Lead Scoring Case Study

Neelam Bairagi, Kameswara Rao, Sriharan

Content

- Problem Statement
- Business Goal
- Strategy
- Data Sourcing, Cleaning & Preparation
- Data Preparation
- EDA
- Univariate Analysis of all Category Feature
- Univariate Analysis for Continuous Features
- Bivariate Analysis – Categorical Features
- Bivariate Analysis – Continuous Features
- Scaling & splitting Train & test sets
- Model Building
- Model Evaluation- Sensitivity & Specificity On Train Data Set
- Model Evaluation- Precision & Recall on Train Dataset
- Model Evaluation – Sensitivity & Specificity on Test Dataset
- Result
- Conclusion

Problem Statement

- An education company named X Education sells online courses and mark their course on several website like google
- Company wants to select most promising leads that can be converted to paying customers
- Company generates a lot of leads but converted few into paying customers.
- Company wants a higher lead conversion.
- Leads comes through numerous mode like email, advertisements on websites
- Company has had 30% conversion rate through the whole process of turning lead into customers by approaching those leads.
- The implementation process of lead generating attributes are not efficient in helping conversions.

Business Goal

- Company requires a model to be built for selecting most promising leads
- Leads score to be given to each such that it indicates how promising the lead could be
- The higher the lead score the more promising the lead to get converted, the lower it is the lesser the chances of conversion
- The model to be built in lead conversion rate around 80% or more.

Strategy

- Import data
- Understand the data
- Clean The data
- EDA
- Model Building
- Model Evaluation
- Making Predictions on the Test set

Data Sourcing, Cleaning & Preparation

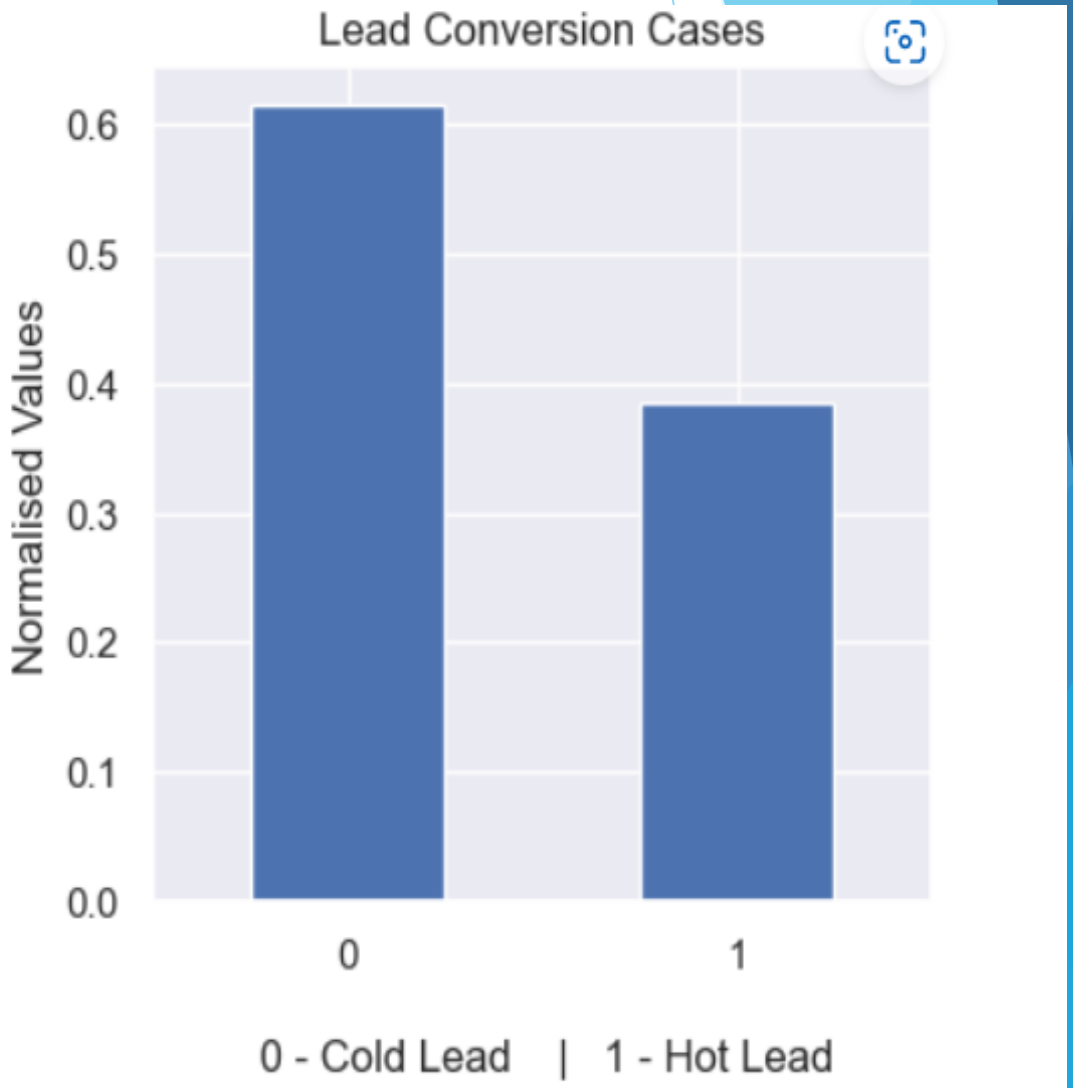
- Read the data from CSV file
- Outlier treatment
- Data cleaning – Handling null values & removing higher null values data
- Removing redundant columns columns in the data
- Imputing null values
- Exploratory null values
- Exploratory data analysis
- Feature standardization

