the journal of financial data science
VOLUME 2, NUMBER 1 WINTER 2020

The four issues of the 2019 inaugural publication of *The Journal of Financial Data Science* by all metrics indicate the success of the journal. Four of the articles published in JFDS were in the top 10 most downloaded articles across the Portfolio Management Research (PMR) platform. This is quite an accomplishment considering that JFDS represented just one year of articles. After publication of the first issue, articles in JFDS were featured in an opinion piece on the challenges of implementing machine learning by David Stevenson ("Machine Learning Revolution is Still Some Way Off") published in the *Financial Times*. One of the articles in the inaugural issue is highlighted by Bill Kelly, the CEO of the CAIA Association, in an August 2019 blog ("Whatfore Art Thou Use of Alt-Data?") in *AllAboutAlpha*. The Financial Data Professional Institute (FDPI), established by the CAIA Association, will be adopting at least five articles from JFDS as required reading for their membership exams. As researchers in this space produce papers, our expectation is that the journal will be well cited.

In the first issue of Volume 2, there are nine articles which are summarized below.

"Machine Learning in Asset Management—*Part 1: Portfolio Construction—Trading Strategies*" is the first in a series of articles by Derek Snow dealing with machine learning in asset management. The series will cover the applications to the major tasks of asset management: (1) portfolio construction, (2) risk management, (3) capital management, (4) infrastructure and deployment, and (5) sales and marketing. Portfolio construction is divided into trading and weight optimization. The primary focus of the current article is on how machine learning can be used to improve various types of trading strategies, while weight optimization is the subject of the next article in the series. Snow classifies trading strategies according to their respective machine-learning frameworks (i.e., reinforcement, supervised and unsupervised learning). He then explains the difference between reinforcement learning and supervised learning, both conceptually and in relation to their respective advantages and disadvantages.

Global equity and bond asset management require techniques that also put effort into understanding the structure of the interactions. Network analysis offers asset managers insightful information regarding factor-based connectedness, relationships, and how risk is transferred between network components. Gueorgui Konstantinov and Mario Rusev demonstrate the relation between global equity and bond funds from a network perspective. In their article, "The Bond–Equity–Fund Relation Using the Fama–French–Carhart Factors: *A Practical Network Approach*," they show the advantages of graph theory to explain the collective

fund dynamics. They report that equity and bond funds have a significant exposure to the well-known Fama–French–Carhart factors. The network, according to Konstantinov and Rusev, is dynamically driven by equity funds with their centrality scores and risk factor exposure and can transmit and amplify system-wide stress or inefficiencies in the factor bets. Applying graph theory, they show that the return-based relationships between bond and equity funds are asymmetrical and the network is sufficiently clustered. In particular, equity funds connect the different clusters. The high book to market minus the low book to market factor in the Fama-French factor model is significant both on a single-fund level and as a web determinant. This suggests that asset managers should pay close attention to it when deriving asset allocations. Finally, the authors show how a machine learning approach can be applied to derive equity–bond allocations based on centrality scores, factor exposure, and hierarchical clustering of asymmetrically connected assets.

Securities and Exchange form 13F provides the public with a quarterly peek at the holdings of top asset managers. Investors have studied these data hoping to figure out what strategies top asset managers are pursuing. Generating investment positions based on these SEC 13F data is known as "alpha cloning." In "Constructing Equity Portfolios from SEC 13F Data Using Feature Extraction and Machine Learning," Alexander Fleiss, Han Cui, Sasha Stoikov, and Daniel M. DiPietro examine two novel quantitative approaches for constructing portfolio-generating models based on these data, as well as guidelines for feature extraction. The first approach uses machine learning (logistic regression and XGBoost) to forecast stock returns based on features that are extracted from 13F holdings. The stocks that are then used to assemble a portfolio are those stocks that are forecasted to have the best performance. The second approach involves selecting top performing funds and replicating their holdings to generate a new, aggregate portfolio. The authors find that both approaches outperformed an S&P 500 benchmark when assessed with backtesting.

Financial data analysis has two primary goals: making a prediction and obtaining information that helps in understanding a system. Along these lines, Yimou Li, David Turkington, and Alireza Yazdani pose two distinct questions in their article "Beyond the Black Box: *An Intuitive Approach to Investment Prediction with Machine Learning*." First, can machine learning algorithms detect patterns in financial data that lead to superior investment returns? Second, how do the algorithms process the data to form predictions? In addressing this second question, which is often neglected in the literature, the authors propose a set of interpretability metrics (collectively named a "model fingerprint") to decompose the behavior of any model's predictions into linear, nonlinear, and interaction effects among predictors. They also demonstrate how to decompose the model's predictive efficacy into these components. Next, they explore these concepts in the context of foreign currency investing. Li, Turkington, and Yazdani provide a case study applying random forest, gradient boosting machine, and neural network algorithms to predict one-month forward currency returns. Their findings for this case study are that machine learning models can be used to reliably identify known effects and to find new nonlinear relationships and interactions.

