LEAD SCORING - Case Study

By,
DRISHYA C
AVI DUTTA
ARTH GANDHI

Problem Statement

An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goals

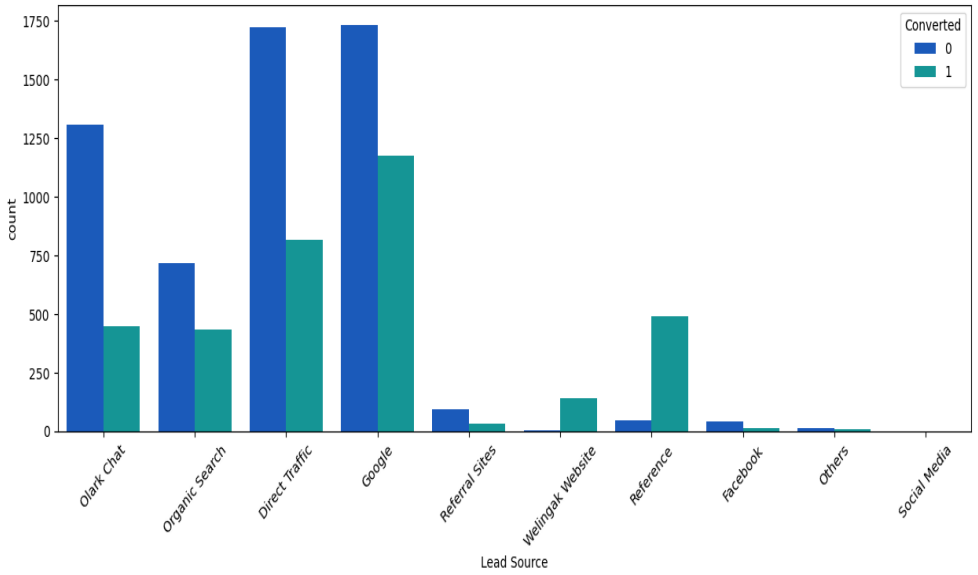
Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

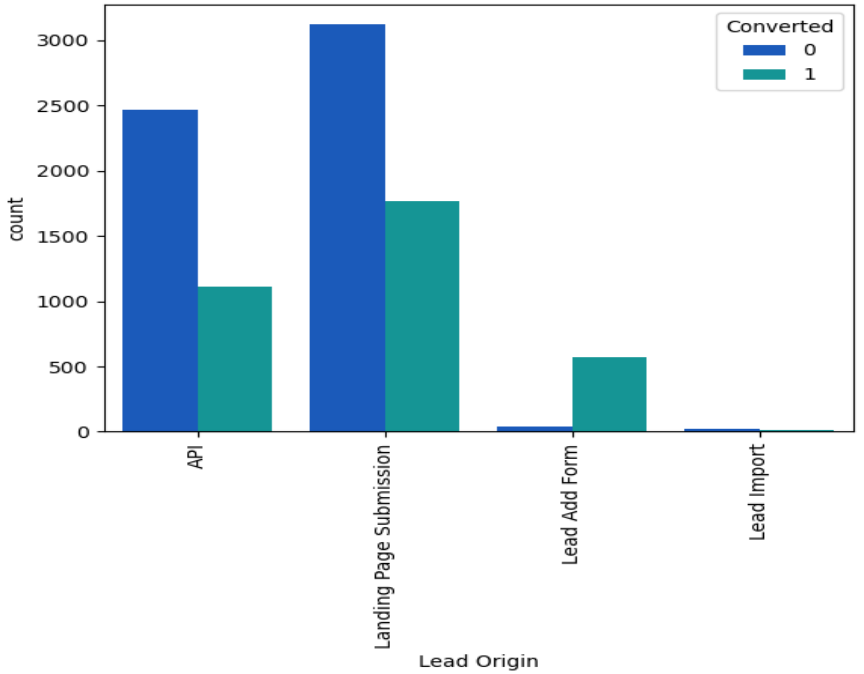
A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Approach

- importing data.
- Clean up and get the data needed for more analysis.
- Exploratory data analysis to identify the traits that are most helpful for conversion aspects of scaling.
- Get the data ready for model building.
- Construct a logistic regression model.
- Each leads should be given a score.
- On a train set, test the model.
- Model evaluation using multiple metrics Use a test set to run the model.
- Check the model's accuracy and other assessment criteria.

