Varun Eknath, Vikas Yadav, Abhilash Srivastava

Lead Scoring Case Study

Agenda

Problem Statement

Dataframe understanding & EDA

Numerical Attribute Analysis

Data Preparation & Train-Test Split

Plotting ROC curve & predictions

Conclusion

Problem Statement

An X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein you need to 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Primary goals

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.

Dataframe understanding & EDA

Dataframe Overview

- The Dataframe contains 37 columns and 9240 rows.
- It describes various prospects with their details like type of lead, contact details, profile activity, payments and payment methods, check to see if the potential customers were converted or not etc.
- All the entries, fortunately are non-null, but on further investigation the ‘Select’ option was found. This meant that the customer had not selected an option. This would be assigned as NaN and will be dropped.
- There were also no duplicate entries found in the dataframe.
- The presence of categorical values in the Datafram was noted, and a dummy value for these entries will be assigned.

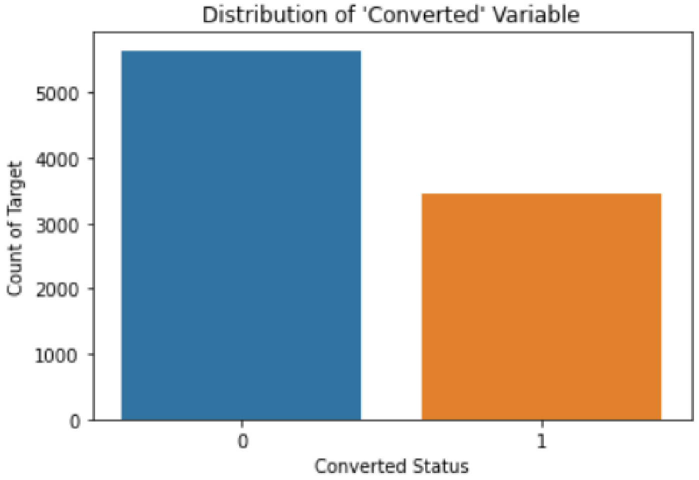
EDA

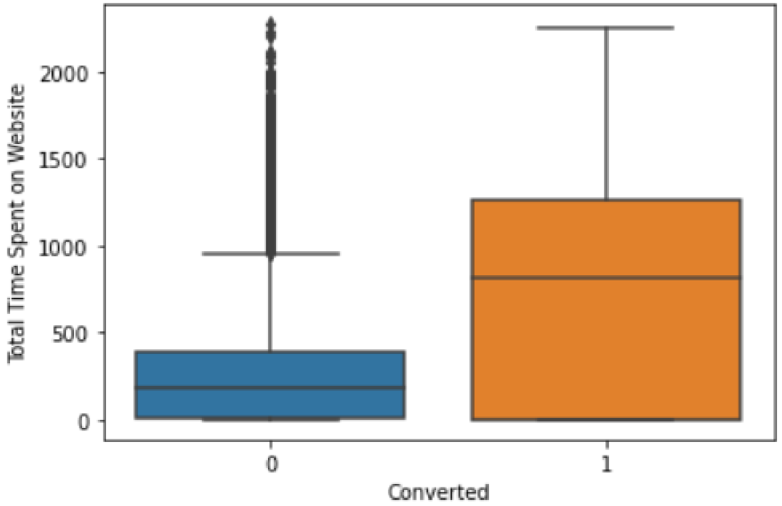
- ‘Select’ option is assigned as NaN, and treated as null-values. This signifies that the customer has not entered any value in the metric. We calculated the percentage of the null values and decided, rows with high percentage of null will be dropped i.e. percentage >=35%.
- Analysing the categorical values by plotting graphs we found that the columns except 'A free copy of Mastering The Interview' data is highly imbalanced. We decided to drop these columns as we deemed it redundant and felt it would throw off the prediction model.
- For missing values in the ‘Lead Source’ column we found that the highest occurring value was ‘ Google’, hence we imputed the same for other missing values in the column.
- We also inferred that Maximum Leads are generated by Google and Direct Traffic & Conversion rate of Reference leads and Welinkgak Website leads is very high.
- Regarding Conversion Rates & Leads, Maximum leads generated are unemployed and their conversion rate is more than 50%. & Conversion rate of working professionals is very high.
- We found that Maximum leads are generated having last activity as Email opened but conversion rate is not too good & SMS sent as last activity has high conversion rate.
- Columns ‘Country’ & ‘Occupation’ were not provided by the customer and were dropped. The ‘Do not call’ column metric was highly skewed and was also subsequently dropped.

