X Education - Lead Scoring Case Study

Identification of Hot Leads to focus more on them and thus enhancing the conversion ratio for X Education

- Shivangi Singh

Background

X Education Company

X Education , An education company named sells online courses to industry professionals

Many interested professionals land on their website

The company markets its courses on several websites 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

Background

X Education Company

When these people fill up a form providing their email address or phone number, they are classified to be a lead

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 Company's Problem

X EDUCATION GETS A LOT OF LEADS BUT ITS LEAD CONVERSION RATE IS VERY POOR

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 Company's Problem

WE WILL HELP THEM TO SELECT THE MOST PROMISING LEADS, I.E. THE LEADS THAT ARE MOST LIKELY TO CONVERT INTO PAYING CUSTOMERS.

WE ARE REQUIRED 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

THE CEO, IN PARTICULAR, HAS GIVEN A BALLPARK OF THE TARGET LEAD CONVERSION RATE TO BE 80%.

Proposed Solution

Selection of Hot Leads

Communicating with Hot Leads

Conversion of Hot Leads

Leads Clustering

We cluster the leads into certain categories based on their tendency or probability to convert, thus, getting a smaller section of hot leads to focus more on.

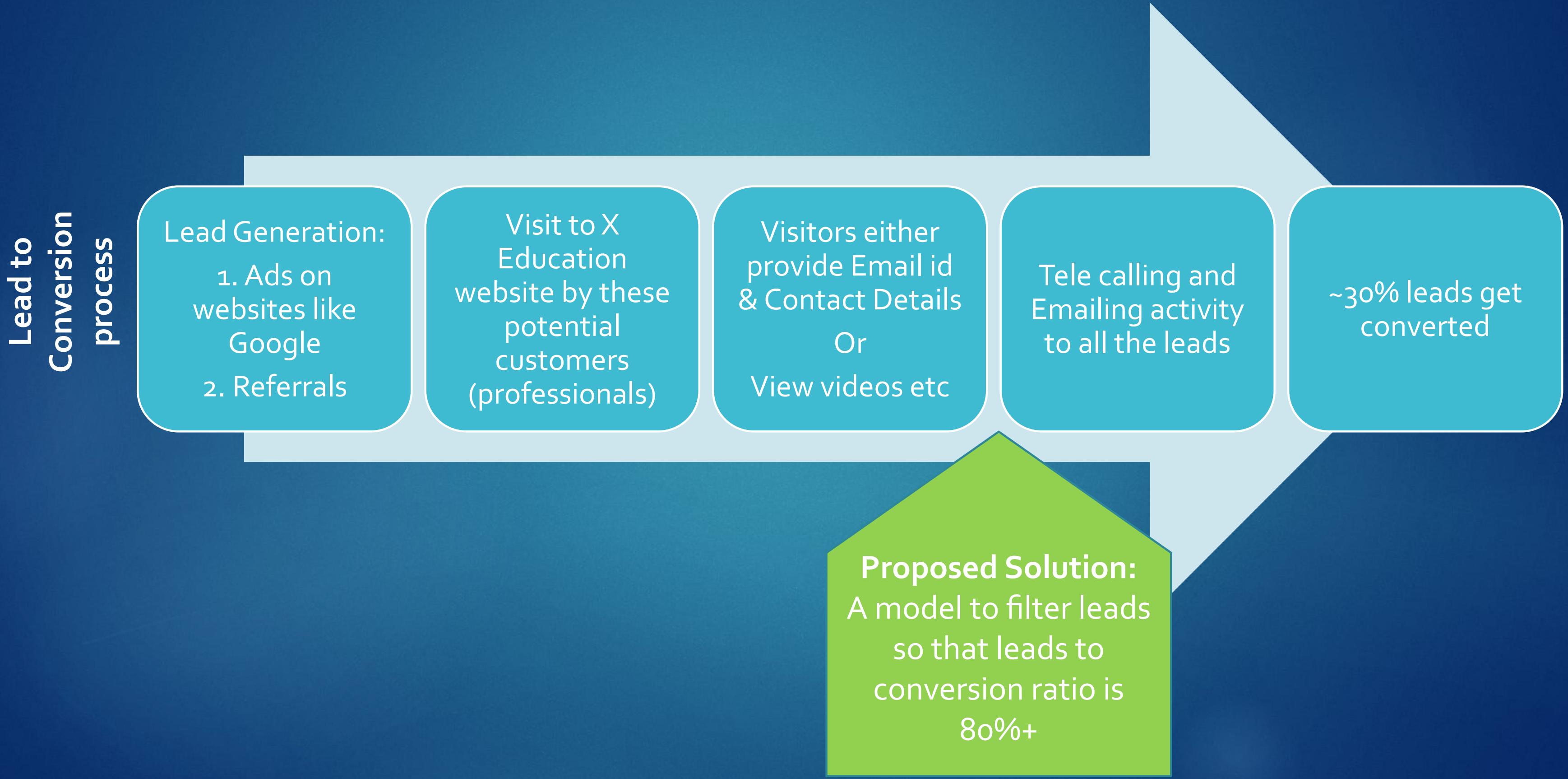
Focus Communication

Since we would have a smaller set of leads to have communication with, we might make more impact with effective communication.

Increase conversion

Since we focussed on hot leads, which were more probable to convert, we would have a better conversion rate, and hence we can achieve the 80% target.

