Features & Importance

Advanced Topics in Financial Machine Learning

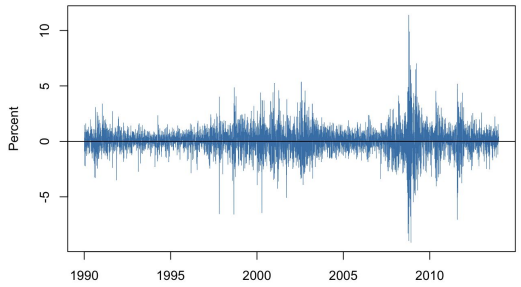
HUDSON
AND THAMES

Backtesting is not a research tool.
Feature importance is.

Marcos Lopez de Prado

Fractional Differentiation

- Inferential analysis requires data with invariant processes, such as:
  - Returns on prices (or log returns)
  - Changes in yield
  - Changes in volatility
- First order differencing makes the series stationary, at the expense of removing all memory from the original series.



- Memory is the basis for the model's predictive power.
  - Example, equilibrium (stationary) models need some memory to assess how far the price process has drifted away from the long-term expected value in order to generate a forecast.
- Problem:
  - Returns are stationary however memory-less.
  - Prices have memory however they are non-stationary.

Fractionally Differentiated Features

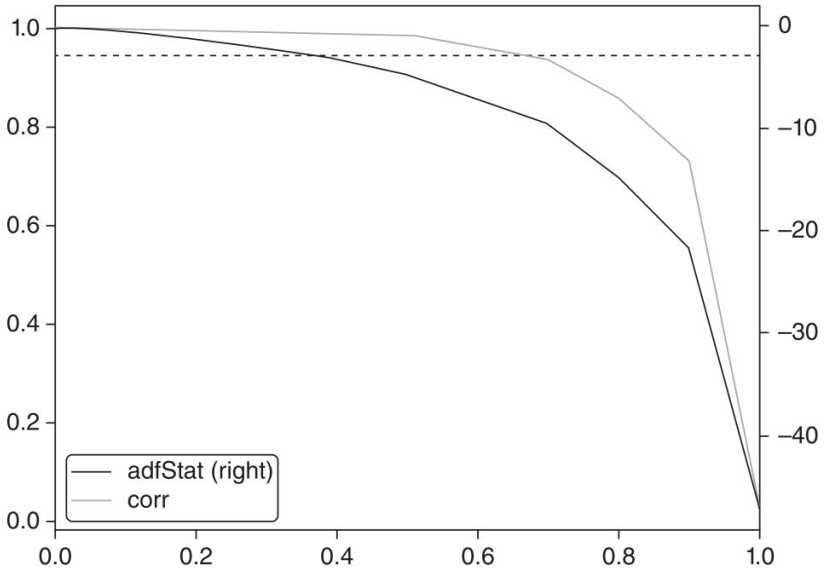


FIGURE 5.5 ADF statistic as a function of *d*, on E-mini S&P 500 futures log-prices

