Final Bootcamp Project

By Rafael Rojas

Table of Contents

Choosing the Dataset

- This dataset was extracted from ‘Kaggle’, it consist on the test scores secured by the students on their midterms.
- One of the goals is to understand the influence of some ’external’ factors on the academic performance.
- We’ll also try to predict which students might be going to summer school with the help of these ‘external’ factors.

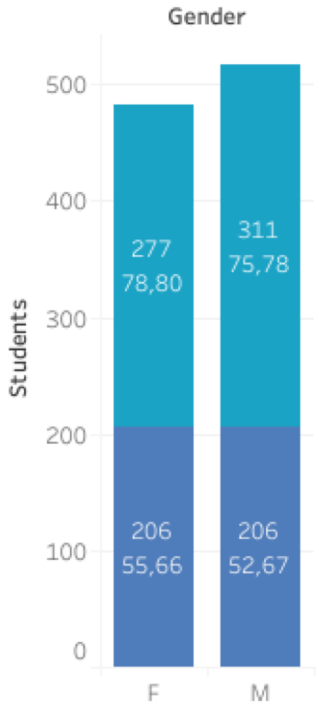
The Dataset

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	male	group A	high school	standard	completed	67	67	63
1	female	group D	some high school	free/reduced	none	40	59	55
2	male	group E	some college	free/reduced	none	59	60	50
3	male	group B	high school	standard	none	77	78	68
4	male	group E	associate's degree	standard	completed	78	73	68
...	...	...	...	...	...	...	...	...
995	male	group C	high school	standard	none	73	70	65
996	male	group D	associate's degree	free/reduced	completed	85	91	92
997	female	group C	some high school	free/reduced	none	32	35	41
998	female	group C	some college	standard	none	73	74	82
999	male	group A	some college	standard	completed	65	60	62

