1. Introduction

As graduate students at Georgia Tech, we found a shared interest in wanting to explore factors that are associated with success in educational environments. As students ourselves, we are well versed in the benefits one can receive from education and we are also aware of some pitfalls that can result in performing poorly or dropping out.

Initially we proposed examining a dataset of secondary school students in Portugal, looking at attrition rates within a dataset of 649 students in relation to 33 attributes/features. Upon further reflection we realized better results would be obtained from examining data from a larger pool of students, and we focused on a dataset of Canadian students taking the OECD's PISA test, an international standardized test and questionnaire.

Our aim is to analyze the correlation between student performance and the multiple demographic and socio-economic factors in the student's background and approach to education that may influence their performance. By determining which factors have a significant impact, we hope to identify significant factors leading to failing students and that impact performance to help provide solutions to improve educational results.

2. Related Research

Research was conducted in 2017 to evaluate factors contributing to poor academic performance of students in higher education in New Zealand. The article published in the Journal of Pedagogical Research regressed various factors such as gender, level of study, age, working status etc. against GPA and found that the average student attendance, if the student had any dependents and the field of study had significant relationship with the GPA.

A similar study was conducted in 2011 to evaluate the socio-economic and demographic factors contributing to 9th-grade exam performance for 600 students across 12 metropolitan students in Pakistan. The study in the Journal of Quality and Technology Management regressed multiple factors such as gender, parent's education, and performance in 9th grade against the student's 9th-grade final performance and found that if a parent has higher education statically significant in predicting a student's performance. The study also found that females tend to outperform their male peers in Math, English, and overall grades.

3. Dataset

The OECD administers the PISA standardized test and questionnaire to students and schools across many countries and collects and tabulates results in the subjects of math, reading and science. The PISA database contains responses from individual students, school principals and parents spanning different years. Students are asked to provide information about their parent's education qualification, occupation, and home dynamics. Student details at school such as the number of periods each student spent in various subjects, the class population and attendance data are also obtained. The questionnaire tries to capture the student's interest in reading books, extracurricular activities, time they spend with friends and their sentiments about school and teachers.

For our research we chose to work with the year 2000 math dataset and student questionnaire data which contains 127,388 observations and 162 features. We further isolated this to Canadian students (16077 students) due to both lack of computational resources and to eliminate between-country variation

The main aim of this analysis is to identify factors that affect students' performance. The feature of interest or the dependent variable for the analysis is the student's subject grade, which is a categorical variable which contains values - "at mark," "above" or "below". We reclassified these categories as "failing" vs "passing" to implement binomial classification models to identify the factors that affect student performance.

4. Approach and Methodology

4.1. Data Cleaning

We encountered some data quality problems. Many columns in our dataset had missing values. When dealing with missing data, there are three approaches: deletion, imputation, or categorical approach (qualitative variables only). For this dataset, we use a combination of deletion and missing value labels. 2.5% of students had no value for math performance, our dependent variable; data for these students was removed entirely as it represented a small percentage of our dataset. The predictor variables in our dataset consist of mostly categorical variables. The categorical approach allows us to add missing values as 0, so we used this categorical approach for missing data in the predictors.

4.2. Exploratory Data Analysis

Before we begin modeling, it is helpful to visualize the data to see if there are any potential pattems or trends in the data. Although correlation does not indicate causation, plotting things out can help us draw insights from the data. Visualizations were created using PowerBI.

Figure 1: Distribution of Math Class Performance

In our models, the students' performance in math class will be the dependent variable. Therefore, it is helpful to look at the distribution of math performance in the dataset. From this, we can see that ~84.98% of the students passed their math class whereas ~15.02% of students failed. Since there is such a big disparity in proportion, it may be helpful to address class imbalance when modeling.

