Quantifying ESG Alpha Using Scholar Big Data: An Automated Machine Learning Approach

Qian Chen
qc2231@columbia.edu
Ion Media Networks & Statistics, Columbia University
New York City, New York

Xiao-Yang Liu*
xl2427@columbia.edu
Electrical Engineering, Columbia University
New York City, New York

ABSTRACT

ESG (Environmental, social and governance) alpha strategy that makes sustainable investment has gained popularity among investors. The ESG fields of study in scholar big data is a valuable alternative data that reflects a company's long-term ESG commitment. However, it is considered a difficulty to quantitatively measure a company's ESG premium and its impact to the company's stock price using scholar big data. In this paper, we utilize ESG scholar data as alternative data to develop an automatic trading strategy and propose a practical machine learning approach to quantify the ESG premium of a company and capture the ESG alpha. First, we construct our ESG investment universe and apply feature engineering on the companies' ESG scholar data from the Microsoft Academic Graph database. Then, we train six complementary machine learning models using a combination of financial indicators and ESG scholar data features and employ an ensemble method to predict stock prices and automatically set up portfolio allocation. Finally, we manage our portfolio, trade and rebalance the portfolio allocation monthly using predicted stock prices. We backtest our ESG alpha strategy and compare its performance with benchmarks. The proposed ESG alpha strategy achieves a cumulative return of 2,154.4% during the backtesting period of ten years, which significantly outperforms the NASDAQ-100 index's 397.4% and S&P 500's 226.9%. The traditional financial indicators results in only 1,443.7%, thus our scholar data-based ESG alpha strategy is better at capturing ESG premium than traditional financial indicators.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Ensemble methods**; *Semi-supervised learning settings*; *Latent Dirichlet allocation.*

KEYWORDS

ESG alpha, scholar data, alternative data, AI in finance, quantitative investment

*Corresponding author.

ICAIF '20, October 15–16, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7584-9/20/10…$15.00
https://doi.org/10.1145/3383455.3422529

ACM Reference Format:
Qian Chen and Xiao-Yang Liu. 2020. Quantifying ESG Alpha Using Scholar Big Data: An Automated Machine Learning Approach. In *ACM International Conference on AI in Finance (ICAIF '20), October 15–16, 2020, New York, NY, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3383455.3422529

1 INTRODUCTION

ESG (Environmental, social, governance) factors are widely known as the three primary factors in measuring the sustainability and societal impacts of an investment in a company or business. ESG investing is the practice of considering ESG factors in the investment process, which helps better predict companies' future financial performance and is now attracting more attention [11]. Nowadays more institutional investors begin to recognize the importance of ESG investing. A range of approaches and investment teams are encouraged to align with the ESG approaches [7, 11].

However, it is considered difficult for investors to quantify the impacts of ESG commitment on companies' stock performance, because the companies' earning calls do not directly reveal their ESG commitments. How to quantify companies' ESG premium that is rewarded by their ESG commitment? How can we leverage the characteristics of ESG factors to quantify the impacts of ESG commitment on companies, and what's more, how to quantify the relationship between ESG research and stock performance?

Traditional approaches of quantifying ESG commitment have many limitations. Here are three approaches that are widely used by investors. First, many investors integrate the companies' sustainable business strategies as ESG factors through their websites, earning calls or news [7]. This approach is straightforward, but subjective, thus is less convincing than quantitative methods. Second, many data providers offer a single ESG score for an individual security [8]. Many investors leverage these ESG scores to evaluate ESG factors of stocks. However, it could be very difficult to justify these scores. The definition and extent of "socially responsible" to one investor may not be the same to another. Third, some investors construct ESG metrics using ESG-related data like carbon emissions or percentage of their energy consumption derived from renewable sources [11]. However, many ESG factors are self-reported, resulting in big difference in data availability. Therefore some ESG factors are not sufficient to cover the investment universe.

ESG scholar data is a good reflection of ESG research commitment. Companies that commit resources into ESG research gain profits from their continuous ESG efforts in the long run [3]. The primary reasons are:

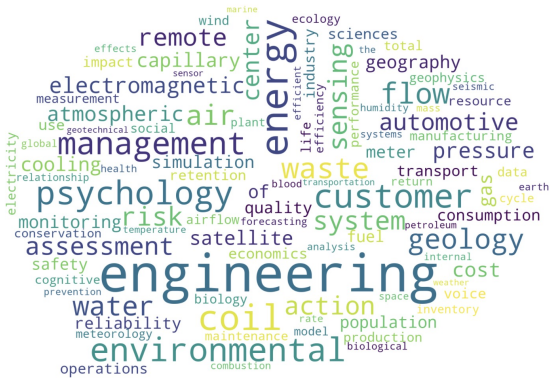

Figure 1: Word cloud of ESG keywords.

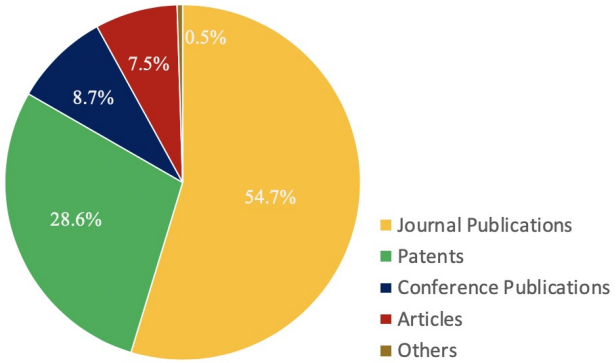

Figure 2: Composition of the scholar dataset.

