Chaire LUSIS - CentraleSupélec

Jeudi 16 janvier 2020

Historique

- 2004 :
  - Détection de fraude - arbres de décision, réseaux bayésiens, HMM, SVM
- 2006 :
  - Simulateur de paris boursiers en ligne - théorie des jeux, espérance de gain
- 2014 :
  - Apprentissage profond pour la détection de fraude
  - Théorie de Dempster-Shafer - application à la prise de décision
- 2015 :
  - Apprentissage automatique pour la détection de fraude
  - Apprentissage automatique pour la prédiction de la durée de transport de marchandises

Historique

- 2017 :
  - Recommandation musicale - Qobuz
  - Apprentissage profond pour la prédiction de marché
- 2018 :
  - Apprentissage automatique pour la détection de fraude
  - Chatbot pour domaine spécifique

Chaire - Moyens

- Durée 4 ans
- 3 enseignants-chercheurs CentraleSupélec rattachés au LRI (UMR8623)
  - Fabrice Popineau - équipe LADHAK
  - Arpad Rimmel - équipe GALAC
  - Bich-Liên Doan - équipe A&O
- Equipe IA Lusis
  - Fabrice Daniel
  - Vincent Reinhard
  - Rosella Martino
  - François de la Bourdonnaye
  - Xavier Farchetto
- 2 thèses sous contrat CIFRE
- 3 contrats d'étude industrielle par an

lusis CentraleSupélec

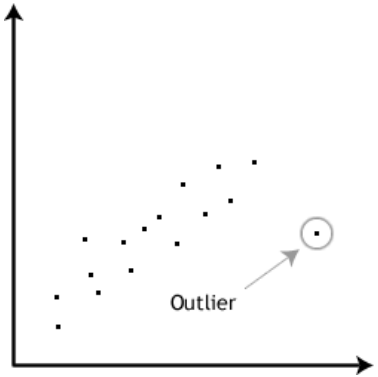
Fraude

lusis CentraleSupélec

Axe de travail - Fraude

Fraude

- Perte sèche, en augmentation
  - En 2018, US$ 24,26 10^9 perdus mondialement
  - Usurpation d'identité
- Eviter les faux positifs!
- Problème
  - de **classification** binaire
  - détection d'**anomalie**



£17 million stolen - biggest UK credit card fraud

In the mid 2000s, a gang of international fraudsters managed to steal the details of over 32,000 credit cards. They used this information to create clone credit cards and scam at least £17million over a period of several years.

