

Random Forest on imbalanced data

Application to credit card default in Taiwan

Matteo Sani

University of Florence
Department of Statistics and Data Science

January 18th, 2022



UNIVERSITÀ
DEGLI STUDI
FIRENZE

DISIA

DIPARTIMENTO DI STATISTICA
INFORMATICA, APPLICAZIONI
"GIUSEPPE PARENTI"

Outline

- ① Underlying Theory
- ② Application to Credit card default
- ③ Conclusion and Discussion
- ④ References

① Underlying Theory

② Application to Credit card default

③ Conclusion and Discussion

④ References

Classification Trees (Breiman Leo et al. 1986)

General idea: Recursively split the covariate space χ into non-overlapping homogeneous partitions R_1, R_2 , stopping when some termination criterion is satisfied.

A splitting point is defined as:

$$(j, s_1) : R_1 = \{(X_1, \dots, X_p) \in \chi : X_j \leq s_1\}, \quad R_2 = \{(X_1, \dots, X_p) \in \chi : X_j > s_1\}$$

where $j = 1, \dots, p$. Given an *impurity measure* ϕ , the best candidate splitting point must satisfy

$$\min_{j, s_1} [p_1 \phi_{R_1} + p_2 \phi_{R_2}] \quad (1)$$

where p_i is the proportion of observations in region i .

Possible choices of ϕ are:

$$\textbf{Gini index : } G = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk}) \quad (2)$$

$$\textbf{Entropy : } D = - \sum_{k=1}^K \hat{p}_{mk} \log \hat{p}_{mk} \quad (3)$$

where K is the number of classes and \hat{p}_{mk} is the proportion of observations of class k in the m -th region.

The prediction is done by taking the most commonly occurring class in that region.

Pros and cons

- ✓ Easy to explain and easily interpretable
- ✗ Lower level of prediction accuracy

Ensemble Methods: Approach that combines many *weak learners* in order to improve the prediction accuracy. In this framework the principals are:

- Bagging
- Boosting
- Random Forests
- BART

Bagging (Bootstrap Aggregation)

- Resample from the training set to obtain B bootstrap samples
- For each sample a deep and not-pruned classification tree is fitted; it will have low bias but high variance.
- Define $\hat{f}^{*b}(x)$ the prediction for a new observation x . The bagging prediction $\hat{f}_{bag}(x)$ is given by the *majority vote rule*

$$\hat{f}_{bag}(x) = \arg \max_k \sum_{b=1}^B 1_{\{\hat{f}^{*b}(x)=k\}} \quad (4)$$

- **Drawback:** Poorer performances if the trees are highly correlated

Random Forests (Breiman 2001)

Bagging procedure, but each tree is built by taking $m < p$ random variables. Generally $m \approx \sqrt{p}$ for classification.

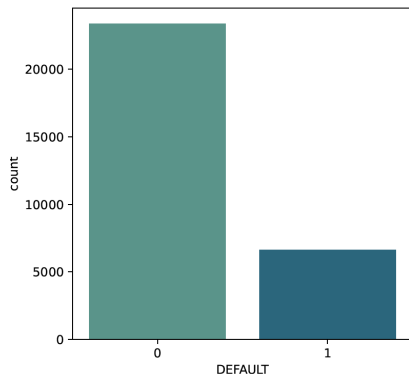
- In all the ensemble procedures, the cost of predictive performance improvement is the lost of interpretability.
- *Variable importance*: Overall measure of the contribute of each variable in the prediction
- Impossible to measure importance in explaining or causing

- ① Underlying Theory
- ② Application to Credit card default
- ③ Conclusion and Discussion
- ④ References

Data¹

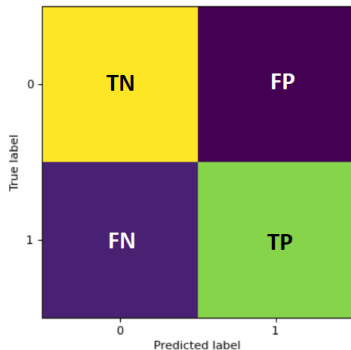
Payment data in October 2005
from a bank in Taiwan: 30000
observations over 23 explanatory
variables.

- Individual characteristics
(sex, age, marital status,
education)
- Loan characteristics
(amount, past payments..)
- Response: Binary Variable
(0: Paid, 1: Default)



¹Yeh 2005.

