Lead Scoring Case Study

Submitted By

1. Krishna Mohan Panguluru
2. Karthik Mishra
3. Komal Votavat

Problem Statement

- This problem statement belongs to an education company named X Education sells online courses to industry professionals.

- Many professionals who are interested in the courses land on their website and browse for courses. The company markets its courses on several websites and search engines like Google.

- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.

- Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Problem Statement

- X Education needs to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company needs to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- The CEO, has given a ballpark of the target lead conversion rate to be around 80%.

Assumptions

- Handling the ‘select’ levels that is present in many of the categorical variables by replacing the ‘select’ values with NaN.
- Treating missing values of both Numeric and categorical columns based on NaN values percentage , EDA and dropping it.
- Checking the number of unique category levels in each categorical columns

Approach & Methodology

- **The following approach is followed to perform Lead Score Case study using Logistic Regression :**

1. Understanding the data domain /Variables by referring to the Data Dictionary.
2. We have been given a Leads dataset to perform the analysis.
3. The analysis is performed by importing Leads dataset.
4. Data Cleaning is performed by analyzing(EDA) and then dropping the features that are not reliable. This is performed based on the following steps:
    - Handling the ‘select’ level that is present in many of the categorical variables by replacing with ‘NaN’ values
    - Dropping columns that are having high percentage of missing values by checking all the columns before dropping them
    - Dealing with the unique categories in the categorical features

Approach & Methodology

5. Data preparation is performed on the following steps:

   - Creating Dummy variables for categorical features
   - Perform train – test split (70 : 30)
   - Scaling the Numeric features using ‘StandardScaler’

6. Building Logistic Regression Model by performing the following techniques on Train data :

   - Initiating the first logistic Model Using the ‘GLM’ method of ‘statsmodel’ library
   - Automatic Feature Selection using RFE (Recursive Feature Elimination)
   - The model is being iterated until the features coefficients are significant based on the P-value < 0.05 and VIF < 5

Approach & Methodology

7. Performing Evaluation metrices on the train data :
   - Confusion Matrix and Accuracy
   - Sensitivity and Specificity
   - Plotting ROC curve
   - Precision and Recall

8. Predicting on the Test data
9. Evaluation Metrices on the Test data

Approach & Methodology

Data Sourcing & Cleaning

1. Importing the data and read the source file.
2. Convert data into clean format suitable for analysis.
3. Remove duplicate data
4. Exploratory Data analysis
5. Treating NaN values and imputing

Data Preparation

1. Creating Dummy variables for categorical features
2. Splitting the data into train and test split
3. Scaling the numeric features

Model Building

1. Feature selection using RFE
2. Determine the optimal model using logistic regression
3. Calculate various metrics like Confusion matrix, accuracy, sensitivity, specificity, precision, recall and evaluate the model

Evaluation

1. Evaluate the final test predictions on the test set using the cutoff threshold from sensitivity and specificity metrics
2. Determine the target lead score conversion rate to be around 80%

