Volume 4 | Number 2 | Spring 2022

the journal of financial data science



Of the eight articles in this issue, four are contributed by practitioners from Allspring Global Investments, JP Morgan Chase, Vanguard, and a Causality Link-Turnleaf Analytics team. There is a joint academic-practitioner (MIT-MSCI) contribution, and three academic contributions from Princeton University, University of Toronto, and Towson University.

The organization of companies into groups, referred to as industry classification, has broad applications in finance and economics. Investors and financial analysts, for example, use industry classification for investment management and competitive industrial diligence. The classification systems are based on different criteria such as production, revenue, earnings, and market perception. Several industry classification systems have been developed since 1937, with the North American Industry Classification System (NAICS): Global Industrial Classification System (GICS), Industry Classification Benchmark (ICB), and Text-Based Network Industry Classifications. Although the ICB does consider market perception as one of the criteria in its industry classification system, there is no system that uses market perception as its sole criterion. In "An Artificial Intelligence-Based Industry Peer Grouping System," George Bonne, Andrew W. Lo, Abilash Prabhakaran, Kien Wei Siah, Manish Singh, Xinxin Wang, Peter Zangari, and Howard Zhang formulate a data-driven industry peer grouping system that clusters similar companies at different levels of granularity and develop an interactive tool to visualize clusters and nearest neighbors of companies. The authors employ artificial intelligence to extract features from various data sources and learn relationships that can identify companies that are similar in terms of their risk–return profile. The features that the authors report contributed the most in grouping companies are historical returns correlation, GICS, 10-K reports, and fundamental factors (e.g., size, momentum, and debt-to-asset ratio).

Sentiment information, derived from views of financial market participants, has played an important role in forecasting financial market conditions. This information reflects a trend in financial data science that has focused on micro-level data obtained from domains such as social networks to enhance investment performance. The improvement of natural language processing (NLP) methods has made it possible to obtain sentiment information from numerous sources of social media as well as articles and reports. Di-Jia Su, John M. Mulvey, and H. Vincent Poor, in their article "Improving Portfolio Performance via Natural Language Processing Methods," start by describing the concepts of NLP and then review the application of NLP to portfolio models via a modern version of sentiment analysis. They describe the advantages of using information from Twitter along with NLP for constructing a portfolio of stocks. Based on their empirical analysis, they find that their proposed techniques perform well, especially during unusual events such as the COVID-19 pandemic.

In their article "Harvesting Multi-Asset Carry, Value, and Momentum: Work Smarter, Not Harder," Brian Jacobsen and Matthias Scheiber investigate whether and when asset managers who pursue a systematic multi-asset strategy should care about carry, value, and momentum. They demonstrate that rather than allowing the signals to determine what trades should be made, a better way to handle trades is to just let

the journal of financial data science

signals inform trades. Using daily data and refreshing their trading rule monthly, they find that different signals have different degrees of persistence and decay across two dimensions (holding period and decay attributable to delaying trading on a signal). As a better strategy than trying to trade signals directly (ranking assets by a signal), signals should be treated as explanatory variables as part of a classification problem. Jacobsen and Scheiber employ classification trees for learning the pattern of horizon decay and delay decay and provide insightful context for monitoring trades.

Institutional investors typically execute large positions on a daily basis. Often a wide variety of algorithmic trading programs are used to trade large positions with various levels of efficiency. A major issue is the impact of trading large positions on the traded stock's share price. Various benchmarks have been adopted by market practitioners to assess the effectiveness of algorithms in trade execution—volume-weighted average price (VWAP), percentage of volume (POV), and time-weighted average price (TWAP). Because the execution of a large position may take several trading days to complete, the concern is that the market microstructure may change during the execution of the order. In "Optimal Trading Algorithms under Regime Switching," Moustapha Pemy addresses this situation assuming that the stock price follows a regime-switching model and investigates how trading algorithms that track market benchmarks perform. Formulating the problem as a discrete-time stochastic optimal control problem with resource constraints, Pemy proposes trading algorithms that break the execution order into small pieces and executing them over a predetermined period of time to minimize the overall execution shortfall or exceed the overall market VWAP. Numerical simulations with real market data are reported to illustrate how the proposed trading algorithms not only track major market benchmarks, but can potentially exceed those benchmarks.

