Predicting Student Dropouts from Socioeconomic Variables

Samdeet Khan, Machine Learning Engineer Intern

BRAC Education Programme (BEP)

September 2024

TABLE OF CONTENTS

CHAPTER Page

1. INTRODUCTION……..……..1
2. LITERATURE REVIEW…….……..…….....2
3. METHODOLOGY…….....8
4. IMPLEMENTATION…….....18
5. RESULTS……..24
6. DISCUSSION……..….30
7. CONCLUSION…….....32

REFERENCES……..…34

Introduction

The increased availability of educational data has presented the opportunity to leverage machine learning techniques to address critical issues in education, including student retention. This summer, I have had the opportunity to apply these methodologies in the context of BRAC's Education Programme (BEP) by leveraging student metrics sourced by its Monitoring & Evaluation team to predict student dropouts. The aim of this project is to identify at-risk students earlier, with the aim of improving educational outcomes. The machine learning models include logistic regression and decision trees, which utilize a variety of academic, demographic, and behavioral variables for their respective training and validation procedures.

The models were assessed based on accuracy, precision, recall, and F1 score to determine the most reliable predictors. Special attention was given to the interpretability of the models, which ensures that the results can be effectively communicated to educators, policymakers, and other relevant stakeholders. A key aspect of interpretability for this project is the provision of feature importance, which indicates which variables have a direct impact on the likelihood of student dropouts.

The results demonstrate that machine learning models provide significant ability to predict student dropouts, offering a valuable tool for educational institutions striving to improve retention rates. This research emphasizes the potential of data-driven approaches to foster positive social change in education, contributing to broader discourse on the ethical and impactful use of artificial intelligence.

Literature Review

Education systems globally face the daunting challenge of dropout rates that threaten the quality and effectiveness of teaching and learning environments. In Bangladesh, despite high initial enrollment rates in primary education, sustaining student engagement through to completion remains a persistent hurdle. This phenomenon is particularly pronounced in contexts where foundational skills in reading and mathematics are not adequately developed early in a child's educational journey, leading to a critical "third zone of exclusion," as identified by Lewin and Little (2011). BRAC's Education Programme (BEP) has been pivotal in addressing such challenges, pioneering initiatives like the Non-Formal Primary Education (NFPE) and the Bridge Programme aimed specifically at children at risk of dropping out.

The Bridge Programme, with its targeted approach, provides a re-entry pathway for students who have previously dropped out, offering an accelerated primary education cycle. This initiative is crucial as it aligns with the United Nations Sustainable Development Goal 4 (SDG 4), which calls for inclusive and equitable quality education and promotes lifelong learning opportunities for all. Notably, the Bridge Programme has enabled students to complete the primary education cycle significantly faster than the conventional route, through tailored curriculum materials that cater to the urgency of student motivation and the necessity for immediate academic progress. Indeed, the programme has enabled students who have been away from school o re-enter and complete the primary school cycle 20% faster than the normal cycle from Grade 1 to 5.

BRAC has outlined many important learning goals within the framework of its educational curricula. The following subjects represent these projected outcomes: reading and writing, numeracy, social values, basic scientific ideas, geography, and history. Special attention

is given to ensuring that numeracy as well as reading and writing skills are well-established within the first five years of a child's education. Evidence has showed that a lack of a proper foundation in these two areas can cause significant hurdles for a student's overall social development and ability to integrate into a productive role as an adult. (Evans and Hares, 2021) Thus, the ability to retain students during these early stages is especially crucial, before their performance and confidence issues become unresolvable. If early identification is possible, teachers can, each day devote attention to specific students' goals. (Chirstenson and Thurlow, 2004)

BRAC has been complementing the Bangladeshi government's efforts for educating eligible children in Bangladesh for over three decades. However, despite all its impactful initiatives (e.g. Non-Formal Primary Education and Education for All), dropouts have become one of the biggest barriers to achieving universal primary education. BRAC targets student dropouts, although the likelihood of students still dropping out of schools can still arise even after rejoining, especially if the students face similar conflicts that prevented their original school attendance. The Non-Formal Primary Education, or NFPE, is a system of education implemented by BRAC that consists of a non-formal one-teacher, one-classroom school setting within local communities. Targeted efforts by the Bangladeshi government, development partners, multilateral and bilateral donors, civil society organizations, and NGOs like BRAC have helped Bangladesh reach over 98% enrollment in primary school. Yet, the completion rate remains unsatisfactory; over 20% of children who enroll drop out before completing he full five-year cycle of primary education.

The United Nations' Development Goal 4 (SDG4-1) requires countries to ensure primary and secondary education with quality and equity. BRAC's Bridge schools can help to achieve

SDG4 by 2030, but halfway through the 2015-2030 period, his goal has seemed impossible to achieve. Dropout at secondary education makes the goal even more challenging to achieve due to the third of secondary students who leave high school without finishing. Additionally, learning loss due to the COVID-19 pandemic has caused lower student achievements, which has caused higher dropout rates. The presence of a system that can help to identify at-risk students and provide relevant interventions early-on becomes an imperative if the SDG4 is still outlined as an achievable outcome by 2030. This technology can complement the presence of programs like BRAC's Bridge Programme which are tailored to students who experience a disruption in their primary school attendance and academic abilities.

Many reasons exist why early disruptions in student learning outcomes occur in South Asian countries (i.e. India, Bangladesh, Pakistan, Nepal, and Afghanistan) and Sub-Saharan African countries, which are major locations of young adults unable to read. These causes include the need to generate income for one's family, family migration to other parts of the country where more income can be found, displacement from natural disasters (e.g. floods, cyclones, and social unrest), and inability to keep pace with the school curricula, which public schools are especially unable to cater to. Despite the public school system's accessibility, the rigidity in schedule and teaching strategies hinders the acceptance of students from challenging circumstances. For example, despite the high enrollment rate for students in Bangladesh, the 2017 World Bank "learning poverty index" reported that 58% of children at the upper primary grade cohort (10 through 13) were unable to read simple text at a basic level. (World Bank, 2021) Furthermore, with the dropout rate from the primary education cycle in Bangladesh being around 20%, many children still require interventions that cater to their specific needs.

