# LEAD SCORE CASE STUDY

SUBMITTED BY
SUJATA JHA

# PROBLEM STATEMENT

- X Education sells online courses on different websites. When people fill up form and share the email or phone number they are considered as leads.
- These leads are followed by sales team and when the customer buys the courses the leads are converted. This conversion rate is 30% which is very low.
- We need to build a logistic regression model to predict the most potential leads(hot leads) on which sales team can work and do follow ups. This will save the effort and time of the company.

# DATA PROVIDED

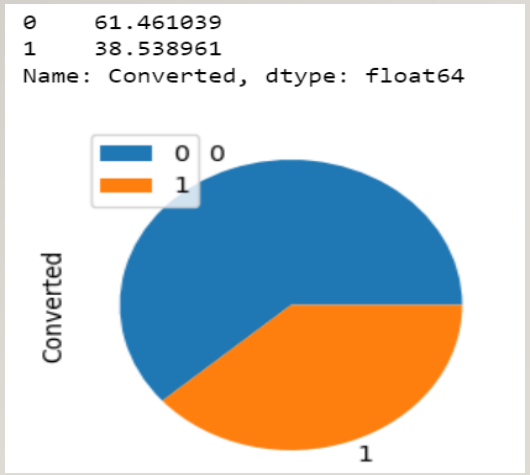
- Leads.csv : This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.
- The target variable, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.
- Leads Data Dictionary.csv : Contains description of all the columns.

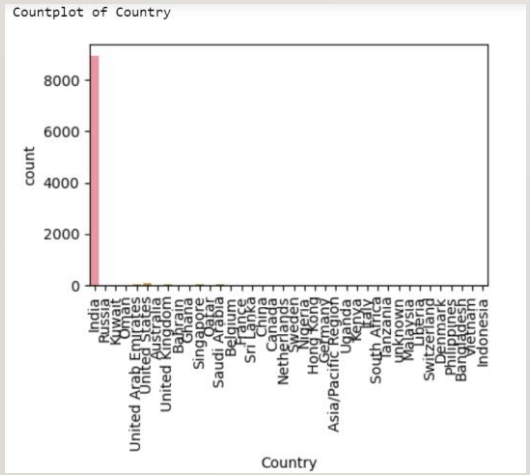
# APPROACH AND METHODOLOGY

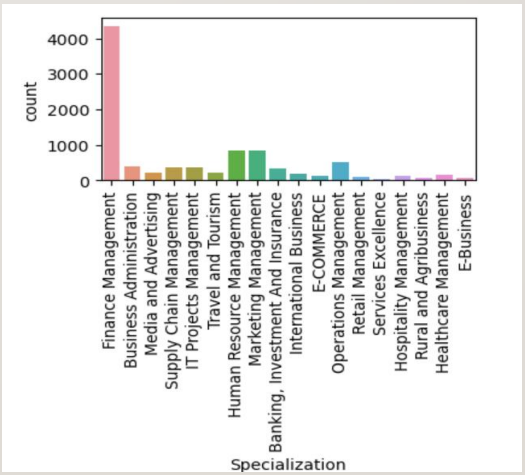
- Data loading and understanding : Loading the data from the Leads.csv file and understanding dataset shape, each columns ,columns type etc.
- Data Cleaning:  If the missing values counts is more than 40% in a column ,then drop the column. Imputing missing values with mode for categorical data and median for continuous data.
- EDA(Univariate, Bivariate and multivariate analysis): Created the list for categorical and continuous data type separately for easy graph plotting.

