COURSE – CS6242 DATA AND VISUAL ANALYTICS

Project Proposal
On
Enhanced Chilean Mutual Fund Data Explorer

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

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).

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. We can gauge the impact of this project (if fully implemented) by observing web traffic data, as a full user study might not be capable due to privacy and geographic limitations.

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.

There are several unsupervised ML techniques[4,8] that, paired with intuitive visualizations, can help Chilean investors make well-informed decisions. We will use reconstruction error, derived from an autoencoder NN architecture and/or principal component analysis, to identify mutual funds that exhibit anomalous behavior as compared to their stated investment objectives or peer group[9]. Utilizing multi-dimensional scaling (MDS) we will visualize a self-organizing map (SOM) of mutual funds, based on their underlying characteristics[10].

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.

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.

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. By midterm, we aim to 1) have the data cleaned and stored in a database; 2) begun writing our algorithms; and 3) completed initial storyboarding for visualizations. At the

final, we will measure success by 1) having completed sufficient ML algorithms; and 2) having an efficient, intuitive and interactive data visualization to serve the people of Chile.

Turning to project costs, we can scrape or download data from publicly available sources. Immediate expenses include a minimal amount (less than $10/month) for cloud hosting and our own time in building the project. Were the project to continue, past this semester, we would incur common software development lifecycle maintenance and improvement costs. We cannot predict, at this time, these associated costs, however, we would assume that they would be similar to our current expenses.

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.

We have proposed a list of tasks and assigned team members responsibility. Pedro, Collin and Nagasree will complete initial database queries; we expect these tasks to take approximately one and a half weeks. Next, Christopher, Collin, Nagasree and Pedro will implement various machine learning algorithms and financial methods. Shannon will, concurrently, complete story boards for visualizations. We expect these tasks to take three weeks. Once the algorithms are complete, Shannon will spend three weeks utilizing the data and storyboards to create interactive visualizations in Tableau. She will consult industry standards and best practices for creating visualizations with quantitative data[17] and decision making[18]. Finally, all team members will compile information and complete a final report and poster, which we expect will take one to two weeks to complete. To date, all team members have contributed similar amounts of effort and we have distributed tasks in a similar manner.

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.

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**