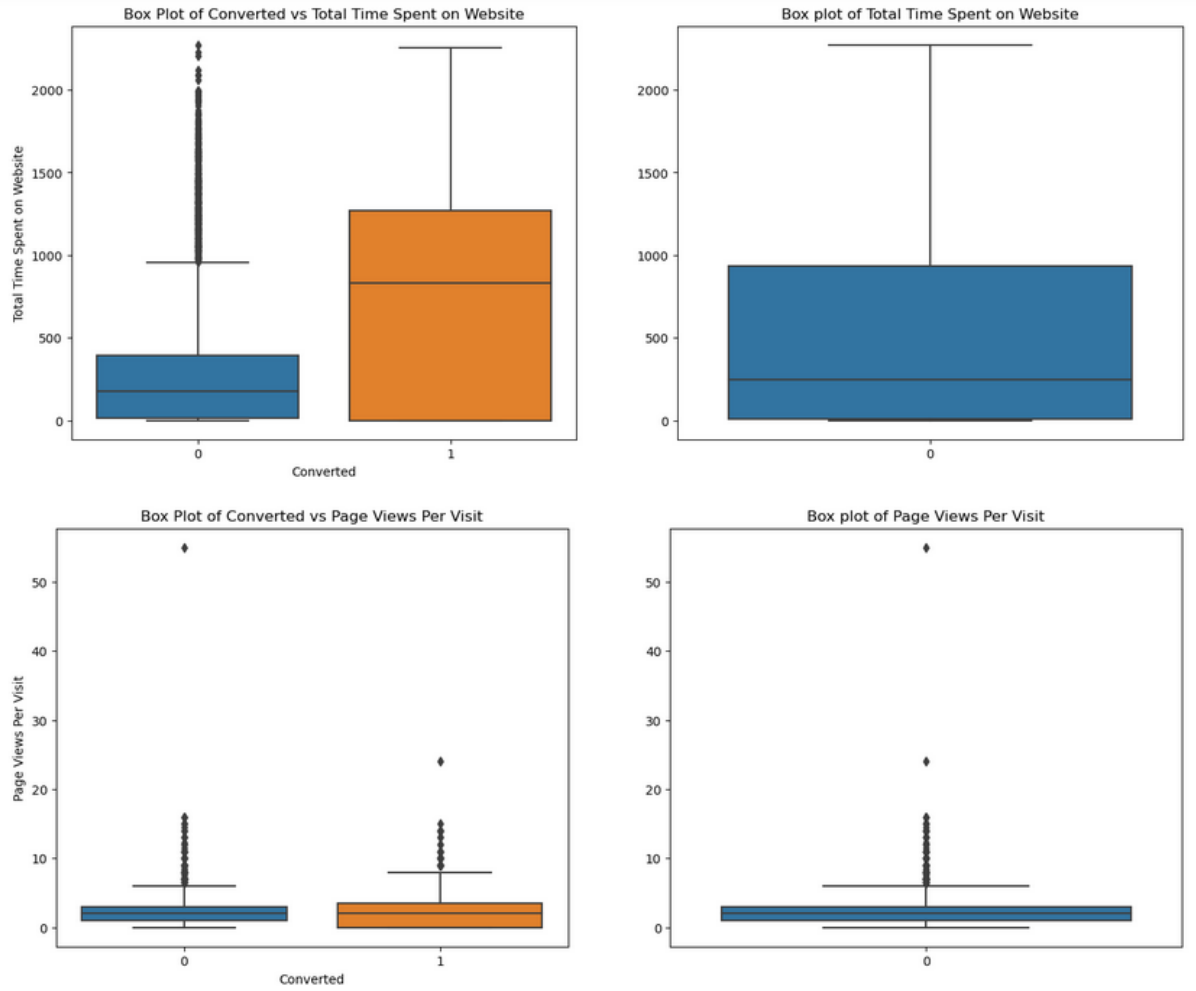
Data Visualisations & Insights (EDA)

EDA was performed using the following steps:

1. **Importing the dataset, Data cleaning and preparation:**
   - Importing the dataset, Data cleaning and preparation
   - Data Cleaning - Identifying NaN values and unique values in each columns
   - Identifying data issues with the columns if any and treating suitably
2. **EDA - Inferences from the Data**
   - Univariate and Bi-variate Analysis for Numerical columns
   - Univariate and Bi-variate Analysis for Categorical Columns
3. **EDA and Data Preparation for the Model**
   - Treating Nan values for numerical columns and dropping some columns based on EDA/NaN values
   - Treating NaN values of categorical columns and dropping some columns based on nan values and EDA
   - Removing Unique columns and creation/modifying existing columns based on EDA
   - Converting Categorical Columns suitably for the model
   - Splitting the data into Train and Test Datasets
   - Scaling the data for Numerical columns

