Lead Scoring Case Study

By
Priyanka Gupta
Astha Srivastava
C Kalarani

Problem statement and Business Objective

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

- Now, 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.

- **The objective** is to build a model to indentify the hot leads and achieve lead conversion rate to 80%

Data Knowledge

- Dataset used : “Leads.csv “
- Total number of customers present : 9240
- Total number of features : 37
- Model used : Logistic Regression

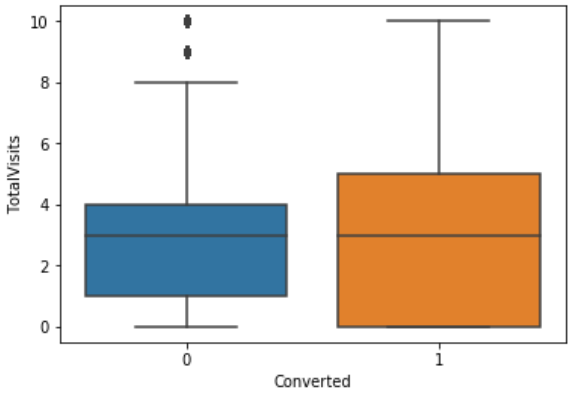
Approach Used to solve Problem

- Importing libraries
- Read the data
- Data Cleaning
- Exploratory data analysis
- Adding dummies
- Split the data into train and test
- Model Building

Data cleaning

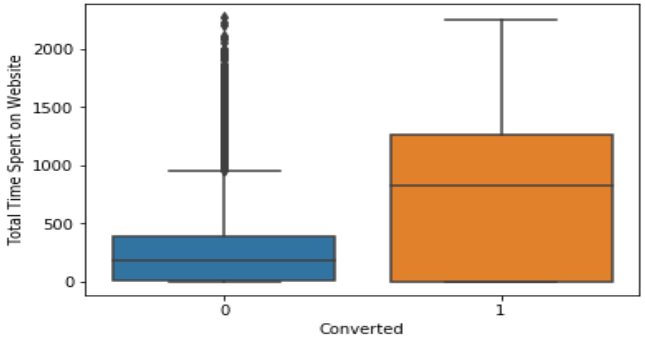
- Replaced “Select” value with NaN.
- Calculation of missing values for each column and dropping Score and Activity variable.
- Dropping the columns who have highest percentage of missing values.

EDA : Numerical Data



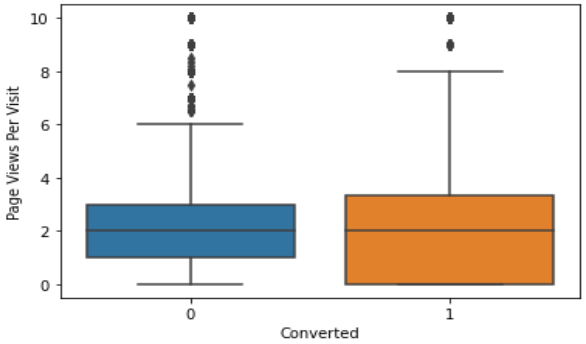
Total Visits

The average total visits for both converted and non converted people is found to be the same.



Total Time Spent On The Website

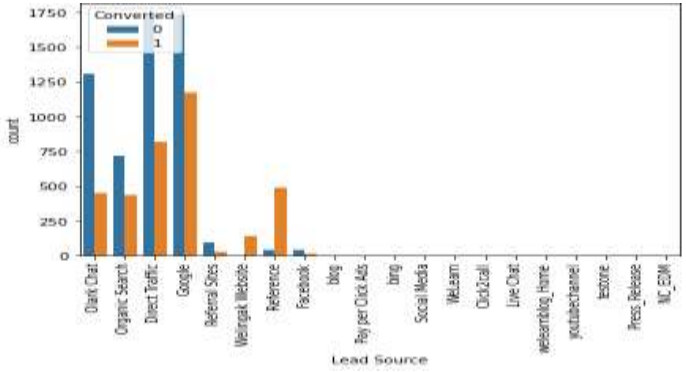
The mean is found to be higher in case of Converted people rather than non converted people.

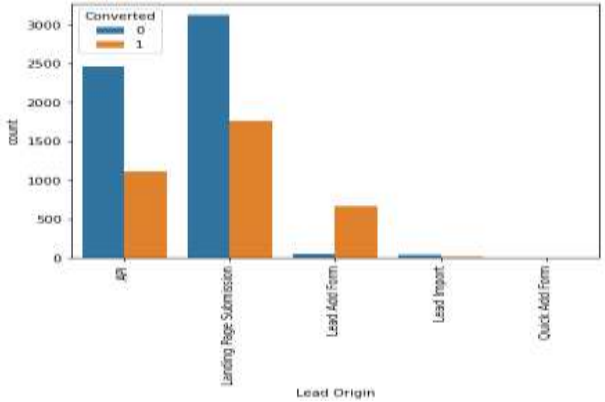


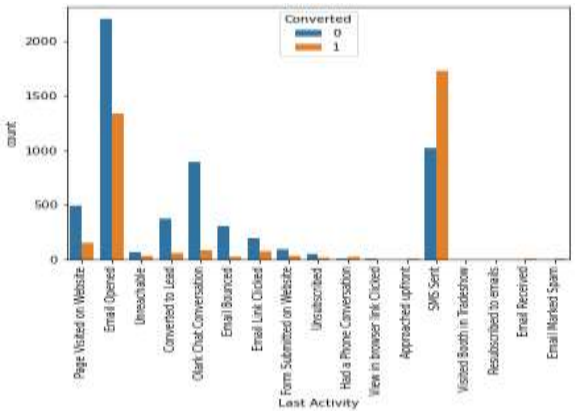
Page Views Per Visit

The average page views for both converted and non converted is found to be the same.

