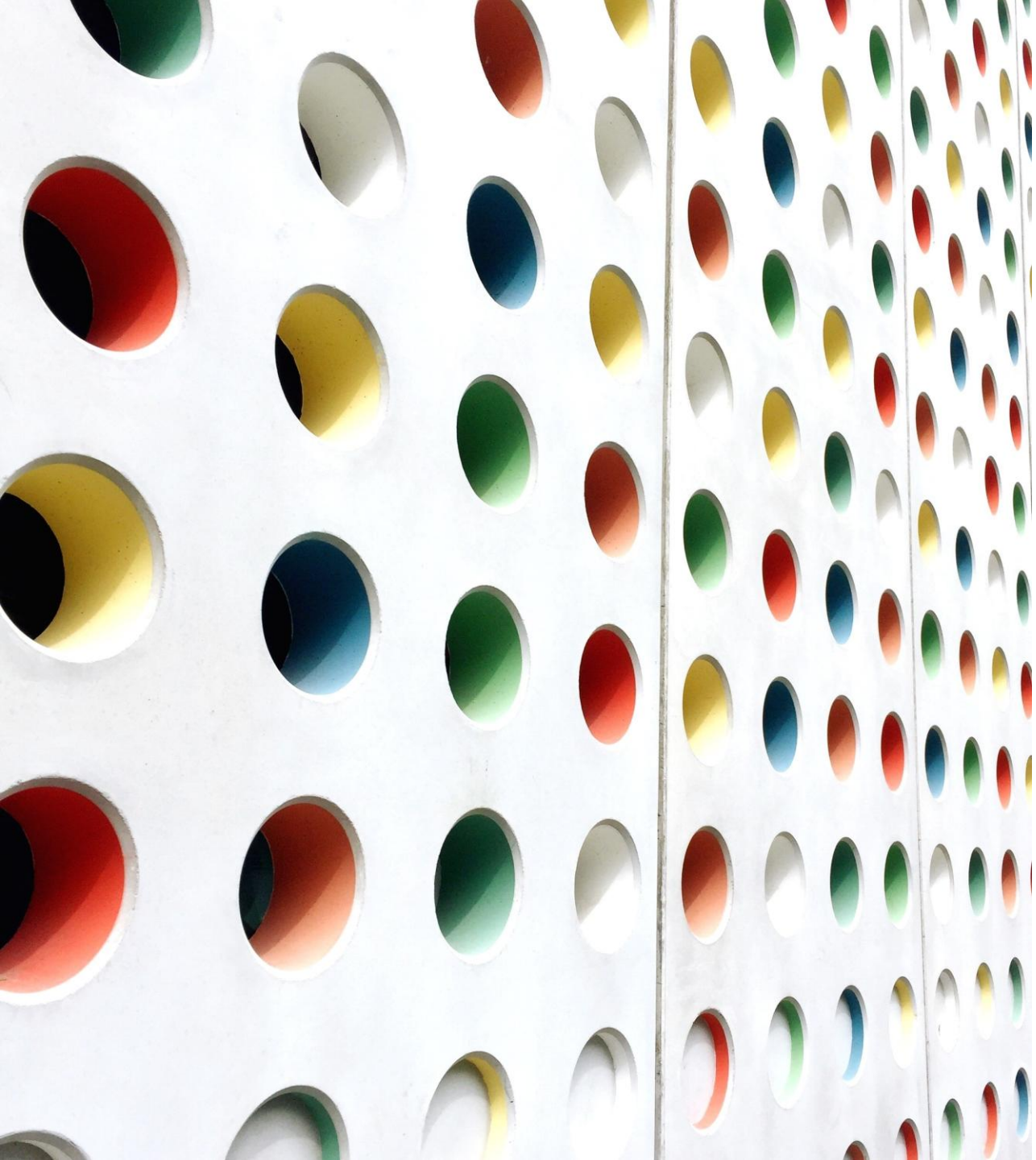


Lead Score Case Study

SUBMITTED BY:

RAHUL BIRHADE

K.R.K.PRASANNA KUMAR

SAI DEEPIKA REDDY

HRITHIK SIVADAS

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. They have process of form filling on their website after which the company that individual as a lead.

Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.

The typical lead conversion rate at X education is around 30%. Now, this means if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as Hot Leads.

