LEAD SCORING STUDY CASE

By Study Group:

Deekshashree KM, Keerthana S, Vignesh Keshavan

PROBLEM STATEMENT:

An education company named X Education sells online courses to industry professionals. 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%`; although X Education gets a lot of leads, its lead conversion rate is very poor.

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.

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 a higher lead score have a higher conversion chance and the customers with a 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%`.

GOALS OF THE ANALYSIS:

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. Such as:

   a. X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, 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 of 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.

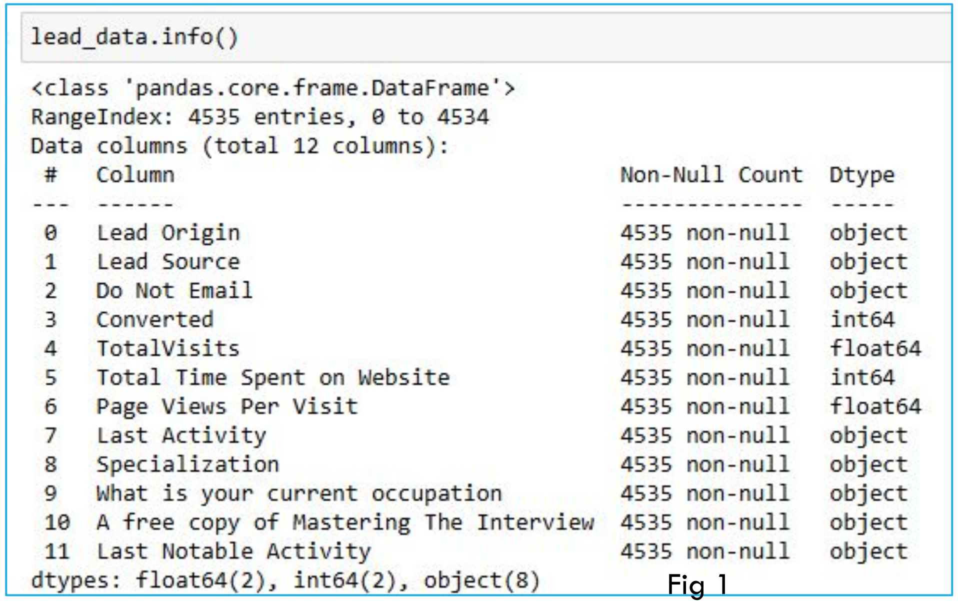
   b. Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize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

DATA EXPLORATION AND PREPROCESSING

The dataset contains 9,240 rows and 37 columns, with few numerical and more categorical variables. We observed that lot of null values were present in 'Asymmetrique' related variables and few other columns. So, we decided to drop all those columns having >30% of missing values.

Then we looked into remaining feature variables individually to understand it's importance in our analysis to retain or to drop. Here, we had 2 observation:

- That many variables showed supremacy towards one level, such columns were dropped as they were insignificant in the analysis.
- In columns with prominent ‘Select’ level data, indicated missing data; here, as-well we considered >30% threshold for missing values and dropped those columns. And for columns < 30% missing values we retained those but we dropped those rows with null values/‘Select’ option.



```
lead_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4535 entries, 0 to 4534
Data columns (total 12 columns):
 #   Column                                   Non-Null Count  Dtype  
---  ------                                   --------------  -----  
 0   Lead Origin                              4535 non-null   object 
 1   Lead Source                              4535 non-null   object 
 2   Do Not Email                             4535 non-null   object 
 3   Converted                                4535 non-null   int64  
 4   TotalVisits                              4535 non-null   float64
 5   Total Time Spent on Website              4535 non-null   int64  
 6   Page Views Per Visit                     4535 non-null   float64
 7   Last Activity                            4535 non-null   object 
 8   Specialization                           4535 non-null   object 
 9   What is your current occupation          4535 non-null   object 
 10  A free copy of Mastering The Interview   4535 non-null   object 
 11  Last Notable Activity                    4535 non-null   object 
dtypes: float64(2), int64(2), object(8)
```

