On Transaction-Based Metrics as a Proxy for Profitability of Financial Asset Recommendations

JAVIER SANZ-CRUZADO, RICHARD MCCREADIE, CRAIG MACDONALD, and IADH OUNIS, University of Glasgow, United Kingdom

NIKOLAOS DROUKAS, National Bank of Greece, Greece

The use of recommender systems to assist in the provision of financial asset and portfolio recommendations to investors is increasing, spanning a wide range of algorithms and techniques. Several strategies have been devised for the evaluation of financial asset recommendations, with the two most prominent strategies measuring (a) the money customers could obtain if they followed the recommendations (profitability-based evaluation) and (b) the ability of models to predict future customer investments (transaction-based evaluation). If customers are effective investors, we would expect these two perspectives to be positively correlated. In this paper, we perform experiments over a new large-scale financial recommendation dataset with real customer investment transactions to validate this assumption. Surprisingly, we find that transaction and profitability-based metrics are in fact negatively correlated and moreover, algorithms that actively try to learn from past customer transactions lose money over the mid-term. A thorough analysis of model performance and customer transaction patterns over time illustrates that this is due to a set of confounding factors, namely: customers failing to beat the market with their investments; a tendency for the customers to favour different investment lengths; and the impact of global events such as the Covid-19 pandemic.

1 INTRODUCTION

The digital transformation of financial organisations, along with the huge increase in the data available to them has created a need for automated analytic and artificial intelligence tools for the financial domain [31]. Under this group of applications, financial asset recommender (FAR) systems have a prominent role, as they are increasingly being used to provide financial investment options to customers and drive automated trading algorithms [14]. FAR algorithms seek to identify a list of financial assets for a customer, ranked by their suitability for investment by that customer. However, the suitability of a financial asset for a customer does not only depend on that customer's preferences (as is the case on movie or music recommendation [27]), but also on external factors, such as the short or long term market returns, the value of the currency used in the trading process, and the impact of governmental regulations or global events [40]. In addition to these external factors, FAR systems need to consider factors related to the customers, like the alignment of the recommendations with their financial risk tolerance. These complexities show that the financial domain is markedly different to traditional recommendation domains, and as such we cannot assume that observations from those domains will generalize to the finance space.

Developing effective strategies for evaluating FAR solutions is fundamental for the advancement of the field, as this enables both the sound comparison of solutions and is also a requirement for training many of those solutions. However, the FAR field is clearly fragmented when it comes to evaluation, with many competing methodologies having been proposed [6, 16, 17, 21, 39]. In this work, we focus on two of these methodologies, namely: *profitability-based evaluation* [7, 21, 23, 39] and *transaction-based evaluation* [6, 16, 18, 38]. Profitability-based evaluation use metrics like return on investment to quantify whether investors would make money by investing in the recommended assets. Meanwhile, transaction-based evaluation derives performance scores by comparing the recommended assets against what the customers chose to invest in (using ranking metrics such as nDCG). In theory, if customers invest intelligently, and thereby profit from the market, a high correlation between these two metrics would be expected – making transaction-based evaluation superior, as it would not only be able to measure profitability, but also incorporate customer preferences.

However, given the complexity of the finance domain, we cannot assume that this hypothesis holds. Hence, in this paper, we compare profitability and transaction-based evaluation methodologies over a new large-scale financial investment dataset to validate this hypothesis. Specifically, we first implement a diverse set of 12 FAR approaches using a range of pricing and transaction features, providing a representative sample of popular solutions. We then evaluate these solutions over a 1-year period using both profitability and transaction-based metrics to see if those metrics are positively correlated, followed by an in-depth analysis of the factors that influence the value-add of real investment transaction data (and hence transaction-based evaluation and models based on this data). The primary contributions of this paper are as follows:

(1) Evaluation of 12 FAR approaches over a novel and recent financial pricing and transaction dataset (that spans the Covid-19 pandemic period), including profitability prediction, personalized collaborative filtering and hybrid strategies that are rarely compared.
(2) We demonstrate that approaches that leverage real customer transaction data perform poorly, and that profitability and transaction-based evaluation metrics are negatively correlated.
(3) Through an in-depth analysis of model effectiveness and customer investment behaviour, we show that customer transactions are problematic as a source of evidence on the suitability of financial assets, since customers investments often lose money in the mid-term; investment success is dependent on the customers (largely unknown) asset holding time; and asset profitability is strongly influenced by global events such as the Covid-19 pandemic.

2 NOTATION AND *FAR* TASK DEFINITION

FAR systems are concerned with two groups of entities: the customers/users who are interested in investing (which we shall denote as $u \in \mathcal{U}$) and the financial products/items they can invest in (that we denote as $i \in \mathcal{I}$). At a given time, t, customers can purchase or sell the different financial assets at a given price, $\text{price}(i, t)$, that varies over time according to supply and demand. We define as $I_u(t) \subset I$ as the set of financial assets a customer u has interacted with at some point before t. We divide this set into two subsets: $I_u^+(t)$ and $I_u^-(t)$, representing the assets that u has bought or sold before t. The goal of a FAR system is then to rank the available financial assets, $R_u \subset \mathcal{I} \setminus \mathcal{I}_u(t)$ that are unknown to the customer u (i.e. those they have not interacted with in the past), based on their investment suitability (where an asset is suitable if the price at selling point is greater than the price at purchase time).

3 *FAR* APPROACHES

The financial domain has inspired a wide variety of techniques for suggesting products on which to invest, based on many sources of information, including investment transactions, pricing data, news and social networks, among others. In our later experiments we will evaluate 12 different recommendation approaches from the literature, hence we summarize the main classes of FAR approach below for reference.

Profitability Prediction algorithms are non-personalized algorithms [40] that aim to predict the future price of products or related key performance indicators about the financial assets to be recommended [7, 29, 30]. Most methods under this category are only based on the pricing data of the assets. For example, Yang et al. [36] combines several regression algorithms like neural networks and decision trees for estimating asset profitability. Other approaches exploit similarities between the pricing time-series of multiple assets to predict key price indicators [8, 23, 39]. Past works have also examined the integration of additional sources of information, such as news [9, 29] or social media [30, 32, 35], providing evidence of major events and trader's views about assets.

Collaborative filtering recommenders are based on the principle that similar customers invest on similar assets, and similar assets are acquired by similar people [25]. These methods require interactions between customers and assets (for example, transactions from investment logs). Some notable collaborative filtering methods which have been used for financial asset recommendation include the work by Lee et al. [16], who introduce a fairness-aware matrix factorization method for suggesting loans to fund, and the work by Zhao et al. [38], combining probabilistic matrix factorization with portfolio optimization techniques to suggest startups on which to invest.

Content-based recommenders extract the investment preferences of customers based on analysis of assets that they have previously invested in, with the aim of identifying similar products that those customers have not seen before [25]. As a representative algorithm in the financial domain, Luef et al. [17] design an algorithm that first builds customer profiles according to features like the market sector or life cycle of the enterprises on which customers invested previously. Then, the customer profile is matched with the financial products using Jaccard similarity to rank those products.

