White Paper

Enhancing forecast availability using machine-learning-driven earnings and revenue models

StarMine SmartForecast Model

Author:

Tarun Sanghi, PhD, CFA, Head of Quant Equity Research, StarMine

LSEG DATA & ANALYTICS

Contents

Abstract

StarMine SmartForecast Model delivers forward-looking one-year and second-year earnings and revenue forecasts for 50,000~ publicly traded companies – including those with limited or no analyst coverage. Forward-one-year annual estimates are further disaggregated into quarterly forecasts using StarMine's proprietary seasonality forecasting algorithm. This whitepaper provides a detailed description of the construction of the StarMine SmartForecast Model as well as historical performance analysis.

Introduction

Forward looking estimates, especially the earnings and the revenue estimates, are extremely useful to investors to gauge a company's on-going performance and assess future prospects. Further, earnings forecasts also serve as key inputs to cost of capital calculations. For most investors, sell-side analysts remain the primary source of these forward-looking estimates, which are collected and aggregated by data providers such as LSEG I/B/E/S and subsequently made available to practitioners.

However, there are two fundamental challenges with analyst forecasts. Firstly, prior research shows that analyst forecasts tend to exhibit significant optimism bias [1]. Secondly, analyst forecasts are not available for the entire universe of publicly traded companies, as analysts tend to follow larger firms with high level of institutional investment. Based on the availability of forward one-year earnings per share (EPS) estimates in I/B/E/S, approximately half of the publicly traded companies worldwide receive no sell-side coverage and therefore lack forward looking earnings estimates available for them.

StarMine developed proprietary SmartEstimates and SmartGrowth algorithms [2-4] to mitigate analyst biases. SmartEstimates algorithm leverages the finding that analyst accuracy persists when measured properly. SmartEstimates improves upon consensus estimates by excluding stale estimates and suspected data errors, and weighting the remaining estimates based on the historical accuracy of each analyst and the recency of each estimate. SmartGrowth algorithm is designed to identify and remove systematic analyst biases from SmartEstimates.

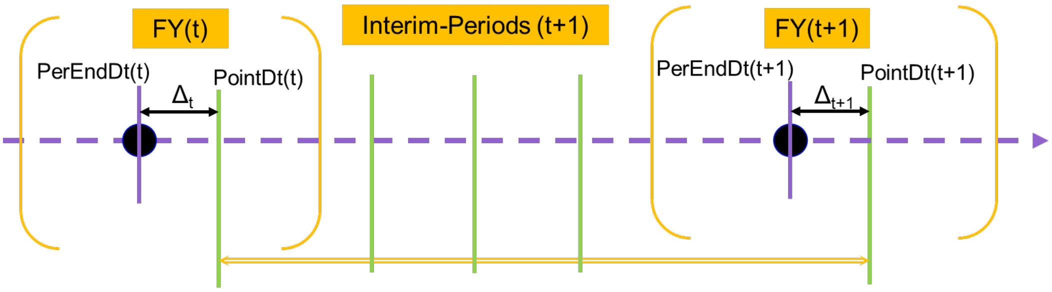
To address the coverage gap, StarMine has developed a SmartForecast Model, a data-driven, machine-learning framework, that produces a forward-one-year (FY1) and forward-second-year (FY2) earnings and revenue forecasts for ~50,000 publicly traded companies.

Model overview

StarMine SmartForecast Model operates in two stages:

1. **Stage I (Initial forecast):** In Stage I, the model uses the fundamental data, market data, and StarMine proprietary analytics to provide an initial FY1 annual forecasts. Stage I forecasts are provided as soon as the most recent fiscal-year-end fundamental data becomes available. Once FY1 forecasts are obtained, they are used to calculate a 2-year compound annual growth rate (CAGR) and the FY2 forecasts. Further, the FY1 annual forecasts are disaggregated into quarterly forecasts using StarMine's proprietary seasonality forecasting algorithm.
2. **Stage II (Revisions):** In Stage II, the model incorporates interim-fiscal-period fundamental data (as soon as it becomes available), updated market data, and StarMine proprietary analytics to revise the initial-forecast/prior-revisions. For companies that report interim-fiscal-period data with quarterly frequency (4Q), three revisions are performed (after first, second and third reporting), and are referred to as Stage II-F[1-3]Q forecasts in the whitepaper. For companies that report the data with semi-annual frequency (2S), one revision is performed after the initial forecast, and it is referred to as Stage II-F1S forecast in the whitepaper.

The two-stage framework is designed to emulate the iterative process used by sell-side analysts to revise their estimates. Figure 1 below shows the two-stage forecasting framework for a 4Q company.



