Lead Scoring Case Study
SUBMISSION REPORT

- *To build a Logistic Regression Model to predict whether a lead for online courses for an education company named X Education would be successfully converted or not.*

Business Objective

- *To help X Education to select the most promising leads(Hot Leads), i.e. the leads that are most likely to convert into paying customers.*
- *To build a logistic regression model to assign a lead score value between 0 and 100 to each of the leads which can be used by the company to target potential leads.*

The objective is thus classified into the following sub-goals:

Create a Logistic Regression model to predict the Lead Conversion probabilities for each lead.

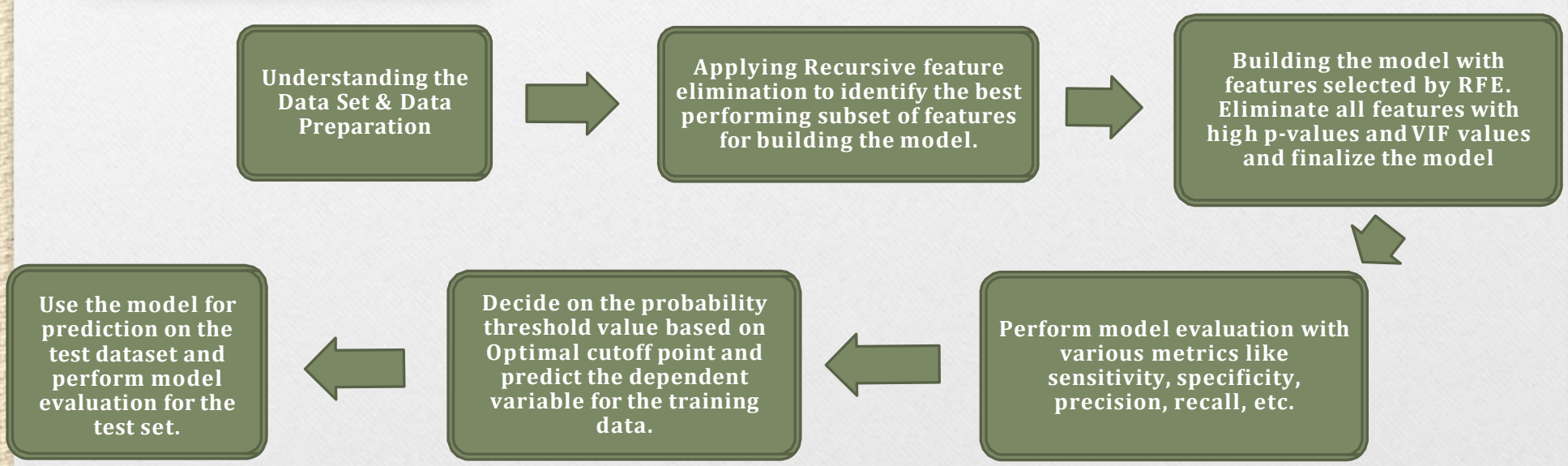
Decide on a probality threshold value above which a lead will be predicted as converted, whereas not converted if it is below it.

Multiply the Lead Conversion probability to arrive at the Lead Score value for each lead.

Problem

Solving Methodology

- The approach for this project has been to divide the entire case study into various checkpoints to meet each of the sub-goals. The checkpoints are represented in a sequential flow as below:



Data Preparation & Feature Engineering

The following data preparation processes were applied to make the data dependable so that it can provide significant business value by improving Decision Making Process:

Remove columns which has only one unique value

- Deleting the following columns as they have only one unique value and hence cannot be responsible in predicting a successful lead case – '**Magazine**', '**Receive More Updates About Our Courses**' , '**Update me on Supply Chain Content**' , '**Update me on Supply Chain Content**' and '**I agree to pay the amount through cheque**'.

Removing rows where a particular column has high missing values

- '**Lead Source**' is an important column for analysis. Hence all the rows that have null values for it were dropped.

Imputing NULL values with Median

- The columns '**TotalVisits**' and '**Page Views Per Visit**' are continuous variables with outliers. Hence the null values for these columns were imputed with the column median values.

