X-EDUCATION LEAD SCORING CASE STUDY

- ANJALI SIKHWAL
- SUPARTH JAIN

Background Information

- An education company named X Education sells online courses to industry professionals. On any given day, 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%.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

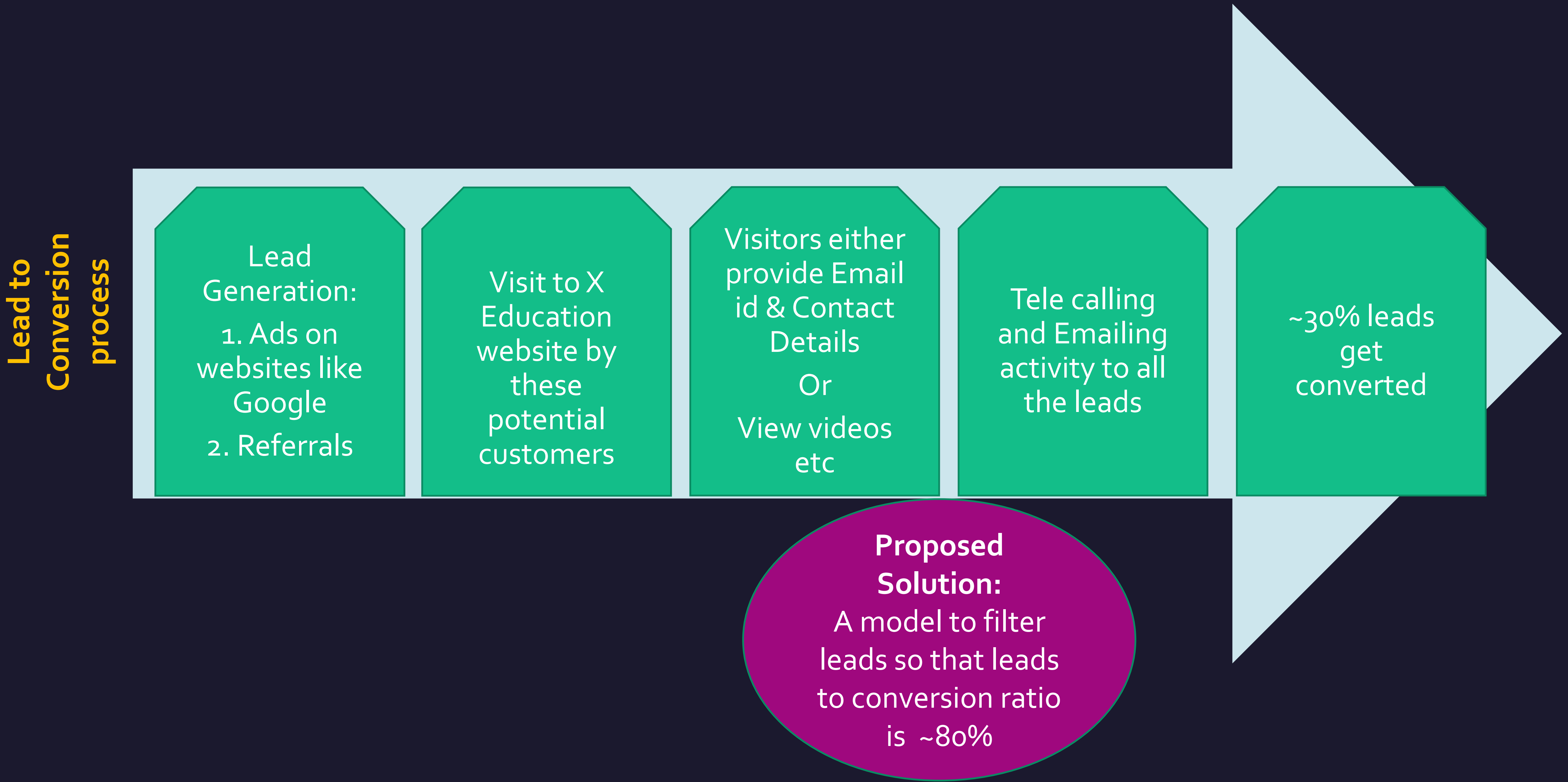
PROBLEM STATEMENT

- X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

- The company requires us 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%.

LEAD – CONVERSION PROCESS



Lead to Conversion process

Lead Generation:

