MICHIGAN ROSS

Machine Learning for Investing, Telecommunications Sector

Judele Bogdan | bobi@umich.edu | +40 733 986 015
Fabio Magnavita | fabiomag@umich.edu | +39 347 186 2338
Nicolas Newberry | nnicolas@umich.edu | +1 734 634 1562

Abstract

This project evaluates the predictive power of various quantitative techniques on future stock returns, focusing on the Communication Services sector, with an emphasis on stock characteristics like dividend yield and other variables. Our goal is to determine whether these characteristics can effectively anticipate stock performance and recommend one stock for the portfolio manager to purchase. We applied a range of statistical and machine learning models, OLS (Ordinary Least Squares), LASSO (Least Absolute Shrinkage and Selection Operator), Decision Tree, Random Forest, Gradient Boosting, and Neural Networks, to capture both linear and non-linear relationships between predictors and future returns. Each model was trained and validated using separate datasets, with hyperparameters optimized on the validation set to maximize predictive accuracy and minimize overfitting. We then used these models to predict stock returns for a specific month within the Communication Services sector, generating individual rankings of top stocks per model, which were combined into a final ranking featuring companies like Verizon, AT&T, GOGO, T-Mobile, and Lumen Technologies.

Our analysis revealed that dividend yield, the ratio of a company's annual dividends to its stock price, serves as a strong predictor of industry-adjusted returns in the Telecommunications subsector. We initially hypothesized that high dividend yields might signal limited reinvestment opportunities, as firms with fewer growth prospects often prioritize returning cash to shareholders over investing in expansion, potentially leading to lower returns. Conversely, we expected low dividend yields to indicate effective reinvestment strategies that could drive higher returns. Contrary to this hypothesis, we found a positive correlation between dividend yield and returns, suggesting that better-performing companies in this sector distribute higher dividends, likely reflecting financial stability and consistent cash flows rather than a lack of growth opportunities. Among the top-ranked stocks, we recommend that the portfolio manager purchase Verizon due to its strong performance across all models, consistently ranking in the top five stocks in the S&P 1500 Communications Sector. Verizon's attractive dividend yield of 6.5% and earnings yield of 4% were associated with high returns in the OLS and Neural Network models, while its ability to beat consensus forecasts with an SUE score (z-scores above analyst earnings estimates) of 2.02 was a key metric in Random Forest, Gradient Boosting, and Neural Network models. These models look at a variety of different factors for predicting high industry-adjusted returns, demonstrating Verizon's consistent outperformance and supporting its selection as a reliable investment in the Telecommunications subsector.

Contents

Tables

Figures

Introduction

This project investigates whether machine learning techniques can effectively predict future stock returns within the Communication Services sector, using a variety of financial and market-based indicators. We leveraged a comprehensive dataset spanning from 2006 to 2023, which includes over 880,000 monthly stock-level observations across a range of financial variables. To evaluate the predictive power of these factors, we implemented a suite of models: Ordinary Least Squares (OLS), LASSO, Decision Trees, Random Forest, Gradient Boosting, and Neural Networks. These models were trained and validated using a time-based data split designed to preserve market volatility characteristics, ensuring robust performance evaluation across economic cycles.

Our objective was not only to assess which methods produce the most accurate return predictions, but also to understand which financial indicators contribute most meaningfully to these forecasts. Ultimately, we used the best-performing models to generate a cross-validated ranking of top-performing stocks, and recommended a stock for an investor to purchase for the month of January 2023.

Data

Introduction

This dataset contains a set of financial and market variables, derived from a wide range of corporate and stock market data, with observations from 2006 to 2023. The dataset includes metrics from 14 teams such as dividends, ESG scores, insider buying, free cash flow yield, research and development expenditures, earnings yield, sales growth, short interest, and various financial ratios. Features like "lag1mcreal," the natural log of market capitalization, and "adjlag1bm," the book-to-market ratio, are widely researched financial indicators associated with future stock returns.

The dataset also includes numerous derived variables such as z-scores and percentile rank, which standardize or normalize raw financial data, enabling better comparisons across companies, time periods and different analysis techniques. Missing data in the dataset has been handled using dummy variables, and the final variables (those with the "fin" prefix) provide the missing values as zeros, while "miss" variables indicate the absence of data. These preprocessing steps ensure the dataset remains consistent and useful for time-series analyses.

Training, Validation and Testing Splits

To ensure a balanced and meaningful training-validation split for our analysis, we selected December 31, 2015, as the cutoff point. This choice provides a 60/40 split of the data while maintaining comparable levels of market volatility on both sides. Our dataset spans 2006 to 2023, covering significant economic events, including the 2008 financial crisis, the European debt crisis (2011–2012), the COVID-19 market shock (2020), and recent post-pandemic fluctuations. By using 2006 to 2015 for training, we capture a diverse range of volatility regimes, including the extreme spike of 2008 (VIX high: 59.89) and the gradual stabilization that followed (VIX average: 22.6). Meanwhile, the 2016 to 2023 validation period reflects more recent market conditions, incorporating both pre-pandemic stability (VIX in 2017: 11.04) and heightened uncertainty post 2018 (VIX high in 2020: 53.54).

Had we chosen a later date as a cutoff, the validation set would have been disproportionately volatile compared to the training data, making model evaluation less reliable. The end of 2015 cutoff balances these effects, ensuring that both training and validation periods include periods of high and low volatility, crucial for testing robustness in our models. Furthermore, our choice aligns with financial data best practices, where models trained on multiple economic cycles tend to generalize better. The volatility data is sourced from CBOE's historical VIX records, a widely used measure of market sentiment and uncertainty (CBOE VIX Index).

Descriptive statistics on the whole sample

Descriptive statistics are presented in Table 1.

