Man vs Machine: predicting bankruptcies in the Taiwanese stock market

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

9 April 2021

Introduction

Is there a way to predict which firms are more prone to going bankrupt? When it comes to predicting bankruptcies, do machines perform better than humans? These are the questions this report seeks to find an answer to. Our project's main goal is to develop a model to use in order to predict which firms are about to go bankrupt and to discover if a data-driven selection of predictive variables can improve models elaborated on the basis of our prior knowledge of accounting and business administration. In order to do so, starting from an original dataset containing information about approximately 7,000 firms, we first construct two distinct datasets: one that includes only those financial indicators that are commonly monitored for our purposes and another one which includes only the first principal components obtained through a Principal Component Analysis of the complete dataset; then, we apply a series of classification techniques to these two dataset with the goal of identifying what are the firms that have actually gone bankrupt. Finally, we cross-validate our models to assess their ability to track down bankrupt firms and ultimately decide which models has proven to be the best.

This report is organized as follows: in section 1 we describe our original dataset into more detail and we illustrate the preprocessing procedures we had to implement before undertaking any classification technique in order to make the data tractable. In section 2 we explain how we selected relevant variables for both the human-driven and the data-driven datasets. In section

3 we present the results of an attempt to find out whether there was a way to subgroup our observations into natural clusters. In section 4 we show the different model performances obtained through a Validation Set approach. Finally, is section 5 we display the final results of a k-fold cross-validation on the classification techniques we employed.

1 The dataset

The original dataset we adopted for our investigations was published in 2016 by [3], who recovered information about 6820 firms which were active in the Taiwanese stock market during a period that goes from 1999 to 2009. The authors managed to collect data about 95 numeric variables which shed light on our observational units' liquidity, solvency, income, capital structure and so on and so forth. [1] Out of almost 7,000 observational units, only 115 had actually gone bankrupt by the end of 2009. This is a fact that poses quite a few problems for our classification techniques; in fact, we will shortly explain how we dealt with this issue and how we tried to solve it.

Fortunately, there are virtually no missing values within our dataset, therefore we did not really have to worry about handling NAs. Nevertheless, there were still some major problems with the original dataset as we initially recovered it; therefore, we were compelled to conduct a preliminary preprocessing of the data, so that it could be effectively used for prediction purposes.

1.1 Preprocessing the data

First of all, many of the variables contained in the dataset revealed a large level of skewness. This is a feature that can be immediately detected with a quick glimpse at Figure 1, which plots distributions for some of the most highly skewed variables. This lead us to think that some of these variables might not be normally distributed. If that were the case, this circumstance could possibly have a negative impact on the quality of our techniques, since some of them rest on the hypothesis of normally distributed features (notably, LDA and QDA). Through a common Shapiro-Wilk test, we discovered that our suspicions were well-founded, as nearly all variables were very distant from normality. Figure 2 shows the distribution of the p-values obtained

[1] For more information, a detailed list of variables is displayed by the authors themselves in the original paper [3]. We chose not to report this table, as not all these variables played a central role in our analysis.

through Shapiro-Wilk tests carried out on all variables. In order to tackle this issue, we considered computing the logarithm of the original variables to partially reduce their skewness. Unfortunately, the outcome we obtained was not as good as we expected. There was an underlying problem which we had not considered: the available data was not actually the raw data, but rather the output of a preliminary standardization, performed by the authors themselves, which forced all entries to range from 0 to 1. This is the formula that was used:

$$\forall x \in F, standardize(x) = \frac{x - min(F)}{max(F) - min(F)}$$

where F is a set of one specific feature (i.e. variable), x is the feature value, and $max(F)$ and $min(F)$ are the maximum and minimum values of the specific feature set, respectively. Of course, calculating the logarithm for such low values produces new measures that tend towards minus infinity, which are interpreted by our statistical software as missing values. For these reasons, we eventually decided to discard logarithms and hold to our initial standardized values.

