Stock Performance Classification in Stock Exchange of Thailand (SET) by Using Supervised Machine Learning Model

Chayanant Kosol and Punnamee Sachakamol

***Abstract*—Most investors decide to invest in a stock market in order to win from an inflation. And, Financial Statement is the top tool that Thai investors have been using a financial statement to support their buying/selling decision in the stock market for a long time. Even in a digital era as nowadays, the financial statement is still in use by many investors. Particularly, they manually review the statement and make a decision based on their own judgement.**

The purpose of this research is to use a proper of technology and financial statement to build classification models to identify a winning stock in the Stock Exchange of Thailand (SET).

***Index Terms*—Classification algorithms, logistic regression, K-nearest neighbors, support vector machine, supervised machine learning, stock market, financial statement.**

I. Introduction

Like many other countries around the world, a cost of living in Thailand increases every year, while the banks deposit interest rate decreases significantly. This implies that the value of money decreases over time.

Fig. 1 and Fig. 2 demonstrate a meal price from McDonald's in Thailand from 2011 to 2017 [1] and the deposits interest rate from the bank in Thailand from 2011 to 2017 [2].

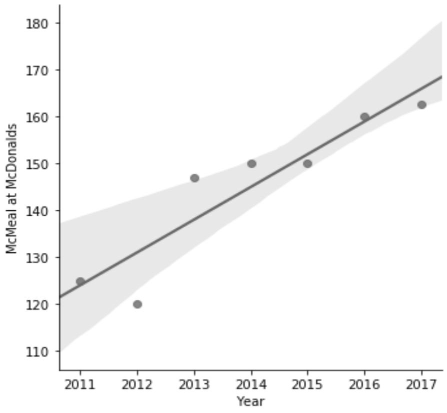


Fig. 1. Line chart of meal price from McDonald's.

From Fig. 2, the line chart, the highest deposit interest rate is 2.5 in 2011 and it have decreased every year around 2 percent, whereas the meal price from Fig.1 has increased about 5 percent every year. Amount of money in the bank will be less when the time has passed each year, this is called "decreasing of buying power" or "increasing of the inflation rate". Instead of saving money in the bank, people choose to invest in a fund in order to avoid this problem.

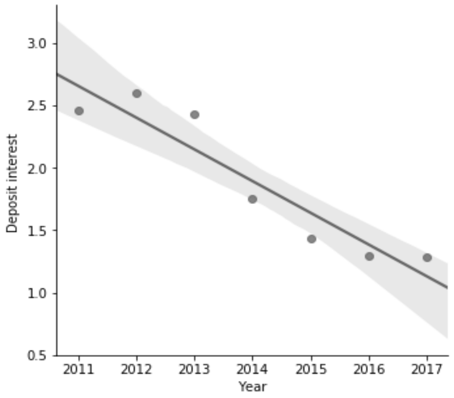


Fig. 2. Deposit interest rate from the bank in Thailand from 2010 – 2017.

A box plot [3] of a performance of 5 stars fund from Morningstar Thailand that has been operated for 7 years [4] is provide in Fig. 3.

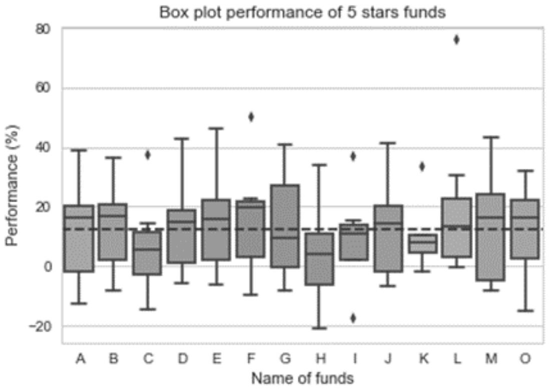


Fig. 3. Box plot performance of 5 stars funds.

Form the boxplot, the average return of almost fund reach is high about 12.5 percent, but the range of return below 0 percent also high either. Therefore, the problem became to handing with the fluctuation of return. To solve this problem, the objective of this paper is to find the method that drives instrument in order to generate the return more than 12.5 percent and reduce the fluctuation below 0 percent.