lusis CentraleSupélec

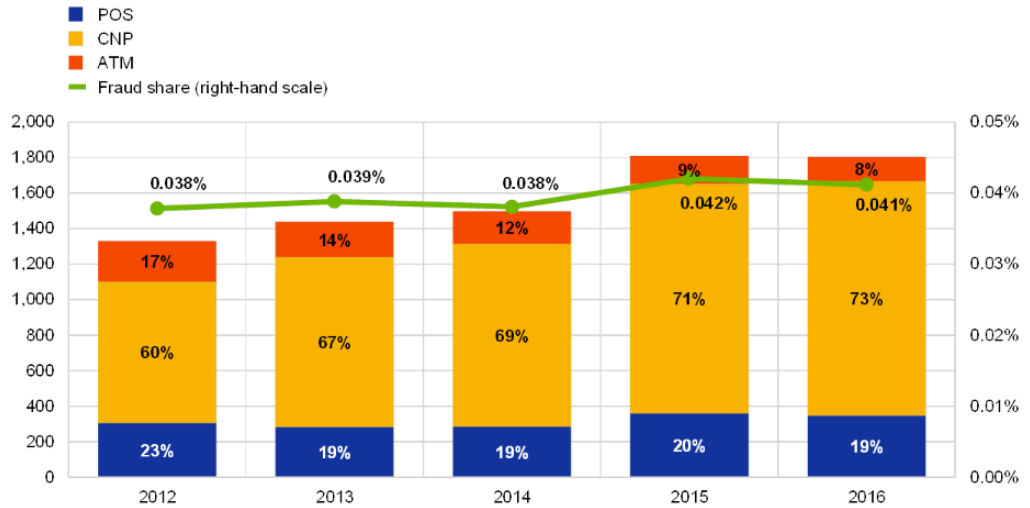
Axe de travail - Fraude

Part de la fraude

SEPA = Single Euro Payments Area

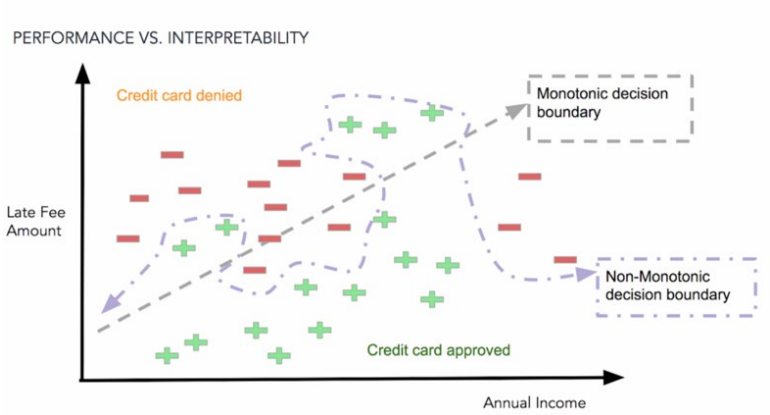
CNP = Card-Not-Present

POS = Point-Of-Sale



Fraude - Verrous

- Données déséquilibrées
  - Pourcentage de transactions frauduleuses < 0,5%
  - Difficulté d'accès à des datasets
- Détection en ligne en temps-réel
- Dérive conceptuelle et temporelle
  - Evolution des habitudes de consommation des clients
  - Etiquetage à retardement des cas de fraude
  - Nouvelles tactiques de fraude
- Explicabilité
  - RGPD : impose de fournir au client un motif de refus compréhensible par l'humain



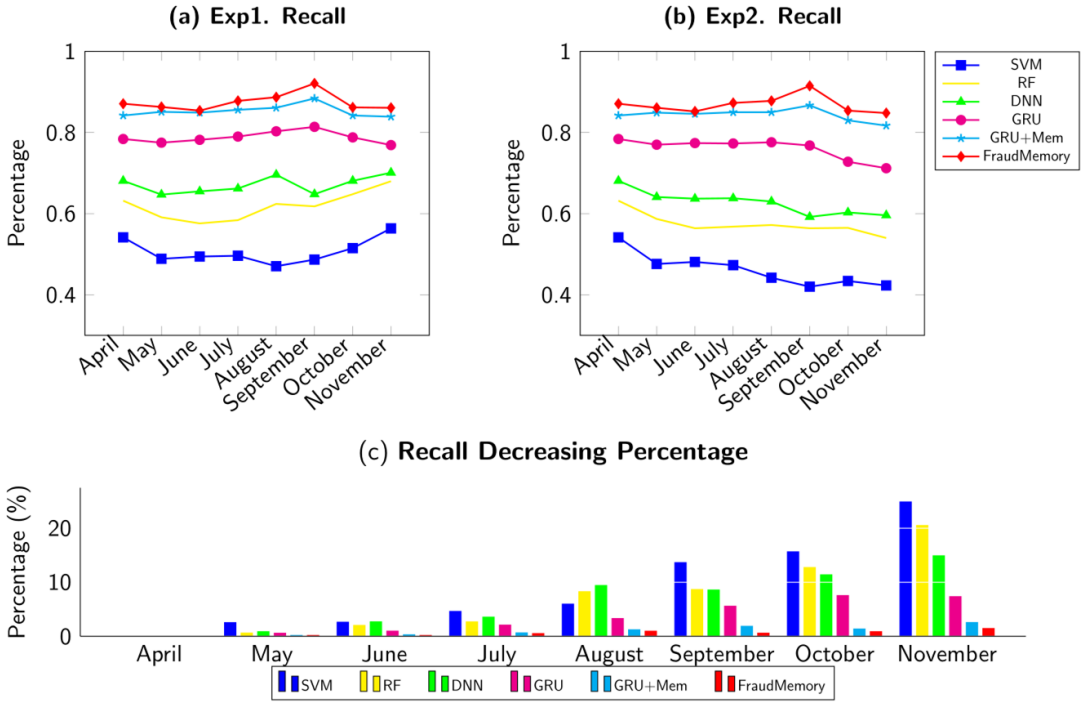
lusis CentraleSupélec

Fraude - Apprentissage automatique

Différentes techniques :

