Lead Scoring: Case Study

By : Arun Dutt and Ramandeep Singh

Problem Statement

What is required from us?

● X Education had ask us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

● For the above statement we have to build a model wherein which we 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.

Analysis Approach

To achieve our problem statement, we did the following steps :

1. Performing EDA
2. Data Preparation
3. Test-Train splitting
4. Scaling the numerical data
5. Model building
6. Plotting the ROC curve
7. Finding optimal cutoff point
8. Finding sensitivity and specificity
9. Precision and Recall
10. Making prediction on test set

PERFORMING EDA

Data cleaning:

- Handling “Select” which is present in many of the categorical variables and replacing select by NaN.
- Checking the missing values column and row wise.
- Dropping the independent variable(i.e score variable created by business team) and dropping the columns with high missing values(i.e missing values >45%)
- Dropping the rows having more than 5 missing values.
- Dropping the Highly skewed columns(i,e NO=100%) which are not required.
- Imputing mean for the missing values in numerical column and for categorical variable column we have imputing 'mode' for missing value.
- Identifying the categories of categorical columns which are having less row count and combining these categories and naming it as „Other‟.
- Before cleaning the data the dimension of data frame was (9240, 37) and after cleaning the dimension became (9191, 14) .
- We are retained with 99% of rows after data cleaning process.

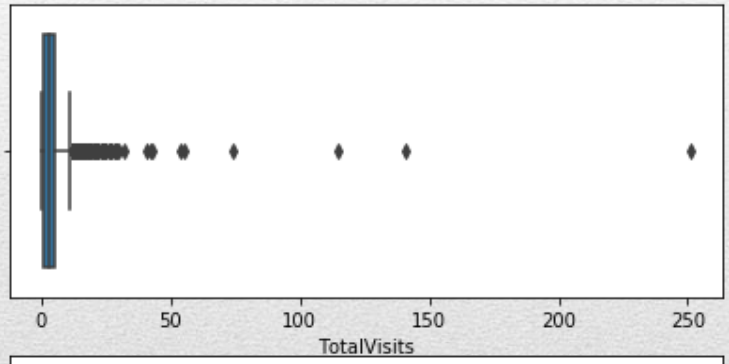
Data Preparation

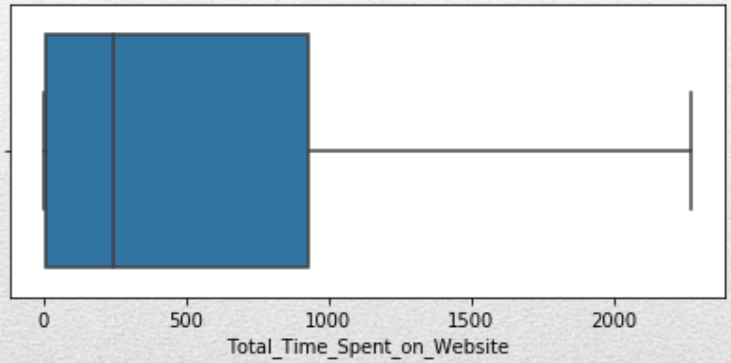
Dummifying Variables :

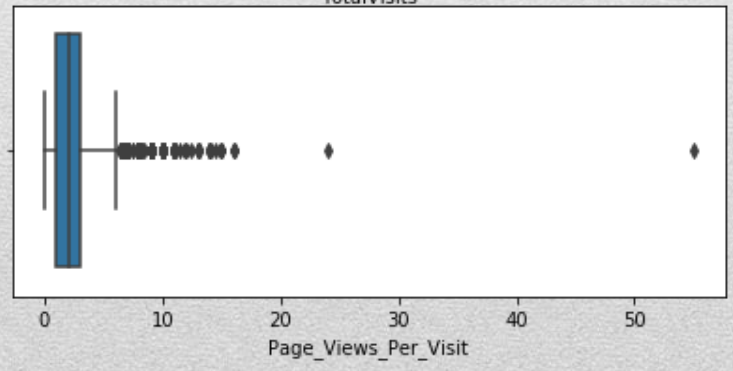
- Creating a dummy variable for all the categorical variables and dropping the first one.
- Adding the dummy results to the main dataframe.
- Dropping the column of repeated variables (i.e. column of the variables that have been dummified.)

