University of Pittsburgh

INFSCI 0530: Decision Making in Sports

Fall 2021

Overfitting

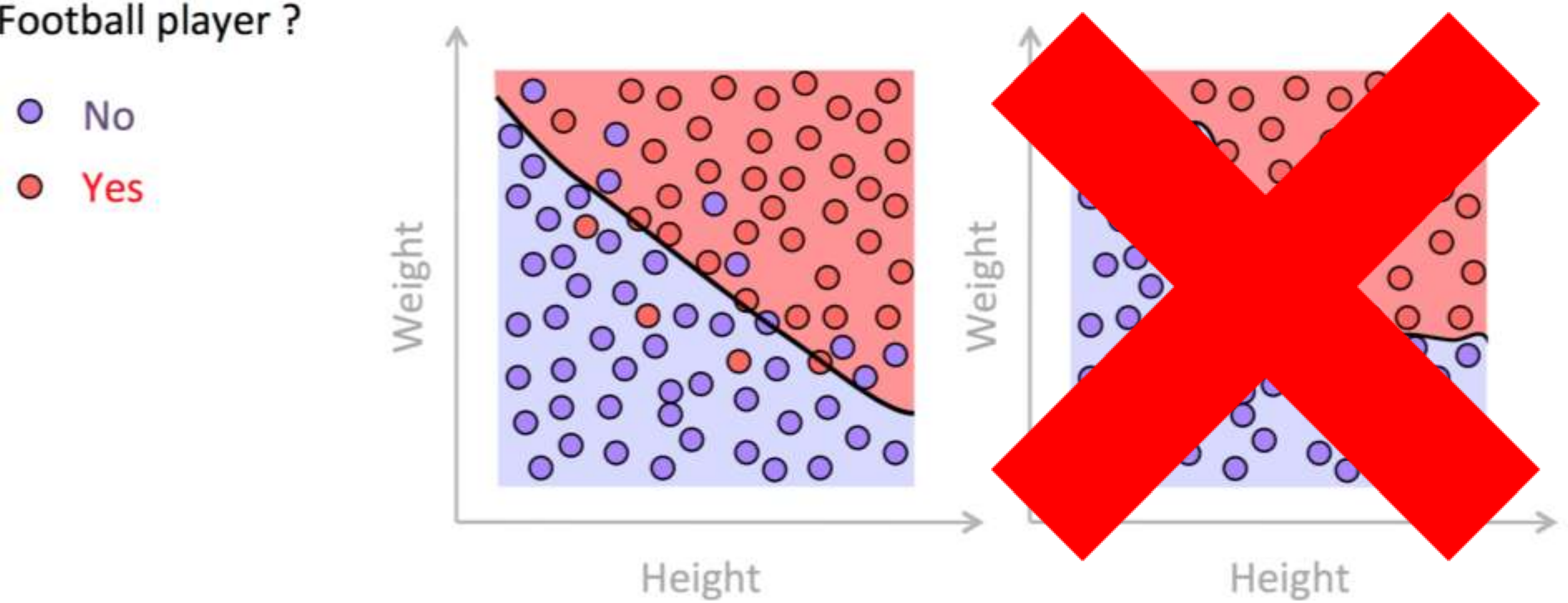
- When learning a model we have a set of data (training set) that we use to learn the model parameters
- The evaluation of the model needs to happen out-of-sample, i.e., on a different set that was not used for learning model parameters
- One of the most common problems during training is tying the model to the training set
  - Overfitting

Overfitting

- When a model is over fitted it is not expected to perform well to new data
  - It is not generalizable
- Overfitting occurs when the model chosen is too complex that ends up describing the noise in the data instead of the trend
  - E.g., too many parameters relative to the size of the training dataset
  - An over fitted model *memorizes* the training instances and does not learn the general trend in them

Overfitting

Football player ?

