Lead Scoring for X Education

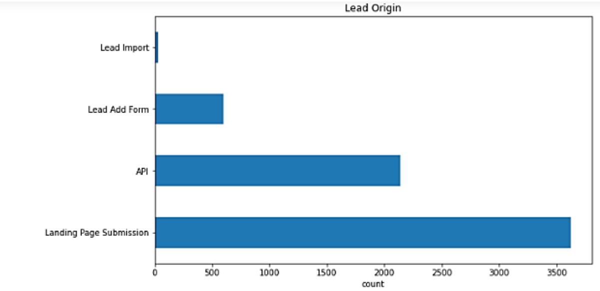
A Case Study by Venkataramanan M & Akshay Anand,

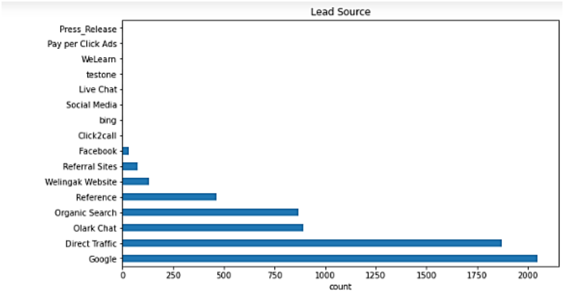
IIIT-B, upGrad

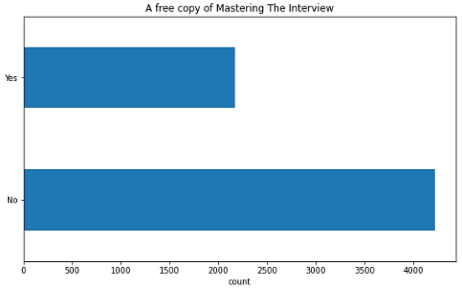
Why Lead Scoring?

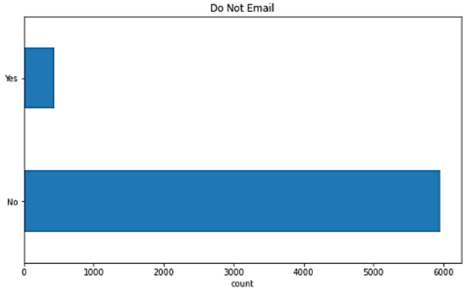
- A person is referred to as a lead if he/she provides their contact details on visiting the website.
- Only 30% of leads get converted into paying customers.
- Assign a score from 0 (hot) to 100 (cold) to identify the most potential clients (hot leads).
- Build a model for assigning the scores dynamically, so as to get accustomed to future changes.
- Nurture more leads to become paying customers.
- Increase the ROI of time & efforts for social media marketing.