Numerical Attribute Analysis

Converted Value Distribution

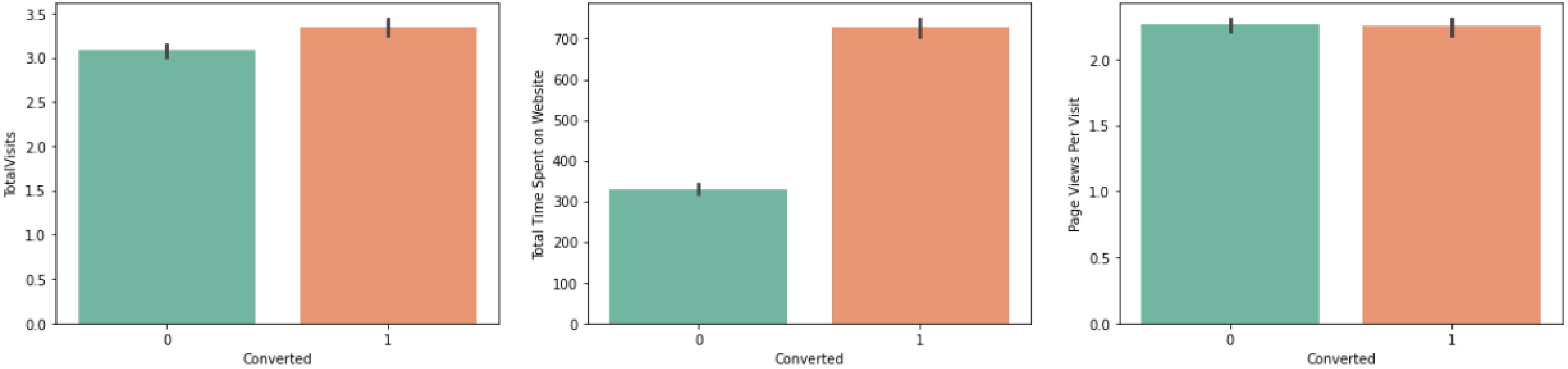
- The target variable distribution shows, that most of the customers were not converted. By taking the average we find Currently, lead Conversion rate is 38% only.





- Making a boxplot of Total time spent on websites vs. Converted values, we inferred leads spending more time on website are more likely to convert , thus website should be made more engaging to increase conversion rate

Converted Value Distribution



- Further investigating the ‘Page Views per visit’ , ‘Total Visits’ & ‘Total Time Spent on Website’ to the target variable by plotting boxplots for the same, we infer that more people who engage with the website more and stay in the website longer are most likely to get converted. This is taken as the Basis for our Machine Learning model.
- NOTE: Numerical Analysis that were inconclusive are not added to the presentation.

Data Preparation & Train-Test Split

Dataframe Preparation

- The columns like ‘Do not Email’ is converted from Yes & No to Binary. Dummy variables are assigned to columns like ‘Lead Origin, ‘ Lead Source’, ‘What is your current occupation’.

