

# Leads Scoring Case Study

To create a Logistic Regression Model to determine whether an education company's lead for online courses the conversion of designated X Education would be successful or not.

By:-

Vasvi Bhangalia

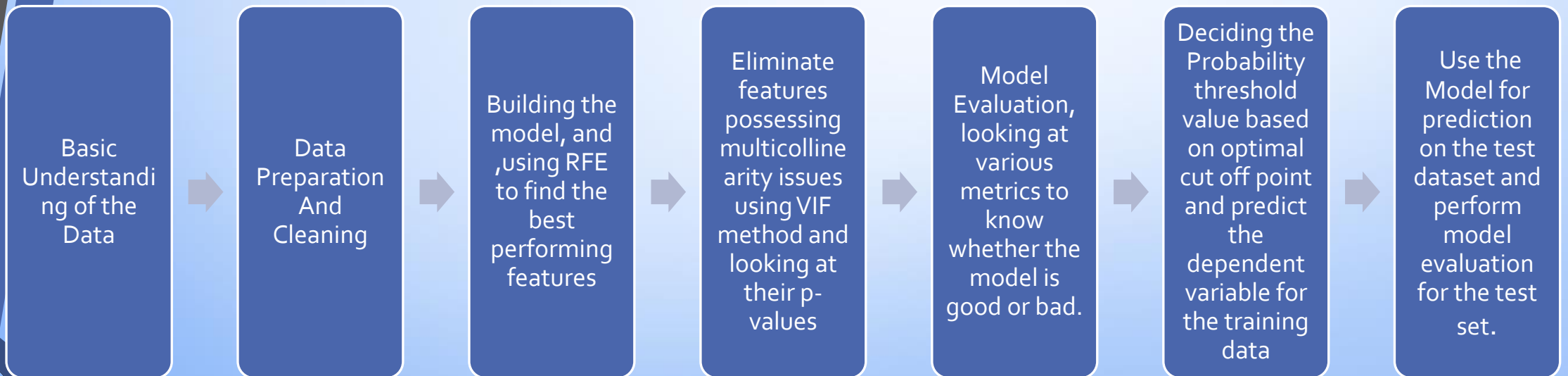
Ashish Prasad Maharana

# Sub Objectives

- A Logistic Regression model which predicts the conversion probabilities of each lead.
- Decide a threshold value, which denotes whether a lead probabilistic value will be classified as converted or not, depending whether the lead probabilistic value is greater than or lower than threshold value.
- Multiply the Probability of each respective lead to get a Lead Score value

**Assumption:** We have created the model assuming that tags and last activity of the lead will not be provided in the future

# Problem Solving Methodology



# Data Preparation & Feature Engineering

## Deleting Redundant columns

- Columns with Single Value, i.e **Receive More Updates About Our Courses**(only no),**Update me on Supply Chain Content**,**I agree to pay the amount through cheque**,**Magazine**,**Get updates on DM Content** were deleted because they contained only one value and not help in predicting cases.

## Removing Rows where columns have high missing values

- Columns **Asymmetrique Activity Index**, **Asymmetrique Profile Index**, **Asymmetrique Activity Score**, **Asymmetrique Profile Score**, **Lead Quality** were having null values with null values greater than 40% hence were deleted.

## Imputing NULL values with mode values

- **Lead Profile**, **What is your current occupation**, **What matters most to you in choosing a course**, **Country**, **Last Activity**, **Lead Source** where null values are replaced with mode values

## Imputing NULL values with median values

- **TotalVisits**, **Page Views Per Visit** were columns where NULL values are replaced with median values.

## Removing unnecessary columns

- **Country, What matters most to you in choosing a course, Last Notable Activity , Last Activity,Tags** were dropped because they were not giving

## Handling 'Select' values in columns

- **Lead Profile , How did you hear about X Education, Specialization , City** Select values were lumped with NULL values and imputed with values unique to their categories.