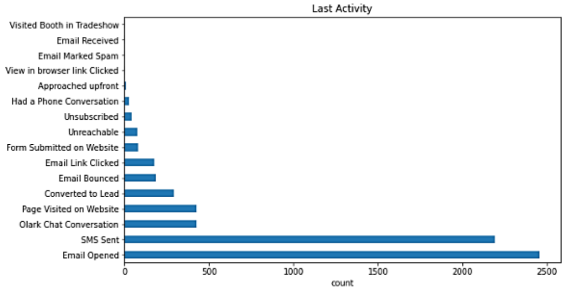
EDA – Univariate – Plots

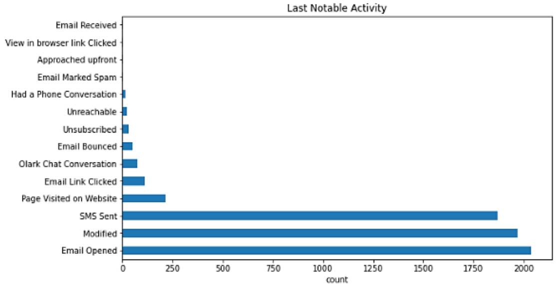


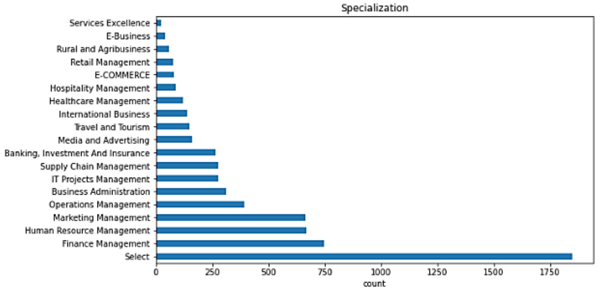


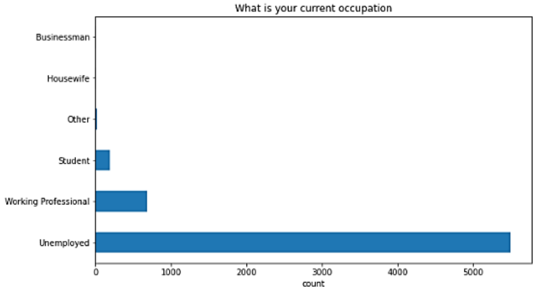


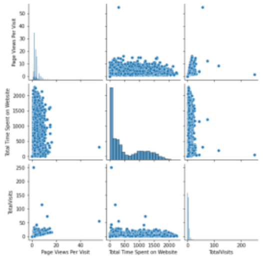








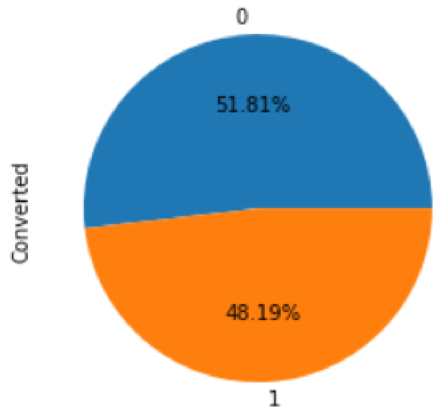




EDA – Univariate – Observations

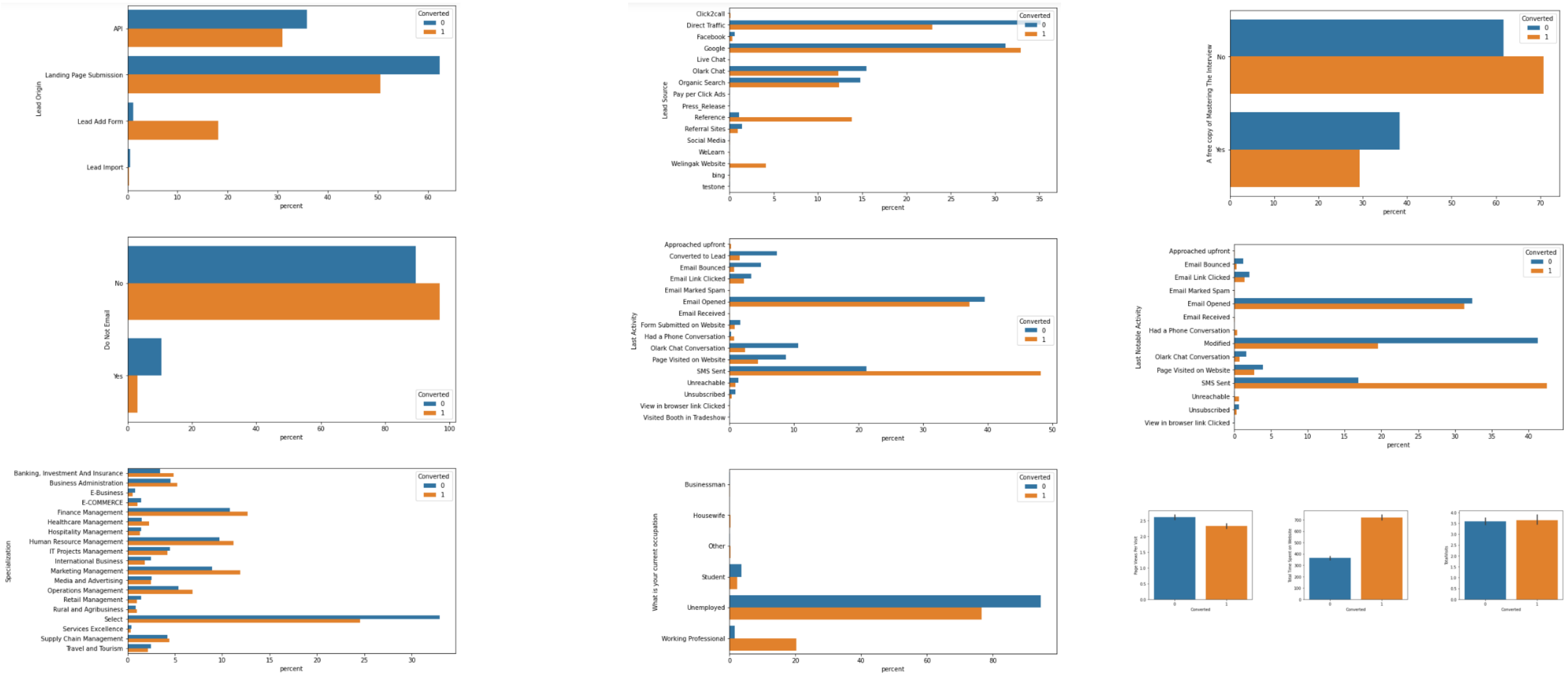
- Highest number of leads were identified from the landing page, who provided the required information via a Google page/form.
- Most leads opted for Email services. Also, their last activity was opening Email, student or not.
- Most people who were working in the roles of Finance, HR and Marketing management, now unemployed, became a lead.
- Most leads did not opt for a free copy of mastering the interview.
- People spending more time on the website have their browsing frequency and, number of pages viewed per visit almost same as that of the people spending far lesser time.
- Browsing frequency increases with an increase in number of pages viewed per visit.

