Lead scoring case study

By : Jayshree Sahoo and Shreeyash

Data sourcing

- Data was provided by Upgrad
- Master data and data dictionary were present
- Master data had 9240 rows and 37 columns
- It had both system and sales generated data

Exploring the data

- X Education sells online courses to industry professionals
- They have sales team that reaches out to the leads and try to sell them the courses
- We have different columns namely, lead origin, specialization, converted, total visit, do not call, do not email etc
- Some of the data is sales generated so we may have to drop them
- Some of the columns are numerical and some are categorical
- There seemed to be outliers in the numerical columns
- Categorical columns had high values as select, per the data dictionary these are null values and need to be replaced by nan

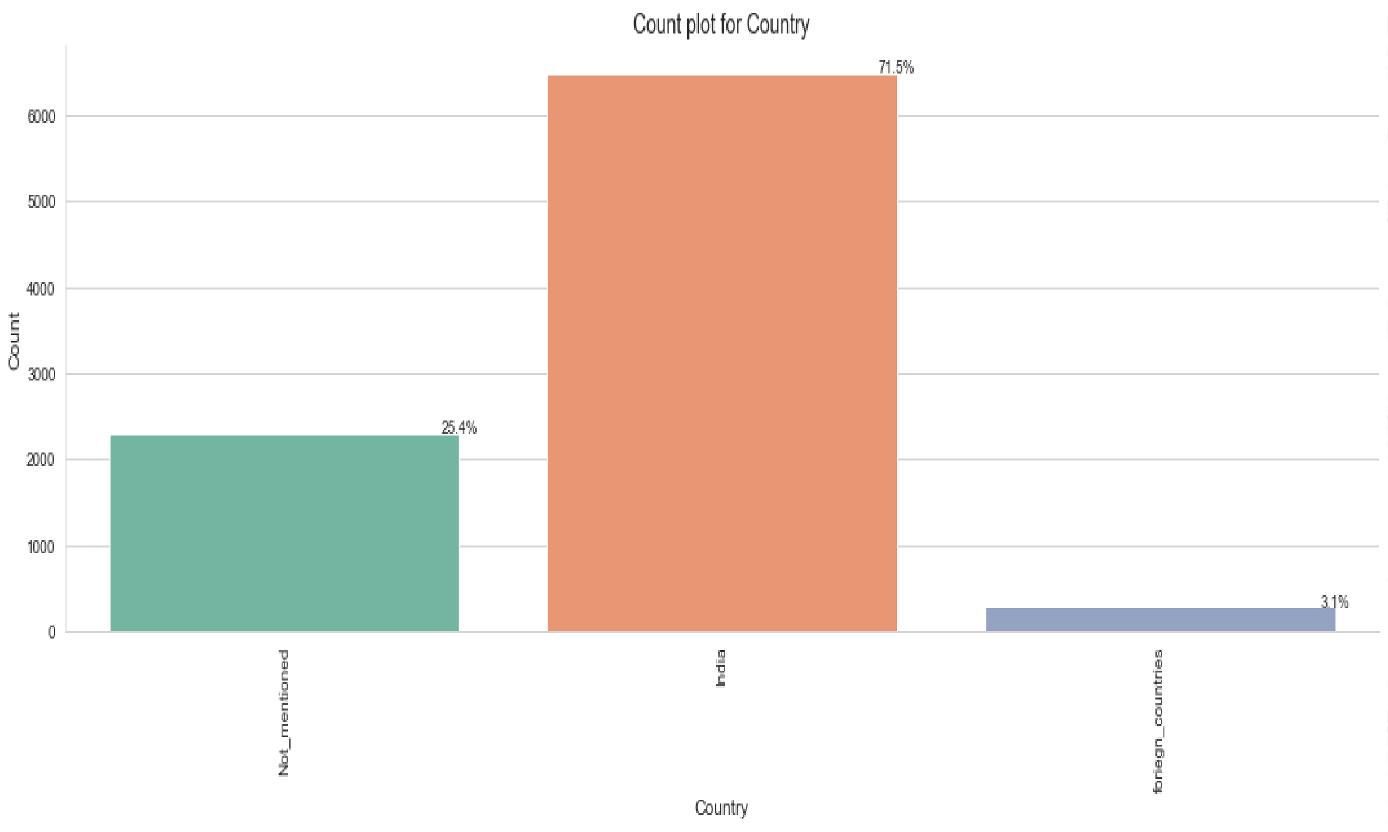
Data cleaning

- Checking of missing values
    - Replacing select value by NaN, as python will not recognize select as null- value
    - Dropping the columns having missing values more than 40%
- For some categorical columns we replaced null values as not mentioned
- Combining the variables of the categorical columns that had low individual value count
- We have renamed some of the columns as their names were too large
- Checking for outliers
    - We used box plots to identify the outliers here
    - We capped all the numerical columns except lead number(as it is unique), between .05 and .95 to remove the outliers

