Supervised Machine Learning to Predict Student Proficiency in Oregon Public Schools

---

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Bhavjot Khurana

May 2023

Approved for the Division
(Mathematics)

_____

Njesa Totty

List of Abbreviations

AYP	Adequate Yearly Progress
ELA	English Language Arts
ESSA	Every Student Succeeds Act
FTE	Full-Time Equivalent
MDI	Mean Decrease Importance
NCLB	No Child Left Behind
NIR	No Information Rate
OOB	Out Of Bag
VIM	Variable Importance Measure

Table of Contents

Introduction **1**

Chapter 1: Background **3**
1.1 Percent Proficient 3
1.1.1 Definition 3
1.1.2 No Child Left Behind Act (2001) 3
1.1.3 Every Student Succeeds Act (2015) 4
1.2 Literature Review 4
1.2.1 Student Proficiency Testing 4
1.2.2 Approaches to Model Student Performance 5

Chapter 2: Data **7**
2.1 Final Data Set 9
2.2 Set Up Modeling 10
2.2.1 Stratified Sampling 10
2.2.2 Create Training and Test Sets 10
2.2.3 Dealing With Missing Values 10
2.2.4 Dealing with Level Differences in the Response 10

Chapter 3: Methods **11**
3.1 Decision Trees 11
3.2 Random Forest 12
3.3 Creating Untuned Random Forest 14
3.4 Best Model Using `train` from Package `caret` 14
3.4.1 Cross Validation 15
3.4.2 Synthetic Minority Oversampling Technique 16
3.4.3 Fitting Best Model 16
3.5 Model Evaluation Metrics 16
3.5.1 Decision Trees 16
3.5.2 Random Forest 17

Chapter 4: Results **19**
4.1 Exploratory Analysis 19
4.2 Decision Trees 25
4.2.1 English Language Arts (ELA) 25

4.2.2 Mathematics 31
4.3 Random Forest 37
4.3.1 English Language Arts (ELA) 37
4.3.2 Mathematics 46

Chapter 5: Discussion **55**
5.1 Results of Random Forest 55
5.2 Decision Trees vs Random Forest 56
5.3 Variable Importance 56
5.4 Missing Data 57

Conclusion **59**

Appendix A: Creating Visualizations **61**

Appendix B: Model Fitting **65**

References **71**

List of Tables

3.1 Complexity Parameter Table for $\texttt{tree}_a sian_e la$ 12

4.1 Average Number of Participants per School by Ethnicity 24
4.2 Average Number of Participants per School by gender 24
4.3 Summary Table 24
4.4 Confusion Matrix Results for Group "Asian" (ELA) 37
4.5 Confusion Matrix Results for Group "White" (ELA) 38
4.6 Confusion Matrix Results for Group "Black" (ELA) 39
4.7 Confusion Matrix Results for Group "Hispanic" (ELA) 40
4.8 Confusion Matrix Results for Group "Multi-Racial" (ELA) 41
4.9 Confusion Matrix Results for Genders (ELA) 42
4.10 Confusion Matrix Results for Group "Asian" (Math) 46
4.11 Confusion Matrix Results for Group "White" (Math) 47
4.12 Confusion Matrix Results for Group "Black" (Math) 48
4.13 Confusion Matrix Results for Group "Hispanic" (Math) 49
4.14 Confusion Matrix Results for Group "Multi-Racial" (Math) 50
4.15 Confusion Matrix Results for Genders (Math) 51

List of Figures

4.1 Proficiency Distributions for Different Ethnicities in English Language Arts 20
4.2 Proficiency Distributions for Different Ethnicities in Math 21
4.3 Proficiency Distributions for Different Genders in English Language Arts 22
4.4 Proficiency Distributions for Different Genders in Math 23
4.5 Decision Tree group "Asian" for ELA (Accuracy = 0.27) 25
4.6 Decision Tree group "White" for ELA (Accuracy = 0.30) 26
4.7 Decision Tree group "Black" for ELA (Accuracy = 0.36) 27
4.8 Decision Tree group "Hispanic" for ELA (Accuracy = 0.27) 28
4.9 Decision Tree group "Multi-Racial" for ELA (Accuracy = 0.30) . . . 29
4.10 Decision Tree for Gender for ELA (Accuracy = 0.24) 30
4.11 Decision Tree group "Asian" for Math (Accuracy = 0.19) 31
4.12 Decision Tree group "White" for Math (Accuracy = 0.31) 32
4.13 Decision Tree group "Black" for Math (Accuracy = 0.62) 33
4.14 Decision Tree group "Hispanic" for Math (Accuracy = 0.36) 34
4.15 Decision Tree group "Multi-Racial" for Math (Accuracy = 0.22) . . . 35
4.16 Decision Tree Gender group for Math (Accuracy = 0.24) 36
4.17 Variable Importance Plot (Race, ELA) 43
4.18 Variable Importance Plot (Gender, ELA) 44
4.19 Variable Importance Plot (Race, Math) 52
4.20 Variable Importance Plot (Gender, Math) 53

Abstract

The Every Student Succeeds Act (ESSA) requires all states to annually measure the academic performance of public schools in the United States, including Oregon. One critical metric of school performance is the "percentage proficient" score, which measures the percentage of students who achieve proficiency in English Language arts (ELA) and Math standardized tests for the 2018-19 year. This thesis compares the use of random forest models and decision trees to predict student proficiency for Oregon public schools using various student and school information. These variables include gender, ethnicity, faculty resources, and others. By comparing random forests and decision trees we aim to provide models that are both interpretable and accurate.

Introduction

The Every Student Succeeds Act (ESSA) mandates that schools provide data on student proficiency levels to assess their educational progress. The data include information on demographics, school environment, and teacher qualifications. The data is collected annually and is used to evaluate schools' effectiveness in achieving their educational goals.

Predicting student performance has always been a significant challenge in education. However, recent advancements in machine learning and data analytics have provided new tools for predicting and analyzing student performance (Namoun, 2020). One such tool is the random forest algorithm, which is a type of supervised statistical learning algorithm used for classification and regression tasks.

This thesis aims to use random forest models to predict student proficiency levels based on data collected under ESSA for the 2018-19 year. The variables used in the models are teacher full-time equivalent (FTE), teacher experience, school poverty rating, student participation rate, race, and gender. The goal is to provide accurate and reliable prediction of student proficiency levels for different demographic groups. This can be used to identify areas that require improvement and inform targeted interventions to support student achievement. In the case of this thesis, the random forest models are applied to both English Language Arts (ELA) assessments and Mathematics assessments. The statewide assessment data is compiled in 2019 for academic year 2018-19.

The study also aims to address the limitations of previous research on predicting student performance by exploring the effectiveness of using a combination of teacher and school-level variables in predicting student proficiency levels. Additionally, the study seeks to investigate how the predictive power of the model varies across different demographic groups, including different ethnicities and gender.

Chapter 1

Background

1.1 Percent Proficient

1.1.1 Definition

Percentage proficient was introduced into education law under the No Child Left Behind Act. It is a metric used to measure the percentage of students who have achieved a certain level of proficiency in a specific subject or skill. In education, it is often used to measure student achievement on standardized tests (Koretz, 2009). According to the U.S. Department of Education (2021), percentage proficient is typically calculated by dividing the number of students who scored at or above the proficiency threshold by the total number of students who took the test, and then multiplying that result by 100. The proficiency threshold is set by each state and represents the minimum level of knowledge or skill that a student is expected to have in a given subject.

According to information found on the Oregon Department of Education website, the Oregon statewide student assessment data for English Language Arts (ELA) and mathematics comes from the Smarter Balanced Assessment Consortium. In Oregon's case, there are four levels of scores a student can obtain - Level 1, 2, 3, or 4. A student is considered proficient if they score a Level 3 or 4. For example, if students of group "White" at Ashland Public School have 50% proficiency in Mathematics, it implies that 50% of these students achieved either a Level 3 or 4 in the statewide and national assessments.

1.1.2 No Child Left Behind Act (2001)

The No Child Left Behind (NCLB) Act was a landmark education reform bill signed into law by President George W. Bush in 2002. The aim of the act was to increase student achievement by holding schools accountable for the academic progress of all students, particularly those who are traditionally underserved (Hanushek, 2009). Under NCLB, states were required to administer annual standardized tests to all students in grades 3-8 in order to measure student proficiency in reading and math. Schools that did not make adequate yearly progress (AYP) towards the goal of 100% proficiency in reading and math by 2014 were labeled as failing and were subject to

a range of sanctions and interventions (Haas et al., 2005).

The law was widely debated, with critics arguing that it placed too much emphasis on testing and that its focus on sanctions failed to address the root causes of low student achievement. Others argued that the law was an important step towards closing achievement gaps and holding schools accountable for their performance (Koretz, 2009). The act was eventually replaced by the Every Student Succeeds Act.

1.1.3 Every Student Succeeds Act (2015)

The datasets used in this study were obtained under the Every Student Succeeds Act (ESSA). It is a federal education law that was signed into law by President Barack Obama in 2015. It replaced the No Child Left Behind Act (NCLB) and shifted more decision-making power back to the states while maintaining a focus on accountability and closing achievement gaps (Heise, 2017).

Under the Every Student Succeeds Act (ESSA) in Oregon, score at grade-level content in English Language Arts (ELA) and mathematics on annual statewide assessments is divided into 4 levels (Level 1, 2, 3, and 4). ESSA requires each state to set its own proficiency targets and long-term goals for academic achievement, graduation rates, and English language proficiency. The goal of ESSA is to ensure that all students, regardless of their background or circumstances, have access to a high-quality education that prepares them for college, careers, and life (Oregon Department of Education, 2021).

In addition to the traditional proficiency measures, ESSA also requires states to use additional indicators of school quality and student success. These indicators may include measures such as chronic absenteeism, college and career readiness, and progress toward English language proficiency for English learners.

1.2 Literature Review

1.2.1 Student Proficiency Testing

Research on student proficiency has been conducted in various educational settings and has explored a multitude of factors that affect academic performance. Among these factors, accountability policies and teacher involvement have been studied extensively to understand their impact on student proficiency (Jennings, 2014; Hashim et al., 2019; Okpala, 2002, Aina et al., 2013).

Accountability policies typically aim to hold schools and teachers responsible for student performance by using various metrics to evaluate their effectiveness. These metrics can include standardized test scores, graduation rates, and other measures of academic achievement. Some studies have suggested that accountability policies can improve student proficiency by creating incentives for schools and teachers to focus more on student performance. For example, a study conducted in New York City by Winters and Cowen, in 2012, found that accountability policies led to increased math proficiency in schools. They used the regression discontinuity approach to study

the influences of New York accountability policies. The study found that accountability policies were associated with increased proficiency in both reading and math, especially in schools serving low-income students.

On the other hand, other studies have shown that accountability policies may have no significant effect on student proficiency, or even have negative consequences. One study found that accountability policies had no impact on reading proficiency and that accountability policies were associated with increased pressure on teachers to "teach to the test," which could undermine other important educational objectives such as critical thinking and problem-solving skills (Bearkar, 2014). Effective teaching practices, such as teacher involvement and experience, have also been studied as potential factors in improving student proficiency. Research has shown that teacher involvement can improve student learning outcomes by promoting student engagement, interest, and motivation (Dana, 2018).

Other studies have examined the impact of various demographic factors on student proficiency, such as race, socioeconomic status, and language proficiency. These studies have found that these factors can have a significant impact on student academic performance, with students from disadvantaged backgrounds often facing greater challenges and experiencing lower proficiency levels (Tate, 1997).

We see in the literature that accountability details, teacher involvement, race, gender, socio-economic status are all important factors in understanding proficiency. The research does show a gap, however, for methods to predict student proficiency in Oregon, especially under ESSA. This is an important consideration because, as mentioned before, states are given more decision-making power under the new law and have a more direct effect on student performance (Heise, 2017).

1.2.2 Approaches to Model Student Performance

The literature review suggests that traditional statistical methods such as descriptive statistics, correlation analysis, and regression analysis have been used to predict and model student performance. These methods help analyze assessment data and identify trends in student learning (Papamitsiou, 2014). For example, descriptive statistics are useful in summarizing the distribution of student performance across various subjects and grade levels (Fisher & Marshall, 2009). Correlation analysis helps identify the relationship between different factors and student performance, such as the correlation between attendance rates and academic performance. Regression analysis is useful in identifying the factors that are most strongly associated with student performance and can be used to develop predictive models (Abbassi et. al, 2011).

The study by Winters and Cohen (2012) used regression discontinuity as their method to understand the effect of New York educational policies on student performance. Regression discontinuity is a statistical method used to estimate the causal effect of a treatment or program when the assignment to the treatment is based on a continuous score or variable. The basic idea is to compare the outcomes of units that are just above and just below a threshold score, assuming that the only difference between them is their treatment status (Imbens & Lemieux, 2007).

One study proposed a student performance prediction model that utilized deci-

sion trees and various features such as student demographics, academic history, and attendance data. The researchers explored the use of decision tree analysis to predict student performance in higher education institutions. The authors used final course grades as the basis for determining academic success, with students who obtained a final grade of C or higher considered successful, while those who obtained a final grade lower than C were considered unsuccessful. The authors used data from a public university in Turkey to develop and evaluate their predictive model. The study proposed a framework for using machine learning algorithms to analyze large-scale student data and assess school effectiveness. The authors found that decision tree analysis was an effective method for predicting student performance, with an accuracy rate of over 80%. They also noted that the predictive power of the model could be improved by incorporating additional factors such as socio-economic status and student engagement. The model outperformed regression models in predicting student performance (Hamoud, 2018).

More recent research also suggests that such machine learning algorithms, particularly decision trees and random forest models, are more effective in predicting and modeling student performance. Random forest models are particularly useful for handling high-dimensional data with complex interdependencies (Hashim, 2020). Random forests work by constructing multiple decision trees, each based on a randomly selected subset of the data, and combining their predictions to produce a final result. This approach can lead to more accurate predictions of student performance, particularly when dealing with complex, multi-dimensional data as mentioned as part of a literature review conducted by Zacharoula Papamitsiou (2014) on learning analytics.

The literature shows that various factors affect student proficiency, including accountability policies, teacher involvement, demographic factors, and others. Studies on accountability policies have yielded mixed results, with some indicating a positive impact on student proficiency, while others suggest no effect or negative consequences. Effective teaching practices, such as teacher involvement and experience, have been shown to improve student learning outcomes. Demographic factors such as race, socioeconomic status, and language proficiency can also have a significant impact on student performance, with students from disadvantaged backgrounds often experiencing lower proficiency levels. It can also be seen that though regression models have been used to predict and model student performance, recent research suggests that machine learning algorithms, such as decision trees and random forest models, may be more effective. Based on this information, this study proposes to predict student proficiency in Oregon while also showing a comparison of methods between decision trees and random forests.

Chapter 2

Data

The 2018-19 statewide assessment and accountability data for Oregon, taken from the Oregon Department of Education website, includes a wide range of information about Oregon's public schools and districts. The data is collected annually and covers various aspects of education, including academic achievement, school climate, school demographics, and teacher qualifications.

The 2018-19 academic year was specifically chosen as the basis for modeling student proficiency in ELA and Mathematics. This decision was made in order to focus on a period before the onset of the COVID-19 pandemic and its subsequent impact on the education system. By using data from this pre-pandemic period, the models are better equipped to capture the nuances of traditional, in-person learning environments. Analyzing data collected after the pandemic would introduce additional complexities, such as the effects of online and hybrid learning, which could potentially confound the models and reduce their predictive accuracy (Vijayan, 2021). By limiting the scope to the 2018-19 academic year, this study aims to provide a more reliable and focused analysis of student proficiency across various racial groups and genders in the context of conventional educational settings.

According to information found on the Oregon Department of Education website, the Oregon statewide student assessment data for English Language Arts (ELA) and mathematics comes from the Smarter Balanced Assessment Consortium. It is a group of states working together to develop and administer high-quality assessments that measure student proficiency in ELA and mathematics. The Smarter Balanced assessments are computer-adaptive, meaning the questions adjust to the student's level of knowledge as they progress through the test. The assessments are aligned with the Common Core State Standards and are designed to measure critical thinking, problem-solving, and communication skills (Sato et al., 2011). In Oregon, all students in grades 3-8 and 11 are required to take the Smarter Balanced assessments in ELA and math each year. The results of these assessments are used to measure student proficiency, evaluate school and district performance, and inform decisions related to instruction and curriculum (Gordon, 2017).

The dataset contains information about 1236 public schools in Oregon, including traditional public schools, charter schools, and alternative education programs (ranging from elementary to high schools). The dataset also includes graduation rates,

chronic absenteeism rates, and other indicators of academic progress.

The teacher qualifications data in the dataset includes information about the education level and experience of teachers in Oregon's public schools. The source of teacher qualification data for Oregon's Department of Education comes from the Teacher and Administrator Data Mart (TAD). TAD is a system that collects data from Oregon's school districts and provides reports to support policy decisions and analysis. The data includes information such as teacher licensure status, years of experience, and educational qualifications. The data is collected annually from each school district in the state and is used to assess the qualifications and effectiveness of teachers and administrators.

The free and reduced-price food data for Oregon DOE comes from the National School Lunch Program (NSLP), which is a federally assisted meal program in public and nonprofit private schools and residential child care institutions. The program provides nutritionally balanced, low-cost or free lunches to eligible children each school day. Schools and institutions submit data to the Oregon Department of Education to determine the poverty rate of their student population, which is then used to allocate resources and funding to support student success (Fazlul et al., 2021).

