LEAD SCORING ASSIGNMENT

Using LOGISTIC REGRESSION

Submitted by:

Nupur Bhui

P A V Prasad

Problem statement:

X Education, an online course provider, seeks to identify the most promising leads from website visitors and referrals. They need a model to assign lead scores, predicting the likelihood of conversion into paying customers. The goal is to achieve an 80% conversion rate, prioritizing leads with higher scores to optimize marketing efforts and increase sales.

Goals of the analysis:

- To identify the variables that help identify the leads that are most likely to convert into paying customers.
- To assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- To calculate the confusion matrix and know the overall accuracy, sensitivity, specificity and recall of the model i.e. how well these variables describe the lead conversion probability.

Analysis Approach

1. **Cleaning the data:**
    - The null values are handled, data is standardized, duplicates are removed and irrelevant tags are turned into relevant and understandable tags.

2. EDA:

- Data visualization is used to identify outliers and handled for categorical & numerical data.
- Correlation analysis is performed to understand the relation between the categories.
- Dummy variables are created for the categorical columns and StandardScaler is used to scale the data into a more compact version for analysis.

4. Train-Test split:

The data is split into 70% and 30% for train and test respectively.

5. Model Building:

- Based on the heatmap a few highly correlated and insignificant dummy variables were removed.
- RFE was performed to attain the top 15 relevant variables. Later the variables with VIF < 5 and p-value < 0.05 were kept.

6. Model Evaluation:

- A confusion matrix is plotted on the train data.
- The optimum cut off value (using ROC curve) is used to find the accuracy, sensitivity and specificity which came to be around 80% each.

7. Prediction:

- Prediction is done on the test data frame with an optimum cut off as 0.35 and the outcome of accuracy, sensitivity and specificity came to be around 79%-80%.
- As the accuracy, sensitivity and specificity are nearly same, the cutoff is optimal.

8. Precision-Recall:

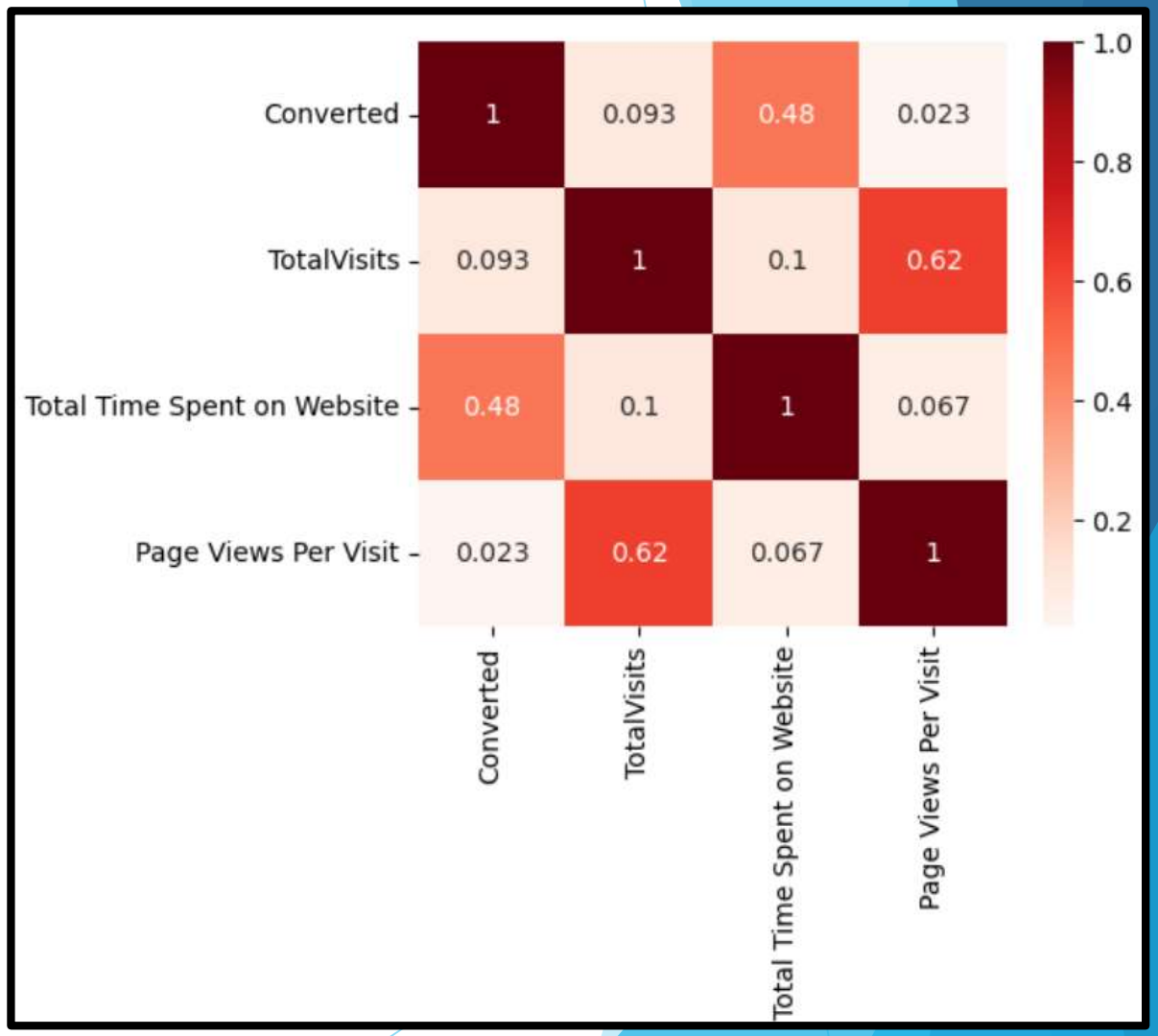
- With a cut off of 0.35, the Precision was found to be around 70%.
- The recall comes to be around 80% on the test data frame. Hence, the model is able to correctly predict 80% of the lead conversions.

