Binary classification of breast cancer data

Project 4 - Statistical learning MT7049

August Jonasson & Martin Löfström

2025-01-14

Introduction and data

- Breast cancer data: 30 covariates - 10 cell nucleus base features
  - *mean*, *standard deviation* and *worst occurence*
- Binary response: Benign or malign
- Tree-based methods: GBM and random forests

Feature	**Description**
Radius	Mean of distances from center to points on the perimeter
Texture	Standard deviation of gray-scale values
Perimeter	Total distance around the cell nucleus
Area	Total area enclosed within the cell nucleus
Smoothness	Local variation in radius lengths
Compactness	(perimeter^2 / area) - 1
Concavity	Severity of concave portions of the contour
Concave points	Number of concave portions of the contour
Symmetry	Symmetry of the cell nucleus
Fractal dimension	Coastline approximation minus 1

Exploratory analysis

- colinearity: several covariates / base feature
  - Does not influence model
  - may influence variable importance

	area1	area2	area3	perimeter1	perimeter2	perimeter3	radius1	radius2	radius3	texture1	texture2	texture3
area1	1.000000	0.800086	0.959213	0.986507	0.726628	0.959120	0.987357	0.732562	0.962746	0.321086	-0.066280	0.287489
area2	0.800086	1.000000	0.811408	0.744983	0.937655	0.761213	0.735864	0.951830	0.757373	0.259845	0.111567	0.196497
area3	0.959213	0.811408	1.000000	0.941550	0.730713	0.977578	0.941082	0.751548	0.984015	0.343546	-0.083195	0.345842
perimeter1	0.986507	0.744983	0.941550	1.000000	0.693135	0.970387	0.997855	0.691765	0.969476	0.329533	-0.086761	0.303038
perimeter2	0.726628	0.937655	0.730713	0.693135	1.000000	0.721031	0.674172	0.972794	0.697201	0.281673	0.223171	0.200371
perimeter3	0.959120	0.761213	0.977578	0.970387	0.721031	1.000000	0.965137	0.719684	0.993708	0.358040	-0.102242	0.365098
radius1	0.987357	0.735864	0.941082	0.997855	0.674172	0.965137	1.000000	0.679090	0.969539	0.323782	-0.097317	0.297008
radius2	0.732562	0.951830	0.751548	0.691765	0.972794	0.719684	0.679090	1.000000	0.715065	0.275869	0.213247	0.194799
radius3	0.962746	0.757373	0.984015	0.969476	0.697201	0.993708	0.969539	0.715065	1.000000	0.352573	-0.111690	0.359921
texture1	0.321086	0.259845	0.343546	0.329533	0.281673	0.358040	0.323782	0.275869	0.352573	1.000000	0.386358	0.912045
texture2	-0.066280	0.111567	-0.083195	-0.086761	0.223171	-0.102242	-0.097317	0.213247	-0.111690	0.386358	1.000000	0.409003
texture3	0.287489	0.196497	0.345842	0.303038	0.200371	0.365098	0.297008	0.194799	0.359921	0.912045	0.409003	1.000000

Model selection GBM

Hyperparameters:

