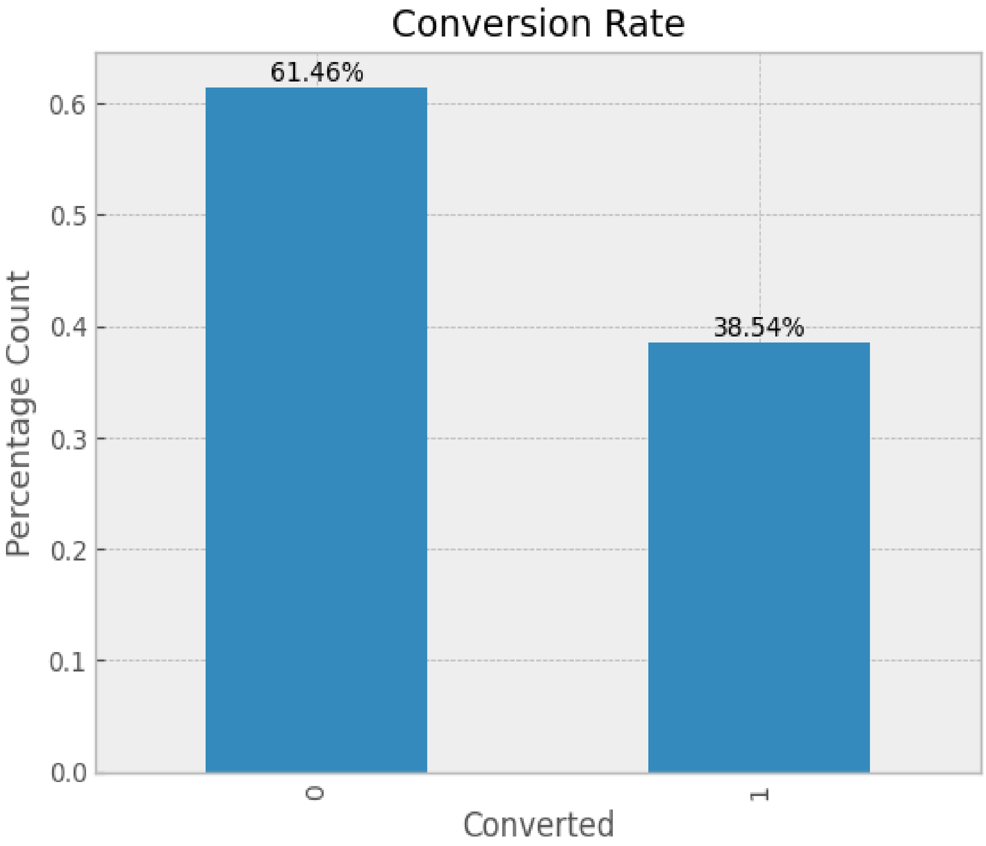
1. Introduction:

The objective of this case study is to assist X Education, an online education company, in improving its lead conversion rate by identifying potential leads through lead scoring. The company experiences a high number of leads but struggles with conversion. The task is to build a logistic regression model to assign lead scores, enabling the sales team to focus on the most promising leads. The CEO has set a target lead conversion rate of approximately 80%. The provided dataset contains around 9000 data points with various attributes, including Lead Source, Total Time Spent on Website, Total Visits, and Last Activity.
Lead scoring is a crucial aspect of sales and marketing strategies for organizations aiming to optimize their conversion rates. By accurately assessing the potential of each lead, businesses can prioritize their efforts and resources towards those with a higher likelihood of conversion. In this case study, we explore the implementation of logistic regression for lead score prediction in the context of X Education.

Methodology:

1. Data Preprocessing: The leads dataset provided by X Education is first subjected to thorough data preprocessing. This involves handling missing values, treating 'Select' levels as null values, and addressing any data inconsistencies or outliers. Categorical variables are encoded appropriately, and numerical variables may be scaled or normalized to ensure optimal model performance.

