LEAD SCORE CASE STUDY

BY

AARSHIA

SAMARTH HEGDE

BINDU KSHTRIYA

PROBLEM STATEMENT

X Education is an organization which provide online courses for industry professional. The company marks its courses on several popular websites like google.

X Education wants to select most promising leads that can be converted to paying customers.

Although 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

Typical lead conversion rate of x education is 30% through the whole process of turning leads into customers by approaching those leads which are to be found having interest in taking the course. The implementation process of lead generating attributes are not efficient in helping conversions.

BUISNESS GOAL

The company requires a model to be built for selecting most promising leads.

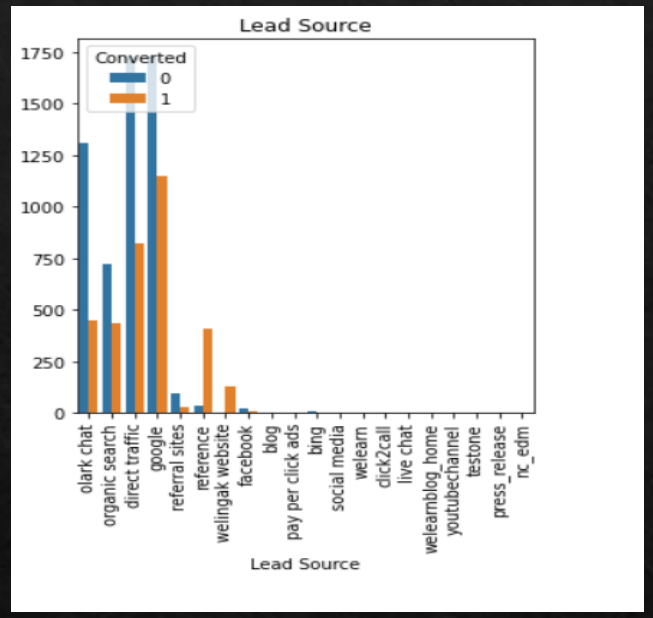
Lead score to be given to each leads such it indicates how promising the lead could be. The higher the lead score the more promising the lead to get converted, the lower it is the lesser the chances of conversion.

The model to be built in lead conversion rate around 80% or more.

STRATEGY

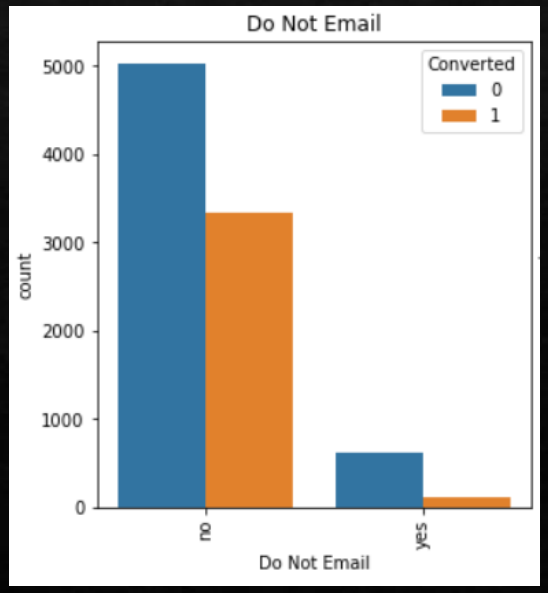
Import data
Clean and prepare the acquired data for further analysis
Exploratory data analysis for figuring out most helpful attributes for conversion
Scaling Features
Prepare the data for model building
Build a logistic regression model.
Assign a lead score for each leads
Test the model on train set
Evaluate model by different measures and metrics
Test the model on test set
Measure the accuracy of the model and other metrics for evaluation

