Predicting The Academic Performance Of American Students

BA305 Team B7

[redacted]
& Deanna Soukhaseum

Meet The Team

[redacted]

[redacted]

[redacted]

Deanna Soukhaseum

Finance & BA

Presentation Agenda

Introduction 01

Data Overview 02

Visualization 03

04 Modeling

05 Challenges

06 Conclusion

Problem: Declining Academic Performance in the U.S.

Why did we decide to focus on academic performance?

- The trend of **growing gaps between high and low performing students** began over a decade ago and worsened due to the effects of the pandemic
- In 2024, low performers **scored 100 points below high performers** on a 500 point scale
- **33%** of 8th graders are **below the NAEP Basic reading level** and are unable to identify basic literary elements (i.e. order of events, character traits, main ideas, etc.)

How can prediction be useful for solving this educational crisis?

- Through prediction modeling, we can help government departments and educational institutions **determine which factors lead to academic success**
- Project aims to figure out how to **maximize academic performance** in underperforming students **measured by exam scores**

The Student Dataset

11 Variables

2,133 Records

No Missing Values

Our Data Cleaning & Preparation Process

- Removed 3 variables due to unnecessary or overlapping information: “student_id,” “classroom,” and “school”
- Transformed categorical variables into dummy variables and binarized them using OneHotEncoder
- Divided the dataset into 60% training data and 40% testing data
- Standardized the variables in the train and test data used in all of our models to maintain a consistent scale across variables
- Resulted in a final dataset of 14 columns

Correlation Matrix

Pretest scores had the **strongest correlation** by far to our target variable at **0.95**

Strongest Positive Predictors Of Posttest

- Qualifies for reduced/free lunch: +0.6
- Suburban school setting: +0.45

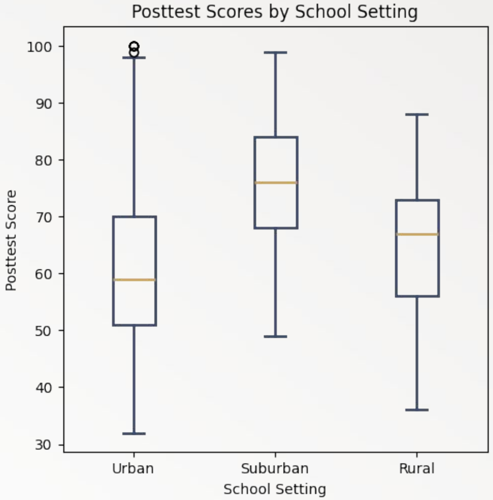
Strongest Negative Predictors

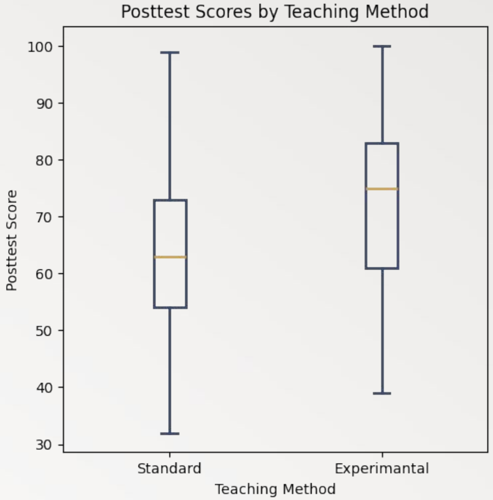
- Class size (n_students): −0.5
- Urban setting: −0.33

