Lead score analysis

Logistic Regression

Agenda

X Education want to find the most promising leads with their data i.e. the leads that are most likely to convert into paying customers. We need to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The ballpark of the target lead conversion rate to be around **80%**.

We will be using **logistic regression** to predict the lead score.

Base Data

- Base file contains history leads with converted column having status of 1 or 0.
- We have different dimensions of the leads like lead source, lead activity, lead origin and nature of the person that lead assigned to it.
- Base data contains null values across many columns and proper data cleaning is required before proceeding to EDA and modeling.

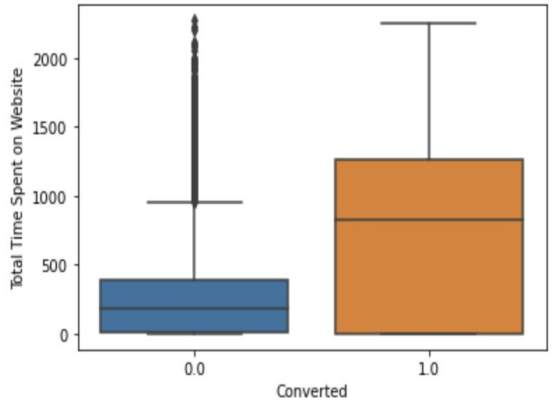
Handling Null Values

- Many columns which take values from website contains values called '**Select**', which essentially is null and we need to impute that with np.nan.
- Removed the columns which has more than **70%** null values.
- For remaining columns imputed the null values with appropriate values. Either mode or 'Other' category.
- Removed Tag column in this process as it talks about lead condition after contacting. Since we are giving lead score to make sure that we need to contact potential customer. More over the data is also not reliable.
- After all these removed some 2% of data which has nulls in some columns.

EDA

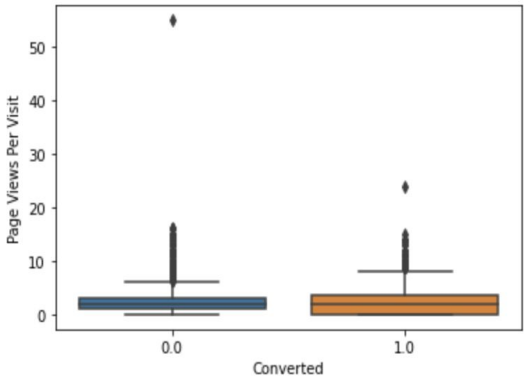
Univariate Analysis

Total Time Spent on Website



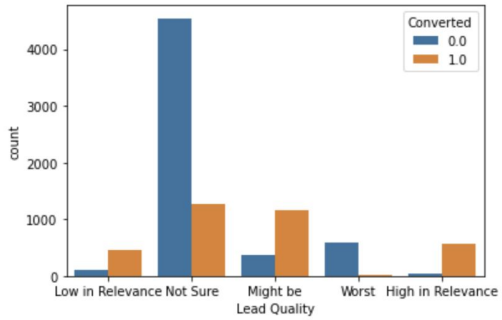
Distribution of Time spend across 2 categories

Page Views Per Visit



Spread of number of visits across converted and unconverted leads

Lead Quality



Converted leads are spread out across all categories in Lead Quality

Redundant variables like country, Magazine, Select etc are removed as they are showing any distribution among the converted and not converted.

RFE and Model Building

- Before going to RFE, dummy variable are created for categorical values.
- Scaling (Normalization) is done for all continuous variable.
- RFE technique is used to reduce the variable from 58 variables to 15 variables.
- Mode 2 results are attached beside.

Model 2

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6335
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2170.9
Date:	Tue, 26 Sep 2023	**Deviance:**	4341.8
Time:	10:07:11	**Pearson chi2:**	6.40e+03
No. Iterations:	21		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.9323	0.224	8.616	0.000	1.493	2.372
Do Not Email	-1.0516	0.179	-5.872	0.000	-1.403	-0.701
Total Time Spent on Website	1.0664	0.044	24.073	0.000	0.980	1.153
Lead Origin_Lead Add Form	5.1413	0.526	9.775	0.000	4.110	6.172
Lead Source_Olark Chat	1.5292	0.116	13.173	0.000	1.302	1.757
Lead Source_Reference	-2.5909	0.576	-4.498	0.000	-3.720	-1.462
Last Activity_Olark Chat Conversation	-0.9580	0.193	-4.972	0.000	-1.336	-0.580
Last Activity_SMS Sent	1.2644	0.084	15.026	0.000	1.099	1.429
What is your current occupation_Housewife	21.4094	1.51e+04	0.001	0.999	-2.97e+04	2.97e+04
What is your current occupation_Working Professional	1.8094	0.217	8.349	0.000	1.385	2.234
What matters most to you in choosing a course_Other	-0.5736	0.097	-5.909	0.000	-0.764	-0.383
Lead Quality_Low in Relevance	-1.0138	0.268	-3.786	0.000	-1.539	-0.489
Lead Quality_Might be	-2.1255	0.240	-8.870	0.000	-2.595	-1.656
Lead Quality_Not Sure	-3.5709	0.231	-15.488	0.000	-4.023	-3.119
Lead Quality_Worst	-5.8643	0.417	-14.075	0.000	-6.681	-5.048
Last Notable Activity_Modified	-0.7353	0.091	-8.107	0.000	-0.913	-0.558

RFE and Model Building

- Model 3 is a stable model with all p values < 0.05 and VIF's are well below 5.

