Trading Strategy - Financial Data Science

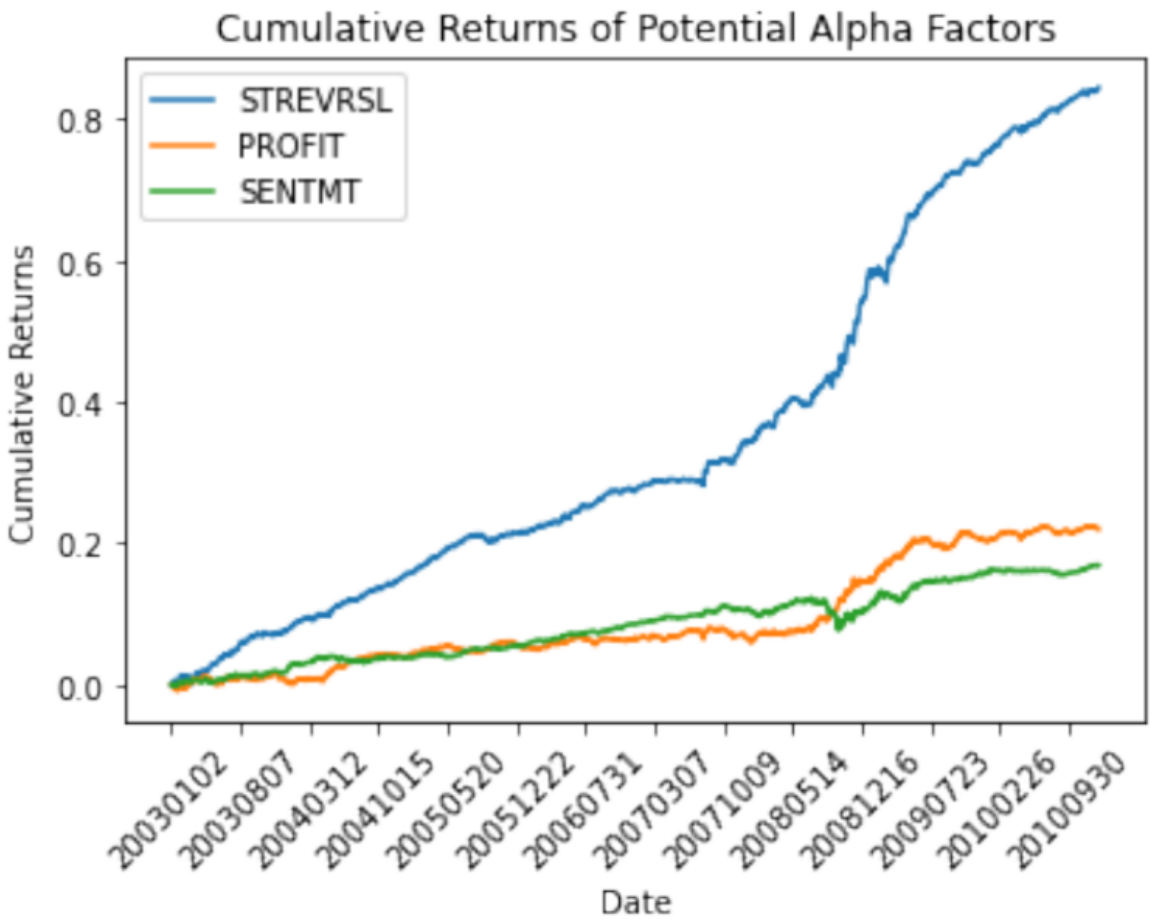
Executive Summary

The primary objective of this project is to apply machine learning techniques for factor analysis, identifying the most influential alpha factors for portfolio optimization. This optimization focuses on stocks with a market capitalization greater than 1 billion, constituting the investable universe. Employing Markowitz's Modern Portfolio Theory for mean-variance optimization, the process is further refined using the Woodbury Inversion Lemma. The crux of the project lies in leveraging machine learning to ascertain key determinants in computing optimal asset weights.

Steps for Data Preparation and Data Split

In preparation for data cleansing and preparing for implementing the machine learning models, we conducted the steps below:

1. Exploratory Data Analysis (EDA) : we conducted EDA on the datasets to look into different alpha factors within the dataset that can potentially indicate returns. Base from the EDA, after analyzing the winsorized cumulative return, STREVRSL has the highest potential to become the alpha factor, given the cumulative return result.



2. Combining dataset: We combined the time series data prior to the pseudo-inverse and winsorized data prior to the splitting of the dataset.
3. Pseudo-inverse and winsorized data: We winsorized the data by removing the extreme values in the dataset to reduce the impact of outliers. The purpose of the pseudo-inverse and winsorizing data is to augment the existing data frames with additional computed information to perform further data analyses.
4. Data Split: Training (70%) and Testing (30%) by dates and join together vertically the frames in the training/validation set *Dtrain* into a single frame called a panel.