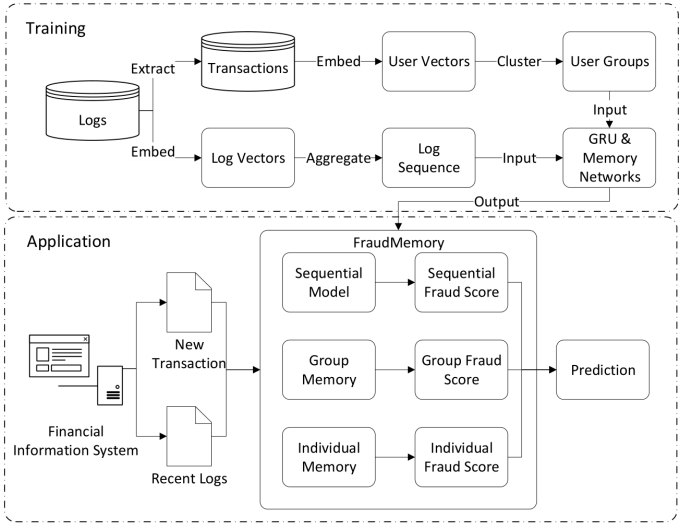
- SVM, DNN, GRU,

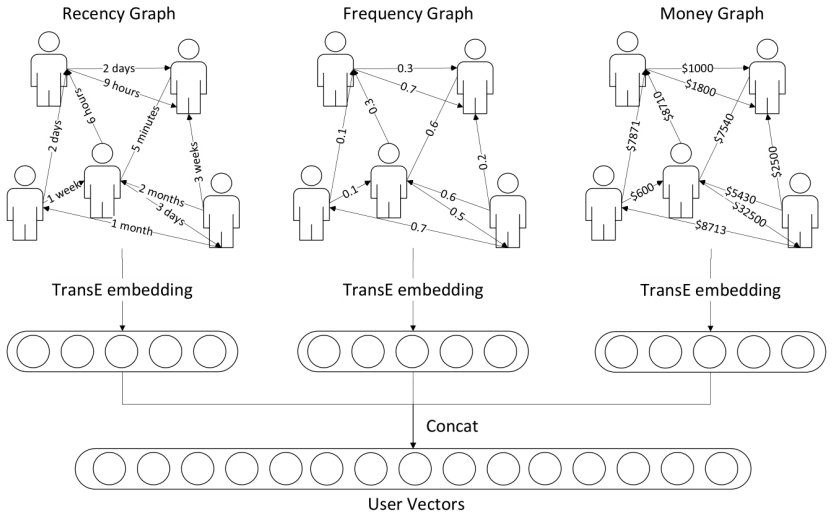
  Random Forests
- Rappel TP/(TP+FN)
- Dérive conceptuelle