Model 3

Generalized Linear Model Regression Results

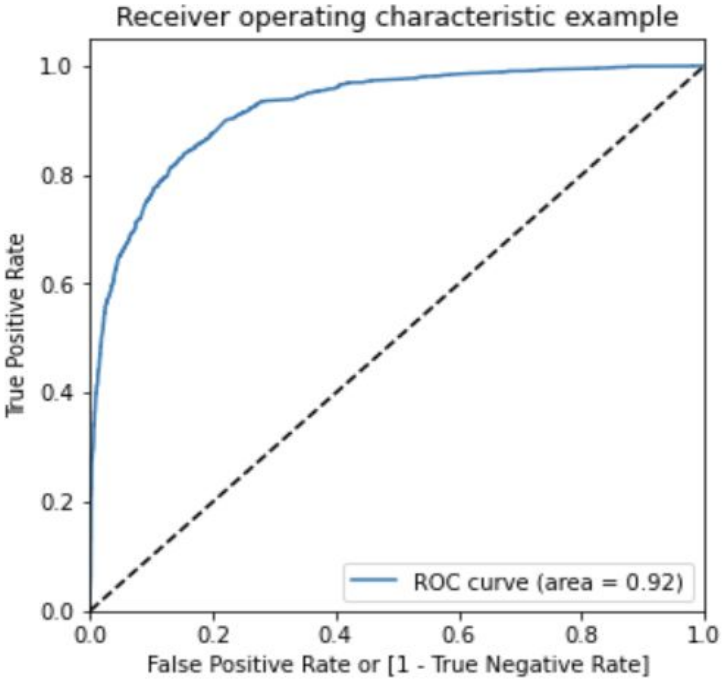
Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6336
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2173.1
Date:	Tue, 26 Sep 2023	**Deviance:**	4346.3
Time:	10:08:09	**Pearson chi2:**	6.42e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.9313	0.224	8.612	0.000	1.492	2.371
Do Not Email	-1.0540	0.179	-5.883	0.000	-1.405	-0.703
Total Time Spent on Website	1.0659	0.044	24.074	0.000	0.979	1.153
Lead Origin_Lead Add Form	5.1406	0.526	9.774	0.000	4.110	6.171
Lead Source_Olark Chat	1.5270	0.116	13.157	0.000	1.300	1.754
Lead Source_Reference	-2.5719	0.576	-4.468	0.000	-3.700	-1.444
Last Activity_Olark Chat Conversation	-0.9606	0.193	-4.984	0.000	-1.338	-0.583
Last Activity_SMS Sent	1.2607	0.084	14.989	0.000	1.096	1.426
What is your current occupation_Working Professional	1.8010	0.217	8.306	0.000	1.376	2.226
What matters most to you in choosing a course_Other	-0.5732	0.097	-5.906	0.000	-0.763	-0.383
Lead Quality_Low in Relevance	-0.9911	0.268	-3.704	0.000	-1.515	-0.467
Lead Quality_Might be	-2.1164	0.240	-8.836	0.000	-2.586	-1.647
Lead Quality_Not Sure	-3.5681	0.231	-15.478	0.000	-4.020	-3.116
Lead Quality_Worst	-5.8615	0.417	-14.070	0.000	-6.678	-5.045
Last Notable Activity_Modified	-0.7346	0.091	-8.106	0.000	-0.912	-0.557

Coefficients

```
const                                                 1.931290
Do Not Email                                         -1.053984
Total Time Spent on Website                           1.065872
Lead Origin_Lead Add Form                             5.140644
Lead Source_Olark Chat                                1.527022
Lead Source_Reference                                -2.571932
Last Activity_Olark Chat Conversation                -0.960626
Last Activity_SMS Sent                                1.260701
What is your current occupation_Working Professional  1.801025
What matters most to you in choosing a course_Other  -0.573189
Lead Quality_Low in Relevance                        -0.991060
Lead Quality_Might be                                -2.116387
Lead Quality_Not Sure                                -3.568061
Lead Quality_Worst                                   -5.861452
Last Notable Activity_Modified                       -0.734616
dtype: float64
```

ROC Curve



- More area under ROC curve indicates it is good model.
- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