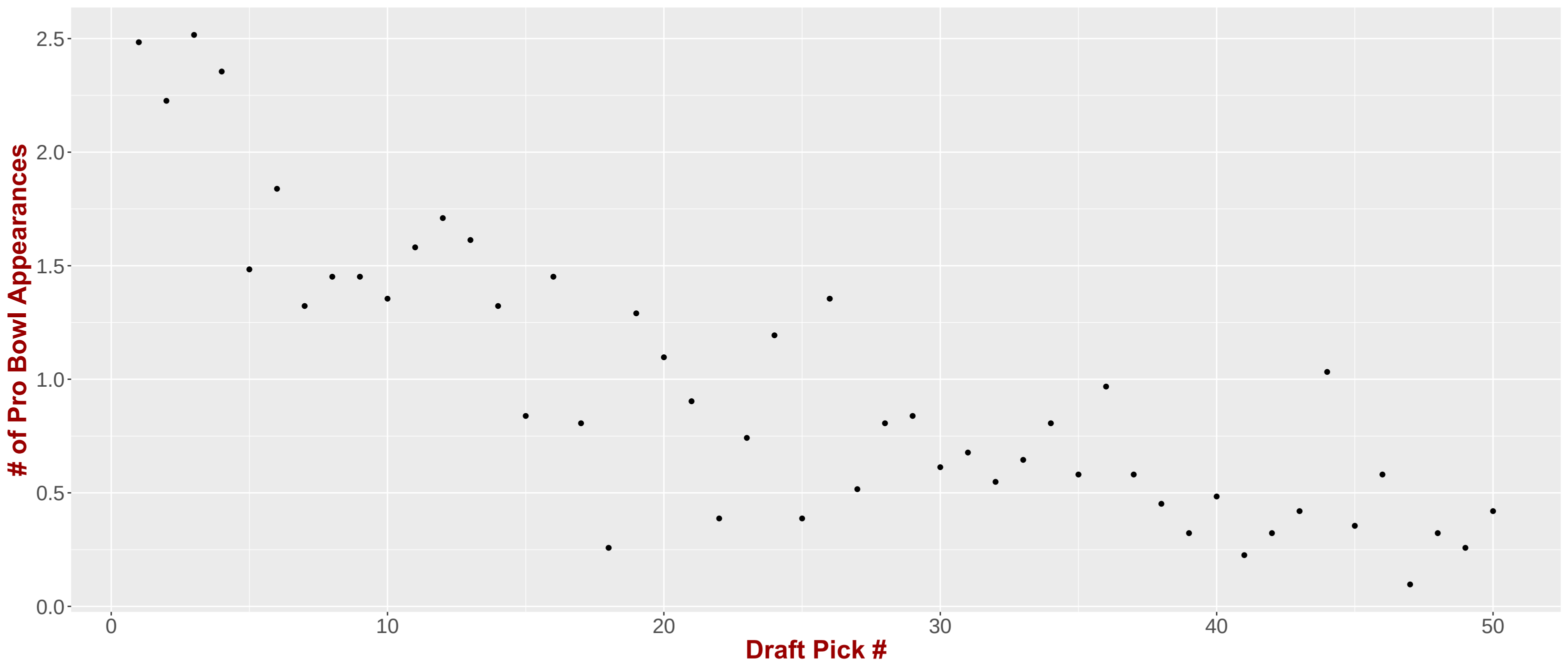


Overfitting

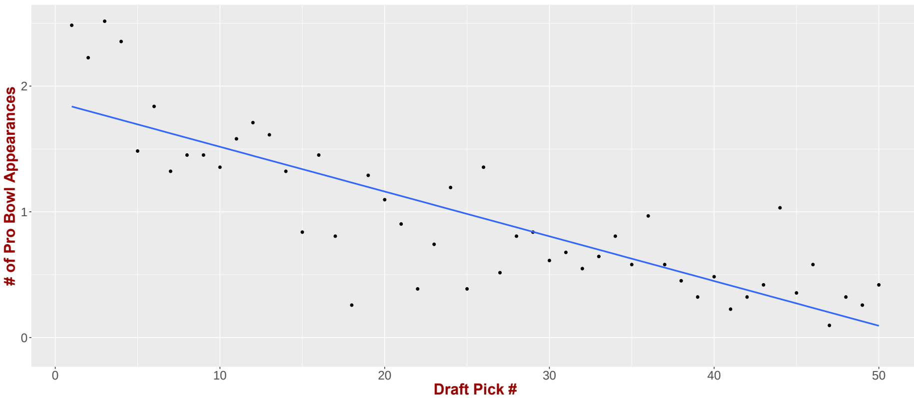
- In a regression model the complexity of the model is captured by the number of parameters
  - If there are n data points in the training set and the number of parameters is also n, then the fitted model line will go through all of the points in the training set
  - Even if we only have one independent variable, we can still have n > 1 parameters for the model through polynomial regression: $y = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n + \varepsilon$

Overfitting

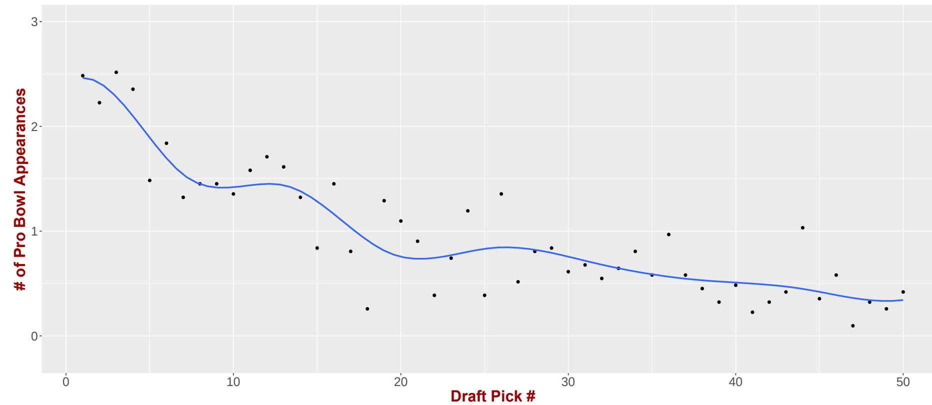
- What is the relationship between number of Pro Bowl appearances for an NFL player and his draft order?



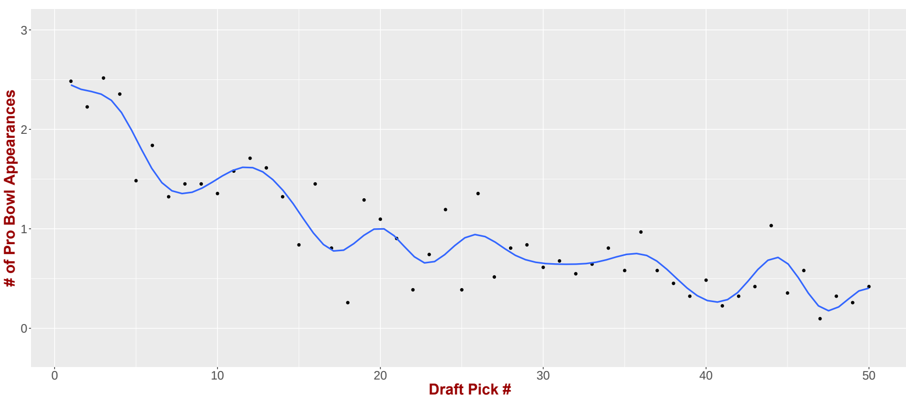
Data: https://www.kaggle.com/ronaldjgrafjr/nfl-draft-outcomes/version/1



