Model to identify the most potential leads.

- HARSHIT PAUNIKAR

(IIIT Bangalore)

Problem Statement

X Education is currently struggling with the problem of very low lead conversion rates. Organization wishes to identify the most potential leads - 'Hot Leads' for the for the sales team to focus on leads which are more likely to convert.

Our goal here is to:

- 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.
- Try to maintain the conversion rate to 80% or above
- Develop a stable and robust model which could handle future requirement changes

Information provided

- We are given the “ Leads.csv ” data set with 9240 rows and 37 columns.

- This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not

- The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.

Approach

We performed following steps working on this problem

- Data Understanding
- Data Exploration
- Data cleaning and preparation
- Model Development
- Model Evaluation
- Model Predictions
- Results & Recommendations

Data Understanding

- The Data Frame contains 9240 rows and 37 Columns.
- Values present in most of the columns are same e.g. Newspaper Article(No -9238, Yes-2), Magazine. etc
- More than 50% values of columns are blank e.g. Lead Quality(51.6% missing) etc.
- Specific variable is derived from other variable e.g. Asymmetrique Activity Index is derived out of Asymmetrique Activity Score.
- Variables having around 20% to 40% missing (NaN) values were imputed with appropriate values.

Data Preparation for Modelling

- Removed the Outliers present in various columns accordingly.
- Changed the categorical variables with ‘Yes’ or ‘No’ to 0 and 1 respectively.
- Created dummy variables for the categorical variables with multiple values.
- We impute the suitable values in place of missing values before starting the model building .

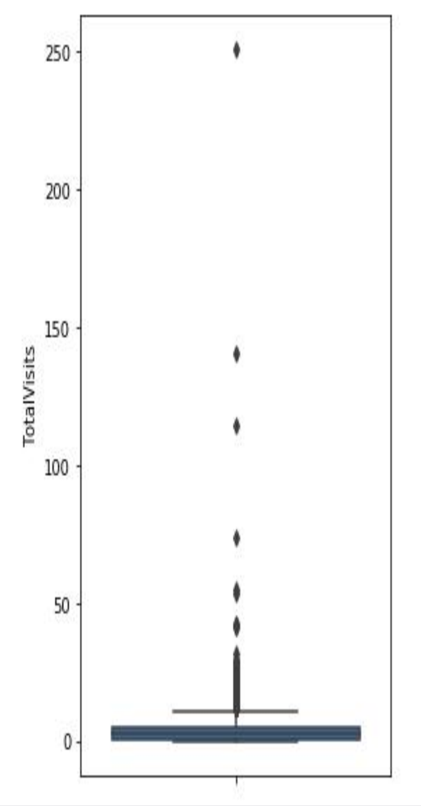
Imputation of Categorical columns

Some columns are having comparatively low missing values such as:

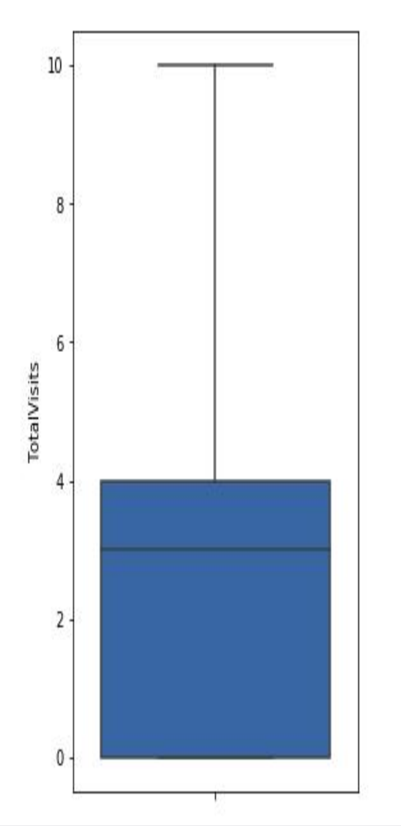
- Lead Source
- What is your current occupation
- Tags
- Lead Quality

Above columns are imputed with the mode value of the data of the column.

