Get a Bit More Familiar with Supervised Learning





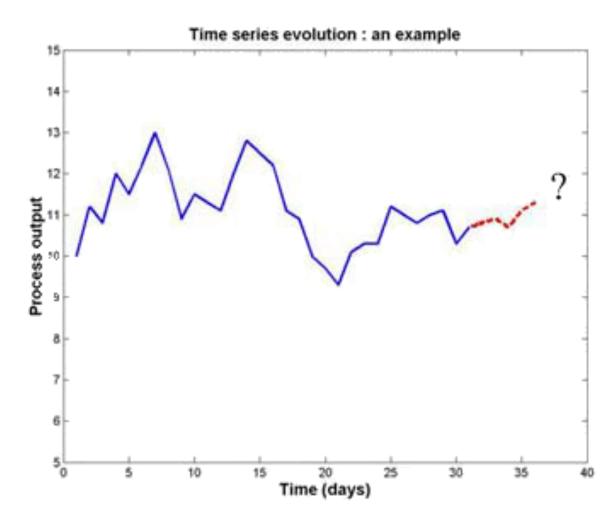
Time Series





Time Series Prediction

Training on previous data, predicting numeric values: Regression.



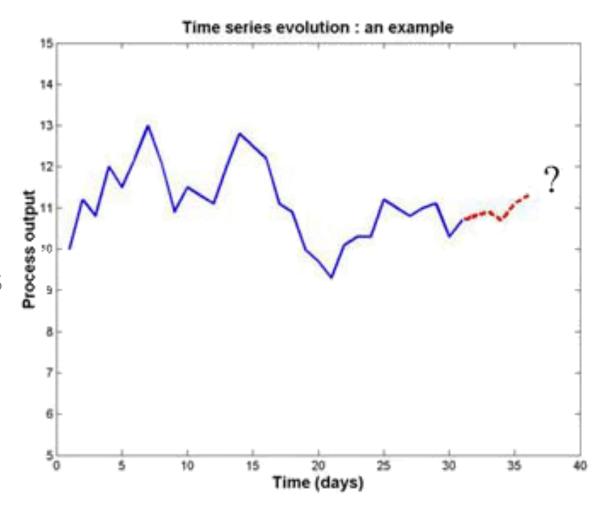
Time Series Prediction

WARNING!

Time Series are autocorrelated.

Your regression is useless.

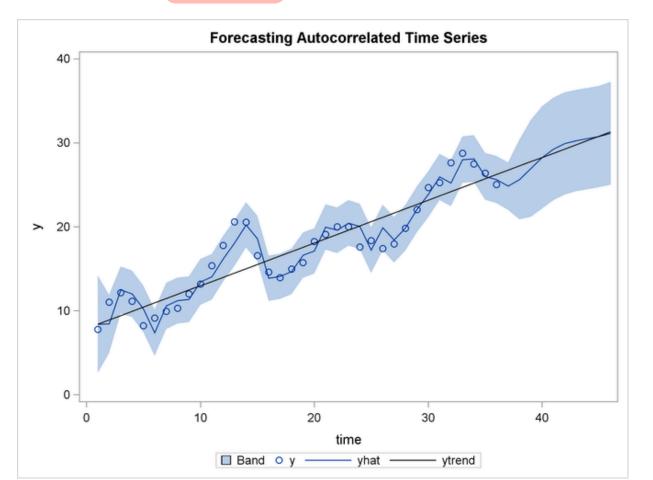
Unless.....



Autoregressive Models



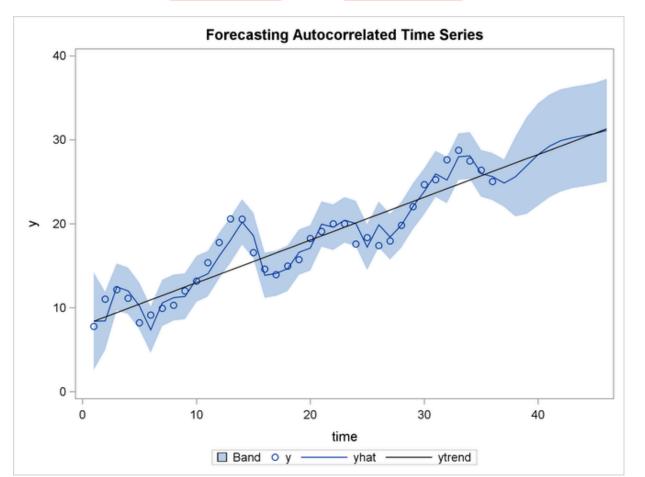
$$y_{\beta}(x_{t}) = \beta_{0} + \beta_{t}x_{t-1} + \beta_{1}x_{1} + \beta_{2}x_{2} + \beta_{3}x_{3} + \varepsilon$$



Autoregressive Models



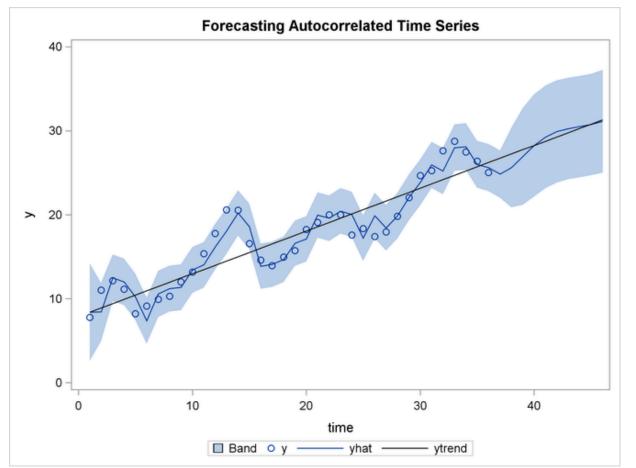
$$y_{\beta}(x_{t}) = \beta_{0} + \beta_{t1}x_{t-1} + \beta_{t2}x_{t-2} + \beta_{1}x_{1} + \dots$$