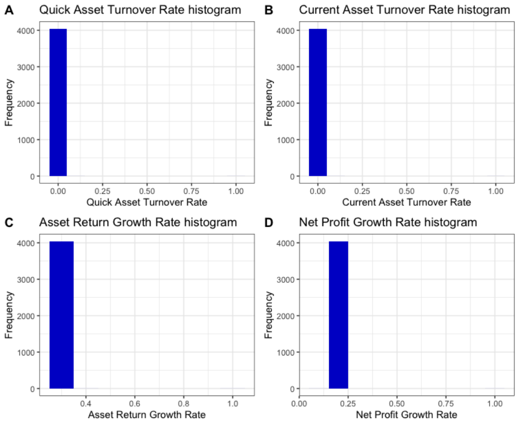


Figure 1: Histograms for highly skewed variables

Secondly, the dataset contained numerous corrupt records. As we previously said, following to the standardization, all entries had to fall within the interval that goes from 0 to 1, or, at least, this is what should have happened. Yet, there were some of the 95 original variables whose order of magnitude was much larger than 1 (some even reached 10^{17}!). The presence of these problematic variables could severely distort the results of the techniques we used, starting from the PCA; hence, we acted some form of data cleansing by

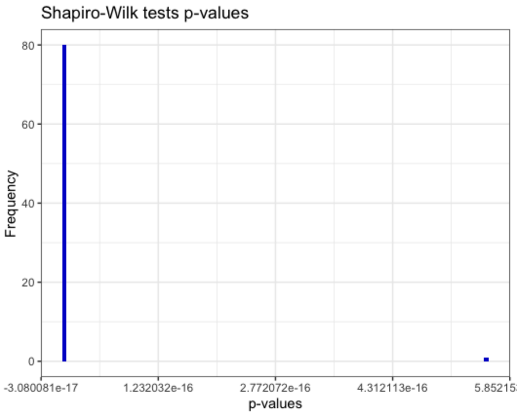


Figure 2: Histograms for Shapiro-Wilk tests p-values

simply removing the problematic variables. As a consequence, our dataset reduced its dimensionality from 95 to 82. We inevitably lost some information, but the number of variables is still very high.

Finally, we had to face the so called Class Imbalance Problem: our aim is to make predictions about bankruptcies based on the information contained in the dataset; however, the number of firms that went bankrupt (115) is very small if compared with the total number of observations (6820). Clearly, estimates resulting from any classification technique employed on the original dataset would be affected by this evident disproportion between the number of observational units that went bankrupt and those that did not. We found a solution to our problem thanks to the Synthetic Minority Oversampling Technique, which will be henceforth referred to as the SMOTE.

The SMOTE is an algorithm that takes 2 parameters: the first one is the number of artificial data points that we intend to create; the second one is the number of k nearest neighbours that it will consider for creating the artificial data points. We decided to create as many artificial data points as were needed to have an equal number of data points for both classes, and we set our knn parameter to 4.

First of all, the algorithm randomly chooses an observation in the underrepresented class and finds its k-nearest neighbours, then it randomly chooses from this neighbours a point to pair it with. Secondly it traces a line between the two points in the feature space and, with the help of a stochastic mechanism, it randomly selects the values for each feature of the newly created artificial data point inside the line traced between the two original points.

This trick of creating the artificial data point starting from two neighbouring points serves as a disruptor, as an alternative to adding a random deviation in the process of creating artificial data points to balance the dataset. Figure 3 shows a comparison between the SMOTED dataset and the unbalanced one.

