Problem Statement

The company X-Education 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%.

Approach

1-Data Cleaning-

Data frame provided to us were having lots of null values.We started data cleaning activity by dropping columns having more that 70% null values. Many categorical variables have level 'Select' which could be due to not mandatory field to be filled by customer while filling the form we replaced 'Select' by nan values to be filled later on.

For categorical columns we plotted bar plots and imputed most frequent values in few columns and for other columns we created levels like 'Not sure' because columns were having lots of null values but since columns were important from business point of view we didn't dropped them.

For numerical columns we imputed null values with median of columns

2-EDA:

We have done univariate analysis of various columns by plotting bar graphs with respect to column 'Convert' on Y-axis to get an idea about how lead conversion is happening as per different attributes.

Few columns has values entered for only 1 field in them and has no variance and few columns has high variance only for 1 field so we dropped these columns as they were not providing much information about lead conversion

3-Data Preparation:

Following activities were as part of data preparation:

- Encoding categorical features
- Rescaling the features
- Splitting the Data into Training and Testing Sets

4-Training the model

We created Logistic regression model by selecting features through RFE. We started with 15 features and after running multiples iterations and dropping variables by checking the VIF score and p-value we got a model having 13 features with vif score less than 5% and p-values less than 0.05.

5- Confusion Matrix calculation with cutoff 0.5

At arbitrary cutoff of 0.5 of threshold probability we predicted probability of conversion.We calculated confusion matrix and it is giving us accuracy of 91%.

6-Optimal Cutoff-

Plotted roc curve and got optimal cutoff of threshold probability of 0.3 to predict final probability.

7- Confusion Matrix calculation with cutoff 0.5

We calculated confusion matrix at optimal cutoff of 0.3 and it gave us following values of different parameters:

```
Logistic regression model Accuracy -  0.9125359499238708

Logistic regression model Precision -  0.8589527027027027

Logistic regression model Recall -  0.9174560216508796

Logistic regression model Sensitivity -   0.9174560216508796

Logistic regression model Specificity -   0.9095831077422848

False positive of model-predicting converted when customer has not converted
-   0.09041689225771521

Logistic regression model Predictive -   0.8589527027027027
```

8- Precision and Recall:

We got 0.42 the tradeoff between Precision and Recall.

Thus we can safely choose to consider any Lead with Conversion Probability higher than 42 % to be a hot Lead

9-Making Predictions on test set Test

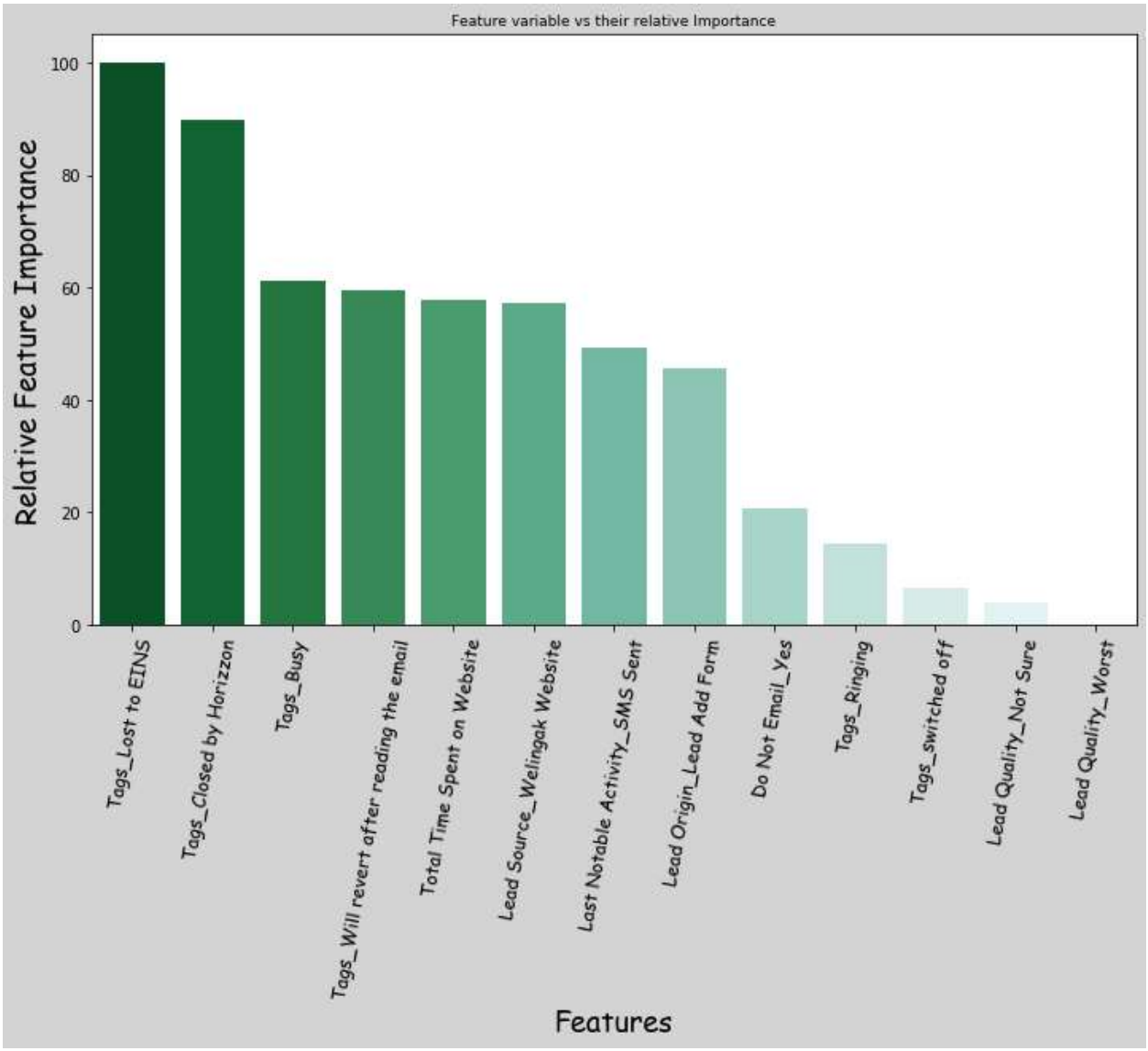
We made prediction on test data set with 0.42 as the cutoff probability and calculated confusion matrix which gave following values for different parameters:

```
Model Accuracy on Test data is  0.9163378058405682
```

Sensitivity of the model on test data is 0.89
Specificity of the model on test data is 0.9293820933165196
Positive predictive 0.8833333333333333
Negative predictive value 0.9364675984752223

Precision values comes out as 0.8833333333333333
Recall values comes out as 0.8945147679324894

10-Features of our final model:



Our model performed well and achived desired outcome of more than 80% Accuracy as mentioned by CEO of X-Education