Lead Scoring: Case Study

By : Abhishek Kumar

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.

PERFORMING EDA

- Data cleaning:
- 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 (9204, 14) .
- We are retained with 99.6% of rows after data cleaning process.

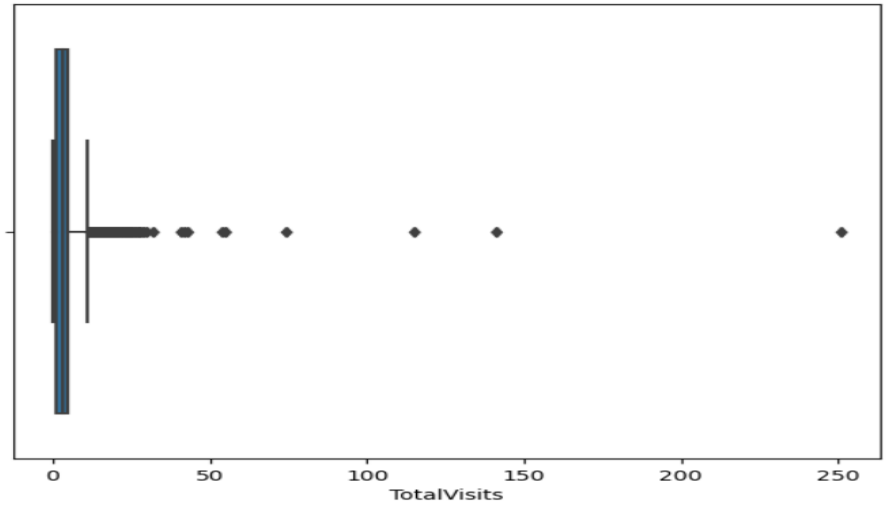
Data Preparation

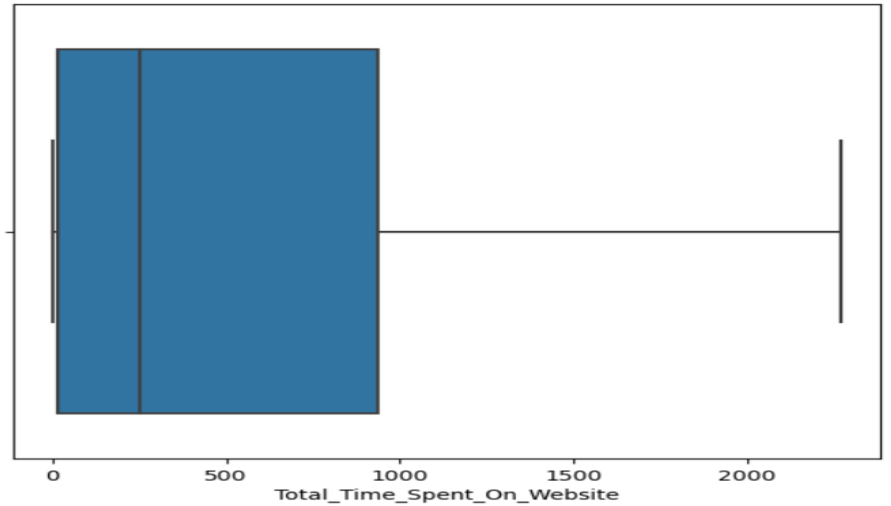
Dummifying Variables :

