Earnings prediction using machine learning: A survey*

Yuanchao Peng†

Abstract

This survey investigates the application of machine learning (ML) techniques in predicting corporate earnings. By reviewing literature spanning 2019 to 2024, this paper aims to provide a comprehensive overview of the methodological trends, strengths, and limitations of current ML approaches in the context of earnings prediction. While most research focuses on U.S. firms, a smaller portion examines international firms, including one study on Japan. A key trend is the preference for predicting directional changes in earnings (binary classification) over actual earnings levels, as classification models leverage high-dimensional data more effectively and yield economically meaningful insights. For example, portfolios based on predicted earnings changes outperform traditional models in generating abnormal returns. This paper also points out the shortcomings of existing research, 1) there is a lack of sufficient international evidence to prove that ML is superior to traditional models, 2) most earnings forecasts are short-term, and 3) there is a lack of exploration of non-financial data or forward-looking data. These shortcomings point to promising directions for future research. Another notable trend is that large language models (LLMs) have been shown to outperform traditional methods and human analysts in predicting the direction of corporate earnings. This emerging approach demonstrates impressive predictive performance in analyzing financial ratios and trends without extensive retraining.

JEL Classification: C53, G17, M41

Keywords: Earnings prediction, Machine learning, Large language models, Classification

1. Introduction

The main purpose of this study is to investigate recent papers that apply machine learning (ML) techniques to forecast corporate earnings. This analysis covers content from leading accounting journals and working papers, most of which focus on U.S. firms. In addition, one paper (Chattopadhyay et al. 2022) analyzes both U.S. and international firms, and one paper (Yakabi et al. 2024) focuses specifically on Japanese firms. This review aims to identify general research and methodological

* I would like to express my heartfelt gratitude to my supervisor, Professor Atsushi Shiiba, at the Graduate School of Economics, Osaka University, for his valuable guidance and support throughout the writing of this paper.

† Graduate Student, Graduate School of Economics, Osaka University

trends, analyze the strengths and weaknesses of existing methods, and propose potential future research directions for the application of ML in earnings prediction.

A major finding of this study is that most papers published in leading accounting journals use ML to predict directional changes in earnings rather than predicting specific level of earnings (Chen et al. 2022; Hunt et al. 2022; Jones et al. 2023). This tendency towards classification reflects the advantages of ML algorithms in extracting information from high-dimensional data sets and achieving higher accuracy than traditional regression models. At the same time, these studies also emphasize that predicting binary outcomes is more economically meaningful than predicting actual values. Hedge portfolios based on predicted results of increasing or decreasing earnings can achieve higher abnormal returns than traditional models.

This paper also highlights the concentration of research on U.S. firms, while international evidence supporting the superiority of ML over traditional methods is relatively limited. In addition, most studies rely on historical financial statement data as input variables, while paying little attention to non-financial or forward-looking data sources, such as textual information, macroeconomic indicators, or management's perceptions on future risks and uncertainties. Therefore, integrating these alternative data types into existing models provides a promising direction for future research. I also noticed that most earnings forecasts are conducted for the short term, specifically one year ahead. Forecasting earnings in the long term (3-5 years) is still an underexplored area.

A notable emerging trend is the use of Large Language Models (LLMs), such as GPT–4, in financial analysis. A recent study by Kim et al. (2024) demonstrate that LLMs, when applied with structured and anonymized financial data, can outperform human analysts in predicting the direction of future earnings. LLMs complement both human analysts and traditional ML models, excelling in scenarios where analysts are prone to bias or disagreement. They also rival advanced ML techniques, such as artificial neural networks, in certain predictive contexts. Moreover, LLMs exhibit unique capabilities in interpreting trends and financial ratios, offering state-of-the-art performance without specialized training. This highlights their potential not only as supportive tools but as central elements in financial decision-making. The inclusion of LLMs marks a significant shift in earnings prediction research, showcasing their ability to democratize financial analysis and opening new pathways for integrating AI-driven methods into finance. These developments suggest a promising and optimistic direction for future research, which will be further discussed in Section 6.

At the end of Section 3, I provide a structured overview of recent advancements in using ML to predict future earnings in Table 3. It highlights both the diversity and evolution of methods and findings in this field. Studies from leading journals, such as *The Accounting Review* and *Journal of Accounting Research*, along with working papers, collectively demonstrate the growing preference for ML over traditional methods like logistic regression and analysts' forecasts. Decision Tree-based methods, particularly Random Forest and Gradient Boosting, emerge as high performers across various studies, often yielding superior predictive accuracy and economic benefits, such as enhanced portfolio returns. Evaluation metrics range from statistical measures (e.g., MAE, RMSE) to economic outcomes (abnormal returns), reflecting the dual focus on accuracy and practical outcome. Notably,

while most studies affirm ML's advantages, exceptions like Campbell et al. (2023) underscore its limitations in specific contexts. Overall, the research captures the field's progress, showing both the promise of innovative techniques like LightGBM and LLMs and the ongoing challenges in consistently outperforming traditional methods. This survey aims to synthesize these findings, offering a comprehensive analysis of ML's role in advancing earnings prediction.

