Problem Statement

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 can. This blanket approach leads to the marketing team emailing and calling every single lead. 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. Their CEO has specified a conversion rate goal of 80%.

My goal is to generate a model or indicator of the chance for each lead converting. This will be in the form of a percentage 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 dataframe.

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 were a useless select value. For "what matters most to you in choosing a course", every value except 3 were either null or were 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”.

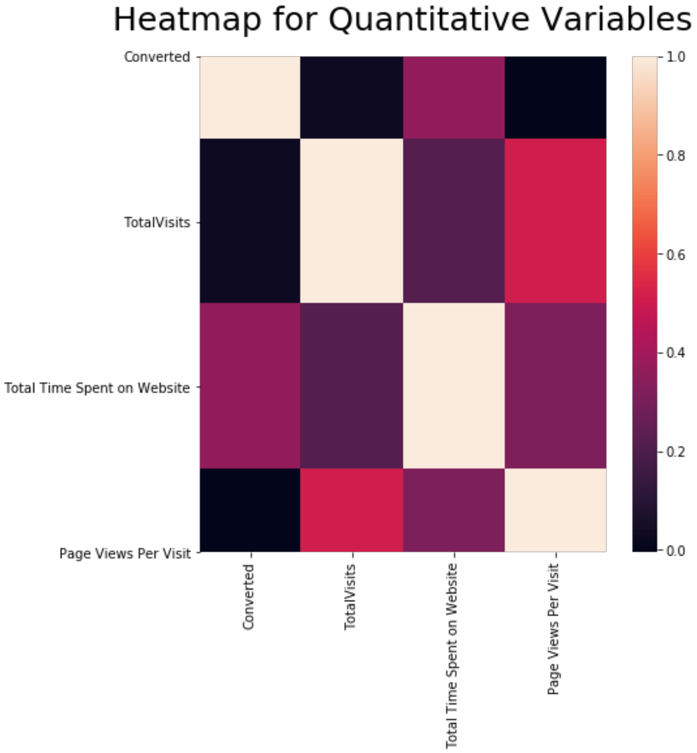
Some columns that had values that were too similar conceptually to other columns were dropped entirely. These included “Asymmetrique Activity Index” and “Asymmetrique Profile Index”, which were intended to capture the same information as “Asymmetrique Activity Score” and “Asymmetrique Activity Score”.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