Two Stage Model

Stage I

— Initial Forecast: FY(t+1) forecast as soon as FY(t) data become available

Stage II

— Revision: Revise FY(t+1) forecast as soon as FY(t+1)'s interim-periods data becomes available

- **FY(t):** Most recent fiscal year (YYYY); **FY(t+1):** Next fiscal year
- **Interim-Periods(t+1;s):** Interim periods of fiscal year FY(t+1)
- **PerEndDt(t):** Date when fiscal year FY(t) is considered to have ended
- **PointDt(t):** Date when fiscal year FY(t) data is made available to public
- **Δ_t:** Difference between PerEndDt(t) and PointDt(t)

Figure 1: Two-stage forecasting framework of StarMine SmartForecast Model.

We use Random Forest algorithm [5], which is a tree-based machine-learning approach, to build non-linear cross-sectional forecasting models. The main advantage of building cross-sectional model is that it can be built without imposing any firm-specific data limitation, i.e., a firm does not need to have existed in the entire estimation period to train the model. In Random Forest algorithm, given a set of data (predictors and response variable), it generates a forest of regression trees, each of which is a weak learner build using subset of the input data. To create the output, it takes the average prediction over all the trees in the forest to make a final prediction. Mathematically, the prediction f(.) for a random forest is thus given by:

$$f(.) = \frac{1}{M}\sum_{m=1}^{M} T_m(X_i, \Theta_m)$$

where M is the number of trees, Tm is an individual tree, X_i are the predictors, and Q_m are the parameters that define that tree. Random forest algorithm has three main hyperparameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. From these, a random forest model can be trained and used to solve for regression problems. For the StarMine SmartForecast model, a hyperparameter search grid was used to train the Random Forest models. The optimal hyperparameter set (or the best model) was obtained based on the minimisation of the “mean-squared-error” metric. A six-year lookback sliding window approach as shown in Figure 2 (the left chart), was used to train both Stage I and Stage II models.

The right-hand-side chart on Figure 2 shows the different datasets that are used to build the models. We took a hypothesis driven approach for the predictor variable selection. We surveyed the academic literature on earnings and revenue forecasting and combined that knowledge with our equity expertise, to construct a pool of predictors. Each selected predictor had a sound economic rationale for why it should be a predictor of future earnings or revenue. Then those selected predictors were tested for their collective value using the Boruta algorithm [6], which is a Random Forest based “all relevant feature” selection algorithm, to obtain an optimal subset that can be used to train random forest models. In the Boruta algorithm, shadow features are generated, which are the randomised permutation of the original features, and a random forest is built using both the original and shadow features. The features that have importance higher than the shadow features are considered important.

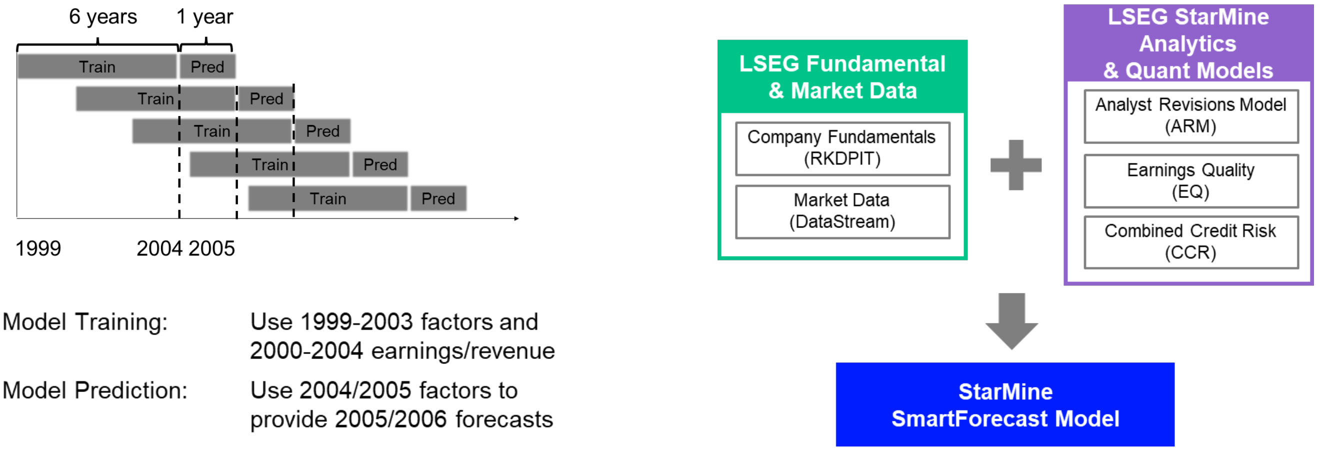

6 years
1 year
Train
Pred
Train
Pred
Train
Pred
Train
Pred
Train
Pred
1999
2004 2005
Model Training:
Use 1999-2003 factors and 2000-2004 earnings/revenue
Model Prediction:
Use 2004/2005 factors to provide 2005/2006 forecasts
LSEG Fundamental & Market Data
Company Fundamentals (RKDPIT)
Market Data (DataStream)
LSEG StarMine Analytics & Quant Models
Analyst Revisions Model (ARM)
Earnings Quality (EQ)
Combined Credit Risk (CCR)
StarMine SmartForecast Model

