SUBMITTED BY:

ASHOK GORANTALA

DEVENDER GAKKULA

LEAD SCORING CASE STUDY SUBMISSION

Submitted for Upgrad Data Science Program

Dated: 18th Nov 2019

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%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel:

Lead Conversion Process - Demonstrated as a funnel

PROBLEM STATEMENT,..CONTD

As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you 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%.

Data

You have been provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted. You can learn more about the dataset from the data dictionary provided in the zip folder at the end of the page. Another thing that you also need to check out for are the levels present in the categorical variables. Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value (think why?).

Goals of the Case Study

There are quite a few goals for this case study.

1. 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.
2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

LOGISTIC REGRESSION – MODEL BUILD – ANALYSIS

Analysing and Understanding raw data, we are having very high NULL values in following columns (especially)

	MIS_PERC
Lead Quality	51.590909
Asymmetrique Activity Index	45.649351
Asymmetrique Profile Score	45.649351
Asymmetrique Activity Score	45.649351
Asymmetrique Profile Index	45.649351
Tags	36.287879
Lead Profile	29.318182
What matters most to you in choosing a course	29.318182
What is your current occupation	29.112554
Country	26.634199
How did you hear about X Education	23.885281
Specialization	15.562771
City	15.367965
Page Views Per Visit	1.482684
TotalVisits	1.482684
Last Activity	1.114719
Lead Source	0.389610

Hence, we pragmatic towards cleaning the NULL entries.

LOGISTIC REGRESSION – MODEL BUILD – ANALYSIS

NULL VALUE HANDLING:

1. We have managed to exclude the dimension having NULL entries more than 40% of data entries.
2. Following this percentage of values we have few other dimension like, TAGS, WHAT MATTERS MOST TO YOU IN CHOOSING A COURSE as a dimensions having very high NULL entries. Even though this kind of dimensions are categorical but these are adding good meaning to the data. So, we intended to mute NULL values by adding “NO_VALUE”; such that label_encoder will assign a tag to this and this will be part of our Logistic Regression.
3. Then we have very less fractionated NULL entries, which we easily handled by dropping them.
4. With all these operations, all NULL values are eliminated, leaving a data to build our logistic regression model.

LOGISTIC REGRESSION – MODEL BUILD – ANALYSIS

OUTLIER ANALYSIS:

- With a clear contemplation on the data; we have quantile distribution on numerical columns as as:

	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820
std	23405.995698	0.486714	4.854853	548.021466	2.161418
min	579533.000000	0.000000	0.000000	0.000000	0.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000
90%	650506.100000	1.000000	7.000000	1380.000000	5.000000
95%	655404.050000	1.000000	10.000000	1562.000000	6.000000
99%	659592.980000	1.000000	17.000000	1840.610000	9.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000

Looking at above values, TotalVisits and 'Per Views Per visit' clearly having outliers. Removing values above 99%

LOGISTIC REGRESSION – MODEL BUILD – ANALYSIS

LABEL CONVERSION

- Excluding already numerical columns, we have initiated and fit the Label Encoder, leaving us the data converting all categorically into numerical value for each categorical value.