Train –Test Split

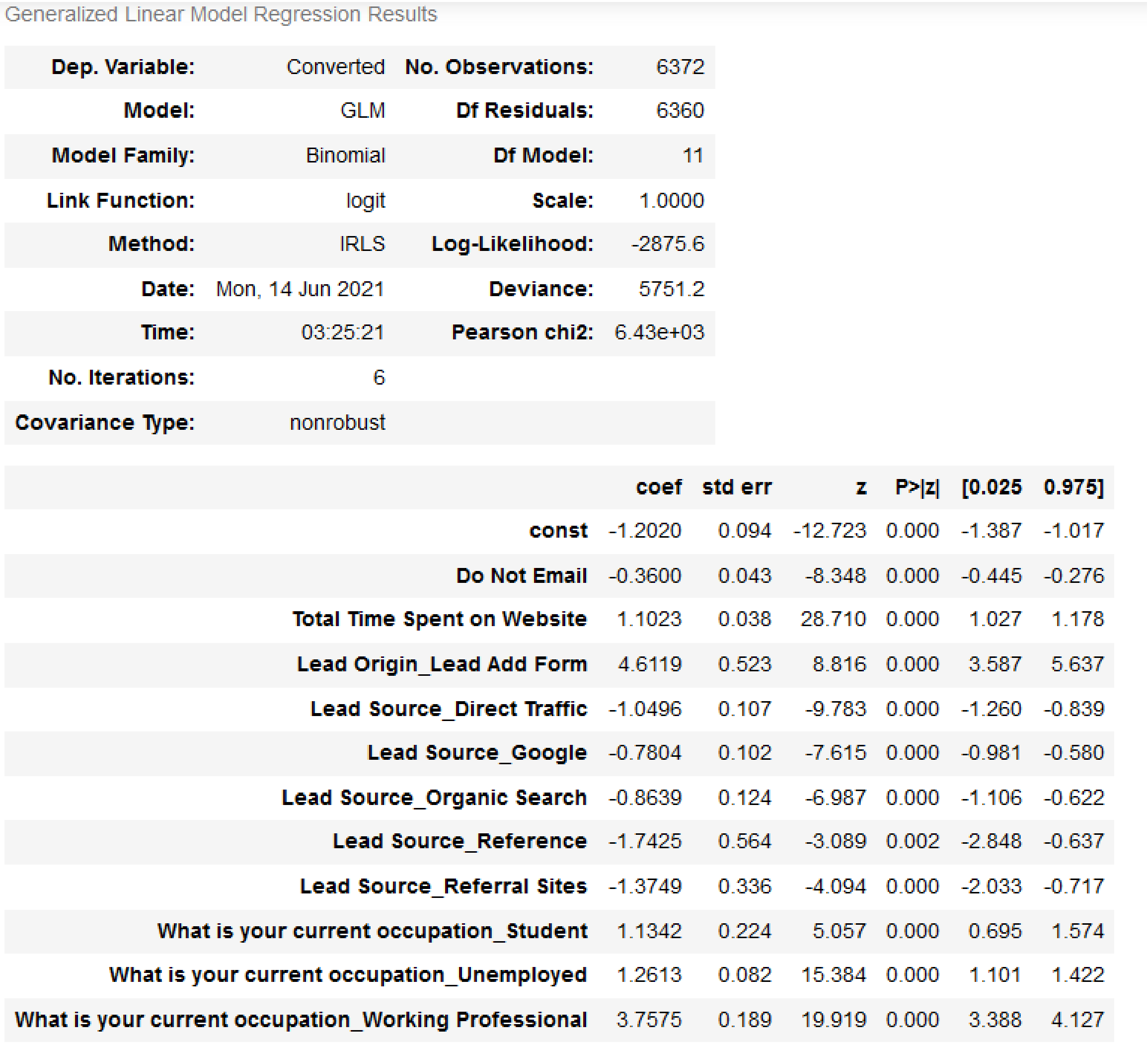
- We created a 70-30 spilt of both X_train, y_train & X_test & y_test cases from the dataframe, with X_train not having the target variable. A standard Scaler was used to Scale the model appropriately.
- We created a correlation and plotted a heatmap for the leads Dataframe. We found that ‘Lead Source_Olark chat’ and ‘Lead Origin_Landing Page Submission’ were highly correlated.
- These columns were subsequently dropped as we felt this would throw the prediction model off.