Figure 2: Construction of StarMine SmartForecast Model. The left chart shows the six-year lookback sliding window approach used to train both Stage I and Stage II Random Forest models. The right chart shows the different datasets that are used to build the model.

Performance measures and benchmark

To quantify the value-add from our machine-learning based non-linear models, we used the Random-Walk (RW) models in earnings and revenue growth ($\Delta x(t+1) = \Delta x(t)$; Δ represents the first order change in x) as the benchmark models. RW models are parameter free models that are used both by practitioners, and in academia, to generate forward-looking estimates, especially for companies that are not covered by analysts.

To demonstrate the value-add from StarMine SmartForecast Model, we use the direction of the earnings change as the primary evaluation metric. This metric compares the forecasted direction of change with the actual realised direction and is widely considered as the most economically meaningful and actionable signal in portfolio construction [7]. If the subsequently reported earnings are higher (or lower) than the last reported earnings, and the model forecast also predicts an increase (or decrease) relative to the last reported earnings, the forecasts are considered directionally correct. We wish to point out that several other quantitative measures such as the mean-absolute-error were utilised to objectively assess the quality of the forecast. Further, goodness-of-fit tests were conducted by regressing model forecasts against reported actuals to detect and correct for any systematic biases in the model outputs.

Figure 3 below shows the mean directional correctness (hit rate) of the FY1 earnings forecasts over 2005-2024 period by year (left charts) and TRBC economic sectors (right charts) in different regions. Most companies in North America, Developed Asia and Emerging Markets report interim-fiscal-period data with quarterly frequency and are thus represented in the 4Q models. Most European companies report interim-fiscal-period data with semi-annual frequency and are represented in the 2S models. Markets such as

Taiwan, Singapore, Hong Kong, Australia, and Japan that follow either a mixed reporting structure (might require quarterly reporting for riskier companies, semi-annual otherwise) or switched reporting structure (switched from semi-annual to quarterly or vice-versa) are represented in both the models.

