LEAD SCORING CASE STUDY REPORT

Group Members: Shyamala Rajasekar

Prashant Kumar

Objective :

- To build a logistic regression model to predict whether a lead for online courses for an education company ‘X Education’ would be successfully converted or not.
- 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 divided into following sub-goals:

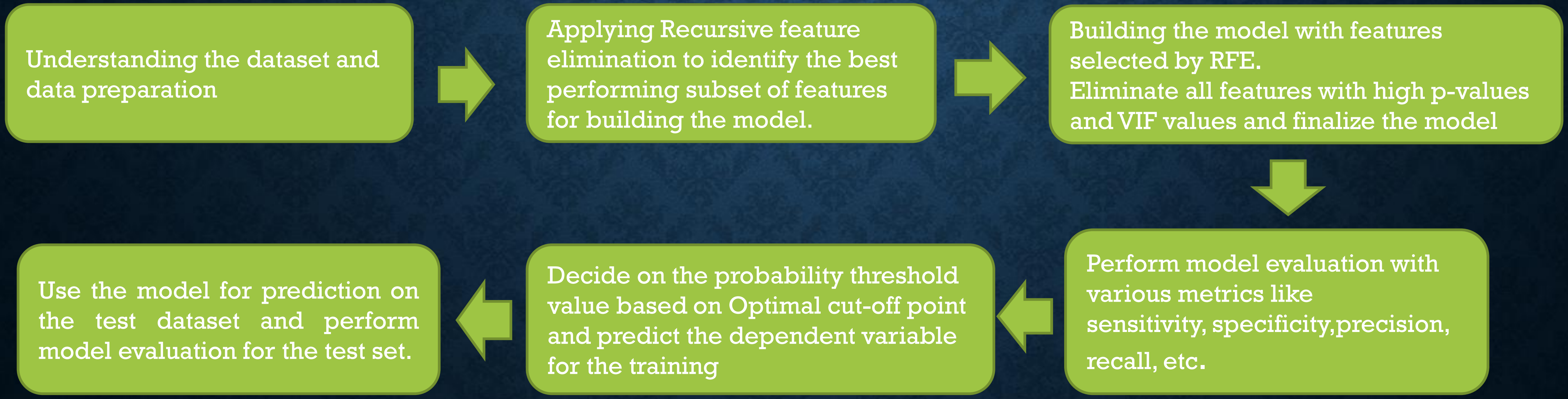
Create a Logistic Regression model to predict the Lead Conversion probabilities for each lead.

Decide on a probability 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:



Understanding the dataset and data preparation

Applying Recursive feature elimination to identify the best performing subset of features for building the model.

Building the model with features selected by RFE.
Eliminate all features with high p-values and VIF values and finalize the model

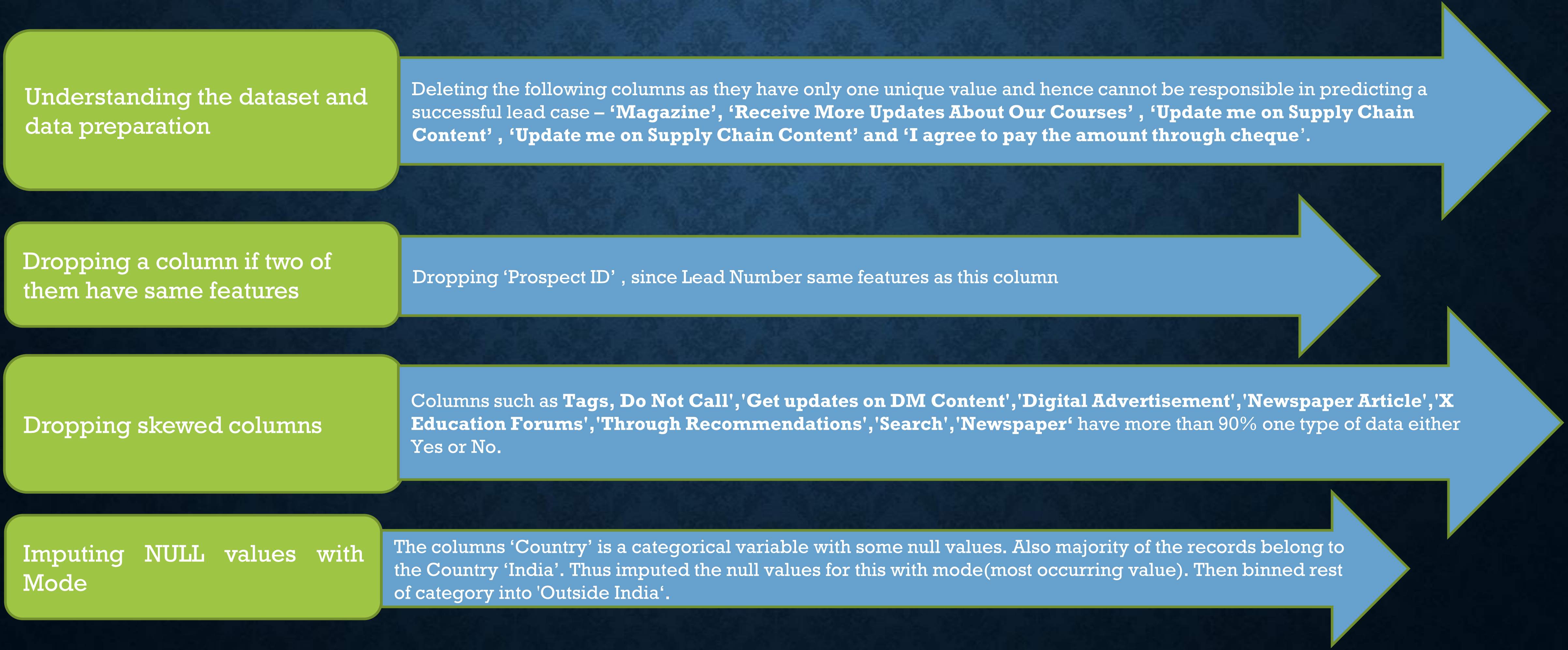
Perform model evaluation with various metrics like sensitivity, specificity,precision, recall, etc.

Decide on the probability threshold value based on Optimal cut-off point and predict the dependent variable for the training

Use the model for prediction on the test dataset and perform model evaluation for the test set.

DATA PREPARATION

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:



Understanding the dataset and data preparation

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**’.

