X EDUCATION

LEAD SCORE CASE STUDY

By:
Simranjeet Singh
&
Kartik Mehra

PROBLEM STATEMNET

- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you 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%

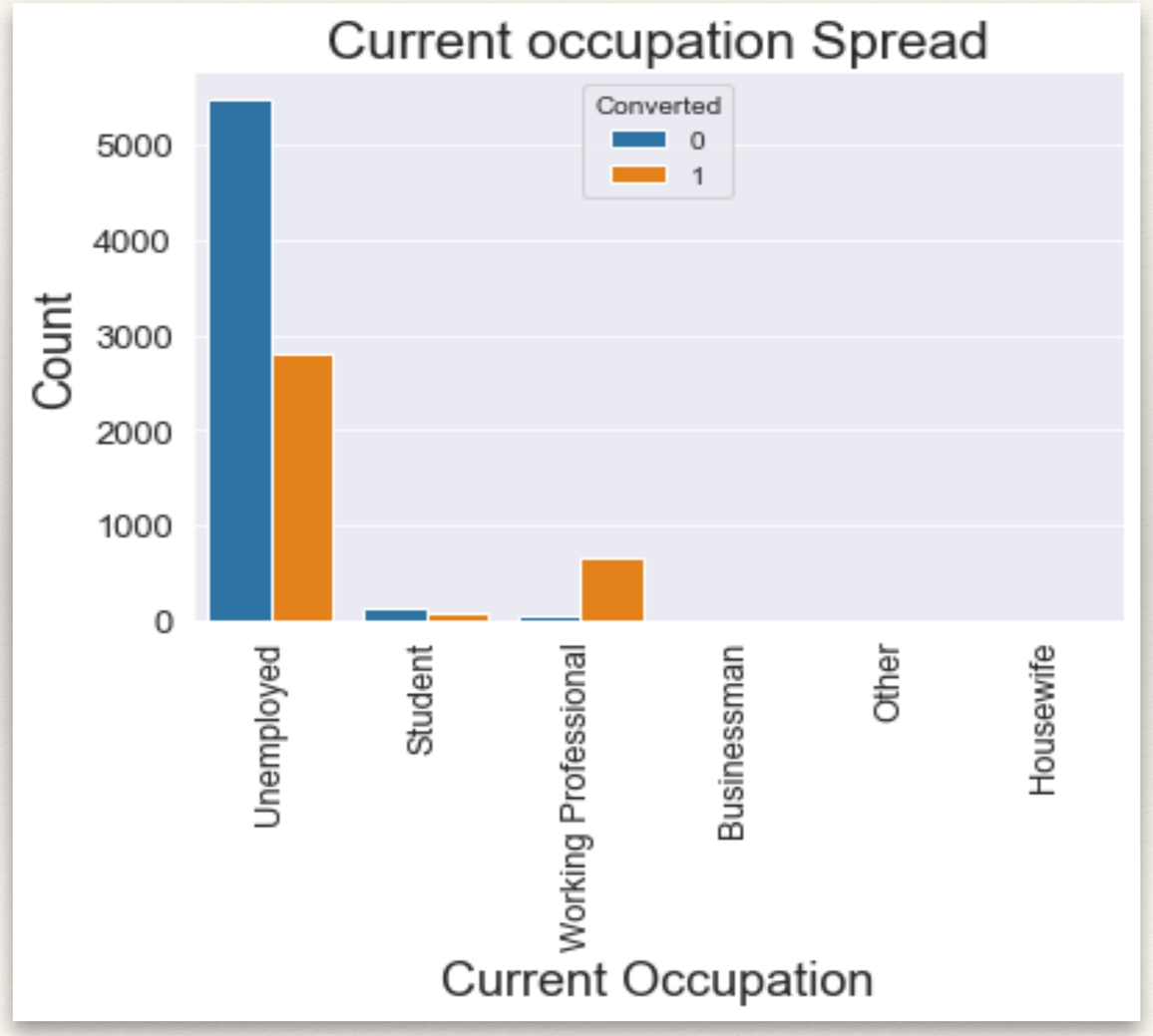
INTRODUCTION OF THE DATA SET

- Data set contains 9240 records.
- Prospect ID and Lead Number are unique columns.
- 7 features have more than 40% missing value. So, will remove those columns as it can mislead our model.
- 14 features have same value for 95% or more records. So, we will drop these features as well.