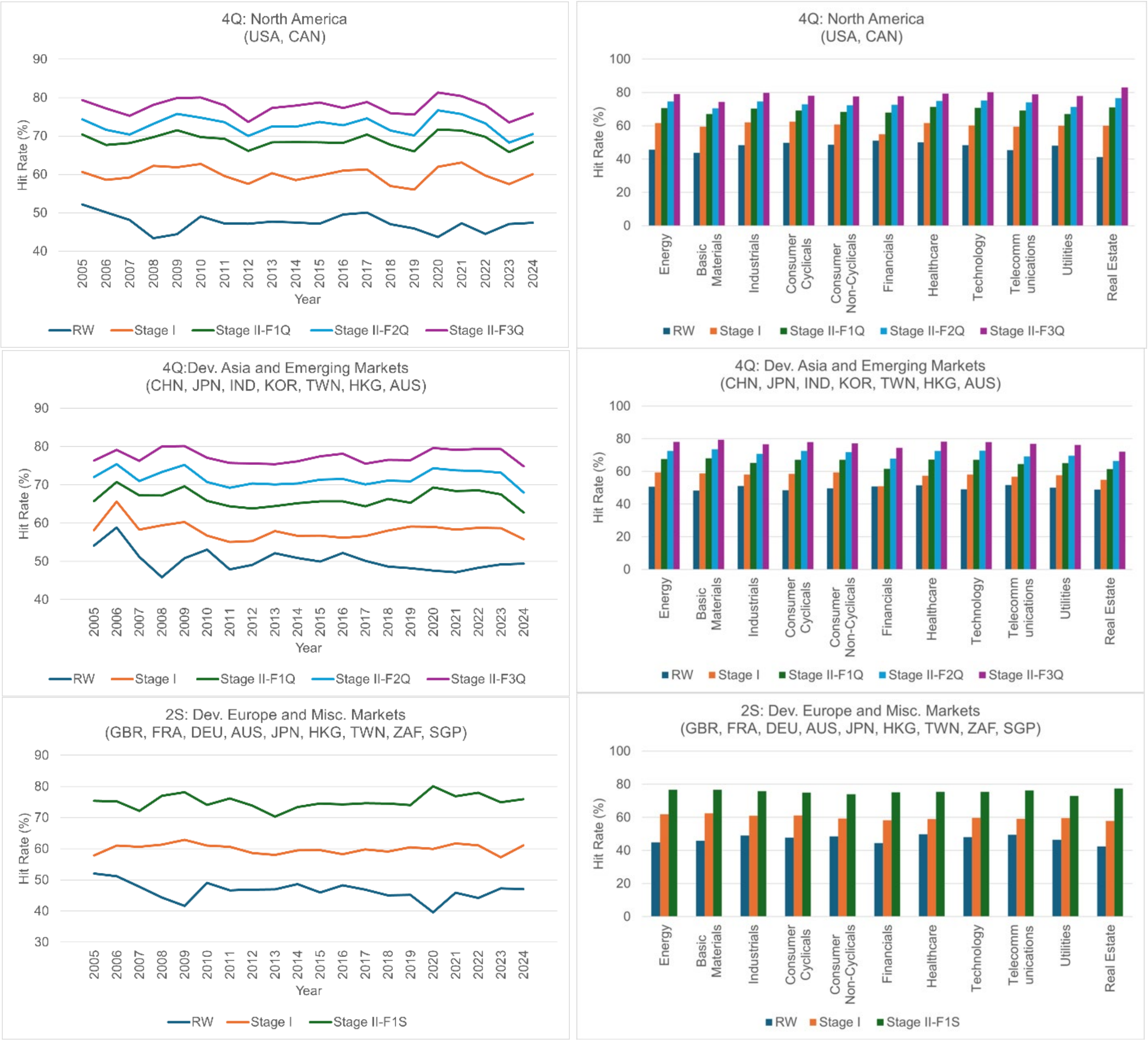


Figure 3: Directional correctness (Hit Rate) of the FY1 earnings forecasts over the 2005-2024 period by year (left charts) and economic sectors (right charts) in different regions.

It can be observed that the hit rate of Stage I forecasts is ~60% (vs ~50% for RW model) over the entire 2005-2024 period in all the regions (left charts). We want to highlight that these hit rates were calculated in a true point-in-time manner. Stage I hit rate for the year 2005 represents the mean directional correctness of 2006 annual forecasts (provided as soon as 2005 fiscal-year-end data becomes available for a company) as obtained from the models trained on 1999-2004 data. Similarly, Stage II-F[1-3]Q and Stage II-F1S hit rates for the year 2005 represent the mean directional correctness of the revisions (of 2006 annual forecasts) performed after the 2006 interim-period-data becomes available for 4Q and 2S companies, respectively. It can be observed that the hit rate increases with each subsequent revision, demonstrating the value-add Stage II revisions provide in improving the directional accuracy of the model forecasts. Further, it can be seen from the charts on the right that both Stage I and Stage II hit rates are very similar for all the 11 TRBC economic sectors, demonstrating the consistency in performance across different sectors.

In Figure 4 we report and compare the mean directional correctness of FY1 earnings forecast by year for analyst-covered (left charts) and no-analyst-coverage (right charts) companies in different regions over the 2005-2024 period.

