# LEAD SCORING CASE STUDY FOR X EDUCATION

BY:

Akshay Tyagi Akshay Sharma

Ajay Solanki

### 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 wants its leads generation rate to be increased. For which they need to identify what are the potential leads aka Hot Leads.
- For this company wants to create a model where we can assign lead score to each of the identified leads for higher chances of conversion. Target for the company is 80 percent.

#### **GOALS TO ACHIEVE**

- 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.

### PROBLEM SOLVING METHODOLOGY

- Reading and Understanding the data
- ♦ Data Cleaning
- Data Visualization using EDA
- Data Preparation for Modelling
- Model Building: Building the model with features selected by RFE. Eliminate all features with high pvalues and VIF values and finalize the model
- Model Evaluation: with various metrics like sensitivity, specificity, precision, recall, etc.

### UNDERSTANDING THE DATA

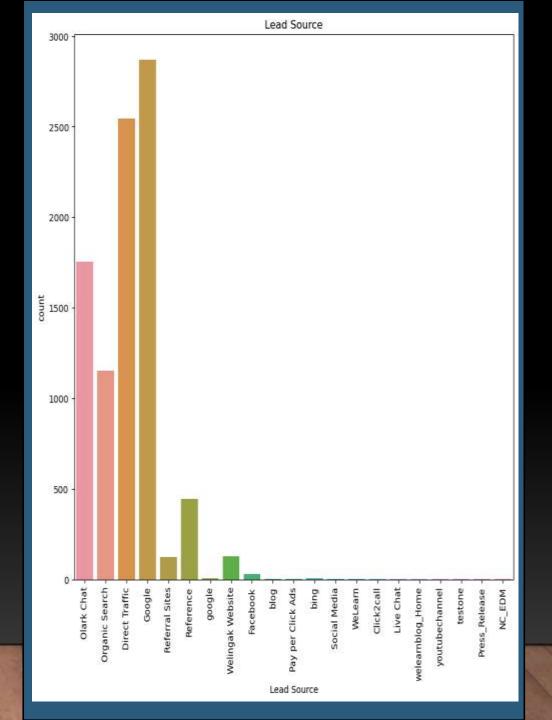
- We import all the necessary libraries for e.g. NumPy, pandas, matplotlib seaborn etc., import all the warnings.
- •We read the data and check the no. of rows and columns. We also, check if there are any missing/null values or not. Afterwards, we see the statistical summary of the data.

### DATA CLEANING

- •We saw there were few columns with high percentage of null values, so we decided to drop those columns.
- few columns had null values, but column's were important for analysis so all null values repaced with 'not provided'.
- Few of the columns had values as 'Select'so we replaced it by 'NaN'.
- Few columns were having outliers and treatment of outlier was performed.

## DATA VISUALIZATION USING EDA

- For data visualization we performed exploratory data analysis (EDA). We first did the univariate analysis of categorical and continuous data.
- Quick check was done on % of null value and we dropped columns with more than 45% missing values.
- Moved ahead with the bi-variate analysis of categorical and continuous data.
- We also found in Lead Source data most of the records were from Google and few were from social media platforms and we classified as same.
- We found correlations between variables by using different plots.