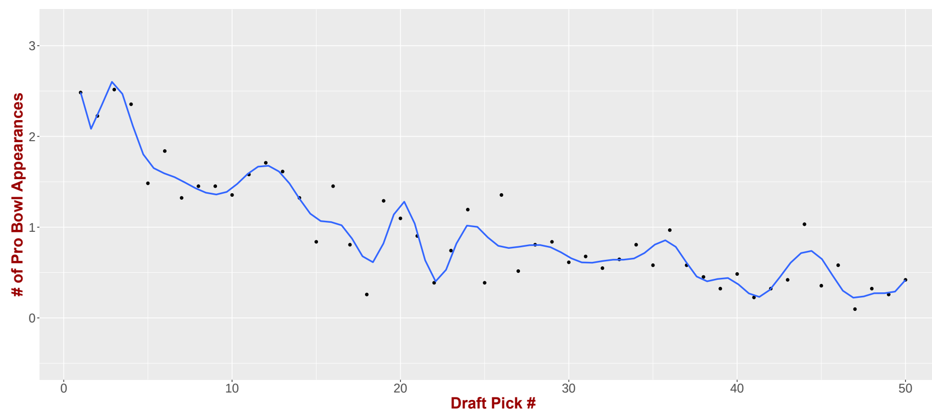
# of Pro Bowl Appearances
Draft Pick #
0
1
2
10
20
30
40
50



# of Pro Bowl Appearances
Draft Pick #
0
1
2
3
10
20
30
40
50



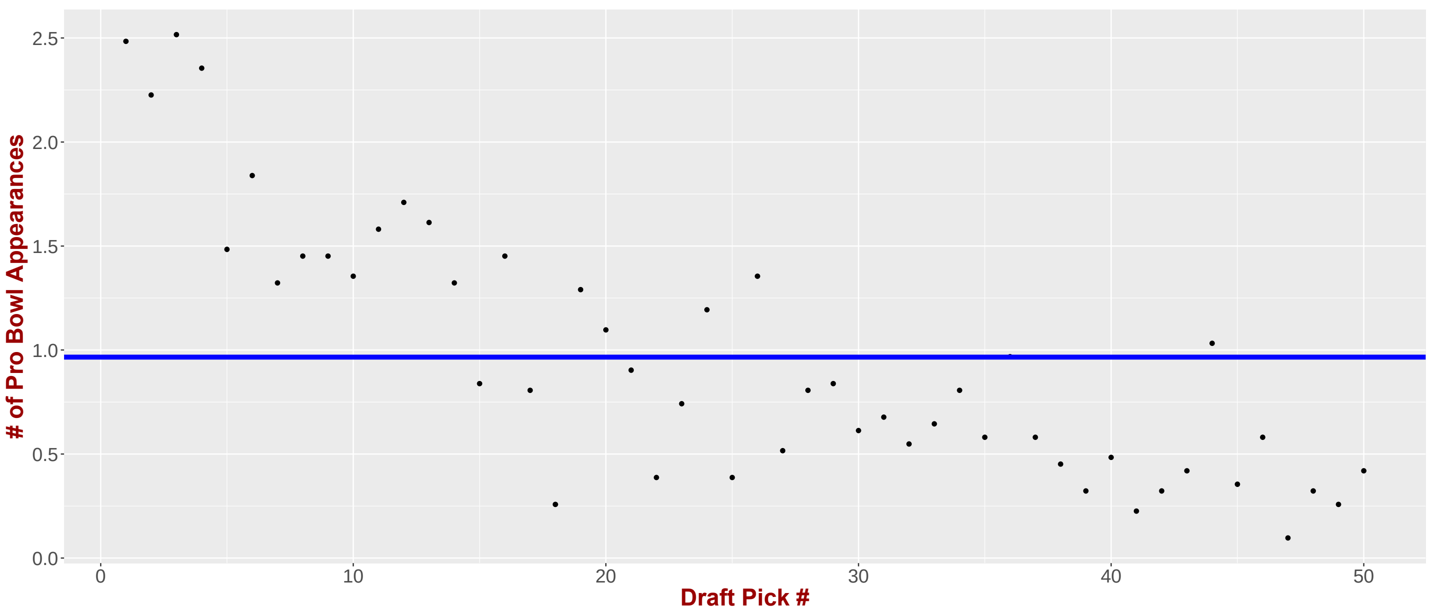
# of Pro Bowl Appearances
Draft Pick #
0
1
2
3
10
20
30
40
50



# of Pro Bowl Appearances
Draft Pick #
0
1
2
3
10
20
30
40
50

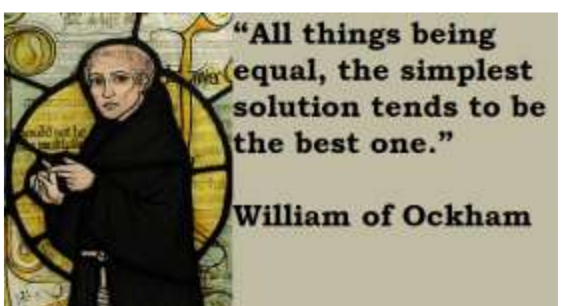
Underfitting

- One might face the opposite problem – underfitting
  - The model is too simplistic to capture any useful information in the data



Occam's razor

- When there are two explanations for an observation, the simpler is *usually* better
- In modeling this means that between two model hypothesis the simpler is preferable
  - The more complex a model is the more probable it is not true, and, thus we have overfitting



