Assignment 1

Machine Learning Applied To Financial Data

MATH 60610A.A2023

`

BY

Prateek Sinha (11306859)

Hair Albeiro Parra Barrera (11315672)

Kriti Bhaya (11321276)

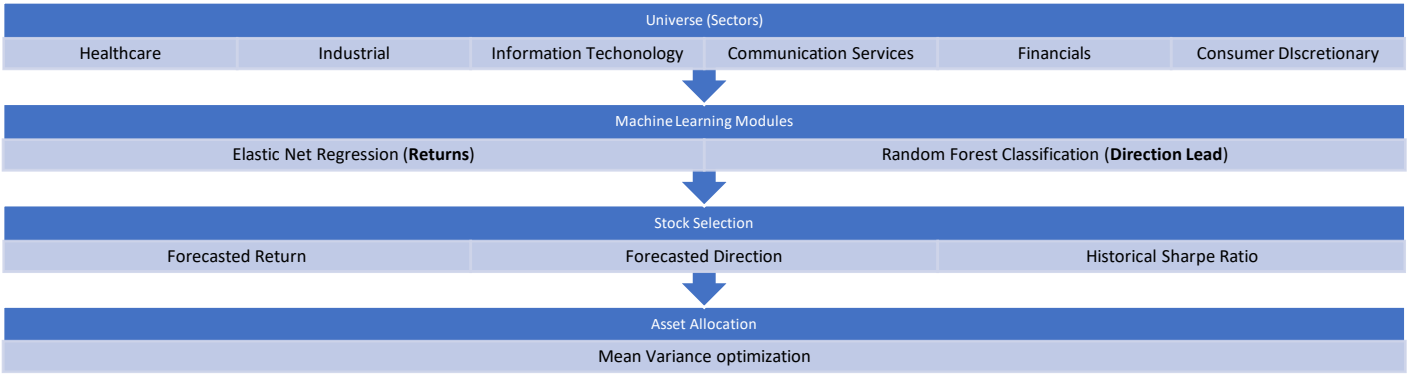
Xiao Xue (11308678)