Demographic recommenders consider personal information about customers as a means to identify similar investors [25]. In stock recommendation, Yujun et al. [37] propose one of these methods, formulated as a user-based kNN on which, instead of finding similarities between past investments, the answers to a risk assessment questionnaire are considered to determine whether pairs of customers are similar to each other or not.

Knowledge-based systems apply specific domain knowledge about how different items meet user needs and preferences [3]. Several approaches have been proposed under this category for producing financial recommendations. Gonzalez et al. [10] propose an investment portfolio advisor based on fuzzy logic for matching customers and assets according to psychological and social characteristics, while Musto et al. [19–21] design investment portfolio case-based recommendation algorithms that factor in the risk aversion level of customers.

Social-based recommenders [34] consider social connections (like follow relations in networks like Twitter) to generate recommendations. For instance, Luef et al. [17] propose a trust-aware strategy, where customers are required to specify other investors they trust, who could then be leveraged to identify assets to recommend.

Hybrid algorithms [4] combine several techniques and information sources to provide recommendations. On financial asset recommendation, Chalidabhongse et al. [6] propose an adaptive model to learn from past investments, financial technical indicators and demographic data about the customers. Meanwhile, Matsatsinis et al. [18] combine collaborative filtering and multi-criteria decision analysis to generate a utility score for equity fund recommendation. Finally, Luef et al. [17] propose a hybrid method that combines both content-based and knowledge-based components.

As we can observe in this short review, many diverse algorithms have been proposed for financial asset recommendation. However, what approaches are the most effective is still largely unknown because approach types are rarely compared, and as we will discuss next, there is little agreement on how success should be defined for these approaches. In our later experiments, we compare 12 distinct approaches, drawn from the profitability prediction, collaborative filtering, demographic, and hybrid classes (the other classes are omitted due to either cost or data unavailability).

4 EVALUATING FINANCIAL RECOMMENDATIONS

In any research environment, a commonly agreed upon and experimentally sound strategy for evaluating the different approaches is critical. For classical recommendation tasks, such as movie recommendation, researchers and practitioners

Table 1. Comparison of recommendation techniques and associated evaluation strategies reported across research papers.

FAR Approach	**Evaluation Method**			
	Transaction-Based	Performance-Based	Expert-based	Hybrid
Collaborative filtering	[16, 18, 38]		[17]	
Content-based			[17]	
Knowledge-based		[19–21]	[10, 17]	
Social-based			[17]	
Profitability-based		[7, 8, 23, 29, 39, 30, 32, 35–37]		[39]
Hybrid	[6]		[17, 33]	

have found that implicit interactions like clicks on movies, or explicit ratings function well as a surrogate for whether a user is satisfied with a recommendation. However, in the financial domain, whether a customer will be satisfied by an asset is more difficult to measure, since it depends on more than the inherent properties of the asset, such as market conditions and the amount of time the customer wants to invest for. This complexity has resulted in a range of competing methods, namely: transaction-based evaluation; profitability/performance based evaluation; expert-based evaluation; as well as hybrid methods that combine one or more of these methods with additional aspects such as the customer risk appetite or asset class preferences. This lack of a standardized and agreed upon evaluation method is problematic when evaluating systems, as prior works tend to only use one or in rare cases two of these methods, as we illustrate in Table 1. Hence, there is a clear need for research efforts toward the understanding and standardization of the use of these methods. Below we summarize these evaluation methods and then discuss transaction and profitability-based evaluations in more detail, as these are the focus of our study.

Profitability/Performance Evaluation: In the specific case of financial asset recommendation, the real-world performance of a recommended asset is a natural proxy to customer satisfaction, since it aligns with the core goal of the customer (to maximise profit). However, profitability is complex to measure, since even if we have future pricing data, when the customer will 'cash-out' is unknown. Metrics used under this type of evaluation attempt to quantify the benefits (or losses) that a customer might obtain by investing in a recommendation. This is usually achieved by directly computing key performance indicators like the net profit and return on investment for a particular time horizon [8, 19, 20, 23, 32], such as 6 months in the future. The primary limitation of this type of metric is that it ignores the customer's situation, and so cannot personalize to them or consider their appetite for risk [39].

Transaction-based Evaluation: For non-cold-start investors, their past transaction history containing buy and sell actions may be available. It has been hypothesised that these transactions are a good alternative measure for customer satisfaction, as if the customer chose to invest in something then this is a strong signal that they like it. Moreover, under the assumption that customers invest intelligently and hence make a profit, metrics based on these transactions should positively correlate with profitability metrics. In this way it is theorized that transaction-based evaluation is a superior method if such transaction data is available. Transaction-based evaluation re-uses metrics from the information retrieval domain, such as precision [6, 18, 38], recall [18, 38] and normalised discounted cumulative gain (nDCG) [38], among others. Notably, transaction-based evaluation is equivalent to classical recommendation evaluation using explicit interactions [5, 11], where buy transactions are similar to positive ratings and sell transactions are similar to negative ratings. In practice, transaction-based evaluation using non-synthetic data is under-researched in the literature, primarily due to the lack of publicly available data (since logs of individual customer investments are considered sensitive).

Expert-based Evaluation: This method involves the participation of domain experts to establish what constitutes a good recommendation for a customer. Experts have a deep understanding of the prevailing market conditions, historical asset performances and the different factors which might influence the market evolution. Consequently, they are capable of providing advice on the long and short term viability of investments. However, it can be difficult and costly to obtain access to such experts. There are many ways to leverage expert judgments for evaluation, such as comparing the recommended assets with the expert asset selection using accuracy metrics like precision, recall or F1 [10]. Past works have also experimented with manually showing recommendations to experts for their assessment [33].

Hybrid Methods: Due to the multiple factors that influence what a customer might value in an asset, hybrid approaches have been proposed that combine multiple asset, customer and market features together to produce a single score for an asset. A simple example is the Sharpe Ratio [39], which represents a ratio between the profitability of a product and its volatility (risk). However, as we show from our literature survey in Table 1 these hybrid measures are rarely reported in the literature, likely due to the additional complexity when attempting to interpret them.

Of these four classes of evaluation method, profitability/performance evaluation is by far the most frequently reported as shown in Table 1 (likely due to the high availability of asset pricing data). However, this method has clear limitations due to its customer-agnostic nature. On the other hand, transaction-based evaluation intuitively appears a more well-rounded metric, as it is based on real customer interactions. However, there are a number of caveats around whether this type of evaluation would be effective in practice – since it assumes the customers are effective investors. Hence, in the remainder of this paper we investigate to what extent this is the case, by comparing how profitability and transaction-based metrics perform over a real pricing transaction dataset when evaluating a wide range of FAR approaches.

