Lead Score Case Study

Group Members

1. Vinod Kumar
2. Srivalli Akella
3. Subham Pant

Problem Statement

- X Education Company Sells online courses to industry professionals.
- X Education company 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.

Business Goal

- X education company wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

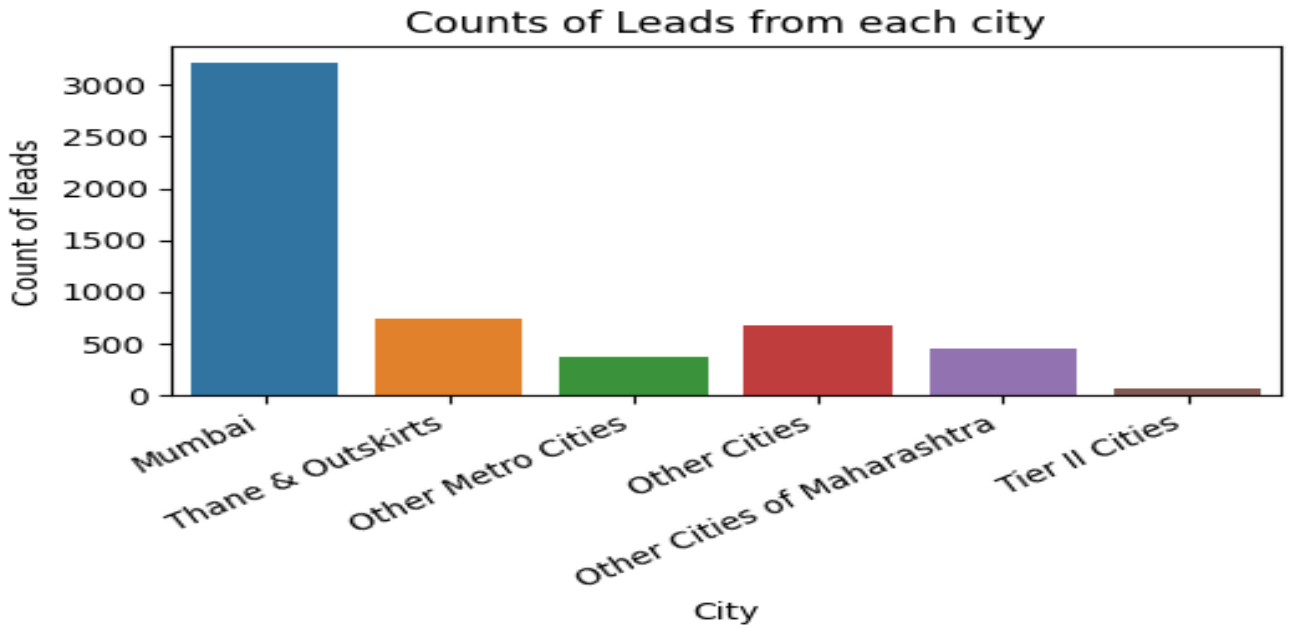
Method of Problem Solving

- Importing of Data.
- Data Preparation and Cleaning of the data for further analysis.
- Exploratory Data Analysis: Univariate Analysis, Bivariate Analysis.
- Feature Scaling and Dummy Variables creation.
- Building a Logistic Regression model a classification Technique used for Making and Prediction.
- Test the Model on Train and Test Split Using RFE Approach.
- Validation and Model Presentation.
- Model Interpretation and Conclusion.

