X Education - Lead Score Case Study

- a model to improve the conversion rate of leads to 80%

Team - Angad Gothe, Anish Mahale and Anadita

Topics inside :

- Company background and Problem Statement
- Data Understanding, cleaning and exploration (Performing EDA)
- Model building and model evaluation
- Recommendations

Company background and Problem Statement

Background :

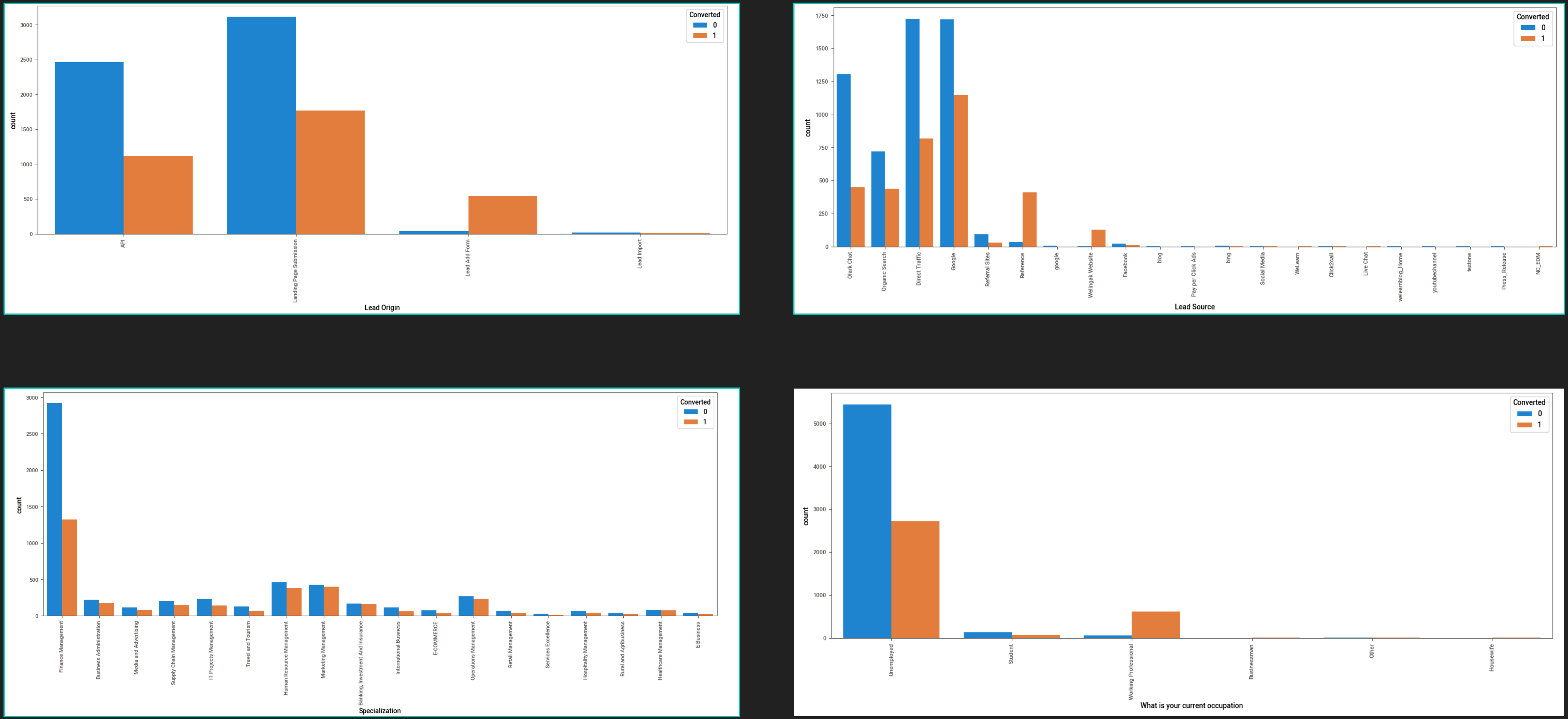
- Company named X Education is a provider of online courses and study options.
- Many websites do have the advertisements of this company which number of customers visit. Some provide basic information and some don't.
- This information is used to collect the data of customers and approach them to offer them courses and some get converted and buy the courses some don't convert.