Data Preparation

- Converted Binary Variables into 0 & 1
- Created dummy variables for Categorical Variables

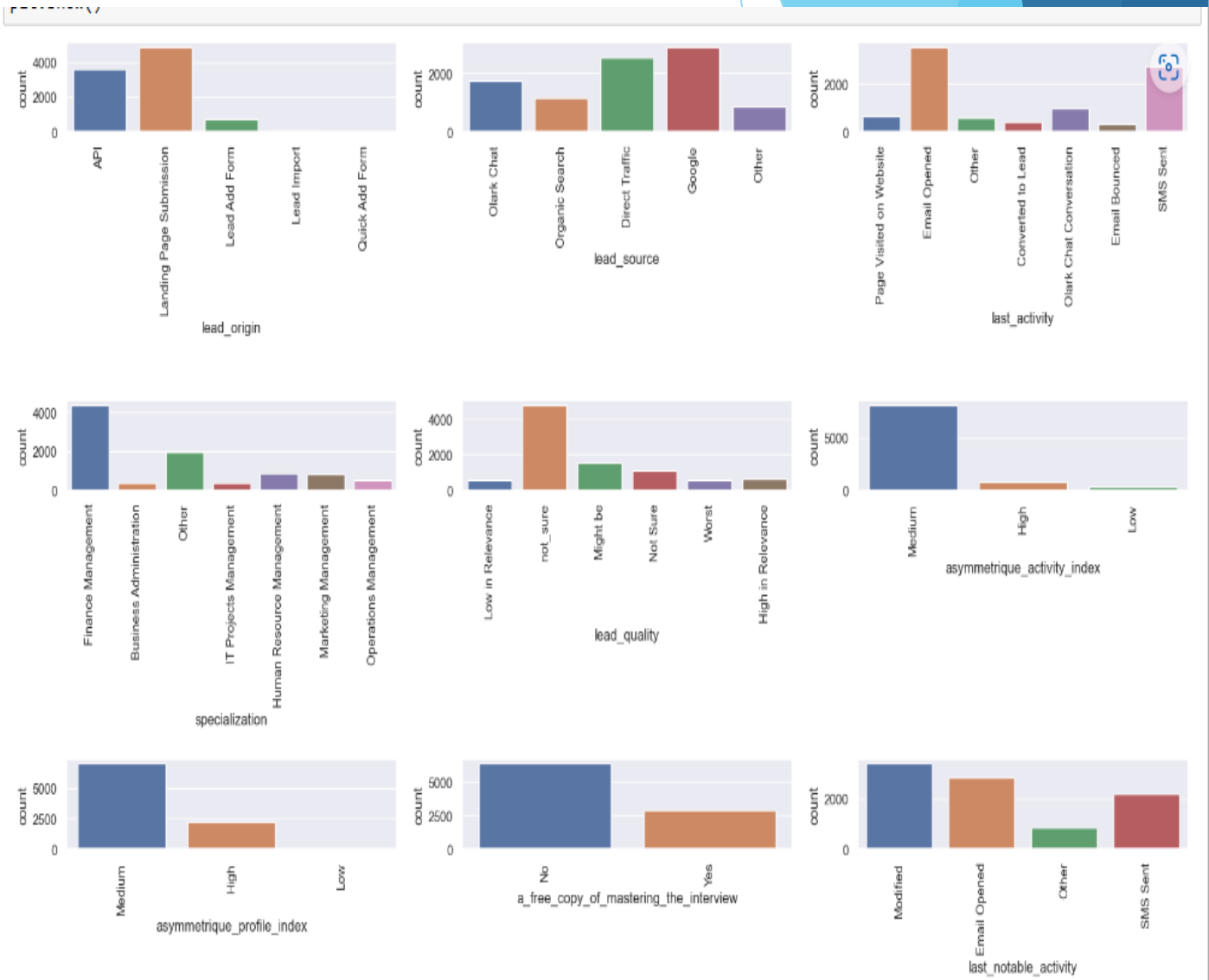
Exploratory Data Analysis

- Univariate Analysis
- Based on the data, 48% of the lead data is converted and 62% of the data is not converted. Now we have to propose recommendations for data to be converted.



