Feature Analysis
and
Comparison of Supervised Models

Mantas Stankevičius

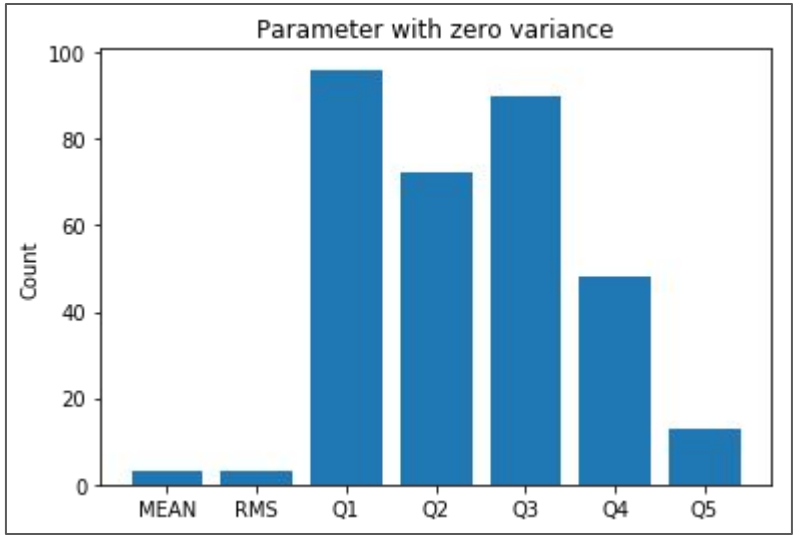
Feature variance

Zero variance

325 features

Highest variance

17838340872397 (~1.78 x 10^13)



https://github.com/mantydze/ml4dc/blob/master/src/analysis/feature_variance.ipynb

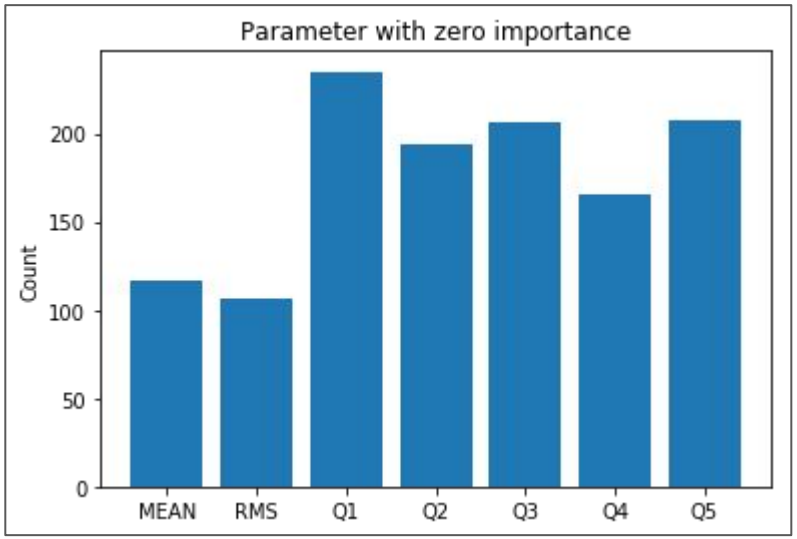
Feature importance (random forest)

