COSC 4557/5557 Practical Machine Learning Spring 2024

Wild Card Project: Student Dropout Analysis for School Education

Submitted by: Iqbal Khatoon

Introduction

Access to quality education is a fundamental human right, and governments around the world are committed to ensuring that all children are enrolled in and complete their education. Despite these efforts, school dropout rates continue to be a significant issue, driven by a variety of social, economic, and demographic factors. To tackle this problem, the Government of Gujarat has embarked on a detailed study of dropout patterns at the school level. Understanding the root causes and identifying the groups most at risk will help the government develop specific interventions aimed at significantly lowering dropout rates. This project offers a detailed examination of trends in student dropouts within the school system, using a dataset named "Predict Students' Dropout and Academic Success - Investigating the Impact of Social and Economic Factors." This dataset, which is available on Kaggle and was provided by thedevastator (https://www.kaggle.com/thedevastator), includes a broad range of variables that illuminate the factors influencing student dropouts.

Project Overview

The primary objective of this project is to conduct a comprehensive analysis of student dropout rates in school education, with a focus on the state of Gujarat, utilizing the available dataset titled "Student-record"' Investigating the Impact of Social and Economic Factors." While the dataset may not include information on schools, areas, or castes, we can still extract valuable insights from the existing attributes.

About the Dataset

This dataset comprised of total 4424 entries and total of 35 columns, offers an extensive overview of students pursuing various undergraduate programs at a higher education institution. It encompasses a wide array of data, including demographic details, socio-economic factors, and academic performance metrics, which are pivotal for analyzing potential predictors of student dropout and academic success. The dataset is comprised of several separate databases, which collectively include vital information available at the time of student enrollment, such as application method, marital status, and chosen course.

Furthermore, the dataset facilitates the evaluation of student performance at the conclusion of each semester by examining the curricular units that students have credited, enrolled in, evaluated, and successfully completed, along with their corresponding grades. It also includes economic indicators like the unemployment rate, inflation rate, and GDP of the region, offering deeper insights into how economic conditions may influence students' decisions to continue their education or discontinue their studies. This robust analytical tool is invaluable for shedding light on the factors that encourage student retention or lead to dropout across a diverse array of fields including agronomy, design, education, nursing, journalism, management, social service, and technology.

Key columns include:

- Demographic information such as marital status, nationality, and age at enrollment.
- Academic performance metrics, including curricular units credited, enrolled, evaluated, and approved.
- Socio-economic factors like parental qualifications, occupation, and financial status.
- Macro-economic indicators such as unemployment rates, inflation rates, and GDP growth from the region.

Methodology

Data Preprocessing

- The dataset was cleaned and prepared for analysis by renaming columns for consistency and checking for missing values.
- Feature engineering included encoding categorical variables into numerical ones to prepare the data for machine learning models.

Correlation analysis was conducted to identify the features most strongly related to the target variable, 'Target', representing the students' status (Dropout, Enrolled, Graduate).

Exploratory Data Analysis (EDA)

- Various visualizations were created to explore the relationship between demographic and economic factors and student dropout rates.
- A correlation heatmap was used to determine the influence of various features on the likelihood of dropping out.

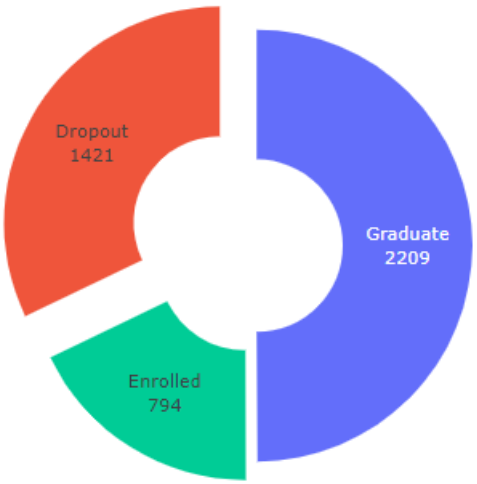

Figure1: 'Target', representing the students' status (Dropout, Enrolled, Graduate).