## Binary Encoding

- Handling Categorical columns with either Yes/No values 0/1
- A free copy of Mastering The Interview, Through Recommendations, Digital Advertisement , Newspaper , Newspaper Article, X Education Forums , Search , Do Not Email, Do Not Call were converted to 0/1.

## Dummy Encoding

- For the following categorical variables with multiple values, dummy features were created
- **Lead Origin, Lead Source ,TotalVisits, Page Views Per Visit , Last Activity , Specialization, How did you hear about X education,, hat is yourtr current occupation, tags, Lead Profile, City**

## Test- Train Split

- The original dataframe was split into Test 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 the categorical features ) 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.

**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 sets of features 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 how they were eliminated.

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

```
Index(['Do Not Email', 'Total Time Spent on Website', 'Newspaper',
      'Digital Advertisement', 'Lead Source_Facebook',
      'Lead Source_Referral Sites', 'Lead Source_Welingak Website',
      'Specialization_E-COMMERCE', 'Specialization_Hospitality Management',
      'Specialization_IT Projects Management',
      'Specialization_Rural and Agribusiness',
      'How did you hear about X Education_Email',
      'What is your current occupation_Housewife',
      'What is your current occupation_Student',
      'What is your current occupation_Unemployed',
      'What is your current occupation_Working Professional',
      'Lead Profile_Dual Specialization Student',
      'Lead Profile_Lateral Student', 'Lead Profile_Potential Lead',
      'Lead Profile_Student of SomeSchool', 'Lead Origin_Lead Add Form',
      'Lead Origin_Lead Import', 'TotalVisits_1-2_Visits',
      'TotalVisits_3-4_Visits',
      'Page Views Per Visit_1-2_Page_views_per_visit',
      'Page Views Per Visit_3-4_Page_views_per_visit',
      'Page Views Per Visit_5-6_Page_views_per_visit',
      'Page Views Per Visit_7-8_Page_views_per_visit',
      'Page Views Per Visit_9-12_Page_views_per_visit',
      'Page Views Per Visit_above-12_Page_views_per_visit'],
      dtype='object')
```