Table 1. Descriptive statistics on the whole sample

Statistic	Lag1mcreal	G01dyadj	G02esg	G03nibadj	G04fcfyadj	G05rdsadj	G06invpegadj
Count	883172	883172	323744	222634	200673	604111	348593
Mean	7450514.49	0.01	38.57	0.00	0.01	0.09	0.07
Standard Dev.	39357040.33	0.02	19.23	0.02	0.08	0.26	0.06

Min	109.28	0.00	0.25	-0.20	-0.19	0.00	0.00
25%	176577.34	0.00	23.65	0.00	-0.01	0.00	0.03
50%	766415.45	0.00	34.69	0.00	0.01	0.00	0.05
75%	3355652.31	0.02	51.15	0.00	0.03	0.05	0.08
Max	3270662740.8	0.11	95.68	0.05	0.95	1.39	0.38
Missing Values	0%	0%	63%	75%	77%	32%	61%

G07epadj	G08sadadj	G09shoadj	G10shiadj	G11ret5adj	G12empadj	G13sueadj	G14erevadj
477188	338909	880908	255050	652817	712611	569037	575311
0.07	1.15	0.04	0.08	0.74	0.98	0.94	0.0
0.06	0.68	0.06	0.10	1.79	0.18	3.56	0.03
0.01	0.18	0.00	0.00	-0.98	0.40	-9.19	-0.19
0.03	0.88	0.00	0.02	-0.29	0.90	-0.70	0.00
0.05	1.02	0.02	0.04	0.33	0.97	0.71	0.00
0.08	1.20	0.05	0.10	1.14	1.04	2.43	0.00
0.35	4.81	0.53	0.65	13.49	2.12	10.81	0.22
46%	62%	1%	71%	26%	19%	36%	35%

Looking at the descriptive statistics table, the maximum number of entries for variables is 883,172. Many variables present a relatively high number of missing values, with the highest being 77% for Free Cash Flow Yield (g04fcfyadj), while variables such as the Market Cap and Dividend Yield present values for all fields. Seen below are the descriptions for the different variables.

- Group 1 (g01dyadj) focuses on Dividend Yield, calculated as the trailing 12-month dividend divided by the stock price at the end of the month.
- Group 2 (g02esg) uses the ESG (Environmental, Social, Governance) score, which is calculated annually and reflects a company's overall commitment to sustainable and ethical practices.
- Group 3 (g03nib) tracks Net Insider Buying, the percentage of shares purchased by company insiders minus those sold.
- Group 4 (g04_fcfyadj) looks at Free Cash Flow Yield, which is the free cash flow as a percentage of market capitalization.
- Group 5 (g05_rdsadj) examines the ratio of Research & Development expenses to sales. A higher ratio indicates that a company is investing more in innovation.
- Group 6 (g06_invpegadj) focuses on the Inverse PEG ratio, calculated as the inverse of the price-to-earnings ratio divided by (1 + future sales growth).
- Group 7 (g07epadj) considers Earnings Yield, which is the earnings per share divided by the stock price.
- Group 8 (g08_sadadj) compares the growth rates of sales and advertising expenses.
- Group 9 (g09shoadj) tracks Short Interest, or the percentage of shares sold short.
- Group 10 (g10shiadj) also measures Short Interest, similarly, reflecting investor pessimism.
- Group 11 (g11ret5adj) measures the trailing 5-year return, including delisting returns. Companies with higher 5-year returns tend to experience lower subsequent returns, based on the idea that past performance may not always predict future success.
- Group 12 (g12empadj) looks at Employee Turnover relative to Revenue Change, comparing the percentage change in the number of employees to the percentage change in revenue. While the relationship with returns isn't clear, high turnover relative to revenue changes may indicate operational inefficiencies.
- Group 13 (g13sueadj) focuses on Standardized Unanticipated Earnings (SUE), which reflects the difference between actual earnings and analysts' consensus forecasts.
- Group 14 (g14erevadj) tracks Analyst EPS Revisions, scaling changes in earnings forecasts by the stock price.

Descriptive statistics on your variable of interest

Our primary variable of interest is dividend yield, which reflects the proportion of a company's stock price returned to investors as dividends. We chose dividend yield because it is commonly used as a signal of a company's growth prospects, capital allocation strategy, and ability to generate shareholder value. Companies with high dividend yields may lack profitable reinvestment opportunities, whereas those with low dividend yields may have better opportunities to invest in projects that yield high returns on capital. By observing firms that reinvest earnings rather than pay high dividends, we are seeking to understand if investors may achieve superior long-term returns. This view is loosely aligned with Lintner (1956), who observed that firms maintain stable dividend policies and often retain earnings to fund growth, implying a trade-off with high payouts. Additionally, Fama and French (2001) suggest that high dividend yields are

characteristic of mature firms with fewer growth prospects, providing further support for exploring dividend yield as a predictor of stock returns in our model.

Descriptive statistics on dividend yield are presented in Table 2.

Table 2. Descriptive Statistics on Dividend Yield

Metric	DIVIDEND YIELD
Mean	0.013
Std Dev	0.022
Observations	883,172
Percentiles:	
Minimum	0.0
25%	0.0
50%	0.0
75%	0.019
Maximum	0.110

Table 2 presents the mean dividend yield is 1.3%, with a standard deviation of 2.2%. However, the distribution of dividend yields is heavily skewed to the left, as evidenced by the percentile values: the minimum is 0.0, the 25th and 50th percentiles are both 0.0, the 75th percentile is 1.9%, and the maximum is 11.0%. The median being 0.0 suggests that at least half of the firms in the sample do not pay dividends, highlighting a significant concentration of non-dividend-paying firms. This left-skewed distribution is further supported by the mean being greater than the median, a characteristic of left-skewness where a small number of firms with higher dividend yields pull the mean upward. The maximum dividend yield of 0.110 indicates the presence of some outliers with relatively high yields, though these are rare given the low values at the 75th percentile.

Results

Regression

Ordinary least squares regression

Looking at the results of the OLS in Table 3, we can observe a relatively low R^2 value of 0.003 which shows us that our chosen variable explains a low percentage of returns, although it represents a statistically significant relationship with a near 0 p-value. One reason for this result is that the stock market is largely irrational and the returns are uncertain and difficult to predict.

Table 3. In-sample OLS Regression Output

OLS Regression Results	
Dependent Variable	indadjret
R^2	0.003
Adjusted R^2	0.003
F-Statistic	86.02
Probability (F-statistic)	0.00

Examining the coefficients in Table 4 reveals that most variables exhibit small values, indicating a minimal influence on industry-adjusted returns. Among them, dividend yield stands out with the largest coefficient. For each 1% rise in dividend yield, returns increase, on average, by 0.0895 (or 8.95%). The intercept, at 0.0024, suggests that when all other variables are held constant, the expected industry-adjusted return is 0.24%.

