Lead Scoring Case Study – X Education

Logistic Regression

Group Members :

1. **Snehal Bhosle**
2. **Swathi Kommana**
3. **Yatin Kanchan**
4. **Tapan Apte**

Date : 3rd March 2019

Abstract

An education company named X Education sells online courses to industry professionals. X education wants to identify the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company wants to build a model wherein they assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Business Objectives of Data Analysis

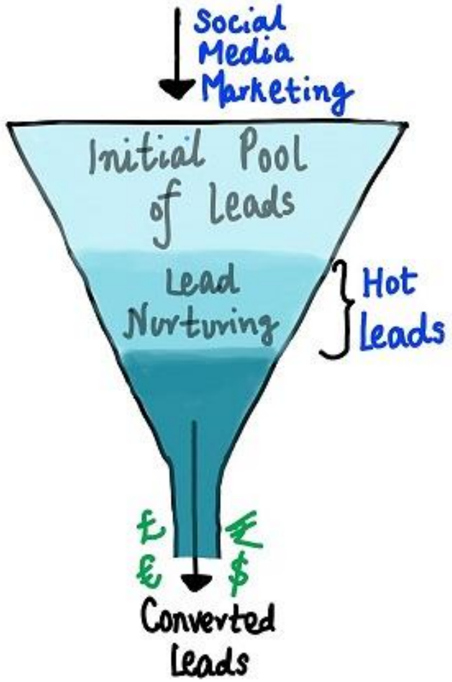
The CEO has given a ballpark of the target lead conversion rate to be around 80% which today is 30%.

Strategy

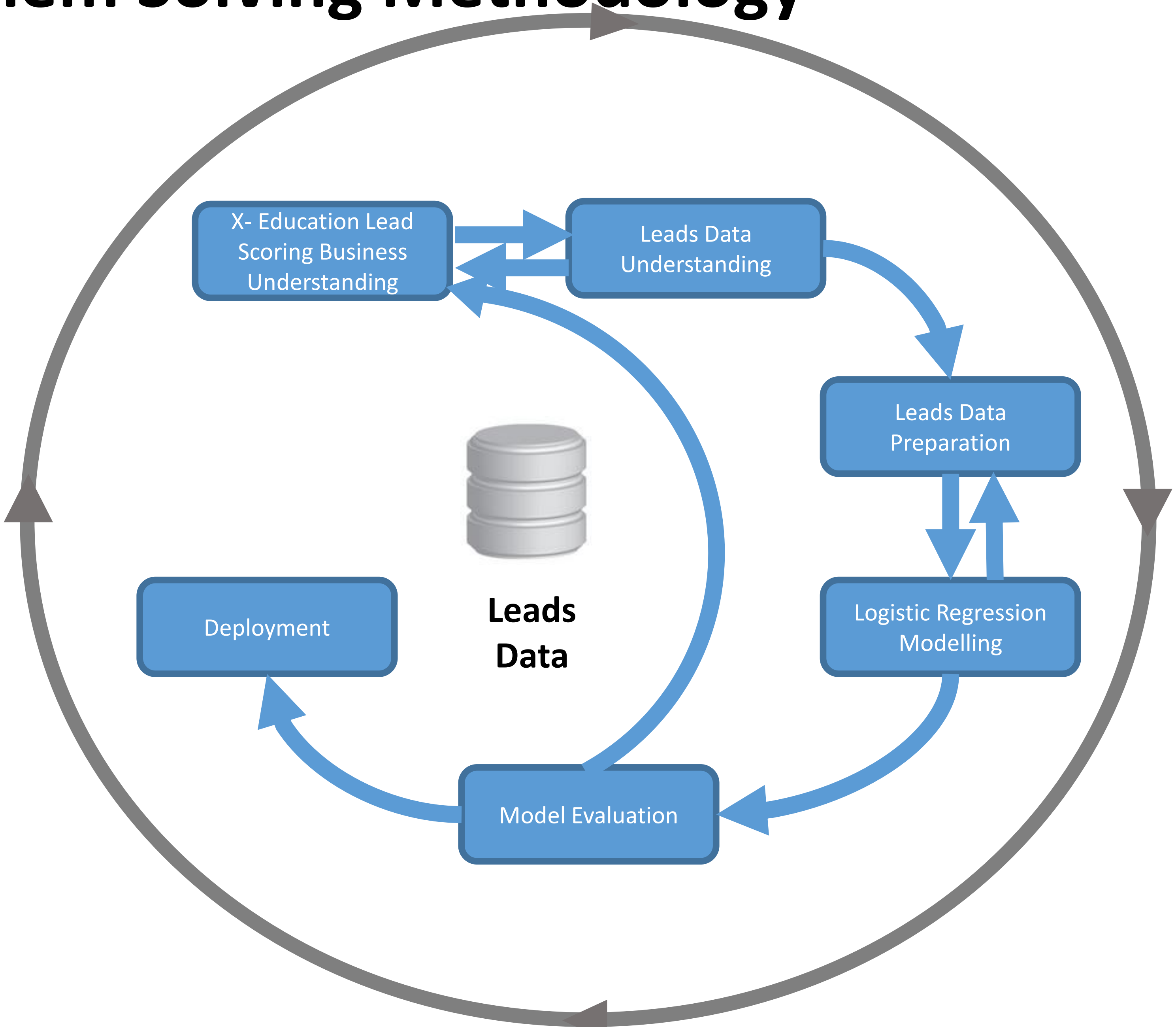
Make this process more efficient by successfully identifying 'Hot Leads' so that the lead conversion rate goes up and the sales team can focus more on communicating with the potential leads rather than making calls to everyone.

