Topics in Statistical Learning: Final presentation

- Bernardo D'Agostino (*II ECO*)
- Matteo Ronga (*III ECO*)
- Laura Pittalis (*II ECO*)
- Pietro Carlotti (*II ECO*)

The dataset

- Taiwan Economic Journal for the years 1999–2009
- 6820 companies: 150 went bankrupt
- 95 financial ratios

Our aim

Bankrupcy prediction

Human-driven vs
Data-driven techniques

Human-driven selection

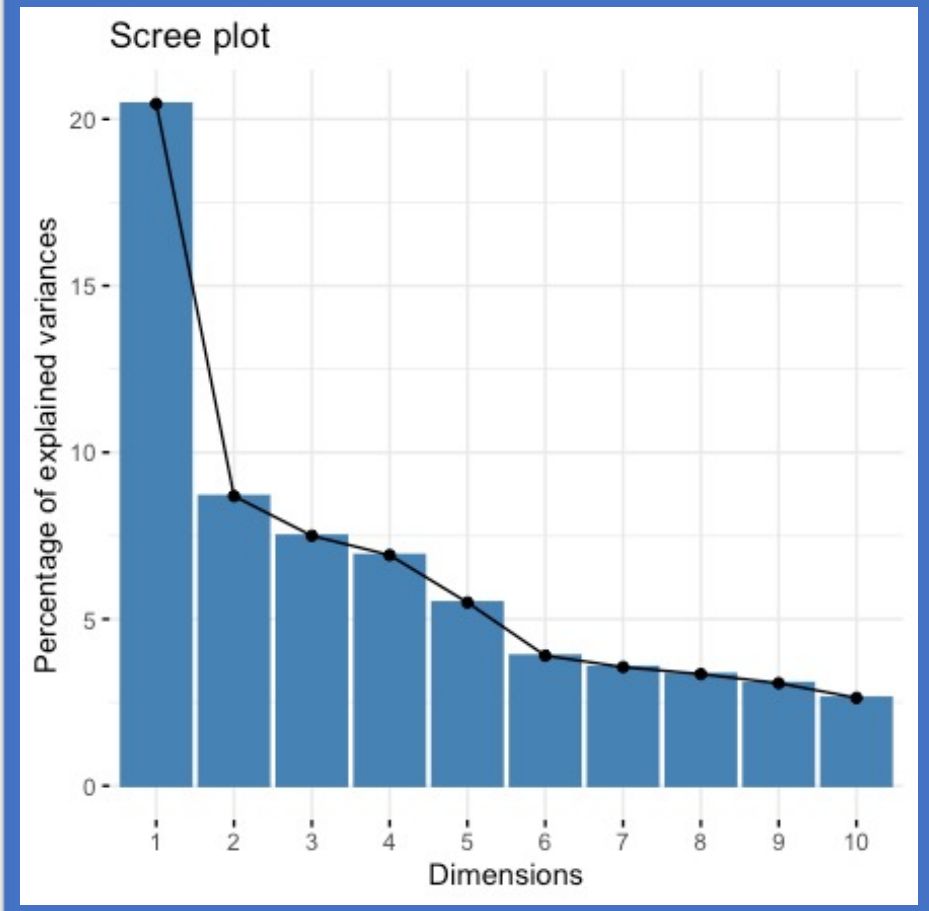
- Cash Flow / Liability
- ROA
- Income / Equity
- Current Liabilities / Assets
- Long Term Liabilities / Assets

