LEAD SCORE CASE STUDY

Group Members
- PRITHVI RAJ CHAUHAN
- SARTHAK GUPTA
- JEEVANJYOT SINGH

PROBLEM STATEMENT

- ✓ X Education sells online courses to industry professionals.
- ✓ 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.
- ✓ To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- ✓ If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective:

- ✓ X education wants to know most promising leads.
- ✓ For that they want to build a Model which identifies the hot leads.
- ✓ Deployment of the model for the future use.

SOLUTION METHODOLOGY

- ✓ Data cleaning and data manipulation.
  1. Check and handle duplicate data.
  2. Check and handle NA values and missing values.
  3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  4. Imputation of the values, if necessary.
  5. Check and handle outliers in data.

- ✓ EDA
  - Bivariate data analysis: correlation coefficients and pattern between the variables etc.
  - Feature Scaling & Dummy Variables and encoding of the data.

- ✓ Classification technique: logistic regression used for the model making and prediction.

- ✓ Validation of the model.

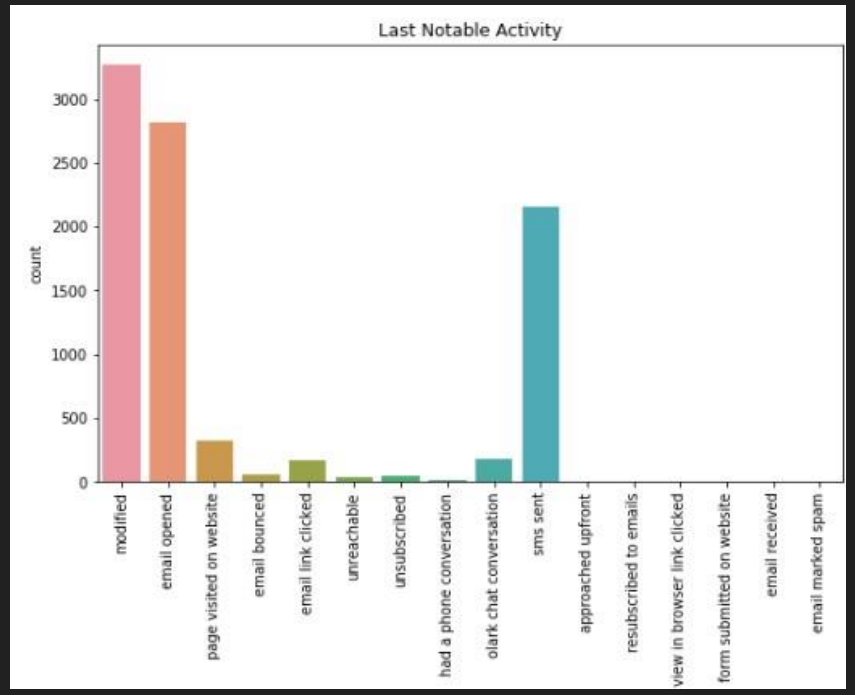
- ✓ Model presentation.

- ✓ Conclusions and recommendations.

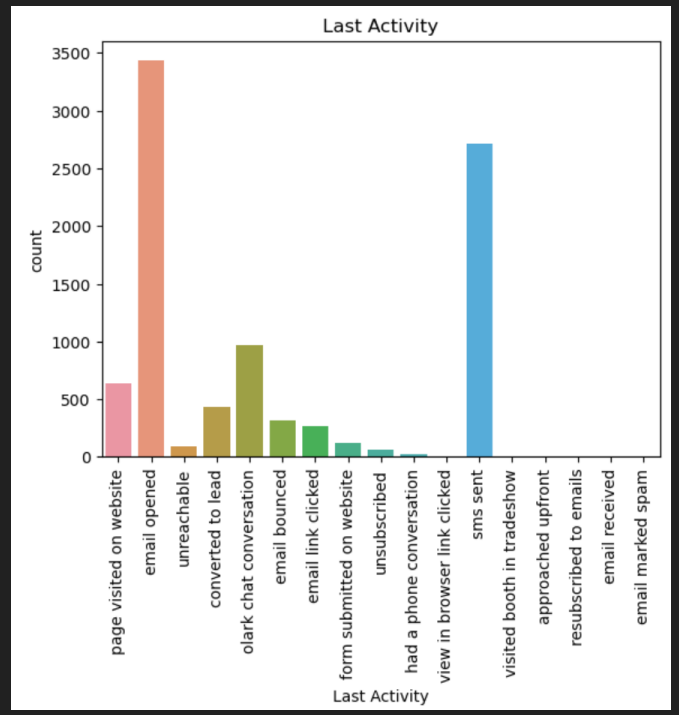
DATA MANIPULATION

- ✓ Total Number of Rows = 37, Total Number of Columns = 9240.
- ✓ Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”
- ✓ Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- ✓ Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- ✓ After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- ✓ Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