Bias-Variance Tradeoff

- Model complexity and the Occam's razor principle can be further explored with the bias-variance tradeoff for a model
- Let's consider a regression model and its evaluation through the mean squared error (MSE): $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
- There are two elements that contribute to this error (apart from the inherent noise)
  – Model bias
  - Model variance

$$MSE = bias^2 + variance + \varepsilon$$

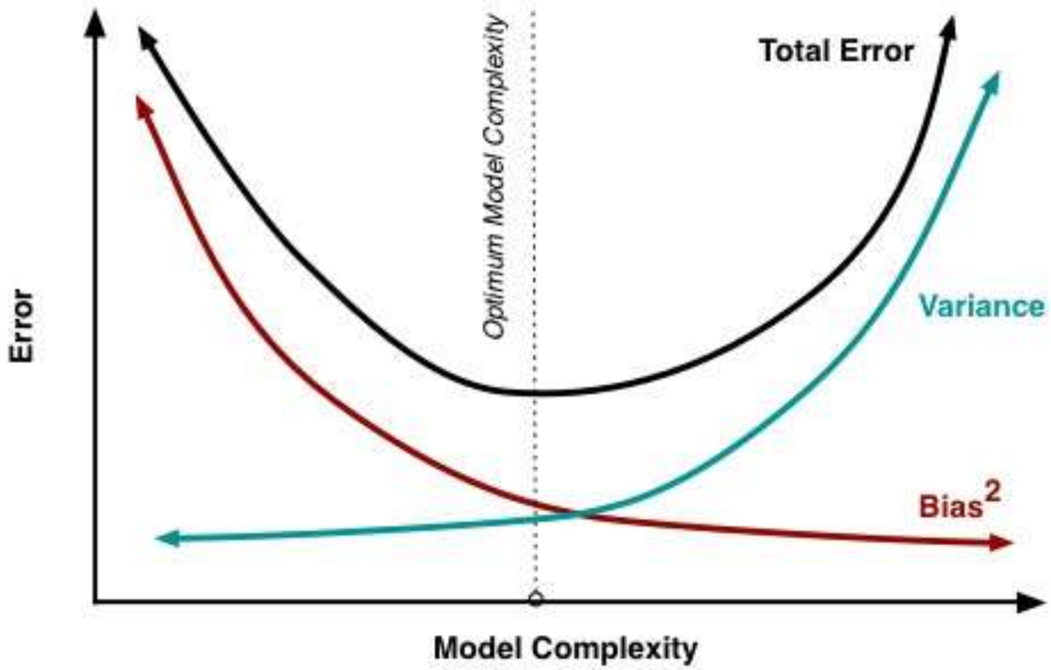
Bias-Variance Tradeoff

- If we want to minimize MSE, we need to minimize both bias and variance
  - However, when bias gets smaller, variance increases and vice versa
- A model that is underfitted has high bias
  - Misses relevant relations between the independent variables and the response variable
  - Bias is reduced by increasing model complexity

Bias-Variance Tradeoff

- If we want to minimize MSE, we need to minimize both bias and variance
  - However, when bias gets smaller, variance increases and vice versa
- A model that is overfitted has high variance
  - The model captures the noise in the training data instead of the trend
  - Variance is reduced by decreasing model complexity

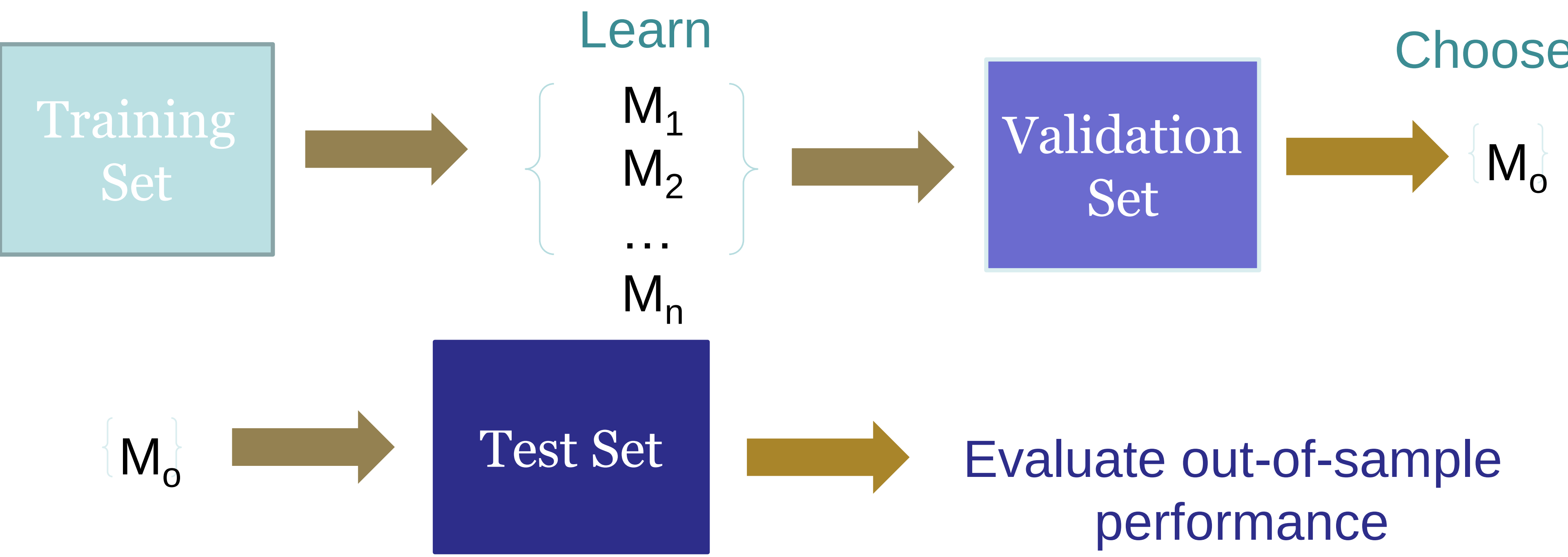
Bias-Variance Tradeoff



Model Selection

- To avoid overfitting and pick the *best* possible model we need three sets:
  - Training set: Identify the weights of different regression models by minimizing the (squared) error on the training set
    - Different regression models can include linear-vs-polynomial regression, different set of features etc.
  - Validation set: Evaluate the performance of the different regression models identified via training & pick the best
  - Test set: Evaluate the performance of the model chosen from the validation set → this is the expected performance for the model