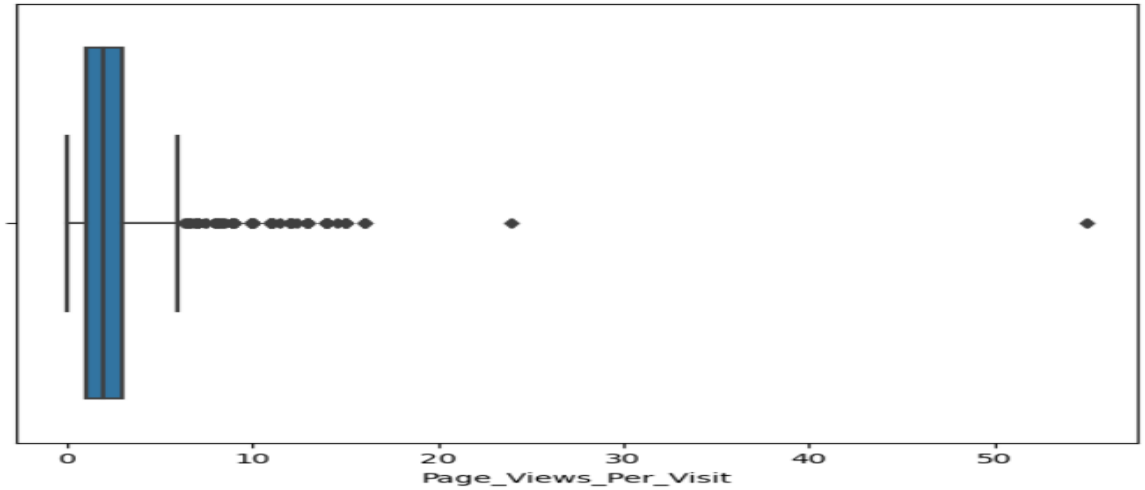
- Creating a dummy variable for all the categorical variables and dropping the first one.
- Adding the dummy results to the main dataframe.
- Dropping the column of repeated variables (i.e. column of the variables that have been dummified.)

