Aalto University
School of Science

AALTO UNIVERSITY

SEMINAR ON CASE STUDIES IN OPERATIONS RESEARCH

MS-E2177

Data-driven categorization of stocks in asset portfolio management: Project plan

Tommi Kantala (Project manager)

Hannu Pantsar

Ilkka Särkiö

Client: Elo

April 11, 2018

1 Background

Finland has a mandatory pension system instated by law which requires employers to fund pension insurance companies. Pension insurance companies have large investment capital compared to other Finnish, and even international, banks and investors. Combined with their legal duties to responsibly ensure stable capital growth the large capital provides interesting viewpoints in equities investments and portfolio management.

Elo was founded in January 2014 as a merger of Eläke-Fennia and LähiTapiola Eläke and is the third largest pension insurance company in Finland by capital. The project is made in cooperation with the equities investment team at Elo, located in Tapiola, Espoo. Elo operates its investments based on systematic strategies where investment decisions are based on quantitative analysis of market, fundamental and estimate data. Risk diversification is an important factor in the investment decision process. To avoid risk concentrations, assets are spread with respect to risk across asset classes, markets, geographic locations and industries. Traditionally diversification across industries relies on recognized classification systems where each stock is categorized in a single industry. It is thus assumed that the risk profile within each industry is similar and different to other industry classes. However, it is easy to come up with examples where this assumption does not hold e.g. in materials industry where products behaving similarly to consumer staples and raw material products which tend to behave cyclically are grouped together.

The goal of the project is to find an alternative categorization method to provide robust measure for risk diversification based on stock returns, financial statements and macroeconomic data. The model is then used to assess the level of diversification of portfolios with respect to the developed new categorization and compare performance against readily available classifications such as GICS.

2 Objectives

During the initial talks within our team and with Elo, several objectives were formed. The main objective is to find a viable technique to predict a clustering of stocks with respect to risk to support portfolio investment decisions. There are a few alternative approaches but the team has decided to focus on first finding ways to model stock covariances data and producing viable forecasts, and then clustering stocks based on the forecasted correlations. This would give us an insight of the similarity on the risk profiles of stocks e.g. in the coming fiscal quarter.

The first part of the main objective has strong foundations in literature as modeling and forecasting financial time series data is literally a "billion dollar question". As a part of reaching the objective, a literature analysis is made to base the methods on a sound foundation. The second part, the clustering of the stocks based on correlations or other similarity measures is close to any clustering problem. Through the increasing popularity of machine learning methods, it should be straightforward to assess different approaches for the clustering objective.

As a follow up objective, the team will evaluate the found clustering scheme and its performance against readily available categorizations such as GICS. Examples of the evaluations are comparisons of previous Elo portfolio allocations with respect to the clusters our team has found. Interpreting implications of why some cluster are or are not represented in the portfolio as well as finding interpretations on the logic behind found clusters is also important.

The third objective to wrap the project into a tool to support portfolio diversification. This objective is given less emphasis during the project to ensure main objective success. The tool would, in minimum, be a wrapper for project code and it would produce forecasted clusters based on provided data. Graphical representations and features to compare clusterings to GICS would be a plus. These features would be useful for the portfolio manager to support the decision making process.

3 Tasks

When having a very strict scope, the project has two core tasks. These are 1) Forecasting the volatility correlation matrix of assets and 2) Performing a clustering of the assets based on the forecasted correlation matrix. These two core tasks are supplemented by performing literature review, cleaning the data and analyzing it, validating the results, writing the report, giving feedback and other general project related tasks. Required tasks and their schedule can be seen in Figure 1.

3.1 Client interaction

The client has a team of six persons handling the project work from their side. Due to proximity of the client, communication is mostly done by meeting the client in person. E-mail is used for arranging meetings and some more pressing questions with smaller scope. Interaction with the client is frequent and every meeting in person with the client has a goal-oriented agenda that is decided before the meeting. In the current project schedule in Figure 1 only meetings which are already booked with the client are shown. More meetings will most definitely be arranged with the client later during spring. In the first project meeting with the client included going over the objectives of the project and what the client expects from us. In the later meeting the dataset received from the client was discussed with the client answering our questions regarding the data. In this meeting the scoping of the project was clarified yet again. Later meetings with the client are steering meetings where the progress of modeling work is shared with client and used models and the received results are evaluated together. Minor changes to scope and direction of the project work are possible as a resolution of these evaluation sessions. (Last meetings with the client in April are meant for discussing the wrapping up of project and what the client wants to see in the final report.)

