Machine Learning of Financial Time Series – a Case Study

Lan Gong

April 6th, 2018

Introduction

- Financial time series is discrete in time but continuous in value
- Asset returns are modeled instead of prices
  - Prices are usually highly correlated from day to day
  - Variance of price can grow over time
  - Returns reflect the change of price over some period (e.g., daily etc.)
  - Returns have more desirable statistical features
- Challenges of Prediction
  - low signal-to-noise ratio ("noisy")
  - Economic uncertainties ("event-driven")
- Predictive methods
  - Statistical analysis using a few fundamental variables (e.g., indices, GDP, unemployment rate, consumer index)
  - Machine learning comes into play as more data are collected and used for prediction (e.g. alternative data such as satellite images)

2Sigma Financial Modeling Challenge

- Predict investment returns has been a central topic in trading and risk management
- Leverage data set of “2sigma financial challenge” as a playground to explore financial time series and apply machine learning methods
- Data sets are not “very clean”, representing some of the real world challenges

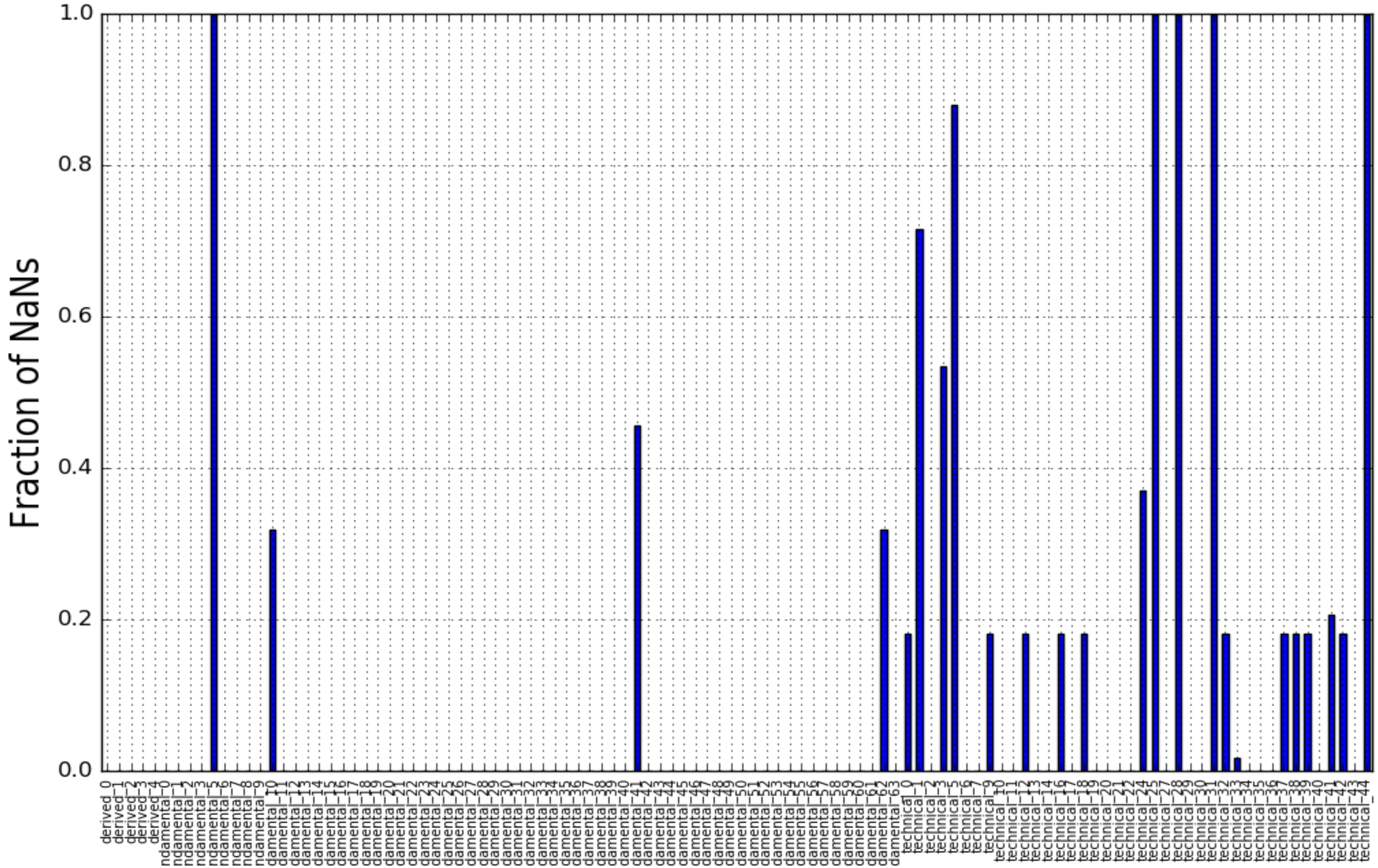
Overview of Data Set

- Description
  - Time series of financial instruments with anonymized features and one target variable for prediction
  - No further information provided on the meaning of the features or transformations applied to them
  - No information about the type of an instrument
- Dimensions
  - 5 years & ~1000 instruments per timestamp: total 1MM+ observations
  - 100+ features
  - The variable to predict is ‘y’ – presumably investment returns