Machine Learning Methods

In order to find the model of the form $Y = f(\text{candidate alphas}) + \epsilon$ to best predict the return by determining candidate alphas, we conducted 3 methods to predict the candidate alphas: two linear regressions, Lasso and Elastic Net, and one non-linear, Support Vector Regression (SVR).

Linear

1. *Lasso*

We first utilized Lasso Regression to conduct our linear regression cross-validation technique. Lasso utilizes shrinkage to prevent overfitting of the model. Using the training and testing data split as performed in the previous steps, we perform cross-validation to evaluate the Root Mean Square Error (RMSE) for accuracy. The number of partitions is represented by "k"; in this case, we use 5-fold cross-validation (cv=5), so the data is split into 5 folds. Four of these folds are used to train the model, and the remaining fold is used to validate the data. The performance of the model is then averaged over the five folds to give a more robust measurement of its effectiveness.

Lasso then predicts the alpha coefficients in the dataset. To choose the best alphas coefficients, we utilized the 'LassoCV()' function from the sklearn library. We set the n_alphas=num_alphas parameter in LassoCV() to 1000, asking the model to consider 1000 equally spaced alpha values within its range and pick the one that results in the best cross-validation score.
Finally, the script prints out the optimal alpha value found, the R-squared score, and the number of input features that were kept or eliminated by the model. Features with non-zero coefficients are those that the model found useful for predicting the target variable, whereas features with zero coefficients have been 'eliminated' as not useful or redundant.

The result of Lasso indicates the alpha factors are 'STREVRSL', 'INDMOM', 'SENTMT', 'SEASON', 'EARNYILD','PROFIT', and 'MGMTQLTY'. The RMSE Score is 0.017909.

2. *Elastic net cross-validation(Sklearns)*

Elastic Net CV was utilized to see if there is a difference between Lasso. Although they are both linear models, Elastic net could be seen as an improvement from Lasso and Ridge regression model through linearly combining L1/L2 penalties to overcome the limitation of both models. Elastic Net addresses the issues of high correlation between predictors. Although the only main downside for elastic net is that it is computationally expensive. After running cross validation on the elastic net model, the best L1 ratio that we received was 0.01 while the best alpha was 0.0011.

The result of the Elastic net indicates 'STREVRSL', "INDMOM', 'SENTMT', 'SEASON', 'EARNYILD', 'PROFIT', 'MGMTQLTY' as the alpha factors.The RMSE Score is 0.0179. The candidate alpha's for Elastic Net and Lasso are identical in this instance.

<u>Non-Linear</u>

3. *Support Vector Regression*

We utilized Support Vector Regression (SVR) from sklearn's GridSearch CV as our non-linear function and cross-validation to optimize the hyper parameters. We then use the SHAP library to interpret the model's predictions. The param_grid dictionary includes the parameters kernel and c for tuning the model, with the parameter grid and crossvalidation folds set to 5.

After fitting the GridSearchCV with the time series pricing data, it conducted a search with the SVR model that satisfies the combination of the kernel and c parameters to identify the best accuracy of the alphas using cross-validation.

The SHAP result of the SVR indicates ‘INDMOM’, ‘EARNYILD’, ‘PROFIT’,’MGMTQLTY’, ‘STREVRSL’, ‘LTREVRSL’,’SEASON’, ‘EARNQLTY’, and ‘SENTMT’ as the alpha factors, with a RMSE score of 0.02079.

