Lead Scoring Case Study using Logistic Regression

SUBMITTED BY TANMOY BERA

Contents

- Problem Statement
- Business objective
- Problem Approach
- Data Cleaning
- EDA
- Data Preparation
- Model Building
- Model Evaluation
- Conclusion
- Recommendation

Problem statement

- X Education, an online course provider for industry professionals, aims to enhance its lead conversion process by singling out potential hot leads. To achieve this goal, the company is considering several strategies. One key approach involves implementing a lead scoring system that evaluates leads based on criteria such as job title, company size, website engagement, and form responses.
- Leads with higher scores are given priority, as they are deemed more likely to convert. Additionally, behavioral analysis of website visitors helps identify patterns indicative of strong interest, enabling the company to pinpoint leads showing elevated engagement. By segmenting leads based on demographics and industry, personalized communication strategies can be employed to address specific needs.
- Analyzing lead sources and leveraging machine learning algorithms further enhances the identification of leads with higher conversion potential. Regular feedback from the sales team aids in refining the lead scoring criteria, ensuring continuous optimization of the overall process.
- The aim is to create a more efficient system where the sales team can focus on communicating with leads that have a higher likelihood of conversion, ultimately increasing the overall lead conversion rate.

Business Objective

- To 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.
- To adjust to if the company's requirement changes in the future so you will need to handle these as well.
- The CEO want to achieve a lead conversion rate of 80%.

Problem Approach

- Importing the data and inspecting the data frame.
- Data preparation
- EDA
- Dummy variable creation
- Test-Train split
- Feature scaling
- Correlations
- Model Building
- Model Evaluation
- Making predictions on test set

Data Cleaning

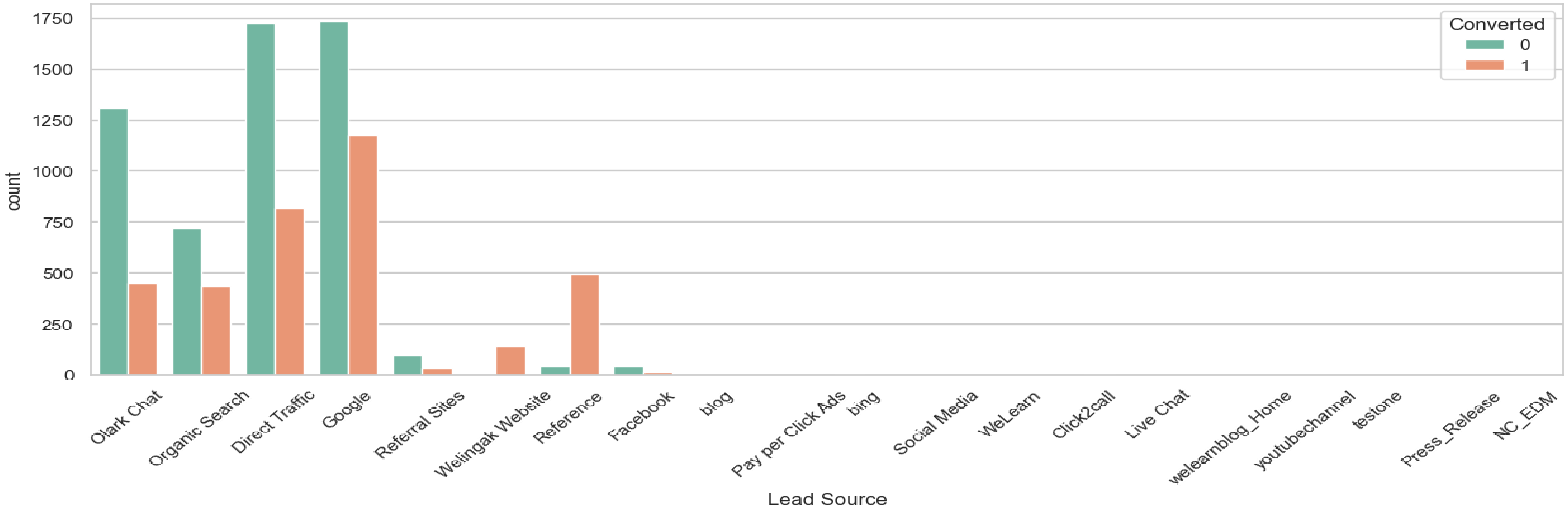
- First step to clean the dataset we dropped the columns having NULL values greater than 30%.