The following list builds on the previously mentioned reasons for early student dropouts, which were primarily geographical and financial in nature. These 17 criteria provide a basis for understanding which metrics, sourced by BRAC's Monitoring & Evaluation team, would be relevant in a dropout prediction model:

1. Financial problems
2. Parents' willingness
3. Distance and lack of basic facilities
4. Inadequate school environment
5. Overloaded classrooms
6. Improper teaching languages
7. Carelessness of teachers
8. Security problems for girls
9. Poor home conditions
10. Grade retention
11. Student's out-of-school companionship
12. Truancy
13. Difficulty in learning
14. Students' choice of labor over studies
15. Psychological problems
16. Parents' illiteracy
17. Poor health

Some of the more powerful interventions for dropouts have involved the concentration on social-emotional learning. (Wang et al. , 2016; Remschmidt et al. 2007) Concentration on social

and emotional status of children may be nested within a package of targeted interventions, which a dropout model would complement by providing comprehension of which students to give more specific focus in these aspects. (Graeff-Martins et al., 2006) In some instances, the social and emotional health of children is described as the "whole school approach", which involves coordinated activities across breadth of the curricula, ethos and positive environments of the schools, partnerships with families and local communities, and an "inclusive education philosophy including social, professional, and school transformations." (Farooq, 2013 p 47) Indeed, a positive atmosphere in school allows the students to associate attending school as a choice to "feel well and not a compulsory requirement." (Frederico, 2019; Batini et al., 2019) Students think about dropping out when they get discouraged by their academic environment and feel that there is little hope to overcome social and performance conflicts. (Zuilkowski, Jukes, and Dubeck, 2016) As shown in the list of dropout factors above, school experience isn't just limited to classroom performance; if a child feels bullied by other children, for example, it can have detrimental effects on their desire to learn. (Townsend et al., 2008) Another significant contributor is parent involvement. Children of parents who did not participate in parent-teacher meetings, discuss academic progress with their child's teacher, or supervise homework were more likely to drop out of school. (Sabates, Hossain, and Lewin, 2013) Interventions that work help establish a moral responsibility in families to send their children to school. (Mishra & Abdul-Azeez, 2014)

The Monitoring and Evaluation team at BRAC provides a method for examining key metrics related to predictive variables like family involvement and social disillusionment that students experience. Monitoring follows a systematic process to gather data regarding progress made by an implemented BRAC program, such as that of a Bridge Program school, while

Evaluation follows a process of determining whether the program has reached its optimal goals. In other words, the M&E team at BRAC takes the lead in data collection, data utilization, and overall supervision to see if the indicators set at the program's inception are met or not. Due to this jurisdiction, the work done by the M&E team is crucial for the purposes of developing machine learning models for student dropout predictions, which is the aim of this project.

Methodology

The first step for this project was to access BRAC's student database to understand which metrics would be most relevant for predicting student dropouts. As mentioned in the literature review, the Monitoring and Evaluation (M&E) team leads data collection and analysis efforts within BRAC's Education Programme (BEP). Thus, collaboration with the M&E team was necessary to identify and gather comprehensive datasets that track academic performance, attendance records, socio-economic background information, and behavioral indicators of students within BEP's school network. [Data privacy was upheld by using student ID's as an identifier rather than students' names and by maintaining strict confidentiality in data sharing with other BEP team members.]

After obtaining access to a snapshot of the database from the IT department. I began configuring the required coding environment to perform data extraction, exploratory data analysis (EDA), and implementation of machine learning models. This step involved the installation of Python 3.12.3, Jupyter Notebook, MySQL 8.0, and git 2.46.0. Additionally, a local SQL server was built to handle the student database from the M&E team. Once these were installed, I downloaded relevant packages in Python: pandas, matplotlib, seaborn, and sqlalchemy. The pandas package allowed me to store the database tables into simplified dataframes that were less intensive to load and manipulate. Matplotlib made it possible to generate a variety of plots for data visualization and other EDA purposes. Seaborn, which build on matlab,provided additional sophisticated visualization options that made it easier to produce more complex charts for detailed EDA. For database access and data sourcing, sqlalchemy provided a means of connecting to the local database and running relevant SQL queries within the Jupyter notebook environment, rather than relying on a separate database interface, like

MySQL Workbench. This integrated setup streamlined the data handling process and allowed for more efficient data manipulation and analysis. As a result, I was able to interact with the student database, perform real-time analysis, and visualize results in a seamless workflow.

```
engine = create_engine('mysql+pymysql://root:[redacted]@localhost:3306/emis')
query1 = "SELECT * FROM student"

# Load data into dataframe
student_data = pd.read_sql_query(query1, engine)
print(student_data)
```

Figure 1. Python code for loading data from 'student' MySQL table into a pandas dataframe

With the technical setup complete and the data accessibility established, the next step was to understand the relevant student data table categories. The dimensions of the raw dataset were 789,129 rows by 106 columns which caused the dataframe output from Figure 1 to be truncated, making it impractical to understand relevant categories from the output alone. Due to the large number of categories, I iterated through the category names in batches of 20 until all of them were visible from the output. The code in Figure 2 below shows my iteration logic, as well as the corresponding output of all categories present in the student data table. The list, although comprehensive, allowed me to compare variables with those known to contribute to student dropout probabilities from the previous literature review section, including financial problems, parents' characteristics, and student sense of belonging inside and outside of the classroom. Among the listed categories also existed the outcome variable, 'dropout', and identifier variable, 'id', which made it feasible to use this data table as a starting point for building a predictive machine learning model. Predictor variables that were filtered for their relevance were 'sex', 'is_orphan', 'is_ethnic', 'is_never_been_to_school', 'father_educational_attainment',

'mother_educational_attainment', 'relation_with_guardian', 'parents_income', 'previous_dropout', 'pwd_degree', 'received_any_treatment', 'marital_status', and 'newly_admitted'.

```
print('Student Data Categories:')
for i in range(0, len(student_data.columns), 20):
    print(student_data.columns[i:i+20])
```

```
Student Data Categories:
Index(['id', 'created_by_date', 'created_by_id', 'created_by_name',
       'last_modified_by_date', 'last_modified_by_id', 'last_modified_by_name',
       'address', 'bkash_number', 'brac_graduate', 'date_of_birth',
       'dialect_spoken', 'dropout', 'family_members_involve_with_brac',
       'family_members_involve_with_brac_service', 'father_dob',
       'father_educational_attainment', 'father_name', 'father_occupation',
       'first_name'],
      dtype='object')
Index(['grade_id', 'guardian_mobile', 'height', 'institute_id',
       'involved_with_chhatrabandhu', 'last_name', 'latitude',
       'location_hierarchy', 'location_id', 'location_type_udv_id',
       'longitude', 'middle_name', 'mother_dob',
       'mother_educational_attainment', 'mother_name', 'mother_occupation',
       'name_of_transferred_school', 'religion', 'replacement',
       'residential_address'],
      dtype='object')
Index(['roll', 'session_end', 'session_start', 'sex', 'student_id',
       'student_type_udv_id', 'transfered_school', 'transferred_school_id',
       'type_of_csn_udv_id', 'type_of_ethnicity_community_udv_id', 'waiver',
       'weight', 'counter', 'closed', 'migration_dates', 'prev_grade_ids',
       'is_updated', 'birth_certificate_id', 'bmi', 'device_type'],
      dtype='object')
Index(['dropped_out_enrolled_class', 'dropped_out_school_name',
       'father_income', 'father_nid', 'graduate_level', 'guardian_name',
       'import_history_id', 'is_ethnic', 'is_ndd', 'is_never_been_to_school',
       'is_orphan', 'is_participate_brac_other_service',
       'is_participate_with_other_ngo', 'is_pwd', 'is_received_device',
       'marital_status', 'mother_income', 'mother_nid', 'name_of_the_cluster',
       'ndd_type'],
      dtype='object')
Index(['operational_approach', 'parents_income', 'pwd_degree', 'pwd_type',
       'received_any_treatment', 'relation_with_guardian', 'section',
       'transferred_institute_address', 'treatment_type',
       'institute_type_udv_id', 'operation_category_udv_id', 'pngo',
       'currently_dropout', 'previous_dropout', 'old_institute_id',
       'section_udv_id', 'old_code', 'old_grade_id', 'admission_date',
       'is_admitted'],
      dtype='object')
Index(['unique_student_id', 'newly_admitted', 'is_from_survey',
       'old_first_name', 'old_last_name', 'old_middle_name'],
      dtype='object')
```

