LEAD SCORING ASSIGNMENT (LOGISTIC REGRESSION)

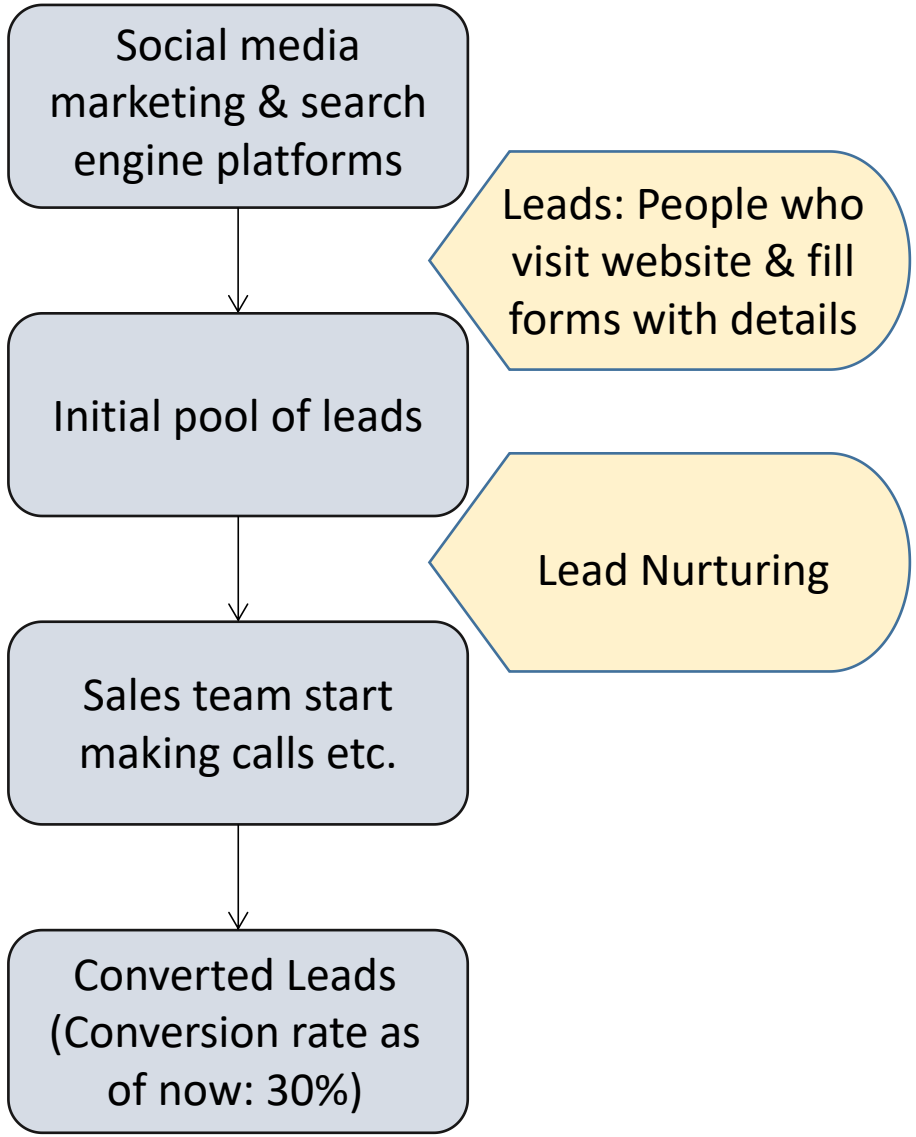
Prakhar Shrivastava

Arpit Vijay

(DS C22 Batch)

Business Understanding

- X Education, an education company, sells online courses to industry professionals.
- The company does it's major marketing campaigns through social media platforms and search engines.
- Once people visit their website and show interest in one of their courses, they are classified as leads and the sales teams work towards converting the leads.