1. Ads on websites like Google
2. Referrals

Visit to X Education website by these potential customers

Visitors either provide Email id & Contact Details

Or

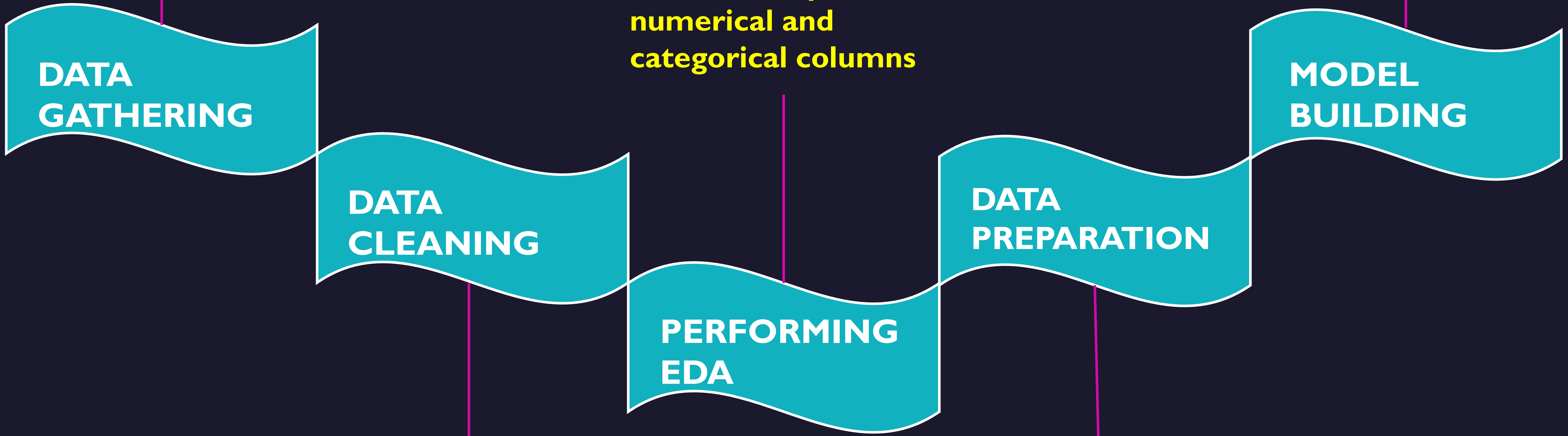
View videos etc

Tele calling and Emailing activity to all the leads

~30% leads get converted

Proposed Solution:
A model to filter leads so that leads to conversion ratio is ~80%

IMPLEMENTATION CYCLE



DATA GATHERING

Loading & Observing the past data provided by the Company

DATA CLEANING

Duplicate removal, null value treatment, unnecessary column elimination, etc.

PERFORMING EDA

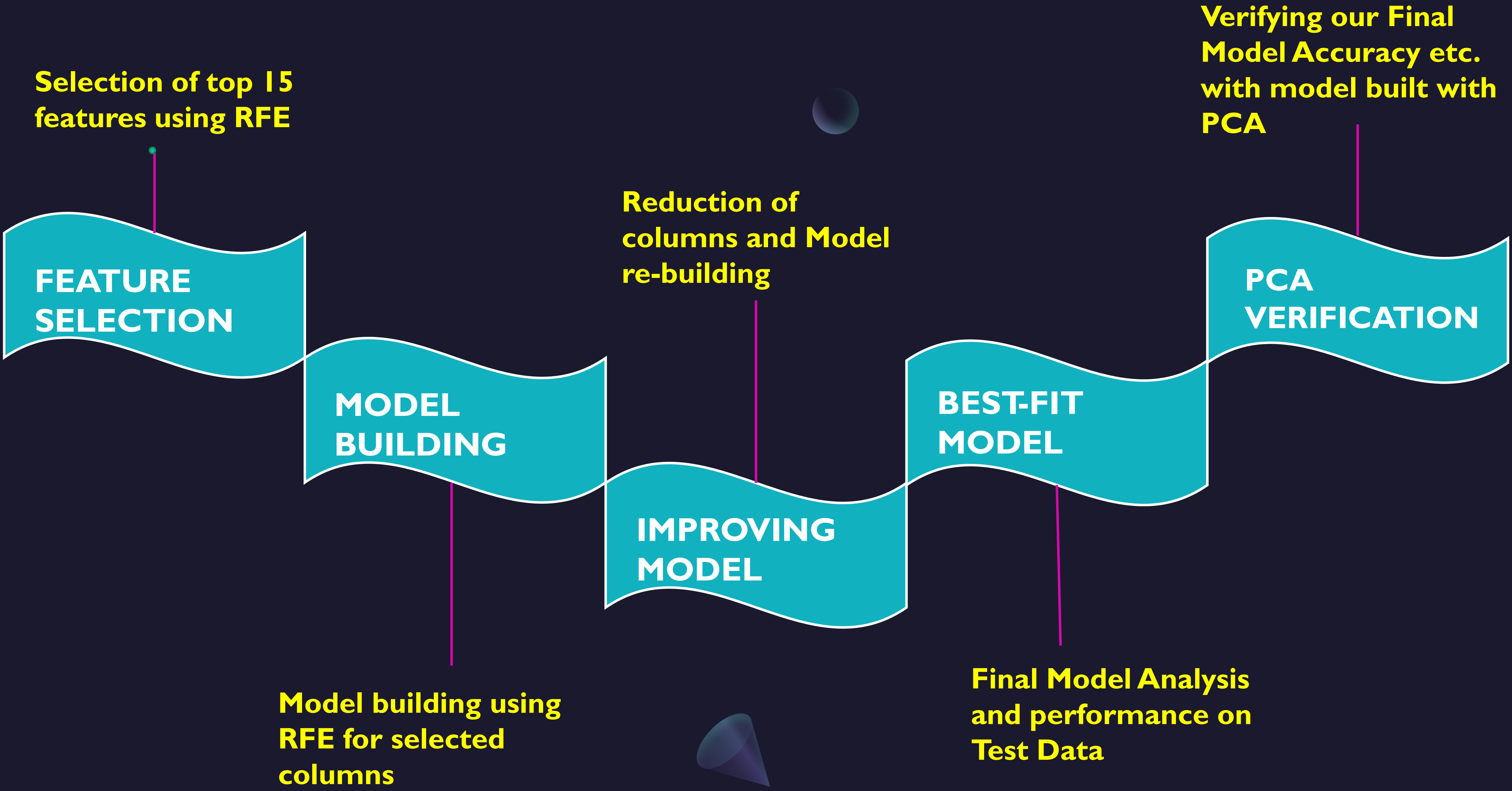
Univariate, Bivariate, and Heatmap for numerical and categorical columns