2.1 Final Data Set

The final wrangled dataset used in the analysis includes the dependent variable "proficiency" and independent variables, including "group", "total_teacher_fte", "teacher_experience_avg , "teacher_grad_fte", "rate_participation", and "poverty_rating". These are defined as follows:

- `proficiency`: This is the dependent variables that refers to "percent proficient" as defined earlier. It is a "factor" type variable that has 10 levels - "<10%", "10-20%", "20-30%"..."90-100%". Each `group` in the dataset is

- `group`: This refers to the group that the students belong to and is a factor type variable. The levels include ethnicities "Asian", "Black/African American", "Hispanic/Latino", "Multi-Racial", "White" and genders "Male" and "Female". The data does not include intersections between gender and ethnicity, for example, it "Male" and "White" are separate groups within the dataset making it hard to analyze assessment data for White Male students.

- `total_teacher_fte`: This represents the sum total of teacher Full-Time Equivalent (FTE) for a specific school and is of type "numeric" . Full-Time Equivalent is calculated as the number of total hours worked by an employee divided by the number of hours that a full-time employee is expected to work in a given period, typically one year. For example, an employee who works 20 hours a week would be considered as 0.5 FTE (20/40), which means that they are half-time equivalent to a full-time employee who works 40 hours a week.

- `teacher_grad_fte`: This refers to the percentage of `total_teacher_fte` for teachers who possess graduate degrees, and is of type "numeric".

- `teacher_experience_average`: This refers to the sum total of years of experience of teachers in the school, and is of type "numeric"/

- `rate_participation`: The rate of participation variable refers to the percentage of students who participate in extracurricular activities, such as sports, clubs, and organizations, in a given school or district. This variable is of type "numeric"

- `poverty_rating`: This represents the percentage of students who are eligible for free or reduced-price meals. It is a "factor" type variable that has three levels: "Low Poverty", "Middle Poverty", and "High Poverty". "High Poverty" includes schools in the upper quartile (25%) of all schools statewide based on the percentage of students eligible for free or reduced priced school meals. "Low Poverty" includes schools in the lower quartile of all schools statewide based on the percentage of students eligible for free or reduced price school meals. All schools identified between the upper and lower quartile are marked "Middle Poverty" (Oregon Department of Education, 2021).

2.2 Set Up Modeling

2.2.1 Stratified Sampling

Stratified sampling is a sampling technique that ensures that the sample represents the population's characteristics in the same proportions as the population. In other words, it divides the population into subgroups, called strata, based on some variable of interest, in our case `proficiency`. Then, it takes a random sample from each stratum, such that the proportions of the strata in the sample are the same as in the population (Lohr, 2019).

2.2.2 Create Training and Test Sets

First, the dataset for a specific `group` is filtered from the big set and using the `create_test_train_strat` function (created with help from Njesa Totty) we split the dataset into training and test sets. The `create_test_train_strat` function implements stratified sampling by using the "sample" function to select a random sample of rows from the dataset, but it also sets the "prob" argument to a vector of equal probabilities for each row. This ensures that each row has an equal chance of being selected in the sample. By default, the function creates a training set with 80% of the rows and a test set with 20% of the rows, but these values can be changed by specifying the `size` argument.

2.2.3 Dealing With Missing Values

To train the model efficiently we need to deal with missing values in the `proficiency` and to do that we employ the use of `na.roughfix`. The `na.roughfix` function is part of the `impute` package in R, which provides several methods for imputing missing values. The `roughfix` method, as the name suggests, is a rough approximation that replaces missing values with their nearest non-missing neighbor's value. The na.roughfix method can handle mixed data types, including continuous, categorical, and ordinal variables which makes it useful in our case.

2.2.4 Dealing with Level Differences in the Response

To be able to model and test proficiency for different groups in ELA and Math, some levels had to be combined for proficiency due to less availability of significant values in each range (<10%, 10-20%,..., 90-100%). For example, in group "Asian" for ELA, the levels has to be combined: 0-40%, 40-50%, 50-60%, 60-70%, 70-80%, 80-90%, 90-100%. Though stratified sampling was used, there were low number of schools that had proficiency for Asian students between 0-40% for ELA. Almost every group had similar level changes (both in ELA and Math) which can be seen in the results section for the confusion matrices.

Chapter 3

Methods

This chapter highlights the procedure used to build the final model that is used to predict `proficiency` for all groups: "White", "Black/African American", "Hispanic/Latino", "Asian", "Multi-Racial", "Male, and "Female". All of the ethnicity groups are modeled separately due to the vast variance in proficiency for Math and ELA but using the same method as listed below. The gender groups ("Male", "Female") are modeled together. The code examples used in this section are from group "Asian" but are standard across all other models for other groups. The modeling starts with building decision trees for the dependent variable `proficiency` against the independent predictors.

3.1 Decision Trees

A decision tree is a hierarchical model that splits the predictor space into a number of rectangular regions, where the response variable is assumed to be constant. It works by breaking down a dataset into smaller and smaller subsets based on criteria that best separates the data. Each subset is then analyzed and broken down further until the algorithm determines that the subsets are sufficiently homogeneous or distinct from each other. Each internal node of the tree specifies a test of one of the predictors, and the branches leaving the node correspond to the possible values of the predictor variable. The terminal nodes (also called leaves) of the tree represent the predicted response for the observations that fall in that region (James et al., 2021). The rpart function was used to create decision trees, which are a simple yet powerful model for making predictions. This results in a tree-like structure that can be easily interpreted and visualized, and to be able to tune the decision trees to maintain low amounts of error we use the complexity parameter (CP).

The complexity parameter is a tuning parameter that determines the level of pruning in a decision tree. It controls the trade-off between model complexity and accuracy. The complexity parameter in decision trees is used to control the complexity of the tree by pruning branches that do not significantly improve the fit of the model. The pruning process stops when the addition of another branch would not significantly increase the fit of the model. The fit of a decision tree is typically mea-

sured using a metric such as accuracy, which calculates the proportion of correctly classified instances divided by the total number of instances. Another commonly used metric is error rate, which is the proportion of misclassified instances divided by the total number of instances.

A smaller value of complexity parameter results in a more complex tree, with more splits and potentially better fit to the training data, but with an increased risk of overfitting and worse performance on new data. Table 3.1 serves as an example of a complexity parameter table used to tune the tree for group "Asian" for ELA.

Table 3.1: Complexity Parameter Table for $\texttt{tree}_{a}sian_{e}la$

	CP	nsplit	rel error	xerror	xstd
1	0.07	0.00	1.00	1.08	0.03
2	0.02	1.00	0.93	1.02	0.04

The value of the complexity parameter (CP) that results in the lowest cross-validation error rate (error) is selected as the optimal value. The idea is to pick a complexity parameter that avoids overfitting the model to the training data. By selecting the optimal complexity parameter, we ensure that the model performs well on new, unseen data.

3.2 Random Forest

Random Forests is a type of ensemble model which is an extension of decision trees, and is a popular method used in machine learning for classification and regression tasks (Segal, 2004). Ensemble models are a type of machine learning model that combine the predictions of multiple individual models to generate a final prediction. The idea behind this is that the combination of multiple models can lead to better accuracy and more robust predictions compared to using a single model (Shaik, 2019). It involves the creation of many decision trees, where each tree is trained on a random subset of the available predictors and a bootstrap sample of the observations (Buvaneshwaran, 2022).

Bootstrap sampling is a statistical technique used to estimate the variability of a sample statistic or model parameter by resampling the original dataset with replacement to generate multiple new datasets of the same size. In other words, the technique involves randomly selecting observations from the original dataset with replacement to create a new dataset, and repeating this process multiple times. Each of the new datasets is a "bootstrap sample" that can be used to compute the sample statistic or model parameter of interest (Kulesa et al., 2015).

The final prediction is made by averaging the predictions of all the individual trees. In more detail, the Random Forest algorithm works as follows (Kulesa et al., 2015):

- From the original data, a large number of bootstrap samples are drawn (i.e., random samples of the same size as the original data but with replacement).
- For each bootstrap sample, a decision tree is grown using a random subset of the available predictors at each split point in the tree. The number of predictors to consider at each split is a tuning parameter that can be optimized using cross-validation. The algorithm will continue to split the nodes until all of the leaves are pure or until they reach a user-defined maximum depth. However, the splitting can also stop when the improvement in impurity measure or other metric falls below a certain threshold value.
- The final prediction for a new observation is made by averaging the predictions of all the individual trees, either by taking a simple majority vote for classification problems or by taking the mean for regression problems.

Hyperparameters are model settings that are fixed before training the model on a dataset. Hyperparameters cannot be learned from the data and are chosen based on the problem being solved and the characteristics of the dataset. For random forests, some of the hyperparameters include the number of trees in the forest (ntree), the minimum number of samples required to split an internal node (minsplit), the number of variables randomly sampled as candidates at each split (mtry), amd 'nodesize' determines the minimum number of samples in a terminal node.

- 'ntree' refers to the number of decision trees generated in the ensemble. Each decision tree is grown independently from the others. Increasing the number of trees can improve the accuracy of the model up to a certain point, beyond which the improvement is negligible or even decreases (Probst, 2019).
- 'minsplit' ensures that each split has enough samples to be statistically meaningful, thus avoiding overfitting. If the number of samples in a node is less than 'minsplit', the algorithm will not attempt to split the node any further, and it will become a terminal node (Probst, 2019).
- 'mtry' refers to the number of predictor variables randomly selected as candidates at each split in the decision tree. This parameter controls the amount of randomness injected into the model, which can help reduce overfitting and increase generalization performance. A small 'mtry' value will result in more randomness and diversity among the decision trees, while a large 'mtry' value will result in less randomness and more similar decision trees (Probst, 2019).
- 'nodesize' ensures that the tree doesn't become too complex and avoids overfitting by setting a threshold on the number of samples in a terminal node. If the number of samples in a terminal node is less than nodesize, the node will not be further divided, and it will become a leaf node (Probst, 2019).

These hyperparameters are important for controlling the size and complexity of the trees in the random forest, which can affect the performance of the model.

3.3 Creating Untuned Random Forest

We use the `randomForest` function in R, found in the `randomForest` package, which is the algorithm that builds an ensemble of decision trees. The algorithm then combines the predictions from all the trees to make a final prediction. This approach is known as "bagging" (short for bootstrap aggregating) and helps to reduce overfitting and improve the accuracy and robustness of the model (Liaw & Wiener, 2002). In an untuned random forest, the algorithm uses default values for the hyperparameters, which are the settings that control the behavior of the algorithm. The hyperparameters in a random forest include the number of trees in the forest (`nTree`), the number of features to consider at each split (`mtry`), and the minimum number of samples required to split a node, among others.

This model is used not for its accuracy but to get the best `nTree` which has the least Out Of Bag (OOB) error rate. The OOB error rate in `randomForest` is an estimate of the error rate that the model would have on new, unseen data, calculated based on the samples in the training set that were not used to build a particular decision tree in the forest.

3.4 Best Model Using `train` from Package `caret`

In the case of this study, the model looks like this for group "Asian" where `Asian_ELA_train_fixed` is the training set for ELA that has been rough fixed using `na.roughfix()`.

```
```{r}
bag_asian_ela <- randomForest(proficiency ~ teacher_grad_fte + total_teacher_fte

 + teacher_experience_avg + rate_participation + poverty_rating,

data = Asian_ELA_train_fixed, importance = TRUE)
```
```

The models for other groups follow the same structure with the training sets specific to the group. `proficiency` is the dependent variable and `teacher_grad_fte`, `total_teacher_fte`, `teacher_experience_avg`, `rate_participation`, `poverty_rating` are the independent explanatory variables (when modeling gender, `group`, divided into "Male" and "Female" is used also as an explanatory variable).

The `train` function in the `caret` (Classification and Regression Training) package is a powerful tool for model training, tuning, and evaluation in machine learning. It provides a unified interface for building and comparing different models using a wide range of algorithms, from simple linear regression to complex deep neural networks. The train function uses a user-defined formula and data set to fit a predictive model, and it can be customized with a range of tuning parameters, cross-validation methods, resampling techniques, and evaluation metrics. It also supports parallel processing for faster model training and can output detailed summaries and visualizations of the modeling process.

To understand the use of `train` in the case of this study we can again use the example from group "Asian" for ELA

```
‘‘‘{r}
asian_rf_train<- train(proficiency_chr ~ teacher_grad_fte + total_teacher_fte
+ teacher_experience_avg + rate_participation + poverty_rating,
data = Asian_ELA_train_fixed_1, method = "rf",
trControl = trainControl(method = "cv", number = 9, classProbs = TRUE,
summaryFunction = MySummary, sampling = "smote"), ntree = nTree,
tuneGrid = data.frame(mtry = c(1:5)), metric = "Accuracy")
‘‘‘
```

The different parameters are explained below:

- `asian_rf_train` is the object that will store the output of the tuning process.
- `Asian_ELA_train_fixed_1` is the name of the training dataset.
- `method = "rf"` specifies that we are using the random forest algorithm.
- `trControl` specifies the settings for the cross-validation process. `method = "cv"` specifies that k-fold cross-validation is used with `number = 9` folds. `classProbs = TRUE` specifies that class probabilities will be calculated. `summaryFunction = MySummary` (code for `MySummary` is available in the appendix) specifies that we will be using a custom summary function for model evaluation. `sampling = "smote"` specifies that SMOTE (Synthetic Minority Over-sampling Technique) will be used to address class imbalance.
- `ntree` specifies the number of trees in the random forest model which was selected based on the lowest OOB error rate.
- `tuneGrid` specifies the grid of values to be tested for the tuning parameter `mtry`.
- `metric = "Accuracy"` specifies the evaluation metric to be used to compare different models during the tuning process. In this case, the metric is accuracy.

3.4.1 Cross Validation

Cross-validation is a crucial technique in random forest modeling that helps to estimate the prediction error of the model on new, unseen data. It involves splitting the original dataset into multiple subsets or folds and iteratively training and testing the model on different subsets of the data. By doing this, cross-validation helps to prevent overfitting and provides a more accurate estimate of the model's performance on new data. K-fold cross-validation, where the original dataset is divided into k non-overlapping folds (9 folds in our case), is the most commonly used approach in random forest modeling. (Ghojogh & Krowley, 2019)

3.4.2 Synthetic Minority Oversampling Technique

We use Synthetic Minority Over-sampling Technique (SMOTE) sampling to help with class imbalance in `proficiency`. SMOTE creates synthetic data points by taking a small subset of the minority class and randomly selecting one of its nearest neighbors. It then generates a new data point along the line connecting the two points. This process is repeated for each sample in the minority class, resulting in a balanced dataset. It improves the performance of machine learning models by reducing bias and improving the accuracy of predictions for the minority class (Chawla et al., 2002). This is important in our case as it addresses the imbalance found in the minority classes for `proficiency`.

3.4.3 Fitting Best Model

```
```{r}
best_tune <- asian_rf_train$bestTune

asian_rf <- randomForest(as.factor(proficiency_chr) ~ teacher_grad_fte +
total_teacher_fte + teacher_experience_avg + rate_participation +
poverty_rating,
data = Asian_ELA_train_fixed_1, ntree = nTree, mtry = best_tune$mtry)
```
```

Here, we obtain the `bestTune` from the model fitted using `train`. The `bestTune` object contains the best tuning parameters from the training model with the least error after performing cross validation and checking for best `mtry`. The number of trees used follows from the train model as it had the lowest OOB error rate as tested earlier.

3.5 Model Evaluation Metrics

This section includes the results collected from the final models as mentioned in the earlier section. To test for model accuracy and understand the importance of the predictors in predicting the response we make use of the `confusionMatrix` function and the `varImp` function.

3.5.1 Decision Trees

r.part

The `rpart.plot()` function from the `rpart` package is what this study uses to make visualization simpler. The interpretation of the decision tree involves following the branches from the root node to the terminal nodes. At each node, the split criterion is evaluated for the incoming data, and the data is sent to the left or right branch based on whether it satisfies the condition of the split. The process continues until

the data reaches a terminal node, where a predicted outcome is given. By examining the path taken through the tree for a particular observation, it is possible to see which variables had the most impact on the prediction.

The numbers found within the boxes of the plot represent the predicted values for `proficiency` based on the set of independent variables we use as predictors. The top number in each box indicates the class that the observation belongs to, while the bottom number represents the proportion of observations in that class in that particular box.

Accuracy

We test the accuracy of the decision tree through a confusion matrix made by comparing the predictions of the decision tree and the values of proficiency in the test set. The accuracy score gives an indication of how well the decision tree is able to classify observations into their respective categories. It is calculated by dividing the number of correct observations by the total number of observations.

3.5.2 Random Forest

`confusionMatrix`

A confusion matrix, also known as an error matrix or contingency table, is a visualization tool used to evaluate the performance of classification models. It presents a summary of the model's predictions compared to the actual target values (ground truth) for a given dataset. The confusion matrix provides insight into the types of errors made by the classifier.

The confusionMatrix function in R, provided by the 'caret' package, is a powerful tool for evaluating the performance of multi-class classification models (Navin & Pankaja, 2016). It produces a variety of metrics, which can be interpreted to understand a model's strengths and weaknesses. In this study, we will discuss the interpretations of the following multi-class result metrics for the random forest models: sensitivity, specificity, positive predictive value (pos.pred.value), negative predictive value (neg.pred.value), F1-score (f1), and balanced accuracy.

- Sensitivity: Also known as the true positive rate or recall, sensitivity measures the proportion of actual positive instances that the model correctly identified as positive, for each class separately. A high sensitivity for a class indicates that the model is effective in detecting instances of that class, while a low sensitivity suggests that the model struggles to identify these instances accurately.

