Lead Scoring Case Study

SHRESTA KOURLA

CHANDRA SHEKHAR BANERJEE

Objective:

- Build a model in order to increase the lead conversion rate for **X Education** i.e.: get in touch with those leads who are more likely to be converted into a customer.
- In order to achieve the objective, we need to build a logistic regression model which can assign a score starting from 0 to 100 next to the lead . A higher score means a lead who is most likely to convert or a hot lead. Whereas, a lower score means a cold lead which is less likely to be converted
  - Predict a Lead Conversion Probability for each lead
  - Decide the cutoff above which a lead will be predicted as converted
  - From Lead Conversion Probability calculate Lead Score for each Lead

Conducted Steps:

- Understand the data frame by conducting EDA and removed the non-required dataset as well as imputing missing value
- Split the data into Train & Test set and scale the features
- Run Logistic Regression Model and use RFE and remove columns with high p-value and VIF
- Evaluate the model with various metrics like Accuracy, Sensitivity, Specificity, Precision, Recall etc.
- Find the Optimal Cutoff point and predict the dependent variable based on probability threshold value
- Use the model on the test dataset and perform the model evaluation
- Calculate the lead score for the dataset in order to meet the final objective

Data Preparation:

Removed Columns

- Columns are removed initially based on multiple criteria:
- Based on missing value - If more than 70% (**'How did you hear about X Education',' Lead Profile')**
- Columns where only one unique value is present (**'Magazine', 'Receive More Updates About Our Courses',' Update me on Supply Chain Content',' Get updates on DM Content',' I agree to pay the amount through cheque', 'What matters most to you in choosing a course'**)
- Columns which are mainly an index value and score based on index and thus difficult to impute (**'Asymmetrique Profile Index',' Asymmetrique Activity Index',' Asymmetrique Profile Score',' Asymmetrique Activity Score'**)
- Columns Based on Geographical Information ('City','Country')

Removed Rows

- Removed rows where more than 20% of total values are missing

Data Preparation:

Replace ‘Select’ with Nan

- ‘Select’ in the dataset is as equal to null in the dataset. Hence it is replaced with null

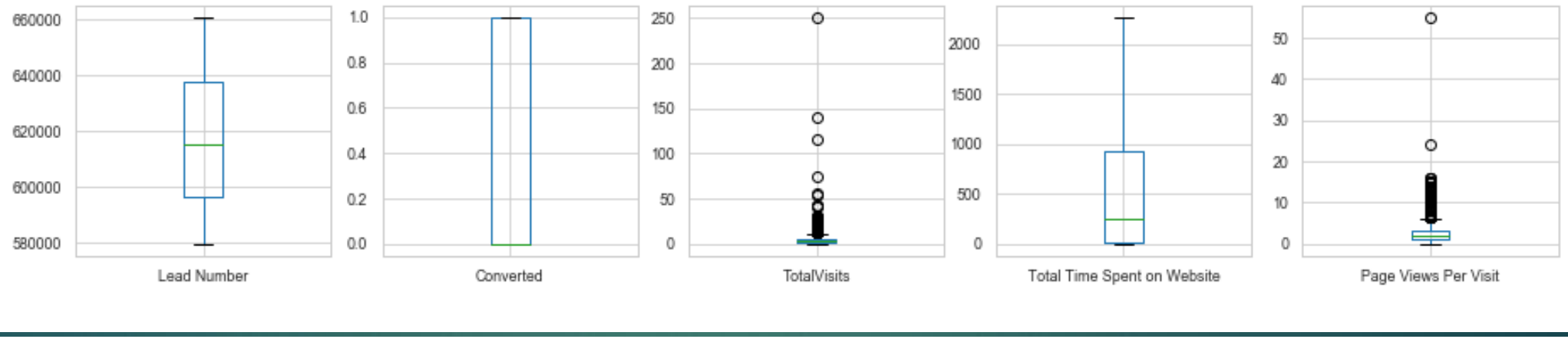
Impute Null values with Median and Mode

- Null values are replaced with Median (**‘TotalVisits’,’ Page Views Per Visit’**) and Mode (**‘Last Activity’,’ Lead Source’)**