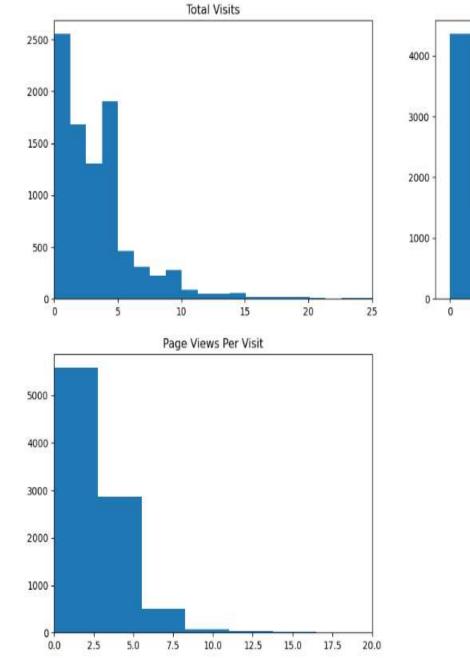


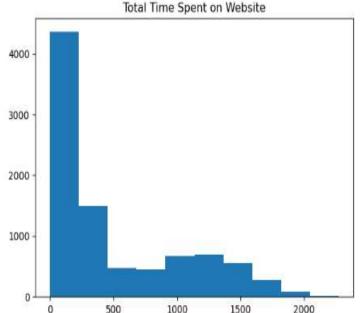


Trom this we inferred that candidates focus on better career prospectus while choosing a course.

□ Very less number of leads saw the advertisement regarding the course which shows that less visibility of the website.

From this data visualisation graph we can infer that more than 2500 candidates visit the website but spend around two minutes and visit only two pages. Therefore, website should be improved and made more enthusiastic so that candidates spend more time on the website and see maximum pages possible.





# DATA PREPARATION FOR MODELLING

- Data preparation for multiple linear regression involves handling the categorical variables first and then performing dummy encoding.
- We then performed the train and test split using 70%-30% rule and then performed the scaling of variables. Since scaling of variables is an important step, we may have different variables of different scales. So, it's important to have everything on the same scale for the model to be easily interpretable.
- Therefore, we used **MinMaxScaler** for the same.

### MODEL BUILDING

- We follow the bottom up approach for this, i.e. we start by building the model with just one variable. Hence, the choice of variables becomes very crucial.
- RFE was done to attain the top 15 relevant variables. First, we will look at the significance of variables and based on the significance.
- We checked VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept). In our model for all the feature VIF < 5 and p-value < 0.05.

### FEATURE SELECTION USING RFE

	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Lead Add Form	Lead Source_Direct Traffic	Lead Source_Olark Chat	Lead Source_Reference	AUDITOR AMBRITUATION	LIA NAP			WI
1289	0.014184	0.612676	0.083333	0	0	0	0	0	0	0		
3604	0.000000	0.000000	0.000000	0	0	- 1	0	0	0	0	-	
5584	0.042553	0.751761	0.250000	0	0	0	0	0	1	0		
7679	0.000000	0.000000	0.000000	0	0	1	0	0	0	0		
7563	0.014184	0.787852	0.083333	0	0	0	0	0	1	0		
5 rows	× 30 colum	ns										
												)

- ♦ Recursive feature elimination is an optimization technique for finding the best performing subset of features.
- ♦ It is based on the idea of repeatedly constructing a model and choosing either the best (based on coefficients), setting the feature aside and then repeating the process with the rest of the features.
- ♦ This process is applied until all the features in the dataset are exhausted. Features are then ranked according to their time of elimination.

