Lead Scoring Case Study

Analysis on the Lead Scoring Case Study

Problem Statement

- To help X Education to select the most promising leads known as ‘hot leads’ who are most highly to convert into paid customers
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads where the leads with higher and lower lead score have a higher and lower conversion chance respectively
- To Identify the contributing variables and understand their significance which are strong indicators of lead conversion
- Identify the outliers, if any, in the dataset and justify the same
- Consider both technical and business aspects while building the model
- Summarize the conversion predictions by using evaluation metrics like accuracy, sensitivity, specificity and precision

Data Exploration

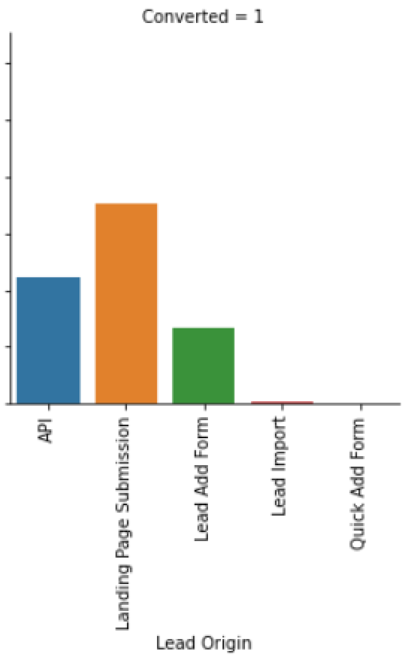
- The ‘**Leads.csv**’ has the dataset present having shape **(9240, 37)**
- The ‘**Leads Data Dictionary.csv**’ is data dictionary which describes the meaning of the attributes present in the “Leads” dataset

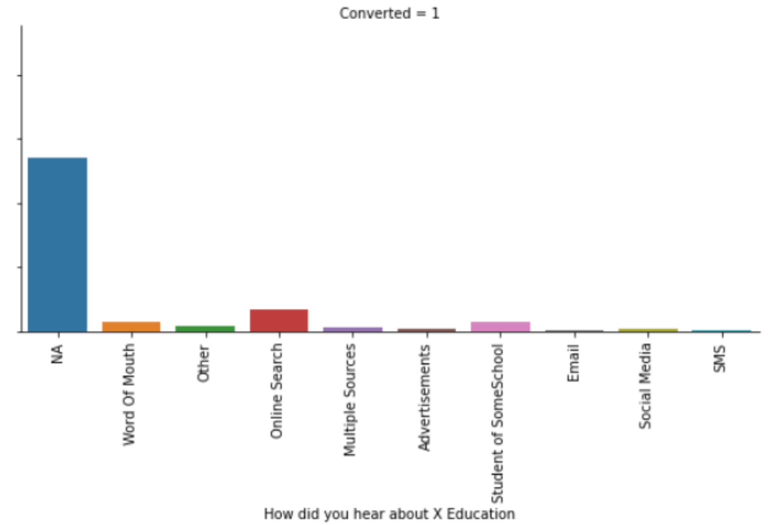
Data Cleaning and Preparation

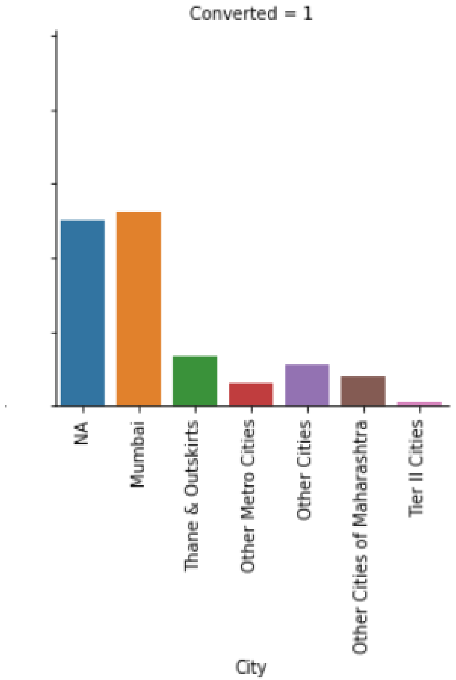
- We did the following
  1. Identified the null values and dropped columns which had more than 40% null values
     1. Lead Quality
     2. Asymmetrique Activity Index
     3. Asymmetrique Profile Index
     4. Asymmetrique Activity Score
     5. Asymmetrique Profile Score
  2. Some column “Select” as entries, converted it to “NA”
  3. Treated Categorical attributes missing value with “NA”
  4. Treated numerical attributes missing values with “Median”
  5. Dropped the un-necessary columns “Prospect Id” and “Lead number”

