Lead Score – Case Study

To build logistic regression Model for an education company named X education to predict most promising lead which will most likely join the online courses

Presented By :-
Rohit Joshi
Akshat Jain

Problem Statement and Objective

Objective

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.

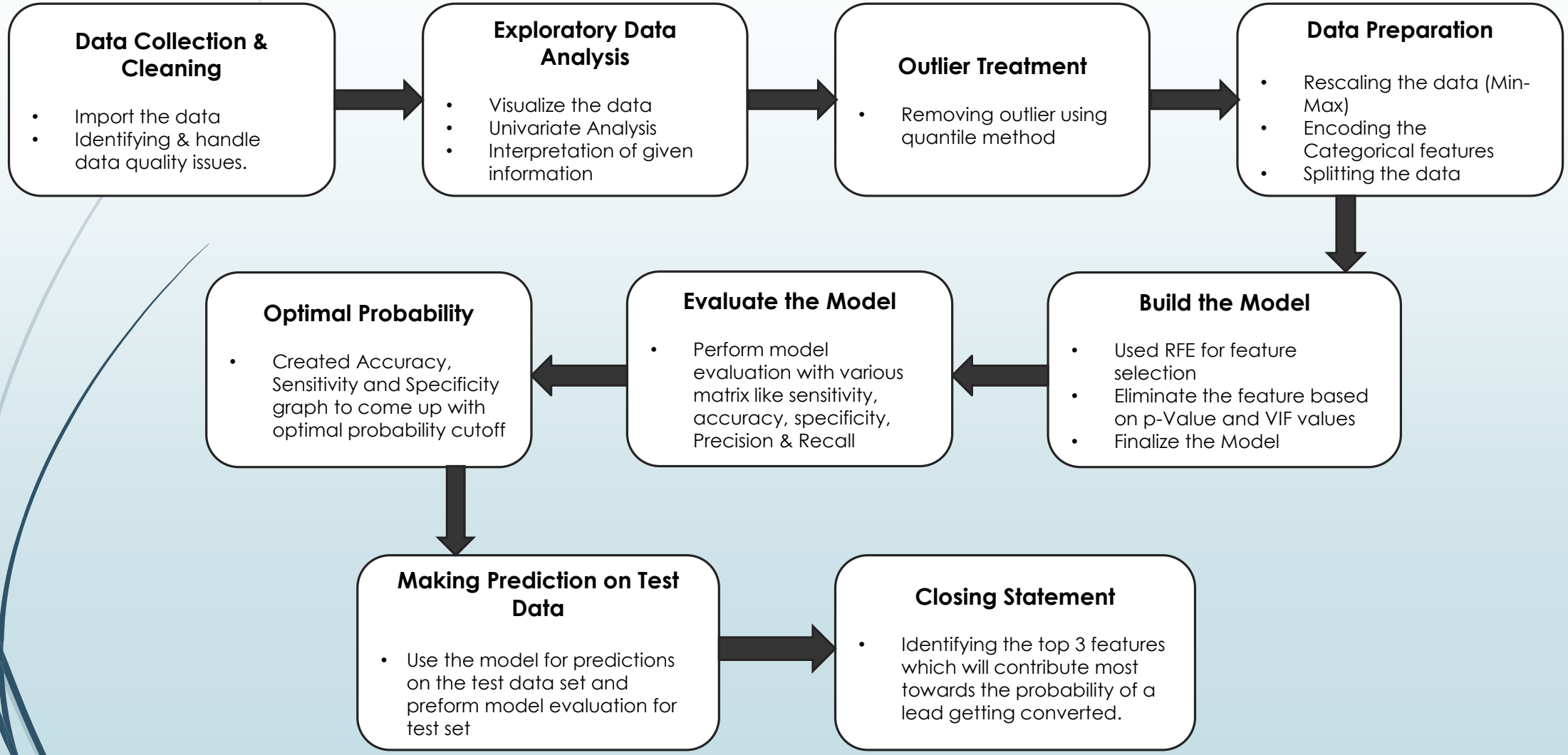
The objective is to help X Education to select the most promising lead i.e. leads that are most likely to convert into paying customer.