(i) learning rate (shrinkage) (ν),

(ii) subsampling fraction (η),

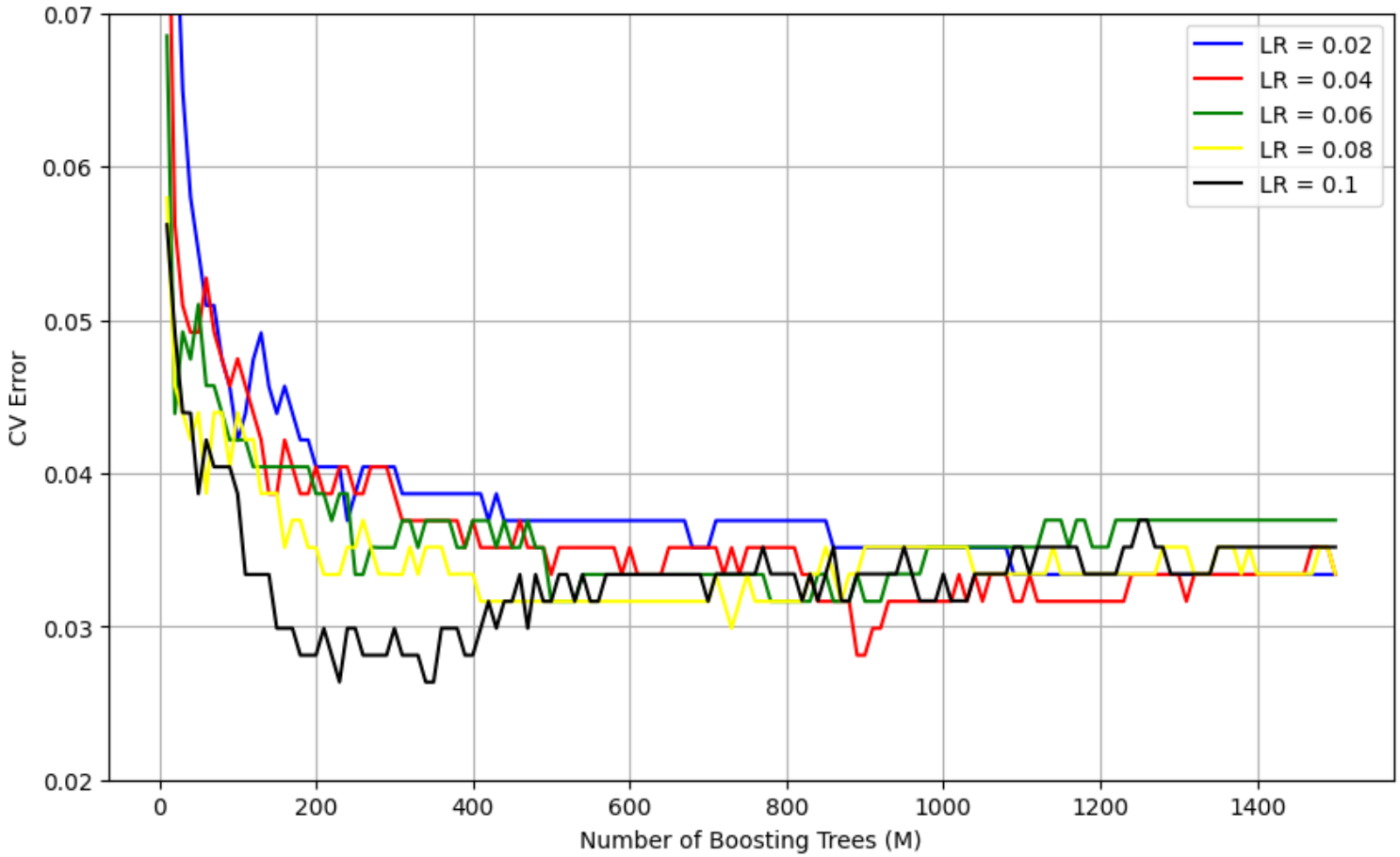
(iii) number of leaves per tree (N), and

(iv) number of trees (M).

- Full 4-dim grid-search optimal but too computationally expensive
- Settle for naive approach of tuning one at a time while following heuristics to fix others

