LEAD SCORING CASE STUDY

By Mona Narula, Nguyen Thi Thu Huong, Kartikey Ishra

Problem Statement

X Education Company wishes to identify the most potential leads, also known as “Hot Leads”.

Hot Leads are basically the people who are engaging with the company’s website/videos/sales calls and have high probability of converting into customers of X Education Company, thus bringing revenue to the company.

That’s why X Education wants to identify all such people who are their potential customer to boost its growth.

Problem Statement

The company needs a model wherein a lead score is assigned to each of the leads such that the customer with higher lead score have a higher conversion chance and customer with lower lead score have a lower conversion chance

The CEO, has given a ballpark number for the lead conversion rate i.e. 80% .

Approach in Nutshell

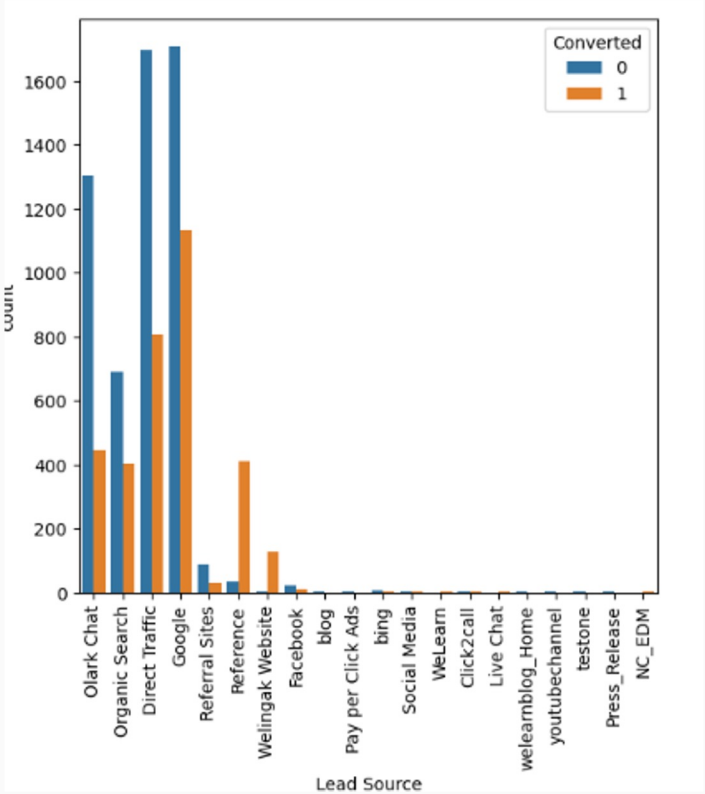
1. Importing Data, Inspecting the Dataframe
2. Data Preparation (Encoding Categorical Variables, Handling Null Values)
3. EDA (univariate analysis, outlier detection, checking data imbalance)
4. Dummy Variable Creation
5. Test-Train Split
6. Feature Scaling
7. Looking at Correlations
8. Model Building (Feature Selection Using RFE, Improvising the model further
9. inspecting adjusted R-squared, VIF and p-vales)
10. Build final model
11. Model evaluation with different metrics Sensitivity, Specificity

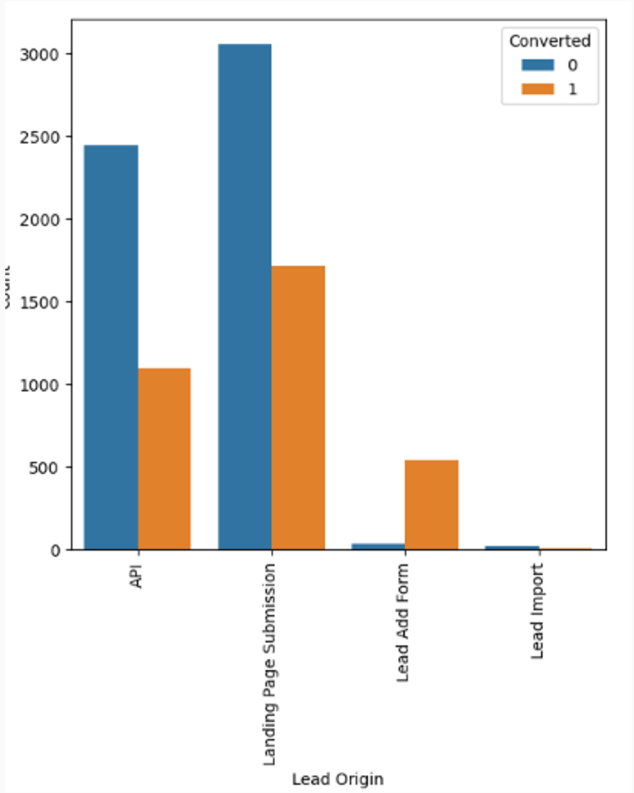
BUSINESS GOAL

Build a logistic regression model to assign a lead score ranging from 0 to 100 to each lead, allowing the company to effectively target potential leads. A higher score indicates a "hot" lead, meaning it is highly likely to convert, while a lower score suggests a "cold" lead with a lower chance of conversion.

