2020

Determining if Marketing Leads Will Convert

SPRINGBOARD CAPSTONE 1

PAUL BUTLER

The Problem

The company X Education has a low conversion rate for the leads its marketing team focuses on. Currently, their lead conversion rate is only 38%. A big struggle for their marketing team is identifying which leads are warm and which ones are cold. Their current strategy is to spend time following up on every lead they possibly can. This wide net approach is not very successful, and the company believes if they could spend their time focusing on the warm leads, then they would have a higher conversion rate.

My goal is to generate a model or indicator of the chance for each lead converting. This will be in the form of a scale indicator of how warm a lead is, with 100 being a lead guaranteed to convert, and 0 being a lead completely unlikely to convert. This information could be used to make better decisions in several ways. On an individual employee basis, it could be used to determine which leads to focus on. On a team level, it could be used by the leadership to determine a cutoff for which leads to focus on and which leads to ignore, change marketing channel strategies to generate better leads, or change staff levels based on the number of good leads they receive and their capacity to handle them.

Data Cleaning

The dataset was provided by the company. It contains 37 variables for 9240 prospects. The key variable is “converted.” Which indicates if a lead converted to be a paying customer. The dataset includes data on which source the lead came from, user behavior on the website, their reason for looking at the courses, which ads they saw outside of the website, and various user preferences for how they want to be contacted. Additionally, it contains various subjective tags used by the marketing team to assess leads. A full dictionary of variables provided by the company can be found in Appendix A.

The dataset I worked with had several problematic columns, which required several different methods to clean properly for analysis. I will discuss these steps for several categories of similar data types below. A table with exact steps for each column can be found in Appendix B.

The dataset included two custom indexes. I dropped “Prospect ID” and kept “Lead Number” which was used as the index.

For the continuous variables “Total Visits”, and “Time spent on the Website”, I replaced any null values with the median for each column and then used IQR to filter out any outliers. “Page View Per Visit” was left unchanged.

Categorical variables that contained only one observation value were dropped since they would not contribute any information to the overall model. These variables included: “I agree to pay the amount through cheque”, “Get updates on DM Content”, “Update me on Supply Chain Content”, “Receive More Updates About Our Courses”, “Newspaper Article”, “Magazine”.

The columns “How did you hear about X Education”, and “What matters most to you in choosing a course”. Contained mostly useless data and were dropped. For the “How did you hear about X Education” column, more than 50% of data was a useless select(NaN) value. For “what matters most to you in choosing a course”, every value except 3 was either null or was the option “Better Career Prospects”. Since the uniformity of this column adds nothing to the analysis it was also dropped.

Columns with yes and no observations were encoded to 1 and 0 for model building. These Columns included “Do Not Email”, “Do Not Call”, “Search”, “X Education Forums”, “Newspaper”, “Digital Advertisement”, “Through Recommendations” and “a free copy of Mastering the Interview”

Categorical variables that contained missing data, or the value “Select” (which indicates that the user could have selected and option but didn’t), were inputted as NaN. These columns included “Lead Source”, “Specialization”, “What is your current occupation”, “Tags”, “Lead Quality”, “Lead Profile”, “City”.

Some columns that had values that were too similar conceptually to other columns were dropped entirely. These included “Asymmetrique Activity Index” and “Asymmetrique Profile Index”, which were intended to capture the same information as “Asymmetrique Activity Score” and “Asymmetrique Activity Score”. Additionally, “Last Activity” was dropped for having almost the exact values as “Last Notable Activity”