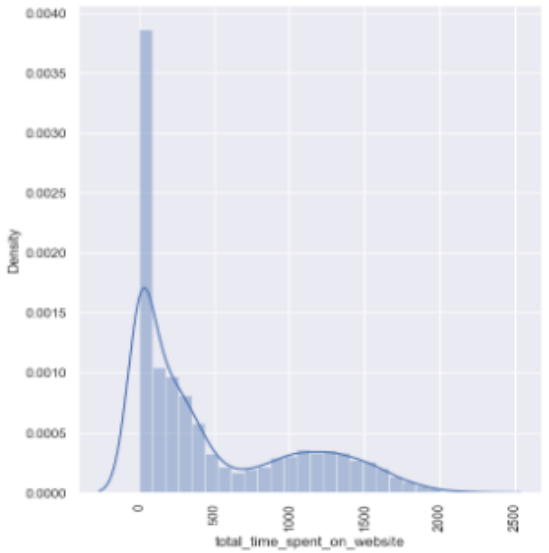
Univariate Analysis of all Category Feature

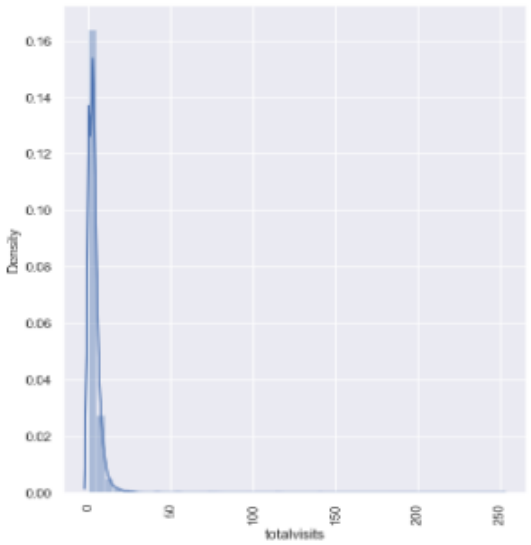
- In Lead Origin API and 'Landing Page submission" are the two main origins for Leads
- The Number of values is High in Email Opened and SMS Sent in Last Activity
- Most of the people chooses Finance Management Specialization rather than othe Specialization
- Most of the unemployed people are preferre for this specialization. This needs to be checked further.
- Most of the people are choosing the specialization for better career proposition.

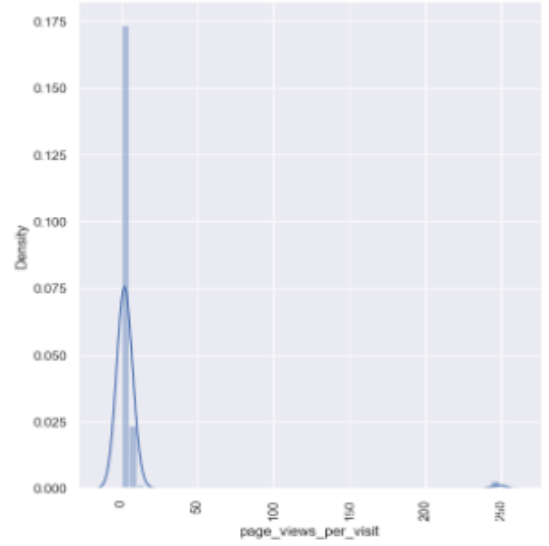


Univariate Analysis for Continuous Features

- None of the Continuous Variables are in Normal distribution and data is skewed right.
- Presence of Outliers in Total Visits and Page Views Per Visit
- In total visits more values is between 0-50 and page views per visits 0-20

