ARIA

APPLIED RESEARCH IN ACTION

Enhancing Forecasting Performance of LightGBM through Feature Screening

A systematic process to mitigate the impact of high feature dimensions while retaining informative features to produce forecasts with better out-of-sample ranking accuracy and volatility

Dixin Mou

Sebastian Jaimungal
ACADEMIC SUPERVISOR

Omid Namvar
INDUSTRY SUPERVISOR

Target	SMAPE: All Features	SMAPE: Current Model	SMAPE: Feature Screening	% Difference for Feature Screening vs. All Features	% Difference for Feature Screening vs. Current Model
Target 1	1.394	1.391	1.390	-0.2%	-0.1%
Target 2	1.414	1.410	1.404	-0.7%	-0.4%
Target 3	1.427	1.416	1.414	-0.9%	-0.1%
Target 4	1.440	1.434	1.430	-0.7%	-0.3%
Target 5	1.423	1.419	1.411	-0.9%	-0.6%
average	1.420	1.414	1.410	-0.7%	-0.3%

Target	Standard Deviation of SMAPE: All Features	Standard Deviation of SMAPE: Current Model	Standard Deviation of SMAPE: Feature Screening	% Difference for Feature Screening vs. All Features	% Difference for Feature Screening vs. Current Model
Target 1	0.600	0.602	0.602	0.2%	0.0%
Target 2	0.603	0.606	0.605	0.4%	0.1%
Target 3	0.600	0.608	0.606	1.1%	0.4%
Target 4	0.594	0.599	0.595	0.0%	0.7%
Target 5	0.596	0.599	0.598	0.4%	0.0%
average	0.599	0.603	0.601	0.4%	0.2%

Target	Ranking Accuracy: All Features	Ranking Accuracy: Current Model	Ranking Accuracy: Feature Screening	% Difference for Feature Screening vs. All Features	% Difference for Feature Screening vs. Current Model
Target 1	0.357	0.357	0.357	0.1%	0.1%
Target 2	0.351	0.350	0.352	0.4%	0.6%
Target 3	0.326	0.323	0.332	1.6%	2.8%
Target 4	0.307	0.300	0.317	3.2%	5.8%
Target 5	0.343	0.339	0.349	1.7%	2.8%
average	0.337	0.334	0.342	1.4%	2.4%

Target	Standard Deviation of Ranking Accuracy: All Features	Standard Deviation of Ranking Accuracy: Current Model	Standard Deviation of Ranking Accuracy: Feature Screening	% Difference for Feature Screening vs. All Features	% Difference for Feature Screening vs. Current Model
Target 1	0.049	0.048	0.048	-2.6%	-0.9%
Target 2	0.072	0.074	0.075	3.4%	0.9%
Target 3	0.104	0.100	0.101	-2.7%	0.9%
Target 4	0.114	0.125	0.109	-4.4%	-12.8%
Target 5	0.118	0.113	0.111	-6.0%	-1.4%
average	0.091	0.092	0.089	-2.5%	-2.7%

PROJECT SUMMARY

This report provides an overview of an advancement that have been made to improve the performance of a forecasting engine. The engine is a Machine Learning model that utilizes Gradient Boosting Decision Tree (GBDT) to forecast company fundamentals for investment strategies. The primary objective of the advancement is to enhance the forecasting accuracy of the engine while maintaining consistent forecasts over time. However, the current model faces a challenge: a dataset with a high feature space and relatively a small number of data instances, which leads to the curse of dimensionality. To address this challenge, a systematic target-specific feature screening process has been proposed to reduce the number of features used in the model effectively. The feature screening process provides a viable solution to mitigate the impact of high feature dimensions while still retaining informative features to produce forecasts with better out-of-sample ranking accuracy and volatility. This report applies two updates in the feature screening process to enable target independent feature screening and Macro feature screening. And a detailed experiment is conducted to assess the efficacy of the feature screening process with various parameters. Empirical data shows strong evidence that the recommended model achieves superior ranking accuracy and volatility consistently at both cross-section and security level. Thus, the proposed advancement is a significant step to boost the forecasting performance of the engine, making it a more valuable to produce high quality signals for downstream investment tasks.

REFERENCES

Daniel, Wayne W. (1990). “Spearman Rank Correlation Coefficient”. In: Applied Nonparametric Statistics. Boston: PWS-Kent, pp. 358–365.
Friedman, Jerome H. (2001). “Greedy Function Approximation: A Gradient Boosting Machine”. In: The Annals of Statistics 29.5, pp. 1189–1232. issn: 00905364. url: http://www.jstor.org/stable/2699986 (visited on 09/28/2023).
Mullner, Daniel (2011). Modern hierarchical, agglomerative clustering algorithms.arXiv: 1109.2378 [stat.ML].
Murtagh, Fionn and Pierre Legendre (Oct. 2014). “Ward’s Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward’s Criterion?” In: Journal of Classification 31.3, pp. 274–295. doi: 10.1007/ s00357-014-9161-z. url: https://doi.org/10.1007%2Fs00357-014-9161-z.
Chen, Tianqi and Carlos Guestrin (Aug. 2016). “XGBoost”. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. doi: 10.1145/2939672.2939785. url:https://doi.org/10.1145%2F2939672.2939785.
Ke, Guolin et al. (2017). “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”. In: Advances in Neural Information Processing Systems.Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. url: https://
proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.
Prokhorenkova, Liudmila et al. (2019). CatBoost: unbiased boosting with categorical features. arXiv: 1706.09516 [cs.LG].

CPP Investments

Computer Science
UNIVERSITY OF TORONTO

Master of Science in Applied Computing