Can PCA extract important information from not significant features? Neural Network case.

Project presentation

Maciej Odziemczyk

Faculty of Economic Sciences, University of Warsaw

Introduction

Project goals:

- Check usefulness of the Principal Component Analysis applied on non-signigicant features in boosting Neural Network performance,
- build good enough bankruptcy prediction model (this kind of dataset was used, Random Forest, Extreme Gradient Boosting and Neural Networks were considered),
- check many feature selection methods (6(5)),
- build a better NN model than Tomczak et al. (the same data was used, authors reported poor NN performance),

Dataset

Polish companies bankruptcy dastaset was used, it contains 64 continuous features (financial indicators) and binary target, 1 if company went bankrupt in choosen horizon 0 if not. The whole dataset contains observations from 2007-2013 period and it was splitted into 5 tasks dependent on forecast horizon, in this study one year was choosen (the biggest potential).

- 5910 observation,
- high level of imbalance (410 samples labeled as "1", near 7%),
- a lot of missings,
- multicoliearity and correlation of features.

Methods

Because dataset was non-trivial, some methods were applied to solve some problems.

- small number of observations - 5-fold Cross Validation,
- imbalance - Startified version of CV, AUC-PR as a main metric, AUC-ROC as a helper metric,
- missings - data imputation Random Forest proximity based algorithm implemented by hand
- bad quaility features: feature selection methods for PCA, such that:
  - Random Forest feature importance score (folds average),
  - Mutual Information,
  - Spearman rank correlation with target, and between features,
  - General to Specyfic procedure based on Logistic Regression and Loglikelihood Ratio test (robust to Lovell's bias),
  - Lasso Logistic Regression (L1 norm penalty)

Models

Random Forest, Extreme Gradient Boosting (with gbtree and dart cores) and Neural Networks were considered. NN strucutres below (number of nodes was reduced).

Figure: 2 layer NN structure

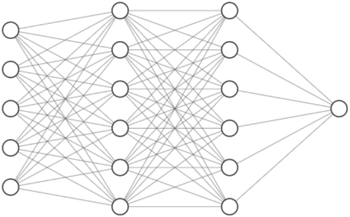
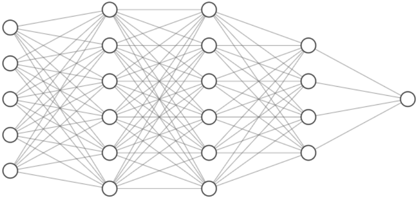


Figure: 3 layer NN structure



List of optimized hyperparameters

Random Forest

- max depth (8)
- class weights (1:1)
- num of trees (100)
- max features (53)
- min samples split (7)
- min samples split (5)

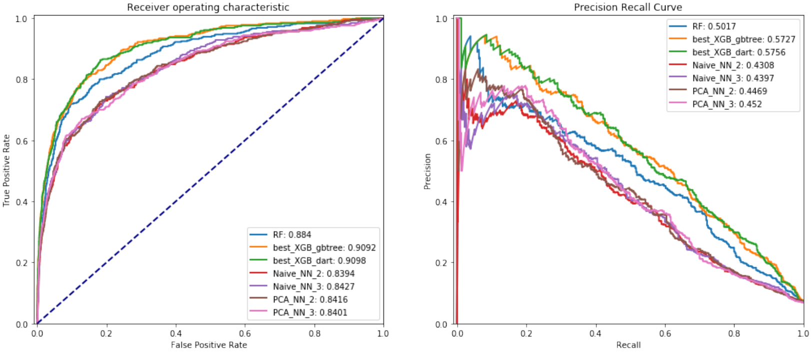
XGB

- max depth (5)
- η (0.08)
- subsample (0.9)
- max features by tree (0.8)
- max features by level (1)
- regularization (0.1 L2)
- γ (4)
- dropout (0.2 - dart)
- dropout skip (0.6 - dart)

Neural Networks

- structure [60,60],[60,60,40],
- no batch norm,
- batch size: 350,
- activations: [tanh, σ],[tanh, tanh, σ],
- dropout: [0.4,0.4],[0.4,0.4,0]
- L2: [0,0],[0,0,0.001]
- optimizer: RMSprop,
- 400 epochs.

Figure: ROC and PR results



Wilcoxon test rejected the null hypotesis about median of folds AUC-PR results equality between 2 layer NN and PCA 2 layer NN ($p = 0.0216$) in favor on the better PCA combined model performance.

Conclusions

- PCA can extract important information from non-important features but it is not a magic trick nor the rule, but it is nice to try especially with high dimensional, correlated data,
- Random Forest impurity based feature importance score is the best feature selection method, but all of presented may be successfully used (except Spearman continous vs bianry),
- If overfitting is not a problem XGB outperforms RF and NNs in this task, but NNs still have potential (some feature engineering, more epochs etc.). NNs are also less overfitted (3-4 times),
- If you don't have time, just use a Random Forest, tune it and it'll be fine.

References

- Maciej Zieba, Sebastian K. Tomczak, Jakub M. Tomczak. Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction. *Expert Systems With Applications*, 58(1): 93–101, 2016
- Mu-Yen Chen. Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert Systems With Applications*, 38: 11261–11272, 2011
- Martin Scholz George Forman. Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*, 12, 2010.

The End

Thank for your attention.