Credit Risk Analysis Using Machine and Deep Learning Models

Peter Martey ADDO
Agence française de développement (AFD)
coauthors: D. Guegan (UP1,LabEx ReFi) - B. Hassani (UP1)

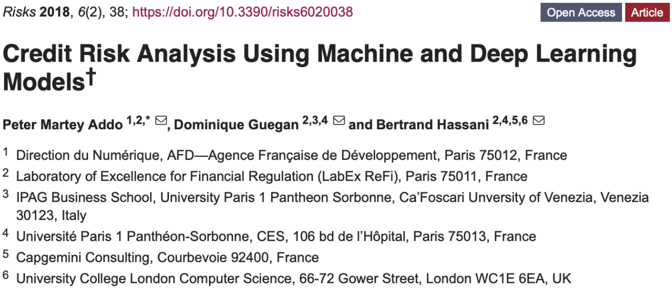
Workshop: "Big Data & Artificial Intelligence : Risks, Challenges and Applications."
Venue: ACPR

March 21, 2019

Disclaimer

The opinions, ideas and approaches expressed or presented are those of the authors and do not necessarily reflect any past or future Agence française de développement (AFD) positions. As a result, AFD cannot be held responsible for them.

Access Article & Source Codes



Risks **2018**, *6*(2), 38; https://doi.org/10.3390/risks6020038 Open Access Article

Credit Risk Analysis Using Machine and Deep Learning Models†

Peter Martey Addo [1,2,*] ✉, **Dominique Guegan** [2,3,4] ✉ and **Bertrand Hassani** [2,4,5,6] ✉

[1] Direction du Numérique, AFD—Agence Française de Développement, Paris 75012, France
[2] Laboratory of Excellence for Financial Regulation (LabEx ReFi), Paris 75011, France
[3] IPAG Business School, University Paris 1 Pantheon Sorbonne, Ca'Foscari Unversity of Venezia, Venezia 30123, Italy
[4] Université Paris 1 Panthéon-Sorbonne, CES, 106 bd de l'Hôpital, Paris 75013, France
[5] Capgemini Consulting, Courbevoie 92400, France
[6] University College London Computer Science, 66-72 Gower Street, London WC1E 6EA, UK

- Open Access: https://doi.org/10.3390/risks6020038
- Source Codes on Github: https://github.com/brainy749/CreditRiskPaper

Introduction

OUR RISK MANAGEMENT SOFTWARE SAYS YOUR IDEA IS TOO RISKY.
www.dilbert.com
scottadams@aol.com
TRY REDUCING ONE OF THE INPUTS.
WHICH ONE?
HONESTY.
I JUST THREW UP IN MY MOUTH.

- Credit risk predictions, monitoring, model reliability and effective loan processing are key to decision-making and transparency.
- In this work, we build binary classifiers based on machine and deep learning models on real data in predicting loan default probability.
- The top 10 important features from these models are selected and then used in the modeling process to test the stability of binary classifiers by comparing their performance on separate data.

Data

- The information set contains 117019 credit records over a year, each of them representing either a default or not a default (binary value) of an enterprise when they ask for a loan from a bank.
- Default and good health are characterised by the same 235 labelled variables that are directly obtained from the companies: financial statements, balance sheets, income statements and cash flows statements where the values are considered at the lowest level of granularity. These represent the credit repayment capability of the enterprise.
- In the year 2016/2017, 115288 lines represented companies in good health and 1731 represented companies in default.

Data

- 181 variables are retained after removing feature with no relevant information. Then we split the data in three subsets, considering 80% of the data (60% for the fitting and 20% for the cross validation) and then 20% of this data used for test purposes.
- We have an imbalanced data: the value 0 represents 98,5% and the value 1, 1,5%. So, the extreme events are less than 2%.
- Using the SMOTE[1] algorithm to obtain a balanced set with 46% of 0 and 53% of 1.
- We provide only results with balanced training data of the binary classes, following the method of Smote.

[1] Synthetic Minority Over-sampling Technique (SMOTE) algorithm developed by Chawla et al. (2002).

Models I

To have a benchmark for comparison and replication of results, a fix seed is set. The models have been fitted using the balanced training data set. We consider seven models.

- The Logistic regression model M1: to fit the logistic regression modeling, we use the elastic net logistic regression and regularization functions. ($\alpha = 0.5$, $\lambda = 1.9210^{-6}$).
- The random forest approach M2: we choose the number of trees $B = 120$, the stopping criterion is equal to 10^{-3} . If the process converges quicker than expected, the algorithms stops and we use a smallest number of trees.
- The gradient Boosting model M3: to fit this algorithm, we use the logistic binomial log-likelihood function: $L(y, f) = log(1 + exp(-2yf))$, $B = 120$ for classification and the stopping criterion is equal to 10^{-3}. We use a learning rate equal to 0.3.

