The Impoverished and School

Predicting Assessment Success for Pakistani Students with the 2003 and 2004 LEAPS Datasets

1. Introduction

LEAPS: Learning and Educational Achievement in Pakistan Schools

- Punjab, Pakistan
- 112 villages, 850 schools, 12,000 children, 5,500 teachers, 800 headmasters
- The 2003 dataset alone has over 150 columns.
- Thriving private school competition
- Despite more public funds, student learning is low, especially in public schools.
- A minority of households and teachers were given questionnaires, meaning that there were some students who had more data than others, which became something to address while cleaning and preparing the data.

My Purpose and Hypothesis

- Students were assessed on Math, English, and Urdu, a local language.
- I build models that predict student success as measured by a **high median of the three graded assessments** in order to identify contributing factors.
- **Problem**: As Pakistani officials reach their goal of putting every citizen in school, my study helps identify factors that contribute to high learning as measured by excellent test scores.
- Because the datasets are so large, I made smaller DataFrames in pandas with factors I determined as important after initial EDA and reading research papers.

Data Acquisition

- As of mid-2020, the website from which one can download the LEAPS datasets as CSV files is down.
- I requested the CSV files from Ayi Chang (ayichang@worldbank.org) via email, who promptly responded with data from 2003-2005.

2. Data Wrangling and Cleaning

Data Wrangling

- I crafted two DataFrames: 2003 is comprised from five separate tables; 2004 is made up from eight tables.
  - My decisions were based on reading papers from scholars and EDA shown below.

Data Wrangling for 2003: Missing Values

Three groups:

a. Target variables (Math, English, Urdu): about 11%.
   - These rows had to be dropped.

b. Survey questions (only given to a minority of students): above 50%
   - These were kept and empty data were filled a number to be ignored: 99.

c. Missing data intended for all students: less than 3%.
   - Most were filled with the most frequent value

	Total Missing	Percent
own_agri_land_last_2_seasons	12741	92.76
i_vis_have	12738	92.74
print_have	12738	92.74
mauzaid	12738	92.74
hhid	12738	92.74
child_studied_at_diff_school	7363	53.61
teacher_rates_child_how_good_in_studies	7347	53.49
math	1625	11.83
urdu	1625	11.83
english	1625	11.83
teacher_training	280	2.04
teacher_days_absent_last_mo	117	0.85
teacher_from_mauza	56	0.41
salary_monthly_Rs	19	0.14
teacher_years_teaching	10	0.07

Data Wrangling for 2004: Missing Values

Three groups:

a. Target variables (Math, English, Urdu): about 41%.
   - These rows had to be dropped.
b. Survey questions (only given to a minority of students): above 29%
   - These were kept and empty data were filled with a number to be ignored: 99
c. Missing data intended for all students: less than 1%.
   - Most were filled with the most frequent value.

	Total	Missing Percent
teacher_years_teaching	108363	99.27
teacher_from_mauza	108362	99.27
teacher_qualifications	108361	99.27
teacher_training	108361	99.27
teacher_survey_absent_other_work	108361	99.27
teacher_survey_absent_office_work	108361	99.27
teacher_survey_absent_emergency	108361	99.27
teacher_sex	108361	99.27
teachercode	108361	99.27
type_of_housework_timeslot_5	107934	98.88
child_helped	107417	98.41
studied_at_same_school_as_last_year	81069	74.27
school_type	81069	74.27
television	64725	59.29
radio	64725	59.29
child_studied_at_diff_school	64725	59.29
teacher_rates_child_how_good_in_studies	62109	56.90
child_days_absent_last_mo	62109	56.90
urdu	44940	41.17
math	44940	41.17
english	44940	41.17
grade	37906	34.73
child_teachercode	37906	34.73
hh_child_in_govt_primary_school	32960	30.19
tehsil_census_code	32745	30.00
supervisor_code	32745	30.00
hhid	32733	29.99
student_sex	59	0.05
grade_median	59	0.05

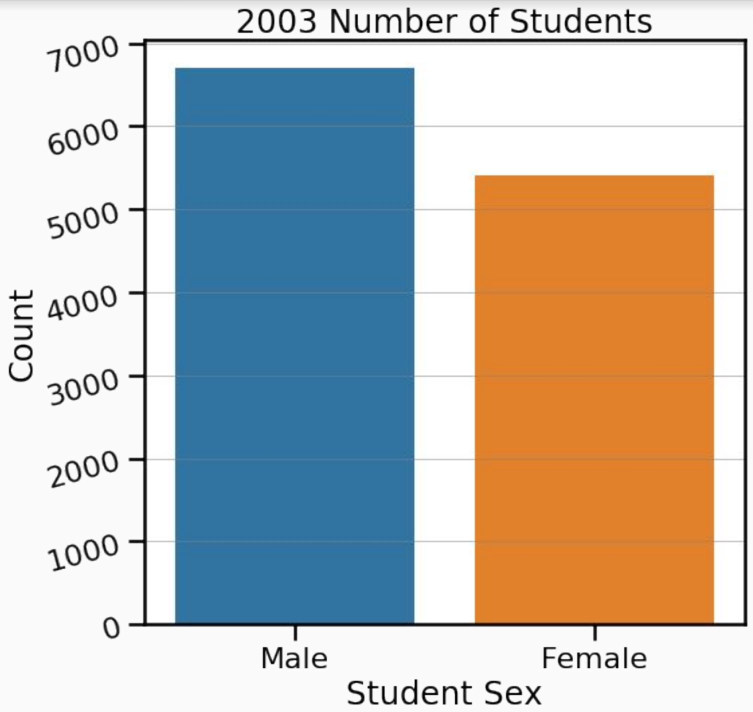
Data Cleaning

- I convert categories into integers in order to prepare them for predictive models
  - For example, a column recording how many years teachers have taught was originally categorical: “< 1 Year”, “1-3 Years”, “< 3 Years.” I changed those to numbers: 1, 2, and 3, respectively, because predictive models only understand numbers, not categories or words.
- I use sklearn’s SimpleImputer to impute some missing values.

3. EDA

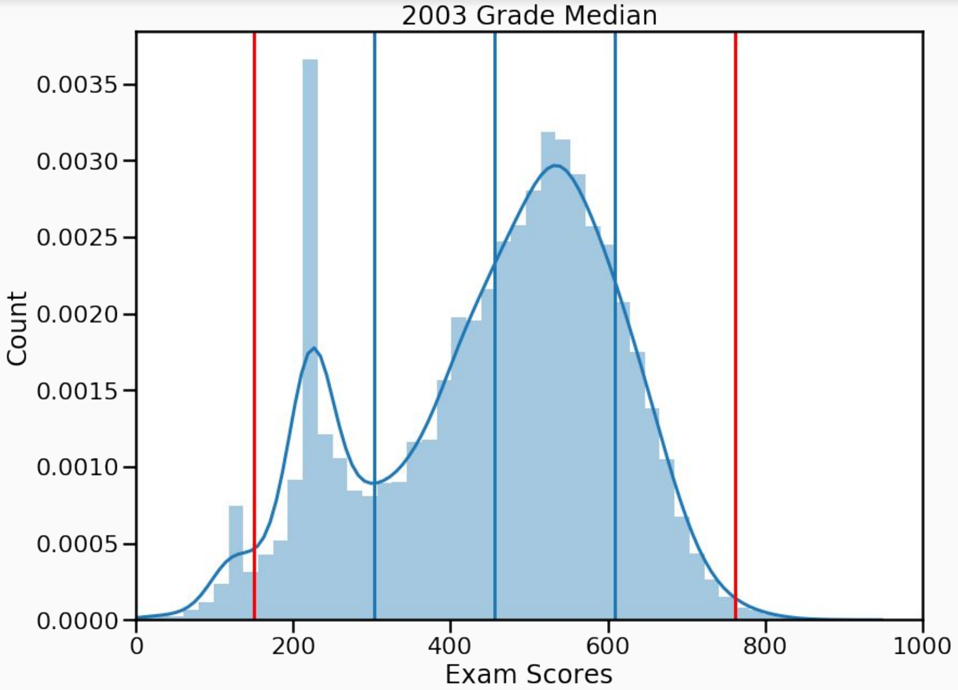
Let's Explore the Data: 2003

12,110 students (with valid grades)



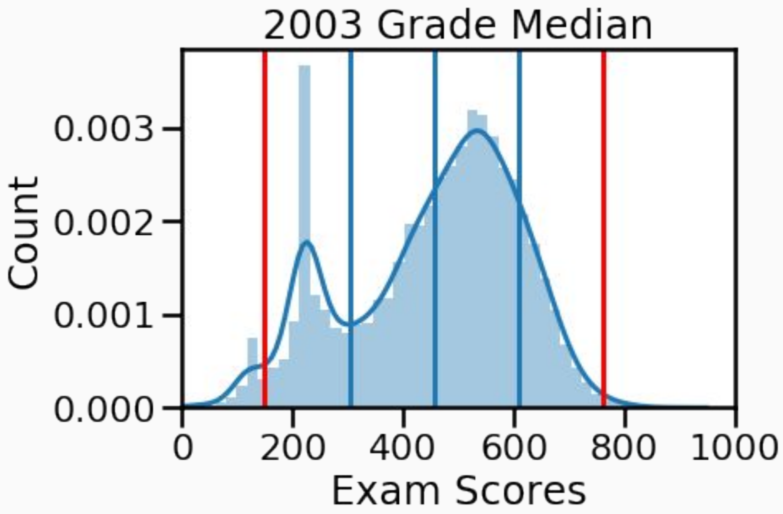
2003 Bell Curve of The Target Variable: Grade Median

