Lead Scoring Case Study

Identifying Hot leads

Group Members:
Diksha Sahu
Saurabh Rungta
Anuraag Nedunoori

Problem Statement

- X Education sells online courses to Industry Professionals.
- Although 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 30 of them gets converted.
- The objective is to build a model to identify the hot leads and achieve conversion rate of around 80%.

Business Objective :

- X Education wants to know most promising leads.
- For which they want a Logistic Regression Model whose conversion rate is around 80%.
- This will help the Sales team to divert their focus on potential leads.

Solution Approach

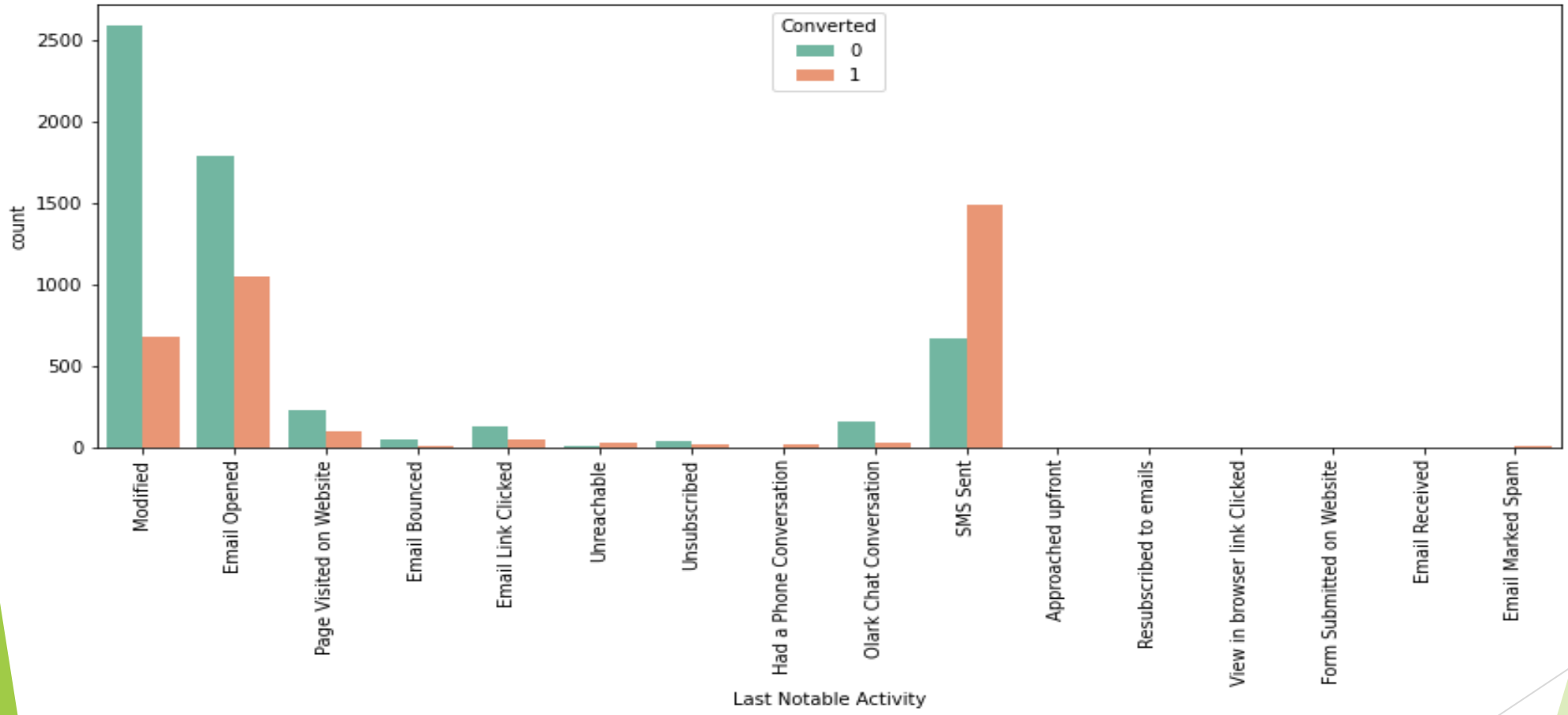
- Data Understanding and Data cleaning -
  a. Check and handle duplicate values if any.
  b. Check and handle missing & unique & ‘select’ values.
  c. If the missing value percentage is > 40 %, drop such columns.
  d. Perform missing value imputations when necessary.
  e. Outlier treatment.
- EDA –
  a. Univariate Analysis of categorical and numerical variables.
  b. Bivariate Analysis of categorical and numerical variable against the target variable.
  c. Build a correlation matrix.