- Development of sustainable business practices increases productivity and reduces costs, which is attractive to both shareholders and consumers [5].
- Encouraging ESG research not only improves companies' brand awareness, but also improves brand image and competitive advantages [1].
- Integrating ESG research commitment into business and investment allows companies to achieve strategic freedom by easing regulatory pressure in a timely manner [7].

Industry leading companies have invested growing amount of resources for ESG purposes that in general could benefit their business. For example, by April 2019 Bank of America had its total commitment in the Environmental Business Initiative over $445 billion [23], in order to maintain a sustainable business development. Apple Inc had reached its target of 100% renewable energy across the company's operations, spanning 43 countries around the globe by the end of 2018 [18]. These ESG commitments and accomplishments empower the sustainable business development, improve brand awareness and make the companies more attractive to investors and consumers, and of course impact the stock performance.

In this paper, we propose a novel approach to quantitatively measuring a company's ESG research commitment and capturing ESG alpha using an automatic machine learning method on a combination of ESG scholar data as alternative data source and traditional financial indicators. First, we create a feature space with 48 scholar features and 10 financial features by performing feature engineering on alternative data and earning calls. Second, we implement a list of machine learning methods to predict the monthly return of all 88 stocks in the selected investment pool and build up our portfolio using top 20% stocks with highest predicted monthly returns, set stock allocation with equal weights. We trade them using 'buy-and-hold-long-only' strategy [9] and re-balance the portfolio monthly. Third, we evaluate our ESG alpha strategy by comparing the cumulative return and annualized Sharpe ratio of the ESG alpha strategy with a baseline strategy that uses financial indicators only and benchmark with both the S&P 500 index and NASDAQ-100 index. The proposed ESG alpha strategy achieves a cumulative return of 2,154.4% during the backtesting period of ten years(2009-2018), which significantly outperforms the NASDAQ-100 index's 397.4% and S&P 500's 226.9%. The traditional financial indicators approach results in only 1,443.7%, which indicates that the scholar data driven ESG alpha strategy is better at capturing ESG premium than traditional financial indicators.

The remainder of this paper is organized as follows. Section 2 describes how alternative data is collected and processed and how the stock pool is determined. Section 3 presents a practical machine learning approach to capture ESG alpha. Section 4 provides backtesting and performance evaluation. We conclude this paper in Section 5.

2 SCHOLAR DATA FOR ESG ALPHA

The ESG alpha in scholar data serves as a quantitative measurement of companies' ESG commitment, which is closely related with their stock prices.

2.1 Scholar Data as ESG Alternative Dataset.

Alternative data is being used by fundamental and quantitative investors to capture innovative sources of alpha. It is able to provide information, such as companies' ESG commitment [16], that we cannot see from earning reports or other traditional metrics.

There are three reasons to leverage scholar data as a valuable alternative data to capture the ESG alpha:

- ESG commitment impacts business gradually, meaning it will take some time before the ESG commitment comes into effect on stock prices [11], while academic research also does not impact business immediately.
- ESG commitment will benefit the company development cumulatively. It can impact business over a long period of time [11, 22]. Similarly academic research also has a cumulative contribution on business growth.
- Companies are now paying more attention to ESG commitment each year since 2000, especially more investment on academic research [4].

The data we use is from the Microsoft Academic Graph database [26]. It is an open resource database with records of publications, including papers, journals, conferences, books, etc. It provides the demographics of the publications like public date, citations, authors and affiliated institutes. The Microsoft Academic Graph database

now indexes more than 221 million publications, 241 million authors and 25,000 institutions, containing companies from a wide range of industries. More importantly, it includes ESG publication records dating back to 1970s - long enough to study the relationship between ESG publications and companies' stock prices.

2.2 Data Collection

The first task is to filter out ESG scholar data. We start by selecting an ESG stock pool to narrow down our investment universe. It is reasonable since one may not have a large capital for hundreds of stocks. The selected stock pool has two characteristics:

- Diverse: The ESG stock pool is selected from a wide range of traded stocks from the New York Stock Exchange (NYSE) and NASDAQ. It is based on a list of ESG related indexes in the market [8], such as iShares Global Telecom ETF (Exchange-traded fund) and iShares Nasdaq Biotechnology ETF, which covers companies from various industries.
- Typical: The ESG stock pool consists of companies that have been investing academic research on a long-term basis. All companies in our ESG stock pool have publication records in the Microsoft Academic Graph database during the back-testing period of 2009-2019 [19].

It turns out that ESG stock pool consists of 88 publicly traded stocks from the New York Stock Exchange (NYSE) and NASDAQ. We provide the list of companies in our Github link [6]. We select the publications when one of the authors is employed by the company in our ESG stock pool. We collect a raw scholar dataset with 722,070 publication records in total.