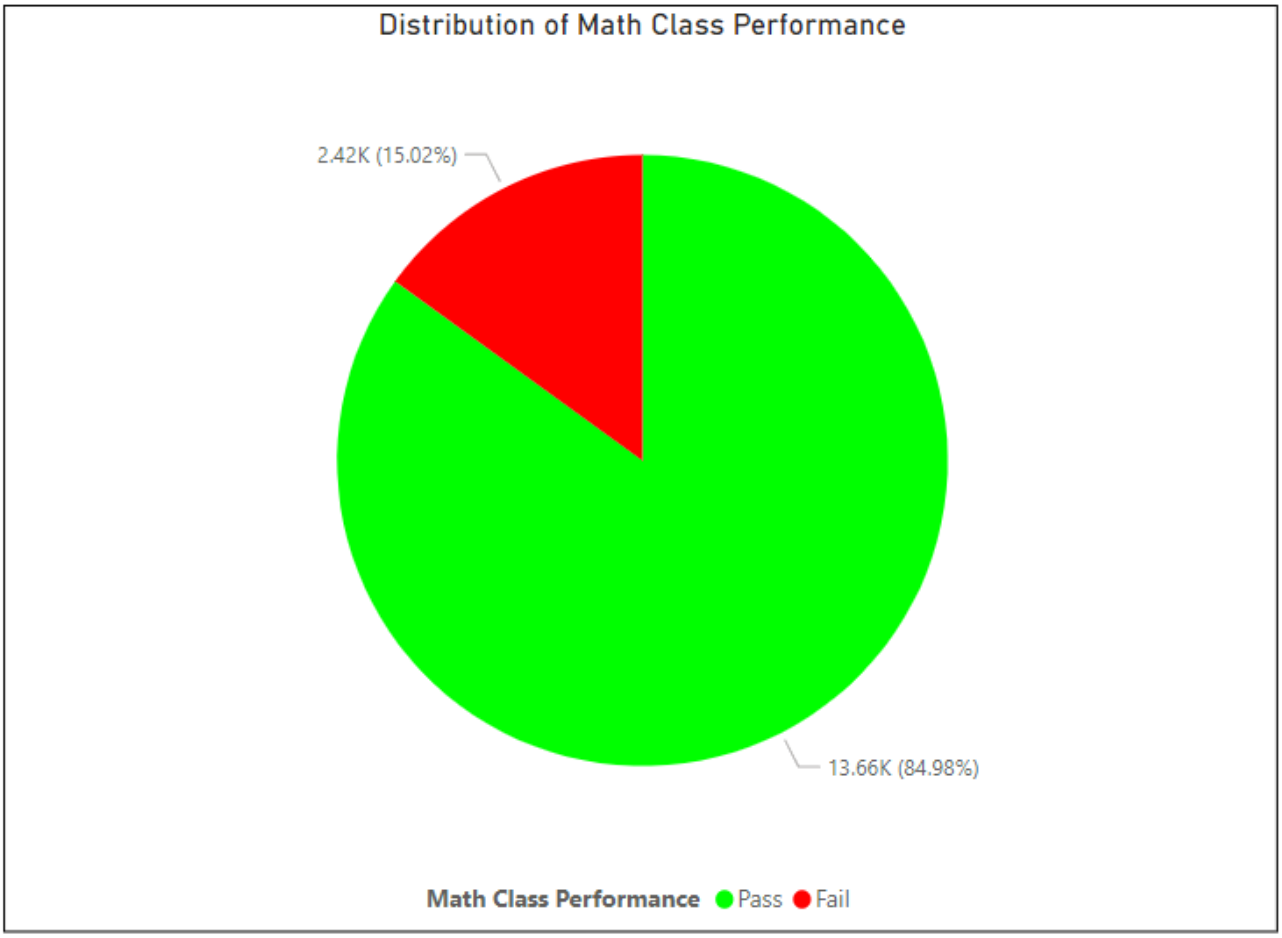


Figure 2: Math Class Performance Distribution by Private Tutoring:

One of the independent variables we are interested in is how often the student went to private tutoring. Typically, kids who need to go to private tutoring are either struggling in class or supplement / enhance their education. Thus, we want to see if there is any correlation between private tutoring and class performance. Based on the graph, it appears that private tutoring has some correlation with class performance. The distribution of failing class performance in student groups that attend private tutoring is higher than the distribution of failing class performance in student groups that never attended private tutoring. However, failing students remain a minority in each breakdown.

