Lead Scoring Case Study

Case Study to build a Logistic Regression model for improving the Lead Conversion Rate

Presenter: Vidhi Surana & Shubham

(Upgrad student Batch : DS C48)

Problem Statement

Domain: Education

Company: X Education

- Improve the Lead Conversion Rate of the education company called X Education that sells online courses to industry professionals

Lead Conversion Rate → Leads that convert / Total number of Leads

- When people fill up a form providing their email address or phone number, they are classified to be a lead or a potential customer.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. Main problem is how to improve the same.

Business Objectives

X Education has appointed you to help them

- identify the most promising leads(hot leads), i.e. the leads that are most likely to convert into paying customers.
- The company requires you to build a model wherein you 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%.**

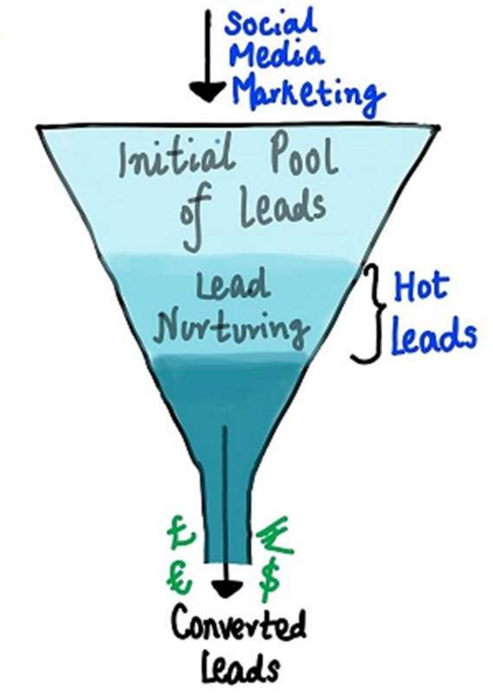
Goal of the Case Study

There are quite a few goals for this case study:

1. 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.
2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

System : Lead Conversion Process

- As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom (Converted Leads).
- In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.



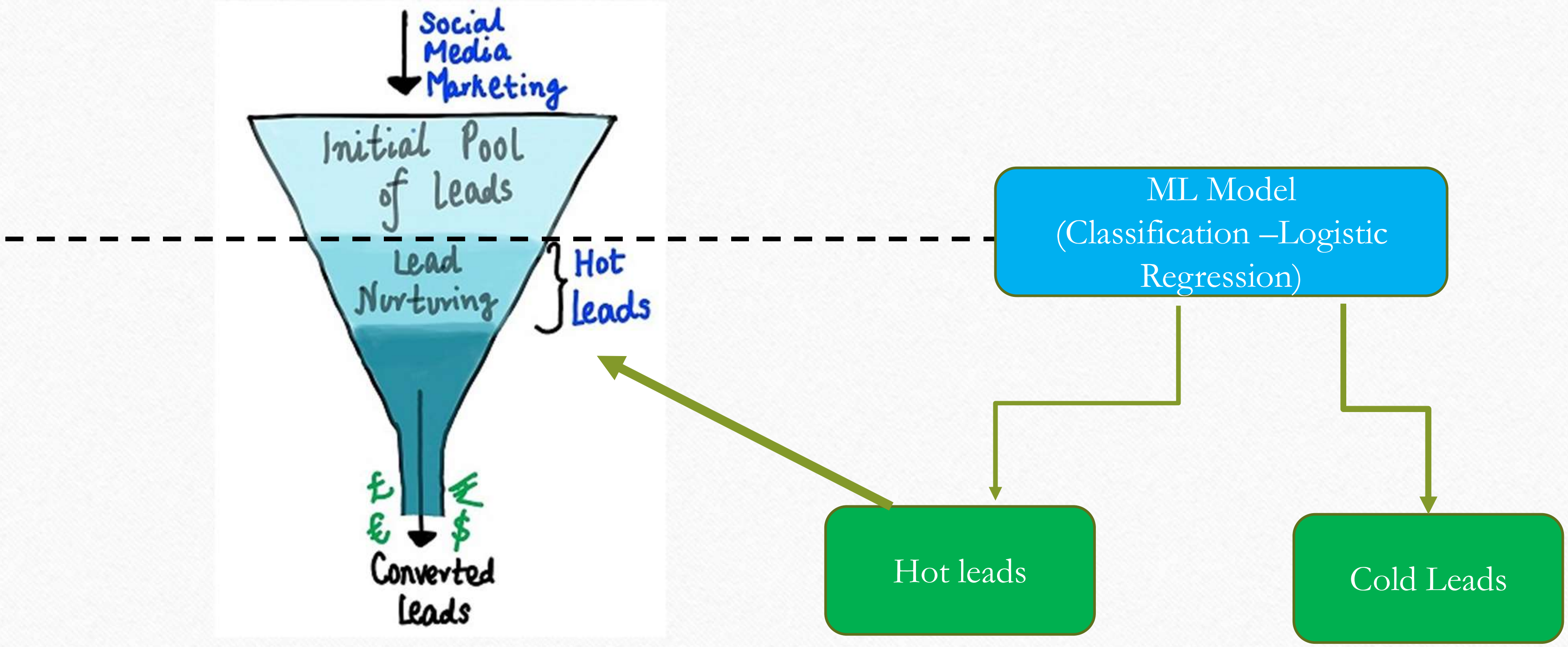
Model Approach

We will be using the following approach for this problem:

- ***The Machine Learning algorithm will work at the level between the Initial Pool of Leads and Leads Nurturing.***
- ***The filter of ML Model provides the Binary Classification Model predicting which leads have high chances of converting into lead(1 or 0)***
- ***Rather than passing all the potential customers to the sales team, only the ones that are filtered are passed.***
- ***With the help of Ml, LCR will increase and Sales team will not reach out to people having lesser chances of Converting into a lead.***

Goal : Build a Logistic Regression Model which will 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. Target lead conversion rate to be around 80%.he middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

ML Model Design



ML Model (Classification –Logistic Regression)

Hot leads

Cold Leads

Dataset

- We have been provided with a leads dataset from the past with around 9000 data points.
- This dataset consists of various attributes such as lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.
- The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

Steps for solving the problem

- Data Understanding
- Data Cleaning
  - Handling Select values, replacing them with NULL
  - Handling Missing values and outliers
  - Imputation of Missing values
  - Binary Encoding (1/0)
- Data Preparation
  - Dummy values creation
  - Train Test Split
  - Perform Scaling – Using Standard Scaler