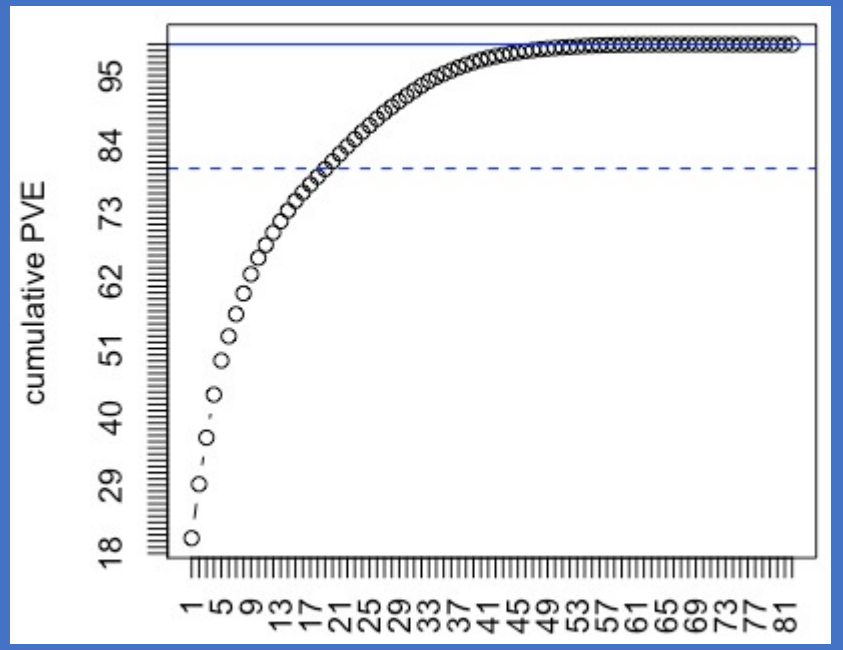
- Cash Flow / Equity
- Liability / Equity
- Interest Coverage Ratio
- Interest Rate
- Current Ratio

Preprocessing the data

- *Corrupt records* → Data cleansing
- *Normalization* → Logarithm
- *Class Imbalance Problem* → SMOTE