LOGISTIC REGRESSION – MODEL BUILD – ANALYSIS

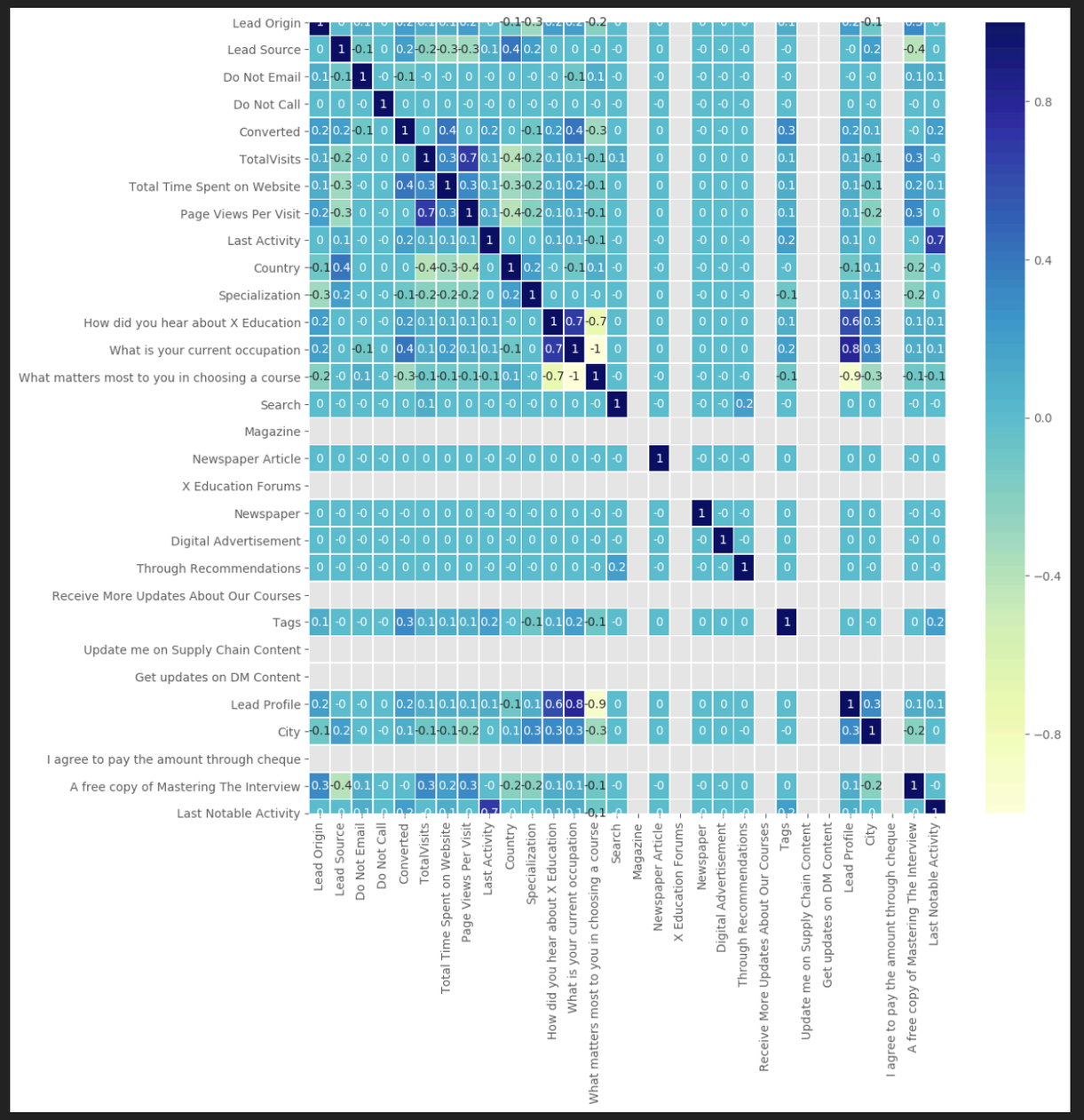
CHECKING DATA CORRELATIONS

Post performing LabelEncoding, we are managed to check the correlation between data dimensions and resulted into.

Heatmap, clearly indicating few of the dimensions are not contributing any data variance.
When checked for the same, we have value_counts for the features as:

```
--------------------------
Magazine
0    8991
Name: Magazine, dtype: int64
--------------------------
X Education Forums
0    8991
Name: X Education Forums, dtype: int64
--------------------------
Receive More Updates About Our Courses
0    8991
Name: Receive More Updates About Our Courses, dtype: int64
--------------------------
Update me on Supply Chain Content
0    8991
Name: Update me on Supply Chain Content, dtype: int64
--------------------------
Get updates on DM Content
0    8991
Name: Get updates on DM Content, dtype: int64
--------------------------
I agree to pay the amount through cheque
0    8991
Name: I agree to pay the amount through cheque, dtype: int64
```

