Use of Machine Learning to Help Investor Choose Stock Securities in Indonesia Capital Market

Anindhito Irmandharu
Computer Science Departement,
School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia
anindhito.irmandharu@binus.ac.id

Dewi Suryani
Computer Science Departement,
School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia
dewi.suryani@binus.ac.id

Buying stock securities by evaluating public companies is no small feat. Not only investor must do due diligence on researching the company they want to buy but they also need to choose from hundreds of public companies that are listed in the Indonesian Stock Exchange (IHSG). In this research we propose a machine learning assisted approach to evaluate public companies listed in IHSG based on their financial statements and uses IHSG as benchmark. Companies that have their stock price perform above IHSG will be marked as "superior stocks" and companies that is underperforming IHSG will be marked as "inferior stocks". Our method can correctly predict whether the company stock price movement will be above or below IHSG in 1 year period in 86,12% of the cases.

Keywords: Machine Learning, Stock Market, Classification

I. INTRODUCTION

The capital market is a market for various long-term financial instruments that can be traded, both debt securities (bonds), equities (stocks), mutual funds, derivative instruments, and other instruments. The capital market is an investment vehicle for companies and other institutions (e.g., the government), and as a means for investing activities. Thus, the capital market facilitates various facilities and infrastructure for buying and selling activities and other related activities (Bursa Efek Indonesia, 2018).

However, the existence of information technology innovation in the financial sector also makes it easier for new companies to go public and list on the Indonesia Stock Exchange. In 2016 there were 16 new companies that went public, and the number of companies that wanted to go public continued to show an upward trend until 2019 with a total of 55 companies. Of course, the increase in the number of companies that are listing is a good sign for the Indonesian capital market, but the more companies that list, the more choices of shares that can be purchased. As of December 2020, there are 713 companies listed on the Indonesia Stock Exchange (CNBC, 2020).

The large number of total shares and the rapid growth of new shares listed each year can overwhelm retail and institutional investors due to the large amount of data that must be digested to decide on which shares to buy among hundreds of companies listed on the Indonesia Stock Exchange. A study conducted by DALBAR which is a leading independent expert in the financial community to evaluate, audit, and assess business practices, customer performance, product quality, and services found that from 1984 to 2019 approximately 70% of individual capital market investors in the United States had returns below the index (S&P500). Research conducted by DALBAR shows that the method of placing their funds in the index and keeping them quiet performs better than the average individual investor in the capital market (DALBAR, 2019).

This problem does not only apply to individual investors but also to professional actors such as mutual funds managed by investment managers. A study conducted by S&P Dow Jones Indices (SPIVA) found that 57% of fund managers in the United States failed to beat the index in a 1-year period, and 73% of mutual funds failed to beat the index in a 5-year period. Studies conducted by SPIVA indicate that investment funds managed by professionals are not guaranteed and have a high probability of beating the index (SPIVA, 2020).

Therefore, the main goal of this paper is to create a machine learning model that will help individual investor and institutional investor in classifying stock securities where their performance is above index or known as "superior stocks" while discarding "inferior stock", a stock which underperform the index.

II. RELATED WORKS

Several research on the use of Artificial Intelligence at the financial industries specifically the stock market had been done and show promising results. A study held by a researcher from Stanford University explores how the short-term predictive power of machine learning models can reap financial benefits for investors who trade based on predictions of future prices. This study focuses on the problem of binary classification, predicting the next minute price movement of the S&P 500 Index and acting on the insights generated from our model. This research implements several machine learning algorithms which are Logistic Regression, Support Vector Machine (SVM), Long short-term memory (LSTM) and Convolutional Neural Network (CNN).

