PREDICITNG STUDENTS' ACADEMIC SUCCESS & DROPOUT USING SUPERVISED MACHINE LEARNING

Divvyam Arora

Student, Introduction to Machine Learning, Stanford Pre-Collegiate Summer Institutes

Introduction & Purpose

Reducing academic dropout and failure in higher education is a crucial goal. The purpose of this research project was to utilize machine learning techniques to identify at-risk students early in their academic journey, enabling timely intervention and support. By utilizing a comprehensive dataset, I developed a classification model using python that analyzes academic and demographic features to effectively identify students in need of assistance. The outcomes of this research demonstrate the potential of machine learning to contribute to proactive measures for student success.

Dataset

A dataset was retrieved from the UCI Machine Learning Repository. The dataset used was created from a higher education institution (acquired from several disjoint databases). It is related to students enrolled in different undergraduate degrees. The dataset includes information known at the time of student enrollment and the students' academic performance at the end of the first and second semesters.

Dataset Characteristics

Tabular | 36 Attributes | 4424 Records

UC Irvine Machine Learning Repository

Attribute Categories

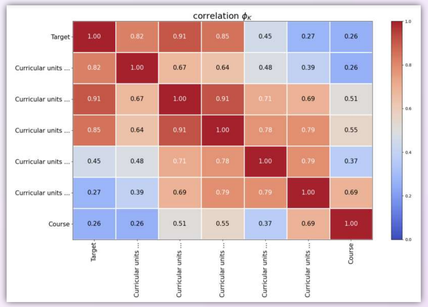
- Academic Data & Field
- Demographic Data
- Socio-Economic Factors
- Personal Details

Outcomes

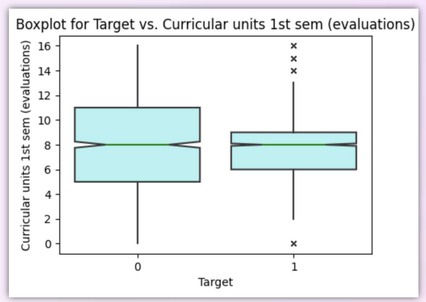
- Graduate [1]
- Dropout [0]

Methodology

- Refined dataset to have a binary outcome.
- Sorted out important attributes through (Φk) correlation coefficient matrice (7 attributes were used in total).



- Removed data outliers and conducted bivariate analysis.

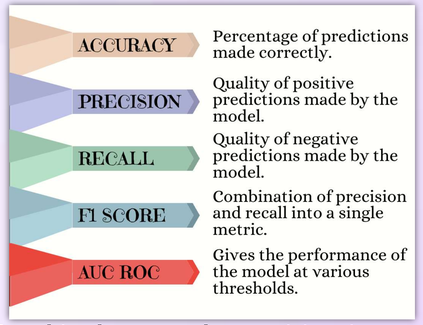


- Split data into train and test sets [70/30], and trained about 18 different models.

Neural Network	Logistic Regression	K Nearest Neighbour
AdaBoost	Gradient Boost	XGBoost
Tuned Random Forest	Support Vector Machine	Tuned Bagging Classifier
Decision Tree (GridSearch)	Stacking Classifier	Combo Stacking Classifier
Tuned XGBoost	Bagging Classifier	Tuned Bagging Classifier
Decision Tree	Tuned Decision Tree	Random Forest

- Selected the final model based on evaluation results.

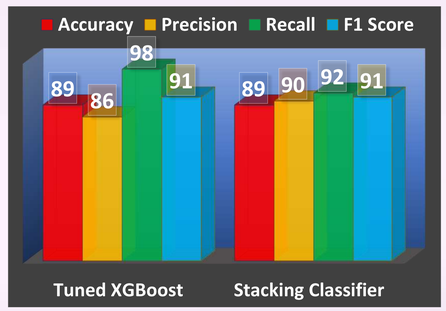
Evaluation Metrics



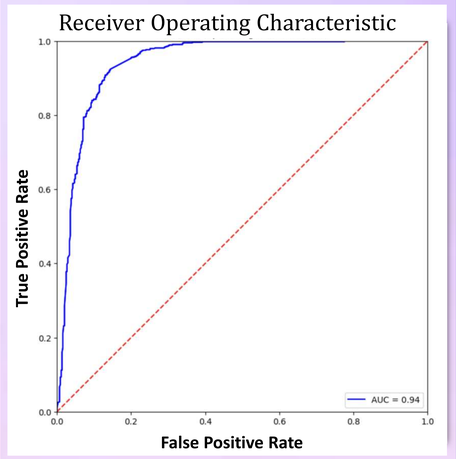
For this data sample, precision is more valuable than recall as predicting that a student will not dropout if in reality they will is worse than predicting that a student will dropout even if they do not. Therefore, I tried to maximize precision. Accuracy and F1 Score still highly valued. AUC ROC indicates my model's overall performance.

Results

Initial results obtained by basic models were satisfactory. To further increase the performance, boosting algorithms and advanced models were used. Ultimately two models offered the best results. One of them was Tuned XGBoost and the other one was Stacking Classifier (preferred due to higher precision) which was trained with the best performing models to improve its performance. Neural Networks gave the highest recall score, but their other metrics were low similar to other advanced models.

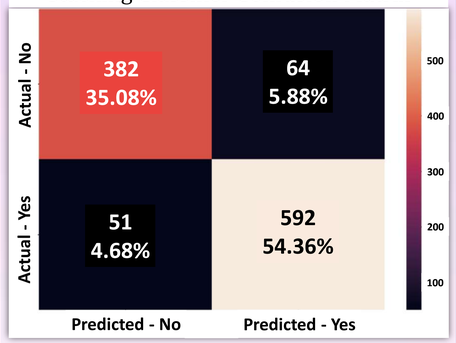


The ROC Graph was very similar for both the metrics and the AUC was the same.

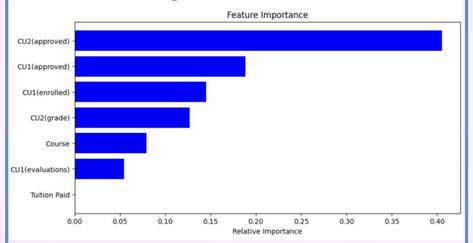


AUC = 0.94

Stacking Classifier Confusion Matrix



Attribute Importance in Classification



Summary & Conclusion

- The project utilizes machine learning techniques to identify at-risk students early in their academic journey, enabling timely intervention and support.
- The dataset was modified to be have a binary output and just 7 of the most important attributes. About 18 models were then trained with these and compared based on precision, accuracy, recall, F1 score and AUC ROC.
- Tuned XGBoost and Stacking Classifier give the best results, but Stacking Classifier is preferred due to its high precision value (XGBoost has higher recall).
- The accuracy, F1 score and AUC ROC was same for both the models.
- It is concluded that Stacking Classifier can be used to reliably make predictions related to student academic success & dropout, however, more data is required to continuously train the model to improve its accuracy and performance.

Acknowledgements

I extend my heartfelt thanks to Dr. Jamie Fairclough for her exceptional teaching, my parents for unwavering support, Stanford Pre-College authorities for the opportunity, dataset creators for their valuable contribution, Mr. Jason Shin for his guidance and peers for enriching my learning experience. Grateful to all who played a role in my academic journey.

References