Problem Statement

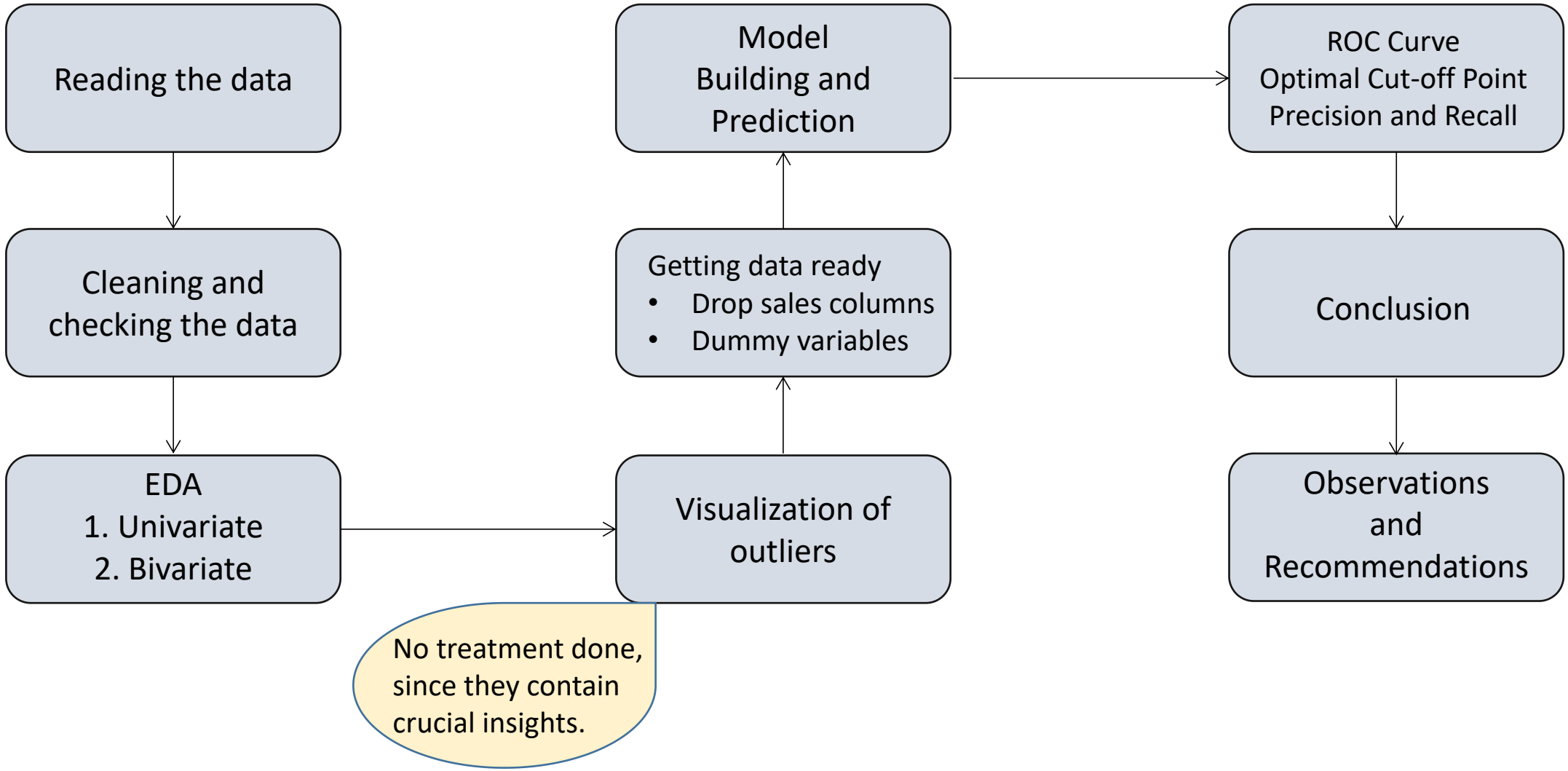
The leads obtained have a very poor conversion rate (30%). To improve the conversion rate, company needs to find out “hot leads”, so that the sales team can focus at the high potential leads rather than wasting time and resources on calling everyone.

Objective

The objective is to come up with a model that successfully assigns a “lead score” to each customer. Higher the lead score, higher the chances of conversion.

Business Objective: Improve the conversion from 30% to around 80%.