2.3 ESG Topic Filtering

In this section we filter out ESG publications from the raw scholar data using semi-supervised learning method. The Microsoft Academic Graph database contains publications from a large variety of fields. We apply ESG topic filtering on the raw scholar dataset from previous step to exclude non-ESG-related publications. However, there is no indicators in the Microsoft Academic Graph database that can filter out ESG related research, since ESG is not a subject with clear definition and keywords, making it a hard target to classify. Semi-supervised learning method can solve this problem by manually add labels to a small proportion of the unlabeled dataset.

First, we use unsupervised clustering to perform topic filtering on the raw scholar data. Unsupervised cluster matching, like topic model, is powerful when filtering publications in different domains without correspondence information. We choose Latent Dirichlet Allocation (LDA) [25], a topic model that is able to find correspondence between publications given their titles and keywords without topic labels. In order to provide guidance to the topic model, we add labels to a small portion of the raw scholar data that we believe are ESG publications. This is done by manually labeling papers and publications with domain knowledge. Those labeled publications boost the accuracy of the ESG filtering model [33]. We tune the LDA model so that the labeled ESG publications can be correctly classified as the same topic.

The ESG topic filtering process separates 722,070 records in the raw scholar dataset into 30 latent groups and 4 of them are related to ESG topics. Figure 1 illustrates the word cloud of keywords in our latent topics. We use publications in these 4 latent topics as our ESG scholar data. It has 65,101 records in total, dating back to 1970. Figure 2 shows the composition of our ESG scholar dataset. 54.7% of the records are journal publications, 28.6% are patents, 8.7% are articles, 7.5% are conference publications and 0.5% are others.

2.4 Feature Engineering for ESG Scholar Data

We perform feature engineering on ESG scholar data to measure the quantity and the quality of ESG research. Since both ESG scholar data and stock prices are time-sensitive, we follow two rules to prevent any causality conflict.

First, for each publication, the number of citations is counted on a monthly basis, and one cannot use the cumulative sum of citations. For each month a citation counts into the monthly citation if only the citation happens within the current month. Therefore, our strategy will not use on citations from a future time. In order to do that, we go through all the publications that have cited publications in our ESG scholar data. Then we perform the aggregation based on the publication date of those cited-by publications.

Second, the scholar data is lagged by one month from the financial data, which is reasonable. Since it usually takes a period of time before any research result goes into production and takes effect on business [17]. It can avoid estimating the impact of ESG commitment too soon and help generate more reliable results.

There are five types of publications. For each type we extract 8 features as follows.

- First, we calculate each company's monthly total number of each type of publications and their sum of citations in order to capture the development of different types of ESG publications. It results in 2 features.
- Second, we bring in relative sum of publications to eliminate the bias of different ESG topics, because we observe imbalanced development among 4 ESG topics, which we get from topic filtering model in Section 2.3. For each publication item, its relative count equals one divided by count of publications last year under the exact same ESG topic. It results in 1 feature.
- Third, we want to describe the general development of ESG research among all companies in ESG stock pool using the average level and the top level of ESG development. Citations can be used as a measurement of quality of a research result. We calculate monthly average number of citations to describe the average level of ESG research each month. And we aggregate maximum number of citations each month to capture breakthrough research, because a publication with great number of citation usually is of big value. Besides, we create features to measure the cumulative effect of ESG research. This results in 2 features.
- We calculate monthly cumulative sum for count of publications, sum of citations and relative count of publications. This results in 3 features.

We calculate 8 features for five publication types separately, and also 8 features for all these publication types. Therefore in total we result in a scholar data feature space with 48 features in total.

3 PRACTICAL MACHINE LEARNING APPROACH

In this section we present a machine learning approach to quantify ESG alpha. First we prepare the training data using ESG scholar data and financial indicators. Then we train 6 machine learning models and integrate the results using an ensemble method. Finally we describes our rolling-window backtesting process.

3.1 Financial Indicators

Besides 48 ESG scholar features, we add 10 financial indicators into the feature space. We select financial indicators that are able to represent a company's fundamental conditions and predict the stock price [13, 24]. The predictive model that trained by using financial indicators alone could serve as a baseline model when creating a strategy. Afterwards a new predictive model trained by using scholar data together with financial indicators will be evaluated and compared with the baseline.

Similar to [10, 32], we select Price-to-Sales ratio (P/S), Price-to-Book ratio (P/B) and Price-to-Equity ratio (P/E) to evaluate if the current stock price is reasonable. We select Current Ratio (CR) and Quick Ratio (QR) to evaluate the short term liquidity. We select Earnings-per-Share (EPS), Return-on-Assets (ROA), Returns-on-Equity (ROE) and Net Profit Margin (NPM) to evaluate the profitability of a company. And we also use Debt-to-Equity ratio (D/E) to check the long term condition of the company.

All financial indicators including stock prices come from Compustat database via Wharton Research Data Services [28, 29] and Yahoo Finance website.

3.2 Machine Learning Methods

Suppose at time step t, the logarithmic ratio of stock return (log return) of the current period r_t is defined as:

$$r_t = \log\left(\frac{P_t}{P_{t-1}}\right), \tag{1}$$

where P_t is stock price in time t. Our task is to predict log return of the following period r_{t+1} given the feature vector X_t. The relationship between r_{t+1} and X_t is described as:

$$r_{t+1} = f(X_t) + \epsilon_t, \tag{2}$$

.