Imputing NULL values with Mode

- The columns '**Country**' is a categorical variable with some null values. Also majority of the records belong to the Country 'India'. Thus imputed the null values for this with mode(most occuring value). Then binned rest of category into 'Outside India'.

Data Preparation & Feature Engineering contd…

Handling 'Select' values in some columns

- There are some columns in dataset which have a level/value called 'Select'. This might have happened because these fields in the website might be non mandatory fields with drop downs options for the customer to choose from. Amongst the dropdown values, the default option is probably 'Select' and since these aren't mandatory fields, many customer might have have chosen to leave it as the default value 'Select'.
- The Select values in columns were **converted to Nulls**.

Assigning a Unique Category to NULL/SELECT values

- All the nulls in the columns were binned into a separate column '**Unknown**'.
- Instead of deleting columns with huge null value percentage(which results in loss of data), this strategy adds more information into the dataset and results in the change of variance.
- The Unknown levels for each of these columns will be finally dropped during dummy encoding.

Outlier Treatment

- The outliers present in the columns '**TotalVisits**' & '**Page Views Per Visit**' were finally removed based on iterquatile range analysis.

Binary Encoding

- Converting the following binary variables (Yes/No) to 0/1:
- '**Search**','**Do Not Email**', '**Do Not Call**', '**Newspaper Article**', '**X Education Forums**', '**Newspaper**', '**Digital Advertisement**','**Through Recommendations**' and '**A free copy of Mastering The Interview**'

Data Preparation & Feature Engineering contd…

Dummy Encoding

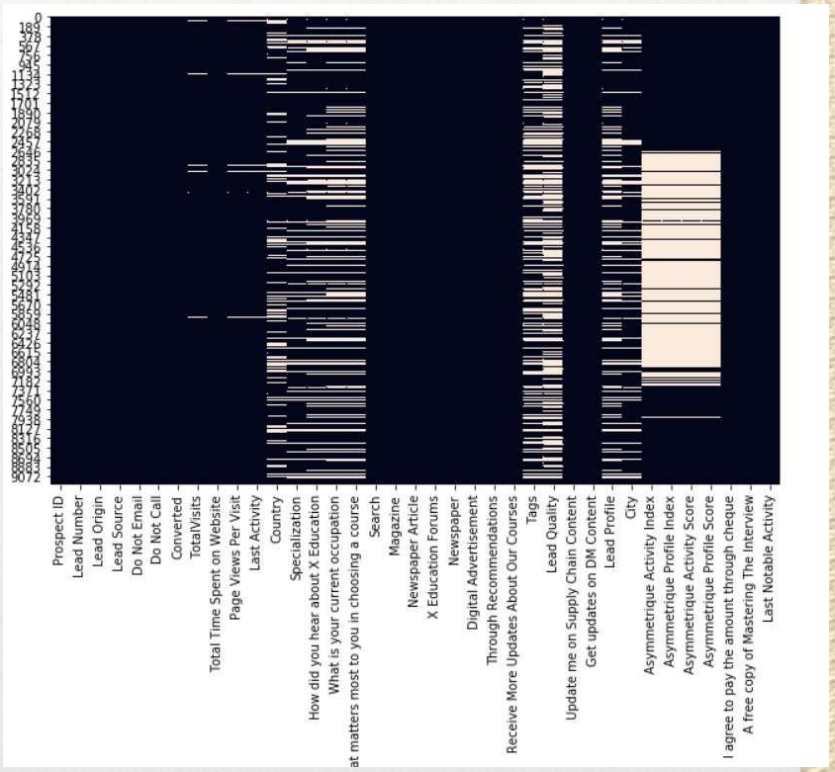
- For the following categorical variables with multiple levels, dummy features (one-hot encoded) were created:
- **'Lead Quality','Asymmetrique Profile Index','Asymmetrique Activity Index','Tags','Lead Profile', 'Lead Origin','What is your current occupation', 'Specialization', 'City','Last Activity', 'Country'** and **'Lead Source','Last Notable Activity'**

Test-Train Split

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

Feature Scaling

- Scaling helps in interpretation. It is important to have all variables(specially categorical ones which has values 0 and 1) on the same scale for the model to be easily interpretable.
- '**Standardisation**' was used to scale the data for modelling. It basically brings all of the data into a standard normal distribution with mean at zero and standard deviation one.