Problem Statement

The company requires 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Analysis Methodology



Data Cleaning & EDA

Removal based on unique values

- **Prospect ID & Lead Number** are just indicative of the ID number of the Contacted People & can be dropped.
- Features like **Magazine, Recieve More updates about the course, I agree to pay the amount through cheque etc…** have only one unique (set **No**)value. These features show no variance and doesn't make any impact or difference on conversion of leads.

Removal based on null values

- We have dropped all the columns which have 70% rows with null values like **"How did you hear about X education, Lead Profile.**

Handling 'Select' values

- Select values are mainly the non mandatory fields which the customer hasn't selected any option for. These has been handled as NAN.

Data Cleaning & EDA

Imputing Null values with Mode

- Around 60% of the **city data is Mumbai** so we can impute Mumbai in the missing values
- Nan in the 'Tags' column can be imputed by **'Will revert after reading the email'.**
- 95% of the data has country as India hence imputing with India
- For Current Occupation, **Unemployed** leads are the most in terms of Absolute numbers and hence can be imputed with the same.

Assigning category to null/unselected values

- As Lead quality is based on the intution of employee, so if left 'Select' we can impute 'Not Sure' in NaN safely.
- All the null value in the columns were categorised under separate column 'others', like null in Specialization,etc.

Removing rows with low NA

- Drop all rows which has NA value less than 2%

Data Cleaning & EDA

Combining low frequency label into one category

- Multiple columns with low frequency leads like ' Last Notable activity' with 1-2 leads are combined to 'Misc_Notable_Activity',
- similarly we have formed Misc_Activity, Misc_Tags and Misc_Lead_Source.

Removal based on Column with negligible Yes

- Newspaper, X Education Forums , Newspaper Articles, Through Recommendations & Digital Advertisements and search have 99% values as 'NO'.

Outlier treatment using Quantile method

- "Total Visits" & "Page Views Per Visit" are having large no. of outliers and hence was cap to 95% value for analysis.

Data Preparation & Feature Engineering

Encoding categorical features

- Following Categorical variables with multiple levels, dummy features were created.
- **Lead Origin', 'Lead Source', 'Do Not Email', 'Last Activity', 'Country', 'Specialization', 'What is your current occupation', 'Tags', 'Lead Quality', 'City', 'A free copy of Mastering The Interview', 'Last Notable Activity')**

Test train Split

- The original data frame was split into test and train dataset. The train dataset was used to train the model and test dataset was used to evaluate the model.

Feature scaling

- It is important that we rescale the feature such that they have a comparable scales. This can lead us to save time consuming during model evaluation.
- We have used Min-Max scaling (Normalization).

Feature selection using RFE

- **Recursive feature elimination (RFE**) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached.
- We have run the RFE on the output number of the variable equal to 15.
- Eliminate the feature based on **p-Value and VIF values**

```
1  cols=X_train.columns[rfe.support_]
2  cols
```

```
Index(['Total Time Spent on Website', 'Lead Origin_Lead Add Form',
       'Lead Source_Welingak Website', 'Do Not Email_Yes', 'Tags_Busy',
       'Tags_Closed by Horizzon', 'Tags_Lost to EINS', 'Tags_Ringing',
       'Tags_Will revert after reading the email', 'Tags_invalid number',
       'Tags_switched off', 'Lead Quality_Not Sure', 'Lead Quality_Worst',
       'Last Notable Activity_Olark Chat Conversation',
       'Last Notable Activity_SMS Sent'],
      dtype='object')
```