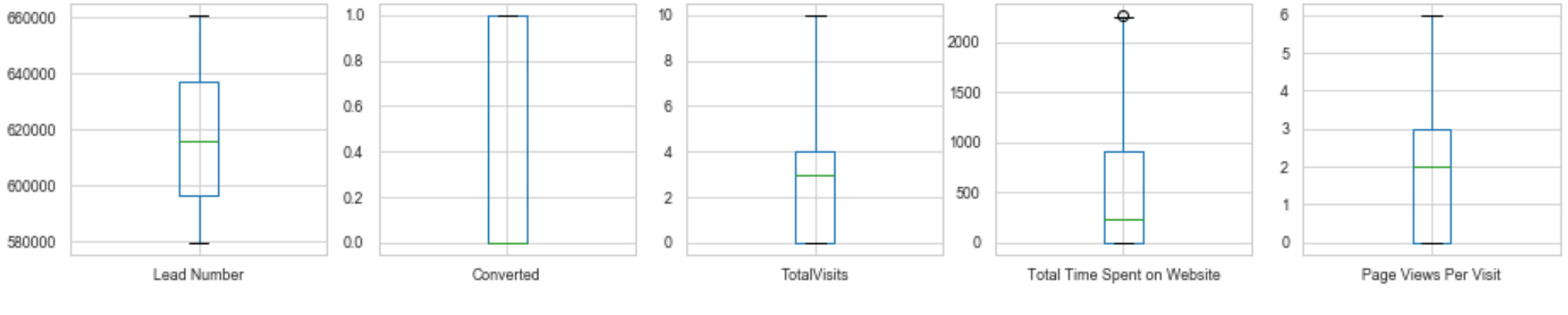
Data Preparation:

- Outlier Treatment:
  - Outliers are removed from the data based on upper quartile(0.75) and lower quartile(0.25)

Before Outlier treatment:



After Outlier treatment:



Data Preparation:

Convert binary variables from ‘Yes/No’ to ‘0/1’:

- List of variables are: **Do Not Email', 'Do Not Call', 'Search', 'Newspaper Article‘ , 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'A free copy of Mastering The Interview‘**

Convert categorical variables into dummy variable

- List of variables are : **'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'Tags', 'Lead Quality', 'Last Notable Activity'**

Feature Removed

- Originating features of dummy variables are removed

Data Preparation:

Train-Test split

- The dataset is divided into training and test dataset by 70:30 ratio.
- Training dataset is used to build the model whereas the test dataset is used to test the model.

Feature Scaling

- Scaling is done for all the features to bring all the numeric features into same scale

Feature Elimination using RFE

- RFE is used for 20 features initially and checked the p-value and VIF based on p-value (<0.05) and VIF(<5).
- Higher p-value features are eliminated from the dataset. However, VIF is always less than 5 in all the cases

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3889	0.078	-17.723	0.000	-1.543	-1.235
Lead Origin_Lead Add Form	0.4011	0.369	1.087	0.277	-0.322	1.124
Lead Source_Welingak Website	4.5684	1.094	4.177	0.000	2.425	6.712
Last Activity_SMS Sent	2.2639	0.114	19.946	0.000	2.041	2.486
Tags_Closed by Horizzon	8.1073	1.026	7.898	0.000	6.095	10.119
Tags_Diploma holder (Not Eligible)	-1.5002	1.085	-1.382	0.167	-3.627	0.627
Tags_Interested in other courses	-1.4893	0.404	-3.684	0.000	-2.282	-0.697
Tags_Lateral student	26.1991	1.19e+05	0.000	1.000	-2.33e+05	2.33e+05
Tags_Lost to EINS	7.0255	0.675	10.405	0.000	5.702	8.349
Tags_Not doing further education	-23.0640	1.83e+04	-0.001	0.999	-3.58e+04	3.58e+04
Tags_Ringing	-3.5426	0.257	-13.808	0.000	-4.045	-3.040
Tags_Will revert after reading the email	4.8939	0.187	26.217	0.000	4.528	5.260
Tags_invalid number	-3.1666	1.040	-3.044	0.002	-5.205	-1.128
Tags_number not provided	-25.2686	4.33e+04	-0.001	1.000	-8.49e+04	8.48e+04
Tags_opp hangup	-2.6714	1.071	-2.493	0.013	-4.771	-0.571
Tags_switched off	-4.9461	1.011	-4.891	0.000	-6.928	-2.964
Tags_wrong number given	-25.1882	3.76e+04	-0.001	0.999	-7.38e+04	7.37e+04
Lead Quality_Worst	-2.6486	0.574	-4.615	0.000	-3.773	-1.524
Last Notable Activity_Email Link Clicked	-1.2958	0.473	-2.742	0.006	-2.222	-0.370
Last Notable Activity_Modified	-1.9620	0.126	-15.533	0.000	-2.210	-1.714
Last Notable Activity_Olark Chat Conversation	-1.4763	0.447	-3.304	0.001	-2.352	-0.601

