Lead Score case Study

Dharmeshnath Mishra

Problem Statement

X Education sells online courses to industry professionals.

X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Goals of the case study

Building a model to help them identify most promising leads which is called hot leads

Model should be able to keep up with company's future requirements and make the predictions accordingly

Analysis approach

We started with data Cleaning and data Manipulation

Inspecting the data manually and using **sweetviz** library for duplicate or missing or NA or something that indicate missing values like in this data select option was used in place of missing data.

Dropping columns having large number of missing values making it useless for analysis.

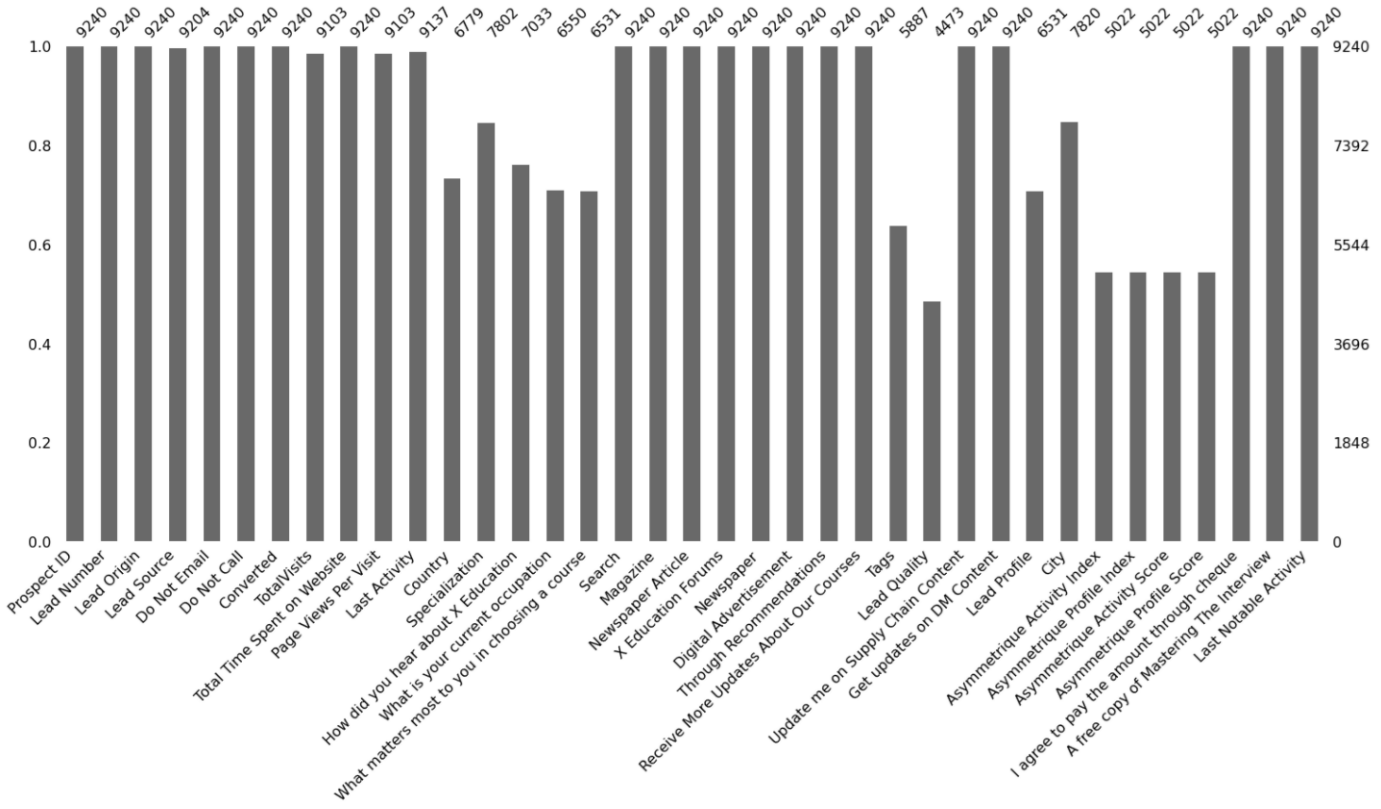
Manipulating columns to solve the missing value problem or to make it more useful.

Analysis approach

- Next, we did EDA
- Both using sweetViz library and manually
- Univariate data analysis
- Bivariate data analysis
- Followed by scaling variables and creating dummy variables
- Followed by using classification technique for model making and prediction
- Followed by finding the best model using p-value and VIF
- Model presentation followed by conclusion.

