Predict or explain: What can machine learning do for me?

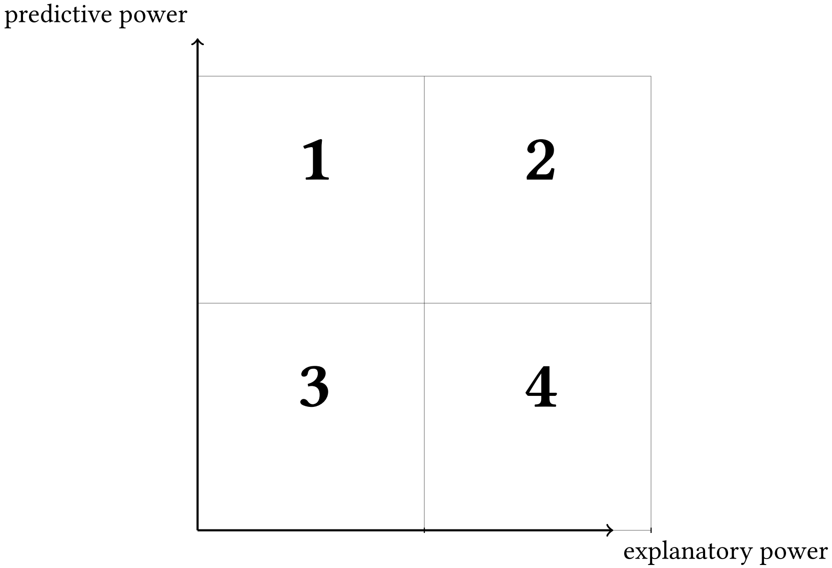
Sahir Rai Bhatnagar

Department of Epidemiology, Biostatistics, and Occupational Health
Department of Diagnostic Radiology

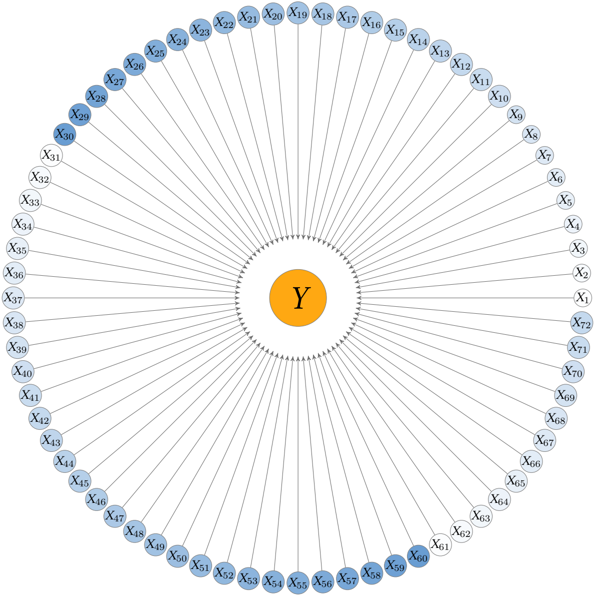
https://sahirbhatnagar.com/

March 16, 2022

CVD calculator: explain or predict?



Setting



Y
X_1
X_2
X_3
X_4
X_5
X_6
X_7
X_8
X_9
X_{10}
X_{11}
X_{12}
X_{13}
X_{14}
X_{15}
X_{16}
X_{17}
X_{18}
X_{19}
X_{20}
X_{21}
X_{22}
X_{23}
X_{24}
X_{25}
X_{26}
X_{27}
X_{28}
X_{29}
X_{30}
X_{31}
X_{32}
X_{33}
X_{34}
X_{35}
X_{36}
X_{37}
X_{38}
X_{39}
X_{40}
X_{41}
X_{42}
X_{43}
X_{44}
X_{45}
X_{46}
X_{47}
X_{48}
X_{49}
X_{50}
X_{51}
X_{52}
X_{53}
X_{54}
X_{55}
X_{56}
X_{57}
X_{58}
X_{59}
X_{60}
X_{61}
X_{62}
X_{63}
X_{64}
X_{65}
X_{66}
X_{67}
X_{68}
X_{69}
X_{70}
X_{71}
X_{72}

Nature functions to associate x with y

- A matrix of input variables **x** go in one side
- On the other side, response variable **y** comes out

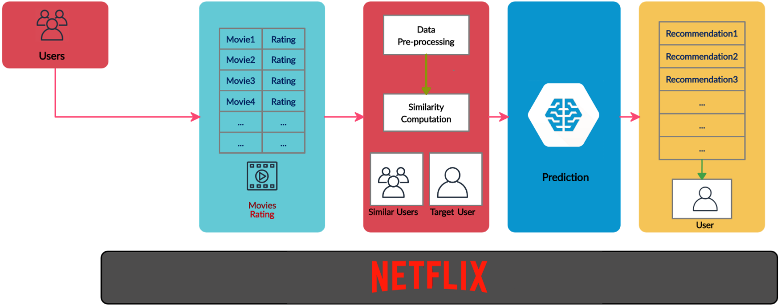
Leo Breiman. *Statistical modeling: The two cultures.* Statistical Science (2001).