ın addition, we need to address other potential challenges outlined in a separate document provided by the company. These challenges should be considered when developing the logistic regression model and will be included in the final presentation (PPT) along with recommendations. Please ensure that the document is filled based on the logistic regression model obtained in the initial step.

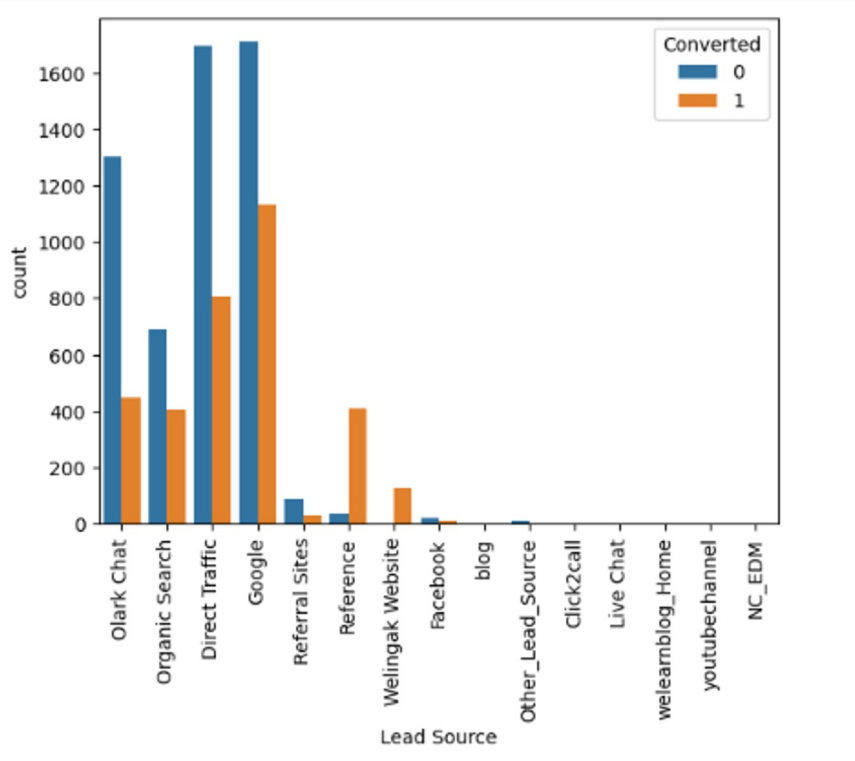
To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'API' and 'Landing Page Submission' and also increasing the number of leads from 'Lead Add Form'