The main purpose of this study is to use the financial data to build classification models to distinguish between a

Manuscript received January 28, 2019; revised December 14, 2019.

Chayanant Kosol is with the Engineering Management, Kasetsart University (KU), Thailand (e-mail: chayanant.ko@ku.th).

Punnamee Sachakamol is with the Department of Industrial Engineering and Engineering Management, Kasetsart University (KU), Thailand (e-mail: fengpmsa@ku.ac.th).

doi: 10.18178/ijmlc.2020.10.2.922

winner stock and a loser stock. The winner stock is the stock whose price increases more than 5 percent in one quarter and the loser stock is the stock whose price increases less than 5 percent in one quarter. Afterward, it could generate the return more than 12.5 percent and reduce the risk that return will less than 0 percent.

II. Data Preparation

We first collet the financial data online by using Python and prepare it in a suitable format, we then split the whole data into 2 part as shown in Fig. 4.

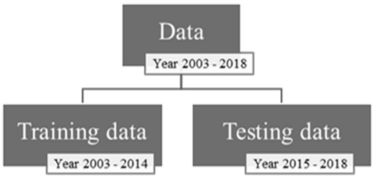


Fig. 4. Splitting the whole data into 2 part.

The training data that is used to fit model is selected from 2003 to 2014 due to a market cycle assumption. In this period there are a downtrend market, an uptrend market and no trend market which cover all the market cycles [5].

As models should be evaluated on any data that has not been used in training process, we ser apart 2015 to 2018 data as a test set. In particular, we apply the trained models on this test data to assess the performance among models.

III. Data Exploration

A. Winning Ratio

According to the assumption of used in this paper, the winning stock is defined as any stock whose price increases more than 5 percent in a quarter. On the contrary, the loser stock's price decreases or increases no more than 5 percent in a single quarter. The ratio of winning stock and loser stock is shown in Fig. 5.

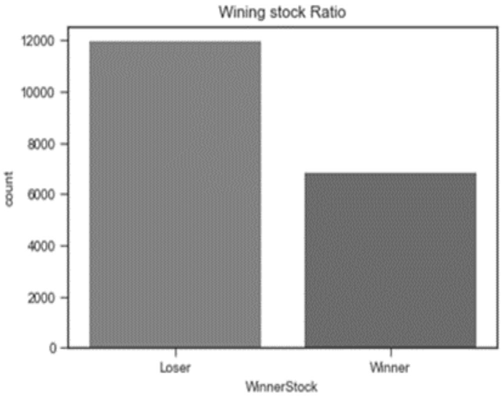


Fig. 5. The ratio of winning stock and loser stock.

The percentage of the loser stock is 63.55 whereas the percentage of winner stock is 36.44. The winning ratio is around 2 loser stock per 1 winning stock.

B. Distribution of Factor

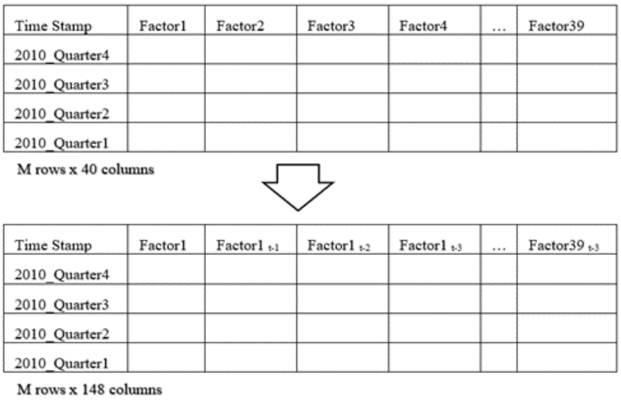
The financial time series data in Thailand is not publicly for an individual investor. Accessing this data incurs a large cost. Due to this limit, the factors that could be scraped from the website are 39 factors and the sample of financial factors as shown in Table I.

To increase the factor in the model, researcher shift the factors back from current quarter to 3 quarter back, the sample of spreading out the factors is shown in Fig. 6. After this step, total factor is 147 factors.

To overview the effect of each factor, the first step is plotting Kernel Density Estimation Plot (KDE) to estimate the probability density function of the factors [6]. The emphasis factors as shown as Fig. 7, Fig. 8, Fig. 9 and Fig. 10.

