Lead Score Case Study – VIRAL SHAH

Presentation to Executive, chief data science officer.

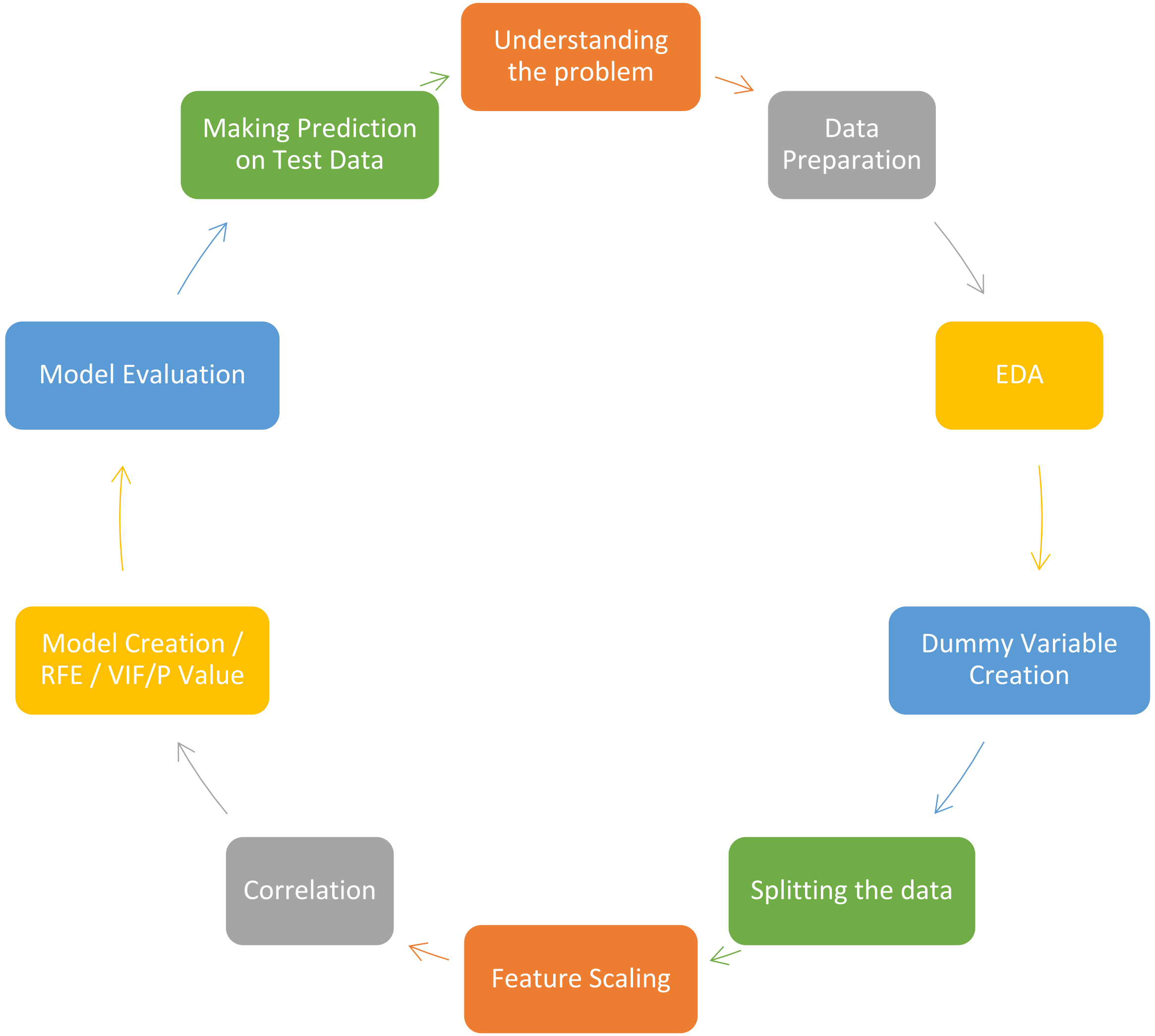
- **Problem Statement:** We were given a problem where we had to build a model where the customer with higher lead score have higher chances of conversion. And lower lead score had lower chances of conversion. This would help the sales team to make a right call to the hot leads.

