Lead Score Case Study

Build a logistic regression model to assign a lead score

- Rahul Biswas
- Suman Prabhakar
- Sonal Padole

Problem Statement

- An education company named X Education sells online courses to industry professionals
- The company markets its courses on several websites, the company also gets leads through past referrals
- Leads are acquired through this process, 30 of the leads get converted while most do not
- 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
- In the middle stage, you need to nurture the potential leads well in order to get a higher lead conversion

Objective

- X Education has appointed you to help them The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well

Steps Taken to study Lead Scoring Case Study

Loading & Reading

- Read the csv into a pandas dataframe
- Understanding the descriptors of the variables using the dictionary provided.

Data Cleaning

- Address data types, missing values, remove unused columns and address outliers

Initiate EDA

- Check Data imbalance using the TARGET columns
- Define Categorical and Numerical column lists
- Univariate Categorical, Numeric Analysis
- Bivariate Categorical, Numeric Analysis
- Correlation heatmaps to see what relations would need further exploration

Logistic Regression Model

- Using logistic regression for model building and making predictions.

Data Cleaning

- Find the percentage of missing values of all the columns
- Remove columns with high missing percentage
- replacing the unique data with relevant category
- Checking and removing outliers
- Values that are not mentioned are replaced with other relevant values.

Steps Taken

- We converted select as NaN as they are as good as null value.
- Dropped *' Asymmetrique Activity Index',' Asymmetrique Profile Index', Asymmetrique Activity Score',' Asymmetrique Profile Score','Leads Quality ','Lead Profile', 'How did you hear about X Education'* as they have almost 40% or more missing values.
- Replacing the unique data with relevant category
- Addressed outliers
- Dropped columns - *‘City’, ‘What matters most to you in choosing a course’,'What is your current occupation', ‘Country’*
- Replaced with common categories in columns like -’Specialization’, ‘Tags’

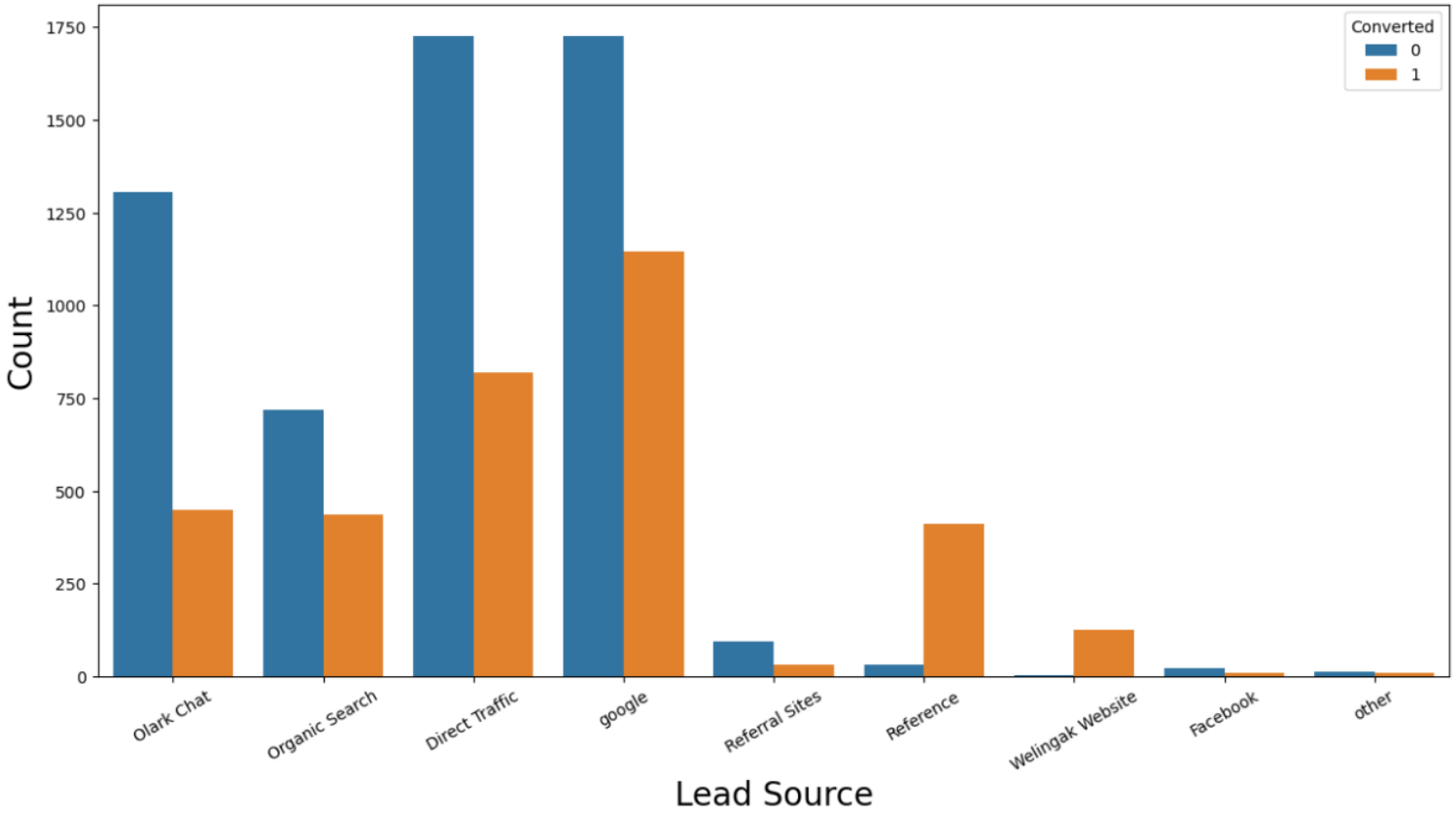
EDA

- Univariate Analysis
- Bivariate Analysis
- Categorical variable relation.