Fig 1

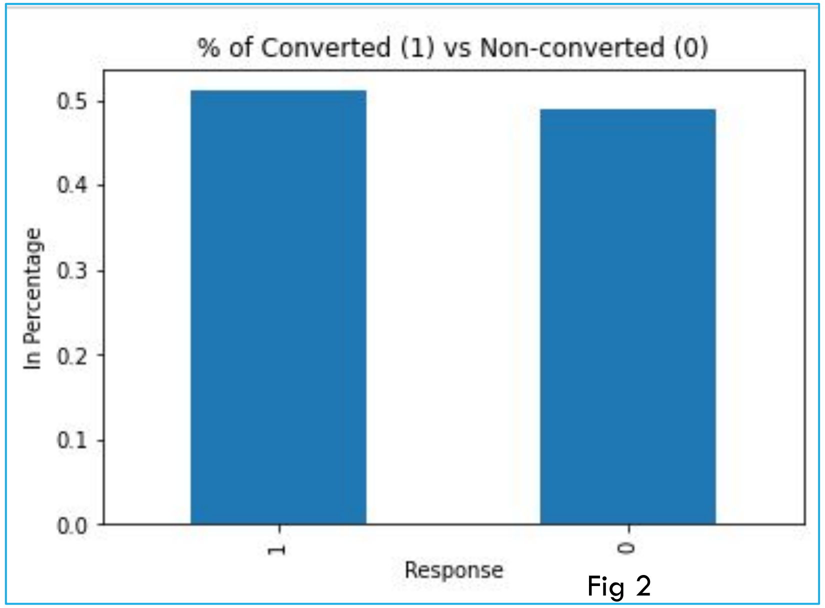
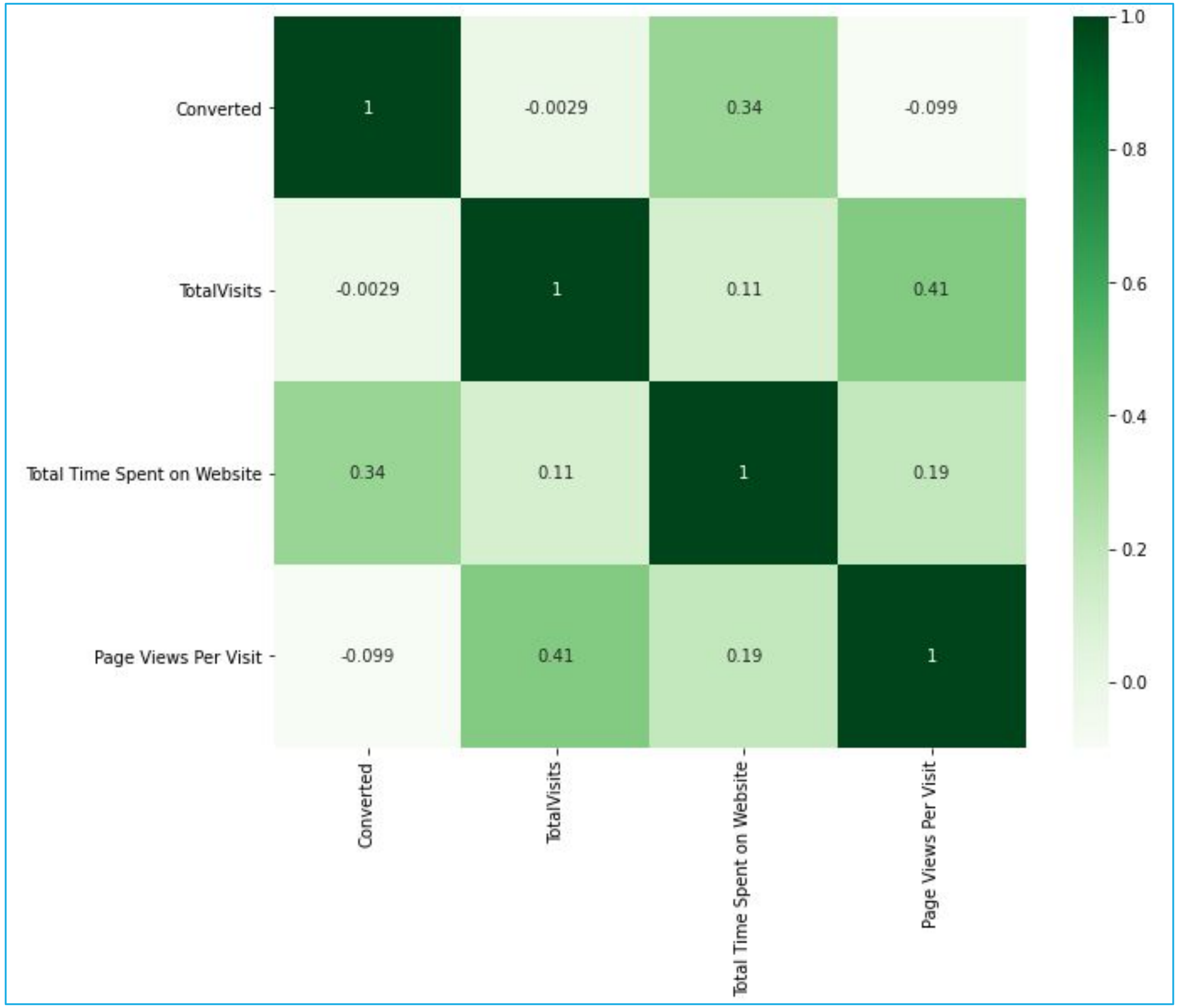


Fig 2

Fig 1: Post cleaning we had 4535 rows and 12 columns dataset. Out of it for 8 categorical col, dummy variables were created resulting in 73 feature variables for analysis.

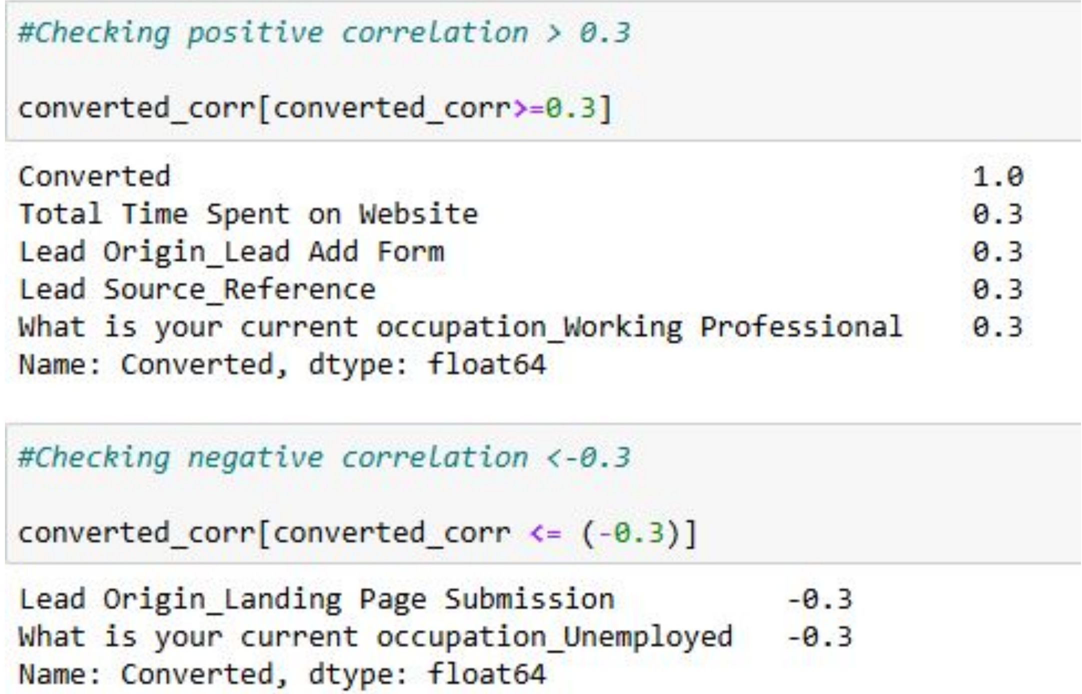
Fig 2: No data imbalance was observed in the given dataset. The percentage of conversion was around 50% and non-conversion 48%, which is near-equal distribution.

