Stock Recommendation System using Machine Learning Approache

Ankit Kumar[1], Nilesh[2]

[1]M.Tech Student, Department of Computer Science Engineering, RAMA University, Kanpur.
[2]Assistant Professor, Department of Computer Science Engineering, RAMA University, Kanpur.

Abstract- Stock Recommendation is obligatory to speculation enterprises and speculators. Notwithstanding, no single stock decision procedure can persistently win though examiners probably won't have sufficient opportunity to determine all S&P 400 stocks (the standard and Poor's 400), during this paper, it's anticipated an affordable subject that implies stocks from S&P four hundred abuse AI. To start with, we will in general pick model image with reasonable explicative power. Also, it's taken 5 off occasions utilized AI ways, edge relapse, stepwise relapse, just as factual relapse, arbitrary woodland and summed up helped relapse, to demonstrate stock markers and quarterly log-return in a moving window. Thirdly, it's picked the model with the base Mean sq. Mistake in each add up to rank stocks. At long last, we will in general check the picked stocks by direct portfolio distribution ways like similarly weighted, mean-uniqueness, and least dissimilarity. Our trial decision show that the anticipated subject beats the long-just methodology on the S&P 400 file as far as Sharpe quantitative connection and aggregate returns.

Keywords— Stock Recommendation System. Model Selection. Machine Learning.

1. INTRODUCTION

Earning reports play a key role available recommendation. Analysts use company earning reports to try to stock purchase and sell recommendation. Future earnings estimates square measure vital factors to worth a firm. Earnings forecasts square measure supported analysts' estimation of company growth and gain. To predict earnings, most analysts build monetary models that estimate prospective revenues and prices. However, it may be terribly tough for analysts to accurately estimate earnings several researchers try to make a sturdy model to predict earnings. for instance, earnings generated by the cross-sectional model square measure thought-about superior to analyst forecasts to estimate the tacit prices of capital (ICC), that play a key role in firm valuation [1]. Moreover, regression-based models [2] is wont to predict scaled and un-scaled profits [3] various recent papers think about using deep learning algorithms to model securities market knowledge [4]. Deep neural networks models can also be trained to predict future fundamentals like book to-market quantitative relation, and as a result, investors will use foreseen fundamentals to rank current stocks [5]. There square measure 2 ancient approaches the primary approach is choosing stocks supported a gift criteria like worth to earnings (P/E) quantitative relation [6]. Stocks square measure hierarchal by P/E ratios using historical knowledge. Then, a portfolio can contain stocks with lowest P/E ratios. This approach is dissatisfactory in sensible things since the choice with P/E ratio solely is unstable (e.g., choosing prime amount stocks might end in a less prophetic power.) The second approach put together uses many criterion to rank stocks, like P/E ratio, worth to sales (P/S) quantitative relation, price/earnings to growth (PEG) quantitative relation, etc. However, this approach doesn't take the correlations among totally different predictor factors into thought. Consequently, weights of those factors square measure assigned comparatively subjective, that will increase the danger worth finance has been wide used these days by investors and portfolio managers. Graham 1st comes up with the thought of Associate in nursing intrinsic worth for a stock that's freelance of the market [7]. He emphasizes the importance of Associate in nursing intrinsic worth that's mirrored by a corporation market size, assets level, dividends, monetary strength, earnings stability, earnings growth that specialize in this worth, he believes would stop Associate in Nursing capitalist from misjudgment and interpretation throughout a bull or securities industry. within the long haul, we have a tendency to expect stock costs ought to eventually be regression towards the company's intrinsic worth several elementary monetary ratios like P/E ratio, earnings per share (EPS), come on equity (ROE), gross margin a fast quantitative relation indicate overall gain, stability, operational potency, capital structure, ability of generating future money flows and alternative valuable data of the corresponding corporations. Thus, these monetary ratios may replicate a corporation intrinsic worth and will have prophetic power on the long run performance [8] to boot, monetary ratios offer standardization in order that all corporations would have constant scale of information and therefore, those with massive capital can have equal influence, during this paper, we have a tendency to propose a completely unique theme that predict stock's future worth come supported earnings factors by machine learning so that all companies would have the same scale of data and thus, those with large capital will have equal influence. In this paper, we propose a novel scheme that predict stock's future price return based on earnings factors by machine learning. we have a tendency to use 5 machine learning algorithms (linear regression, random forest, ridge, stepwise regression, and generalized boosting regression) to assign weights to every issue dynamically, and choose prime 2 hundredth stocks every quarter supported the ranking of foreseen returns generated by the simplest acting formula over the past coaching periods before every re-balancing day on a rolling basis. The 5 models even have high instructive power. By victimization he lowest MSE to decide on the simplest model, we offer responsibility to business call, therefore increase