EDA : Categorical Data







Lead Source
Google is found to be the important source for Lead Conversion.
Direct Traffic also proves to be important to secure leads.

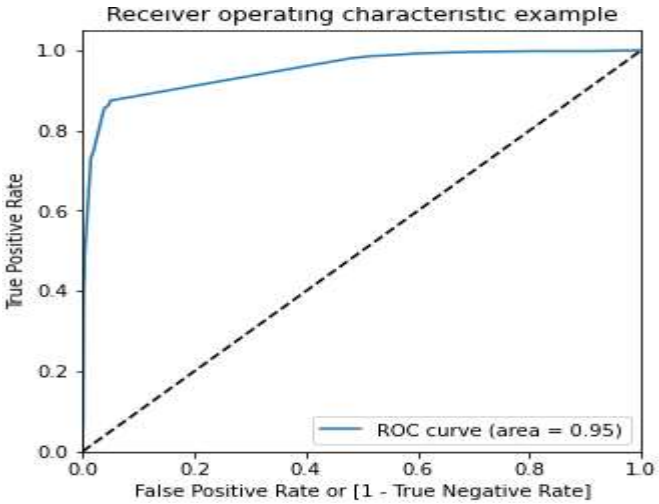
Lead Origin
The percentage of Converted people is found to be greater for Landing Page Submission.
We can also see that if Lead source is Add Form, the ratio of lead conversion is very high (almost not converted is very less).

Last Activity
We need to target people via Emails and SMS as it is found that the probability of response in case Converted leads is found to be higher.

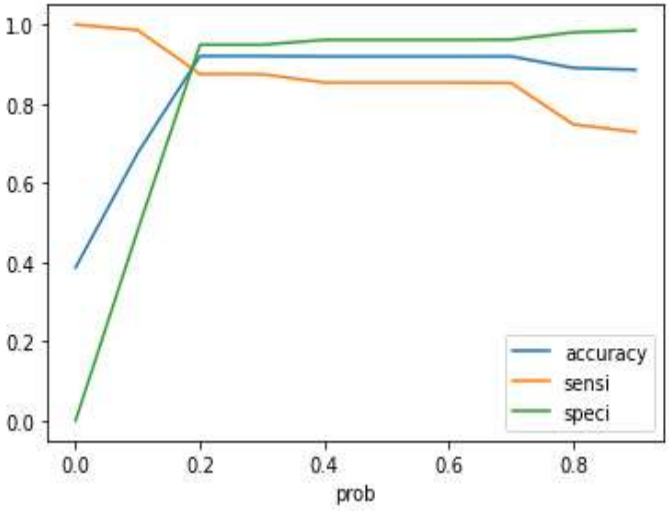
Model Building

- Model – I: We build a basic model using different features. Since it is not efficient we perform RFE to obtain a model. There are so many variables with high p-values and VIF value, we need to remove them.
- Model – II: Removing variable with if p value>=0.5 and VIF >= 5, if p value>=0.5 and VIF<=5, if p value <=5 and VIF>=5, If p value <=5 and VIF<=5(Significant). vif of all the features is above 5 means there is no multicollinearity between independent variables. But p value is high . so, on the basis of 2nd condition .we delete the column with high p value
- Model III : VIF is significant. so on the basis of p value drop Tags_invalid number. here p value of Tags_wrong number given is high.
- Model IV: we can see that all the p value and VIf is significant . so, we can consider it as a final model.
- Final model has 11 features in total.

ROC Curve And Optical Cut-Off Probability



The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. on the basis of above condition the obtained ROC curve is good in terms of accuracy



When we plot the sensitivity, accuracy and specificity of the model together, the optimal cut off point is found to be at 0.2. This means the sensitivity and specificity are found to be balanced.

Model Performance Train

- ACCURACY - 91.96%
- SENSITIVITY - 85.36%
- SPECIFICITY - 96.99%
- PRECISION- 93.26%
- RECALL- 85.36%

Model Performance Test

- ACCURACY – 92.27%
- SENSITIVITY – 89.53%
- SPECIFICITY – 93.98%

Conclusion

- A customer/lead who fills out a form is a potential lead.
- You should focus primarily on working professionals.
- You should primarily focus on prospects whose last activity was SMS sent or Email opened.
- It's always good to focus on the customers who spend a lot of time on our site.
- It's better not to focus too much on the customers whose sent emails are returned.
- If the lead source is a referrer, it can be a potential lead.
- If a lead doesn't fill specialization, they may not know what to study and may not be the right audience. Therefore, it is better not to focus too much on such cases.

Recommendations

- We recommend collecting data and running the model frequently to keep your potential leads up to date.
- It is believed that the best time to call a potential leads is within a few hours of their interest in the course.
- Along with phone calls, it’s good to mail the leads also to keep them reminding as email is as powerful as cold calling.
- Reducing the number of phone attempts to 2-4 and using other media such as Google ads or email to reach out to your leads more frequently can save you a lot of time.
- Focusing on Hot Leads will increase the chances of obtaining more value to the business as the numbers of people we contact are less but the conversion rate is high.