Lead Scoring Case Study

Dam Xuan Long - longdamxuan06@gmail.com

Table of Contents

Exploratory data analysis

Train-Test split & Scaling

Model Building

Model Evaluation

Exploratory data analysis

- No instances of duplication are observed in the Prospect ID and Lead Number variables. It is evident that Prospect ID and Lead Number merely serve as identifiers for the contact individuals and can be omitted.
- Significantly, certain columns exhibit a substantial number of null values, as previously observed. However, opting to remove rows with null values would result in a significant loss of valuable data, particularly in essential columns. Instead, a pragmatic approach is adopted wherein NaN values are replaced with 'not provided.' This strategy ensures the retention of all data with minimal null values. Any potential emergence of these values during modeling would render them ineffectual, allowing for their subsequent removal.
- Moreover, the column indicating the country 'India' is overwhelmingly predominant, constituting nearly 97% of the dataset. Consequently, this column is deemed dispensable and can be excluded from further analysis.

Train-Test split & Scaling

- The Variance Inflation Factor (VIF) values exhibit no concerns, indicating acceptable levels. However, certain p-values are notably high, reaching 99%. Consequently, the decision is made to eliminate the variables 'What is your current occupation_Housewife' and 'Last Notable Activity_Had a Phone Conversation' from the model.
- Upon this adjustment, all VIF values remain satisfactory, and all p-values are below the 0.05 threshold. The model is deemed to be in good standing and can be considered finalized.

Model Evaluation

- Achieving an accuracy of approximately 82% is commendable, indicating robust model performance.
- At the current cut-off of 0.5, the accuracy stands at 82%, with a sensitivity of around 70% and a specificity of around 89%. These metrics collectively highlight the effectiveness of the model in making accurate predictions across the dataset.
- Achieving an area under the ROC curve of 0.90 signifies excellent model performance.
- Visual inspection of the graph reveals that the optimal cut-off is situated at 0.35.
- At this threshold, the model yields an accuracy, sensitivity, and specificity of approximately 80%.
- This reinforces the model's effectiveness in accurately predicting outcomes based on the chosen cut-off value.

Precision-Recall

- At the current cut-off of 0.35, the model exhibits a precision of approximately 80% and a recall of around 70%. This emphasizes the model's capability to achieve a high precision rate in correctly identifying positive instances, while also effectively capturing a substantial proportion of actual positive instances, as reflected by the recall value.
- The optimal cut-off is discernible at 0.41, we have Precision around 76% and Recall around 76% and accuracy 82 %.
- At the present cut-off of 0.65, the model achieves a precision of approximately 82%, a recall of around 60%, and an accuracy of 80%. The model demonstrates a proficient prediction of the conversion rate, instilling confidence in the CEO to make informed decisions based on its reliable performance.