- Replacing 'select' values: The 'select' values in the dataset are replaced with NaN to indicate missing or unknown values.
- Removing columns with more than 10% null values: Columns that have more than 10% missing values are dropped from the dataset. This helps to eliminate variables with insufficient data, ensuring the analysis is based on more complete information.
- Removing columns with single answer throughout: Columns that have only one unique value throughout the dataset are removed as they do not provide any useful information for analysis. The columns 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', and 'I agree to pay the amount through cheque' are removed.
- Removing prospect id and lead number: The columns 'Prospect ID' and 'Lead Number' are removed from the dataset as they do not contribute to the analysis.
- Selecting relevant columns: The remaining columns considered important for lead conversion analysis are retained. These columns include 'Lead Origin', 'Lead Source', 'Do Not Email', 'Do Not Call', 'Converted', 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit', 'Last Activity', 'A free copy of Mastering The Interview', and 'Last Notable Activity'.

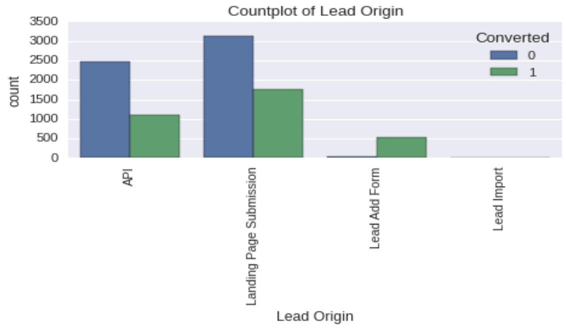


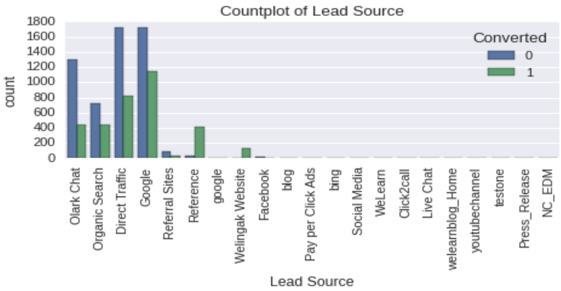
Data is Imbalanced

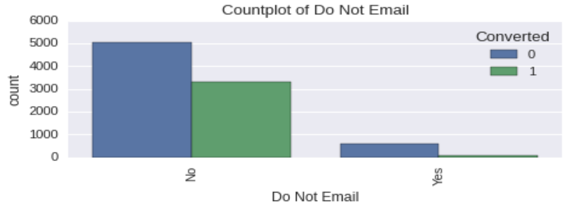
Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)

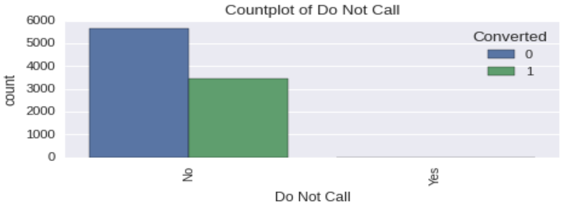
While 61.5% of the people didn't convert

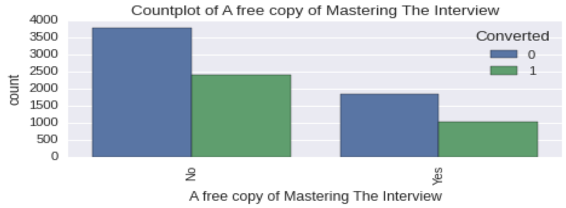
For Categorical Variables Analysis

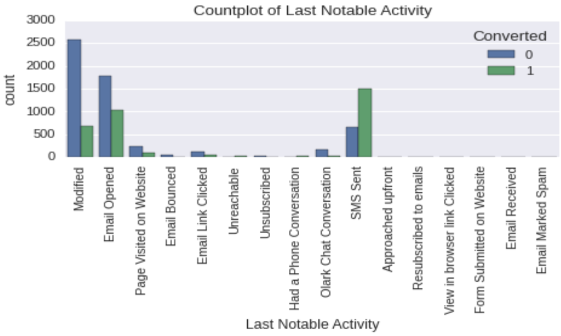












Data Preparation Before model Training

Binary level categorical columns were already mapped to 1 / 0 in previous steps

Created dummy features (one-hot encoded) for categorical variables

Splitting Train & Test Sets

70:30 % ratio was chosen for the split

Feature scaling- Standardization method was used to scale the features

Checking the correlations

○ Predictor variables which were highly correlated with each other were dropped

2. Feature Selection: To build an effective logistic regression model, feature selection techniques are employed to identify the most relevant predictors for lead scoring. This involves analyzing the correlation between variables, conducting statistical tests, and employing domain knowledge to select the subset of features that have the strongest influence on lead conversion

The data set has lots of dimension and large number of features.