Checking for Outliers

Plotting the boxplot graph to find outliers

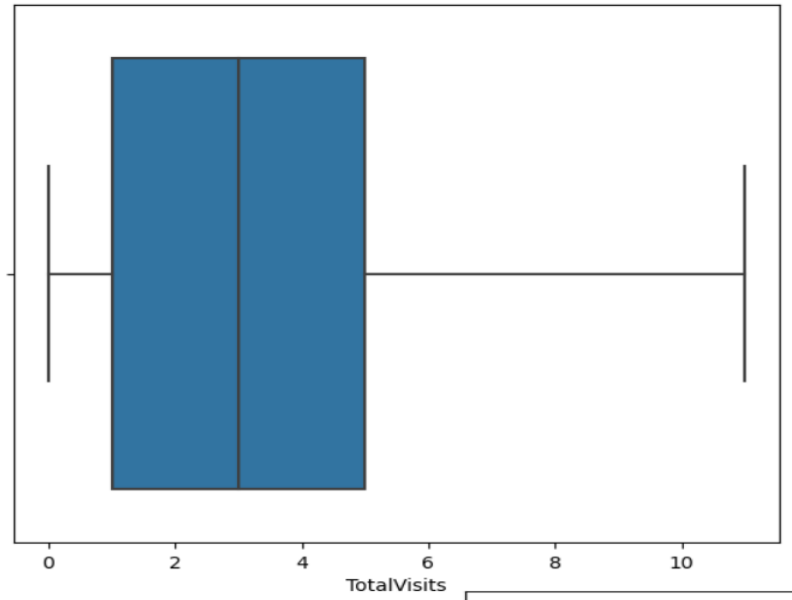


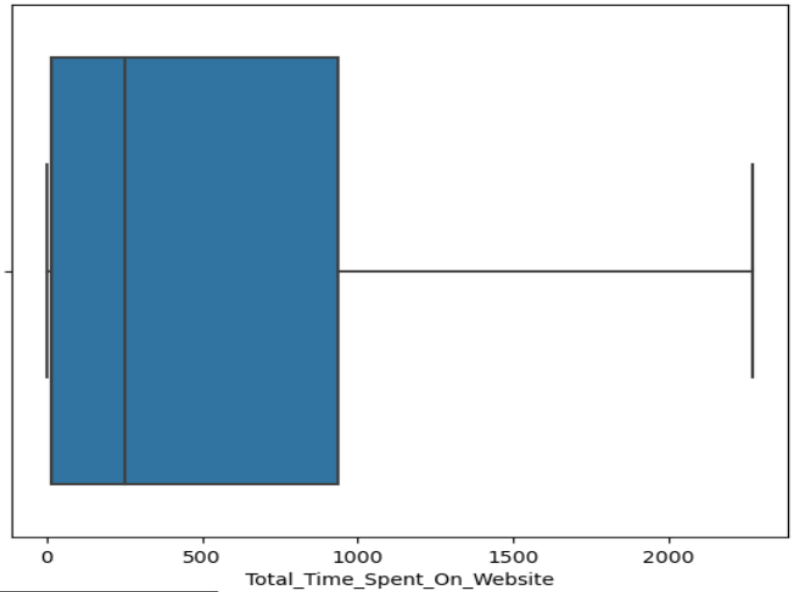


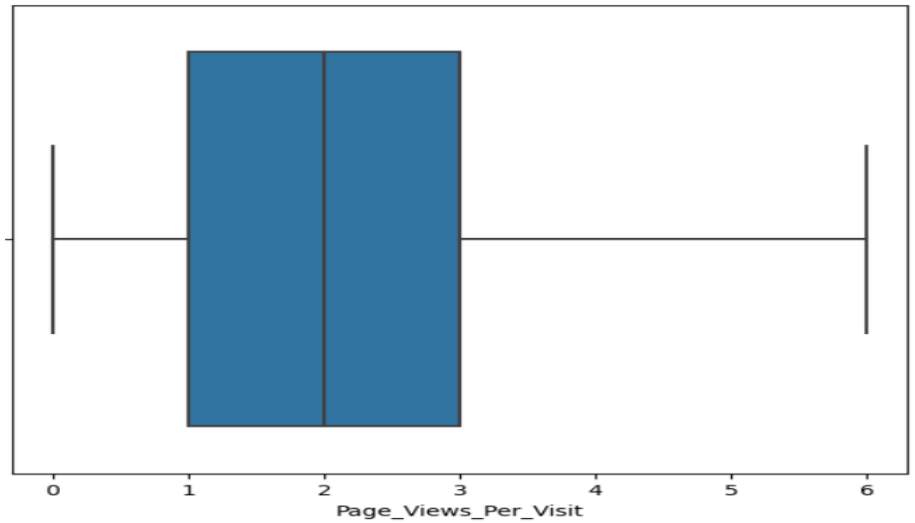


Outlier treatment

Capping the outlier and replotting box plot after outlier







Test-Train Splitting

- 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-values (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.

Model Building

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6442
Model:	GLM	**Df Residuals:**	6370
Model Family:	Binomial	**Df Model:**	71
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1696.5
Date:	Sun, 31 Mar 2024	**Deviance:**	3393.0
Time:	18:09:51	**Pearson chi2:**	1.29e+04
No. Iterations:	24	**Pseudo R-squ. (CS):**	0.5533
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.8547	0.318	-8.979	0.000	-3.478	-2.232
Lead_Origin_Lead Add Form	2.0024	0.254	7.874	0.000	1.504	2.501
Lead_Source_Welingak Website	2.2634	1.045	2.166	0.030	0.215	4.311
Last_Activity_Email Bounced	-0.6827	0.305	-2.239	0.025	-1.280	-0.085
Last_Activity_Email Opened	0.8321	0.128	6.504	0.000	0.581	1.083
Last_Activity_OTHER	1.2870	0.273	4.716	0.000	0.752	1.822
Last_Activity_SMS Sent	1.2242	0.168	7.300	0.000	0.896	1.553
Country_United Kingdom	1.7972	1.109	1.621	0.105	-0.376	3.970
Current_Occupation_Unemployed	-1.6588	0.272	-6.088	0.000	-2.193	-1.125
Current_Occupation_Working Professional	0.9985	0.334	2.988	0.003	0.344	1.654
Tags_Busy	3.3835	0.290	11.655	0.000	2.815	3.952
Tags_Closed by Horizzon	9.1042	1.027	8.868	0.000	7.092	11.116
Tags_Lost to EINS	8.4516	0.746	11.322	0.000	6.988	9.915
Tags_Ringing	-1.0375	0.288	-3.599	0.000	-1.603	-0.473
Tags_Will revert after reading the email	3.5033	0.197	17.779	0.000	3.117	3.889
Tags_switched off	-1.3105	0.549	-2.386	0.017	-2.387	-0.234
Last_Notable_Activity_Modified	-0.8767	0.124	-7.086	0.000	-1.119	-0.634
Last_Notable_Activity_Olark Chat Conversation	-1.1532	0.355	-3.247	0.001	-1.849	-0.457