EXPLORATORY DATA ANALYSIS



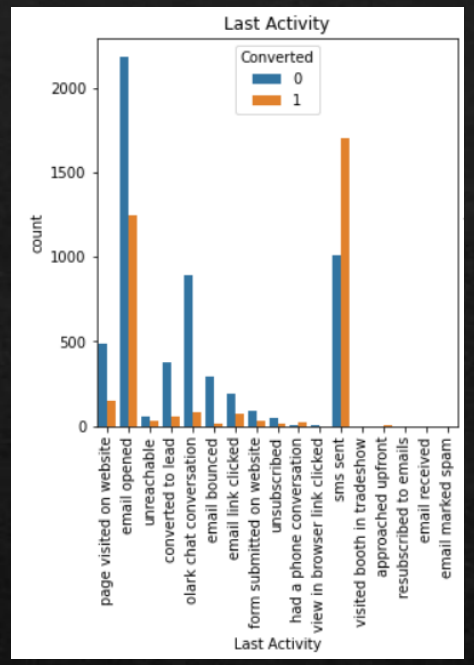
LEAD SOURCE VS CONVERTED

Google searches has had high conversions compared to other models, whilst references has had high conversion rate.



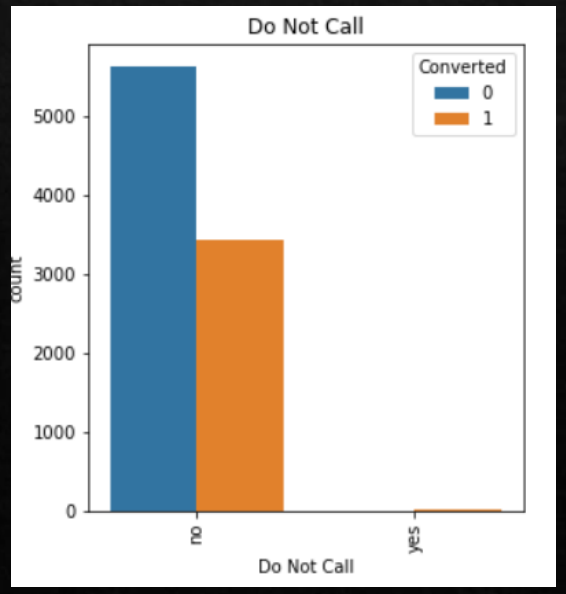
DONOT EMAIL VS CONVERTED

Most leads preferred not to informed through email.



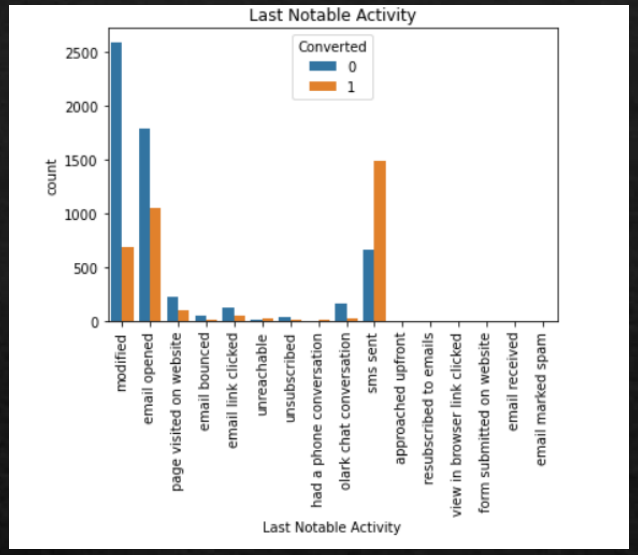
LAST ACTIVITY VS CONVERTED

SMS has shown to be a promising method for getting higher confirmed leads, emails also has high conversions.



