LEAD SCORING CASE STUDY

Building a Logistic Regression Model for X Education, an online education company using Python libraries.

GROUP

By – Anupam Vaibhav, R V Gokul and Nisha Venkatesh

Cohort – DSC44

Batch ID- 3161

Approach

- Datasets were imported and Python libraries – Pandas, Numpy, Matplotlib.plotly, Seaborn, Statsmodels and sklearn were used for analysis.
- Data was observed, cleaned and analysed using imputations, deletion, outlier treatments, data imbalance, Univariate and Multivariate analysis.
- Model was built using MultiClass Logistic Regression

Data Understanding

1. There were 9240 rows and 37 columns initially.
2. 30 of the columns contained string/object type data, rest were numerical.
3. ‘Converted’ column as Target column
4. Initial conversion rate at 38%.

Data Cleaning

- 5 columns, eg. Magazine with no variance were deleted.
- Columns with more than 30% values either missing or having ‘Select’ as value were deleted. Eg. How did you hear about X Education
- Columns having 2% to 30% missing values were imputed with ‘Other’ or ‘Unknown’ for categorical columns.
- Rows of variables such as TotalVisits with less than 2% missing values were deleted.

Data Cleaning

- Categorical variables such as ‘Country’ having large number of unique value but having less than 8% contribution were merged as Other.

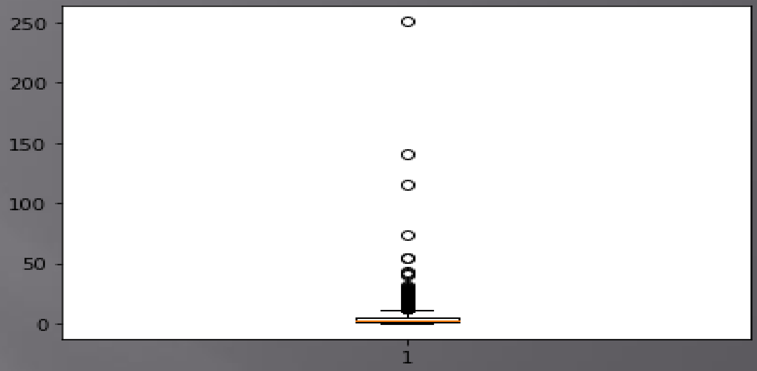
Outliers

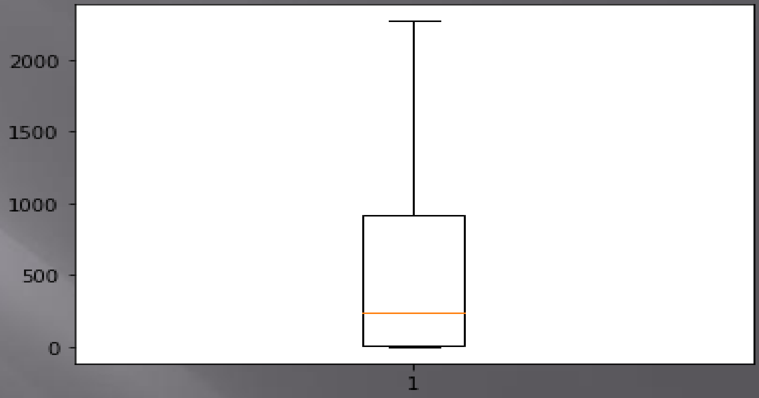
For numerical columns , Boxplot were made and outliers were analysed.

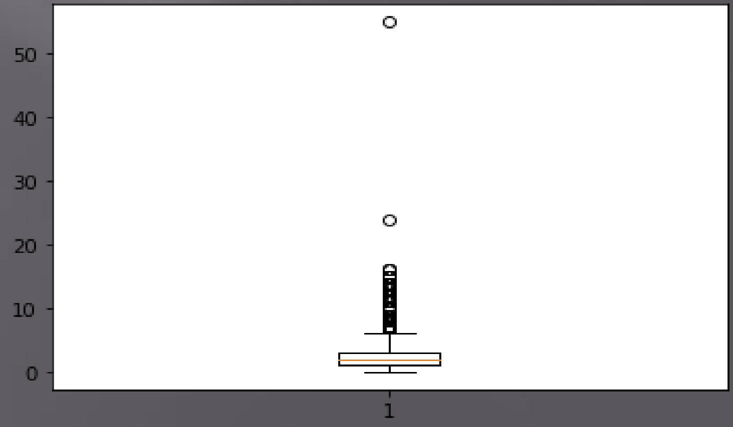
For, TotalVisits column, 25 was set as cutoff.