the safety to monetary investments. When we have a tendency to choose stocks for our portfolio on eac re-balancing day, we have a tendency to check quality allocation methodologies: mean-variance, min-variance, and equally-weighted allocation on selected stocks victimization in sample knowledge (1990-2007). Risk management is concerned by the tactic of most Sharpe quantitative relation employed in portfolio allocation methodologies. We have a tendency to square measure getting to balance the expected come and therefore the variance of the portfolio to realize the simplest risk to reward. Finally, we have a tendency to compare the P&L of our strategy with S&P five hundred index one, all 3 portfolio allocation strategies surpass the market, summarize the competence of our strategy. This paper take as follows. Section II describes the rolling window, mercantilism time, the data, and additionally presents the methodology and implementation of our theme. Section III contains the portfolio allocation strategies, risk management and dealing value. Section IV presents the performance and Section V concludes the paper.

2. PRPPOSED STOCK RECOMMENDATION SCHEME

A. Rolling Window Based Data Separation

Rolling windows is utilized to divide knowledge for multiple functions (i.e., coaching and testing). Rolling windows for coaching ranges from 16-quarter (4-year) to a most of 40-quarter (10-year). This coaching rolling window is followed by a annual window for take a look a thing and that we trade in line with the test results. The training-testing-trading cycle of our strategy is summarized. we tend to conjointly extend the trade date by 2 months lag on the far side the quality quarter finish date just in case some firms have a non-standard quarter finish date, e.g. Apple free its profit-and-loss statement on 2010/07/20 for the second quarter of year 2010. Therefore for the quarter between 04/01 and 06/30, our trade date is adjusted to 09/01 (sum methodology for different 3 quarters).

B. Data Pre-processing

The data for this project is especially taken from computation info accessed through Wharton analysis knowledge Services (WRDS) [9]. The dataset used here consists of the info over the amount of twenty seven years (from 06/01/1990 to 06/01/2017). We have a tendency to use all historical &P five hundred element stocks (about 1142 stocks) because the S&P five hundred pool square measure updated quarterly. The adjusted shut value goes on a day to day (trading days) and generates half-dozen,438,964 observations the elemental knowledge goes on a quarterly basis and generates ninety one,216 observations, additionally, we have a tendency to delete out line records that indicate a unleash date (rdq) once the trade date, that embody regarding zero.84% of the dataset. We have a tendency to assure that on our trade date, ninety nine of the businesses have their earnings reports able to be used. so as to preserve Associate in Nursing out-of-sample amount sufficiently long for back-testing the connection, the dataset has been divided into 3 periods in Fig. 2. to create the dataset for coaching, we have a tendency to choose high twenty most well liked money ratios in table I [2] and calculated the quality Poor's five hundred is Associate in Nursing yankee exchange index supported the market capitalizations of five hundred massive corporations having common shares listed on the N. Y. Stock Exchange or NASDAQ.