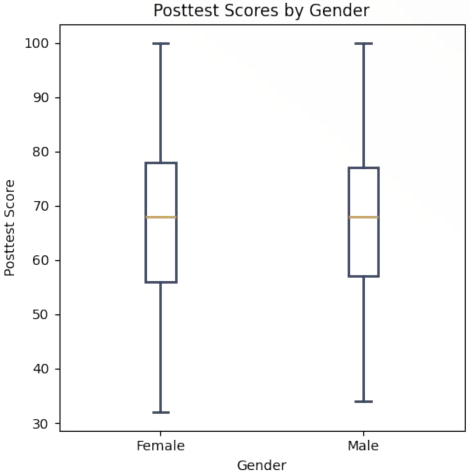
Categorical Variables

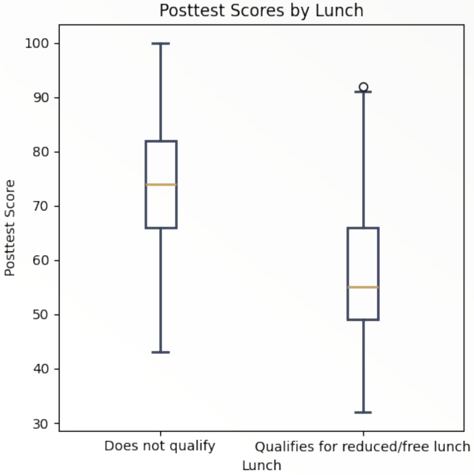
Finding 1: Students perform better in suburban school settings compared to rural and urban school settings

Finding 2: Experimental teaching methods correlate with higher test scores than standard methods





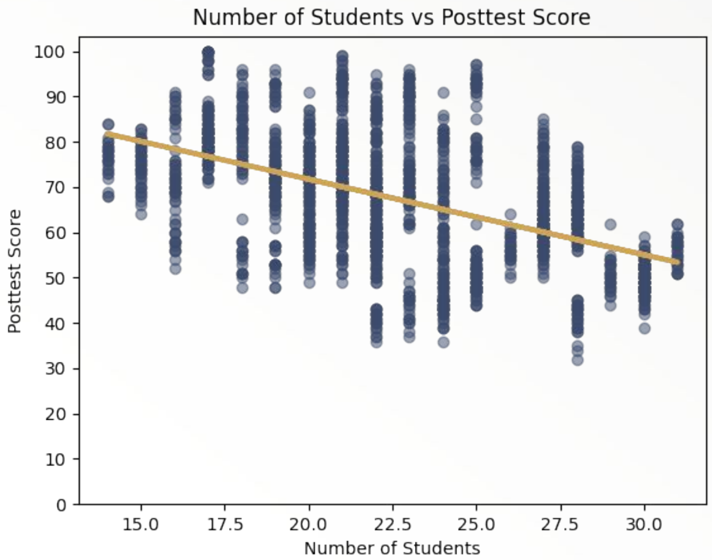


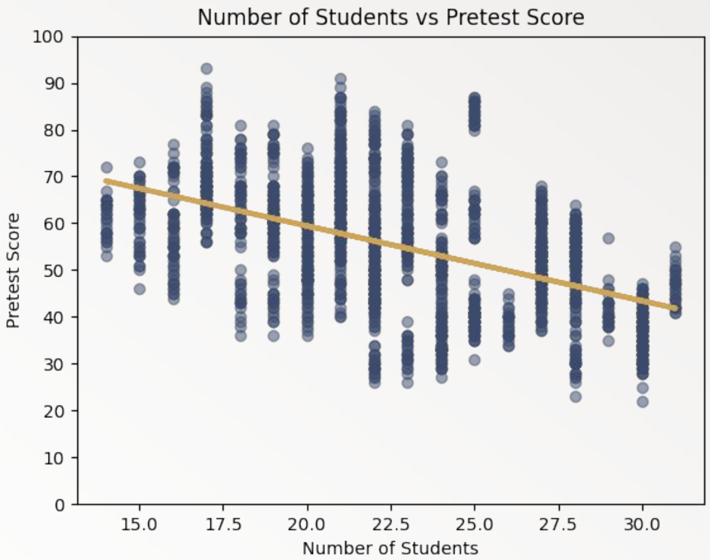


Finding 3: Minimal differences exist between the distribution of test scores of male and female students

Finding 4: Students from a higher socioeconomic background (indicated by subsidized school lunch qualification) tend to have higher scores

Numeric Variables





Finding 5: There is a negative relationship between class size and posttest performance, as well as class size and pretest performance

Finding 6: There is less of a disparity across different class sizes for students' grades before taking the exam compared to after

