Positive Unlabeled Learning

A Three-way Benchmark Against CatBoost

Mohsen Karshenas
Label Spreading, CatBoost
CPSC 6300
Applied Data Science
Fall 2023

Jacob Martin
Random Forest Classifier, Logistic Regression Model, Writing
CPSC 4300
Applied Data Science
Fall 2023

Rithika Kuchanpally
Self-Training, Logistic Regression Model
CPSC 6300
Applied Data Science
Fall 2023

ABSTRACT

This research addresses the critical challenge of predicting and identifying students at risk of academic dropout and failure in the early stages of their academic journey. Early detection enables proactive intervention, optimizing resource allocation and fostering the well-being of at-risk students. The project contributes actionable insights to inform decision-making for educators, administrators, and policymakers.

1. INTRODUCTION

Leveraging a comprehensive dataset on student dropout predictors, we employ various positive-unlabeled classification methods to discern patterns and relationships within the data. Our overarching question is: How can we predict and identify students facing academic challenges, and what effective strategies can be implemented to provide timely support and intervention for these individuals? By formulating our project around this question, we aim to develop insights that not only enhance our understanding of dropout predictors but also contribute to the development of targeted interventions for at-risk students. The question of predicting and identifying students at risk of academic dropout and failure holds paramount significance in the realm of education, and addressing this question is crucial. Early detection allows educational institutions to intervene proactively and provide timely support to students who may be facing academic challenges. This, in turn, increases the likelihood of preventing dropout and fostering academic success. By understanding the predictors of academic difficulties, we can develop strategies that not only address immediate academic concerns but also contribute to the overall well-being of students. Understanding the factors influencing dropout rates equips stakeholders with the knowledge needed to implement targeted policies and initiatives. All in all, our project seeks to provide actionable insights that go beyond mere identification; it aims to empower educational institutions to implement effective strategies, ultimately supporting student success and contributing to the broader societal goal of enhancing education outcomes.

2. DATA SET

Throughout our work, we made use of a dataset created from a higher education institution (acquired from several disjoint databases) 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. The dataset is available for download from the UCI Machine Learning Repository: here, and was donated on December 12, 2021. The dataset contains 4,424 instances, where each instance represents a unique student.

3. EXPLORATORY DATA ANALYSIS

3.1 Cleaning

As previously stated, each row corresponds to a unique student. However, the time period of the data, in terms of when it was all collected, is unknown. The dataset has 36 features, providing information known at the time of student enrollment, including academic path, demographics, and social-economic factors. The final feature, known as “Target,” indicates whether they are currently enrolled, graduated, or dropped out of their institution. When it comes to data cleaning, we first transformed the “Target” feature to be numeric rather than categorical so that the entire data set is in a numerical format. Instead of having possible values of “Dropout,” “Enrolled,” or “Graduate” we instead changed the values to 0, -1, and 1 respectively. Then we removed all rows from the data set that contained a value of -1 in the “Target” column since we are more concerned with whether a student has graduated or dropped out, rather than if they are simply enrolled. In later algorithms, the data set is then scaled again with a standard scaler to avoid overfitting the data.

3.2 DATA VISUALIZATION

The outcome, or target variable, in the project is related to predicting students' academic outcomes, specifically their likelihood of dropout and academic success. The problem is formulated as a binary classification task, where the possible outcomes are:

Dropout (0): Students who leave the academic program before completion.

Graduate (1): Students who successfully complete the academic program and graduate.

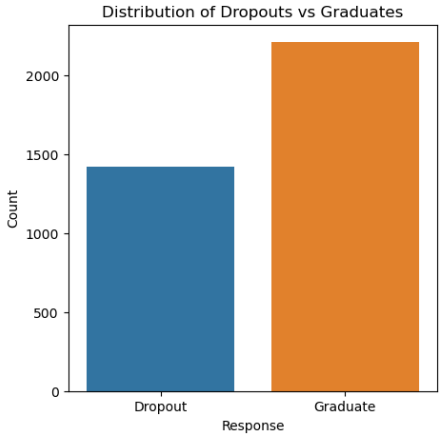

Figure 1. Count plot showing the number of times each response type appears in the data set.