For, Total Time Spent on Website , no outlier treatment was required.

For, Page Views Per Visit col, 10 was set as cutoff.



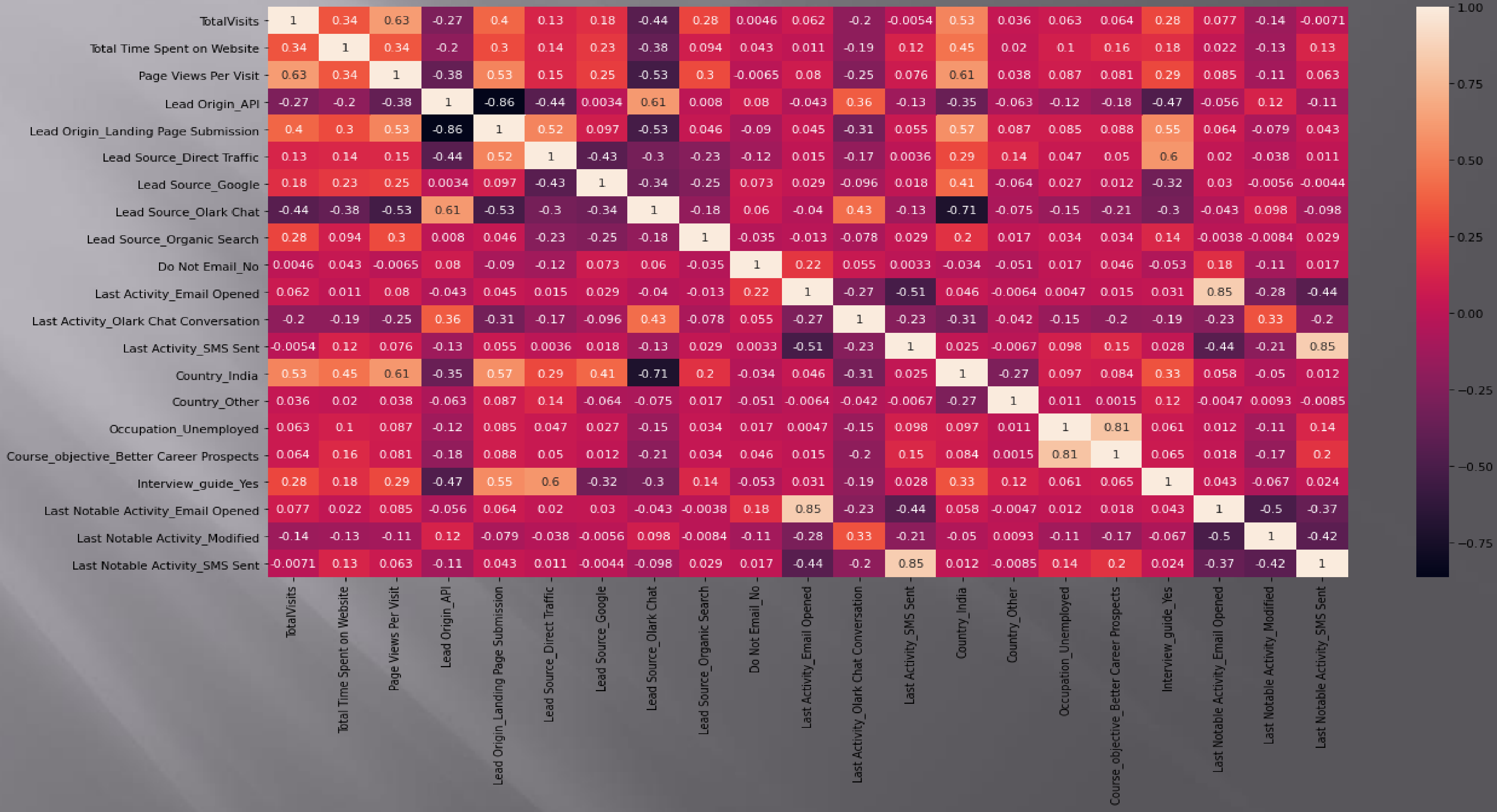




Splitting, Dummy Creation and Scaling

- Test-Train split was done as 70-30%.
- Train set numerical variables were fit and scaled using Standardisation.
- Test set numerical variables were only scaled using Standardisation.
- Dummy variables were created for rest of the categorical columns and one was dropped from each.

Correlations



	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_API	Lead Origin_Landing Page Submission	Lead Source_Direct Traffic	Lead Source_Google	Lead Source_Olark Chat	Lead Source_Organic Search	Do Not Email_No	Last Activity_Email Opened	Last Activity_Olark Chat Conversation	Last Activity_SMS Sent	Country_India	Country_Other	Occupation_Unemployed	Course_objective_Better Career Prospects	Interview_guide_Yes	Last Notable Activity_Email Opened	Last Notable Activity_Modified	Last Notable Activity_SMS Sent
TotalVisits	1	0.34	0.63	-0.27	0.4	0.13	0.18	-0.44	0.28	0.0046	0.062	-0.2	-0.0054	0.53	0.036	0.063	0.064	0.28	0.077	-0.14	-0.0071
Total Time Spent on Website	0.34	1	0.34	-0.2	0.3	0.14	0.23	-0.38	0.094	0.043	0.011	-0.19	0.12	0.45	0.02	0.1	0.16	0.18	0.022	-0.13	0.13
Page Views Per Visit	0.63	0.34	1	-0.38	0.53	0.15	0.25	-0.53	0.3	-0.0065	0.08	-0.25	0.076	0.61	0.038	0.087	0.081	0.29	0.085	-0.11	0.063
Lead Origin_API	-0.27	-0.2	-0.38	1	-0.86	-0.44	0.0034	0.61	0.008	0.08	-0.043	0.36	-0.13	-0.35	-0.063	-0.12	-0.18	-0.47	-0.056	0.12	-0.11
Lead Origin_Landing Page Submission	0.4	0.3	0.53	-0.86	1	0.52	0.097	-0.53	0.046	-0.09	0.045	-0.31	0.055	0.57	0.087	0.085	0.088	0.55	0.064	-0.079	0.043
Lead Source_Direct Traffic	0.13	0.14	0.15	-0.44	0.52	1	-0.43	-0.3	-0.23	-0.12	0.015	-0.17	0.0036	0.29	0.14	0.047	0.05	0.6	0.02	-0.038	0.011
Lead Source_Google	0.18	0.23	0.25	0.0034	0.097	-0.43	1	-0.34	-0.25	0.073	0.029	-0.096	0.018	0.41	-0.064	0.027	0.012	-0.32	0.03	-0.0056	-0.0044