Models II

We consider four versions of the deep learning models.

1. D1: This model considers 2 hidden layers, and 120 neurons. This number of neurons depends on the number of features and we take $2/3$ of this number.
2. D2: Three hidden layers have been used, each composed of 40 neurons and a stopping criteria equal to 10^{-3} has been added.
3. D3: 3 hidden layers with 120 neurons each have been tested. A stopping criteria equal to 10^{-3} and ℓ_1 and ℓ_2 regularization functions have been used.
4. D4: Uses the best model after considering a grid of hyper-parameters: the drop out ratio, the activation functions, the ℓ_1 and ℓ_2 regularization functions, the hidden layers. We also use a stopping criterion. The best model's parameters yields a dropout ratio of 0, $\ell_1 = 6,8.10^{-5}$, $\ell_2 = 6,4.10^{-5}$, hidden layers $= [50, 50]$, activation function is the rectifier ($f(x) = 0$ if $x < 0$, if not $f(x) = x$).

Remark

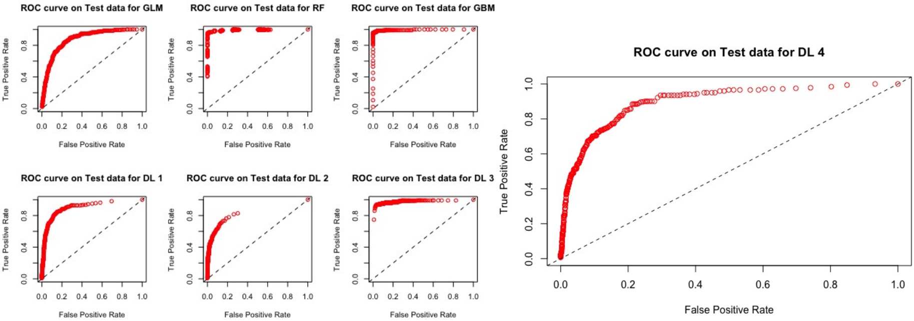
1. The regularization penalties are introduced to the model-building process to avoid over-fitting, reduce the variance of the prediction error and handle correlated predictors.
2. In addition, we include early stopping criteria to avoid this issue of overfitting.

The results with 181 features

Models	AUC	RMSE	Mean Per Class Error
M1	0.876280	0.245231	0.308058
M2	**0.993066**	0.096683	**0.061315**
M3	**0.994803**	0.044277	**0.051617**
D1	0.904914	0.114487	0.337806
D2	0.841172	0.116625	0.348201
D3	0.975266	0.323504	0.177175
D4	0.897737	0.113269	0.346406

Table: Models' Performances on test dataset with 181 variables using AUC, RMSE, and Mean Per Class Error values for the seven models.

ROC Curves


ROC curve on Test data for GLM
ROC curve on Test data for RF
ROC curve on Test data for GBM
ROC curve on Test data for DL 1
ROC curve on Test data for DL 2
ROC curve on Test data for DL 3
ROC curve on Test data for DL 4
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

The Top ten variables for each model I

- We now investigate the performance of these algorithms using only the ten variables selected by each algorithm. We do this work for the seven models.
- The whole set of features retained by the seven models correspond to 54 different variables.
- M2 selects 3 variables already provided by model M1. Model M3 selects only one variable provided by M1. The model D1 uses three variables of model M1. The model D2 selects one variable selected by model M2. Model D3 selects one variable used by model D1. The model D4 selects one variable selected by M1.

The Top ten variables for each model II

- Model M1 correspond to flow data and aggregated balance sheets (assets and liabilities). It concerns financial statement data.
- The model M2 selects financial statement data and detail financial statements (equities and debt).
- The model M3 selects detail financial statements (equities and debt).
- The model D1 selects financial statement data and a lowest level of granularity of financial statement data (long term bank debt). The models D2 and D3 select an even lower level of granularity of financial statement data (short term bank debt and leasing). The model D4 has selected the most granular data, for instance, the ratio between elements as the financial statements.

The results with Top 10 variables from Model M1

Models	AUC	RMSE	Mean Per Class Error
M1	0.638738	0.296555	0.392486
M2	0.98458	0.152238	0.091032
M3	0.975619	0.132364	0.090014
D1	0.660371	0.117126	0.430840
D2	0.707802	0.119424	0.483188
D3	0.640448	0.117151	0.437238
D4	0.661925	0.117167	0.477215

Table: Performance for the seven models using the Top 10 features from model M1 on Test dataset.