Distribution of NULL values in the different columns of the original dataset before null-handling.

Feature Selection Using RFE

•**Recursive feature elimination** is an optimization technique for finding the best performing subset of features. It is based on the idea of repeatedly constructing a model and choosing either the best (based on coefficients), setting the feature aside and then repeating the process with the rest of the features. This process is applied until all the features in the dataset are exhausted. Features are then ranked according to when they were eliminated.

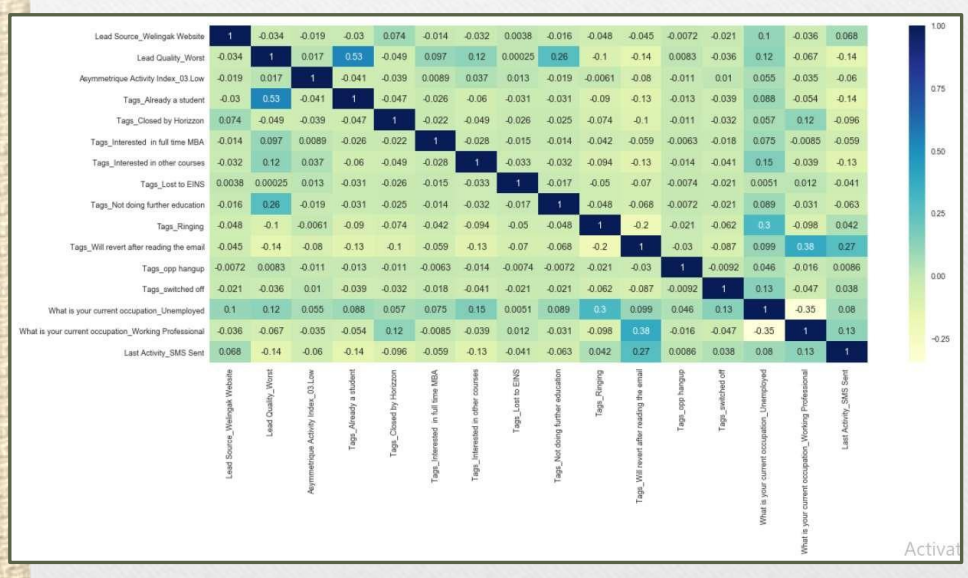
```
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()

from sklearn.feature_selection import RFE
rfe = RFE(logreg, 20)             # running RFE with 30 variables as output
rfe = rfe.fit(X_train, y_train)

col = X_train.columns[rfe.support_]
col
```

```
Index(['Lead Source_Welingak Website', 'Lead Quality_Worst',
       'Asymmetrique Activity Index_03.Low', 'Tags_Already a student',
       'Tags_Closed by Horizzon', 'Tags_Diploma holder (Not Eligible)',
       'Tags_Interested  in full time MBA', 'Tags_Interested in other courses',
       'Tags_Lost to EINS', 'Tags_Not doing further education', 'Tags_Ringing',
       'Tags_Will revert after reading the email', 'Tags_invalid number',
       'Tags_number not provided', 'Tags_opp hangup', 'Tags_switched off',
       'Tags_wrong number given', 'What is your current occupation_Unemployed',
       'What is your current occupation_Working Professional',
       'Last Activity_SMS Sent'],
      dtype='object')
```

•**Running RFE with the output number of the variable equal to 20.**

Building the model

- Generalized Linear Models from StatsModels is used to build the Logistic Regression model.
- The model is built initially with the 20 variables selected by RFE.
- Unwanted features are dropped serially after checking p values (< 0.5) and VIF (< 5) and model is built multiple times.
- The final model with 16 features, passes both the significance test and the multi-collinearity test.



