, which is currently 30%.

Business Problem: *Current Model*



- Currently, users land on the website from multiple marketing channels and submit their contact information using the form.
- The lead is then sent to the sales team, which contacts users and attempts to convert them.
- The lead to sale conversion rate is about 30% at present.

Business Problem: *Required Model*



- The goal is to build a logistic regression model, which will score all incoming leads and classify them as ‘Hot’ or ‘Cold’ before they are sent to the sales team.
- The sales team will prioritise the ‘Hot’ leads, which must have a ballpark conversion rate of 80% according to the CEO.

Approach to Analysis

Analysis is performed according to the following steps:

- Data Understanding, Cleaning & Visualisation
- Data Preparation
- Model Building (Logistic Regression)
- Model Evaluation
- Model Interpretation & Lead Score Assignment

Approach to Analysis

Data Understanding, Cleaning & Visualisation

1. Firstly, the data dictionary was reviewed, the data was loaded into a dataframe and rows and columns were inspected.
2. Data cleaning was started by removing all columns with more than 45% missing values, checking for duplicate rows (there were none) and checking descriptive statistics of numerical variables to find anomalies.
3. The next step involved printing and inspecting unique values of categorical variables using DataFrame.value_counts(), which surfaced a number of anomalies.
4. These were fixed by (i) imputing 'NaN' for the value 'Select', (ii) subsequently dropping columns with more than 45% missing values, (iii) dropping columns with low variability, and (iv) imputing missing values in columns with less than 45% missing values.
5. The next step was the grouping of levels under the category 'Other' in categorical variables with many levels, followed by dropping of variables like 'Tags' which are not available during lead scoring.
6. Finally, outliers were treated and univariate/bivariate analyses were performed using data visualisation.

Approach to Analysis

Data Preparation

This step involved (i) encoding of categorical variables with dummy variables, (ii) splitting of the dataset into training and testing sets, and (iii) scaling of numerical variables using standard scaling

Model Building

1. First of all, a preliminary logistic regression model was built using all **32 predictor variables** obtained from the data preparation step and then, recursive feature elimination was performed to select **20 predictors**.
2. Thereafter, manual elimination was performed by removing variables with large p-values and variance inflation factors. A total of **5 models** were built during manual elimination.
3. Then the ROC curve was plotted (AUC score = 0.89) and the optimal probability threshold was identified as **0.4** by adopting the precision-recall view.

Approach to Analysis

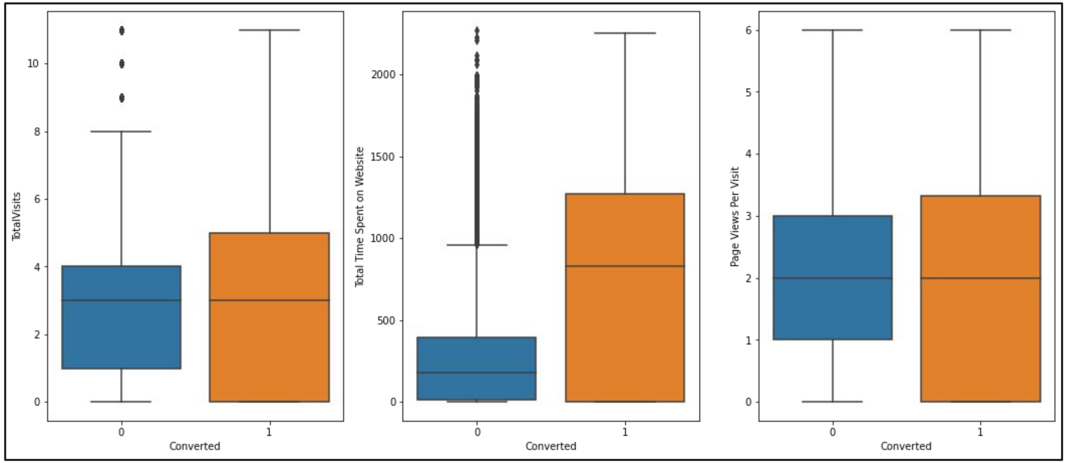
Model Evaluation

Model evaluation was done by calculating metrics such as accuracy, sensitivity, specificity, precision and recall for both training and test sets. These were very satisfactory and the model was found to be robust.

Model Interpretation & Lead Score Assignment

1. Model interpretation involved writing the equation of logistic regression and interpreting some of the coefficients.
2. The last step was assignment of lead scores to all of the 9240 leads and filtering of hot leads (about 3603 i.e. almost 40%).