Table 4. OLS In-Sample Coefficients

Feature	Coefficient	Feature	Coefficient	Feature	Coefficient
intercept	0.0024	fing05rdsadjmiss	-0.0046	fing10shiadjmiss	-0.0022
lag1cmreal	-1.17e-11	fing06_invpegadj	-0.0246	fing11retSadj	-0.0011
fing01dyadj	0.0895	fing06_invpegadjmiss	0.0056	fing11retSadjmiss	-0.0033
fing02tseq	-6.9e-05	fing07epadj	0.0725	fing12empadj	-0.0037
fing02esmigss	-0.0006	fing07epadjmiss	0.0013	fing12empadjmiss	-0.001
fing03nibadj	-0.0273	fing08sadadj	0.0004	fing13useadj	0.0023
fing03nibadjmiss	-0.0051	fing08sadadjmiss	0.0008	fing13useadjmiss	0.0008
fing04fcfyadj	0.0589	fing09shoadj	-0.0159	fing14revadj	-0.0109

fing04fcfyadjmiss	0.0033	fing09shoadjmiss	-0.004	fing14revadjmiss	-0.0049
fing05rdsadj	0.0046	fing10shiadj	0.0071		

Looking at the output of tests in Table 5, The Omnibus and Jarque-Bera tests both have p-values of 0.000, meaning the residuals deviate significantly from a normal distribution. The Durbin-Watson statistic is close to 2, meaning there is little to no autocorrelation in the residuals. However, the condition number (1.03e+16) is extremely high, pointing to severe multicollinearity, meaning that some independent variables are highly correlated.

Table 5. Tests Results

Omnibus	646116.66	Jarque-Bera (JB)	108381586.2
Prob (Omnibus)	0.00	Prob (JB)	0.00
Durbin-Watson	1.938	Condition Number	1.03E+16

The out-of-sample R^2 is 0.004, versus an in-sample R^2 of 0.003. Which suggests that the model is not overfitting the data and is able to predict slightly better out-of-sample.

To predict the industry adjusted return for January 2023 and find out which stocks are predicted to perform the best using this model, we fit the data and obtain the results presented in Table 6.

Table 6. Stock Predictions

CUSIP	Ticker	Company Name	Predicted Return
00206R102	T	A T & T INC	1.73%
550241103	LUMN	LUMEN TECHNOLOGIES INC	1.71%
92343V104	VZ	VERIZON COMMUNICATIONS INC	1.18%
38046C109	GOGO	GOGO INC	0.60%
872590104	TMUS	T MOBILE U S INC	0.40%

Looking at the top 5 companies, we see names such as Gogo Inc (0.6%), T-Mobile Inc (0.4%), AT&T (1.73%), Lumen Technologies (1.71%) and Verizon (1.18%). The average predicted returns for the top 5 companies in January 2023 is 0.011 (1.1%).

Penalized regression

LASSO regression is a linear regression technique that adds a penalty to the size of coefficients, effectively shrinking less important ones to zero to prevent overfitting and enhance model interpretability. We included LASSO in our analysis of industry-adjusted stock returns because it excels at feature selection, identifying the most relevant financial metrics among a large set of predictors. This makes it a valuable addition to our methodology, offering a simpler, more interpretable linear approach compared to complex models like neural networks, while providing insights into variable importance that complement the non-linear methods such as Decision Trees, Random Forest, and Gradient Boosting.

We used normalized variables for the right-side of our regression, to account for differences in scale. We also tested different learning rates and tested the performance of the penalized regression against that of the OLS. The following tables shows our results in terms of R^2 within the sample (rsq_lasso_train), outside of the sample and fine-tuned (rsq_lasso_valid) and finally the results of running a LASSO regression on the validation dataset (rsq_lasso_valid2). We also included the performance of the OLS for comparison.

When looking at the overall performance of the LASSO regression in Table 7, we can see that it performs worse than the OLS regression, even though there is high multicollinearity, and a high number of variables included in the regression. We can see that when we increase the penalty factor, the model's explanatory power decreases both in and out of the sample. Usually, we might be looking at worse in sample predictive power that is “traded” for a higher out of sample predictive power, however, that does not happen in our case. As such, LASSO seems to be performing worse on our dataset than the standard OLS (or LASSO with the learning rate set to 0).

Table 7. LASSO Results

Variable	OLS (penalty = 0)	Penalty = 0.0001	Penalty = 0.0003	Penalty =0.0006
rsq_lasso_train	0.00330	0.00308	0.00288	0.00254
rsq_lasso_valid	0.00353	0.00315	0.00275	0.00219
Rsq_lasso_valid2	0.00495	0.00462	0.00442	0.0039

Looking at the coefficients in Table 8, we can see that most of the variables have small coefficients, meaning their impact on industry adjusted returns are low. The variable with the most significant coefficient is Standardized Unanticipated Earnings (SUE). For each 1 percentage increase in SUE, returns increase, on average, by 0.008 (0.8%). The intercept has a value of 0.004, meaning that while all other variables remain constant, the expected industry adjusted return is 0.4%.

Table 8. LASSO In-Sample Coefficients

Feature	Coefficient	Feature	Coefficient	Feature	Coefficient
intercept	0.0040	zfing05rdsadjmiss	0.0034	zfing10shiadjmiss	-0.0002
lag1cmreal	-3.4e-11	zfing06_invpegadj	-0.0004	zfing11retSadj	-0.0016
zfing01dyadj	0.0018	zfing06_invpegadjmiss	0.0021	zfing11retSadjmiss	0.0024
zfing02tseq	0.0	zfing07epadj	0.0030	zfing12empadj	0.0011
zfing02esmigss	0.0005	zfing07epadjmiss	0.005	zfing12empadjmiss	0.0002
zfing03nibadj	0.0016	zfing08sadadj	0.0	zfing13useadj	0.0083
zfing03nibadjmiss	0.0037	zfing08sadadjmiss	0.0003	zfing13useadjmiss	0.0
zfing04fcfyadj	0.0132	zfing09shoadj	-0.0007	zfing14revadj	9.5e-05
zfing04fcfyadjmiss	0.0	zfing09shoadjmiss	0.0	zfing14revadjmiss	-0.0012
zfing05rdsadj	0.0010	zfing10shiadj	0.0		

To predict the industry adjusted return for January 2023, we fit the data and obtain the results presented in Table 9. The stocks are listed from highest predicted industry adjusted return to lowest. We see names such as AT&T (0.36%), Frontier Communications (1.18%), Verizon (1.36%), T-Mobile (0.96%) and Gogo Inc (0.28%). We observe that the stocks that ranked high using the OLS are also present here with small changes in their order. The average predicted returns for the top 5 companies in January 2023 is 0.0083 (0.8%), using a penalty factor of 0.0001.

