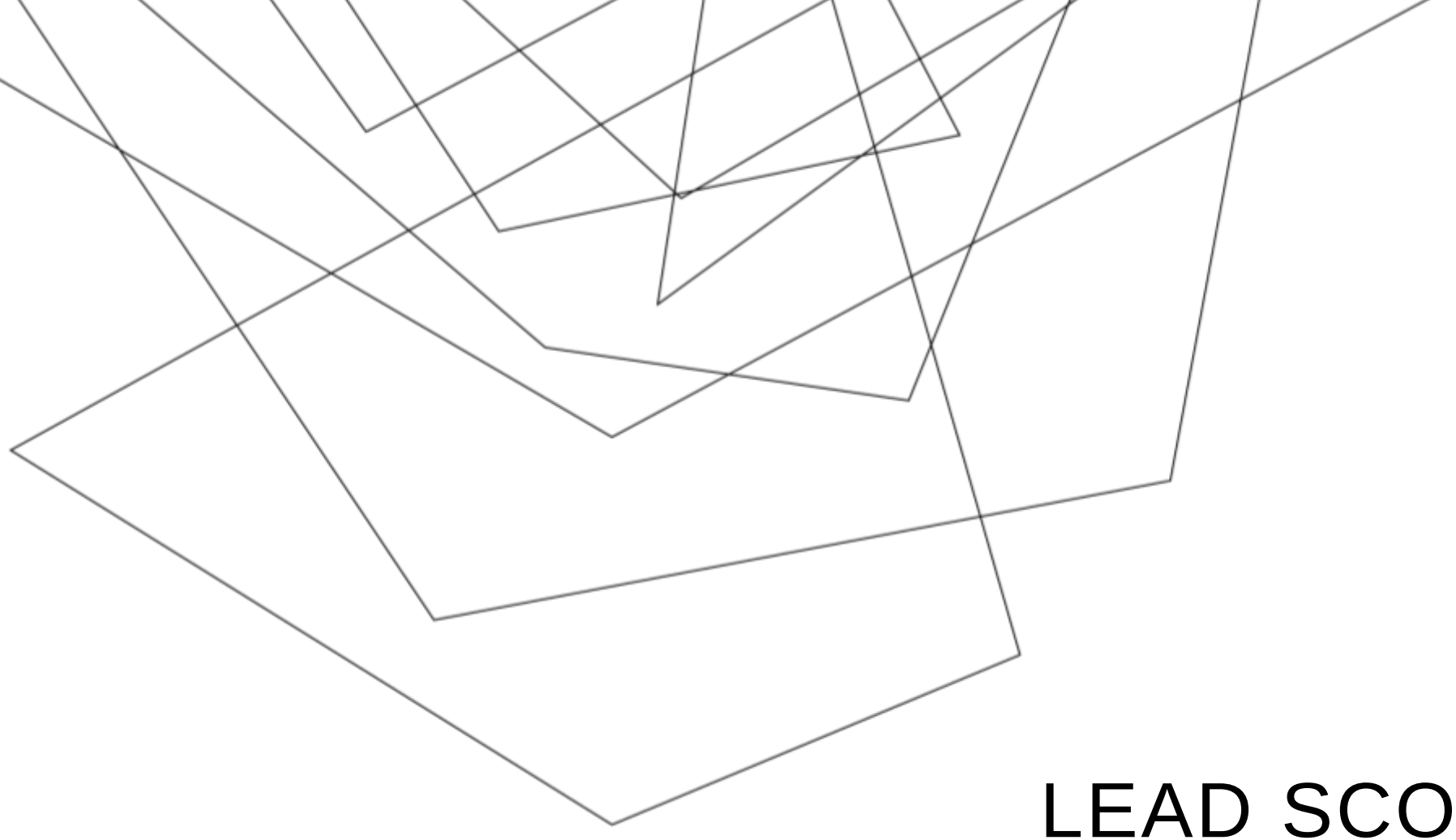


LEAD SCORING CASE STUDY

Venkatasubramanian Sundaramahadevan

AGENDA

Introduction

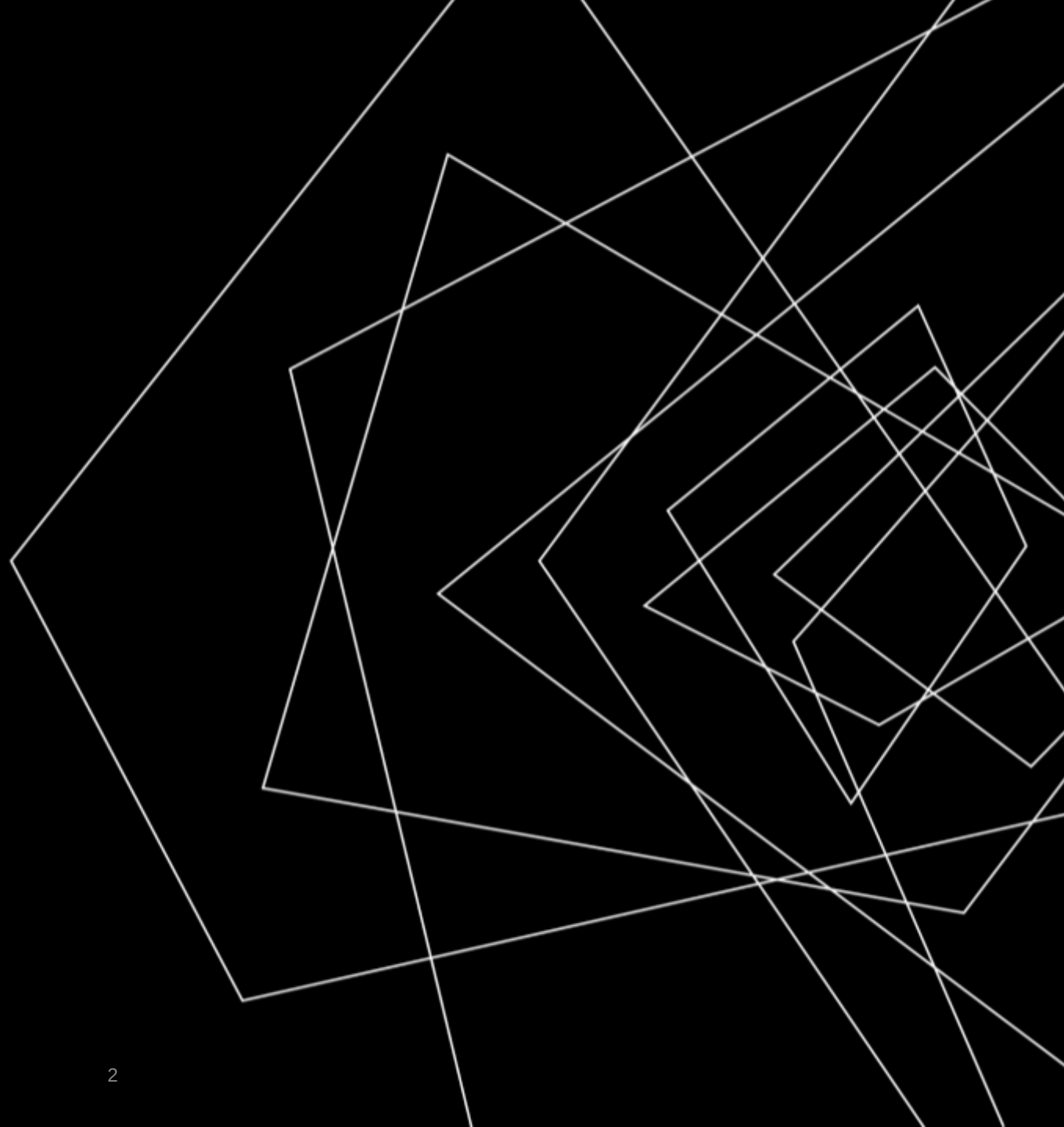
Problem Statement

Analysis Approach

Observations & Results

Recommendations

Summary



INTRODUCTION

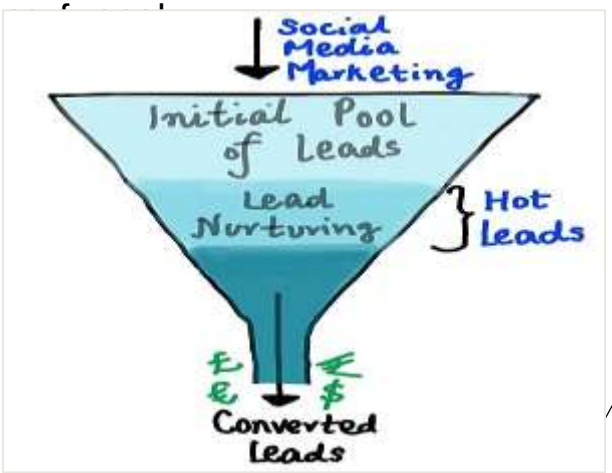
In this Case Study, we will be focusing on improving the Lead Conversion Rate of an Education company “X Education”. We will analyze the data provided and find the important features that are of significance for Lead Conversion and the process will be discussed in the upcoming slides

