

FAR-Trans: An Investment Dataset for Financial Asset Recommendation

Javier Sanz-Cruzado¹, Nikolaos Droukas², Richard McCreadie¹

¹University of Glasgow

²National Bank of Greece

javier.sanz-cruzadopuig@glasgow.ac.uk, droukas.nikolaos@nbg.gr, richard.mccreadie@glasgow.ac.uk

Abstract

Financial asset recommendation (FAR) is a sub-domain of recommender systems which identifies useful financial securities for investors, with the expectation that they will invest capital on the recommended assets. FAR solutions analyse and learn from multiple data sources, including time series pricing data, customer profile information and expectations, as well as past investments. However, most models have been developed over proprietary datasets, making a comparison over a common benchmark impossible. In this paper, we aim to solve this problem by introducing FAR-Trans, the first public dataset for FAR, containing pricing information and retail investor transactions acquired from a large European financial institution. We also provide a bench-marking comparison between eleven FAR algorithms over the data for use as future baselines. The dataset can be downloaded from <https://doi.org/10.5525/gla.researchdata.1658>.

1 Introduction

Recent advances in the automated analysis of financial content and artificial intelligence are driving a digital transformation of financial services. These technologies represent an opportunity for banks, fund operators and fintech companies to improve their business processes, improve the quality of their decisions, and increase financial inclusion [Soldatos and Kyriazis, 2022]. The investment advice sector has been disrupted by these technologies, resulting in a transition from customers only receiving assistance from certified financial advisors to new scenarios on which the advisors' decisions are supported by automated systems or where customers are directly served by robo-advisors.

Financial asset recommendation (FAR) lies at the core of these automated advice tools. For an investor, FAR identifies a list of financial securities or assets (such as stocks, bonds or funds) ranked by their suitability for the customer. This suitability is not only driven by the investor, but also by external factors like asset returns, currency value and inflation. FAR methods also need to consider the personal situation, needs and preferences of the user, represented by past investment

transactions and explicit customer information (e.g. risk tolerance and investment capacity). Therefore, effective recommendations should model pricing data to distinguish highly performing and under-valued assets, while also identifying those assets on which the investor might be interested [McCreadie *et al.*, 2022; Sanz-Cruzado *et al.*, 2022].

There is growing interest into the research and development of FAR technologies, as demonstrated by workshops in prominent conferences like RecSys [Bogers *et al.*, 2022], ICAIF¹ and IJCAI². However, the absence of public datasets with realistic customer transactions that can be used to train and evaluate approaches under a common benchmark is a significant barrier to research. A majority of past research has focused on the study of profitability prediction methods [Paranjape-Voditel and Deshpande, 2013; Schumaker and Chen, 2009; Sehgal and Song, 2007; Song *et al.*, 2017; Zheng *et al.*, 2020] as asset pricing data is freely available. However, these approaches are fundamentally limited by their inability to model the customer. Although some works do consider complex investor information, such as customer profiles or transactions [Barreau and Carlier, 2020; Chalidabhongse and Kaensar, 2006; Gonzales and Hargreaves, 2022; Lee *et al.*, 2014; Musto and Semeraro, 2015; Takayanagi *et al.*, 2023; Zhao *et al.*, 2015], they rely on proprietary or simulated datasets that are not publicly available.

This work aims to solve this limitation by proposing a novel dataset for the FAR task, provided by a large European financial institution. As far as we are aware, this dataset represents the first dataset in this domain containing pricing time series for multiple asset types (stocks, bonds and mutual funds), asset descriptions, as well as most importantly (anonymised) customer information and investment transactions. In this paper, we provide a description of the dataset, recommended experimental setup and an initial comparison of 11 baseline FAR approaches to support future work.

2 Related work

2.1 Financial asset recommendation

The particular nature of the financial domain has inspired a variety of recommendation techniques taking advantage

¹<https://sites.google.com/view/ml-for-investor-recsys>

²<https://sites.google.com/view/fin-recsys2024/>

of multiple data sources, such as: pricing data, investment transactions, news, social media, etc. [Zibriczky, 2016; McCreddie *et al.*, 2022]. According to their main data source, we can categorize these methods in three primary groups: based on price, based on transactions and hybrid models.

Price-based methods

The first category of FAR algorithms establishes price time series as their primary source of information to identify investment opportunities. These methods, based on the prediction of the price or performance of the securities, are not personalized [Zibriczky, 2016].

Most works are based on regression techniques. The simplest methods use one or several regression models (such as a Random Forest or SVM) to estimate asset profitability [Sanz-Cruzado *et al.*, 2022; Yang *et al.*, 2018] based on price or technical indicators. More complex models explore similarities between the time series of multiple assets [Feng *et al.*, 2022; Paranjape-Voditel and Deshpande, 2013; Zheng *et al.*, 2020] or incorporate information from other data sources to generate the prediction, such as news with evidence of major events [Song *et al.*, 2017] or trader's views about assets on social media [Sun *et al.*, 2018; Tu *et al.*, 2018].

More recently, some works have addressed this problem as a stock ranking selection task where the goal is to select a list of assets maximizing some utility function (for example, the combined predicted returns). [Feng *et al.*, 2019] represents the first work in this area, using a temporal graph convolutional network to combine asset prices and knowledge graphs. More recently, [Alsulmi, 2022] combined pricing and fundamental asset information to train multiple learning-to-rank methods [Liu, 2009] for selecting stocks in the Saudi market.

Transaction-based methods

Following the classic methodology of recommender systems [Ricci *et al.*, 2022], the second category of FAR algorithms uses investment transactions as the core data source. These methods assume that investors follow patterns, and hence past investments can be used to model customers (either individually or as groups).

Some of these works rely only on investment logs, developing collaborative filtering approaches based on matrix factorization [Lee *et al.*, 2014; Zhao *et al.*, 2015], convolutional networks [Barreau and Carlier, 2020] or customer clustering [Gonzales and Hargreaves, 2022]. Other methods incorporate other information sources. For example, [Musto *et al.*, 2014; Musto and Semeraro, 2015; Musto *et al.*, 2015] design investment portfolio case-based recommendations factoring in the risk aversion of customers. Meanwhile, [Takayanagi and Izumi, 2024] proposes a demographic kNN method where user similarity is computed according to personality traits. Finally, [Luef *et al.*, 2020] develop content-based methods by adding asset information like market sector or enterprise life cycle, as well as a social recommendation approach based on trust between investors.