Autoregressive Models



$$y_{\beta}(x_{t}) = \beta_{0} + \beta_{t1}x_{t-1} + \beta_{t2}x_{t-2} + \dots + \frac{\beta_{t}}{\beta_{t}} + \beta_{t} + \dots + \frac{\beta_{t}}{\beta_{t}} + \dots$$



Decision Trees





Decision Trees

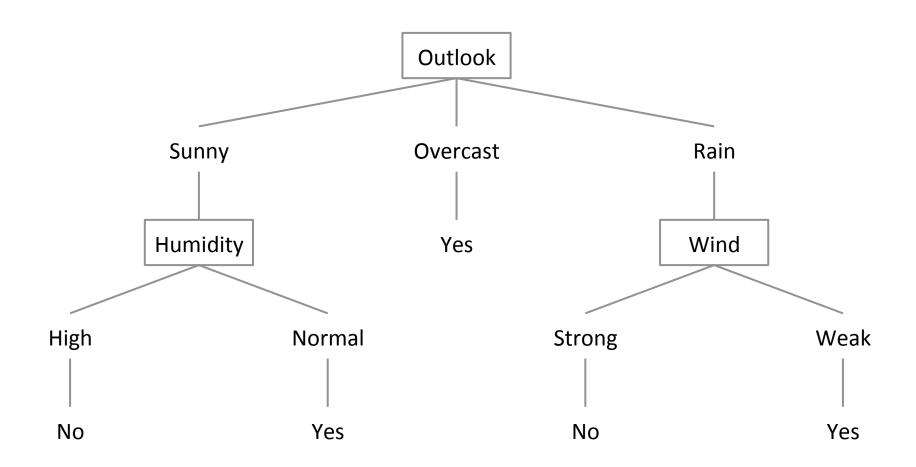
Prediction:

Based on a bunch of IF ... THEN ... ELSE ... rules

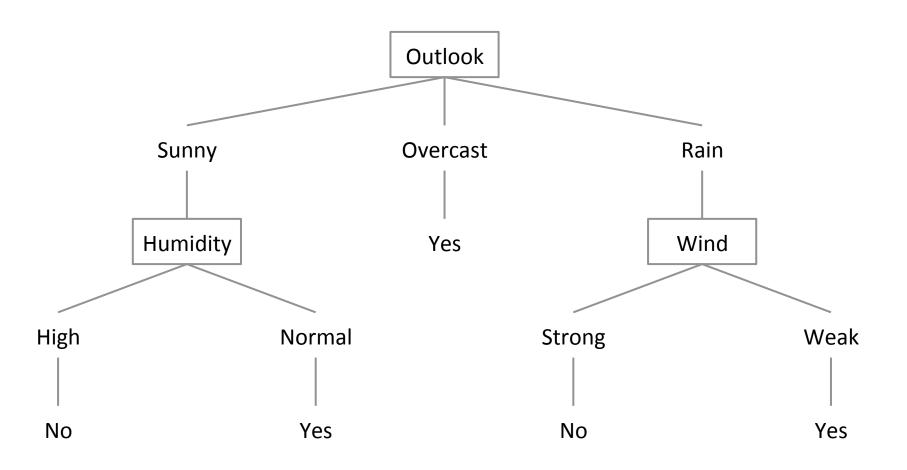
Fitting:

Find a bunch of IF ... THEN ... ELSE ... rules to cover all cases as best you can.

Day	Outlook	Temp	Hum.	Wind	Play Tennis
DI	Sunny	Hot	High	Weak	No
02	Sunny	Hot	High	Strong	No
03	Overcast	Hot	High	Weak	Yes
04	Rain	Mild	High	Weak	Yes
DS	Rain	Cool	Nml	Weak	Yes
06	Rain	Cool	Nml	Strong	No
07	Overcast	Cool	Nml	Strong	Yes
DB	Sunny	Mild	High	Weak	No
09	Sunny	Cool	Nml	Weak	Yes
DIO	Rain	Mild	Nml	Weak	Yes
DII	Sunny	Mild	Nml	Strong	Yes
012	Overcast	Mild	High	Strong	Yes
013	Overcast	Hot	Nml	Weak	Yes
014	Rain	Mild	High	Strong	No

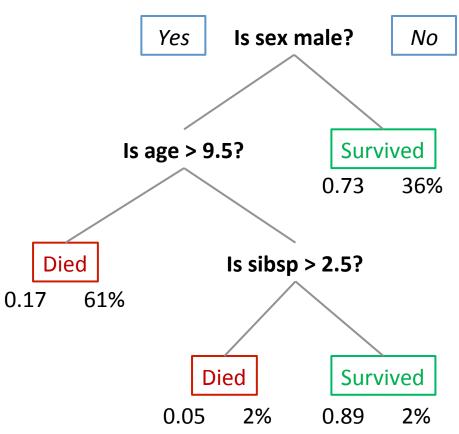


Each inner node is a decision based on a feature Each leaf node is a class label



Each inner node is a decision based on a feature Each leaf node is a **class label**