5 EXPERIMENTAL SETUP

In order to understand the utility of the profitability and transaction-based evaluation strategies, we perform a comparison study of 12 FAR approaches over a new large-scale financial asset pricing and transaction dataset. In this section, we summarize this dataset and its statistics, the cleaning techniques employed, how we split this dataset into temporal settings, as well as discuss the FAR approaches tested and evaluation metrics used. We conclude the section with an overview of the performance of the FAR approaches under a range of metrics, before our primary analysis in the following section.

5.1 Dataset

Pricing and Transaction Data: One of the novelties of this work is that we compare both (personalized) collaborative filtering and demographic-based recommenders to (un-personalised) content-based recommenders that are more common in the financial domain. To enable this comparison, we require a dataset that provides (private) financial transaction data. Hence, we use a proprietary dataset, provided by a large European financial institution. This dataset represents a snapshot of the Greek market, and covers a range of different securities: stocks, bonds, mutual funds and other banking products for the period between January 2018 and March 2021. In addition to security pricing data for that period, it also includes investment transaction logs (asset buy and sell actions) handled by the institution. Table 2 summarizes the properties of the dataset. [1]

Dataset Cleaning (Pre-Split): In order to avoid outliers or invalid values in our data, we preprocess our dataset.
Historical pricing data: Since pricing data collection is not perfect, it is common to find time-series gaps, where market data is missing for some assets over short periods of time. While these gaps are realistic, they add a confounding variable when performing analysis, and as such we clean our data to minimise their impact. First, we remove any assets with gaps in the time series greater than a week. Second, we fill gaps less than a week by applying a moving average over

[1] A sample of the transaction data used here is available for free at https://marketplace.infinitech-h2020.eu/assets/nbg-datasets

Table 2. Dataset description.

Market data		Customer data	
Property	Value	Property	Value
Unique assets	5,371	Unique customers	52,390
Assets with investments	2,025	Transactions	313,004
Price data points	1,768,128	Acquisitions	269,931
Average return (by assets, whole period)	23.67%	% Average return (by customers, whole period)	18.41%
% profitable assets	53.08%	% customers with profits	58.00%

the previous five days. Finally, we remove any assets having a closing price equal to 0 within their time series – we assume that acquiring or selling an asset involves monetary exchange, and zero-valued assets can lead to profitability values equal to infinity.

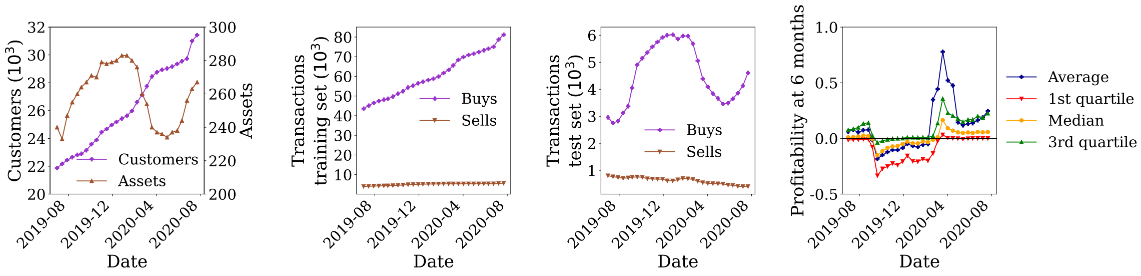
Transaction data: Collaborative filtering algorithms typically receive as input a rating matrix – where each user-item pair is represented by a numerical value representing the interest of the user on the item. In our experiments, we consider that a customer has interest on a financial asset ($Rel(u, i) = 1.0$) if she has acquired instances of the asset. Otherwise, it is considered that the customer is not interested on that product ($Rel(u, i) = 0.0$). Whether a customer is considered to have acquired instances of an asset for the purposes of training/testing each model is based on the temporal split, discussed next.

Dataset Temporal Splitting: This dataset spans 39 months (just over 3 years). The effectiveness of different recommendation algorithms will naturally vary as market conditions change (as we will demonstrate later). Hence, it is important to examine how performance varies over time if we are to gauge more accurately when and where different recommendation strategies succeed and fail. To this end, we divide our dataset into 29 distinct variants, each representing a recommendation setting for a different point in time. Each variant defines a time point when recommendations are produced $t \in T$, with a pricing data and investment transactions recorded prior to t available for model training/validation, and the pricing data and investment transactions made after t being used for evaluating the resulting recommendations. Our first time point t_0 is the 1st of July 2019 (providing 1.5 years worth of training data in the first instance). Time points $t \in T$ are spaced two weeks apart, so t_1 is mid July, t_2 is the beginning of August, and so on. When reporting results, we chart recommendation model performance over time for all 29 time points.

Dataset Cleaning (Post-Split): After we have generated a dataset variant for a time point t, we next subject it to a second-stage cleaning process to remove inconsistencies between users and items across the training and test periods. First, we only keep those customers with at least one interaction in the training period . Second, our test set is restricted to assets that (a) have at least one interaction in both training and test periods and (b) have pricing information during the test period. This post filtering is important, as otherwise the pricing-based metrics and transaction-based would be calculated over different customer and asset subsets, which would make them non-comparable.

Content-based Model Recommendation Horizon: The most common types of content-based recommendation models aim to predict how asset prices will change in the future, if the price is predicted to go up faster than the market as a whole then it should be a good investment. How far into the future the model tries to predict is known as the time horizon, which we denote as Δt. For our experiments, we use a fixed Δt of six months, as a mid-term investment horizon.

Dataset Statistics: Figure 1 summarizes the statistics of each split post cleaning in terms of the number of customers, the number of financial assets, the number of transactions in the training and test sets and the profitability of the assets



(a) Num. customers/assets (b) Training transactions (c) Test transactions (d) Market Profitability

Fig. 1. Basic properties of the dataset.

for our selected time horizon (i.e. profitability at t+6 months). As we can observe in Figure 1 (d), the studied period is not stable: from September 2019 to March 2020, the market loses money when we look six months into the future and only a few assets provide positive returns during this period. This is primarily due to the Covid-19 pandemic, which had its greatest economic impact in Europe from March 2020 (six months after September 2019). Having such an unstable period allows us to analyze how this market instability impacts our algorithms over time.

5.2 Metrics

Primary Metrics: The primary focus of this paper is a contrastive study between transaction-based evaluation and profitability-based evaluation. As such, our primary metrics representing these two evaluation types are as follows:

- **Transaction-based Evaluation:** We employ the normalised cumulative discounted gain (nDCG) metric [13] to measure how close the recommendations produced by each FAR approach are to the investments made by the customers. This metric prioritizes having relevant assets (i.e. assets acquired during the test period) in the top ranks. The formulation for this metric is:

$$\text{nDCG@}k(u, R_u) = \frac{\text{DCG@}k(u, R_u)}{\text{IDCG@}k(u)} \tag{1}$$

  where

$$\text{DCG@}k(u, R_u) = \sum_{j=1}^{k} \frac{g_u(i_j)}{\log_2(j+1)} \qquad \text{and} \qquad \text{IDCG@}k(u, R') = \max_{R'} \text{DCG@}k(u, R') \tag{2}$$

  and $g_u(i)$ is the grade of relevance of item i for user u and i_j is the j-th item in ranking R_u. In our experiments, we consider $g_u(i) = Rel(u, i)$, i.e. 1 if u acquires i during the test period and 0 otherwise.

