X Education – Lead Scoring Case Study

Soumyashree Behera

Vedhavathi Nanjappa

Shravani Peddagari

Problem Statement

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
- X Education gets a lot of leads, its lead conversion rate is very poor at around 30%
- X Education wants to make lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads
- The sales team wants to know these potential set of leads, which they will be focusing more on communicating rather than making calls to everyone.
- The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Objective

- To help X Education select the most promising leads, i.e., the leads that are most likely to convert.
- The company requires us to build a logistic regression model wherein we 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.

Approach

- Data Reading and Understanding
- Data Cleaning
- Data Visualization (EDA)
- Feature Scaling & Model Building
- Model evaluation on Train Set
- Predictions on Test Dataset
- Recommendations

Data Reading and Understanding

Analyse following features

- Number of rows and columns
- Data types of each columns
- Checking how the data is spread
- Checking for any duplicate columns or dummy columns

Data Cleaning

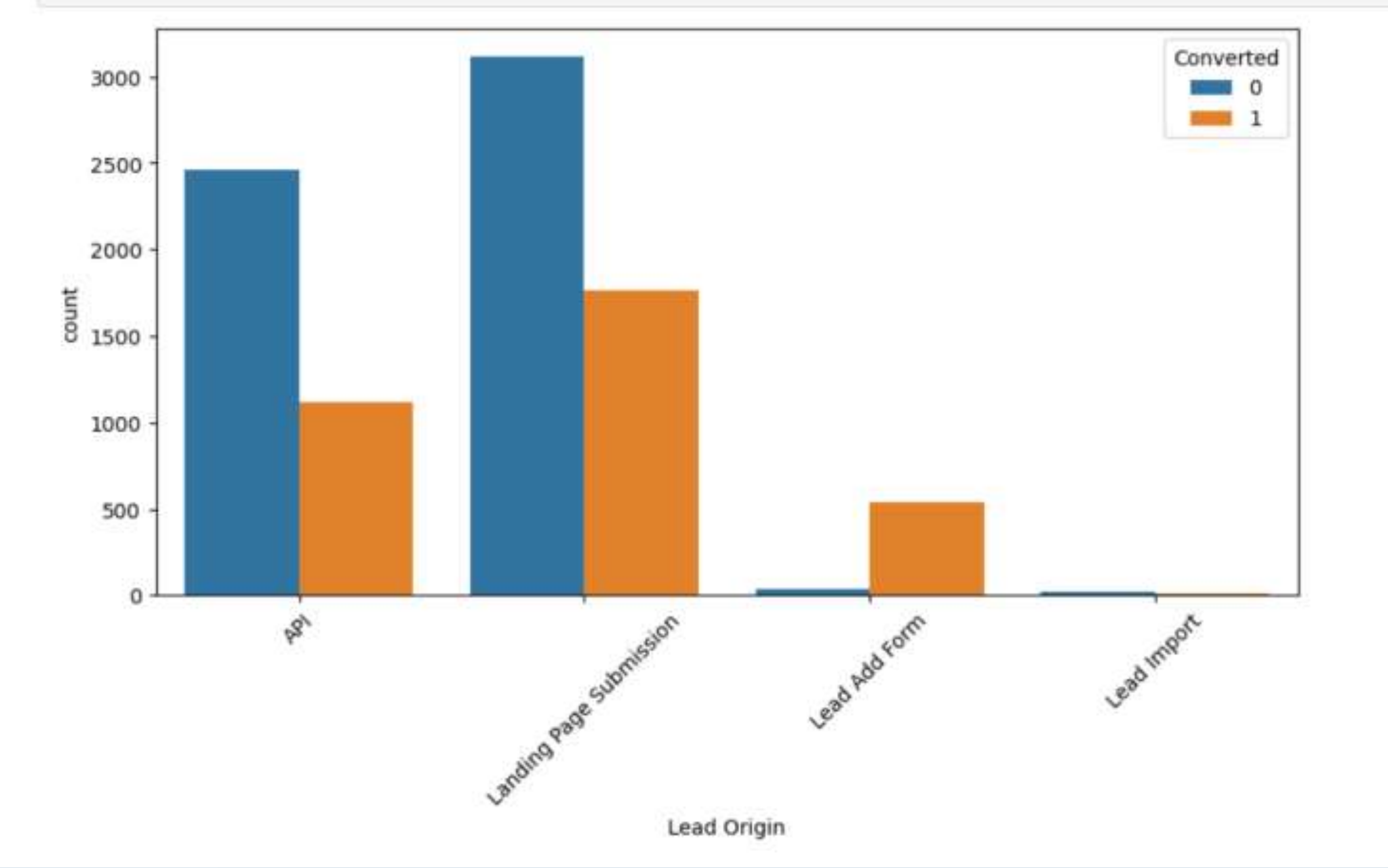
Check for discrepancies in the dataset.

- Checking for null values and imputing them with appropriate values
- Columns having >40 % null values are dropped
- We used mode imputation for categorical columns (Eg : City column)
- Drop the columns which are not of any importance in our analysis and visualization
- We have replaced the values of columns with low frequency into “Others” (Eg : Country column)