Predicting Titanic Survivors



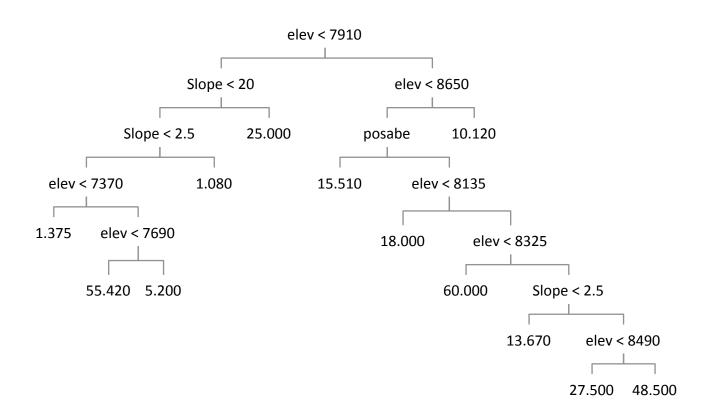
It can also be used for regression

CART

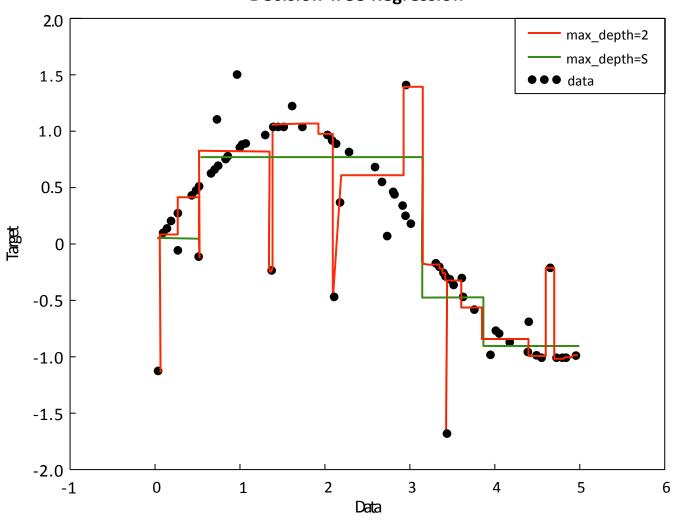
(Classification and Regression Tree)

Each inner node is a decision based on a feature Each leaf node is a <u>predicted value</u>

Predicting precipitation in Himalayas from elevation, slope and position



Decision Tree Regression



from sklearn.tree import DecisionTreeClassifier

from sklearn.tree import DecisionTreeRegressor

Advantages:

The decisions are easy to understand and interpret.

Both numerical and categorical features can be used naturally.

Natural multiclass classifier.

Advantages:

The decisions are easy to understand and interpret.

Both numerical and categorical features can be used naturally.

Natural multiclass classifier.

Disadvantages:

Can overfit to training data with complex trees.

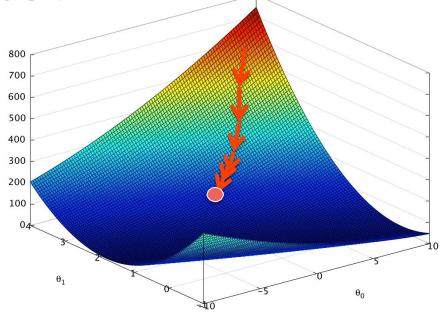
Small changes in input data can result in totally different trees.

Can make mistakes with unbalanced classes.

No confidence intervals (regression).

So, how do we build the best tree?

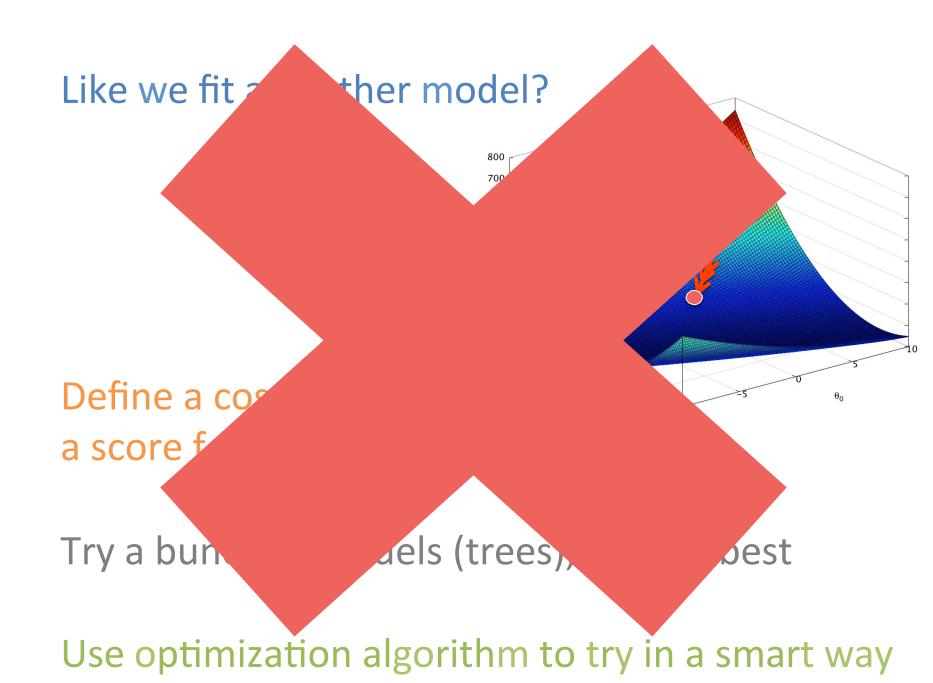
Like we fit any other model?