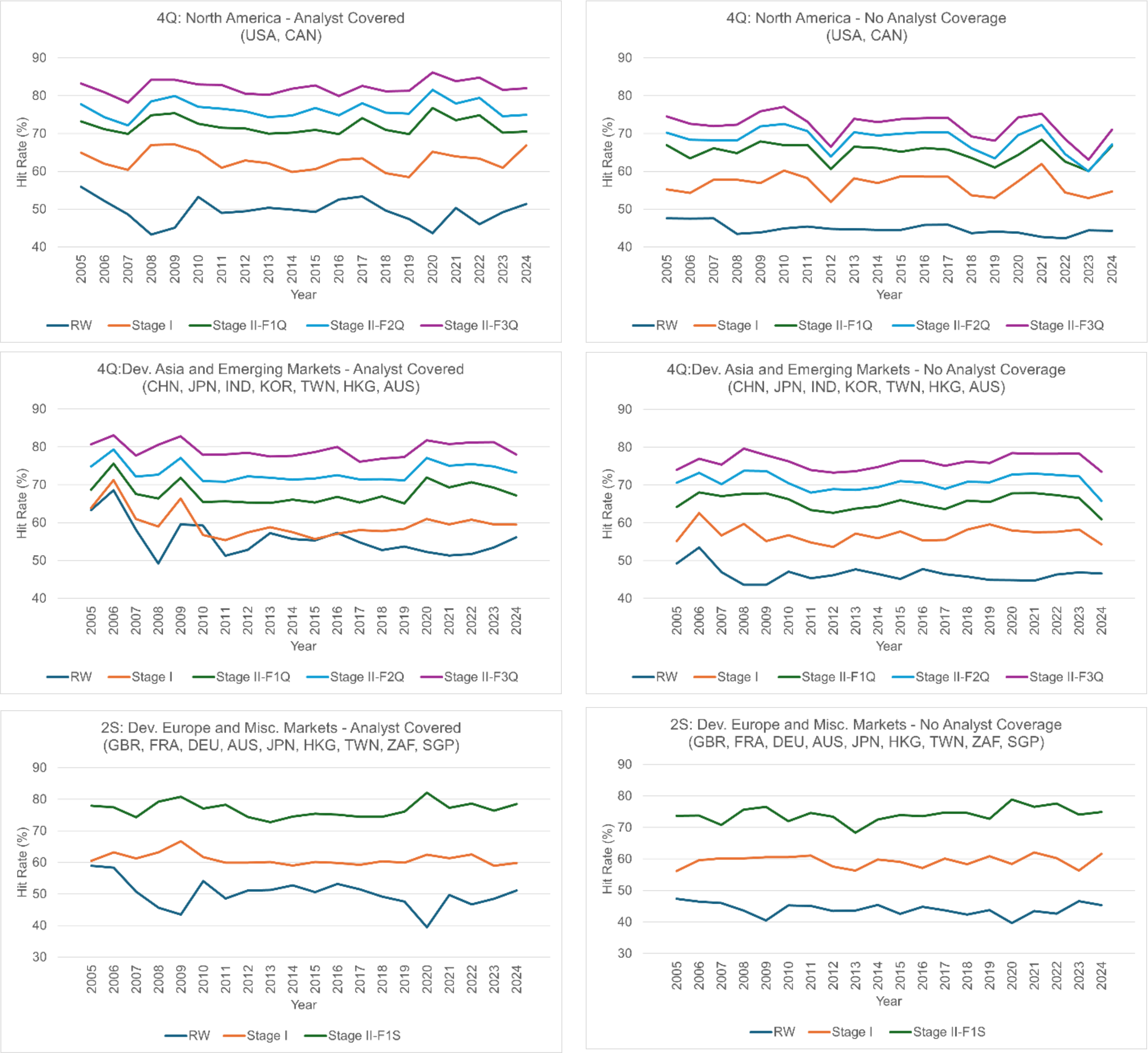


Figure 4: Directional correctness (Hit Rate) of the FY1 earnings forecasts over the 2005-2024 period by year for analyst-covered (left charts) and no-analyst-coverage (right charts) companies in different regions.

The following observations were made from the charts above:

- For no-analyst-coverage companies, the directional correctness of the RW model-based forecasts (Δearnings(t+1) = Δearnings(t)) was less than 50% (i.e., worse than a random guess) in all the regions. It is not surprising because earnings are the bottom-line item on the income statement and typically subjected to significant accounting noise. Further, earnings are known to exhibit mean-reversion.
- For both analyst-covered and no-analyst-coverage companies, StarMine SmartForecast Model hit rates for Stage I forecasts were consistently greater than 55% in all regions. Further, the hit rate increased with each subsequent revision (Stage II forecasts) demonstrating the value-add of the two-stage framework over the static RW model based forecasts.

- Model outputs were found to perform better for analyst-covered companies compared to no-analyst-coverage companies. This is not surprising as the no-analyst-coverage companies typically operate in weak information environments and tend to exhibit higher volatility in earnings than those covered by analysts. We wish to point out that this is an “after-the-fact” observation, and no special treatment was given either to analyst-covered or no-analyst-coverage companies while training the model.

We now shift our attention on the revenue forecasts. Although expected earnings attract the most attention, expected revenues also play an important role in the investor’s valuation process and investment decisions. Figure 5 reports and compares the directional correctness of FY1 revenue forecasts by year for analyst-covered (left charts) and no-analyst-coverage (right charts) 4Q and 2S companies.

