Project 2: Data Analysis and Machine Learning Report

Pranav Surampudi

GT ID: 903948185

Data Collection and Cleaning

In the process of collecting and preparing data for financial analysis and model training, I employed a comprehensive approach that adhered closely to the standard practices of data collection, cleaning, and preprocessing. The initial step involved gathering the necessary stock data, which was efficiently accomplished using the yfinance package. This package is particularly advantageous for financial data retrieval as it provides access to a vast repository of historical market data from Yahoo Finance, thereby ensuring the data's accuracy and reliability.

After acquiring the data, the next critical phase was data cleaning and normalization. For this, I utilized the Robust Scaler method. The rationale behind choosing Robust Scaler over other normalization techniques is its resilience against outliers, which are relatively common in financial datasets. Stock market data often contains extreme values due to market volatility, corporate events, or economic news. The Robust Scaler, by scaling data according to the interquartile range, minimizes the influence of these outliers, leading to a more robust model performance.

After gathering the data, the next crucial step involved data cleaning and normalization. For this purpose, I employed the Min-Max Scaler method from sklearn, with the range set to (0,1). The choice of Min-Max Scaler was guided by its effectiveness in transforming features by scaling each feature to a given range, here between 0 and 1. This scaling technique is particularly beneficial in contexts where the parameters need to be on a positive scale. Unlike other normalization methods, the Min-Max Scaler preserves the shape of the original distribution and doesn't reduce the importance of outliers. This characteristic is especially advantageous in financial datasets where the range of values can be pivotal and indicative of underlying economic factors or market sentiments. By rescaling the data within the range of 0 to 1, the Min-Max Scaler ensures that the model is fed with consistently scaled features, enhancing model performance and interpretability.

Feature Selection and Feature Generation

In the realm of financial analysis and predictive modelling, the selection and generation of features play a pivotal role in understanding market dynamics and forecasting future trends. The intricate nature of financial markets demands a multifaceted approach, where various indicators are employed to decode the complex interplay of market forces. The following section delves into the specific features used in our analysis, each meticulously chosen for its proven effectiveness in financial literature. These features range from moving averages that smooth out price fluctuations to oscillators that gauge momentum and volatility. By carefully integrating these diverse indicators, we aim to construct a comprehensive and nuanced view of market behaviours, enhancing the predictive power of our models. This blend of technical indicators, each with its unique strengths and insights, forms the backbone of our analytical framework, providing a rich tapestry of data points from which meaningful market insights can be extracted and reliable predictions can be made.

The following are the features that I have used from the existing data and a few I have generated using the existing feature set, so this portion includes feature selection as well as feature generation along with the rationale of selecting and generating these features

1. **Open Price:** The opening price of a stock is the price at which a stock first trades upon the opening of an exchange on a trading day. It's significant because it reflects the market's sentiment and expectations following all developments since the previous closing bell. The open price can be influenced by after-hours trading, news, or events that occur between trading sessions. This price is critical for setting the tone of the trading day and can be a predictor of daily market trends. It's especially useful in gap analysis, where the difference between the previous day's close and the current day's open indicates market sentiment.
2. **Close Price:** The closing price is the last price at which a stock trades during a regular trading session. For many financial analyses and models, the close is considered the most important price of the day because it is the final valuation by the market of the stock for that day. It's used as a benchmark for the next day's trading and in calculating various other technical indicators.
3. **High Price:** This is the highest price at which a stock traded during the course of the trading day. The high price is significant as it represents the maximum level of bullish sentiment during the day. It's essential for understanding the range within which a stock is trading. Analysts often look at the high price in conjunction with other indicators to gauge resistance levels and to identify potential breakout points.

4. **Low Price:** Conversely, the low price is the lowest price at which a stock is traded during the day. It reflects the point of maximum bearish sentiment. This metric is crucial for determining support levels. A consistently low price over a period can indicate a solid support level, below which the stock rarely falls.
5. **Volume:** Trading volume represents the total number of shares or contracts traded for a specified security during a given period. It is a significant indicator of the strength or intensity behind price trends. High volume often signifies that a market move is gaining support, which is crucial for validating price movements. For instance, an upward price trend accompanied by high volume is typically seen as more robust and likely to continue than the same trend with low volume.
6. **Adjusted Close Price:** The adjusted closing price amends a stock's closing price to reflect that stock's value after accounting for any corporate actions. It's vital for historical data analysis, particularly when looking at long-term performance. Adjustments can include dividends, stock splits, and new stock offerings. This metric provides a clearer picture of a stock's value over time and is often used in performance analysis.
7. **Simple Moving Average (SMA):** SMA is a widely used technical indicator that smoothens price data to identify trends. It's calculated as the average of a selected range of prices, typically closing prices, over a specific period. You've used various time windows (3, 10, 20, 50, 100, 200 days). The rationale is that different time windows can capture short-term, medium-term, and long-term trends, making SMAs crucial in trend analysis.
8. **Exponential Moving Average (EMA):** EMA is similar to SMA but gives more weight to recent prices. This makes it more responsive to new information. You've used EMAs over 10, 20, 30, and 50 days, useful for highlighting recent trends in price movements. EMAs are often used in conjunction with SMAs to determine potential price reversals.
9. **Moving Average Convergence Divergence (MACD):** This is a trend-following momentum indicator that shows the relationship between two EMAs of a security’s price. The MACD is calculated by subtracting the 26-period EMA from the 12-period EMA. The resulting MACD line is then plotted alongside a signal line, which helps identify turning points. The MACD is a well-established tool in technical analysis for identifying buy or sell signals.

