VIETNAM NATIONAL UNIVERSITY – HO CHI MINH CITY
UNIVERSITY OF ECONOMICS AND LAW
FACULTY OF FINANCE AND BANKING

FINAL REPORT

SUBJECT: MACHINE LEARNING

APPLICATION OF MACHINE LEARNING IN DIVIDEND PAYOUT POLICY: THE CASE OF NON-FINANCIAL FIRMS IN VIETNAM

Lecturer: Nguyen Anh Phong – Phan Huy Tam

Student: Do Thi Lan Phuong

MSSV: K194141741

Ho Chi Minh City, 2022, June

Table of Contents

I. Introduction

Dividends are the after-tax profits of a company paying to current shareholders (Allen & Michaely, 1995). Dividend policy is used to decide whether the profits earned by company will be distributed to shareholders as dividends or will be retained as retained earnings to finance investments in the future. In addition, the high or low dividend payment will affect the current real income and the potential for future earnings growth of shareholders. Dividend policy decision is considered one of the three most important decisions in corporate financial management (besides investing and financing decisions) (Won et al., 2012). Dividend policy has great impacts on the operation, existence and development of enterprises in many aspects. The company can pay dividends to shareholders in three forms: cash dividends, stock dividends and asset dividends. Of the three types above, the cash dividend payment is the most interested by researchers because it is the only form that shareholders can get cash and immediately invest in other companies (Denis et al. Osobov, 2008; Kania and Bacon, 2005; Kazmiers-Jozwiak, 2015). Therefore, this study will focus on the cash dividend payment policy of enterprises.

Various theories have been developed to explain the behavior of corporate dividend policies. These theories include: A bird in the hand (Gordon, 1959), (Lintner J., 1956); Agency costs (Jensen & Meckling, 1976); Signaling (Bhattacharya, 1979); Behavioral explanation (H. M. Shefrin & Thaler, 1988), (H. Shefrin & Statmant, 1985); Life Cycle (Grullon, Michaely, & Swaminathan, 2002) and Taxation (Miller & Scholes, 1978). Besides, there are also many studies conducted to analyze the factors affecting dividend payment policy. These factors include profitability, liquidity, efficiency, leverage, lifecycle, ownership and size of the company (Lestari, 2018); (Franc-Dąbrowska, Mądra-Sawicka and Ulrichs, 2020). In Vietnam, the research of Le Thao Vy (2010) also shows that earnings per share (EPS) and profitability (ROE) are positively correlated with dividend policy of enterprises. The researches all say that the source of cash to pay dividends depends on the income and the level of profitability that the business achieves in the year. The need to be able to predict whether a company will pay a dividend is essential for investors to determine their investment decisions. Dividend-paying companies are considered companies with good future prospects. Through the process of research and review, the studies in Vietnam are stopping at traditional statistical and regression methods. Therefore, this study will use Machine Learning algorithms to estimate whether the company will pay dividends or not. Because of the application of algorithms to classification problems, so some models can be used such as: Logistic Regression, Decision Tree, Random Forest, SVM, KNN... Comparing and linking with previous studies, this article will use 3 algorithms: Logistic Regression, KNN and XGBoost with the expectation of achieving a level of efficiency, accuracy and can help investors have more reference sources to make investment decisions.

II. Theoretical framework

1. Theory

1.1 Dividend payout policy

According to the theory of Miller & Modigliani (M&M), dividend policy had absolutely no effect on the value of the firm, under the condition of assuming perfect capital markets. Changes in dividend payments were a signal to investors of management's assessment of a company's future earnings and cash flows. M&M relied on the customer effect argument to support its conclusion. If a company changed its dividend policy, it would lose some shareholders who choose to invest in other companies with more attractive dividends. As a result, share prices fell temporarily. However, other investors who prefered the new dividend policy would think the company's shares were underpriced and buy more shares. These transactions were assumed to be instantaneous and cost-free, resulting in the value of the shares remaining unchanged.