Random forest model

Mean ROC AUC 0.971

Zero importance

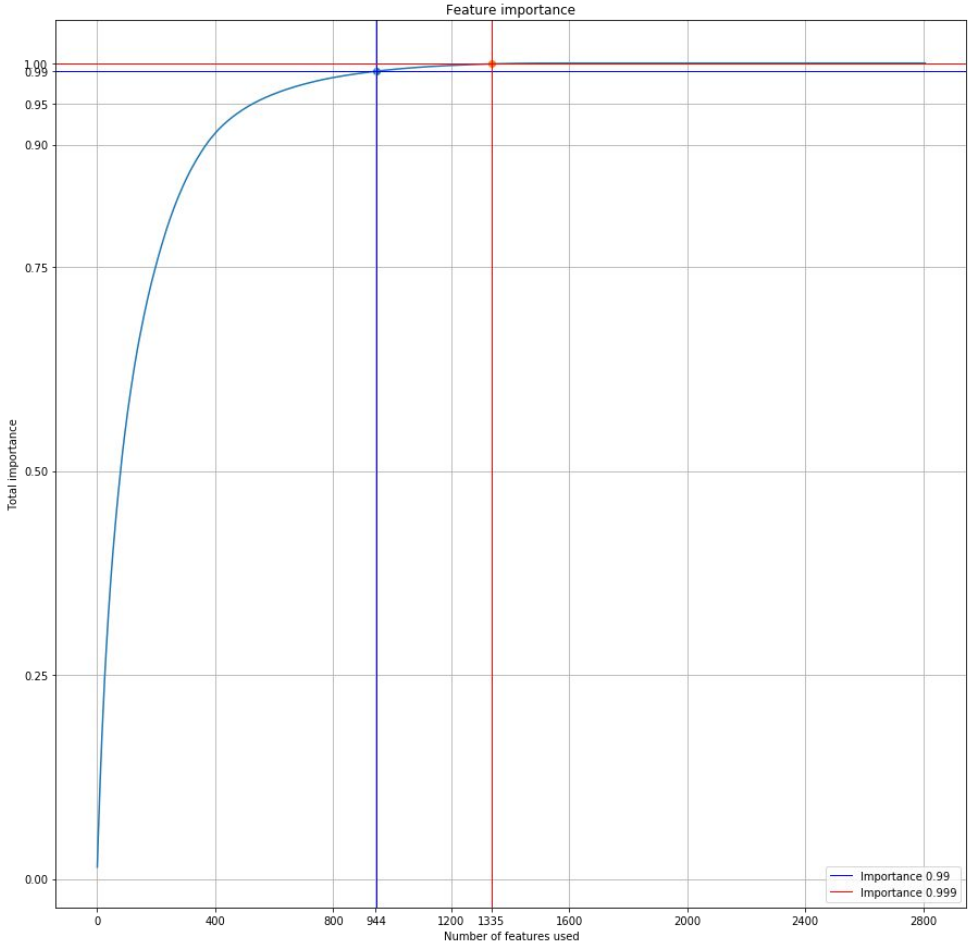
1232 features



https://github.com/mantydze/ml4dc/blob/master/src/analysis/random_forest_feature_importance.ipynb

Feature importance (random forest)

Total importance %	# of features
0.97	527
0.98	769
0.99	945
0.999	1335



Supervised models

Compare different types of methods: probabilistic, ensemble and hierarchical:

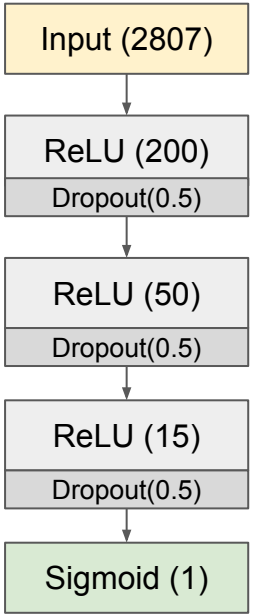
- Neural Network (ANN)
- Naive Bayes (NB)
- Random Forest (RF)
- XGBoost (XGB)

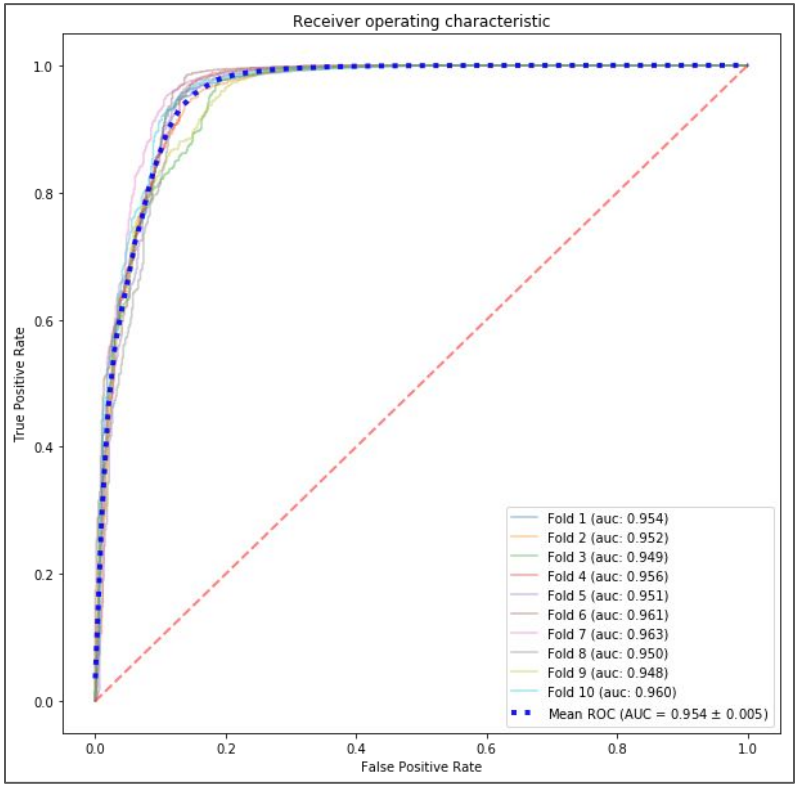
https://github.com/mantydze/ml4dc/tree/master/src/models