- Univariate, Bi-variate Analysis & Multivariate analysis
- Model Building & Evaluation
  - Feature selection using RFE with output number of variables = 20
  - Build Logistic Regression Model with good sensitivity
  - Evaluating the model with metrics
  - Find out the optimal probability Cutoff
  - Model diagnosis using ROC Curve, Precision-Recall curve, and probability calibration curve
  - Use the model prediction on the test dataset and perform model evaluation for the test dataset

- Final Step:
  - After having LCR > 80%, Assign Lead Score
  - Predict Hot leads

Data Cleaning and Preparation

- **Data Cleaning**

After data cleaning, following columns were removed during univariate analysis

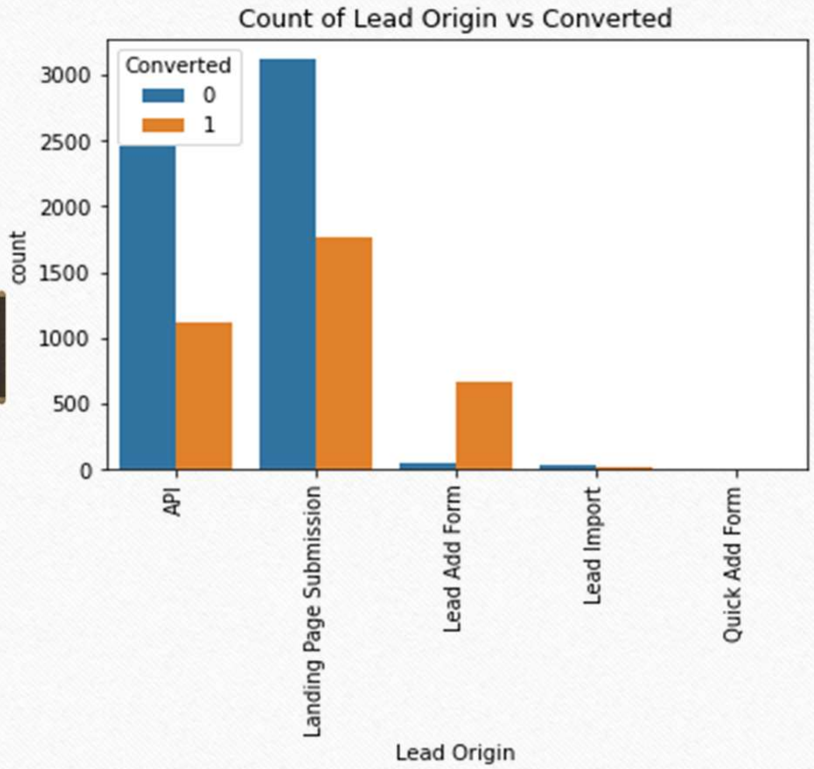
- Lead Number
- What matters most to you in choosing a course
- Receive More Updates About Our Courses
- Update me on Supply Chain Content
- Search, Magazine, Newspaper Article, X Education Forums,
- Newspaper, Digital Advertisement , Through Recommendations
- Update me on Supply Chain Content
- Get updates on DM Content
- I agree to pay the amount through cheque
- Tags, A free copy of Mastering The Interview Country

- **Data Preparation**

Dummy variables are created for

- Lead Origin
- Lead Source
- Last Activity
- Specialization
- What is your current occupation
- City
- Last Notable Activity

Analysis using graphs

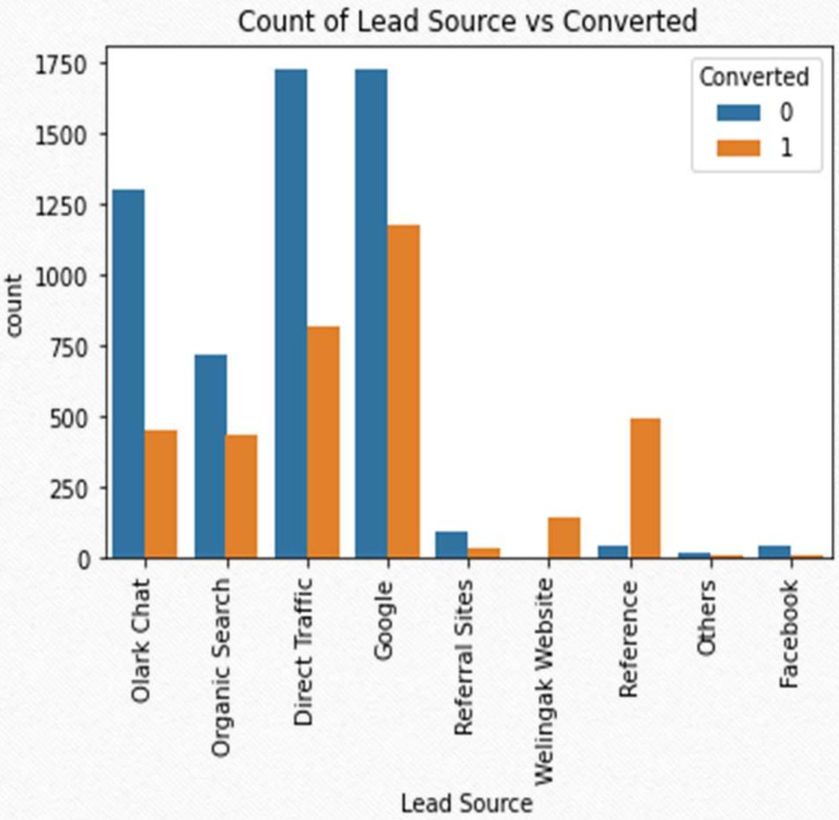


Converted is the target variable, Indicates whether a lead has been successfully converted (1) or not (0)

Lead Origin –

- Landing Page Submission and API are the two origin identifiers having almost 30-50% of conversion rate.
- Though Lead Add Form count is very less, the conversion rate is pretty high almost 90%
- Lead Import is minimal
- Quick Add form have no counts.

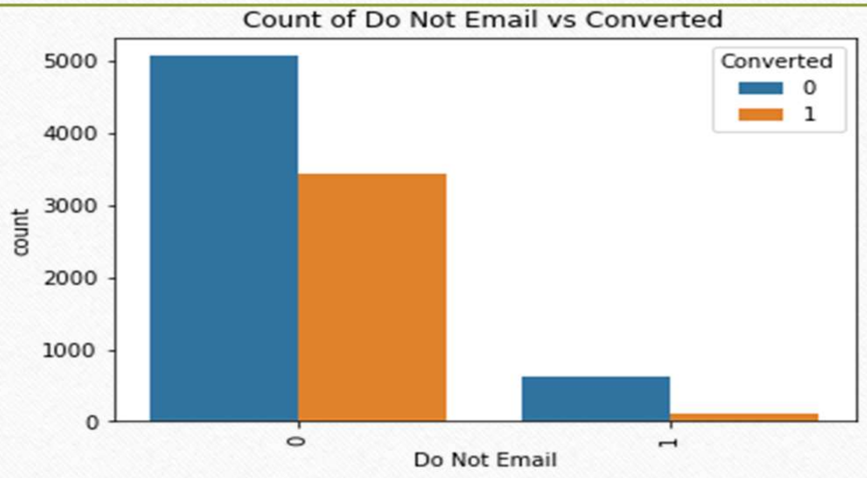
For better results, Should focus more on how to increase lead conversion rate from API and Landing Page Submission and also get more leads from Add Form.

