COURSE – CS6242 DATA AND VISUAL ANALYTICS

Progress Report
for
Enhanced Chilean Mutual Fund Data Explorer

By,
Pedro Pablo Ramirez,
Christopher J Santiago,
Collin G Kruger,
Shannon R Flynn,
Nagasree Chelamalla.

Introduction

In Chile, limited tools and static reports exist to help Chilean retail investors understand their investment options[1, 2]. We recognize this community as underserved and aim to help retail investors make well-informed decisions by serving intuitive data visualizations, informed by a combination of machine learning methods and investment theory. Our solution will coalesce multi-platform data sets to hasten, and make possible, insights that are difficult to achieve with the status quo. Initial research indicates that the results of this project would be highly impactful, as mutual funds are one of the primary investment avenues in Chile[3], representing nearly 20% of total assets under management (AUM) in Chile (see Table 1 in the Appendix).

Problem Definition

This project has four objectives: 1) consolidate disparate sources of Chilean financial data; 2) apply unsupervised machine learning (ML) techniques to identify funds with abnormal risk/return profiles compared to their stated objective[4]; 3) apply semi-supervised ML techniques[5] to re-classify funds based on their historic performance[6], vice their stated objectives, for portfolio optimization via Modern Portfolio Theory[7] (MPT); and 4) create a user interface (UI) to visualize distributions and key statistics of Chilean mutual funds, any identified abnormalities, and correlations of newly-clustered funds.

To provide an interactive UI to the people of Chile, we are using Tableau as the software for implementation. Tableau is beneficial because it's dashboards and packaged workbooks can be accessed anywhere by anyone through utilizing available platforms, such as Tableau Reader, Tableau Server, Tableau Public, and embedded workbooks in personal websites. For this reason, we decided Tableau would be an excellent resource to enable the people of Chile for their investment fund decision making. The UI created in Tableau is going to consist of three major dashboards for interaction:

1. Chilean Mutual Funds - Anomaly Detection: This dashboard will be used to allow investors to view a self-organized map of the available funds and observe outliers and fund misclassifications. They will be able to view the reclassifications of each fund to enable them to understand which funds have been misclassified, and the proper classification of each. Filters will be available to allow users to choose what information they are seeing, such as filters for asset categories and fund names.
2. Chilean Mutual Funds – Similarity (or Dissimilarity): This dashboard will be used to allow users to view funds that have been classified as similar (or dissimilar) to a selected fund from a table. They will be able to select a parameter to choose whether they are viewing similar funds or different funds, which will be determined from the covariance matrix. The information that will be displayed will include a line chart of risk/return of the selected fund over time, scatter plots to show how other funds compare to the chosen fund for risk/return and risk/price, and various other features. Filters will be available to allow users to choose what information they are seeing, such as filters on date, country, asset classification, bank, etc.
3. Chilean Mutual Funds – Covariance Matrix: This dashboard will be used to display the updated and improved covariance matrix. Users will be able to quickly visualize how funds compare to each other utilizing color and size. Filtering and highlighting functions will be available to provide the user with more in-depth information where needed.

Survey

In industry we see a range of state-of-the-art practices in finance, from "robo-advisors" that employ MPT for the masses, to natural language processing (NLP) sentiment analysis with neural networks (NN)

deployed by hedge funds with unfathomable resources. Chilean investors, however, are limited to price monitoring and investment consulting services. In Chile, much of the reasoning behind specific asset allocations are hidden behind "black boxes" that lack online, consumer-led interaction[1]. Further, investment companies often display bias in their recommendations, pushing their own funds onto investors. Overall, Chilean retail investors lack access to open and transparent investment analytics.

MPT provides a method for finding an optimal combination of assets that maximizes an investor's return, given their specific risk tolerance[7]. MPT can fall short, especially during financial crises, if it over-weights an investor's exposure to correlated risk assets. This happens if the covariance matrix is mis-specified, which can arise when assets are arbitrarily grouped based on investment styles or descriptions, vice underlying risk and return profiles. We plan to use semi-supervised clustering methods[6,11] to better classify Chilean mutual funds and improve MPT optimization.

Proposed Method

There are several unsupervised ML techniques[4,8] that, paired with intuitive visualizations, could help Chilean investors make well-informed decisions. Originally, we hypothesized Autoencoders and Principle Component Analysis would yield good results for finding outliers in mutual fund classification; however, at least for the nature of the data that we have, this was not the case due to the final useful data scale being smaller than anticipated (we started with ~4,000,000 records and have cleaned and transformed it down to 312 records).[9] We discovered a fairly recent technique of clustering called t-SNE which became our resolution to the Autoencoder and PCA problem. Using pairwise distances from our t-SNE embedding we were able to identify funds whose Euclidean distance was far from their respective category mean. Along the way to this resolution, we experimented with several other dimensionality reduction embedding techniques including Multi-Dimensional Analysis[10], Isomap, and Uniform Manifold Approximation and Projection (UMAP)[20]. Ultimately, we decided that t-SNE best embedded our data for visualization in a two-dimensional space.