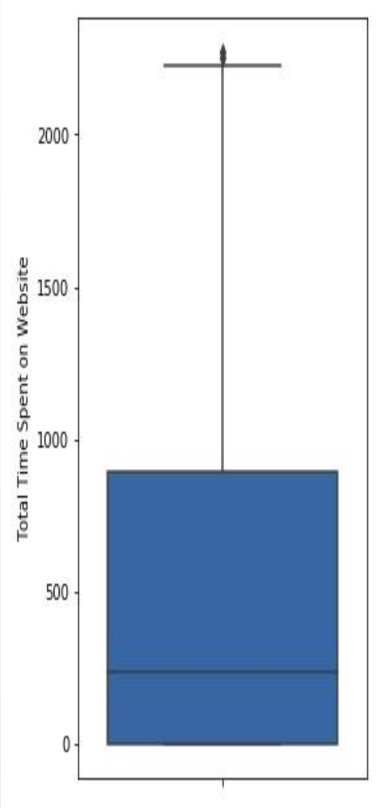
Outliers analysis and removal of outliers



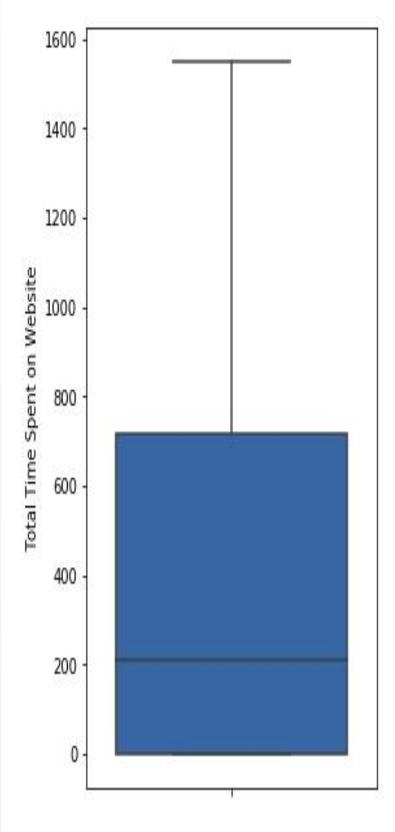
“Total visits”column
having outliers



“Total visits”column after
removing outliers



“Total time spent on
websites”column haivng
outliers



“Total time spent on
websites”column after
removing outliers

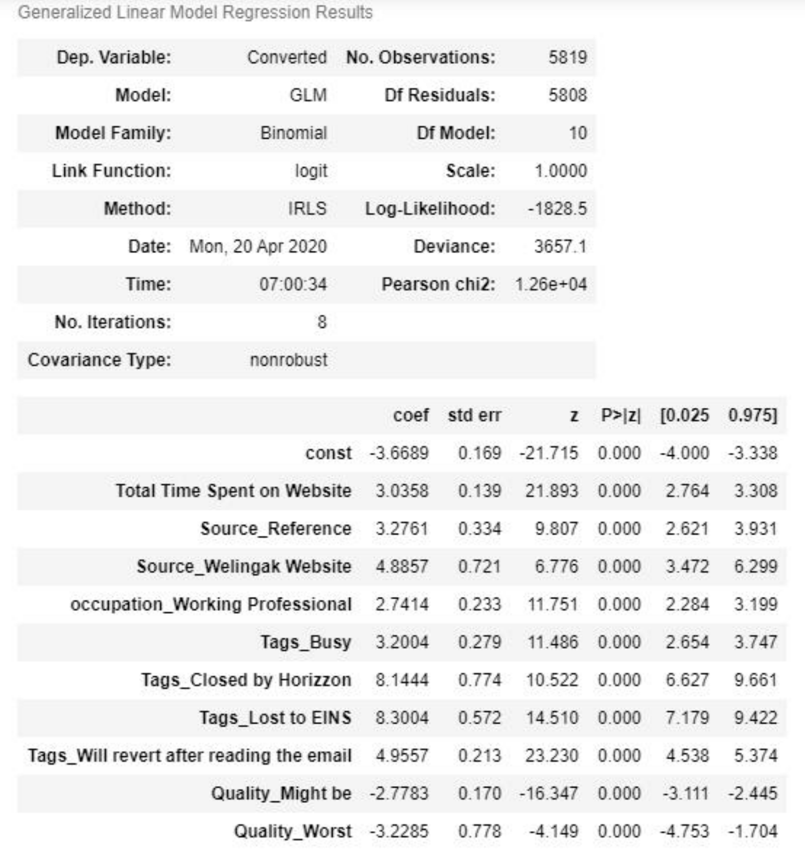
Model Development

Below are the steps performed for model development:

- Identified predictor and target variables
- Splitted Data into Test and Train datasets
- Re-scaled predictor variables
- Run the first iteration of model on train dataset
- Select the features for modeling using RFE feature selection technique
- Run deferent iterations of model removing one variable at a time based on P-values and VIFs.

