

LEAD SCORE CASE STUDY

Team Member

❖ Aditi Deshpande

❖ Akshay Bankar

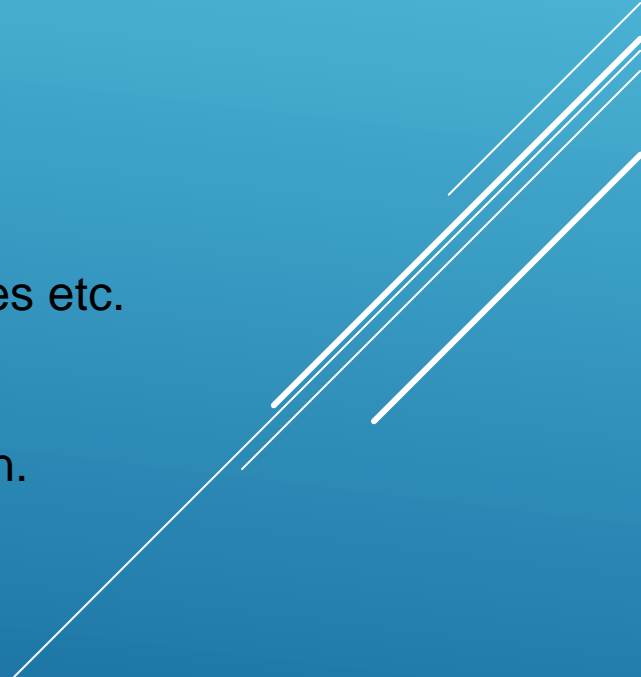
PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google. Through marketing company 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.

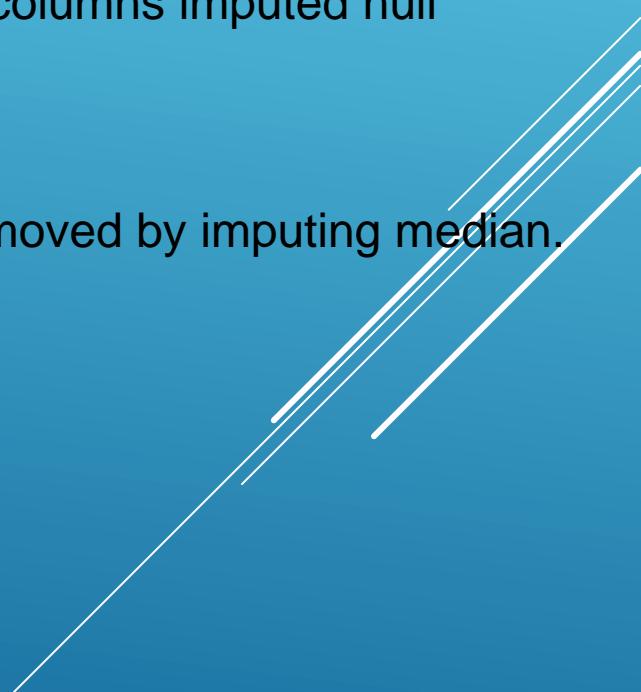
Business Objective

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads and provide lead score to each lead

