Stroll through the forest: Applying Random Forest to predict Credit Risk

Maisa Aniceto

PUC-Rio
Brazil

May, 2019

Introduction

Credit risk evaluation $\rightarrow$ financial risk management

- Bankruptcy and insolvency prediction
- 80% of bank loss with financial risk is the result of credit risk exposure (Xu, 2017)
- It's necessary to use models and algorithms that avoid human failure in each credit grant.

Aim

- The use of ML techniques to measure credit risk is an obvious benefit for financial institutions (Wall,2018)
- Credit risk is usually the main concern for banks

Propose

This work proposes to deepen the understanding of the Random Forest algorithm and apply it on a Brazilian dataset.

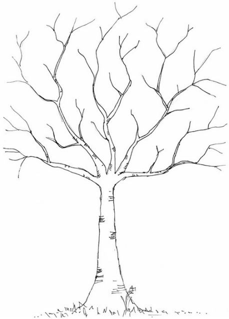
Decision Tree

Decision Tree is the base for Random Forest.
The performance of a DT based credit scoring model is often relatively poorer than other techniques (Wang, 2012).
Decision Tree is easily affected by :

1. the noise in the data,
2. the redundant attributes of data under the circumstance of credit scoring.

Random Forest

- Combinations of decision trees.
- It requires just a small random part from a complete set of observations and manipulates big sets of data (Lantz, 2013).
- Its performance is constantly better than other algorithms (Wall, 2018).
- Its has high prediction accuracy, it is more tolerant to outliers and noise and is less likely to have overfitting issues (Tang, 2018).



Random Forest

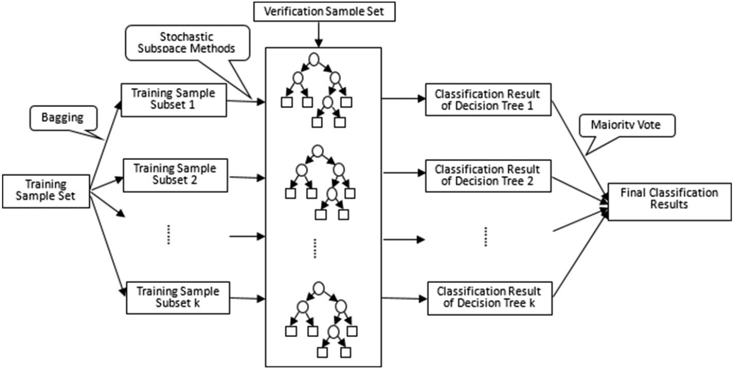


FIGURE – Basic RF framework by Tang, 2018.

Data

Database

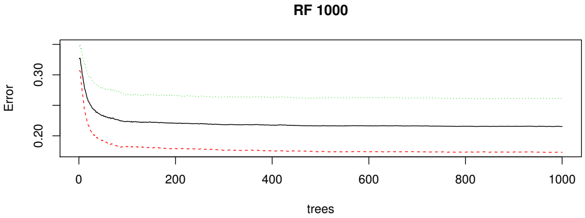
- More than 100,000 consumers.
- Line of credit to individuals.
- Tenor of 24 months.
- A pre-approved limit.
- Fixed interest rate.
- 21 variables
  (income, past loans, savings amount, marital status, type of job, number of dependents, etc).
- High level of credit risk.

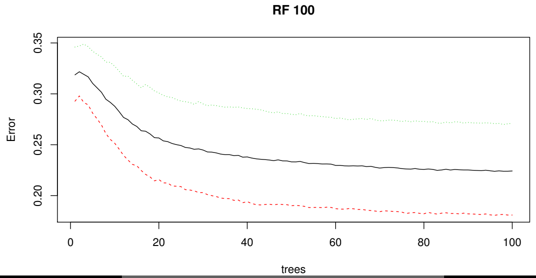
Method

Standard metrics established for credit classification (Wang, 2011, Huang, 2018) :

- Mean accuracy
- Sensitivity (1 - Type I error)
- Specificity (1 - Type II error)
- AUC

Classification error





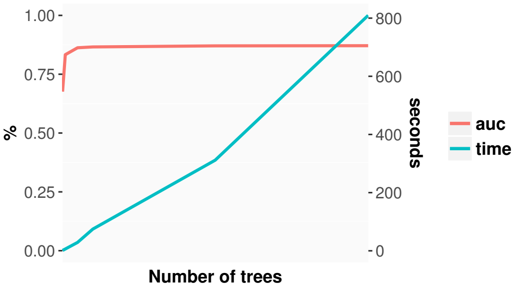
Validation Measures

Number of Trees	Accuracy	Sensitivity	Specificity
1	67.69%	68.90%	66.34%
10	75.69%	75.09%	76.43%
50	77.82%	76.87%	79.03%
100	78.17%	77.34%	79.20%
500	78.41%	77.65%	79.36%
1,000	78.41%	77.62%	79.40%

AUC versus Elapsed Time

Number of Trees	AUC	Time (sec)
1	67.60%	0.92
10	83.31%	5.83
50	86.24%	28.44
100	86.60%	74.37
500	87.05%	311.43
1000	87.12%	809.50

AUC versus Elapsed Time



Conclusion

- The increase in the numbers of trees improves the accuracy of the model.
- The increase of the number of trees increases the elapsed time to gather results.
- Around 100 trees, for this study, posed as the best alternative, presented consistent results throughout the measures applied.

Next step :

- To use validation measures to compare the results (BRIER score, Kolmogorov-Smirnov statistic, CIER measure, among others).
- To analyze different costs of misclassification.

Thank you !

maisa.c.aniceto@gmail.com
Brazil