If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone

Goals and Objectives

- Lead X wants us to build a model to give every lead a lead score between 0 -100 . So that they can identify the Hot leads and increase their conversion rate as well.
- The CEO want to achieve a lead conversion rate of 80%.
- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches

Problem Approach

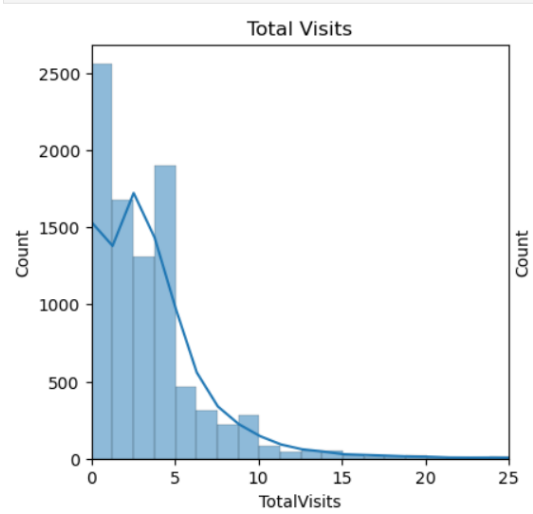
- Importing the data and inspecting the data frame
- EDA
- Dummy variable creation
- Test-Train split
- Model Building
- Creating Predictions
- Model Evaluation
- Optimizing cut off (ROC Curve)
- Prediction of Test set
- Precision – Recall
- Prediction on Test set

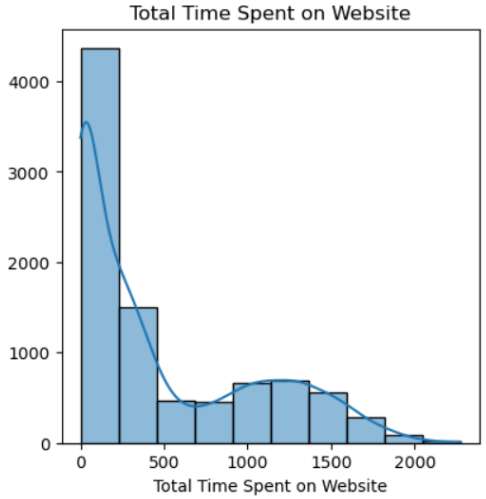
EDA

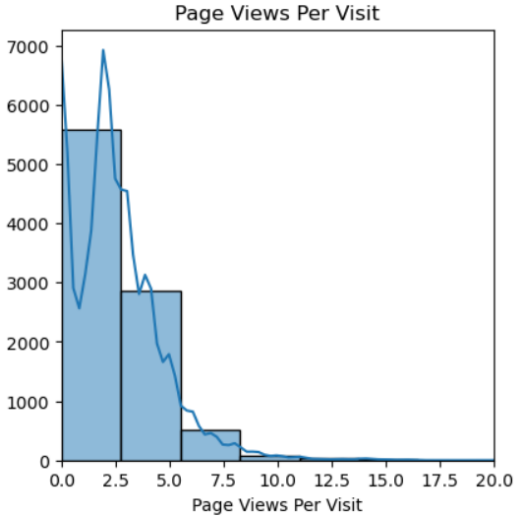
Data Cleaning and Preprocessing

- A quick check was performed to identify columns with a high percentage of missing values. Columns with more than 35% missing data were dropped.
- Upon further review, we noticed that the rows with missing values belonged to important columns, and dropping them would result in significant data loss. Instead of removing them, we replaced the missing values with 'Not Provided'.
- Given that "India" was the most frequent value among the non-missing entries, we imputed all 'Not Provided' values with "India".
- As a result, we observed that nearly 97% of the data in this column was for India, leading us to drop this column due to its lack of variability.
- Additionally, we addressed issues with numerical variables, outliers, and created necessary dummy variables for categorical features