Building the model

- Generalized Linear models from **statsModel** is used to build the Logistic regression model.
- We started the model with 15 variables selection with RFE.
- Multiple variables were dropped in multiple iteration based on high P-value and High VIF.
- The final model was created with 13 essential features.
- The final Error distribution is more close to a normal distribution and more centered at 0.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7491	0.241	-11.409	0.000	-3.221	-2.277
Total Time Spent on Website	3.5840	0.212	16.900	0.000	3.168	4.000
Lead Origin_Lead Add Form	2.0492	0.373	5.492	0.000	1.318	2.780
Lead Source_Welingak Website	3.5157	1.088	3.231	0.001	1.383	5.648
Do Not Email_Yes	-1.1405	0.226	-5.055	0.000	-1.583	-0.698
Tags_Busy	4.0042	0.344	11.654	0.000	3.331	4.678
Tags_Closed by Horizzon	7.6453	0.787	9.719	0.000	6.104	9.187
Tags_Lost to EINS	8.9437	0.766	11.676	0.000	7.442	10.445
Tags_Ringing	-1.9216	0.373	-5.151	0.000	-2.653	-1.190
Tags_Will revert after reading the email	3.8040	0.248	15.318	0.000	3.317	4.291
Tags_switched off	-2.9499	0.795	-3.713	0.000	-4.507	-1.393
Lead Quality_Not Sure	-3.2813	0.138	-23.855	0.000	-3.551	-3.012
Lead Quality_Worst	-3.7712	0.847	-4.454	0.000	-5.431	-2.112
Last Notable Activity_SMS Sent	2.4903	0.127	19.614	0.000	2.241	2.739

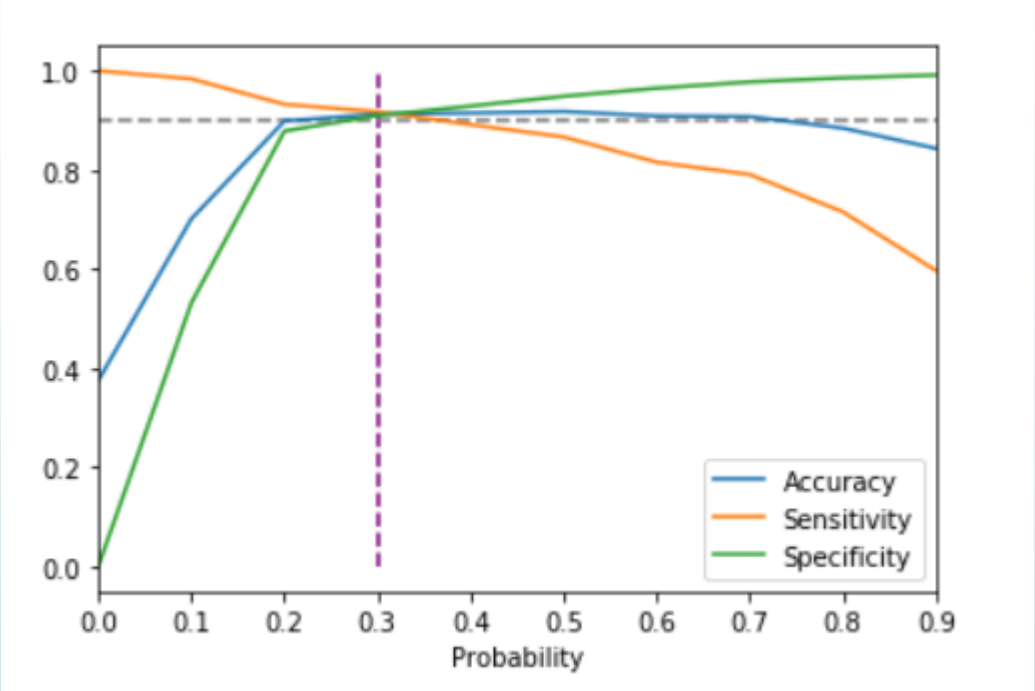
Getting Predicted Probability for Positive labels

- Creating a data frame with the actual converted flag and the predict probabilities.
- Creating new column 'Predicted' with 1 if **Conversion_Rate > 0.5** else 0
- Showing top 5 record of the data frame