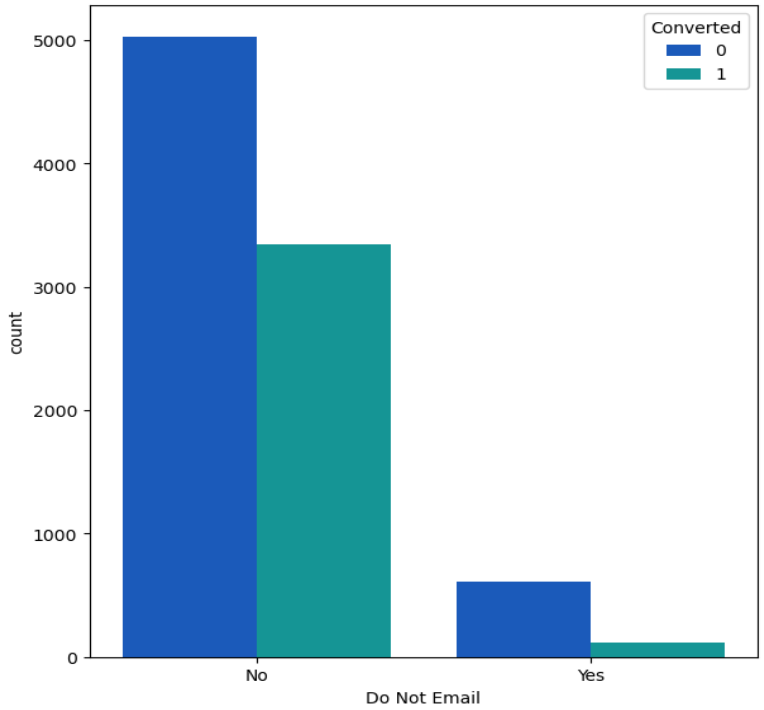
EXPLORATORY DATA ANALYSIS

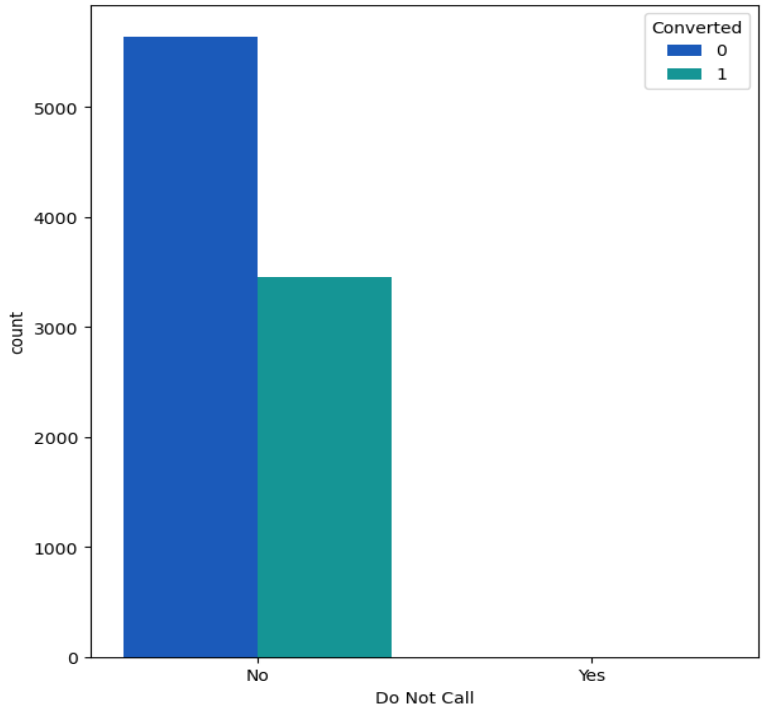




- Maximum Leads are generated by Google and Direct Traffic. - Conversion rate of Reference leads and Welinkgak Website leads is very high.
- Landing page submission is the highest converted lead.

