Lead Scoring Case Study – DSC54

Submitted by:

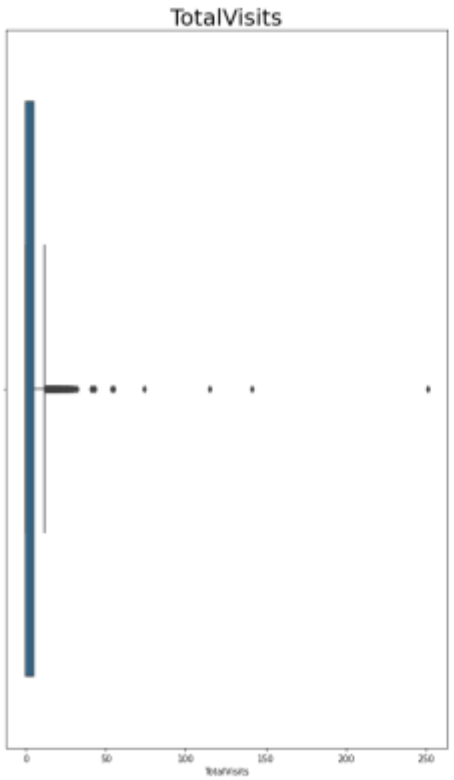
1. Pankaj Sharma
2. Parinita Padge
3. Sritam Krishna Panigrahi

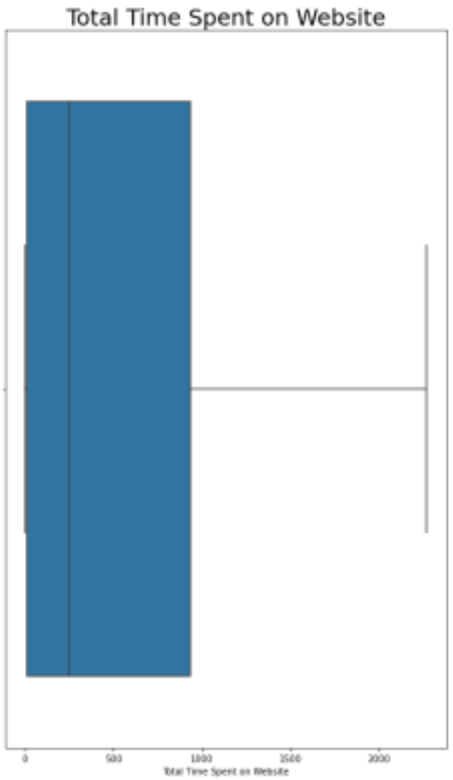
Problem Statement

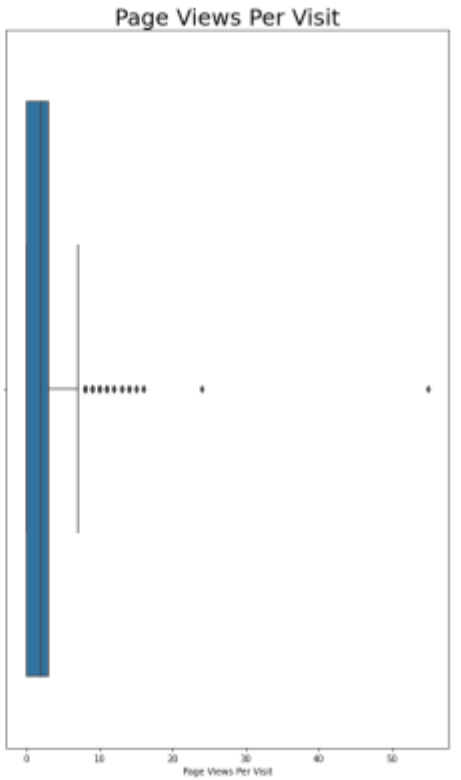
X Education, an online course provider for industry professionals, faces a challenge with its lead conversion process. While they attract numerous potential customers through website visits, form submissions, and referrals, their current lead conversion rate is just 30%. To improve efficiency, the company seeks to identify 'Hot Leads' – those with the highest conversion potential. We have to develop a lead scoring model that assigns scores to leads, enabling the sales team to prioritize leads with higher scores, aiming for a target conversion rate of approximately 80%.