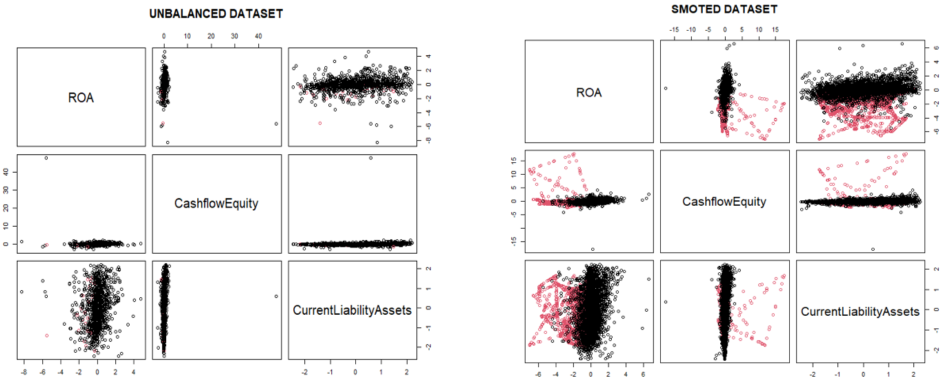


Figure 3: Red points represent bankrupt firms, whereas black points represent non-bankrupt firms

2 Variable selection

2.1 Human-driven selection of variables

In order to carry out a comparison between human-driven and data-driven techniques, we had to choose fitting criteria to select relevant variables. As far as the human-driven selection is concerned, we chose to consult two professors who teach accounting and business administration ad Scuola Superiore Sant'Anna [2]. Based on their suggestions and on our prior knowledge of these subjects, we managed to compile a list of variables which may be significant when assessing the likelihood that a firm goes bankrupt. Table 1 lists the 10 variables we selected. One may argue that the indicators we chose are not really relevant for our purposes, but we beg to differ, as they take into account all the most crucial dimensions of a firm's existence: income, liquidity, solvency and capital structure. Anyway, this is just a first attempt to develop

[2] We do not intend to divulge their names for privacy reasons.

Variable	**Variable**
Return On Assets	*Cost of interest-bearing debt*
Current Ratio	*Current Liabilities / Assets*
Long-term Liabilities / Assets	*Cash Flow / Liabilities*
Cash Flow / Equity	*Income / Equity*
Liability / Equity	*Interest Coverage Ratio*

Table 1: Human-driven selection of relevant variables

our model. We leave the task to further explore this issue for future research.

2.2 Data-driven selection of variables

Instead, for our data driven selection of variables, we relied on a PCA conducted on the SMOTED dataset. Just by looking at the resulting scree plot shown in Figure 4, it is clear that there is not a sharp elbow in the percentage of variance explained by each principal component and, actually, each principal component explains quite a little proportion of total variance. Figure 5 plots the cumulative percentage of variance explained by each principal component. We tried setting our threshold for CPVE at 80%, but in this case we needed up to 20 principal components to reach that level; so, we faced the trade-off between CVPE and model parsimony, and eventually we chose to consider only the first 8 principal components, which explain approximately 60% of total variance within the dataset.

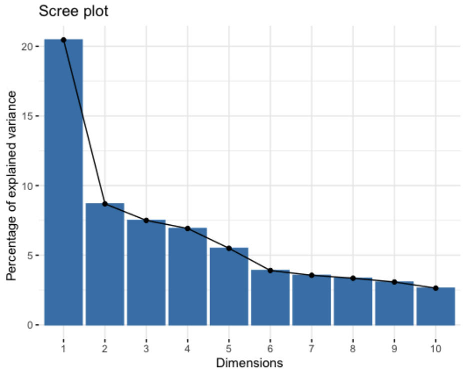


Figure 4: PCA scree plot

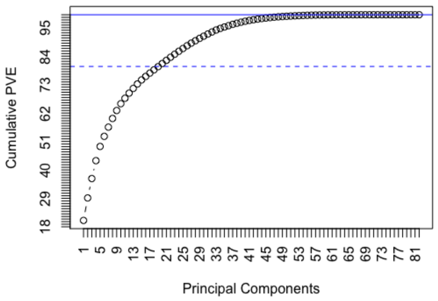


Figure 5: Cumulative PVE

3 Unsupervised learning techniques

3.1 Clustering

Before trying out any classification technique on our data, we wondered if it were possible to extract some ”natural” clusters from our observations. Could it be possible that those firms that went bankrupt already stand out from the rest? We tried running both a k-means and a hierarchical clustering, but the outcome was not nearly as good as we hoped. As it can be seen from the Figure 7, the average silhouette width is quite small. Nevertheless, we discovered that through the k-means algorithm it was possible to trace a consistent four-fold segmentation of the data.