Although machine learning techniques have been revolutionary in recent years due to their unparalleled proficiency at solving many tasks previously thought impossible to automatize, applications in finance such as price forecasting, risk analysis, and portfolio construction have been impaired due to the unique characteristics of financial data. As a result, overfitting is an inevitable phenomenon when applying deep learning techniques to financial data due to the relative scarcity of available historical data and the ever-changing nature of financial series. One way to deal with this problem is data augmentation. More specifically, Fernando De Meer Pardo and Rafael Cobo López show how generative adversarial networks (GANs) can be used to deal with this shortcoming in their article "Mitigating Overfitting on Financial Datasets with Generative Adversarial Networks." Through adversarial training, GAN—a type of neural network architecture that focuses on sample generation—can implicitly learn the underlying structure inherent to the dynamics of financial series. By doing so, it can acquire the capacity to generate scenarios that share many similarities to those seen in the historic time series. The data augmentation technique proposed by the authors is the Wasserstein GAN with gradient penalty.

De Meer Pardo and Cobo López demonstrate how training deep models on synthetic data mitigates overfitting, improving the model's performance on test data when compared to models trained solely on real data.

In the formulation of investment strategies, there are four flaws: (1) domain-specific knowledge barrier, (2) budgetary constraints and confidentiality restrictions, (3) inability to monetize the value of data, and (4) backtest overfitting. Because the majority of data scientists cannot overcome one or more of these barriers, investment opportunities are not researched by the data science community. In recent years, there has been a movement toward crowdsourcing research through the development of backtesting platforms and investment tournaments. A backtesting platform provides unaffiliated researchers (i.e., non-employees) with data, computing resources, and software for the backtesting of various investment strategies. They enable undirected crowdsourcing, whereas researchers are not told what kind of strategies they should develop. The output delivered by the crowd is trades or portfolios, not forecasts. In a tournament, an organizer formalizes an investment challenge in terms of a forecasting problem, hence addressing the experience and data barriers. Tournaments enable directed crowdsourcing, whereas researchers are told the exact problem they must focus all their effort and attention on. Tournaments allow data scientists without an investment background to contribute forecasts to a systematic asset manager. In "Crowdsourced Investment Research Through Tournaments," Marcos López de Prado and Frank J. Fabozzi discuss the merits of tournaments as a viable crowdsourcing paradigm that fosters the research of investment strategies while avoiding the four flaws of the silo paradigm: (1) domain-specific barrier, (2) data barrier, (3) data monetization, and (4) backtest overfitting. As the authors point out, tournaments do not obviate the need for financial expertise. Rather, deep financial expertise is required to cast the problem crowdsourced through the tournament.

A major drawback of traditional time-series models for predicting equity price movements is that they restrict their ability to learn latent patterns in the data. Due to their universal approximation properties, artificial neural networks (ANNs) offer more flexibility in understanding stock price movements. However, most experiments applying neural networks require a considerable amount of time to search a suitable network and subsequently train the network. In "Deep Learning Classifier with Piecewise Linear Activation Function: *An Empirical Evaluation with Intraday Financial Data*," Soham Banerjee and Diganta Mukherjee develop a deep multilayer perceptron (MLP) formalization for stock trend prediction using minute-by-minute price–volume data and Twitter data, with the belief that patterns do exist in equity price movements and that these can be captured more effectively by neural networks than conventional machine learning models, especially with large datasets. Banerjee and Mukherjee explain the theoretical properties and advantages of their proposed model. To illustrate their model, they apply it to a large high-frequency dataset on selected bank shares from the Indian stock market.

Since the Chinese A-shares market is characterized by heavy retail investor participation and driven by retail trading sentiment, it is essential to understand the sentiment of retail investors. In "Teaching Machines to Understand Chinese Investment Slang," Mike Chen, Jaime Lee, and George Mussalli discuss their approach of combining a recently developed natural language processing (NLP) technique with online Chinese-language investment blogs favored by Chinese retail investors to understand their views toward various A-share stocks. The Chinese language poses several distinct challenges that makes application of traditional NLP techniques not particularly fruitful. Utilizing the latest, neural-network based NLP approach, the authors can overcome these challenges and gain insights into the A-shares market. The authors also show how adapting the standard neural network underpinning their NLP approach makes it more suitable for application to the financial domain.

Geolocation data can be used to capture the relative influx and outflux of consumers from department stores. This is an extremely useful feature in training machine learning models which can be used in a popular hedge fund trading strategy known as pairs trading. In their article, Jim Kyung-Soo Liew, Tamas Budavari, Zixiao Kang, Fengxu Li, Xuzhi Wang, Shihao Ma and Brandon Fremin use Under Armour (UA) and Nike stock market price and volume to illustrate their methodology for

pairs trading. They argue in "Pairs Trading Strategy with Geolocation Data—*The Battle between Under Armour and Nike*" that there is the temptation in the era of big data to toss every feature into a model and assume that the model will always be capable of determining which features are most important. However, the model will usually perform better when trained on a subset of distinct features. The two principal findings of their study are (1) geolocation information is an important factor in pairs trading strategy between UA and Nike (based on the results of feature selection and prediction accuracy for price ratio change), and (2) ensemble methods are an effective means of eliminating bias in machine learning models because they use the output from a series of independently trained submodels to generate one balanced result.

Francesco A. Fabozzi
Managing Editor