Hybrid algorithms

The last family of algorithms [Burke, 2007] combines several information sources to provide recommendations. For

FAR, [Chalidabhongse and Kaensar, 2006] propose an adaptive model to learn from past investments, financial technical indicators and demographic data about the customers. Meanwhile, [Matsatsinis and Manarolis, 2009] combine collaborative filtering and multi-criteria decision analysis to generate a utility score for equity fund recommendation. [Swezey and Charron, 2018] rerank the output of a collaborative filtering matrix factorization approach using the weights obtained in a portfolio optimization process. Luef *et al.* [Luef *et al.*, 2020] propose a hybrid method that combines both content-based and knowledge-based components. Finally, Kubota *et al.* [Kubota *et al.*, 2022] leverage card transactions and mobile usage app statistics from customers to identify companies they have interacted with in the past, and recommend them to invest on their stocks.

As we can see, many diverse algorithms have been proposed for the FAR task. However, the lack of a common dataset and evaluation methodology makes it impossible to fairly compare approaches in terms of effectiveness for the task. Therefore, in Section 5 we provide an evaluation benchmark comparing 11 models over our new dataset, drawn from the three algorithm classes.

2.2 Existing Recommender Systems Datasets

The development of recommendation technologies has been assisted by the availability of public resources for researchers and practitioners. One of the earliest efforts in the area is the original MovieLens dataset released in 1997 [Harper and Konstan, 2015], which provided customer ratings for movies. Since their original release, multiple data collections have been published for different recommendation domains, including movies and TV series [Pérez Maurera *et al.*, 2020], music [Bertin-Mahieux *et al.*, 2011], videogames [Pathak *et al.*, 2017], books [Wan and McAuley, 2018] and points of interest [Yang *et al.*, 2015].

However, there is not a standard dataset for developing and comparing novel approaches in the investment domain. Besides those studies using only public pricing information [Chong *et al.*, 2017; Feng *et al.*, 2022; Yang *et al.*, 2018], some works have evaluated algorithms on datasets containing customer and transaction information. However, customer and transaction information is commonly subject to privacy concerns [Thompson *et al.*, 2021], so these works use private datasets, collected in agreement with banks or brokerage firms [Barreau and Carlier, 2020; Kubota *et al.*, 2022; Gonzales and Hargreaves, 2022; Takayanagi *et al.*, 2023].

The only exception to this is the dataset introduced in [Musto *et al.*, 2014; Musto *et al.*, 2015]³. Created in agreement with ObjectWay Financial Software, this dataset is publicly accessible and collects the investment portfolios of 1,172 users between June 2011 and 2013. Besides the portfolios, it includes information about customer needs and asset types. However, it does not provide pricing information about the assets or information which can be used to identify them, preventing researchers from testing price-based approaches.

In this paper, we aim to provide a new dataset which can be used to develop and evaluate novel FAR models, either

³http://bit.ly/financialRS_data_uniba (Accessed 19/04/2024)

focused on profitability prediction, investment transactions, or hybrid models combining both. We provide a description of the dataset in the next section.

3 Dataset

We introduce in this work a novel dataset for financial recommendation, which we shall name FAR-Trans. As far as we are aware this dataset represents the first public dataset containing both asset pricing information and investment transactions for FAR. The data has been provided by a large European financial institution, representing a snapshot of the market available to Greek investors between January 2018 and November 2022. FAR-Trans covers pricing data for stocks, bonds and mutual funds, as well as investment transaction logs (asset buy and sell actions) handled by the institution, customer, market and asset information. This section provides a description of the dataset and the acquisition and cleaning methodology. The dataset is available from <https://doi.org/10.5525/gla.researchdata.1658>. Table 1 summarizes its global properties.

3.1 Prices

Prices indicate variations in the value of financial securities. Therefore, the past prices of financial assets represent an important source of information in the development and evaluation of FAR approaches. Pricing time-series have multiple uses: asset analysis through the computation of technical indicators, risk estimation, development of content-based FAR methods or evaluation according to the profitability of assets [Sanz-Cruzado *et al.*, 2022].

Cleaning and pre-processing

When acquiring pricing data for financial assets, it is not uncommon to find small gaps or invalid values in them caused by problems in the data collection. While these problems are realistic, they add a confounding variable for the asset analysis. Thus, we need to clean and pre-process our data to minimise the impact of these errors.

A potential source of error involves the collection of pricing information from multiple data sources or markets. We clean our dataset so every asset has, at most, a single price value at any given date. We first remove pure duplicates from our data. Then, for those assets that still have multiple values on a date (a) we remove values equal to 0 or (b) we keep the value which is closer to the price of the previous 5 days. In cases where the price time series changes trading currency mid-day, we keep the value closer to the price of the following 5 days instead (as these days use the new currency).

We next treat major errors within the data: as investment transactions require capital exchange, we removed from our dataset those assets with closing price values equal to 0 at some point in their time series. We also observed assets with time gaps. For shorter periods, we can estimate the missing points, but, the longer the gap, the more inaccurate the estimation will be. We therefore then remove assets with large time gaps (longer than 10 days).

Another aspect to study are sudden variations in price, as they might lead to outliers affecting what FAR algorithms learn. We consider as outliers those values where the price

is increased by 10 times or loses 90% of the value on a single day, and price reverts to its original value range on the following day. As those cases are mostly due to errors on data collection, we estimate the correct price by a moving average of the previous five days to prevent undesirable effects.