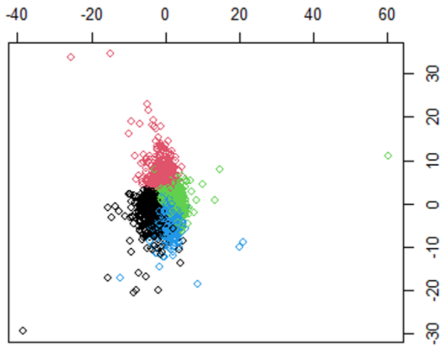


Figure 6: The 4 clusters found by the k-Means algorithm

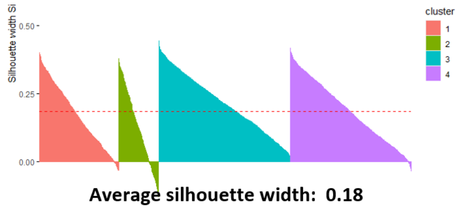


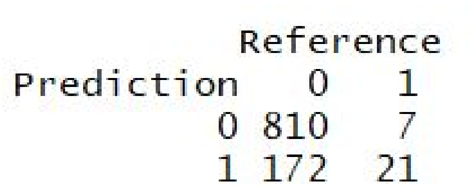
Figure 7: Silhouette widths for each cluster

4 Supervised learning techniques

4.1 Classification

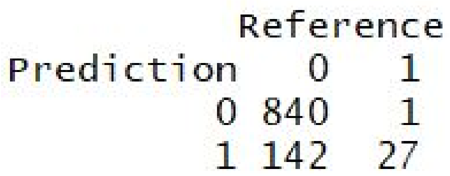
In this section, we present the outcome of the classification techniques we have experimented on our dataset. Indeed, it is possible to compare the different results of many classification methods, which are statistical procedures of supervised learning, on human-selected variables and on data-selected ones.

First of all, in accordance with the validation set approach, we have divided our rebalanced datasets, both human-driven and data-driven, into training and test set, containing 1010 and 5820 observations respectively. Then, we have implemented a logit regression on the ten variables of the human-driven dataset and another logit regression on the eight principal components of data-driven dataset. In addition to this, we have implemented the stepwise methodology on the logit regressions based on the AIC, the Akaike Information Criterion, which is a measure of fitting model performance founded on information theory. Although we have obtained high values of accuracy in the logit regressions, as it is reported in the following figures, these results do not provide us interesting proofs to analyse the difference performance of data-driven and human-driven selection.



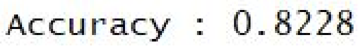
```
          Reference
Prediction   0   1
         0 810   7
         1 172  21
```

Figure 8: Logit Human-driven Confusion Matrix



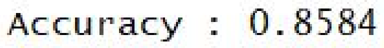
```
          Reference
Prediction   0   1
         0 840   1
         1 142  27
```

Figure 9: Logit Data-driven Confusion Matrix



```
Accuracy : 0.8228
```

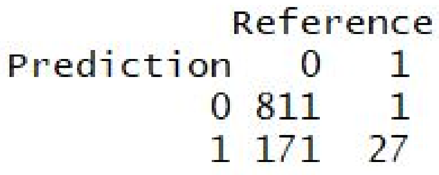
Figure 10: Accuracy Logit Human-driven



```
Accuracy : 0.8584
```