EDA

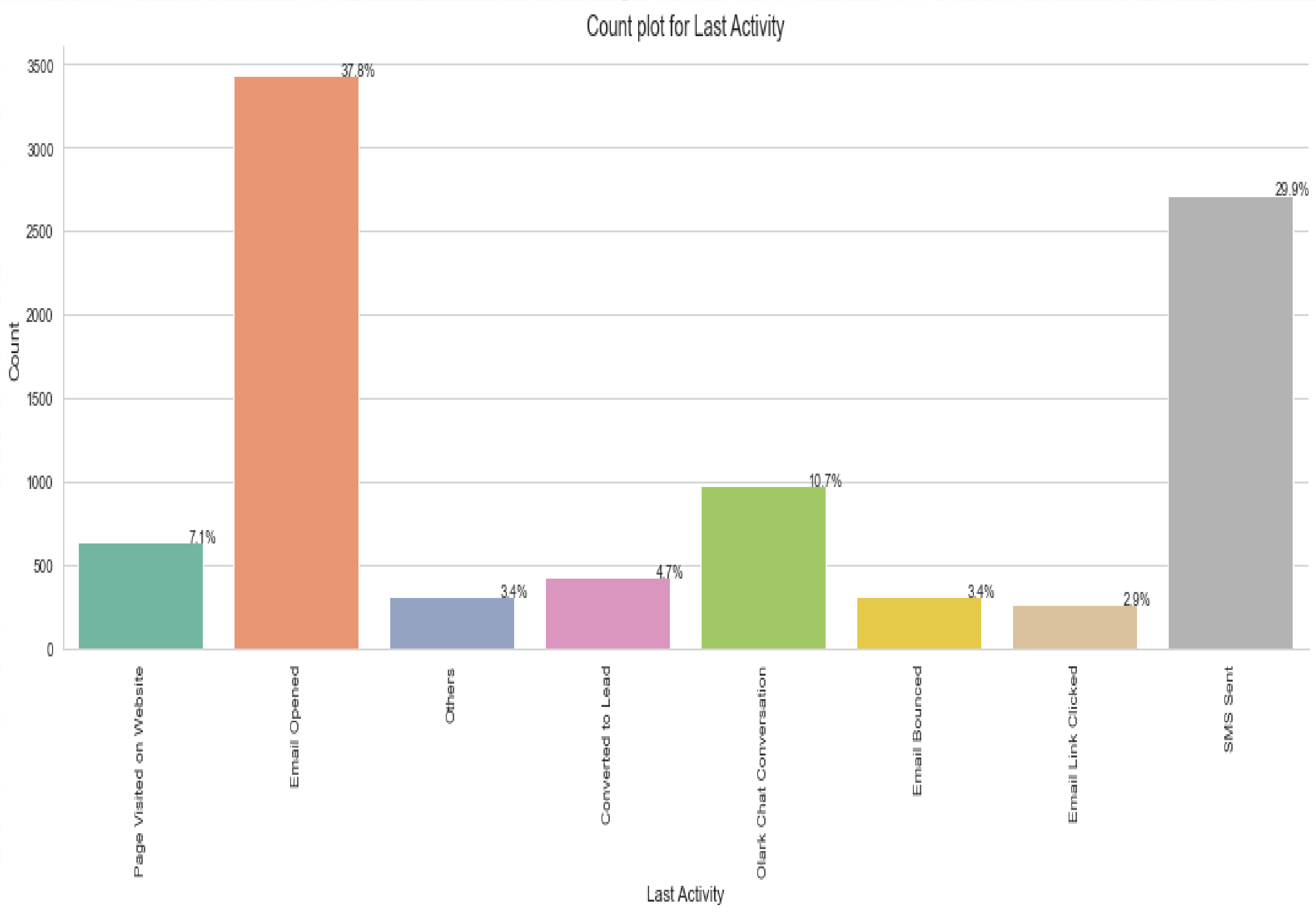
Univariate Analysis

- Country
  - Here we can see that most of the leads are from one category i.e India, we may not need this column for our model



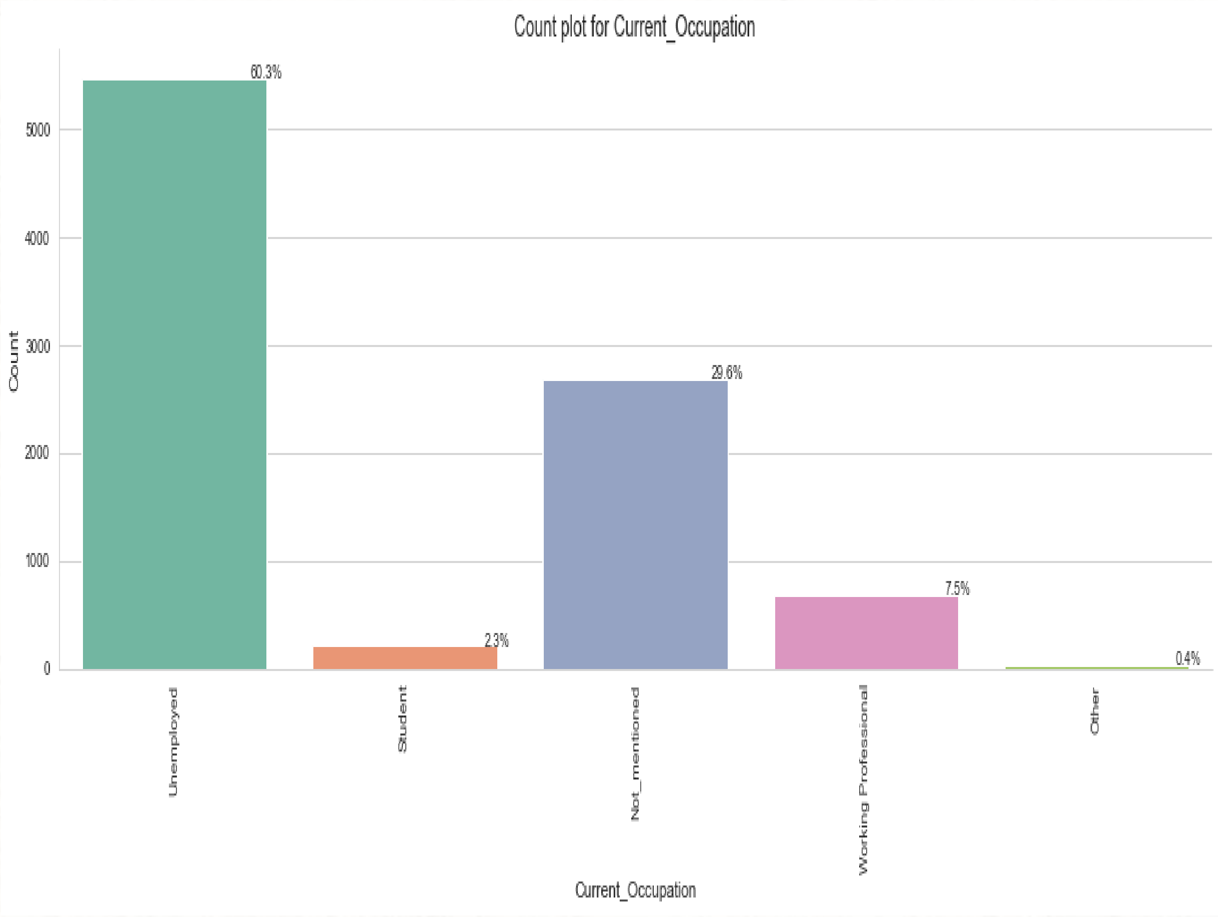
Last Activity

- Here we combined all the variables that had very low value count to others
- However this is a sales generated data so we may not consider this as well for our model



Current Occupation

- Here we can see that most of the leads are unemployed.
- Also most of them have not mentioned there current occupation