Submitted to

Professor David Ardia

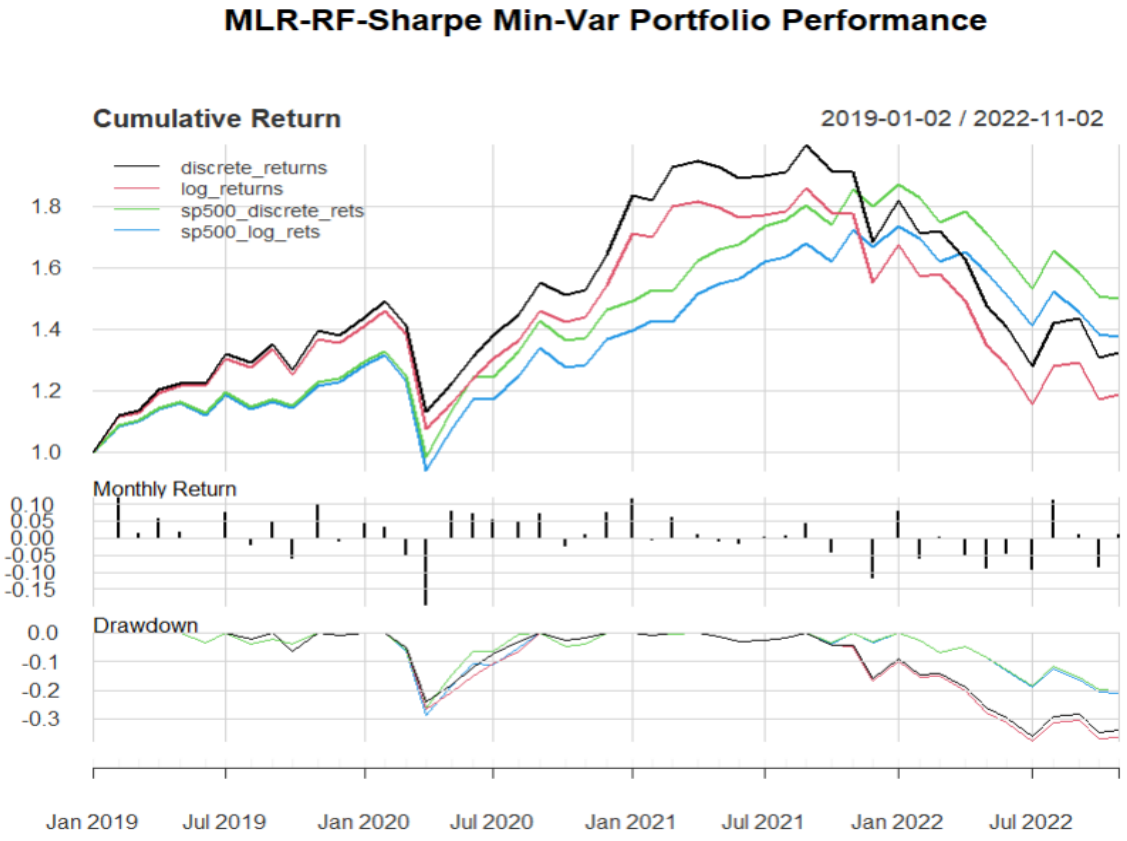
TRADING STRATEGY

We use an automatic S&P 500 sub-universe selection algorithm based on economic sectors and stocks with largest market capitalisation. At the beginning of every month, we subset a window of "recent" past data (e.g., two years), and for each combination of sector and corresponding stocks, we perform backward linear feature selection, followed by Elastic Net Regression to anticipated returns, and a Random Forest to anticipate their trajectory (up or down). We filter assets emphasizing on anticipated returns, direction and the Sharpe ratio. With the help of a modified minimum-variance optimization framework based on a different set of time-series-based forecasted returns and volatility, we rebalance the portfolio by longing based on the optimized weights. We then hold until the beginning of next month, where we close all positions, and perform the strategy again.



PERFORMANCE (JAN 2019 – NOV 2022)

The backtest spans over a 5-year period, using a rolling 24-month window to forecast the metrics for the subsequent month. Our data, however, consists of weekly Wednesday adjusted close. We benchmarked the performance against the S&P 500 index. As evident from the graph below, the portfolio has consistently outperformed the index on a monthly basis. The portfolio, however, also displays significant drawdowns indicating high volatility and large declines from peaks. Towards the latter part of the backtest, portfolio's performance appears to align closely with that of the S&P 500 before eventually intersecting with it. This pattern suggests an asymmetric response of the portfolio to market dynamics. Specifically, the portfolio doesn't demonstrate equal sensitivity to market uptrends as it does to downtrends.



Notes:

- Sharpe ratio (2019-2020): **0.72**
- Pandemic at the beginning of 2020.

MACHINE LEARNING STRATEGIES

As mentioned in the Trading Strategy section, every month, we leverage a fixed window of past data, in this case 2 years, or 24 months. For each of these, we extract the features for that range, but we also compute a couple of "dynamic features", given by the following:

1. **Shifted SARIMA(p,d,q) features** for different combinations of *p,d,q.* The goal is to forecast the realized returns 4-weeks ahead, so that these are introduced to the historical data as predictive features. The SARIMA (Seasonal Autoregressive Integrated Moving Average) is a statistical time-series model that combines the AR (autoregressive on the target) models, the MA (moving average of past noise), differencing and seasonality modelling to approximate the behaviour of the stochastic process in question. The mathematical expression for this model is complex in nature, and therefore omitted here for brevity.
2. **Shifted GARCH(1,1) features.** This feature reflects a 4-weeks ahead volatility forecast for the asset. Generalized Autoregressive Conditional Heteroskedasticity (GARCH) with specification (1,1) can be described as follows:

$$\begin{cases} y_t = \mu + \epsilon_t & \text{, (observed data as function of mean and error term)} \\ \sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 & \text{, (recursive model for the conditional variance)} \\ \epsilon_t = \sigma_t z_t & \text{, (innovations term)} \\ z_t \sim N(0,1) & \text{, (white-noise term)} \end{cases}$$

These features have two goals: (i) to create informative predictive features used in the regression and classification models and (ii) function as the future-shifted mean and volatility return vectors used in the min-variance framework, as opposed to the raw historical data.

