LEAD SCORING
CASE STUDY

By - Vyankatesh Kale

INTRODUCTION

Company named X Education gets a lot of leads. However, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, 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 as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

PROBLEM STATMENT

X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

BUSINESS OBJECTIVE

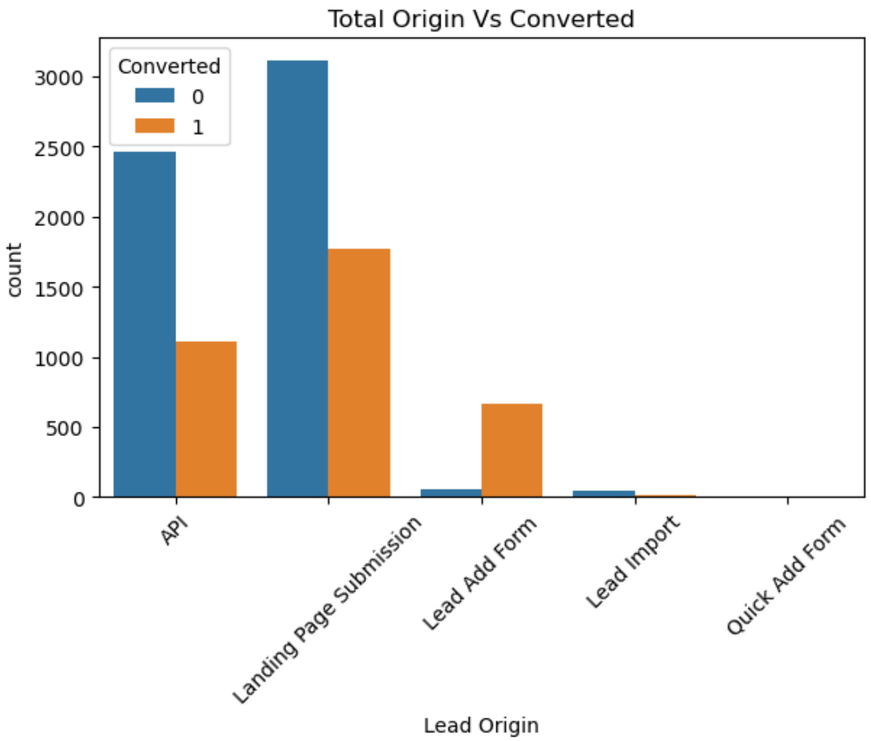
Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

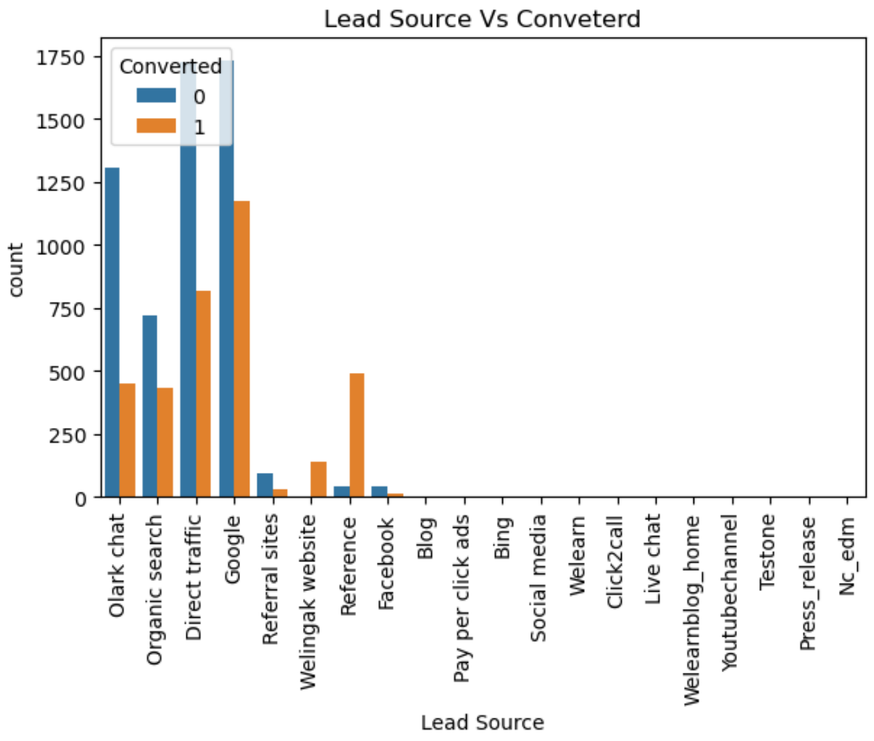
There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step