In contrast to Miller & Modigliani (1961), Gordon (1963) and Lintner (1956) argued that when the perfect market assumptions of the M&M model did not exist, dividend policy become more important and had a significant impact on firm value. With the Bird-in-the-hand policy theory, Gordon (1963) argued that there existed a relationship between firm value and dividend policy because:

- Investors' risk aversion: According to Gordon (1963), dividends reduced shareholder uncertainty, allowing a lower rate of future profits to be discounted, thereby increasing the value of the firm and vice versa.
- Taxes: In a taxed environment, higher dividend payments will reduce the value of the business.
- Transaction costs: In fact, the existence of transaction costs makes investors care about whether they receive cash dividends or capital gains.
- Issuance costs: The presence of issuance costs when selling new shares also tends to make companies prefer to retain profits.
- Agency costs: Dividend payments reduce the amount of retained earnings available for reinvestment and require the use of more external equity capital to finance growth.

1.2 Machine Learning algorithm

1.2.1 Logistic Regression

Logistic Regression model predicts a dependent variable by analyzing the relationship between one or more existing independent variables and is used mainly in classification problems. Logistic Regression is a statistical analysis method to predict the probability of occurrence for an observation by adjusting the data using a Logistic curve. From there, the model gives a binary result based on the probability of each class. With a threshold of 0.5 (default), cases with a higher probability than this threshold will be classified as 1, while cases lower than this will be classified as 0. For example, a regression Logistics can be used to predict whether a high school student will be admitted to a particular university. These binary outcomes allow for simple decisions between two alternatives. Here is the Logistic regression equation:

$$\ln\left(\frac{p}{p-1}\right) = b_0 + b_1x$$

From this equation, the probability of the output is calculated by the formula:

$$p = \frac{1}{1+e^{-y}}$$

In there:

p is the probability that it occurs from 0 to 1.
b_0 is the intercept coefficient
b_1 is the slope of the input variable x
x is the independent variable

1.2.2 KNN

KNN (K-Nearest Neighbors) is a simple algorithm in the group of supervised learning algorithms. KNN algorithm assumes that similar data will exist close to each other in a space, from which its task is to find the output of a new data based on the output of the K nearest points around it. The class (label) of a new data object can be predicted from the classes (labels) of its K nearest neighbors. The algorithm of KNN can be described as follows:

- Determine the parameter K the number of nearest neighbors
- Calculate the distance of the object to be classified to all the objects in the training set
- Get top K for the smallest (or largest) value
- In the top K values just taken, we statistics the number of each class, choose the classifier for the largest number

One question that it is better to choose K as large as possible. That answer depends on how the data is. Not always the larger K gives good results and vice versa. The choice of parameter K of the model will proceed through many experiments to choose the best results.

1.2.3 XG Boost

XGBoost (Extreme gradient boosting algorithm) is a general Tree Boosting algorithm, which is an improvement on Gradient Boosting Decision Tree (GBDT) (Friedman 2001), proposed by Chen and Guestrin in 2016. XGBoost is made up of several decision trees, each of which is a weak learner, and Boosting technology promotes it to a strong learner. In the classification problem, although the global prediction accuracy of each weak classifier is not high, it may have a very high prediction accuracy in some aspect of data. XGBoost training, for the production of each tree method, through calculating node split and classification first whether to produce “gain” to determine whether the node is divided, and through parameter controls the depth of the tree, when a tree is generated after pruning were needed to prevent a fitting, the first m round of the generated tree will learn the real value and m-1 wheel model forecasts “residual”, So that the model prediction results gradually close to the real value.

2. Literature review

Currently, there are many studies examining the factors affecting the dividend policy of the company. Amidu (2007) conducted a study on a sample of companies listed on the Ghana Stock Exchange over a period of 8 years. Regression analysis is used as a statistical method to determine the positive relationship between dividend policy, ROA and sales growth. Husam & Al-Malkawi (2007) conducted a study on the dividend payment behavior of firms in Jordan and noted the positive relationship between firm size and dividend policy. Enterprises whose capital is financed mainly by debt are those that are under pressure to pay interest and principal. The empirical research results of Asif et al. (2011) have shown the relationship between financial leverage and dividend policy of listed Pakistani companies in the period 2002-2008. Liquidity is also a factor positively correlated with dividend payout. Alli, Kahn and Ramirez (1993) analyzed a sample of 105 companies listed on the New York Stock Exchange and found that firms with high liquidity had lower systemic risk, and they also reported signal to investors that they can pay high dividends. Several other studies analyze both dividend trends and dividend payout ratios (Floyd, Li and Skinner 2015; Brockman and Unlu 2009; Arko et al. 2014). These studies often use the same set of explanatory variables for both dividend propensity and dividend payout regression. Arko et al. (2014) consider profitability, cash flow, investment opportunities, taxes, institutional ownership, firm size, leverage, and business risk as independent variables in their model. Brockman and Unlu (2009) include retained earnings ratio (RE), equity ratio (TE), profitability (ROA), revenue growth (SGR), market capitalization (LOGSIZE) and cash holdings ratio (CASH HOLDING) in studying both dividend trends and dividend payout ratios. Inheriting from previous studies, the article will use 11 variables to explain the model including: ROA, Cash ratio, Current ratio, Capex, Leverage, Firm size, D/E, RETA, FCF per share, Profit margin and DSO.