Figure 2. The list of all categories within the 'student_data' dataframe

From the filtered dataset, I observed the values for relevant categories. A pattern among the binary variables (e.g. dropout) was the designation of byte values rather than integers; this variation required conversion of the byte values to integers to ensure compatibility with data analysis algorithms. To perform this conversion, I implemented a Python function named 'convert_byte' that applied a lambda function to handle each value with Python's built-in 'int.from_bytes' method with the 'little' byte order. Beyond compatibility for EDA, this adjustment was essential for data consistency and insurance that the machine learning algorithms could accurately interpret and process these binary variables.

The first metric I wanted to understand from the dataset was the percentage of students who were dropouts, as their characteristics would be the basis for making the predictive models. I calculated the counts and percentages of students classified as dropouts versus those who were not. This data was summarized in a table and then visualized using a pie chart to provide a clear representation of the dropout distribution within the dataset (Figure 3).

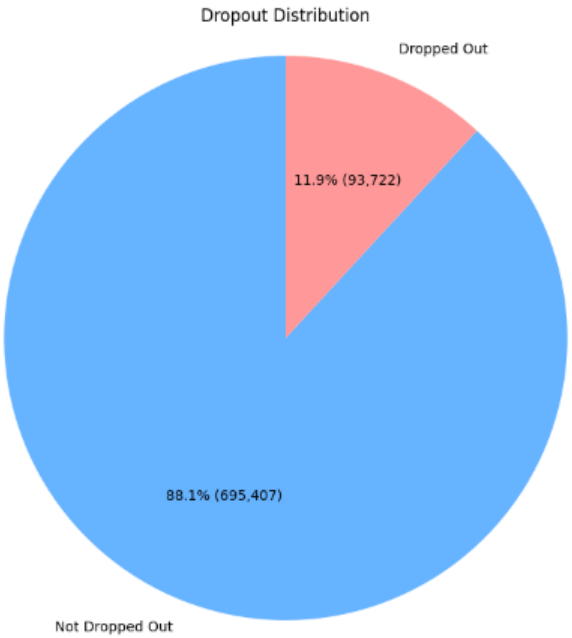

Figure 3. Pie chart that shows the distribution of dropouts and non-dropouts in 'student_data'

The pie chart shows that of the 789,129 students from the dataset, 11.9%, or 93,722 students, were current dropouts. This statistic represents a significant challenge for educational retention and highlights the benefit of a successful predictive model that can promote early interventions for at-risk students. By leveraging the characteristics common among the dropout population, educational programs can be tailored to address the underlying issues contributing to student disengagement and attrition. Based on this insight, I wanted to explore descriptive statistics for a significant predictor of student success, the household income. These included mean, median, standard deviation, skewness, and kurtosis, all of which provided a baseline comprehension of the income distribution. The output, shown in Figure 4, indicated that only 89,565 students, or 11.3%, reported a household income, and of those, the data was heavily skewed toward lower-income backgrounds.

```
Descriptive Statistics for Parents Income:
count     89565.000000
mean       1435.064573
std       11041.000919
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max      908000.000000
Name: parents_income, dtype: float64

Additional Statistics for Parents Income:
median    0.000000e+00
std       1.104100e+04
var       1.219037e+08
skew      4.622647e+01
kurt      3.009101e+03
Name: parents_income, dtype: float64
```

Figure 4. Descriptive statistics for 'parents_income' category

The high standard deviation of 11,041 indicates significant variability among students based on household incomes, while the median value of 0 showed that most came from households that generated no income. Furthermore, I segmented the income data into various

income brackets, or bins, using pandas' 'cut' feature, which helped to examine the distribution across defined ranges.The bins spanned from low to high incomes to ensure an inclusive analysis of all economic backgrounds.

```
Frequency of Each Bin:
parents_income
(0, 50000]           354595
(50000, 100000]        1474
(100000, 200000]        613
(200000, 500000]        158
(500000, 1000000]        31
Name: count, dtype: int64

Frequency and Percentage of Each Bin:
                    Count  Percentage
parents_income
(0, 50000]         354595   99.362235
(50000, 100000]      1474    0.413034
(100000, 200000]      613    0.171771
(200000, 500000]      158    0.044274
(500000, 1000000]      31    0.008687

Frequency and Percentage of Each Small Bin within 0 to 50000:
                  Count  Percentage
parents_income
(0, 10000]       195190   55.045897
(10000, 20000]   141405   39.877889
(20000, 30000]    14603    4.118219
(30000, 40000]     2419    0.682187
(40000, 50000]      978    0.275808
```

Figure 5. Bin distribution and frequency for parents_income

As the output in Figure 5 shows, almost the entirety of the income data (99.36%) fell into the lowest income bracket of 0 to 50,000. To address this, I performed a more granular analysis within the range of 0 to 50,000 to gain deeper insights into the lower end of the income spectrum. The output indicates that more than half of the reported student incomes (55.04%) within the 0 to 50,000 range were between 0 and 10,000; the high presence of 0 values for household income, as indicated by the median of the entire dataset, contributed to this majority. About 39.87% of students in the 0 to 50,000 range were in the 10,000 to 20,000 range, showing

that the bulk of incomes in the dataset are concentrated closer to the poverty threshold. This profound level of economic strain is a major factor influencing the observed high dropout rates.