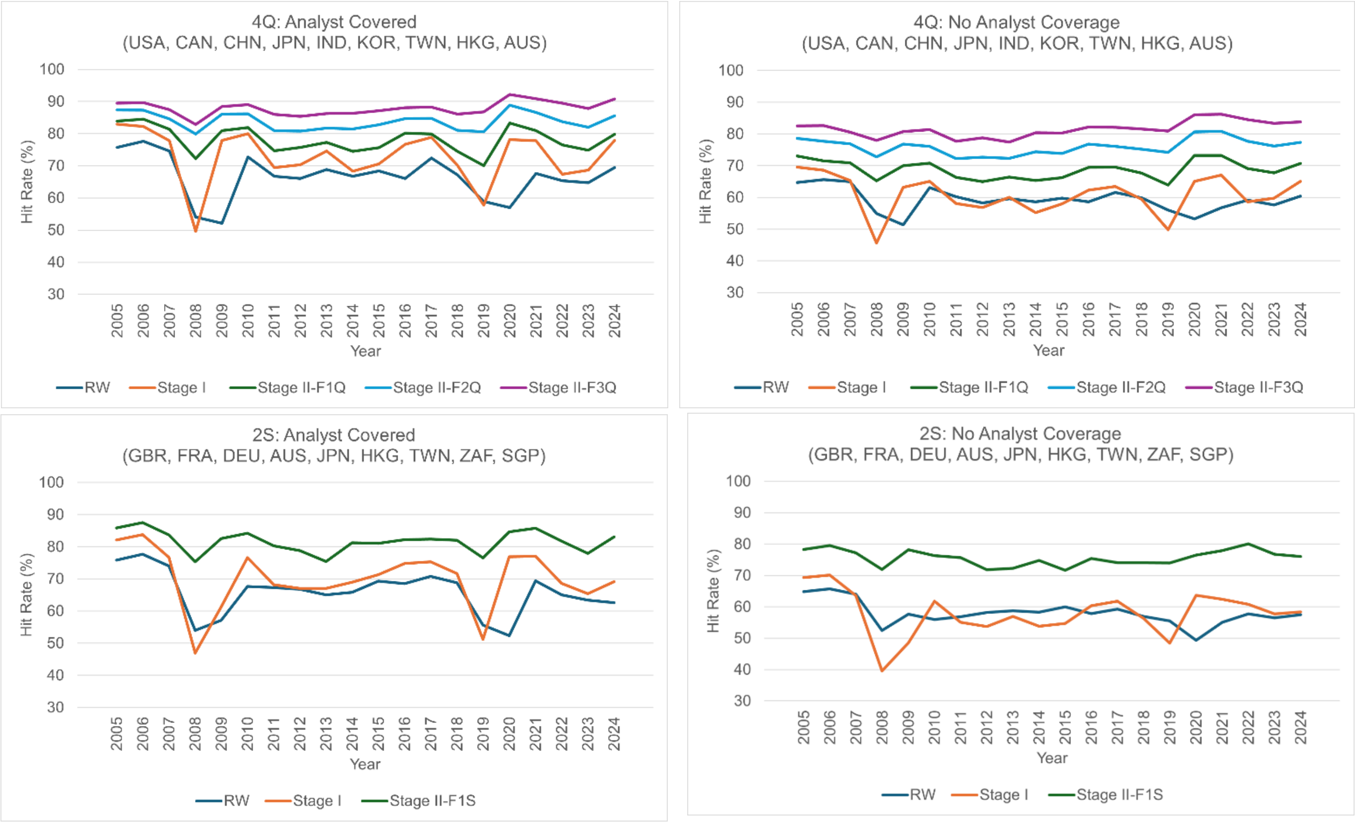


Figure 5: Directional correctness (Hit Rate) of the FY1 revenue forecasts over the 2005-2024 period by year for analyst-covered (left charts) and no-analyst-coverage (right charts) 4Q and 2S companies.

The following observations were made from the charts above:

- Unlike earnings, for revenue, the hit rate of the RW model-based forecasts (Δrevenue(t+1) = Δrevenue (t)) was more than 50% in most years. It is again not surprising as revenue being the top-line item on the income statement, tends to be more stable and persistent over time, making the RW model a hard benchmark to beat.
- For both analyst-covered and no-analyst-coverage companies, the hit rate of Stage I forecasts were similar to those obtained from the RW model-based forecasts. However, the hit rates significantly improved during the Stage II revisions, again demonstrating the value-add of our two-stage framework over the static RW model-based forecasts.

Quant use-case

To demonstrate the utility of model forecasts in designing quantitative strategies, we used the FY1 earnings forecasts to build a simple expected earnings-growth signal (calculated as FY1 earnings forecast minus current reported earnings, scaled by current market value) and used it to construct long and short portfolios. Figure 6 shows the cumulative return of top (Top) and bottom (Bot) decile portfolios built using the expected earnings-growth signal on a US-equity universe of top 3000 companies by market capitalisation. The portfolios are rebalanced monthly, and their performance is compared against an equal-weighted market portfolio (EqWgtRet – Top 3000) over January 2005 to December 2025 period. Performance statistics are reported in Table 1.