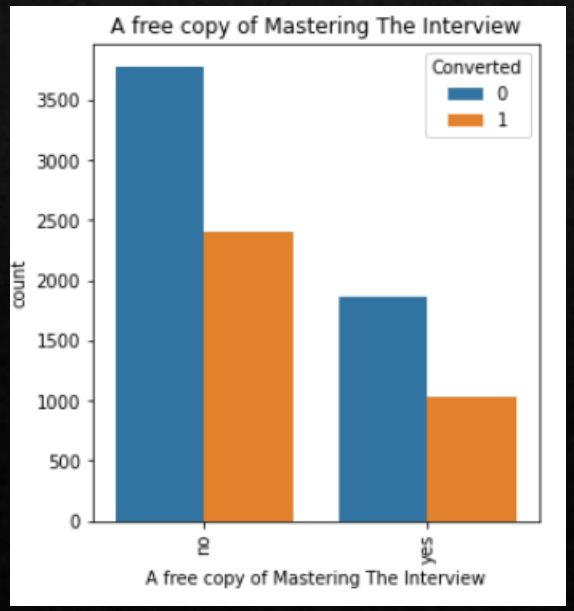
DO NOT CALL VS CONVERTED

Most leads prefer not to informed through phone.



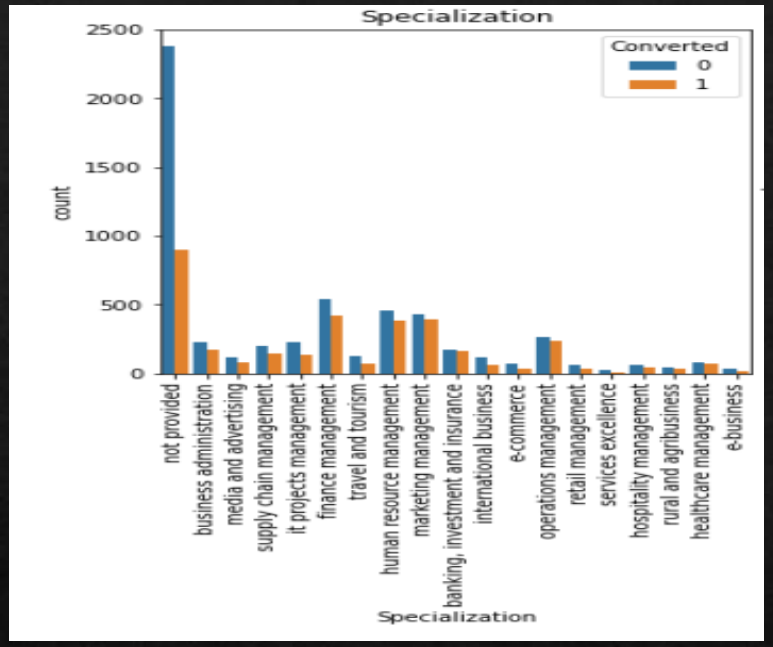
LAST NOTABLE ACTIVITY VS CONVERTED

Most leads are converted with messages.
Emails also induce leads.



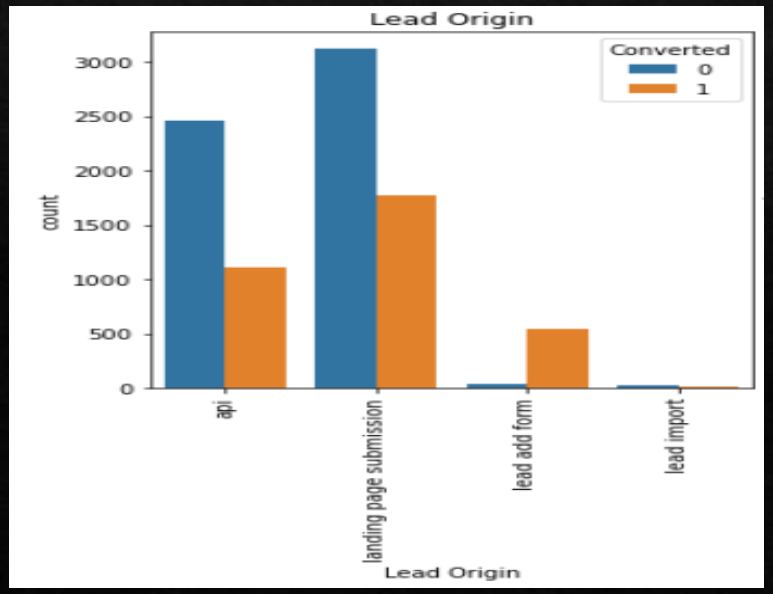
A FREE COPY OF MASTERING THE INTERVIEW VS CONVERTED

Leads prefer less copies of interviews.



