Lead-Scoring Case Study

- Harshavardhini.R

Business objective and problem statement

- Business Objective:
  - To help X Education select the most promising leads with likely conversion rate of ~80%
  - To build a model to assign a lead score to each of the leads, such that customers with high lead score are hot leads most likely to convert, and customers with low score means cold leads and not likely to convert
- Goal of Data analysis:
  - To 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
  - To address few more problems presented by the company, which the model should be able to adjust to in case the company's requirement changes in the future

Data understanding & EDA

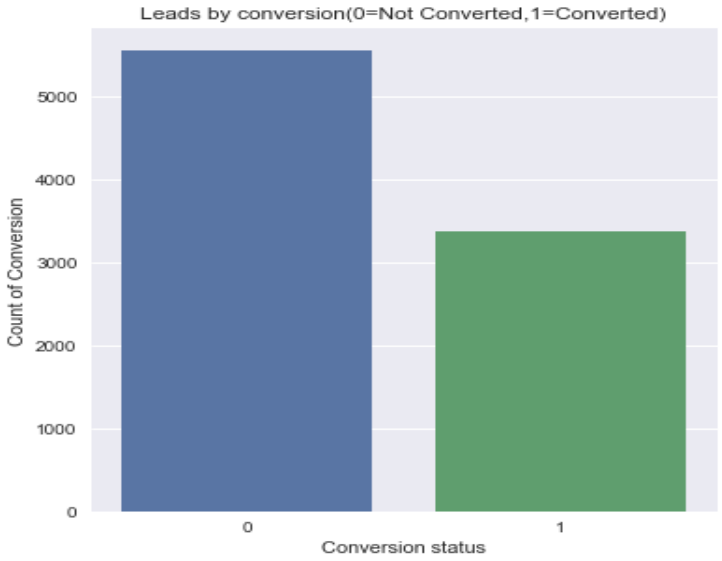
- Lead dataset available for **9,240 leads consists of 37 attributes** such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity etc.

- Data cleaning:
  - Replaced 'Select' with Null values *('Select' values exist as the customer would have not selected anything from drop down when filling web form)*
  - Dropped columns with more than 3,000 Null values, as they may not be significant for analysis
  - Deleted columns with single values or 99%+ single values as they don't show any variance
  - Removed ID columns like prospect ID and Lead number
  - Imputed Null values with 'Other' in specific columns
  - In Lead Source column, 'google' replaced with 'Google', as both represent a single value
  - Removed remaining Null value rows from columns
  - Renames columns for better readability

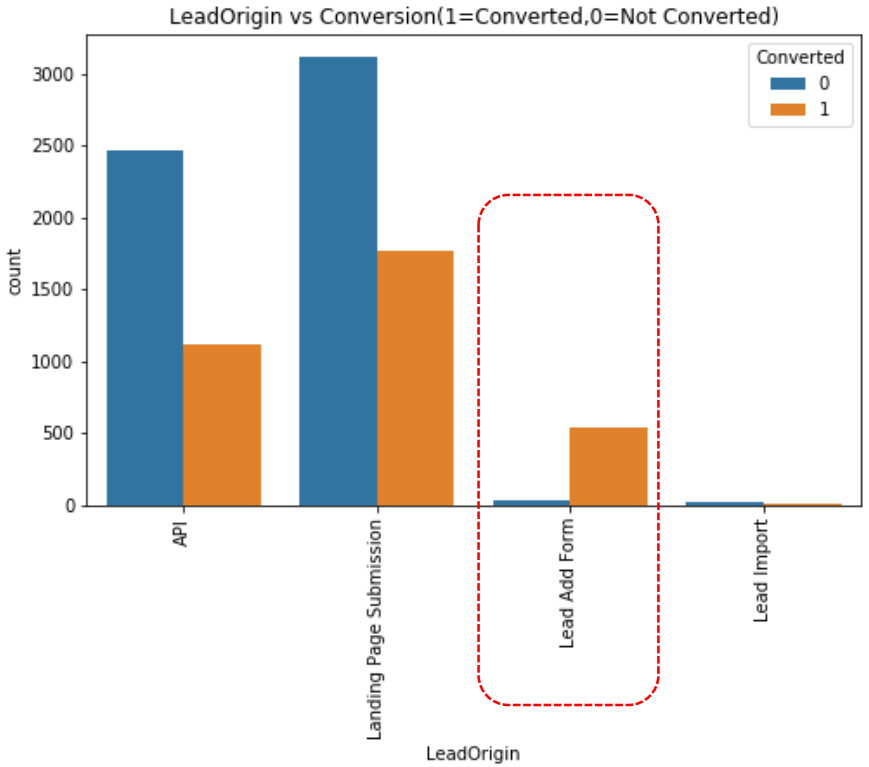
- Post data cleaning, we are left with dataset for **9,074 points with 11 attributes**