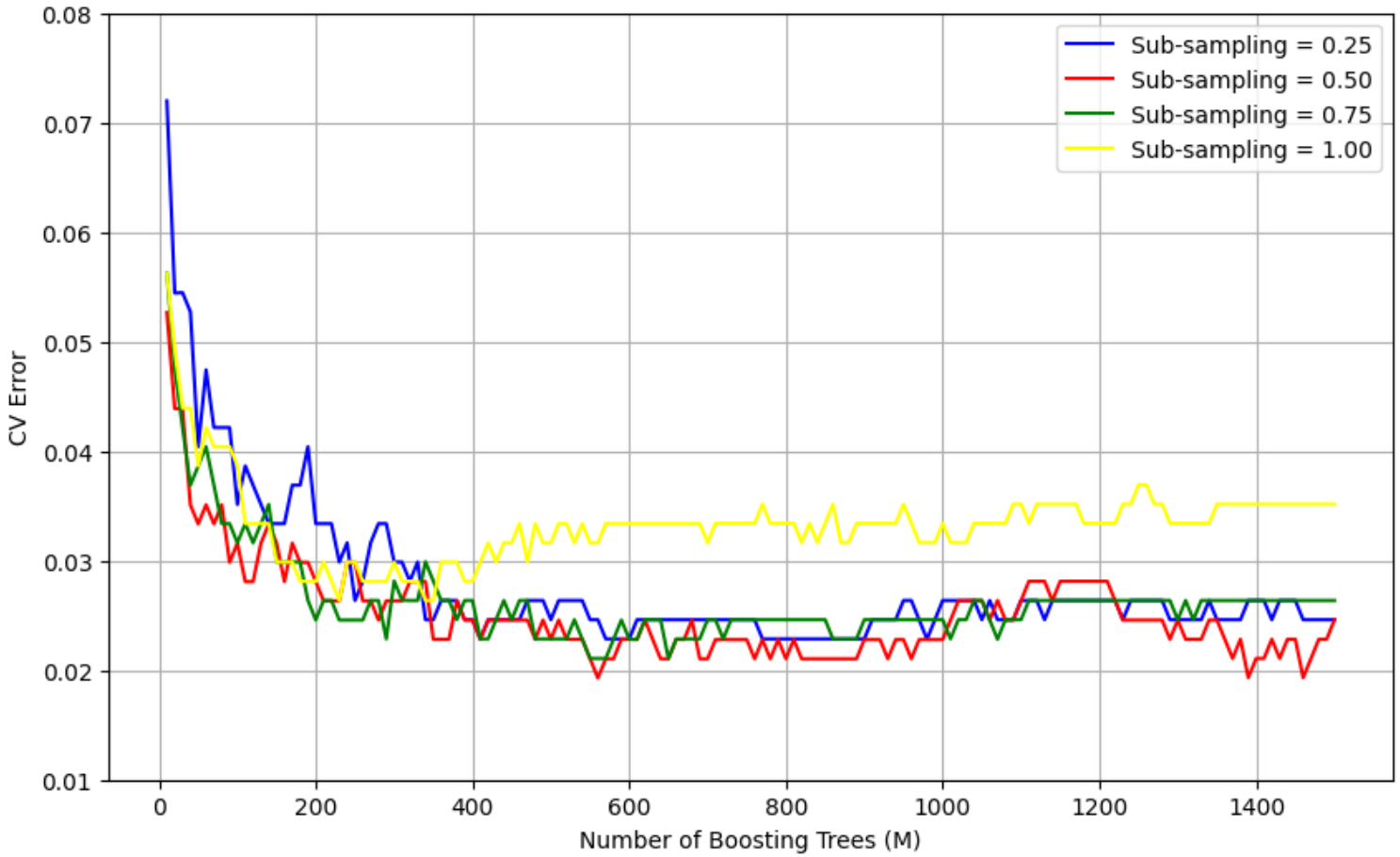
Model selection GBM

(i) learning rate (ν),