- Specificity: Also known as the true negative rate, specificity measures the proportion of actual negative instances that the model correctly identified as negative, for each class separately. A high specificity for a class implies that the model is effective in not misclassifying instances of other classes as belonging to the class in question, while a low specificity indicates that the model has difficulty identifying instances of other classes accurately.

- Positive Predictive Value (pos.pred.value): Also known as precision, this metric measures the proportion of predicted positive instances that were actually positive for each class. A high positive predictive value for a class indicates that when the model predicts an instance to belong to that class, it is likely to be correct.

- Negative Predictive Value (neg.pred.value): This metric measures the proportion of predicted negative instances that were actually negative for each class. A high negative predictive value for a class indicates that when the model predicts an instance not to belong to that class, it is likely to be correct.

- F1-score: The F1-score is the harmonic mean of precision and recall for each class separately, combining both metrics into a single value. It is particularly useful when dealing with imbalanced datasets, as it balances the importance of both false positives and false negatives. A high F1-score for a class indicates that the model has a good balance between precision and recall for that class.

- Balanced Accuracy: This metric is the average of sensitivity and specificity for each class, providing a balanced measure of a model's performance on both positive and negative instances for that class. Balanced accuracy is especially useful when dealing with imbalanced datasets, which is true in our case, as it accounts for the unequal distribution of classes.

`varImp`

`varImp` function in R is used to estimate the importance of predictor variables in a model. The importance score, also known as variable importance measure (VIM), is calculated for each predictor. The score shows the amount of information that each predictor contributes to the model.

One common method for calculating VIM is mean decrease impurity (MDI), which measures the reduction in impurity achieved by each predictor. Impurity is calculated using a metric such as Gini index, which measures the degree of randomness or impurity in a dataset.

Mean decrease Gini is a commonly used MDI method that measures the average reduction in Gini index across all trees in a random forest model. In a random forest, the MDG score is calculated for each variable by measuring the reduction in impurity (i.e., Gini index) achieved by splitting the data based on that variable. The importance of a variable is then estimated by averaging its MDG score across all decision trees in the forest. A higher mean decrease Gini score indicates that the predictor variable is more important for predicting the outcome variable (Calle, 2011). The `varImp` in R function is a useful tool for feature selection, as it helps to identify which variables are most important for making accurate predictions based on their MDG. (James et al., 2017)

Chapter 4

Results

4.1 Exploratory Analysis

This section explores the independent and dependent variables at a glance. The distribution plots (Figures, 4.1, 4.2, 4.3, 4.4) show proficiency distributions in comparison to other groups - the race/ethnicities are compared together and the genders are compared together. The plots are split by subject between ELA and Math.

Tables 4.1 and 4.2 show average number of participants per school, split again by ethnicity and gender. This helps provide a broader picture of school diversity in Oregon Public Schools for the 2018-19 year.

Table 4.3 uses the R `summary` function to give a glimpse into the independent variables that we use to model our dependent variable proficiency. For numeric variables, `summary` provides basic statistics such as the minimum and maximum values, median, mean, and quartiles. Additionally, it provides information on the number of missing values in each variable. For categorical variables it provides a table of counts.

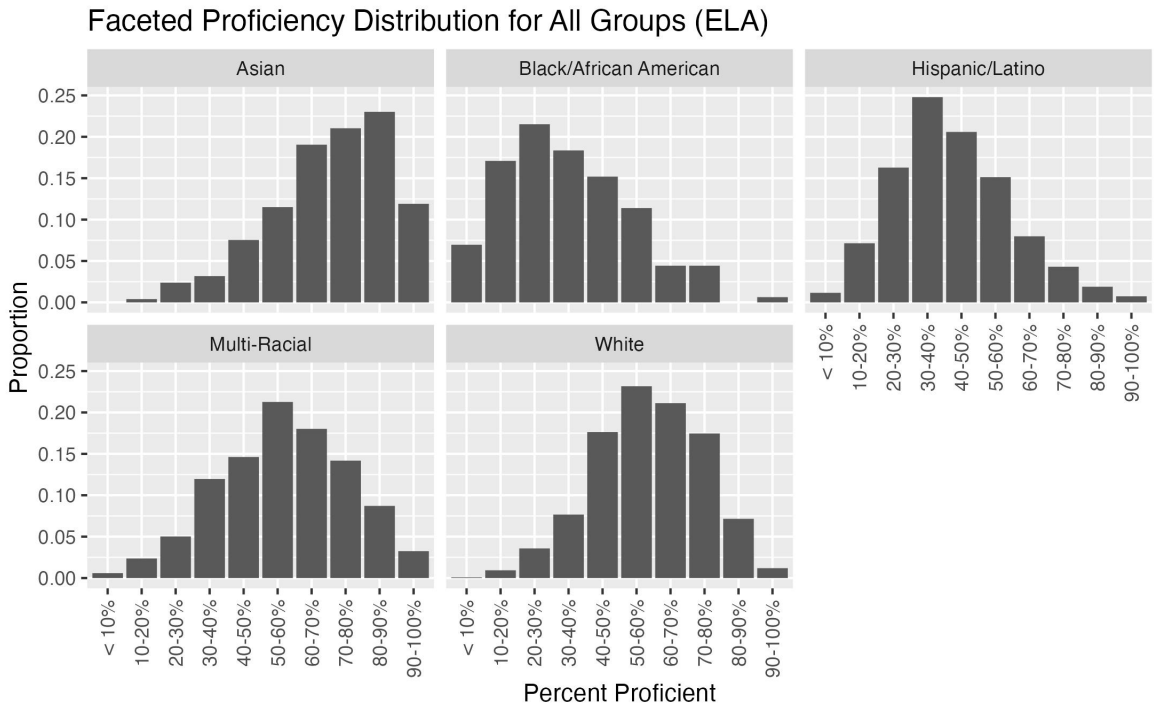

Figure 4.1: Proficiency Distributions for Different Ethnicities in English Language Arts

The plot above shows the proficiency distributions for all ethnicity groups in the dataset for ELA assessments. It can be seen in the plots that distribution Asian students have a left skew signalling generally higher proficiency. For Black and Hispanic students we see a sort of right skew towards the lower end of proficiency. White and Multi-Racial students have generally normal distributions with White students having some left skew.

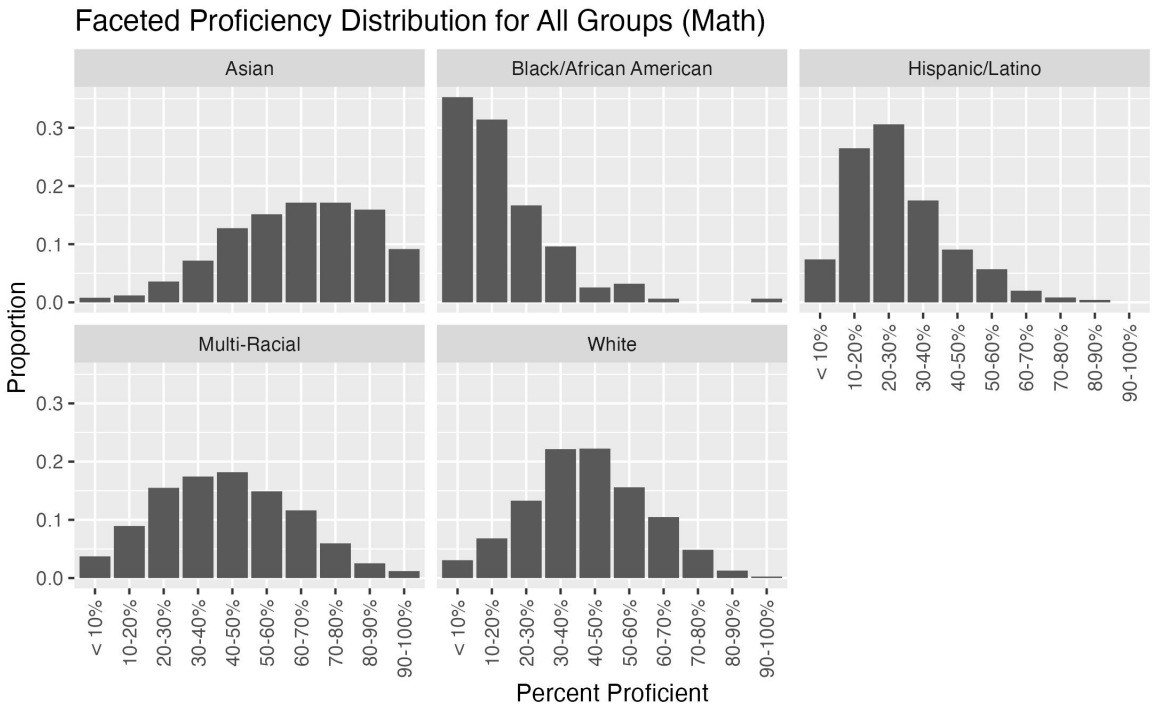

Figure 4.2: Proficiency Distributions for Different Ethnicities in Math

The plot above shows the proficiency distributions for all ethnicity groups in the dataset for Math assessments. It can be seen in the plots that distribution Asian students have a similar left skew signalling generally higher proficiency. For Black and Hispanic students we see a significantly heavier right skew towards the lower end of proficiency. White and Multi-Racial students have generally normal distributions again with some minimal right skew.

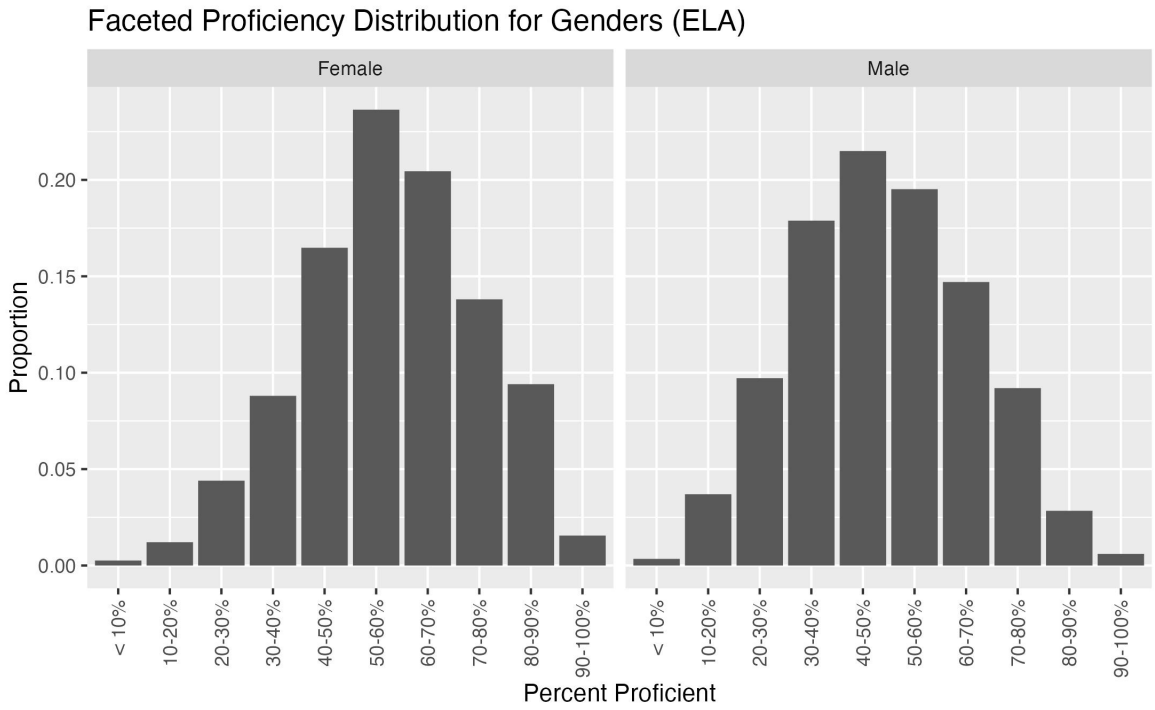

Figure 4.3: Proficiency Distributions for Different Genders in English Language Arts

The plot above shows the proficiency distributions for male and female in the dataset for ELA assessments. There is not a significant discernible difference in proficiency for either group but it seems that female students have a slightly higher proficiency trend.

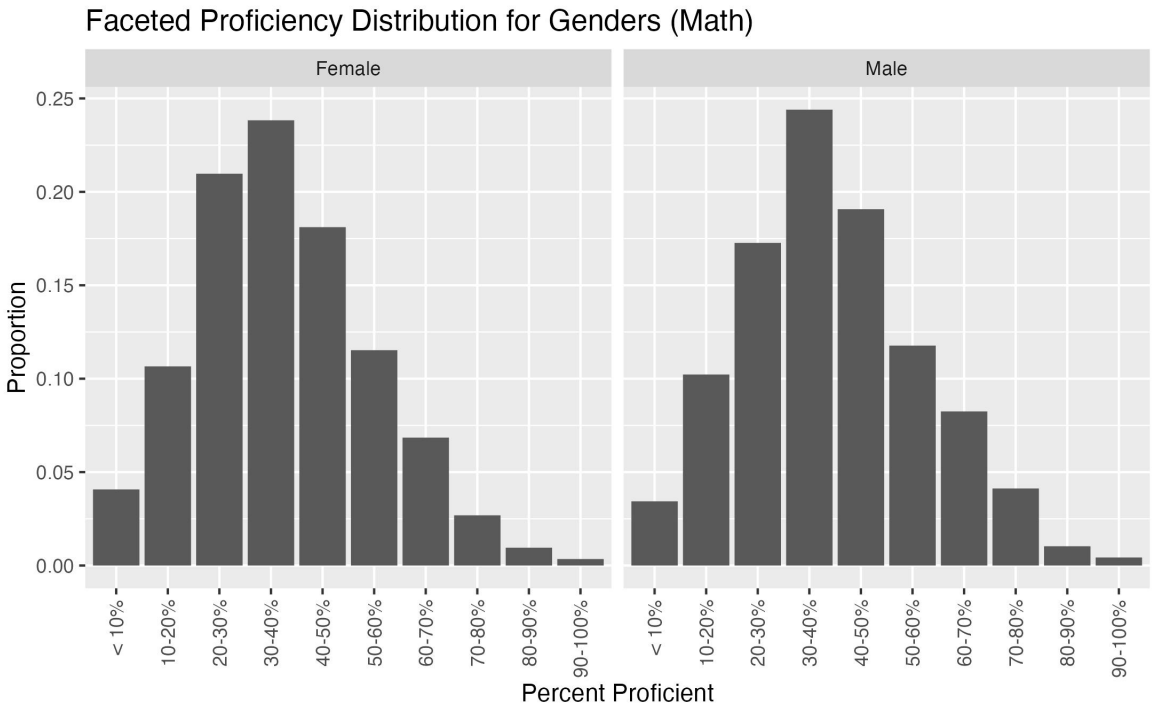

Figure 4.4: Proficiency Distributions for Different Genders in Math

The plot above shows the proficiency distributions for male and female in the dataset for Math assessments. There is no significant discernible difference between the groups but it seems that overall there is a some right skew signalling lower proficiency in math more generally.

Table 4.1: Average Number of Participants per School by Ethnicity

Group	Average Participants Per School
Asian	31.98
Black/African American	15.10
Hispanic/Latino	143.20
Multi-Racial	69.51
White	171.28

Table 4.2: Average Number of Participants per School by Gender

Group	Average Participants Per School
Female	180.38
Male	179.68

Table 4.3: Summary Table

Avg Teacher Experience	Teacher FTE (Graduate)	Total Teacher FTE	Poverty Rating
Min. : 0.000	Min. : 0.00	Min. : 0.01	High Poverty :312
1st Qu.: 9.797	1st Qu.: 63.80	1st Qu.: 11.55	Low Poverty :222
Median :12.035	Median : 75.03	Median : 18.18	Middle Poverty:702
Mean :11.807	Mean : 72.63	Mean : 20.75	
3rd Qu.:13.883	3rd Qu.: 84.50	3rd Qu.: 25.17	
Max. :36.000	Max. :100.00	Max. :183.67	
NA's :12	NA's :12	NA's :6	

4.2 Decision Trees

4.2.1 English Language Arts (ELA)

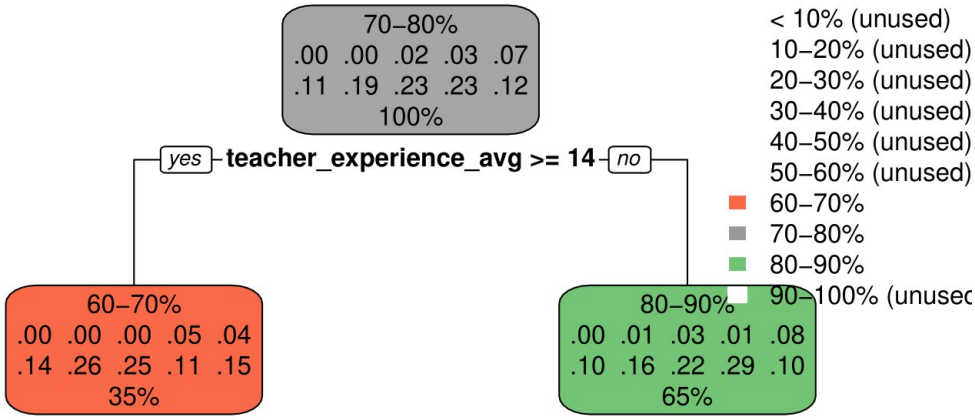

Figure 4.5: Decision Tree group "Asian" for ELA (Accuracy = 0.27)

The plot above shows the decision tree (made using `rpart.plot()`) for Asian students with response proficiency for English Language Arts. It uses the proficiency group "70-80%" as its root node and makes its first and only split if teacher experience average is greater than or equal to 14 years. The accuracy, when tested on the test set, for this model comes out to 0.27 which is fairly low. It doesn't use most of the proficiency buckets as can be seen in the legend of the plot. It classifies 35% of students in the 60-70% proficiency bucket and the rest of the 65% students in the 80-90% bucket, implying that a lower teacher experience average leads to increased proficiency for Asian students.

