Lead Scoring Case Study

PRESENTED BY –

SHYAMLI KUMARI

Business Objective

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

The aim of this case study is to Build a logistic model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

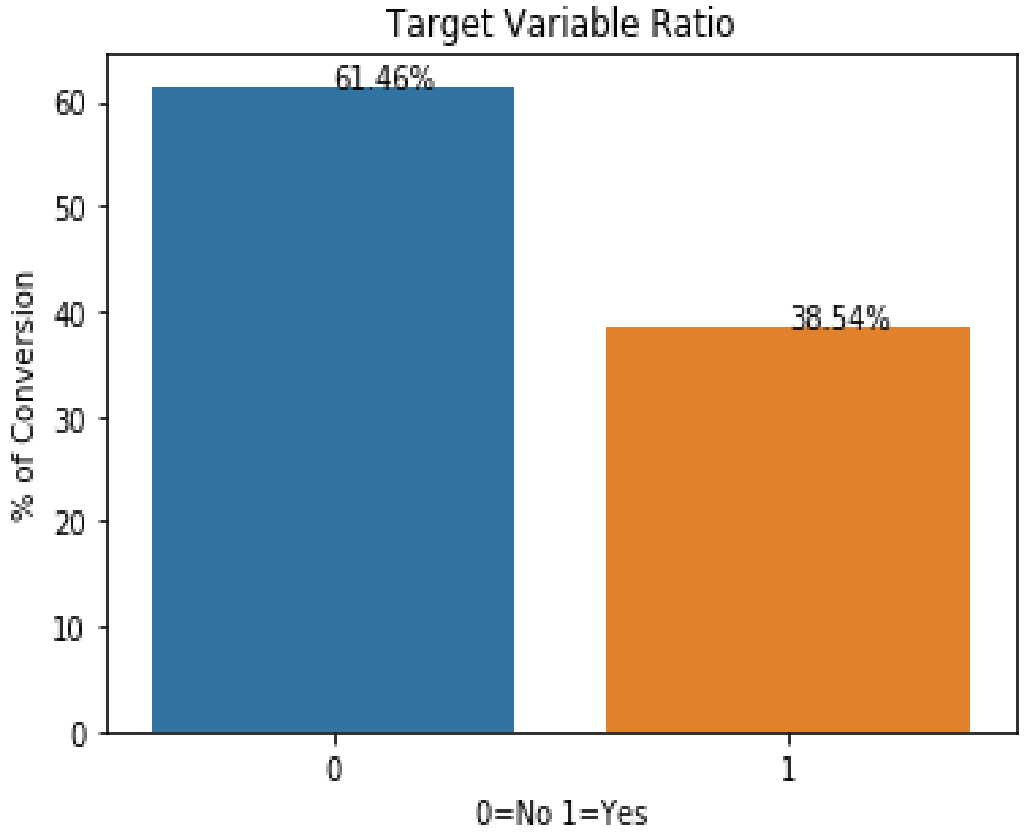
Solution Approach

The model is developed after following the below steps –

1. Data Understanding and Data Exploration
2. Data Cleaning
3. Data Visualization and Analysis
4. Data Preparation for model building
5. Model Building and Model Evaluation

