CUNYlife

Jennifer

Weiming

Anastasiya

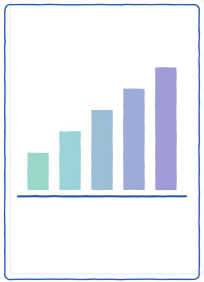
Our Project

Purpose

Understand what factors can influence a students performance on a standardized test

Target Audience

Parents and school representatives to provide better ed. programs



Data Source

Data set is used from Kaggle

Three Main Problems

1. Is there a correlation between test prep and a higher total score?
    a. **Total score = reading + math + writing**
2. Which variable highly contributes to students exam performance?
3. Predict the overall test score from inputted variables

Exploratory Data Analysis

How we analyzed our data

Exploratory Data Analysis

1. Used Kaggle dataset and made questions on what to do with this data
2. Do we predict each score? Or do we aggregate?
3. Decided to **sum up all the scores** so that we cover all the variables and see its effect across all the provided variables

	A	B	C	D	E	F	G	H
1	gender	race/ethnicity	parental level of education	lunch	test preparation	math	reading	writing
2	female	group B	bachelor's degree	standard	none	72	72	74
3	female	group C	some college	standard	completed	69	90	88
4	female	group B	master's degree	standard	none	90	95	93
5	male	group A	associate's degree	free/reduced	none	47	57	44
6	male	group C	some college	standard	none	76	78	75

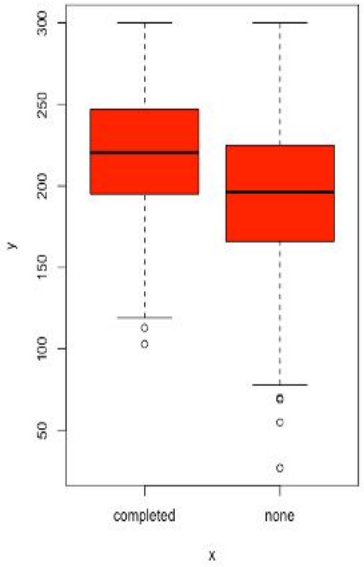
Original data file with categorical variables and 3 exam scores

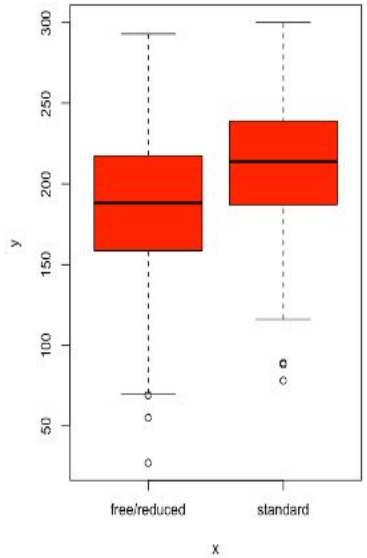
+

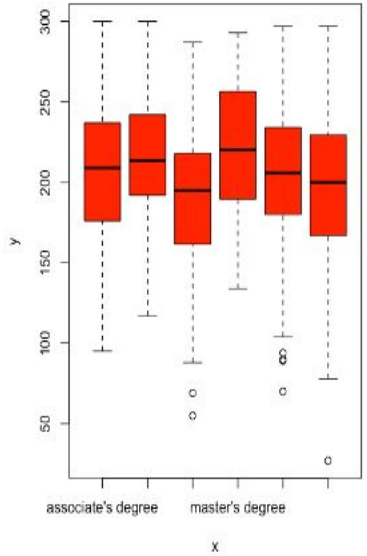
	I
	overallscore
	218
	247
	278
	148
	229

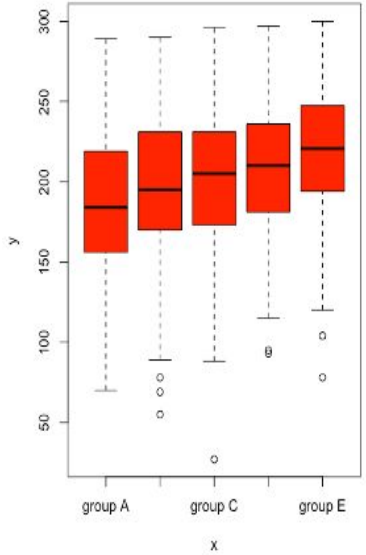
Added overall score
(math+ reading + writing)