TABLE I: Sample of Financial Factor

Number	Factor
1	Accounts Receivable Turnover
2	Average Collection Period
3	Average Sale Period
4	Book Value per Share
5	Cash
6	Close Price
7	Cost of Sales
8	Current Liabilities
9	Current Ratio
10	Debt to Equity
11	Dividend Yield
12	Earnings Before Interest and Taxes
13	Earnings per Share
14	Equities
15	Fixed Asset Turnover



Time Stamp	Factor1	Factor2	Factor3	Factor4	...	Factor39
2010_Quarter4						
2010_Quarter3						
2010_Quarter2						
2010_Quarter1						

M rows x 40 columns

Time Stamp	Factor1	Factor1 $_{t-1}$	Factor1 $_{t-2}$	Factor1 $_{t-3}$	...	Factor39 $_{t-3}$
2010_Quarter4						
2010_Quarter3						
2010_Quarter2						
2010_Quarter1						

M rows x 148 columns

Fig. 6. The sample of spreading out the factors.

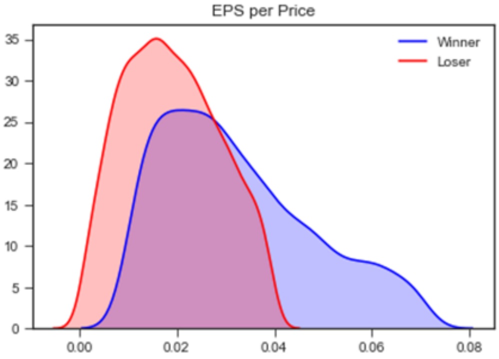


Fig. 7. KDE plot of EPS per Price factor.

From difference of all distribution between winner stock and loser stock above, the factors could be a good predictor of the outcome variable.

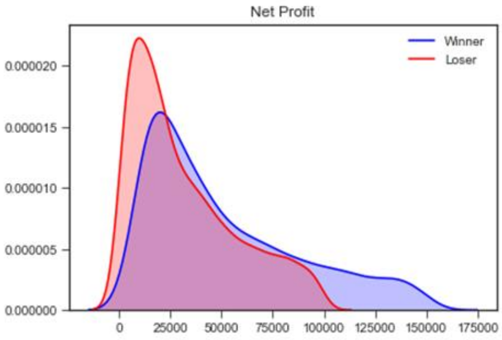


Fig. 8. KDE plot of net profit factor.

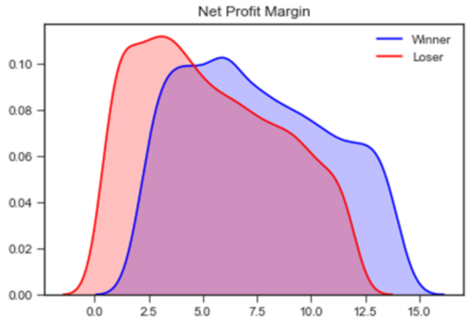


Fig. 9. KDE plot of net profit margin.

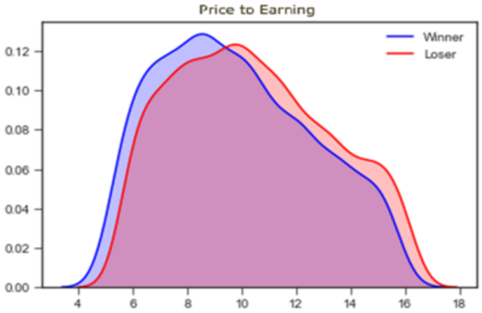


Fig. 10. KDE plot of price to earning factor.

IV. Modeling

After data exploration step is choosing the proper classification model such as Logistic Regression, K-Nearest Neighbors and Support vector machine [7]. The researcher selects these methods because of, it has been widely used and simple to understand.

A. *Logistic Regression*

The importance advantage of this model is probability of prediction of a winning stock. The sample of the prediction probability as shown in Table II.

