Supervised Learning and Rare Events

Tobias Cagala / Deutsche Bundesbank

October 22, 2019

The presentation represents the author's personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or its staff.

Econometrics v Machine Learning

Data generating process $y = \alpha + x'\beta + \epsilon$:

Econometrics: Identify causal effects $\Rightarrow \widehat{\beta}$

Machine Learning: Make accurate predictions $\Rightarrow \widehat{y}$

"... Applying machine learning to economics requires finding relevant $\widehat{y}$ cases."

Mullainathan and Spiess, 2017

Use supervised learning algorithms to approximate human decisions in the DQM process.

Use supervised learning algorithms to approximate human decisions in the DQM process.

Relevance for improvement of decision making process and ultimately, automation.

Use supervised learning algorithms to approximate human decisions in the DQM process.

Relevance for improvement of decision making process and ultimately, automation.

Methodological aspects related to empirical modelling of rare events lend themselves to application of Machine Learning methods.

Machine Learning Modul

Goal

– Use supervised learning algorithms to predict reporting errors

Challenge: Rare events data

– Availability of training data

– Modelling uncommon structure of events (reporting errors)

German Securities Holdings Statistics

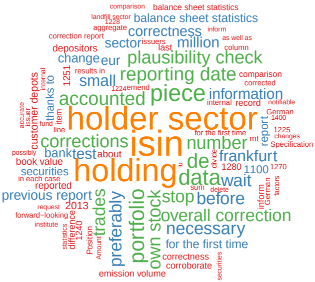
Securities Holdings Statistics

- German banks provide monthly reports of securities holdings (security-by-security)
- DQM with labor intensive manual case-by-case evaluations

Machine Learning: Data Sources

Two Data Sources

1. Text Mining of Written Inquiries

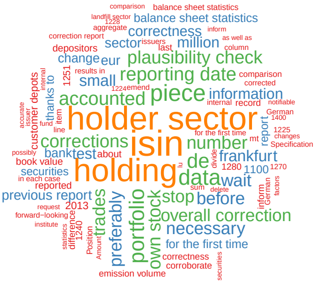


- $\approx$ 10 000 Documents
- Unstructured Data
- Automated isolation of keys in text data
- Takes tables and pictures in documents into account

Machine Learning: Data Sources

Two Data Sources

1. Text Mining of Written Inquiries



- ≈ 10 000 Documents
- Unstructured Data
- Automated isolation of keys in text data
- Takes tables and pictures in documents into account

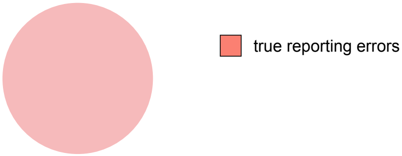
2. Different Versions of the Report

Compare first (pre-DQM) to final (post-DQM) report

Machine Learning: Data Sources

Combination of both data sources

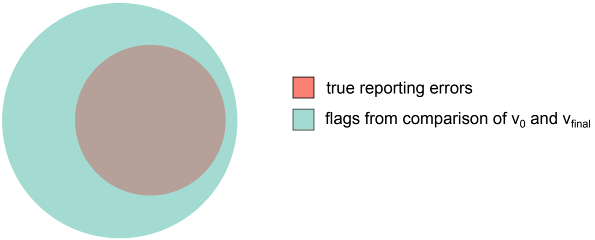
- Both data sources have strengths and weaknesses
- Improve training data by combining both data sources



Machine Learning: Data Sources

Combination of both data sources

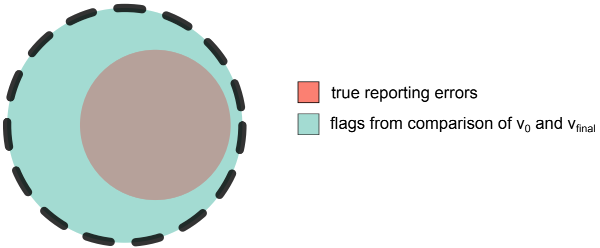
- Both data sources have strengths and weaknesses
- Improve training data by combining both data sources



