Celebrating the Contributions of the M6 Forecasting Competition

Presentation at the M6 Conference

Chris Fry (chrisfry@google.com), 2023-11-07 (12:00 PM EST)

With special thanks to the following contributors:
Ruming Wang (rumingw@google.com), Kashif Yousuf (kyousuf@google.com), Greyson Liu (greysonliu@google.com), Chris Dobronyi (dobronyi@google.com)

Recommendations to M6 Organizers - 2022

Spryos Makridakis, C. Fry, F. Petropoulos, E. Spiliotis, “The Future of Forecasting Competitions: Design Attributes and Principles”, INFORMS Journal on Data Science, Vol. 1, No. 1, April–June 2022, pp. 96–113.

Theme	What we recommended
Replicability	Submission of source code or executable that can be rerun on the same or new inputs / datasets.
Representativeness	Represents real-world challenge(s); similar to problems faced by other forecasters; broad coverage across time series feature space and range of possible external effects.
Robust Evaluation	Rolling origin evaluation, covering different times of year; multiple evaluation rounds in a live setup.
Measuring Decision Impact	Measure utility of forecasts directly through their impact on decisions, reflected in monetary terms when possible.
Showcase Forecast Value-Added (FVA)	Demonstrate value of winning solutions against state-of-the-art benchmarks on multiple dimensions including point estimates, uncertainty, computational complexity, and cost.
Enhancing Knowledge	Contribute new learnings and insights to the forecasting community.
Open Sharing	Encourage open participation and sharing of information, e.g. through code-only competitions requiring model sharing, limit model tweaks, and/or enforce computation time limits.
Multiple Challenges (forecasting athlons)	Incorporate multiple challenges into a forecasting competition to enable comprehensive evaluation across multiple skill areas or application areas.

M6 incorporated the majority of our recommendations

Spryos Makridakis, C. Fry, F. Petropoulos, E. Spiliotis, “The Future of Forecasting Competitions: Design Attributes and Principles”, INFORMS Journal on Data Science, Vol. 1, No. 1, April–June 2022, pp. 96–113.

Theme	What we recommended	Assessment (M6 “report card”)
Replicability	Submission of source code or executable that can be rerun on the same or new inputs / datasets.	F - Not achieved - participants only submit model outputs; forecasts cannot be replicated.
Representativeness	Represents real-world challenge(s); similar to problems faced by other forecasters; broad coverage across time series feature space and range of possible external effects.	B+ - Stock price prediction and investment decisions reflect a highly realistic real-world challenge; unlimited external features; rebalancing limited to 1x/month; forecasting not critical to investment challenge.
Robust Evaluation	Rolling origin evaluation, covering different times of year; multiple evaluation rounds in a live setup.	A+ - 4 quarterly evaluations covering a full year; live setup, with each quarter requiring 3 rolling submissions.
Measuring Decision Impact	Measure utility of forecasts directly through their impact on decisions, reflected in monetary terms when possible.	A+ - Information ratio metric links contest result directly to decision impact (risk-adjusted financial return)
Showcase Forecast Value-Added (FVA)	Demonstrate value of winning solutions against state-of-the-art benchmarks on multiple dimensions including point estimates, uncertainty, computational complexity, and cost.	B - Benchmarks used were *equal probabilities* and *equal long positions*. No standard forecasting model benchmarks were reported or used for FVA.
Enhancing Knowledge	Contribute new learnings and insights to the forecasting community.	B+ - Several innovations (live contest with rolling evaluation; decision impact metric). Non-standard forecasting contest metric.
Open Sharing	Encourage open participation and sharing of information, e.g. through code-only competitions requiring model sharing, limit model tweaks, and/or enforce computation time limits.	C- - Submission questionnaire allowed teams to voluntarily share data sources used and methodology; limited discussion on forum
Multiple Challenges (forecasting athlons)	Incorporate multiple challenges into a forecasting competition to enable comprehensive evaluation across multiple skill areas or application areas.	B - Duathlon format incorporated two challenges (forecasting and investment).

Major wins of the M6 competition

1. Measures a real business outcome in addition to measuring forecasting quality
2. Answers questions directly relevant to the finance industry
3. Great example of a rolling evaluation setup (more time needed to fully separate luck vs skill?)
4. Live forecasting challenge
5. Comparison to benchmarks
6. Contest results showed the value of linking decisions with uncertainty modeled in forecasts --> risk model / diversification approach
7. Reminder of the value of flexibility - strategies that took advantage of short positions outperformed strategies limited to long positions only (e.g. the investment benchmark)

Some learnings (my opinions)

