Quality over Quantity

Analysing models and features predicting bankruptcy

Davide Bacigalupi and Pietro Pianini

Sant'Anna
Scuola Universitaria Superiore Pisa

Outline

1. Dataset description
2. Research questions
3. Data manipulation
4. Analysing Models
5. Solving the unbalanced dataset problem
6. Analysing Features
7. Conclusions

Dataset description

- Data on listed companies collected from the Taiwan Economic Journal for the years from 1999 to 2009

- Target binary variable: company bankruptcy

- Explanatory features: 94 (28 on company's solvency; 9 on capital structure; 19 on profitability; 13 turnover ratios; 5 cash flow ratios; 8 on growth; 12 others)

- Instances: 6819

Research questions

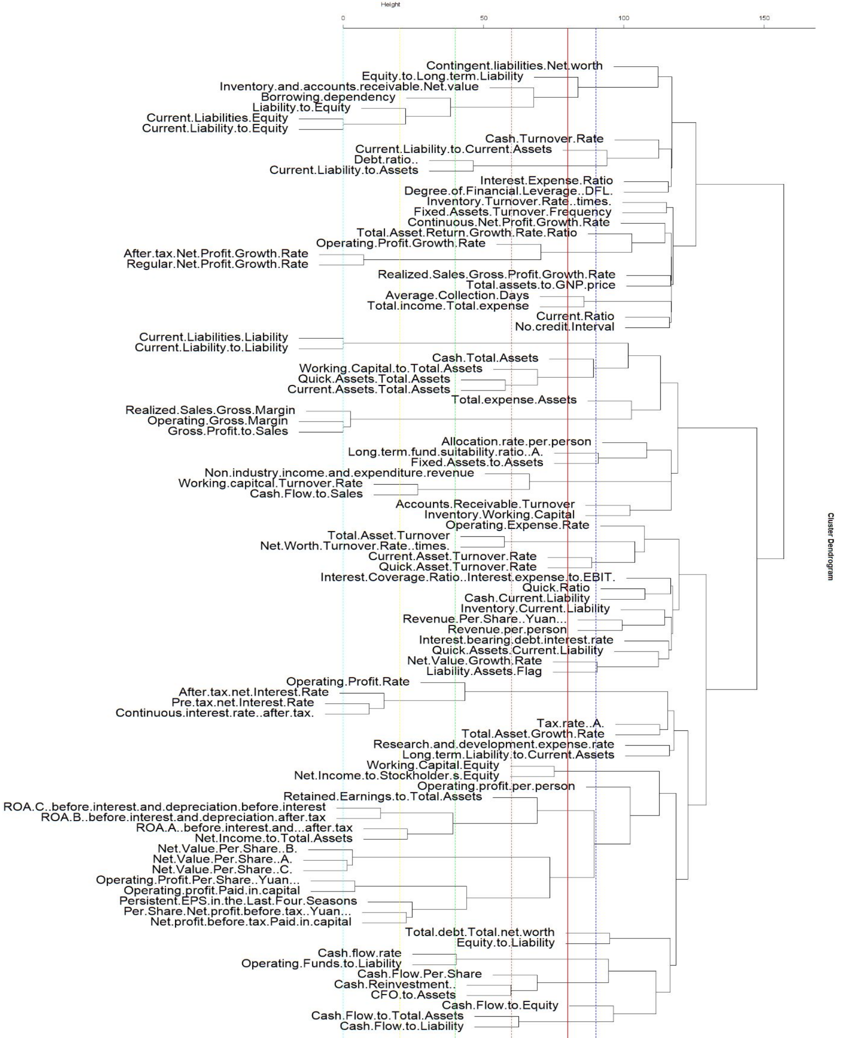
- On methods: how to deal with an unbalanced dataset with many correlated features

- On models: what is the best feature selection model (the one with the highest performance) ?