Machine Learning: Data Sources

Combination of both data sources

- Both data sources have strengths and weaknesses
- Improve training data by combining both data sources



Machine Learning: Data Sources

Combination of both data sources

- Both data sources have strengths and weaknesses
- Improve training data by combining both data sources

Machine Learning: Data Sources

Combination of both data sources

- Both data sources have strengths and weaknesses
- Improve training data by combining both data sources

Machine Learning: Algorithmus

Data Driven Approach

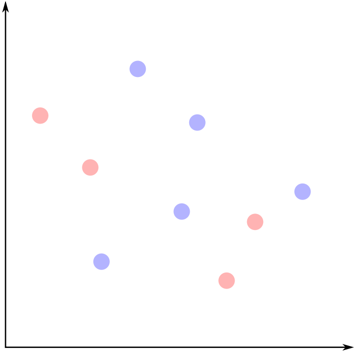
- Randomized-Grid Search for optimal Hyperparameters
- Combination of different algorithms (Stacking)

Method

- Ensemble of algorithms: Nearest-neighbour Clustering, Random Forest, Logit
- Weighting of algortithms' predictions with Stacking (Logit)
- Measurement errors as a Rare Event:
  1. Oversampling of erroneous data points (SMOTE)
  2. Undersampling of correct data points
  3. Weighted Loss Function

Method: SMOTE and Weights

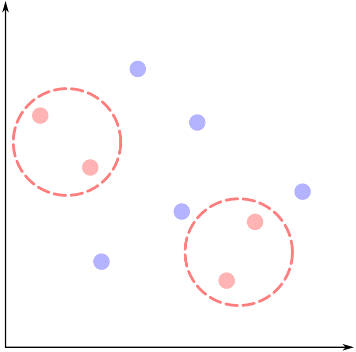
SMOTE



Weighted Loss Function

Method: SMOTE and Weights

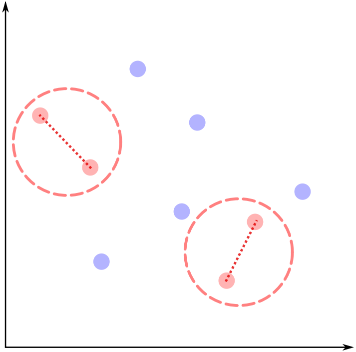
SMOTE



Weighted Loss Function

Method: SMOTE and Weights

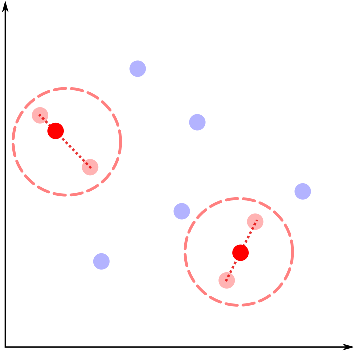
SMOTE



Weighted Loss Function

Method: SMOTE and Weights

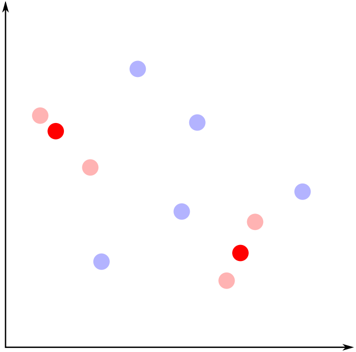
SMOTE



Weighted Loss Function

Method: SMOTE and Weights

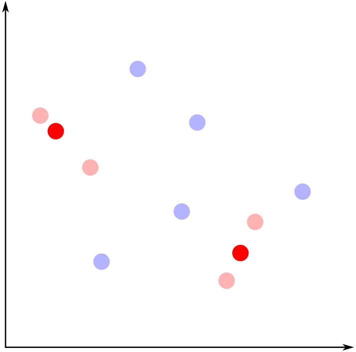
SMOTE



Weighted Loss Function

Method: SMOTE and Weights

SMOTE

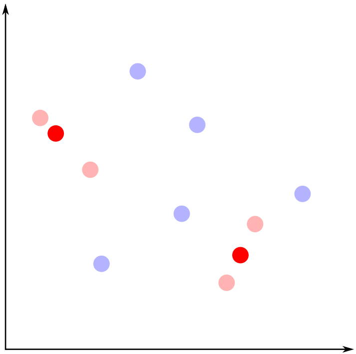


Weighted Loss Function

$$L(X, y, \beta) = \sum_{i=1}^{n} \color{red}{w_i} * f\left(\underbrace{y_i - g(x_i, \beta)}_{\text{prediction error}}\right)$$