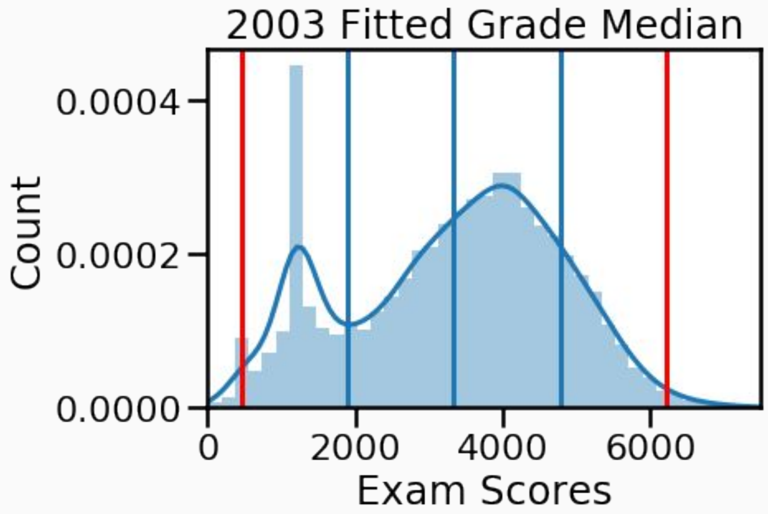
- Low grades (out of 1,000)
- Abnormal bell curve: it's steep and there is a large skew to the left.



2003 Distribution

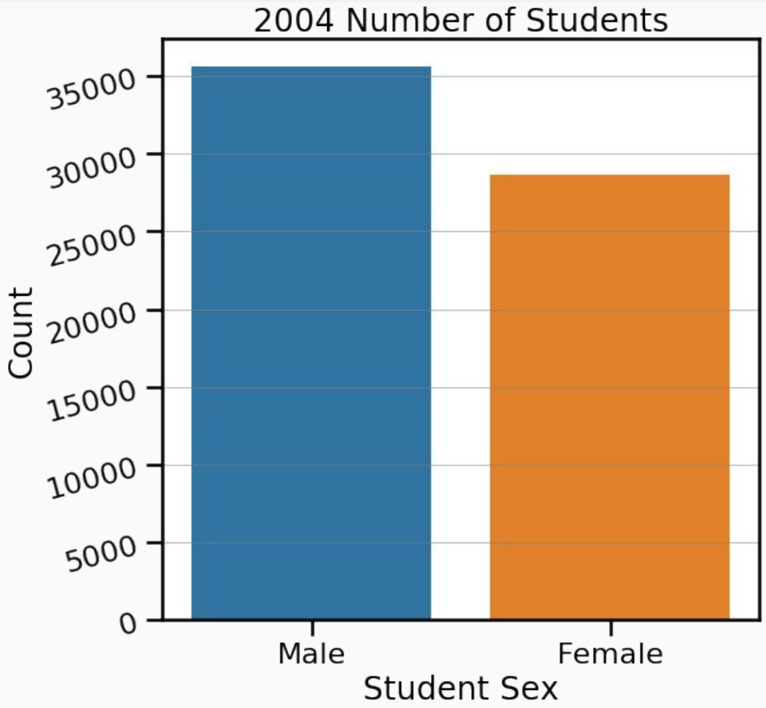
The left shows the original bell curve of the data for 2003 and 2004. I crammed all the data into a more normal looking bell curve using a box cox transformation. Notice the heights are lower and the curves are more rounded. Making them approximate a normal distribution increases model accuracy.





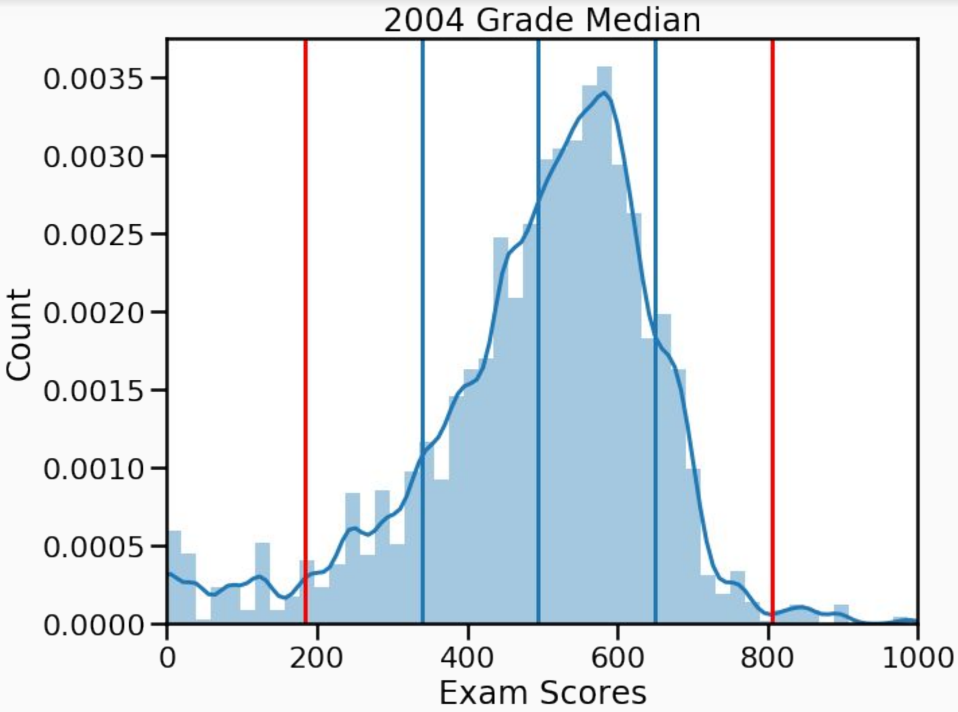
Let’s Explore the Data: 2004

64,218 students (with valid grades)



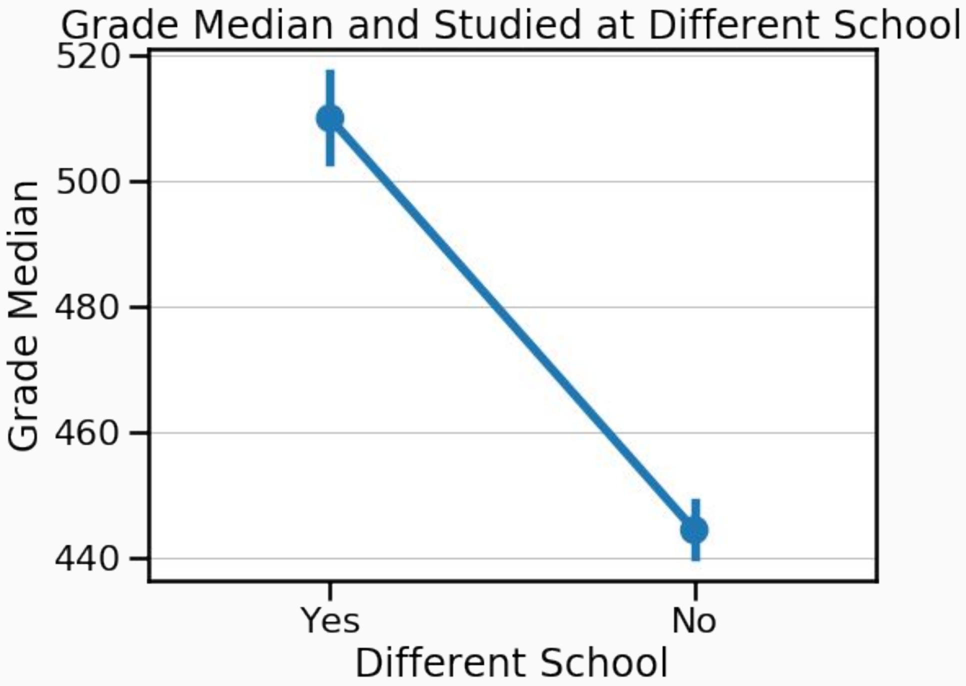
2004 Bell Curve of Grade Median

- Low grades (out of 1,000)
- Abnormal bell curve: it's steep and there is a large skew to the left.