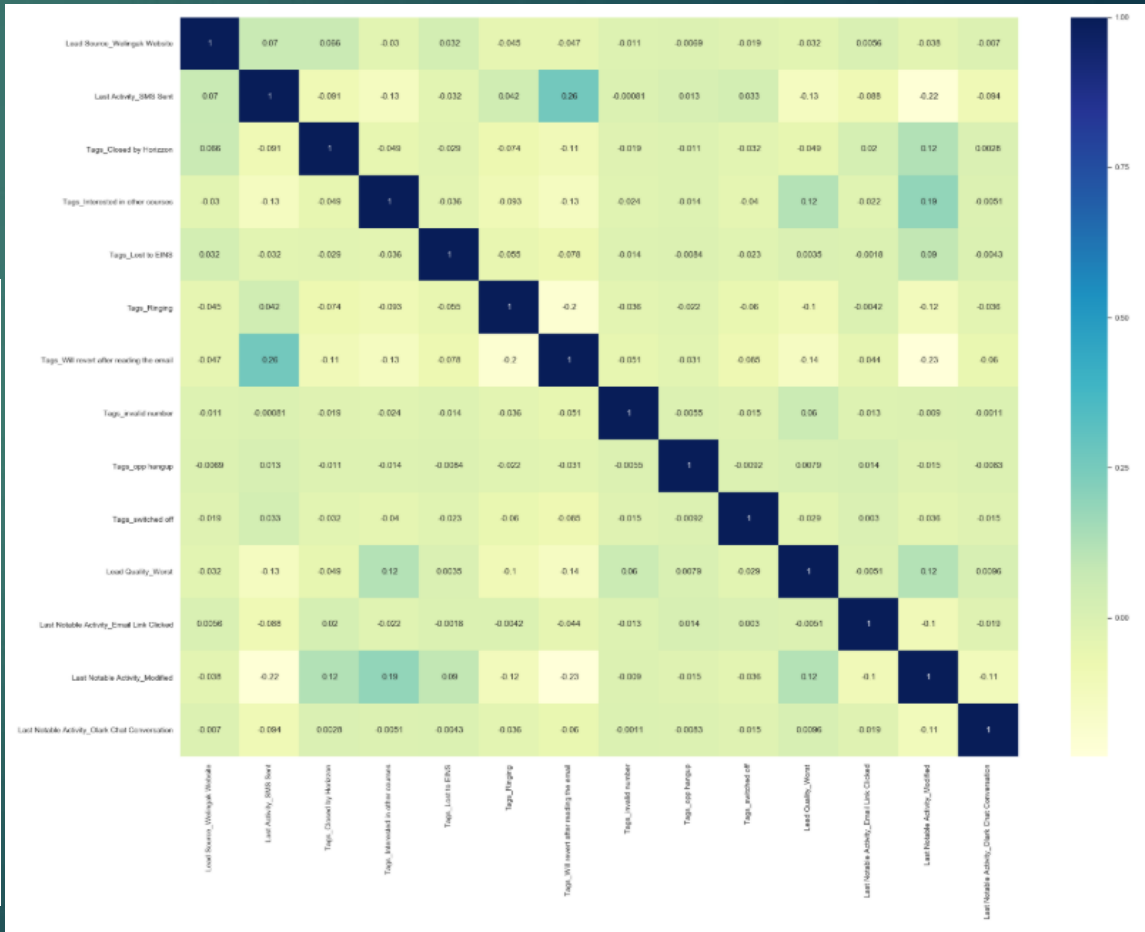
	Features	VIF
3	Tags_Closed by Horizzon	1.29
1	Lead Source_Welingak Website	1.26
7	Tags_Not doing further education	1.10
6	Tags_Lost to EINS	1.04
4	Tags_Diploma holder (Not Eligible)	1.03
13	Tags_switched off	1.02
18	Last Notable Activity_Olark Chat Conversation	1.01
16	Last Notable Activity_Email Link Clicked	1.01
14	Tags_wrong number given	1.01
10	Tags_invalid number	1.01
12	Tags_opp hangup	1.00
11	Tags_number not provided	1.00
0	Lead Origin_Lead Add Form	0.77
15	Lead Quality_Worst	0.42
8	Tags_Ringing	0.34
5	Tags_Interested in other courses	0.30
17	Last Notable Activity_Modified	0.14
9	Tags_Will revert after reading the email	0.06
2	Last Activity_SMS Sent	0.01