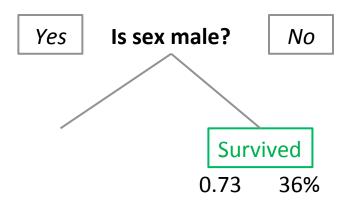


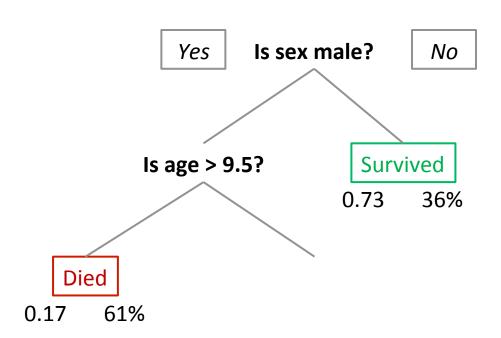
Define a cost function: a score for a specific model (tree)

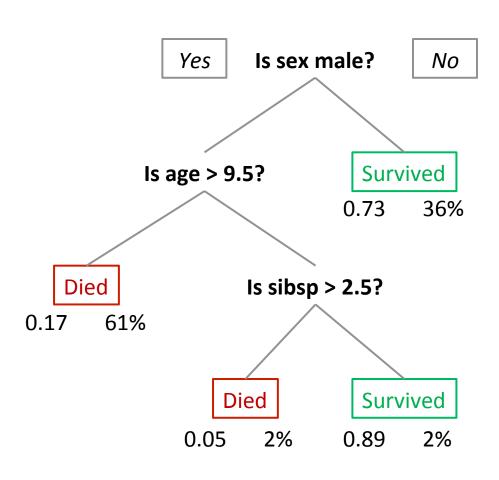
Try a bunch of models (trees), choose best

Use optimization algorithm to try in a smart way





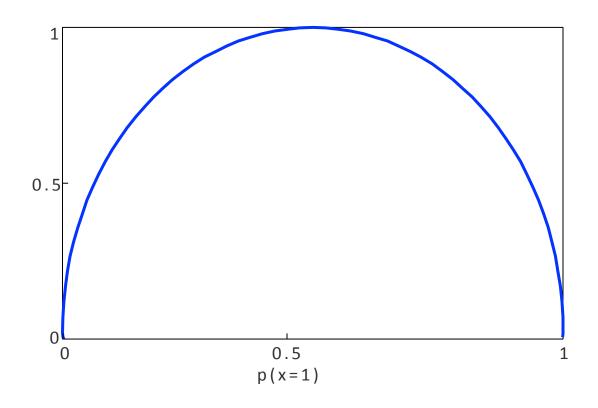




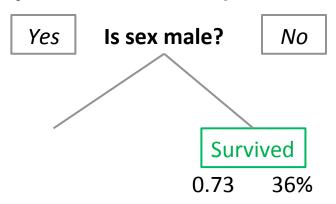
Use a heuristic: Greedy search

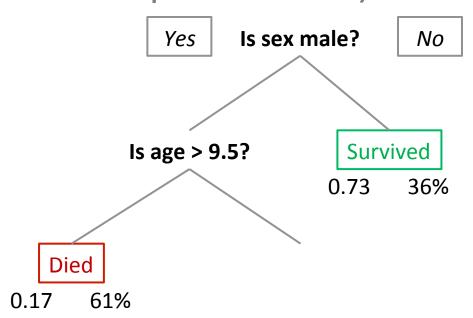


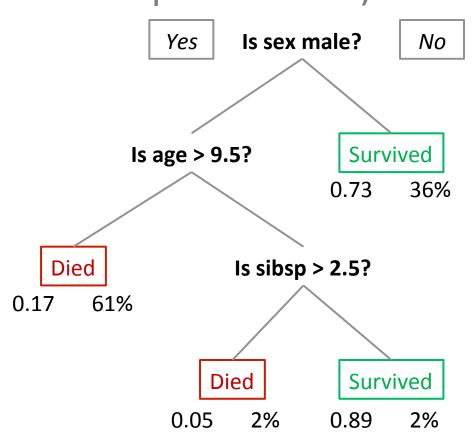
Information Entropy



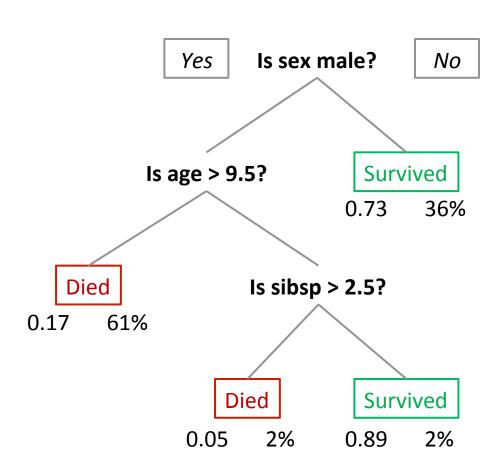
$$H = \sum_{i} p(x_i) \log p(x_i)$$



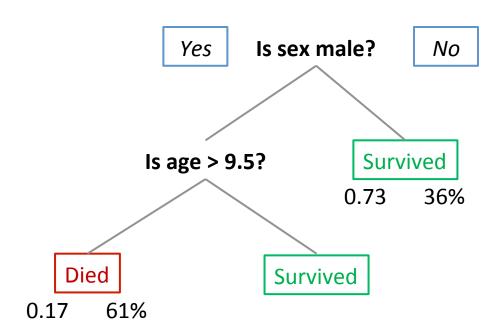




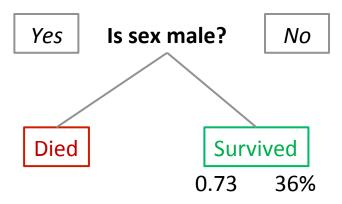
May overfit. Prune.