Dropping a column if two of them have same features

Dropping ‘Prospect ID’ , since Lead Number same features as this column

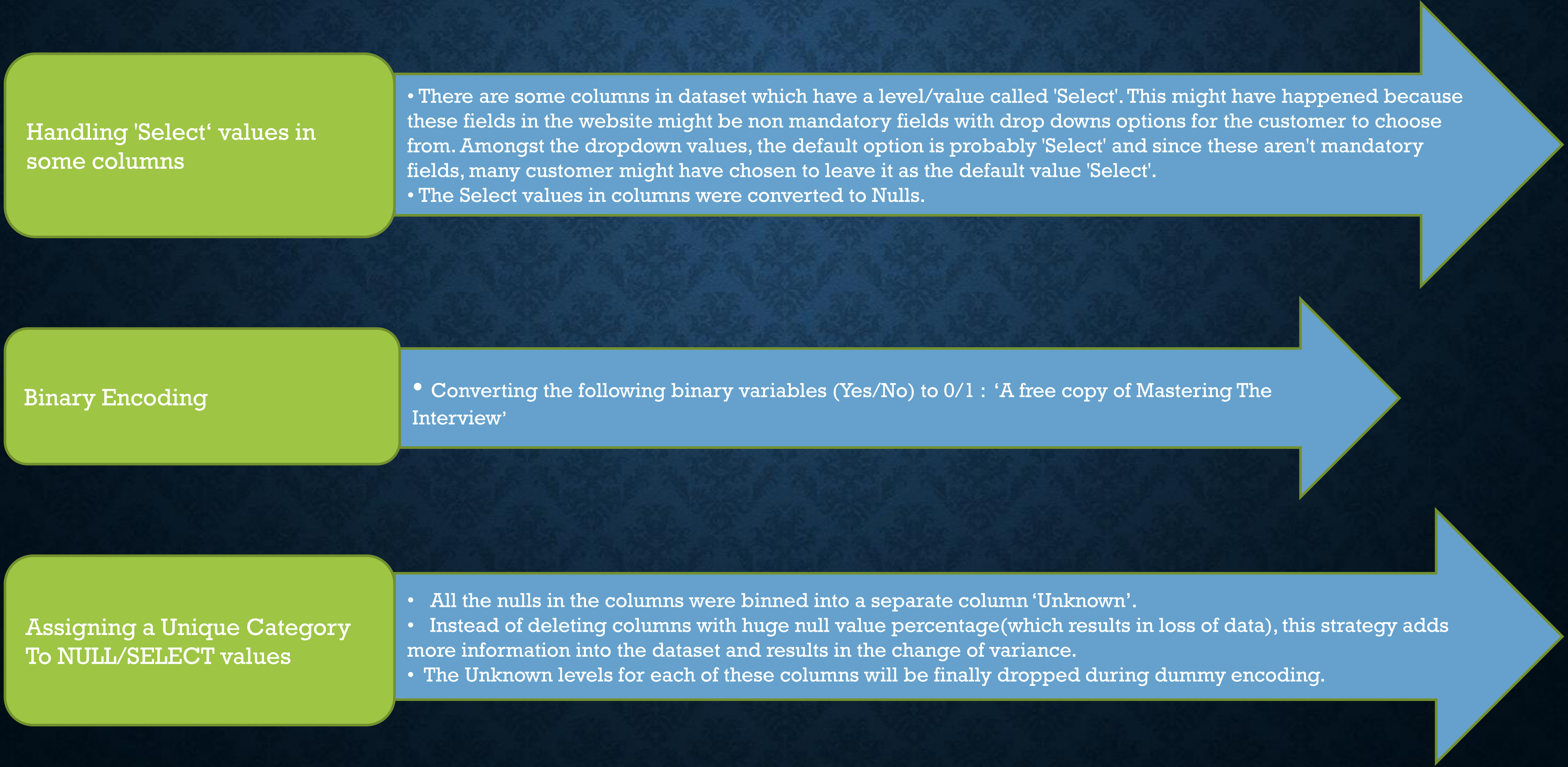
Dropping skewed columns

Columns such as **Tags, Do Not Call','Get updates on DM Content','Digital Advertisement','Newspaper Article','X Education Forums','Through Recommendations','Search','Newspaper‘** have more than 90% one type of data either Yes or No.

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 occurring value). Then binned rest of category into 'Outside India‘.

DATA PREPARATION 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 chosen to leave it as the default value 'Select'.
- The Select values in columns were converted to Nulls.

