# 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

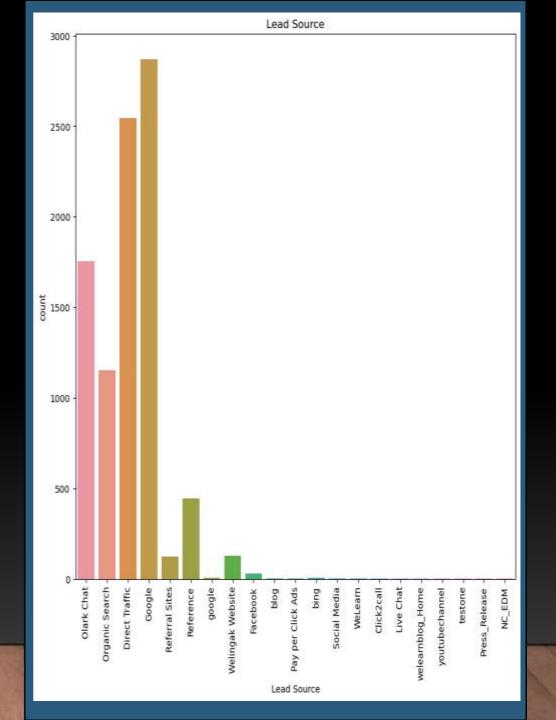
- We import all the necessary libraries for e.g. NumPy, pandas, matplotlib seaborn etc., import all the warnings.
- •We read the data and check the no. of rows and columns. We also, check if there are any missing/null values or not. Afterwards, we see the statistical summary of the data.

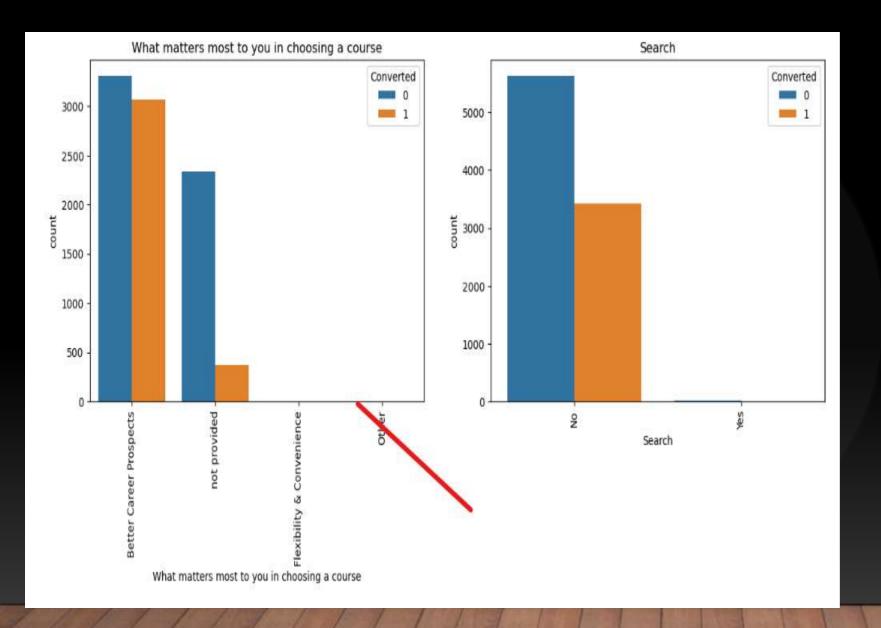
### DATA CLEANING

- •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.