Model Selection



Regularization

- In order to avoid overfitting we can slightly alter the optimization problem we have to solve for training the model
  - Implicitly constraint the values that the model parameters can take
- Key idea: Penalize overly complicated answers
  - Extreme curves/models typically require extreme values → susceptible to high variance

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{N} (y_i - \boldsymbol{\alpha}^T \cdot \boldsymbol{x_i})^2 + \lambda f(\boldsymbol{\alpha})$$

Regularization term

Regularization

- The regularization term can take different forms

$$f(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_2^2 \quad Ridge\ regression$$
$$f(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1 \quad Lasso\ regression$$

Ridge Regression

- The solution obtained depends on the shrinkage parameter λ
  - λ controls the size of the coefficients, i.e., the amount of regularization
  - Reducing λ leads to solutions closer to the least squares (λ=0)
  - Increasing λ will give us an intercept only
- How to choose λ?
  - Use a validation set!

Lasso Regression

- Very similar to ridge regression but with subtle and important differences
  - The optimization problem is not linear anymore
- Ridge regression forced the square of the coefficients to be less than a fixed value
  - This shrinks the size of the coefficients but does not set any of them exactly equal to 0
- Lasso forces the sum of the absolute values of the coefficients to be less than a fixed value
  - This can force some of the coefficients to be equal to 0 → essentially this chooses a simpler model that does not include these features

Model selection example

- Let's consider NBA team ratings
- We have seen that we can build a simple regression rating by minimizing the sum of the squared differences of the actual and predicted score differential
- What are some alternatives?
  - Rergularization (both ridge and lasso)
- How can we choose among the three?
  - Model selection!

Model selection example

Team	No regularization	Ridge	Lasso
Atlanta Hawks	-5.243140245	-2.0562291	-4.3249679
Boston Celtics	3.856650794	1.6242642	3.26436923
Brooklyn Nets	-4.208863962	-1.701494	-3.4989667
Charlotte Hornets	-0.533159455	-0.2522544	0.00032652
Chicago Bulls	-5.682990076	-2.3266832	-4.879608
Cleveland Cavaliers	-0.260973016	-0.0530343	0.00019336
Dallas Mavericks	-2.044925396	-0.9574353	-1.5515693
Denver Nuggets	1.160528605	0.46653658	0.52620444
Detroit Pistons	-1.006147943	-0.4221928	-0.4287926
Golden State Warriors	8.077495223	3.20263193	7.30864545
Houston Rockets	8.962682316	3.62637896	8.26129563
Indiana Pacers	0.824543413	0.49708523	0.36985976
Los Angeles Clippers	1.070307928	0.45459214	0.458719
Los Angeles Lakers	-1.453617886	-0.5765	-0.811317
Memphis Grizzlies	-4.915058078	-2.0645623	-4.1691074
Miami Heat	0.171037061	0.10242633	0.00191281
Milwaukee Bucks	-0.606079482	-0.2400093	-0.0001204
Minnesota Timberwolves	2.732240184	1.07748441	2.06317797
New Orleans Pelicans	0.461259049	0.18946809	0.21150404
New York Knicks	-3.378411257	-1.3762296	-2.6505121
Oklahoma City Thunder	2.824586502	1.19360069	2.2265779
Orlando Magic	-4.267844453	-1.6897891	-3.5112646
Philadelphia 76ers	2.769041804	1.01520704	2.0225851
Phoenix Suns	-8.555117647	-3.5410105	-7.7856664
Portland Trail Blazers	2.152887726	0.92334278	1.56961878
Sacramento Kings	-7.593330583	-3.1265078	-6.8146881
San Antonio Spurs	2.300144236	0.97229131	1.70809857
Toronto Raptors	8.238598805	3.41081269	7.62263935
Utah Jazz	2.925182942	1.08682161	2.18062719
Washington Wizards	1.22247289	0.54098773	0.6302253

$\lambda=100$

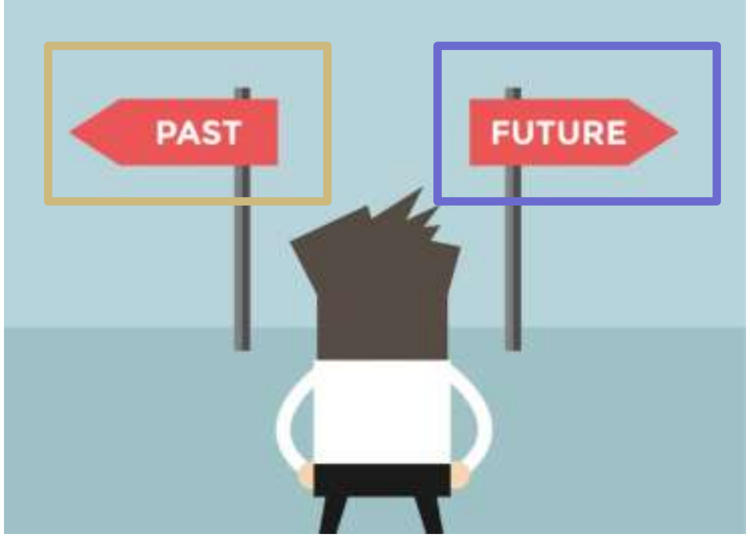
- Notice the shrinkage of the coefficients in the regularized regressions
- For lasso, a few coefficients have been shrunk almost all the way to 0 (e.g., Cleveland and Milwuakee)

Model selection example

	No regularization	Ridge	Lasso
Train MSE	146.2844626	159.4338315	147.1360025
Validation MSE	**132.9994492**	143.751361	134.5141575
Test average MSE	241.265375	-	-

Descriptive & predictive models

- Many times the two are confused and assumed to be the same
- Descriptive models tell us *what has happened*
- Predictive models tell us *what might happen*