lusis CentraleSupélec

Fraude - Apprentissage automatique

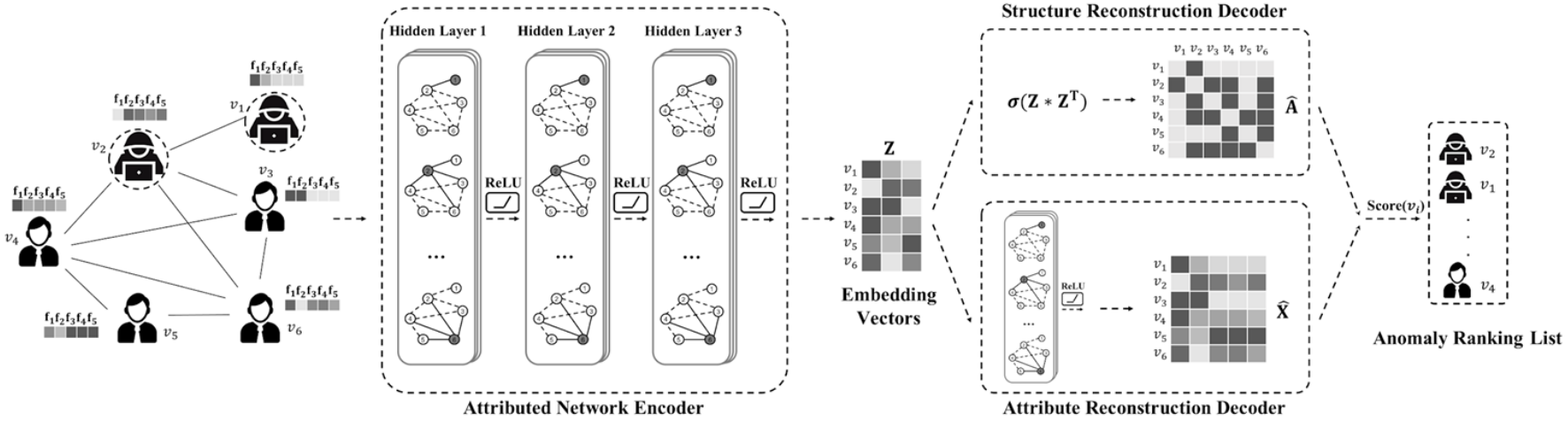




Architectures complexes

Robustesse des réseaux de neurones

Fraude - Réseaux attribués

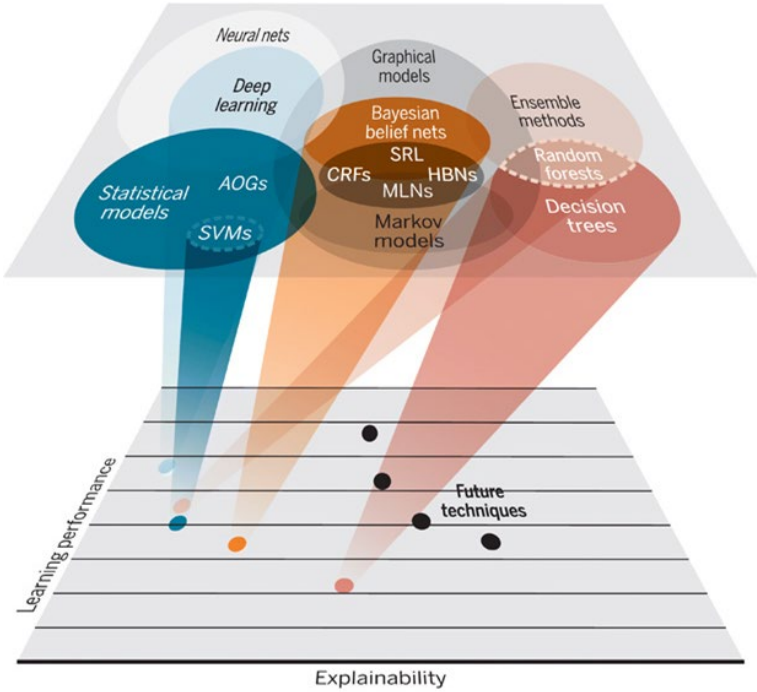


Fraude - Explicabilité

Les meilleures techniques d'apprentissage automatique sont des boîtes noires (+/-)

Exigence réglementaire : fournir une explication à l'utilisateur

Construire une explication à partir des éléments qui ont présidé à la décision



Performance vs. explainability

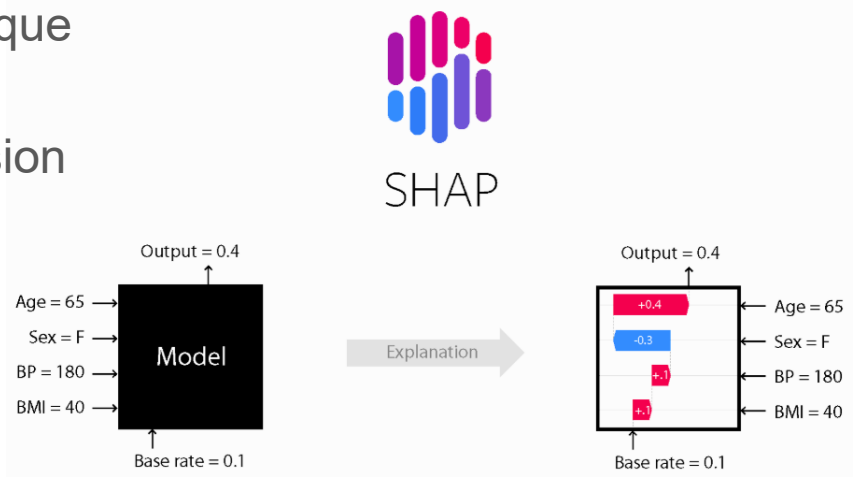
lusis CentraleSupélec

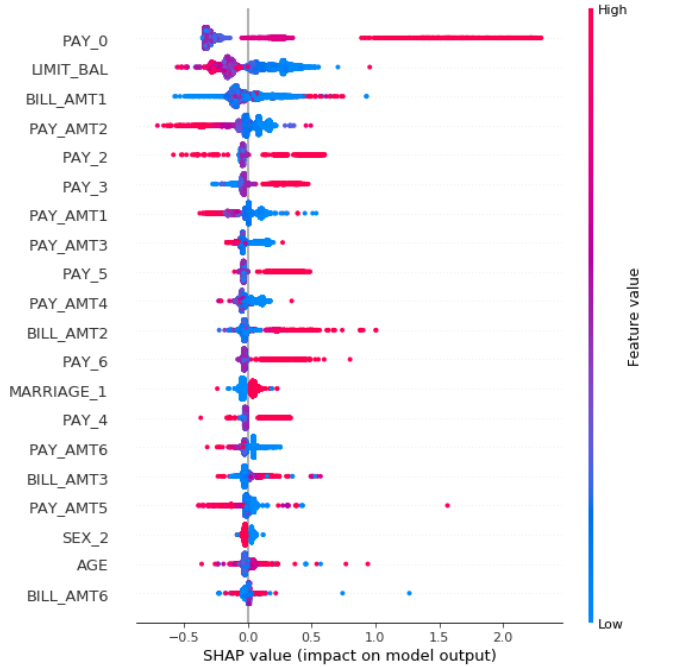
Fraude - Explicabilité

SHAP = SHapley Additive exPlanations.