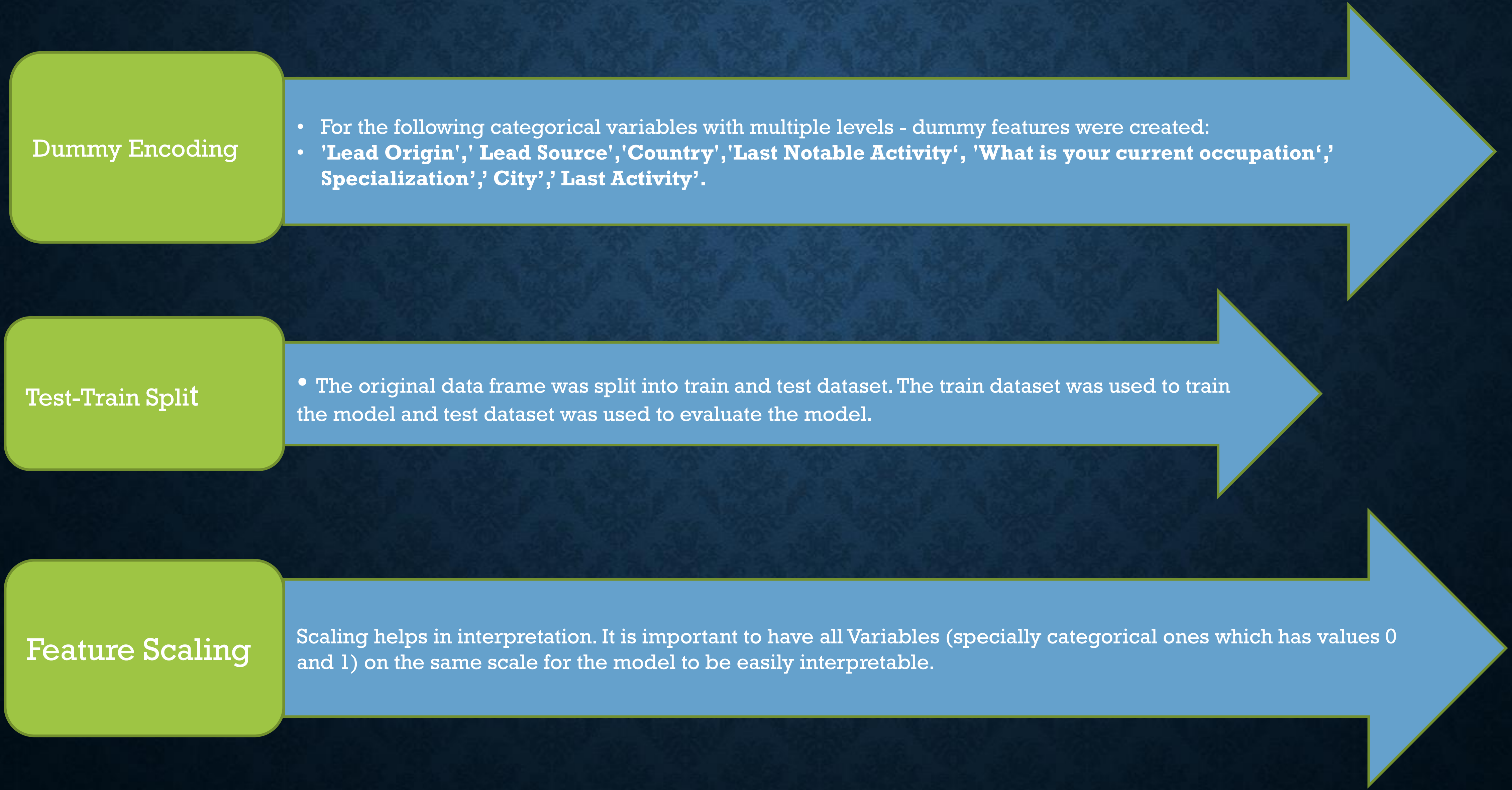
Binary Encoding

- Converting the following binary variables (Yes/No) to 0/1 : ‘A free copy of Mastering The Interview’

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.

DATA PREPARATION CONTD…



Dummy Encoding

- For the following categorical variables with multiple levels - dummy features were created:
- **'Lead Origin',' Lead Source','Country','Last Notable Activity‘, 'What is your current occupation‘,' Specialization',' City',' Last Activity'.**

Test-Train Split

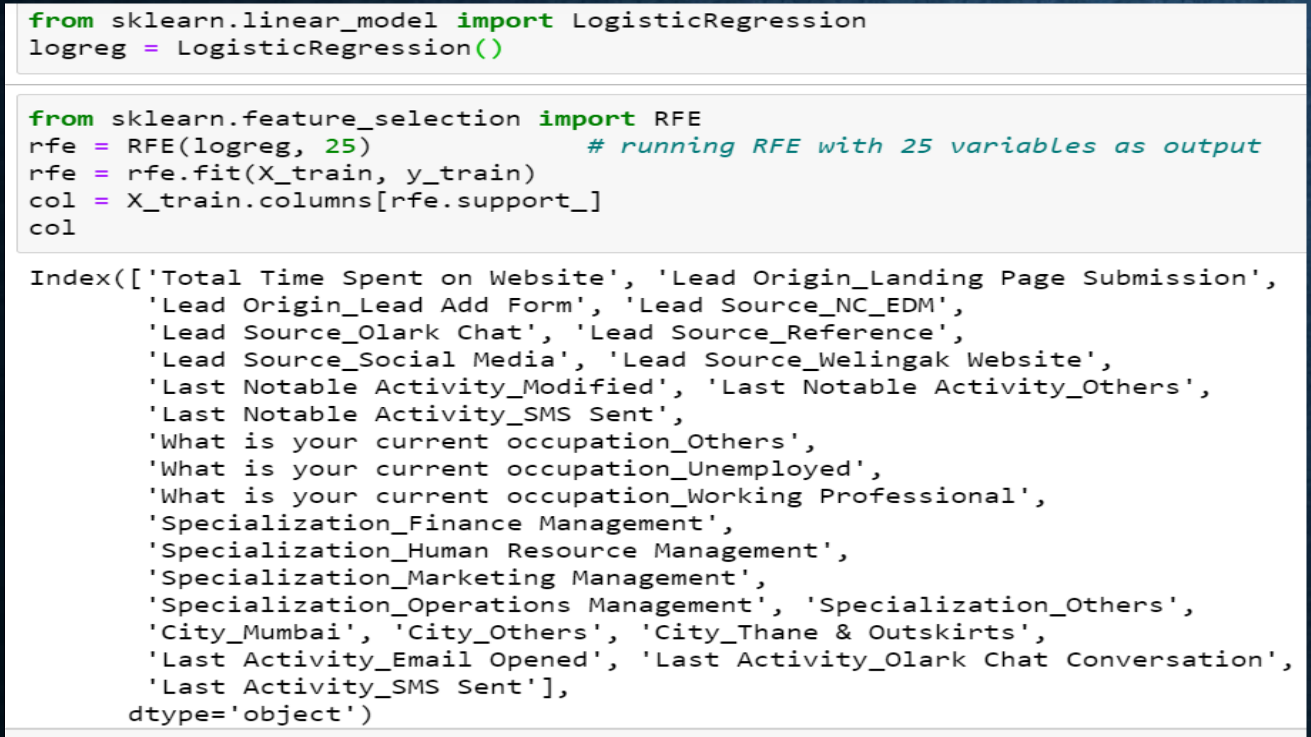
- The original data frame was split into train and test dataset. The train dataset was used to train the model and test dataset was used to evaluate the model.

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.