A more complex case occurs when the price never reverts to its original scale – as this might reflect a currency change or a stock split. In the first case, we apply price transformations to ensure all time series are represented in euros. In the second case, stock splits represent corporate actions changing the number of shares on which a stock is divided. For instance, a company might divide every share into two – causing every investor to own twice the number of shares, but every individual share halved in value. There are two types of splits: direct stock splits increase the number of shares, whereas reverse stock splits diminish it. To prevent variations in our data, we check those assets with increases or decreases of a third of their value in a single day. Then, with the assistance of Yahoo! Finance⁴, we identify (a) whether a stock split occurred, (b) its date and (c) the split ratio. Then, for every stock split, we divide the prices previous to the split date by the split ratio – for example, if a company performs a 2-for-1 (2:1) stock split, all prices previous to the split date are halved. Finally, we finish our cleaning process by closing the remaining gaps by applying a moving average over the previous five days.

Statistics

Figure 1(a) illustrates the average price of the assets included in our dataset. As we can observe, our data covers both bullish and bearish market periods – including recent economic recess periods like the Covid-19 pandemic in March 2020, and the effects of the Ukraine-Russian war at the beginning of 2022.

3.2 Assets

We collect further information about the financial securities beyond the pricing data. For each of the 807 assets with pricing data, we obtain from public sources their asset type (stock, bond or mutual fund) and sub-type (for instance, bonds can be government or company bonds), their names, the market where they are traded and, where available, their sector and industry.

Figures 1(b) and (c) provide some statistics about the assets. Figure 1(b) illustrates the number of assets of each type (stocks, bonds and mutual funds). As it can be observed, the three categories are well represented, with at least 200 assets on each of them – with mutual funds representing the majority of the collection. Following Figure 1(c), we also observe that our assets are not just restricted to the Greek market – they represent the assets on which customers could invest through the financial institution. Therefore, although a large fraction of the assets come from Greek markets, there are also assets from other European markets (e.g. Luxembourg and Germany) and some US securities.

3.3 Transactions

The main novelty of this dataset is the availability of investment interactions between banking customers and financial

⁴Yahoo! Finance: <https://finance.yahoo.com/>

Table 1: Description of the FAR-Trans dataset.

Market data		Customer data	
Property	Value	Property	Value
Unique assets	806	Unique customers	29,090
Assets with investments	321	Transactions (unique)	388,049 (154,103)
Unique markets	38	Acquisitions (unique)	228,913 (89,884)
Price data points	703,303	Sales (unique)	159,136 (64,219)
Average return (by assets, whole period)	37.16%	Average return (by customers, whole period)	22.89%
% profitable assets	54.28%	% customers with profits	54.56%

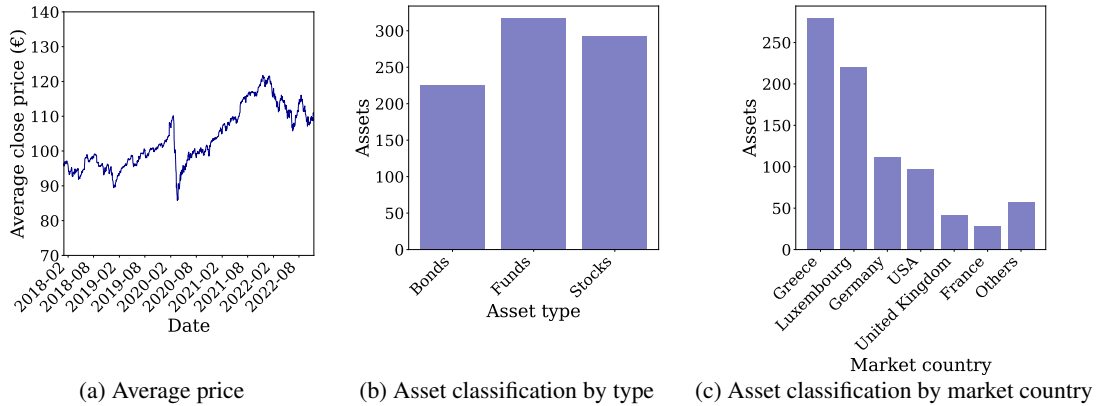


Figure 1: Financial asset statistics.

assets. These interactions represent acquisitions and sales of the securities by individual investors which have been managed by the financial institution. Investment transactions can be used for modelling the past behaviour of customers and develop personalized recommendation approaches. They can also be used to evaluate investment prediction using classical recommender systems and information retrieval metrics, which consider these algorithms as a predictor for future customer behaviour [Sanz-Cruzado *et al.*, 2022].

Cleaning and pre-processing

The raw data includes customer and asset identifiers, the type (buy, sale) and date of the transaction, the number of shares bought/sold by the investor, the total amount of money involved in the transaction, and the channel the customer used to execute that transaction. However, this raw data needs to be aligned with the rest of the dataset (and, specifically, with the price time series described in Section 3.1). We therefore perform some cleaning transformations over the initial log.

First, every transaction needs an associated customer to it, so we removed those with a blank customer. Once this is done, we consider the effects that stock splits have on the number of shares every customer owns by multiplying the number of acquired/sold shares by the split ratio. If a customer owns fractional shares after a reverse stock split, we assume that the company provides cash instead of those fractional shares⁵ and we add a transaction selling those fractional shares at the date of the split.

⁵<https://finance.yahoo.com/news/why-investors-cash-lieu-fractional-140004745.html>

Another observation over the raw data is that customers sell assets which they never acquired during the 2018-2022 period, indicating that they acquired them earlier and had them in their portfolio. To ensure that every asset sale is backed by a purchase, we recreate those asset buys. For every customer, we compute the number of shares she owns of every asset. If the investor owns a negative number of shares (meaning that she has sold more shares than she has bought) at the end of 2022, we add a buy transaction at the earliest point in time for which we have pricing data for the asset (in a majority of cases, 2nd January 2018). We assume that the customer acquires the number of shares which were sold in excess.

Afterwards, we fix those cases where customers interact with assets at times when the pricing data is not available. In case the transaction is outside the range of dates for which we have pricing data, we move the transaction to the closest date where the price exists. Then, we check if customers have shares of an asset after the end of the pricing time series of the security. If they do, we add a transaction selling those assets.