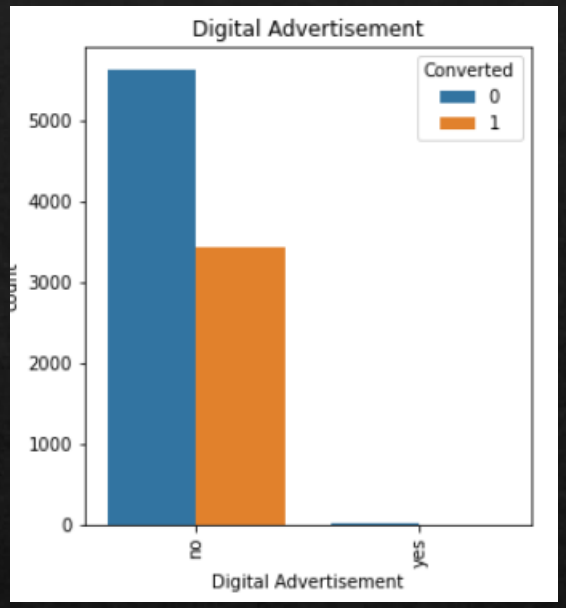
SPECIALIZATION VS CONVERTED

Most of the leads have no information about the specialization. On the other hand, finance management, marketing management, human resources management has high conversion rates. People from these specializations can be promising leads



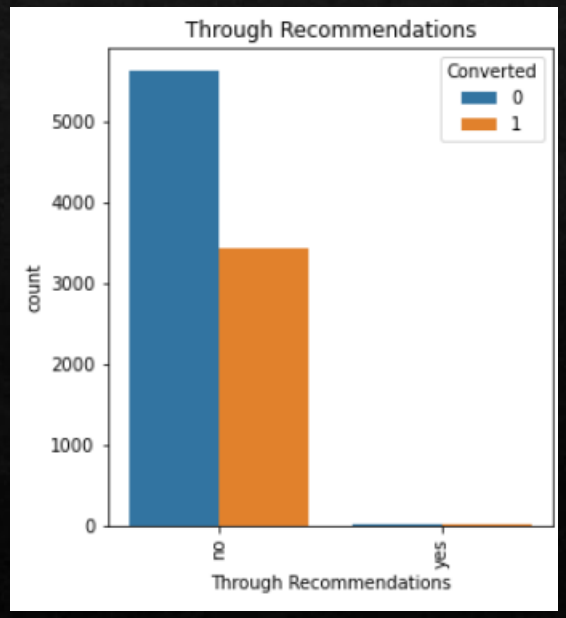
LEAD ORIGIN VS CONVERTED

Landing page submissions has had high lead conversions.