TABLE II: Sample of the Prediction Probability

Time Stamp	Symbol	Probability of Winning Stock
2010_Quarter4	Stock A	0.6070
2010_Quarter3	Stock A	0.5891
2010_Quarter2	Stock A	0.6065
2010_Quarter4	Stock B	0.6565
2010_Quarter3	Stock B	0.6294

After trained the model with default parameters, the classification report as shown in Table III.

When considering the average of Precision, Recall, and F1-Score. All the parameters above are acceptable but if the purpose of this study is predicting the winning stock in the stock market, the number in Table III are all rejected. Because the value of recall parameter for winning stock is low about 0.06 which means the model cannot reach the winning stock in training data.

TABLE III: Classification Report of Training Data Using Logistic Regression with Default Parameters

	Precision	Recall	F1-Score	Support
Loser	0.65	0.98	0.78	11952
Winner	0.62	0.06	0.12	6854
Average / Total	0.63	0.64	0.54	18806

To solve the problem, changing the Probability Thresholds is the first way to consider. The model cannot reach the winning stock in the dataset because by default, the Probability Thresholds is set to 0.5, which is if the predicting Probability is higher than 0.5 that observation will be the winning stock but if not it will be loser stock. Recall parameter for winning stock is low because the ratio between the number of observations that higher than 0.5 and the number of observations that higher than 0.5 is definitely low.

To change the Probability Thresholds to obtain a better model, the Overall F1-Score versus each Probability Thresholds and Winner Recall by each Probability Thresholds as shown as Fig. 11 and Fig. 12.

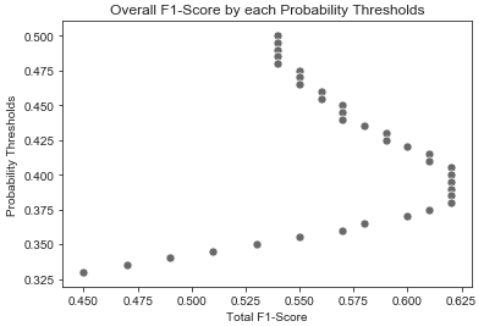


Fig. 11. Overall F1-score versus each probability thresholds.

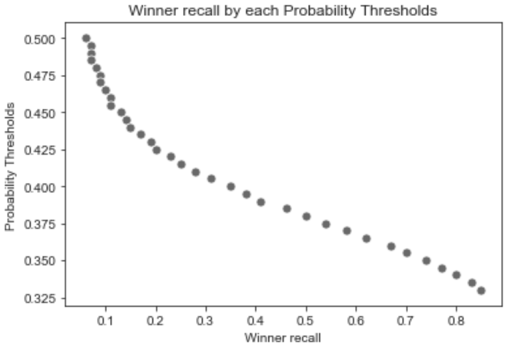


Fig. 12. Winner stock recall versus each probability thresholds.

From Fig. 11, the Probability Thresholds that give highest overall F1-Score are 0.405, 0.400, 0.395, 0.390, 0.385 and 0.380. From Fig. 12, the decreasing in Probability Thresholds is associated with an increase in Recall parameter for winning stock. Therefore, when considering specifically 6 numbers above, the number that gives the winning stock recall highest is 0.385. Finally, after changing the Probability Thresholds to 0.385. The classification report of training data by using Logistic Regression as shown in Table IV.

TABLE IV: CLASSIFICATION REPORT OF TRAINING DATA USING LOGISTIC REGRESSION

	Precision	Recall	F1-Score	Support
Loser	0.68	0.81	0.74	11952
Winner	0.52	0.35	0.42	6854
Average / Total	0.62	0.64	0.62	18806

At present, the classification report seems better than before.

For any winner stock, the Precision is defined as the ratio of true positives [8] with respect to all instances predicted as positive. The Precision of the winner stock in the training data is 0.52 by following equation (1):

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \qquad (1)$$

For any winner stock, the recall is defined as the ratio of true positives with respect to a total number of true instances. The recall of the winner stock in the training data is 0.35 and Recall score is defined as follows:

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \qquad (2)$$

F1- Score is q weighted harmonic mean of Precision and Recall. It is ranging between 0 (worst) and 1 (best). The F1-Score of winner stock in training data is 0.42 by following equation (3):

$$F1 - \text{Score} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right) \qquad (3)$$

From all of the 3 measurement parameters above, the model seems to be reasonable enough. However, what happens if the model meets the data that it never has seen before by using the model that trained in training data with testing data.