Checking for Outliers

Plotting the boxplot graph to find outliers

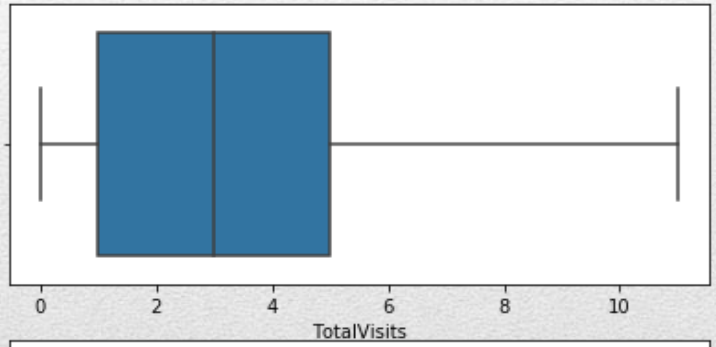


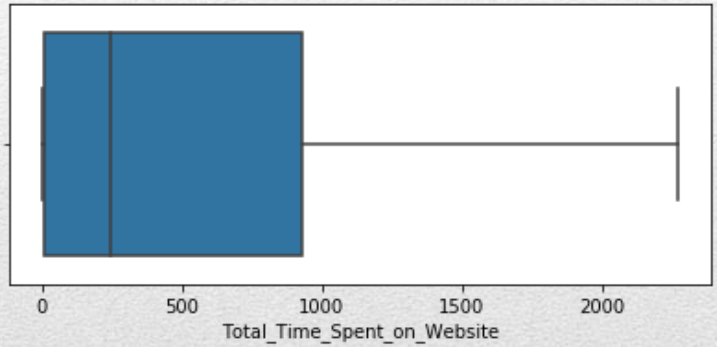


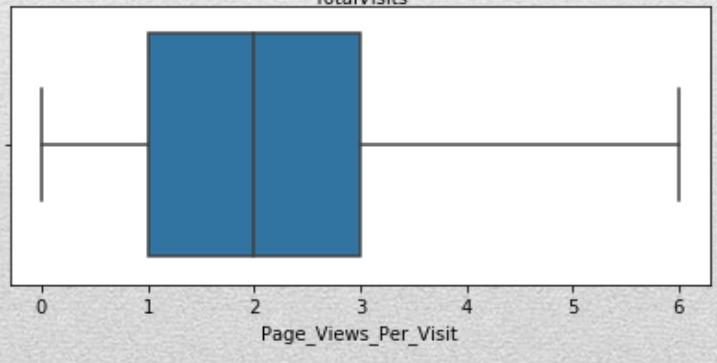


Outlier treatment

Capping the outlier and replotting box plot after outlier treatment.







Test-Train Spliting

- Putting feature variable to X
- Putting response variable to y
- Splitting the data into train and test(i.e.70% -30%)

Scaling the numerical data

- Scaling of numerical variable in train set using Standard Scalar method.
- We have almost 38% converted rate.

Model Building

- Feature are selected using RFE method and we selected 20 columns as output for starting model building.
- Building the model using stats models and assessing it.
- Dropping the columns which are having high P-vlaues (i.e p>5%) one at a time.
- After that ,dropping the column with high VIF value(VIF>5%) one at a time.
- Re-running the model without the dropped column and performing the above 2 steps .
- Till we get all variables with good value of VIF and P-value.
- Now we can proceed with making predictions using this final model.