Features	VIF
Tags_Closed by Horizzon	1.26
Tags_Not doing further education	1.23
Tags_switched off	1.17
Tags_Interested in full time MBA	1.10
Lead Source_Welingak Website	1.08
Asymmetrique Activity Index_03.Low	1.07
Tags_Lost to EINS	1.06
Tags_opp hangup	1.02
What is your current occupation_Working Profes...	0.77
Lead Quality_Worst	0.67
Tags_Ringing	0.58
Tags_Interested in other courses	0.38
Tags_Already a student	0.36
Tags_Will revert after reading the email	0.09
What is your current occupation_Unemployed	0.01
Last Activity_SMS Sent	0.00

Features	p-Value
const	0.00
Lead Source_Welingak Website	0.00
Lead Quality_Worst	0.00
Asymmetrique Activity Index_03.Low	0.00
Tags_Already a student	0.00
Tags_Closed by Horizzon	0.00
Tags_Interested in full time MBA	0.00
Tags_Interested in other courses	0.00
Tags_Lost to EINS	0.00
Tags_Not doing further education	0.00
Tags_Ringing	0.00
Tags_Will revert after reading the email	0.00
Tags_opp hangup	0.00
Tags_switched off	0.00
What is your current occupation_Unemployed	0.00
What is your current occupation_Working Profes...	0.00
Last Activity_SMS Sent	0.00

- A heat map consisting of the final 16 features proves that there is no significant correlation between the independent variables.

Predicting the Conversion Probability and Predicted column

Creating a dataframe with the actual Converted flag and the predicted probabilities.

Showing top 5 records of the dataframe in the picture on the right.

	Converted	Conversion_Prob	LeadID
0	0	0.064688	8529
1	0	0.009566	7331
2	1	0.762190	7688
3	0	0.077626	92
4	0	0.077626	4908

	Converted	Conversion_Prob	LeadID	predicted
0	0	0.064688	8529	0
1	0	0.009566	7331	0
2	1	0.762190	7688	1
3	0	0.077626	92	0
4	0	0.077626	4908	0

Creating new column 'predicted' with 1 if Conversion_Prob > 0.5 else 0

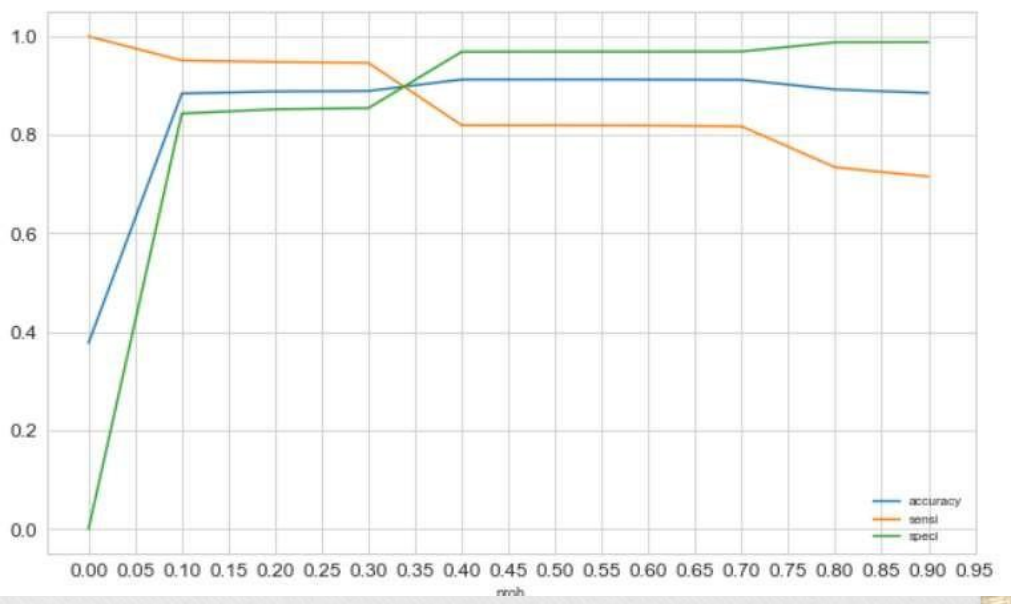
Showing top 5 records of the dataframe in the picture on the left.