- Dummy Variable creation, Feature scaling & Train-Test split of the data.
- Model Building – logistic regression
- Validation and Representation of the final Model.
- Conclusion and Deriving factors responsible in driving leads.

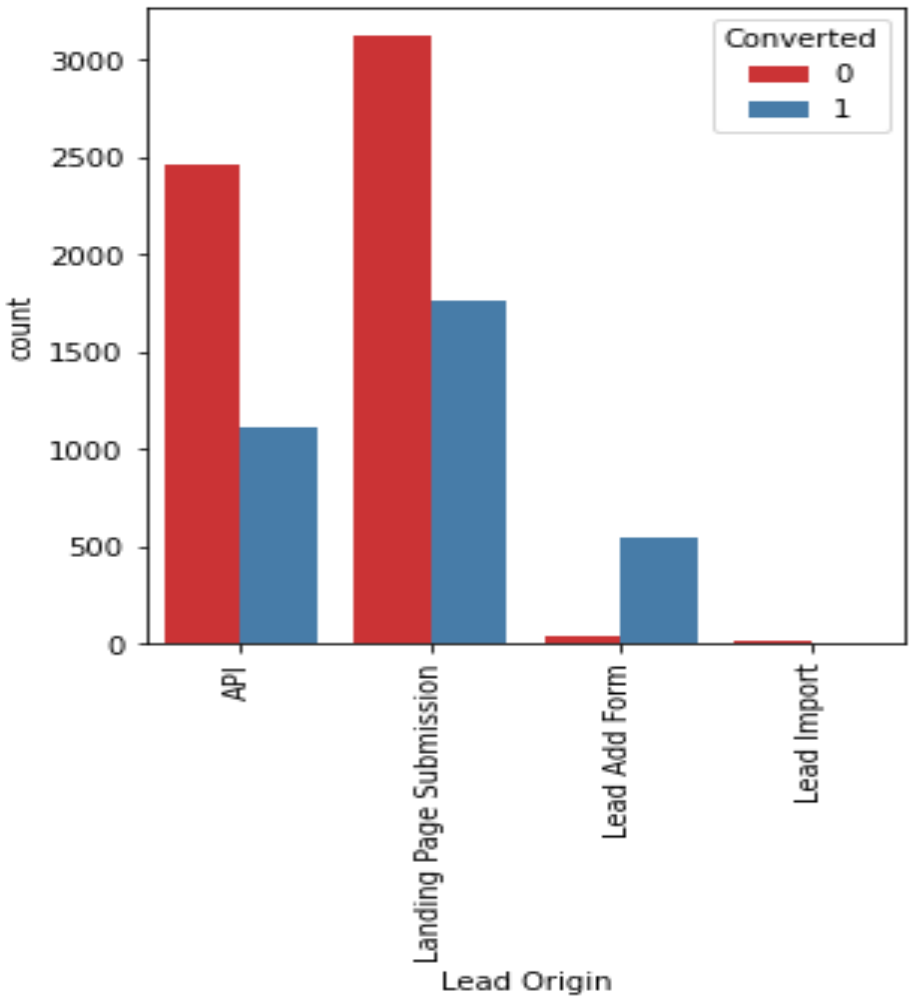
Data Cleaning & Manipulation

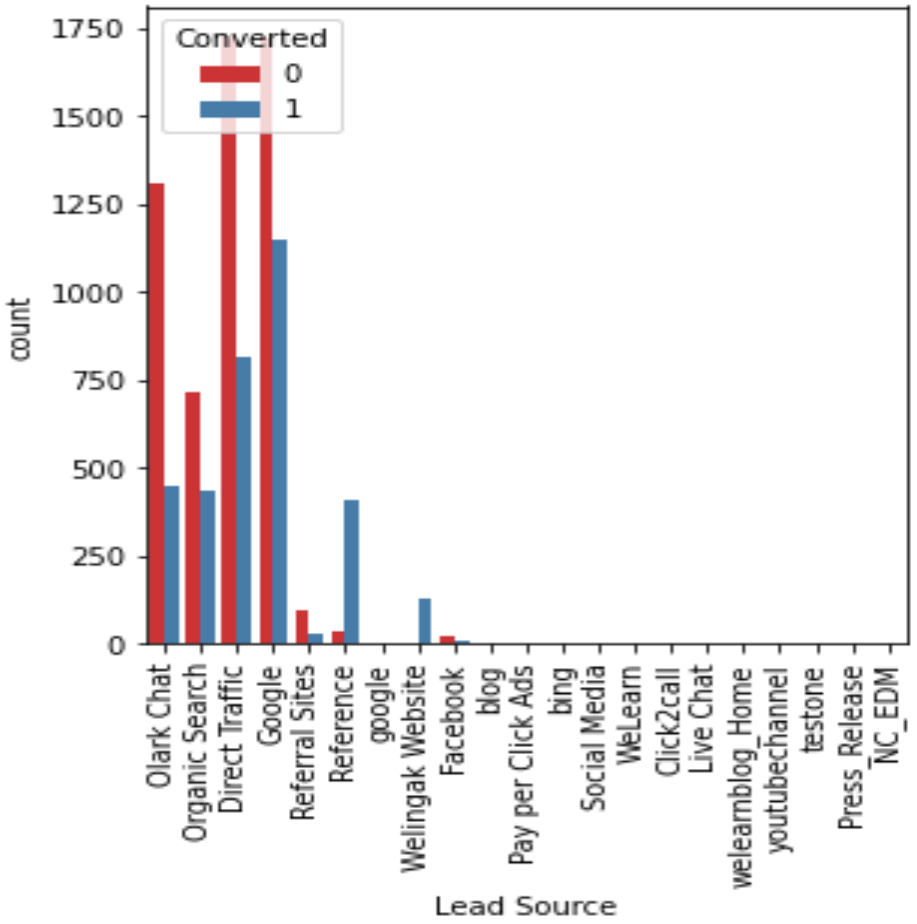
- We checked if the dataset had any duplicate value; no such discrepancies were found.
- Handling of ‘Select’ variable – this value implies that user has not selected any option. Hence, it was imputed with null values.
- Removed columns ‘Prospect ID’ and ‘Lead Number’ as these were just indicative of user has been contacted.
- Dropped the column with null values > 40 % such ‘Lead Profile’ and ‘Lead Quality’
- Columns having only one unique value for all the leads such as ‘Magzine’ , ‘Get updates on DM Content’ etc.
- Missing value imputation for columns such as ‘Country’ , ‘What is your current occupation’, ‘Specialization’ etc.
- Analyzed the Categorical feature ‘Country’ after checking the value counts.
- So, after performing all above steps we had Total No. of Rows = 9074 & Total No. of Columns = 23.