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 followi



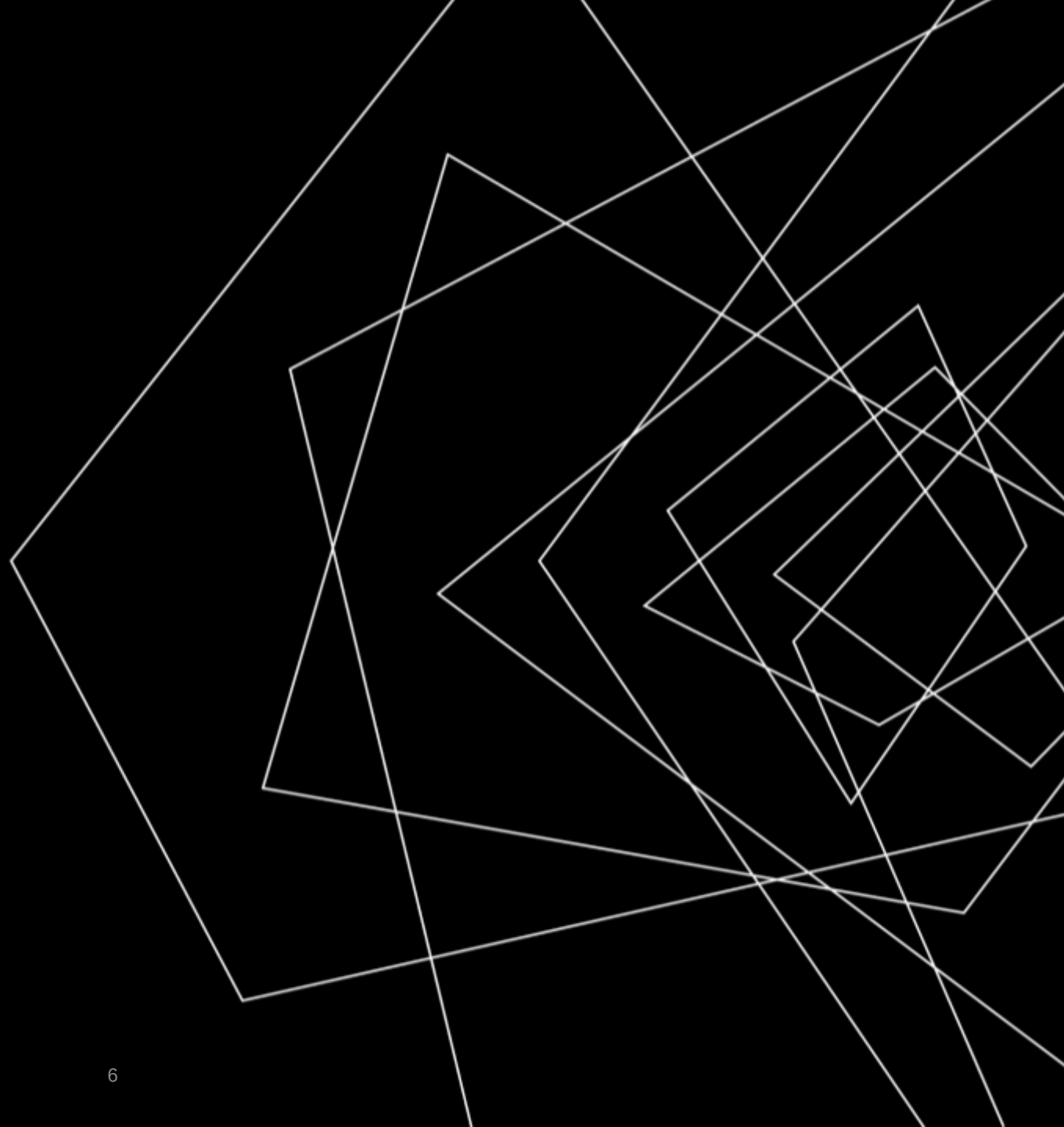
BUSINESS OBJECTIVE

X Education has appointed us to help them 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 we 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 has given a ballpark of the target lead conversion rate to be around 80%.

ANALYSIS APPROACH

THE FOLLOWING FOUR STEPS ARE THE INTEGRAL PART OF END-TO-END ANALYSIS OF THE CASE STUDY AT HAND

1. Data Cleaning
2. Data Preparation
3. Modeling
4. Evaluation



DATA CLEANING

In this step, we will focus on:

- Importing the data
- Handling NULL values
  - Remove the fields that have NULL values above 32%
- Removing unnecessary columns
  - Remove columns that don’t provide any business value such as Country/City/Lead Number etc.,
  - Remove columns that have very less variance such as ‘Do Not Call’
  - Remove columns with more than 50% ‘select’ values (this indicates that user didn’t select any value)
- Handling NULL values < 32%
  - Remove the rows for all columns with < 32% NULL values