Risks

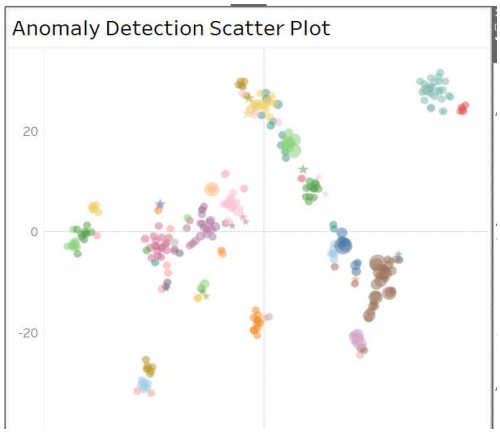
Bugs are always a present danger in software development. Given that our software is ambitious and that we are writing it outside of our daily jobs, there is significant risk that we implement an algorithm incorrectly, transform data incorrectly, or suffer similar errors. Studies in software development have found that, typically, "about 15 – 50 errors [occur] per 1,000 lines of delivered code[12]." This means that we should be very conscious about delivering accurate software, and that our initial users should not use our software as their only source of information—perhaps using it for simulated trades, only, until they feel comfortable with the data and software reliability. However, if this project is successful, it will greatly benefit the people of Chile and therefore we believe the payoffs will outweigh the risks.

Continued Research

Due to time and immediate data availability constraints, there were several areas of research that we did not include in this project: 1) how political parties affect the selection of mutual funds[13]; 2) how predicting the performance of constituent stocks can be used to predict mutual fund performance[14]; 3) application of evolution-inspired algorithms to make predictions[15]; and 4) using support vector machines (SVM) to identify types of assets and groups for building a covariance matrix[16]. These ideas would be solid options for follow-on work, given more time and data.

Checkpoint

The initial version of our software will take the remainder of the semester to complete. We plan to give the completed software to our Chilean teammate, so that he can open source and improve upon it, as he sees fit, to maximize value to Chilean citizens. At this point in the semester we have 1) cleaned and transformed the data into a form useful for storage and retrieval 2) implemented some of our algorithms (discussed further below) 3) completed initial storyboarding of our visualizations 4) completed a first draft of one dashboard in Tableau. This places us on track for delivery at the end of the semester.



List of Innovations

This project is unique because it will enable the people of Chile to make informed decisions about their retail investments with data and visualizations that have not been available previously. The highlights of this project are:

- Reclassifying asset classes that are potentially misclassified to allow for a better understanding of available funds and how they behave.
- Creating an updated covariance matrix for an improved MPT to ensure that retail investors are getting the most out of their investment portfolios.
- Open sourcing the software.
- Coalescing disparate Chilean data sources.

Design of upcoming experiments / evaluation

As stated previously, a full user study likely will not be possible due to privacy and geographic limitations. Therefore, in order to gauge the impact of this project (if fully implemented), we could observe web traffic data, including geographic information, to understand how many people are utilizing the tool and where they are located. Our Chilean teammate plans to expand on this semester's work by extending the software where appropriate, embedding it in a public facing website, and evangelizing its usefulness to the Chilean retail investment community through his work at Itau Chile Bank.

To lieu of a full user study, a usability analysis will be conducted before the end of the semester to gauge the ease of use of the new tool compared to the currently available platform. Our Chilean teammate will recruit participants from his community to use the current tool, and the new tool, and then complete a System Usability Scale (SUS) industry standard survey for both to determine whether the new tool improves usability compared to the current tool.

Completed and Planned List of Activities

From our initial plan the following is a condensed list of tasks that were accomplished.

- Pedro completed scraping mutual fund data from publicly available sources.
- Christopher, Collin, and Nagasree completed extracting, cleaning, and transforming the raw scraped data into usable data.
- Pedro, and Christopher have begun work on ML and statistical models.
- Shannon has prototyped several Tableau dashboards.
- The entire team contributed to both the proposal and this progress report.
- Collin prepared the presentation video and slides.

The following is a condensed list of our next tasks.

- Christopher, Pedro, Nagasree, and Collin will continue development of ML and statistical methods.
- Shannon will continue development of Tableau dashboards.
- Shannon, and Pedro will work on usability assessments.
- Nagasree, Shannon, and Collin will work on the final presentation.

