Can Machines Understand Human Decisions: Dissecting Stock Forecasting Skill

Sean Cao[a] Xuxi Guo[a] Houping Xiao[a] Baozhong Yang[a]

[a]J. Mack Robinson College of Business, Georgia State University

FMA 2021

Motivation

Human decisions

- 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
- Text interpretation and translation
- Game playing
- Automatic driving

Can machines understand and evaluate human decisions?

Research Objectives

Research Questions

1. Can machines identify skilled analysts by analyzing their forecasting behavior?
2. What are the differences between machines and human experts when they evaluate analyst skill?
3. Can we extract valuable information from individual and collective analyst forecasts?
4. Can we provide a framework for analyzing other types of human decisions and behavior?

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
- Firm and macro-level factors
- High-dimensional, nonlinear interactions

Why Do We Use Machine Learning

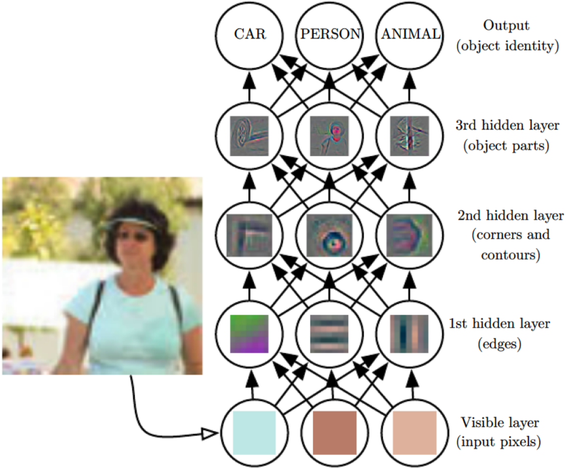
Traditional Econometrics, e.g., OLS

- 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

Example of ML: Convolutional Neural Networks (CNN)



Contributions

Machine-identified Analyst Skill

- ML-star (machine picked star) analysts outperform All-Star (human-picked star) analysts and historically accurate analysts in future analyst forecasts
- Persistent skill measure
  - Machines and human experts rely on different dimensions in evaluating analyst skill
- Explains the post-revision drift anomaly for analysts
- ML predicts skills better when the information environment is more transparent

A "smart" analyst consensus measure

- "Smart consensus" can be formed by aggregating forecasts by machine-identified skilled analysts
- Better predicts earnings surprise
- Predicts stock returns around earnings announcements
- Generates profitable trading strategies for investors
- AI provides significant incremental information to common consensus

Contributions

Methodological contributions

- Feature and model selection in Machine Learning
- Grouping and ordering of features can be important
- 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

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

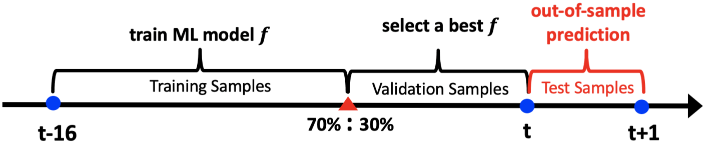
Note: analyst i, firm j, and time t

First Step: Prediction Model of Analyst Skill

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

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 $Star_{ijt} = 0$.



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

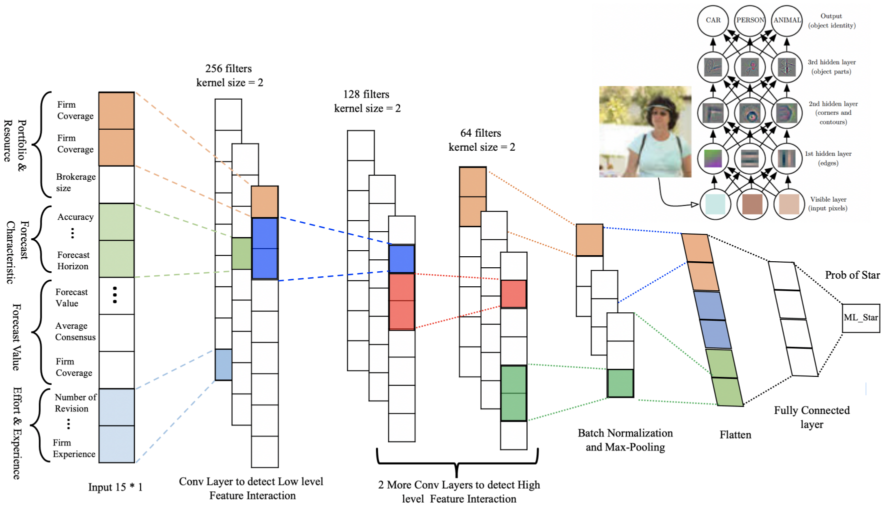
Feature	Accuracy	Precision	Recall	F1 Score
$[Analyst, Firm, Macro]$	68.60%	68.37%	77.45%	72.63%
$[Analyst, Macro]$	68.56%	68.53%	77.29%	72.65%
$[Analyst, Firm]$	68.62%	68.44%	77.30%	72.60%
$[Firm, Macro]$	53.93%	53.96%	99.66%	70.01%
$[Analyst]$	69.03%	69.10%	77.04%	72.86%
$[Firm]$	53.93%	53.94%	99.77%	70.03%
$[Macro]$	54.16%	54.16%	99.97%	70.26%