The classification report of testing data by using Logistic Regression as shown in Table V.

From Table V, Precision of winning stock decrease 25 percent from 0.52 to 0.39 and Recall of winning stock decrease 14.29 percent from 0.35 to 0.3, it demonstrates that the ability of the model to reach winning stock is decreased. And, F1-Score of winning stock decrease 19 percent from 0.42 to 0.34.

Fig. 13 illustrates the confusion matrix [9] of the Logistic Regression model in training data.

Finally, Table VI is the highlight of the Logistic Regression model, the 10 samples of the coefficient factor as shown in Table VI which is from 5 highest coefficient and 5 lowest coefficient.

TABLE V: CLASSIFICATION REPORT OF TESTING DATA USING LOGISTIC REGRESSION

	Precision	Recall	F1-Score	Support
Loser	0.75	0.82	0.78	5812
Winner	0.39	0.3	0.34	2308
Average / Total	0.65	0.64	0.66	8120

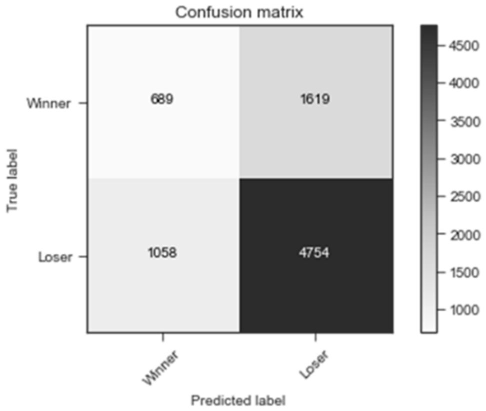


Fig. 13. The confusion matrix of logistic regression model.

TABLE VI: 14 SAMPLE OF COEFFICIENT OF EACH FACTOR

Factor	Coefficient
Return on Asset	0.3607
Earnings per Share	0.2692
Revenue per Market Capitalization t-1	0.1615
Dividend Yield t-1	0.1373
Value/Day t-3	0.1171
Dividend Yield t-2	0.0931
Gross Profit Margin	0.0921
Close Price t-2	-0.1039
Earnings per Share t-2	-0.1108
Net Profit Margin	-0.1121
Fixed Asset Turnover t-1	-0.1135
Net Profit Margin t-1	-0.1244
Total Asset Turnover	-0.1245
Return on Asset t-1	-0.1807

From Table VI, the top 3 factors that impact the model the most are Return on Asset, Earnings per Share and Revenue per Market Capitalization from the previous quarter. The interesting is if consider only current period people will surprise that the popular factor such as Total Asset Turnover and Net profit have a negative coefficient in the model.

TABLE VII: COEFFICIENT OF POPULAR FINANCIAL FACTORS

Financial Factor	Period t	Period t-1	Period t-2	Period t-3
Total Assets	-0.0212	0.0191	0.009	0.0167
Total Liabilities	-0.0143	0.0236	0.015	0.0077
Equities	-0.0487	-0.0068	-0.0374	0.0372
Revenue	-0.0565	-0.0025	0.0321	0.0321
Net Profit	0.0045	-0.0128	-0.0063	0.01786
Earnings per Share	0.269	0.0369	-0.1108	0.0125
Return on Asset	0.3608	-0.1807	-0.041	-0.0815
Return on Equity	0.0033	-0.0225	-0.024	-0.0197
Net Profit Margin	-0.1122	-0.1244	-0.0039	0.0388
Price to Earning	0.0019	0.0571	0.0208	0.02659
Price to Book Value	0.0448	-0.0483	-0.0529	0.0239
Dividend Yield	-0.086	0.1374	0.0931	-0.081

To dive deeper into the truth, Table VII contains the coefficient from a different time period of the popular financial parameter that Thai investors have used for a long time [10]. And, the red text represents the negative coefficient and if the coefficient is low the probability to be a winning stock will lower either.

