Lead Source Case Study

GROUP MEMBER
1. Yathestha Siddh
2. Puneet Sharma

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads but the conversion rate is very low. Let's say on particular day 100 leads, from them only 30% are getting converted
- The company wishes to identify the most potential leads and make the process efficient to identify those potential customers.
- This will lead to more conversion of leads as sales team will now be more focused on the commutating with the potential buyer.

Business Objective:

- X Education wants to identify the most potential leads
- For this they want a model which can identify the HOT LEADS
- Deployment of the model for future use

Solution Steps:

- Data understanding & Data importing:
- Data Cleaning & Handle the “Select” level that is present in many of the categorical variables.
- Drop columns that are having a high percentage of missing values & Impute which are having less missing values
- Check the number of unique categories in each categorical column.
- Check & Handle Outliners
- EDA
- Create dummies for all categorical columns.
- Perform train-test split & Perform Feature scaling.

- ***Model Building:***
- Build a Logistic Regression model for predictions.
- Check the model performance
- Conclusion

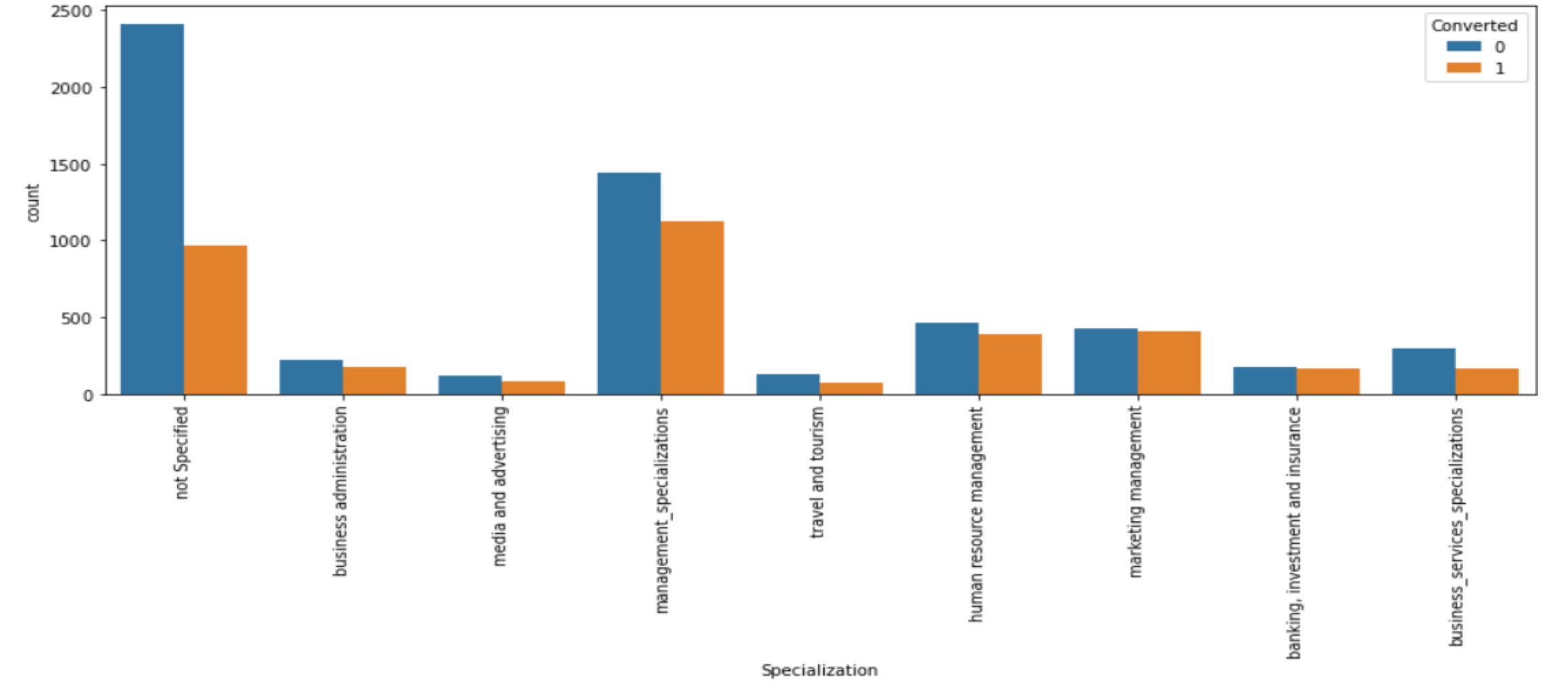
.