Approach towards the problem

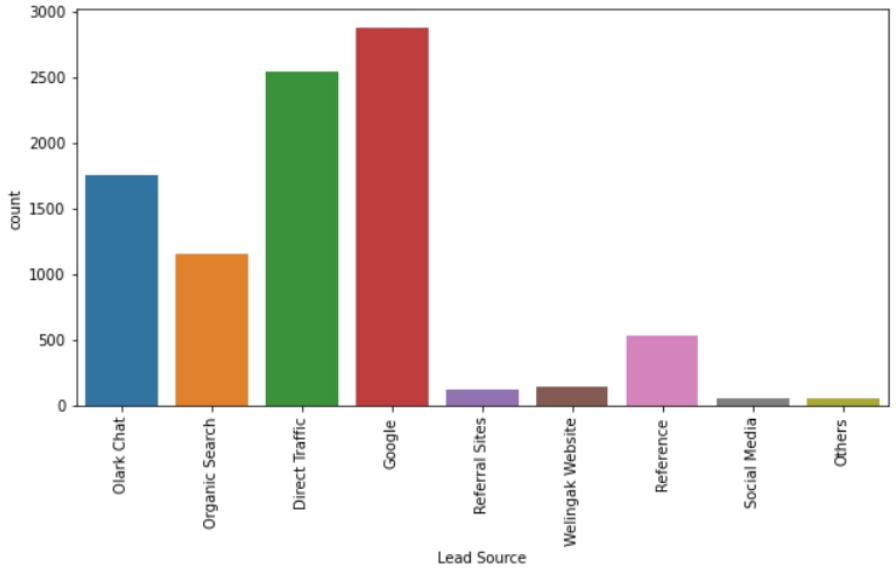


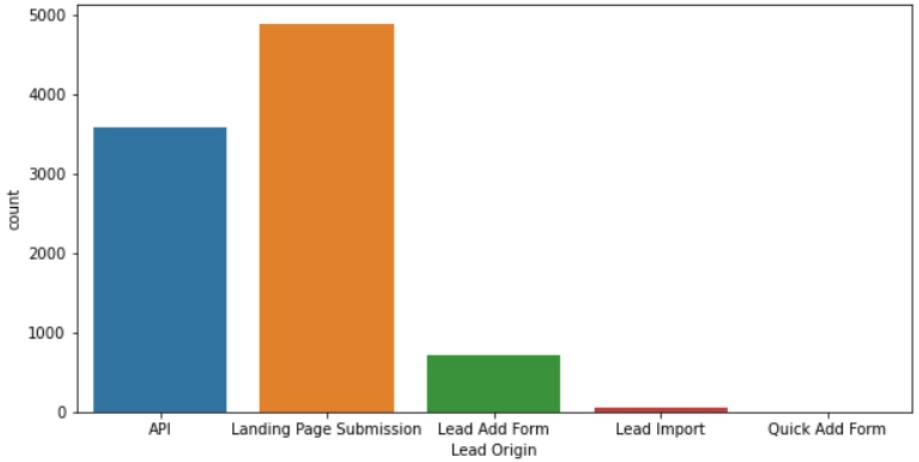
Cleaning data

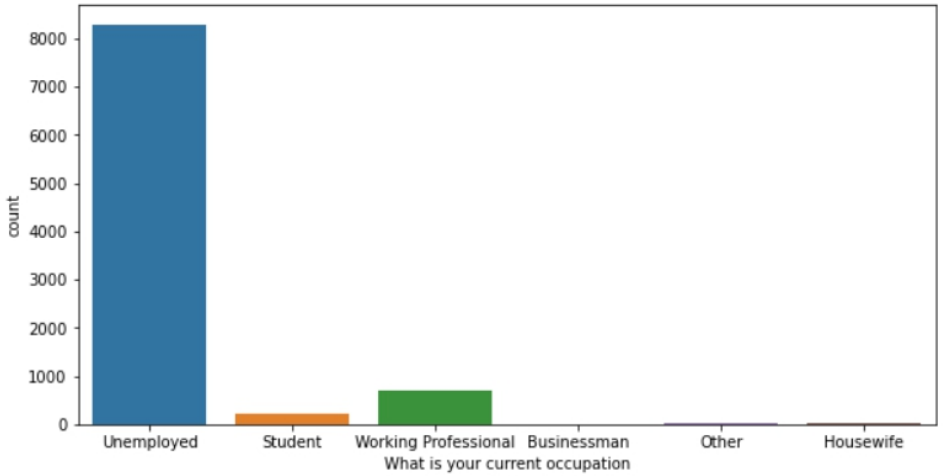
- Removed columns “Prospect ID” and “Lead Number”, since they aren’t useful in the modelling.
- Removed columns like “Last Activity”, “Tags” etc., since they are provided by the sales team.
- There is “Select” entry in a lot of fields, which is replaced by np.nan (Since we’re considering all such entries as null values).
- Getting rid of columns having more than 30% null values.
- Imputing null values of other columns with mean, mode etc.
- Dropping columns that are unbalaced and skewed (when the whole column is majorly represented by only one value).
- Cleaning other important columns like “Lead Source”, by grouping a lot of low frequency values under “Others” category, to avoid huge number of dummy variables.

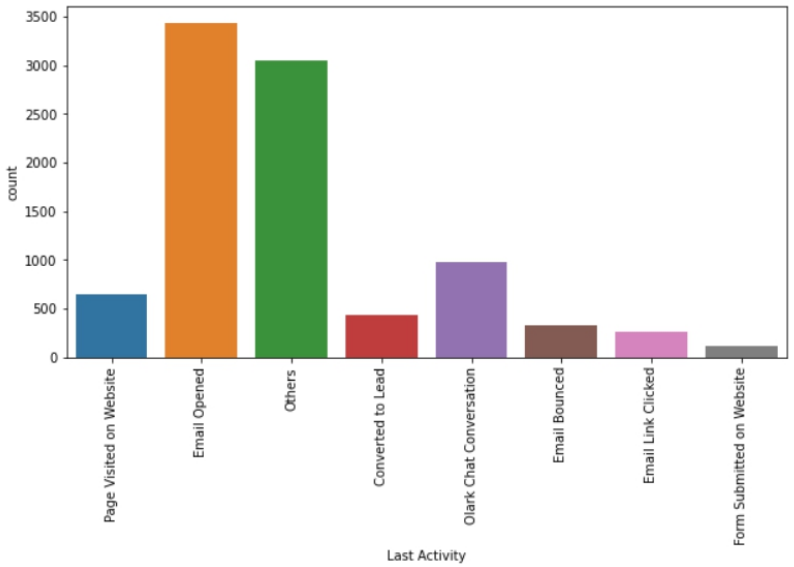
Exploratory Data Analysis

- Univariate
  - Most leads obtained from Google search.
  - Most leads originated from Landing Page Submission.
  - Most leads are people who are currently unemployed.
  - Most leads prefer emails as their point of contact.