Problem Statement :

- Here the problem statement is to identify the key features which can help identify the potential customers. Currently the lead conversion rate is very low around 30%
- Hence build a model which can help predict the potential customer from the sea of visitors.
- The CEO's vision is to build to model to identify the Hot Leads and take the lead conversion rate to 80%.

Data Understanding, cleaning and exploration (Performing EDA)

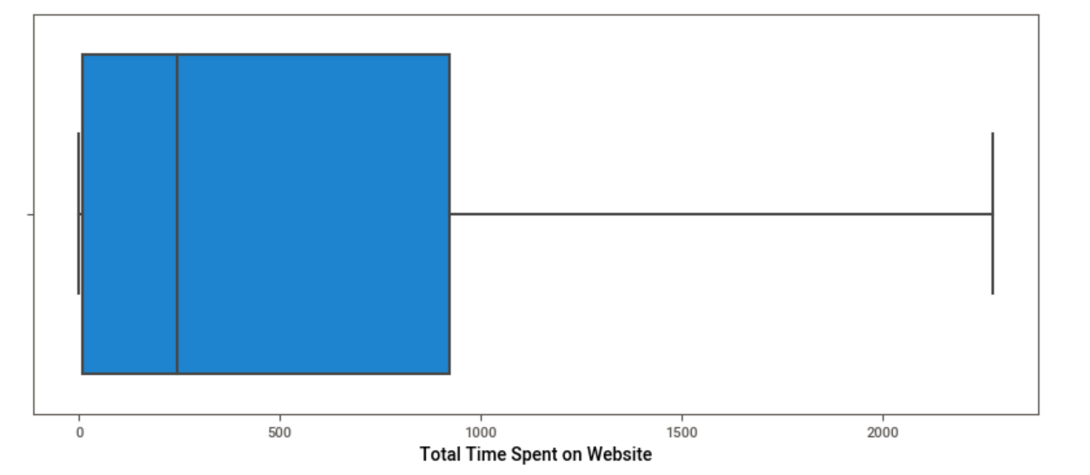
- Used the Data dictionary to understand the details about each columns in the given csv.
- Dropped the columns where the NULL values are greater that 40% (Example : How did you hear about X Education, Lead Quality, Lead Profile, Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score','Asymmetrique Profile Score. As these NULL values is nothing but user has preferred not to answer. And very less values available will disturb the model.
- Checked the value counts and used Mean, Median and Mode to replace the NULL.
- Ensured that all the NULL values are now either dropped or replace.
- As a next step to identify the relation plotted graphs for pictorial understanding. (Uni/Bivariate analysis)
- Some graphs are shown here :

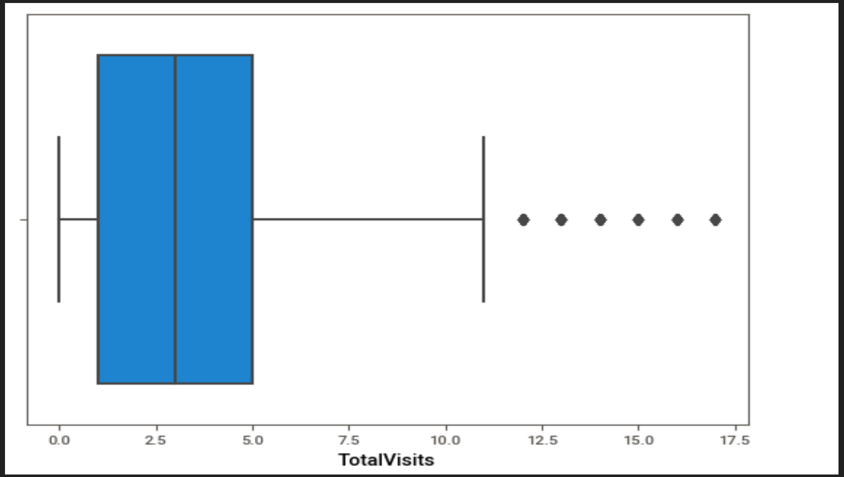


Examples : These are plots for Lead Origin, Lead Source, Specialization, Current Occupation vs customers who Converted or not.