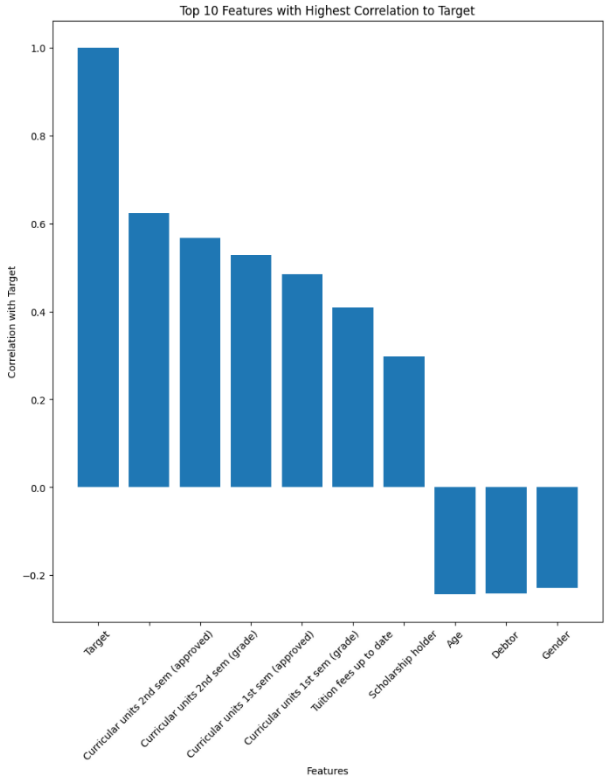

Figure2: top features with high correlation with target

We also visualized the Distribution of age of students at the time of enrollment to understand the data in better way, figure 3 shows the age distribution.

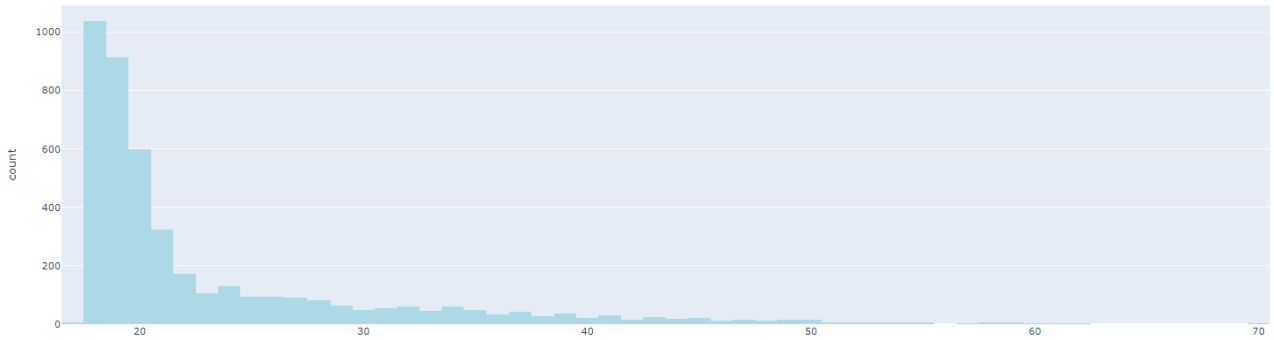

Figure3: Students age Distribution at the time of enrollment

Results

Following machine learning models were employed to predict dropout rates:

- Decision Trees
- Random Forest
- Logistic Regression
- KNN
- AdaBoost
- SVM

After training and testing the model performance we decided to perform Hyperparameter tuning on selected models using GridSearchCV to optimize their performance. Figure 4 shows the before and after tunning results.

