Client Name: X Educations
Sector: Education
Prepared By: Vinay Chawla
Submitted on: 1st March 2025

Lead Score Analysis – A Logistic Regression based ML Modelling Project.

Problem Statement: Company X Educations is struggling with low lead conversion rate of 30% and wants to increase the same above 80%.

Project Objective: Calculate Lead score for each lead. find the right threshold above which if calls are prioritized, the success rate is higher than 80%.

Input Information:

- What is a Lead? The company markets its courses on several websites, search engines, and even social media sometimes. Once these people land on the website, they might browse the courses, fill out a form for the course, or watch some videos. When these people fill out a form with their email address or phone number, they are classified as leads.
- Hot Leads: Most Potential leads that have high potency to be converted to Successful sales.
- present success sale conversion of leads: 30%
- Usual input for leads: social media marketing such as:
  - website visits. visitors fill some forms which generates leads.
  - Past referrals.
- Usual Lead to sales conversion process: Social media marketing input > initial pool of leads > HOT leads > Converted sales.

Summary of the major steps followed:

Lead score analysis Project – Summary Report

Based on the Machine Learning Model built on the Leads dataset

Client Name: X Educations
Sector: Education
Prepared By: Vinay Chawla
Submitted on: 1st March 2025

- Performed EDA. left with 13 clean columns
- applied dummy conversion to categorical columns. converted to a decent sized 48 columns.
- applied VIF function to remove multicollinear features left with 40 features.
- Standardizing the features
- applied logistic regression
- Searching the right value of lambda (C value) by trials using Grid search
- Refitting the model as per the best C value

model building finished here. Evaluation and testing of model started here afterwards:

- Predictions. Putting Test data to model and getting the predicted value from this model.
- Model evaluation by comparing Actual Y value (Test dataset) vs predicted values of Y
  - Establishing model capability with help of ROC curve.
  - Calculating Predict probabilities for ROC Curve using x test scaled data on predict_proba function.
  - Compute Confusion Matrix to validate the model capability to distinguish the good and bad calls.
  - Calculating precision, recall

(Also, Extract features that impact the Model significantly.

Bigger the bar (Positive or Negative), higher the impact on results.)

  - Model Evaluation by finding accuracy, recall, precision, F1Score, AUC, confusion matrix

Lead score analysis Project – Summary Report

Based on the Machine Learning Model built on the Leads dataset

Client Name: X Educations
Sector: Education
Prepared By: Vinay Chawla
Submitted on: 1ˢᵗ March 2025

- o Optimize Threshold Trade-off based on Precision Recall Curve
- o Compute Precision Recall curve
- o checking Accuracy with Optimized Threshold

- OUTCOME Check:
  - o Calculating Lead score
  - o Using lead scores >50 to convert to a recommended HOT LEAD.
 - o compare actual "Converted" to HOT LEAD to get success rate.

Visuals from the Project

➢ Exploratory Data Analysis

Observations : EDA

➢ Preliminary Data Cleaning
  ▪ Removed Features with Nulls >45%
  ▪ Removed Index columns
➢ Exploratory Data Analysis
  ▪ Individual Column wise approach
  ▪ Checking Categorical data
  ▪ Removing Outliers
  ▪ Grouping categorical data if needed

Examples of Category Grouping
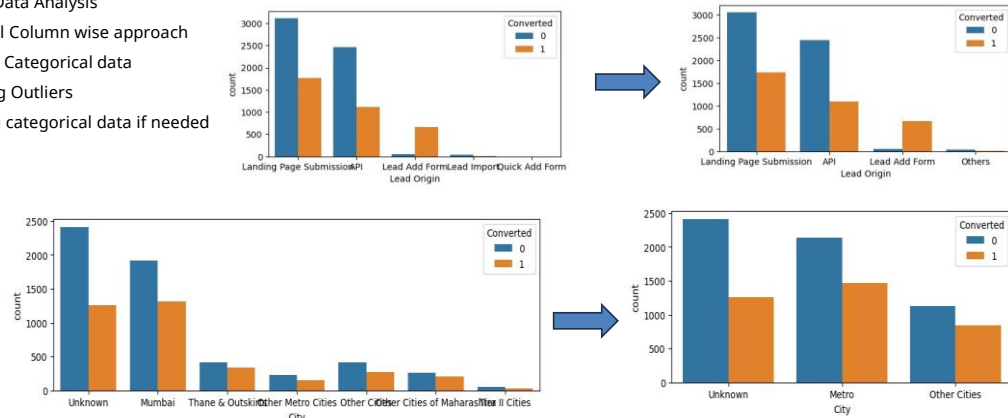
Lead score analysis Project – Summary Report

