Machine Learning and Predictive Analytics in the Context of Portfolio Building and Optimisation

DBA3803 Predictive Analytics in Business

October 9th 2023

Group 19

Michail Eric Marechal (e1120253), Zofia Maria Sosnowska (e1120255), Sofia Roche Vidaurre (e1120271)

Introduction

This paper is to enhance predictive modelling against an equal weighted portfolio and data exploration and manipulation. Firstly, overfitting concerns will be dealt with using a Lasso and Ridge regression. Given the limited number of observations (252) and numerous covariates (99), a reduction in covariates will be performed. Then, using Fama and French's 3-factor model as inspiration, we will focus on narrowing down the covariates in the lower median for the size factor (ME1-5) and upper median for the value factor (BM6-10). Consequently, the refined selection should adhere to the rule of thumb, mitigating overfitting risks. Next, the study will delve further by restricting asset weights depending on observed extremes, to mitigate volatility. This should result in performance improvements, yielding better cumulative returns and a superior Sharpe ratio. Thus, while LASSO and Ridge aim to address overfitting, additional constraints on asset weights will be used to enhance the model's predictive power in the investment landscape.

Data[1]

The investigation began with questioning how the data could be manipulated to improve our predictive model while maintaining the aim of beating the Equal Weighted portfolio (EW). First, problems of over-fitting were addressed using the, generally accepted, rule of thumb; #observations should be greater or equal to the #covariates multiplied by ten. With 252 observations, representing daily returns for the past 252 trading days, and 100 portfolio assets representing 99 covariate the model would certainly overfit. With the given data, the rule of thumb resulted in 252 > 9900 confirming a simple Lasso/Ridge regression would overfit. The first possible solution would be to increase the number of observations, however, the assignment asked us to perform the prediction with a training set of only 252 observations. Additionally, in the case of predicting results for publicly traded securities, the stock market is subject to quickly changing trends, so it is not logical to add more past data, as it would not bring much additional predictive value.

Therefore, we were left with our second solution of decreasing the number of covariates. This would entail reducing the potential number of assets our model could invest in and calculating the respective weights. The next logical step was to decide which criteria to use to select the assets which would remain in our investable portfolio. To make this decision we relied on the 3-factor model developed by Fama and

[1] We used a refined Excel Sheet comprised of only the relevant data observations for this prediction model. Therefore, in order to execute our code please use the attached Excel file.

French in their paper "*The cross sectional of Expected Stock Returns"* (1992)[2]. The starting point for their work was the Capital Asset Pricing Model, mostly known simply as the CAPM model[3]. This model places securities along a vertical line and explains the difference in the securities returns by one factor, which is a linear relation with market risk characterised by β, which resulted in the following equation: Cost of equity = r(the risk free rate) + β * (Market excess). However, Fama and French, when analysing data from previous decades of securities returns, observed that there were two additional factors that explained a significant part of the errors obtained with CAPM. These two new factors were Small minus Big, which represented that small cap stocks historically outperformed big cap stocks and High minus Low, which represented that value factor expressed by the book to market equity ratio.

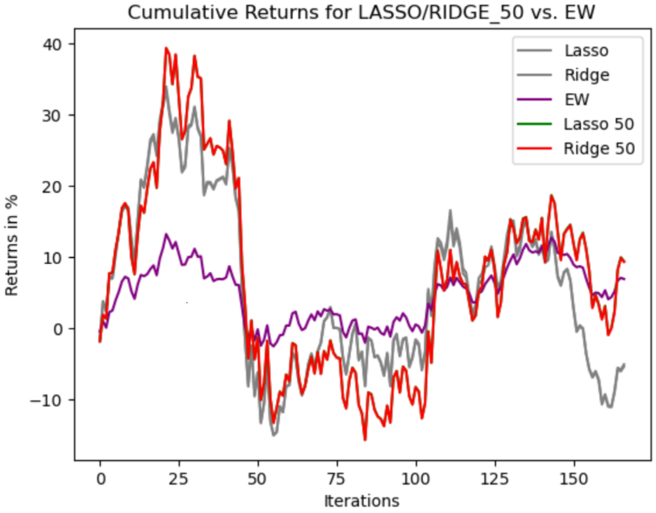
Based on this research, we focused on selecting assets in the lower median for size factor and upper median for the value factor. As the data was obtained from Fama and French's website, the titles of the columns allowed us to locate the assets of interest efficiently. The columns are named: ME number 1 to 10 and BM number 1 to 10. ME stands for market equity and represents the size factor and its corresponding number expresses the deciles in which it falls. BM stands for book to market value and the juxtaposed number represents the deciles in which the particular asset falls in. Therefore, the assets in the lower median for the size factor, with an index of ME1-5, and assets in the upper median for the value factor, BM6-10, were selected for our portfolio. Only assets that respected both conditions were chosen to take full advantage of the supposed better historical performance of the size and value factors. This left us with 24 assets, meaning 23 covariates, which meant that our Lasso and Ridge regression would not overfit as the rule of thumb of 252 > 23.10 is true.