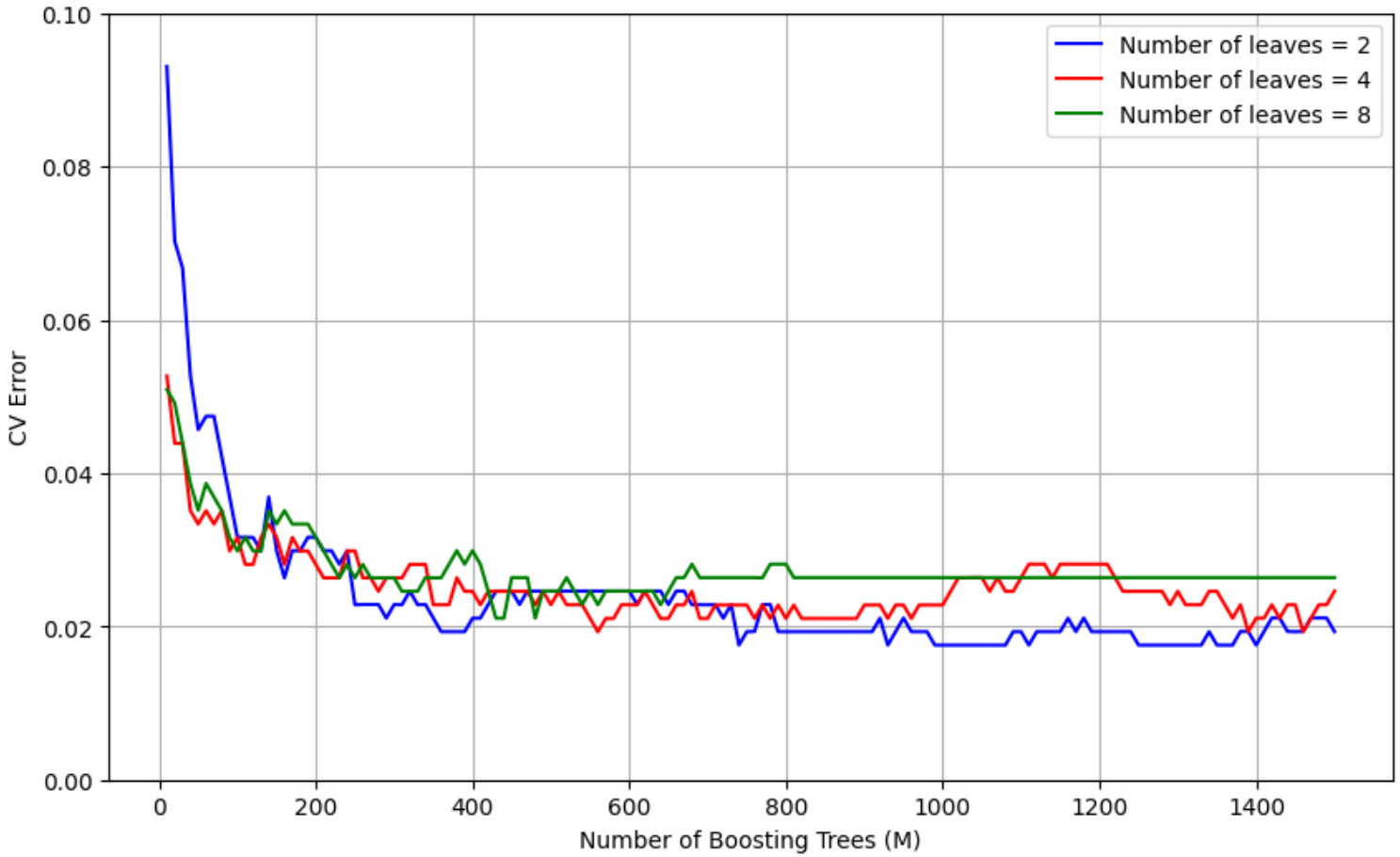
Model selection

(ii) subsampling fraction (η)



