Lead Score Case Study using Logistic Regression

Submitted by:

1. Vyankatesh Kshatriya
2. Nikunj Garg
3. Rishabh Zanwar

Content

- Problem Statement
- Problem Approach
- Exploratory Data Analysis
- Correlations
- Model Evaluation
- Observation
- Conclusion

Problem Statement

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. The individuals visit the website and fill up mandatory including few details, based on those details company contacts individuals term them as lead
- The tasks of sales/marketing team is to converted these cold leads to hot leads. The model to be build based on this information such as Total site visits, employment status, education background, etc.
- The model build will help company identify which factors are driving force in conversion of cold lead to hot leads and hence, company can focus more on those factors in future in order to increase the conversion rate to 80% as termed by CEO which currently hanging around 30%
- This will directly impact there sales, as more conversions from cold to hot leads indicates more sales and higher growth rate of company

Business Objective

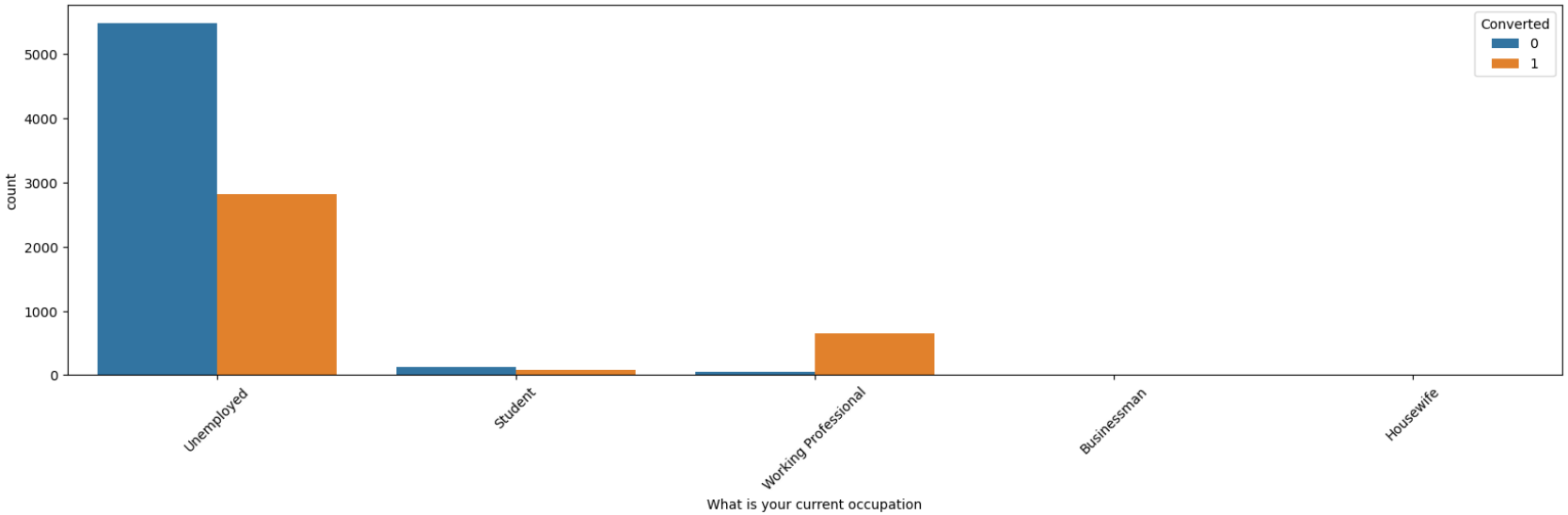
- The lead getting from data should have a score ranging from 0 to 100, as 100 indicated higher conversion rate and classified as hot lead
- The CEO of the company expected the conversion rate of 80% which is currently at 30%
- The company wants a model which ready for ingestion of data and as a output will provide the lead score for the future aspects of company as well as the past data

Problem Approach

- The choice of model building technique is Logistic Regression as the target variable is categorical binomial data
- Importing of the necessary libraries and importing of data
- Exploratory Data Analysis
  - Data understanding an data preprocessing
  - Data quality checks, data imbalance study, checking of duplicates, checking for missing values and imputation of missing values
- Dummy Variable creation, binary mapping and finding correlation
- Splitting of data into train and test
- Feature scaling
- Model building using RFE and VIF techniques
- Making prediction of train set onto the test set
- Model Evaluation