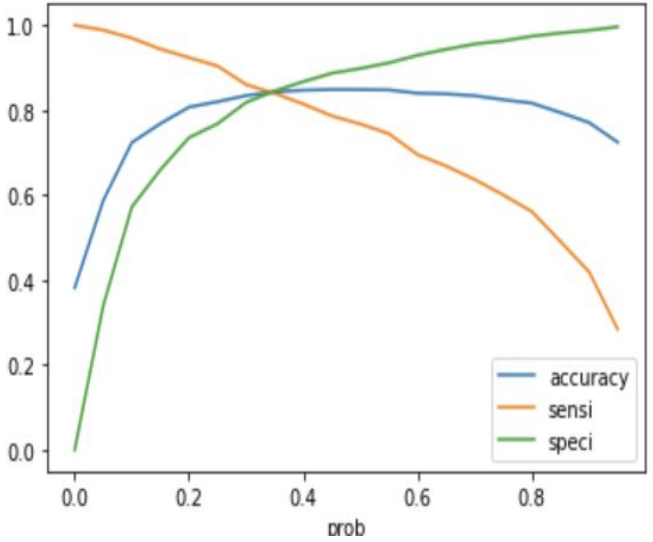
Optimal cutoff point

accuracy sensitivity and specificity for various probability cutoffs.

Values

	prob	accuracy	sensi	speci
0.00	0.00	0.381672	1.000000	0.000000
0.05	0.05	0.587152	0.988449	0.339445
0.10	0.10	0.723666	0.969472	0.571938
0.15	0.15	0.768225	0.943894	0.659791
0.20	0.20	0.807117	0.923680	0.735167
0.25	0.25	0.819871	0.903878	0.768016
0.30	0.30	0.834199	0.860149	0.818182
0.35	0.35	0.843017	0.839521	0.845174
0.40	0.40	0.846953	0.814356	0.867074
0.45	0.45	0.848528	0.786304	0.886937
0.50	0.50	0.848213	0.767327	0.898141
0.55	0.55	0.847583	0.744224	0.911383
0.60	0.60	0.839868	0.695132	0.929208
0.65	0.65	0.838293	0.668317	0.943214
0.70	0.70	0.833884	0.636551	0.955691
0.75	0.75	0.824595	0.600248	0.963076
0.80	0.80	0.816407	0.561056	0.974026
0.85	0.85	0.794206	0.490512	0.981665
0.90	0.90	0.770745	0.419142	0.987777
0.95	0.95	0.724453	0.284241	0.996180

Graph



*Looks like **0.35** is optimal cutoff point*

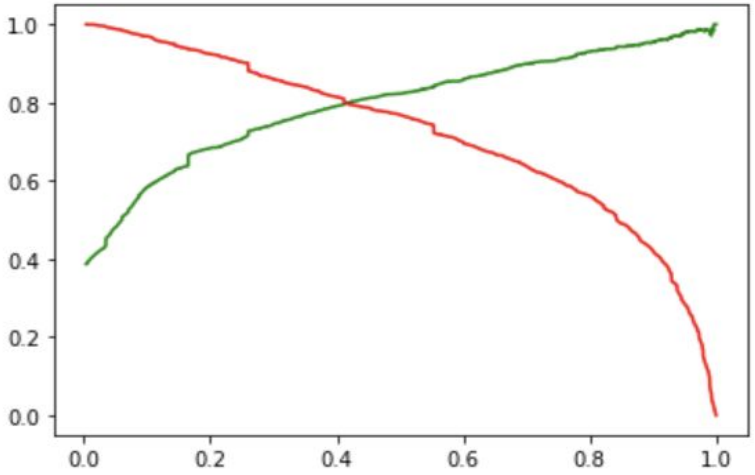
Metrics at 0.35 cutoff

```
sensitivity:
0.8395214521452146
Specificity:
0.8451744334097275
FPR:
0.15482556659027247
TPR:
0.8395214521452146
```

Good Model

TPR ↑ FPR ↓

Precision Recall Curve



Precision Recall

```
# Precision - Chances of you predicting 1
print(TP/float(TP+FP))
```

```
0.7699583806280742
```

```
#recall - Out of all yes how many you got correct
print(TP/float(TP+FN))
```

```
0.8395214521452146
```

Model Accuracy

```
accuracy
0.8430168477405133
```

Prediction on Test set

sensitivity:
0.847675568743818
Specificity:
0.8405373831775701
FPR:
0.1594626168224299
TPR:
0.847675568743818

Similar pattern is observed for Test Data set

TPR ↑ FPR ↓

#sensitivity or Recall - Out of all yes how many you got correct

Specificity - Out of all No, How many you got correct

False positive rate - predicting wrong customer as converted

True positive rate - predicting converted customer as converted

Precision Recall

Precision
0.7584070796460177
recall
0.847675568743818

Accuracy is > 80%

accuracy

0.8431876606683805