Using Sats model & RFE to create a predictive model on Train Set

- We created 5 models on the Train set to find the attributes of the dataset that might have high p-value and maybe redundant. The Models yielded the following results. The high p-value attributes were dropped:

- **Model 1:** p-value of the attribute 'What is current occupation – Housewife' was high
- **Model 2:** p-value of the attribute 'Lead_Source – Welingak Website' was high
- **Model 3:** p-value of the attribute 'What is current occupation – Businessman' was high
- **Model 4:** p-value of the attribute 'What is current occupation – Other' was high
- **Model 5:** p-value of all the attributes was nominal. The model details is shown in the picture.

- We VIF for the model attributes were also nominal as none of them were higher than 5, highest being 3.81.



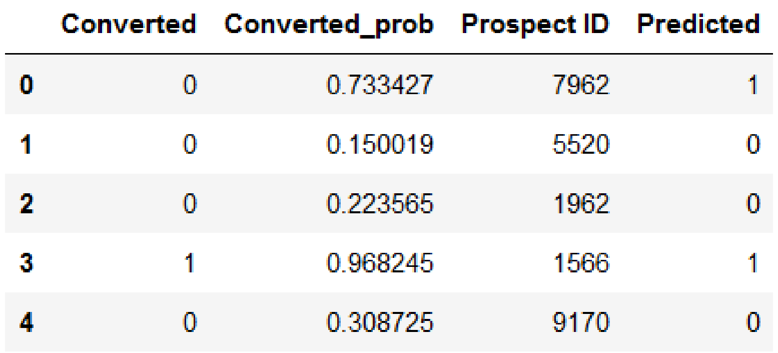
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6372
Model:	GLM	**Df Residuals:**	6360
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2875.6
Date:	Mon, 14 Jun 2021	**Deviance:**	5751.2
Time:	03:25:21	**Pearson chi2:**	6.43e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-1.2020	0.094	-12.723	0.000	-1.387	-1.017
Do Not Email	-0.3600	0.043	-8.348	0.000	-0.445	-0.276
Total Time Spent on Website	1.1023	0.038	28.710	0.000	1.027	1.178
Lead Origin_Lead Add Form	4.6119	0.523	8.816	0.000	3.587	5.637
Lead Source_Direct Traffic	-1.0496	0.107	-9.783	0.000	-1.260	-0.839
Lead Source_Google	-0.7804	0.102	-7.615	0.000	-0.981	-0.580
Lead Source_Organic Search	-0.8639	0.124	-6.987	0.000	-1.106	-0.622
Lead Source_Reference	-1.7425	0.564	-3.089	0.002	-2.848	-0.637
Lead Source_Referral Sites	-1.3749	0.336	-4.094	0.000	-2.033	-0.717
What is your current occupation_Student	1.1342	0.224	5.057	0.000	0.695	1.574
What is your current occupation_Unemployed	1.2613	0.082	15.384	0.000	1.101	1.422
What is your current occupation_Working Professional	3.7575	0.189	19.919	0.000	3.388	4.127

Predicting a Train Model & finding Prediction Metrics

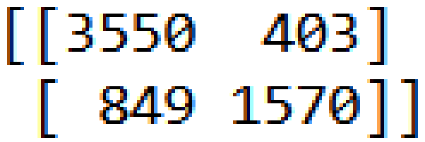
- We used the previously created Train model and with sklearn made predictions on the same. This yielded the below values for the first 5 entries:



	Converted	Converted_prob	Prospect ID	Predicted
0	0	0.733427	7962	1
1	0	0.150019	5520	0
2	0	0.223565	1962	0
3	1	0.968245	1566	1
4	0	0.308725	9170	0

- The metrics for the predicted model, i.e. **Accuracy, Sensitivity, Specificity, False Positive Rate, Positive Predictive Value and Negative Predictive Value** are as follows:
- Accuracy: **0.8035**
- Sensitivity: **0.649**
- Specificity: **0.898**
- False Positive Rate: **0.102**
- Positive Predictive Value: **0.795**
- Negative Predictive Value: **0.807**