DATA PREPARATION

In this step, we will focus on:

- Converting below binary categorical variables to 1 & 0 values
  - 'Do Not Email' & 'A free copy of Mastering The Interview'
- For multi-level categorical variables, we will perform One-Hot Encoding
  - 'Lead Origin', 'Lead Source', 'Last Activity', 'What is your current occupation', 'Last Notable Activity'
  - Drop the first column for each dummified variable as we need only n-1 dummies to represent unique n values
  - Drop all the categorical variables for which we performed dummification

MODELING

In this step, we will focus on:

- Split data into train and test with ratio 0.7 to 0.3
- Do Feature Scaling for numerical variables with MinMaxScaler
- Look at Correlations using Heatmap and make a note of correlation above 60% as they indicate multicollinearity
- Create 1st model with all features
- As we have 74 features, we will use RFE (Recursive Feature Elimination) to narrow down to 15 important features. Using this create 2nd model
- Check VIFs (Variance Inflation factor) for all features. This helps us to identify features which have highly collinear (i.e., VIF > 5)
- Check p-value for all features (p-value should be < 0.05)
- From hereon, we repeatedly create models (after removing high VIFs and p-value variables one by one)
- The model was finalized at 6th iteration where we all p-values < 0.05 & VIFs < 5

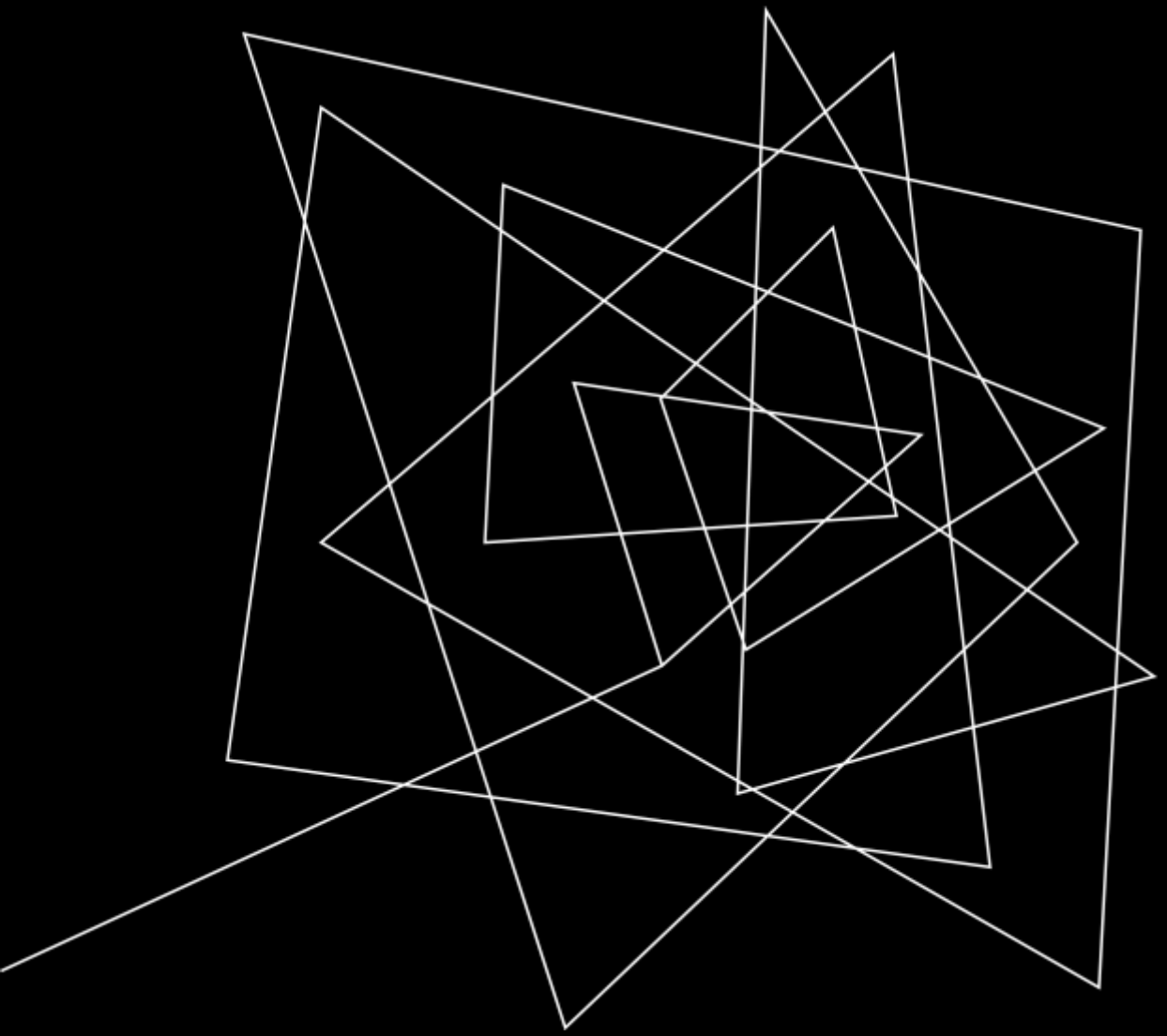
EVALUATION

In this step, we will be focus on:

