# Lead Score Case study

Group Members:

Jaswanth NA

Sonu Patil

# Problem Statement

- X Education sells online courses to industry professionals.
- 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.

## Business Objective:

- X education 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.

Sensitivity: Internal & Restricted

# 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.
- Conclusions and recommendations.

Sensitivity: Internal & Restricted

# Data Manipulation

Total Number of Rows =33, Total Number of Columns =9204.

TotalVisits , Page Views Per Visit imputed using the median values

'How did you hear about X Education', What matters most to you in choosing a course dropped.

'Magazine','Receive More Updates About Our Courses','Update me on Supply Chain Content','Get updates on DM Content',I agree to pay the amount through cheque' dropped

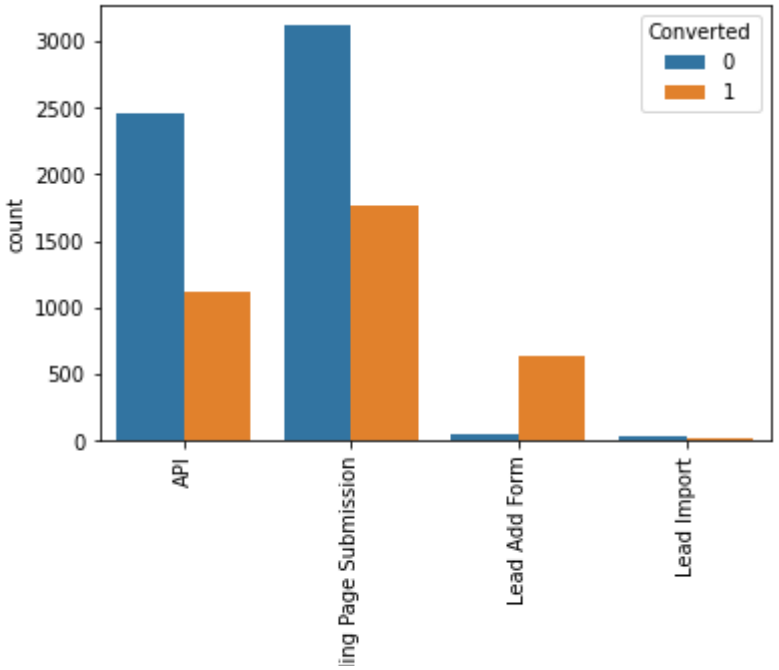
City, Lead Profile , What is your current occupation , Specialization , Last Activity, Tags, Lead Quality, Asymmetrique Profile Score , Asymmetrique Activity Score replaced null values to 'Unknown'

Replacing 90% countries to Foreign Country other than India

Converted is the target variable, Indicates whether a lead has been successfully converted (1) or not (0).

