Lead Score Case Study

Submitted by:-

Adarsh Singh

Asif Iqbal

Aryan Upadhyay

INTRODUCTION

In this case study, we will build a logistic regression model for X Education to assign a lead score between 0 and 100 to each lead. The lead score will help the company identify potential leads and prioritize them based on their likelihood of conversion. Our aim is to help X Education achieve their target conversion rate of 80%. Additionally, we will address the other problems presented by the company and provide recommendations on how to utilize the lead scoring model effectively to achieve their business goals. The model should also be able to adjust to any changes in the company's requirements in the future.

BUSINESS UNDERSTANDING

X Education is an online education company that offers courses to industry professionals.

- The Company promotes its courses through various online channels, including search Engines like Google.
- Prospective customers who are interested in the courses visit the X Education website And browse through the available courses.
- Some of these visitors may fill out a form on the website with their email address or Phone number to express interest in the courses. These visitors are classified as leads.
- X Education's sales team contacts the leads via phone or email to try and convert them Into paying customers.

Data under Data Understanding

The dataset consists of two files: 'Leads.csv' and 'Leads Data Dictionary.xlsx'.

* The 'Leads.csv' file contains around 9000 data points. The target variable of the dataset is the column 'Converted', which indicates whether a past lead was converted or not. The values in the 'Converted' column are binary, where 1 means the lead was converted and 0 means it wasn't converted.
* The 'Leads Data Dictionary.xlsx' file provides a data dictionary that explains the meaning of the variables in the 'Leads.csv' file.

Steps of Steps of Analysis

* DATA IMPORTING AND CLEANING
* EXPLORATORY DATA ANALYSIS
* DATA PREPARATION
* MODEL BUILDING AND WVALUTION
* MAKING PREDICTION ON TEST DATASET

Data Clean Data Cleaning

* Missing values in categorical columns were handled based on value counts and certain considerations.

*Drop columns that don't add any insight or value to the study objective.

*Imputation was used for some categorical variables.

*Columns with no use for modeling were dropped.

* Numerical data was imputed with mode after checking distribution.

*Skewed category columns were checked and dropped to avoid bias in logistic regression models.

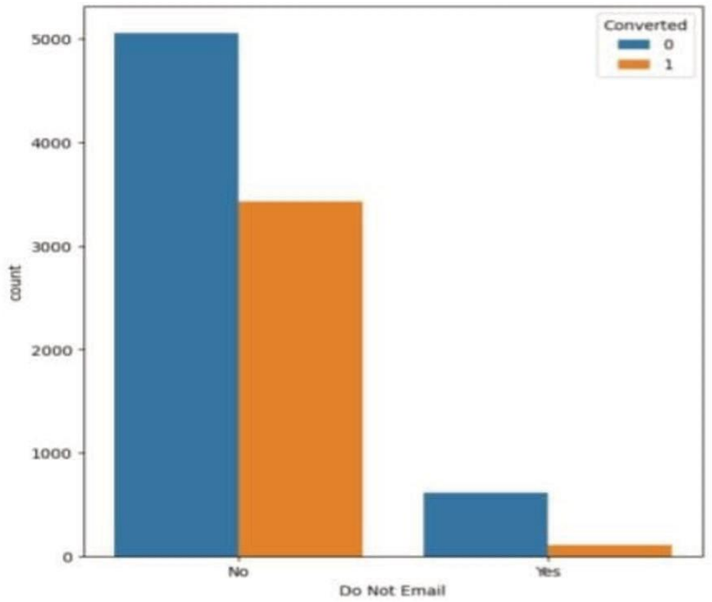
*Outliers in "TotalVisits', "Total Time Spent on Website' and 'Page Views Per Visit' were treated and capped.