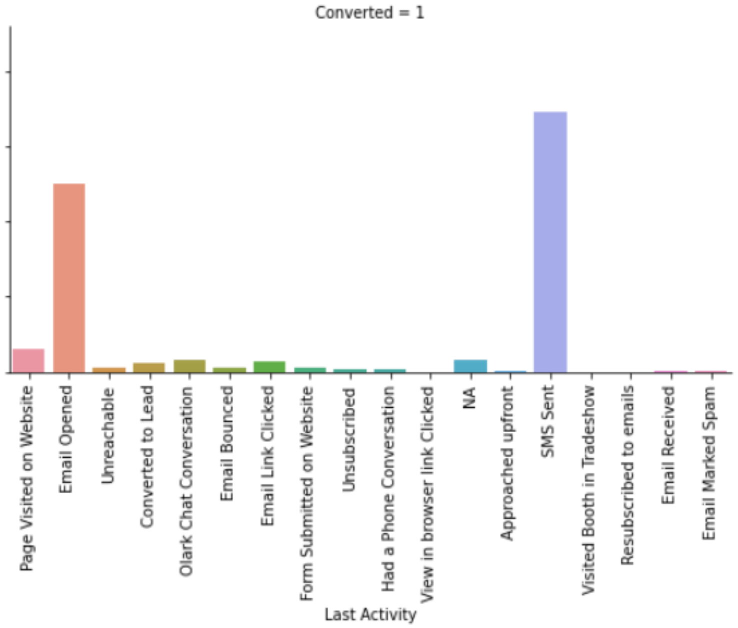
Data Visualization and Analysis

- The “Converted” column has been chosen as target
- Here we have counted the “Converted=1” for each attribute to find majority of the attributes which contributed in it
- Lead originated through “Lead Add Form” and “Quick Add Form” has high possibility of getting converted.