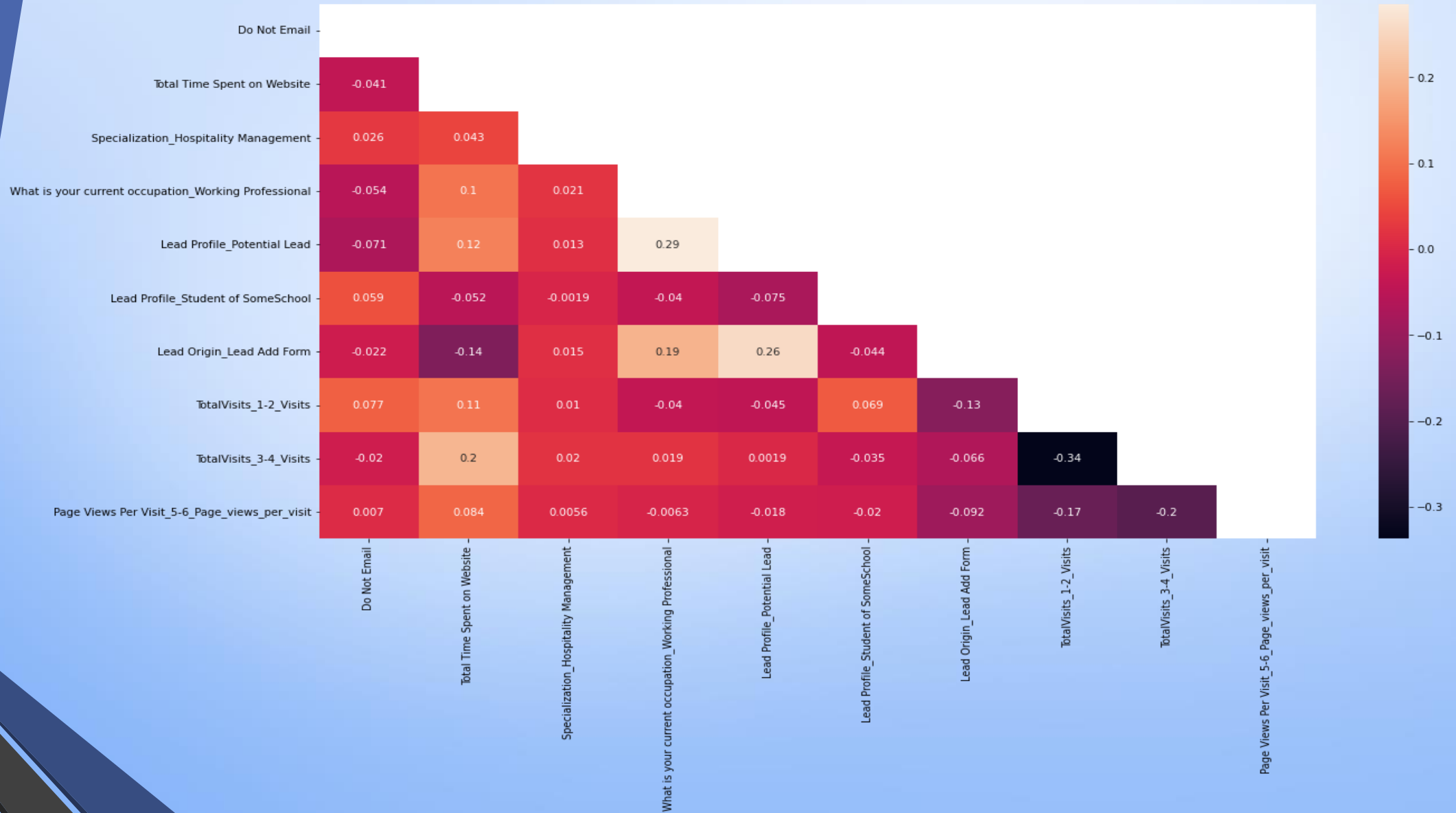
Running RFE with the output number of the variables equal to 30.

- Generalized Linear Models from StatsModel is used to build the Logistic Regression Model
- The Model is built initially with the 30 variables selected by the 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 17 features, passes both the significances test and the multi-collinearity test

	Features	VIF
4	Lead Profile_Potential Lead	1.34
6	Lead Origin_Lead Add Form	1.21
3	What is your current occupation_Working Profes...	1.20
8	TotalVisits_3-4_Visits	1.12
7	TotalVisits_1-2_Visits	1.11
1	Total Time Spent on Website	1.10
0	Do Not Email	1.09
5	Lead Profile_Student of SomeSchool	1.03
9	Page Views Per Visit_5-6_Page_views_per_visit	1.03
2	Specialization_Hospitality Management	1.02



# Heatmap of Final features



# Predicting the Conversion Probability and Predicted Column

Creating a dataframe with the actual Converted Flag and the predicted Probabilities

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

	Converted	Converted_Prob	Lead_Id
0	0	0.164195	1871
1	0	0.212349	6795
2	0	0.193974	3516
3	0	0.582508	8105
4	0	0.164195	3934