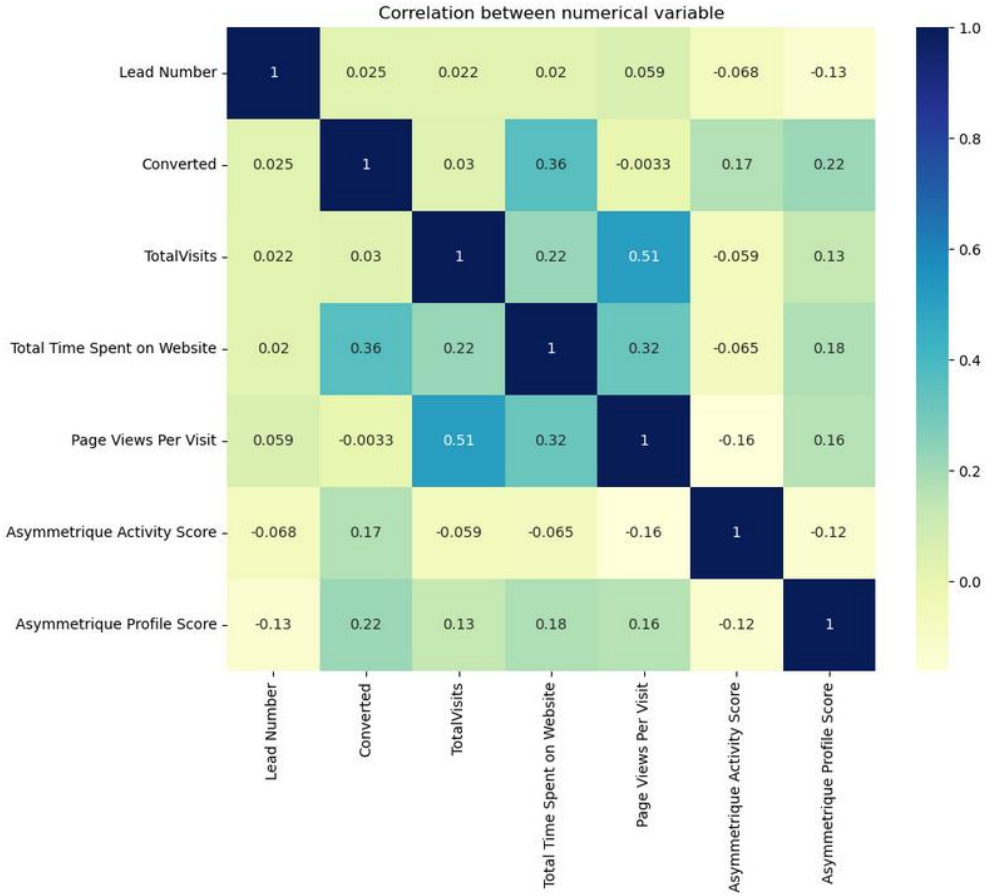
Data Visualisations & Insights (EDA)

The following graph highlights the boxplot of different columns wrt converted status and just the column