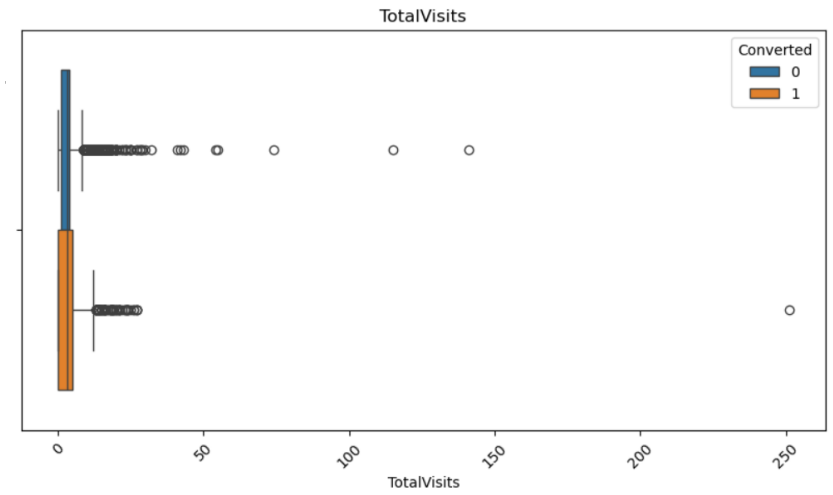
Hist plot on Numerical Columns

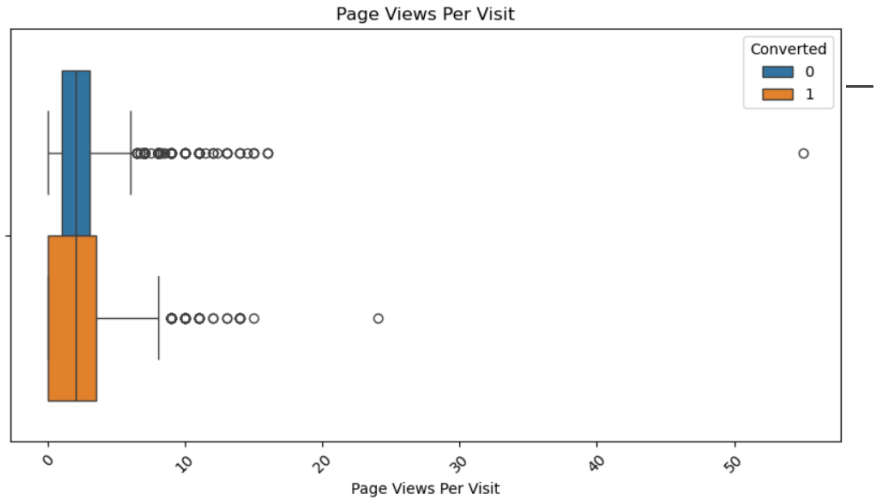


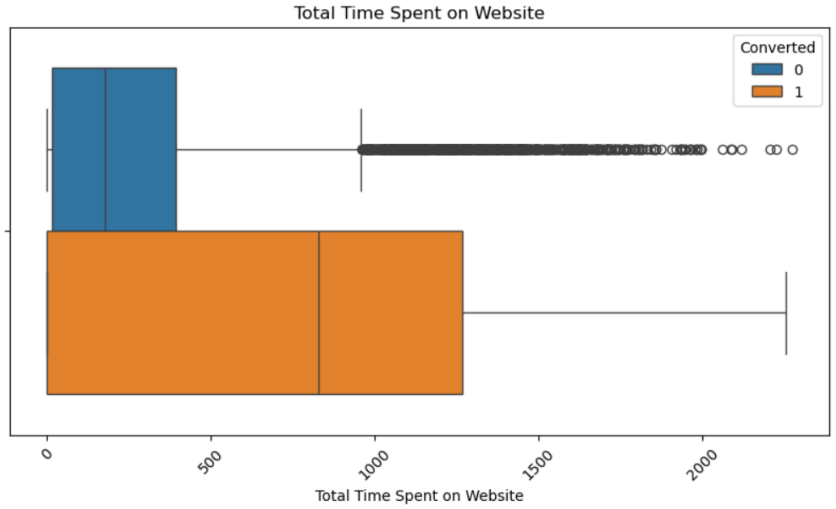




Outlier Identification







Data Split and Scaling

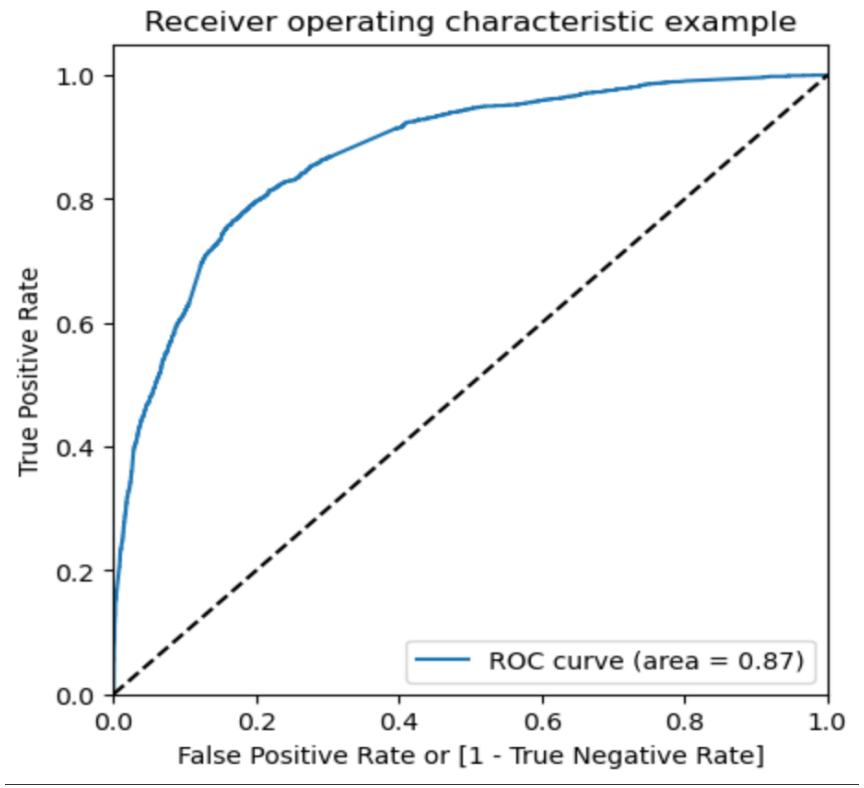
- The data was split into training (70%) and testing (30%) sets.
- Min-max scaling was applied to the following variables.
    1. Total Visits
    2. Page Views Per Visit
    3. Total Time Spent on Website

Model Building

Feature Selection and Model Evaluation

- Recursive Feature Elimination (RFE) was employed for feature selection, which helped identify the most relevant features for the model.
- Initially, RFE was used to reduce the feature set, and the top 15 most relevant variables were selected based on their importance to the model's predictive power.