Analysing model based on P-values

- We can observe in the given summary that p values of the variables are zero this means they are playing significance role in model.



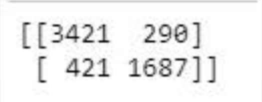
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	5819
Model:	GLM	Df Residuals:	5808
Model Family:	Binomial	Df Model:	10
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1828.5
Date:	Mon, 20 Apr 2020	Deviance:	3657.1
Time:	07:00:34	Pearson chi2:	1.26e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.6689	0.169	-21.715	0.000	-4.000	-3.338
Total Time Spent on Website	3.0358	0.139	21.893	0.000	2.764	3.308
Source_Reference	3.2761	0.334	9.807	0.000	2.621	3.931
Source_Welingak Website	4.8857	0.721	6.776	0.000	3.472	6.299
occupation_Working Professional	2.7414	0.233	11.751	0.000	2.284	3.199
Tags_Busy	3.2004	0.279	11.486	0.000	2.654	3.747
Tags_Closed by Horizzon	8.1444	0.774	10.522	0.000	6.627	9.661
Tags_Lost to EINS	8.3004	0.572	14.510	0.000	7.179	9.422
Tags_Will revert after reading the email	4.9557	0.213	23.230	0.000	4.538	5.374
Quality_Might be	-2.7783	0.170	-16.347	0.000	-3.111	-2.445
Quality_Worst	-3.2285	0.778	-4.149	0.000	-4.753	-1.704

Confusion Matrix, Sensitivity, Specificity, False positive rate, Positive predictive value, Negative predictive value:

Confusion matrix



```
[[3421  290]
 [ 421 1687]]
```

Sensitivity	0.80
Specificity	0.92
False Positive Rate	0.07
Positive Predictive Value	0.85
Negative predictive value	0.89

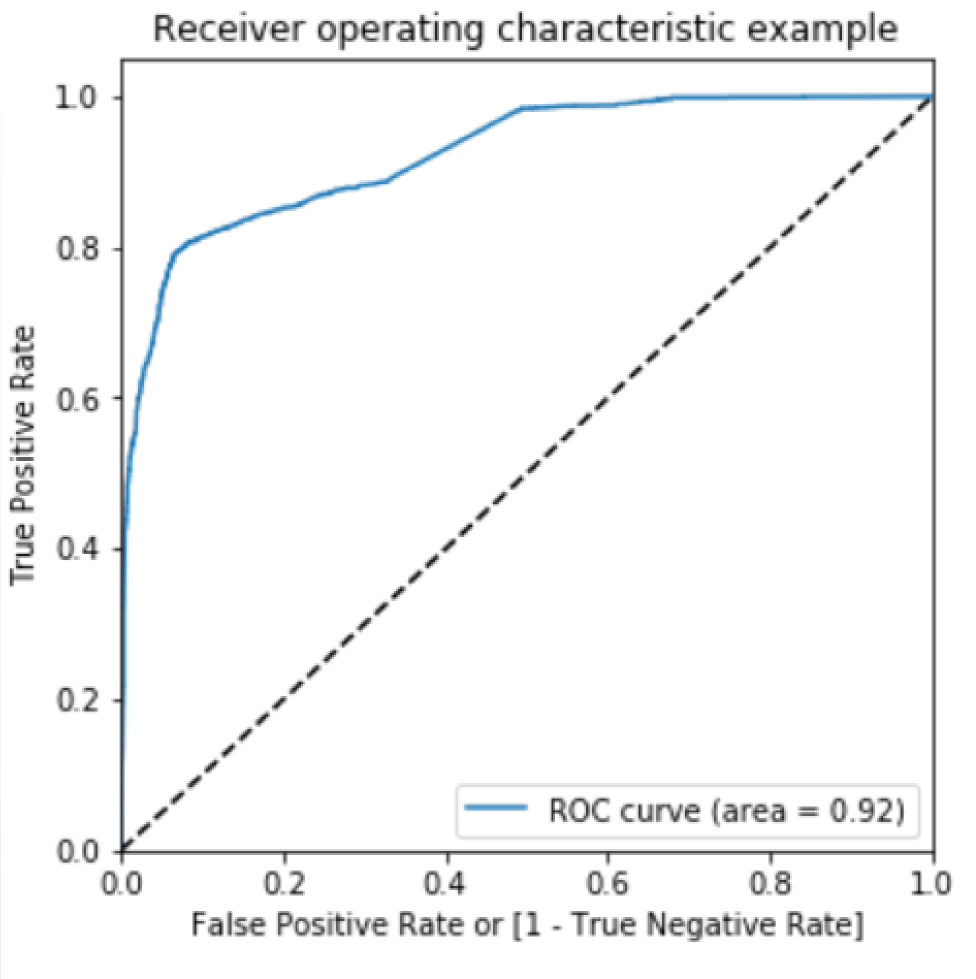
- Here we have calculated the values of Sensitivity, specificity, positive predictive value, Negative predictive value, False predictive value etc. to access the model that we made.