With this reason, removed the columns



LOGISTIC REGRESSION – MODEL BUILD – ANALYSIS

TRAIN-TEST SPLIT

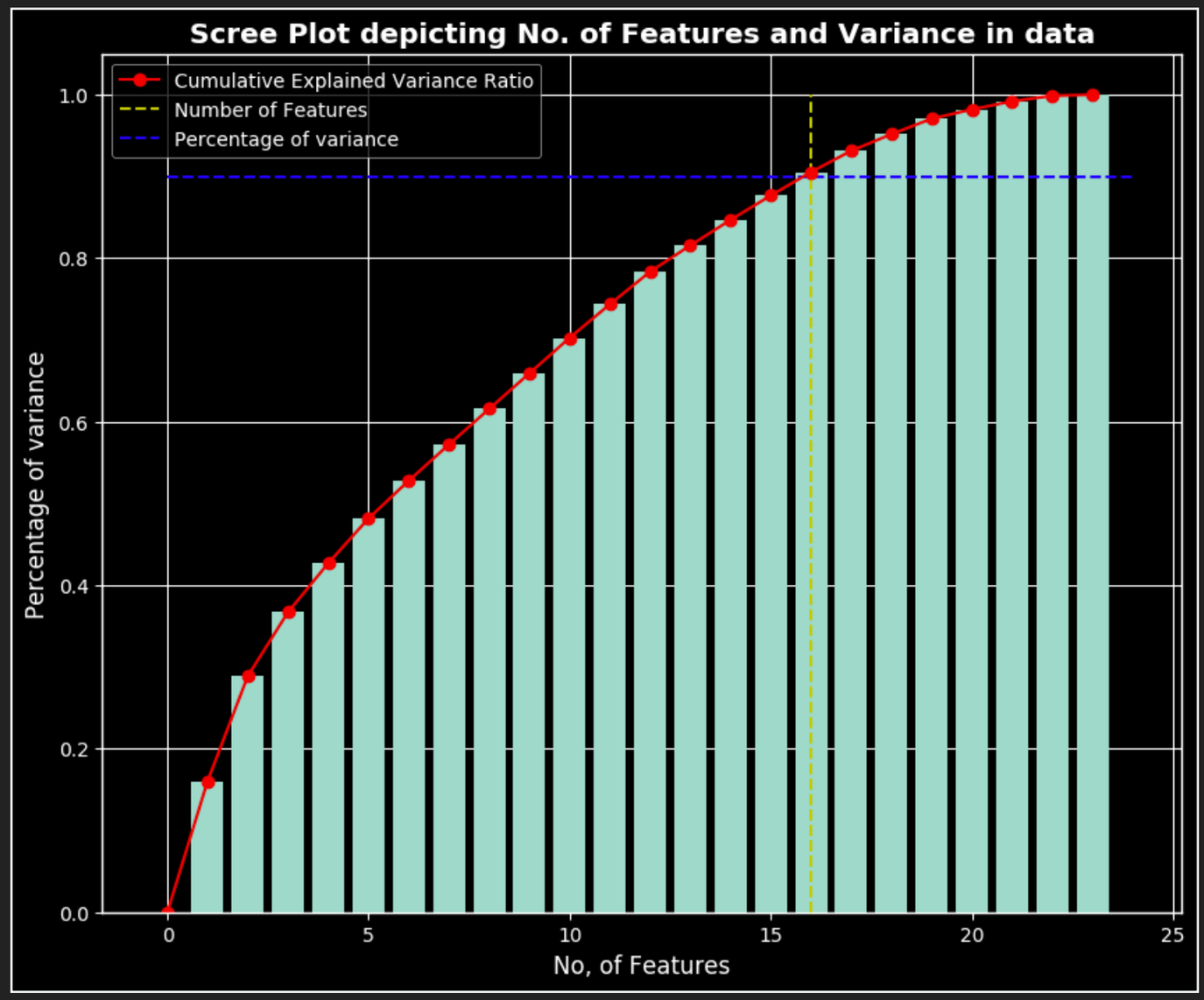
- In order to built the logistic regression, we segregated the data into train and test modules.

FEATURE STANDARDISATION

- Using the standard-scaler from sclera we have applied scaler on Train data and performed scaling conversion.