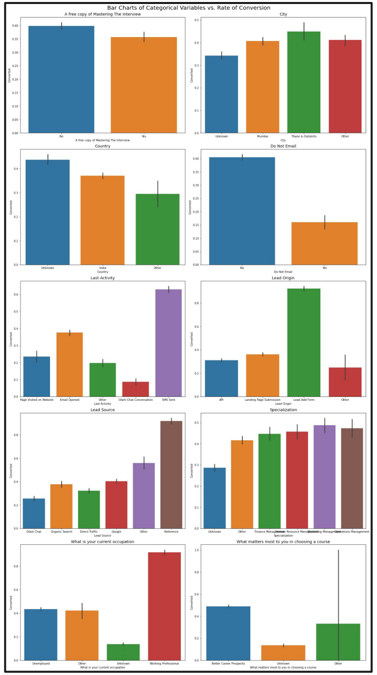
Results: *EDA - Bivariate Analysis*



Insights

1. Leads that did not convert had fewer 'TotalVisits' than those that converted.
2. Leads that did not convert spent less time on the website than those who converted.
3. Somewhat counterintuitively, there was no major difference in the 'Pages Views Per Visit' of those who converted and those who did not convert.

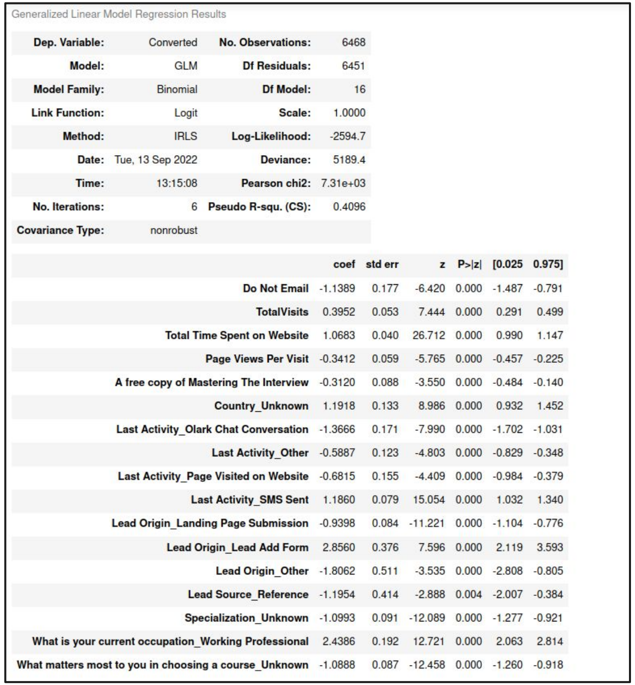
Results: *EDA - Bivariate Analysis*



Key Insights

1. ‘City' - Thane & Outskirts had the highest conversion rate, followed by Mumbai and Other.
2. 'Country' - Leads whose country was Unknown had the highest conversion rate followed by India.
3. 'Do Not Email' - Leads who wanted emails had a significantly higher conversion rate.
4. 'Last Activity' - Leads whose last activity was 'SMS Sent' had the highest conversion rate, followed by 'Email Opened' and 'Page Visited on Website'
5. 'Lead Origin' - Leads which originated from 'Lead Add Form' had the highest conversion rate by far.
6. 'Lead Source' - The lead source 'Reference' had the highest conversion rate, followed by 'Other'.
7. 'What is your current occupation' - Working professionals had the highest conversion rate by far and 'Unknown' had the lowest conversion rate.
8. 'What matters most to you in choosing a course' - Those looking for better career prospects had the highest conversion rate.