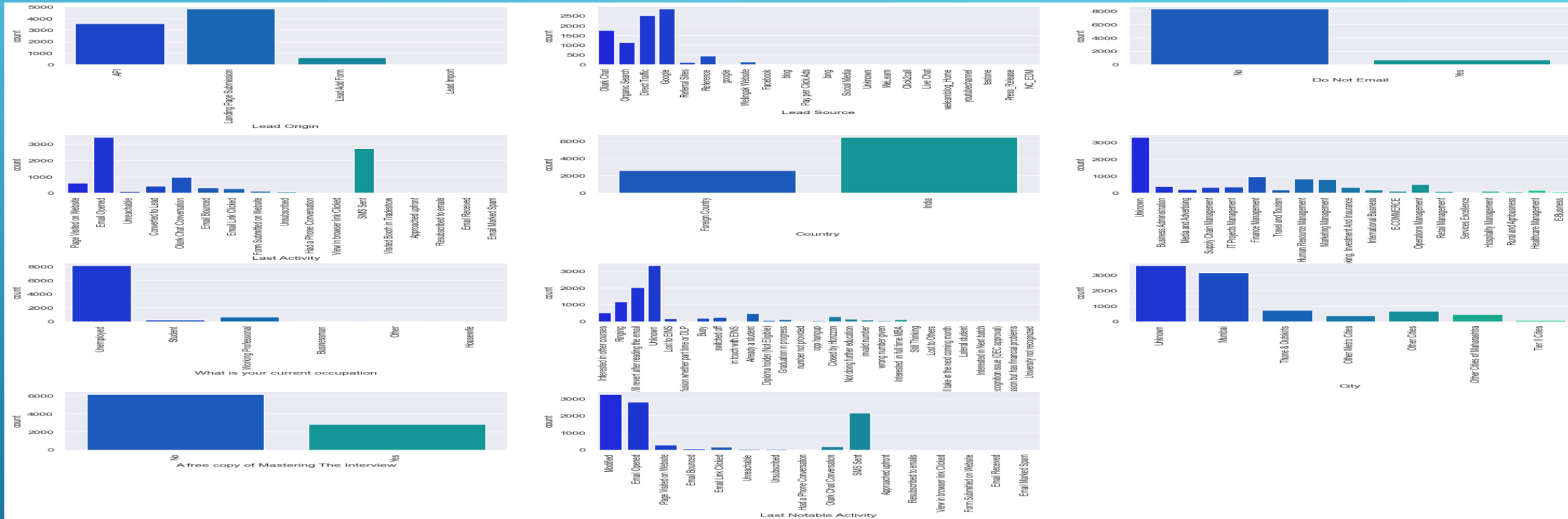
SOLUTION METHODOLOGY

- Data cleaning and data manipulation.
 1. Check and handle duplicate data.
 2. Check and handle NA values and missing values.
 3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
 4. Imputation of the values, if necessary.
 5. Check and handle outliers in data.
 - EDA
 1. Univariate data analysis: value count, distribution of variable etc.
 2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
 - Feature Scaling & Dummy Variables and encoding of the data.
 - Classification technique: logistic regression used for the model making and prediction.
 - Validation of the model.
 - Model presentation.
 - Conclusion.
- 

DATA MANIPULATION

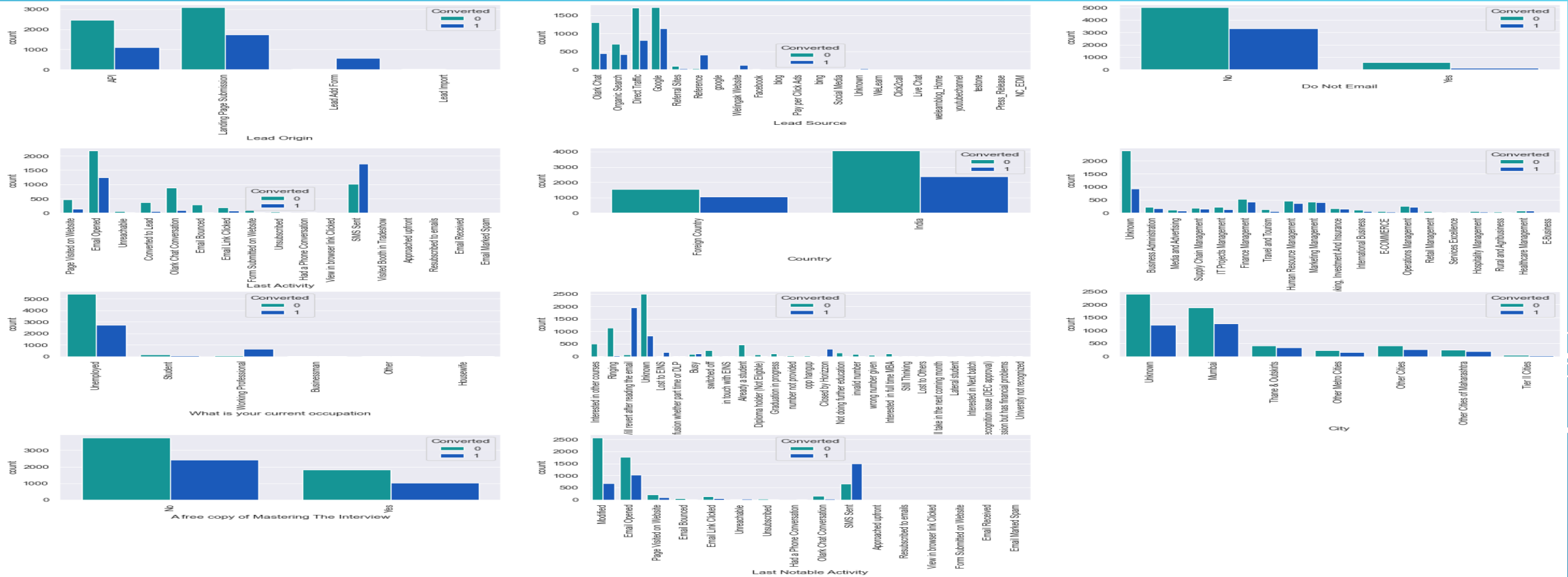
- Total Number of Rows =37 and Total Number of Columns =9240.
 - Single value features like 'Magazine', 'Receive More Updates About Our Courses', 'I agree to pay the amount through cheque', 'Get updates on DM Content', 'Update me on Supply Chain Content' etc. have been dropped.
 - After checking null percentage dropped column which has null values greater than 45%.
 - For other categorical columns impute null values with 'Unknown' while for numerical columns imputed null values with median.
 - Also removed rows with null values.
 - Outliers are Present in the Variables TotalVisits and Page Views Per Visit and it is removed by imputing median.
- 
- A series of three parallel white diagonal lines extending from the bottom right corner towards the center of the slide.