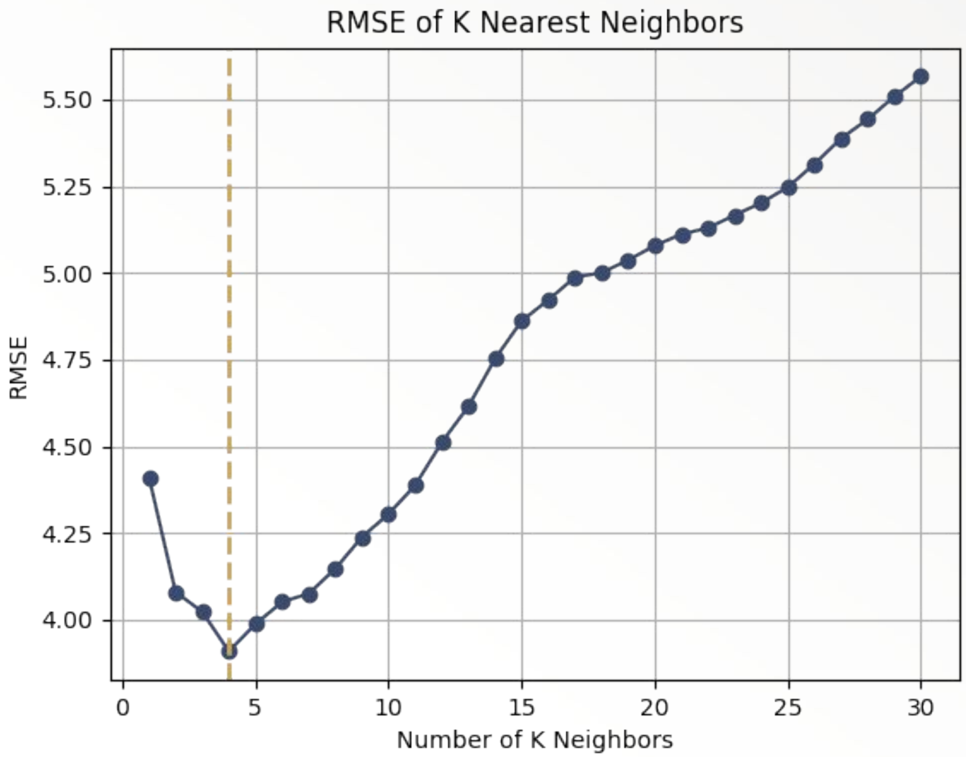
Model Overview

Model	Baseline	Linear Regression	KNN Regressor	Decision Tree	Random Forest	XGBoost	Neural Network
RMSE	14.209	3.206	3.909	3.371	3.41	3.28	3.204

Key Takeaways

- We chose to use RMSE to measure model performance since this is a regression problem rather than a classification problem
- Our **neural network model performed the best** with the lowest RMSE
- In our models, the RMSE indicates, on average, how much **each model's predictions deviated from the student's actual grades** after they took the exam using a 100-point scale
- The RMSE that we used as our **naive baseline** was calculated using the **mean of the posttest scores in our training data**