Visualization

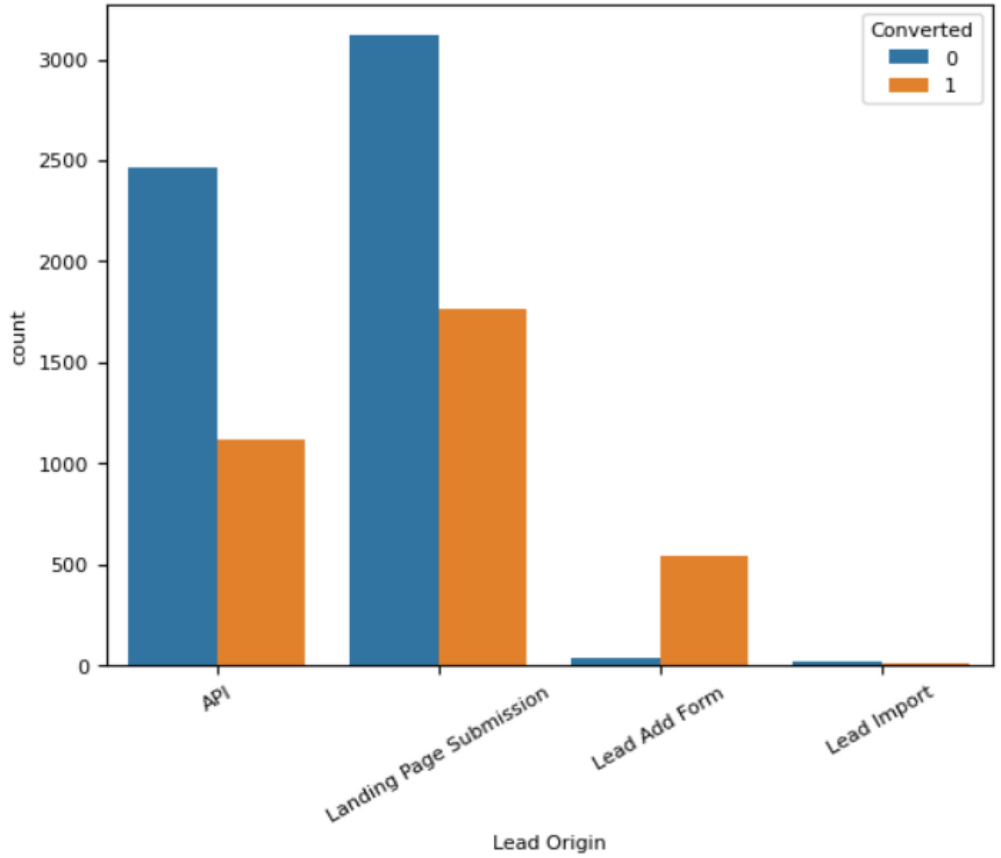
- Google & Direct traffic generates maximum number of the leads.
- Conversion rate of the welingak website and reference leads is high
- To improve the overall lead conversion rate, we should focus on the Organic
- Search,Olark Chat,Direct Traffic and google leads in the Lead Source and generates more leads.

