Human Skill Through a Digital Lens: Evaluating Analysts with Machine Learning

Sean Cao[a] Norman Guo[b] Houping Xiao[c] Baozhong Yang[c]

[a]University of Maryland

[b]Saint Louis University

[c]Georgia State University

University of Louisville
Feb 2024

Motivation

Human skills

- Important to the economy and markets
- Difficult to analyze
  - Rational paradigm (e.g., Muth 1961; Lucas 1987)
  - Cognitive heuristics (e.g.,Kahneman and Tversky 1979)

Machines have succeeded in many tasks

- Image recognition
- Natural language processing
- Game playing
- Automatic driving

Can machines understand and evaluate human skills?

Overview

1. Design a human-friendly AI model for financial data.
    - Integrate domain knowledge.
    - Facilitate local non-linear interactions.
    - Convert images to numerical data.
2. Evaluate analysts' skills from a machine perspective.
    - Machine vs. human assess human skills differently in important dimensions
    - Answer the puzzle of post-revision drift
3. Extract valuable information from individual and collective analyst forecasts.
    - Generate significant abnormal returns from machine-selected analysts
    - Create a "smart" analyst consensus that better proxies for earning news before earnings announcements than the traditional analyst consensus

Research Objectives

Why the analyst setting

- Analysts are important financial intermediaries
- Earning forecasts are measurable individual opinions
- Observable features from analysts, firms and economy
- Past realized earnings can serve as benchmark for evaluation of performance
  - Manual labelling is labor-intensive
  - Learning from labels: Which analyst has information or is more skilled

Challenges

- Each analyst's private information and expertise
- High-dimensional, nonlinear interactions

Why Do We Use Machine Learning

Traditional Econometrics, e.g., OLS

- Have difficulty dealing with a large number of variables
- Cannot handle complicated nonlinear relations
- Optimized for in-sample interpretation, not out-of-sample prediction

Machine Learning Methods, e.g., Neural Networks

- Built-in dimension reduction to focus on more important variables
- Incorporate highly flexible nonlinear relations
- Model designs are optimized for out-of-sample predictions

Data and Features

Data Sources: IBES, Compustat, CRSP, Fed St. Louis, Thomson 13F, etc.

- **Analyst-level features**, $A_{i,j,t}$
  - 15 features
  - including Firm Experience, Forecast Horizon, Effort, Consensus (IBES), etc
- **Macro-level features**, T_t
  - 12 features
  - including Inflation, Oil Prices, Term Spread, Default Spread, VIX, etc
- **Firm-level features**, $F_{j,t}$
  - 40 features
  - including Size, Book to Market, Momentum, Accruals, Profit Margin, Asset Liquidity, Closed Price, Turnover, Institutional Ownership, etc

Note: analyst i, firm j, and time t

Construct Target Variable

$$\boldsymbol{Star_{i,j,t+1} = f(A_{i,j,t}, F_{j,t}, T_t) + \epsilon_{i,j,t+1}}$$

for analyst i, firm j, and time t.

- **Classification**: $Star_{ijt} = 1$ if the absolute forecast error of analyst i in the quarter t is lower than median of all analysts covering the firm j; otherwise $Star_{ijt} = 0$.

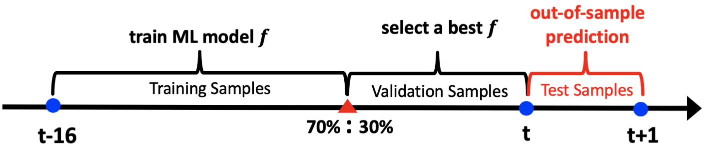
Construct Target Variable

Analysts	FE
John	0.15
Mary	0.2
Sarah	0.3
Lenard	-0.1
Brooke	0.5
Clifford	0.45
Emerson	-0.2
Olive	-0.4
Shelia	-0.35

Construct Target Variable

Analysts	Abs FE	Star
Lenard	0.1	1
John	0.15	1
Mary	0.2	1
Emerson	0.2	1
Sarah	0.3	0
Shelia	0.35	0
Olive	0.4	0
Clifford	0.45	0
Brooke	0.5	0