The rest of this paper is organized as follow. In Section 2, I review papers that apply ML to predict earnings directional changes. Those papers of using ML to predict level of earnings are discussed in Section 3. At the end of Section 3, I provide an overview of studies investigated in this paper. From Section 4 to Section 5, I discuss the potential challenges and opportunities of using ML to forecast future earnings. In Section 6, I present an emerging trend of using LLMs to predict earnings. Section 7 concludes and provide future research path on this topic.

2. Predicting Changes of Earnings Ratios: Classification

In this section, I will first summarize the ML methods used in this field, as well as the strengths and shortages of each method. Then I will summarize the conclusions of the main papers in this field.

2.1 Main Algorithms and Methodology

Among the classification algorithms of ML, decision tree-based algorithms have been widely used in recent years. The most popular algorithms are Random Forest and Gradient Boosting Machine. Both algorithms belong to ensemble algorithms, but they have significant differences in the construction of decision trees and the aggregation of final results. Therefore, many studies use these two algorithms at the same time to test the credibility of the results. Random forest builds multiple decision trees on different bootstrap samples of training data and randomly assigns predictor variables to each decision tree. The number of decision trees in the model and the number of predictor variables assigned to each decision tree are the two most critical parameters. The determination of these parameters usually requires a series of tests to find the best parameter combination suitable for a specific data set. This process is called parameter tuning. Random forest reduces overfitting of data by averaging the prediction results of these different decision trees and further reduces variance by reducing the influence of the main variables. The final prediction is a simple average of the predictions from each of the individual trees.

The Gradient Boosting algorithm builds decision trees sequentially, and each subsequent tree focuses on correcting the mistakes made by the previous tree. Therefore, there is a dependency between each decision tree in the Gradient Boosting algorithm, which is the biggest difference from the Random Forest. Specifically, the Gradient Boosting algorithm initially starts with a weak learner (decision tree) and iterates for each sample, calculating the residual between the predicted value of the current model and the true value. The residual represents the part that the current model failed to predict correctly. Therefore, these residuals will serve as the training target for the next decision tree. This process is iterated until the performance of the model no longer improves. In the process of training the model, the hyperparameter-learning rate is usually adjusted, which controls the degree of influence

of each decision tree on the final model. A smaller learning rate makes the model more robust, but converges more slowly. The final prediction result of the Gradient Boosting model is the weighted sum of the prediction results of all decision trees. The weight is usually related to the performance of the decision tree, and the decision tree with better performance will get a larger weight. In order to prevent overfitting, the Gradient Boosting algorithm usually introduces regularization terms, such as limiting the depth of the decision tree or the number of nodes per leaf.

In recent years, many studies have used optimized Gradient Boosting algorithms to improve data processing efficiency or reduce memory usage. For example, LightGBM optimizes data processing speed and model memory consumption. XGBoost significantly improves data processing speed without reducing the reliability of the model.

I summarize the advantages and disadvantages of these decision-tree based algorithms in Table 1. From Table 1, we can see that 7 papers use the Random Forest algorithm, far more than other algorithms. A main reason is that the Random Forest effectively reduces the risk of overfitting while maintaining a high model robustness. For the same reason, 4 papers use the StochasticGBM algorithm. Table 1 also includes paper indices that indicate which studies employed these methods. For details on the papers referred to by the indices, please refer to Table 3.

Table 1: Comparison of Decision-Tree based Algorithms

Feature	LightGBM	StochasticGBM	XGBoost	GBRT	Random Forest
Data Sampling	GOSS, EFB	Random Subsampling	Full Dataset	Full Dataset	Random Subsampling
Feature Selection	Random Splits	Pre-processing	Pre-processing	Pre-processing	Random Splits
Tree Growth	Leaf-wise	Level-wise	Level-wise	Level-wise	Level-wise
Regularization	L1		L1 and L2	L1 (Optional)	
Speed	Very Fast	Moderate	Fast	Moderate	Moderate
Overfitting Risk	Higher	Lower	Moderate	Higher	Lower
Robustness	Moderate	High	High	Moderate	High
Memory Usage	Low	Moderate	Moderate to High	Moderate	Moderate to High
Interpretability	Moderate	Low	Moderate	Moderate	Moderate
Paper Index	F	C, E, I, K	D	B	A, B, C, E, H, J, K

Notes: Gradient-based One-Side Sampling (GOSS) is a technique used to speed up gradient boosting algorithms. GOSS prioritizes data points with large gradients while randomly sampling from those with smaller gradients. This selective focus ensures the algorithm learns more effectively from challenging cases while reducing the computational burden. Exclusive Feature Bundling (EFB) is a method for reducing the dimensionality of datasets with a large number of sparse features. EFB groups features that are mutually exclusive into a single "bundle." This bundling reduces the number of features without losing significant information, making the algorithm faster and more efficient while preserving predictive accuracy. For details on the papers referred to by the indices, please refer to Table 3.