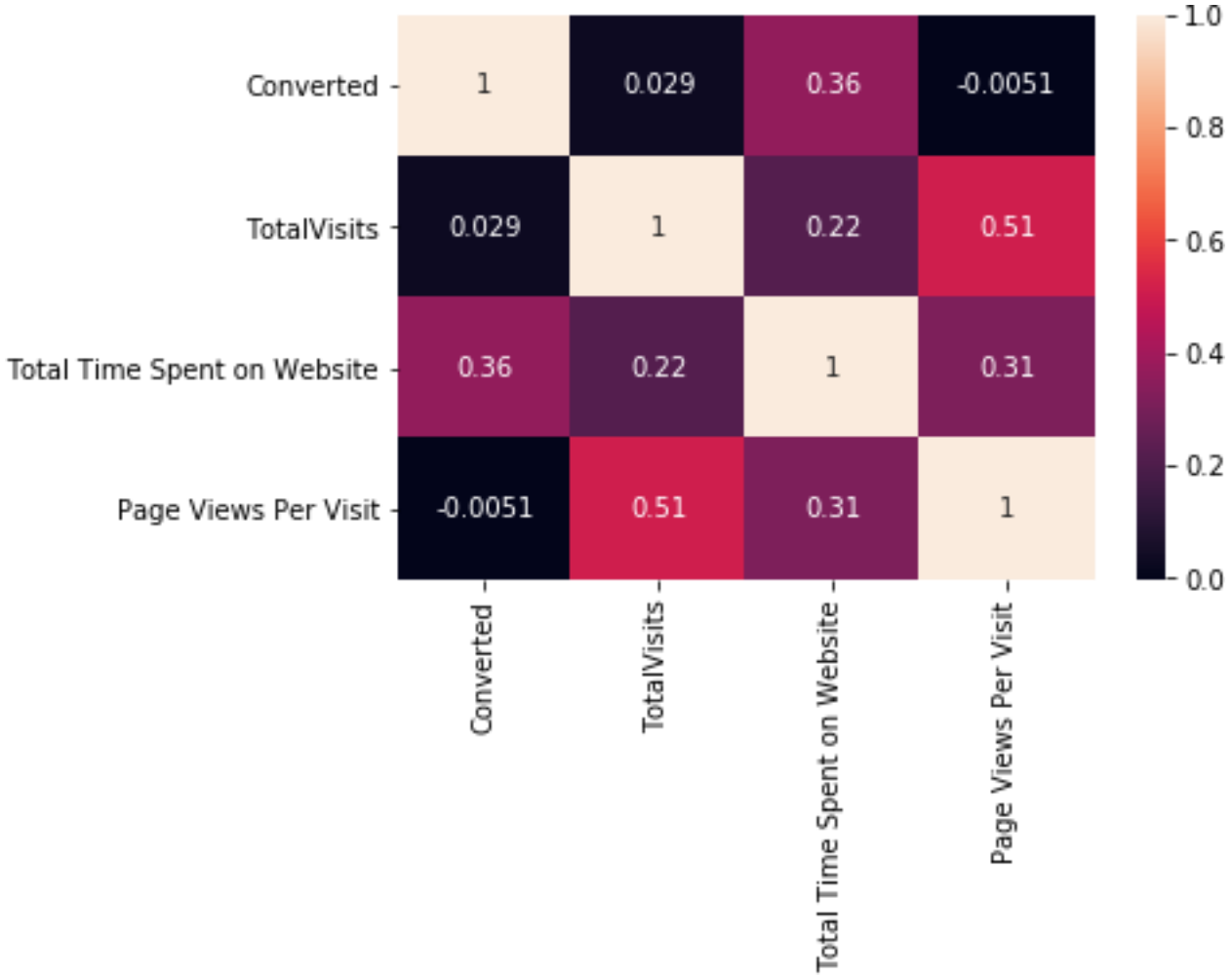
Data Distribution of target variable ‘Converted)

- The leads converted are less and are 38.54% of the total leads and the leads not converted are 61.46%