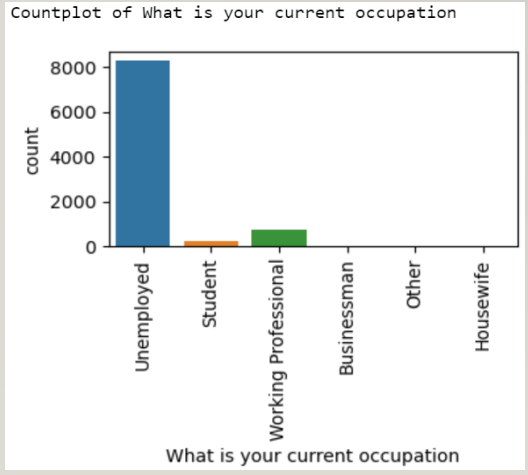
# APPROACH AND METHODOLOGY
# UNIVARIATE ANALYSIS(GRAPHS)

- Conversion rate is only 38%.
- Mostly customers are from India.
- Finance Management people are more interested in buying course.
- Mostly customers are Unemployed



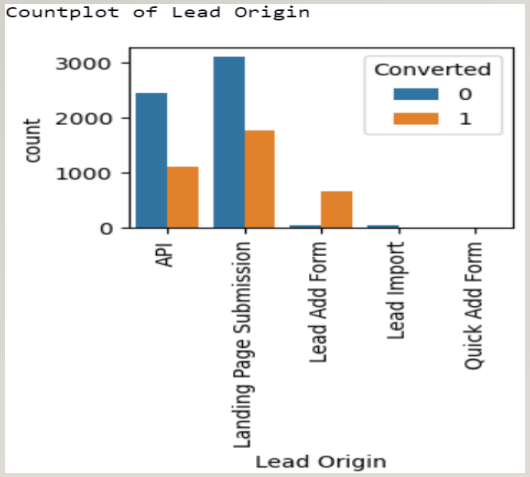



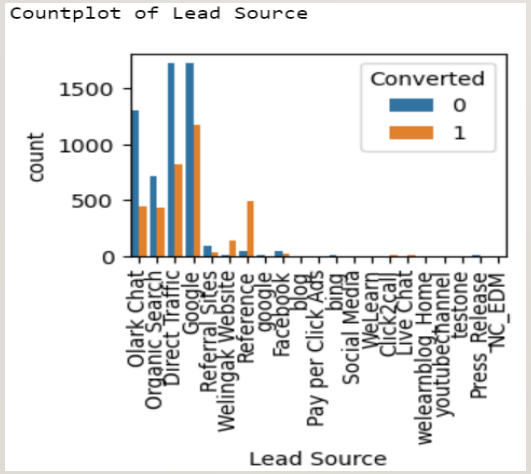



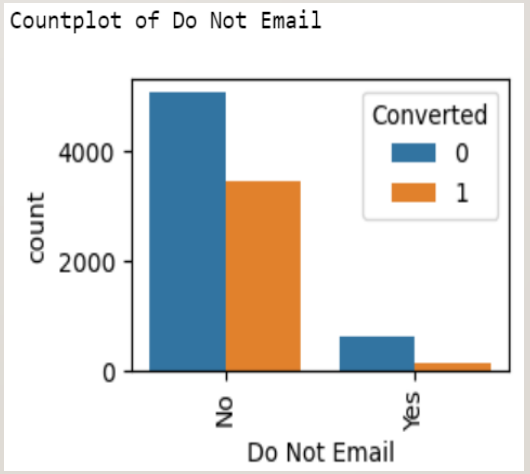



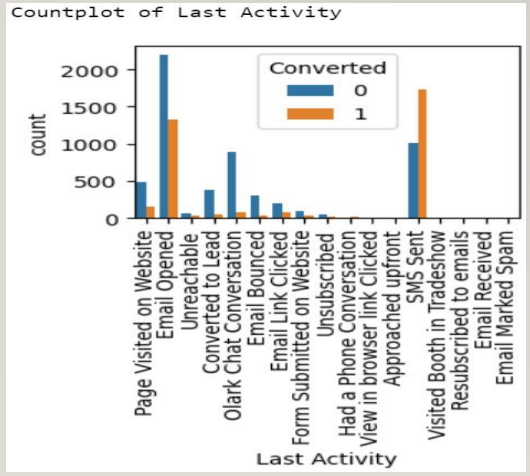
# APPROACH AND METHODOLOGY
# BIVARIATE ANALYSIS(GRAPHS)

- Lead origin as Lead add form ,lead source as Google , permitting to mail, SMS sent as Last activity are showing max. conversion rates.