In this section we choose 6 machine learning algorithms to model the relationship, inspired by [10, 32]. Three of them are linear regression and its improvements, and the other three are LSTM, Random Forest (RF) and Support Vector Regression (SVR). The essential task of these complementary models are to capture the patterns from different perspectives of the data so that we can use an ensemble method to aggregate the best prediction of stock prices.

3.2.1 Linear Regression & Improvements. Linear regression [12] is simple to implement and easy to interpret. We can describe the underlying relationship between log return r_{t+1} and X_t with an error term ϵ_t

$$r_{t+1} = \alpha + X_t\beta + \epsilon_t, \tag{3}$$

and use residual sum of squares

$$\text{RSS} = \sum_{t=1}^{n}(r_{t+1} - f(X_t))^2, \tag{4}$$

as the cost function. In order to cast penalty to the feature space to avoid over-fitting, we add two regularizers *Lasso* and *Ridge* into the cost function, which are specified as follows,

$$\text{RSS}_{\text{Lasso}} = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|,\ \text{RSS}_{\text{Ridge}} = \text{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2, \tag{5}$$

where β is the vector of coefficients and p is the number of features in the training data. Since the feature matrix is large and sparse, Lasso regression can reduce the dimension of feature space by leaving out insignificant features.

3.2.2 Support Vector Regression. Support Vector Regression (SVR) with Radial Basis Function (RBF) as kernel can deal with training data of high dimensions [2]. The SVR model depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction. The RBF kernel works well when decision boundary is not linear and the parameters are easy to tune. Since we have already implemented linear regression, we use SVR to deal with data that is not linearly predictable.

3.2.3 Random Forest. Random Forest is an ensemble method aggregating decision trees using sampling feature space and sampling training data [21]. We use regression trees to form random forest, which also use RSS as the cost function. The biggest advantage of random forest model is because it prevents over-fitting so it can deal with large feature space and reduce out-of-sample variance with high flexibility.

3.2.4 LSTM. Recurrent neural network (RNN) is powerful when dealing with sequential and time series data. Long Short Term Memory networks - usually referred as 'LSTM' - are an effective gradient-based method, capable of dealing with the vanishing gradient problem in long term dependencies of RNN. They were introduced by Hochreiter & Schmidhuber (1997) [14], and were refined and popularized by many people in following works [15, 30]. A typical LSTM network contains an input layer, an output layer and one or more hidden layers. The hidden layer does have the ability to remove or add information to the cell state, carefully regulated by structures called gates.

Now we want to predict stock return r_{t+1} with feature vector X_t which consists of scholar data and financial indicators. We use the forget gate:

$$\Gamma_t^f = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f), \tag{6}$$

to remove the previously stored memory state of stock return, where W_f and b_f are weights and bias of the gate, $[h_{t-1}, X_t]$ is the concatenate of h_{t-1} and X_t, and σ is the Sigmoid function that controls the memory state. The equation will output a vector Γ_t^f which is the forget state at time step t with values between 0 and 1. Second we use input gate:

$$\Gamma_t^i = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i), \tag{7}$$

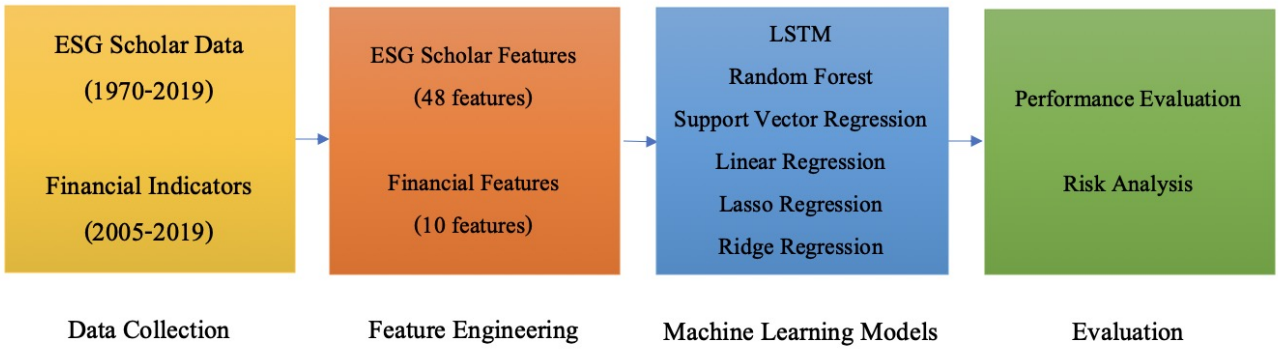

Figure 3: Overview of our ensemble scheme.

to import information from current step into the memory cell. Similarly, W_i and b_i are weights and bias of the input date and the output vector Γ_t^i sits between 0 and 1, which stands for how much we decided to update each state value. Then a tanh layer will add a candidate value to the state, and it is calculated as:

$$\tilde{c_t} = \tanh(W_c \cdot [h_{t-1}, X_t] + b_c). \tag{8}$$

Next we can update the old cell state to the new one:

$$c_t = \Gamma_t^f \cdot c_{t-1} + \Gamma_t^i \cdot \tilde{c_t}. \tag{9}$$

Finally we put the state through output gate:

$$\Gamma_t^o = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o), \qquad h_t = \Gamma_t^o \cdot \tanh(c_t), \tag{10}$$

For evaluation we use Mean Absolute Error as loss function because of its robustness to outliers.

