LEAD SCORING CASE STUDY

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 company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Goal

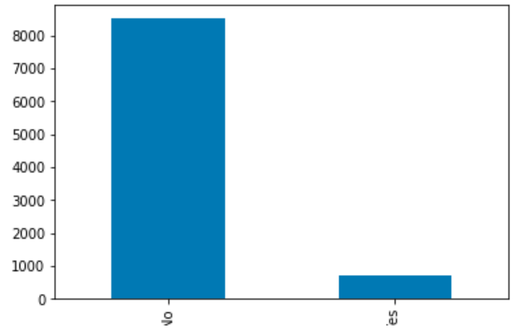
Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

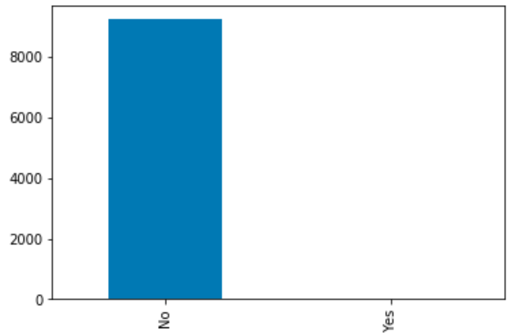
Steps Performed

- Data Import
  - Reading and Visualizing the Data
- Data Preparation and Cleaning
  - Treating Missing Values (Dropping columns with high missing percentage)
  - Removing Highly Skewed Columns
  - Treating Outliers
- Data Modelling
  - Train Test Split
  - Scaling of data
  - Iterative Modelling using coarse tuning (RFE + Manual Tuning)
  - Modelling using Logistic Regression's GLM
  - Checking key parameters like Accuracy, Sensitivity, Specificity, ROC Curve, Optimal Cutoff, Precision and Recall
- Prediction using the derived model
  - Running the model on Test Data Set
  - Checking the key parameters whether model performs good on unseen data

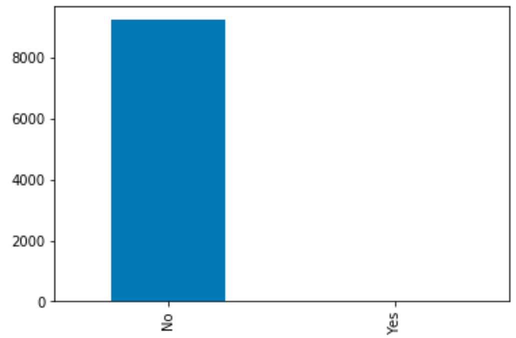
```
Name: Do Not Email, dtype: int64
```



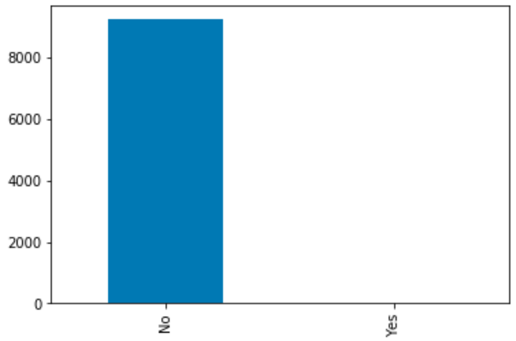
```
Name: Newspaper Article, dtype: int64
```



```
Name: X Education Forums, dtype: int64
```



```
Name: Do Not Call, dtype: int64
```