Results

We successfully tested our model with truncated data against the original EW. We kept the original EW invested equally in the original 100 assets and did not create a new equally weighted portfolio only comprised of the 24 selected assets to compare against. This is because we assumed EW to be a naive benchmark and therefore it does not perform factor selecting. Consequently we kept the original 100 assets EW portfolio as our benchmark. The obtained results were satisfactory, as we managed to beat EW in terms of cumulative returns. Our portfolio delivered cumulative returns of 9.4159% for Lasso and

[2] Fama, E. F., & French, K. R. (1992). The Cross-Section of Expected Stock Returns. *The Journal of Finance*, *47*(2), 427–465. https://doi.org/10.2307/2329112
[3] Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, *19*(3), 425-442.

9.3211% for Ridge, whereas EW returned 6.9126% cumulatively, and -5.3721% and -5.0489% for Lasso and Ridge, respectively, for the entire data sample. [4]



However, our graphic for cumulative returns showed that our portfolio did not consistently beat EW and only managed to outperform it by the end of the training sample while bringing a lot of volatility to the returns. This observation was confirmed by the Sharpe ratios, where our portfolio returned 0.0329 for Lasso and 0.0328 for Ridge, whereas EW yielded 0.0412, and 0.0053 and 0.006 for Lasso and Ridge for the full portfolio. This data shows our truncated portfolio although it brought better returns it also brought more than proportional risk. Therefore, for the metric of returns adjusted for risk(volatility) EW is still more performant.

More about our of usage of Lasso and Ridge in the different models in the constraints section.

Loss Function

The two metrics used to evaluate the accuracy of our predictive model were median absolute error (L1) and mean squared error (L2). Given the nature of the risky assets in the portfolio employing the L2 function was better suited for our data. This is due to median absolute error being robust to outliers while MSE penalises them heavily. In investing extreme events are particularly important as they can greatly impact an investment portfolio. The outliers can either indicate above average returns or substantial losses. Therefore, it is important to assign proper weights and penalties, according to the results, to manipulate the data effectively.

Structure

To estimate the future returns, based on our model, a linear structure was used. This structure provided us with necessary complexity a simple one-number structure would not. The equation corresponding to our model consists of the summation of returns of particular assets multiplied by their corresponding

[4] Lasso/Rigde stand for a simple Lasso/Ridge regression
EW stand for the Equally Weighted Portfolio, which is our benchmark of reference
Lasso/Ridge 50 stand for a Lasso/Ridge regression on the truncated data, as we used the 50th percentile as the demarcation line

coefficients, which depict the degree of influence the return of a given asset has on the prediction of the model overall.

Constraints

Lasso:

The first attempt at restricting the model, addressing the overfitting issue, and making better predictions was through the use of LASSO regularisation. Restricting the region of the coefficients present in the linear equation depiction of our model resulted in some of them being equalised to zero. The coefficients that became zero contributed to decreasing the amount of noisy data affecting the outputs and the inferences coming from the model.

When running an analysis after performing the LASSO regularisation on the entire dataset, it was found that the returns coming from the regression performed better than a simple equal-weighted portfolio in only some of the iterations. Moreover, the overall cumulative returns proved to be negative when LASSO was introduced and equalled -5.3572%, while the cumulative returns of the model without constraints were equal to 6.9126%. To adjust for the risk connected with the investments, Sharpe ratios should be compared, in the case of LASSO-adjusted portfolios, said ratio will be 0.0053, while the one for the equal-weighted portfolio will be higher at 0.0412. This result points towards the EW being the superior portfolio type.

To explore the implications of using LASSO further, an analysis was performed on a dataset truncated in accordance with the theory put forward by E. Fama and K. French. Decreasing the size of the dataset and choosing the well-curated data points proved to be beneficial as the average cumulative return of the model with LASSO regularisation was higher than the one of the equal-weighted portfolio. The corresponding cumulative returns in the two aforementioned cases were the following: 9.4159% and 6.9129%. To further evaluate the outcomes of the analysis once again the Sharpe ratio comes into play. For the former LASSO-adjusted portfolio decreased in size it is equal to 0.0329, which is still lower than the one of the EW portfolio (0.0412).

Performing the LASSO regularisation did in fact contribute to the alleviation of the overfitting problem in the dataset at hand. Therefore, it can be concluded that the inferences drawn from the adjusted model could be generalised more freely and the risk of the outcomes mirroring the training dataset too closely was minimised to a large extent.

Ridge:

The second attempt at restricting the model was through the use of Ridge regularisation. Restricting the region of coefficients, in this case, resulted in some of them decreasing significantly, hence making the impact of some of the realisations less impactful on the overall output of the model. It can be inferred that the noise variables should now have less influence on the returns and the inferences drawn from the model will be based on significant data points to a large extent.