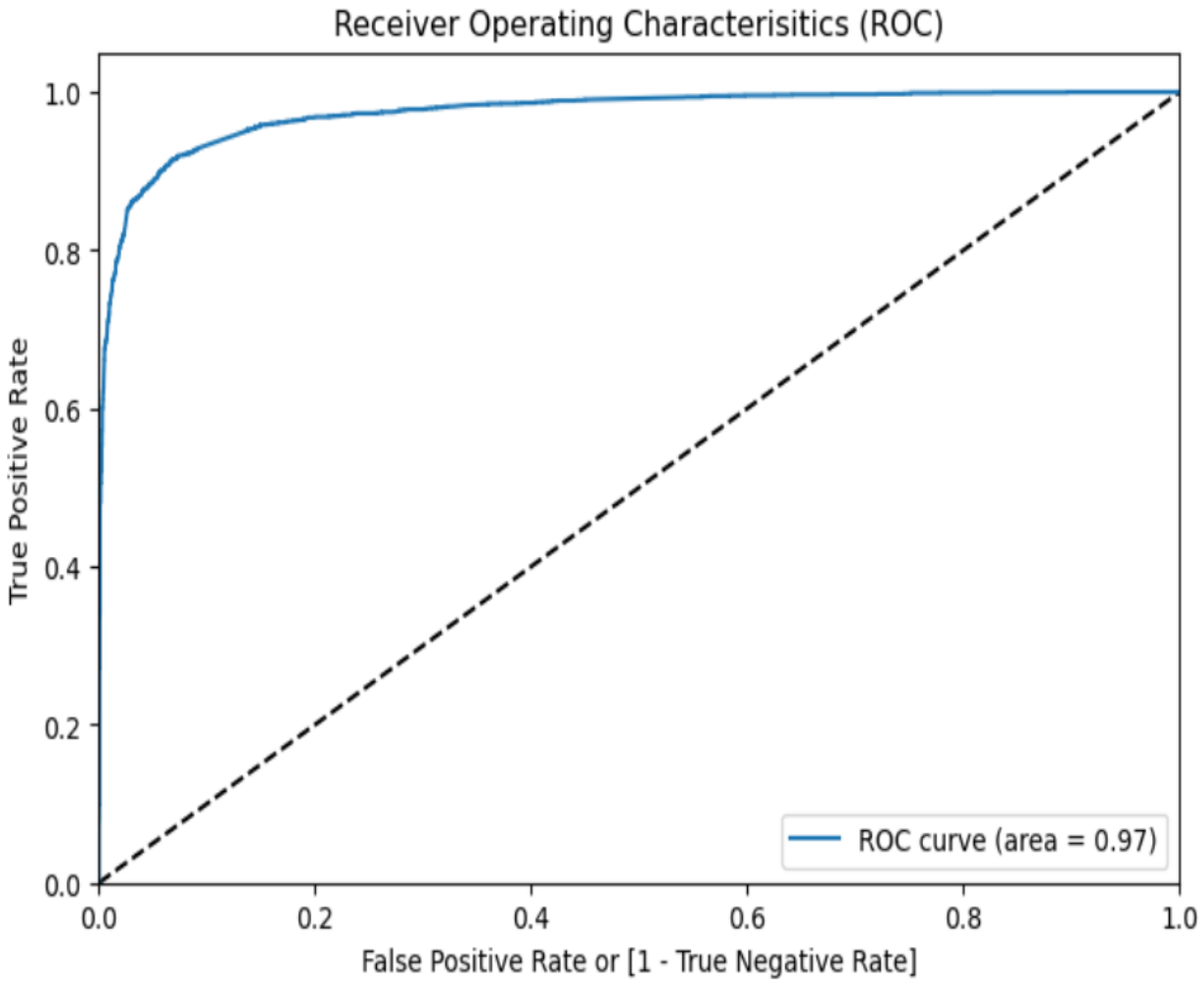


Data Visualisations & Insights (EDA)

The following image shows the heat map among the numeric variables:



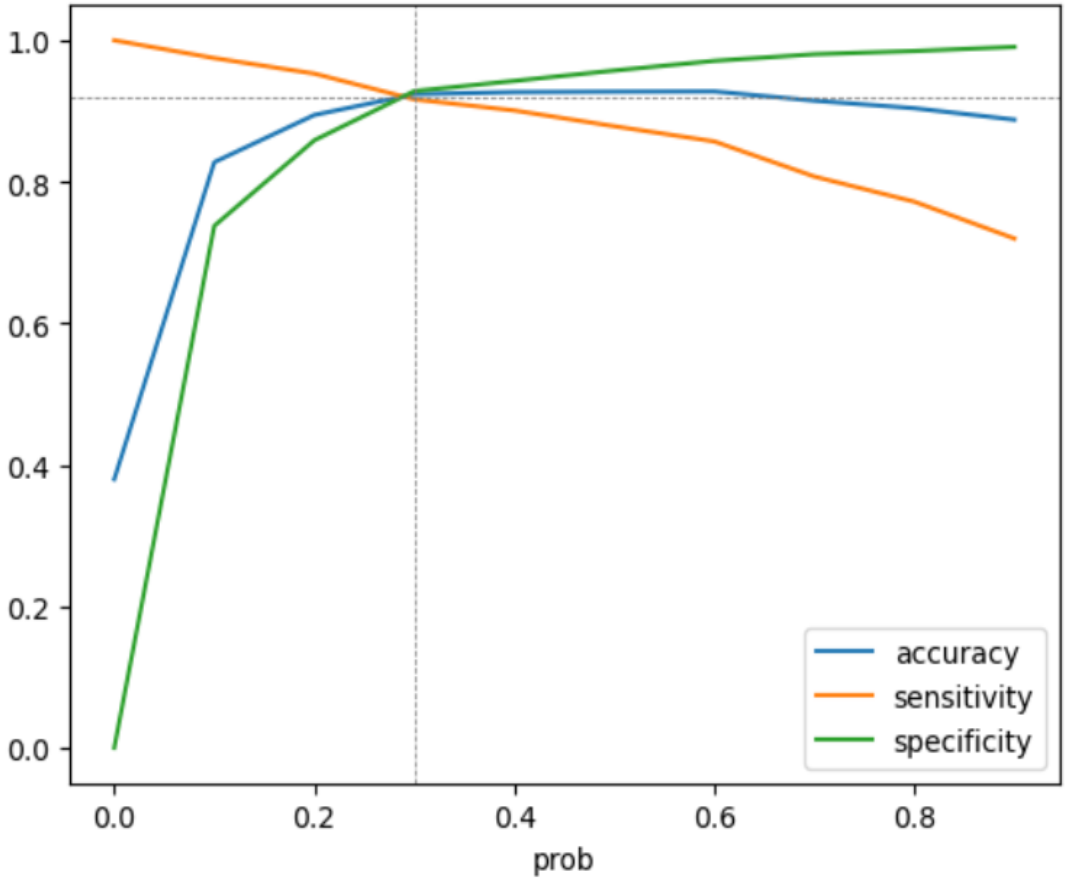
Model Building – RoC Curve



The area that the RoC (Receiver Operating Characteristics) curve is 0.97

- It shows the tradeoff between Sensitivity and Specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45 degrees diagonal of the ROC space, the less accurate the test.

Model Evaluation Metrics on Train data



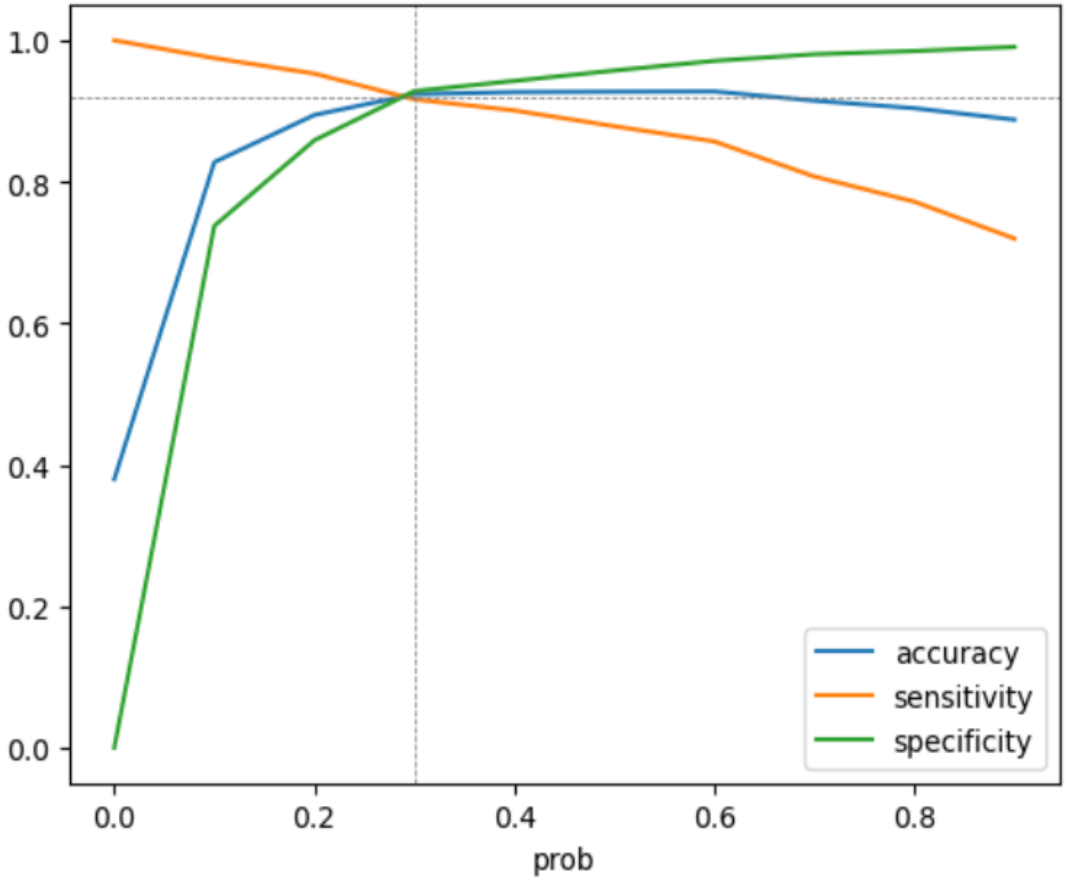
The graph depicts an optimal probability cutoff of '0.3' based on Accuracy, Sensitivity, Specificity.