Finally, we solve inconsistencies on asset prices by providing an estimate of the total value of the transaction. We estimate the value by multiplying the number of shares by the closing price of the asset on the date of the transaction. In the end, we have 388,049 transactions in our dataset, corresponding to 29,090 customers.

Statistics

We summarize in Figure 2 the statistics of the transaction data. First, 2(a) displays the number of investment transactions registered on every month of the studied period. In

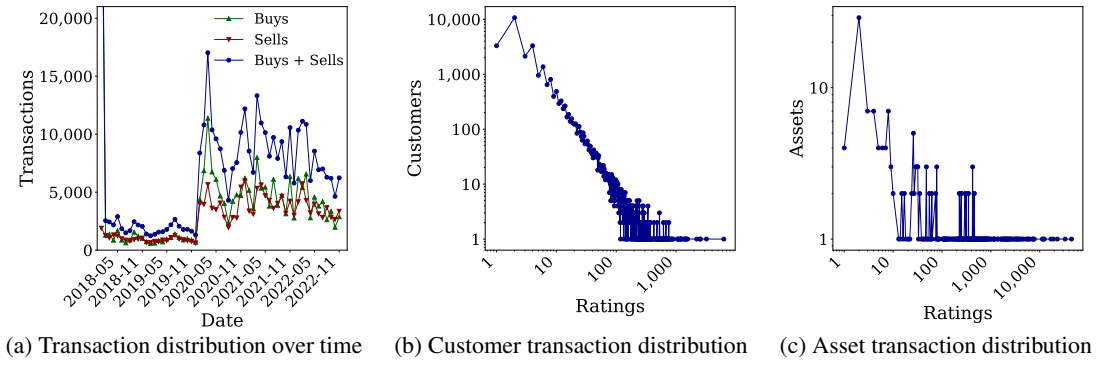


Figure 2: Transaction statistics.

the figure, the x axis shows the dates, whereas the y axis represents the number of transactions. The green line represents the asset purchases, the red line represents the sales and the blue line represents the combined number of transactions. The first observation from this figure is that most of the transactions occur in the period between January 2020 and November 2022, with over 5,000 interactions happening every month. The previous period, between January 2018 and December 2019 only receives between 1,000 and 2,000 trades per month, with the exception of January 2018. However, the large number of transactions at that date is due to the creation of asset buys representing what customers had in their portfolio before the beginning of the period covered by our dataset. The actual largest number of transactions occurs in March 2020, corresponding with the time the Covid-19 pandemic hit Europe, likely due to the huge drop in market prices that occurred during that period.

Figures 2(b) and (c) represent, respectively, the transaction distribution over customers and assets. The y axis represents the number of customers/assets which have associated the number of transactions as indicated on the x axis. Due to the skewness of the distributions, we represent both figures in log-log scale.

The investment distribution by customers in Figure 2(b) resembles a long-tail distribution, where a majority of customers only modify their portfolios a few times over the whole period (over 50% of the users have 3 or less transactions between 2018 and 2022) whereas only a few customers modify their investments continuously (around 650 customers have more than 100 transactions). This long-tail distribution is similar to other recommender system datasets like MovieLens.

A different pattern is observed when we examine the asset distribution however. First, as indicated in Table 1, less than half of the assets (321 out of 807) have ever been bought or sold in our dataset. Second, although the distribution illustrated in Figure 2 is skewed, this is due to a few assets concentrating lots of transactions: even when most of the interacted assets have a reasonable number of transactions (75% of them have more than 20 interactions, and 58% of them have more than 100), the top 12 assets concentrate more than 50% of the dataset interactions. This indicates that there is an important popularity factor over the collected transactions.

3.4 Customers

Financial asset recommendation needs to consider the specific needs and preferences of the customers. However, all that information is not only hidden in the past customer transactions: explicit information about customer investment capacity or risk profile can be considered to identify more relevant investment opportunities for retail customers. As such, we include in the dataset information about the classification of customers within the bank, their investment risk profile and their investment capacity. We collect that information from the 29,090 customers within the bank who have, at least, one investment reflected in our cleaned transaction data. All customer information has been thoroughly anonymized and does not contain sensitive data to satisfy regulations. We provide further descriptions for the customer classifications below:

Customer segments

Customer segments represent the internal classification of customers within the bank. We consider five different segments in our data:

- **Mass:** The majority of the customers. This category represents customers with less than €60,000 of managed assets (investments, deposits and insurance products).
- **Premium:** Individual customers with more than €60,000 on managed assets.
- **Professional:** Sole proprietorship. Individual customer exercising their activity without having created a legal person.
- **Legal Entity:** This category represents legal entities with services within the bank.
- **Inactive:** Customers without available segment.

Figure 3(a) shows the distribution of customers over the different categories: as we can observe individual retail investors represent the majority of the dataset, with 18,610 mass customers, 8,906 premium customers, followed by the business customers (1,327 professional and 39 legal entities). The segment of 208 customers remains unknown.

Investment risk profile

The investment risk profile categorises customers according to the amount of risk they would accept on their investments. To assess if the offered investment assets are suitable for their

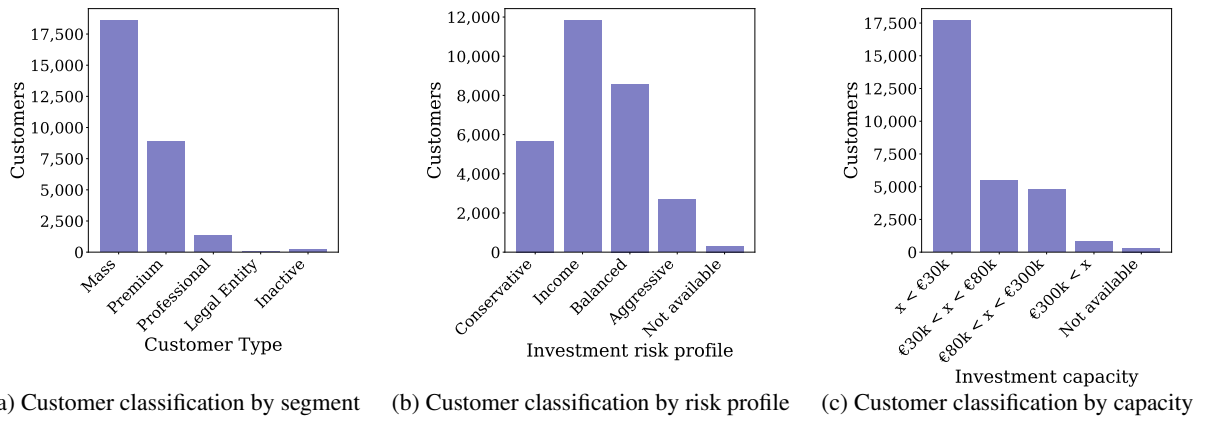


Figure 3: Customer statistics.