Visualization - Lead Source



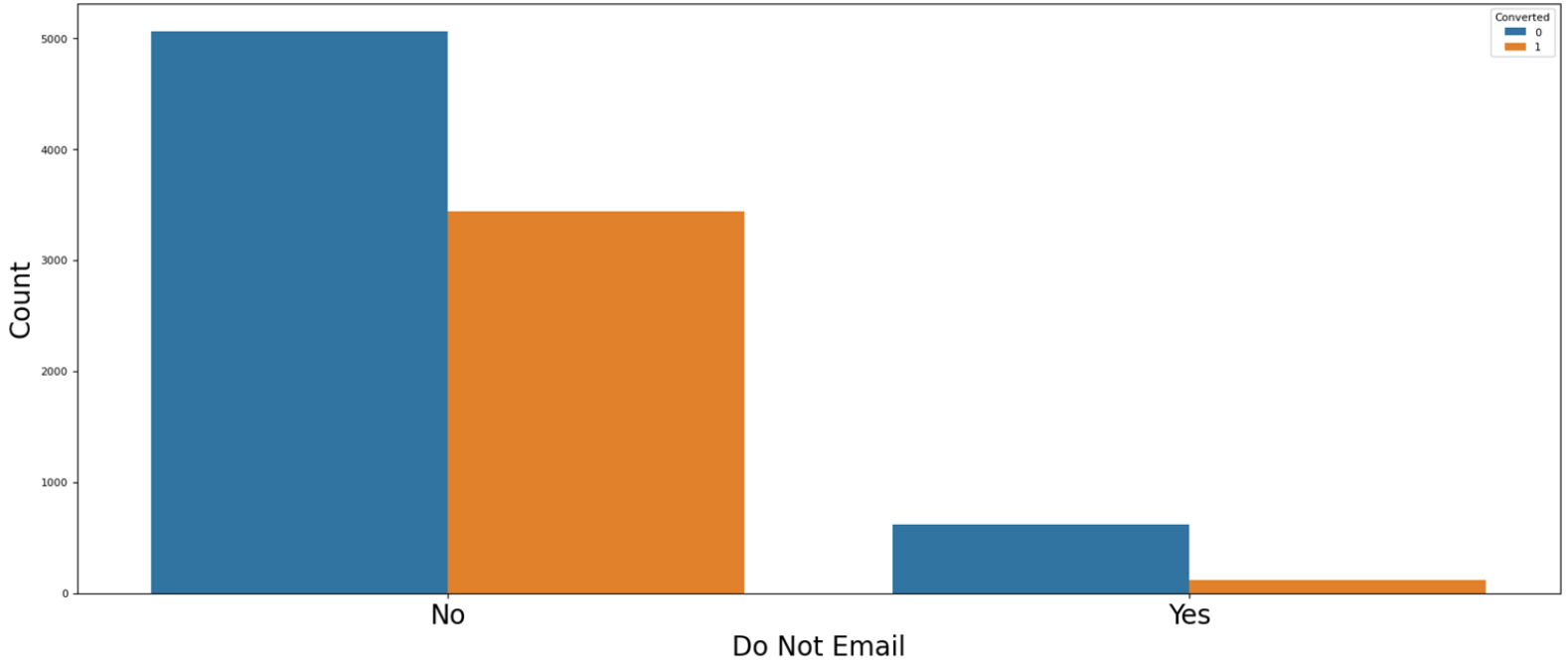
- Google & Direct traffic generates maximum number of the leads, having a good conversion rate.
- Conversion rate of the welingak website and reference leads is high
- To improve the overall lead conversion rate, we should focus on the Organic Search,Olark Chat,Direct Traffic and google leads in the Lead Source and generates more leads.
- Leads by reference is mostly converted and same goes with Welingak Website.

Visualization - Lead Origin



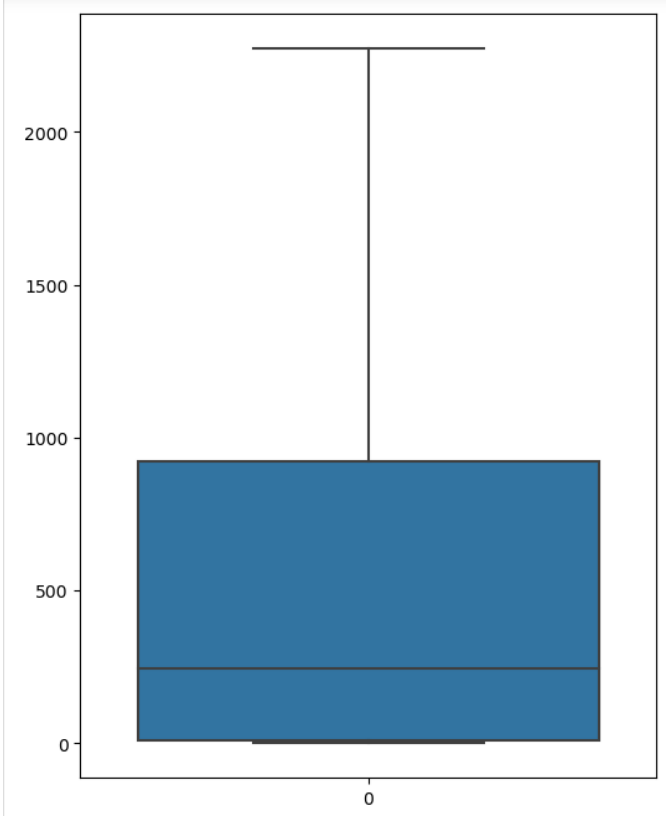
- API and Landing page show a good conversion rate and are good enough in number.
- While Lead Add form has high conversion rate but do not have enough counts.

