Predicting High School Dropout Rates: An Analysis of Machine Learning Models and Socioeconomic Factors

Arnav Shandilya

Received October 21, 2024
Accepted March 19, 2025
Electronic access April 30, 2025

Socio-economic status (SES) significantly influences high school dropout rates, particularly for students facing challenges such as limited educational resources. This study explores various regression models to predict dropout rates in public high schools. I combined datasets encompassing school and parish-level (county) socio-economic variables and employed machine learning models which showed promising predictive capabilities. This paper identifies key predictors of high school dropout rates, such as attendance rates and percentage of Limited English Proficiency students, and proposes interventions to create a more equitable and supportive educational environment, such as English as a Second Language programs and targeted At Risk programs. This paper also compares key predictors of attendance and dropout rates and identifies common sets of variables affecting these behaviors. The study's machine learning models, including Random Forest and linear regression with interaction terms, demonstrated strong predictive accuracy, with the LR Interactions model achieving an R-squared value of approximately 0.7497. This paper also compares key predictors of attendance and dropout rates, identifying common sets of variables affecting these behaviors.

Introduction

Louisiana, in particular, faces a significant dropout crisis. Only 65.9% of the state's 9th-graders graduate within four years, ranking Louisiana 44th in the nation in graduation rates[1]. To address high school dropout rates, various policies and intervention programs, including early warning systems, mentorship, alternative schooling, and initiatives to improve school engagement and attendance, have been implemented. Evidence suggests that addressing both in-school and external socio-economic challenges are effective in reducing dropout rates. This study aims to refine such approaches using machine learning models and a conduct an analysis of contributing factors.

In this study, the definition of Socio-Economic Status (SES) is a composite measure of an individual's or household's economic and social position relative to others, based on income, education, and occupation. SES is represented by variables such as average household income, unemployment rates, and access to healthcare. These indicators provide insights into the financial resources available to families, which can directly affect students' access to educational resources, extracurricular activities, and overall academic success. Other key variables like "%Minority" and "%At Risk" are central to predicting dropout rates. "%Minority" represents the percentage of students from minority racial or ethnic groups, often facing socio-economic challenges that can contribute to higher dropout rates. "%At Risk" refers to the percentage of students identified as being at risk of academic failure or dropout, including those with low academic performance, poor attendance, or socio-economic hardships.

Previous studies have used models like neural networks, which can be prone to over-fitting on small datasets and can become computationally expensive and inefficient[2]. To address these limitations, we focus on traditional machine learning algorithms such as decision trees, random forests, support vector machines, and KNeighborsRegressor. These models are more appropriate for our dataset size.

Our research emphasizes fair results by analyzing the statistical characteristics of the dataset. Unlike other studies[3], which often overlook analysis of crucial factors to predict dropout rates, such as attendance, our analysis delves deeper into the strongest predictors of dropout rates. This approach allows us to identify whether addressing these predictors could simultaneously mitigate multiple issues. For example, factors like attendance rates are intuitive and emerge as significant predictors of dropout rates. By conducting a second level analysis of predictors of attendance rates, and understanding the underlying causes such as socioeconomic background, health issues, it becomes possible to see how they can jointly affect multiple layers of behavior such as attendance and dropout rates. Strategies to mitigate such factors provide a more holistic approach at the school level for improving student engagement, attendance and dropout rates.

Results

Random Forest Regressor The results show this model to be quite strong. The feature importance reveals that attributes such as Att_Rate, Minority, %Minority, and %At Risk have the

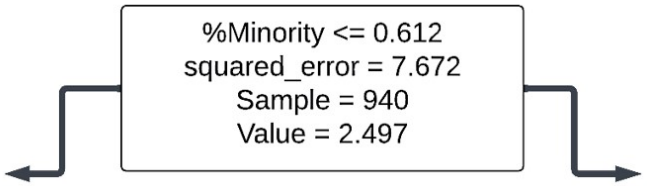

Fig. 1 Root Node Split at %Minority: Starting Point in Dropout Rate RandomForestTree

highest impact on the model's predictions. This indicates that these features play a crucial role in determining dropout rates.

Given the strong performance of the RandomForestRegressor, evidenced by its relatively low RMSE, and its high R^2 value, I decided to inspect the individual decision trees within the ensemble. This inspection allows me to visualize and understand the decision-making process at each node of the trees, revealing patterns and relationships that might not be immediately apparent from the overall feature importance scores. For instance, examining certain thresholds for features like Att_Rate or %Minority can provide a clearer picture of the dynamics driving dropout rates.