- **Business Objective:** The X education company wants to have a model where they can generate higher lead to have higher conversion.

Lead Score Case Study – VIRAL SHAH

The Process

Overall Process



Understanding the problem

Data Preparation

EDA

Dummy Variable Creation

Splitting the data

Feature Scaling

Correlation

Model Creation / RFE / VIF/P Value

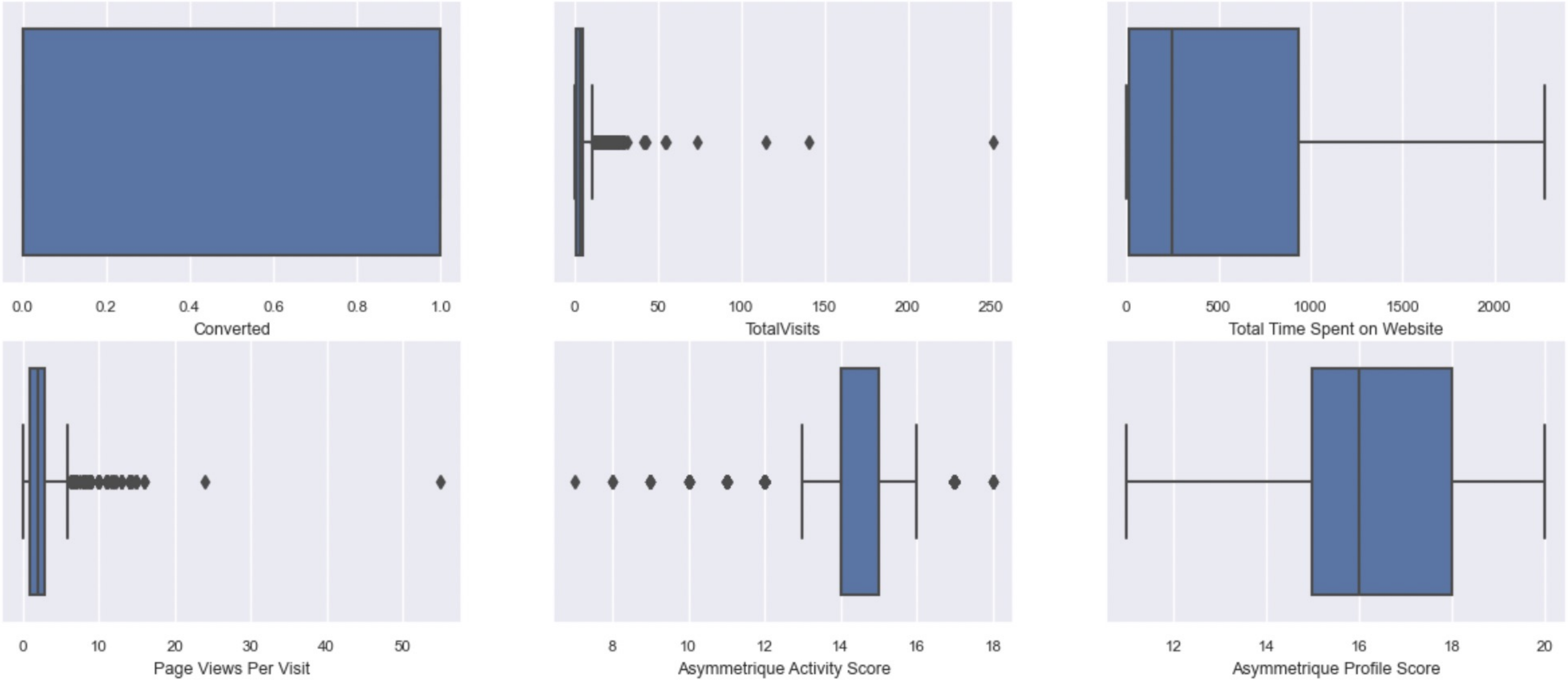
Model Evaluation

Making Prediction on Test Data

Lead Score Case Study – VIRAL SHAH

Identifying Outliers

- **Process / Data Analysis / Model Building / Model Evaluation:** To start the analysis we first started with box plots of Numeric data variables to understand the outliers in the data.