Target variable



- We could see that only 48.19% of the leads have been converted into paying customers.
- 51.81% of them continue to be just leads.
- Our aim is to build a Logistic Regression model, so as to increase lead conversion rate to 80%.

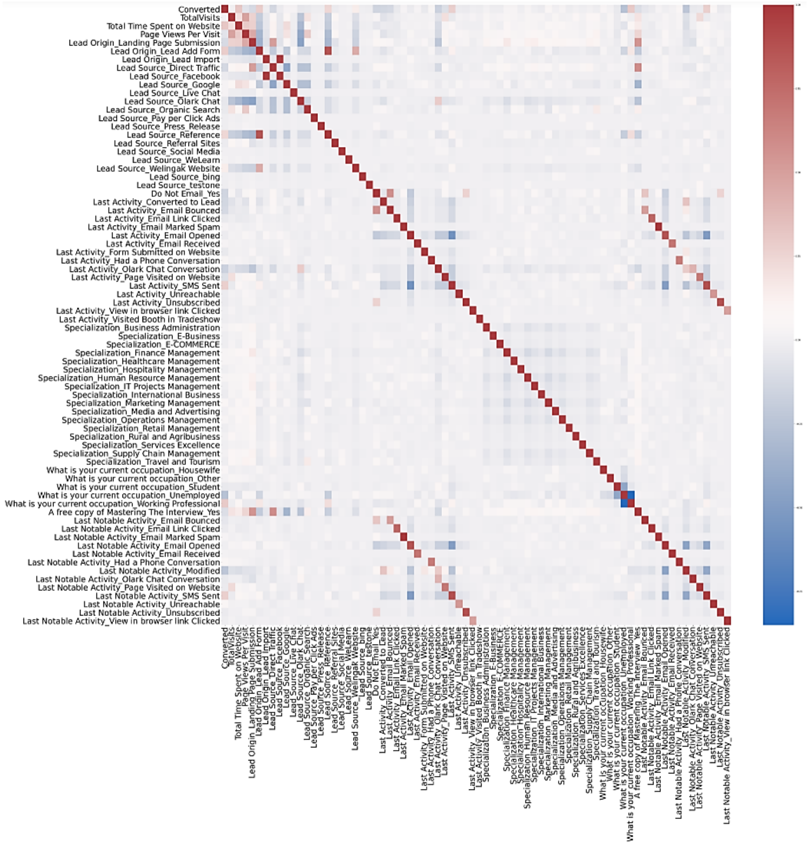
EDA – Bivariate – Plots



EDA – Bivariate – Observations

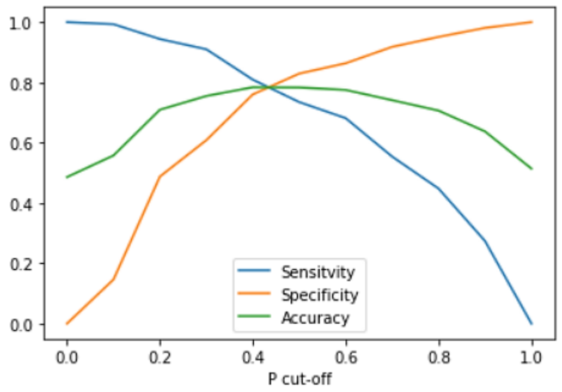
- Highest lead conversion rate is acquired by the leads identified from the landing page, who provided the required information via a Google page/form.
- Leads who opted for Email services have a much higher conversion rate. However, their last activity was sending an SMS, be a student or not.
- Leads who were working in the roles of Finance, HR and Marketing management, now unemployed, have a much higher conversion rate.
- Leads who did not opt for a free copy of mastering the interview, have a much higher conversion rate.
- Leads who spend more spend more time on visiting the website have a high conversion rate. Whereas, those who visit more number of pages have a low conversion rate.
- Leads who visit the site frequently have an equal chance of getting converted and not.

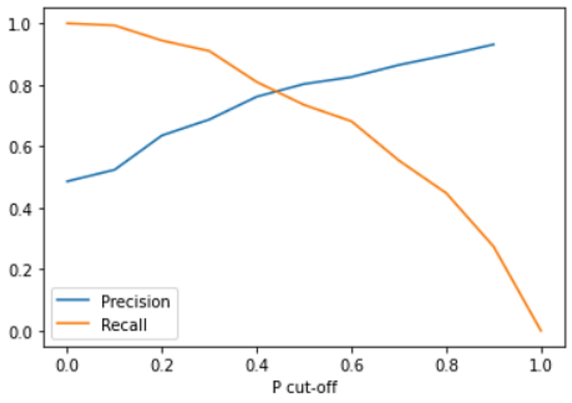
Correlation Matrix - Heatmap



- A vast majority of the variables form weak positive to weak negative correlations with each other.
- The variable 'Converted' alone forms strong correlations with all the variables since, it is the target variable.
- A minority of the variables form strong positive and strong negative correlations with each other.
- Correlation values varies from -0.75 (strong negative correlation) to +1.00 (strong positive correlation).

Finding optimal probability cut-off





- Optimal point is where all the performance metrics plotted against the range of probability distribution intersects each other.
- With high sensitivity, our model will consider almost all leads for conversion, including those having marginal likelihood. Can be employed when having sufficient man-hours for lead nurturing.
- With high specificity, our model will identify only those leads having high likelihood for conversion, leaving behind the ones with moderate likelihood. Can be employed when having less man-hours.
- In our model, recall metric is of more importance since, our target is to achieve a lead conversion rate of 80%. Also, correctness of prediction is not of a risk.
- Optimal probability cut-off = 0.44 (from both graphs)