Assessing the model with StatsModels

Running 1st test model :

Generalised Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6433
Model:	GLM	**Df Residuals:**	6412
Model Family:	Binomial	**Df Model:**	20
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2713.9
Date:	Sat, 05 Sep 2020	**Deviance:**	5427.9
Time:	12:05:54	**Pearson chi2:**	7.27E+03
No. Iterations:	7		
Covariance Type:			nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5108	0.266	-5.679	0.000	-2.032	-0.989
Total_Time_Spent_on_Website	1.0895	0.039	27.665	0.000	1.012	1.167
Lead_Origin_Landing Page Submission	-0.6025	0.103	-5.863	0.000	-0.804	-0.401
Lead_Origin_Lead Add Form	3.2337	0.212	15.256	0.000	2.818	3.649
Lead_Source_Olark Chat	0.8518	0.116	7.315	0.000	0.624	1.080
Lead_Source_Welingak Website	2.8476	1.031	2.763	0.006	0.827	4.868
Last_Activity_Email Bounced	-1.0543	0.371	-2.844	0.004	-1.781	-0.328
Last_Activity_Email Link Clicked	1.0900	0.226	4.832	0.000	0.648	1.532
Last_Activity_Email Opened	1.4875	0.134	11.064	0.000	1.224	1.751
Last_Activity_Form Submitted on Website	0.9759	0.337	2.897	0.004	0.316	1.636
Last_Activity_OTHER	1.1002	0.319	3.448	0.001	0.475	1.726
Last_Activity_Page Visited on Website	0.9407	0.184	5.102	0.000	0.579	1.302
Last_Activity_SMS Sent	1.5221	0.179	8.514	0.000	1.172	1.873
Specialization_Finance Management	-0.4163	0.090	-4.642	0.000	-0.592	-0.241
Specialization_Hospitality Management	-0.4911	0.325	-1.511	0.131	-1.128	0.146
Specialization_Retail Management	-0.6584	0.328	-2.004	0.045	-1.302	-0.015
Specialization_Services Excellence	-1.0891	0.710	-1.535	0.125	-2.480	0.302
What_is_your_current_occupation_Unemployed	-0.6568	0.221	-2.967	0.003	-1.091	-0.223
What_is_your_current_occupation_Working Professional	2.0292	0.288	7.041	0.000	1.464	2.594
Last_Notable_Activity_OTHER	1.1701	0.351	3.332	0.001	0.482	1.858
Last_Notable_Activity_SMS Sent	1.3060	0.143	9.140	0.000	1.026	1.586

CHECKING THE VIF

	Features	VIF
1	Lead_Origin_Landing Page Submission	1.90
7	Specialization_Finance Management	1.87
6	Last_Activity_Email Opened	1.84
3	Lead_Source_Olark Chat	1.75
10	Last_Notable_Activity_SMS Sent	1.64
2	Lead_Origin_Lead Add Form	1.58
0	Total_Time_Spent_on_Website	1.28
4	Lead_Source_Welingak Website	1.26
8	What_is_your_current_occupation_Working Profes...	1.21
5	Last_Activity_Email Bounced	1.15
9	Last_Notable_Activity_OTHER	1.09

- **After dropping the column which are having high VIF value the final model VIF**

Dropped the column one at a time which had the highest P-value (P-value > 5%) until we achieve test model were all the P-values are <5%