EDA – Data Cleaning

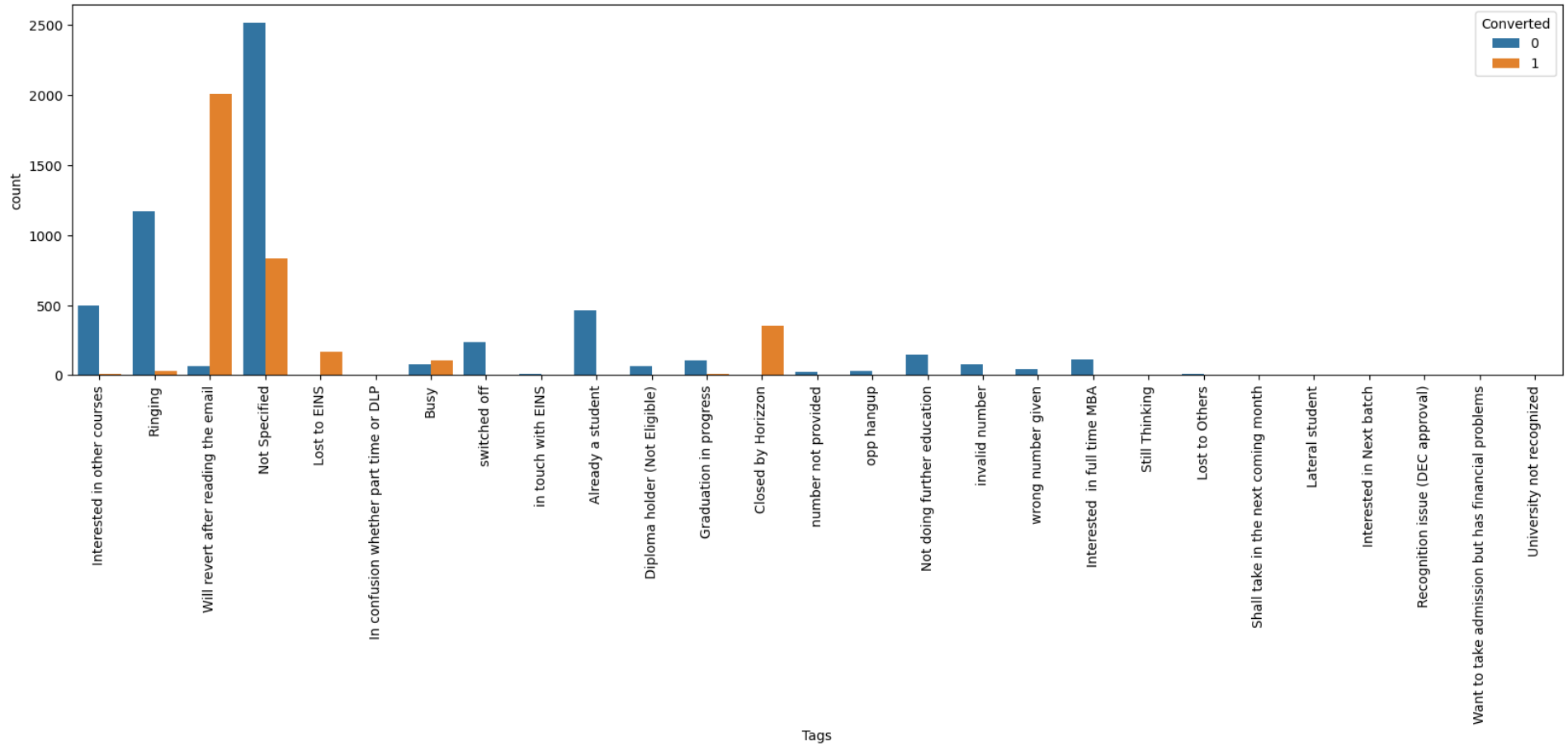
- Taking of care of columns which includes Select as the data value which is as important as Null value




```
What is your current occupation
Unemployed                  5600
NaN                         2690
Working Professional         706
Student                      210
Other                         16
Housewife                     10
Businessman                    8
Name: count, dtype: int64
```