Feature Elimination using RFE

- After removing 6 features the final dataset is having 14 features

	P>|z|
const	0.000
Lead Source_Welingak Website	0.000
Last Activity_SMS Sent	0.000
Tags_Closed by Horizzon	0.000
Tags_Interested in other courses	0.000
Tags_Lost to EINS	0.000
Tags_Ringing	0.000
Tags_Will revert after reading the email	0.000
Tags_invalid number	0.003
Tags_opp hangup	0.016
Tags_switched off	0.000
Lead Quality_Worst	0.000
Last Notable Activity_Email Link Clicked	0.007
Last Notable Activity_Modified	0.000
Last Notable Activity_Olark Chat Conversation	0.001

	Features	VIF
2	Tags_Closed by Horizzon	1.07
4	Tags_Lost to EINS	1.04
0	Lead Source_Welingak Website	1.03
9	Tags_switched off	1.02
7	Tags_invalid number	1.01
11	Last Notable Activity_Email Link Clicked	1.01
13	Last Notable Activity_Olark Chat Conversation	1.01
8	Tags_opp hangup	1.00
10	Lead Quality_Worst	0.39
5	Tags_Ringing	0.34
3	Tags_Interested in other courses	0.30
12	Last Notable Activity_Modified	0.13
6	Tags_Will revert after reading the email	0.06
1	Last Activity_SMS Sent	0.01



Conversion Probability and Predicted Column

- Created data frame with converted information conversion probability
- Created a new column in the dataset as 1 if the probability >0.5 else 0.

	Converted	Conversion_Prob	LeadID
0	1	0.97	1490
1	1	0.68	4901
2	1	0.97	1804
3	1	1.00	3411
4	0	0.19	642

	Converted	Conversion_Prob	LeadID	predicted
0	1	0.97	1490	1
1	1	0.68	4901	1
2	1	0.97	1804	1
3	1	1.00	3411	1
4	0	0.19	642	0

Conversion Probability and Predicted Column

- Created data frame with converted information conversion probability
- Created a new column in the dataset as 1 if the probability >0.5 else 0.

	Converted	Conversion_Prob	LeadID
0	1	0.97	1490
1	1	0.68	4901
2	1	0.97	1804
3	1	1.00	3411
4	0	0.19	642

	Converted	Conversion_Prob	LeadID	predicted
0	1	0.97	1490	1
1	1	0.68	4901	1
2	1	0.97	1804	1
3	1	1.00	3411	1
4	0	0.19	642	0

KPIs for Model Evaluation

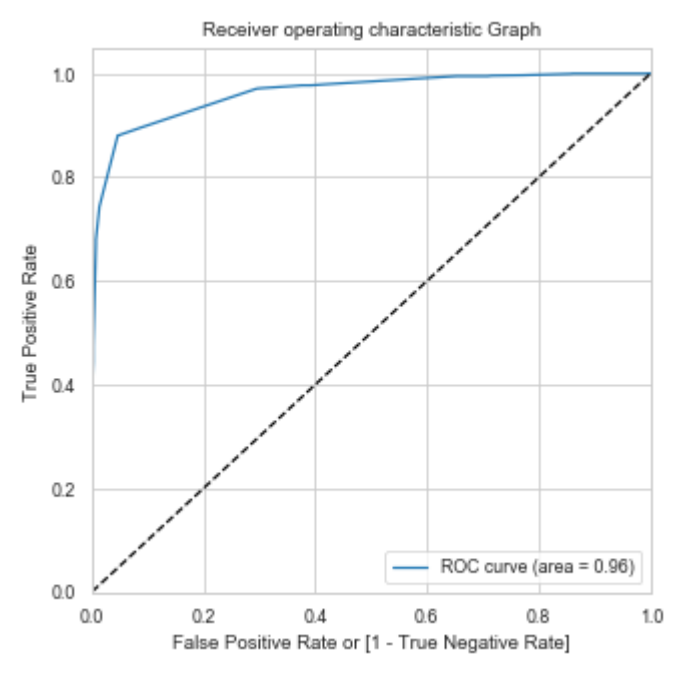
Confusion Matrix

actual	predicted	
	not_converted	converted
not_converted	3547 (TN)	170 (FP)
converted	280 (FN)	2016 (TP)

