MS Big Data 2019-2020

SES721 – Ecosystème Big Data

Case Large company :

European Central Bank

1. Context and Organisation

Context :

- established in 1998
- capital of €11 Bn, from E.U. Member States only
- 3600 staff from over 27 nationalities
- yearly budget of over €450 Mn

Organisation :

- Executive Board : 6 ECB staff, including president Christine Lagarde
- Governing Council : Executive Board + governors of the national central banks of the Eurozone's 19 member states (25 members)
- 29 « Directorate General » (divisions), divided between :
  - core ECB : e.g. DG-Economics, DG-International, DG-Monetary Policy, DG-Research
  - SSM (Single Supervisory Mechanism) : Microprudential Supervision I-IV

1.1 Main objectives and tools

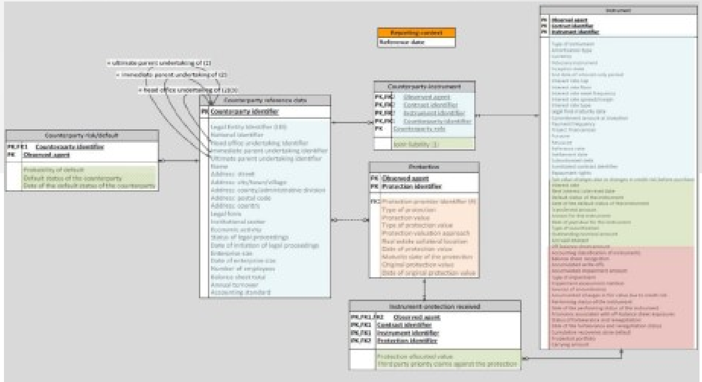
Main objectives and tasks :

- achieve and maintain price stability
- support economic growth
- define and implement monetary policy
- forecast and analyse key macro-financial variables : GDP, inflation, interest rates...
- support financial stability and supervise credit institutions

Tools (regarding data science):

- macro side (monetary policy) :
  - classic econometrics models (linear regressions, Vector Autoregressions)
  - theoretical side : DSGE models (Dynamic and Stochastic General Equilibrium)
- micro side (supervision of credit institutions) :
  - no models : credit institutions self-evaluate ! (IMM : Internal Model Methods)
  - loose verification of compliance with ECB guidelines

1.2 Where is Big Data ?



Big Data :

- fledging concept at the ECB
- first Data Lab in 2017
- first parallel computing and machine learning models in 2018
- first data science team established in 2019 !

Remote platforms and databases

- remote storage and analytics platforms : DISC, ORBIS
- EMIR database (2017)
  - transaction-level data on derivatives contracts
  - 90K files, 25Bn observations, Hadoop storage and Spark treatment
- Anacredit database (2012)
  - detailed records on 60 million individual bank loans in the Euro Area
- MMSR database (2016)
  - loan information from 52 largest Euro Area banks
  - daily record of 45000 transactions on unsecured/secured loans, total volume €600Bn

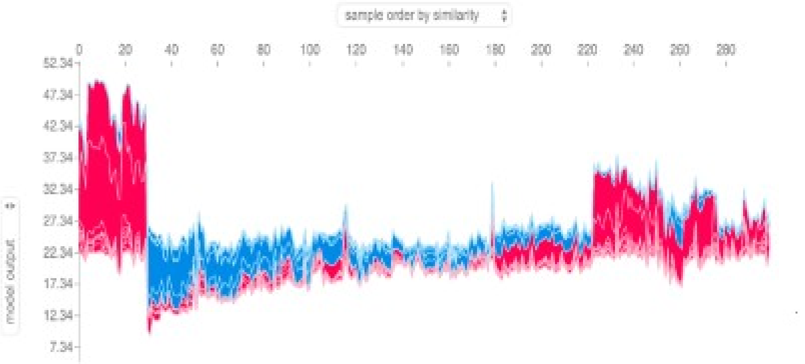
1.2 Where is Big Data ?

Machine learning

- Anacredit
  - outlier detection (isolation forest)
  - data compression (auto-encoders)
  - feature selection (XGboost)
- SSM FAQ content: NLP for topic classification of legal opinions
- DISC Cloud environment: web scraping of online stores for inflation nowcasting

Data visualisation

- SUBA : querying and visualisation on Hive/Tableau
- Anacredit : ELI5, DeepVis



1.3 Where is Big Data ? (again)

Big data mostly remarkable for its *absence* of core areas !