Based on the Machine Learning Model built on the Leads dataset
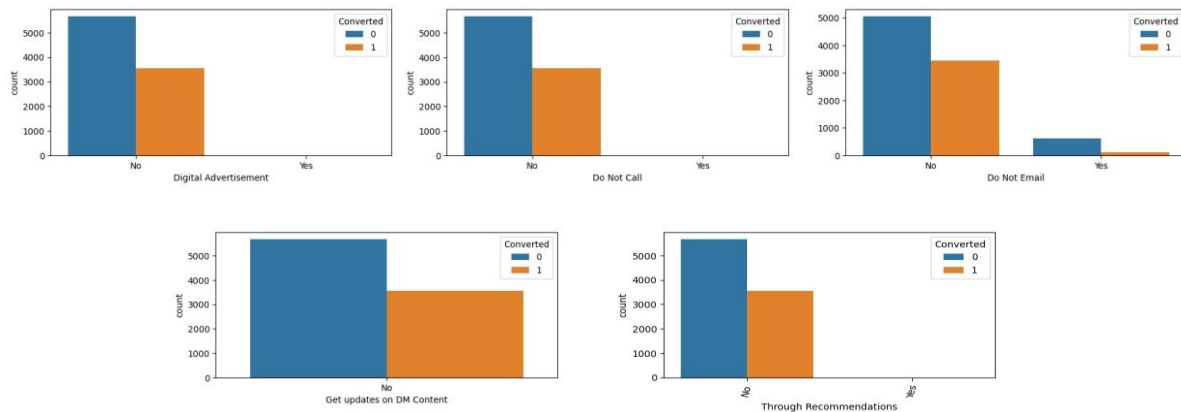
Client Name: X Educations
Sector: Education
Prepared By: Vinay Chawla
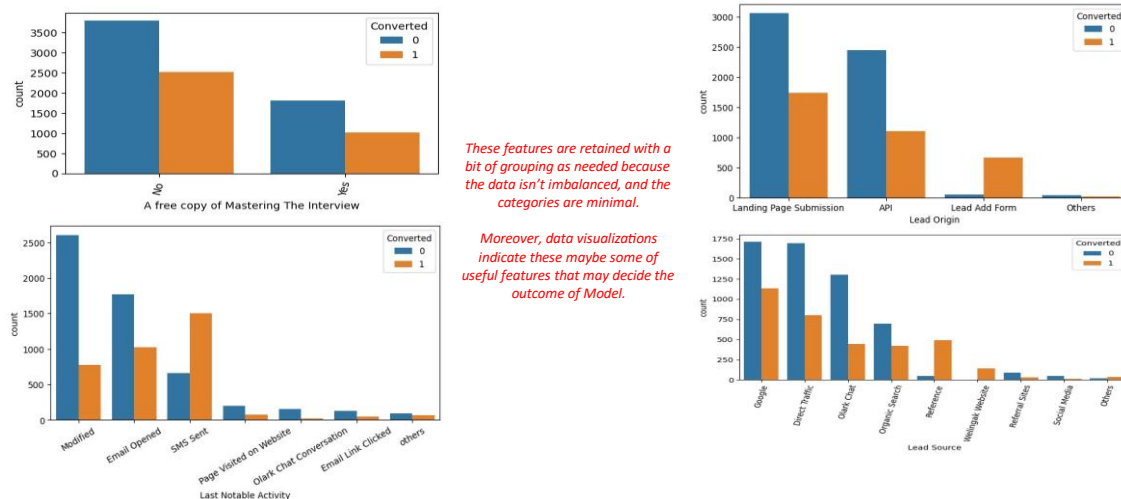*Submitted on: 1st March 2025*

Observations : EDA

Feature dropped based on their data set nature

Such Features are dropped due to their Biased (highly imbalanced data) nature. Such data can cause wrong predictions in model.