EDA

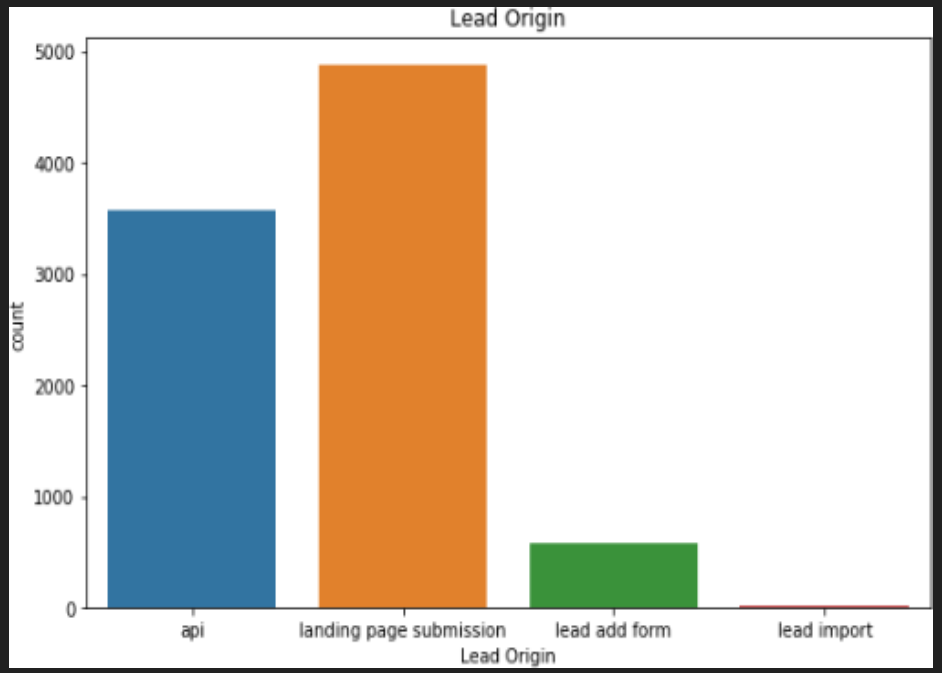


Opening & Modifying the email is the most common last notable activity

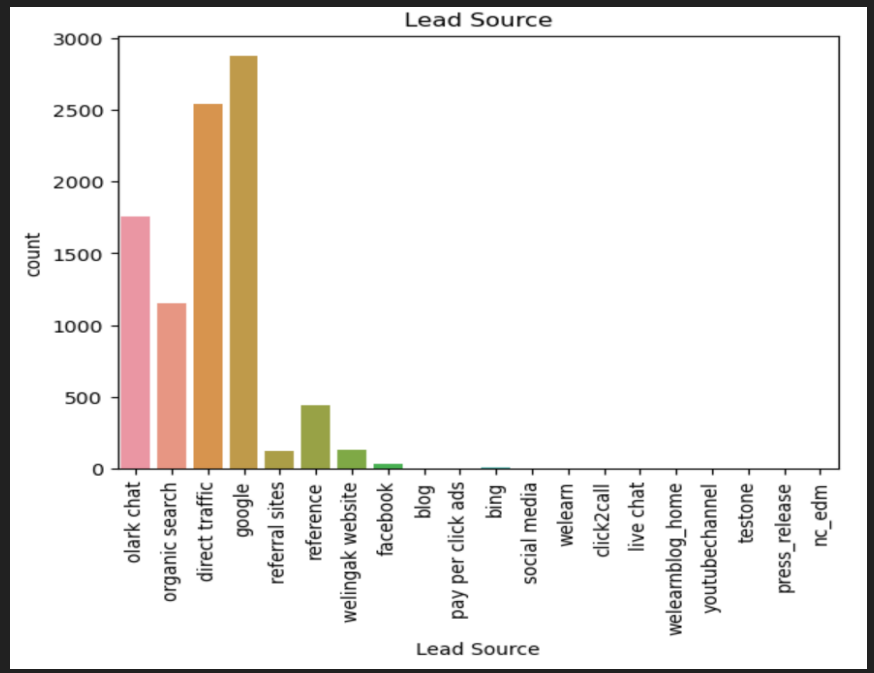


Opening the email is the most common last activity

EDA