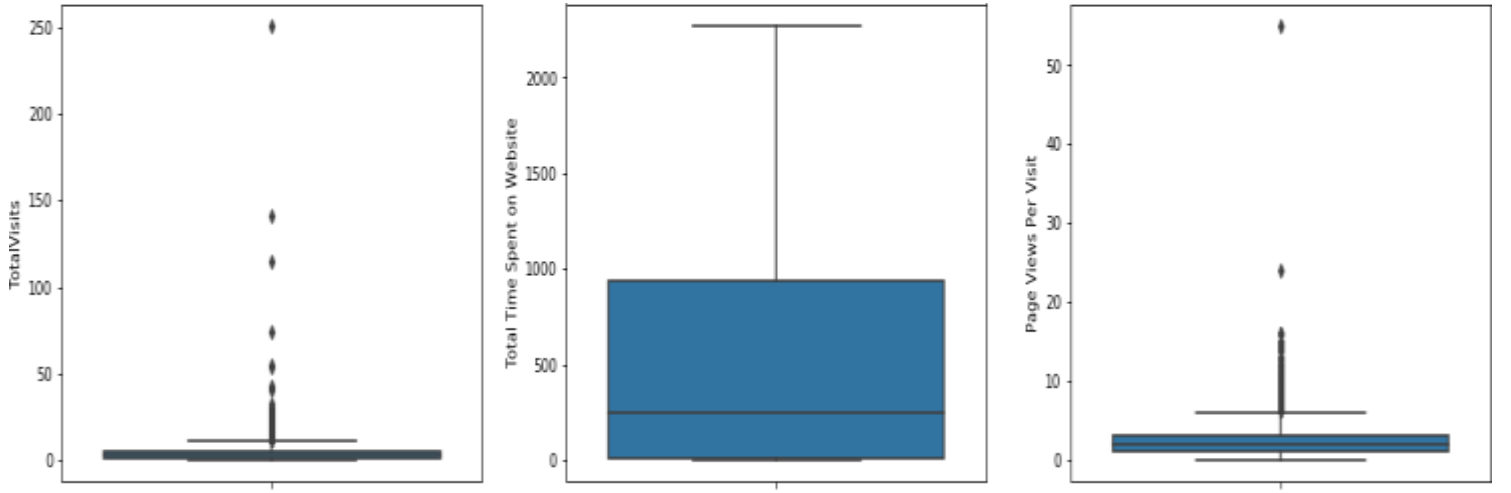
Correlation of numerical variables with Converted (target variable)

- It can be seen that for the target variable converted, the highest correlation is 0.36 with Total Time Spent on Website and the least is -0.0051 with Page Views Per Visit
- The highest correlation is between Total Visits and Page Views per visit hence it can be used to convert to leads