- The confusion Matrix using sklearn gives the below matrix:



```
[[3550  403]
 [ 849 1570]]
```

Plotting ROC curve & predictions

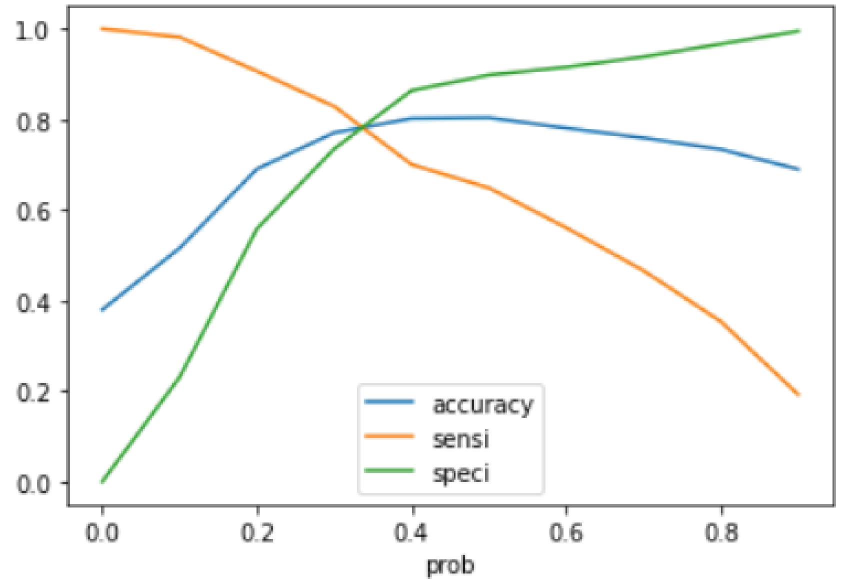
Requirement for ROC Curve

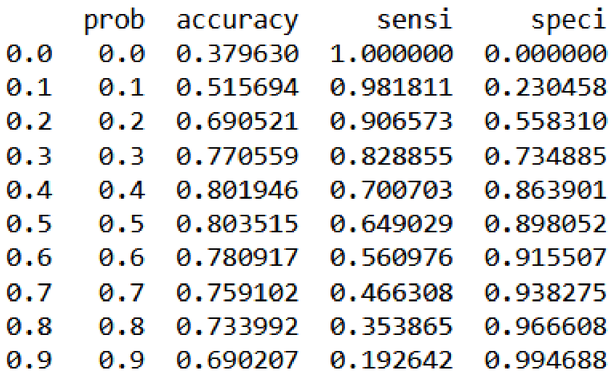
- An ROC curve demonstrates several things:
- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

.

