Bankruptcy prediction

Georgi Emilov Hristov

1) Objective of the analysis

The objective of the analysis is to establish whether an entity is going-concern (not going to be declared bankrupt) or non-going-concern with the idea if improving financial information provided by entities. Financial information is generated through accounting records which follow accounting standards (for example, International Financial Reporting Standards (IFRS)). According to IFRS, entities must disclose whether they expect to continue operating in the next reporting period (going-concern) or not (non-going-concern). Usually, this assumption is made by management and can be subjective. The objective of this report is to create a classification model which uses financial ratios as inputs and predicts whether an entity is going-concern (GC) or not (NGC). Therefore, the main goal of the report is prediction, rather than interpretation.

2) Data

The sample is created by the author of this report. It consists of 100 entities from the construction sector in the Republic of Bulgaria. 50 of them are declared bankrupt in 2021 (NGC) and 50 of them are financially stable (GC). In other words, the working sample is balanced. For each entity 4 financial ratios are calculated: RETA (Retained Earnings/Total Assets) as a representative of efficiency; NITA (Net Income/Total Assets) also as a representative of efficiency; QACL (Quick Assets/Current Liabilities) as a representative of liquidity; and TLTA (Total Liabilities/Total Assets) as representative of indebtedness. In addition to those four input variables, the sample consists of three more features – CLASS which is going to be the target variable (it can either be GC or NGC); ID_COM_REGISTER – a unique identification number of each company in the commercial register – it cannot be used as input variable due to its high entropy; and COMPANY – which is the name of the respective company – also ignored in the analysis due to the same reason.

3) Data preprocessing

No missing values are found in the sample, so no imputing is performed. Input variables are fit to normal distribution (via np.log1p()). Where log1p returns -inf and NaN, those values are replaced with 0. Features are normalized to scale 0 to 1. Target variable is one-hot-encoded (from GC and NGC to 1 and 0). Figure 1 presents the distribution of feature before and after the preprocessing.

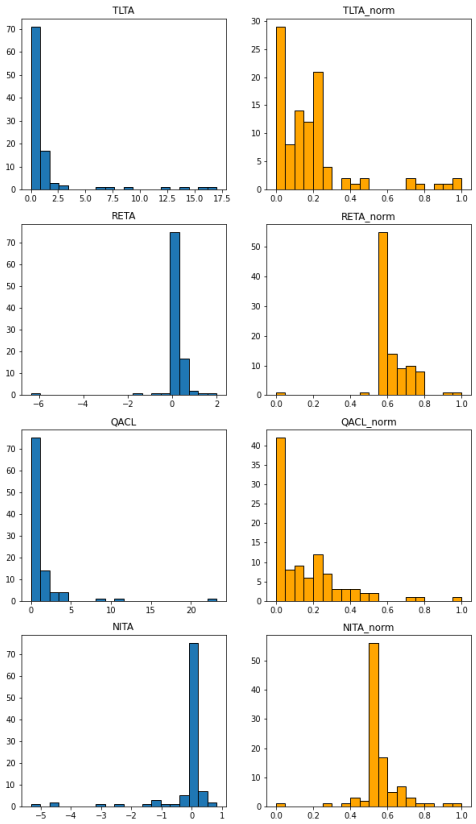

Figure 1 - Features before and after preprocessing

4) Exploratory Data Analysis

After the data has been preprocessed, EDA is performed with the idea of gaining more insight on the data that we work with. From the point of view of correlation (Figure 2), most correlated features (negatively) are QACL and NITA → therefore, companies which are illiquid and inefficient are more prone to file for bankruptcy.

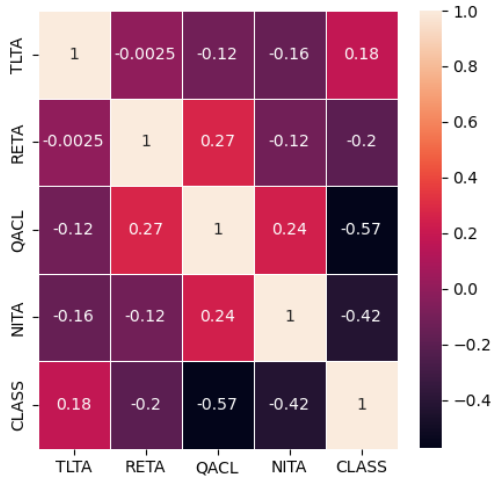

Figure 2 - Correlation matrix

5) Model selection

Three classification algorithms are tested to establish what algorithm (and with what hyperparameters) yields the best result. Models are chosen baring in mind that we work with small training sample.

a) K-Nearest-Neighbors

GridSearchCV is applied to establish what number of neighbors yields highest classification accuracy. Parameter grid consists of number of neighbors between 1 and 24. 10-fold CV is applied. Results show that highest classification accuracy (0.875) is achieved with 3 neighbors. This is backed by calculating the silhouette score for the sample, which is highest at 3 neighbors (0.718).

b) Ridge Regression

The second model tested is Ridge Regression (with different levels of regularization). The highest

classification accuracy (0.8375) is achieved at alpha=0.1.

c) Decision Tree

Once again GridSearchCV is used for tuning the hyperparameters of the algorithm (min_samples_leaf, max_depth and max_features). Highest accuracy achieved is 0.9. Which makes the decision tree the best model for the purposes of this analysis (highest accuracy). The tree is visualized on Figure 3.

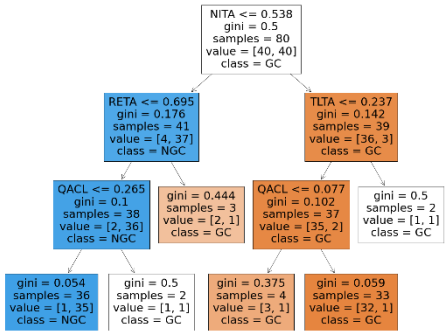

Figure 3 - Visual representation of the best performing model

6) Disadvantages of the models

kNN is well-suited for this task, because of the small training sample (with big dataset, prediction stage is usually very slow). However, as with other models, accuracy is dependent on the quality of the input data, which, in this case, can be improved as discussed later on.

Taking into account the purpose of Ridge Regression (trading variance for bias) and the objective of this analysis, it is not well-suited, because our main goal is to achieve highest accuracy possible (while avoiding overfitting, of course).

Although it performs the best, decision trees can be hugely unstable when new data is introduced into the analysis. This should be taken into account in future work on the models.

7) Suggestions for improvement of the analysis

A bigger sample can be collected with the idea of avoiding non-representative data. With more data, more complex algorithms can be used (like neural networks) with the idea of achieving deeper insight into what drives companies to bankruptcy. Surrogate models should be used, if complex models are applied and if the goal of the future analysis is interpretation.

In addition to that, more algorithms can be tested in order to achieve higher accuracy of the predictions. Different metrics of accuracy should also be taken into account.

8) Conclusion

The best performing model in the analysis is the decision tree. Such high classification accuracy is enough for the purposes of the analysis. Such model can be used as support both for managers and auditors in the process of making the decision whether a company is going-concern or not.