Model Building

Generalized Linear Model Regression Results

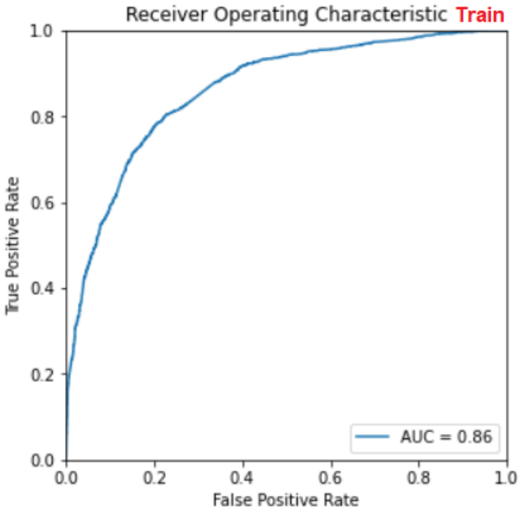
Dep. Variable:	Converted	**No. Observations:**	4476
Model:	GLM	**Df Residuals:**	4463
Model Family:	Binomial	**Df Model:**	12
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2079.3
Date:	Wed, 08 Dec 2021	**Deviance:**	4158.7
Time:	19:01:04	**Pearson chi2:**	5.05e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.4893	0.181	8.244	0.000	1.135	1.843
Total Time Spent on Website	1.1257	0.046	24.245	0.000	1.035	1.217
Lead Origin_Lead Add Form	3.5627	0.229	15.527	0.000	3.113	4.012
Lead Source_Olark Chat	1.3931	0.118	11.787	0.000	1.161	1.625
Lead Source_Welingak Website	2.3844	1.033	2.309	0.021	0.360	4.409
Do Not Email_Yes	-1.5002	0.196	-7.638	0.000	-1.885	-1.115
Last Activity_Converted to Lead	-1.2448	0.238	-5.219	0.000	-1.712	-0.777
Last Activity_Olark Chat Conversation	-1.2469	0.182	-6.851	0.000	-1.604	-0.890
Last Activity_SMS Sent	1.1110	0.084	13.243	0.000	0.947	1.275
Last Activity_Unsubscribed	1.0648	0.508	2.096	0.036	0.069	2.061
What is your current occupation_Student	-2.2471	0.269	-8.359	0.000	-2.774	-1.720
What is your current occupation_Unemployed	-2.3953	0.182	-13.153	0.000	-2.752	-2.038
Last Notable Activity_Unreachable	3.3800	1.069	3.161	0.002	1.284	5.476

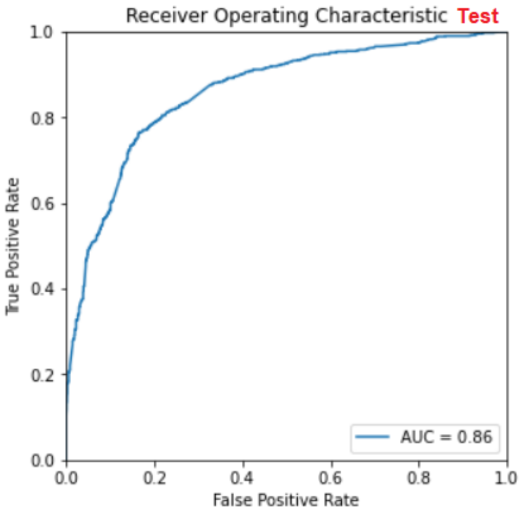
	Features	VIF
10	What is your current occupation_Unemployed	1.91
7	Last Activity_SMS Sent	1.52
2	Lead Source_Olark Chat	1.47
1	Lead Origin_Lead Add Form	1.44
3	Lead Source_Welingak Website	1.28
0	Total Time Spent on Website	1.23
6	Last Activity_Olark Chat Conversation	1.22
4	Do Not Email_Yes	1.18
5	Last Activity_Converted to Lead	1.09
8	Last Activity_Unsubscribed	1.09
9	What is your current occupation_Student	1.04
11	Last Notable Activity_Unreachable	1.01

- Performed logistic regression using Generalized Linear Model (GLM) to fit the model following Binomial distribution of families.
- Obtained the best features for our model in 4th model re-build iteration, after performing RFE.
- All the selected features p-value < 0.05 & VIF < 5

