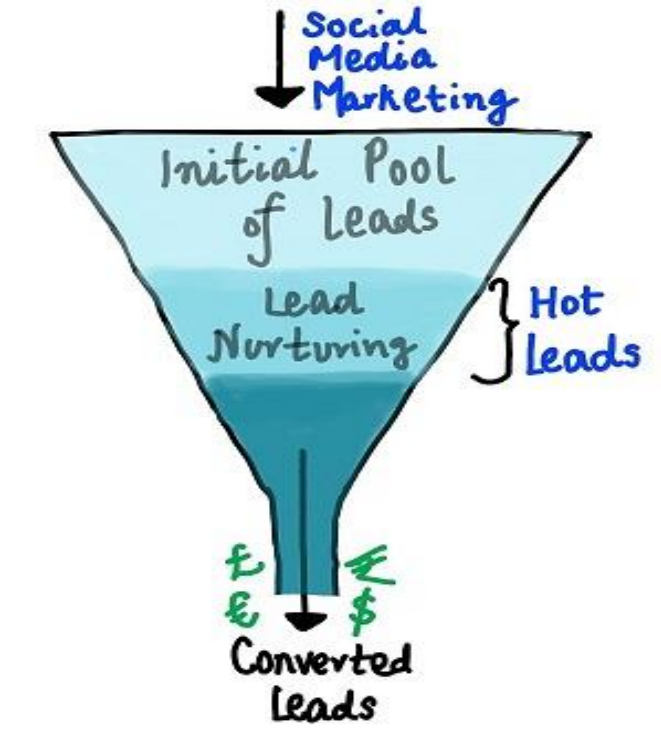
LEAD SCORING LOGISTIC REGRESSION ASSIGNMENT

SUBMITTED BY: ROHIT BHANDARI

EMAIL-ID: ROHSINGHIT@GMAIL.COM

PROBLEM STATEMENT / BUSINESS PROBLEM

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. 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.



Data Preparation, Preprocessing & Missing value treatment

- **Since this dataset contains certain columns with missing values, outliers, simple “Select” field, binaries in the form of yes/no etc. So in order to fit them into our logistic regression model, these were treated in the manner required for our model.**
- **Mode is used for imputation in the categorical variables columns.**
- **The categorical binaries in the form of yes and no were converted as 0/1 to fit in the model.**
- **The select fields in the dataset was replaced with Nan and handled accordingly.**
- **Columns with missing values more than 35% were removed.**
- **The columns where only one value was dominating were also removed as they were of no use in the model.**