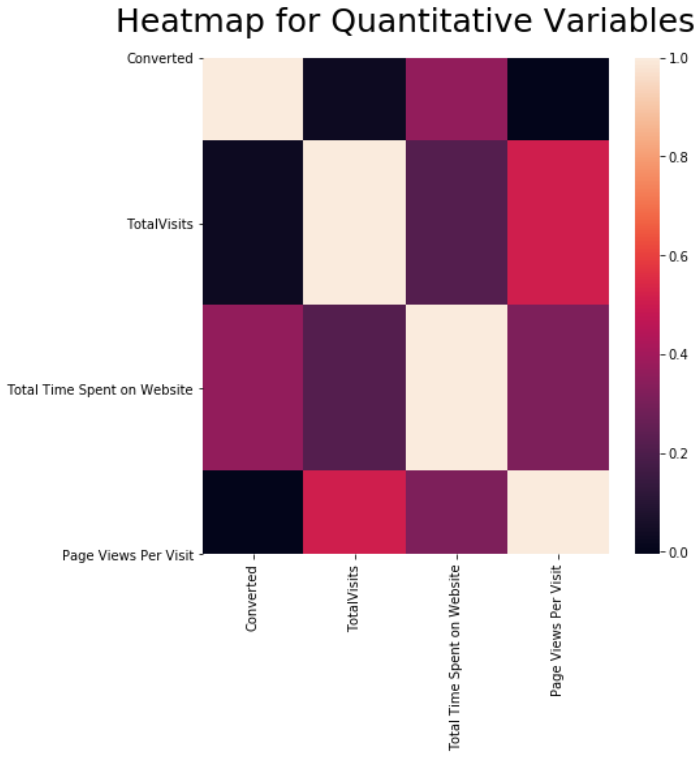
All categorical variables with 3 or more labels were hot encoded to dummy variables and the original column and any created “NaN” columns were dropped in preparation for model building.

Subsequent EDA of the dataset confirmed that there no obvious outliers, null values, or data that should significantly impair or contribute to misleading analyses or conclusions. Some additional data cleaning to check for multicollinearity would be performed before creating the logistic regression model.

