X Education Online Course

SUBMISSION

- ANSHUL TOMAR
- BINDHU BALASUBRAMANIAN
- ASHWINI VATSA
- GAURAV JAWRANI

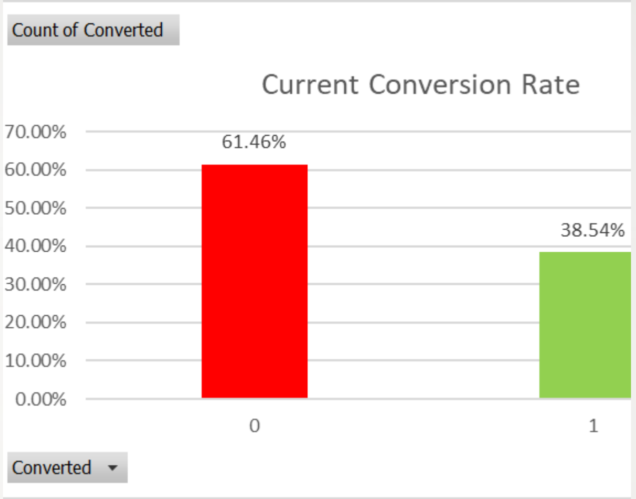
X Education

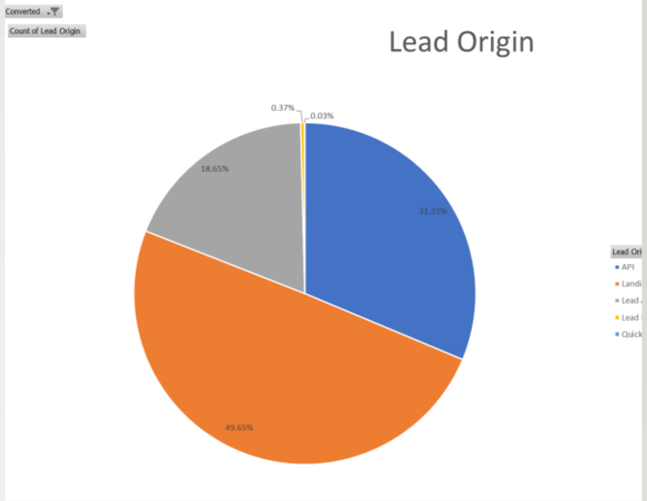
An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.

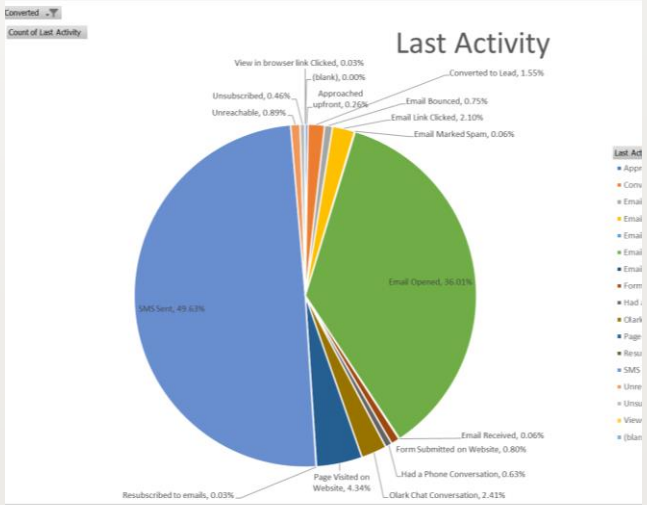
Business Objective: The lead conversion rate is very poor and company wants to make the process efficient to identify the most potential leads, also known as 'Hot Leads'. It will help sales team to focus more on communicating with the potential leads rather than making calls to everyone.

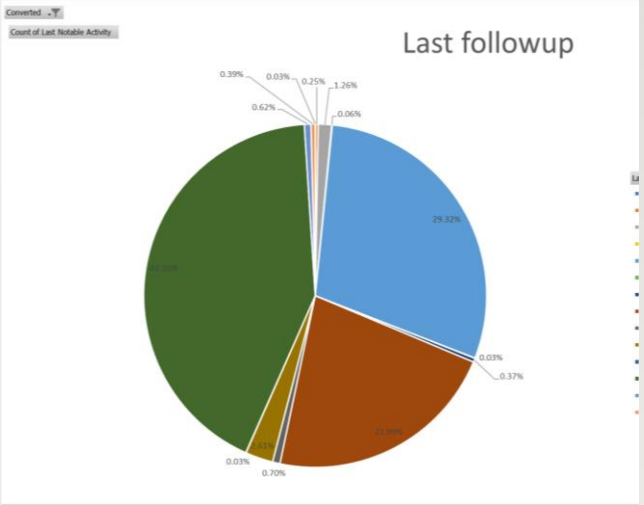
Overview

- The dataset has 9240 records.
- The current Lead conversion rate is below 39%.