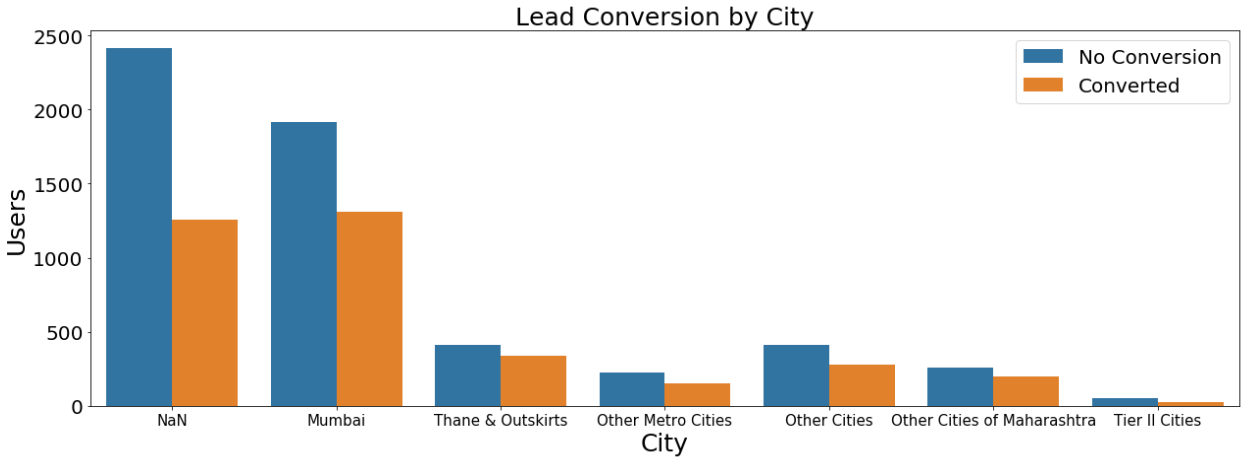
Data storytelling

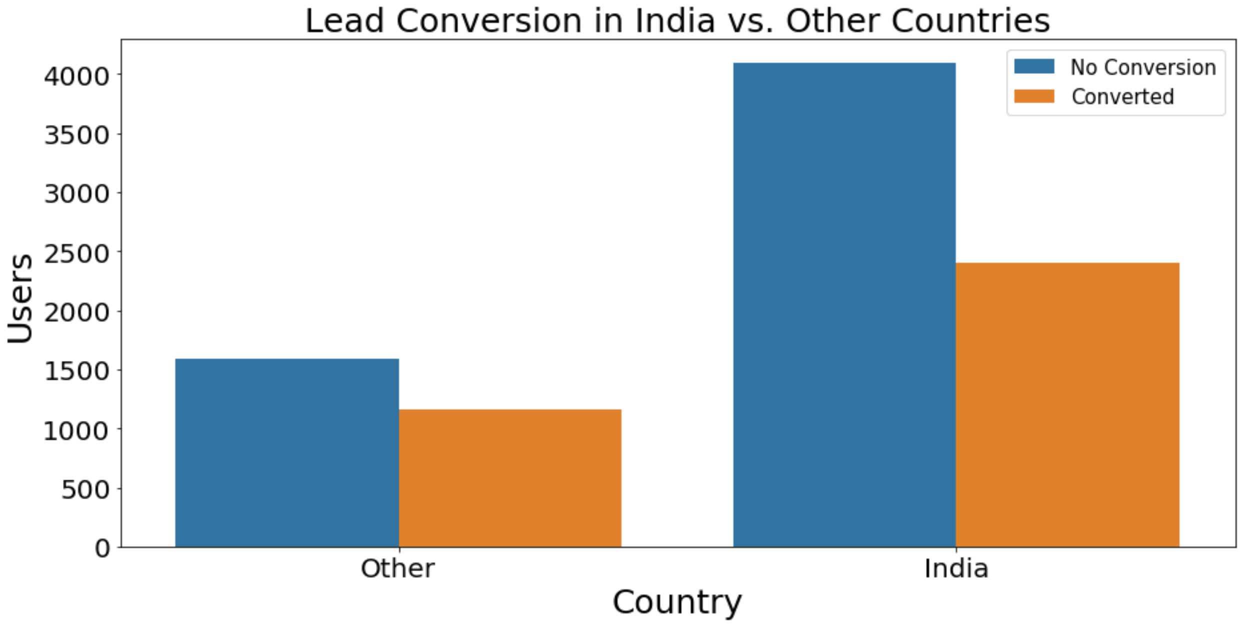
The first comparison I did was to check correlations between the quantitative variables in the dataset. This dataset is mostly limited to categorical data, but they do record the amount of times a user visits their site, how much time the user spends on the site, and how many pages they view per visit.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