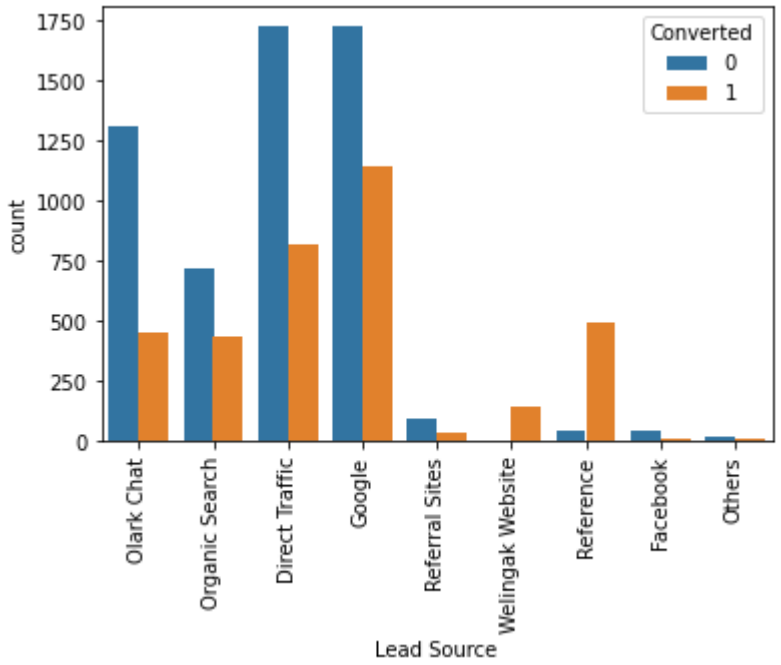
Sensitivity: Internal & Restricted

# Univariate Analysis




## Inference

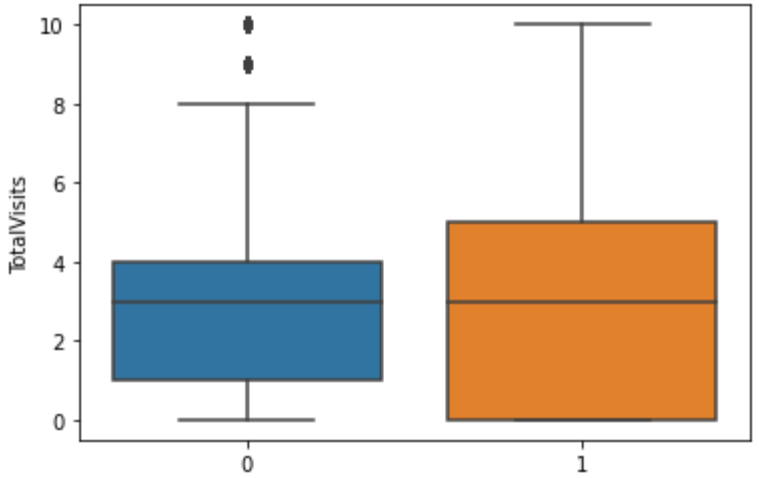
- API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
- Lead Add Form has more than 90% conversion rate but count of lead are not very high.
- Lead Import are very less in count.
- To improve overall lead conversion rate, we need to focus more on improving lead converion of API and Landing Page Submission origin and generate more leads from Lead Add Form




## Inference

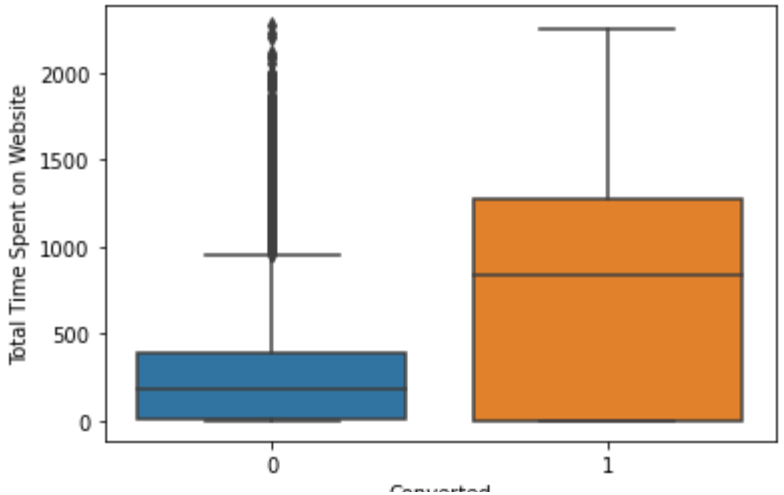
- Google and Direct traffic generates maximum number of leads.
- Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, focus should be on improving lead conversion
- of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