With a basic understanding of the student data table's structure and distribution of key predictors, I began investigating which machine learning model would best suit the prediction of dropouts from the filtered categories. Given the binary nature of the dropout indicator and the need to balance model interpretability with overall accuracy, I first considered logistic regression. This model provides a simple and powerful tool for binary classification problems and, unlike linear and polynomial regression, outputs probabilities for its outcomes, which in this case is a student dropping out. These probabilities are calculated through the logistic function, also known as the sigmoid, which outputs values between 0 and 1. (Figure 6) The sigmoid function combines the input features (e.g. household income) each multiplied by a coefficient that represents the weight or importance of that feature. These weights are adjusted during the training process to best fit the model to the data. The function then takes the results of this linear combination and transforms them into a probability score. For example, a higher probability close to 1 might indicate a higher likelihood of a student dropping out, while a score closer to 0 would indicate the opposite. This probabilistic output can be useful because it not only provides the binary prediction but also gives a measure of certainty about that prediction, which can be crucial for making informed interventions.

The ease with which logistic regression models can be interpreted would allow stakeholders to understand the contributing factors behind the predictions that the model generates. The coefficients of the model can be directly related to the influence of corresponding features; if the coefficient for attendance is negative, it suggests that higher attendance is associated is with a lower likelihood of dropping out, which intuitively makes sense.

Stakeholders in an educational setting can use these insights to know which factors are most predictive of dropouts and guide resources and interventions to students who are most at risk based on quantifiable metrics. Moreover, logistic regression does not require large computational resources, unlike more complex models, making it a practical choice especially when quick and clear insights are required.

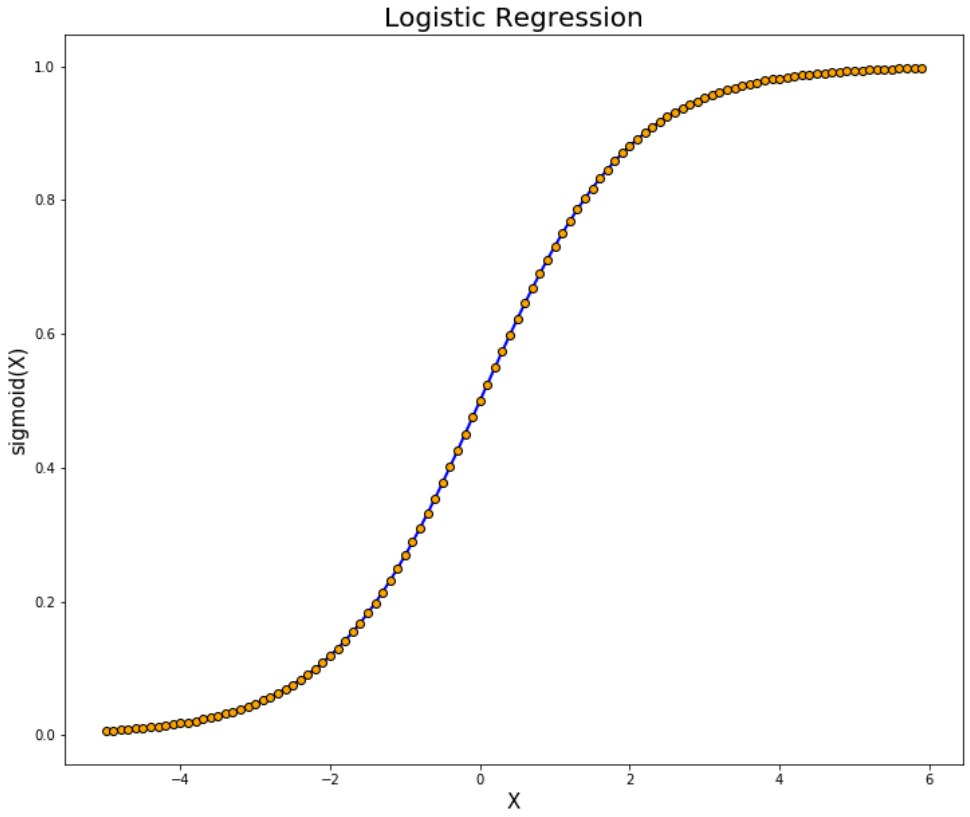

Figure 6. Graph of sigmoid function for Logistic Regression models

I next considered Random Forest due to its robustness and ability to manage overfitting, which can be a common issue with simpler models like logistic regression when dealing with complex or noisy data like the student dataset. Random Forest operates by constructing multiple decision trees during the training process and outputting the mode of the classes predicted by the individual trees. (Figure 7) This method is known as ensemble learning, where multiple models combine to improve the overall result. Random Forest's strength lies in its ability to handle a

large number of input features, as is the case with the number of predictor features for this project, and its capacity to identify the most influential features through a built-in feature importance metric. By examining the importance scores assigned to each feature, stakeholders can gain more nuanced insights into which factors most strongly predict dropout.