Methodology Overview

Data loading

Data cleaning

Standardization

Data scaling

Dummy Variable creation

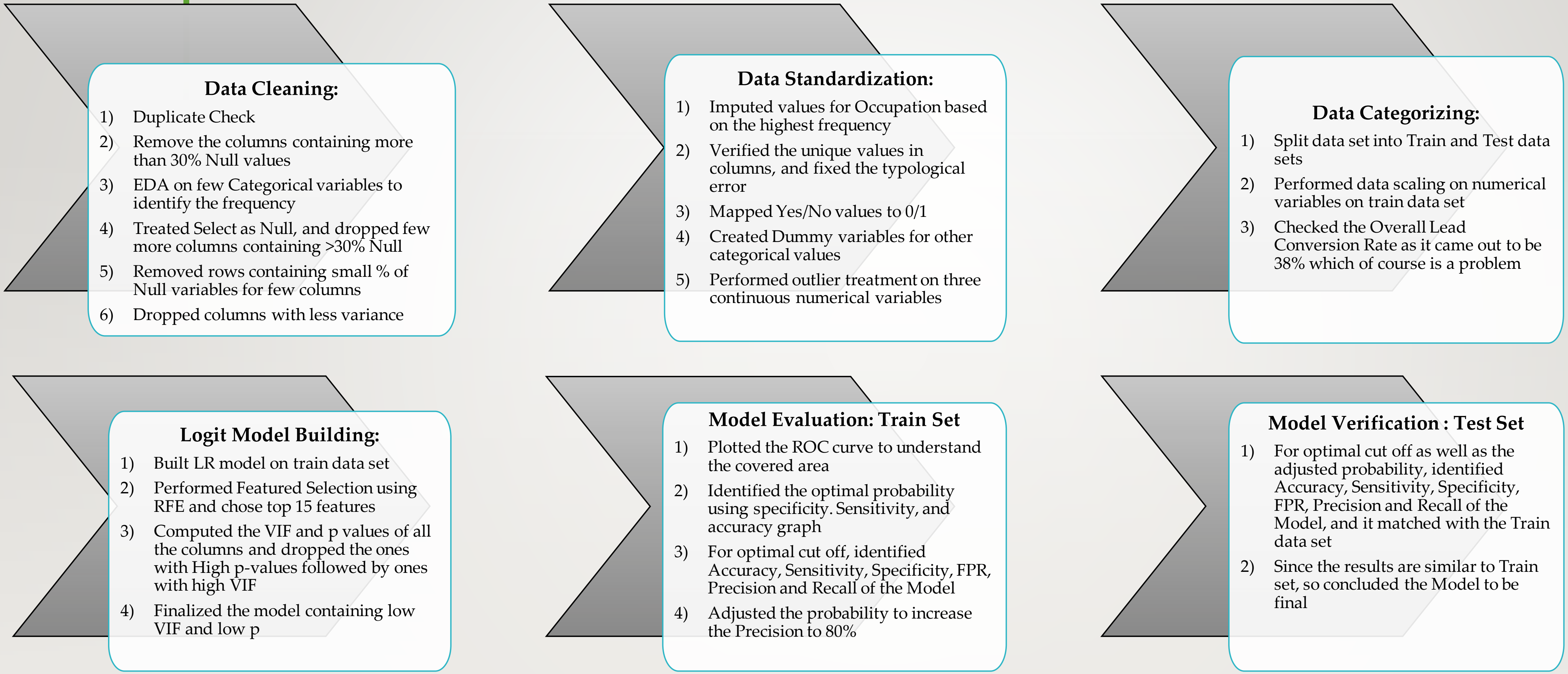
Outlier check

Prediction

Precision and Recall

Lead Conversion

Detailed Methodology



Data Cleaning:

1) Duplicate Check
2) Remove the columns containing more than 30% Null values
3) EDA on few Categorical variables to identify the frequency
4) Treated Select as Null, and dropped few more columns containing >30% Null
5) Removed rows containing small % of Null variables for few columns
6) Dropped columns with less variance

Data Standardization:

1) Imputed values for Occupation based on the highest frequency
2) Verified the unique values in columns, and fixed the typological error
3) Mapped Yes/No values to 0/1
4) Created Dummy variables for other categorical values
5) Performed outlier treatment on three continuous numerical variables

Data Categorizing:

1) Split data set into Train and Test data sets
2) Performed data scaling on numerical variables on train data set
3) Checked the Overall Lead Conversion Rate as it came out to be 38% which of course is a problem

Logit Model Building:

1) Built LR model on train data set
2) Performed Featured Selection using RFE and chose top 15 features
3) Computed the VIF and p values of all the columns and dropped the ones with High p-values followed by ones with high VIF
4) Finalized the model containing low VIF and low p

Model Evaluation: Train Set

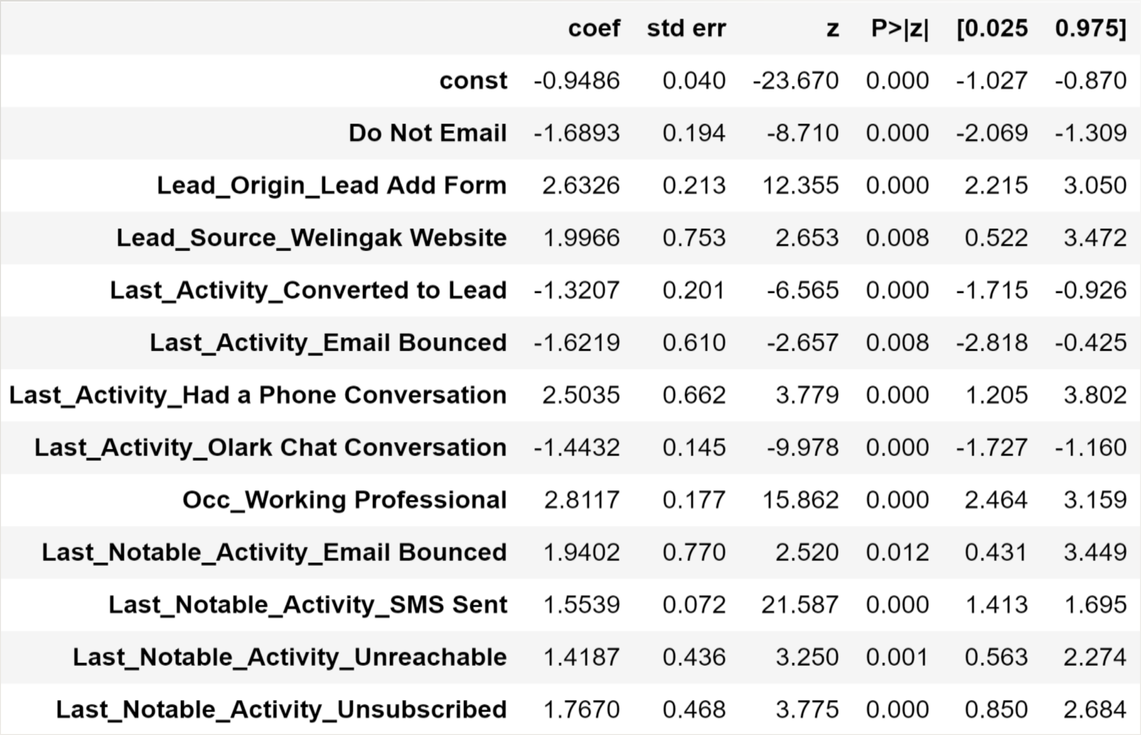
1) Plotted the ROC curve to understand the covered area
2) Identified the optimal probability using specificity. Sensitivity, and accuracy graph
3) For optimal cut off, identified Accuracy, Sensitivity, Specificity, FPR, Precision and Recall of the Model
4) Adjusted the probability to increase the Precision to 80%