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |r_{t+1} - f(X_t)|, \tag{11}$$

where n is number of training data points, r_{t+1} and $f(X_t)$ are true values and predicted values of the stock return at time step $t+1$.

3.3 Ensemble Method and Backtesting Methodology

We use an ensemble method to aggregate all six models and select the model with best performance to trade. Ensemble learning is primarily used to improve the performance of a model, or reduce the likelihood of an unfortunate selection of a poor one. In our scenario the ensemble method can automatically form the best portfolio allocation by selecting best model for prediction. During the back-testing period, we first train and validate all algorithms concurrently with a rolling window, and then use the best model as the predictive model for the trading month after the rolling window. The evaluation metrics that we use to decide the best model is Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |r_{t+1}^{val} - f(X_t^{val})|, \tag{12}$$

which is a general metric for out-of-sample predictive ability [27].

We use a rolling backtesting window for both training and validation process as shown in Fig. 4. The training window is 48 months long and the validation window is 12 months long. We use the following month to trade, and re-balance our ESG portfolio by the end of each trading month. The lag between two successive rolling window is 1 month. The ensemble method will provide the aggregated estimation of the monthly return of each stock in the ESG stock pool. We use the monthly return estimation to build up our ESG portfolio. We select top 20% performance stocks and then trade using equal weights.

Our rolling window backtesting processes as follows:

- Step 1. In each rolling window we train six machine learning models on the 48-month training window and validate the models on the 12-month validation window followed by the training window. For each model we calculate the Mean Absolute Error (MAE) of the validation window.
- Step 2. We choose the best model which has the lowest MAE of the validation window to predict the stock prices of the following backtesting window. Then we use the selected model to predict the stock return of the backtesting month for all ESG stocks. Fig. 5 shows the selection of best model among all backtesting months. From the table we can see that LSTM has been selected the most times as the best model.
- Step 3. We select top 20% performance stocks provided by the best predictive model as our ESG portfolio. Then we trade our ESG portfolio with equal weights.

Table 1 shows the model usage in rolling-window process while training our ESG alpha strategy. From the result we can see LSTM model is selected as the best model in 36.7% of the backtesting period, which ranks the first in among all models.

4 PERFORMANCE EVALUATIONS

In this section we evaluate the performance of ESG alpha. First we introduce the benchmark portfolio we use to compare with ESG alpha strategy. And we describe the evaluation metrics and the trading setup. Then we compare ESG alpha strategy with benchmark portfolios and indexes in Section 4.2 and 4.3.

4.1 Trading Setup & Benchmark Portfolios

The backtest trading period contains 121 months, which goes from 2009-01 to 2019-01. We compare the strategies on two perspectives:

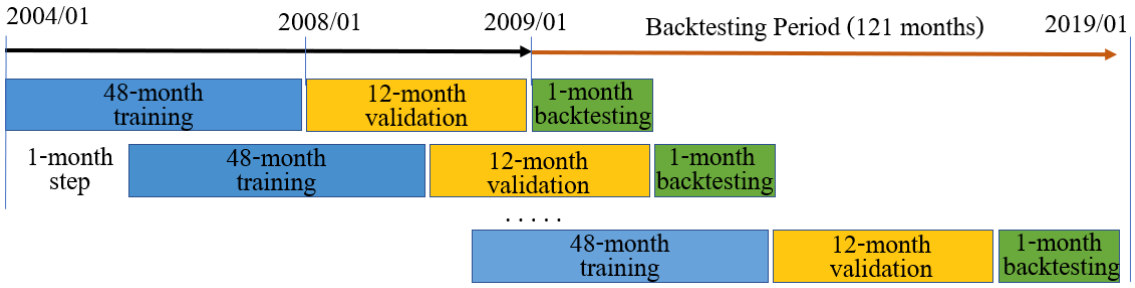

Figure 4: Rolling window in the backtesting procedure.

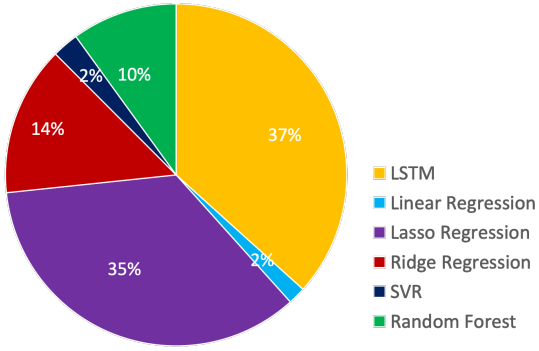

Figure 5: Model usage for trading.

return and risk. The metrics we use are log cumulative return and annualized Sharpe ratio. We also use other metrics including annual return, volatility, max drawdown and daily value at risk to compare the performance between strategies. Table 1 shows the performance evaluation metrics of all five strategies during our backtesting period 2009-2019 and Figure 6 shows the comparison of log cumulative returns.