EDA: Key insights (1/2)

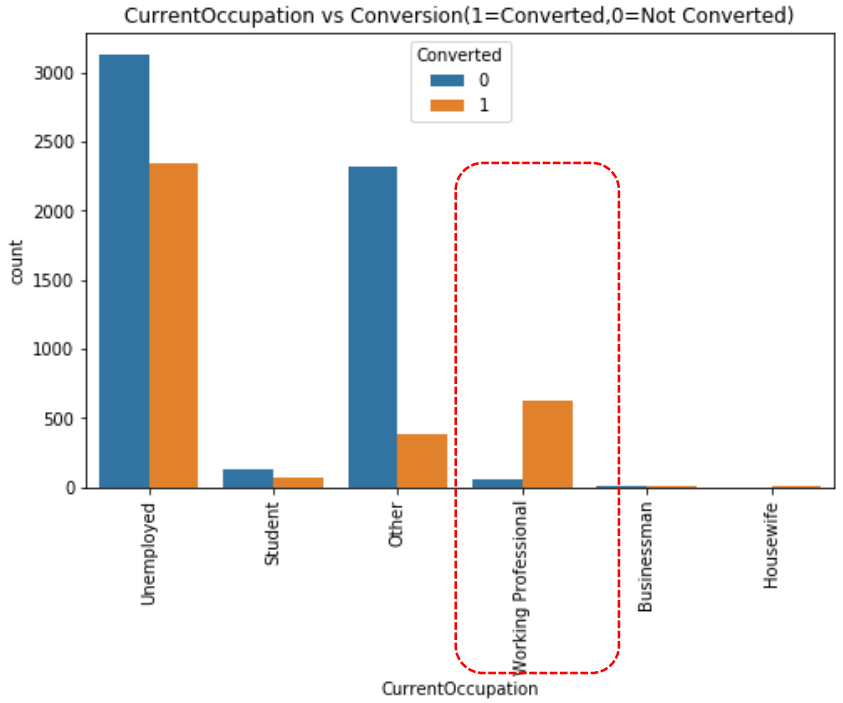
38.5 % lead conversion on existing dataset; objective is to increase this to 80%