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 (professionals)

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%+

Solution

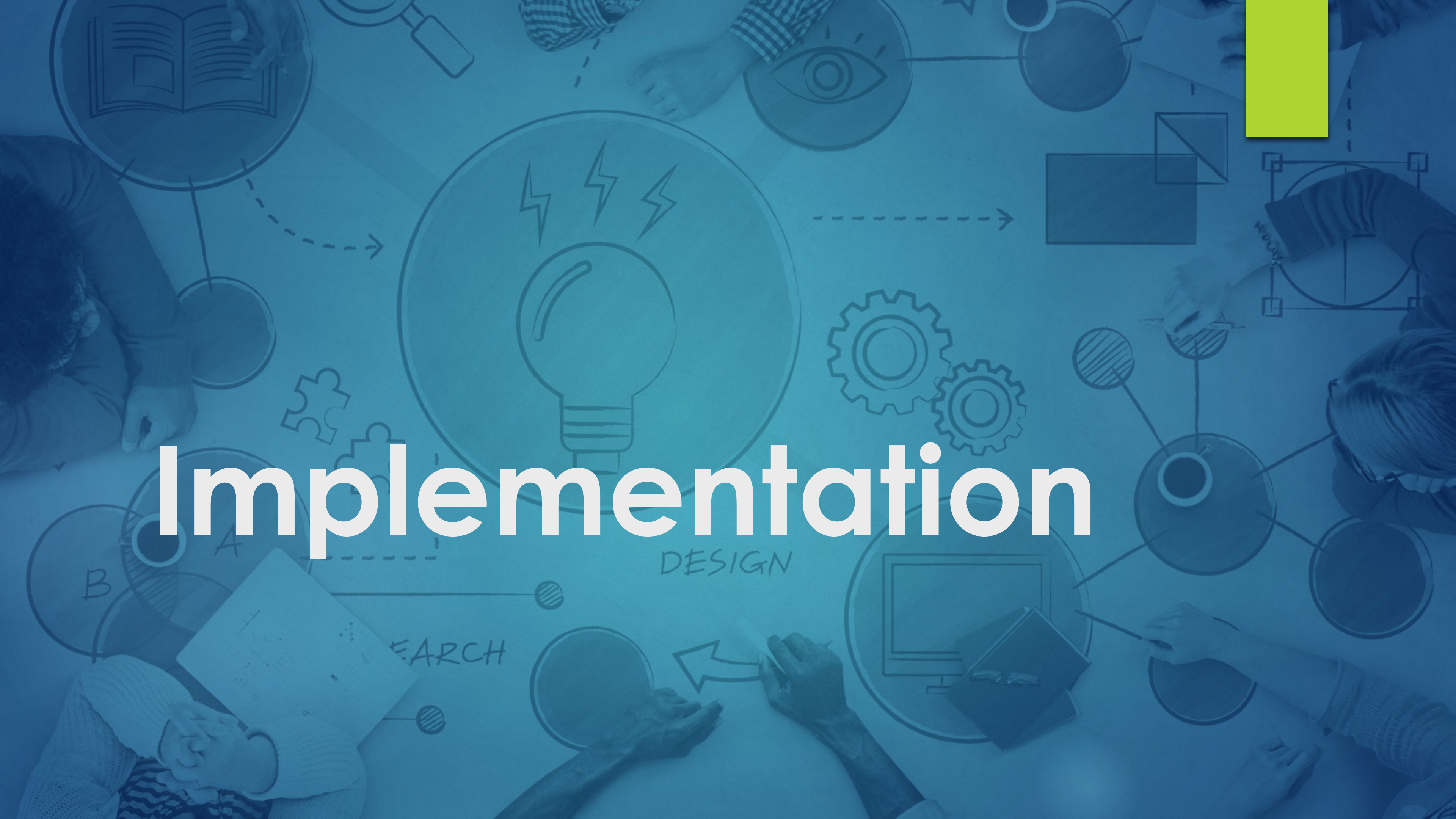
Selection of Hot Leads

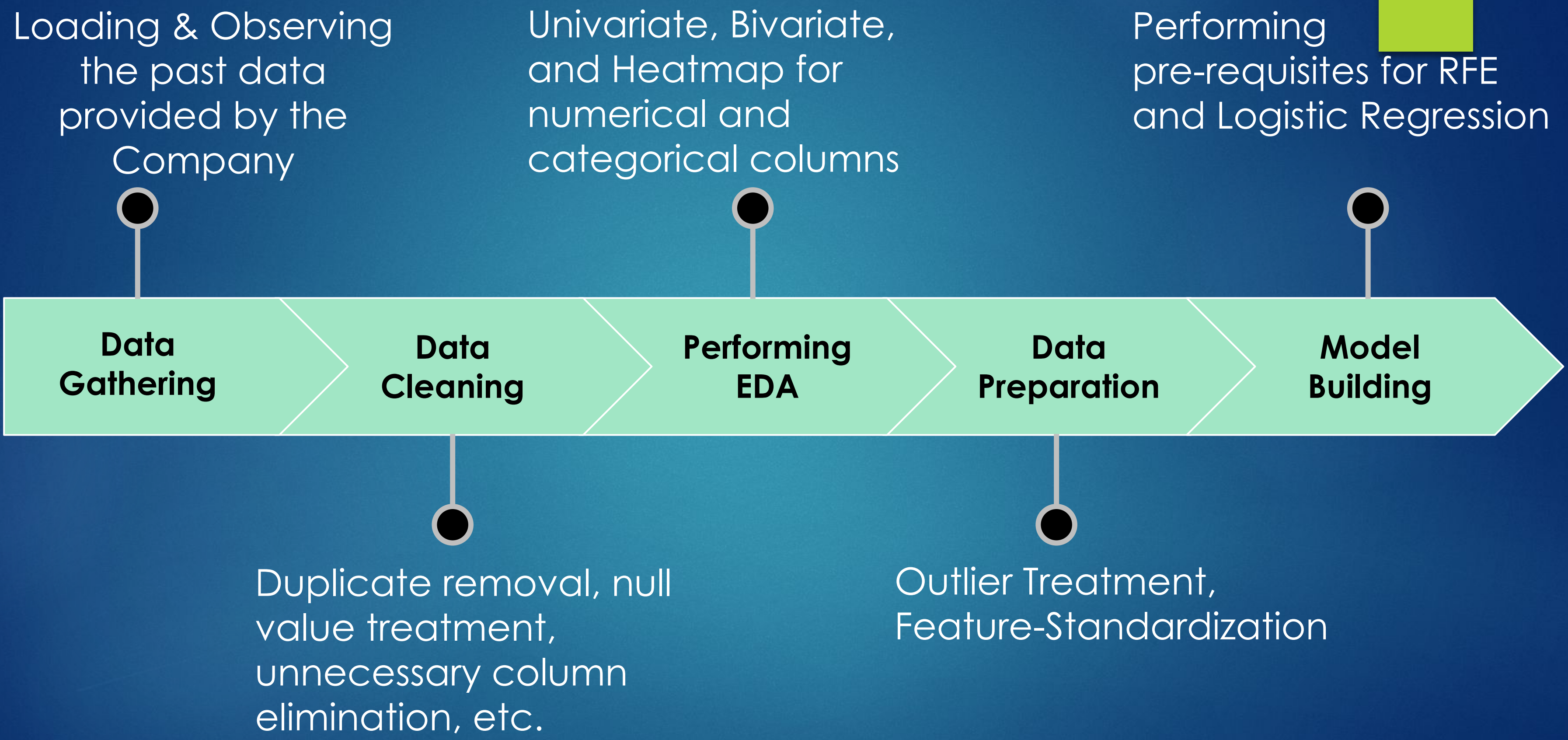
FOR OUR PROBLEM SOLUTION, THE CRUCIAL PART IS TO ACCURATELY IDENTIFY HOT LEADS.