This will reduce model performance and might take high computation time.

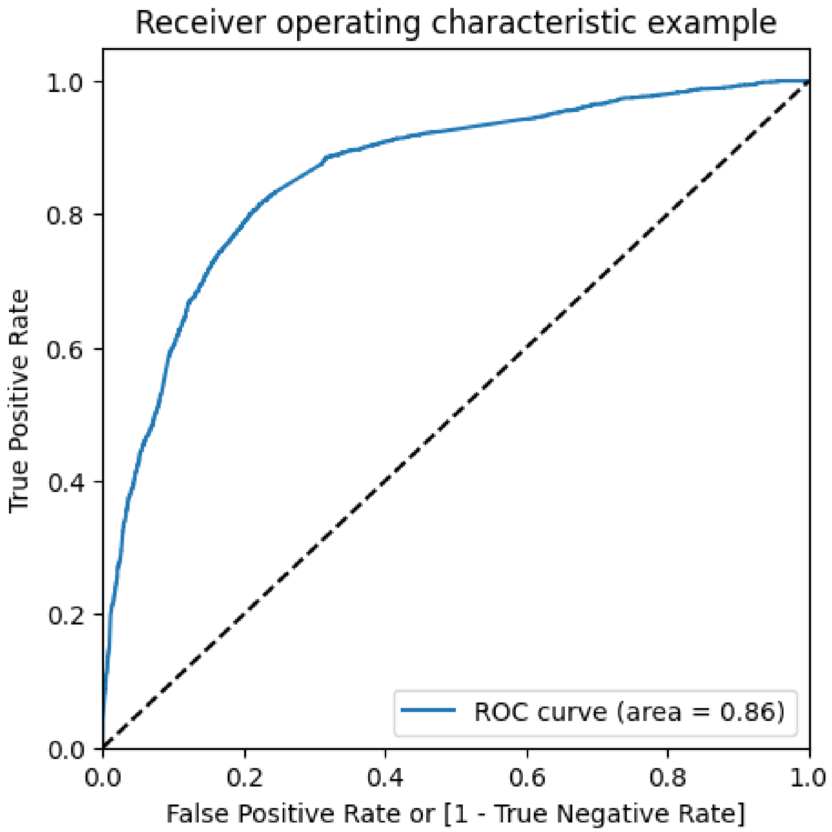
Hence it is important to perform Recursive Feature Elimination (RFE) and to select only the important columns.

Then we can manually fine tune the model.