Two goals in analyzing the data

1. **Prediction**: To be able to predict what the responses are going to be to future input variables

Two goals in analyzing the data

1. **Prediction**: To be able to predict what the responses are going to be to future input variables



Two goals in analyzing the data

2. **Explanation**: To extract some information about how nature is associating the response variables to the input variables.

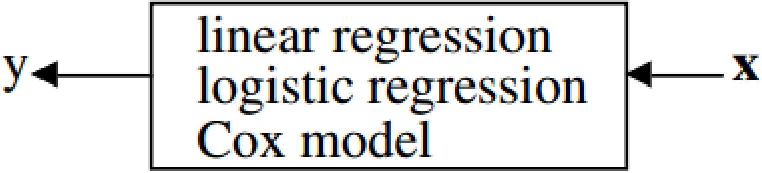
Two goals in analyzing the data

2. **Explanation**: To extract some information about how nature is associating the response variables to the input variables.

```
##
## Call:
## lm(formula = y.train ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.65540 -0.39856  0.02914  0.43816  1.81211
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.91081    0.12694  -7.175 3.33e-12 ***
## X15E1.2      0.28324    0.05369   5.275 2.14e-07 ***
## X2..PDE      0.25930    0.08366   3.099  0.00207 **
## X7A5        -0.07482    0.02419  -3.094  0.00211 **
## A1BG        -0.13033    0.04920  -2.649  0.00838 **
## A2BP1        0.05182    0.05127   1.011  0.31271
## A2M         -0.18041    0.03579  -5.040 6.95e-07 ***
## A2ML1       -0.08147    0.04788  -1.701  0.08960 .
## A3GALT2      0.09927    0.09471   1.048  0.29519
## A4GALT       0.09667    0.04494   2.151  0.03204 *
## A4GNT        0.01535    0.06841   0.224  0.82252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6503 on 417 degrees of freedom
## Multiple R-squared:  0.2052,^^IAdjusted R-squared:  0.1862
## F-statistic: 10.77 on 10 and 417 DF,  p-value: 2.514e-16
```

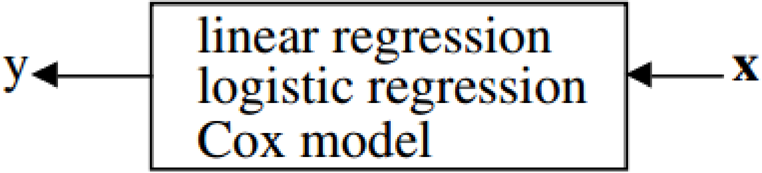
Two different approaches toward these goals