- On features: 94 features are a lot, are they all really necessary? Are we able to select very few very indicative ones?

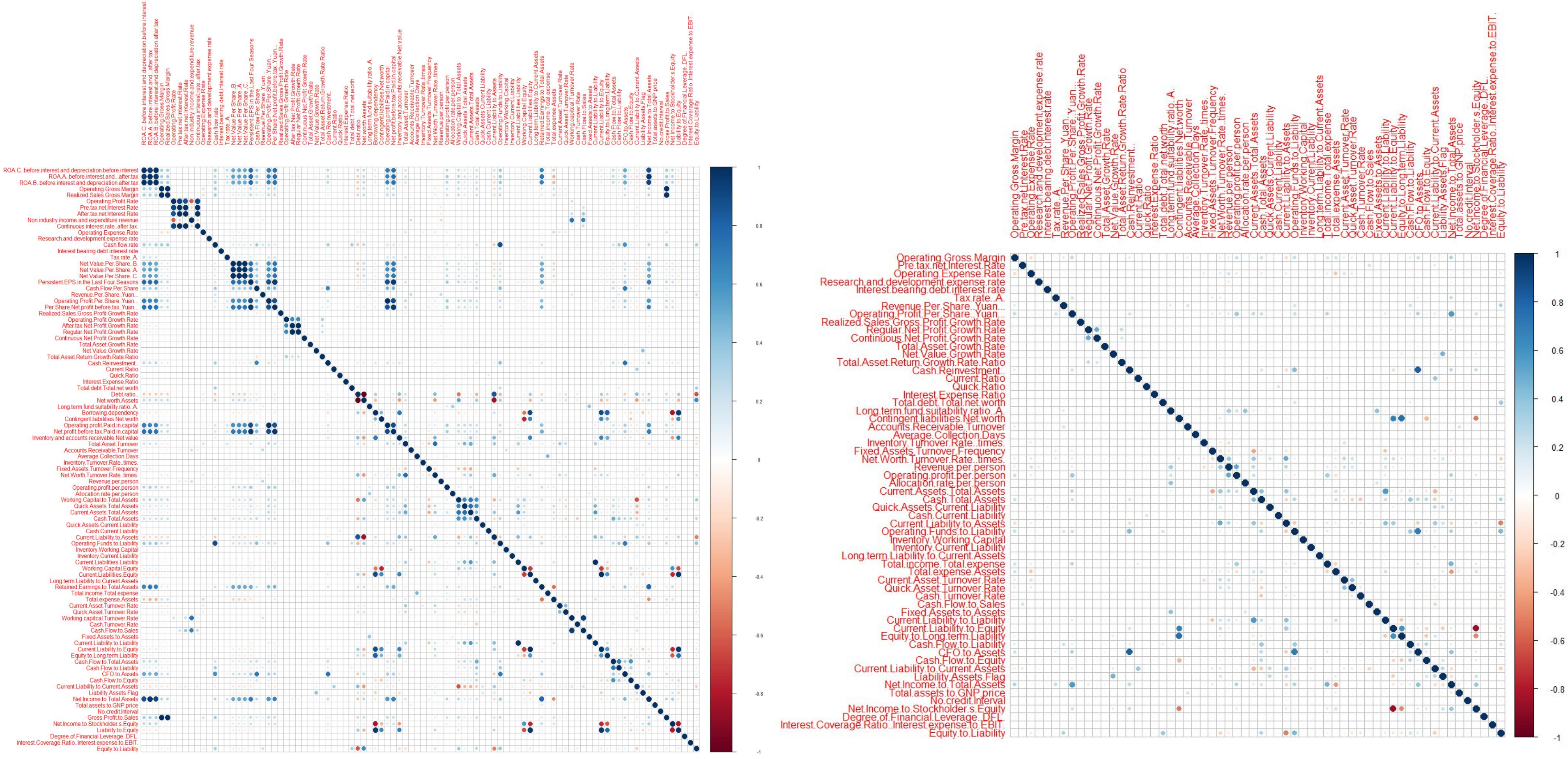
Data manipulation: cutting the dendrogram