- **Profitability-based Evaluation:** In our experiments we report the average return on investment (ROI) of the top k recommended assets after a fixed time Δt as our measure of profitability. The ROI of an asset is defined as the relative difference between the future and present pricings of the asset:

$$\text{ROI}(i, t, \Delta t) = \frac{\text{price}(i, t + \Delta t) - \text{price}(i, t)}{\text{price}(i, t)} \tag{3}$$

  As indicated earlier, Δt is equal to six months in our experiments.

Secondary Metrics: In addition to the above primary metrics we also report the following secondary metrics to support our analysis in this paper:

- **Profitable Asset Ratio (%prof)**: The proportion of the top-k recommended assets with a ROI ≥ 0.
- **Volatility**: The standard deviation of the daily returns for an asset, averaged over the top-k recommended assets.
- **ROI-nDCG**: This is the nDCG score as calculated above, with the difference that any assets with a ROI ≤ 0 contribute a gain of 0 even if the customer invested into them while profitable assets contribute a gain of ROI. This provides a measure of whether the algorithms were recommending assets that both the customer invested in and were profitable.
- **nROI-nDCG**: This is the nDCG score as calculated above, with the difference that any assets with a ROI ≥ 0 contribute a gain of 0 even if the customer invested into them while lossy recommendations contribute a gain of $-$ROI. This provides a measure of whether the algorithms are recommending non-profitable assets that the customer invested in (i.e. those the customer liked but ultimately lost value). Following [26], we report (1-nDCG), so algorithms recommending more non-profitable assets receive lower values.

5.3 Algorithms

To provide a meaningful comparison of evaluation methods, we need to apply these methods over a range of different FAR approaches, hence, we implement a diverse suite of 12 FAR approaches from the literature, summarized below:

- **Random recommendation:** As a simple, sanity-baseline, we include an algorithm that randomly selects the assets to recommend.
- **Profitability-based models:** As representative algorithms which only consider the pricing history algorithm of the assets, we test three regression approaches, predicting return at $t+6$ months: support vector regression (SVR), random forest and LightGBM regression, a method using gradient boosted regression trees [15]. As featured, we use a selection of technical indicators based on closing price: average price, return on investment, volatility, moving average convergence divergence, momentum, rate of change, relative strength index, detrended close oscillator, Sharpe ratio, and maximum and minimum values over a time period prior to prediction.
- **Transaction-based models:** We choose several methods exploiting investment transactions to generate recommendations. We divide these approaches in three categories:
  - **Non-personalized:** As a basic, not personalized baseline, we consider popularity-based recommendation, which ranks assets according to the number of times they have been purchased in the past.
  - **Collaborative filtering:** As collaborative filtering methods, we test three proposals: LightGCN [12], matrix factorization (MF) [24] and user-based kNN (UB kNN) [22]. We also add the Apriori association rule mining (ARM) algorithm [1], which identifies groups of assets which are commonly acquired together, and establishes rules for recommending assets according to the past investments of the customers.
  - **Demographic methods:** We add another method based on user-based kNN, which instead of using the past customer investments to compute the similarities between customers uses the demographic profile of the customers. In this case , our features are derived from a questionnaire regarding their risk appetite (similarly to [37]). We denote this method as customer profile similarity (CPS).
- **Hybrid methods:** Finally, we test two hybrid methods, based on gradient boosting regression trees [15]: a regression LightGBM algorithm, targeting the profitability at six months in the future (Hybrid-regression), and, second, the LightGBM implementation of the LambdaMART learning to rank algorithm [2], optimizing nDCG (Hybrid-nDCG). As features, we use the outcome of all the previous listed recommendation algorithms.

Table 3. Effectiveness of the compared models at cutoff 10. A cell color goes from red (lower) to blue (higher values) for each metric, with the top value both underlined and highlighted in bold. In the case of ROI, %prof and volatility, blue cells indicate an improvement over the average market value.

Data source	Category	Algorithm	nDCG	ROI	%prof	Volatility	ROI-nDCG	nROI-nDCG
None	–	Random	0.0223	0.0118	0.4879	**0.3895**	0.0094	0.9852
Prices	Regression	SVR	0.0041	0.1212	**0.6415**	0.5045	0.0039	**0.9985**
		LightGBM	0.0599	**0.1423**	0.4914	0.7400	0.0350	0.9661
		Random forest	0.0570	0.0583	0.4314	0.6619	0.0297	0.9644
Transactions	Non-personalized	Popularity	**0.3374**	-0.0628	0.3951	0.5147	**0.1206**	0.7481
	Collaborative filtering	LightGCN	0.3081	-0.0643	0.3620	0.5336	0.1151	0.7772
		ARM	0.2687	-0.0647	0.3619	0.5316	0.0928	0.7950
		MF	0.0812	-0.0460	0.4033	0.4803	0.0301	0.9401
		UB kNN	0.1428	-0.0344	0.4197	0.4303	0.0499	0.8960
	Demographic	CPS	0.3003	-0.0544	0.3853	0.5162	0.1093	0.7791
Hybrid	–	Hybrid-nDCG	0.2454	-0.0466	0.3571	0.5028	0.0880	0.8220
		Hybrid-regression	0.0220	0.0382	0.5199	0.4124	0.0090	0.9848
Market average				0.1026	0.5023	0.4654		

For each algorithm, we select as the optimal hyperparameters those maximizing the ROI at 6 months at three dates: April 1st 2019, October 1st 2019 and January 31st 2020.

5.4 FAR Approach Effectiveness

As discussed, the primary goal of this paper is to analyse the differences between profitability and transaction-based evaluation methods, hence our later results focus exclusively on this. On the other hand, as we study a somewhat rare dataset that has both pricing and transaction data, there is value in reporting the performances of the different FAR approaches more broadly and highlighting patterns of interest. As such, Table 3 reports the performance of all 12 FAR approaches listed above under the six evaluation metrics when averaging over all the considered time points. The highest performing model under each metric is highlighted in bold, and the performance distribution for each metric is colour coded (blue for highly performing and red for poorly performing).

From Table 3 we observe the following points of interest. First, we observe that of the algorithms tested, only the profitability prediction algorithms are capable of suggesting profitable assets, with both SVR and LightGBM being able to beat the average profitability of the market. Second, although transaction-based algorithms are able to reasonably predict customer preferences (as shown by their high nDCG values), they show an overall poor performance in terms of the ROI profitability metric (which we will analyse further in the main experiments in the next section). These methods are, however, able to recommend a minority of profitable assets, something that it is shown by %prof (although not as many as price-based models). When looking at hybrid metrics, ROI-nDCG results are very similar in ranking to nDCG ones, while nROI-nDCG metric mostly inverses nDCG results. This is indicative of how these models only consider whether customers bought the financial assets and not their profitability, as they retrieve both profitable and lossy securities acquired by customers. Finally, FAR methods recommend far more volatile assets than the market average – with the only exceptions of the Hybrid-regression, user-based kNN and random recommenders.