Conversion basis Lead origin: High conversion from Lead Add form

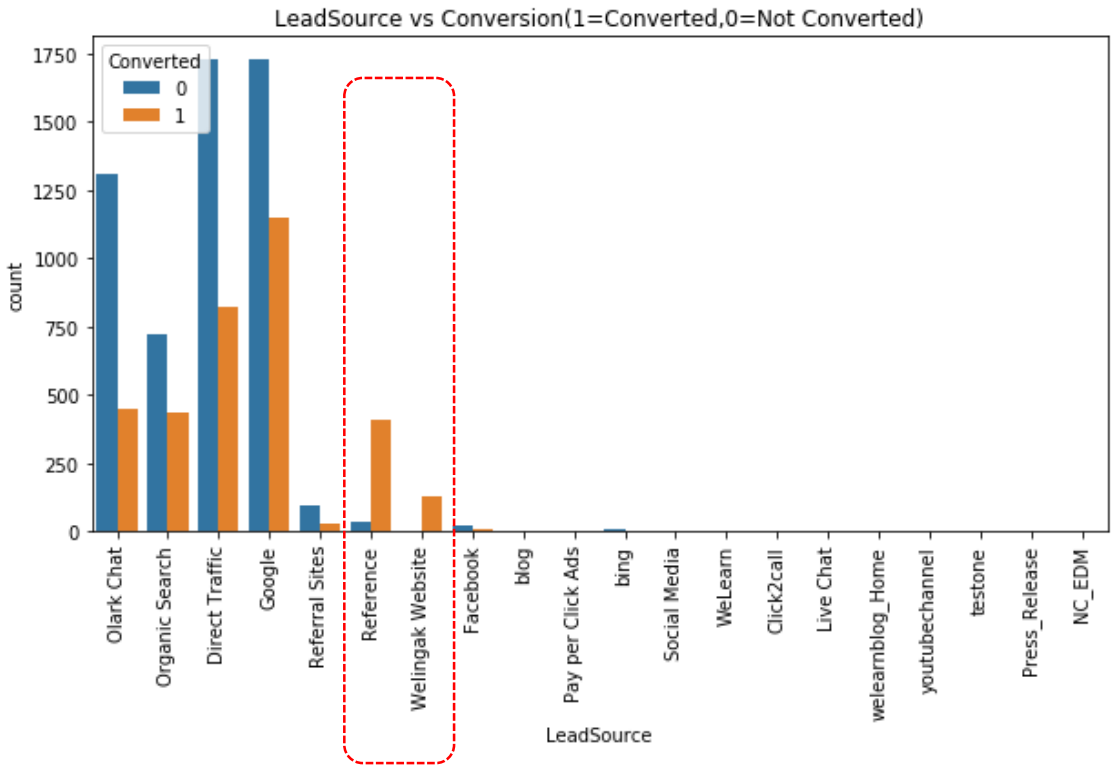


Conversion basis occupation: Very High conversion from working professionals

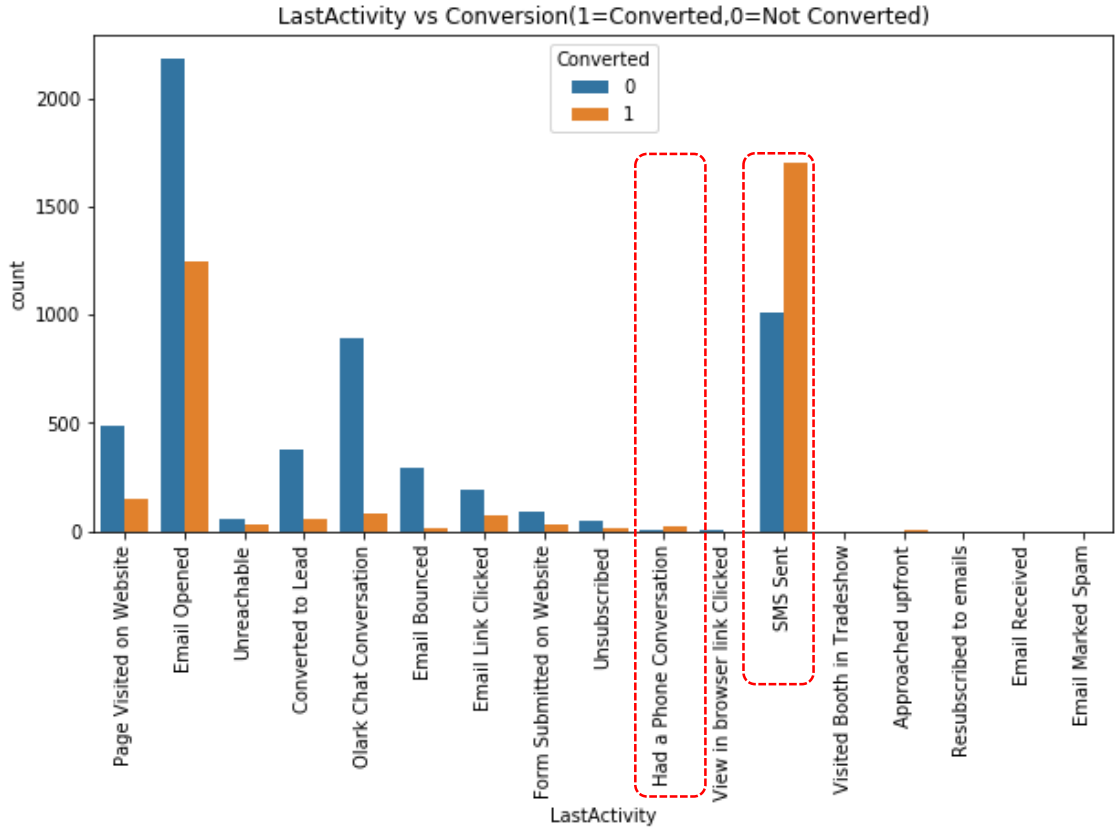


EDA: Key insights (2/2)

Conversion basis lead source: Very High conversion from reference and Welingak Website



Conversion basis last activity: High conversion from SMS sent and Had a phone call conversations



Data preparation for Logistic regression model

- Created Dummy variables for columns with categorical values; total attributes increased to 64
- Outlier treatment for continuous variables: ‘Total visits’, ‘PageviewsPerVisit’
  - Removed values falling in > 99% percentile (~150 rows deleted)
- **The final dataset has 64 attributes for 8,924 leads**
- Train-test split to fit our model on the train data, in order to make predictions on the test data
- Feature scaling to standardize the 3 continuous variables
- Removed highly correlated variables (correlation > 0.7) from both train and test data set in order to capture maximum variance
- **The final dataset has 1 predicted variable, 53 fairly independent features for 8,924 leads**

Model building using Recursive Feature Elimination (RFE)

- RFE with 30 independent variables
- **First model has 30 features: 16 variables with VIF> 2 and 7 features with p-values > 5%**