THE MORE ACCURATE WE OBTAIN THE HOT LEAD, THE MORE CHANCE WE GET OF HIGHER CONVERSION RATIO.

SINCE WE HAVE A TARGET OF 80% CONVERSION RATE, WE WOULD WANT TO OBTAIN A HIGH ACCURACY IN OBTAINING HOT LEADS.

Implementation





Data Gathering

Loading & Observing the past data provided by the Company

Data Cleaning

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

Performing EDA

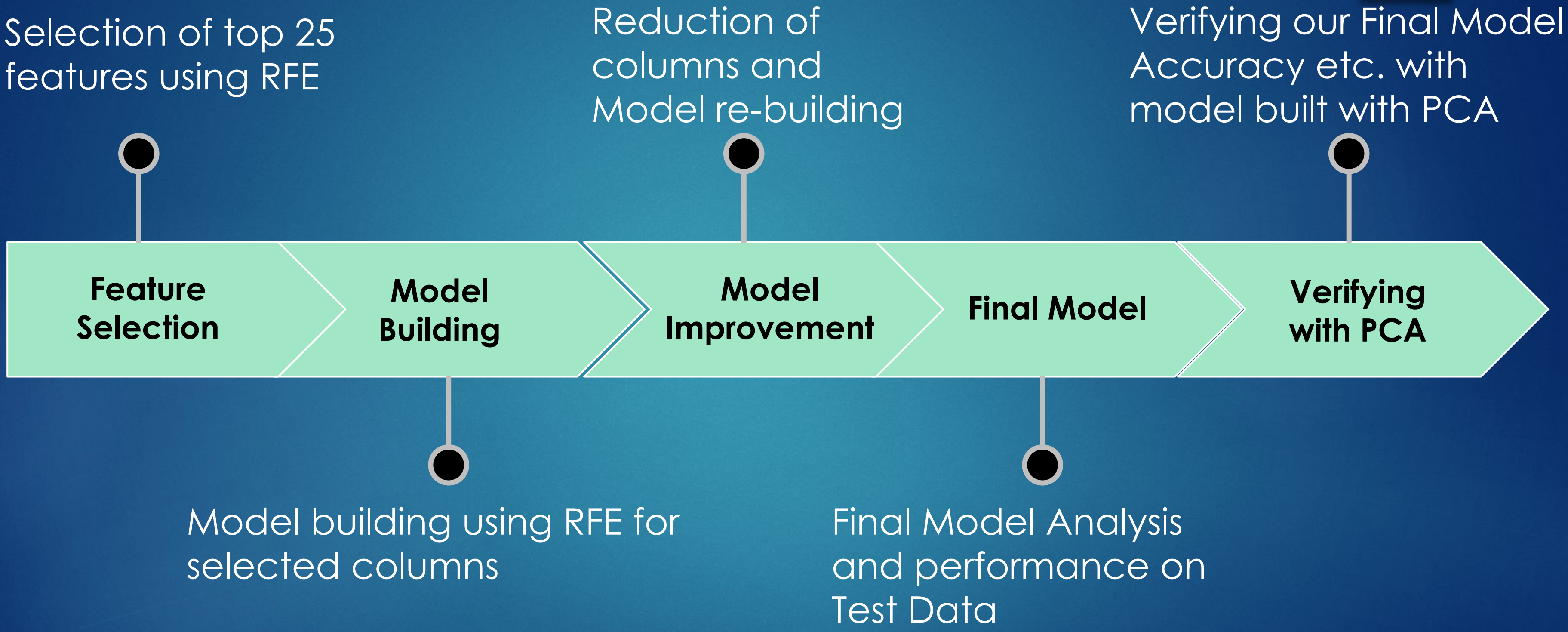
Univariate, Bivariate, and Heatmap for numerical and categorical columns