KEY INSIGHTS FROM EDA



We could observe slight correlation of 0.34 between 'total time spent on website' vs 'converted'.

- We could observe:
  - ~0.3 weak positive correlation between 'Total Time Spent on Website', 'Lead Origin_Lead Add Form', 'Lead Source_Reference' & 'What is your current occupation_Working Professional' vs the 'Converted'.
  - ~0.3 weak negative correlation between 'Lead Origin_Landing Page Submission' & 'What is your current occupation_Unemployed' vs the 'Converted'



```
#Checking positive correlation > 0.3

converted_corr[converted_corr>=0.3]
```

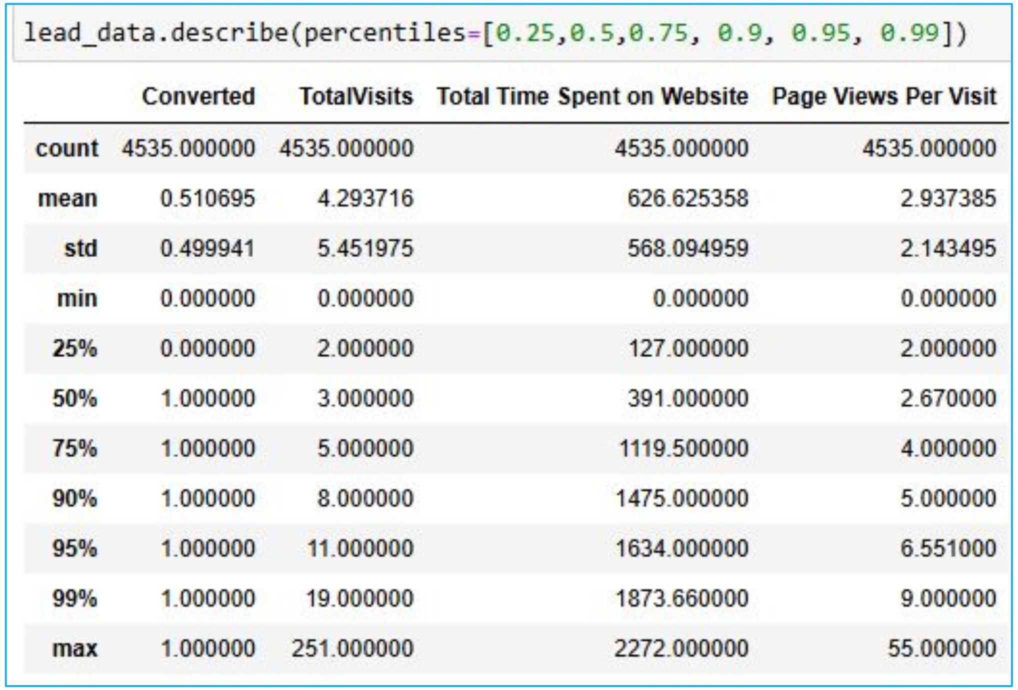
```
Converted                                           1.0
Total Time Spent on Website                         0.3
Lead Origin_Lead Add Form                           0.3
Lead Source_Reference                               0.3
What is your current occupation_Working Professional    0.3
Name: Converted, dtype: float64
```

```
#Checking negative correlation <-0.3

converted_corr[converted_corr <= (-0.3)]
```

```
Lead Origin_Landing Page Submission       -0.3
What is your current occupation_Unemployed   -0.3
Name: Converted, dtype: float64
```