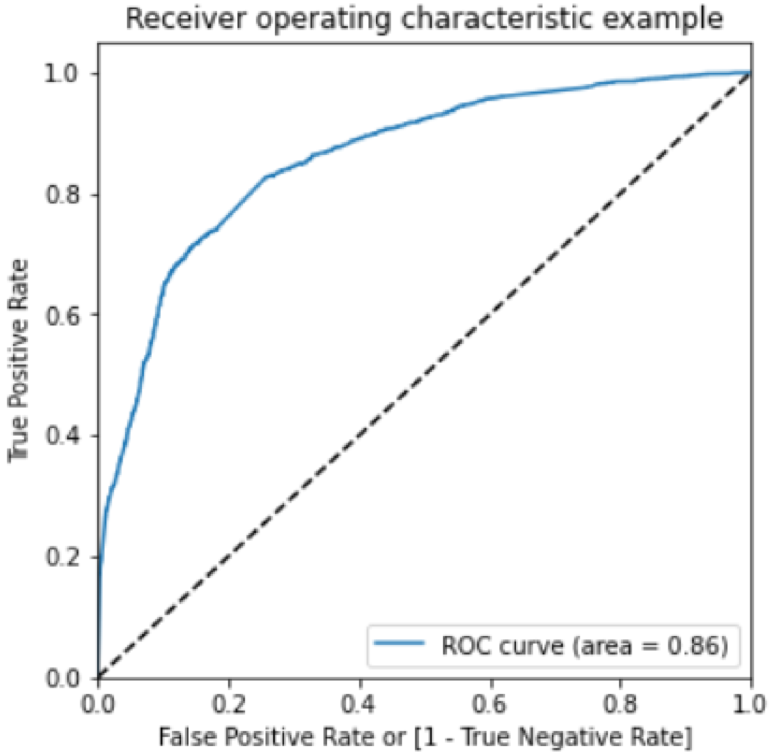
ROC Curve Inference

- By plotting ROC Curve for the model we find the curve value as 0.86 which is close to 1. This signifies this is a reliable model
- We also find the Optimal Cutoff point which is the is that probability where we get balanced sensitivity and specificity





```
     prob  accuracy     sensi     speci
0.0   0.0  0.379630  1.000000  0.000000
0.1   0.1  0.515694  0.981811  0.230458
0.2   0.2  0.690521  0.906573  0.558310
0.3   0.3  0.770559  0.828855  0.734885
0.4   0.4  0.801946  0.700703  0.863901
0.5   0.5  0.803515  0.649029  0.898052
0.6   0.6  0.780917  0.560976  0.915507
0.7   0.7  0.759102  0.466308  0.938275
0.8   0.8  0.733992  0.353865  0.966608
0.9   0.9  0.690207  0.192642  0.994688
```



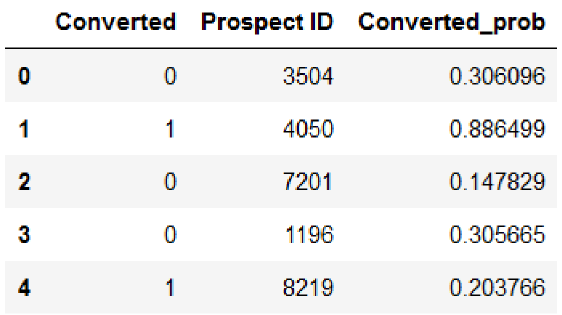
From the curve above, 0.3 is the optimum point to take it as a cutoff probability.

Train Test Conclusion:

By making a Train Prediction from the above inference we can calculate the total of final predicted conversion / non conversion counts from the actual converted rates for which we get check the precentage of final_predicted conversions as **0.828.** This above the 80% conversions needed by X educations.

Predictions on the test set

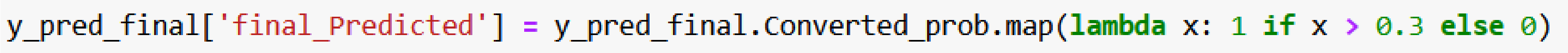
- We now run the predictive test model to find the capability of the test cases. The test case is first scaled and made to fit the Train model by using scaler.fit_transform. For X_Test constant is assigned.
- From the newly modified X_Test, y_test is predicted. y_pred_1 dataframe is created from the y_test prediction.
- The conversion target attribute predicted this way is given below for the first 5 entries



	Converted	Prospect ID	Converted_prob
0	0	3504	0.306096
1	1	4050	0.886499
2	0	7201	0.147829
3	0	1196	0.305665
4	1	8219	0.203766

