IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Machine learning for anomaly detection in financial regulatory data[1]

Maryam Haghighi, Colin Jones and James Younker,
Bank of Canada

[1] This presentation was prepared for the Workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the central banks and other institutions represented at the event.

Machine Learning for Anomaly Detection on bank regulatory data

October 2021

Maryam Haghighi, Colin Jones and James Younker, Bank of Canada

The views expressed in this presentation are solely those of the authors/presenters and may differ from official Bank of Canada views. No responsibility for them should be attributed to the Bank.

Context

- Rapid growth in data, tools, and analytic techniques to leverage data.
- Opportunity to drive efficiency in our internal operations using non-traditional techniques such as machine learning.
- Transforming to exploit these opportunities.

Key Messages from this presentation

1. We developed a novel method based on machined learning, for detecting anomalies in data from financial institutions.
2. We have operationalized this method to gain efficiencies and better detect anomalies.
3. A robust pipeline built for business users, currently in-use on a daily basis.

Power of Operationalized Machine Learning

60-80 FIs filing returns

From 100's to 10,000's data points/return

26 returns (out of 40) operationalized so far

Millions of datapoints every month

Impossible to examine every single one with traditional ways.
Instead, analysts focused on only a few variables. Even then, *it took significant time*.
Risk that critical anomalies may be missed.

Need high quality data used daily in sensitive economic policy analysis
Over the past year, developed a ML model for anomaly detection.
Now operationalized, running daily @1 am
Detecting anomalies we couldn't detect before + Saving significant time

Formalized with modern Data Science standards: reliable, scalable, fully explainable.

Moving over to cloud environment- data lake

The challenge:

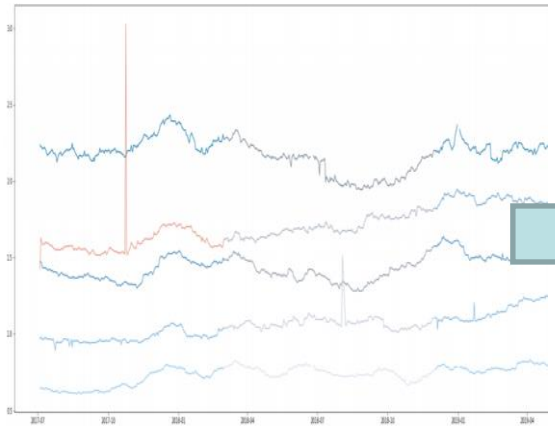
- Financial institutions send us the data (weekly, biweekly, monthly etc.)
- Not every anomaly is an error. How do we know what is an error and what is not?
- Even small numbers of errors can undermine the usefulness of the data
- Traditionally, we used rule-based approaches for anomaly detection.

Our minimum requirements:

1. Function with little labelled data
2. Manage incorrect labelling
3. Avoid reputational risk of false positives
4. Detect anomalies conditional on the activities of that bank
5. Handle regime changes
6. Be useable by non-data scientists
7. Be reliable, automated and scalable

Binary Classification

- Tries to predict if a given time series has at least one anomaly
- Imbalance issue
- Accuracy is not a good performance metric
- Instead use precision (control number of false positive to true positives), recall (identification of true positives) and F_β statistics



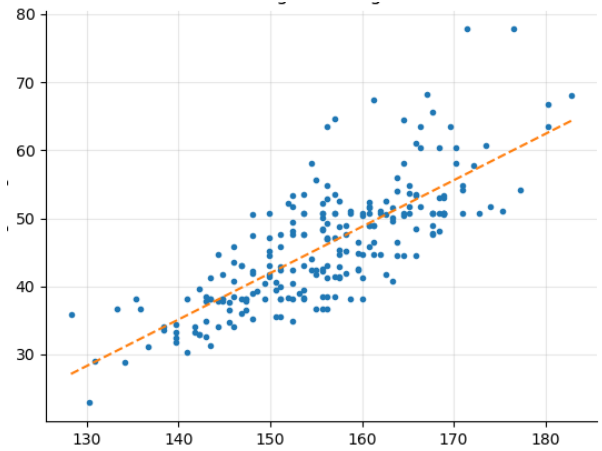
$\{0,1\}$

Questions:

- Null values may or may not be an anomaly
  - Is the bank known to be involved in the specific category, for example in collateral swap market?
- Usually large/small volumes
  - How large is large? How small is small? May vary across financial institutions
- Volatility: How volatile is volatile?
- Spikes: How big of a spike is unusual?