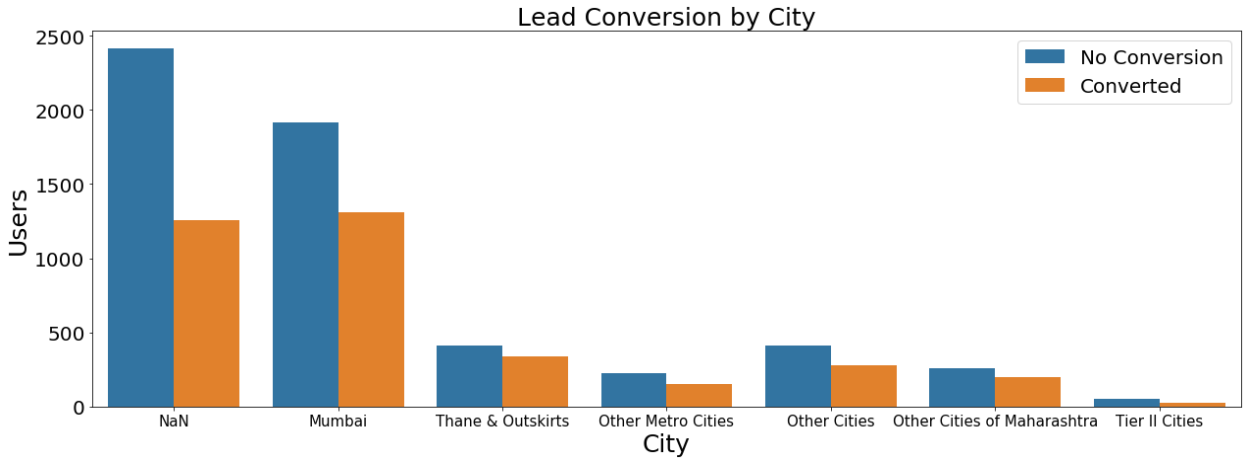
Data storytelling

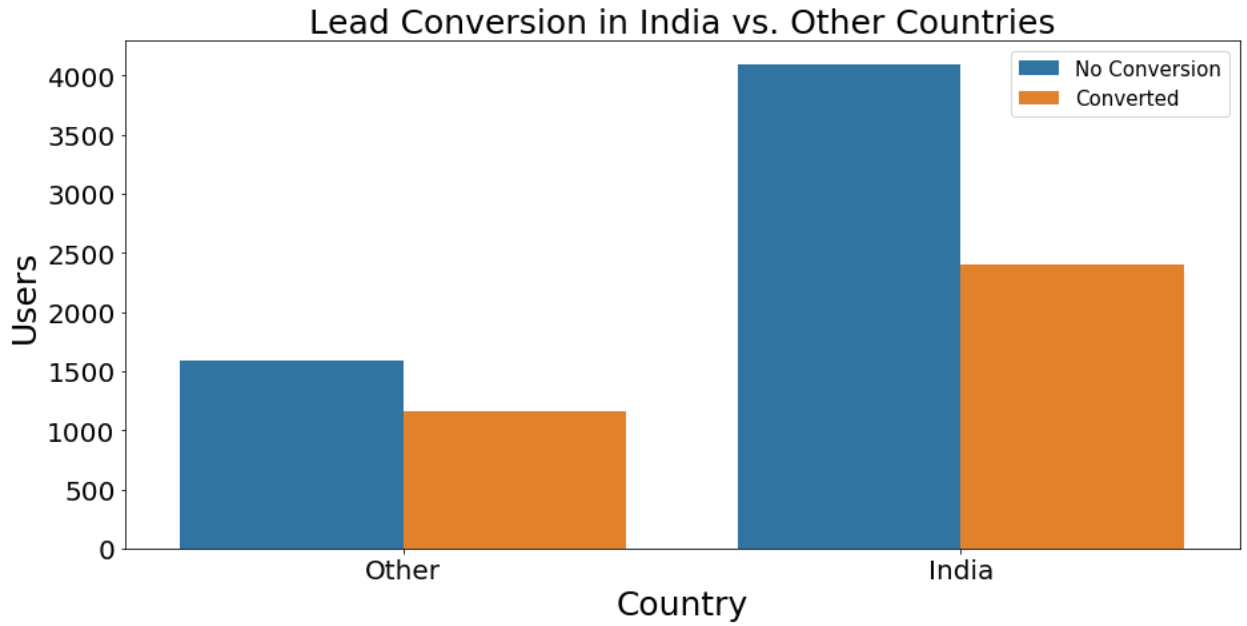
The goal of this project is to help the company X Education predict which sales leads are likely to convert to paid customers. The company uses a variety of methods for tagging their users or “leads”. Some of these are objective and others are subjective. I wanted to see if there were any obvious relationships between the different ways the company labeled users and see if there was an effect on conversion.

The first comparison I did was to check correlations between the quantitative variables in the dataset. This dataset is mostly limited to categorical data, but they do record the number of times a user visits their site, how much time the user spends on the site, and how many pages they view per visit. This simple heatmap shows a modest relationship between page views per visit and total visits, but not much else.