Finding Optimal Probability Threshold

Optimal cutoff probability is that prob where we get balanced sensitivity and specificity.

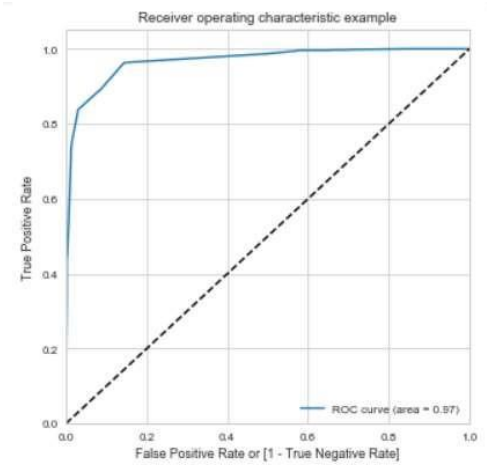
Optimal Probability Threshold

- The accuracy sensitivity and specificity was calculated for various values of probability threshold and plotted in the graph to the right.
- From the curve above, **0.33** is found to be the optimum point for cutoff probability.
- At this threshold value, all the 3 metrics - accuracy sensitivity and specificity was found to be well above 80% which is a well acceptable value.



Receiver Operating Characteristics (ROC) Curve

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).



Area under the Curve (GINI)

- By determining the Area under the curve (AUC) of the ROC curve, the goodness of the model is determined. Since the ROC curve is more towards the upper-left corner of the graph, it means that the model is very good. The larger the AUC, the better will is the model.
- The value of AUC for our model is 0.9678.

As a rule of thumb, an AUC can be classed as follows,

- 0.90 - 1.00 = excellent
- 0.80 - 0.90 = good
- 0.70 - 0.80 = fair
- 0.60 - 0.70 = poor
- 0.50 - 0.60 = fail

Since we got a value of 0.9678, our model seems to be doing well on the test dataset.

Evaluating the model on train datatset

Confusion Matrix

# Predicted	Not Converted	Converted
# Actual		
Not Converted	3411	325
Converted	256	2010

Probability Threshold = 0.33

Accuracy TP +TN/ (TP+TN+FN+FP)	Sensitivity TP / (TP+FN)	Specificity TN / (TN+FP)	False Positive Rate FP/ (TN+FP)	Positive Predictive Value TP / (TP+FP)
0.903	0.887	0.913	0.087	0.860

Negative Predictive Value TN / (TN+ FN)	Precision TP / TP + FP	Recall TP / TP + FN	*F1 score = 2×(Precision*Recall)/(Precision+ Recall)*	Area under the cuve
0.930	0.861	0.887	0.874	0.962

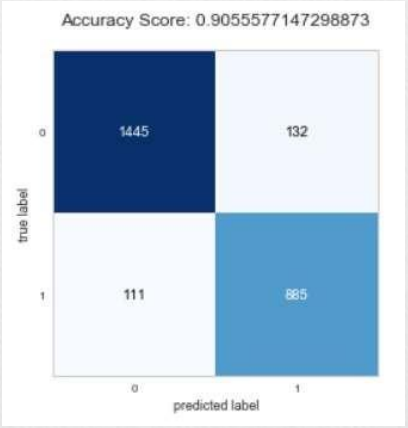
Making predictions on the test set

- The final model on the train dataset is used to make predictions for the test dataset
- The train data set was scaled using the scaler.transform function that was used to scale the train dataset.
- The Predicted probabilities were added to the leads in the test dataframe.
- Using the probability threshold value of 0.33, the leads from the test dataset were predicted if they will convert or not.

- The Conversion Matrix was calculated based on the Actual and Predicted 'Converted' columns.