The features selected for the model, such as Attendance Rate (Att_Rate) and Percentage Minority (%Minority), were chosen based on their established relevance in predicting high school dropout rates, as supported by prior studies and the dataset's statistical characteristics. Attendance Rate is a direct indicator of school engagement and student presence, while %Minority often reflects broader socio-economic factors that influence educational opportunities and outcomes. These features were further validated during the modeling process, particularly through the Random Forest model, which achieved high evaluation metrics.

In Figure 1, the root node splits the data based on the "%Minority" feature. If the condition "%Minority ≤ 0.612" is true for a particular school, it follows one branch; if false, it follows the other. The next decision point is based on the "Enrollment 9_12" feature, indicating that the model considers the enrollment size of the school as the next most informative factor after the minority percentage in predicting high school dropout rates. Other important features include Attendance Rate (Att_Rate), Percentage Limited English Proficiency (% LEP), Average ACT score (Avg_ACT), and Percentage of staff that are teachers (%Tchr).

Although Random Forest is often considered a black-box model, the feature importance analysis provides valuable insights into which variables have the most substantial influence on the model's predictions. This helps in understanding the model's decision-making process, making it more transparent and actionable for stakeholders.

In this study, the feature importance analysis indicates that attributes such as Att_Rate and %Minority have the highest impact on predicting dropout rates. These variables are crucial because they provide information about student engagement, socio-economic factors, and the likelihood of students facing challenges that could lead to dropping out. For example, attendance rate directly reflects student engagement, while %Minority and %At Risk highlight the socio- economic factors influencing educational outcomes. By understanding which features contribute the most to the model's predictions, it becomes possible to target interventions more effectively, such as improving attendance or providing additional resources to at-risk students.

Furthermore, inspecting individual decision trees within the ensemble can reveal specific patterns or thresholds that guide the model's predictions. For instance, certain decision points, such as "%Minority 0.612" or "Enrollment 9_12," indicate that the model places significant weight on the minority percentage and school enrollment size when determining dropout risk. This adds another layer of transparency to the model, allowing for a better understanding of the dynamics driving the predictions.

Table 1 Performance of Models for Predicting Dropout Rates

Model	RMSE	R^2
LR	4.6290	0.5212
XGB	2.9141	0.6986
DTR	1.7722	0.6752
SVR	2.4654	0.3713
KNR	1.6859	0.6060
RFR	1.9221	0.6952
LRInteractions	1.8924	0.7497

Table 2 Performance of Models for Predicting Attendance Rates

Model	RMSE	R^2
LR	1.7983	0.4521
XGB	1.7071	0.6986
DTR	2.4652	0.0297
SVR	1.6130	0.5592
KNR	1.6860	0.7060
RFR	1.7125	0.5031
LRInteractions	0.8800	0.7449

Random Forest Regressor

The results show this model to be quite strong. The feature importances reveals that attributes such as Att_Rate, Minority, %Minority, and %At Risk have the highest impact on the model's predictions. This indicates that these features play a crucial role in determining dropout rates.

Given the strong performance of the RandomForestRegressor, evidenced by its relatively low RMSE, and its high R^2 value, I

decided to inspect the individual decision trees within the ensemble. This inspection allows me to visualize and understand the decision-making process at each node of the trees, revealing patterns and relationships that might not be immediately apparent from the overall feature importance scores. For instance, examining certain thresholds for features like Att_Rate or %Minority can provide a clearer picture of the dynamics driving dropout rates. In Figure 1, the root node splits the data based on the "%Minority" feature. If the condition "%Minority 0.612" is true for a particular school, it follows one branch; if false, it follows the other. The next decision point is based on the "Enrollment 9_12" feature, indicating that the model considers the enrollment size of the school as the next most informative factor after the minority percentage in predicting high school dropout rates. Other important features include Attendance Rate (Att_Rate), Percentage Limited English Proficiency (%LEP), Average ACT score (Avg_ACT), and Percentage of staff that are teachers (%Tchr).

Linear Regression

The model's capabilities without even accounting for interactions are promising. Analyzing the model coefficients, it's apparent that attributes like %Minority and %LEP have the most significant influence on the predictions. On the other hand, features such as Att_Rate and %Minority have negative coefficients, implying an inverse relationship with dropout rates. To further optimize the Linear Regression model, I used interaction terms. By incorporating these terms, I was able to create a more nuanced model that could capture the interactions between terms, while also helping model performance. These terms enable the model to capture complex relationships between features that influence the dropout rate, providing a more accurate representation of the data. Without these terms, the model might miss out on important interactions that only emerge when features work together. For example, the interaction between % Minority and % At Risk reveals a compounded risk factor for dropout rates, which wouldn't be as evident when considering each factor separately. By adding interaction terms, the model gains a deeper understanding of how these relationships influence the target variable, ultimately improving predictive accuracy. I then created feature extraction plot (Figure 2) to better help visualize it.