Observations : EDA

Features that were retained

These features are retained with a bit of grouping as needed because the data isn't imbalanced, and the categories are minimal.

Moreover, data visualizations indicate these maybe some of useful features that may decide the outcome of Model.

Lead score analysis Project – Summary Report

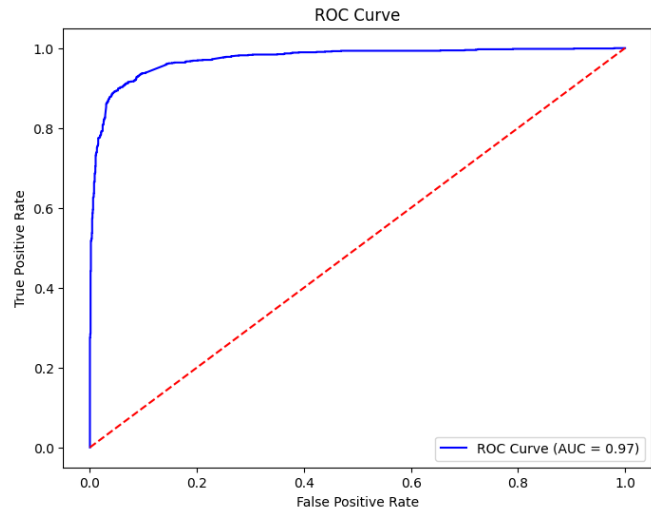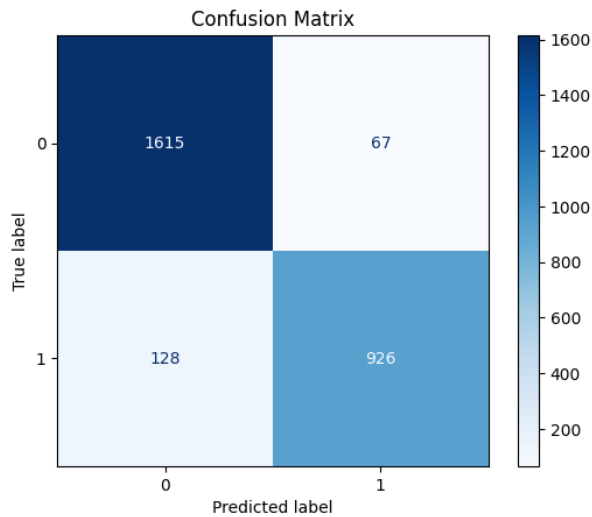Based on the Machine Learning Model built on the Leads dataset

Client Name: X Educations
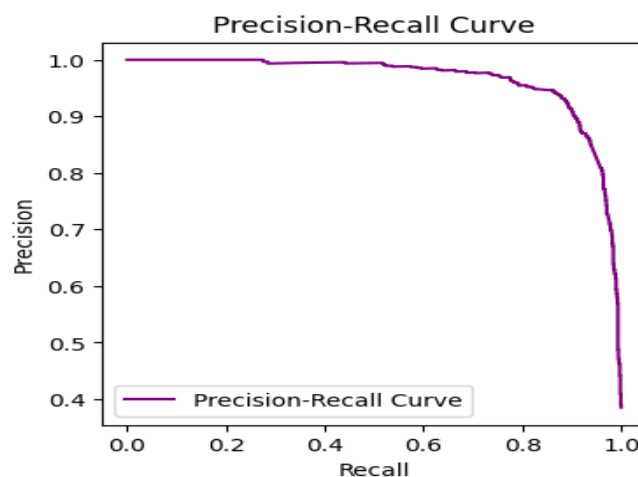*Sector: Education*
Prepared By: Vinay Chawla
Submitted on: 1st March 2025

➢ Model Outcome and Plots

Classification Report:

	precision	recall	f1-score	support
0	0.93	0.96	0.94	1682
1	0.93	0.88	0.90	1054
accuracy			0.93	2736
macro avg	0.93	0.92	0.92	2736
weighted avg	0.93	0.93	0.93	2736

Lead score analysis Project – Summary Report

Based on the Machine Learning Model built on the Leads dataset

Major Takeaways for Client

➢ The model that is built for the customer as per this data achieves a success rate of 87.3 %. In real life, considering unknown situations, we can safely say that this model would perform its job well.