KNN Regressor



The optimal number of neighbors is **4**

Our KNN model's RMSE is **3.909**

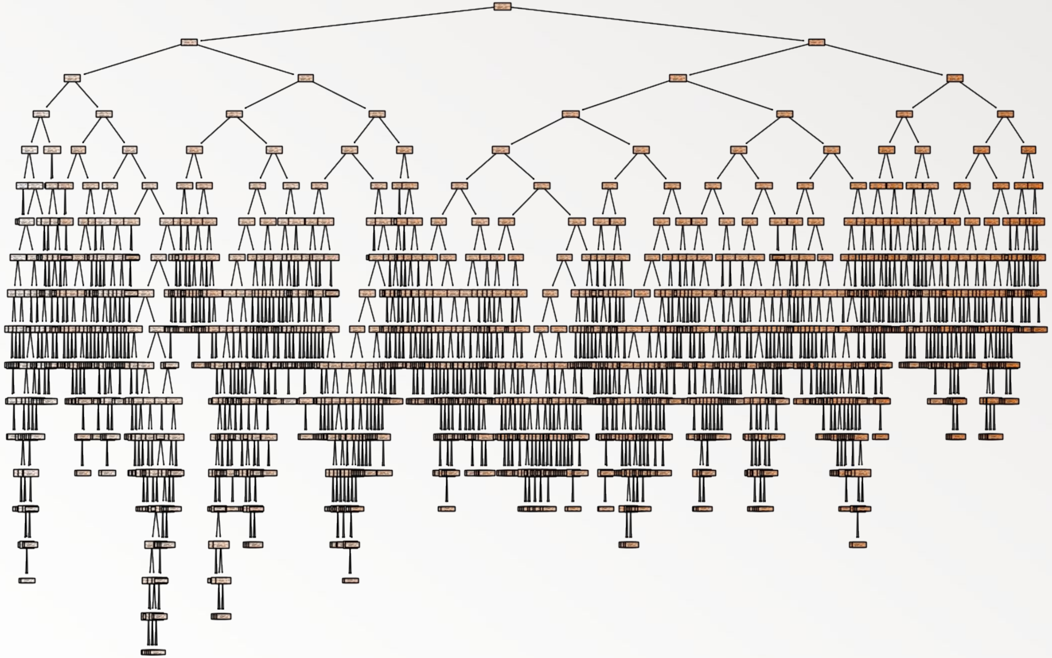
Our KNN model **improved from the baseline RMSE** of 14.209

Full Decision Tree

Full Tree Structure

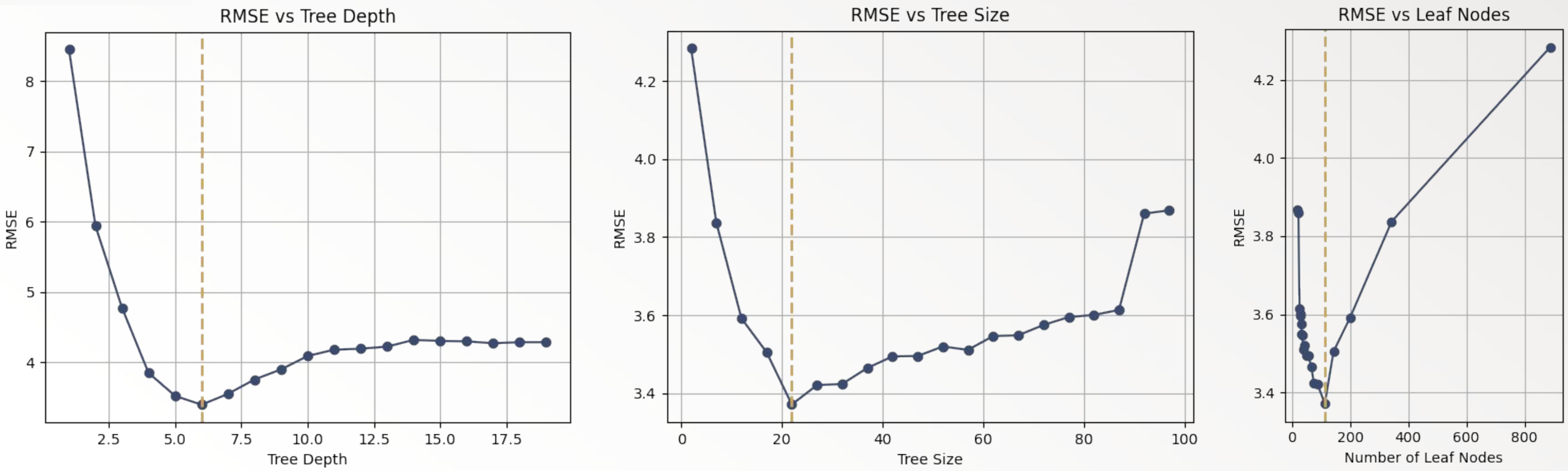
- Number of nodes: **1775 nodes**
- Number of leaves: **888 leaves**
- Tree depth: **88 layers**