In recent years, machine learning algorithms have been widely applied in the financial sector. Bae (2010) tested the prediction of dividend policy decisions of Korean companies using Support vector machine (SVM), Decision tree and ANN. The results showed that SVM performs better than other techniques in forecasting dividend policy. According to research by Triasesiarta Nur (2019), Logistic Regression and NN models were used to forecast whether to pay or not to pay dividends. The results showed that the NN model had higher accuracy than Logistic Regression. Meanwhile, the research results of Branko SORIĆ & Toni ŠUŠAK (2015) on the dividend policy of companies in Croatia had an accurate result of 90.5% for the Logistic Regression model. The study of Şirin ÖZLEM & Ömer Faruk TAN (2021) used two models XGBoost and MLNN to evaluate dividend policy. In particular, the XGBoost model had higher accuracy than the MLNN model with 82%. Precision and Recall of the two layers 0-1 of the XGBoost model also outperformed the MLNN. Besides, KNN is also considered a simple and easy-to-use algorithm. The model also predicts well with small and low-dimensional data sets. KNN methods have also been used successfully in distribution-based problems such as outliers detection (Divya & Kumaran, 2016; Hautamaki et al., 2004). Therefore, the study will use 3 models Logistic Regression, KNN and XGBoost to evaluate dividend policy and compare the performance of the three models.

III. Data and Research method

1. Data and variables

The dataset used in this study includes 594 companies listed on the Ho Chi Minh City Stock Exchange (HOSE) and the Hanoi Stock Exchange (HNX). In which, HOSE has 299 companies and HNX has 295 companies. The data includes non-financial companies (excluding companies in the financial, banking, securities, insurance, real estate and investment funds sectors). Data samples were taken in 2021 and collected from Thomson reuters.

In which, data calculated for building model are ROA, Cash holding, Current ratio, Capital Expenditure (Capex), Leverage, Debt to Equity ratio (D/E), Firm size, Retained earnings to total assets ratio (RETA), FCF per share, Profit margin and DSO. With dependent variable, cash dividend is as the proxy for dividend payout policy. Structure of dependent variable between two classes is 2:1.

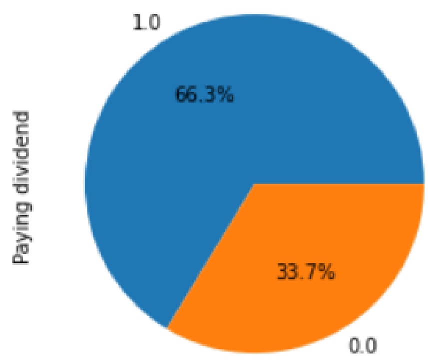

Figure 1: The ratio between the two classes of the dependent variable

Variable	Definition
Cash dividend	= 0 if not pay cash dividend = 1 if pay cash dividend
ROA	Net income/Total asset
Cash ratio	Cash/Total current liabilities
Current ratio	Total current assets/Total current liabilities
Capital Expenditure	Fixed assets/Total assets
Leverage	Total debt/Total assets
D/E	Total debt/Total equity
Firm size	Log(Total assets)
RETA	Retained earnings/Total assets
FCF per share	FCF/Total common shares outstanding
Profit margin	Net income/Total revenue
DSO	Account receivable/Total revenue

Table 1: Attribute variables collected and calculated in dataset

2. Model building

This study will test the predictive accuracy of the factors affecting the company's cash dividend policy whether to pay or not to pay. Three models are used in this paper: Logistic Regression, KNN and XGBoost. Inputs of the model are variables related to financial indicators including liquidity, efficiency, profitability, leverage ratio. In addition to these ratios, factors related to firm size, free cash flow, and retained earnings are also included. The output of the model is a categorical variable that pays cash dividends (1) – no cash dividends (0). To get the best prediction results, the model is built according to the following steps:

Step 1: Split the dataset

The data will be divided into two parts, the part of data for training and the other part of data for testing after training with the ratio of 80:20. The inputs and output of the two sets have the same variables.