Data Preparation

Outlier Treatment, Feature-Standardization

Model Building

Performing pre-requisites for RFE and Logistic Regression



Feature Selection

Selection of top 25 features using RFE

Model Building

Model building using RFE for selected columns

Model Improvement

Reduction of columns and Model re-building

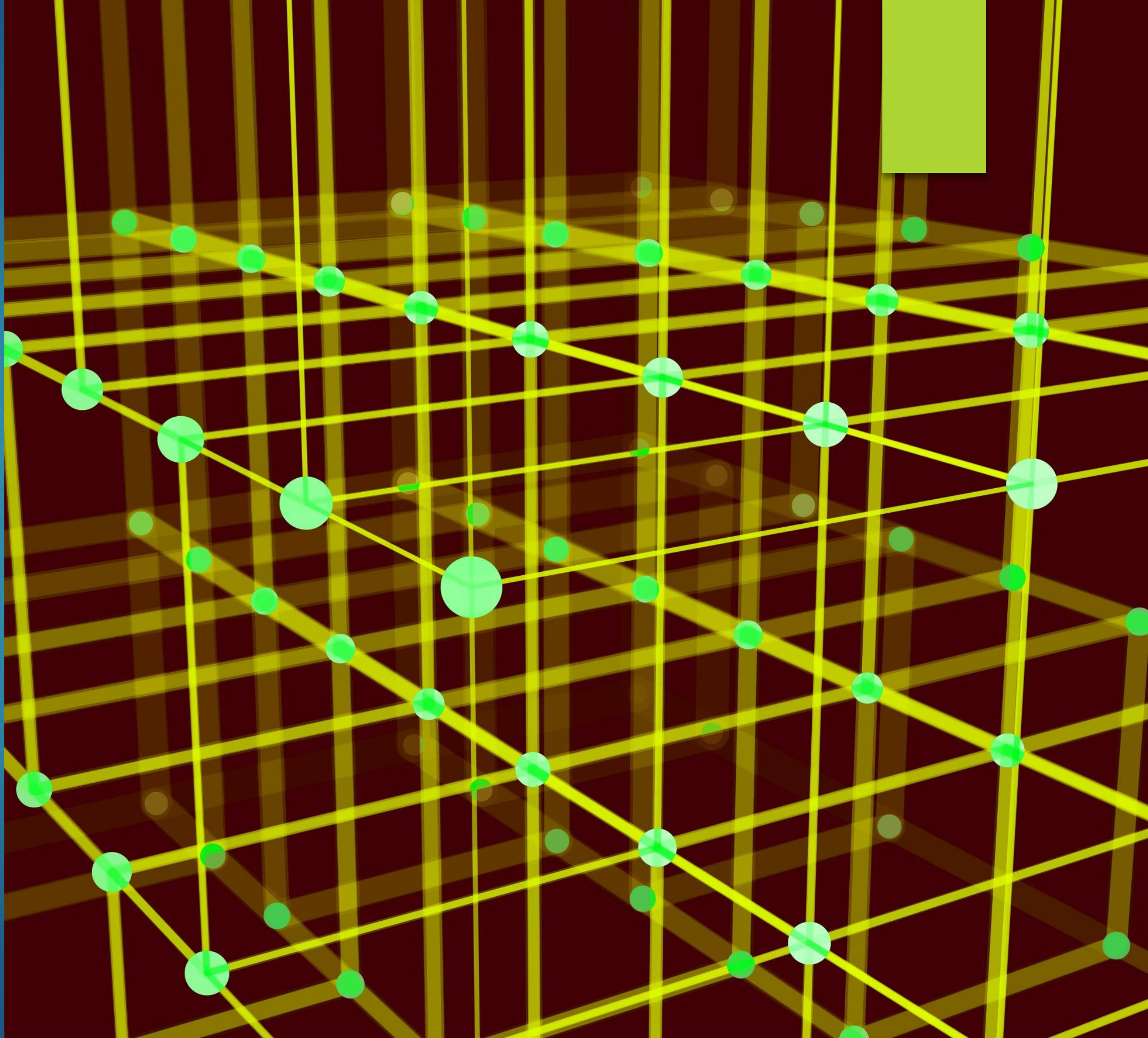
Final Model

Final Model Analysis and performance on Test Data

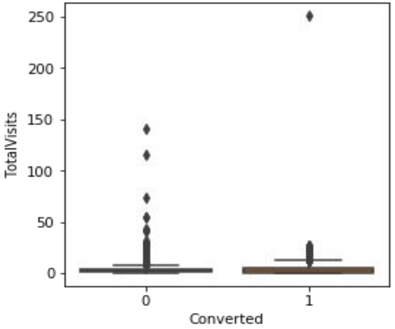
Verifying with PCA

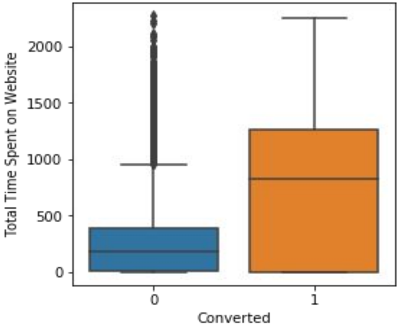
Verifying our Final Model Accuracy etc. with model built with PCA

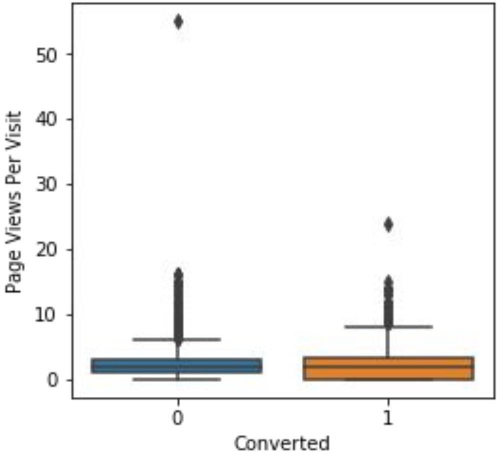
Plots
(Visualization)



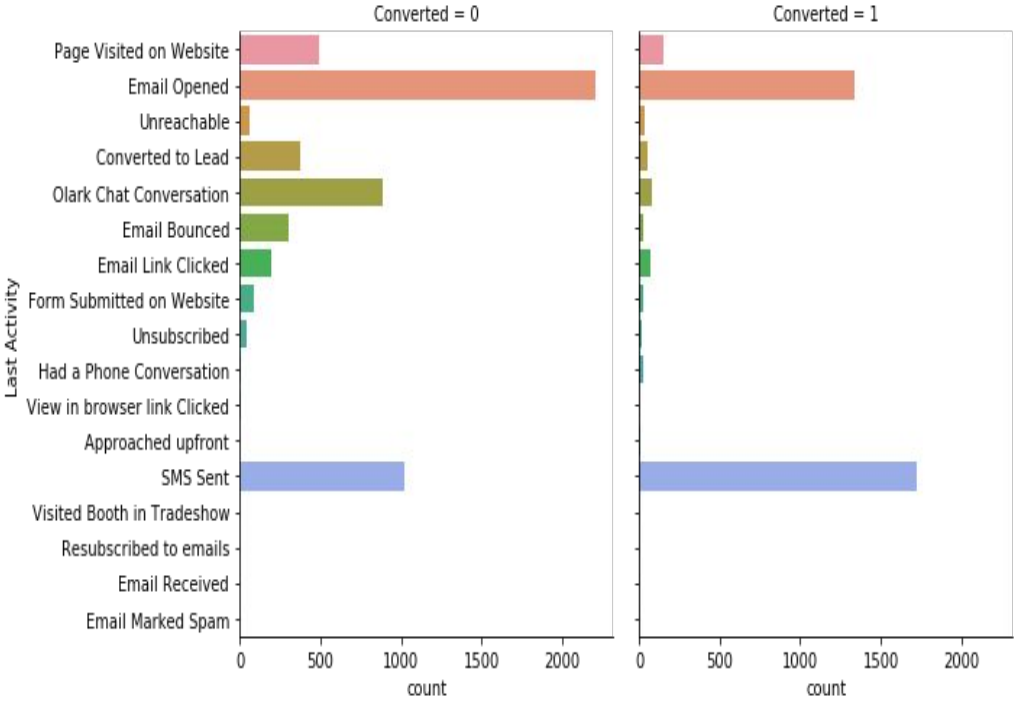
- EDA plots depicting variation in numerical columns for those who converted and those who didn't.