Explications locale + théorie des jeux

Calcule une importance pour chaque caractéristique
à l'origine
d'une décision





lusis

CentraleSupélec

Trading

lusis CentraleSupélec

Axes de travail - Trading

“It’s one of the most difficult problems in applied machine learning,”
Ciamac Moallemi, professor at Columbia Business School

“Using machines to beat the markets is a really difficult challenge,
But I don’t think it’s impossible.”
Jon McAuliffe, professor at the University of California at Berkeley and chief investment officer at Voleon Capital Management LP

Source : Article Bloomberg, mai 2019

https://www.bloomberg.com/news/articles/2019-05-21/computer-models-won-t-beat-the-stock-market-any-time-soon

lusis CentraleSupélec

Axes de travail - Trading

Problème difficile mais possible

- Portefeuille de modèles sur les devises mis au point par Lusis

Approches peu satisfaisantes

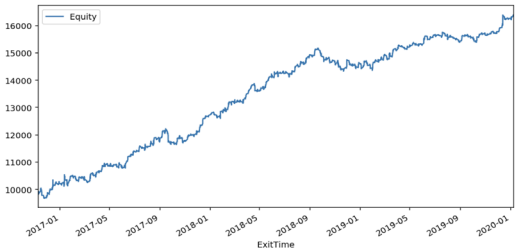
- sensibles aux conditions initiales
- non généralisables
- non transférable

Performance Report

	All	Long	Short
Net P&L	6,391.80	3,050.00	3,341.80
Gross Profit	27,939.50	10,163.50	17,776.00
Gross Loss	-21,547.70	-7,113.50	-14,434.20
Profit Factor	1.30	1.43	1.23
Total # of Trades	2,392.00	812.00	1,580.00
Number Winning Trades	1,289.00	440.00	849.00
Number Losing Trades	1,103.00	372.00	731.00
Percent Profitable	0.54	0.54	0.54
Avg Trade Win Loss	2.67	3.76	2.12
Avg Winning Trade	21.68	23.10	20.94
Avg Losing Trade	-19.54	-19.12	-19.75
Ratio Avg Win Loss	1.11	1.21	1.06
Largest Winning Trade	352.00	352.00	171.20
Largest Losing Trade	-189.70	-141.00	-189.70
Opened P&L	0.00	0.00	0.00

Performance/Risk Metrics

slope	2.7164	
r2	0.9493	
stderr	0.0128	
quality ratio	200.7603	
Max Drawdown	842.7	2018-11-21 00:00:00
Max Drawdown %	6.0 %	2018-11-21 00:00:00

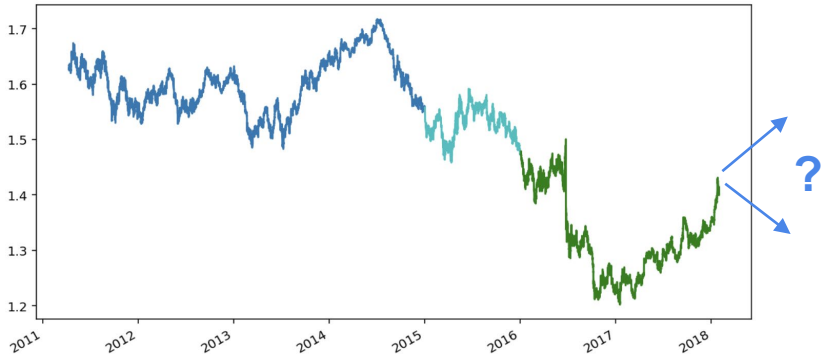


lusis CentraleSupélec

Axes de travail - Trading

Prévision des marchés financiers

- Grand nombre de publications
  - Google Scholar : 488 000 articles
  - souvent mises en échec ou inapplicables
- Problème
  - classification
  - régression