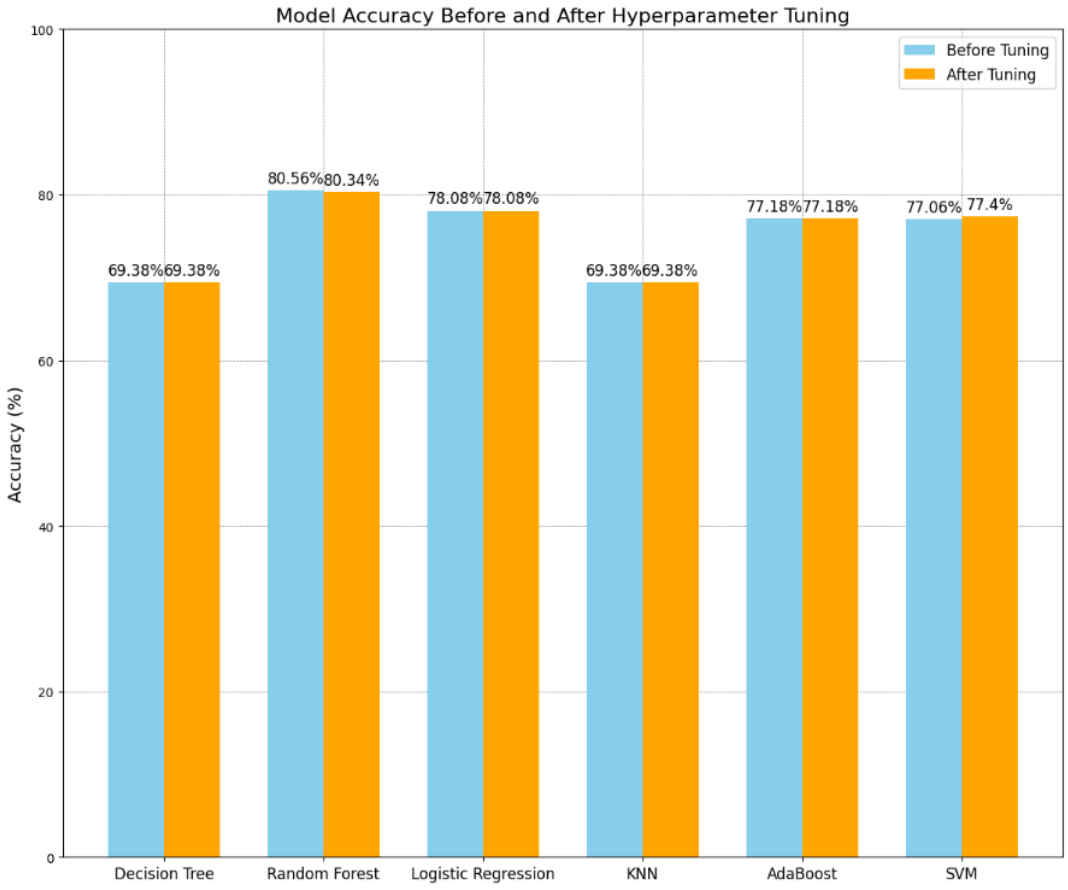

Figure4: All Model Accuracies before and after Hyperparameter Tuning

In order to enhance model's performance, ensemble methods, specifically Voting Classifiers with soft and hard voting, were utilized by combining the predictions of individual models.

Ensemble Techniques

The ensemble models with both soft and hard voting showed an accuracy of around 80%, indicating a robust model performance compared to individual classifiers.

Conclusion and Recommendations

The analysis revealed several key factors influencing student dropout rates, including socio-economic status, parental education, and individual academic performance. While some models showed significant predictive power, ensemble methods provided the most reliable results.

Recommendations for Policy Makers:

- Intervention Programs: Develop targeted interventions focusing on students identified at high risk based on the model predictions.
- Continuous Monitoring: Implement systems for continuous monitoring of student performance and socio-economic conditions.
- Policy Adjustments: Adjust educational policies to provide more support where it is needed, such as scholarships and financial aid, especially for vulnerable groups identified through the analysis.

This report provides a foundation for understanding the complex factors contributing to student dropouts and offers insights into potential areas for policy intervention to enhance student retention rates in Gujarat's schools.

References

[1] https://www.kaggle.com/thedevastator