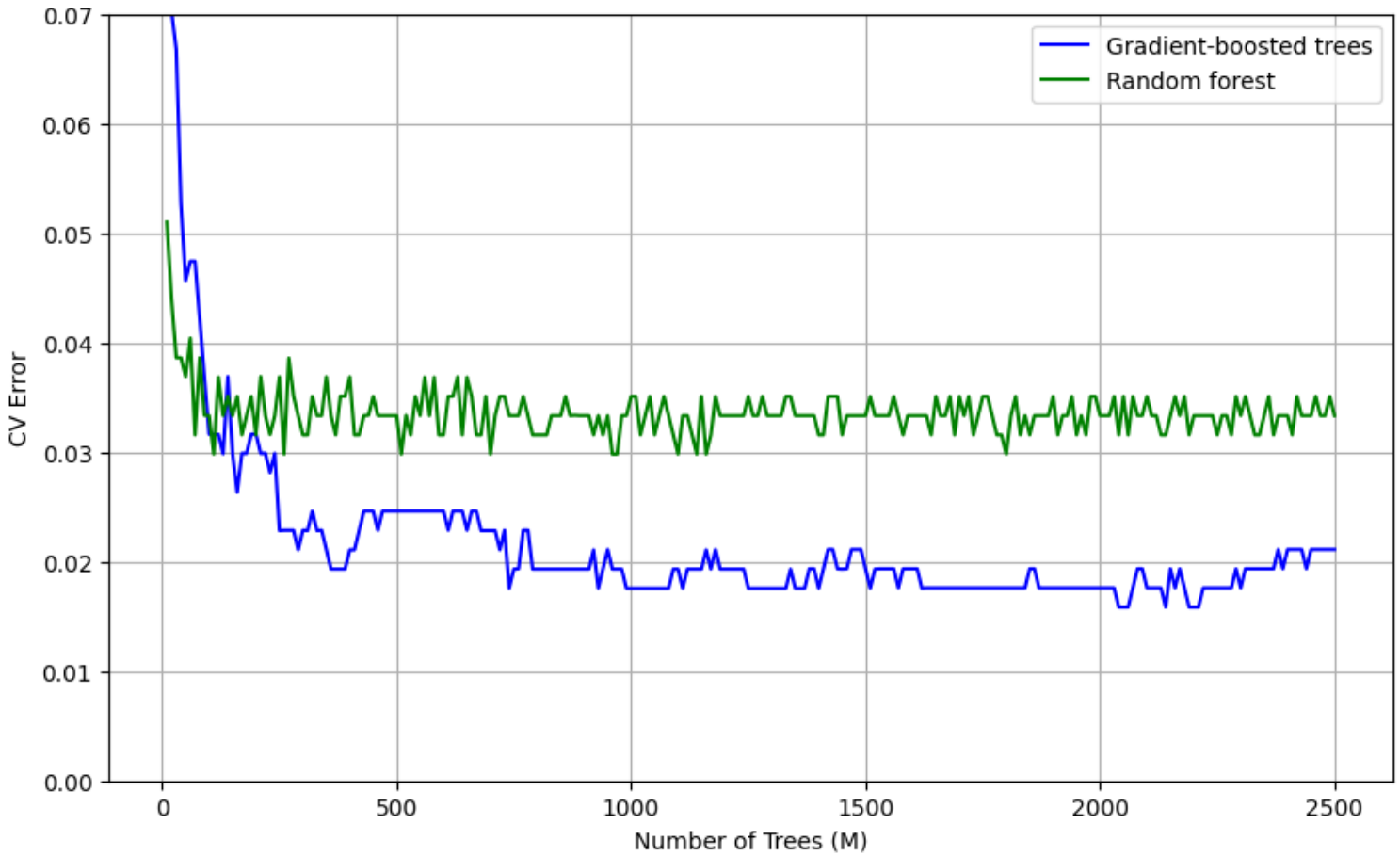
Model selection

(iii) number of leaves per tree (N)