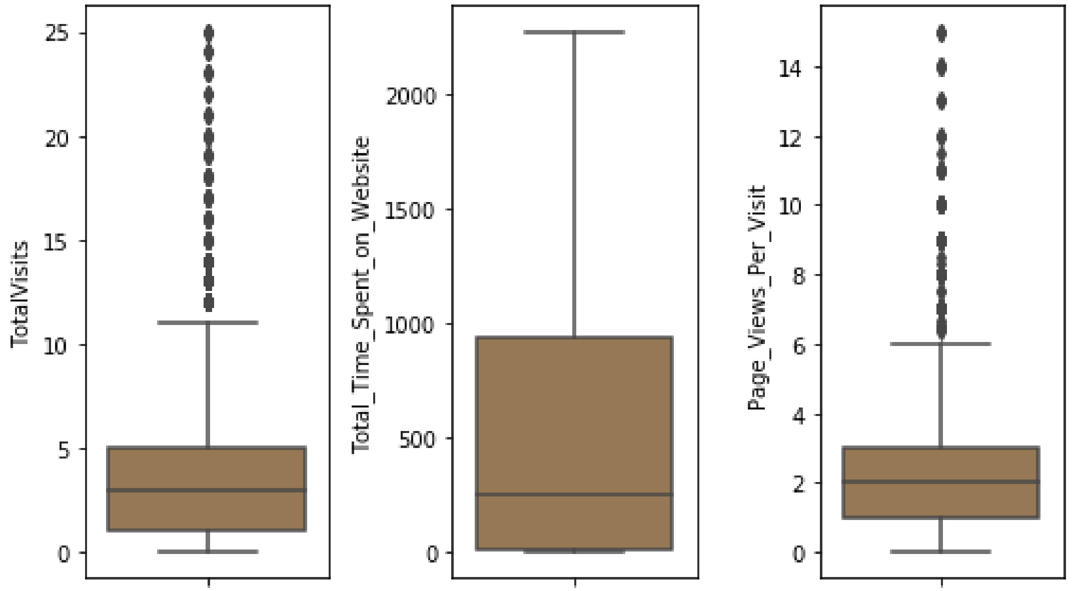


Skewness Check

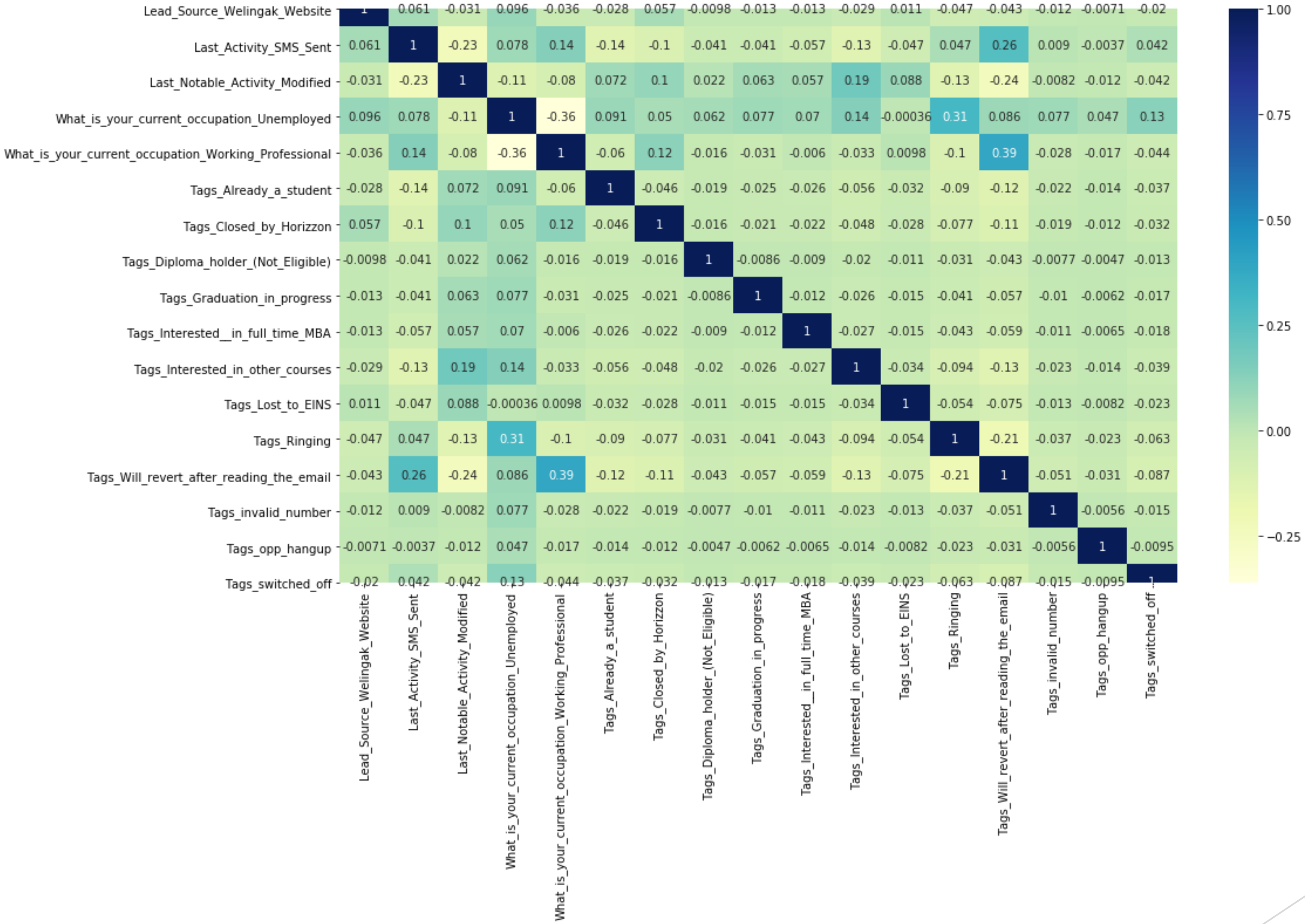
- As seen in the adjacent images, some of the columns shown here are highly skewed.
- Hence it is best to drop them