1st model with 30 features

Features	Variables	coefficient
0	const	(0.1043)
1	LastActivity_Resubscribed to emails	27.3467
2	CurrentOccupation_Housewife	25.8245
3	LastActivity_Email Marked Spam	25.0597
4	LastNotableActivity_Had a Phone Conversation	23.8813
5	LeadOrigin_Lead Add Form	2.8209
6	LastActivity_Had a Phone Conversation	2.4054
7	CurrentOccupation_Working Professional	2.3826
8	LastNotableActivity_Unreachable	1.7881
9	LeadSource_Welingak Website	1.7084
10	LastActivity_SMS Sent	1.3036
11	TotTimeSpent	1.0996
12	LastActivity_Unsubscribed	0.8482
13	LeadSource_Olark Chat	0.5103
14	TotalVisits	0.3357
15	CurrentOccupation_Student	(0.0909)

Features	Variables	coefficient
16	CurrentOccupation_Unemployed	(0.0972)
17	PageViewsPerVisit	(0.3245)
18	LeadSource_Google	(0.5526)
19	LastNotableActivity_Olark Chat Conversation	(0.5560)
20	LastActivity_Email Bounced	(0.6033)
21	LastNotableActivity_Modified	(0.6443)
22	LastNotableActivity_Page Visited on Website	(0.6530)
23	LastActivity_Converted to Lead	(0.6979)
24	LeadSource_Organic Search	(0.7230)
25	LeadSource_Referral Sites	(0.8250)
26	LeadSource_Direct Traffic	(0.9802)
27	LastActivity_Olark Chat Conversation	(0.9867)
28	CurrentOccupation_Other	(1.2432)
29	DoNotEmail	(1.3599)
30	LastActivity_View in browser link Clicked	(24.4066)

