Hackathon

Anastasia Demina, Elina Harutyunyan
15.457 - Advanced Analytics of Finance
ademina@mit.edu, elinahar@mit.edu

May 6, 2020

1 Project Description

The investment objective is the following:

$$U_t = \min_{w_t} E_t[r_{p,t+1}] - \gamma Var_t[r_{p,t+1}]$$

where $E_t(r_{p,t+1}) = w_t' E_t(r_{i,t+1})$ and $Var_t(r_{p,t+1}) = w_t' cov_t(r_{t+1}) w_t$.

Maximizing the objective we get the optimal weights expression:

$$w^*_{i,t} = \frac{E_t(r_{i,t+1}) cov_t(r_{t+1})^{-1}}{\mathbb{1}' E_t(r_{i,t+1}) cov_t(r_{t+1})^{-1}}$$

$E_t(r_{i,t+1})$ is the predicted return for each stock for the next month produced by Ridge regression with λ=0.316.
$cov_t(r_{t+1})$ is estimated using the following formula: $cov_t(r_{t+1}) = \beta_i cov_t(f_{t+1}) \beta_i' + \Omega$.

To calculate predicted returns $E_t(r_{i,t+1})$ for each stock, we need to model the factors in order to calculate $E_t(f_{i,t+1})$, the predicted value of each factor for month t. We looked at the autocorrelation functions of factors and found that none of them exhibit strong autocorrelation. We decided to model each factor using constant mean model:$R_{i,t} = \mu_i + \epsilon_{i,t}$

We use the optimal weights found above to construct monthly portfolio (weight at time t is applied to the returns at $t+1$). Every month we calculate portfolio returns and store them. We rebalance the portfolio every month.

2 Feature Selection

We use 3 different ways to select features to include in our model: regularized regressions (Ridge and Lasso), Random Forest and Recursive Feature Elimination. First, we divide our sample into 2 parts: training (70%) and validation(30%). After that we run models on our training set and perform feature selection.

2.1 Ridge and Lasso Regressions

Ridge and Lasso are regularized regression models that help to prevent over-fitting out-of-sample by penalizing the magnitude of the coefficients.We perform time-series cross-validation to choose the "best" lambda parameter. We do it in the following way: 1. for each stock for each lambda perform cross validation and calculate the average MSE for each lambda; 2. find the mean of the average error for each lambda across stocks; 3. choose lambda that produces the lowest average error across all stocks.
The best parameters are: $\lambda_{Lasso} = 0.316$ and $\lambda_{Ridge} = 0.008$
After that we run Ridge and Lasso regressions on all 500 stocks separately using "best" lambdas. After that we find mean and median of coefficients of each of the factors across stocks. Based on Ridge and Lasso regressions, we find that the following factors are important for prediction of the stock returns: industries (Money, Manufacturing, Utilities, Shops, Business Equipment, Other) and market variables (Market-Rf, SMB, HML, MOM, RMW).

2.2 Random Forest

We run a random forest on the training set on each of the stocks with the full set of factors. We obtain the relative importance of each factor for each stock, after which we calculate the average or median values for each factor across stocks. We obtain that according to the average value the factors that we most important were Utilities, Money, Manufacturing and Energy. Using median we obtain that SMB, Chemicals, Utilities, Manufacturing, Money and Momentum are the top important factors.

2.3 Recursive Feature Elimination

Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. We perform time series cross validation for RFE in the same manner as the one for Ridge and Lasso. The results showed that for each stock the optimal number of factors to use is 1. We then proceed to run the RFE model with 1 factor on each stock and obtain which factor is chosen. We conclude with a list of all the factors and number of times it was chosen by the RFE model. The factors that occurred the most number of times were Manufacturing, Money, Market, Utilities, Shops and Other.
We combine all of the results obtained from the models and choose these factors to include in our analysis: Market, SMB ,HML, MOM, Money, Manufacturing, Utilities, Shops and Other. Even though in some cases the Fama French factors were not included in the top, we still believe that they should be included in the analysis. As a robustness check we also included Durables, Energy, Robust minus weak and Business Equipment, however they did not impact the results, hence we exclude them from our final model.

3 Model Selection

- We fit a Ridge regression model with $\lambda = 0.316$ and obtain coefficients for each stock as well as the residuals from the models for both training and testing periods.

- We forecast the expected factor returns for the test set using the constant mean model, whereby we calculate the mean of the training set of the factor up to each month and use that value as the prediction for the next month. The first value of the test set is the mean of the training set.

- The expected returns of each stocks for each month in the test set are obtained using the coefficients and the forecasets of expected factor returns.

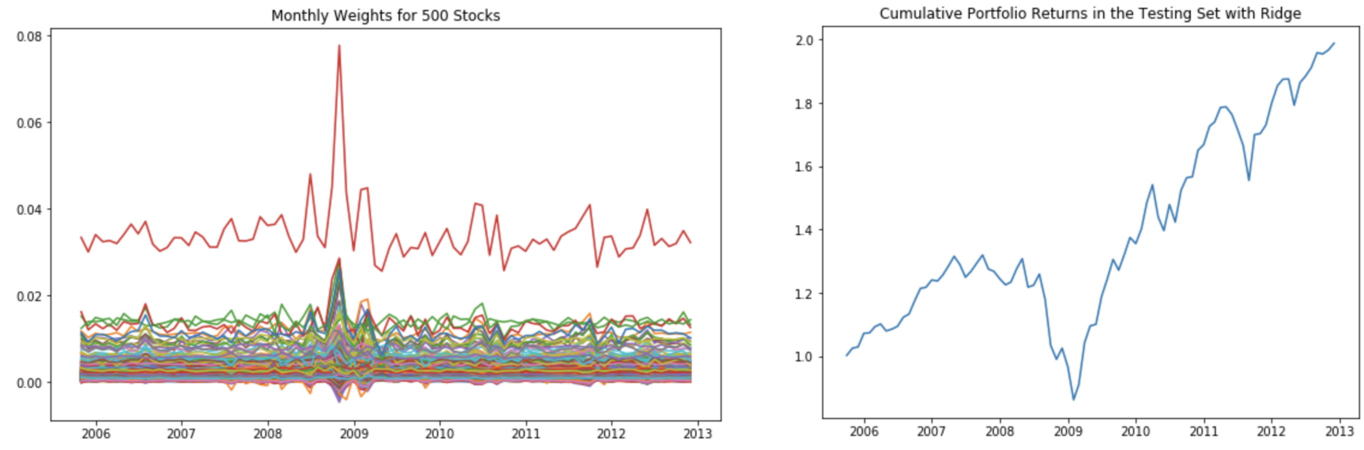
- For the covariance matrix of factors, we use the same logic, where we calculate the covariance of factors up till a given month and use that as prediction for the next month. Afterwards we obtain the term $\beta_i cov_t(f_{t+1})\beta_i'$ for each month of the testing set.

- The prediction for variance of residuals are calculated by taking the variance of residuals up to month t, which is then summed to the covariance term calculated above to get the forecast of the covariance of the stock return for a given month.

- We obtain weights by dividing the forecasted returns by the forecasted variance of returns for each stock for each month and normalize by dividing it by the sum of weights for that given month.

- We apply the weights obtained in a given month to the returns of the next month in the test set and calculate portfolio returns.

- Since, everything was done in terms of excess returns, we add back the risk free rate for the period and calculate the cumulative returns. We report the graph of cumulative returns, monhtly weights of stocks and the performance indicator $U_p = \bar{r}_p - \gamma\bar{\sigma}_p^2 = 0.000639$.



(a)

(b)

Figure 1: Weights and Cumulative Returns