LEAD SCORING CASE STUDY

-Aishwarya Bankar

Problem Statement

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

Goals

- 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. The model should be able to adjust to if the company's requirement changes in the future

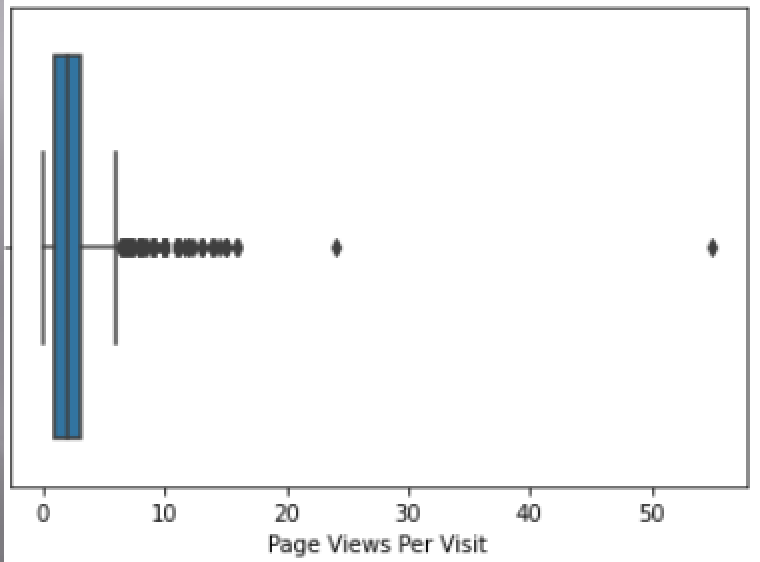
Steps followed

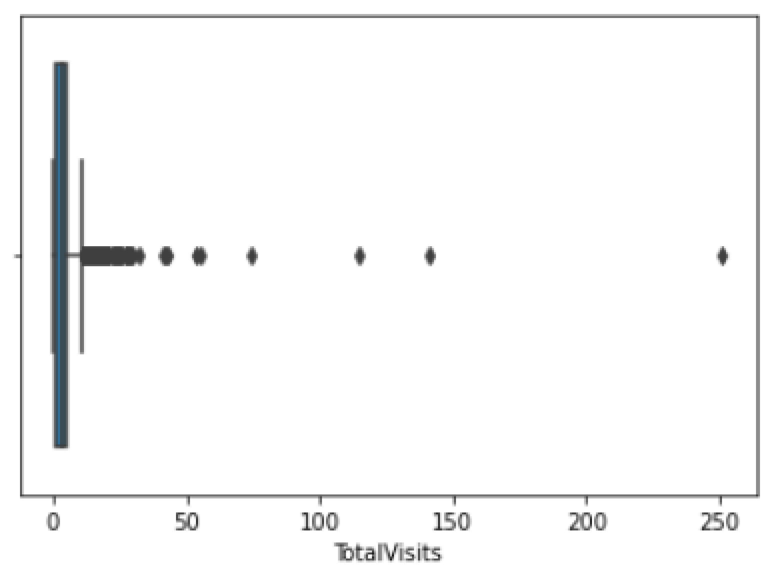
1. Importing the necessary libraries and reading the dataset.
2. Understanding the data to gain some insights
3. Performing univariate analysis to check for any outliers
4. Data cleaning
5. Splitting the data into train and test set
6. Feature standardization to scale the features
7. Removing highly correlated variables
8. Model building using RFE the manual tuning of model based on p-value and VIF
9. Model evaluation
10. Conclusion

Outlier detection/treatment

•Outliers present in two columns 'Page Views Per Visit & 'TotalVisits'

•We created bins for these columns as these outliers may affect our model prediction.