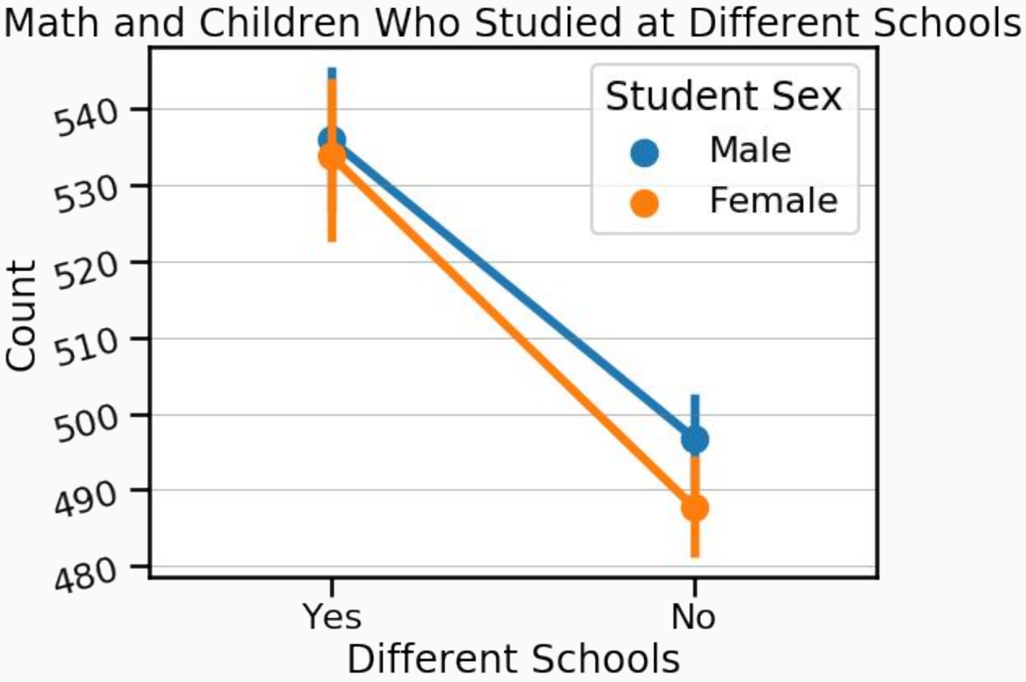
Changing Schools

- Students who changed schools scored higher by about 5%. It seems parents are changing schools intentionally, and with good results.



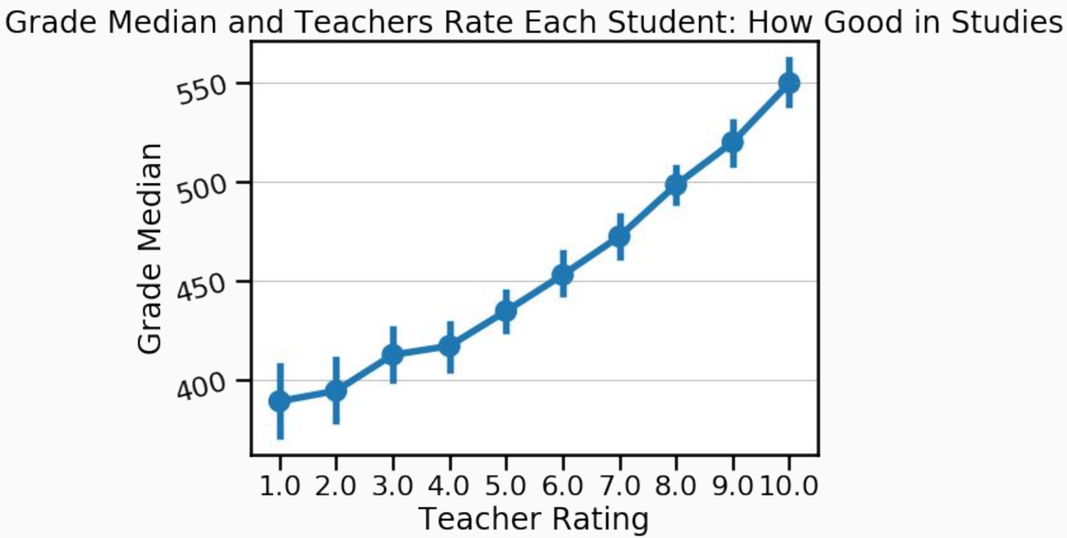
Grades: Males and Females

- Females performed better in English and Urdu, while boys scored slightly higher in Math, unless (for 2003 only) girls changed schools.
- The narrow margin between the sexes in math is also present in 2004.



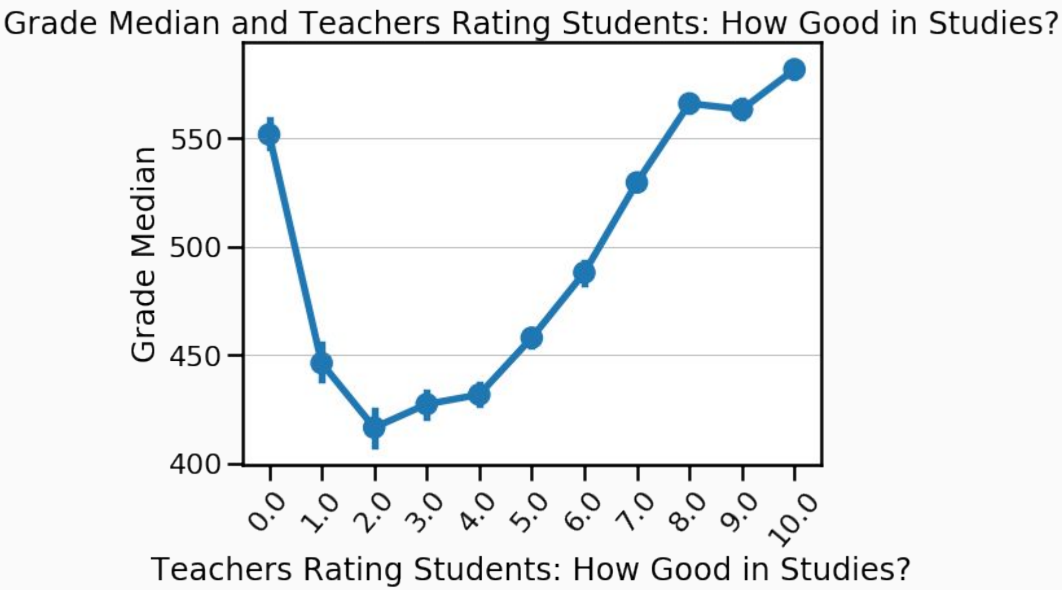
2003: Teacher's Ratings of Students

Teachers rated the academic performance of each student on a scale of 1 to 10, and seemed to be correct. [Teacher expectations are powerful](), and it is clear that there may be a causal, two-way street.



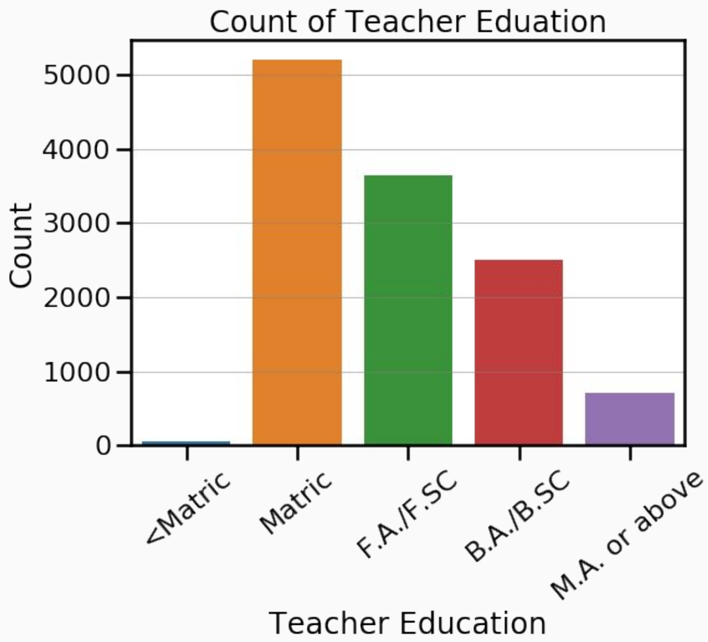
2004: Teacher Ratings of Students

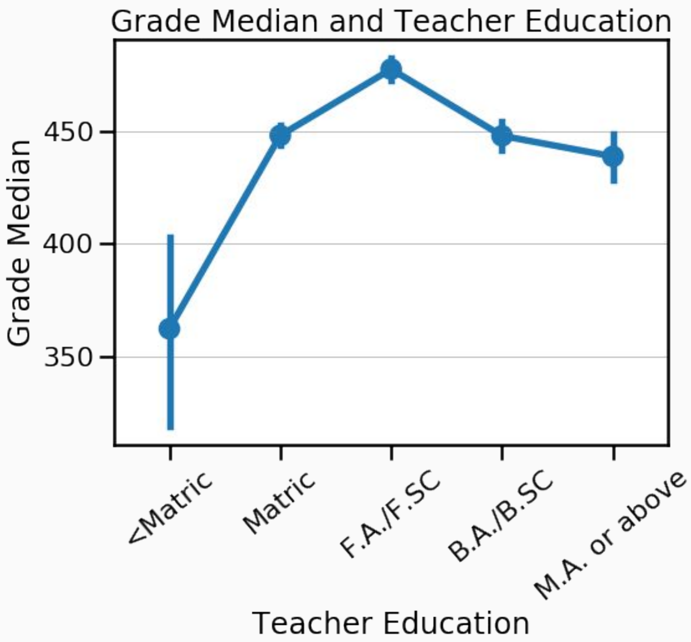
Aside from the odd jump in 0 (which seems to be where teachers did not rate children) the same holds true for th 2004 dataset.