DATA Manipulation

- Total No of Rows are 9240 & Columns are 37.
- Dropping the columns which are having null values more than 45% like ‘City’, ‘Lead Quality’ etc as we can not get anything from this data.
- The columns which are having less than 45% missing values imputation is the best to avoid any data loss.
- Drop the categorical columns which are having more than 99% data with single value.
- Removing Prospect ID & Lead Number as they are of no use in the analysis.

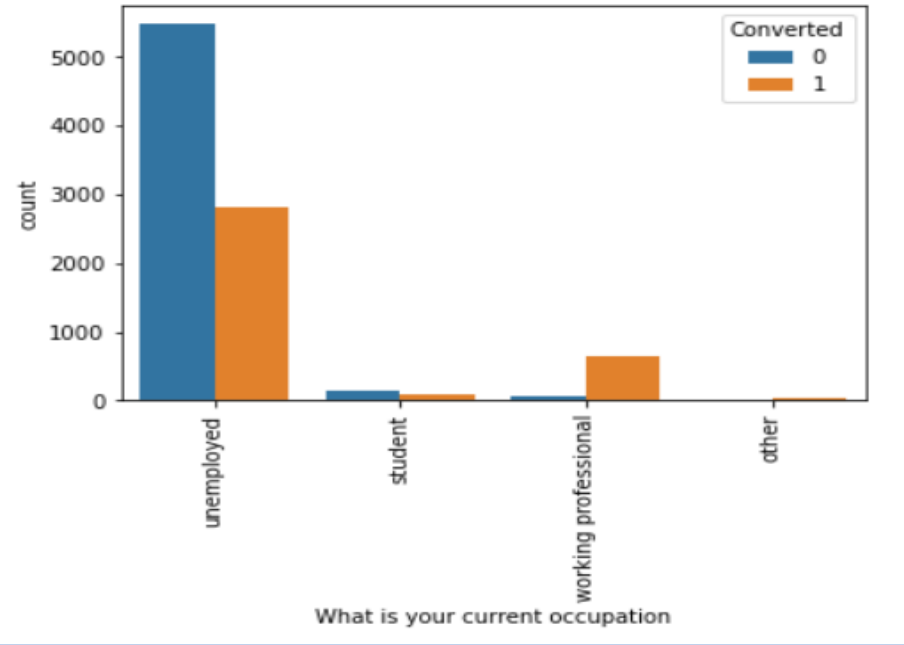
EDA

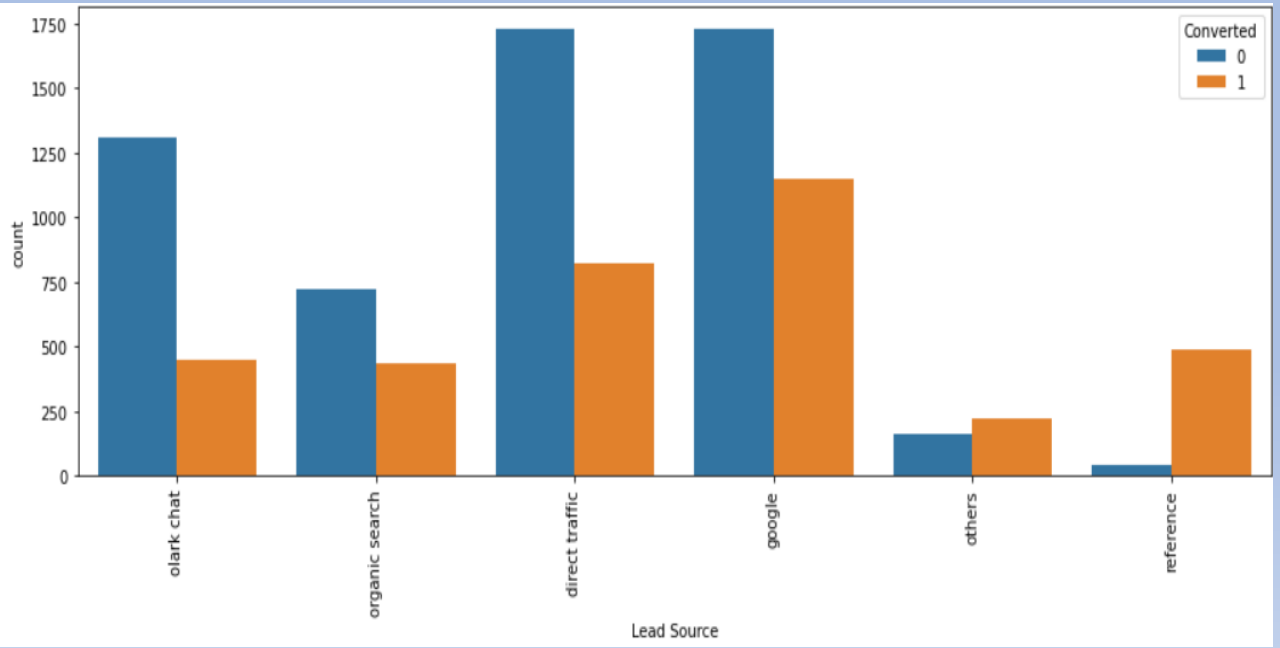
From the below graph of specialization that management is the highly converted variable.