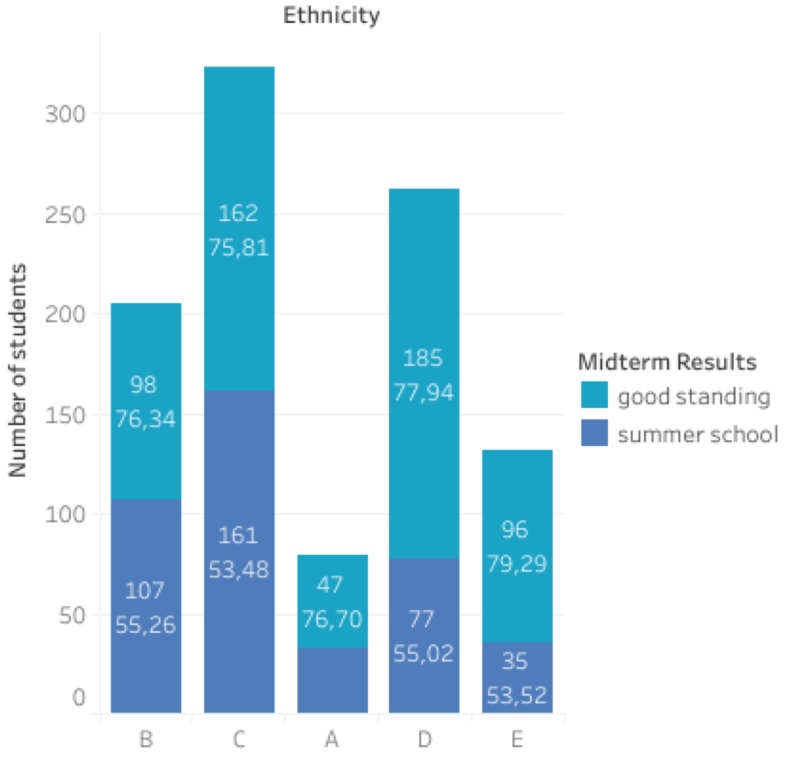
1000 rows × 8 columns

01
Exploratory Data Analysis

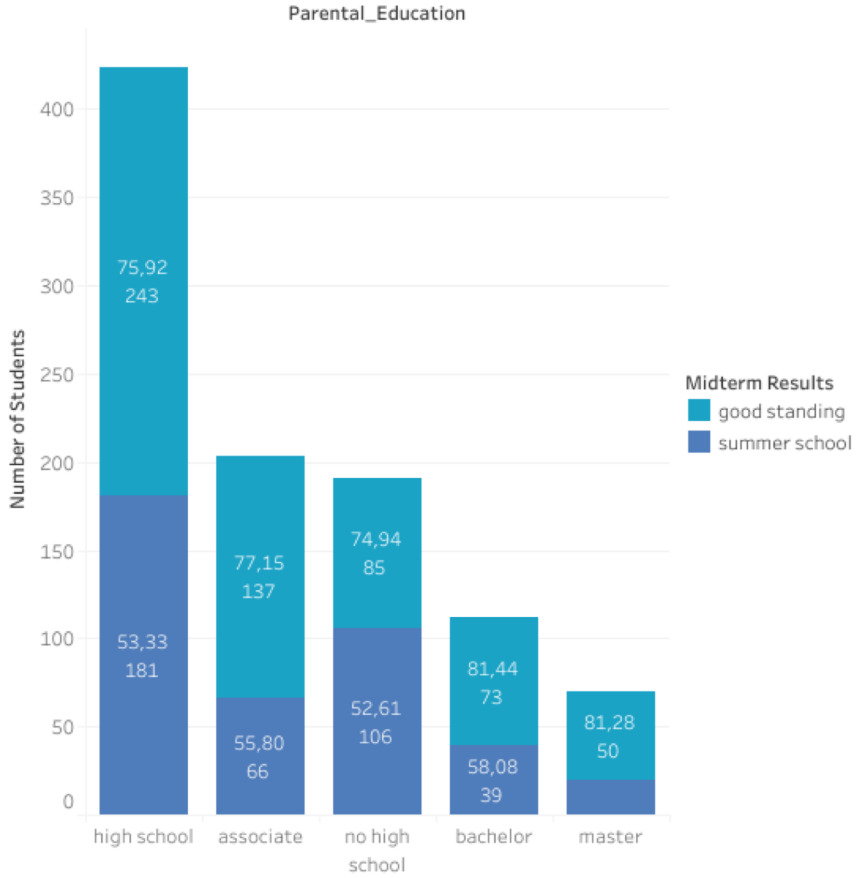
Number of students by Gender.