2.2 Main Results of Related Papers

Many studies use ML methods to predict the direction of earnings changes. One of the reasons is that a large number of studies construct hedge portfolios based on the predicted direction of earnings changes.

Chen et al. (2022) use Random Forest and Stochastic Gradient Boosting to forecast the direction of one-year-ahead earnings change. They obtain detailed financial ratios from XBRL filings and apply a large set of variables including 4,000 distinct financial items with their current, lagged and percentage changes value. They solve the class imbalance problem (earnings increase sample outnumber decrease sample) by adjusting earnings changes for the average change in EPS (Earnings Per Share) over the past four years. They obtain 3,610 earnings increase samples and 4,539 earnings decrease samples during year 2012 to 2018. Instead of using standard cross-validation, the study uses a rolling sample splitting approach that training and validation samples are gradually shift forward in time. This approach ensures that predictions rely only on the most recent data without backward-looking biases. Their models achieved an area under the curve[1] (AUC) between 67.52% and 68.66%, significantly outperforming random walk model (50%). The annual size-adjusted returns to hedge portfolios formed based on predictions range from 5.02% to 9.74%. The superior performance compared to traditional logistic regression models and analyst forecasts is attributed to both the nonlinear interactions captured by ML and the use of more detailed financial data. These findings underscore the value of ensemble learning and detailed financial data for binary earnings change predictions.

Jones et al. (2023) uses Gradient Boosting Machine to predict next period change in profitability based on a model proposed by Penman and Zhang (2004). Changes in profitability is defined as the difference between return on net operating assets (RNOA) at year t+1 with RNOA at year t. To avoid look-ahead bias, the dataset is divided into seven distinct training and test periods, ensuring that no future data from the test samples influences the training process. They find that Gradient Boosting Machine and Random Forests, consistently outperformed traditional models across various metrics (R^2, MAE[2], RMSE[3]). They identified both asset turnover and profit margin (components of the DuPont decomposition) as strong predictors, contradicting to the results of prior research. The study also found that the PZ model's key variables (e.g., growth in net operating assets and RNOA) remained robust predictors even in high-dimensional settings. The research suggests that while ML models enhance interpretability and accuracy through nonlinear interactions and high-order effects, they may not always translate to superior economic returns in portfolio applications compared to traditional regression models. Future research is encouraged to explore when ML's predictive gains lead to economic benefits.

---

1 The Area Under the Curve (AUC) is a metric used to evaluate the performance of binary classification models. It is derived from the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds. AUC ranges from 0 to 1. The closer the AUC is to 1, the better the model's ability to separate positive and negative classes.

2 MAE represents the average absolute difference between predicted and actual values.

3 RMSE represents the square root of the average squared differences between predicted and actual values.

Hunt et al. (2022) evaluate ML's effectiveness in predicting the direction of earnings changes and its utility in returns prediction. They found that while Elastic Net Regression did not outperform traditional Stepwise Logit models, Random Forest significantly improved out-of-sample prediction accuracy across all subsamples and time periods. Additionally, trading strategies based on Random Forest predictions yield higher abnormal returns than those based on other models, suggesting a practical advantage in financial contexts. The study advocates for exploring other nonparametric methods (e.g., Neural Networks, Support Vector Machines[4]) as they may offer even better performance. Random Forest's flexibility and ability to handle raw data without preprocessing (like standardization) highlight its utility for practical applications in predicting binary earnings outcomes.

I also surveyed two working papers on this topic. Cui et al. (2020) evaluated the application of LightGBM combining the dimension reduction technique-Principal Component Analysis (PCA) for forecasting directional changes of earnings. Their study compared the model's performance against analysts' consensus estimates and traditional logistic regression models. While the proposed model outperformed logistic regression in prediction accuracy and computational speed, it fell short of matching the performance of analysts, who benefit from broader information, including qualitative and potentially insider insights that are difficult to quantify. The study highlights the limitations of relying solely on structured data from public databases like Compustat and Thomson Reuters but emphasizes the potential for improvement. The authors suggest that incorporating non-quantitative data through advanced techniques such as Natural Language Processing (NLP) could enable the model to extract valuable insights from market news and textual disclosures.

Anand et al. (2019) investigate the effectiveness of classification trees in generating out-of-sample profitability forecasts. Using data from U.S. firms (1963-2017), the study evaluates directional changes in five profitability measures: return on equity (ROE), return on assets (ROA), return on net operating assets (RNOA), cash flow from operations (CFO), and free cash flow (FCF). The ML method achieves classification accuracies between 57% and 64%-significantly better than the random walk's 50%. Notably, its performance remains stable over a five-year forecast horizon. The study finds higher classification accuracy for cash flow measures (CFO, FCF), especially when accruals are included, compared to earnings-based measures (ROE, ROA, RNOA). However, in extreme portfolios of conditioning variables, earnings-based measures often outperform cash flow measures, indicating that no single profitability metric is superior under all conditions.