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	68.27%	68.02%	77.72%	75%
	Convolutional Neural Network	69.03%	69.10%	77.04%	72.86%

CNN Architecture

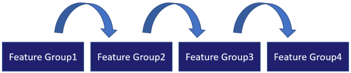


Grouping Analyst Features

Analyst features come naturally in 4 groups

- **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

Importance of the Order of Feature Categories in CNN



- The order of different feature groups matters slightly for the results

Feature	Accuracy	Precision	Recall	F1 Score
[*Portfolio&Resource*, *ForecastChar*, *ForecastValue*, *Effort&Experience*]	70.30%	70.72%	78.59%	74.44%
[*Effort&Experience*, *ForecastChar*, *ForecastValue*, *Portfolio&Resource*]	70.25%	70.62%	78.70%	74.44%
[*ForecastValue*, *Effort&Experience*, *Portfolio&Resource*, *ForecastChar*]	70.15%	70.68%	78.22%	74.26%
...				
[*Effort&Experience*, *Portfolio&Resource*, *ForecastValue*, *ForecastChar*]	69.68%	70.10%	78.32%	73.97%
[*ForecastValue*, *ForecastChar*, *Portfolio&Resource*, *Effort&Experience*]	69.67%	70.19%	78.07%	73.92%
[*Effort&Experience*, *ForecastChar*, *Portfolio&Resource*, *ForecastValue*]	69.63%	69.81%	79.10%	74.15%

Note: Results based on 5 randomly selected quarters

Order within Each Feature Category

Feature Group1 | Feature Group2 | Feature Group3 | Feature Group4

- The order of different variables within each group matters slightly

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%

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%

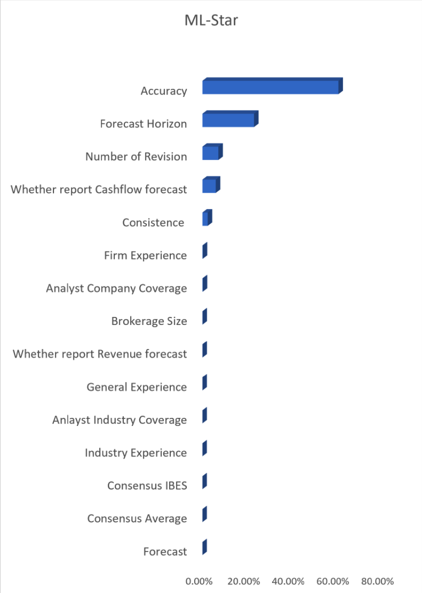
Second Step: Crowd Wisdom/Information Aggregation

- **ML-Star Analyst**: Analyst predicted to be a "Star" by the ML model
- **ML Earnings Consensus**: We compute the ML consensus as the average of predicted ML-Star analysts' forecasts