Timeline


train ML model *f*
select a best *f*
out-of-sample prediction
Training Samples
Validation Samples
Test Samples
t-16
70% : 30%
t
t+1

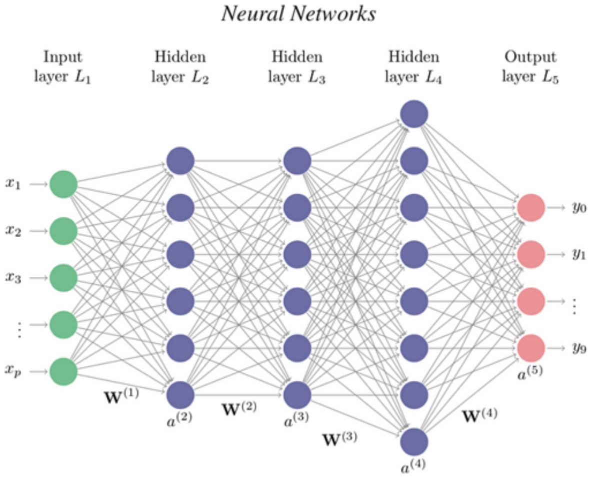
Issues of Machine Learning Models in Finance

AI Can Write a Song, but It Can't Beat the Market

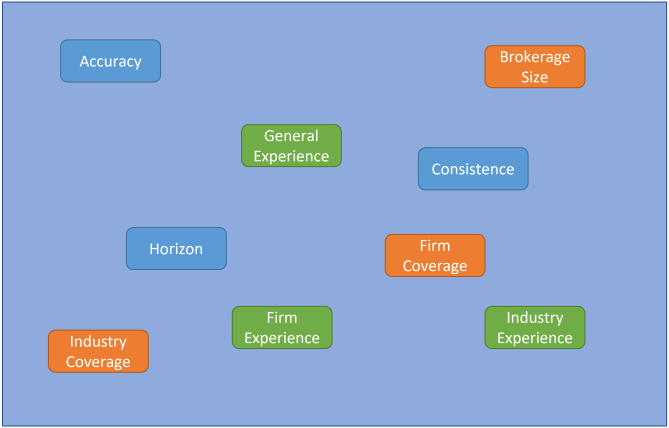
Quants have tried for decades with limited success at their biggest challenge

Market data is smaller in size and "noisier" than language and other data, making it harder to use it to explain or predict market moves

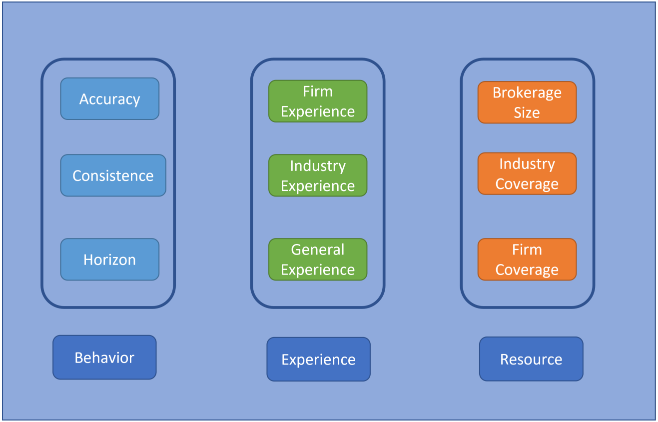
Example of ML: Feed Forward Neural Networks



