Leonard De Vinci Graduate School Of Engineering

Machine Learning Report

Predicting end-of-session returns for the US equity market

Students :
Enzo Delgado
Raphael Demare
Mohammed Drici

Contents

1 Introduction

The US equity market is one of the most liquid in the world, providing numerous investment opportunities. The last two hours of the session, between 2pm and 4pm, are the most liquid of all.
This is the preferred time to buy or sell a large quantity of assets, as market conditions are better (lower transaction costs, lower volatility, etc.). This is why estimating an asset's behavior over this period in advance helps you optimize your entire portfolio.

We've chosen this topic because it has a real application for the various players in the sector. In fact, this subject is proposed as part of a data challenge co-organized by the ENS and CFM (Capital Fund Management: one of France's biggest hedge funds).

The aim of this challenge is to estimate the direction of a stock's price during the last two hours of trading, knowing how it behaved at the start of the day.
To avoid the usual noise of financial movements, we consider only three directions:

- an outright drop in price
- a minor change in either direction
- a clear price rise

Figure 1: Capital Fund Managment Logo

Figure 2: ENS Logo

2 Data Description and Analysis

2.1 Description

We have price trends (yields) with a granularity of 5 minutes, resulting in 53 values over these 4.5 hours, for each day and share. As these yields are low, we express them in basis points (bps), i.e. $\frac{P_{t+5minutes}-P_t}{P_t} * 10^4$.

We then have the columns :

- 'ID': the unique identifier of this line
- 'day': the identifier of the day concerned (not unique in the dataset)
- 'equity': the identifier of the share concerned (not unique in the dataset)
- 'r0': $\frac{P_{09:35}-P_{09:30}}{P_{09:30}} * 10^4$, the yield over the first 5 minutes
- 'r1': $\frac{P_{09:40}-P_{09:35}}{P_{09:35}} * 10^4$, the yield over the next 5 minutes
- $\vdots$
- 'r52': $\frac{P_{14:00}-P_{13:55}}{P_{13:55}} * 10^4$, the last yield provided

To reduce the difficulty of the task, the prediction is limited to estimating the direction of output over the last two hours, grouped into 3 cases:

- -1 if $\frac{P_{16:00}-P_{14:00}}{P_{14:00}} * 10^4 < -25bps$
- 0 if $-25bps < \frac{P_{16:00}-P_{14:00}}{P_{14:00}} * 10^4 < -25bps$
- $+1$ if $\frac{P_{16:00}-P_{14:00}}{P_{14:00}} * 10^4 > 25bps$

The prediction file includes the columns:

- 'ID': the unique identifier of the predicted line
- 'reod': the estimated end-of-session yield class, included in $[-1, 0, 1]$ according to the above formula.

It should be noted that, to avoid any leakage of information, the training and test datasets have no shares or days in common. However, the stocks in both datasets have exactly the same financial characteristics, so that their behavior can be predicted. Nevertheless, the two periods concerned are totally different, so as to correspond to real prediction work, on real data from a world whose characteristics may change.

2.2 Data Analysis

Now that we've explained our data formats, we can get down to the nitty-gritty. The first step is to clean up our data to make sure that no parasites will interfere with our predictions. In our case, we're pretty lucky - the datasets are pretty clean. All we needed to do was remove the rows containing NaN and rename a column.

The analysis of the U.S. equity market returns confirmed that they do not adhere to a Gaussian distribution. This observation aligns with common characteristics of financial datasets, which often display heavy tails and skewness, thereby deviating from the assumptions of normality. The implications of a non-Gaussian distribution are significant as they can lead to the presence of outliers and extreme values, which, in turn, may substantially affect both the model's performance and its predictive accuracy.

Additionally, the dataset exhibited characteristics of non-independence, a trait commonly seen in financial time series due to factors such as market trends, volatility clustering, and macroeconomic influences. The assumption of independence, a cornerstone in many statistical models, does not hold in such cases, necessitating careful consideration in the selection and validation of predictive models.

A further layer of complexity is introduced by the Hurst exponent values observed in the dataset. Typically, a Hurst exponent of 0.5 indicates a random walk, consistent with a Gaussian distribution. However, our analysis revealed Hurst exponents ranging from 0.41 to 0.45, deviating from the expected 0.5 value. This deviation confirms the non-random behavior of the time series and suggests a propensity for mean reversion. Specifically, the negative bias in the Hurst exponent implies a tendency for the time series to invert its trajectory in the future, hinting at anti-persistent behavior. This insight provides a pivotal consideration for predictive modeling, as it suggests an inherent reversal tendency in price movements that must be factored into any forecasting or trading strategies.

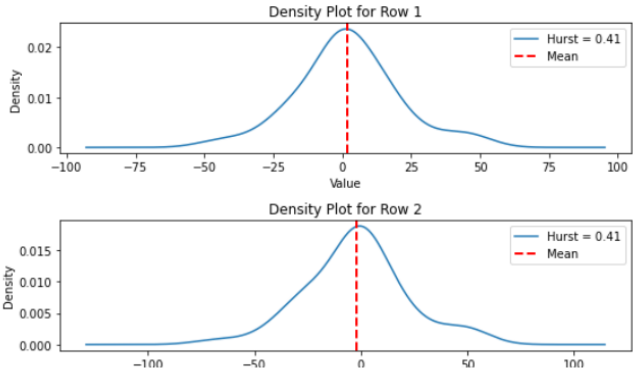

Figure 3: Density plot for a few rows and Hurst exponent association.

3 Random Forest

3.1 Configuration and Training

A RandomForestClassifier was implemented with the following parameters: n_estimators set to 50 and max_depth set to 10. The model was trained on the training dataset using these parameters. The trained model was then saved for future use.