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()
```

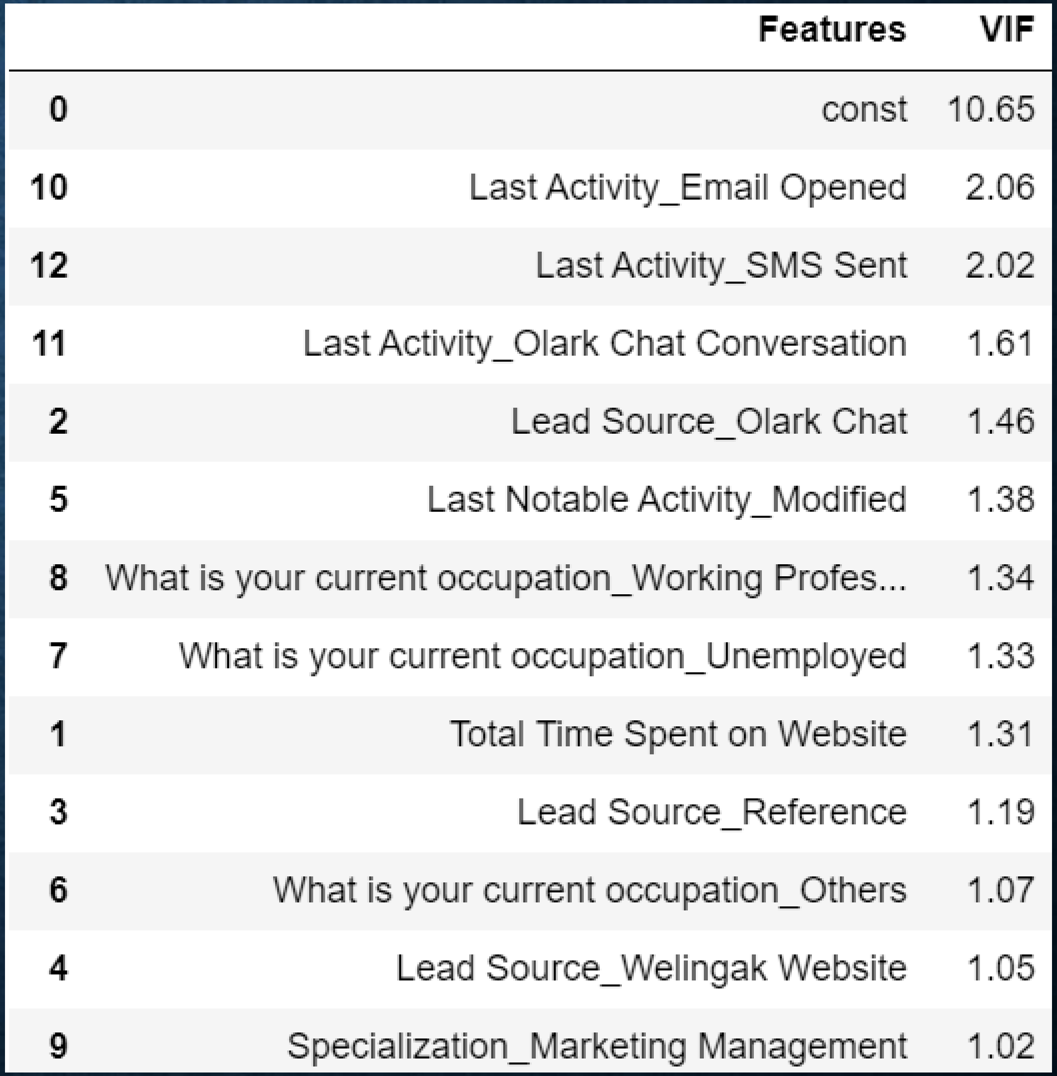
```
from sklearn.feature_selection import RFE
rfe = RFE(logreg, 25)              # running RFE with 25 variables as output
rfe = rfe.fit(X_train, y_train)
col = X_train.columns[rfe.support_]
col
```