EDA : Categorical Features



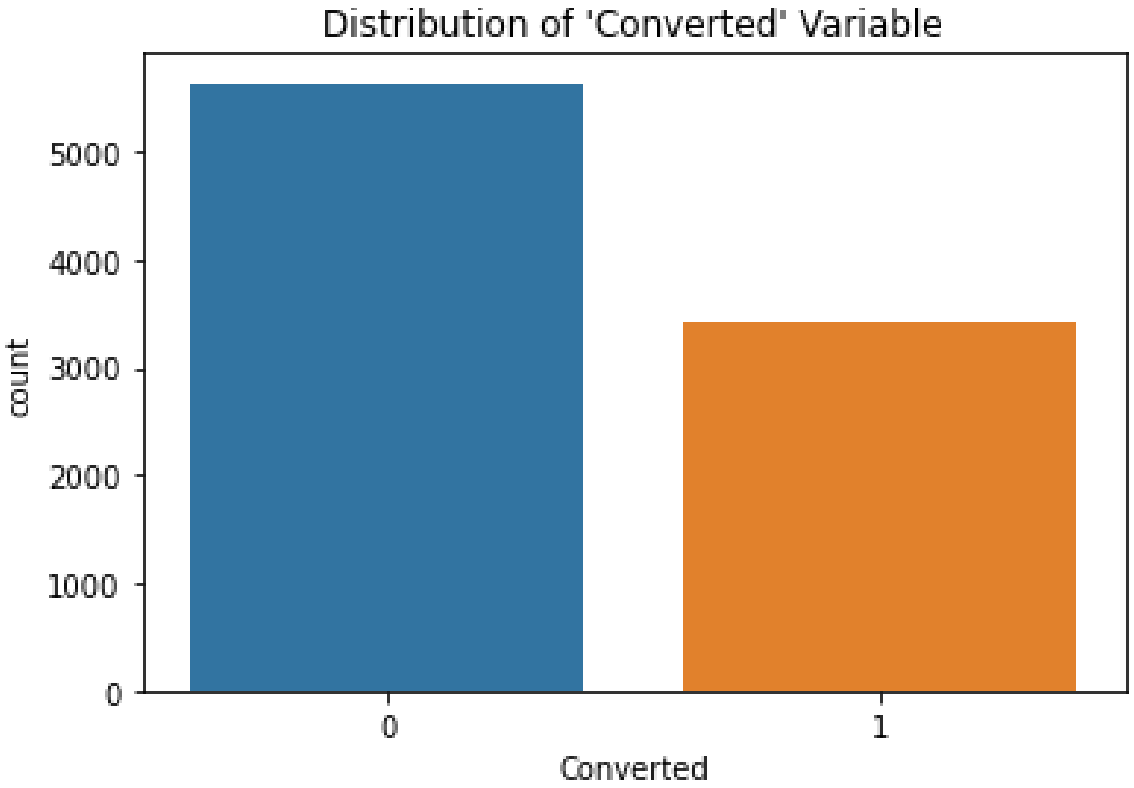
SMS sent as last activity has the highest conversion rate