1. Low signal to noise ratio shifts challenge from prediction to risk management
   - *Difficult to confidently project trend from history*
   - *High variance estimates require portfolio approach (and a bit of luck)*
2. Importance of aligning contest structure to real-world challenge
   - *Monthly portfolio rebalancing is unrealistic - limits flexibility; doesn’t consider transaction costs*
   - *Difficult to draw conclusions with single year of data*
3. Importance of consistency between forecast and outcome metrics
   - *Strong performance on RPS score did not correlate with strong outcomes*

Performance on RPS score did not correlate with outcomes*

		RPS	IR
M6 submissions	Scores directly from competition leaderboard**.	**0.156 ~ 0.32**	**-29 ~ 33**
Benchmark	Equal probability in each quintile. Equal weight among the asset (1%).	**0.16**	**-2.7**

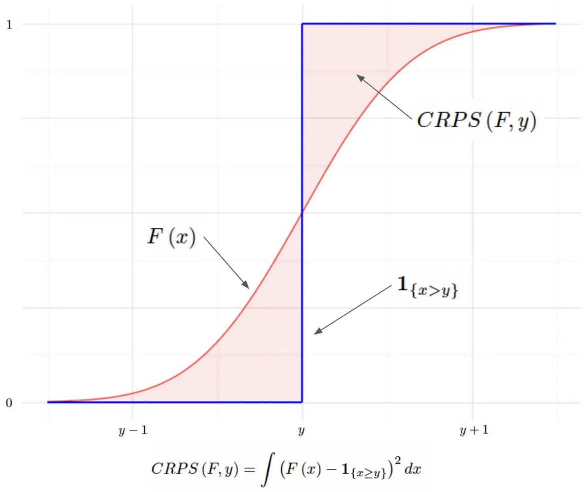
Benchmark (equal probability) strategy achieved a high RPS score, with the highest-performing team only beating it by 2.2%...

...but the benchmark **returned a negative IR.**

* Reported correlation between RPS and IR was 0.04 overall, and 0.12 for the top 5% highest-performing teams. Maximum correlation observed was 0.7 for top 20% highest-performing teams.
Source: Makridakis et al, "The M6 forecasting competition: Bridging the gap between forecasting and investment decisions", Preprint submitted to International Journal of Forecasting
** Source: m6competition.com/Leaderboard. Used with permission.

CRPS a better metric?

		RPS	IR	CRPS
M6 submissions	Scores directly from competition leaderboard*.	**0.156 ~ 0.32**	**-29 ~ 33**	
Benchmark 1	Equal probability in each quintile. Equal weight among the asset (1%). (reshuffled last month returns)	**0.16**	**-2.7**	**0.106**
Benchmark 2	Equal probability in each quintile. Equal weight among the asset (1%). (random sample, Uniform[-1.0, 1.0])	**0.16**	**-2.7**	**0.504**



1
0
$CRPS(F,y)$
$F(x)$
$\mathbf{1}_{\{x>y\}}$
$y-1$
y
$y+1$
$CRPS(F,y) = \int \left(F(x) - \mathbf{1}_{\{x\geq y\}}\right)^2 dx$

Benchmark (equal probability) strategy achieved a poor CRPS score based on our assessment

Multiple forecasts can generate the Benchmark strategy; most have poor CRPS scores.

* Source: m6competition.com/Leaderboard. Used with permission.

Optimizing for asset value prediction accuracy drove higher returns in our toy example

		RPS	IR	CRPS
M6 submissions	Scores directly from competition leaderboard*.	**0.156 ~ 0.32**	**-29 ~ 33**	
Benchmark 1	Equal probability in each quintile. Equal weight among the asset (1%). (reshuffled last month returns)	**0.16**	**-2.7**	**0.106**
Benchmark 2	Equal probability in each quintile. Equal weight among the asset (1%). (random sample, Uniform[-1.0, 1.0])	**0.16**	**-2.7**	**0.504**
Toy model univariate forecast**	Quintile = quintile distribution based on range forecasts for each timeseries. Portfolio weight ~ P50 forecast	**0.186**	**6.9**	**0.08**

**** Recency-weighted linear regression model (half-life = 30 days), with empirically-generated prediction intervals.**

* Source: m6competition.com/Leaderboard.   Used with permission.