Macroeconomic forecasting and modelling

- core activity of DG-Economics (including modelling division) and DG-research
- no modern machine learning models used in this division, nor in any core ECB division
- exclusive use of DSGE, Bayesian VARs and other semi-structural econometric models
- yet all quantitative ML algorithms *could* and *should* be used for possible performance and accuracy improvement

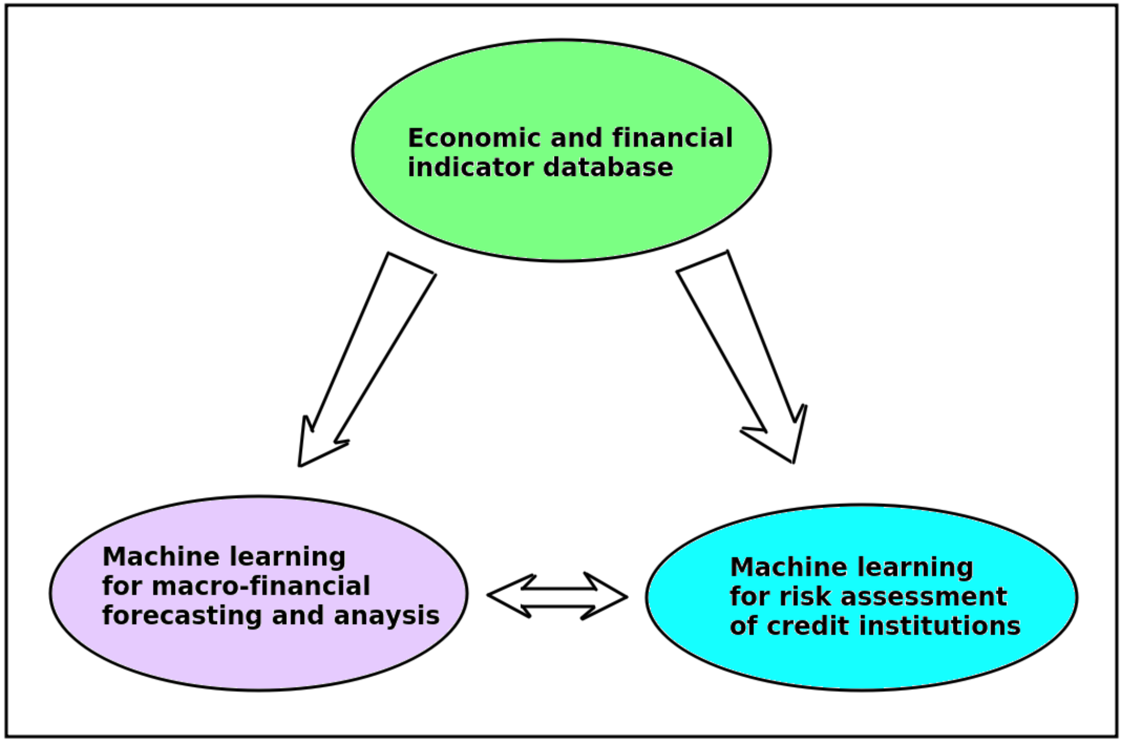
Risk assessment of credit institutions

- core activity of SSM (DG-Microprudential Supervision I-IV)
- no machine learning algorithms used since no internal modelling at all !
- risk assessement realised by institutions themselves, mostly on Excel models
- yet very strong potential for ML algorithms on classification and outlier detection

2. Big Data case

Value proposition : big data project based on 3 pillars

- expanded microtransaction database for economic indicators
- machine learning models for macro-financial forecasting and analysis
- machine learning models for risk assessment of credit institutions



2.1 Pillar 1 : database for economic indicators

Massive databases already exist

- EMIR, Anacredit, MMSR, DISC
- focus only on financial transaction data
- need for more economic-oriented indicators
- proposition : create new massive database with micro-transaction economic data

Possibility 1 : web-based data

- systematic use of web scraping
- possible data : hotel booking, online prices on Amazon and retail stores, job offers, Google search, social medias for agent sentiment, real estate offers, ...

Possibility 2 : national administrative data

- national administrations collect *a lot* of individual data (e.g. INSEE)
- possible data : monthly VAT records, employer contributions, notarial acts, vehicle registration, income tax statements, oil tax, electricity consumption, ...

2.2 Pillar 2 : machine learning for economic forecasting

Economic forecasting :

- key for decision-taking : interest rate primary decided by Board based on GDP/inflation forecasts
- proposition : supplement classical DSGE and econometrics model with modern machine learning algorithms

Possibility 1 : standard economic datasets

- use regular time series and economic indicators : real GDP, inflation, interest and exchange rates, consumption, investment, financial index (e.g. VIX), …
- candidate algorithms : SVM, KNN-TSPI, random forest regression (and all bagging algorithms) ; boosting algorithms (XGBoost, AdaBoost, CatBoost, Light GBM) ; deep neural network, in particular RNN and LSTM.