- Predict target probability values based on X_train (feature train data set)
- Build base predictions based on 0.5 probability cut-off point and calculate metrics such as accuracy, sensitivity, specificity, False Positive rate, Positive Predictive rate, & Negative Predictive rate
- Plot ROC (Receiver Operating Characteristic) curve to determine trade-off b/w Sensitivity and Specificity. We get AUC (Area under the curve) as 0.86 which indicates a good model
- We will then determine the optimal probability cut-off point based on various probabilities in the range 0, 0.1, 0.2,…,0.9 by plotting the accuracy, sensitivity and specificity curve. The intersection point determines the optimal cut-off. In our case it's 0.43
- We create precision and recall curve as well to validate the optimal cut-off. We observe 0.43 in this case as well
- Predict target probability values based on X_test (feature test data set)
- Calculate accuracy, sensitivity and specificity for both train and test data based on 0.43 probability cut-off. They are as follows:

Train/Test	Accuracy	Sensitivity	Specificity
Train	0.7895	0.7854	0.7932
Test	0.7850	0.7751	0.7941

- Calculate Lead Score based on the Conversion Probabilities on the test data set by multiplying the probability values by 100



OBSERVATIONS & RESULTS

1. Correlations
2. Final Model Summary with important features
3. ROC Curve
4. Optimal cut-off (Sensitivity vs Specificity vs Accuracy)
5. Precision vs Recall trade-off
6. Model Performance
7. Lead Score

HIGHLY CORRELATED FEATURES

Below the features that have a correlation of over 60%

Lead Origin_Lead Import	Lead Source_Facebook	0.981903
Lead Source_Facebook	Lead Origin_Lead Import	0.981903
Last Activity_SMS Sent	Last Notable Activity_SMS Sent	0.890591
Last Notable Activity_SMS Sent	Last Activity_SMS Sent	0.890591
Last Activity_Unsubscribed	Last Notable Activity_Unsubscribed	0.879716
Last Notable Activity_Unsubscribed	Last Activity_Unsubscribed	0.879716
Last Activity_Email Opened	Last Notable Activity_Email Opened	0.866192
Last Notable Activity_Email Opened	Last Activity_Email Opened	0.866192
Lead Source_Reference	Lead Origin_Lead Add Form	0.862980
Lead Origin_Lead Add Form	Lead Source_Reference	0.862980
Last Activity_Email Link Clicked	Last Notable Activity_Email Link Clicked	0.781836
Last Notable Activity_Email Link Clicked	Last Activity_Email Link Clicked	0.781836
Last Activity_Had a Phone Conversation	Last Notable Activity_Had a Phone Conversation	0.751218
Last Notable Activity_Had a Phone Conversation	Last Activity_Had a Phone Conversation	0.751218
Last Activity_Email Received	Last Notable Activity_Email Received	0.707051
Last Notable Activity_Email Received	Last Activity_Email Received	0.707051
Last Activity_Page Visited on Website	Last Notable Activity_Page Visited on Website	0.693902
Last Notable Activity_Page Visited on Website	Last Activity_Page Visited on Website	0.693902

FINAL MODEL SUMMARY WITH IMPORANT FEATURES

We have 11 important features highlighted in the final model along with their corresponding coefficients (Higher values indicate more impact on the model outcome), p-value and VIFs

Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4449
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2079.1
Date:	Tue, 14 Sep 2021	**Deviance:**	4158.1
Time:	00:03:28	**Pearson chi2:**	4.80e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2040	0.196	1.043	0.297	-0.179	0.587
Do Not Email	-1.5037	0.193	-7.774	0.000	-1.883	-1.125
TotalVisits	11.1489	2.665	4.184	0.000	5.926	16.371
Total Time Spent on Website	4.4223	0.185	23.899	0.000	4.060	4.785
Lead Origin_Lead Add Form	4.2051	0.258	16.275	0.000	3.699	4.712
Lead Source_Olark Chat	1.4526	0.122	11.934	0.000	1.214	1.691
Lead Source_Welingak Website	2.1526	1.037	2.076	0.038	0.121	4.185
Last Activity_Had a Phone Conversation	2.7552	0.802	3.438	0.001	1.184	4.326
Last Activity_SMS Sent	1.1856	0.082	14.421	0.000	1.024	1.347
What is your current occupation_Student	-2.3578	0.281	-8.392	0.000	-2.908	-1.807
What is your current occupation_Unemployed	-2.5445	0.186	-13.699	0.000	-2.908	-2.180
Last Notable Activity_Unreachable	2.7846	0.807	3.449	0.001	1.202	4.367

