Lead Scoring Analysis

Submitted By: Ujjwal Rastogi
(DS C57)

Index

- Objective
- Analysis approach
- Data preparation: Categorial columns
- Data preparation: Numeric columns
- Dummy variables preparation
- Model building: Baseline model
- Model building: RFE
- Model Evaluation

Index

- Finding the cutoff
- Making predictions
- Conclusion

Objective

- The objective of this analysis is to train a logistic regression model and assign a lead score between 0 and 100 to each of a lead coming in to X Education, where higher score means a higher probablity of conversion and vice versa.
- This score will then be used by the sales team to target leads that have higher chance of conversion.

Analysis Approach

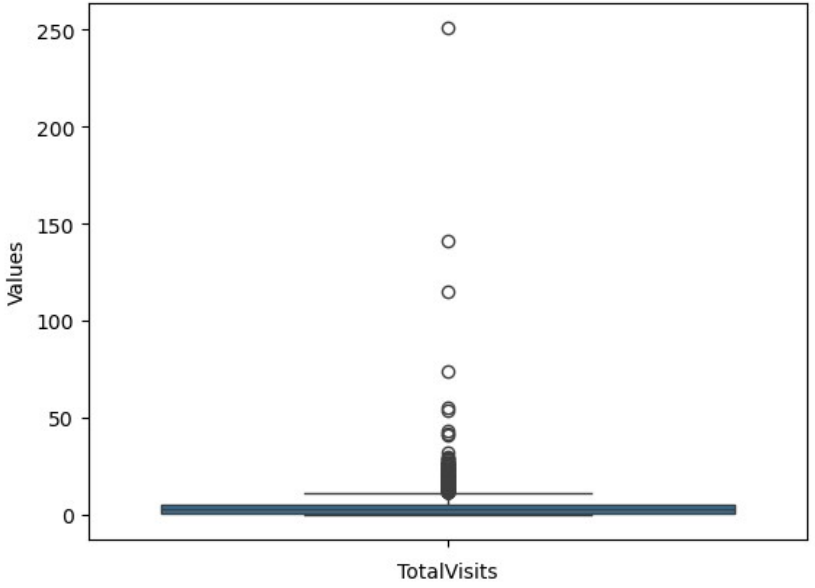
- The data is studied, Select value treated, then the columns that have high null or low variance are removed.
- Columns that have outliers get imputed with a maximum and minimum values.
- Numeric columns with low percentage of nulls are imputed with median.
- Dummy variables are then made and numeric value scaled.
- Then a model is build using RFE to identify the best features.

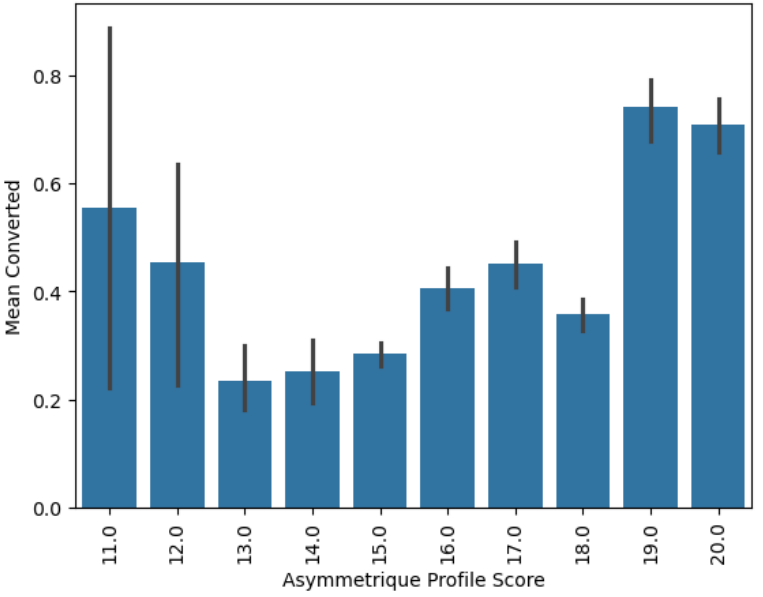
Data Preparation: Categorical columns

- Some variables contains a keyword “Select” which is replaced with null.
- Variables with high null percentage and low variance are removed.
- Tags column, which seems to have information that we don’t get when a lead comes in, is also removed.
- Some columns that have high null value but have a high converted percentage variance are not removed.

