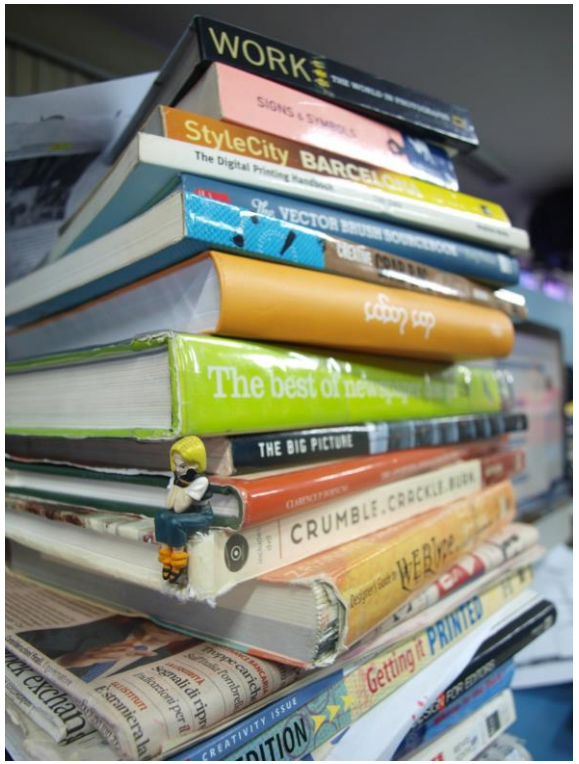
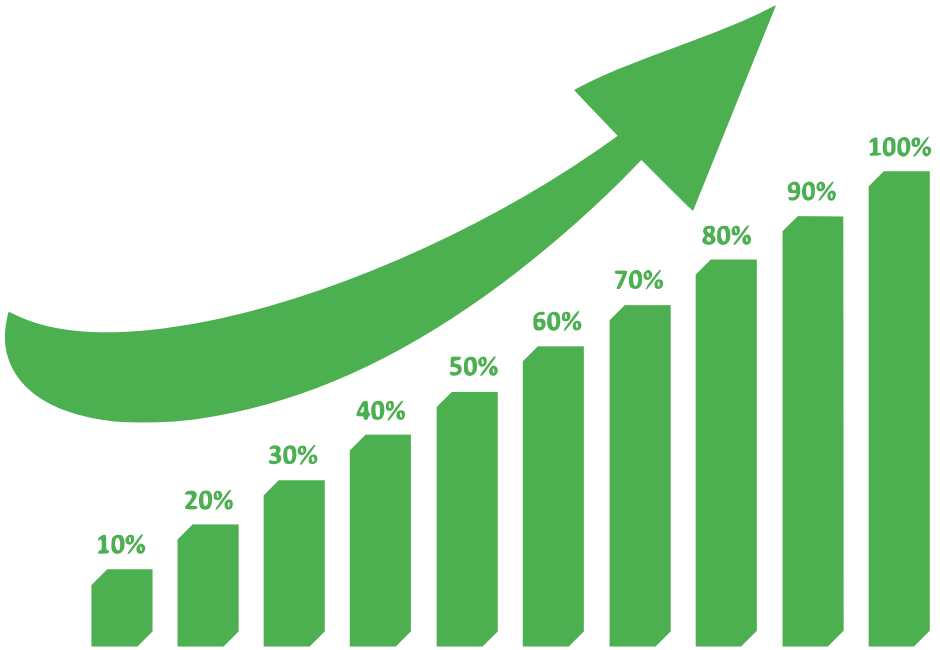
Lead-Scoring Case Study