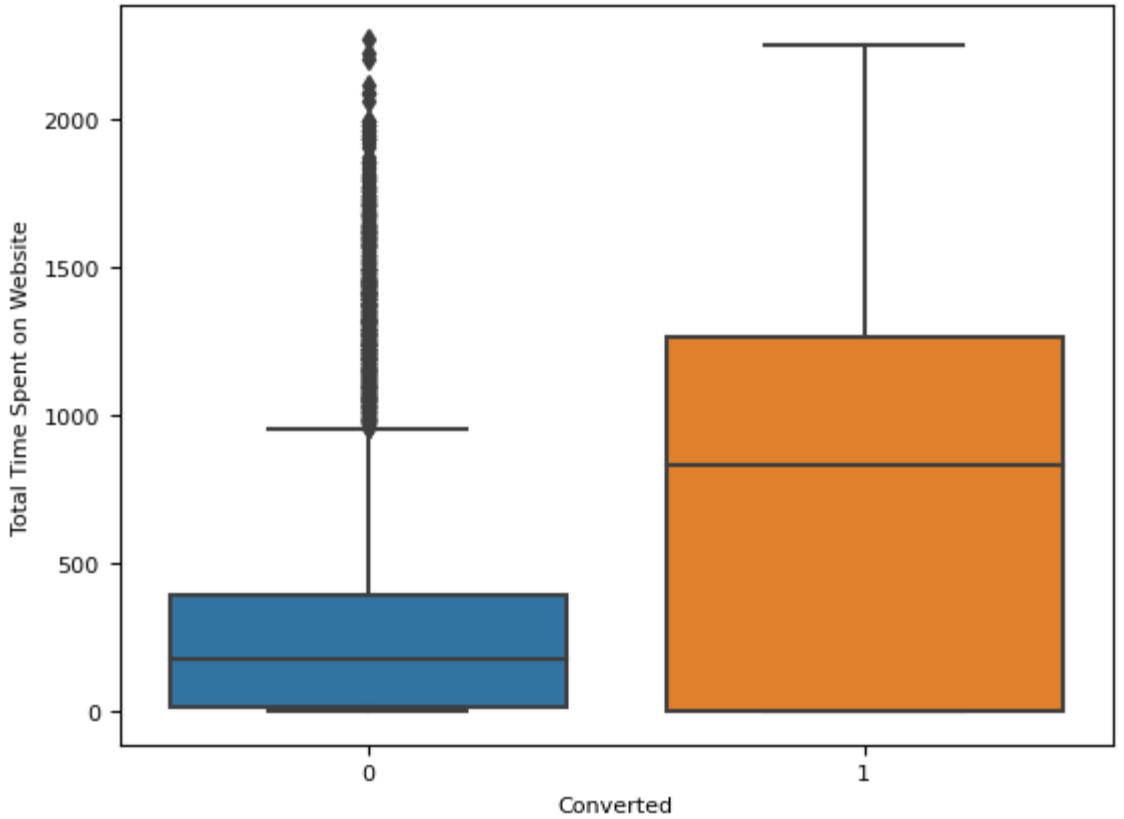
Visualization - Do Not Email



- Most customers don’t want to get emailed about the course but still get converted more as compared to people who choose to get emailed.