In an effort deeply rooted in probability theory and therefore well suited to express uncertainty, many domains have applied Bayesian networks to incorporate knowledge about why things happen (i.e., what causes what). A stock market crash, for example, will likely cause a spike in implied volatilities, but the opposite is not necessarily true. That is, Bayesian networks incorporate causation as a primary concept and correlation as a derived one. In "Building Probabilistic Causal Models Using Collective Intelligence," Olav Laudy, Alexander Denev, and Allen Ginsberg show how to derive Bayesian networks about variables of interest to asset managers by applying natural language processing (NLP) techniques to millions of digitally published news articles in which different authors express their views on the future states of economic and financial variables, as well as geopolitical events. In addition to deriving forward-looking, point-in-time views on various variables that are of interest to asset managers in enhancing their understanding of the current drivers of an economy, the Bayesian networks derived can be fed to an optimization engine to construct a forward-looking optimal portfolio given constraints imposed by the asset manager.

The field of interpretability is concerned with identifying the top drivers of a model's output. In the lending industry there are three reasons why interpretability is important. First, US lending laws require creditors to provide potential borrowers who are denied credit with reasons for their decline decision. Second, lenders who use a model for making lending decisions need to understand the reasons underlying model prediction. Finally, interpretability allows identification of model bias and reinforces stakeholders' trust in the model. In "Interpretability of Machine Learning versus

the journal of financial data science

Statistical Credit Risk Models," Anand K. Ramteke, Pavan Wadhwa, and Monica Yan compare the interpretability of a machine learning XGBoost algorithm versus a logistic model when predicting the probability of default for a credit card customer. The conclusions of the authors are that while reason codes generated from an XGBoost model and a comparable logistic model are somewhat similar, reason codes generated by XGBoost are more trustworthy from the applicant's perspective. Additionally, they find that the nonlinearity of XGBoost is unlikely to have a significant impact on reason code(s).

The volatility surface depicts the implied volatility of an option on an asset as a function of the option's strike price and maturity. It is an important tool for pricing and hedging derivatives. When market data are incomplete, it becomes necessary to estimate missing data points on partially observed surfaces. In "Variational Autoencoders: A Hands-Off Approach to Volatility," Maxime Bergeron, Nicholas Fung, John Hull, Zissis Poulos, and Andreas Veneris demonstrate how a deep learning approach using variational autoencoders can be used to construct a complete volatility surface when only a small number of points are available without making assumptions about the process driving the underlying asset or the shape of the surface. A variational autoencoder is a special type of neural network where the output layer is the same as the input layer and whose objective is to determine probability distributions that can be sampled from in order to create data that is indistinguishable from historical data. Using foreign exchange data, the authors demonstrate their proposed approach and empirically show that pooling volatility surface data from multiple currency pairs improves the results.

Researchers have yet to fully explore individual health care cost risks and their role in personalized investment and retirement planning strategies. This analysis is critical for individuals in countries that do not have a universal health care system because uncertainty about future health care costs can have a material impact in formulating a retirement investment strategy. While many factors impact health care costs, one of the key factors is the individual's health status and, therefore, it is important to categorize individuals into meaningful health risk types in formulating a personalized investment and retirement strategy. Using individuals' self-rated health state categorization has been the traditional approach employed. In "Health State Risk Categorization: A Machine Learning Clustering Approach Using Health and Retirement Study Data," Fu Tan and Dhagash Mehta propose an objective and data-driven machine learning-based approach (the K-modes clustering method) to algorithmically cluster heath state risk based on the Health and Retirement Study, the most widely used US household surveys on older Americans. In contrast to the conventional five-state self-rated health state categorization, which Tan and Mehta find offers only a weak association with some observed health behaviors and medical care utilization, they note that the machine-learning approach offers an objective, interpretable, and practical health state risk categorization. Moreover, the authors illustrate the difficulty in predicting out-of-pocket costs derived from self-rated health status and explain how categorizations obtained by applying machine-learning based methods can produce more accurate health care cost estimates for personalized retirement planning.

Francesco A. Fabozzi
Managing Editor