- In our dataset there are a lot of redundant and almost multicollinear variables!
- The names of these variables are very similar to each other… (like, for example, “Current Liabilities Liability” and “Current Liability to Liability”)
- We cut the dendrogram at height = 80 and we select only one variable per group (this is needed in order to avoid multicollinearity issues and the <u>irrepresentability problem</u> of methods like the Ridge and LASSO)
- After the preprocessing, we scale the dataset and we divide it into a training set and a test set (70% - 30%)

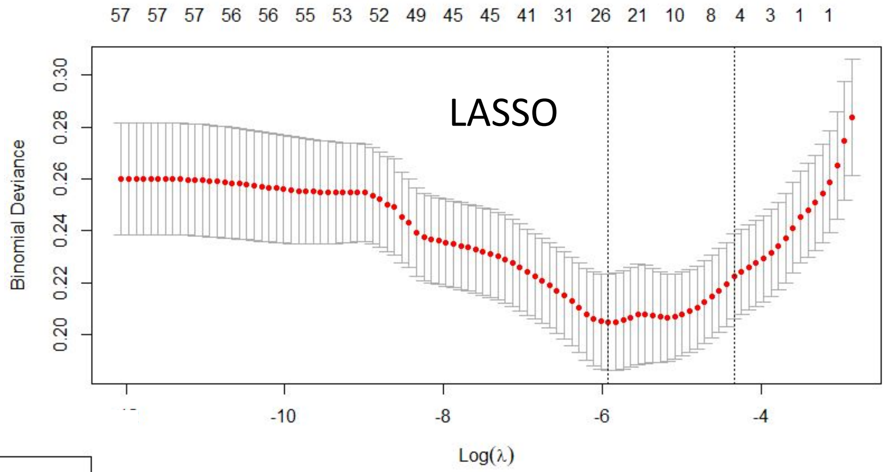


Sant’Anna
Scuola Universitaria Superiore Pisa

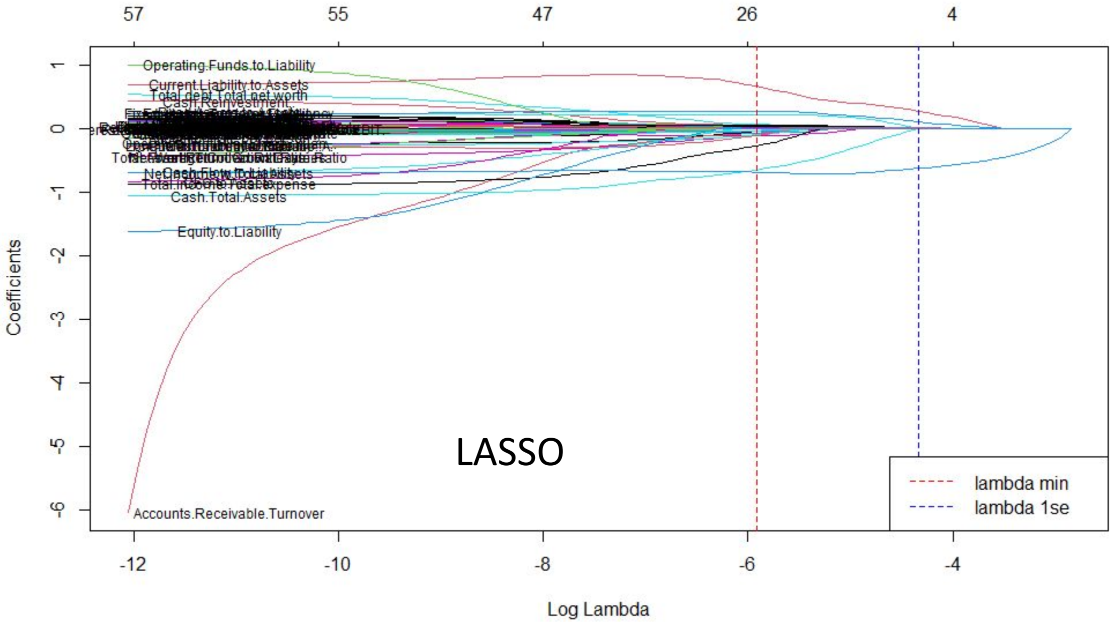
Corrplots: before and after preselection