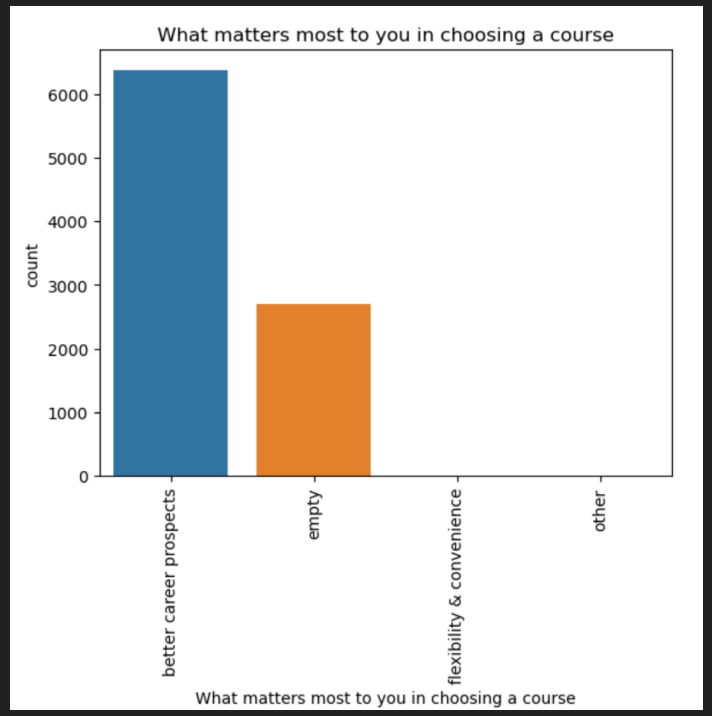


The landing page submission have the most count amongst the lead origin

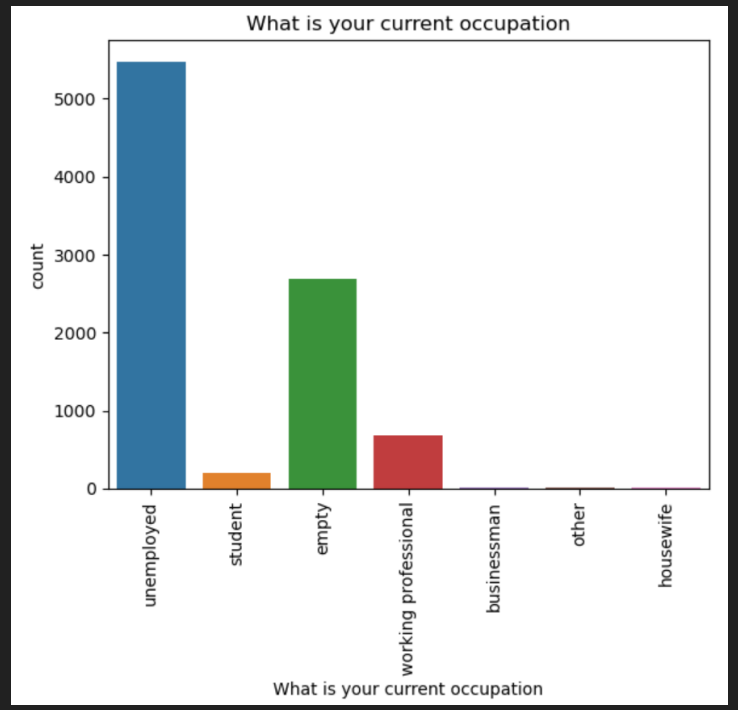


Google is the main lead source

EDA

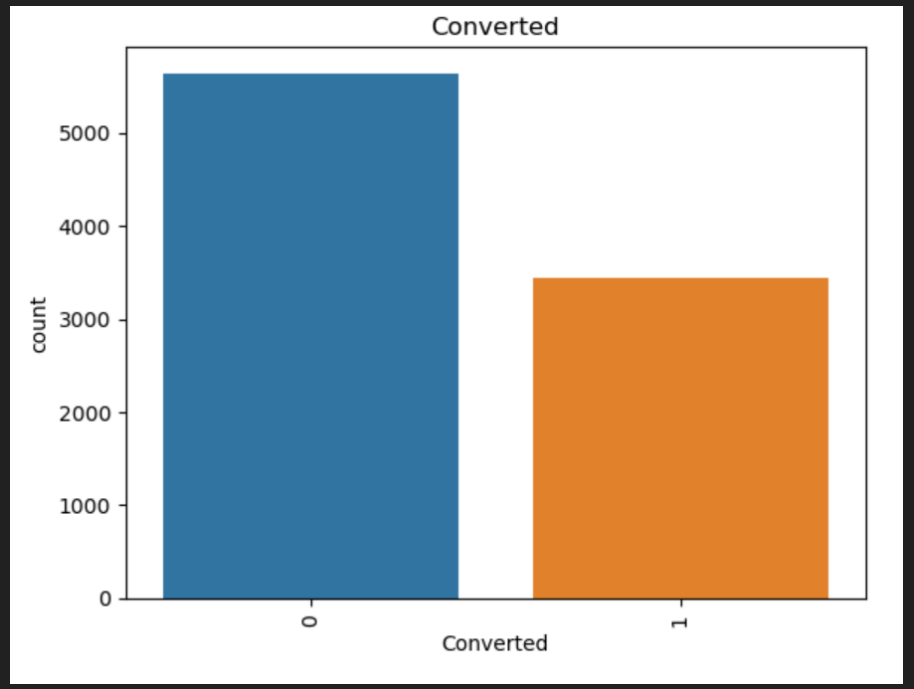


