SME default prediction using novel data sources and machine learning techniques

Fabio Sigrist
Lecturer & Project Leader, fabio.sigrist@hslu.ch
Lucerne University of Applied Sciences and Arts

Christoph Hirnschall
Lead Credit Analytics, christoph@advanon.com
Advanon AG

Stijn Pieper
CTO, stijn@advanon.com
Advanon AG

Bern, June 7, 2018

Outline

- Applying machine learning to scale and digitize at a fast-growing startup
- Novel data sources for credit risk modeling
- The challenge of small data
- A novel machine learning model for default prediction

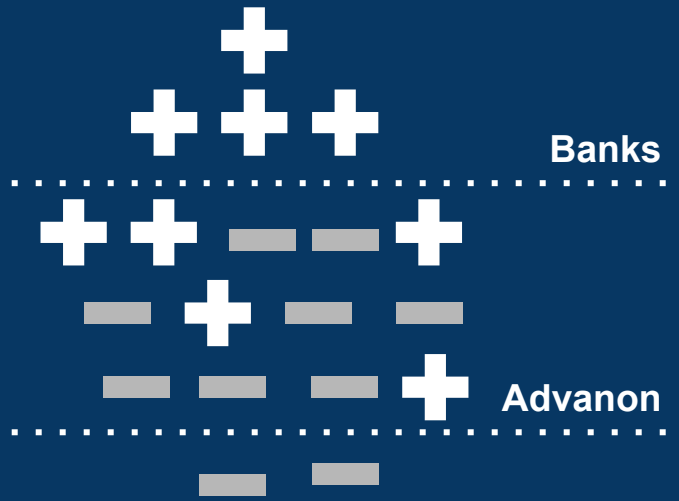
Providing Liquidity to SMEs
Hidden Champions

Predicting **Short Term Repayment** Behavior

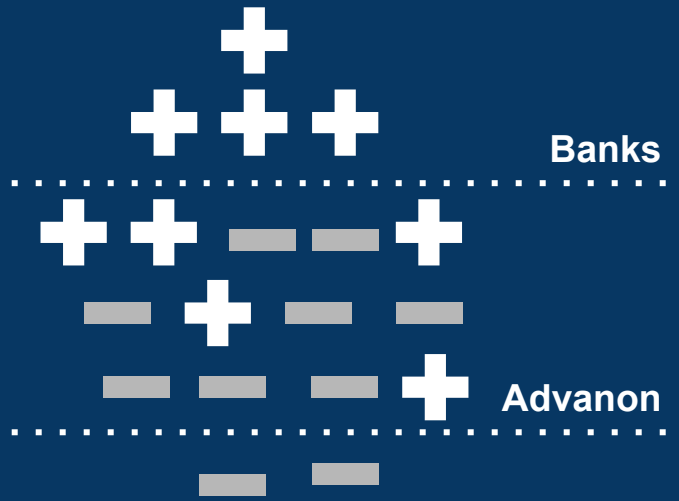
Automation and **Scalability**

- **Time/credit decision:** ~~1-2 weeks~~ <1 day
- **Cost/credit decision:** ~~2000 CHF~~ <50 CHF

Predicting **Short Term Repayment** Behavior

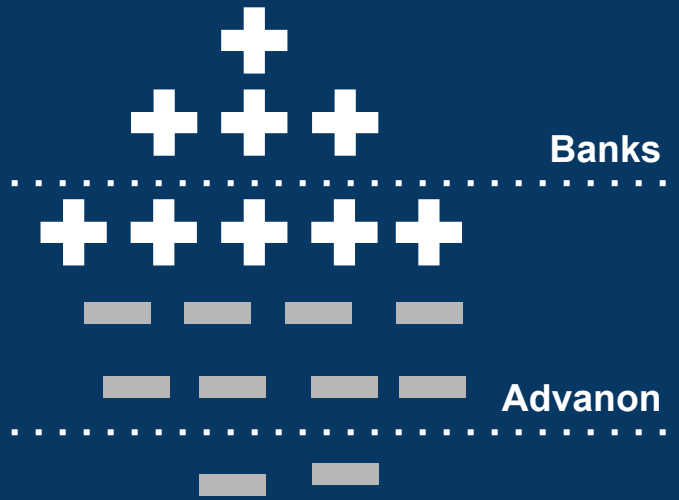


1. Public Websites and APIs
2. Social Media
3. Repayment Behavior
4. Accounting Softwares
5. Financial Statements
6. Bank Statements
7. Credit rating reports