KPIs	value
Accuracy **(TP+TN)/(TN+FP+FN+TP)**	0.9251
Sensitivity **TP / (TP+FN)**	0.8780
Specificity **TN / (TN+FP)**	0.9542
Precision **TP /(TP + FP)**	0.8386
Recall **TP /(TP + FN)**	0.8780

KPIs for Model Evaluation

Receiver Operating Characteristic Curve (ROC Curve)



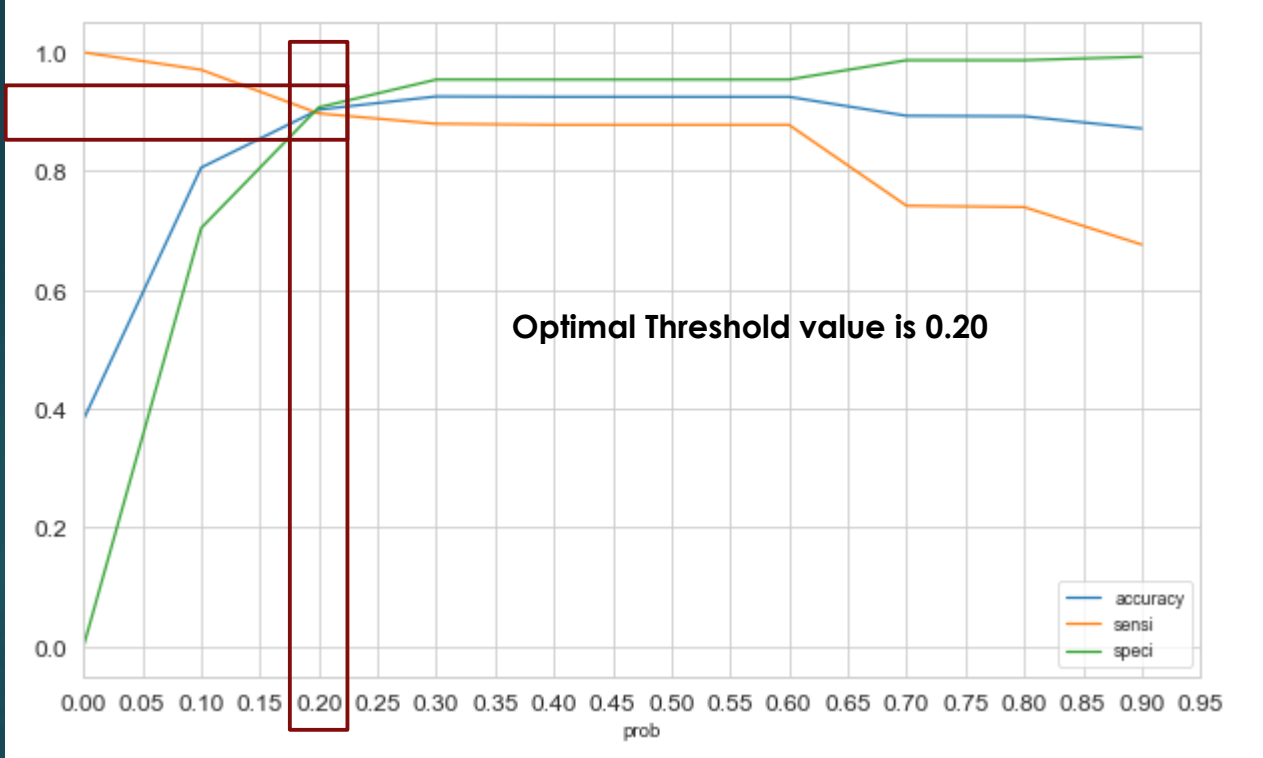
KPIs	value
False Positive Rate **FP/ (TN+FP)**	0.0457
Area Under the Curve**	0.9627

** True Positive Rate value can also found from the formula of **sensitivity***

***From area under the curve (AUC) of a ROC curve, one can determine how good the model is. The larger the AUC, the better will be the model.*

KPIs for Model Evaluation

Finding the Optimal Threshold



- It is required to balance the sensitivity and specificity and hence required a threshold point.
- Hence, we ran accuracy, sensitivity and specificity for various probability cut-off value to determine the same
- From the left side graph it can found that the threshold value is 0.20
- Using the threshold value, we can find that the model accuracy is 0.9035 *