Goals of Data Analysis

- Build a 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 lead.
- The model should be able to adjust such that it is able to handle company's changing requirements in the future.



Problem Solving Methodology



X- Education Lead Scoring Business Understanding

Leads Data Understanding

Leads Data Preparation

Logistic Regression Modelling

Model Evaluation

Deployment

Leads Data

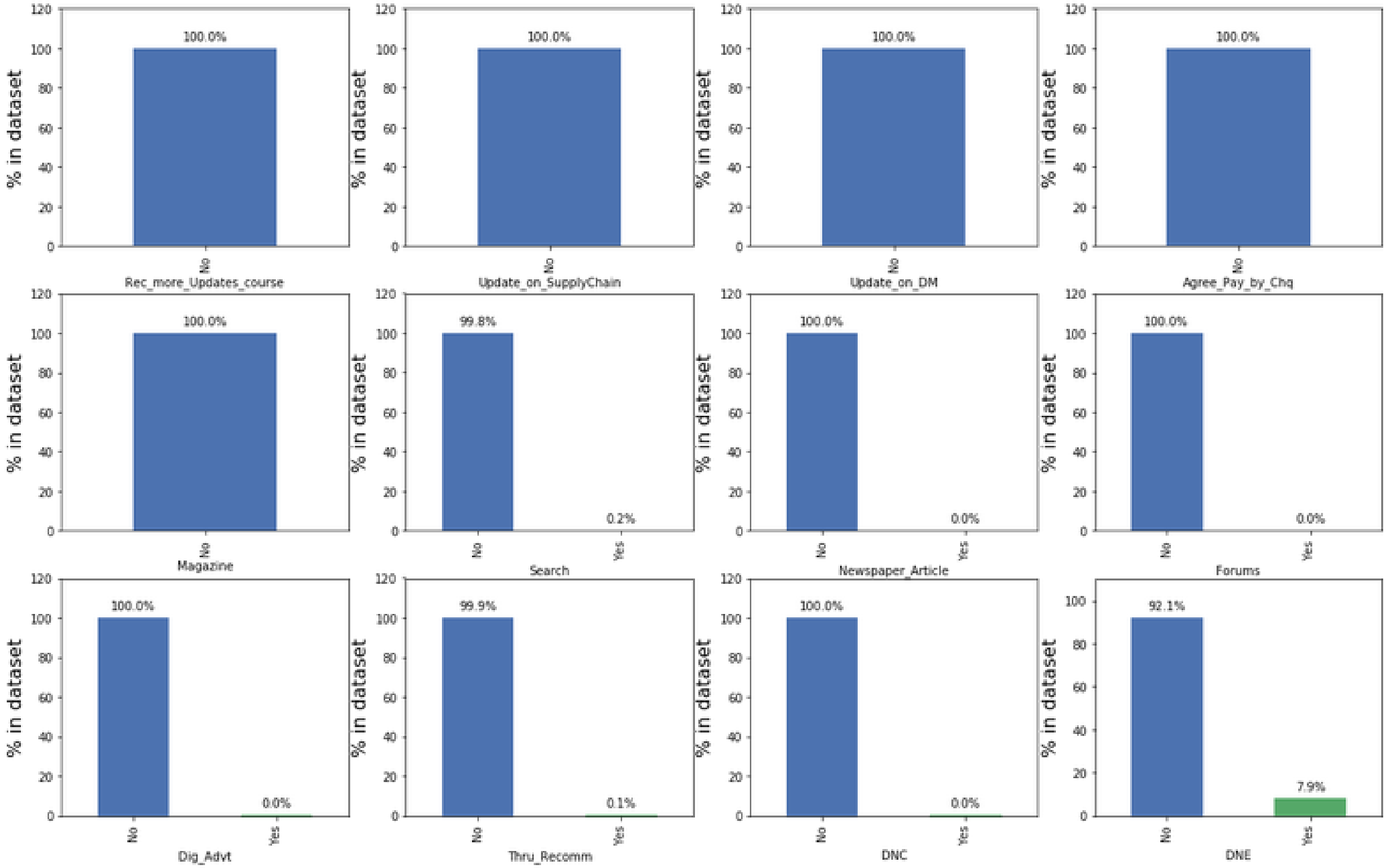
Data Inspection

Observation :

1) Total # of Records : **9240**

2) List of Fields that had below 0.5% distinct values are :

- Receive More Updates About Our Courses.
- Update me on Supply Chain Content
- Get updates on DM Content
- I agree to pay the amount through cheque
- Magazine
- Search
- Newspaper Article
- X Education Forums
- Digital Advertisement
- Through Recommendations
- Do Not Email
- Do Not Call



Recommendation : Hence it is recommended to drop the above the listed columns from the Data.

