ML-MODEL

LEAD SCORING CASE STUDY

Submitted by

Neha

Sunava Neogy

Naveen Kumar Upadhyay

PROBLEM STATEMENT

- To develop a ML model to help education company to filter their prospect lead and to convert maximum of them with their minimum effort and cost.
- To achieve the sales target of 80%.

GOALS AND OBJECTIVES

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

APPROACH AND METHODOLOGY

- Data Preparation (EDA) – Finding missing values and removing the column which is having>40% missing values. Option select had to be replaced with a null value. Imputing the rest with level ‘Not Provided’. Data to be converted to a suitable format. Perform Univariate, Bivariate and Multivariate Analysis.
- Model Building-Dummy variable creation, Scaling of variables and splitting in train test data. Variable selection with correct techniques and choosing best model out of it. Model evaluation using confusion metrics and compare given target.

EDA

- Finding missing values and removing the column which is having>40% missing values. Option select had to be replaced with a null value. Imputing the rest with level ‘Not Provided’. Data to be converted to a suitable format. In the columns ‘Select’ to be treated as null.
- There is a huge value of null variables in 4 columns. But removing the rows with the null value will cost us a lot of data and they are important columns. So, instead we are going to replace the NaN values with ‘Not Provided'. This way we have all the data and almost no null values. In case these come up in the model, it will be of no use and we can drop it off then.
- Perform univariate, Bivariate and Multivariate Analysis.
- There are many elements that have very little data and so will be of less relevance to our analysis. Eg: the columns having one unique value.
- There aren't any major outliers.

DUMMY VARIABLES

- The dummy variables were created where no of category>2 in categorical variables. Remove the dummy column manually where it is suffix is “Not Provided”
- For numeric values we used the MinMaxScaler.

TRAIN TEST SPLIT AND SCALING

- The split was done at 80% and 20% for train and test data respectively.
- Scaling of features on train and test data using MinMaxScaler.

MODEL BUILDING

- RFE done to attain the top 20 relevant variables.
- Later the rest of the variables were removed manually depending on the VIF values and p-value.(variables with VIF < 5 and p-value < 0.05 to be considered)

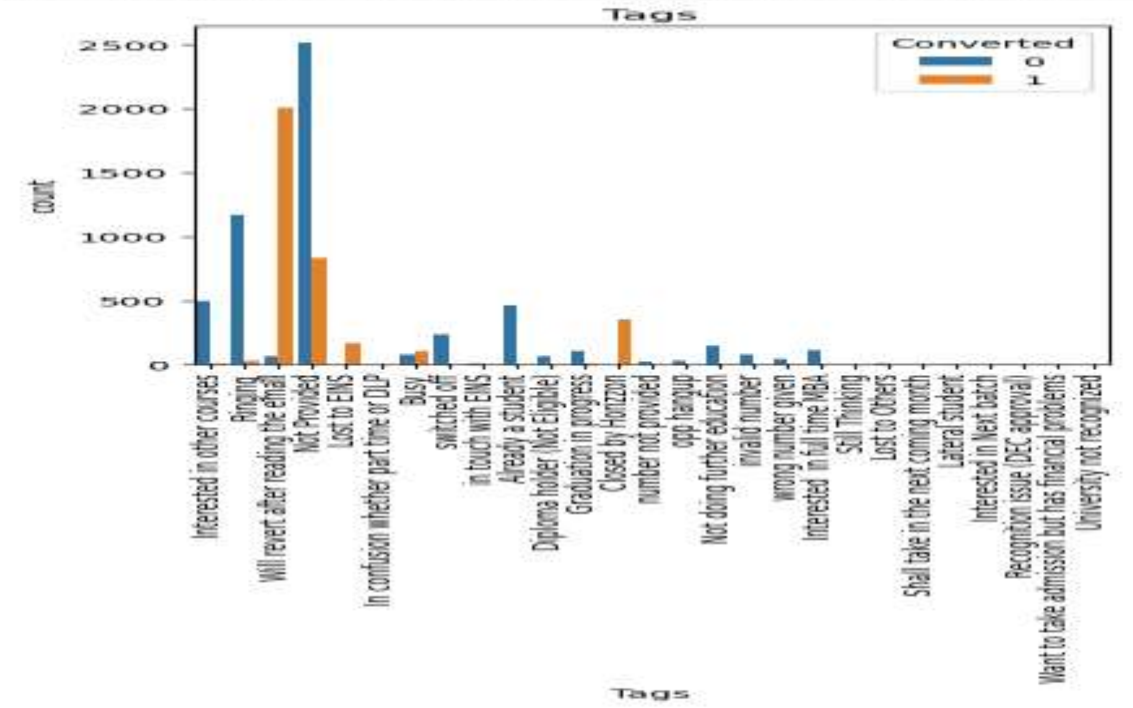
MODEL EVALUATION

- Result evaluated on the basis of confusion matrix.
- With the cut off as 0.5 we have around 93% accuracy, sensitivity of around 88% and precision is 93%,specificity of around 96%.
- The area under ROC curve is 0.97 which is a very good value.
- From the graph it is visible that the optimal cut off is at 0.30.
- With the current cut off as 0.30 we have accuracy ,sensitivity ,specificity all 91% and precision 93% on train data set.
- **Prediction on test set:** With the current cut off as 0.30 we have accuracy-91%, sensitivity-92% ,precision-93% and specificity -90**%.**
- **Precision – Recall :** This method was also used to recheck and a cut off 0.39 was found with Precision around 91.5% and recall around 90% on the train data frame.
- **Prediction on Test set:** With the current cut off as 0.39 we have Precision 92% and Recall around 90.5%.