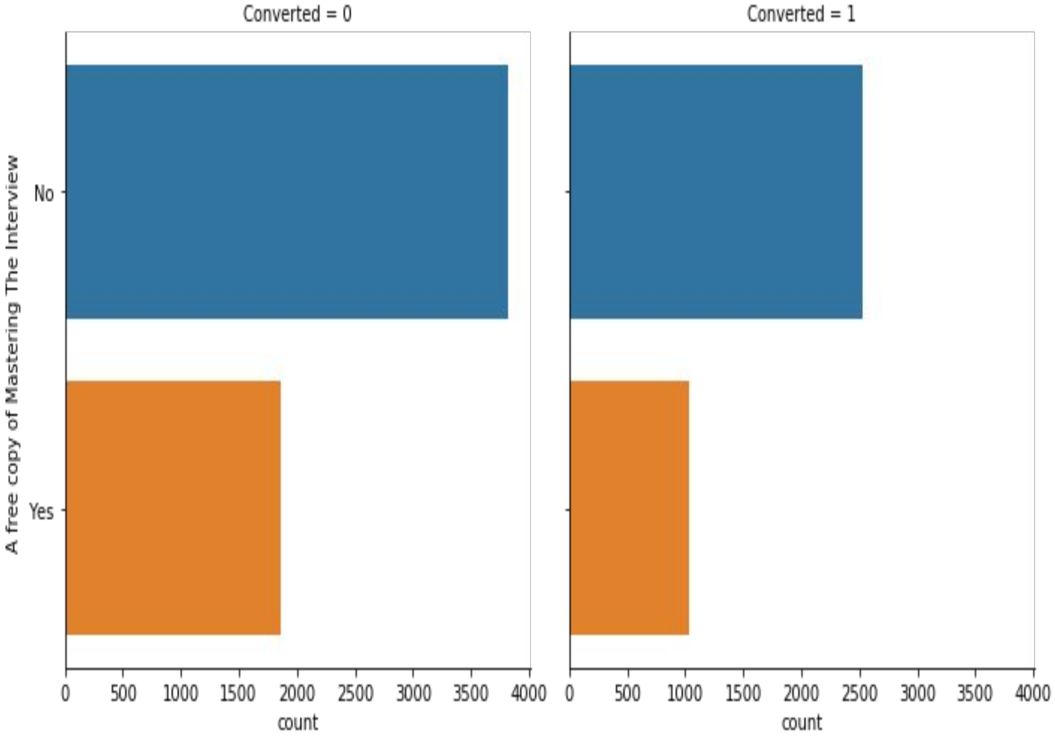




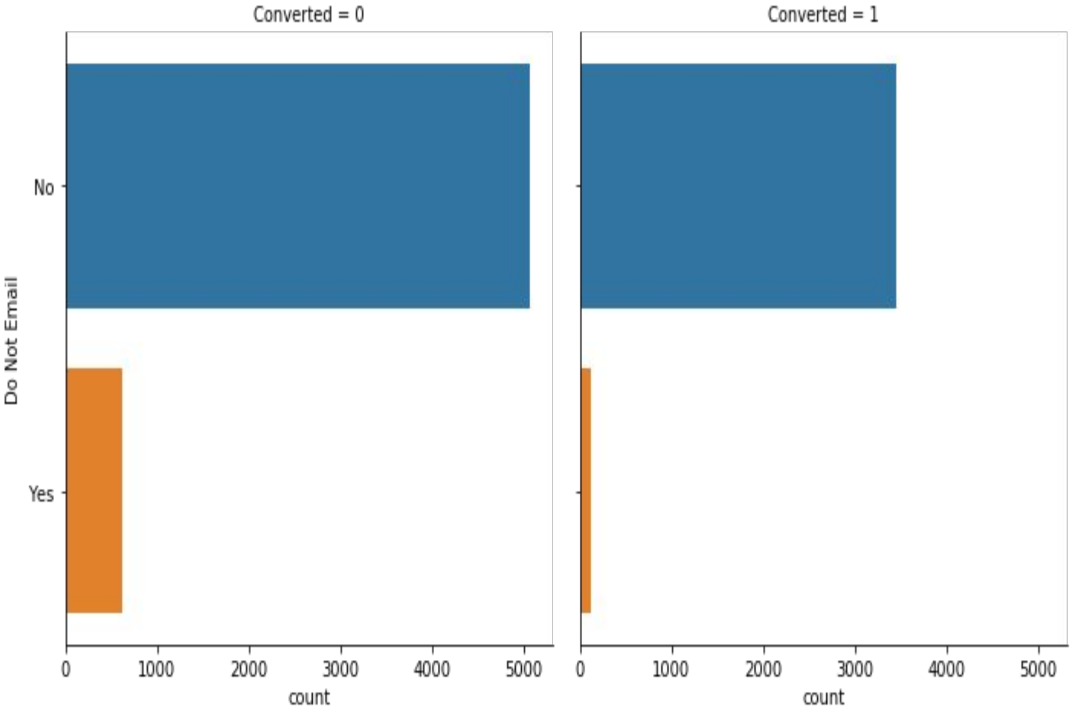
- EDA plots depicting variation in categorical column (Last Activity) for those who Converted and those who didn't.