EXPLORATORY DATA ANALYSIS (UNIVARIATE ANALYSIS)



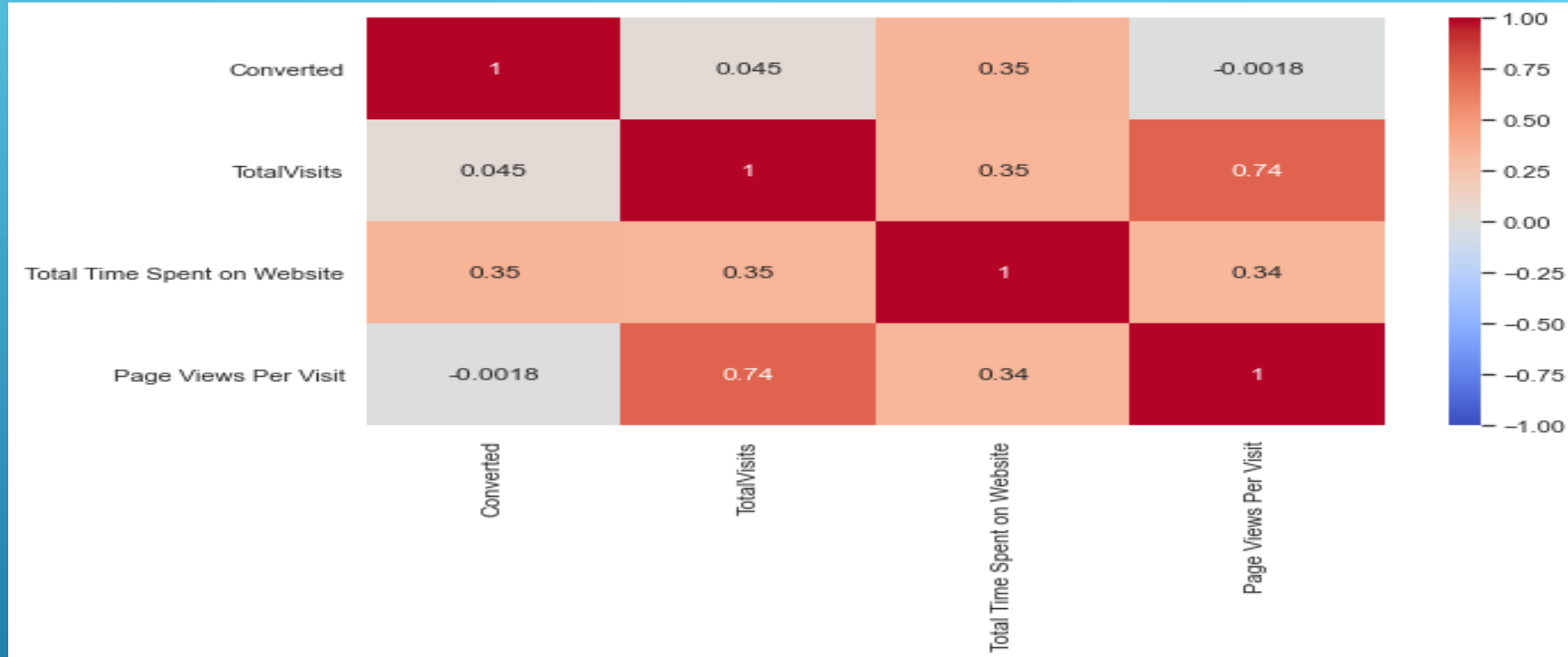
- In Lead Source Direct Traffic and Google are the two main source for Leads
- The Number of values is High in Email Opened and SMS Sent in Last Activity
- Most of the people chooses Finance Management Specialization rather than other Specialization and also most of the values are Unknown.

EXPLORATORY DATA ANALYSIS (BIVARIATE ANALYSIS)




- 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 Specialization the most of the leads are comes from Finance management but here Hot leads are lesser than Cold leads
- In City the most of the hot leads are from Mumbai.

EXPLORATORY DATA ANALYSIS (CORELATION)




As from heatmap can be seen that TotalVisits and Page Views Per Visit are highly correlated.