The benchmark indexes we use are S&P 500 and NASDAQ-100 so we can compare with the market. We also build up another two benchmark portfolios. One is called equal weight benchmark. It longs and holds all 88 stocks in our investment universe with equal weights each month. The other one is called financial indicators only benchmark. It only uses financial indicators to predict stock prices then selects top 20% performance stocks and trades with equal weights. We can see how much ESG scholar data contributes to the performance of ESG alpha strategy by comparing ESG alpha strategy to financial indicators only portfolio. Our codes are available online [6].

4.2 Comparison with Equal Weights Portfolio and Benchmark Indexes

We first compare ESG alpha strategy with benchmark indexes S&P 500 and NASDAQ-100 to show the overall profitability and risk by evaluating the cumulative return and Sharpe ratio. The ESG alpha strategy outperforms the cumulative return of NASDAQ-100 index by 397.4% and S&P 500 by 226.9% respectively. The annualized Sharpe ratio of ESG alpha greatly outperforms the benchmarks. The ESG alpha reaches an annualized Sharpe ratio of 1.97, which is greater than equal weights baseline of 1.75 and much greater than benchmark indexes.

Then we compare ESG alpha strategy with equal weight benchmark. In Figure 6, the cumulative return curve of ESG alpha is lower than the equal weights portfolio in the first two years. But ESG alpha develops more steadily and the curve is more flat than the equal weights benchmark. Then after 2011-06, the cumulative return of ESG alpha starts to beat the equal weights portfolio. It aligns with our intuition of ESG alpha, that ESG scholar research helps companies develop a sustainable business and steady growth trend. Figure 7 shows the detailed return by month and year and shows the return distribution of the scholar driven strategy. It is apparent that except 2013, the rest of the annual returns are steady and consistent in terms of direction, which usually indicates more value for a stable portfolio than other portfolios with unstable factors that lower the actual Sharpe ratio because of higher risks.

4.3 Comparison with Financial Indicators Only Portfolio

We compare ESG alpha strategy with the financial indicators only portfolio to see the contribution of ESG features in the proposed strategy. In Table 1, ESG alpha provides a huge increase in the cumulative return and annualized Sharpe ratio of the portfolio to the financial indicators only portfolio. In Figure 6, the cumulative return curve of ESG alpha starts to beat the financial indicators only portfolio since 2011-06. By the end of the backtesting period the ESG alpha achieves an annualized Sharpe ratio of 1.97 while financial indicators only portfolio gets 1.77. Figure 8 shows the rolling Annualized Sharpe Ratio comparison of ESG alpha and financial indicators only portfolio and during the majority of backtesting period the rolling annualized Sharpe ratio of ESG alpha is above the financial indicators only benchmark. This proves that ESG alpha achieves higher risk-adjusted return compared to strategy that use financial indicators only.

5 CONCLUSION

The efforts and policies made for ESG purposes may not be tracked by traditional financial indicators, but they still have a huge impact on the business development and stock performance of the company. Making good use of ESG components will make a big

Table 1: Performance metrics

(2009-2019)	Scholar Alpha	Equal Weights	Financial Indicators Only	NASDAQ-100	S&P 500
Annual Return	37.0%	29.5%	31.8%	17.4%	12.6%
Cumulative Return	2,154.4%	1,221.8.0%	1,443.7%	397.4%	226.9%
Log Cumulative Return	3.11	2.58	2.73	1.60	1.18
Annual Volatility	24.8%	21.5%	24.5%	18.5%	16.4%
Annualized Sharpe Ratio	1.97	1.75	1.56	1.36	1.13
Max Drawdown	-29.4%	-26.3%	-29.1%	-23.0%	-27.1%
Daily Value at Risk	-3.0%	-2.6%	-3.0%	-2.3%	-2.0%

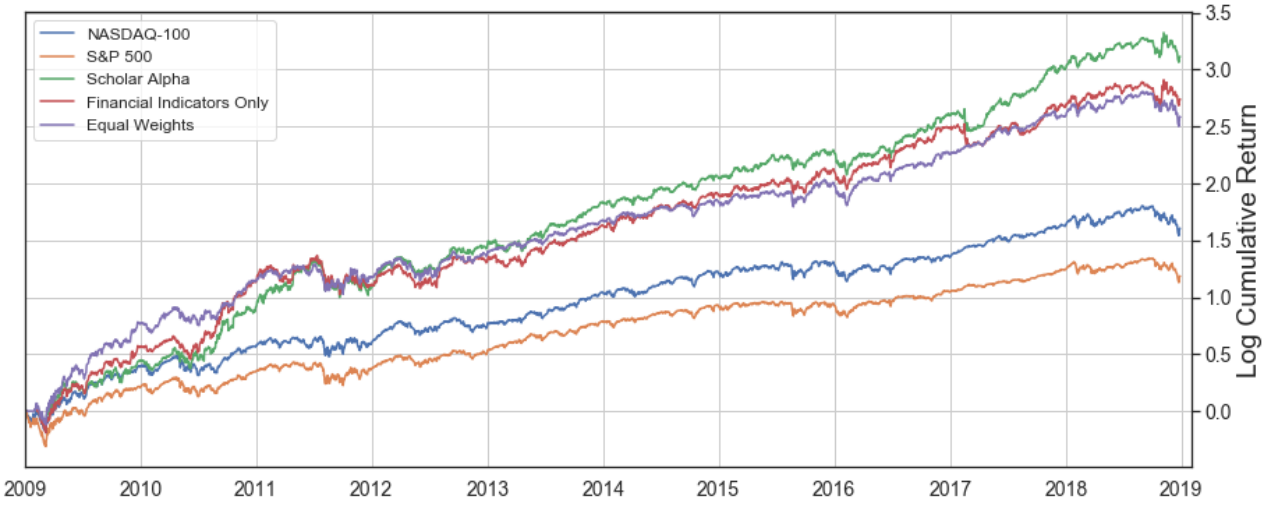

Figure 6: Log cumulative return during 2009-2019.