Agenda

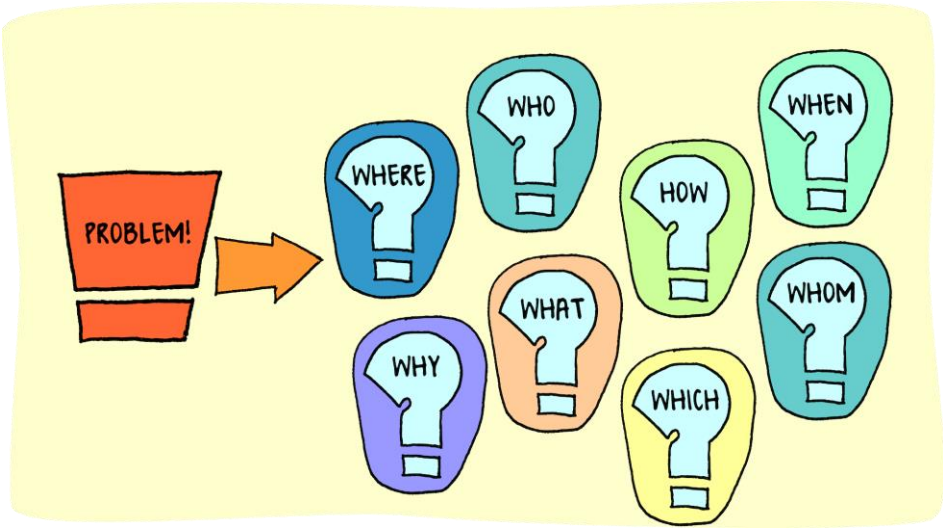
- Business Problem
- Business Objective
- Technical Approach
- Conclusion
- Final takeaways

Business Problem

An education company named 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 about 30 of them are converted.

The objective is to build a model to identify the hot leads and achieve lead conversion rate to 80%.



Business Objective

The Business Objective Is To Build A Logistic Regression Model To Identify The Hot/Potential Leads And Achieve The Lead Conversion Rate To 80%.

