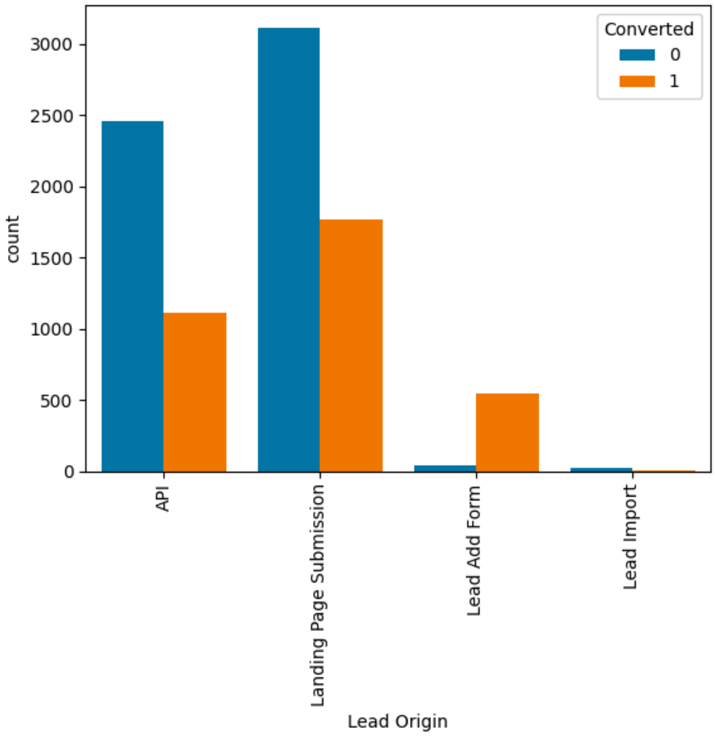
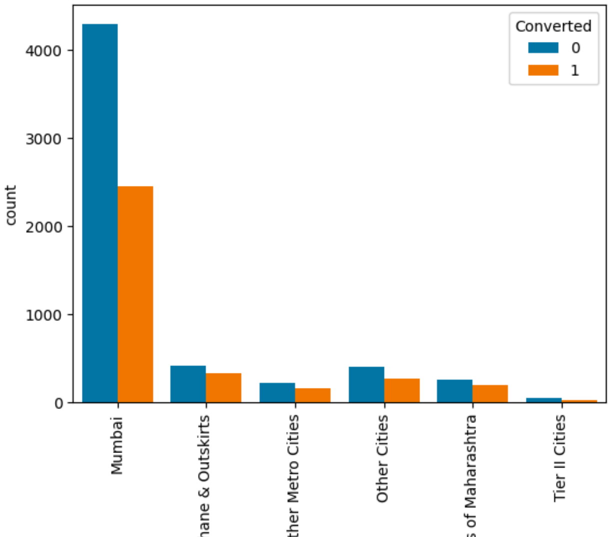
Summary:

The problem statement for this assignment is that there is an education company which is trying to sell the online courses . There are marketing on various search engines and when people fill in their details they are considered to be a lead . With leads multiple phone calls and emails are sent . After all these process the leads who get converted to enrolling the course is only 30% and this needs to be improved . To prevent wasting time and effort they want to maximise the potential leads by classification .