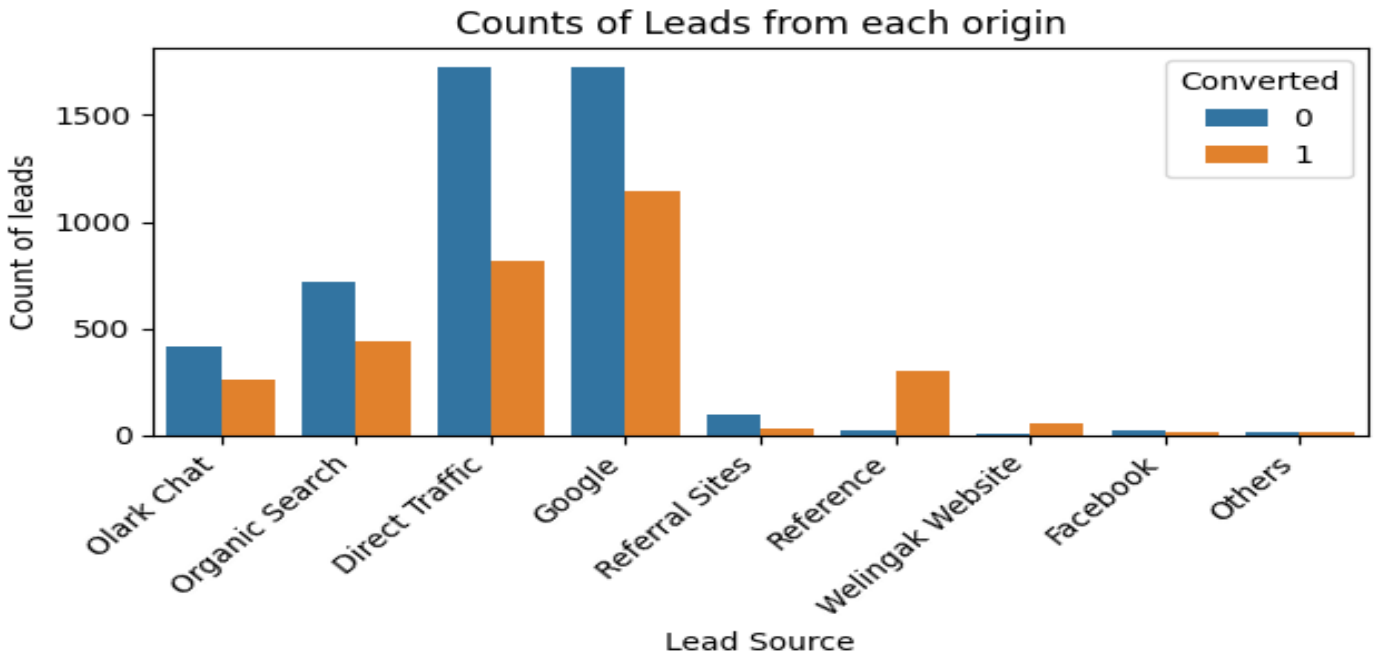
Data Understanding and EDA

- The size of the Data Set is (9240, 37).
- Number of Rows = 9240, Number of Columns = 37.
- Dropping ['Prospect ID', 'Lead Number’] which are not helpful for the analysis.
- Dropping Duplicates also, if any.
- Treating the Columns that has null values.

