

## SUMMARY REPORT:

### Problem Statement:

An education company X Education sells online courses for industry professionals on their website and acquire leads through website forms, referrals & marketing campaigns. The potential leads or else known as Hot leads have a conversion rate of 30% which is low. So, they want a model which identifies potential leads which gives them high conversion rates.

So, we have built a logistic regression model to help the company in identifying the potential leads.

The model building process followed the bellows steps:

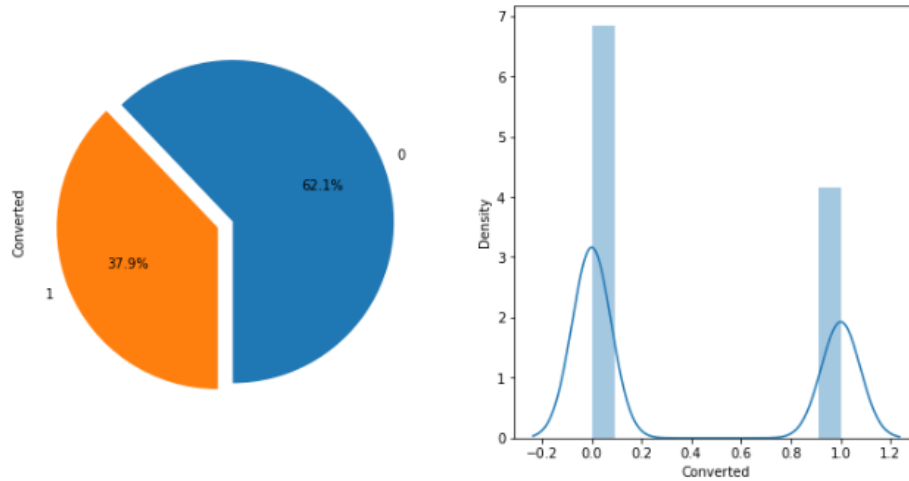
#### 1. Data Cleaning & Manipulation:

- Identified the unique values in the columns of data & dropped which are less important for the analysis.
- Some of the category columns have values as SELECT which are as good as nulls were replaced by Not Available.
- Columns having the null values are identify those having >35% nulls were dropped to make the data precise.
- The rows that having less % of nulls were dropped.

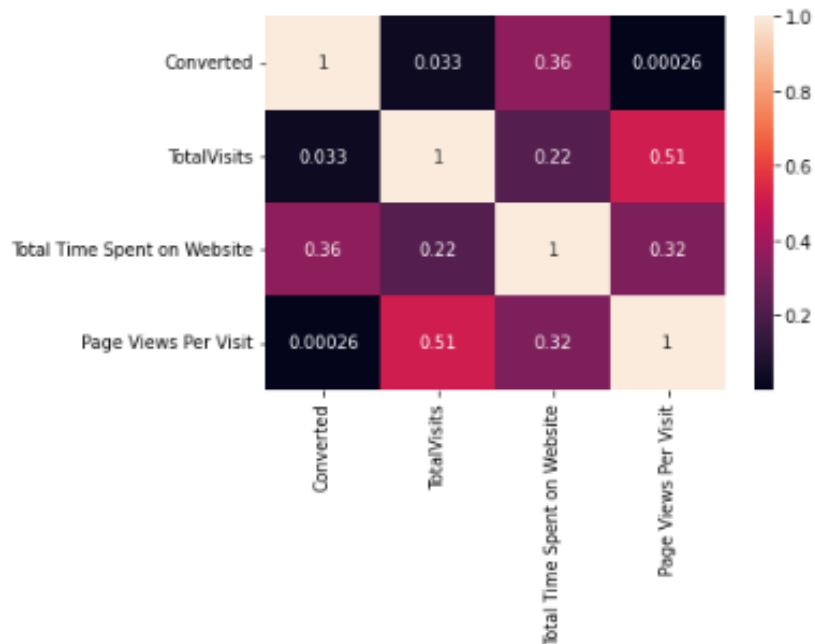
#### 2. Exploratory data analysis:

- The probability of lead getting converted is high when 'Lead Origin' is from 'Lead add form'.
- The Large number of leads comes from Google and direct traffic but the referral sites convert most of the leads.
- Leads opting for emailing option have more probability of getting converted.
- The Conversion rate is higher when the information is sent through SMS
- Unemployed people have more conversion rate as well as more count.

# LEAD SCORING CASE STUDY



From the plot we can observe that 37.9% are converted & rest are not converted



From the heat map above we can say Total visits shows high correlation with pages per visit

## 3. Data Modelling:

## LEAD SCORING CASE STUDY

---

- The dummy variables were created and the Train - Test split was done at 70% and 30% respectively.
- Model Building: RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value.
- Model Evaluation: A confusion matrix was made. Later, the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.
- Prediction: Prediction was done on the test data frame and got the optimal cut-off of 0.35 and the final model gave the accuracy of 80% with 79% sensitivity and 80% specificity.
- Precision – Recall: This method was also used to recheck and a cut off of 0.41 was found with precision of 79% and Recall of 69% on the test data frame.

### **Conclusion:**

- 1) From the given data the conversion rate was 39% .
- 2) When 'Lead Origin' is 'Lead add form', the probability of lead getting converted is high.
- 3) Google and direct traffic generating majority of leads, but referral sites shows high Conversion rate.
- 4) Most of the leads are opting for mailing them.
- 5) Conversion rate is higher when the information is sent through SMS .
- 6) Unemployed people have more conversion rate as well as more count.
- 7) Our model gives accuracy of 81% with our selected cut-off of 0.5.
- 8) An optimal cut-off comes out to be 0.35 which gives the accuracy of 80% with sensitivity and specificity at 79% and 80%.
- 9) Lead Origin Lead Add Form, Last Notable Activity Had a Phone Conversation, What is your current occupation, Working Professional are the features shows high promising leads.
- 10) Last Activity Converted to Lead, Last Activity\_Olark Chat Conversation shows less lead conversion rate.

LEAD SCORING CASE STUDY

---