Lead Score Case Study

Group Members
1. Ayush Maheshwari
2. Akash Kaushik

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 Select/NA values and missing values.
  2. Drop columns, if it contains large number of missing values and not useful for the analysis.
  3. Imputation of the values, if necessary.
  4. Check and handle outliers in data.
- EDA
  1. Univariate data analysis: value count, distribution of variable etc.
  2. Bivariate data analysis: correlation and distribution of variables against the target, etc.
- Feature Scaling & Dummy Variables of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

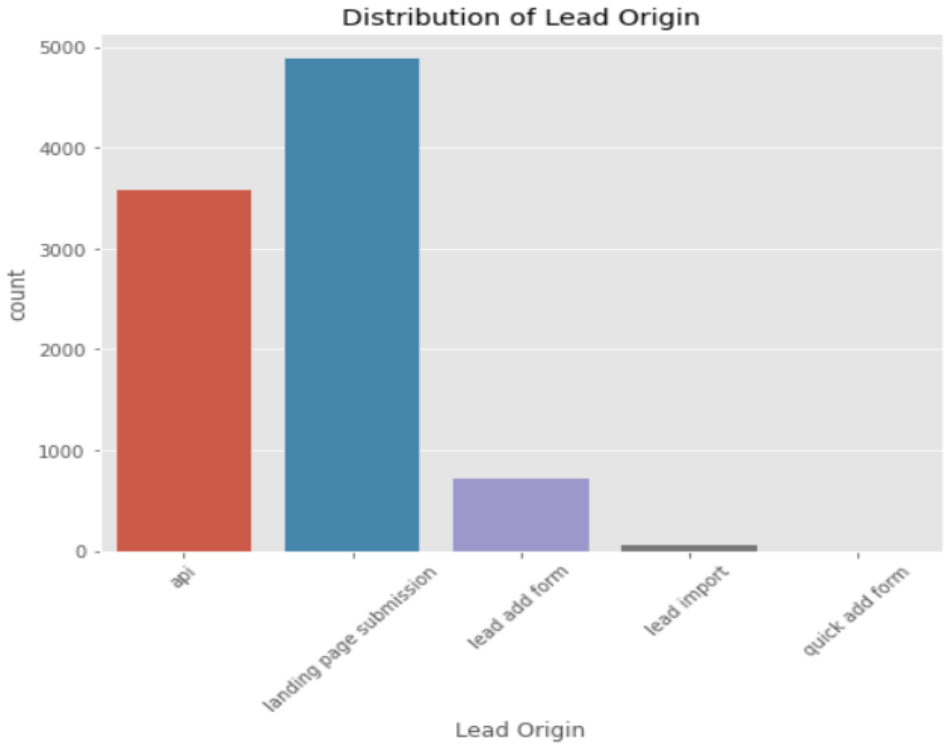
Data Manipulation

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “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” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the null value counts for some of the object type variables, we find some of the features which had large number of null values, which we have dropped, the features are: "How did you hear about X Education”, “Lead Quality”, “Lead Profile”, “Asymmetrique Profile Score”, “Asymmetrique Activity Score”,etc.
- For the columns containing values like Not Available, even though it looks like null value, we have kept it as a separate category .

EDA

Distribution of Lead Origin:

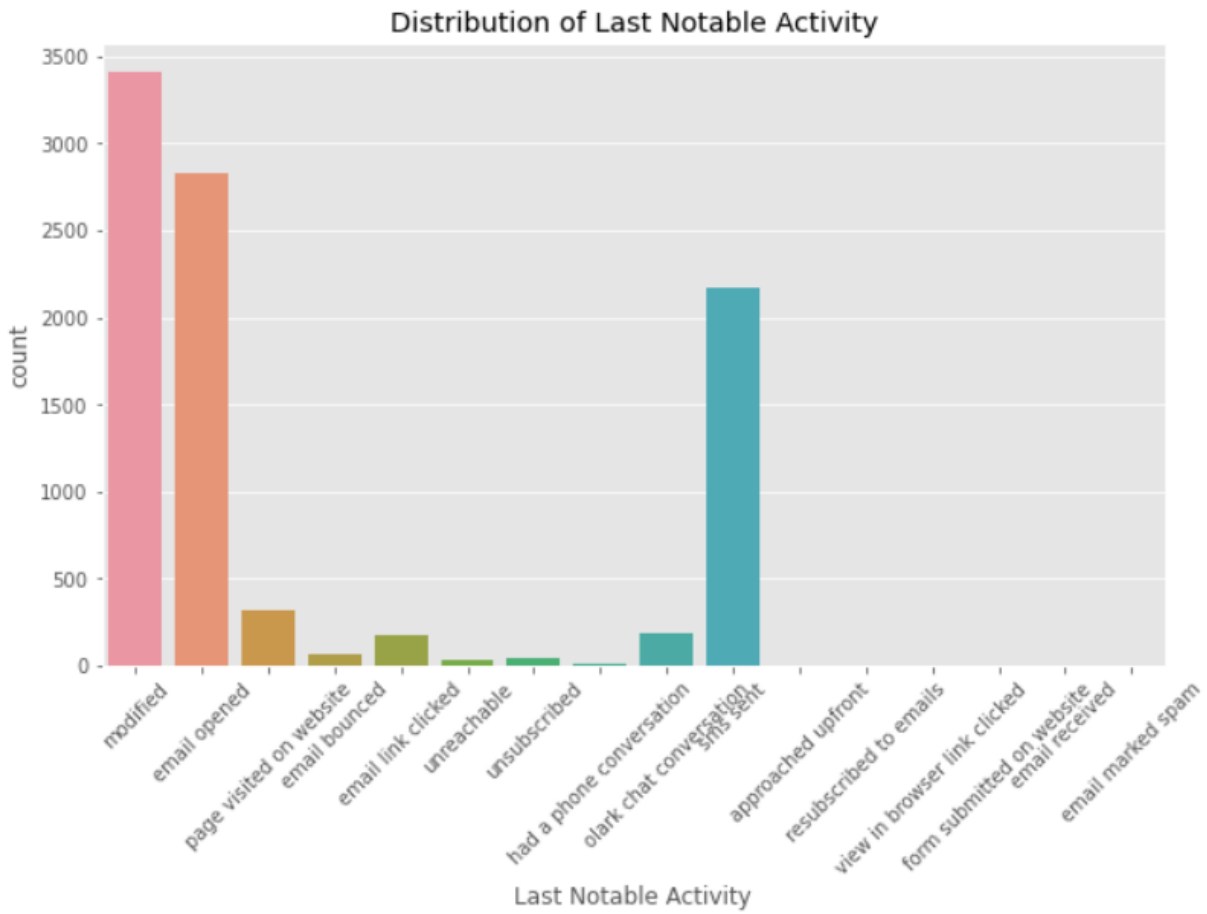
Most of the leads have originated by landing page submission, followed by api, lead add form and so on.



EDA

Distribution of Last Notable Activity:

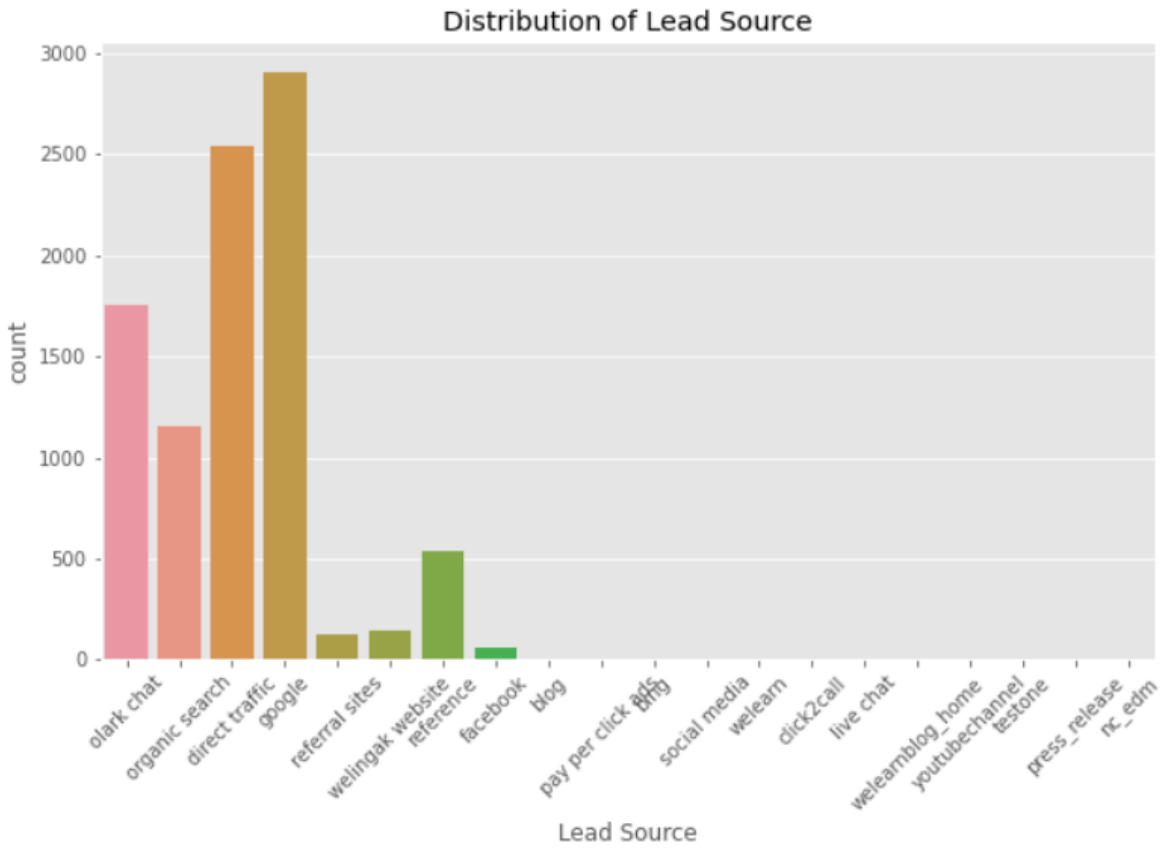
Most of the leads modified their account as a last activity followed by opening the email and doing a chat conversation, probably with a chatbot.



