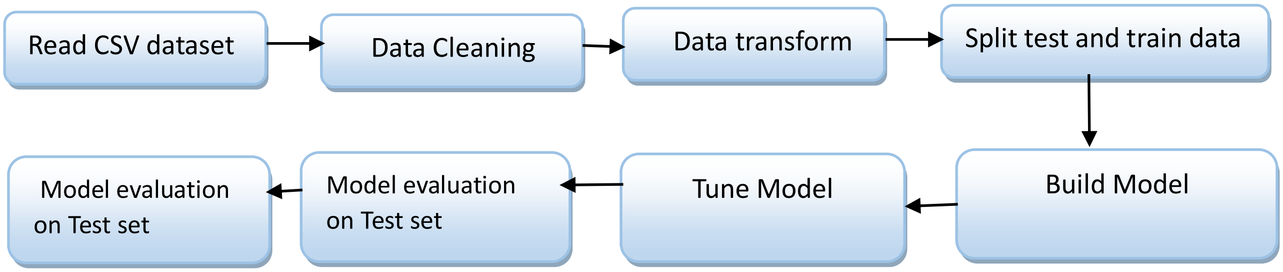
SUMMARY

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

Flow Diagram:


Read CSV dataset
Data Cleaning
Data transform
Split test and train data
Build Model
Tune Model
Model evaluation on Test set
Model evaluation on Test set

Cleaning data:

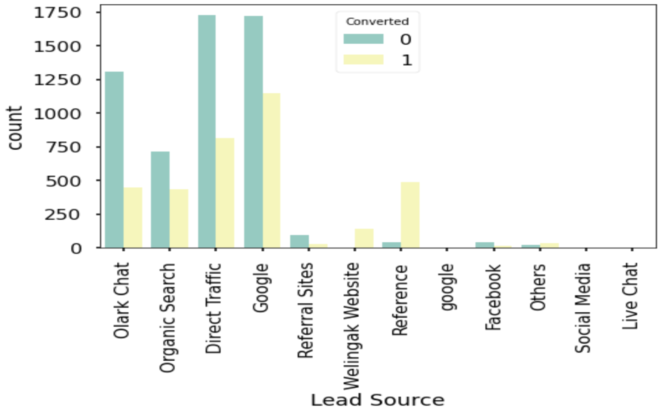
The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to 'not provided' so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided'.

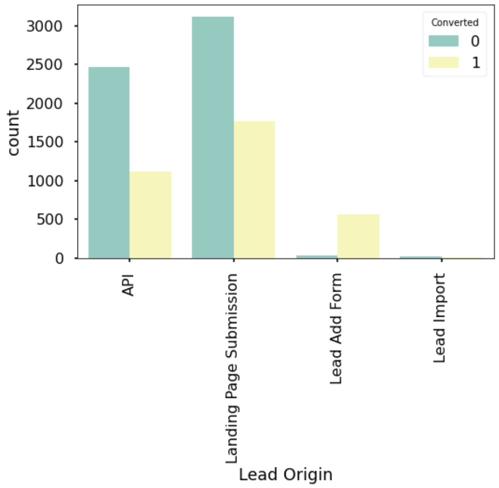
The below some highlights about data cleaning:

- Dataset contain 37 column and 9240 rows include all NaN
- Converted the categorical field NaN as 'select'
- Observed and dropped the column has more than 45 % of NaN
- Dropped the column has Imbalanced data such as ('Through Recommendations', 'Receive More Updates About Our Courses',etc).
- Removed "Prospect ID" and "Lead Number" for analysis.

EDA:

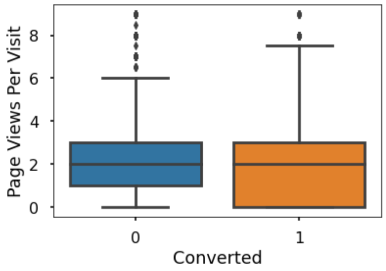
A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good and no outliers were found.

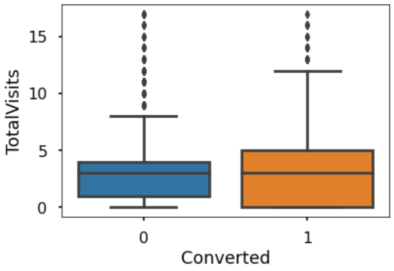




Observation,

- Max number of lead from Google link and followed by Direct traffic
- Most conversion of leads from the welingak website
- API and Landing Page Submission bring higher number of leads as well as conversion.





Numerical Value Observation,

- Median of Page views per visit for both leads are about same
- Same for the total visits for both leads.

