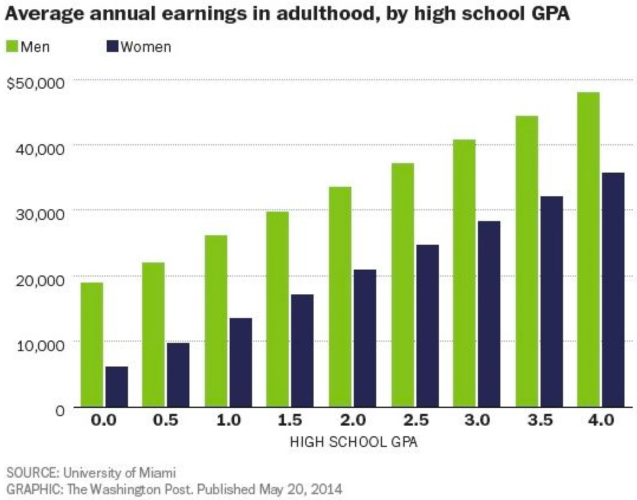
An Analysis of Student Grades

Student Grades

Student high school grades are important because they increase chances of college admission, number of colleges they can get into, merit based scholarships, and lifetime earnings.



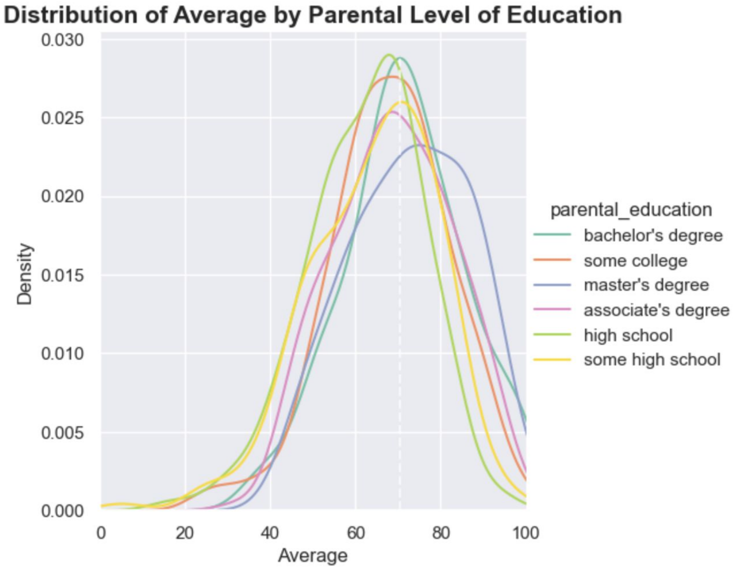
What affects student grades?

It's no doubt that grades play a role in a student's future, but what can we do to help students optimize their grades?

To help answer this question, I am analyzing a dataset from Kaggle containing information on math, reading, writing scores, as well as other factors, such as parental education, gender, standard lunch, race/ethnicity, and completion of a test prep program.

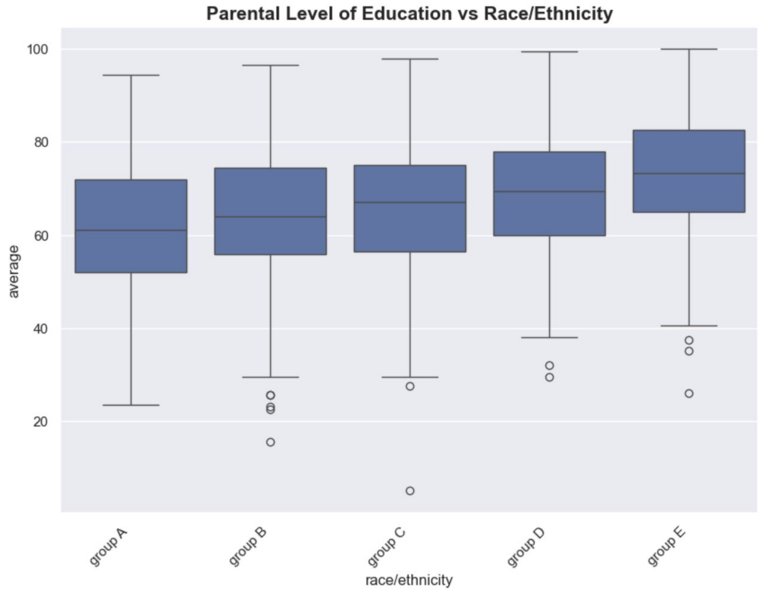
How does parental level of education affect grades?