6 RESULTS

This paper investigates two primary research questions, each in a separate section:

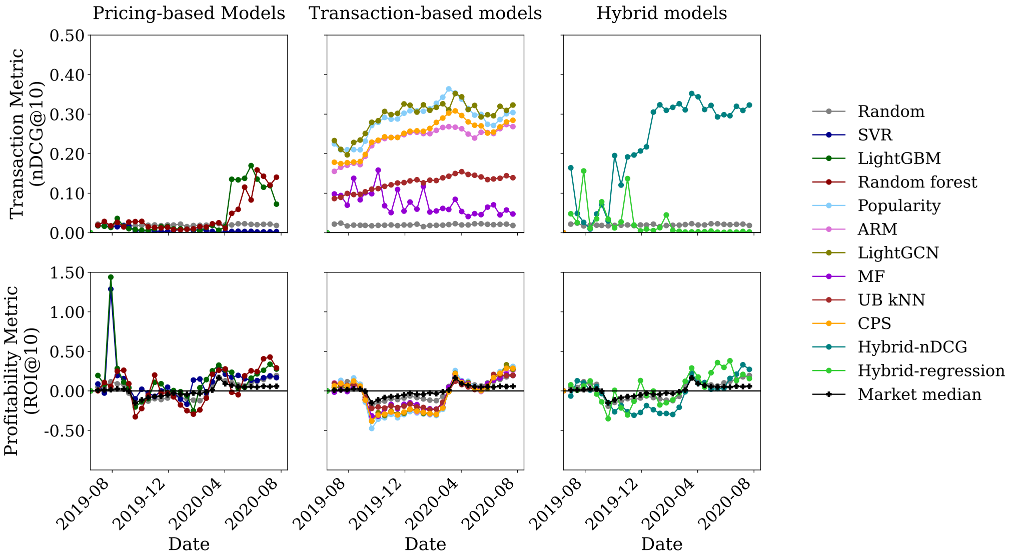


Fig. 2. Comparison of performances reported by the Transaction-Based nDCG@10 and Profitability-based ROI@10 metrics for 12 recommendation algorithms over time.

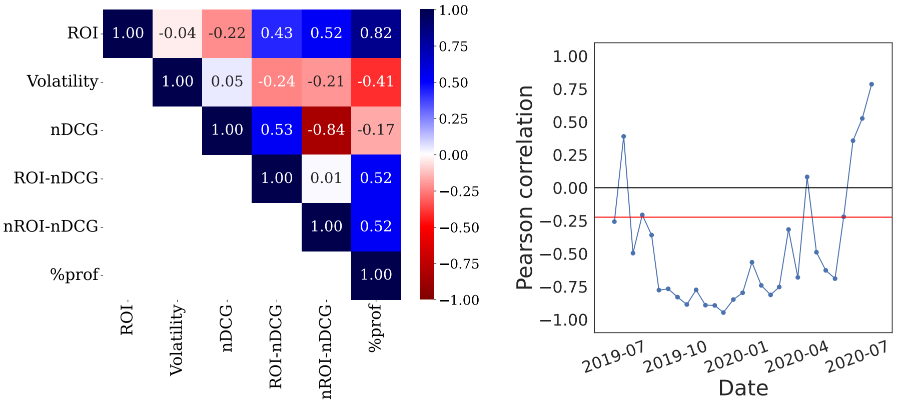
- RQ1: Are transaction-based and profitability-based metrics interchangeable when evaluating financial asset recommendation systems? (Section 6.1)
- RQ2: What are the main factors that influence transaction-based metrics? (Section 6.2)

6.1 RQ1: Are transaction-based and profitability-based metrics interchangeable?

As we discussed in the related work, there have been a range of prior works that have used transaction-based metrics to evaluate whether their asset recommendation technologies are effective [16, 18, 38]. Unlike metrics that measure profitability (e.g. ROI), transaction-based metrics like nDCG do not directly measure whether the customer would make money by investing in the recommended assets, but rather whether the model is recommending assets that the customer invested in later. The underlying assumption behind using transaction-based metrics is that if our customers are intelligent actors that can profit from the market, then transaction-based metrics should correlate with profitability metrics, while also capturing customer preferences towards particular types of assets (making them a superior overall metric).

But is this the case in practice? To answer this question we compare the performance of 12 recommendation strategies when used to produce asset recommendations for between 20-30k customers (this varies over time) for each of the 29 time points in our dataset, under both profitability (ROI@10) and transaction-based (nDCG@10) metrics. If these metrics are interchangeable, we should observe a similar performance pattern produced by both metrics across recommendation strategies. Figure 2 illustrates the performance of the recommendation strategies over time, divided by broad model type (pricing-based, transaction-based or hybrid) for readability.

As we can observe from Figure 2, it is clear that the performance trends as measured by the transaction-based metric (top row of graphs) and those measured by the profitability-based metric (bottom row of graphs) are very different. For the models that derive their recommendations from past pricing history, the profitability metric is reporting variances in returns over time with overall positive returns toward the beginning and end of the dataset and fluctuating returns



(a) Correlation between metrics

(b) Correlation between nDCG and ROI

Fig. 3. Pearson correlation between the different metrics

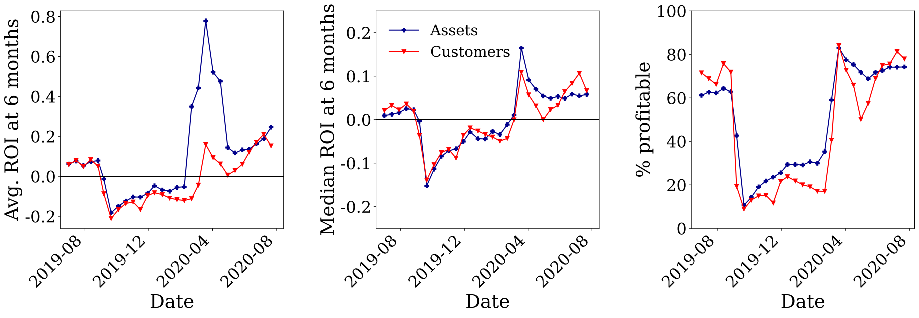
during the Covid downturn, while the transaction-based metric remains around 0 until the final 3-4 months. Meanwhile, for the models that use past transactions during training, there is a clear downward trend in profitability due to the on-set of the Covid period, which is not reflected under the transaction-based metric. Hence, we can conclude that these metric types are not interchangeable.