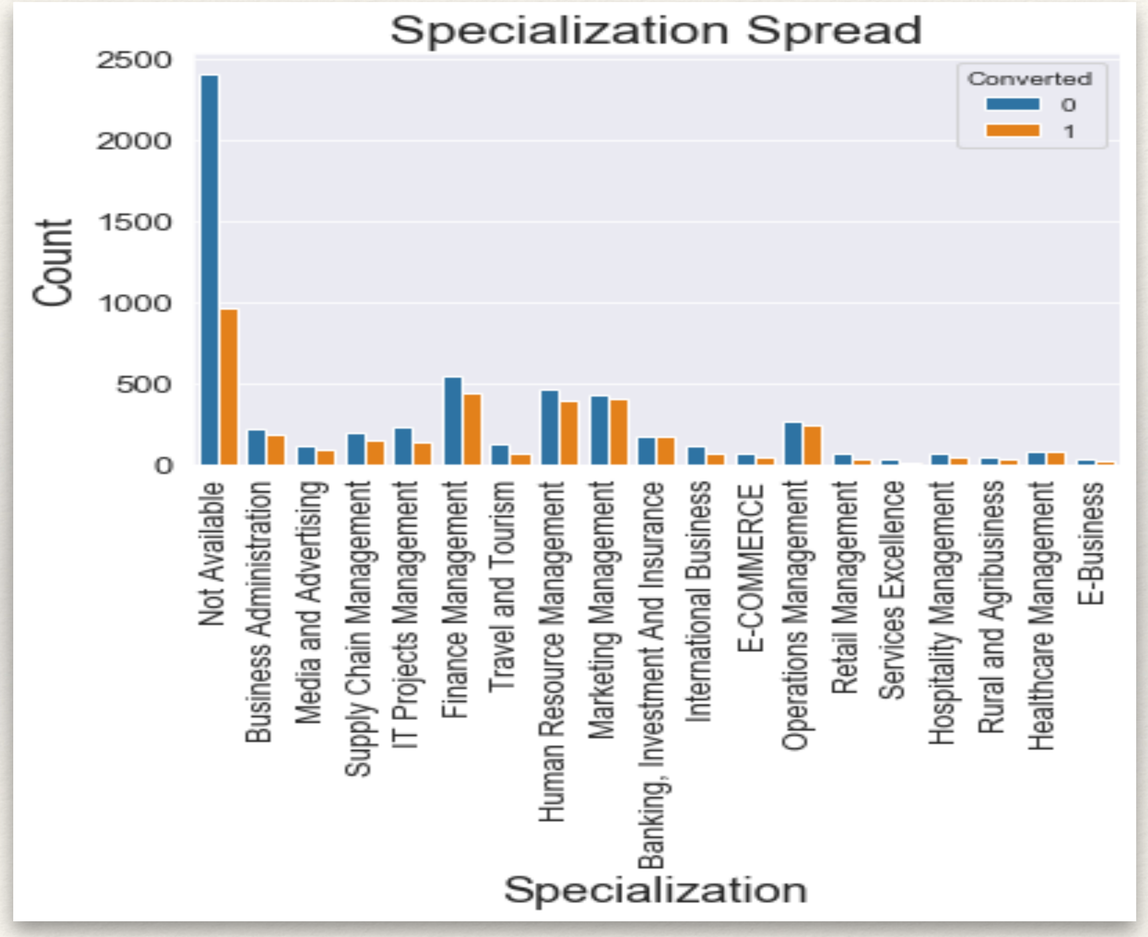
EXPLORATORY DATA ANALYSIS (EDA)