Step 2: Standardize the data

Before being included in the model, the data will be normalized using a standardizing technique called Standard Scaler.

Step 3: Train the model

- For Logistic Regression model: The model will train data with a default threshold of 0.5. Then, test the model for the test dataset and run the performance evaluation results. Next, adjust the model with the best threshold. Finally, test the dataset again with the best adjusted threshold and run the performance evaluation results.
- For KNN model: The model will be trained with k=5 (default). Then, test the model for the dataset and run the performance evaluation results. Next, adjust the model to find the k with the highest accuracy. Finally, test the dataset with k just found and run the performance evaluation results.
- For the XGBoost model: The train set will be trained with the default model, without any adjustment parameters and test the datatest. Then adjust the model with the new learning_rate, max_depth, min_child_weight, n_estimators, and nthread metrics. Finally, test the dataset with the adjusted model to evaluate the performance.

Step 4: Evaluate model performance and compare results among models

The test results of the models will be saved for evaluation and comparison. Compare the accuracy of dividend-paying or non-dividend-paying company predictions from three models using three metrics: accuracy, precision and recall (Devi & Radhika, 2018).

IV. Results

1. Correlation analysis and descriptive statistic

When developing a dividend payout prediction model, it is important to check for multicollinearity and make sure that there is no high correlation among the independent variables. “Multicollinearity exists when the independent variables are highly correlated (r = 0.9 or more).” (Pallant, 2007)

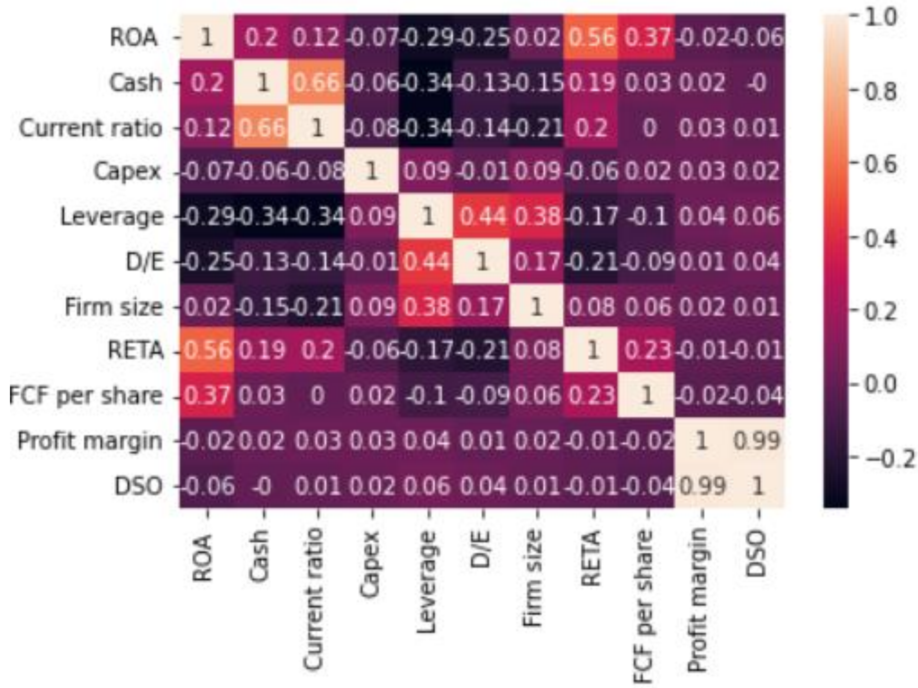

Figure 2: Heat map for the independent variables

In general, the independent variables have a small correlation. However, there are three pairs of variables with correlations greater than 0.5, including Cash and Current ratio, ROA and RETA, Profit margin and DSO. In which, two variables Profit margin and DSO have a correlation of 0.99. Therefore, the two above variables exist at multicollinearity, so they will be excluded from the model. The other variables are kept unchanged and are inputs of the model.