1. **Data Modelling Culture**

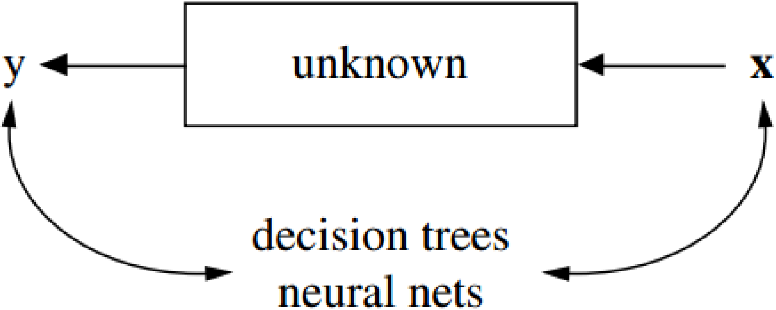


Two different approaches toward these goals

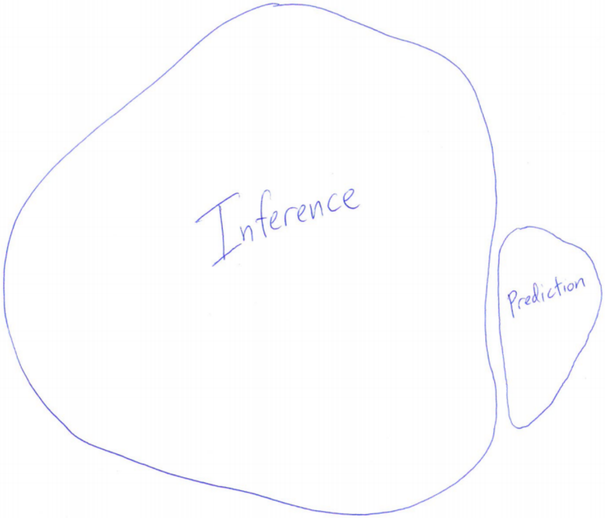
1. **Data Modelling Culture**



2. **Algorithmic Modelling Culture**



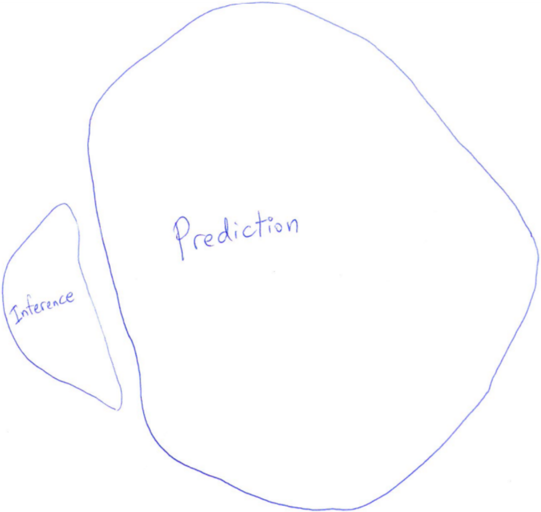
Statistics vs. Machine Learning



How statisticians see the world?

source: http://statweb.stanford.edu/ tibs/ftp/nips2015.pdf

Statistics vs. Machine Learning



How machine learners see the world?

source: http://statweb.stanford.edu/ tibs/ftp/nips2015.pdf

The focus is different

- $\mathbf{d} = \{\mathbf{x}, \mathbf{y}\}$

$$\mathbf{y}_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{x}_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{12} & \cdots & x_{1p} \\ x_{31} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{12} & \cdots & x_{np} \end{bmatrix}$$

Surface plus noise models

- Traditional regression model:

$$y_i = \underbrace{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}_{surface} + \underbrace{\varepsilon_i}_{noise} \qquad i = 1, \ldots, n$$

$$\mathbf{y} = \mathbf{x}\beta + \varepsilon$$

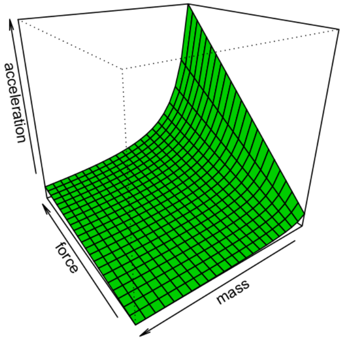
Efron B. Prediction, estimation, and attribution. International Statistical Review. 2020 Dec;88:S28-59.

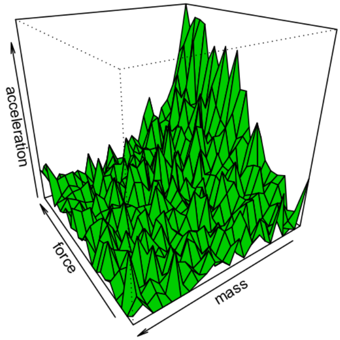
Surface plus noise models

- Traditional regression model:

$$y_i = \underbrace{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}_{surface} + \underbrace{\varepsilon_i}_{noise} \qquad i = 1, \ldots, n$$

$$\mathbf{y} = \mathbf{x}\beta + \varepsilon$$





Efron B. Prediction, estimation, and attribution. International Statistical Review. 2020 Dec;88:S28-59.

CVD risk model

Table 2. Regression Coefficients and Hazard Ratios

Variable	β*	P	Hazard Ratio	95% CI
Women [So(10)=0.95012]				
Log of age	2.32888	<0.0001	10.27	(5.65–18.64)
Log of total cholesterol	1.20904	<0.0001	3.35	(2.00–5.62)
Log of HDL cholesterol	−0.70833	<0.0001	0.49	(0.35–0.69)
Log of SBP if not treated	2.76157	<0.0001	15.82	(7.86–31.87)
Log of SBP if treated	2.82263	<0.0001	16.82	(8.46–33.46)
Smoking	0.52873	<0.0001	1.70	(1.40–2.06)
Diabetes	0.69154	<0.0001	2.00	(1.49–2.67)
Men [So(10)=0.88936]				
Log of age	3.06117	<0.0001	21.35	(14.03–32.48)
Log of total cholesterol	1.12370	<0.0001	3.08	(2.05–4.62)
Log of HDL cholesterol	−0.93263	<0.0001	0.39	(0.30–0.52)
Log of SBP if not treated	1.93303	<0.0001	6.91	(3.91–12.20)
Log of SBP if treated	1.99881	<0.0001	7.38	(4.22–12.92)
Smoking	0.65451	<0.0001	1.92	(1.65–2.24)
Diabetes	0.57367	<0.0001	1.78	(1.43–2.20)