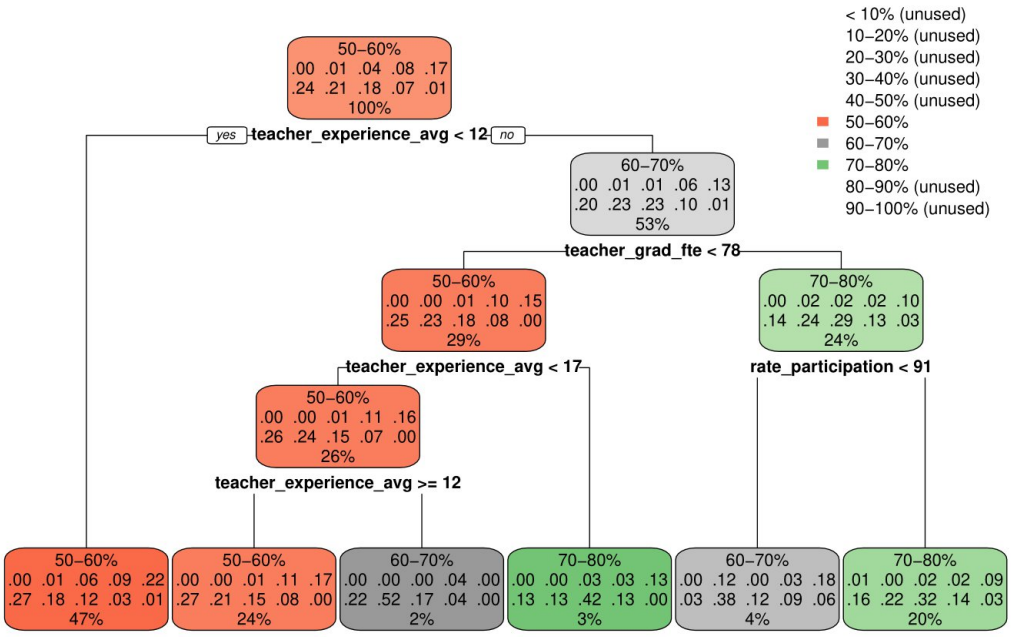

Figure 4.6: Decision Tree group "White" for ELA (Accuracy = 0.30)

The plot above shows the decision tree for White students with response proficiency for English Language Arts. It uses the proficiency group "50-60%" as its root node. The plot shows that generally higher teacher experience and teacher FTE for those with graduate degrees leads to higher proficiency for White students in ELA. The accuracy, when tested on the test set, for this model comes out to 0.3 which is fairly low. It again doesn't use most of the proficiency buckets as can be seen in the legend of the plot.

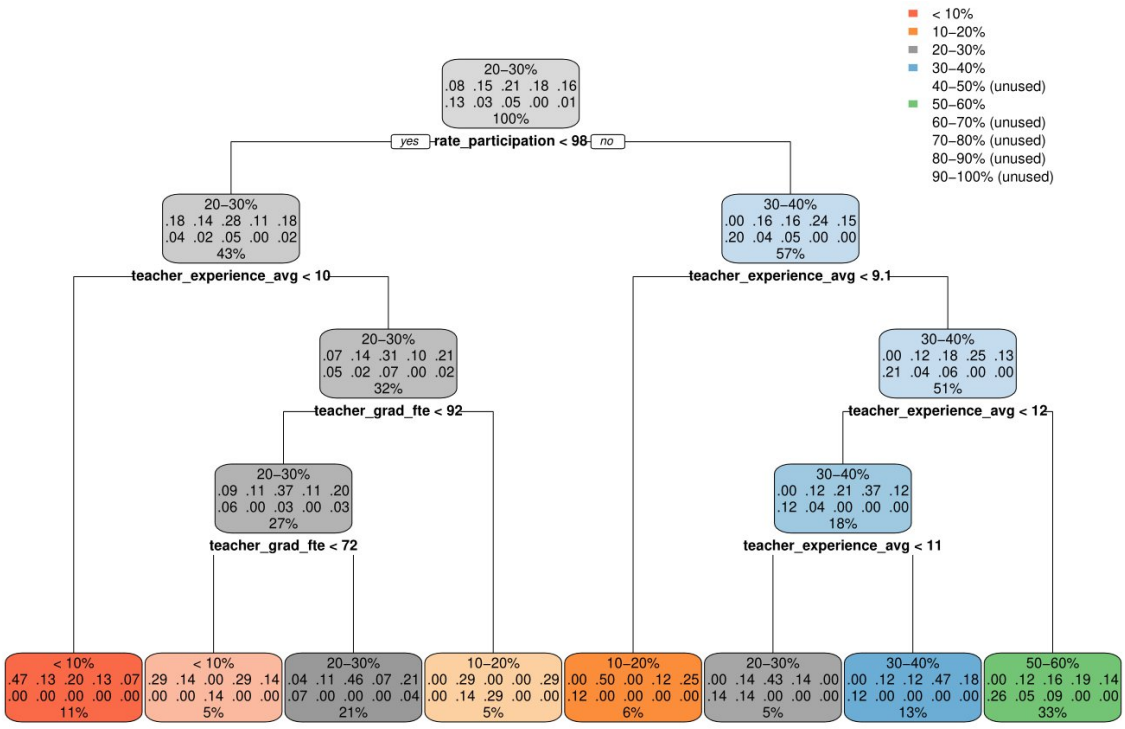

Figure 4.7: Decision Tree group "Black" for ELA (Accuracy = 0.36)

The plot above shows the decision tree for Black/African American students with response proficiency for English Language Arts. It uses the proficiency group "20-30%" as its root node. The plot shows that generally higher teacher experience and teacher FTE for those with graduate degrees leads to higher proficiency. The accuracy, when tested on the test set, for this model comes out to 0.36 which is fairly low. It again doesn't use most of the proficiency buckets as can be seen in the legend of the plot.

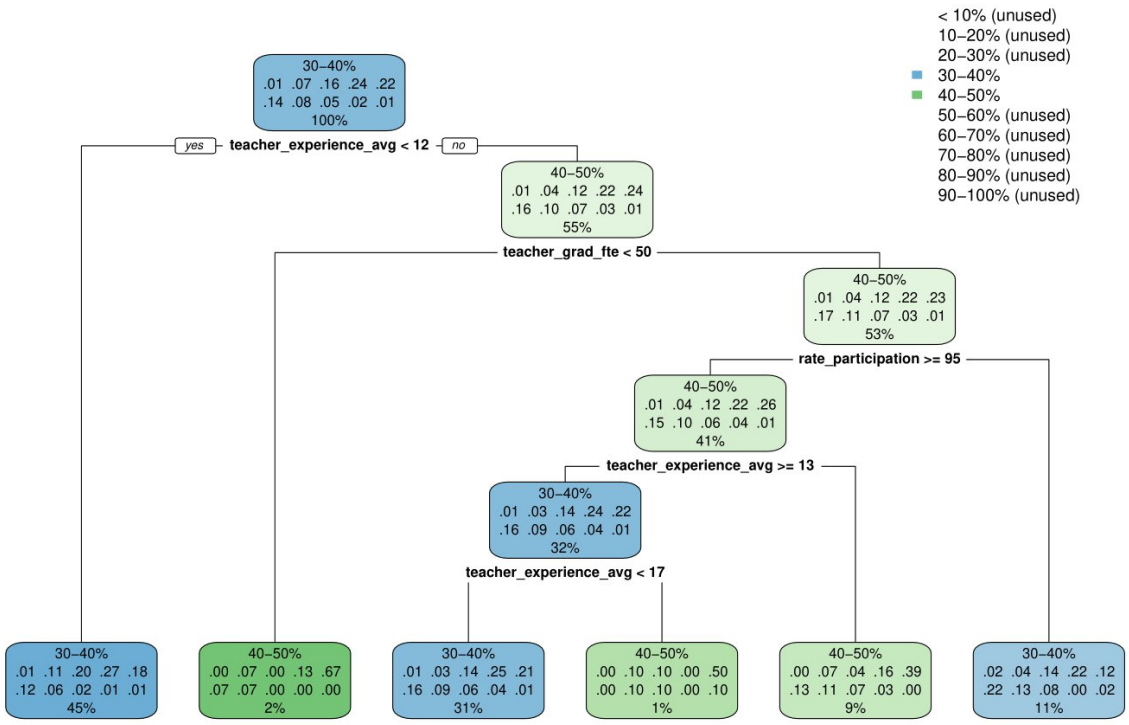

Figure 4.8: Decision Tree group "Hispanic" for ELA (Accuracy = 0.27)

The plot above shows the decision tree for Hispanic students with response proficiency for English Language Arts. It uses the proficiency group "30-40%" as its root node. The plot shows that generally higher teacher experience and teacher FTE for those with graduate degrees leads to higher proficiency. The accuracy, when tested on the test set, for this model comes out to 0.27 which is fairly low. It again doesn't use most of the proficiency buckets as can be seen in the legend of the plot.

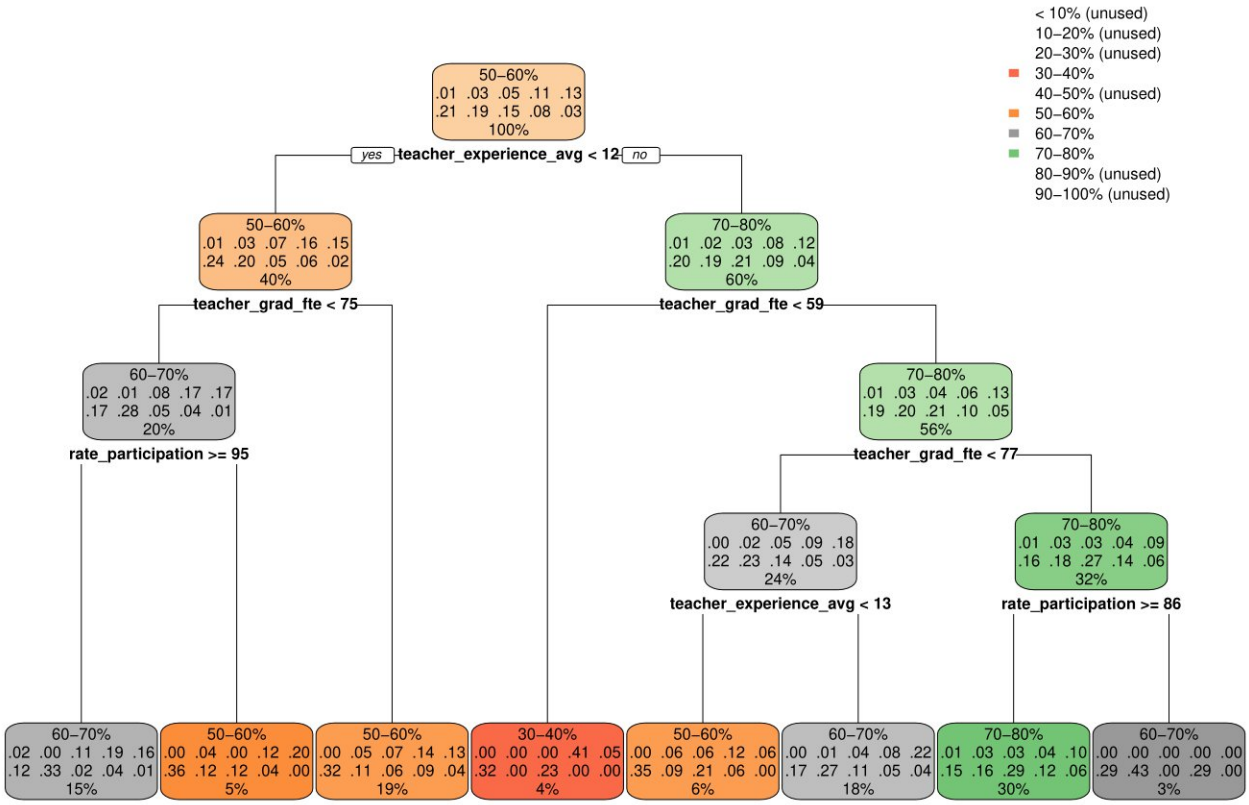

Figure 4.9: Decision Tree group "Multi-Racial" for ELA (Accuracy = 0.30)

The plot above shows the decision tree for Multi-Racial students with response proficiency for English Language Arts. It uses the proficiency group "50-60%" as its root node. The plot shows again that generally higher teacher experience and teacher FTE for those with graduate degrees leads to higher proficiency. It also shows that a higher rate of participation has a positive impact on proficiency. The accuracy, when tested on the test set, for this model comes out to 0.30 which is fairly low. It again doesn't use most of the proficiency buckets as can be seen in the legend of the plot.

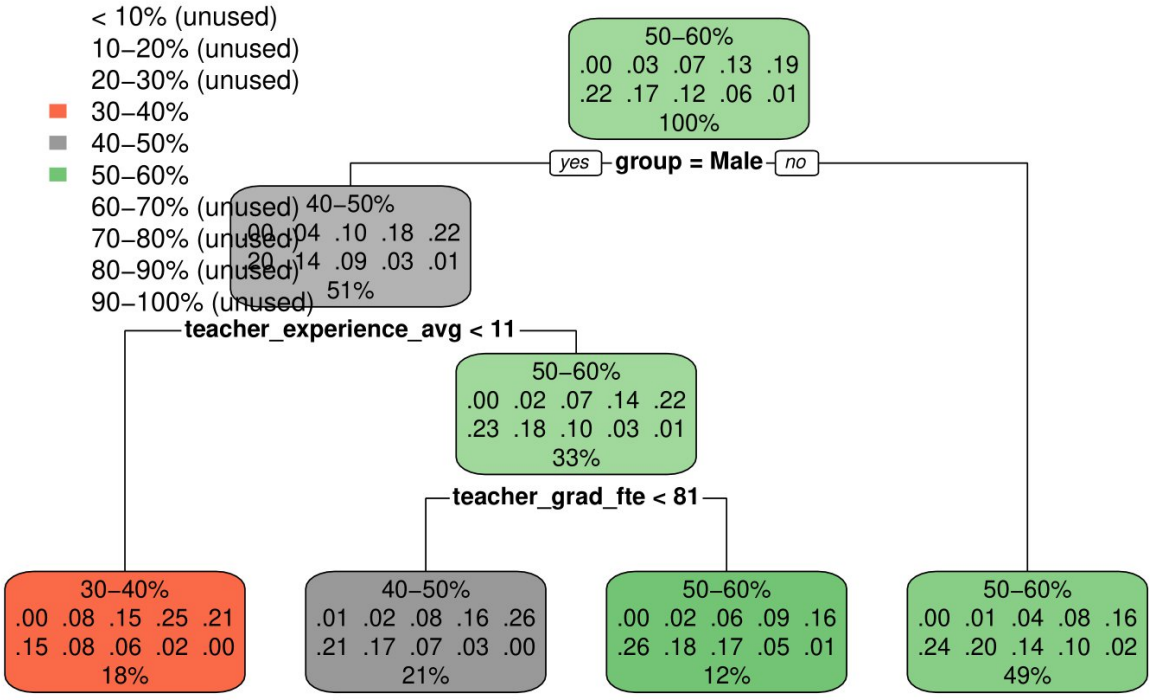

Figure 4.10: Decision Tree for Gender for ELA (Accuracy = 0.24)

The plot above shows the decision tree for male and female with response proficiency for English Language Arts. It uses the proficiency group "50-60%" as its root node. It shows that male students tend to have lower proficiency but depending on teaching resources they can match the proficiency of female students. The accuracy, when tested on the test set, for this model comes out to 0.24 which is fairly low. It again doesn't use most of the proficiency buckets as can be seen in the legend of the plot.