Data Transform:

Dropping Column: Few columns are dropped due to imbalance in the dataset

Dummy Variables: The dummy variables were created and later on the dummies with 'not provided' elements were removed. For numeric values we used the MinMaxScaler. Few example variables are ['Lead Origin','What is your current occupation','City']

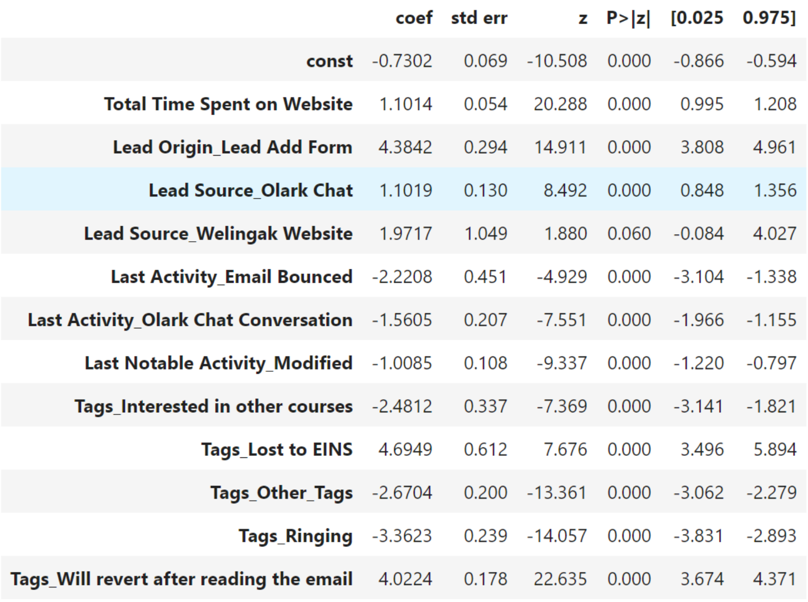
After the transformation,

- Total rows for analysis 8953
- Total column for analysis 59

Model Building:

Train and test data split: The split was done at 70% and 30% for train and test data respectively.

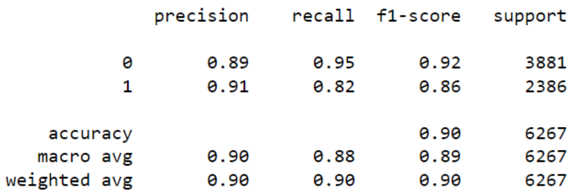
Model configuration: Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).



	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.7302	0.069	-10.508	0.000	-0.866	-0.594
Total Time Spent on Website	1.1014	0.054	20.288	0.000	0.995	1.208
Lead Origin_Lead Add Form	4.3842	0.294	14.911	0.000	3.808	4.961
Lead Source_Olark Chat	1.1019	0.130	8.492	0.000	0.848	1.356
Lead Source_Welingak Website	1.9717	1.049	1.880	0.060	-0.084	4.027
Last Activity_Email Bounced	-2.2208	0.451	-4.929	0.000	-3.104	-1.338
Last Activity_Olark Chat Conversation	-1.5605	0.207	-7.551	0.000	-1.966	-1.155
Last Notable Activity_Modified	-1.0085	0.108	-9.337	0.000	-1.220	-0.797
Tags_Interested in other courses	-2.4812	0.337	-7.369	0.000	-3.141	-1.821
Tags_Lost to EINS	4.6949	0.612	7.676	0.000	3.496	5.894
Tags_Other_Tags	-2.6704	0.200	-13.361	0.000	-3.062	-2.279
Tags_Ringing	-3.3623	0.239	-14.057	0.000	-3.831	-2.893
Tags_Will revert after reading the email	4.0224	0.178	22.635	0.000	3.674	4.371

Model Evaluation: A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around average 80% each.

Evaluation against Trains data Set:



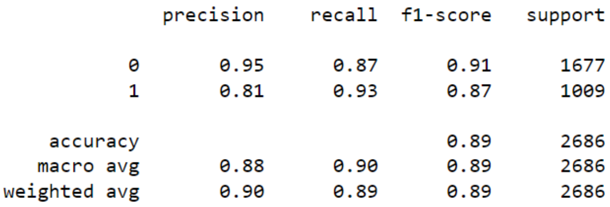
```
              precision    recall  f1-score   support

           0       0.89      0.95      0.92      3881
           1       0.91      0.82      0.86      2386

    accuracy                           0.90      6267
   macro avg       0.90      0.88      0.89      6267
weighted avg       0.90      0.90      0.90      6267
```

Accuracy : 90.00%
Sensitivity : 81.60%
Specificity : 95.33%

Evaluation against Test data Set:



```
              precision    recall  f1-score   support

           0       0.95      0.87      0.91      1677
           1       0.81      0.93      0.87      1009

    accuracy                           0.89      2686
   macro avg       0.88      0.90      0.89      2686
weighted avg       0.90      0.89      0.89      2686
```

Accuracy : 89.00%
Sensitivity : 92.96%

Specificity : 86.88%

Conclusion :

We were able to achieve **90% accuracy average on both train and test set**. With the use of this model, X Education will be able to better plan their company plans and develop by concentrating more on the elements that are most important to them.

It was discovered that the following factors, in descending order, were the most important to the potential buyers:

- Welingak website
- Google
- Direct traffic
- Organic search