People chose the course for a better future prospect the most



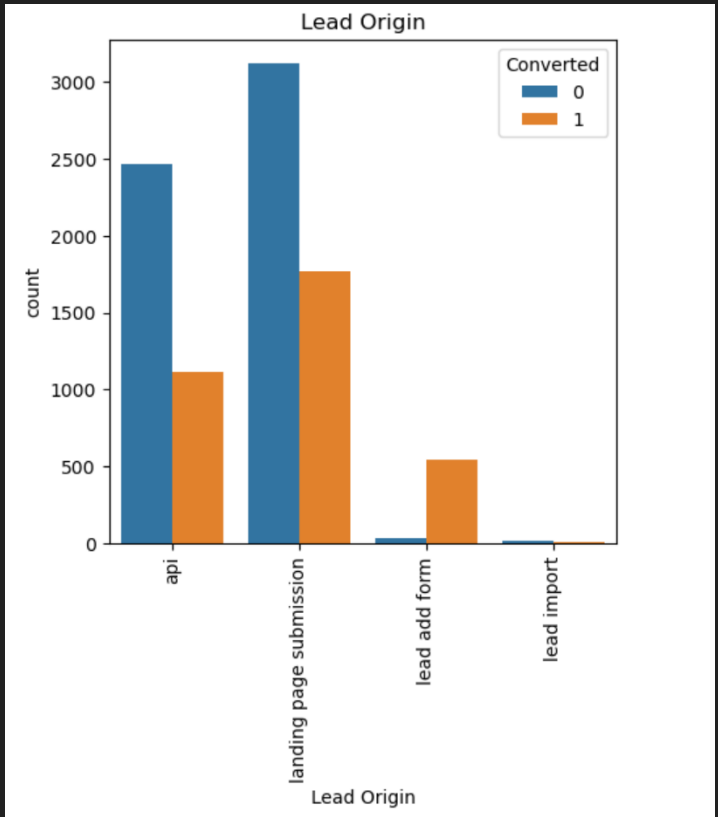
Most leads are unemployed

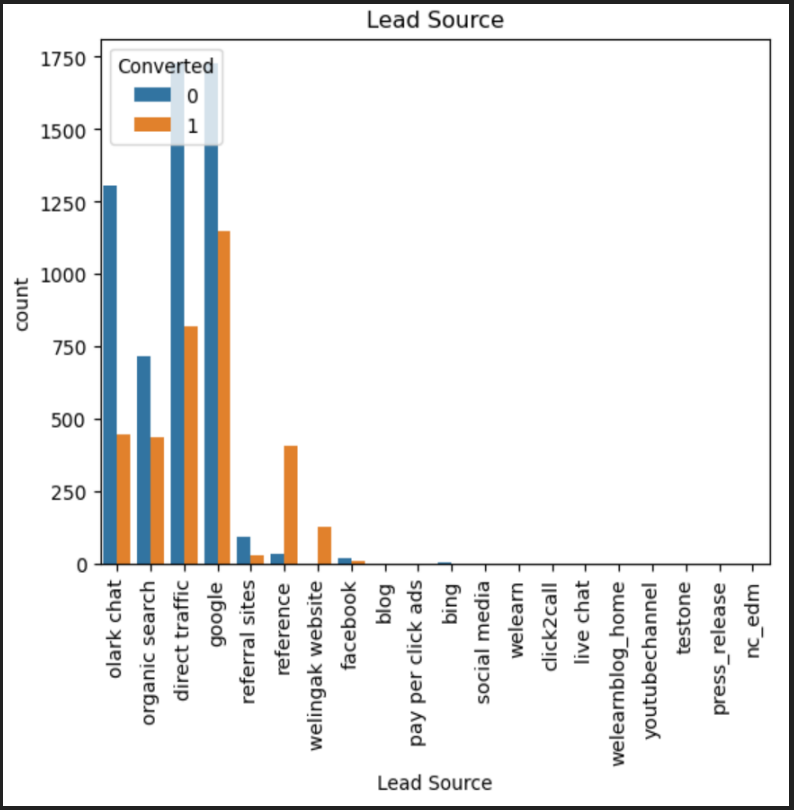
TARGET VARIBLE ANALYSIS