LASSO and Elastic Net



We first tried to use LASSO to select relevant features and then we tried an Elastic Net with alpha=0.5



We use the lambda 1se (1-standard error distant from the minimum) to select features

We selected only 7 features out of 60!

(Current.Liability.to.Assets,Fixed.Assets.to.Assets,Equity.to.Long.term.Liability,Current.Liability.to.Current.Assets,Liability.Assets.Flag,Net.Income.to.Total.Assets,Net.Income.to.Stockholder.s.Equity)

Feature Screening and Feature Selection

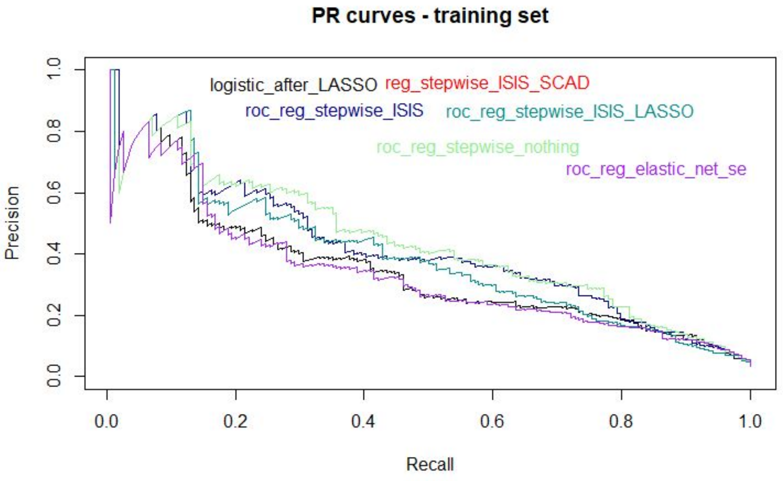
- Afterwards, we performed on the data some Feature Selection algorithms

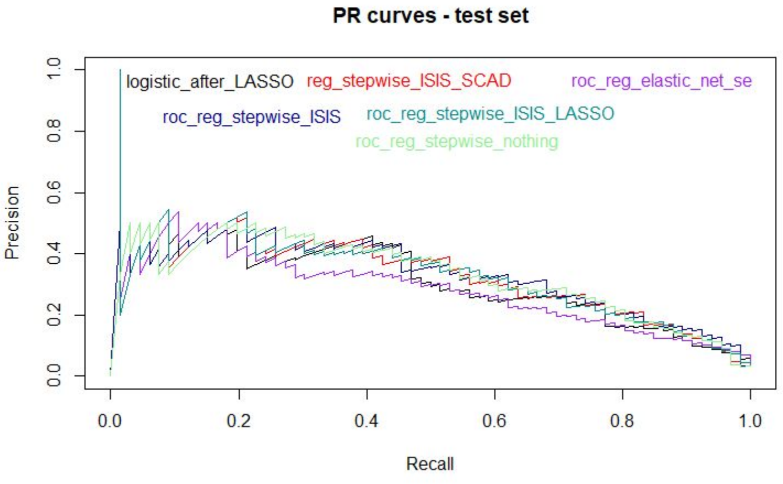
- Since the algorithms are quite computationally expensive, we choose to do only the stepwise selection algorithm (not the best-subset one) after having applied to the data some preliminary screening algorithms (i.e., ISIS + SCAD, ISIS + LASSO and only ISIS)