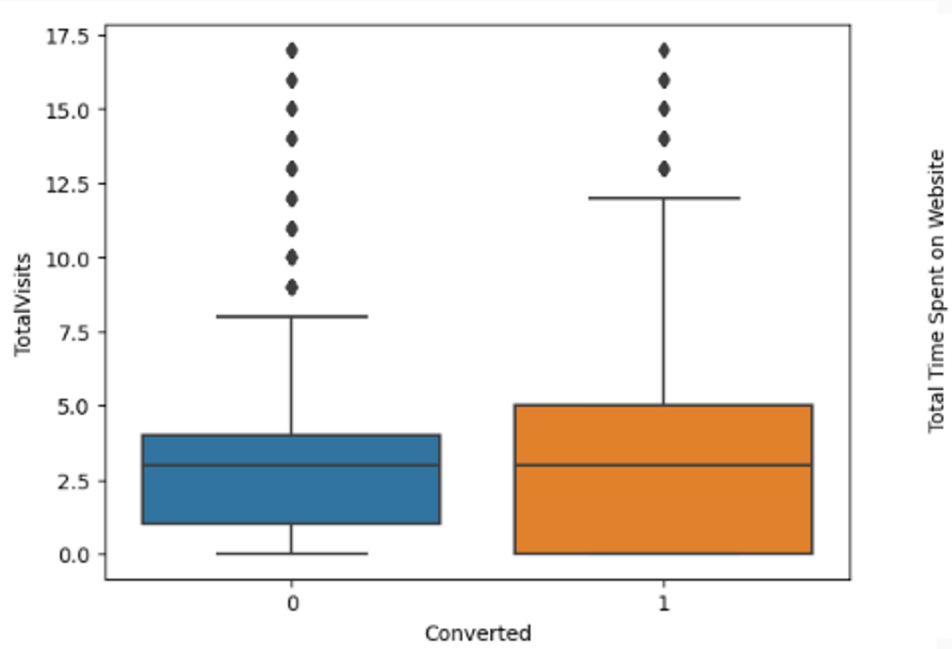


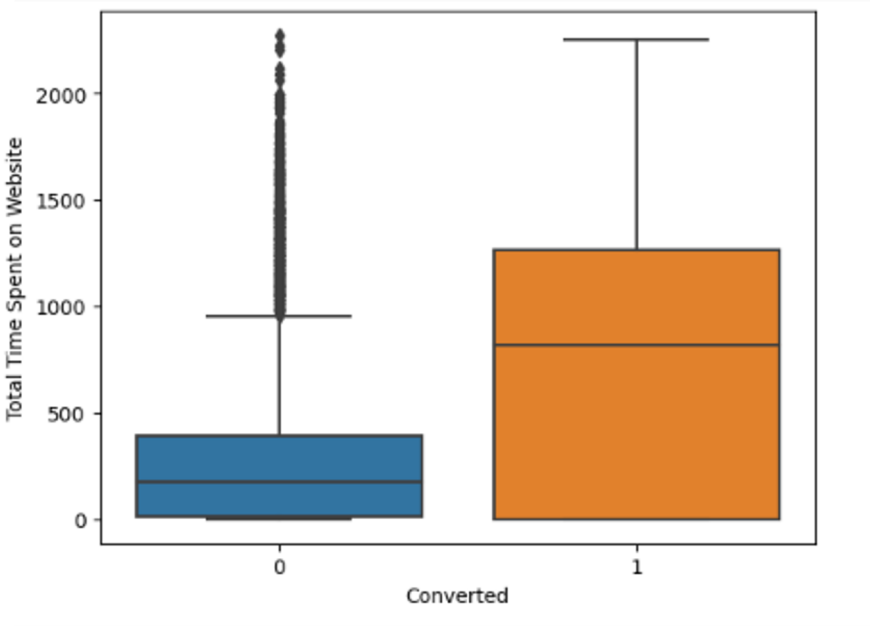
To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'Google', 'Olark Chat', 'Organic Search', 'Direct Traffic' and also increasing the number of leads from Referral sites, 'Reference' and 'Welingak Website'

'

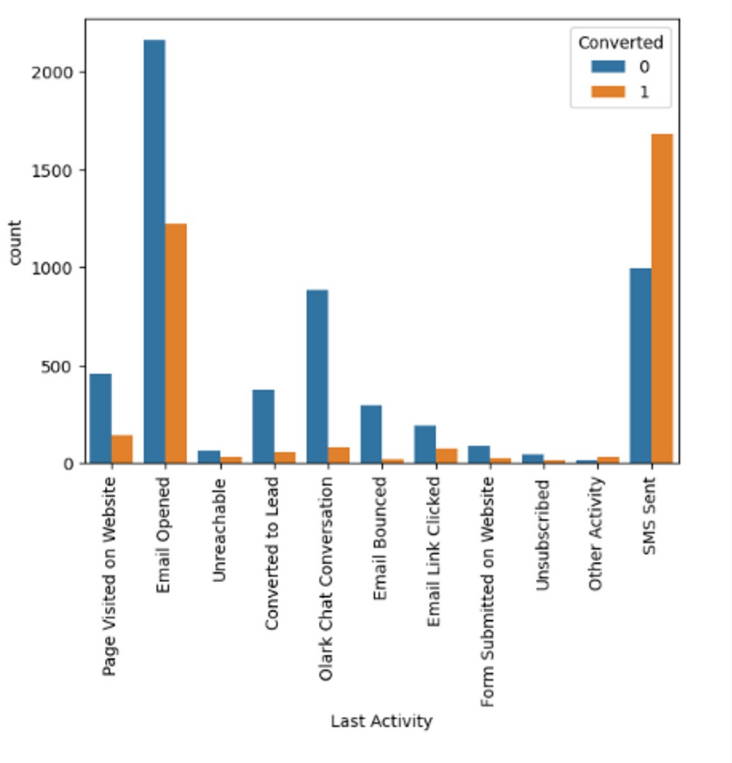


Websites can be made user friendly and more appealing so as to increase the time of the Users on websites