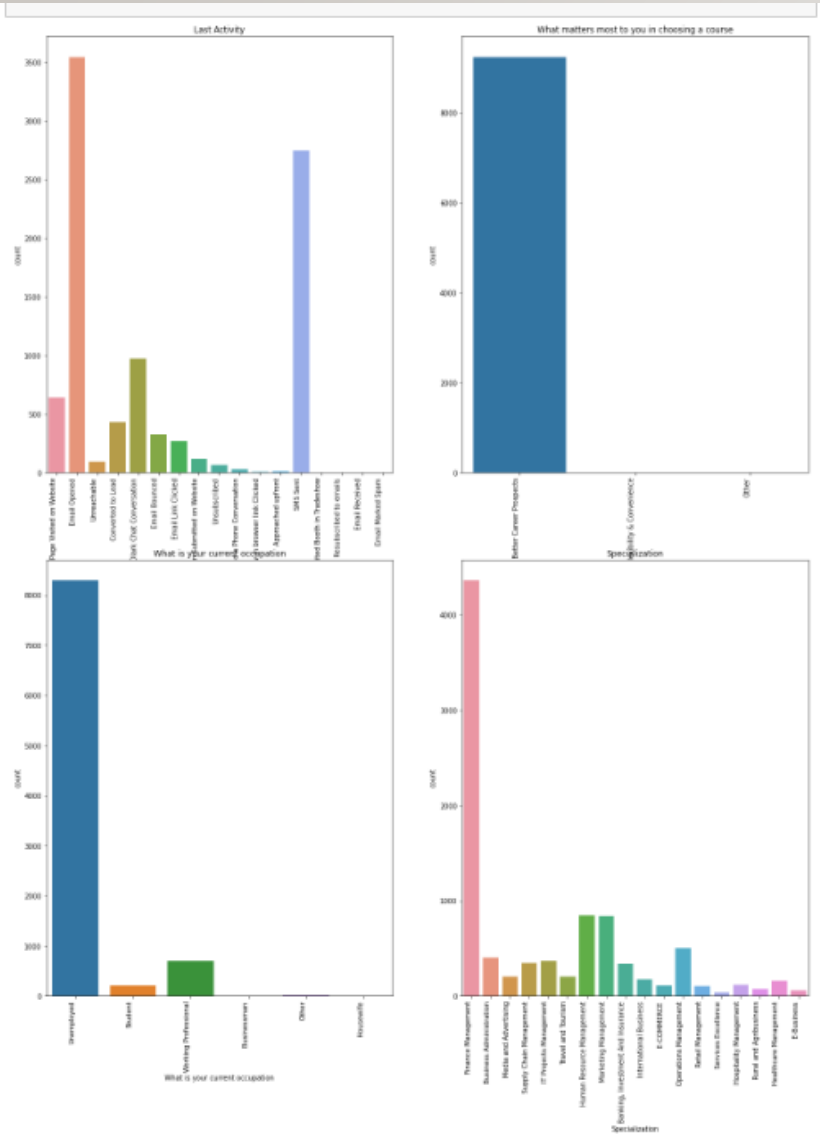
ASSIGNMENT APPROACH

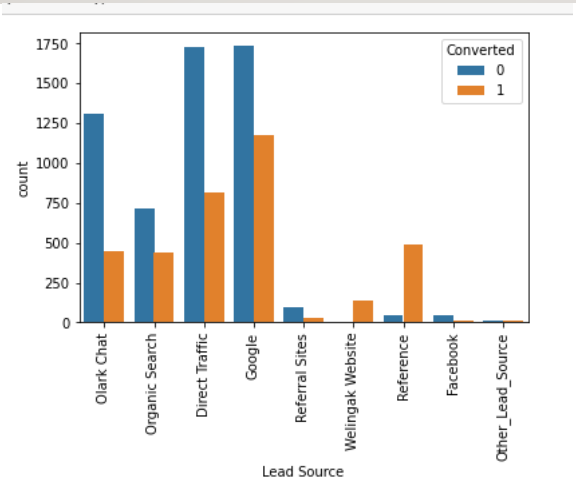
- Data Preparation, Preprocessing & Missing value treatment
- Exploratory Data Analysis and visualization
- Dummy Variable Creation
- Test-Train Split
- Model Building
- Feature Selection Using RFE
- Checking VIFs
- Plotting the Confusion Matrix, Accuracy, Sensitivity, Specificity
- Plotting ROC Curve
- Finding optimal value of the cut off
- Precision-Recall Trade off
- Model Building