Exploratory Data Analysis

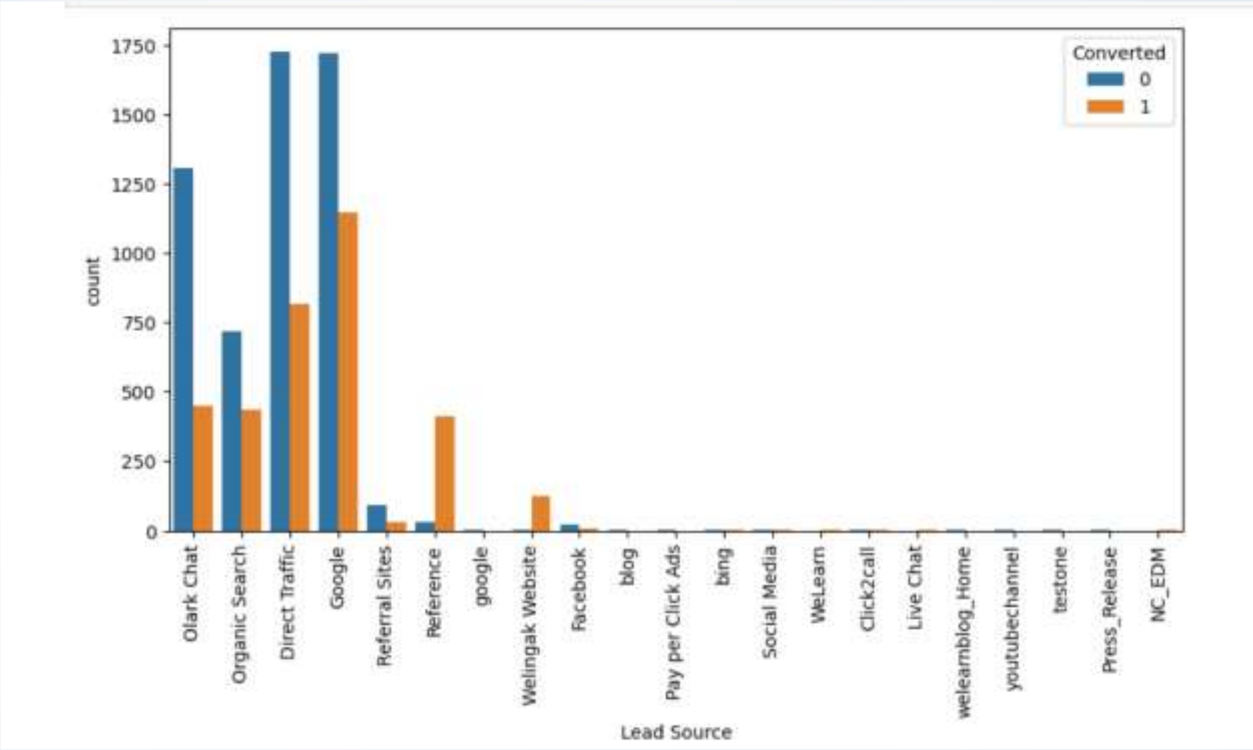
1. 'API' and 'Landing Page Submission' have a conversion rate of 30-36%, but there is a considerable amount of customer originated from this.
2. 'Lead Add Form' have a conversion rate of more than 90%, but the originated customer from this is less.
3. 'Lead Import' has a very low count and a low conversion rate as well.

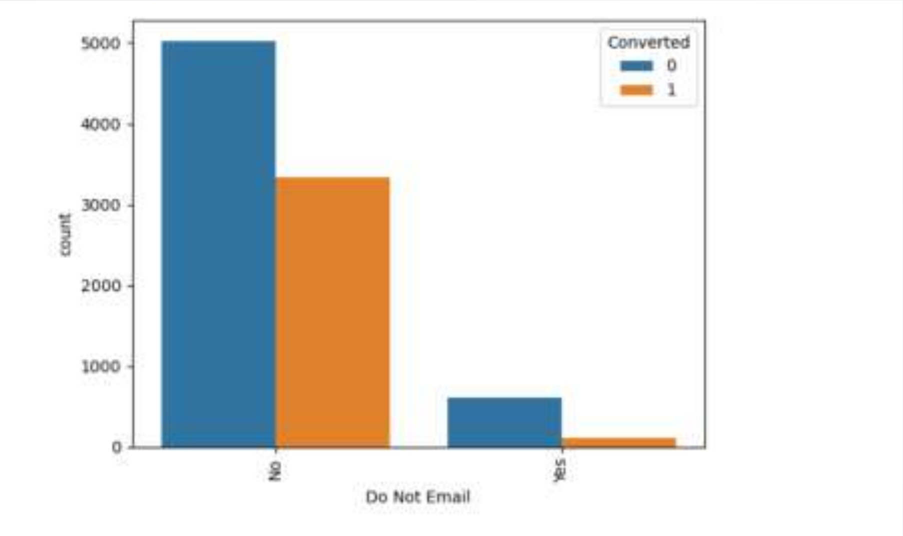
To improve the overall lead conversion rate, we need to focus more on improving the lead conversion from "API" and "Landing Page Submission" origin and generate more leads from "Lead Add Form".

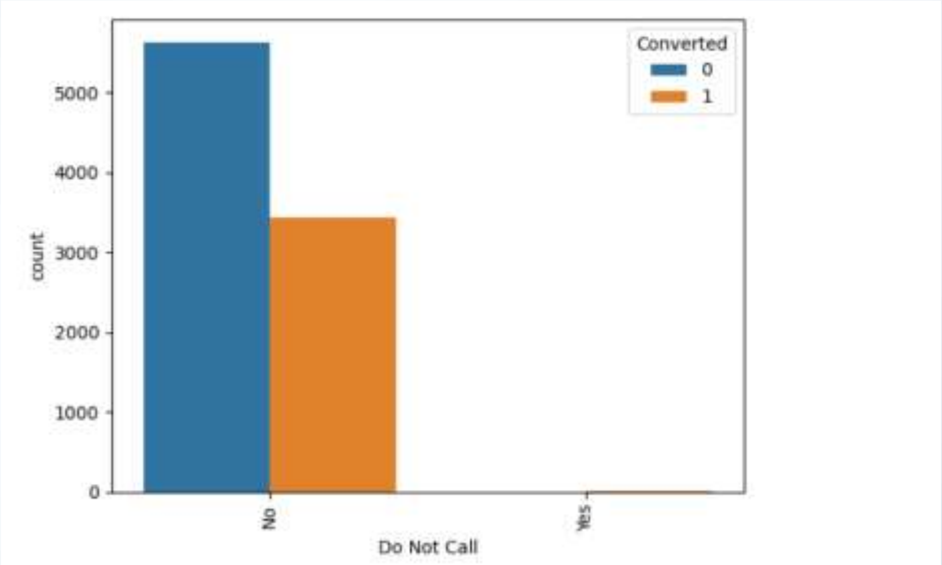


1. The top three lead sources are Google, Direct Traffic, and Olark Chat.
2. Conversion Rate of reference leads and leads through welingak website is high.