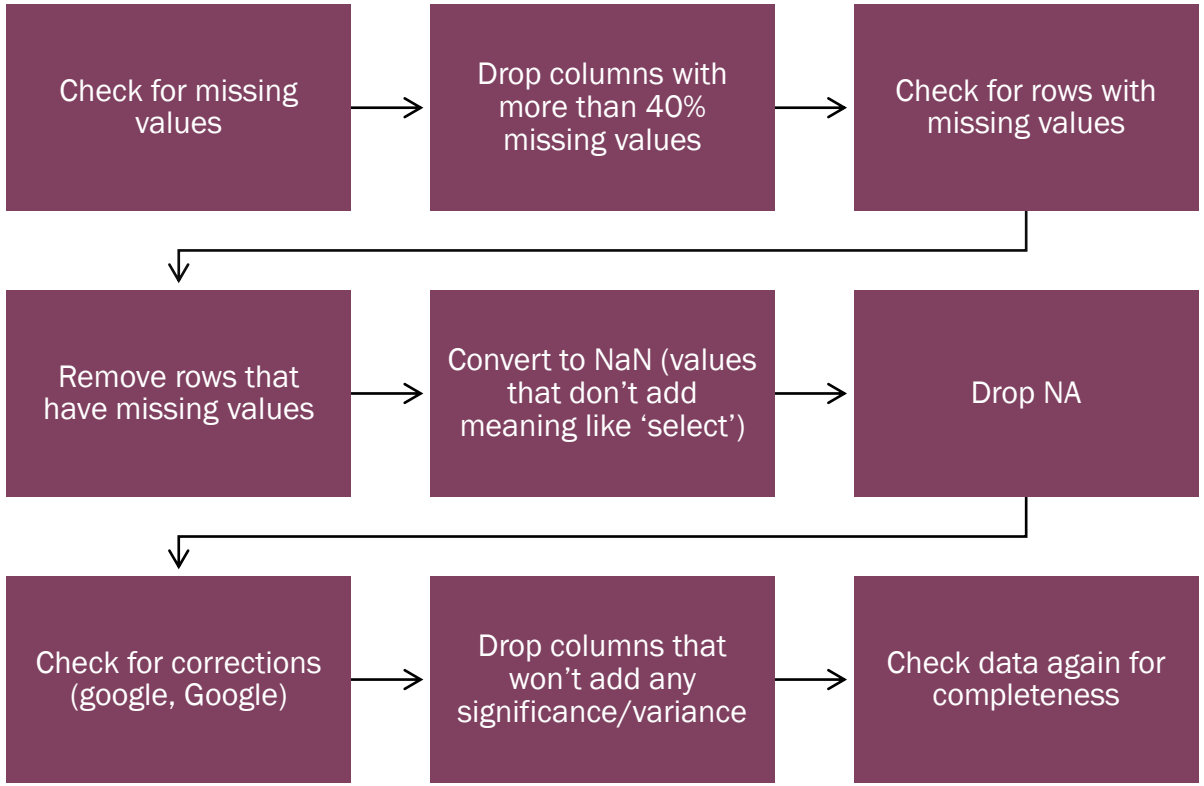
Technical Approach

- Dataset
- Data Cleaning
- EDA
- Model building
- Performance

Dataset

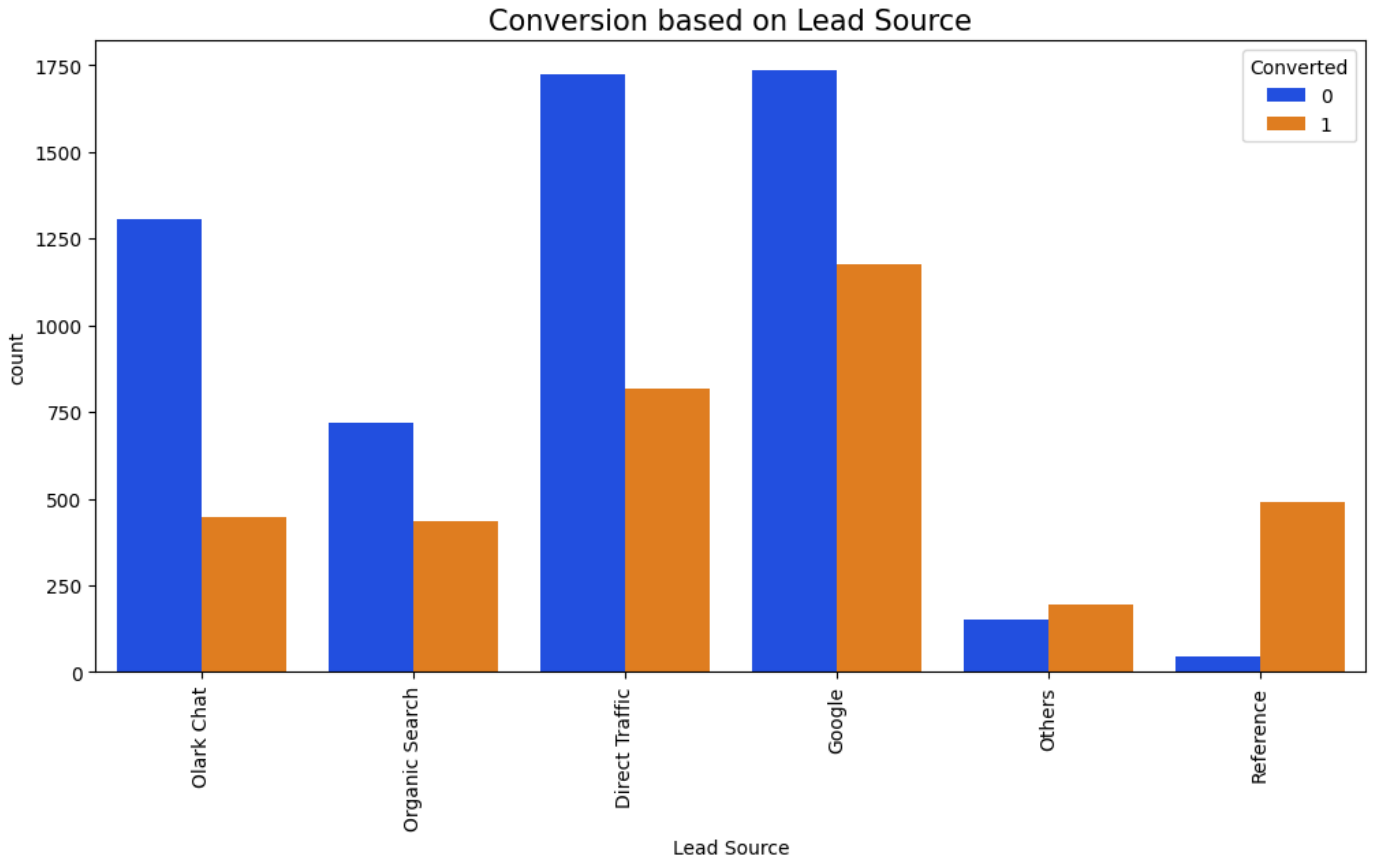
- Dataset used : “Leads.csv “
- Total number of customers present : 9240
- Total number of features : 37
- Model used : Logistic Regression
- After initial analysis, we see that there are multiple factors that influence conversion rate.
- The target column in our dataset : “Converted” .
- We need to reduce the features to maximize the conversion rate.
- Current Conversion Rate = 38.53%