Possibility 2 : micro-transactions datasets from pillar 1

- offers strong potential for nowcasting (thanks to high-frequency data like daily datasets)
- potential need for dimensionality reduction : PCA, UMAP, t-SNE
- then run same ML algorithms as mentioned above

2.3 Pillar 3 : machine learning for risk assessment

Risk assessment of credit institutions :

- main task of SSM, but no formal machine learning algorithms due to IMM
- proposition : integrate machine learning to internalize the SSM assessment process, exploiting the massive financial transaction databases

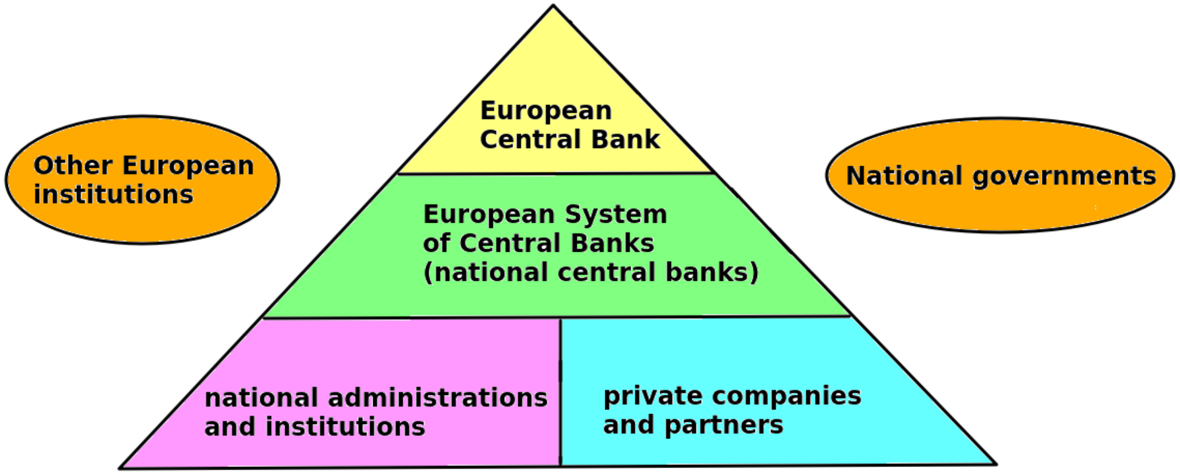
Strategy :

- exploit already existing massive financial databases : EMIR, Anacredit, MMSR
- if needed : dimensionality reduction of datasets, using previously mentioned techniques
- classification algorithms to assess the solvability of credit institutions and borrowers risk : KNN, SVM, bagging and boosting algorithms, deep neural networks, ...
- outlier detection algorithms to detect non-normal and risky behaviours : SVMRank, RankBoost, plug-in techniques (KNN and kernel smoothing), isolation forests, one-class SVM, …
- interpretable algorithms to analyse agent decision-making processes : decision trees, random forest with feature importance and Shapley value.

3. Further considerations

ECB : part of a complex hierarchical system

- ECB is not autonomous : decisions are taken both at the international (European Union) level and at national level (national governements in Governing Council, national administrations and national central banks)
- necessary to account for this specificity in business plan, cost estimation and deployment calendar



3.1 Pillar 1 : database for economic indicators

Key resources and partners : technical structure

- embryo of infrastructure already exists : DISC database
- already established a 5-year partnership with T-systems and Cloudera

Key resources and key partners : content

- massive database of micro-transaction economic data cannot be collected by ECB alone
- part can be done at ECB level (web scraping and web exploration)
- but most of it : information collected by national central banks (e.g. Banque de France) and national administrations (e.g. INSEE) then channelled to ECB database
- possible cooperation with corporate sector (e.g. Altares)

3.1 Pillar 1 : database for economic indicators

Cost structure :

- limited infrastructure costs as cloud database already exists (DISC)
- other costs : extra staff for part realised by ECB , extra data storage
- precise storage cost impossible to evaluate (business confidentiality + unknown storage needs)
- rough estimate : €200-300K/year for storage, €300K/year for 3 data scientists
- extra costs : data collection at national level, including additional staff and computer equipments
- also : data possibly collected already but need to centralise, clean and format datasets
- bulk of the project cost : possibly millions of Euros, especially for poorest E.U. countries with limited infrastructures (very rough estimate : €10Mn/year for whole E.U.)