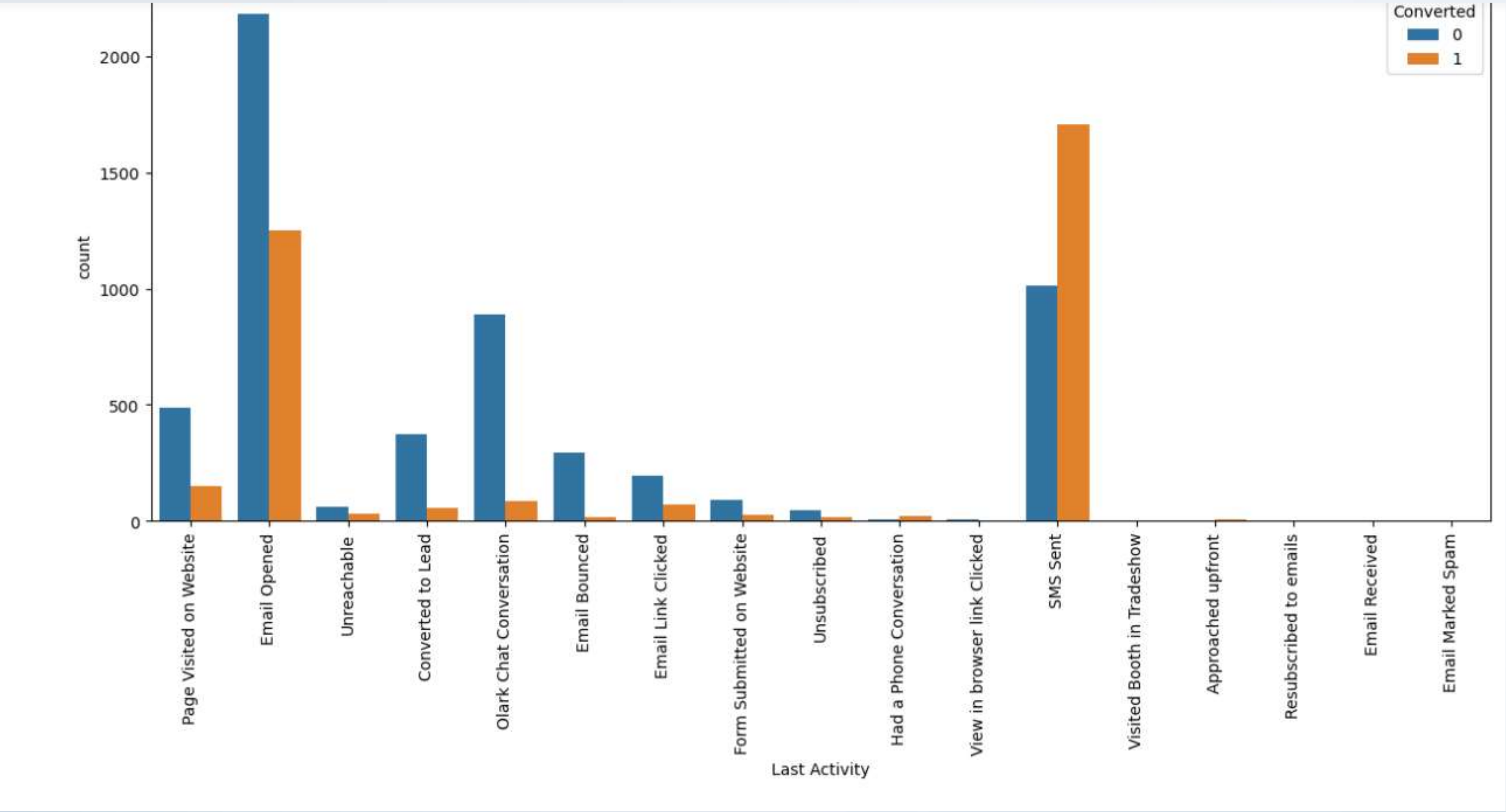
To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads source. Focus on generating more leads from reference and welingak website.