EDA

Distribution of Lead Source:

Most of the leads are generated from Google followed by organic search, olark chat and so on.

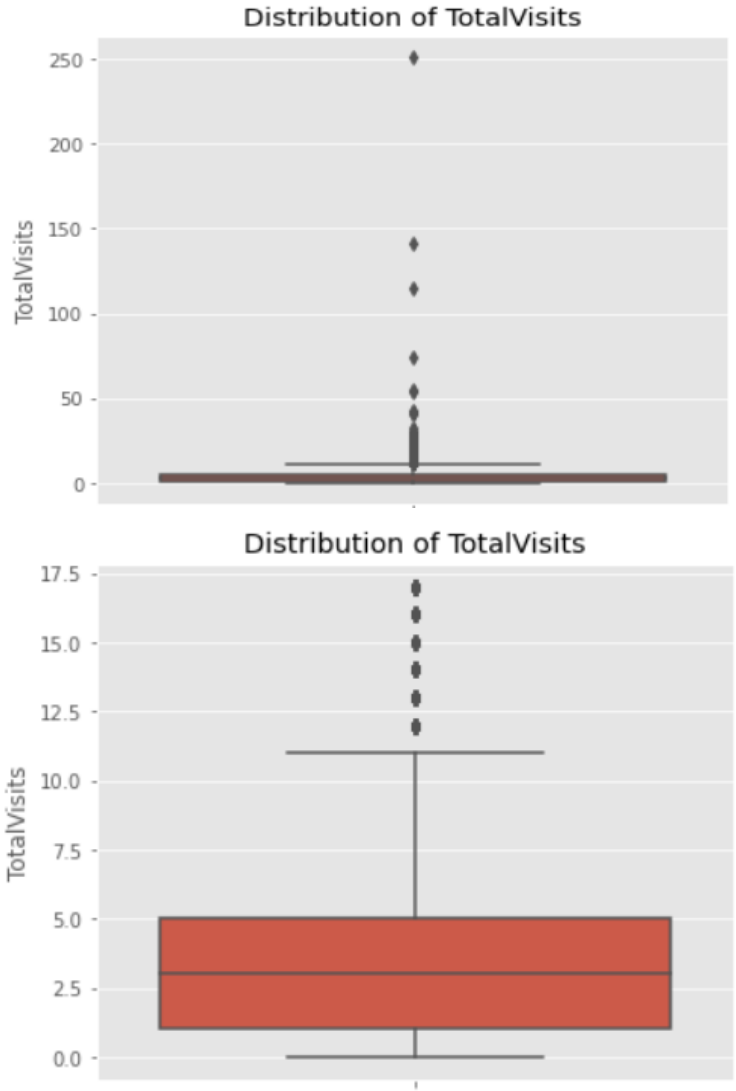


EDA

Distribution of Total Visits:

In the plot 1 we see there are some outliers, moslty between the 99th quantile and the max Value.