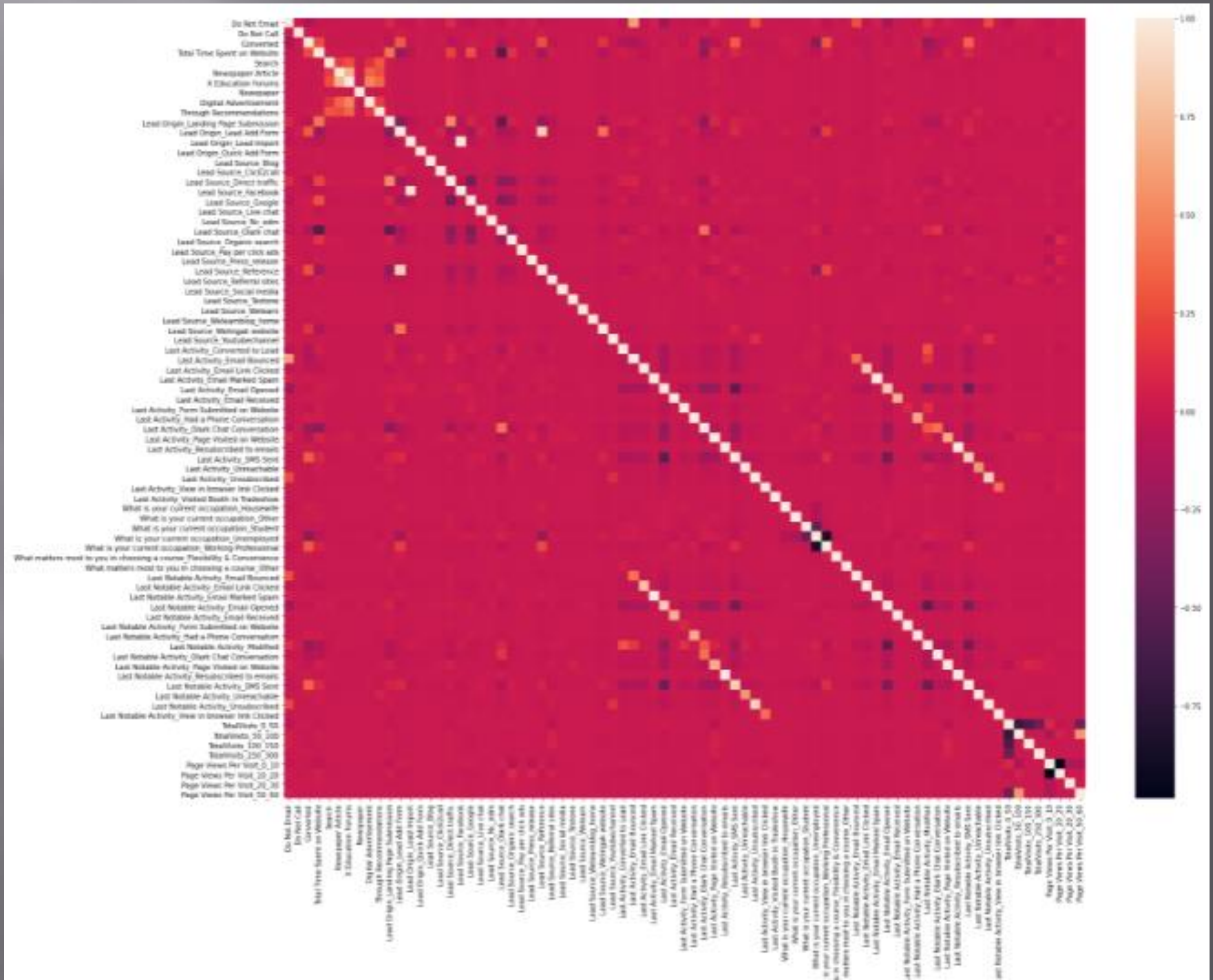


Data Cleaning

- Replacing the value ‘Select’ with null as ‘select’ means that no value was selected by the user
- Check the columns with null percentage below 30 and impute them with their mode.
- Checked the duplicate values ,column Lead Source has duplicate values 'google' & 'Google‘. So we will capitalized the values.
- Creating dummy variables for categorical columns

Correlation Matrix

There are two variables having high correlation namely 'Lead Source_Olark chat','What is your current occupation_Unemployed', so we going to drop them



Building a model using RFE

- From the sklearn library we have used logistic regression to solve this problem as it is a classification problem.
- We used RFE to select top 19 features as follows

```
col
```

```
Index(['Do Not Email', 'Total Time Spent on Website',
       'Lead Origin_Lead Add Form', 'Lead Source_Direct traffic',
       'Lead Source_Facebook', 'Lead Source_Google',
       'Lead Source_Organic search', 'Lead Source_Referral sites',
       'Lead Source_Welingak website', 'Last Activity_Converted to Lead',
       'Last Activity_Email Bounced', 'Last Activity_Had a Phone Conversation'
       'Last Activity_Olark Chat Conversation',
       'What is your current occupation_Housewife',
       'What is your current occupation_Working Professional',
       'Last Notable Activity_Email Bounced',
       'Last Notable Activity_Had a Phone Conversation',
       'Last Notable Activity_SMS Sent', 'Last Notable Activity_Unreachable'],
      dtype='object')
```

- Now that we have selected the columns with RFE , we will use stats model to add/remove features.
- We use the GLM model ,below is the summary.