- The highest average comes from parents with master's degrees. This can be seen in the navy bell curve. It's center is more to the right, which indicates a higher average for that group.
- The lowest average comes with parents with only high school degrees.



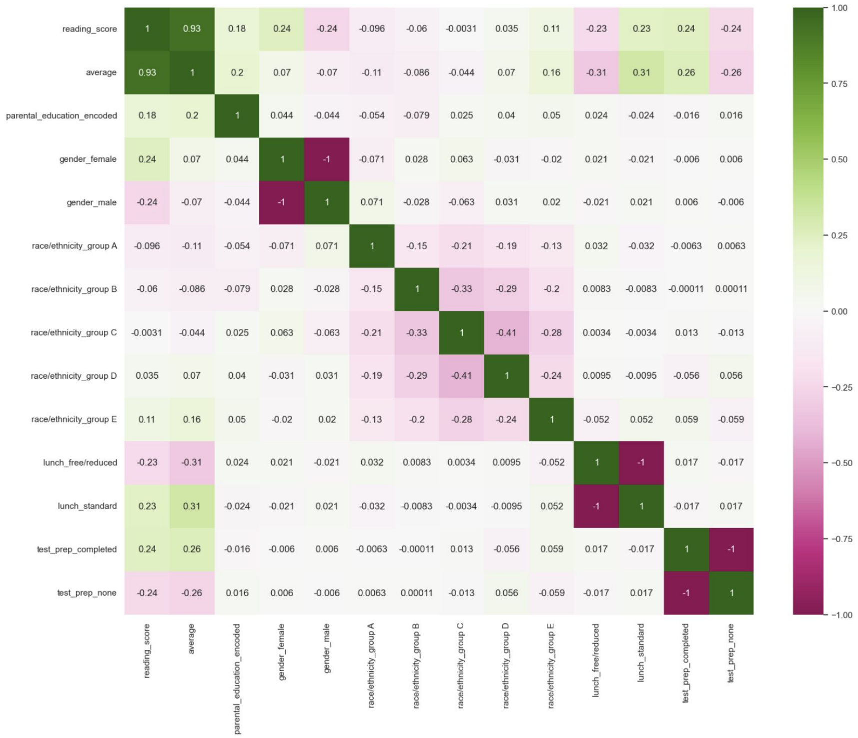
Do race/ethnicity affect student grades?

- By comparing averages across 5 different race/ethnicity we can see that group E has highest median average and group A has lowest median average.



Correlation Matrix

- The correlation matrix tells us how strongly each variable relates to one another.
- How does each variable relates to average?
    - Reading score has the highest correlation with average at 0.93
    - Having standard lunch, completion of test prep, parental education, and being part of race/ethnicity group E have positive, but low correlations with average.

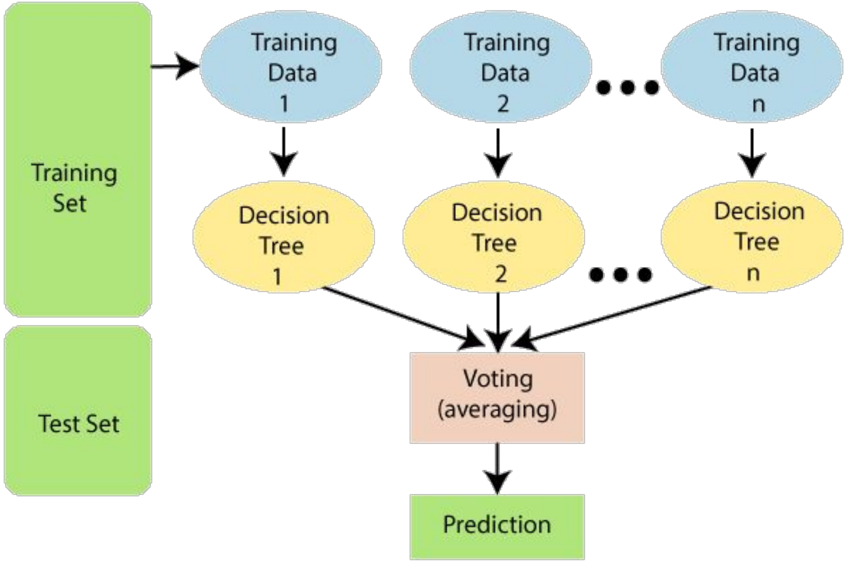


How well can we predict a student's grade average?