CHECKING THE VIF

	VIF	Feature
10	1.453123	Tags_Closed by Horizzon
1	1.298685	Lead_Source_Welingak Website
14	1.178276	Tags_switched off
16	1.148181	Last_Notable_Activity_Olark Chat Conversation
9	1.143750	Tags_Busy
2	1.138524	Last_Activity_Email Bounced
11	1.111557	Tags_Lost to EINS
4	1.104310	Last_Activity_OTHER
17	1.007925	Last_Notable_Activity_SMS Sent
6	1.002801	Country_United Kingdom
8	0.930160	Current_Occupation_Working Professional
0	0.773659	Lead_Origin_Lead Add Form
5	0.345525	Last_Activity_SMS Sent
7	0.269098	Current_Occupation_Unemployed
13	0.260159	Tags_Will revert after reading the email
3	0.247826	Last_Activity_Email Opened
12	0.165517	Tags_Ringing
15	0.048263	Last_Notable_Activity_Modified

1. **After dropping the columns which high p-values and vif.**
2. **dropping the column which has highest P-value (<5%) or (.05).**

Model Building

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.8422	0.317	-8.964	0.000	-3.464	-2.221
Lead_Origin_Lead Add Form	1.9999	0.254	7.869	0.000	1.502	2.498
Lead_Source_Welingak Website	2.2645	1.045	2.167	0.030	0.217	4.312
Last_Activity_Email Bounced	-0.6858	0.305	-2.250	0.024	-1.283	-0.089
Last_Activity_Email Opened	0.8297	0.128	6.489	0.000	0.579	1.080
Last_Activity_OTHER	1.2834	0.273	4.705	0.000	0.749	1.818
Last_Activity_SMS Sent	1.2203	0.168	7.280	0.000	0.892	1.549
Current_Occupation_Unemployed	-1.6518	0.272	-6.074	0.000	-2.185	-1.119
Current_Occupation_Working Professional	1.0013	0.334	3.002	0.003	0.347	1.655
Tags_Busy	3.3654	0.290	11.621	0.000	2.798	3.933
Tags_Closed by Horizzon	9.0863	1.026	8.852	0.000	7.075	11.098
Tags_Lost to EINS	8.4330	0.746	11.302	0.000	6.971	9.896
Tags_Ringing	-1.0542	0.288	-3.663	0.000	-1.618	-0.490
Tags_Will revert after reading the email	3.4872	0.196	17.787	0.000	3.103	3.871
Tags_switched off	-1.3292	0.549	-2.422	0.015	-2.405	-0.254
Last_Notable_Activity_Modified	-0.8761	0.124	-7.086	0.000	-1.118	-0.634
Last_Notable_Activity_Olark Chat Conversation	-1.1566	0.355	-3.257	0.001	-1.853	-0.461
Last_Notable_Activity_SMS Sent	1.6850	0.206	8.177	0.000	1.281	2.089