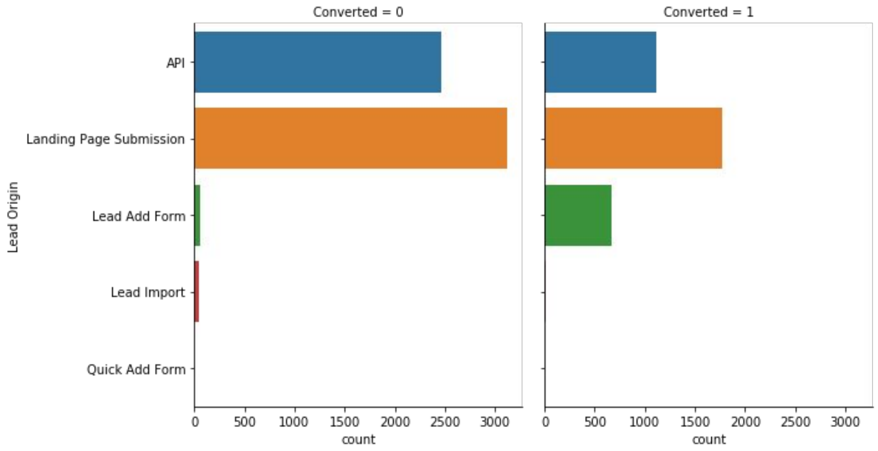
- EDA plots depicting variation in categorical column (A free copy of Mastering The Interview) for those who Converted and those who didn't.



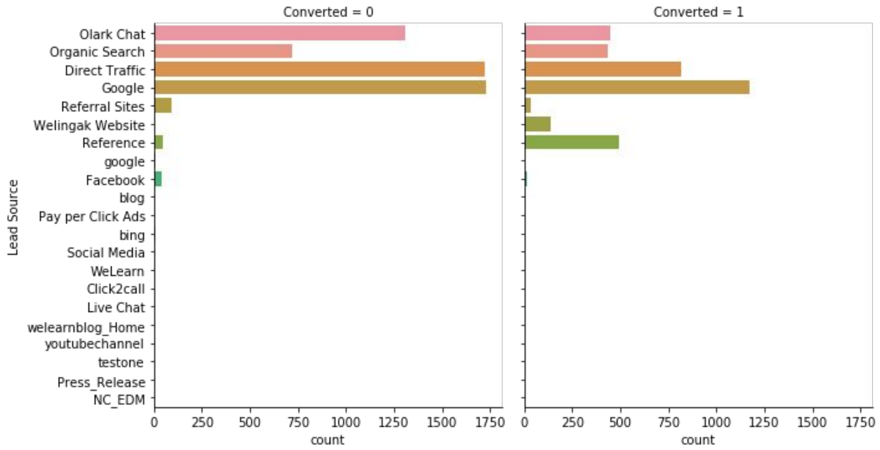
- EDA plots depicting variation in categorical column (Do Not Email) for those who Converted and those who didn't.



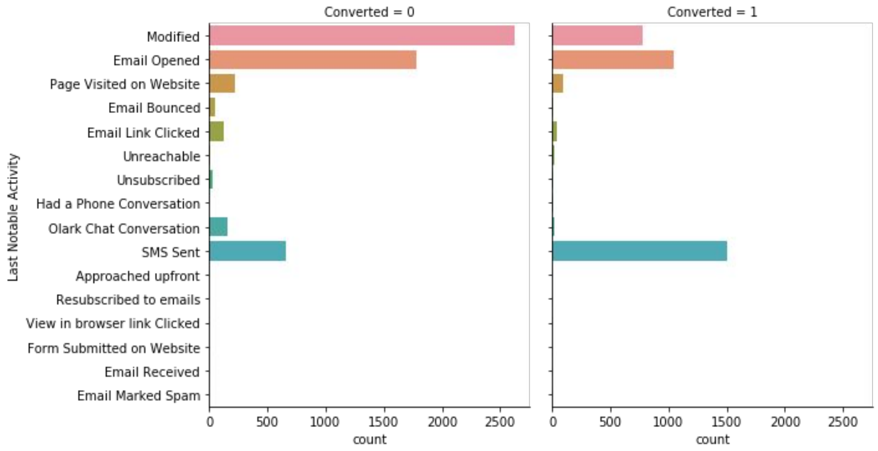
- EDA plots depicting variation in categorical column (Lead Origin) for those who Converted and those who didn't.

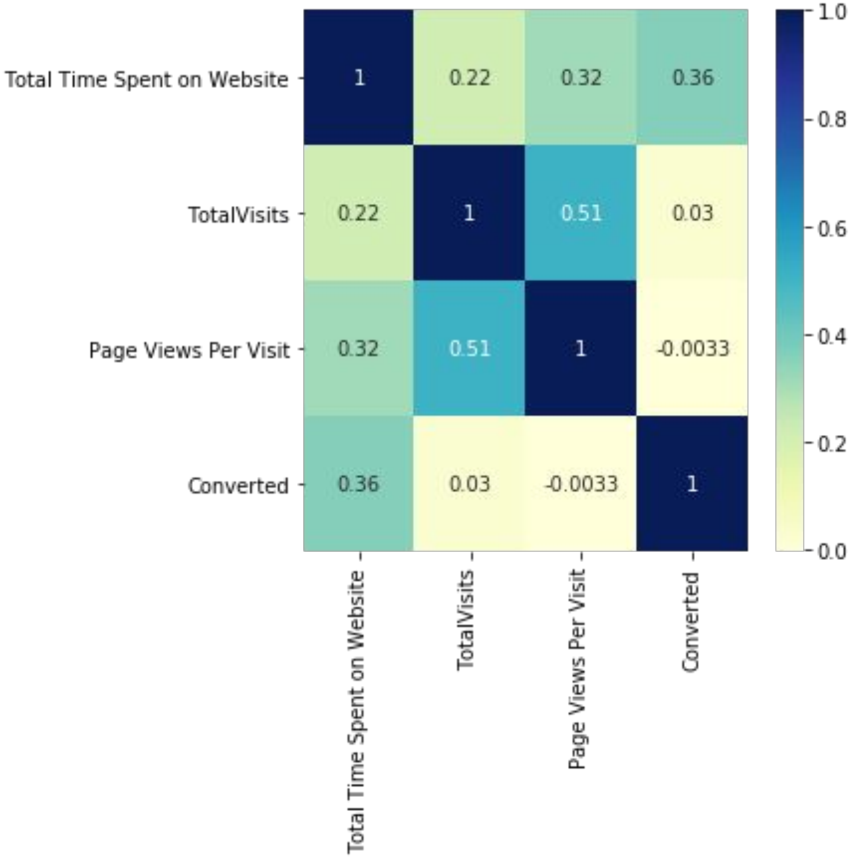


- EDA plots depicting variation in categorical column (Lead Source) for those who Converted and those who didn't.