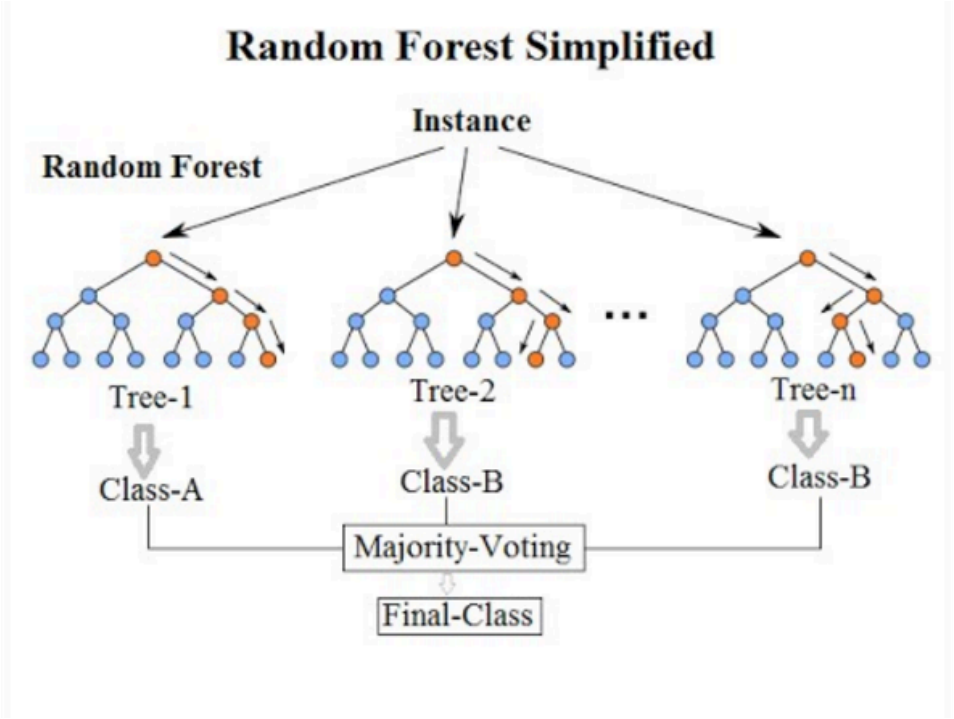

Figure 7. Diagram of the mechanism for Random Forest models

Furthermore, Random Forest provides a measure of uncertainty for each prediction, which can offer insights into model confidence that can be pivotal for applying practical interventions. Unlike logistic regression, which struggles with complex pattern recognition in high-dimensional data, Random Forest can detect subtle interactions and nonlinear relationships without explicit feature engineering. This aspect makes it powerful for exploring educational datasets where interactions between variables, such as socioeconomic status combined with academic performance, might significantly impact the likelihood of student dropouts. Random Forest also benefits from its non-parametric nature, meaning it does not make strong assumptions

about the form of the data distribution. This flexibility allows it to adapt more freely to the actual nuances present in the data, even in the presence of skewness like that of relevant categories within the student dataset like household income. Given these advantages, I chose Random Forest as the first model to implement, followed by a Logistic Regression analysis.

An important consideration of Random Forest when compared to Logistic Regression, however, is the additional computational resources required to make multiple decision trees from the large student dataset. This demand required me to consult cloud-based GPU's rather than relying on local computing resources to improve the processing speed and efficiency of the model training sessions. Fortunately, as a current student at Brown University, I had free access to Brown's OSCAR cluster which provided the necessary computational power to manage the dataset's size and complexity, ensuring swift and scalable model development.

Implementation

When building the initial predictive models, I was less interested in the probability of 'dropout' being true, so the Random Forest model was trained first to handle the high-dimensional data and any potential overfitting. The ensemble nature of Random Forest, which constructs multiple decision trees to arrive at a more stable and accurate prediction, made it suitable for the complex nature of the student dataset.

After separating the outcome variable, dropout, from the several predictor variables, data preprocessing involved handling missing values in the dataset (e.g. blank values for parents_income and father_educational_attainment). At first, I omitted any rows with missing values for any of the predictor variables, but this approach removed more than half of the recorded entries in the dataset; the omission left an insufficient dataset for training the Random Forest model. The outcome necessitated imputation for the missing numerical, ordinal, and categorical values. The ordinal values in the dataset (e.g. father_educational attainment) were already mapped to corresponding numerical rankings, so for both numerical and ordinal values, I applied the median of the respective columns. This method preserved the central tendency of the data without interference from outliers. To achieve this imputation, I applied a pipeline that included a SimpleImputer with a strategy of 'median', followed by scaling using StandardScaler to standardize the feature values, ensuring that all numeric inputs contribute equally to the prediction without any single attribute dominating due to its scale.

For categorical data (e.g. is_never_been_to_school), missing values were imputed with the most frequent category within each feature. This approach helped maintain the statistical distribution of each category within the dataset. These imputed categories were then transformed using OneHotEncoder, a method for converting each categorical column into a new binary

column. This step is crucial for machine learning models as it allows them to process categorical data by creating a distinct feature for each possible category value. For example, if a feature like 'is_never_been_to_school' has three categories, 'yes', no', and 'unknown', then OneHotEncoder creates three new columns, one for each category and mark them as 1 (i.e. true) or 0 (i.e. false). This encoding clarifies the presence or absence of a condition and ensures that the model interprets the absence of data as a separate category, preventing any misinterpretation of missing data as a 'no' or 'yes'. These preprocessing steps for numerical, ordinal, and categorical data ensured that they reflected the nuances of the educational contexts within the student dataset.

The next step for the Random Forest implementation was to configure the model to address the class imbalance present in the dataset, as dropout rates represented a small proportion, 11.9%, of the total entries (Figure 3). Therefore, I set the 'class_weights' parameter for the RandomForestClassifier to 'balanced'. This setting adjusted the weights inversely proportional to the class frequencies in the input data which allowed the model to give higher priority to the minority class (i.e. 'dropout') during training. To further enhance the model's ability to deal with the imbalanced dataset, I incorporated SMOTE (Synthetic Minority Over-Sampling Technique) into the data preprocessing pipeline. SMOTE works by creating synthetic samples from the minority class to balance the class distribution, which helps prevent the model from being biased toward the majority class. Thus, the full pipeline, shown in Figure 8 below, included the preprocessing steps for handling missing values and encoding, the application of SMOTE, and parameter tuning for the Random Forest model itself. The configuration of the Random Forest model was further specified with 100 trees (i.e. 'n_estimators=100') and a maximum depth of 100 (i.e. 'max_depth= 100') to allow the model to

sufficiently explore and build predictions from the dataset without overfitting and extensive computational resources.