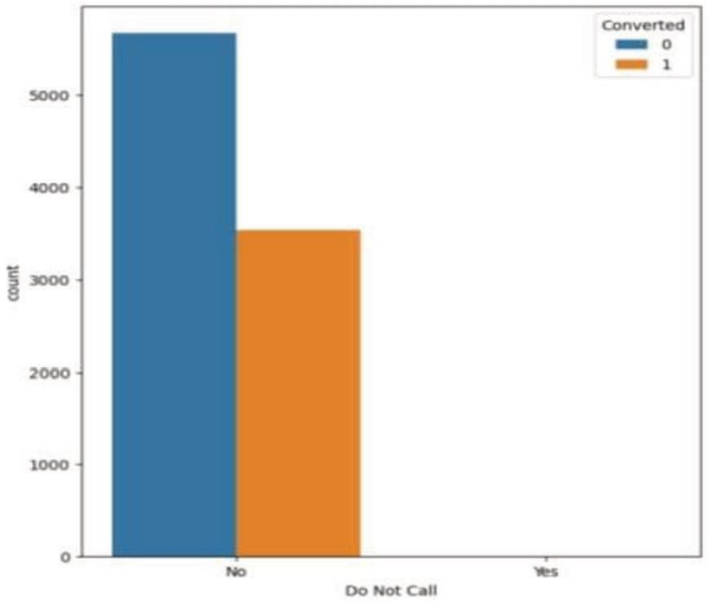
*Low frequency values were grouped together to "Others".

*Standardizing Data in columns by checking casing styles, etc.

EXPLORATORY DATA ANALYSIS

UNIVARIATE ANALYSIS



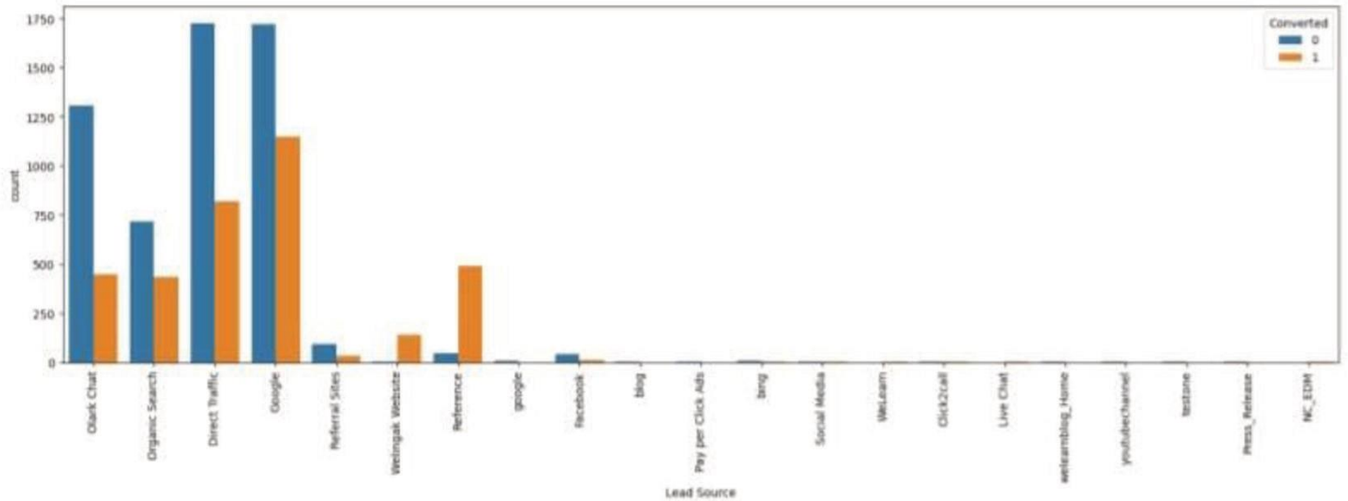


Inference:

• Lead Origin: The majority of customers, 52.9%, were identified through 'Landing Page Submission' as the lead origin, followed by 'API' at 38.7%.