3.2 Model Evaluation

The model's performance was evaluated on the test dataset. Key metrics used for evaluation included a classification report and a confusion matrix. The classification report provided insights into the precision, recall, and F1-score of the model, while the confusion matrix offered a detailed view of the model's prediction accuracy.

3.3 Classification Report

The classification report provides detailed insights into the model's performance for each class:

- **Precision**: This metric indicates the proportion of correct predictions among all predictions made for a given class. The precisions for classes -1, 0, and 1 are 0.42, 0.44, and 0.47, respectively.

- **Recall**: It measures the proportion of actual positive cases correctly identified by the model. For classes -1, 0, and 1, the recalls are 0.37, 0.74, and 0.17, respectively.

- **F1 Score**: This is the harmonic mean of precision and recall, offering a balance between these two metrics. The F1 scores for classes -1, 0, and 1 are 0.39, 0.56, and 0.24.

- **Support**: This represents the number of actual cases for each class in the test dataset. The supports for classes -1, 0, and 1 are 39353, 43294, and 37541, respectively.

- **Accuracy**: The overall accuracy of the model is 0.44.

- **Macro and Weighted Averages**: These averages provide an overview of the model's performance across all classes.

These metrics offer a comprehensive perspective on the model's ability to correctly classify different classes, as well as the errors it makes.

3.4 Rationale for Using Random Forest

Random Forest was chosen due to its robustness in handling large datasets and its ability to manage overfitting. It is particularly effective in scenarios where the relationship between features and the target variable is complex and non-linear. The ensemble nature of Random Forest, which combines multiple decision trees, contributes to its high accuracy and generalization ability, making it well-suited for predicting stock price movements.

4 XGBoost

4.1 Configuration and Training

An XGBoost classifier was configured with parameters tailored to the dataset's specific characteristics: n_estimators was set to 50, max_depth was chosen as 10, and the learning_rate was set to 0.1. Additionally, subsample and colsample_bytree were both set to 0.8 to ensure that each tree uses a subset of the data and features, reducing the risk of overfitting. The tree_method was set to 'hist', an optimized algorithm for XGBoost. The labels in the training set were transformed to start from 0 for compatibility with the model's requirements.

The model was then trained on the transformed dataset. After training, the model was saved as a .pkl file for later use or further analysis.

4.2 Model Evaluation

The XGBoost model's performance was evaluated on the test dataset. The classification report and confusion matrix were used to assess the model's predictive capabilities.

- **Precision**: Measures the proportion of correct positive identifications made by the model out of all positive identifications. Precisions for classes -1, 0, and 1 are 0.44, 0.47, and 0.45, respectively.

- **Recall**: Reflects the proportion of actual positives that were correctly identified. The model's recall for classes -1, 0, and 1 are 0.38, 0.67, and 0.29, respectively.

- **F1 Score**: The harmonic mean of precision and recall, representing the balance between them. The F1 scores for classes -1, 0, and 1 are 0.41, 0.55, and 0.35.

- **Support**: Indicates the number of true instances for each class. The support numbers for classes -1, 0, and 1 are 39353, 43294, and 37541, respectively.

- **Accuracy**: The overall accuracy of the XGBoost model is 0.46, demonstrating the percentage of total correct predictions.

- **Macro Avg**: The unweighted average of precision, recall, and F1 score across all classes, which is 0.45 for precision and recall, and 0.44 for the F1 score.

- **Weighted Avg**: The average of precision, recall, and F1 score across all classes weighted by the support, with values of 0.45 for precision, 0.46 for accuracy, and 0.44 for the F1 score.

4.3 Rationale for Using XGBoost

XGBoost was chosen for its efficiency and effectiveness in handling various types of data. Its ability to perform both linear and non-linear classifications makes it particularly versatile for complex datasets like those found in financial markets. XGBoost's use of gradient boosting offers a sequential learning process that improves accuracy and reduces errors iteratively. Moreover, its robust handling of missing values and regularization to prevent overfitting are essential for achieving reliable predictions in stock price movement forecasting.

5 Conclusion & Follow-up

5.1 Future Directions: Utilizing the Hurst Exponent

The Hurst exponent has provided us with valuable insights into the memory and autocorrelation properties of the equity market returns. Given that the exponent values suggest a potential for mean reversion, one could leverage this information to forecast future price movements. A logical extension of our current model could involve integrating the Hurst indicator to identify when a reversal in trend is likely to occur.

5.2 Volatility Prediction with GARCH or ARCH Models

Another promising direction could be to use volatility modeling techniques, such as Generalized Autoregressive Conditional Heteroskedasticity (GARCH) or Autoregressive Conditional Heteroskedasticity (ARCH), which are particularly well-suited for financial time series data. These models excel in capturing the 'clustering' effect in volatility, a common phenomenon where large changes in asset prices tend to be followed by further large changes, and periods of tranquility tend to persist.

- **GARCH and ARCH for Volatility Prediction**: By predicting volatility at time $t - 1$, we could obtain a more nuanced understanding of the risk associated with the asset for the next return. This prediction could then be used as an input to our return estimation models, potentially enhancing their predictive accuracy.
- **Implications for Return Estimation**: The expected volatility could serve as an important feature in our models, especially if we consider that periods of high volatility may precede reversals. This could significantly inform our trading strategies, particularly in a high-frequency setting where anticipating short-term movements is crucial.

By continuously refining our models with relevant indicators such as the Hurst exponent and incorporating volatility predictions, we can aspire to create a robust system that not only anticipates directional movements but also quantifies the associated risk, thereby offering a comprehensive solution for portfolio optimization and risk management.

6 Appendix

In the project files you will find:

- The notebook containing all our code
- The notebook exported in html
- The pdf report
- The various datasets
- The .pkl models