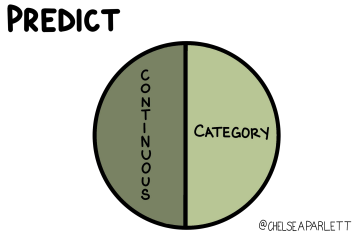


Linear Regression

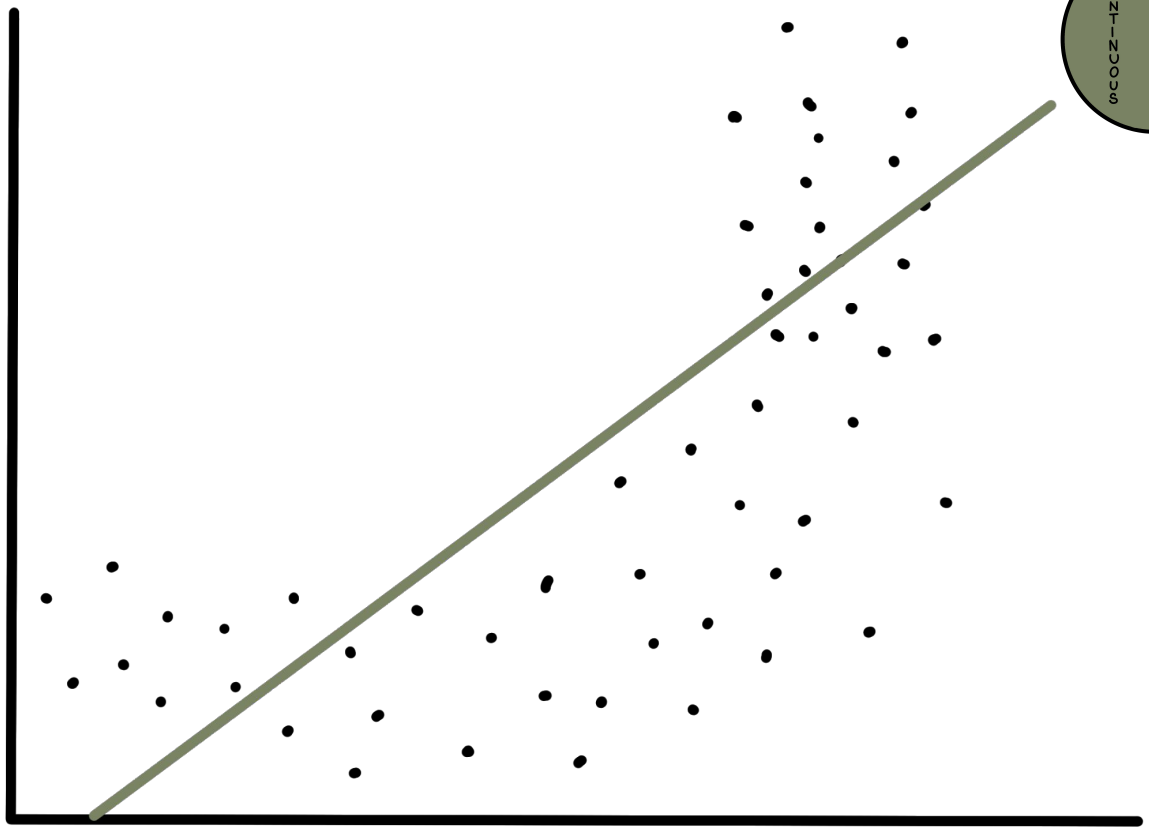
Chelsea Parlett-Pelleriti

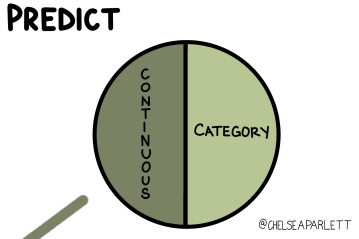
What

- Use **multiple variables** (can be continuous, categorical, or both) to predict a **continuous variable**.
- Use a line (or a plane) to describe the relationship between these variables.