lusis CentraleSupélec

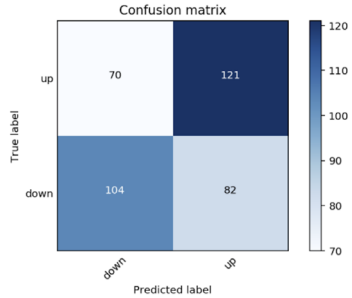
Axes de travail - Trading

Classification

- Prédire une hausse/baisse à un horizon

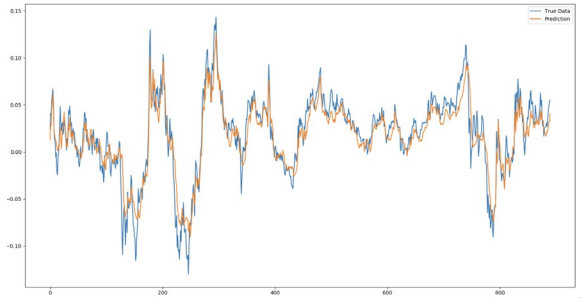
Accuracy : 0.59682

	precision	recall	f1-score	support
0	0.597701	0.559140	0.577778	186
1	0.596059	0.633508	0.614213	191



Régression

- Prédire l'amplitude d'une variation à un horizon



lusis

CentraleSupélec

Axes de travail - Trading

Pourquoi les marchés financiers

- couvre de multiples problèmes théoriques existant dans d'autres domaines
- immense variété d'approches applicables
- données abondantes facilement disponibles
- labels automatiques

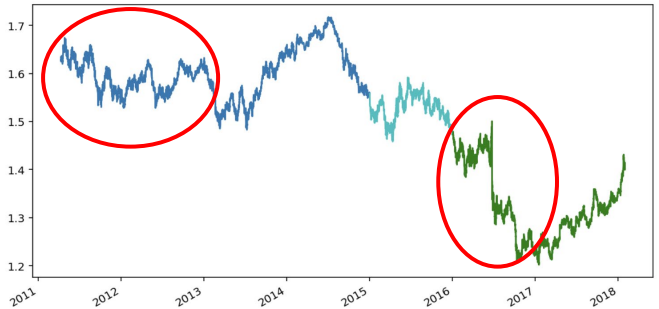
Enjeu

- apport à tous les domaines aux comportements **probabilistes**
  - biologie/médecine, météorologie, sismologie, consommation d'énergie, économie, assurance, gestion de stocks

Trading - Verrous

Domaine probabiliste

- Processus stochastiques non stationnaires
  - conditions changeantes au cours du temps (volatilité, acteurs, règles)

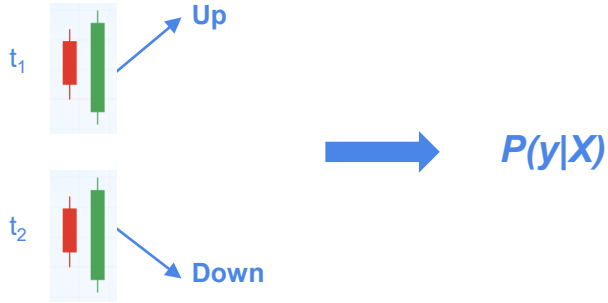


lusis CentraleSupélec

Trading - Verrous

Domaine probabiliste

- Incertitude à la fois aléatoire et épistémique
  - facteur humain, information partielle, influences exogènes
  - variabilité des labels → **probabilité conditionnelle**

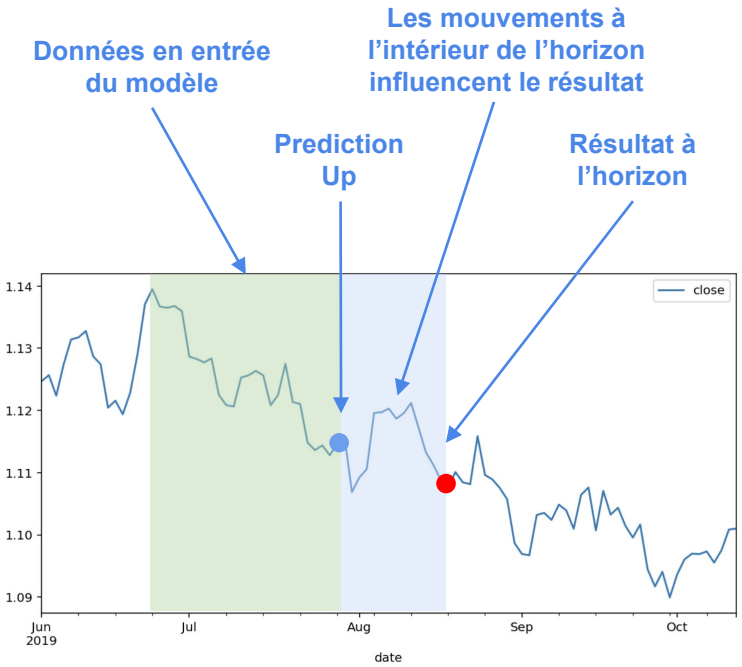


conditions identiques à deux instants différents, résultat différent

Trading - Verrous

Influence des événements post prédiction

- Problèmes propres aux time series
  - les mouvements à l'intérieur de l'horizon influencent le résultat