OUTLIERS TREATMENT

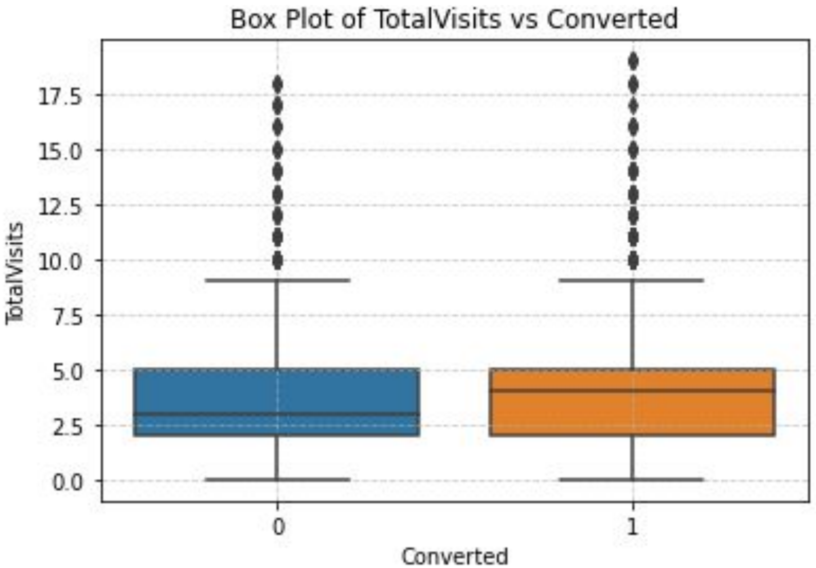


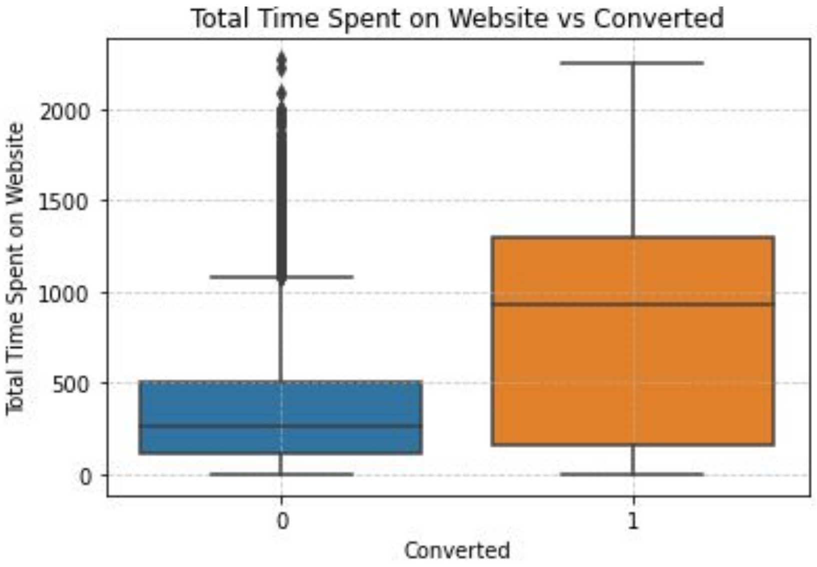
```
lead_data.describe(percentiles=[0.25,0.5,0.75, 0.9, 0.95, 0.99])
```

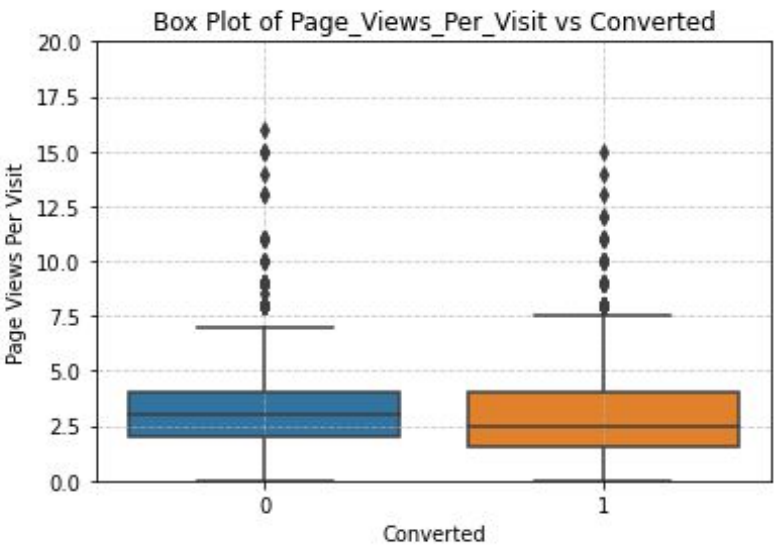
	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	4535.000000	4535.000000	4535.000000	4535.000000
mean	0.510695	4.293716	626.625358	2.937385
std	0.499941	5.451975	568.094959	2.143495
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	127.000000	2.000000
50%	1.000000	3.000000	391.000000	2.670000
75%	1.000000	5.000000	1119.500000	4.000000
90%	1.000000	8.000000	1475.000000	5.000000
95%	1.000000	11.000000	1634.000000	6.551000
99%	1.000000	19.000000	1873.660000	9.000000
max	1.000000	251.000000	2272.000000	55.000000

Per the stats value, we could observe high difference between mean and max values in 'TotalVisits' & 'Total Time Spent on Website' feature data.

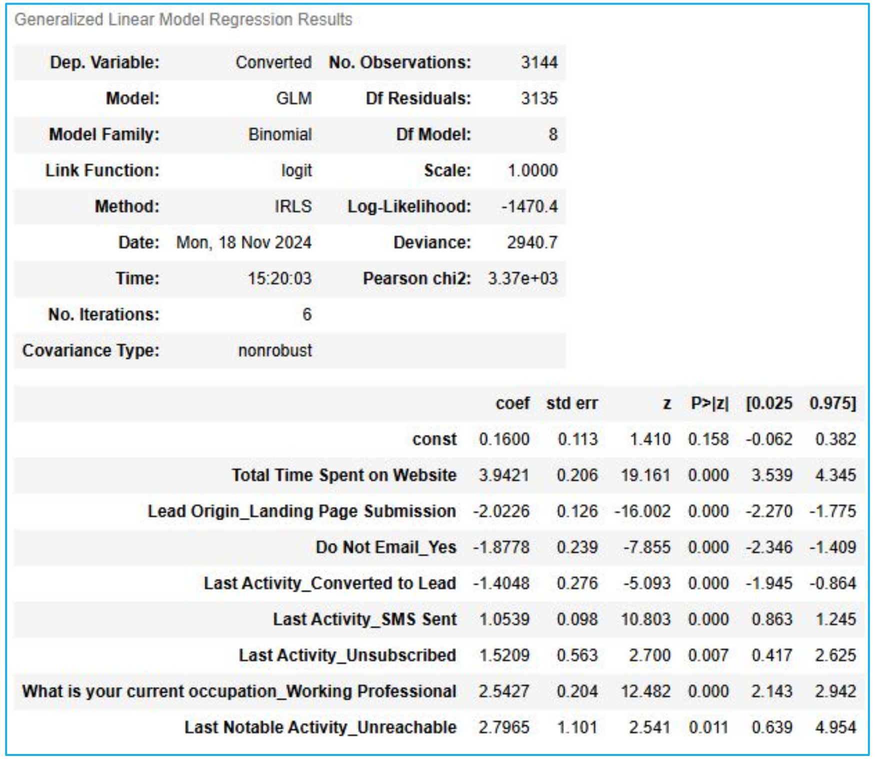
- For “TotalVisits”: There is no much significance in retaining the outliers in 'TotalVisits’. So, values up to 99% quantile were selected. Candidate visiting the website up to 250 times is out of the ordinary.
- For ‘Total Time Spent on Website’ : Spending more amount of time on the website translates to taking a concrete decision to pursue a course, the values are retained.