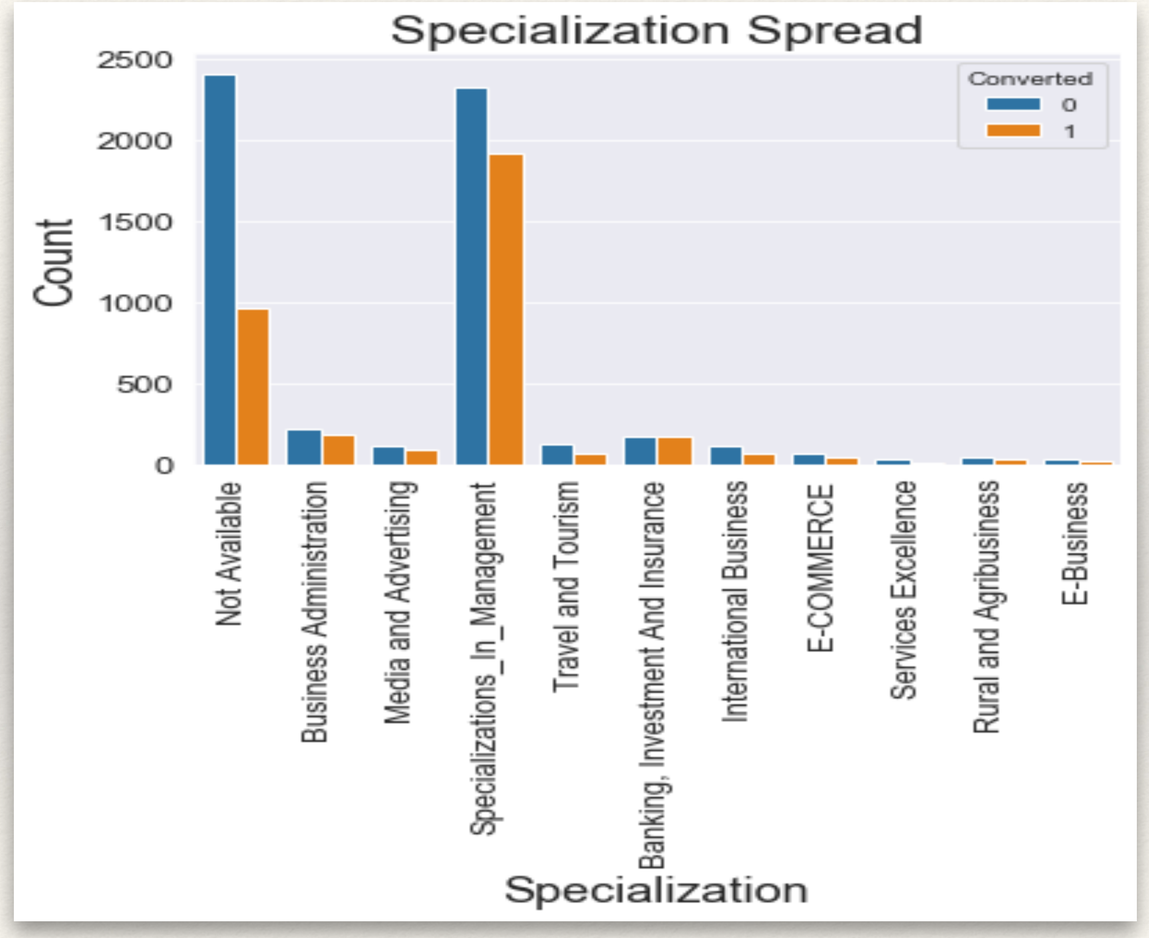
FEATURE: What is your current occupation

- Working Professionals have high chance of conversion.
- Unemployed sector has more initial leads but has less conversation rate i.e., approx. 50%