Confusion Matrix :

3669 (TN)	284 (FP)
202 (FN)	2217 (TP)

- Accuracy – 92.37 %
- Sensitivity – 91.65 %
- Specificity – 92.82 %
- False Positive rate – 7.18 %
- Positive predictive value – 88.64 %
- Negative Predictive value – 94.78 %

Model Evaluation Metrics on Train data



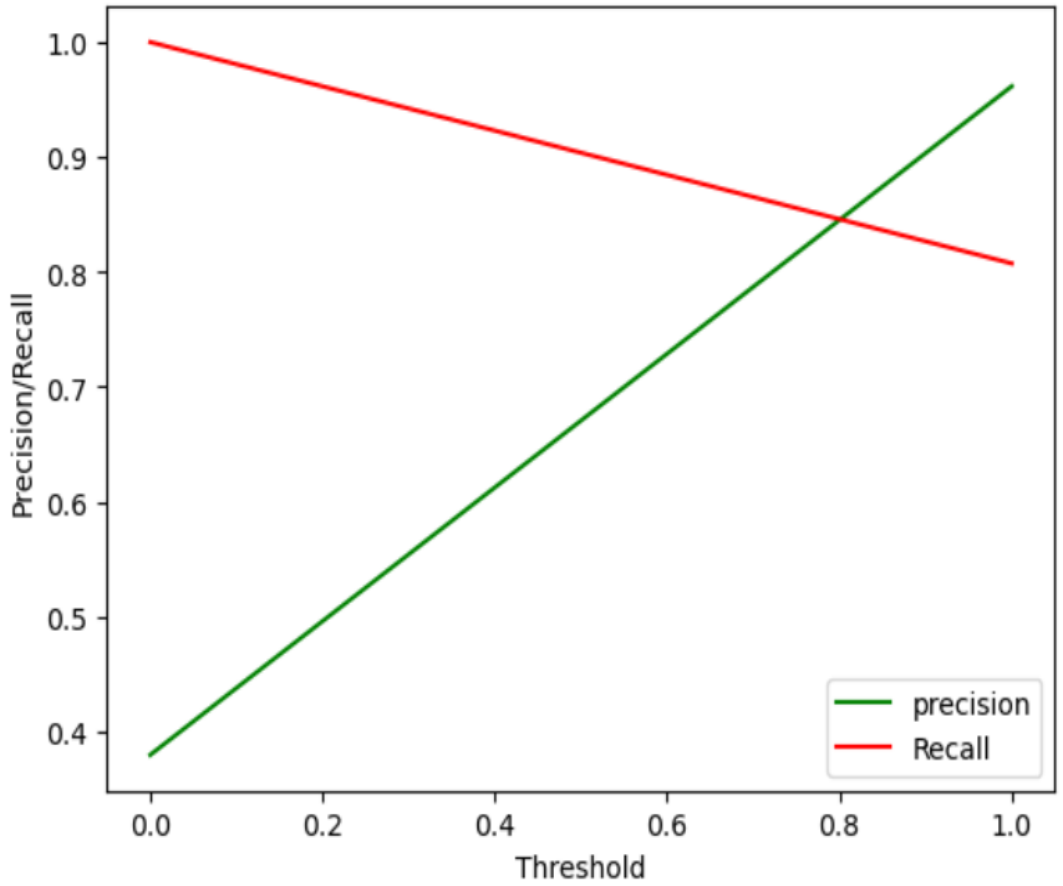
When an optimal probability cutoff of '0.7' based on Accuracy, Sensitivity, Specificity.