Table 9. LASSO Stock Predictions

CUSIP	Ticker	Company Name	Predicted Return
00206R102	VZ	VERIZON COMMUNICATIONS INC	1.36%
550241103	FYBR	FRONTIER COMMUNICATIONS PRINT	1.18%
92343V104	T	T MOBILE U S INC	0.96%
872590104	TMUS	A T & T INC	0.36%
38046C109	GOGO	GOGO INC	0.28%

Decision trees

Vanilla decision tree

Decision Trees are a valuable tool in this analysis due to their ability to capture non-linear relationships and interactions between variables, which may not be adequately modeled by traditional linear approaches like OLS. This makes them well suited for predicting industry-adjusted stock returns, as complex patterns are likely to exist due to varying growth prospects and market dynamics across firms. Additionally, Decision Trees offer interpretability through a clear hierarchical structure, allowing us to identify key features influencing returns and setting the stage for more advanced ensemble methods like Random Forest and Gradient Boosting.

To find the optimal values for these hyperparameters, we used automated search via Grid Search in Python, which explores predefined combinations of parameters and selects the one that maximizes R^2 on the validation set.

We used the following hyperparameters:

- **max_depth:** This parameter limits the number of levels in the Decision Tree, controlling its complexity to prevent overfitting while capturing relevant patterns. We tested values of 5,7,9,11,13.
- **min_weight_fraction_leaf:** This parameter sets the minimum fraction of the total sample weight required at a leaf node, ensuring leaves have sufficient samples to prevent overfitting. We tested values of 0.05, 0.10, 0.15, 0.20, and 0.25.

After testing 25 combinations of hyperparameters, we found the optimal combination of hyperparameters leading to a 0.00349 R^2 in the training set, and an 0.00233 R^2 in the validation set, as can be seen in Table 10. After running the model, a maximum depth of 7 and a minimum weight per leaf of 0.05

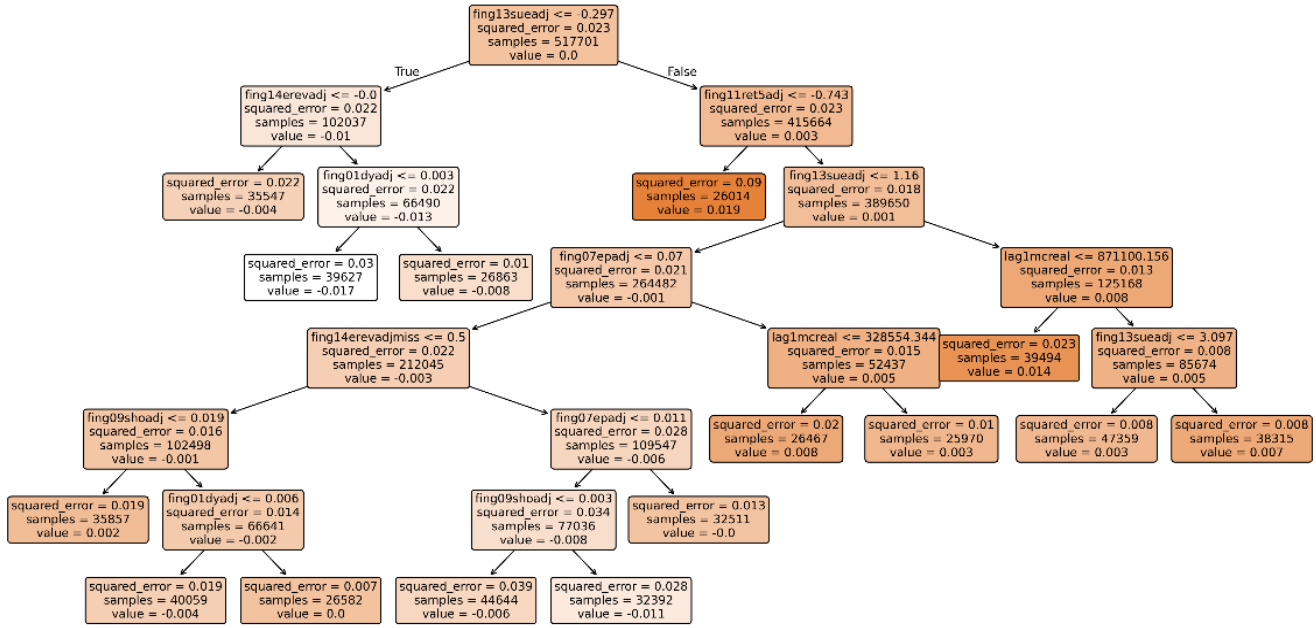
produced the highest out of sample R^2. The top 5 combinations of results are ranked from highest to lowest R^2 in the validation set. The top five combinations of hyperparameters are presented in Table 10.

Table 10. Vanilla Decision Tree: Optimal Hyperparameters and R^2

Max Depth	Min Weight Fraction Leaf	Training R^2	Validation R^2
7	0.05	0.0035	0.0023
9	0.05	0.0035	0.0023
11	0.05	0.0035	0.0023
5	0.05	0.0033	0.0023
5	0.10	0.0026	0.0021

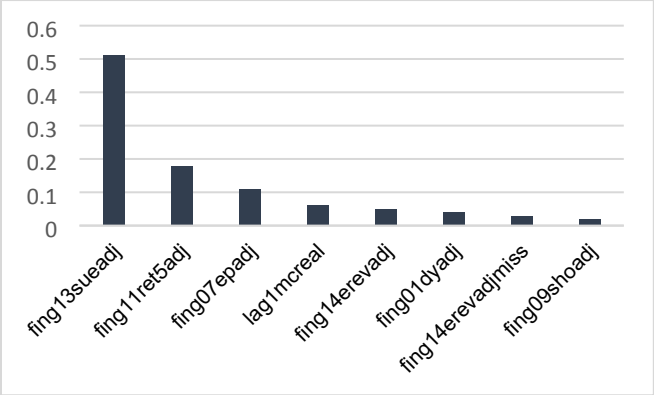
The Decision Tree (**Error! Reference source not found.**) identifies fing13sueadj (standardized unexpected earnings) as the most important feature, with the root split at ≤ -0.297, separating firms with significant negative earnings surprises. Other key splits include fing14erevadj (earnings revisions) at ≤ 0.0, fing11ret5adj (5-year return) at ≤ 0.743, and lag1mcreal (lagged market capitalization) at ≤ 871100.156, with additional splits on fing13sueadj (e.g., ≤ 1.116) and fing07epadj (earnings-to-price ratio) at ≤ 0.021. Notably, the lag1mcreal split predicts higher returns for larger firms (> 871100.156) at 0.014 compared to 0.005 for smaller firms (≤ 871100.156), which contrasts with our broader analysis where smaller firms showed higher industry-adjusted returns, suggesting the tree may be capturing a specific pattern in this subset of the data.