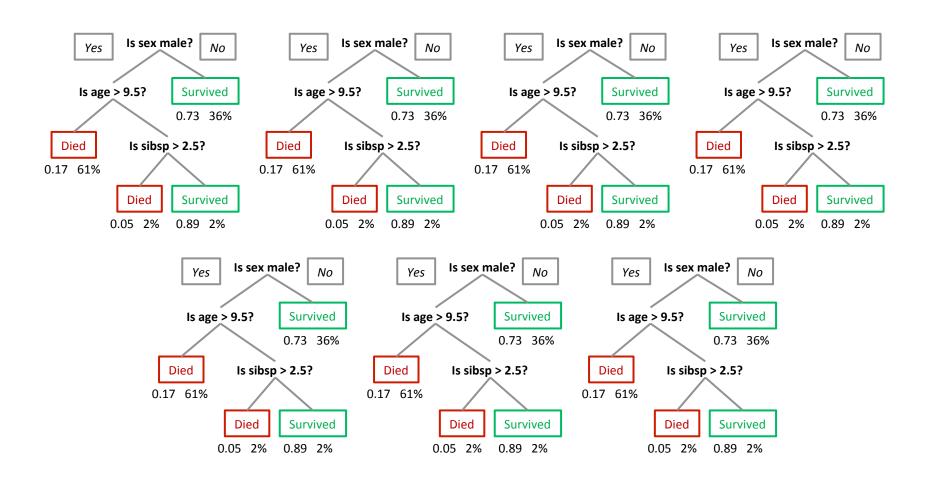
May overfit. Prune.



May overfit. Prune.



Better yet: Use ensemble methods



Ensemble Methods: Bagging





Bootstrap aggregating

Problem:

Overfitting to training set

Solution:

Bootstrap training set into multiple sets

Fit a model to each random set

Each model has one vote, choose max vote

	Date	Title	Budget	DomesticTotalGross	Director	Rating	Runtime
0	2013-11-22	The Hunger Games: Catching Fire	130000000	424668047	Francis Lawrence	PG-13	146
1	2013-05-03	Iron Man 3	200000000	409013994	Shane Black	PG-13	128
2	2013-11-22	Frozen	150000000	400738009	Chris BuckJennifer Lee	PG	108
3	2013-07-03	Despicable Me 2	760000000	368061265	Pierre CoffinChris Renaud	PG	98
4	2013-06-14	Man of Steel	225000000	291045518	Zack Snyder	PG-13	143
5	2013-10-04	Gravity	100000000	274092705	Alfonso Cuaron	PG-13	91
6	2013-06-21	Monsters University	NaN	268492764	Dan Scanion	G	107
7	2013-12-13	The Hobbit: The Desolation of Smaug	NaN	258366855	Peter Jackson	PG-13	161
8	2013-05-24	Fast & Furious 6	160000000	238679850	Justin Lin	PG-13	130
9	2013-03-08	OZ The Great and Powerful	215000000	234911825	Sam Raimi	PG	127
10	2013-05-16	Star Trek Into Darkness	190000000	228778661	J.J. Abrams	PG-13	123
11	2013-11-08	Thor: The Drak World	170000000	206362140	Alan Taylor	PG-13	120
12	2013-06-21	World War Z	190000000	202359711	Marc Forster	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MiccoChris Sanders	PG	98
14	2013-06-28	The Heat	43000000	159582188	Paul Feig	R	117
15	2013-08-07	We're the Millers	37000000	150394119	Rawson Marshall Thurber	R	110
16	2013-12-13	American Hustle	40000000	150117807	David O. Russell	R	138
17	2013-05-10	The Great Gatsby	105000000	144840419	Baz Luhrmann	PG-13	143

Training set

Test set

Bootstrap: Sample with replacement

	Date	Title	Budget	DomesticTotalGross	Director	Rating	Runtime
)	2013-11-22	The Hunger Games: Catching Fire	130000000	424668047	Francis Lawrence	PG-13	146
1	2013-05-03	Iron Man 3	200000000	409013994	Shane Black	PG-13	128
2	2013-11-22	Frozen	150000000	400738009	Chris BuckJennifer Lee	PG	108
3	2013-07-03	Despicable Me 2	760000000	368061265	Pierre CoffinChris Renaud	PG	98
1	2013-06-14	Man of Steel	225000000	291045518	Zack Snyder	PG-13	143
5	2013-10-04	Gravity	100000000	274092705	Alfonso Cuaron	PG-13	91
6	2013-06-21	Monsters University	NaN	268492764	Dan Scanion	G	107
7	2013-12-13	The Hobbit: The Desolation of Smaug	NaN	258366855	Peter Jackson	PG-13	161
В	2013-05-24	Fast & Furious 6	160000000	238679850	Justin Lin	PG-13	130
9	2013-03-08	OZ The Great and Powerful	215000000	234911825	Sam Raimi	PG	127
10	2013-05-16	Star Trek Into Darkness	190000000	228778661	J.J. Abrams	PG-13	123
11	2013-11-08	Thor: The Drak World	170000000	206362140	Alan Taylor	PG-13	120
12	2013-06-21	World War Z	190000000	202359711	Marc Forster	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MiccoChris Sanders	PG	98
14	2013-06-28	The Heat	43000000	159582188	Paul Feig	R	117
15	2013-08-07	We're the Millers	37000000	150394119	Rawson Marshall Thurber	R	110
16	2013-12-13	American Hustle	40000000	150117807	David O. Russell	R	138
17	2013-05-10	The Great Gatsby	105000000	144840419	Baz Luhrmann	PG-13	143

