Predicting Student Success and Foiling Dropout in Polytechnic Institute of Portalegre (IPP), Portugal

Lakshmi Chirumamilla
Pace University, Seidenberg School Of Computer Science and Information Systems
Github: https://github.com/Sudhachirumamilla1029

ABSTRACT

The increasing concern regarding student dropout rates in educational institutions prompts the exploration of predictive analytics to identify students at risk and facilitate timely interventions. This project presents a novel predictive framework using machine learning (ML) techniques to distinguish students likely to succeed academically from those at risk of dropping out. Leveraging a comprehensive dataset obtained from the Polytechnic Institute of Portalegre (IPP), Portugal, encompassing student demographics, academic performance, behavioral patterns, and socioeconomic factors. The results show that the Random forest performs better than XGBoost, Logistic regression , SVM, Decision Tree, Naive Bayes. Used SHAP analysis for finding best parameters and Fine-tune the models accordingly.

RESEARCH QUESTION

What strategies and techniques can be effectively employed to achieve precise predictions regarding students at risk of dropout and those on track for academic success?

RELATED WORK

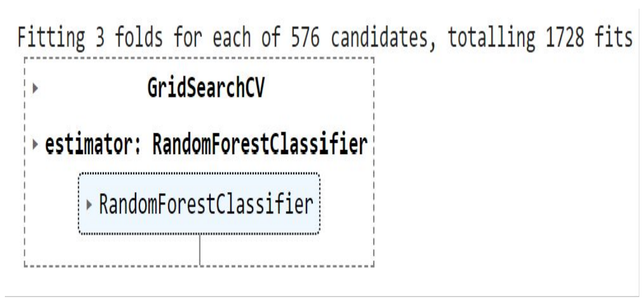
In [1] and [2], extensively explores predictive analytics using machine learning models. These studies emphasize the persistent challenge of class imbalances in student datasets, addressed through innovative techniques like SMOTE and its variants to rectify skewed class distributions and enhance predictive accuracy [4]. Evaluation metrics such as precision, recall, F1-score, and AUC-ROC play pivotal roles in gauging model performance, revealing insights into student persistence and dropout rates. This collectively emphasize the evolution of predictive analytics, stressing the need to address class imbalances and emphasizing the critical roles of algorithm selection and nuanced evaluation metrics in predicting student success and dropout likelihood in higher education. There is a shortage of Hyperparameter tuning of the ML models in context of this work.

DATASET

A dataset created from a higher education institution related to students enrolled in different undergraduate degrees, such as agronomy, design, education, nursing, journalism, management, social service, and technologies. The dataset includes information known at the time of student enrollment (academic path, demographics, and social-economic factors) and the students' academic performance at the end of the first and second semesters [3].

METHODOLOGY

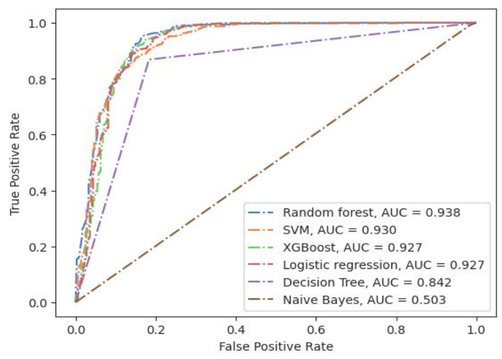
- Addresses class imbalances. It experiments with various balancing methods to assess their impact on model performance.
- Multiple machine learning algorithms—Random Forest, Gradient Boosting, Logistic Regression, and Neural Networks—are implemented and fine-tuned using techniques like grid search.
- The dataset is split into training, validation, and test sets for model training and evaluation. Evaluation involves diverse metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.
- Additionally, SHAP analysis interprets model predictions, while confusion matrices identify prediction errors for improvement insights.
- Hyper-parameter tuning




```
{'max_depth': 10,
 'max_features': 'sqrt',
 'max_samples': 0.7,
 'min_samples_leaf': 1,
 'min_samples_split': 0.001,
 'n_estimators': 75}
```

