

Random Forest

Basic steps - Classification algorithms

Profiling Differentiation Classification

Should I invest in a company – ask the experts

Employee of XYZ

Knows internal functionality

insider information

lacks a broader perspective on competitors

has been right 70% times.

Financial Advisor of XYZ

perspective on companies vs competition

lacks a view on internal policies

has been right 75% times.

Stock Market Trader

observed company's stock price over past 3 years

knows seasonality trends and market performance

has been right 70% times.

Employee of acompetitor

internal functionality of the competitor firms

lacks a sight of company in focus and the external factors

has been right 60% of times.

Market Research team

analyzes the customer preference of XYZsproduct

unaware of the changes XYZwill bring

have been right 75% of times.

Social Media Expert

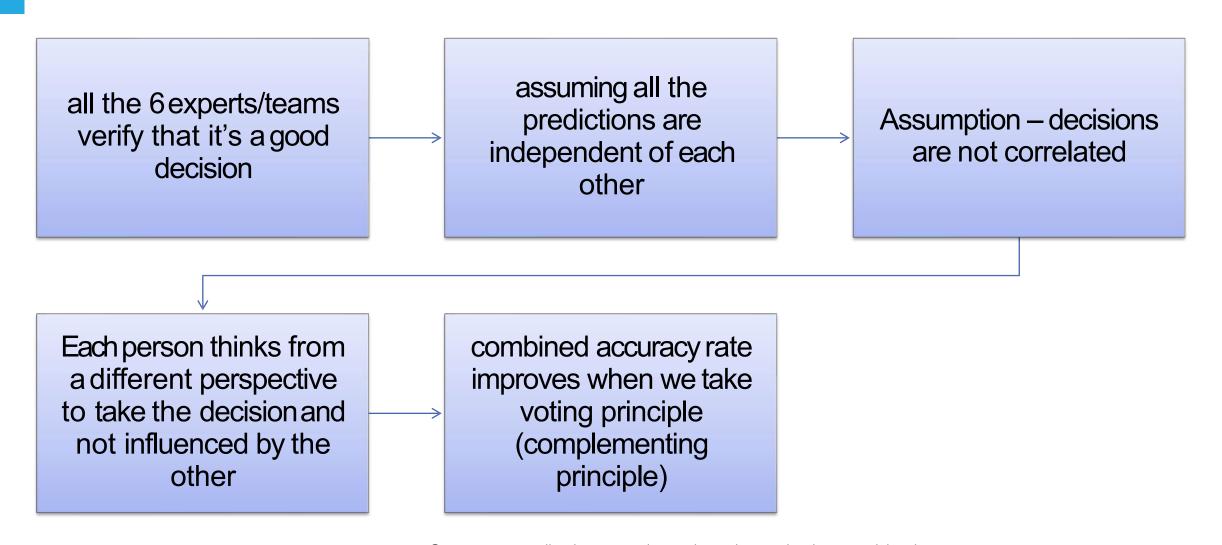
understand product positioning

Changes in customer sentiment overtime

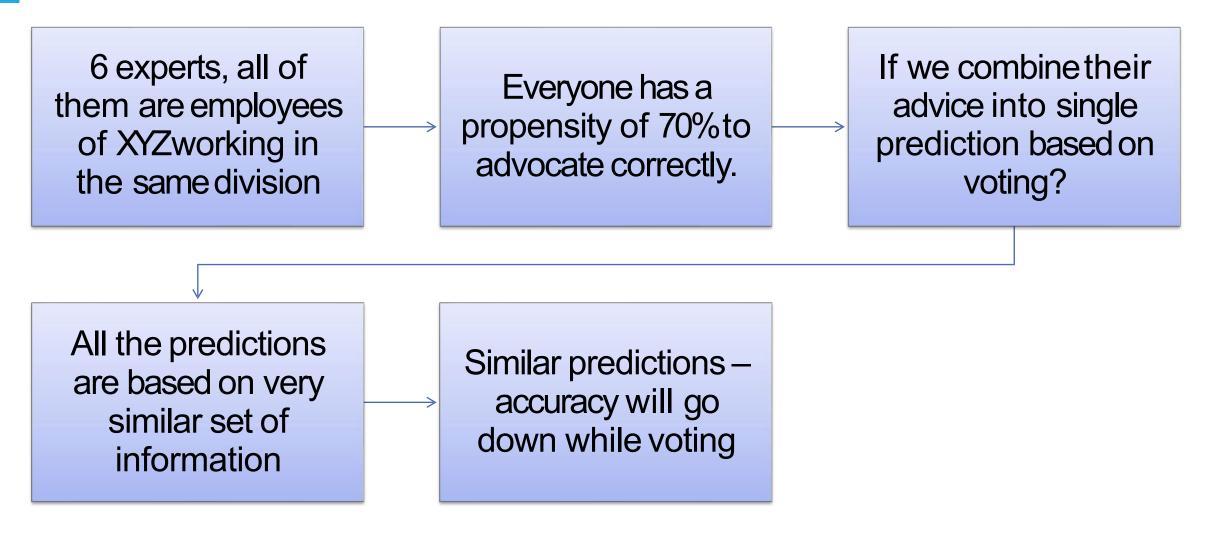
unaware of details beyond digital marketing

has been right 65% of times.