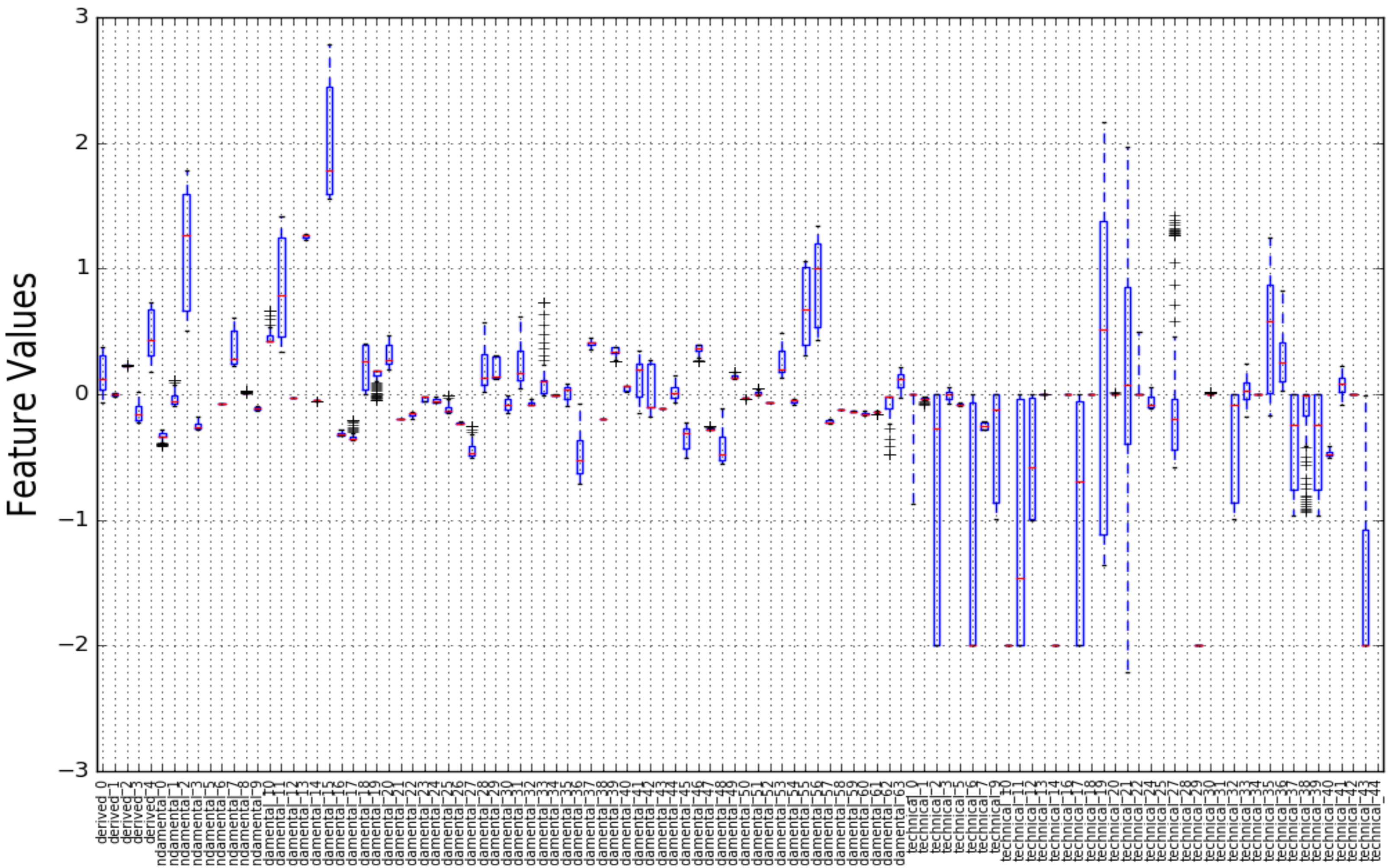
More about the Data Set

- Each instrument is labeled with a unique id
- An instrument doesn't need to have values for all the features, e.g., stocks and bonds differ in the available features
- Feature values are not centered and can have outliers
- Collinearity among the features
- 'y' is approximately normal but may differ slightly among instruments