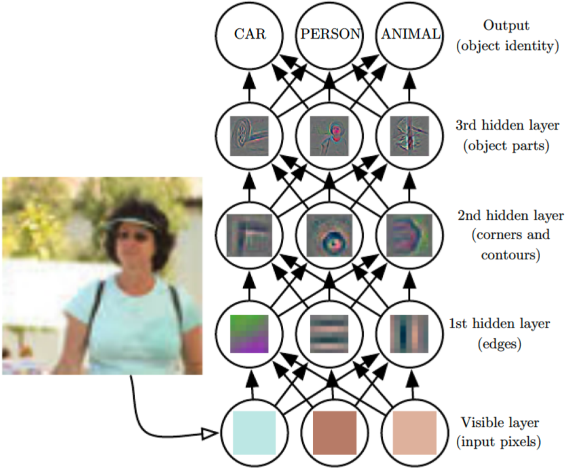
Example of ML: Convolutional Neural Networks (CNN)



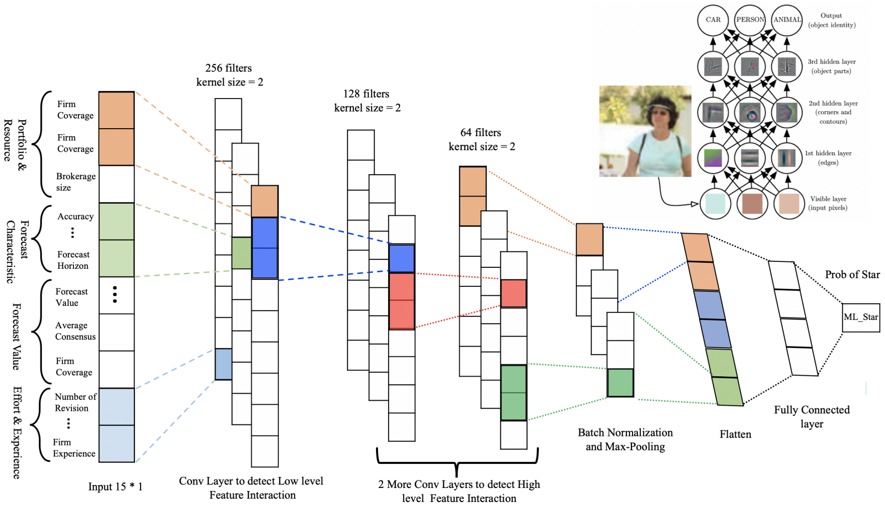
Example of ML: Convolutional Neural Networks (CNN)



Example of ML: Convolutional Neural Networks (CNN)



CNN Architecture



Explanation of Machine Learning Metrics

Metrics used to evaluate ML models:

- **Accuracy:** $\frac{True\ Positives}{Total\ Sample}$
- **Precision:** $\frac{True\ Positive}{True\ Positive + False\ Positive}$, measures Type I error
- **Recall:** $\frac{True\ Positive}{True\ Positive + False\ Negative}$, measures Type II error
- **F1 Score:** Average of precision and recall (harmonic average)

Feature Selection

- Not always "the more the better"
- Analyst features are the most important ones

Feature	Accuracy	Precision	Recall	F1 Score
$[Analyst]$	**61.57%**	62.89%	83.06%	71.10%
$[Analyst, Firm]$	60.80%	61.60%	83.33%	70.49%
$[Analyst, Macro]$	60.30%	60.57%	87.70%	71.26%
$[Analyst, Firm, Macro]$	59.49%	60.74%	82.92%	69.85%
$[Firm]$	56.33%	56.38%	99.68%	72.02%
$[Macro]$	56.38%	56.38%	100.00%	72.10%
$[Firm, Macro]$	56.29%	56.35%	99.66%	71.99%

Bringing in Domain Knowledge

Group Analyst features in four categories based on literature:

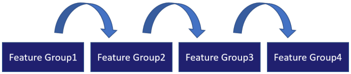
- **Forecast Values**: Forecast, Consensus from I/B/E/S, Average Consensus
- **Forecast Characteristics**: Accuracy, Consistence, Horizon
- **Effort & Experience**: Number of Revisions, Whether Report Revenue Forecast, Whether Report Cash flow Forecast, General Experience, Industry Experience, Firm Experience
- **Portfolio & Resource**: Analyst Firm Coverage, Analyst Industry Coverage, Brokerage Size

Comparison with Cases with Random Orders

- Random orders of variables do not work well with CNN
- Grouping of features are important!