Data Cleaning



Exploratory data analysis

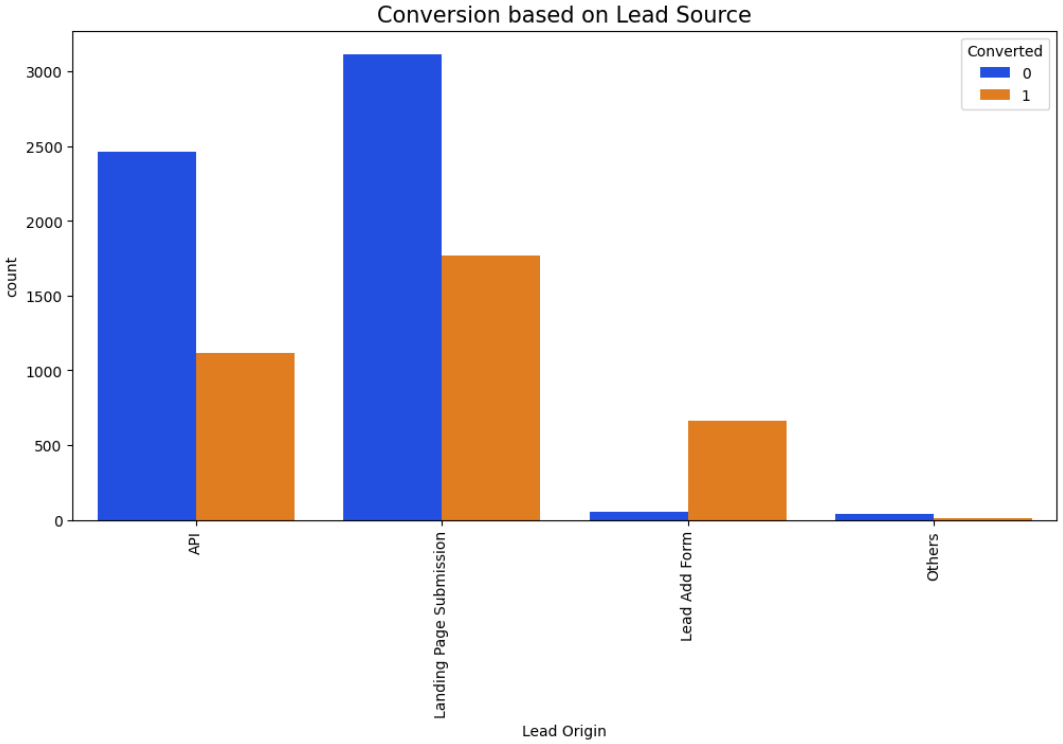
Categorical Data



- Google is found to be the important source for Lead Conversion.

- Direct Traffic also proves to be important to secure leads.