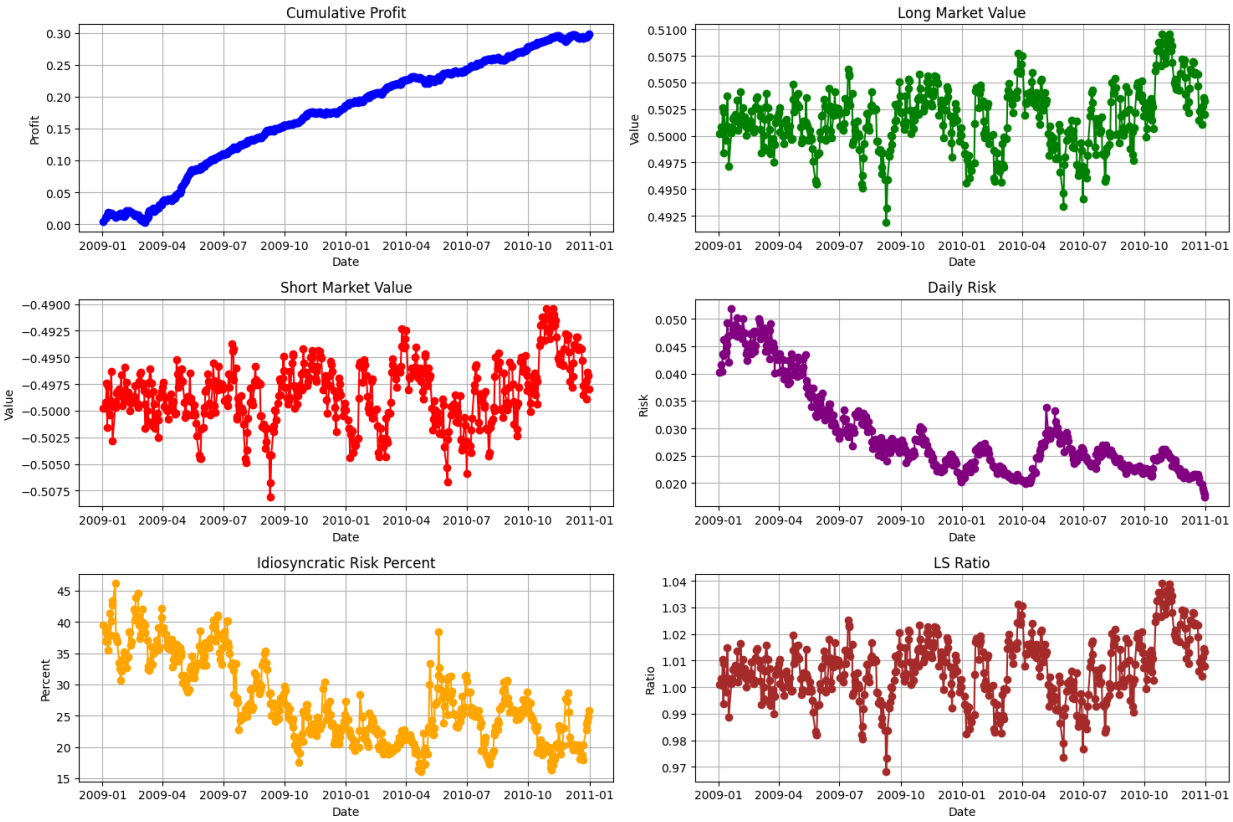
Results

Comparing the results of the three models as discussed above, we note that the Lasso Regression and Elastic Net Regression have the lowest and the same RMSE score of 0.0179, hence we adapted the alpha factors indicated in the two models, ‘STREVRSL’, ‘INDMOM’, ‘SENTMT’, ‘SEASON’, ‘EARNYILD’, ‘PROFIT’, ‘MGMTQLTY’ , for the woodbury matrix optimization to determine the weights of the portfolio, and the time series plotting which shows the long and short market values, and cumulative profit of the portfolio.

Please see below the details of the selected alphas:

STREVRSL: short-term reversal.
INDMOM: industry momentum (defined as relative historical outperformance or underperformance of the other stocks in the same industry).
SENTMT: news sentiment.
SEASON: seasonality-based alpha factor.
EARNYILD: earnings yield.
PROFIT: profitability.
MGMTQLTY: management quality, looks at quantitative measures of how well-run a company is by its management.

Please see below the time series plot showing the long and short market values



- **Cumulative Profit:**
  - Upward trend; The return base from the suggested variable weight from Lasso suggested 30% cumulative return.
- **Long Market Value:**
  - Stable fluctuations; The value seems to fluctuate around a central value without a clear long-term trend, which suggests that the value of assets held has been relatively stable.
- **Short Market Value:**
  - Stable negative values; The values for the short position remain constant around -0.5, showing that the short positions have also been relatively stable in value.
- **Daily Risk:**
  - Decreasing trend; Initially, the risk is higher and then gradually decreases. This could mean that the portfolio is becoming less volatile or that risk management strategies are effectively reducing risk.
- **Idiosyncratic Risk Percent:**
  - Variable, with no clear trend; The percentages fluctuate with no clear trend, implying that the idiosyncratic risk varies but does not show a systematic increase or decrease. The reason for this maybe due to lack of observed data.
- **LS Ratio:**
  - The ratio fluctuates around 1. This suggests that the portfolio maintains a balance between the value of its long and short positions.

Key Takeaway and Limitation:

On determining the most effective model for our alpha factors, the two models that stood out from the rest are Elastic Net and Lasso. These two struck the right balance and consistently came through with the lowest RMSE scores, signaling that they were on point in predicting outcomes without getting bogged down by complexity.

On the other hand, the SVR model did not have as good of the performance. The reason for the lower performance can be due to lack of a bit more fine-tuning, and perhaps not having enough data to effectively train the model. Moreover, SVR is a very complex model, so the two aforementioned reasons are the important factors that could impact the poor performance of the model.

Subsequent to model selection, portfolio optimization techniques were employed to ascertain the optimal weights for the portfolio. Through the utilization of time series plots, we gained insights into the stock performance over time, accounting for various risk factors.

However, it is important to note the limitations encountered during the testing phase. The complexity of more advanced models, coupled with high data volume, led to instances of kernel crashes, which slowed down the process. To mitigate this, we resorted to reducing the size of the data, which, while enabling us to complete the exercise, also resulted in a compromise on the quality of the optimization. Specifically, this reduction in data size may have led to a less accurate representation of idiosyncratic risk within the optimization process.