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6448
Model Family:	Binomial	Df Model:	19
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2686.9
Date:	Sun, 10 Oct 2021	Deviance:	5373.7
Time:	17:24:03	Pearson chi2:	7.60e+03
No. Iterations:	21		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.2700	0.088	-3.066	0.002	-0.443	-0.097
Do Not Email	-1.0697	0.189	-5.654	0.000	-1.441	-0.699
Total Time Spent on Website	1.0855	0.039	27.771	0.000	1.009	1.162
Lead Origin_Lead Add Form	2.6685	0.196	13.645	0.000	2.285	3.052
Lead Source_Direct traffic	-1.3295	0.115	-11.608	0.000	-1.554	-1.105
Lead Source_Facebook	-1.2205	0.523	-2.336	0.019	-2.245	-0.196
Lead Source_Google	-0.9459	0.107	-8.810	0.000	-1.156	-0.735
Lead Source_Organic search	-1.1054	0.133	-8.314	0.000	-1.366	-0.845
Lead Source_Referral sites	-1.1998	0.314	-3.825	0.000	-1.815	-0.585
Lead Source_Welingak website	1.8182	0.743	2.447	0.014	0.362	3.275
Last Activity_Converted to Lead	-1.2310	0.217	-5.664	0.000	-1.657	-0.805
Last Activity_Email Bounced	-1.4848	0.419	-3.547	0.000	-2.305	-0.664
Last Activity_Had a Phone Conversation	0.4273	0.950	0.450	0.653	-1.434	2.288
Last Activity_Olark Chat Conversation	-1.3968	0.163	-8.578	0.000	-1.716	-1.078
What is your current occupation_Housewife	22.9006	1.36e+04	0.002	0.999	-2.67e+04	2.67e+04
What is your current occupation_Working Professional	2.8077	0.188	14.943	0.000	2.439	3.176
Last Notable Activity_Email Bounced	1.7415	0.602	2.894	0.004	0.562	2.921
Last Notable Activity_Had a Phone Conversation	3.1106	1.456	2.137	0.033	0.257	5.964
Last Notable Activity_SMS Sent	1.4748	0.079	18.603	0.000	1.319	1.630
Last Notable Activity_Unreachable	1.7651	0.518	3.410	0.001	0.751	2.780

	Features	VIF
11	Last Activity_Had a Phone Conversation	2.02
16	Last Notable Activity_Had a Phone Conversation	2.01
10	Last Activity_Email Bounced	1.94
0	Do Not Email	1.84
2	Lead Origin_Lead Add Form	1.41
17	Last Notable Activity_SMS Sent	1.38
3	Lead Source_Direct traffic	1.26
5	Lead Source_Google	1.25
8	Lead Source_Welingak website	1.24
15	Last Notable Activity_Email Bounced	1.21
14	What is your current occupation_Working Profes...	1.18
1	Total Time Spent on Website	1.16
6	Lead Source_Organic search	1.12
9	Last Activity_Converted to Lead	1.10
12	Last Activity_Olark Chat Conversation	1.08
13	What is your current occupation_Housewife	1.01
7	Lead Source_Referral sites	1.01
18	Last Notable Activity_Unreachable	1.01
4	Lead Source_Facebook	1.00

- As the VIF of all variables is below 5 , we will drop features with high p-value one by one.
- Below is the final model we get.