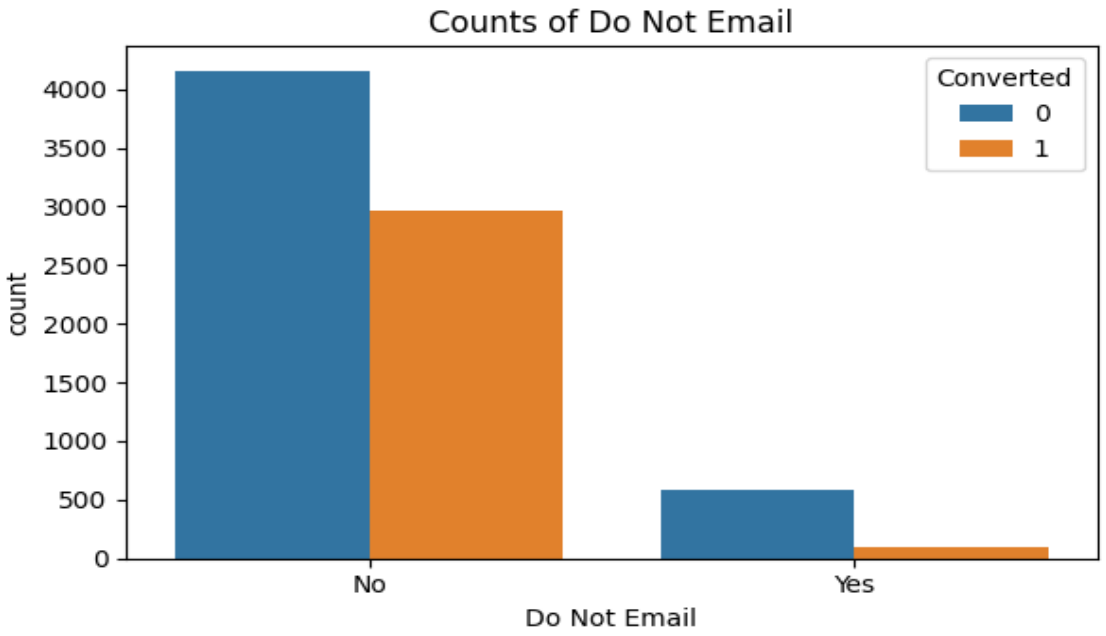
EDA and Categorical Variable Relation

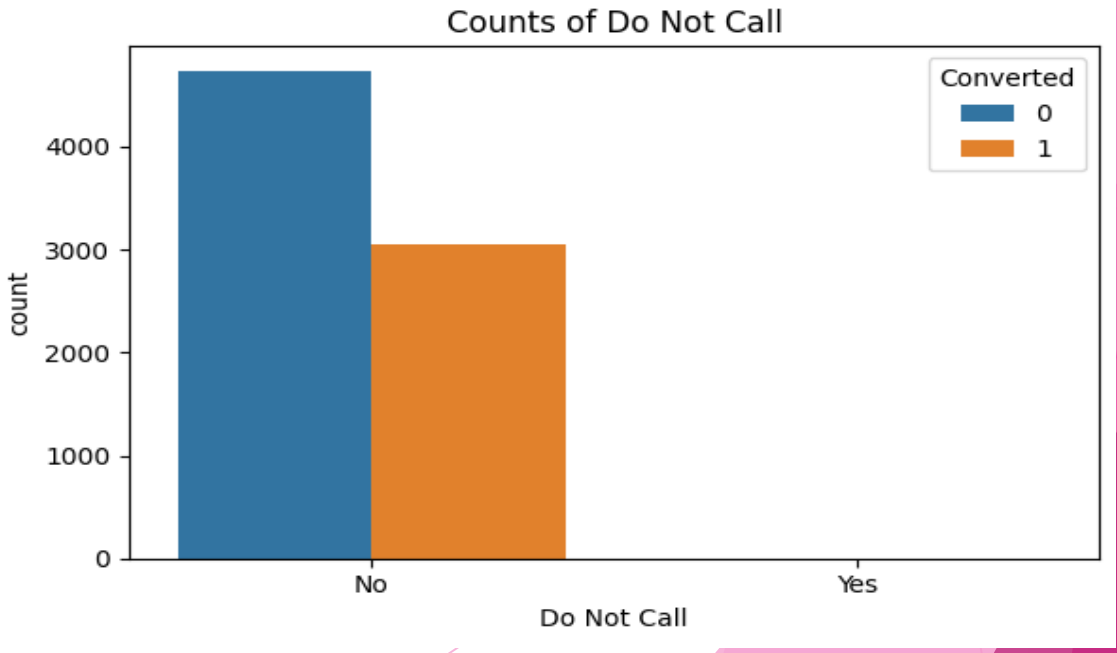


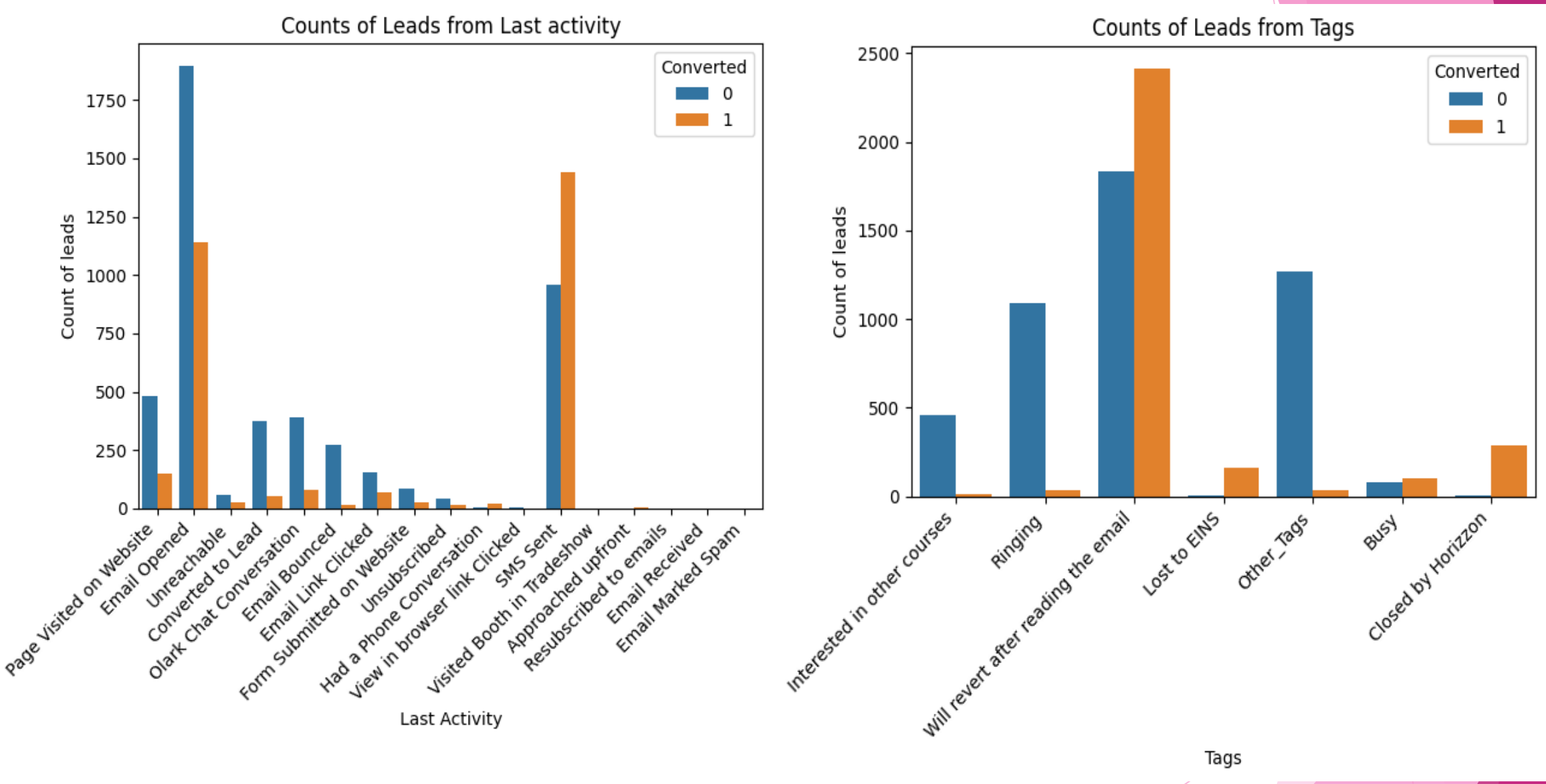
- Out of all the cities, Mumbai is the big market for customers to the company.



- Google & Direct search are the main sources of leads origin to the company.
- Out of all the leads, who opted for email & call communication results to lead conversion eventually.