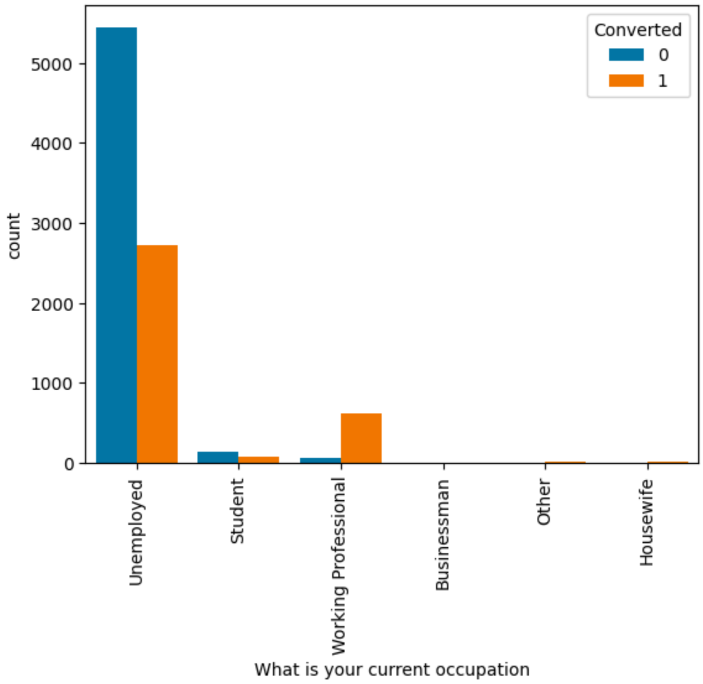
In order to do that first the data needs to be analysed . The data set provided has 36 columns all with various data types . The first step is to clean the data . First look at the data we see that there are many columns called as select . This means that the user has not filled data for that columns , so in the cleaning this is converted to a null value . Next is to check for all the null values in the data . We see that there are multiple null values . First that data which have majority as null value needs to be removed , so deleted all the columns where the null value was more than 70%.For values where the null values are less the rows with that column having null values was removed . For the other cases the null values were imputed . Is there were not much variation the null values were renamed as another category . Once all the null values were treated the Exploratory data analysis is performed . Checked how the lead counts are getting rate as converted .



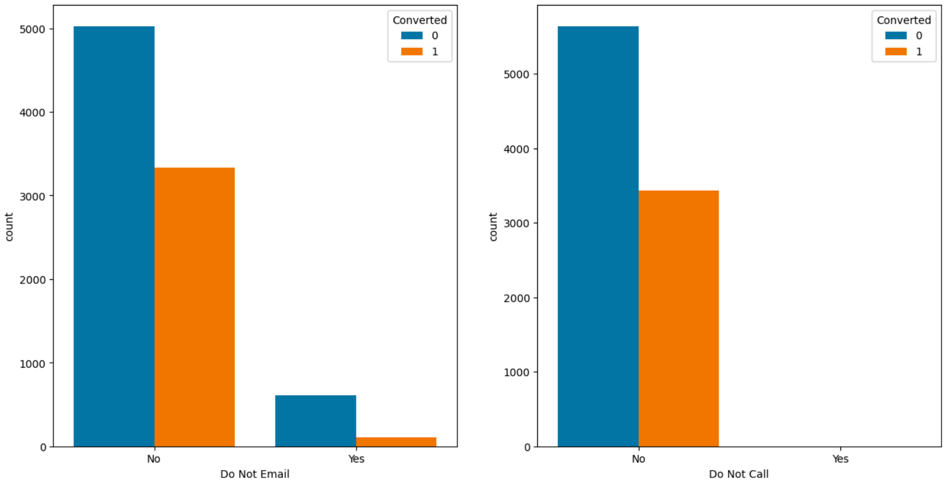
Here we see that the landing page submission has most number of conversions.
Based on cities



We see that Mumbai is having the best conversion rate maybe due to large amount of data from Mumbai



Based on the above graph we see that the conversion rate of working professionals is more than not getting converted . So this is a good data point .



Email and calls have a better conversion rate the done . The next step after the basic analysis is to prepare the data . Here we see multiple binary values to the columns that is mapped to 0 and 1 . The categorical data is converted to dummy value . The next step is outlier treatment . Removed all the outliers from the continuous values and made the data stay within the interquartile range . Once that was done the the data is converted , the data is split to train and test with 70% train and 30% test. Once the data was split , the next is to normalise the value which was done using standard scaler here . The next is to remove the unwanted features and use the ones that matter the most . The was done using recurrent feature alimentation method and top 20 features were then selected . The accuracy on the training set was 92.4% . But we need to check if there is a variance inflation factor and remove unnecessary correlation between independent variables .This helped in removing more unwanted features and the final train set had an accuracy of 92.8%. With this we also need to find the optimal cut off for which roc curve was drawn , along with that checked how the specificity , sensitivity and the accuracy vary with respect to the cutoff. And the optimal cutoff was found , which gave a final accuracy of 92.4 % on the training set . The same model was used to predict the output of the test set which had an accuracy of 92.9% which means that there is no overfitting of the model .