First, If a focus on period t, from 12 popular financial factors. Only 6 factors that positive coefficient namely Net profit, Earnings per Share, Return on Asset, Return on Equity, Price to Earning and Price to Book Value. Moreover, the top of financial factors like Net Profit Margin has a negative coefficient for period t, period t-1, and period t-2.

From Table VII, the safes factor to be used in every period is Price to Earning because it has a positive coefficient in every period of time. The most dangerous factor is Net Profit Margin because from the highest negative coefficient number one and two of the table are from Net Profit Margin. The highest positive coefficient is Return on Asset from period t.

B. K-Nearest Neighbors

After optimizing K value from 1 to 10, the highest accuracy is 0.6532 with k equal to 8. The accuracy in K-Nearest Neighbors model from K 1 to 10 as shown as Fig. 14.

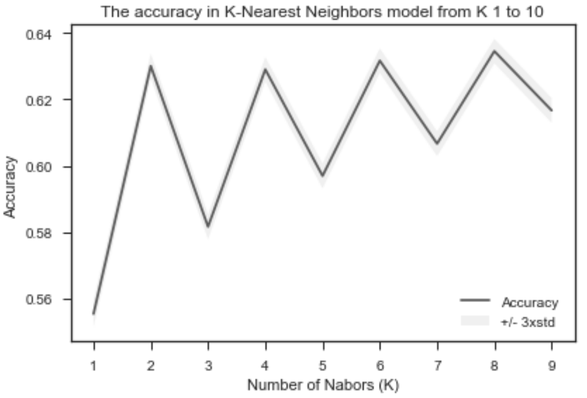


Fig. 14. The accuracy in K-Nearest Neighbors model from K 1 to 10.

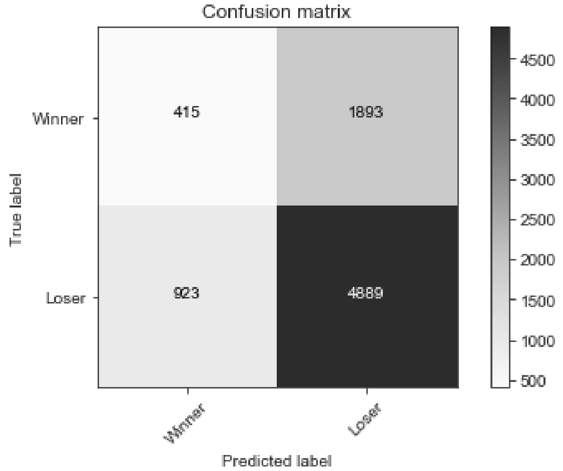


Fig. 15. The confusion matrix of K-Nearest Neighbors model.

After running the model with K equal to 8, the classification report of training data and testing data by using K-Nearest Neighbors model as shown in Table VIII and Table IX.

Average Precision of training data is 0.71, however, it decreased to 0.60 in the testing data, especially Precision of winning stock decreased by 50%, from 0.72 to 0.31. It means, the model is definitely fit with the training and definitely sensitive when it faces to the data it hasn't seen before.

In the same way, the recall of the winner stock in the training data is 0.29 which is already low, however, it decreases to 0.18 in the testing data.

F1-Score of the winner stock in training data is 0.41 but it decreases to 0.23 in the testing data.

The confusion matrix of K-Nearest Neighbors model is shown in Fig. 15.

TABLE VIII: CLASSIFICATION REPORT OF TRAINING DATA USING K-NEAREST NEIGHBORS MODEL

	Precision	Recall	F1-Score	Support
Loser	0.70	0.94	0.80	11952
Winner	0.72	0.29	0.41	6854
Average / Total	0.71	0.70	0.66	18806

TABLE IX: CLASSIFICATION REPORT OF TESTING DATA USING K-NEAREST NEIGHBORS MODEL

	Precision	Recall	F1-Score	Support
Loser	0.72	0.84	0.78	5812
Winner	0.31	0.18	0.23	2308
Average / Total	0.60	0.65	0.62	8120

C. Support Vector Machine

The advantage of this model is accurate in high dimensions space. We could reference this model to cause approaching overfitting.

The classification report of training data and testing data by using Support Vector Machine as shown in Table X and Table XI.