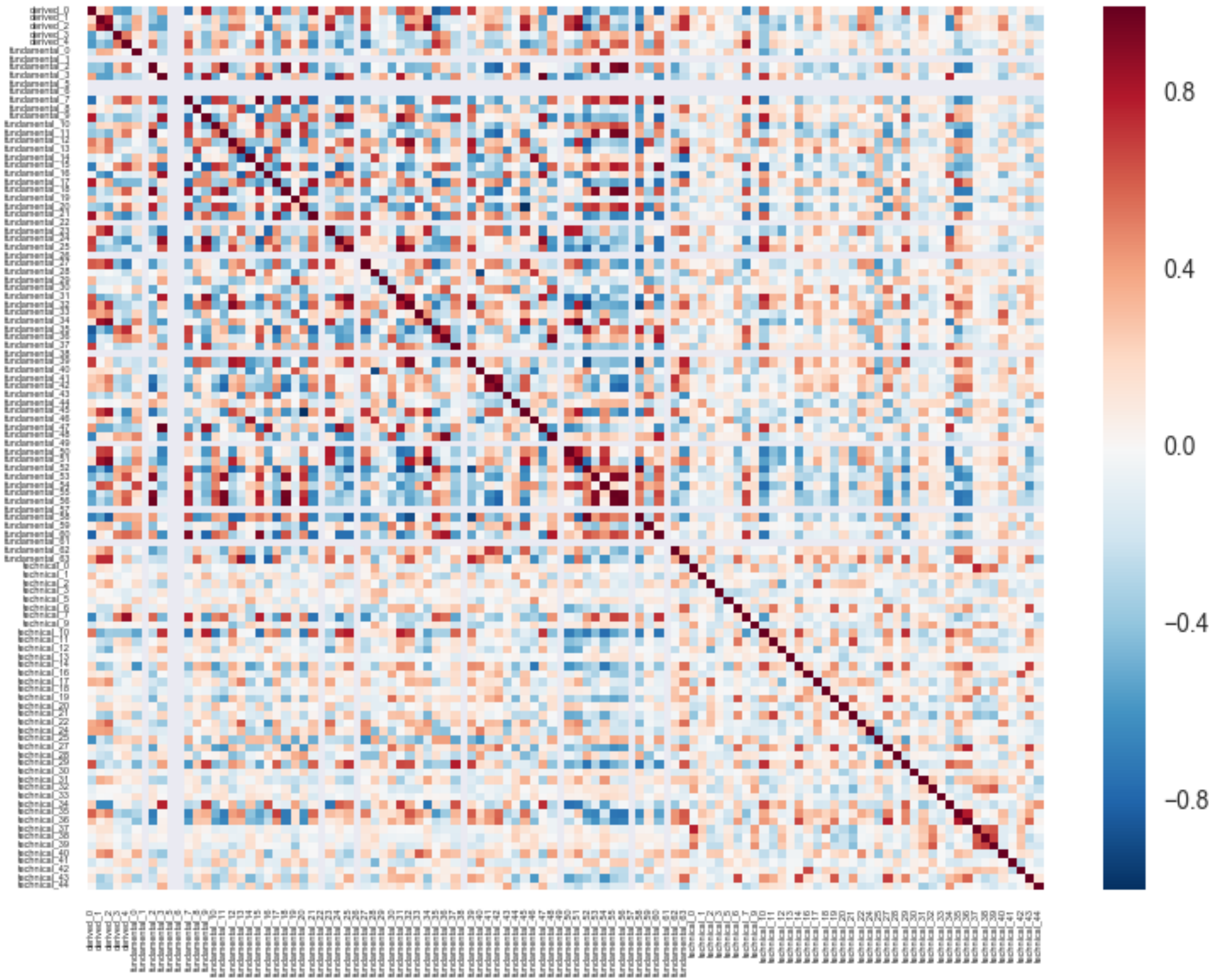
Percent of NaNs for one instrument



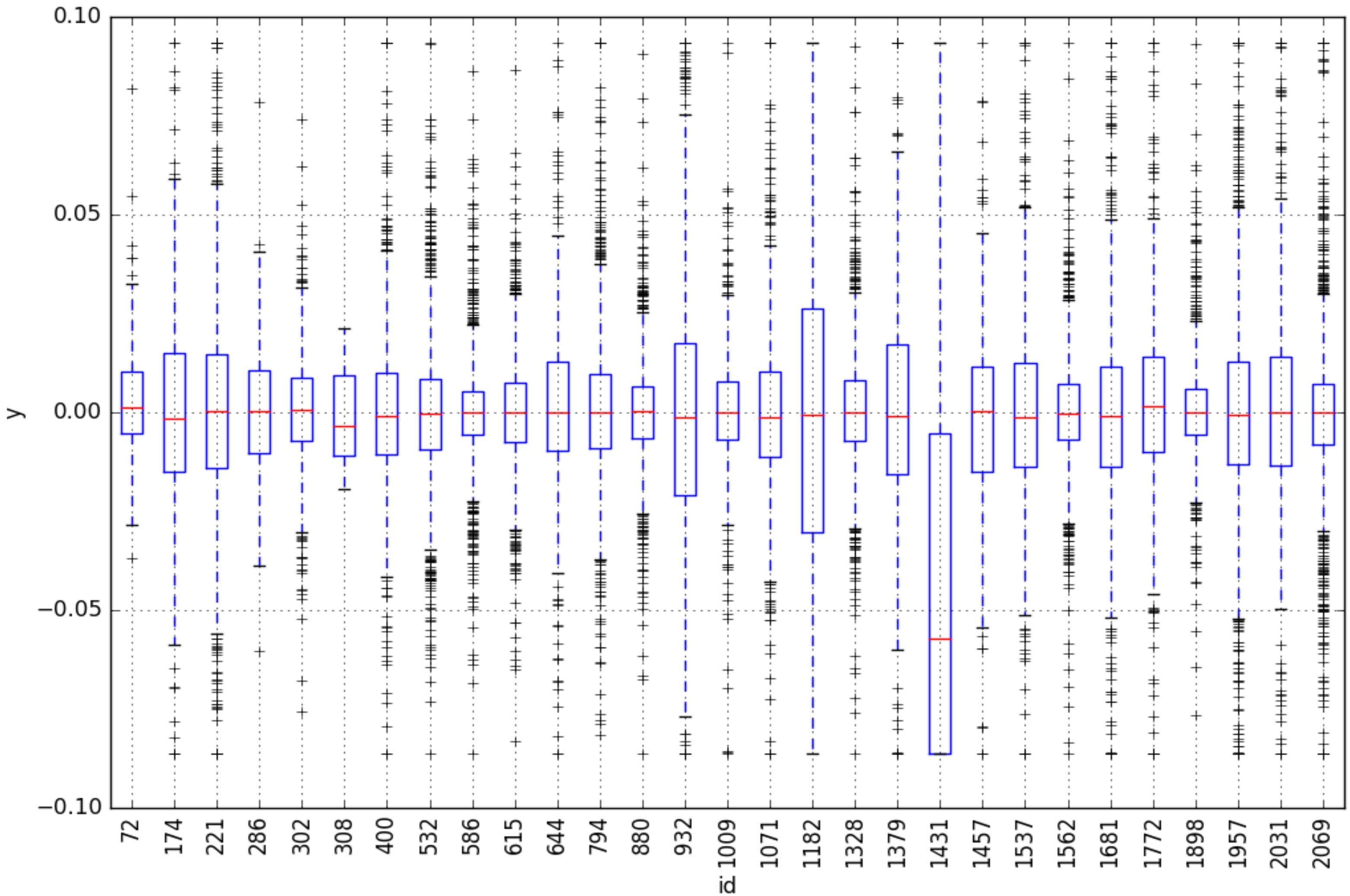
Distribution of feature values for one instrument