Visualizations

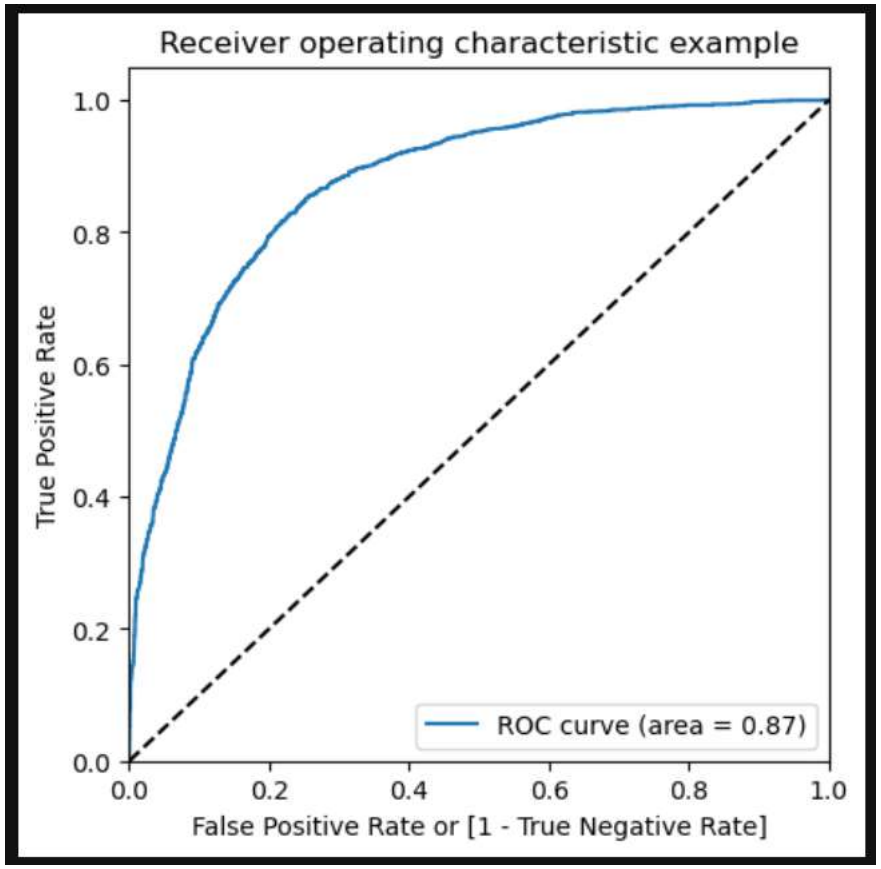
Heatmap of Numerical columns

- As seen in the matrix, the variables “Total Visits” and “Page views per visit” are highly correlated with a value of 0.62.
- The leads that visit the site a lot tend to go through more pages to explore about the various course details, which is an indicator of interest in finalizing the one that are matching their need.