But are these metrics correlated at all? To evaluate this, we aggregate the performances of all recommendation strategies for all time periods per metric, and then compute the Pearson correlation between pairs of metrics. Figure 3 (a) visualises the resultant correlation for all pairs of metrics described previously in Section 5.2. As we can see from Figure 3 (a), the two metrics we used earlier (ROI@10 and nDCG@10) are in-fact negatively correlated (-0.22), meaning that recommendation models that perform well under this transaction-based metric are likely to lose the customer money! From this result, it appears that the underlying assumption behind using transaction-based metrics does not hold, calling into question the validity of these types of metrics. Hence, in the next section we examine why this is the case.

6.2 RQ2: What Factors Influence Transaction-based Metrics?

From the above analysis, it is clear that these transaction-based metrics are not a proxy for profitability as we might expect. We identified three hypotheses for why this might be the case: 1) our customers are not in fact effective actors and are losing money on the markets; 2) the transaction and profitability metrics do correlate, but only under certain market conditions not prominent in our dataset; and 3) we are measuring profitability incorrectly for our customer-base. We examine each of these hypotheses below:

Are our customers effective investors?: Our first hypothesis for why the transaction-based metrics perform poorly (and also why the models trained with transaction-based data do not make money) is that our customers might not be able to effectively navigate the market and so lose money on average. We can evaluate this by comparing the return on investment of our actual customer investments over time against the market. If our customer investments are under-performing the market then this would explain why transaction-based metrics are not correlated with profitability. To analyse this, we compute the average return on investment obtained in the following 6 months by the customer portfolios for each time point. Figure 4 (a) and (b) compare the average and mean return on investment for the market (Assets, in blue) and the customers (Customers, in red), respectively. Data points below the '0.0' line indicates the market/customer is losing value.



(a) Average ROI at 6 months

(b) Median ROI at 6 months

(c) % profitable customers/assets at 6 months

Fig. 4. Comparison between the profitability of the market and the customers.

Examining Figure 4 (b), we observe that the market and customer curves are very close together, indicating that the customers in this dataset are unable to beat the market median. Moreover, markets like these tend to have a few 'big winners', which skew the average market ROI upward, as we can see expressed as the large spike in ROI starting around April 2020 in Figure 4 (a). Contrasting against the customer average ROI, we do not see a similar spike, indicating that the customers largely did not invest in these winning assets. Hence, we conclude that the customers in this dataset do not seem to be particularly effective investors, so this is clearly one reason that explains why the transaction-based metrics are not correlated with the profitability metrics. However, recall that we did not simply observe no correlation, but a negative correlation, which we cannot fully explain from this return on investment data.

Was the time period of this dataset a-typical?: Our second hypothesis is that the lack of correlation is a side effect of unusual behaviour during the time period examined. We previously illustrated in Figure 1 (c) that there was an unusual spike in the number of assets purchased in the first half of our dataset, and this coincides with a marked drop in market profitability during the same period illustrated in Figure 1 (d). This unusual behavior is due to the Covid-19 pandemic, and its down-stream effect on businesses. If these adverse market conditions had a strong impact on our customer investments (and hence the transaction-based metrics derived from them) then this should be apparent if we contrast the correlation between the profitability and transaction-based metrics over time. If the time period has no impact, then the correlation should remain roughly constant, however if the pandemic had a large impact then we should see a marked drop in correlation when the pandemic starts to impact the market. Figure 3 (b) charts the Pearson correlation between the transaction-based (nDCG@10) and profitability-based (ROI@10) metrics over time. The horizontal red line indicates the average correlation discussed previously in Section 6.1. Note that profitability is calculated 6 months into the future, hence the impact of the pandemic will appear 6 months before its actual onset in this chart.

As we can see from Figure 3 (b) we do see the expected marked decrease in correlation between nDCG@10 and ROI@10 when the pandemic starts to affect the market. This indicates that the customers ability to select profitable assets to invest in was negatively impacted by the pandemic. Moreover, this was not a short term issue, as it took over half a year before the correlation returned to pre-pandemic levels. There are two factors that seem to have caused this sudden drop in investment effectiveness of the customers. First, there were far fewer assets that were profitable during this period, as illustrated in Figure 4 (c), making investment decisions more difficult. Second, as noted earlier, there was a large increase in the number of buy transactions when the pandemic hit (Figure 1 (c)) indicating customers

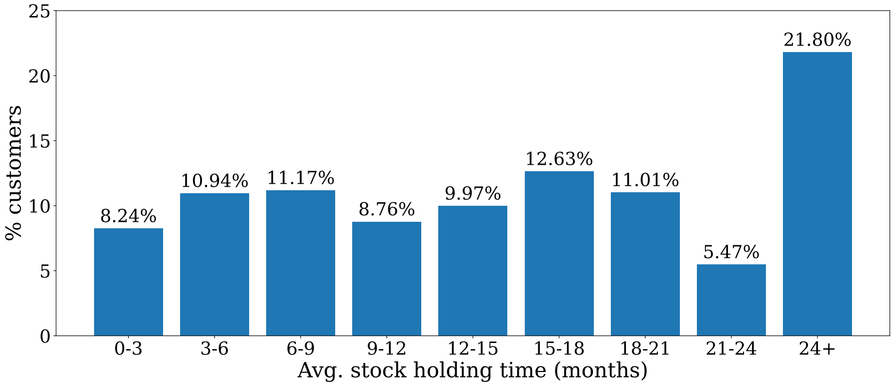


Fig. 5. Classification of customers according to the average time they hold each stock unit.

were purchasing assets when they were under-valued. Hence, we can conclude that Covid-19 was a factor in why the transaction metrics and profitability metrics are not correlated here.

Is ROI after 6 months a good profitability metric?: All of our analysis up-to this point have assumed that we can judge the suitability of an asset for investment based on whether investing in it would result in a profit 6 months later. However, we noted above that customers appeared to be buying assets when they were under-valued due to the pandemic and that these were predominantly not profitable short term - but what if these were longer term investments? If that is the case then we would not necessarily expect such assets to return a profit in only 6 months. To determine the proportion of short and long term investments held by the customers, we calculate the average stock holding time of our customers in the dataset. If ROI after 6 months is a reasonable metric, then we would expect our customers to hold assets for around 6 months on average. Figure 5 reports the stratified average stock holding time of the customers in this dataset. Note that our dataset is only a snapshot of investment transactions, meaning that we do not necessarily have both the buy and sell transactions for each asset. As such, to perform this calculation we assume any assets that the customers held at the start of the dataset were bought on day 1 of the dataset and that all customers holding assets sell those assets on the final day of the dataset. This will skew the data towards a shorter holding time, as some customers may have held an asset for a long time before the start of the dataset, and may want to continue to hold that asset for a long time after the end of the dataset.

As we can see from Figure 5, counter to our expectations (and despite the skew inherent to this analysis), the customers in this dataset appear to favour longer term investing rather than short term investments, with a peak around 15-18 months of holding time.[2] This may be because the asset mix in this dataset is not only stocks, but also covers mutual funds and bonds that customers are likely to hold onto for extended periods. This also raises an important point about working with real transaction data either when training models or evaluating - we need to factor in the customers investment strategy and time horizon, otherwise it is difficult to interpret whether investors are succeeding or not.