We capped the outliers to the 99th quantile and Now there are no outliers in the Total Visits Variable as can be seen in the second plot.

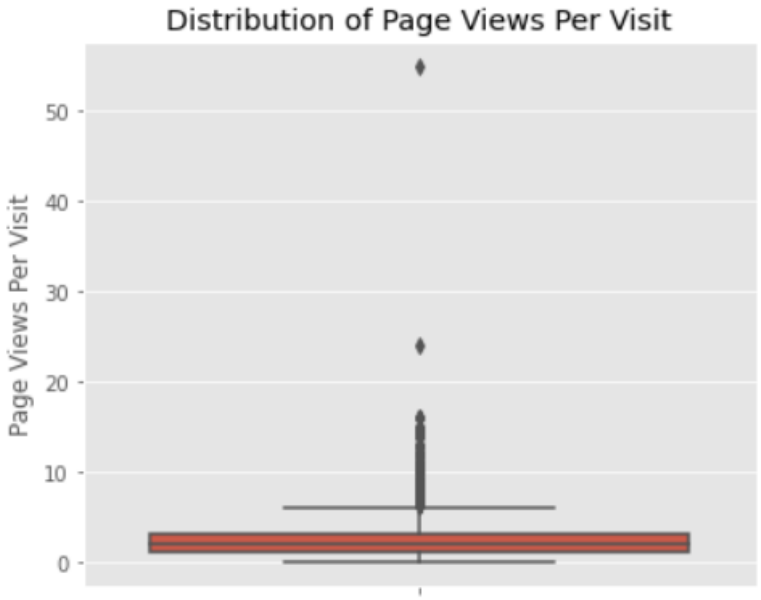


EDA

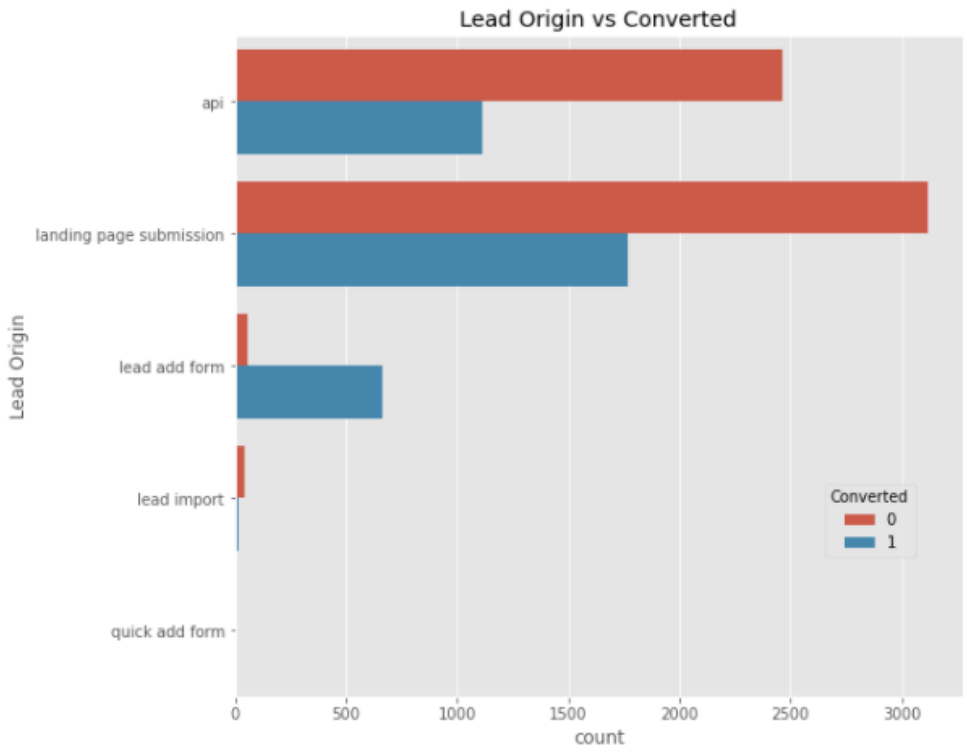
Distribution of Page Views Per Visit :