- To further refine the model, additional variables were removed manually. This was done by analyzing the Variance Inflation Factor (VIF) values to detect multicollinearity and checking the p-values of the variables to assess their statistical significance.
- Once the optimal feature set was finalized, a confusion matrix was created to evaluate the performance of the model. The overall accuracy achieved was 80.77%, indicating a strong model fit.

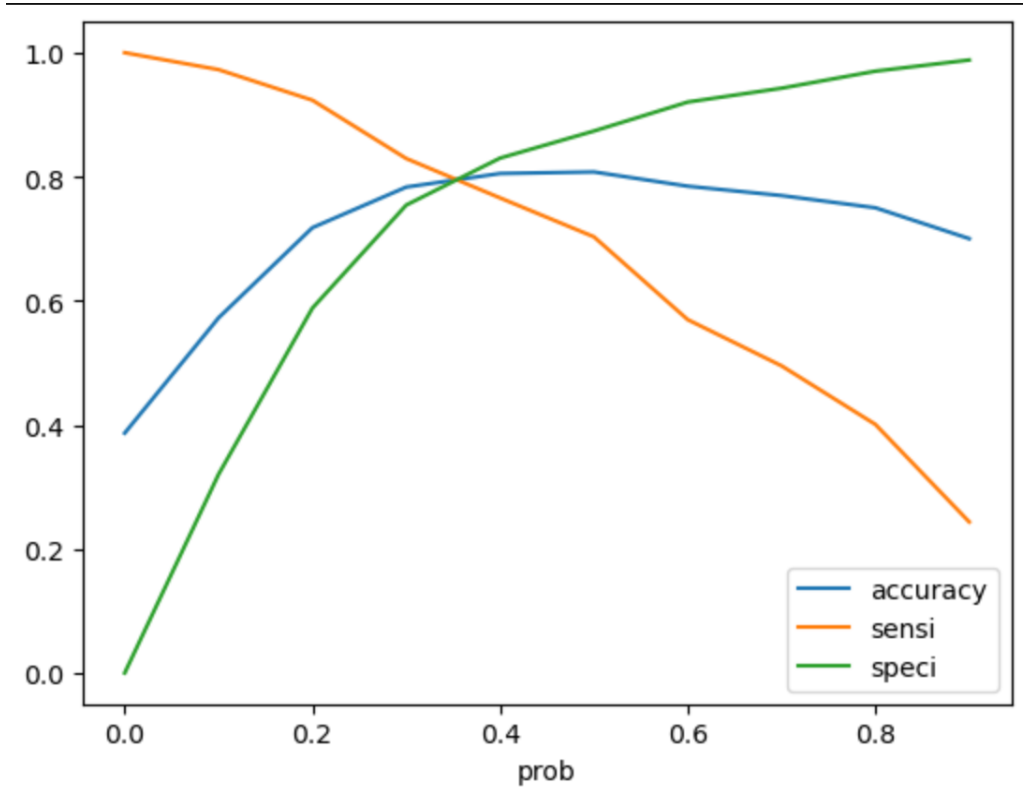
Roc Curve



Model Evaluation - ROC Curve

- The **optimum cutoff value** for classification was determined using the **ROC (Receiver Operating Characteristic) curve**, which helps to find the threshold that best balances sensitivity and specificity.
- The **Area Under the ROC Curve (AUC)** was calculated to evaluate the model's overall performance. The AUC value of **0.87** indicates that the model has good discriminatory power, with a high likelihood of correctly distinguishing between the classes.

