# Lead Score Case study

Group Members:

Sayanwita

Sandeep

# 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.

# Solution Methodology

- Data cleaning and data manipulation
  1. Check and handle duplicate data.
  2. Check and handle garbage 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.

# Data Manipulation

Total Number of Rows =37, Total Number of Columns =9240.

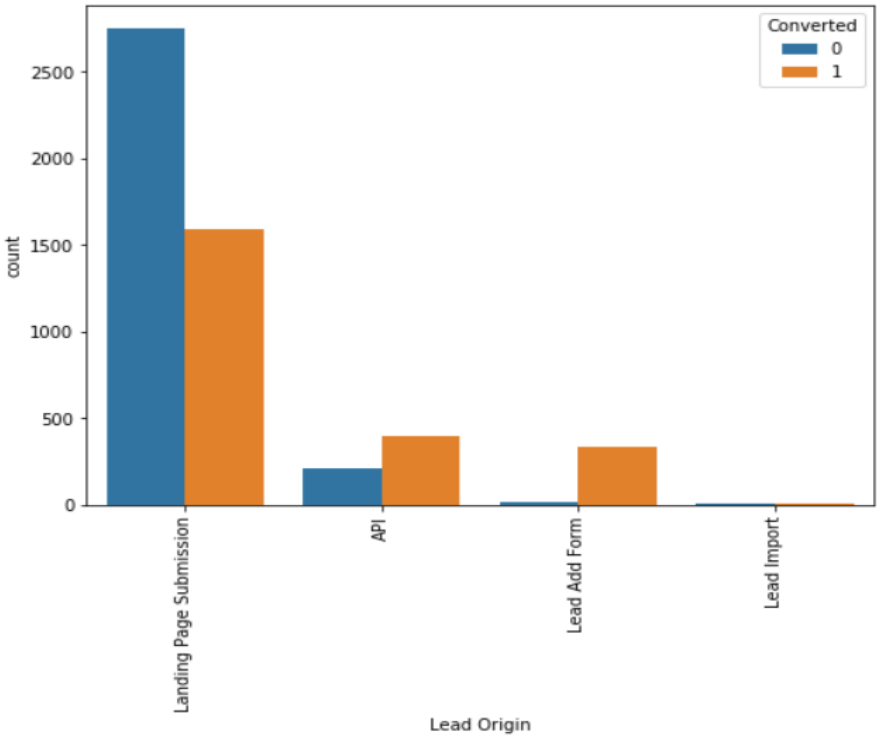
Dropping the columns having more than or equal to 45% NA values.

Replaced NAN values of City,Tags,Specialization,Country and 2 other columns.

“Lead Number, Search, Do Not Call, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, 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” are the uneccessary columns which were dropped.

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

For outlier handling we removed the top & bottom 5% of the Column Outliers.