	Features	VIF
9	What is your current occupation_Unemployed	2.82
2	Total Time Spent on Website	2.00
1	TotalVisits	1.54
7	Last Activity_SMS Sent	1.51
3	Lead Origin_Lead Add Form	1.45
4	Lead Source_Olark Chat	1.33
5	Lead Source_Welingak Website	1.30
0	Do Not Email	1.08
8	What is your current occupation_Student	1.06
6	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

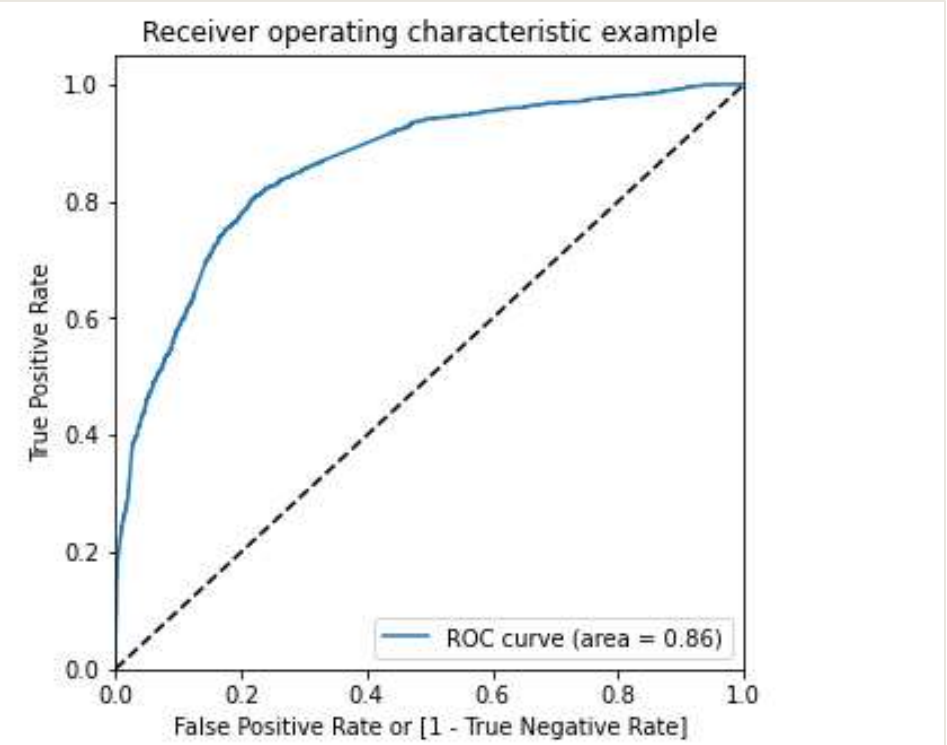
ROC CURVE

An ROC (Receiver Operating Characteristic) 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.

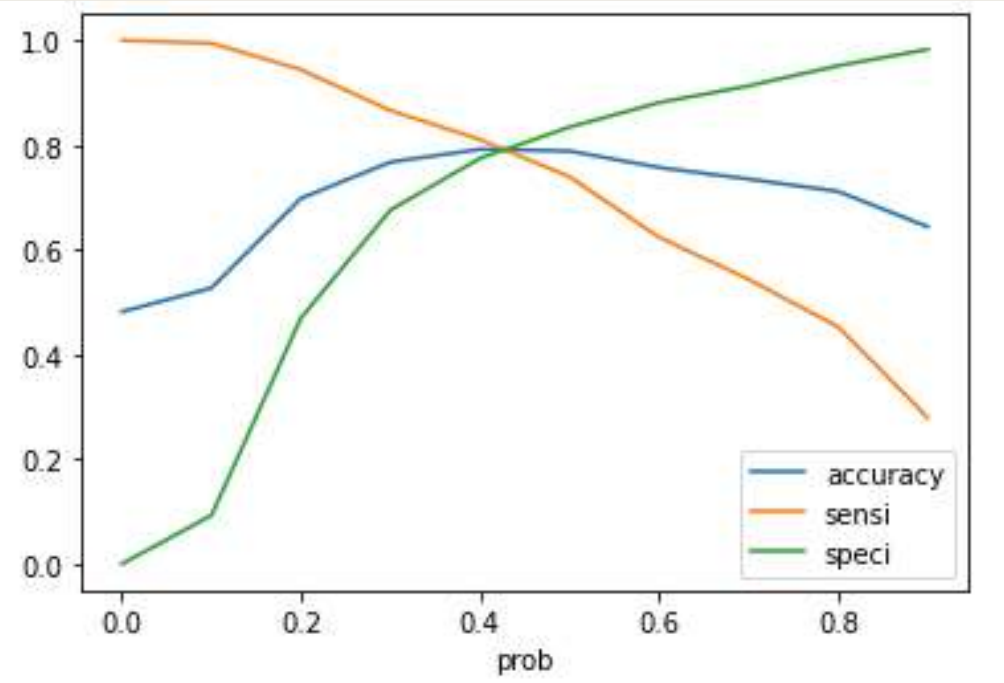
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test