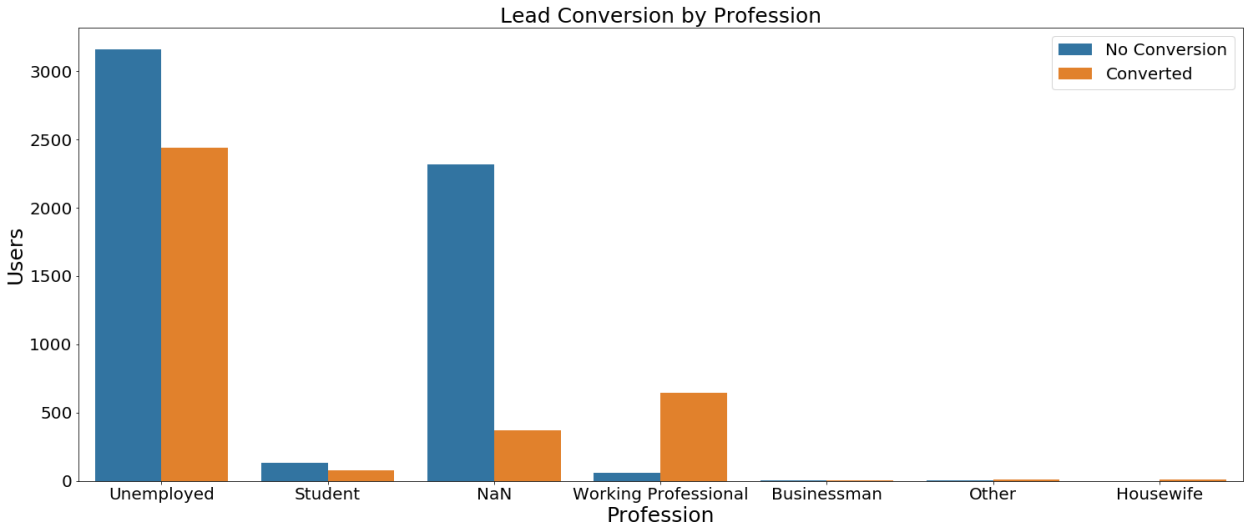


I wanted to see if there were differences between success regionally, nationally, or internationally. This revealed that most of their traffic came from Mumbai, and from within India, but there was not much difference between conversion between cities within India, or between India and conversions outside of India.

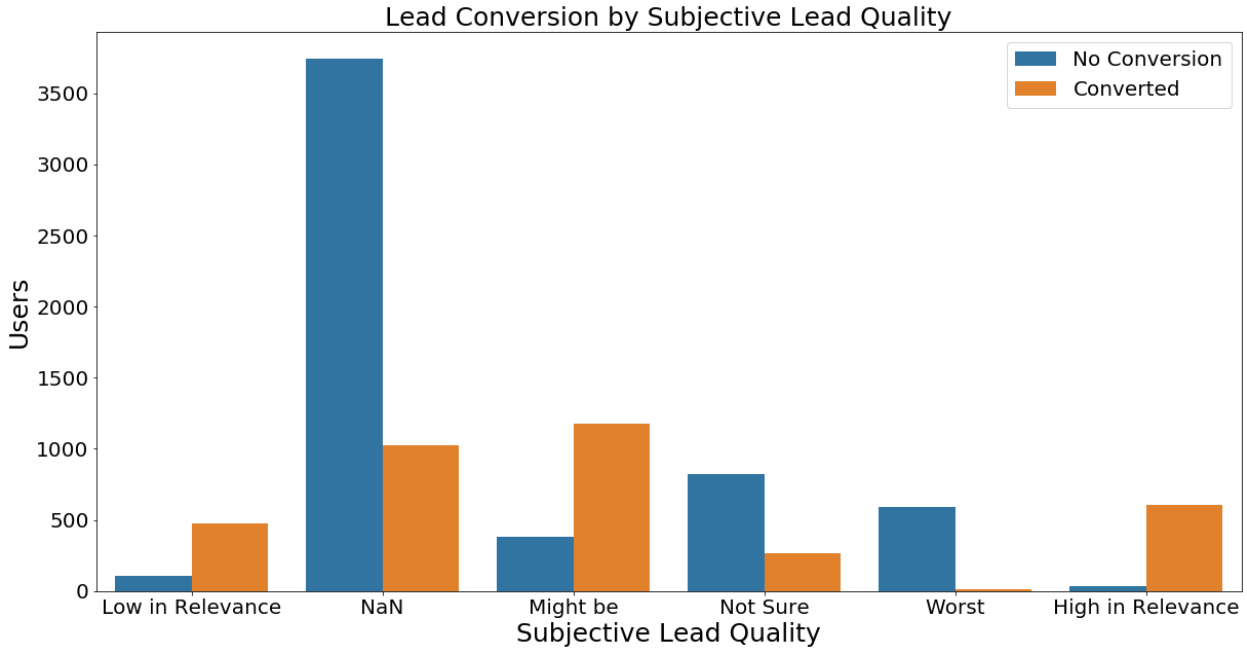




Next, I wanted to see if there was an obvious relationship between different professions and conversion. I discovered that while most users were unemployed, users who described themselves as working professionals were the most likely to convert.



Lastly, I wanted to check if the subjective labels from the marketing department were accurately predicting lead conversion. The graph below indicates that they can predict conversion if certain leads are terrible or great, but not when their subjective rating is in between. This chasm in predictive ability will need to be explored using another method.