So(10) indicates 10-year baseline survival; SBP, systolic blood pressure.
*Estimated regression coefficient

CVD risk model

Table 2. Regression Coefficients and Hazard Ratios

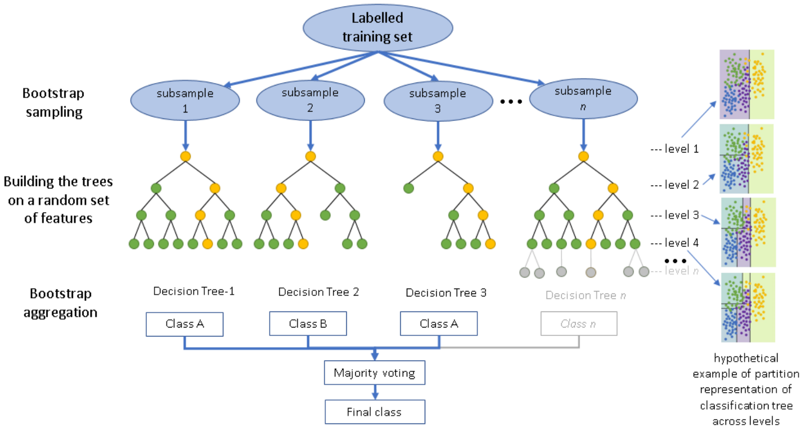
Variable	β*	P	Hazard Ratio	95% CI
Women [So(10)=0.95012]				
Log of age	2.32888	<0.0001	10.27	(5.65–18.64)
Log of total cholesterol	1.20904	<0.0001	3.35	(2.00–5.62)
Log of HDL cholesterol	−0.70833	<0.0001	0.49	(0.35–0.69)
Log of SBP if not treated	2.76157	<0.0001	15.82	(7.86–31.87)
Log of SBP if treated	2.82263	<0.0001	16.82	(8.46–33.46)
Smoking	0.52873	<0.0001	1.70	(1.40–2.06)
Diabetes	0.69154	<0.0001	2.00	(1.49–2.67)
Men [So(10)=0.88936]				
Log of age	3.06117	<0.0001	21.35	(14.03–32.48)
Log of total cholesterol	1.12370	<0.0001	3.08	(2.05–4.62)
Log of HDL cholesterol	−0.93263	<0.0001	0.39	(0.30–0.52)
Log of SBP if not treated	1.93303	<0.0001	6.91	(3.91–12.20)
Log of SBP if treated	1.99881	<0.0001	7.38	(4.22–12.92)
Smoking	0.65451	<0.0001	1.92	(1.65–2.24)
Diabetes	0.57367	<0.0001	1.78	(1.43–2.20)

So(10) indicates 10-year baseline survival; SBP, systolic blood pressure.
*Estimated regression coefficient

10-year Risk:

$$1 - S_0(t)^{\exp(2.32\cdot\log(age)+1.2\cdot\log(chol)-0.708\cdot\log(HDL)+\cdots+0.53\cdot smoker+0.69\cdot diabetic)}$$

Random forests



Random Forest Algorithm

Algorithm 17.1 RANDOM FOREST.

1 Given training data set $\boldsymbol{d} = (\boldsymbol{X}, \boldsymbol{y})$. Fix $m \leq p$ and the number of trees B.

2 For $b = 1, 2, \ldots, B$, do the following.

(a) Create a bootstrap version of the training data $\boldsymbol{d}_b^*$, by randomly sampling the n rows with replacement n times. The sample can be represented by the bootstrap frequency vector $\boldsymbol{w}_b^*$.

(b) Grow a maximal-depth tree $\hat{r}_b(x)$ using the data in $\boldsymbol{d}_b^*$, sampling m of the p features at random prior to making each split.

(c) Save the tree, as well as the bootstrap sampling frequencies for each of the training observations.

3 Compute the random-forest fit at any prediction point x_0 as the average

$$\hat{r}_{\text{rf}}(x_0) = \frac{1}{B} \sum_{b=1}^{B} \hat{r}_b(x_0).$$

4 Compute the OOB_i error for each response observation y_i in the training data, by using the fit $\hat{r}_{\text{rf}}^{(i)}$, obtained by averaging only those $\hat{r}_b(x_i)$ for which observation i was *not* in the bootstrap sample. The overall OOB error is the average of these OOB_i.

Example: Microarray study of prostate cancer

- The study involved $n = 102$ men, 52 cancer patients and 50 normal controls. Each man's genetic expression levels were measured on a panel of $p = 6033$ genes

$$\mathbf{X}_{n \times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{12} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & x_{np} \end{bmatrix}$$

Efron B. Prediction, estimation, and attribution. International Statistical Review. 2020 Dec;88:S28-59.