Teacher Factors

Surprisingly, teacher education did not seem too important for assessment scores in 2003 and 2004.



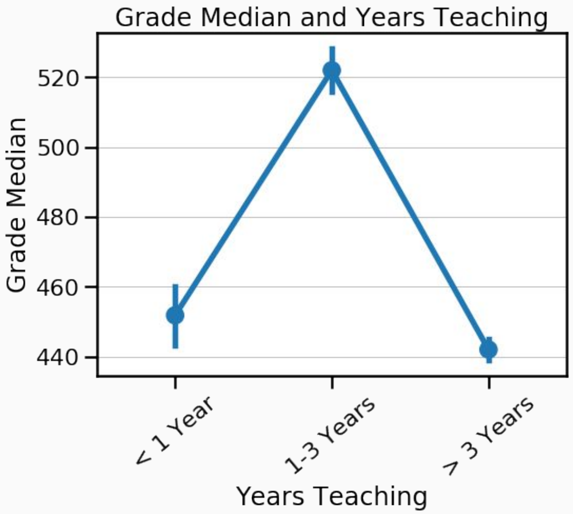


Teacher Factors

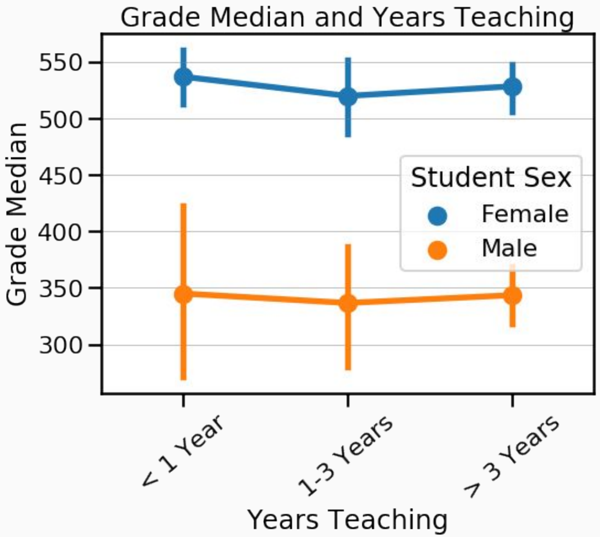
In 2003 alone, teachers with 1 - 3 years of experience teaching have students performing higher on exams.

2004 does not have such a drastic difference, which may be just as surprising.

2003



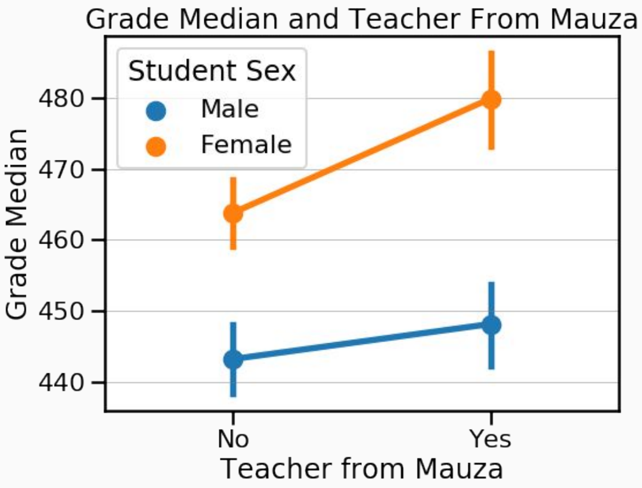
2004



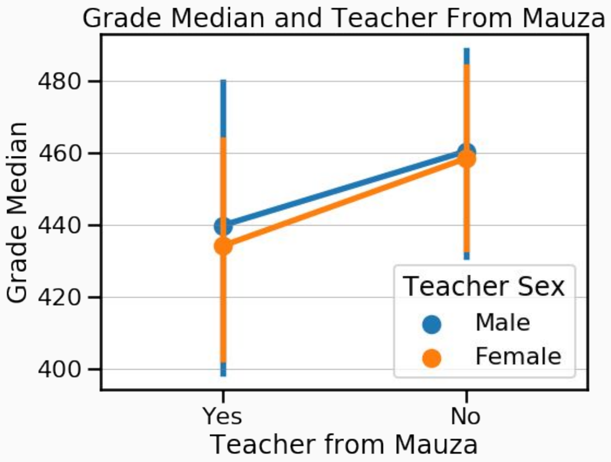
Teacher Factors

If the teacher is from the mauza in which she is teaching, students score slightly higher.