Conclusion

Currently, at the midterm checkpoint, we are on track with the schedule and plan of activities from above. We have successfully completed extraction of data, all required data cleaning, translation from Spanish to English, removing funds that do not currently exist, conducting mean replacement for missing time series data, joining disparate data sets, consolidating asset categories into fewer groups to allow for a larger within-groups sample size, and various other methods. We have started with around 4 million fund records initially in ETL, upon joining with time series returns and cleaning the data finally ended up with 312 records. We have also implemented many of the machine learning algorithms and financial methods as planned and have pivoted away from some algorithms due to poor performance on the data (Principal Component Analysis and Autoencoders). We have started using the t-SNE[19] clustering algorithm (new since we submitted the proposal) as a visualization technique for projecting 72 periods (dimensions) of mutual fund returns into a 2-dimensional space. We then used the embedding matrix from t-SNE to identify outliers/anomalies based on Euclidean distance from their respective category means. We next plan to test K-means and other clustering methods to re-classify funds based on either their t-SNE embeddings and financial performance metrics or underlying fund holdings (dependent on our analysis). The clustering results will be used to group the funds for later use in MPT optimization. We also are performing time series analysis of risk and return for each fund, and creating an updated covariance matrix to improve MPT optimization. Additionally, storyboard ideas for three dashboards have been completed, and the first draft of one dashboard has been developed in Tableau. The remaining tasks to complete are importing the remaining data into Tableau to create interactive user visualizations, add remaining ML/statistical algorithms for clustering/outlier analysis, and compile information to complete a final report and poster. Each team member has completed tasks as anticipated and stated in the proposal plan above. Additionally, to date, all team members have contributed similar amounts of effort in their own areas of expertise.

References:

1. Informes de Recomendacion (2020). Sura. SURA Asset Management. https://inversiones.sura.cl/nosotros/Paginas/informes-de-recomendacion.aspx
2. IPSA, Chile (2021). Btg Pactual. CME Group. https://www.mercadosenlinea.cl/www/chile/resume.html
3. Ahumada, L., Alvarez, N., & Diego, S. (2011). Valorización de Fondos Mutuos Monetarios y su Impacto sobre Estabilidad Financiera. Central Bank of Chile.
4. Hastie, T., Tibshirani, R. & Friedman, J. (2009). *Unsupervised learning.* The elements of statistical learning (pp. 501-528). New York, New York: Springer Science+Business Media, LLC.
5. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An introduction to statistical learning with applications in R. New York, N.Y: Springer.
6. Mehta, D., Desai, D. & Pradeep, J. (2020). *Machine learning fund categorizations.* ArXiv.
7. Markowitz, H. (1952). Portfolio selection. Journal of Finance, vol 7(1), (pp. 77-91).
8. Dixon, R., Halperin, I. & Bilokon, P. (2020). *Autoencoders*. Machine learning in finance: From theory to practice (pp. 266-271). Cham, Switzerland: Springer Nature Switzerland AG.
9. Kim, M., et al. (2000). Mutual Fund Objective Misclassification. Journal of Economics and Business (pp. 309–323.).
10. Aflalo, Y., Dubrovina, A. & Kimmel, R. (2016). Spectral Generalized Multi-dimensional Scaling. Int J Comput Vis 118, 380–392. https://doi.org/10.1007/s11263-016-0883-8
11. Pattarin, F., et al. (2004). Clustering Financial Time Series: an Application to Mutual Funds Style Analysis. *Computational Statistics & Data Analysis (*pp. 353–372).
12. McConnell, S. (2004). Code Complete, Second Edition. Microsoft Press, USA.
13. Bubb, R. & Catan, E. (2020). The Party Structure of Mutual Funds. European Corporate Governance Institute - Law Working Paper 560/2020. http://dx.doi.org/10.2139/ssrn.3124039
14. Li, B. & Rossi, A. G. (2020). Selecting Mutual Funds from the Stocks They Hold: A Machine Learning Approach http://dx.doi.org/10.2139/ssrn.3737667
15. Kyong Joo Oh, Tae Yoon Kim, Sungky Min (2005). Using genetic algorithm to support portfolio optimization for index fund management. Expert Systems with Applications, Volume 28, Issue 2, Pages 371-379, ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2004.10.014
16. Guglietta, J. (2018). *Support vector machine-based global tactical asset allocation.* Big data and machine learning in quantitative investment (pp. 211-224). John Wiley & Sons, Ltd.
17. Tufte, E. R. (2001). The visual display of quantitative information. Graphic Press.
18. Geraldi, J., & Arlt, M. (2015). *Visuals Matter! Designing and using effective visual representations to support project and portfolio decisions.* Project Management Institute.
19. van der Maaten, L. & Hinton, G. (2008). Visiualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov):2579-2605.
20. Leland McInnes, John Healy, James Melville (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://arxiv.org/abs/1802.03426

Appendix

TABLE 1: Data obtain from the Chilean Financial Regulator (Financial Market Commission "CMF")

Assets Under Management

Institution	AUM	Market Share
Mutual Funds	62.500	18%
Pension Funds	200.000	58%
Insurance Companies	60.000	17%
Private Funds	25.000	7%

***MMUSD**