Application-specific challenges :

- political : convince Executive Board and Governing Council of project utility, unlock budget
- enforce data collection : adoption by Board → regulation by Commission → translation in national legislations → assessment and control by national administrations and ECB
- minor : privacy and reputational risks with use of personal micro-data
- minor : possible issues with data quality and selection bias compared to official data

3.1 Pillar 1 : database for economic indicators

Deployment calendar :

- at least one year for negociation and transcription in law at E.U. and national levels
- at least another year for technological deployment and data collection at national and ECB level
- up to one year to collect sufficient data for machine learning/data analysis
- minimum of 2 years before first assessment

Measuring impact and success :

- control compliance of national authorities to integrate national data within planned timeline
- evaluate volume / number of observations generated by the new database
- estimate use of new database by ECB staff for machine learning applications
- compare requests of new database with existing macro-financial databases (Bloomberg/Thomson)
- recommandation : initial evaluation report after first 3 years, then renew report every 2 years

3.2 Pillar 2 : machine learning for economic forecasting

Key resources and partners :

- databases : commercial databases (Thomson, Bloomberg) and massive database from pillar 1
- standard libraries for machine learning : Scikit Learn, TensorFlow, SparkML
- at least one cluster for Hadoop/Spark treatment
- academic partnerships : to recruit data scientists and remain up-to-date for ML algorithms

Cost structure :

- recruit 4 data scientists, 1 data engineer, 1 lawyer : 6 × €100K = €600K/year
- Spark/Hadoop cluster : €100K/year for deployment and maintenance
- total cost about €700K/year

Deployment calendar :

- 4-6 months to estimate the first models on regular data
- one year to deploy Hadoop/Spark cluster and estimate models on massive micro-data

3.2 Pillar 2 : machine learning for economic forecasting

Application-specific challenges :

- lack of understanding of big data and its potential for economists :
  - Benoît Coeuré (2017) : Big Data reduces to « timelier and richer data »
  - big data limited to DG-statistics and DG-IT : unknown in DG-Economics (« geek thing »)
  - no ML culture due to long-time presence of DSGE/econometrics : more a curse than a blessing !
- ideological reluctance to use machine learning :
  - genuine « religion » of DSGE modelling at ECB and DG-Economics
  - Christiano et al. (2017) : « People who don't like dynamic stochastic general equilibrium (DSGE) models are dilettantes »
  - need to first to convince executives that machine learning can improve on modelling

Measuring impact and success :

- compare forecasting performance of machine learning models over traditional DSGE and VAR models (RMSE, MAE, Theil's coefficient)
- analyse performance of nowcasting algorithms and their contribution to decision making

3.3 Pillar 3 : machine learning for risk assessment

Key resources and partners :

- resources already exist : databases (EMIR, Anacredit, MMSR) and remote platform (DISC)
- everything can be done internally without further ado

Cost structure and deployment calendar:

- recruit 3 data scientists : 3 × €100K : total cost of €300K/year
- allow for one year to produce first machine learning algorithms and risks assessments

Application-specific challenges :

- political challenge : switch from IMM (credit institutions evaluate their own risk level) to internal risk assessment by SSM
- need to convince Governing Council and national States that delegating to the ECB generates added value and more reliable risk evaluation (thanks to ECB independance)

3.3 Pillar 3 : machine learning for risk assessment

Measuring impact and success :

- compare risk assessment produced internally with that estimated by IMM : are they consistent ?
- on historical data, compare default prediction of machine learning models over IMM models (ROC, AUC)
- assess whether ECB machine learning models identify new outliers or institutions with underestimated risk
- evaluate the contribution of machine learning algorithms to improve understanding of lender/borrower decisions.

Conclusion : key messages

- big data remains significantly under-developed at ECB
- mostly used for massive databases at DG-Statistics and DG-Information Systems
- current databases focus on financial transactions : no massive database for economic indicators
- machine learning is completely absent from DG-Economics
- yet machine learning could and should be used to improve on macoreconomic forecasting and risk assessment of credit institutions
- proposition : 3 pillars to develop big data at ECB : creation of a massive database for economic indicators ; creation of machine learning team for economic forecasting; creation of machine learning team for risk assessment of credit institutions
- cost is not a major problem for pillars 2 and 3 : total yearly cost < €1Mn, i.e. 0.2 % of ECB yearly budget
- cost may be an issue for pillar 1 : possibly up to €10Mn/year or more, to be shared among NCB's
- main challenges are political :
- internally, convince Executive Board and division managers that big data and machine learning can generate added value ; create awareness and culture of big data
- externally, convince national governments that additional expenses and delegation of control over the risk of credit institutions are justified and serve the common interest