Method: SMOTE and Weights

SMOTE



Weighted Loss Function

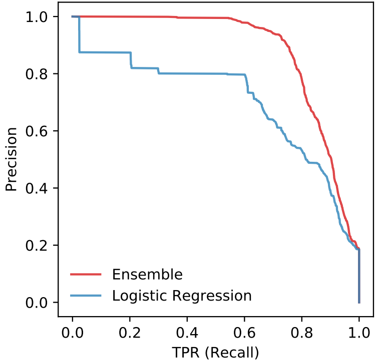
$$L(X, y, \beta) = \sum_{i=1}^{n} w_i * f\big(\underbrace{y_i - g(x_i, \beta)}_{\text{prediction error}}\big)$$

with

$$w_i = \begin{cases} n_{y_i=1}^{-1} & if \quad y_i = 1 \\ n_{y_i=0}^{-1}, & if \quad y_i = 0. \end{cases}$$

Results

Results for 2017: Out-of-sample Performance



- Recall and Precision of 80%
- Performance is superior to Logit
- Performance is even better if we consider the size of the reporting errors

Alternative: (Automated) Rule Based Approach

Advantages and Disadvantages in Comparison to a Rule Based approach

+ Algorithm learns rules
+ Complexity of rules (very) weakly related to cost
+ Automated update of rules
+ Optimization for out-of-sample performance
+ Prediction of (continuous) probability allows for prioritization

– Potential for overfitting
– Transparency
– **Requires data on measurement errors**
– **Learning on historic data**

Discovering New Types of Measurement Errors

Goal

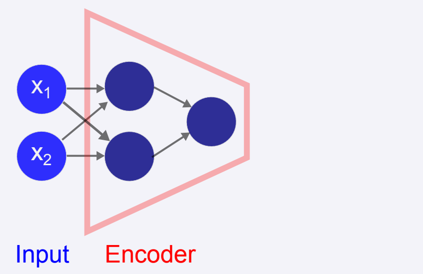
- Isolation of anomalous data points that were not recognized as measurement errors in the past

Challenges

- Interpretation of results (Unsupervised Approach)

Discovering New Types of Measurement Errors

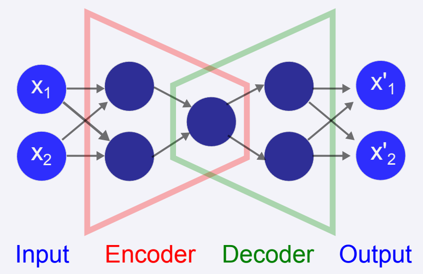
Idea: Autoencoder



- Deep Learning approach
- Simplified 'synthetic' version of the dataset contains information on structure of the data
- Use reconstruction error to discover anomalies

Discovering New Types of Measurement Errors

Idea: Autoencoder



- Deep Learning approach
- Simplified 'synthetic' version of the dataset contains information on structure of the data
- Use reconstruction error to discover anomalies

Discovering New Types of Measurement Errors

First Results

- Successful isolation of anomalous data points in German securities holdings data
- Some of the anomalous data points were classified as measurement errors in the past
- Potential for discovery of novel types of measurement errors
- Largest potential if there is little domain knowledge regarding the data

Conclusion

- Deriving training data from written inquiries and reports is feasible
- Data-driven approach to model selection
- Implementation (ongoing)
  - Microservice architecture with Docker Container
  - Deployment of trained model
- External validity: Similar results with simulated data
- Extension with autoencoder can address issues related to data availability and novel errors

Feature Engineering