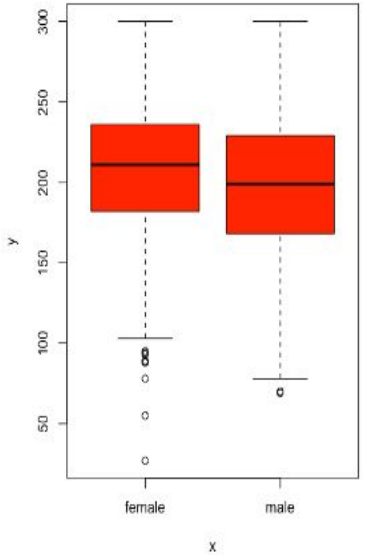
Boxplot


y
300
250
200
150
100
50
completed
none
x


y
300
250
200
150
100
50
free/reduced
standard
x


y
300
250
200
150
100
50
associate's degree
master's degree
x


y
300
250
200
150
100
50
group A
group C
group E
x


y
300
250
200
150
100
50
female
male
x

Data Mining Methods

Which methods we chose

1. Linear Regression
2. Regression Tree
3. Random Forest

Linear Regression

- Used **70%** of the data for **training** and the **30%** for **testing**
- Used the **overallscore** as the response variable for the linear model
-

```
Call:
lm(formula = overallscore ~ ., data = student.train)

Residuals:
     Min       1Q   Median       3Q      Max
-146.013  -24.159    1.298   25.817   82.352

Coefficients:
                                          Estimate Std. Error
(Intercept)                                197.683      6.491
gendermale                                 -10.588      2.848
race.ethnicitygroup B                        9.340      5.905
race.ethnicitygroup C                       12.811      5.552
race.ethnicitygroup D                       20.256      5.678
race.ethnicitygroup E                       23.553      6.197
parental.level.of.educationbachelor's degree 10.088      5.121
parental.level.of.educationhigh school      -16.984      4.357
parental.level.of.educationmaster's degree    8.239      6.479
parental.level.of.educationsome college      -1.704      4.287
parental.level.of.educationsome high school -13.429      4.510
lunchstandard                               26.226      2.952
test.preparation.coursenone                 -24.052      2.980
                                          t value Pr(>|t|)
```