Data Preparation

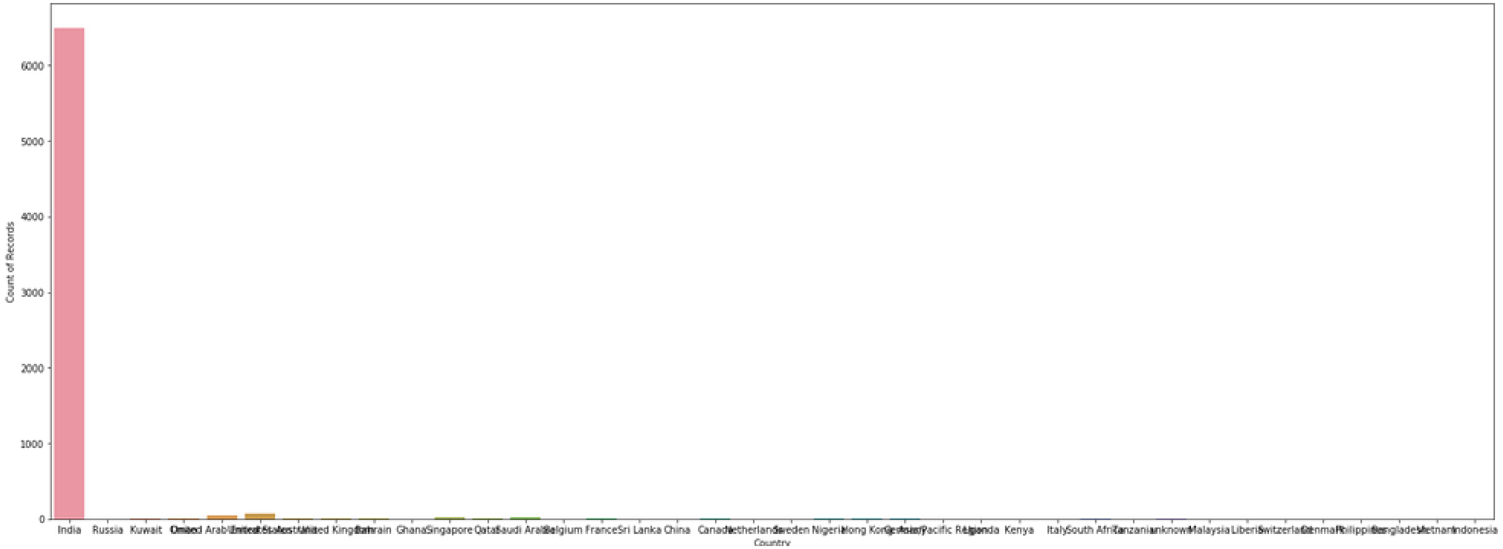
Observation

1) Below listed fields had many “Select” as values. They appear to be values that were not correctly entered when the data were captured from the UI.

Variable	Count of Records containing 'Select' values
Lead Profile	4146
How did you hear about X Education	5034
City	2249
Specialization	1942

Recommendation :
It is recommended to replace those “Select” values with NULL (blanks).

2) Most of the Leads appear to be from “India”. The count of Leads were too low when compared with other Countries



Recommendation :
It is suggested to Impute the Null values with “India” for the records with blank Countries

Data Preparation Contd…

Observation

1) There were many Fields with more than **3000** values being null.
2) Total Visits has Outlier values

Recommendation :

- It is recommended to drop those columns.

  Below are the list of Fields that were dropped

  - Lead Profile
  - Lead Quality
  - How did you hear about X education
  - Asymmetric Activity Index
  - Asymmetric Activity Score
  - Asymmetric Profile Index
  - Asymmetric Profile Score

- Rows with Total Visits Outlier values can be excluded from our Model.

Fields	NULL value Count
Lead_Num	**0**
Lead_Origin	**0**
Lead_Source	**36**
DNE	**0**
Converted	**0**
TotalVisits	**137**
Tot_Time_Website	**0**
Page_Views_per_Visit	**137**
Last_Activity	**103**
Country	**0**
Specialization	**3380**
How_did_you_know	**7250**
Current_Occup	**2690**
Matters_Most	**2709**
Tags	**3353**
Lead_Quality	**4767**
Lead_Profile	**6855**
City	**3669**
Asymm_Activity_Idx	**4218**
Asymm_Profile_Idx	**4218**
Asymm_Activity_Score	**4218**
Asymm_Profile_Score	**4218**
Free_copy_Intvw	**0**
Last_Notable_Activity	**0**

Data Preparation Contd…

Observation

1) Post Null Treatment below listed Fields continued to still have Null values.

Fields	NULL value Count
Lead_Num	0
Lead_Origin	0
Lead_Source	23
DNE	0
Converted	0
TotalVisits	0
Tot_Time_Website	0
Country	0
Specialization	1373
Current_Occup	77
Tags	0
Free_copy_Intvw	0
Last_Notable_Activity	0

Recommendation :

- Categorical Fields like Lead Source, Specialization and Current Occupation can be imputed with Mode values for each of those fields.
- Segment Countries other than India under “Others” category.
- To proceed further for Modelling create “Hot Encoded” dummy variables.