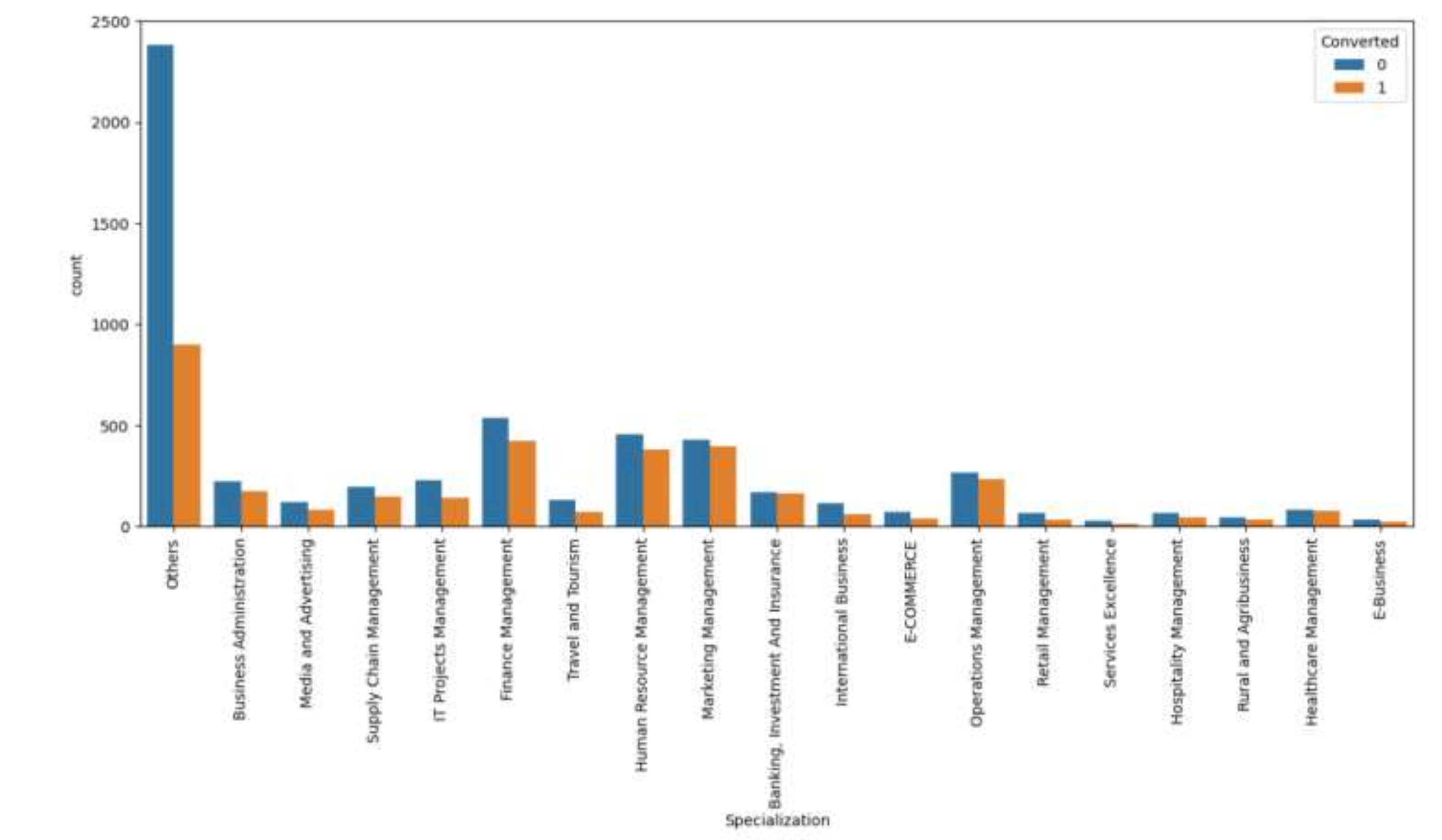




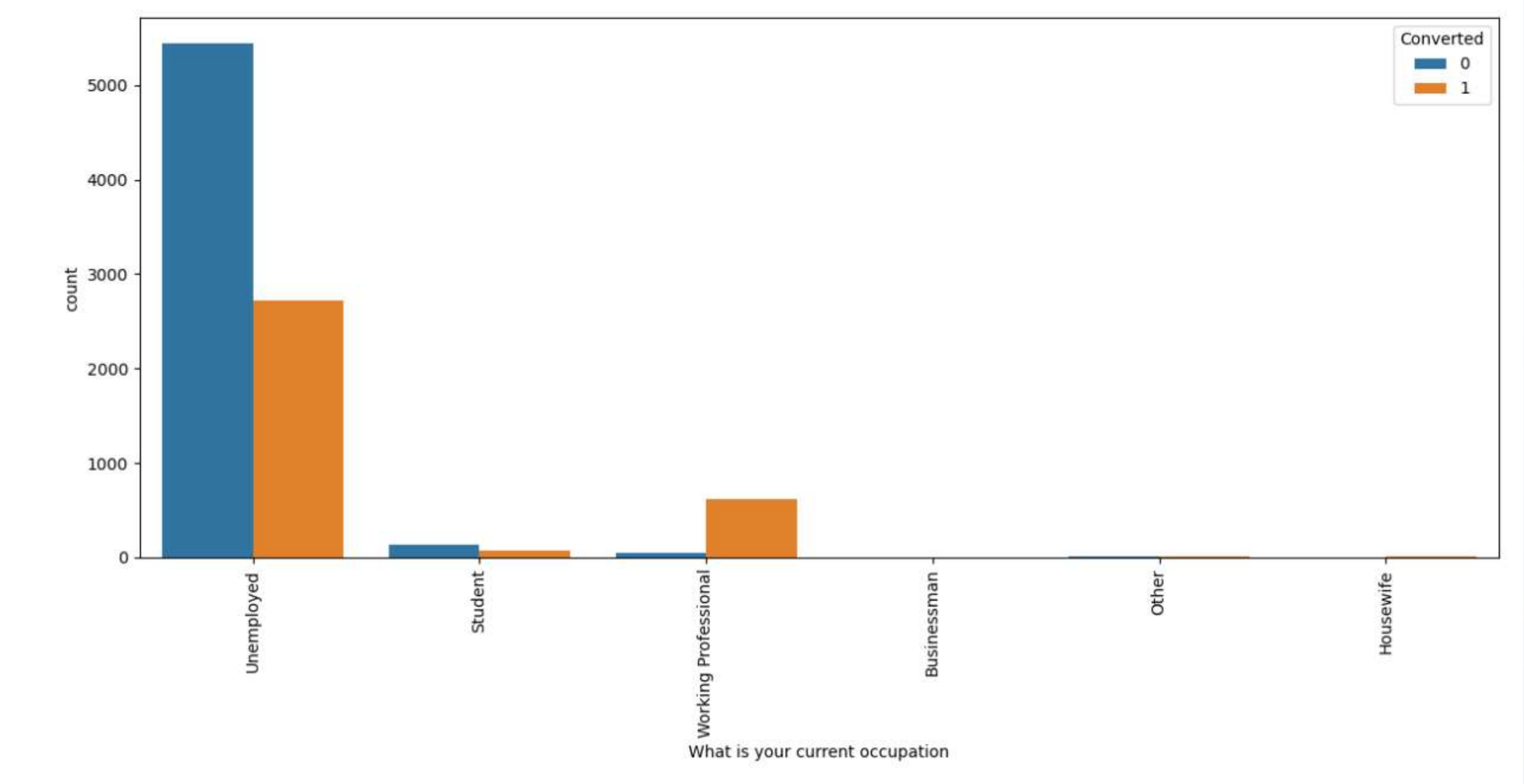
- The conversion rate for People who do not want an email is significantly lower than those who do want an Email



1.Email Opened and SMS Sent are the most common last activities before conversion.
2.Other_Activity has the lowest conversion rate among all last activities.
3.Email Link Clicked, Form Submitted on Website, and Page Visited on Website have moderate conversion rates.