- The stepwise regression performed after ISIS + SCAD and ISIS + LASSO gives us 7 features (same number as baseline LASSO, but different ones!), while the stepwise selection after only ISIS gives us much more features (17)

Feature Selection: performances

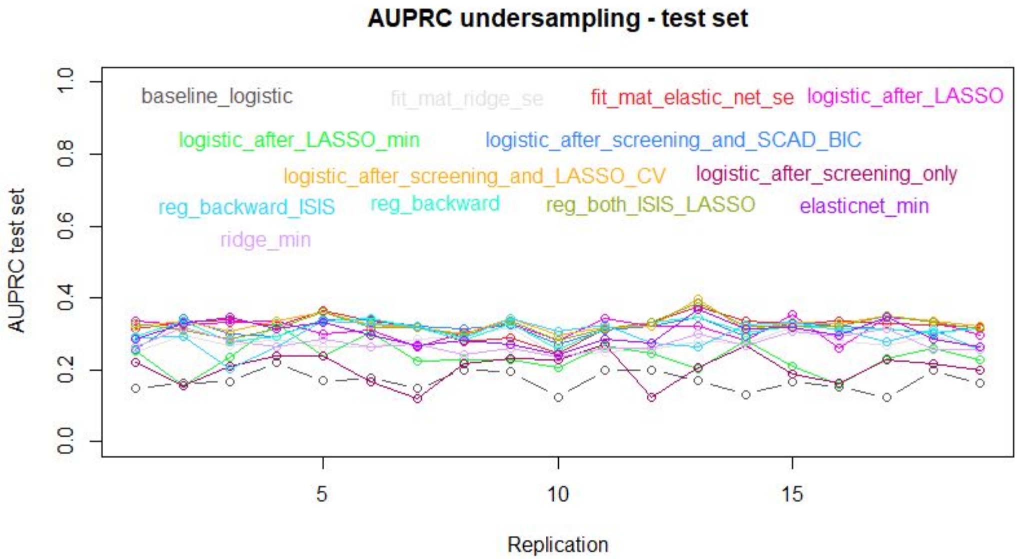
- We measure the performance of our models so far (we use the Precision - Recall curve since it is robust to imbalanced datasets)
- Models chosen by stepwise regression are better than the ones chosen by LASSO and Elastic Net, in both training set and test set!
- The best one is stepwise selection after ISIS (Area under PR curve : 0.34504 on test set, 0.42029 on training set)
- In any case, the performances of the models chosen here are all very poor! (particularly on the test set)
- This is probably because of the fact that we have a VERY IMBALANCED dataset (220 bankrupt firms over 6819!)





Undersampling

- We try to resolve this issue by performing the models seen so far on new balanced samples made by randomly undersampling the more abundant class (non-bankrupt firms) and join them with all bankrupt firms
- This is done on 19 (perfectly balanced) under-samples from the training set
- We then select the best model by looking for each model at the mean and the maximum of the Area under the PR curve for every replication



The best model here is logistic after ISIS + LASSO (mean AUPRC on test set = 0.3298, max AUPRC on test set = 0.39605)

Max performance better than the best model chosen using the original data!