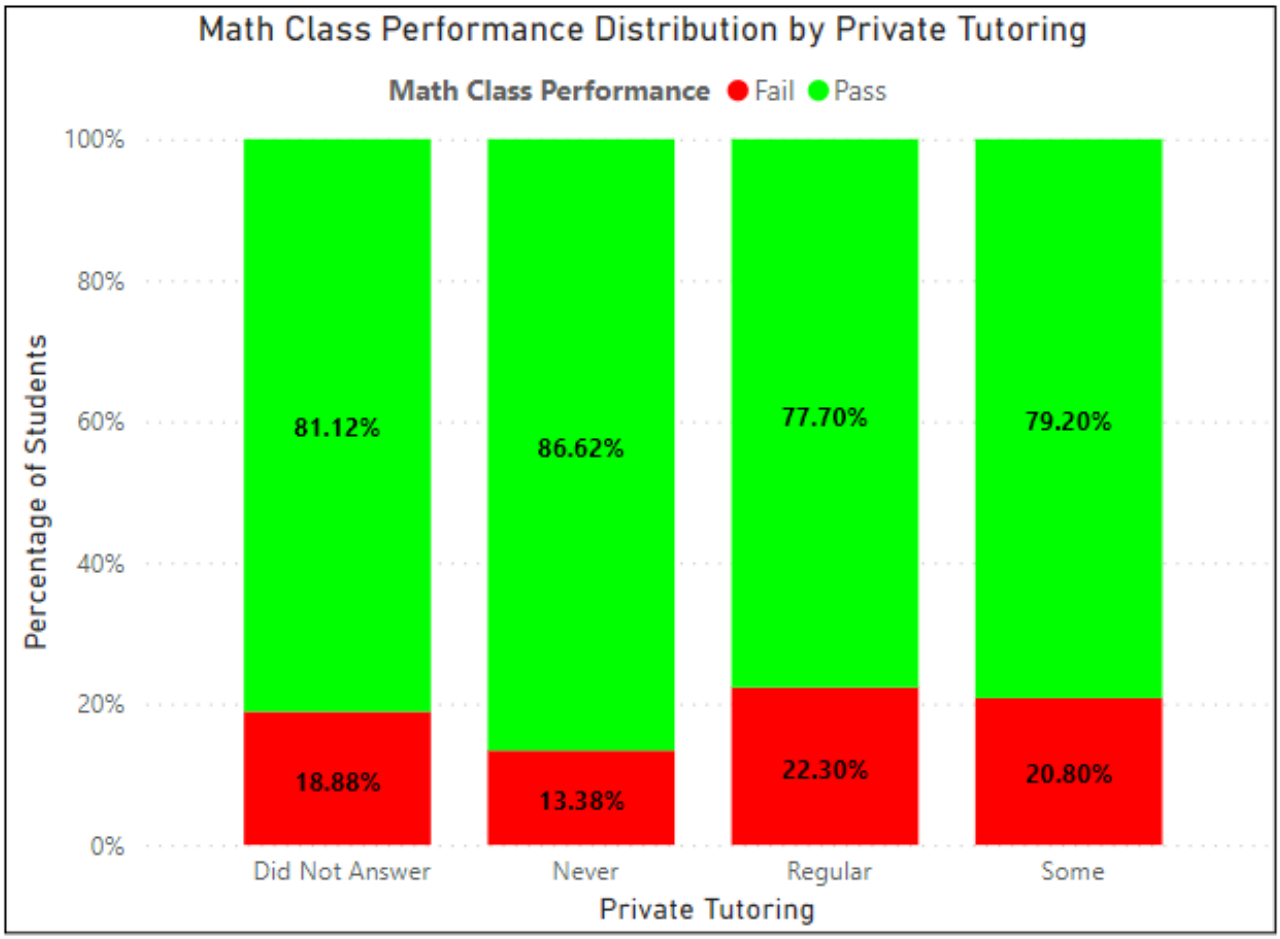


Figure 3: Math Class Performance Distribution by Number of Computers Owned

The number of computers owned is another interesting feature. Nowadays, computers play a very important role in education. Computers provide access to a huge variety of information and tools that can be used to aid education. Therefore, it is useful to plot this out to see if there is any correlation with class performance. Based on the graph, student groups with computers have a similar distribution of failing students. However, the student group with no computers have a slightly higher distribution of failing students. Owning a computer seems to correlate with better class performance.