LOGISTIC REGRESSION WITH PCA

USING PRINCIPAL COMPONENT ANALYSIS – FINDING BEST DIMENSIONS EXPLAINING MAXIMUM VARIANCE IN DATA



From this Scree Plot:

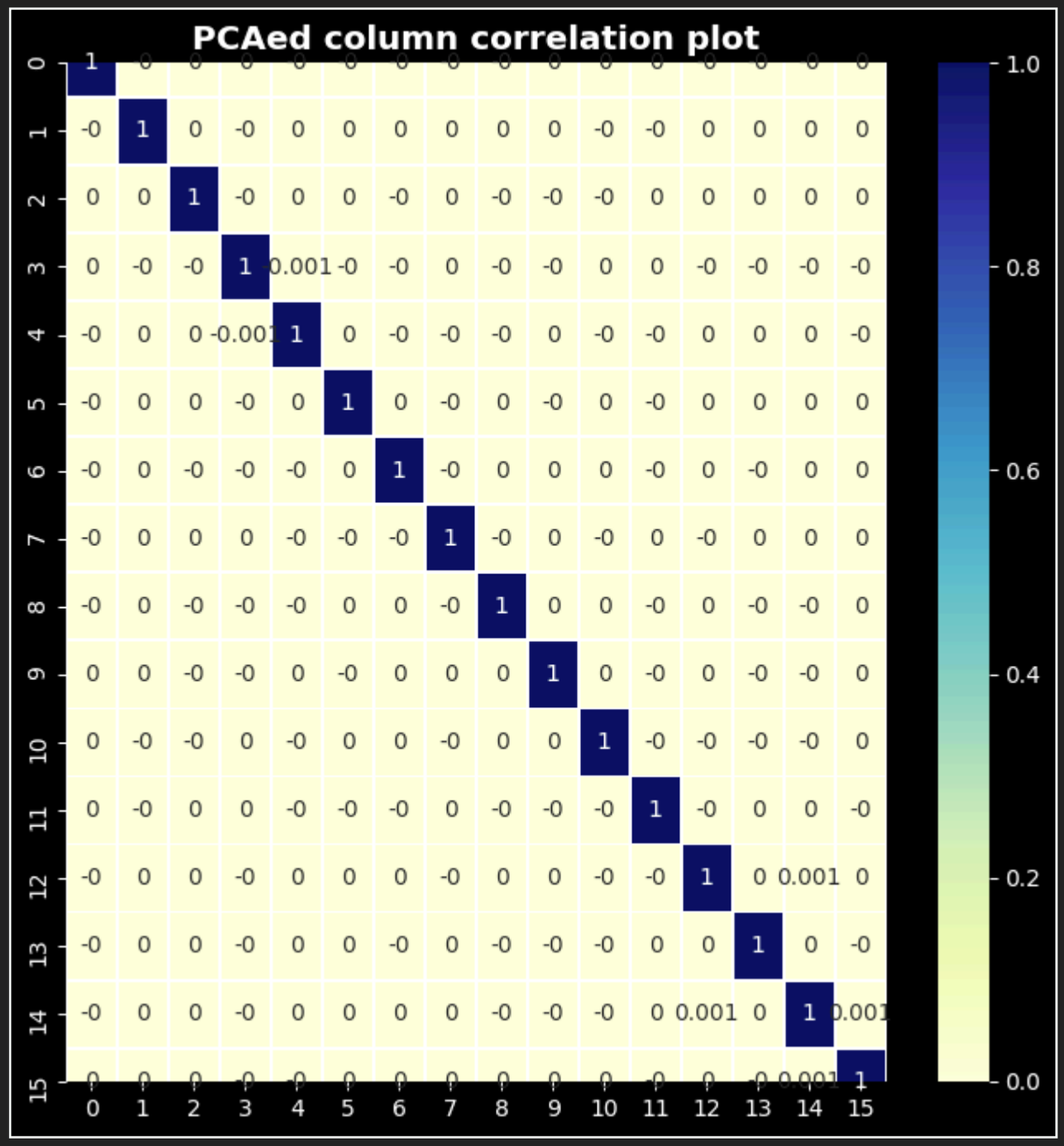
1. Features are contributing equal amount of variance in data
2. That is why, we are having gradual increasing curve
3. Considering, 16 features are showing 90% of data variance.
4. Finally considering 16 features for analysis

LOGISTIC REGRESSION WITH PCA

PCA WITH DECIDED 16 DIMENSIONS

With application of PCA model and converted to new basis, we have heat map as

This heatplot, clearly identifies that new basis of features are clearly having no inter-dependency on other features.

