FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Diplomarbeitspräsentation

Cost-based statistical methods for fraud detection.Prediction of never paying customers considering individual risk

Georg Heiler

Masterstudium:

Business Informatics

Vienna University of Technology
Institute of Statistics and Mathematical Methods in Economics Vienna University of Technology
Research group: Computational Statistics
Supervisor: Univ.-Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser

Context of problem

Telecommunication providers not only offer services but increasingly finance consumer devices. Credit scoring and the detection of fraud for new account applications gained importance as standard credit approval processes showed to fall short for new customers as there is only scarce information available in internal systems. Modern machine learning algorithms, however, can still infer intricate patterns from the data and thus can efficiently classify customers.

Research goal

Cost-sensitive methodologies can even enhance the savings. In this thesis, we develop a cost matrix which allows evaluating the individual risk of accepting a new customer and therefore helps to prevent new account subscription fraud optimally. As a nice side effect results can be easier communicated to a non technical audience. More importantly, the value at risk can be assessed better and the cost matrix can aid in handling the imbalance of the dataset.

Methods & formulation of cost matrix

Our best model is using gradient boosted trees (Chen et al., 2016) based on *lightGBM* and is enhanced with our own cost matrix which suites the business needs of the telecommunication industry to factor in the individual value at risk, see Table 1 and Figure 1 Bahnsen et al. (2015) and Vadera (2010).

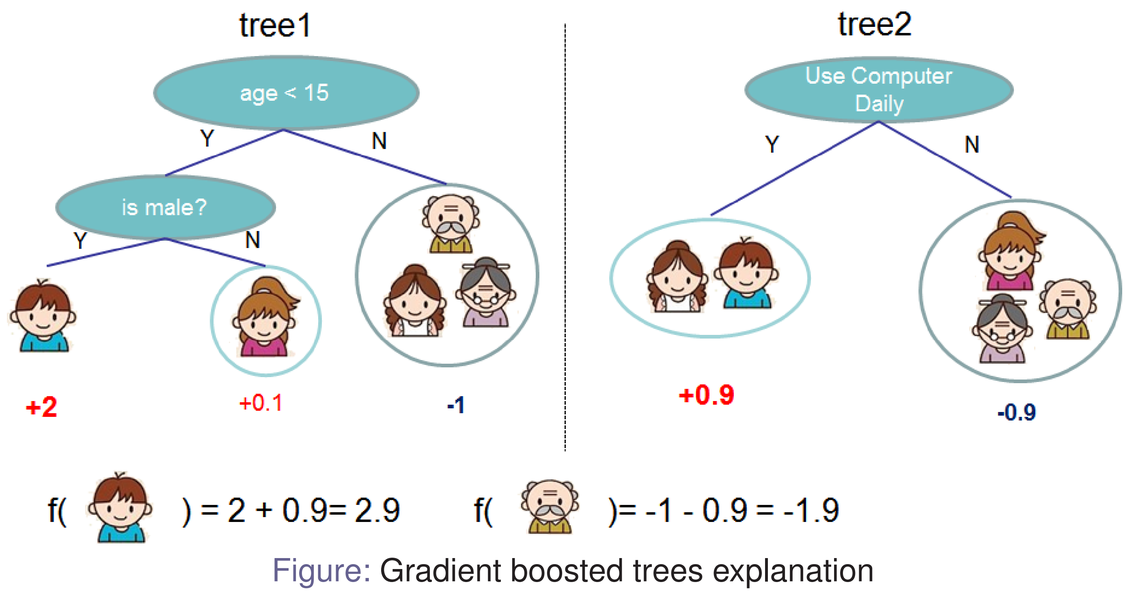

Figure: Gradient boosted trees explanation

Table: Cost matrix proposed for telecommunication industry to price individual risk.

	Actual Pos. $(y_i = 1)$	Actual Neg. $(y_i = 0)$
Predicted Pos. $(c_i = 1)$	$C_{TP_i} = 0$	$C_{FP_i} = r_i + C^m_{FP_i}$
Predicted Neg. $(c_i = 0)$	$C_{FN_i} = C_{device_i} + C_{marketInvest_i} + C_{usage_i} - D_i$	$C_{TN_i} = 0$

False positives C_{FP_i} are the sum of the opportunity financial cost and median risk cost, r_i and $C^m_{FP_i}$, where r_i describes the loss in profit if it had been a good customer.
False negative per customer C_{FN_i} consist of the losses if the customer never pays a single bill.

Results

The current approach for credit scoring at our partner is a traffic light system. To make the models comparable, we can only compare the automated part of the current credit check process and must ignore the manual actions. Therefore, there are two cases to differentiate:

- *red* predictions are assumed to be a neverpayer and *green* a regular customer (*TMA (current)*)
- *red* and *yellow* predictions are assumed to be a neverpayer and *green* a regular customer (*TMA (current, assuming Yellow as neverpayer)*)

Having a closer look at the second case and the F_2 score 2 we see that our model is better than the current process.