Data Preparation: Numeric columns

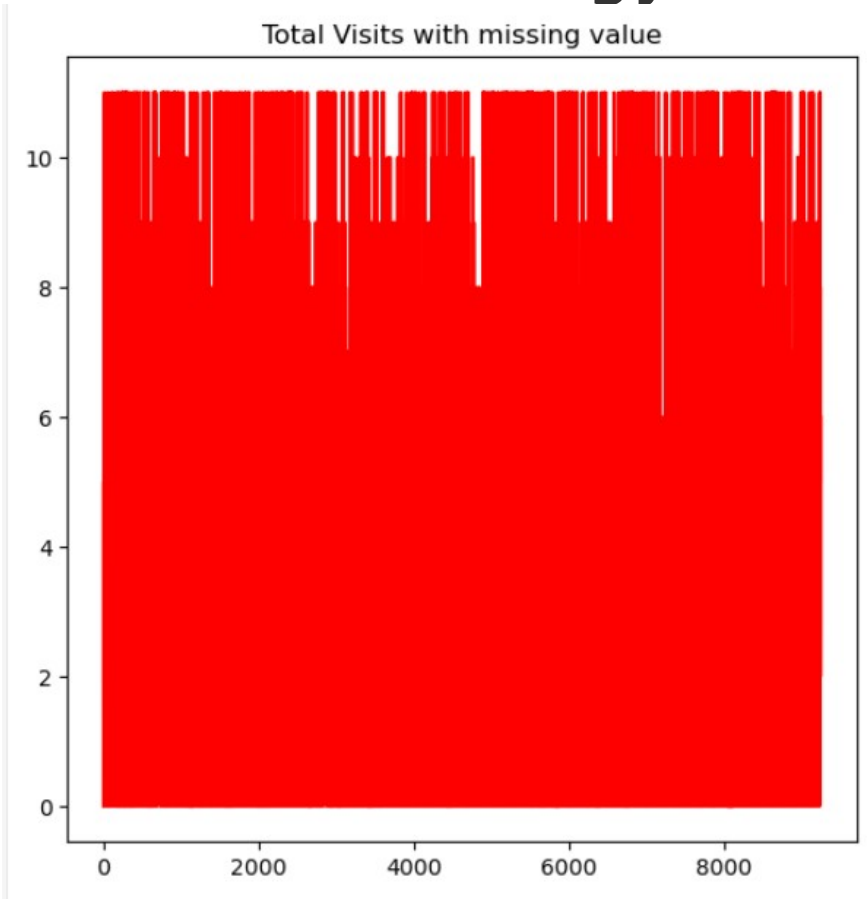
- Some columns have outliers, such columns have values capped based on a set floor and ceiling.
- “Asymmetrique Activity Score” and “Asymmetrique Profile Score” are added into a new column named “Asymmetrique Score” and the original columns are dropped.