The full decision tree has an RMSE of **4.283**



The model's RMSE improved from the naive baseline but **we can still improve this model**

Pruned Decision Tree



Optimal Parameters For Tuning Individually

- Maximum tree depth: **6 layers**
- Minimum tree size: **22 nodes**
- Minimum leaf nodes: **113 nodes**

The best model that adjusts these parameters individually is the model **pruned by tree size** with an **RMSE of 3.371**

Pruned Decision Tree

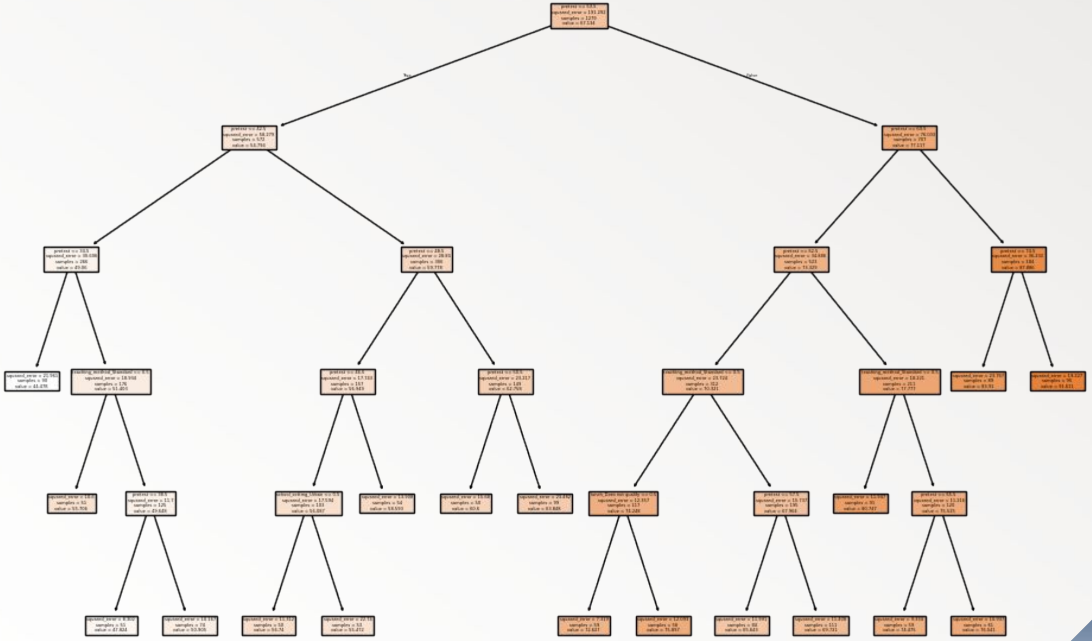
When tuning multiple parameters in the same model, the best model had an **RMSE of 4.016**

Model Parameters

- Maximum tree depth: **5 layers**
- Minimum tree size: **10 nodes**
- Minimum leaf nodes: **50 nodes**
- Minimum impurity decrease: **0**