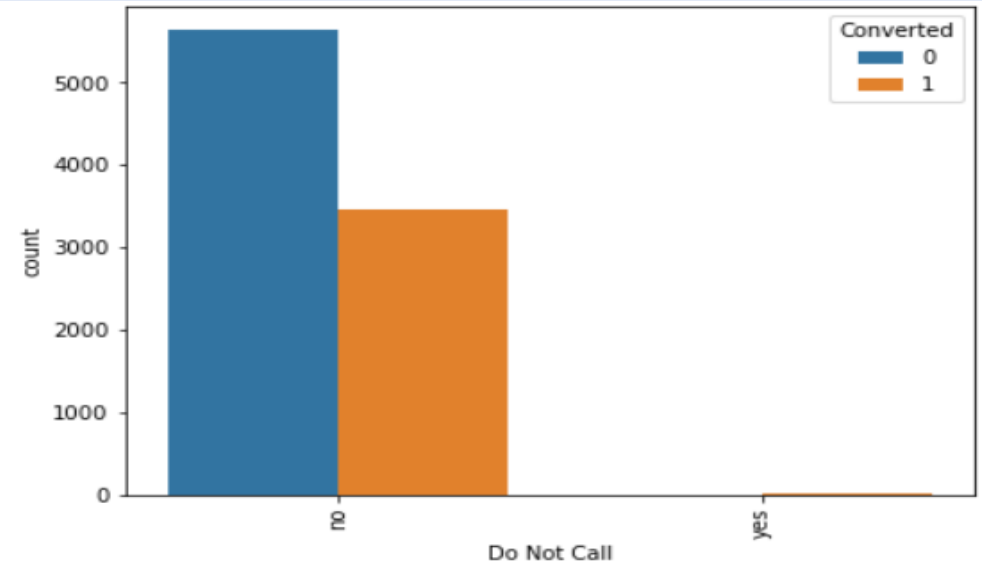
Occupation & Lead Source Wise Distribution

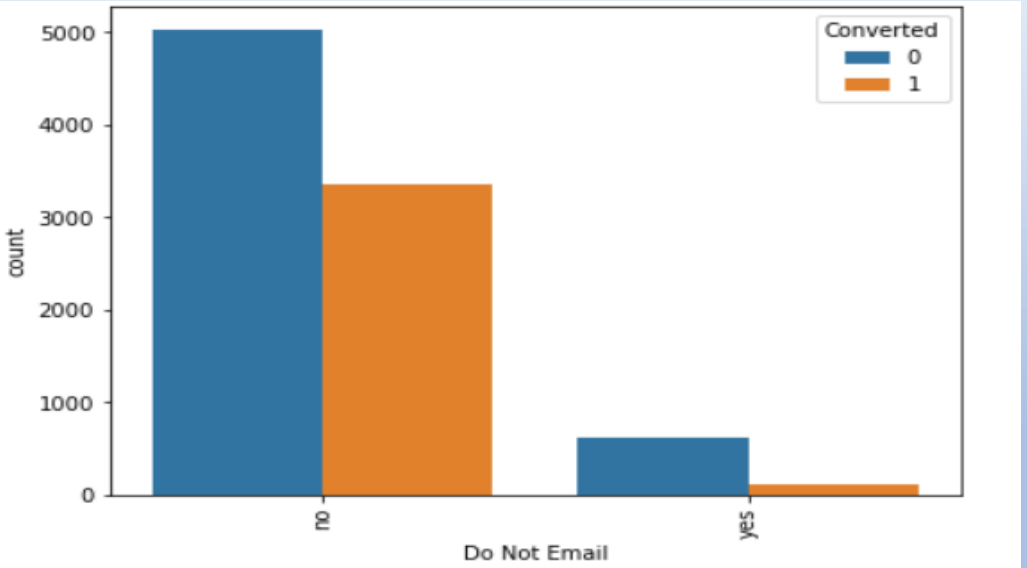
- Occupation Wise: We can Clearly observe that the conversion rate for the working professional is high. Though unemployed also showed high but the leads also high.
- Lead Source Wise: It is observed that reference has the higher conversion rate than comparison to the other fields. As Google & Direct traffic also seems high but the main reason is the no of leads also high in the source