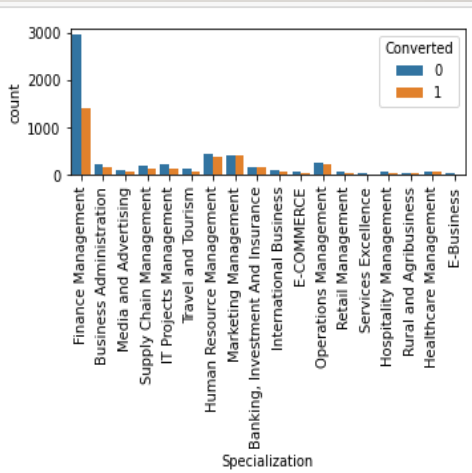
EDA OUTCOMES

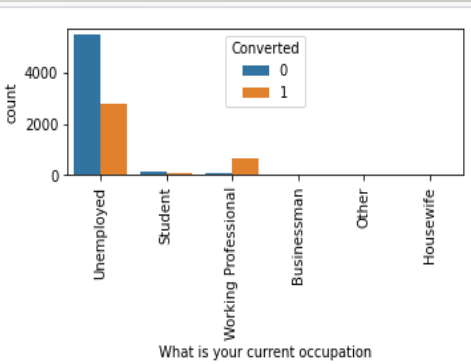
- To increase overall conversion rate, we need to increase the number of Working Professional leads by reaching out to them through different mediums and also on increasing the conversion rate of Unemployed leads.
- The count of leads from the Google and Direct Traffic is maximum.
- The conversion rate of the leads from Reference and Welingak Website is maximum
- To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'API' and 'Landing Page Submission' and also increasing the number of leads from 'Lead Add Form’.
- The count of last activity as "Email Opened" is max
- The conversion rate of SMS sent as last activity is maximum
- To increase overall conversion rate, we need to increase the number of Working Professional leads by reaching out to them through different mediums and also on increasing the conversion rate of Unemployed leads

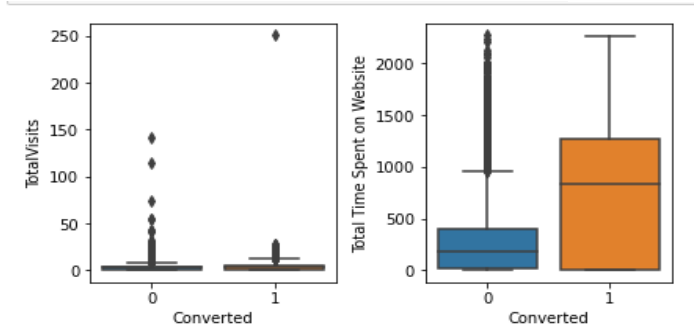
EDA

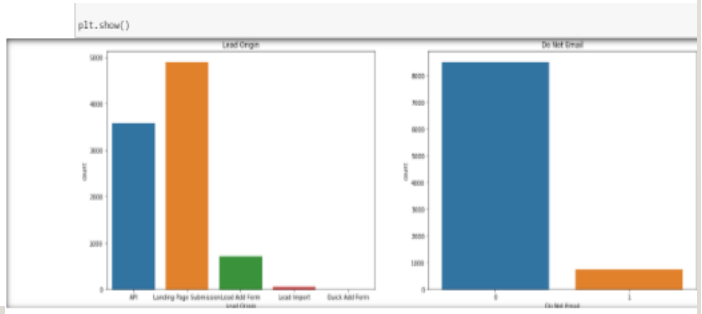

Last Activity
What matters most to you in choosing a course
What is your current occupation
Specialization

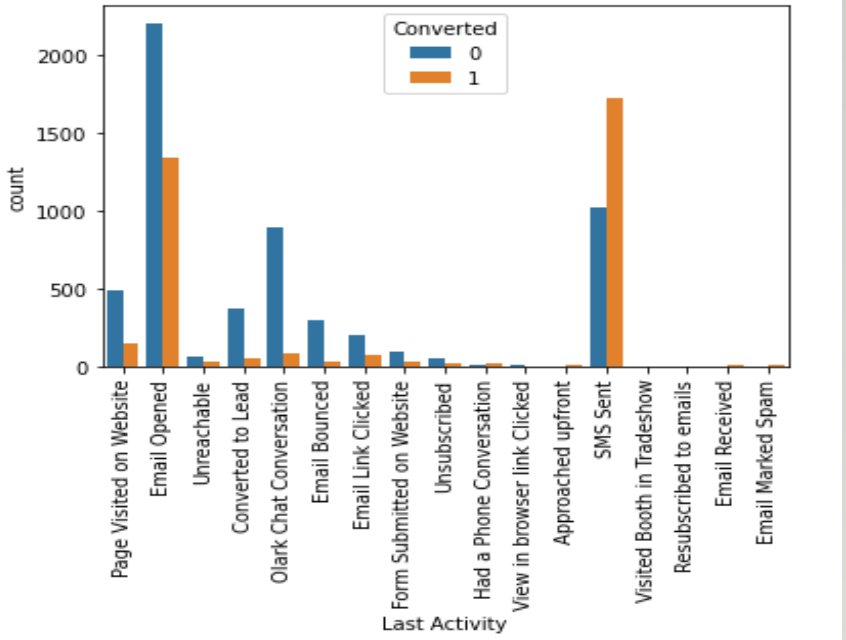

Converted
0
1
count
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
Facebook
Other_Lead_Source
Lead Source