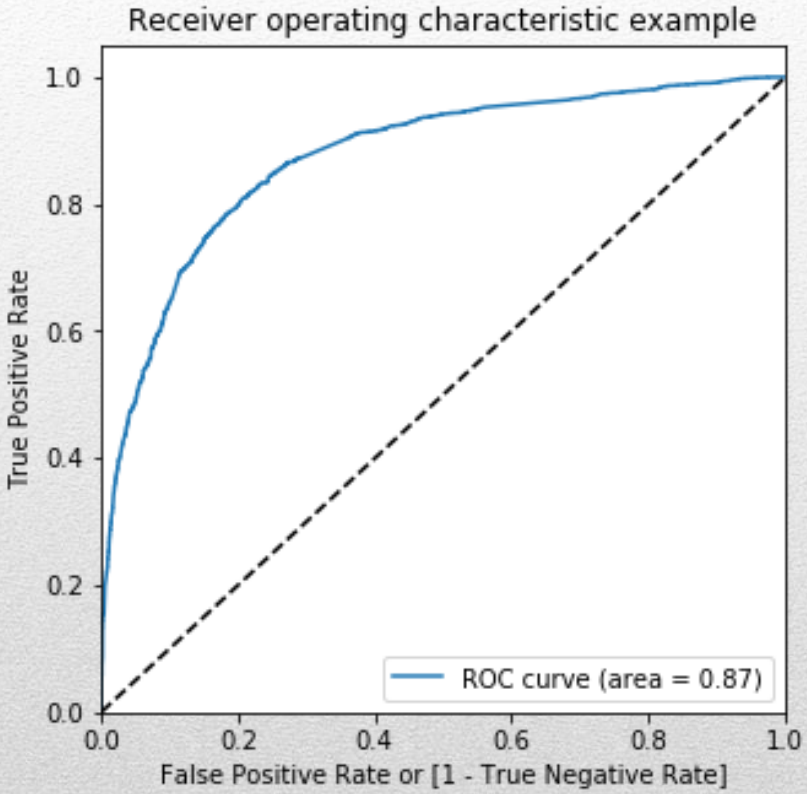
Running Final test model :

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6433
Model:	GLM	**Df Residuals:**	6421
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2762.7
Date:	Sat, 05 Sep 2020	**Deviance:**	5525.5
Time:	12:05:57	**Pearson chi2:**	7.49E+03
No. Iterations:	7		
Covariance Type:			nonrobust

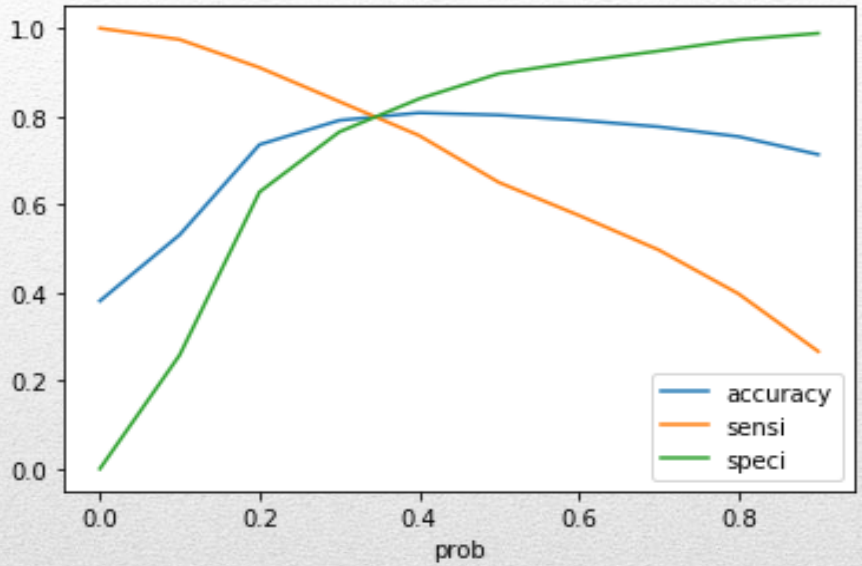
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3942	0.115	-12.161	0.000	-1.619	-1.170
Total_Time_Spent_on_Website	1.0954	0.039	27.954	0.000	1.019	1.172
Lead_Origin_Landing Page Submission	-0.5582	0.102	-5.480	0.000	-0.758	-0.359
Lead_Origin_Lead Add Form	3.4136	0.212	16.072	0.000	2.997	3.830
Lead_Source_Olark Chat	0.7691	0.114	6.763	0.000	0.546	0.992
Lead_Source_Welingak Website	2.8529	1.030	2.770	0.006	0.834	4.872
Last_Activity_Email Bounced	-1.9720	0.344	-5.741	0.000	-2.645	-1.299
Last_Activity_Email Opened	0.7161	0.082	8.739	0.000	0.556	0.877
Specialization_Finance Management	-0.4374	0.089	-4.931	0.000	-0.611	-0.264
What_is_your_current_occupation_Working Professional	2.6274	0.189	13.883	0.000	2.256	2.998
Last_Notable_Activity_OTHER	1.4271	0.258	5.526	0.000	0.921	1.933
Last_Notable_Activity_SMS Sent	2.0358	0.092	22.173	0.000	1.856	2.216

Plotting ROC curve



Since the curve is towards the upper-left corner and the area under the curve (AUC) is more so we have a better model.