Use of Correlation

- We use correlations between banks and inside the return itself
- Use the data from other FIs that have similar behaviour
  - This multivariate approach allows us to leverage more information as compared to just considering one FI on its own.



- Use the correlation or similarity structure of time series inside the return or across returns.
- Suppose that time series X and time series Y are known historically to be very correlated.
- If one time series behaves significantly differently then the other, we may conclude that the time series in question may be an anomaly.

Two step procedure

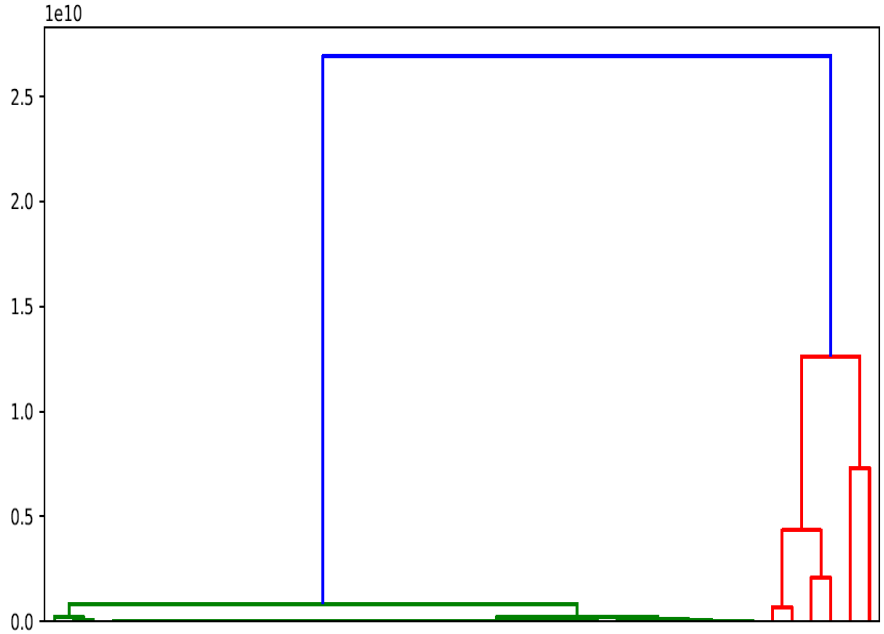
Step 1

- Cluster the Financial Institutions (FIs) based on the raw time series

Step 2

- Implement a supervised ML algorithm using time series features that include covariates from all FIs in the cluster

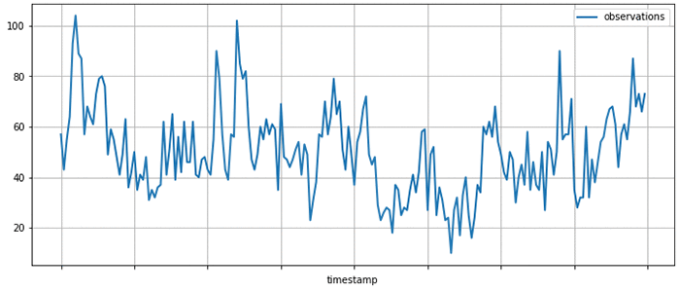
Clustering (Dendrogram)



$FI_1, \ldots, FI_n$

Time series features

- Using the time series features approach we considered many features. These include
  - Mean, max of a rolling standard deviation, max first difference, proportion of zeros; for all banks
  - Mean and standard deviation of correlated time series



$$Features = \{F_1, \dots, F_n\}$$

In production

- The model needs to be maintained, documented, governed and managed to ensure that the model is still performing at its best over time.
- Code requirements (run on several computers, and documented packages required to run ML code), troubleshooting and error management, step-by-step analysis, use of Git and other version control technologies, production schedules, and a centralized tracking log for communication with the FIs.
- Periodic review of training sets, metrics of performance need to be continually examined to further ensure the efficacy of the model.
- Moving the project over cloud environment- ML Ops

Running the model on each FI individually

Worse Performance

Running the two-step model

Better Performance

Conclusion

- Novel method of multivariate anomaly detection
  - Clustering time series by banks that have a strong correlation and then applying a supervised classification ML algorithm.
  - Used correlation among time series within a given return to enhance the detection algorithm.

  These uses of correlation amplified our detection power compared to evaluating each time series on its own.

- This model is actively in production mode. Running daily at BoC, Detecting anomalies otherwise not detectable before + Saving significant time
- reliable, explainable, scalable and robust.
- Moving the model over to cloud environment via MLOps framework.