Feature	Accuracy	Precision	Recall	F1 Score
[*FirmExperience*, *FirmCoverage*, *Accuracy*, *ReportCashflow*, *Consistency*, *IndustryExperience*, *BrokerageSize*, *IndustryCoverage*, *ForecastHorizon*, *ReportRevenue* *consensus_avg*, *ForecastValue*, *NumberofRevision*, *meanest_ibes*, *GeneralExperience*]	68.83%	71.58%	74.62%	73.05%
[*GeneralExperience*, *IndustryCoverage*, *FirmExperience*, *consensus_avg*, *ReportCashflow*, *Accuracy*, *ForecastValue*, *ReportRevenue*, *BrokerageSize*, *ForecastHorizon*, *IndustryExperience*, *meanest_ibes*, *NumberofRevision*, *Consistency*, *FirmCoverage*]	66.92%	69.02%	75.57%	72.13%
[*ForecastHorizon*, *NumberofRevision*, *Accuracy*, *ReportCashflow*, *ReportRevenue*, *consensus_avg*, *FirmExperience*, *IndustryCoverage*, *ForecastValue*, *meanest_ibes*, *Consistency*, *IndustryExperience*, *GeneralExperience*, *FirmCoverage*, *BrokerageSize*]	66.16%	66.84%	80.01%	72.83%
. . .				
[*ForecastValue*, *ReportCashflow*, *IndustryExperience*, *GeneralExperience*, *FirmCoverage*, *Consistency*, *consensus_avg*, *NumberofRevision*, *ForecastHorizon*, *IndustryCoverage*, *meanest_ibes*, *BrokerageSize*, *FirmExperience*, *ReportRevenue*, *Accuracy*]	57.14%	57.95%	89.00%	70.19%
[*ForecastHorizon*, *consensus_avg*, *GeneralExperience*, *FirmExperience*, *meanest_ibes*, *IndustryCoverage*, *IndustryExperience*, *ForecastValue*, *Consistency*, *Accuracy*, *ReportRevenue*, *NumberofRevision*, *FirmCoverage*, *BrokerageSize*, *ReportCashflow*]	57.04%	58.09%	87.22%	69.72%
[*ForecastHorizon*, *FirmCoverage*, *Accuracy*, *ReportRevenue*, *BrokerageSize*, *IndustryCoverage*, *ForecastValue*, *GeneralExperience*, *consensus_avg*, *Consistency*, *IndustryExperience*, *ReportCashflow*, *FirmExperience*, *NumberofRevision*, *meanest_ibes*]	56.83%	57.97%	86.87%	69.52%

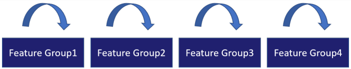
Importance of the Order of Feature Categories in CNN



- The order of different feature groups matters

Feature	Accuracy	Precision	Recall	F1 Score
[*Porfolio&Resource*, *Effort&Experience*, *ForecastChar*, *ForecastValue*]	69.79%	71.83%	76.35%	74.02%
[*Porfolio&Resource*, *ForecastChar*, *Effort&Experience*, *ForecastValue*]	69.58%	70.23%	79.92%	74.76%
[*Effort&Experience*, *ForecastValue*, *ForecastChar*, *Porfolio&Resource*]	69.57%	71.58%	76.30%	73.87%
...				
[*ForecastChar*, *ForecastValue*, *Porfolio&Resource*, *Effort&Experience*]	68.61%	69.49%	79.01%	73.94%
[*ForecastChar*, *Effort&Experience*, *ForecastValue*, *Porfolio&Resource*]	68.00%	69.33%	77.56%	73.21%
[*Effort&Experience*, *ForecastValue*, *Porfolio&Resource*, *ForecastChar*]	67.27%	67.10%	82.27%	73.92%