It is clear from these plots that EDA is not enough to solve the question sufficiently, and a more advanced approach must be used. There are a few obvious relationships that jump out, but most observations do not seem to predict lead conversion. The story so far is that company is gathering a lot of data but has not managed to determine what data helps predict a successful outcome except in a few rare cases.

The Problem Revisited

Using a machine learning method is essential for solving the problem of lead conversion prediction since EDA revealed few obvious trends. For this case, I used logistic regression to classify leads as converting or not. Using logistic regression also allows the reporting of the prediction values as a "Lead Score" between 1-100, which will give an intuitive understanding of the model's result for the marketing team. Tree-based methods were also considered to check the maximum accuracy potential of the dataset, but this was not done due to the high accuracy achieved by the logistic regression model (90%).

Data Preparation

The dataset was further prepared by removing outliers beyond the IQR, and mapping all yes and no categorical variables to 1 and 0, respectively. All other categorical variables featuring two or more responses were hot encoded to dummy variables, with the missing values dropped. Multicollinearity was then assessed and all similar variables conceptually were removed. I decided at this point to wait before cleaning further variables, to see what result would be achieved by analyzing the data in statsmodel. If multicollinearity was still a problem, this would be solved by testing for variable inflation factor (VIF).

Model Building

After the data was prepared, the data was split into testing and training data using a stratified 80/20 split. The model was then first evaluated using statsmodel with a binomial generalized linear model since it is equivalent to logistic regression. This revealed that the coefficient values for each feature had extreme values, and the log-likelihood and deviance were not computable, signaling that the model had too many features with dependencies or multicollinearity. I then decided to implement recursive feature elimination (RFE) to help eliminate noncontributing features.

I implemented RFE with 20 features and analyzed the result using a binomial GLM in statsmodels. This brought the coefficients for the remaining features into a reasonable range, indicating a working model, but several features were also nonsignificant. These were dropped manually, and the model was evaluated using stats model until there were 11 remaining significant features (highest $p = 0.002$). The model was further evaluated using VIF with the highest value being 1.70. Indicating that multicollinearity was not an issue in the refined model.

5 Features with the Highest VIF values

Columns	VIF
What is your current occupation_Working Profes...	1.70
Tags_Ringing	1.70
Tags_Already a student	1.55
Lead Quality_Worst	1.53
Tags_Interested in other courses	1.28

Remaining Significant Model Features

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.5636	0.158	-16.239	0.000	-2.873	-2.254
Last Notable Activity_SMS Sent	2.1505	0.219	9.819	0.000	1.721	2.580
Asymmetrique Activity Score_15.0	1.5885	0.255	6.240	0.000	1.090	2.087
What is your current occupation_Unemployed	1.4446	0.195	7.419	0.000	1.063	1.826
What is your current occupation_Working Professional	2.8423	0.938	3.030	0.002	1.004	4.681
Lead Quality_Worst	-2.3113	0.721	-3.204	0.001	-3.725	-0.897
Tags_Already a student	-3.2659	1.060	-3.082	0.002	-5.343	-1.189
Tags_Closed by Horizzon	5.0557	1.024	4.939	0.000	3.049	7.062
Tags_Interested in other courses	-3.9045	1.030	-3.793	0.000	-5.922	-1.887
Tags_Lost to EINS	4.4834	0.847	5.296	0.000	2.824	6.143
Tags_Ringing	-3.7259	0.428	-8.714	0.000	-4.564	-2.888
Tags_Will revert after reading the email	4.1441	0.390	10.639	0.000	3.381	4.908

The coefficients for the logistic regression model indicate a few general patterns. First a set of features that demonstrate an obvious negative relationship. The tags “Ringing”, “Already a Student”, and “Interested in other Courses” are predictive of a failure to convert and may represent a factor of users who expressed initial interest, but later decided not to convert. The feature “Tag Lost to Eins” is also likely in this factor, since it a feature which where a positive observation predicts a negative real-world outcome. The remaining negative feature “Lead Quality Worst” is also predictive of a user not converting.