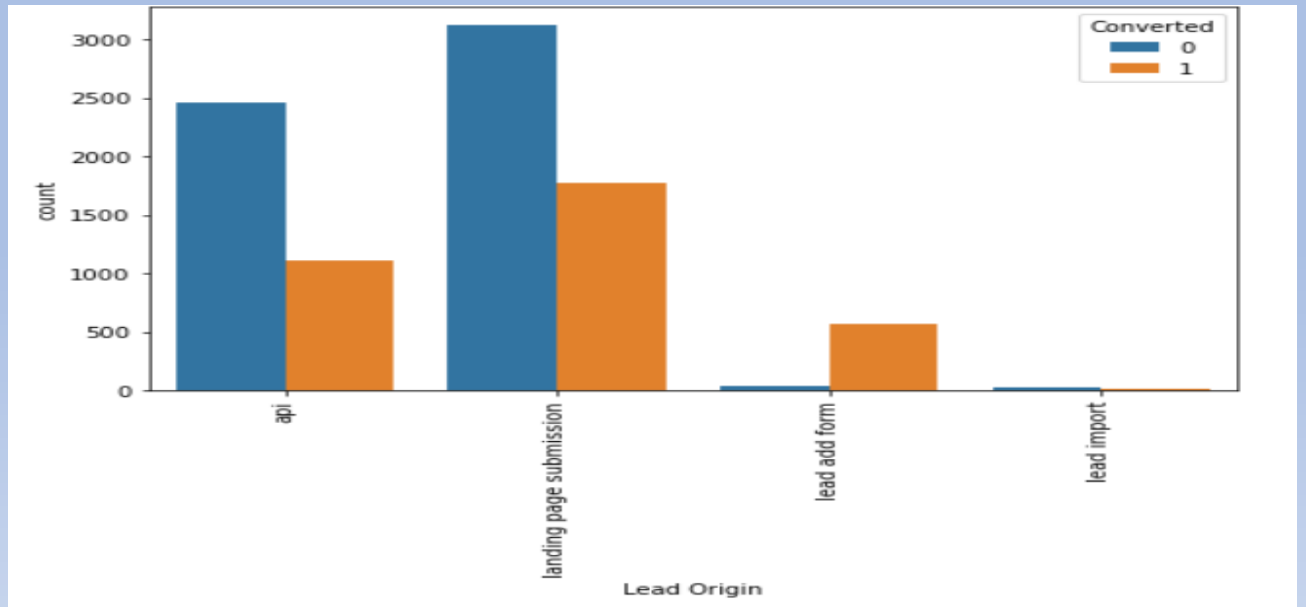




Do Not Call, Do Not Mail & Lead Origin wise

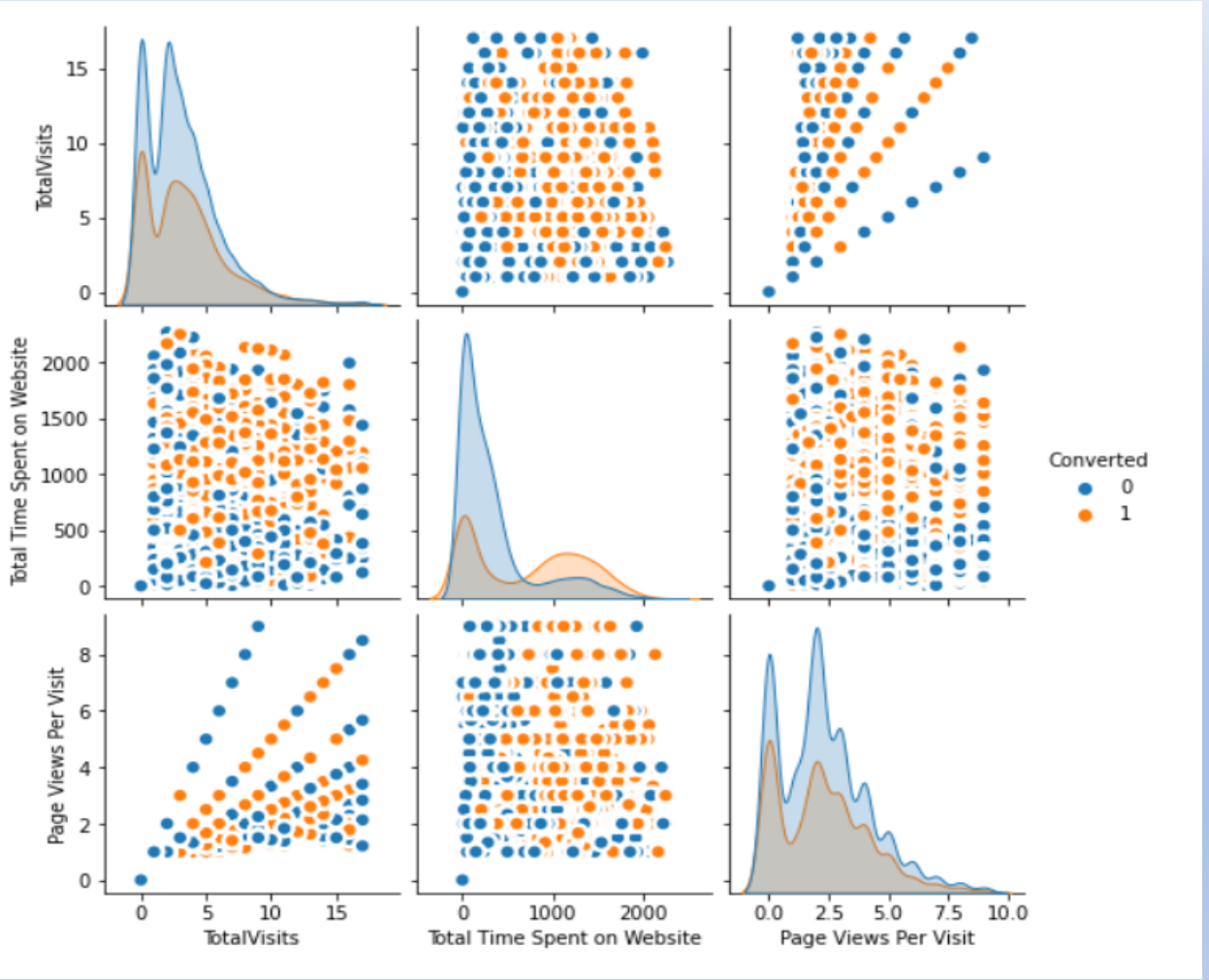






Correlation Matrix:

Scatter Plot & Heat Map :



