BANCA D'ITALIA

Quality checks on granular banking data: an experimental approach based on machine learning

Zambuto F., Buzzi M.R., Costanzo G., Di Lucido M., La Ganga B., Maddaloni P., Papale F., Svezia E.

Bank of Italy

22nd October 2019

Outline

- Context and Motivations
- Data
- The Algorithm
- Results
- Conclusions and Future Work

Context and Motivations (1)

- Central Banks collect, process and disseminate a wide set of statistical data: Data Quality Management (DQM) is crucial to support decision making.

- DQM in Bank of Italy: **automated** checks to verify **pre-determined relationships** in the data (e.g. accounting, logical and mathematical relationships).

- When deterministic relationships are **weak** DQM entails plausibility checks (trend-based) that rely on “acceptance regions” to isolate outliers.

Context and Motivations (2)

- Shortcomings of plausibility checks:
  - Calibration not straightforward**,**
  - Periodical revision and update needed,
  - Large number of acceptance thresholds.

- Complex and time-consuming system with **highly granular** data and **heterogeneous** reporting patterns.

- **Aim**: employ ML to improve plausibility checks in granular databases.

- Approach: a supervised learning algorithm (Quantile Regression Forests) employed to detect potential outliers.

Findings

- Application to payment services data reported by banks. Outliers cross-checked with reporting agents.

- Empirical findings:
  - **New outliers** detected (not identified by the current DQM system).
  - **High accuracy** (77% precision; reduced “false positives”).

- Improvements:
  - Thresholds tailored to the characteristics of banks and to the degree of granularity of the data.
  - Dynamic thresholds that are automatically updated as new data are reported. Reduced involvement of analysts.

Data

- Focus on debit cards issued:
  - Unit of analysis = n. of cards issued by bank (***i***), at the end of the semester (***t***), for a given province (***p***).
  - Data extracted from DWH. Period: Dec-2014 to Jun-2018.

- Additional data on bank features:
  - n. of customers by province of the counterparty,
  - type of customer accounts,
  - other payment services offered (business model).

- Final sample: **18,000 observations** corresponding to **213** banks.

The Algorithm (1)

- Analysis of the empirical distribution of the n. of debit cards (Y) conditional on bank characteristics (Xs).

- Estimation of quantile functions $q_\tau\ (Y|X)$:

$$Prob(Y < q_\tau(X)) = F(q_\tau(X)) = \tau$$

- Quantile functions combined to form **prediction intervals** (acceptance thresholds) associated with a given probability (α):

$$PI(X) = [q_{\frac{\alpha}{2}}(X), q_{1-\frac{\alpha}{2}}(X)]$$

- <u>Outliers</u>**:** values outside the intervals; unlikely to occur (too high/too low) given the reporting context.

The Algorithm (2)

- Sampling:
  - **Train** set to estimate **quantile functions** $q_\tau(x)$ for different τs.
  - **Test** set to compute **intervals** $[\hat{q}_{\tau 1}(x), \hat{q}_{\tau 2}(x)]$ and detect **outliers**.

- Training:
  - Algorithms: Quantile Regression Forest, Linear Quantile Model, Linear Quantile Model with Fixed-Effects.
  - Model selection with 10-folds cross validation.

- Testing:
  - Rolling window with two snapshots of data. Last two semesters in each snapshot as test set.
  - Outliers communicated to banks for cross-check.

The Algorithm (3)

- Model:

$$q_\tau(x_{ipt}) = \beta_0 + \beta_1 depositors_{ipt} + \beta_2 perc\_ca_{ipt} + \beta_3 size_{it} + \beta_4 iss\_acq\_ratio_{it} + \beta_5 trend + \beta_6 sem + \alpha_i + \mu_p$$

- Predictors:
  - $depositors_{ipt}$= N. of depositors (of a bank in a given province)
  - $perc\_ca_{ipt}$= % of depositors with current accounts
  - $size_{it}$= Total transacted amounts (as an issuer and as an acquirer)
  - $iss\_acq\_ratio_{it}$= Balance between issuing and acquiring services
  - sem = Semester dummy
  - t = N. of semesters starting from the first period in the dataset
  - α_i= Bank fixed effects
  - μ_p = Province fixed effects

The Algorithm (4)

- Estimated acceptance thresholds:

$$PI_1(x) = [q_{0.01}(x), q_{0.99}(x)]$$

$$PI_2(x) = [q_{0.025}(x), q_{0.975}(x)]$$

$$PI_3(x) = [q_{0.25}(x) - 1.5 \cdot (q_{0.75}(x) - q_{0.25}(x)), q_{0.75}(x) + 1.5 \cdot (q_{0.75}(x) - q_{0.25}(x))]$$

- Observations falling outside **any** of the intervals flagged as potential outliers.

Results

Cross check of outliers with banks

Prediction intervals:	PI1 $[q_{0.01}, q_{0.99}]$	PI2 $[q_{0.025}, q_{0.975}]$	PI3 Inter-quartile range
a-Total number of potential outliers	**373**	**489**	**457**
b-Anomalies detected and revised ("true positives")	289	312	292
c-Confirmed observations ("false positives")	84	177	165
d-Precision b/a (%)	**77.5%**	**63.8%**	**63.9%**

Concluding Remarks

- Potential to improve DQM: more precise quality checks to detect outliers at a fine grained level with reasonable level of accuracy.
- Maintanance of DQM system: dynamic thresholds and periodical training of the algorithm vs manual update of acceptance thresholds.
- Additional challanges:
  - New processes and IT solutions for the production phase.
  - Communication of anomalies to banks becomes more complex.

Future Work

- Extensions:
  - Application to other payment services data (e.g. credit cards).
  - Analysis of data at the collection stage (i.e. before delivery to the DWH).
  - Classification algorithms (exploiting variations to reported data).
  - Unsupervised algorithms for outlier detection.
- In perspective: extend the ML approach to other granular data collections (in particular when current checks are weak).

Thank you for your attention!

fabio.zambuto@bancaditalia.it