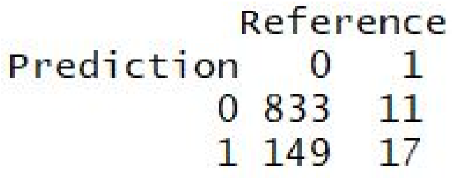
Figure 11: Accuracy Logit Data-driven

Moreover, we have implemented linear discriminant analysis and quadratic discriminant analysis on our datasets. It is essential to highlight, as discussed in detail in the previous sections, that our observations were already standardized. LDA and QDA have been implemented on the eight principal components of data-driven dataset and on the ten variables of human-driven dataset. Then, we have compared the results in terms of recall (or sensitivity) and precision, classification measures that respectively stand for the ratio of true positives and true positive plus false negative and the ratio of true positive and true positive plus false positive. As it is reported in the confusion matrixes below, these results are particularly useful to compare the performance of our different dataset. In fact, data-driven selection determines higher values of recall and precision than the ones obtained through the human-driven selection. Nevertheless, the use of cross-validation, in the next section, will strength these concepts.



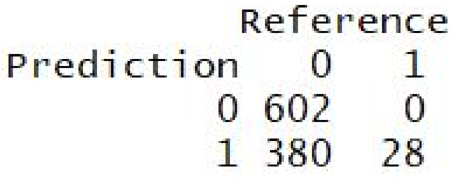
```
          Reference
Prediction   0   1
         0 811   1
         1 171  27
```

Figure 12: LDA Data Driven



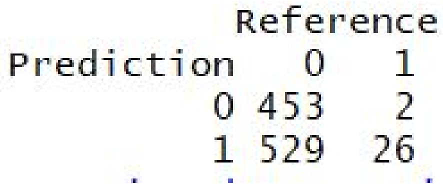
```
          Reference
Prediction   0   1
         0 833  11
         1 149  17
```

Figure 13: LDA Human Driven



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

Figure 14: QDA Data Driven



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

Figure 15: QDA Human Driven

Finally, we have implemented a KNN (k-Nearest-Neighbours) analysis on the eight PC and on the ten human-driven variables of the data-driven and human-driven datasets. Then, we have cross-validated our results in order to determine the best value of the tuning parameter, which in this case is k. Although we have obtained that the highest accuracy is predicted with $k = 1$, we have compared the performance of KNN in terms of recall and precision and, as reported in the matrixes below, we have obtained further proof that a data-driven selection performs better than a human-driven selection.