Model Evaluation

- Predicted Target Variable using the final iteration of model
- Created a dataset with the actual conversion flag and the predicted probabilities
- Created Confusion Matrix to understand Model's accuracy, sensitivity & Specificity
- Created RoC Curve
- The True Positive Rate (Sensitivity) is plotted in function of the false positive rate for different cut-off points of a parameter. It shows the trade-off between sensitivity and specificity
- Found the optimal probability cutoff Sensitivity, Specificity and accuracy. Point of intersection when all three metric are plotted gives the optimal cut-off. It is 0.3 for our model.

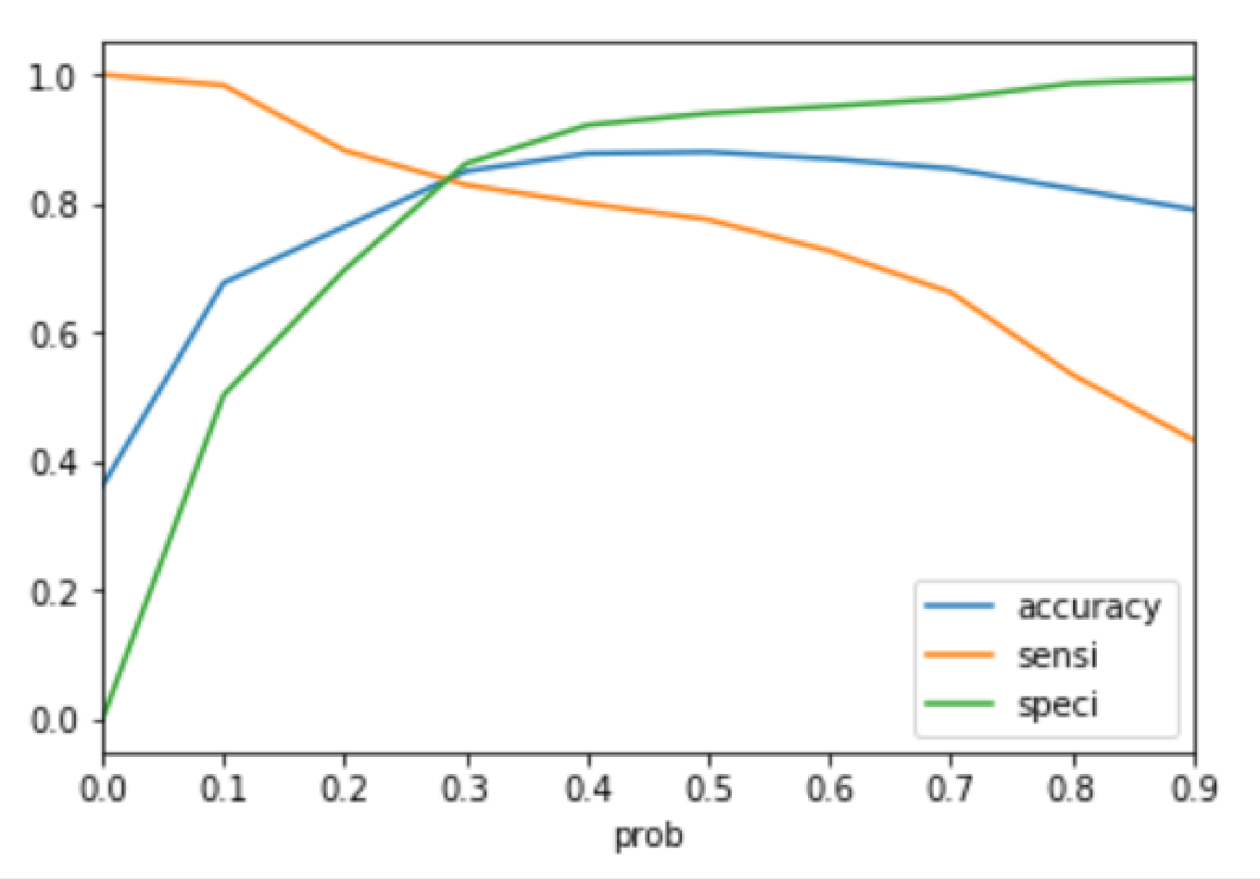
ROC Curve

- We can observe from the graph that area under the curve is more so the model is good.