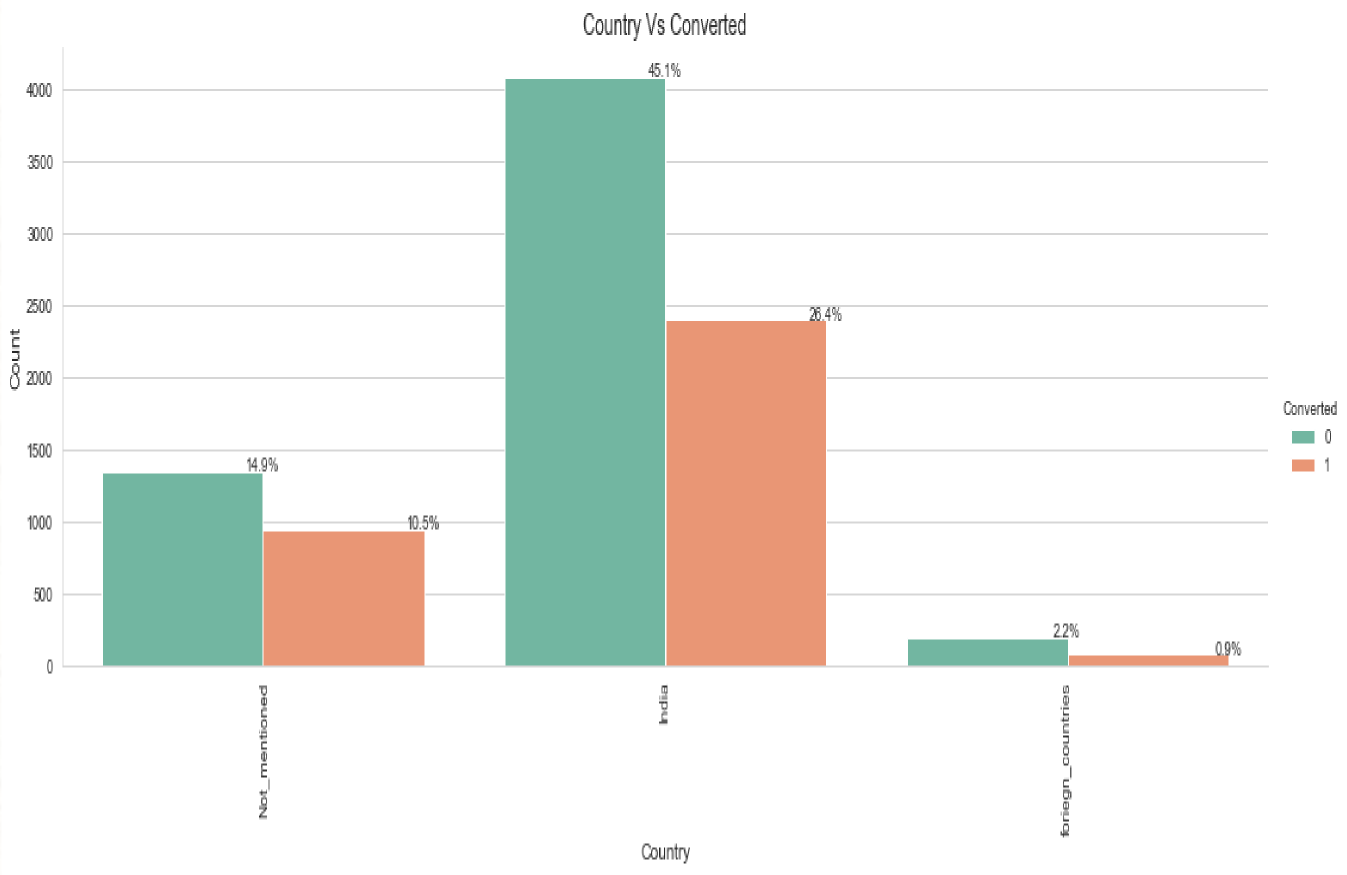


- As visible in the graph the data is highly skewed so these won't be of any help for our model
- We will not consider these columns for our model
- Also there are columns that are sales generated we will drop those columns to avoid ambiguity.