PRINCIPAL COMPONENT ANALYSIS

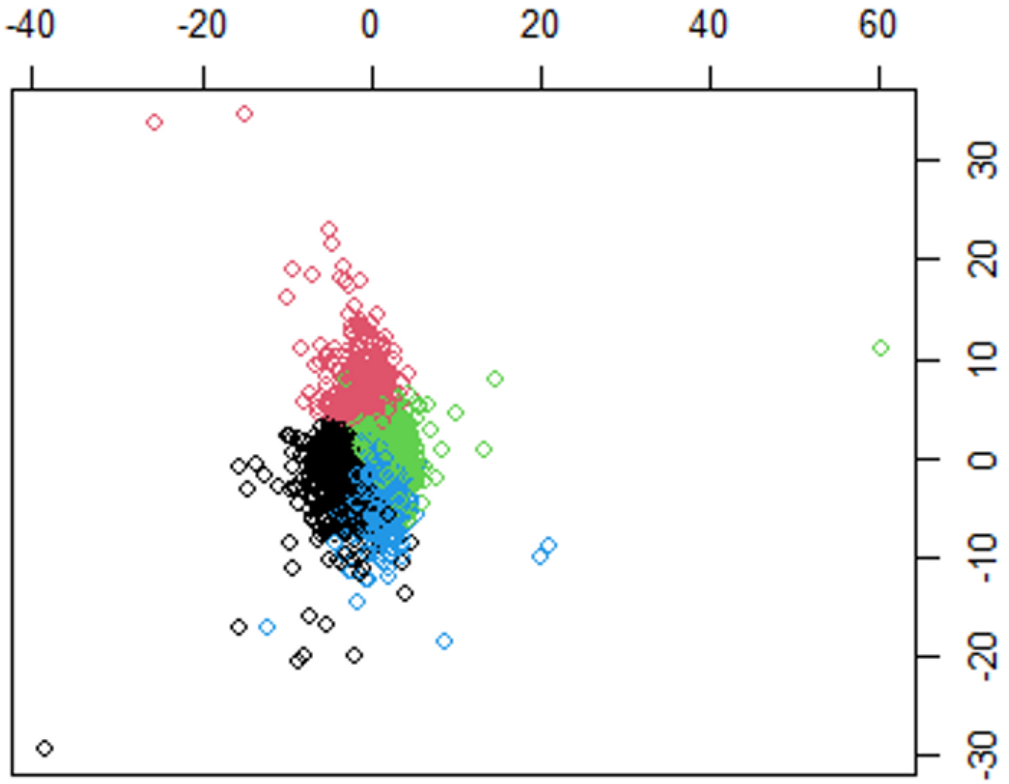


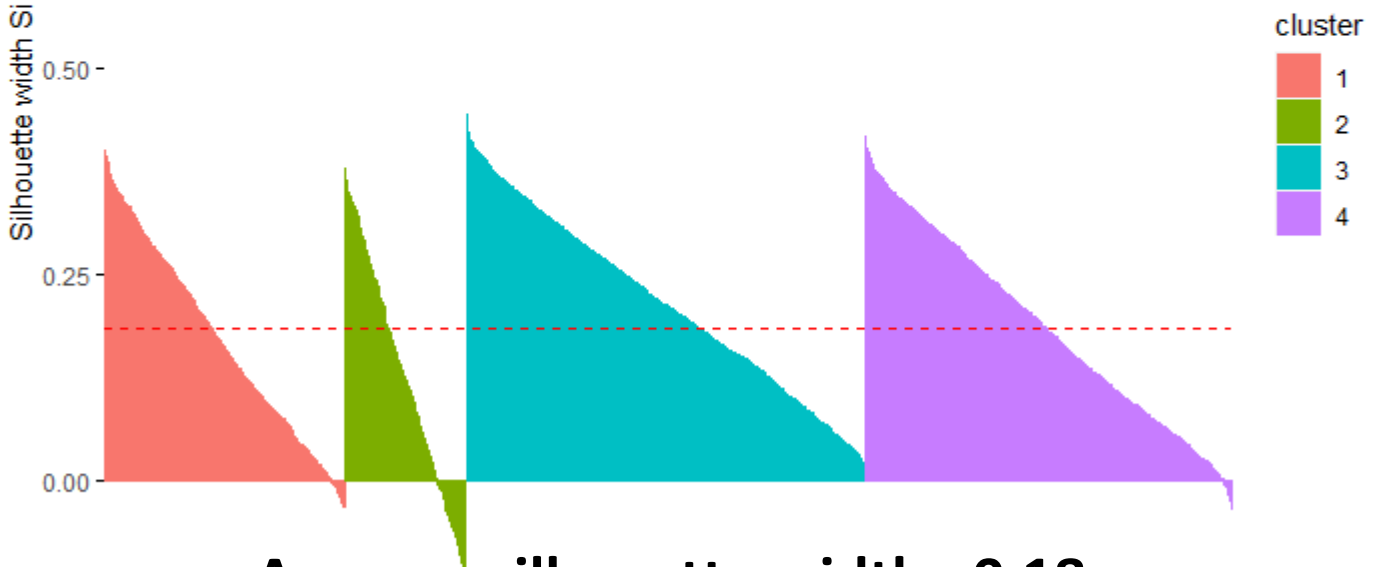


	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	1.677692e+01	2.045965e+01	20.45965
Dim.2	7.125846e+00	8.690056e+00	29.14971
Dim.3	6.150276e+00	7.500337e+00	36.65005
Dim.4	5.675413e+00	6.921235e+00	43.57128
Dim.5	4.508612e+00	5.498307e+00	49.06959
Dim.6	3.200993e+00	3.903650e+00	52.97324
Dim.7	2.920068e+00	3.561059e+00	56.53430
Dim.8	2.748191e+00	3.351452e+00	59.88575