FEATURE: Specialization

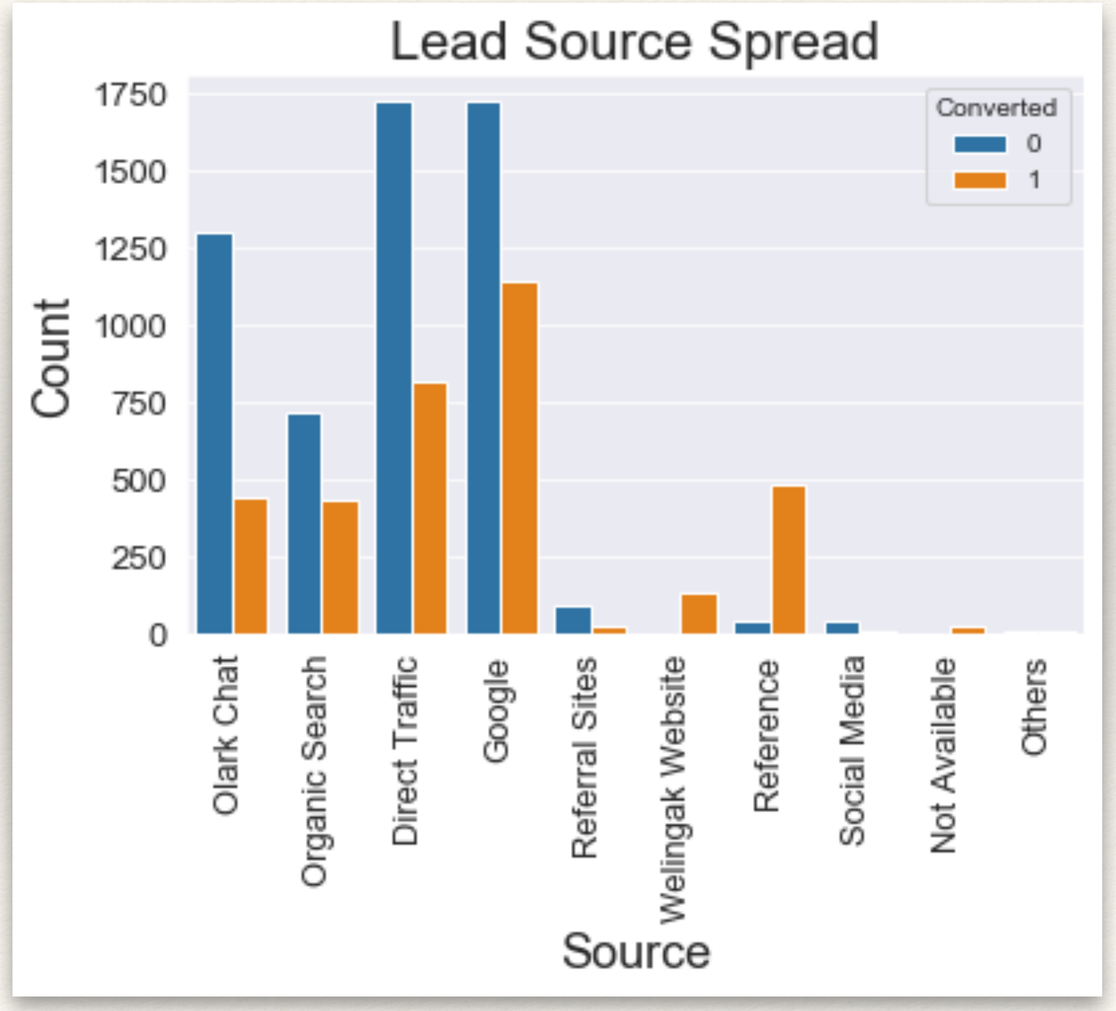




- Here we have 36% of missing data but not any value is highly frequent so we will create another value for this N/A
- We will merge the values which have term Management because they have higher number of lead conversion as compared to other specialisation