Although, most of the studies are using samples from the U.S., there is one study provides international evidence from Japan. Yakabi et al. (2024) examines the predictability of the direction of future earnings changes using ML techniques applied to Japanese companies' financial data based on the methodology of Chen et al. (2022). They find that Random Forest and Gradient Boosting outperformed Logistic Regression in terms of prediction accuracy and portfolio return. The abnormal

[4] Support Vector Machines (SVM) is a supervised machine learning algorithm used for classification and regression tasks. The primary goal of SVM is to find the best decision boundary (or hyperplane) that separates data points of different classes in a feature space. For specific explanation, please refer to: https://link.springer.com/chapter/10.1007/978-1-4899-7641-3_9.

returns generated by portfolios based on ML model predictions are statistically significant, indicating that the market does not fully incorporate information available in the financial statements. This finding challenges the efficient market hypothesis. They use 62 financial indicators as features, derived from Japanese companies' financial statements. Predictive performance is evaluated using the area under the ROC curve (AUC) and abnormal returns from hedge portfolios constructed based on the predictions. They also conduct a preliminary analysis using a Large Language Model (LLM), specifically GPT–4, to assess its potential in predicting earnings changes. The LLM (GPT–4) showed mixed results. While achieving a lower AUC compared to other models, it generated the highest abnormal return (AR). This suggests the LLM might provide valuable insights by incorporating qualitative factors alongside quantitative data, though further research is needed to confirm its reliability.

3. Predicting Level of Earnings Ratios: Regression

3.1 Main Algorithms and Methodology

The OLS, LASSO, RIDGE are the most popular algorithms in this task. The OLS model aims to estimate parameters by minimizing the sum of squared differences between observed and predicted values:

$$\beta^{OLS} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2.$$

When the number of parameters increases, OLS is prone to overfitting the model in-sample, leading to poor predictive performance out-of-sample (Chattopadhyay et al. 2022). To address this problem, penalized models, also referred to as "Shrinkage" methods, are designed to give the highest weights to a subset of predictors that demonstrate the strongest predictive power. RIDGE minimizes the sum of squared deviations while adding a penalty proportional to the square of the coefficients' magnitudes.

$$\beta^{RIDGE} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

LASSO also minimizes the sum of squared deviations but adds a penalty proportional to the absolute values of the coefficients.

$$\beta^{LASSO} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

Elastic Net combines the penalty of LASSO (L1) and RIDGE (L2) to enable it to handle high-dimensional data. At the same time, as the complexity of the model increases, the interpretability of the results decreases. From Table 2, we can see that there is no preference of specific model among papers investigated here. In general, there is no best model that is suitable for all situations. We need to select the most proper algorithm according to different data and purposes.

K-Nearest Neighbors (KNN) is a simple yet powerful supervised learning algorithm used for classification and regression tasks. For a given input data point, KNN calculates its distance to all the data points in the training dataset. Then, it identifies the k closest data points (neighbors) based on a distance metric (e.g., Euclidean, Manhattan, or Minkowski distance). Lastly, KNN predicts the output

Table 2: Comparison of OLS-based Algorithms

Feature	OLS	LASSO	RIDGE	Elastic Net
Penalty	None	L1	L2	Combination of L1 and L2
Feature Selection	No	Yes	No	Yes
Multicollinearity	Sensitive	Robust	Robust	Robust
Interpretability	High	High	High	Moderate
Flexibility	Low	Moderate	Moderate	High
Paper Index	B, C, I	B, C, D	B, C, D	B, H

Notes: For details on the papers referred to by the indices, please refer to Table 3.

by averaging the values of the k neighbors. One of the key steps in KNN is choosing the proper value of k, which determines the number of neighbors considered for making predictions. A small k can make the algorithm sensitive to noise, while a large k can dilute the influence of nearby neighbors, making predictions less specific. In recent study, Easton et al. (2024) introduced this method into earnings prediction task.

Artificial Neural Networks (ANNs) is another powerful ML methods that can handle high-dimensional dataset and complex relationships between features. However, ANNs are often criticized for being difficult to interpret compared to simpler models like linear regression or decision trees. Despite the fact that there is only one study used this method in earnings prediction, I will still briefly introduce how ANNs works. ANNs typically consist of three types of layers: 1) the input layer, 2) hidden layers, and 3) the output layer. Information flows from the input layer through the hidden layers to the output layer. Each neuron in the layers computes a weighted sum of its inputs:

$$z = \sum_{i=1}^{n} w_i x_i + b,$$

where w_i is the weight of i-th input, x_i is the value of i-th input, and b is the bias term. The weighted sum z is passed through an activation function to determine the neuron's output:

$$a = \mathrm{f}(z).$$

The activation functions include Sigmoid: $\sigma(z) = \frac{1}{1+e^{-z}}$, ReLU: $\mathrm{ReLU}(z) = \max(0, z)$, and Tanh: $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$. Activation functions impact how gradients are calculated and propagated. Each activation function has its advantages and disadvantages. For example, Sigmoid and Tanh may cause gradients to shrink (vanishing gradient problem), and ReLU ensures gradients flow effectively for positive z, improving training in deep networks. The output a becomes the input for the next layer or the final prediction. a is compared to the true value using a loss function, which quantifies prediction errors. The algorithm calculates the gradient of the loss with respect to the weights and biases using

the chain rule. Gradients are propagated backward through the network to update weights and biases. Weights and biases are adjusted to minimize the loss:

$$w_i = w_i - \eta \frac{\partial \text{Loss}}{\partial w_i},$$

where η is the learning rate, controlling the step size of updates. Lastly, the above steps are repeated across multiple iterations until the loss converges to a satisfactory level.