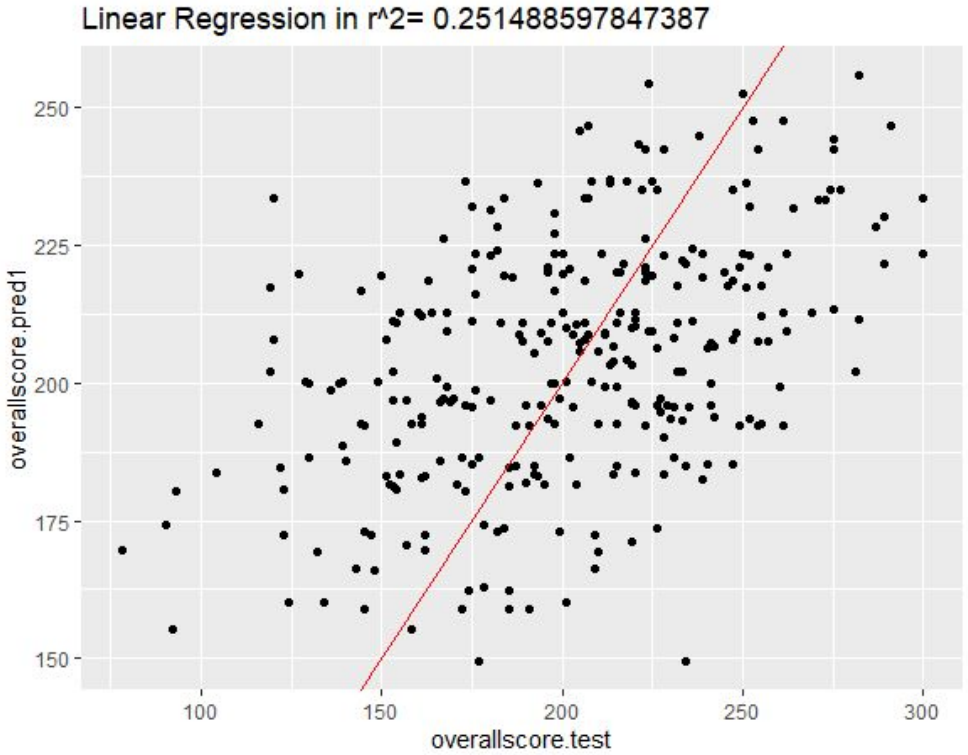
Stepwise Forward Selection

```
> step(fitstart, direction = "forward", scope = formula(M1))
Start:  AIC=5262.93
overallscore ~ 1

                             Df Sum of Sq     RSS    AIC
+ lunch                       1     98514 1187037 5209.1
+ test.preparation.course     1     95251 1190300 5211.0
+ parental.level.of.education 5     75034 1210517 5230.8
+ race.ethnicity              4     50528 1235023 5242.9
+ gender                      1     18207 1267344 5254.9
<none>                                   1285551 5262.9

Step:  AIC=5209.12
overallscore ~ lunch

                             Df Sum of Sq     RSS    AIC
+ test.preparation.course     1    102365 1084673 5148.0
+ parental.level.of.education 5     78595 1108442 5171.2
+ race.ethnicity              4     48484 1138553 5187.9
+ gender                      1     20568 1166470 5198.9
<none>                                   1187037 5209.1

Step:  AIC=5148
```

```
Call:
lm(formula = overallscore ~ lunch + test.preparation.course +
    parental.level.of.education + race.ethnicity + gender, data = student.train)
```

Linear Regression with Plot

The **MSE** is **1428**



Regression Tree

- **1000 records** with 1 continuous response "**overallscore**" (the total score of Math, Reading and Writing
- 70% training and 30% testing
- Predict "overallscore" using Input - 5 categorical variables

	gender	race.ethnicity	parental.level.of.education	lunch	test.preparation.course	overallscore
1	female	group B	bachelor's degree	standard	none	218
2	female	group C	some college	standard	completed	247
3	female	group B	master's degree	standard	none	278
4	male	group A	associate's degree	free/reduced	none	148
5	male	group C	some college	standard	none	229
6	female	group B	associate's degree	standard	none	232