2003

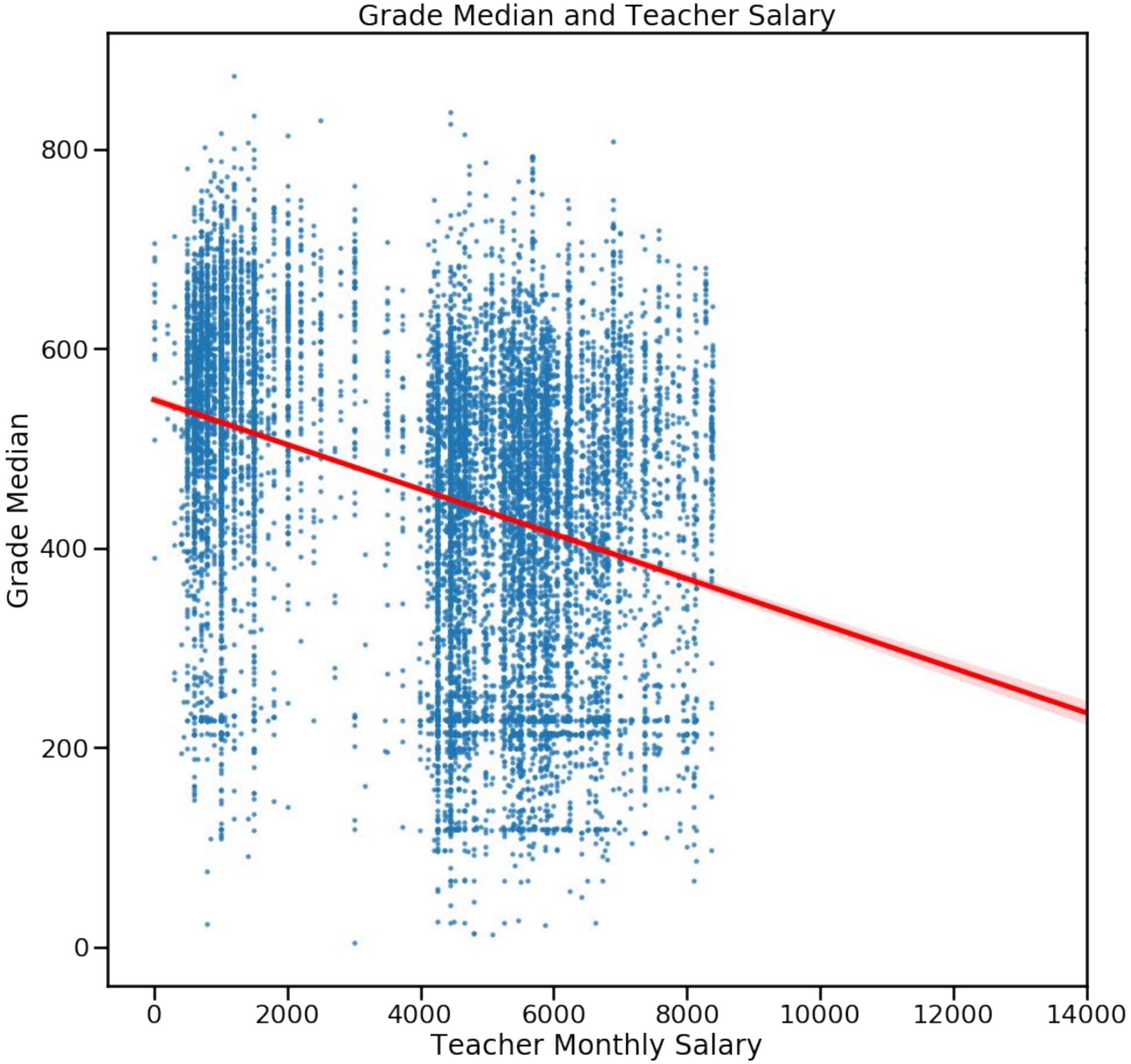


2004



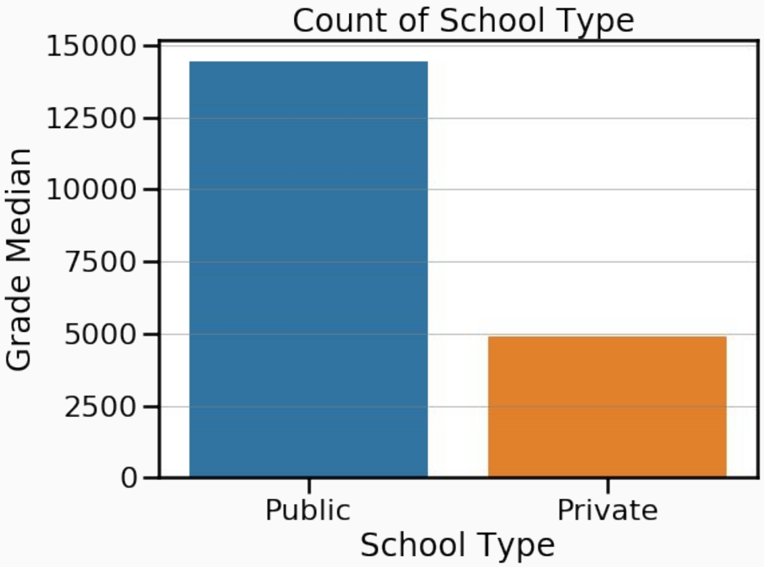
Teacher Factors

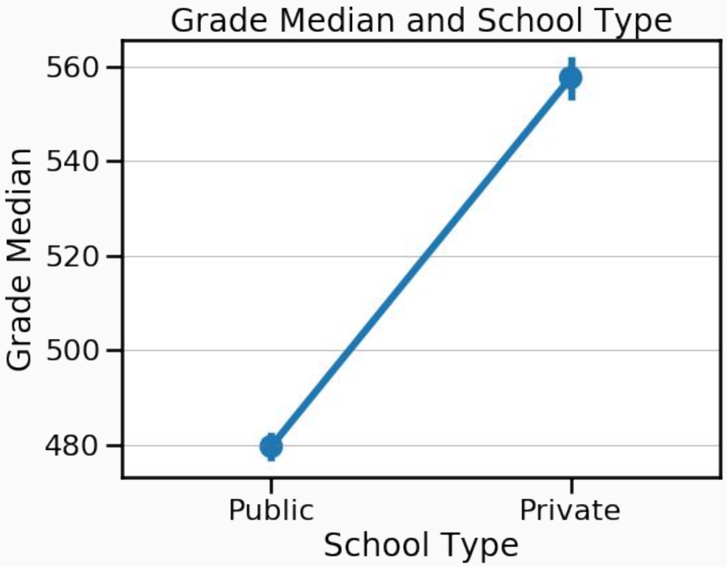
There is a negative correlation between teacher pay and student success.



Private and Public Schools: 2004

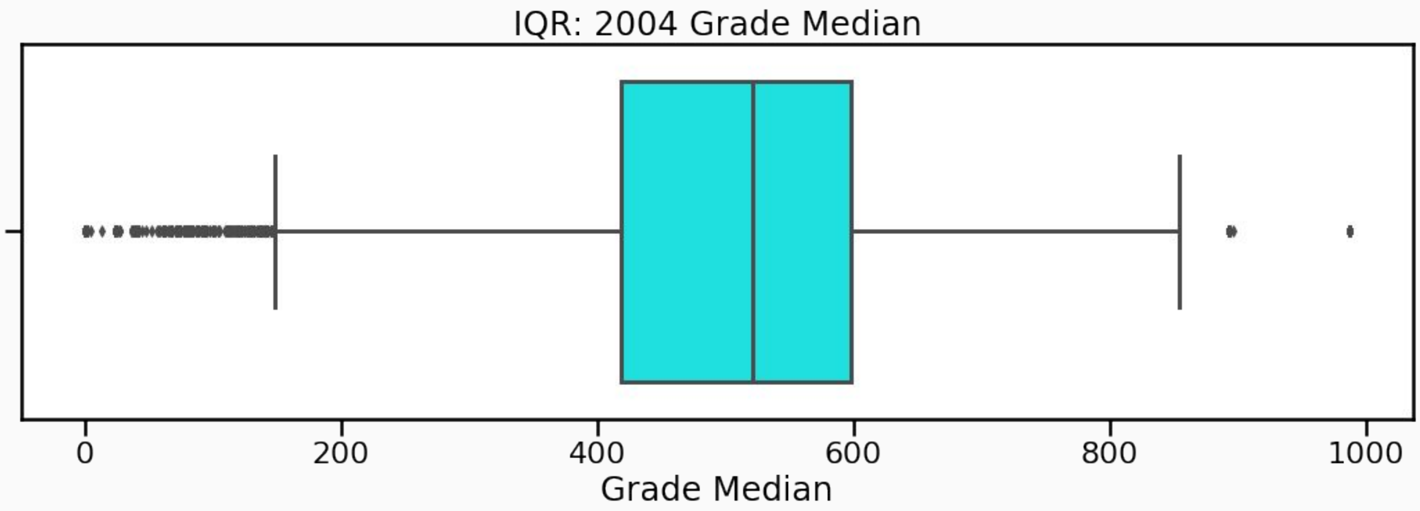
Students perform better in private rather than public schools.





Outliers

The datasets are odd in that each had a hundreds of instances below two standard deviations and less than 100 above. They were skewed to the left.



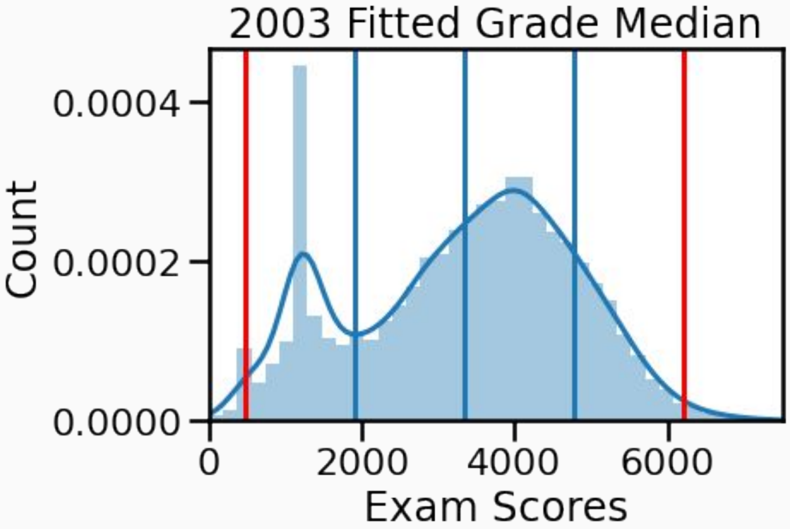
Outliers (fitted data)

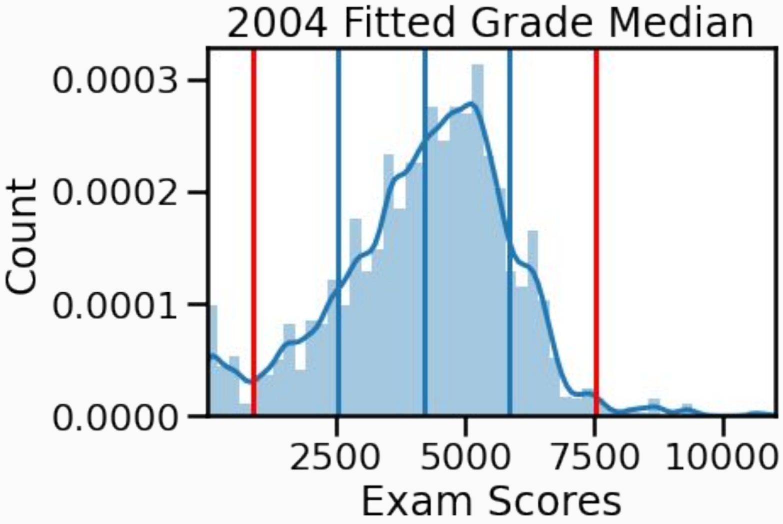
```
df2003.grade_median:
    Q1: 352.0
    Q3: 570.75
    IQR: 218.75

df2004.grade_median:
    Q1:418.0
    Q3:598.0
    IQR:180.0
```

```
2003: Instances Two STDs above: 46
2003: Instances Two STDs below: 11775
```

```
2004: Instances Two STDs above: 611
2004: Instances Two STDs below: 61049
```

Quick Look at the Distribution (fitted data)





Predictive Power Score

2003

- Math (a derivative) was beaten by child_teachercode.
- Monthly salary and student success was a negative correlation.