Order within Each Feature Category



- The order of different variables within each group matters

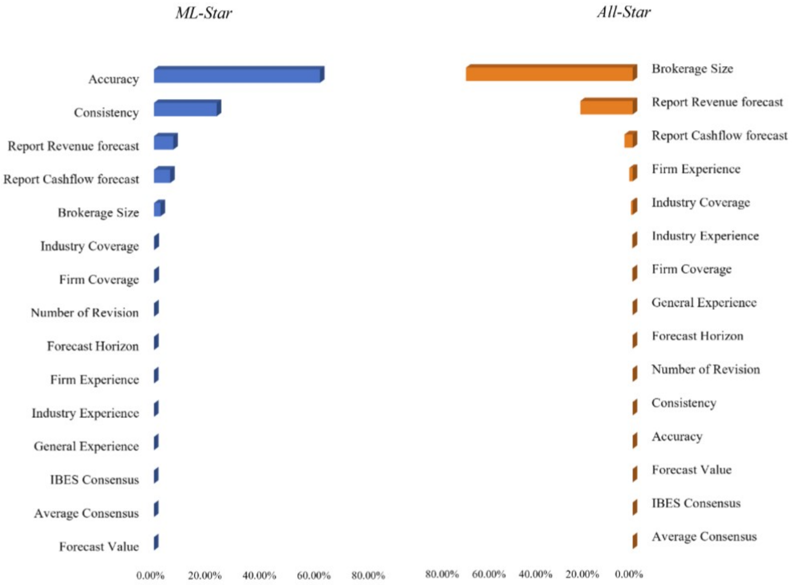
Features	Accuracy	Precision	Recall	F1 Score
[*IndustryCoverage*, *FirmCoverage*, *BrokerageSize*, *Accuracy*, *ForecastHorizon*, *Consistency*, *meanest_ibes*, *consensus_avg*, *ForecastValue*, *ReportCashflow*, *GeneralExperience*, *IndustryExperience*, *ReportRevenue*, *NumberofRevision*, *FirmExperience*]	70.33%	71.92%	78.33%	74.97%
[*FirmCoverage*, *BrokerageSize*, *IndustryCoverage*, *Consistency*, *ForecastHorizon*, *Accuracy*, *consensus_avg*, *ForecastValue*, *meanest_ibes*, *FirmExperience*, *GeneralExperience*, *ReportRevenue*, *ReportCashflow*, *IndustryExperience*, *NumberofRevision*]	70.06%	71.64%	78.15%	74.74%
[*BrokerageSize*, *FirmCoverage*, *IndustryCoverage*, *Accuracy*, *Consistency*, *ForecastHorizon*, *ForecastValue*, *meanest_ibes*, *consensus_avg*, *FirmExperience*, *IndustryExperience*, *NumberofRevision*, *ReportCashflow*, *ReportRevenue*, *GeneralExperience*]	69.96%	71.55%	78.09%	74.66%
...				
[*FirmCoverage*, *IndustryCoverage*, *BrokerageSize*, *Accuracy*, *ForecastHorizon*, *Consistency*, *meanest_ibes*, *consensus_avg*, *ForecastValue*, *ReportCashflow*, *ReportRevenue*, *IndustryExperience*, *NumberofRevision*, *FirmExperience*, *GeneralExperience*]	69.55%	71.32%	77.50%	74.20%
[*IndustryCoverage*, *BrokerageSize*, *FirmCoverage*, *ForecastHorizon*, *Accuracy*, *Consistency*, *consensus_avg*, *meanest_ibes*, *ForecastValue*, *NumberofRevision*, *IndustryExperience*, *ReportRevenue*, *FirmExperience*, *ReportCashflow*, *GeneralExperience*]	69.49%	71.52%	76.75%	74.04%
[*FirmCoverage*, *BrokerageSize*, *IndustryCoverage*, *Consistency*, *ForecastHorizon*, *Accuracy*, *meanest_ibes*, *consensus_avg*, *ForecastValue*, *ReportRevenue*, *NumberofRevision*, *ReportCashflow*, *GeneralExperience*, *IndustryExperience*, *FirmExperience*]	69.43%	71.09%	77.62%	74.20%