Plotting Accuracy, Sensitivity and Specificity



Model Performance Evaluation

After plotting the ROC curve, we determined that the optimum cutoff value was **0.35**, resulting in the following metrics:

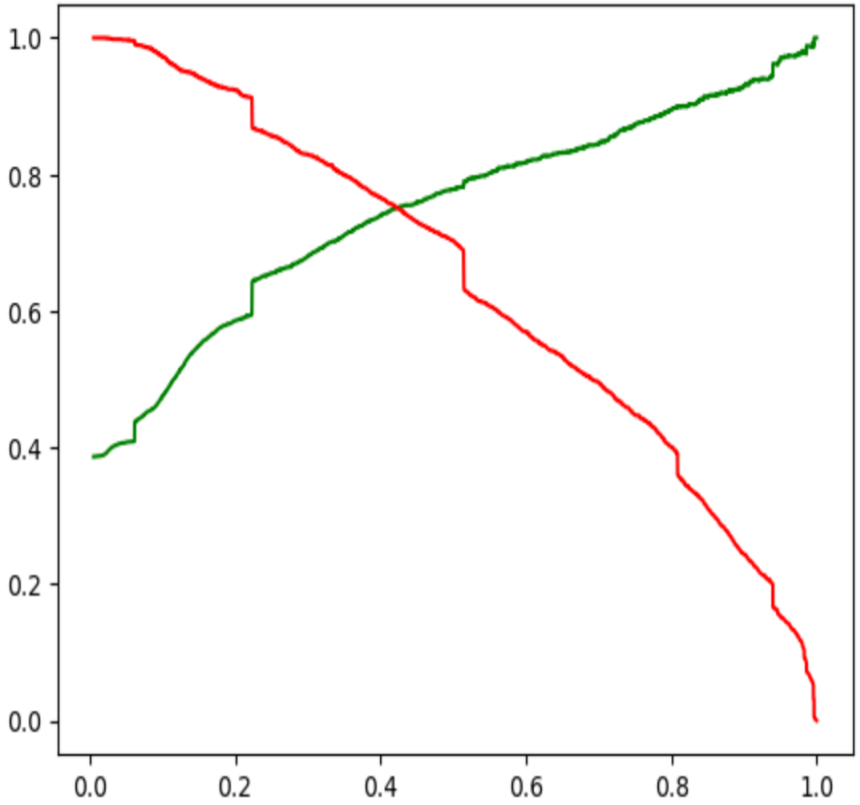
- **Accuracy**: 79.67%
- **Sensitivity**: 79.92%
- **Specificity**: 79.51%

Prediction on Test Data:

When predicting on the test data, the model achieved:

- **Accuracy**: 80.05%
- **Sensitivity**: 80.28%
- **Specificity**: 79.93%

Precision & Recall



Precision-Recall Evaluation

•On Training Data:

- With a cutoff value of **0.35**, the model achieved:
  - **Precision**: 77.82%
  - **Recall**: 70.31%

- To improve these values, the cutoff value was adjusted. After plotting, the optimum cutoff value of **0.44** resulted in the following metrics:
  - **Accuracy**: 80.60%
  - **Precision**: 74.46%
  - **Recall**: 75.85%

•Prediction on Test Data:

- With the optimized cutoff value, the model performed as follows on the test data:
  - **Accuracy**: 80.86%
  - **Precision**: 72.53%
  - **Recall**: 75.28%

Conclusion

- We see that the conversion rate is 30-35% (close to average) for API and Landing page submission. But very low for Lead Add form and Lead import. Therefore, we can intervene that we need to focus more on the leads originated from API and Landing page submission.
- We see max number of leads are generated by google / direct traffic. Max conversion ratio is by reference and Welingak website.
- Leads who spent more time on website, more likely to convert.
- Most common last activity is email opened. highest rate = SMS Sent. Max are unemployed. Max conversion with working professional.

THANK YOU