x	y	ppscore
english	grade_median_fitted	0.590254
urdu	grade_median_fitted	0.581644
child_teachercode	grade_median_fitted	0.274411
math	grade_median_fitted	0.257869
salary_monthly_Rs	grade_median_fitted	0.176480
childcode	grade_median_fitted	0.095510
teacher_training	grade_median_fitted	0.040940
teacher_rates_child_how_good_in_studies	grade_median_fitted	0.016774
teacher_sex	grade_median_fitted	0.016630
teacher_years_teaching	grade_median_fitted	0.013918

Predictive Power Score

2004

- More variables beat derivatives.

x	y	ppscore
childcode	grade_median_fitted	0.750722
math	grade_median_fitted	0.721916
hhid	grade_median_fitted	0.675942
english	grade_median_fitted	0.641680
urdu	grade_median_fitted	0.528018
child_teachercode	grade_median_fitted	0.389172
teacher_rates_child_how_good_in_studies	grade_median_fitted	0.070731
tehsil_census_code	grade_median_fitted	0.030817
hh_child_in_govt_primary_school	grade_median_fitted	0.024000
supervisor_code	grade_median_fitted	0.014571

4. Machine Learning (ML)

ML

I considered four datasets:

1. 2003
2. 2003 without derivatives
3. 2004
4. 2004 without derivatives

ML: No Parameter Tuning

2003 No Derivatives:

- Linear Regression Score: 0.144
- Decision Trees (XGB): 0.488

2004 No Derivatives:

- Linear Regression Score: 0.144
- Decision Trees (Random Forest): 0.86

```
regression_models = [
    LinearRegression(),
    Ridge(),
    Lasso(),
    ElasticNet(),
    LinearSVR(),
    RandomForestRegressor(),
    GradientBoostingRegressor(),
    xgb.XGBRegressor()
]

for regression_model in regression_models:
    loop_pipe = make_pipeline(regression_model)
    loop_pipe.fit(X_train04n, y_train04n)
    print(f'2004 No Derivatives\n\
    {regression_model} \n\
    model score: {loop_pipe.score(X_test04n, y_test04n):.4f}')
```