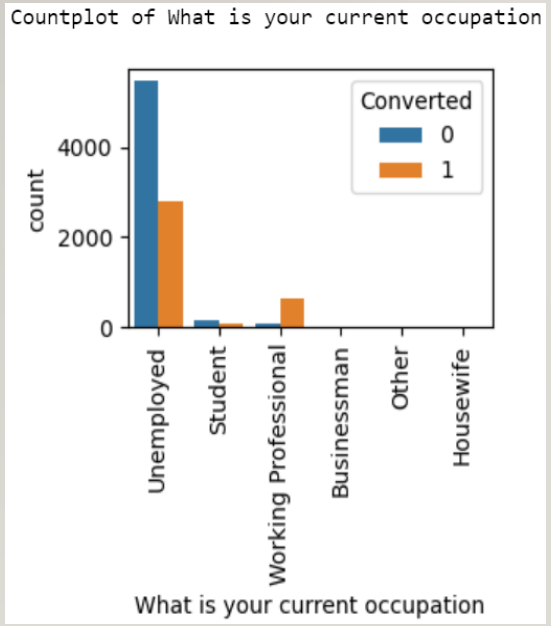



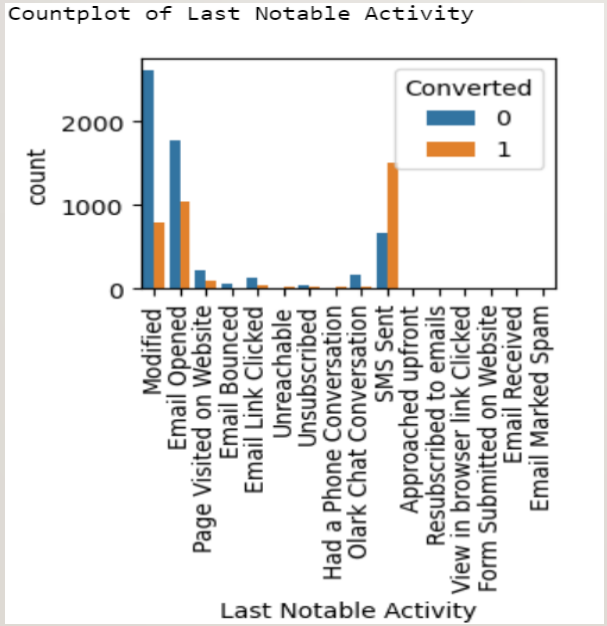



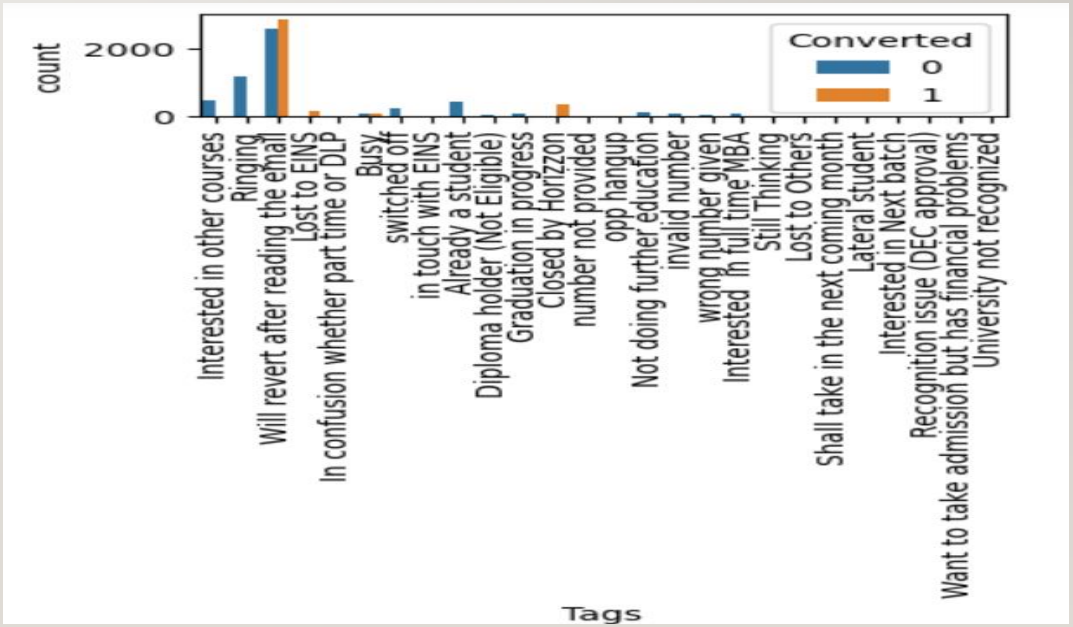



# APPROACH AND METHODOLOGY
# BIVARIATE ANALYSIS(GRAPHS)

- Working professionals have high conversion rate. Last Notable activity as SMS sent , tags as will revert After reading mail