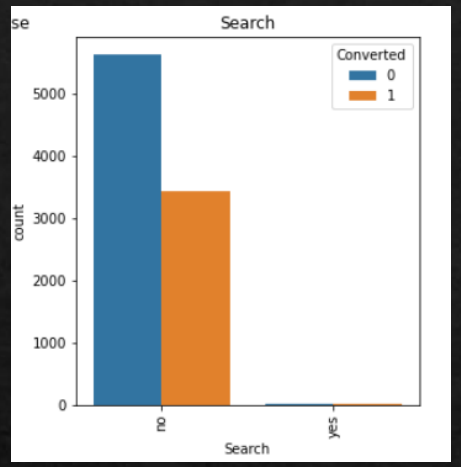
DIGITAL ADVERTISEMENTS VS CONVERTED

Based on the graph digital advertisements do not have promising leads,



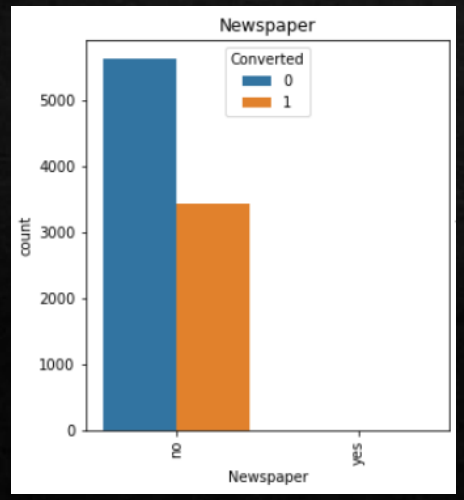
THROUGH RECOMMENDATIONS VS CONVERTED

Based on graph, recommendations are not a good source of promising leads.



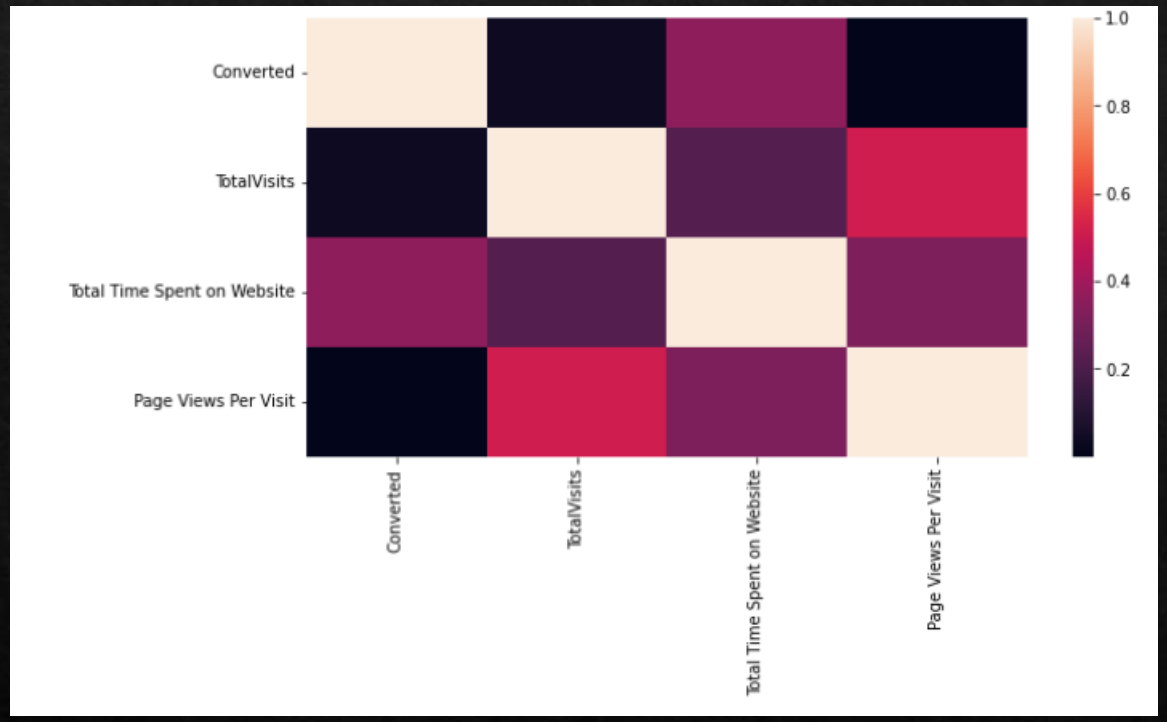
SEARCH VS CONVERTED

Based on the graph shows searches are not good source of leads.



NEWSPAPER VS CONVERTED

Newspaper don’t have high conversion rate.