3.2 Main Results of Related Papers

I find that relatively few studies use ML to predict specific level of earnings. Even though it is more challenging to provide accurate prediction of exact level of earnings than to predict the direction of earnings changes. However, this does not mean that predicting level of earnings is meaningless or impossible. Several studies has provide insights on this topic.

Easton et al. (2024) utilize KNN approach to predicting one-year-ahead earnings. They developed a simple and effective method to predict the future earnings of a target firm by identifying the firms with similar history earnings. For instance, the Euclidean distance between firm i's M-year earnings history ending in year t and firm j's M-year earnings history ending in year s is calculated as:

$$\text{Distance}^{M}_{(i,t,j,s)} = \sqrt{\sum_{m=1}^{M} \left(EARN_{i,t-m+1} - EARN_{j,s-m+1}\right)^2}.$$

This method is based on the assumption that historical earnings serve as a reliable indicator of future performance; firms with similar past performance are likely to exhibit similar future performance. The earnings prediction for the subject firm-year is derived from the median of the lead earnings observed among its identified nearest neighbors. Easton et al. (2024) advocate for the simplicity of the KNN algorithm in forecasting corporate earnings, drawing on comparisons with firms that have similar current and past earnings. They assert that a simpler forecasting method is easier to interpret, modify, and less prone to overfitting. Their findings indicate that KNN significantly outperforms more complex models, such as advanced KNN variations, random walk models, and existing regression models in terms of accuracy. Moreover, KNN forecasts of longer-term earnings per share (EPS) and aggregate EPS were found to be more precise than those generated by professional analysts. A distinct advantage of the KNN algorithm is its ability to self-assess accuracy through the Mean Absolute Deviation[5] (MAD) metric, which effectively predicts forecast accuracy and provides investors with an indication of reliability. Their research underscores the notion that increasing the number of variables or extending the historical data scope does not enhance forecast accuracy. Instead, it affirms that recent earnings history is a robust predictor of future earnings when contextualized appropriately. This study highlights the practical value of a simple, comparable-firm-based method for earnings prediction, advocating for both simplicity and the careful selection of relevant historical data.

[5] In Easton et al. (2024), MAD is the median of the absolute values of the differences between each nearest neighbor's lead earnings and the median of the nearest neighbors' lead earnings.

Cao and You (2024) find non-linear ML models significantly outperform traditional earnings prediction models. This improved performance is largely due to the models' capacity to identify economically important predictors and capture nuanced, nonlinear relationships within financial data. The information collected from ML forecasts has substantial economic value for investors, as it also demonstrates predictive power for future stock returns. The study further compares ML-based forecasts with analysts' consensus forecasts, noting that ML models perform comparably to analyst forecasts over a one-year horizon and surpass them over longer periods. Moreover, ML models provide incremental information beyond analyst forecasts, even when analysts have access to comprehensive financial statement data. The ML models also help in detecting optimistic biases in analysts' predictions, supporting investors' need for objective data analysis. Overall, their results underscore that ML is an effective tool for deriving relevant insights for investors from financial statements and highlight the continued value of fundamental analysis. This reinforces the potential for ML in financial analysis by enabling sophisticated pattern recognition and data utilization for improved earnings prediction.

Campbell et al. (2023) conclude that while ML methods have the potential to improve earnings forecasts, their effectiveness is highly dependent on model specification choices. Specifically, they found that 90% of ML models evaluated did not outperform analysts' forecasts. However, the best-performing ML forecasts consistently correct for predictable analyst biases related to past errors and stock prices, leading to statistically significant accuracy improvements, particularly for small-cap firms and over longer forecasting horizons. Additionally, the study reveals that investors' earnings expectations, as reflected in stock prices, partially account for these biases but do not fully correct them, with price realizations often lagging by up to nine months. Overall, the findings indicate that the most accurate ML forecasts can mitigate predictable biases in analyst forecasts and align more closely with investors' expectations, particularly for large-cap firms with significant institutional ownership.

Van Binsbergen et al. (2023) conclude that the pricing of assets heavily relies on earnings forecasts, which are often upward-biased. They introduce a novel ML forecasting algorithm that is statistically optimal and resistant to variable selection bias, demonstrating its effectiveness in out-of-sample contexts compared to traditional linear forecasts. This new benchmark serves not only as a valuable input for asset pricing but also as a real-time tool for evaluating analyst earnings forecast biases over time and across different stocks. Their analysis reveals significant variation in these biases, and they find that stocks with the most upward-biased earnings forecasts tend to experience lower future returns, while those with downward biases generate higher returns. This suggests that analysts' forecast errors can significantly influence asset prices.