Most leads were not converted.so a little imbalance in data exists

BI VARIATE ANALYSIS

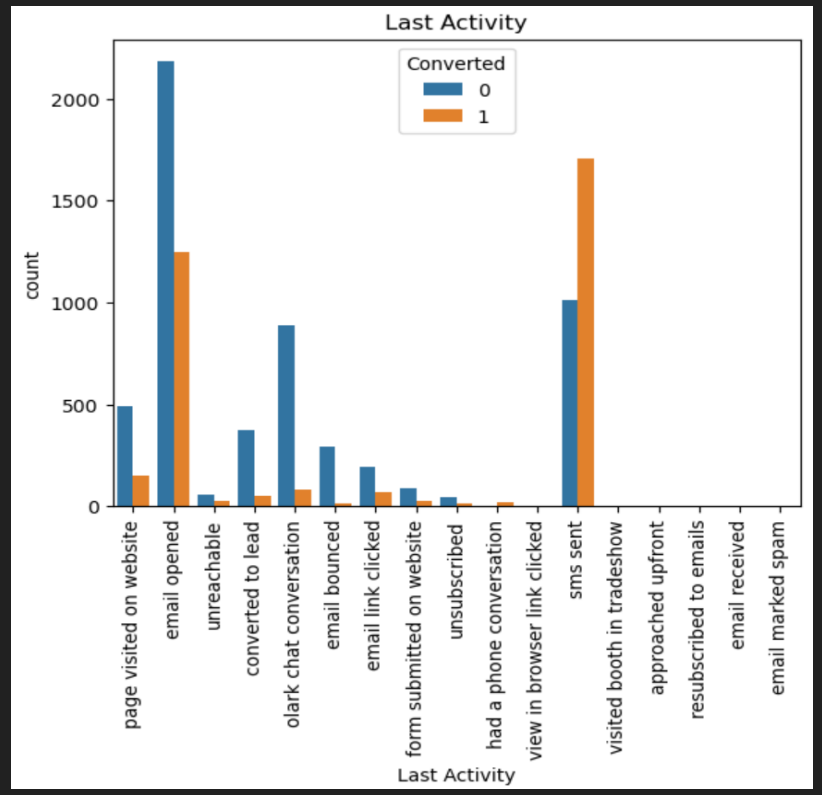


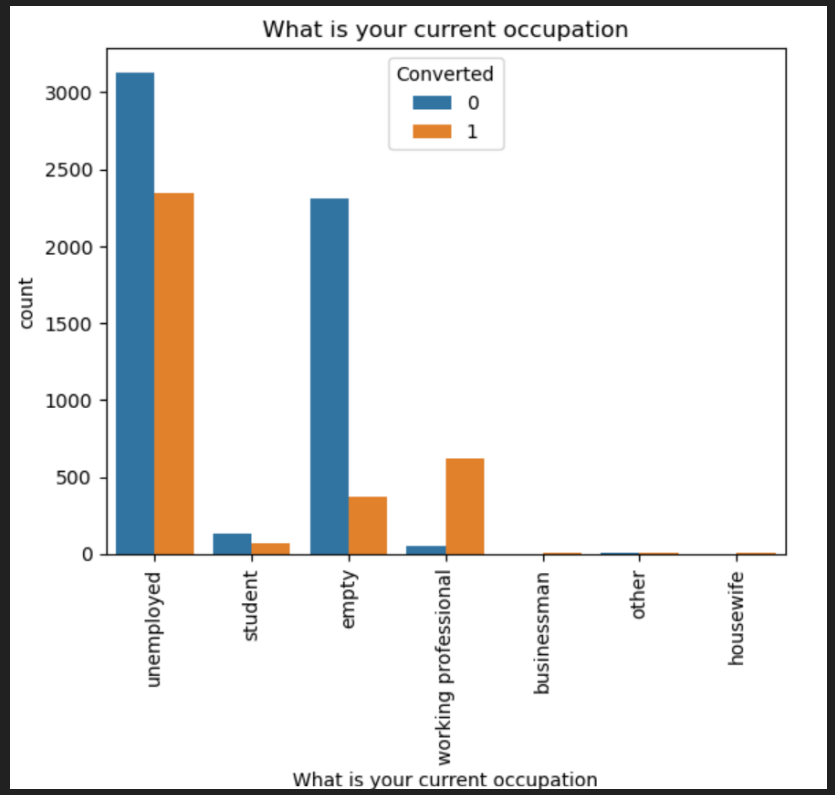


- We see leads with lead add form as lead origin were converted way more than others.