4.2.2 Mathematics

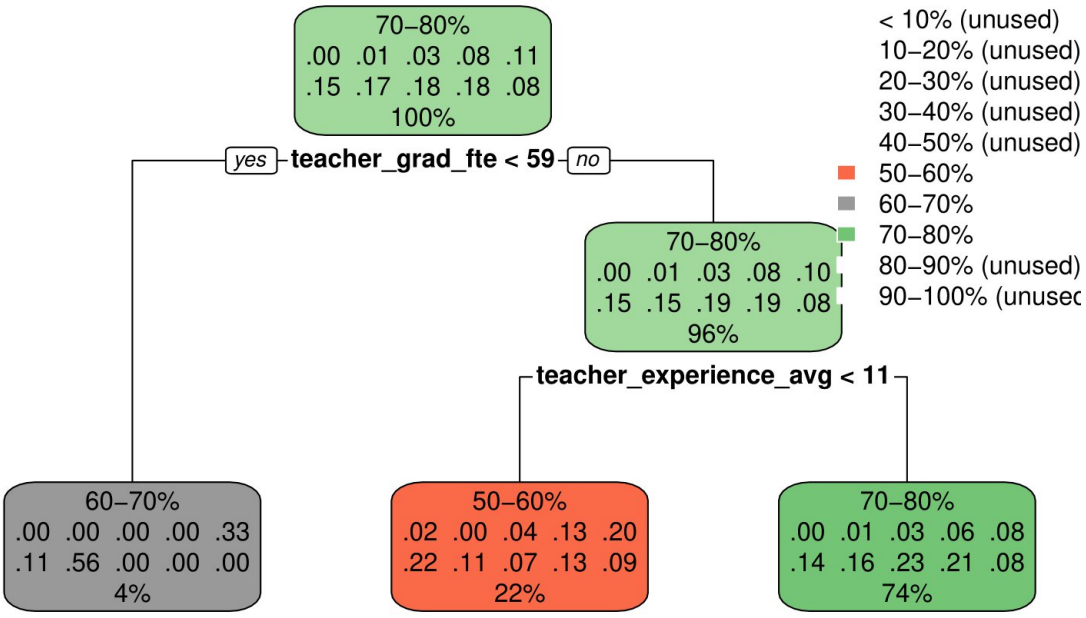

Figure 4.11: Decision Tree group "Asian" for Math (Accuracy = 0.19)

The plot above shows the decision tree (made using `rpart.plot()`) for Asian students with response proficiency for Mathematics. It uses the proficiency group "70-80%" as its root node. The plot shows that generally higher teacher experience and teacher FTE for those with graduate degrees leads to higher proficiency. The accuracy, when tested on the test set, for this model comes out to 0.19 which is fairly low. It again doesn't use most of the proficiency buckets as can be seen in the legend of the plot.

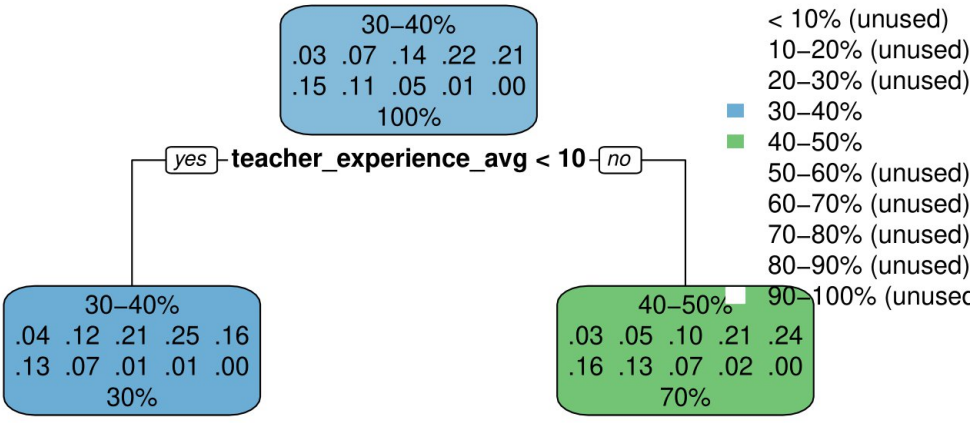

Figure 4.12: Decision Tree group "White" for Math (Accuracy = 0.31)

The plot above shows the decision tree for White students with response proficiency for Mathematics. It uses the proficiency group "30-40%" as its root node. The plot shows that generally higher teacher experience leads to higher proficiency. The accuracy, when tested on the test set, for this model comes out to 0.31 which is fairly low. It again doesn't use most of the proficiency buckets as can be seen in the legend of the plot.

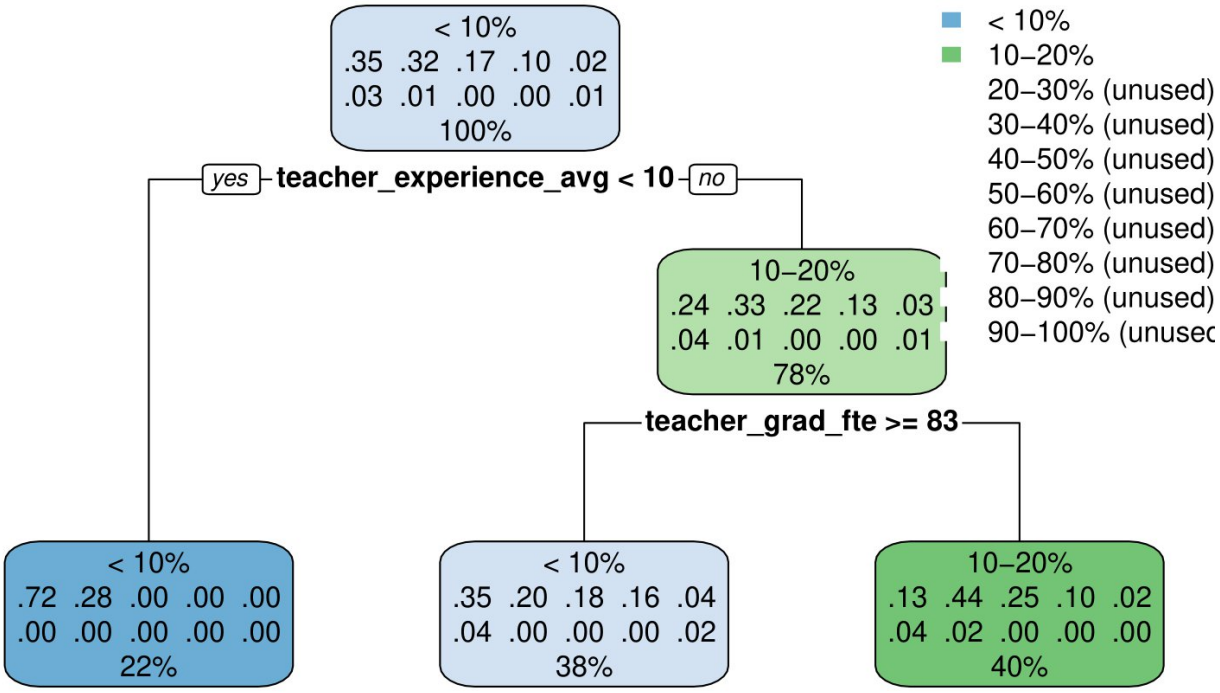

Figure 4.13: Decision Tree group "Black" for Math (Accuracy = 0.62)

The plot above shows the decision tree for Black/African American students with response proficiency for Mathematics. It uses the proficiency group "<10%" as its root node. The plot shows that generally higher teacher experience leads to higher proficiency but the proficiency buckets used are on the lowest end of proficiency. Surprisingly, higher FTE from teachers with graduate degrees leads to lower proficiency. The accuracy, when tested on the test set, for this model comes out to 0.62 which is reasonably higher than what was presented earlier. It again doesn't use most of the proficiency buckets as can be seen in the legend of the plot.

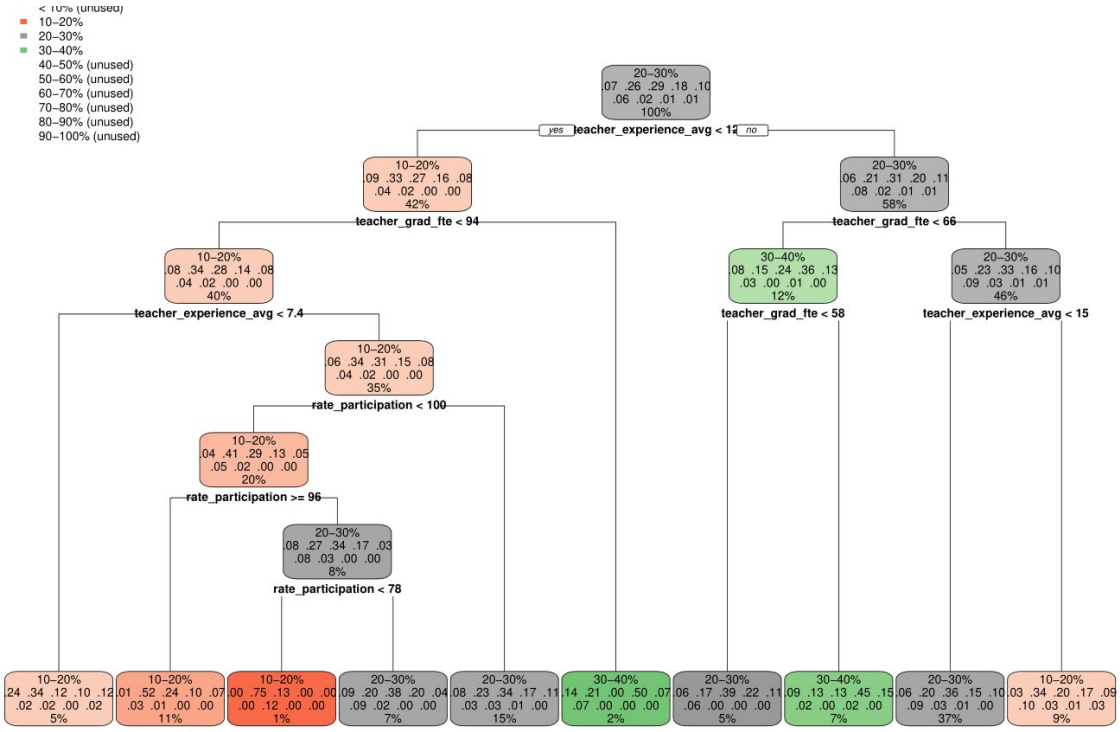

Figure 4.14: Decision Tree group "Hispanic" for Math (Accuracy = 0.36)

The plot above shows the decision tree for Hispanic students with response proficiency for Mathematics. It uses the proficiency group "20-30%" as its root node. The plot shows that generally higher teacher experience leads to higher proficiency. The accuracy, when tested on the test set, for this model comes out to 0.36 which is fairly low. It again doesn't use most of the proficiency buckets as can be seen in the legend of the plot.

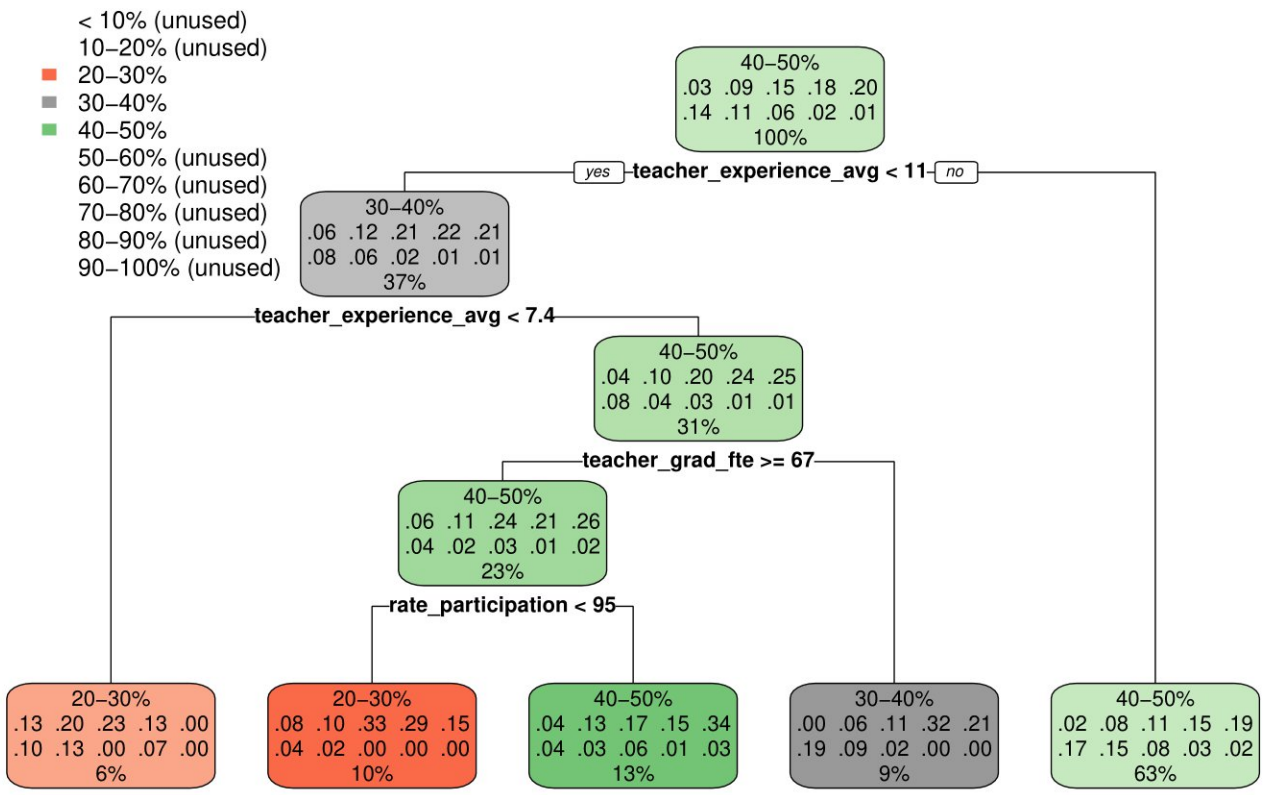

Figure 4.15: Decision Tree group "Multi-Racial" for Math (Accuracy = 0.22)

The plot above shows the decision tree for Multi-racial students with response proficiency for Mathematics. It uses the proficiency group "40-50%" as its root node. The plot shows that generally higher teacher experience leads to higher proficiency. The accuracy, when tested on the test set, for this model comes out to 0.22 which is fairly low. It again doesn't use most of the proficiency buckets as can be seen in the legend of the plot.

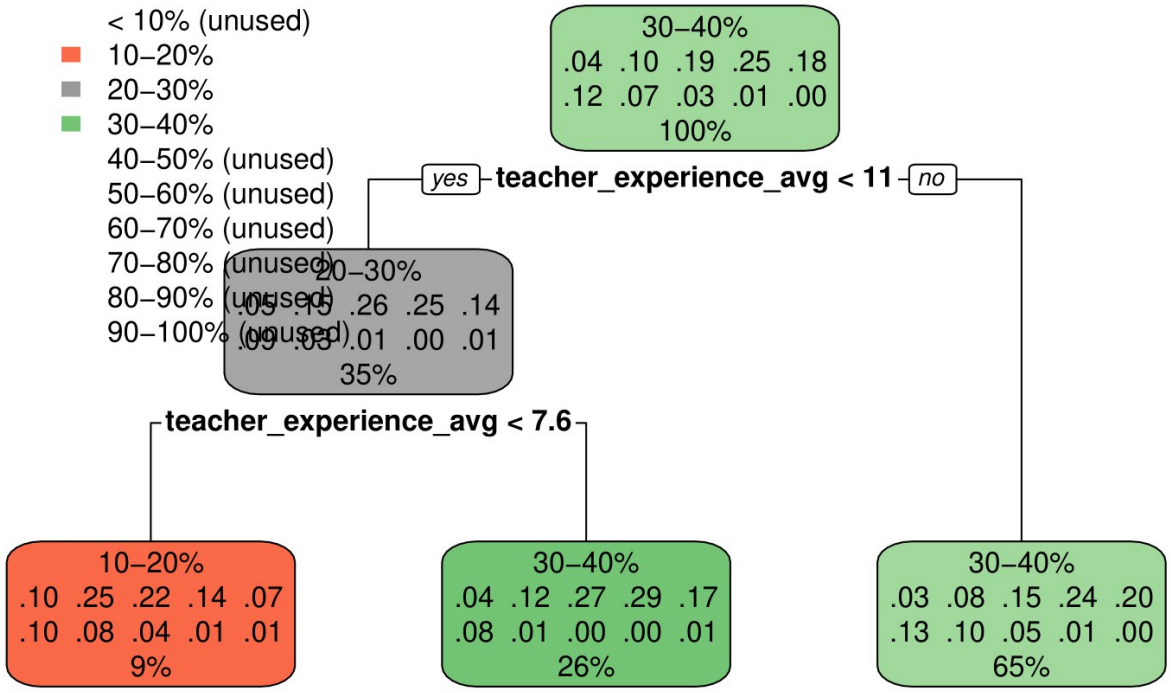

Figure 4.16: Decision Tree Gender group for Math (Accuracy = 0.24)

The plot above shows the decision tree for male and female students with response proficiency for Mathematics. It uses the proficiency group "30-40%" as its root node. The plot shows that generally higher teacher experience leads to higher proficiency. Interestingly, the tree doesn't make use of the student's gender in determining proficiency. = The accuracy, when tested on the test set, for this model comes out to 0.24 which is fairly low. It again doesn't use most of the proficiency buckets as can be seen in the legend of the plot.

4.3 Random Forest

4.3.1 English Language Arts (ELA)

Table 4.4: Confusion Matrix Results for Group "Asian" (ELA)

Class	Sensitivity	Specificity	Pos.Pred.Value	Neg.Pred.Value	F1	Balanced.Accuracy
0-40%	0.75	1.00	1.00	0.98	0.86	0.88
40-50%	1.00	1.00	1.00	1.00	1.00	1.00
50-60%	1.00	1.00	1.00	1.00	1.00	1.00
60-70%	0.86	0.97	0.86	0.97	0.86	0.92
70-80%	1.00	0.86	0.69	1.00	0.81	0.93
80-90%	0.76	1.00	1.00	0.88	0.87	0.88
90-100%	1.00	1.00	1.00	1.00	1.00	1.00

The confusion matrix results for the group "Asian" in the English Language Arts (ELA) proficiency classes (Table 4.4) indicate that the classification model performs exceptionally well for most classes, achieving perfect performance for the 40-50%, 50-60%, and 90-100% classes. High specificity is observed across all classes, ranging from 0.97 to 1.00. However, sensitivity varies, with lower values in the 0-40% (0.75) and 80-90% (0.76) classes. Additionally, the positive predictive value is notably lower for the 70-80% class (0.69). Overall, the F1-scores and balanced accuracy values are relatively high for all classes, suggesting a good balance between precision and recall, as well as sensitivity and specificity.