• Current Occupation: A significant proportion of customers, 89.7%, are unemployed based on the current occupation information

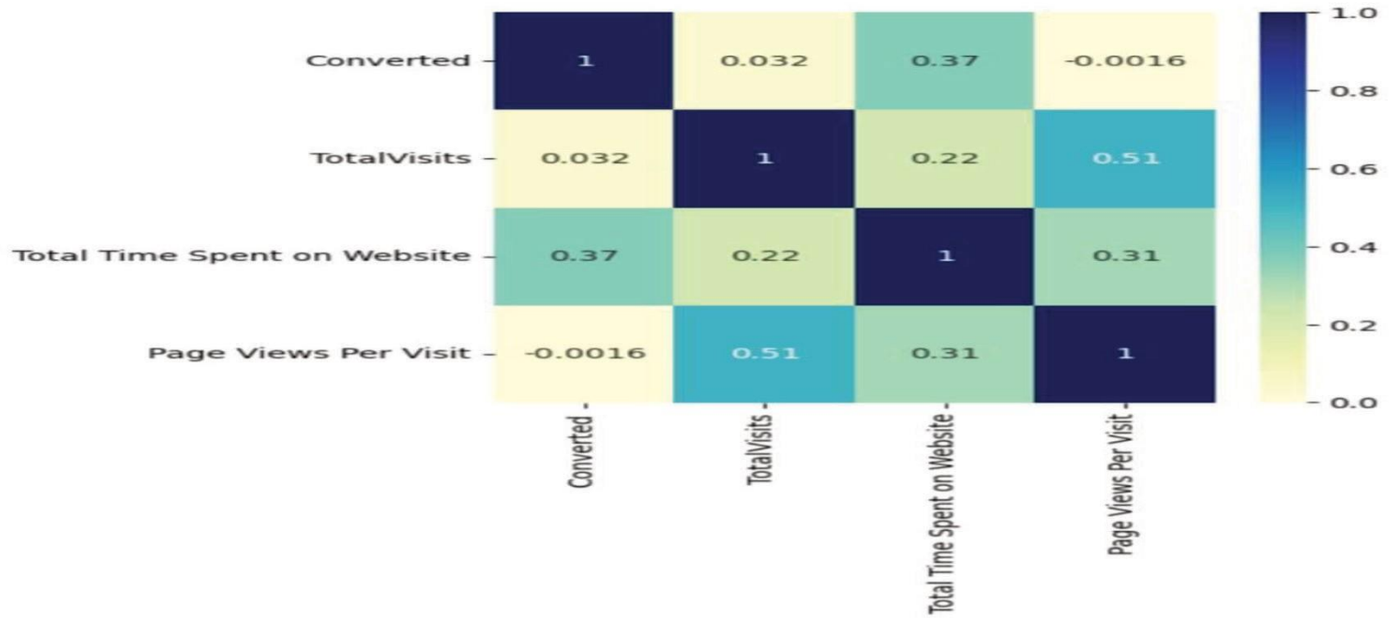
UNIVARIATE ANALYSIS



Inference:

• Specialization: The 'Others' specialization category is the most common among customers at 36.6%, followed by Finance Management at 10.6%, HR Management at 9.2%, marketing Management at 9.1%, and Operations Management at 5.4%.

CORRELATION ANALYSIS



Inference:

- There is a strong positive correlation between 'Total Visits' and 'Page Views per Visit', indicating That customers who visit the website more frequently tend to view more pages per visit.
- Customers who spend more time on the website have a higher LCR, indicating that increasing the Time spent on the website can lead to higher conversion rates.

DATA PREPARATION

* Correlated predictor variables, such as Lead Origin Lead Import and Lead Origin Lead Add Form, were dropped to avoid multicollinearity issues.

* Dummy features were created for categorical variables such as Lead Origin, Lead Source, Last Activity, Specialization, and Current Occupation, using one-hot encoding

* Feature scaling was performed using the standardization method to ensure that all features were on the same scale and no feature dominated the others.

* Binary level categorical columns were mapped to 1/0 in previous steps to make them compatible with the logistic regression model.