- All specializations have a conversion rate in between 30-50%
- Focus can be given to those specialistion for those whose conversion rate is greater than 40%



Most responses are from the Unemployed Category, however their conversion rate is around 35%

Working Professional and Housewives have low count but high conversion rates

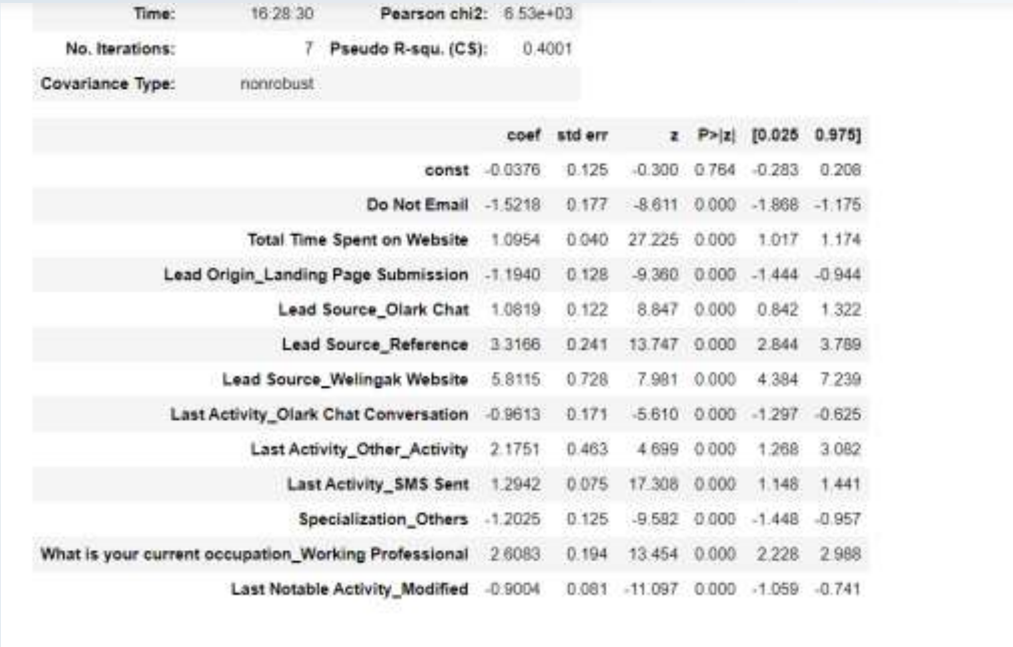
Data Preparation before Model building

- Binary level categorical columns are mapped to 1 / 0
- Created dummy features for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, Current occupation
- Splitting Train & Test Sets. 70:30 % ratio was chosen for the split
- Standardization method was used to scale the features

Feature Scaling & Model Building

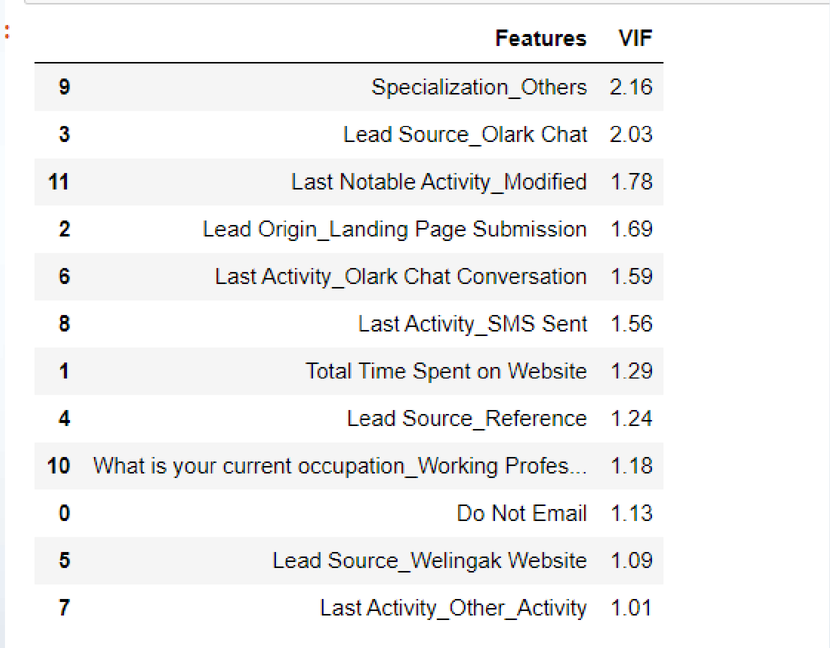
- Used Recursive Feature Elimination Technique to select top 20 featured based on which model has been built. RFE uses the model accuracy to identify which attributes contribute the most to predicting the target attribute. The model stability has been evaluated by making sure features have p value < 0.05 and VIF values < 5.
- Variance inflation factor(VIF) is used to treat multicollinearity. We achieved a stable model at the end of creation of 9th model.
- Once the stable model was created, we predicted probabilities on the train set and created a new column predicted with 1 if probability is greater than .5 else 0.
- We calculated the confusion matrix on this predicted column to the actual converted column. We also calculated the evaluation metrics Accuracy, Sensitivity, Specificity, Precision & Recall.
- We also plotted ROC curve to find the area under the curve (0.89 value we got).
- ROC curve shows the tradeoff between sensitivity and specificity

P-Value and VIF value



Time:	16:28:30	Pearson chi2:	6.53e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.4001
Covariance Type:	nonrobust		