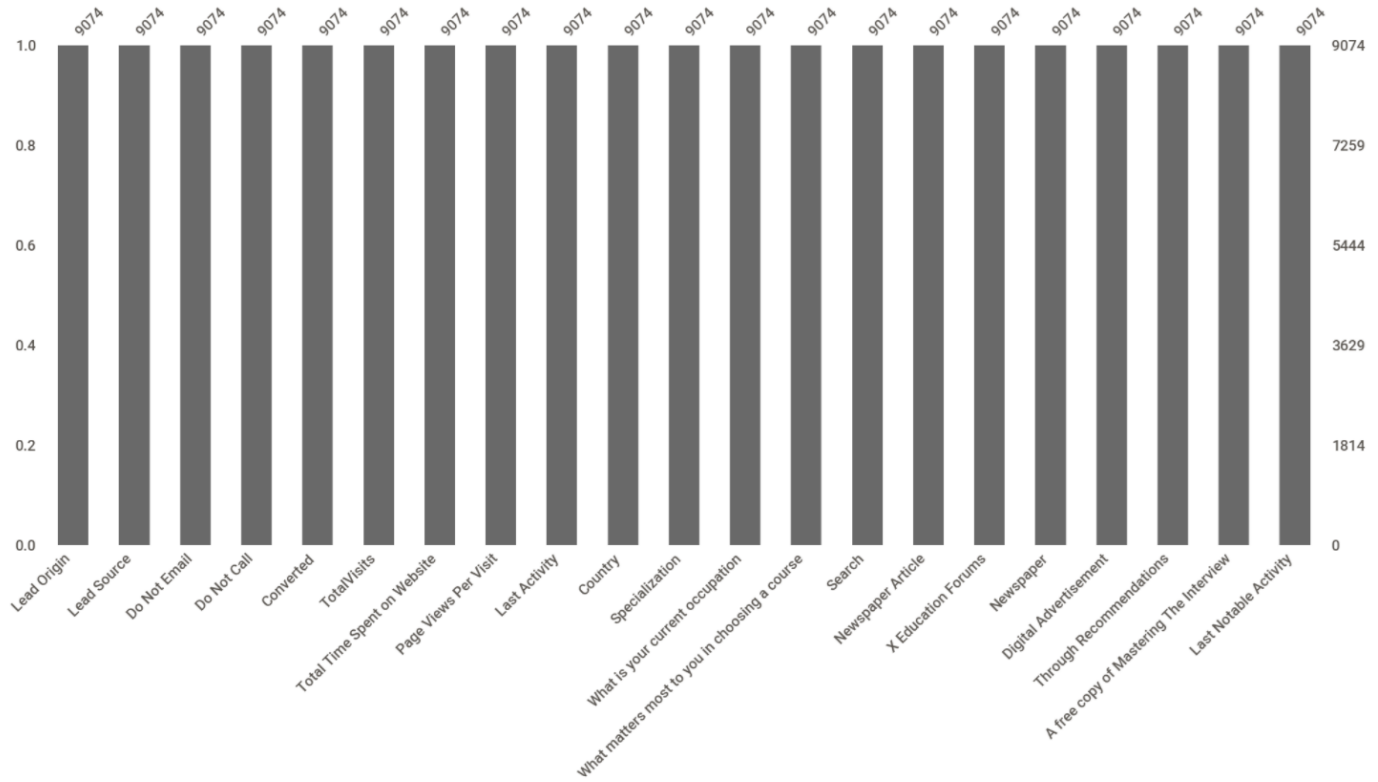
Data Manipulation

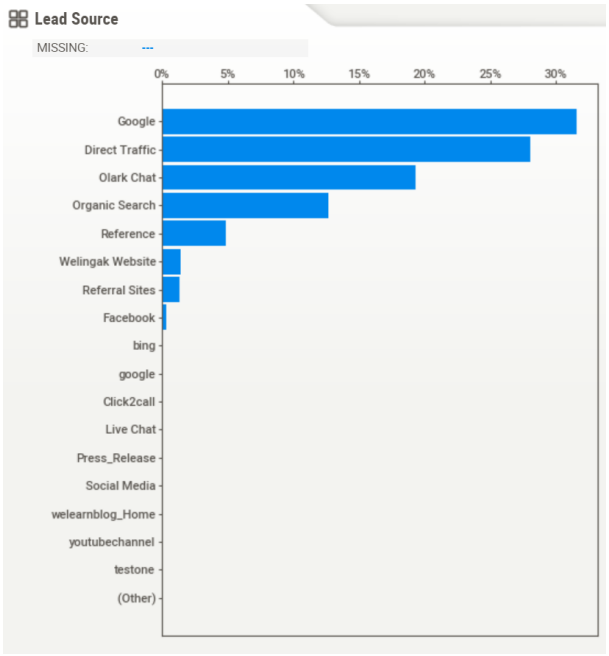
- Missing values before any manipulation