10. **Rate of Change (ROC):** ROC measures the percentage change in price between the current price and the price a certain number of periods ago. It's a momentum oscillator, used to identify overbought or oversold conditions.
11. **On-Balance-Volume (OBV):** OBV is a technical trading momentum indicator that uses volume flow to predict changes in stock price. Increasing OBV suggests positive buying pressure and bullish trends, while decreasing OBV indicates bearish trends.
12. **Money Flow Index (MFI):** The Money Flow Index is a unique indicator that combines volume and price data to identify buying or selling pressure. Functioning as a volume-weighted RSI, the MFI takes into account not only the price changes but also the transaction volume, providing a more comprehensive view of market sentiment. An MFI value above 80 typically indicates that the security is overbought, suggesting that the price might drop soon, while a value below 20 indicates it is oversold, suggesting a potential price increase. The MFI is especially useful in highlighting divergences with price, which can signal potential reversals.
13. **Standard Deviation:** This statistical measure is essential in assessing the volatility of a stock. A higher standard deviation indicates greater variability in the stock's price, thus implying higher risk. In the context of financial markets, understanding volatility is crucial, as it affects the decision-making process for risk management and trading strategies. Standard deviation is often used to calculate other indicators like Bollinger Bands and is a key component in portfolio management theories, such as Modern Portfolio Theory (MPT).
14. **Average Directional Index (ADX):** The ADX is a trend strength indicator, measuring the strength of a price trend. A high ADX value (typically above 25) suggests a strong trend, which can be either up or down, while a low ADX value (below 20) indicates a weak or non-trending market. The ADX helps traders differentiate between times when a market is trending and when it is ranging, enabling them to apply the appropriate trading strategy for the market condition.
15. **Bollinger Bands:** Bollinger Bands are used to measure market volatility and are particularly effective in identifying overbought and oversold conditions. These bands consist of a middle band (usually a simple moving average), an upper band, and a lower band. The spacing between the bands varies based on the volatility, calculated using the standard deviation of the same data used for the average. Traders use these bands

to identify potential points where the market may be overstretched, either to the upside or downside.

16. **Relative Strength Index (RSI):** The RSI is a popular momentum oscillator that measures the speed and change of price movements. It oscillates between zero and 100 and is used to identify overbought (typically above 70) or oversold (typically below 30) conditions. The RSI can also be used to identify divergences where the price makes a new high or low, but the RSI does not, often indicating a potential reversal in the current trend.
17. **Volume Changes:** Volume changes are critical for confirming the strength of a trend. An upward price trend with increasing volume is generally seen as more robust and likely to continue, compared to an upward trend with decreasing volume. Volume analysis is a key aspect of market analysis, providing insights into the conviction behind price movements and potential reversals.
18. **Price Rate of Change (Price_ROC):** This indicator measures the percentage change in price between the current price and the price in a previous period. The Price ROC is a pure momentum oscillator that reflects the rate at which prices are changing. This can be especially useful for identifying trend strength and potential reversals when the ROC diverges from the price.
19. **Stochastic Oscillator (%K and %D):** This is a momentum indicator comparing a particular closing price of a security to a range of its prices over a certain period. The sensitivity of the oscillator to market movements is adjustable by altering the period or taking a moving average of the result. It helps identify overbought and oversold conditions and can provide insights into potential trend reversals.

Each of these features plays a crucial role in analyzing the complex dynamics of financial markets. From capturing basic market sentiments with price and volume to unravelling deeper trends and signals through technical indicators, they collectively enable a holistic view of market behaviour, aiding in making more informed and strategic investment decisions.

Models and Training

As I embarked on my recent project, I made a deliberate choice to use SVM and Extra Trees Classifier as my primary models. My decision was driven by a combination of literature evidence, industry practices, and specific characteristics of my training data.

Why SVM?