MODEL DEVELOPMENT: LOGISTIC REGRESSION

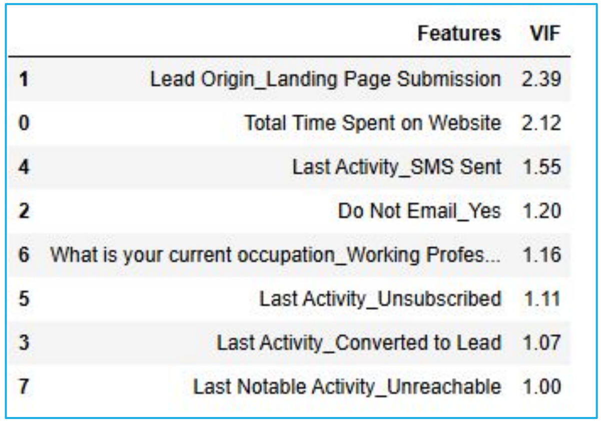


Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	3144
Model:	GLM	Df Residuals:	3135
Model Family:	Binomial	Df Model:	8
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1470.4
Date:	Mon, 18 Nov 2024	Deviance:	2940.7
Time:	15:20:03	Pearson chi2:	3.37e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1600	0.113	1.410	0.158	-0.062	0.382
Total Time Spent on Website	3.9421	0.206	19.161	0.000	3.539	4.345
Lead Origin_Landing Page Submission	-2.0226	0.126	-16.002	0.000	-2.270	-1.775
Do Not Email_Yes	-1.8778	0.239	-7.855	0.000	-2.346	-1.409
Last Activity_Converted to Lead	-1.4048	0.276	-5.093	0.000	-1.945	-0.864
Last Activity_SMS Sent	1.0539	0.098	10.803	0.000	0.863	1.245
Last Activity_Unsubscribed	1.5209	0.563	2.700	0.007	0.417	2.625
What is your current occupation_Working Professional	2.5427	0.204	12.482	0.000	2.143	2.942
Last Notable Activity_Unreachable	2.7965	1.101	2.541	0.011	0.639	4.954

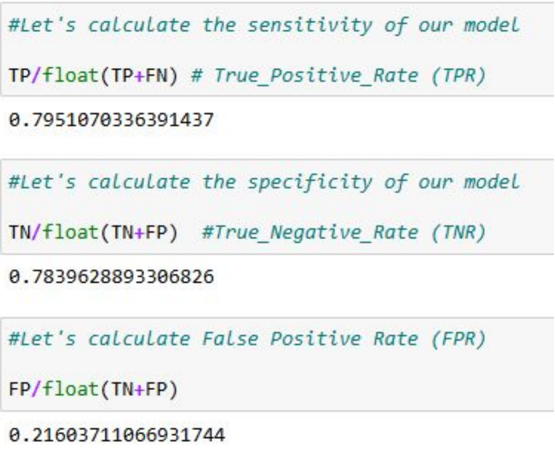
Final variables with p-values are <0.05 used for model evaluation



	Features	VIF
1	Lead Origin_Landing Page Submission	2.39
0	Total Time Spent on Website	2.12
4	Last Activity_SMS Sent	1.55
2	Do Not Email_Yes	1.20
6	What is your current occupation_Working Profes...	1.16
5	Last Activity_Unsubscribed	1.11
3	Last Activity_Converted to Lead	1.07
7	Last Notable Activity_Unreachable	1.00

Final variables with VIF < 5

- The target variable for this logistic regression model is 'Converted', which indicates whether a lead was converted (1) or not (0).
- 'MinMaxScaler' used as part of the data preprocessing steps to scale numerical features before feeding them into the logistic regression model.
- The data was split into training and testing datasets to evaluate model performance and generalization. 70-30% split is used, where 70% of the data is used for training and 30% for testing.
- For feature selection- RFE technique used to identify the most important 15 features for our model
- A logistic regression model was built using the training data to predicts the optimal probability cutoff point for lead conversion.



```
#Let's calculate the sensitivity of our model

TP/float(TP+FN) # True_Positive_Rate (TPR)
```

```
0.7951070336391437
```

```
#Let's calculate the specificity of our model

TN/float(TN+FP)  #True_Negative_Rate (TNR)
```

```
0.7839628893306826
```

```
#Let's calculate False Positive Rate (FPR)

FP/float(TN+FP)
```

```
0.21603711066931744
```

- To evaluate the model and to capture the incorrectly classified errors 'Confusion matrix' is used. And other metrics such as Accuracy, Sensitivity/Recall, Specificity, False Positive Rate, Positive Predicting value (Precision) were calculated.

MODEL PERFORMANCE:

- We observe high sensitivity and low FPR and plotted ROC curve and got 0.86 area under curve (AUC) [fig 1]

- Considering AUC-ROC of 0.86 to be sign of good model, we calculated optimal probability cutoff point to get balanced sensitivity and specificity.

- In 'Sensitivity-Specificity' view (fig 2), intersection was observed at ‘0.5’ probability. Similar optimum cutoff probability was also observed in Precision-Recall view (fig 3) as-well.

- At 0.5 cutoff probability, we got TPR= 79% and TNR= 78% and low FPR. This was good to proceed with test set prediction