Bivariate Analysis

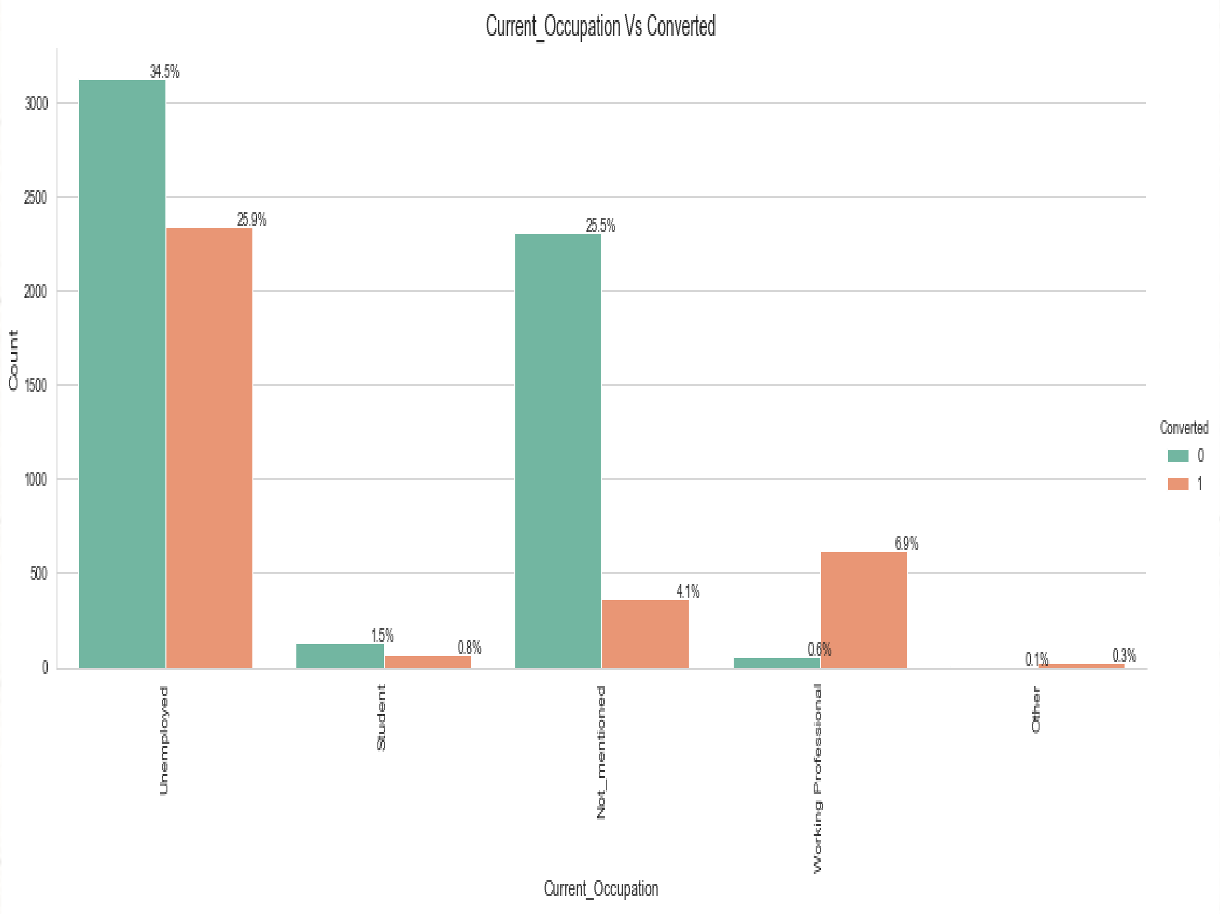
Country

- Here we can observe that the data is skewed so we will be dropping this column



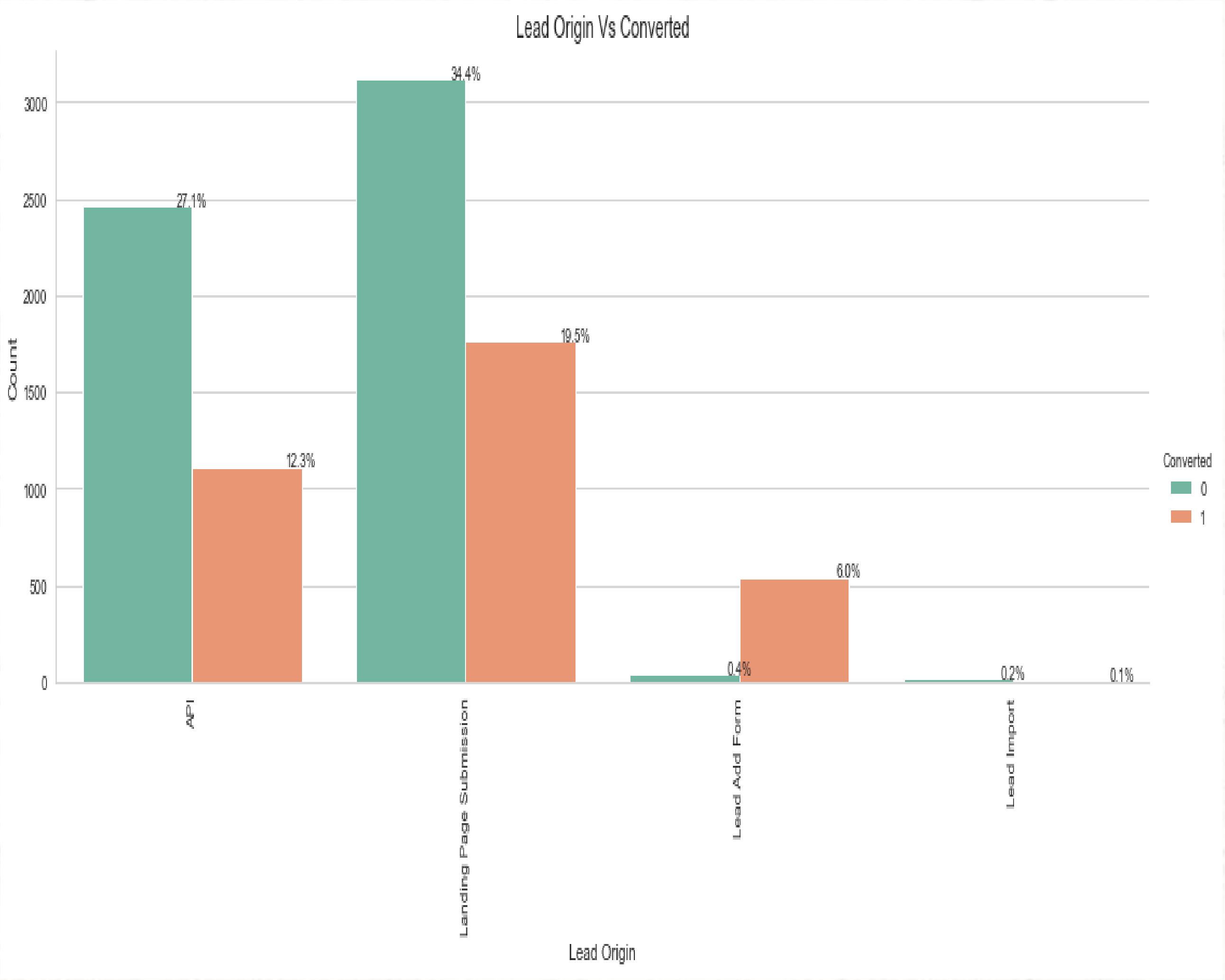
Current Occupation

- Here we can see that the maximum conversion is for the unemployed group
- Leads who have not mentioned their occupation have a very less conversion rate.