TABLE X: CLASSIFICATION REPORT OF TRAINING DATA USING SUPPORT VECTOR MACHINE

	Precision	Recall	F1-Score	Support
Loser	0.73	0.81	0.77	11952
Winner	0.59	0.47	0.52	6854
Average / Total	0.68	0.69	0.68	18806

TABLE XI: CLASSIFICATION REPORT OF TESTING DATA USING SUPPORT VECTOR MACHINE

	Precision	Recall	F1-Score	Support
Loser	0.76	0.64	0.70	5812
Winner	0.35	0.49	0.41	2308
Average / Total	0.64	0.60	0.61	8120

Table X and Table XI, Precision of the winner stock decrease 41 percent from 0.59 in training data to 0.35 in the testing data. And, the Recall of the winner stock increase 4 percent from 0.47 in the training data to 0.49 in the testing data. From these two parameters, it demonstrates that the accuracy of the model is definitely dropped, however, the ability of the model to reach winning stock is increasing which is unusual.

To clear this point, we required checking F1-Score. F1-Score of the winner stock in training data is 0.52 and it decreases to 0.41 in the testing data which is 21 percent decreasing.

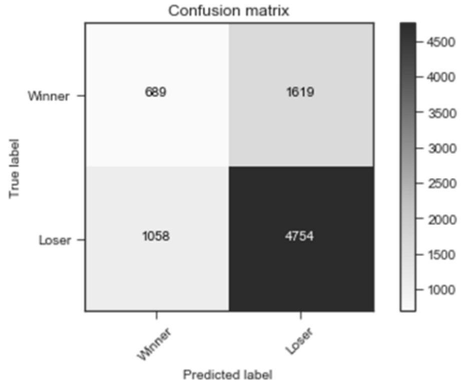


Fig. 16. The confusion matrix of support vector machine model.

The confusion matrix of Support Vector Machine model is shown in Fig. 16.

D. Summary

After using the testing data to test the models, a summary of average F1-Score and F1-Score of winning stock in each model as shown in Table XII.

TABLE XII: SUMMARY OF F1 SCORE OF EACH MODEL

Algorithm	Average	Winning Stock
Logistic Regression	0.66	0.34
K-Nearest Neighbors	0.62	0.23
Support Vector Machine	0.61	0.41

Form Table XII, the highest F1-Score of average is Logistic Regression, however, if the highest F1-Score of winning stock will be Support Vector Machine.

V. EVALUATION

If consider from F1-Score, Logistic Regression and Support Vector Machine seem to have predictive power more than the K-Nearest Neighbors model. However, the purpose of this study is building the model that could generate more than 12.5 percent of return and reduce the fluctuation of negative rate of return.

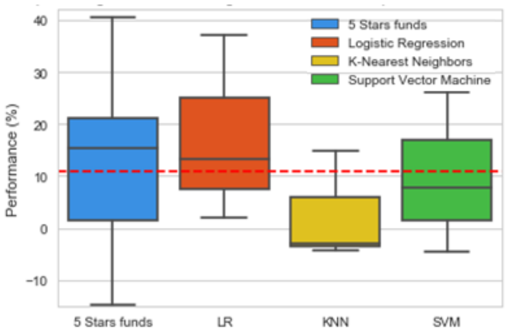


Fig. 1. The box plot of algorithm backtesting benchmark with the perform.

To consider this topic in the models, backtesting becomes an important section for considering that if we ran the model in the past from 2015 quarter 1 to 2018 quarter 3 what is the result of the models.

A. Backtesting Condition

1) Buy all stock that the algorithm predicts that it will be winning stock.
2) Buy all stock with total money in a portfolio.
3) Buy each stock with an equal amount of money.
4) Use only the time period in testing data (2015 quarter 1 to 2018 quarter 3).
5) Test in each yearly period separately.

After using all of the condition above, the backtesting result of each algorithm from the year 2015 quarter 1 to 2018 quarter 3 as shown as Table XIII and the box plot of algorithm backtesting benchmark with the performance of 5 stars funds as shown in Fig. 17.

From Fig. 17, all of the algorithms achieve to reduce fluctuation of negative rate of return below 0 percent, but from Table XIII, only the Logistic Regression model could generate positive rate of return in every years and it could achieve the 12.5 percent rate of return condition with 14.02% rate of return.