Chattopadhyay et al. (2022) explore the effectiveness of ML techniques in forecasting future earnings and estimating implied cost of capital (ICC). The study evaluates three ML models-LASSO regression, RIDGE regression, and Extreme Gradient Boosting (XGBoost). They compare the results with the random walk model, the Hou et al. (2012)'s model, and the earnings persistence and residual income models (Li and Mohanram, 2014). Additionally, the ML models are benchmarked against a simple linear model with an augmented set of predictors. They are the first study that investigate both U.S. and international firms. In the U.S. sample, they find that XGBoost generates

Table 3: Overview of Machine Learning Applications in Earnings Prediction

Author (Year)	Journal	Conclusion	ML method	Evaluation Metric	Paper Index
Anand et al. (2019)	Working Paper	ML achieves classification accuracies ranging from 57-64%, better than the random walk.	Random Forest	Out-of-sample Accuracy	A
Campbell et al. (2023)	Working Paper	90% of ML models evaluated did not outperform analysts' forecasts	OLS, LASSO, RIDGE, Elastic Net, Random Forest, GBRT	MAE, MSE	B
Cao and You (2024)	Financial Analysts Journal	ML models generate more accurate forecasts than state-of-the-art earnings prediction models	OLS, LASSO, RIDGE, Random Forest, StochasticGBM, ANNs	MAFE	C
Chattopadhyay et al. (2022)	Working Paper	Nonlinear tree-based models outperform extant models over different horizons	LASSO, RIDGE, XGBoost	MAFE	D
Chen et al. (2022)	Journal of Accounting Research	ML methods outperform logistic regressions and analysts' forecasts	Random Forest, StochasticGBM	ROC, AUC, Hedge portfolio return	E
Cui et al. (2020)	Atlantic Economic Journal	ML models outperform logistic regression but unable to surpass analysts	LightGBM	Out-of-sample Accuracy	F
Easton et al. (2024)	The Accounting Review	KNN forecasts are more accurate than extant approaches for one-, two-, and three-year-ahead earnings	KNN	MAFE, MSE, MAD	G
Hunt et al. (2022)	Accounting Horizons	Random forest provides better out-of-sample accuracy and higher abnormal returns	Stepwise Logit Regression, Random Forest, Elastic Net	Out-of-sample accuracy, Abnormal return	H
Jone et al. (2023)	Contemporary Accounting Research	ML methods predict out of sample better than traditional regression methods	StochasticGBM, OLS	MSE, RMSE, MAPE, R^2	I
Van Binsbergen et al. (2023)	The Review of Financial Studies	Introduce a novel ML algorithm resistant to variable selection bias, demonstrating its effectiveness in out-of-sample contexts compared to traditional linear forecasts	Random Forest	Forecasting Bias	J
Yakabi et al. (2024)	Working Paper	Random Forest and Gradient Boosting outperformed Logistic Regression in terms of prediction accuracy and portfolio return	Random Forest, StochasticGBM, LLMs	ROC, AUC, Hedge portfolio return	K

the most accurate forecasts, particularly for small firms and firms with volatile earnings. However, the improvements in forecast accuracy are modest. For the international sample, XGBoost demonstrates significantly superior performance, highlighting its robustness in settings with sparse coverage and volatile earnings. ICC tests corroborate these findings, with XGBoost consistently outperforming other models, especially for international firms where traditional cross-sectional models underperform. The paper highlights methodological contributions by demonstrating XGBoost's ability to deliver accurate forecasts with relatively low computational demands compared to other ML models like Random Forest or Gradient Boosting. However, they also acknowledge limitations, noting that their analysis relies on a static set of explanatory variables. Future research could enhance forecast accuracy by incorporating non-financial and market-based signals. The findings emphasize the potential of ML models in advancing earnings forecasting and ICC estimation.

4. Challenges in using ML to Predict Future Earnings

Although ML models have been proven to be more efficient and accurate than traditional models in many fields, their application in earnings prediction still faces many challenges.

i. Data Quality and Complexity

ML models require extensive and detailed data to accurately predict earnings. Studies (e.g., Chen et al., 2022) leverage large sets of detailed financial information, yet this dependency can introduce challenges related to data collection, preparation, and ensuring consistency across time periods and different firms. Complex and high-dimensional data also increase the likelihood of overfitting, especially with certain algorithms. Although decision tree-based ML algorithms are good at processing high-dimensional data, it is not easy for researchers to collect, clean, and integrate such high-dimensional data. In order to obtain more abundant predictive variables, more observations need to be sacrificed most of the time. Therefore, when using such algorithms for profit forecasting, how to ensure that the entire data processing process is controllable is still a challenge for researchers. Many ML models are trained on historical data and assume that past relationships will hold in the future. In rapidly changing markets, this reliance on historical patterns may not always yield accurate predictions.