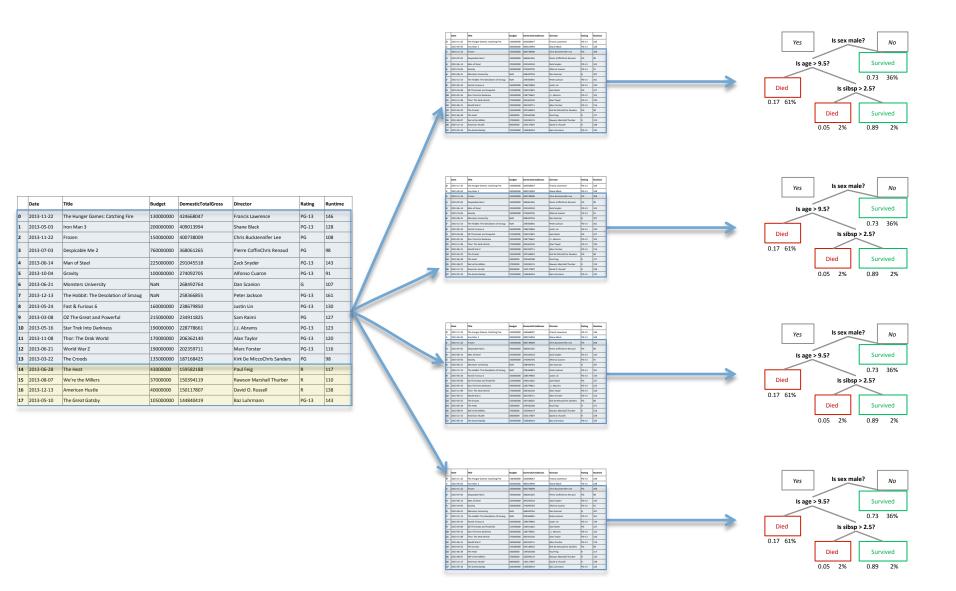
	Duse	Title	Budget	DomesticTutalStress	Director	Rating	Runding
0	2019-11-22	The Hunger Games Catching Fire	2 800000000	620668067	Francis Lawrence	PG-13	106
	2013-05-08	Son Man 3	200000000	039313994	Stare Black	PG-13	128
2	2013-11-22	frizies	150000000	030738009	Chris Bucklennifer Lee	PG .	108
	2013-07-08	Despisable Me 2	260000000	16806265	Piece CuffinChris Resaud	PG	100
	2013-06-14	Man of Steel	225000000	290045518	Zack Snyder	PG-13	163
	2013-10-06	Gravity	200000000	270092305	Alfonsa Cuaran	PG-13	91
	2013-06-21	Manders University	NaN	268192764	Dan Scanion	G G	107
,	2013-12-13	The Hobbit: The Desolution of Smaug	NaN	218166855	Peter Jackson	PG-13	565
	2013-05-21	Farz & Furious 6	560000000	238679850	pagin sin	PG-13	130
,	2013-03-08	CZ The Great and Powerful	215000000	230911825	Sam Raimi	PG	127
10	2013-05-36	Star Trektisto Darliness	190000000	228778661	LL Abrams	PG-13	123
13	2013-11-08	Than The Disk World	170000000	206162100	Alan Taylor	PG-13	130
12	2013-06-21	World War 2	190000000	202399711	Marc Forder	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MiczoChris Sanders	PG	100
14	2013-06-28	The Hest	£3000000	158582188	Paul Feig	R.	117
15	2053-08-07	We're the Millers	17000000	150396119	Rawson Mandall Thurber	k	130
16	2013-12-13	American Huczle	03000000	150117907	David O. Russell	k.	238
17	2013-05-22	The Great Gottley	205000000	100800019	Bac Lubranann	PG-13	100

	Date	Title	Budget	DomesticTstaltiress	Director	Rating	Runtime
0	2013-11-22	The Hunger Games Catching Fire	180000000	620668067	Francis Lawrence	PG-13	166
	2013-05-08	Son Man 3	200000000	039313994	Stone Black	PG-13	128
r	2013-11-22	frizes	150000000	002738009	Chris Bucksennfer Lee	PG	108
	2019-07-08	Despisable Me 2	260000000	168062365	Pierre CuffinChris Resaud	PG.	100
	2013-06-10	Man of Steel	225/000000	290085518	Zack Snyder	PG-13	16)
	2013-10-06	Gravity	100000000	270092705	Alfonsa Cuaran	PG-13	91
	2013-06-21	Manders University	NaN	268492764	Dan Scanion	6	107
,	2013-12-13	The Hobbit: The Desolution of Smaug	NaN	218366855	Peter Jackson	PG-13	262
	2013-05-26	Fact & Furlous 6	140000000	238679850	suctin Ein	PG-13	130
,	2013-03-08	CZ The Grout and Powerful	215000000	230911825	Sam Kaimi	PG	127
10	2013-05-26	Star Trek Into Darliness	190000000	228778661	LL Abrams	PG-13	123
13	2013-11-08	than the brok World	170000000	206362160	Alan Taylor	PG-13	130
12	2013-06-21	World War 2	190000000	202399711	MarcForder	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MicroChris Sanders	PG	98
14	2013-06-28	The Heat	£30000000	159582188	Paul Feig	t.	117
15	2053-08-07	We're the Millers	17000000	152396119	Kawson Mandali Thurber	t.	110
16	2013-12-13	Assertion Huttle	03000000	193117907	David O. Russell		138
17	2013-05-22	The Great Gottley	505000000	245840419	Mac Lubranann	PG-13	100