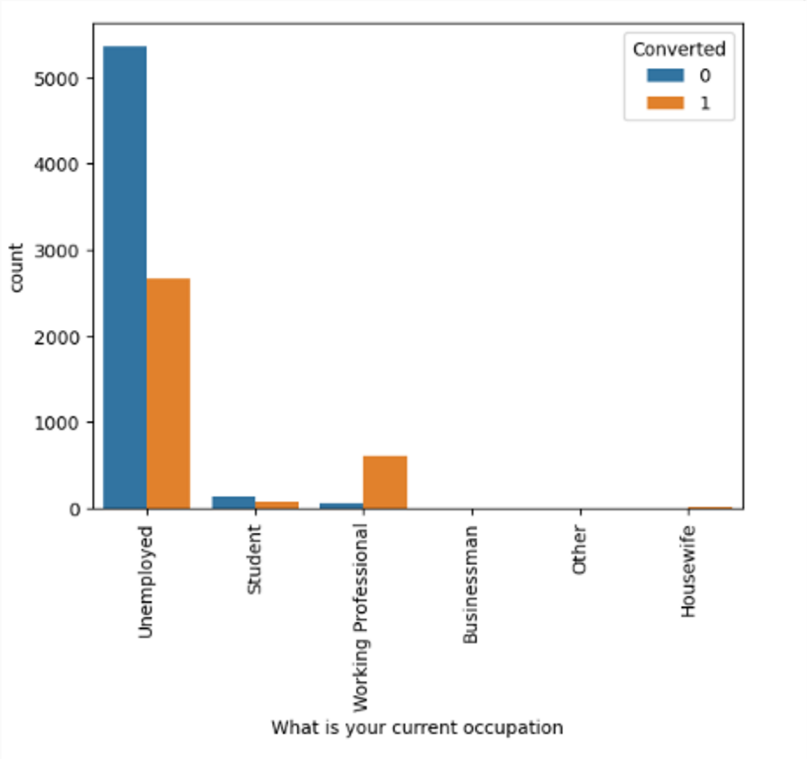




We should focus on increasing the conversion rate of those having last activity as Email Opened by making a call to those leads and also try to increase the count of the ones having last activity as SMS sent



To increase overall conversion rate, we need to increase the number of Working Professional leads by reaching out to them through different social sites such as LinkedIn etc. and also on increasing the conversion rate of Unemployed leads



SUMMARY

- To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'API' and 'Landing Page Submission' Lead Origins and also increasing the number of leads from 'Lead Add Form'
- To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'Google', 'Olark Chat', 'Organic Search', 'Direct Traffic' and also increasing the number of leads from 'Reference' and 'Welingak Website'
- Websites can be made more appealing so as to increase the time of the Users on websites
- We should focus on increasing the conversion rate of those having last activity as Email Opened by making a call to those leads and also try to increase the count of the ones having last activity as SMS sent

SUMMARY

- To increase overall conversion rate, we need to increase the number of Working Professional leads by reaching out to them through different social sites such as LinkedIn etc. and also on increasing the conversion rate of Unemployed leads
- We also observed that there are multiple columns which contains data of a single value only. As these columns do not contribute towards any inference, we can remove them from further analysis

Train-Test Split & Scaling

- Split the dataset into a 70% training set and a 30% test set.
- Applied min-max scaling to the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website'] for uniform scaling.

Model Building

- Split the dataset into a 70% training set and a 30% test set.
- Applied min-max scaling to the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website'] for uniform scaling.

Model Evaluation

- Creating a data frame to store the actual converted flag and the predicted Lead Score probabilities.
- Reshaping the predicted probabilities array for further analysis.
- Classifying the prospects as likely to convert (1) or not likely to convert (0) based on a cutoff probability of 0.5.
- Creating a confusion matrix to evaluate the model's performance.
- Calculating the accuracy score of the model.
- Checking the Variance Inflation Factors (VIF) for the predictor variables to identify multicollinearity.
- Calculating sensitivity, specificity, false positive rate, positive predictive value, and negative predictive value.
- Plotting the Receiver Operating Characteristic (ROC) curve to evaluate the model's performance.
- Finding the optimal cutoff point by calculating accuracy, sensitivity, and specificity for different probability cutoffs.
- Making the final predictions using a cutoff of 0.3.

Results Achieved

The final model achieved an accuracy score of 91.77% on the training dataset, indicating a high level of accuracy in classifying leads as potential or non-potential. This evaluation metric provided confidence in the model's predictive capabilities. The model's precision, recall, and F1 score were also assessed, indicating its effectiveness in identifying 'Hot Leads' and reducing false positives or false negatives.

Conclusion

This summary report outlined the methodology employed in the data analysis assignment for X Education. By leveraging data cleaning, exploratory analysis, feature engineering, and model building, model evaluation the company aimed to improve its lead conversion rate by identifying 'Hot Leads'. The insights gained from the analysis, along with the achieved results, can help X Education optimize their sales efforts, focus on potential leads, and increase the overall conversion rate.

THANK YOU