Predicting Market Directions

By John Lee

What is a Classifier?

- A type of supervised learning algorithm that learns from labeled training data to assign input instances to one or more predefined classes.
- The goal is to learn a decision boundary that separates different classes based on the input features.
- Can be used on binary class or multi-class problems.
- Choice of classifier depends on various factors such as the nature of the data, the complexity of the problem, the interpretability requirements, and the performance goals.

Classifier Use Cases

- Recognize economic cycles (e.g., expansion, peak, through, recession)
- Classify credit card transactions as legitimate or fraudulent
- Identify early warning signs of a company's bankrupt based on financial indicators, market trends, and other relevant factors.
- Determining market trend structures (e.g., uptrend, congestion, downtrend)
- Categorizing financial news, blogs, or letter to shareholders as having different sentiments (e.g., positive, neutral, or negative sentiments)
- Clustering traders or investors into different groups based on their trading preferences, risk profiles, or investment strategies.

Logistic Regression (Classification)

- The gogal is to estimate the probabilities of the binary outcome based on the input features.
- Fits a logistic curve to the data, which separates the two classes by finding an optimal decision boundary.
- This boundary is determined by estimating the coefficients (weights) associated with each predictor variable.
- Like other regression models, can use MLE to find the optimal values for the coefficients.

Converting Response into Binary

- Convert the IWM log returns into binary values, 0 represents a day with negative return (downward direction) and 1 represents a day with positive return (upward direction).
- The problem now becomes using the BMED and IVEG daily log return values to predict IWM daily log return directions.

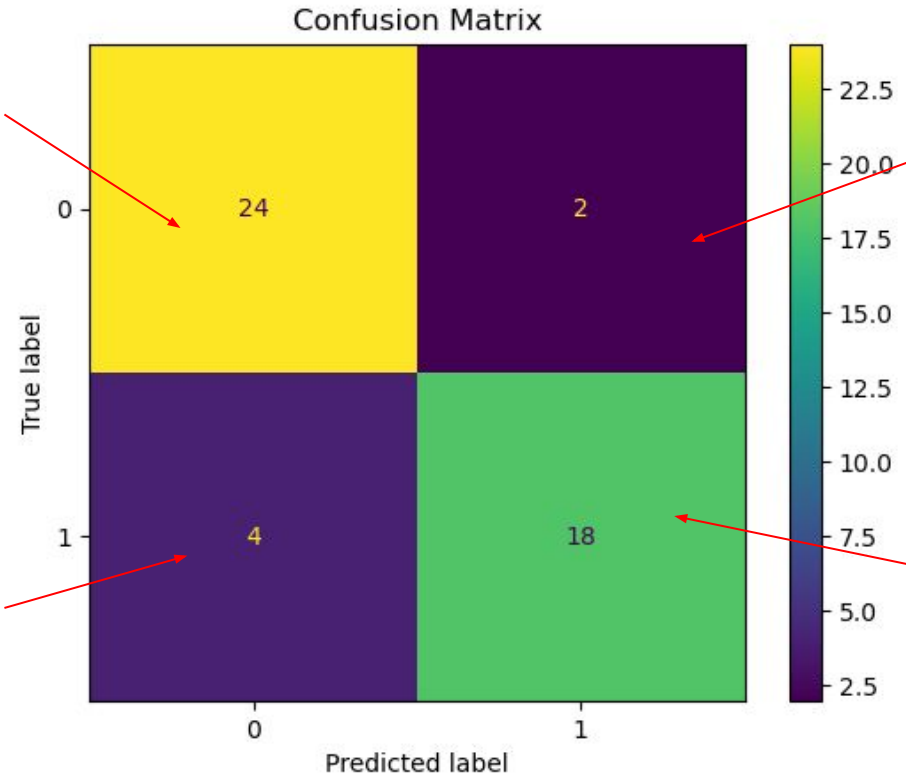
	BMED	IVEG	IWM
2022-06-02	1.519343	1.553735	0.024129
2022-06-03	-0.709552	-0.918115	-0.009047
2022-06-06	-0.212229	0.228417	0.004321
2022-06-07	1.059938	0.144687	0.015999
2022-06-08	-0.339705	-1.122713	-0.015574

BInary conversion

	BMED	IVEG	IWM
2022-06-02	1.519343	1.553735	1
2022-06-03	-0.709552	-0.918115	0
2022-06-06	-0.212229	0.228417	1
2022-06-07	1.059938	0.144687	1
2022-06-08	-0.339705	-1.122713	0

Fitting a Logistic Regression

of correctly predicted downward days



# of doward days predicted as upward days

# of upward days predicted as downward days

# of correctly predicted upward days

Training and Testing Accuracies

Training Accuracy:

- Measures the accuracy of the model on the data it was trained on.
- Can be misleading as it may not reflect the model's ability to generalize to unseen data. If the model overfits the training data, it may achieve high training accuracy but perform poorly on new, unseen data.

Testing Accuracy:

- Measures the accuracy of the model on the testing dataset or validation dataset.
- More reliable as it assesses the model's performance on data it has not seen during training.

```
training accuracy score: 0.9157894736842105
testing accuracy score: 0.875
```

Advantages of Logistic Regression

- **Simplicity:** doesn't require assumptions like normally distributed feature values for linear regression.
- **Interpretable:** predicts probabilities rather than just class labels. This allows for a more nuanced understanding of the uncertainty and confidence associated with the predictions.
- **Efficiency:** can handle large datasets with a relatively small number of predictors.
- **Feature Importance:** by examining the magnitude of the coefficients, you can determine which predictors have a stronger influence on the outcome.

Disadvantages of Logistic Regression

- **Linearity assumption:** assumes a linear relationship between the predictors and the log-odds of the outcome. Otherwise, it may not perform well.
- **Limited complexity:** too simple and may not capture complex relationships between predictors (e.g., may struggle with datasets that have nonlinear decision boundaries)
- **Sensitivity to outliers:** outliers with extreme values can have a large impact on the model's coefficients and predictions.
- **Independence of observations:** assumes that the observations are independent of each other. Otherwise, it may not perform well.
- **Imbalanced classes:** may not perform well when dealing with imbalanced classes

Predicting Market Directions

By John Lee