Figure 1. Decision Tree Output



As shown in Figure 2, Fing13sueadj (standardized unanticipated earnings) has the highest feature importance value at 51% which is as expected as it was the first node to split. Fing11ret5adj (trailing 5-year return) followed by fing11ret5adj at 20% and fing07epadj variable (earnings per share/price) at 11%.

Figure 2. Decision Tree Importance Values



After testing the Decision Tree model, we applied it to the data for January 2023 to predict the stock returns in the Communication Services sector. The detailed results are shown in Table 11. The best performing company is TMUS, with a predicted return of 0.73%. The following two predicted return values are identical (0.33% for T and GOGO), as well as the last two (0.30% for FYBR and VZ). This result is a result of the decision tree model structure, which identified only eight relevant variables to explain the returns. As a result, if these companies fit within the constraints of these variables, the model assigns the same predicted return to each of them.

Table 11. Decision Tree Predictions

CUSIP	Ticker	Company Name	Predicted Return
872590104	TMUS	T MOBILE U S INC	0.73%
00206R102	T	A T & T INC	0.33%
38046C109	GOGO	GOGO INC	0.33%
35909D109	FYBR	FRONTIER COMMUNICATIONS PRINT	0.30%
92343V104	VZ	VERIZON COMMUNICATIONS INC	0.30%

Random forest

The Random Forest analysis addresses the limitations of single decision trees by using randomness and diversity across multiple trees to limit the impact of overfitting. This method improves predictive accuracy and generalization, making it particularly effective for complex datasets where non-linear relationships are prevalent. In predicting industry-adjusted stock returns, Random Forest' ability to capture intricate, context-dependent patterns without requiring assumptions about the data's underlying structure makes it an important addition to our methodology, complementing simpler models like OLS and Decision Trees.

To ensure the Random Forest model performed optimally, we tuned its hyperparameters in the validation set. These are settings that govern the structure and behavior of the individual decision trees within the forest. The key hyperparameters we focused on included:

- **max_depth:** This parameter restricts the maximum depth of each tree, preventing excessive growth that could lead to overfitting. We tested values of 3, 5, and 7 to balance model complexity and generalization.
- **min_weight_fraction_leaf:** This specifies the minimum weighted fraction of the total sum of weights (of the input samples) required to be at a leaf node. We evaluated values of 0.05, 0.10, and 0.15 to control the robustness of the leaves and reduce sensitivity to small sample variations.
- **n_estimators:** This determines the number of trees in the forest. We tested 3, 5, and 10 trees to assess the trade-off between computational efficiency and the reduction in variance achieved with a larger ensemble.
- **max_features:** This sets the maximum number of features considered for splitting at each tree. We explored values of 4, 6, and 8 to introduce randomness and ensure diverse trees within the forest.
- **max_samples:** This controls the fraction of the original dataset sampled (with replacement) for each tree. We tested values of 0.6, 0.8, and 1.0 to evaluate the impact of bootstrap sample size on model performance.

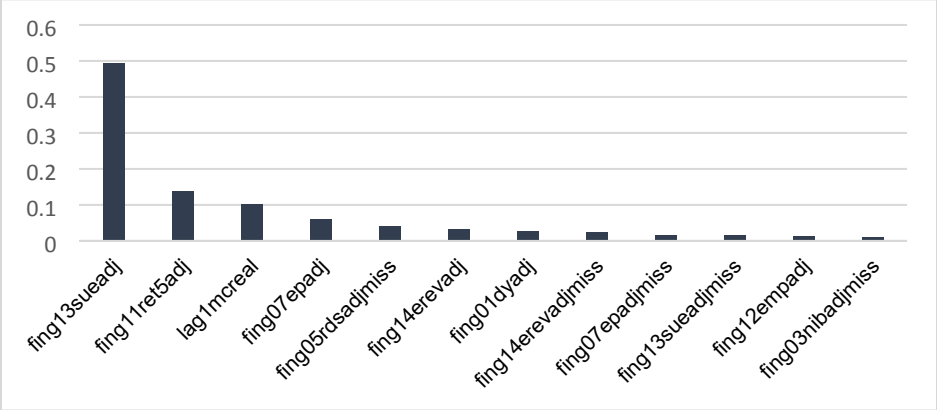
To determine the optimal combination of these hyperparameters, we employed a Grid Search approach in Python. This automated technique evaluated all predefined combinations of parameter values and selected the configuration that maximized the out-of-sample R^2 on a validation set. After testing 243 combinations of hyperparameters, we found the optimal result to have a combination of 7 on the max depth, 0.05 min_weight_fraction_leaf, 5 n_estimators, 6 max_features, and 0.8 max_samples to be the best performing hyperparameters leading to a 0.00356 R^2 in the training set (r2_train), and an 0.00296 R^2 in the validation set (r2_valid). In Table 12, the top five combinations of results are ranked from highest to lowest R^2 in the validation set.

Table 12. Hyperparameter Optimization Results

max_depth	min_weight_fraction_leaf	n_estimators	max_features	max_samples	r2_train	r2_valid
7	0.05	5	6	0.8	0.0036	0.0030
7	0.05	10	8	0.6	0.0039	0.0028
7	0.05	3	6	0.8	0.0034	0.0028
7	0.05	5	8	0.8	0.0038	0.0028
7	0.05	10	8	1	0.0039	0.0028

Feature importance scores, derived from the final trained model, measure the relative contribution of each input variable to reducing mean squared error across the forest. These scores, normalized to sum to 1, provide insight into which features most influenced the model's predictions during the validation period. Figure 3 presents the feature importance scores for the features used in the model that had greater than 1% importance. The variable with the highest importance factor by a large margin was fin13sueadj (standardized unexpected earnings), with 49.4% importance. This was followed by fing11ret5adj (5-year historical return) with a 13.8% importance followed later by lag1mcreal (market capitalization) at 6.1% and fin07epadj (earnings-to-price ratio) at 4.0%.