1. Effective with High-Dimensional Data: Given the high-dimensional nature of my dataset, SVM emerged as a suitable choice. Its capability to handle high-dimensional spaces is well-documented in academic circles, including in the works of Cortes and Vapnik.
2. Margin Maximization for Clear Decision Boundaries: SVM's principle of maximizing the margin between classes offers a robust way to establish clear decision boundaries, an essential feature for my dataset's characteristics. This approach's effectiveness is well-established in machine learning literature.
3. Kernel Trick for Non-Linear Data: The kernel trick in SVM allows it to efficiently handle non-linear relationships in data, a feature prominently discussed in Schölkopf and Smola's work on kernel methods.
4. Versatility with Different Kernel Functions: The flexibility to choose from various kernel functions in SVM meant I could tailor the model more precisely to fit the specific patterns and complexities of my data.

Why Extra Trees Classifier?

1. Handling Random Data Variations: Extra Trees Classifier is particularly effective for datasets with substantial variability, as it builds multiple trees and averages their results, reducing the impact of random fluctuations. This approach is supported by empirical evidence in ensemble learning research.
2. Efficiency with Large Datasets: Its efficiency in processing large datasets was crucial, considering the scale of data in my project. Extra Trees Classifier, part of the ensemble methods family, is known for its scalability, as emphasized in various machine learning publications.
3. Feature Selection and Dimensionality Reduction: The inherent feature selection capabilities of Extra Trees Classifier make it advantageous for high-dimensional data.

It helps in identifying the most informative features, reducing dimensionality without significant loss of information.

4. Robustness to Overfitting: Compared to other models like Random Forest, Extra Trees Classifier is less prone to overfitting, especially when dealing with noisy data. This robustness is particularly beneficial in complex datasets common in my domain.

Grid Search for Hyperparameter Tuning:

Using grid search was a strategic choice for hyperparameter tuning. The systematic exploration of parameters it offers is critical for optimizing both SVM and Extra Trees Classifier, as suggested by Bergstra and Bengio's work on random search vs. grid search.

Conclusion – Effectiveness of Models

Support Vector Machine (SVM): Effectiveness and Considerations

1. Training Phase:

The SVM model exhibits commendable effectiveness in the training phase. It typically achieves high accuracy, as evidenced by scores like 87.28% for the 'FIX' ticker in the small universe dataset. This indicates its strong capability in correctly classifying instances.

The balance between precision and recall is noteworthy, with values such as 89.05% precision and an F1 Score of 86.99% for the same ticker, reflecting a well-calibrated model.

The AUC ROC scores, often above 95%, suggest excellent separability and a strong model fit.

2. Test Phase:

On the test data, SVM generally maintains robust performance, with a slight variation across metrics. For instance, for 'DAL.csv' in the big universe dataset, the accuracy is 87.28%, with a precision of 84.04%.

The model's ability to generalize well is indicated by consistent F1 scores and AUC ROC values in the test phase.

3. Advantages:

SVM's strength lies in its versatility and effectiveness in high-dimensional spaces. It's particularly powerful in scenarios where the distinction between classes is clear.

It's effective in cases where the number of dimensions exceeds the number of samples, making it suitable for many real-world problems.

4. Disadvantages:

However, SVM models can be computationally intensive, especially with large datasets. Their performance heavily depends on the selection of the kernel.

They are less intuitive to interpret compared to simpler models, which can be a drawback in applications requiring model interpretability.

Extra Trees Classifier: Effectiveness and Considerations

1. Training Phase:

The Extra Trees Classifier demonstrates impressive performance in training, often surpassing SVM. For example, it shows an accuracy of 90.38% for the 'FIX' ticker in the small universe.

It exhibits high precision and recall values, leading to strong F1 scores. The AUC ROC values, often nearing or exceeding 97%, indicate an exceptional fit to the training data.

2. Test Phase:

In the test dataset, Extra Trees tends to maintain high performance but can show slight variability. The accuracy for 'DAL.csv' in the big universe dataset is 90.75%, with similar robustness in other metrics.

This model shows good generalization capabilities, but careful tuning is needed to avoid overfitting.

3. Advantages:

Extra Trees are known for their speed and efficiency in training, as they randomly select splits, which can be advantageous over more complex models.

They handle non-linear data well and are less prone to overfitting compared to other ensemble methods like Random Forests.

4. Disadvantages:

However, they may not perform well with very small datasets and can be less intuitive to interpret.

The randomness can sometimes lead to lower model performance if not properly tuned and validated.

Comparative Analysis: Support Vector Machine (SVM) vs. Extra Trees Classifier

Data Dimensionality and Complexity: For high-dimensional datasets, especially where the boundary between classes is distinct, SVM often outperforms Extra Trees. However, when dealing with complex, non-linear relationships in data, Extra Trees may have an edge due to its ensemble approach.