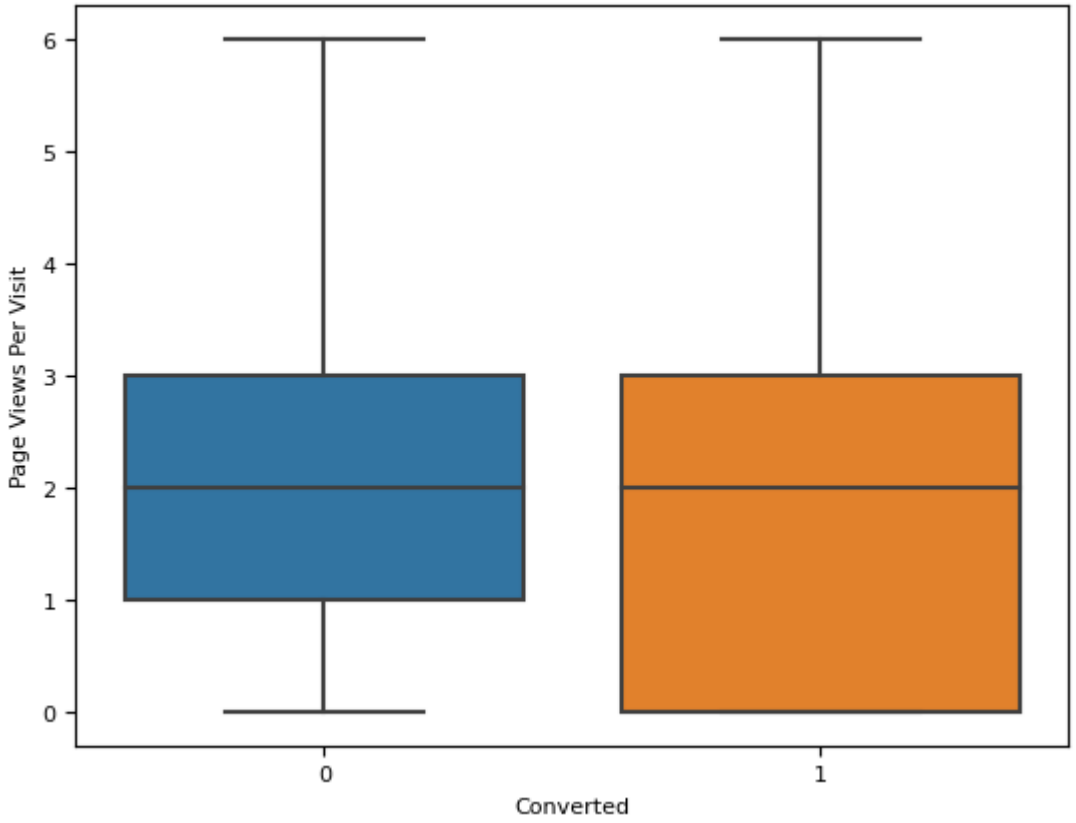
Visualization - Time Spent on Website





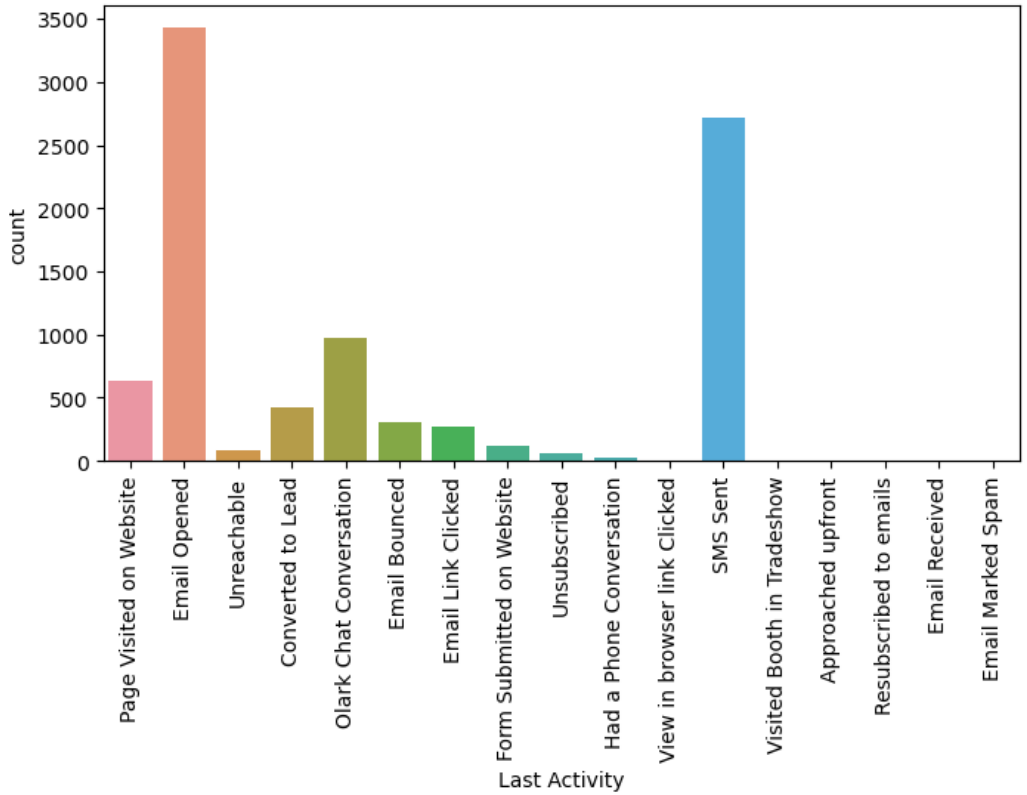
- The average time spent on website is around 250.
- We can see that when the total time spent by a person on website is around 1000 , the lead shows a positive conversion sign.