Sensitivity: Internal & Restricted



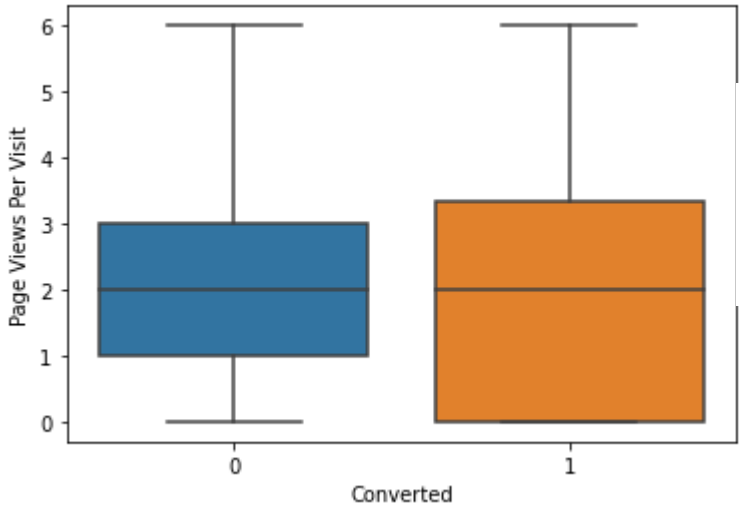

## Inference

- Median for converted and not converted leads are the same.
- Nothng conclusive can be said on the basis of Total Visits.




## Inference

- Leads spending more time on the weblise are more likely to be converted.
- Website should be made more engaging to make leads spend more time




## Inference

- Median for converted and unconverted leads is the same.
- Nothing can be said specifically for lead conversion from Page Views Per Visit

Sensitivity: Internal & Restricted

# Data Preparation

- Numerical Variables are Normalized
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 9204
- Total Columns for Analysis: 27

