CHAIRE MACHINE LEARNING FOR BIG DATA

Paris, le 13/03/2018

# *Offre de stage*

*Sujet : Stock Prediction: a Machine Learning Approach*

## *Encadrement*

Robert M. Gower, Alexandre Rubesam

## *Lieu et dates du stage*

Télécom ParisTech, 46 rue Barrault, 75013 Paris

Date de début du stage : April 2018

## *Équipe(s) d'accueil de la thèse*

département IDS, équipe Signal, Statistique et Apprentissage ($S^2A$)

## *Mots clés*

Stock return predictability, machine learning, Fama-Macbeth regression, Lasso, Kernel Methods

## *Sujet détaillé*

Can stock returns be predicted using publicly available information, such as market data (past prices and volumes), firm characteristics and accounting information from firm's financial statements, or forecasts from financial analysts? As more data have become available, and computational costs have decreased, the finance literature has uncovered hundreds of variables, or factors, purported to be predictive of future returns [1]. While many of these variables might be spurious results due to data mining, as argued by [4], the predictive power of many characteristics appears to persist.

Explicit attempts to predict stock returns using linear cross-sectional regression methods, such as [2] and [3], have provided evidence that (i) it is possible to build a linear factor model to predict with a surprisingly high level of accuracy the expected return of stocks and (ii) that the stocks with higher expected returns are unambiguously associated with lower risk, which directly contradicts basic models in Finance. However, [3] have shown that the returns to strategies that attempt to exploit these predictive signals using linear models appear to have diminished in the last 15 years.

While [2] and [3] use linear models estimated using simple Ordinary Least Squares, we propose to investigate whether standard machine learning methods and variable selection strategies improve the quality of the predictions. This will include using sparsity-inducing regularisation methods, kernel-

based methods, cross-validation and others. These methods and good practices in Machine Learning will be fundamental in avoiding selection bias and data leakage [5].

This internship will be based at Telecom Paristech but will be jointly supervised by Alexandre Rubesam at the IESEG School of Management in La Défense and Robert Gower at Télécom ParisTech. The internship will provide the perfect platform for a student who wishes to work in the data science of finance. The project will explore a readily available and rich dataset of stock returns and firm characteristics, obtained from the merged CRSP/Compustat database, comprising all stocks available in the U.S. markets. This amounts to over a hundred characteristics measured for thousands of stocks over a period of almost 40 years. Furthermore, the project is within the *Axis 5. Large Dimension Learning and Series/Time Data Streams* of the big data chair.

## *La Chaire Machine Learning for Big Data*

Le traitement statistique des masses de données convoque à la fois mathématiques appliquées et informatique, à travers une discipline en plein essor : le Machine Learning ou apprentissage statistique.

Créée en septembre 2013 avec le soutien de la Fondation Télécom et financée à hauteur de près de 2 M€ par quatre entreprises partenaires : Criteo, PSA Peugeot Citroën, Safran et BNP Paribas, la Chaire Machine Learning for Big Data est portée par le mathématicien Stéphan Clémençon, Enseignant-Chercheur, Professeur au sein du Département du Traitement du Signal et des Images à Télécom ParisTech.

Proposant cinq axes de recherche méthodologiques, enrichis par des applications industrielles concrètes, cette Chaire a pour objectif d'animer, en interaction avec ses partenaires, une activité de recherche de pointe en Machine Learning, ainsi que de proposer des programmes de formation.

La variété des données aujourd'hui disponibles (nombres, images, textes, signaux), leur grande dimension et leur volumétrie rendent souvent inopérantes les méthodes statistiques traditionnelles reposant sur le prétraitement humain et un long travail de modélisation. Le Machine Learning vise donc à élaborer et étudier des algorithmes, à vocation prédictive le plus souvent, permettant à des machines d'apprendre automatiquement à partir des données et à effectuer des tâches de façon performante.

Les avancées technologiques, l'omniprésence des capteurs (systèmes embarqués, objets connectés, Internet…) et l'explosion des réseaux sociaux s'accompagnent d'un véritable déluge de données, propulsant les sciences de l'information au centre du processus de valorisation des masses de donnés. En plus de la collecte et du stockage, l'enjeu est de pouvoir analyser ces données afin d'optimiser les décisions et mettre au point de nouvelles applications.

Au-delà du buzz médiatique dont il fait l'objet, le Big Data est donc un sujet stratégique majeur, au cœur d'enjeux économiques et sociétaux considérables. Son impact est désormais perçu dans presque tous les secteurs de l'activité humaine : de la recherche scientifique à la médecine en passant, entre autres, par la finance, le bâtiment, l'e-commerce, la défense ou les transports.

En savoir plus sur la Chaire, ses axes de recherche, ses activités, ses publications :

**http://machinelearningforbigdata.telecom-paristech.fr**

## *Profil du candidat*

Etudiant titulaire d'un master 2 recherche

- Statistics and Probability

- Machine Learning course

- A good level in programming (Java, C/C++, Python)

- A high level in English

## *Candidatures*

à envoyer à robert.gower@telecom-paristech.fr :

- Curriculum Vitae

- Lettre de motivation personnalisée expliquant l'intérêt du candidat sur le sujet (directement dans le corps du mail)

- Relevés de notes des années précédentes

- Contact d'une personne de référence

Les candidatures incomplètes ne seront pas examinées.

## *Référence*

[1] Harvey, Campbell R, Liu, Yan, & Zhu, Heqing, . . . . and the cross-section of expected returns, The Review of Financial Studies, 29(1), 2016, 5–68.

[2] Robert A. Haugen, Nardin L. Baker, Commonality in the determinants of expected stock returns, Journal of Financial Economics, Volume 41, Issue 3, 1996, Pages 401-439,

[3] Green, Jeremiah and Hand, John R. M. and Zhang, Frank, The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns, Volume 30, n 12, 2017.

[4] Harvey, Campbell R. and Liu, Yan, Lucky Factors (January 15, 2018). Available at SSRN: https://ssrn.com/abstract=2528780 or http://dx.doi.org/10.2139/ssrn.2528780

[5] Feng, Guanhao and Giglio, Stefano and Xiu, Dacheng, Taming the Factor Zoo (August 31, 2017). Fama-Miller Working Paper Forthcoming; Chicago Booth Research Paper No. 17-04.