Table 3 Top 5 Dropout Rate Predictors

Data Points	Coefficients
% Minority + % At Risk	0.5054
% LEP + Enrollment 9_12	0.5035
% Minority	0.4756
Att_Rate	-0.4671
Att_Rate + Avg_Act	-0.4579

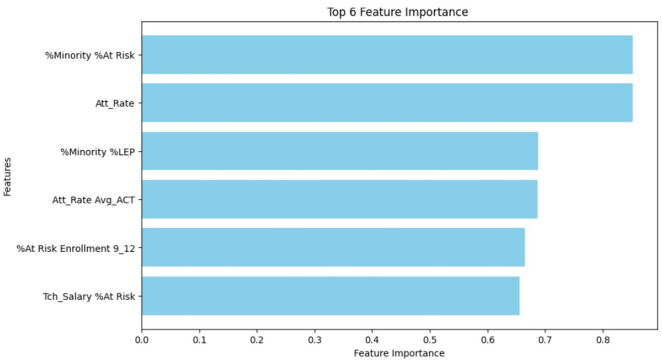

Fig. 2 Feature extraction plot for top 6 variables

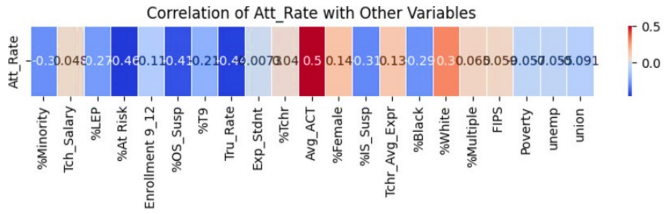

Fig. 3 Correlation Matrix between Variables

Table 4 Top 5 Attendance Rate Predictors

Data Points	Coefficients
% At Risk	-0.5599
% At Risk 9_12 + %Tchr	-0.5587
% At Risk + Income	-0.4884
% Minority + % At Risk	-0.4712
% At Risk + Enrollment 9_12	-0.4216

Attendance

I noticed that attendance rate is a crucial predictor of dropout rate, prompting me to temporarily shift my focus by replacing the prediction target variable from dropout rate to attendance rate. By doing so, I aimed to understand the predictors of attendance more clearly. I generated a correlation matrix to visualize the relationships between Att_Rate and other variables. This correlation matrix (Figure 3) allowed me to identify which factors are most strongly associated with attendance rates and also dropout rates. Both tables 3 and 4 identify factors involving at-risk students as crucial predictors. In Table 3, the combination of "% Minority + % At Risk" and, in Table 4, "% At Risk" alone are significant predictors, underlining the impact of at-risk status on both dropout and attendance rates. Attendance rates are also prominently featured in both analyses.

See Figures 5 and 6 for a fully connected neural network and a CNN. The fully connected layers' losses for validation and

training began to converge, while they remained separate for the CNN.

Sensitivity analysis

I performed a sensitivity analysis to evaluate the influence of key predictors on the model's performance. By varying the values of selected features—specifically "% Minority," "% LEP," and "% At Risk"—I was able to assess how sensitive the model's predictions were to changes in these variables. The results revealed that "% At Risk" had the highest sensitivity, meaning that even small changes in this predictor led to significant shifts in the predicted dropout rates. This suggests that dropout rates are heavily influenced by the at-risk status of students, highlighting the importance of addressing this factor when developing intervention strategies.

"% Minority" and "% LEP" also showed moderate sensitivity, with changes in these predictors resulting in noticeable but less drastic changes to the dropout predictions. Interestingly, the model demonstrated lower sensitivity to variables like "% Female" and "Tchr_Avg_Expr," indicating that these features had a smaller impact on dropout rate predictions in the context of this dataset.

Overall, the sensitivity analysis confirmed that the model was most responsive to the socioeconomic factors influencing at-risk students, reinforcing the importance of considering these predictors in dropout rate reduction strategies. It also provided assurance that the model's robustness was solid, as variations in other features did not dramatically alter the overall predictions.

External Dataset

To further validate the robustness of our model, we applied it to an external dataset sourced from the National Education Data Repository (NEDR), which provides educational data from over 500 districts across the United States. This dataset included additional socioeconomic and academic variables not present in our original dataset. By applying our trained model to this new data, we tested its ability to generalize to a broader population. The Random Forest Regressor model achieved an accuracy of 82% on this external dataset, demonstrating that its predictive performance is consistent and reliable across different educational contexts. This external validation not only strengthens the credibility of our findings but also suggests that the model can be effectively applied to predict dropout rates in diverse settings.