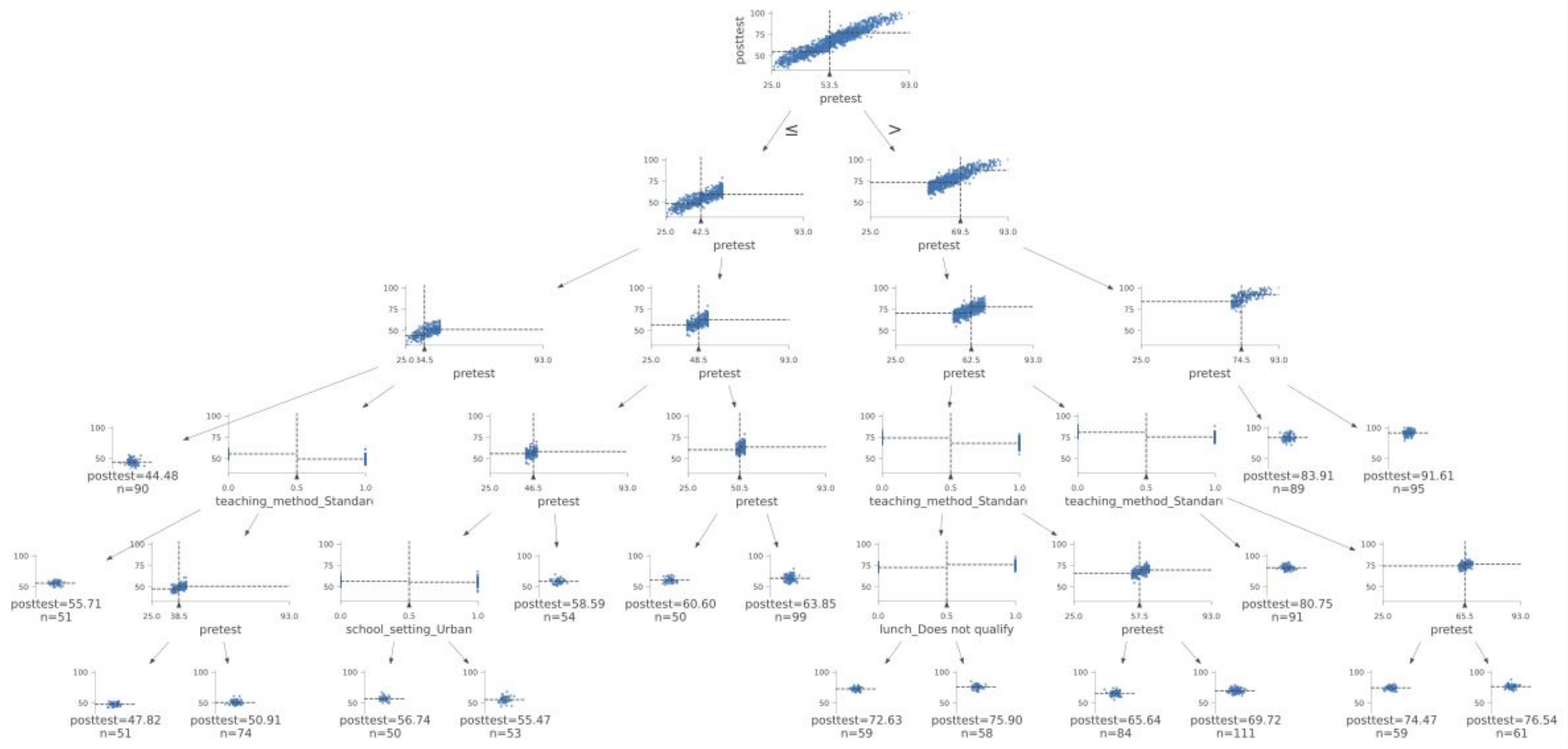
Model Structure

- Number of nodes: **35 nodes**
- Number of leaves: **18 leaves**
- Tree depth: **5 layers**



Model performed better than the full tree, but **performed worse than the previous pruned tree**

Pruned Decision Tree (dtreeviz)



Decision Tree Extensions

K-Fold Cross Validation

- **More reliable evaluation** of our model's performance since it averages results over multiple splits of the data
- Cross validation **resulted in a lower RMSE** for both the full tree and pruned tree models
- The full tree's cross validation RMSE is **3.309**, while the original full tree's RMSE is 4.283
- Pruned tree's cross validation RMSE is **4.22**, while the original pruned tree's RMSE is 4.016