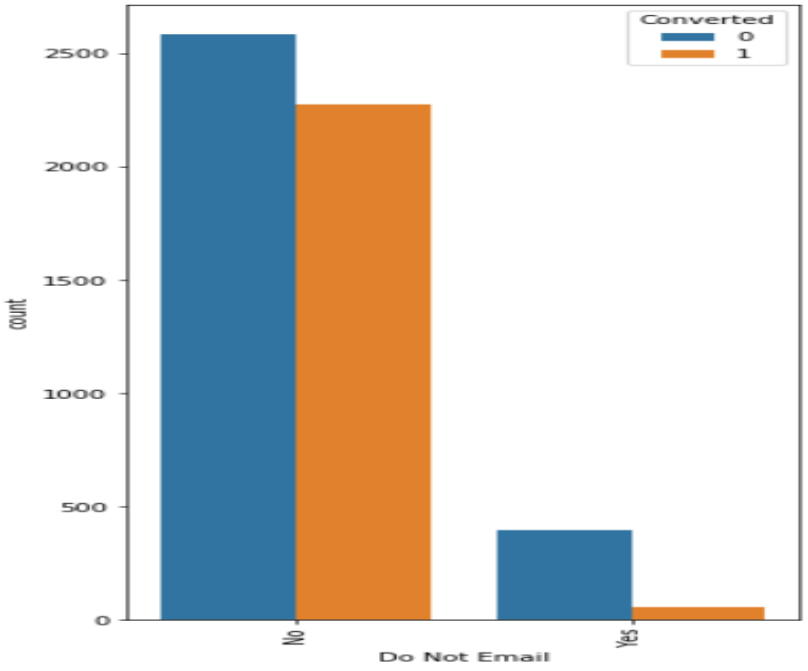


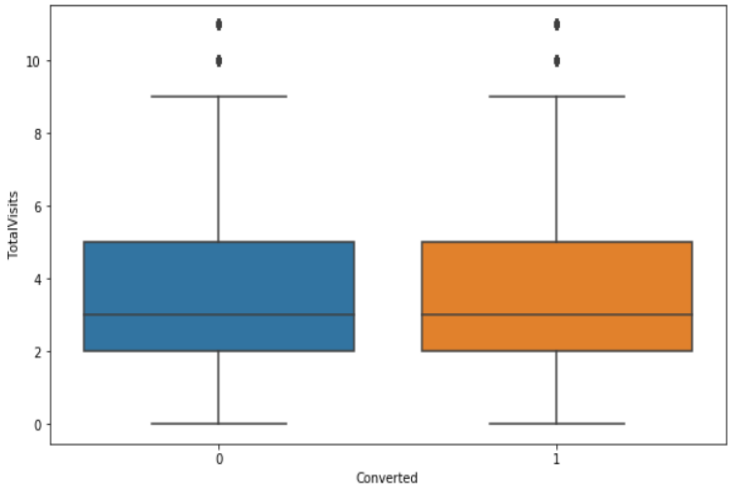

## Inference

- Landing Page Submission have nearly 55% conversion rate but count of lead originated from them are considerable.
- Lead Add Form has a very high 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 Landing Page Submission origin and generate more leads from Lead Add Form

## Inference

- The number of leads from NO option in Do Not Mail is very high as compared to the YES option.
- Conversion Rate of NO leads is around 86%
- Whereas the conversion rate for YES option is low around 20%
- To improve overall lead conversion rate, focus should be on improving lead conversion.



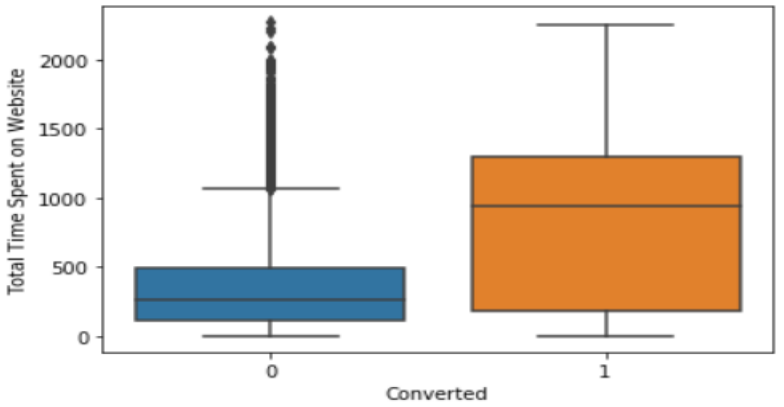


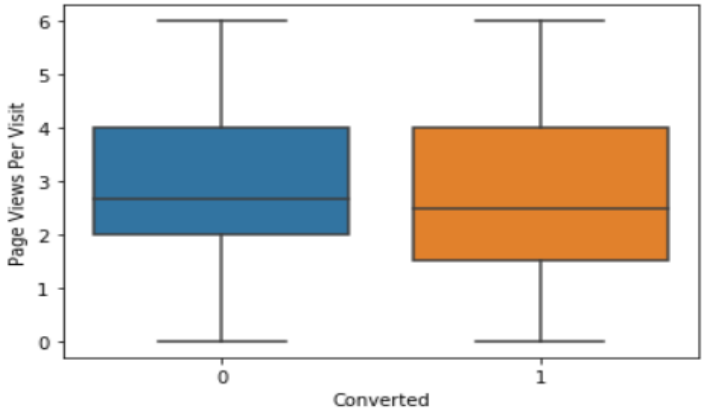

## Inference

- 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 website 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