*Code mentioned below:

```
metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.final_predicted)
```

KPIs for Model Evaluation

confusion matrix after using the cut-off 0.20

Confusion Matrix

actual	Predicted	
	not_converted	converted
not_converted	3373 (TN)	344 (FP)
converted	236 (FN)	2060 (TP)

KPIs	value
Accuracy **(TP+TN)/(TN+FP+FN+TP)**	0.9035
Sensitivity **TP / (TP+FN)**	0.8972
Specificity **TN / (TN+FP)**	0.9074

```
confusion1 = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.final_predicted)
confusion1
```

Use the model on test dataset

Use probability threshold value of 0.20 on the test dataset to predict if a lead will convert or not

	LeadID	Converted	Conversion_Prob	final_predicted
0	2695	1	0.97	1
1	7431	0	0.19	0
2	6242	1	1.00	1
3	2871	0	0.03	0
4	7560	0	0.00	0

Use the model on test dataset

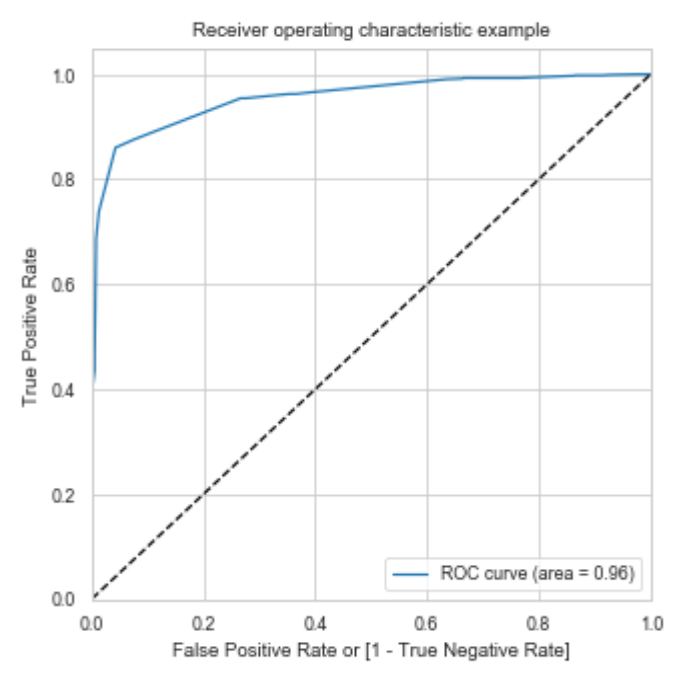
Confusion Matrix

actual	predicted	
	not_converted	converted
not_converted	1470 (TN)	122 (FP)
converted	122 (FN)	864 (TP)

KPIs	value
Accuracy **(TP+TN)/(TN+FP+FN+TP)**	0.9053
Sensitivity **TP / (TP+FN)**	0.8762
Specificity **TN / (TN+FP)**	0.9233
Precision **TP /(TP + FP)**	0.8762
Recall **TP /(TP + FN)**	0.8762

Use the model on test dataset

Receiver Operating Characteristic Curve (ROC Curve)



KPIs	value
False Positive Rate **FP/ (TN+FP)**	0.0766
Area Under the Curve**	0.9555

** True Positive Rate value can also found from the formula of **sensitivity***

***From area under the curve (AUC) of a ROC curve, one can determine how good the model is. The larger the AUC, the better will be the model.*

Lead Score Calculation:

Lead Score Formula: 100*Conversion Probability

- Since, we divided the actual dataset into train and test at the beginning, we append them again to get the entire list of leads
- Conversion probability is multiplied by 100 to get the score
- Higher lead score denotes that the lead is more likely to convert

	Lead Number	Converted	Conversion_Prob	final_predicted	Lead_Score
0	660737	0.00	0.01	0.00	1.00
1	660728	0.00	0.01	0.00	1.00
2	660727	1.00	0.97	1.00	97.00
3	660719	0.00	0.00	0.00	0.00
4	660681	1.00	0.84	1.00	84.00
5	660680	0.00	0.03	0.00	3.00
6	660673	1.00	0.84	1.00	84.00
7	660664	0.00	0.03	0.00	3.00
8	660624	0.00	0.19	0.00	19.00
9	660616	0.00	0.19	0.00	19.00

*Lead score with >=20 will have a final prediction of 1 as we consider the threshold value of 0.20

Thank You