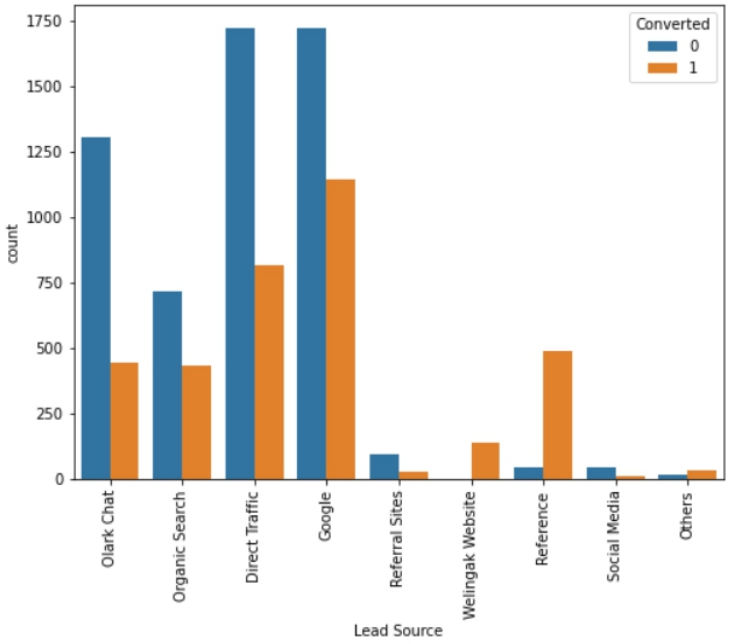


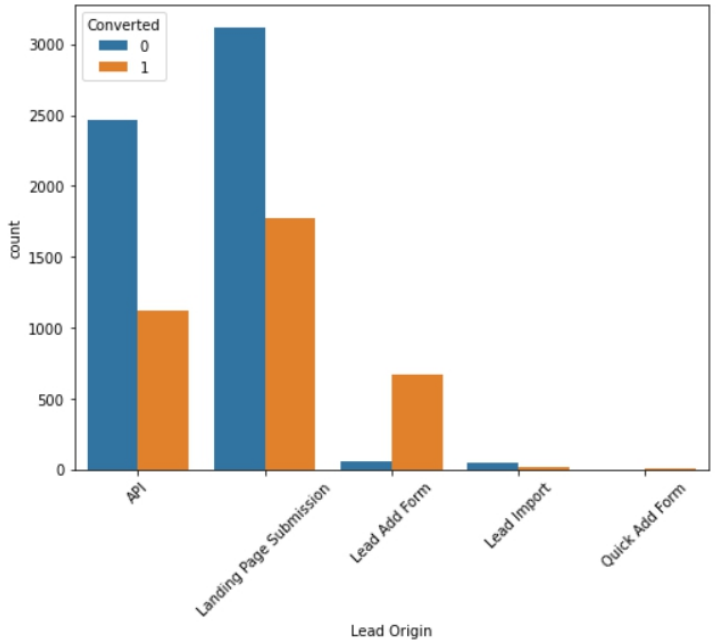


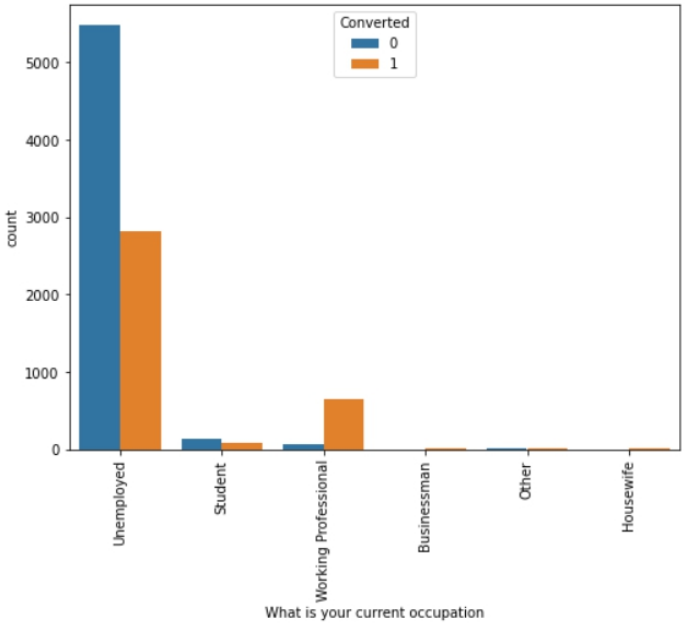


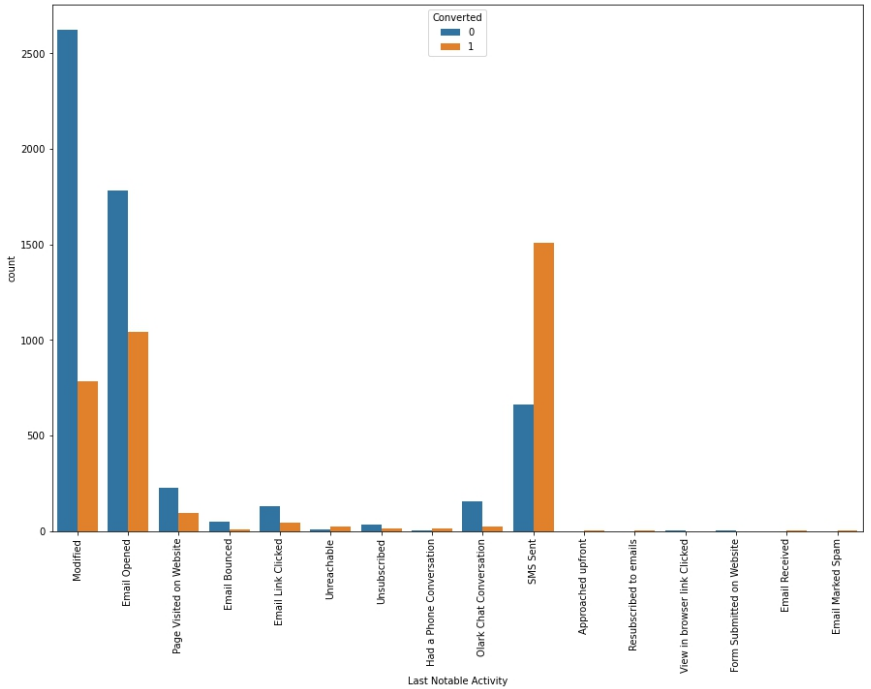
- Bivariate

- People who are approached on direct messages seem to buy the courses on a higher rate.
- For people approaching through Google search, references and Orgnic search, there is a high chance that they will opt for a course
- It is evident, that people who are filling the forms and providing their details are highly probable to buy a course.