Clustering

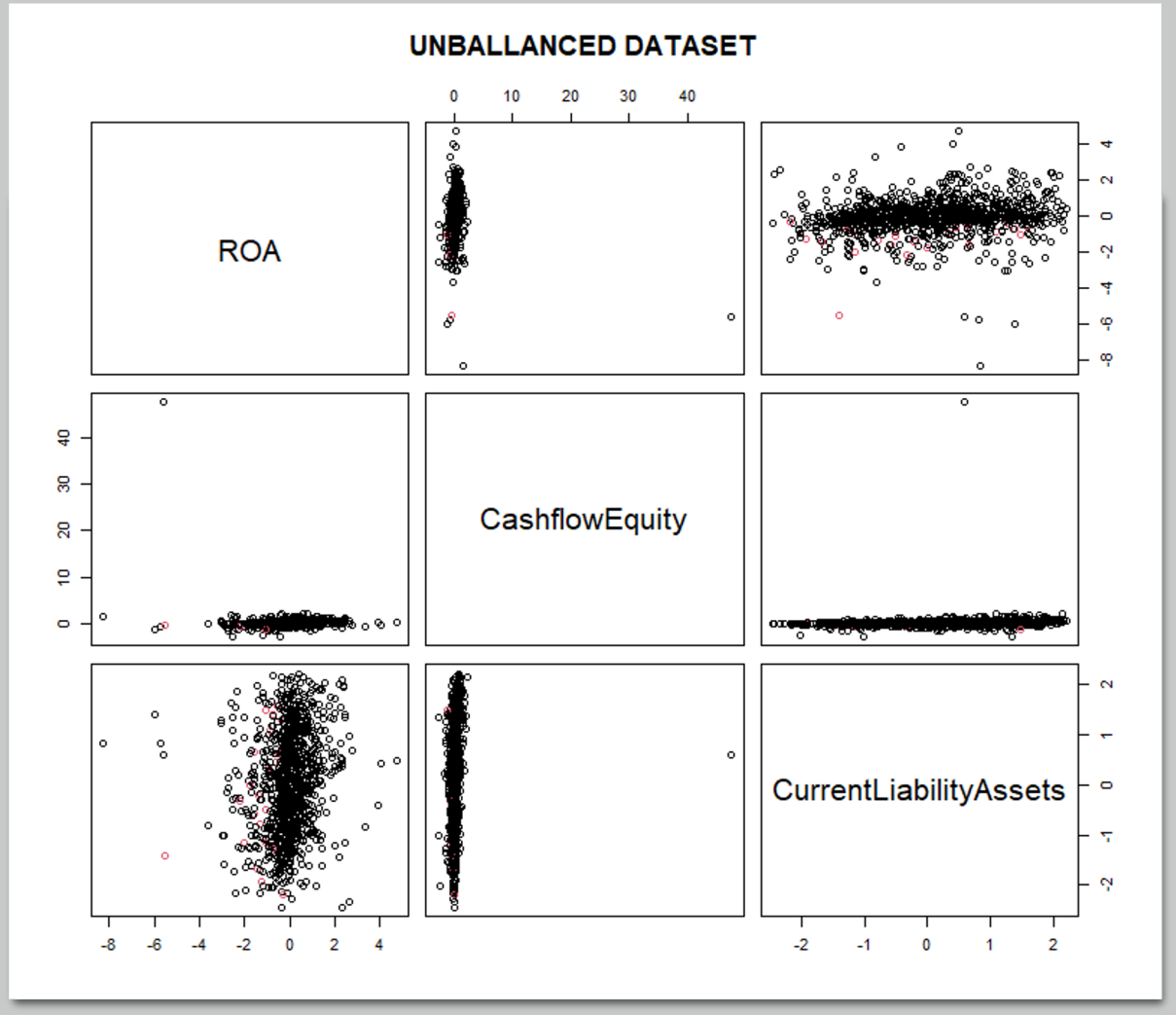
Our results were <u>NOT</u> satisfactory