Heat map of pairwise correlation among features



Distribution of returns by id
(only a subset of ids is shown)



Data Preprocessing

For each id:

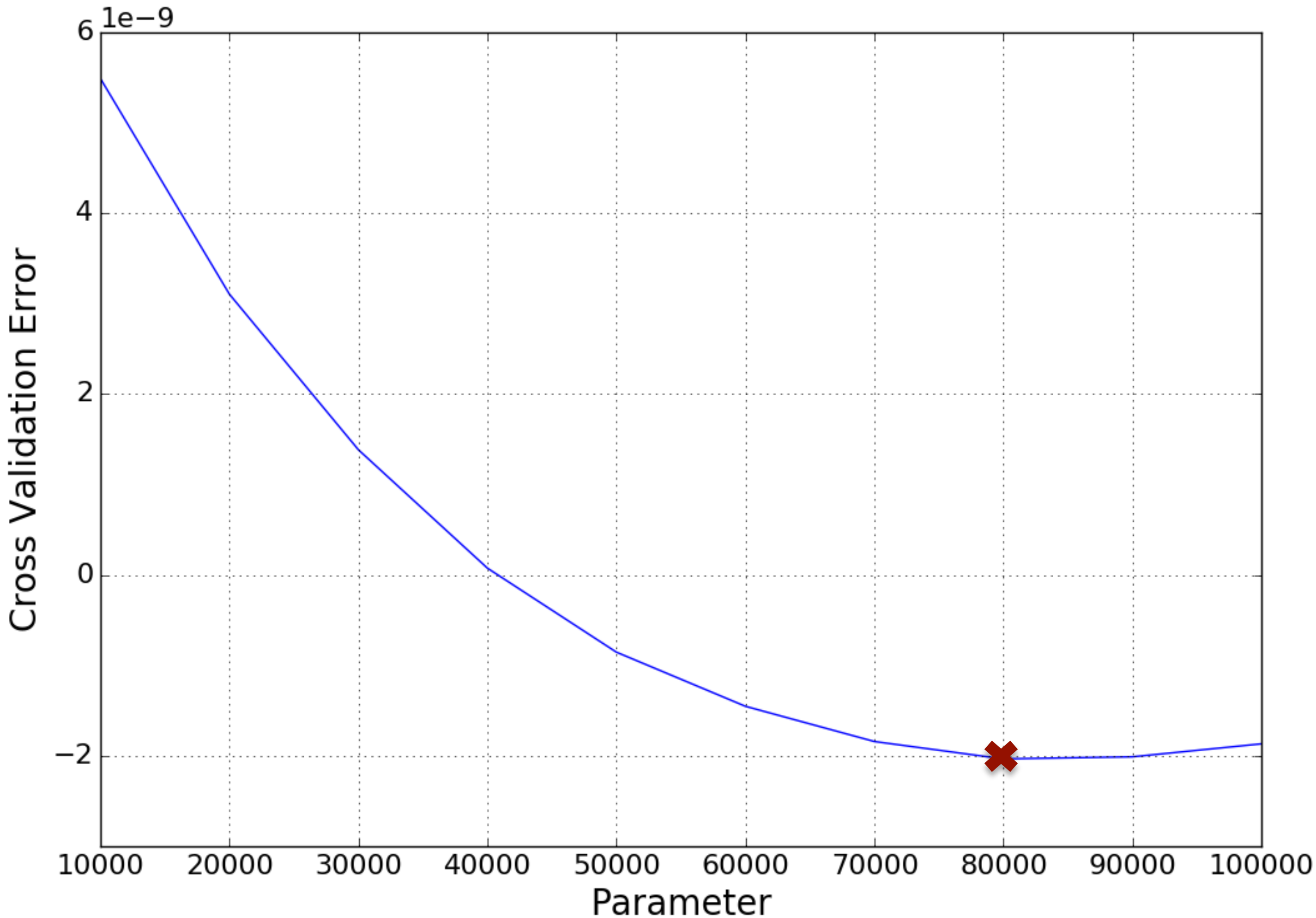
- Clip the outliers
- Standardize the feature values
- Fill the NaNs with means