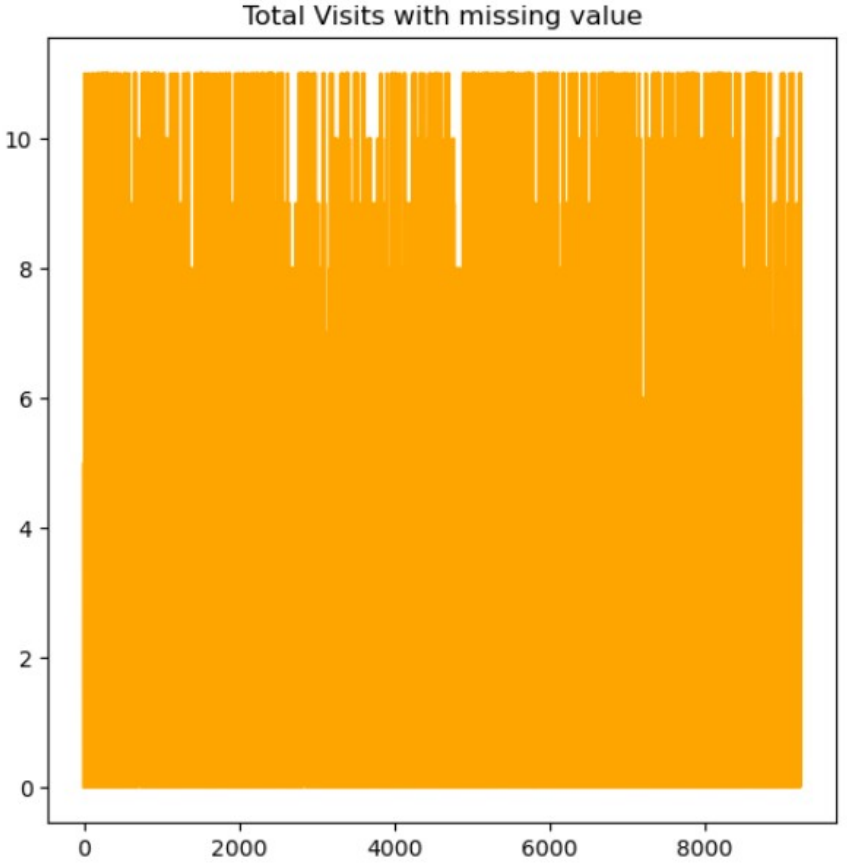




Data Preparation: Numeric columns

- Some variables with low percentage of nulls are imputed using SimpleImputer with median strategy.





Dummy Variables Preparation

- Total categorical columns are divided into two parts, one with high null values and one with no or low null values.
- In categories with high null values, null is treated as a separate column by not adding “drop_first=True” parameter.
- For low or no null categories, dummies are prepared normally, with “drop_first=True”.
- Then all these are joined back to the original dataset and the original columns are dropped.

Model building: Baseline model

- Data is split into test and train dataset using a technique called “Test-train split”. 80% of the data is used to train while 20% is used to verify the model.
- Numeric Data is then scaled to avoid any bias due to scale of data.
- After that, a baseline model is trained to get an idea about the accuracy and other metrics of the dataset.
- In this baseline mode, the train accuracy is 0.852 while test accuracy is 0.849.

Model building: RFE

After building a baseline model, Recursive Feature Elimination (RFE) is used to find out the most important feature.

```
col = X_train.columns[rfe.support_].tolist()
col
```

```
['Total Time Spent on Website',
 'What is your current occupation_Housewife',
 'What is your current occupation_Other',
 'What is your current occupation_Working Professional',
 'Lead Quality_High in Relevance',
 'Lead Quality_Low in Relevance',
 'Lead Quality_Might be',
 'Lead Quality_Worst',
 'Lead Origin_Lead Add Form',
 'Lead Source_Olark Chat',
 'Lead Source_Welingak Website',
 'Do Not Email_Yes',
 'Last Activity_Email Opened',
 'Last Activity_Resubscribed to emails',
 'Last Activity_SMS Sent',
 'Last Notable Activity_Email Opened',
 'Last Notable Activity_Had a Phone Conversation',
 'Last Notable Activity_Modified',
 'Last Notable Activity_Olark Chat Conversation',
 'Last Notable Activity_Resubscribed to emails']
```

Model building: RFE

Dep. Variable:	Converted	**No. Observations:**	7392
Model:	GLM	**Df Residuals:**	7376
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2592.3
Date:	Tue, 26 Dec 2023	**Deviance:**	5184.6
Time:	21:40:44	**Pearson chi2:**	7.99e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.4691
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.1942	0.121	-18.120	0.000	-2.432	-1.957
Total Time Spent on Website	1.3735	0.059	23.399	0.000	1.258	1.489
What is your current occupation_Working Professional	1.7989	0.198	9.070	0.000	1.410	2.188
Lead Quality_High in Relevance	3.8700	0.221	17.523	0.000	3.437	4.303
Lead Quality_Low in Relevance	2.7111	0.151	17.971	0.000	2.415	3.007
Lead Quality_Might be	1.6447	0.094	17.488	0.000	1.460	1.829
Lead Quality_Worst	-2.0191	0.335	-6.032	0.000	-2.675	-1.363
Lead Origin_Lead Add Form	3.7639	0.224	16.817	0.000	3.325	4.203
Lead Source_Olark Chat	2.6421	0.146	18.131	0.000	2.356	2.928
Lead Source_Welingak Website	3.7462	0.748	5.010	0.000	2.281	5.212
Do Not Email_Yes	-1.0375	0.175	-5.930	0.000	-1.380	-0.695
Last Activity_Email Opened	0.9506	0.164	5.795	0.000	0.629	1.272
Last Activity_SMS Sent	1.4374	0.111	12.912	0.000	1.219	1.656
Last Notable Activity_Email Opened	-0.9347	0.185	-5.049	0.000	-1.298	-0.572
Last Notable Activity_Modified	-1.1507	0.109	-10.568	0.000	-1.364	-0.937
Last Notable Activity_Olark Chat Conversation	-1.3685	0.344	-3.978	0.000	-2.043	-0.694