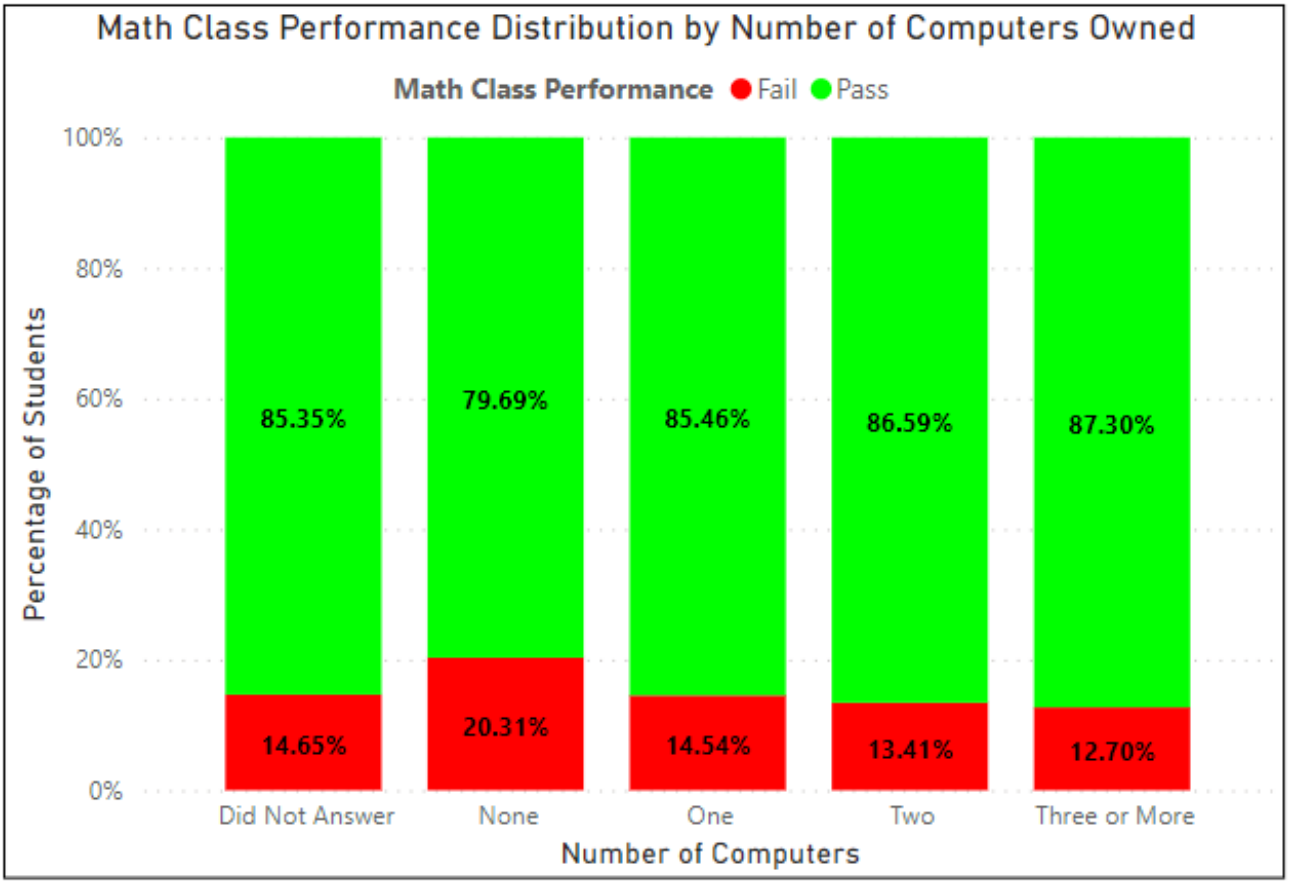
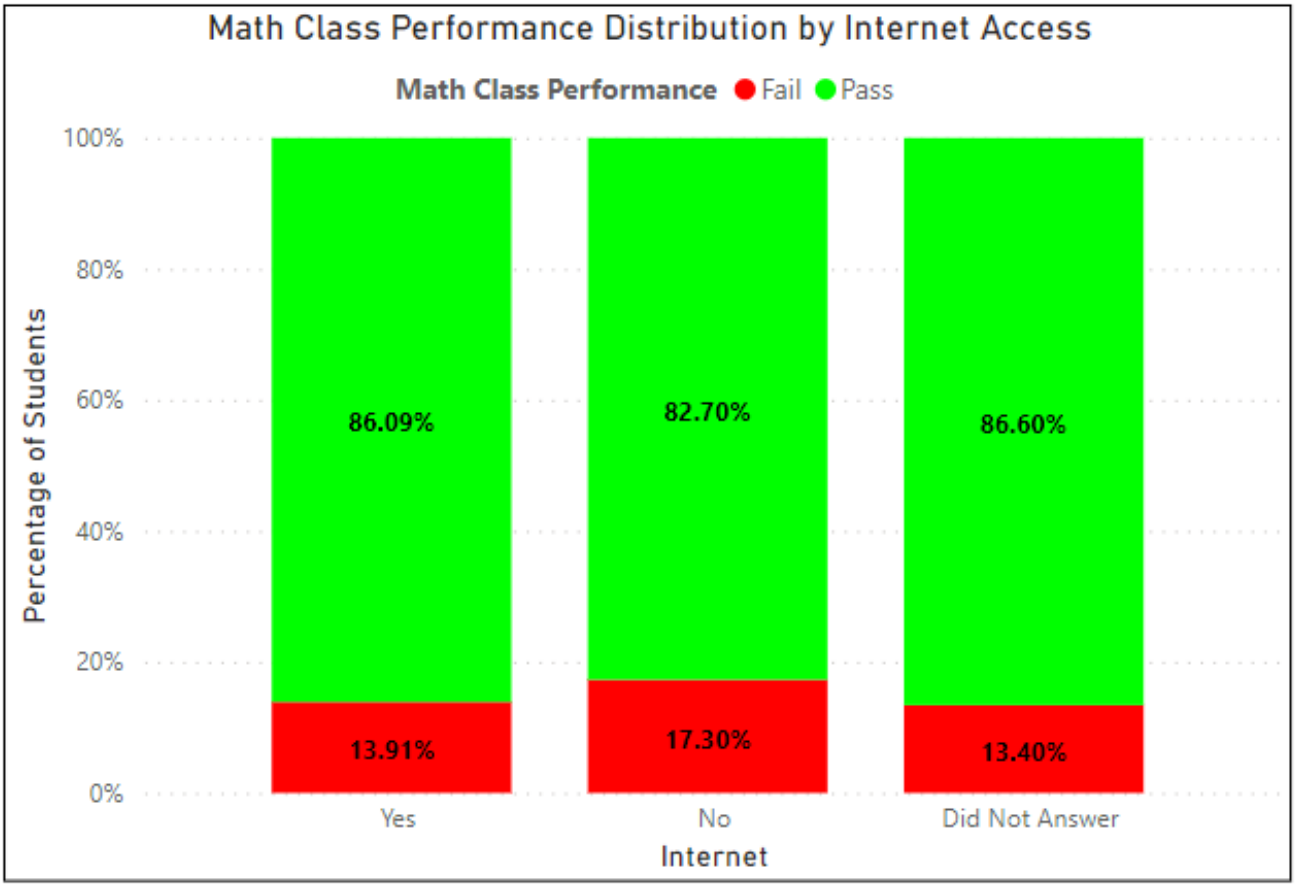