Backward Stepwise Regression for Feature Selection

Navigating the multifaceted domain of stock prediction mandates the meticulous selection of features to judiciously mitigate risks associated with overfitting and enhance overall model performance. Within this context, we strategically employed Backward Stepwise Regression, wherein a carefully curated subset of maximum 35 features was randomly selected, subsequently undergoing the rigor of backward stepwise regression. This approach was adopted to ascertain that only variables of substantial predictive merit were retained, thereby optimizing the model's capability to generate precise and reliable forecasts while maintaining a streamlined and computationally efficient feature set.

Model Training with Elastic Net Regression

Subsequent to feature selection, an Elastic Net regression model was implemented, which seamlessly blends the penalties from both Lasso and Ridge regression, succinctly expressed by the following objective function:

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} l(y_i, \beta_0 + \beta' x_i) + \lambda \left[\frac{1}{2} (1-\alpha) \parallel \beta \parallel_2^2 + \alpha \parallel \beta \parallel_1 \right]$$

Where $l(y,\eta)$ is the negative log-likelihood over contribution for observation i. α controls elastic-net penalty and λ controls the overall strength of the penalty.

The model was meticulously trained using 10-fold cross-validation, ensuring the derivation of optimal hyperparameters $\alpha^*$and λ^* whilst safeguarding against overfitting.

Random Forest Model for Classification

Mathematically, a Random Forest can be described as a collection of decision trees $T(x, \Theta_k)$parameters for the k -th tree. These trees are trained independently on different bootstraps samples of the data (bagging), and their predictions are aggregated to

produce a final prediction. This modified bagging methods reduces variance in the predictions, making it a more robust prediction algorithm.

In the case of classification, with C classes, the prediction of a Random Forest Model $F(x)$ for input features x is given by:

$$\mathcal{F}(x) = \arg\max_{c \in C} \sum_{k=1}^{K} I(T(x; \Theta_k) = c)$$

That is, this a majority vote of the predictions for each class, over the random forest parameter space.

In our case, $C = 2$, since we are predicting the direction of the historical returns for this period: i.e., "up" or "down".

Ensemble Model Prediction

To choose the best performing stocks, we combine these methods along with the historical Sharpe ratio using the following heuristic:

$$\text{Choose stock } x \text{ if } \begin{cases} \texttt{forecasted\_return > 0 \& forecasted\_direction == "up"} \\ \texttt{OR} \\ \texttt{sharpe\_ratio > 0.3 \& forecasted\_direction == "up"} \\ \texttt{OR} \\ \texttt{sharpe\_ratio > 0.5} \end{cases}$$

The logic behind this heuristic is simply that we want to have stocks with high confidence in the positive returns and in increase, without neglecting historical performance based on the Sharpe ratio, which might also reflect in the future.

STRATEGY: PROS & CONS

1. **Investment horizon and risk profile**: The portfolio has demonstrated positive returns over a short-term horizon (2 years), although with periods of significant volatility. Therefore, this strategy is apt for investors seeking to generate income within a shorter timeframe (as the portfolio does outperform the S&P 500 index) and who can tolerate occasional downturns, indicating a lower risk aversion. Conversely, investors with a more conservative risk appetite and a longer investment horizon might not find this strategy suitable.

2. **Modelling:** The portfolio strategy employs multiple regression and classification techniques to ensure robustness. Specifically, random forest is a fairly good model which can handle complex, non-linear relationships to provide accurate classification predictions. However, though robust, using several regression & classifications-based signals can lead to complexity. Exploring other combinations of more streamlined regression and classification models might yield even more effective outcomes. Further, incorporating additional data features could further refine prediction accuracy.

3. **Assumptions:** By eliminating short sales, we avoid potential losses that might occur in case stock prices rise substantially. This also simplifies the portfolio's management, making it more direct and user-friendly. However, the approach operates on the premise that the returns and momentum are persistent, implying that historical trends will reappear. A 2-year window may not adequately represent long-term market volatility trends. Also, if markets are in a mean-reverting stage, a momentum-based strategy can lead to potential losses. The assumption of zero transaction costs is not reflective of real-world scenarios, which should be accounted while designing an actionable stock market strategy.