Of all the models, the SVM with a polynomial kernel performed the best across all metrics. Most of the SVM methods outperformed deep learning methods (LSTM, GRU,

and CNN) with this study's data set (approximately 20,000 data points). All models except SVM (Sigmoid Kernel) model, Single-layer LSTM, and CNN model posted positive gains. Using 15 basis points (BPS) transaction fees, we find that transaction fees erode all our profits, resulting in negative returns for all strategies (models). This is mainly because the transaction fee associated with the complete process of entering and exiting positions is 30 bps or 0.3%. However, because we trade in a minute range, short term price movements in minutes rarely exceed 0.3%. That is, even if our predictions are correct, the researchers are still at a loss (Art Paspanthong, 2019)

Another study which focuses on the selection of stocks for the long term based on past financial reports has been carried out by Yuxuan Huang from the University of Western Ontario (Huang, 2019). This study uses the Feed-Forward Neural Network (FNN), Adaptive Neural Fuzzy Inference System (ANFIS) and Random Forest (RF) algorithms to predict the relative returns of stocks with quarterly periods. This research dataset covers 22 years of financial statements and consists of companies listed on the S&P100 index and selected 70 stocks in the index. In this study the researcher argues for choosing a fundamental and long-term approach because the fractional or commission fees of trading can deplete the potential returns. Our target variable in this study is quarterly relative returns, while many features of the raw data set have global trends that have time to go.

As researchers transfer these time series problems into supervised learning problems, these globally trending features can hinder the machine learning model's ability to generalize and provide reliable predictions. The researchers therefore took the percentage change (delta) between successive observations. After the dataset is processed, a total of 21 features are used and each stock has 88 observation points, starting from Q1 1996 to Q4 2017, with an interval of one quarter between two consecutive observations. To evaluate the performance of various machine learning methods, researchers ranked 70 stocks based on their predicted relative returns. Portfolios are built based on ratings and the actual relative results of the portfolios are used as evaluation criteria. In addition, the relative compounded return over the 18-quarter test period is calculated for each method as an additional metric. All Buy portfolios managed to beat the S&P100 by a large margin, but when compared to other models, RF had the best portfolio performance followed by FNN, while ANFIS had the worst performance when compared to other models. The detail about return of investment based on the machine learning implementation can be seen at Table. 1.

TABLE 1: PORTFOLIO RETURN

Model	Portfolio Return
FNN	14.4%
ANFIS	8.85%
RF	32.1%
SP 100 Index	1.35%

Another study to classify and select stocks for long-term investment was conducted by researchers from the University of Manchester, UK (Milosevic, 2016). In this study 1739 stocks were selected with a dataset starting from Q1 2012 to Q4 2015. The main hypothesis of this research is that by applying machine learning and training it on past data, stock price movements can be predicted, as well as the ratio of movements over a certain period. This study aims to predict stock price movements as a classification task, in which they classify stocks that will have a price 10% higher in one year period as “Good” stocks and others as “Bad”. The data that will be used to train machine learning algorithms is financial statement data and is fundamental. There are several classification algorithms used to train machine learning models. The researcher trained the model using C4.5 Decision Tree, Support Vector Machines with Sequential Minimal Optimization, JRip, Random Trees (RT), Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB), and Bayesian Networks (BN).

First, cross-validation is carried out 10 times on all the algorithms that have been mentioned. Then they perform manual feature selection by removing features and evaluating whether algorithm performance improves or decreases. They do this process iteratively, until they don't get the optimal model with the least number of features and the best performance. When all implementation has been tested RF performed the best compared to all the implementation that have been tested. The detail of testing result can be seen on Table. 2 below.

TABLE 2: MODEL PERFORMANCE

Model	Precision	Recall	F-Score
C4.5 Decision Trees	0.660	0.660	0.660
SVM with SMO	0.636	0.629	0.624
JRip	0.640	0.639	0.639
Random Tree	0.700	0.700	0.700
Random Forest	**0.765**	**0.765**	**0.765**
Logistic Regression	0.638	0.630	0.625
Naive Bayes	0.526	0.515	0.453
Bayesian Networks	0.641	0.626	0.615

III. DATASET

In the data collection stage, companies that have financial statements above or equal to 10 years are recorded, from the company selection process, 283 companies that meet the criteria are obtained. After getting a list of companies that meet the criteria, a web scraping process is carried out to obtain data on the company's financial statements. The financial report data taken is on annual format and can be divided into three main parts, namely Income Statement, Balance Sheet and Cashflow Statement. In total there are 62 features obtained for the financial statements. On the next phase which are the cleaning stage of company data, the company's financial statements are loaded into the Dataframe and transposed so that the data has the appropriate axis when training. After performing the transposition process, the character manipulation process is carried out to remove the string with the help of regex. Once the character manipulation process is complete, all financial statement data is converted to numeric data type so that it can be read by machine learning algorithms later. From the cleaning stage it is found that

several features have data that is Missing Not at Random (MNAR) which later are dropped.