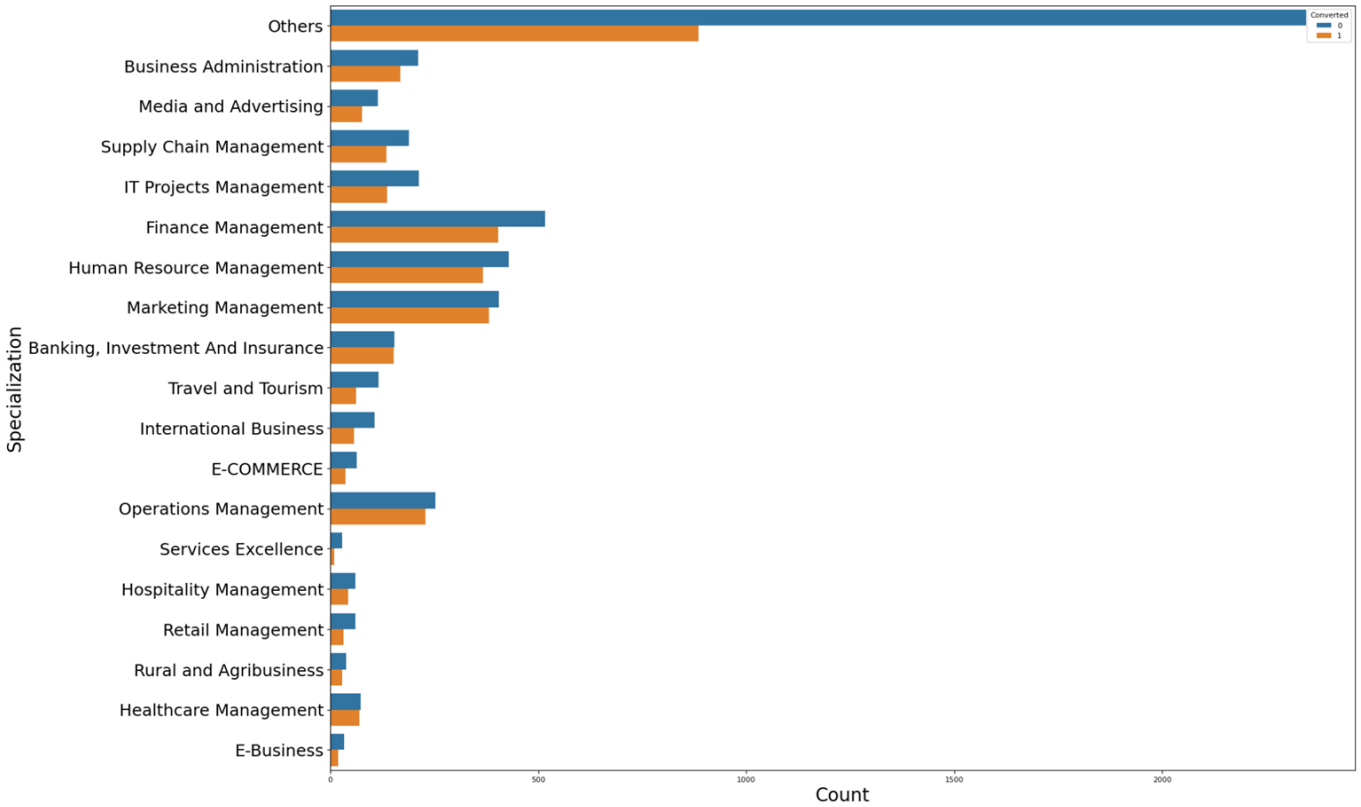
Visualization - Page Views per Visit



- There is no difference in median of the page views for converted and not converted leads.