- Created 15 model iterations, eliminating variables based on high VIF and p-values > 5% and at the same time ensuring accuracy of ~80%

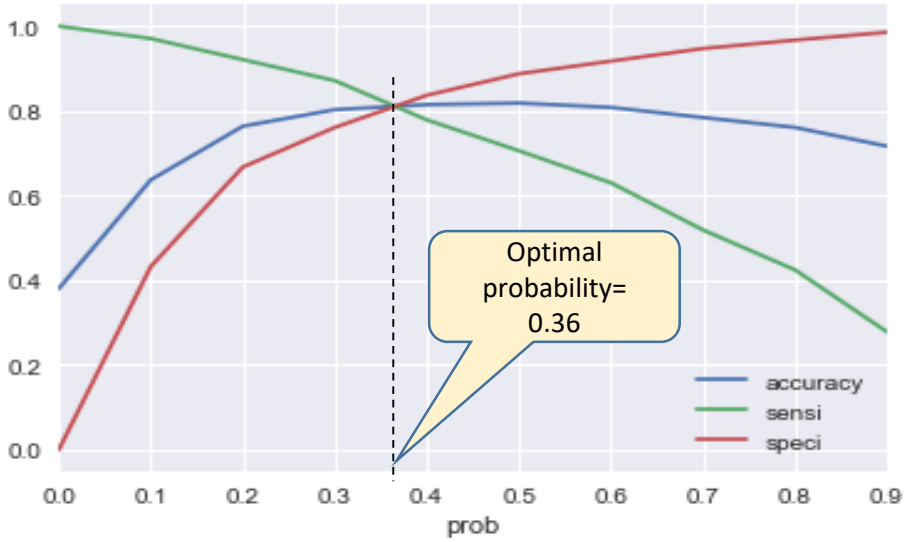
Final Model and Optimal cut-off point

- **Final model has 16 features, all with VIF < 2 and p-values < 5%**

Final model with 16 features

Features	Variables	coefficient
0	const	0.3408
1	LastActivity_Had a Phone Conversation	2.8224
2	CurrentOccupation_Working Professional	2.4766
3	LeadOrigin_Lead Add Form	2.3376
4	LastNotableActivity_Unreachable	1.9670
5	LeadSource_Welingak Website	1.6684
6	LastActivity_SMS Sent	1.3422
7	TotTimeSpent	1.0932
8	TotalVisits	0.1899
9	LastActivity_Converted to Lead	(1.0698)
10	CurrentOccupation_Other	(1.1875)
11	LeadSource_Google	(1.3616)
12	LastActivity_Olark Chat Conversation	(1.3658)
13	DoNotEmail	(1.5145)
14	LeadSource_Organic Search	(1.5818)
15	LeadSource_Referral Sites	(1.6922)
16	LeadSource_Direct Traffic	(1.7428)