Metrics



$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

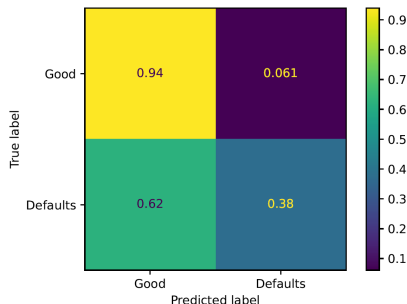
$$Precision = \frac{TP}{TP + FP}$$

$$F1 = 2 * \frac{Sensitivity * Precision}{Sensitivity + Precision}$$

Random Forest Classifier

- 70% of the sample for the training set, 30% for the test set
- Hyperparameters tuning via grid search CV
 - $B = 2000$
 - $m = 5$ (rounded square root)

Accuracy $\approx 82\%$. It's a good performance?



- Excellent performance in predicting good creditors (94% specificity) but on the other hand only 38% of the bad borrowers are classified as such (sensitivity).
- Imbalanced class distribution (77,8% are good clients) has a role
- Re-balance the training set with resampling techniques
 - Undersampling: remove observations from the majority class
 - Oversampling: create new observations in the minority class
 - Hybrid

SMOTE: Synthetic Minority Oversampling TEchnique

Data augmentation technique introduced by Chawla et al. 2002 which generate new synthetic samples.

For each minority sample X in the training set, compute its K -Nearest Neighbors (generally 5). Take only $N < K$ of them where N depend on the amount of oversampling desired.

A new synthetic sample X_1 is obtained as

$$X_1 = X + \delta(T_1 - X) \quad (5)$$

where $\delta \in (0, 1)$ randomly drawn and T_1 is a nearest neighbor.

Performances after resampling

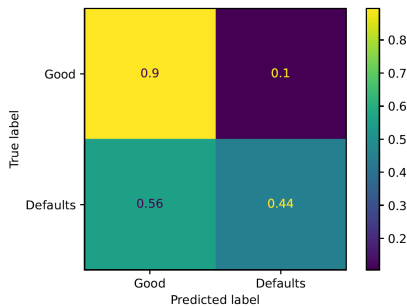


Figure 1: SMOTE

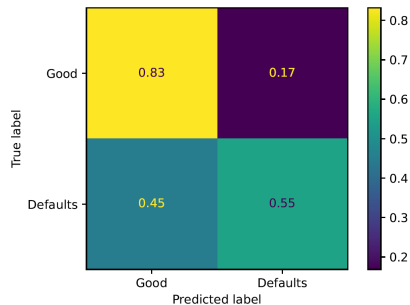


Figure 2: SMOTE+RUS

Comparison

	RF IMB	RF SMOTE	RF RUS	RF SM+RUS
Accuracy	0.816333	0.797667	0.741444	0.779556
Specificity	0.939205	0.895597	0.776563	0.842187
Sensitivity	0.375000	0.445918	0.615306	0.554592
Precision	0.631986	0.543195	0.433969	0.494540
F1-score	0.470701	0.489773	0.508968	0.522848
AUC-score	0.657102	0.670757	0.695934	0.698390

- 1 Underlying Theory
- 2 Application to Credit card default
- 3 Conclusion and Discussion**
- 4 References

Conclusion

Banks would like to prevent defaults on payments with the minimum margin of uncertainty. However, it can be seen that this is accomplished with an increasing number of false positives (good borrowers classified as defaults).





Keeping this in mind, the combination of SMOTE and RUS as proposed in the original paper seems to achieve the best trade-off.





Further Developments

- Features Engineering: extract more relevant information by manipulating the variables
- Improvement of resampling techniques (Fernández et al. 2018) to reduce the amount of overlapping samples
- Different classification algorithm

- ① Underlying Theory
- ② Application to Credit card default
- ③ Conclusion and Discussion
- ④ References**

References

-  James, Gareth et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.
-  Breiman Leo, Grajski et al. (1986). "Classification of EEG spatial patterns with a tree-structured methodology: CART". In: *IEEE transactions on biomedical engineering* 12, pp. 1076–1086.
-  Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.
-  Chawla, Nitesh V et al. (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16, pp. 321–357.

-  Fernández, Alberto et al. (2018). “SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary”. In: *Journal of artificial intelligence research* 61, pp. 863–905.
-  Tan, Xiaopeng et al. (2019). “Wireless sensor networks intrusion detection based on SMOTE and the random forest algorithm”. In: *Sensors* 19.1, p. 203.
-  Malekipirbazari, Milad and Vural Aksakalli (2015). “Risk assessment in social lending via random forests”. In: *Expert Systems with Applications* 42.10, pp. 4621–4631.
-  Namvar, Anahita et al. (2018). “Credit risk prediction in an imbalanced social lending environment”. In: *arXiv preprint arXiv:1805.00801*.



Yeh, I-Cheng (2005). *Default of credit card clients*. URL:
<https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>.