BOX Plots : to find the distribution and outliers.

Some examples are as below :





- Continued the analysis by creating Dummy variable for some columns to make them usable for model building. Example : Lead Score, Last Activity, Current Occupation etc.
- Dropped the duplicate columns created by Dummy.

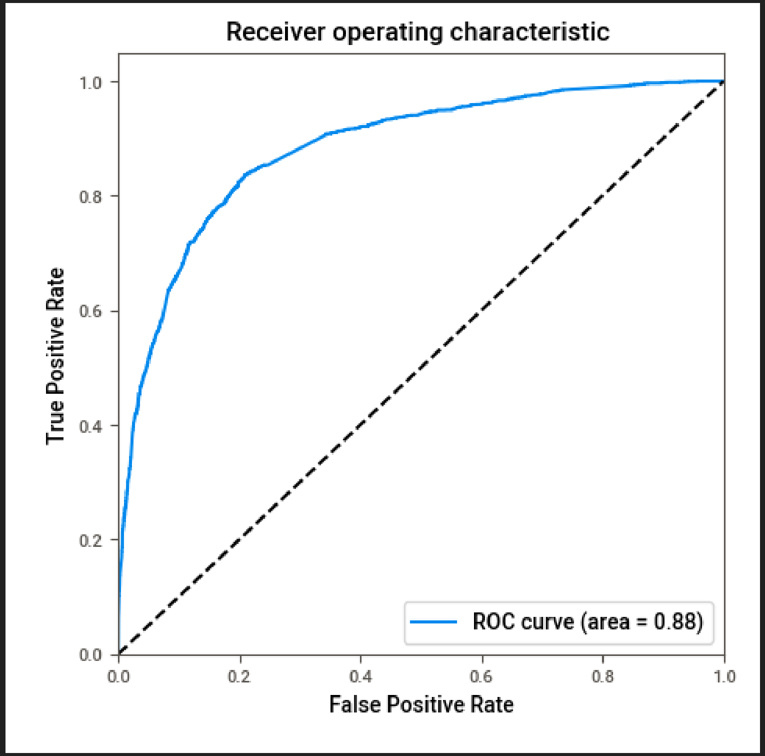
Model building and model evaluation

- Started with dividing the data into Train and Test Model. (70 – 30 percent.)
- Performed scaling on the columns to bring all the variable at same scale to remove bias due to unscaled features.
- Eliminated the not so significant features by using RFE automated feature selection and built a Logistic Regression Model.
- Started building model on 20 selected features, then performed the elimination of features based on P value and VIF values.
- Finally got a model where all P Values are less than 0.05.
- Categorized the predicted values as Yes if probability was greater than 0.5.
- Used the above prediction to built a confusion matrix. The actual vs predicted looks like below