Assumptions

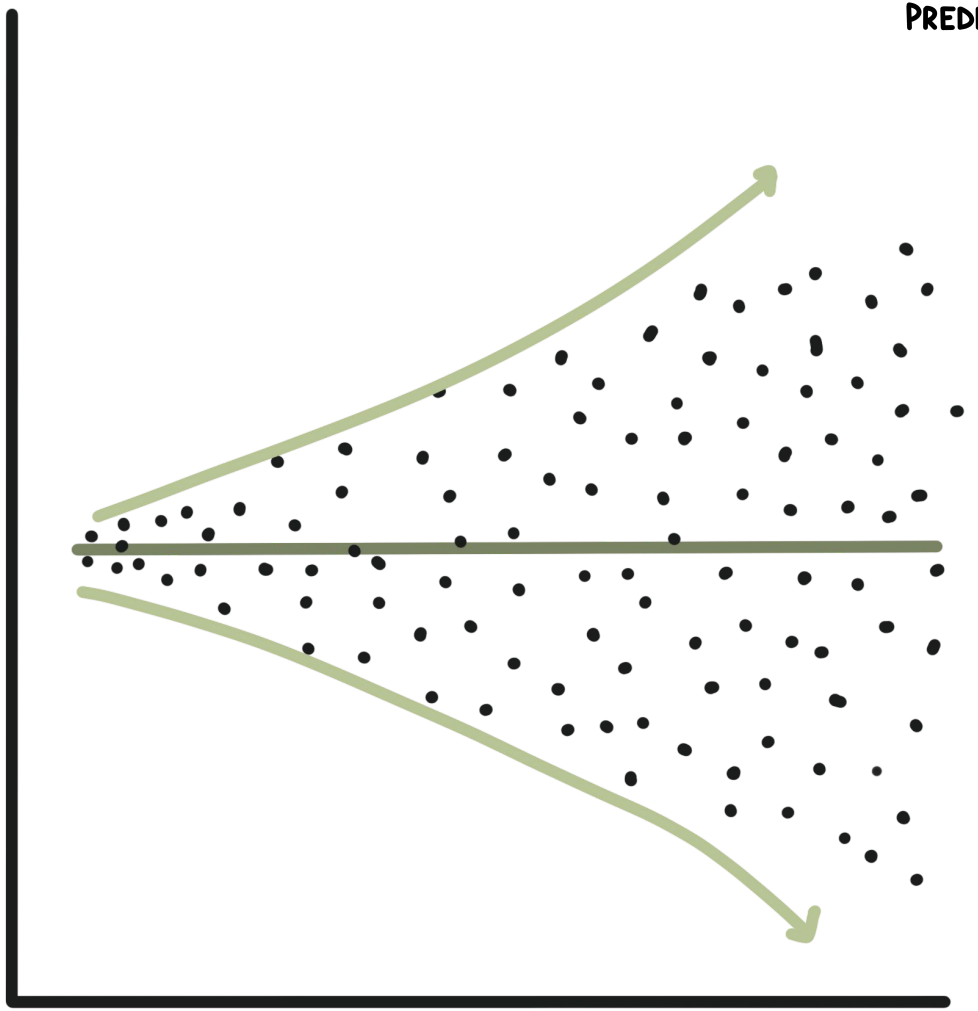
- The relationship between your variables is linear

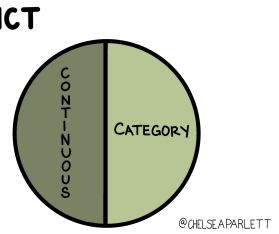




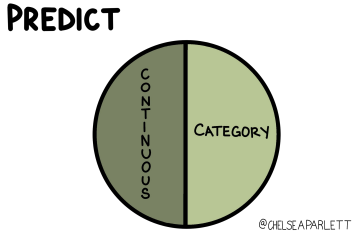
Assumptions

- Homoscedasticity
  - Is the mode worse in some areas than others?