Converted
0
1
count
Finance Management
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization


Converted
0
1
count
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation


TotalVisits
Converted
Total Time Spent on Website
Converted


plt.show()


Converted
0
1
count
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity

FURTHER STEPS

- Dummy variable creation
- Train-Test Split
- Feature Scaling : using StandardScaler()
- Checking Correlation between different numerical variables
- Dropping the high correlated variables
- Then, going ahead with logistic regression model building to get a Generalized Linear Model Regression Results.
- Feature selection using RFE.
- Then further dropping variables based on p value and VIF.
- Creating a dataframe with the true conversion status and the predicted probabilities.
- Then, again building the model till we get the final model.

FINAL MODEL

Out[83]:

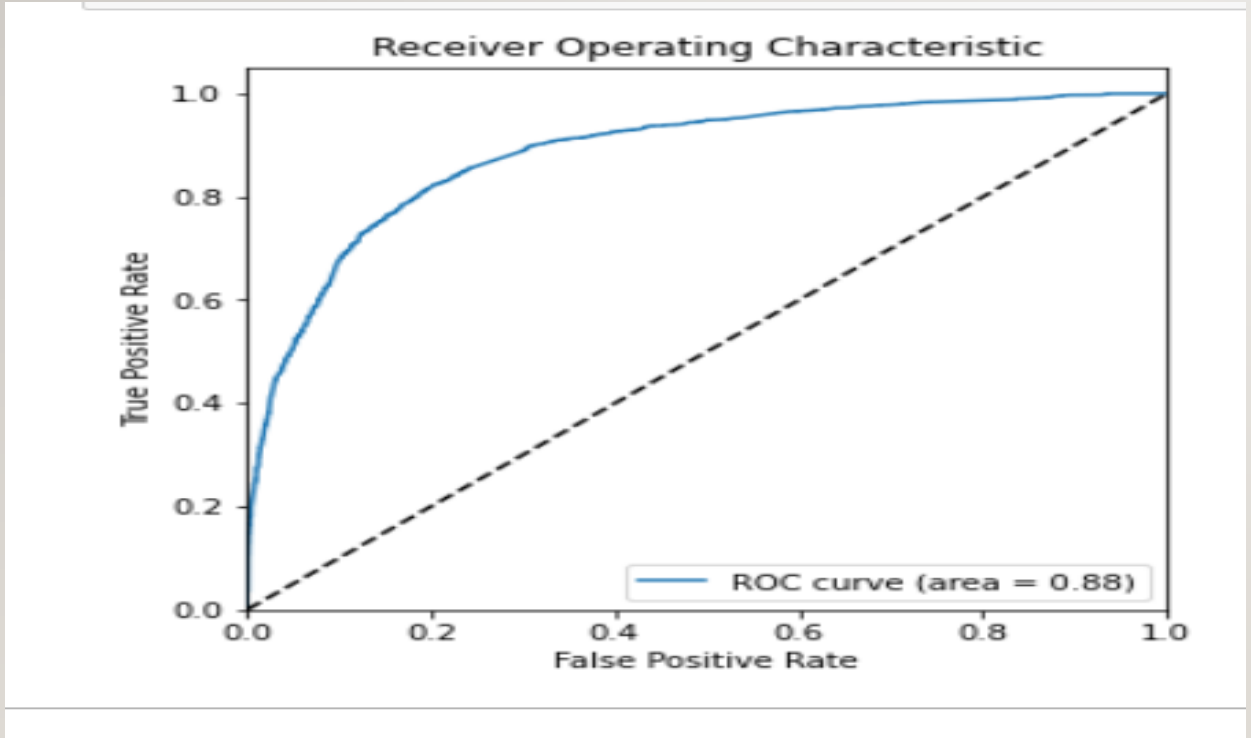
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6453
Model Family:	Binomial	**Df Model:**	14
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2677.2
Date:	Mon, 21 Nov 2022	**Deviance:**	5354.5
Time:	10:47:15	**Pearson chi2:**	7.82e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3944
Covariance Type:	nonrobust		