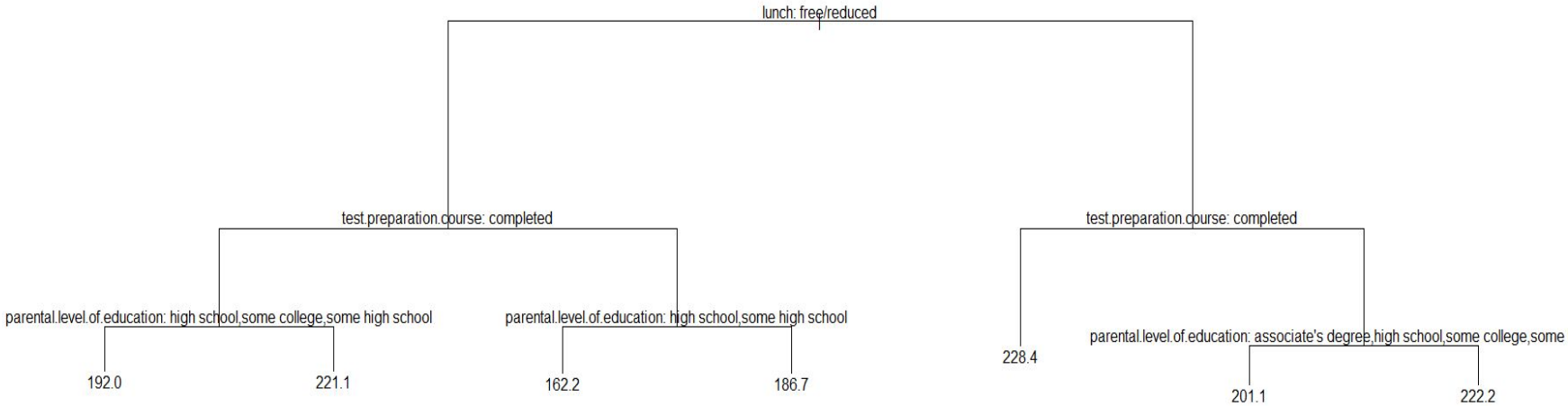
Regression Tree

- Use **K-fold cross-validation** to determine the optimal subtree
- **cv.tree()** function reports the number of terminal nodes of each tree considered and the corresponding error rate.

```
$size
[1] 7 6 5 4 3 2 1

$dev
[1] 1153629 1196723 1184913 1181097 1187226 1218758 1297831

$k
[1]     -Inf 18286.39 20924.49 22009.07 47050.62 55993.23 98513.72

$method
[1] "deviance"

attr(,"class")
[1] "prune"         "tree.sequence"
```

Regression Tree

- **MSE** between **True value** and **Predicted value** is 1517.304



Two Ensemble Algorithms: Bagging

Bagging: improves the accuracy & stability of our Regression Tree model

- **Bootstrap Aggregating** - sample with replacement

Bagging uses all 5 predictors:

- Race
- Ethnicity
- Parental Level of Education
- Gender
- Lunch

Mean value: shows the performance of bagging

```
>
> bag.model=randomForest(overallscore~ gender+ lunch + race.ethn
icity+ parental.level.of.education+ test.preparation.course,data
=survey, subset=train, mtry=5, importance=TRUE)
> bag.model

Call:
 randomForest(formula = overallscore ~ gender + lunch + race.eth
nicity +      parental.level.of.education + test.preparation.cou
rse, data = survey,      mtry = 5, importance = TRUE, subset = t
rain)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 5

          Mean of squared residuals: 1826.7
                    % Var explained: 0.53
```

```
> #performance of the bagging
> yhat.bag = predict(bag.model,newdata=survey[-train,]
> mean((yhat.bag-model.test)^2)
[1] 1669.201
>
```

Two Ensemble Algorithms: Random Forest

Random Forest: Bagging + Decision Tree

- uses a random split with replacement at 2 predictors (5 / 3 = 2)

```
> #random forest
> set.seed(1)
> rf.model=randomForest(overallscore~ gender+ lunch +
 test.preparation.course,data=survey,subset=train,mtry
 mtry should be 3
> yhat.rf = predict(rf.model,newdata=survey[-train,])
> mean((yhat.rf-model.test)^2)
[1] 1484.04
>
```

Mean value: shows the performance of Random Forest

- Mean is lower for **Random Forest => less errors**

Random Forest: was used to help evaluate which variable was more important for the total test score