Dispatching Outliers

We transformed binary variables into 0s and 1s and converted multiple categories into dummy variables. While we identified outliers in the dataset, it's important to note that removing them may negatively impact the model-building process.

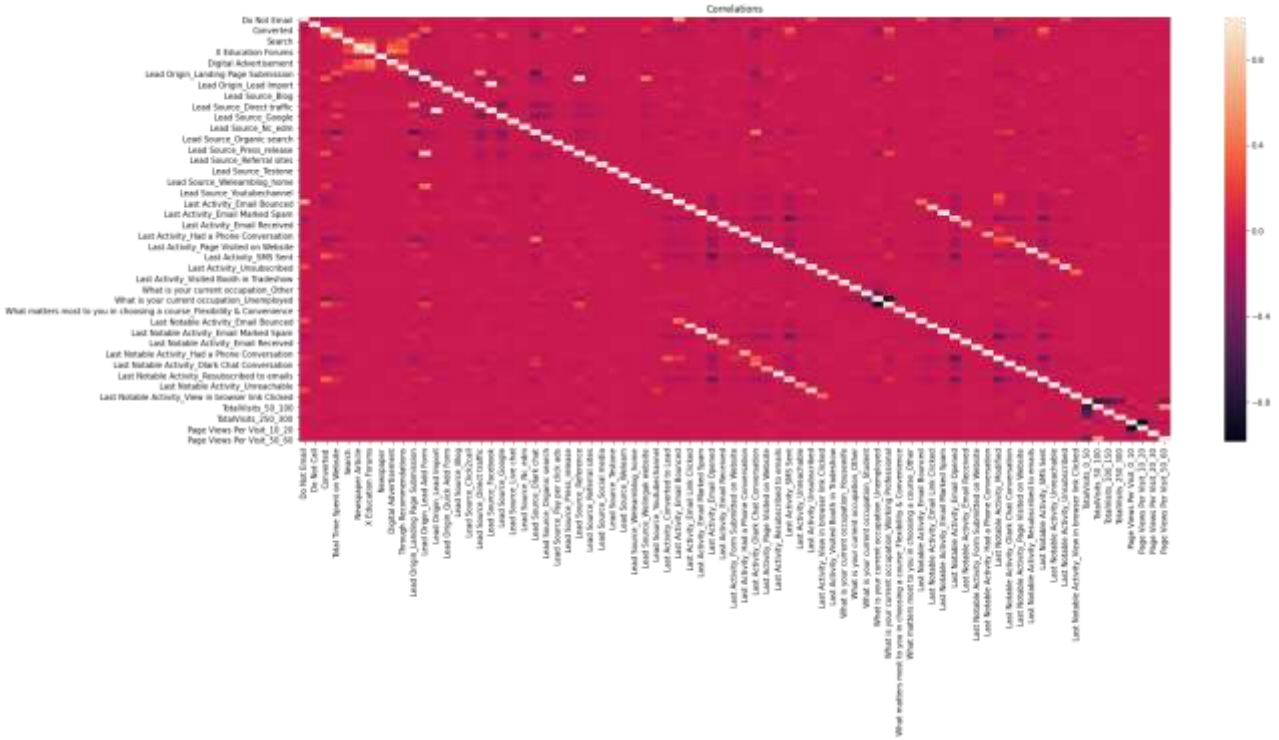






Correlation

Following the resolution of outliers and the creation of dummy variables, we moved on to the data preparation phase. This involved dividing the data into training and testing sets, standardizing it for meaningful analysis, and assessing correlations by generating a heat map.