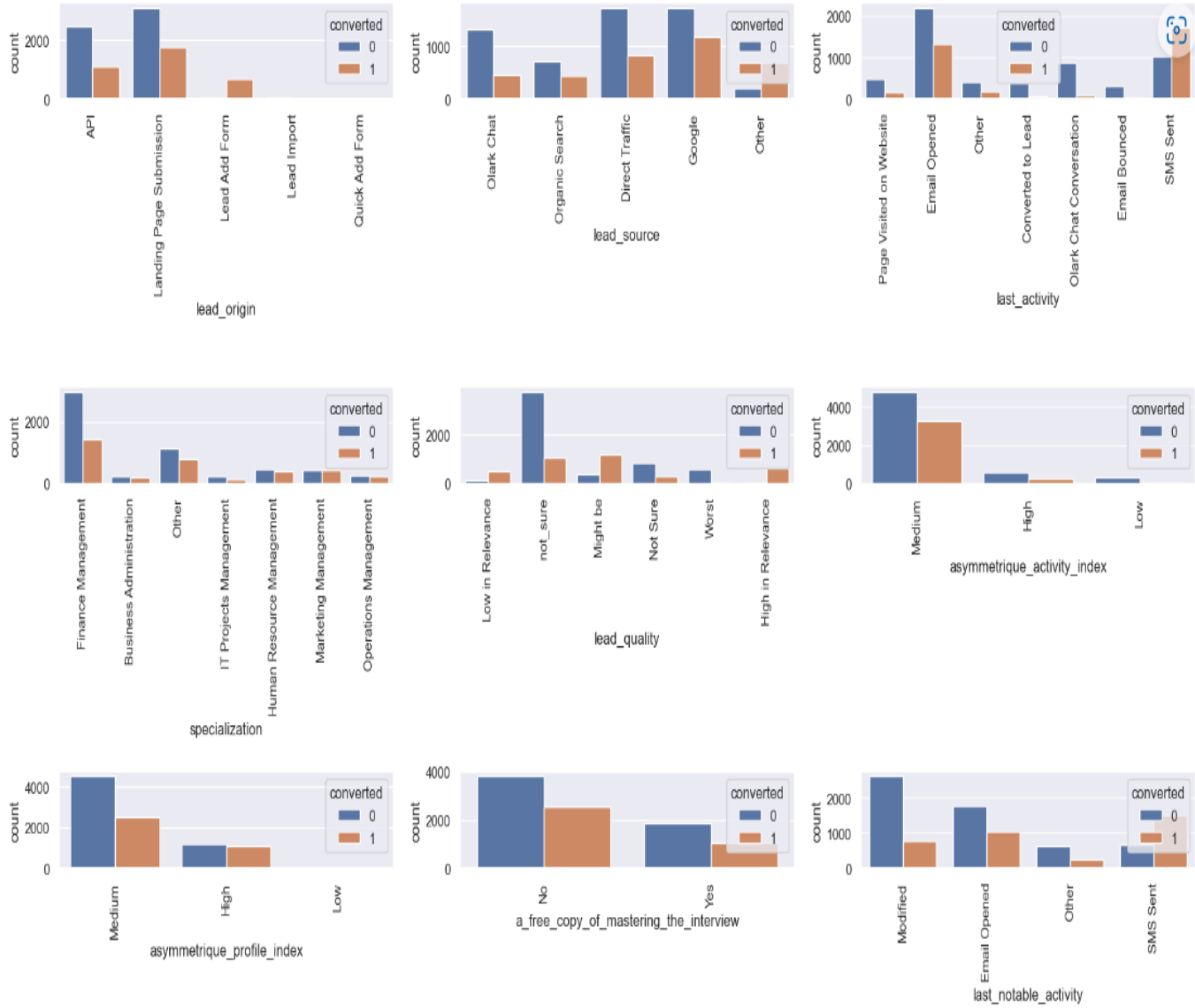






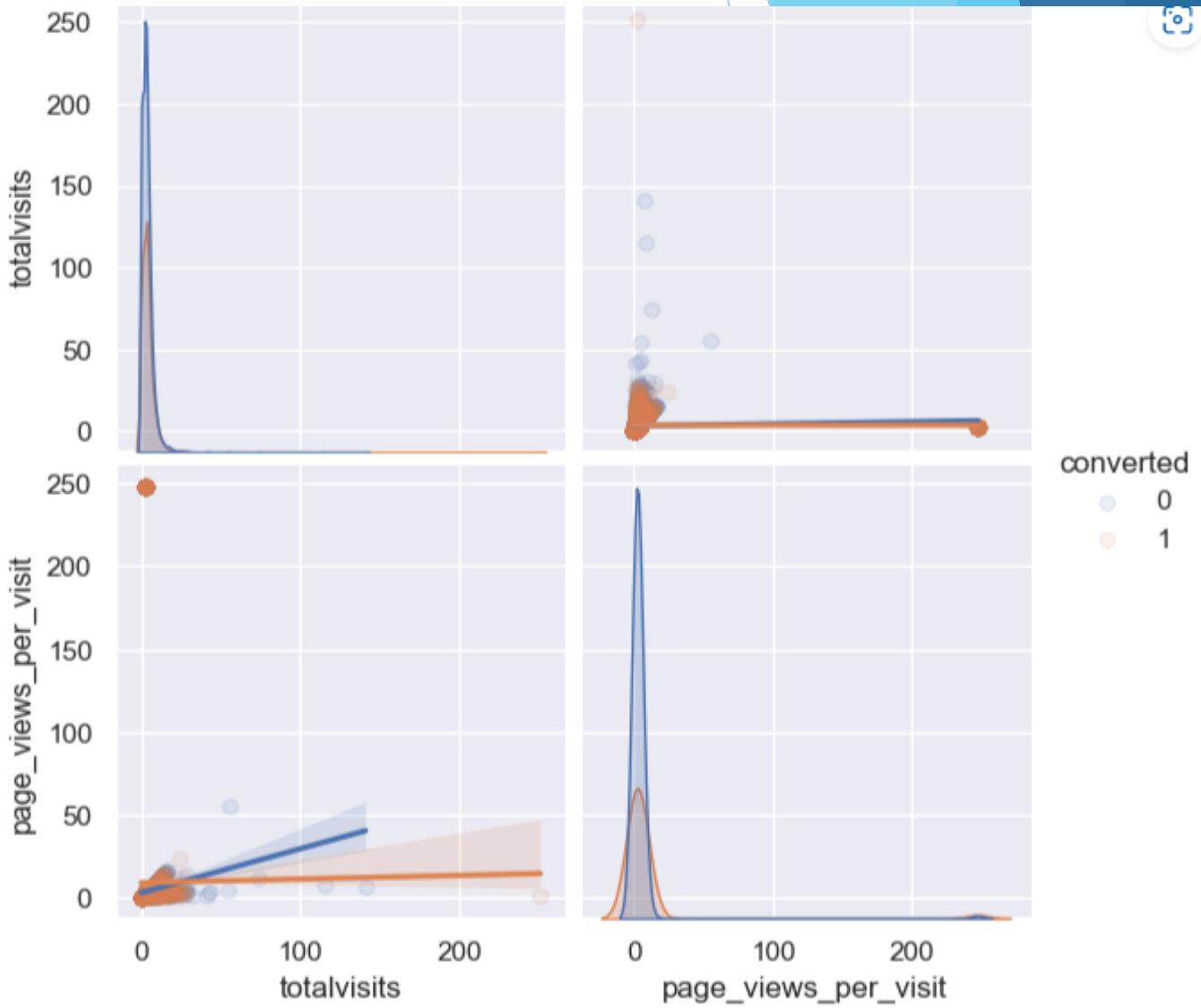
Bivariate Analysis – Categorical Features

- In Lead Source The number of Hot leads is higher in Direct Traffic and Google less in Other Category
- In Last Activity the number of Hot leads is higher in SMS and in EMAIL cold leads is higher than hot leads.
- In Last Notable Activity it's mostly same as Last Activity.
- In Specialization the most of the leads are comes from Finance management but here Hot leads are lesser than Cold leads.