investment goals and aligned with their risk aversion, every customer who is interested in investing is asked by the bank to complete an investment profile questionnaire with 25 questions. We provide these questions on the dataset. Following the MiFID II [Council, 2014] regulatory framework, this risk assessment aims to ensure that the financial instruments provided to investors are compatible with their needs, characteristics and goals. According to their answers, their risk profile is in one of the following four categories:

- **Conservative:** Conservative customers prioritize protecting their capital. Their portfolios should be easily liquidated and present extremely low investment risk. An example portfolio might include short-term placements and fixed-income securities.
- **Income:** Customers with this risk profile aim at generating a fixed income arising from bond coupons, dividends and short-term placements. Their portfolios should present very low investment risk.
- **Balanced:** Accepting possible fluctuations on the invested capital, these customers aim at generating fixed income from coupons and dividends, as well as medium-term capital gains. An example portfolio would contain both bonds and stocks.
- **Aggressive:** This profile aims at significant long-term gains, which come with high risks.

The risk profile is precise for those customers who have already answered the questionnaire. In case they have not, the score is simulated through an automated process: first, given the answered questionnaires, linear regression is used to determine which questions have the largest weight. Then, estimations for the most important scores are obtained from alternative customer data (e.g., yearly salary estimation). Using estimations on those basic components and their weights, it is possible to simulate the risk tolerance for those customers without questionnaires. A risk profile is estimated following this method for 7,141 investors in our dataset.

Figure 3(b) displays the distribution of customers according to their investment risk profiling. We find that most customers favor intermediate risk profiles (income and bal-

anced), with a minority willing to risk their capital on aggressive investments.

Investment capacity

The last customer categorization divides customers by the amount of money they can invest: we consider four different segments according to their investment limit: (a) less than €30,000, (b) between €30,000 and €80,000, (c) between €80,000 and €300,000 and (d) more than €300,000. These values are obtained from the risk assessment questionnaire. A similar procedure to the one used for risk profile is used for those customers without assessment (in this case, this is estimated for 7,318 customers).

The customer distribution is illustrated in Figure 3(c). In that figure, we can observe that a majority of the investors in our dataset have a low investment capacity (less than €30k). This is consistent with our customer segmentation, where more than 18,000 customers were identified as mass customers, with less than €60k on investments. The number of customers in each category diminishes as the investment capacity increases (with those customers capable of investments above €300k representing a minority of the dataset).

4 Potential use cases

Considering the information included in the FAR-Trans dataset, we envision several potential use cases for researchers in the recommender systems and investment spaces. These use cases include (but are not limited to):

- **Investor modelling:** the customer information and investment transactions might be useful to develop new models of investor behaviour for banking customers [Thompson *et al.*, 2021]
- **Financial asset recommendation:** FAR represents the main use case for which the dataset was built. The availability of customer and asset information, pricing data and transactions allow the development of price-based, transaction-based and hybrid models for the task [McCreddie *et al.*, 2022; Sanz-Cruzado *et al.*, 2022]
- **Portfolio management:** this task involves building an investment portfolio for the customers: not only identifying investments, but also estimating how much capital

they should invest on each asset and how they should modify their current investments [Markowitz, 1952].

As the main use case considered during the construction of this dataset, we provide a recommended experimental setup for the FAR task, as well as algorithmic benchmarks for future comparison.

5 Example Use Case: Financial Asset Recommendation

We provide an example use case for this dataset, where we identify potential investments for retail investors using FAR algorithms. This example provides a benchmark for assessing new developments in the FAR domain. In this work, we aim to answer the following research questions:

- **RQ1:** Which algorithms are best at identifying profitable assets for investors?
- **RQ2:** Which algorithms are best at identifying future customer investments?

5.1 Task definition

FAR systems consider two types of entities: investors (denoted as \mathcal{U}) and financial securities (denoted as \mathcal{I}). At a given time t , customers buy or sell financial assets at a price that varies according to the asset supply and demand. If we define the set of assets which a customer u has bought before time t as $\mathcal{I}_u(t) \subset \mathcal{I}$, a FAR system generates a ranking $R_u \subset \mathcal{I} \setminus \mathcal{I}_u(t)$ of those assets who the user has not interacted with in the past, based on their suitability for the customer.

5.2 Experimental setup

Dataset post-processing

We modify our transaction data so our algorithms can receive it as input. We transform them into a binary rating matrix Rel where every user-item pair represents the interest of the user on the item. Following common practice in implicit recommender systems [Ricci *et al.*, 2022], we consider that a customer u has interest on a financial asset i ($Rel(u, i) = 1.0$) if she has acquired instances of the asset. Otherwise, it is considered that the customer is not interested in that product ($Rel(u, i) = 0.0$).

Then, as the effectiveness of different recommendation algorithms naturally varies as market conditions change [Sanz-Cruzado *et al.*, 2022], it is important to examine the performance over different market conditions. To this end, we generate 61 distinct variants of the dataset, each representing a setting for a different point in time. Each variant defines a time t when recommendations are provided, and takes pricing data and investment transactions prior to t as the training data, and the pricing data and transactions in the following $(t, t + \Delta t)$ period as test. We choose Δt equal to 6 months in our experiments (i.e. we predict prices/interactions 6 months into the future). Our first time point t_0 is August 1st 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 August, t_2 is the beginning of September, and so on. To avoid contamination of the test set, if a customer acquires an asset in both training and test sets, we only keep

the training interactions. We also keep only those customers with at least one interaction in the training and test sets and assets that have pricing in the complete test period. This post filtering is important, as otherwise the pricing-based metrics and transaction-based metrics would be calculated over different customer and asset subsets, which would make them non-comparable.