Table 4.5: Confusion Matrix Results for Group "White" (ELA)

Class	Sensitivity	Specificity	Pos.Pred.Value	Neg.Pred.Value	F1	Balanced.Accuracy
0-20%	0.75	0.99	0.67	0.99	0.71	0.87
20-30%	0.77	1.00	1.00	0.98	0.87	0.89
30-40%	0.78	0.97	0.81	0.97	0.79	0.88
40-50%	0.88	0.92	0.82	0.95	0.85	0.90
60-70%	0.90	0.97	0.90	0.97	0.90	0.94
70-80%	0.85	0.97	0.85	0.97	0.85	0.91
80-100%	0.89	0.99	0.89	0.99	0.89	0.94

The confusion matrix results for the group "White" in the English Language Arts (ELA) proficiency classes (Table 4.5) show that the classification model generally performs well across classes. Sensitivity ranges from 0.75 to 0.90, with the lowest value in the 0-20% class. Specificity is consistently high, between 0.92 and 1.00. The positive predictive value is the lowest for the 0-20% class (0.67). F1-scores and balanced accuracy values are relatively strong, indicating a balanced performance between precision and recall, as well as sensitivity and specificity. The lowest F1-score and balanced accuracy are observed in the 0-20% class (0.71 and 0.87, respectively).

Table 4.6: Confusion Matrix Results for Group "Black" (ELA)

Class	Sensitivity	Specificity	Pos.Pred.Value	Neg.Pred.Value	F1	Balanced.Accuracy
0-40%	1.00	0.78	0.92	1.00	0.96	0.89
40-50%	0.67	1.00	1.00	0.97	0.80	0.83
50-60%	1.00	1.00	1.00	1.00	1.00	1.00
60-100%	0.67	1.00	1.00	0.97	0.80	0.83

The confusion matrix results for the group "Black" in the English Language Arts (ELA) proficiency classes (Table 4.6) reveal that the classification model performs well in some classes, while there is room for improvement in others. The model achieves perfect performance for the 50-60% class, with sensitivity, specificity, positive predictive value, negative predictive value, F1-score, and balanced accuracy all equal to 1.00. Sensitivity varies across classes, with perfect sensitivity for the 0-40% class (1.00) and lower sensitivity for the 40-50% and 60-100% classes (both 0.67). Specificity is high for all classes, ranging from 0.78 to 1.00. The positive predictive value is notably lower for the 0-40% class (0.92) compared to other classes. The F1-scores and balanced accuracy values indicate a balanced performance between precision and recall, as well as sensitivity and specificity, but are lower for the 40-50% and 60-100% classes (both 0.80 and 0.83, respectively).

Table 4.7: Confusion Matrix Results for Group "Hispanic" (ELA)

Class	Sensitivity	Specificity	Pos.Pred.Value	Neg.Pred.Value	F1	Balanced.Accuracy
<10%	1.00	1.00	1.00	1.00	1.00	1.00
10-20%	0.79	0.99	0.85	0.98	0.81	0.89
20-30%	0.84	0.98	0.84	0.98	0.84	0.91
30-40%	0.81	0.95	0.85	0.94	0.83	0.88
40-50%	0.95	0.92	0.74	0.99	0.83	0.93
50-60%	0.79	0.99	0.94	0.95	0.86	0.89
60-70%	0.94	0.99	0.89	0.99	0.91	0.96
70-80%	0.78	1.00	1.00	0.99	0.88	0.89
80-100%	1.00	1.00	1.00	1.00	1.00	1.00

The confusion matrix results for the group "Hispanic" in the English Language Arts (ELA) proficiency classes (Table 4.7) display generally strong performance across classes. The model achieves perfect performance for the <10% and 80-100% classes. Sensitivity ranges from 0.78 to 1.00, with the lowest values in the 10-20% and 70-80% classes. Specificity remains consistently high, between 0.92 and 1.00. The lowest positive predictive value is observed in the 40-50% class (0.74). F1-scores and balanced accuracy values indicate a balanced performance between precision and recall, as well as sensitivity and specificity, with the lowest F1-score in the 10-20% class (0.81) and the lowest balanced accuracy in the 30-40% class (0.88).

Table 4.8: Confusion Matrix Results for Group "Multi-Racial" (ELA)

Class	Sensitivity	Specificity	Pos.Pred.Value	Neg.Pred.Value	F1	Balanced.Accuracy
0-20%	0.00	0.96	0.00	0.98	NaN	0.48
20-30%	0.09	0.94	0.11	0.92	0.10	0.51
30-40%	0.06	0.89	0.07	0.86	0.06	0.47
40-50%	0.65	0.31	0.14	0.84	0.23	0.48
50-60%	0.07	0.98	0.50	0.80	0.12	0.53
60-70%	0.08	0.95	0.25	0.83	0.12	0.52
70-80%	0.05	0.99	0.50	0.86	0.09	0.52
80-90%	0.00	1.00	NaN	0.92	NA	0.50
90-100%	0.00	1.00	NaN	0.99	NA	0.50

The confusion matrix results for the group "Multi-Racial" in the English Language Arts (ELA) proficiency classes (Table 4.8) show varying performance across classes. Sensitivity values are generally low, ranging from 0.00 to 0.65. Specificity values exhibit a broader range, from 0.31 to 1.00. The positive predictive value has NaN (not a number) values for the 80-90% and 90-100% classes, indicating a lack of true positive and false positive cases, making it impossible to calculate the values. Similarly, F1-scores have NA (not available) values for the same classes due to the absence of true positive cases.

The lowest F1-score is observed in the 30-40% class (0.06), and the lowest balanced accuracy values are seen in the 30-40% and 40-50% classes (both 0.47). The balanced accuracy values range between 0.47 and 0.53, indicating relatively weak performance in distinguishing between classes.

Table 4.9: Confusion Matrix Results for Genders (ELA)

Class	Sensitivity	Specificity	Pos.Pred.Value	Neg.Pred.Value	F1	Balanced.Accuracy
0-40%	0.97	0.94	0.83	0.99	0.89	0.96
40-50%	0.80	0.97	0.85	0.96	0.82	0.88
50-60%	0.86	0.96	0.87	0.96	0.86	0.91
60-70%	0.83	0.99	0.96	0.96	0.89	0.91
70-80%	0.94	0.99	0.91	0.99	0.92	0.96
80-90%	0.79	0.99	0.83	0.99	0.81	0.89
90-100%	0.80	1.00	1.00	1.00	0.89	0.90

The confusion matrix results for gender in the English Language Arts (ELA) proficiency classes (Table 4.9) show relatively high performance in most classes. Sensitivity values range from 0.79 to 0.97, and specificity values are consistently high, ranging from 0.94 to 1.00. Positive predictive values and negative predictive values are also strong, with most values above 0.80.

The F1-scores range from 0.81 to 0.92, indicating that the model performs relatively well in classifying students into the right proficiency groups. Balanced accuracy values are consistently high, ranging from 0.88 to 0.96, which suggests that the model effectively differentiates between classes for both genders. Overall, the model demonstrates good performance in predicting ELA proficiency classes based on gender.

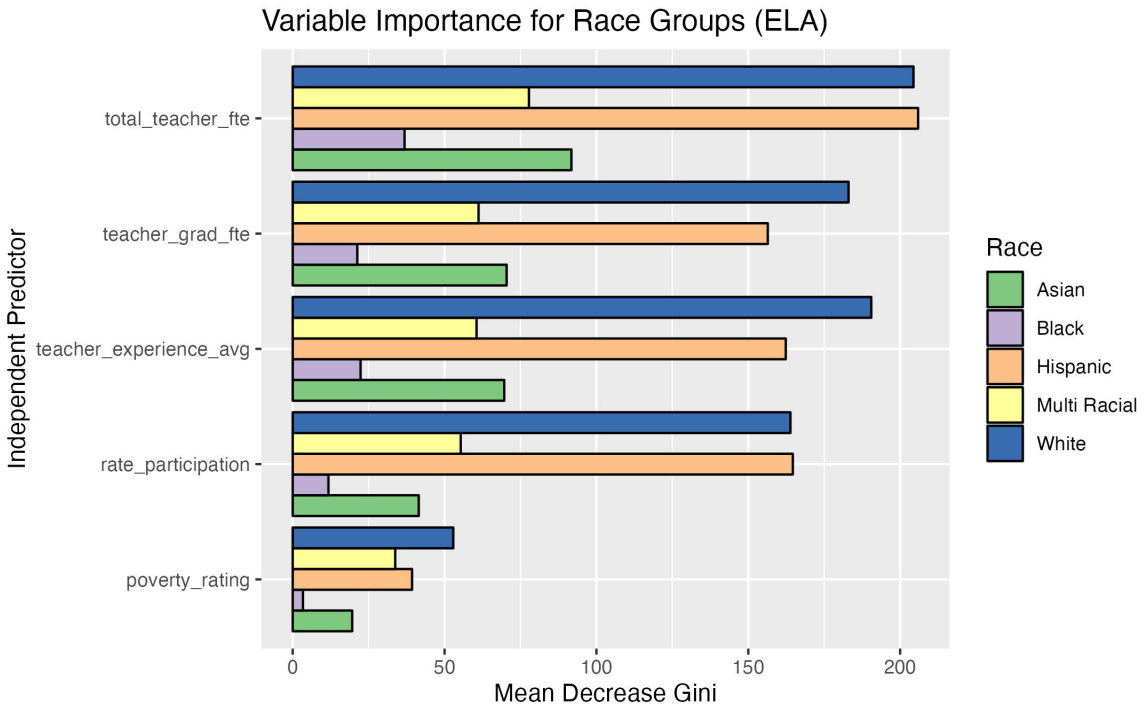

Figure 4.17: Variable Importance Plot (Race, ELA)

This plot presents the variable importance scores for different groups (Asian, Black, Hispanic, Multi Racial, and White) in predicting student proficiency for ELA. Looking at the plot, we can see that for all groups, the variables related to teacher characteristics, including total_teacher_fte, teacher_grad_fte, and teacher_experience_avg, have the highest Mean Decrease Gini (MDG) scores. This suggests that these variables are critical factors that affect the outcome variable for all groups. In particular, total_teacher_fte has the highest MDG scores for all groups except for group Black, where it is second to teacher_grad_fte. On the other hand, the poverty_rating variable has the lowest MDG scores across all groups, suggesting that it is the least important predictor of the outcome variable. The rate_participation variable has relatively lower MDG scores compared to the teacher-related variables for all groups except for Hispanic, where it has a relatively high MDG score.

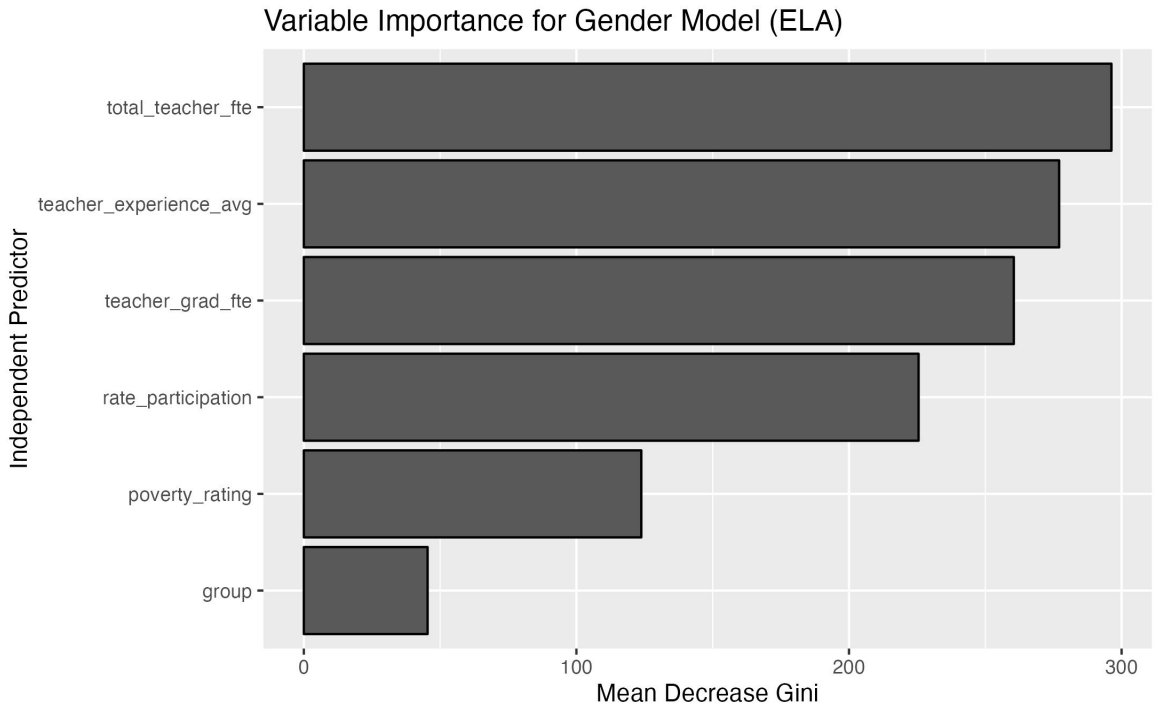

Figure 4.18: Variable Importance Plot (Gender, ELA)

This plot presents the variable importance scores for predicting a proficiency, separated by gender. The MDG scores are provided for each variable, with higher scores indicating more important predictors. The results suggest that the teacher variables are the most important predictors of proficiency for both male and female students. The rate_participation variable is also an important predictor for both genders, while poverty_rating, as measured by the poverty_rating variable, has the least impact on the outcome variable for both genders.

4.3.2 Mathematics

Table 4.10: Confusion Matrix Results for Group "Asian" (Math)

Class	Sensitivity	Specificity	Pos.Pred.Value	Neg.Pred.Value	F1	Balanced.Accuracy
0-40%	0.83	1.00	1.00	0.98	0.91	0.92
40-50%	1.00	1.00	1.00	1.00	1.00	1.00
50-60%	0.71	1.00	1.00	0.95	0.83	0.86
60-70%	1.00	0.81	0.59	1.00	0.74	0.91
70-80%	0.89	1.00	1.00	0.97	0.94	0.94
80-90%	1.00	1.00	1.00	1.00	1.00	1.00
90-100%	0.40	1.00	1.00	0.93	0.57	0.70

The confusion matrix results for the "Asian" group in Math proficiency classes (Table 4.10) demonstrate strong performance in several classes. Sensitivity values range from 0.40 to 1.00, and specificity values are consistently high at 0.81 to 1.00. Both positive predictive values and negative predictive values are mostly high, with most values above 0.90. F1-scores range from 0.57 to 1.00, indicating that the model performs well in several classes but struggles with the 90-100% class. Balanced accuracy values show varying performance, ranging from 0.70 to 1.00. Overall, the model demonstrates strong performance in predicting Math proficiency classes for the "Asian" group, with some room for improvement in the highest proficiency class.

Table 4.11: Confusion Matrix Results for Group "White" (Math)

Class	Sensitivity	Specificity	Pos.Pred.Value	Neg.Pred.Value	F1	Balanced.Accuracy
0-20%	0.72	1.00	1.00	0.98	0.84	0.86
20-30%	0.78	1.00	1.00	0.97	0.88	0.89
30-40%	0.84	0.93	0.77	0.95	0.80	0.88
40-50%	0.87	0.88	0.73	0.95	0.79	0.87
50-60%	0.63	0.99	0.92	0.93	0.75	0.81
60-70%	0.74	0.92	0.47	0.97	0.57	0.83
70-80%	0.60	1.00	1.00	0.98	0.75	0.80
80-100%	0.50	1.00	1.00	1.00	0.67	0.75

The confusion matrix results for the "White" group in Math proficiency classes (Table 4.11) display a variety of performances across classes. Sensitivity values range from 0.50 to 0.87, while specificity values are high, with most values at 0.88 or higher. Both positive predictive values and negative predictive values are generally high, with most values above 0.90.

F1-scores vary from 0.57 to 0.88, indicating a mix of strong and weaker performances across classes. Balanced accuracy values range from 0.75 to 0.89, also reflecting a range of performances. Overall, the model demonstrates a relatively strong performance in predicting Math proficiency classes for the "White" group, but there are a few areas where improvements could be made, particularly in the 60-70% and 80-100% classes.

Table 4.12: Confusion Matrix Results for Group "Black" (Math)

Class	Sensitivity	Specificity	Pos.Pred.Value	Neg.Pred.Value	F1	Balanced.Accuracy
<10%	0.75	1.00	1.00	0.90	0.86	0.88
10-20%	0.92	0.69	0.75	0.90	0.83	0.81
20-30%	0.67	1.00	1.00	0.96	0.80	0.83
30-100%	0.50	0.96	0.50	0.96	0.50	0.73

In the confusion matrix results for the "Black" group in Mathematics (Table 4.12), the Sensitivity values range from 0.50 to 0.92, while Specificity values range from 0.69 to 1.00, indicating good model performance. The F1 scores range from 0.50 to 0.86, and the balanced accuracy values range from 0.73 to 0.88, which further supports the strong performance of these models. It's important to note that the 30-100% class exhibits a lower Sensitivity and F1 score compared to the other classes, which may indicate room for improvement in this particular class.