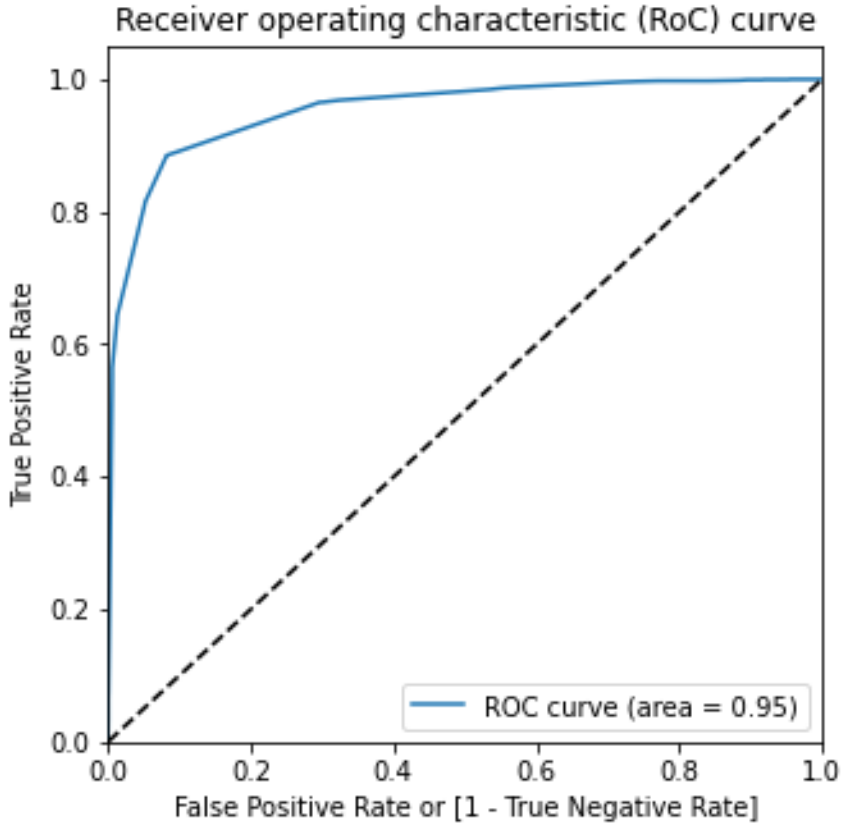
Sensitivity: Internal & Restricted

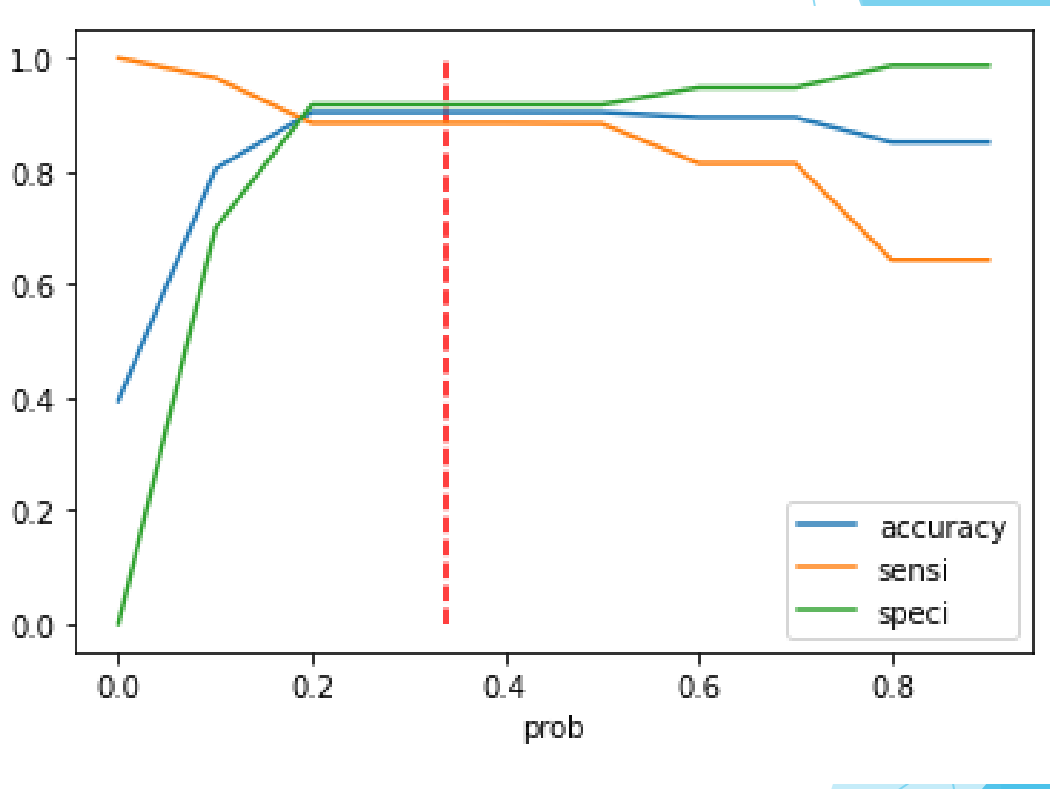
# Model Building

- 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 13 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 90%

Sensitivity: Internal & Restricted

# 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 0.35.

Sensitivity: Internal & Restricted

# Conclusion

- 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.
- the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 90%, 89.9%; Sensitivity=87%; Specificity= 93%.
- Also the lead score calculated shows the conversion rate on the final predicted model is around 80% (in train set) and 79% in test set
- The top 3 variables that contribute for lead getting converted in the model are
  - Tags_Lost to EINS
  - Lead Source_Welingak Website
  - Tags_Will revert after reading the email

Overall. Our model looks good.

Sensitivity: Internal & Restricted

# Conclusion

- 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.
- the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 90%, 89.9%; Sensitivity=87%; Specificity= 93%.
- Also the lead score calculated shows the conversion rate on the final predicted model is around 80% (in train set) and 79% in test set
- The top 3 variables that contribute for lead getting converted in the model are
  - Tags_Lost to EINS
  - Lead Source_Welingak Website
  - Tags_Will revert after reading the email

Overall. Our model looks good.

Sensitivity: Internal & Restricted