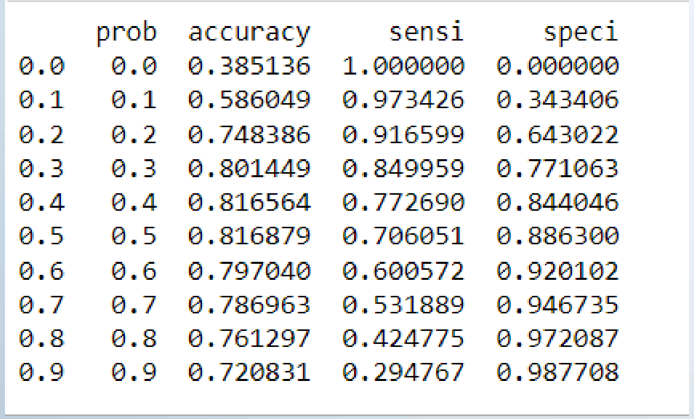
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0376	0.125	-0.300	0.764	-0.283	0.208
Do Not Email	-1.5218	0.177	-8.611	0.000	-1.868	-1.175
Total Time Spent on Website	1.0954	0.040	27.225	0.000	1.017	1.174
Lead Origin_Landing Page Submission	-1.1940	0.128	-9.360	0.000	-1.444	-0.944
Lead Source_Olark Chat	1.0819	0.122	8.847	0.000	0.842	1.322
Lead Source_Reference	3.3166	0.241	13.747	0.000	2.844	3.789
Lead Source_Welingak Website	5.8115	0.728	7.981	0.000	4.384	7.239
Last Activity_Olark Chat Conversation	-0.9613	0.171	-5.610	0.000	-1.297	-0.625
Last Activity_Other_Activity	2.1751	0.463	4.699	0.000	1.268	3.082
Last Activity_SMS Sent	1.2942	0.075	17.308	0.000	1.148	1.441
Specialization_Others	-1.2025	0.125	-9.582	0.000	-1.448	-0.957
What is your current occupation_Working Professional	2.6083	0.194	13.454	0.000	2.228	2.988
Last Notable Activity_Modified	-0.9004	0.081	-11.097	0.000	-1.059	-0.741



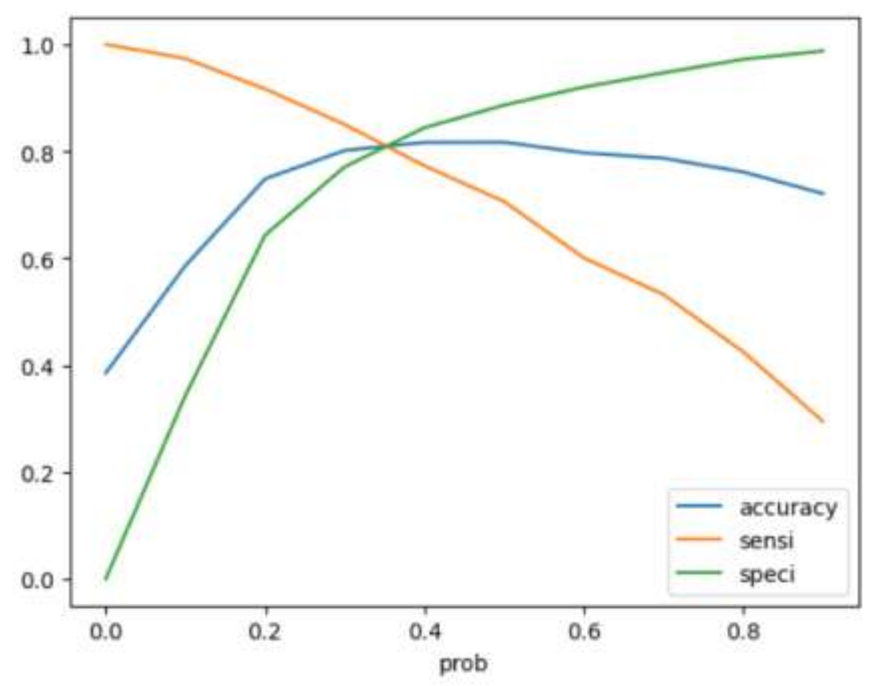
	Features	VIF
9	Specialization_Others	2.16
3	Lead Source_Olark Chat	2.03
11	Last Notable Activity_Modified	1.78
2	Lead Origin_Landing Page Submission	1.69
6	Last Activity_Olark Chat Conversation	1.59
8	Last Activity_SMS Sent	1.56
1	Total Time Spent on Website	1.29
4	Lead Source_Reference	1.24
10	What is your current occupation_Working Profes...	1.18
0	Do Not Email	1.13
5	Lead Source_Welingak Website	1.09
7	Last Activity_Other_Activity	1.01

P values of all variables is 0 and VIF values are low for all the variables, model9 is our final model. There are 12 variables considered in our final model.

Model Evaluation



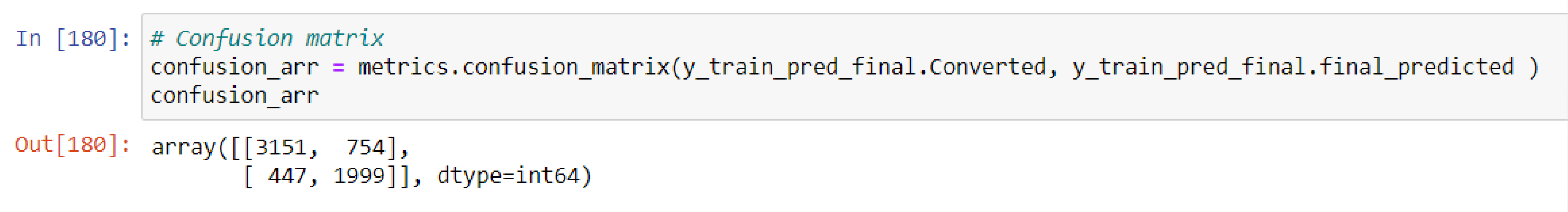
```
     prob  accuracy     sensi     speci
0.0   0.0  0.385136  1.000000  0.000000
0.1   0.1  0.586049  0.973426  0.343406
0.2   0.2  0.748386  0.916599  0.643022
0.3   0.3  0.801449  0.849959  0.771063
0.4   0.4  0.816564  0.772690  0.844046
0.5   0.5  0.816879  0.706051  0.886300
0.6   0.6  0.797040  0.600572  0.920102
0.7   0.7  0.786963  0.531889  0.946735
0.8   0.8  0.761297  0.424775  0.972087
0.9   0.9  0.720831  0.294767  0.987708
```



From the above curve, 0.34 is the optimal point to take it as a cutoff probability.