Generalized Linear Model Regression Results

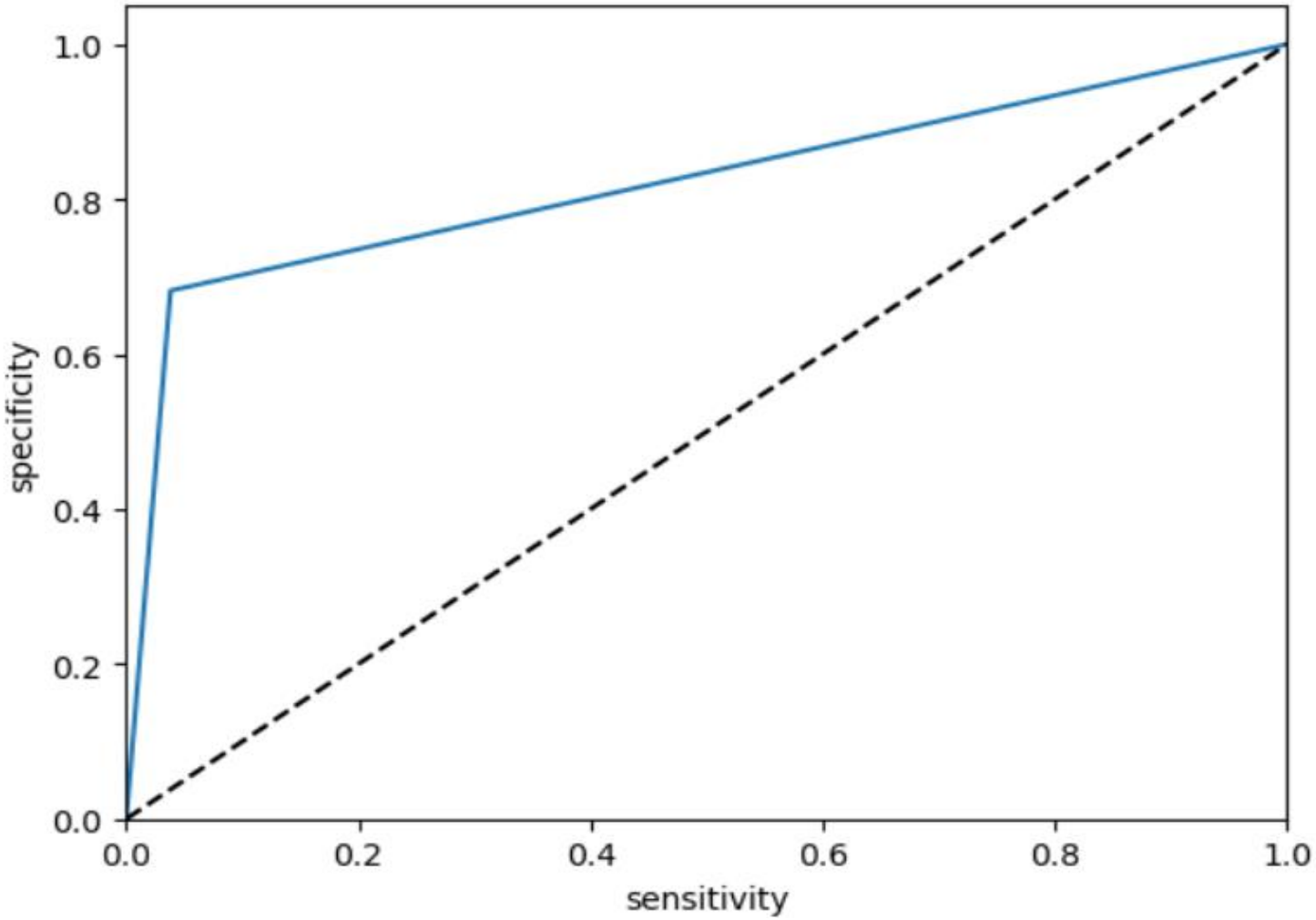
Dep. Variable:	Converted	**No. Observations:**	6442
Model:	GLM	**Df Residuals:**	6422
Model Family:	Binomial	**Df Model:**	19
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2079.0
Date:	Sun, 31 Mar 2024	**Deviance:**	4157.9
Time:	18:09:59	**Pearson chi2:**	1.04e+04
No. Iterations:	23	**Pseudo R-squ. (CS):**	0.4969
Covariance Type:	nonrobust		

CHECKING THE VIF

	VIF	Feature
10	1.469934	Tags_Closed by Horizzon
1	1.298712	Lead_Source_Welingak Website
15	1.190913	Tags_switched off
2	1.164206	Last_Activity_Email Bounced
9	1.152295	Tags_Busy
17	1.148194	Last_Notable_Activity_Olark Chat Conversation
11	1.118877	Tags_Lost to EINS
12	1.115427	Tags_Not doing further education
4	1.104316	Last_Activity_OTHER
18	1.007932	Last_Notable_Activity_SMS Sent
6	1.002908	Country_United Kingdom
8	0.942363	Current_Occupation_Working Professional
0	0.773711	Lead_Origin_Lead Add Form
5	0.345541	Last_Activity_SMS Sent
7	0.277845	Current_Occupation_Unemployed
14	0.275395	Tags_Will revert after reading the email
3	0.247878	Last_Activity_Email Opened
13	0.170764	Tags_Ringing
16	0.048277	Last_Notable_Activity_Modified