Outlier Treatment

- As seen from adjacent box plots, Total Visits and Page Views Per Visit have outliers.
- It is imperative to remove them.

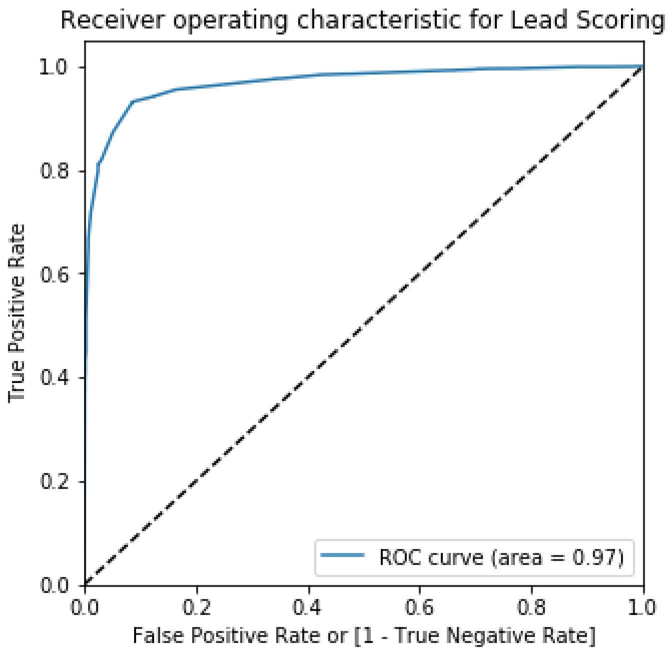


Correlations on Chosen Features

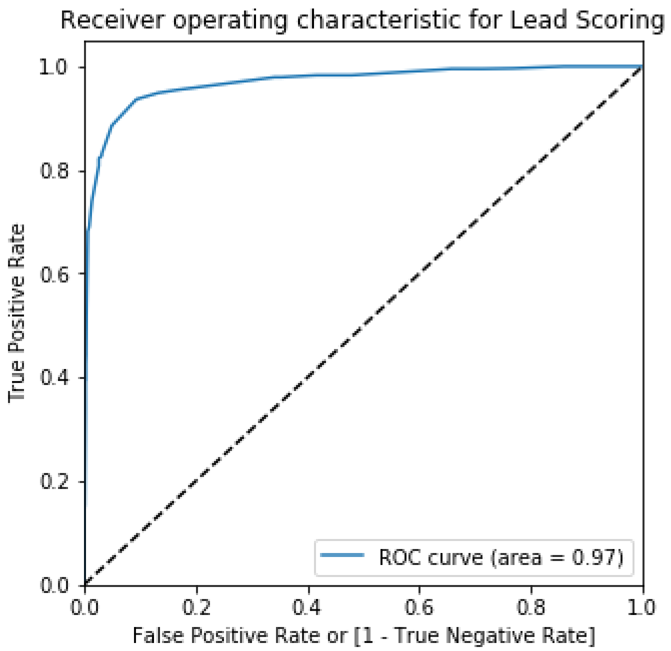


	Lead_Source_Welingak_Website	Last_Activity_SMS_Sent	Last_Notable_Activity_Modified	What_is_your_current_occupation_Unemployed	What_is_your_current_occupation_Working_Professional	Tags_Already_a_student	Tags_Closed_by_Horizzon	Tags_Diploma_holder_(Not_Eligible)	Tags_Graduation_in_progress	Tags_Interested__in_full_time_MBA	Tags_Interested_in_other_courses	Tags_Lost_to_EINS	Tags_Ringing	Tags_Will_revert_after_reading_the_email	Tags_invalid_number	Tags_opp_hangup	Tags_switched_off
Lead_Source_Welingak_Website	1	0.061	-0.031	0.096	-0.036	-0.028	0.057	-0.0098	-0.013	-0.013	-0.029	0.011	-0.047	-0.043	-0.012	-0.0071	-0.02
Last_Activity_SMS_Sent	0.061	1	-0.23	0.078	0.14	-0.14	-0.1	-0.041	-0.041	-0.057	-0.13	-0.047	0.047	0.26	0.009	-0.0037	0.042
Last_Notable_Activity_Modified	-0.031	-0.23	1	-0.11	-0.08	0.072	0.1	0.022	0.063	0.057	0.19	0.088	-0.13	-0.24	-0.0082	-0.012	-0.042
What_is_your_current_occupation_Unemployed	0.096	0.078	-0.11	1	-0.36	0.091	0.05	0.062	0.077	0.07	0.14	-0.00036	0.31	0.086	0.077	0.047	0.13
What_is_your_current_occupation_Working_Professional	-0.036	0.14	-0.08	-0.36	1	-0.06	0.12	-0.016	-0.031	-0.006	-0.033	0.0098	-0.1	0.39	-0.028	-0.017	-0.044
Tags_Already_a_student	-0.028	-0.14	0.072	0.091	-0.06	1	-0.046	-0.019	-0.025	-0.026	-0.056	-0.032	-0.09	-0.12	-0.022	-0.014	-0.037