The model is then built using these columns and features with high VIF and p values are removed.

This is the final model we get.

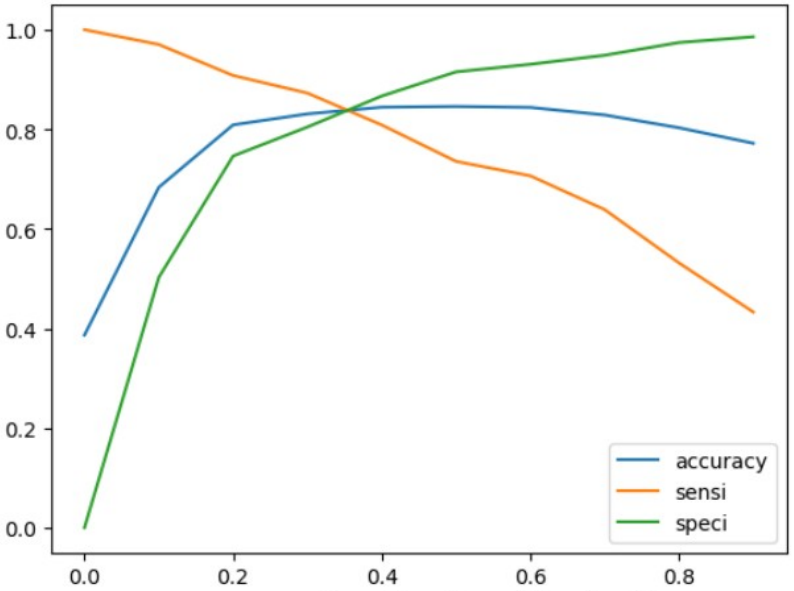
Model Evaluation

Model is evaluated using not just accuracy but other important matrices too.

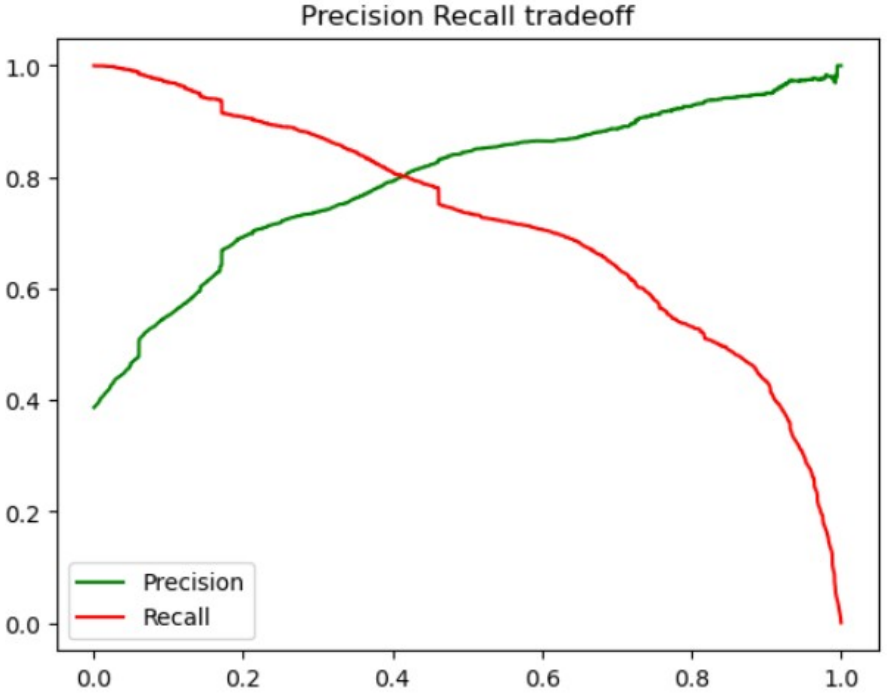
Train	precision	recall	f1-score	support
0	0.85	0.92	0.88	4533
1	0.85	0.74	0.79	2859
accuracy			0.85	7392
macro avg	0.85	0.83	0.83	7392
weighted avg	0.85	0.85	0.84	7392