Modeling Approaches

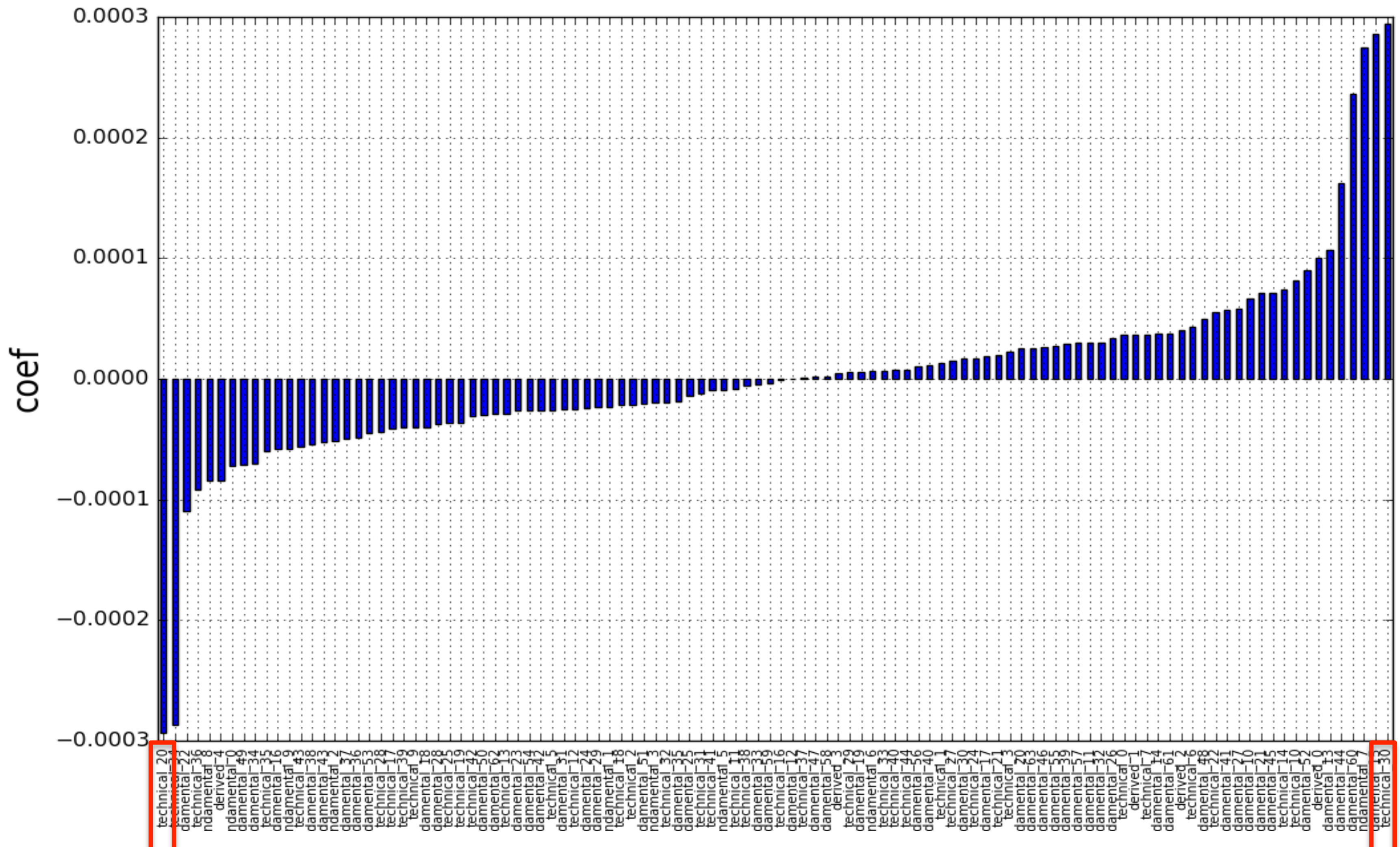
Ridge Regression

- To address the issue of collinearity
- Penalty tuning: Cross validation on training set
- Coefficients: a few features have larger coefficients than the rest