Figure 3. Random Forest Importance Values



In Table 13, the stocks with the highest predicted industry adjusted returns for the test set in January are ranked from best at the top of the table, to worst at the bottom. VZ was predicted to perform the best with a predicted industry adjusted return of 0.75% for the month, followed by FYBR at 0.62% and TMUS at 0.62%.

Table 13. Random Forest Predictions

CUSIP	Ticker	Company Name	Predicted Return
92343V104	VZ	VERIZON COMMUNICATIONS INC	0.75%
35909D109	FYBR	FRONTIER COMMUNICATIONS PRINT	0.62%
872590104	TMUS	T MOBILE U S INC	0.62%
00206R102	T	A T & T INC	0.57%
38046C109	GOGO	GOGO INC	0.45%

Gradient boosting

Gradient Boosting builds decision trees sequentially, with each tree designed to correct the errors of its predecessors by minimizing the mean squared error through gradient descent optimization. Unlike Random Forest, which primarily reduce variance through averaging, Gradient Boosting also considers bias, enabling it to uncover intricate, non-linear patterns and higher-order feature interactions within the data. This adaptability makes it well-suited for predicting industry-adjusted stock returns, where

relationships between financial indicators and performance are rarely linear and often shift across diverse market conditions, such as the 2008 financial crisis, the 2020 COVID-19 downturn, and subsequent recovery phases. By iteratively refining predictions to focus on challenging cases, Gradient Boosting excels at modeling these dynamic complexities without imposing assumptions about the data's structure.

To ensure the Gradient Boosting model performed optimally, we tuned its hyperparameters using the validation set (January 2016 to December 2022). These hyperparameters govern the structure, learning rate, and behavior of the model, balancing its complexity and generalization ability. The key hyperparameters we focused on included the following:

- **max_depth:** This parameter restricts the maximum depth of each tree, controlling model complexity to prevent overfitting. We tested values of 3, and 7 to balance the trade-off between capturing patterns and maintaining generalization.
- **min_child_weight:** This specifies the minimum weighted fraction of the total sum of weights (of the input samples) required to be at a leaf node. We evaluated values of 0.05, and 0.15 to ensure robustness at the leaf level and reduce sensitivity to small sample variations.
- **n_estimators:** This determines the number of trees (boosting stages) in the model. We tested 50, and 150 trees to assess the trade-off between predictive accuracy and computational efficiency, as more trees can improve performance but increase training time.
- **max_features:** This sets the maximum number of features considered for splitting at each tree. We explored values of 4, and 8 to introduce randomness and ensure diverse trees, similar to Random Forest, while allowing the model to focus on the most relevant features.
- **subsample:** This controls the fraction of the training data sampled (without replacement) for each tree. We tested values of 0.6, and 1.0 to evaluate the impact of stochastic Gradient Boosting, which can further reduce overfitting by introducing randomness in the training process.
- **learning_rate:** This scales the contribution of each tree to the final prediction. We tested values of 0.01, and 0.3 to balance the trade-off between learning speed and model stability, as a smaller learning rate requires more trees but can lead to better generalization.

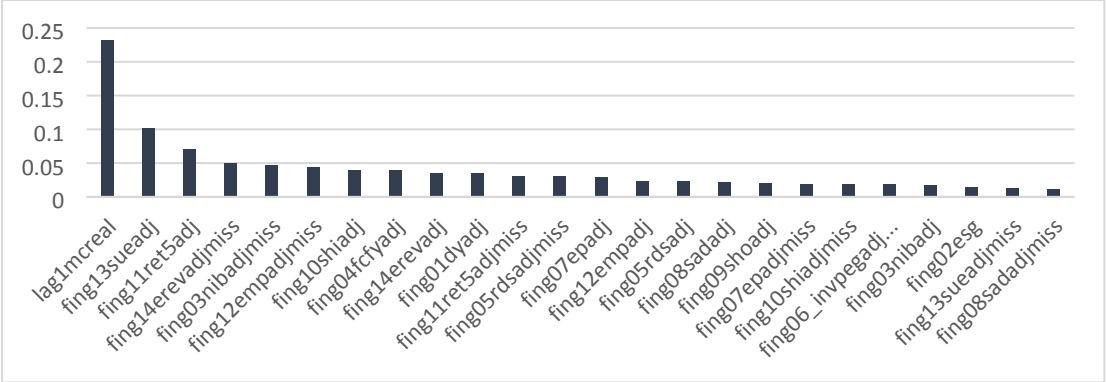
To determine the optimal combination of these hyperparameters, we employed a Grid Search approach in Python. This automated technique systematically evaluated all 64 combinations of parameter values and selected the configuration that maximized the out-of-sample R^2 on the validation set. Displayed in Table 14 are the top 5 best performing set of parameters ranked from best at the top to worst at the bottom. The best performing set of parameters had a max_depth of 3, min_child_weight of 0.05, max_features of 4, subsample value of 6, and the learning rate was 0.3 resulting in a R^2 value of 0.004375 in the validation set.

Table 14. Hyperparameter Optimization Results

max_depth	min_child_weight	n_estimators	max_features	subsample	learning_rate	r2_train	r2_valid
3	0.05	50	4	0.6	0.3	0.00988	0.00438
3	0.15	50	4	0.6	0.3	0.00988	0.00438
3	0.05	50	4	1	0.3	0.01014	0.00426
3	0.15	50	4	1	0.3	0.01014	0.00426
3	0.05	50	8	0.6	0.3	0.01296	0.00384

In contrast to the previous models, the GBM model produced a more balanced distribution of feature importances, with predictive power drawn from a wider variety of input variables. As shown in Figure 4, the most influential feature was lag1mcreal, accounting for 23.1% of the total importance. This was followed by fin13sueadj (standardized unexpected earnings) at 10.2%, fing11ret5adj (5-year historical return) at 7.0%, and fin14erevadjmiss (missing indicator for earnings revisions) at 5.0%. While these features stood out, a larger number of variables contributed non-trivially to the predictions compared to the more concentrated importance seen in the random forest and decision tree models. This broader spread suggests that the GBM model leveraged a more diverse set of signals to optimize predictive performance, potentially reflecting the impact of regularization techniques, shallow tree depth, and a richer feature space that prevented the model from over-relying on just a few dominant predictors.