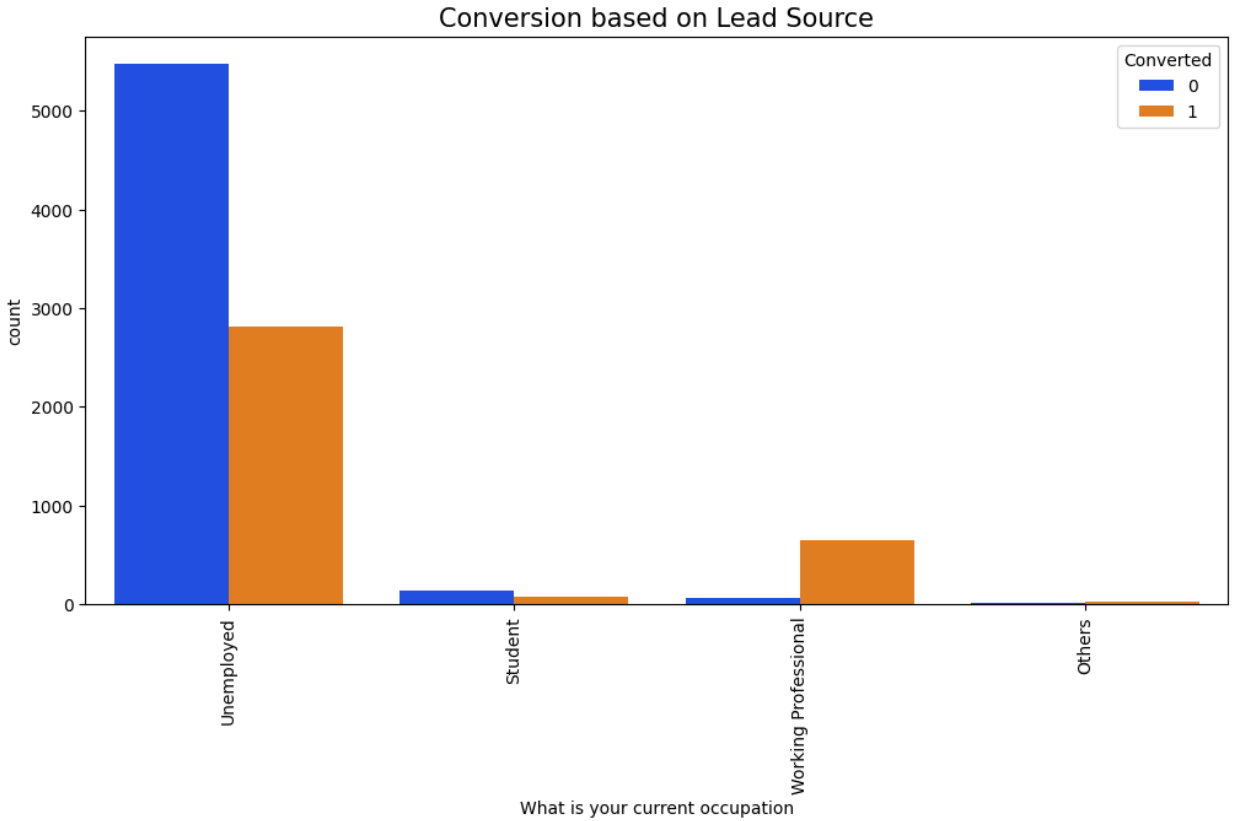
Categorical Data



- The percentage of Converted people is found to be greater for Landing Page Submission.

- We can also see that if Lead source is Add Form, the ratio of lead conversion is very high (almost not converted is very less).