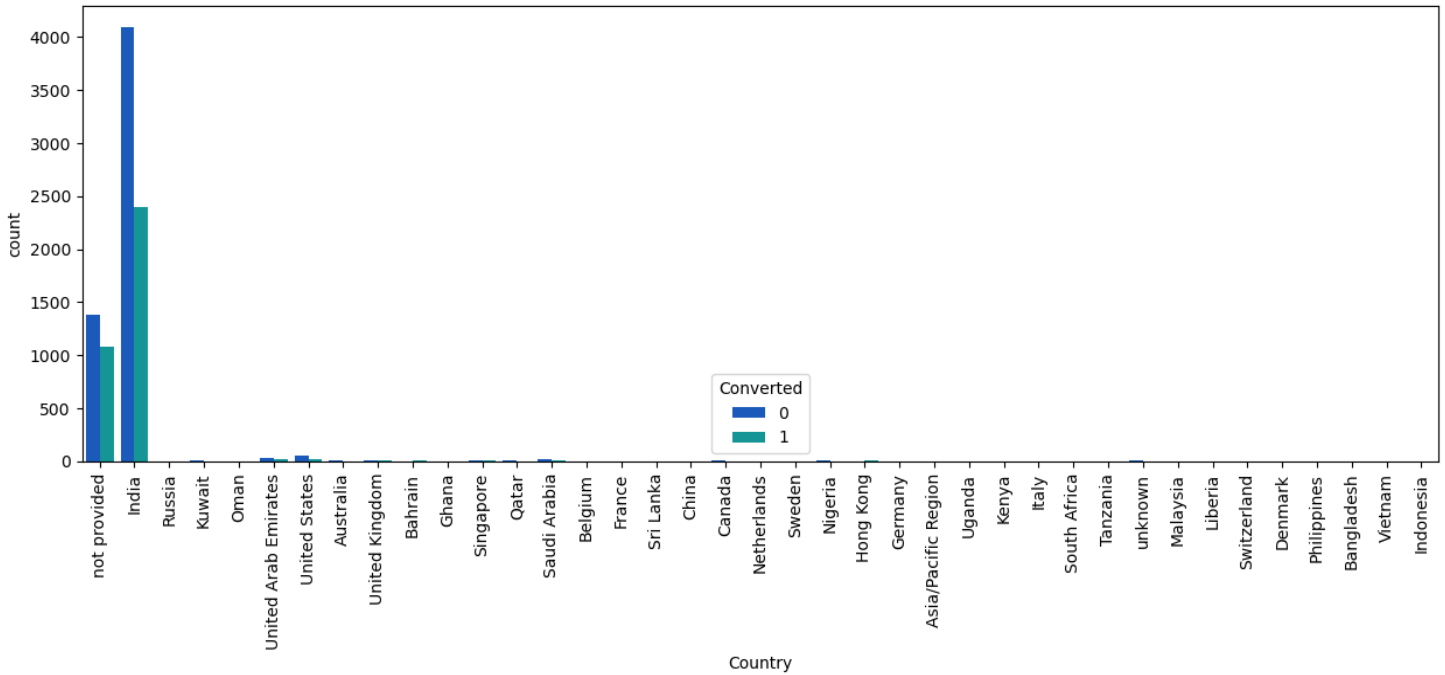
Do not Email Vs Do not Call





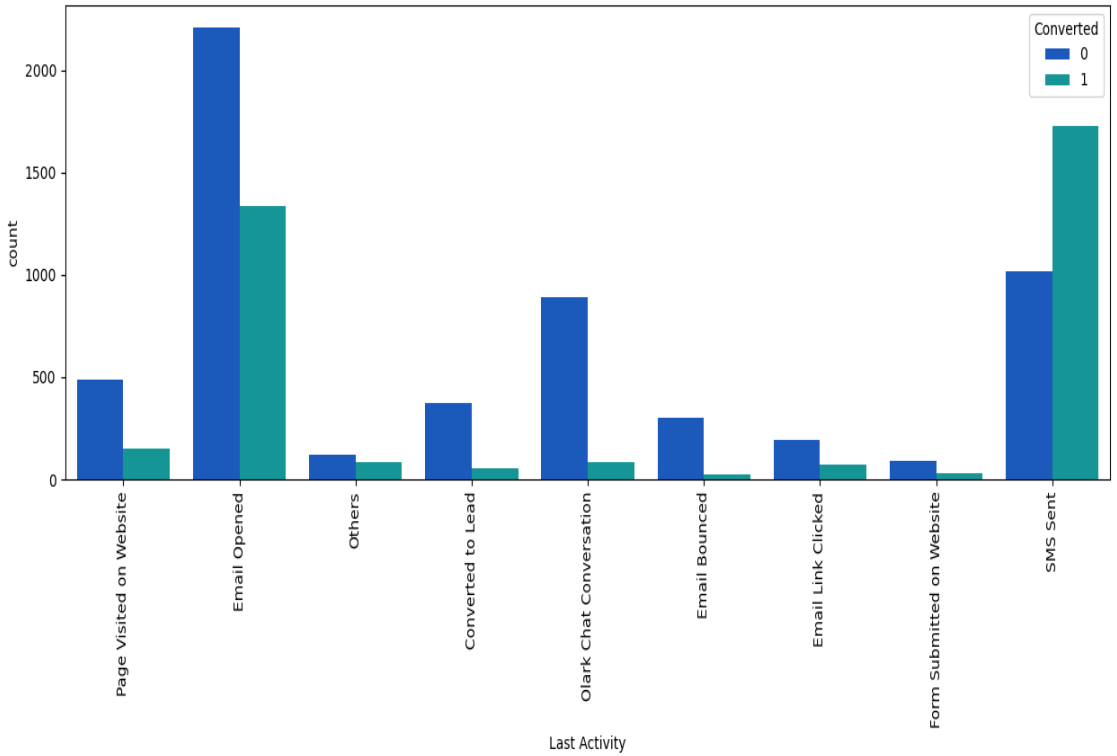
People who do not want to receive emails and people who want to receive emails