	Duse	Title	Budget	DomesticTstaltiress	Director	Rating	Burdine
0	2019-11-22	The Hunger Games Catching Fire	2 800000000	620668067	Francis Lawrence	PG-13	166
1	2053-05-08	ton Man 3	200000000	009018994	Store Black	PG-13	128
2	2013-11-22	frizes	150000000	002728009	Chris Bucksennfer Lee	PG	108
	2019-07-08	Despisable Me 2	260000000	168062365	Pierre CuffinChris Resaud	PG	100
	2013-06-10	Man of Steel	225/000000	29008518	Zack Snyder	PG-13	16)
	2013-10-06	Gravity	200000000	270092305	Alfonsa Cuaran	PG-13	91
	2013-06-21	Mandays University	Nav	268192764	Dan Scanion	6	107
,	2013-12-13	The Hobbit The Desolution of Smarg	NaN	218366855	Peter tackson	PG-13	565
	2013-05-26	Fact & Furlous 6	140000000	238679950	audin bir	PG-13	130
,	2013-03-08	CZ The Grout and Powerful	215000000	230911825	Sam Kaimi	PG	127
10	2013-05-26	Star Trek Into Darliness	190000000	228778661	LL Abrams	PG-13	123
13	2013-11-08	than the box world	170000000	206362160	Alan Taylor	PG-13	130
12	2013-06-21	World War 2	190000000	202399711	MarcForder	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MicroChris Sanders	PG	98
14	2013-06-28	The Heat	£30000000	159582188	Paul Feig		117
15	2013-08-07	We're the Millers	17000000	192896119	Rawson Manhall Thurber	R.	130
16	2013-12-13	Assertion Huttle	03000000	190117907	David O. Russell		238
17	2013-05-22	The Great Gataby	505000000	100800019	Bac Lubranann	PG-13	100

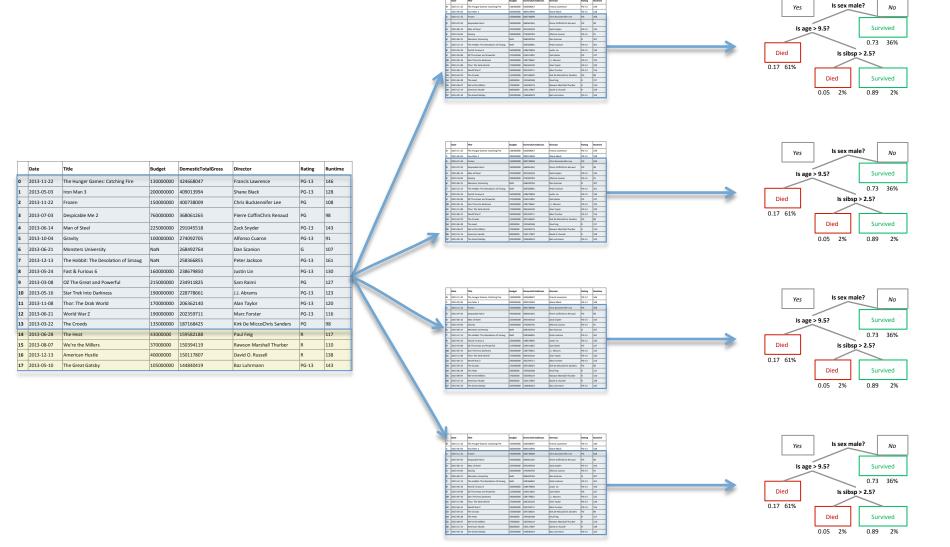
ì											
Г	Date	Title	Budget	DomesticTutalStress	Director	Rating	Runtime				
	2019-11-22	The Hunger Games Catching Fire	1100000000	620668067	Francis Lawrence	PG-13	106				
1	2013-05-08	Son Man 3	200000000	039313994	Share Mack	PG-13	128				
	2019-11-22	frizes	150000000	030738009	Chris Bucksennifer Lee	PG.	108				
	2013-07-08	Despirable Me 2	260000000	168061365	Piece CuffinChris Resaud	PG	100				
	2013-06-16	Man of Steel	225000000	290085518	Zack Snyder	PG-13	16)				
	2013-10-06	Gravity	500000000	270092705	Alfonsa Cuaran	PG-13	91				
	2013-06-21	Mandars University	Nav	268192764	Dan Scanion	6	107				
,	2019-12-18	The Hobbit: The Desolution of Smaug	NaN	218366855	Peter Jackson	PG-13	262				
	2013-05-26	Fact & Furlous 6	540000000	238679850	audin bir	PG-13	130				
,	2013-03-08	CZ The Grout and Powerful	215000000	230911825	Sam kaimi	PG	127				
10	2013-05-36	Star Trek Into Darliness	190000000	228778661	LL Abrams	PG-13	123				
11	2013-11-08	than the box world	170000000	206362160	Alan Taylor	PG-13	130				
12	2013-06-21	World War 2	190000000	202399711	Marcfortw	PG-13	116				
13	2013-03-22	The Croods	115000000	187168425	Kirk De MicsoChris Sanders	PG	98				
14	2013-06-28	The Heat	£30000000	159582188	Paul Feig		117				
15	2013-08-07	We're the Millers	E7000000	193996119	Rawson Mandall Thurber	t.	110				
16	2013-12-13	Assertion Huttle	03000000	150117907	David O. Russell		133				
17	2013-05-20	The Great Gotsby	209/000000	100800019	Rad Luftranann	PG-13	163				