Summary Statistics

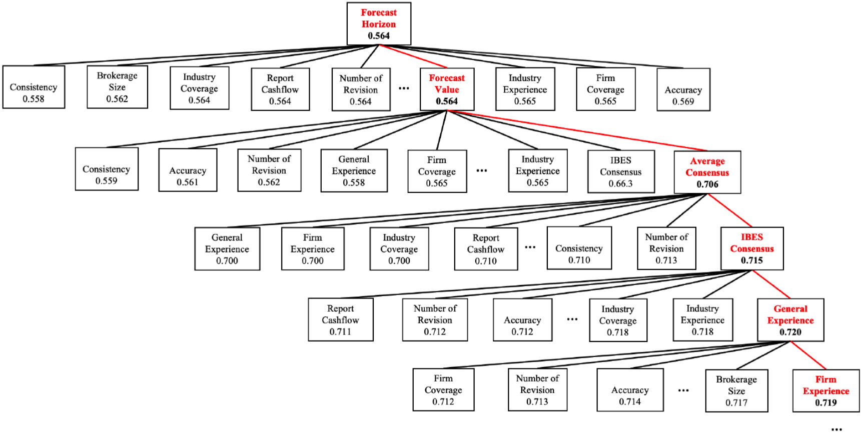
Variables	Mean	Median	Std	P25	P75	N
			Analyst Level			
Forecast	0.47	0.35	0.47	0.11	0.70	1,400,341
ML_Star	0.60	1.00	0.60	0.00	1.00	1,400,341
General Experience	37.05	30.00	37.05	15.00	53.00	1,400,341
Industry Experience	32.71	26.00	32.71	13.00	47.00	1,400,341
Firm Experience	16.03	11.00	16.03	5.00	22.00	1,400,341
Whether report Cashflow	0.10	0.00	0.10	0.00	0.00	1,400,341
Whether report Revenue	0.66	1.00	0.66	0.00	1.00	1,400,341
Number of Revision	1.43	1.00	1.43	1.00	2.00	1,400,341
Analyst Firm Coverage	18.56	18.00	18.56	14.00	23.00	1,400,341
Analyst Industry Coverage	2.56	2.00	2.56	1.00	3.00	1,400,341
Brokerage Size	61.19	54.00	61.19	24.00	96.00	1,400,341
Forecast Horizon	62.93	77.00	62.93	28.00	90.00	1,400,341
Consistency	0.58	0.60	0.58	0.33	0.83	1,400,341
Accuarcy	0.54	0.54	0.54	0.30	0.79	1,400,341
			Firm Level			
Earning	0.32	0.27	0.32	0.05	0.57	147,825
Consensus	0.34	0.26	0.34	0.06	0.56	147,825
CAR[-1,+1]	0.00	0.00	0.00	-0.04	0.04	147,825
CAR[+2,+7]	0.00	0.00	0.00	-0.03	0.03	147,825
CAR[+8,+14]	0.00	0.00	0.00	-0.03	0.03	147,825
Liquidity	0.15	0.04	0.15	0.01	0.11	147,825
Size	5892	1289	5892	443	4120	147,825
BM	0.54	0.44	0.54	0.25	0.72	147,825
Analyst Coverage	8.92	7.00	8.92	4.00	12.00	147,825

Linear Variance Importance: ML Star vs All Star



Nonlinearity of Feature Importance: CNN Model

- Forward stepwise selection strategy



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)
Historical Star			0.018***	0.018***
			(19.72)	(16.41)
All-Star			0.009***	0.006***
			(6.29)	(3.06)
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)
Historical 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)
Time 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

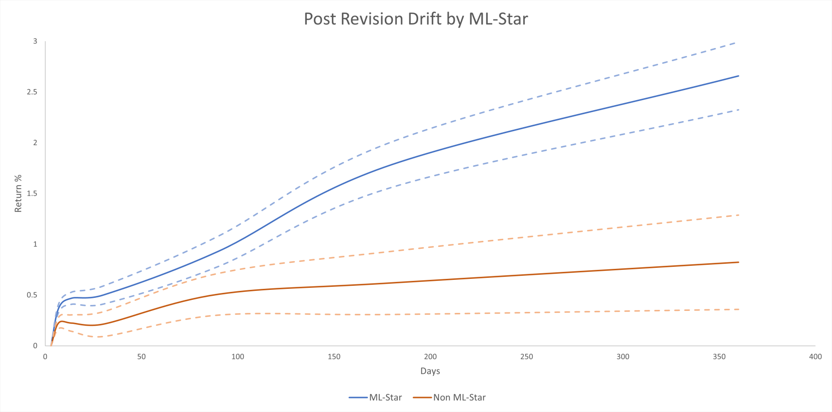
Analyst skill can be more accurately predicted by machines when