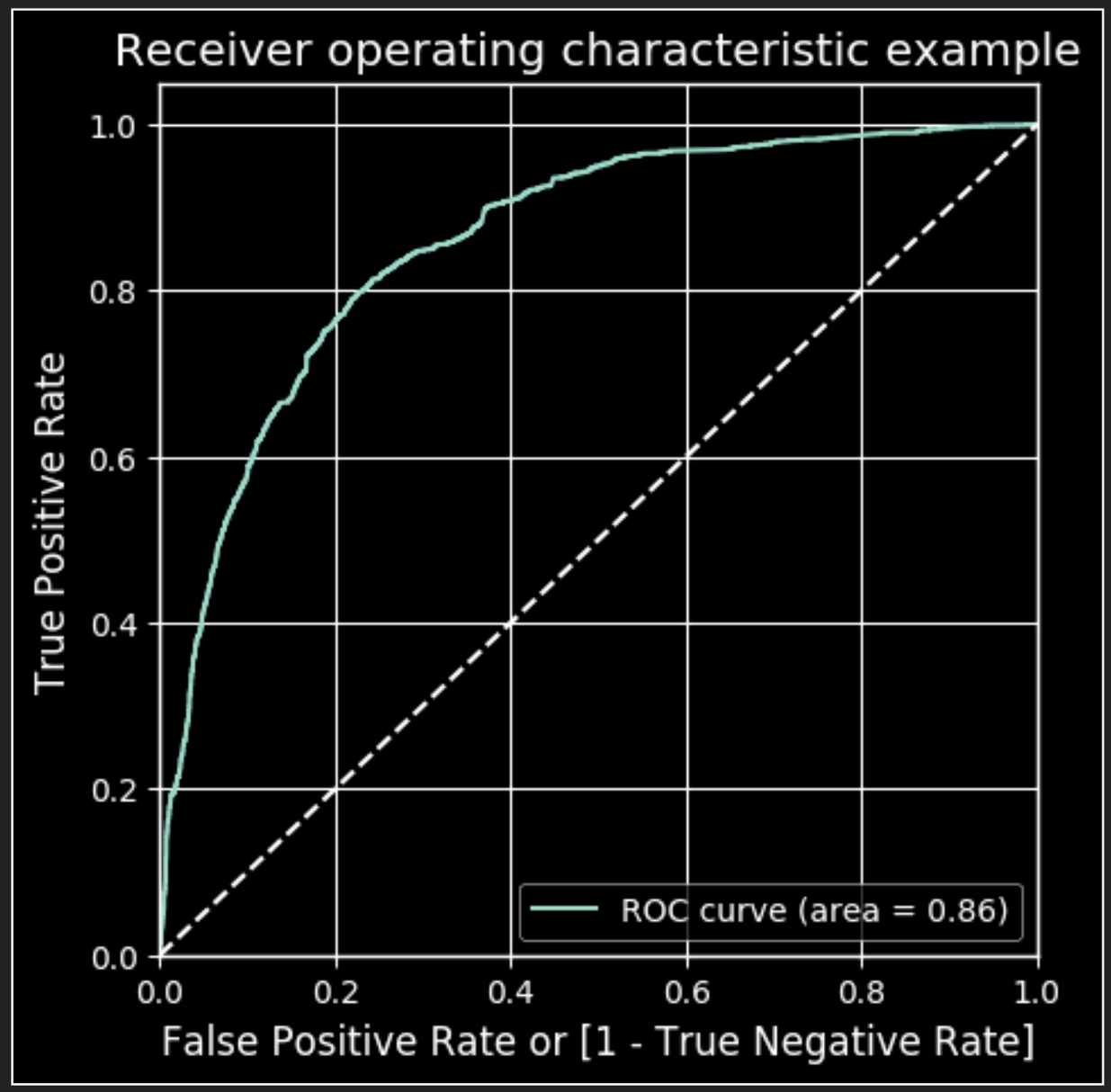


LOGISTIC REGRESSION WITH PCA

APPLYING LOGISTIC REGRESSION OVER TRAIN DATA

With successful application of logistic regression over PCAed data and finding probabilities of conversion, we have plotted ROC curve against Converted values, and we have.