- Leads who are working professionals show a higher chance of conversion.

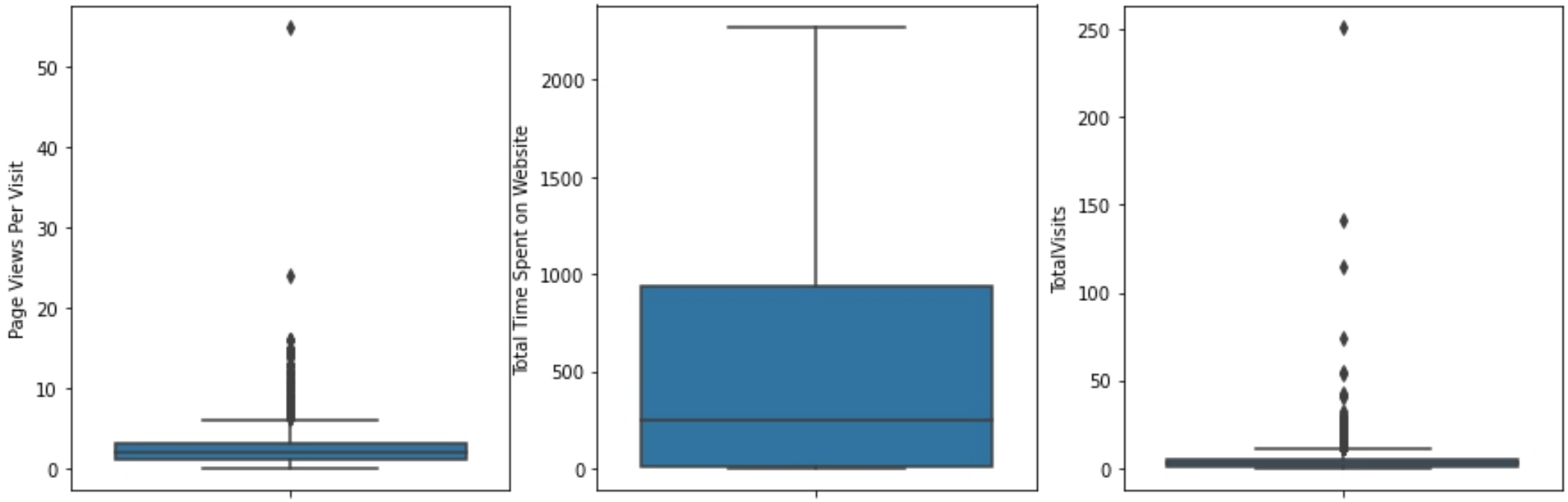








Visualizing outliers



Here, we can see there are outliers in the numerical variables, but we won’t treat them since they carry important insights in this case.

Logistic Regression Model Building

- Once EDA was done, we built Logistic regression model using **GLM()** function, under statsmodel library.
- As we can see from the result, a lot of variables were insignificant (as per P values).
- These variables were removed using -

  (1) Automated approach (RFE)