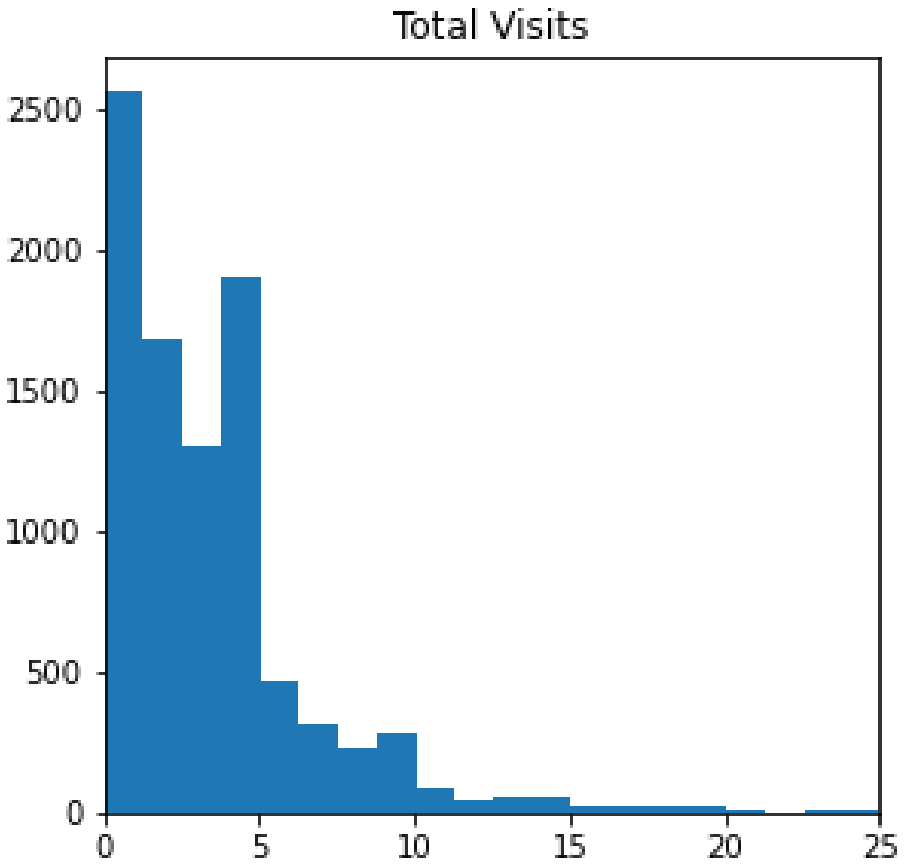


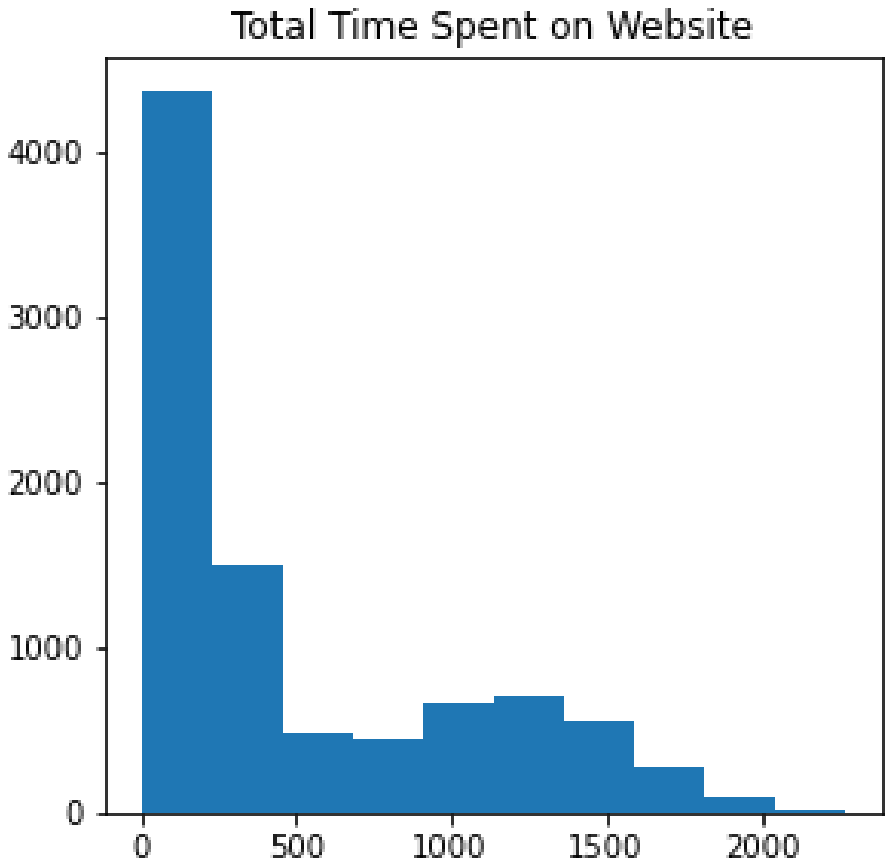
- Landing Page Submission and Lead Add Form attract most promising leads with high conversion rate.
- Google, Direct Traffic and Olark Chat are found important to secure hot leads.