Model Verification : Test Set

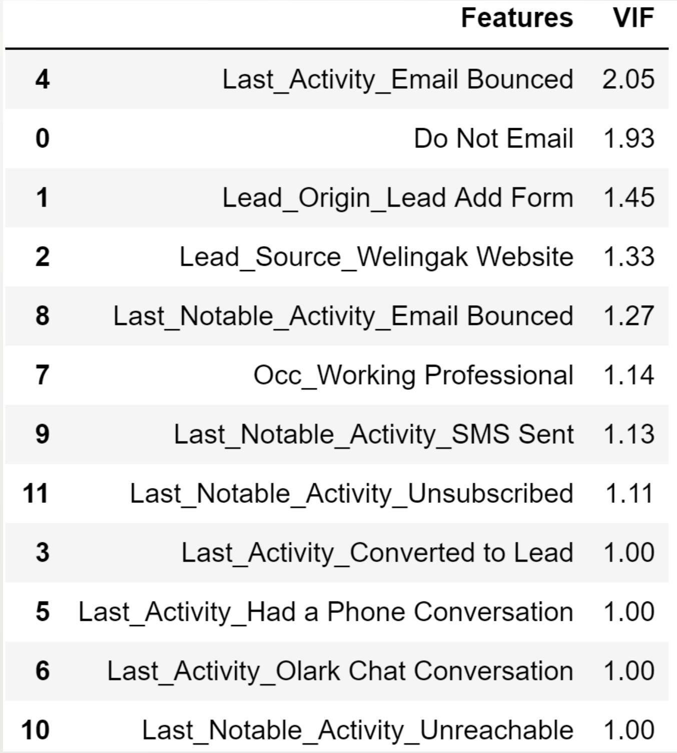
1) For optimal cut off as well as the adjusted probability, identified Accuracy, Sensitivity, Specificity, FPR, Precision and Recall of the Model, and it matched with the Train data set
2) Since the results are similar to Train set, so concluded the Model to be final

Final Model

- Attached Tables depicts the final variables and the coefficient associated to each variable
- It is clear that all these variables are significant as the p value of each variables if very less
- VIF is also very low of each variable indicating that there is no collinearity among the selected variables



	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9486	0.040	-23.670	0.000	-1.027	-0.870
Do Not Email	-1.6893	0.194	-8.710	0.000	-2.069	-1.309
Lead_Origin_Lead Add Form	2.6326	0.213	12.355	0.000	2.215	3.050
Lead_Source_Welingak Website	1.9966	0.753	2.653	0.008	0.522	3.472
Last_Activity_Converted to Lead	-1.3207	0.201	-6.565	0.000	-1.715	-0.926
Last_Activity_Email Bounced	-1.6219	0.610	-2.657	0.008	-2.818	-0.425
Last_Activity_Had a Phone Conversation	2.5035	0.662	3.779	0.000	1.205	3.802
Last_Activity_Olark Chat Conversation	-1.4432	0.145	-9.978	0.000	-1.727	-1.160
Occ_Working Professional	2.8117	0.177	15.862	0.000	2.464	3.159
Last_Notable_Activity_Email Bounced	1.9402	0.770	2.520	0.012	0.431	3.449
Last_Notable_Activity_SMS Sent	1.5539	0.072	21.587	0.000	1.413	1.695
Last_Notable_Activity_Unreachable	1.4187	0.436	3.250	0.001	0.563	2.274
Last_Notable_Activity_Unsubscribed	1.7670	0.468	3.775	0.000	0.850	2.684