APPROACH

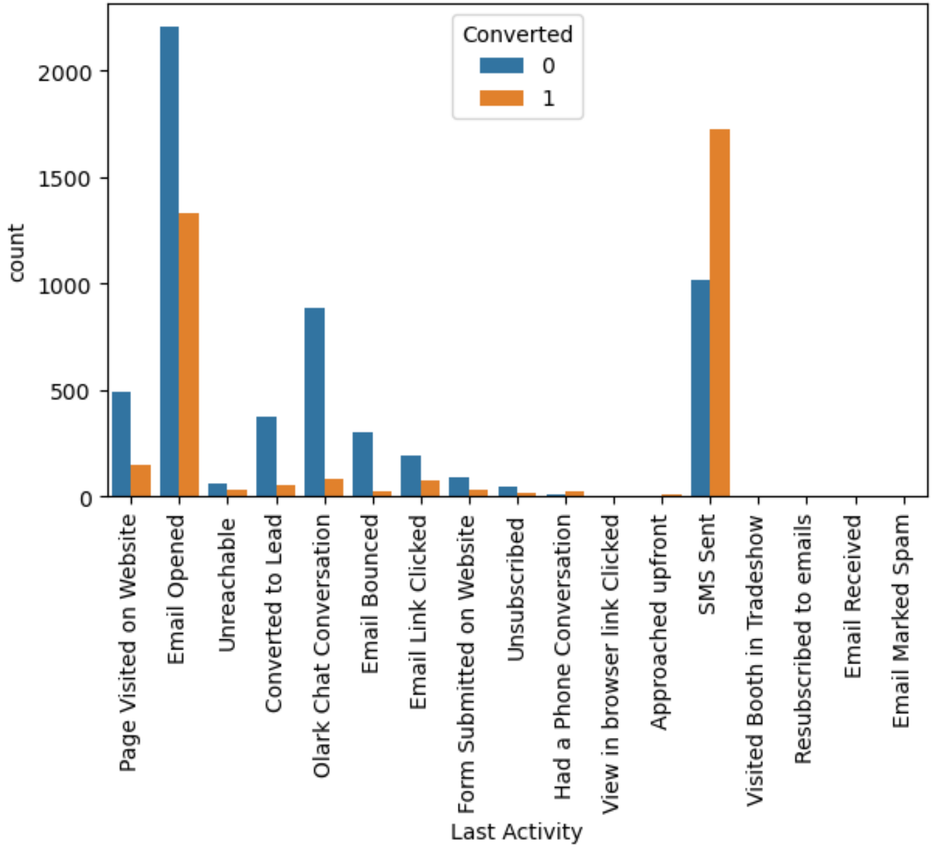
1. Import the packages
2. Read the data (Leads.csv)
3. Check the null/missing values
4. Replace null values with ‘Not Provided’
5. Remove columns having 35% null values and remaining missing values were imputed using Mode and Median
6. Perform Exploratory Data Analysis
7. Change the multicategorical lebels into dummy variables and binary
8. Split data into Train set and Test set
9. Build a Model
10. Firstly RFE was done to attain the top 15 variables
11. Calculate Accuracy for both the Train and Test set data
12. Calculate Precision and Recall Score
13. Conclusion