1. Public Websites and APIs
2. Social Media
3. Repayment Behavior
4. Accounting Softwares
5. Financial Statements
6. Bank Statements
7. Credit rating reports

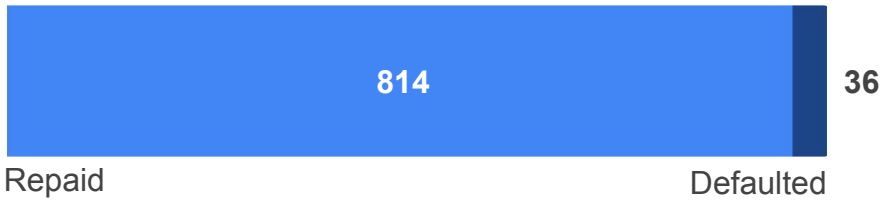
Machine Learning



THE CHALLENGE OF SMALL DATA

Our data[1]

- 850 loans between 2016 and 2017
- 36 loans were not repaid

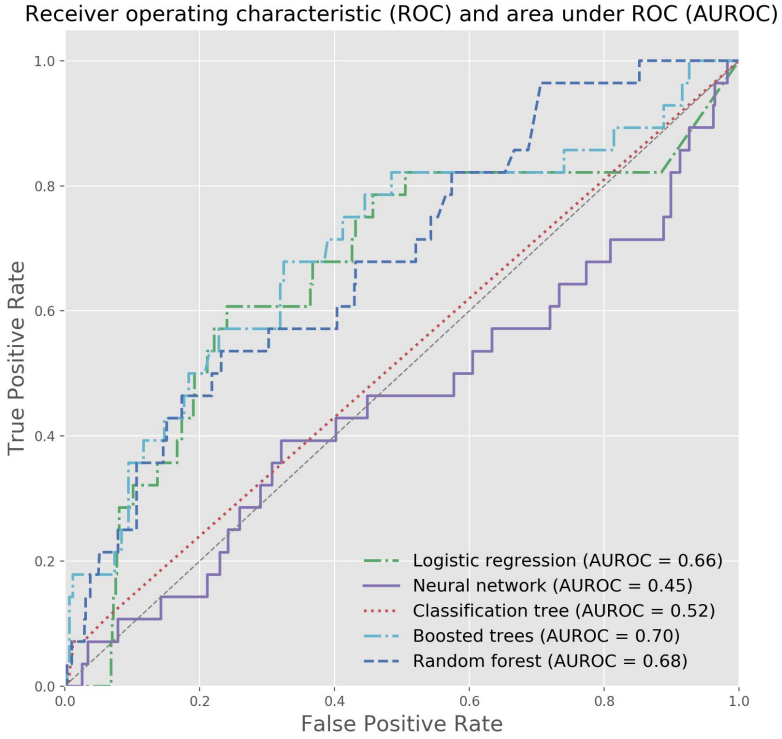


- Approx. 100 predictors (features)

[1]The data presented here consists of a random subsample of all loans made on Advanon's platform. The subsample contains all default events but only a random selection of all non-defaulted loans. This means that the true default rate is different from the one shown here. However, the results change only marginally when using the original data set.

Comparing state-of-the-art classification methods

(temporal leave-one-out cross-validation)