Figure 4. Gradient Boosting Importance Values



As shown in Table 15, the stocks with the highest predicted industry adjusted returns from the test set in January 2023 are ranked from best to worst with Lumen, Verizon, and AT&T, TMUS, and GOGO having the highest expected industry adjusted returns with 3.3%, 1.2%, 1.0%, 0.5%, and 0.3% expected returns, respectively.

Table 15. Gradient Boosting Predictions

CUSIP	Ticker	Company Name	Predicted Return
550241103	LUMN	LUMEN TECHOLOGIES INC	3.25%
92343V104	VZ	VERIZON COMMUNICATIONS INC	1.24%
00206R102	T	A T & T INC	1.0%
872590104	TMUS	T MOBILE U S INC	0.46%
38046C109	GOGO	GOGO INC	0.33%

To evaluate the predictive potential of our models, we compared the performance of a single Decision Tree, Random Forest, and Gradient Boosting on the validation set, as shown in Table 16. The Decision Tree yielded a validation R^2 of 0.0023, while the Random Forest improved upon this with an R^2 of 0.003, and Gradient Boosting outperformed both with an R^2 of 0.0044. The Decision Tree's lower R^2 suggests it struggled to explain return variability compared to a simple mean-based prediction, whereas the positive R^2 values for Random Forest and Gradient Boosting indicate modest but meaningful improvements. Random Forest's ensemble approach mitigates overfitting by averaging predictions across multiple trees, enhancing its predictive power over a single Decision Tree. Gradient Boosting, however, further improves accuracy by sequentially optimizing weak learners to correct prior errors, capturing a greater portion of return variability through its ability to handle complex, non-linear relationships. This makes both ensemble methods, Random Forest and Gradient Boosting, more suitable than a standalone Decision Tree for predicting industry-adjusted stock returns, with Gradient Boosting showing the strongest performance due to its iterative refinement.

Table 16. Decision Tree Comparison

R^2	Training	Validation
Gradient Boosting	0.0099	0.0044
Random Forest	0.0036	0.0030
Decision Tree	0.0035	0.0023

Neural networks

Neural networks complement our analysis because of how they excel at modeling highly complex and non-linear relationships through their layered structure, which can uncover patterns that simpler models might miss. For predicting industry-adjusted stock returns, neural networks are particularly valuable due to their ability to learn intricate dependencies among financial metrics, such as dividend yield and earnings surprises, by adjusting weights during training to optimize predictions. Their adaptability to large, high-dimensional datasets makes them a powerful tool in our methodology, offering a complementary perspective to tree-based models like Decision Trees, Random Forest, and Gradient Boosting, and providing insights into the potential of deep learning for financial forecasting.

For our neural network architecture, we used a two-layer black box with 64 neurons per layer and batch normalization. Our chosen batch size is 32. We experimented with other architectures but ultimately found similar results.

To set the number of epochs, we used early stopping to monitor the validation loss (val_loss) with a "patience" of 10 epochs. If after 10 epochs the validation loss does not decrease by more than 0.00001, the model stops training and reverts to its best weights. For the learning rate, we started with 0.001 and decreased it by a factor of 0.1 each time the validation loss did not change for more than 5 epochs. The minimum learning rate we set is 0.0000001.

As the model "chooses" its best learning rate and maximum number of epochs, we set the epochs to 100 to allow the model to keep running until it finds its best combination of hyperparameters. From our observations, the model would train on average for 24-30 epochs until it found the right combination. This generated an in-sample R^2 of 0.00336, an out-of-sample R^2 of 0.00304 and an R^2 on the whole dataset of 0.00404.

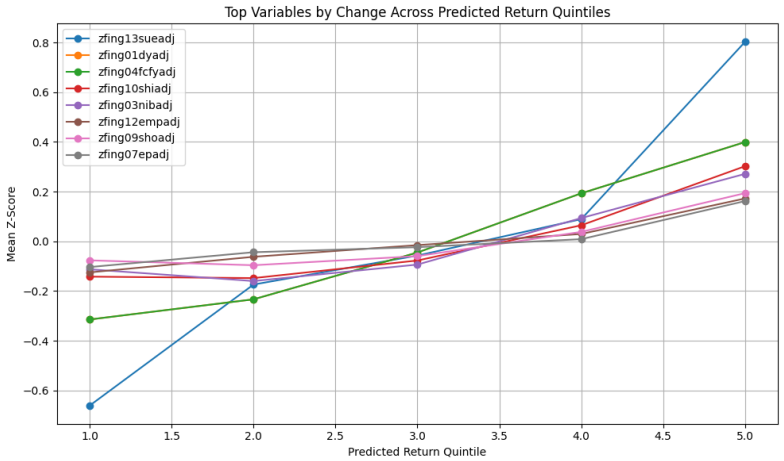
To better understand which features the neural network relied on most heavily, we analyzed the change in mean z-score values of each input variable across predicted return quintiles. This approach reveals how sensitive the model's predictions are to specific financial signals.

Looking at the most significant changes in Table 17 below, zfing13sueadj, which captures standardized unexpected earnings (SUE), showed the most significant increase from the lowest to the highest quintile, rising by 1.46 standard deviations. This dramatic shift indicates that the model strongly favors companies that have recently delivered positive earnings surprises. Additionally, dividend yield (zfing01dyadj) and free cash flow yield (zfing04fcfyadj) also ranked highly in importance, each increasing by 0.71 standard deviations. These findings suggest the model rewards firms with strong shareholder distributions and solid cash generation capabilities. Other meaningful contributors include short interest (zfing10shiadj), which increased by 0.44 standard deviations, and net income before extraordinary items (zfing03nibadj), which rose by 0.38. We also graphed these changes in Figure 5 to show the evolution of these variables across quintiles.

Table 17. Significant Variable Changes Across Quintiles

Variable	Description	Change (Q5 - Q1)
zfing13sueadj	Standardized Unexpected Earnings (SUE)	+1.46
zfing01dyadj	Dividend Yield	+0.71
zfing04fcfyadj	Free Cash Flow Yield	+0.71
zfing10shiadj	Short Interest	+0.44
zfing03nibadj	Net Income Before Extraordinary Items	+0.38

Figure 5. Significant Variable Changes Graph



We then computed the best companies in terms of predicted returns, which we can observe in Table 18.