The features “Last Notable Activity SMS Sent” and “Tags Will revert after reading the email” are associated with conversion and likely indictors of users who have decided to convert but are still discussing details with staff. “Tags Closed by Horizzon”, has the highest positive coefficient, and likely indicates a partner that closes users for X Education. “What is your current occupation unemployed”

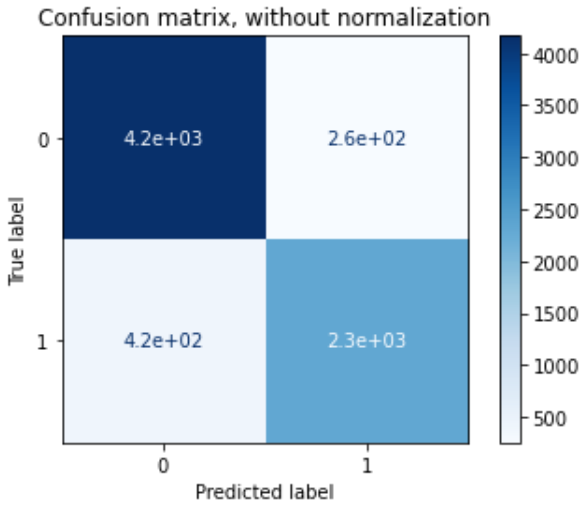
and “what is your current occupation working professional” both have a positive relationship with conversion and indicate both features represent groups which are associated with taking courses from X Education.

The last remaining feature with a modest contribution to the model is “Asymmetrique Activity Score 15.0” since this is a proprietary measure, it is not obvious why this score on the scale is predictive while others are not.

Assessing Model Accuracy

I then created the model in scikit-learn using the 11 selected features, training accuracy was ~91%, and testing accuracy was ~90%, indicating that the model was neither underfitting nor overfitting. Area under the curve was 0.95, which indicates the model overall is an exceptional classifier. To further assess model performance a confusion matrix was generated. The confusion matrix scores were very good overall for both classification conditions, with a relative drop in f1 score and recall for the converted condition. Indicating the model has a relative weakness in finding all the cases that will convert. A recall score of 0.85 is still very good and is likely not worth further parameter tuning.

Confusion Matrix



Cross-validation was also performed to further test the model accuracy. Cross-validation using 10 folds was performed and the results were compared using minimum, mean, and max. The minimum accuracy was ~89%, the mean accuracy was ~91% and the maximum accuracy was 94%, indicating that accuracy did not vary significantly enough during cross-validation to indicate a problem with the model.

Closing Statement

The result is a model that not only predicts classification with 90% accuracy on the testing data but also reports the predicted chance of conversion on a scale of 1-100. While in a typical logistic regression model the predicted chance is not normally reported in favor of only the classification label. It is employed in this case to meet the needs of the marketing team so they can have an ordinal ranking of leads from the warmest to the coldest. This allows for a more intuitive understanding of the model results, which will enable them to use it to make better decisions. Results in this format will help make decisions such as the following with less uncertainty:

- Whether to increase or decrease staffing levels based on the volume of warm leads.
- Which leads to focus time and resources on and which to ignore.
- Which leads are the most likely to convert and need little encouragement, and which marginal leads may need additional attention to convert.
- Which team members should be paired with which leads (a lead that is on the margin might be given to the best performing team members, to see if they can convert them. Etc.)

As noted during EDA, the current subjective observations in the dataset are useful at predicting extremely warm or cold leads at first glance but do not help predict the leads between these two extremes. The biggest benefit of this model is that it creates useful information for the leads within this gap. The marketing team can now better interpret the quality of leads across the entire range. Instead of just the worst or best leads.