	ROA	Cash	Current ratio	Capex	Leverage	D/E	Firm size	RETA	FCF per share	Profit margin	DSO	Paying dividend
ROA	1.000000	0.197762	0.123807	-0.065471	-0.293180	-0.249402	0.021976	0.558372	0.370342	-0.015664	-0.057618	0.297210
Cash	0.197762	1.000000	0.658139	-0.062329	-0.335517	-0.131722	-0.145915	0.190160	0.029757	0.019874	-0.003427	0.036011
Current ratio	0.123807	0.658139	1.000000	-0.082712	-0.343159	-0.135777	-0.207885	0.195524	0.004367	0.026085	0.012613	-0.046080
Capex	-0.065471	-0.062329	-0.082712	1.000000	0.089985	-0.011908	0.090547	-0.055552	0.023449	0.031299	0.024022	-0.030124
Leverage	-0.293180	-0.335517	-0.343159	0.089985	1.000000	0.436888	0.377045	-0.174079	-0.098800	0.044050	0.057458	-0.120291
D/E	-0.249402	-0.131722	-0.135777	-0.011908	0.436888	1.000000	0.165267	-0.212761	-0.087014	0.010258	0.039851	-0.114232
Firm size	0.021976	-0.145915	-0.207885	0.090547	0.377045	0.165267	1.000000	0.081426	0.057686	0.017551	0.006484	0.095639
RETA	0.558372	0.190160	0.195524	-0.055552	-0.174079	-0.212761	0.081426	1.000000	0.229278	-0.007593	-0.010289	0.327327
FCF per share	0.370342	0.029757	0.004367	0.023449	-0.098800	-0.087014	0.057686	0.229278	1.000000	-0.020360	-0.037697	0.066278
Profit margin	-0.015664	0.019874	0.026085	0.031299	0.044050	0.010258	0.017551	-0.007593	-0.020360	1.000000	0.988927	0.060262
DSO	-0.057618	-0.003427	0.012613	0.024022	0.057458	0.039851	0.006484	-0.010289	-0.037697	0.988927	1.000000	0.051301
Paying dividend	0.297210	0.036011	-0.046080	-0.030124	-0.120291	-0.114232	0.095639	0.327327	0.066278	0.060262	0.051301	1.000000

Figure 3: Correlation between independent variables and dependent variable

Looking at the above table, it can be seen that the dependent variable is negatively correlated with Capex, Leverage and D/E. The remaining variables are positively correlated with the dependent variable.

	ROA	Cash	Current ratio	Capex	Leverage	D/E	Firm size	RETA	FCF per share
count	594.000000	594.000000	594.000000	594.000000	594.000000	594.000000	594.000000	594.000000	594.000000
mean	0.055573	0.904468	2.573264	0.261925	0.215318	0.871100	27.557860	0.069279	1160.129524
std	0.074659	2.042584	3.978887	0.330885	0.184415	2.255323	1.618848	0.145943	3229.121111
min	-0.255507	-0.007250	0.033829	0.000000	0.000000	0.000000	23.636331	-1.075446	-17316.031192
25%	0.014641	0.090711	1.146372	0.061577	0.034995	0.057592	26.418246	0.025094	-51.265394
50%	0.039981	0.299411	1.453585	0.178920	0.185025	0.371069	27.513169	0.064008	834.644353
75%	0.081675	0.748361	2.388351	0.366220	0.349911	1.110378	28.499764	0.132464	2234.599531
max	0.449104	21.287109	39.059628	4.919161	0.954313	40.341193	32.814132	0.722730	27174.541064

Figure 4: Descriptive statistics of the independent variables

The table provides descriptive statistics about the independent variables for the machine learning model including: ROA, Cash, Current ratio, Capex, Leverage, D/E, Firm size, RETA and FCF per share. The dataset has a total of 594 observations. ROA values ranged from -0.255 to 0.449 with an average of 0.055. The Cash variable ranges from 2,042 to 21,287. Meanwhile, Current ratio is distributed from 0.033 to 39,059 but the average is at 2,573, showing that the liquidity is still quite low. Capex ranges from 0 to 4,919 and the average is 0.261 indicating that capital expenditure is not high. Similar to Capex, Leverage and RETA show averages below 0.2. The average D/E ratio of 0.871 is quite small, showing that the business is less dependent on debt financing. Firm size of companies allocated from 27,557 to 32,814. FCF per share ranges from -17,316 to 27,174. In which, ¼ of the number of companies with FCF per share is -51.

2. Model results

The results of the three models before adjusting are shown in performance evaluation tables.