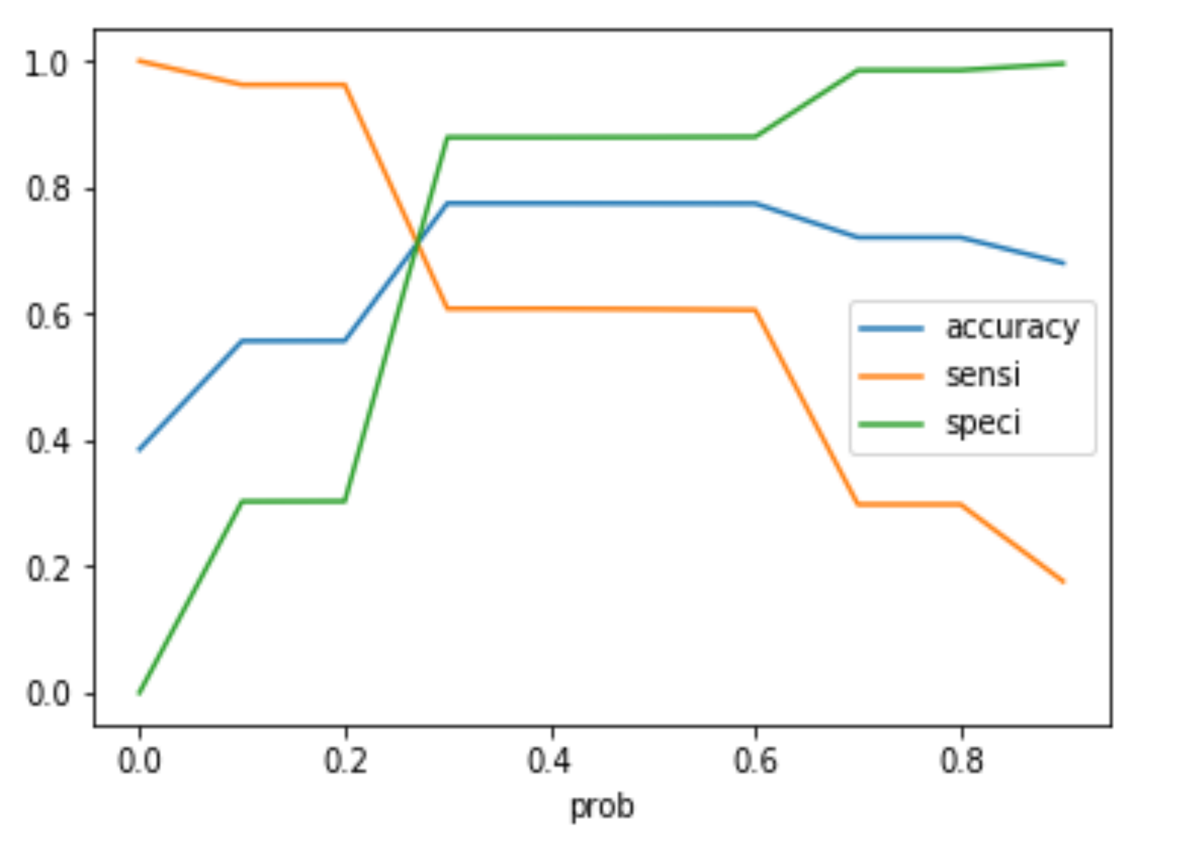


	Features	VIF
4	Last_Activity_Email Bounced	2.05
0	Do Not Email	1.93
1	Lead_Origin_Lead Add Form	1.45
2	Lead_Source_Welingak Website	1.33
8	Last_Notable_Activity_Email Bounced	1.27
7	Occ_Working Professional	1.14
9	Last_Notable_Activity_SMS Sent	1.13
11	Last_Notable_Activity_Unsubscribed	1.11
3	Last_Activity_Converted to Lead	1.00
5	Last_Activity_Had a Phone Conversation	1.00
6	Last_Activity_Olark Chat Conversation	1.00
10	Last_Notable_Activity_Unreachable	1.00

Process to chose the Optimal Cut off

- Attached graphs shows the different curve for Accuracy, Sensitivity, and Specificity
- Attached graphs gives a very good reference point of what cut off to be chosen so that all these three metrics remain within the acceptable range
- For this particular dataset, 0.28 is the cut off probability and below are the details of each metrics

Metric	Percentage
Accuracy	77%
Sensitivity	61%
Specificity	87%
False Positive Rate	13%
Positive Predicted Value	75%
Negative Predicted Value	78%
Precision Score	75%
Recall Score	61%



Choosing the right cutoff to have 80% precision

- Since the desired result is to increase the precision to at least 80%, so we need to increase the probability of the customer who can be a potential lead
- After choosing the probability as 0.65, below tables represents the value of each metric
- **Precision Score is 93% which surpasses the Goal of 80%**

Metric	Percentage
Overall Accuracy	72%
Sensitivity	30%
Specificity	99%
False Positive Rate	1%
Positive Predicted Value	93%
Negative Predicted Value	69%
Precision Score	93%

Important Variables which leads towards Hot lead

- Occupation : Working Professional
- Last Activity : Had a phone conversation
- Last Notable Activity : SMS Sent
- Lead Origin : Lead Add Form