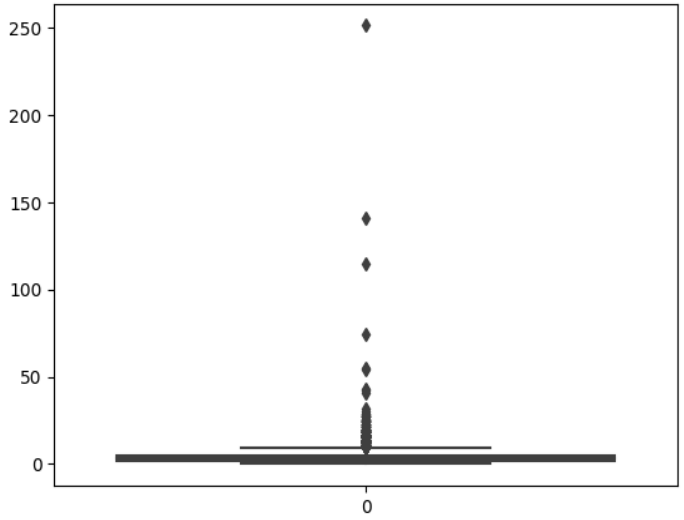


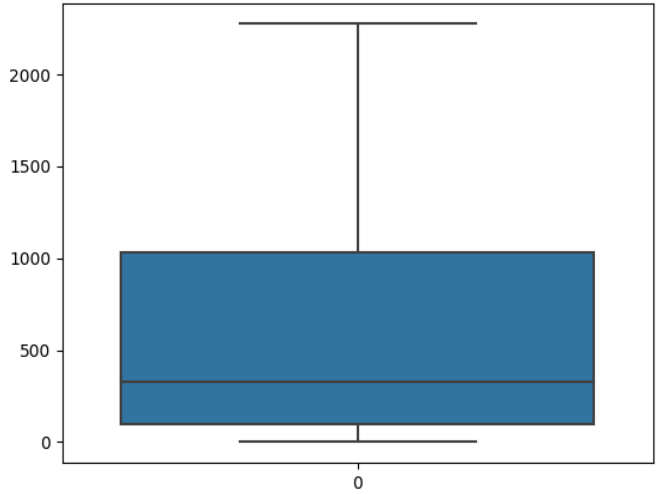


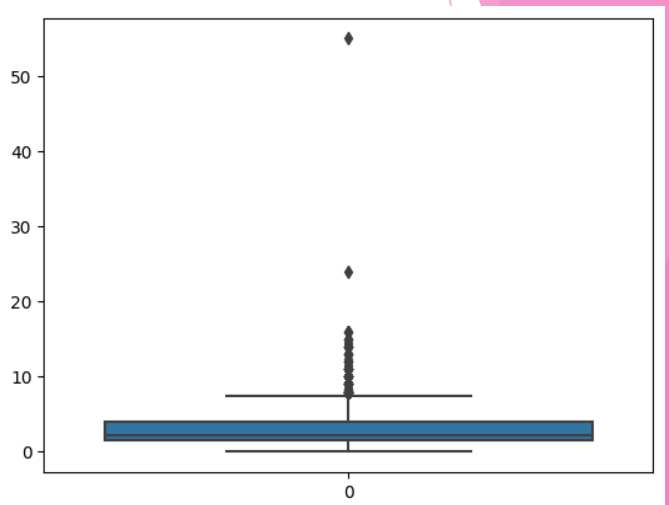


- Leads who open the email and receive the SMS are most likely to convert.
- Also Leads who will respond to the email results to the most potential lead conversion.