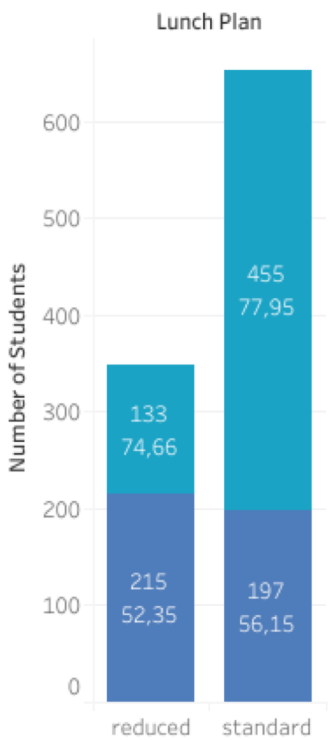
Number of students by Ethnicity



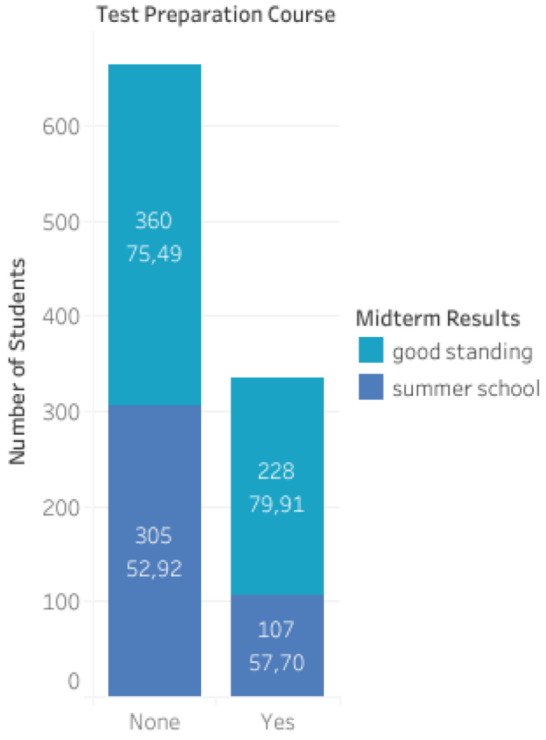
Number of students by PE



Number of Students by LP



Number of Students by TPC



02

Data Cleaning

Cleaning the Dataset

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	male	group A	high school	standard	completed	67	67	63
1	female	group D	some high school	free/reduced	none	40	59	55
2	male	group E	some college	free/reduced	none	59	60	50
3	male	group B	high school	standard	none	77	78	68
4	male	group E	associate's degree	standard	completed	78	73	68

	gender	ethnicity	parental_education	lunch	test_preparation_course	math_score	reading_score	writing_score	pass/fail_math	pass/fail_reading	pass/fail_writing	midterm_results	failed_courses
0	M	A	high school	standard	yes	67	67	63	passed	passed	passed	good standing	good standing
1	F	D	no high school	reduced	no	40	59	55	failed	failed	failed	summer school	math/reading
2	M	E	high school	reduced	no	59	60	50	failed	passed	failed	summer school	math/writing
3	M	B	high school	standard	no	77	78	68	passed	passed	passed	good standing	good standing
4	M	E	associate	standard	yes	78	73	68	passed	passed	passed	good standing	good standing

03
Create and validate the model

Creating the model:

Regression

- Linear Regression
- KNN Regressor
- Decision Tree Regressor

Classification

- Logistic Regression
- KNN Classifier
- Random Forest Classifier

Validate the models with Error Metrics

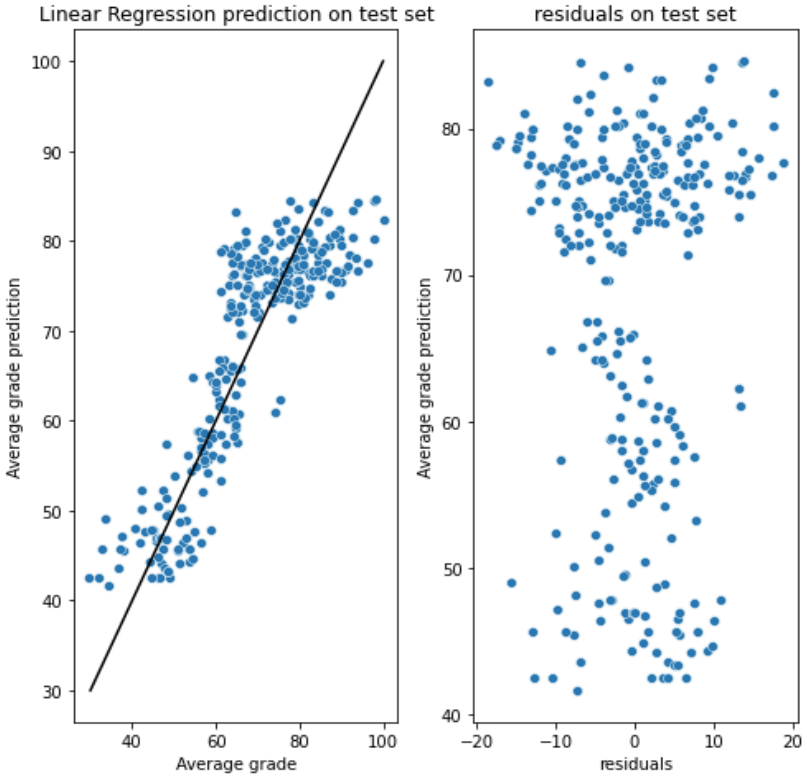
Linear Regression - Error metrics		
Error Metric	**Train**	**Test**
Mean Error	0,01	-0,07
Mean Absolute Error	5,63	5,81
Mean Squared Error	50,3	53,19
Root Mean Squared Error	7,09	7,29
Mean Absolute Percentual Error	8,67	8,86
R^2	0,76	0,74