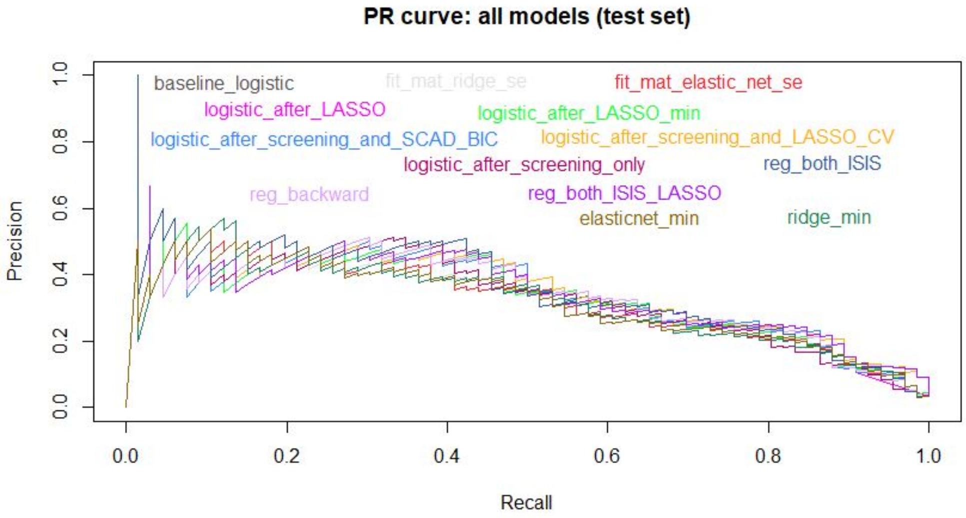
Sant'Anna
Scuola Universitaria Superiore Pisa

Oversampling

- Another approach to solve the problem of imbalanced dataset: oversample the less abundant class (bankrupt firms…)

- We did this using the SMOTE method on the training set, obtaining a new dataset of 8777 observations, and then performing all the models seen before on this new dataset

- For each model we measured the PR curve on training and test set



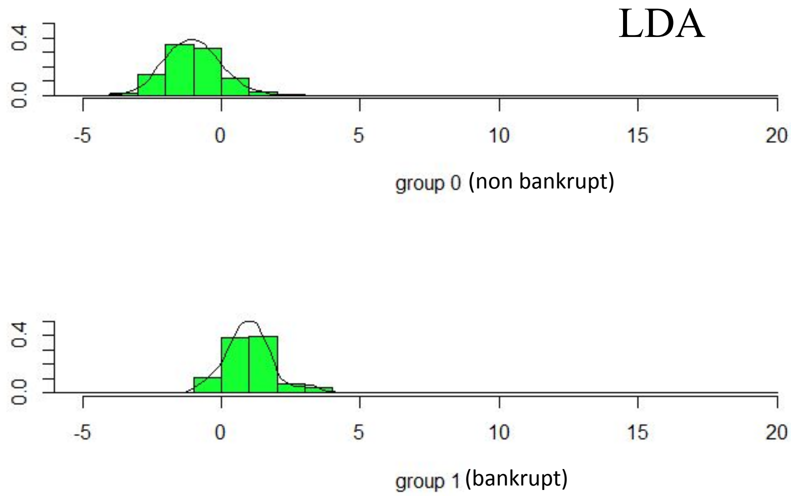
The best model chosen here is stepwise selection after ISIS (AUPRC on test set = 0.36931)

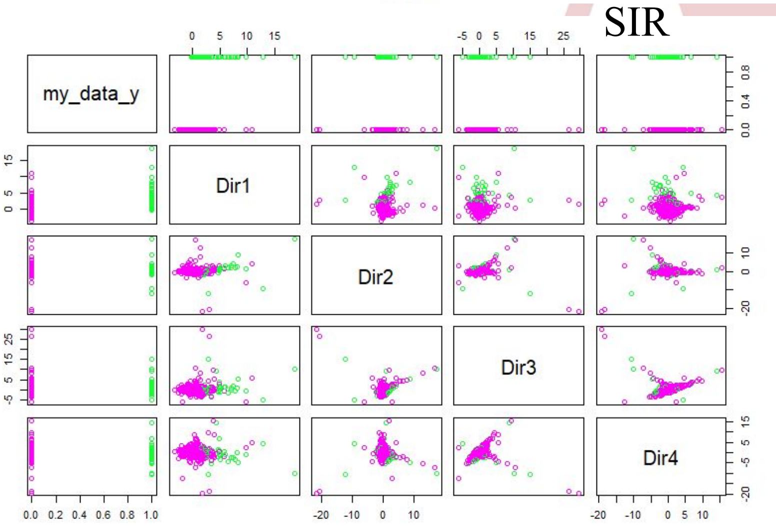
Performance better than the best model chosen using the original data!