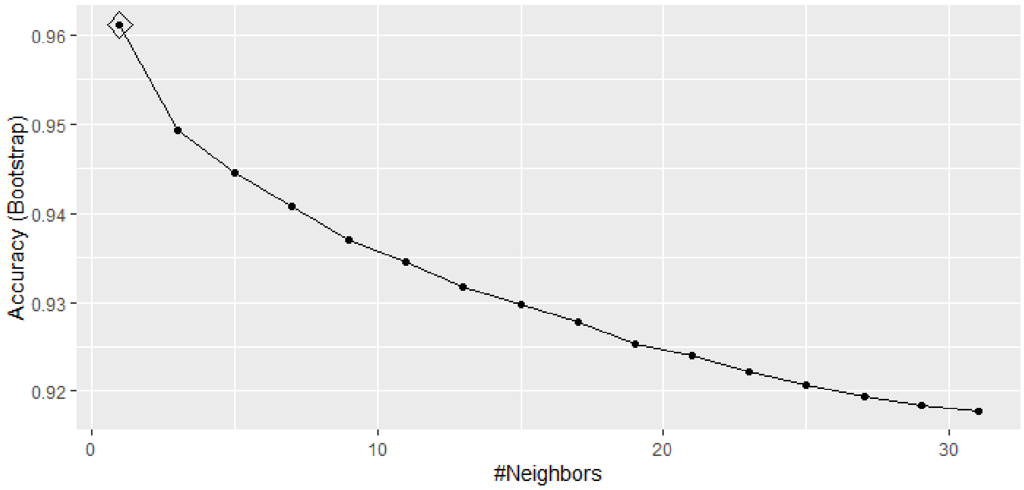


Figure 16: Knn Data driven with k=1

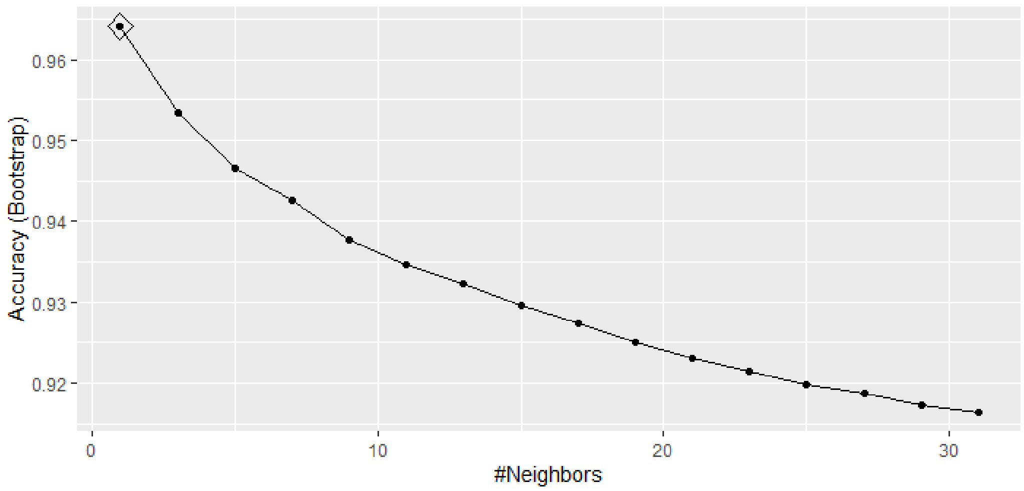
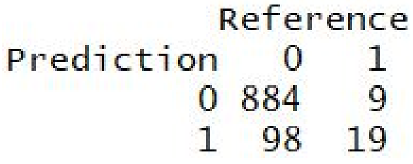
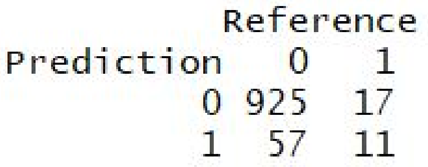


Figure 17: Knn Human Driven with k=1



```
            Reference
Prediction   0   1
         0 884   9
         1  98  19
```



```
            Reference
Prediction   0   1
         0 925  17
         1  57  11
```

Figure 18: Knn Data driven confusion matrix

Figure 19: Knn Human drive confusion matrix

4.2 Cross-validation

At this point, we have to find a way to compare the five different classification approaches and to decide which one performs better in predicting bankruptcies. In order to do that, we apply a 10-fold cross validation for the five methods. This allows us to evaluate the models on out-of-sample data and to obtain distributions of recall and precision for each technique. But how to perform cross validation on our unbalanced dataset? For ten times, the $k-1$ folds, used as training set, were treated with the SMOTE algorithm so as to solve the imbalance problem. Then, the model obtained on SMOTE data was used to predict bankruptcy on the 1-fold not-SMOTE test set. In this way, we get ten different models for each technique and just as many out-of-sample confusion matrices.

As a criterion of comparison, we chose two important measures of classification performance: recall and precision. The recall gauges how many bankruptcies were correctly predicted by the models among the total number of bankruptcies.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{1}$$

Since our aim was to find the best model in business failures prediction, this was clearly the most important measure in our analysis evaluation. In figure 20 the boxplots for the recall distributions obtained with the cross-validation are reported. The results show high values of recall for both dataset. In particular, the best performance is attained by the k-nearest neighbors and the logistic regression. In the data-driven dataset, the models resulting from these techniques have a recall close to 1: this means that almost all the bankruptcies were correctly predicted. Finally, the low result obtained by the QDA is probably due to the linear true decision boundaries.