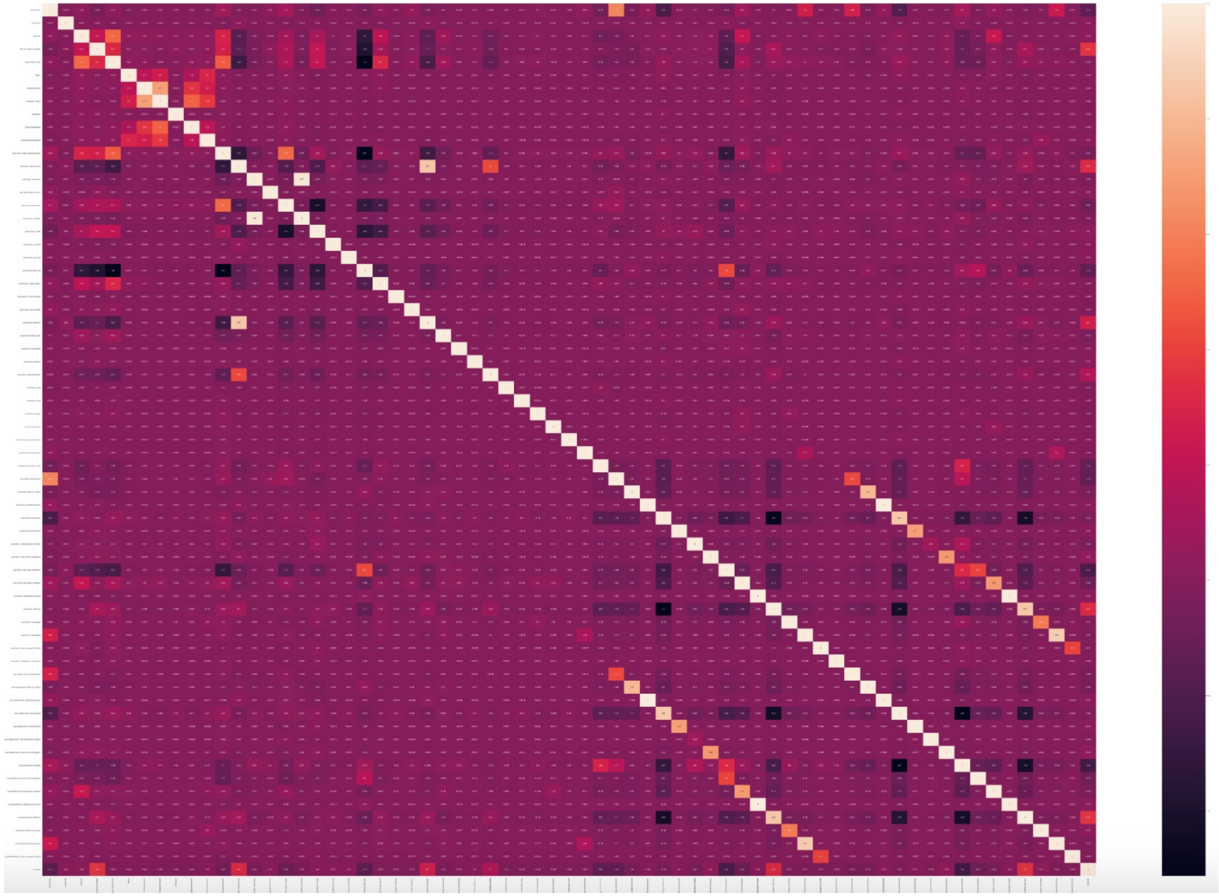


Lead Score Case Study – VIRAL SHAH

Correlations Plot

- **Process / Data Analysis / Model Building / Model Evaluation:**

  1. To begin with there were in total 37 variables with 9240 rows of data. We had a given condition to remove 'Select' as 'NaN' as they don't serve any purpose. We had address 'Select' scenario in columns like : Specialization, how did you hear about X Education, Lead Profile, City.
  2. The second problem that was address was the missing values. We had ensured that columns which had more than 25% of missing values are removed, so it doesn't impact our model building analysis.
  3. The second scenario which was checked is the outliers, we analysed it's impact and realized that we can retain the outliers in the data.
  4. The categorical variables were there. We implemented one-hot encoding and obtain the dummy variables
  5. Correlation analysis was performed and the plot of the same is displayed here.