Example: Microarray study of prostate cancer

- The study involved $n = 102$ men, 52 cancer patients and 50 normal controls. Each man's genetic expression levels were measured on a panel of $p = 6033$ genes

$$\mathbf{X}_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{12} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & x_{np} \end{bmatrix}$$

- Random forests was used to predict normal or cancer from a man's microarray measurements. The 102 men were **randomly** divided into training and test sets of size 51 each having 25 normal controls and 26 cancer patients.

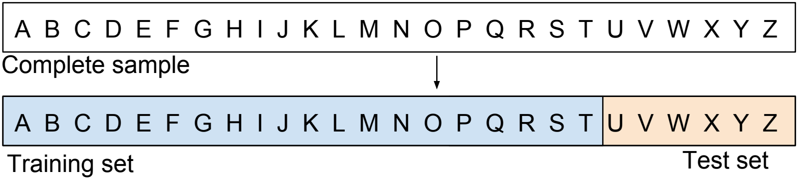
Efron B. Prediction, estimation, and attribution. International Statistical Review. 2020 Dec;88:S28-59.

Background on Train-Test split

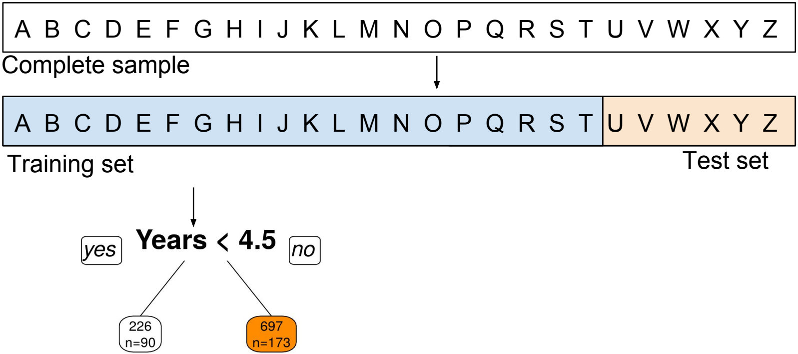
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Complete sample

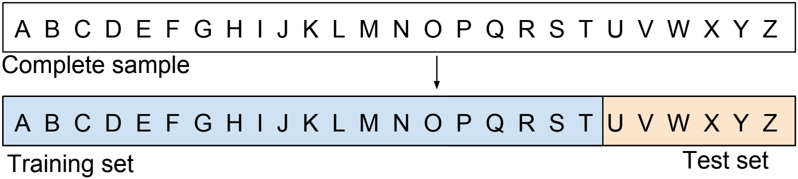
Background on Train-Test split

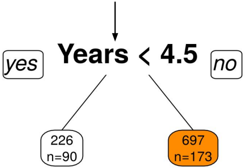


Background on Train-Test split



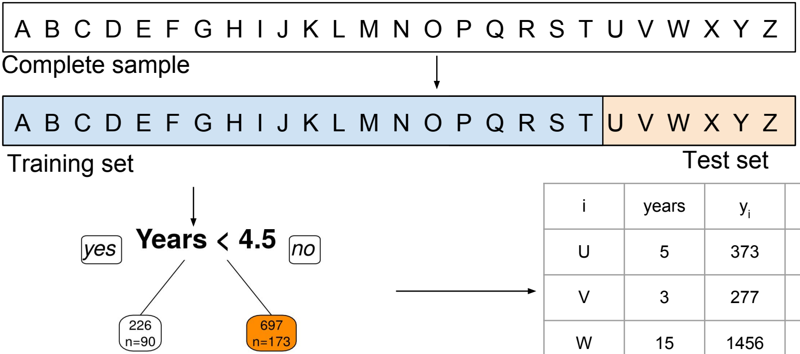
Background on Train-Test split





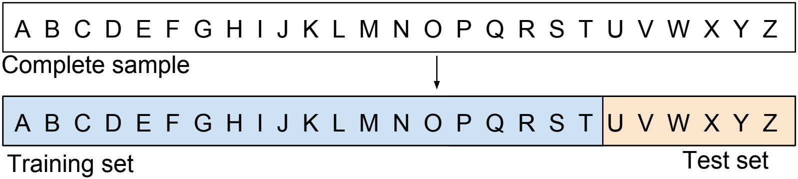
i	years	y_i	$y_i^{(pred)}$
U	5	373	
V	3	277	
W	15	1456	
X	4	455	
Y	1	235	
Z	9	987	

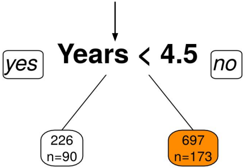
Background on Train-Test split



i	years	y_i	$y_i^{(pred)}$
U	5	373	697
V	3	277	226
W	15	1456	697
X	4	455	226
Y	1	235	226
Z	9	987	697

