LEAD SCORE CASE STUDY

ANAS TOPIA

ANITHA

AMRUTHA

STRATERGY FOLLOWED

- Understanding the data
- Clean the data
- EDA
- Prepare the data for Model Building(Binary mapping and including dummy for category)
- Model Building
- Model Evaluation
- Making Predictions on the Test Set

- UNDERSTANDING THE DATA:
  - IMPORT DATA AND ANALYSE ITS SHAPE , SIZE , NUMBER OF OBJECTS AND NUMERIC COLUMNS AND ALSO BUSINESS OBJECTVE OF EACH COLUMNS
- DATA CLEANING:

REMOVING COLUMNS WITH NULL VALUE> 30%

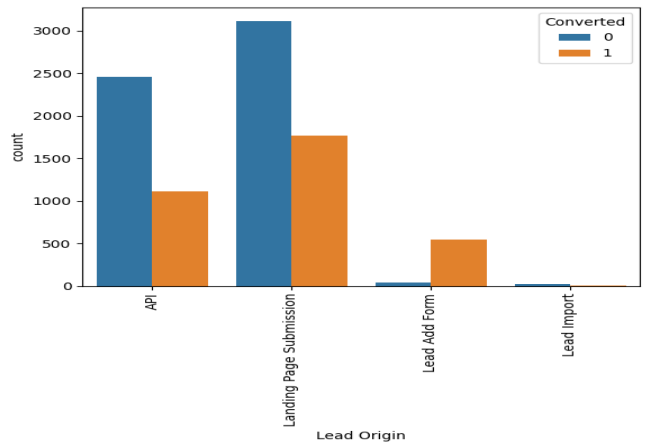
REMOVING COLUMNS WHICH DOESNOT SHOW ANY VARIATIONS

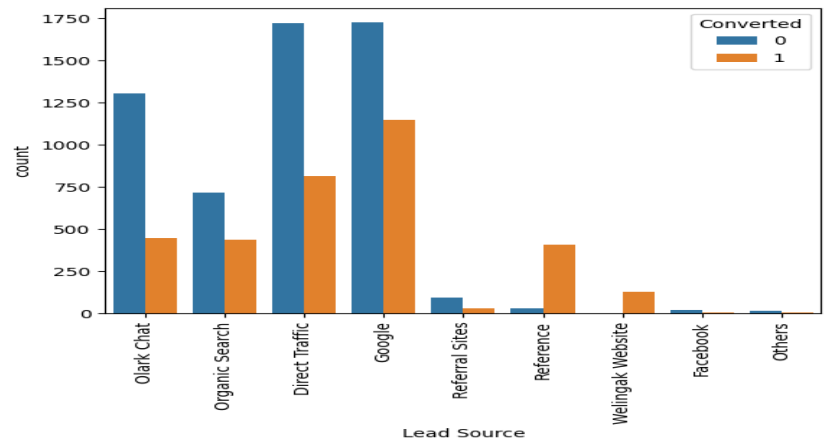
CITY AND COUNTRY ARE REMOVED

OUT OF ALL COLUMNS, VARIABLES Lead Profile, How did you hear about X Education AND Specialization HAVE SELECT VALUES. DROPPING OTHER 2 COLUMNS BASED ON ITS VALUE COUNTS SHOWING VERY HIGH SELECT VALUE COMPARED TO OTHERS

CAP THE OUTLIERS TO 95% IN PAGE VIEWS PER VISIT

- EDA





Univariate and bivariate analysis shows below inference:

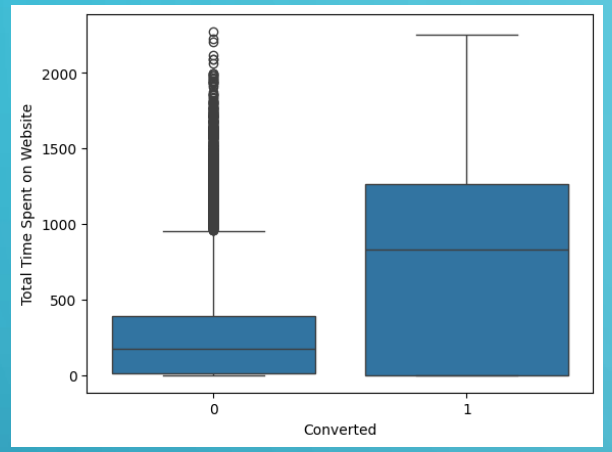
Lead origin:

1. Landing Page Submission have higher count of lead originated
2. Lead Add Form has high conversion rate but count of lead are less

Lead source:

1. Google have maximum number of leads.
2. reference leads and welingak website , conversion is highh

- EDA



Time spent:

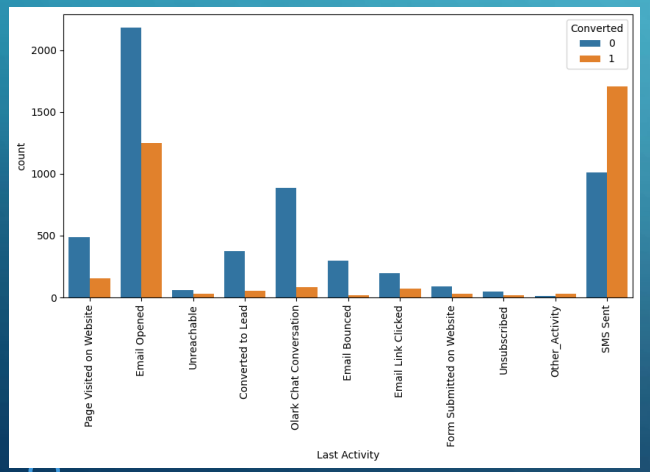
1. Leads spending more time on the weblise are more likely to be converted.