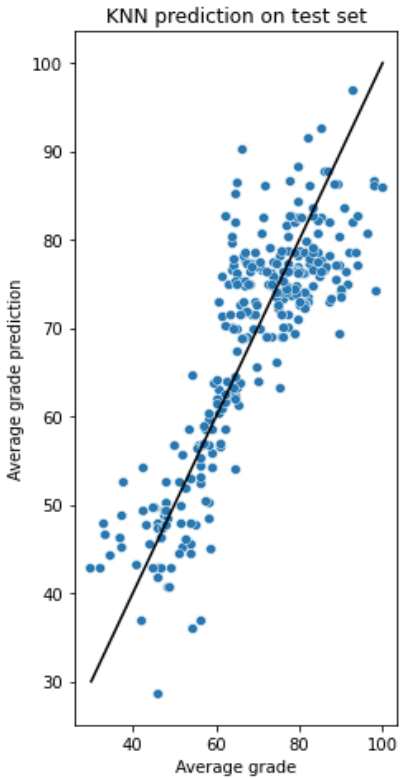
KNN Regressor - Error metrics		
Error Metric	**Train**	**Test**
Mean Error	-0,06	0,19
Mean Absolute Error	3,77	6,08
Mean Squared Error	32,39	62,06
Root Mean Squared Error	5,69	7,88
Mean Absolute Percentual Error	5,63	9,31
R^2	0,85	0,70

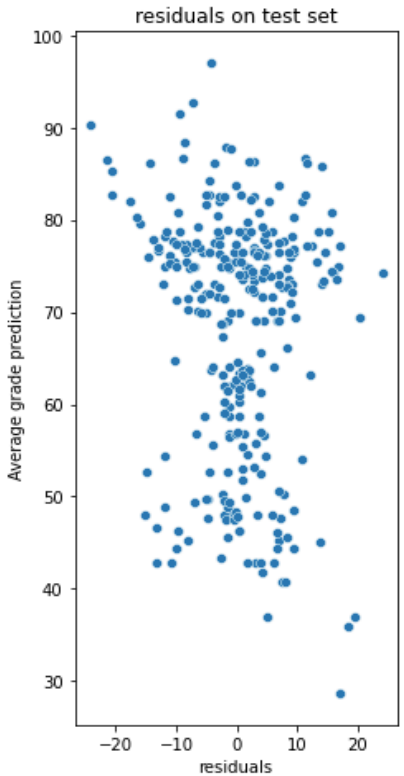
Logistic Regression - Error metrics		
Error Metric	**Train**	**Test**
Accuracy score	0,69	0,68
Precision score	0,69	0,74
Recall score	0,82	0,82
F1 score	0,75	0,75
Kappa score	0,35	0,31

Random Forest Classifier - Error metrics		
Error Metric	**Train**	**Test**
Accuracy score	0,7	0,67
Precision score	0,69	0,55
Recall score	0,55	0,55
F1 score	0,61	0,61
Kappa score	0,37	0,30