Metrics

Following [Sanz-Cruzado *et al.*, 2022], we provide results for two evaluation metrics: one measuring the profitability of the provided recommendations, and another one measuring the capacity of the model to predict customer preferences.

- **ROI@k:** As a measure of profitability of the recommendations, we report the monthly average return on investment (ROI) of an equally weighted portfolio containing the top k recommended assets after a fixed time Δt . This measures how much our money would increase (or decrease) every month if we invested on it at time t .
- **nDCG@k:** We measure how close the recommendations are to the investments made by customers in the $(t, t + \Delta t)$ period using the normalised cumulative discounted gain (nDCG) metric [Järvelin and Kekäläinen, 2002] over the top k recommendations. It prioritizes relevant assets (i.e. assets acquired during the test period) in the top ranks. We consider that an asset is relevant only if u acquires i during the $(t, t + \Delta t)$ period.

For both metrics, we use $k = 10$ and Δt equal to six months.

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 11 FAR approaches from the literature, summarized below:

- **Random recommendation:** As a sanity-check baseline, we include an algorithm recommending assets randomly to customers.
- **Profitability-based models:** We test three regression algorithms, predicting ROI at $t + 6$ months: linear regression, random forest and LightGBM regression [Ke *et al.*, 2017]. We craft a selection of technical indicators based on close price as features: average price, return on investment, volatility, moving average convergence divergence, momentum, rate of change, relative strength index, detrended close oscillator, return on investment/volatility 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 them in two categories:
 - **Non-personalized:** As a 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 [He *et al.*, 2020], matrix factorization (MF) [Rendle *et*

Table 2: 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. For ROI, blue cells show an improvement over the average market value.

Data source	Algorithm	nDCG@10	ROI@10
None	Random	0.0106	0.0071
Prices	Random forest	0.0237	0.0259
	Linear regression	0.0215	0.0249
	LightGBM	0.0221	0.0225
Transactions	Popularity	0.2710	0.0006
	LightGCN	0.3404	0.0004
	ARM	0.2556	0.0007
	MF	0.1780	0.0038
	UB kNN	0.1599	0.0119
Hybrid	Hybrid-nDCG	0.2313	0.0063
	Hybrid-regression	0.0261	0.0132
Market average		-	0.0079
Customer average		-	0.0018

al., 2020] and user-based kNN (UB kNN) [Nikolopoulos *et al.*, 2022]. We also add the Apriori association rule mining (ARM) algorithm [Agrawal and Srikant, 1994].

- **Hybrid methods:** Finally, we test two hybrid methods, based on gradient boosting regression trees [Ke *et al.*, 2017; Sanz-Cruzado *et al.*, 2022]: 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 [Burgess, 2010], optimizing nDCG (Hybrid-nDCG). As features, we use the outcome of all the previous listed recommendation algorithms.

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.

5.3 Experimental results

In order to foster research over this dataset, we provide a benchmark of multiple FAR models on the dataset. We therefore report the performance of the 11 FAR approaches in Table 2 where every column represents one evaluation metric averaged over all the considered time points. The highest performing model under each metric is highlighted in bold and underlined, and the performance distribution for each metric is colour coded (blue for highly performing and red for poorly performing). From Table 2 we observe the following points of interest:

First, we observe that, in general, only a few of the algorithms are able to provide a set of assets which are profitable above the market ($\text{ROI}@10 > 0.0079$): the price-based algorithms, the hybrid model optimizing a profitability regression function and the user-based kNN collaborative filtering algorithm. Among these, the best alternatives are notably the profitability prediction models, with the three of them (linear regression, random forest and LightGBM) being able to

beat the monthly profitability of a market fund where all assets are equally weighted. From these three models, random forest regression appears as the best alternative. However, these methods fail to identify assets on which customers are interested (achieving nDCG values barely above random recommendation).

Second, transaction-based algorithms are able to reasonably predict customer preferences (as shown by their high nDCG values). We observe that the algorithm with the highest nDCG value is the most advanced LightGCN algorithm. However, we can also notice that the rest of the approaches tested are not able to outperform the non-personalized popularity-based recommendation algorithm. This follows our previous observation that 10 assets concentrate around 50% of the investment transactions in our dataset. Although collaborative filtering approaches achieve high nDCG values in our comparison, they show an overall poor performance in terms of the ROI profitability metric.

These observations allow us to answer RQ1 and RQ2: *those methods targeting a particular evaluation objective are the best optimizing that perspective at evaluation time, with price-based methods like random forest or LightGBM achieving high profitability values, whereas collaborative filtering transaction-based approaches like LightGCN stand out against other algorithms in terms of nDCG.*

6 Conclusions

In this work, we have introduced FAR-Trans, a novel dataset for financial asset recommendation that includes customer and asset information, asset pricing time series and investment transaction data from a large financial institution. The dataset spans the period between January 2018 and November 2022, covering not only bullish periods, but also periods of time impacted by external events such as the Covid-19 pandemic or the Ukraine-Russia war.

We also compare 11 recommendation algorithms in terms of their capability for recommending profitable assets and their capacity for predicting future customer investments. We find that the non-personalized profitability prediction algorithms are more capable of navigating the market prices and are therefore able to provide asset recommendations above the average market profitability. On the other hand, they fail at predicting customer investments, a task on which collaborative filtering models excel.

As future work, we shall explore the use of this dataset for multiple tasks, not only including financial asset recommendation, but also portfolio construction and optimisation or investor and asset modelling.

Acknowledgments

The work introduced in this paper was in part carried out within the Infintech project which is supported by the European Union’s Horizon 2020 Research and Innovation programme under grant agreement no. 856632. Subsequent development was also financially supported via Engineering and Physical Sciences Research Council (EPSRC) Impact Accelerator, part of UK Research and Innovation (UKRI) with grant ref. number EP/X525716/1.