MODEL BUILDING

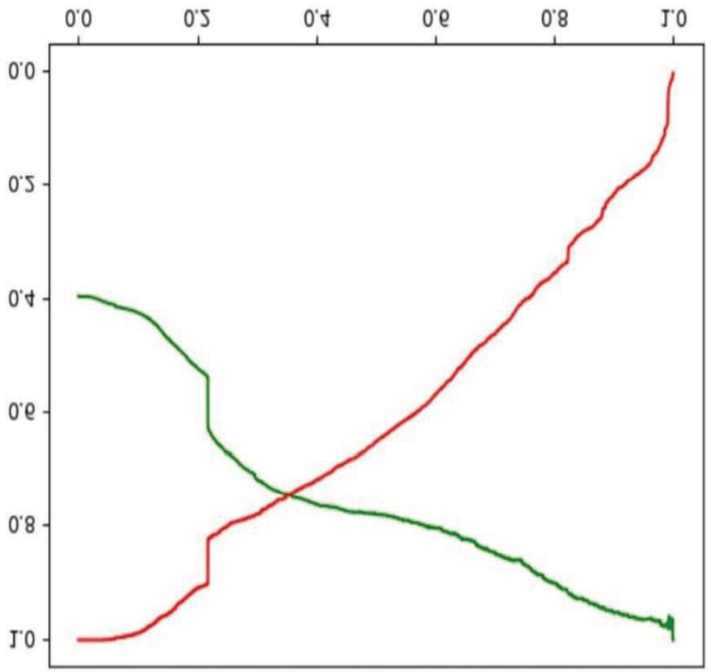
* The data set has a large number of features and dimensions which can reduce model performance and increase computation time.

* Manual feature reduction process was used in Logistic Regression Model-2 and 3 to build models by dropping variables with p-value greater than 0.05.

* Recursive Feature Elimination (RFF) is performed to select only the important columns,

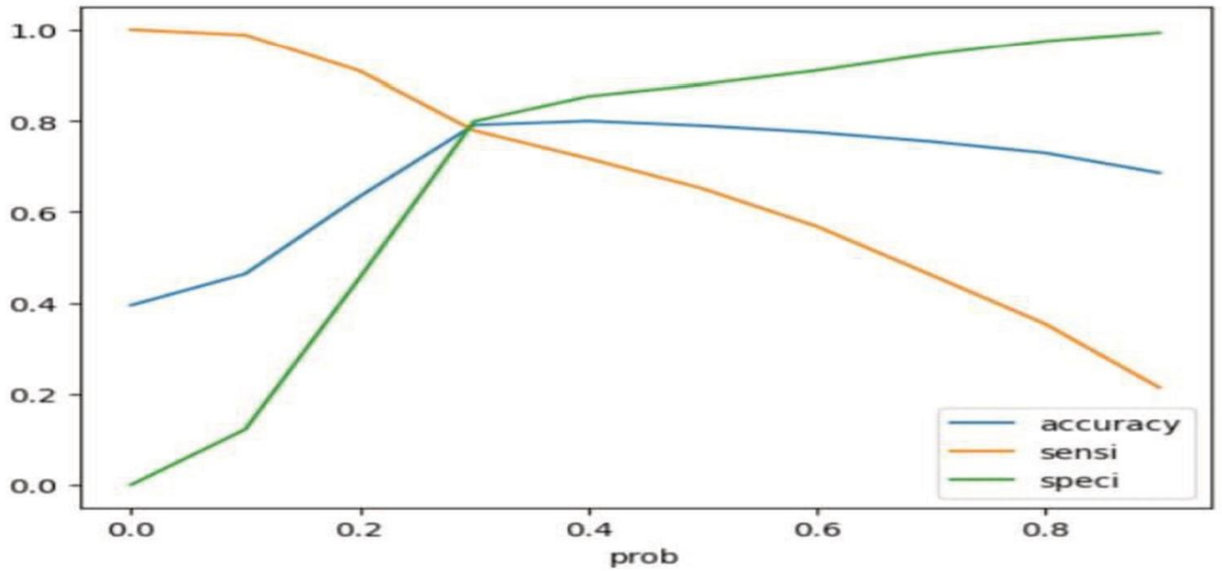
* Logistic Regression Model - 1 is a basic model.