Tags_Closed_by_Horizzon	0.057	-0.1	0.1	0.05	0.12	-0.046	1	-0.016	-0.021	-0.022	-0.048	-0.028	-0.077	-0.11	-0.019	-0.012	-0.032
Tags_Diploma_holder_(Not_Eligible)	-0.0098	-0.041	0.022	0.062	-0.016	-0.019	-0.016	1	-0.0086	-0.009	-0.02	-0.011	-0.031	-0.043	-0.0077	-0.0047	-0.013
Tags_Graduation_in_progress	-0.013	-0.041	0.063	0.077	-0.031	-0.025	-0.021	-0.0086	1	-0.012	-0.026	-0.015	-0.041	-0.057	-0.01	-0.0062	-0.017
Tags_Interested__in_full_time_MBA	-0.013	-0.057	0.057	0.07	-0.006	-0.026	-0.022	-0.009	-0.012	1	-0.027	-0.015	-0.043	-0.059	-0.011	-0.0065	-0.018
Tags_Interested_in_other_courses	-0.029	-0.13	0.19	0.14	-0.033	-0.056	-0.048	-0.02	-0.026	-0.027	1	-0.034	-0.094	-0.13	-0.023	-0.014	-0.039
Tags_Lost_to_EINS	0.011	-0.047	0.088	-0.00036	0.0098	-0.032	-0.028	-0.011	-0.015	-0.015	-0.034	1	-0.054	-0.075	-0.013	-0.0082	-0.023
Tags_Ringing	-0.047	0.047	-0.13	0.31	-0.1	-0.09	-0.077	-0.031	-0.041	-0.043	-0.094	-0.054	1	-0.21	-0.037	-0.023	-0.063
Tags_Will_revert_after_reading_the_email	-0.043	0.26	-0.24	0.086	0.39	-0.12	-0.11	-0.043	-0.057	-0.059	-0.13	-0.075	-0.21	1	-0.051	-0.031	-0.087
Tags_invalid_number	-0.012	0.009	-0.0082	0.077	-0.028	-0.022	-0.019	-0.0077	-0.01	-0.011	-0.023	-0.013	-0.037	-0.051	1	-0.0056	-0.015
Tags_opp_hangup	-0.0071	-0.0037	-0.012	0.047	-0.017	-0.014	-0.012	-0.0047	-0.0062	-0.0065	-0.014	-0.0082	-0.023	-0.031	-0.0056	1	-0.0095
Tags_switched_off	-0.02	0.042	-0.042	0.13	-0.044	-0.037	-0.032	-0.013	-0.017	-0.018	-0.039	-0.023	-0.063	-0.087	-0.015	-0.0095	1