➢ Following this model, the company can utilize their resources on right points and get their lead success to a very good level Presumably above 80% as the model is suggesting.

➢ One more thing that this model allows Company to do is to understand which of the leads are at the borderline case. means which of the leads could be converted to Success rate (but presently the chances of their conversion may not be very bright). These leads are those where the lead score is probably between 50% to 70%.

Above point could be very useful to company to decide where to put extra efforts to increase their sales.

➢ Last but not the least Following plot tells user which parameters are the most important ones that are majorly deciding how a lead could be converted both, in favour or against.

Lead score analysis Project – Summary Report

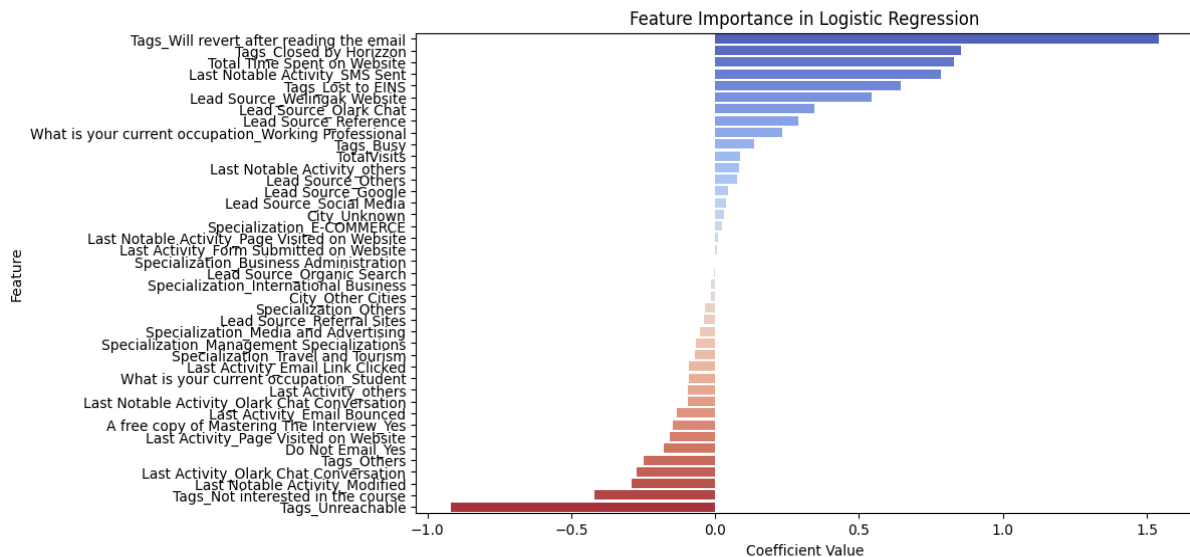Based on the Machine Learning Model built on the Leads dataset

Client Name: X Educations
Sector: Education
Prepared By: Vinay Chawla
Submitted on: 1ˢᵗ March 2025

Feature Importance in Logistic Regression

Major Takeaways from the Project for learner like me

➢ Usually, the data that we would get won't always be clean. Cannot be directly fed to model.

➢ Nulls and Outliers deteriorate ML models badly. Hence, need to be handled carefully; while ensuring we do not lose vital information that could potentially change the model capability.

➢ Many of the features needed optimization like categorization of the data, grouping multiple small categories to one. etc. Usually such information would be from Business or would vary by case to case.

➢ Independent variables could have collinearity among themselves which would be detrimental to the Model outcome. Hence, needs to be removed.

➢ Since we're dealing with many variables at same time, regularization was important.

➢ Usage of right limit to decide the best conversion percentage was important to bring a desired outcome.

Conclusion:

Lead score analysis Project – Summary Report

Based on the Machine Learning Model built on the Leads dataset

Client Name: X Educations
Sector: Education
Prepared By: Vinay Chawla
Submitted on: 1ˢᵗ March 2025

This project gave a practical understanding about how an ML Model could bring wonderful benefits to a company's business if rightly used by reducing their resources cost and maximizing their profitability.