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. And the main purpose of analysis is to find way or factors that influence the lead score the most and can help achieve the lead conversion score of 80% which is currently 30%.

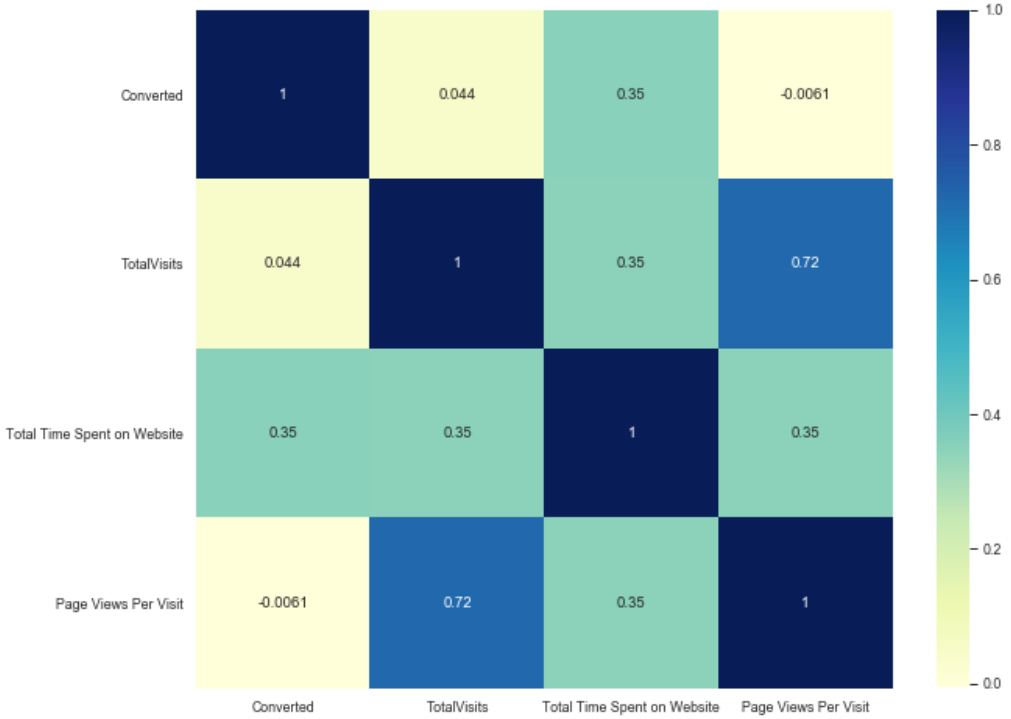
The following are the steps used:

1. **Cleaning data:**

   Following steps were taken for cleaning of data:
   1. Some column have 'Select' value and this means that nothing was selected for this column and is equivalent to Null, hence 'Select' was replaced by null.
   2. There are few columns that have relatively high value of nulls and thus these columns cannot help in prediction, hence columns with more that 39% null values were dropped.
   3. There were few categorical columns that were highly skewed, like 'What matters most to you in choosing a course' were dropped.
   4. Few categorical columns which have a high number of category but some categories had a relatively less number of rows and hence we combined categories with less percentage of rows into a single category.
   5. Columns with less percentage of missing values were imputed.
   6. Columns with outlier were subjected to outlier treatment.

2. **EDA:**

   Following steps were taken for quick EDA check:
   1. Numerical variables are not highly correlated, very less correlated.



   2. We analysed categorical and numerical variable and compared with 'Converted rate' and few conclusions were drawn like :

- Maximum number of leads is generated by Google and Direct traffic.
- Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.
- Leads spending more time on the website are more likely to be converted. Thus, Website should be made more engaging to make leads spend more time.
- Unemployed people are mostly the most approachable one.
- Working Professionals are most likely to be converted and reaching out them will increase the lead score.

3. Dummy Variables:

Following steps were taken for dummy variable creation:
1. The dummy variables were created for all categorical columns and later original categorical columns were removed.
2. For numeric values we used the MinMaxScaler for columns, 'TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website'

4. Train-Test split:

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

5. Feature Selection:

Following steps were taken for Feature selection:
1. We use RFE for variable selection and with the help of RFE we selected top 20 columns.
2. Secondly, we used manual approach for Feature selection and eliminated all the column that p-value more than 0.05
3. We also involved VIF for checking multicollinearity and also eliminated columns with VIF more than 3.

6. Model Building:

Following steps were taken for Model Building:
1. We used logistic regression approach for model building and predicted the probability of each lead to be converted to 1.
2. We choose multiple cut off for probability score and selected 0.34 cut off for probability

7. Model Evaluation:

Following steps were taken for Model Evaluation:
1. We made use of confusion matrix was made.
2. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity for Train set.
3. Final made prediction on test set as well and for test set also on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity

8. Final Conclusion:

Following conclusion can be drawn:

1. If Lead Origin is add form then probability of lead getting converted increases.
2. The students spending more time on the website have the higher chances of being converted.
3. If Lead source is welingak Website and olark chat, then probability of conversion is high.
4. The students visiting the platform more no of times have the higher chances of being converted.
5. If last activity is email opened, that means lead have shown some interest and is more likely to be converted.
6. If the specialization is not specified then chances of conversion to Lead become very low but from EDA we found out that Working professionals are more likely to join the course so company can create some campaign for working professionals.
7. Those candidate who are opting out of email service are less likely to be converted

Train Data Observation:
So as we can see above the model seems to be performing well. The ROC curve has a value of 0.87, which is very good.
We have the following values for the Train Data:

- Accuracy : 78.45%
- Sensitivity : 80.83%
- Specificity : 77.74%

Test Data Observation: So as we can see above the model seems to be performing well. The ROC curve has a value of 0.80, which is very good.
We have the following values for the Test Data:

- Accuracy : 79.15%
- Sensitivity : 81.45%
- Specificity : 77.74%

The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model