- Converting null values and the other category values as unemployed

Tags

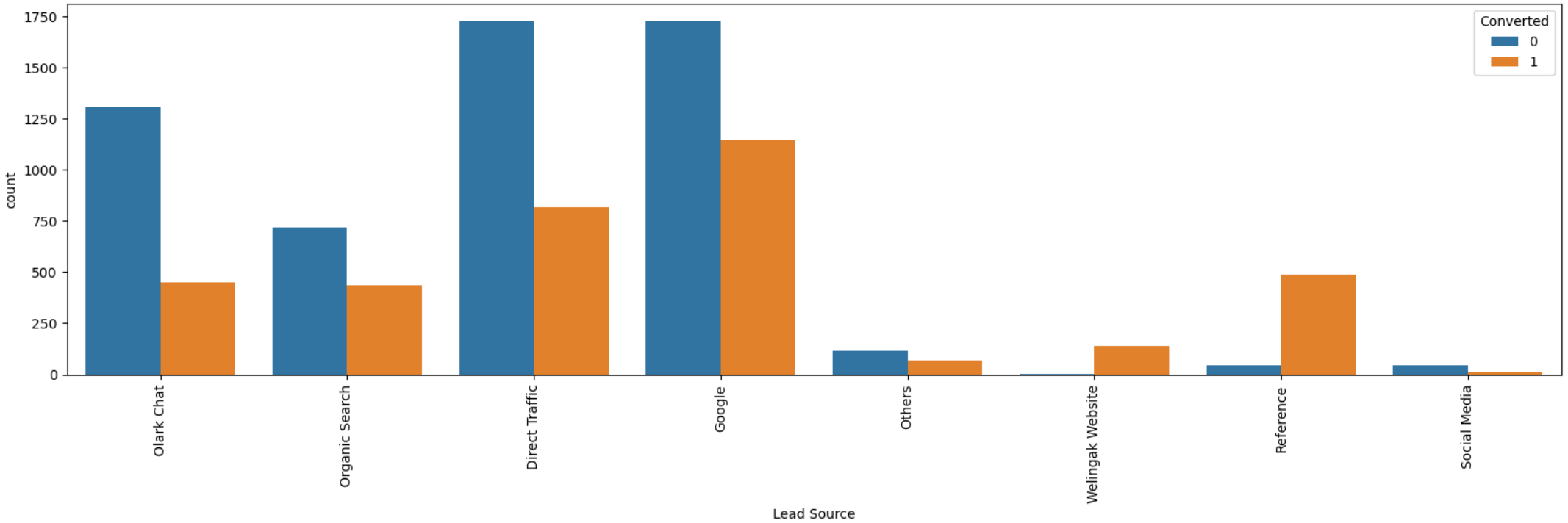


Tags	
NaN	3353
Will revert after reading the email	2072
Ringing	1203
Interested in other courses	513
Already a student	465
Closed by Horizzon	358
switched off	240
Busy	186
Lost to EINS	175
Not doing further education	145
Interested in full time MBA	117
Graduation in progress	111
invalid number	83
Diploma holder (Not Eligible)	63
wrong number given	47
opp hangup	33
number not provided	27
in touch with EINS	12
Lost to Others	7
Still Thinking	6
Want to take admission but has financial problems	6
In confusion whether part time or DLP	5
Interested in Next batch	5
Lateral student	3
Shall take in the next coming month	2
University not recognized	2
Recognition issue (DEC approval)	1