Confusion Matrix :

3875 (TN)	78 (FP)
466 (FN)	1953 (TP)

- Accuracy – 91.46 %
- Sensitivity – 80.74 %
- Specificity – 98.03 %
- False Positive rate – 1.97 %
- Positive predictive value – 96.16 %
- Negative Predictive value – 89.27 %

Model Evaluation Metrics – Precision & Recall on Train data



When an optimal probability cutoff of '0.7' based on Accuracy, Sensitivity, Specificity.

Confusion Matrix :

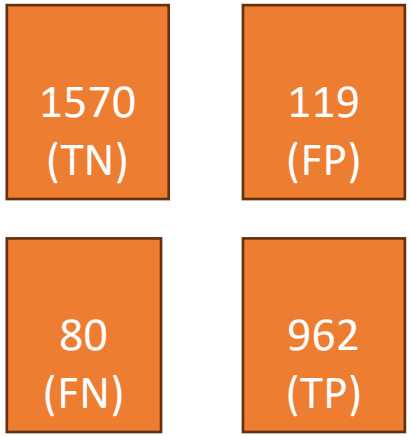
3875 (TN)	78 (FP)
466 (FN)	1953 (TP)

- Precision – 96.16 %
- Recall – 80.74 %

Model Evaluation Metrics – Sensitivity and Specificity on Test data

When an optimal probability cutoff of ‘0.3’ based on Accuracy, Sensitivity, Specificity.

Confusion Matrix :

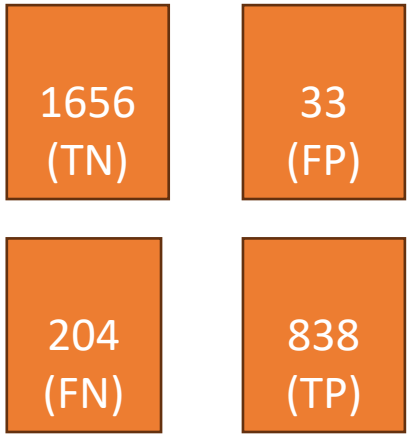


- Accuracy – 92.71 %
- Sensitivity – 92.32 %
- Specificity – 92.95 %

Model Evaluation Metrics – Sensitivity and Specificity on Test data

When an optimal probability cutoff of ‘0.7’ based on Accuracy, Sensitivity, Specificity.

Confusion Matrix :



- Accuracy – 91.32 %
- Sensitivity – 80.42 %
- Specificity – 98.05 %

Conclusions

- After Building the model, we have performed evaluation metrics on both train and test data. Checked based on the Sensitivity and Specificity as well as Precision and Recall. We have considered the optimal cutoff based on Sensitivity and Specificity for calculating the final predictions.
- On the train dataset, for the optimal threshold of 0.3 we got the sensitivity or Recall as 91.65 % while the optimal cutoff chosen as 0.7 we got the sensitivity or Recall as 80.74 %.
- On the test dataset, for the optimal threshold of 0.3 we got the sensitivity or Recall as 92.32 % while the optimal cutoff chosen as 0.7 we got the sensitivity or Recall as 80.42 %.
- Also the lead score calculated shows the conversion rate on the final predicted model is above 80% in train and test set

Conclusions

- The top 3 variables that contribute for lead getting converted in the model based on the highest coefficients that are obtained after the final model
  - Tags_Lost to EINS – 6.6208
  - Tags_Closed by Horizzon – 6.0712
  - Tags_Will revert after reading the email – 4.4926

- Based on the above summary of model prediction and evaluation it was concluded that this model seems to be good with good sensitivity or recall which satisfies our requirement.