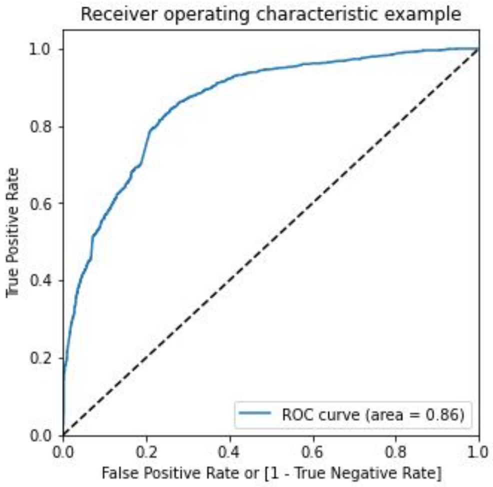


Fig 1: AUC-ROC

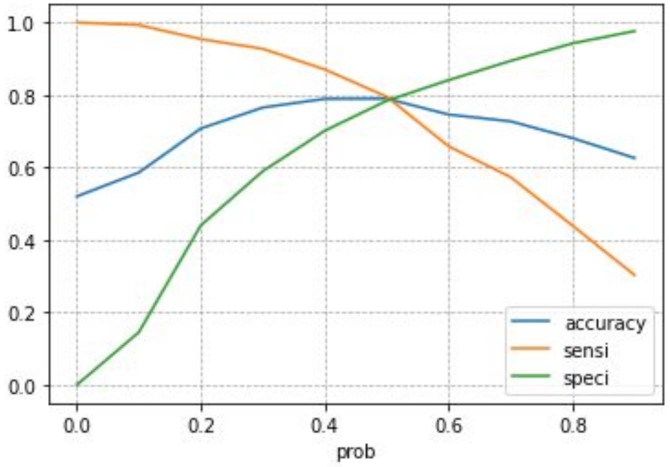


Fig 2: Sensitivity-Specificity view

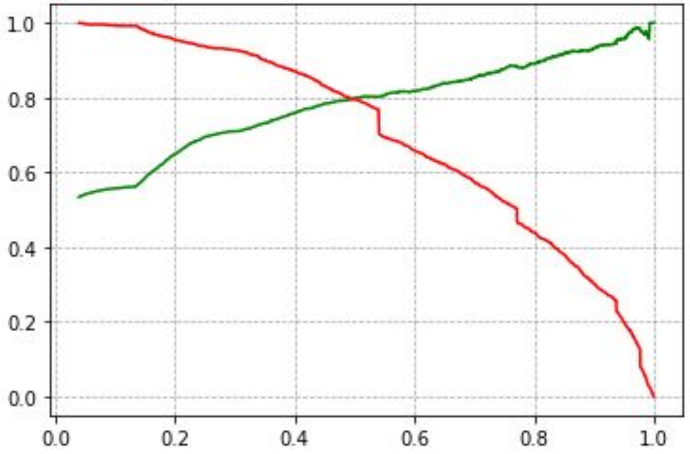
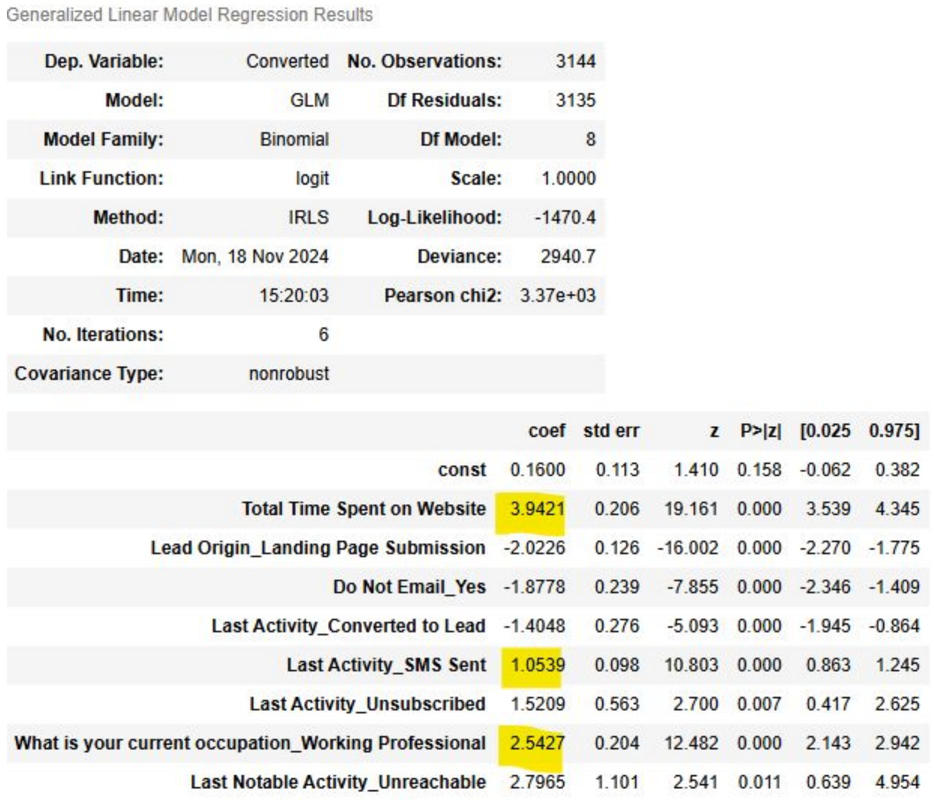


Fig 3: Precision-Recall view

DRIVERS OF LEAD CONVERSION:

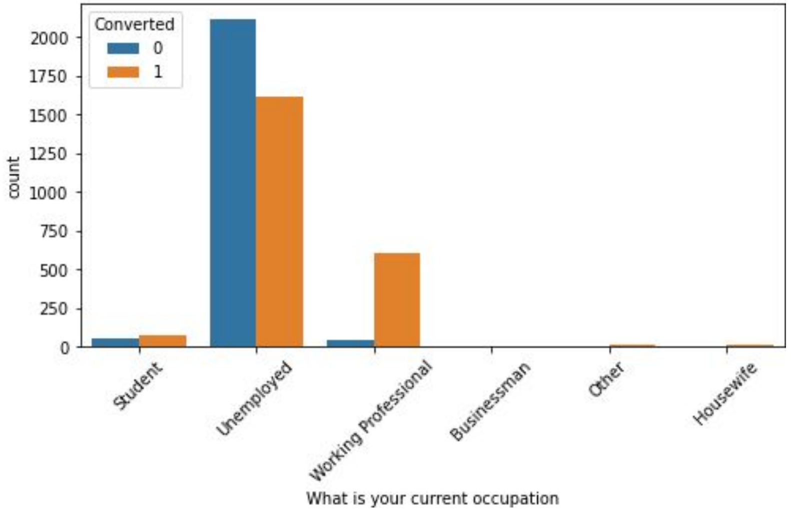
- To know the top three variables in our model which contribute most towards the probability of a lead getting converted were identified by checking the coefficient values in the our last model with p<0.05 and VIF<5. Higher the coefficient more probable variable is for lead conversion.