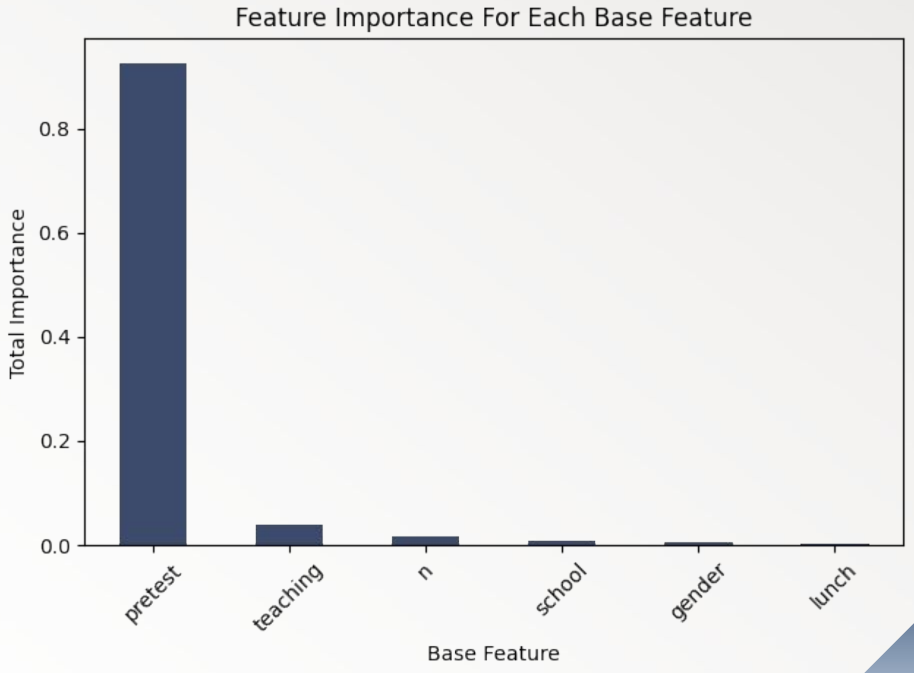
XGBoost RMSE of 3.28

Random Forest

We averaged a set of **1000 decision trees** to reduce variance

Key Takeaways

- Our random forest model resulted in a RMSE of **3.41** → this model **performed better than the full tree but not the pruned tree**
- Students' **grades before taking the exam** was by far the **most important variable** used in these decision trees to predict students scores after the exam
- Feature importance aligned with our findings through the correlation matrix



Neural Network

This model performed the best with **2 hidden layers**: the first with **10 neurons** and the second with **5 neurons**

Hidden Layer #	# Of Nodes	RMSE
1	9	3.246
	10	**3.213**
	11	3.224
2	4	3.271
	5	**3.224**
	6	3.268

With **2 hidden layers** and **(10, 5) neurons**, the RMSE of the neural network model is **3.224**

K-Fold Cross Validation

- Cross validation **resulted in a slightly lower RMSE** than the original neural network model
- The RMSE for the neural network cross validation model is **3.204**

Models Excluding The Pretest Variable

Model	Baseline	Linear Regression	KNN Regressor	Decision Tree	Random Forest	XGBoost	Neural Network
RMSE	14.209	8.270	6.403	5.357	5.462	5.3	7.763

Key Takeaways

- Due to concerns that "pretest" overshadowed the effects of our other variables, we ran the same models without this variable
- Models excluding "pretest" performed worse with a **higher RMSE across all models**
- From the models excluding "pretest," **XGBoost performed the best** with the lowest RMSE
- The most important features became **socioeconomic status** ("lunch" variable), **public vs private school**, and **number of students per class**

RMSE Issues

What went wrong?

- Used the wrong column as our model's target → using "gain" instead of "posttest"
- All RMSE comparisons were therefore invalid

How did this impact our results?