Sant'Anna
Scuola Universitaria Superiore Pisa

Supervised Dimension Reduction

- We try here to find the best linear combination of features to explain the dependent variable; we used both LDA and SIR approaches (LDA is far more suitable since it is tailored to classification problems)
- We apply LDA both to original (training) dataset and to oversample dataset
- LDA gives us only one dimension since the problem involves a binary classification
- Even here, performances on the test set are better for the oversample application (AUPRC = 0.3765 for oversample, AUPRC = 0.299 for original data)!





Main Features

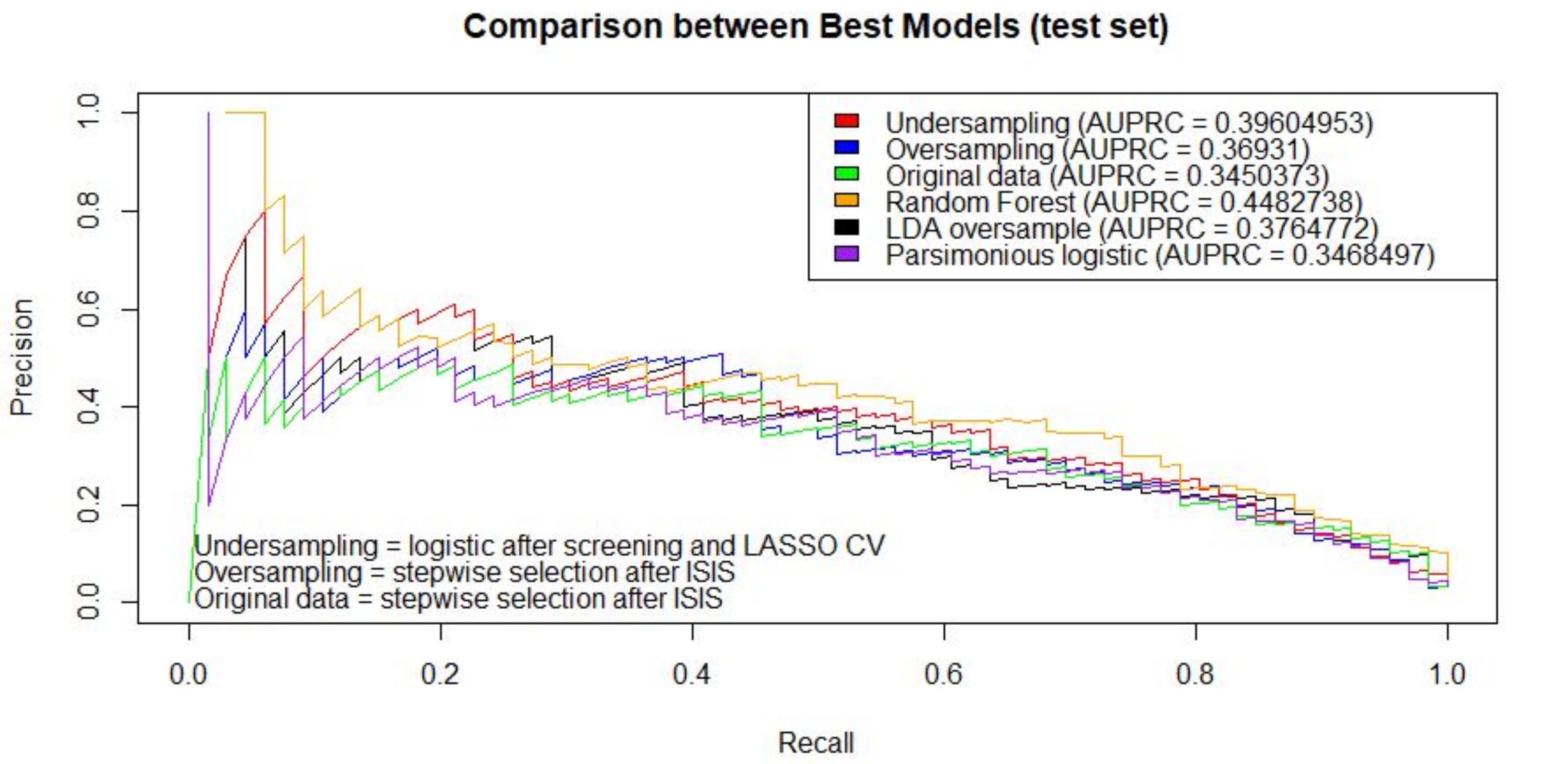
Comparing the 7 main models we noticed that only 6 features are selected in at least 4 of them!