- Alse leads with reference as lead source were mostly converted.

BI VARIATE ANALYSIS

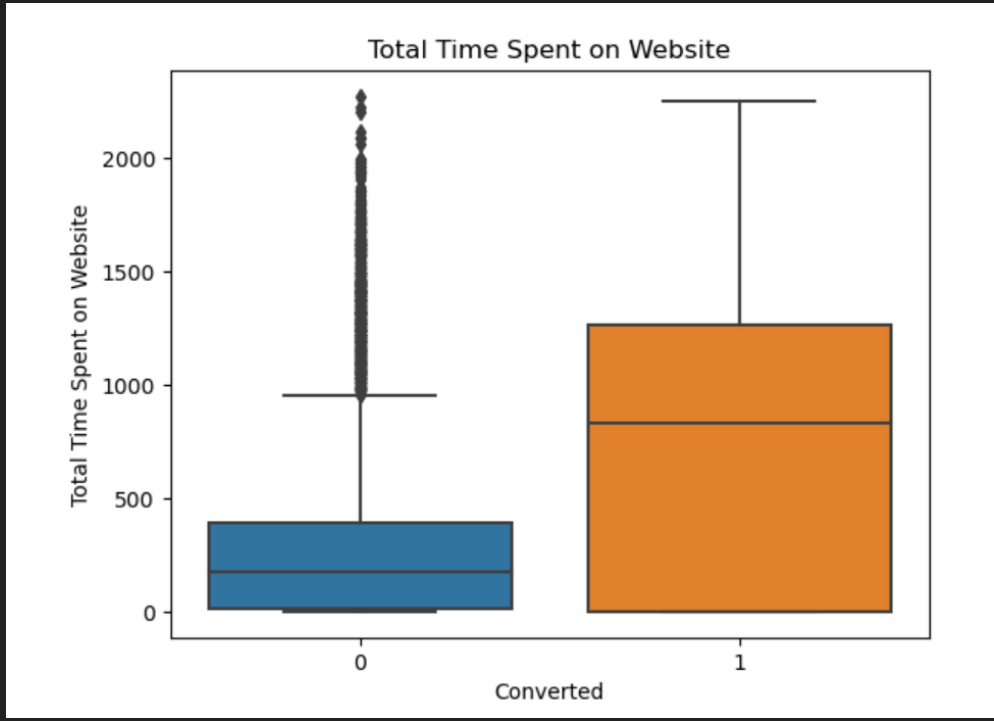




- Leads with last activity of sms sent were converted more.