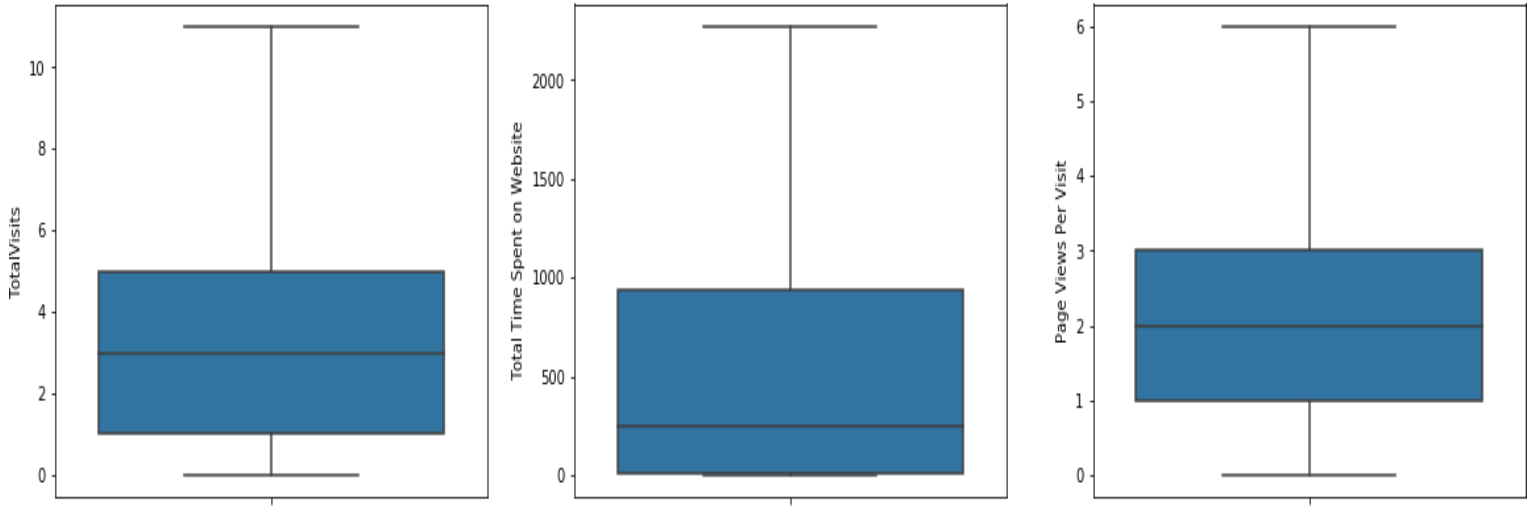


Outliers

- There are outliers present in Total Visits and Page Views per visit
- These are handled using imputing outliers with upper bound values