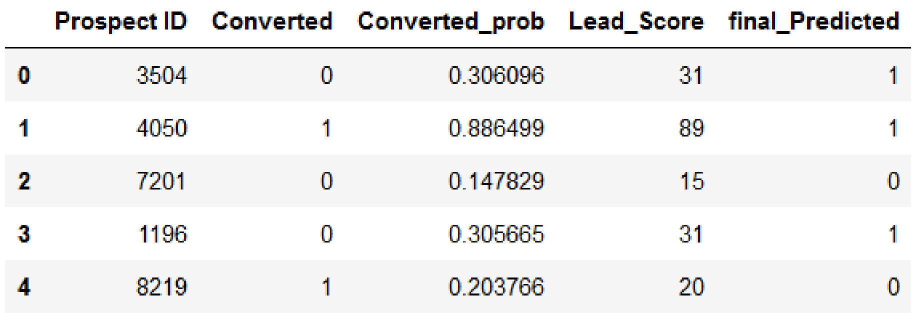
Assigning Lead Score

- To check whether the 80% is achieved for test cases we make a prediction based on the modified y_test for the columns 'Prospect ID’, 'Converted’, 'Converted_prob’\
- We use the following condition for the y_test_final



```
y_pred_final['final_Predicted'] = y_pred_final.Converted_prob.map(lambda x: 1 if x > 0.3 else 0)
```

- We the above operation we get the below dataset:

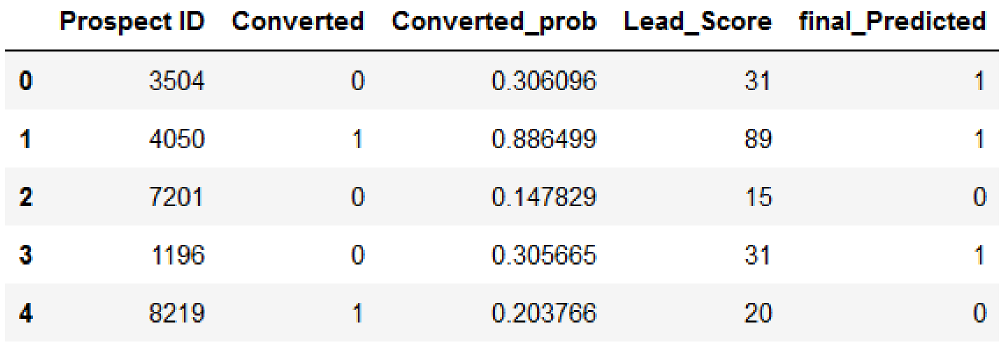


	Prospect ID	Converted	Converted_prob	Lead_Score	final_Predicted
0	3504	0	0.306096	31	1
1	4050	1	0.886499	89	1
2	7201	0	0.147829	15	0
3	1196	0	0.305665	31	1
4	8219	1	0.203766	20	0

Test Case Conclusion: By making a Train Prediction from the above inference we can calculate the total of final predicted conversion / non conversion counts from the actual converted rates for which we get check the precentage of final_predicted conversions as 0.83. Hence we can see that the final prediction of conversions have a target rate of 83% (same as predictions made on training data set)

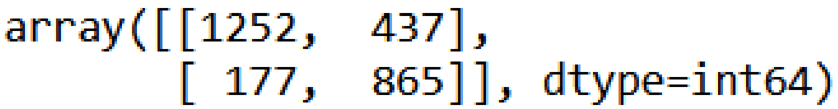
Test Prediction Metrics

- The metrics for the (test case)predicted model, i.e. **Accuracy, Sensitivi** **Specificity** are as follows:
- Accuracy: **0.775**
- Sensitivity: **0.83**
- Specificity: **0.74**



	Prospect ID	Converted	Converted_prob	Lead_Score	final_Predicted
0	3504	0	0.306096	31	1
1	4050	1	0.886499	89	1
2	7201	0	0.147829	15	0
3	1196	0	0.305665	31	1
4	8219	1	0.203766	20	0

- The confusion Matrix using sklearn gives the below matrix:



```
array([[1252,  437],
       [ 177,  865]], dtype=int64)
```

Conclusion

- An While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 77%, 83% and 74% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- Hence overall this model seems to be good.

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :

1. Lead Origin_Lead Add Form
2. What is your current occupation_Working Professional
3. Total Time Spent on Website

Thank you

Varun Eknath

Vikas Yadav

Abhilash Srivastava