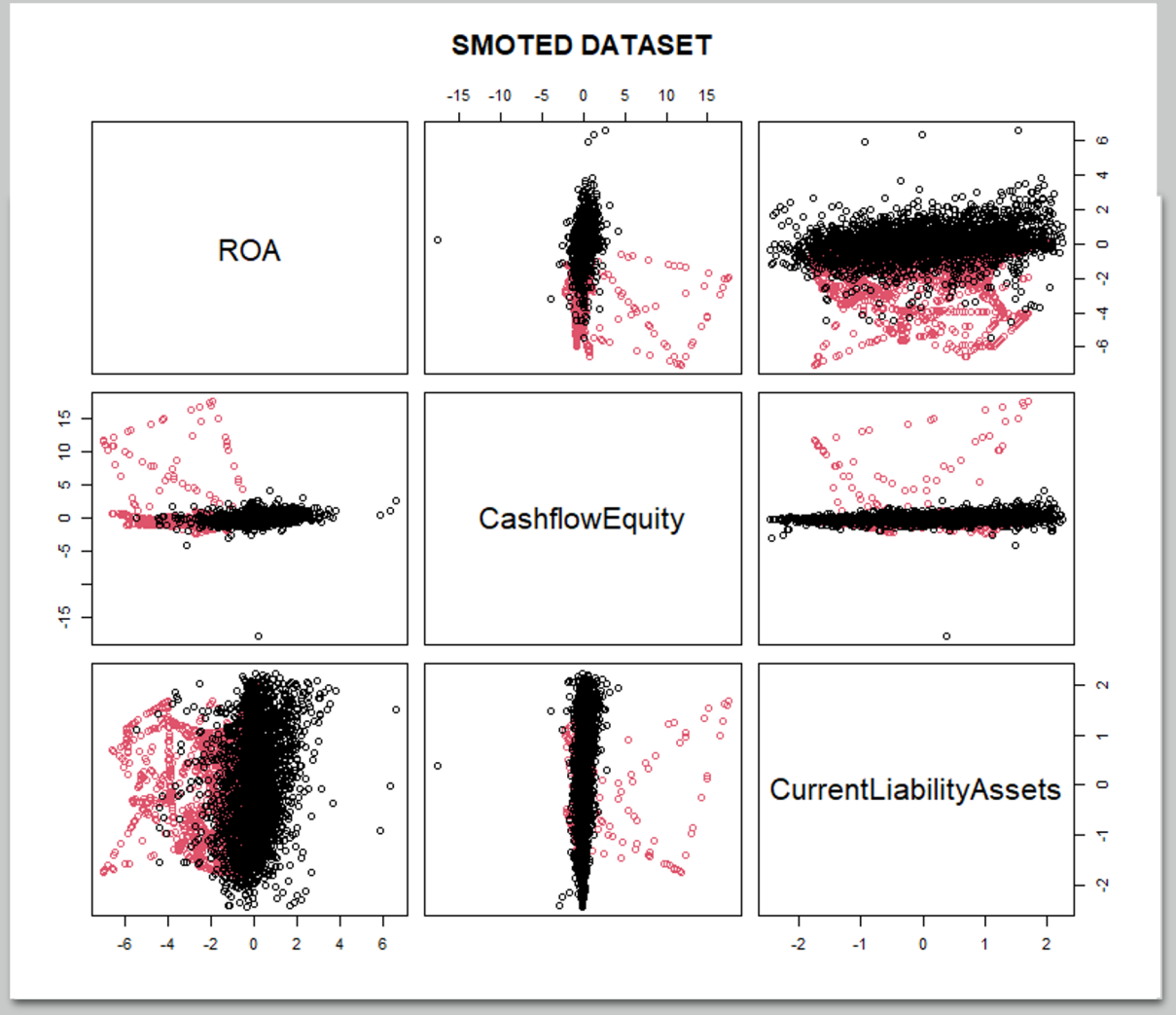




Average silhouette width: 0.18

The SMOTE algorithm





Classification techniques

- Logit → Stepwise methodology
- LDA
- QDA
- KNN

Classification techniques

LOGIT

LOGIT STEPWISE

LDA

QDA

DATA-DRIVEN

KNN

DATA-DRIVEN

AIC	
LOGIT	4315.201
LOGIT STEPWISE	4313.646

HUMAN-DRIVEN

HUMAN-DRIVEN

```
            Reference
Prediction    0   1
         0  811   1
         1  171  27
```

```
            Reference
Prediction    0   1
         0  602   0
         1  380  28
```

```
            Reference
Prediction    0   1
         0  876   6
         1  106  22
```

```
            Reference
Prediction    0   1
         0  833  11
         1  149  17
```

```
            Reference
Prediction    0   1
         0  453   2
         1  529  26
```

```
            Reference
Prediction    0   1
         0  884  13
         1   98  15
```