Model Comparison

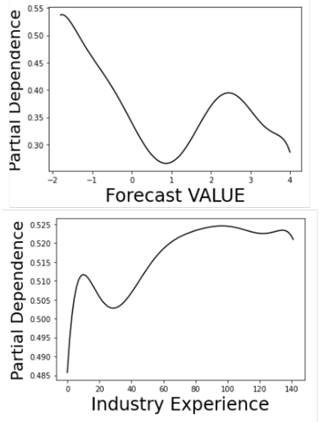
- Non-linear models outperform
- Convolutional Neural Networks (CNN), which proceeds from low-dimensional interactions of features to high-dimensional interactions, excels

	Models	Accuracy	Precision	Recall	F1 Score
Linear	Logistic Regression	53.81%	54.33%	90.23%	67.84%
	Logistic LASSO	55.49%	55.90%	82.93%	66.78%
Non-Linear	Gradient Boost	58.14%	57.70%	83.95%	68.40%
	Neural Network	59.81%	58.16%	90.86%	70.93%
	Convolutional Neural Network	70.33%	71.92%	78.33%	74.97%

Feature Importance



Partial Dependence



Forecast Accuracy

- ML predicted star analysts outperform historically accurate analysts and (human-labeled) all star analysts

	(1)	(2)	(3)	(4)
Variables		*Star*		
ML-Star	0.381***	0.382***	0.380***	0.380***
	(123.66)	(81.42)	(123.65)	(81.02)
Prior Star			0.018***	0.018***
			(19.72)	(16.41)
All-Star			0.009***	0.006***
			(6.29)	(3.06)
Year-Quarter FE	No	Yes	No	Yes
Firm FE	Yes	Yes	Yes	Yes
Observations	1,488,430	1,488,430	1,488,430	1,488,430
R-squared	0.145	0.145	0.145	0.145

Forecast Persistence

- The predictive power of the ML-Star is persistent

Variables	(1)	(2)	(3)	(4)	(5)
			Star		
	1 Qtr	2 Qtr	3 Qtr	4 Qtr	8 Qtr
ML-Star	0.056***	0.042***	0.038***	0.035***	0.025***
	(29.43)	(25.39)	(24.08)	(20.95)	(17.03)
Prior Star	0.036***	0.032***	0.028***	0.026***	0.022***
	(23.98)	(20.21)	(18.99)	(18.14)	(13.11)
All-Star	-0.001	-0.003	-0.006	-0.007	-0.008*
	(-0.23)	(-0.99)	(-1.55)	(-1.63)	(-1.90)
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes
Fund FE	Yes	Yes	Yes	Yes	Yes
Observations	1,308,358	1,172,893	1,054,670	951,168	640,661
R-squared	0.014	0.013	0.013	0.013	0.014

Forecast Accuracy: Subsample Analysis

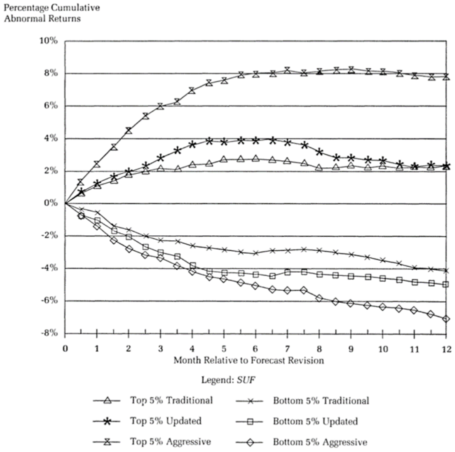
Analyst skill can be more accurately predicted by machines when

