Big Data in Finance 1 - Assignment

MSc Financial Technology: Group 10

January 2021

Prediction: Regression

SMA tends to outperform rolling regressions. RF better with PCA & additional features

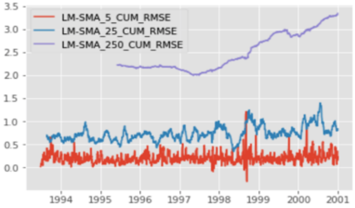

Figure 1: Cumulative RMSE differential between rolling linear regression and historical mean return models for w=5, 25, 250

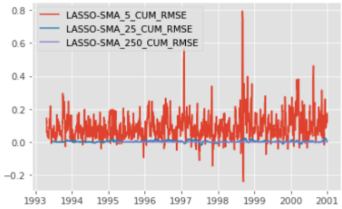

Figure 2: Cumulative RMSE differential between rolling LASSO with λ=0.5 and historical mean return models for w=5, 25, 250

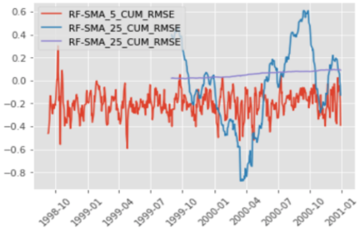

Figure 3: Cumulative RMSE differential between the Random Forest Regressor (70:30 split) and historical mean return models for w=5, 25, 250

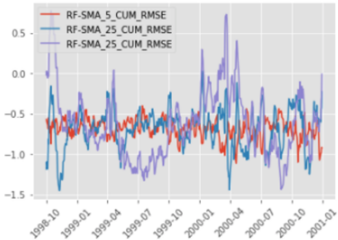

Figure 4: Cumulative RMSE differential between the Random Forest Regressor (PCA + Feature space) and historical mean return models for w=5, 25, 250

Prediction: Regression - First Thoughts

Propsed methods were slow and did not have much prediction power

- **No long-term linear auto-correlation in financial returns**. Already known → Efficient Market Hypothesis /Random Walk Theory
- ML models tend to **perform generally poorly** solely on return data (also cross-sections), see Chinco, Clark-Joseph, and Ye → $R^{2*} = 0.08$
- Hyperparameter optimization is **computationally infeasible** with rolling windows, i.e. $(1958 - w) * 100$ CV cycles . Even with dimensionality reduction (PCA) and adjustable step-sizes.
- Financial returns are **very** noisy!
- In financial settings minimising mean squared error is not always a proxy for returns, optimisation methods need to be geared towards more meaningful objectives

Classification & Performance: Our Take

Instead of predicting returns themselves, we predict market states and optimal action

- Rather than directly predicting future prices or returns, we classify the **state** of a **market** based on a **feature space**, e.g. moving averages. Follows a recent trend in AM & was popularized by de Prado (2018)
- Market states can be buy, hold & sell.These are seperated through the treshhold labels τ:E.g. if $\tau > |r_{i,i+1}|$: $y_i = (\frac{r_{i,i+1}}{\tau})^3$ or 0, else $y_i = sgn(r_{i,i+1})$
- We utilised $SMA_5(r_{i,i+5})$ instead of $r_{i,i+1}$ to identify market state. This is to smooth the prediction and to reduce the noise in financial data and to properly capture the market state for a holding period
- Huge potential: τ could be treated as a hyperparameter and optmised for every stock, be a stock selection criterion or both
- Implementation:github.com/Parhamallboje/BigDatainFinance1

Label Prediction Rationale

Balancing the right threshold parameter to optimal action not necessarily reducing RMSE

- A SVM Classifier can learn a set of market states, if τ per stock was chosen somewhat optimally (Figure 5). Even though we the learner tends to hold in buy states, when the learner buys, it buys correctly. We could call this a **careful** learner. We do not lose money this way.
- If τ is too low, the learner will overestimate buy/sell market states and act too rash. We could call this a **noisy** learner (Figure 6). Anything else would be just a line at 0.

Figure 5: Actual vs predicted market states (τ = 0.01) using SVC (γ=auto) for the Walmart stock with technical indicators, no PCA

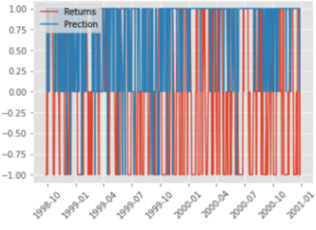

Figure 6: Actual vs predicted market states (τ = 0.003) using SVC (γ=auto) for the ALD stock with technical indicators, no PCA

Back-test

Careful learner with careful return dynamics with no transaction costs and infinite cash

- With a preset threshold parameter of 0.012 (stocks that did not get any market state label were dropped for training/prediction) the strategy outperforms mean returns of all stocks.
- Note: Profit from the Strategy is computed by multiplying signal with $SMA_5(r_{i,i+5})$, which is the minimum return case. These returns in theory can be achieved by limit order for $price * \tau$.

Conclusion

Market State Classification is the way to move forward and holds a lot of potential

- Key Findings & Contributions:
  - Only the RandomForest outperformed SMA during this time period
  - Threshold parameter (τ) optimisation offers a different valuable point of view for investment decisions and a new parameter that can be optimized.
- Limitations:
  - Apart from technical indicators, no we do not have additional data
  - Provided data set is from 1993-2000, which is quite the limited time frame. Today, there is much more noise given more market participants enabled through technological advances (algorithms & apps)
- Future Work:
  - Optimization algorithm for τ towards careful learner for all stocks
  - Include additional data sources, i.e. company or alternative data
  - Start backtesting models & strategy for other periods, e.g. 2001-2020

References

- Alex Chinco, Adam D. Clark-Joseph, and Mao Ye, 2018, **Sparse Signals in the Cross-Section of Returns**, Journal of Finance, here
- Marcos Lopez de Prado, 2018, **Advances in Financial Machine Learning**, Wiley Publishing
- Michal Balcerak and Thomas Schmelzer, 2020, **Constructing trading strategy ensembles by classifying market states**, arXiv, here

- SMA - Simple Moving Average, VOL - Simple Moving Volatility, UBB - Upper Bollinger Band, LBB - Lower Bollinger Band, s - cross returns
- Linear Regression (Figure 1)

$$\begin{vmatrix} R_{t-1} & \dots & R_{t-w} & Flow_{t-1} & \dots & Flow_{t-w} \end{vmatrix}$$

- ML with no features (Figure 2 & 3)

$$\begin{vmatrix} R_{t-1} & \dots & R_{t-w} & Flow_{t-1} & \dots & Flow_{t-w} \\ R_{s1t-1} & .. & R_{s99t-1} & Flow_{s1t-1} & \dots & Flow_{s99t-1} \end{vmatrix}$$

- ML with features (Figure 4 (no PCA case) & Backtest)

$$\begin{vmatrix} R_{t-1} & \dots & R_{t-w} & Flow_{t-1} & \dots & Flow_{t-w} \\ R_{s1t-1} & .. & R_{s99t-1} & Flow_{s1t-1} & \dots & Flow_{s99t-1} \\ SMA_5(r_t) & .. & SMA_{250}(r_t) & UBB_5(r_t) & .. & UBB_{250}(r_t) \\ VOL_5(r_t) & .. & VOL_{250}(r_t) & LBB_5(r_t) & .. & LBB_{250}(r_t) \end{vmatrix}$$