The top 5 records from the final test data set

	LeadID	Converted	Conversion_Prob	final_predicted
0	6190	0	0.000591	0
1	7073	0	0.077626	0
2	4519	0	0.309185	0
3	607	1	0.999825	1
4	440	0	0.077626	0



Evaluating the model on test datatset

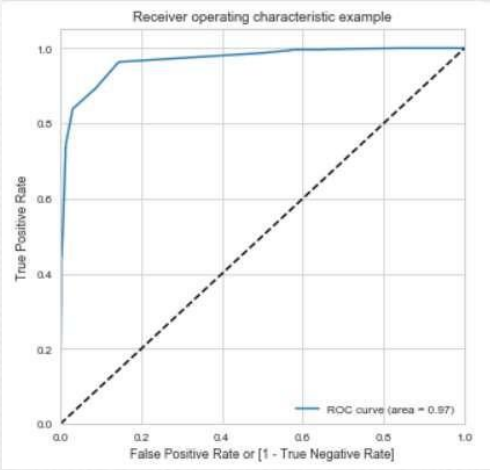
The following evaluation metrices were recorded for the test dataset.

Accuracy	Sensitivity	Specificity	Area under the cuve
TP +TN/ (TP+TN+FN+FP)	TP / (TP+FN)	TN / (TN+FP)	
0.906	0.889	0.916	0.968

Negative Predictive Value	Precision	Recall	*F1 score = 2×(Precision*Recall)/(Precision+Recall)*
TN / (TN+ FN)	TP / TP + FP	TP / TP + FN	
0.928	0.870	0.889	0.879

False Positive Rate	Positive Predictive Value	C ross Validation Score
FP/ (TN+FP)	TP / (TP+FP)	
0.084	0.870	0.913

Area under the Curve



Classification Report

	precision	recall	f1-score	support
0	0.93	0.92	0.92	1577
1	0.87	0.89	0.88	996
avg / total	0.91	0.91	0.91	2573

Lead Score Calculation

Lead Score is calculated for all the leads in the original dataframe.

Formula for Lead Score calculation is:

Lead Score = 100 * Conversion Probability

LeadID	Lead Number	Converted	Conversion_Prob	final_predicted	Lead_Score
0	660737	0	0.03	0	3
1	660728	0	0.01	0	1
2	660727	1	0.80	1	80
3	660719	0	0.01	0	1
4	660681	1	0.96	1	96
5	660680	0	0.08	0	8
6	660673	1	0.96	1	96
7	660664	0	0.08	0	8
8	660624	0	0.08	0	8
9	660616	0	0.08	0	8

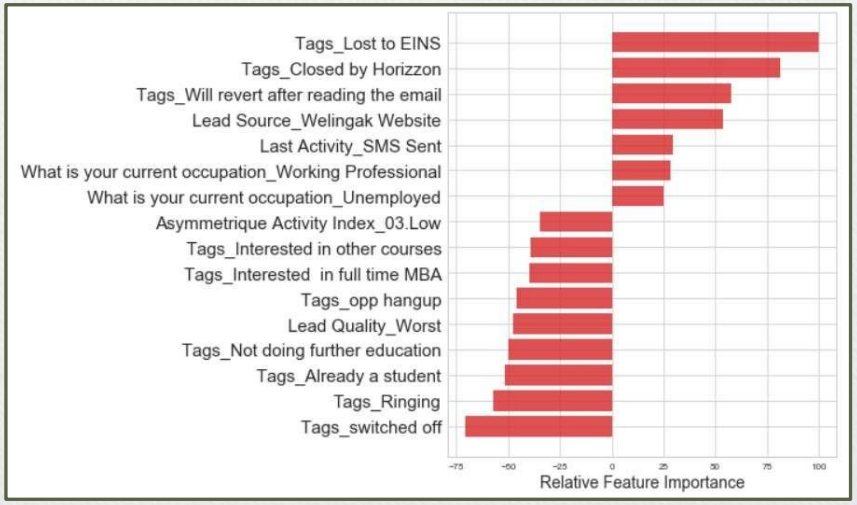
The figure showing Lead Score for top 10 records from the data set.

- The train and test dataset is concatenated to get the entire list of leads available.
- The Conversion Probability is multiplied by 100 to obtain the Lead Score for each lead.
- Higher the lead score, higher is the probability of a lead getting converted and vice versa,
- Since, we had used 0.33 as our final Probability threshold for deciding if a lead will convert or not, any lead with a lead score of 34 or above will have a value of '1' in the final_predicted column.