Name: count, dtype: int64

- Here, converted the null values as something meaningful as Not specified

Lead Source

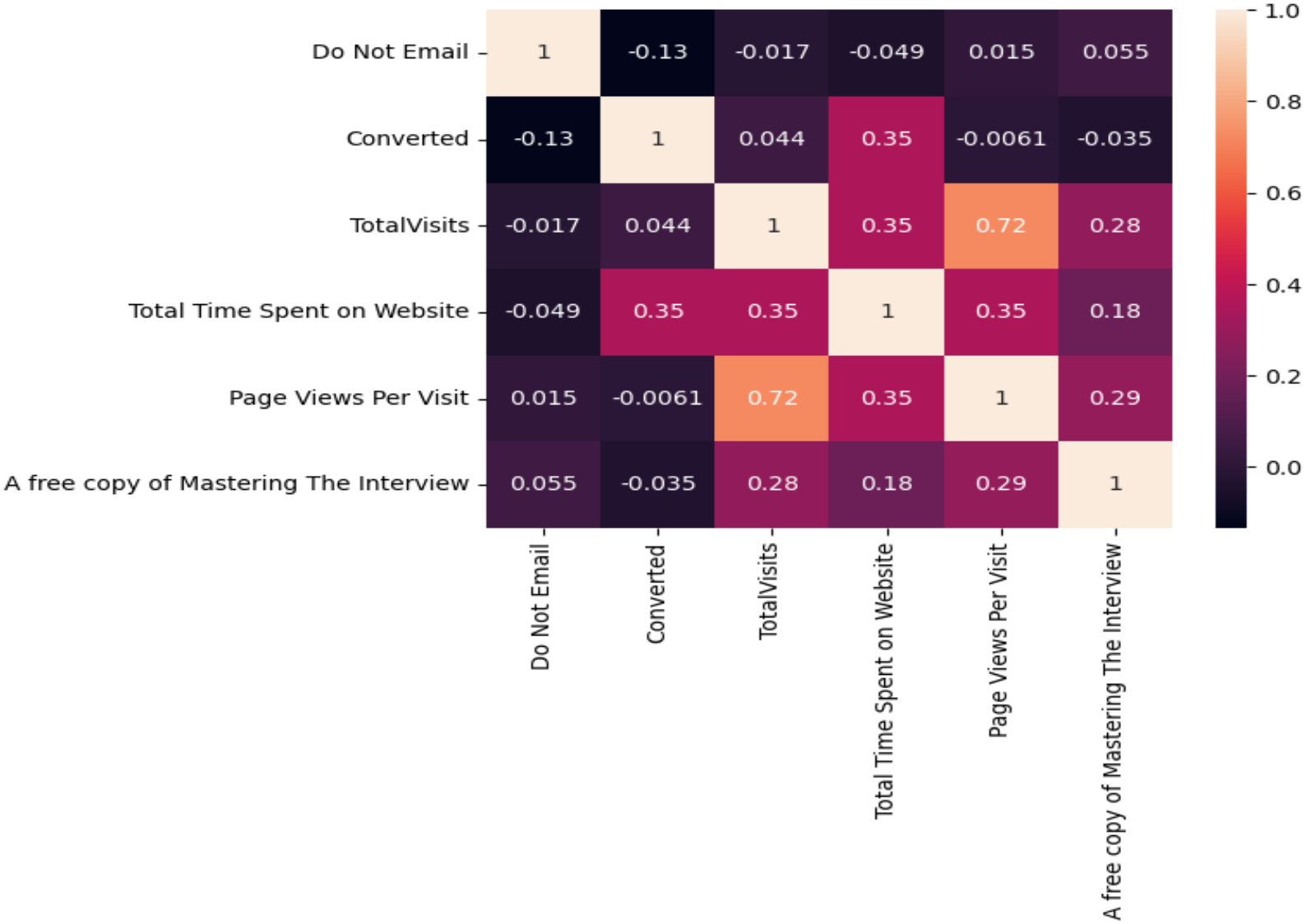


Lead Source	
Google	2868
Direct Traffic	2543
Olark Chat	1755
Organic Search	1154
Reference	534
Welingak Website	142
Referral Sites	125
Facebook	55
NaN	36
bing	6
google	5
Click2call	4
Press_Release	2
Social Media	2
Live Chat	2
youtubechannel	1
testone	1
Pay per Click Ads	1
welearnblog_Home	1
WeLearn	1
blog	1
NC_EDM	1

Name: count, dtype: int64

- There are a lot of values which are not significant when analysis taken into consideration
- The values which very less count replaced by Others

Correlation



- Correlation is plotted using heatmap method of seaborn library
- There is no significant correlation between columns except Page vies per visit and Total Visites which means same
- In the model building RFE eliminated one column as it was very much correlated with other column