We have the Area Under the Curve as 0.86 which is a good model.

OPTIMAL CUT-OFF (SENSITIVITY VS SPECIFICITY VS ACCURACY)

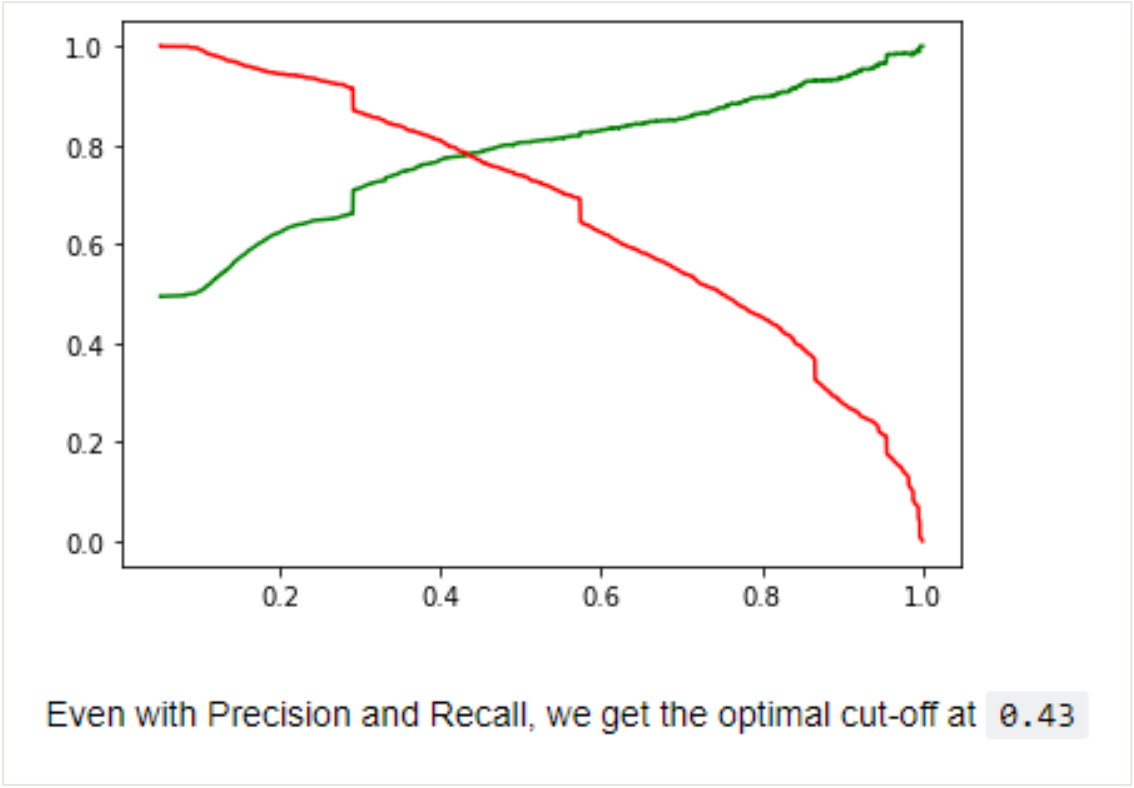
- We can see that the cut-off point is 0.43 which is at the intersection of accuracy, sensitivity and specificity.
- We can also see the accuracy, sensitivity and specificity values for various probabilities



prob	accuracy	sensi	speci
0.0	0.481731	1.000000	0.000000
0.1	0.527012	0.994416	0.092561
0.2	0.698274	0.944160	0.469723
0.3	0.767541	0.865984	0.676038
0.4	0.791975	0.810610	0.774654
0.5	0.788612	0.739414	0.834343
0.6	0.757229	0.624011	0.881055
0.7	0.735037	0.543509	0.913062
0.8	0.711500	0.453234	0.951557
0.9	0.644026	0.279665	0.982699