both have negative conversion rates

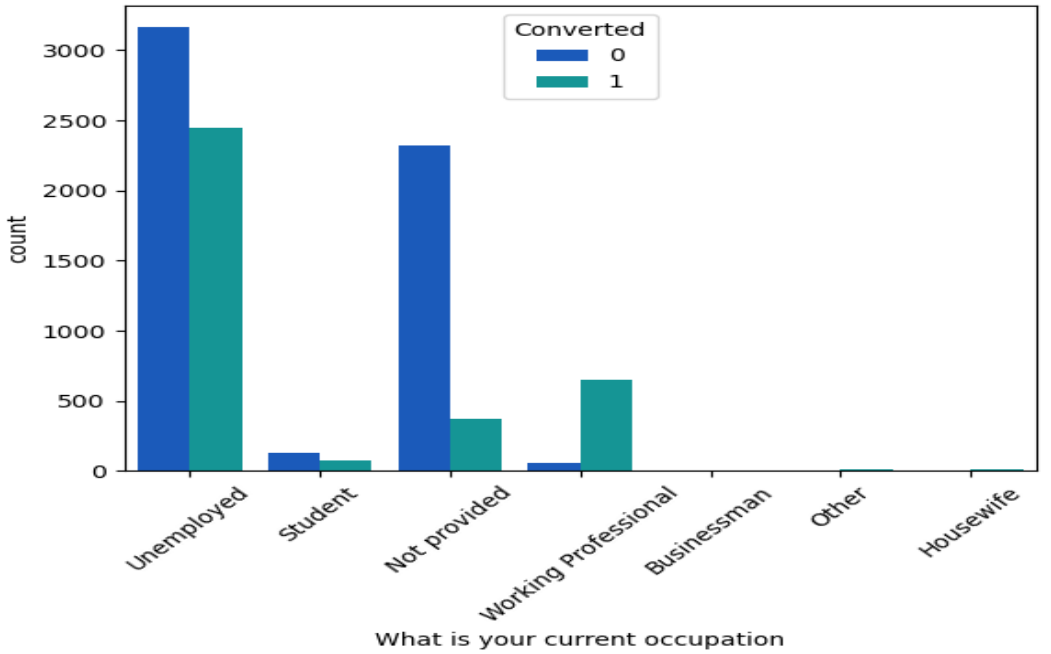
COUNTRY



we can see that most of the data consists of value 'India', no inference can be drawn from this parameter.