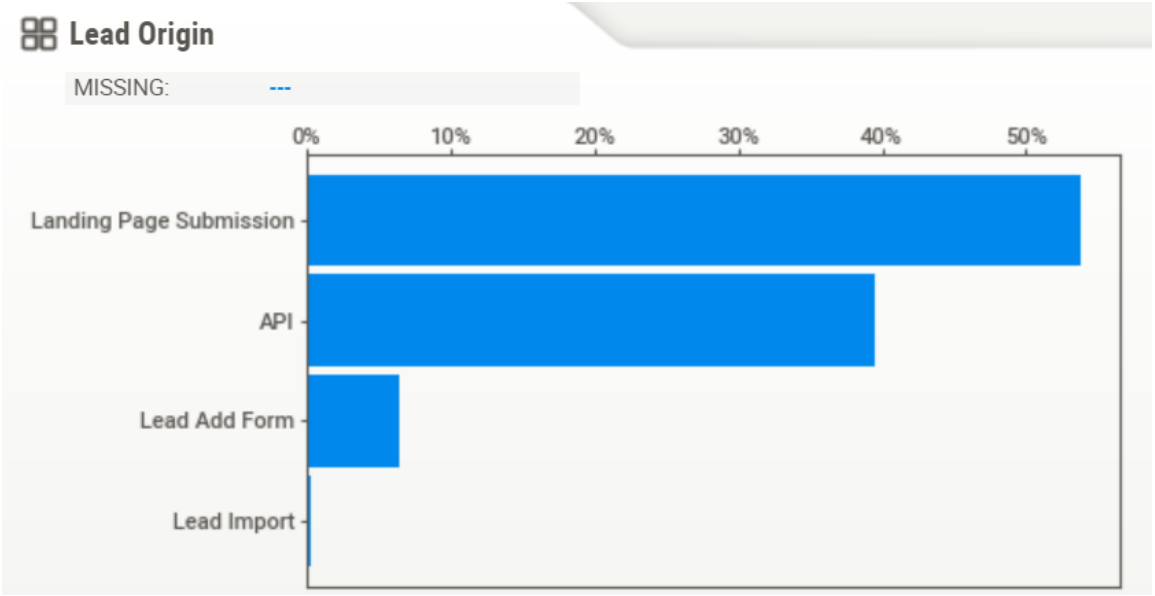


Data Manipulation

- Missing values after doing all the manipulation we end up retaining around 98.2% of the original data before dropping two columns ‘prospect ID’ and ‘Lead Number’ since they were not of any use in the analysis