The results with Top 10 variables from Model M2

Models	AUC	RMSE	Mean Per Class Error
M1	0.595919	0.296551	0.477611
M2	0.983867	0.123936	0.199454
M3	0.983377	0.089072	0.296917
D1	0.596515	0.116444	0.479501
D2	0.553320	0.117119	0.471134
D3	0.585993	0.116545	0.468640
D4	0.622177	0.878704	0.456312

Table: Performance for the seven models using the Top 10 features from model M2 on Test dataset.

The results with Top 10 variables from Model M3

Models	AUC	RMSE	Mean Per Class Error
M1	0.667479	0.311934	0.375959
M2	0.988521	0.101909	0.156578
M3	0.992349	0.077407	0.269011
D1	0.732356	0.137137	0.481022
D2	0.701672	0.116130	0.412747
D3	0.621228	0.122152	0.468326
D4	0.726558	0.120833	0.472218

Table: Performance for the seven models using the Top 10 features from model M3 on Test dataset.

The results with Top 10 variables from Model D1

Models	AUC	RMSE	Mean Per Class Error
M1	0.669498	0.308062	0.418021
M2	0.981920	0.131938	0.185291
M3	0.981107	0.083457	0.112385
D1	0.647392	0.119056	0.438926
D2	0.667277	0.116790	0.393139
D3	0.6074986	0.116886	0.476628
D4	0.661554	0.116312	0.455485

Table: Performance for the seven models using the Top 10 features from model D1 on Test dataset.

The results with Top 10 variables from Model D2

Models	AUC	RMSE	Mean Per Class Error
M1	0.669964	0.328974	0.3917951
M2	0.989488	0.120352	0.3283074
M3	0.983411	0.088718	0.404613
D1	0.672673	0.121265	0.445209
D2	0.706265	0.118287	0.42686
D3	0.611325	0.117237	0.459714
D4	0.573700	0.116588	0.477995

Table: Performance for the seven models using the Top 10 features from model D2 on Test dataset.

The results with Top 10 variables from Model D3

Models	AUC	RMSE	Mean Per Class Error
M1	0.640431	0.459820	0.451520
M2	0.980599	0.179471	0.356908
M3	0.985183	0.112334	0.365164
D1	0.712025	0.158077	0.417753
D2	0.838344	0.120950	0.428768
D3	0.753037	0.117660	0.458827
D4	0.711824	0.814445	0.463790

Table: Performance for the seven models using the Top 10 features from model D3 on Test dataset.

The results with Top 10 variables from Model D4

Models	AUC	RMSE	Mean Per Class Error
M1	0.650105	0.396886	0.388119
M2	0.985096	0.128967	0.171076
M3	0.984594	0.089097	0.131012
D1	0.668186	0.116838	0.419290
D2	0.827911	0.401133	0.383895
D3	0.763055	0.205981	0.461856
D4	0.698505	0.118343	0.441699

Table: Performance for the seven models using the Top 10 features from model D4 on Test dataset.

Insights on Results

- The class of tree-base algorithms (M2 and M3) outperforms (in terms of AUC, Mean Per Class Error, and RMSE) the logistic regression model (M1) and the multilayer neural network models (deep learning D1-D4)) considered in this study in both the validation and test datasets using all the **181** features. We observe that Gradient Boosting model (M3) seem to demonstrate slightly high performance for the binary classification problem compared to Random Forest model (M2), given lower Mean Per Class Error values.
- Upon selection of the top 10 variables from each model to be used for modelling, we obtain same conclusion of higher performance with models M2 and M3.
- The Random Forest model (M2) recorded the highest performance in terms of Mean Per Class Error on test dataset on the top 10 variables selected, out of 181 variables by this model M2.

What are the variables selected by models?

- We observe an important difference in the way the models select and work with the data considered for scoring a company, and as a result accepting to provide them with a loan.
- The model M1 selects more global and aggregated financial variables. The models M2 and M3 select detailed financial variables. The models relying on deep learning select more granular financial variables, which provide more detailed information on the customer.
- There is no appropriate discrimination among the deep learning models of selected variables and associated performance on test set.
- It appears that the model M2 is capable of distinguishing the information provided by the data and only retains the information that improves the fit of the model.

What do we learn from tree-based models?

- The tree-based models, M2 and M3, turn out to be best and stable binary classifiers as they properly create split directions, thus keeping only the efficient information.
- From an economic stand point, the profile of the selected top 10 variables from the M2 and M3 models will be essential in deciding whether to provide a loan or not.

Conclusion

- Despite the hype in using deep learning for credit risk analysis by financial institutions (firms, banks, insurance), it is important to evaluate & monitor the robustness of the performance of these models compared with tree-based models.
- We observed that tree-based models seem to outperform elastic-net logistic regression in binary classification problems.
- It is important to properly evaluate: the quality of data, the transparency of the way in which the algorithms work, the choice of the parameters, the role of the variables (features) to avoid bias, the role of the evaluation criteria and the importance of the humans in terms of final decision.