- Firm information is more transparent
- Analyst is more experienced, more focused, and has more resources
- The economy is in a normal state

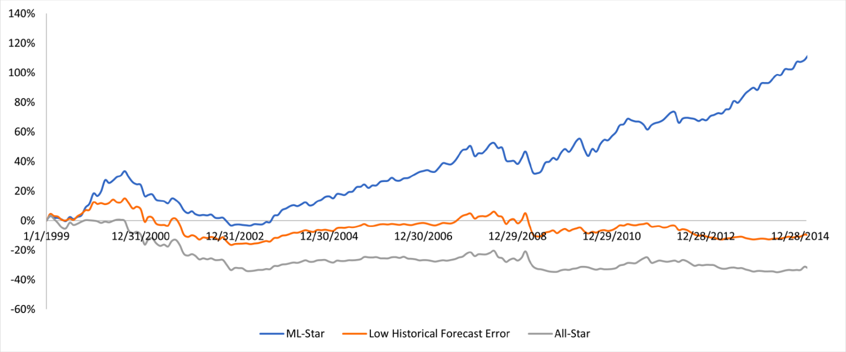
	High	Low	Diff	t-Stat
Information Asymmetry				
Bid Ask Spread	0.364***	0.415***	-0.051	(-5.70)
Adj probability of informed trading	0.389***	0.431***	-0.042	(-3.08)
Return Volatility	0.362***	0.417***	-0.055	(-6.64)
Cashflow Volatility	0.365***	0.403***	-0.038	(-6.15)
Earning Quality	0.416***	0.384***	0.032	(4.49)
Firm Age	0.397***	0.380***	0.017	(2.56)
Analyst Characteristics				
General Experience	0.395***	0.383***	0.012	(4.70)
Analyst Firm Coverage	0.385***	0.393***	-0.008	(-2.86)
Brokerage Size	0.395***	0.383***	0.012	(4.49)
Market Condition				
NBER Crisis Dummy	0.366***	0.393***	-0.027	(-1.96)

Post Revision Drift

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



Trading Strategy Returns: Post Analyst Revision Drift



Earning Forecast

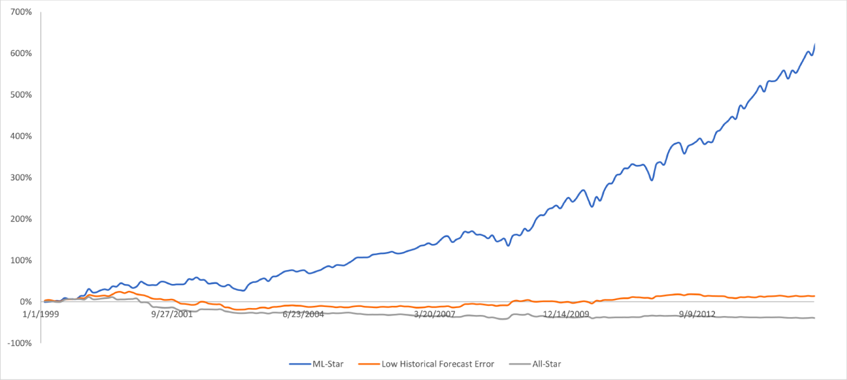
	(1)	(2)	(3)	(4)
Dependent Variable		*Earnings*		
Consensus_ML - Consensus	2.133***	2.142***	2.058***	2.034***
	(4.25)	(4.32)	(4.73)	(4.93)
Consensus	1.064***	1.059***	1.112***	1.091***
	(41.69)	(48.97)	(22.46)	(25.18)
Liquidity	-0.003		0.004**	
	(-1.32)		(2.55)	
Momentum	0.030***		0.012**	
	(6.32)		(2.13)	
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)	
Quarter FE	Yes	Yes	Yes	Yes
Fund FE	No	No	Yes	Yes
Observations	156,635	203,759	156,158	203,118
Adj R-squared	0.790	0.771	0.808	0.791

Market Expectation

- ML consensus predicts returns around earnings announcements

Variables	(1) *CAR [-1, +1]*	(2) *CAR [+2, +7]*	(3) *CAR [+8, +14]*
Consensus_ML - 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)
Quarter FE	Yes	Yes	Yes
Fund 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



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