**Title**:

Failure prediction models for Tunisian companies: Development and comparison between multivariate discriminant analysis, logistic regression, multilayer perceptron, and support vector machines

**Dataset size:**

The dataset size severely compromises the hypothesis under testing. In statistical learning, each methodology is to be applied according to the dataset availability (size, format, panel data), the target (nominal, binary, quantitative), and type of features (quantitative, nominal, image matrix, text).

In the case of the paper, the dataset size of 212 compromises the use of multilayer perceptron to validate the hypothesis stated in section 2.5.2 (pg. 6). The performance of multilayer perceptron is to be improved with the amount of data available, and given the number of features reported, the good performance measures can even be attributed to possible overfitting.

In the case of small data sets, there are some data augmentation techniques reported by Brownlee (2019) that can be used. The data augmentation techniques usually rely on creating new and different training examples that can be useful to artificially create new training data from existing training data.

For a discussion on the sample size requirements see Hastie, et al. (2008) and Jain and Chandrasekaran (1981), and Douarrea (2019).

For a strategy on how to use machine learning in small samples see Zhang & Ling (2018) and Vabalas et al (2019).

**Overfitting diagnosis:**

There is not any misclassification in tables 7, 8, and 9. Instead of a good sign, this is indicative of the presence of Overfitting. That is to say, the presented models work well in this sample, but if we update the data with new observations, new individuals, or more recent data on the same individuals, the model will no perform well. To see more how to avoid and deal with overfitting see Koehrsen, W. 2018.

Panda (2016) compares Deep Neural Network performance for three completed different datasets: Bioinformatics datasets, Handwritten dataset, and the Japanese vowel datasets**.**All those datasets are available in the internet platforms like Kaggle, TensorFlow Datasets, and UCI Machine Learning Repository. And the criterium for choosing those datasets is that it consists of more observations than the number of instances.This way the author can provide sound evidence that the proposed algorithms and architectures generalizes to different situations: Computer Vision, Image processing, Speech recognition.

The analogy with credit risk is to provide the performance measures for completely different datasets: retail credit, microcredit in developing countries, Interbank lending market.

Some free available credit score datasets for different countries are provided in this github:

<https://github.com/JLZml/Credit-Scoring-Data-Sets>

**Train/Test Split**

Given the small sample data set available, the authors should split the sample into only two parts (train/test). Classical train/test split choices in literature are 75%-25%, 70%-30%, 80%-20%.

**Alternatives for MLP**

Given the small data set sample, I would eliminate the MLP algorithm or consider other alternatives more fit to small sample datasets, such as:

* KNN
* Decision tree
* Random Forrest
* Gradient boosting

**Features Selection**

While developing the explanatory variables the authors mention the absence of a theory of business distress, that they can rely on to build the features. They should also take advantage of the fertile possibilities of features engineering provided by the recent machine learning literature. (see Ray; 2015, and Rençberoğlu; 2019)

Since machine learning algorithms do not rely on a lack of multicollinearity assumption, practitioners use transformations on available features, such as pair-wise difference, quadratic and square root transformations of quantitative variables, and using encodings of categorical variables. Keep in mind that when you have as many features as possible then your observations, again the overfitting can be an issue.

**Software tool**:

I found the software choice is problematic since SPSS is one of the least cited tools in machine learning literature and does not carry many of the diagnosis tools commonly used in machine learning literature. I would suggest the authors use sci-kit-learn python, with is a completely free tool, and widely used in this field. (see Garreta and Moncecchi, 2013).

**References**:

Jain and Chandrasekaran. 39 Dimensionality and sample size considerations in pattern recognition practice, 1982. <https://doi.org/10.1016/S0169-7161(82)02042-2>

Jason Brownlee. How to Use Test-Time Augmentation to Make Better Predictions, 2019 in Deep Learning for Computer Vision. <https://machinelearningmastery.com/how-to-use-test-time-augmentation-to-improve-model-performance-for-image-classification/>

Douarrea, et al. Novel data augmentation strategies to boost supervised segmentation of plant disease images, 2019. Computers and Electronics in Agriculture. <https://hal.archives-ouvertes.fr/hal-02330900>

Vabalas, A., et al. Machine learning algorithm validation with a limited sample size, 2019. Plos One. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0224365>

Trevor Hastie, Robert Tibshirani and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics). 2008. <https://web.stanford.edu/~hastie/Papers/ESLII.pdf>

Koehrsen, W. 2018. Overfitting vs. Underfitting: A Complete Example. Towards data Science, 2018. <https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765>

Ray, S. Feature Engineering: How to transform variables and create new ones? Analytics Vidhya. 2015-03-12. http://www.analyticsvidhya.com/blog/2015/03/feature-engineering-variable-transformation-creation

Emre Rençberoğlu, 2019. Fundamental Techniques of Feature Engineering for Machine Learning. All required methods for comprehensive data preprocessing with Pandas examples. <https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114>

Garreta, R; Moncecchi, G. Learning scikit-learn: Machine Learning in Python, 2013. Packt Publishing. <https://www.amazon.com/Learning-scikit-learn-Machine-Python/dp/1783281936>

Ying Zhang & Chen Ling. A strategy to apply machine learning to small datasets in materials science. NPJ Computational Materials. V. 4, nº 25, 2018. <https://www.nature.com/articles/s41524-018-0081-z>