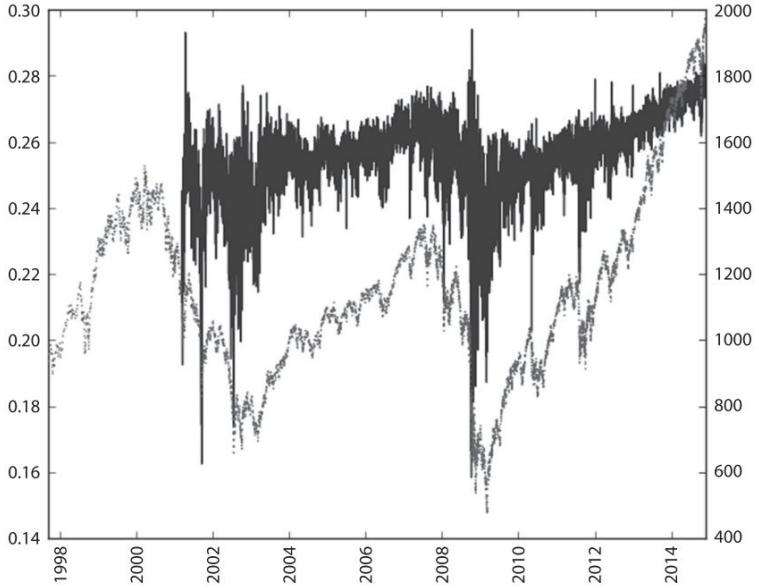
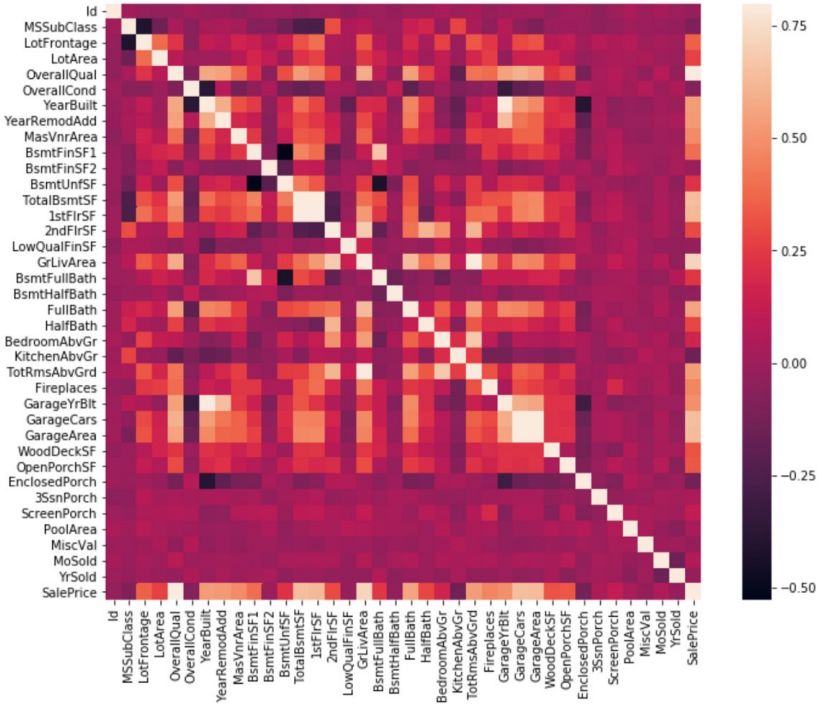


FIGURE 5.4 Fractional differentiation after controlling for weight loss with a fixed-width window

HUDSON
AND THAMES

Substitution Effects



- **Multicollinearity** is a **problem** because it undermines the statistical significance of an independent variable.
- MDI & MDA dilutes the importance of substitute features, because of their interchangeability: The importance of two identical features will be halved, as they are randomly chosen with equal probability.

Feature Importance: Mean Decrease Impurity (MDI)

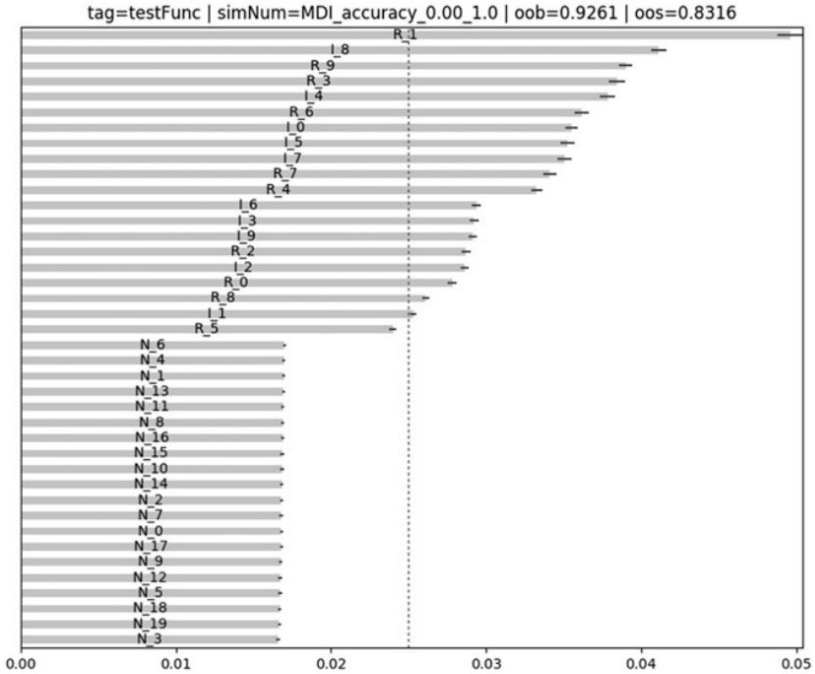


FIGURE 8.2 MDI feature importance computed on a synthetic dataset

- Uses in-sample performance to estimate feature importance.
- Standard sklearn's feature importance method.
- Better to use with **max_features = 1** to address the problem of masking effects.
- Can be applied to tree-based classifiers only.
- Suffers from substitution effects.