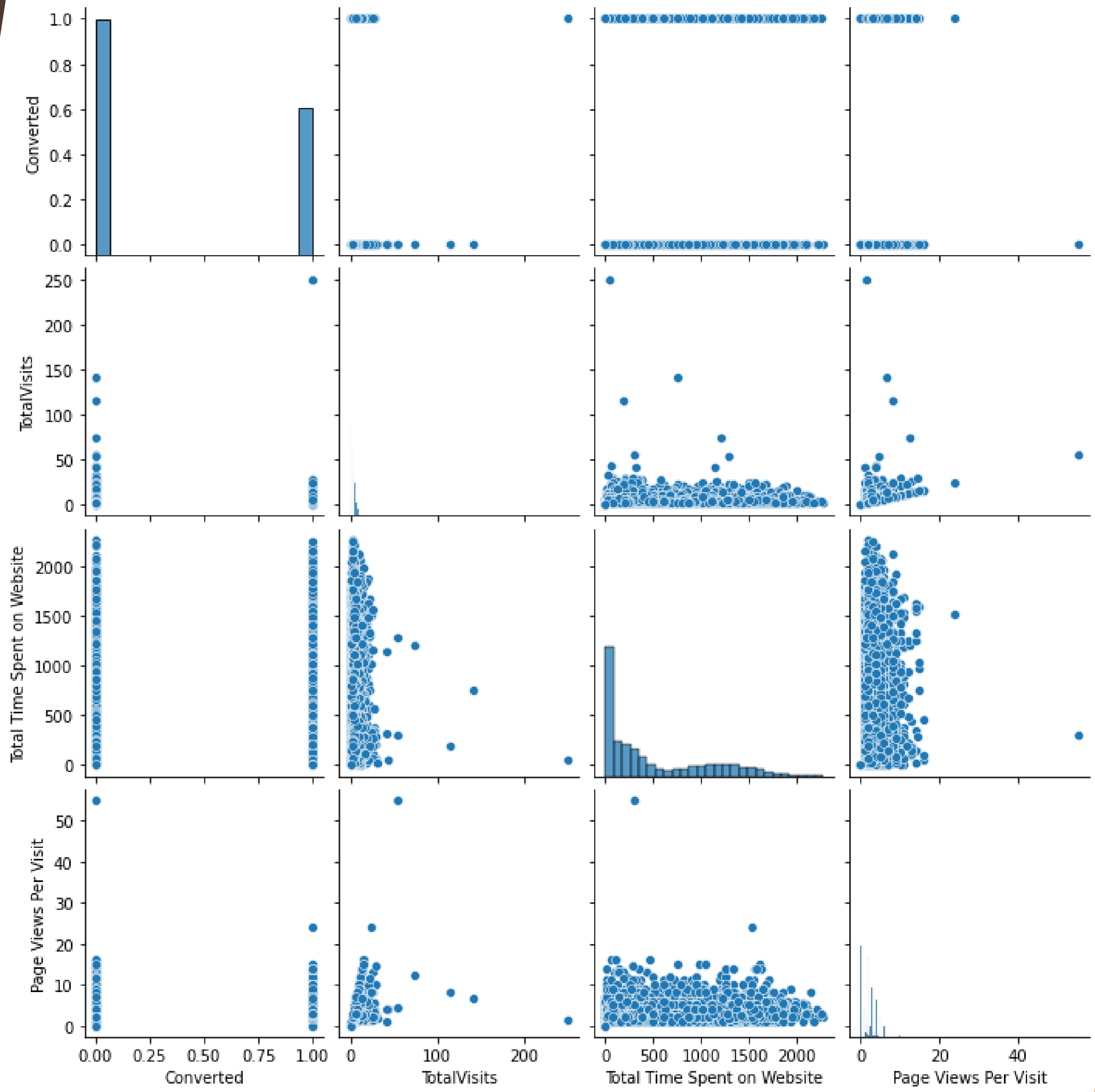






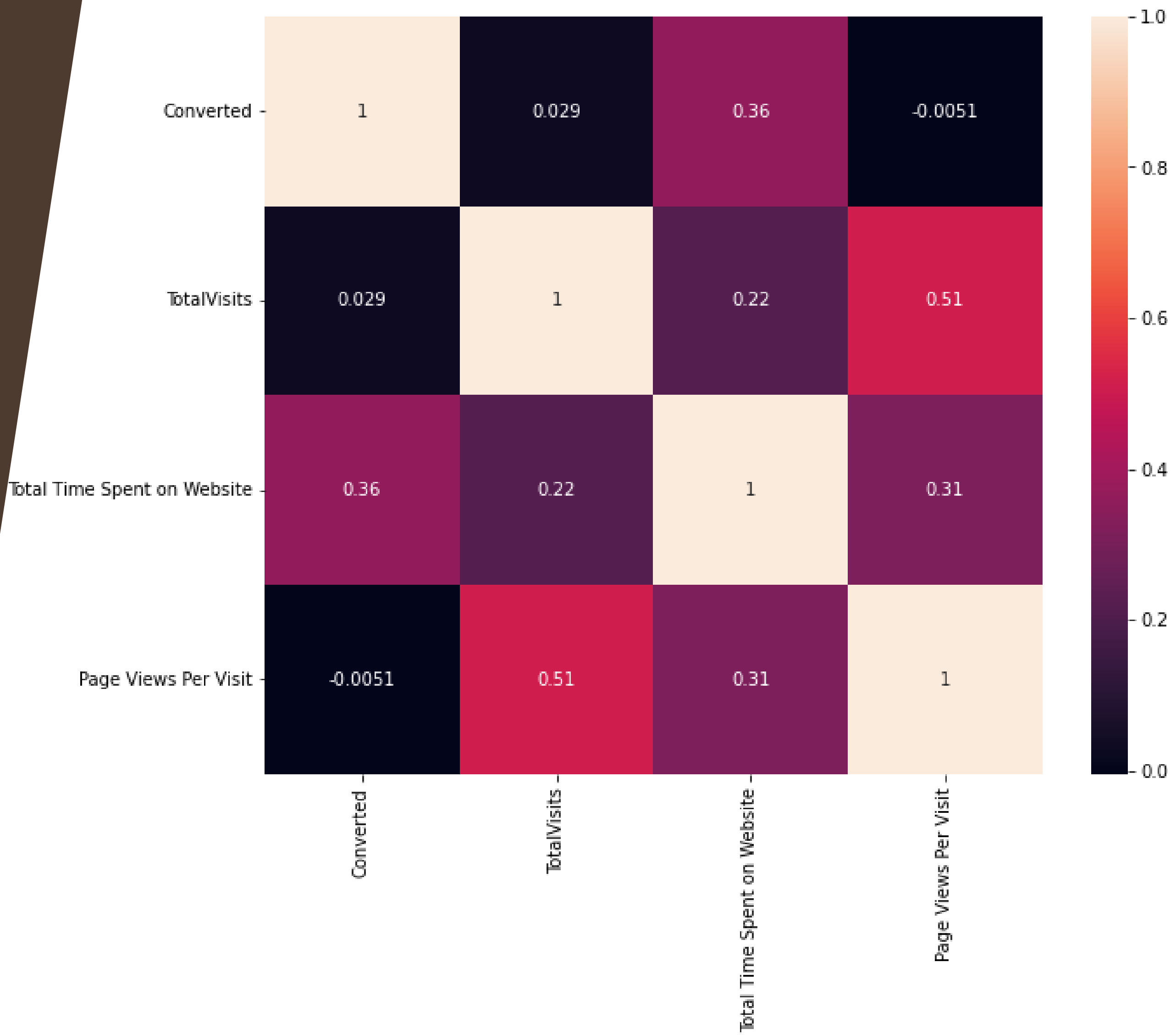
Data Visualization and Analysis

- For numerical category we plotted pairplot and heatmap



Data Visualization and Analysis

- As per paiplot and Heat map it shows that Total votes and Page views per visit shows max correlation
- Total time spent on website has some correlation with the person getting converted