3.2 Reporting

Reporting of the project work consists of Project plan, Interim report and the Final report. In the Interim report the completed activities so far are summarized and updates to the initial project plan are documented. The Final report is much larger than the other two deliverables and includes reporting on the main results of the project. In our case we assess whether our model is reasonable and if the results of our model bring value to the client-side objective of using the clusters as a tool in performing portfolio allocation. The model and other methods used in the project (i.e. how did we get to the result we have) are documented in the Final report. In addition, we self-assess the execution of the project and document which things could have been done better for the optimal project success.

3.3 Modeling

3.3.1 Literature review

The core of the project work is reviewing literature related to the task at hand. Publications on performing financial time series forecasting and clustering are sought after and read. Specifically those kind of publications where the methods are used for stock volatility assessment and clustering of stocks are prioritized. Literature on the subject is searched from the university and internet databases and requested from the client side.

3.3.2 Exploratory data analysis

After receiving the required dataset of stock returns from the client side, the data is analyzed and reformatted so that we can use our modeling software to process the data. A smaller but still representative subset of data is extracted from the main data set for future use of testing different kind of forecasting and clustering models. The whole dataset is cleaned from data that would cause too much effort to testing the model within the scope of the project work. This can be performed due to the excessively large size of the dataset. Possible shortcomings and bias in the data is thoroughly analyzed.

3.3.3 Forecast + clustering model prototype

A prototype forecast+clustering model combination that is able to process the data input and output even somewhat reasonable results is created.

3.3.4 Improved forecast + clustering model

A different kind of forecast + clustering model combinations are iteratively tested for the dataset and the results are compared with each other. The best model to fulfill the set objectives is chosen as the final model which is used to get the final results presented in the Final report. The output clusters of the models are compared to the current industry standard sectors for assets. The output clusters are also validated against clusters drawn from historical data.

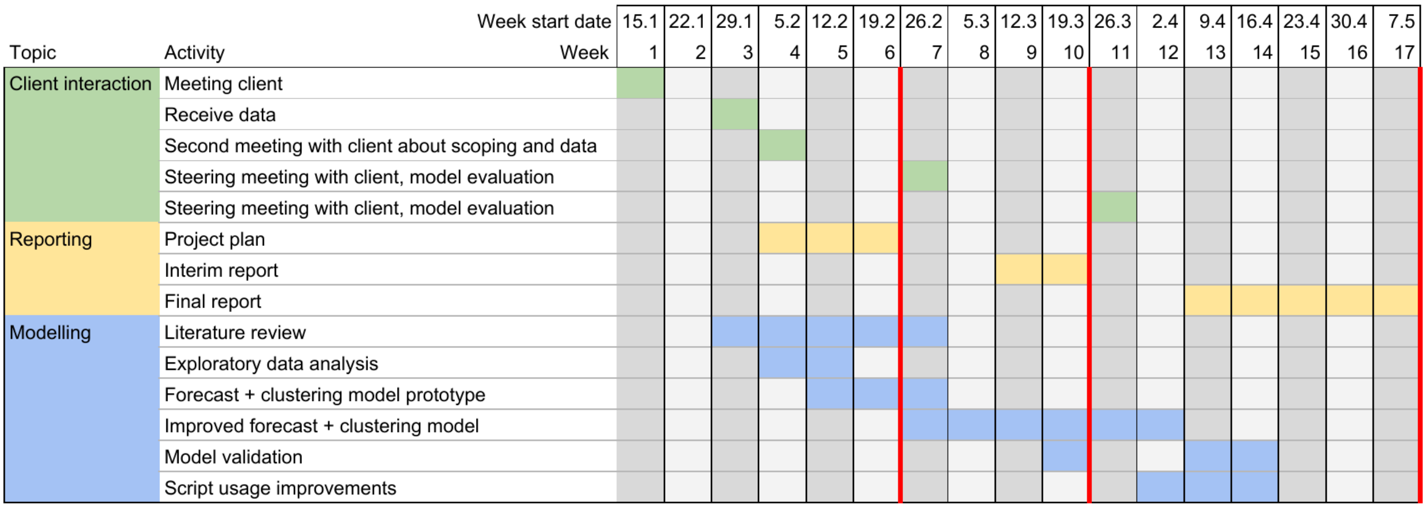
3.3.5 Model validation

The models being tested for the dataset are validated together with the client to ensure the optimal result that fulfils the set objectives.