Scalability and Efficiency: In scenarios where processing large datasets swiftly is crucial, Extra Trees stand out due to their faster training times. Conversely, SVM may become impractical in large-scale applications due to its computational intensity.

Robustness and Overfitting: Extra Trees generally provide a more robust solution against overfitting compared to SVM, especially in cases where data might contain noise or irrelevant features.

Interpretability and Transparency: Both models present challenges in interpretability, but this issue is more pronounced with SVM when using non-linear kernels. In contexts where explaining the model's decisions is vital, simpler or more transparent models might be preferred.

Conclusion: Strategic Model Selection

In conclusion, the choice between SVM and Extra Trees should be a deliberate decision, guided by the specific characteristics of the dataset, the computational resources available, and the requirements of the problem at hand. While SVM excels in high-dimensional, clearly separable data, its scalability and interpretability issues may limit its applicability in certain scenarios. On the other hand, Extra Trees offer a balance of speed, efficiency, and robustness, making them suitable for a broader range of applications, albeit with some limitations in small datasets and interpretability. As with any machine learning task, careful consideration of these factors, along with rigorous validation and tuning, is crucial to harness the full potential of these powerful models.

Accuracy? Real World Considerations

The Allure of High Accuracy:

High accuracy in a trading model suggests a strong ability to correctly predict market movements. In theory, this should translate to successful trades and profits.

However, accuracy alone can be misleading. It doesn't account for the magnitude of gains or losses, nor does it consider the cost of trades and market impact.

The Reality of Market Dynamics:

Financial markets are complex adaptive systems influenced by a multitude of factors, including economic indicators, company performance, geopolitical events, and trader psychology.

This complexity means that even a high-accuracy model might struggle to consistently predict market movements, especially in volatile or unprecedented scenarios.

Challenges in Real-World Trading

Overfitting and Market Adaptability:

A common pitfall is overfitting, where a model performs well on historical data but fails to adapt to new market conditions.

Real-world markets continuously evolve, and a model that's not regularly updated or that doesn't account for this dynamic nature can quickly become obsolete.

Transaction Costs and Slippage:

Trading incurs costs such as brokerage fees, bid-ask spreads, and taxes, which can erode profits.

Slippage, the difference between the expected price of a trade and the price at which the trade is executed, can further impact trading outcomes, especially in fast-moving markets.

Risk and Reward Trade-Offs:

High returns often come with high risks. A model that aims for high accuracy might be overly conservative, missing profitable opportunities, or take excessive risks that can lead to substantial losses.

The Sharpe ratio, a measure of risk-adjusted return, is crucial in this context. A low Sharpe ratio indicates that the returns are not commensurate with the risks taken.

Enhancing Realism and Effectiveness in Trading Models

Diversification and Risk Management:

Diversifying trades across different assets, sectors, and geographies can reduce risk. It's not just about picking winners but also about spreading risk.

Effective risk management strategies, such as setting stop-loss orders and position sizing based on the volatility of assets, are essential.

Incorporating Multiple Data Sources:

Beyond price and volume data, incorporating alternative data sources like news sentiment, economic indicators, and social media trends can provide a more holistic view of the market.

Machine learning models can be trained to process and extract meaningful patterns from these diverse data sets, potentially improving predictive power.

Dynamic and Adaptive Models:

Models should be adaptive, capable of learning from new data and adjusting to changing market conditions.

Techniques like rolling window analysis, continual retraining, and incorporating feedback loops can help models stay relevant.

Robust Backtesting and Forward Testing:

Rigorous backtesting, simulating trades with historical data, is crucial. However, it's important to avoid hindsight bias and overfitting.

Forward testing, or paper trading, where the model's predictions are tested in real-time without actual financial risk, can provide additional insights into its real-world performance.

Understanding Model Limitations:

It's vital to recognize that no model can predict the future with certainty. Markets can be influenced by unforeseen events (e.g., geopolitical crises, economic shocks).

Traders should be wary of relying solely on model predictions and consider combining quantitative models with qualitative analysis.

Continuous Learning and Improvement:

Financial modelling and trading is an iterative process. Continuous learning from past trades, both successful and unsuccessful, is key to improvement.

Keeping abreast of the latest developments in financial markets, trading strategies, and machine learning algorithms can provide a competitive edge.

Conclusion

In conclusion, while a high-accuracy trading model can be an attractive proposition, its real-world application requires careful consideration of various factors beyond mere accuracy. Challenges such as overfitting, transaction costs, market volatility, and the need for dynamic adaptability must be addressed. Enhancing the realism and effectiveness of a trading model involves a holistic approach, integrating robust risk management, diverse data sources, continuous learning, and a deep understanding of market dynamics. As the trading landscape evolves, so must the models and strategies employed

Note: For Trading Reports and Graphs please see the reports Folder