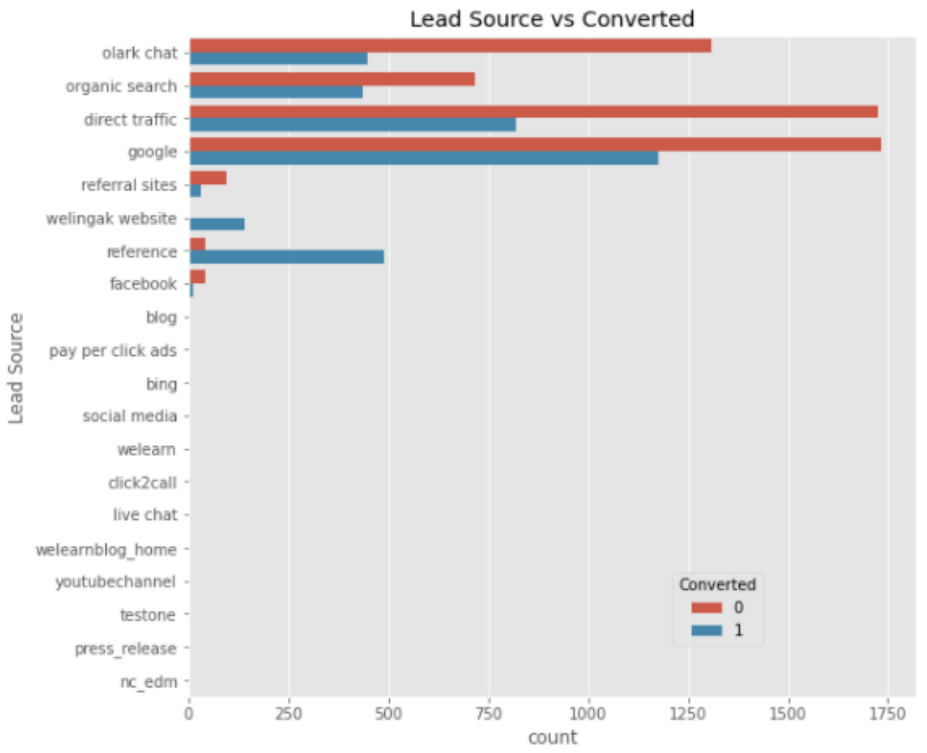
In the plot 1 we see there are some outliers, between the 99th quantile and the max Value.

We capped the outliers to the 99th quantile and Now there are no outliers in the Page Views Per Visit Variable as can be seen in the second plot.