EDA : Numerical Features

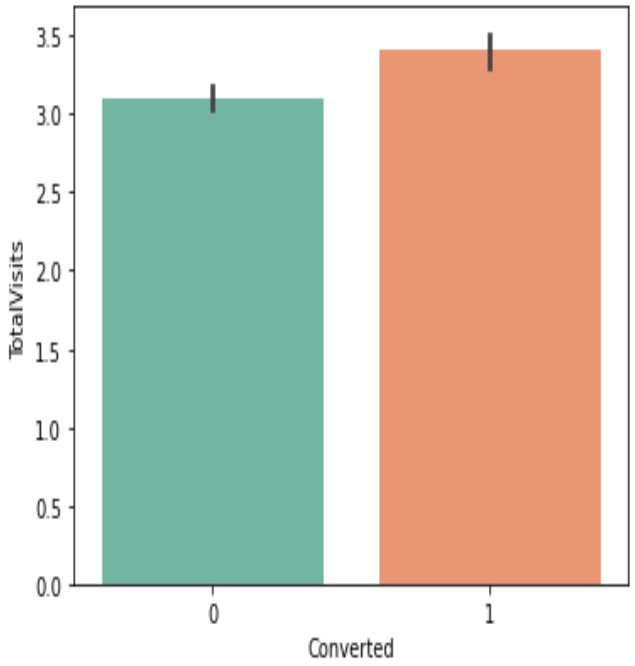


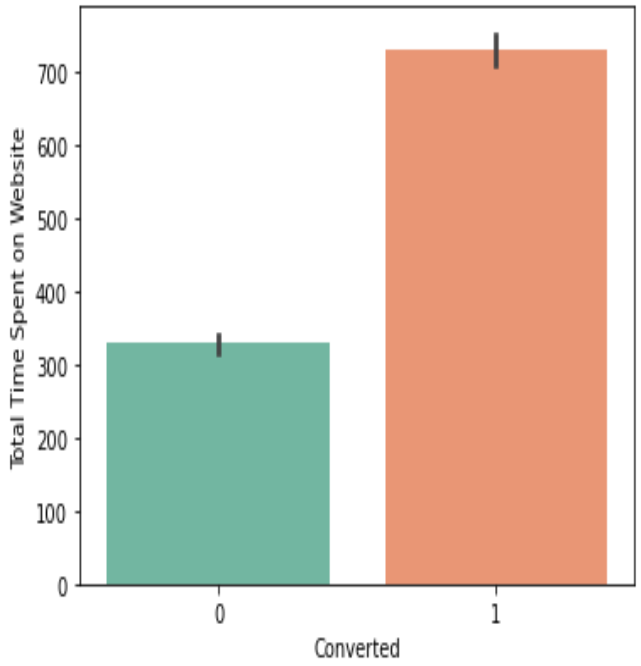
Current spread of lead conversion rate is poor. We need to increase it by building Regression model.

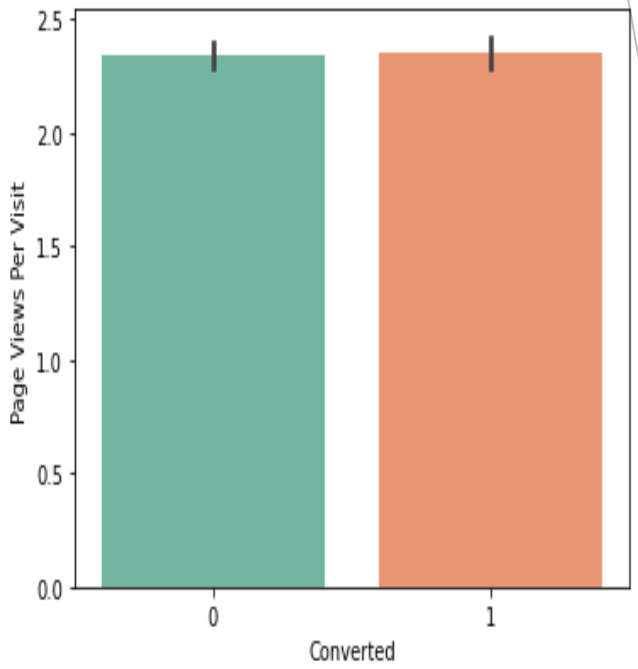




We observe our data is skewed.