References

- [Agrawal and Srikant, 1994] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB 1994)*, page 487–499, Santiago de Chile, Chile, 1994. Morgan Kaufmann Publishers Inc.
- [Alsulmi, 2022] Mohammad Alsulmi. From Ranking Search Results to Managing Investment Portfolios: Exploring Rank-Based Approaches for Portfolio Stock Selection. *Electronics*, 11(23):4019, 2022.
- [Barreau and Carlier, 2020] Baptiste Barreau and Laurent Carlier. History-Augmented Collaborative Filtering for Financial Recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems (RecSys 2020)*, page 492–497, Virtual Event, Brazil, 2020. Association for Computing Machinery.
- [Bertin-Mahieux *et al.*, 2011] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, pages 591–596, Miami, Florida, USA, 2011.
- [Bogers *et al.*, 2022] Toine Bogers, Cataldo Musto, David Wang, Alexander Felfernig, Simone Borg Bruun, Giovanni Semeraro, and Yong Zheng. Finrec: The 3rd international workshop on personalization & recommender systems in financial services. In *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys 2022)*, page 688–690, Seattle, WA, USA, 2022.
- [Burges, 2010] Chris Burges. From RankNet to LambdaRank to LambdaMART: An Overview. Microsoft Research Technical Report MSR-TR-2010-82, Microsoft, 2010.
- [Burke, 2007] Robin D. Burke. Hybrid Web Recommender Systems. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *The Adaptive Web: Methods and Strategies of Web Personalization*, pages 377–408. Springer, Berlin, Heidelberg, Germany, 2007.
- [Chalidabhongse and Kaensar, 2006] Thanarat H. Chalidabhongse and Chayaporn Kaensar. A Personalized Stock Recommendation System using Adaptive User Modeling. In *Proceedings of the 2006 International Symposium on Communications and Information Technologies (ISCIT 2006)*, pages 463–468, Bangkok, Thailand, 2006.
- [Chong *et al.*, 2017] Eunsuk Chong, Chulwoo Han, and Frank C. Park. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*, 83:187–205, 2017.
- [Council, 2014] European Parliament & Council. Directive 2014/65/EU of the European Parliament and of the Council of 15 May 2014 on markets in financial instruments and amending Directive 2002/92/EC and Directive 2011/61/EU. *Official Journal of the European Union*, 57:349–396, 2014.
- [Feng *et al.*, 2019] Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. Temporal Relational Ranking for Stock Prediction. *ACM Transactions on Information Systems*, 37(2):1–30, 2019.
- [Feng *et al.*, 2022] Shibo Feng, Chen Xu, Yu Zuo, Guo Chen, Fan Lin, and Jianbing Xiahou. Relation-aware dynamic attributed graph attention network for stocks recommendation. *Pattern Recognition*, 121:108119, 2022.
- [Gonzales and Hargreaves, 2022] Reyes Michaela Denise Gonzales and Carol Anne Hargreaves. How can we use artificial intelligence for stock recommendation and risk management? A proposed decision support system. *International Journal of Information Management Data Insights*, 2(2):100130, 2022.
- [Harper and Konstan, 2015] F. Maxwell Harper and Joseph A. Konstan. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems*, 5(4), 2015.
- [He *et al.*, 2020] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, YongDong Zhang, and Meng Wang. 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)*, page 639–648, Virtual Event, China, 2020. ACM.
- [Järvelin and Kekäläinen, 2002] Kalervo Järvelin and Jaana Kekäläinen. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems*, 20:422–446, October 2002.
- [Ke *et al.*, 2017] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS 2017)*. Curran Associates, Inc., 2017.
- [Kubota *et al.*, 2022] Kohsuke Kubota, Hiroyuki Sato, Wataru Yamada, Keiichi Ochiai, and Hiroshi Kawakami. Content-based stock recommendation using smartphone data. *Journal of Information Processing*, 30:361–371, 2022.
- [Lee *et al.*, 2014] Eric L. Lee, Jing-Kai Lou, Wei-Ming Chen, Yen-Chi Chen, Shou-De Lin, Yen-Sheng Chiang, and Kuan-Ta Chen. Fairness-Aware Loan Recommendation for Microfinance Services. In *Proceedings of the 2014 International Conference on Social Computing (SocialCom 2014)*, page 1–4, Beijing, China, 2014. ACM.
- [Liu, 2009] Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends on Information Retrieval*, 3(3):225–331, 2009.
- [Luef *et al.*, 2020] Johannes Luef, Christian Ohrfandl, Dimitris Sacharidis, and Hannes Werthner. A Recommender System for Investing in Early-Stage Enterprises. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing (SAC 2020)*, page 1453–1460, Online, 2020. ACM.