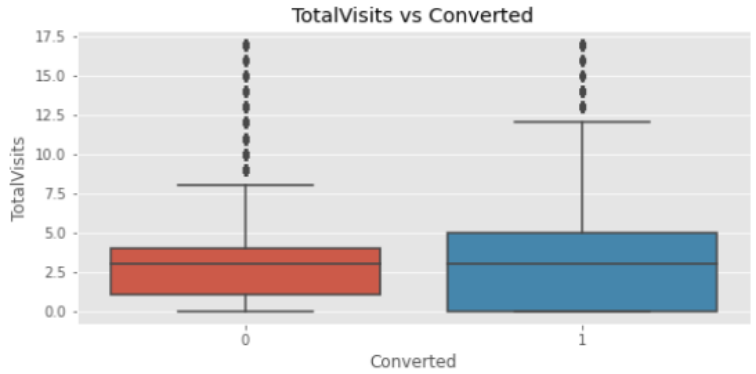
EDA

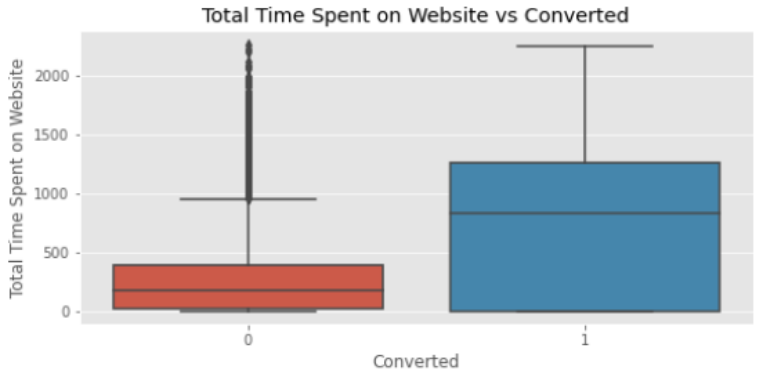


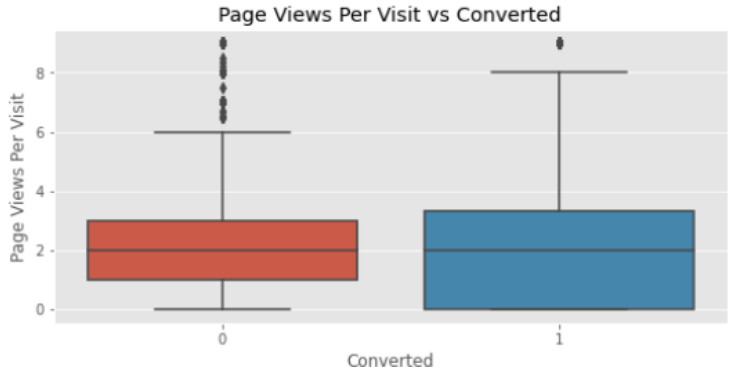


- Most of the Leads originating from Landing Page Submission are successfully converted followed by those originating from api.
- Also, there are a large number of leads who came through Google are converted followed by direct trafic, olark chat and organic search.

EDA







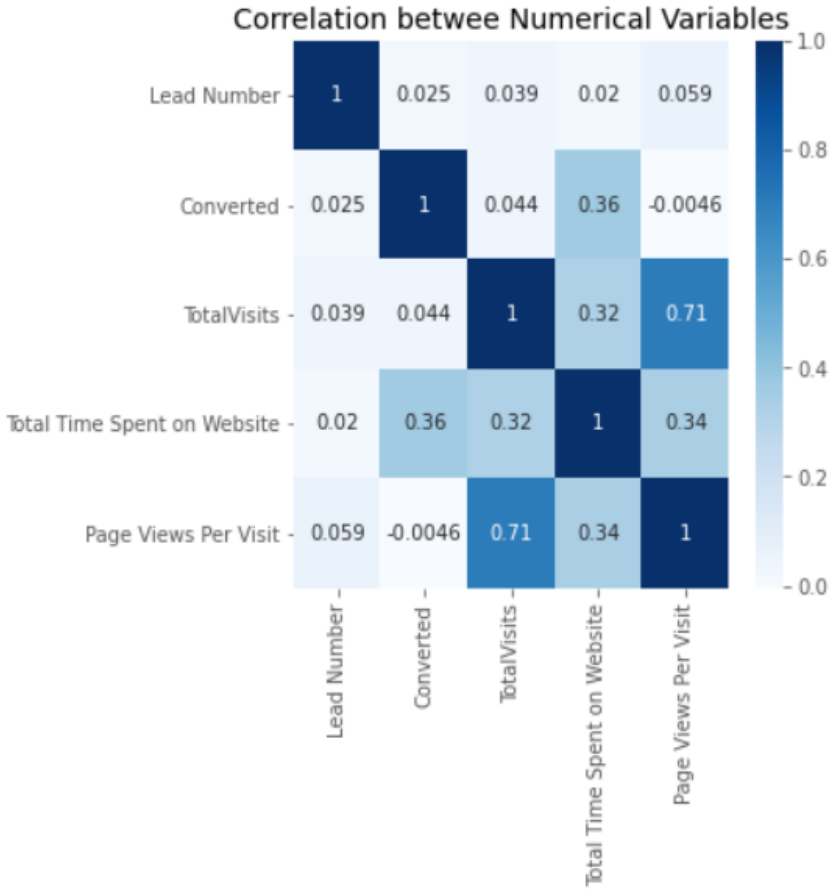
Leads Spending more time on website are likely to be converted, on the other hand since median of the converted and not converted leads for the columns Total Visits and Page Views Per Visit are almost same, nothing conclusive can be said based on this variables.

EDA

•We see few patches of notable correlation, there are correlation between:

- Page Views Per Visit and TotalVists
- Page Views Per Visit and Total Time Spent on Website
- Total Time Spent on Website and TotalVisits

Now all this are self explanatory as more the lead visists the website the more he/she will spend time and view pages.