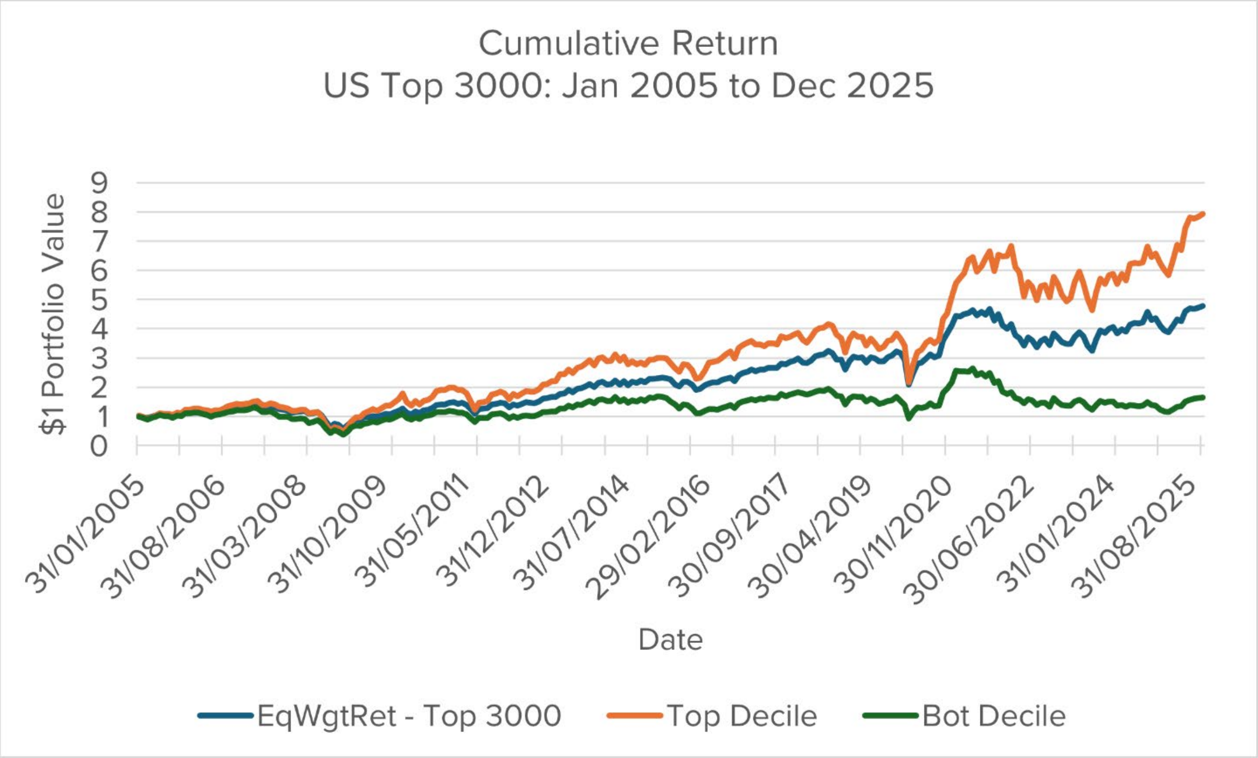
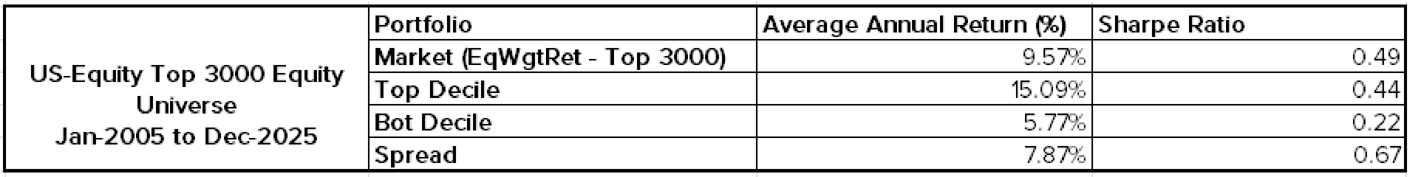


Figure 6: Cumulative return of monthly rebalanced top and bottom decile portfolios constructed using expected earnings-growth signal on US Top 3000 equity universe.



US-Equity Top 3000 Equity Universe Jan-2005 to Dec-2025	Portfolio	Average Annual Return (%)	Sharpe Ratio
	Market (EqWgtRet - Top 3000)	9.57%	0.49
	Top Decile	15.09%	0.44
	Bot Decile	5.77%	0.22
	Spread	7.87%	0.67

Table 1: Performance statistics of monthly rebalanced top and bottom decile portfolios constructed using expected earnings-growth signal on US Top 3000 equity universe.

Conclusions

StarMine SmartForecast Model is a powerful, non-linear cross-sectional framework, that can be used to generate forward-one-year and forward-second-year earnings and revenue forecasts for publicly traded companies. As of December 2025, StarMine SmartForecast Model provides FY1 and FY2 earnings and revenue forecasts for ~50k+ publicly listed companies, including 25k+ companies that have zero sell-side coverage. StarMine SmartForecast Model is delivered via secure SFTP – Secured File Transfer Protocol, for seamless integration into existing workflows. Historical point-in-time forecasts are also available for those who wish to backtest the forecasts.