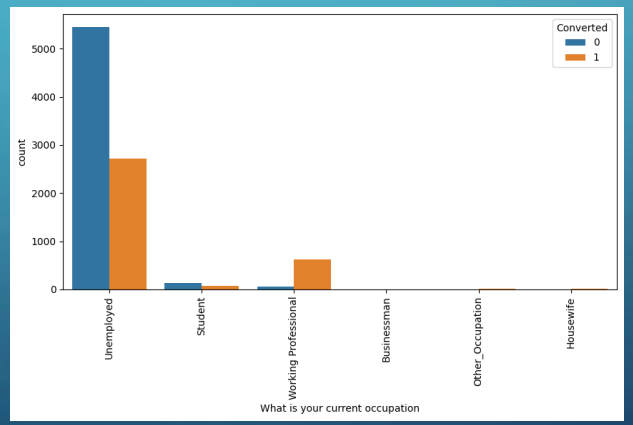
Last Activity:

1. Email opened has highest lead
2. Conversion rate for leads SMS Sent is high

Occupation:

1. Working Professionals going for the course have high conversion
2. Unemployed leads are the high





PREPARING DATA FOR MODEL

Binary mapping of data and adding dummy columns for categorical columns

Scaling of numeric features for normalizing independent variables using fit_transform

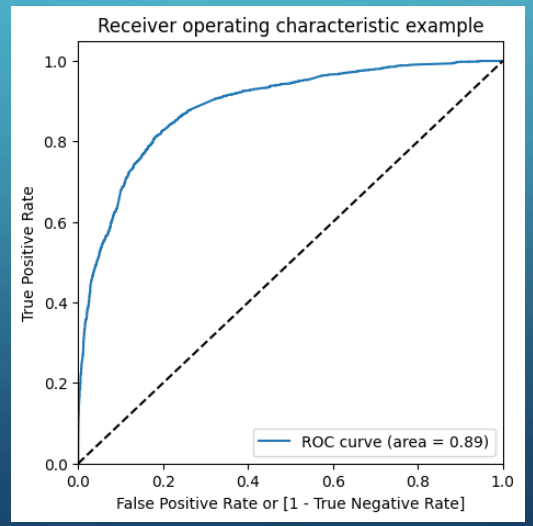
MODEL BUILDING

After splitting test train, using statsmodels glm top 15 RFE variables are identified and used for building model

dropping 'What is your current occupation_Housewife‘ as its P>|z| is 0.999 again buiding the model with dropped data

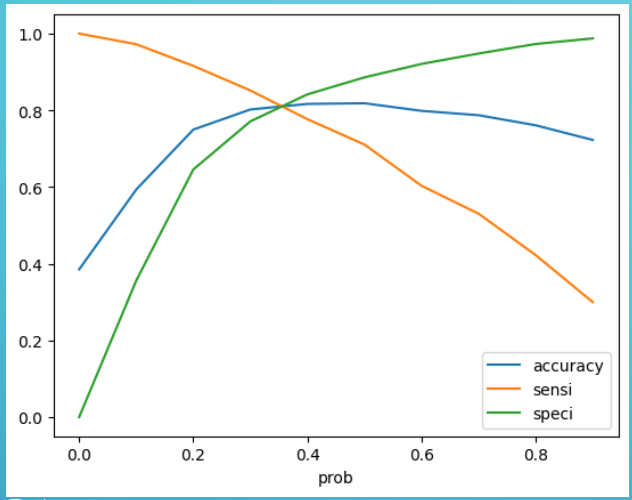
checking VIF for multicolinearity and found everything to be below <5

MODEL EVALUATION



MODEL EVALUATION

MEETING POINT IS 0.38



PREDICTION

FOUND THE MEETING POINT TO BE 0.38 AND REVISING THE PREDICTION

RUN THE MODEL FOR XTEST , FIND YTEST PRED(PROBABILITY OF CONVERSION)

ASSIGN THE SCORE FOR EACH PROBABILITY – LEAD SCORE

METRICS

- CONFUSION MATRIX
- array([[1096, 638],
- [97, 892]]
- Specificity –.78
- Sensitivity -.83
- ACCURACY -.81

INFERENCE & CONCLUSION

- To improve the lead conversion we should analys occupation, Total Time Spent on website and lead sources of the lead and proceed with them
- People who are unemployed and having specialization ‘Other’ can be ignored for this process. Also Person with Email bounced and status of Do Not Email set also will not convert to Hot leads. So based on the correlation we can conclude this
- Univariate and bivariate analysis shows below inference: Lead origin:
- Landing Page Submission have higher count of lead originated
- Lead Add Form has high conversion rate but count of lead are less Lead source:
- Google have maximum number of leads.
- reference leads and welingak website , conversion is high Time spent:
- Leads spending more time on the weblise are more likely to be converted. Last Activity:
- Email opened has highest lead
- Conversion rate for leads SMS Sent is high Occupation:
- Working Professionals going for the course have high conversion
- Unemployed leads are the high
- Intern should work on conversion of lead variable having high lead numbers He should also work on increasing the lead numbers of leads with highest conversion rate