```
Index(['Total Time Spent on Website', 'Lead Origin_Landing Page Submission',
       'Lead Origin_Lead Add Form', 'Lead Source_NC_EDM',
       'Lead Source_Olark Chat', 'Lead Source_Reference',
       'Lead Source_Social Media', 'Lead Source_Welingak Website',
       'Last Notable Activity_Modified', 'Last Notable Activity_Others',
       'Last Notable Activity_SMS Sent',
       'What is your current occupation_Others',
       'What is your current occupation_Unemployed',
       'What is your current occupation_Working Professional',
       'Specialization_Finance Management',
       'Specialization_Human Resource Management',
       'Specialization_Marketing Management',
       'Specialization_Operations Management', 'Specialization_Others',
       'City_Mumbai', 'City_Others', 'City_Thane & Outskirts',
       'Last Activity_Email Opened', 'Last Activity_Olark Chat Conversation',
       'Last Activity_SMS Sent'],
      dtype='object')
```

Running RFE with the output number of the variable equal to 25.

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 12 features, passes both the significance test and the multi-collinearity test.

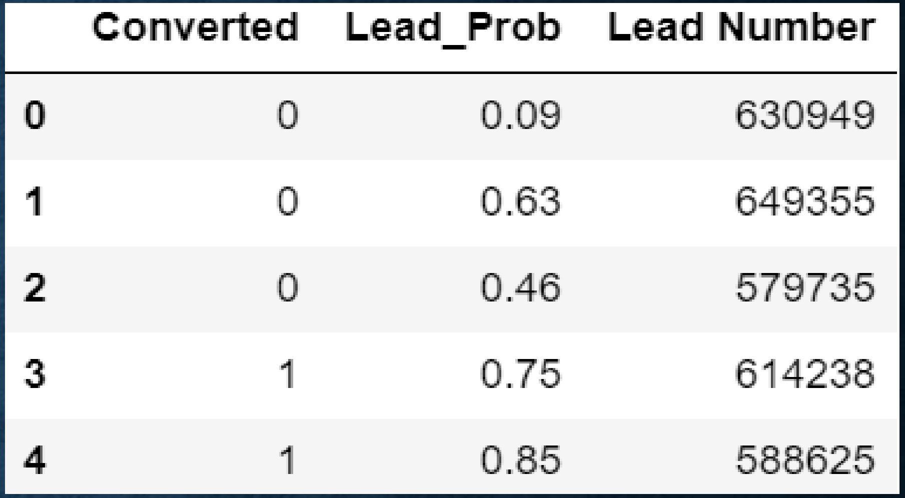


	Features	VIF
0	const	10.65
10	Last Activity_Email Opened	2.06
12	Last Activity_SMS Sent	2.02
11	Last Activity_Olark Chat Conversation	1.61
2	Lead Source_Olark Chat	1.46
5	Last Notable Activity_Modified	1.38
8	What is your current occupation_Working Profes...	1.34
7	What is your current occupation_Unemployed	1.33
1	Total Time Spent on Website	1.31
3	Lead Source_Reference	1.19
6	What is your current occupation_Others	1.07
4	Lead Source_Welingak Website	1.05
9	Specialization_Marketing Management	1.02

PREDICTING THE CONVERSION PROBABILITY AND PREDICTED COLUMN

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

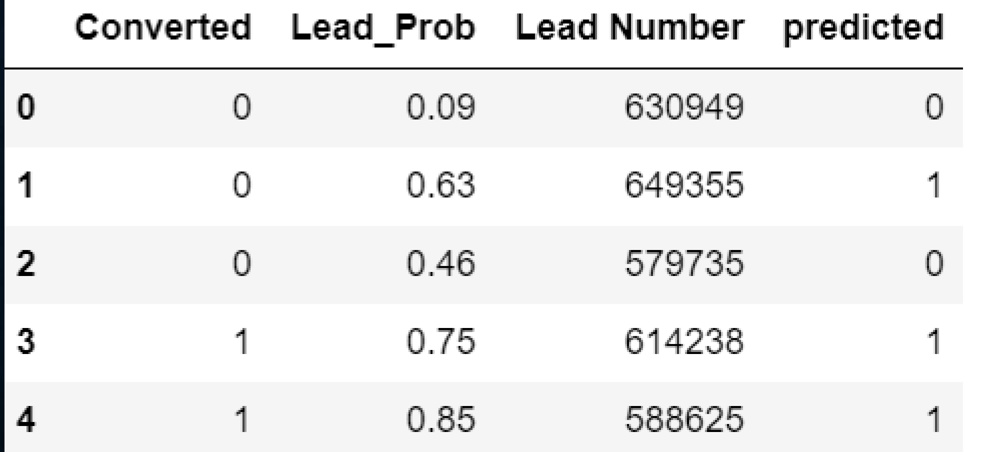
Showing top 5 records of the dataframe in the picture on the right.



	Converted	Lead_Prob	Lead Number
0	0	0.09	630949