TABLE I

20 FINANCIAL INDICATORS

Revenue Growth	Price to cash flow ratio
Earnings per share (EPS)	Cash ratio
Return on asset (ROA)	Enterprise multiple
Return on equity (ROE)	Enterprise value/cash flow from operations
Price to earnings (P/E) ratio	Enterprise value/cash flow from operations
Price to sales (P/S) ratio	Working capital ratio
Net profit margin	Debt to equity ratio
Gross profit margin	Quick ratio
Operating margin	Days sales of inventory
Price to book (P/B) ratio	Days payable of outstanding

These factors from the fundamental raw data from the WRDS. Also, in order to build a sector-neutral portfolio, we split the dataset by the Global Industry Classification Standard (GICS) sectors. We handle missing data separately by sector: if one factor has more than 5% missing data, we delete this factor; if a certain stock generates the most missing data, we delete this stock. In this way, we've removed 46 stocks and the overall missing data is reduced to less than 7% of each sector. Finally, we delete this 7% missing data.

C. Methodology

Our goal is to predict S&P 500 forward quarter log-return rqtr T +f given predictors XT constructed from historical data of the twenty financial factors over a particular quarter T and S&P 500 horizon f. At a given time T of the financial horizon, the 1-quarter forward log-returns of a certain stock price S are defined as: rqtr T +f,i = ln(ST +f,i/ST,i), i = 1, ..., nT , (1) where nT is the companies whose stock price and earnings factors are available at time T. A general estimator is the ordinary least square: r qtr T +f,i = β0 + p j =1 βjXT,i,j + , j = 1, ..., 20, (2) where j is the number of the twenty financial ratios, p is the total factors we used in the model, β0 is the intercept of the model, Xj corresponds to the jth predictor variable of the model, βj is the coefficients of the predictor variable and is the random error with expectation 0 and variance σ2. Moreover, regularized linear OLS estimators have a higher accuracy in many aspects [10]. We need to use multiple regression estimators to increase accuracy. [3] has summarized the prediction rule and estimator selection rule for using multiple estimators: rqtr T +f,i|XT,i,j, θ, i = 1, ...nT , j = 1, ..., 20, (3) rqtrt+f,i = gθ(Xt,i,j) + , t = T, ..., T − h, i = 1, ..., nt, (4) where h is the historical estimation period, gθ(Xt,i,j) is used to estimate θ through historical regressions. It is noticeable to point out that (2) is the basic estimator of (4). We pick five models for gθ: linear regression, forward and backward stepwise regression under Akaike information criterion (AIC), regularized linear OLS estimator ridge regression, tree based nonlinear model random forest and generalized boosted regression model (GBM) using gaussian distribution which implements A da Boost algorithm and Friedman's gradient boosting machine. All of the algorithms are facilitated by standard R packages [11]. For linear regression and stepwise regression we use lm and step and for ridge we use glmnet and MASS [12], [13]. For random forest we use random Forest [14]. For gbm we use gbm [15]. The reason of using these five models is that we need feature selection methods to remove undesirable features, thus reducing the over fitting issues, improving model accuracy and expediting the training procedure. We also have a white-box model that we can observe every single factor with its coefficients in our model. Mean Squared Error (MSE) [3] is used as the metric for our evaluation.

D. Implementation

Our implementation is summarized because the following four steps: Step one. Train and check the model to urge the MSE for every of the 5 models. Our current methodology primarily selects the minimum MSE. We tend to assign one to the chosen model and zero to alternative models. Step 2 select the model that has the bottom MSE therein bound amount for instance, in Table II, we decide Ridge regression as our model to pick stocks on Jun. 1st,

TABLE II

MODEL ERROR AND SELECTED MODEL FOR SECTOR 10, ENERGY

Trading Date	MSE Linear	MSE RF	MSE RIDGE	MSE Step	MSE gbm
19950601	**0.02238**	**0.02180**	**0.02161**	**0.02205**	**0.02443**
1995091	**0.01908**	**0.01828**	**0.01870**	**0.01841**	**0.02098**
19951201	**0.01852**	**0.01641**	**0.01820**	**0.01855**	**0.01996**
19960301	**0.02040**	**0.01822**	**0.01981**	**0.01879**	**0.02192**
19960603	**0.02442**	**0.01885**	**0.02394**	**0.02340**	**0.02210**