Adjusting for risk further improved performance

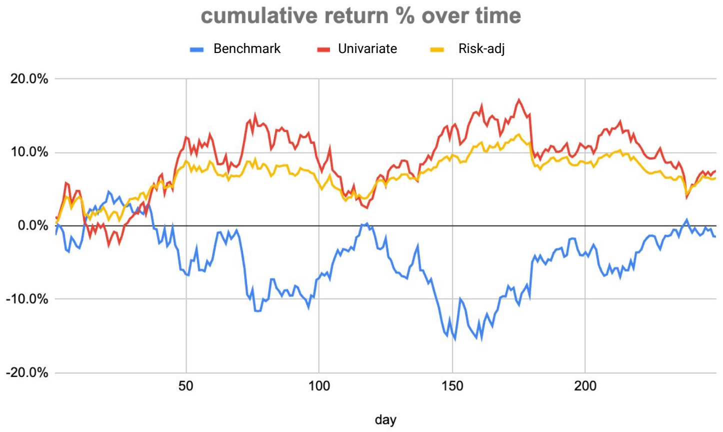
		RPS	IR	CRPS
M6 submissions	Scores directly from competition leaderboard*.	**0.156 ~ 0.32**	**-29 ~ 33**	
Benchmark 1	Equal probability in each quintile. Equal weight among the asset (1%). (reshuffled last month returns)	**0.16**	**-2.7**	**0.106**
Benchmark 2	Equal probability in each quintile. Equal weight among the asset (1%). (random sample, Uniform[-1.0, 1.0])	**0.16**	**-2.7**	**0.504**
Toy model univariate forecast**	Quintile = quintile distribution based on range forecasts for each timeseries. Portfolio weight ~ P50 forecast	**0.186**	**6.9**	**0.08**
Univariate forecast with risk-adjusted investment**	Quintile = quintile distribution based on range forecasts for each timeseries. Portfolio weight ~ P50 forecast / sigma*0.5	**0.186**	**11.5**	**0.08**

* Source: m6competition.com/Leaderboard. Used with permission.

** Recency-weighted linear regression model (half-life = 30 days), with empirically-generated prediction intervals.

Adjusting for risk further improved performance

		RPS	IR	CRPS
M6 submissions	Scores directly from competition leaderboard*.	**0.156 ~ 0.32**	**-29 ~ 33**	
Benchmark 1	Equal probability in each quintile. Equal weight among the asset (1%). (reshuffled last month returns)	**0.16**	**-2.7**	**0.106**
Benchmark 2	Equal probability in each quintile. Equal weight among the asset (1%). (random sample from Uniform[-1.0, 1.0])	**0.16**	**-2.7**	**0.504**
Toy model univariate forecast**	Quintile = quintile distribution based on range forecasts for each timeseries. Portfolio weight ~ P50 forecast	**0.186**	**6.9**	**0.08**
Univariate forecast w/ risk-adjusted investment**	Quintile = quintile distribution based on range forecasts for each timeseries. Portfolio weight ~ P50 forecast / sigma*0.5	**0.186**	**11.5**	**0.08**

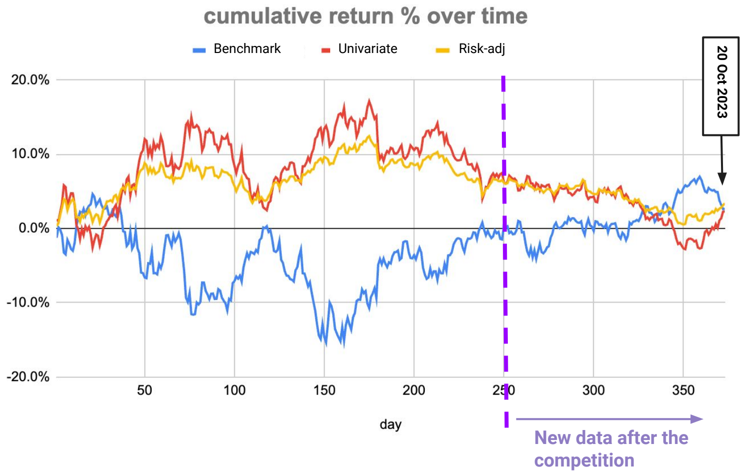


* Source: m6competition.com/Leaderboard. Used with permission.

** Recency-weighted linear regression model (half-life = 30 days), with empirically-generated prediction intervals.

Playing out the game a bit longer...

		RPS	IR	CRPS
M6 submissions	Scores directly from competition leaderboard*.	**0.156 ~ 0.32**	**-29 ~ 33**	
Benchmark 1	Equal probability in each quintile. Equal weight among the asset (1%). (reshuffled last month returns)	**0.16**	**-2.7**	**0.106**
Benchmark 2	Equal probability in each quintile. Equal weight among the asset (1%). (random sample from Uniform[-1.0, 1.0])	**0.16**	**-2.7**	**0.504**
Toy model univariate forecast**	Quintile = quintile distribution based on range forecasts for each timeseries. Portfolio weight ~ P50 forecast	**0.186**	**6.9**	**0.08**
Univariate forecast w/ risk-adjusted investment**	Quintile = quintile distribution based on range forecasts for each timeseries. Portfolio weight ~ P50 forecast / sigma*0.5	**0.186**	**11.5**	**0.08**



20 Oct 2023

New data after the competition

* Source: m6competition.com/Leaderboard.   Used with permission.

** Recency-weighted linear regression model (half-life = 30 days), with empirically-generated prediction intervals.