Background on Train-Test split

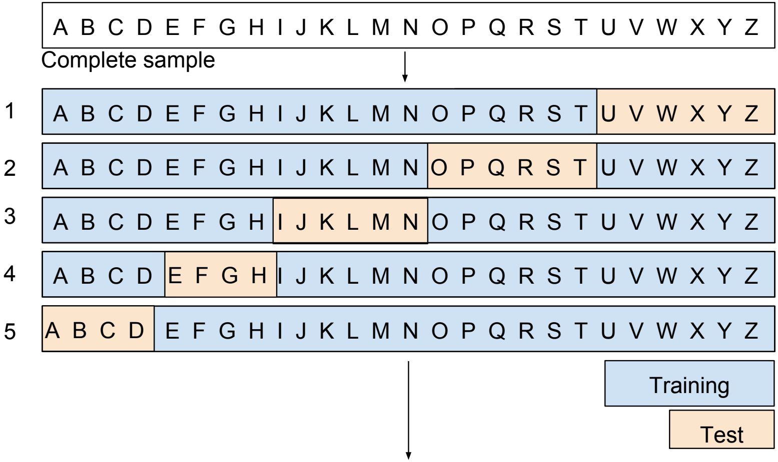




$$MSE^{(test)} = \sum_{i=1}^{6} \left(y_i - y_i^{(pred)}\right)^2 + \alpha|T|$$

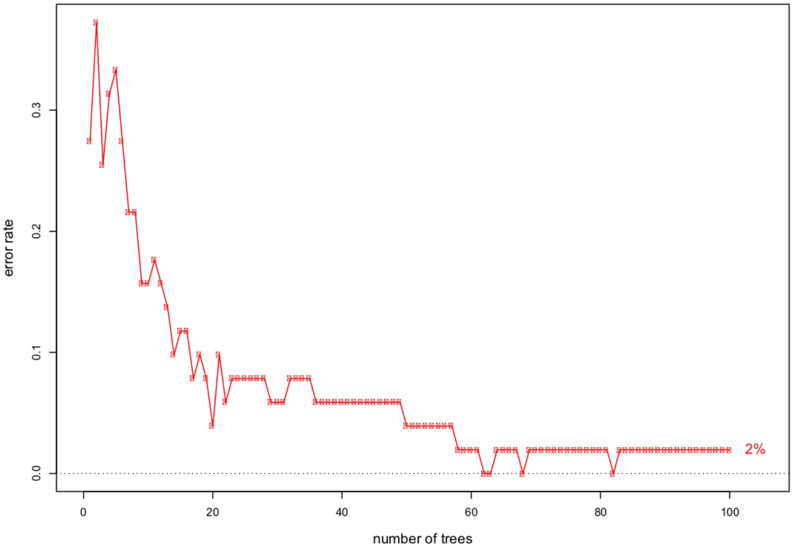
i	years	y_i	$y_i^{(pred)}$
U	5	373	697
V	3	277	226
W	15	1456	697
X	4	455	226
Y	1	235	226
Z	9	987	697