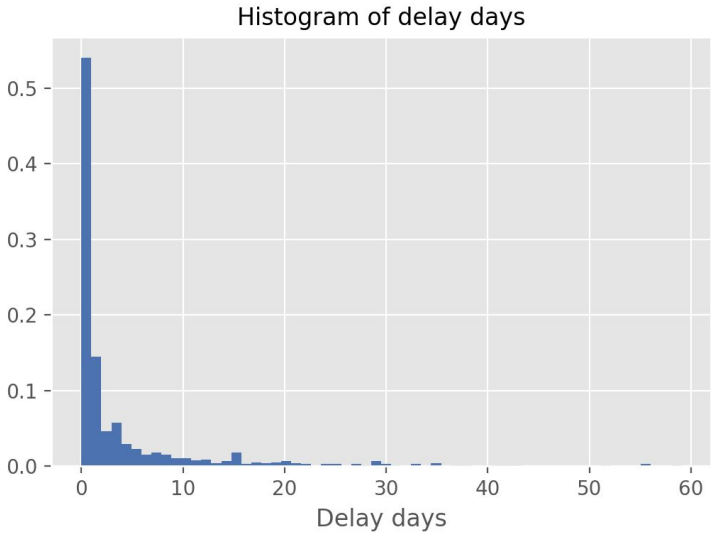


Problem:
small data and
class imbalance

OVERCOMING THE SMALL DATA AND CLASS IMBALANCE CHALLENGE

Using delay in repayment as additional information

- Loans can be repaid with a certain delay (max 60 days) without resulting in a default

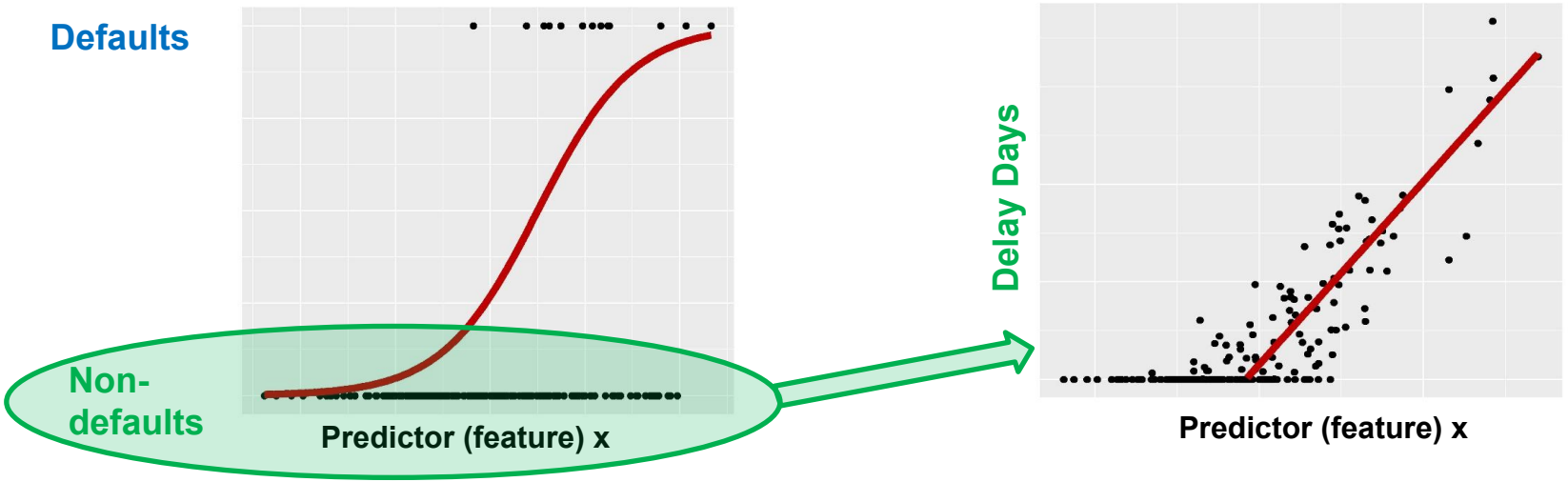


- Is this **additional data useful?** If yes, **how** can it be used**?**

Combination of classification and regression task

We have **two types of response variables** (or labels)