EVALUATION

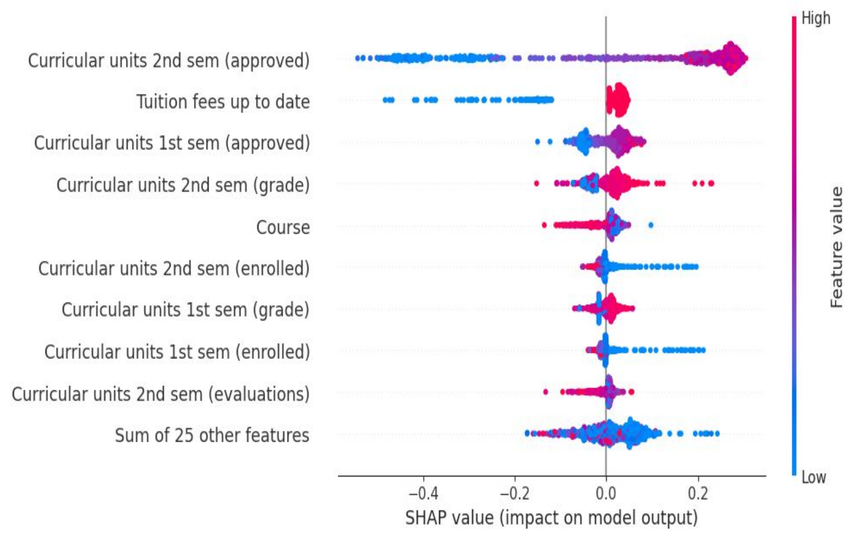
Algorithm	Accuracy	Precision	Recall	F1 Score
Random forest	0.904185	0.891003	0.955473	0.922113
XGBoost	0.888767	0.876289	0.946197	0.909902
Logistic regression	0.886564	0.873288	0.946197	0.908281
SVM	0.867841	0.844007	0.953618	0.895470
Decision Tree	0.845815	0.874296	0.864564	0.869403
Naive Bayes	0.594714	0.595318	0.990724	0.743733



The AUC curves reveal that Random Forest achieving the highest AUC of 0.938 and accuracy of 0.904 outperforming remaining algorithms.

RESULTS

The performance of six machine learning algorithms—Logistic Regression, SVM, Decision Tree, Naive Bayes, Random Forest, and XGBoost-was assessed for a classification task. Initial evaluation without hyperparameter tuning revealed varied performance metrics. Notably, Random Forest outperformed others with an accuracy of 90.42%, precision of 89.10%, recall of 95.55%, and F1 score of 92.21%. Subsequently, hyperparameter tuning using GridSearchCV for the Random Forest model. This refinement resulted in a marginal improvement, maintaining the accuracy at 90.42% while enhancing precision, recall, and F1 scores to 89.10%, 95.55%, and 92.21%, respectively. These findings highlight the initial strength of Random Forest and the limited impact of hyperparameter tuning on its performance in this specific classification task. Using SHAPley, the Circular units of 1st and 2nd semester have the most impact on the ML model.



CONCLUSION & FUTURE WORK

The study investigated the use of machine learning (ML) tools to predict the academic performance of first-year computer science (CS) students and identify key factors affecting their success. It found that Random Forest classifiers were effective in this prediction task. While previous work centered on GPA-based performance prediction, focusing on academic probation status allows for aiding struggling students beyond just those with low grades.. Future steps proposed involve leveraging Natural Language Processing (NLP) to analyze text materials from college applications.

REFERENCES

1. M.V.Martins, et al., Trends and Applications in Information Systems and Technologies, vol.1, in Advances in Intelligent Systems and Computing series. Springer.
2. Tang, Z, et al., Exploring Individual Feature Importance in Student Persistence Prediction. Journal of Higher Education Theory and Practice, 23(6). doi:10.33423/jhetp.v23i6.5957
3. B. M. Neda, M. Wang, A. Singh, S. Gago-Masague and J. Wong-Ma, "Staying Ahead of the Curve: Early Prediction of Academic Probation among First-Year CS Students," 2023 3rd International Conference on Applied Artificial Intelligence (ICAPAI), Halden, Norway, 2023.
4. Llauró A, Fonseca, et al., Identification and comparison of the main variables affecting early university dropout rates according to knowledge area and institution.