lusis CentraleSupélec

Trading - Verrous

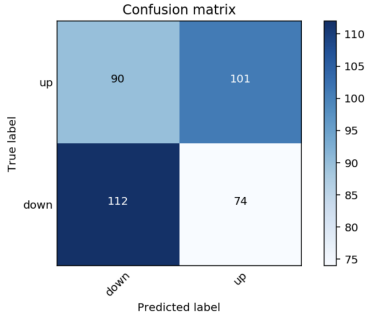
Sensibilité aux conditions initiales

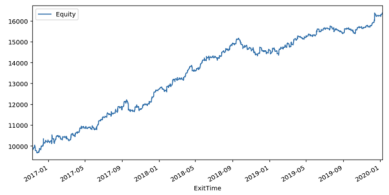
- modification de seed
- quelques données différentes

Seed = 1000

Accuracy : 0.56499
Classification Report

	precision	recall	f1-score	support
0	0.554455	0.602151	0.577320	186
1	0.577143	0.528796	0.551913	191

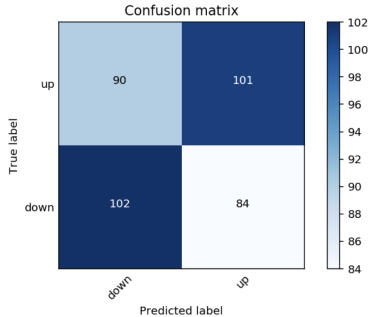


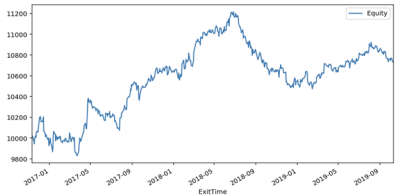


Seed = 1001

Accuracy : 0.53846
Classification Report

	precision	recall	f1-score	support
0	0.531250	0.548387	0.539683	186
1	0.545946	0.528796	0.537234	191



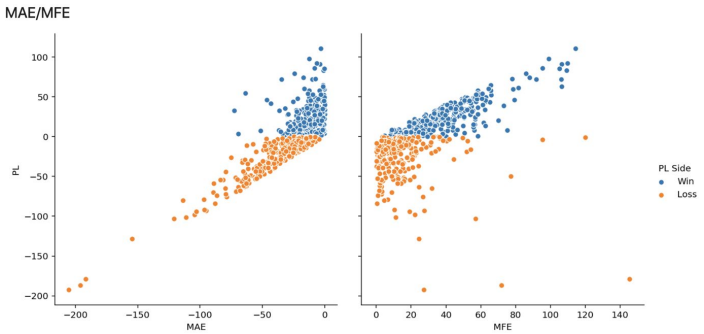


lusis CentraleSupélec

Trading - Verrous

Inefficacité des métriques classiques

- Bonne métrique ≠ modèle exploitable
  - trajectoires et amplitudes variables à l'intérieur de l'horizon de prédiction



- Implique
  - backtesting
  - création de métriques spécifique

Performance/Risk Metrics

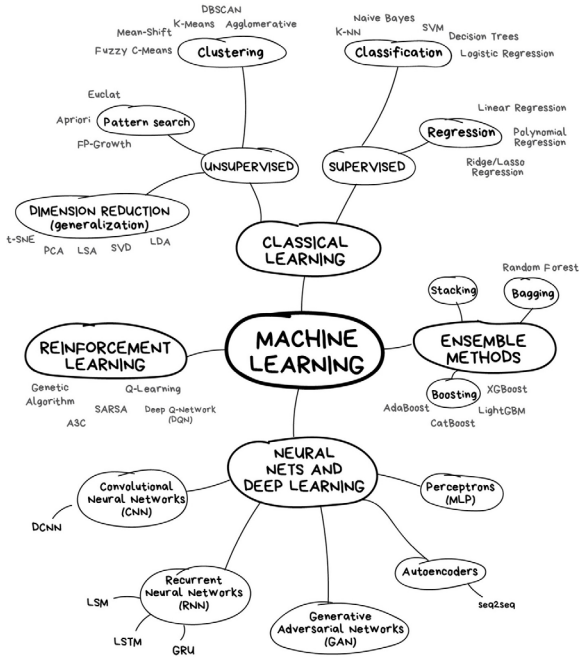
slope	2.7164	
r2	0.9493	
stderr	0.0128	
quality ratio	200.7603	
Max Drawdown	842.7	2018-11-21 00:00:00
Max Drawdown %	6.0 %	2018-11-21 00:00:00