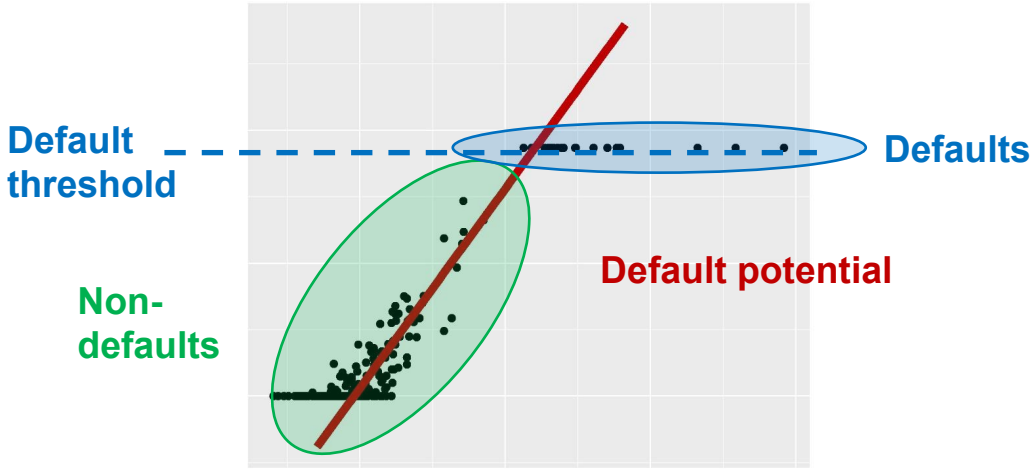
- **Binary default** variable
- **Continuous delay days** $\in [0,60]$ (for non-defaults only)



→ **Combination of classification and regression task**

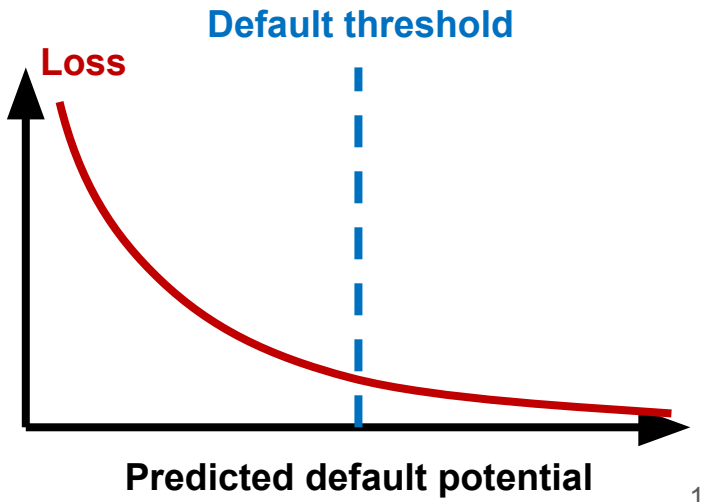
Tobit model can combine both types of labels

- Tobit model
  - o “Combines” discrete and continuous variables
  - o Assumes that a latent variable (**default potential**) drives both default events and delay in repayment

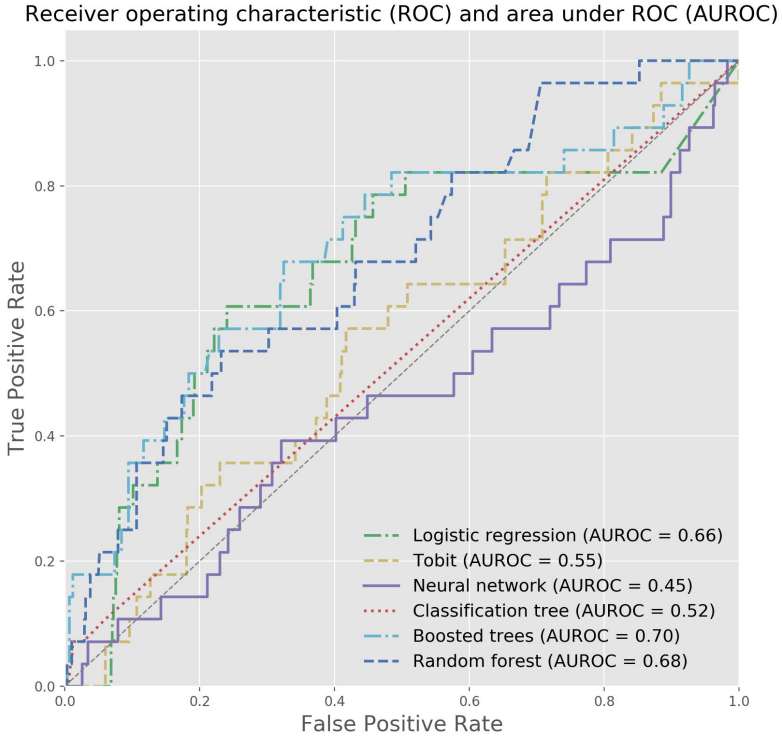


Tobit model can combine both types of labels

- Tobit model
  - o “Combines” discrete and continuous variables
  - o Assumes that a latent variable (**default potential**) drives both default events and delay in repayment
  - o **Asymmetric loss** for defaults
    - “If prediction too low -> large loss”
    - “If prediction too high -> small loss”