and lost to ENS ,closed by horizon have good conversion rate .

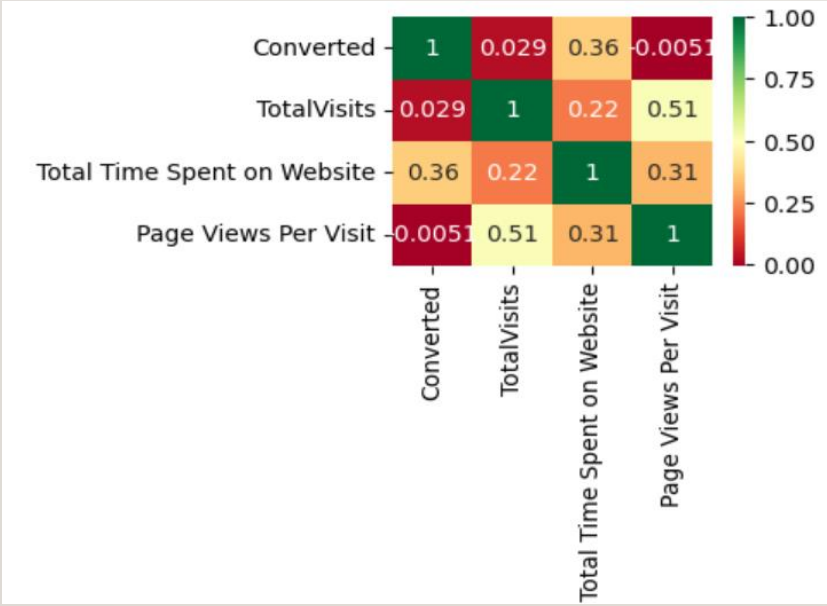


# APPROACH AND METHODOLOGY MULTIVARIATE ANALYSIS(GRAPHS)

- Converted (Target variable) has maximum correlation with Total time spent on websites. Page Views per visit is also correlated with Total visits.