Categorical Variable Relation

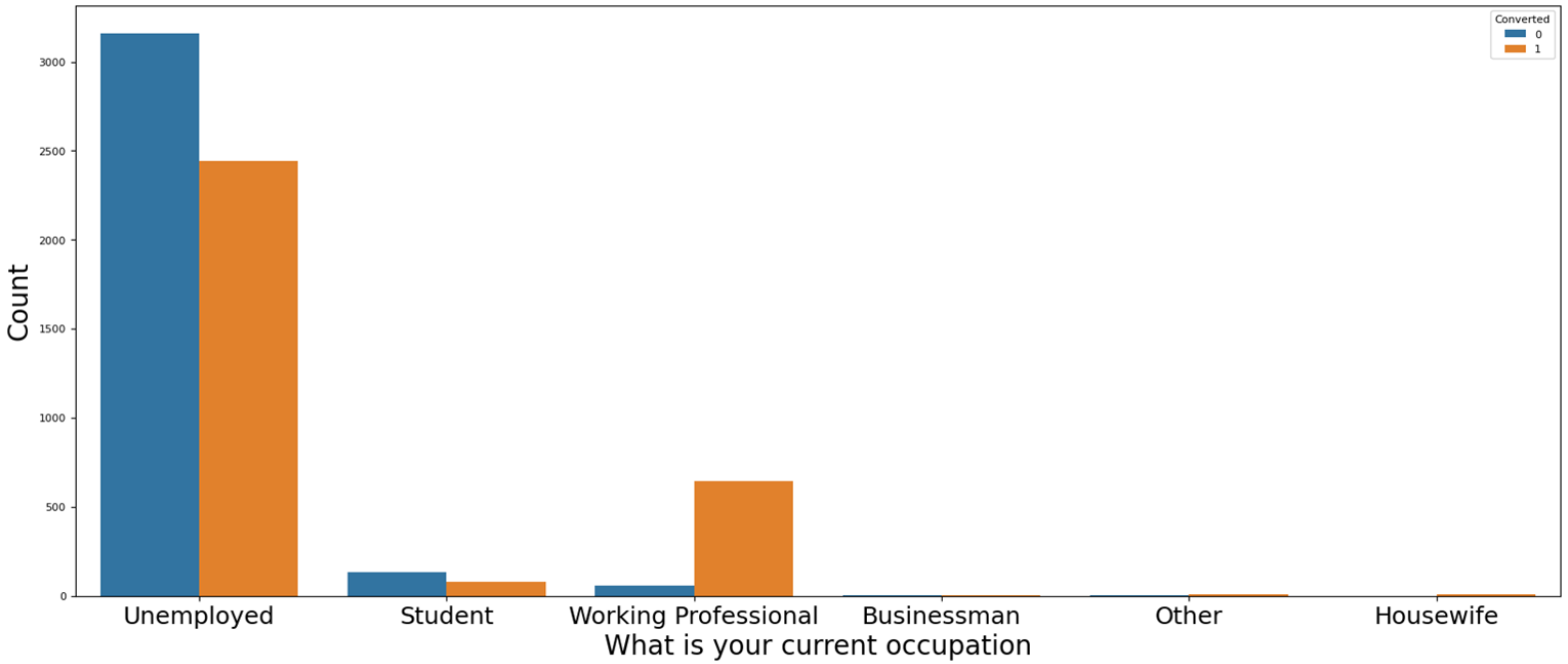




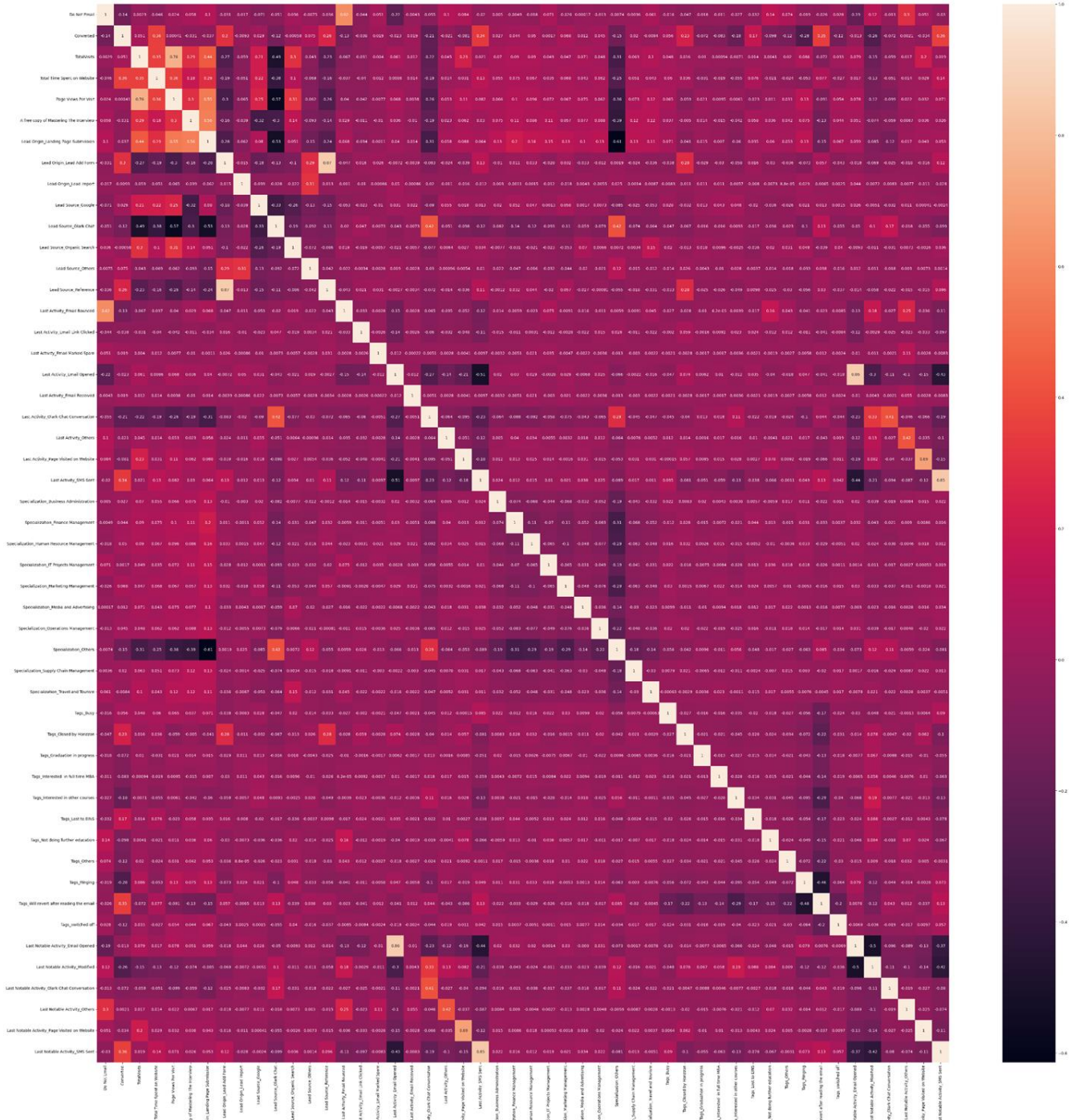
- Most of the leads are from email opened.
- SmS sent has the highest conversion rate among all of them

- As we can see that those who have not specified their industry domain in which they worked before have the highest conversion rate followed by Finance management ,HR management and marketing management.

Categorical Variable Relation



- Working Professionals have high conversion rate as compared to others.
- Most of the leads are generated by Unemployeds.



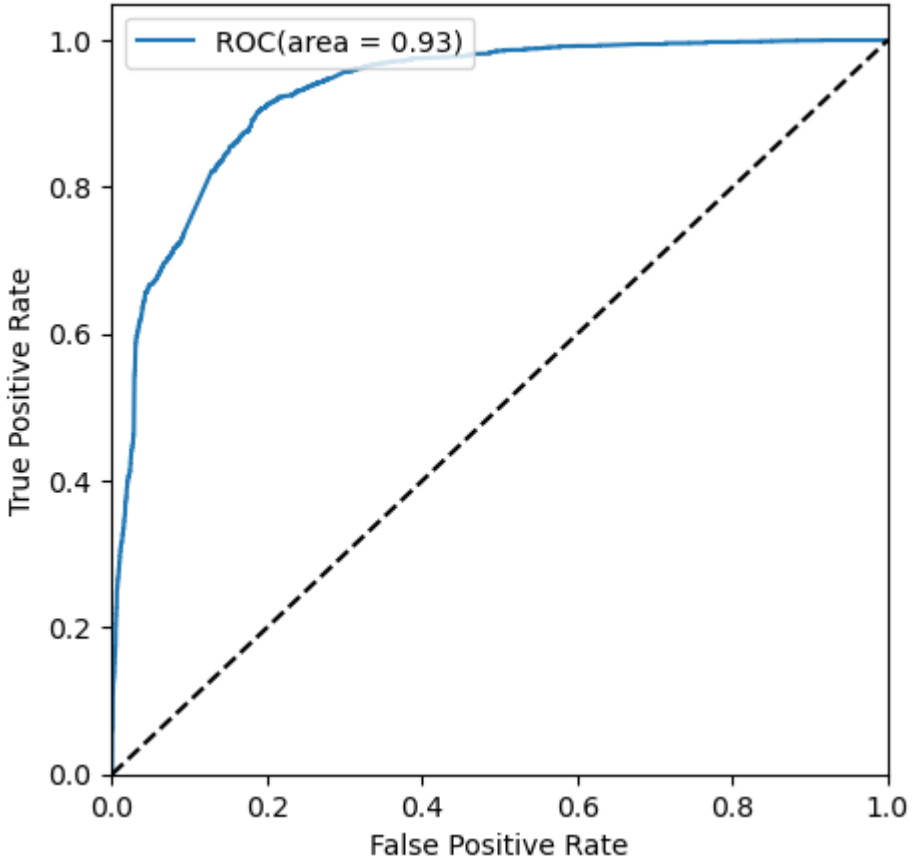
Heatmap after Encoding ,Scaling and creating dummy variables.