- Next, we removed the imbalanced and redundant variables. This step also included imputing the missing values as and where required. The outliers were identified and removed. Also, in one column was having identical label in different cases (first letter small and capital respectively). We fixed this issue by converting the label with first letter in small case to upper case.

- Then we chose to drop the variables having unique values.

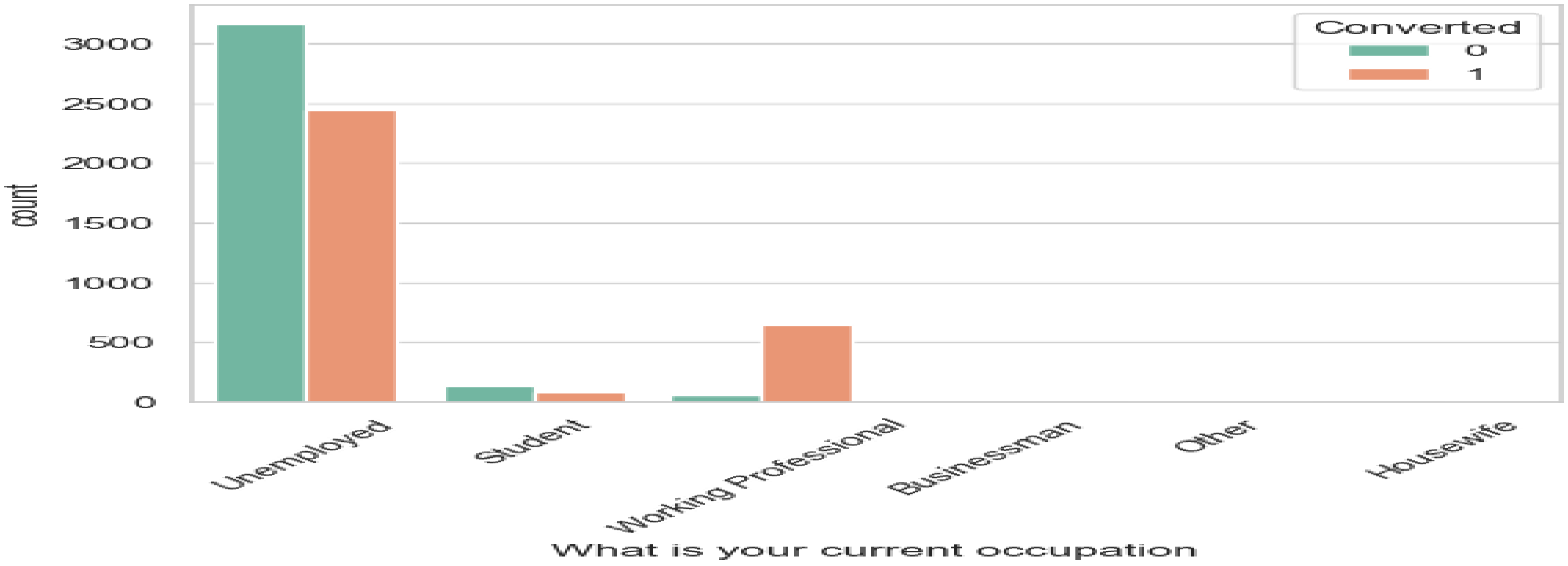
EDA

- Maximum Leads are generated by Google and Direct Traffic. Conversion rate of Reference leads and Welinkgak Website leads is very high.