### TO CHECK CORRELATION AMONG VARIABLES

Tetafologie	100	(6.29)	0.00	0.00	0.11	H.16	1016	3.00	0.000	0.00)	0.002	640	0.00	1039	30.033	*******	0.001	0.041	0.000	0.001	0.011	0.034	0.000	637	6.001	934	0.001	0.0070
fotal time Spent on Website	0.01	- 1	0.37		910			0.001	****	30,000	11 040	8.000		0.017	0.011	0.14	111144	0.002	0.012	0.000	0.095	(0.000)	8003	9.13	0.00	0.007	0.001	0.003
Fage Vision For Visit -	16.60	SAME	100	0.00	(839)			(833)	(889)	0.000	41.042	NAME OF		8.0	00.004	BHT.	(0000	0.107	4.00	0.07W	0.0000	+077	0.074	937	0.000	3000	10,000	0.034
Laud Origin, Lead Add Form			2.00			0.17	2.07	0.46	-0.005	-0.090	9.996	*****	9.000	0.033	0.0010	(0.14)	0.0004	-0.01	8019	9100	8.59	0.0014	9.039	0.04	5555	-016	0.0000	0.010
Least Source_Orrect Traffic	9.41	0.37	0.03	SEMI	1.0		-14	4.076	910	0.010	0.001	0.010		0.000		W.1000	-0.038	0.010	0.00	9.594	0.0179	10.0010	0.007	141145	40.000	0.043	30.0000	0.034
tand boyen, Dlark Char-			2000	11.4.4		1	m 4.1.	-0.1110	manu.	9.34	9919	-0.010	0.02	77,000	***	-910		0.02	0.017	0.40	*****	200	-11.001	0.094	0.67	0.000	93376	
Level Stearce, Reference				0.07	9.75	0.11	1/10/2	4401	0.009	0.003	0.049	0.0033	0.073	0.000	0.0020	933	amote	0.0001	-	0.0010	0.29	0.0077	0.014	0.000	0.027	0.013	namon	0.00%
Level Source_Websquir Websites -	41	41.095	0.13	0.40	4.075	0.010	0.017		0.00025	-0.027	0.015	-	-0.016	anir	0.0015	0.08	-n ni i	-0.0048	4 1000	2000	-0.011	-manual	-	41444	-0.0076		minus	0.000
Exs Poot E-mail_Year -	****	-0.959	0.00	4.025	9.11	4054	Asses	-control		4.062		- W/95	4 952	0.079	4.01154	4.616	-0.9495	+011	41110	4194	-0.044	-0.5078		445	++10	0.099	40.000	-0.057
Last activity_Converted to Lead -	400	*******	-0.001	10.0000	0.016	-617	0.001	4.411	40.000	- 4	48.000	****	4.076	-0.04	@ (MAIN	9.18	-0.002	-0.0008	9.01	-01010	-0.008	9.01	928	4.8	0.010	40.00	-0.00000	-0.013
sant Activity Small Bounced -	4.001	0.040	-0.042	18.046	11.003	4000	0.043	Aus	HAR	-0.042	- 11		0.004	0.011	40 00000	-0.11	-0.010	-0.0075	2045	0.034	0.04	10.005	40.89	0.10	0.621	0.030	0.0034	-0.011
Last Activity_Had a Plane Conversation	0.007	0.000	0.022	0.0000		-0.000	0.0012	11 00061	-1111000	-0.002		L Y	970	11044	-	40.004	-0.000	0.0001	-	10022	0.0001	0.007	****	11114	-coin			0.000
Last Activity_Olask Chat Consumation				4-983		0.42	0.072	ot title		-0.076	10014	+111+		0.000	0.0042		-0.011	10.00040	1007	0.14	10.001	0.046		0.03	0.40	*100	0.0049	402
Linit Activity_Page Visited on Website	0.09	0.017	0.0	ALC: N	11000	-0.007	0.022	-0.017	0.070	+14	Al and	0.014	4.000		0.0014		-0.036	0.001	9.29.00	Links	2366	- me		0.000	0.034		0.0034	-0.034
Last Activity Resultscotted to similar.	-014	16004	-0.000	0.00000	0.000	9300	400000	0.0031	.0.0000	distant.	-0.0074	0.06000	0.0000	400004	K1.	-0.000	0.000.0	0.0008	0.0010	0.014	40,000	0.0017	-0.0000	10,6600.5	-0000	0.0024	, K	0.50075
Last Activity Inthis Inert	4900	0.54	0.047	3666	30000	war.	2012	0.00	30.004	439	36.44	***			0.0000	*	-0.000	eme.	30000	2.004	36930	+1110			0.001	9.88	H 2001	-0.000
Last Activity_timesaltuints -	0000	0.0044	0.052	10.0004	0.000	0.020	0.00054	-011	0.000)	0.013	41.0330	+301	0.031	0.000	0.0011	0.004		-0.0636	0.005	6407	0.0017	0.000	11.040	0.000	0.044	0.000	0.0012	121