- The variables “Converted” and “Total time spent on the website” are also fairly correlated with a value of 0.48.
- The leads that spend more time on the website directly relate to conversions as they are exploring enough and choosing the course they see fit for their need.

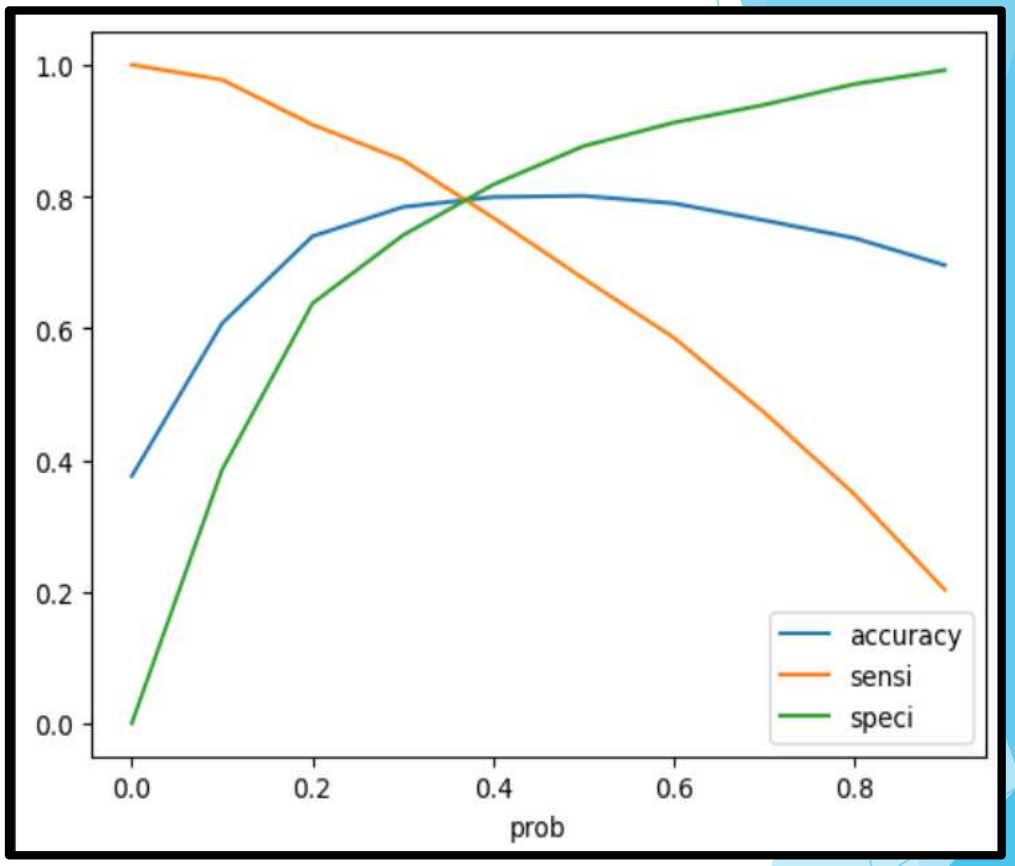


ROC Curve



- The ROC curve shows the tradeoff between sensitivity and specificity. The more the area under the curve, the more accurate the model is.
- We got an area of 0.87 which indicates our model is performing well.

Accuracy/Sensitivity/Specificity curves



- This plot captures the accuracy sensitivity and specificity for various probabilities.
- The intersection of the three metrics is the optimal cutoff for the model

Summary

- The conversion rate of working professionals is high hence the company should focus on personalizing the enquiry experience for working professionals to enhance customer experience and further increased conversions.
- The search history of the leads spending more time on the website can be tracked and targeted with ads based on their most surfed courses. This will make it easier for the customers to look for what they need.
- We got that the last notable activity of leads is one of the significant features that tells us if they are actively browsing the website and can be reached out directly via calls or emails to further know their need and suggest the best course that matches it.
- We also saw that the reason for exploring the website gives us a hint on whether the lead is actually interested in looking into the courses. Hence, popping a poll when they open the website to ask for the reason and their response will give a better idea about the intention of the lead.