  (2) Manual elimination

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.4205	1.008	-1.410	0.159	-3.395	0.554
TotalVisits	0.1171	0.044	2.681	0.007	0.031	0.203
Total Time Spent on Website	1.1004	0.037	29.774	0.000	1.028	1.173
Page Views Per Visit	-0.0182	0.044	-0.414	0.679	-0.104	0.068
rigin_Landing Page Submission	0.0020	0.092	0.022	0.983	-0.178	0.182
Lead Origin_Lead Add Form	3.0671	0.749	4.097	0.000	1.600	4.534
Lead Origin_Lead Import	-0.5748	1.297	-0.443	0.658	-3.116	1.967
Lead Origin_Quick Add Form	20.0799	4.82e+04	0.000	1.000	-9.44e+04	9.45e+04
Lead Source_Google	0.3950	0.089	4.441	0.000	0.221	0.569
Lead Source_Olark Chat	1.0614	0.138	7.689	0.000	0.791	1.332
Lead Source_Organic Search	0.2277	0.117	1.948	0.051	-0.001	0.457
Lead Source_Others	0.3962	0.646	0.613	0.540	-0.870	1.662
Lead Source_Reference	1.0273	0.769	1.335	0.182	-0.481	2.535
Lead Source_Referral Sites	-0.2740	0.306	-0.896	0.370	-0.873	0.325
Lead Source_Social Media	0.6669	1.208	0.552	0.581	-1.700	3.034
Lead Source_Welingak Website	3.0059	1.034	2.907	0.004	0.979	5.033
r current occupation_Housewife	22.7085	1.35e+04	0.002	0.999	-2.64e+04	2.65e+04
s your current occupation_Other	-0.4392	1.274	-0.345	0.730	-2.936	2.058
our current occupation_Student	0.2346	1.026	0.229	0.819	-1.776	2.246
urrent occupation_Unemployed	0.0216	1.005	0.022	0.983	-1.948	1.991
cupation_Working Professional	2.9273	1.020	2.870	0.004	0.928	4.926
ourse_Flexibility & Convenience	-1.9085	3.780	-0.505	0.614	-9.318	5.501

Final Model

- As we can see, all variables are now significant and VIF < 3.

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.5344	0.174	8.838	0.000	1.194	1.875
Total Time Spent on Website	1.1009	0.037	29.935	0.000	1.029	1.173
Lead Origin_Lead Add Form	3.8597	0.185	20.874	0.000	3.497	4.222
Lead Source_Google	0.3151	0.073	4.310	0.000	0.172	0.458
Lead Source_Olark Chat	0.9165	0.097	9.471	0.000	0.727	1.106
Lead Source_Welingak Website	2.0659	0.737	2.805	0.005	0.622	3.509
What is your current occupation_Other	-3.2455	0.808	-4.017	0.000	-4.829	-1.662
What is your current occupation_Student	-2.6378	0.269	-9.789	0.000	-3.166	-2.110
What is your current occupation_Unemployed	-2.8399	0.174	-16.364	0.000	-3.180	-2.500

	Features	VIF
7	What is your current occupation_Unemployed	2.15
2	Lead Source_Google	1.66
3	Lead Source_Olark Chat	1.59
1	Lead Origin_Lead Add Form	1.32
4	Lead Source_Welingak Website	1.22
0	Total Time Spent on Website	1.20
6	What is your current occupation_Student	1.03
5	What is your current occupation_Other	1.00

Prediction

- “Conversion_Probability” has probability (of getting a lead converted), as calculated by our model.
- With initial cut-off value of 0.5 (to be calculated further), we have got the “Predicted” column.

	Converted	Conversion_Probability	ID	Predicted
0	0	0.203646	1871	0
1	0	0.214303	6795	0
2	0	0.240464	3516	0
3	0	0.625416	8105	1
4	0	0.203646	3934	0
5	1	0.973019	4844	1
6	0	0.108228	3297	0
7	1	0.993978	8071	1
8	0	0.345594	987	0
9	1	0.765558	7423	1