Scenario1 - Combine all the info - informed decision



Scenario 2 – info from similar sources

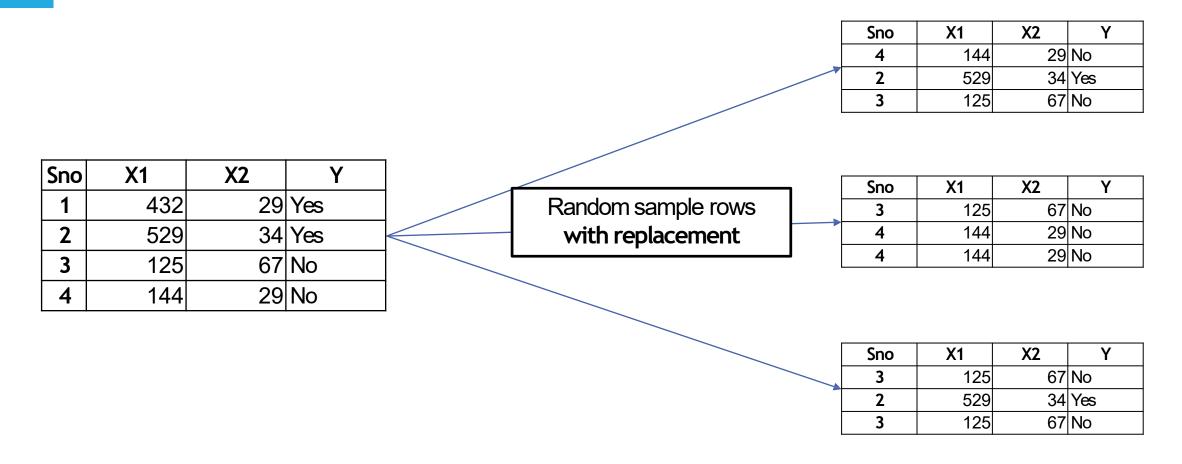


Ensemble learning

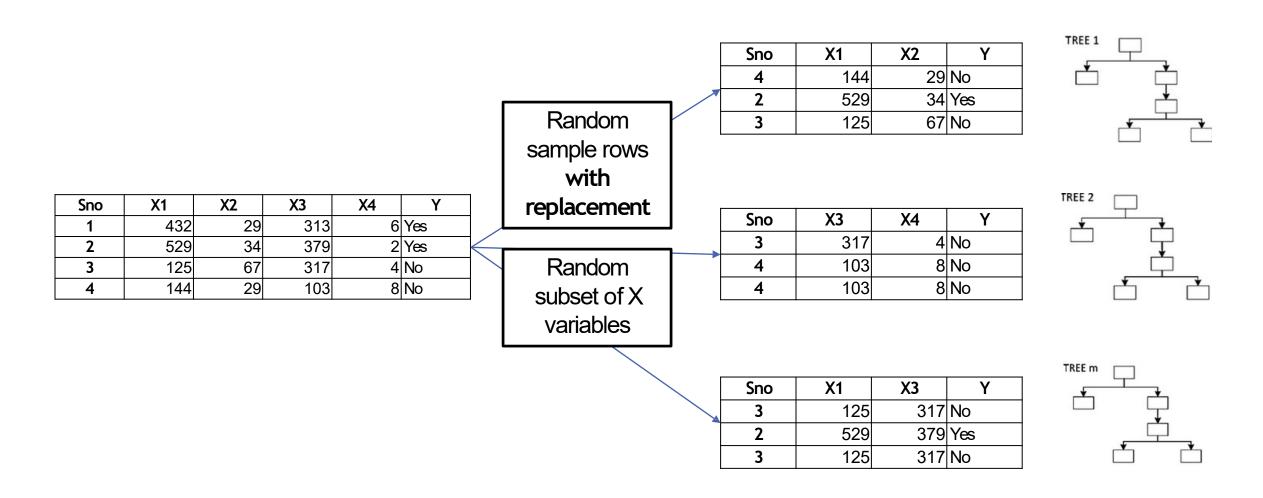
- Machine learning technique that combines several base models in order to produce one optimal predictive model.
- Weak classifiers
- Different set of variables for each classifier
- Combine into single prediction



What is a boot strapped dataset



Using a random set of variables every time



Basic idea of random forest

Draw multiple random samples, with replacement, from the data

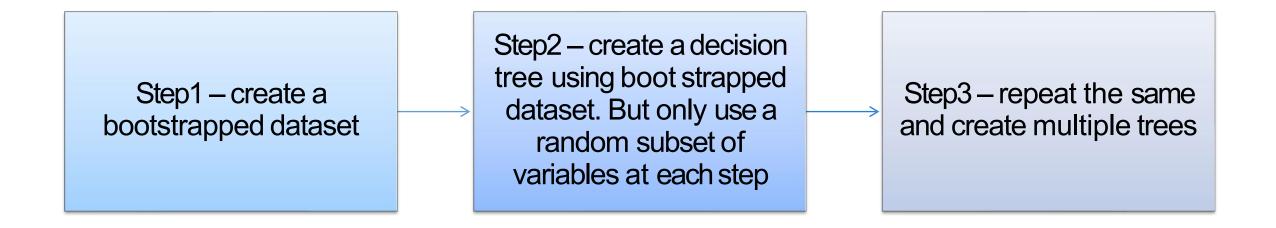
• (this sampling approach is called the *bootstrap*).

Using a random subset of predictors at each stage, fit a classification (or regression) tree to each sample (and thus obtain a "forest").

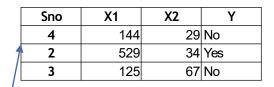
Combine the predictions/dassifications from the individual trees to obtain improved predictions.

Use voting for classification and averaging for prediction.

Steps in random forest algorithm



Out of bag data points



Sno	X1	X2	Х3	X4	Υ
1	432	29	313	6	Yes
2	529	34	379	2	Yes
3	125	67	317	4	No
4	144	29	103	8	No

Sno	Х3	X4	Υ
3	317	4	No
4	103	8	No
4	103	8	No

Sno	X1	Х3	٧
3	125	,	No
2	529	379	
3	125	317	

- When we create a bootstrapped dataset, ~1/3 of the original data does not end up in the boot strapped dataset
- This is called out-of-bag dataset

How to calculate accuracy

- OOB samples used to measure how accurate our random forest is
- by the ratio of out of bag samples correctly classified by the random forest model
- Proportion of OOB samples incorrectly classified out of bag error

How to decide on how many variables to use per step?

- Compare OOB error for using 2 variables per step, 3 variables and so on
- Choose the most accurate set of variables
- Typically we start by using square root of number of variables
- Then try a few settings above and below the value

Summary of Random forest

Consists of a large number of individual decision trees that operate as an ensemble Each tree in the random forest spits out a class prediction

dass with most votes becomes model's prediction

fundamental concept - wisdom of crowds

Alarge number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual models.

Overall flow of the RF classification process