- Working professionals were the most converted leads amongst occupations

BI VARIATE ANALYSIS



Leads which convert spent more time on the website

MODEL BUILDING

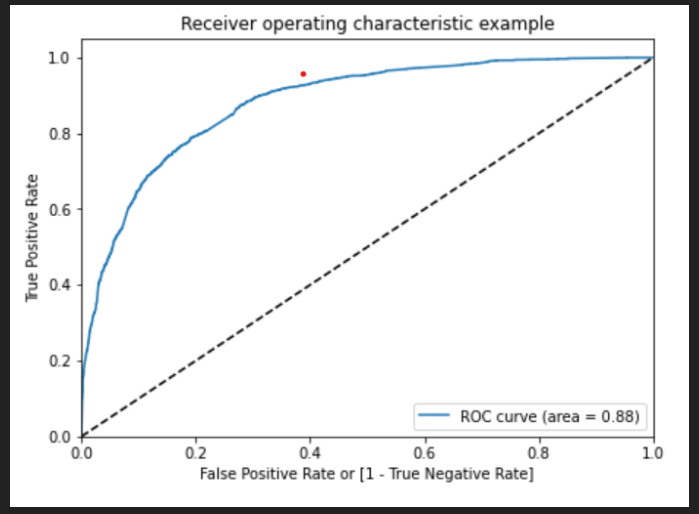
- ✓ Splitting the Data into Training and Testing Sets
- ✓ The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- ✓ Use RFE for Feature Selection
- ✓ Running RFE with 15 variables as output
- ✓ Building Model by removing the variable whose p- value is greater than 0.05 and VIF value is greater than 5.
- ✓ Predictions on test data set
- ✓ Overall accuracy 81% on test data set.

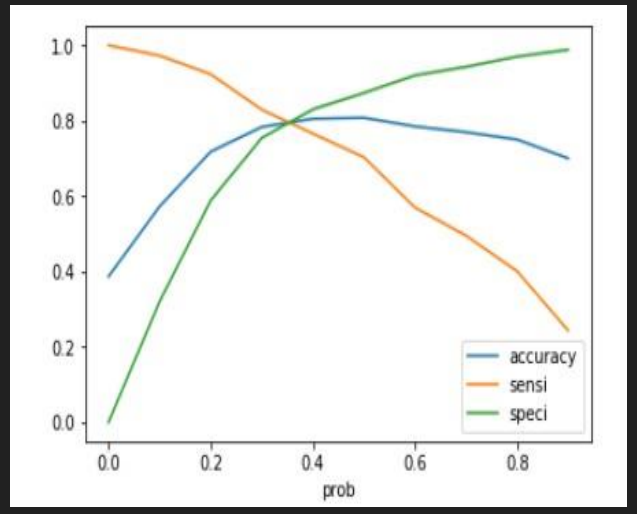
MODEL BUILDING

- ✓ we can increase the recall for the model so as to catch more clients who will convert
- ✓ we have a precision and recall od 0.74 and 0.75 for a cut off of 0.4
- ✓ For a cut off of 0.4 we have an accuracy of 0.819 and sensitivity and specificity of 0.762 and 0.851 respectively on test set

ROC CURVE

- ✓ Finding Optimal Cut off Point
- ✓ Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
- ✓ From the second graph it is visible that the optimal cut off is near 0.4.





CONCLUSION

It was found that variables that mattered the most in the potential buyers are:

1. Total time spends on the Website.
2. Total number of visits.
3. When the lead source was
    a. Welingak website
4. When the last activity was:
    a. SMS
    b. Olark chat conversation
5. When their current occupation is as a working professional.
6. When the lead origin is Lead add format.
7. When the last notable activity was:
    a. Had a phone conversation
    b. Unreachable

So overall we built a logistic regression model with a good accuracy of 81.9% with a cut off of 0.4.
From EDA we found most leads were from india and they were unemployed.
Working professionals were the most converted leads and referred leads had good chances of conversion.
People with sms as the last activity and spend more time on website too had good chances of conversion.
Welingak website provides very good leads and must be worked upon.
Most people opt for the course for better future prospects.
Overall the company must target referred working professionals from india who will take the course for better future prospect

THANK YOU