Similarly, the outcome of the analysis run on the whole dataset after the LASSO regularisation, after performing Ridge, the returns yielded from the regression performed better than a simple equal-weight portfolio in some but not the majority of iterations. With the cumulative returns of the Ridge model being -5.0489%, and the cumulative returns of the equal-weight portfolio being 6.9126%, we can see that the simple model performed better overall. The plot of returns actually proved to be almost identical to the one from the post-LASSO analysis when data points from the entire dataset were analysed. To further deepen the applicability of the model, a risk-adjusted performance measure known as the Sharpe ratio will be introduced. It is equal to 0.00598 for the post-Ridge portfolio, and 0.0412 for the equal-weighted portfolio. Highlighting the supposed superiority of the inferences drawn from the simpler portfolio design.

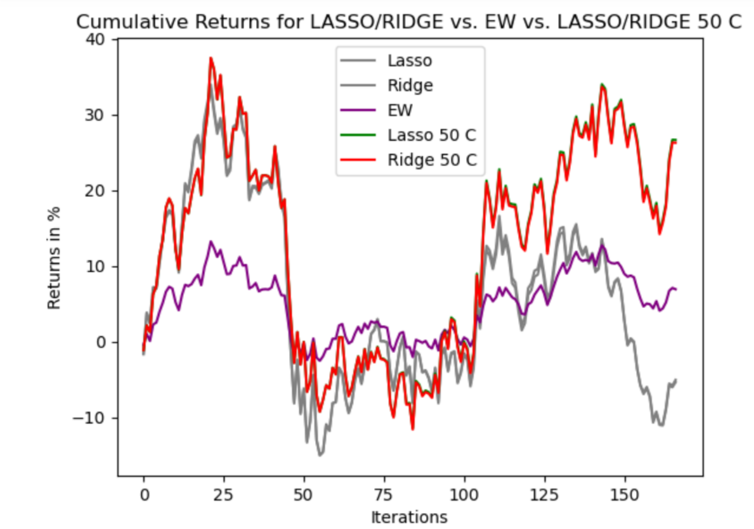
To make for a cohesive analysis of the problem, Ridge regularisation was also performed on the truncated dataset. Again the regression of the dataset cut down in accordance with the Fama-French three-factor showed that the average cumulative return of the model with Ridge regularisation was higher than the equal-weight portfolio one. After the introduction of the constraints, the cumulative returns of the model equalled 9.3211%. In comparison, the unconstrained simple model's cumulative returns remained at 6.9126%. Not omitting the risk associated with the prediction of the return, the comparison of the Sharpe ratios between the two portfolio designs was also performed. In the case of the portfolio constrained by the Ridge regularisation, the value of the aforementioned statistics was 0.03278 and for the EW portfolio, it was again 0.4121.

When talking about the predictions yielded from the analysis performed on the truncated dataset, it is worth mentioning that, just like in the case of the analysis run on the entire dataset, the plots of the post-Ridge and post-LASSO analyses are almost identical. However, both of the portfolios based on the smaller sample size performed better than both the ones based on the full ones and the equal-weight portfolio.

Further identification of the pitfalls of the model were conducted with the help of the restriction of weights imposed on the included assets. This was introduced after realising that the model would take on extreme positions in certain assets which brought about large volatility. In order to combat the issue at hand, a limitation of weights approximately within the 5th and 95th percentile. The analysis run after the restrictions had been imposed yielded the following results: the post-LASSO cumulative returns equalled 26.6088%, the post-Ridge ones were 26.2420%, while the EW portfolio's cumulative returns amounted to 6.9126%. Therefore, it can be inferred that both times the portfolios with the added constraints outperformed the equal-weighted portfolio. In terms of the Sharpe ratios, the outputs were as follows: for the LASSO regression 0.05395; for the Ridge regression 0.0534 while for the EW portfolio the ratio amounted to 0.0412. It may be hence inferred that the introduction of a weight constraint on the truncated dataset resulted in model predictions that significantly outperformed the equal-weighted portfolio.

Conclusion[5]

In conclusion, after the first run the resulting truncated model showcased promising cumulative returns, surpassing EW's performance over the training sample. Despite better returns, the truncated model showed more than proportional risk, highlighting the nuanced trade-off between returns and volatility. Then, through additional constraint imposed on the weights of the assets included in the model, the overall variance of the portfolio was brought down. Through the changes introduced in data and constraints the issue of overfitting present in the original model was alleviated, and the overall performance was significantly improved and managed to outperform the Equally Weighted Portfolio on all the measures.



This analysis underscored the intricacies of model refinement in the dynamic landscape of financial forecasting.

[5] Lasso/Ridge 50 C stand for a Lasso/Ridge regression on the truncated data with a constraint limiting the weights of the assets.