DATA PREPARATION

Outlier Treatment, Feature-Standardization

MODEL BUILDING

Performing pre-requisites for RFE and Logistic Regression



FEATURE SELECTION

Selection of top 15 features using RFE

MODEL BUILDING

Model building using RFE for selected columns

IMPROVING MODEL

Reduction of columns and Model re-building

BEST-FIT MODEL

Final Model Analysis and performance on Test Data

PCA VERIFICATION

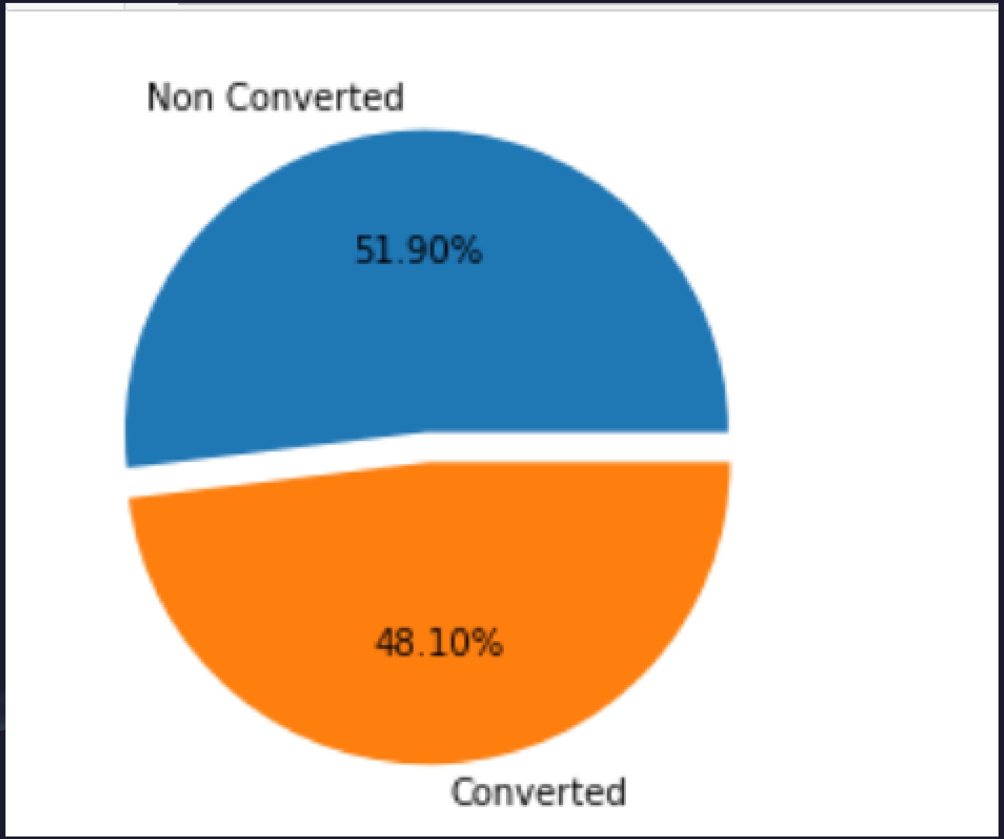
Verifying our Final Model Accuracy etc. with model built with PCA

Solution Methodology

- Data Cleaning and Manipulation
- Exploratory Data Analysis
- Feature Scaling and creating dummy variables
- Logistic Regression: Model Making and Prediction
- Validation of Model
- Model Presentation
- Conclusions and Recommendations