Model Building

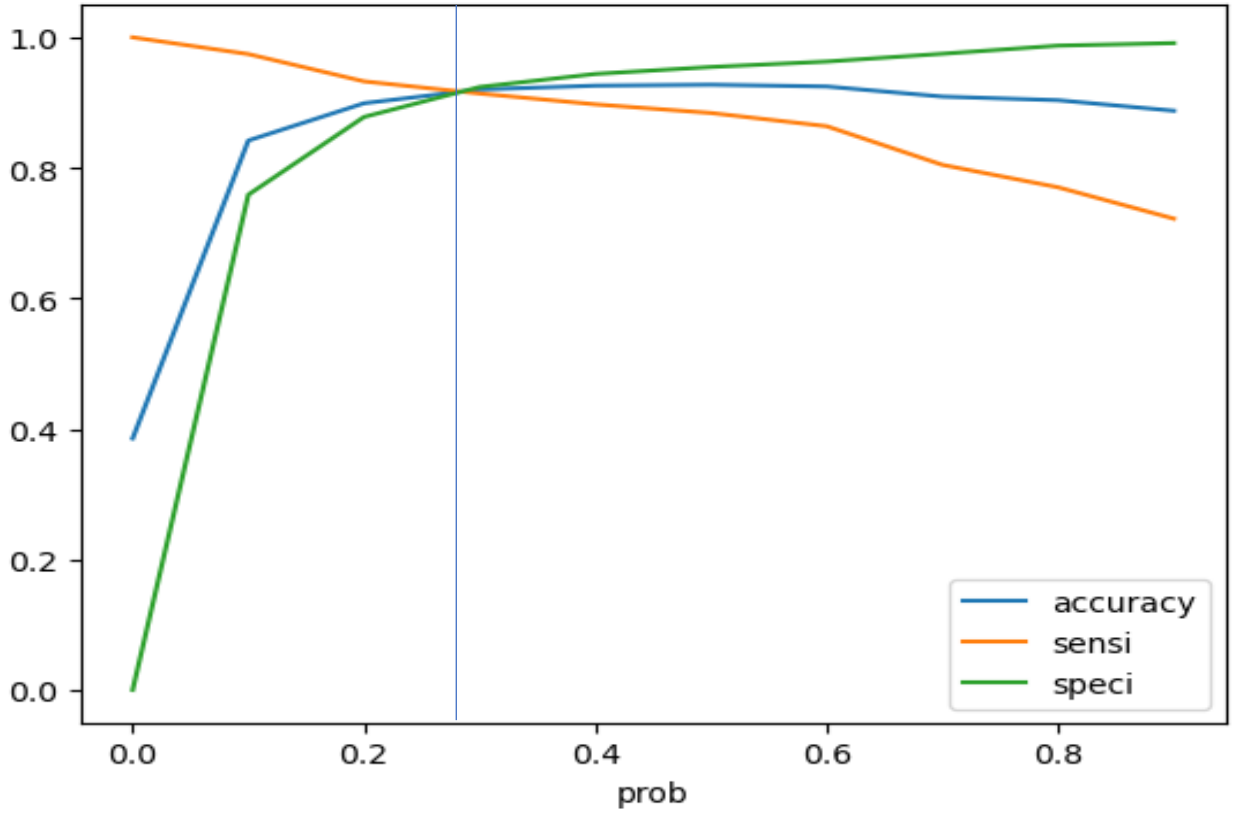
Final Model:

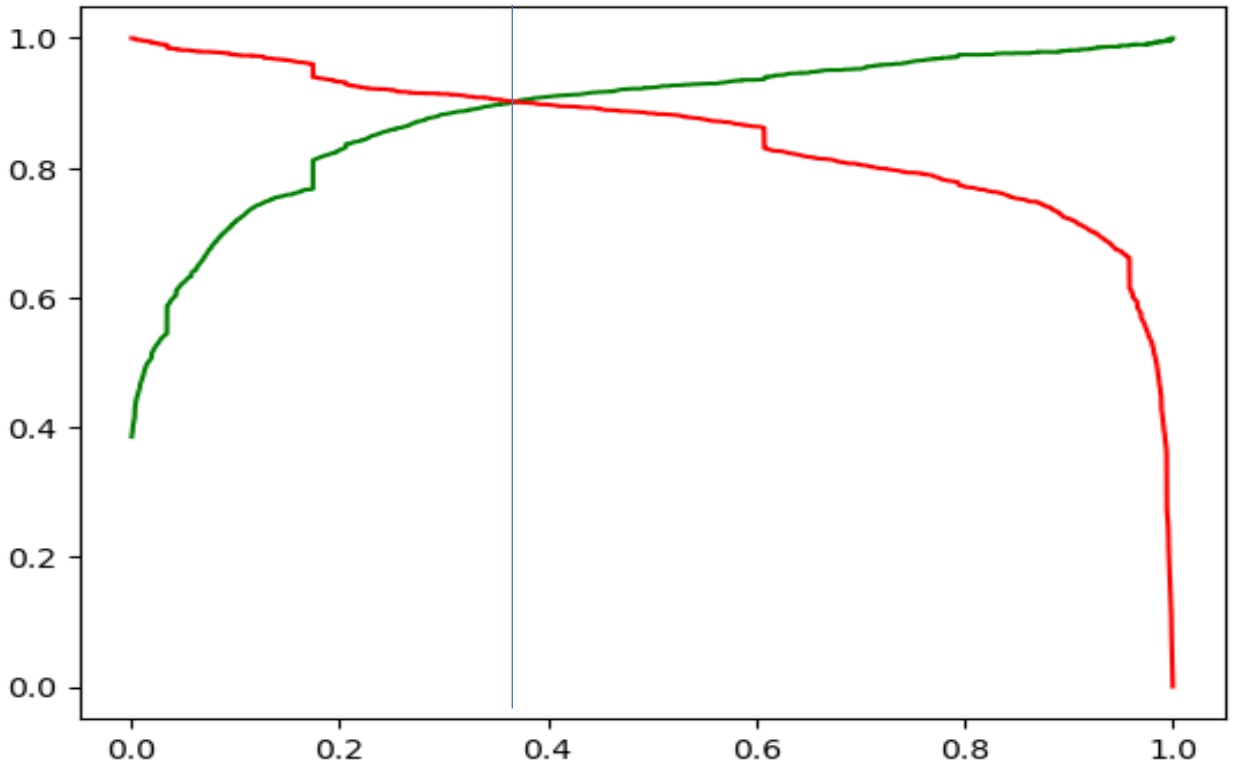
	Features	VIF
1	Lead Origin_Landing Page Submission	2.47
0	Total Time Spent on Website	2.14
6	Tags_Not Specified	1.95
9	Tags_Will revert after reading the email	1.84
10	Last Notable Activity_Modified	1.81
7	Tags_Other	1.69
3	Last Activity_SMS Sent	1.60
8	Tags_Ringing	1.40
4	Tags_Closed by Horizzon	1.13
5	Tags_Lost to EINS	1.12
2	Lead Source_Welingak Website	1.07
11	Last Notable Activity_Olark Chat Conversation	1.07