Acknowledgement

The author gratefully acknowledges the contributions of Nelson Chin from the StarMine Research team, as well as Gain Davis and Maryia Vauchkova from the StarMine Product and Operations team, for their assistance and support in running an early-access-program for this research.

References

[1] Birman, S., Malinak, S. and Bonne, G., 2009. StarMine Intrinsic Valuation Model (IV): Overview and global performance, StarMine Whitepaper.

[2] Stauth, J. and Bonne, G., 2009, SmartEstimates and the Predicted Surprise: Construction and Accuracy, StarMine Whitepaper.

[3] Vieira, M., Genin, H. and Birman, S., 2018, An Update on the Performance of SmartEstimate and Predicted Surprise, StarMine Whitepaper.

[4] Vieira, M., Genin, H. and Birman, S., 2019, A Deeper Analysis of the Performance of SmartEstimates and Predicted Surprise, StarMine Whitepaper.

[5] Breiman, L. 2001, Random Forests. Machine Learning 45, 5–32.

[6] Kursa, M. B., & Rudnicki, W. R., 2010, Feature Selection with the Boruta Package, Journal of Statistical Software, 36(11), 1–13.

[7] Chen, X., Cho, Y. H. T., Dou, Y., & Lev, B., 2022, Predicting Future Earnings Changes Using Machine Learning and Detailed Financial Data. Journal of Accounting Research, 60(2), 467–515

Disclaimer

The content of this publication is provided by London Stock Exchange Group plc, its applicable group undertakings and/or its affiliates or licensors (the “**LSE Group**” or “**We**”) exclusively.

The content of this publication is for informational purposes only. All information and data contained in this publication is obtained by LSE Group from sources believed by it to be accurate and reliable. Because of the possibility of human and mechanical error as well as other factors, however, such information and data are provided "as is" without warranty of any kind. You understand and agree that this publication does not, and does not seek to, constitute advice of any nature. You may not rely upon the content of this document under any circumstances and should seek your own independent legal, tax or investment advice or opinion regarding the suitability, value or profitability of any particular security, portfolio or investment strategy. Neither We nor our affiliates shall be liable for any errors, inaccuracies or delays in the publication or any other content, or for any actions taken by you in reliance thereon. You expressly agree that your use of the publication and its content is at your sole risk.

Neither We nor our affiliates guarantee the accuracy of or endorse the views or opinions given by any third party content provider, advertiser, sponsor or other user. We may link to, reference, or promote websites, applications and/or services from third parties. You agree that We are not responsible for, and do not control such non-LSE Group websites, applications or services.

To the fullest extent permitted by applicable law, LSE Group, expressly disclaims any representation or warranties, express or implied, including, without limitation, any representations or warranties of performance, merchantability, fitness for a particular purpose, accuracy, completeness, reliability and non-infringement. LSE Group, its subsidiaries, its affiliates and their respective shareholders, directors, officers employees, agents, advertisers, content providers and licensors (collectively referred to as the “LSE Group Parties”) disclaim all responsibility for any loss, liability or damage of any kind resulting from or related to access, use or the unavailability of the publication (or any part of it); and none of the LSE Group Parties will be liable (jointly or severally) to you for any direct, indirect, consequential, special, incidental, punitive or exemplary damages, howsoever arising, even if any member of the LSE Group Parties are advised in advance of the possibility of such damages or could have foreseen any such damages arising or resulting from the use of, or inability to use, the information contained in the publication. For the avoidance of doubt, the LSE Group Parties shall have no liability for any losses, claims, demands, actions, proceedings, damages, costs or expenses arising out of, or in any way connected with, the information contained in this document.

No part of this information may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior written permission of the applicable member of LSEG. Use and distribution of LSEG data requires a licence from the relevant member of LSEG, and/or their respective licensors. Republication or redistribution of LSE Group content, including by framing or similar means, is prohibited without the prior written consent of LSEG.

Past performance is not indicative of future results. Charts and graphs are provided for illustrative purposes only. This document may contain forward-looking assessments, which are based on assumptions regarding future conditions that may prove to be inaccurate. Such forward-looking assessments are subject to risks and uncertainties, and actual outcomes may differ materially.

LSE Group is the owner of various intellectual property rights ("**IPR**”), including but not limited to, numerous trademarks that are used to identify, advertise, and promote LSE Group products, services and activities. Nothing contained herein should be construed as granting any licence or right to use any of the trademarks or any other LSE Group IPR for any purpose whatsoever without the written permission or applicable licence terms.

lseg.com

LSEG DATA & ANALYTICS