ii. Model Selection and Overfitting Risks

Although ML models can capture complex, nonlinear relationships (as noted by Cao and You, 2024), this complexity can also make models prone to overfitting. When there are too many predictors or a high-dimensional feature space, as in the study by Jones et al. (2023), models might capture noise instead of true patterns, reducing predictive accuracy in out-of-sample tests. Different studies highlight the effectiveness of various models (e.g., Random Forests, Gradient Boosting Machine, KNN), but there is no one-size-fits-all solution. Choosing the most appropriate model is challenging, as performance can vary based on the dataset, feature selection, and model hyperparameters, which requiring extensive testing and optimization.

iii. Interpretability and Transparency

Machine learning models, especially nonparametric ones like ANNs and ensemble methods, often lack transparency and can be difficult to interpret for stakeholders. As explained in Jones et al., (2023), ML models uncover complex interactions among predictors that may not be straightforward to explain.

iv. Practical Application and Economic Value

While ML models may yield more accurate predictions than traditional models, translating these predictions into economically significant gains is not guaranteed (Jones et al., 2023). For example, in portfolio return analysis, ML predictions did not always result in superior abnormal returns compared to traditional regression-based models. This limitation indicates that improved forecast accuracy does not always correlate with better investment outcomes. Financial markets are dynamic, and earnings predictors may change in relevance over time. While ML models can adapt to changes, the robustness of these models across different economic conditions and market cycles remains a concern, as highlighted by studies like Hunt et al. (2022), which suggest the need for ongoing refinement and testing over time.

5. Opportunities of using ML to Predict Future Earnings

Challenges also mean opportunities. Next, I will discuss where future research can be carried out.

i. Exploration of New Data Types

When exist studies mainly focused on financial data, future studies could investigate the inclusion of alternative data sources such as text sentiment, customer reviews, and macroeconomic indicators. Such data has shown promise in other areas of finance and could enrich earnings forecasts by capturing more dimensions of market and firm sentiment. Future research can explore techniques to analyze these data types, perhaps using natural language processing (NLP) alongside traditional financial metrics.

ii. Forecasting Earnings in Longer Horizon

Many current models focus on short-term (one-year-ahead) earnings predictions. Research can explore ML methods in a longer forecast horizons, which allowing analysts to consider broader economic cycles.

iii. Incorporate Forward-looking Information

Most studies reviewed in this paper rely on historical financial ratio to predict future earnings and earnings changes. With more and more firms disclose their perceptions of future risk and opportunities in their annual report, researchers should utilize those forward-looking information to strengthen the prediction of earnings. The most promising approach might lie in hybrid models that combine both historical financial ratios and forward-looking information. By integrating structured historical data with unstructured forward-looking disclosures, such models can leverage the consistency of past performance data and the adaptability of current expectations.

6. Emerging Trend in Earnings Prediction: Large Language Models

With the development of Large Language Models (Hereafter, LLMs), many studies incorporate this method into financial analysis. [6]

Kim et al. (2024) investigate the capabilities of LLMs, such as GPT–4, in financial statement analysis. The study provides LLMs with structured, anonymized financial statements and uses a Chain-of-Thought prompting technique to simulate the analytical process of human financial experts, excluding any narrative inputs. Specifically, they evaluate the performance of LLMs in predicting the direction of future earnings using a two-step approach. First, corporate financial statements are anonymized and standardized to eliminate potential bias from the model's memory of specific companies or time periods. Company names are removed, years are replaced with labels (e.g., t and t–1), and financial statements are reformatted to align with Compustat's balancing model, ensuring consistency across firm-years. In the second step, they employs carefully designed prompts to guide the LLMs in performing financial analysis. Alongside a simple prompt, a Chain-of-Thought[7] (CoT) prompt is introduced to emulate the analytical process of human financial experts. The CoT prompt directs the model to identify trends in financial statement line items, compute key financial ratios (e.g., operating efficiency, liquidity, leverage), synthesize this information, and predict whether next year's earnings will increase or decrease. This structured prompting effectively mirrors the reasoning process used by professional analysts, enabling the LLMs to simulate complex financial analysis tasks.

Kim et al. (2024) demonstrate that LLMs can outperform analysts in predicting the direction of future earnings, particularly in scenarios where analysts are prone to bias or disagreement. LLMs complement both human analysts and ML models. They perform better than humans when additional narrative context is unnecessary and outperform quantitative ML models in areas like analyzing loss-making firms, showing "human-like" qualities. Conversely, they exhibit "machine-like" tendencies by excelling with larger firms. Surprisingly, GPT–4's performance rivals advanced ML models like ANNs and exceeds them in certain contexts. Additionally, the narrative analysis generated by the model adds substantial informational value.

Kim et al. (2024) highlights GPT–4's ability to derive insights from trends and financial ratios, emphasizing its broad reasoning capabilities over memory-based performance. A trading strategy based on GPT–4's predictions outperformed strategies using traditional ML models, yielding higher Sharpe ratios and alphas. The findings suggest that general-purpose LLMs can democratize financial analysis by offering state-of-the-art performance without specialized training. While LLMs have the potential to act as central elements in financial decision-making, the study calls for further exploration into the broader implications of AI-driven financial analysis. At the end of their study, they also commented that while LLMs can mimic human reasoning through chain-of-thought prompts, the underlying mechanics of their decision-making are not always clear, particularly when predicting

[6] Recent studies using LLMs to imply a wide range of tasks, including summarization of complex disclosures, sentiment analysis, information extraction, report generation, compliance verification, etc. Please refer to Kim et al. (2024) for comprehensive review of studies on this topic.