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.7414	0.235	-3.158	0.002	-1.202	-0.281
Total Time Spent on Website	3.9606	0.229	17.274	0.000	3.511	4.410
Lead Origin_Landing Page Submission	-1.1562	0.118	-9.838	0.000	-1.387	-0.926
Lead Source_Welingak Website	5.0139	1.028	4.876	0.000	2.999	7.029
Last Activity_SMS Sent	1.9909	0.115	17.241	0.000	1.765	2.217
Tags_Closed by Horizzon	7.0140	1.033	6.788	0.000	4.989	9.039
Tags_Lost to EINS	5.7405	0.760	7.549	0.000	4.250	7.231
Tags_Not Specified	-0.8128	0.222	-3.661	0.000	-1.248	-0.378
Tags_Other	-3.1765	0.278	-11.414	0.000	-3.722	-2.631
Tags_Ringing	-4.0459	0.306	-13.241	0.000	-4.645	-3.447
Tags_Will revert after reading the email	3.8731	0.274	14.154	0.000	3.337	4.409
Last Notable Activity_Modified	-1.7855	0.126	-14.126	0.000	-2.033	-1.538
Last Notable Activity_Olark Chat Conversation	-1.5443	0.410	-3.768	0.000	-2.348	-0.741

- Notably, the final model had very less (zero) p value as it shows high significance of columns
- The VIF score also shows there are no columns in the model which shows multi-collinearity of columns

Model Evaluation





- Before, finding cutoff values an arbitrary value 0.5 was taken into consideration for model evaluation
- The best cut-off values for accuracy, sensitivity and specificity was near 0.3
- After this point the probability of conversion of lead decrease significantly
- Hence, for further analysis the probability of lead conversion cut off was chose to be at 0.3

Observation

Final Observation of mode:

Test:

- Accuracy is 92%
- Specificity is 92.39%
- Sensitivity is 91.38%

Train:

- Accuracy is 92.1%
- Specificity is 91.55%
- Sensitivity is 93.06%

Final feature list in model:

- Total Time Spent on Website
- Lead Origin_Landing Page Submission
- Lead Source_Welingak Website
- Last Activity_SMS Sent
- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Tags_Not Specified
- Tags_Other
- Tags_Ringing
- Tags_Will revert after reading the email
- Last Notable Activity_Modified
- Last Notable Activity_Olark Chat Conversation

Conclusion

- The model built was above 90% accuracy which more than what CEO of X company expected
- The conversion rate of candidates source for reference was higher than any other lead source
- The leads which spent more time on website and opt for SMS notification are more prone to convert as a lead and take the course
- In the final mode, there were total of 12 features selected to build a model. The features selected were all very significant for model building and when compared to target variable (Converted)
- The VIF also showed that the features had no multi-collinearity with each other and working independently
- The accuracy, specificity and sensitivity of train and test was very close to each other at the same it above 90 which shows the quality of model which was built