Best performing models



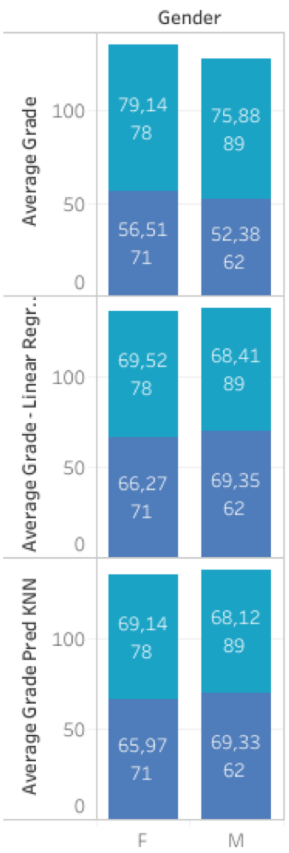




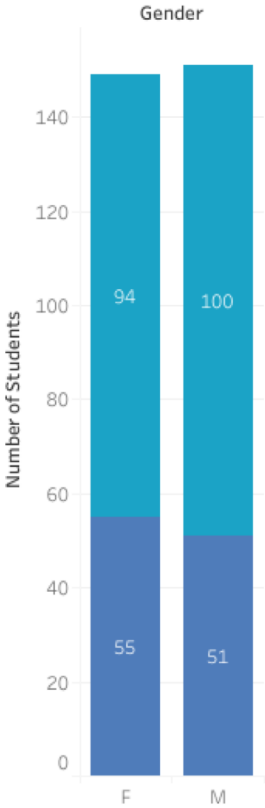
04

Analyze predictions

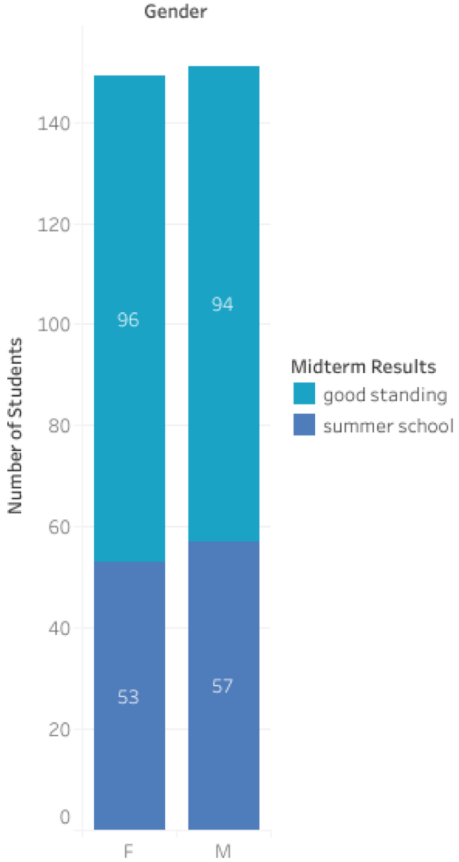
Average Grade by Gender



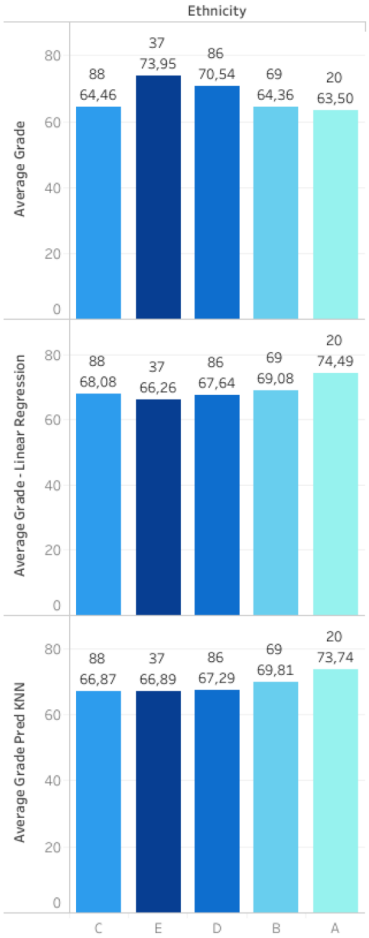