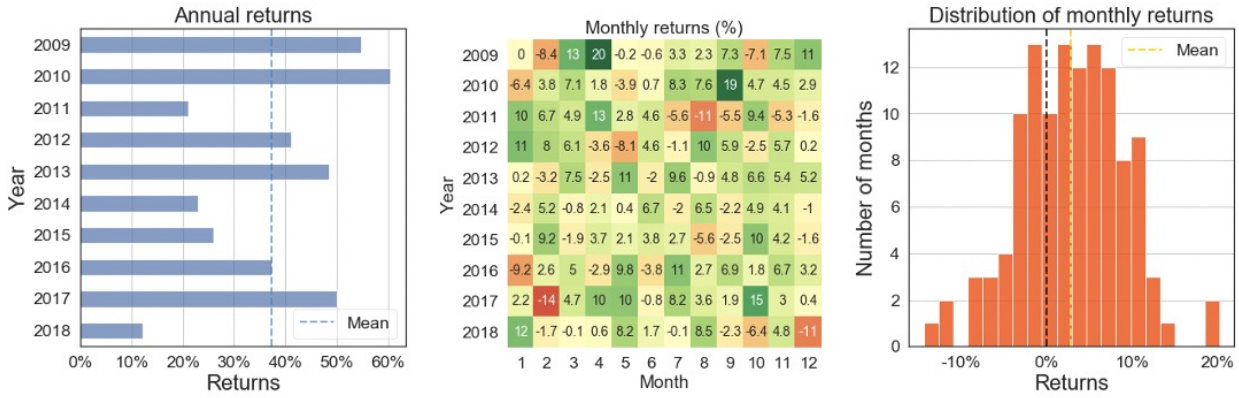

Figure 7: Return distribution.

difference on the investment. And ESG scholar data is a good reflection of the sustainable practice of business development and a good measurement of the ESG premium of a company. In this paper, we utilize ESG scholar data to generate alpha signals using practical machine learning approaches. The portfolio built by our ESG alpha strategy beats the benchmark indexes and baseline portfolio built based on traditional financial indicators only on cumulative return and rolling Sharpe ratio. The result shows that the ESG scholar alternative data can broaden the horizon of traditional investment space and generate greater profit by extracting ESG alpha from the scholar data. In the future we can expand the alternative data feature space. For example, news data also contains a lots of ESG

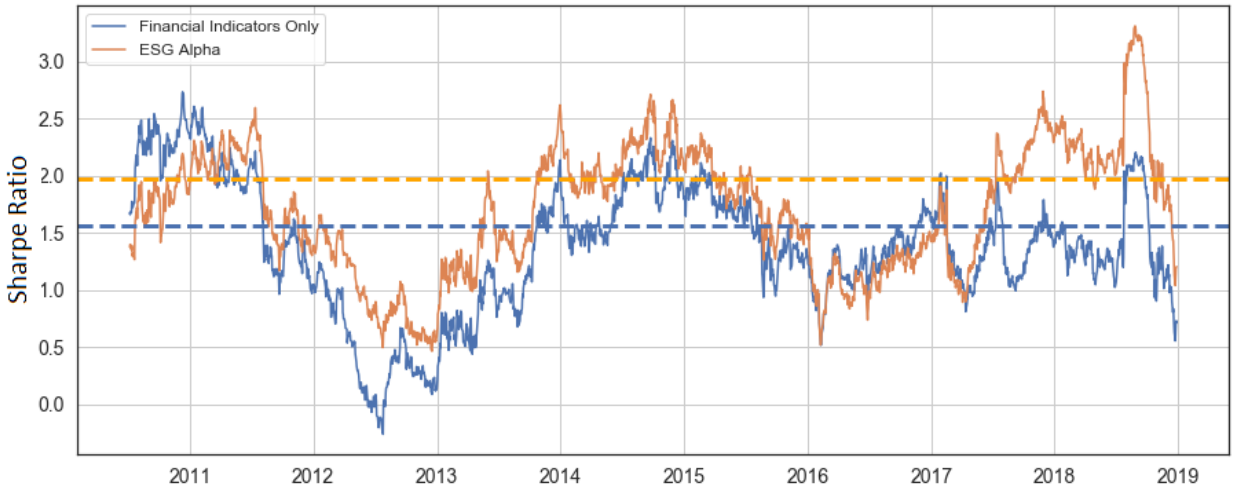

Figure 8: Rolling annualized Sharpe ratio.

topics[20]. We can also implement Deep Reinforcement Learning to do better trading, by using scholar data features in the state space or in the reward function[31], and achieving better performance by implementing advanced feature engineering and making better predictions of the market.