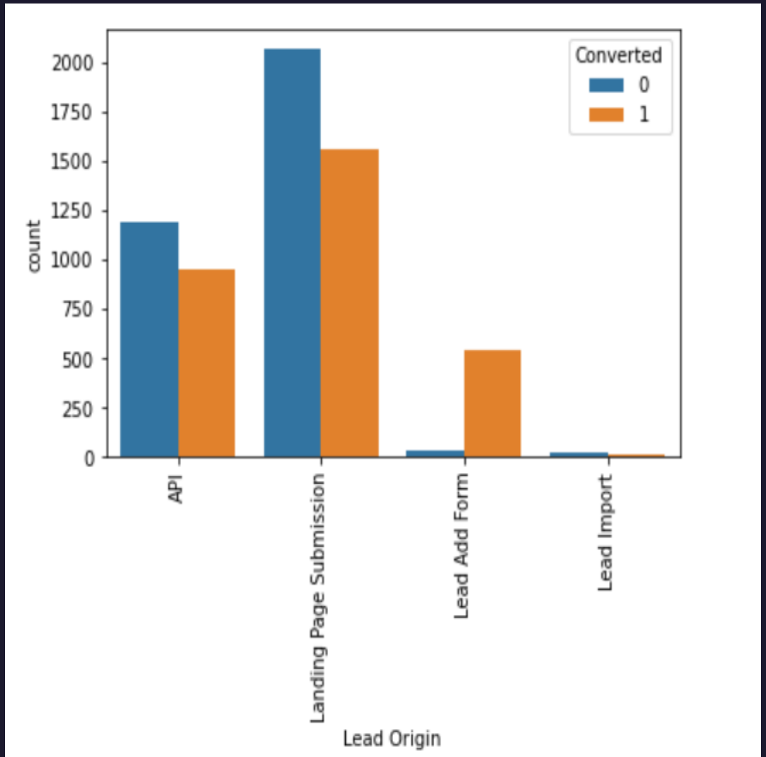
VISUALISATION

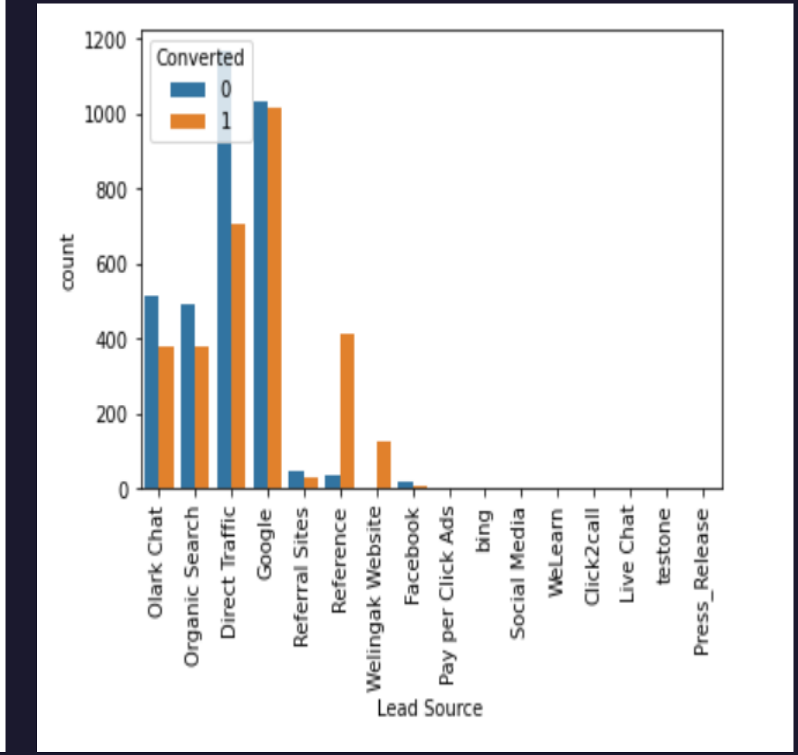
- **DATA IMBALANCE**

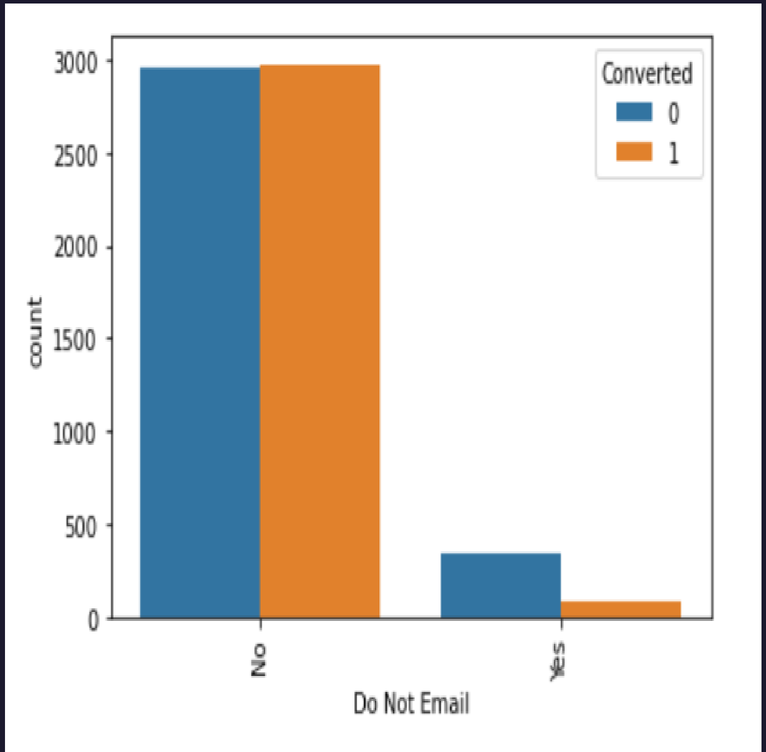


51.90% leads turn out to be non-converted and 48.10% leads were successfully converted

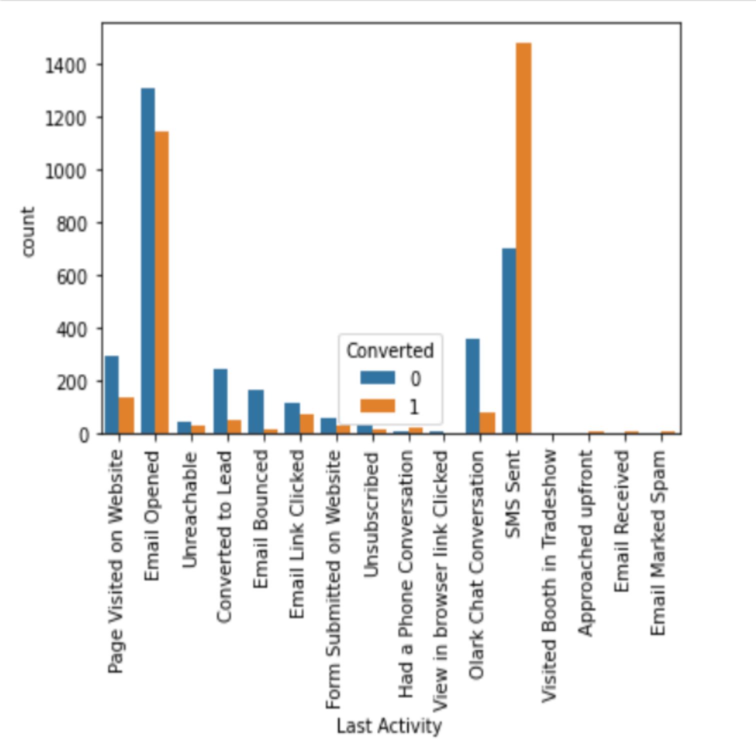
VARIATION IN CATEGORICAL COLUMNS (Converted v/s Non-converted Leads)