Neural Network

Mean AUC = 0.954

- ☑ Average predictions
- ☑ Slow search of hyper-parameters



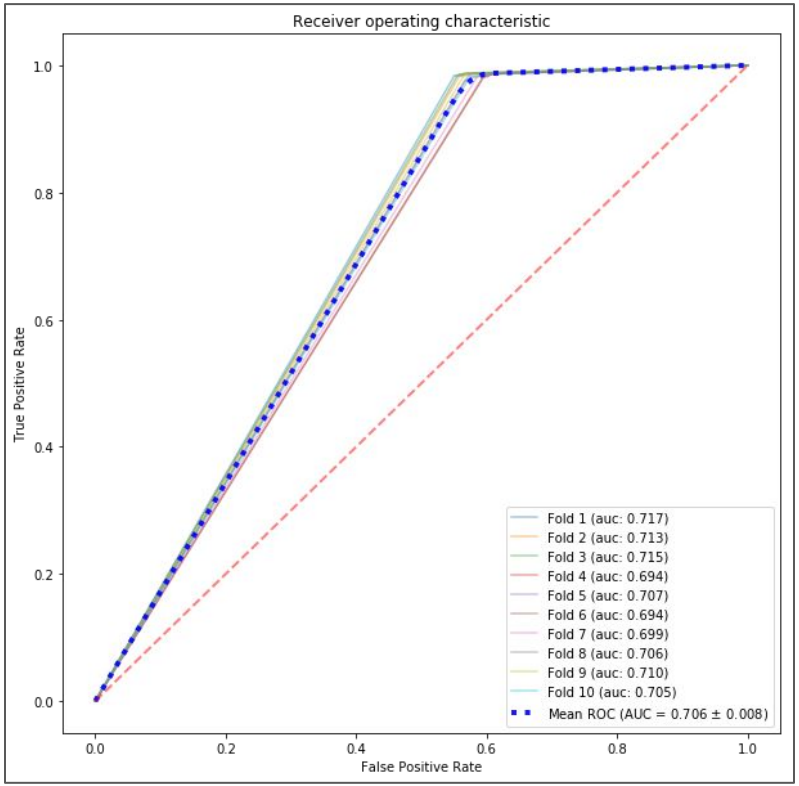


Naive Bayes

Probabilistic method.