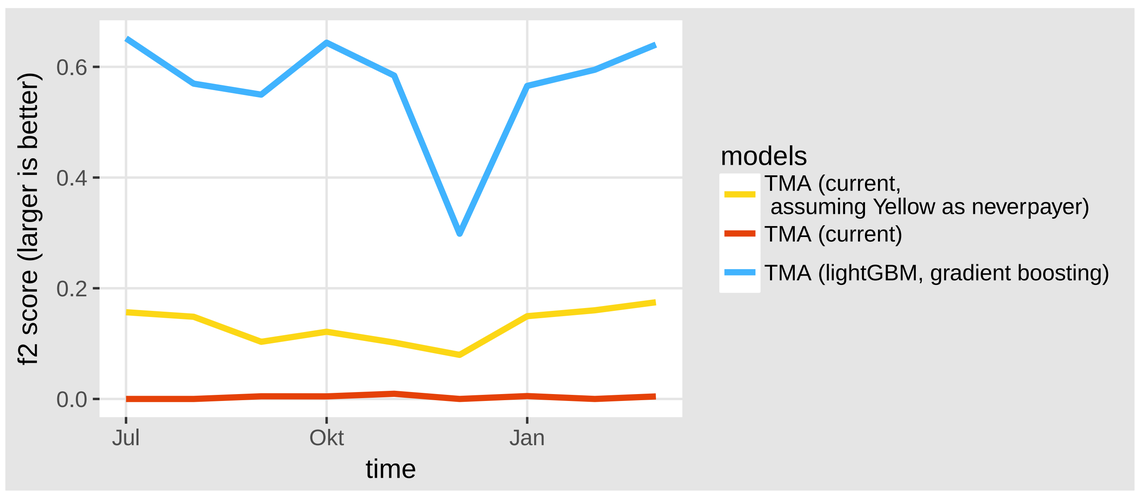

Figure: F2 compared with current approach *red and yellow*

Even though the F_2 score is better for the current approach (*red and yellow*) when additionally computing the savings criterion as shown in Figure 3 it gets apparent that the current approach would be by far too aggressive.

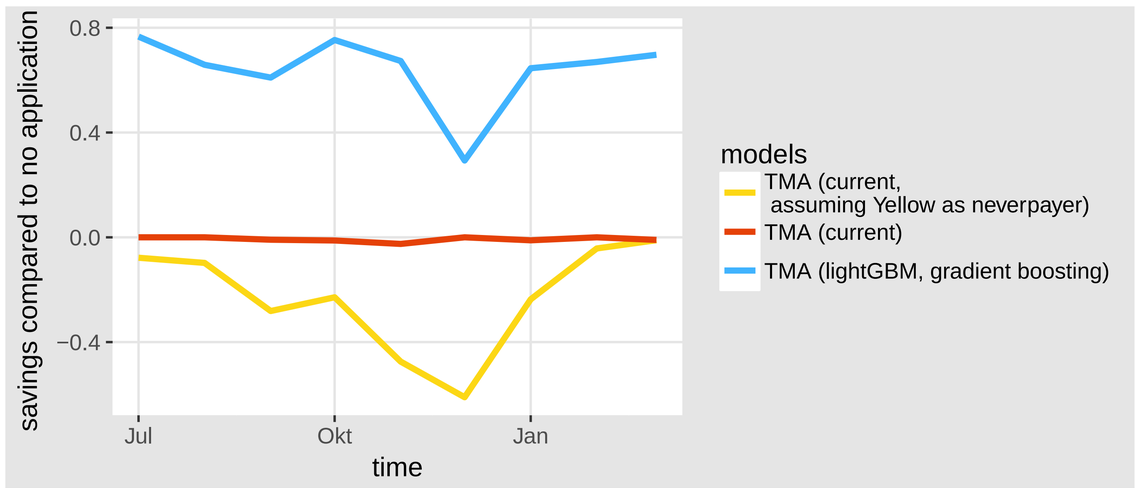

Figure: Savings compared with current approach *red and yellow*

References

- Bahnsen, A. C., D. Aouada, and B. Ottersten (2015). Example-Dependent Cost-Sensitive Decision Trees. *Expert Systems with Applications* 42 (19), 6609–6619.
- Chen, T. and C. Guestrin (2016). XGBoost: Reliable Large-Scale Tree Boosting System. *Arxiv*, 1–6. arXiv: 1603.02754.
- Vadera, S. (2010). CSNL: A Cost-Sensitive Non-Linear Decision Tree Algorithm. *Acm Transactions on Knowledge Discovery from Data* 4 (2), 1–25.

Kontakt: georg.kf.heiler@gmail.com