Confusion Matrix on train data set

Consider the Optimal cut-off value as 0.34. Below is the confusion matrix created in train data set.



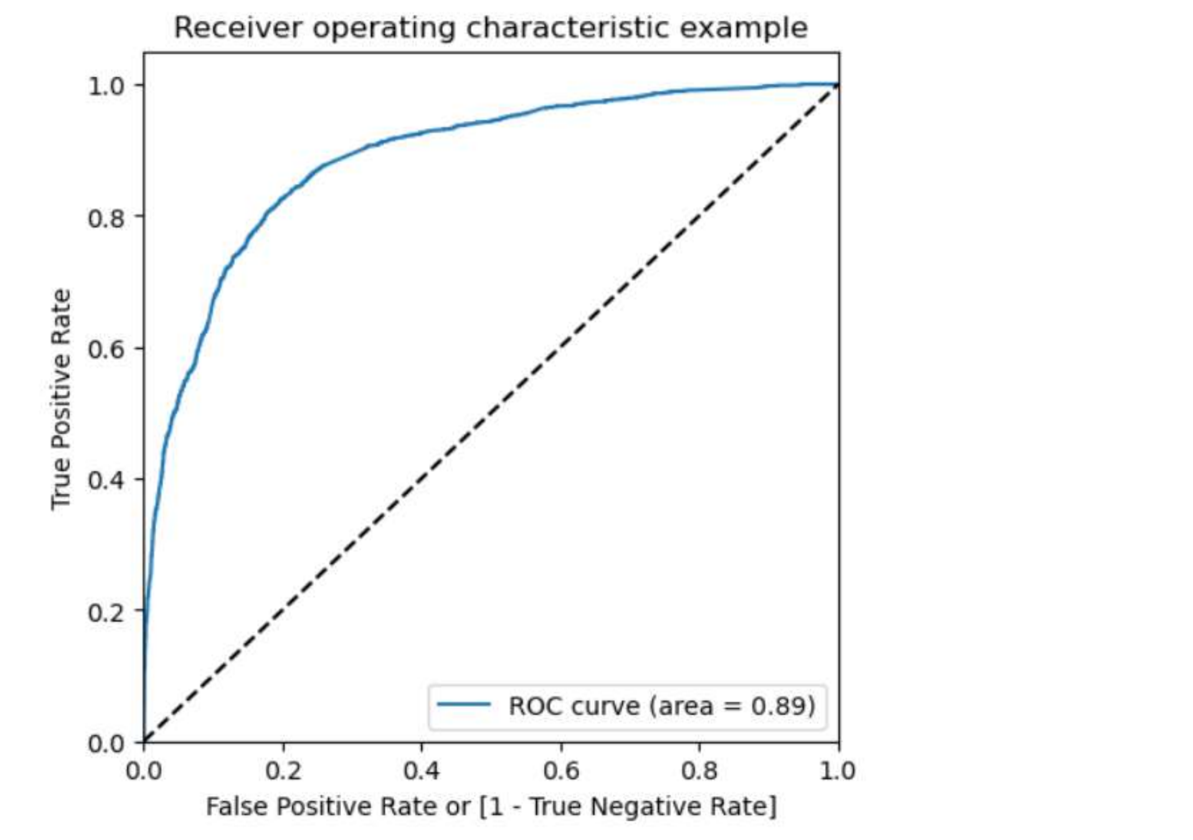
```
In [180]: # Confusion matrix
          confusion_arr = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.final_predicted )
          confusion_arr

Out[180]: array([[3151,  754],
                 [ 447, 1999]], dtype=int64)
```