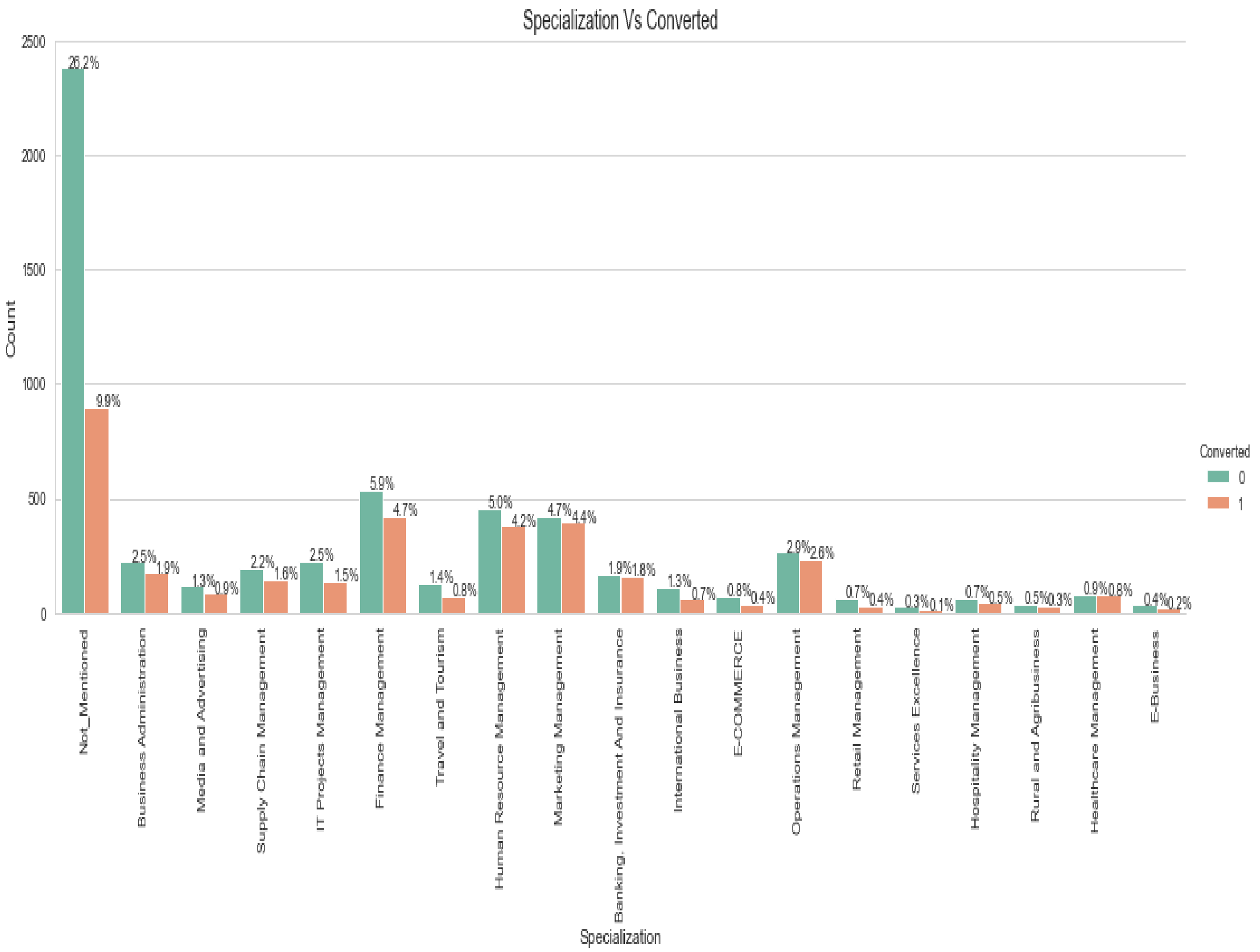
Lead Origin

- The leads originated from ‘Landing Page Submission’ contributes the most.
- The leads originated from ‘Lead Add Form ’ have very high conversion rate



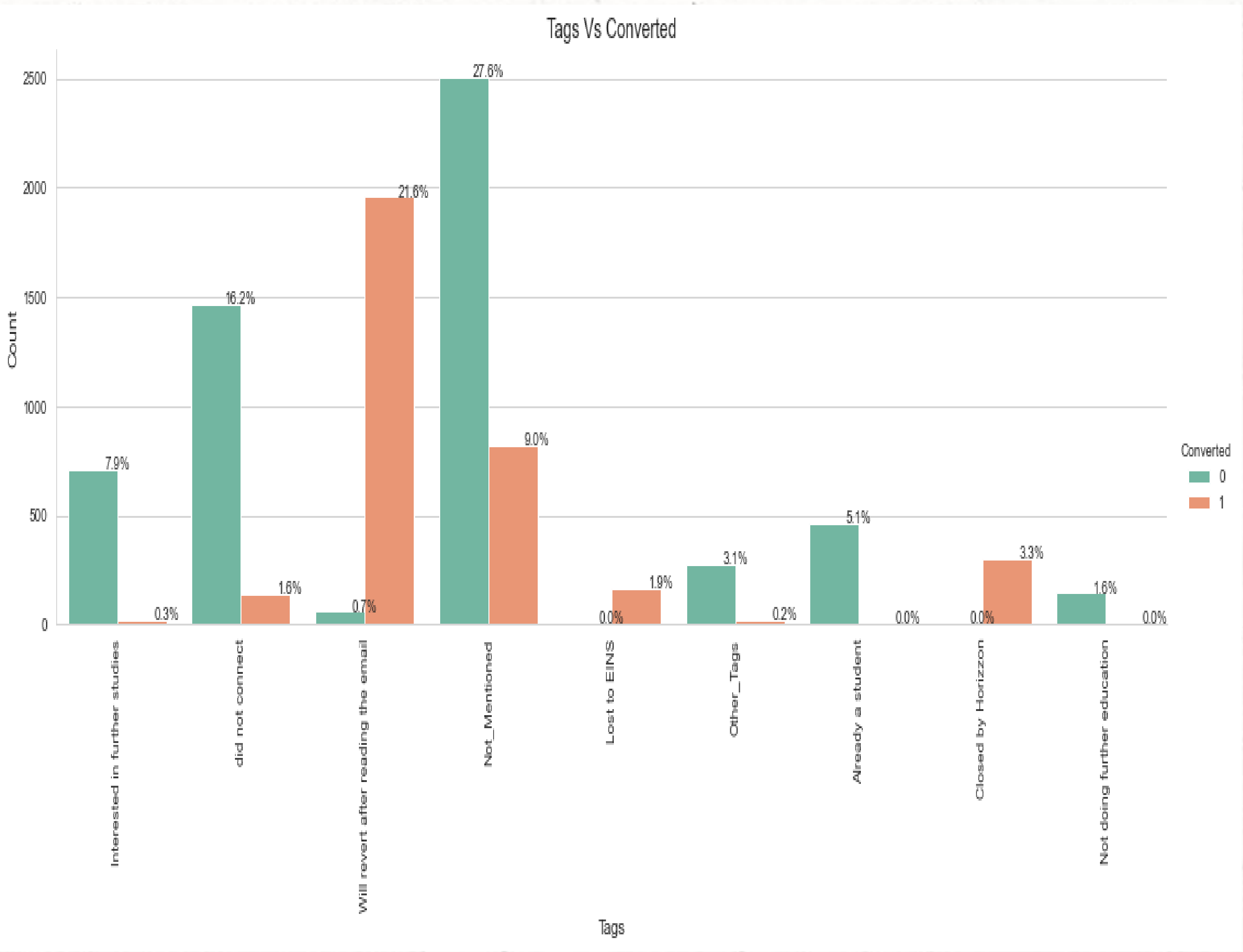
Specialization

- Leads in management courses have a very good conversion
- Leads with not mentioned specialization have very less conversion rate



Tags

- This is highly skewed and sales generated data so we will remove this column.