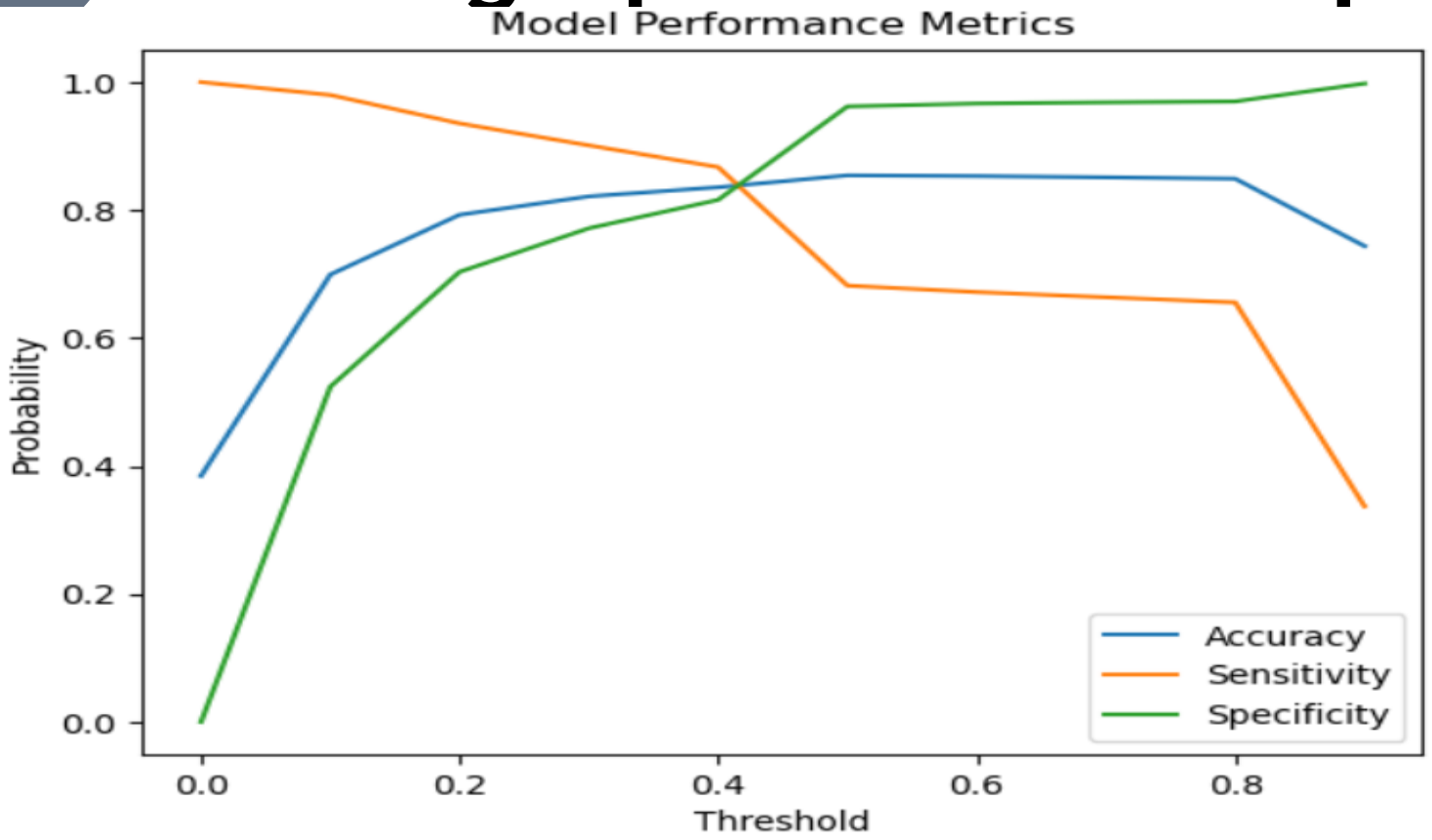
After dropping the columns which high p-values and vif.
dropping the column which has highest P-value (<5%) or (.05).