$$QR = r^2 \frac{\beta_1}{\sigma}$$

lusis CentraleSupélec

Trading - apprentissage automatique

- Toute approche de machine learning
  - Random Forest, XGBoost
  - MLP, LSTM, CNN 1D, TCN, GAN
  - **Deep learning probabiliste, VAE**

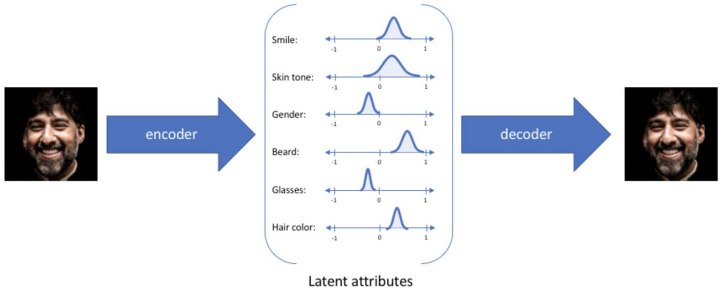


lusis CentraleSupélec

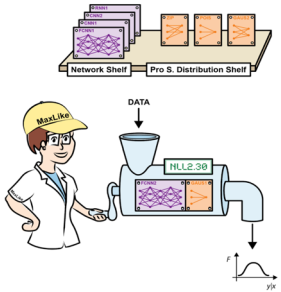
Trading - apprentissage automatique

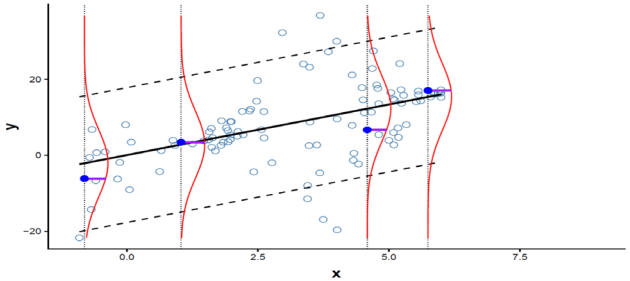
Les approches probabilistes

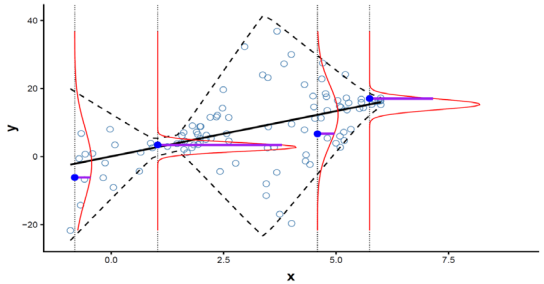
- VAE : extraction de features



- Probabilistic Deep Learning Models







lusis CentraleSupélec

Résumé

Fraude

- Enjeu Thèse : **explicabilité**
- CEI/Lusis : support à l'enjeu direct, données déséquilibrées, dérive conceptuelle et temporelle

Trading

- Enjeu Thèse : prédiction dans un **domaine probabiliste**
- CEI/Lusis : support à l'enjeu direct, exploitabilité, backtesting, sensibilité aux conditions initiales

lusis CentraleSupélec

Kick-off!

lusis CentraleSupélec

Chaire - premiers travaux

- Novembre 2019 à mars 2020
  - Détection de fraude et explicabilité
  - Détection de fraude - approches par graphes
  - Trading automatique
  - Constitution de playlists musicales (Qobuz)
- Proposition de sujets de stage de M2
- Rédaction des sujets de thèse
- Arrivée d'un nouveau partenaire associé

lusis CentraleSupélec

Le Futur

La collaboration CentraleSupelec / Lusis au delà de la chaire

- Lors de CEI et/ou de thèse(s)

Sur les grands enjeux du Machine Learning

- Domaine probabiliste
- Encodage de raisonnements
- Apprentissage avec peu de données
- Travaux sur la robustesse

lusis CentraleSupélec


Merci