In the context of predicting students' academic outcomes (Y), the key predictors (X) are likely to include a combination of demographic information, academic history, and socio-economic factors. A brief comparison of some features to the Target feature can be seen below. For example, “Tuition Fees Up to Date” and “Curricular Units 2nd Sem (Approved)” are of interest.

Figure 2. A heatmap comparing predictors to our targeted response feature.

4. MODELS

4.1 Initial Modeling

For our initial analysis, we have chosen to employ a logistic regression model, primarily driven by the nature of our response variable and key observations made during the prior EDA. Logistic regression is well-suited for binary classification, providing a probability estimate and facilitating clear interpretation of the relationship between predictors and the binary outcome. Our exploratory analysis revealed patterns and relationships in the data that align with the assumptions of logistic regression. Specifically, we observed non-linear trends and the presence of categorical variables, reinforcing the appropriateness of it over linear regression for our binary response variable. Furthermore, it provides coefficients that offer straightforward interpretability. This is crucial for extracting insights into how each predictor contributes to the likelihood of the binary outcome. Ultimately, logistic regression emerges as the clear choice for our model given the binary classification nature of our problem, the need for probability estimates, interpretability, and alignment with key observations. To assess the performance of our logistic regression model, we examined key metrics on the test set. The model demonstrated notable effectiveness, as evidenced by the following evaluation metrics:

• Accuracy: The accuracy of the logistic regression model on the test set was found to be 84.71%. This metric reflects the proportion of correctly classified instances and serves as a high-level indicator of overall model performance. It is well above what would be considered a “good” score of around 70-75%.

• Precision: Precision, a crucial metric in binary classification, measures the accuracy of positive predictions. In our case, the model achieved precision scores exceeding 80% for both classes (0 and 1). This implies that when the model predicts a particular class, it is correct at least 75% of the time.

- o Precision for 0: 84%
- o Precision for 1: 85%

The consistently high performance aligns with our initial choice of logistic regression. The model, designed for binary classification and probability estimation, has demonstrated its suitability for our problem.

```
Accuracy: 0.8471074380165289
Confusion Matrix:
[[213  71]
 [ 40 402]]
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.75      0.79       284
           1       0.85      0.91      0.88       442

    accuracy                           0.85       726
   macro avg       0.85      0.83      0.84       726
weighted avg       0.85      0.85      0.85       726
```

Figure 3. A more detailed view of our logistic model’s performance.

The observed test error rate, represented by the accuracy metric, indicates the proportion of correctly classified instances. Our high accuracy suggests that the model generally aligns well with the structure of the data. The high accuracy indicates successful overall classification, while precision scores shed light on the model's ability to make accurate positive predictions. To better visualize how our data fits, we leveraged the use of a Receiver Operating Characteristic Curve and a Precision-Recall Curve. The ROC Curve plots the true positive rate against the false positive rate at various threshold settings. A good model will have an ROC curve that hugs the top-left corner, which you can see demonstrated in our curve. A perfect P-R Curve would reach the top-right corner of the plot, and while ours does not achieve that, it does still stay close to the top portion of the graph. Overall, our model fits the data quite well.

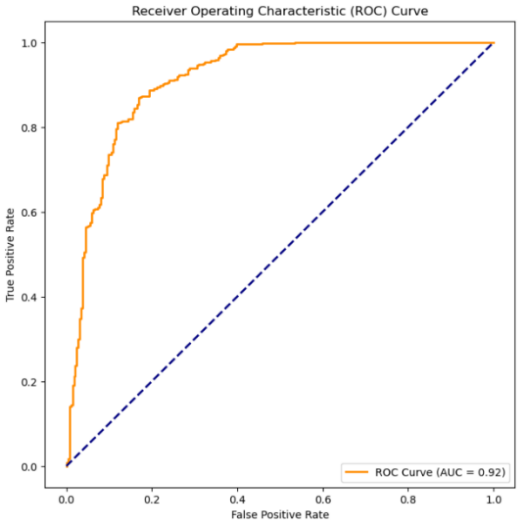

Figure 4. Resulting ROC Curve from our logistic model.