INTERPRETING RESULTS

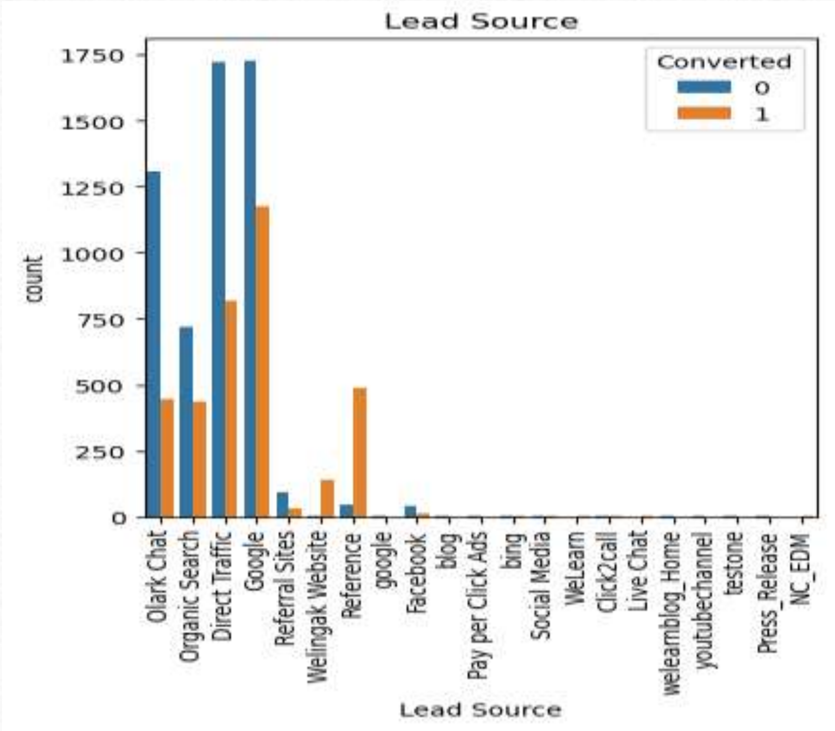
As per graph, we can see some of the variables of Tag column is highly responsible to determine target. Some are +vely related which means converted and some are –vely related which means not-converted.

1 Tags_Closed by Horizzon(converted)
2 Tags_Lost to EINS(converted)
3 Tags_invalid number(not converted)
4 Tags_switched off(not converted)
5 Tags_Already a student(not converted)



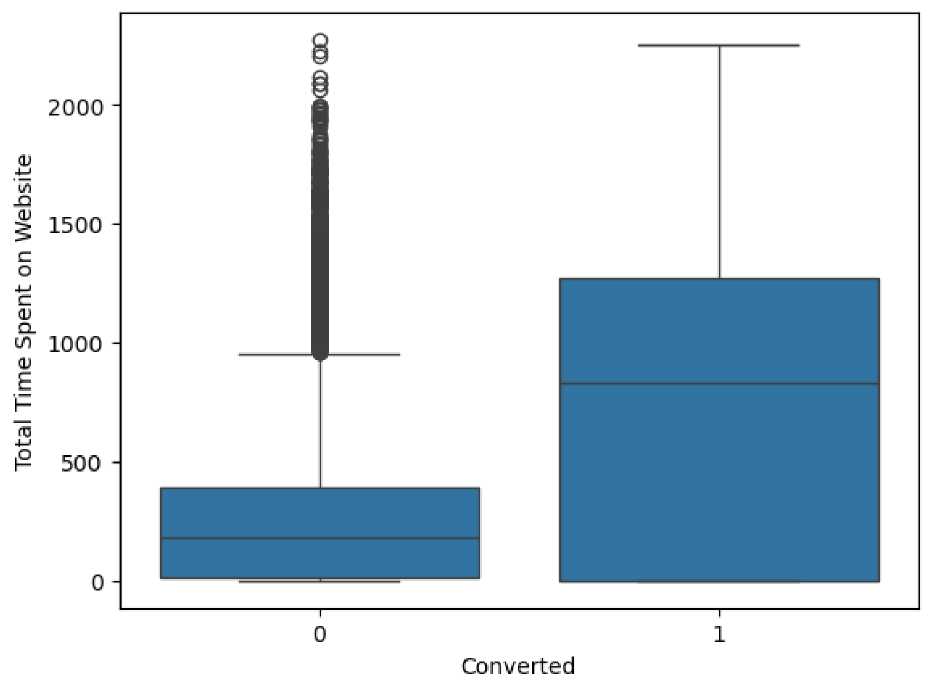
INTERPRETATION ON LEAD SOURCE

- It is clearly seen from the graph Lead Source having category ‘Welingak Website’ is positively related with the target hence we can consider it for achieving the target.



TOTAL TIME SPENT ON WEBSITE

- The medians of both the boxplot is showing that the who spent more time on website the chances of converting them to lead is higher, since the converted one is having high median value.



CONCLUSION

It was found from the model that the variables that mattered the most to decide top 10 in the potential buyers are (In descending order) :

1. Tags_Closed by Horizzon
2. Tags_Lost to EINS
3. Tags_invalid number
4. Tags_switched off
5. Tags_Already a student
6. Tags_Ringing
7. Tags_Not doing further education
8. Lead Source_Welingak Website
9. Tags_Will revert after reading the email
10. Total Time Spent on Website

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.