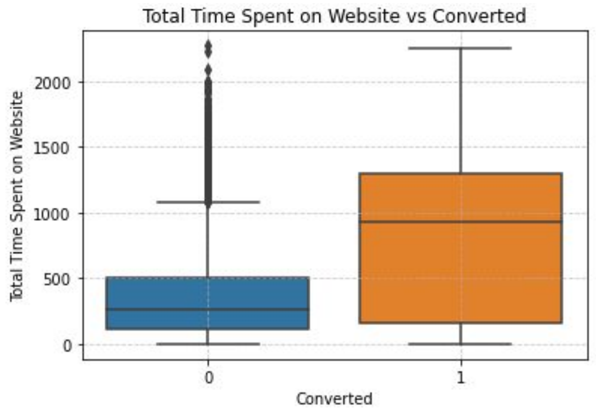


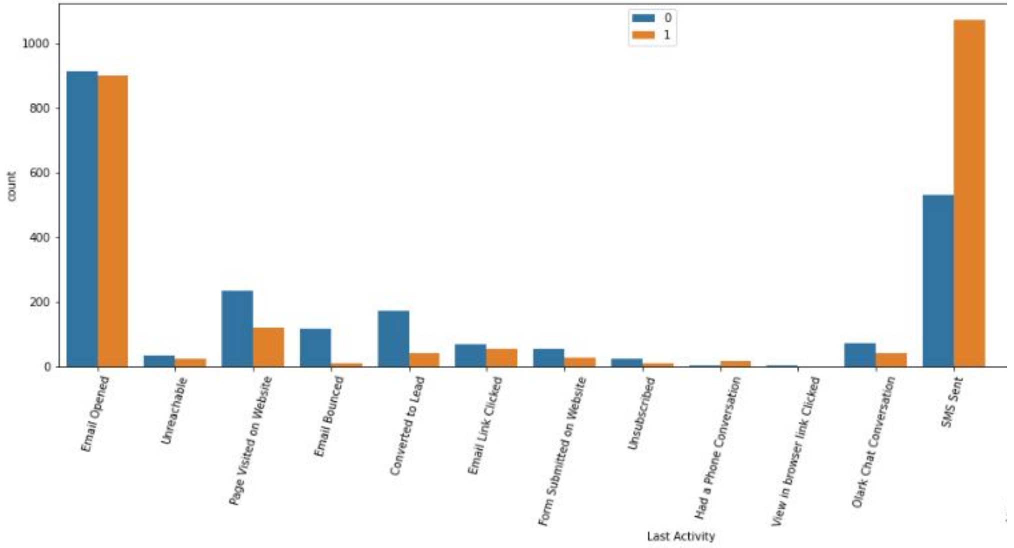
```
Generalized Linear Model Regression Results

   Dep. Variable:          Converted   No. Observations:      3144
           Model:                GLM       Df Residuals:      3135
    Model Family:           Binomial           Df Model:         8
   Link Function:              logit              Scale:    1.0000
          Method:               IRLS     Log-Likelihood:   -1470.4
            Date:   Mon, 18 Nov 2024           Deviance:    2940.7
            Time:           15:20:03      Pearson chi2:   3.37e+03
  No. Iterations:                  6
 Covariance Type:          nonrobust

                                               coef   std err        z   P>|z|   [0.025   0.975]
                                     const   0.1600     0.113    1.410   0.158   -0.062    0.382
                  Total Time Spent on Website   3.9421     0.206   19.161   0.000    3.539    4.345
          Lead Origin_Landing Page Submission  -2.0226     0.126  -16.002   0.000   -2.270   -1.775
                            Do Not Email_Yes  -1.8778     0.239   -7.855   0.000   -2.346   -1.409
                Last Activity_Converted to Lead  -1.4048     0.276   -5.093   0.000   -1.945   -0.864
                       Last Activity_SMS Sent   1.0539     0.098   10.803   0.000    0.863    1.245
                   Last Activity_Unsubscribed   1.5209     0.563    2.700   0.007    0.417    2.625
What is your current occupation_Working Professional   2.5427     0.204   12.482   0.000    2.143    2.942
           Last Notable Activity_Unreachable   2.7965     1.101    2.541   0.011    0.639    4.954
```

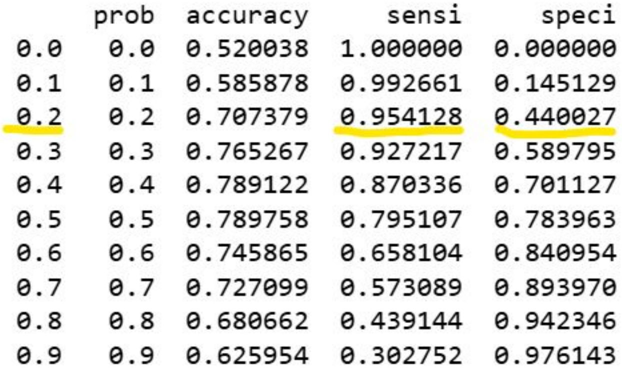






STRATEGIES FOR AGGRESSIVE AND LOW-PRIORITY PERIODS:

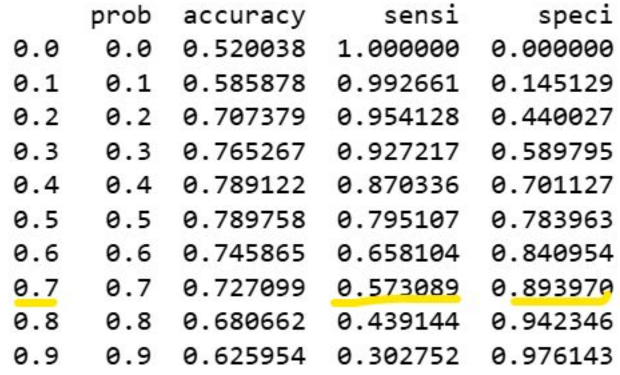
The goal of X Education's sales team to maximize lead conversion during their 2-month intern hiring period, where they have 10 interns available to aggressively pursue potential leads. So, our Strategy for Aggressive Lead Conversion is to Lower The Probability Cutoff value.



```
     prob  accuracy     sensi     speci
0.0   0.0  0.520038  1.000000  0.000000
0.1   0.1  0.585878  0.992661  0.145129
0.2   0.2  0.707379  0.954128  0.440027
0.3   0.3  0.765267  0.927217  0.589795
0.4   0.4  0.789122  0.870336  0.701127
0.5   0.5  0.789758  0.795107  0.783963
0.6   0.6  0.745865  0.658104  0.840954
0.7   0.7  0.727099  0.573089  0.893970
0.8   0.8  0.680662  0.439144  0.942346
0.9   0.9  0.625954  0.302752  0.976143
```