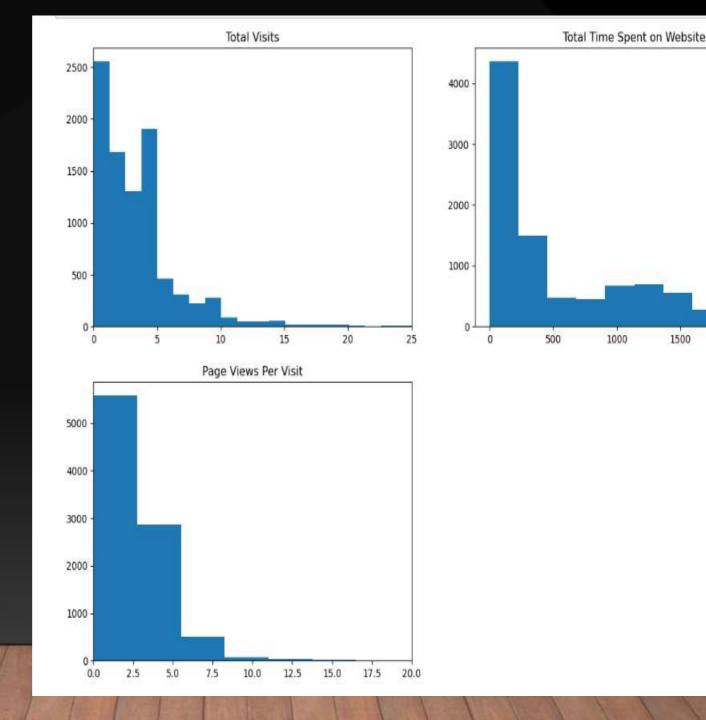




Trom this we inferred that candidates focus on better career prospectus while choosing a course.

□ 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.



2000

# 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	Lead Source_Welingak Website	Do Not Email_Yes		۱	WI
1289	0.014184	0.612676	0.083333	0	0	0	0	0	0	0		i
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										
												þ