The area under the curve, signifies the ratio between True Positive Rate to the False Positive Rate. A value of 86% signifies that we have more True Positive and our model built to be giving good results.



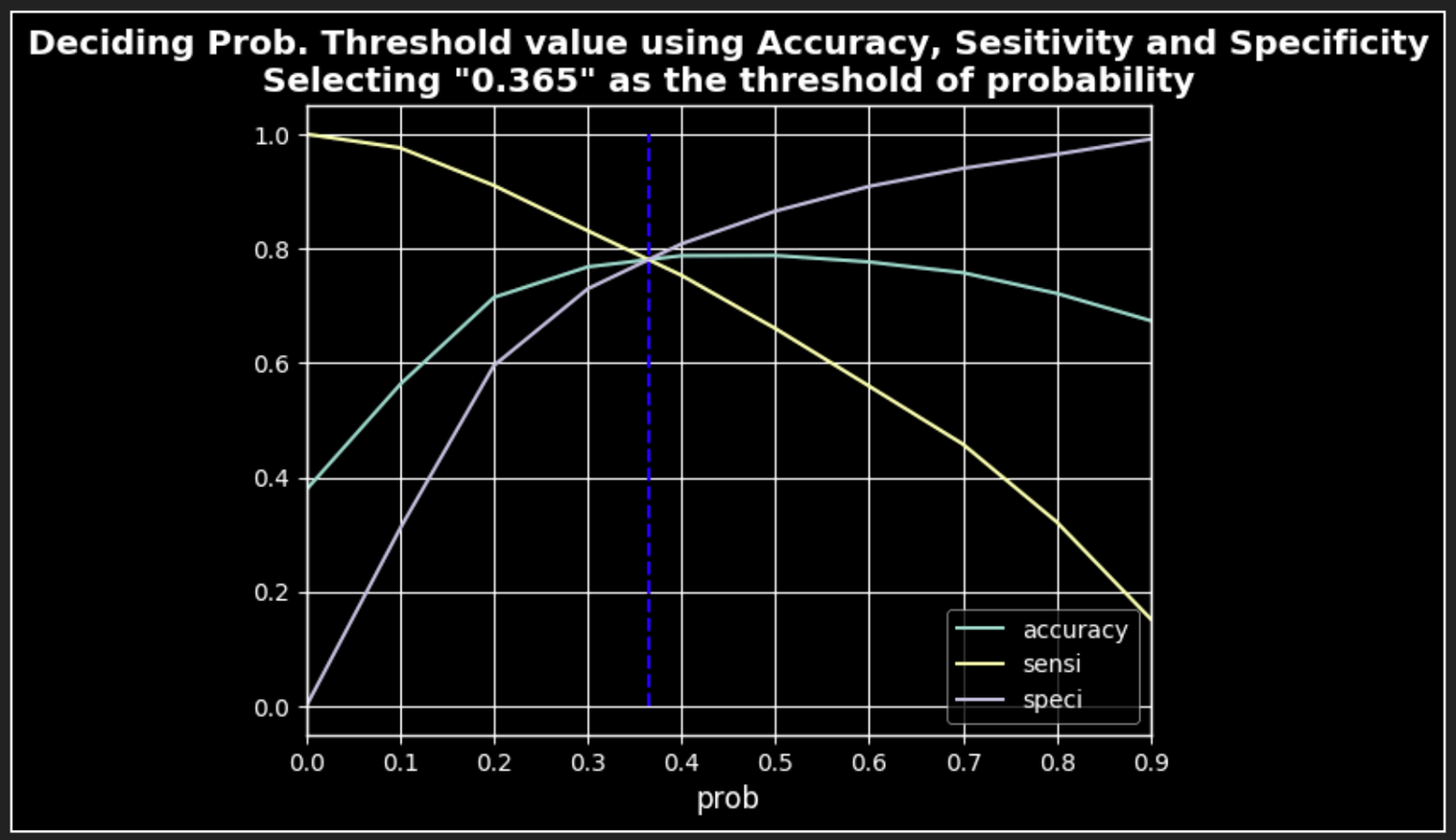
LOGISTIC REGRESSION WITH PCA

FINDING OPTIMAL THRESHOLD PROBABILITY

For different probability values, computed Accuracy, Sensitivity and Specificity. Results computed to be:

	prob	accuracy	sensi	speci
0.0	0.0	0.378318	1.000000	0.000000