DATA
VISUALIZATION

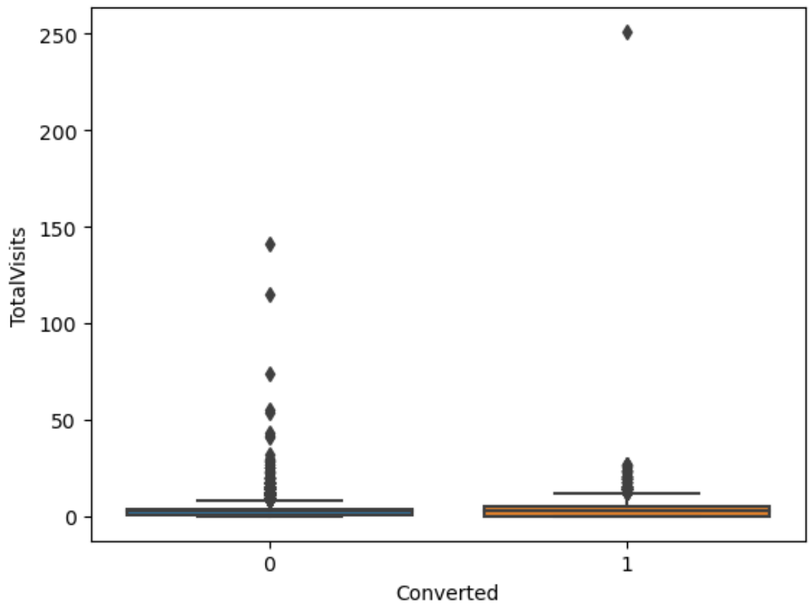


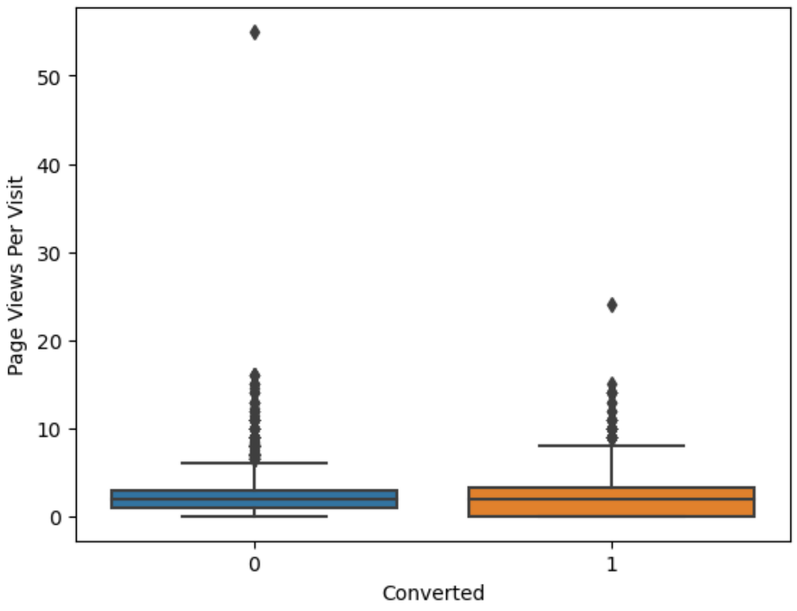


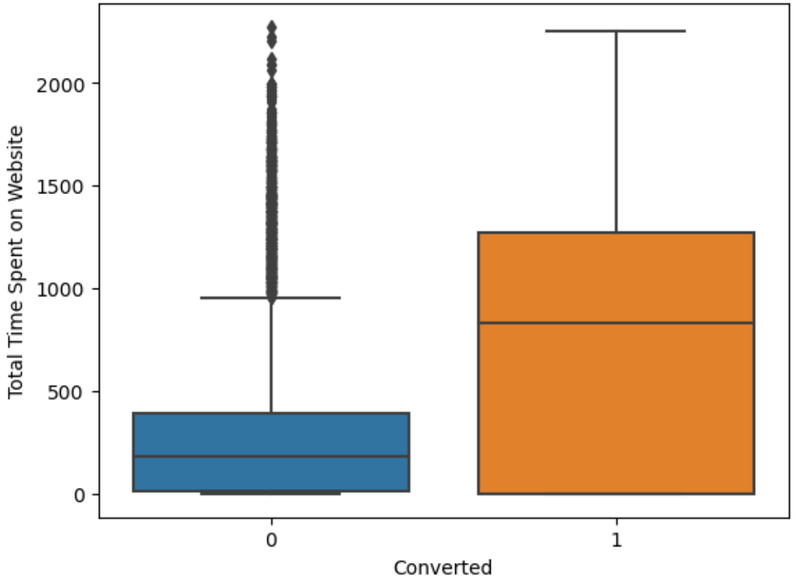
- Maximum leads are from Landing page submission but Lead Add Form have a good conversion rate
- Maximum leads are genrerated from Lead Source "Google"
- "Reference" and "Welingak website" have the higher conversion rate



- Last Activity as "Email Opened" have the highest lead but conversion rate is not good
- "SMS sent" have high conversion rate
- Most of the data is from Country "India" so it is not that insightful.