1	0	0.63	649355
2	0	0.46	579735
3	1	0.75	614238
4	1	0.85	588625

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.



	Converted	Lead_Prob	Lead Number	predicted
0	0	0.09	630949	0
1	0	0.63	649355	1
2	0	0.46	579735	0
3	1	0.75	614238	1
4	1	0.85	588625	1

FINDING OPTIMAL PROBABILITY THRESHOLD

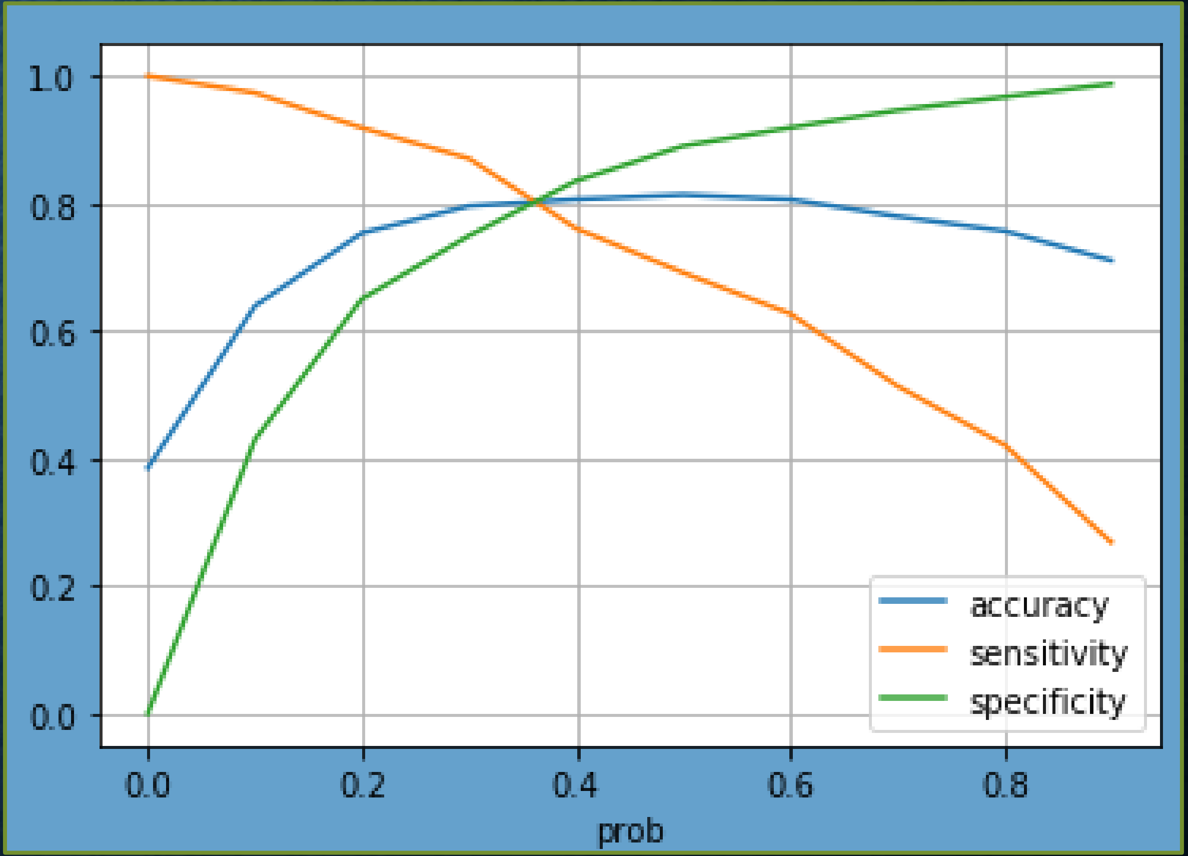
Optimal cutoff probability is that prob. where we get balanced sensitivity and specificity.

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, 0.35 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 about 80% which is a well acceptable value.

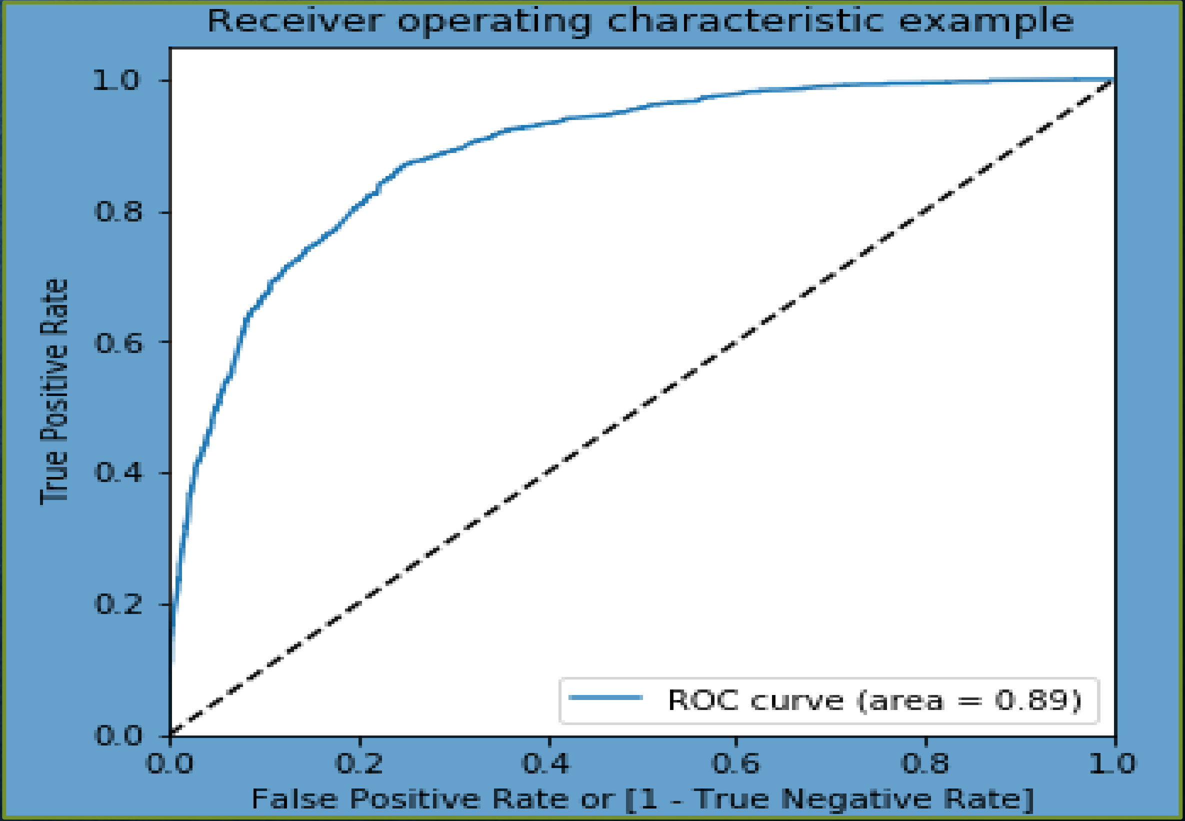


PLOTTING ROC CURVE

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).

- 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 0.89



EVALUATING THE MODEL ON TRAIN DATASET

Confusion Matrix

# Predicted	Not-Converted	Converted
# Actual		
Not-Converted	3061	844
Converted	429	2017

Accuracy
TP +TN/
(TP+TN+FN+FP)
79.95

Sensitivity
TP / (TP+FN)
82.36

Specificity
TN / (TN+FP)
78.38

Precision
TP / (TP + FP)
70.50

Recall
TP / (TP+FN)
82.36

False Positive Rate
FP/ (TN+FP)
21.61

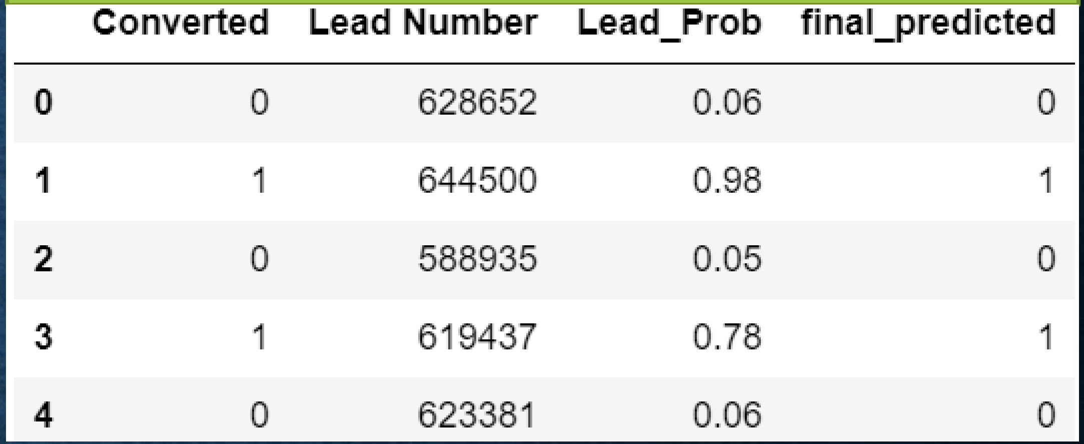
Positive Predictive Value TP / (TP+FP)
70.5

Negative Predictive Value
TN / (TN+ FN)
87.70

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 0.35, the leads from the test dataset were predicted if they will convert or not.

The top 5 records from the final test data set



	Converted	Lead Number	Lead_Prob	final_predicted
0	0	628652	0.06	0
1	1	644500	0.98	1
2	0	588935	0.05	0
3	1	619437	0.78	1
4	0	623381	0.06	0

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

EVALUATING THE MODEL ON TEST DATASET

Confusion Matrix

# Predicted	Not-Converted	Converted
# Actual		
Not-Converted	1351	383
Converted	178	811

Accuracy
TP +TN/
(TP+TN+FN+FP)

79.39

Sensitivity
TP / (TP+FN)

82.00

Specificity
TN / (TN+FP)

77.91

Precision
TP / (TP + FP)

67.92

Recall
TP / (TP+FN)

82.00

False Positive Rate