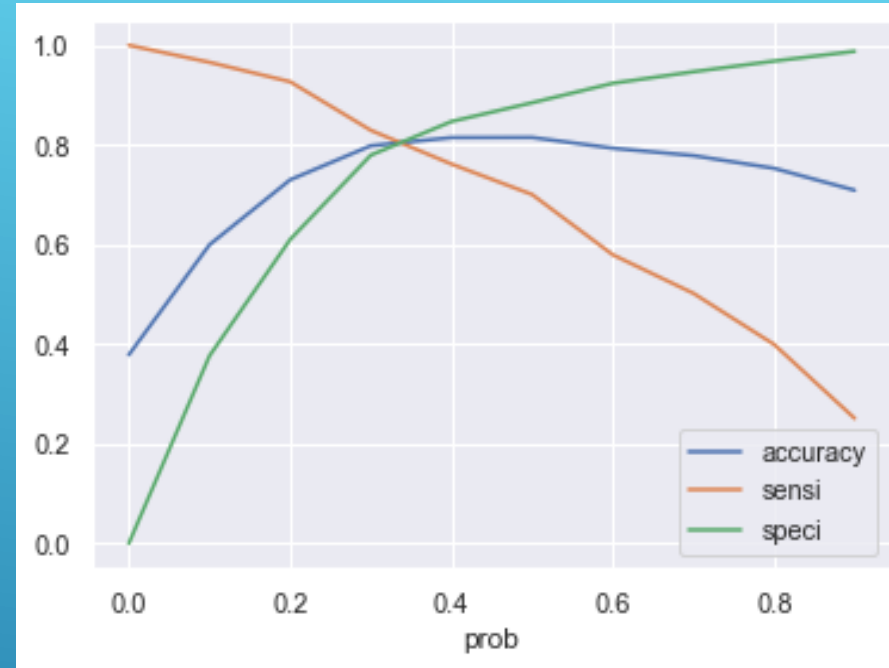
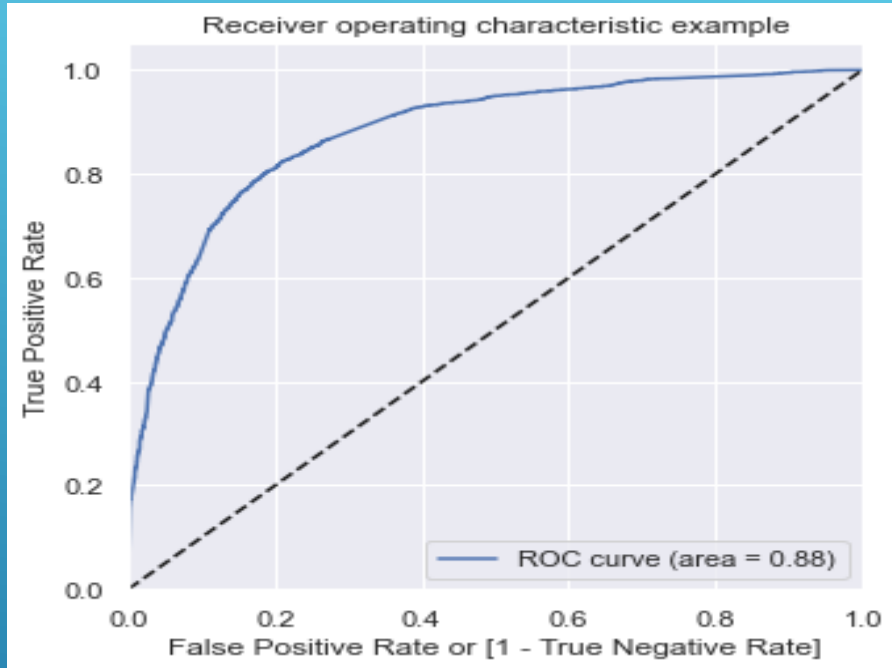
DATA PREPARATION

- Mapped categorical column to 0 and 1.
 - Also dropped some of columns ('Tags', 'Last Notable Activity') which has same value as other column.
 - Dummy Variables are created for object type variables.
- 
- A series of several parallel white diagonal lines of varying lengths and positions, located in the bottom right corner of the slide, creating a modern, abstract graphic element.

MODEL BUILDING

- Splitting the Data into Training and Testing Sets. Splitted data into 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
 - Predictions on test data set.
 - Overall accuracy achieved 81% .
- 
- A series of three parallel white diagonal lines in the bottom right corner of the slide, extending from the bottom edge towards the right edge.

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 0.37 .

CONCLUSION

It was found that the variables that mattered the most in the potential buyers :

- Total number of visits.
- The total time spend on the Website.
- Page Views Per Visit
- When the lead origin is Lead add format.
- When the lead source was:
 - a. Olark Chat
 - b. Welingak website
- When the last activity was:
 - a. Had a Phone Conversation
 - b. Olark chat conversation
 - c. SMS
- 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