ML: with Parameter Tuning

2003 No Derivatives

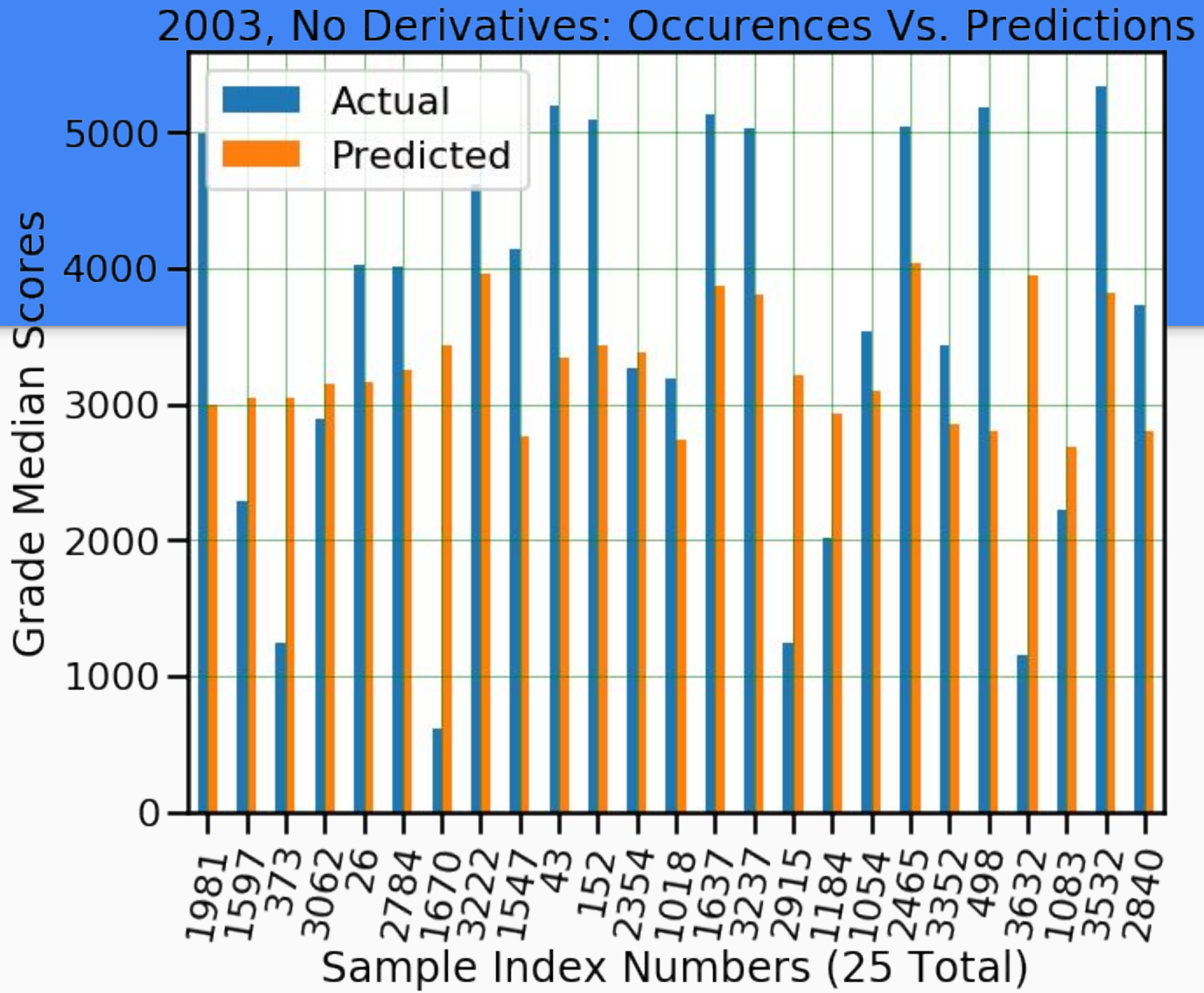
- Random Forest: .49

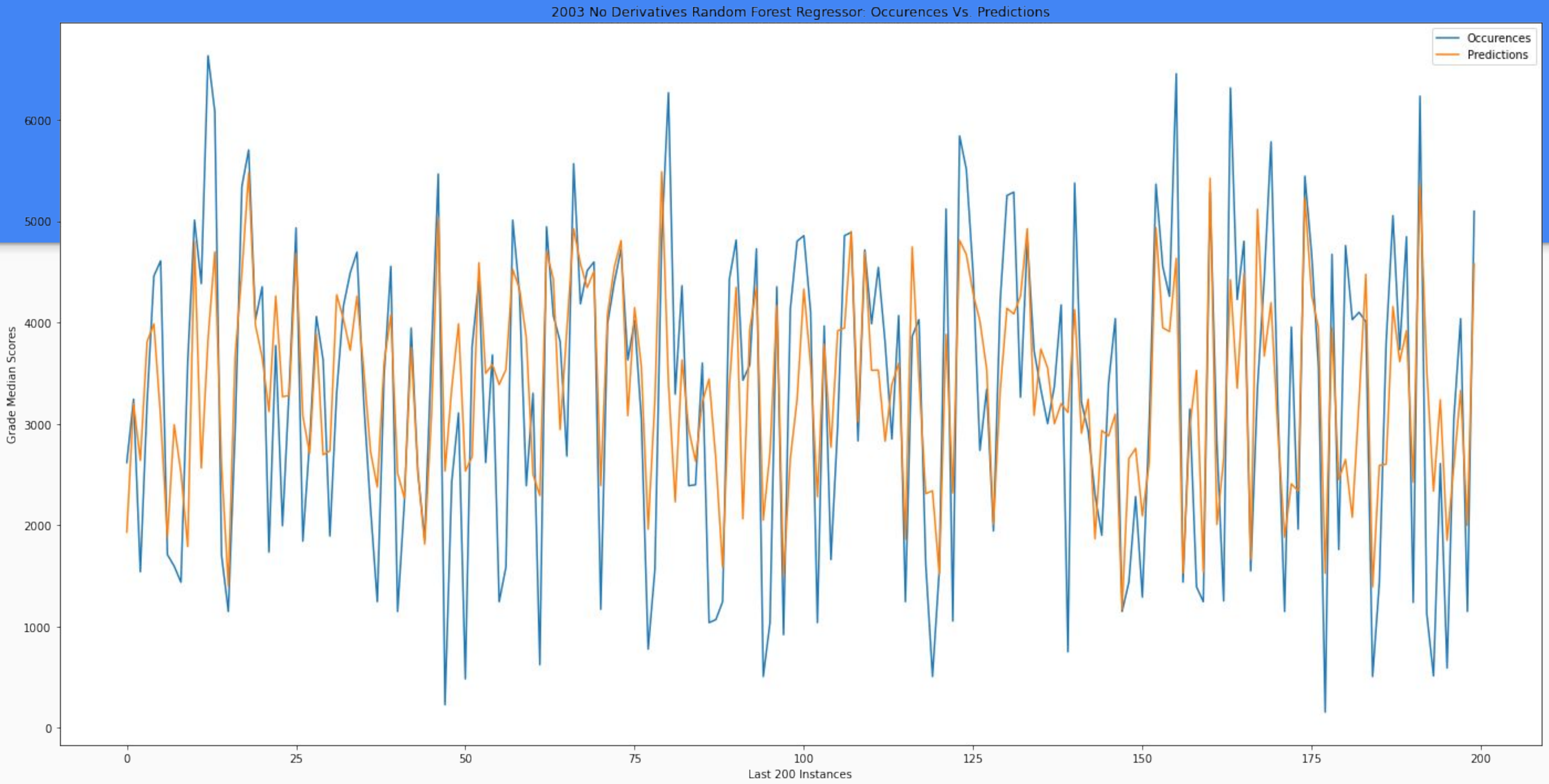
2004: No Derivatives

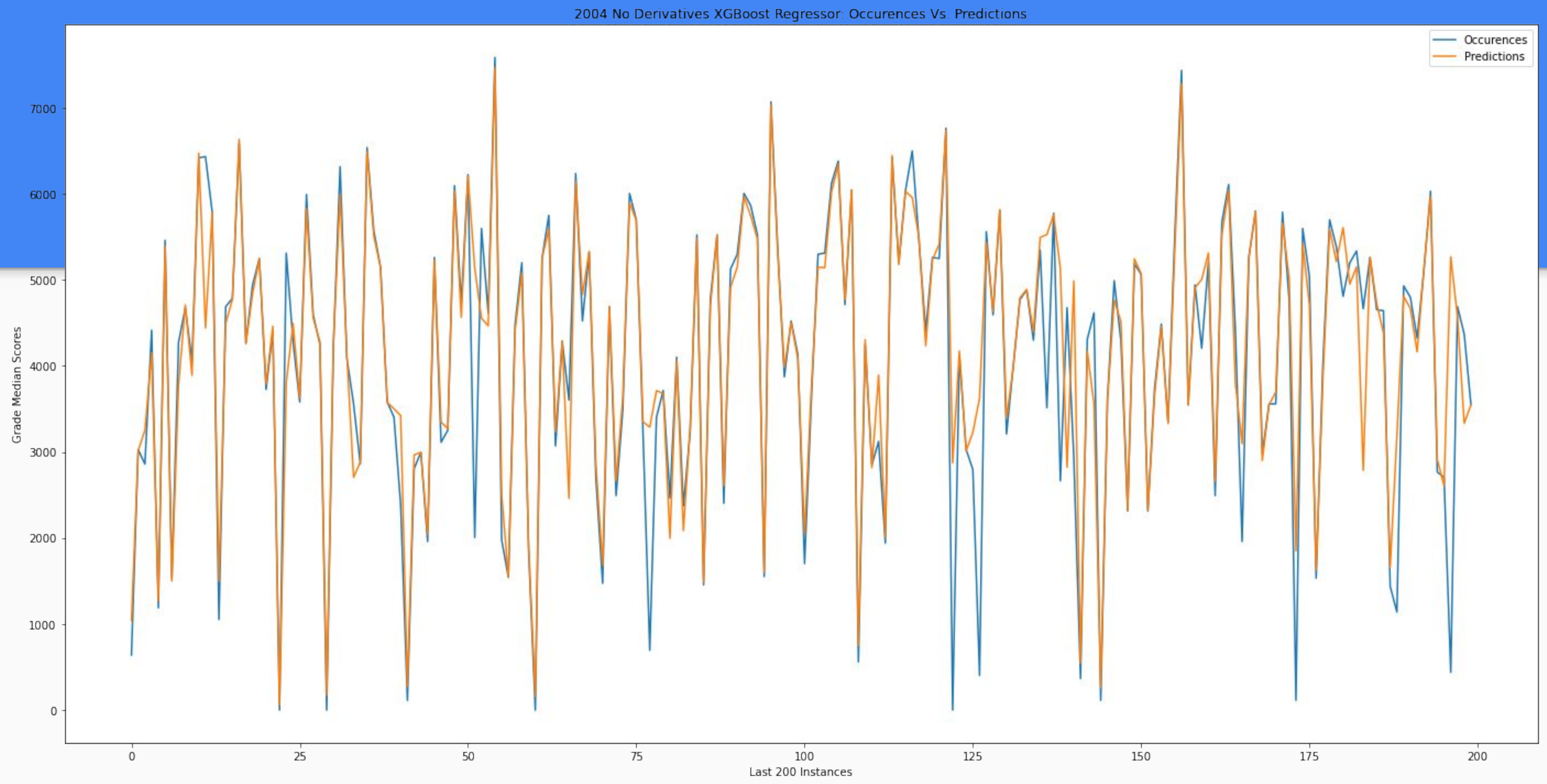
- XGB: .88

ML

Random Forest: .49




2003 No Derivatives Random Forest Regressor: Occurences Vs. Predictions
Occurences
Predictions
Grade Median Scores
0
1000
2000
3000
4000
5000
6000
0
25
50
75
100
125
150
175
200
Last 200 Instances


2004 No Derivatives XGBoost Regressor: Occurences Vs. Predictions
Occurences
Predictions
Grade Median Scores
0
1000
2000
3000
4000
5000
6000
7000
0
25
50
75
100
125
150
175
200
Last 200 Instances

ML: 2004 Feature Importance

The most important column for 2004 is 'hh_child_in_govt_primary_school.' Perhaps the dataset for 2003 does not score as well because it is missing this column.

The next column is Tehsil Census Code, which is also absent in df2003. It is unclear why the Tehsil Census Code is an important predictor when similar features like teachercode or hhid (household ID) are not.

Next is grade, which is weak. Just to convey how little information the grade column gives, let's discuss it. It describes what grade a child is in.

features	importance
hh_child_in_govt_primary_school	0.214284
tehsil_census_code	0.128346
grade	0.088576

ML: 2004 Feature Importance

About 96% of students in the “grade” column are in 4th grade.

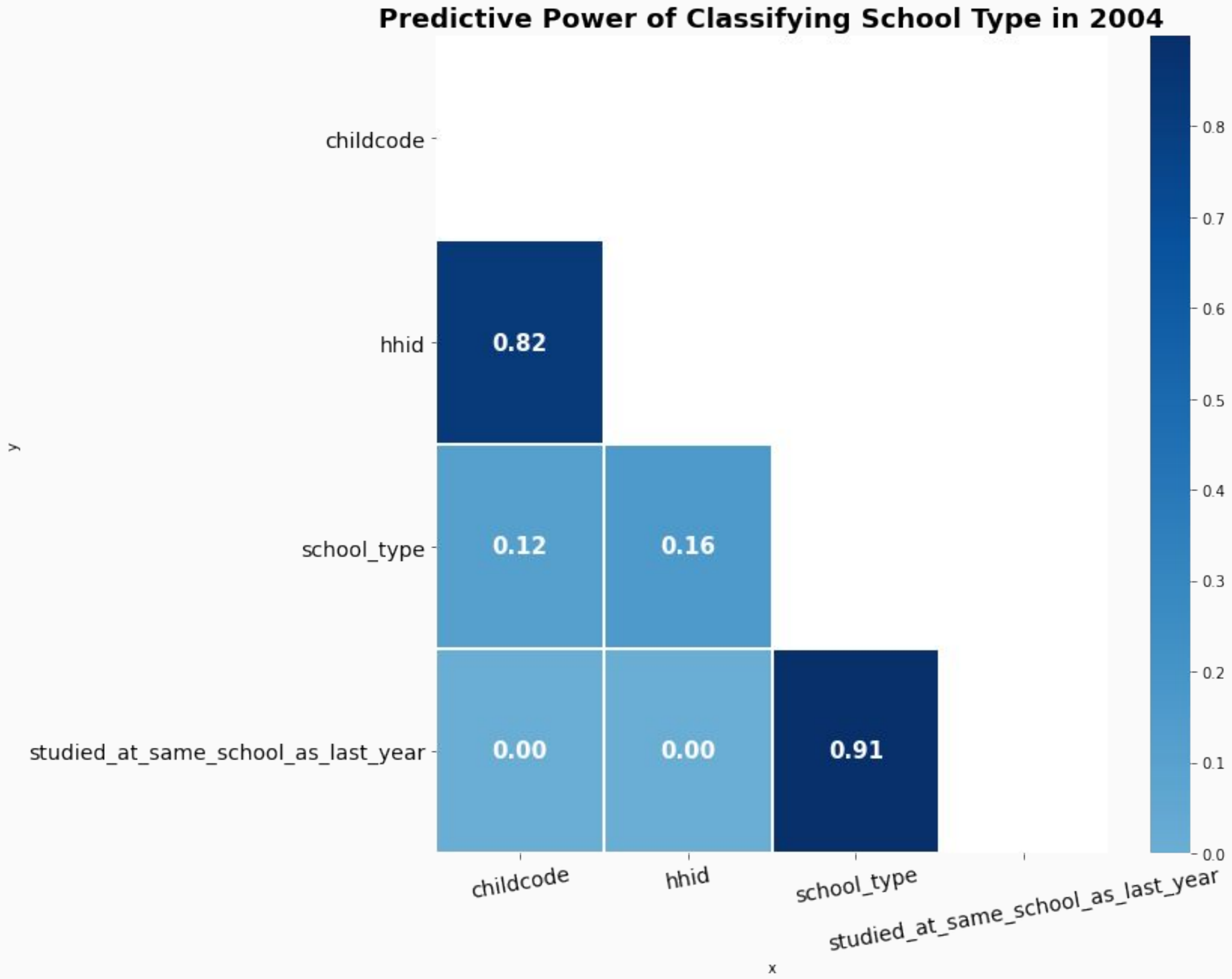
```
There are 64,218 total students.
61,373 students are fourth graders, leaving only 2,845 non-fourth graders.
```