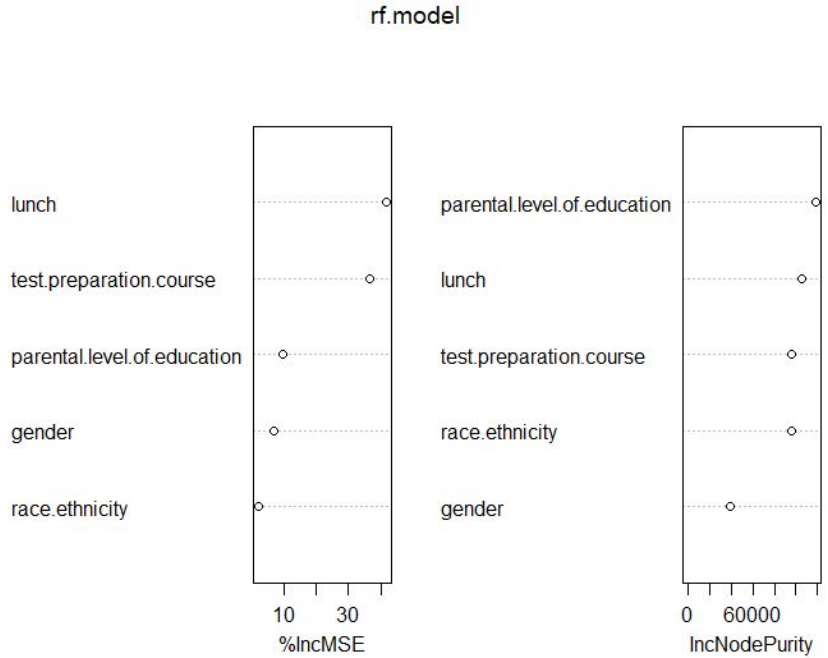
Importance of Each Variable

The importance of each variable was evaluated using:

- **importance()** - mean decrease of accuracy in predictions
- **varImpPlot()** - total decrease in node impurity

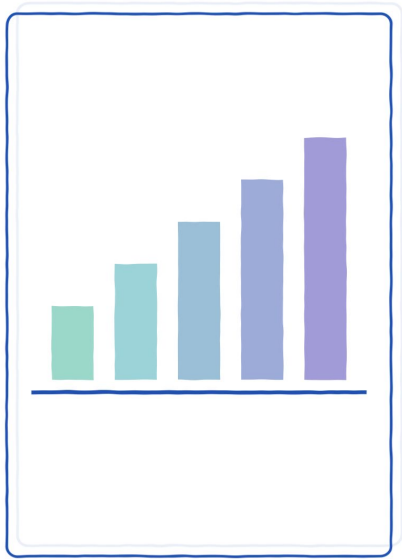
Lunch, test prep, parental level of education give the best prediction and contribute most to the model

```
> importance(rf.model)
                             %IncMSE IncNodePurity
gender                      6.778136      38306.71
lunch                      41.911091     105822.48
race.ethnicity              2.037405      96102.57
parental.level.of.education 9.508892     118778.41
test.preparation.course    36.468568      96367.22
>
```



Final Model

Choosing the optimal model



Final Model & Results

- **Linear regression model** is the better model with **lowest MSE**

```
Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  197.683      6.491  30.455  < 2e-16 ***
gendermale                                   -10.588      2.848  -3.717 0.000218 ***
race.ethnicitygroup B                          9.340      5.905   1.582 0.114151
race.ethnicitygroup C                         12.811      5.552   2.308 0.021315 *
race.ethnicitygroup D                         20.256      5.678   3.567 0.000386 ***
race.ethnicitygroup E                         23.553      6.197   3.801 0.000157 ***
parental.level.of.educationbachelor's degree  10.088      5.121   1.970 0.049257 *
parental.level.of.educationhigh school       -16.984      4.357  -3.898 0.000106 ***
parental.level.of.educationmaster's degree     8.239      6.479   1.272 0.203903
parental.level.of.educationsome college       -1.704      4.287  -0.397 0.691206
parental.level.of.educationsome high school   -13.429      4.510  -2.978 0.003006 **
lunchstandard                                 26.226      2.952   8.886  < 2e-16 ***
test.preparation.coursenone                  -24.052      2.980  -8.071 3.11e-15 ***
---
```

Increase	Decrease
race.ethnicity groupB	gender male
race.ethnicity groupC	parental.level.of.education high school
race.ethnicity groupD	parental.level.of.education some college
race.ethnicity groupE	parental.level.of.education some high school
parental.level.of.education bachelor's degree	test.preparation.course none
parental.level.of.education master's degree	
lunch standard	

1. Is there a correlation between test prep and a higher total score? **Yes**
2. Which variable highly contributes to students exam performance? **Multiple variables**
3. Predict the overall test score from inputted variables. **With all other predictors held fixed, an increase to a predictor in the increase column or a decrease to a predictor in decrease column would result to an increase to the student's overall score.**

CUNYlife

Thank You!