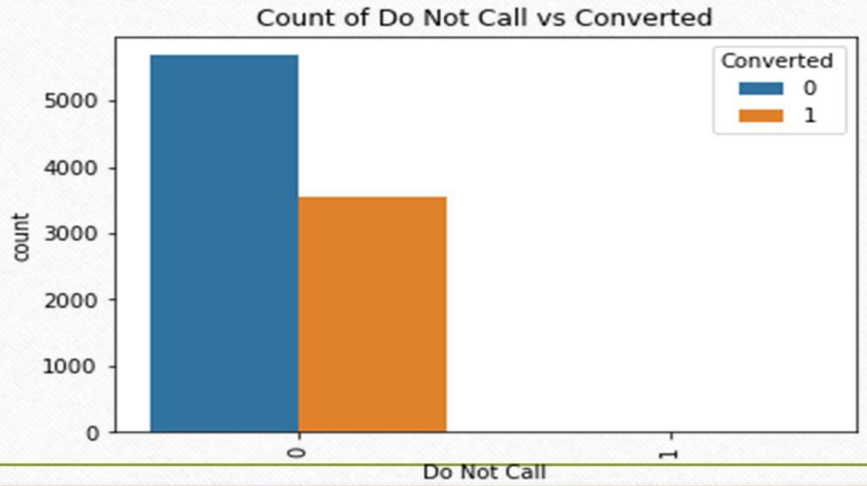


Lead Source –

- Organic Search, Google and Direct Traffic provide most number of Leads having Lead conversion rate less than 50%.
- Clients coming from Reference and Welingak Website mostly convert into a lead. There conversion rate is very high

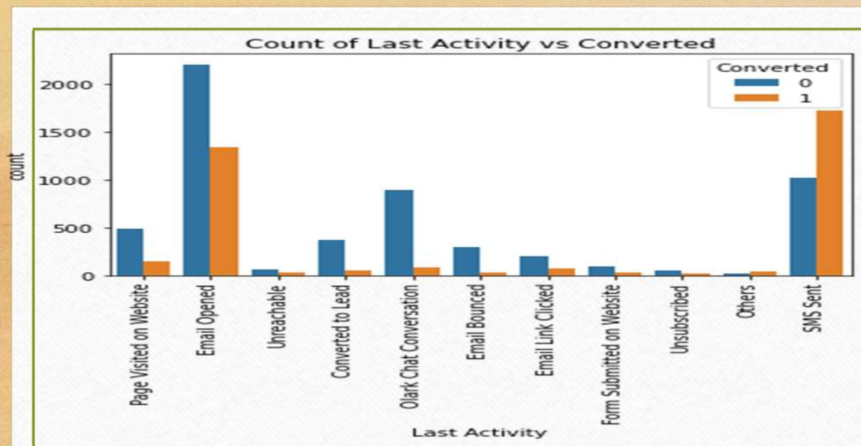
Hence, for better results need to focus on the Lead Sources coming from Reference and Welingak Website

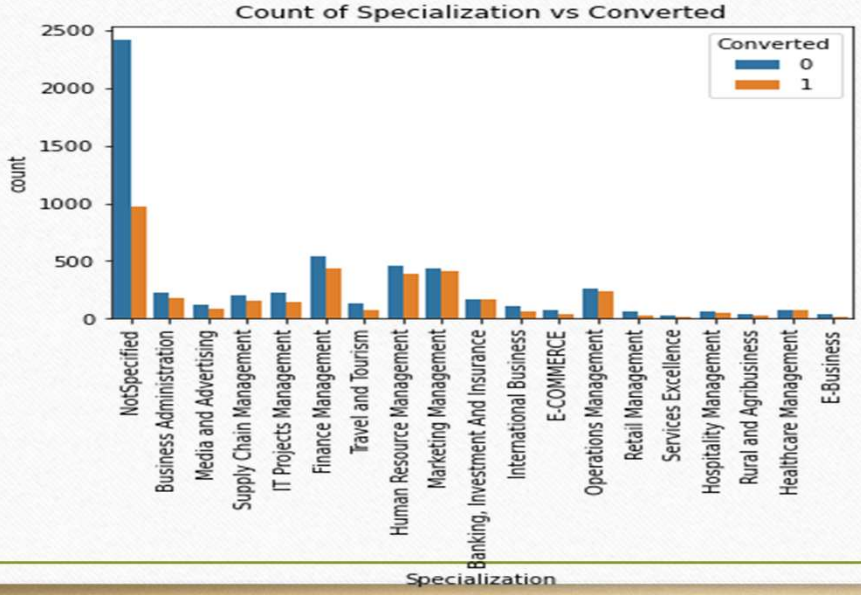




Do Not Email and Do Not Call.

- Most of the clients almost 90% say Do Not Email and Do not Call. Hence no conclusions can be drawn from this.





Last Activity –

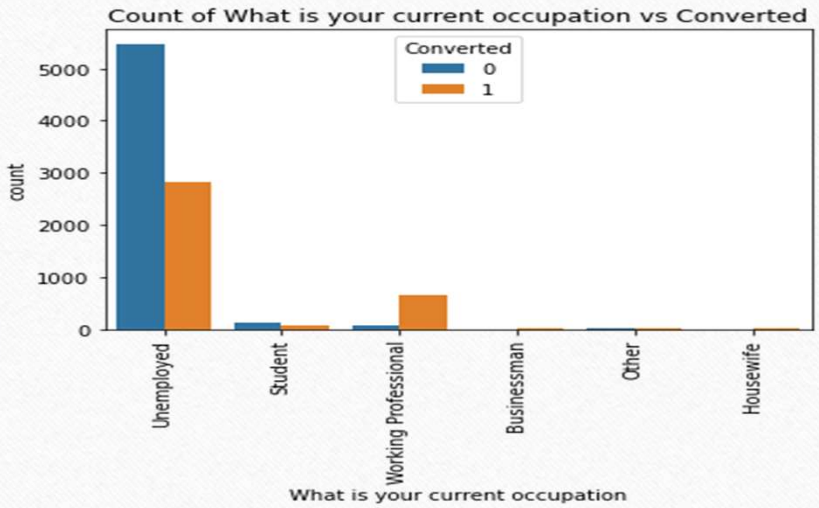
- Last Activity for most of the Clients is Email Opened.
- Last Activity of SMS Sent has highest conversion rate

Hence, for better results need to focus on people whose last activity is SMS Sent.