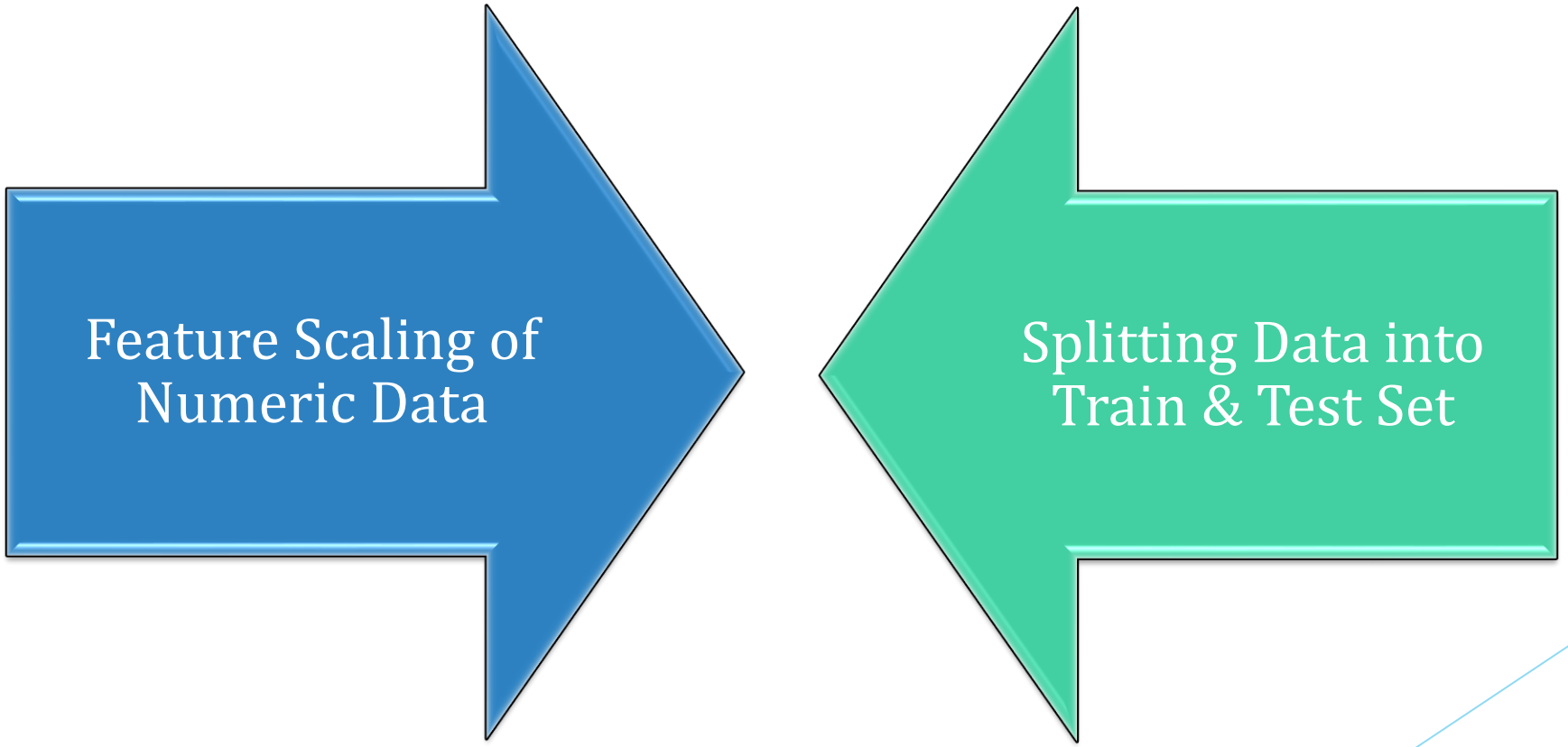


Bivariate Analysis – Continuous Features

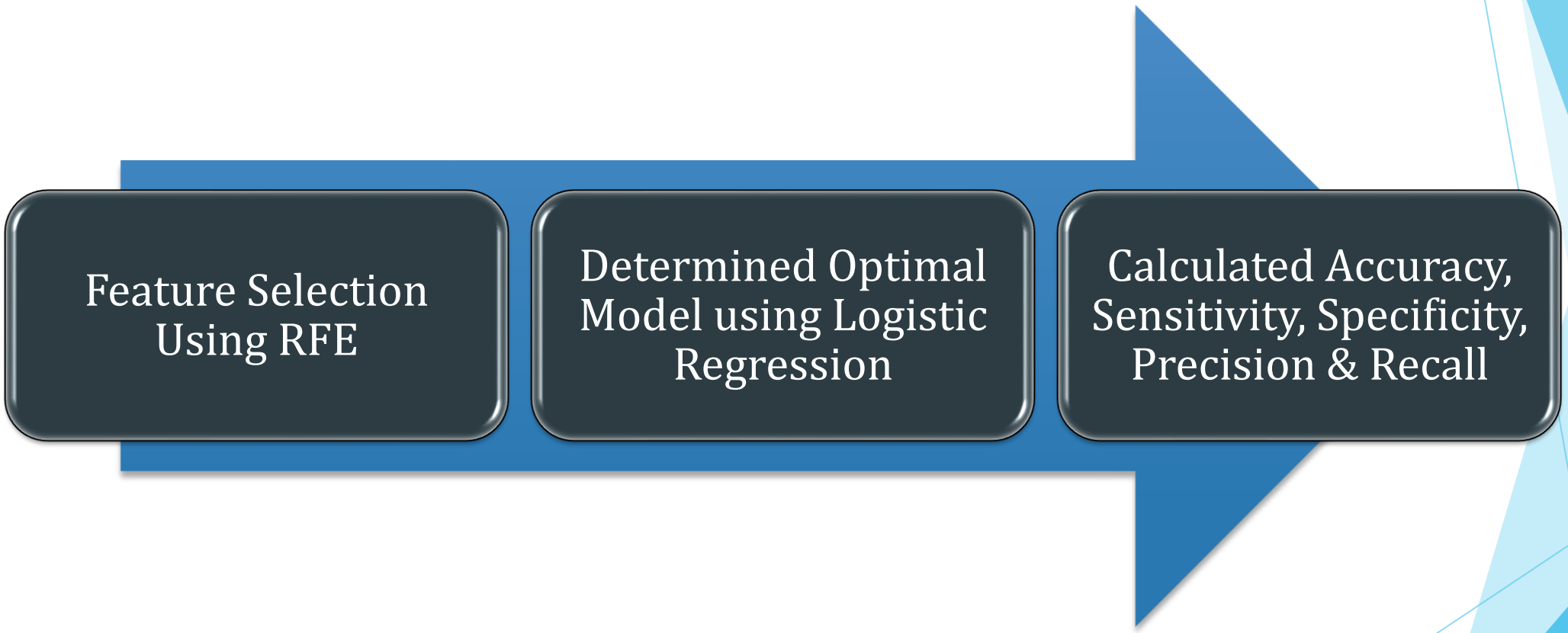
- Data transformation will be require to get the more clarity