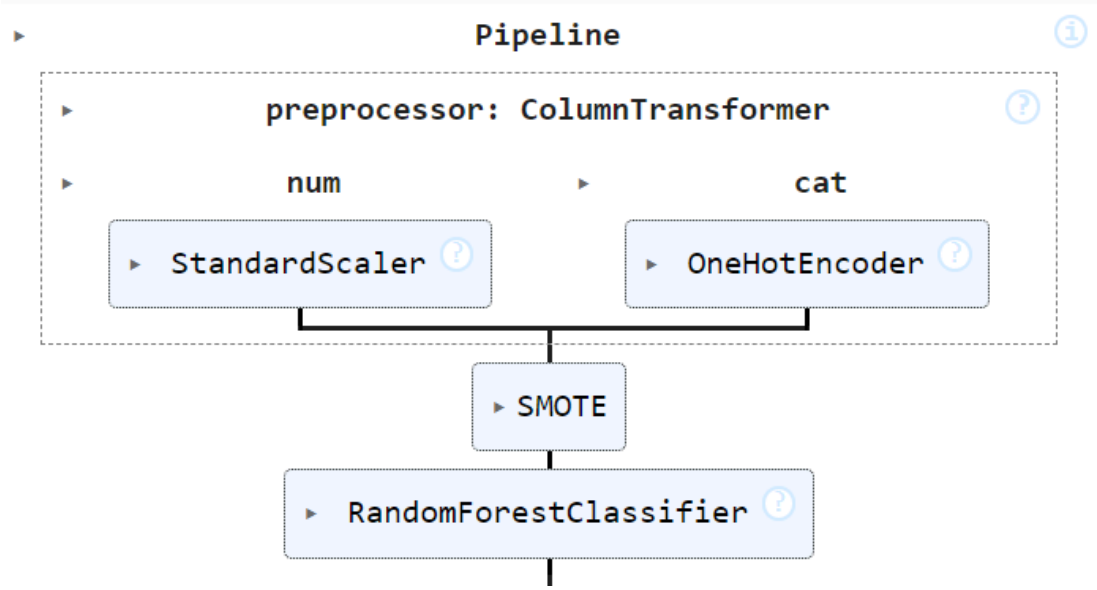

Figure 8. Data pipeline for Random Forest model

The next step involved splitting the data into training and test sets, with 20% of the data reserved for testing to evaluate the model's performance on unseen data (i.e. 'test_size=0.2'). This allocation was achieved with the 'train_test_split' function from the scikit-learn library. By assigning 20% of the data for testing, the setup ensured an unbiased evaluation of the Random Forest model's ability to generalize to new, unseen data. The model was then trained on the 'X_train' (predictor variables training) and y_train (outcome variable training) datasets, where it learned to recognize patterns that indicate a higher likelihood of dropout based on the several predictor variables in the filtered student dataset. Upon completion of the training phase, the model's performance was evaluated using the 'X_test' (predictor variables testing) and 'y_test' (outcome variable testing). This evaluation provided specific insights into the accuracy and reliability of the Random Forest model for predicting the outcome of 'dropout' based on the predictor variables like 'is_orphan', 'is_never_been_to_school', and 'marital_status'.

After implementation and evaluation of the Random Forest model, my attention shifted to creating the Logistic Regression model. Although I knew it would be less capable than the Random Forest at capturing nuances within the high-dimensional dataset due to its more simplistic approach, Logistic Regression provides the probability of the outcome variable and provides a more feasible foundation for interpretability by outputting coefficients that show the relationship between each predictor variable and the outcome. This model provides a simpler basis for performing predictions compared to Random Forest by estimating the probability of dropout as a function of a linear combination of input features (Figure 9).

$$\text{logit}(P(y = 1|X)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

Figure 9. Sigmoid function for Logistic Regression model

For overall model performance, Random Forest is still preferable due to its higher accuracy, feature importance outputs, relative lack of overfitting, and superior ability to handle missing data, but both approaches were combined in this project to provide a balanced perspective. Logistic regression provided a simpler means of understanding which variables influenced the probability of dropout with less focus on the specific interactions between them, giving educational stakeholders straightforward insights that could be acted upon quicker.

The implementation began with similar preprocessing steps to those used for Random Forest, including imputation and encoding to prepare the data for the Logistic Regression model. Numerical data were also imputed with the median to maintain distribution, and categorical data with the most frequent entries. These variables were then one-hot encoded, translating categories into binary columns to enable Logistic Regression processing. With the preprocessing pipeline established (Figure 10), the Logistic Regression model was configured with Elastic Net regularization, which combines the properties of both lasso (L1) and ridge (L2) penalties. This

approach helps in managing multicollinearity and feature selection by shrinking some coefficients to a potential value of zero, like in lasso, and others more smoothly, like in ridge. The 'l1_ratio' was set to 0.5 to balance between the two regularization types, which provided flexibility in how the model penalizes the complexity. The use of the 'saga' solver facilitated the efficient handling of the large dataset with potentially many sparse values from the one-hot encoding.

The model was then trained using the 'X_train' and 'y_train' datasets. Training a logistic model tends to be faster than a Random Forest due to its simpler calculations and fewer parameters to tune. After training, the model was evaluated using the 'X_test' and 'y_test' datasets. The Logistic Regression's performance metrics (i.e. accuracy, precision, recall, and F1 sore) were calculated to assess its effectiveness in predicting the student dropouts. While the Logistic Regression model may not have captured the complex interactions between features as effectively as the Random Forest, its output was helpful for its clarity and ease of interpretation. The coefficients provided by the model offered direct insights into how each feature influenced the likelihood of a student dropping out, which is particularly useful for stakeholders who need to understand the impact of specific variables on educational outcomes.

The juxtaposition of the two models, Random Forest and Logistic Regression, highlighted their respective strengths and weaknesses. The combined analysis provided the depth of insight provided by the Random Forest with the interpretability of the Logistic Regression model. This approach ensured that decision-makers could choose from a range of interventions based on robust data-driven insights, catering to both immediate needs and more strategic long-term planning. The successful integration of both models into the educational strategy framework demonstrated the practical utility of machine learning in educational settings and

reinforced the importance of using a variety of analytical techniques to address diverse and complex challenges in student retention and education management.

Results

For the Random Forest model, several metrics were incorporated to evaluate its effectiveness at predicting the outcome of 'dropout' from the selected predictor variables. A classification report incorporated all of these statistics (i.e. the accuracy, precision, recall, f1-score, confusion matrix, and ROC AUC score) into the model's output.

```
Accuracy: 0.7025458416230532
Classification Report:
              precision    recall  f1-score

           0       0.96      0.69      0.80
           1       0.26      0.80      0.39

    accuracy                           0.70
   macro avg       0.61      0.75      0.60
weighted avg       0.88      0.70      0.75

Confusion Matrix:
 [[95924 43245]
 [ 3701 14956]]
ROC AUC Score: 0.7454460551148095
```

Figure 11. Accuracy, classification report, confusion matrix, and ROC AUC score for Random Forest model

- Accuracy (0.7025): This metric indicates that the model correctly predicted the dropout status for about 70% of the students in the testing set. While this is a relatively decent accuracy score, evaluation of accuracy on its own can be misleading because it does not account for the distribution of predictions across all classes.
- Precision and Recall
    - For class 0 (non-dropout), the precision is high at 0.96, which indicates that the model is very reliable when it predicts that a student will not drop

out. The recall value of 0.69 suggests that it misses about 31% of the actual non-dropout cases.

    - For class 1 (dropout), the precision is significantly lower at 0.26, which implies that only about 26% of the the predictions made by the model for dropout are correct. The high recall of 0.8 for dropouts indicates that the model is able to identify most of the true dropout cases, but it does so at the expense of many false positives.

- F1-Score: This metric helps balance the trade-off between precision and recall and is particularly useful when the cost of false positives and false negatives varies. For non-dropouts, the F1-score is 0.8, which is relatively high. For dropouts, however, the F1-score is 0.39, highlighting the model's difficulty in accurately predicting dropouts without generating many false alarms.
- Confusion Matrix
    - True Negative (TN): 59,924 students were correctly identified as non-dropouts.
    - False Positive (FP): 43,245 students were incorrectly identified as dropouts.
    - False Negative (FN): 3,701 students who did dropout were not identified by the model.
    - True Positives (TP): 14,956 students were correctly identified as dropouts.
- ROC AUC Score (0.75): This score, which measures the area under the ROC curve, is a performance measurement for the classification at various threshold settings. This particular score indicates a good ability to distinguish between

dropout and non-dropout students, but there is room for improvement, specifically in how the model handles class separation thresholds.