Pass/Fail Logistic Regression by Gender



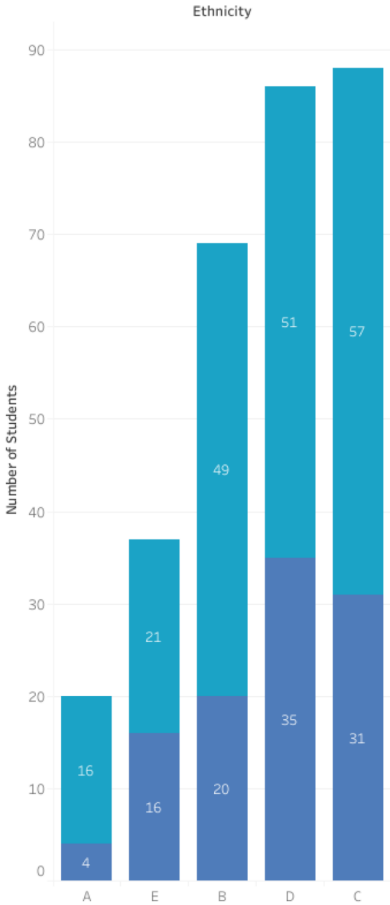
Pass/Fail Random Forest by Gender



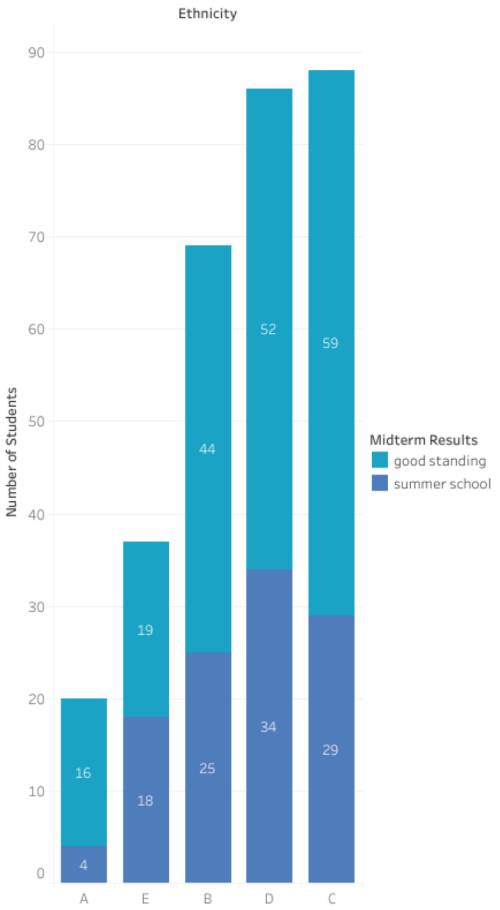
Average Grade by Ethnicity



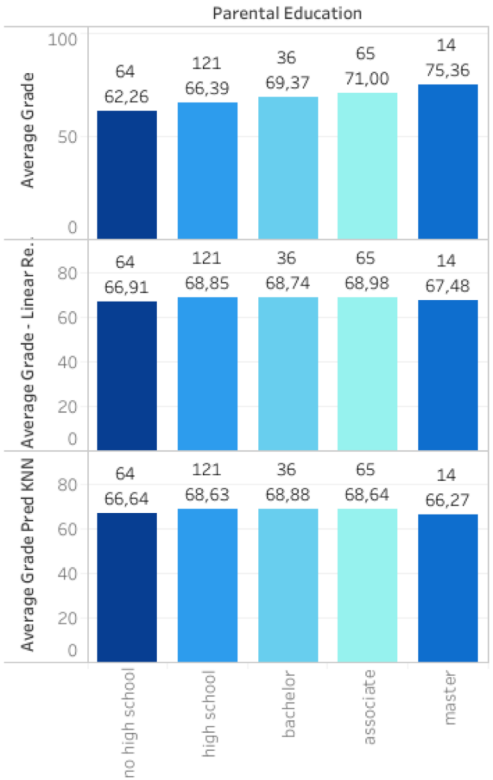
Pass/Fail Logistic Regression by Ethnicity