	Logistic Regression		KNN		XGBoost	
Binary State	0	1	0	1	0	1
Precision	0.88	0.68	0.74	0.72	0.76	0.73
Recall	0.17	0.99	0.33	0.94	0.38	0.94
F1-Score	0.28	0.81	0.46	0.81	0.51	0.82
Accuracy	69.75%		72.27%		73.95%	

Table 2: Performance metrics for three algorithms

The performance metrics of three algorithms to correctly predict the class of Paying dividend are summarized in Table 2. For class 0 (not paying dividend), the precision of Logistic Regression model is the highest at 88%. This means that the number of observations that the model predicts correctly for non-dividend cases reaches 88%.

Meanwhile, the KNN and XGBoost algorithms have lower precision than Logistic Regression by 74% and 76%, respectively. However, the recall of Logistic Regression model for class 0 is very low, only 17%. It means that omission rate of observations of class 0 is very high. Meanwhile, recall of the other two models is 33% and 38% higher than Logistic Regression. Thus, it can be seen that there is a trade-off between the two indexes of precision and recall at class 0 in all three models. For class 1 (paying dividend), the precision of the three models is 68%, 72% and 73%, respectively. This high index shows a high ratio of correctly predicting dividend cases to the total number of cases actually paying dividends. Different from class 0, class 1 has high recall in all 3 models at 99%, 94% and 94%, respectively.

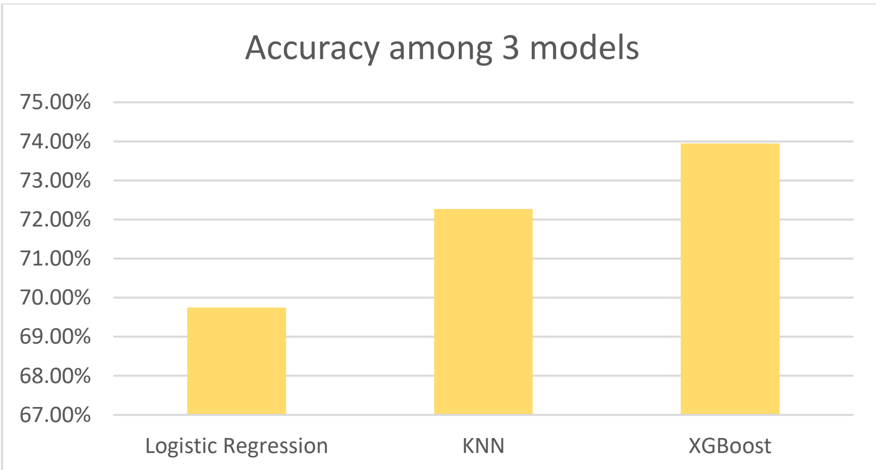

Figure 5: Accuracy of three models before adjusting

Comparing the three models when running with default parameters, XGBoost model gives the highest accuracy with 73.95%. Next is KNN with 72.27% and Logistic Regression with 69.75%. Thus, if compared in terms of accuracy, XGBoost model proved to be more effective than the other two models.

After building the model with default parameters, three models will be set up with adjustments to give more consistent and accurate results. For Logistic Regression, the model will show a Roc curve plot to find the best threshold (this dataset has the best threshold of 0.673993). With the KNN model, after plotting the graph showing the level of accuracy and corresponding k, I found k for the highest accuracy of 5, so no further adjustment was made. Finally, XGBoost will be adjusted by selecting the most appropriate learning rate and min_child_weight.

	Logistic Regression		KNN		XGBoost	
Binary State	0	1	0	1	0	1
Precision	0.6	0.83	0.74	0.72	0.78	0.72
Recall	0.71	0.74	0.33	0.94	0.33	0.95
F1-Score	0.65	0.78	0.46	0.81	0.47	0.82
Accuracy	73.11%		72.27%		73.11%	

Table 3: Performance metrics for three algorithms after adjusting

.

After adjusting, Logistic Regression model gives much better results than the model with default parameters. Although precision decreased to 60%, recall increased to 71% and F1-score increased to 0.65 in class 0. For class 1, Logistic Regression has a decrease in recall but a significant increase in precision indicating a trade-off between two indexes. The KNN model does not change because k remains the same. Meanwhile, the XGBoost model does not have too much change in both classes and two indices of precision and recall. Therefore, if comparing precision and recall, Logistic Regression model shows more efficiency than the other two models.