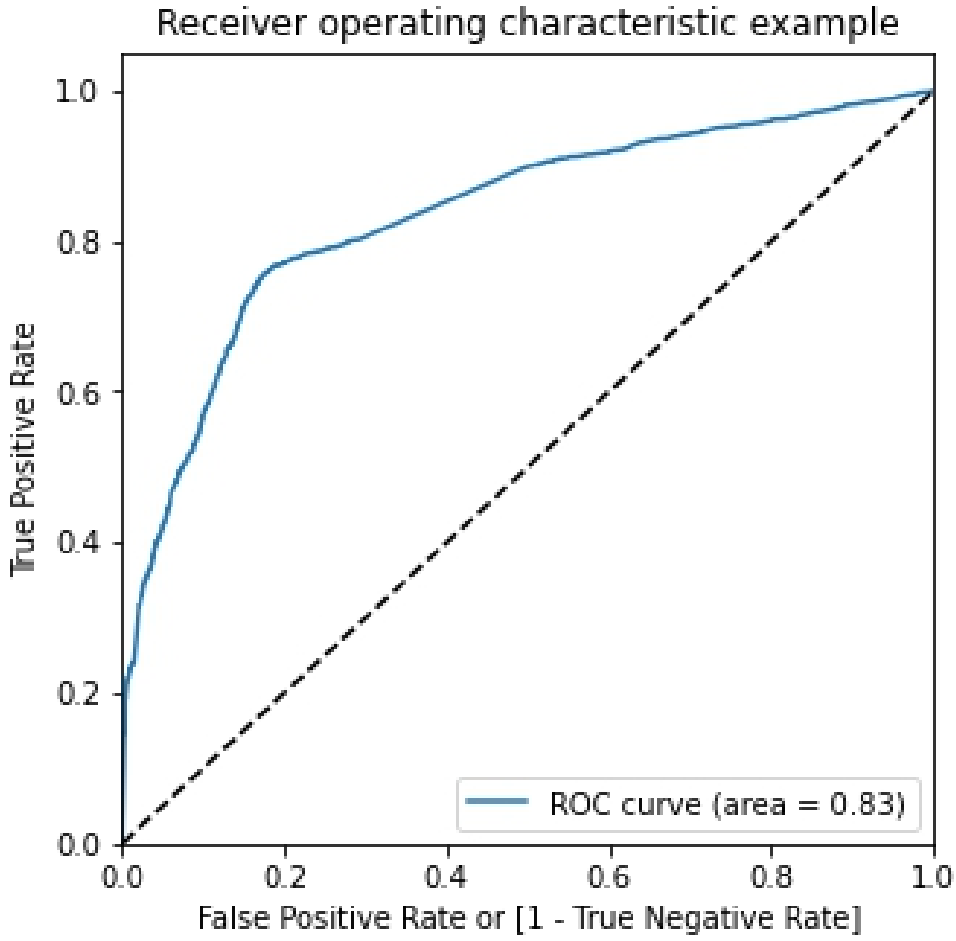
ROC Curve

This curve shows-

- Tradeoff between sensitivity and specificity
- Closer the curve is to the left and the upper border, more accurate is the test.
- Closer the curve to the line, lesser is the accuracy

In our model, the ROC curve is inline with the above 3 points mentioned.

Hence, we can use this model for our further analysis.

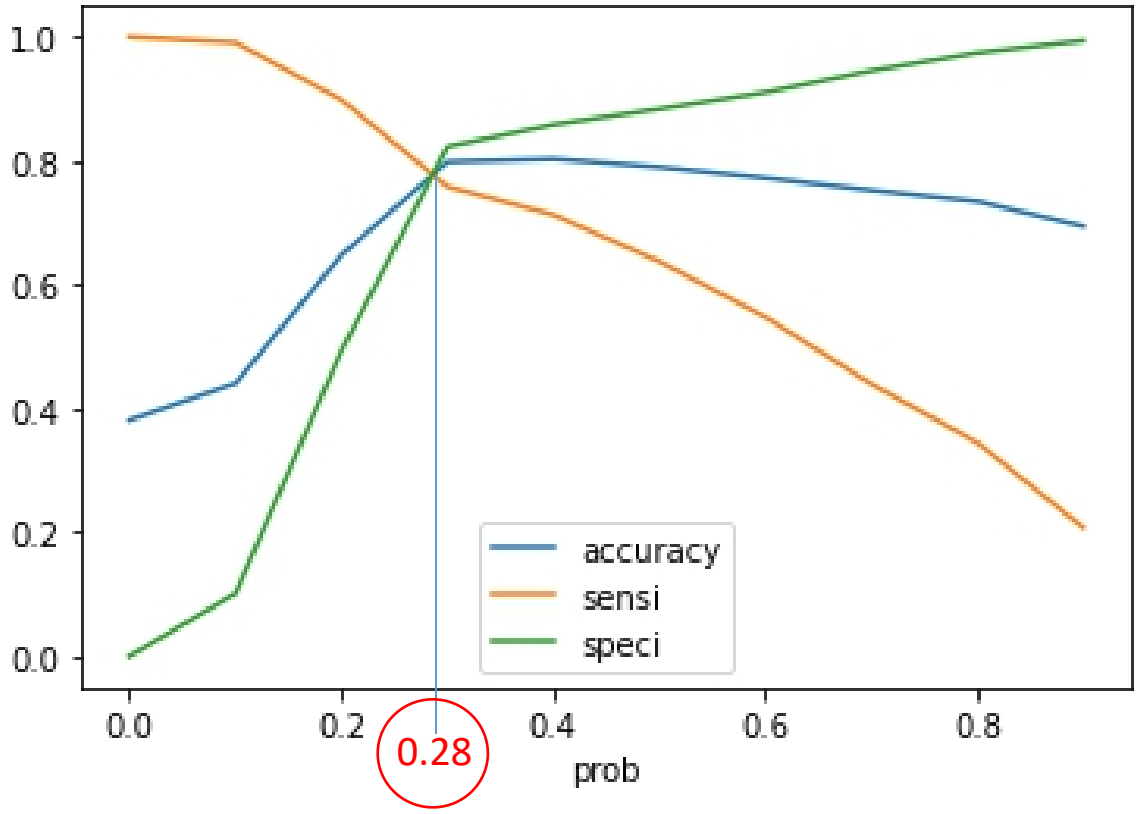


Optimal Cut-off Point

On plotting accuracy, sensitivity and specificity for various probabilities, we get the Cut-off point ~ 0.28.