Cross-validation

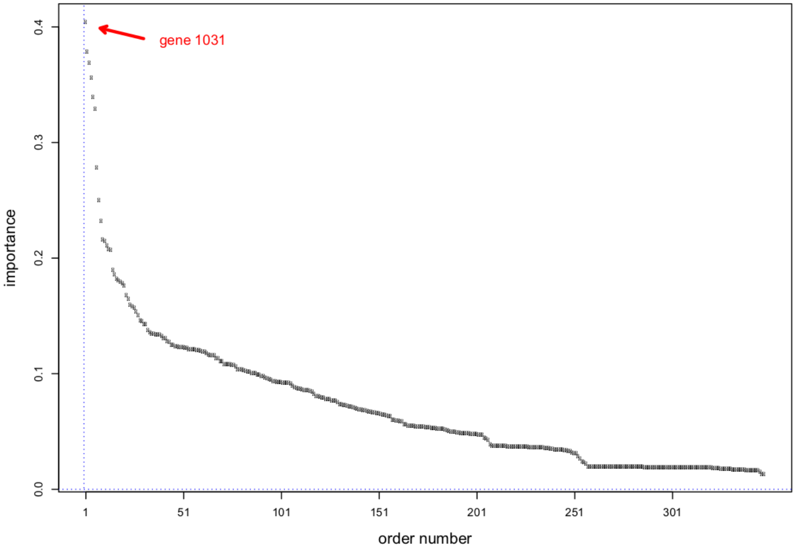


$$CV(\alpha) = \frac{1}{5} \sum_{v=1}^{5} MSE_v^{(test)}$$

Test set error rate for random forests



Variable Importance



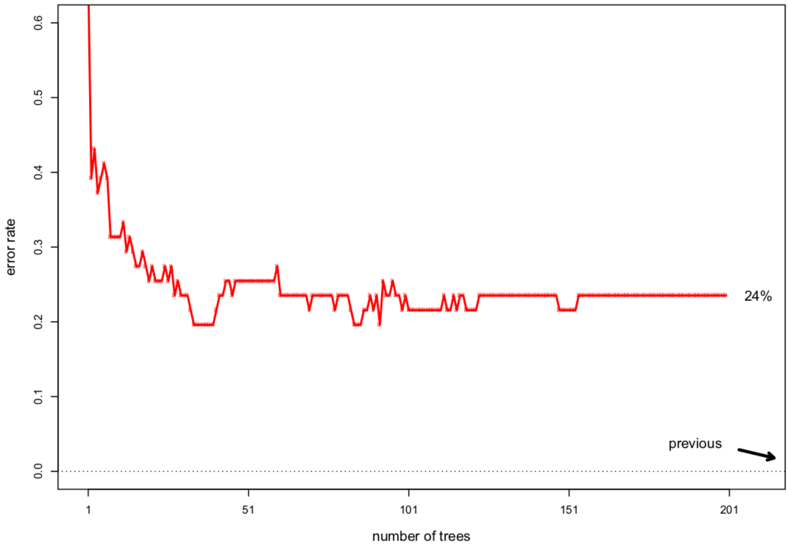
Removing the most important variables

# removed	0	1	5	10	20	40	80	160	348
# errors	1	0	3	1	1	2	2	2	0

Split train/test by early/late ID number

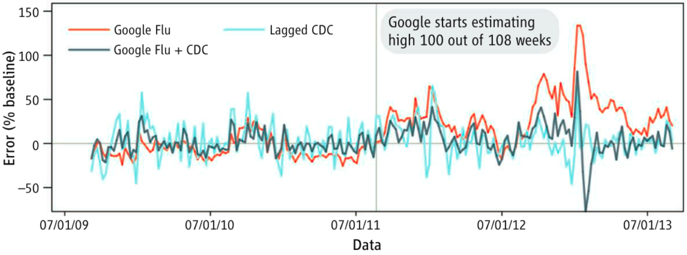
ID	gene1	gene2	...	gene 6033
1	·	·	·	·
2	·	·	·	·
3	·	·	·	·
4	·	·	·	·
5	·	·	·	·
6	·	·	·	·
7	·	·	·	·
8	·	·	·	·
9	·	·	·	·
10	·	·	·	·

Split train/test by early/late ID number



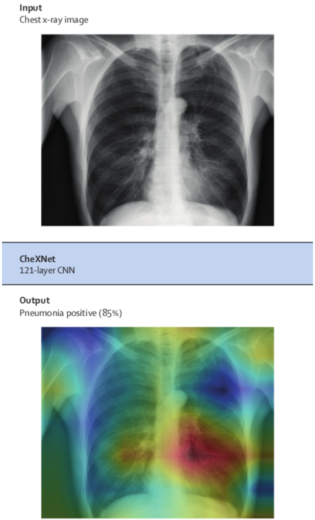
Google Flu Trends

- A machine-learning algorithm for predicting influenza outbreaks introduced in 2008 based on counts of internet search terms, outperformed traditional medical surveys in terms of speed and predictive accuracy.
- Four years later, however, the algorithm failed, badly overestimating what turned out to be a nonexistent flu epidemic.

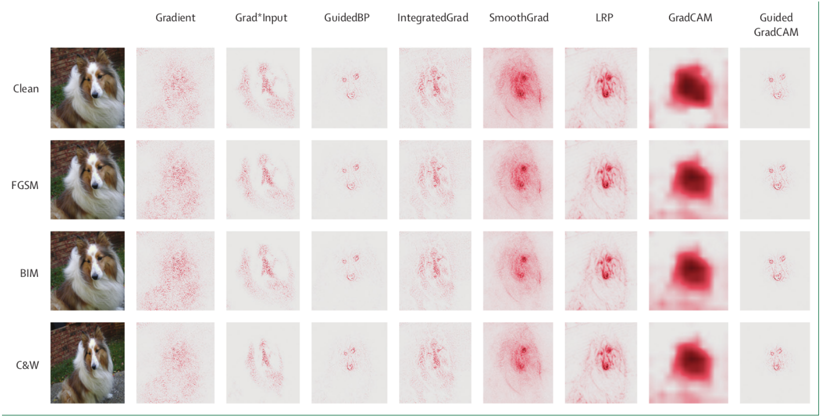


Should prediction models be interpretable?

The search for interpretable prediction models



Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health. 2021 Nov 1;3(11):e745-50.



Gu J, Tresp V. Saliency methods for explaining adversarial attacks. arXiv 2019; published online Aug 22. http://arxiv.org/ abs/1908.08413.