These results suggest that while the Random Forest model is proficient in identifying true non-dropout cases, its predictive performance for actual dropouts, although strong in recall (0.8), is hampered by a large number of false positives, as evidenced by the low precision for dropouts (0.26). Future model tuning will focus on strategies to improve the precision while not compromising the high recall value, such as adjusting thresholds or exploring different feature engineering strategies to better understand the underlying patterns associated with dropouts. By addressing these areas, the effectiveness of the dropout prediction model can be improved, leading to more accurate and actionable insights for educational interventions designed to reduce dropout rates.

```
Accuracy: 0.8831941505201931
Classification Report:
              precision    recall  f1-score   support

           0       0.89      1.00      0.94    139169
           1       0.60      0.04      0.07     18657

    accuracy                           0.88    157826
   macro avg       0.74      0.52      0.50    157826
weighted avg       0.85      0.88      0.83    157826

Confusion Matrix:
 [[138728    441]
 [ 17994    663]]
```

Figure 11. Accuracy, classification report, and confusion matrix for Logistic Regression model

Surprisingly, despite its lack of ability to capture subtle nuances between the several predictor variables, the Logistic Regression model performed better on the student dataset for predicting dropouts. This capability makes it a potentially valuable tool for educational settings,

where quick and accurate predictions can lead to impactful interventions, alongside the interpretability of which features from the predictor variables contribute to the predictions that are made. The performance metrics in Figure 12 illustrate the superior performance relative to Random Forest, based on the corresponding accuracy, precision, recall, f1-score, and confusion matrix.

- Accuracy (0.8831): This metric indicates that the model correctly predicted the dropout status for about 88.3% of the students in the testing set. This higher accuracy, compared to the Random Forest model, suggests that Logistic Regression was more effective at general classification for the dataset.
- Precision and Recall
    - For class 0 (non-dropout), the precision score of 0.89 indicates that when the model predicts a student will not dropout, it is correct 89% of the time. The recall value of 1.00 suggests that the model identifies 100% of the non-dropout cases correctly, which is optimal for ensuring that students who are not at risk are not mistakenly targeted by interventions.
    - For class 1 (dropout), the precision score of 0.60 indicates that when the model predicts dropout, it is correct 60% of the time. This also implies the presence of a number of false positives, which is room for improvement. The model's recall score of 0.84 suggests that the model successfully identifies 84% of the actual dropout cases, which demonstrates strong sensitivity compared to the Random Forest model and confirms its utility in identifying at-risk students based on the predictor variables.

- F1-Score
    - For non-dropouts (0.94), the high score indicates a strong balance between precision and recall.
    - For dropouts (0.7), while lower than for non-dropouts, this score reflects a reasonable balance between precision and recall for dropout predictions, indicating the model's utility in identifying true dropout cases despite some false positives.
- Confusion Matrix
    - True Negatives (TN): 138,728 students were correctly identified as not dropping out.
    - False Positives (FP): 441 students were incorrectly identified as dropouts.
    - False Negatives (FN): 17,994 students were dropouts that the model failed to identify.
    - True Positives (TP): 6,631 students were correctly identified as dropouts.

These results indicate that the Logistic Regression model, although simpler, provides a strong basis for predicting student dropouts with considerable accuracy and recall. The model's ability to capture a significant portion of true dropout cases (recall) while maintaining a reasonable precision suggests that it can be highly effective for educational interventions aimed at reducing dropout rates. Moreover, the high recall ensures that most at-risk students are identified, which is crucial for implementing preventative measures. Future model tuning for the Logistic Regression model will focus on strategies to improve the precision score without a significant sacrifice for recall, similar to the goal for the Random Forest model's projected enhancements. This process will involve refining the feature selection process, adjusting the

regularization parameters, and experimenting with different threshold settings for classification decisions. The clear interpretability of the Logistic Regression model also ensures that educational stakeholders can understand and utilize the insights derived from the model, which can support informed decision-making.

To contribute to the overall interpretability of the Logistic Regression model, I also performed a feature importance analysis which allowed me to understand which variables have the most effect on the predicted value for 'dropout'. Due to the high-dimensional nature of the dataset, it is impossible to paste the full output within this report, but notable values were extracted to understand which categories had the most impact on the outcome variable. Parents' income output a significantly negative coefficient of -1.295, indicating that higher income is strongly associated with lower dropout rates. This was shown to be the most impactful continuous variable. The educational attainment, particularly that of the father, also plays a crucial role in determining the likelihood of a dropout. The 'father_educational_attainment_11' variable has a positive coefficient of 0.9285, indicating that a father with at least an educational attainment until grade 11 correlates with a higher likelihood of continuing education for their children. On the other hand, 'father_educational_attainment_8' has a high negative coefficient of -1.0599, showing a strong association with increased dropout rates if the father completed grade 8 but did not progress beyond this stage. Furthermore, the 'is_ethnic' variable has a notable negative coefficient of 0.242, which indicates that ethnic minorities may have higher dropout rates. This effect is intuitive due to the feeling of exclusion that foreign students may experience within the classroom setting based on factors like discrimination and inadequate accessibility. The values for these features suggest that socio-economic factors, such as parents' education and income, are pivotal in influencing dropout rates.

Discussion

The Random Forest model showed a strong ability to identify non-dropout cases with high accuracy, as reflected by the precision and recall values (Figure 10). However, its performance in predicting actual dropouts was less satisfactory due to low precision, despite high recall. This disparity indicates that while the model is effective in recognizing most true dropout cases, it also misclassifies a substantial number of non-dropout students as at-risk. This high false positive rate can lead to inefficient allocation of resources, where interventions are unnecessarily directed at students not truly at risk.

In contrast, the Logistic Regression model outperformed the Random Forest in overall accuracy and demonstrated a better balance between precision and recall for dropout predictions. This model's higher precision for dropout predictions means fewer resources wasted on false positives, making it more suitable for targeted interventions. The model's coefficients also provide clear insights into the influence of various predictors on dropout likelihood, which enhances the interpretability crucial for strategic decision-making by educational administrators. Thus, despite initial predictions that the Random Forest would outperform the Logistic Regression analysis, Logistic Regression appears more advantageous for practical application within BEP due to its higher accuracy and balanced metrics. However, the choice between models should consider the specific strategic goals of dropout prevention initiatives. For instance, if the utmost priority is to ensure no at-risk student is overlooked, the higher recall of the Random Forest model might be preferred despite its lower precision.