PRECISION VS RECALL TRADE-OFF

We get the same probability cut-off as 0.43 with precision and recall trade-off curve



MODEL PERFORMANCE

Final model results:

Train

Accuracy = 78.95%

Sensitivity = 78.54%

Specificity = 79.32%

Test

Accuracy = 78.5%

Sensitivity = 77.51%

Specificity = 79.41%

LEAD SCORE

The below table contains the lead score 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

	Converted	Conversion_Prob	Final_Predicted	Lead Score
0	1	0.996296	1	99.6
1	0	0.129992	0	13.0
2	0	0.703937	1	70.4
3	1	0.299564	0	30.0
4	1	0.720796	1	72.1
5	1	0.792250	1	79.2
6	0	0.704038	1	70.4
7	1	0.464521	1	46.5
8	0	0.282978	0	28.3
9	1	0.786460	1	78.6
10	1	0.987981	1	98.8
11	0	0.351053	0	35.1
12	0	0.189840	0	19.0
13	1	0.472712	1	47.3
14	1	0.876810	1	87.7
15	0	0.185066	0	18.5
16	1	0.390269	0	39.0
17	1	0.935454	1	93.5
18	0	0.316030	0	31.6
19	0	0.400550	0	40.1

BUSINESS OPPORTUNITIES ARE LIKE BUSES. THERE'S ALWAYS ANOTHER ONE COMING.

Richard Branson

RECOMMENDATIONS

The attributes mentioned below will helps us select the most promising leads that are most likely to convert into paying customers. They are listed in their order of significance.

1) The **total number of visits** made by the customer on the website
2) The **total time spent** by the customer on the website

3) The origin identifier with which the customer was identified to be a lead was through **'Lead Add Form'**

4) Last activity performed by the customer was a **phone conversation**

5) The source of the lead is via **'Welingak Website'**
6) The source of the lead is via '**Olark Chat**'

7) Last activity performed by the customer was **'SMS Sent'**

8) The customer is a **'Student'**

9) Current Occupation of the customer is **'Unemployed'**

MODELING FOR DIFFERENT SCENARIOS

X Education has a period of 2 months every year during which they hire some interns. The sales team has around 10 interns allotted to them. So, during this phase, they wish to make the lead conversion more aggressive. So, they want almost all the potential leads (i.e., the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.

The below table shows the Probability cut-off and its relevant model accuracy, sensitivity, and specificity. We can interpret Sensitivity as the % of customers to call and Specificity as the % of Customers that make a successful lead conversion. As we have more manpower, we must strategize to reach more customers with a good successful conversion rate. Hence, we can choose the probability cut-off as 0.4, where we can reach 81% of customers to get a 77.5% conversion.

	prob	accuracy	sensi	speci
0.0	0.0	0.481731	1.000000	0.000000
0.1	0.1	0.527012	0.994416	0.092561
0.2	0.2	0.698274	0.944160	0.469723
0.3	0.3	0.767541	0.865984	0.676038
0.4	0.4	0.791975	0.810610	0.774654
0.5	0.5	0.788612	0.739414	0.834343
0.6	0.6	0.757229	0.624011	0.881055
0.7	0.7	0.735037	0.543509	0.913062
0.8	0.8	0.711500	0.453234	0.951557
0.9	0.9	0.644026	0.279665	0.982699

At times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So, during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e., they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.

The below table shows the Probability cut-off and its relevant model accuracy, sensitivity, and specificity. We can interpret Sensitivity as the % of customers to call and Specificity as the % of Customers that make a successful lead conversion. As we want to minimize phone calls, we must strategize to reach less customers with high conversion rate. Thus, we can choose the probability cut-off as 0.8, where we can reach only 45% of customers to see a 95% conversion.

	prob	accuracy	sensi	speci
0.0	0.0	0.481731	1.000000	0.000000
0.1	0.1	0.527012	0.994416	0.092561
0.2	0.2	0.698274	0.944160	0.469723
0.3	0.3	0.767541	0.865984	0.676038
0.4	0.4	0.791975	0.810610	0.774654
0.5	0.5	0.788612	0.739414	0.834343
0.6	0.6	0.757229	0.624011	0.881055
0.7	0.7	0.735037	0.543509	0.913062
0.8	0.8	0.711500	0.453234	0.951557
0.9	0.9	0.644026	0.279665	0.982699

SUMMARY

We performed an end-to-end case study on Lead Scoring where we identified the important attributes for pursuing hot leads and arrive at the lead score along with recommendations. From the model, we were able to arrive at an accuracy of 78.5%.