Mean AUC = 0.706

- ☑ Fast training
- ☑ Poor predictions

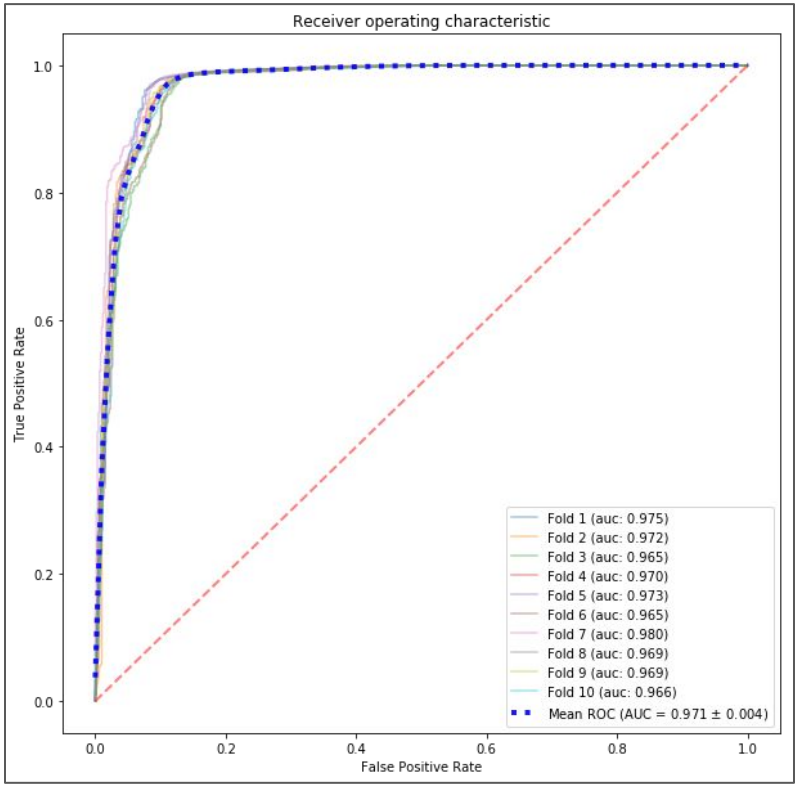


Random Forest

Ensemble method

Mean AUC = 0.971

- ☑ Fast training
- ☑ Good predictions
- ☑ Large forest may be slow predictor

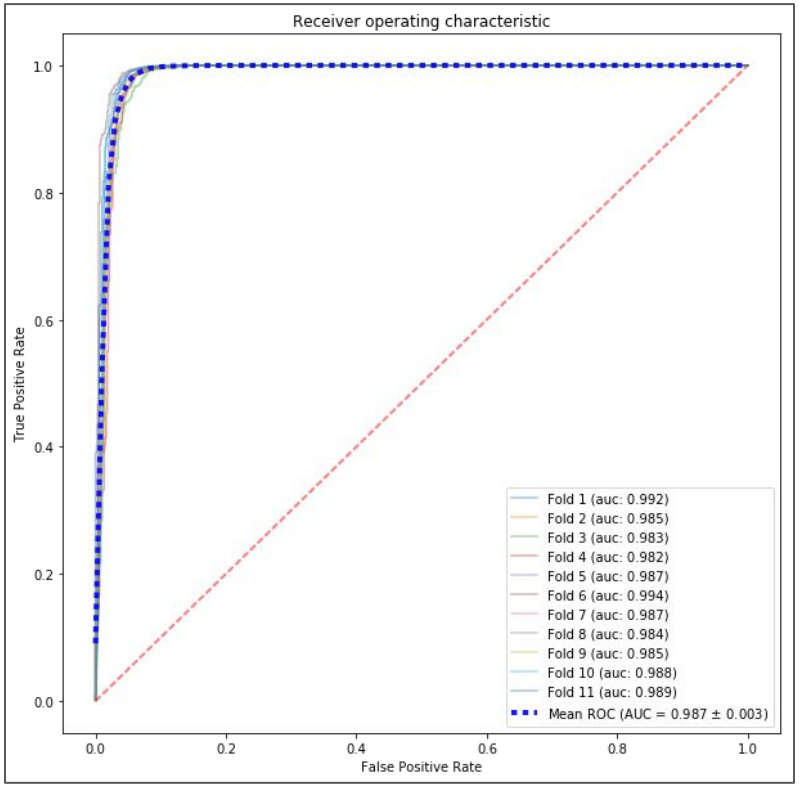


XGBoost

Gradient boosted decision trees.

Mean AUC = 0.987

- ☑ Good predictions
- ☑ Average training speed
- ☑ High memory usage during training



Comparison

Cross Validation

- StratifiedShuffleSplit
- 10 times
- 80:20

	AUC	AUC (std)	ACC	ACC (std)	F1	F1 (std)	sec	sec (std)
XGB	**0.987**	0.004	**0.997**	0.000	**0.998**	0.000	108.090	2.621
RF	0.970	0.004	0.980	0.001	**0.990**	0.000	44.925	2.490
ANN	0.954	0.005	0.961	0.015	0.979	0.008	130.236	38.413
NB	0.706	0.008	0.971	0.002	0.985	0.001	10.529	1.289