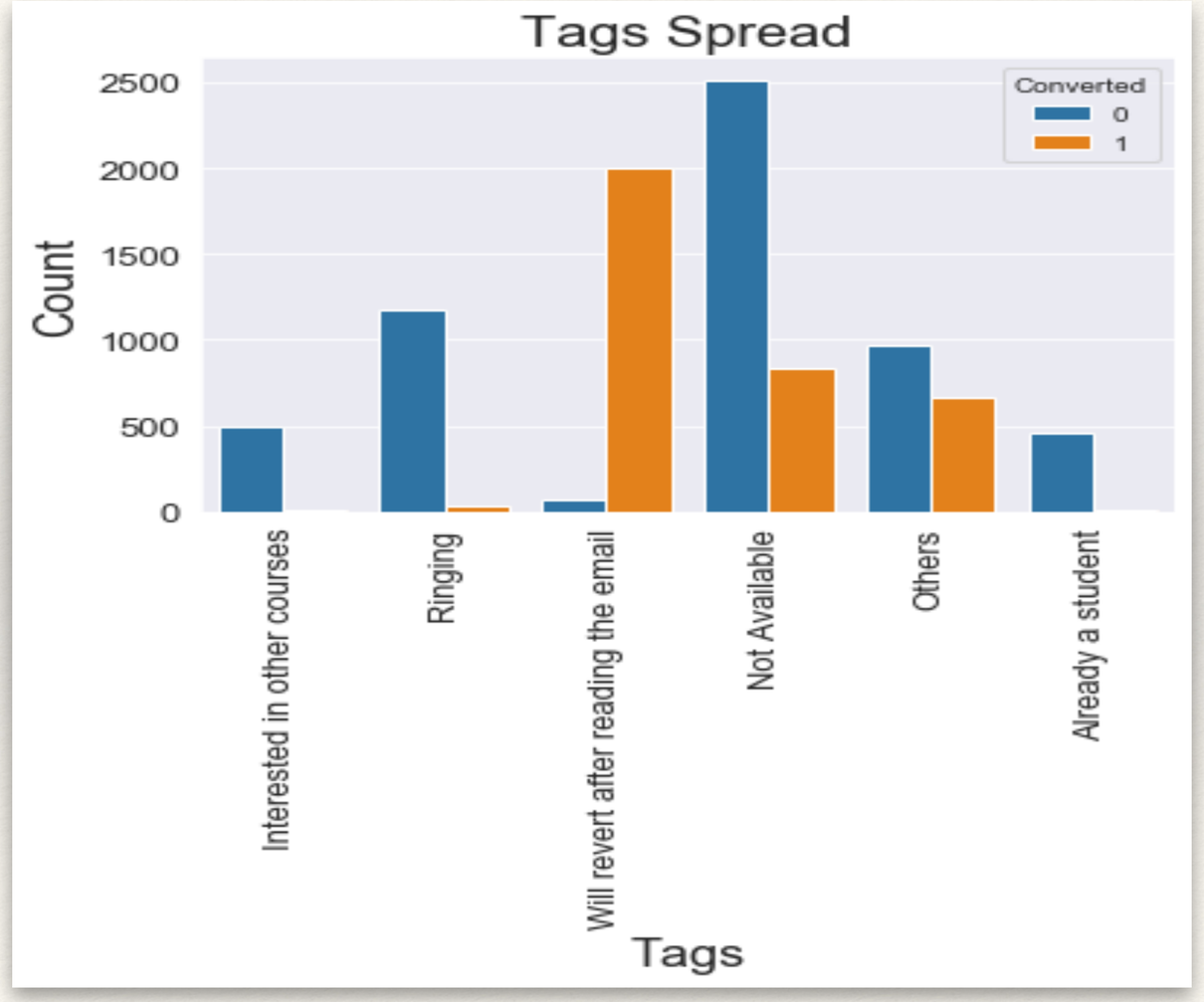
FEATURE: Lead Source

- Maximum number of leads are generated by Google and Direct traffic.
- Conversion Rate of reference leads and leads through welingak website is high than other sources.