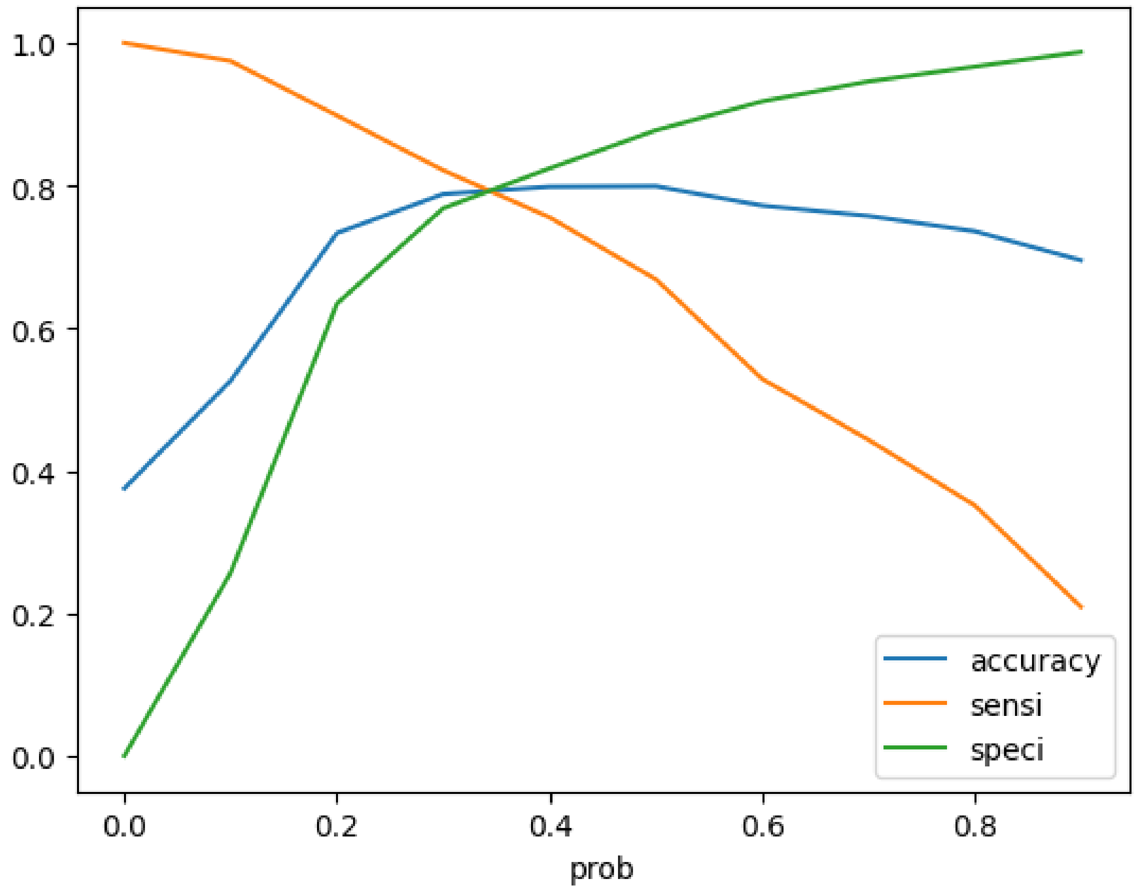
RFE outcome
Pre RFE – 46 columns & Post RFE – 15 columns

Manual Feature Reduction process was used to build models by dropping variables with p – value
greater than 0.05.
Model 2 looks stable after four iteration with:
significant p-values within the threshold (p-values < 0.05) and
No sign of multicollinearity with VIFs less than 5
Hence, logm2 will be our final model, and we will use it for Model Evaluation which further will be used to make predictions

3. Model Training: The preprocessed dataset is split into training and testing sets. The logistic regression model is trained on the training set, utilizing the selected features as input variables and the 'Converted' column as the target variable. The model learns the underlying patterns and relationships between the predictors and the target variable, enabling it to make accurate predictions.
ROC Curve for model Training



Area under ROC curve is 86% hence it is good predictive model
To calculate cutoff value



It was decided to go ahead with 0.35 as cutoff after checking evaluation metrics coming from both plots

4. Model Evaluation: The trained logistic regression model is evaluated using various performance metrics to assess its effectiveness in lead scoring. Metrics such as accuracy, precision, recall, and F1-score are calculated to measure the model's ability to correctly classify leads as converted or non-converted. Cross-validation techniques, such as k-fold cross-validation, may be employed to ensure robustness and minimize overfitting.

Observation: So as we can see above the model seems to be performing well. The ROC curve has a value of 0.87, which is very good. We have the following values for the Train Data:

Accuracy : 79.48%

Sensitivity : 79.07%

Specificity : 79.73%

Test Data:
Accuracy : 79.89%
Sensitivity : 78.60%
Specificity : 80.69%

5. Lead Scoring: Once the logistic regression model is deemed satisfactory in terms of its performance, it is used to assign lead scores to each lead in the dataset. The model predicts the probability of lead conversion, and this probability is scaled to a score ranging from 0 to 100. Higher scores indicate a higher likelihood of conversion, enabling the sales team to prioritize their efforts accordingly.