0.1	0.1	0.562361	0.976088	0.310592
0.2	0.2	0.714519	0.910231	0.595420
0.3	0.3	0.767907	0.831047	0.729485
0.4	0.4	0.787335	0.753038	0.808206
0.5	0.5	0.787780	0.660133	0.865458
0.6	0.6	0.776509	0.559388	0.908635
0.7	0.7	0.757675	0.457860	0.940124
0.8	0.8	0.721489	0.321835	0.964695
0.9	0.9	0.673884	0.152489	0.991174

When plotted over graph:



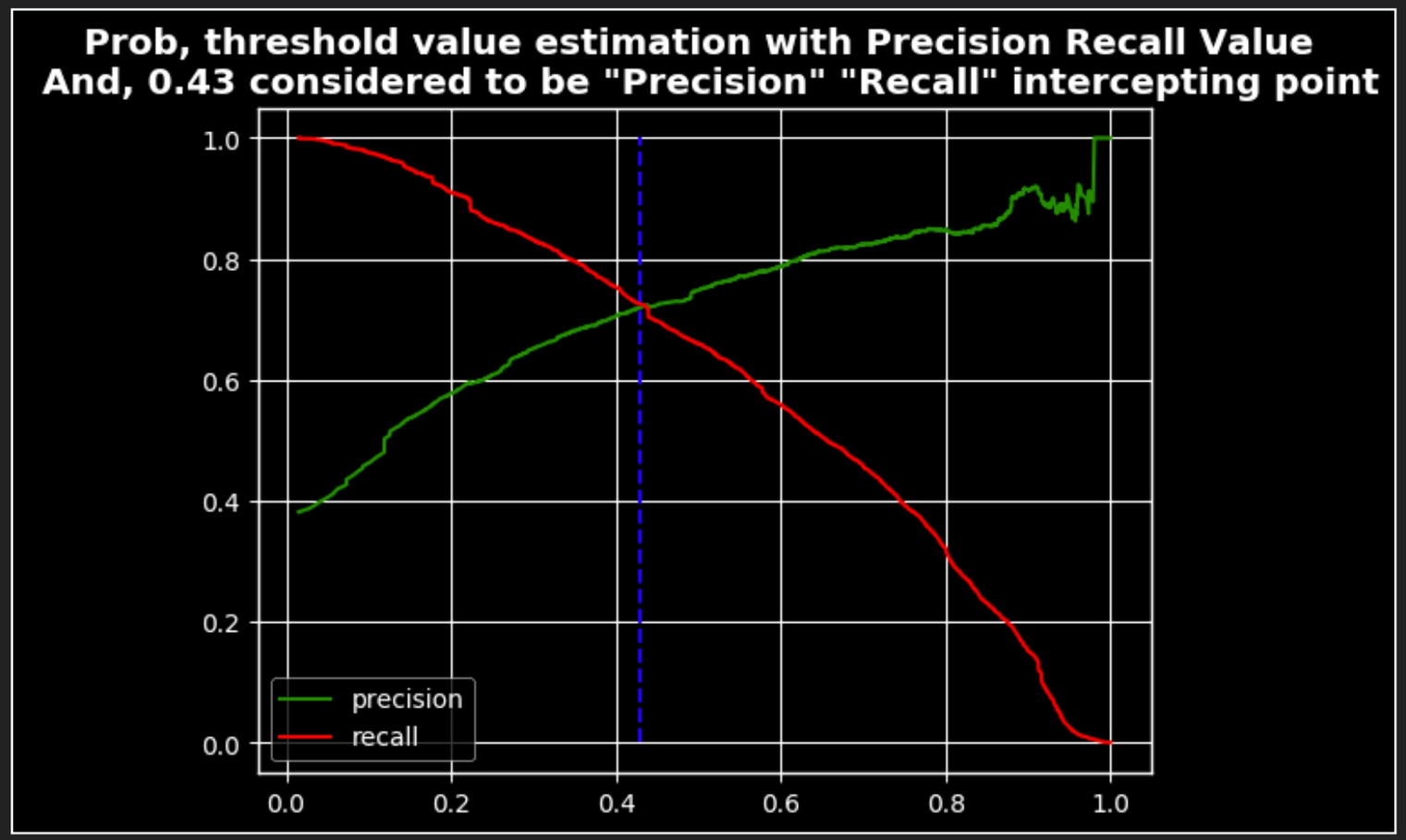
The intersecting point considered to be optimal point for threshold value, which is: 0.365