5. Excursus: A Classification Task

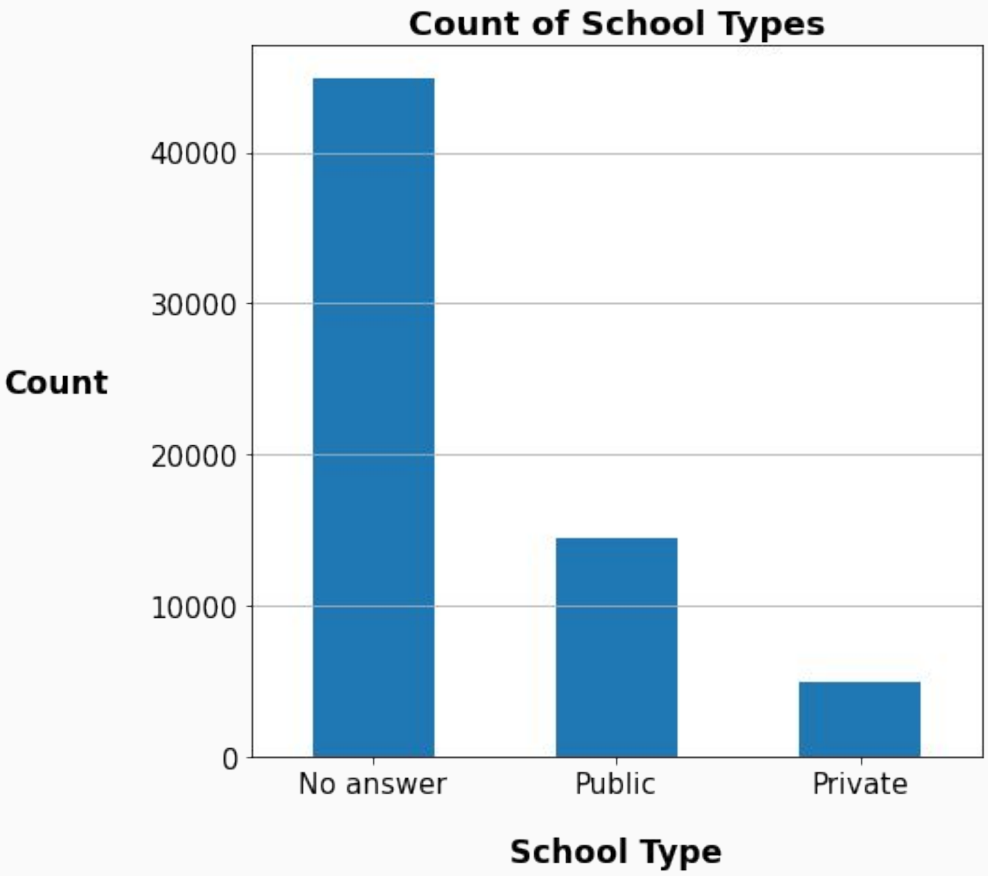
Excursus: Classify Whether a Student is in a Public or Private School for 2004

- Convert ‘school_type’ column back into categorical dtype.
- Instantiate a Random Forest Classifier.
- Score: .98

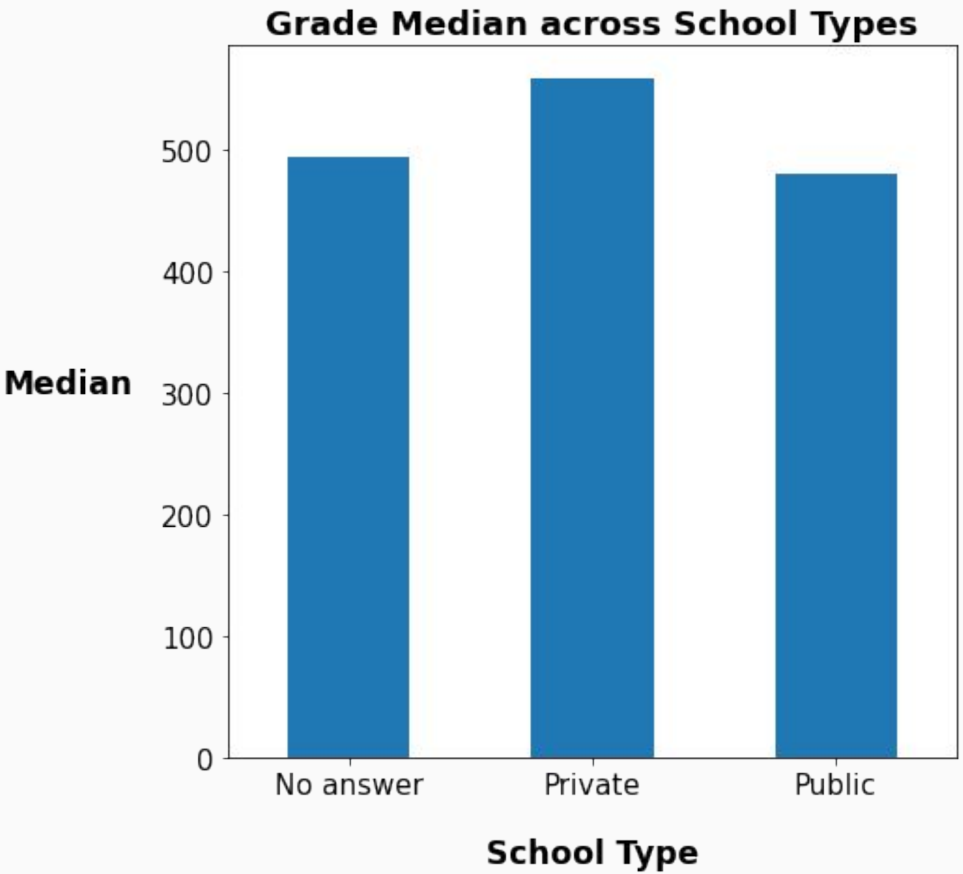
	feature	importance
13	studied_at_same_school_as_last_year	0.742122
3	hhid	0.033000


Predictive Power of Classifying School Type in 2004
childcode
hhid
school_type
studied_at_same_school_as_last_year
0.82
0.12
0.16
0.00
0.00
0.91
y
x
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

Let's Consider School Type



What's the Median of Grade Median by School Types?



6. Recommendations

Recommendations

1. Encourage the growth of private schools.
    - It may be dangerous to offer private schools money if regulations accompany it because regulations may reduce school student success.
2. Investigate why private schools have more success.
3. Encourage teachers why they teach: because they love the children.
    - Teacher pay will probably not rise.
4. Consider incentivizing teachers to stay in their mauza of upbringing because their students tend to have better grades.
5. Investigate why teachers with higher educations do not have students scoring better than those without.

7. What I Would Do Differently If ____.

I had more time or computational power

What I Would Change

- Use PPS and RFE on all the columns (over 150) before forming a DataFrame.
- Data Imputation would use MICE.
- Data cleaning section would use CatBoost for categories.
- Add feature selection and extraction.
- Use more models to get a baseline for different types of algorithms.
- Implement a stacking regressor.
- Utilize Hyperopt to tune the parameters.

8. Conclusion: Steps Taken

Conclusion

- Request data via email.
- Form a small DataFrame for 2003 and for 2004 using pd.merge().
- Clean datasets
    - Drop rows with NaN proto-target variables.
    - Impute missing data less than 3%.
    - Convert categorical data into integers.
    - Cram data into Box-Cox a tranformation.

Conclusion

- EDA
  - Student success factors: Changing Schools and Sex
  - Teacher Factors: Educational Qualifications, Sex, Years Teaching, Teacher from Mauza
  - Negative correlation between teacher pay and student success
    - Probably due to low payment in most private schools.
  - PPS: school type and different school
  - Tehsil Census Code is a strong predictor, which is surprising.
    - This does not seem related to income or locale.

Conclusion

- ML
  - Linear models performed poorly: 0.144 on each year.
  - Decision trees performed well
    - 2003: XGBRegressor: 0.48
    - 2004: Random Forest Regressor: 0.88.
  - Feature importance for 2004
    - Whether a household had one student in government primary school
- ML: Classification Excursus
  - RFClassifier: 0.98
  - Highest feature: whether a student changed schools last year.
    - This seems importance because it invokes competition among schools and enables students to find a better fit among many schools.

9. Questions?