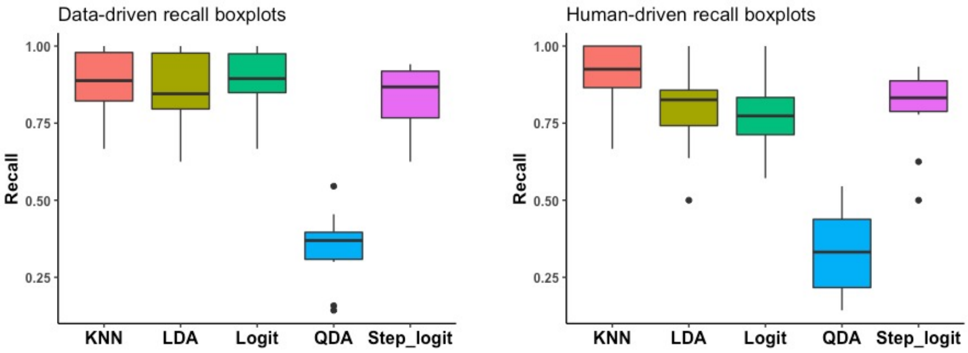


Figure 20: Recall boxplots

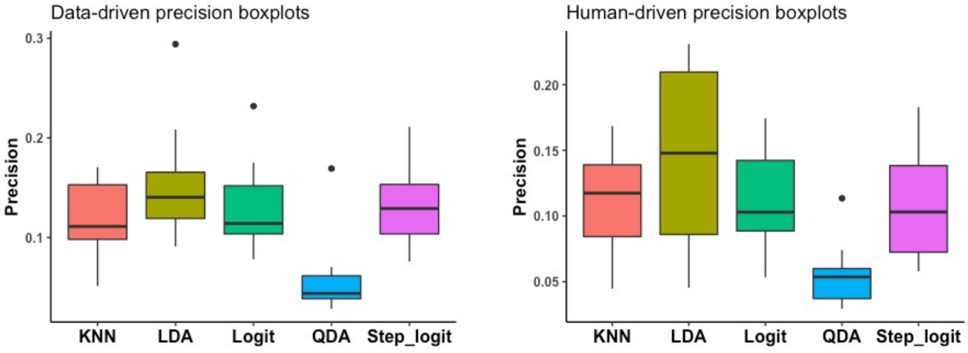


Figure 21: Precision boxplots

In figure 21 the boxplots for precision are shown. This measures how many bankruptcies were correctly predicted among the total numbers of predicted ones.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{2}$$

In this case, the overall performance is quite low: there is a fair number of false positives, scenarios predicted as bankruptcies though they are not. However, especially in our context, false positives are with no doubt less risky as prediction errors than false negatives. In addition, the precision values, unlike the case of the recall, do not show a huge difference in performance for the two dataset.

5 Conclusion

On a final note, what conclusions may we draw from our research? Surely, we managed to develop several models that can be used to detect bankruptcies, which distinguish themselves both by the variables considered and by the classification technique applied. But, can we give a definitive answer to our doubt on whether letting a machine make predictions about what businesses are about to fail based on principal components yields better results than selecting relevant variables "by intuition"? Inferring from the recall and precision boxplots obtained through the cross-validation, it can be certainly asserted that data-driven models appear to have better performances than human-driven ones, but the difference is not as evident as we initially hoped. There are numerous ways we can think of to improve our research; for instance, we could check if using a different dataset or different classification techniques we obtain better results; or else, we could try to experiment a new solution to solve the class imbalance problem, such as undersampling. Nevertheless, these are all research projects we shall leave for the future.

References

[1] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[2] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[3] Deron Liang, Chia-Chi Lu, Chih-Fong Tsai, and Guan-An Shih. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, 252(2):561–572, 2016.