Lead Source_Olark Chat	-0.44	-0.38	-0.53	0.61	-0.53	-0.3	-0.34	1	-0.18	0.06	-0.04	0.43	-0.13	-0.71	-0.075	-0.15	-0.21	-0.3	-0.043	0.098	-0.098
Lead Source_Organic Search	0.28	0.094	0.3	0.008	0.046	-0.23	-0.25	-0.18	1	-0.035	-0.013	-0.078	0.029	0.2	0.017	0.034	0.034	0.14	-0.0038	-0.0084	0.029
Do Not Email_No	0.0046	0.043	-0.0065	0.08	-0.09	-0.12	0.073	0.06	-0.035	1	0.22	0.055	0.0033	-0.034	-0.051	0.017	0.046	-0.053	0.18	-0.11	0.017
Last Activity_Email Opened	0.062	0.011	0.08	-0.043	0.045	0.015	0.029	-0.04	-0.013	0.22	1	-0.27	-0.51	0.046	-0.0064	0.0047	0.015	0.031	0.85	-0.28	-0.44
Last Activity_Olark Chat Conversation	-0.2	-0.19	-0.25	0.36	-0.31	-0.17	-0.096	0.43	-0.078	0.055	-0.27	1	-0.23	-0.31	-0.042	-0.15	-0.2	-0.19	-0.23	0.33	-0.2
Last Activity_SMS Sent	-0.0054	0.12	0.076	-0.13	0.055	0.0036	0.018	-0.13	0.029	0.0033	-0.51	-0.23	1	0.025	-0.0067	0.098	0.15	0.028	-0.44	-0.21	0.85
Country_India	0.53	0.45	0.61	-0.35	0.57	0.29	0.41	-0.71	0.2	-0.034	0.046	-0.31	0.025	1	-0.27	0.097	0.084	0.33	0.058	-0.05	0.012
Country_Other	0.036	0.02	0.038	-0.063	0.087	0.14	-0.064	-0.075	0.017	-0.051	-0.0064	-0.042	-0.0067	-0.27	1	0.011	0.0015	0.12	-0.0047	0.0093	-0.0085
Occupation_Unemployed	0.063	0.1	0.087	-0.12	0.085	0.047	0.027	-0.15	0.034	0.017	0.0047	-0.15	0.098	0.097	0.011	1	0.81	0.061	0.012	-0.11	0.14
Course_objective_Better Career Prospects	0.064	0.16	0.081	-0.18	0.088	0.05	0.012	-0.21	0.034	0.046	0.015	-0.2	0.15	0.084	0.0015	0.81	1	0.065	0.018	-0.17	0.2
Interview_guide_Yes	0.28	0.18	0.29	-0.47	0.55	0.6	-0.32	-0.3	0.14	-0.053	0.031	-0.19	0.028	0.33	0.12	0.061	0.065	1	0.043	-0.067	0.024
Last Notable Activity_Email Opened	0.077	0.022	0.085	-0.056	0.064	0.02	0.03	-0.043	-0.0038	0.18	0.85	-0.23	-0.44	0.058	-0.0047	0.012	0.018	0.043	1	-0.5	-0.37
Last Notable Activity_Modified	-0.14	-0.13	-0.11	0.12	-0.079	-0.038	-0.0056	0.098	-0.0084	-0.11	-0.28	0.33	-0.21	-0.05	0.0093	-0.11	-0.17	-0.067	-0.5	1	-0.42
Last Notable Activity_SMS Sent	-0.0071	0.13	0.063	-0.11	0.043	0.011	-0.0044	-0.098	0.029	0.017	-0.44	-0.2	0.85	0.012	-0.0085	0.14	0.2	0.024	-0.37	-0.42	1