Model Evaluation



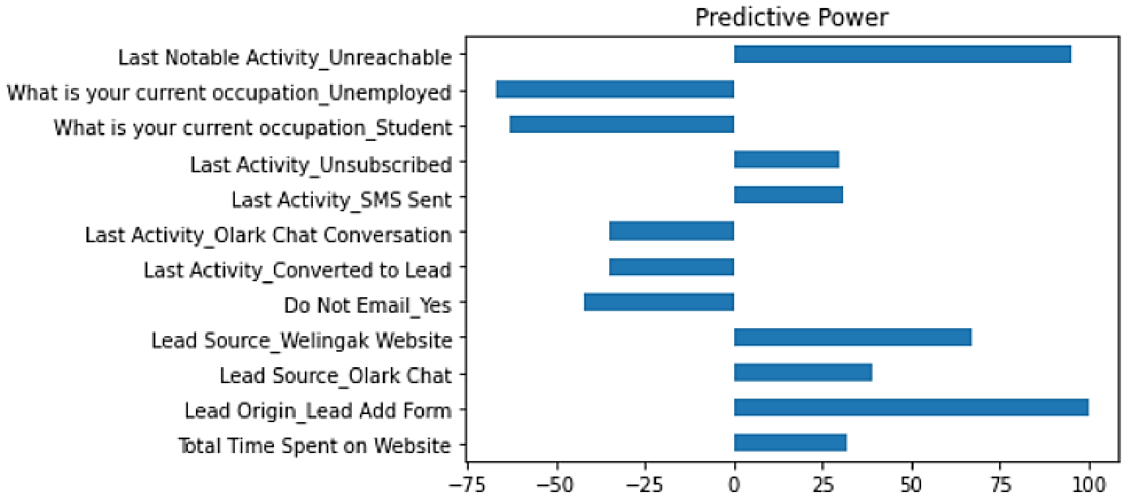

```
F1 Score =  0.78
Sensitivity =  78.0 %
Specificity =  79.0 %
Accuracy =  79.0 %
Precision =  78.0 %
Recall =  78.0 %
True Positive rate =  78.0 %
False Positive rate =  21.0 %
True Negative rate =  84.0 %
False Negative rate =  20.0 %
Positive Predictive value =  78.0 %
Negative Predictive value =  79.0 %
```




```
F1 Score =  0.78
Sensitivity =  79.0 %
Specificity =  80.0 %
Accuracy =  79.0 %
Precision =  78.0 %
Recall =  79.0 %
True Positive rate =  79.0 %
False Positive rate =  20.0 %
True Negative rate =  89.0 %
False Negative rate =  19.0 %
Positive Predictive value =  78.0 %
Negative Predictive value =  81.0 %
```

- AUC under the ROC curve, F1 score and all performance metrics yield good & similar results in both train & test data.
- Our model can now provide weightage to correctness of either sensitivity or specificity, without affecting each other.
- Our model is good to assign the lead score dynamically to each ID.

Final Model – workaround required



P(Converted:X)= 53 %

X = {Total Time Spent on Website = 0.32, Lead Origin_Lead Add Form = 1.00, Lead Source_Olark Chat = 0.39, Lead Source_Welingak Website = 0.67, Do Not Email_Yes = -0.42, Last Activity_Converted to Lead = -0.35, Last Activity_Olark Chat Conversation = -0.35, Last Activity_SMS Sent = 0.31, Last Activity_Unsubscribed = 0.30, What is your current occupation_Student = -0.63, What is your current occupation_Unemployed = -0.67, Last Notable Activity_Unreachable = 0.95}

- Top 3 variables that contribute (positive) most towards the probability of lead conversion are:
  1. Lead Origin_Lead Add Form
  2. Last Notable Activity_Unreachable
  3. Lead Source_Welingak Website
- Selected feature variables influences the predictive power of our model both positively and negatively.
- Probability of conversion rate = 53%.
- Evaluation metrics are good & comparable.
- Variables dropped in bulk even before the feature selection process, due to many missing values. Hence, we need more data.
- We can dynamically add more data to improve the results, as model predictability is already good.