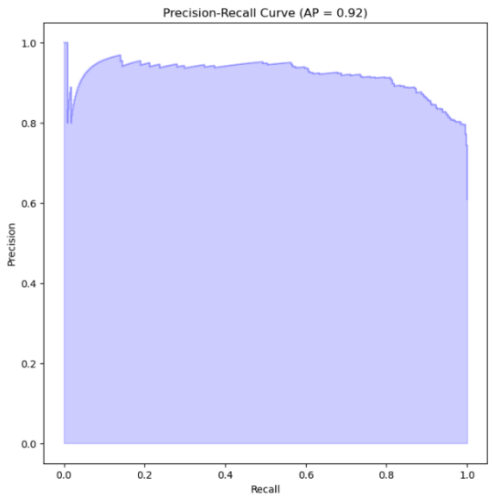

Figure 5. Resulting P-R Curve from our logistic model.

For the following sections, we will bring our implementation of CatBoost into play along with an implementation of a Random Forest Classifier, Label Spreading, and Self-Training algorithms. These algorithms were chosen as our three semi-supervised learning methods because of the nature of our dataset, when a portion of the labels is known or labeled. They utilize the available labeled data and propagate information from these labels to the unlabeled instances, using the underlying data structure to make predictions. We hope to compare the accuracy of all these different visualizations and see how they compare to each other along with the data sets initial accuracy scores and such provided by our logistic regression model. We sought to continue our trend of high accuracy scores throughout our testing.

4.2 Random Forest Classifier

We decided to make use of a Random Forest Classifier model (RFC) as the data is inherently imbalanced as you have positive instances and a larger set of unlabeled instances (presumed negatives). An RFC is known for its ability to handle imbalanced datasets effectively. The algorithm can adapt to different class distributions, making it suitable for situations where positive instances are scarce compared to negatives. An RFC can handle the absence of explicitly labeled negative instances during training. The decision trees in the ensemble consider subsets of features and data, and they implicitly learn patterns that discriminate between positive and unlabeled instances.

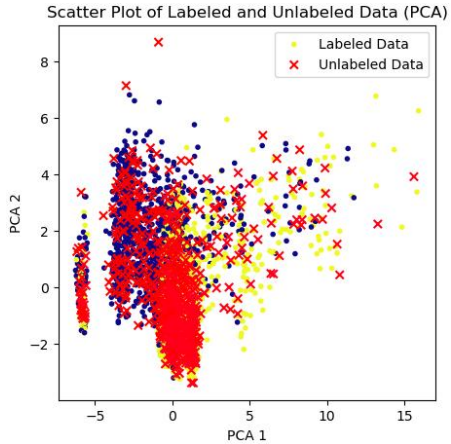

Figure 6. Initial scatter plot of our data before an RFC is used.

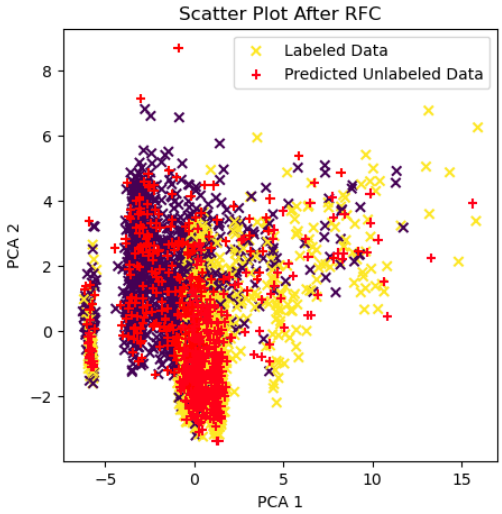

Figure 7. The resulting scatter plot after an RFC is applied.