MODEL EVALUTION



Inference:

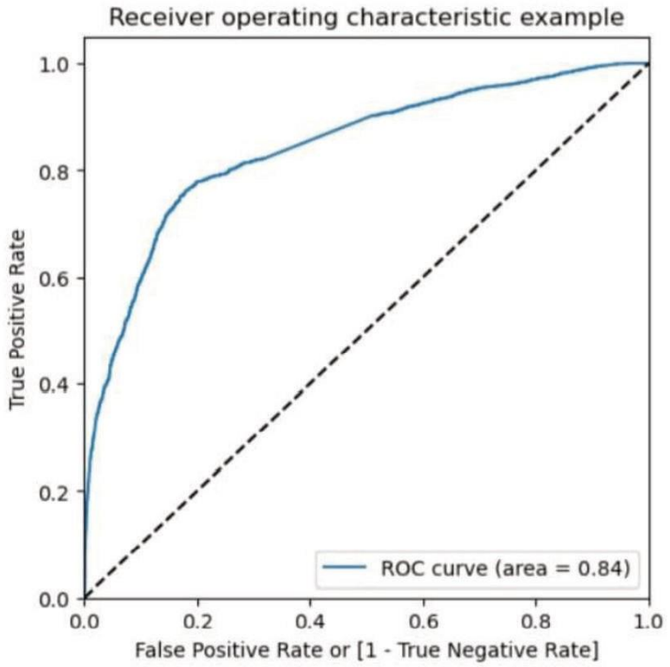
Based on the precision-recall curve, a threshold of 0.4 provides a good balance between precision and recall.



Inference:

Based on the curve analysis, a cutoff probability of 0.35(approx.) is suggested as the optimal point for classification.

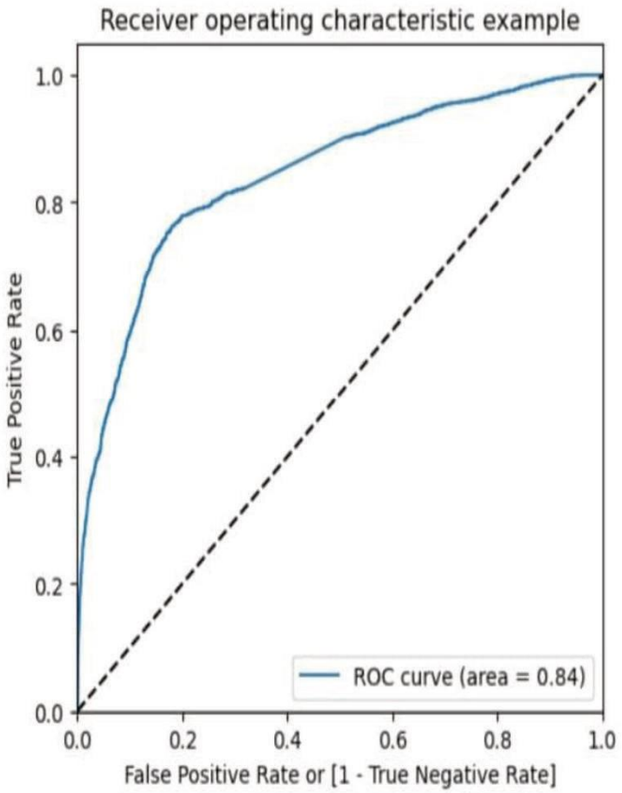
Making Predictions On Test Dataset



ROC Curve - Train Data Set:-

The Area under ROC curve was found to be 0.88 out of 1, indicating that the model is a good predictor.

The curve is plotted as close to the top left corner of the plot as possible, which indicates that the model has a high true positive rate and a low false positive rate at all threshold values.



ROC Curve - Test Data Set

The Area under ROC curve was found to be 0.87 out of 1, indicating that the model is a good predictor.

The curve is plotted as close to the top left corner of the plot as possible, which indicates that the model has a high true positive rate and a low false positive rate at all threshold values.

CONCLUSION

Accuracy, Sensitivity and Specificity values of test set are around 81%, 79% and 82% which are approximately closer to the respective values calculated using trained set.

Also the lead score calculated shows the conversion rate on the final predicted model is around 80% (in train set) and 79% in test set

The top 3 variables that contribute for lead getting converted in the model are

- Total time spent on website
- Lead Add Form from Lead Origin
- Had a Phone Conversation from Last Notable Activity

Hence overall this model seems to be good

Thank You