1995. We decide Random Forest as our model to pick stocks on Sept. 1st, 1995 Step three. Use the anticipated come within the hand-picked model to choose up prime two hundredth stocks from every sector. We tend to predict next quarter come (predicted y) exploitation current data (test Xs) supported the trained model. during this example, in Table III we tend to use ridge foreseen come to choose stocks for the trade amount 1995/06/01, the chosen prime two hundredth stocks are: WMB, OKE, RRC, PXD, VLO, EQT, HES, BHI, MUR, and NE. we tend to then trade these stocks throughout the amount between 1995/06/01 and

TABLE III

PREDICTED RETURN ON TRADE DATE: 1995/06/01 SECTOR 10, ENERGY

	Linear return	RF return	Ridge return	Step return	GBM return
WMB	10.42%	5.12%	9.24%	9.01%	3.51%
OKE	7.55%	4.12%	7.42%	8.53%	2.56%
RRC	4.16%	7.01%	3.74%	4.84%	1.83%
PXD	4.63%	0.96%	3.66%	3.91%	0.23%
VLO	3.48%	2.99%	3.47%	4.04%	2.56%

EQT	2.47%	4.78%	2.34%	2.36%	1.83%
HES	1.80%	4.30%	1.61%	1.81%	0.38%
BHI	1.33%	-0.70%	1.15%	1.99%	-0.27%
MUR	1.11%	1.07%	1.01%	0.49%	0.38%
NE	1.16%	-6.33%	0.94%	0.85%	-2.21%

1995/09/01 within the second trade amount of 1995/09/01, in Table IV we tend to use random forest to choose the stocks, the highest two hundredth stocks are: CHK, SFS.1, WMB, RRC, VLO, BJS.1, MDR, PZE.1, HP, and CVX. We tend to then trade these stocks from 1995/09/01 to 1995/12/01. As for the stocks closely-held at previous quarters, like WMB, RRC, and VLO, we tend to simply got to use the portfolio weights to regulate their shares. Step 4. We tend to check on the corresponding models options and its coefficients or importance level to confirm that there are not any

TABLE IV
PREDICTED RETURN ON TRADE DATE: 1995/09/01 SECTOR 10, ENERGY

	Linear return	RF return	Ridge return	Step return	**GBM return**
CHK	0.97%	12.45%	2.17%	4.86%	0.03%
SFS.1	4.69%	7.35%	4.27%	4.49%	0.78%
WMB	5.87%	6.37%	5.15%	5.46%	0.77%
RRC	1.78%	6.13%	1.52%	1.76%	0.09%
VLO	2.50%	4.83%	2.65%	3.01%	0.77%
BJS.1	5.36%	4.23%	5.28%	6.38%	-0.71%
MDR	1.54%	3.96%	1.26%	1.14%	0.77%
PZE.1	0.56%	3.78%	0.55%	0.78%	-1.71%
HP	-1.63%	3.50%	-1.44%	-1.62%	0.77%
CVX	-0.73%	3.19%	-0.71%	-1.12%	0.09%

TABLE V
RIDGE COEFFICIENTS: 1995/06/01

Factor	Coefficient	Factor	Coefficient
ROA	0.205677	DPO	0.000004
GPM	0.116841	DSI	-0.000018
REVGH	0.081318	EM	-0.000405
(Intercept)	0.049237	WCR	-0.000422
NPM	0.015640	PS	-0.002042
CR	0.004123	PCFO	-0.003140
EPS	0.002562	LTDTA	-0.026227
QR	0.002322	PB	-0.032136
DE	0.000612	OM	-0.055619
EVCFO	0.000013	ROE	-0.078489
PE	0.000004		

a normal results. in Table V and VI, e.g. assign to any or all options. We tend to end these steps for all eleven GICS sectors. Then we tend to get a final table of all hand-picked stocks with its vellication name, foreseen returns for next quarter, and therefore the corresponding mercantilism periods to conduct portfolio allocation.