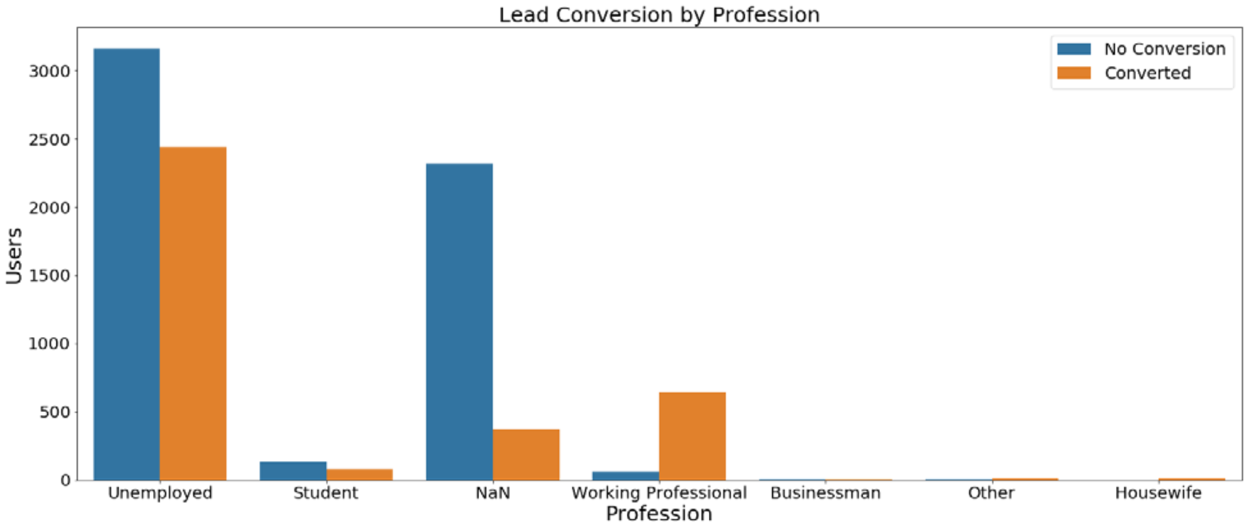
The goal of this project is to help the company X Education predict which sales leads are likely to convert to paid customers. The company uses a variety of methods for tagging their users or “leads”. Some of these are objective and others are subjective. I wanted to see if there were any obvious relationships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

Next, 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.

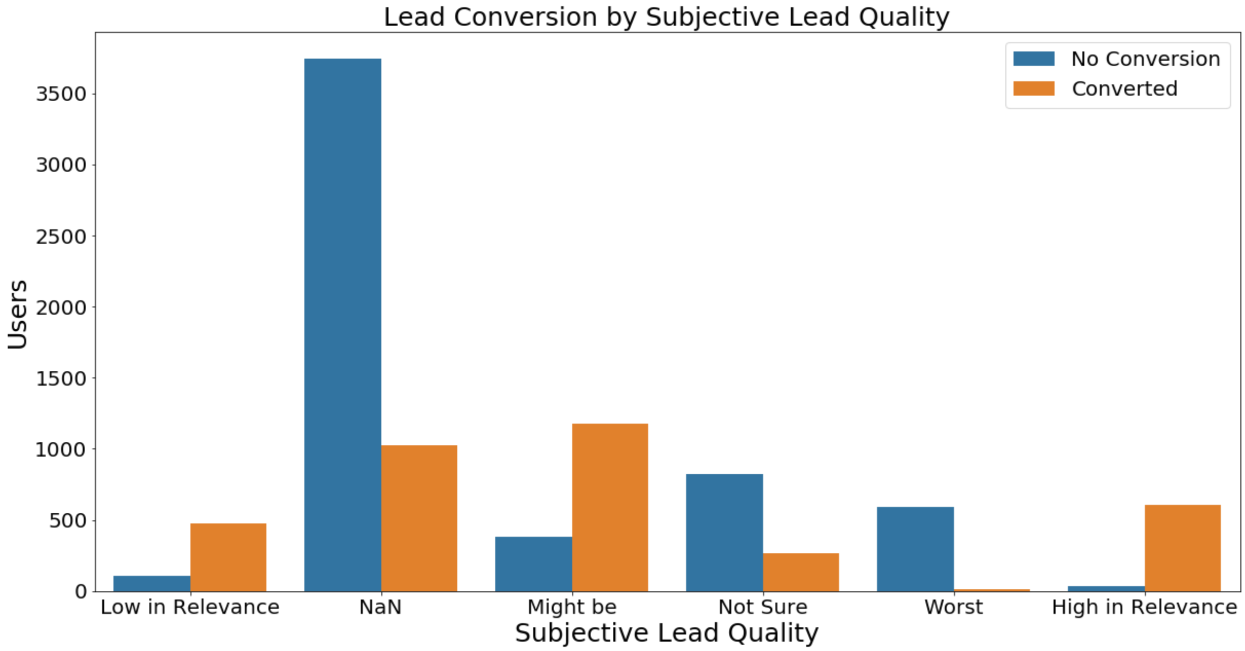




Next, I wanted to see if there was an obvious relationship between different professions and conversion. I discovered that while most users were unemployed, users who described themselves as working professionals were the most likely to convert.



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.

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