- Based on some EDA from previous slides, it looked like there was a relationship between certain features and student's average, but how well do these predict grades?
- To answer this we are using three models:
    - Random Forest
        - Averages the results of a group of decision trees.
    - Gradient Boosting
        - Reduces bias of weak learner (underfit).
    - Linear Regression

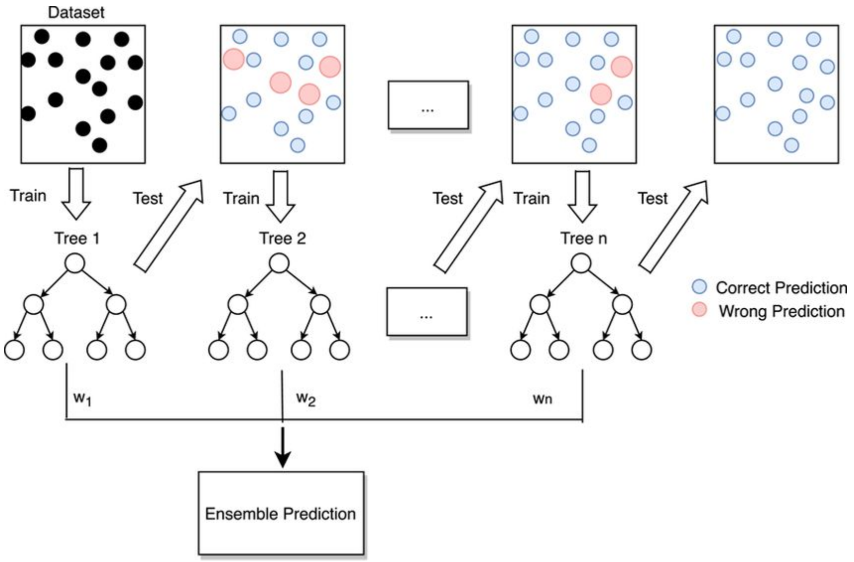
What is Random Forest?

- A group of decision trees that are forced to be as unique as possible through bootstrap sampling and random prediction selection.
  - Bootstrap Sampling: sampling with replacement.
  - Random Prediction Selection: RF selects random sample of predictors instead of using all predictor variables.
- The averaging of decision trees makes Random Forest successful because it reduces the high variance of decision trees.



What is Gradient Boosting?

- A type of machine learning boosting which is used to solve regression and classification problems.
- Target outcomes for each case are set based on the gradient of the error with respect to the prediction. Each new model takes a step in the direction that minimizes prediction error.
- The target outcome for each case in the data depends on the how much a change in prediction affects overall error:
    - If a small change in the prediction for a case causes a large drop in error, then next target outcome of the case is a high value.
    - If a small change in the prediction for a case causes no change in error, then next target outcome of the case is zero. Changing this prediction does not decrease the error.



Results

Random Forest (all predictors)	Random Forest (top predictors	Gradient Boosting (all predictors)	Gradient Boosting (top predictors)	Linear Regression (only Reading Score)
RMSE: 4.95	RMSE: 5.55	RMSE: 4.79	RMSE: 5.31	RMSE: 5.125
MSE: 24.54	MSE: 30.81	MSE: 22.97	MSE: 28.18	MSE: 26.27
R-Squared : 0.889	R-Squared : 0.860	R-Squared : 0.896	R-Squared: 0.872	R-Squared: 0.881

- Gradient Boosting Consistently performed better.
- The best scoring model included all predictors and gradient boosting, however the simple linear regression is favored due to its simplicity and high r-squared.
- 88.1% of variation in student averages can be explained by the relationship between reading score and student average.