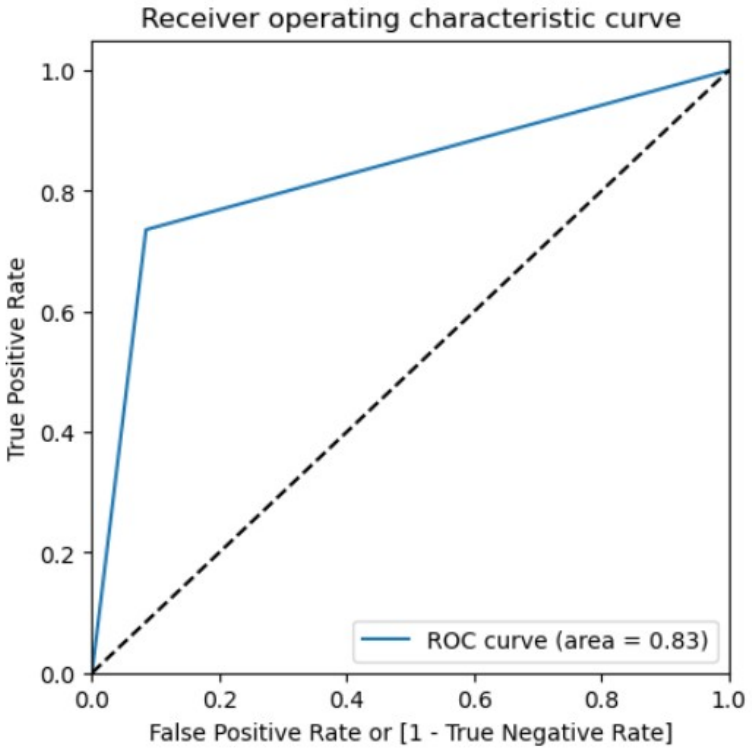
Test	precision	recall	f1-score	support
0	0.85	0.91	0.88	1146
1	0.83	0.74	0.78	702
accuracy			0.84	1848
macro avg	0.84	0.82	0.83	1848
weighted avg	0.84	0.84	0.84	1848

Finding the cutoff



By default, the cutoff is 0.5, but that may or may not be true. These plots will help us understand that better.





Finding the cutoff

Using the above plots, we can see that value of 0.4 or 40% should be the cutoff point for our model.

Train	precision	recall	f1-score	support
0	0.88	0.87	0.87	4533
1	0.79	0.81	0.80	2859
accuracy			0.84	7392
macro avg	0.84	0.84	0.84	7392
weighted avg	0.84	0.84	0.84	7392

Test	precision	recall	f1-score	support
0	0.89	0.87	0.88	1146
1	0.79	0.82	0.80	702
accuracy			0.85	1848
macro avg	0.84	0.84	0.84	1848
weighted avg	0.85	0.85	0.85	1848

Making predictions

- The above model is used to assign a Lead Score between 0 and 100 to every lead, where a high lead score means the lead is more likely to be converted.
- The cutoff can be changed depending on the organization’s goal at any given time. Higher cutoff means higher chances of conversion while a lower cutoff means a large number of leads.

Conclusion

- Lead Quality is one of the most important columns and sales team should be trained on how to assign lead quality.
- Lead coming from “Lead Add Form” and “Welingak website” has more probablity to convert. These sources should be prioritised and marking team can also use this infomation to tune their campaigns.
- Last activity and last notable activities are also a good indication of the quality of a lead.
- The organization can use different cutoff for the lead score according to their goals.

Thank you