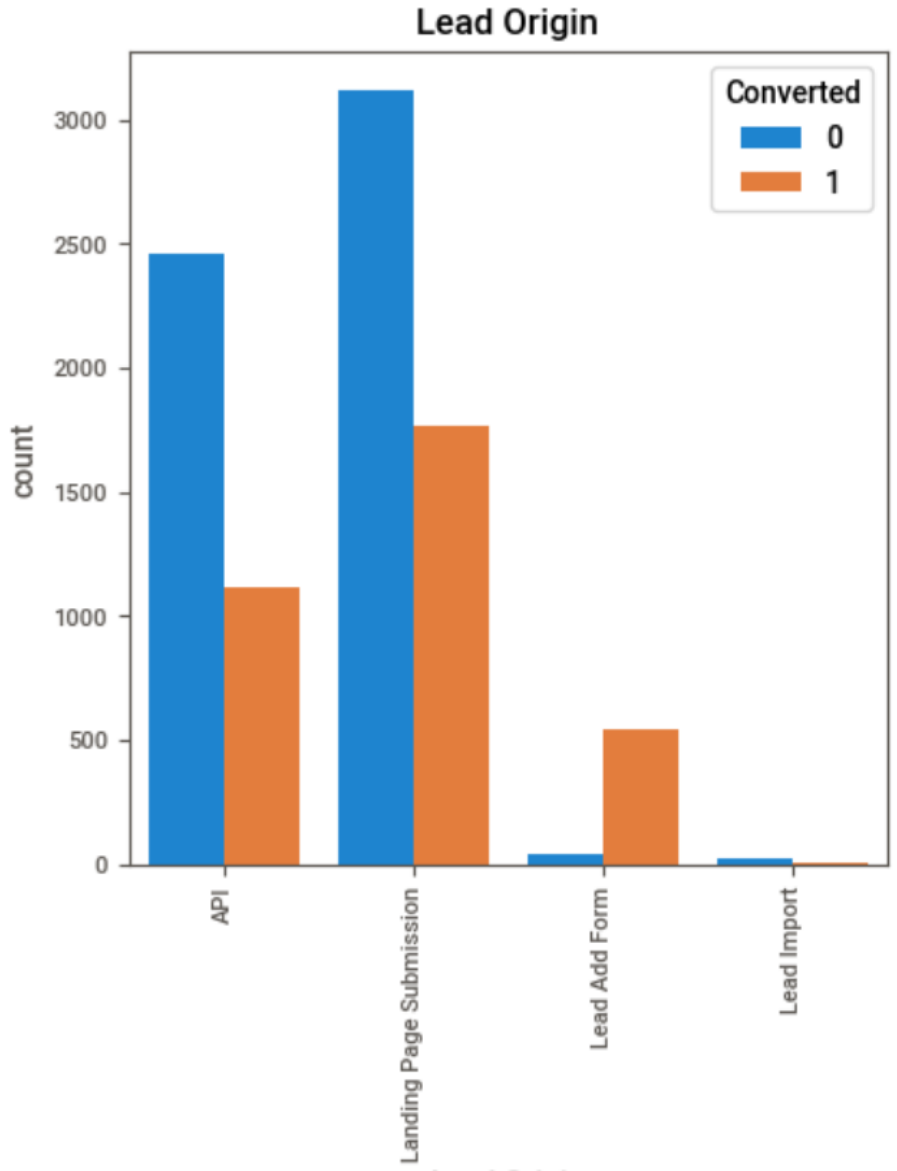
EDA

By looking at these two graph we may think that the category with maximum percentage of presence should have highest conversion rate.

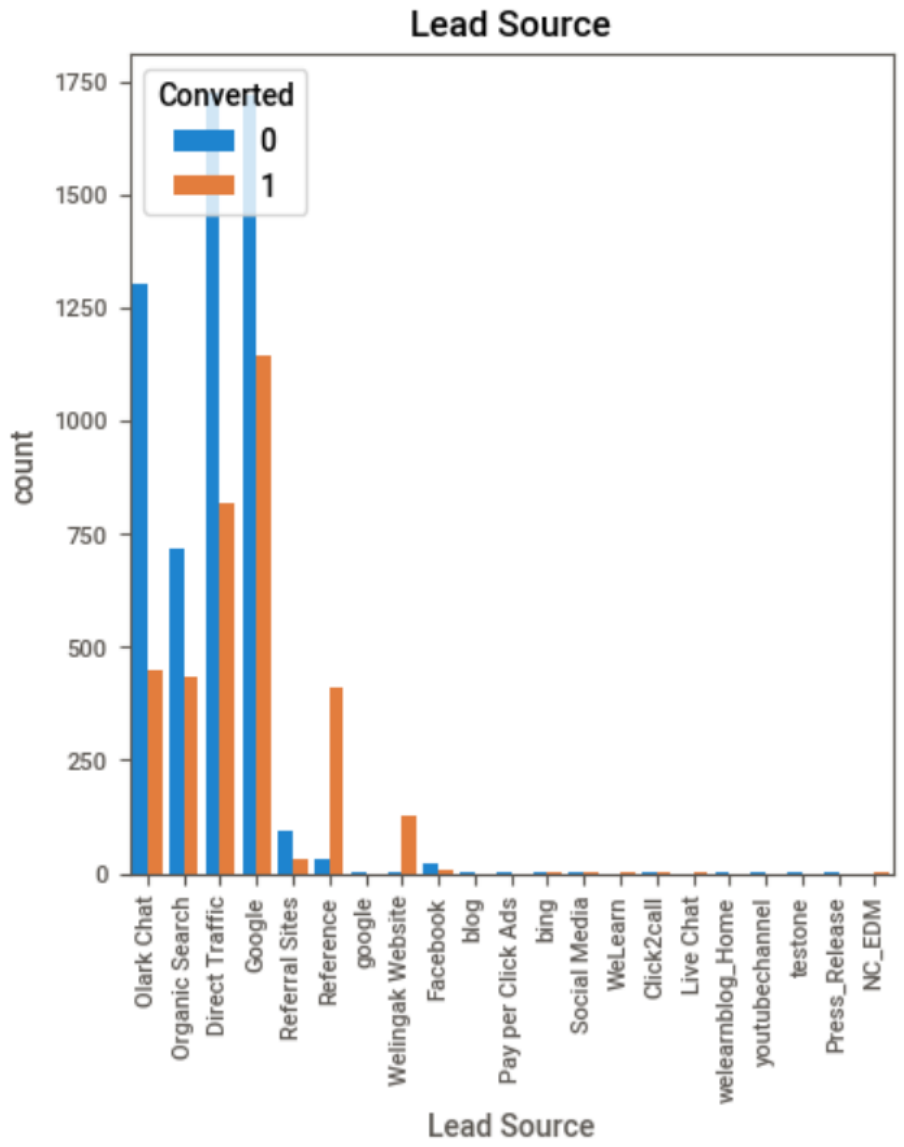
We will go further and see if that is the case.