Cross validation error vs. Penalty Parameter



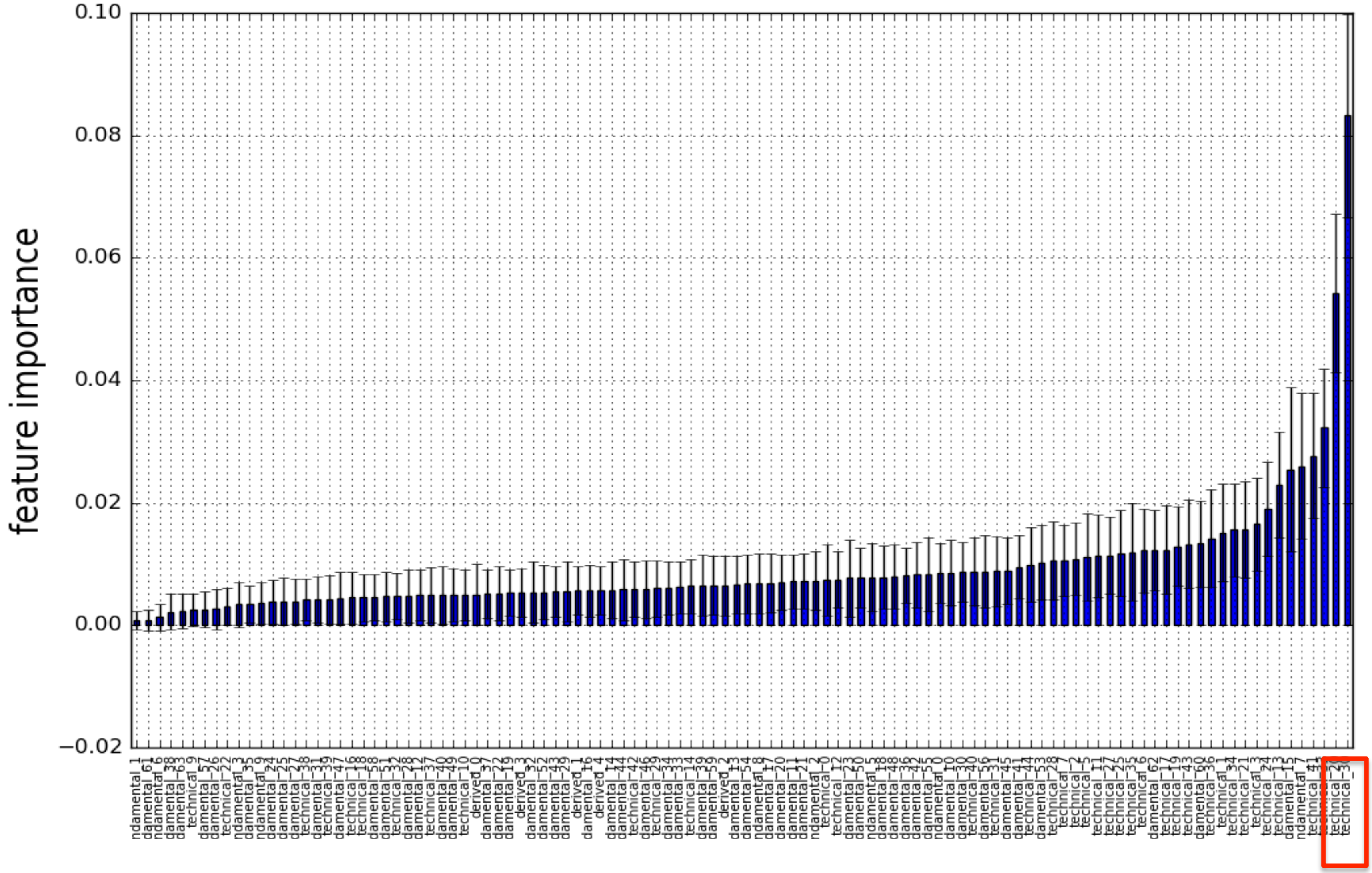
Coefficients of Features



Random Forest

- To capture non-linear behaviors
- Cross validation to find the optimal parameters, e.g., number of trees, tree depth
- Feature importances echo the implication from ridge regression

Feature importance



Mixed model

- Intuition:
  - Ridge regression captures the linear influence from a few dominant features
  - Random forest reflects the hierarchical influence of all variables “conditioned” on prior splits of more dominant ones
  - Adding historical mean return to ridge results to reflect variability among instruments
- Weighted sum of prediction results from each model.