Categorical Data

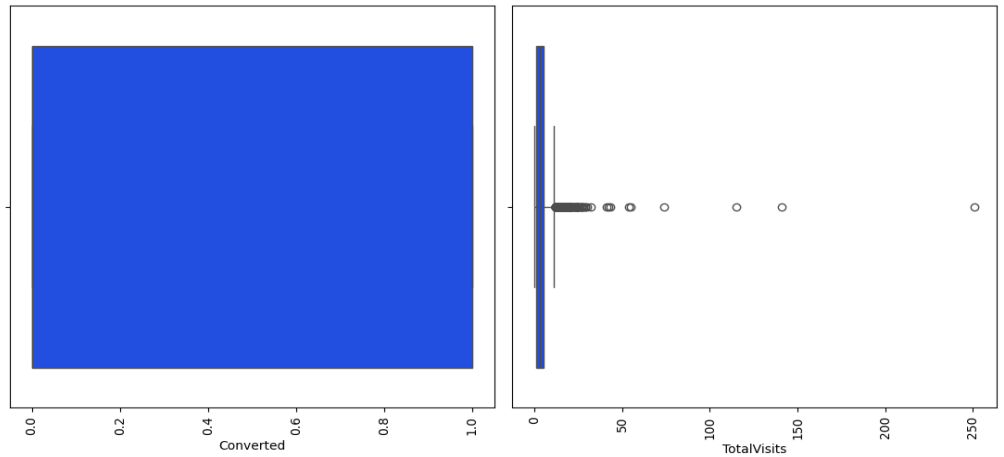


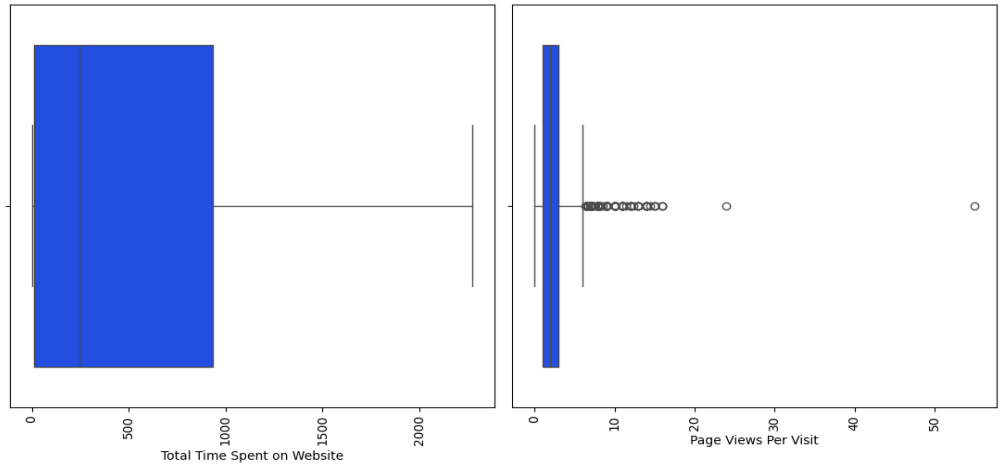
- The Unemployed are interested in doing the online course and improve their skill .

- We can also see working professionals are potential leads, who like to upskill themselves with latest trend

