Predictive Model for Student Retention

Project Overview

The purpose of this project is to develop a predictive model to calculate the likelihood of student retention for the next fall term at Pitt University. With the use of historical data and key demographic variables known to impact the likelihood of student retention, I trained multiple machine learning models and selected the one with the best performance for making predictions on second tab dataset. The likelihood score generated by this model will be used to identify students who may need additional support and resources.

Methodology

1. **Data Pre-processing:** The initial data cleaning involved handling missing values, detecting outliers, and removing duplicated rows. During the feature engineering phase, statistical hypothesis tests were used for feature selection. I then transformed categorical variables using one-hot encoding and scaled the numerical variable using the StandardScaler from the sklearn library.

2. **Model Training:** Four machine learning models were trained on the preprocessed data: Logistic Regression, Decision Tree, Random Forest, and CatBoost (a gradient boosting model). Each model was tuned using hyperparameters to optimize performance.

3. **Model Testing:** The performance of each model is evaluated using the AUC (Area Under the ROC Curve) score. Choosing the AUC score as a performance metric allows for effective model selection without needing to predetermine the best threshold. This is because the AUC score considers all possible thresholds by measuring the ability of the model to distinguish between classes at various levels. Thus, it provides a comprehensive overview of model performance, making it an apt choice for model selection in our project.

The AUC scores for the four models were as follows:

- Logistic Regression: 0.5593
- Decision Tree: 0.5643
- Random Forest: 0.5762
- CatBoost: 0.5734

4. **Model Selection:** Given that the Random Forest model achieved the highest AUC score, it was selected to predict the likelihood of retention for students who started in Fall 2021.

5. **Cut-Point Selection:** In selecting an optimal cut-point for student intervention, I recommend using a likelihood score of 0.88 as the threshold. This decision has been carefully made considering two critical aspects: the university's capacity to provide resources and the performance of the Random Forest model.

Firstly, the capacity of the university to provide support is a significant consideration. The ideal cut-point will classify a manageable number of students as 'not retained' that aligns with the resources available to advisors. A threshold that identifies too many students for intervention might strain resources and prove unrealistic. Therefore, we assume the university can adequately support a maximum of 200 students. With a 0.88 cut-point, our model classifies 184 students as 'not retained,' a number within the assumed capacity.

Secondly, the model's performance is a vital factor in choosing a cut-point. We used three metrics—accuracy, negative predictive value, and total number of negative predictions—to comprehensively evaluate the model's performance across various likelihood scores.

Accuracy measures the overall correctness of the model's predictions. Negative Predictive Value (NPV) indicates the probability that a student predicted as 'not retained' is indeed not retained. And finally, the total number of negative predictions, which equals the sum of True Negatives and False Negatives, represents the total number of students predicted as 'not retained.'

The optimal cut-point should ensure a high degree of accuracy and NPV, while the total number of negative predictions should not exceed the university's assumed maximum capacity (200 students). A threshold of 0.88 satisfies these conditions, making it our recommended cut-point.

Conclusion

In conclusion, the Random Forest model was our preferred predictive tool for assessing student retention into the next academic year, employing a cut-off score of 0.88. At this threshold, the model demonstrated a commendable accuracy of 0.86 and a positive predictive value of 0.91, indicating strong predictive capabilities for retained students.

However, the model exhibited a limitation in its negative predictive value, which stood at 0.11. This implies that our confidence in accurately identifying students who are not retained is only 11%, pointing towards potential room for improvement. The current data features did not provide enough distinction to reliably predict students unlikely to be retained.

To address this issue, a more in-depth understanding of the reasons for student non-retention could prove beneficial. Conducting surveys with non-retained students could reveal additional factors influencing their decision not to return, thereby providing more informative variables. Incorporating these new insights into the model could enhance its performance, allowing for a more precise identification of students who might require extra resources and support.

Note:

1. The updated likelihood score can be found in the second tab of “Data Prediction Model - Likelihood Update.xlsx”

2. Please check “Data Prediction Model - Coding Works.pdf” for more detailed working process