# Data Preparation

- Numerical Variables are Normalized
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 5312
- Total Columns for Analysis: 14

# 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 to achieve a minimum p- value and keeping VIF value is greater than 5
- Predictions on test data set
-

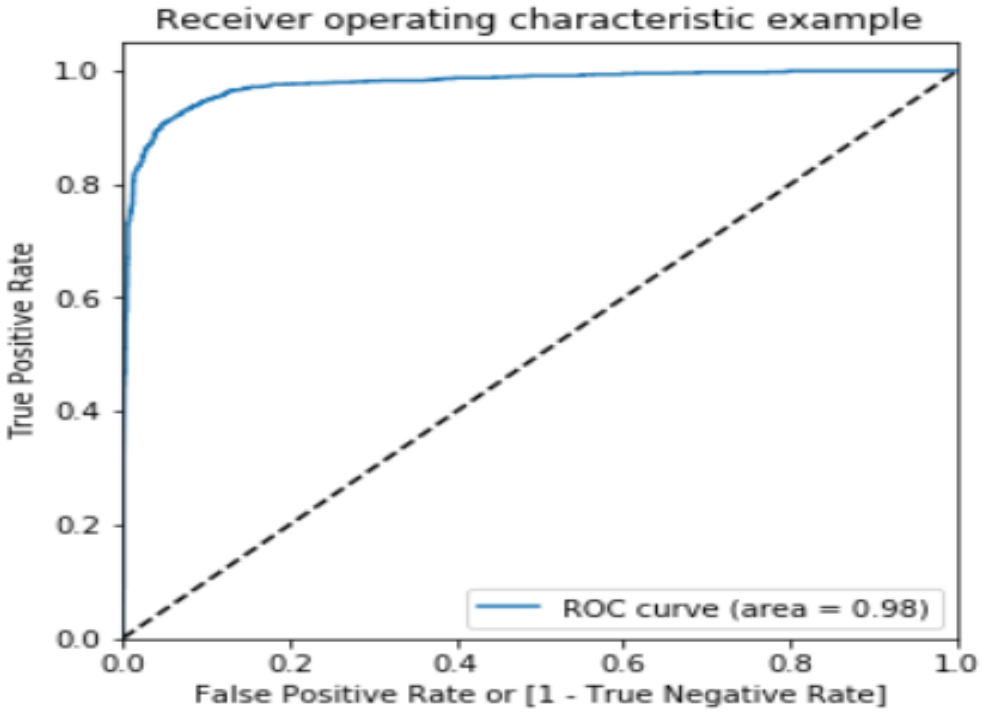
# Final Model

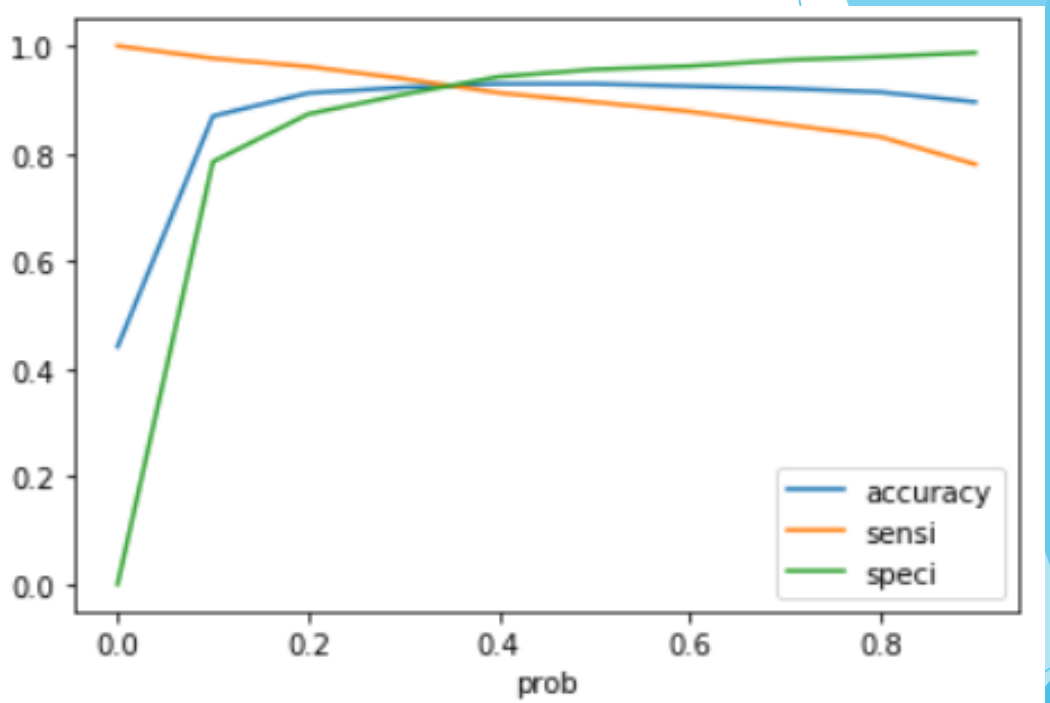
| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -0.8225 | 0.210 | -3.914 | 0.000 | -1.234 | -0.411 |
| Do Not Email | -1.2741 | 0.345 | -3.698 | 0.000 | -1.949 | -0.599 |
| Total Time Spent on Website | 1.0190 | 0.073 | 13.893 | 0.000 | 0.875 | 1.163 |
| Lead Origin_Landing Page Submission | -1.0803 | 0.211 | -5.118 | 0.000 | -1.494 | -0.667 |
| Last Activity_SMS Sent | 1.8862 | 0.154 | 12.288 | 0.000 | 1.585 | 2.187 |
| Tags_Closed by Horizzon | 7.6582 | 1.024 | 7.480 | 0.000 | 5.652 | 9.665 |
| Tags_Interested in other courses | -2.3695 | 0.652 | -3.632 | 0.000 | -3.648 | -1.091 |
| Tags_Lost to EINS | 6.1834 | 0.748 | 8.269 | 0.000 | 4.718 | 7.649 |
| Tags_Ringing | -2.6859 | 0.277 | -9.685 | 0.000 | -3.230 | -2.142 |
| Tags_Will revert after reading the email | 4.9137 | 0.212 | 23.176 | 0.000 | 4.498 | 5.329 |
| Tags_switched off | -3.1681 | 0.746 | -4.247 | 0.000 | -4.630 | -1.706 |
| Last Notable Activity_Modified | -1.4657 | 0.170 | -8.613 | 0.000 | -1.799 | -1.132 |

## Inference:

- As we can see in the table, the highlighted coef values show the variables responsible for the lead conversion
- The other variables which contribute, but not as much as the 3 highlighted ones, are “Total time spent on website” & “Last Activity SMS sent”
- The negative sign on the remaining variables signifies the variables to be affecting inversely in lead conversion.

# ROC Curve




- We can observe from the first graph that the curve is inclined towards left hand corner showing a high True Positive Rate.
- 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

# 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.

<u>Train Data</u>
Specificity: 94.2%
Sensitivity: 91.2%
Accuracy: 92.95%

<u>Test Data</u>
Specificity: 91.37%
Sensitivity: 92.17%
Accuracy: 91.71%

- Also the lead score calculated shows the conversion rate on the final predicted model is around 43.88%
- The top 3 variables that contribute for lead getting converted in the model are
- i)Tags_Closed by Horizzon

ii)Tags_Lost to EINS

iii)Tags_Will revert after reading the email

Overall. Our model looks good.