[7] Chain-of-thought in LLMs refers to a technique that involves prompting the model to break down complex reasoning tasks into a series of intermediate steps, mimicking human-like logical reasoning.

complex financial outcomes. It remains unclear which specific elements within the prompt are essential for achieving great performance (Kim et al. 2024).

7. Conclusion

This study reviewed recent accounting literature to explore the application of ML in earnings prediction. The findings indicate that most studies concentrate on using ML to predict changes in earnings, with portfolio returns based on these predictions significantly outperforming those derived from traditional regression methods. In contrast, fewer studies focus on forecasting exact earnings levels, likely due to the complexity and lower predictive accuracy of such tasks in prior research. This divergence highlights a critical debate on whether predicting directional changes provides more economic value than forecasting specific level of earnings.

The use of ML for earnings prediction is not without challenges. Key obstacles include data quality and consistency issues, the risk of overfitting in high-dimensional datasets, and the limited interpretability of complex models. Moreover, translating improved prediction accuracy into economic gains remains an open question, warranting further investigation.

Despite these challenges, this field presents exciting opportunities for future research. One can leverage new data types, such as textual information from corporate disclosures or macroeconomic indicators, to enhance model performance. Additionally, extending forecasting horizons to incorporate long-term trends and integrating forward-looking information, such as management forecasts and risk disclosures, could further boost predictive accuracy and relevance.

An emerging and promising trend in earnings prediction research is the application of Large Language Models (LLMs), such as GPT–4. Kim et al. (2024) shown that LLMs can outperform human analysts in predicting earnings direction, particularly in scenarios prone to analyst biases or disagreements. LLMs also complement traditional ML models, excelling in specific contexts like analyzing loss-making firms or larger companies. Their ability to process structured financial data and derive insights without specialized training underscores their potential as transformative tools in financial analysis. Overall, LLMs represent a significant innovation in earnings prediction, offering a path to democratize financial analysis and bridge gaps in traditional methods. By addressing existing challenges and exploring these new opportunities, the integration of advanced ML techniques and LLMs could fundamentally reshape the landscape of earnings prediction and financial decision-making.

References

[1] Anand, V., Brunner, R., Ikegwu, K., & Sougiannis, T. (2019). Predicting Profitability using Machine Learning. *SSRN Journal.* 3466478.

[2] Campbell, J. L., Ham, H., Lu, Z., & Wood, K. (2023). Expectations Matter: When (not) to Use Machine Learning Earnings Forecasts. *SSRN Journal.* 4495297.

[3] Cao, K., & You, H. (2024). Fundamental Analysis via Machine Learning. *Financial Analysts Journal*, 80(2), 74–98.

[4] Chattopadhyay, A., Fang, B., & Mohanram, P. (2022). Machine Learning, Earnings Forecasting, and Implied Cost of Capital – US and International Evidence. Working Paper.

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

[6] Cui, X., Xu, Z., & Zhou, Y. (2020). Using Machine Learning to Forecast Future Earnings. *Atlantic Economic Journal*, 48, 543–545.

[7] Easton, P. D., Kapons, M. M., & Monahan, S. J. (2024). Forecasting Earnings Using k-Nearest Neighbors. *The Accounting Review*, 99(3), 115–140.

[8] Hou, K., Van Dijk, M. A., & Zhang, Y. (2012). The Implied Cost of Capital: A New Approach. *Journal of Accounting and Economics*, 53(3), 504–526.

[9] Hunt, J. O. S., Myers, J. N., & Myers, L. A. (2022). Improving Earnings Predictions with Machine Learning. *Accounting Horizons*, 36(1), 131–149.

[10] Jones, S., Moser, W. J., & Wieland, M. M. (2023). Machine Learning and the Prediction of Changes in Profitability. *Contemporary Accounting Research*, 40(4), 2643–2672.

[11] Kim, A., Muhn, M., & Nikolaev, V. (2024). Financial Statement Analysis with Large Language Models. *arXiv preprint arXiv: 2407.17866*.

[12] Penman, S. H., & Zhang, X. (2004). Modeling Sustainable Earnings and P/E Ratios using Financial Statement Information. Working Paper, Columbia University and University of California, Berkeley.

[13] Van Binsbergen, J. H., Han, X., & Lopez-Lira, A. (2023). Man versus Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases. *The Review of Financial Studies*, 36(6), 2361–2396.

[14] Yakabi, K., Kuroki, Y., & Nakagawa, K. (2024). Predicting Earnings Change Using Machine Learning with Data from Japanese Companies. *Jsaisigtwo.* Fin–033, 68–75. (In Japanese)