Box Plot

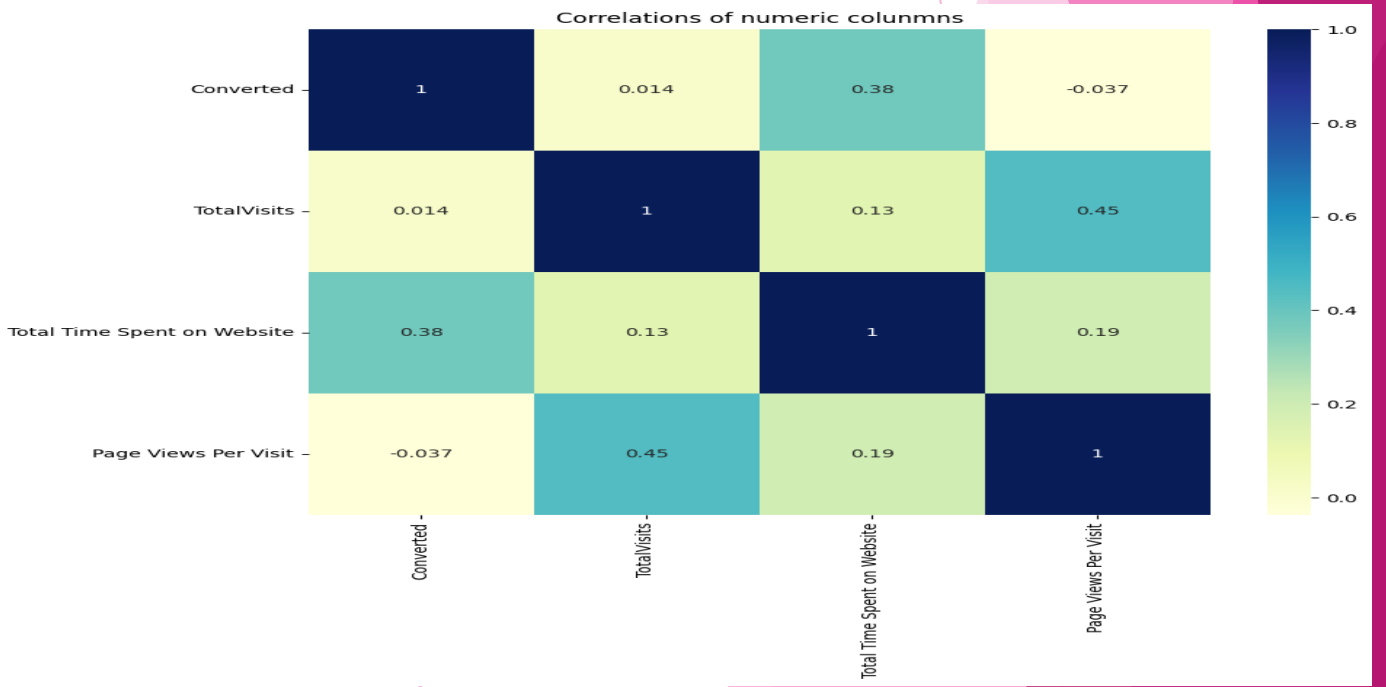






Correlation

- Leads who spend majority of time on website tend to convert the most.