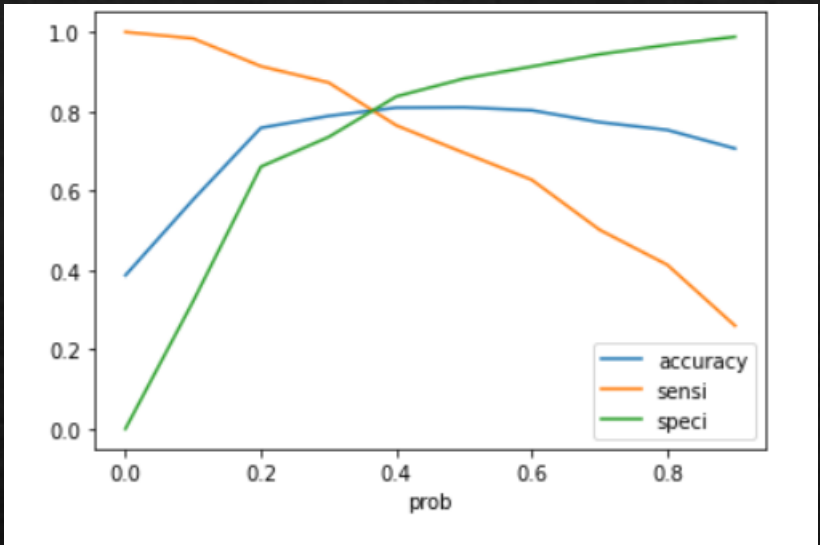
TOTAL TIME SPENT ON WEBSITES VS CONVERTED

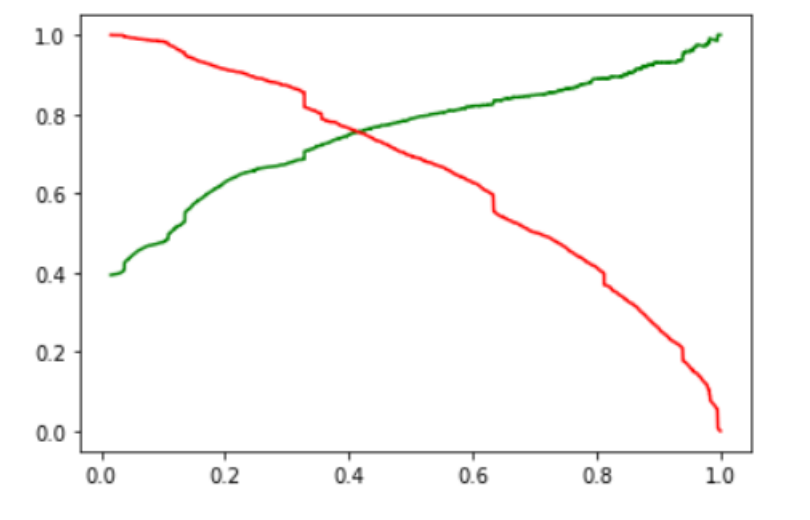
People spending higher than average time are promising leads.

TOTAL VISITS VS CONVERTED

Higher total visits have a slight higher chances of being a promising lead.

MODEL EVALUATION(Optimum cutoff)





ACCURACY SENSITIVITY AND SPECIFICITY

80.3 % Accuracy
80.4% Sensitivity
80.2% Specificity

PRECISION AND RECALL

78.8% Precision
70% Recall

Summary

This analysis is done for X Education and to find ways to get more users to join their courses. The dataset which was provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

The following are the steps used:

1. Cleaning data: The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to ‘not provided’ to not lose much data. Although they were later removed while making dummies.
2. EDA: EDA was done to check the condition of dataset. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good and no outliers were found.
3. Dummy Variables: The dummy variables were created and later the dummies with ‘not provided’ elements were removed. For numeric values we used the MinMaxScaler.
4. Train-Test split: The split was done at 70% and 30% for train and test data respectively.
5. Model Building: Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).
6. Model Evaluation: A confusion matrix was made. Later, the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.
7. Prediction: Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity, and specificity of 80%.
8. Precision – Recall: This method was also used to recheck and a cut off 0.41 was found with Precision around 73% and recall around 75% on the test data frame