What is your current occupation_Other:	0.040	9.037	11127	0.00	0.010	686	(0.000)	0.0044	mint.	0.0000	4.00%	+100	0.01049	0.071	*****	0.000	(0.0000	10.	0.0001	-0.00	10.000	10000	0.020	8.000	0.0017	0.0009	0.000	10.000
What is your current occupation_Student -	-011	(9)00	0.000	8.000	1690	93007	9,000	0.0000	0.004	9.99	0.040	4.00	9907	0.1602	4.0010	14.841	-0.013	-0.0661	11		1100	(6,959))	0.017	10,0000	0.035	0.0000	0.0019	0.000
What is your surrent socupation, directipleyed	0.011	9.016	0.076	35898	Hore	8.00	0.0010	0.000	10.004	010010	200	1000	mae.	1,000	H.038	8998	(state)	635	12.0		30.00	****	eme.	938	0.088	0.00000	9999	10000
What is your current incorporate, Working Professional	9.033	0.000	0.0000	0.10	0.0079	0.00	9.55	-	0.044	-0.038	10.04	0.0004	0.001	0.000	+000	032	0.0017	0.003	10.040		13.5	-0.015	0.000	0.000	0.061	0.0070	0.000	1004
Last feetable followly, from Link Clinbed	-0.004	10000	11.07		*****	0.050	0.0077	-	-	***	9.995	9.807	-	0.000		10.000	-91.011.5	-0.0051	0.000	0.017	0.000	11	10.000		0.010	0.000	******	
Last Watable Activity_Email Opered	6.003	0.013	0.03%	41404	0.037	0.083	0.034	-0.015		4.81	0.33	0.000			0.0000	9.00	-0.0000	0.034	****	0.000	6.603	or once			0.007	0.15	0.0000	-0.010
Last Instable Activity Modified:	0.35	-0.12	0.11	0.00	0.045	a nan	0.049	-0.631	0.13	0.0	W 9-8-	0.019	0.90	0.000	0.0000		0.009	0.0000	0.0008	0.13	4.019	0.3.	-44	12	0.43	-0.34	name	-0.083
Last Nobelie Activity Stark Chat Conversation	400	-0.00	4.002	4.000	-0.000	11.57	41027	40016	-010	-0.052	4477	-	11.42	411159	4.000.0	of 2004	+414	-0.0057	ents-	-0 1000	4921	+1119	-0.000	441	- 0	+100	-0.001	-0.0002
Last potable delivity_Page Visited an Website .	N24	9447	0.04	= 010	0.002	0.012	0.011	49.656.00	Mane	-0.021	400	9113	-0.088		0.0018	41.11	-u.u.tu	-0.0016	0.0010	M. MILDER	0.0076	- tue	41.0	0.14	6.004	4.0	0.0004	-0.011
Lest Motable Activity_Results on bed to simally -	-0.011	0.011	-0.014	-0.0033	0.0079	9.020	-0.0000	0.0013	-0.0026	0.000#	-0.0024	0.00000	+.0043	0.0004	( A)	6.6083	0.0013	-0.0005	0.0018	4.010	0.0025	0.0017	4.005	-0.0083	+10016	0.0004	1	0.00015
Last Minable Activity Demachable	110076	9103	1115.0	0.00.0	mnre	at more	11.0025	110000	-0.003	0.003	#411	-	0.00	nnin	atimer	11.1111	11911	0.0003	44 (1999)	1111123	it size:	Anera	10.00	0.043	*****	0.001	+ 10072	116
	-	4	1	1	é	ž	i	é	*	3	Ý	1	i	ė	į.	*		i	ž	ž.	ż	1	ż	ji	i	é	·	4
	2000	rs Spet is Anh	Page lives Per	Spirit and sec	Soors Deed to	of Source, Chart C	of Sarry Leben	of the last	Do Not fissel	N_Conversal brio	and land, find lane	Phone Consetta	A Dar Coversa	day so below a	e Chichelle	at Amint, pts 5	Schill (Immily	and acceptation of	t ettipalen 30d	dust, lead	Startory Polymer	P <sub>1</sub> (Free Line Co.	dvit, trsi Ope	per bearing	A Section	present or Me	saborbetts en	desity (meath
		579		in and an	200	in	3	and part		Last Activity	Table 1	Last Attroly, Placia	Last Actually, Disc	PA JAMPA TAL	M. Herselly M.	3	THE STREET	Walls prof state	Wat A per cares	What is your named our	Main par ceret acquise.	Lachterin Actor	Last Notable A.	And the	Last Notable Activity, Clar	LATT NAME ATT PLANT.	120 Marie Actes Pe	Ladden's Lad