REFERENCES

[1] Güler Aras and David Crowther. 2016. *The durable corporation: Strategies for sustainable development.* CRC Press.

[2] Mariette Awad and Rahul Khanna. 2015. Support vector regression. In *Efficient Learning Machines.* Springer, 67–80.

[3] Jincheol Bae, Wonchang Hur, Jaehong Lee, and Jaimin Goh. 2017. Patent citations and financial analysts' long-term growth forecasts. *Sustainability* 9, 5 (2017), 846.

[4] Markus Beckmann, Stefan Hielscher, and Ingo Pies. 2014. Commitment strategies for sustainability: How business firms can transform trade-offs into win–win outcomes. *Business Strategy and the Environment* 23, 1 (2014), 18–37.

[5] Robert Boutilier. 2017. *Stakeholder politics: Social capital, sustainable development, and the corporation.* Routledge.

[6] Qian Chen. 2020. Quantifying ESG Alpha in Scholar Big Data An Automated Machine Learning Approach. https://github.com/chenqian0168/Quantifying-ESG-Alpha-in-Scholar-Big-Data-An-Automated-Machine-Learning-Approach

[7] John Elkington. 1994. Towards the sustainable corporation: Win-win-win business strategies for sustainable development. *California management review* 36, 2 (1994), 90–100.

[8] ETFdb. 2018. Top EST Investing ETFs to Buy. https://etfdb.com/esg-investing

[9] Eugene F Fama and Kenneth R French. 1996. Multifactor explanations of asset pricing anomalies. *The journal of finance* 51, 1 (1996), 55–84.

[10] Yunzhe Fang, Xiao-Yang Liu, and Hongyang Yang. 2019. Practical Machine Learning Approach to Capture the Scholar Data Driven Alpha in AI Industry. In *2019 IEEE International Conference on Big Data (Big Data).* IEEE, 2230–2239.

[11] Gunnar Friede, Timo Busch, and Alexander Bassen. 2015. ESG and financial performance: aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment* 5, 4 (2015), 210–233.

[12] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning.* Vol. 1. Springer series in statistics New York.

[13] Joseph J Gerakos and Robert Gramacy. 2013. Regression-based earnings forecasts. *Chicago Booth Research Paper* 12-26 (2013).

[14] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with LSTM. (1999).

[15] Felix A Gers, Nicol N Schraudolph, and Jürgen Schmidhuber. 2002. Learning precise timing with LSTM recurrent networks. *Journal of machine learning research* 3, Aug (2002), 115–143.

[16] Nina Gupta and Terry A Beehr. 1982. A test of the correspondence between self-reports and alternative data sources about work organizations. *Journal of Vocational Behavior* 20, 1 (1982), 1–13.

[17] Kathryn Rudie Harrigan and Yunzhe Fang. 2019. The financial benefits of persistently high forward citations. *The Journal of Technology Transfer* (2019), 1–29.

[18] Apple Inc. 2018. Environmental Responsibility Report. https://www.apple.com/environment/pdf/Apple_Environmental_Responsibility_Report_2018.pdf

[19] Microsoft Inc. 2020. ETAP-Institutions Microsoft Academic. https://academic.microsoft.com/institutions

[20] Xinyi Li, Yinchuan Li, Hongyang Yang, Liuqing Yang, and Xiao-Yang Liu. 2019. DP-LSTM: Differential privacy-inspired LSTM for stock prediction using financial news. *arXiv preprint arXiv:1912.10806* (2019).

[21] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

[22] Edwin Mansfield. 1972. *Research and innovation in the modern corporation.* Springer.

[23] Bank of America. 2018. Corporate Environmental Sustainability at Bank of America. https://newsroom.bankofamerica.com/press-releases/environment/bank-america-commits-300-billion-2030-low-carbon-sustainable-business

[24] Joseph D Piotroski. 2000. Value investing: The use of historical financial statement information to separate winners from losers. *Journal of Accounting Research* (2000), 1–41.

[25] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2012. The author-topic model for authors and documents. *arXiv preprint arXiv:1207.4169* (2012).

[26] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web.* 243–246.

[27] Cort J Willmott and Kenji Matsuura. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research* 30, 1 (2005), 79–82.

[28] WRDS. 2020. Compustat Daily Updates - Fundamentals Quarterly. https://wrds-web.wharton.upenn.edu/wrds/ds/compd/fundq

[29] WRDS. 2020. Financial Ratios Firm Level by WRDS. https://wrds-web.wharton.upenn.edu/wrds/ds/wrdsapps/finratiofirm

[30] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems.* 802–810.

[31] Zhuoran Xiong, Xiao-Yang Liu, Shan Zhong, Hongyang Yang, and Anwar Walid. 2018. Practical deep reinforcement learning approach for stock trading. *arXiv preprint arXiv:1811.07522* (2018).

[32] Hongyang Yang, Xiao-Yang Liu, and Qingwei Wu. 2018. A practical machine learning approach for dynamic stock recommendation. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE).* IEEE, 1693–1697.

[33] Xiaojin Zhu and Andrew B Goldberg. 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning* 3, 1 (2009), 1–130.