Performance of the Tobit model



Drawback of Tobit model: **linearity**

Tree-boosted Tobit model

- Use **gra**dient **b**oosted tree to generalize the Tob**it** model → **Grabit** model
  - Can account for general forms of **non-linearities** and **interactions**
  - **Robust against outliers** in features
  - **Scale invariant** to monotonic transformations for the features
  - **Multicollinearity** not a problem for prediction

- **Less…**
  - data preprocessing and cleaning
  - feature engineering and selection
  - undesired surprises in operational mode

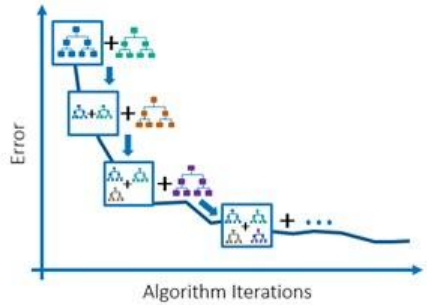
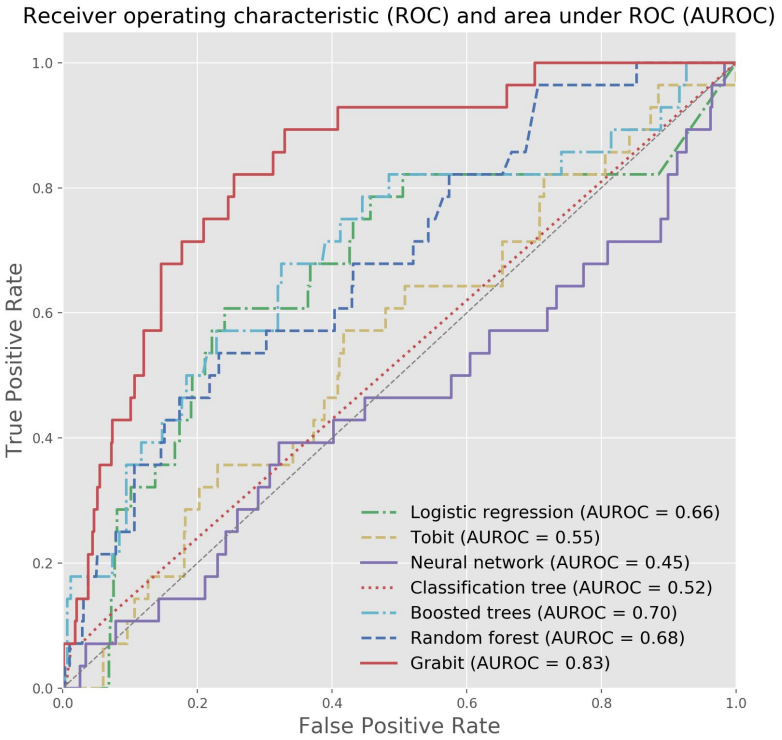


Figure source:
Kdnuggets, Brett Wujek

Performance of the Grabit approach



Large increase in predictive accuracy

Conclusions

- Small data: significant performance gains over out-of-the-box solutions possible in some cases
- Early investment in data quality/storage potentially offers large returns
- Investment in machine learning: tradeoff between short-term and unknown long-term benefits/high uncertainty of outcome
- Machine learning solution provides significant reduction in cost and time for credit risk, while increasing performance
- Limitations: not all manual credit risk activities can be fully automated

References & code

- Preprint available on arXiv
  - https://arxiv.org/abs/1711.08695

- Grabit open source code available on github
  - https://github.com/fabsig/scikit-learn.git (branch ‘grabit’)