Furthermore, the data imputation process is carried out with KNNImputation which is obtained from the sklearn library with a neighbor value = 4 to fill in the remaining missing data. In the data exploration process, it was found that most of the data had an increasing trend (Example: The income of most companies had an increasing trend of data), therefore a replacement variable was made which was the delta of the percentage change for some data that was considered to have an upward trend.

After the preprocessing process is complete and the data is in the desired format, then the data labeling process is carried out. If in a certain year a company has a stock price movement (Performance) above the Composite Stock Price Index (IHSG) then the previous year's financial statement data will be labeled True, and if the performance is positive but still below the IHSG then the company will still be labeled False. To find out if a company has a performance above or below the IHSG, it is determined from the percentage increase/decrease in the adjusted close share price in January of year N to January of year N-1. If the percentage increase in stock prices is above the IHSG, it will be labeled True and False if the percentage increase is below the IHSG. If in a certain year the IHSG has a negative return, the companies whose decline is smaller than the IHSG will be labeled True and those with a greater decline than the IHSG will be labeled False.

Because the data obtained from web scraping have unbalanced classes, a balanced dataset variation is made with the help of Synthetic Minority Oversampling Technique (SMOTE) to balance the number of classes for the preprocessed dataset. This class balancing process is carried out for training data only. The dataset is divided into two parts, namely training and testing. The training dataset consists of the years 2009 – 2016 and 2017 – 2020 for data testing

IV. Machine Learning Implementations

In this study there are 3 variations of the supervised learning algorithm used: Support Vector Machine (SVM), Random Forest (RF) and Deep Neural Network (DNN). These three algorithms were chosen because some of them are the best algorithms in previous studies.

The first algorithm used to classify stocks is Support Vector Machine (SVM). In the SVM algorithm, there are several parameters that can be changed for which the results can be compared later to choose the best parameters with a grid search cv. Parameters that can be changed are kernel and gamma. To find the best hyperparameter combination of the SVM algorithm, a gridsearch-cv process where k=5 is executed based on the variation of hyperparameter that can be seen on Table. 3. This hyperparameter searching process is repeated for both unbalanced dataset and balanced dataset.

TABLE 3: SVM Hyperparameter variation

Hyperparameter	Variation
Kernel	Linear | Poly | RBF | Sigmoid
C	10 | 20 | 100 | 500
Gamma	0.1 | 0.5

The next machine learning algorithm that is used in this research is Random Forest (RF). RF has parameters n_estimators, min_samples_split, max_leaf_nodes, max_features, max_depth, and boostrap. A gridsearch-cv process where k=5 is carried out to find best hyperparameter. The variation of hyperparameter that is used on RF during the gridsearch-cv process can be seen on Table. 4. Just like SVM implementation this process is of searching hyperparameter is executed both for the unbalanced and balanced dataset.

TABLE 4: RF Hyperparameter variation

Hyperparameter	Variation
max_depth	2 | 4 |8
max_features	Sqrt | Log2
min_samples_leaf	3 | 4 | 5
min_samples_split	8 | 10 | 12
n_estimators	100 | 400 | 800

In addition to using SVM and RF, this study also uses the Deep Neural Network (DNN) algorithm to classify stocks. The DNN algorithm training process begins by using the Tensorflow 2 and PyTorch libraries. Furthermore, the process of making the DNN architecture in both libraries uses the same hyperparameters with the aim of comparing the performance of PyTorch and Tensorflow 2. After the architecture is created, a cross-validation process is carried out with k=10 to assess its performance. From the cross-validation results, it was found that Tensorflow 2 has a slightly superior performance compared to PyTorch, therefore this research will use Tensorflow 2. Several

After the comparison process of Tensorflow 2 versus PyTorch is complete, the final DNN model will be created using Tensorflow 2. The DNN model will be based on the Multi-Layer Feed Forward Neural Network architecture. This DNN has hyperparameters consisting of a hidden layer, activation function, optimizer, epoch, learning rate, loss function and batchsize with variations for several attributes. Each combination of hyperparameters will be tried to be trained to find the best combination of hyperparameters with the help of gridsearch-cv with the value k=5.Due to the complex architecture of DNN algorithm the gridsearch-cv are separated into several iteration where optimizer and activation function are searched separately. Once the best optimizer and activation function is obtained the process of gridsearch-cv for epoch, learning rate, batchsize and layer is executed where the variation detail can be found at Table. 5. This process is carried out for both variations of the dataset, namely the balanced dataset and the unbalanced dataset.