Scaling & Splitting Train & Test sets

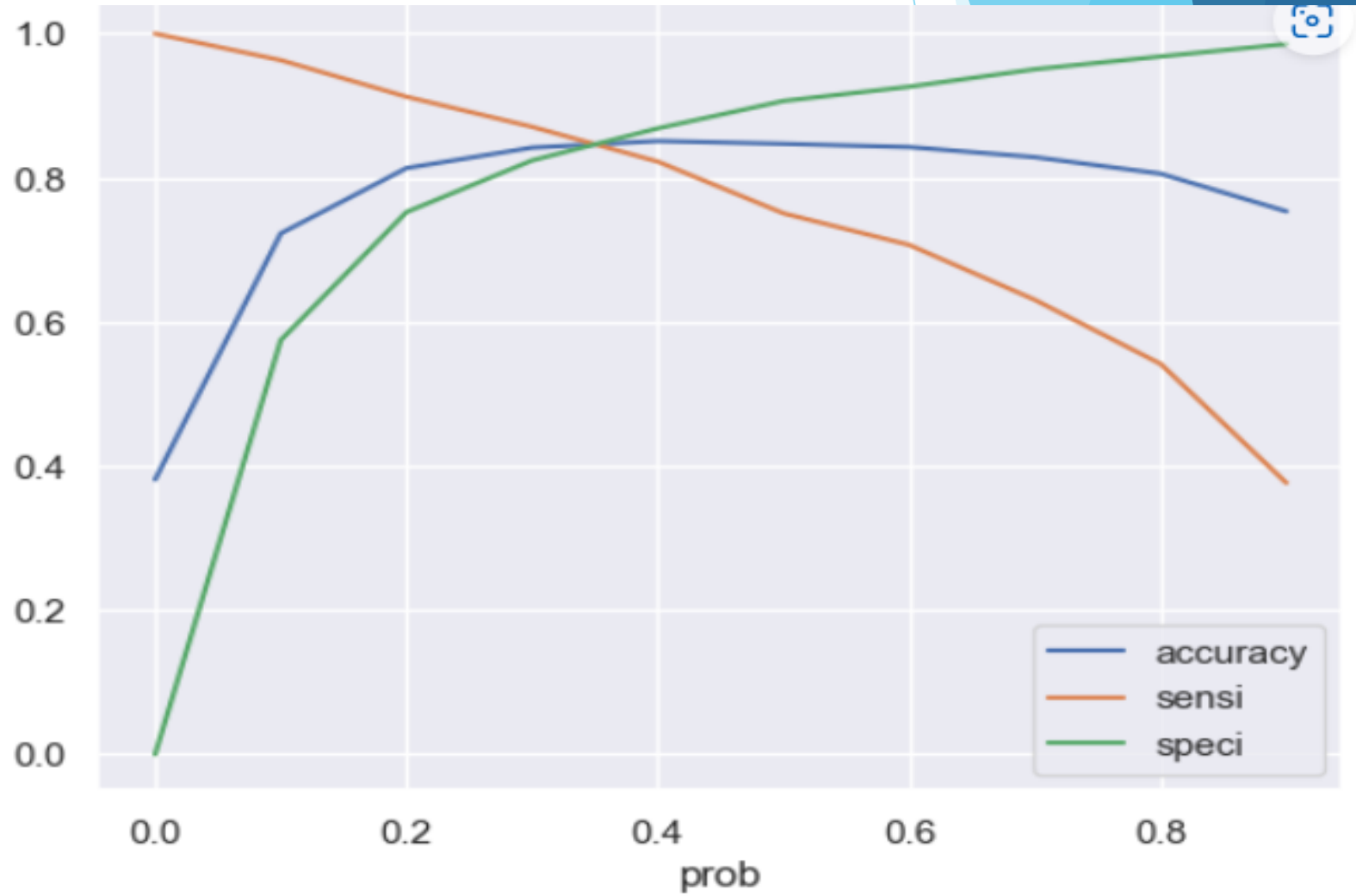


Model Building



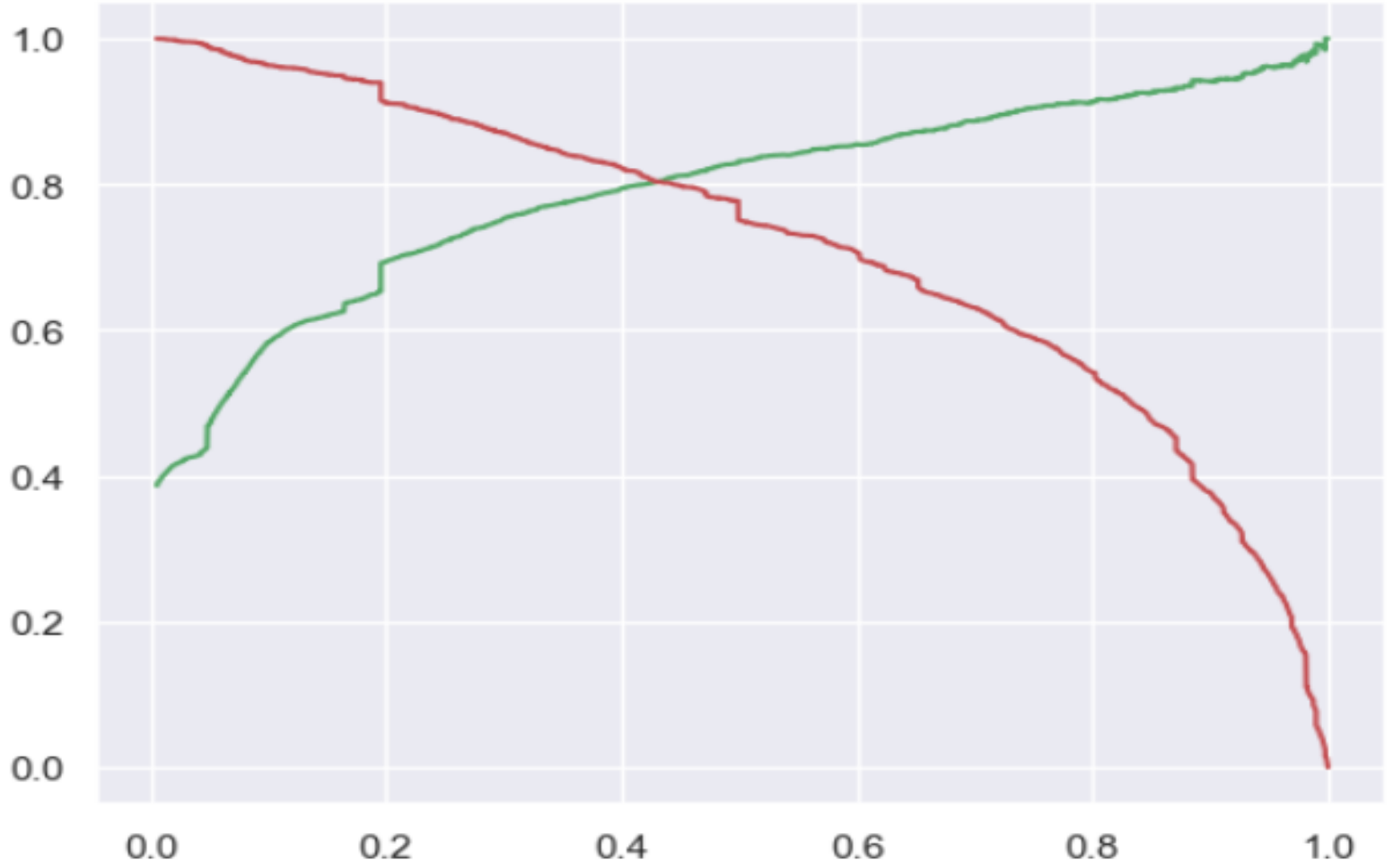
Model Evaluation- Sensitivity & Specificity On Train Data Set