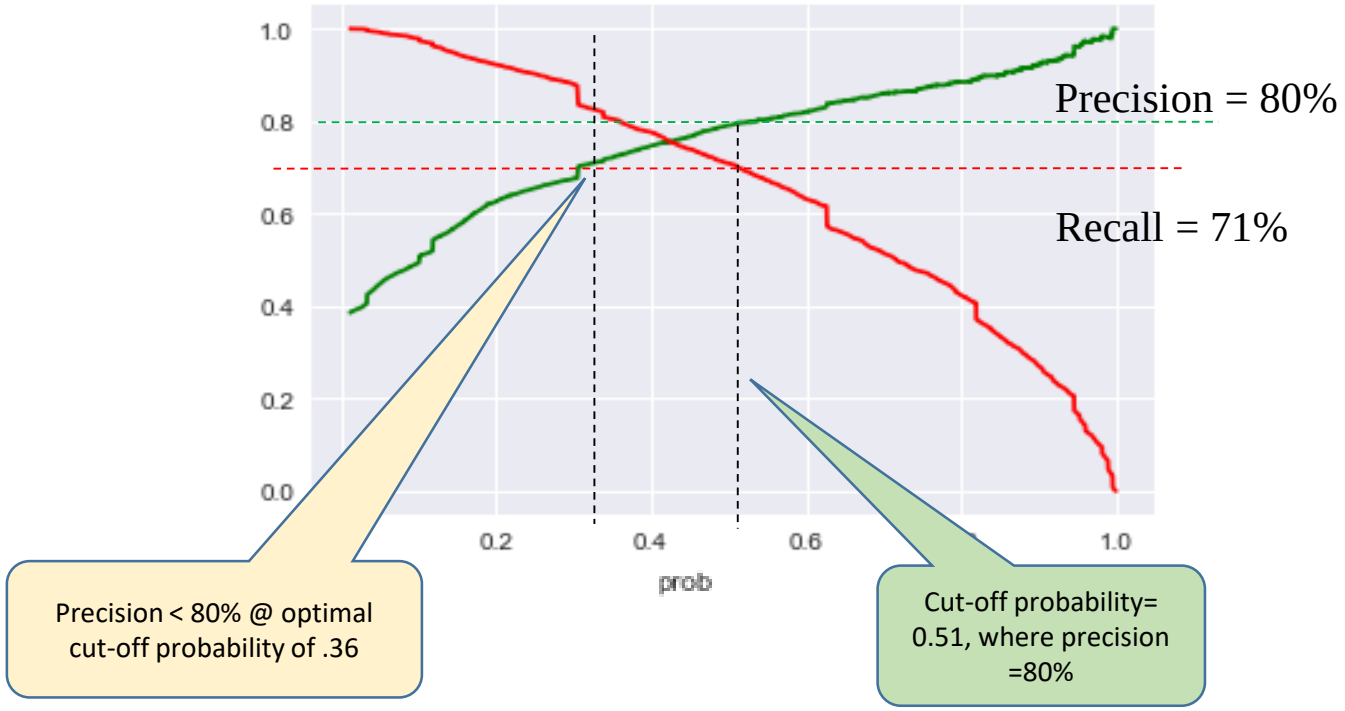
Optimal cut-off point = 0.36



- A plot of accuracy, sensitivity & specificity suggests **0.36 as the optimum probability for conversion**;
- **Precision @ 73% < 80%** conversion target as required by X-education, hence, we need to **identify another threshold point**

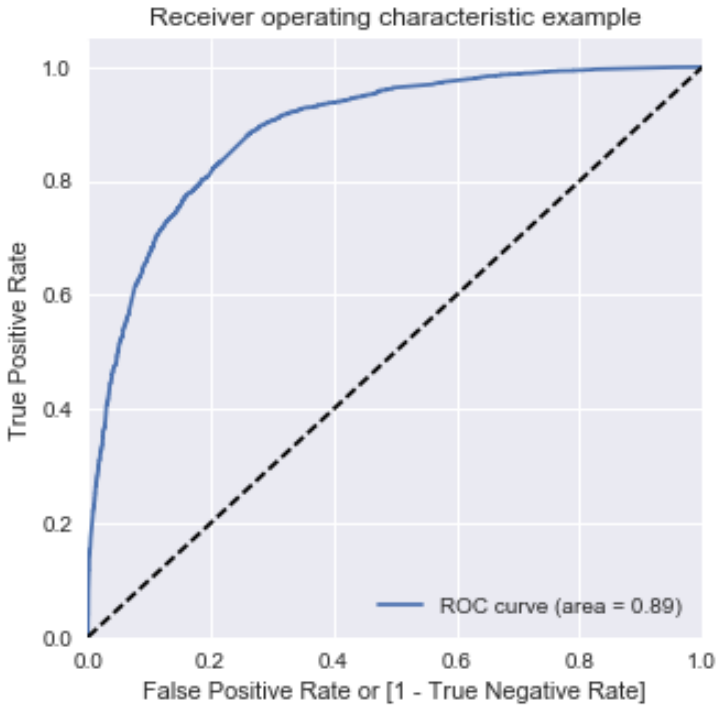
New threshold probability for 80% conversion and Model testing