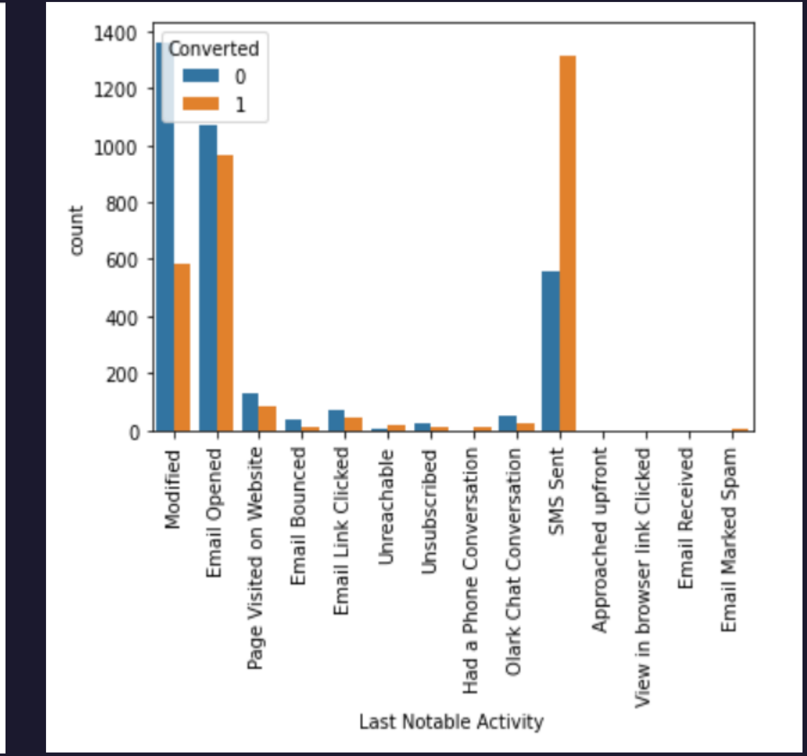


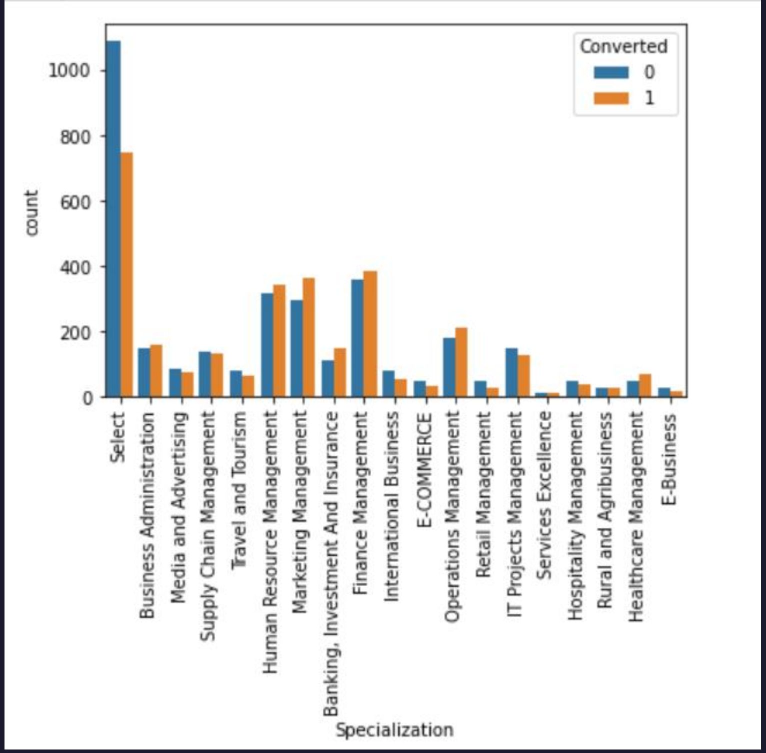




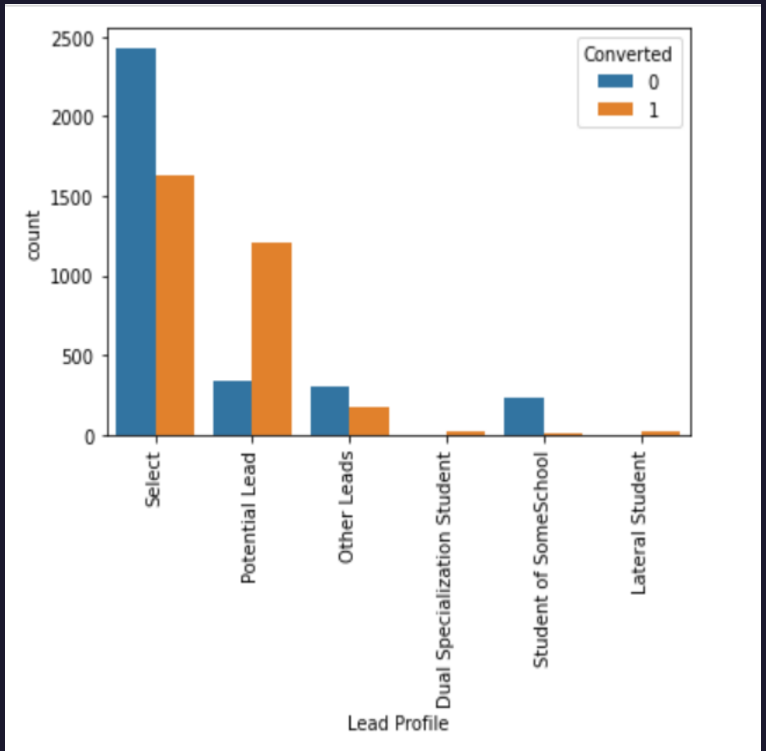
VARIATION IN CATEGORICAL COLUMNS (Converted v/s Non-converted Leads)