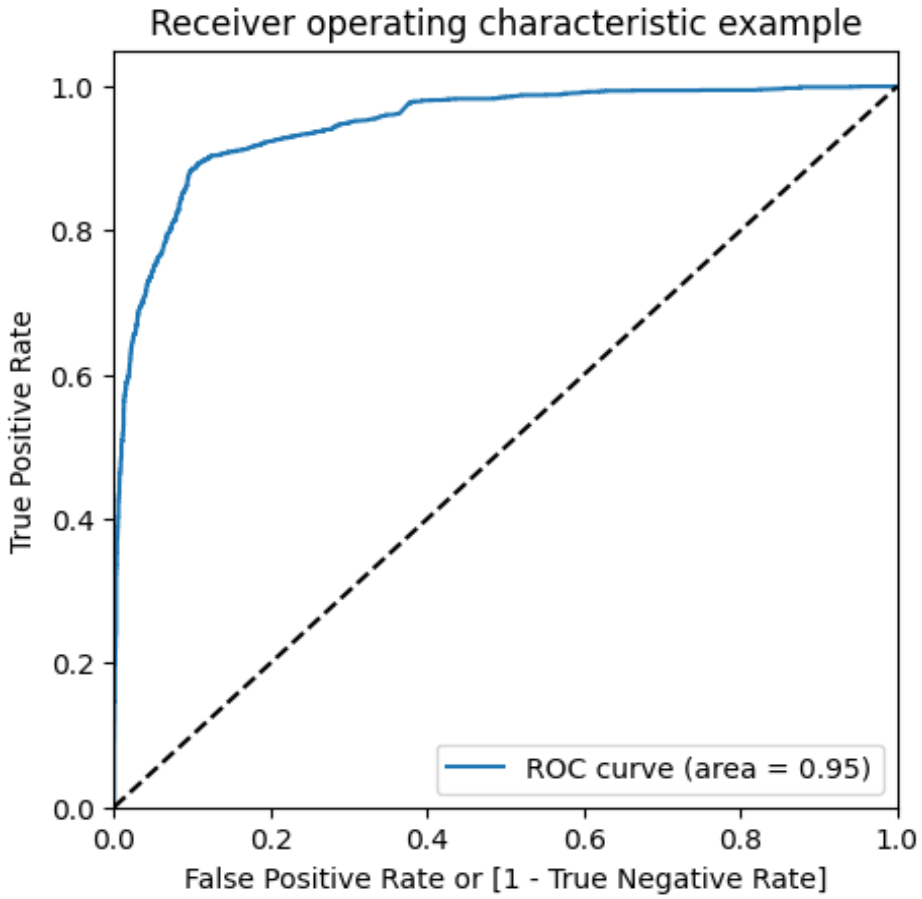
Data Conversion

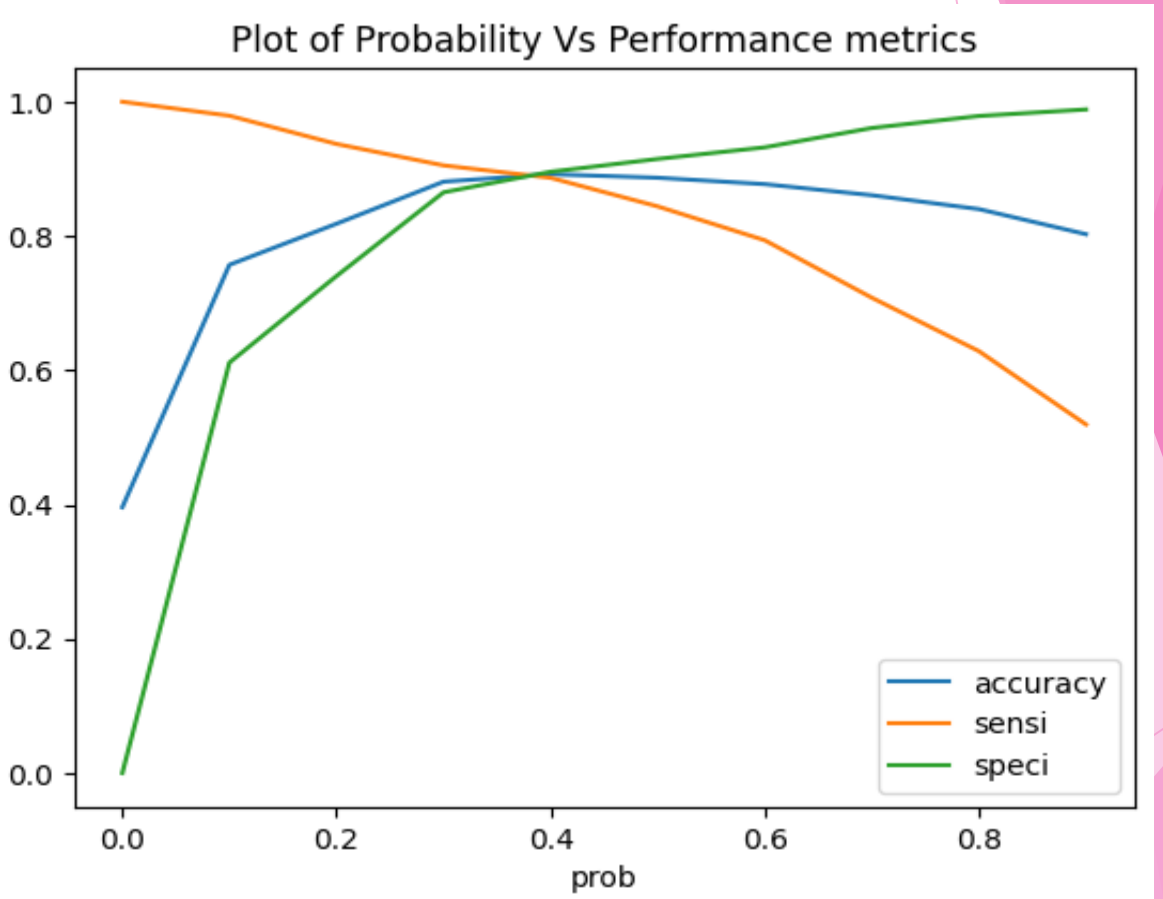
- Numerical Variables are Normalised.
- Dummy Variables are created for object type variables.
- Total Rows for Analysis: 7796
- Total Columns for Analysis: 28

Model Building

- Splitting the Data into Training and Testing Sets.
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection.
- Running RFE with 15 variables as output.
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5.
- Area under the ROC curve usually quantifies the model accuracy in case of classification.
- Predictions on test data set.
- Overall accuracy 88.67%.

ROC Curve





- At 0.4 probability, categorizing the leads gives much better accuracy.
- So any lead with score > 0.4 acts as a potential lead.

ROC Curve

Finding Optimal Cut off Point:

- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.4

Model Interpretation

- After running the model on the Test Data these are the figures we obtain:
  - **Accuracy** : 89.20%
  - **Sensitivity** : 88.66%
  - **Specificity** : 89.56%
- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- Hence overall this model seems to be good.

Suggestions & Recommendations

Company can spend optimizing below parameters to get the most leads into conversion:

- The total time spend on the Website.
- Total number of visits.
- When the lead source was:
  - Google
  - Direct traffic
  - Organic search
  - Welingak website
- When the last activity was:
  - SMS
  - Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

THANK YOU