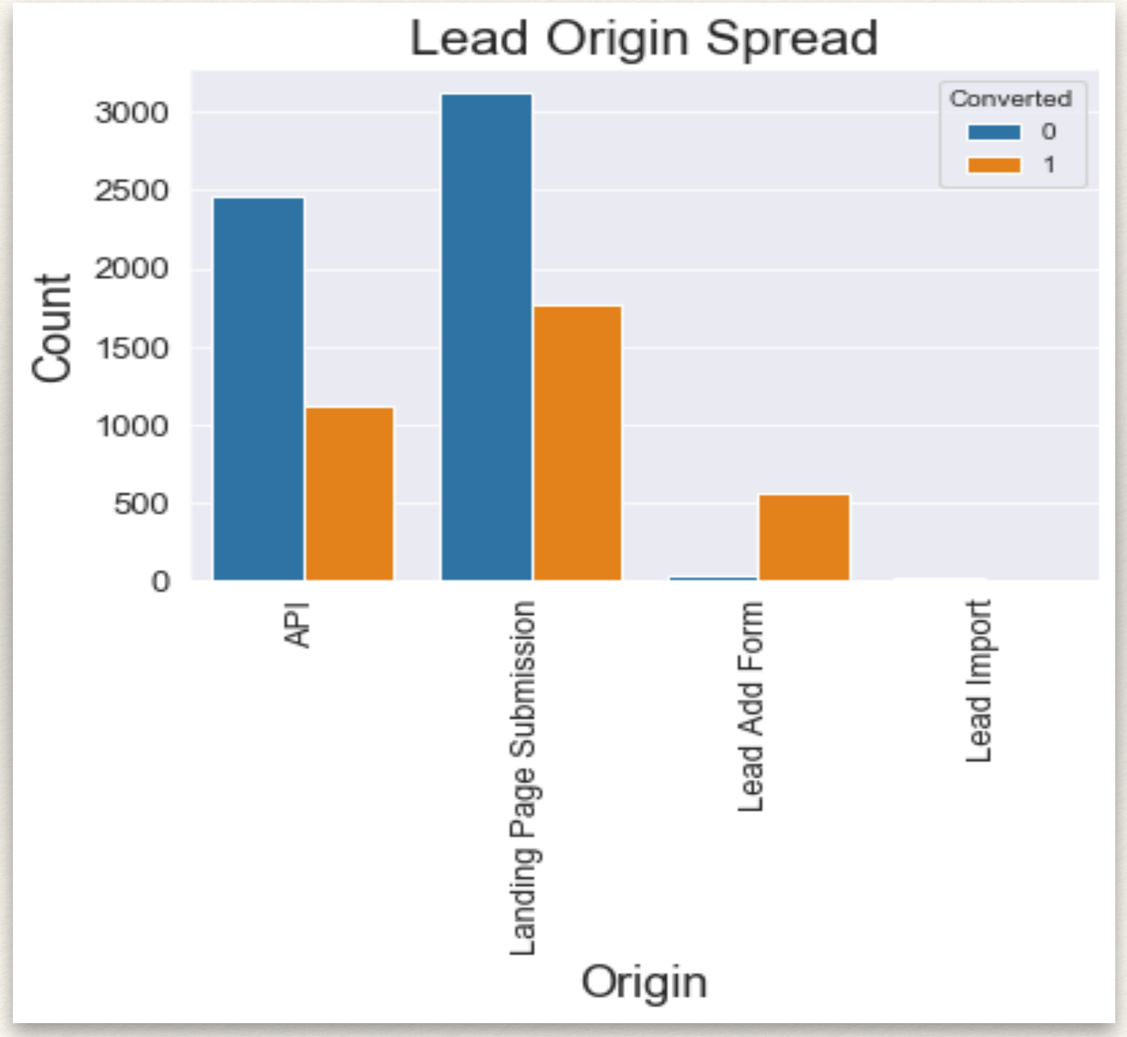
FEATURE: Tags

- Customer who said will revert after reading email have very high conversion rate.
- Those who are already a student have approx. 0% conversion rate.
- Ringing activity has very low conversion rate, maybe they are avoiding calls.
- Those who are interested in other courses have very low conversion rate.



FEATURE: Lead Origin

- API and Landing Page Submission bring higher number of leads.
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- Lead Import Form get very few leads.
- In order to improve overall lead conversion rate, we have to improve lead conversion of API and Landing Page Submission origin and increase number of lead add form as they have higher conversion rate.

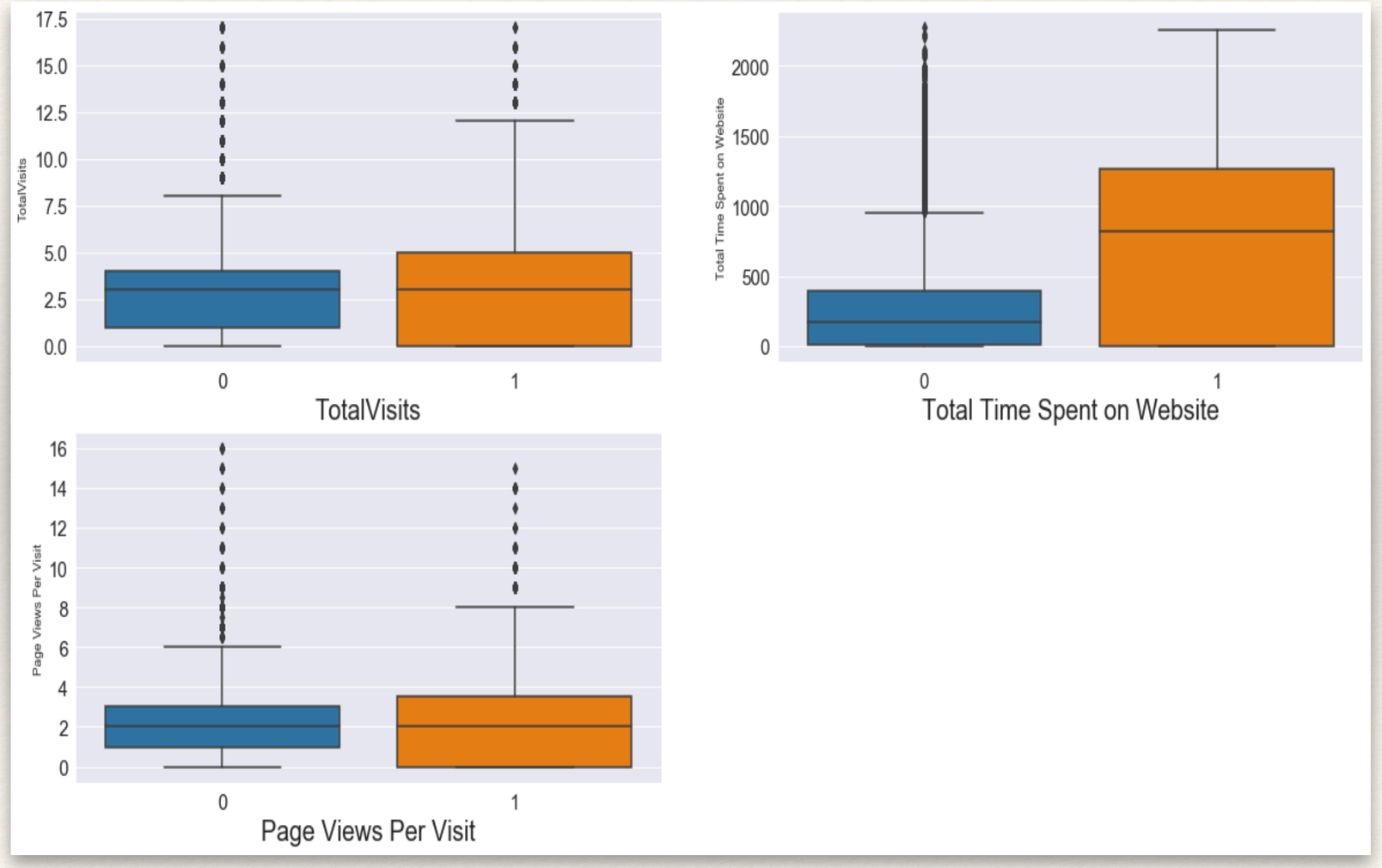


CORRELATION

- We don't have any high correlation.
- So, that's good

FEATURE: TotalVisits, Total Time Spent on Website & Page Views Per Visit

- Leads spending more time on the website are more likely to be converted.
- Median of TotalVisits and Page Views per visit is same for converted and non-converted