As we can see, the two scatter plots look extremely similar to each other which is what we want to see. This lets us know that our RFC can accurately predict points from our data set when compared to where each point should be.

Test Statistics:

CatBoost Accuracy: 0.9036

Random Forest Classifier Accuracy: 0.8126721763085399

F1-Score: 0.81

NMI: 0.2820962115327092

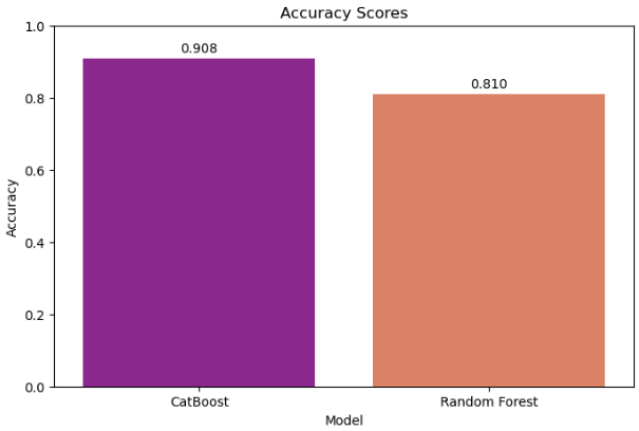

Figure 8. Bar graph comparing the accuracy of CatBoost vs our RFC model.

The RFC demonstrates a reasonably high accuracy, indicating a generally good fit to the test data. Furthermore, the F1-Score also emphasizes the balanced performance of the model in terms of precision and recall, which is essential in our PU Study where the positive class is of interest. The NMI, while not as strong as the accuracy and F1 scores, still indicates some level of agreement between our predicted and true labels.

4.3 Self-Training

In Self-Training algorithm we make use of the labelled and unlabeled data. With the labelled data, it is first split into train and test sets and a classification model, here, a logistic regression model has been utilized since the task at hand is of binary classification and trained.

Using the newly trained model, we generate “Pseudo-labels “for the unlabeled data which are compounded with the labelled training set that is used to train the model again. This model is used now on the labelled test data and the results are evaluated. Here, we are utilizing the accuracy score, t1 test and NMI Score to evaluate the results.

Test Statistics:

CatBoost Accuracy: 0.9036

Self-Training Model Accuracy: 0.7687224669603524

F1-Score: 0.8311897106109325

NMI:0.17486941161017483

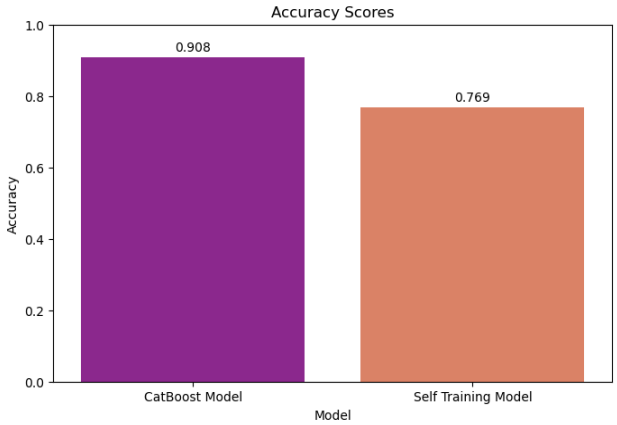

Figure 9. Bar Graph comparing the accuracy of the CatBoost Model vs Self Training Model.

For the self-training model, its accuracy is not as good as the RFC’s, but is still at an acceptable level, and it is important to keep other metrics in mind due to the imbalanced nature of our dataset. On the other hand, the F1-Score of this model is above that of our RFC which means that the model does a better job at minimizing both false positives and false negatives. This also may imply better performance overall when we are dealing with our imbalanced class distributions. However, its NMI is lacking when compared to our RFC which indicates a lower level of agreement between our predicted and true labels. This may suggest that there is a difference in the labeling pattern between our self-training model and the true labels.