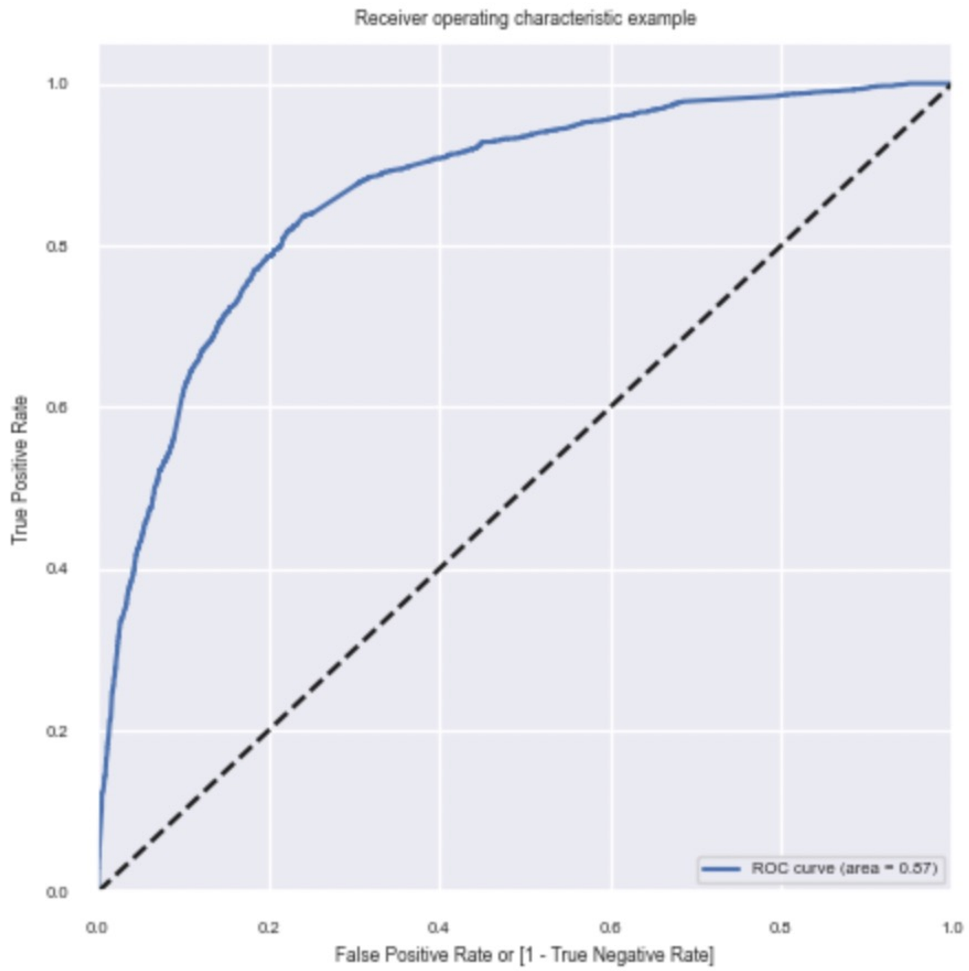


Lead Score Case Study – VIRAL SHAH

Overall Process and ROC Curve

- **Process / Data Analysis / Model Building / Model Evaluation:**

1. To analyze the data better, we had split the data into 70-30 and performed the standardscaler () function
2. Recursive feature Elimination was performed on the data, and the top 15 variables were chosen.
3. Model iterations were then performed keeping in mind the p-value and Variable Inflation Score. The threshold kept in mind was that p-value is supposed to be below 0.05 and the VIF score should not more than 10.
4. Overall Accuracy of 80% was achieved on the training data set, post several model iterations.
5. The ROC curve was then plotted to analyse the performance of the chosen model