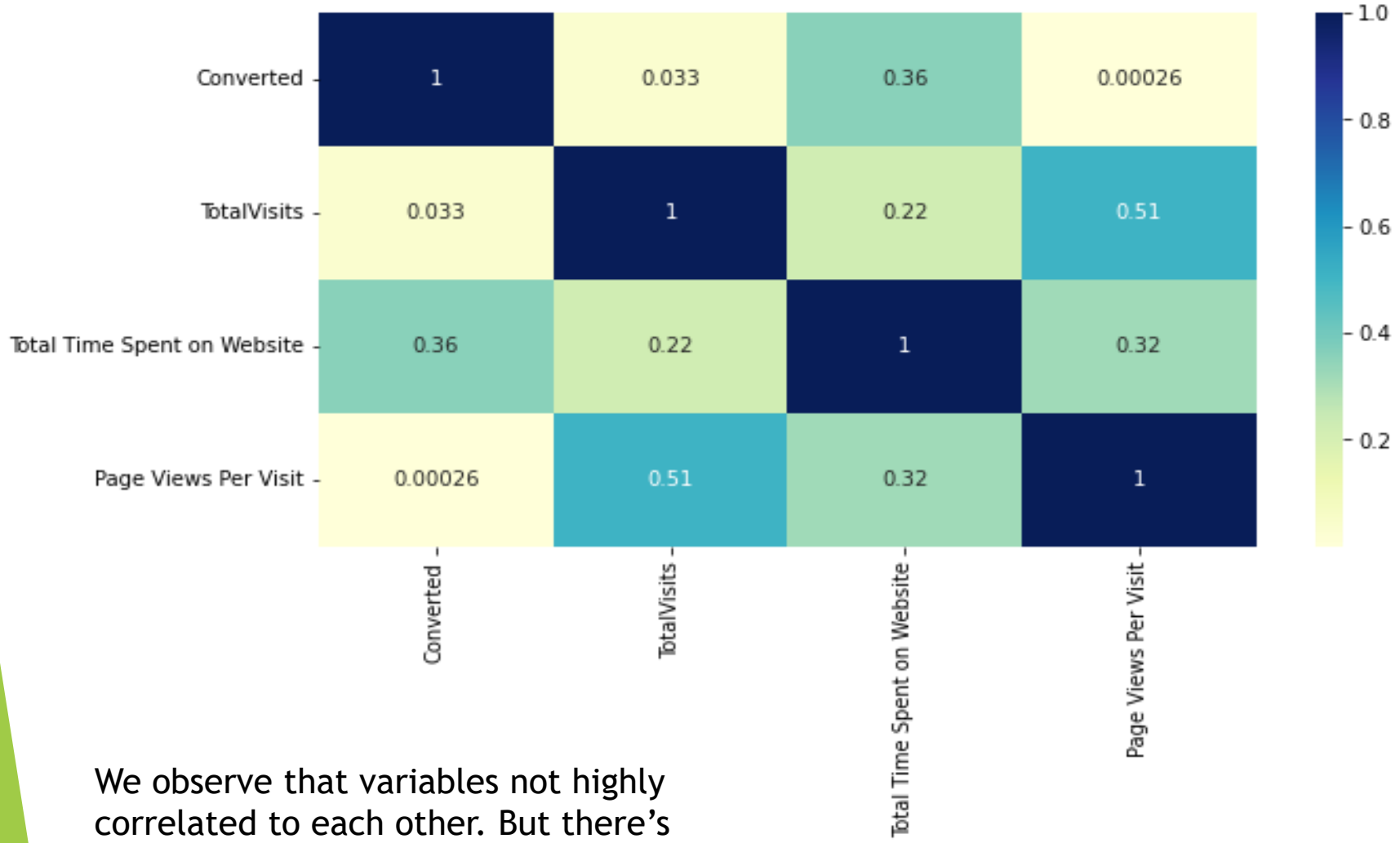






We observe conversion rate is highest for ‘Total Time spent on website’, followed by Total Visits.

Correlations Matrix



We observe that variables not highly correlated to each other. But there's multicollinearity among some features.