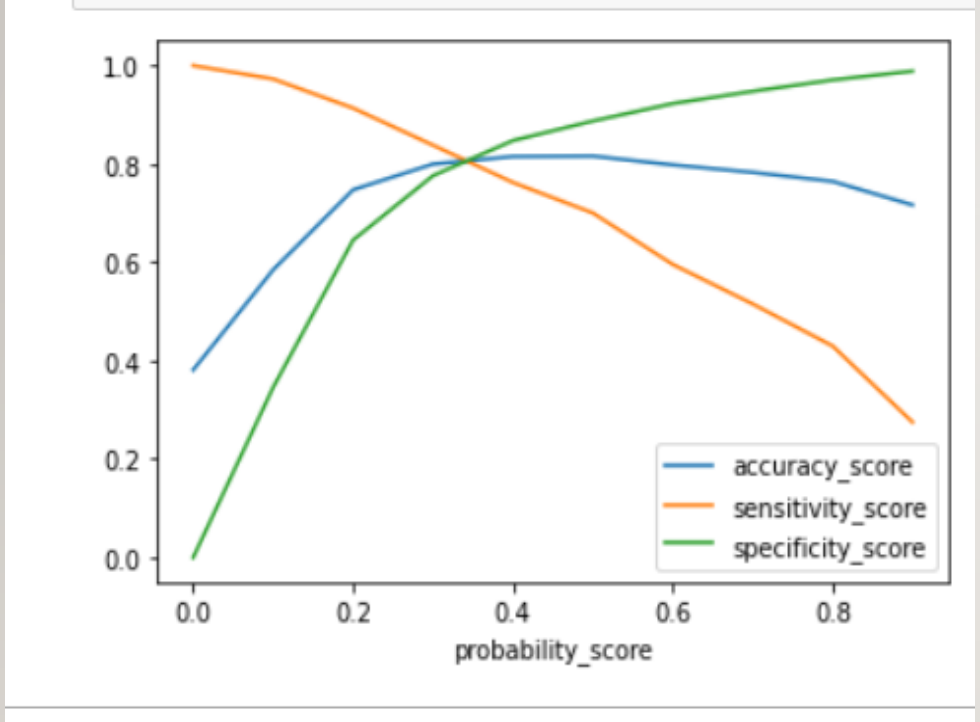
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.1607	0.244	-0.659	0.510	-0.639	0.317
Do Not Email	-1.2430	0.166	-7.493	0.000	-1.568	-0.918
Total Time Spent on Website	1.0909	0.039	27.731	0.000	1.014	1.168
Lead Origin_Lead Add Form	3.6704	0.190	19.343	0.000	3.299	4.042
Lead Source_Olark Chat	1.1621	0.102	11.366	0.000	0.962	1.362
Lead Source_Welingak Website	1.9421	0.742	2.616	0.009	0.487	3.397
Last Activity_Olark Chat Conversation	-0.8691	0.170	-5.099	0.000	-1.203	-0.535
Last Activity_Other Activity	2.0130	0.478	4.213	0.000	1.077	2.949
Last Activity_SMS Sent	1.2056	0.074	16.369	0.000	1.061	1.350
What is your current occupation_Unemployed	-1.0497	0.244	-4.309	0.000	-1.527	-0.572
What is your current occupation_Working Professional	1.7505	0.303	5.785	0.000	1.157	2.344
Lead Profile_Lateral Student	2.6532	1.098	2.417	0.016	0.502	4.804
Lead Profile_Student of SomeSchool	-2.4572	0.454	-5.409	0.000	-3.348	-1.567
How did you hear about X Education_Email	1.1561	0.600	1.926	0.054	-0.021	2.333
Last Notable Activity_Modified	-0.9157	0.080	-11.399	0.000	-1.073	-0.758