To conclude on the analysis of transaction-based vs. profitability-based metrics, we have demonstrated that for this time period, transaction-based metrics were not a good proxy for short term (6 month) profitability. This appears to be caused by a mixture of compounding factors, namely: the Covid-19 pandemic making identification of profitable assets more difficult; customers failing to beat the market in terms of finding profitable assets; and a tendency for the customers to favour longer term investments (that are not well captured by ROI after 6 months).

[2] We note that 24+ months has a higher proportion, but it covers more than a 3-month period and so is not directly comparable.

7 CONCLUSION AND RECOMMENDATIONS

Enabling sound and interpretable evaluation is a critical component of financial asset recommender (FAR) systems. However, the community of FAR researchers and developers are currently presented with multiple competing evaluation methods, with little in the way of guidance regarding when and where they should be used. This paper aims to provide a better understanding of one such evaluation method – transaction-based evaluation, by contrasting it against simpler profitability-based evaluation techniques. Experiments over a large financial asset pricing and transaction dataset demonstrated a negative correlation between profitability and transaction-based metrics across a diverse array of 12 FAR approaches, highlighting that we cannot assume that customers invest effectively and hence models that use those transactions may also not be effective. Through analysis of these models and customer investment behaviour over time, we show that customer investment transactions are a problematic data source for multiple reasons, specifically customers consistently underperforming the market average, the impact of global events leading to changing profitability patterns, and the challenges of leveraging data from a diverse user set with varying trading strategies and investment time horizons.

While it would be premature to suggest that transaction-based evaluation should be abandoned for FAR systems, our results demonstrate that transaction-based metrics have important limitations that need to be understood if they are to be useful. Hence, we provide the following recommendations for researchers and practitioners:

- **Consider changing market conditions:** Global events like pandemics or wars have a huge impact over the market. Major events influence the expectations people have on market segments, prompting customers to change their investment positions. Models trained using transaction-based metrics will perform poorly during such times, as past and current investment behaviour are no longer similar. Hence, it is important to report performance over time to reveal when these changes occur, and solution developers may wish to consider fall-back strategies based on profitability prediction during such times.
- **Investment Horizons are a Confounding Variable:** Different customers plan for different investment time horizons (how long they want to hold an asset for). Analysis of our dataset indicates that these time horizons are markedly longer than we anticipated, with the peak between 15-18 months, but with a wide range of horizons being observed. This has several important consequences for evaluation. First, individual customer transactions become difficult to interpret, as we cannot know in advance the customer's envisaged investment horizon. Second, aggregate metrics like nDCG conflate customers with different horizons, so models trained based on such metrics will likely perform poorly in practice (since we don't know how long to hold a recommended asset for).

As future work, we envision the creation of an adequate and robust framework for FAR evaluation, which puts the focus on the customers and their trading strategies. To develop that framework, it is necessary to understand what role customer features – such as patterns of spending, relationship with the financial institutions, risk aversion, trading platform or sector interest might have on FAR evaluation. Another line of research might address how the past actions of financial institutions might affect the evaluation, as past actions of financial advisors might introduce some biases on the collected datasets (similarly to how the action of past recommendation policies introduce selection biases on offline datasets for general domain recommendation [28]).

ACKNOWLEDGEMENTS

This work was carried out as part of the Infinitech project which is supported by the European Union's Horizon 2020 Research and Innovation programme under grant agreement no. 856632.

REFERENCES

[1] Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB 1994)*. Morgan Kaufmann Publishers Inc., Santiago de Chile, Chile, 487–499.

[2] Chris Burges. 2010. *From RankNet to LambdaRank to LambdaMART: An Overview.* Microsoft Research Technical Report MSR-TR-2010-82. Microsoft.

[3] Robin D. Burke. 2000. Knowledge-based Recommender Systems. *Encyclopedia of Library and Information Systems* 69, Supplement 32 (2000).

[4] Robin D. Burke. 2007. Hybrid Web Recommender Systems. In *The Adaptive Web: Methods and Strategies of Web Personalization*, Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl (Eds.). Springer, Berlin, Heidelberg, Germany, 377–408. https://doi.org/10.1007/978-3-540-72079-9_12

[5] Rocío Cañamares, Pablo Castells, and Alistair Moffat. 2020. Offline evaluation options for recommender systems. *Information Retrieval Journal* 23, 4 (2020), 387–410. https://doi.org/10.1007/s10791-020-09371-3

[6] Thanarat H. Chalidabhongse and Chayaporn Kaensar. 2006. A Personalized Stock Recommendation System using Adaptive User Modeling. In *Proceedings of the 2006 International Symposium on Communications and Information Technologies (ISCIT 2006)*. Bangkok, Thailand, 463–468. https://doi.org/10.1109/ISCIT.2006.339989

[7] Eunsuk Chong, Chulwoo Han, and Frank C. Park. 2017. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications* 83 (2017), 187–205. https://doi.org/10.1016/j.eswa.2017.04.030

[8] Shibo Feng, Chen Xu, Yu Zuo, Guo Chen, Fan Lin, and Jianbing XiaHou. 2022. Relation-aware dynamic attributed graph attention network for stocks recommendation. *Pattern Recognition* 121 (2022), 108119. https://doi.org/10.1016/j.patcog.2021.108119

[9] Tomer Geva and Jacob Zahavi. 2014. Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. *Decision Support Systems* 57 (2014), 212–223. https://doi.org/10.1016/j.dss.2013.09.013

[10] Israel Gonzalez-Carrasco, Ricardo Colomo-Palacios, Jose Luis Lopez-Cuadrado, Ángel García-Crespo, and Belén Ruiz-Mezcua. 2012. PB-ADVISOR: A private banking multi-investment portfolio advisor. *Information Sciences* 206 (2012), 63–82. https://doi.org/10.1016/j.ins.2012.04.008

[11] Asela Gunawardana, Guy Shani, and Sivan Yogev. 2022. Evaluating Recommender Systems. In *Recommender Systems Handbook, 3rd edition*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, New York, NY, USA, 547–601. https://doi.org/10.1007/978-1-0716-2197-4_15

[12] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, YongDong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*. ACM, Virtual Event, China, 639–648. https://doi.org/10.1145/3397271.3401063

[13] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems* 20 (October 2002), 422–446. Issue 4. https://doi.org/10.1145/582415.582418

[14] Dominik Jung, Verena Dorner, Florian Glaser, and Stefan Morana. 2018. Robo-Advisory. *Business and Information Systems Engineering* 60 (2018), 81–86.

[15] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS 2017)*. Curran Associates, Inc.