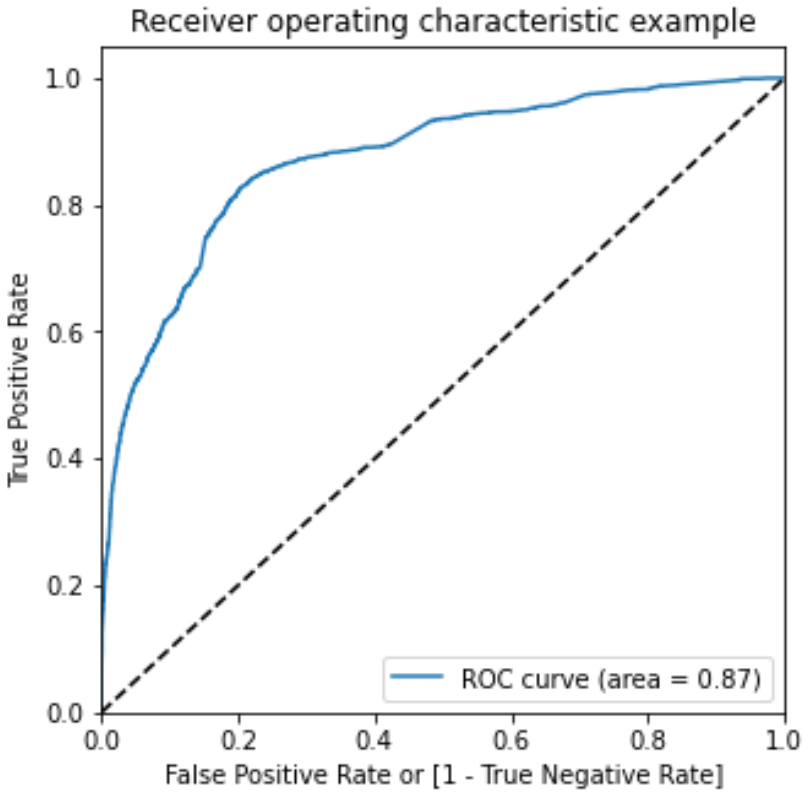
Data Processing

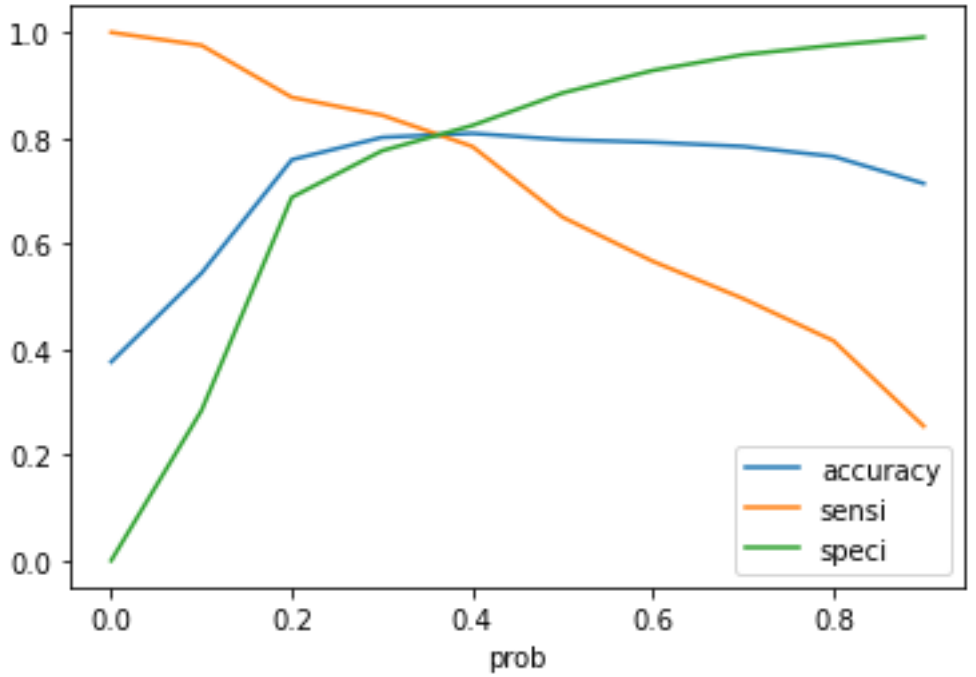
- Outlier Treatment – Total Visits and Total Time spent on website had some outliers. We performed capping by checking for 99th percentile.
- Dummy variables are created for categorical features as they enable us to use a regression equation on multiple groups.
- Numerical features are Normalized using MinMaxScaler.

Model Building

- Train – Test split was performed, we chose 70:30 ratio.
- Used RFE for feature selection with 15 variables as output.
- Model Building was done by removing variables with p-value > 0.05 and vif >5.
- Prediction test data set resulted Sensitivity = 80%

ROC Curve & Optimal Cut-Off Probability





ROC Curve represents how much model is able to distinguish between classes. Optimal cut-off probability is when specificity, accuracy and sensitivity gets balanced i.e. 36 %

Conclusion

- The customer who fills the form are potential leads.
- It’s better to focus on leads that spend significant time on our website.
- X Education must focus on Working professionals as they have high conversion rate.
- If the last activity is SMS sent or Email opened, he/she could prove a potential lead.
- If the lead source is Google, Direct Traffic or Olark Chat, customer is more likely to be a potential lead.
- It’s better to focus least on customers to whom the sent mail has bounced back.

Keeping above factors in mind, X Education could increase their conversion rate significantly.