3.3.6 Script usage improvements

The used computer scripts are modified to output the results in a way that the results can be presented in a clear way in the Final report and presentation. The model should be usable with new datasets without a need for major redo of the code.

4 Schedule



		Week start date	15.1	22.1	29.1	5.2	12.2	19.2	26.2	5.3	12.3	19.3	26.3	2.4	9.4	16.4	23.4	30.4	7.5
Topic	Activity	Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Client interaction	Meeting client		🟩																
	Receive data				🟩														
	Second meeting with client about scoping and data					🟩													
	Steering meeting with client, model evaluation								🟩										
	Steering meeting with client, model evaluation												🟩						
Reporting	Project plan					🟨	🟨	🟨											
	Interim report										🟨	🟨							
	Final report														🟨	🟨	🟨	🟨	🟨
Modelling	Literature review				🟦	🟦	🟦	🟦	🟦										
	Exploratory data analysis					🟦	🟦												
	Forecast + clustering model prototype						🟦	🟦	🟦										
	Improved forecast + clustering model								🟦	🟦	🟦	🟦	🟦	🟦					
	Model validation											🟦			🟦	🟦			
	Script usage improvements													🟦	🟦	🟦			

Figure 1: Project schedule.

5 Resources

Our team consists of three Systems and Operations Research students. Our background in financial modeling is not extensive, but we are all more or less familiar with a range of mathematical and statistical modeling techniques as well as being fluent in R programming. We also have some experience with machine learning algorithms and working with large data sets. Hannu will be absent in May, which may increase the workload of the rest of the team or delay the process, if the project is not nearly finished in the beginning of May.

Our client has a team of quantitative analysts with Systems and Operations Research or Industrial Engineering and Management background. The client team is eager to help and very willing to have discussion sessions as often as needed during the process. They also work on somewhat similar problems in their day-to-day work giving us useful tips. The initial data provided covers daily returns for European stock market for the last 20 years, with some size limits. The list of companies covers 3050 tickers. The client has also provided us with the classifications date of the stocks, countries and broad list of financial statements on as a sparser time series. The data suffer form survivor bias, since only companies existing at the end date are included. To counter this, the company will later provide a time-series with several end dates, which yields more accurate info about the evolution of the stock market.

On a more technical side, we have a possibility to use the university's servers Brute and Force to run R scripts. This may be much needed, since the datasets are quite large and we will needed thousands of time-series forecasts. We also have access to scientific journals through the university's library.

6 Risks

All projects have risks, some general and others stemming from the topic itself. We have tabulated some of the risks our project may face, their probabilities, impacts and mitigation plans in Table 1.

Table 1: Risks, impacts and mitigation plans

Risk	Probability	Impact level	Impact	Precautions and mitigation
Team member inactivity	Low	Medium to High	Project delays	Member know each other quite well which reduces the risk, early intervention if problems arise
Too great workload	Medium	Medium	Project delays, quality suffers	Align project goals with client, have active discussion with client, focus on main tasks
Forecasting inaccurate	Medium	High	Project delays, model has to be changed, whole project at risk	Understand and agree on the assumptions and limitations of the model
Clustering analysis doesn't generalize	High	Medium	Project delays, clustering strategy might be changed	Agree with client that this might be viable result
Data delivery delays	High	Low to high	Depends on missing data, delays project	Active communication with client
Computational resource insufficiency	Low	Medium	Slows down project progress or shrinking of project scope/model complexity	Ensure online computational resources early on (e.g. Aalto University servers) and pay attention to program efficiency during project implementation
Team member absence in late May	Certain	Medium	If project is late in May, smaller team size may result in delays and excessive workload	Schedule must be observed to ensure fast enough progress