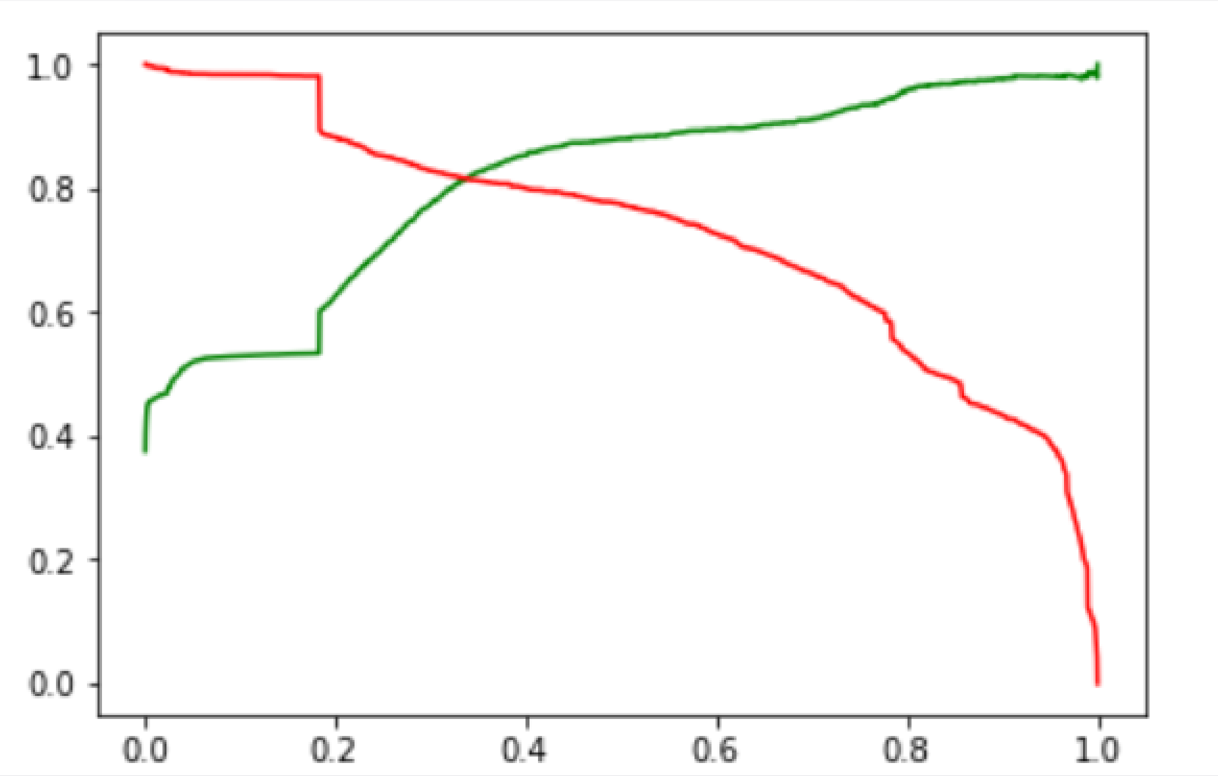
Accuracy, Specificity & Sensitivity

- We observe the graph between accuracy, specificity & sensitivity and selected 0.30 as the best probability value.



Precision vs Recall graph:

- We can observe from the graph that precision and recall are best when they intersect at a point. At that point their tradeoff is minimum.



Model Predictions

- Scaled the predictor variables of test data
- Assigned the probabilities using the model designed and predicted the target variable
- Checked the Sensitivity, Specificity and accuracy on test data
- Reviewed the precision and recall trade-off

Recommendations

- Model developed is predicting the hot leads with 84.9% accuracy and 85.31% conversion rate on test data
- Top 3 variables that contribute most towards the probability of a lead conversion are:
    - Tags
    - Lead Source
    - Total Time Spent on Website
- Model is robust enough to handle requirement changes in future – Management can change the cut-offs based on requirement. In order to meet the aggressive targets, cut-offs could be lowered down and for selective contacts, cutoff could be made more stringent
- Data had lot of missing values. Model would have been more robust if the data points were populated. May be a feedback for data collection agents