Columns with correlation >|0.7| were deleted

Statistical limits

- We were left with 17 columns.
- RFE was used to keep 12 columns and rest were dropped.
- P-value < 0.02 and VIF < 5 was used as cutoff to keep remaining variables thus, 3 more were deleted after a few iterations.
- Total 9 variables were left for model to be trained upon

Model Building

- Logistic Regression was used to teach the model.
- Confusion matrix was made on the predicted and actual conversions at cutoff probability of 0.5 initially.
- Recall, precision and accuracy were calculated for multiple cutoff probability.
- Final cutoff was set at 0.4 while optimising the model for Recall.
- Score column was made as 100 times probability.

Model Performance

- On Train Set: Recall = 75.4%, Accuracy = 77%, Precision = 68% and F1 Score = 71.5%.
- On Test Set: Recall = 75%, Accuracy = 76%, Precision = 68% and F1 Score = 71.8%.

CONCLUSION

Most important Variables for Model:

1. Do Not Email_No
2. Occupation_Unemployed
3. Lead Source_Google
4. Lead Source_Direct Traffic
5. Last Notable Activity_Modified
6. Last Activity_SMS Sent
7. Lead Source_Organic Search
8. Last Activity_Olark Chat Conversation
9. Total Time Spent on Website

Final F1 score was 0.71 with recall of 0.75, precision of 0.68 and accuracy of 0.76.