Finding Optimum cutoff point



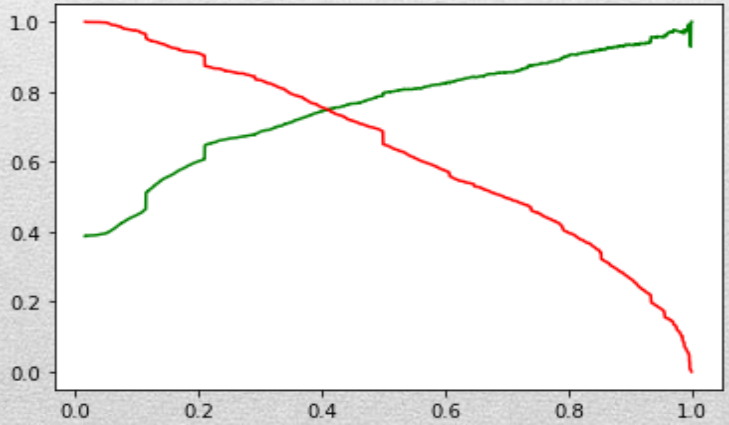
- To we find best cutoff point we calculated and plotted accuracy, sensitivity and specificity for various probabilities.
- From the curve above we see that all the three parameter are coinciding at a point i.e. 0.34 which our optimum point to take it as a cutoff probability.

Finding sensitivity and specificity

- After getting optimum cutoff probability point we did our final prediction.
- We checked for overall accuracy of the model (~80%)
- We calculated the confusion matrix.
- Using the confusion matrix we calculated sensitivity and specificity
- We got Sensitivity = ~80% and specificity = ~80%.
- We Calculate false postive rate - predicting convertion but the customer does not convert (20%).
- We calculated Positive predictive value(70%) and Negative predictive value(86%).

Finding Precision and Recall

- Precision and Recall which are another pair of industry-relevant metric used to evaluate the performance of a logistic regression module.
- Using the confusion matrix we calculated Precision and Recall
- We got Precision = ~79% and Recall = ~64%.
- We used sklearn utility to calculate the same and we got same value as above.
- Then plotting Precision and Recall tradeoff point to get optimum cutoff point.



As we see the curve the point where precision and recall line meet is our Cutoff point (0.4).

Making prediction on test set

- Scaling of numerical variable in test set using Standard Scalar method.
- After scaling we add the selected columns from RFE method.
- Making the prediction on the test set using stats models.
- When comparing the cutoff point of Accuracy, Sensitivity and Specificity(0.34) with cutoff point of precision and recall curve(0.4). We take the optimal cut-off point for our model 0.34 as we are getting decent values of all the three variable Accuracy, Sensitivity and Specificity as ~80% on train set ,so we use same cutoff probability point(0.34) to make our final prediction on test set .
- We generated Lead Score variable for the converted probability.
- We evaluate the test model by calculating ‘Overall accuracy’, ‘Sensitivity’ and ‘Specificity’
- We got Overall Accuracy =80% , Sensitivity = 80% and Specificity =80% .

Conclusion

- We have build a model where we have assigned a lead score to each of the leads so that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- We hav created the model which has Prediction rate of converting lead in to customer is around 80% and it has Predicting rate non converting lead is around 80% .
- Our final model has an accuracy of 80% of Converting lead into customer.
- We have improved the lead conversion rate from 30% to 80%.


Thank
you!