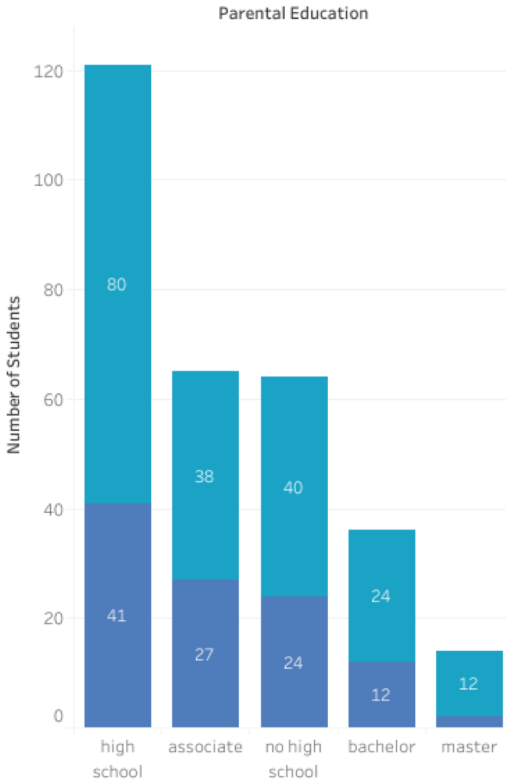
Pass/Fail Random Forest by Ethnicity



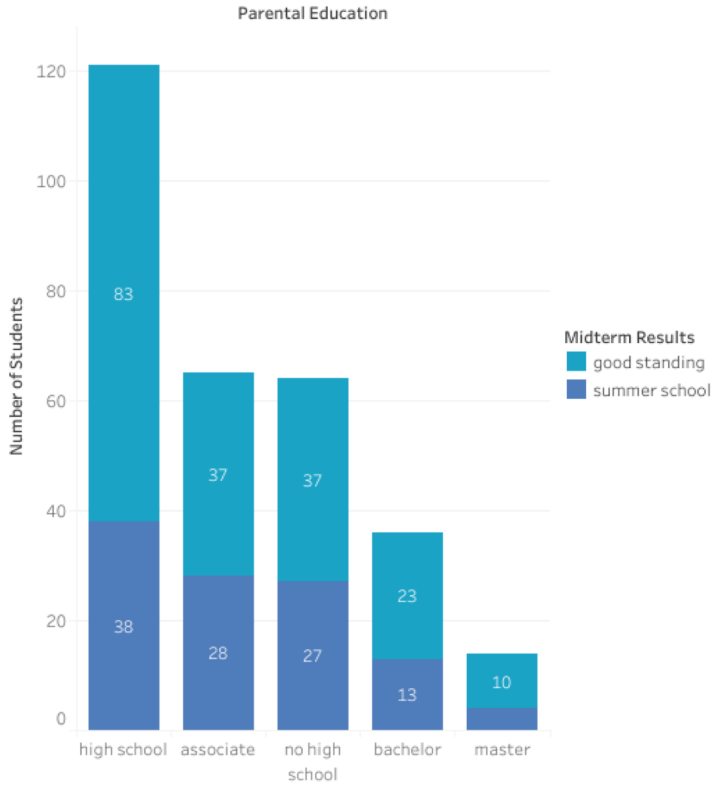
Average Grade by Parental Education



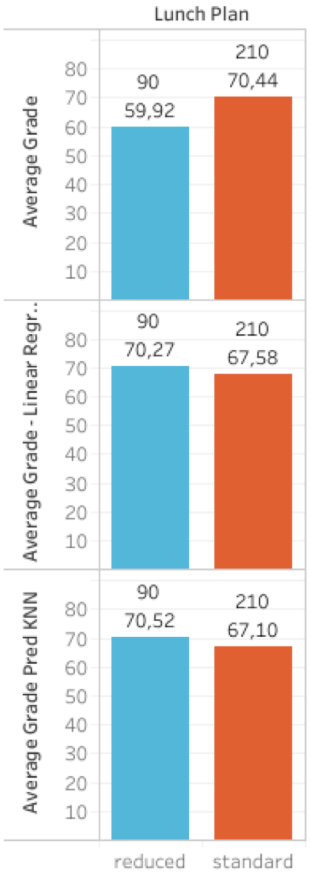
Pass/Fail Logistic Regression by Parental Education



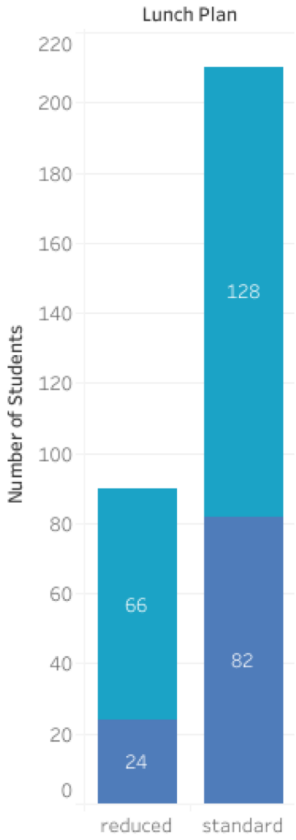
Pass/Fail Random Forest by Parental Education