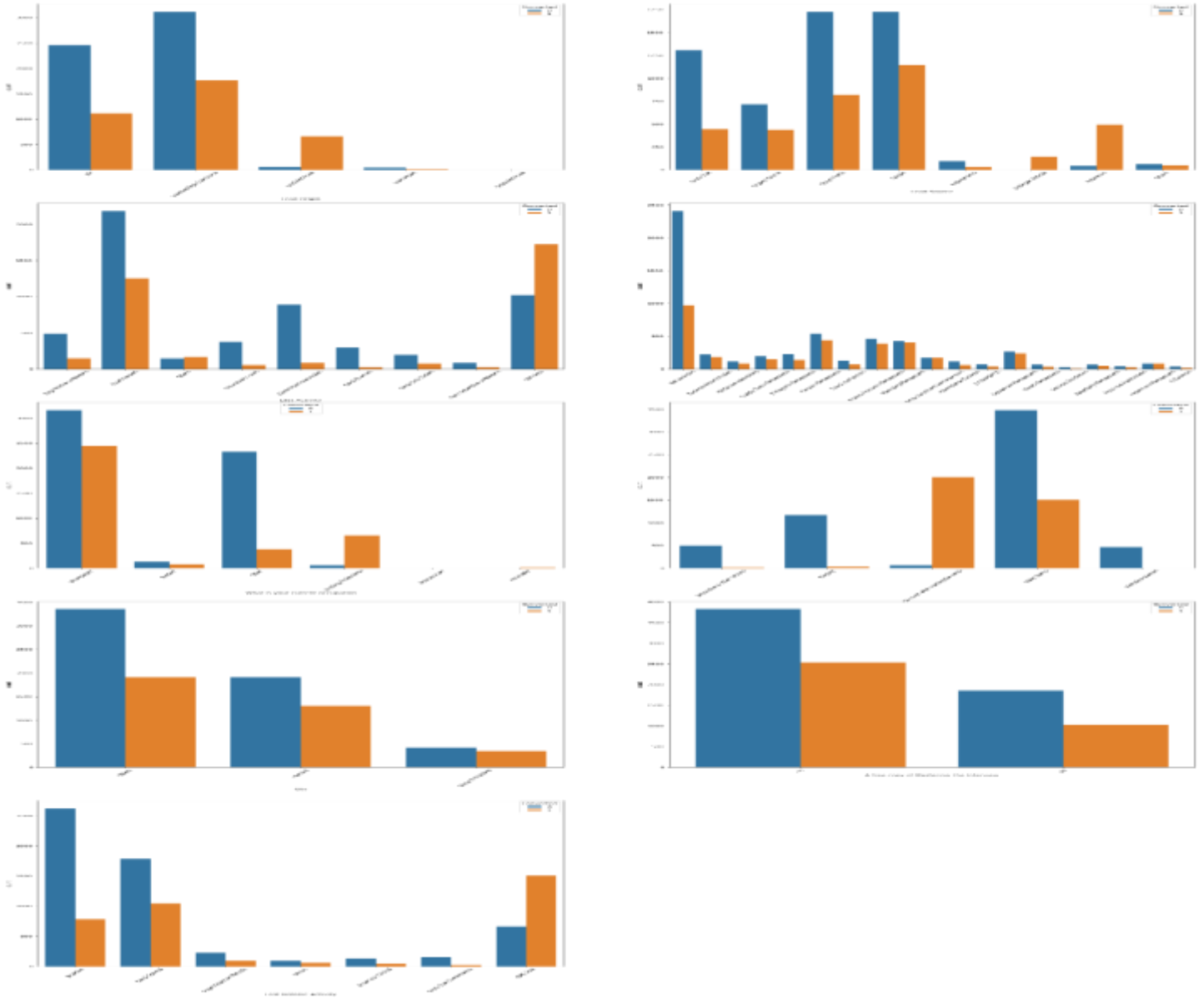


Box Plot before handling outliers



Box Plot after handling outliers

Categorical Variable Analysis with Converted (target variable)



Inferences

1. Lead Origin

- API and Landing Page Submission have most lead originated from them
- Lead Add Form has highest conversion rate but count of lead are not very high

2. Lead Source

- Direct Traffic and Google are generating high number of leads

3. Last Activity

- The leads from SMS sent and email open are higher and conversion rate is highest for SMS sent

4. Specialization

- The specialization not available has most leads and conversion rate is higher hence should be focused

Inferences

5. What is your current occupation

- Unemployed has the highest leads and conversion rate is also higher and Working Professional also have higher conversion rate

6. Tags

- Will Revert after reading the email have high conversion and Unspecified Tags have highest leads

7. City

- Other cities have highest leads but Mumbai and Thane and outskirts have highest conversion rate

8. A free copy of Mastering The Interview

- Leads from those who do not ask for free copy of Mastering Interviews are higher so should be focused for conversion.

9. Last Notable Activity

- SMS sent have high conversion rate

Model Building and Evaluation

1. Splitting of data is done in 70:30 ratio for training and test data respectively

2.For feature selection RFE was chosen as 15, multiple models were created taking under consideration – p value should be less than 0.05 and VIF should be less than 5

3. The final model has 11 features

4. The accuracy for model on training data is 86.42%

5. The average precision score is 0.87 and recall is 0.86 and f1 score is 0.86

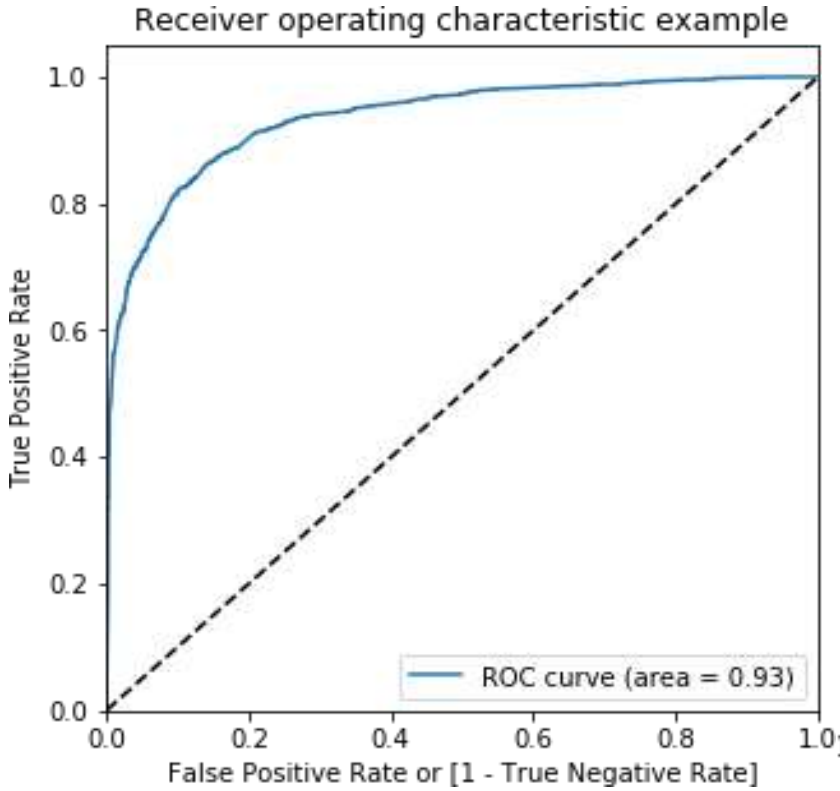
6. The optimal cutoff found out y plotting accuracy sensitivity and specificity for various probabilities is 0.27