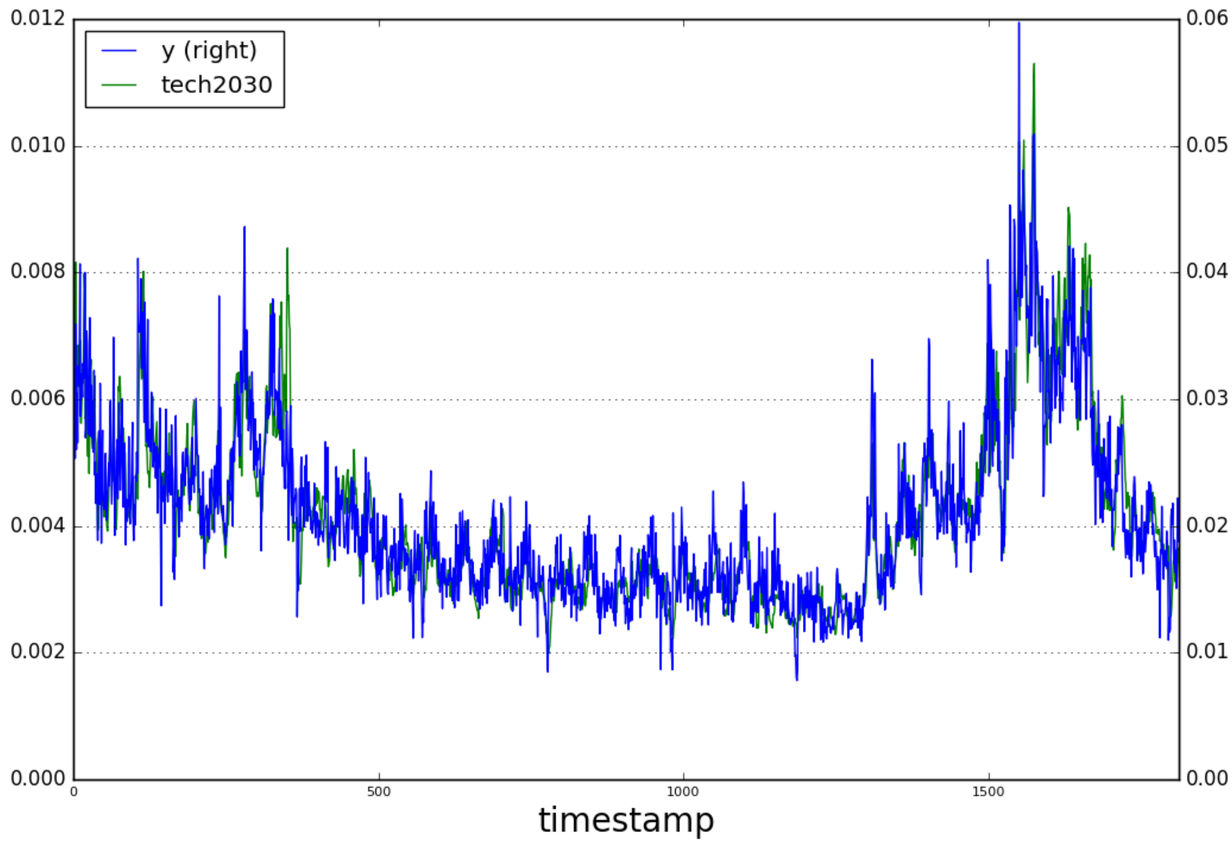
Summary of Testing Results

Modeling Approach	R^2	Improvement
Principal Component	0.095%	Baseline
Ridge Regression	0.10%	5%
Random Forest	0.14%	47%
Mixed Model	0.2%	111%

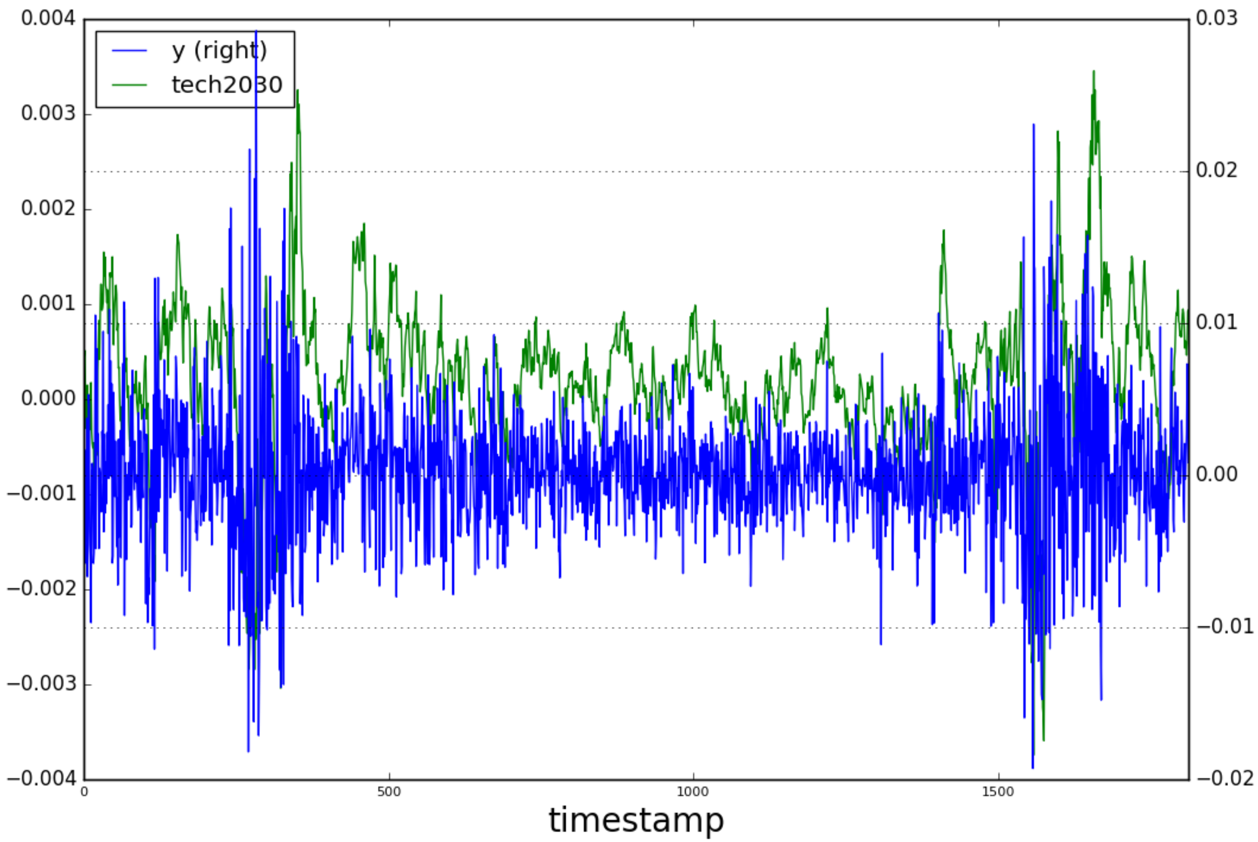
Further Analysis of Important Features

- 'technical_20' and 'technical_30' have been shown as important features by both Ridge and Random Forest regressions.
- Picking instruments in a strategic way to replicate market movement, e.g. S&P500, is a common practice in portfolio management.
- Each instrument can have its own behavior but it is the collective movement of all that matters.
- The grouped instruments forms a "*portfolio*".
- Method of analysis:
  - Construct a new feature: 'tech2030' = technical_20 – technical_30
  - Compute the portfolio-level statistics of the new feature and 'y'