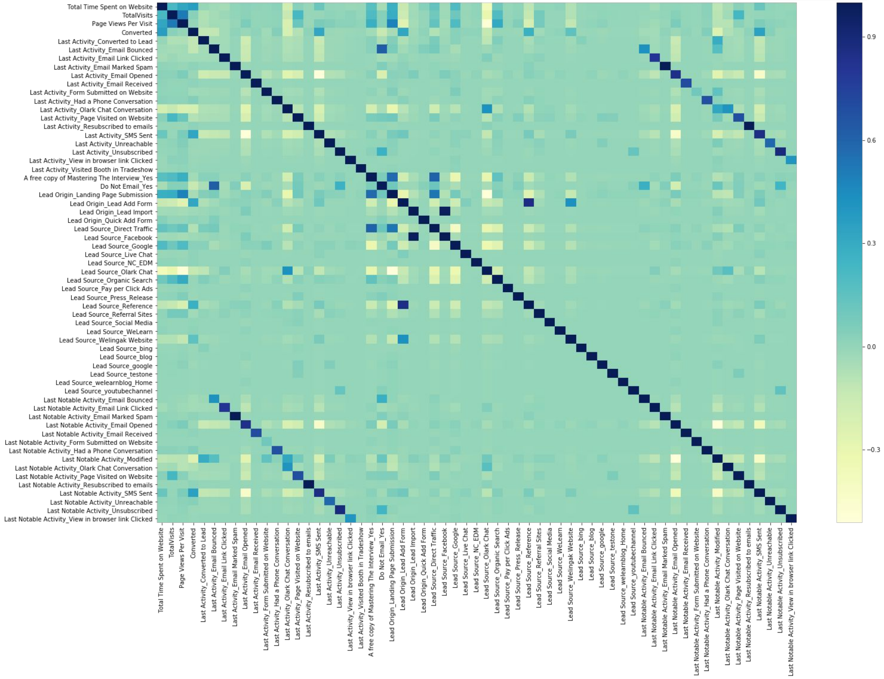
- ►EDA plots depicting variation in categorical column (Last Notable Activity) for those who Converted and those who didn't.



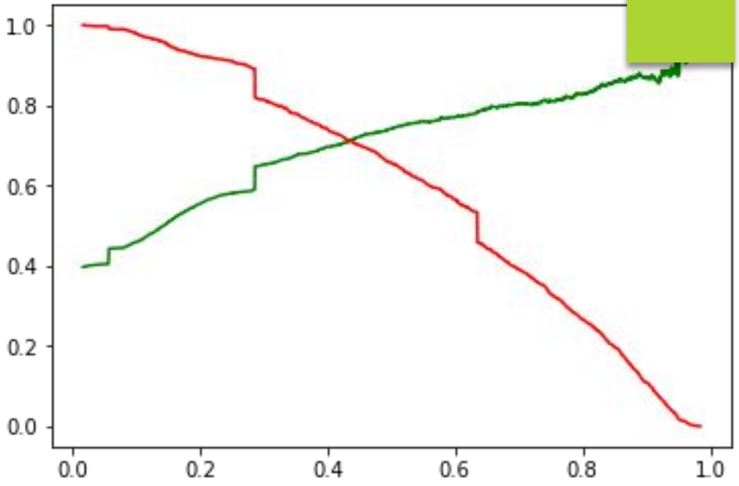


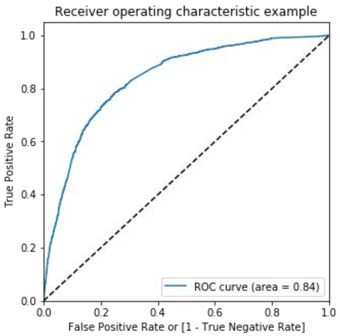
EDA plots depicting correlation (Heat Map) of all selected numerical columns.

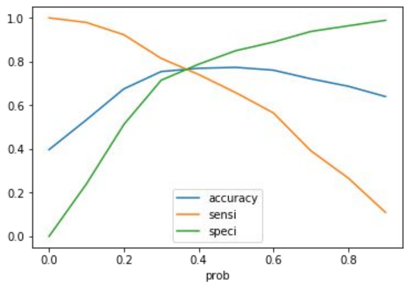
- EDA plots depicting correlation (Heat Map) of all selected columns (numerical columns and dummy columns).



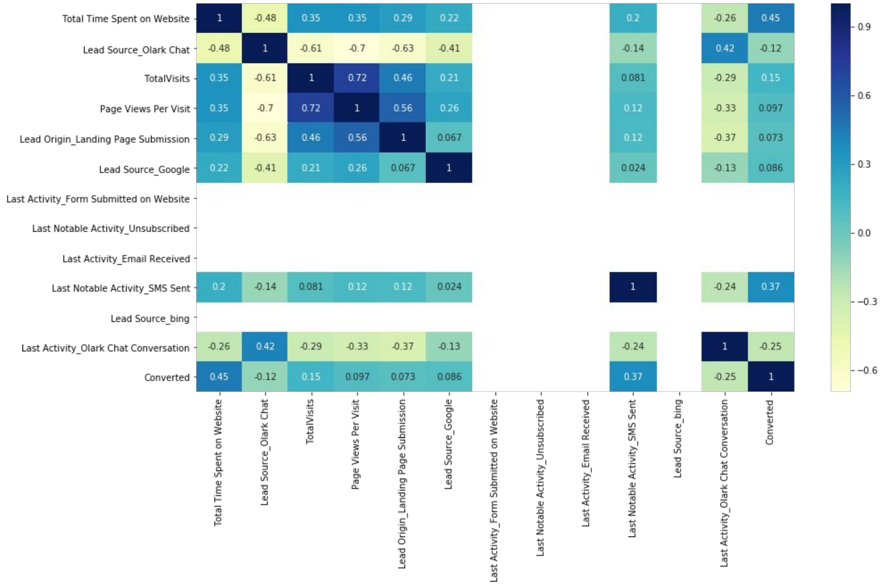
- Linear Regression Final Model Parameters
- Area under ROC = 0.84
- Intermediate cut-off = 0.35
- Final cut-off = 0.42