Sample Outcome Format

Lead Number	Lead Score	Predicted Outcome	Actual Outcome
641725	9	0	0
613102	39	0	0
598020	98	1	1
651611	97	1	1
595073	93	1	1
657686	97	1	1
593831	25	0	0
586770	1	0	0
624430	99	1	1
585273	70	1	0

Appendix A

Variables	Description
Prospect ID	A unique ID with which the customer is identified.
Lead Number	A lead number assigned to each lead procured.
Lead Origin	The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.
Lead Source	The source of the lead. Includes Google, Organic Search, Olark Chat, etc.
Do Not Email	An indicator variable selected by the customer wherein they select whether of not they want to be emailed about the course or not.
Do Not Call	An indicator variable selected by the customer wherein they select whether of not they want to be called about the course or not.
Converted	The target variable. Indicates whether a lead has been successfully converted or not.
Total Visits	The total number of visits made by the customer on the website.
Total Time Spent on Website	The total time spent by the customer on the website.
Page Views Per Visit	Average number of pages on the website viewed during the visits.
Last Activity	Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.
Country	The country of the customer.
Specialization	The industry domain in which the customer worked before. Includes the level 'Select Specialization' which means the customer had not selected this option while filling the form.
How did you hear about X Education	The source from which the customer heard about X Education.
What is your current occupation	Indicates whether the customer is a student, unemployed or employed.
What matters most to you in choosing this course	An option selected by the customer indicating what is their main motto behind doing this course.
Search	Indicating whether the customer had seen the ad in any of the listed items.
Magazine	
Newspaper Article	
X Education Forums	

Newspaper	Indicating whether the customer had seen the ad in any of the listed items.
Digital Advertisement	
Through Recommendations	Indicates whether the customer came in through recommendations.
Receive More Updates About Our Courses	Indicates whether the customer chose to receive more updates about the courses.
Tags	Tags assigned to customers indicating the current status of the lead.
Lead Quality	Indicates the quality of lead based on the data and intuition the employee who has been assigned to the lead.
Update me on Supply Chain Content	Indicates whether the customer wants updates on the Supply Chain Content.
Get updates on DM Content	Indicates whether the customer wants updates on the DM Content.
Lead Profile	A lead level assigned to each customer based on their profile.
City	The city of the customer.
Asymmetrique Activity Index	An index and score assigned to each customer based on their activity and their profile
Asymmetrique Profile Index	
Asymmetrique Activity Score	
Asymmetrique Profile Score	
I agree to pay the amount through cheque	Indicates whether the customer has agreed to pay the amount through cheque or not.
a free copy of Mastering The Interview	Indicates whether the customer wants a free copy of 'Mastering the Interview' or not.
Last Notable Activity	The last notable activity performed by the student.

Appendix B

Variable	Cleaning Outcome
Prospect ID	Dropped
Lead Number	Used as dataframe index
Lead Origin	Unchanged
Lead Source	Missing values inputted using NaN
Do Not Email	Recoded Yes and No to 1 and 0
Do Not Call	Recoded Yes and No to 1 and 0
Converted	Unchanged
Total Visits	Missing values inputted using median
Total Time Spent on Website	Missing values inputted using median
Page Views Per Visit	Unchanged
Last Activity	Dropped-Almost identical values as “Last Notable Activity”
Country	Recoded values to be “India” “Other”
Specialization	Missing values & select values inputted using NaN
How did you hear about X Education	Dropped – too many missing values
What is your current occupation	Missing values inputted using NaN
What matters most to you in choosing this course	Dropped – almost all observations are the same value
Search	Recoded Yes and No to 1 and 0
Magazine	Dropped-Only one value
Newspaper Article	Dropped-Only one value
X Education Forums	Recoded Yes and No to 1 and 0
Newspaper	Recoded Yes and No to 1 and 0
Digital Advertisement	Recoded Yes and No to 1 and 0
Through Recommendations	Recoded Yes and No to 1 and 0
Receive More Updates About Our Courses	Dropped-Only one value
Tags	Missing values inputted using NaN
Lead Quality	Missing values inputted using NaN
Update me on Supply Chain Content	Dropped-Only one value
Get updates on DM Content	Dropped-Only one value
Lead Profile	Missing values & select values inputted using NaN
City	Missing values & select values inputted using NaN
Asymmetrique Activity Index	Dropped
Asymmetrique Profile Index	Dropped
Asymmetrique Activity Score	Missing values inputted using NaN
Asymmetrique Profile Score	Missing values inputted using NaN
I agree to pay the amount through cheque	Dropped-Only one value
a free copy of Mastering The Interview	Recoded Yes and No to 1 and 0
Last Notable Activity	Unchanged