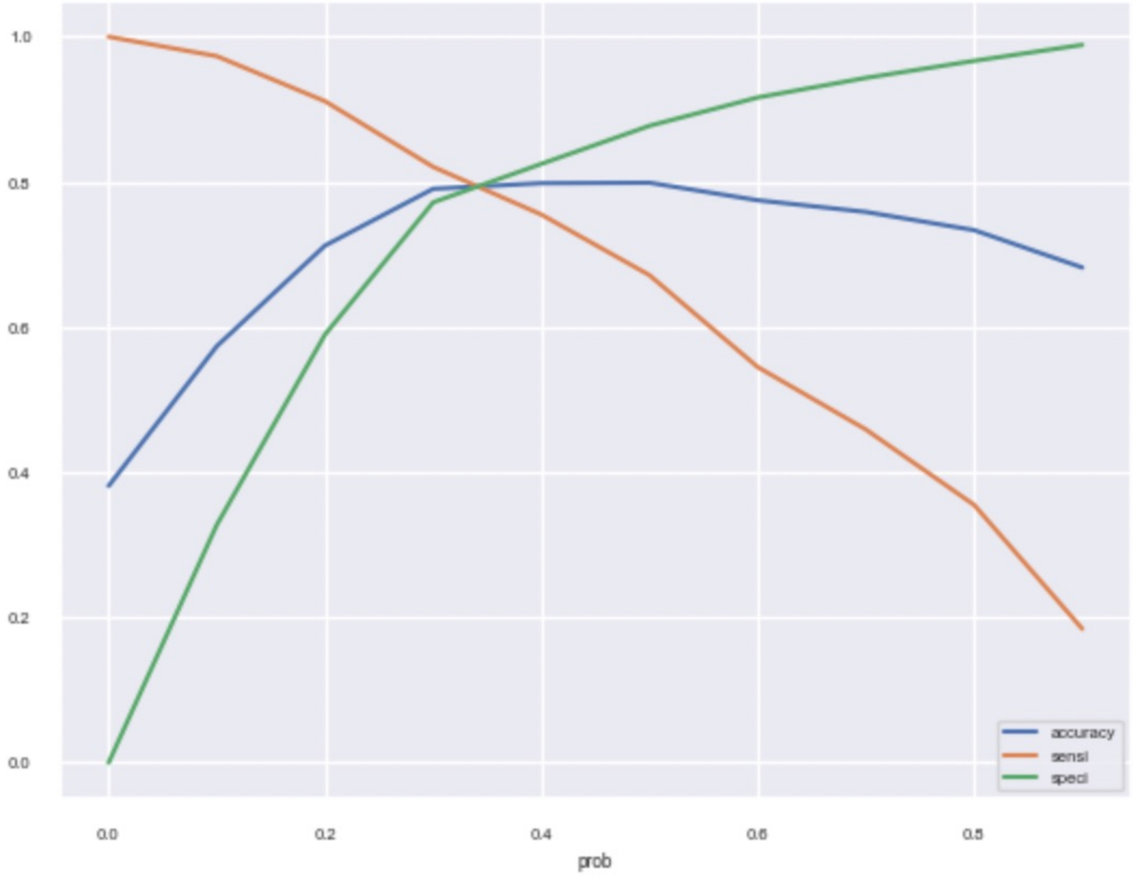


Lead Score Case Study – VIRAL SHAH

Sensitivity and Specificity

- **Process / Data Analysis / Model Building / Model Evaluation:**

  1. Sensitivity-Specificity analysis was performed to obtain an optimal probability of 0.3



Lead Score Case Study – VIRAL SHAH

Precision Vs Recall

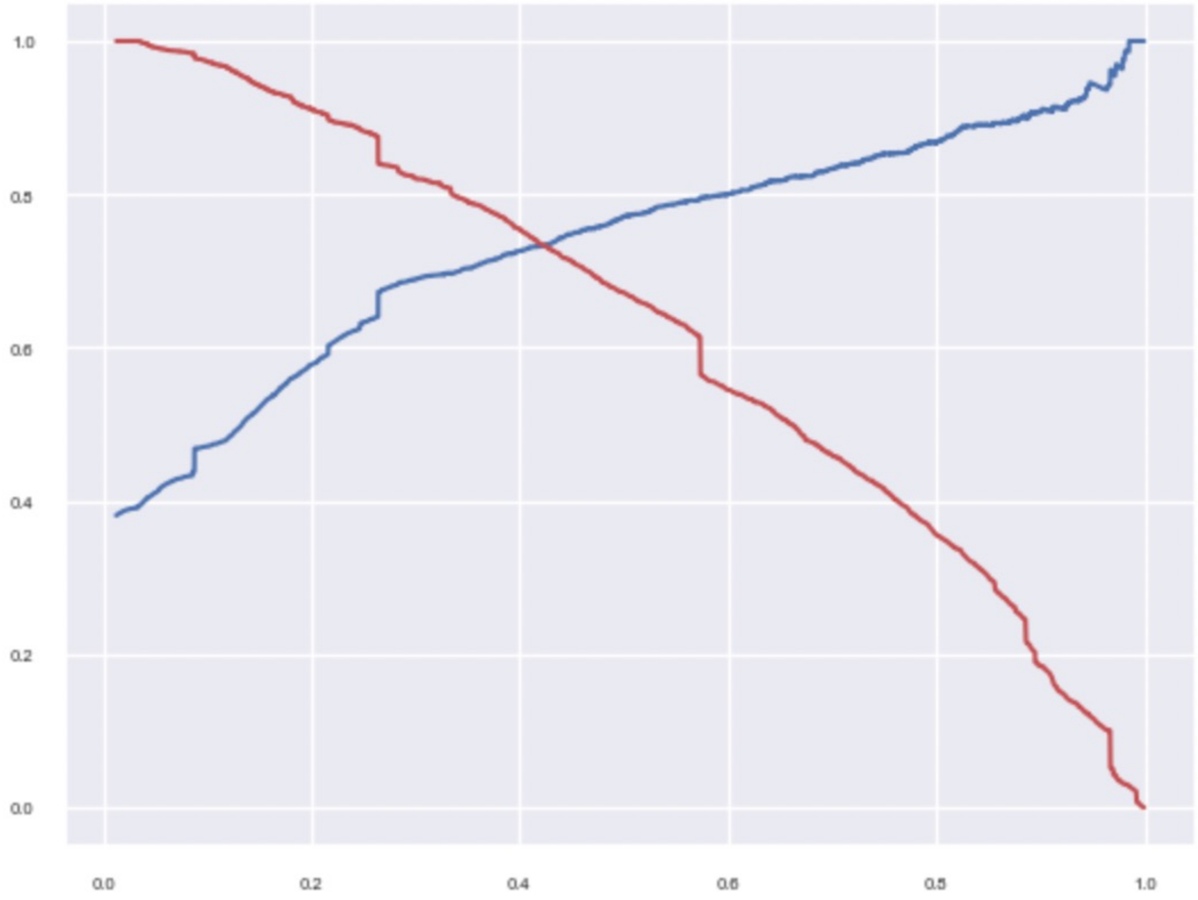
- **Process / Data Analysis / Model Building / Model Evaluation:**

1. The Precision-Recall trade-off was analysed via a plot as well to ensure that there was no bias in the mode
2. From this, we understand that the optimal threshold would be between 0.3 and 0.4.
3. The model was tested on the test data to obtain the following evaluation metrics (Test Data):

o Accuracy:**80.41%**

o Sensitivity:**74.79%**

o Specificity:**84.07%**



Lead Score Case Study – VIRAL SHAH

Conclusion

Hence, a successful model-building exercise was carried out and the sales representative shall be able to understand the leads to target to achieve 80% conversion for the leads pursued. If more leads are to be captured with lesser conversion rates, the threshold can be reduced and vice-versa.

- **Accuracy:80.41%**
- **Sensitivity:74.79%**
- **Specificity:84.07%**

Lead Score Case Study – VIRAL SHAH

Thank You