Model Building: Data preparation

- Created Dummy variables for below attributes:
    1. Specialization
    2. What is your current occupation
    3. City
    4. Lead Origin
    5. Lead Source
    6. Last Activity
- We dropped the above columns post dummy variables creation
- We split the ‘Leads.csv’ dataset into train and test by the ratio of 70:30
- Train data will be used to train the model and test data will be used to test the model

Model Building: Data preparation

- We split the ‘Leads.csv’ dataset into train and test by the ratio of 70:30
- Train data will be used to train the model and test data will be used to test the model
- Using StandardScaler() function we did Feature Scaling so that all variables are on the same scale
- In order to avoid the dominance of any variable over another

Model Building: RFE

- Using RFE we shortlisted 15 out of 97 variables and dropped the rest of them:
  1. Total Time Spent on Website
  2. What is your current occupation_Housewife
  3. What is your current occupation_NA
  4. What is your current occupation_Working Professional
  5. Lead Origin_Lead Add Form', 'Lead Source_Direct Traffic
  6. Lead Source_Organic Search', 'Lead Source_Referral Sites
  7. Lead Source_Welingak Website', 'Last Activity_Converted to Lead
  8. Last Activity_Email Bounced', 'Last Activity_Had a Phone Conversation
  9. Last Activity_NA', 'Last Activity_Olark Chat Conversation
  10. Last Activity_SMS Sent

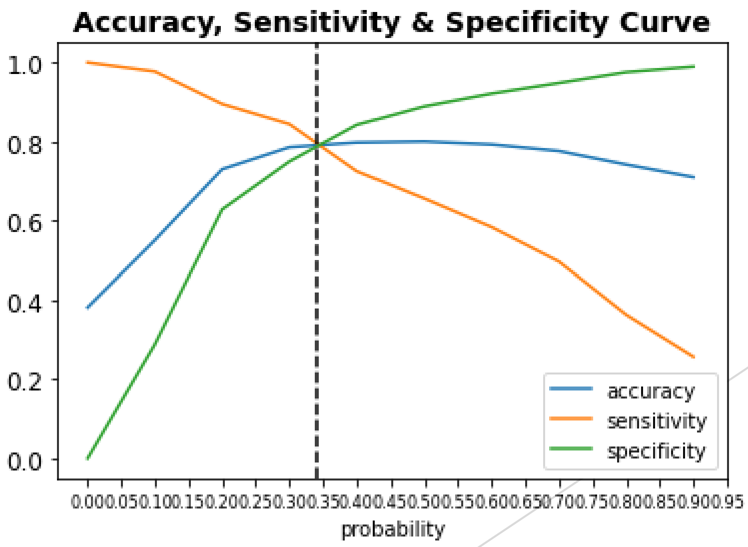
Model Building: Logistic Regression

- Using GLM (Generalised Linear ModeL) from StatsModels library, we built the Logistic Regression Model.
- Accepted P-value should be kept below 0.05 and VIF should be less than 5
- For the first Model #1 we got an attributes with high p-value “What is your current occupation_Housewife”, so we eliminated it
- For Model #2 we got an attributes with high p-value “Lead Source_Referral Sites”, so we eliminated it
- Likewise, till we got Model #5, we eliminated attributes and reached to final 11 attributes.
- As both p-values and VIF scores within their respective thresholds.