This cut-off point results in:

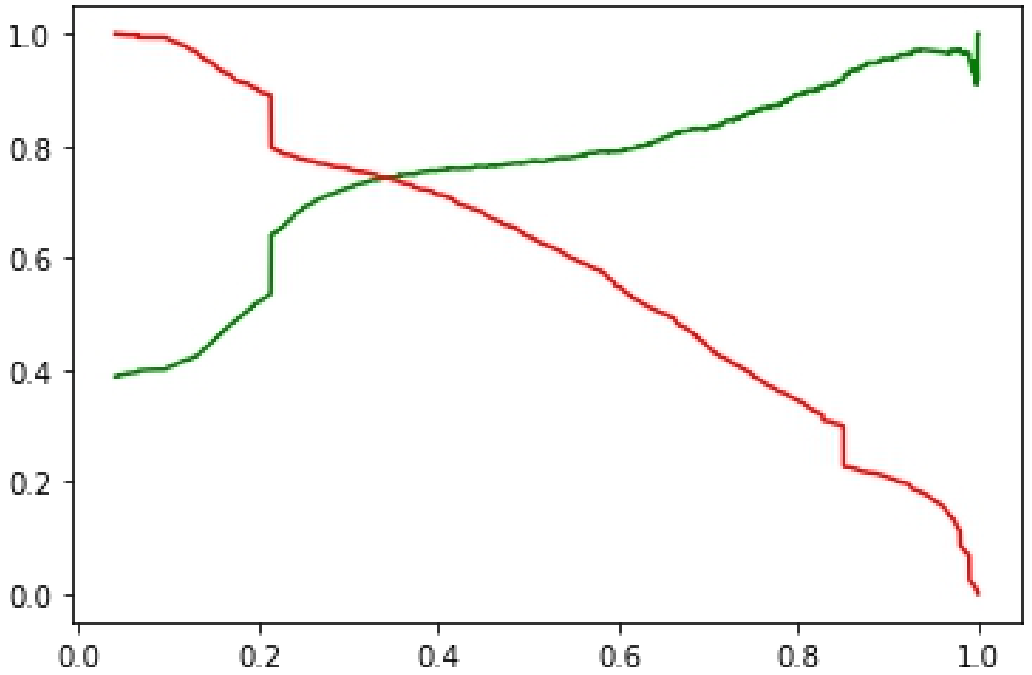
- Accuracy Score %= 79.26
- Sensitivity %= 76.39
- Specificity %= 81.03
- False Positive Rate %= 18.96
- Positive Predictive Rate %= 71.28
- Negative Predictive Rate %= 84.78
- Precision %= 77.13
- Recall %= 63.74



Model Evaluation: Precision and Recall

- This graph shows a trade-off between Precision and Recall.
- We have chosen the cut-off value of 0.28 which gives us perfectly balanced values of:

  - Accuracy Score %= 79.26
  - Sensitivity %= 76.39
  - Specificity %= 81.03
  - False Positive Rate %= 18.96
  - Positive Predictive Rate %= 71.28
  - Negative Predictive Rate %= 84.78
  - Precision %= 77.13
  - Recall %= 63.74



Conclusion

- Our model has accuracy up to 79% on train data set and 79% on test data set.
- **We can now confidently say that the lead conversion will be around 80% as required by the CEO of the company.**
- Columns: "final_predicted", "converted", "Conversion_Probability" are used to identify hot leads and cold leads.
- Cut off value used: **0.28**

- Any probability below 0.28 is considered to have a lower chance of getting converted (Cold Lead).

- Any probability above 0.28 is considered to have a higher chance of getting converted (Hot Lead).

Final Model Statistics (Train Set):

- Accuracy Score %= 79.26
- Sensitivity %= 76.39
- Specificity %= 81.03
- False Positive Rate %= 18.96
- Positive Predictive Rate %= 71.28
- Negative Predictive Rate %= 84.78
- Precision %= 77.13
- Recall %= 63.74

Final Model Statistic (Test Set):

- Accuracy Score %= 79
- Sensitivity %= 75.98
- Specificity %= 80.97
- False Positive Rate %= 19.02
- Positive Predictive Rate %= 72.28
- Negative Predictive Rate %= 83.77
- Precision %= 77.13

Observations and Recommendations

- Observations -

Top 3 variables:

- Lead Origin
- Current Occupation
- Lead Source

Top 3 categorical variables:

- Lead Origin_Lead Add Form
- Current occupation_Working Professional
- Lead Source_Welingak Website

	Features	Coeff
3	Lead Origin_Lead Add Form	3.9665
8	What is your current occupation_Working Profes...	2.8671
7	Lead Source_Welingak Website	2.1256
2	Total Time Spent on Website	1.0973
5	Lead Source_Olark Chat	0.9261
4	Lead Source_Google	0.3319
6	Lead Source_Organic Search	0.2409
9	Do Not Email_Yes	-1.3498

Other observations:

- People who fill the form with details have a very high positive correlation with conversion. (coeff = 3.96)
- Working professionals are highly probable to buy a course. (coeff = 2.86)
- Welingak Website is a good source to get leads. (coeff = 2.12)
- People who do not want to receive emails are not to be targetted as they won't be interested. (coeff = -1.34)

- Recommendations -

- Consider **Lead Score > 75** as a high score.
- Target only the leads having a **score more than 75**.
- Leads with a score of 75-85 should be given to **more experienced and senior sales employees**.
- Leads with score of **more than 85** can be given to lesser experienced or junior employees, since the chances of conversion is high.

Thank you!