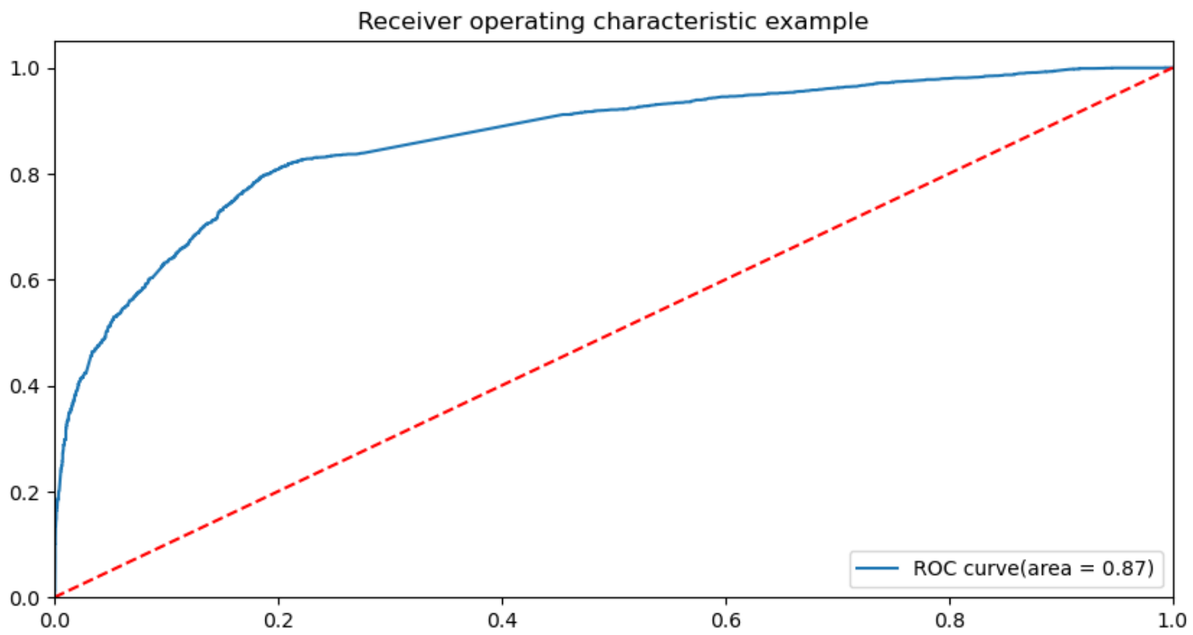


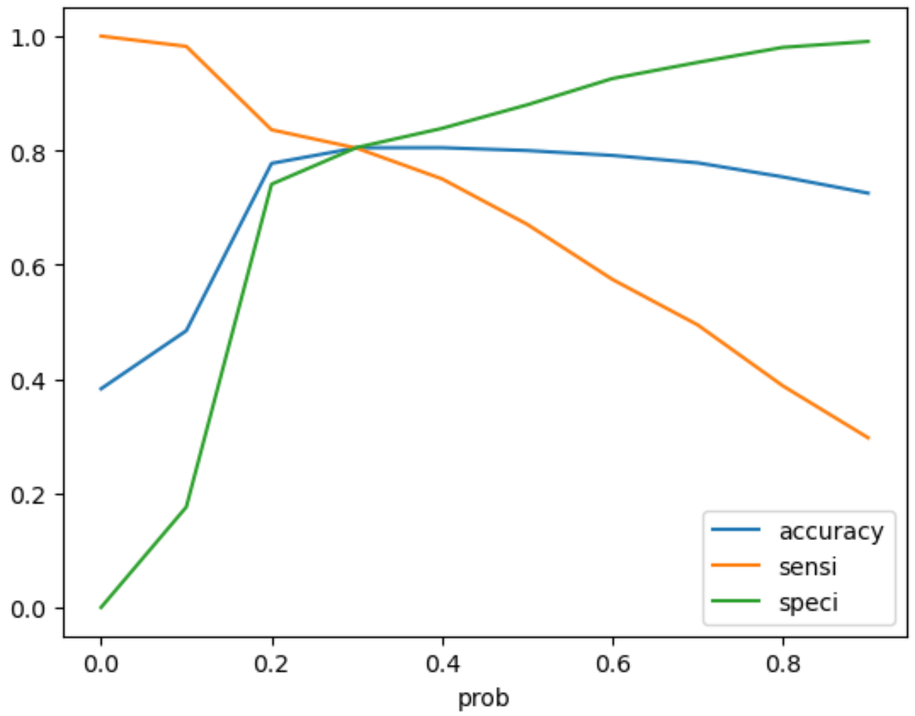




- Totalvisit and Page Views Per Visit have same median for converted non converted so we can not conclude anything
- Total Time Spent on Website : Median of converted is higher than non converted
- Converted leads are those who have spend more time on the website

- Highest correlation is between 'Lead Import' and 'Lead Source_Facebook' - (0.96)
- correlation between Lead Add Form and Lead Source_Referance - (0.86)
- correlation between TotalVisits and Page Views per visits - (0.49)





- The value of ROC Curve should be near to 1 and we are getting 0.87 which is good predictive model
- Optimal Cutoff for Accuracy, Sensitivity, Specificity is 0.35

Out[117]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	7392
Model:	GLM	Df Residuals:	7379
Model Family:	Binomial	Df Model:	12
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3201.2
Date:	Wed, 23 Aug 2023	Deviance:	6402.4
Time:	16:03:04	Pearson chi2:	1.07e+04
No. Iterations:	7	Pseudo R-squ. (CS):	0.3716
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.6390	0.217	-2.946	0.003	-1.064	-0.214
Do Not Email	-0.3580	0.041	-8.763	0.000	-0.438	-0.278
Total Time Spent on Website	1.1005	0.036	30.485	0.000	1.030	1.171
Lead Add Form	3.4340	0.185	18.568	0.000	3.072	3.797
Lead Source_Click2call	-2.6434	1.314	-2.012	0.044	-5.218	-0.069
Lead Source_Olark chat	0.8365	0.089	9.421	0.000	0.662	1.011
Lead Source_Welingak website	2.7576	0.740	3.729	0.000	1.308	4.207
Specialization_Hospitality Management	-0.7487	0.300	-2.494	0.013	-1.337	-0.160
What is your current occupation_Unemployed	-0.7821	0.217	-3.601	0.000	-1.208	-0.356