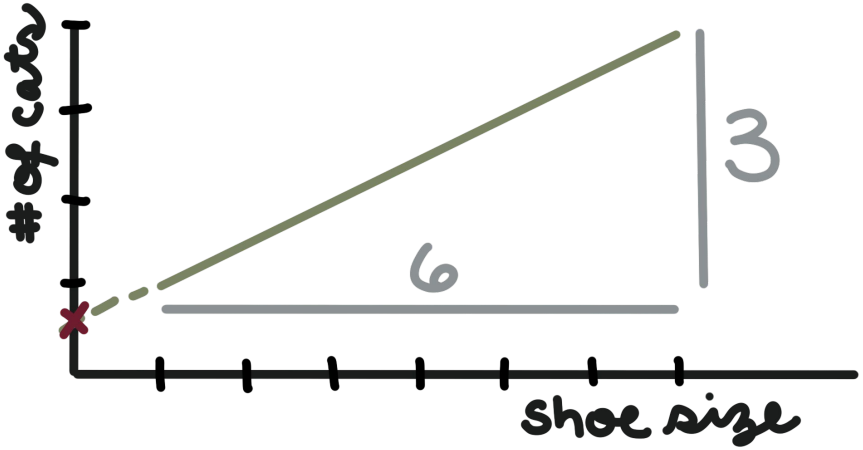


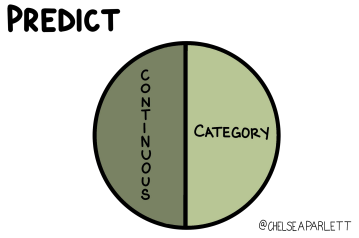


How



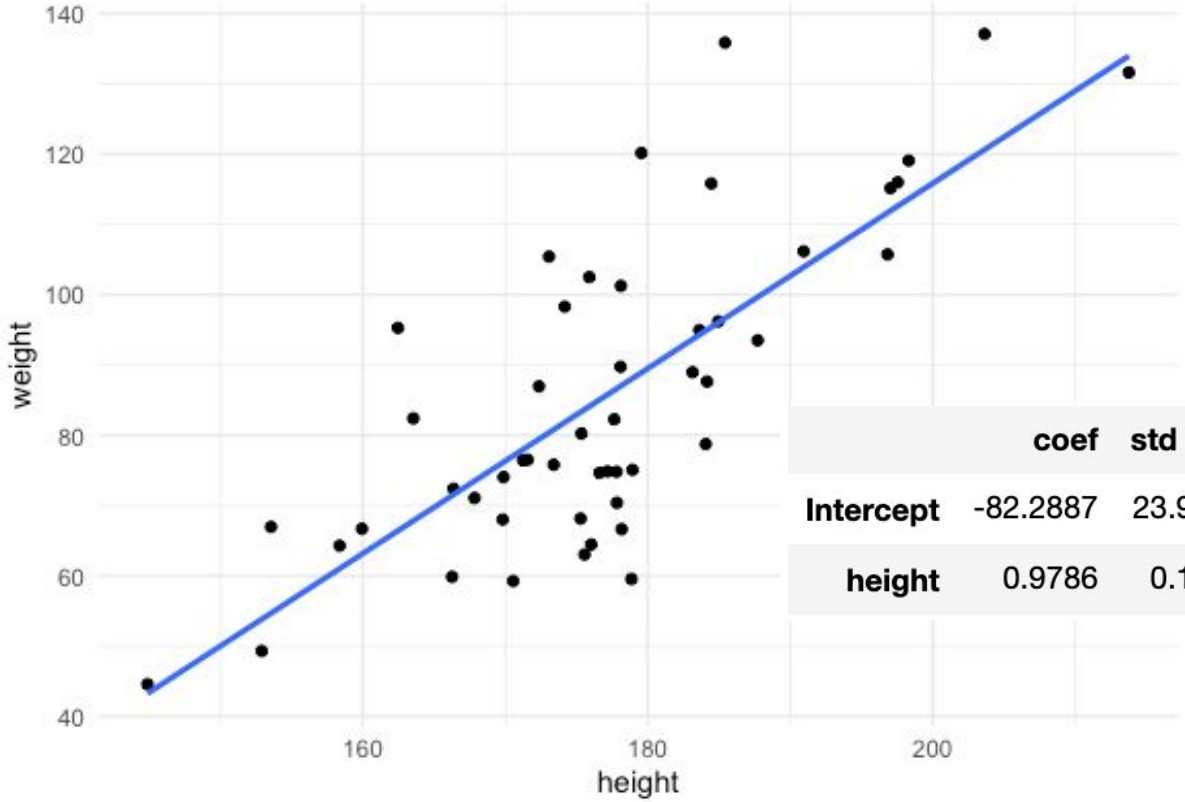
- Y = mx + b
- Y = mx + nz + b
- Slope tells you how variables change together
- Intercept tells you what would happen if all your predictors were 0.