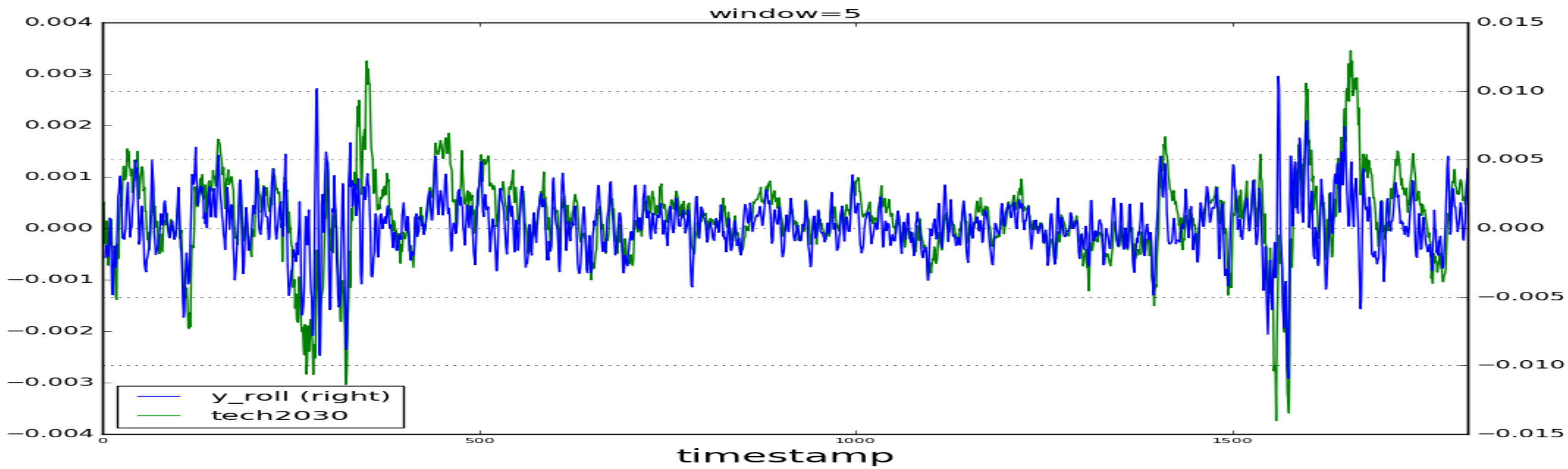
Standard deviation of the constructed feature and 'y'

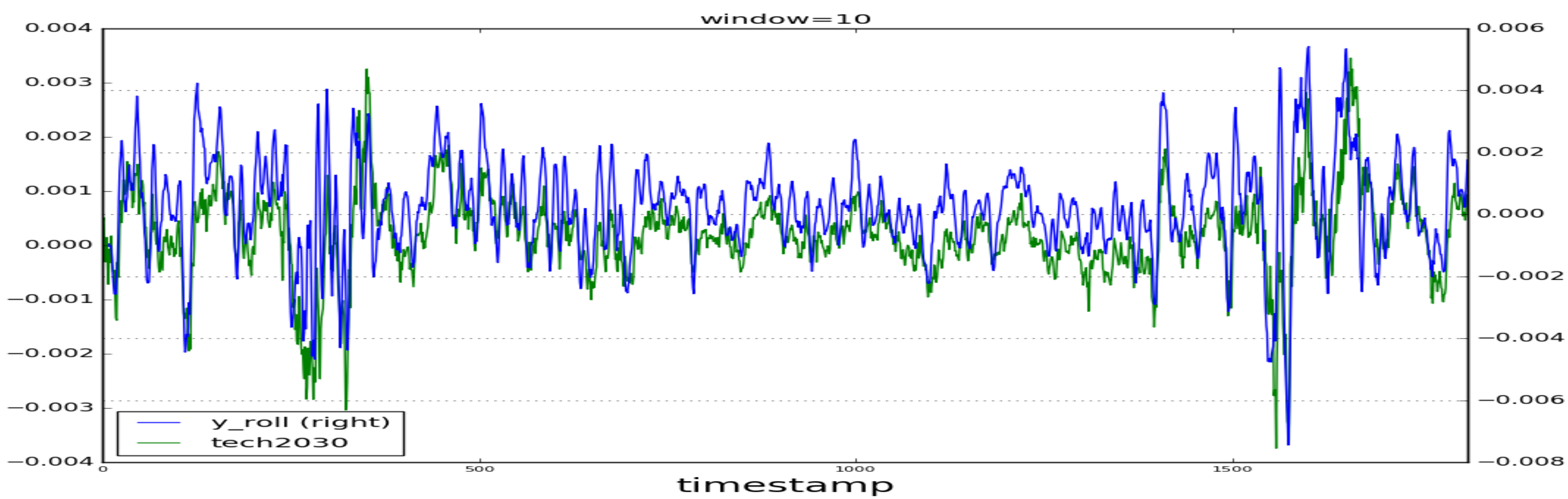


Mean of the constructed feature and 'y'



Mean of the constructed feature vs. *n-day* rolling mean of 'y'





Conclusions

- Initial data exploration helps to choose the proper prediction models
- Feature selection using ridge and random forest regression helps to identify a few dominant variables
- Mixed model can predict better than the single best model
- The constructed feature (*i.e., technical_20 – technical30*) can be a general market index that the portfolio-level returns try to track

Thank you!