Handling Outliers

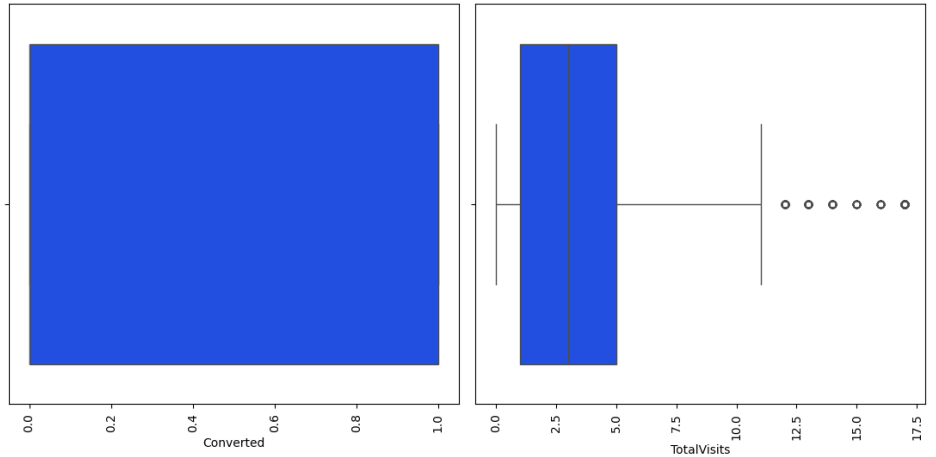
Before

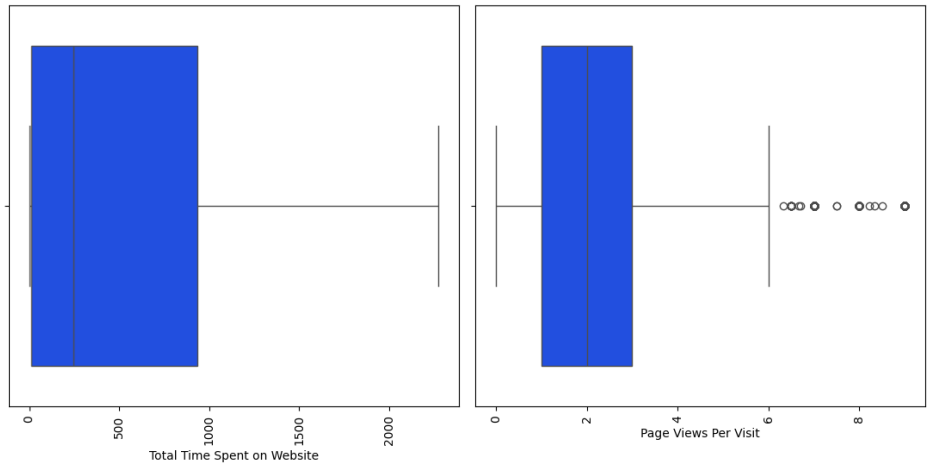




Quantile

After





Model Building

Dummy Variable Creation: We need to create dummy variables for all the categorical columns as they enable us to use a regression equation on multiple groups.
Test Train Split: Division of data into test data and train data to check the stability of the model.We have randomly sampled 70% of the data as the test data and 30% of the data as test data.
Scaling: Division of Train Data into X and Y where X has all the features and Y has the target variable – Converted.We perform scaling to normalize the data within a particular range

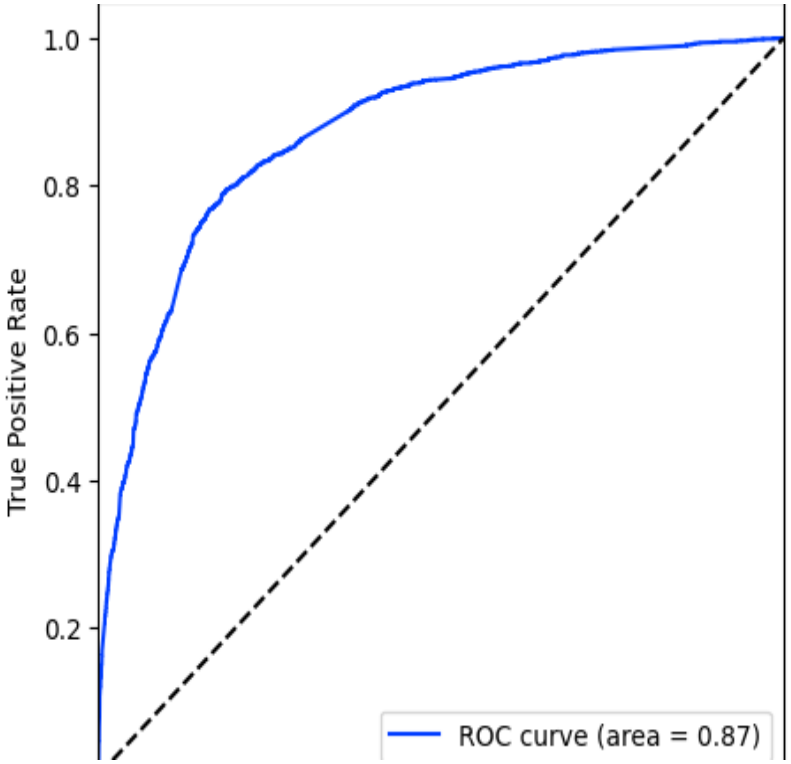
Model – I and II: Basic Model - We build a basic model using 35 features. Since it is not efficient we perform RFE to obtain a model with Top – 20 features. There are so many variables with high VIF value, we need to remove them

Model - III : Removing variables having high VIF

Model – VIII : The Final Model
All p-values < 5% All VIF values are < 5. Hence the dependency of variable with another is tolerable.

- Final model has 18 features in total.