- ♦ 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

Telarinote -	1.00	659	0.00	0.00	0.11	***	39.16	(000)	0.000	0.00)	0.00	1407	0.00	16.219	****	= 11111	0.001	0.041	10.000	(660)	0.011	0.004	0.011	0.00	0.001	834	6003	0.0076
Total Surve Spenit on Website -	8.04	0.0	0.07		910			-0.00	9309	0.0000	400	10.000		0.017	0.033	014	331144	0.007	0.000	0.000	0.099	(0.000)	0.000	437	10.00	HARY.	0.001	0.003
Page Views for Visit -	9.69	578.99	-01	10.10	933	896		431	7.8087	0.00%	11 042	18000		9.0	01.034	ner.	176665	0.107	440	0.079	0.0076	+101	0.074	9.34	*1000	9.00	9.00	0.004
Linux Origin_Load Add Form -				111		4.10	200	0.40	0.000	-0.010	8.995	# 001E	-0.040	0.023	+ 2013	634	(0.0004	0.01	8018	(8488)	0.09	0.000+	100.000	0.04	0.004	-0.00	0.0003	0.014
Court Squirco_Orient TreMis	0.11	0.12	0.1.0	5.46	1.	0.31	-14	-0.070	933	0.010	0.000	(0.000)		0.000		0.0000	-0.000	0.010	-0.01	0.000	9999	or person	0.037		41.000	0.000		0.024
Level Source_Olark Chor.		-570	930	11.8.3			-0.11.		0.004	400	4410	0.01E	0.60	0.000	ware	-0.13	-0.000	-0.62	0.013	0.00	9.99	m 010	11.000	0.090	HAY.	-0.000	pare.	-0.000
(med Source_Reference -				1941	496	0.11	and the	-0.027	-0-00W	nmı	0.049	0.0012	-0.011	0.011	0.0019	913	Emote	-0.0001	-	none	0.29	0.0077	-0.004	9.949	0.021	-0.015	nama	0.004
Level Source, Welengov Westons -	411	-0.055	4112	0.46	-0.075	41014	41.022		0.00025	-ninz?	4.415	+1062	-0.336	41017	e mus	0.00	-0.012	-0.0048	e made	0.001	-0.011	-erminis	0.015	41.025	-0.0076		a-mm1	-0.0000
Env hove Ennact_Ven	- Harris	3150	0.02	4-225	an.	arona	0.078	cours.	W.,	0.063	464	di Maria	0.012	0.079	0.0014	4556	0.0001	4001	8048	10 204	4.044	-0.010		0.18	0.010	a pase	ar out the	0.011
Last Activity_Converted to Lead -	-0.001	-0.0030	-0.001	41.006	0.058	-011	0.011	-0.001	-0.004	1	0.042	4.012	-0.079	40.000	-0.0018	0.11	-0.003	-0.0008	9.01	-0.0014	-0.004	48.000	0.11	4.5	0.030	-0.042	0.0008	-0.013
Likit Activity Privat Business	-	and an	11.114.0	****	it met	1000	0.044	sems.	0.60	armax	- 10	-	10 0014	ams		0.11	01.013.00	name.	2000		41.004		4111	210	*****	-	******	amx
Cook Activity_Hed a Phone Conversation	4.017	10000	0.022	month.	-	-0.010	0.0012	m 00001	-111121	-0.012		T.		0.014	-0.00001	10000	-0.000	-0.0011	*****	6422	0.0001	-0-2017	-0.000	0.020	-0.007.6			-0.088
Last Activity_Chark Chait Convertagion				40000		0.13	0309	10,000	1000	0.090	41000	0.018	100	20.000	0.0000		-0.011	0.00040	8417	0.14	10.000	-man-		0.00	0.45		*****	4.00
Last Activity_Page Walted on Welladia -	0.25	9.007	0.0	H-935 / .	0.000	0.041	0.000	-0.003	0.079	4.00	H-006	0.004	10 3144.0	1000	0.0024		-0.000	0.103	636-60	3699360	4 625	0.030		0.068	40,000		0.003+	-0.010
Last Activity Passales—Baid to emists	mental	00000	0.034	4.0000	manery.	0.036	40,000	40,000,0	1111104	(4000)	4.0024	2.06095	0.1004.1	dimia	0.0	0.0001	300000	0.0000	0.0000	0.114	0.0035	20013	m.4080	10.64953	-0.30330	41,00004	1	0.00076
Link Activity_SMS (land -	0.0001	(0.04	-0.067	0.14	0.0000	10,000	.010	0.00	-0.006	4.16	3033	0.004			0.0000	T.	-0.004	0.010	0.00	0.000	1018	-0.000	0.00		40,004	9.13	0.0001	-0.030
Last Activity_Vermachamin	0.000	0.0044	3.032	e priva l	-0.000	0.000	amore.	-0110	0.0909	0.002	0.010	9.000	0.011	0.000	0.0015	11.004	- 69	-0.0116	H-058	19487	0.0017	0.015	0.000	name.	0.014	0.000	0.0003	239
What is your correct occupation officer-	0.040	9307	OWNER	70.00	note	500	49,000	0.0000	90X	9,000	4.0074	+000	0.09050	0.00%	earns.	DRIVER.	0.0000	100	4.0001	0.0%	19.98X	0.0000	0.004	N DOUBLE	-0.0017	0.0079	9.000	000000
What is your current occupation, Muderit	4.414	19999	9.93	35930	1999	ниу/	9.995	01.000mg	10000	9.00	2010	* 2015	(9888)	0.1005	0.0019	190.000	-0.015	0.0000	* E	1100	0.011	2500 Y	0.017	10.00000	9333	9.0008	H-0010	9.0000
What is your current occupation, charrighty ed-	1999	0.000	0.070	0.055	0.050	*100	0.0010	0.003	0.034	0.0014	0.000	0.000	0100	1.0+00	0.000	0.000	0.001	4.65	SE		***	-0-017	0.038	9.94	0.000	-0.00000	9.00	000003
What is your current incorporation, working Professional		-	0.0079	9538	MARKET .	11.04	10.00	-0.001	0.000	-0.000	****			a nen		9.17	0.000	0.001	-0.000	11.00	17.5	-0.015		-0.070	11/97	******	***************************************	2.000
Last Rotable Activity_Dreaf Link Clicker	# 60×	0.041	0.033	0.0004	0.0018	0.054	0.0077	0.0007	0.034	0.01	0.025	0.007	2000	9.0m	0.0007	0.000	-0.003.0	-0.0003	0.003	4517	0.011	100	-mmq	-1	0.010	0.026	0.000.7	a.nore
Last Notable Activity_Email Opened	nme	0.015	0.014	0.000	8.037	0.011	47.034	-0.015		6.55	6.51	0.000			mannes.	et.ab	-0.2968	0.034	BALT	mate	-0.011	0.000		4.5	0.001	411	-mannes	0.039
Lant Nobalds Activity_Modified -	415	46.60	411	40.000	0.045	0.17911	A1.044		013	0.0	0.19	2010	MAR	0.000	- man		0.059	stream	441114	M 5.2	4974	-0.6	-44		46.11	4.14	41.0004	-0.044
Last Hotable Activity_Otark Chat Conversation		-0.06	-0.961	48.026	9.049	0.17	0.027	0.0076	-0.018	-0.013	0.027	waite	0.42	0.030	e outs	42,004	-0.014	-0.0007	8.016	9.566	4927		- NEW MARK	0.11	-10	-0.038	-0 mile	0.0003
Last Notable Activity Page Vistori on Wellinks -	0.84	mass.	0.01	41.014	0.003	40.052	0.033	0.0067	nme	-0.040	41.056	0.03	2.000		40.000.00	411	-0.013	0.0078	0.0000	W. Harrison	0.0019		40.00	0.34	0.018	1.	41,000.0	-0.011
Last Notative Activity_Resolver their to errors.	-mi	00000	111114	0.0000	MARKET .	9.020	40.002.0	-	-1111000	0.0020	-0.0024	or instead.	0.1004.1	4010004	100	S. 0118.1	*********	0.0000	44444	-	-0.0025	******	30,000		-0.0010	41 01124	- 1	* 20075
Cast Notable Activity_Unites hable	0.0070	9.019	0.014	N.OLD	-0.004	-0.0074	0.005	0.0000	-0.017	-0.013	4011	-0.001	-0.01	0.016	-0.00072	44.074	0.00	-0.0111	0.0000	0.0011	0.020	0.0010	-0.000	41142	- mea	-001	440072	20
	Makes	- Male	The last	100	ectally.	Olum Char.	1	4	Pack.	divised.	Ţ,	į	ě	and a	4	N N	mode.	90.00	, Soler.	Plate	desard	a Cidel	- Daved	į	ĺ	v Vertoday.	1	medalin
	1623	Thirt fee bash	Page Veet	beet degre lead	tand Source Day	0 emos per	enter best	tind foury beings	Take 0	Lat Athely Constra	Last Activity, Small	Last Activity, suit a Phone Can	Last Activity, Diate Chall Can	e protes adel y felich zum	Last Achies, Number thed	Cast Activity,	Lett Article you	Walk is pred harmed account.	MARK Symptometry conjugate	Wat your careet suspense, Un	hat is parturbed ompation, library Po	Contraction activities (Straft)	Lad Worldis Activity, Driv	Care Metable Actives	Last tenddis Activity, Olak Olat Con	Lat Notable Activity. Page 100000 is	Lad Nazalia Accideg. Seaskucibad	no famous sections of