3. PORTFOLIO ALLOCATION AND RISK MANAGEMENT

Portfolio allocation is crucial to associate degree investment strategy as a result of its balance risk and come by modeling individual asset's weights. Mean-variance and minimum-variance square measure 2 typical strategies for portfolio allocation. They perform diversification by limiting mean, volatility and correlation inputs to scale back sampling error [16]. In our portfolio we have a tendency to used mean-variance and min-variance to make your mind up the weights of every stock, so use equal-weighted portfolio as our benchmark. We have a tendency to perform these strategies by Matlab monetary Toolbox-Portfolio Object [17].

A. Mean-Variance and Minimum-Variance Constraints

We 1st use mean-variance optimisation to portion the stocks we've got picked. In. Fig. 3. The yellow star on the curve is that the mean-variance result throughout our 1st trade time; the remainder of the points area unit stocks aforethought supported its foreseen come and variance. The result shows that our approach is legitimate. We tend to set the subsequent constraints for mean-variance:

- Expected come: foreseen return of next quarter.
- Variance matrix: use one year historical daily come.
- Long only: bound five-hitter and bound 1/3.
- Absolutely invest our capital: total of weights=100%.
- Take no leverage: Lower Budget = higher Budget = one.

B. Transaction Costs

Generally, fees for each trade are measured based on broker fees, exchange fees and SEC fees. In the real world scenarios, a fund or trading firm might have different execution costs for many reasons. Despite these possible variations in cost, after going

TABLE VI
RANDOM FOREST IMPORTANCE TABLE: 1995/09/01

Factor	Importance	Factor	Importance
PB	14.2818	PE	5.6404
EPS	9.0556	CR	5.5080
ROE	8.6929	ROA	5.3972
PS	8.4976	PCFO	5.1313
WCR	8.4904	EVCFO	4.6672
EM	7.3808	LTDTA	4.6447
GPM	7.2091	OM	4.2249
QR	7.1337	DE	3.3296
DPO	6.6917	DSI	2.4621
NPM	5.9743	REVGH	0.1632

several scenarios we consider our transaction cost to be 1/1000 of the value of that trade. We believe our fee assumption to be sufficient and reasonable for the study. We use the following formula to calculate the transaction cost: ni =1 $/St,i$ – St−1,$i/$ Pi 0.1% (5) where St,i is the shares we need to buy or sell share based on the portfolio weights at current time t and St−1,i is the shares left at previous time t − 1. Pi is the current stock price of stock i.

Risk Management

After the procedure of building portfolio and structuring with appropriate functions, we equip decision rules that would be applied to risk management of each trade. Fundamentally, due to the nature of long-only strategy, the risk was controlled internally through our portfolio optimization methods. We minimize variance and maximum Sharpe ratio, have limits on position sizes (maximum of position size is 5% of portfolio value), and don't take any leverage.

TABLE VII
IN SAMPLE DATA RESULT: 1995-2007

	Mean-Var	**Equally**	**Min-Var**	**S&P 500**
Annualized Return	13.17%	16.12%	13.29%	7.12%
Annualized Std	17.0%	16.4%	12.9%	13.8%
Sharpe Ratio	0.687	0.887	0.917	0.406

4. PERFORMANCE EVALUATION

We conclude that our theme for stock picks will generate a stronger result than the market portfolio will. If we tend to solely analyze the portfolio price performance from the figure, it's noticeable that the equally weighted portfolio, the benchmark, has higher price than min-variance and mean-variance portfolio. However, the portfolio price isn't the sole thought once choosing the best portfolio. We've 2 reasons to conclude that the min-variance portfolio could be a higher methodology within the real trade amount. First, the equally-weighted portfolio isn't sturdy enough. We tend to notice that the performance of this methodology totally depends on the expected returns that we tend to calculate from our model the expected returns can vary each time we tend to run our model. Secondly, min-variance portfolio allocation takes the chance issue into thought, creating it additional reliable in real trade. From the Table VII, we discover that the min-variance allocation contains a higher Sharpe quantitative relation than that of the equally weighted allocation throughout the in sample amount. We tend to so opt for the min-variance as our portfolio allocation methodology.