ROC Curve And Optical Cut-Off Probability



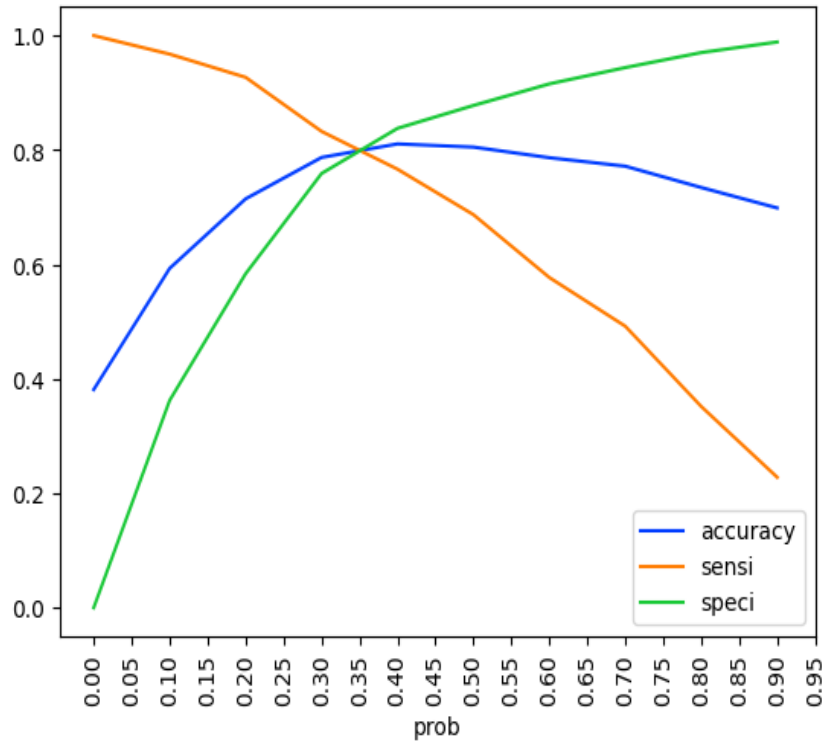
The ROC curve illustrates the model's ability to differentiate between the classes.

On plotting the ROC curve for our data we see that, AUC is around 0.88 which means at around 88% of the times, the model is able to distinguish the 1's as 1's and 0's as 0's.

AUC of 0.88 is found to be very stable model.

When we plot the sensitivity, accuracy and specificity of the model together, the optimal cut off point is found to be at 0.35.

With probability = 0.35 , we predict y-values with XTrain, in such a way that, any conversion prob > 35% is said to be converted to a lead



Model Performance Test

Train Set
ACCURACY - 80%
SENSITIVITY - 71%
SPECIFICITY - 81%

Vs

Test Set
ACCURACY – 79.9%
SENSITIVITY - 71%
SPECIFICITY - 80%

Hot Leads

- Hot leads are people who have a high probability to be converted as a Lead and thus needs to be identified. They have a higher conversion rate.
- The leads whose lead score is greater than 35% are considered as potential leads. The conversion rate is around 73%. When we increase this threshold from 35% to 90% we get Hot Leads.
- Conversion Rate for hot leads is increases from 73% to 92%. This means they have a 92% probability of getting converted to a lead.
- Focusing on Hot Leads will increase the chances of obtaining more value to the business as the number of people we contact are less but the conversion rate is high.

Hot Leads

- Hot leads are people who have a high probability to be converted as a Lead and thus needs to be identified.
- The leads whose lead score is greater than 35% are considered as potential leads. The conversion rate is around 72%. When we increase this threshold from 35% to 90% we get Hot Leads.
- Conversion Rate for hot leads is increases from 72% to 92%. This means they have a 92% probability of getting converted to a lead.
- Focusing on Hot Leads will increase the chances of obtaining more value to the business as the number of people we contact are less but the conversion rate is high.

Conclusion

- The customer/leads who fills the form are the potential leads.
- We must majorly focus on working professionals.
- We must majorly focus on leads whose last activity is SMS sent or Email opened.
- It’s always good to focus on customers, who have spent significant time on our website.
- It’s better to focus least on customers to whom the sent mail is bounced back. If the lead source is referral, he/she may not be the potential lead.
- If the lead didn’t fill specialization, he/she may not know what to study and are not right people to target. So, it’s better to focus less on such cases.

Thank you