[16] Eric L. Lee, Jing-Kai Lou, Wei-Ming Chen, Yen-Chi Chen, Shou-De Lin, Yen-Sheng Chiang, and Kuan-Ta Chen. 2014. Fairness-Aware Loan Recommendation for Microfinance Services. In *Proceedings of the 2014 International Conference on Social Computing (SocialCom 2014)*. ACM, Beijing, China, 1–4. https://doi.org/10.1145/2639968.2640064

[17] Johannes Luef, Christian Ohrfandl, Dimitris Sacharidis, and Hannes Werthner. 2020. A Recommender System for Investing in Early-Stage Enterprises. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing (SAC 2020)*. ACM, Online, 1453–1460. https://doi.org/10.1145/3341105.3375767

[18] Nikolaos F. Matsatsinis and Eleftherios A. Manarolis. 2009. New Hybrid Recommender Approaches: An Application to Equity Funds Selection. In *Proceedings of the 1st International Conference on Algorithmic Decision Theory (ADT 2009)*. Springer Berlin Heidelberg, Venice, Italy, 156–167. https://doi.org/10.1007/978-3-642-04428-1_14

[19] Cataldo Musto and Giovanni Semeraro. 2015. Case-based Recommender Systems for Personalized Finance Advisory. In *Proceedings of the 1st International Workshop on Personalization & Recommender Systems in Financial Services (FinRec 2015)*. Graz, Austria, 35–36.

[20] Cataldo Musto, Giovanni Semeraro, Pasquale Lops, Marco de Gemmis, and Georgios Lekkas. 2014. Financial Product Recommendation through Case-based Reasoning and Diversification Techniques. In *Poster Proceedings of the 8th ACM Conference on Recommender Systems (RecSys 2014)*. Foster City, Silicon Valley, CA, USA.

[21] Cataldo Musto, Giovanni Semeraro, Pasquale Lops, Marco de Gemmis, and Georgios Lekkas. 2015. Personalized finance advisory through case-based recommender systems and diversification strategies. *Decision Support Systems* 77 (2015), 100–111. https://doi.org/10.1016/j.dss.2015.06.001

[22] Athanasios N. Nikolakopoulos, Xia Ning, Christian Desrosiers, and George Karypis. 2022. Trust Your Neighbors: A Comprehensive Survey of Neighborhood-Based Methods for Recommender Systems. In *Recommender Systems Handbook, 3rd Edition*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, 39–89. https://doi.org/10.1007/978-1-0716-2197-4_2

[23] Preeti Paranjape-Voditel and Umesh Deshpande. 2013. A stock market portfolio recommender system based on association rule mining. *Applied Soft Computing* 13, 2 (2013), 1055–1063. https://doi.org/10.1016/j.asoc.2012.09.012

[24] Steffen Rendle, Walid Krichene, Li Zhang, and John Anderson. 2020. Neural Collaborative Filtering vs. Matrix Factorization Revisited. In *Proceedings of the 14th ACM Conference on Recommender Systems (RecSys 2020)*. ACM, Virtual Event, Brazil, 240–248. https://doi.org/10.1145/3383313.3412488

[25] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2022. Recommender Systems: Techniques, Applications, and Challenges. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, 1–35. https://doi.org/10.1007/978-1-0716-2197-4_1

[26] Pablo Sánchez and Alejandro Bellogín. 2018. Measuring Anti-Relevance: A Study on When Recommendation Algorithms Produce Bad Suggestions. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys 2018)*. ACM, Vancouver, British Columbia, Canada, 367–371. https://doi.org/10.1145/3240323.3240382

[27] Markus Schedl, Peter Knees, Brian McFee, and Dmitry Bogdanov. 2022. Music Recommendation Systems: Techniques, Use Cases, and Challenges. In *Recommender Systems Handbook, 3rd Edition*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, 927–971. https://doi.org/10.1007/978-1-0716-2197-4_24

[28] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML 2016)*. JMLR, New York, NY, USA, 1670–1679.

[29] Robert P. Schumaker and Hsinchun Chen. 2009. Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System. *ACM Transactions on Information Systems* 27, 2, Article 12 (2009). https://doi.org/10.1145/1462198.1462204

[30] Vivek Sehgal and Charles Song. 2007. SOPS: Stock Prediction Using Web Sentiment. In *Proceedings of the 7th IEEE International Conference on Data Mining Workshops (ICDMW 2007)*. IEEE, Omaha, NE, USA, 21–26. https://doi.org/10.1109/ICDMW.2007.100

[31] John Soldatos and Dimosthenis Kyriazis (Eds.). 2022. *Big Data and Artificial Intelligence in Digital Finance*. Springer. https://doi.org/10.1007/978-3-030-94590-9

[32] Yunchuan Sun, Mengting Fang, and Xinyu Wang. 2018. A novel stock recommendation system using Guba sentiment analysis. *Personalized Ubiquitous Computing* 22, 3 (2018), 575–587. https://doi.org/10.1007/s00779-018-1121-x

[33] Robin M. E. Swezey and Bruno Charron. 2018. Large-Scale Recommendation for Portfolio Optimization. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys 2018)*. ACM, Vancouver, British Columbia, Canada, 382–386. https://doi.org/10.1145/3240323.3240386

[34] Jiliang Tang, Xia Hu, and Huan Liu. 2013. Social recommendation: a review. *Social Network Analysis and Mining* 3, 4 (2013), 1113–1133. https://doi.org/10.1007/s13278-013-0141-9

[35] Wenting Tu, Min Yang, David W. Cheung, and Nikos Mamoulis. 2018. Investment recommendation by discovering high-quality opinions in investor based social networks. *Information Systems* 78 (2018), 189–198. https://doi.org/10.1016/j.is.2018.02.011

[36] Hongyang Yang, Xiao-Yang Liu, and Qingwei Wu. 2018. A Practical Machine Learning Approach for Dynamic Stock Recommendation. In *Proceedings of the 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE 2018)*. IEEE, New York, NY, USA, 1693–1697. https://doi.org/10.1109/TrustCom/BigDataSE.2018.00253

[37] Yang Yujun, Li Jianping, and Yang Yimei. 2016. An Efficient Stock Recommendation Model Based on Big Order Net Inflow. *Mathematical Problems in Engineering* 2016, Article 5725143 (2016). https://doi.org/10.1155/2016/5725143

[38] Xiaoxue Zhao, Weinan Zhang, and Jun Wang. 2015. Risk-Hedged Venture Capital Investment Recommendation. In *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys 2015)*. ACM, Vienna, Austria, 75–82. https://doi.org/10.1145/2792838.2800181

[39] Zeqi Zheng, Yuandong Gao, Likang Yin, and Monika K. Rabarison. 2020. Modeling and analysis of a stock-based collaborative filtering algorithm for the Chinese stock market . *Expert Systems with Applications* 162 (2020), 113006. https://doi.org/10.1016/j.eswa.2019.113006

[40] Dávid Zibriczky. 2016. Recommender Systems meet Finance: a Literature Review. In *Proceedings of the 2nd International Workshop on Personalization & Recommender Systems in Financial Services (FinRec 2016)*. CEUR Workshop Proceedings, Bari, Italy, 3–10.