- Graph depicts an optimal cutoff of 0.38 based on Accuracy, sensitivity, Specificity
- Accuracy – 84.9%
- Sensitivity- 83.1%
- Specificity – 86%

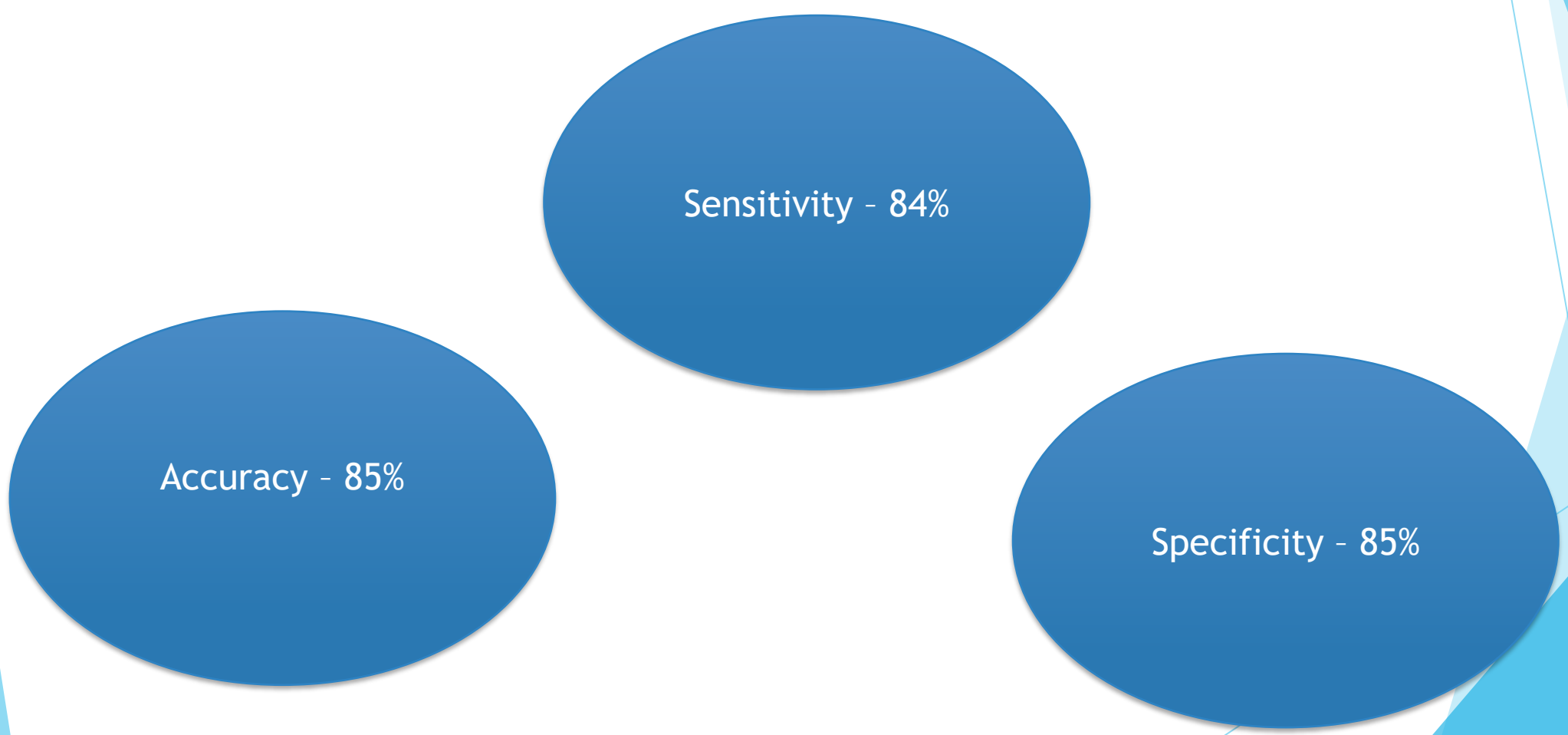


Model Evaluation- Precision & Recall on Train Dataset

- Graph depicts an optimal cutoff of 0.42 based on Accuracy, sensitivity, Specificity.



Model Evaluation – Sensitivity & Specificity on Test Dataset



Result

- Accuracy, Sensitivity and Specificity values of training and test set are close to training set.
- Accuracy, Sensitivity and specificity values of training set are 85%, 83% & 86% respectively
- Accuracy, Sensitivity and specificity values of test are 85%, 84% & 85% respectively.
- Conversion rate for Train & Test Dataset is 83% & 84%.
- We have done the prediction on the test set using cut off threshold from sensitivity & specificity metrics.

Conclusion

- We have checked both sensitivity as well as Precision & recall metrics.
- We have considered the optimal cut off based on sensitivity & Specificity for calculating the final prediction.
- Accuracy, Sensitivity and specificity values of test set are around 85%, 84% & 85% which closer to values calculated using trained Dataset
- Lead score calculated for the conversion rate final model on train & Test dataset is 83% & 84%.
- Overall model seems to be good.

Thank You