What is your current occupation_Working Professional	1.8335	0.275	6.678	0.000	1.295	2.372
Lead Profile_Lateral Student	2.7712	1.081	2.564	0.010	0.653	4.890
Lead Profile_Potential Lead	1.7203	0.088	19.563	0.000	1.548	1.893
Lead Profile_Student of SomeSchool	-2.4538	0.431	-5.699	0.000	-3.298	-1.610

Calculating VIF

In [118]:

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
vif = pd.DataFrame()
vif['Features'] = X_train[col].columns
vif['VIF'] = [variance_inflation_factor(X_train[col].values,i) for i in range(X_train[col].shape[1])]
vif['VIF'] = round(vif['VIF'],2)
vif = vif.sort_values(by = 'VIF' ,ascending= False)
vif
```

Out[118]:

	Features	VIF
2	Lead Add Form	1.56
7	What is your current occupation_Unemployed	1.51
4	Lead Source_Olark chat	1.48
10	Lead Profile_Potential Lead	1.42
5	Lead Source_Welingak website	1.27
8	What is your current occupation_Working Profes...	1.26
1	Total Time Spent on Website	1.24
0	Do Not Email	1.02
6	Specialization_Hospitality Management	1.02
11	Lead Profile_Student of SomeSchool	1.02
3	Lead Source_Click2call	1.01
9	Lead Profile_Lateral Student	1.01

- All the variables have good p-value (p-value < 0.05)
- Also all variables have a good VIF value(VIF value < 5)

CONCLUSION

Accuracy Score in predicting Train dataset - 79.96%
Accuracy Score in predicting Test dataset - 79.54%
Precesion Score in predicting Test dataset - 73.22%
Recall Score in predicting Test dataset - 76.22%

The Accuracy , Presicion ,Recall Score we got on Test set is in Acceptable range
We were looking a high recall Score than Precision score and we got that
The accuracy from Train set and Test set both are close to each other
Hence the Model looks good

THANK YOU !