	Lead_ID	Converted	Converted_Probability	Predicted
0	5279	0	0.054239	0
1	3099	0	0.637556	1
2	91	1	0.407737	0
3	1577	1	0.994512	1
4	487	0	0.133646	0

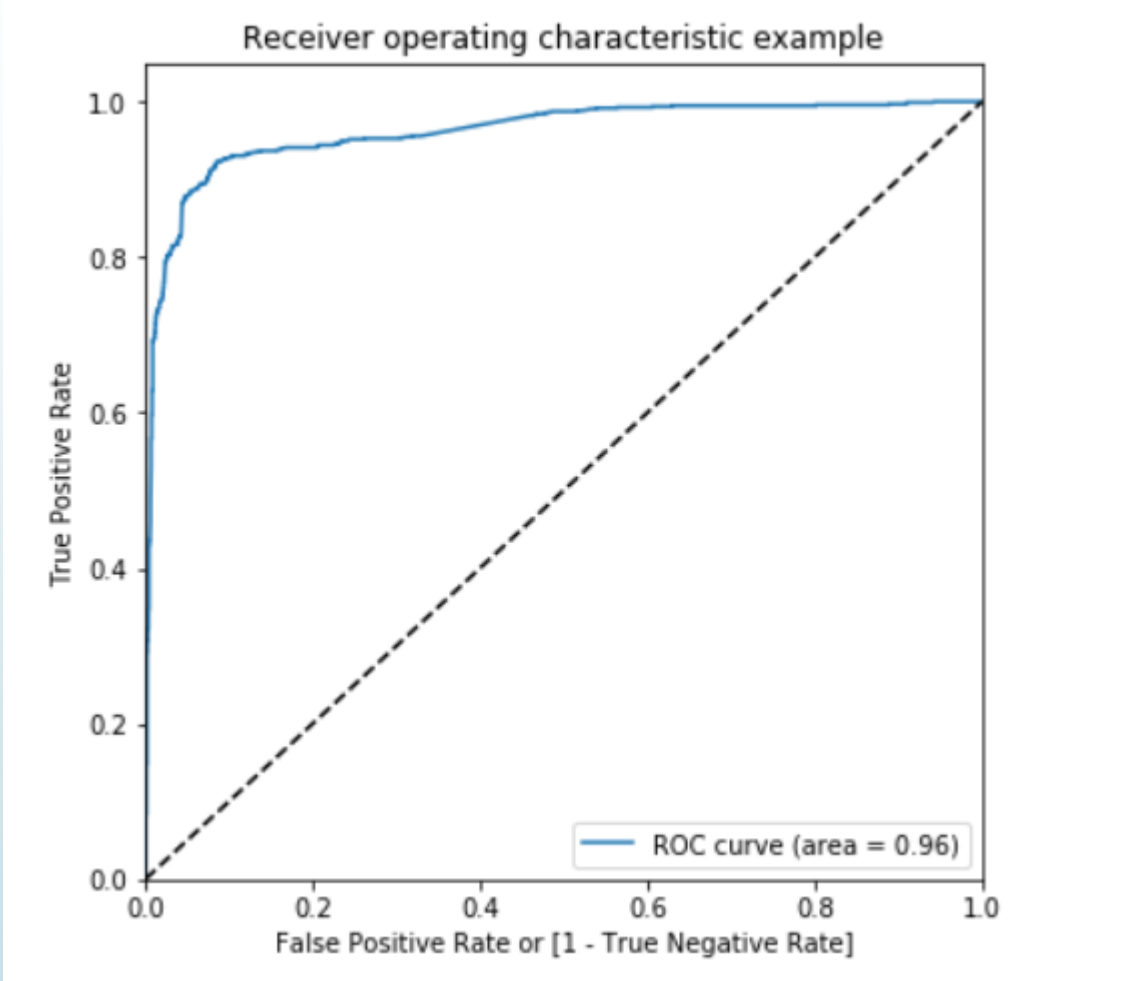
Finding Optimal Probability Threshold

- Optimal Probability is the intersection point at which there is a balance between sensitivity and specificity
- Accuracy sensitivity and specificity were calculated for various of probability threshold and plotted as graph.
- 0.3 was the optimum point found as a cutoff probability to predict the final probability.
- At this threshold value , all the metrics were found to be above 90% which is acceptable.