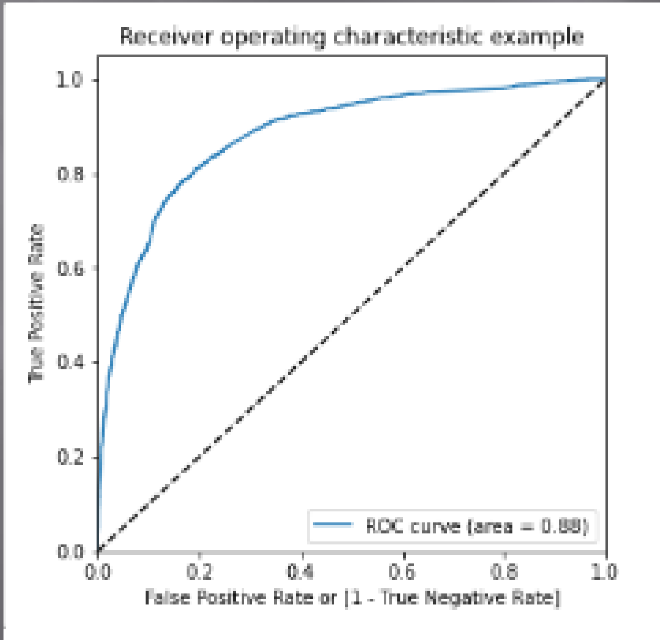
Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6453
Model Family:	Binomial	Df Model:	14
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2714.0
Date:	Sat, 09 Oct 2021	Deviance:	5427.9
Time:	10:50:46	Pearson chi2:	7.24e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5654	0.120	-4.713	0.000	-0.801	-0.330
Total Time Spent on Website	1.1036	0.039	28.175	0.000	1.027	1.180
Lead Origin_Lead Add Form	2.9862	0.185	16.109	0.000	2.623	3.350
Lead Source_Direct traffic	-1.3035	0.111	-11.740	0.000	-1.521	-1.086
Lead Source_Google	-0.8786	0.104	-8.412	0.000	-1.083	-0.674
Lead Source_Organic search	-1.0713	0.130	-8.261	0.000	-1.325	-0.817
Lead Source_Referral sites	-1.0459	0.313	-3.337	0.001	-1.660	-0.432
Last Activity_Email Bounced	-1.0577	0.301	-3.515	0.000	-1.648	-0.468
Last Activity_Email Opened	0.4806	0.105	4.569	0.000	0.274	0.687
Last Activity_SMS Sent	1.5678	0.107	14.718	0.000	1.359	1.777
What is your current occupation_Working Professional	2.8508	0.190	15.037	0.000	2.479	3.222
Last Notable Activity_Had a Phone Conversation	3.7897	1.109	3.416	0.001	1.615	5.964
Last Notable Activity_Modified	-0.8782	0.086	-10.212	0.000	-1.047	-0.710
Last Notable Activity_Olark Chat Conversation	-0.9435	0.334	-2.827	0.005	-1.598	-0.289
Last Notable Activity_Unreachable	1.9693	0.519	3.792	0.000	0.951	2.987

	Features	VIF
0	const	12.32
4	Lead Source_Google	2.06
8	Last Activity_Email Opened	2.04
9	Last Activity_SMS Sent	2.02
3	Lead Source_Direct traffic	2.00
5	Lead Source_Organic search	1.56
12	Last Notable Activity_Modified	1.46
2	Lead Origin_Lead Add Form	1.35
1	Total Time Spent on Website	1.24
13	Last Notable Activity_Olark Chat Conversation	1.14
7	Last Activity_Email Bounced	1.11
6	Lead Source_Referral sites	1.07
10	What is your current occupation_Working Profes...	1.07
14	Last Notable Activity_Unreachable	1.02
11	Last Notable Activity_Had a Phone Conversation	1.01

Model Evaluation

- Plotting ROC curve for evaluation of our model
- The roc curve is towards the left meaning our has good accuracy
- Area under the curve is 0.88 which is a good score .