Feature	Selected in … models	Sign	Significant?
Net.Income.to.Total.Assets	7	-	always
Current.Liability.to.Assets	7	+	always
Equity.to.Long.term.Liability	6	+	always
Cash.Total.Assets	5	-	always
Fixed.Assets.to.Assets	5	+	never
Fixed.Assets.Turnover.Frequency	4	+	always

Parsimonious Regressions

Feature	Very	Standard	Less
Net.Income.to.Total.Assets	-0.8335*	-0.9154*	-0.8800*
Current.Liability.to.Assets	0.6344*	0.5080*	0.5921*
Equity.to.Long.term.Liability		0.4128*	0.4022*
Cash.Total.Assets		-1.7365*	-1.5382*
Fixed.Assets.to.Assets			0.2244
Fixed.Assets.Turnover.Frequency			0.3289*
Performance on test set	*0.2873*	*0,3468*	*0,3447*

*significance at 1%


Comparison between Best Models (test set)
Undersampling (AUPRC = 0.39604953)
Oversampling (AUPRC = 0.36931)
Original data (AUPRC = 0.3450373)
Random Forest (AUPRC = 0.4482738)
LDA oversample (AUPRC = 0.3764772)
Parsimonious logistic (AUPRC = 0.3468497)
Undersampling = logistic after screening and LASSO CV
Oversampling = stepwise selection after ISIS
Original data = stepwise selection after ISIS
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

Conclusions and what's next?

- Undersampling and oversampling help address the imbalanced dataset problem
- ISIS+LASSO on an undersampled dataset is the best feature selection model
- 94 features are too many! By looking at just 4 of them you can get an idea of a company's bankruptcy risk better than by looking at all 94:
    - Net Income to Total Assets (the higher it is, the lower the risk) → profitability
    - Current Liability to Assets (the higher it is, the higher the risk) → solvency
    - Equity to Long term Liability (the higher it is, the higher the risk) → capital structure
    - Cash/Total Assets (the higher it is, the lower the risk) → solvency

- Can the results we obtained be extended to other contexts? (other types of companies, other countries, other years)
- How do other prediction models perform on this dataset? (Gradient boosting, neural network…)

References

- *Dataset:* Taiwanese Bankruptcy Prediction. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5004D.
- *Most methods:* James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). “An introduction to statistical learning: With applications in R.” Springer Nature.
- *Precision-Recall Plot:* Saito, T., & Rehmsmeier, M. (2015). “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets.”
- *Undersampling:* Prusa, J., Khoshgoftaar, T. M., Dittman, D. J., & Napolitano, A. (2015). “Using random undersampling to alleviate class imbalance on tweet sentiment data”. IEEE.
- *Oversampling:* N. V. Chawla, K. W. Bowyer, L. O.Hall, W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” Journal of artificial intelligence research, 321-357, 2002.

Thank you!