Specialisation –

- Most of the clients do not select this field (Not specified) – could be an undergrad student or who didn't do any specialization.
- Management people, business and Finance have good conversion rate close to 50%

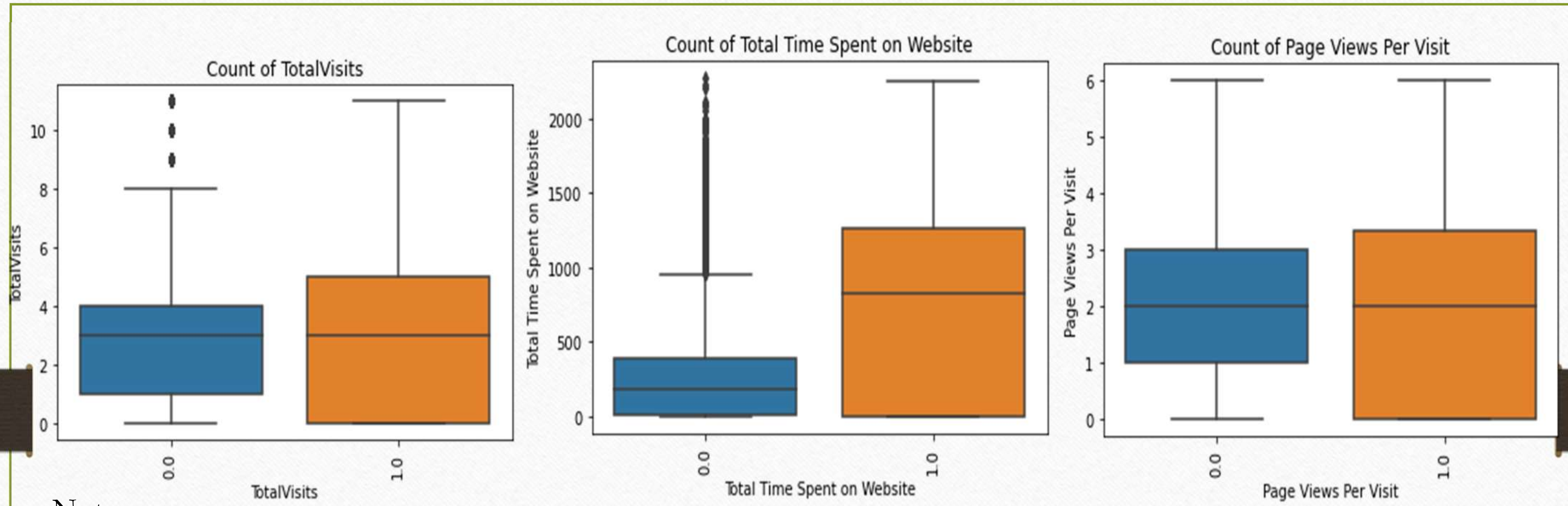
Hence, for better results need to focus more on above people that can give high conversion rate.



Current Occupation –

- Most of the clients are Unemployed.
- Working Professionals have very high conversion rate.

Hence, for better results need to focus on people who are Unemployed and more on working professionals.

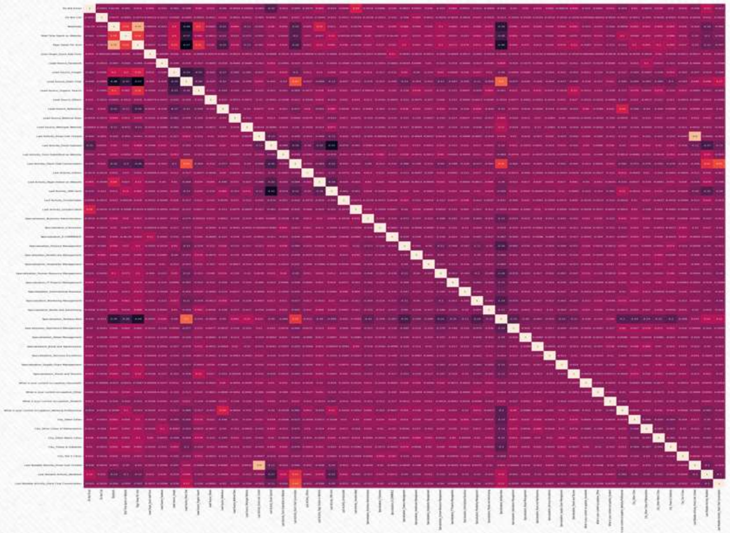


Note:

1. By Total Visits and Pages Views per Visit, no inference can be drawn as median is same for Converted = 0 or 1
2. By Total Time Spent on Website, we can infer that clients who spent more time on the website convert into the lead

Hence , for better results designing the website in such a way that clients spent more time on the website (like can put videos of alumni or quotes) will help.

Heatmap



Columns Removed after seeing Correlation in heatmap

Lead Origin_Landing Page Submission

Last Notable Activity_Email Opened

Lead Origin_Lead Add Form

Lead Origin_Lead Import

What is your current occupation_Unemployed

Last Notable Activity_SMS Sent

Last Activity_Email Bounced

Last Notable Activity_Page Visited on Website

Last Notable Activity_Others

Last Notable Activity_Unsubscribed

Last Notable Activity_Unreachable

Building Logistic Regression Model

- $\text{Sensitivity} = \dfrac{\text{Number of Actual Yeses correctly predicted}}{\text{Total Number of Actual Yeses}}$

We need to select model with good Sensitivity or Recall to make sure maximum leads are correctly identified as converted

By experimentation, its seen that as the

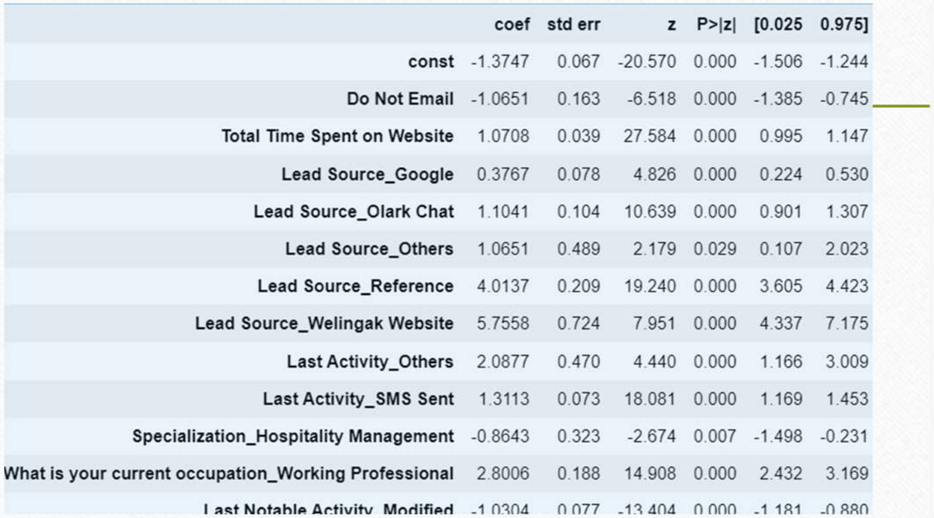
- Optimal Threshold value increases , Sensitivity decreases and specificity increases
- Optimal Threshold value decreases, Sensitivity increases and specificity decreases