Relationship between categorical variables and converted

- Looking at this graph we can see that rate of conversion is highest in the Lead Add From though the total count is low.



We can see that though the count is low in Reference and Welingak Website but the rate of convestion is best in these.



Data Conversion

Creating dummy variables for categorical data and deleting redundant columns like while working on this data one dummy variable of Specialisation was of no use so we have to drop that column with all the other columns

Scaling numerical variables

Model Building

Split data into 70:30 ration of train and test respectively

We will use RFE to narrow down the list of columns to 15

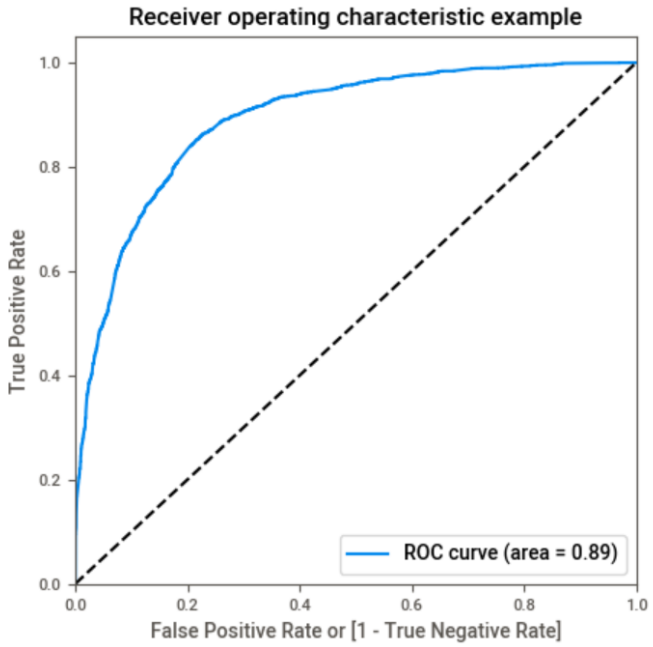
To find the best model we will keep dropping columns with p-value is greater than 0.05 or VIF greater than 4

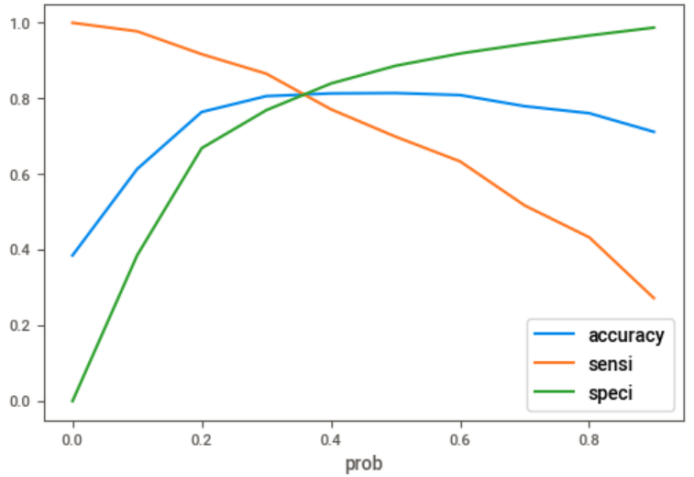
Prediction on test data

We got overall accuracy of 81.4%

Roc and Finding optimal cut off point

- Since area under the ROC curve is 0.89 infers we have a good model
- Looking at the second graph we can see that at optimal cut off 0.38 we have balanced accuracy, sensitivity and specificity.





Conclusion

It can be seen in our final model that there are lot of features which will impact the potential buyer's decision to enrol in the course, we will see few variables which have positive effect on our target variable in descending order:

- Total number of visits made by the customer on the website.
- The total time spent on the Website.
- When their current occupation is working professional
- Lead origin is Lead Add Form
- Lead source is Welingak website
- Lead activity is had a phone conversation

Keeping these in mind the X Education can look into their leads and focus more time on clients with these features in the data to increase their conversion rate.

In Future company can increase or decrease the threshold depending on their requirement of how aggressive they want to be which will increase or decrease the number of Leads.