Results: *Logistic Regression Model*



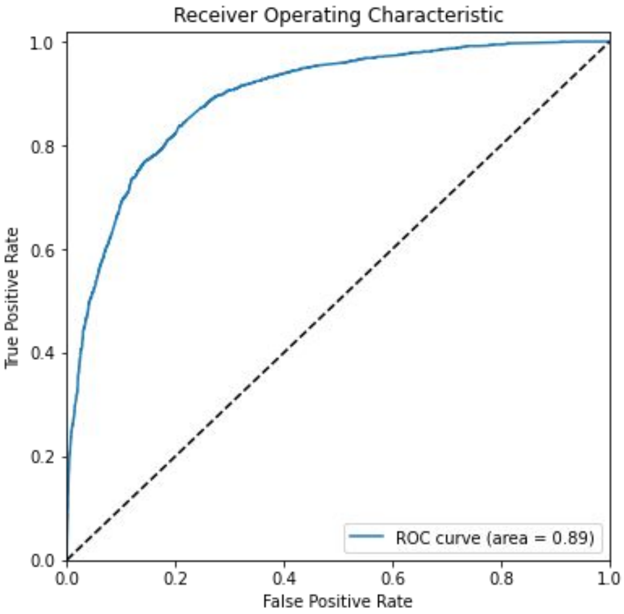
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6451
Model Family:	Binomial	**Df Model:**	16
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2594.7
Date:	Tue, 13 Sep 2022	**Deviance:**	5189.4
Time:	13:15:08	**Pearson chi2:**	7.31e+03
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.4096
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Do Not Email	-1.1389	0.177	-6.420	0.000	-1.487	-0.791
TotalVisits	0.3952	0.053	7.444	0.000	0.291	0.499
Total Time Spent on Website	1.0683	0.040	26.712	0.000	0.990	1.147
Page Views Per Visit	-0.3412	0.059	-5.765	0.000	-0.457	-0.225
A free copy of Mastering The Interview	-0.3120	0.088	-3.550	0.000	-0.484	-0.140
Country_Unknown	1.1918	0.133	8.986	0.000	0.932	1.452
Last Activity_Olark Chat Conversation	-1.3666	0.171	-7.990	0.000	-1.702	-1.031
Last Activity_Other	-0.5887	0.123	-4.803	0.000	-0.829	-0.348
Last Activity_Page Visited on Website	-0.6815	0.155	-4.409	0.000	-0.984	-0.379
Last Activity_SMS Sent	1.1860	0.079	15.054	0.000	1.032	1.340
Lead Origin_Landing Page Submission	-0.9398	0.084	-11.221	0.000	-1.104	-0.776
Lead Origin_Lead Add Form	2.8560	0.376	7.596	0.000	2.119	3.593
Lead Origin_Other	-1.8062	0.511	-3.535	0.000	-2.808	-0.805
Lead Source_Reference	-1.1954	0.414	-2.888	0.004	-2.007	-0.384
Specialization_Unknown	-1.0993	0.091	-12.089	0.000	-1.277	-0.921
What is your current occupation_Working Professional	2.4386	0.192	12.721	0.000	2.063	2.814
What matters most to you in choosing a course_Unknown	-1.0888	0.087	-12.458	0.000	-1.260	-0.918

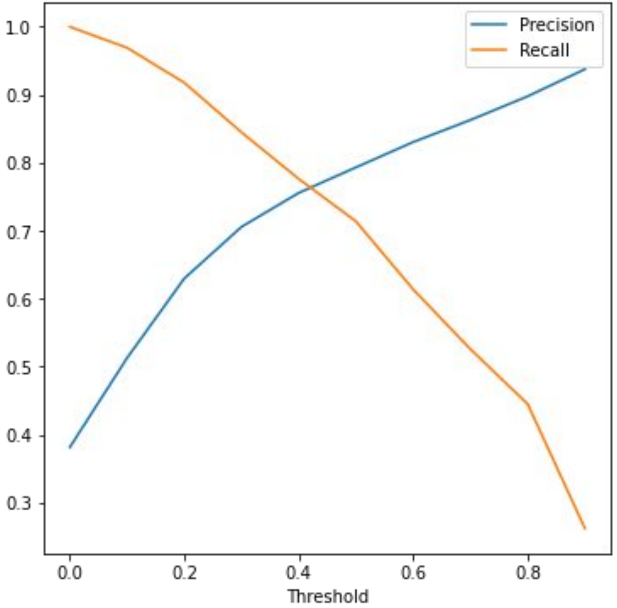
- First of all, a preliminary logistic regression model was built using all **32 predictor variables** obtained from the data preparation step and then, recursive feature elimination was performed to select **20 predictors**.
- Thereafter, manual elimination was performed by removing variables with large p-values and variance inflation factors. A total of **5 models** were built during manual elimination and one was finalized with **17 predictor variables**.
- The coefficients, their p-values and other summary metrics from statsmodel is given in the adjacent table.

Results: *ROC Curve*



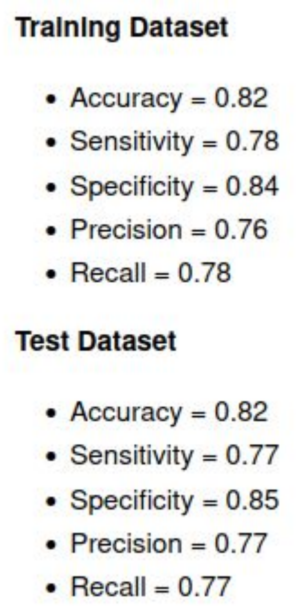
- The ROC curve of the logistic regression model has an **AUC of 0.89**.
- This indicates that the model has correctly classified the vast majority of cases.