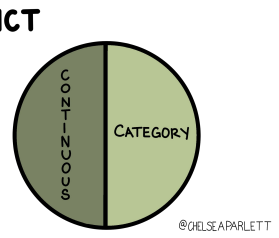


Simple example

Predict weight by height

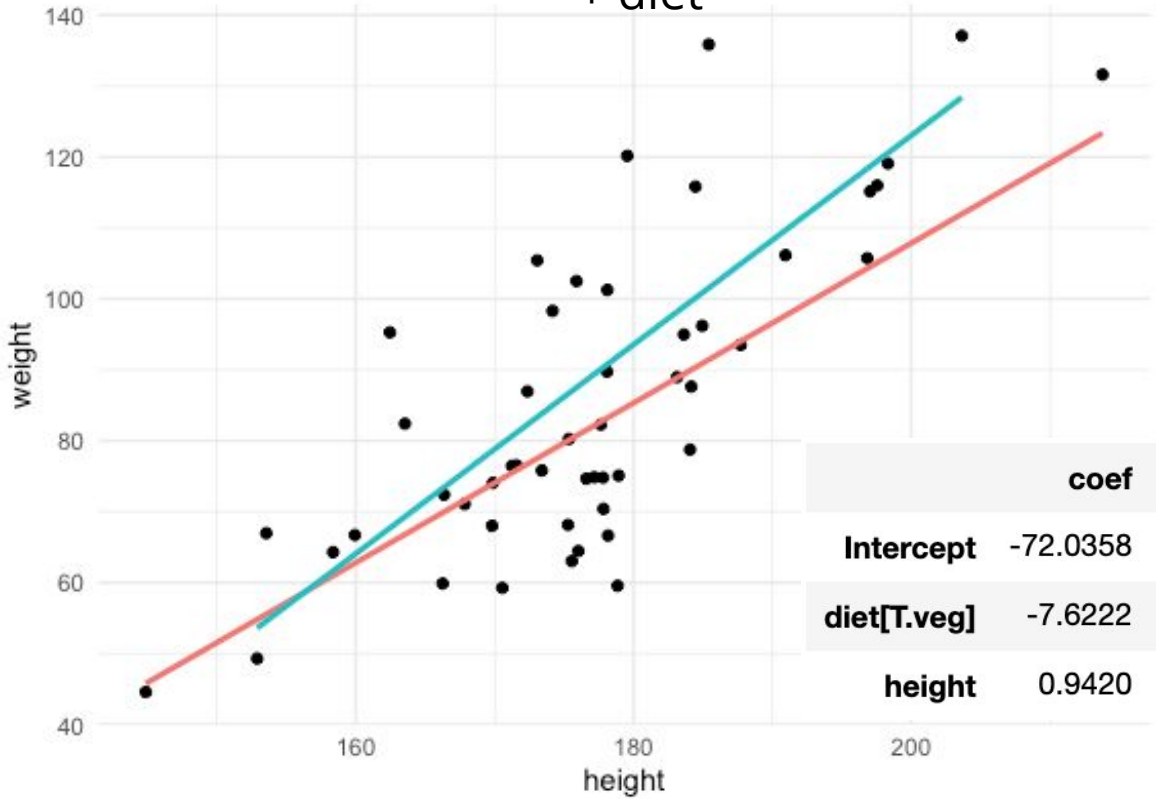


	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-82.2887	23.932	-3.438	0.001	-130.408	-34.170
height	0.9786	0.136	7.207	0.000	0.706	1.252



Simple example

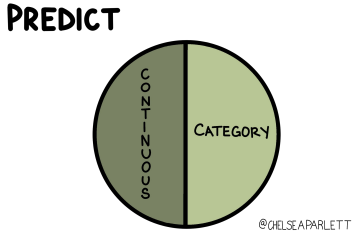
Predict weight by height + diet



	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-72.0358	23.933	-3.010	0.004	-120.182	-23.889
diet[T.veg]	-7.6222	4.025	-1.894	0.064	-15.719	0.474
height	0.9420	0.134	7.046	0.000	0.673	1.211