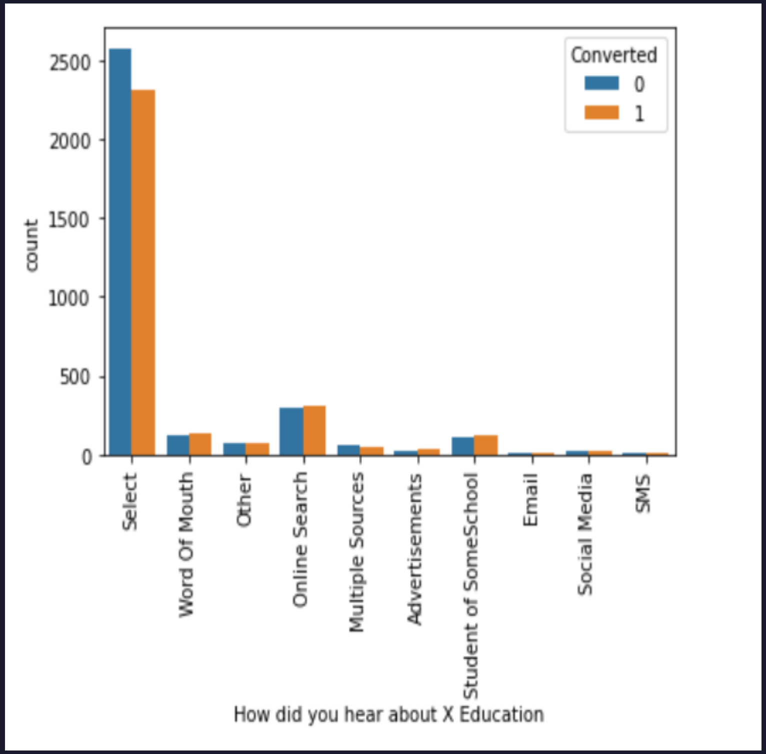


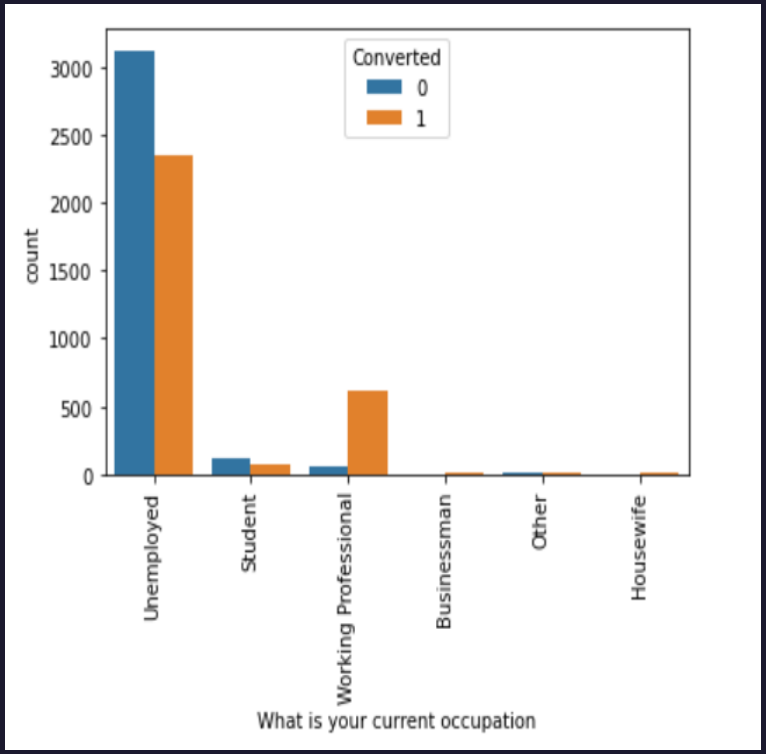




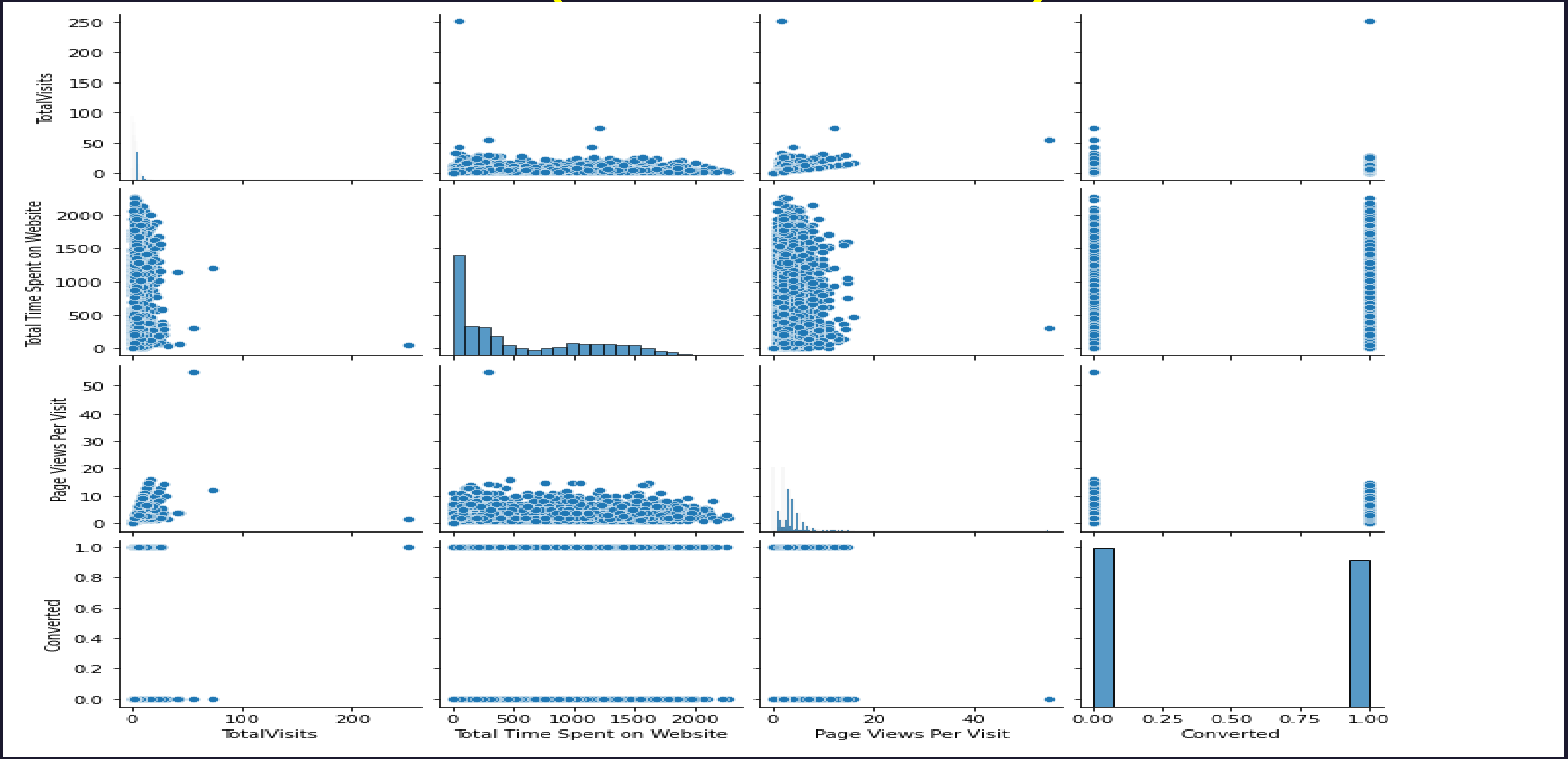
VARIATION IN CATEGORICAL COLUMNS (Converted v/s Non-converted Leads)



