Summary of Lead-Scoring Case-Study Assignment

Step 1 - Business Problem

The X Education company requires to build a model wherein we need to assign a lead score to each of the leads such that the customers with the higher lead score have a higher conversion chance and the customers with the lower lead score have a lower conversion chance.

Step 2 – Data Preparation

- We started with importing basic libraries and performing basic checks on the data. There were 9240 rows and 37 columns initially. Also, we deleted duplicate rows wherever it was required.
- We started with Outlier treatment in the Numerical Variables. We capped the values using Quantiles (0.01 and 0.99) in order to get rid of outliers.
- We performed data cleaning and imputation methods on almost every categorical variable.
- Created Dummy variables for all categorical variables and replaced the original columns.

Step 3 – Model Building

- Divided the data set into Test and Train Dataset
- Put all the feature variables in X, and target variable in y i.e. “Converted”.
- Scaled the three numeric features present in the dataset by using StandardScaler() method.
- Built the logistic regression model using the function GLM() under Statsmodel library. This model contained all the variables, some of which had insignificant coefficients. Hence, some of these variables were removed first based on an automated approach, i.e. RFE and then a manual approach based on the VIFs and p-values.
- Also, with each model we created a data frame with the actual "Converted" flag and the predicted probabilities.

Dep. Variable:	Converted	No. Observations:	6372
Model:	GLM	Df Residuals:	6362
Model Family:	Binomial	Df Model:	9
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1978.3
Date:	Sun, 31 May 2020	Deviance:	3956.7
Time:	23:37:14	Pearson chi2:	7.85e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

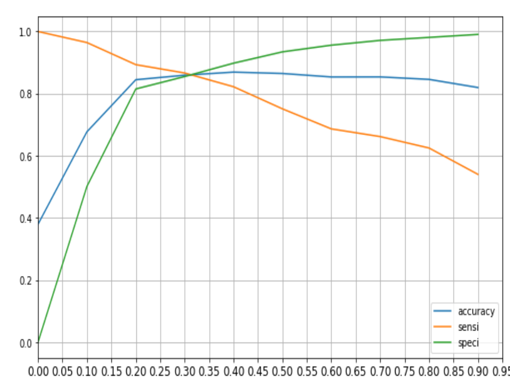
	coef	std err	z	P>|z|	[0.025	0.975]
const	1.1846	0.138	8.581	0.000	0.914	1.455
Total Time Spent on Website	0.9932	0.047	21.110	0.000	0.901	1.085
Ld_Sr_Direct Traffic	-3.0726	0.165	-18.609	0.000	-3.396	-2.749
Ld_Sr_Google	-2.7877	0.160	-17.440	0.000	-3.101	-2.474
Ld_Sr_Olark Chat	-2.1771	0.158	-13.813	0.000	-2.486	-1.868
Ld_Sr_Organic Search	-2.9580	0.186	-15.864	0.000	-3.324	-2.593
Lst_Act_L_A_Otr	-0.6330	0.109	-5.786	0.000	-0.847	-0.419
Tags_Ringing	-3.3150	0.239	-13.866	0.000	-3.784	-2.846
Tags_Will revert after reading the email	4.2976	0.172	24.996	0.000	3.961	4.635
Lst_Ntbl_Act_SMS Sent	1.8270	0.106	17.174	0.000	1.619	2.036

	Features	VIF
8	Lst_Ntbl_Act_SMS Sent	1.56
7	Tags_Will revert after reading the email	1.49
1	Ld_Sr_Direct Traffic	1.46
2	Ld_Sr_Google	1.45
5	Lst_Act_L_A_Otr	1.42
0	Total Time Spent on Website	1.27
6	Tags_Ringing	1.25
3	Ld_Sr_Olark Chat	1.21
4	Ld_Sr_Organic Search	1.21