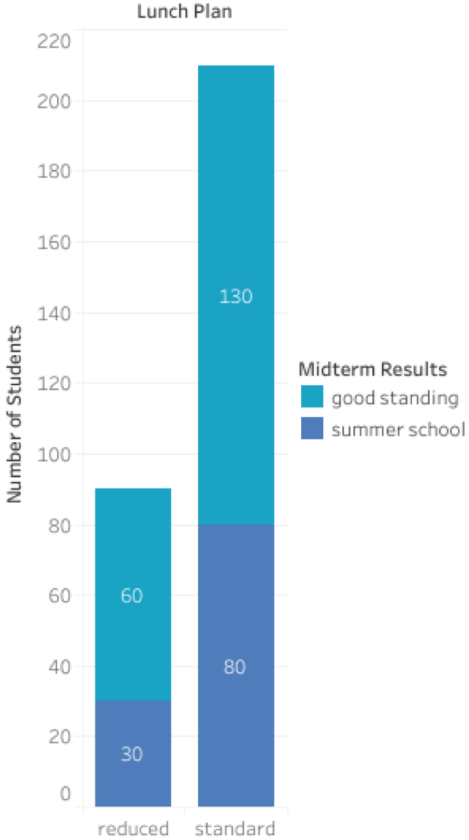
Average Grade by Lunch Plan



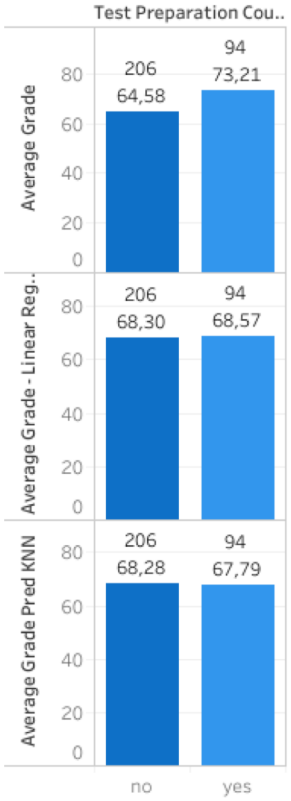
Pass/Fail Logistic Regression by Lunch Plan



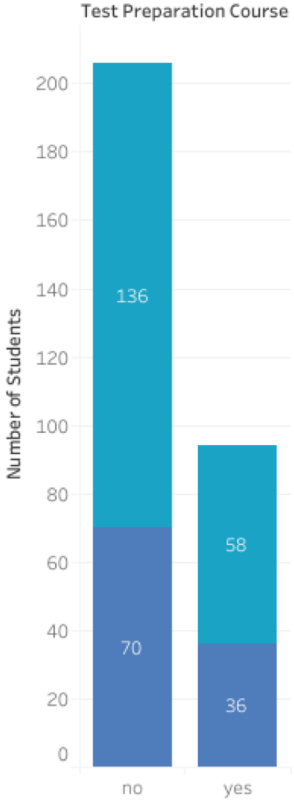
Pass/Fail Logistic Regression by Lunch Plan



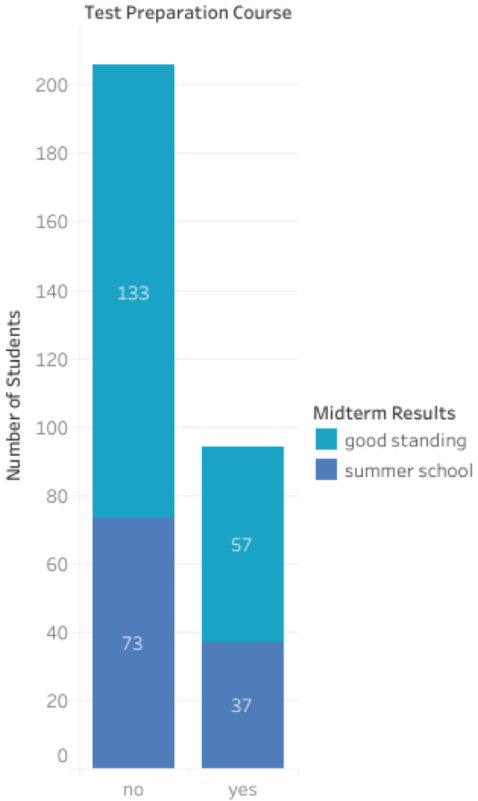
Average Grade by Test Preparation Course



Pass/Fail Logistic Regression by Test Preparation Course



Pass/Fail Random Forest by Test Preparation Course



Conclusions

THANK YOU FOR LISTENING!