Precision vs Recall trade-off



Based on precision-recall curve and **precision = 80%**, **we have chosen the threshold probability for the model @ 0.51**; **recall** is also good **@ 71%** at this cut-off

Model testing using ROC curve



ROC curve with area under the curve @ 89% (@ threshold probability of 0.51) suggests that **there is a good 89% chance that the model will be able to distinguish between true positives and negatives**

Testing the model fit on train data and test data: Confusion matrix

Train data

- 6,246 data points
- Cut-off probability: 0.51

Key metrics	%
Accuracy	81.73%
Specificity: TN/ (TN + FP)	88.84%
Sensitivity/ Recall : TP/ (TP + FN)	70.15%
Positive predictive value/ Precision	79.41%
Negative predictive value	82.91%

Test data

- 2,678 data points
- Cut-off probability: 0.51

Key metrics	%
Accuracy	81.78%
Specificity: TN/ (TN + FP)	89.13%
Sensitivity/ Recall : TP/ (TP + FN)	69.32%
Positive predictive value/ Precision	79.01%
Negative predictive value	83.11%

- **Overall accuracy** for both train and test model is **c. 82%**
- **Precision and recall value** for both train and test data are in the range of **~80% and ~70% respectively**
- Based on the above, we can infer that the model is **a good fit model**, **trained well** on the train data and **able to generalize the results on the test data**

Lead scoring regression model and top 3 variables

- The logistic regression model calculates lead score based on 16 features, as explained earlier:

Features	Symbols
const	c
LastActivity_Had a Phone Conversation	x1
CurrentOccupation_Working Professional	x2
LeadOrigin_Lead Add Form	x3
LastNotableActivity_Unreachable	x4
LeadSource_Welingak Website	x5
LastActivity_SMS Sent	x6
TotTimeSpent	x7
TotalVisits	x8
LastActivity_Converted to Lead	x9
CurrentOccupation_Other	x10
LeadSource_Google	x11
LastActivity_Olark Chat Conversation	x12
DoNotEmail	x13
LeadSource_Organic Search	x14
LeadSource_Referral Sites	x15
LeadSource_Direct Traffic	x16