Cross-validation

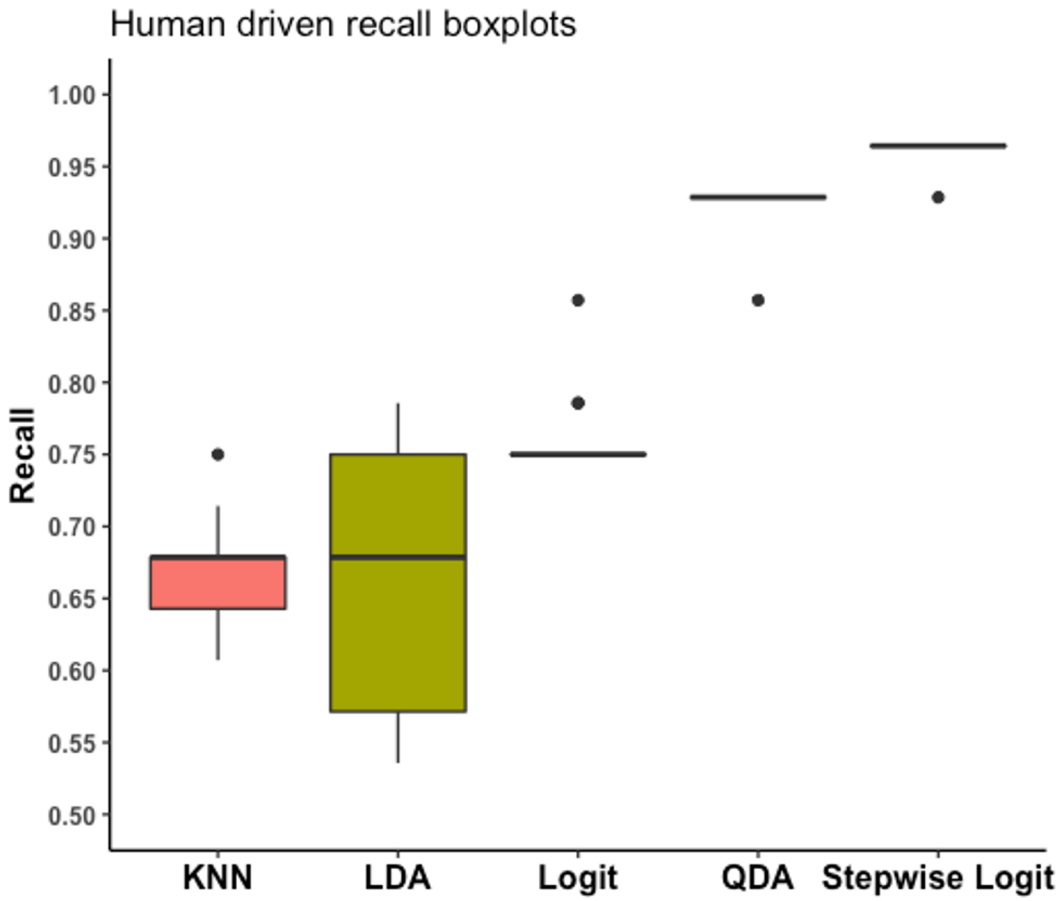
SET	NUMBER OF FOLDS	NUMBER OF OBSERVATIONS
TRAINING SET SMOTE	20 folds	291 observations each
TEST SET	1 fold	1010 observations

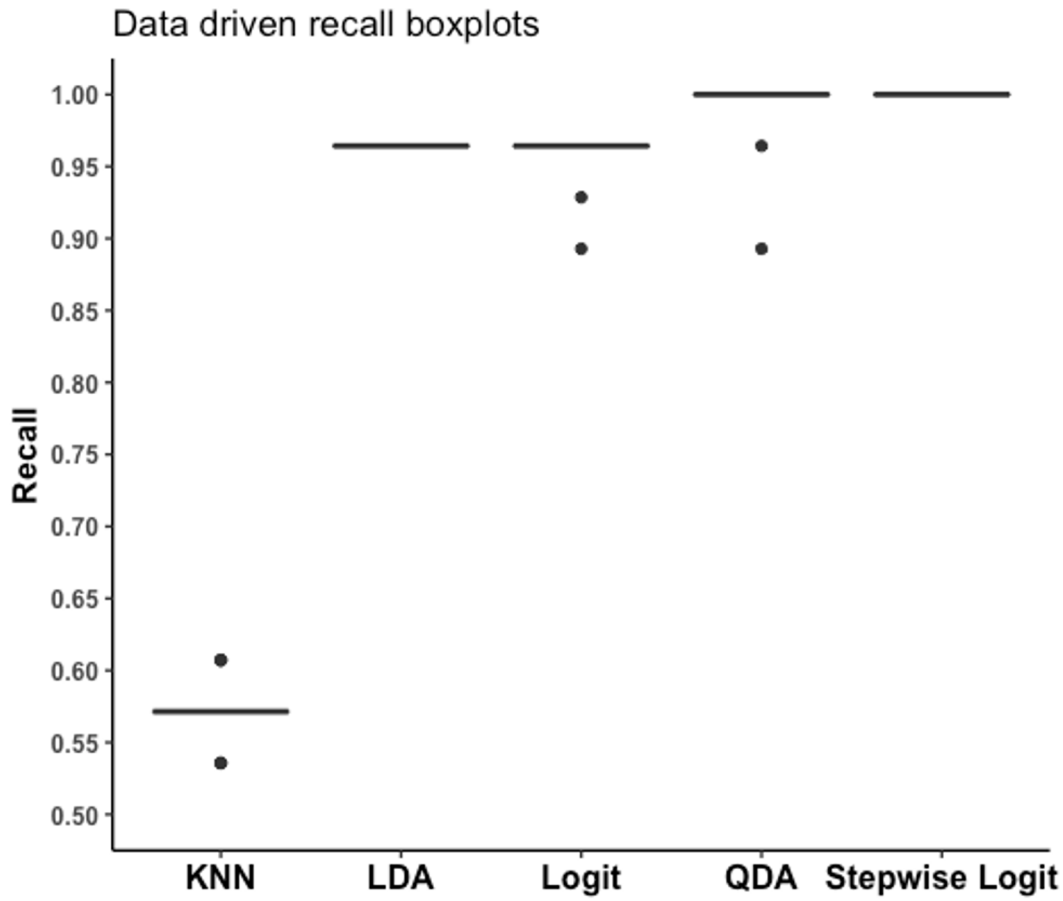
$$\mathbf{RECALL} = \frac{True Positives}{True Positives + False Negatives}$$