Steps in Logistic Regression

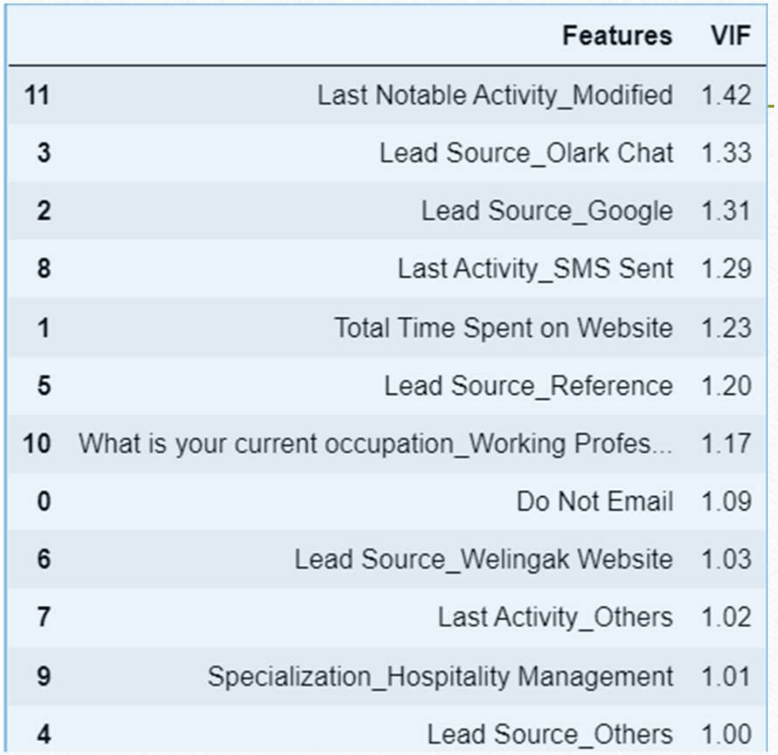
Model Building & Evaluation

- Feature selection using RFE
- Build Logistic Regression Model with good sensitivity
- Evaluating the model with metrics (Accuracy, Precision and Recall)
- Find out the optimal probability Cutoff using Probability curve.
- Model diagnosis using ROC Curve, Precision-Recall curve and probability calibration curve
- Adjusting the optimal threshold to increase or decrease the sensitivity and specificity.