Acetaminophen

ANALGESICS

New clues in the acetaminophen mystery

Although acetaminophen (paracetamol) has been used clinically for more than a century, its mode of action is still not clear. Writing in the *Journal of Biological Chemistry*, Zygmunt and colleagues have now provided evidence for a new and unexpected mechanism through which acetaminophen could exert its analgesic effects.

Acetaminophen differs significantly from aspirin and other non-steroidal anti-inflammatory drugs (NSAIDs), with which it is often was this path of investigation that was followed by Zygmunt and colleagues. The stimulus for their studies was the striking relationship between the structures of acetaminophen and the *N*-acyl phenolamine AM404, which is both a potent activator of the ion channel $TRPV_1$ and has effects on cannabinoid CB_1 receptors. Both $TRPV_1$ and CB_1 receptors are involved in pain and thermoregulatory pathways and are viewed as promising targets for the treatment of pain and inflammation.

synthesize AM404 from *p*-aminophenol and arachidonic acid *in vitro*. In addition, no formation of AM404 was

Kirkpatrick P. New clues in the acetaminophen mystery. Nature Reviews Drug Discovery. 2005 Nov;4(11):883-.

Comparison

	Traditional regressions methods	Pure prediction algorithms
1.	Surface plus noise models (continuous, smooth)	Direct prediction (possibly discrete, jagged)
2.	Scientific truth (long-term)	Empirical prediction accuracy (possibly short-term)
3.	Parametric modeling (causality)	Nonparametric (black box)
4.	Parsimonious modeling (researchers choose covariates)	Anti-parsimony (algorithm chooses predictors)
5.	$\mathbf{x}$ $p \times n$: with $p \ll n$ (homogeneous data)	$p \gg n$, both possibly enormous (mixed data)
6.	Theory of optimal inference (mle, Neyman–Pearson)	Training/test paradigm (Common Task Framework)

Efron B. Prediction, estimation, and attribution. International Statistical Review. 2020 Dec;88:S28-59.

Message # 1

Explanatory power $\neq$ Predictive power

Shmueli G. To explain or to predict?. Statistical science. 2010 Aug;25(3):289-310.

Message # 1

Explanatory power $\neq$ Predictive power

Best explanatory model $\neq$ Best predictive model

Shmueli G. To explain or to predict?. Statistical science. 2010 Aug;25(3):289-310.

Message # 2: Explain vs. Predict

In-sample vs. Out-of-sample

Shmueli G. To explain or to predict?. Statistical science. 2010 Aug;25(3):289-310.

Message # 2: Explain vs. Predict

In-sample vs. Out-of-sample

- Interpretation
- Statistical Significance
- Goodness of fit

- Prediction accuracy

Shmueli G. To explain or to predict?. Statistical science. 2010 Aug;25(3):289-310.

Message # 2: Explain vs. Predict

In-sample vs. Out-of-sample

- Interpretation
- Statistical Significance
- Goodness of fit

- Prediction accuracy

- Type I, II errors

- Over-fitting

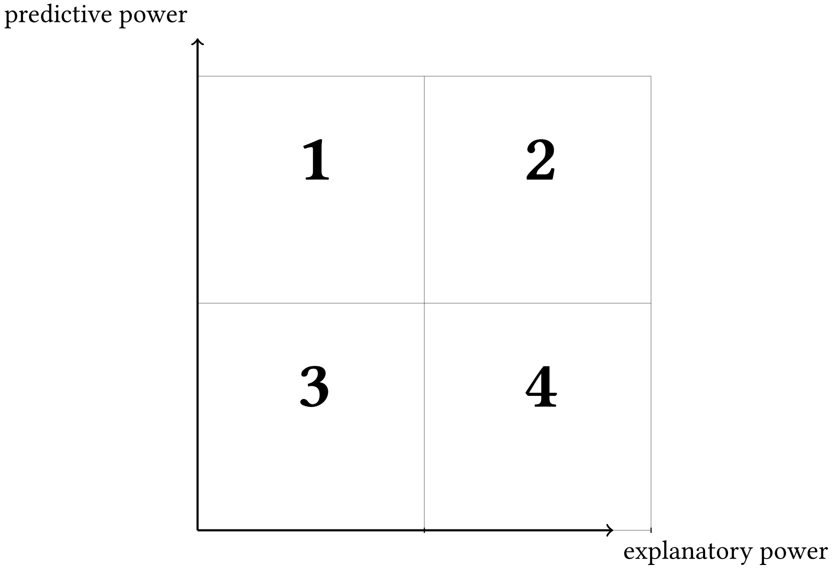
Shmueli G. To explain or to predict?. Statistical science. 2010 Aug;25(3):289-310.

Message # 3

Explaining is harder than Predicting

Eternal vs. Ephemeral

CVD calculator: explain or predict?



CVD calculator: explain or predict?