MTH9899 Machine Learning Final Project

Hongshan Chu, Yuchen Qi, Linwei Shang, ShengQuan Zhou

Baruch MFE

May 24, 2017

Overview of Dataset

Training dataset

- $\sim$ 140,000 rows;
- key: stock ID and a timestamp;
- 27 features: 17 quantitative and 10 categorical;
- A *weight* column and an output column.

General guidelines

- Predictions are made based on the information contained in each row, without cross-row reference.
- No time series modelling is explored, partly because the data points per ID along the time axis are inhomogeneous and incomplete.
- No stock-specific modeling is explored given that the data points per ID are inhomogeneous and incomplete.
- Timestamp information is used only for dividing the dataset into training set and test set.
- Weighted R^2 is used as the final benchmark.

Data Preprocessing

The provided dataset is divided into two parts according to timestamp:

- The first 2/3 are used for training and testing;
- The remaining 1/3 are reserved for a production run.

A series of split points are chosen to divide the first 2/3 of the complete dataset into two parts:

- The first part for training;
- The second part for testing.

Selection of Quantitative Features

Three tests are performed to select quantitative features:

- Pearson correlation coefficient;
- Kendall's rank correlation coefficient, also known as Kendall's τ;
- Spearman's rank correlation coefficient, also known as Spearman's ρ;

with respect to the training set of output data, based on the criterion that the p-value for correlation coefficients being less than 3%.

Feature	Pearson	p-value	Kendall's τ	p-value	Spearman's ρ	p-value
x0	-0.019	10^{-6}	-0.013	10^{-6}	-0.019	10^{-6}
x17	$+0.016$	10^{-5}	$+0.009$	10^{-3}	$+0.013$	10^{-3}
x22	$+0.026$	10^{-11}	$+0.014$	10^{-7}	$+0.020$	10^{-7}
x49	$+0.015$	10^{-4}	$+0.009$	10^{-4}	$+0.013$	10^{-4}
x53	$+0.018$	10^{-6}	$+0.012$	10^{-6}	$+0.018$	10^{-6}
x61	-0.009	0.03	-0.009	10^{-3}	-0.013	10^{-4}

Selection of Categorical Features

Similar tests are performed on categorical features:

Feature	Pearson	p-value	Kendall's τ	p-value	Spearman's ρ	p-value
x2	-0.038	10^{-22}	-0.026	10^{-21}	-0.037	10^{-21}
x6	-0.014	10^{-4}	-0.007	0.02	-0.010	0.01
x30	$+0.020$	10^{-7}	$+0.014$	10^{-6}	$+0.018$	10^{-6}
x46	$+0.026$	10^{-11}	$+0.018$	10^{-10}	$+0.026$	10^{-10}
x51	$+0.017$	10^{-5}	$+0.011$	10^{-4}	$+0.015$	10^{-4}

A selected list of categorical features are treated as

- ordinal numbers; or
- one-hot dummy variables.

Inspection of Period-by-Period Correlations

Period	x17	x49	x53	x22	x46	x61	x0	x30	x42	x51
1	.007	.014	.022	.018	.027	.008	-.013	.005	-.006	.003
2	.014	.019	.010	.020	.020	-.009	-.029	.012	-.003	-.007
3	.021	.006	.016	.015	.014	-.014	-.027	-.002	.014	-.027
4	.018	.017	.025	.045	.032	-.021	-.006	.017	-.017	-.002
5	.016	.012	.009	.044	.033	-.002	-.007	.019	-.019	-.009
6	.009	.017	.012	.009	.032	-.008	-.014	-.0004	-.011	.0005
7	.045	.008	.024	.045	.007	-.004	-.008	.011	-.016	-.003
8	.015	.012	.011	.021	-.012	.003	-.009	.013	-.014	-.003
9	.021	-.010	.006	.019	.002	.004	-.016	-.026	.027	-.013

Removal of Outliers

As a final step of data cleaning, outliers observed in features and outputs in the training set are removed:

- Remove all data rows with output $|y| > 0.05$;
- Remove all data rows that have outliers in at least one column, based on the criterion z-score > 3.

Linear Regression & Random Forest

- Linear Regression

Method	Feature	In-Sample R^2(bps)	Out-of-Sample R^2(bps)
OLS	all	+47.20	−23.37
OLS	selected	+6.28	+7.33
Ridge	selected	+3.91	+7.19
Lasso	selected	+3.10	+7.01

- Tree-based Methods

Method	In-Sample R^2(bps)	Out-of-Sample R^2(bps)
Random Forest	+5.3	+3.9
Boosting Trees (2:1)	+10	+4.7
Boosting Trees (3:1)	+10	+12.7

Selected Details on Testing Boosting Trees

Out-of-sample R^2 for the method of boosting trees:

Train # / Test #	R^2(bps) No Outlier Removal	R^2(bps) Outlier Removal
1:1	+3	+4
2:1	+4	+5
3:1	+15	+13
4:1	+15	+13
5:1	+7	~ 1
6:1	~ 1	−5
7:1	−2	−3
8:1	−1	−2

Conclusion and Outlook

- The methods of linear regressions, random forests, and gradient boosting trees are tested.
- With a series of procedures including feature selection and data cleaning, a range of values $5 \sim 15$(bps) can be reached for the out-of-sample R^2 for boosting trees.
- Regime shifting behavior are observed where the correlation between the features and the output vary over time.
- A mixture of models, for example, boosting trees combind with random forest, is expected to improve the performance.