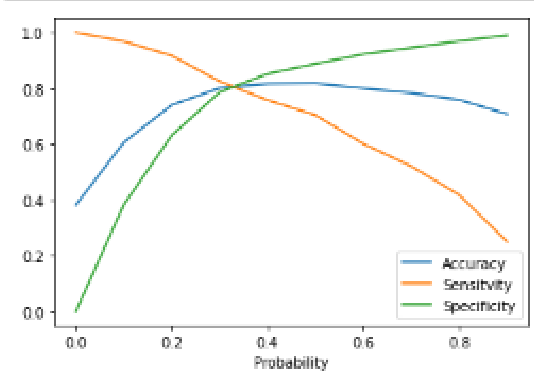


Choosing Threshold value

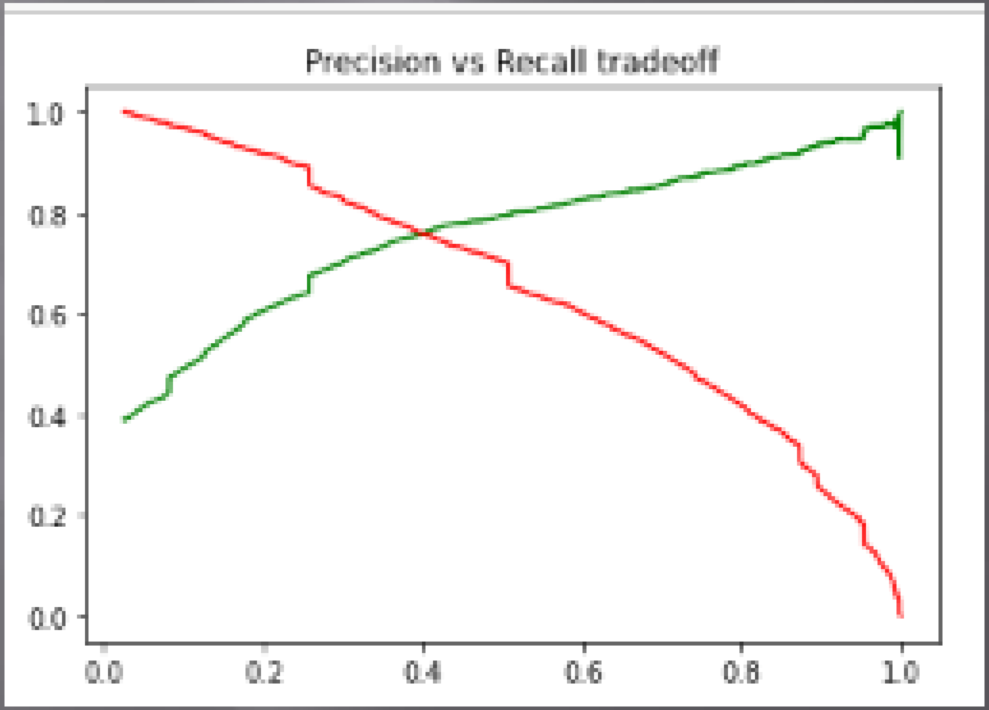
- Choosing a threshold of 0.3 for our model,after checking the accuracy, precision and recall for various thresholds.
- We choose the cutoff of 0.3 as for this value Accuracy ,Sensitivity ,Specificity are nearly equal

	Probability	Accuracy	Sensitvity	Specificity
0.0	0.0	0.381262	1.000000	0.000000
0.1	0.1	0.605751	0.968370	0.382309
0.2	0.2	0.740105	0.917275	0.630935
0.3	0.3	0.800866	0.824818	0.786107
0.4	0.4	0.816017	0.757908	0.851824
0.5	0.5	0.817718	0.703974	0.887806
0.6	0.6	0.799938	0.600973	0.922539
0.7	0.7	0.783395	0.520276	0.945527
0.8	0.8	0.758967	0.417680	0.969265
0.9	0.9	0.707792	0.250203	0.989755

```
#Plotting a graph
cutoff.plot.line(x='Probability',y=['Accuracy','Sensitvity','Specificity'])
plt.show()
```



- We have a good precision and recall value of ~76% Our model is able to explain relevancy of 76% and true relevant results around 76% for both train and test datasets.
- Also the accuracy is ~81% for both train and test data



Conclusion

- We have a good model with test data recall and precision similar to training data So we can say that our model is predicting the conversions correctly in future even when the company's requirement changes.

- The variables with highest coefficient are more significant and help determine probability of conversion :

  Last Notable Activity_Had a Phone Conversation

  Lead Origin_Lead Add Form

  What is your current occupation_Working Professional

- All the metrics are in acceptable range ,our model is stable and will be able to predict conversions correctly

Recommendations

- In order to increase the probability of lead conversion Last Activity_SMS Sent,Last Notable Activity_Unreachable,Last Activity_Email Opened , these variables should be focused.
- Focusing more on working professionals , people who spent more time on website/interacted with our team via SMS,email or phone , would lead to a higher probablity of lead conversion.
- Concentrate more on working professionals as they can spend money on course and people who have had phone coversation earlier who seem to be more interested .Here we will only be checking the hot leads having conversion score above 90 ,so that we can minimize the rate of useless phone calls