Simple example

Predict weight by height + diet + age



	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-57.4078	26.662	-2.153	0.037	-111.076	-3.740
diet[T.veg]	-8.2640	4.038	-2.046	0.046	-16.393	-0.135
height	0.8948	0.139	6.460	0.000	0.616	1.174
age	-0.1298	0.106	-1.219	0.229	-0.344	0.085

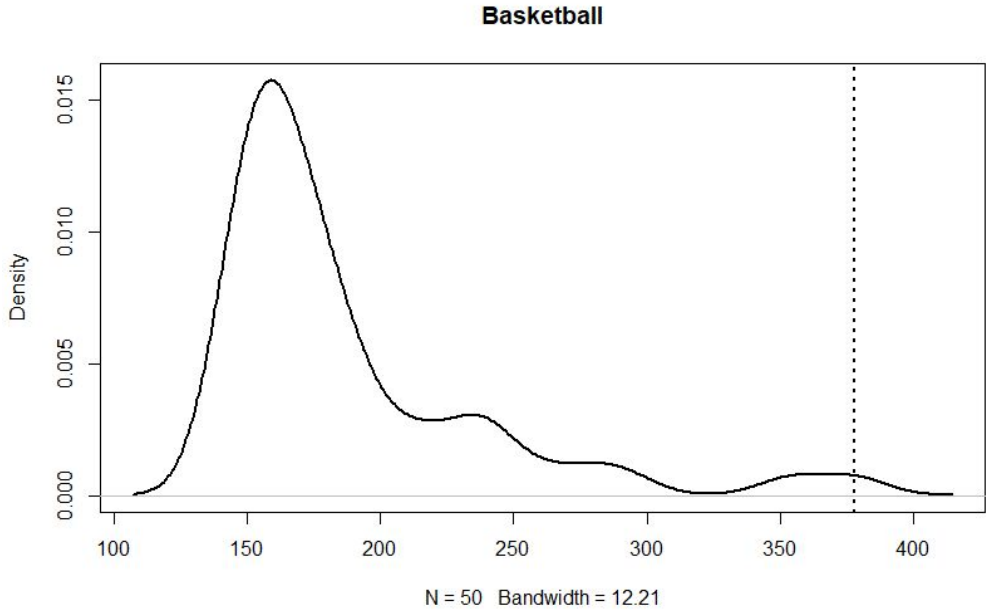
Who is the GOAT? ⚾ 🏀

378 three-pointers

53 home-runs

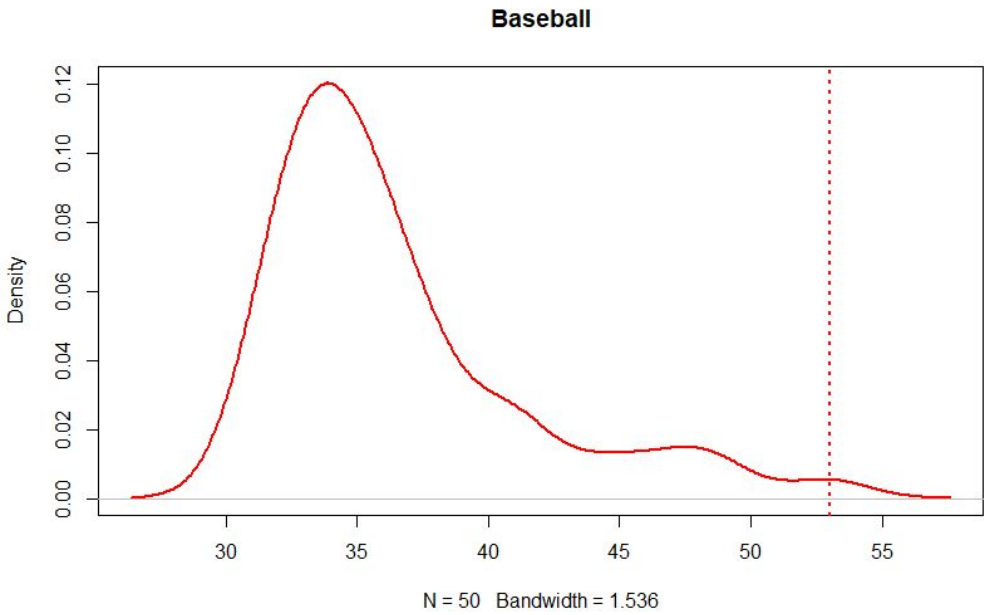
2018-19 NBA Regular Season: Total 3-Pointers Made Leaders
2019 MLB Player Batting Stats | Home Runs

Who is the GOAT?