Determining Feature Importance

•16 features have been used by our model to successfully predict if a lead will get converted or not.
•The Coefficient (beta) values for each of these features from the model parameters are used to determine the order of importance of these features.
•Features with high positive beta values are the ones that contribute most towards the probability of a lead getting converted.
•Similarly, features with high negative beta values contribute the least.

Lead Source_Welingak Website	3.61
Lead Quality_Worst	-3.18
Asymmetrique Activity Index_03.Low	-2.34
Tags_Already a student	-3.45
Tags_Closed by Horizzon	5.44
Tags_Interested in full time MBA	-2.66
Tags_Interested in other courses	-2.63
Tags_Lost to EINS	6.71
Tags_Not doing further education	-3.35
Tags_Ringing	-3.84
Tags_Will revert after reading the email	3.87
Tags_opp hangup	-3.08
Tags_switched off	-4.73
What is your current occupation_Unemployed	1.67
What is your current occupation_Working Professional	1.89
Last Activity_SMS Sent	1.97



The **Relative Importance** of each feature is determined on a scale of 100 with the feature with highest importance having a score of 100.

feature_importance = 100.0 * (feature_importance / feature_importance.max())

The features are then sorted using Quick Sort algorithm.

Finally the sorted features are plotted in a bar graph in descending order of their relative importance.

Inference

After trying several models, we finally chose a model with the following characteristics:

- All variables have p-value < `0.05` .
- All the features have very low VIF values, meaning, there is `hardly any muliticollinearity` among the features. This is also evident from the heat map.
- The overall accuracy of `0.9056` at a probability threshold of 0.33 on the test dataset is also very acceptable.

Using this model, the dependent variable value was predicted as per the following threshold values of Conversion probability:

Dataset	Threshhold value	Accuracy	Sensitivity	Specificity	False Postive Rate	Positive Predictive Value	Negative Predictive value	Precision	Recall	F1 value	Cross Validation Score	AUC
train	0.50	0.9125	0.8195	0.9690	0.0310	0.9412	0.8985					0.9624
train	0.33	0.9032	0.8870	0.9130	0.0870	0.8608	0.9302	0.8608	0.8870	0.8737		0.9624
test	0.33	0.9056	0.8886	0.9163	0.0837	0.8702	0.9287	0.8702	0.8886	0.8793	0.9123	0.9679

Inference contd...

Based on our model, some features are identified which contribute most to a Lead getting converted successfully.

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

Features with Positive Coefficient Values
Tags_Lost to EINS
Tags_Closed by Horizzon
Tags_Will revert after reading the email
Lead Source_Welingak Website
Last Activity_SMS Sent
What is your current occupation_Working Professional
What is your current occupation_Unemployed

The conversion probability of a lead increases with decrease in values of the following features in descending order:

Features with Negative Coefficient Values
Tags_switched off
Tags_Ringing
Tags_Already a student
Tags_Not doing further education
Lead Quality_Worst
Tags_opp hangup
Tags_Interested in full time MBA
Tags_Interested in other courses
Asymmetrique Activity Index_03.Low

Recommendations & Problem Solution

Which are the top three variables in your model that contribute most towards the probability of a lead getting converted?	• **Tags_Lost to EINS** • **Tags_Closed by Horizzon** • **Tags_Will revert after reading the email**
What are the top 3 categorical/dummy variables in the model which get maximum focus in order to increase the probability of lead conversion?	• **Tags_Lost to EINS** • **Tags_Closed by Horizzon** • **Tags_Will revert after reading the email**
X Education has a period of 2 months every year during which they hire few 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 will choose a lower threshold value for Conversion Probability. This will ensure the Sensitivity rating is very high which in turn will make sure almost all leads that are likely to Convert are identified correctly and the agents can make phone calls to as much of such people as possible.**
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 will choose a higher threshold value for Conversion Probability. This will ensure the Specificity rating is very high, which in turn will make sure almost all leads that are on the brink of the probability of getting Converted or not are not selected. As a result the agents won't have to make unnecessary phone calls and can focus on some new work.**