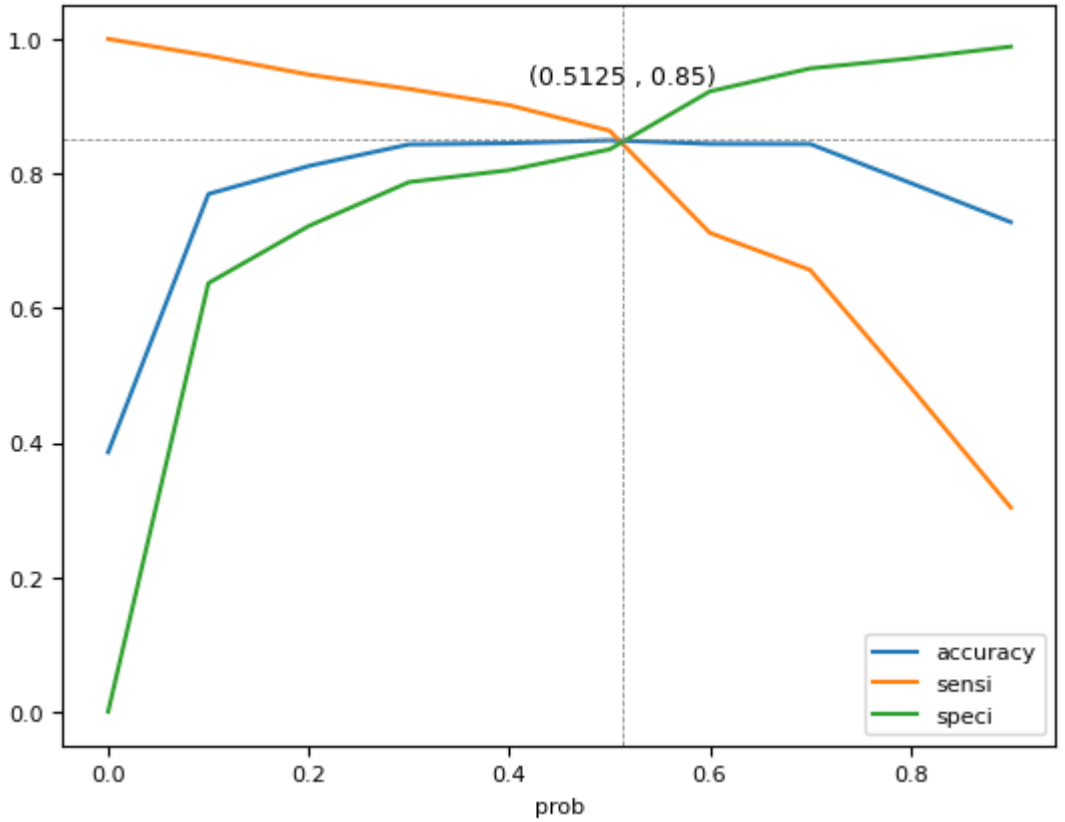
Since this heatmap is very crowded and the correlations are not perfectly visible so we refer to RFE.

Building Our Model

- Splitting the data into training and test sets.
- While performing test train split , we choose 70 : 30 ratio.
- Using RFE for feature selection.
- Running RFE with 20 variables as output.
- Building model by removing the variables whose p value is greater than 0.05 and VIF value is greater than 5.
- Predictions on test dataset.
- Creation of confusion matrix.
- Calculation of accuracy, sensitivity , specificity, precision and Recall.

Building Our Model





- The area under ROC curve is 0.93 , which is very good.
- The optimal cutoff probability from second graph is 0.512

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

- “true positive“ for correctly predicted event values.
- “false positive“ for incorrectly predicted event values.
- “true negative“ for correctly predicted no event values.
- “false negative“ for incorrectly predicted no event values.

```
Confusion Matrix
[[3263  660]
 [ 319 2130]]
```

- The Precision and Recall comes as 0.76 and 0.87 respectively.

Model Evaluation

Observations for Train - Test

Training Data:

- Accuracy: 84.64 % | Sensitivity: 86.97 % | Specificity: 83.18 %

Test Data:

- Accuracy: 84.64 % | Sensitivity: 86.97 % | Specificity: 83.18 %

Conclusion

Learning gathered are below:

- Test set has accuracy, recall/sensitivity all in an acceptable range.
- In business terms, our model is having stability and accuracy with adaptive environment skills which means it will adjust easily with the company's requirement changes made in the coming future.
- Top features for good conversion rate:
  - Tags_Closed by Horizon
  - Tags_Lost to EINS
  - Tags_Will revert after reading the email
- The top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion:
  - Tags,
  - Lead Source,
  - Lead Origin

Thank You