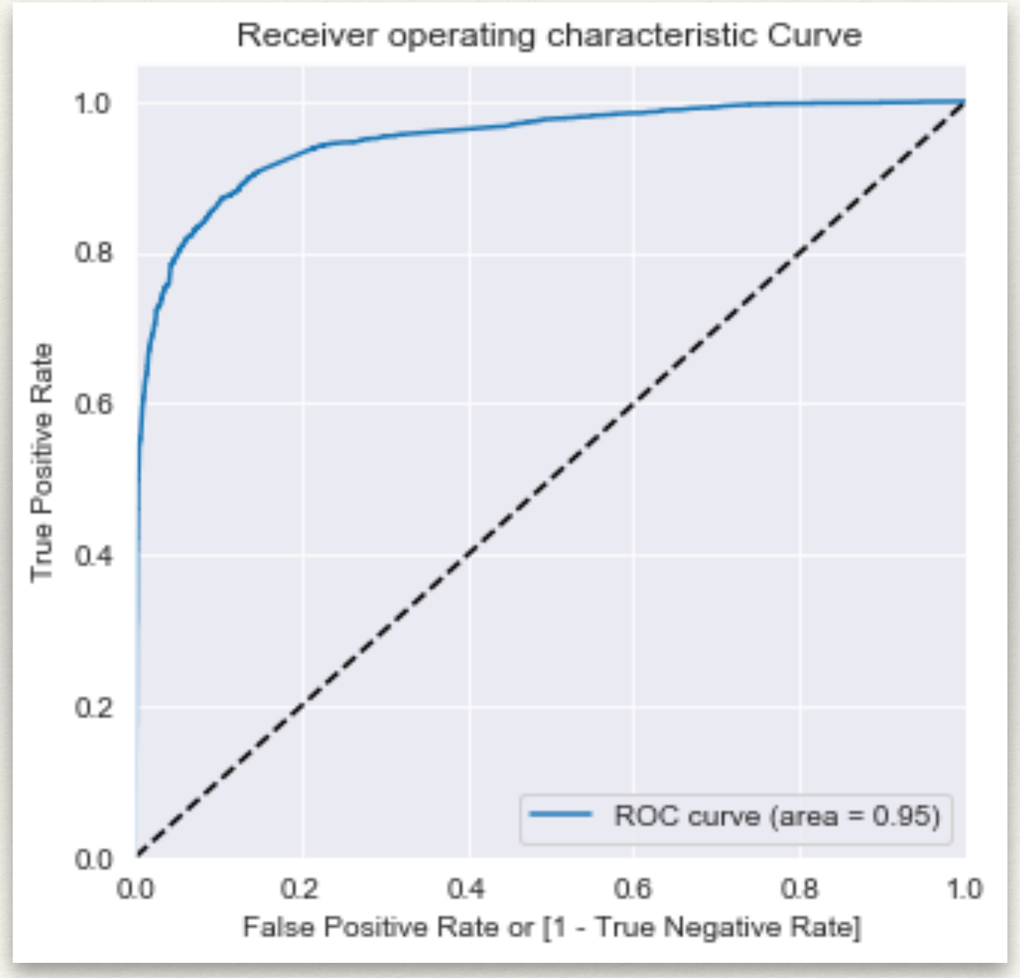
MODEL BUILDING

STEPS FOR MODEL BUILDING

- Divided the data set into Train and Test data sets in 70:30 ratio.
- Then will do scaling on 3 features (TotalVisits, Page Views Per Visit and Total Time Spent on Website).
- Now will create the model using Logistic Regression.
- Running RFE with 15 features.
- Will keep on removing one feature at a time until all the remaining features are significant (p-value should be less) and there VIF values should be less than 2%.