Model Building and Evaluation: Final Model interpretation

- Found the 11 most important attributes
- Assigned **Predicted to** 1, to conversion probability having greater than 0.5
- Created confusion matrix with the cut off to 0.5 but gave poor sensitivity
- Hence, found different probability cut offs to plot the Accuracy, Sensitivity & Specificity Curve. In order to find the threshold values
- Assigned **Predicted to** 1, to conversion

probability having greater than 0.35



Model Accuracy : 79.6 %

Model Sensitivity : 82.2 %

Model Specificity : 78.1 %

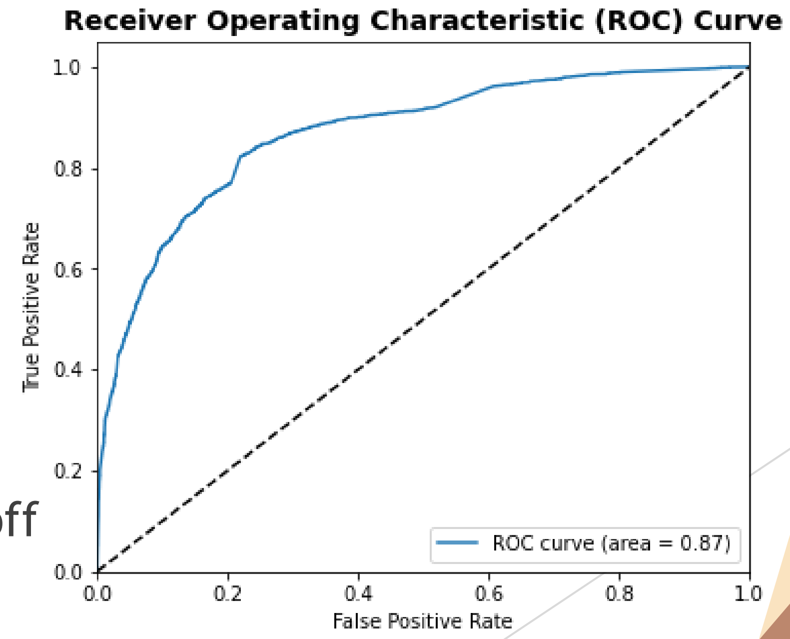
Model Precision : 69.8 %

Model Evaluation: ROC Curve

- Created new column 'Predicted_PRT' with value 1 if Lead_Score_Prob greater than 0.41
- Which gave below evaluation metrics with poor sensitivity

Model Accuracy	: 79.9 %
Model Sensitivity	: 72.0 %
Model Specificity	: 84.7 %
Model Precision	: 74.4 %

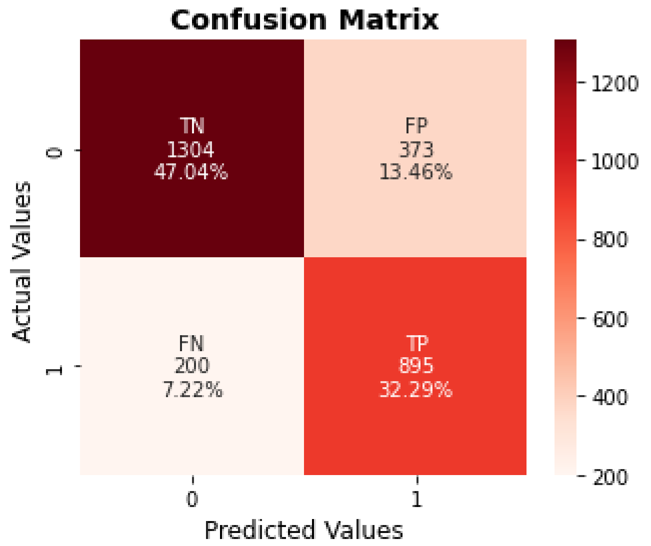
- Hence, selected 0.35 as the conversion cut off



Model Evaluation: Prediction on test data

- Assigned **Predicted to** 1, to conversion probability having greater than 0.35
- Which gave the below evaluation metrics

Model Accuracy	: 79.3 %
Model Sensitivity	: 81.7 %
Model Specificity	: 77.8 %
Model Precision	: 70.6 %



- We got minimum difference on train and test data's performance metrics, showing that our final model **didn't overfit training data** and is performing well as of now.
- High Sensitivity will ensure that almost all leads who are likely to Convert are correctly predicted.

Conclusion

- The top three variables in the model which contribute most towards the probability of a lead getting converted:
  1. Lead Origin
  2. What is your current occupation
  3. Last activity

- Top 3 categorical/dummy variables in the model which should be focused the most on to increase the probability of lead conversion are:
  1. **Lead Origin_Lead Add Form** – The leads who added the form
  2. **What is your current occupation_Working Professional** – The working professionals are having more chances for taking up the courses
  3. **Last Activity_SMS Sent** – The leads who were sent messages are having more chances

Conclusion

- In order to take at-most advantage of the model, below is necessary:
  1. The lead Score which is assigned by calculating ['Convert_Prob']*100 are highes potential customer and should be focused on, i.e. the “**HOT LEADS**”
  2. Approaching the hot lead would result in making the conversion rate high
  3. Better forecasting of the courses being sold
  4. Increase in the revenue of the company