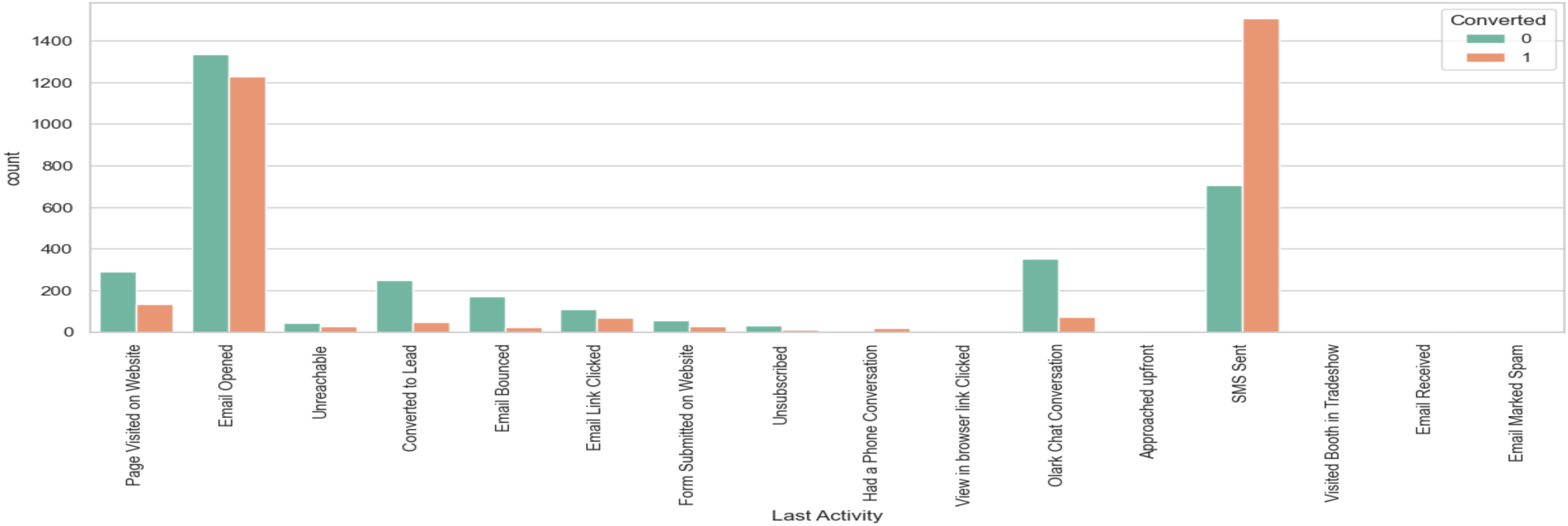
EDA

- Maximum leads generated are unemployed and their conversion rate is more than 50%.Conversion rate of working professionals is very high.



EDA

- Maximum leads are generated having last activity as Email opened but conversion rate is not too good. SMS sent as last activity has high conversion rate.



Data Preparation

- Changed the binary variables into ‘0’ and ‘1’.
- Dummy variables were generated for categorical features.
- Remove all the redundant and repeated variables.
- The dataset was divided into training and testing sets with a split ratio of 70:30.
- Feature scaling was conducted using Min Max Scaler methods to normalize the features.
- Checked for correlation among the variables.
- Removed the highly correlated variables such as 'Lead Source_Olark Chat','Lead Origin_Landing Page Submission’.

Model Building

- Recursive Feature Elimination (RFE) was employed to streamline the dataset, initially consisting of 26 columns, down to 15 through a careful selection process.
- The primary goal of this strategic reduction was to enhance both the efficiency and accuracy of the model. Following RFE, additional refinement was performed through manual fine-tuning.
- In the pursuit of model optimization, a manual feature reduction technique was applied, leading to the exclusion of variables with a p-value greater than 0.05. This step aimed to ensure that only statistically significant features were retained.
- After undergoing four iterations, Model 4 demonstrated stability with p-values consistently below 0.05.
- Then, a manual feature reduction technique was applied, leading to the exclusion of variables with a VIF value greater than 5. This step aimed to ensure that no multicollinearity present there.
- Then after 5th iteration we got the final model, which is selected for Model Evaluation, and further predictions.

Model Evaluation

- Area under ROC curve is 0.84 out of 1 which indicates a good predictive model.
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.
- At a cut-off value of 0.39, the model exhibited a sensitivity of 75.19% , accuracy of 77.83% and specificity of 80.26% for the train set .
- Sensitivity, in this context, signifies the model's capacity to accurately identify converting leads out of the total potential leads.

Conclusions

- While we have checked both Sensitivity-Specificity, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 77.36%, 76.97% and 77.73% which are approximately closer to the respective values calculated using trained set.
- Hence overall this model seems to be good.

Recommendations

To improve our lead conversion rates, here are some recommendations:

- Utilize features with positive coefficients in targeted marketing strategies to effectively attract potential leads.
- Focus efforts on acquiring high-quality leads from the most successful lead sources identified.
- Optimize communication channels by assessing their impact on lead engagement for more effective outreach.
- Consider allocating a higher budget for advertising and related activities on the Welingak Website.
- Adopt an aggressive approach in targeting working professionals, capitalizing on their high conversion rates and potentially stronger financial capabilities that allow for higher fees.

THANK YOU