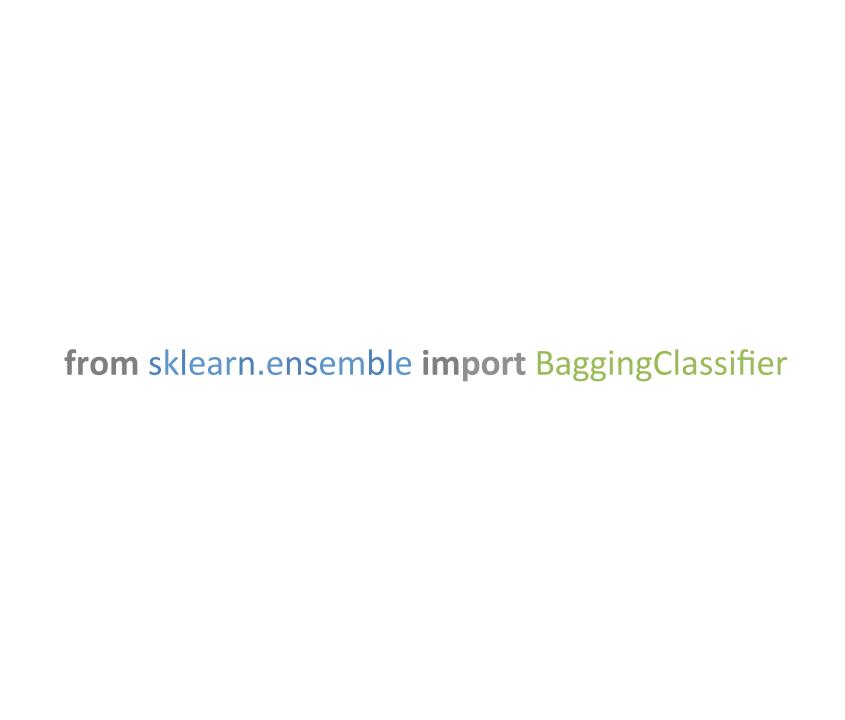
Bootstrap: Sample with replacement



Classification: Major vote

Regression: Average outcome



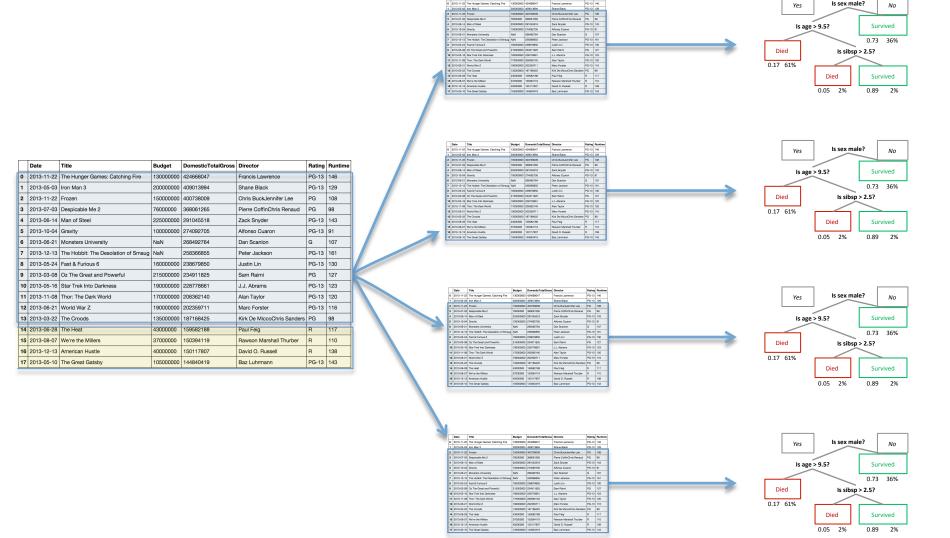


Ensemble Methods: Random Forests



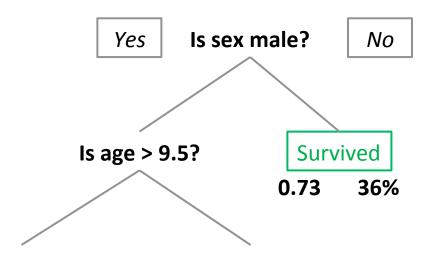


Tree Bagging with a twist

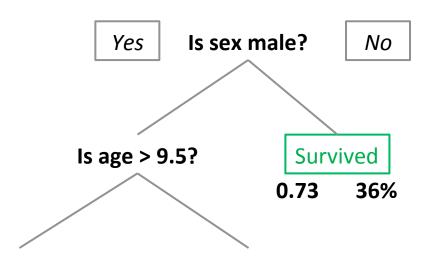


Is sex male?

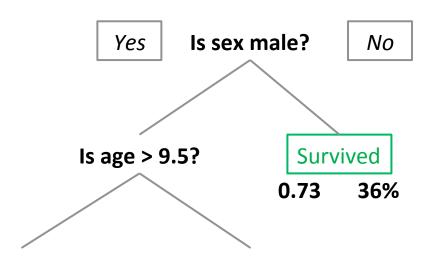
Introduce randomness when building each tree.



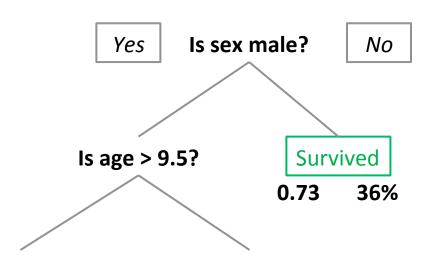
Introduce randomness when building each tree. For a split, do not take the "best" feature split.



Introduce randomness when building each tree. For a split, do not take the "best" feature split. First, randomly choose sqrt(n_feat) features.



Introduce randomness when building each tree. For a split, do not take the "best" feature split. First, randomly choose sqrt(n_feat) features. Only choose the best split among these.



from sklearn.ensemble import RandomForestClassifier from sklearn.ensemble import RandomForestRegressor

from sklearn.ensemble import RandomForestClassifier

from sklearn.ensemble import RandomForestRegressor

from sklearn.ensemble import ExtraTreesClassifier?

from sklearn.ensemble import RandomForestClassifier

from sklearn.ensemble import RandomForestRegressor

from sklearn.ensemble import ExtraTreesClassifier Even more randomized:

For each feature, split rule is random, not optimal