- Firm information is more transparent
- The economy is in a normal state

	ML-Star on Analyst Forecast Accuracy			
	High	Low	Diff	t-Stat
Bid Ask Spread	0.364***	0.415***	-0.051***	(-5.70)
Adj probability of informed trading	0.389***	0.431***	-0.042***	(-3.08)
Flesch-Kincaid Grade Level	0.398***	0.383***	0.015***	(2.28)
Accruals Quarlity	0.386***	0.409***	-0.023***	(-2.64)
Earning Quality	0.416***	0.384***	0.032***	(4.49)
Cashflow Volatility	0.365***	0.403***	-0.038***	(-6.15)
Return Volatility	0.362***	0.417***	-0.055***	(-6.64)
Firm Age	0.397***	0.380***	0.017***	(2.56)
NBER Crisis Dummy	0.366***	0.393***	-0.027**	(-1.96)

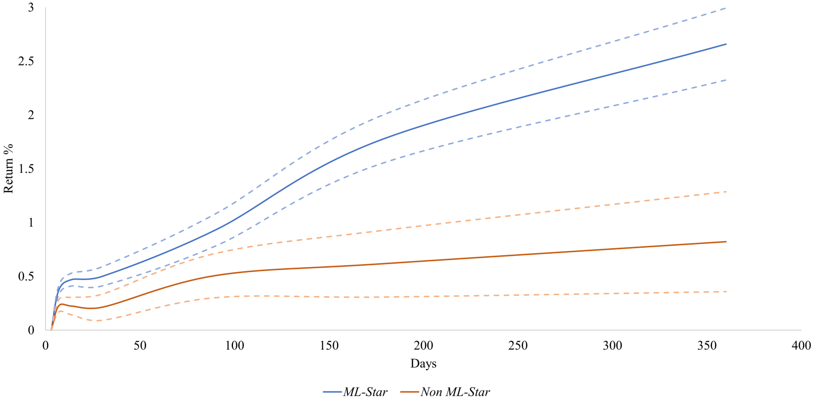
Post Revision Drift

- The phenomenon of delayed stock price reactions to analyst forecast revisions, is a well-documented market anomaly. (Stickel (1999))



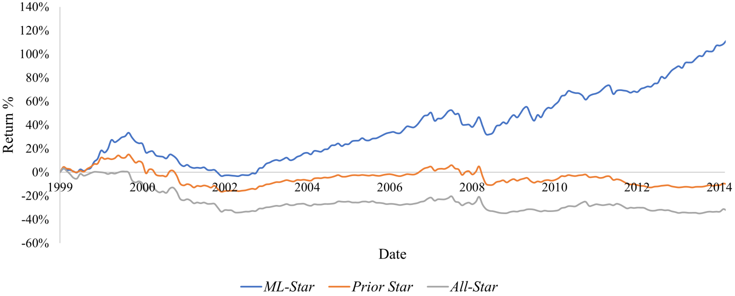
Post Revision Drift

- ML predicted star analysts explain the bulk of post analyst revision drifts.



Trading Strategy Returns: Post Analyst Revision Drift

Long positive revision, short negative revision



Generate Cloud Wisdom

- **ML Earnings Consensus**: We compute the ML consensus as the average of predicted ML-Star analysts' forecasts

Earning Forecast

- ML consensus provide additional predicting future earnings

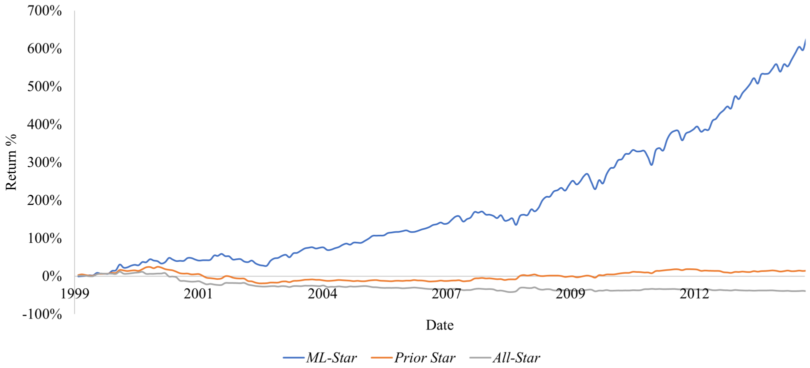
	(1)	(2)	(3)	(4)
Dependent Variable		Earnings		
ML-Consensus - Consensus	2.142***	2.034***	2.133***	2.058***
	(4.32)	(4.93)	(4.25)	(4.73)
Consensus	1.059***	1.091***	1.064***	1.112***
	(48.97)	(25.18)	(41.69)	(22.46)
Liquidity			-0.003	0.004**
			(-1.32)	(2.55)
Momentum			0.030***	0.012**
			(6.32)	(2.13)
Log_Size			-0.004	-0.007
			(-0.80)	(-0.73)
Book to Market			-0.010*	0.028*
			(-1.97)	(1.90)
Coverage			0.001**	-0.001**
			(2.24)	(-2.23)
Year-Quarter FE	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes
Observations	203,759	203,118	156,635	156,158
Adj. R-squared	0.771	0.791	0.790	0.808