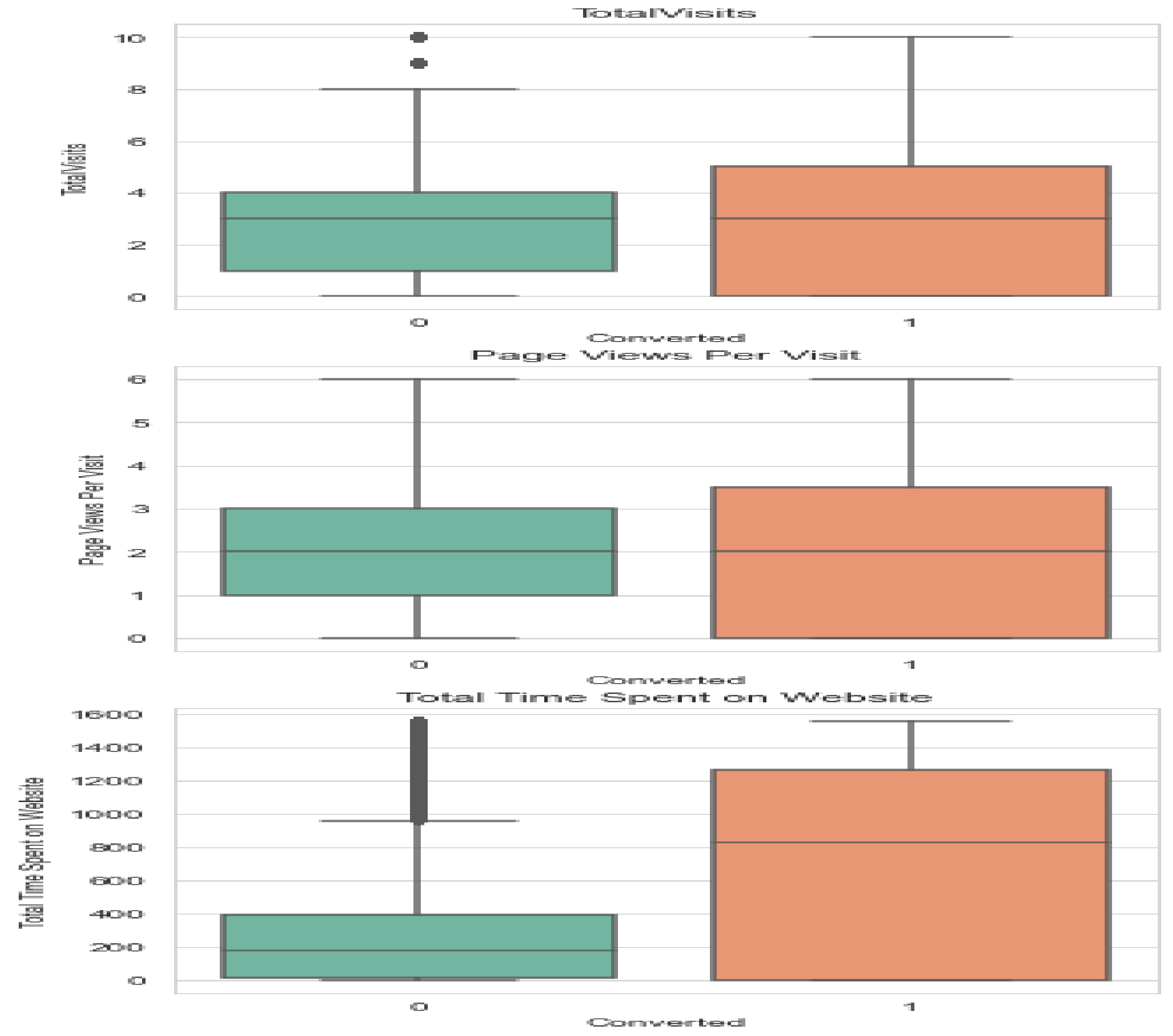


Sales generated columns

- These columns are sales generated and is of no use to our model also they are highly skewed
- Also most of the columns have “No” value