P value and VIF of Final Model

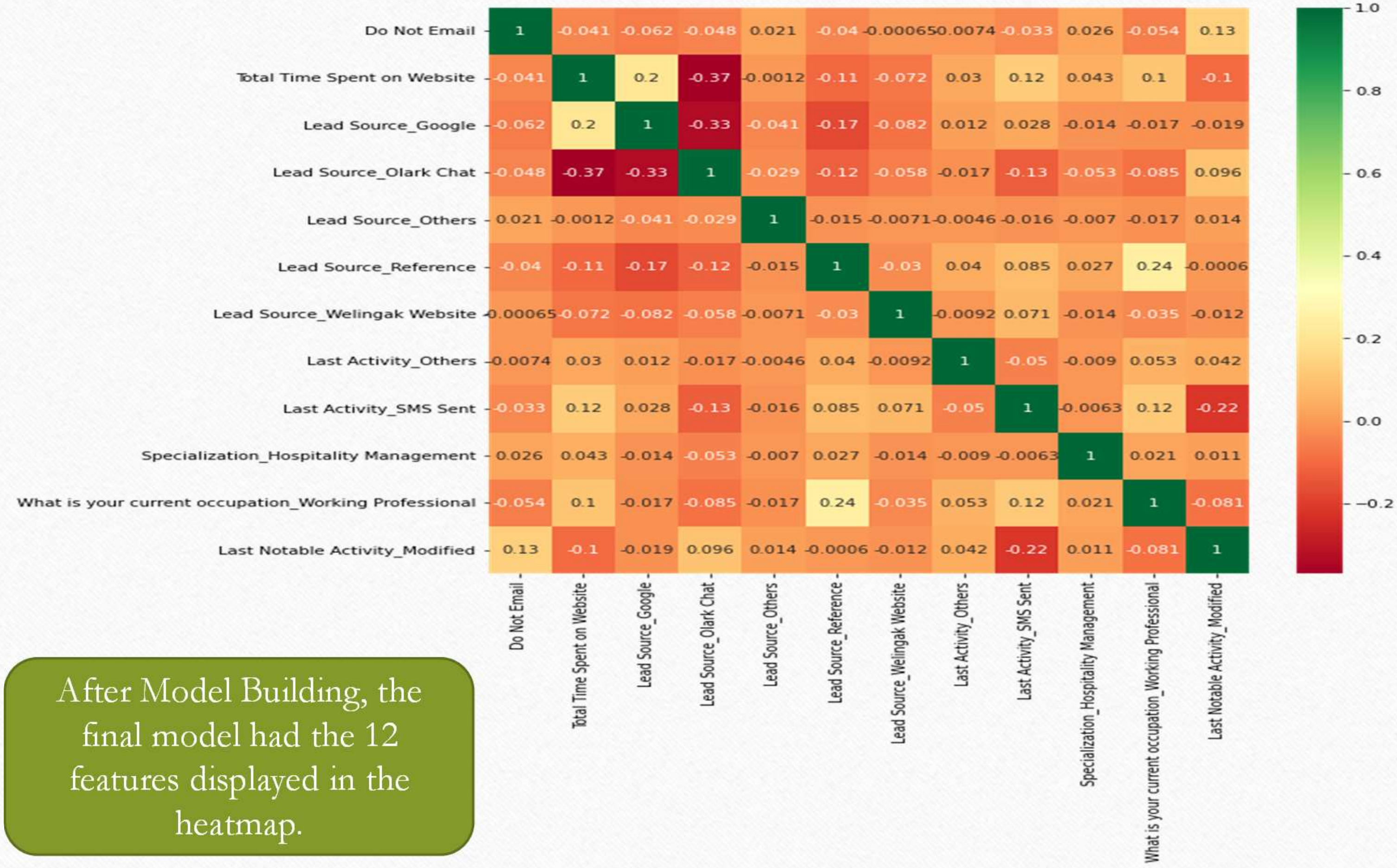


	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3747	0.067	-20.570	0.000	-1.506	-1.244
Do Not Email	-1.0651	0.163	-6.518	0.000	-1.385	-0.745
Total Time Spent on Website	1.0708	0.039	27.584	0.000	0.995	1.147
Lead Source_Google	0.3767	0.078	4.826	0.000	0.224	0.530
Lead Source_Olark Chat	1.1041	0.104	10.639	0.000	0.901	1.307
Lead Source_Others	1.0651	0.489	2.179	0.029	0.107	2.023
Lead Source_Reference	4.0137	0.209	19.240	0.000	3.605	4.423
Lead Source_Welingak Website	5.7558	0.724	7.951	0.000	4.337	7.175
Last Activity_Others	2.0877	0.470	4.440	0.000	1.166	3.009
Last Activity_SMS Sent	1.3113	0.073	18.081	0.000	1.169	1.453
Specialization_Hospitality Management	-0.8643	0.323	-2.674	0.007	-1.498	-0.231
What is your current occupation_Working Professional	2.8006	0.188	14.908	0.000	2.432	3.169
Last Notable Activity_Modified	-1.0304	0.077	-13.404	0.000	-1.181	-0.880



	Features	VIF
11	Last Notable Activity_Modified	1.42
3	Lead Source_Olark Chat	1.33
2	Lead Source_Google	1.31
8	Last Activity_SMS Sent	1.29
1	Total Time Spent on Website	1.23
5	Lead Source_Reference	1.20
10	What is your current occupation_Working Profes...	1.17
0	Do Not Email	1.09
6	Lead Source_Welingak Website	1.03
7	Last Activity_Others	1.02
9	Specialization_Hospitality Management	1.01
4	Lead Source_Others	1.00

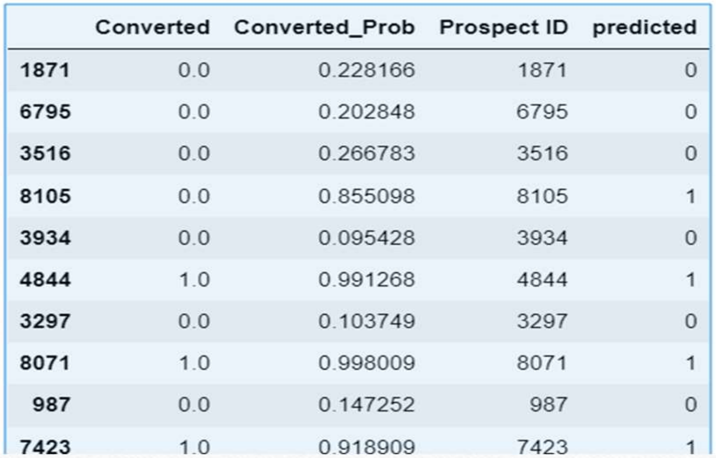
All p values < 0.05 and VIF < =5, hence



	Do Not Email	Total Time Spent on Website	Lead Source_Google	Lead Source_Olark Chat	Lead Source_Others	Lead Source_Reference	Lead Source_Welingak Website	Last Activity_Others	Last Activity_SMS Sent	Specialization_Hospitality Management	What is your current occupation_Working Professional	Last Notable Activity_Modified
Do Not Email	1	-0.041	-0.062	-0.048	0.021	-0.04	-0.00065	-0.0074	-0.033	0.026	-0.054	0.13
Total Time Spent on Website	-0.041	1	0.2	-0.37	-0.0012	-0.11	-0.072	0.03	0.12	0.043	0.1	-0.1
Lead Source_Google	-0.062	0.2	1	-0.33	-0.041	-0.17	-0.082	0.012	0.028	-0.014	-0.017	-0.019
Lead Source_Olark Chat	-0.048	-0.37	-0.33	1	-0.029	-0.12	-0.058	-0.017	-0.13	-0.053	-0.085	0.096
Lead Source_Others	0.021	-0.0012	-0.041	-0.029	1	-0.015	-0.0071	-0.0046	-0.016	-0.007	-0.017	0.014
Lead Source_Reference	-0.04	-0.11	-0.17	-0.12	-0.015	1	-0.03	0.04	0.085	0.027	0.24	-0.0006
Lead Source_Welingak Website	-0.00065	-0.072	-0.082	-0.058	-0.0071	-0.03	1	-0.0092	0.071	-0.014	-0.035	-0.012
Last Activity_Others	-0.0074	0.03	0.012	-0.017	-0.0046	0.04	-0.0092	1	-0.05	-0.009	0.053	0.042
Last Activity_SMS Sent	-0.033	0.12	0.028	-0.13	-0.016	0.085	0.071	-0.05	1	-0.0063	0.12	-0.22
Specialization_Hospitality Management	0.026	0.043	-0.014	-0.053	-0.007	0.027	-0.014	-0.009	-0.0063	1	0.021	0.011
What is your current occupation_Working Professional	-0.054	0.1	-0.017	-0.085	-0.017	0.24	-0.035	0.053	0.12	0.021	1	-0.081
Last Notable Activity_Modified	0.13	-0.1	-0.019	0.096	0.014	-0.0006	-0.012	0.042	-0.22	0.011	-0.081	1

After Model Building, the final model had the 12 features displayed in the heatmap.

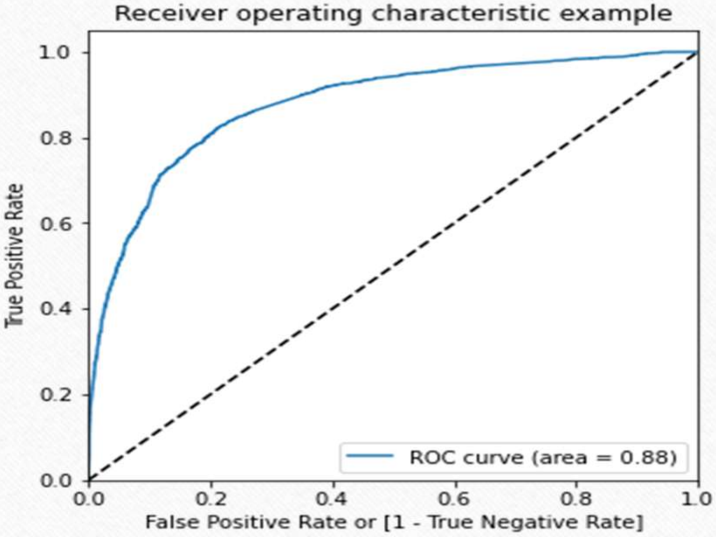
Creating a dataframe with the actual Converted variable and the predicted probabilities with predicted values 'predicted' with 1 if Converted_Prob > 0.5 else 0