Figure 4: Math Class Performance Distribution by Internet Access:

Another independent variable we are interested in is internet access. The internet plays an important role in education by facilitating the sharing of information and communication. A student that is struggling in class may access free online lessons on sites like Khan Academy. If a student wishes to dive deeper into a subject, there are tons of free resources available. Based on the graph, the proportion of failing students without internet is higher than the proportion of failing students with internet. It seems that not having internet access correlates with worse class performance.



4.3. Modeling

4.3.1. Validation

Data consists of two types of patterns, real effects, and random effects. Real effects are actual relationships between the predictors and the response variables. These relationships are present even when the dataset is changed to a new one. Unlike real effects, random effects have a relationship, but may not be present in a different dataset. When we train models, fitting matches both real and random effects. Validation can help ensure that the output is accurate and accurately representative of the situation.

For this project, we will be assessing and comparing multiple models. Therefore, training, validation, and test set are needed for the validation process. These sets are created by randomly splitting the original dataset. Training sets are used to fit the models. Validation sets are used to pick the best model. Test sets are used to estimate the quality and performance. However, this method may introduce bias due to

randomness. Points in certain sets can occur too early or too late. Important data may only appear in one set, but not the other.

To address that concern, cross validation may be combined with the previous approach. K-fold cross validation will be used to train and pick the best model. We only need to split the data randomly into training and test sets. The training dataset will be split into k parts. For each of the k parts, it will select 1 part as the validation data and the remaining parts as the training data. It trains the model using the training set and picks the best model based on its performance on the validation set. This will ensure that important data does not just appear in a single set. After K-fold cross validation picks the best model, we can then estimate the quality using the test set.

4.3.2. Feature Selection and Models

According to the Law of Parsimony, the simplest explanation of an event or observation is preferred. When building a model in real life, it is rare that all the predictors in the dataset are useful. Therefore, simpler models are often better than complex ones. Models with many factors might fit too closely to the random effects, resulting in overfitting. By limiting the number of factors in our models, the model becomes easier to interpret, less likely to include insignificant factors, and reduce the amount of required data and computational power. For our analysis, we will consider the following approaches:

Feature Selection Methods:

Boruta:

- Feature selection algorithm based on random forest algorithm
- Creates shadow attributes and shuffle the values to create randomness in the dataset. Then, it creates a classifier with original and shadow attributes and assesses the importance.
- Heuristic

Lasso Regression:

- Regression analysis method that performs variable selection and regularization
- Adds constraint to the standard regression equation and shrinks the coefficient estimates towards zero when lambda is large enough.
- By shrinking the coefficient, prediction accuracy may improve, and variance may decrease.
- As penalty increase, the number of important variables the model selects will decrease
- Very selective

Elastic Net:

- Regression analysis method that linearly combines the L1 and L2 penalties of Ridge and Lasso Regression
- Finds the ridge regression coefficients and then shrinks the coefficients in a similar fashion to Lasso Regression.
- Have variable selection benefits of lasso and predictive benefits of Ridge Regression
- Underestimates coefficients of predictive variables and arbitrarily rules out some correlated variables
- Moderately selective

Domain Knowledge

Models:

Logistic Regression

- Used to model the probability of a certain class or event occurring
- Does not handle multicollinearity very well. However, this can be addressed by using lasso or elastic net regularization

Random Forest

- Ensemble learning method that uses the multitude of decision trees. For classification problems, it uses the most predicted response.
- Selects random subset of predictors each time model splits a tree to help decorrelate trees
- Using a multitude of different trees neutralizes overfitting, but makes it harder to interpret and explain results
- Can be used to compute feature importance

4.3.3. SMOTE: Synthetic Minority Oversampling Technique

SMOTE is an oversampling technique where synthetic samples are generated for the minority class by using bootstrapping and k-nearest neighbor. This algorithm helps neutralize the overfitting problem posed by random oversampling and class imbalance. Most variable selection methods assume that the samples are independent. Therefore, feature selection before smote is used.

In our training dataset, only 18% of students have below average marks in their math class. While this may be representative of the distribution in real life, the imbalance of positive cases may reduce the generalization capability of the model. The models may just classify most cases as negative if there are not enough positive cases to discover a pattern. To account for this, we will use two different versions of our models. One version will use smote to create a balanced training dataset, and the other will keep its original class distribution. The testing dataset consists of 17% of positive cases and will remain unchanged. We want the testing dataset to be representative of the distribution in real life. Typically, only a minority of students fail their classes.

4.3.4. Technology

To clean the data and perform the analysis, we will be using R, a programming language for statistical computing. The main R packages we will use are dplyr for data manipulation, themis for SMOTE, Boruta for feature selection, and caret for fitting models.

5. Results

Model	Feature Selection Method	Combined Methods	Accuracy	Specificity	Sensitivity	Number of Significant Features
Logistic Regression	Boruta	N/A	~86.63%	~95.66%	~33.43%	45
Logistic Regression	Boruta	SMOTE	~78.03%	~80.55%	~63.14%	163
Logistic Regression	Lasso	N/A	~87.54%	~97.48%	~29.00%	42
Logistic Regression	Lasso	SMOTE	~79.00%	~80.94%	~67.57%	27
Logistic Regression	Elastic Net	N/A	~87.62%	~97.89%	~27.14%	68
Logistic Regression	Elastic Net	SMOTE	~79.21%	~81.33%	~66.71%	48
Random Forest	N/A	N/A	~86.65%	~98.40%	~17.43%	20
Random Forest	N/A	SMOTE	~85.99%	~96.78%	~22.43%	20

*Number of significant features may be greater than the number of columns in the dataset due to dummy variables created by the models

Sensitivity measures how well a model can predict positive cases whereas specificity measures how well a model can predict negative cases. Sensitivity and specificity are inversely proportional. As one increases, the other decreases. The first thing that we noticed is that models trained without SMOTE are terrible at predicting positive cases (below average marks in math class) as indicated by the low sensitivity rate. Models created with SMOTE are trained using a balanced dataset. This means that there is a balanced proportion of positive and negative cases. After addressing the class imbalance in the dataset, the sensitivity rate increased at the cost of specificity.

Based on the results, the random forest model doesn't seem to be as good as the logistic regression model at predicting positive cases. The random forest models have the lowest sensitivity rate out of all the models. Another interesting thing to note is that the sensitivity rate of the random forest model has gone up very little when combined with SMOTE whereas the logistic regression models have seen a greater increase when SMOTE was used. I believe the poor performance in our random forest model may be caused by a limited hyperparameter grid. Due to lack of computation resources, we were only able to test a selected number of values for the hyperparameter. The random forest model may see an increase in performance with further tuning.

In real life, the cost of falsely classifying a performing student as failing is less costly than the cost of falsely classifying a failing student as passing. It is more useful to have a model that can accurately identify the positive cases, so we can recognize the signs of performance decline early on. Let's take a look at what the best performing models consider as significant predictors of class performance.