During the internship, when the selling skills of the interns are at their lowest, candidates with a lead score between 20% and 30% can be used for training. From the above table, we can observe that any candidate with a lead score at around 20%, has 95% (sensitivity) of lead conversion out of which 45% leads would not convert; Still, company will be able to target 50% lead conversion at this probability cutoff. This is fine by company as they want to make phone calls to as many people as possible.

During Low-Priority phase, minimizing the rate of "useless" phone calls can be achieved by Increase the Probability Cutoff Threshold. In this approach, company can prioritize quality over quantity, thus reducing the likelihood of unsuccessful phone calls and saving time.



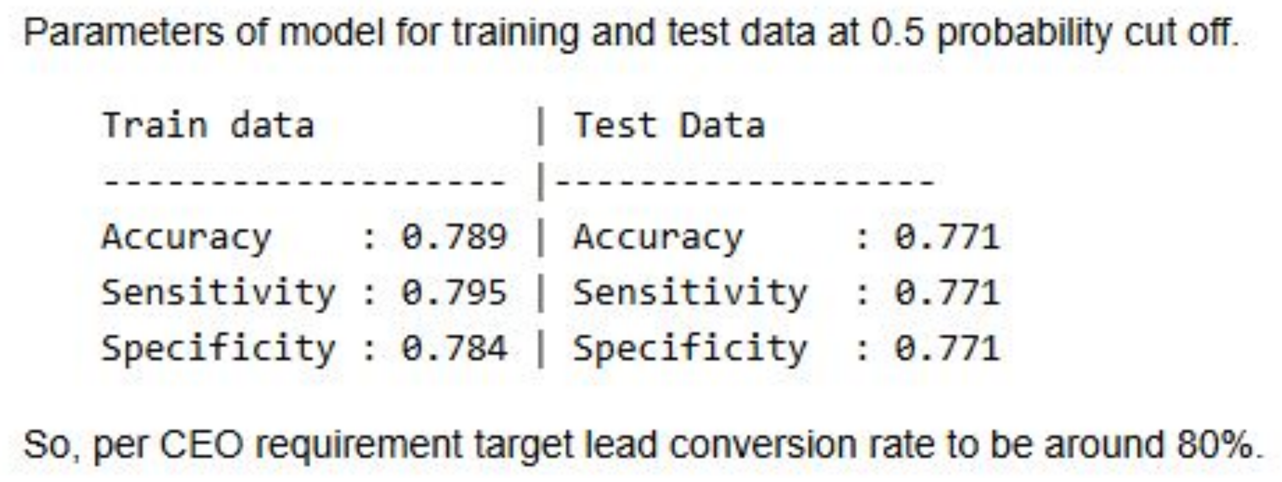
```
     prob  accuracy     sensi     speci
0.0   0.0  0.520038  1.000000  0.000000
0.1   0.1  0.585878  0.992661  0.145129
0.2   0.2  0.707379  0.954128  0.440027
0.3   0.3  0.765267  0.927217  0.589795
0.4   0.4  0.789122  0.870336  0.701127
0.5   0.5  0.789758  0.795107  0.783963
0.6   0.6  0.745865  0.658104  0.840954
0.7   0.7  0.727099  0.573089  0.893970
0.8   0.8  0.680662  0.439144  0.942346
0.9   0.9  0.625954  0.302752  0.976143
```

From the above table, we can observe that by focusing on candidates with lead score >70% (at highest probability of conversion) can get 57% leads that would convert with higher specificity of 89%. Allowing the sales team to avoid unnecessary interactions with leads by making fewer but more targeted calls.

RECOMMENDATIONS AND ACTION PLAN:

- Company can train interns on effective sales pitches, in understanding customer needs and to handle objections. And equip them with scripts tailored for different customer personalities to make calls more personalized and impactful.
- The company can evaluate the effectiveness this proposed strategy (lowered lead score threshold for aggressive period) by weekly reviews of call records, feedbacks and target achieved. Based on the regular evaluation feedback from the sales team, company can adjust the threshold by looking into success rates.
- Similarly, company can evaluate the effectiveness of higher lead score threshold for low-priority period by monitoring conversion rates from the limited phone calls that are made and adjust the lead probability cutoff based on ongoing performance data to ensure phone calls continue to be highly impactful.
- In order to ensure team productivity during low-priority period, company can encourage the sales team to engage in valuable tasks, such as improving their customer relationships, participating in training, creating new content on their website, conducting market research or supporting new initiatives.

CONCLUSION:



Parameters of model for training and test data at 0.5 probability cut off.

```
Train data          | Test Data
------------------- |------------------
Accuracy    : 0.789 | Accuracy     : 0.771
Sensitivity : 0.795 | Sensitivity  : 0.771
Specificity : 0.784 | Specificity  : 0.771
```

So, per CEO requirement target lead conversion rate to be around 80%.

By following structured and data-driven strategy, X Education can make the most of their interns' efforts during the aggressive lead conversion period. This approach ensures optimal focus on high-priority leads, enhances the chances of conversions through targeted and personalized interactions, and provides ongoing adaptation based on real-time lead data and performance metrics.

Raising the lead score threshold and using predictive engagement tools allows the company to minimize unnecessary calls during downtime periods. This targeted approach optimizes efficiency, leverages automated communications, and enables the sales team to contribute to broader company goals beyond direct lead conversion.