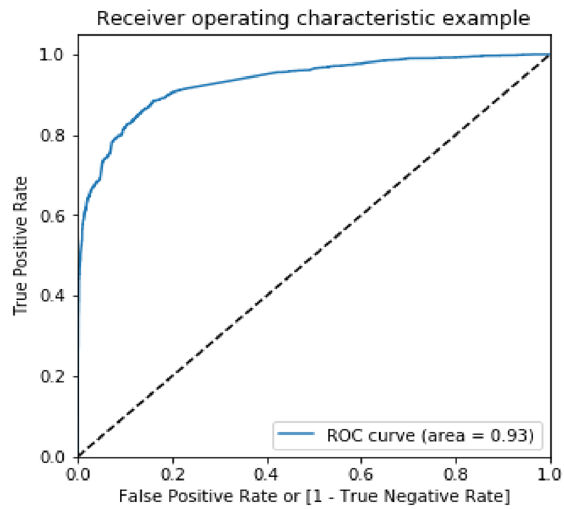
Step 4 - Model Evaluation: Accuracy, Sensitivity, and Specificity

- We first calculated confusion matrix. It was basically a matrix showing the number of all the actual and predicted labels.
- 0.3 is taken as the optimum cutoff point as all the 3 metrices (Sensitivity, Specificity and Accuracy) seems to be doing pretty good at this point.
- Accuracy – 85.9%
- Sensitivity – 86.6%
- Specificity – 85.5%
- FPR– 14.5%
- Positive predictive value – 78%
- Negative predictive value - 91%

Probability	Accuracy	Sensitivity	Specificity
0.0%	38.0%	100.0%	0.0%
10.0%	67.7%	96.4%	50.1%
20.0%	84.5%	89.3%	81.5%
30.0%	85.9%	86.6%	85.5%
40.0%	86.9%	82.2%	89.8%
50.0%	86.5%	75.1%	93.4%
60.0%	85.4%	68.7%	95.6%
70.0%	85.4%	66.2%	97.1%
80.0%	84.6%	62.5%	98.1%
90.0%	82.0%	54.0%	99.1%



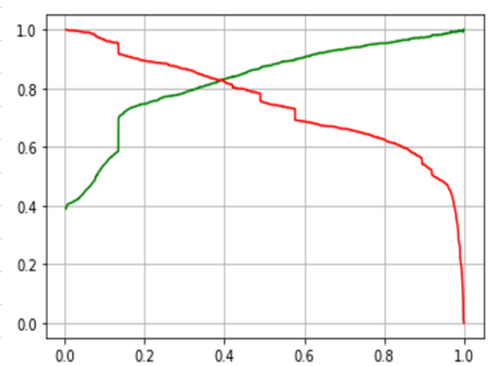
- **ROC Curve** - shows the tradeoff between sensitivity and specificity. It specifies that the test is very likely to be accurate.



- **Precision and Recall Tradeoff -** At 40% Probability we got decent Precision and Recall tradeoff with all the other metrics.

Precision – Recall Trade Off

Probability	Precision	Recall	F1 Score
0%	38%	100%	55%
10%	54%	96%	69%
20%	75%	89%	81%
30%	79%	87%	82%
40%	**83%**	**82%**	**83%**
50%	88%	75%	81%
60%	91%	69%	78%
70%	93%	66%	77%
80%	95%	63%	75%
90%	97%	54%	69%



---Precision --- Recall

Step 5 - Conclusion

- We concluded that in both the datasets – Train and Test, we got decent values of all the three metrics – Accuracy (~86.7%), Sensitivity (~88.1%), and Specificity (~85.5%).

	Probability	Accuracy	TPR (Sensitivity)	Specificity	FPR (1-Specificity)
Train Model	30.0%	85.9%	86.6%	85.5%	14.5%
Test Model	30.0%	86.70%	88.10%	85.90%	14.0%

- It defines we are around 85.9% sure that the model is accurately predicting the customers who are converted and correctly predicting non converted customers as non-converted only.
- If Company want the Leads to get converted, then Leads with below features should be targeted-
  - Tags_Will revert after reading the email
  - Lst_Ntbl_Act_SMS Sent
  - Total Time Spent on Website
- If Company want the Leads to get converted, then Leads with below features should not be targeted-
  - Lst_Act_L_A_Otr
  - Ld_Sr_Olark Chat
  - Ld_Sr_Google
  - Ld_Sr_Organic Search
  - Ld_Sr_Direct Traffic
  - Tags_Ringing