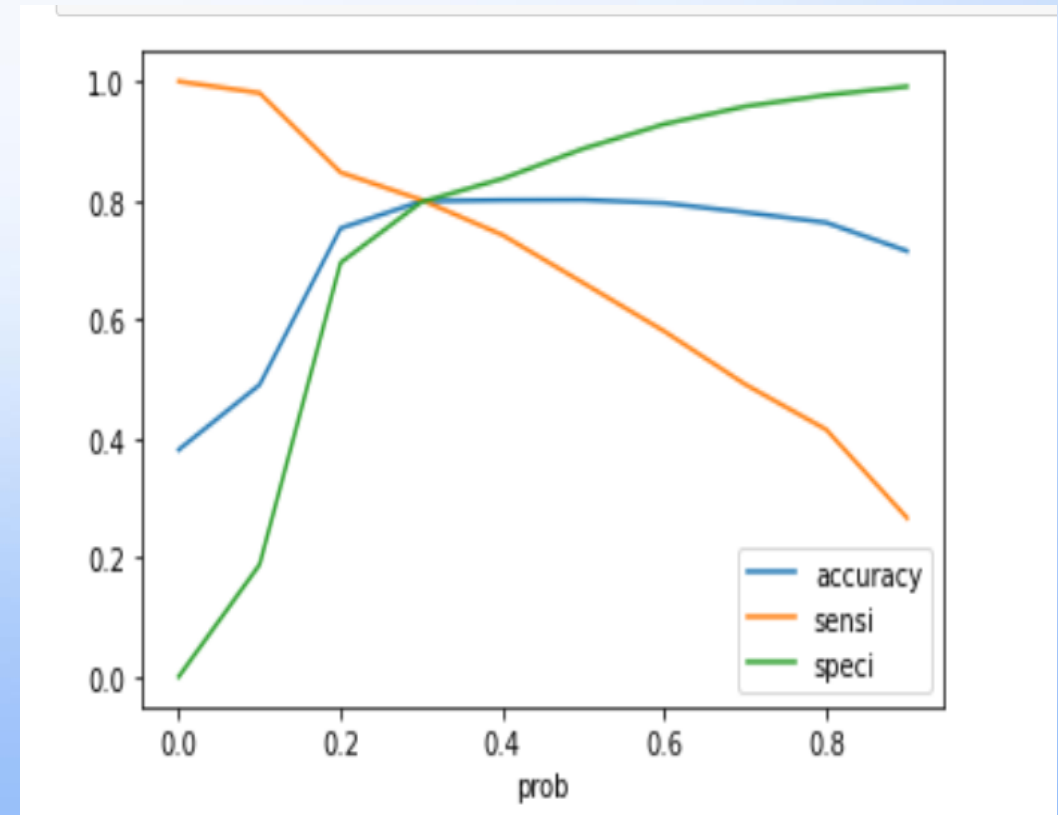
Lead_ID	Converted	Converted_Prob	final_predicted	Lead Score
4269	1	0.576482	1	58
2376	1	0.843123	1	84
7766	1	0.659188	1	66
9199	0	0.495191	1	50
4359	1	0.964076	1	96
...	...	...	...	...
8649	0	0.255004	0	26
2152	1	0.843123	1	84
7101	0	0.164195	0	16
5331	0	0.400582	1	40
2960	1	0.843123	1	84

Creating new column 'Final Predicted' with 1 if Conversion\_Prob>0.3 else 0

Showing all the records of the dataframe in the picture on the left.

# Finding Optimal Probability Threshold

- The accuracy, sensitivity and specificity was calculated for various values of probability threshold and plotted in the graph to the right
- For the curve above 0.3 it is found to be the optimum point for cut off frequency.
- At this threshold value, all the 3 metrics- accuracy, sensitivity and specificity were found above 80% which is a well accepted value