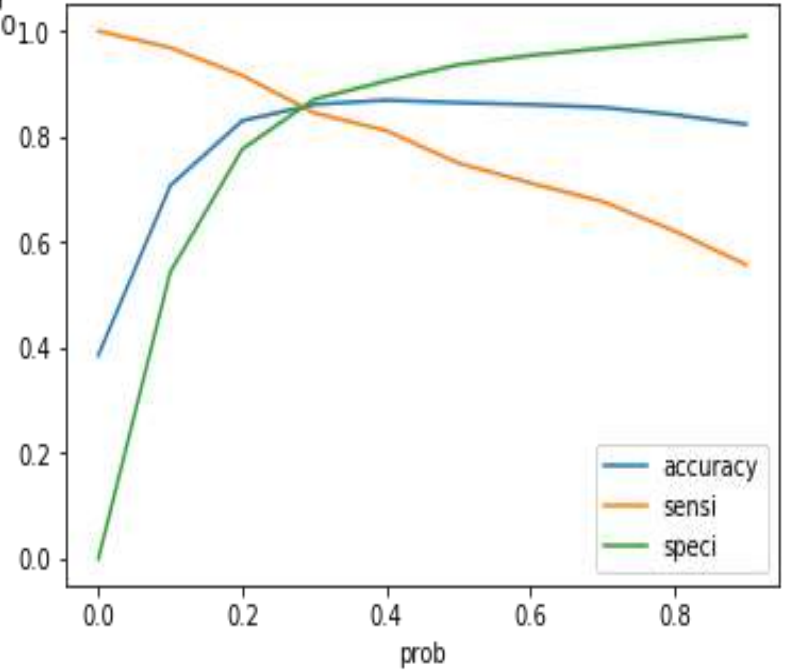
7. The accuracy for our model on test data is 86.04%

8. The average precision score is 0.86 and recall is 0.86 and f1 score is 0.86

ROC and Optimal Cut off

- The area under ROC is 0.93
- The optimal cut off is a probability at which there is a balance between sensitivity, specificity and accuracy – for our model it is 0.27





Conclusion

The features which are used in final model building with an accuracy of 86.42% are –

1. ‘What is your current occupation_Unemployed’
2. ‘TotalVisits’
3. ‘What is your current occupation_Other’ i.e. for which occupation detail is missing
4. ‘Total Time Spent on Website’
5. ‘Last Activity_Email Opened’
6. ‘Last Activity_SMS Sent’
7. ‘Lead Source_Olark Chat’
8. ‘Tags_Will revert after reading the mail’

Conclusion

9. ‘Lead Origin_Lead Add Form’
10. ‘Last Activity_Email Bounced’
11. ‘What is your current occupation_Student’

Recommendation

Based on Model Analysis

- The top features based on the coefficient values of the final model are – Total Time Spent on Website, Lead Origin_Lead Add Form, Tags_Will revert after reading the mail
- The company can decide a threshold time and if a lead spends more time than that then that lead should be chased
- If a customer tag is revert after reading mail then in this scenario company can set up follow up calls in order to get the lead conversion

Based on EDA

- Lead Add Form has highest conversion rate but count of lead are not very high so company should focus on acquiring more leads using that

Recommendation

- Direct Traffic and Google are generating high number of leads, so company should focus on how to increase higher lead conversion rate
- The leads from SMS sent and email open are higher and conversion rate is highest for SMS sent so company can focus on this
- The company should focus on unemployed customers as ell as the conversion rate is highest for this

Thank You