Plotting ROC curve and Calculating AUC

- Receiver operating characteristic (ROC curve): It shows the tradeoff between Sensitivity and Specificity (Both of them being inversely proportional)
- Area under cover (AUC): The goodness of the model can be determined by AUC. The larger the AUC, the better the model. The value of AUC for our model is **0.96**.



Evaluating the model on **train dataset**

- Logistic regression model Accuracy - 0.91
- Logistic regression model Precision - 0.85
- Logistic regression model Recall - 0.91
- Logistic regression model Sentivity - 0.91
- Logistic regression model Specifictiy - 0.90
- False positive of model-predicting converted when customer has not converted - 0.09
- Logistic regression model Predictive - 0.85

Making Predictions on the **test dataset**

- Model Accuracy on Test data is 0.91
- Sensitivity of the model on test data is 0.89
- Specificity of the model on test data is 0.92
- Precision : 0.88
- Recall : 0.89
- Positive predictive 0.88
- Negative predictive value 0.93

Lead Score Calculation

- The relative importance of each feature is determined on a scale of 100 with the feature with highest importance having a score of 100

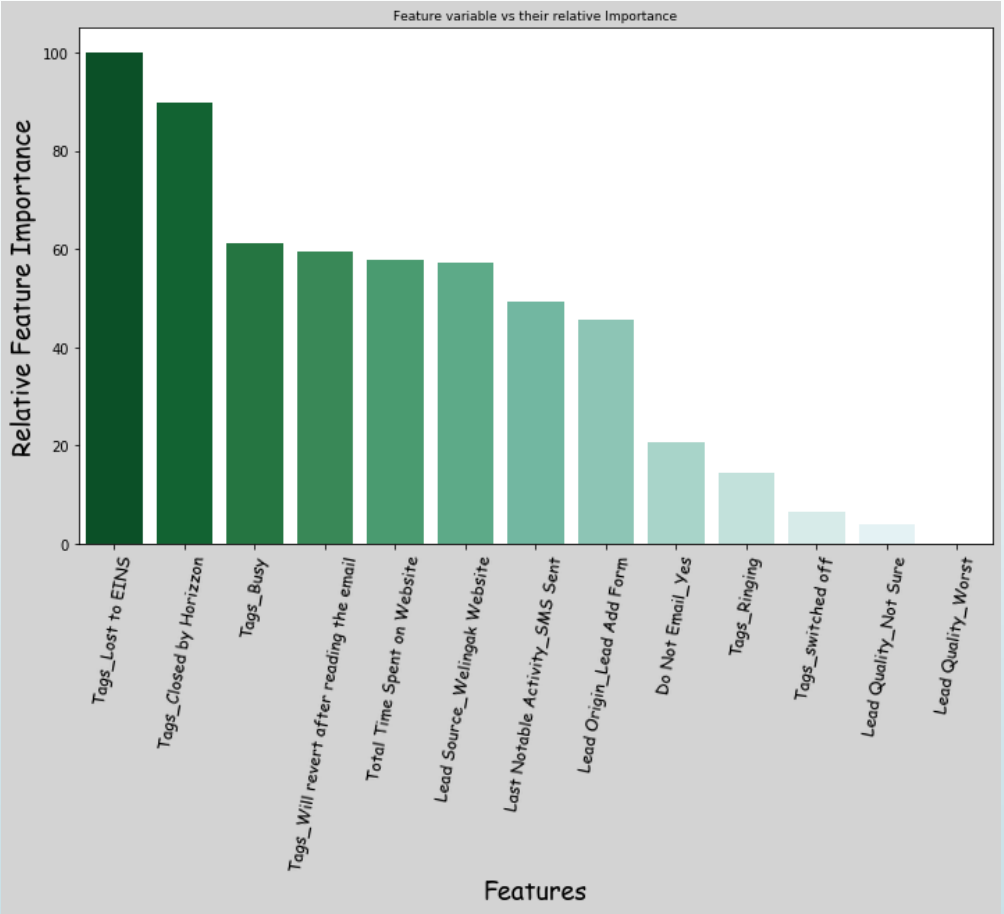
$$\text{Lead Score} = 100 * \frac{FI - \min(FI)}{\max(FI - \min(FI))}$$