ROC Curve



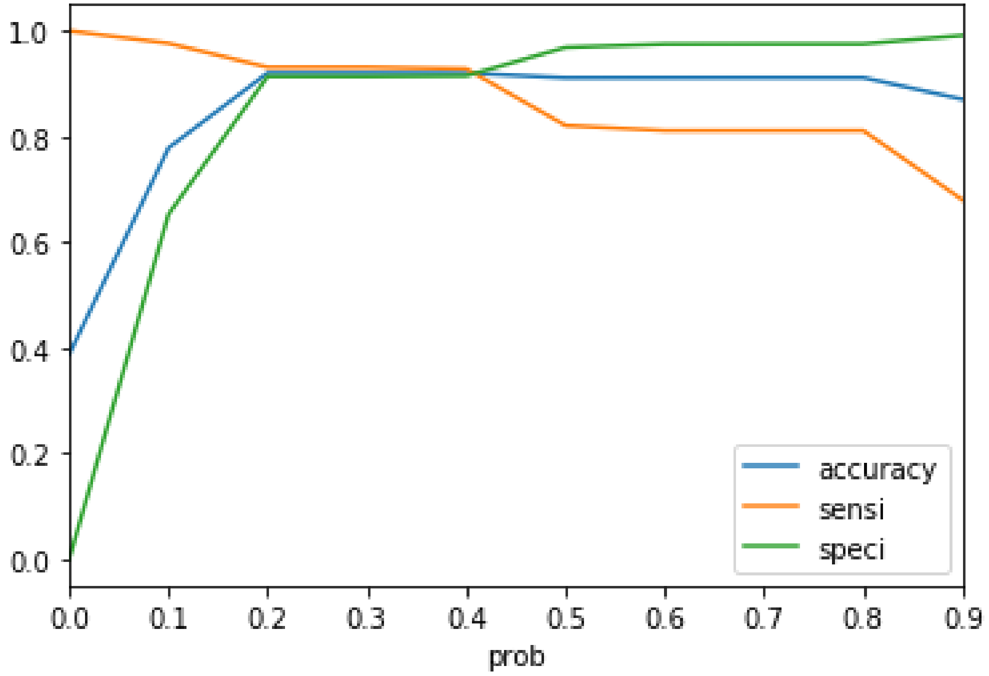
Train Data Set



Test Data Set

ROC Curve is pretty good for both train and test data

Optimal Cutoff



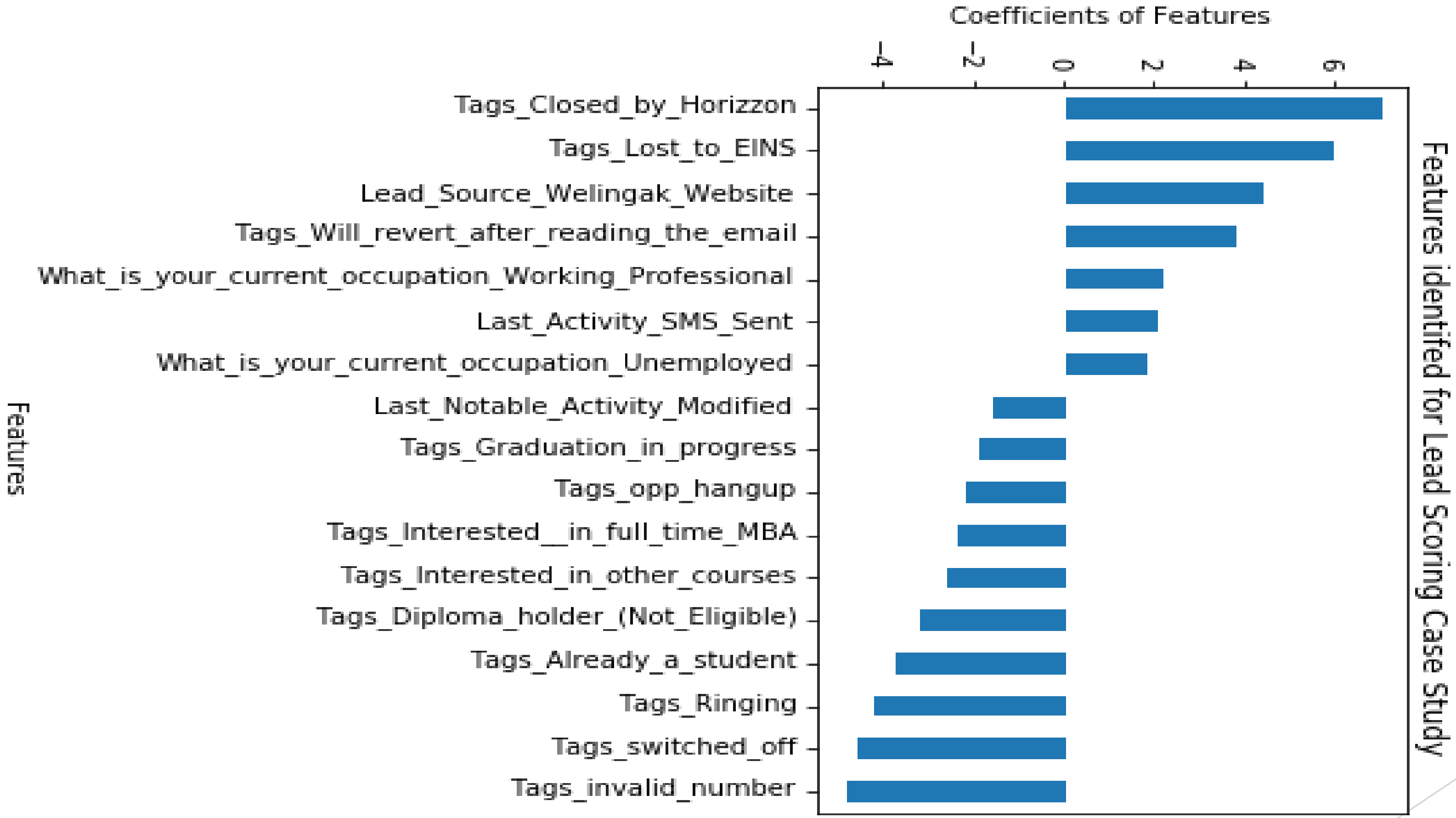
As the sensitivity, specificity and accuracy intersects from x: 0.2 to x:0.4, we are choosing 0.3 as optimal cutoff point

Key Parameters Compared

Parameters	Train Data	Test Data
Accuracy	92.07%	91.78%
Sensitivity	93.12%	93.64%
Specificity	91.42%	90.64%
Precision	94.30%	86.03%

F1 Score = 0.8967

Feature Importance Decoded



Recommendations

- As seen from the bar plot, Tags_Closed_by_Horizzon feature has highest conversion rate
- It is followed by Tags_Lost_to_EINS
- Lead_Source_Welingak_Website contributes the next most to the successful conversion rate
- On the other hand, whoever provided an invalid number, they tend to never convert into a student.
- Most of the respondents whose contact number is invalid or the phone is switched off or never take a ring, may not be converted.
- So the X Education marketing representatives should not call these respondents more than once, which results in better focus on others
- Instead they should focus more on people, who report 'Closed by Horizzon' or 'Lost to EINS' for better conversion rate.
- They can also promote the Welingak Website on various platforms including Social media, which will significantly increase their conversion rate