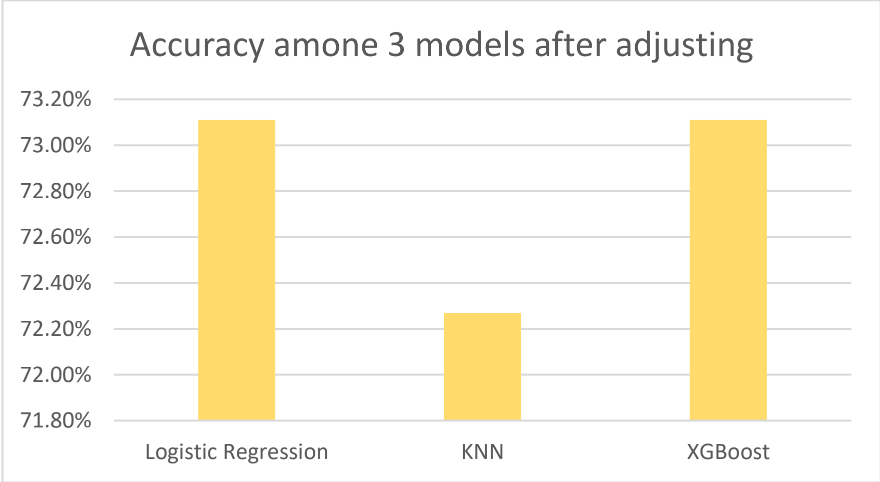

Figure 6: Accuracy of three models after adjusting

Logistic Regression shows a significant effect after adjusting the model when building with the best threshold. Meanwhile, the XGBoost model has not changed much compared to before adjusting. The accuracy of Logistic Regression model has increased from 69.75% to 73.11%. XGBoost's accuracy again reduced from 73.95% to 73.11%.

V. Conclusion

This study tries to predict the dividend payment policy of non-financial companies on HOSE and HNX in 2021 based on 11 selected financial factors such as ROA, cash, current ratio, capex... After reviewing and inheriting the quintessence of previous researches, three machine learning algorithms are used to predict the classification of a company's dividend payment policy including Logistic Regression, KNN and XGBoost. Both XGBoost and Logistic Regression show high performance (average accuracy of 73.11% after adjusting). But in terms of precision and recall, the performance of Logistic Regression is considered the best of three models. Thus, the application of a machine learning model to financial issues in general and dividend payment policy in particular has partly brought accuracy, speed and flexibility to both the company's management and the investors. However, the study still has certain limitations, such as the research period is too short (1 year) or the independent variables do not explain much for the dependent variable.

REFERENCES

1. Kosala, R. (2017). Predicting the likelihood of dividend payment from Indonesian public companies with data mining methods. International Journal of Business Information Systems, 26(2), 139-150.
2. Elyasiani, E., Jia, J., & Movaghari, H. (2019). Determinants of dividend payout and dividend propensity in an emerging market, Iran: an application of the LASSO. Applied Economics, 51(42), 4576-4596.
3. Nguyen, T. T. N., & Bui, P. K. (2019). Dividend policy and earnings quality in Vietnam. Journal of Asian Business and Economic Studies.
4. Ngọc, Đ. B., & Cường, N. C. (2014). Các nhân tố ảnh hưởng đến chính sách cổ tức của các doanh nghiệp niêm yết trên thị trường chứng khoán Việt Nam. Tạp chí Phát triển kinh tế, 290, 42-60.
5. Ozlem, S., & Tan, O. (2021, March). Predicting Dividend Payout Policy of Turkish Firms Using XGBOOST and MLNN Algorithms. In 8th. INTERNATIONAL MANAGEMENT INFORMATION SYSTEMS CONFERENCE.
6. Sorić, B., & Šušak, T. (2015). Development of Dividend Payout Model Using Logistic Regression: The Case of Croatian Non-Financial Companies. Economy Transdisciplinarity Cognition, 18(1), 117-123.
7. Nur, T. (2019). Comparing the Accuracy of Multiple Discriminant Analyisis, Logistic Regression, and Neural Network to estimate pay and not to pay Dividend. INOBIS: Jurnal Inovasi Bisnis dan Manajemen Indonesia, 3(1), 97-105.
8. Laoh, L. C. (2019). Dividend payout forecast: multiple linear regression vs genetic algorithm-neural network. CogITo Smart Journal, 5(2), 252-265.