HUDSON
AND THAMES

Feature Importance: Mean Decrease Accuracy (MDA)

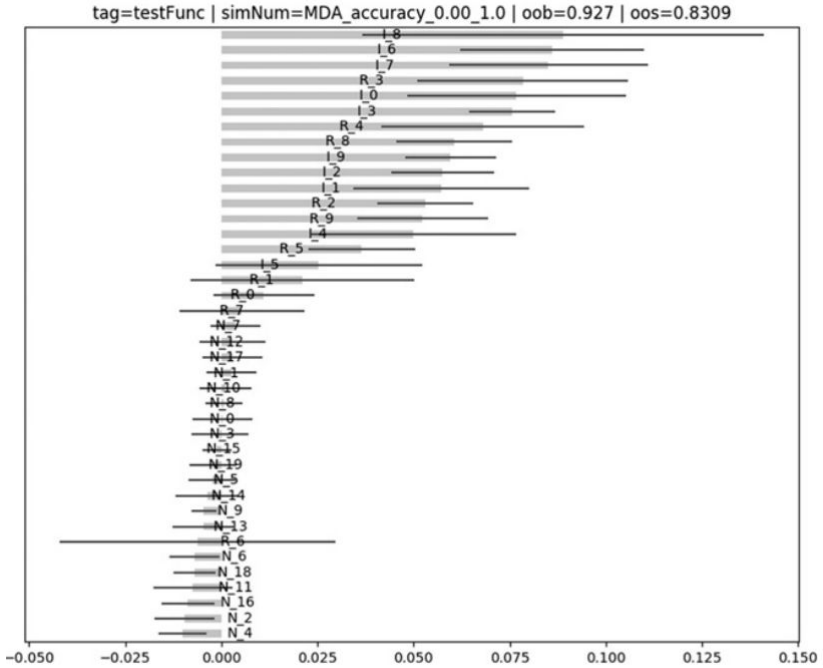


FIGURE 8.3 MDA feature importance computed on a synthetic dataset

- Uses out-of-sample performance to estimate feature importance.
- Any score (F1, ROC-AUC, Precision/Recall) can be used as estimation metric.
- Can be applied to any classifier.

HUDSON
AND THAMES

Feature Importance: Single Feature Importance (SFI)

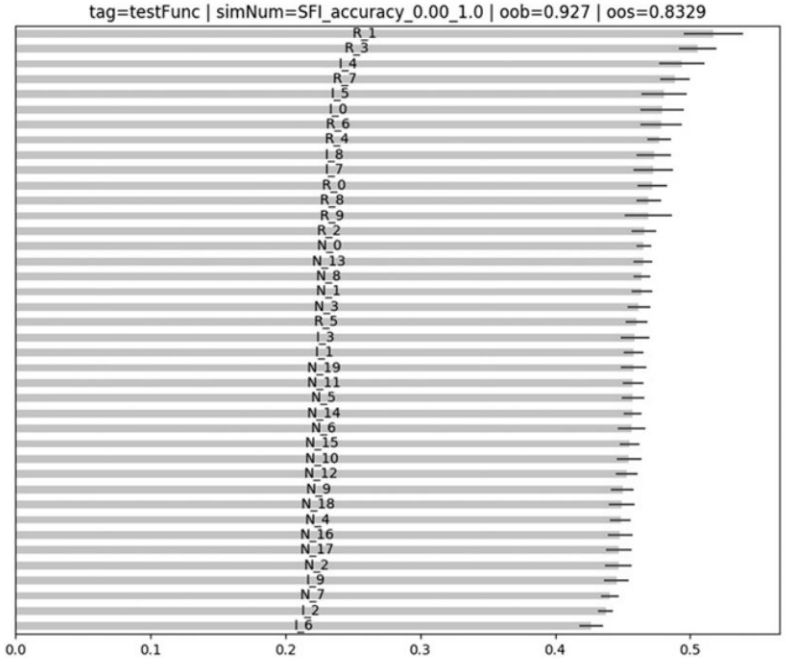


FIGURE 8.4 SFI feature importance computed on a synthetic dataset

- Uses out-of-sample performance to estimate feature importance.
- Any score (F1, ROC-AUC, Precision/Recall) can be used as estimation metric.
- Does not account for joint effects.
- Doesn't suffer from substitution effect.


HUDSON
AND THAMES
Thank You