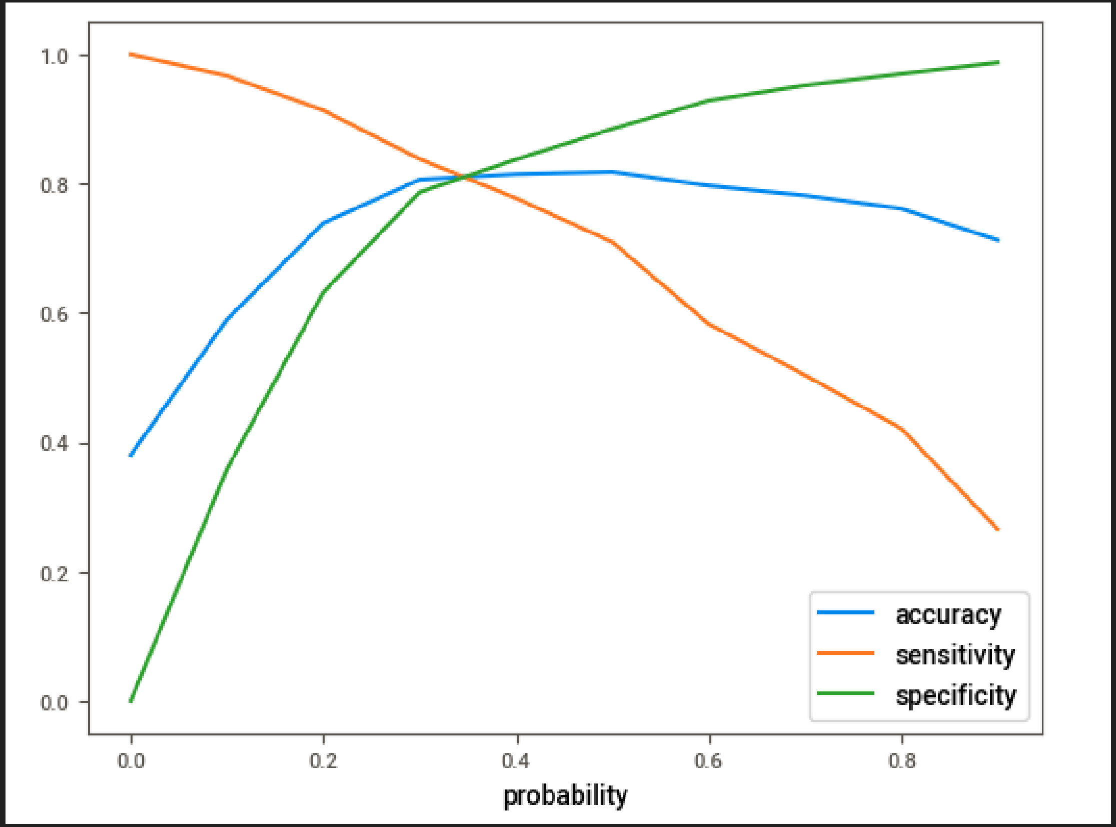
True Negative	**False Positive**
False Negative	**True Positive**

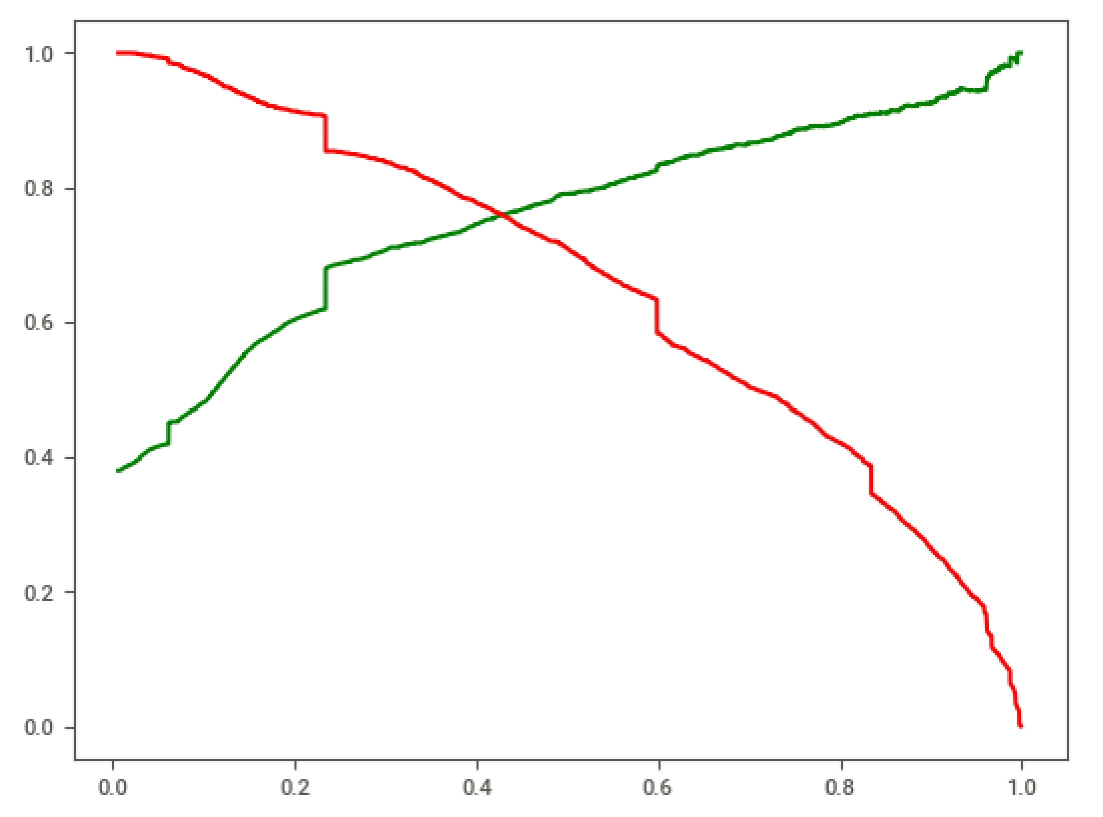
Formula for building model:

- Sensitivity > True Positive / (True Positive +False Negative)
- Specificity > True Positive / (True Negative +False Positive)
- Precision > True Positive / (True Positive +False Positive)
- Recall > True Positive / (True Positive +False Negative)
- Plotted the ROC graph for Trade off between Sensitivity and Specificity (as in adjacent graph)



Plotting accuracy sensitivity and specificity for various probabilities.





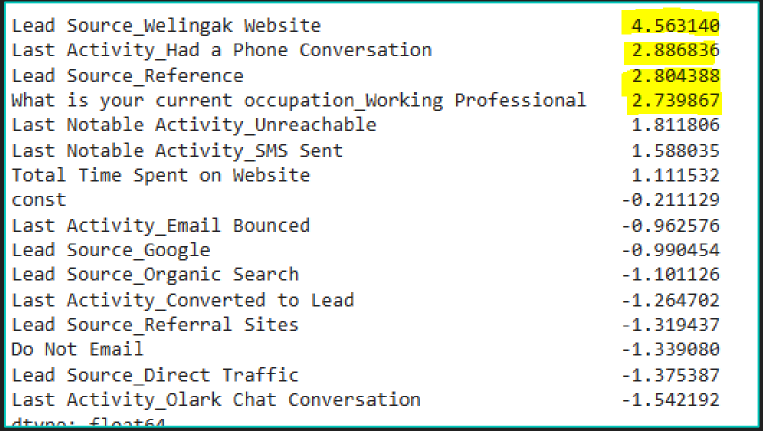
Plot gives an optimum cut-off value of 0.35.

Plot as per new Cut-off value (0.35).

Recommendations

The key features where the coefficients have highest positive values are here. These can be used while approaching the customers. As they will be the potential customers to get converted.

- Company can reach to the customers who have known about the course from Website or from references as the Lead source.
- The customers who had their last activity registered as "Phone conversation" are mostly good leads to convert as they are serious and hence providing their numbers and having conversations.
- The customers whose current occupation is Working professionals are having high changes of taking the courses.



```
Lead Source_Welingak Website                            4.563140
Last Activity_Had a Phone Conversation                  2.886836
Lead Source_Reference                                   2.804388
What is your current occupation_Working Professional    2.739867
Last Notable Activity_Unreachable                       1.811806
Last Notable Activity_SMS Sent                          1.588035
Total Time Spent on Website                             1.111532
const                                                  -0.211129
Last Activity_Email Bounced                            -0.962576
Lead Source_Google                                     -0.990454
Lead Source_Organic Search                             -1.101126
Last Activity_Converted to Lead                        -1.264702
Lead Source_Referral Sites                             -1.319437
Do Not Email                                           -1.339080
Lead Source_Direct Traffic                             -1.375387
Last Activity_Olark Chat Conversation                  -1.542192
dtype: float64
```

With a diligent effort to connect with the customers with these significant features can help X Education company to reach out to right potential customers and hence help to achieve the goal of 80% conversion rate.