Results: *Probability Threshold*



The optimal probability threshold according to the precision-recall view is about **0.4.** This is selected because it yields a precision between 0.75-0.80, which corresponds to the ballpark target conversion rate of 80% given by the CEO.

Results: *Model Metrics*



Training Dataset

- Accuracy = 0.82
- Sensitivity = 0.78
- Specificity = 0.84
- Precision = 0.76
- Recall = 0.78

Test Dataset

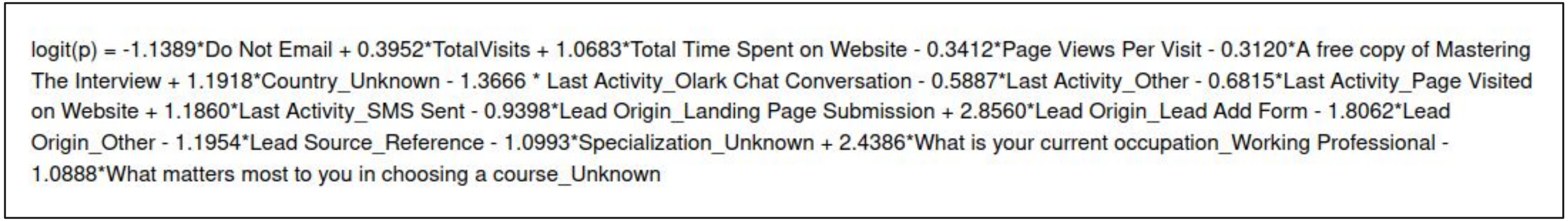
- Accuracy = 0.82
- Sensitivity = 0.77
- Specificity = 0.85
- Precision = 0.77
- Recall = 0.77

In this case,

- Accuracy is the proportion of leads that were classified correctly.
- Sensitivity is the proportion of converted leads that were correctly classified.
- Specificity is the proportion of non-converted leads that were correctly classified.
- Precision is the proportion of leads classified as converted that actually converted.
- Recall is the proportion of converted leads that were correctly classified.

Since the model metrics were very similar on the training and test datasets, the model is robust.

Results: *Logistic Regression Equation*

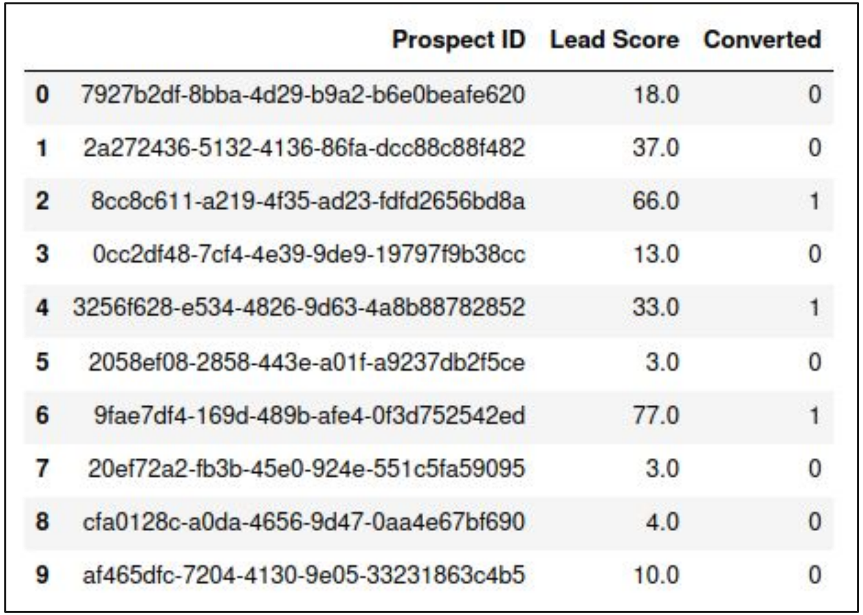


logit(p) = -1.1389*Do Not Email + 0.3952*TotalVisits + 1.0683*Total Time Spent on Website - 0.3412*Page Views Per Visit - 0.3120*A free copy of Mastering The Interview + 1.1918*Country_Unknown - 1.3666 * Last Activity_Olark Chat Conversation - 0.5887*Last Activity_Other - 0.6815*Last Activity_Page Visited on Website + 1.1860*Last Activity_SMS Sent - 0.9398*Lead Origin_Landing Page Submission + 2.8560*Lead Origin_Lead Add Form - 1.8062*Lead Origin_Other - 1.1954*Lead Source_Reference - 1.0993*Specialization_Unknown + 2.4386*What is your current occupation_Working Professional - 1.0888*What matters most to you in choosing a course_Unknown

As per the final model, the top 3 variables that predict whether a lead will convert are as follows:

1. Lead Origin_Lead Add Form - when 'Lead Origin_Lead Add Form' takes the value 1 (compared to the reference 0), log odds of conversion increases by 2.856
2. What is your current occupation_Working Professional - when 'What is your current occupation_Working Professional' takes the value 1 (compared to the reference 0), log odds of conversion increases by 2.4386
3. Lead Origin_Other - when 'Lead Origin_Other' takes the value 1 (compared to the reference 0), log odds of conversion decreases by 1.8062

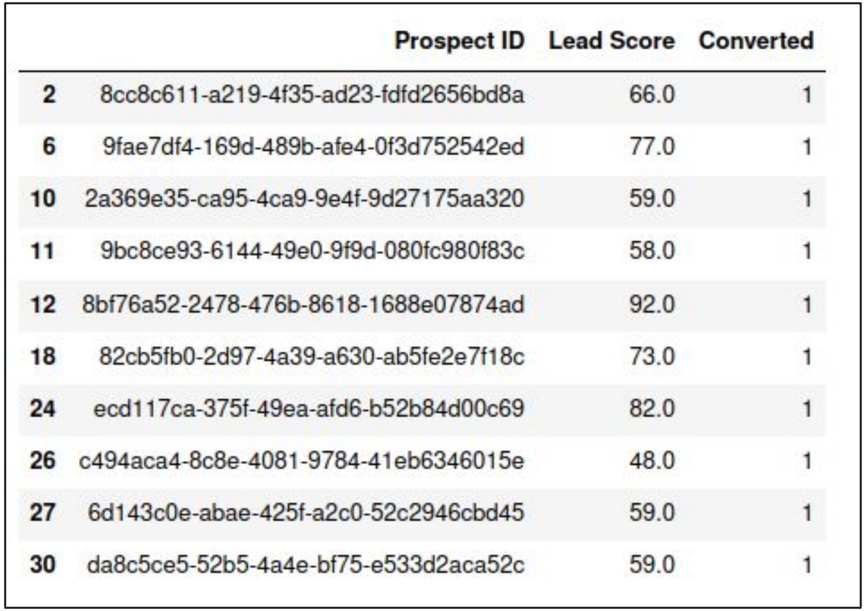
Results: *Lead Score Assignment*



	Prospect ID	Lead Score	Converted
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	18.0	0
1	2a272436-5132-4136-86fa-dcc88c88f482	37.0	0
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	66.0	1
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	13.0	0
4	3256f628-e534-4826-9d63-4a8b88782852	33.0	1
5	2058ef08-2858-443e-a01f-a9237db2f5ce	3.0	0
6	9fae7df4-169d-489b-afe4-0f3d752542ed	77.0	1
7	20ef72a2-fb3b-45e0-924e-551c5fa59095	3.0	0
8	cfa0128c-a0da-4656-9d47-0aa4e67bf690	4.0	0
9	af465dfc-7204-4130-9e05-33231863c4b5	10.0	0

- Lead scores were assigned to each of the 9240 leads as shown in the adjacent chart.
- The scores are intended for the sales team to distinguish between ‘hot’ and ‘cold’ leads.

Results: *Hot Leads*



	Prospect ID	Lead Score	Converted
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	66.0	1
6	9fae7df4-169d-489b-afe4-0f3d752542ed	77.0	1
10	2a369e35-ca95-4ca9-9e4f-9d27175aa320	59.0	1
11	9bc8ce93-6144-49e0-9f9d-080fc980f83c	58.0	1
12	8bf76a52-2478-476b-8618-1688e07874ad	92.0	1
18	82cb5fb0-2d97-4a39-a630-ab5fe2e7f18c	73.0	1
24	ecd117ca-375f-49ea-afd6-b52b84d00c69	82.0	1
26	c494aca4-8c8e-4081-9784-41eb6346015e	48.0	1
27	6d143c0e-abae-425f-a2c0-52c2946cbd45	59.0	1
30	da8c5ce5-52b5-4a4e-bf75-e533d2aca52c	59.0	1

- The adjacent chart shows a sample of ‘hot leads’.
- ‘Hot leads’ were obtained by extracting the 3603 / 9420 leads whose lead scores were greater than 40.
- These are the leads that are to be prioritised by the sales team.

Conclusion

Thus, the logistic regression model is built and lead scores are assigned to all of the leads. The model has identified 3603 ‘hot’ leads from the total of 9240. The metrics of the model are similar in the training and test datasets, which indicates that the model is robust.

The model yields a precision of about 0.77 which corresponds to a lead conversion rate of 77% and meets the CEO’s ballpark target of 80%.