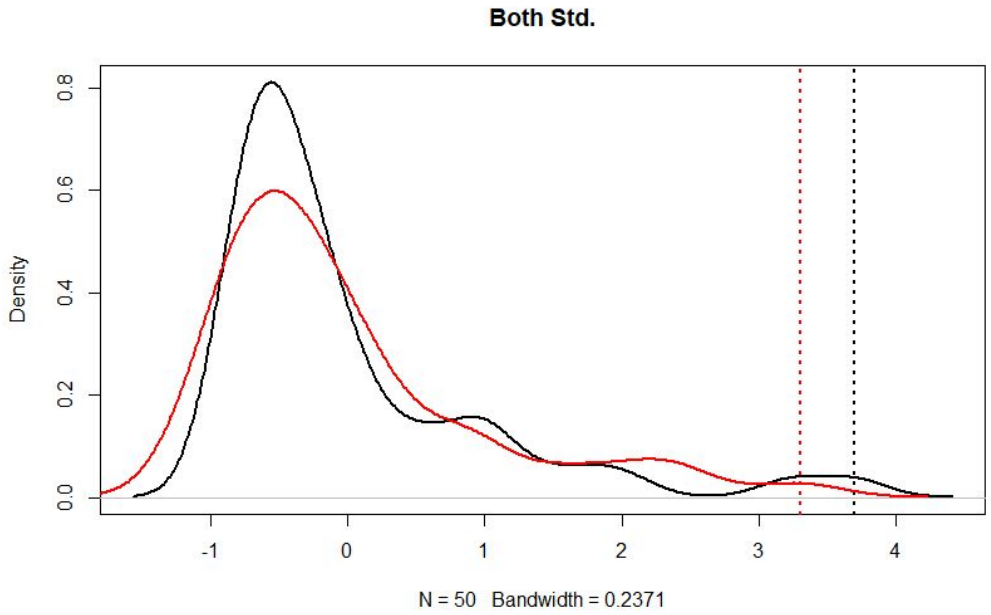
2018-19 NBA Regular Season: Total 3-Pointers Made Leaders
2019 MLB Player Batting Stats | Home Runs

Who is the GOAT?



2018-19 NBA Regular Season: Total 3-Pointers Made Leaders
2019 MLB Player Batting Stats | Home Runs

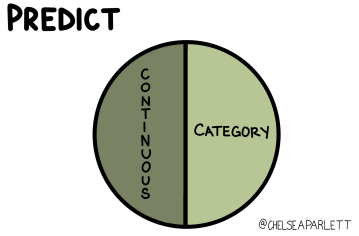
Who is the GOAT?



2018-19 NBA Regular Season: Total 3-Pointers Made Leaders
2019 MLB Player Batting Stats | Home Runs

Simple example

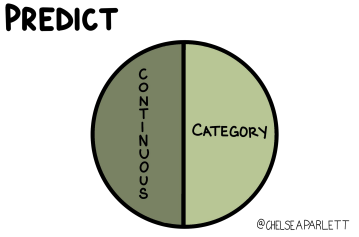
Predict weight by height + diet + age



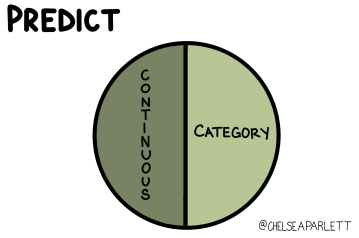
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-57.4078	26.662	-2.153	0.037	-111.076	-3.740
diet[T.veg]	-8.2640	4.038	-2.046	0.046	-16.393	-0.135
height	0.8948	0.139	6.460	0.000	0.616	1.174
age	-0.1298	0.106	-1.219	0.229	-0.344	0.085

Standardizing variables

For understanding and for model convergence



Inferential way to do LR



How is it different from prediction?

How to interpret coefficients (same in both)

http://www.statsmodels.org/stable/index.html

https://towardsdatascience.com/bayesian-linear-regression-in-python-using-machine-learning-to-predict-student-grades-part-2-b72059a8ac7e