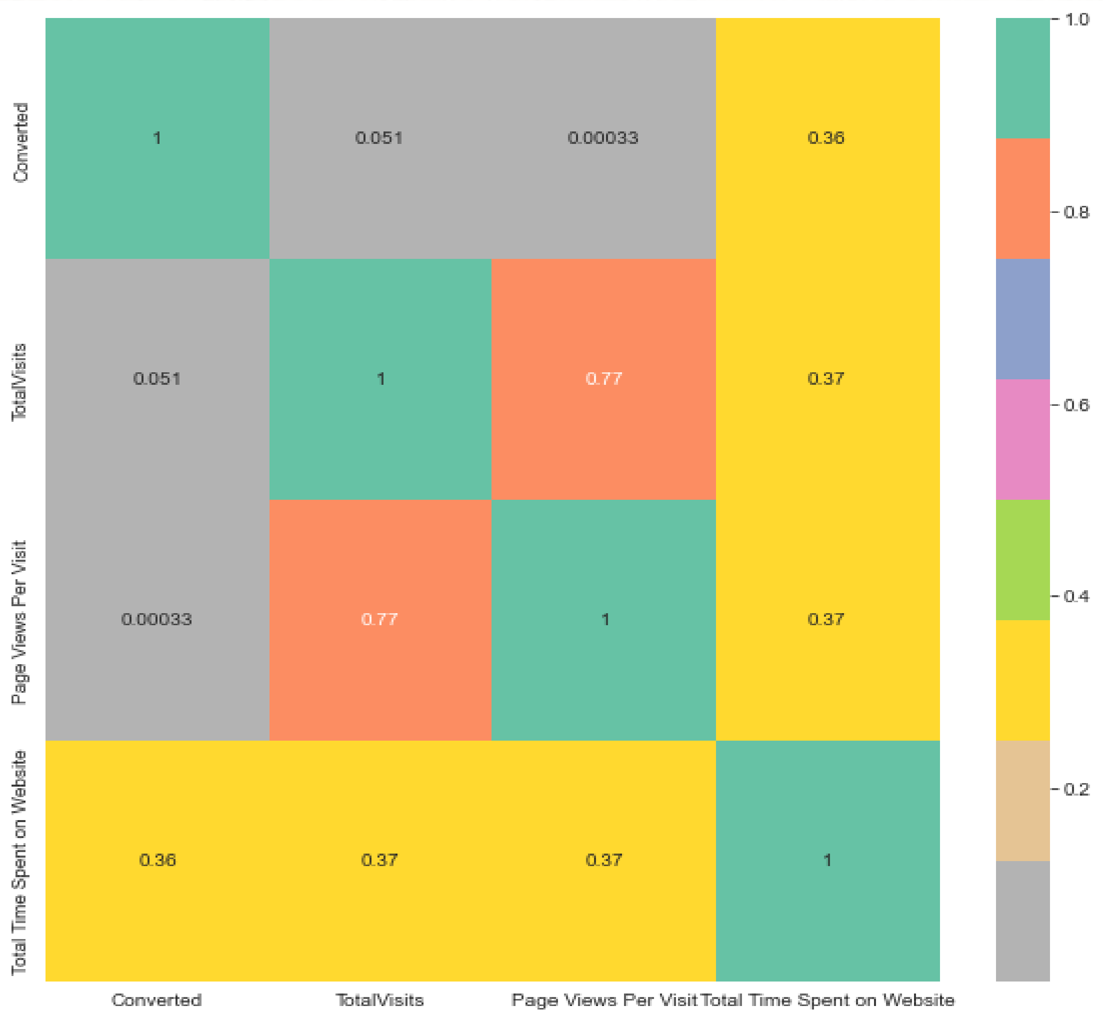
Numerical Columns

- Data looks good
- No outliers and the columns can be used for our final model



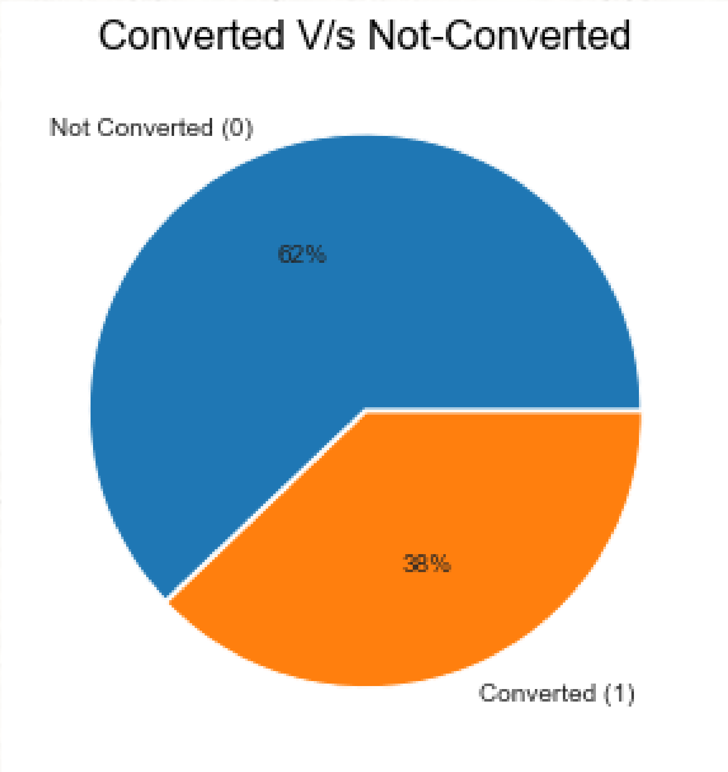
Heatmap for numerical variables

- Here we can see that total time spent on the website has a maximum correlation with the converted columns which is our target column.
- Also, TotalVisits and Page Views Per Visit are highly correlated to each other.



Converted vs Not - converted

- Here we can see that 62% of the data belongs to the leads which were not converted
- The conversion rate was only 38%.



Data preparation

- Here we dropped the columns which were not important for our model, as discovered during the EDA, data exploring, and cleaning
- Then we created the dummy columns for the all the remaining categorical columns
- Then we checked the correlation between dummy columns using heatmap and dropped the highly correlated columns
- We did the train-test split in the ratio of 7:3
- We did the scaling for the numerical column of the train data using Standard Scaler method.

Model building

- We used logistic regression for our model using sklearn and stats model
- We used RFE to select the top 15 features for our model
- Then we manually dropped the features with p-values greater than .05 and VIF values greater than 5.0 one by one until we had a model with features having desired p-values and VIF value