Table 4.13: Confusion Matrix Results for Group "Hispanic" (Math)

Class	Sensitivity	Specificity	Pos.Pred.Value	Neg.Pred.Value	F1	Balanced.Accuracy
<10%	0.11	1.00	1.00	0.96	0.20	0.56
10-20%	0.80	0.96	0.84	0.95	0.82	0.88
20-30%	0.99	0.57	0.55	0.99	0.71	0.78
30-40%	0.47	0.99	0.89	0.89	0.62	0.73
40-50%	0.27	1.00	1.00	0.91	0.43	0.64
50-60%	0.40	0.99	0.67	0.97	0.50	0.69
60-70%	0.00	1.00	NaN	0.99	NA	0.50
80-100%	0.00	1.00	NaN	0.98	NA	0.50

The confusion matrix results for the "Hispanic" group in Mathematics (Table 4.13) display varying performance for different proficiency classes with sensitivity values range from 0.00 to 0.99, with the 20-30% class exhibiting the highest Sensitivity (0.99), while the 60-70% and 80-100% classes have the lowest (0.00). Specificity values are generally high across all classes, with most classes achieving a value of 1.00, indicating that the models perform well in correctly identifying students not belonging to these specific classes.

However, the Pos.Pred.Value and F1 scores reveal areas for improvement. For the 60-70% and 80-100% classes, the Pos.Pred.Value is not applicable (NaN) due to the absence of true positive cases, suggesting that the model might not be accurately predicting these classes. Additionally, the F1 scores range from 0.20 to 0.82, with the lowest value observed for the less than 10% class (0.20), indicating that the models may struggle to predict this class effectively. The Balanced Accuracy values range from 0.50 to 0.88, reflecting the overall mixed performance of the model in some classes.

Table 4.14: Confusion Matrix Results for Group "Multi-Racial" (Math)

Class	Sensitivity	Specificity	Pos.Pred.Value	Neg.Pred.Value	F1	Balanced.Accuracy
0-20%	0.45	1.00	1.00	0.95	0.62	0.73
20-30%	0.43	1.00	1.00	0.90	0.60	0.71
30-40%	0.43	0.98	0.87	0.85	0.58	0.71
40-50%	0.92	0.41	0.28	0.96	0.42	0.66
50-60%	0.35	1.00	1.00	0.88	0.52	0.67
60-70%	0.17	0.97	0.33	0.92	0.22	0.57
70-80%	0.29	1.00	1.00	0.96	0.44	0.64
80-100%	0.00	1.00	NaN	0.98	NA	0.50

The confusion matrix results for the "Multi-Racial" group in Mathematics (Table(4.14)) reveal that the single random forest model exhibits mixed performance across different proficiency classes. Sensitivity values range from 0.00 to 0.92, with the 40-50% class achieving the highest Sensitivity (0.92), whereas the 80-100% class has the lowest (0.00). Specificity values are generally high, with most classes reaching a value of 1.00, suggesting that the model performs well in correctly identifying students who do not belong to these specific classes.

However, some areas of the model's performance need improvement. The Pos.Pred.Value is not applicable (NaN) for the 80-100% class due to the absence of true positive cases, which implies that the model may not accurately predict this class. Additionally, F1 scores range from 0.22 to 0.62, with the lowest value seen for the 60-70% class (0.22), indicating that the model's performance in predicting this class is relatively weak. Balanced Accuracy values vary between 0.50 and 0.73, which further highlights the model's inconsistent performance across different proficiency classes for the Multi-Racial group in Mathematics.

Table 4.15: Confusion Matrix Results for Genders (Math)

	Sensitivity	Specificity	Pos.Pred.Value	Neg.Pred.Value	F1	Balanced.Accuracy
0-10%	0.90	1.00	1.00	1.00	0.95	0.95
10-20%	0.78	1.00	0.95	0.98	0.85	0.89
20-30%	0.86	0.95	0.82	0.96	0.84	0.90
30-40%	0.88	0.94	0.81	0.96	0.85	0.91
40-50%	0.81	0.97	0.85	0.96	0.83	0.89
50-60%	0.85	0.97	0.82	0.98	0.83	0.91
60-70%	0.81	0.99	0.83	0.99	0.82	0.90
70-80%	0.79	1.00	0.92	0.99	0.85	0.89
80-90%	1.00	1.00	1.00	1.00	1.00	1.00
90-100%	1.00	1.00	1.00	1.00	1.00	1.00

The confusion matrix results for the Genders group in Mathematics (Table (4.9)) show that the random forest model generally performs well across different proficiency classes. Sensitivity values range from 0.78 to 1.00, indicating a high true positive rate for most classes. Specificity values are also consistently high, with most classes achieving a value of 1.00, suggesting that the model effectively identifies students who do not belong to these specific classes.

Positive Predictive Value (Pos.Pred.Value) and Negative Predictive Value (Neg.Pred.Value) are also strong, with values close to or equal to 1.00 for most classes, indicating accurate predictions for both positive and negative cases. F1 scores range from 0.82 to 1.00, demonstrating the model's balanced performance in terms of precision and recall. Balanced Accuracy values range from 0.89 to 1.00, reflecting the model's overall strong performance across different proficiency classes for the Genders group in Mathematics.

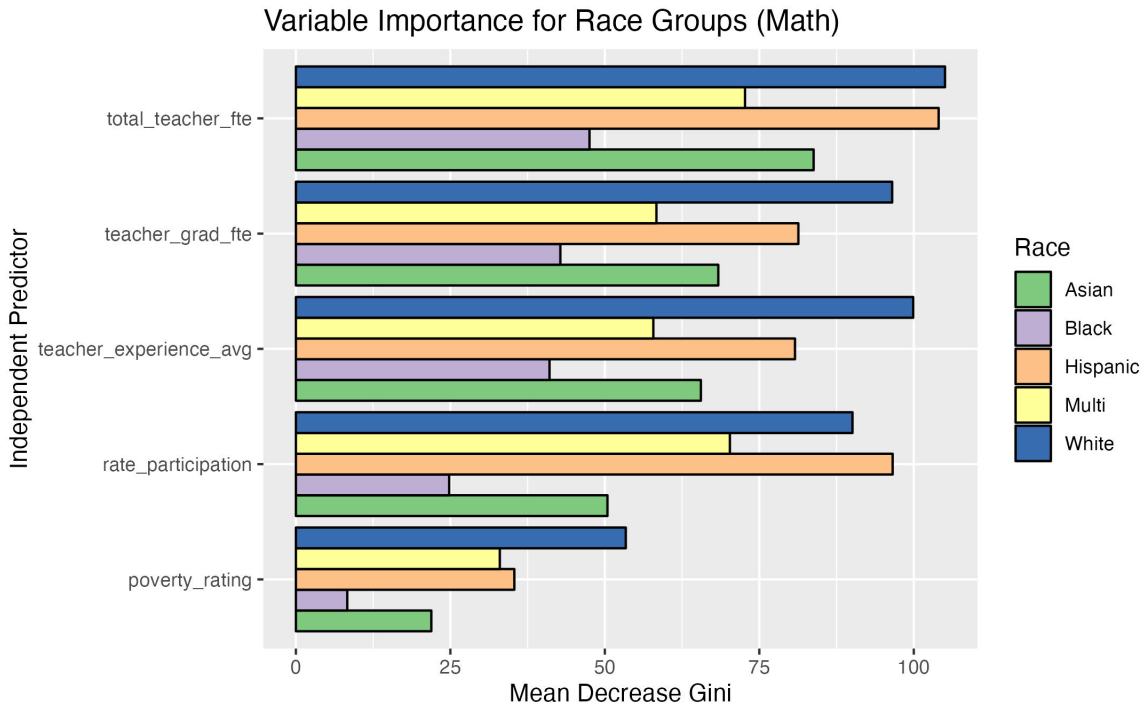

Figure 4.19: Variable Importance Plot (Race, Math)

This plot provides the variable importance scores for different races (Asian, Black, Hispanic, Multi, and White) in predicting proficiency. Looking at the plot, we can see that for all groups, the variables related to teacher characteristics, including total_teacher_fte, teacher_grad_fte, and teacher_experience_avg, have the highest MDG scores. On the other hand, the poverty_rating variable has the lowest MDG scores across all groups, suggesting that it is the least important predictor of profiency. The rate_participation variable has relatively lower MDG scores compared to the teacher-related variables for all groups except for Hispanic, where it has a relatively high MDG score.

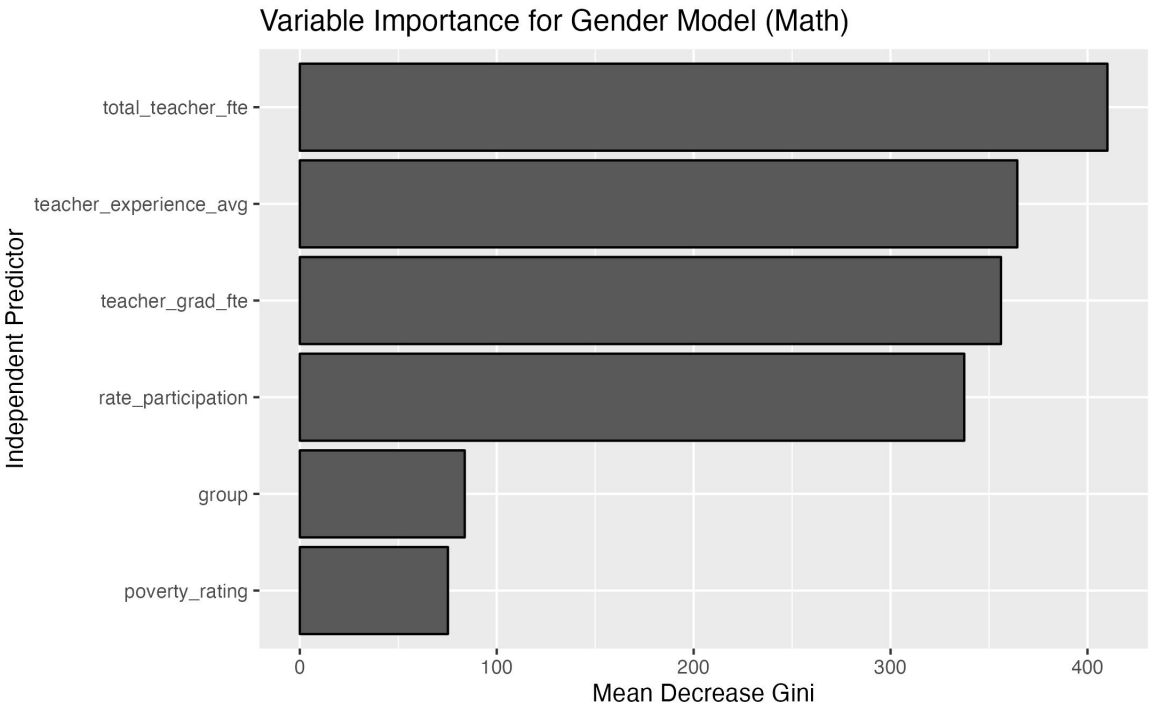

Figure 4.20: Variable Importance Plot (Gender, Math)

This plot shows the variable importance scores for predicting student proficiency for Math for the gender group. We see that the variables with the highest MDG scores are total_teacher_fte, teacher_grad_fte, and teacher_experience_avg. The rate_participation variable has a relatively high MDG score for both genders, although lower than the teacher-related variables. The group and poverty_rating variables have the lowest MDG score, similar to the findings in the ELA case.

Chapter 5

Discussion

5.1 Results of Random Forest

The random forest models used to predict student proficiency in both English Language Arts (ELA) and Mathematics for different racial groups and genders have demonstrated varying levels of performance. For most racial groups and both genders, the models exhibit strong performance in predicting proficiency levels, with high Sensitivity, Specificity, Positive Predictive Value (Pos.Pred.Value), Negative Predictive Value (Neg.Pred.Value), F1 scores, and Balanced Accuracy values. This indicates that the models have the potential to accurately identify and classify students' proficiency levels in ELA and Mathematics across a range of demographic factors.

However, some discrepancies in performance have been observed, particularly for the Multi-racial group. In this group, the random forest models exhibit lower accuracy and performance metrics compared to the other racial groups. This inconsistency may be due to the complexity and diversity of the Multi-racial group, which could pose challenges for the model in accurately predicting proficiency levels. Further investigation and model refinement may be necessary to improve the model's performance for this particular group.

Another area of interest is the varying performance across different proficiency classes within each racial group or gender. In some cases, the models demonstrate strong performance in predicting certain classes while struggling with others. This observation suggests that additional fine-tuning and feature engineering may be needed to enhance the model's ability to accurately predict proficiency levels across all classes.

In conclusion, the random forest models show promise in their ability to predict student proficiency in ELA and Mathematics for various racial groups and both genders. Nevertheless, there is room for improvement, particularly in addressing the performance disparities observed for the Multi-racial group and specific proficiency classes. Future research should focus on refining the models and investigating potential strategies for enhancing their predictive accuracy and generalizability across different demographic factors and proficiency levels. This could ultimately contribute to a more comprehensive understanding of student proficiency and support more effective educational interventions tailored to the diverse needs of students.

5.2 Decision Trees vs Random Forest

The results of our analysis indicate that the random forest model outperforms the decision tree model in all cases. The accuracy of the random forest model is consistently higher than that of the decision tree model for all subgroups of ethnicity, gender, and subject (ELA and Math). This finding suggests that the random forest model is better suited for predicting the percentage proficient score for Oregon public schools. The superiority of the random forest model can be attributed to its ability to reduce overfitting and handle high-dimensional data with complex interactions among variables.

Furthermore, the higher accuracy of the random forest model can be attributed to its use of multiple decision trees and the averaging of their predictions. This approach enables the model to reduce variance and avoid overfitting while maintaining a low bias. In contrast, the decision tree model only uses a single tree, which leads to overfitting and high variance.

However, it is important to note that for the random forest models to work the levels had to be combined due to low availability of data for certain proficiency levels which loses nuance for certain groups that have lower data availability. This is an issue that the decision trees also struggle with. They tend to overfit to the majority classes in all cases pertaining to lower accuracy. However, an advtange decision trees have over random forest models is visual interpretability. They provide a clear visualization of the decision-making process, showing which variables are important and how they contribute to the final decision.

5.3 Variable Importance

The variable importance analysis shows that total teacher FTE, percentage FTE for teachers with a graduate degree, and average teacher experience are more important in predicting student proficiency in both ELA and math compared to poverty rating and participation rate. This highlights the importance of having high-quality teachers in schools, as they have a significant impact on student learning outcomes. The results suggest that policymakers should focus on increasing the number of experienced teachers with graduate degrees and encouraging higher rates of FTE for teachers with these qualifications in order to improve student proficiency.

Interestingly, gender did not appear to play a significant role in predicting student proficiency in either subject. This may suggest that gender equity in hiring practices is already being achieved, or that other factors such as teacher qualifications and experience have a more significant impact on student learning outcomes.

It is important to note that these results are based on the 2018-19 dataset and methodology used in this study, and may not necessarily apply to other contexts. Additionally, while variable importance analysis provides insight into which variables are most influential in predicting outcomes, it does not provide information on causality or directionality. Further research is needed to fully understand the relationship between these variables and student proficiency.

5.4 Missing Data

As mentioned, in this study, we faced several issues with the availability of data. The ESSA data collection practices are not standardized across states, which leads to missing data points, especially in less populated areas. Moreover, many schools may not provide data due to privacy concerns or other reasons, leading to incomplete data (Briones, 2019) . As a result, we had to use various rough fixes to handle the missing values, which might have affected the accuracy of our models.

One of the most common ways to handle missing data is to use mean imputation or median imputation as we do by rough fixing the dataset, where the missing values are replaced with the mean or median value of the respective variable. However, this method can introduce bias and reduce the variability of the data. Another approach is to delete observations or variables that have missing values, but this can lead to a loss of information and reduce the sample size (Peng et al., 2006).

Despite the challenges associated with handling missing data, our results still show that the model can predict student proficiency reasonably well, except in the case of multi-racial students for ELA. Future studies can improve data collection practices to reduce missing data or explore more advanced imputation methods to handle missing values. Additionally, collecting more detailed data on students and schools, such as student attendance, teacher-student ratios, and parental involvement, can provide a more comprehensive understanding of the factors that affect student proficiency (Peng et al., 2006).

Conclusion

The machine learning model developed in this study successfully predicted the percentage proficient score for Oregon public schools using a variety of data sources including teacher FTE, experience, participation rates, poverty rating, ethnicity, and gender. Variable importance plots showed that teacher FTE and experience were the most important predictors of proficiency for both ELA and math, while poverty rating and participation rates had a slightly lower importance, which is in tune with studies conducted previously. Gender also did not play a significant role in predicting proficiency.

Confusion matrices were used to evaluate the model's performance, with overall accuracy rates above 80% for most demographic groups, except for the multi-racial group, where accuracy rates were around 50%. These results suggest that the model is reliable and effective in predicting proficiency scores for most groups and perform better than decision trees in every case, but may need further refinement by using more descriptive predictors and more available data.

Overall, the findings of this study suggest that machine learning models, particularly random forests, can be effective tools for predicting academic performance in public schools. Further research is needed to address the challenges faced in this study and to refine the model for greater accuracy. However, if such models can be improved, they could have important implications for identifying and addressing educational disparities and improving student outcomes.