- [Markowitz, 1952] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.
- [Matsatsinis and Manarolis, 2009] Nikolaos F. Matsatsinis and Eleftherios A. Manarolis. New Hybrid Recommender Approaches: An Application to Equity Funds Selection. In *Proceedings of the 1st International Conference on Algorithmic Decision Theory (ADT 2009)*, pages 156–167, Venice, Italy, 2009. Springer Berlin Heidelberg.
- [McCreadie et al., 2022] Richard McCreadie, Konstantinos Perakis, Maanasa Srikrishna, Nikolaos Droukas, Stamatis Pitsios, Georgia Prokopaki, Eleni Perdikouri, Craig Macdonald, and Iadh Ounis. Next-generation personalized investment recommendations. In John Soldatos and Dimosthenis Kyriazis, editors, *Big Data and Artificial Intelligence in Digital Finance: Increasing Personalization and Trust in Digital Finance using Big Data and AI*, pages 171–198. Springer, 2022.
- [Musto and Semeraro, 2015] Cataldo Musto and Giovanni Semeraro. Case-based recommender systems for personalized finance advisory. In *Proceedings of the 1st International Workshop on Personalization & Recommender Systems in Financial Services (FinRec 2015)*, pages 35–36, Graz, Austria, 2015.
- [Musto et al., 2014] Cataldo Musto, Giovanni Semeraro, Pasquale Lops, Marco de Gemmis, and Georgios Lekkas. 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, 2014.
- [Musto et al., 2015] Cataldo Musto, Giovanni Semeraro, Pasquale Lops, Marco de Gemmis, and Georgios Lekkas. Personalized finance advisory through case-based recommender systems and diversification strategies. *Decision Support Systems*, 77:100–111, 2015.
- [Nikolakopoulos et al., 2022] Athanasios N. Nikolakopoulos, Xia Ning, Christian Desrosiers, and George Karypis. Trust Your Neighbors: A Comprehensive Survey of Neighborhood-Based Methods for Recommender Systems. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook, 3rd Edition*, pages 39–89. Springer US, 2022.
- [Paranjape-Voditel and Deshpande, 2013] Preeti Paranjape-Voditel and Umesh Deshpande. A stock market portfolio recommender system based on association rule mining. *Applied Soft Computing*, 13(2):1055–1063, 2013.
- [Pathak et al., 2017] Apurva Pathak, Kshitiz Gupta, and Julian McAuley. Generating and personalizing bundle recommendations on steam. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*, page 1073–1076, Shinjuku, Tokyo, Japan, 2017.
- [Pérez Maurera et al., 2020] Fernando B. Pérez Maurera, Maurizio Ferrari Dacrema, Lorenzo Saule, Mario Scriminaci, and Paolo Cremonesi. Contentwise impressions: An industrial dataset with impressions included. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM 2020)*, page 3093–3100, Online, 2020.
- [Rendle et al., 2020] Steffen Rendle, Walid Krichene, Li Zhang, and John Anderson. Neural collaborative filtering vs. matrix factorization revisited. In *Proceedings of the 14th ACM Conference on Recommender Systems (RecSys 2020)*, pages 240–248, Virtual Event, Brazil, 2020. ACM.
- [Ricci et al., 2022] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender Systems: Techniques, Applications, and Challenges. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 1–35. Springer, 2022.
- [Sanz-Cruzado et al., 2022] Javier Sanz-Cruzado, Richard McCreadie, Nikolaos Droukas, Craig Macdonald, and Iadh Ounis. On Transaction-Based Metrics as Proxy for Profitability of Financial Asset Recommendations. In *Proceedings of the 3rd International Workshop on Personalization & Recommender Systems in Financial Services (FinRec 2022)*, 2022.
- [Schumaker and Chen, 2009] Robert P. Schumaker and Hsinchun Chen. Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System. *ACM Transactions on Information Systems*, 27(2), 2009.
- [Sehgal and Song, 2007] Vivek Sehgal and Charles Song. SOPS: Stock Prediction Using Web Sentiment. In *Proceedings of the 7th IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, pages 21–26, Omaha, NE, USA, 2007. IEEE.
- [Soldatos and Kyriazis, 2022] John Soldatos and Dimosthenis Kyriazis, editors. *Big Data and Artificial Intelligence in Digital Finance*. Springer, 2022.
- [Song et al., 2017] Qiang Song, Anqi Liu, and Steve Y. Yang. Stock portfolio selection using learning-to-rank algorithms with news sentiment. *Neurocomputing*, 264:20–28, 2017.
- [Sun et al., 2018] Yunchuan Sun, Mengting Fang, and Xinyu Wang. A novel stock recommendation system using Guba sentiment analysis. *Personalized Ubiquitous Computing*, 22(3):575–587, 2018.
- [Swezey and Charron, 2018] Robin M. E. Swezey and Bruno Charron. Large-Scale Recommendation for Portfolio Optimization. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys 2018)*, page 382–386, Vancouver, British Columbia, Canada, 2018. ACM.
- [Takayanagi and Izumi, 2024] Takehiro Takayanagi and Kiyoshi Izumi. Incorporating Domain-Specific Traits into Personality-Aware Recommendations for Financial Applications. *New Generation Computing*, 2024.
- [Takayanagi et al., 2023] Takehiro Takayanagi, Kiyoshi Izumi, Atsuo Kato, Naoyuki Tsunedomi, and Yukina

- Abe. Personalized stock recommendation with investors' attention and contextual information. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*, page 3339–3343, 2023.
- [Thompson *et al.*, 2021] John R.J. Thompson, Longlong Feng, R. Mark Reesor, and Chuck Grace. Know Your Clients' Behaviours: A Cluster Analysis of Financial Transactions. *Journal of Risk and Financial Management*, 14(2):50:1–50:29, 2021.
- [Tu *et al.*, 2018] Wenting Tu, Min Yang, David W. Cheung, and Nikos Mamoulis. Investment recommendation by discovering high-quality opinions in investor based social networks. *Information Systems*, 78:189–198, 2018.
- [Wan and McAuley, 2018] Mengting Wan and Julian McAuley. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys 2018)*, page 86–94, Vancouver, British Columbia, Canada, 2018.
- [Yang *et al.*, 2015] Dingqi Yang, Daqing Zhang, Vincent W. Zheng, and Zhiyong Yu. Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1):129–142, 2015.
- [Yang *et al.*, 2018] Hongyang Yang, Xiao-Yang Liu, and Qingwei Wu. 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)*, pages 1693–1697, New York, NY, USA, 2018. IEEE.
- [Zhao *et al.*, 2015] Xiaoxue Zhao, Weinan Zhang, and Jun Wang. Risk-Hedged Venture Capital Investment Recommendation. In *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys 2015)*, pages 75–82, Vienna, Austria, 2015. ACM.
- [Zheng *et al.*, 2020] Zeqi Zheng, Yuandong Gao, Likang Yin, and Monika K. Rabarison. Modeling and analysis of a stock-based collaborative filtering algorithm for the Chinese stock market. *Expert Systems with Applications*, 162:113006, 2020.
- [Zibriczky, 2016] Dávid Zibriczky. Recommender Systems meet Finance: a Literature Review. In *Proceedings of the 2nd International Workshop on Personalization & Recommender Systems in Financial Services (FinRec 2016)*, pages 3–10, Bari, Italy, 2016. CEUR Workshop Proceedings.