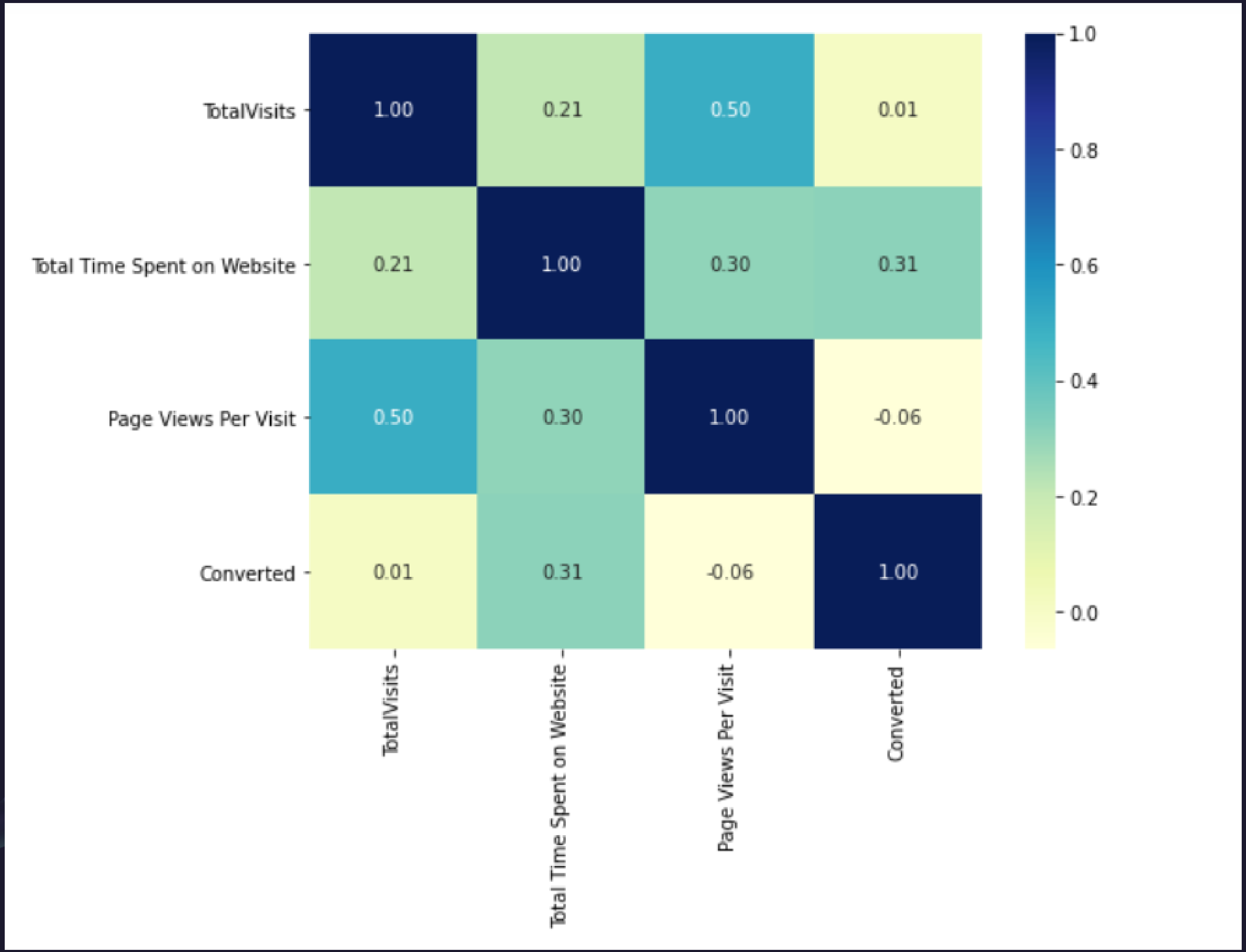




PAIRPLOTS FOR CONTINUOUS VARIABLES (Converted Leads)