TABLE 5: DNN Hyperparameter variation

Hyperparameter	Variation
Epoch	100 | 200 | 400 | 1000
Learning rate	0,00001 | 0,00003 | 0,00005
Batchsize	16 | 64 | 128
Layer	[128,64] | [128,64,32] | [256, 128, 64]
Loss function	Binary Cross-Entropy

V. TRAINING RESULT

Once the hyperparameter search is done an the best hyperparameter combination obtained, the model of each implementation is trained with the training dataset using the best hyperparameter combination. During the training process cross-validation k=10 is implemented. The results achieved are presented in Table. 6.

TABLE 6: TRAINING DATASET RESULT WITH CROSSVALIDATION K=10

Model	Accuracy	F1
SVM (Original Dataset)	51.34%	35.63%
SVM (Balanced Dataset)	58.96%	28.38%
RF (Original Dataset)	74.68%	68.76%
RF (Balanced Dataset)	74.86%	68.81%
DNN (Original Dataset)	81.09%	76.92%
DNN (Balanced Dataset)	84.24%	80.18%

From Table. 6 DNN model that is trained with balanced dataset has the highest accuracy and F1 when compared to other model variation. Once the best model is found, the model is tested with a dataset that has never been used, namely testing dataset. The model will be tested with testing dataset without cross validation and the result can be seen on Table. 7

TABLE 7: TESTING DATASET RESULT

Model	Accuracy	F1
DNN (Balanced Dataset)	86.12%	90.44%

VI. CONCLUSION

With machine learning algorithm that has been made, the model classifies 73 superior stocks in 2018, 50 superior stocks in 2019, and 83 superior stocks in 2020. This means the machine learning model has reduced about 70% - 80% of the number of stock options so that capital market investors can focus more on the classified superior stocks.

Although there are still False Positive and False Negative results from the machine learning model choices, but if a collection of stock is made based on the recommendation DNN model and the funds are divided equally among the selected stocks, it can beat the Composite Stock Price Index (IHSG) in terms of Return on Investment (ROI) for 2018, 2019 and 2020 where the choice of models resulted in an ROI of 58%, 20.2% and 54% while the JCI performed -2.54%, 1.70% and -5.09% in the same year.

It is concluded that cclassification of stocks that have above average returns can be done by machine learning where the Composite Stock Price Index (IHSG) is used as a benchmark. The Deep Neural Network (DNN) model is the stock classification model with the highest accuracy and F1 followed by Random Forest (RF) and Support Vector Machine (SVM) in the last position.

The limitation of this study is that the dataset used for this study can be considered short with financial data up to 10 years however as noted in chapter 2 there are previous work which has over twice the amount when compared to this study. Besides the dataset this study only focuses on 3 machine learning implementations therefore there is room improvement should a different or new machine learning algorithm is tested.

REFERENCES

[1] Art Paspanthong, N. T. (2019). Machine Learning in Intraday Stock Trading. 1-4.

[2] Bursa Efek Indonesia. (2018). *Pengantar Pasar Modal.* Retrieved from IDX: https://www.idx.co.id/investor/pengantar-pasar-modal/

[3] CNBC. (2020, 12 30). *CNBC.* Retrieved 03 05, 2021, from https://www.cnbcindonesia.com/market/20201230142438-17-212584/rekor-tutup-tahun-ipo-tembus-51-investor-capai-38-juta

[4] DALBAR. (2019). *2020 QAIB Report.* DALBAR.

[5] Huang, Y. (2019). Machine Learning for Stock Prediction Based on Fundamental Analysis. *Machine Learning for Stock Prediction Based on Fundamental Analysis*, i-50.

[6] Milosevic, N. (2016). Equity forecast: Predicting long term stock price movement using machine learning. *Equity forecast: Predicting long term stock price movement using machine learning*.

[7] SPIVA. (2020). *SPIVA® U.S. Scorecard.* S&P Dow Jones Indices.