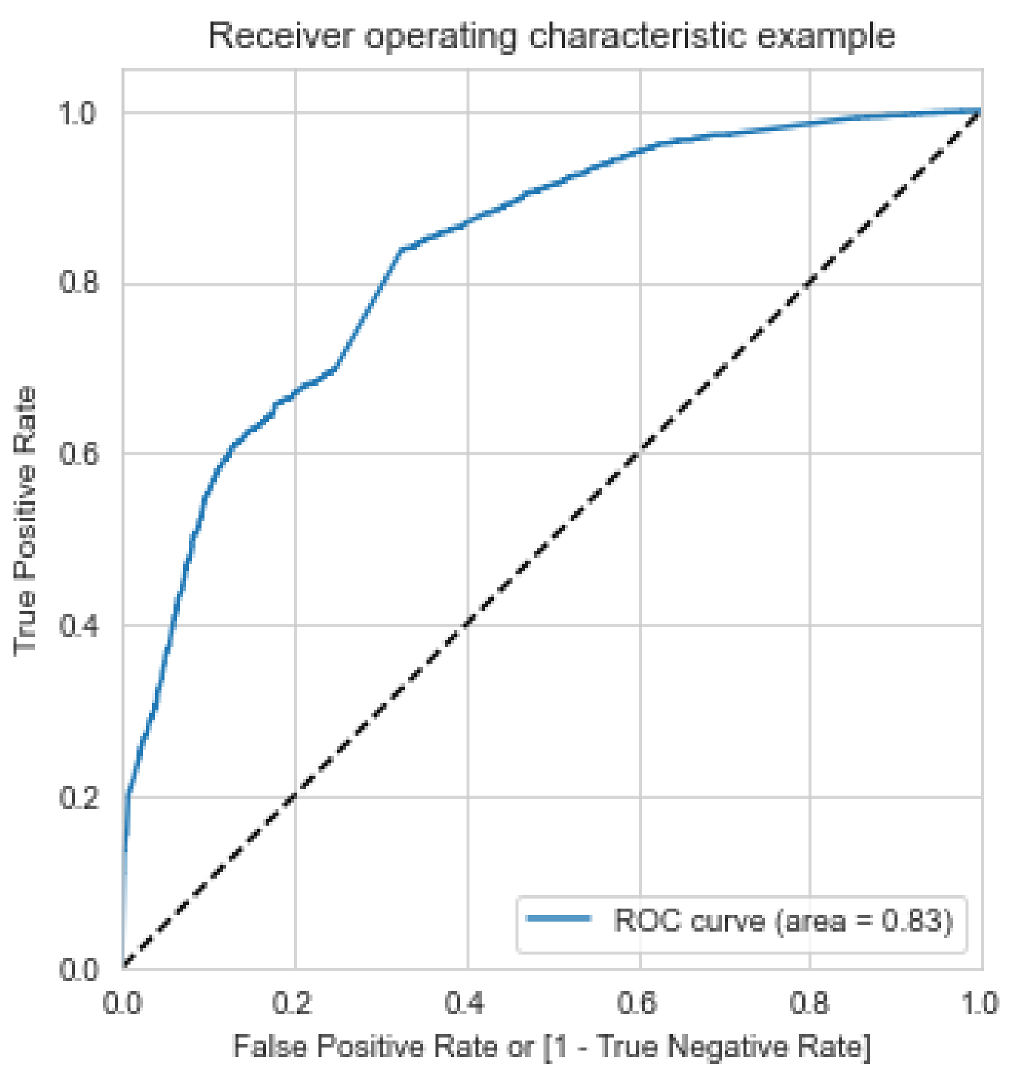
Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6342
Model Family:	Binomial	Df Model:	8
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3105.6
Date:	Tue, 10 May 2022	Deviance:	6211.2
Time:	11:43:08	Pearson chi2:	6.20e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9046	0.072	-26.292	0.000	-2.047	-1.763
Total Time Spent on Website	0.9620	0.035	27.531	0.000	0.893	1.030
Page Views Per Visit	-0.4195	0.037	-11.484	0.000	-0.491	-0.348
Lead Source_Welingak Website	4.6006	0.717	6.418	0.000	3.196	6.006
Specialization_Marketing Management	0.2600	0.107	2.439	0.015	0.051	0.469
Current_Occupation_Other	2.7336	0.525	5.207	0.000	1.705	3.763
Current_Occupation_Student	1.2981	0.201	6.460	0.000	0.904	1.692
Current_Occupation_Unemployed	1.4935	0.081	18.522	0.000	1.335	1.652
Current_Occupation_Working Professional	4.2231	0.183	23.107	0.000	3.865	4.581

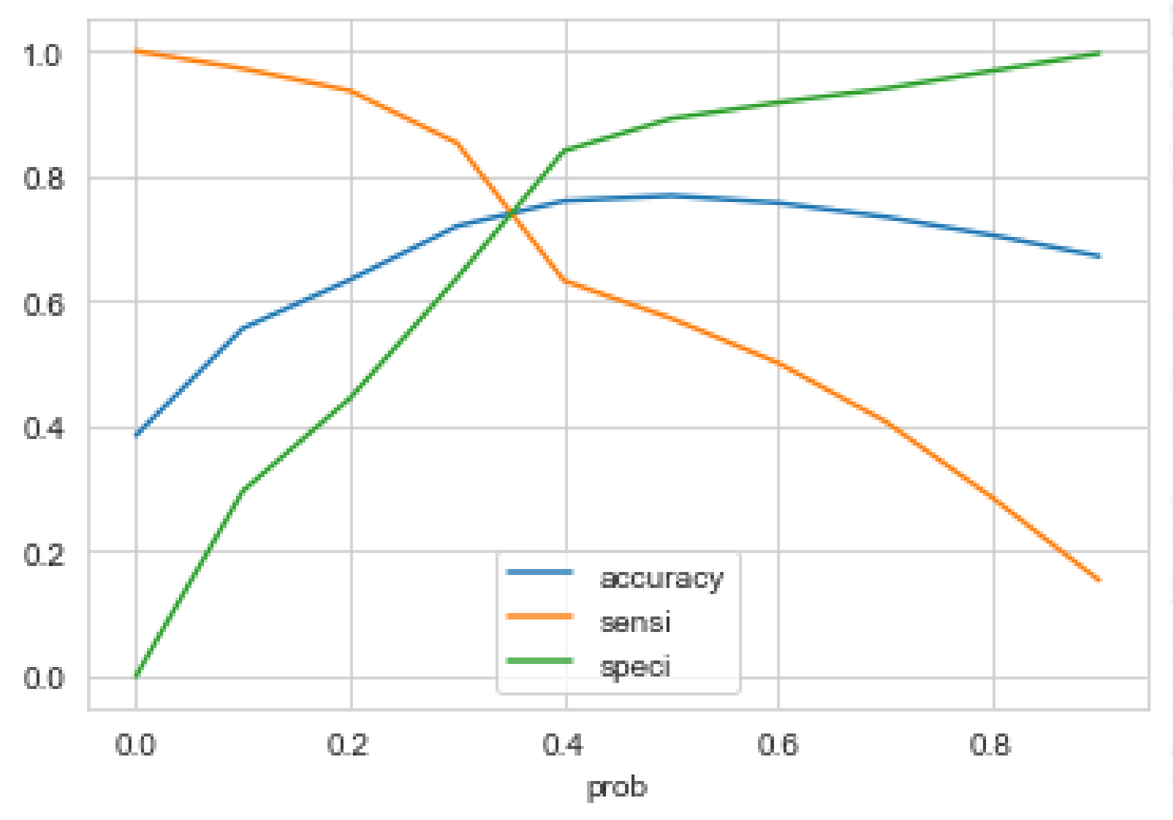
Model Evaluation

- Here the area under the ROC Curve is .83 which is good.
- The specificity, sensitivity and accuracy is around 63%, 85% and 70% respectively
- We found the optimal cut off specificity and sensitivity plot which was around 0.3
- The conversation rate for the test data set was around 85%

ROC Curve



- We will consider 0.3 as the optimal cut off for the final prediction on test data set
- After the prediction of the test data set we obtained the specificity as 62%, sensitivity is around 84% and accuracy is around 70%
- Also the final conversation rate is around 84% which was the target



Conclusion

- The features that contribute the most towards the probability of leads getting converted are :
  - Lead Source_Welingak Website
  - Current_Occupation_Working Professional
  - Current_Occupation_Other

- The lead score calculated for the test dataset and train dataset shows the conversion rate of 84% and 85%,