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

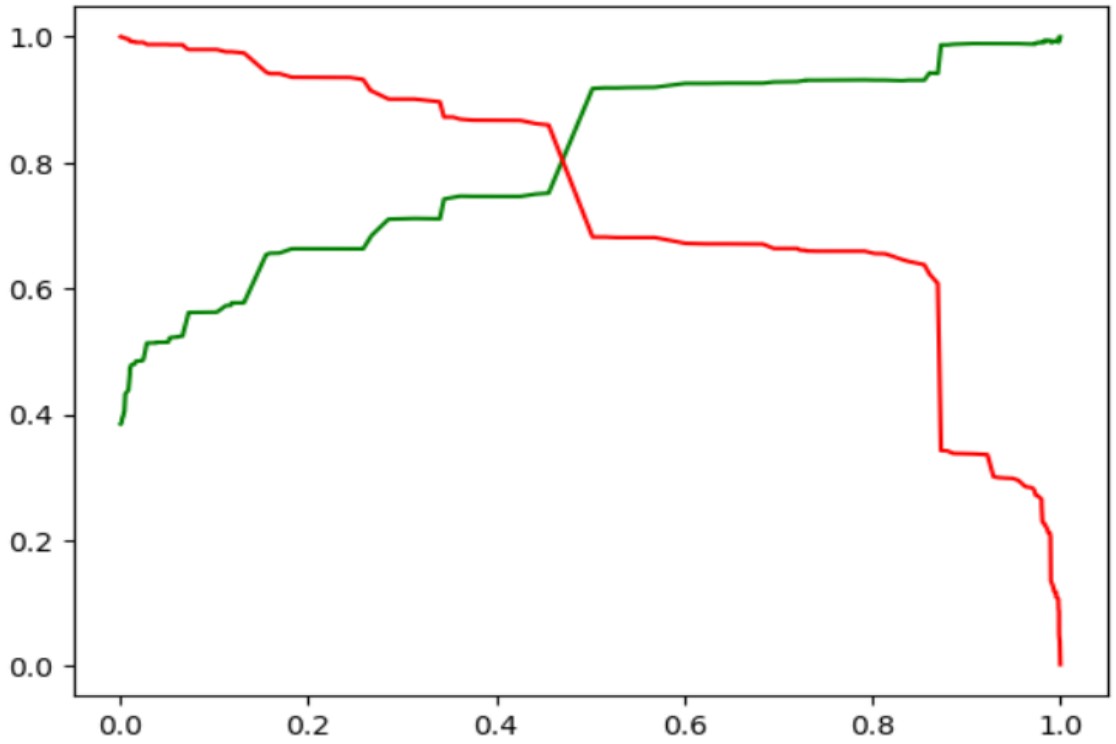
Finding sensitivity and specificity

- 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 = ~86% and specificity = ~81%.
- We Calculate false positive rate - predicting conversion but the customer does not convert (18%).

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 = ~74% and Recall = ~86%.
- 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.

Finding Precision and Recall



When comparing the cutoff point of Accuracy, Sensitivity and Specificity(0.4) with cutoff point of precision and recall curve(0.49). We take the optimal cut-off point for our model 0.4 as we are gettting decent values of all the three variable Accuracy, Sensitivity and Specificity as ~82%

As we see the curve the point where precision and recall line meet is our Cutoff point (0.49).

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(0.49). 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 = 84% , Sensitivity = 88% and Specificity = 81% .

Conclusion

1. 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.
2. We have created the model which has Prediction rate of converting lead in to customer is around 88% and it has Predicting rate non converting lead is around 81% .
3. Our final model has an accuracy of 84% of Converting lead into customer.
4. We have improved the lead conversion rate from 30% to 84%

Thank you