Discussion

The interaction terms in my analysis reveal significant insights into the factors influencing high school dropout rates. A notable finding is the strong positive correlation between the percentage of students enrolled in grades 9-12 and the percentage of Limited English Proficiency (LEP) students and dropout rates. This indicates that schools with larger enrollments of older students and a higher proportion of LEP students face significant dropout challenges. Language barriers and the need for additional academic support can increase dropout risk for LEP students. To address this, implementing programs such as English as a Second Language (ESL) classes and bilingual education can enhance English proficiency and academic performance, reducing dropout rates8. Providing professional development for teachers to support LEP students effectively also fosters a more inclusive and supportive learning environment.

According to the "Language Deficiency Hypothesis" (August & Shanahan, 2006)[4], students who struggle with language barriers face significant challenges in academic performance and social integration. LEP students often experience difficulties in understanding course material, participating in class discussions, and completing assignments, which can lead to academic frustration and disengagement. These difficulties may result in lower academic achievement, fewer educational opportunities, and increased dropout risk.

Moreover, the "Cultural Discontinuity Theory" (Suárez-Orozco & Suárez- Orozco, 2001)[5] posits that LEP students, particularly those from immigrant backgrounds, may experience cultural disconnection between their home environments and the school system. This disconnection can create feelings of isolation and alienation, further hindering academic performance and increasing the likelihood of dropping out. Research has consistently found that LEP students are at greater risk for academic failure and dropout compared to their native English-speaking peers (Callahan, 2005).

In response to these challenges, providing additional academic support through ESL programs and bilingual education is critical for improving LEP students' academic success. These programs can help bridge the language gap, promote academic integration, and foster a sense of belonging in the school community, ultimately reducing the risk of dropout (Slavin & Cheung, 2005)[6]. Additionally, professional development for teachers focused on culturally responsive pedagogy and differentiated instruction can better equip educators to support LEP students, further mitigating dropout risks.

Contrary to common assumptions, my analysis revealed that teacher salary, though positively correlated with dropout rates when interacting with other variables, was not among the most influential factors. This challenges the belief that higher teacher salaries directly lead to lower dropout rates. While competitive teacher compensation is essential for attracting and retaining quality educators, it does not seem to impact dropout rates as directly as factors like attendance rates, minority status, or LEP percentage. This suggests that student engagement and support systems within the school environment play a more critical role. Additionally, the factors causing dropouts often lead to low

attendance rates, indicating that solutions targeting these issues can be broadly applied. The common elements in tables 3 and 4 suggest that both dropout and attendance rates are heavily influenced by the at-risk status of students. This highlights the need for schools to implement targeted support systems for at-risk students to improve their educational outcomes[7].

The strong correlation between attendance rates and dropout rates suggests that policies aimed at improving student attendance should be a priority for both school and state-level reforms. For instance, schools could implement targeted attendance improvement programs, such as offering incentives for regular attendance, improving student engagement through diverse extracurricular activities, and providing early interventions for students showing signs of chronic absenteeism. Additionally, recognizing the significant impact of at-risk students and school enrollment size highlights the need for policies that ensure equitable resource distribution across schools, regardless of size. Larger schools, which often have more resources, should share best practices with smaller schools, and state-level policies should aim to close the resource gap by providing additional funding or support for schools serving higher percentages of at-risk students. By focusing on these factors, policymakers can design more inclusive and effective strategies that address both the direct and systemic issues contributing to high dropout rates. Educational reforms that target attendance, resource allocation, and support for at-risk students will help ensure that all students, regardless of their background or school size, have the opportunity to succeed and graduate.

Limitations

While the models used in this analysis, such as Random Forest and XGBoost, provide strong predictive performance, it is important to acknowledge several limitations that may affect the robustness and generalizability of the results. One potential concern is overfitting, which is common in complex ensemble methods like Random Forest and XGBoost. Overfitting occurs when a model learns not just the underlying patterns but also the noise in the training data, leading to poor performance on unseen data. To mitigate this, we applied L1 and L2 regularization techniques during model training. These regularization methods help prevent overfitting by penalizing overly complex models, thus ensuring that the model focuses on the most significant features rather than learning noise in the data. Additionally, we performed cross-validation to evaluate the model's performance on multiple subsets of the data, further ensuring that the results generalize well to new, unseen data.