4.4 Label Spreading

The dataset used contains a substantial number of unlabeled or weakly labeled data points, **so** semi-supervised methods like Label Spreading can effectively make use of this data to improve predictions:

X labeled: 3539 data points and 36 features

X unlabeled: 885 data points and 36 features

y labeled: 861 positives out of 3539 total

y unlabeled: 231 positives out of 885 total

If the dataset has a scarcity of labeled data but an abundance of unlabeled data, semi-supervised learning becomes advantageous. Label Spreading maximizes the use of available information by considering both labeled and unlabeled instances during training.

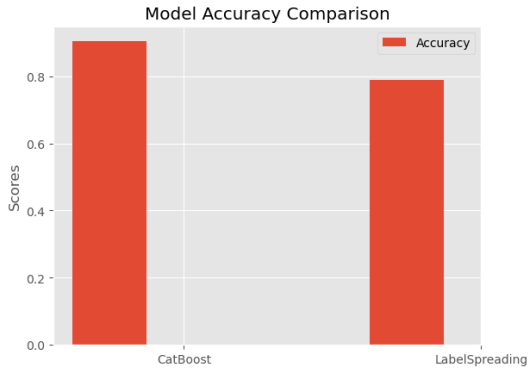

Figure 10. Initial performance of Label Spreading vs CatBoost.

Test Statistics:

CatBoost Accuracy: 0.9036

Label Spreading Accuracy: 0.7396694214876033

F1-Score: 0.7974276527331191

NMI: 0.1645409210854779

Our Label Spreading model appears to be our worst performing model for the task at hand. The model has a moderate level of accuracy which indicates a reasonable overall performance and has a relatively high F1-Score along with our other models. However, its NMI is comparable to the weak score provided by the Self-Training Model. Ultimately, Label Spreading underperforms in every metric when compared to our other models.

4.5 A General View of Our Models

When trying to decide which model is “best” we concluded that the RFC is likely our most suitable model due to it having the highest accuracy and NMI, along with a comparable F1-Score to our Self-Training model. However, none of the models come close to beating CatBoost when it comes to accuracy, but accuracy may not always provide a complete picture due to it being a very high-level overview of the model’s predictions. Lastly, we want to narrow down our dataset by quite a bit and only focus on the three features with the highest correlation to our response variable (see figure 2). We will focus on the following three features: Curricular units 2nd sem (approved), Course, and Tuition fees up to date. As you can see in the metrics below, narrowing our focus down to only three features resulted in an improvement in every single tested metric by quite a great margin.

```
Accuracy: 0.8939393939393939
Confusion Matrix:
[[221  63]
 [ 14 428]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.78      0.85       284
           1       0.87      0.97      0.92       442

    accuracy                           0.89       726
   macro avg       0.91      0.87      0.88       726
weighted avg       0.90      0.89      0.89       726

NMI: 0.518982844120032
```

Figure 11. Performance metrics of RFC after narrowing number of features tested.

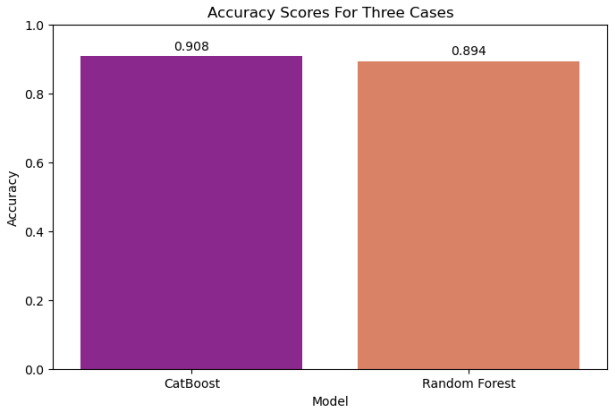

Figure 12. Resulting accuracy score of RFC vs CatBoost after narrowing number of features tested.