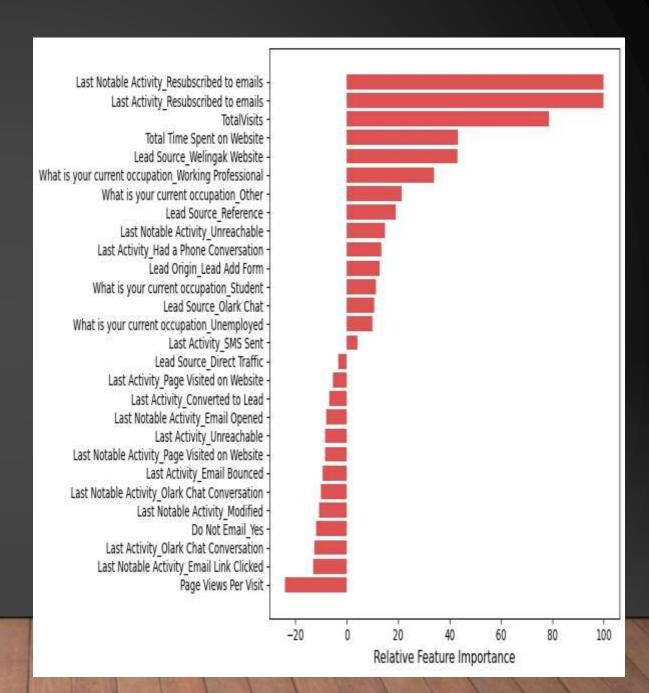
### LEAD SCORE CALCULATION

- Lead score is calculated for all the leads in the original data frame.
- Formula for Lead Score calculation is Lead Score = 100\* Conversion Probability.
- Higher the lead score, higher is the probability of a lead getting converted and vice versa.
- Since, we had used 0.42 as our final probability threshold for deciding if a lead will convert or not, any lead with a lead score of 34 or above will have a value of '1' in the final predicted column.

	Lead Number	Converted	Conversion_Probability	final_predicted	Lead_Score
0	660737	0.0	0.190369	0.0	19.0
1	660728	0.0	0.422565	1.0	42.0
2	660727	1.0	0.751855	1.0	75.0
3	660719	0.0	0.063021	0.0	6.0
4	660681	1.0	0.542724	1.0	54.0
5	660680	0.0	0.035828	0.0	4.0
6	660673	1.0	0.772298	1.0	77.0
7	660664	0.0	0.035828	0.0	4.0
8	660624	0.0	0.043511	0.0	4.0
9	660616	0.0	0.054540	0.0	5.0

#### FEATURE DETERMINATION

- ♦ The Coefficient (beta) values for each of these features from the model parameters are used to determine the order of importance of these features.
- ♦ Features with high positive beta values are the ones that contribute most towards the probability of a lead getting converted.
- ♦ Similarly, features with high negative beta values contribute the least



#### MODEL EVALUATION

- We first created confusion matrix to find the TP,TN, FP and FN.We calculated the sensitivity and specificity.
- We plotted the graph of sensitivity, accuracy and specificity for each level of probability. We found that 0.42 was the cut-off point.
- We then used the cut-off point 0.42 to select the person and see if he would be converted or not then ,
- We again created the confusion matrix to calculate TP,TN,FP & FN calculated the sensitivity and specificity and got 0.84984 and 0.78669 respectively.