Appendix A

Creating Visualizations

This section presents code used to make the explanatory plots for all groups for proficiency along with summary tables for the dependent variables. To make the plots the study uses the R ggplot function and for the summaries it uses the summary() function.

```
      ```{r}
level_order <- c('<␣10%', '10-20%', '20-30%', '30-40%','40-50%','50-60%','60-70%','70-80
```

#Exploratory plots

##ELA
```{r}
plot_prof_ethnicity_ELA <- Ethnicity_All_ELA %>%
  drop_na(group, proficiency) %>%
  ggplot(aes(x = factor(proficiency, level = level_order))) + geom_bar(aes(y = (..count.
plot_prof_ethnicity_ELA
ggsave("~/Desktop/plot_prof_ethnicity_ELA.jpg")
```

```{r}
plot_prof_gender_ELA <- Gender_ELA %>%
  drop_na(group, proficiency) %>%
  ggplot(aes(x = factor(proficiency, level = level_order))) + geom_bar(aes(y = (..count.
plot_prof_gender_ELA
ggsave("~/Desktop/plot_prof_gender_ELA.jpg")
```

##Ethnicity Math
```{r}
plot_prof_ethnicity_Math <- Ethnicity_All_Math %>%
  drop_na(group, proficiency) %>%
  ggplot(aes(x = factor(proficiency, level = level_order))) + geom_bar(aes(y = (..count.
```
```

```
plot_prof_ethnicity_Math
ggsave("~/Desktop/plot_prof_ethnicity_Math.jpg")
```

```{r}
plot_prof_gender_Math <- Gender_Math %>%
  drop_na(group, proficiency) %>%
  ggplot(aes(x = factor(proficiency, level = level_order))) + geom_bar(aes(y = (..count..)/tapp
plot_prof_gender_Math
ggsave("~/Desktop/plot_prof_gender_Math.jpg")
```
#Summary Statistics ELA

##ethnicity
```{r}
summary_dat_ethnicity_ELA <- Ethnicity_All_ELA %>%
  subset(select = c(school, group, num_participants, proficiency,teacher_experience_avg, teache
  mutate(num_participants = as.numeric(num_participants))

summary_dat_gender_ELA <- Gender_ELA %>%
  subset(select = c(school, group, num_participants, proficiency,teacher_experience_avg, teache
  mutate(num_participants = as.numeric(num_participants))
```

```{r}
rp_summary_ethnictiy_ELA <- summary_dat_ethnicity_ELA %>%
  group_by(group) %>%
  drop_na(rate_participation) %>%
  summarize(avg_rp = mean(rate_participation))

count_by_ethnicity_ELA <- aggregate(summary_dat_ethnicity_ELA[, c("num_participants")],
                                    by = list(summary_dat_ethnicity_ELA$group), sum)

avg_num_participants_eth_ela <- count_by_ethnicity_ELA
avg_num_participants_eth_ela$x <- (avg_num_participants_eth_ela$x)/1236

summary(summary_dat_ethnicity_ELA)
```

##gender
```{r}
rp_summary_gender_ela <- summary_dat_gender_ELA %>%
  group_by(group) %>%
  drop_na(rate_participation) %>%
  summarize(avg_rp = mean(rate_participation))

count_by_gender_ELA <- aggregate(summary_dat_gender_ELA[, c("num_participants")],
```

```
                                by = list(summary_dat_gender_ELA$group), sum)

avg_num_participants_gen_ela <- count_by_gender_ELA
avg_num_participants_gen_ela$x <- (avg_num_participants_gen_ela$x)/1236

summary(summary_dat_gender_ELA)
```

```
#Summary Statistics Math

##ethnicity
```{r}
summary_dat_ethnicity_math <- Ethnicity_All_Math %>%
  subset(select = c(school, group, num_participants, proficiency,teacher_experience_avg,
  mutate(num_participants = as.numeric(num_participants))

summary_dat_gender_math <- Gender_Math %>%
  subset(select = c(school, group, num_participants, proficiency,teacher_experience_avg,
  mutate(num_participants = as.numeric(num_participants))
```

```{r}
rp_summary_ethnictiy_math <- summary_dat_ethnicity_math %>%
  group_by(group) %>%
  drop_na(rate_participation) %>%
  summarize(avg_rp = mean(rate_participation))
latex_rp_summary_ethnictiy_math <- xtable(rp_summary_ethnictiy_math)

count_by_ethnicity_math <- aggregate(summary_dat_ethnicity_math[, c("num_participants")]
                                by = list(summary_dat_ethnicity_math$group), sum)


avg_num_participants_eth_math <- count_by_ethnicity_math
avg_num_participants_eth_math$x <- (avg_num_participants_eth_math$x)/1236

latex_avg_num_participants_eth_math <- xtable(avg_num_participants_eth_math)
```

##gender
```{r}
rp_summary_gender_math <- summary_dat_gender_math %>%
  group_by(group) %>%
  drop_na(rate_participation) %>%
  summarize(avg_rp = mean(rate_participation))

count_by_gender_math <- aggregate(summary_dat_gender_math[, c("num_participants")],
                                by = list(summary_dat_gender_math$group), sum)
```
```

```
avg_num_participants_gen_math <- count_by_gender_math
avg_num_participants_gen_math$x <- (avg_num_participants_gen_math$x)/1236

latex_avg_num_participants_gen_math <- xtable(avg_num_participants_gen_math)

summary_dat <- summary_dat_gender_math %>%
  filter(group == "Male") %>%
  subset(select = c(teacher_experience_avg, teacher_grad_fte, total_teacher_fte, poverty_rating
summ_latex <- xtable(summary(summary_dat))
```

Appendix B

Model Fitting

This section presents R chunks from code that was described in the methods section. The code is from group "Asian" to model proficiency for Mathematics and is annotated with "" wherever needed. The differences that arise as part of the code: "group" predictor is added to the models when modeling for Gender groups, there is a change in levels for different different groups based on availability of proficiency values.

```
     ```{r}
#calculate accuracy for decision trees

calculate_accuracy <- function(tbl) {
  diag_sum <- sum(diag(tbl))
  total_sum <- sum(tbl)
  accuracy <- diag_sum / total_sum
  return(signif(accuracy, digits = 2))
}
```

```{r}
#Train Test set creation using stratified sampling

create_train_test_strat <- function(data, size = 0.8, train = TRUE){

  trainind <- sample(1:nrow(data),
                     nrow(data)*size,
                     replace = FALSE,
                     prob = rep(1/nrow(data), nrow(data)))

  list(data[trainind,], data[-trainind,])

}

```
```

```{r}
#summary function for further tuning and summary abilities
MySummary <- function(data, lev = NULL, model = NULL){

  a1 <- defaultSummary(data, lev, model)
  b1 <- multiClassSummary(data, lev, model)

  out <- c(a1, b1)
  out

  }
```

#Modeling Group "Asian"

```{r}
set.seed(1)
#create train and test sets (stratified sampling)
Asian_Math <- Ethnicity_All_Math %>%
  filter(group == "Asian")
Asian_Math_train <- create_train_test_strat(Asian_Math)[[1]]
Asian_Math_test <- create_train_test_strat(Asian_Math)[[2]]
```

```{r}
set.seed(1)
tree_asian_Math <- rpart(proficiency ~ teacher_grad_fte + teacher_experience_avg + rate_partici
rpart.plot(tree_asian_Math)
predict_unseen <-predict(tree_asian_Math, Asian_Math_test, type = "class")
table_mat <- table(Asian_Math_test$proficiency, predict_unseen)
accuracy_asian_math_tree <- calculate_accuracy(table_mat)
```

```{r}
set.seed(1)
Asian_Math_train_fixed <- na.roughfix(Asian_Math_train) #roughfix for training set
Asian_Math_train_fixed$proficiency <- factor(Asian_Math_train_fixed$proficiency)
#create untuned randomForest model
bag_asian_Math <- randomForest(proficiency ~ teacher_grad_fte + total_teacher_fte + teacher_exp
```

```{r}
#plot Out Of Bag error rate with number of trees
err.dat.asian <- bag_asian_Math[["err.rate"]]
err.dat.asian <- as.data.frame(err.dat.asian)
ggplot(err.dat.asian, aes(x = 1:500, y = OOB)) + geom_line()
```

```
```

```{r}
err.dat.asian$tree <- 1:500
err.dat.asian.long <- gather(err.dat.asian, key = "type", value = "error", 1:10)
ggplot(err.dat.asian.long, aes(x = tree, y = error)) + geom_line() + facet_wrap(~type)
```

```{r}
importance_asian <- bag_asian_Math[["importance"]]
importance_asian <- as.data.frame(importance_asian)
```

```{r}
#Create plot to understand variable importance (Mean Decrease Gini)
ib_mdg_asian <- data.frame(var = row.names(importance_asian), mdg = importance_asian$Mea
ggplot(data = ib_mdg_asian, aes(x = mdg, y = reorder(var, mdg))) + geom_col()
```

```{r}
#Create plot to understand variable importance (Mean Decrease Accuracy)
ib_mda_asian <- data.frame(var = row.names(importance_asian), mda = importance_asian$Mea
ggplot(data = ib_mda_asian, aes(x = mda, y = reorder(var, mda))) + geom_col()
```

```{r}
#Make character variable for proficiency
Asian_Math_train_fixed_1 <- Asian_Math_train_fixed %>%
  mutate(proficiency_chr = as.character(proficiency))
Asian_Math_train_fixed_1 <- within(Asian_Math_train_fixed_1, {
  proficiency_chr[proficiency_num < 40] <- "lev0to3"
  proficiency_chr[proficiency_num >= 40 & proficiency_num < 50] <- "lev4"
  proficiency_chr[proficiency_num >= 50 & proficiency_num < 60] <- "lev5"
  proficiency_chr[proficiency_num >= 60 & proficiency_num < 70] <- "lev6"
  proficiency_chr[proficiency_num >= 70 & proficiency_num < 80] <- "lev7"
  proficiency_chr[proficiency_num >= 80 & proficiency_num < 90] <- "lev8"
  proficiency_chr[proficiency_num >= 90 & proficiency_num < 101] <- "lev9"
  })
```

```{r}
#Build matching test set
Asian_Math_test_clean<- Asian_Math_test
Asian_Math_test_clean <- Asian_Math_test_clean %>%
  mutate(proficiency_chr = as.character(proficiency))
Asian_Math_test_clean <- within(Asian_Math_test_clean, {
  proficiency_chr[proficiency_num < 40] <- "lev0to3"
```

```
  proficiency_chr[proficiency_num >= 40 & proficiency_num < 50] <- "lev4"
  proficiency_chr[proficiency_num >= 50 & proficiency_num < 60] <- "lev5"
  proficiency_chr[proficiency_num >= 60 & proficiency_num < 70] <- "lev6"
  proficiency_chr[proficiency_num >= 70 & proficiency_num < 80] <- "lev7"
  proficiency_chr[proficiency_num >= 80 & proficiency_num < 90] <- "lev8"
  proficiency_chr[proficiency_num >= 90 & proficiency_num < 101] <- "lev9"
  })


Asian_Math_test_clean <- Asian_Math_test_clean %>%
  filter(!is.na(teacher_grad_fte), !is.na(teacher_experience_avg), !is.na(total_teacher_fte), !
        !is.na(proficiency_chr), !is.na(rate_participation))
```

```{r}
#extract ntree with minimum OOB error rate
min_row <- err.dat.asian[which.min(err.dat.asian$OOB), ]
nTree <- min_row$tree
```

```{r, cache=TRUE}
#using train function from 'caret' to find best tuning for final model
asian_rf_train <- train(proficiency_chr ~ teacher_grad_fte + total_teacher_fte + teacher_experi
                        rate_participation + poverty_rating, data = Asian_Math_train_fixed_1, m
                      trControl = trainControl(method = "cv", number = 9, classProbs = TRUE,
                      summaryFunction = MySummary, sampling = "smote"), ntree = nTree,
                      tuneGrid = data.frame(mtry = c(1:5)), metric = "Accuracy")
```

```{r}
#final best model from 'train'
best_tune <- asian_rf_train\$bestTune

asian_rf <- randomForest(as.factor(proficiency_chr) ~ teacher_grad_fte +
                           total_teacher_fte + teacher_experience_avg + rate_participation + pov
                         data = Asian_Math_train_fixed_1, ntree = nTree, mtry = best_tune$mtry)
```

```{r}
predictions <- predict(asian_rf, Asian_Math_test_clean)
cm_asian_math <- confusionMatrix(predictions, as.factor(Asian_Math_test_clean\$proficiency_chr)
cm_asian_math
```

```{r}
#used to provide the variable importance plot for the final model
```

```
varImpPlot(asian_rf)
```

References

[1] Anthony Kulesa, Martin Krzywinski, Paul Blainey, and Naomi Altman. Sampling distributions and the bootstrap. 12(6):477–478.

[2] Philipp Probst, Marvin N. Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. 9(3):e1301. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1301.

[3] Tzu-Ling Lai. Effects of student-teacher congruence on students' learning performance: A dyadic approach. 96(5):1424–1435.

[4] Ann Owens, Sean F. Reardon, and Christopher Jencks. Income segregation between schools and school districts. 53(4):1159–1197.

[5] John Conlisk. Determinants of school enrollment and school performance. 4(2):140–157.

[6] R Marc Brodersen, Douglas Gagnon, Jing Liu, and Tony Moss. Steps to develop a model to estimate school- and district-level postsecondary success.

[7] Gary L. Marco, Richard T. Murphy, and Thomas J. Quirk. A classification of methods of using student data to assess school effectiveness. 13(4):243–252.

[8] Daniel F. McCaffrey and J. R. Lockwood. Missing data in value-added modeling of teacher effects. 5(2):773–797.

[9] Everett Weber. Quantifying student learning: How to analyze assessment data. 90(4):501–511.

[10] Jennifer Briones. Data and the every student succeeds act (ESSA).

[11] Eric A. Hanushek. Building on no child left behind. 326(5954):802–803.

[12] Daniel Koretz. Moving past no child left behind. 326(5954):803–804.

[13] Jennifer L. Jennings and Jonathan Marc Bearak. "teaching to the test" in the NCLB era: How test predictability affects our understanding of student performance. 43(8):381–389.

[14] Zacharoula Papamitsiou and Anastasios A. Economides. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. 17(4):49–64.

[15] Ali Salah Hashim, Wid Akeel Awadh, and Alaa Khalaf Hamoud. Student performance prediction model based on supervised machine learning algorithms. 928(3):032019.

[16] Comfort O. Okpala. Educational resources, student demographics and achievement scores. 27(3):885–907.

[17] Paul E Peterson and Matthew Ackerman. States raise proficiency standards in math and reading.

[18] Mathew D. Knepper. Shooting for the moon: The innocence of the no child left behind act's one hundred percent proficiency goal and its consequences teaching federal courts - comment. 53(3):899–926.

[19] Sean Flaherty. Does money matter in pennsylvania? school district spending and student proficiency since no child left behind. 39(2):145–171.

[20] Andrew Dean Ho. The problem with "proficiency": Limitations of statistics and policy under no child left behind. 37(6):351–360.

[21] Eric Haas, Glen Wilson, Casey Cobb, and Sharon Rallis. One hundred percent proficiency: A mission impossible. 38:180–189.

[22] William F Tate Iv. Race, SES, gender, and language proficiency trends in mathematics achievement: An update.

[23] Dana Uerz, Monique Volman, and Marijke Kral. Teacher educators' competences in fostering student teachers' proficiency in teaching and learning with technology: An overview of relevant research literature. 70:12–23.

[24] Simon Cassidy and Peter Eachus. Learning style, academic belief systems, self-report student proficiency and academic achievement in higher education. 20(3):307–322.

[25] Jacob Kola Aina, Alexander Gbenga Ogundele, and Shola Sunday Olanipekun. Students' proficiency in english language relationship with academic performance in science and technical education.

[26] Marcus Winters and Joshua Cowen. Grading new york: Accountability and student proficiency in america's largest school district.

[27] Scott W. Rogers and Recep K. Goktas. Exploring engineering graduate student research proficiency with student surveys. 99(3):263–278.

[28] Oregon department of education.

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. 16:321–357.

[30] Fariha Sohil, Muhammad Umair Sohali, and Javid Shabbir. An introduction to statistical learning with applications in r: by gareth james, daniela witten, trevor hastie, and robert tibshirani, new york, springer science and business media, 2013, $41.98, eISBN: 978-1-4614-7137-7. 6(1):87–87.

[31] Keerthana Buvaneshwaran. Decision tree vs random forest (10 differences).

[32] Murray J. Fisher and Andrea P. Marshall. Understanding descriptive statistics. 22(2):93–97.

[33] Alireza Abbasi, Jörn Altmann, and Liaquat Hossain. Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. 5(4):594–607.

[34] Edynn Sato, Rachel Lagunoff, and Peter Worth. SMARTER balanced assessment consortium common core state standards analysis: Eligible content for the summative assessment. final report.

[35] Nora Gordon. How state ESSA accountability plans can shine a statistically sound light on more students.

[36] Ishtiaque Fazlul, Cory Koedel, and Eric Parsons. Free and reduced-price meal eligibility does not measure student poverty: Evidence and policy significance.

[37] Anjaneyulu Babu Shaik and Sujatha Srinivasan. A brief survey on random forest ensembles in classification model. In Siddhartha Bhattacharyya, Aboul Ella Hassanien, Deepak Gupta, Ashish Khanna, and Indrajit Pan, editors, *International Conference on Innovative Computing and Communications*, Lecture Notes in Networks and Systems, pages 253–260. Springer. Place: Singapore.

[38] Anthony Kulesa, Martin Krzywinski, Paul Blainey, and Naomi Altman. Sampling distributions and the bootstrap. 12(6):477–478.

[39] Philipp Probst, Marvin N. Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. 9(3):e1301.