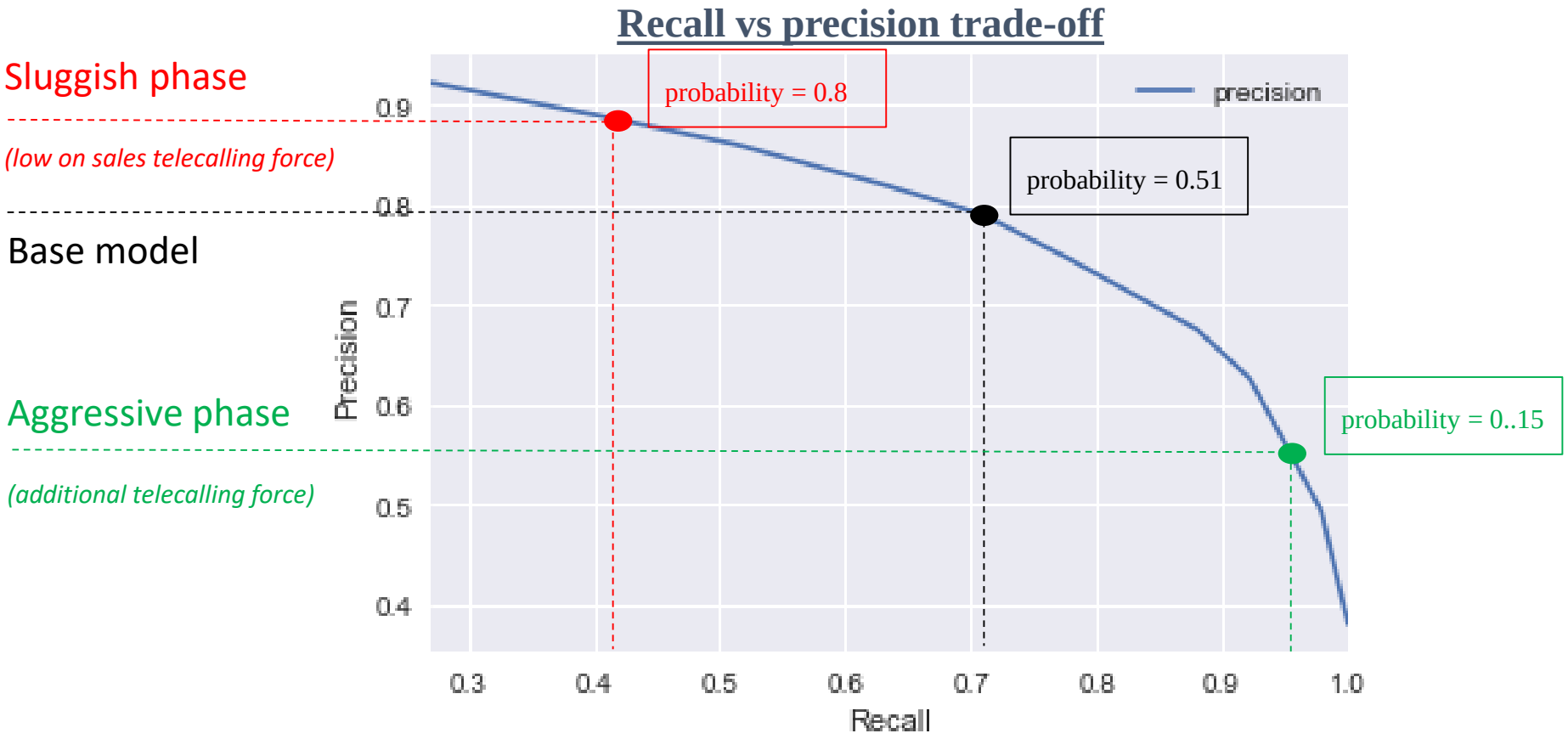
- The lead score to each of the leads based on the below equation:

Lead score = (0.34c + 2.82x1 + 2.48x2 + 2.34x3 + 1.97x4 + 1.674x5 + 1.34x6 + 1.09x7 + 0.19x8 - 1.07x9 - 1.19x10 - 1.36x11 - 1.37x12 - 1.51x13 - 1.58x14 - 1.69x15 - 1.74x16)*100

- Top 3 variables that contribute to most towards the probability of a lead getting converted:

Variables	Category/ Dummy variables
Last Activity	LastActivity_Had a Phone Conversation
What is your current occupation	CurrentOccupation_Working Professional
Lead Origin	LeadOrigin_Lead Add Form

Adjusting the model during aggressive and sluggish phase using interplay between precision & recall



- During **aggressive phase,** the model may be tuned in a way that no potential lead is missed out. i.e **reducing the precision** (increasing the input funnel) and **increasing recall value** (more HOT leads from total relevant leads)- may be achieved by **lowering the cut-off probability** to 0.15
- During **sluggish phase,** the model may be tuned in a way that to minimize useless calls. i.e **increasing the precision** (more relevant leads) and **reducing recall value** (less conversion (HOT leads) from total relevant leads)- may be achieved by **increasing the cut-off probability** to 0.8

Thank You