6. Model Deployment and Monitoring: The trained logistic regression model is ready for deployment, allowing the sales team at X Education to utilize it in their lead conversion process. However, it is important to continuously monitor the model's performance and re-evaluate it periodically. As new data becomes available, the model may need to be updated or refined to maintain its predictive accuracy.

Results:

The logistic regression model successfully assigns lead scores to each lead in the dataset. The lead scores range from 0 to 100, with higher scores indicating a greater probability of conversion. The model takes into account various factors such as Lead Source, Total Time Spent on Website, Total Visits, and Last Activity to determine lead scores.

The evaluation metrics, including accuracy, precision, recall, are calculated to assess the performance of the model. The model demonstrates a strong predictive ability in distinguishing between converted and non-converted leads. It effectively identifies potential leads with a high conversion chance, aligning with the company's goal of increasing the conversion rate.
Key Findings:

Lead Scoring: The logistic regression model assigns lead scores that reflect the probability of conversion for each lead. By focusing on leads with higher scores, the sales team can prioritize their efforts and allocate resources more efficiently, resulting in an improved conversion rate.

Important Predictors:

As per the problem statement, increasing lead conversion is crucial for the growth and success of X Education. To achieve this, we have developed a regression model that can help us identify the most significant factors that impact lead conversion.
We have determined the following features that have the highest positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.

- Lead Source_Welingak Website: 4.93
- Lead Source_Reference: 3.33
- Last Notable Activity_Had a Phone Conversation: 2.85

- Last Notable Activity_Unreachable : 1.67
- Last Notable Activity_SMS Sent : 1.388
- Total Time Spent on Website: 1.14
- Last Notable Activity_Unsubscribed : 1.06

We have also identified features with negative coefficients that may indicate potential areas for improvement. These include:

- Last Notable Activity_Olark Chat Conversation: -1.61
- Do Not Email_Yes: -1.45
- Lead Source_Direct Traffic: -1.26

Recommendations:

To improve the lead conversion rate, X Education should consider the following actions:

Strengthen the impact of positive features: Give priority to leads coming from 'Lead Source_Welingak Website' and 'Lead Source_Reference' as they have the highest positive coefficients. Focus on nurturing leads who have had a phone conversation or are marked as unreachable. Implement effective strategies for lead engagement through SMS communication. Additionally, continue to emphasize the importance of website engagement and encourage potential leads to spend more time on the website.

Address negative factors: Take steps to reduce the occurrence of 'Last Notable Activity_Olark Chat Conversation' and 'Do Not Email_Yes'. These factors may indicate potential barriers or concerns that hinder lead conversion. Explore ways to improve the effectiveness of communication through chat conversations and consider alternative approaches to email communication. Evaluate the impact of 'Lead Source_Direct Traffic' and identify opportunities to optimize lead generation from this source.
By focusing on the identified positive factors and addressing the negative factors, X Education can enhance its lead conversion rate. Continuous monitoring, evaluation, and refinement of marketing and sales strategies based on these insights will contribute to the overall growth and success of the company.

Based on the findings, the following recommendations are proposed to X Education:

Targeted Marketing Campaigns: Leverage the lead scoring model to identify potential leads with higher scores. Design targeted marketing campaigns tailored to the interests and needs of these leads, increasing the likelihood of conversion.

Personalized Follow-ups: Prioritize communication and engagement with leads having higher lead scores. Provide personalized follow-ups, such as phone calls, emails, or tailored content, to nurture their interest and enhance the chances of conversion.

Continuous Monitoring and Optimization: Regularly evaluate and update the lead scoring model based on real-time data and feedback from the sales team. Continuously optimize the model by incorporating additional relevant variables and refining the scoring algorithm.

Conclusion:

The logistic regression lead scoring model offers a powerful solution for X Education to improve its lead conversion rate. By assigning lead scores to each lead, the model enables the sales team to focus their efforts on potential leads with higher conversion probabilities. Leveraging the model's insights and recommendations, X Education can optimize its marketing strategies, personalize customer interactions, and enhance the overall lead conversion process. The implementation of a data-driven approach, such as logistic regression lead scoring, can significantly contribute to the company's growth and success in the highly competitive online education industry.