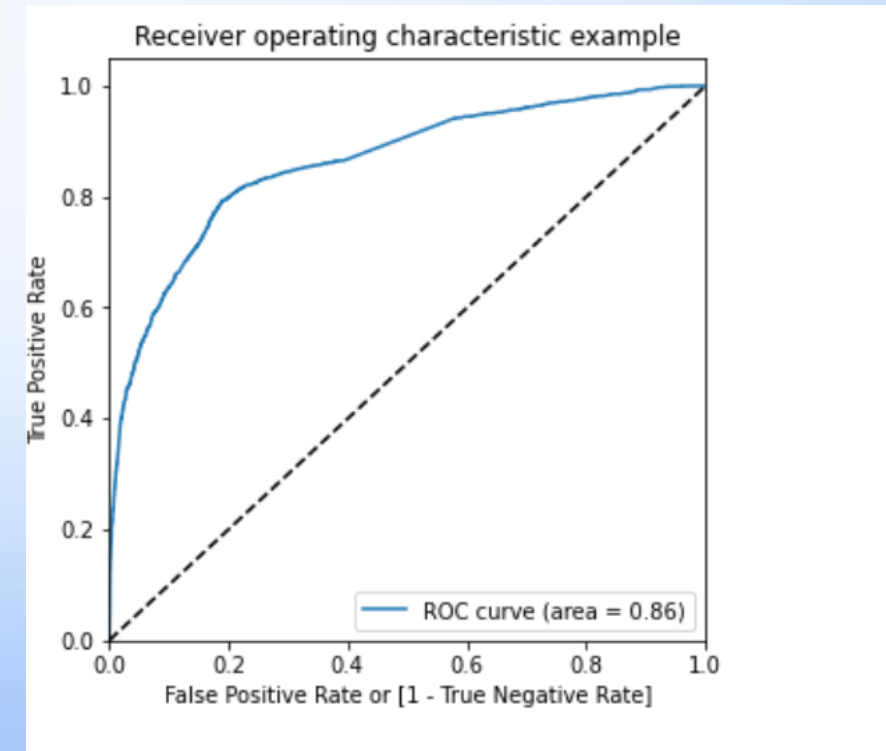
# Plotting the ROC Curve & Calculating AUC

## Receiver Operating Characteristics (ROC) Curve

- It Shows the tradeoff between sensitivity and specificity

## Area under the curve (GINI)

- By determining the area under the curve 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 be the model
- The value of AUC for our model is **0.86**



# Evaluating the model on the Train Set

## Confusion Matrix

#Predicted #Actual	Not Converted	Converted
Not Converted	3191	811
Converted	491	1975

Probability  
Threshold  
= 0.33

Accuracy

• 79.87

Sensitivity

• 80.09

Specificity

• 79.74

Precision

• 70.89

Recall

• 80.09

# 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 data frame
- Using the probability threshold value of 0.3, the leads from the test dataset were predicted if they will convert or not.

	Converted	Converted_Prob	final_predicted
Lead_ID			
4269	1	0.576482	1
2376	1	0.843123	1
7766	1	0.659188	1
9199	0	0.495191	1
4359	1	0.964076	1

# Lead Score Calculation

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

Formula for the Lead Score conversion is:

$$\text{Lead Score} = 100 * \text{Conversion Probability}$$

Lead_ID	Converted	Converted_Prob	final_predicted	Lead Score
4269	1	0.576482	1	58
2376	1	0.843123	1	84
7766	1	0.659188	1	66
9199	0	0.495191	1	50
4359	1	0.964076	1	96
...	...	...	...	...

- The train and test data set 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.
- As the threshold value is 0.3. Any lead with a lead score of 30 or above will have a value of 1 in the final predicted column.

# Deciding Feature Importance

- A total of 10 features has been used by our model to successfully predict if a lead will get converted or not.
- The coefficient 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 values contribute the least.

	coef
const	-0.7009
Do Not Email	-1.2385
Total Time Spent on Website	1.0464
Specialization_Hospitality Management	-0.9593
What is your current occupation_Working Professional	2.5686
Lead Profile_Potential Lead	1.6081
Lead Profile_Student of SomeSchool	-1.9887
Lead Origin_Lead Add Form	3.3090
TotalVisits_1-2_Visits	-0.8588
TotalVisits_3-4_Visits	-0.6159
Page Views Per Visit_5-6_Page_views_per_visit	-0.3946



# Inference

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

1. The conversion probability of a lead increases in the values of the following features:-  
What is your current occupation \_Working Professional, Lead Profile \_ Potential Lead ,  
Lead Origin \_ Lead Add Form
2. The conversion probability of a lead decreases in the values of the following features:-  
Do not email, Specialization\_Hospitality Management, Lead Profile\_student of some  
school,Toatal Vists\_1-2\_Visits, TotalVisits\_3-4\_Visits,Page Views Per Visit\_5-  
6\_Page\_views\_peer\_visit