Figure 5: Significant Features Selected by Logistic Regression with Lasso Regularization and SMOTE

```
28 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept) 0.365403530
ST06Q013    0.150190184
ST16Q011    0.086952529
ST16Q021    0.110539271
ST17Q011    0.055211835
ST18Q021    0.007417425
ST18Q061    0.002819843
ST19Q031    0.110148692
ST19Q045    0.050727190
ST21Q061    0.011888783
ST22Q072    0.049224384
ST23Q011    0.422010048
ST23Q032    0.041938223
ST24Q031    0.045941112
ST24Q073    0.001582676
ST26Q074    0.019988150
ST26Q122    0.012328841
ST26Q132    0.082269919
ST28Q0170   0.891226150
ST28Q0228   0.006920845
ST30Q011    0.025070239
ST32Q042    0.063445419
ST35Q071    0.035485153
ST36Q041    0.076720345
ST37Q014    0.054542615
ST39Q051    0.030756606
ST41Q043    0.997220182
ST41Q063    0.954803675
```

Out of all the models we have so far, the logistic regression model with SMOTE and lasso regularization has the highest true positive rate. ~67.57% of failing students in math class were identified correctly. This is decent since there are a lot of random effects and noise in real life situations. Some of the features the model considers significant for predicting performance in math class are mother's job status, performance in other classes, special course attendance, class size, and teacher characteristics.

Figure 6: Significant Features Selected by Logistic Regression with Elastic Net Regularization and SMOTE

```
## 49 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept) 0.1916909144
## ST04Q052    0.0006821542
## ST06Q013    0.0853688456
## ST15Q012    0.0049145983
## ST16Q011    0.0472233012
## ST16Q021    0.0613756077
## ST16Q031    0.0364087000
## ST18Q021    0.0523458009
## ST18Q051    0.0343622424
## ST18Q061    0.0479237750
## ST19Q011    0.0174204044
## ST19Q031    0.0549927654
## ST19Q045    0.0490027028
## ST20Q025    0.0213323975
## ST21Q021    0.0106335763
## ST21Q061    0.0632454389
## ST22Q061    0.0475915179
## ST22Q072    0.0625801946
## ST23Q011    0.2346502787
## ST23Q032    0.1094586057
## ST23Q033    0.0145598304
## ST24Q031    0.0893049503
## ST24Q052    0.0199137959
## ST24Q072    0.0810059014
## ST24Q073    0.1456910998
## ST26Q074    0.0273370153
## ST26Q074    0.0273370153
## ST26Q122    0.0098767713
## ST26Q132    0.0695315242
## ST27Q0313   0.3281335878
## ST28Q012    0.0450055927
## ST28Q0170   0.8662525579
## ST28Q0228   0.0102262414
## ST29Q034    0.0371755469
## ST30Q011    0.0209276167
## ST31Q081    0.0198187508
## ST32Q012    0.0056020721
## ST32Q042    0.0368068401
## ST35Q044    0.0172696299
## ST35Q071    0.0396481239
## ST35Q094    0.0390737754
## ST36Q021    0.0003085383
## ST36Q041    0.0724417987
## ST37Q013    0.0060480053
## ST37Q014    0.0356663967
## ST39Q045    0.0074100409
## ST39Q051    0.0628265659
## ST41Q043    0.7008678078
## ST41Q062    0.0279497002
## ST41Q063    0.7615726708
```

The model with the second highest sensitivity rate is the logistic regression with elastic net regularization and SMOTE. Out of the 162 features, it found 48 features to be significant at predicting performance in class. Some of the variables it found to be significant are having your own room, having a quiet place to study, taking special courses outside of school in other subjects in the past three years, and not owning comic books.