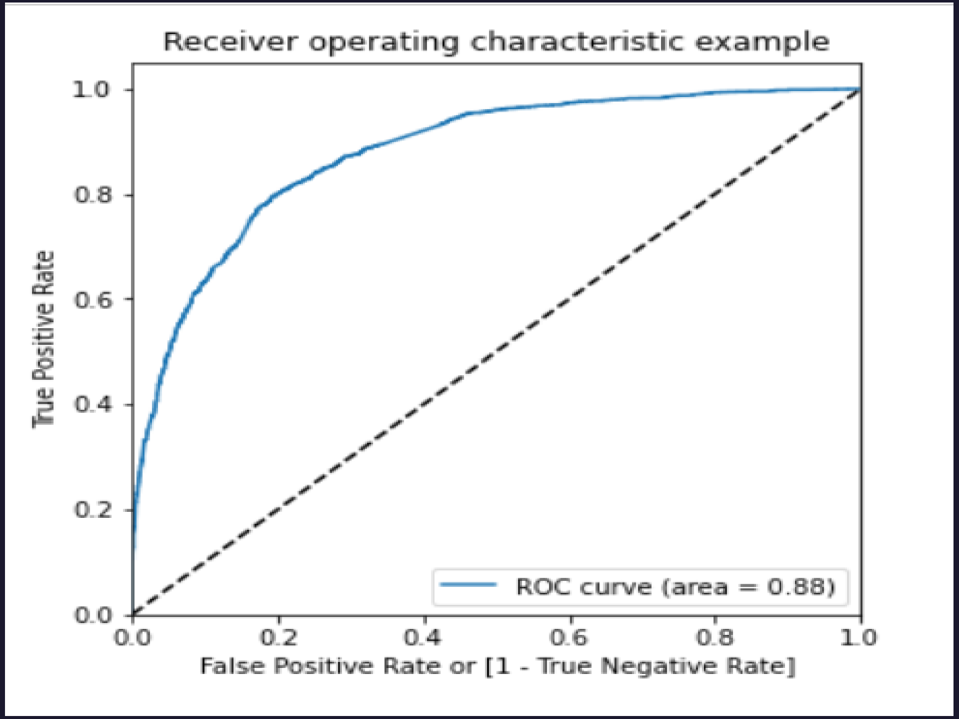
CORRELATION b/w CONTINUOUS VARIABLES

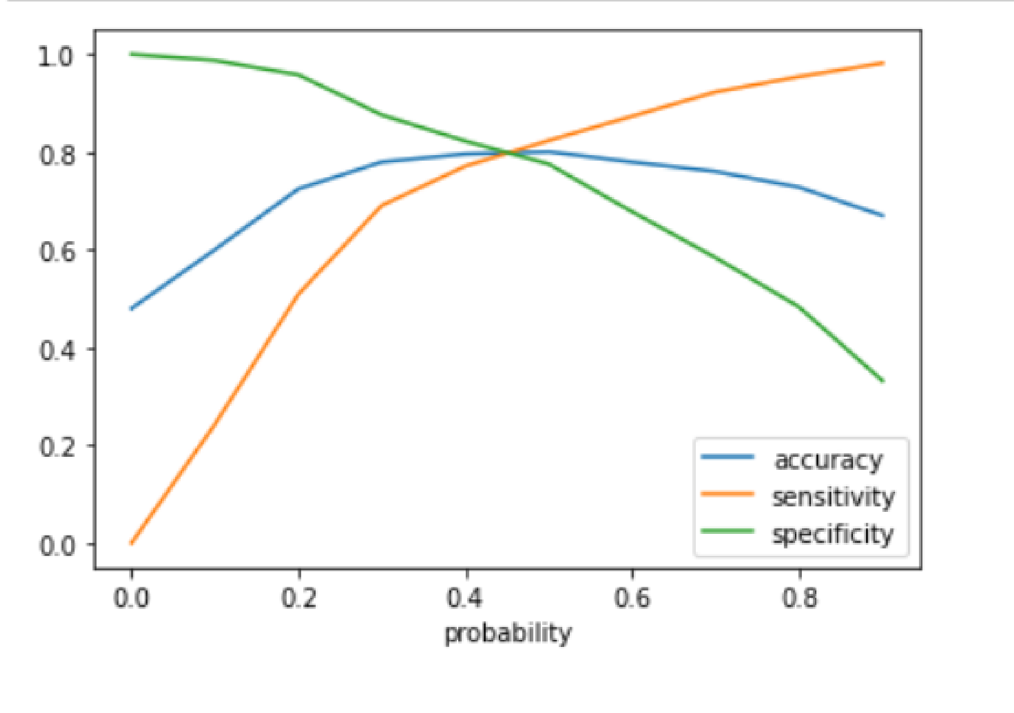


MODEL BUILDING

- Splitting the given dataset into Training and Testing Sets.
- The primary basic step for regression is to perform a train-test split. For this process, we have chosen a ratio of 70:30
- Using RFE for Feature Selection
- Running RFE test with 15 variables as output and building the model by removing all the variables having p-value higher than 0.05 and VIF value greater than 5
- Predictions on test data set
- Overall accuracy of the model is 80.17%

ROC CURVE





Area under ROC Curve is 0.88

Optimum Point of cut-off Probability is 0.45

CONCLUSIONS

- While we have checked both sensitivity- specificity and precision- recall metrics, we have considered optimal cutoff based on sensitivity- specificity for final predication. Accuracy, sensitivity, specificity values for test dataset are round 80.17%, 80.08%, 80.26% respectively.

- Lead score of train and test dataset shows conversion rate for final predicated model is 80%.

INFERENCES

- The top three variables in model that contribute the most probability of a lead getting converted are

  - TotalVisits

  - Total Time spent on website

  - Lead Origin_Lead Add Form.

- The top 3 categorical/dummy variables in the model that increase the probability of lead conversion are

  - Lead Origin_Lead Add Form

  - Lead Source_Welingak website

  - Lead Source_Olark Chat

INFERENCES

- To make the lead conversion more aggressive we need to choose probability cutoff at 0.5 where we can get maximum accuracy 80%, sensitivity 82%, and specificity 77% so that in 82% calls 77% conversion become possible.

- The Company reaches its target for a quarter before the deadline only when to choose probability cutoff at 0.2 where we can get maximum accuracy 72%, sensitivity 50%, and specificity 95% so that in minimum calls (50%) maximum conversion is possible (95%).

THANK YOU!