Last activity and what is your current occupation



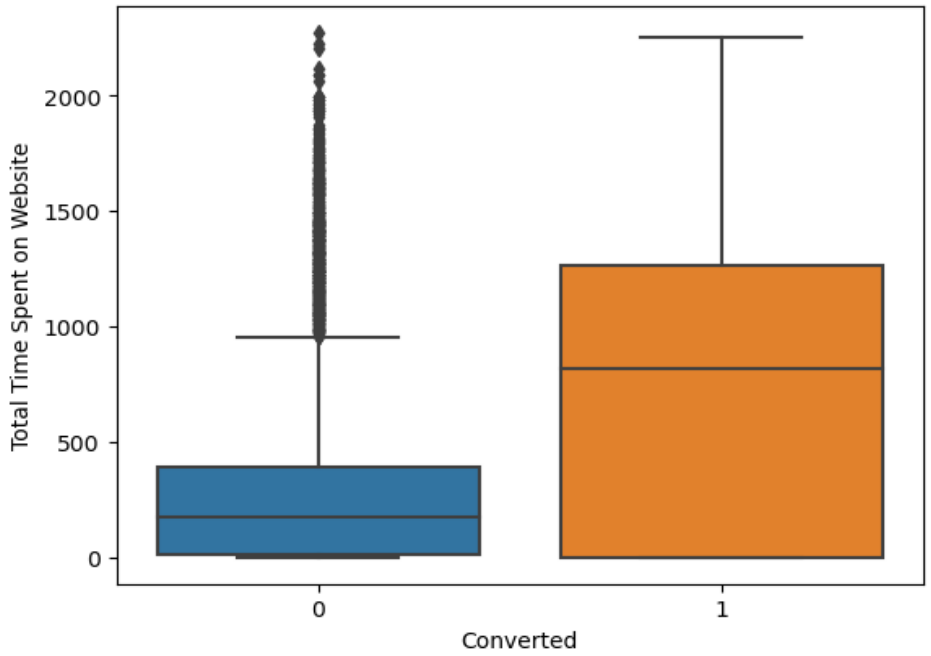


Maximum leads are generated having last activity as Email opened but conversion rate is not too good. - SMS sent as last activity has high conversion rate.
Maximum leads generated are unemployed and their conversion rate is more than 50%. - Conversion rate of working professionals is very high.

Numeric variable analysis

INFERENCE:
The target variable that is converted and Total time spent on website have negative correlation.

Total time spent on website

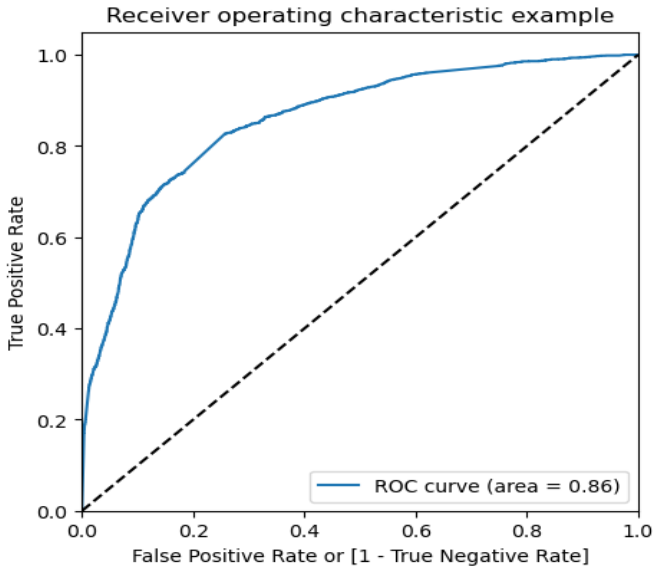


leads spending more time on website are more likely to convert , thus website should be made more engaging to increase conversion rate