Data Retained

63%

Conversion Rate

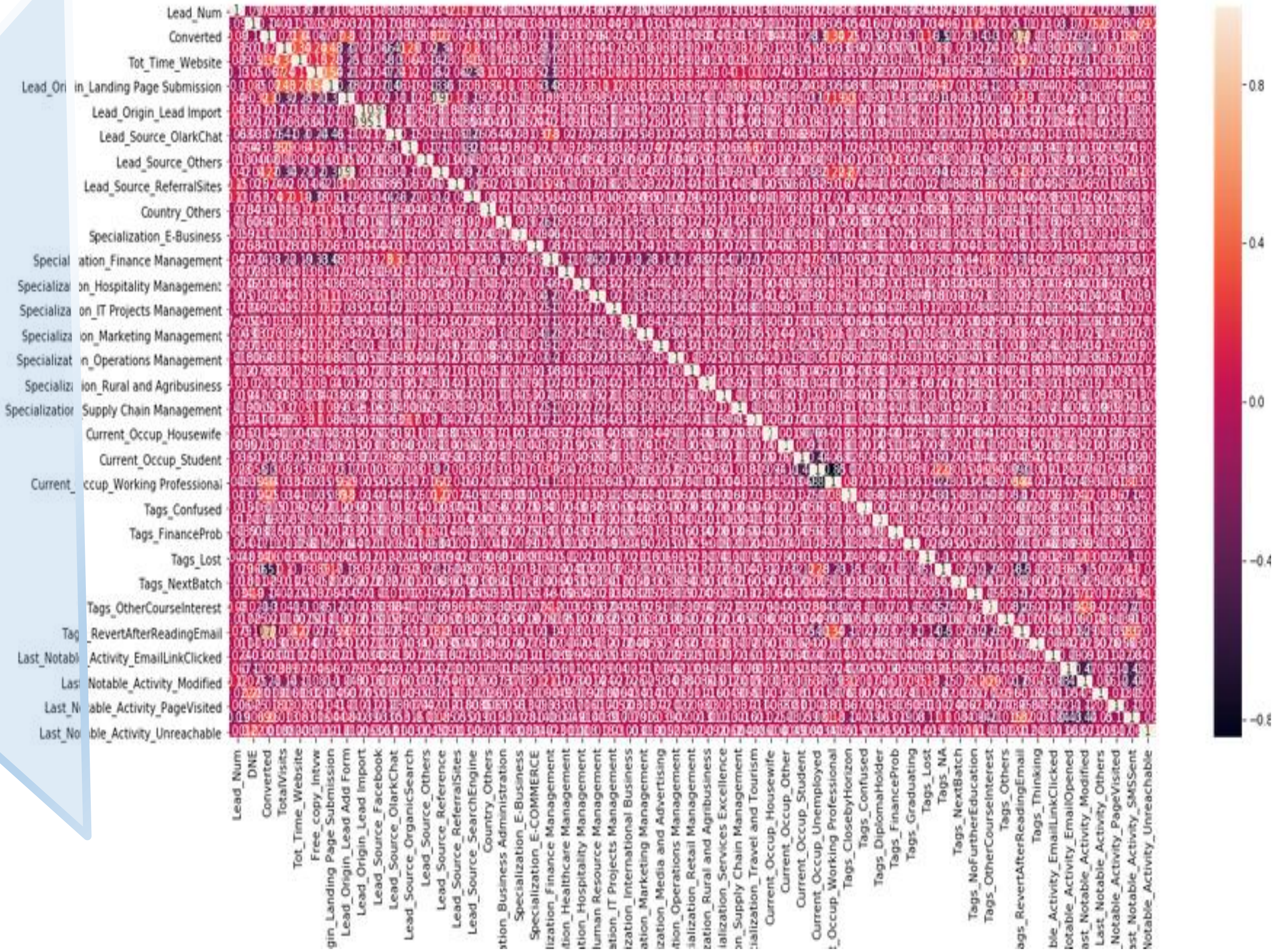
46%

Correlation

Observation

1. Post Feature Scaling below are the Top 4 fields that has High Correlation

Dummy Variables	Dummy Variables	Absolute Correlation
Lead_Origin_Lead Import	Lead_Source_Facebook	95%
Lead_Origin_Lead Add Form	Lead_Source_Reference	93%
Current_Occup_Unemployed	Current_Occup_Working Professional	85%
Converted	Tags_RevertAfterReading Email	77%



Recommendation :

Drop **Lead Source as Facebook**, **Lead Source as Reference**, **Current Occupation as Working** and **Tags having Revert After Reading Email**.

Model Building

- For the Logistic Regression, we are using **70% of Train Data** for Building the Model.
- Remaining **30% of data** is on which we will test our Model.
- Building the Model on 54 variables may not provide Accurate Lead Conversion score. Hence we use **Recursive Feature Elimination (RFE)** to build our Model on reduced variables.
- Below are the List of Variables Identified post RFE treatment.