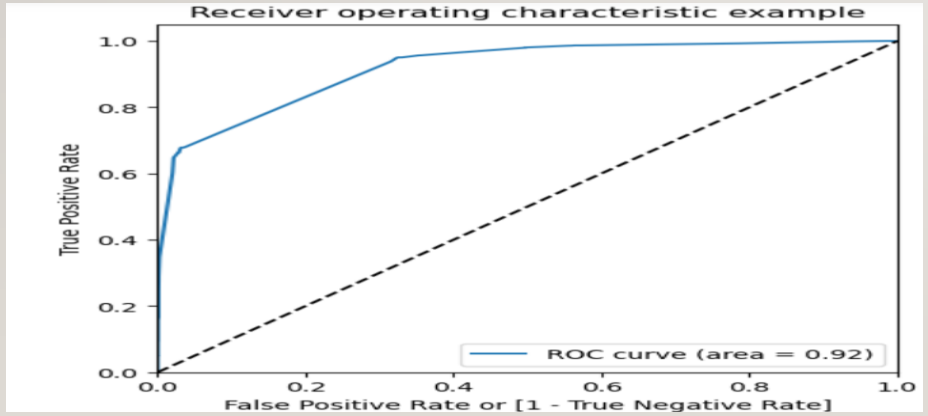


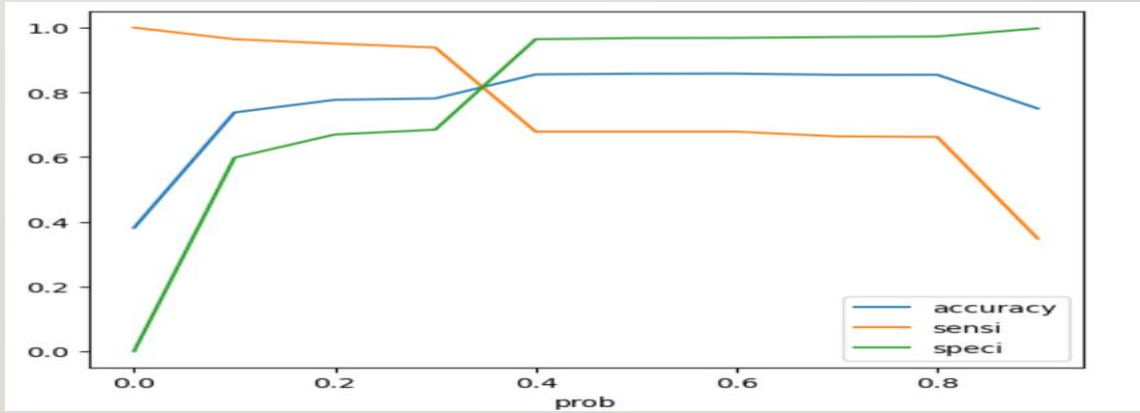
# APPROACH AND METHODOLOGY

- Dummy variables : For logistic regression model all categorical data are split into dummy variables to change the values to numeric form(1,0).
- Train –Test data : Data is split into train and test dataset by 70:30 ratio. Train and test data are normalised using standard scaler.
- Feature selection and Model building: Using hybrid approach to select feature. RFE is used to select 15 features and then manually features are dropped by analysing p-value(less than 0.05) and VIF(less than 5). Finally 11 features are obtained in Model5.

# APPROACH AND METHODOLOGY

- Model evaluation: Predicting from train data and evaluating Accuracy, Recall and Precision. Plotting ROC curve to check the model. We got Optimum cutoff at 0.35.
- Prediction on test set: Normalising numeric data types in test set. Predicting from test data and evaluating Accuracy, Recall and Precision. On comparing the metrics from test and train data set were almost same depicting the model efficiency.



# CONCLUSION

Key variables to identify the hot leads are:

- When the customer is a working professional it has high chance of conversion.
- When the Lead origin is Lead add form.
- Last activity is identified as Olark Chat Conversation and SMS sent.
- When customer has permitted for email.
- When customer is tagged as ‘lost to EINS’, ‘closed by horizon’,’Will revert after reading mail', 'In touch with EINS’.
- Through EDA we can see the ‘time spent in websites’, ‘total visits’, ‘lead source as Google’ seems to give fruitful result.

# SUGGESTION

Some quick suggestions for reducing effort for sales team:

- Phone calls must be done to people who are spending more time on the websites, filling forms, coming back with queries.
- The one who have permitted to call as well as email.
- Reverting back as ping you after reading mail.
- Target the working professionals.
- Use the automated SMS service which can reduces the calls count.