FP/ (TN+FP)

22.08

Positive Predictive Value TP / (TP+FP)

67.92

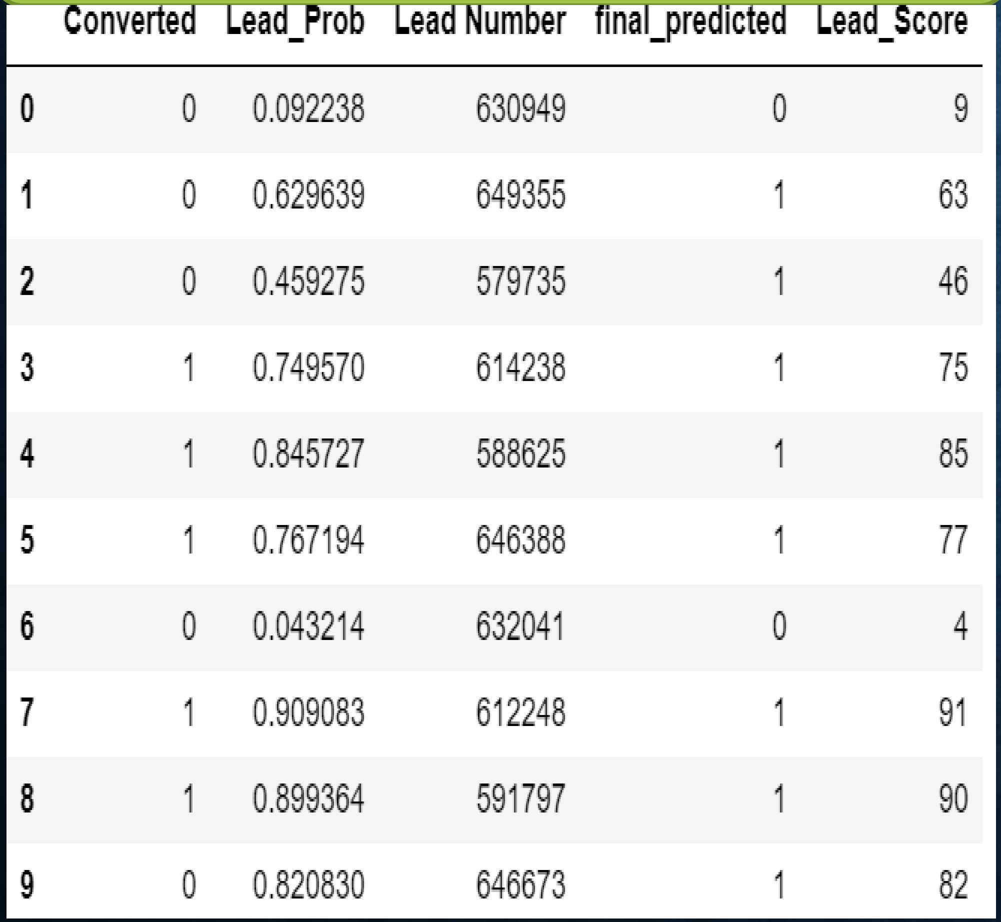
Negative Predictive Value
TN / (TN+ FN)

88.35

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



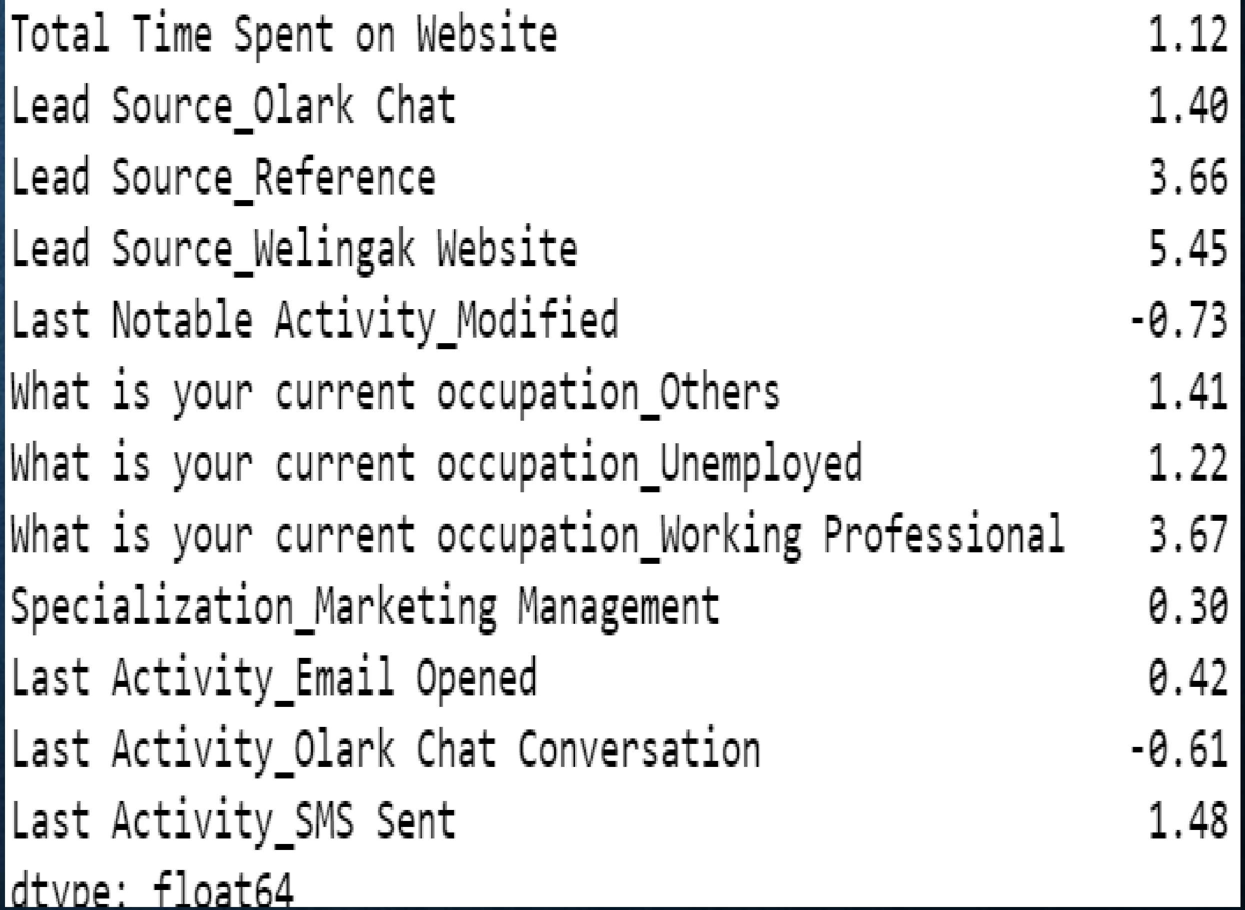
	Converted	Lead_Prob	Lead Number	final_predicted	Lead_Score
0	0	0.092238	630949	0	9
1	0	0.629639	649355	1	63
2	0	0.459275	579735	1	46
3	1	0.749570	614238	1	75
4	1	0.845727	588625	1	85
5	1	0.767194	646388	1	77
6	0	0.043214	632041	0	4
7	1	0.909083	612248	1	91
8	1	0.899364	591797	1	90
9	0	0.820830	646673	1	82

- 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 0.35 as our final Probability threshold for deciding if a lead will convert or not, any lead with a lead score of 36 or above will have a value of '1' in the final_predicted column.

Lead Score for top10 records from dataset.

DETERMINING FEATURE IMPORTANCE

- 25 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.



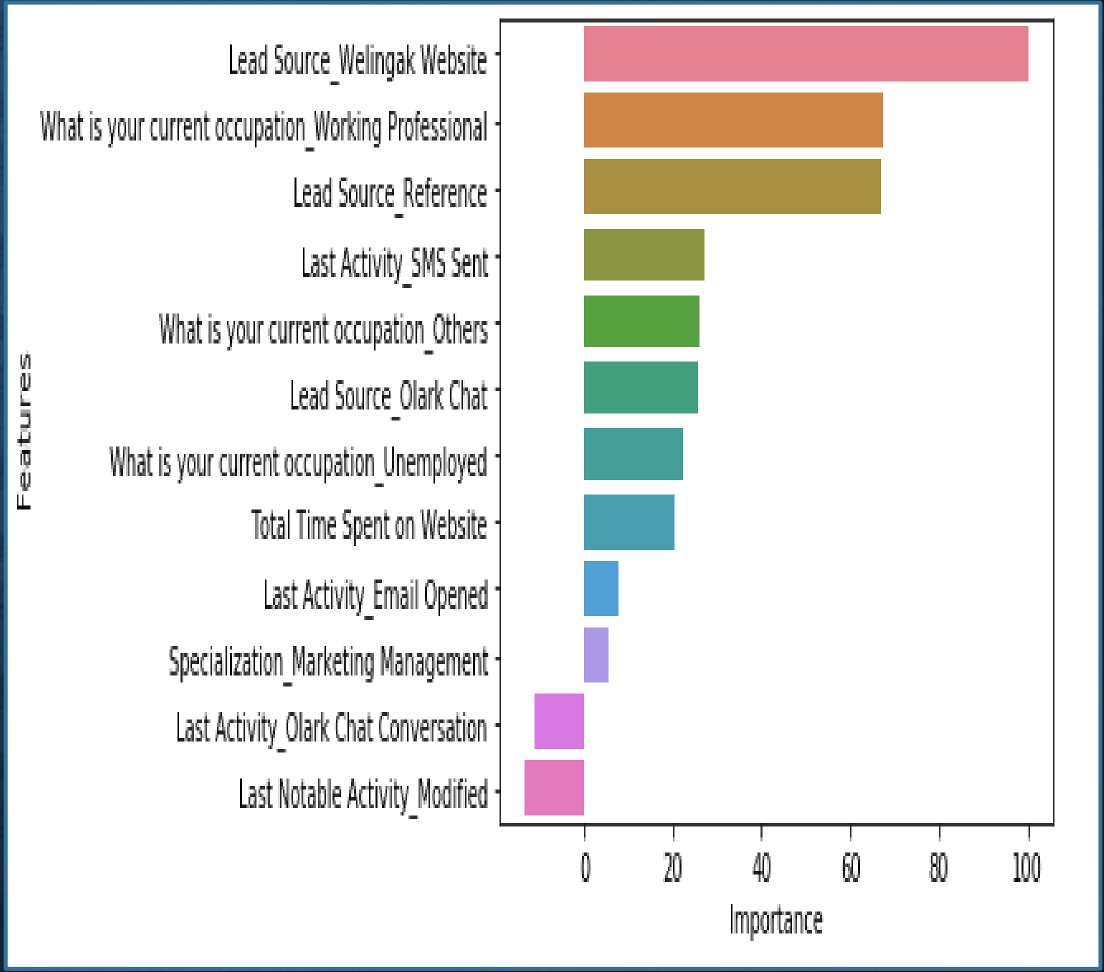
```
Total Time Spent on Website                              1.12
Lead Source_Olark Chat                                   1.40
Lead Source_Reference                                    3.66
Lead Source_Welingak Website                             5.45
Last Notable Activity_Modified                          -0.73
What is your current occupation_Others                   1.41
What is your current occupation_Unemployed               1.22
What is your current occupation_Working Professional     3.67
Specialization_Marketing Management                      0.30
Last Activity_Email Opened                               0.42
Last Activity_Olark Chat Conversation                   -0.61
Last Activity_SMS Sent                                   1.48
dtype: float64
```

DETERMINING FEATURE IMPORTANCE CNTD…

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



CONCLUSION

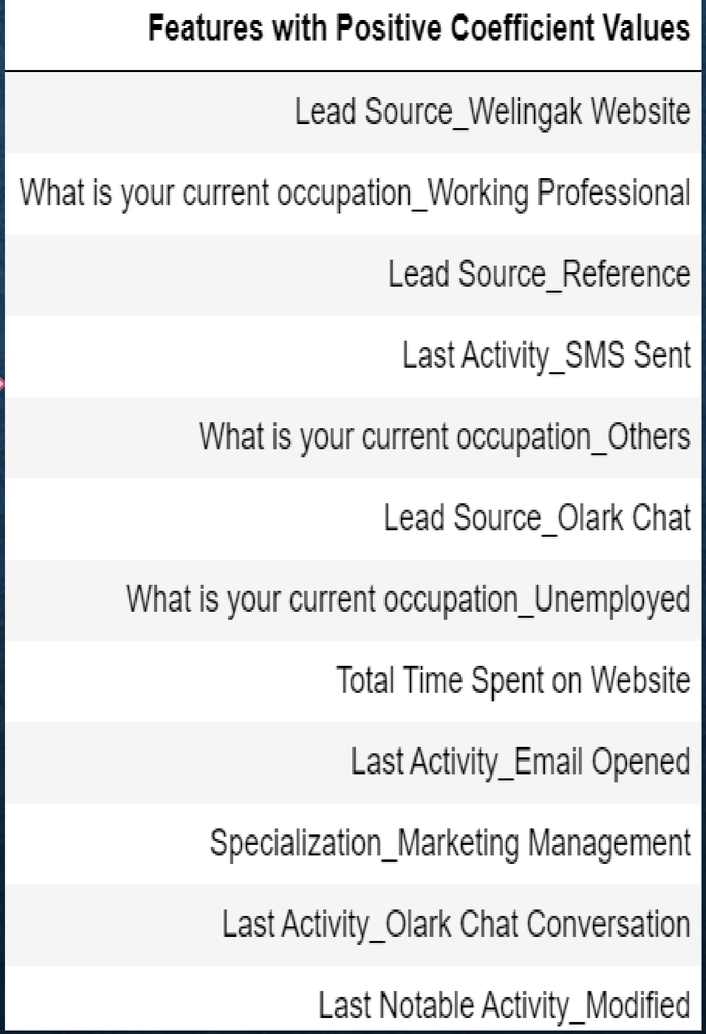
After trying several models, we finally chose a model with the following characteristics:

- All variables have p-value less than 0.05
- All the features have very low VIF values, meaning, there is hardly any multi-collinearity among the features. This is also evident from the heat map.
- The overall accuracy of 79.39% at a probability threshold of 0.35 on the test dataset is also very acceptable.

CONCLUSION CONTD….

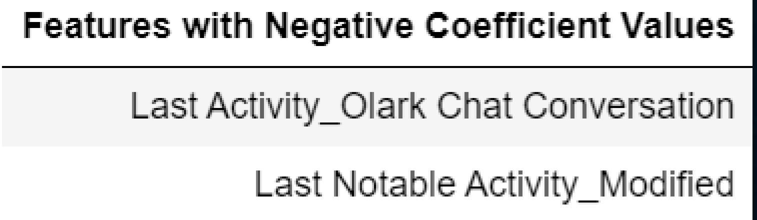
Based on our model, some features are identified which contribute most to a Lead getting converted successfully

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



Features with Positive Coefficient Values
Lead Source_Welingak Website
What is your current occupation_Working Professional
Lead Source_Reference
Last Activity_SMS Sent
What is your current occupation_Others
Lead Source_Olark Chat
What is your current occupation_Unemployed
Total Time Spent on Website
Last Activity_Email Opened
Specialization_Marketing Management
Last Activity_Olark Chat Conversation
Last Notable Activity_Modified

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



Features with Negative Coefficient Values
Last Activity_Olark Chat Conversation
Last Notable Activity_Modified

Recommendations to X Education:

1. Leads with Lead Source as reference to be contacted frequently.

2. Leads from reference are low as compared to others such as Google but have high conversion rate. Hence, efforts must be put in to generate more leads through reference. Those who provide reference can be given incentives. This could generate more leads through reference

3. Leads from Welingak website have high rate of conversion but welingak websites has low number of leads overall. Efforts must be put to generate more leads through Welingak website. This could be through blogs, videos etc. on Welingak Website.

4. Target working professionals as working professional category has a high rate of conversion.

5. Minimize or curb all efforts on Olark Chat conversation as it has a negative correlation on conversion.