Market Expectation

- ML consensus predicts returns around earnings announcements

Variables	(1) *CAR [-1, 1]*	(2) *CAR [2, 7]*	(3) *CAR [8, 14]*
ML-Consensus - Consensus	0.019**	-0.001	-0.003
	(2.60)	(-0.16)	(-0.51)
Consensus	0.002***	0.003***	0.002***
	(2.70)	(3.02)	(3.46)
Liquidity	-0.001	-0.001*	0.000
	(-1.46)	(-1.96)	(0.54)
Momentum	0.001	-0.004**	-0.002
	(0.86)	(-1.99)	(-1.24)
Log_Size	-0.012***	-0.007***	-0.007***
	(-11.45)	(-5.98)	(-6.86)
Book to Market	-0.001	-0.001	-0.002*
	(-1.26)	(-0.84)	(-1.82)
Coverage	-0.000	0.000	0.000
	(-0.81)	(0.82)	(0.06)
Year-Quarter FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
Observations	154,783	154,767	154,662
Adj. R-squared	0.0293	0.0410	0.0446

Trading Strategy Returns: Post Earning Drift



Factor Analysis

- Machine learning-based strategy captures unique insights not fully integrated into the market

	(1)	(2)	(3)	(4)	(5)
	Algorithm Return Based on PEAD				
Variables	FF3	FF4	FF5	Q4	SY4
Mkt-RF	0.479***	0.491***	0.534***	0.495***	0.544***
	(13.60)	(12.93)	(12.75)	(11.57)	(11.59)
SMB	0.004	-0.002	0.031	0.018	0.025
	(0.08)	(-0.04)	(0.58)	(0.35)	(0.48)
HML	-0.113**	-0.103**	-0.230***		
	(-2.46)	(-2.16)	(-3.50)		
Mom		0.025			
		(0.84)			
RMW			0.122*		
			(1.69)		
CMA			0.179*		
			(1.95)		
R_IA				-0.018	
				(-0.24)	
R_ROE				0.059	
				(0.93)	
MGMT					0.018
					(0.31)
PERF					0.095**
					(2.52)
Constant	0.007***	0.007***	0.006***	0.007***	0.006***
	(4.62)	(4.50)	(3.79)	(4.17)	(3.24)
Observations	236	236	236	236	216
adj R-squared	0.468	0.468	0.478	0.440	0.460

Conclusion

- **A ML measure of analyst skill**
  - A persistent skill measure that outperforms human-labeled star analysts and historically accurate analysts in future analyst forecasts
  - Explains the post-revision drift anomaly for analysts
  - Skill prediction is more accurate in a transparent information environment
- **A new earnings expectation measure from ML analyst consensus**
  - Better predicts earnings surprise
  - Predicts stock returns around earnings announcements
  - Generates profitable trading strategies for investors
  - AI provides significant incremental information to common consensus
- **Methodological contribution**
  - Feature and model selection in Machine Learning
  - CNN can capture subtle variable interactions by grouping and ordering of features
  - Interpretation of non-linear relations in deep-learning models
  - A new ML method to aggregate information from heterogeneous agents: Applicable to general settings, e.g., online forums, political opinions, and macroeconomic outlooks