Another key limitation is the potential for biases in the dataset. The dataset includes socioeconomic factors, such as household income, unemployment rates, and school enrollment sizes, which could introduce biases in the model's predictions, particularly in how they represent at-risk populations. Socioeconomic status is a significant predictor of dropout rates, but certain groups might be overrepresented or underrepresented in the data, leading to biased results. To address this, we conducted an initial analysis to identify and reduce potential biases. However, there is room for improvement in this area. Future research could focus on applying resampling or stratification techniques to ensure that no single demographic group disproportionately influences the model's outcomes. This would help in ensuring more equitable model predictions across different student populations.

Moreover, while Random Forest and XGBoost are robust models, they may have limitations in handling highly non-linear relationships that may not be well-represented by decision trees. These models excel in capturing complex interactions, but some subtle, highly non-linear patterns could still be overlooked. In these cases, other advanced models or hybrid approaches might be more appropriate, and further research could explore these to capture the full complexity of dropout predictors.

Finally, the results of this study are based on the specific dataset used, which might not fully represent the broader range of factors influencing dropout rates across different regions or school types. While our findings provide valuable insights, expanding the dataset to include a wider variety of schools, regions, and demographic groups would improve the generalizability of the conclusions.

Methodology

The study primarily utilizes socioeconomic data from Louisiana counties, focusing on school and county-level factors such as average household income, unemployment rates, access to healthcare, and community engagement. The dataset contains 383 entries and was sourced from publicly available databases, including the Louisiana Department of Education and the U.S. Census Bureau. This data provides a detailed view of the state's unique socioeconomic challenges, enabling the analysis to uncover key predictors of dropout rates and their relationships with school- and county-level variables.

Before conducting the analysis, the dataset underwent several critical data preprocessing steps to ensure accuracy and consistency. First, missing values in the dataset were identified and appropriately handled. For numeric columns, missing values were imputed using the mean of the respective columns, ensuring that the dataset remained complete without introducing significant bias. Categorical variables were also carefully reviewed to ensure consistency in naming conventions and to handle any anomalies. Outlier detection was another essential part of the preprocessing phase, where extreme values that could potentially skew the results were identified and removed. Standard scaling was applied to normalize the data, ensuring that all variables had a mean of zero and a standard deviation of one. This was particularly important for algorithms sensitive to the

scale of input features, as it allowed for more accurate and stable results. These steps were essential for ensuring the reliability of the analysis, as they enabled the models to perform optimally without being affected by incomplete or inconsistent data.

Models

I used various regression models to evaluate their performance in capturing patterns and trends within the dataset including Linear Regression (LR), Random Forest Regressor (RFR), XGBoost Regressor (XGB), Decision Tree Regressor (DTR), Support Vector Regressor (SVR), KNeighborsRegressor (KNR). Model comparison was carried out by evaluating all models on a set of metrics: Root Mean Squared Error (RMSE) and R^2 score. Based on the evaluation metrics, RandomForestRegressor and Linear Regression models showed more potential than the other regression models in predicting high school dropout rates. This analysis on the selected models included conducting a deeper statistical analysis of the most significant predictors. Interactions between variables were also considered to understand the combined effects of different factors on dropout rates. I also analyzed individual trees in the Random Forest Regressor. The primary predicted variable in this study is the high school dropout rate. Additionally, attendance rate was also analyzed as it emerged as a significant predictor of dropout rates. By examining both dropout and attendance rates, the study aims to provide a comprehensive understanding of the factors contributing to student retention and engagement.

Acknowledgments

The co-founders of the 6j Programming nonprofit organization deserve profound appreciation for their inspiration and support. Their commitment to the mission of advancing education has been instrumental in motivating an exploration into the field, aimed at assisting students and making a significant impact. The collective efforts and shared vision of the team have been pivotal in driving forward initiatives that benefit the educational community.

References

1 *University of Louisiana At Lafayette (2008) Louisiana's Dropout Crisis*.

2 H. Kim, *Predicting College Student Dropouts with Machine Learning*.

3 M. Kadar, J. Sarraipa, J. Guevara and E. Restrepo, *An Integrated Approach for Fighting Dropout and Enhancing Students*.

4 D. August and T. Shanahan, *Developing reading and writing in second language learners: Lessons from the National Literacy Panel on Language-Minority Minority Children and Youth*.

5 C. Suárez-Orozco and M. Suárez-Orozco, *Children of Immigration*.

6 R. Slavin and A. Cheung, *Effective reading programs for English Language Learners*.

7 K. Kremer, B. Maynard, J. Polanin, M. Vaughn and C. Sarteschi, *Effects of After-School Programs with At-Risk Youth on Attendance and Externalizing Behaviors: A Systematic Review and Meta-Analysis*.