While our tests were mainly focused on the RFC, we were surprised to see that our Label Spreading model now outperformed CatBoost in terms of accuracy. The model went from being our worst performing to our best performing in accuracy all due to narrowing down our data set to key predictors only.

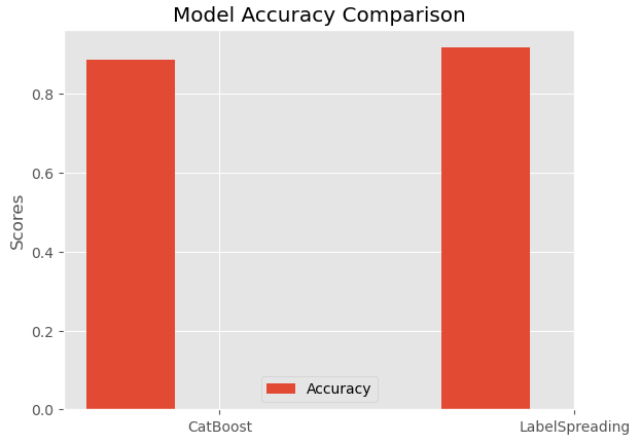

Figure 13. Bar graph comparing the accuracy of the Label Spreading model to the accuracy of CatBoost.

All tests now have scores comparable to CatBoost or even better than CatBoost in metrics such as overall accuracy. We can now conclude that the decision of whether we should simply default to CatBoost or not is not as black and white as it first seemed. Perhaps if our dataset is full of features like before then CatBoost can be a clear winner, but when narrowing our focus on a few features there is much more flexibility in which models to choose.

5. CONCLUSION

5.1 Retrospective

Given the results of our analysis, we have uncovered valuable insights into predicting and identifying students at risk of academic dropout and failure. The positive-unlabeled classification methods applied to the comprehensive dataset on student dropout predictors have allowed us to discern nuanced patterns and relationships within the data. Our findings suggest that specific factors, such as tuition fee status, course number, and approved course units, play significant roles in predicting students who may face academic challenges leading to dropout. By identifying these predictors, we can develop a targeted approach for early intervention. In response to the overarching question of how to predict and identify students facing academic challenges, our analysis highlights the importance of a multifaceted approach. Early detection can only be achieved through looking at a multitude of factors, as no single predictor can provide a concrete answer to our questions.

5.2 Usage and Application

Domain experts in the field of education and student support can glean valuable insights from our project that directly inform and enhance their work. For example, the project's findings can guide the development of targeted intervention strategies. Experts can use the identified predictors to design customized support programs that address the unique needs of at-risk students. This may involve creating academic assistance initiatives, mentorship programs, or some form of an economic support system tailored to the specific challenges identified in the analysis. The insights gained from our analysis can contribute to the enhancement of early warning systems as well. Those in the field can integrate the identified predictors into existing or developing early warning systems, improving their accuracy and effectiveness in flagging students at risk. This, in turn, enables timely and targeted interventions. Also, for those solely interested in the machine learning aspect of things and not the dropout intervention information, there is plenty to be learned from our study as well. Most importantly, we have shown that the model used for a particular study may not be as clear cut as we first thought. For example, label spreading, was our worst performing model by a clear margin but then shot to first place after simply reducing our number of predictors. Meanwhile CatBoost, our best model, was hardly affected at all by this change to our number of features studied. Therefore, future data scientists may want to use a variety of models in their studies to find what fits their work best, rather than going with what seems the best at first glance.

5.3 Improvements

If more time and resources were available for this project, one significant improvement could involve the incorporation of additional data sources to enrich the predictive modeling. For example, while the current analysis considers mostly academic factors, there could be additional non-academic variables that significantly influence student outcomes. Gathering data on aspects such as students' extracurricular activities, engagement in community programs, or health-related factors could provide a more holistic understanding of the challenges students face. Integrating such data could enhance the model's predictive accuracy and broaden the scope of intervention strategies.