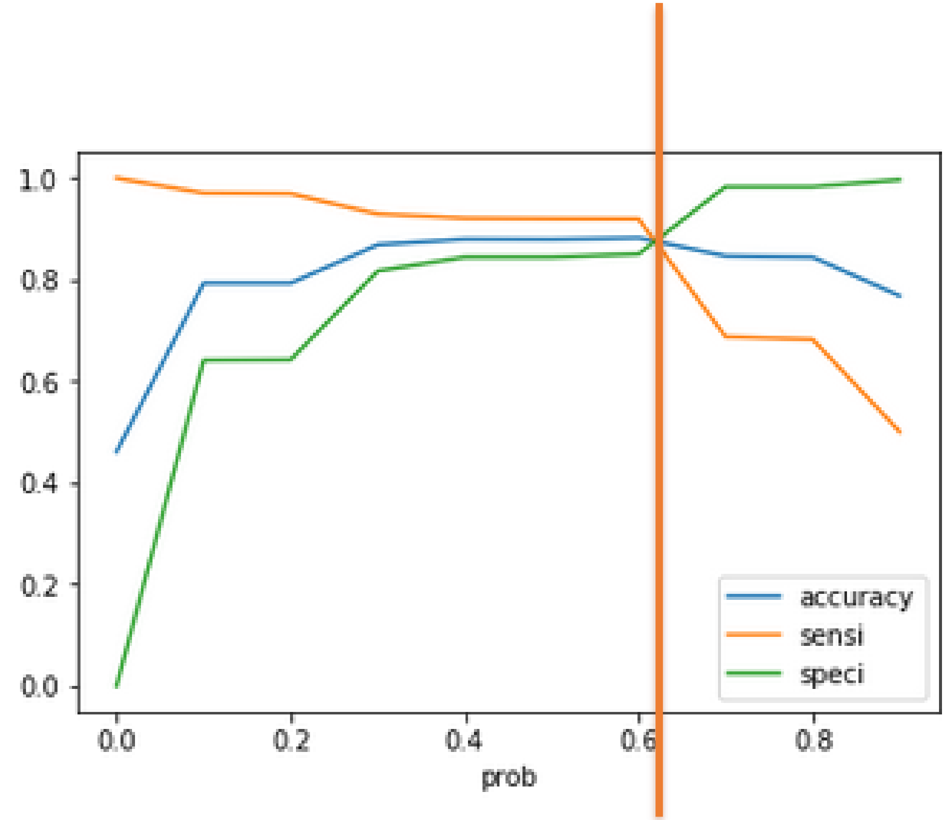
VIF < 5

P < 5%

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.1634	0.225	9.618	0	1.723	2.604
DNE	-1.494	0.239	-6.254	0	-1.962	-1.026
Lead_Origin_Lead Add Form	1.62	0.269	6.013	0	1.092	2.148
Specialization_Rural and Agribusiness	1.5456	0.608	2.543	0.011	0.354	2.737
Current_Occup_Student	-2.913	0.313	-9.306	0	-3.527	-2.3
Current_Occup_Unemployed	-1.653	0.232	-7.126	0	-2.107	-1.198
Tags_ClosebyHorizon	3.4793	0.717	4.85	0	2.073	4.885
Tags_FinanceProb	-2.585	1.234	-2.095	0.036	-5.004	-0.166
Tags_Graduating	-3.4	0.484	-7.03	0	-4.349	-2.452
Tags_Lost	1.5243	0.287	5.305	0	0.961	2.087
Tags_NA	-4.406	0.179	-24.63	0	-4.757	-4.056
Tags_NoFurtherEducation	-5.156	1.036	-4.976	0	-7.187	-3.125
Tags_OtherCourseInterest	-4.38	0.32	-13.67	0	-5.008	-3.751
Tags_Thinking	-2.969	1.234	-2.407	0.016	-5.386	-0.551
Last_Notable_Activity_SMSSent	2.5135	0.17	14.793	0	2.181	2.847

Features	VIF
Current_Occup_Unemployed	2.99
Tags_NA	1.91
Last_Notable_Activity_SMSSent	1.44
Tags_OtherCourseInterest	1.33
Lead_Origin_Lead Add Form	1.22
Tags_ClosebyHorizon	1.21
DNE	1.13
Tags_NoFurtherEducation	1.12
Tags_Lost	1.1
Tags_Graduating	1.06
Current_Occup_Student	1.02
Specialization_Rural and Agribusiness	1.01
Tags_FinanceProb	1.01
Tags_Thinking	1.01