$$\mathbf{PRECISION} = \frac{True Positives}{True Positives + False Positives}$$

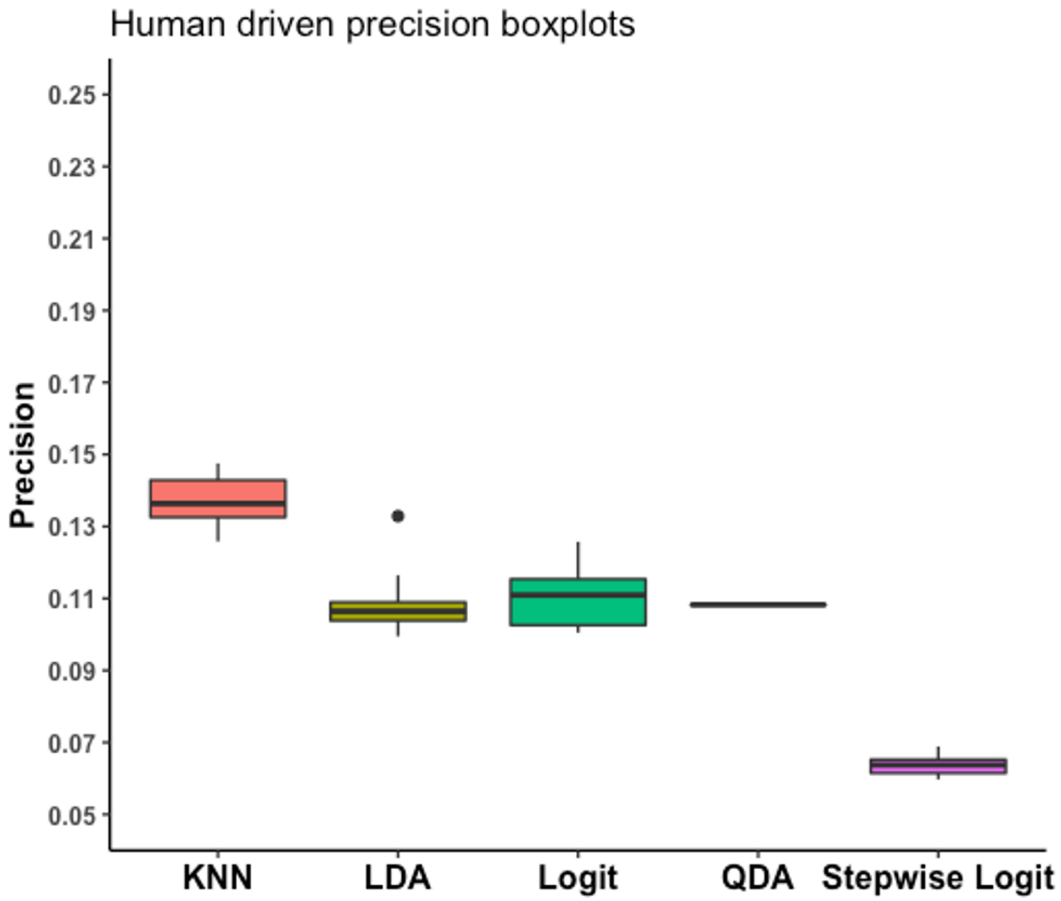
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 797   7
         1 185  21