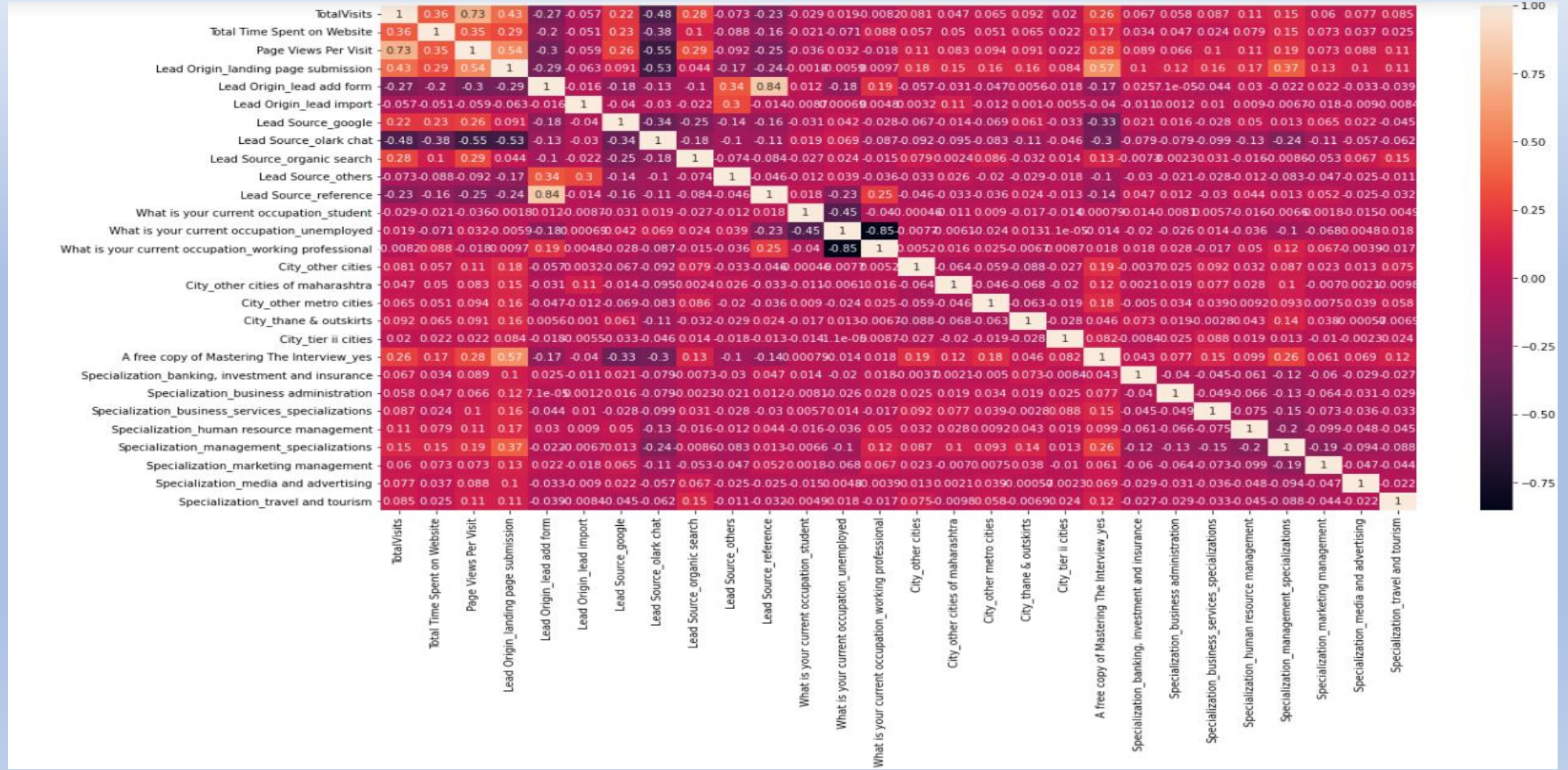
Data Conversion:

- Dummy Variable are created for Object type variable
- Total Rows for analysis are 8953
- Total columns for analysis are 29

Model Building:

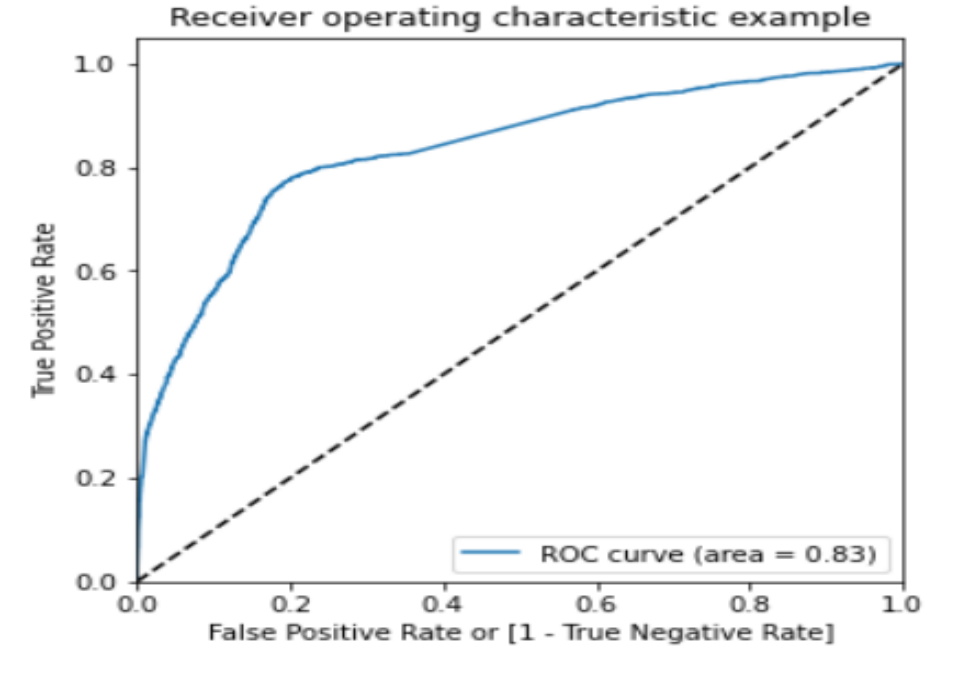
- Split the data into Train & Test sets
- First step towards the regression model is to do train-test-split.
- Feature Scaling
- Check the correlation matrix and drop the variables which are highly correlated for model performing well.
- Use RFE for Feature Selection & Running with 14 variables
- Building model by dropping the variables which are having P-value >.05 and VIF greater than 5.
- Prediction on Test data
- Overall accuracy is 79.63%