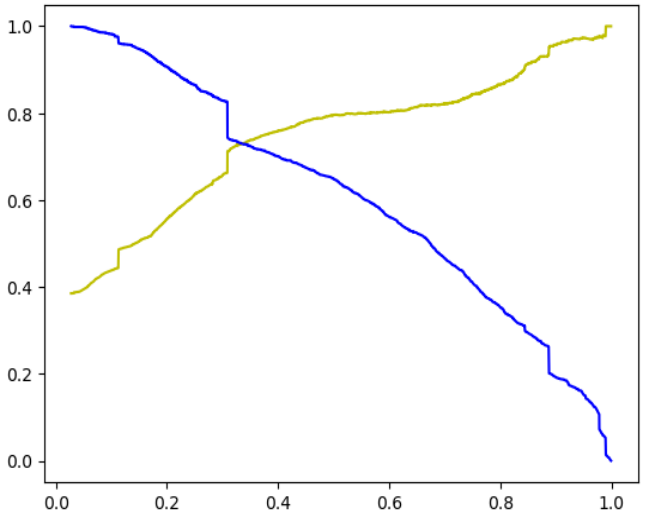
MODEL BUILDING

- Dividing the data into train and test sets as 70:30; scaling the variables in the train set.
- Model Building using Stats Model & RFE
- removing variables based on high p-values;
- checking the value of the VIF for all the existing columns;
- predicting using the train set; assessing accuracy and other metrics;
- predicting using the test set; and performing precision and recall analysis on test predictions.

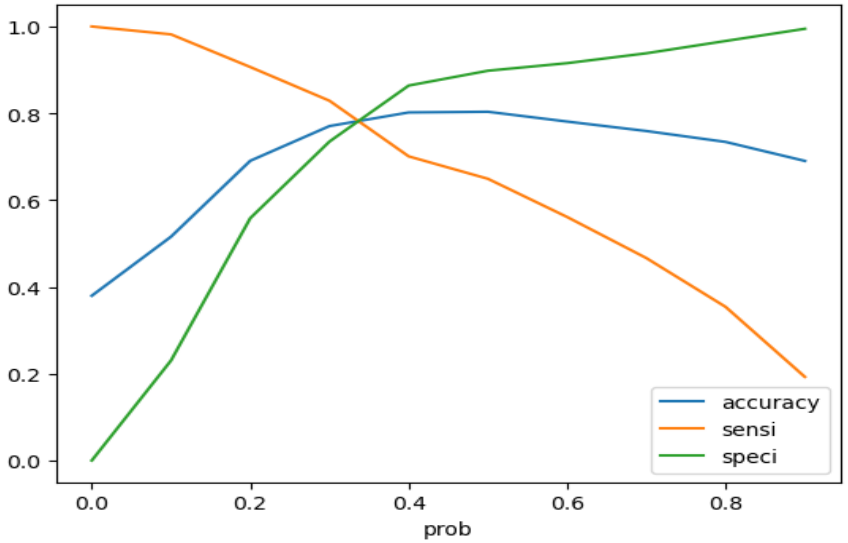
ROC Curve



Precision Recall curve



Accuracy, **sensitivity** and specificity



	prob	accuracy	sensi	speci
0.0	0.0	0.379630	1.000000	0.000000
0.1	0.1	0.515694	0.981811	0.230458
0.2	0.2	0.690521	0.906573	0.558310
0.3	0.3	0.770559	0.828855	0.734885
0.4	0.4	0.801946	0.700703	0.863901
0.5	0.5	0.803515	0.649029	0.898052
0.6	0.6	0.780917	0.560976	0.915507
0.7	0.7	0.759102	0.466308	0.938275
0.8	0.8	0.733992	0.353865	0.966608
0.9	0.9	0.690207	0.192642	0.994688

A good value of 0.86 indicating a good predictive model. Means the model is good to go.

Result From Train and Test set

Train set

- Accuracy – 77.05%
- Sensitivity : 82.89%
- Specificity : 73.49%

Test set

- Accuracy – 77.52%
- Sensitivity : 83.01%
- Specificity : 74.13%

CONCLUSION

-Below listed are Important features responsible for good conversion rate

- Lead Origin_Lead Add Form
- What is your current occupation_Working Professional
- Total Time Spent on Website

-Accuracy, Sensitivity and Specificity values of test set are around 77%, 83% and 74% which are approximately closer to the respective values calculated using trained set.

- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- Hence overall this model seems to be good.

Recommendation

X Education Company needs to focus on the following main aspects to increase the conversion rate :

- Increase on sending SMS notification since this helps in higher conversion.
- Get total visits increased by advertising etc, since this tends to give high conversion.