               Accuracy : 0.8099
                 95% CI : (0.7843, 0.8337)
    No Information Rate : 0.9723
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1374

 Mcnemar's Test P-Value : <2e-16
```

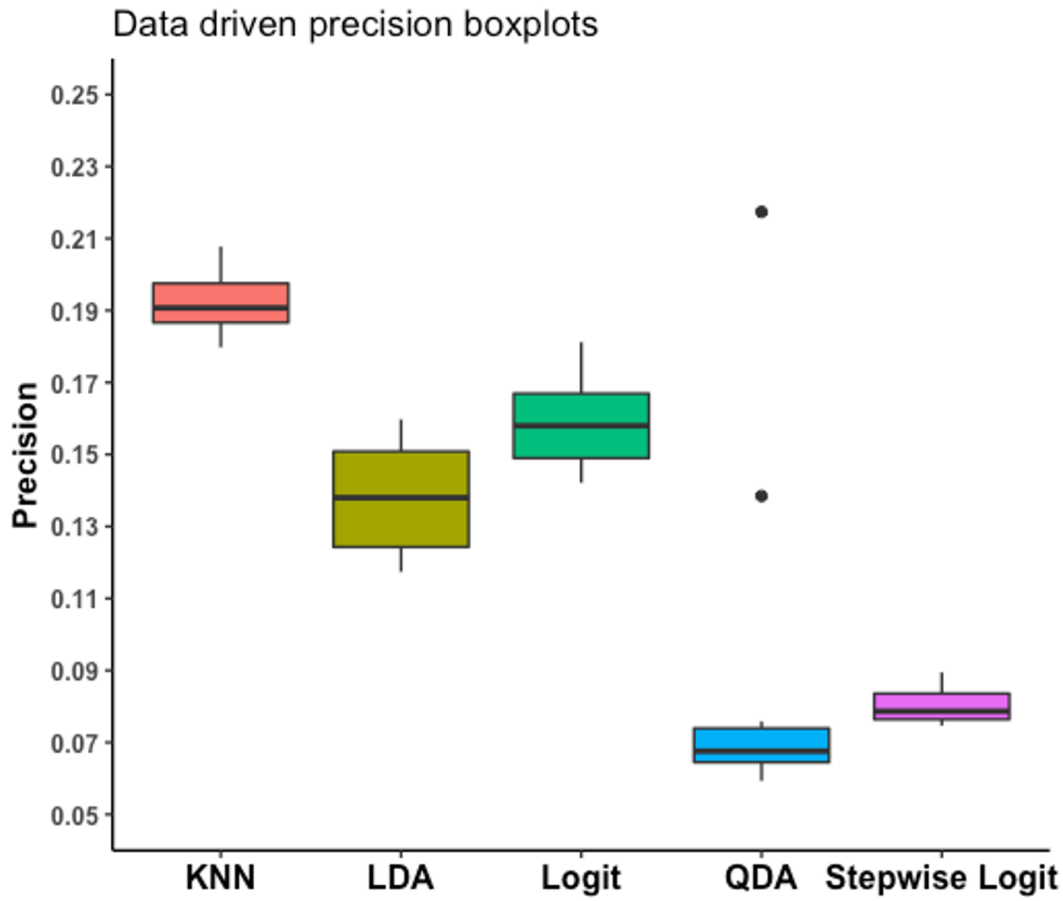
Recall boxplots





Precision boxplots





Future research paths

Experiment alternative preprocessing solutions

Check for robustness with a different dataset

Explore new classification techniques

References

- Liang D., Lub C., Tsai C. and Shiha G. (2016), 'Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study', *European Journal of Operational Research*, Research 252, pp. 561-572.

- Chawla N. V., Bowyer K. W., Hall L. O. and Kegelmeyer W. P. (2002), 'SMOTE: Synthetic Minority Over-sampling Technique', *Journal of Artificial Intelligence*, Research 16, pp. 321–357.

- James G., Witten D., Hastie T. and R. Tibshirani (2013), *An Introduction to Statistical Learning with Applications in R, (*New York: Springer Science+Business Media).

Thank you for your attention