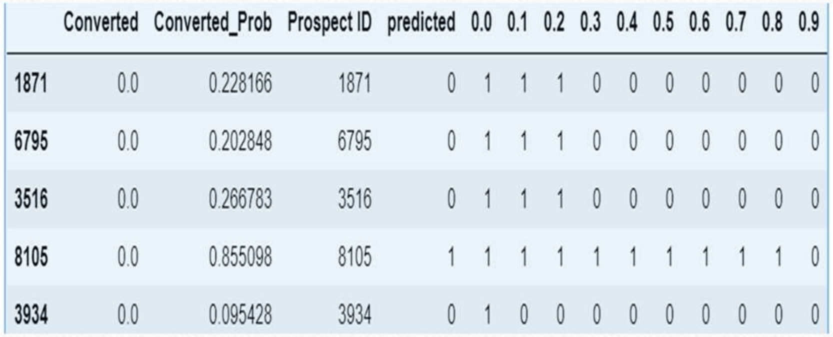
	Converted	Converted_Prob	Prospect ID	predicted
1871	0.0	0.228166	1871	0
6795	0.0	0.202848	6795	0
3516	0.0	0.266783	3516	0
8105	0.0	0.855098	8105	1
3934	0.0	0.095428	3934	0
4844	1.0	0.991268	4844	1
3297	0.0	0.103749	3297	0
8071	1.0	0.998009	8071	1
987	0.0	0.147252	987	0
7423	1.0	0.918909	7423	1

Next is Plotting an ROC curve demonstrates several things:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



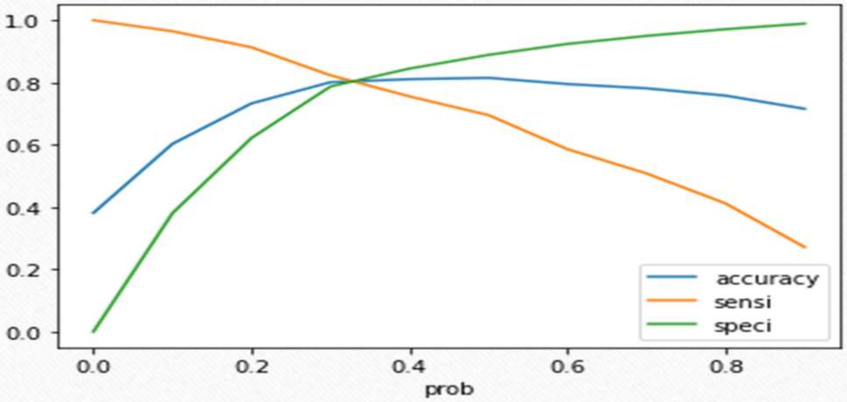
Probability Curve



	Converted	Converted_Prob	Prospect ID	predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1871	0.0	0.228166	1871	0	1	1	1	0	0	0	0	0	0	0
6795	0.0	0.202848	6795	0	1	1	1	0	0	0	0	0	0	0
3516	0.0	0.266783	3516	0	1	1	1	0	0	0	0	0	0	0
8105	0.0	0.855098	8105	1	1	1	1	1	1	1	1	1	1	0
3934	0.0	0.095428	3934	0	1	0	0	0	0	0	0	0	0	0

Creating columns with different probability cutoffs and calculating the accuracy sensitivity and specificity for various probability cutoffs.

Draw a probability curve using the above



From the curve above, 0.32 is the optimum point to take it as a cutoff probability.

Optimal cutoff probability is that prob where we get balanced sensitivity and specificity

Metrics value for Train and Test Data

Train Data

- Accuracy – 80.12%
- Sensitivity – 81%
- Specificity – 79.59%

Test Data

- Accuracy - 81%
- Sensitivity – 81.64%
- Specificity – 80.32 %

The CEO had has given a ballpark of the target lead conversion rate to be around 80%. The Model seems to have achieved that.
Selection of the model here is on the basis of Sensitivity/Recall which is above 80%.

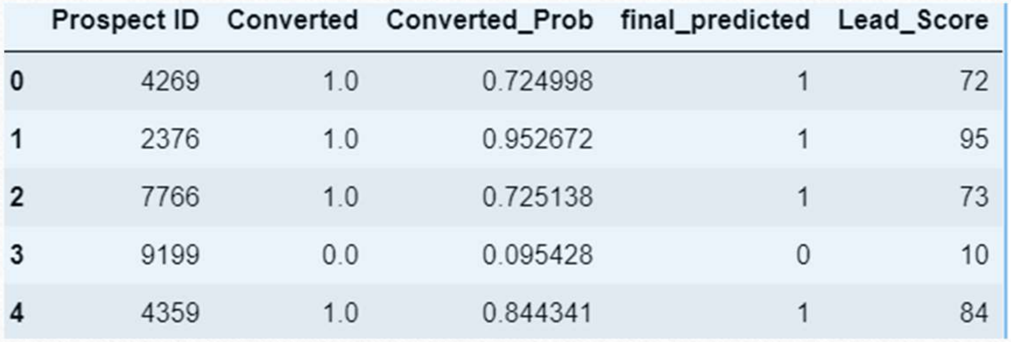
Making Predictions on the test Data

- The final model on the train dataset is used to make
- predictions for the test dataset
- The train data set was scaled using the scaler.transform
- The Predicted probabilities were added to the leads in the test dataframe.
- Using the probability threshold value of 0.32, the leads from the test dataset were predicted if they will convert or not.
- **NOTE : This threshold value can be adjusted to get more sensitivity (decrease the optimal threshold) or get more specificity(increase the optimal threshold)**

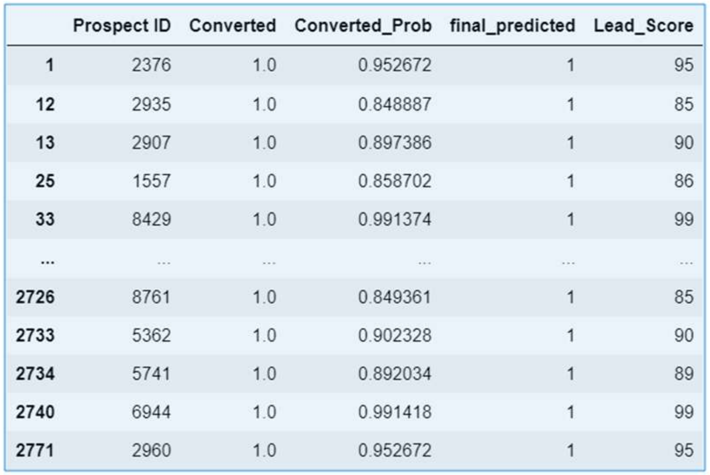
Lead Score Calculation

Lead Score is calculated by the following formula:

Lead Score for each lead = Lead Conversion Probability for the lead X 100



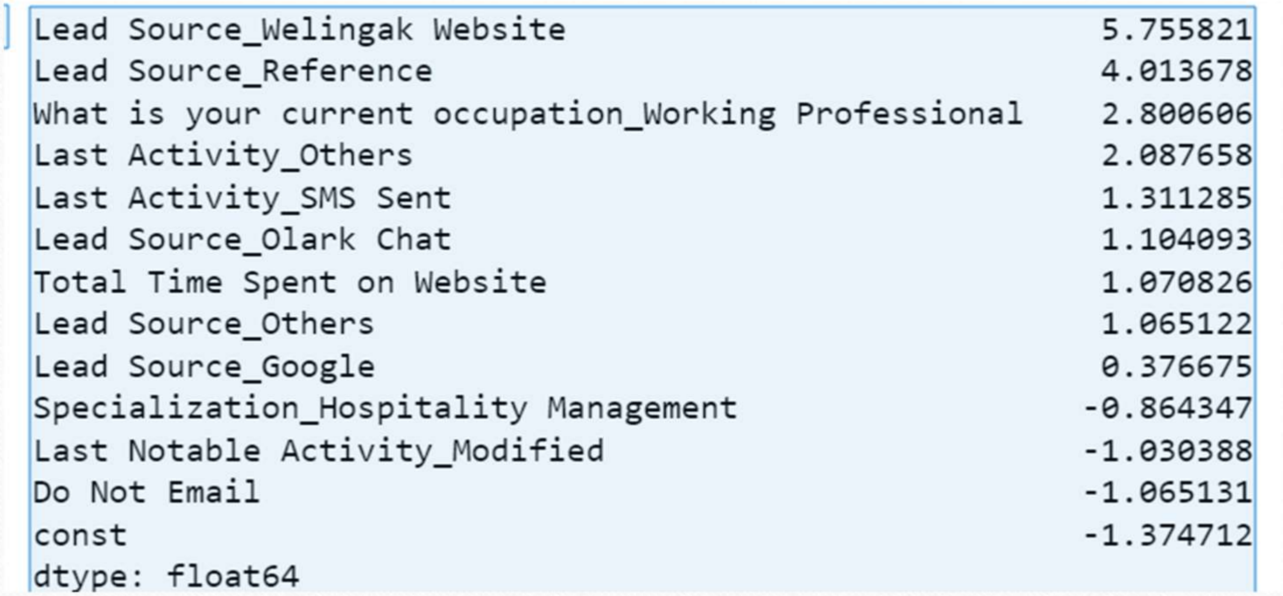
	Prospect ID	Converted	Converted_Prob	final_predicted	Lead_Score
0	4269	1.0	0.724998	1	72
1	2376	1.0	0.952672	1	95
2	7766	1.0	0.725138	1	73
3	9199	0.0	0.095428	0	10
4	4359	1.0	0.844341	1	84



	Prospect ID	Converted	Converted_Prob	final_predicted	Lead_Score
1	2376	1.0	0.952672	1	95
12	2935	1.0	0.848887	1	85
13	2907	1.0	0.897386	1	90
25	1557	1.0	0.858702	1	86
33	8429	1.0	0.991374	1	99
...	...	...	...	...	...
2726	8761	1.0	0.849361	1	85
2733	5362	1.0	0.902328	1	90
2734	5741	1.0	0.892034	1	89
2740	6944	1.0	0.991418	1	99
2771	2960	1.0	0.952672	1	95

Hot Leads are shown above whose Lead Score > 85
(375 Hot leads were identified)

Final Feature List with coefficients



```
Lead Source_Welingak Website                             5.755821
Lead Source_Reference                                    4.013678
What is your current occupation_Working Professional     2.800606
Last Activity_Others                                     2.087658
Last Activity_SMS Sent                                   1.311285
Lead Source_Olark Chat                                   1.104093
Total Time Spent on Website                              1.070826
Lead Source_Others                                       1.065122
Lead Source_Google                                       0.376675
Specialization_Hospitality Management                   -0.864347
Last Notable Activity_Modified                          -1.030388
Do Not Email                                            -1.065131
const                                                   -1.374712
dtype: float64
```

Recommendations

1. The X Education company should be contacting / calling those leads whose **Lead Source** is **'Reference' or 'Wellingak Website',** since they are more likely to convert.
2. The company should contact more of the leads who are **Working Professionals**. Since, they are more likely to convert.
3. The company should contact more of the leads whose **Last Activity** is **SMS Sent or Others**. Since, they are more likely to convert.
4. The company should be contacting / calling those leads whose **Lead Source** is **'Olark Chat, Google and others'** since they are more likely to convert.
5. The company should be calling those leads for whom **Total Time Spent on Website** is high, meaning leads who spent more time on the company's website. Since, they are more likely to convert.
6. The company should not be calling those leads who have their **specialization** in Hospital Management. Since, they are not likely to convert.
7. The company should not be calling those leads whose **Last Notable Activity as 'Modified'**. Since, they are not likely to convert.
8. The company should not be calling those leads who have chosen the option of **Do Not Email as "yes"** as they are less likely to convert.

Recommendations Contd.

In case of problems where we need to aggressively contact the leads that are going to convert

So, Sensitivity is a measure of predicting yes when its actually yes. It is defined as

$$\text{Sensitivity} = \frac{\text{Number of Actual Yeses correctly predicted}}{\text{Total Number of Actual Yeses}}$$

- As the optimal threshold decreases, sensitivity increases and specificity decreases and vice versa.
- To ensure that almost all the potential leads are converted we need to ensure the sensitivity is more and for that we will have to adjust the threshold. If we experiment and *decrease the optimal threshold,* sensitivity keeps on increasing. In turn increasing the number of hot leads identified.

In case of problems where we need not waste time at all in contacting the leads that are not going to convert

- For this, we should make sure that the specificity of the model is high,

$$\text{Specificity} = \frac{\text{Number of actual Nos correctly predicted}}{\text{Total number of actual Nos}}$$

This can happen only when we *increase the optimal threshold*, which in turn decreases the sensitivity and increases the specificity. This will make sure that people that will not convert will not get selected at all.

Reference

- Model Building & Case Study live sessions – Upgrad
- Logistic Regression modelling Theory – Upgrad
- Logistic Regression Telecom Churn Study - Upgrad

Thank You