Optimal Cut Off – Conversion Probability



Confusion Metrics		Predicted	
		Not Converted	Converted
Actual	Not Converted	1791	315
	Converted	146	1657

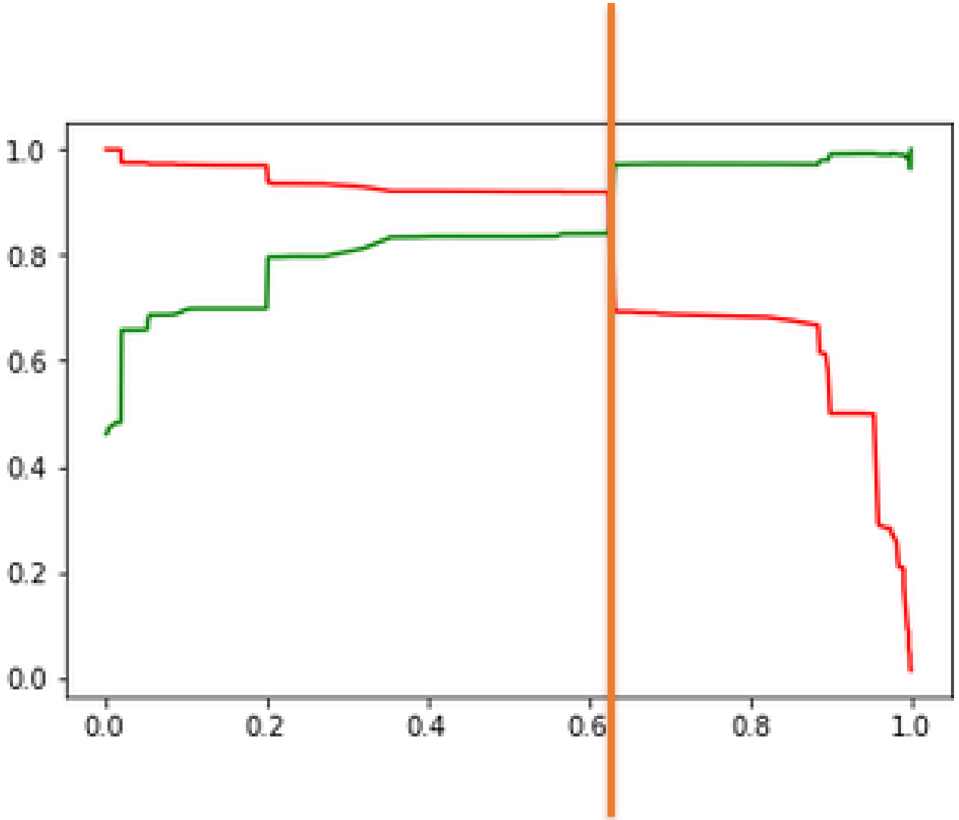
Train Data – Accuracy	**88%**
Train Data – Sensitivity	**92%**
Train Data – Specificity	**85%**

Train Data – Precision	**83%**
Train Data – Recall	**92%**

From the curve above, 0.6 is the optimum point to take it as a cutoff probability.

Train Data Probability

Optimal Cut Off – Conversion Probability



Confusion Metrics		Predicted	
		Not Converted	Converted
Actual	Not Converted	801	136
	Converted	59	680

Test Data – Accuracy	**88%**
Test Data – Sensitivity	**92%**
Test Data – Specificity	**85%**

Test Data – Precision	**83%**
Test Data – Recall	**92%**

Test Data Probability

Conclusions

The top three variables in our model which contribute most towards the probability of a lead getting converted?

- Tags
- Last Notable Activity
- Lead Origin

The top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion?

- Tags - Closed by Horizzon
- Last Notable Activity – SMS Sent
- Lead Origin – Lead Add Form

A good strategy to tap Leads through maximum phone calls would be by looking leads above **40** lead score.

If the company's aim is to not make phone calls unless it's extremely necessary, the Sales team should tap those leads with leads score above **60**.