5. CONCLUSION

Applying machine learning algorithms to the basic monetary knowledge will separate stocks with a relative dangerous earnings, therefore providing a far better thanks to choose stocks. Minimum-variance technique, the five % holding rule, no short and leverage rule give risk management and diversification, scale back the portfolio risk and therefore yield a better Sharpe magnitude relation. Compared to the benchmark, our commerce strategy outperforms the S&P five hundred index, additional significantly, combined with our commerce strategy, the portfolio allocation technique is well-tried to boost the performance.

TABLE VIII
OVERALL PERFORMANCE

(Risk-Free: 1.5%)	**Mean-Var**	**Equally**	**Min-Var**	**S&P 500**
Start Value in million	**1**	**1**	**1**	**1**
End Value in million	10.9917	22.1383	12.81498	1.933153
Total Return	999.17%	2113.83%	1181.50%	93.32%
Maximum Drawdown	−56.89%	−57.63%	−46.30%	−66.73%
Annualized Return	8.29%	10.77%	9.87%	5.22%
Annualized Std	23.6%	26.4%	18.1%	19.1%
Sharpe Ratio	0.287	0.351	0.462	0.195

.

Finally, the Sharpe ratios of the 3 portfolio strategies indicate that our strategy additionally outperforms the market. Future work would be handling a normally knowledge [19] within the knowledge pre-processing stage, and applying correct prediction schemes by modelling stock indicators as tensor statistic [20] with meagreness in rework domains.

6. REFERENCES

1. Joseph J. Gerakos and Robert Gramacy, “Regression-based earnings forecasts,” Chicago Booth Research Paper No. 12-26., 2013.

2. Jagadeesh D. Pujari, Rajesh Yakkundimath, Abdulmunaf S. Byadgi. 2015.

3. Radim Gottwald, “The use of the p/e ratio to stock valuation,” vol. 415, 2012.

4. Benjamin Graham, Sidney Cottle, Roger F. Murray, and Frank E. Block, Security Analysis, Fifth Edition, McGraw-Hill, 1988.

5. Maria Crawford Scott, “Value investing: A look at the benjain graham approach,” AAII, 1996.

6. Compustat Industrial [daily and quarterly Data]. (2017). Available: Standard Poor's/Compustat [2017]. Retrieved from Wharton Research Data Service., ," .

7. Trevor Hastie, Robert Tibshirani, and Jerome Friedman, The Elements of Statistical Learning, 2nd Edition, Springer, New York, NY,2009.

8. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, An Introduction to Statistical Learning with Applications in R,Springer, New York, NY, 2013.

9. Iain Johnstone Bradley Efron, Trevor Hastie and Robert Tibshirani, "Least angle regression," The Annals of Statistics 2004, Vol. 32, No. 2, 407499.

10. Andy Liaw and Matthew Wiener, "Classification and regression by randomforest," R News 2 (3), 1822., 2002.

11. Greg Ridgeway, "Generalized boosted models: A guide to the gbm package," 2007, URL https://CRAN.R-project.org/package=gbm.

12. Andrew Ang, "Mean-variance investing," Columbia Business School Research Paper No. 12/49., August 10, 2012.

13. MathWorks, "Matlab financial toolbox: Portfolio object," 2017, https://www.mathworks.com/help/finance/portfolio-object-mv.html.

14. "Our codes," http://www.tensorlet.com/.

15. Xiao-Yang Liu and Xiaodong Wang, "Ls-decomposition for robust recovery of sensory big data," IEEE Transactions on Big Data, 2017.