where **FI is Feature Importance**

	Relative Feature Importance	Feature
0	100.00	Tags_Lost to EINS
1	89.79	Tags_Closed by Horizzon
2	61.15	Tags_Busy
3	59.58	Tags_Will revert after reading the email
4	57.85	Total Time Spent on Website
5	57.31	Lead Source_Welingak Website
6	49.25	Last Notable Activity_SMS Sent
7	45.78	Lead Origin_Lead Add Form
8	20.69	Do Not Email_Yes
9	14.55	Tags_Ringing
10	6.46	Tags_switched off
11	3.85	Lead Quality_Not Sure
12	0.00	Lead Quality_Worst

Determining the Feature Importance

- 13 features have ben used by the model to predict whether a lead will get converted or not.
- The coefficient values for each of these features are used to determine the order of importance.
- Features with high positive coefficients are the ones that contributed the most towards the probability of lead getting converted.



Closing Statement

- After several iterations, our model has following characteristics:
  - All variables have low P-values (i.e <0.05)
  - All features have very low VIF values which means , there is hardly any multicollinearity among the features.
  - The overall accuracy of 0.91 at a probability threshold of 0.3 is also very good.

Closing Statement

- The conversion probability of a lead increases with increase in values of the following features in descending order

	Relative Feature Importance	Feature
Tags_Lost to EINS	8.94	Tags_Lost to EINS
Tags_Closed by Horizzon	7.65	Tags_Closed by Horizzon
Tags_Busy	4.00	Tags_Busy
Tags_Will revert after reading the email	3.80	Tags_Will revert after reading the email
Total Time Spent on Website	3.58	Total Time Spent on Website
Lead Source_Welingak Website	3.52	Lead Source_Welingak Website
Last Notable Activity_SMS Sent	2.49	Last Notable Activity_SMS Sent
Lead Origin_Lead Add Form	2.05	Lead Origin_Lead Add Form

- The conversion probability of a lead increases with decrease in values of the follwing features in descending order

Do Not Email_Yes	-1.14	Do Not Email_Yes
Tags_Ringing	-1.92	Tags_Ringing
Tags_switched off	-2.95	Tags_switched off
Lead Quality_Not Sure	-3.28	Lead Quality_Not Sure
Lead Quality_Worst	-3.77	Lead Quality_Worst

Recommendation and Problem Solution

- Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?
  - Tags_Lost to EINS
  - Tags_Closed by Horizzon
  - Tags_busy

- What are the top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion?
  - Tags_Lost to EINS
  - Tags_Closed by Horizzon
  - Tags_busy

Recommendation and Problem Solution

- 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.
  - We can suggest X education to decrease probability threshold due to which the sensitivity of model becomes high and specificity becomes low.
  - High sensitivity will lead to classify some leads which will not going to convert as converted but since X-Education has leverage of extra employee's for 2 months they could make as many calls as possible to potential leads and convert them.

Recommendation and Problem Solution

- 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.

  - We can suggest X education to increase probability threshold due to which the sensitivity of model becomes low and specificity becomes high.
  - High specificity will lead to classify some borderline leads which might convert or not as non-converted but since X-Education has achieved their target for quarter it shouldn't call them and minimize their rate of useless phone calls.

Thanks