Descriptive models

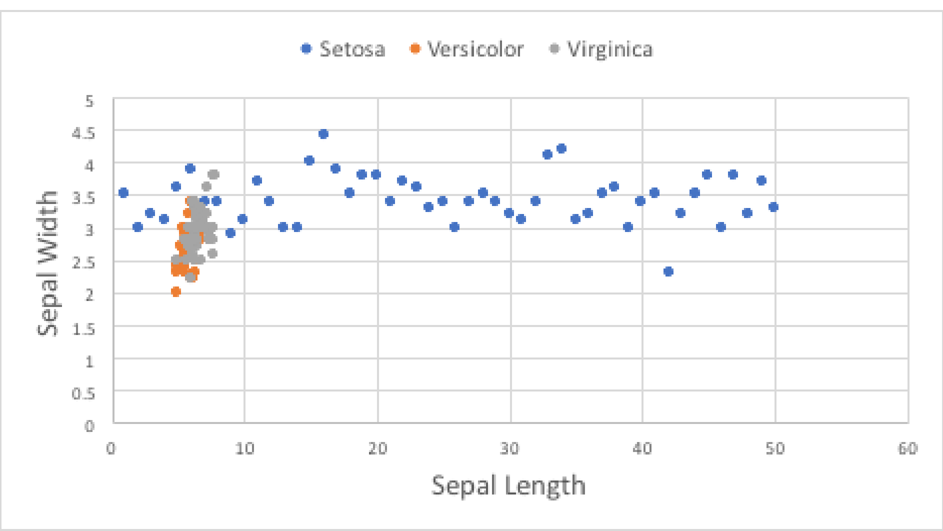
- Descriptive models and analytics in general help us understand what has happened in the past
- They present the main *features* of the data
  - A summary of the data
  - Clustering is most probably the best example
- Data that are generated from a *good* descriptive model will have the same characteristics as the real data

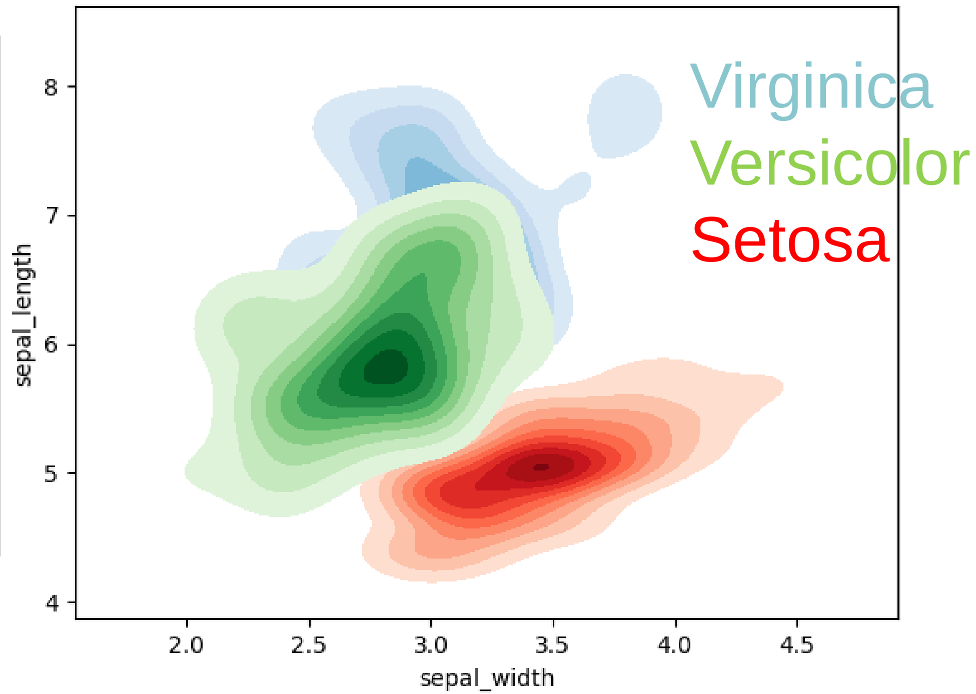
Descriptive models

- Descriptive models can be as simple as a kernel density estimation
  - Mutlivariate or univariate
  - Parametric or non-parametric
- For example, the Iris dataset includes information from 50 samples of the Iris flower
  - Length and width of sepals and petals

Descriptive models

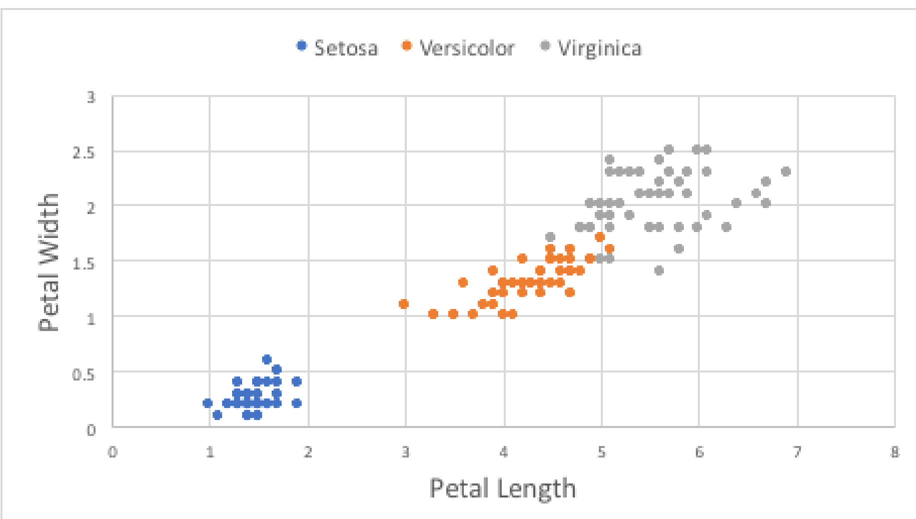
- What is the distribution of the flower's sepal width and length for the different species in the dataset?