Model Building

1. Initially, a model was constructed using all available features, revealing numerous insignificant variables.

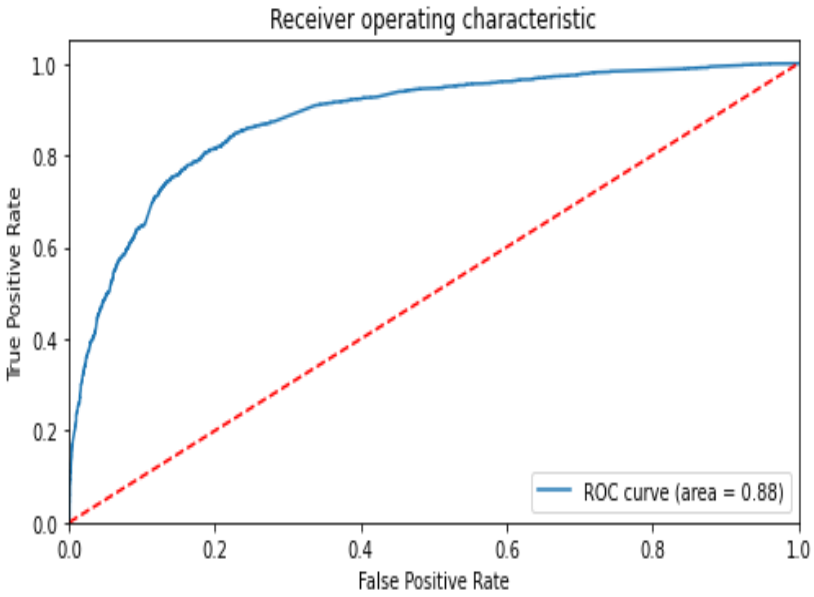
2. To address this issue, the Recursive Feature Elimination (RFE) method was employed, as manual elimination of variables was impractical.

3. Two different RFE counts were considered to ensure the stability of the final model.

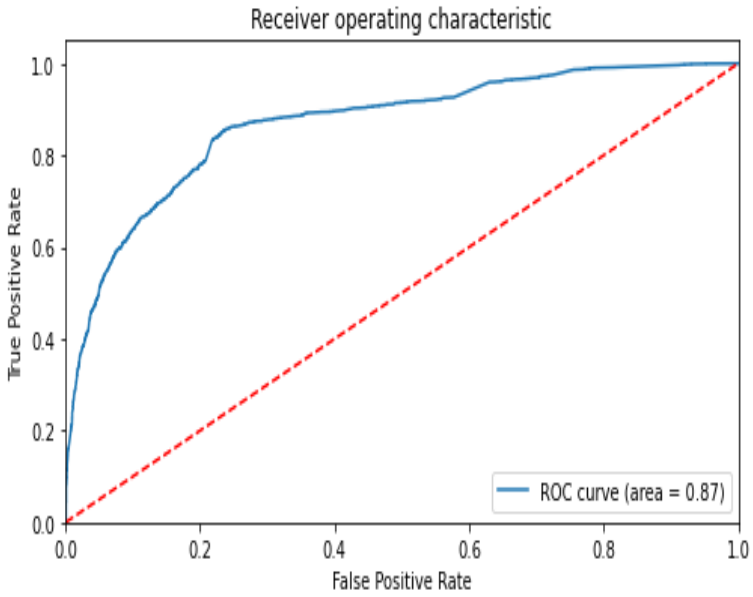
4. The model-building process commenced with an RFE count of 19, and variables were systematically dropped until the model comprised only significant variables with low Variance Inflation Factor (VIF) values.

5. Finally, the model's performance was assessed through predictions and evaluations.

RFE 1 vs RFE 2



RFE 1



RFE 2

- We found that RFE 1 has more AUC score than RFE 2. So , we conclude that RFE 1 is more stable than RFE 2

Conclusion

Analysis Findings:

- 1. Our model exhibits a higher recall score than precision, aligning with our desired outcome.
- 2. This model possesses the flexibility to adapt to the company's future needs.
- 3. The features most influential in achieving a higher conversion rate, or those with a significant impact on the probability of lead conversion, include:

  - Current occupation_Working Professional
  - Last Notable Activity_Had a Phone Conversation
  - Lead Origin_Lead Add Form