TABLE XIII: THE BACKTESTING RESULT OF EACH ALGORITHM FROM THE YEAR 2015 QUARTER 1 TO 2018 QUARTER 3

Year	Logistic Regression	K-Nearest Neighbors	Support Vector Machine
2015	1.95%	-4.27%	-4.67%
2016	37.11%	14.95%	26.13%
2017	13.19%	-2.88%	7.63%
2018	3.82%	-2.90%	0.75%
Average	14.02%	1.23%	7.46%

At the end of the evaluation section, the researcher chose the Logistic Regression algorithms to be the winning algorithms in this study because of, highest average F1-Score, probability of prediction advantage, knowing coefficient of each factor for advantage to explain the impact of each factor and the most important reason is insensitivity of each measurement parameters between training data and testing data.

VI. CONCLUSION

Using Supervised Machine Learning Model in the Stock Exchange of Thailand (SET) is possible, in this study, all of the algorithms that the researcher use could generate a positive rate of return. Especially, the Logistic Regression algorithms that the researcher chose to be the winning algorithms in this study. It could generate a 14.02 percent rate of return which is equivalent to the rate of return of 5 stars funds. And could reduce the fluctuation of negative rate of return below 0 percent if compared with 5 stars funds in the case that the amount of money does not have an effect of investing in the Stock Exchange of Thailand (SET).

This paper used a classification model to predict the category of winning and loser stock in the Stock Exchange of Thailand (SET) by using financial variables from the financial statement. The key advantage of this paper is a benefit of an investor to study and apply their research or be a piece of information for investors to make a decision that the traditional knowledge about the impact of each financial factor in the Stock Exchange of Thailand (SET) is reasonable enough to risk their money in the stock market.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Chayanant Kosol and Punnamee Sachakamol designed the research framework; Chayanant Kosol collected the data, analyzed the data and wrote the paper; Punnamee Sachakamol had approved the final version.

REFERENCES

[1] Numbeo. (2009-2019). Historical Prices of Restaurants in Bangkok. [Online]. Available: https://travel.usnews.com/gallery/the-worlds-30-best-places-to-visit-in-2017-18?slide=13

[2] Trading Economics. *Deposit Interest Rate in Thailand*. [Online]. Available: https://tradingeconomics.com/thailand/deposit-interest-rate

[3] A. Kumar and J. Babcock, *Python: Advanced Predictive Analytics*, Packt Publishing, pp. 51-52, 2017.

[4] Morningstarthailand. List of fund ranking by Morningstar. [Online]. Available: http://tools.morningstarthailand.com/th/fundquickrank/default.aspx?Site=th&LanguageId=th-TH

[5] C. D. Kirkpatrick *et al*., *Technical Analysis*, second ed, FT Press, ch. 2, pp. 11-13, 2009.

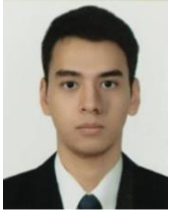
[6] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, 1962.

[7] G. Hackeling, *Mastering Machine Learning with Scikit-Learn*, second ed, Packt Publishing, 2017, pp. 91-92.

[8] T. W. Miller, "Web and network data science: Modeling techniques in predictive analytics," *Pearson Education*, pp. 247-248, 2015.

[9] R. Kumar, "Machine learning quick reference," *Packt Publishing*, pp. 31-39, 2019.

[10] Financial Factor. Popular Financial Factor in Thailand. [Online] Available: https://www.set.or.th/set/companyhighlight.do?symbol=PTT&ssoPageId=5&language=en&country=US



Chayanant Kosol is currently a postgraduate student at the Engineering Management (IGP) from Kasetsart University. He received his B.E. degree in industrial engineering from Kasetsart University, Thailand in 2016. His research interests include computer and information engineering and machine learning.

Punnamee Sachakamol is currently a lecturer at the Industrial Engineering Department, Kasetsart University, Thailand. He received his B.E. degree in industrial engineering from Thammasat University, Thailand. He received the B.S., M.S. and Ph.D. degree in industrial systems engineering from University of Regina, Canada. His research interests include financial analysis, business evaluation and manufacturing process.