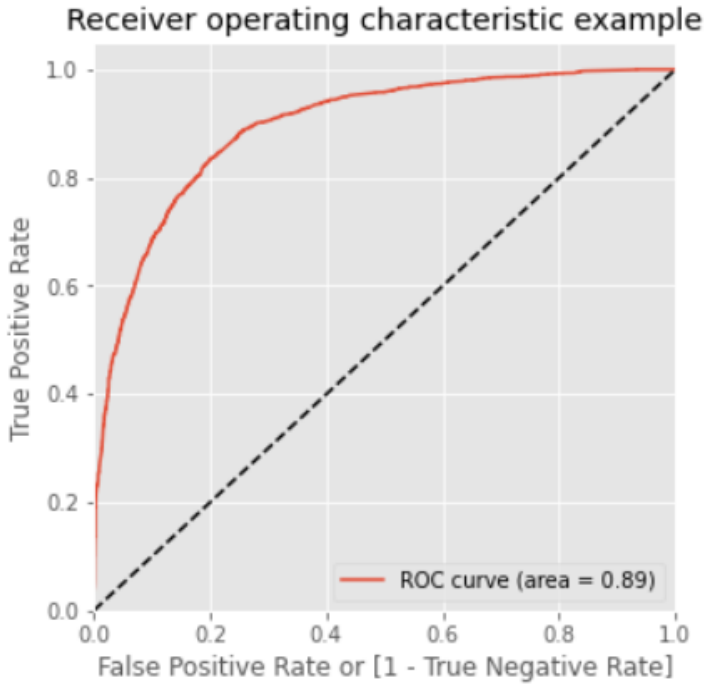
Data Conversion

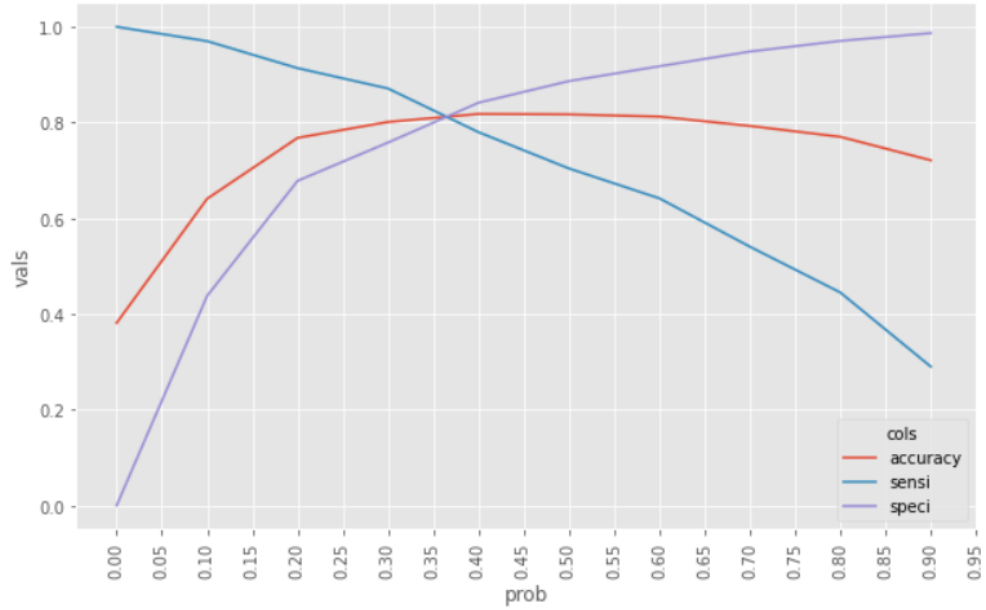
- Binary encoding was done for variables with binary categories like yes and no.
- Dummy Variables are created for object type variables
- Numerical Variables are Normalised

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 20 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 train data set
- Overall accuracy 81.33%, after selecting the optimal cutoff probability.
- Model Sensitivity/Recall : 80.57%, Model Specificity : 81.80% , Model Precision : 73.18% and F1 score 0.767 after selecting the optimal cutoff probability.

ROC Curve and Optimal cutoff





- **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 0.36.

Evaluation on Test Set

- We did the same Normalization on the test set.
- Using the model we predicted the Conversion probabilities.
- We then used the optimal cutoff probability to predict the Lead Conversion.
- Overall accuracy on test set 81.42%.
- Sensitivity/Recall : 80.09%, Specificity : 82.28% , Precision : 74.70% and F1 score 0.773 on the test set.

Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- The total time spend on the Website.
- Having a phone conversation as Last Notable Activity.
- When the lead is a working professional.
- When the lead source is Welingak Website.
- When the last activity was:
  a. SMS
  b. Olark chat conversation
- When the lead origin is Lead add format.

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.