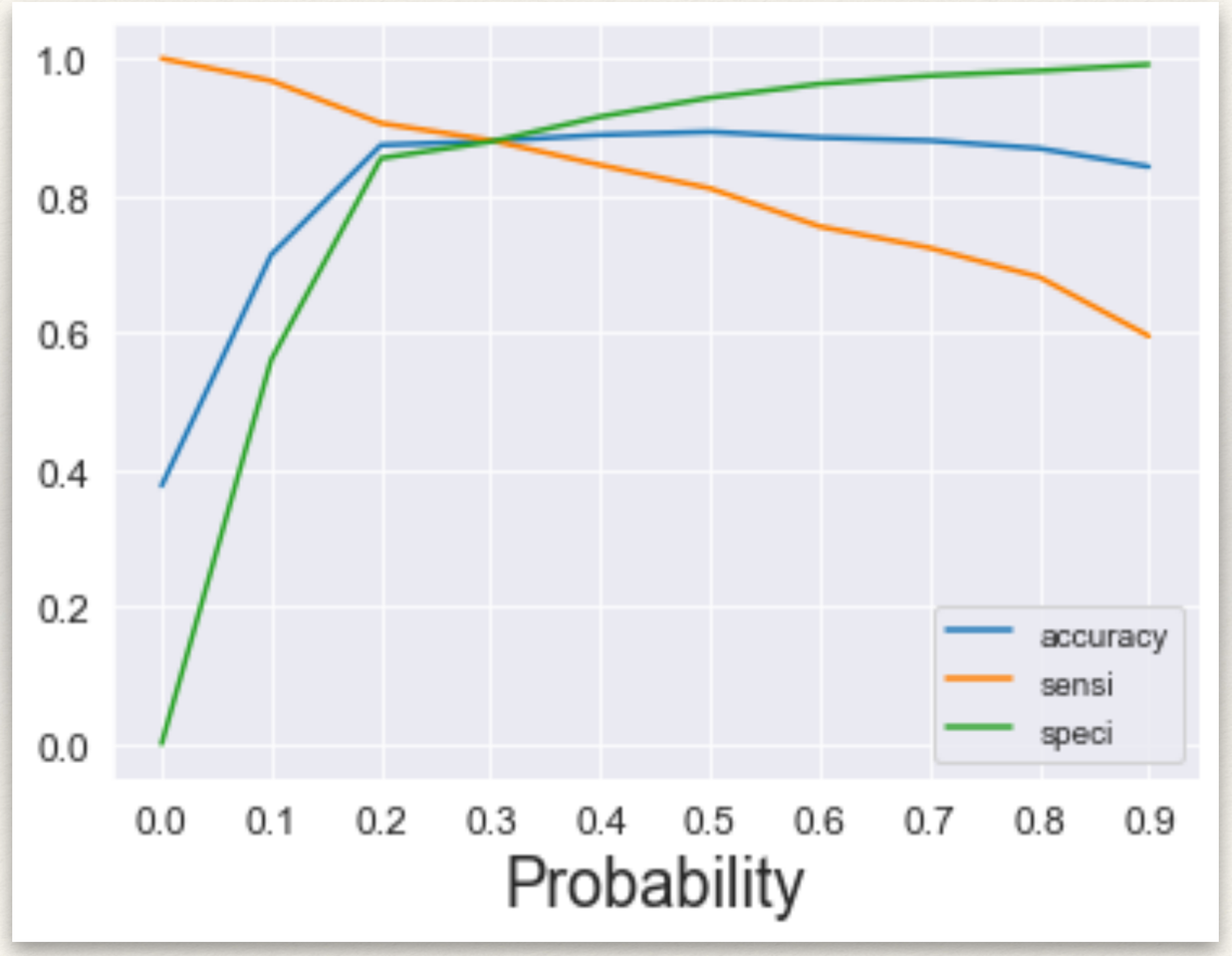
ROC CURVE

- Area under the ROC curve is 0.95 so we have pretty good model



OPTIMAL CUTOFF

- Optimal cutoff value comes out to be 0.3



PREDICTION

- After using 0.3 as cutoff we got following key points:
  - True Positive: 2093
  - True Negative: 3458
  - False Positive: 479
  - False Negative: 284

PREDICTION

	TRAIN DATA SET	TEST DATA SET
ACCURACY	87.91%	87.87%
SENSITIVITY	88.05%	87.78%
SPECIFICITY	87.83%	87.93%
PRECISION	81.37%	82.14%

CONCLUSION

- Features which are having positive coefficients:
  - When the lead origin is lead add form.
  - When the tags is will revert after reading the email.
  - When the lead source was:
    - Olark Chat
    - Welingak Website
  - When there current occupation is working professional.
  - When the last notable activity is SMS sent.
  - The total time spend on website is more.

CONCLUSION

- Features which are having negative coefficients:
  - When do not send email is yes.
  - When the tags are:
    - Ringing.
    - Interested in other courses.

THANK YOU