Both models can benefit from further refinement to enhance their predictive accuracy and utility:

- Feature engineering: The incorporation of more granular data about student behavior and academic performance might reveal deeper insights into dropout triggers, thus improving the models' overall accuracy scores.
- Model tuning: Adjusting classification thresholds and exploring more sophisticated ensemble techniques could reduce false positives in the Random Forest model and enhance the Logistic Regression model's sensitivity.
- Real-time data integration: The implementation of a feedback system that allows for real-time data updates from educational staff within BEP can help in dynamically adjusting the models to better reflect current student conditions.
- User application: The ultimate goal of this project is to incorporate the predictive models into an intuitive user application that handles inputs for various student metrics. The output would be based on whether the student is classified as a future dropout and the probability of occurrence.

The findings from this project highlight the potential of machine learning to significantly impact educational strategies aimed at reducing dropout rates. While both models provide valuable predictions, their effective deployment depends on continuous improvement and a clear understanding of their operational limitations. For example, a student may drop out due to a host of reasons that are impossible to capture via the metrics that the Evaluation and Monitoring team evaluates for each student within the vast network of primary schools. Future strategies should, therefore, focus on leveraging the strengths of predictive analytics to optimize intervention efforts in a hands-on capacity (e.g. the user application for educators) which will ensure that resources are efficiently allocated to students who are most in need.

Conclusion

My summer internship at the BRAC Education Programme (BEP) leveraged the increased availability of educational data to tackle significant challenges in student retention through the application of targeted machine learning analyses. The project specifically aimed to predict student dropouts by employing logistic regression and random forest models based on a broad spectrum of academic, demographic, and behavioral variables. The project's objective was to enhance early identification of at-risk students to improve educational outcomes at the primary school level, a crucial period of learning for adolescents in developing countries. The implementation of machine learning models showed the potential of data-driven approaches to foster substantial social change within the educational sector. These models, assessed on accuracy, precision, recall, and F1 scores, demonstrated substantial capability to predict dropouts. While the Random Forest model achieved an accuracy of around 70%, the surprising performance of the Logistic Regression analysis with an accuracy close to 90% showed the feasibility of integrating these tools into BEP's intervention strategies.

Globally, education systems face the daunting challenge of dropout rates that compromise the quality and effectiveness of educational environments. This challenge is particularly pronounced in Bangladesh, where despite high enrollment rates, sustaining student engagement through completion remains a persistent hurdle. Initiatives like the Non-Formal Primary Education (NFPE) and the Bridge Programme, aimed at children at risk of dropping out, have been crucial in addressing these challenges. The Bridge Programme, notably, aligns with the United Nations Sustainable Development Goal 4 (SDG-4), which advocates for inclusive, equitable education that promotes lifelong learning opportunities for all. The outcomes of this project suggest that machine learning can further help to complement BEP's goals in maintaining

alignment with SDG-4 by improving overall retention and resource allocation for students within the Bridge Programme and other educational initiatives.

Overall, the integration of logistic regression and random forest models into BEP's strategy demonstrated the practical utility of machine learning in educational settings and reinforced the importance of using a variety of analytical techniques to address complex challenges in student retention and educational management. The insights can contribute to a broader understanding of how data-driven strategies can be effectively implemented to improve student retention and support the achievement of BEP's broader educational goals.

References

Christenson, Sandra L., and Martha L. Thurlow. “School Dropouts: Prevention considerations, interventions, and challenges.” *Current Directions in Psychological Science*, vol. 13, no. 1, Feb. 2004, pp. 36–39, https://doi.org/10.1111/j.0963-7214.2004.01301010.x.

Farooq, Muhammad Shahid. “An inclusive schooling model for the prevention of dropout in primary schools in pakistan.” *Bulletin of Education and Research*, vol. 35, no. 1, June 2013, pp. 47–74

Federico, Batini, et al. “I feel good at school! reducing school discomfort levels through integrated interventions.” *ATHENS JOURNAL OF EDUCATION*, vol. 6, no. 3, 22 May 2019, pp. 209–222, https://doi.org/10.30958/aje.6-3-3.

Graeff-Martins, Ana Soledade, et al. “A package of interventions to reduce school dropout in public schools in a developing country.” *European Child & Adolescent Psychiatry*, vol. 15, no. 8, 6 June 2006, pp. 442–449, https://doi.org/10.1007/s00787-006-0555-2.

Lewin, Keith M., and Angela W. Little. “Access to education revisited: Equity, drop out and transitions to secondary school in South Asia and Sub-Saharan africa.” *International Journal of Educational Development*, vol. 31, no. 4, May 2011, pp. 333–337, https://doi.org/10.1016/j.ijedudev.2011.01.011.

Mishra, Pratibha, and Abdul Azeez. “Family etiology of school dropouts: A psychosocial study.” *International Journal of Language & Linguistics*, vol. 1, no. 1, June 2014, pp. 45–50, https://doi.org/https://ijllnet.com/journals/Vol_1_No_1_June_2014/6.pdf.

Remschmidt, H., et al. “Forty-two-years later: The outcome of childhood-onset schizophrenia.” *Journal of Neural Transmission*, vol. 114, no. 4, 10 Aug. 2006, pp. 505–512, https://doi.org/10.1007/s00702-006-0553-z.

Sabates, Ricardo, et al. “School drop out in Bangladesh: Insights Using Panel Data.” *International Journal of Educational Development*, vol. 33, no. 3, May 2013, pp. 225–232, https://doi.org/10.1016/j.ijedudev.2012.09.007.

Townsend, Loraine, et al. “The relationship between bullying behaviours and high school dropout in Cape Town, South Africa.” *South African Journal of Psychology*, vol. 38, no. 1, Apr. 2008, pp. 21–32, https://doi.org/10.1177/008124630803800102.

Wang, Huan, et al. “Can social–emotional learning reduce school dropout in developing countries?” *Journal of Policy Analysis and Management*, vol. 35, no. 4, 16 May 2016, pp. 818–847, https://doi.org/10.1002/pam.21915.

World Bank. (2021). Learning Poverty Index. Accessed 20240814 at https://databank.worldbank.org/ id/c755d342?Code=SE.LPV.PRIM&=&report_ name=EdStats_Indicators_Report&populartype=series

Zuilkowski, Stephanie Simmons, et al. “‘I failed, no matter how hard I tried’: A mixed-methods study of the role of achievement in primary school dropout in rural Kenya.” *International Journal of Educational Development*, vol. 50, Sept. 2016, pp. 100–107, https://doi.org/10.1016/j.ijedudev.2016.07.002.