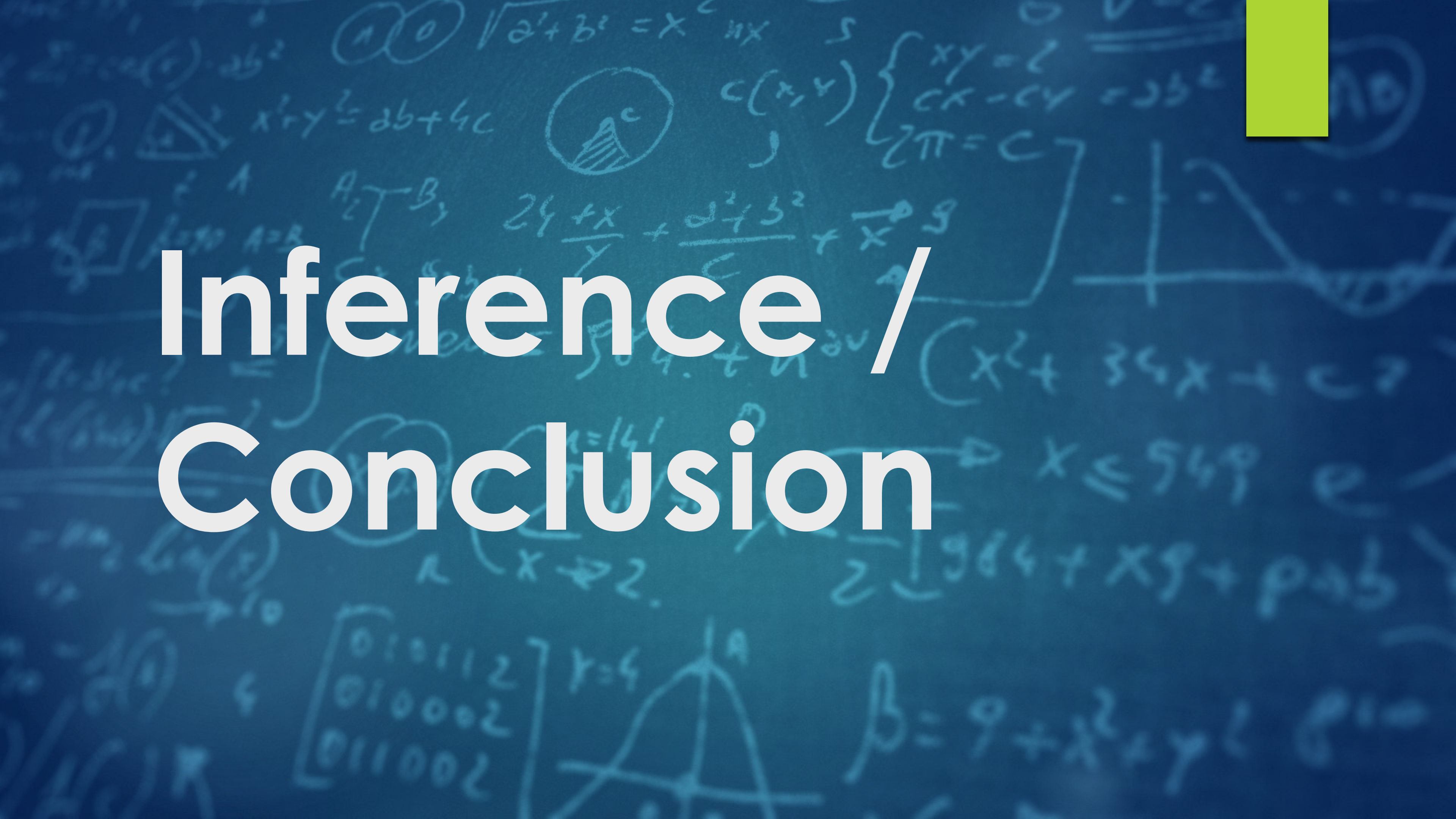






- EDA plots depicting correlation (Heat Map) of all selected columns (numerical columns and dummy columns) in our final Model.





Inference / Conclusion

Model Analysis

Performance of our Final Model

Overall accuracy on Test set: 0.786

Sensitivity of our logistic regression model: 0.733

Specificity of our logistic regression model: 0.823

Inferences from Model

Business Insights Derived from our Model

Top 3 variables in the model, that contribute towards lead conversion:

1. Total Time Spent on Website
2. Last Notable Activity_SMS Sent
3. Total Visits

Inferences from Model

Business Insights Derived from our Model

Top 3 variables in my model, that should be focused are:

1. Last Activity_SMS Sent (positively impacting)
2. Last Activity_Olark Chat Conversation (negatively impacting)
3. Lead Source_Olark Chat (negatively impacting)

Conclusion 1 (LR Model)

Our Logistic Regression Model is decent and accurate enough, when compared to the model derived using PCA, with 78.6 % Accuracy on Test Set, 73.3 % Sensitivity and 82.3 % Specificity.

We can vary these parameters by varying the cut-off value and thus predict Hot leads based on scenarios like availability of extra resources and vice-versa.

Conclusion 2 (Recommendation)

Focus on	X Education Company needs to focus on following key aspects to improve the overall conversion rate:
Increase	Increase user engagement on their website since this helps in higher conversion
Increase on	Increase on sending SMS notifications since this helps in higher conversion
Get	Get Total visits increased by advertising etc. since this helps in higher conversion
Improve	Improve the Olark Chat service since this is affecting the conversion negatively