Table 18. Neural Networks Results

CUSIP	Ticker	Company Name	Predicted Return
550241103	LUMN	LUMEN TECHNOLOGIES INC	1.61%
00206R102	T	A T & T INC	1.48%
92343V104	VZ	VERIZON COMMUNICATIONS INC	1.02%
38046C109	GOGO	GOGO INC	0.55%
872590104	TMUS	T MOBILE U S INC	0.42%

The top 5 companies are, in order of best predicted performance to worst:

- Lumen Technologies Inc. (1.61%)
- AT&T Inc. (1.48%)
- Verizon (1.02%)
- Gogo Inc. (0.55%)
- T-Mobile Inc. (0.42%)

On average, these companies are predicted to have a 0.01 (1%) increase in returns compared to the industry average.

Reconciliation of results based upon different estimation techniques

Table 19. Stock rankings under different estimation techniques

	Regression		Decision trees			
Rank	OLS	LASSO	Vanilla	Random Forest	Gradient Boosting	Neural networks
Panel A: Stock						
1	T	VZ	TMUS	VZ	LUMN	LUMN
2	LUMN	FYBR	T	FYBR	VZ	T
3	VZ	T	GOGO	TMUS	T	VZ
4	GOGO	TMUS	FYBR	T	TMUS	GOGO
5	TMUS	GOGO	VZ	GOGO	GOGO	TMUS
Panel B: Industry-adjusted predicted return (%)						
1	1.73%	1.36%	0.73%	0.75%	3.25%	1.61%
2	1.71%	1.18%	0.33%	0.62%	1.24%	1.48%
3	1.18%	0.96%	0.33%	0.62%	1.00%	1.02%
4	0.60%	0.36%	0.30%	0.57%	0.46%	0.55%
5	0.40%	0.28%	0.30%	0.45%	0.33%	0.42%

To determine the relevant techniques for our final stock recommendation we will be looking at how the models perform in and out of sample. Looking at the regression techniques, OLS will serve as our base linear model, showing an out-of-sample R^2 of 0.00353, while LASSO performed comparably worse with and R^2 value of 0.00315. The vanilla decision tree performed the worst of the three models, with an out-of-sample R^2 of 0.00233. Random Forest and Gradient Boosting showed promising results with an out-of-sample R^2 of 0.00296 and 0.00440 respectively. Finally, Neural Networks also showed good performance in and out of sample, with an out-of-sample R^2 of 0.00304.

In analyzing our results, we categorized stocks into preferred and non-preferred groups, highlighting them in green and gray, respectively, based on their predicted returns across models. We discounted the Decision Tree outputs due to their tendency to overfit and the fact that they had the lowest R-squared value (0.0023) among all models, which led us to mark T-Mobile (TMUS) as gray despite its high predicted returns in that model. Verizon (VZ) and AT&T (T), both large conglomerates with high dividend yields and stable free cash flow yields, exhibited consistently high predicted returns across models, earning them a green highlight. Lumen (LUMN), a smaller company, also performed strongly, particularly in Gradient Boosting, due to its low market cap and high dividend yield. Dividend yield and earnings yields were metrics that aligned with Verizon and AT&T and proved significant across multiple models, reinforcing our confidence in our predictions. Conversely, we marked Gogo (GOGO) as gray due to its

consistently lower rankings, and Frontier Communications (FYBR) was also grayed out, as it only appeared in the top five for predicted returns in three out of six models.

Conclusion

Considering model performance, predicted return magnitude, and cross-model consensus, we propose five final stock recommendations: AT&T, Verizon, Lumen Technologies, Gogo, and T-Mobile. These stocks were consistently among the top-ranked across models and presented favorable predicted returns. Finally, our ranking, from best performing to worst, is the following: Verizon (1), Lumen Technologies (2), AT&T (3), T-Mobile (4), GOGO (5).

Looking at which features were of most importance in our models, OLS emphasizes earnings strength through dividend yield (g01dyadj), earnings yield (g07epadj), and free cash flow yield (g04fcfyadj). The Random Forest model, along with Gradient Boosting, focuses on capturing unexpected performance and stability via earnings surprises (g13sueadj), five-year returns (g11ret5adj), and market capitalization (lag1mcreal), meanwhile, the Neural Networks model integrates the focus on earnings surprises (g13sueadj) with dividend yield (g01dyadj) and free cash flow yield (g04fcfyadj) to balance growth potential with steady cash generation.

We chose Verizon as our top spot and final recommendation to purchase as it ranks consistently high across all our models and because it is an established company with a strong financial record. The data indicates that Verizon consistently delivers a steady dividend yield (which ranks as an important feature in both OLS and Neural Network models) with a value of 6.5% and which positions it as an attractive option for income-focused investors. Favorable metrics like an earnings yield of 4% (which ranks highly in OLS and Neural Networks) and a SUE of 2.02 (which is important for Random Forest, Gradient Boosting and Neural Networks), Verizon's overall financial health and consistent performance have led to its top ranking in both traditional and machine learning models.

AT&T (T), Lumen Technologies (LUMN), T-Mobile (TMUS), and Gogo (GOGO) each stood out for slightly different reasons, but all performed well according to the key drivers identified by our models. AT&T demonstrated a notably strong dividend yield (0.11) and healthy free cash flow yield (0.05), both of which boosted its rankings in the OLS and Neural Network frameworks. Lumen's low market capitalization (5 million $), high dividend yield (0.11) and respectable values in other fields, helped it score well across models. T-Mobile's combination of SUE (4.5) and 5-year trailing return (1.21) made it attractive to the decision tree-based models, which reward consistent returns and ongoing improvements. Lastly, Gogo's 5-year trailing return (0.3) and moderate SUE value (0.08) made it a good, although lower ranked choice on the decision tree-based models.

References

- Lintner, J. (1956). Distribution of Incomes of Corporations Among Dividends, Retained Earnings, and Taxes. American Economic Review, 46(2).
- Fama, E. F., & French, K. R. (2001). "Disappearing Dividends: Changing Firm Characteristics or Lower Propensity to Pay?" Journal of Financial Economics, 60(1), 3–43.
- CBOE VIX Index: https://finance.yahoo.com/quote/%5EVIX/
- GitHub Code Repository: https://github.com/Boby0230/ML_In_Investing