- Accuracy: 81%
- Sensitivity: 81.7 %
- Specificity : 80.6 %

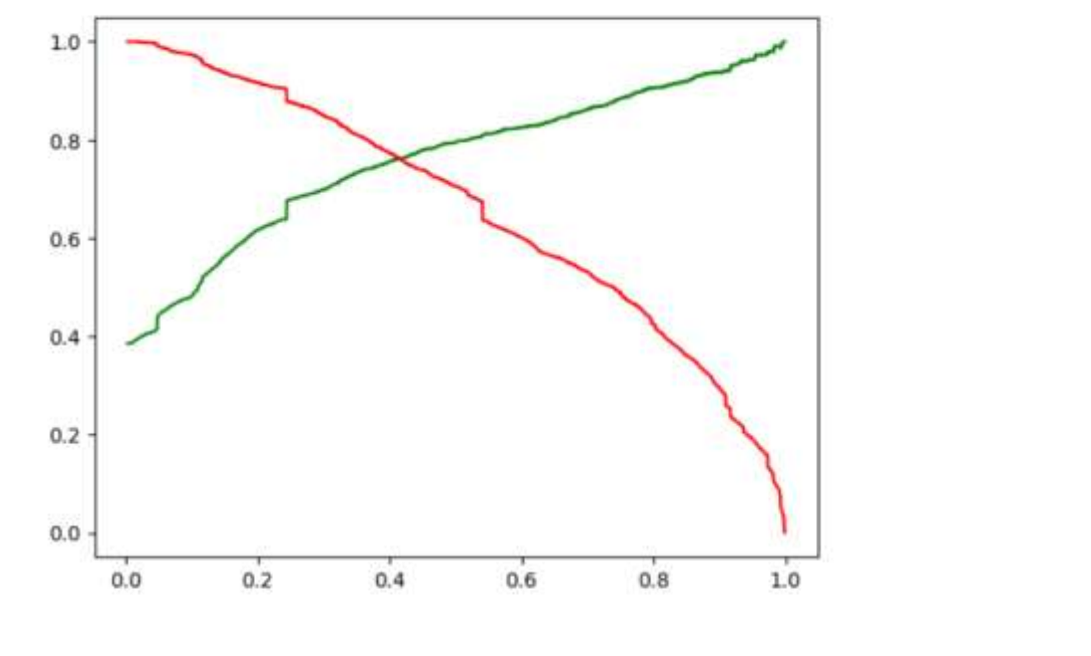
ROC Curve

- ROC curve shows the tradeoff between sensitivity and specificity



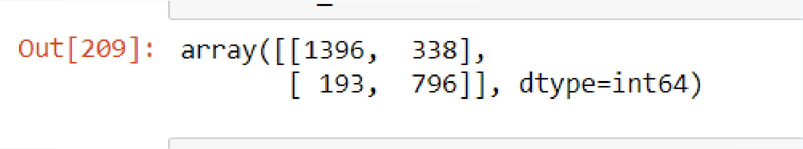
This model has an area under the ROC curve of 0.89, indicating that the model is good.

Plot Trade-off between Precision and Recall



Predictions on Test Dataset

Confusion Matrix



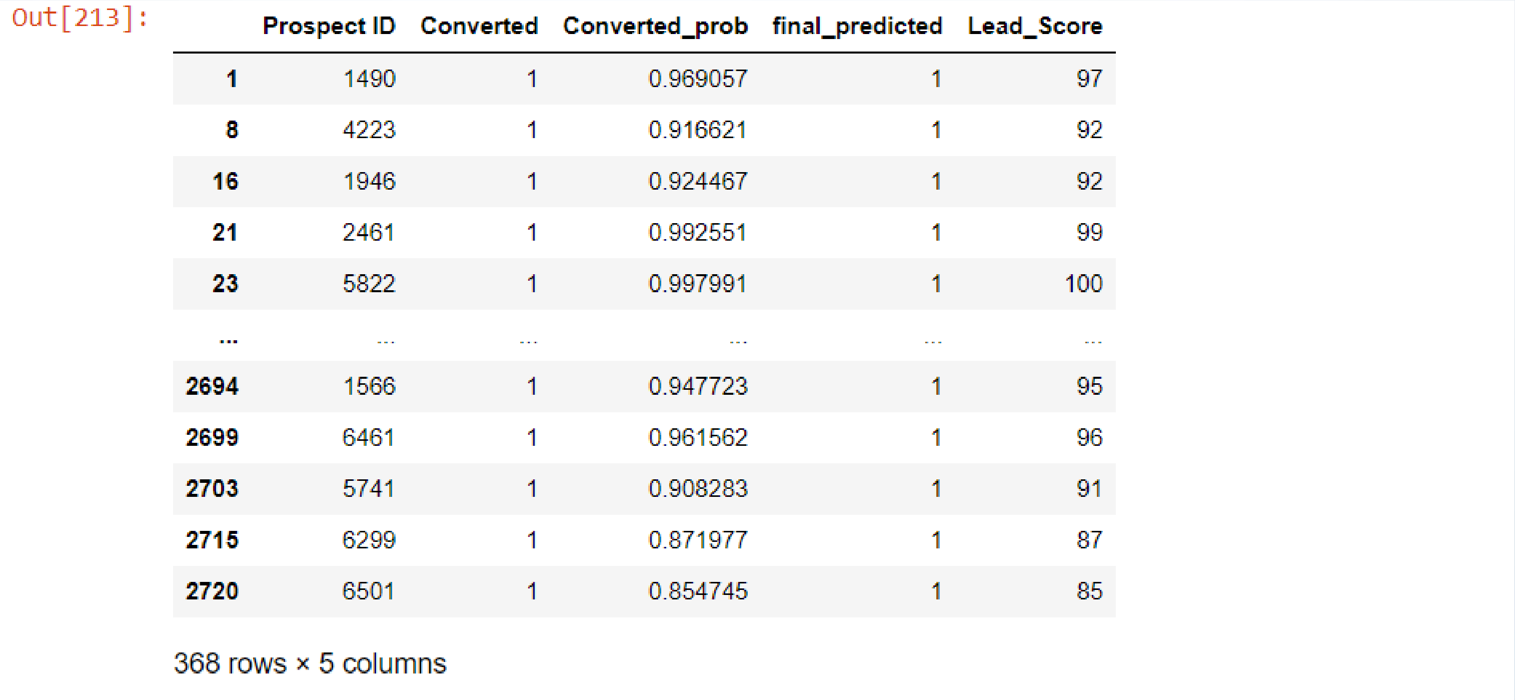
```
Out[209]: array([[1396,  338],
                 [ 193,  796]], dtype=int64)
```

Evaluation metrics on test data set

- Accuracy: 80.4%
- Sensitivity: 80.4%
- Specificity: 80.5%

Finding out the leads which should be contacted

- The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model to get a higher lead conversion rate of 80%.
- The customers which should be contacted are the customers whose "Lead Score" is equal to or greater than 85. They can be termed as 'Hot Leads'.
- There are 368 leads which can be contacted and have a high chance of getting converted.



Out[213]:

	Prospect ID	Converted	Converted_prob	final_predicted	Lead_Score
1	1490	1	0.969057	1	97
8	4223	1	0.916621	1	92
16	1946	1	0.924467	1	92
21	2461	1	0.992551	1	99
23	5822	1	0.997991	1	100
...	...	...	...	...	...
2694	1566	1	0.947723	1	95
2699	6461	1	0.961562	1	96
2703	5741	1	0.908283	1	91
2715	6299	1	0.871977	1	87
2720	6501	1	0.854745	1	85

368 rows × 5 columns

Recommendations

- The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
- The company should make calls to the leads who are the "working professionals" as they are more likely to get converted.
- The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.
- The company should make calls to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.
- The company should make calls to the leads whose last activity was SMS Sent as they are more likely to get converted.
- The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
- The company should not make calls to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.
- The company should not make calls to the leads whose Specialization was "Others" as they are not likely to get converted.
- The company should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.