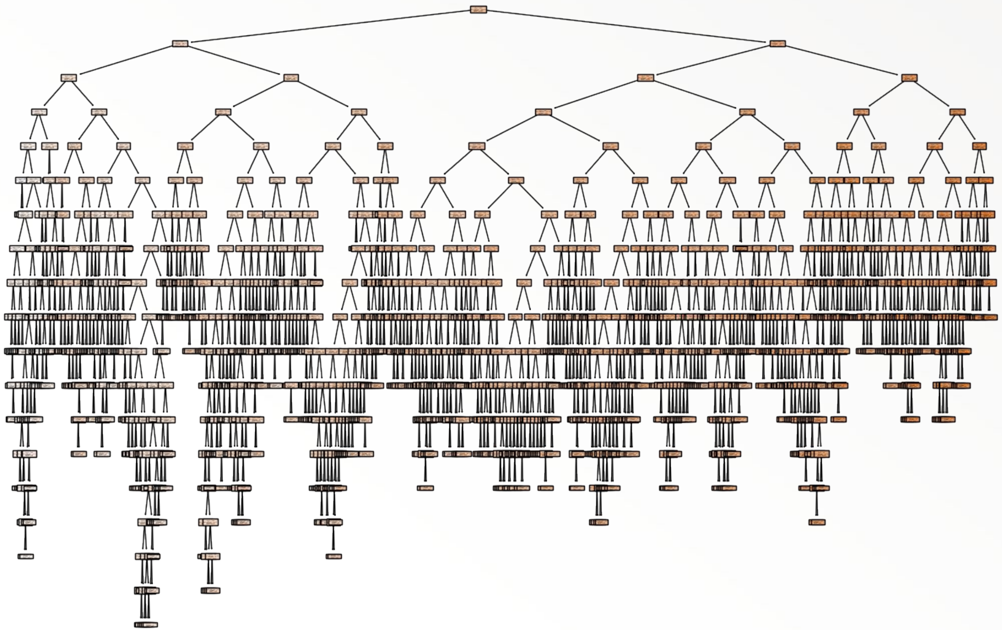
- Models learned to predict the wrong scores → **inflated errors vs. naive baseline**
- Our models RMSE were incorrect: Decision Tree was **14.8** and Random Forest was **14.1**, while the Naive Baseline is **14.2**

Lack of Variables

What went wrong?

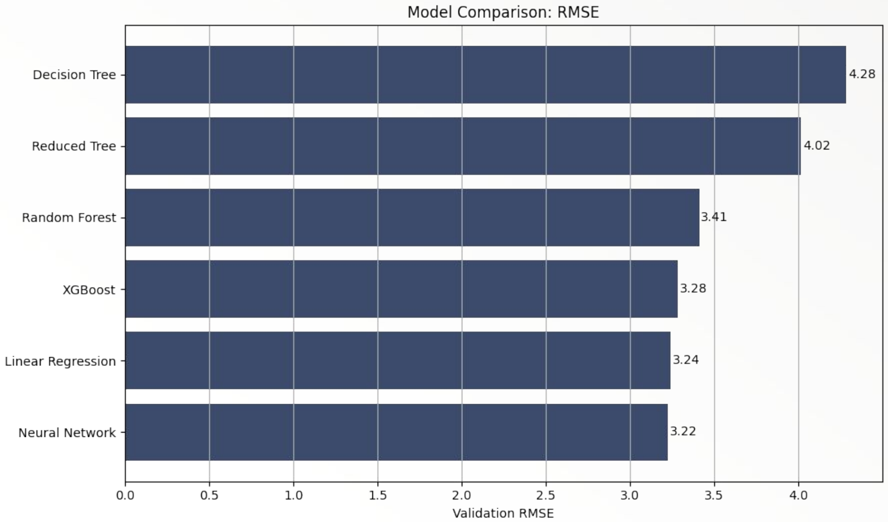
- Our dataset only contained a few variables
- This resulted in a **limited number of features** and **constrained** the scope and depth of our **analysis**
- May have led us to overlook **other important drivers** of posttest performance
- Possibly explore adding more relevant variables in future studies

Tree Overfitting



- Over 1,000 before pruning → **hard to read and shows indications of overfitting**
- The model **memorizes random noise** in the training data instead of learning the true underlying patterns
- The model's predictive **accuracy deteriorates** on unseen data
- Training and prediction with the model are **slow** and require **substantial computational resources**
- As tree depth increases, both training and prediction **times become significantly slower**

Conclusion



The **optimal prediction model** is the **neural network model** enhanced by k-fold cross validation

- The optimal neural network model for predicting academic performance had **two hidden layers** with **(10, 5) neurons**
- Applying cross validation to the model further minimized RMSE → model **generalizes well**

When analyzing individual contributors of academic success, we consistently found **previous academic success** to be the **strongest indicator** of future success

THANK YOU

Q&A