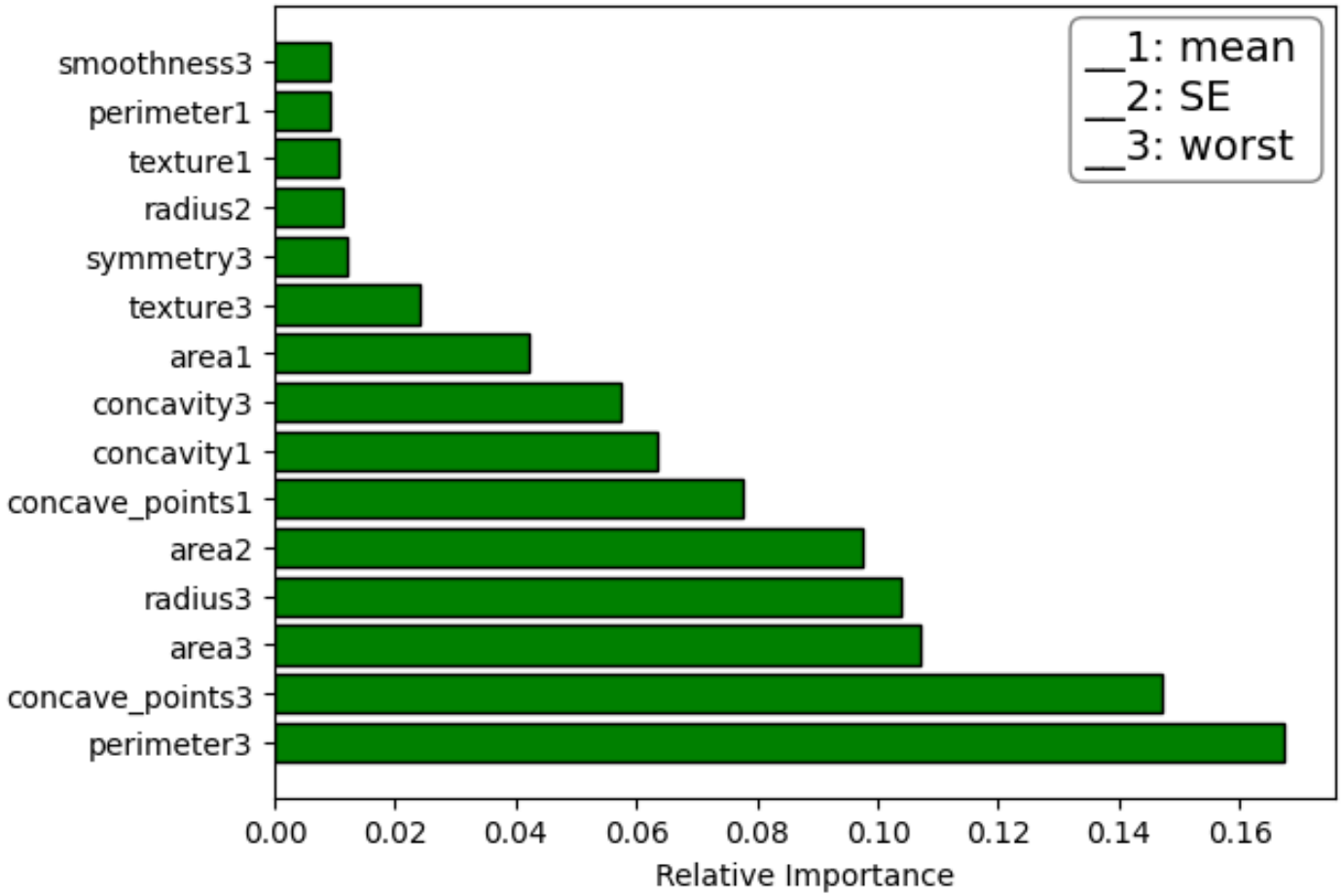
Model selection

Tuned GBM model vs standard random forest



Analysis and interpretation

- **Worst occurence** most important
- false negatives/positives $\sim$ split data



Possible improvements

- Generalisation error
- No inherent way of addressing interaction effects (as opposed to SVM)
  - would have to be added manually into the covariates
- exists models that generally performs better on prediction tasks (such as SVM)