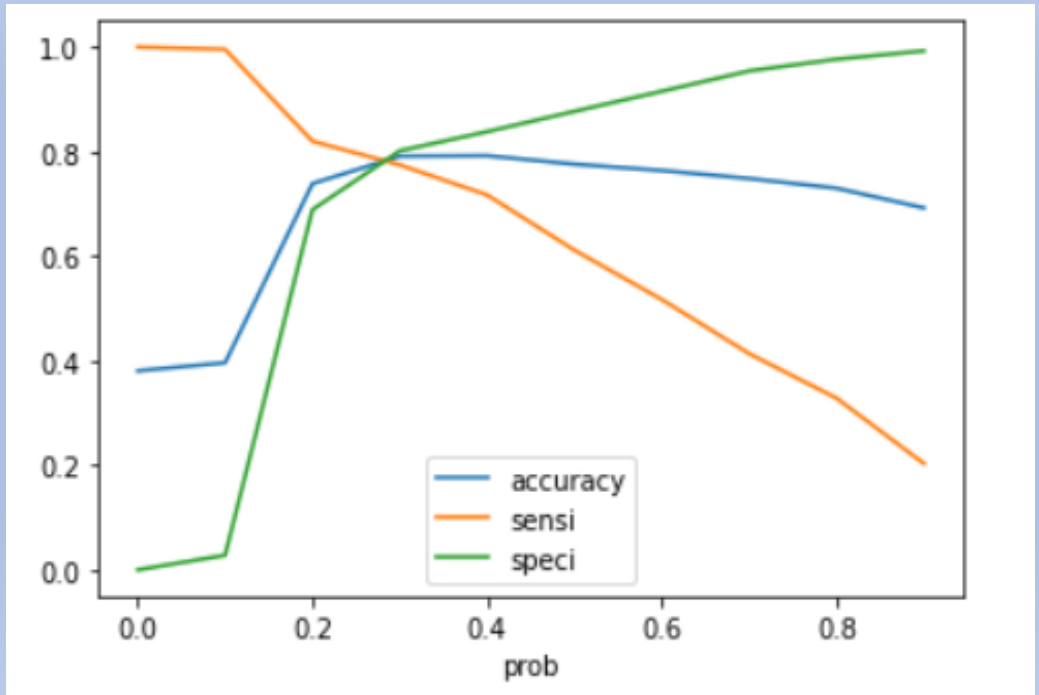
Correlation Matrix with Dummy Variables:



ROC Curve:

From the below graph we can find the optimum point to get the cut-off probability which is showing as .3 from the second graph





Conclusion:

- From the regression model and predictive analysis it is found that the variable which are mattered most in the potential buyer are listed below-
- Total Time spent on website
- Lead origin
- Lead source
- Top categorical/dummy variable which should focused more to get the high conversion of leads
- Lead Origin lead add form
- Current occupation working professional
- Total time spent on website
- Model Observation: After running the model on the Test Data these are the figures we obtain:
- Observation: Let us compare the values obtained for Train & Test:
- Train Data: Accuracy : 79.19% Sensitivity : 71.69% Specificity : 83.79% Test Data: Accuracy : 79.63% Sensitivity : 76.33% Specificity : 81.62% The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model

To optimize the lead conversion process, sales team should focus more on the highly potential leads which are predicted by the model.