Figure 7: Significant Features Selected by Logistic Regression with Boruta and SMOTE

```
##     ST04Q041     ST05Q022     ST05Q023     ST05Q024     ST05Q025     ST06Q011
## 4.574525e-03 2.146011e-02 4.409681e-05 3.404014e-02 2.045803e-02 1.977827e-03
##     ST06Q012     ST06Q014     ST16Q011     ST16Q012     ST18Q042     ST18Q044
## 8.843918e-06 2.526952e-04 1.033328e-04 2.741117e-02 1.192303e-02 1.465021e-03
##     ST19Q061     ST19Q062     ST19Q063     ST19Q064     ST19Q065     ST20Q031
## 1.245167e-02 4.862322e-03 3.397312e-03 5.370493e-03 2.625124e-03 3.451884e-10
##     ST20Q032     ST20Q033     ST20Q034     ST20Q035     ST20Q051     ST20Q052
## 3.383332e-15 2.391051e-17 4.226256e-14 1.548933e-10 4.808010e-02 4.092239e-02
##     ST21Q031     ST21Q062     ST21Q101     ST21Q102     ST22Q041     ST22Q042
## 4.908331e-02 9.207940e-03 2.224891e-02 7.837310e-03 2.125663e-02 5.364910e-03
##     ST22Q043     ST22Q044     ST23Q011     ST24Q061     ST24Q062     ST24Q063
## 2.210911e-03 1.125419e-02 2.315891e-03 1.100600e-06 3.198412e-03 4.378295e-02
##     ST26Q051     ST26Q052     ST26Q053     ST26Q054     ST26Q071     ST26Q072
## 6.488559e-03 3.763540e-03 2.233473e-03 1.252790e-03 2.610701e-02 1.252231e-02
##     ST26Q073     ST26Q091     ST26Q092     ST26Q093     ST26Q121     ST26Q122
## 1.471923e-02 1.184700e-02 2.964658e-02 2.857722e-02 1.951356e-03 1.420164e-04
##     ST26Q123     ST26Q124     ST26Q141     ST26Q142     ST26Q143     ST26Q144
## 7.756052e-03 1.962689e-02 4.022230e-03 1.196002e-02 3.692931e-02 4.170093e-02
##     ST26Q151     ST26Q152     ST26Q153     ST26Q154     ST26Q171     ST26Q172
## 3.154816e-03 1.935970e-03 2.395296e-02 3.856515e-02 4.934880e-04 1.910831e-04
##     ST26Q173     ST26Q174     ST27Q053     ST27Q056    ST27Q0510    ST27Q0520
## 3.483022e-04 3.787980e-04 2.017103e-02 9.629448e-04 1.573812e-03 4.124867e-02
##     ST27Q062     ST28Q012     ST28Q015    ST28Q0110    ST28Q0113    ST28Q0115
## 5.497403e-03 2.781552e-03 3.766971e-03 2.726696e-05 4.054938e-07 5.912455e-06
##    ST28Q0116    ST28Q0118    ST28Q0119    ST28Q0120    ST28Q0121    ST28Q0122
## 1.148535e-02 2.238547e-02 4.290946e-02 3.334764e-11 3.341680e-10 1.767058e-08
##    ST28Q0123    ST28Q0124    ST28Q0125    ST28Q0126    ST28Q0127    ST28Q0128
## 3.038932e-14 4.417866e-08 8.260635e-05 4.801224e-04 1.070344e-03 6.967899e-03
##    ST28Q0129    ST28Q0130    ST28Q0132    ST28Q0135    ST28Q0160     ST39Q031
## 4.838646e-04 1.309967e-07 1.048469e-02 1.037287e-06 4.216618e-02 3.313158e-04
##     ST39Q032     ST39Q033     ST39Q034     ST39Q035     ST39Q041     ST39Q042
## 3.273130e-04 5.685223e-04 4.106020e-03 2.244053e-03 8.449269e-03 1.293676e-02
##     ST39Q043     ST39Q044     ST39Q045     ST41Q043     ST41Q061     ST41Q063
## 9.866916e-03 9.451521e-03 3.790006e-03 1.393942e-08 1.451225e-18 4.194572e-14
```

The logistic regression model with boruta feature selection and SMOTE has the third highest true positive rate. Some of the features it considers significant are whether students are living at home with certain family members, how often parents interact with the student in certain activities, how often parents work with student on schoolwork, how often certain activities happen during class, and mother's education.

6. Future Improvements

While this project was able to shed some light on the significant features that affect student performance, numerous improvements can be made to enhance the results. For starters, we would also like to know whether these significant variables apply to other populations. It would be bias to say that the results drawn from a Canadian student dataset can be extrapolated to European students. It would be interesting to repeat the analysis on different student populations from various countries. It would also be interesting to plot more variables for exploratory data analysis.

Another potential improvement is to improve model sensitivity rate further. Currently, the best model has a sensitivity rate of 67.57%. Due to the limitations of time and computational resources, we had to limit our hyperparameter grid during training. If possible, we would like to expand our hyperparameter grids in future iterations as that may improve the performance of the model.

Lastly, we can also look into additional models. In this analysis, we only used logistic regression and random forest models. In future iterations, we should also consider other models like k-nearest neighbor and neural networks. It is also important to note that KNN suffers greatly from high dimensionality, so it may be helpful to perform feature selection beforehand.