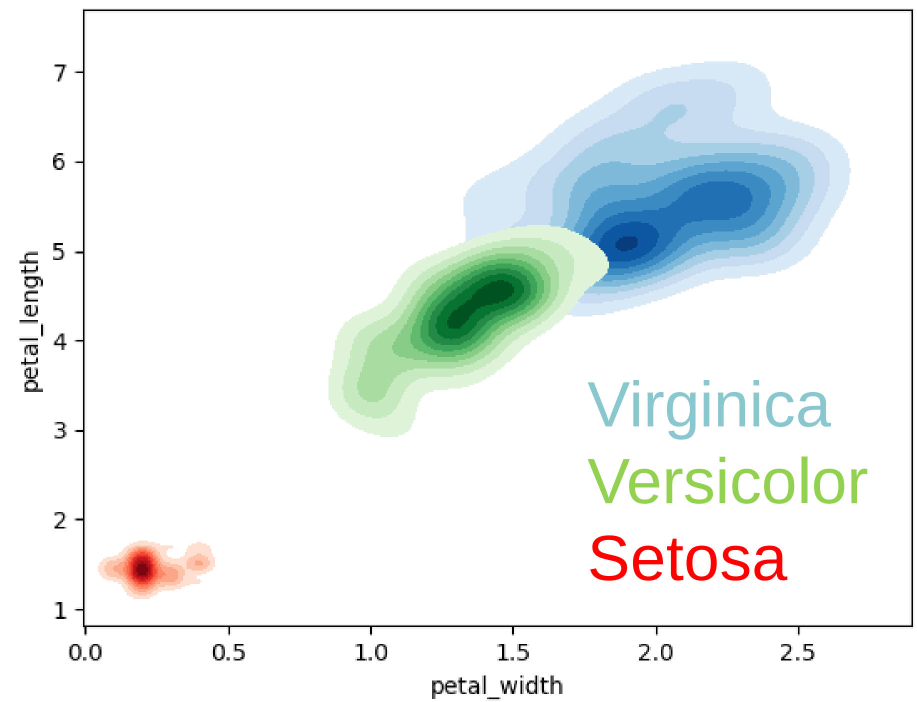




Descriptive models

- What is the distribution of the flower's petal width and length for the different species in the dataset?





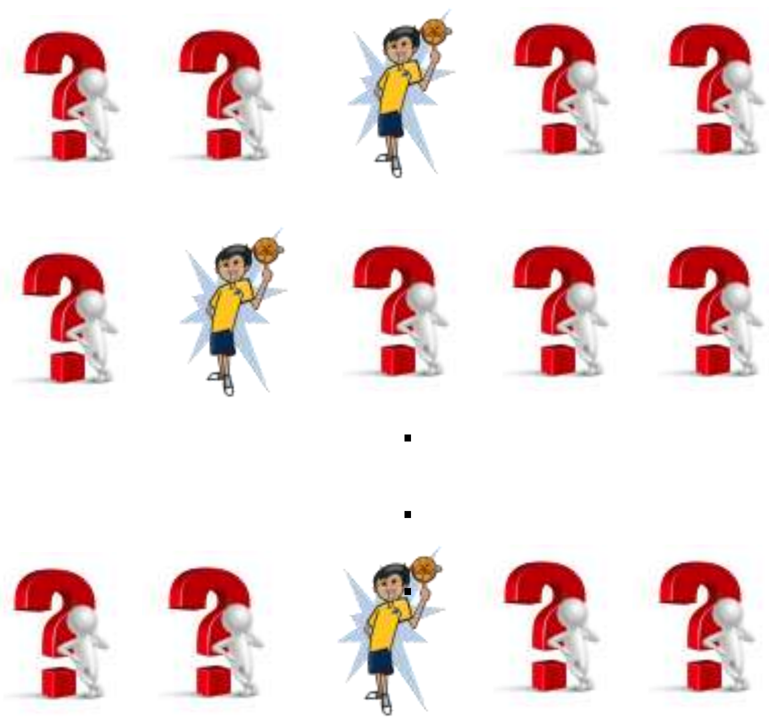
Descriptive models

- A field where the distinction is clear is sports
- Descriptive models describe how a player performed over the season
  - E.g., used for end-of-season awards (MVP etc.)
- Predictive models aim at projecting future player performance
  - E.g., for player trades and acquisition

Descriptive models

- How can we quantify the contributions of a basketball player to his team during the past season?
- Typical way to do so is with the +/- metric
  - Captures the point margin for the team when the player is on the field
  - This point margin can then be translated to wins-contributed

Plus-Minus (+/-)