LOGISTIC REGRESSION WITH PCA

PRECISION AND RECALL SCORE

Even though, we have “Optimal” value of probability of threshold, considering PRECISION and RECALL SCORE for further verification of probability of threshold.

The intersecting point considered to be optimal point for threshold value, which is: 0.43



Decision on Probability Threshold confirmation:

1. Using Accuracy, Sensitivity and Specificity method, we are having 0.365 as probability threshold
2. Whereas using, Precision & Recall method, we have 0.43 as probability threshold for Conversion analysis.

LOGISTIC REGRESSION WITH PCA

USING OPTIMAL THRESHOLD VALUE OF 0.365...

Accuracy Score on Test Data: 0.788
Precision Score On Test Data: 0.69
Recall Score on Test Data: 0.791

In Business Terms:
Accuracy defined to be, out of all labels correctly predicted Converted labels.
Precision defined to be, out of all Positive Converted assumed labels, how many Converted labels are correct.
Similarly, **Recall** score defined to be, of all converted labels (including Falsely identified), how many correctly identified.

With all above definitions, the model we built are yielding good results to be perceived as we are ready to deploy the model.

CONCLUSION

USING OPTIMAL THRESHOLD VALUE OF 0.365...

From the basic user data, we have around 30% of Conversion rate, and we found that probability threshold value of 0.365 predicting almost same amount of Conversion rate.
If we are supposed to improve the Conversion rate, we are supposed to increase the Probability threshold value much less than 0.365 and build the conversion labels.

That being said, if we increase probability threshold value to **0.56**, we are getting Precision and Recall values as following.

```
[485] # Creating new column 'predicted' with 1 if Churn_Prob > 0.365 else 0
      y_train_pred_final['predicted_2'] = y_train_pred_final.Converted_prob.map(lambda x: 1 if x >= 0.56 else 0)
      y_train_pred_final.head()
```

	Converted	Converted_prob	Lead Number	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	predicted	predicted_2
89	0	0.547449	659630	1	1	1	1	1	1	0	0	0	0	1	0
377	1	0.782794	656469	1	1	1	1	1	1	1	1	0	0	1	1
949	1	0.162144	640752	1	1	0	0	0	0	0	0	0	0	0	0
555	0	0.185440	591682	1	1	0	0	0	0	0	0	0	0	0	0
488	0	0.015293	592207	1	0	0	0	0	0	0	0	0	0	0	0

```
# Precision Score
precision_score(y_train_pred_final.Converted, y_train_pred_final.predicted_2)
```

```
0.7943085371942087
```

```
[487] # Recall Score
      recall_score(y_train_pred_final.Converted, y_train_pred_final.predicted_2)
```

```
0.6236769894159153
```

That means, the data points predicted to be above 0.56, having very high chances of getting converted into the program. These people should be considered to HIGH LEAD people.