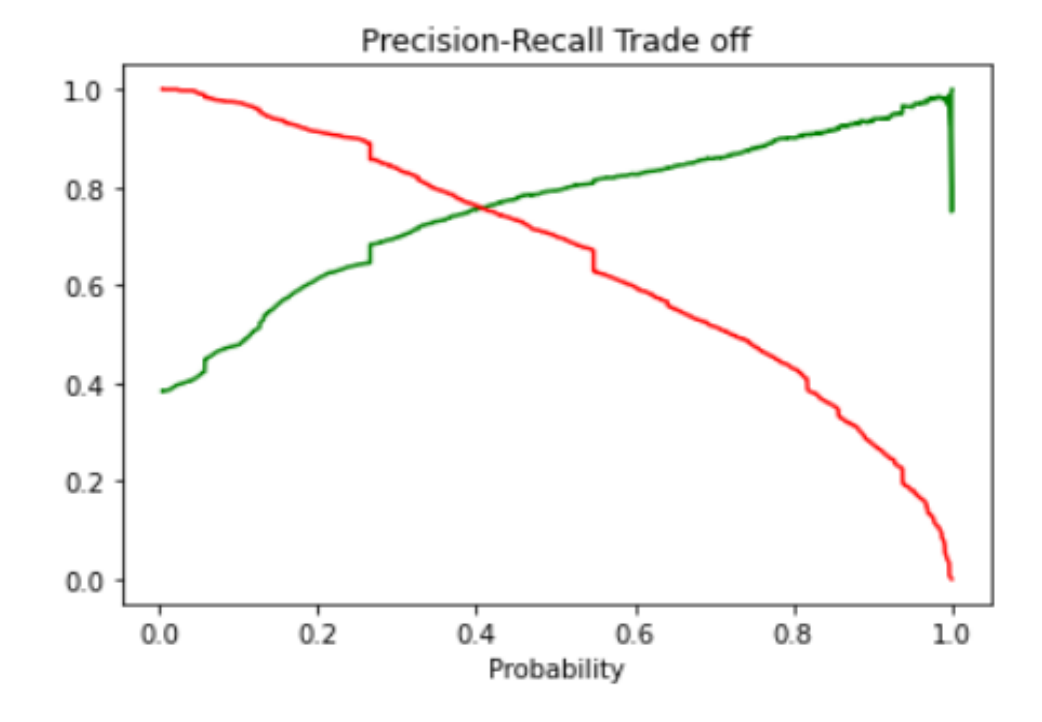
Model Evaluation : Plotting the ROC Curve

- Shows tradeoff between sensitivity and specificity (increase in one will cause decrease in other).
- The closer the curve follows the y-axis and then the top border of the ROC space, means more area under the curve and the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space i.e. the reference line, means less area and the less accurate is the test.



Finding optimal value of the cut off





Observation:

- In Sensitivity-Specificity-Accuracy plot 0.35 probability looks optimal. In Precision-Recall Curve 0.4 looks optimal.
- We are taking 0.35 is the optimum point as a cutoff probability and assigning Lead Score in training data

CONCLUSION

- The logistic regression model predicts the probability of the target variable having a certain value, rather than predicting the value of the target variable directly. Then a cutoff of the probability is used to obtain the predicted value of the target variable. Here, the logistic regression model is used to predict the probabilty of conversion of a customer. Optimum cut off is chosen to be 0.27 i.e. any lead with greater than 0.27 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.27 or less probability of converting is predicted as Cold Lead (customer will not convert) Our final Logistic Regression Model is built with 14 features

- The final model has Sensitivity of 0.978, this means the model is able to predict 98% customers out of all the converted customers, (Positive conversion) correctly.

- The final model has Precision of 0.427, this means 43% of predicted hot leads are True Hot Leads.

FEATURES USED IN THE MODEL

- 'Do Not Email',
- 'Total Time Spent on Website',
- 'Lead Origin_Lead Add Form',
- 'Lead Source_Olark Chat',
- 'Lead Source_Welingak Website',
- 'Last Activity_Olark Chat Conversation',
- 'Last Activity_Other Activity',
- 'Last Activity_SMS Sent',
- 'What is your current occupation_Unemployed',
- 'What is your current occupation_Working Professional',
- 'Lead Profile_Lateral Student',
- 'Lead Profile_Student of SomeSchool',
- 'How did you hear about X Education_Email',
- 'Last Notable Activity_Modified

The top three categorical/dummy variables in the final model and which are important from the business perspective are: 'Lead Origin_Lead Add Form', 'Lead Profile_Lateral Student', 'Last Activity_Other Activity'

The above variables are with respect to the absolute value of their coefficient factors.