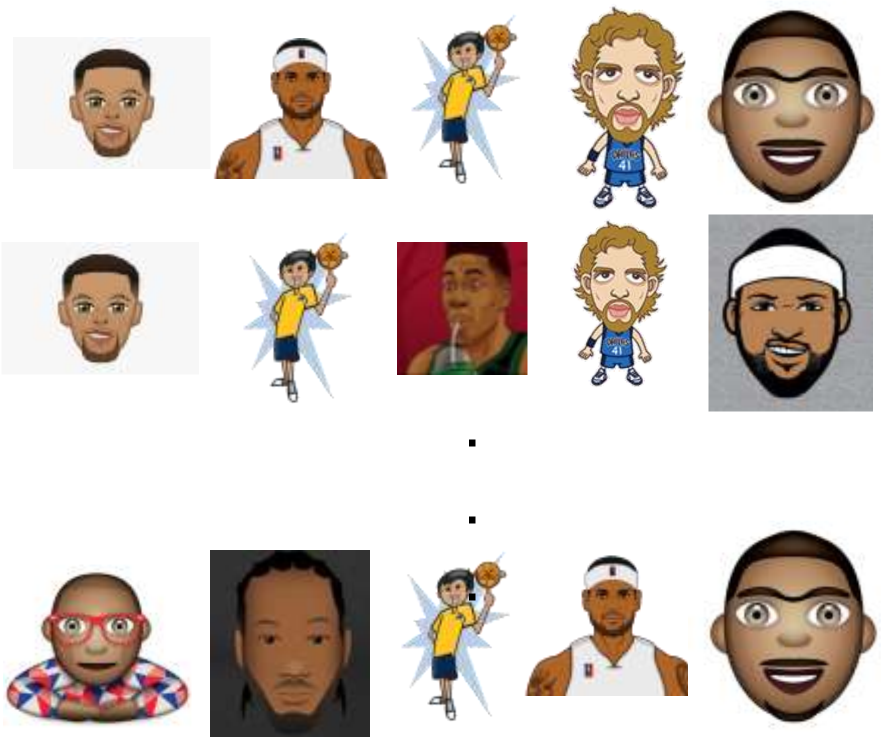
Points scored: s_1
Points allowed: a_1

Points scored: s_2
Points allowed: a_2

Points scored: s_n
Points allowed: a_n

$$\sum_{i=1}^{n} (s_i - a_i)$$

Plus-Minus (+/-)



Adjusted +/-

- Controls for the presence of other players on the court
  - Both offense and defense
- Each *stint* is a data point
  - DV: PM/possession
  - IVs: Dummy variables for all players
    - 1 for home team players in the stint, -1 for visiting team players in the stint and 0 for the rest

Adjusted +/-

- Pass all the stints through a linear regression

- The coefficient for each player α_i is the adjusted plus-minus of the player

$$y = a_1 x_1 + a_1 x_1 + \ldots + a_r x_r + \varepsilon$$

Adjusted +/-

Team 1/Us (P1-P5): Players 1 through 9

Team 2/Them (P6-P10): Players 10 through 18

Stints are full games (i.e., 48 minutes)

Assume no home edge (neutral court)

Game	Result	P 1	P2	P3	P4	P5	P6	P7	P8	P9	P10
1	-13	4	1	7	5	2	15	16	10	17	14
2	19	1	6	2	5	4	11	17	14	15	18
3	-4	1	9	2	8	4	15	14	10	17	13
4	29	1	6	5	3	2	16	17	18	14	11
5	-3	9	7	1	5	6	17	15	12	18	10
6	12	7	2	5	1	4	17	11	15	16	18
7	-5	6	5	8	9	1	13	16	12	15	10
8	-32	4	2	9	5	3	17	12	10	18	15
9	18	8	3	9	1	7	17	16	15	14	11
10	17	1	2	9	6	4	13	16	10	11	18
11	-11	7	3	2	5	6	14	17	15	12	15
12	-14	7	8	4	6	3	18	11	12	17	15
13	29	4	5	9	2	6	11	13	14	17	18
14	17	1	8	4	2	7	13	12	14	17	18
15	0	6	9	8	7	10	15	12	10	17	14
16	-7	6	3	2	1	8	17	18	16	14	10
17	9	3	2	5	6	7	13	16	14	10	11
18	24	1	7	6	7	4	18	13	18	15	11
19	18	1	2	5	8	6	14	13	12	15	18
20	-24	2	4	3	8	5	11	18	16	17	10

$$\min_{a} \sum_{i=1}^{20} \left(\sum_{j=1}^{5} a_{Pj(i)} - \sum_{j=6}^{10} a_{Pj(i)}\right)^2$$

Player	Adjusted +/
1	12.78637207
2	2.919687658
3	-6.185552324
4	-10.09502237
5	-0.121270455
6	0.878532834
7	1.917570362
8	-6.064612857
9	5.972176048
10	16.90654413
11	-13.07998337
12	0.88523712
13	-8.991225193
14	-6.212323616
15	7.866337403
16	1.809264884
17	0.008932283
18	-1.200664607

Adjusted +/-

- P1 has an adjusted +/- of +12.8 points
  - Whenever P1 is on the court his team is expected to outscore the opponent +12.8 points/game
- Adjusted +/- is not stable from season to season
  - Cannot be used to predict a players future +/-
- It is a descriptive metric!
  - Assignment of credit

Predictive models

- Predictive models and analytics in general aim at forecasting the future
  - These forecasts are probabilistic
- **Predictive models do not identify causes**!
- They are similar to descriptive models in the sense that they are looking for patterns in past data, **but** these patterns need to be persistent to provide predictive power

Predictive models

- For predictive models is absolutely crucial to examine their quality out-of-sample
- We need to make sure that the patterns identified from the training set are *generalizable*
  - Models need to be evaluated regularly to ensure they are still predictive

Predictive models: example

- While the adjusted +/- that we saw earlier is a descriptive model, teams are certainly interested in a predictive *version* of it
  - Regularization can help
- Ridge regression is usually used to improve the out-of-sample predictive power of the model

$$\min_{a} \sum^{20} \left(\sum^{5} a_{Pj(i)} - \sum^{10} a_{Pj(i)} \right)^2$$

Predictive models: example

- Now in our case with this toy-example we cannot really make a meaningful evaluation of the predictive power since we have very few data (and artificially generated) data
- However, it is worth noting the shrinkage of the coefficients as compared to the non-regularized version

Player	RAdjusted+/
1	7.76836091
2	0.136819927
3	-5.415001027
4	-4.370679634
5	-0.191613197
6	3.282882849
7	1.320318543
8	-3.645487267
9	2.987361165
10	12.0503124
11	-7.565657477
12	3.412795014
13	-6.42204394
14	-6.959639575
15	4.847689719
16	-0.266711362
17	1.767711392
18	-2.737418437