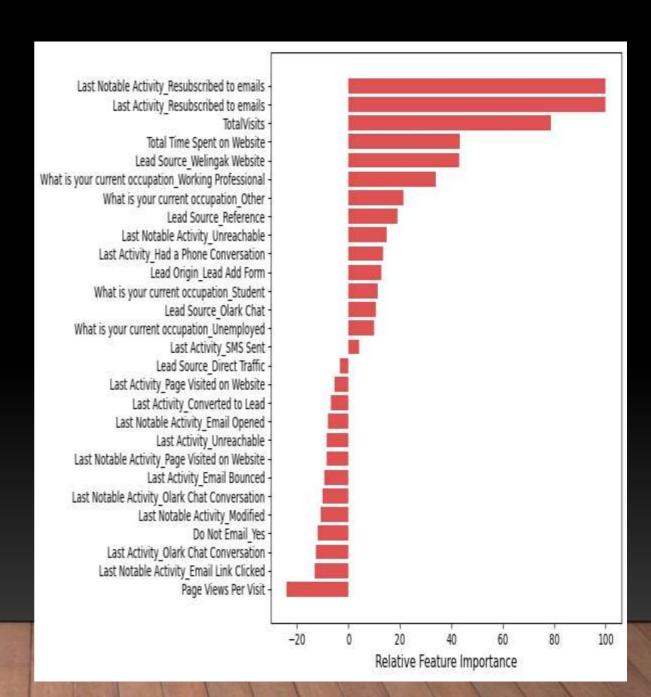
### LEAD SCORE CALCULATION

- 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.

