Stock Prediction Challenge

Georgios Gkolemis

The initial set of 47 features was reduced to 15 to minimize noise and capture market dynamics.

Dataset

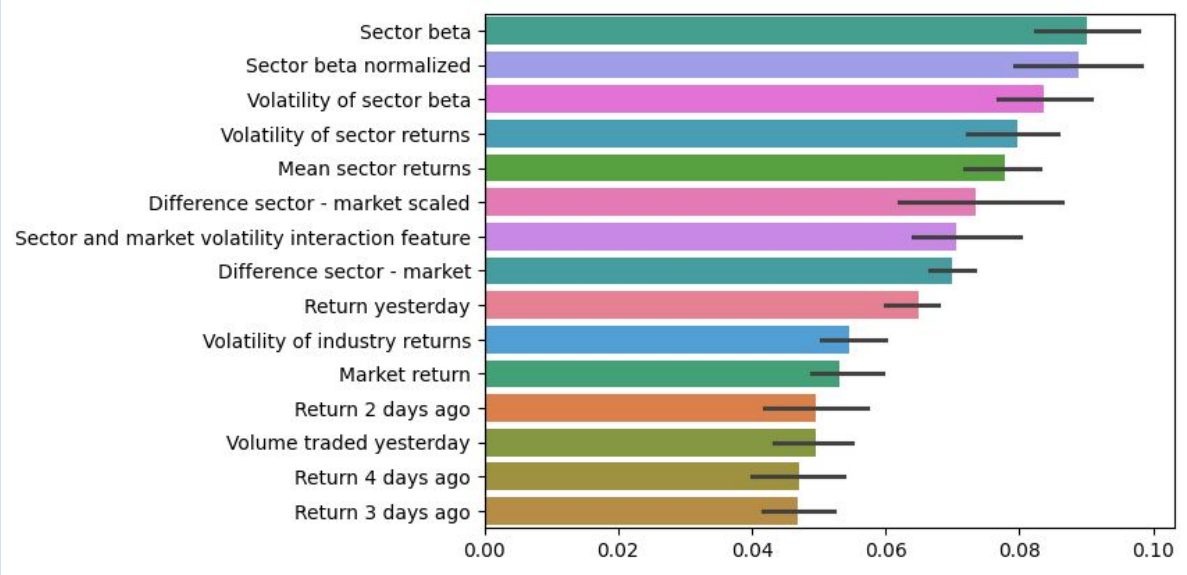
- Date indexes (randomized)
- Sector, industry, stock indexes
- Residual returns over 20 days
- Volumes traded over 20 days
- Sign of residual stock return

Feature engineering

- Residual returns over 5 days and volume traded over 1 day
- Sector mean return
- Market mean return
- Sector beta
- Difference between sector and market mean return
- Standard deviation (std) of sector and industry returns
- Three more interaction features through multiplication/ scaling

Sector-wide features show highest significance.

- The data was fitted to a Random Forest model in order to identify the most significant features.
- Aggregations by date and sector, industry, industry group, sub-industry were tested for significance.
- Sector-wide statistics such as mean returns, volatility and beta proved most relevant.



The model is a fully connected neural network that yields an accuracy of 51.64% on the test set.

Neural net architecture

1. The input layer receives the 15 features and passes them through **three hidden layers**.
2. **ReLu activation function** in the first two hidden layers captures non-linear relationships.
3. **Batch normalization** helps convergence and **dropout** applied to hidden layers prevents overfitting.
4. Output layer is a sigmoid function producing a **probability score** for the positive (True) class.

➔ Hyperparameter tuning informed the number of neurons, hidden layers, dropout rate and batch size.

Training and prediction

1. **Five epochs** of training for each fold of a 4-fold cross validation process.
2. Training and validation sets are scaled separately to avoid data leakage.
3. Validation accuracy is equal to 51.49%
4. This generalizes to the test set with a **prediction accuracy of 51.64%.**