International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 6 Issue 9, September - 2022, Pages: 131-134

The Performance of Ridge Regression, LASSO and Elastic Net in Modeling Market Value Data*

Nur Izzati Nabeela Adenan[1], Mazni Mohamad[2*], Adzhar Rambli[3]
[1]Department of Household Income and Basic Amenities (BPHPP), Federal Government Administrative Centre Putrajaya, Malaysia
[2,3]Centre of Statistics and Decision Science, Faculty of Computer and Mathematical Sciences, Shah Alam, Selangor, Malaysia
[2*]mazni825@uitm.edu.my

Abstract: *Modelling market value is important in determining contributing factors and ensuring the well-being of a business or a company. Many studies employed ordinary least squares (OLS) regression in modelling market value. OLS regression is no longer appropriate when data is high dimensionality. An alternative to OLS regression, known as penalized regression, is utilized in this study since high-dimensionality market value data is to be analyzed. Ridge regression, LASSO, and elastic-net are the three techniques of penalized regression employed in this research. Their performance is compared by means of root mean square error (RMSE) and coefficient of determination,* R^2. *Elastic net outperforms other techniques as it has the smallest RMSE and largest* R^2.

Keywords—ridge regression; LASSO; elastic-net; high-dimensional; market value

1. Introduction

Market value is the current price of an asset, market-traded security, or company. In the context of companies, market value is equivalent to market capitalization. It is computed based on the current market price of the company's shares [1]. Market value is also commonly referred to as the market capitalization of a publicly traded company and is calculated by multiplying the total number of its outstanding shares by the current market price [2]. It allows investors to understand the relative size of one company to the others. Market value measures what a company is worth on the open market, as well as the market's perception of its prospects as it reflects what investors are willing to pay for its stock. Many approaches have been utilized to examine market capitalization, including cash-flow-based analysis [3]. The market value is defined by the ratings or multiples agreed upon by corporate investors, such as the selling price, profit price, and company value to Earnings Before Interest, Taxes, Depreciation, and Amortization (EBITDA). For example, the greater the valuations, the higher the market value [1]. As mentioned in [4], market value also tells a company's current value. If there are more buyers than sellers, the share price will increase. If there are more vendors than buyers, the price will fall.

Understanding and modeling market value is important in ensuring the good health of a company's financial well-being. Researchers in this area, [5], [6], and [7] generally utilized Multiple Linear Regression (MLR) in modeling market values. Factors that contribute to the behavior of market values can be easily identified through MLR analysis. In line with the big data era, MLR which is based on ordinary least squares (OLS) could no longer be an appropriate tool when a dataset contains more variables, x than observations, n. A dataset with $x > n$ is known as high-dimensional data. High dimensionality poses numerous challenges to statistical theory, methods, and implementations [11]. For example, in a linear regression model with error variance σ^2, when the dimensionality x is comparable to or exceeds sample size n, the OLS estimator is not well behaved or even no longer unique due to the (near) singularity of the design matrix. Hence, this study aims to analyze high-dimensional market value data using a better alternative to OLS which is known as penalized regression. Ridge, Least Absolute Shrinkage and Selection Operator (LASSO), and Elastic net are the penalized regression techniques to be applied in this research. Their performance will then be compared.

2. METHODOLOGY

2.1 Data

In this study, the high-dimensional market value dataset of Tenaga Nasional Berhad (TNB) is to be utilized. TNB is the major electricity utility in Malaysia. The dataset which contains 33 variables, x with only 22 observations, n was obtained from the Thomson Reuters Datastream. It is a monthly dataset from January 2020 until October 2021.

2.2 Penalized Regression

According to [8] and [9] as cited in [10], penalized regression introduces a constraint in the equation model for having too many variables in the model. A coefficient that is close to zero or equal to zero will be assigned to the less contribute variables in the model. Thus, it allows the development of a linear regression model that is penalized.ck price and

Ridge Regression

Ridge regression (RR) modifies the OLS method to allow biased [12]. For OLS, the parameter estimates are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{1}$$

where $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$ $\boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & \dots X_{1p} \\ \vdots & \vdots & \dots \vdots \\ 1 & X_{n1} & \dots X_{np} \end{bmatrix}$ and $\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$ such that

$$\widehat{\boldsymbol{\beta}} = \text{argmin } RSS(\beta) \quad (2)$$

where $RSS(\beta)$ is the sum of squared residuals and is defined as follows:

$$RSS(\beta) = \left[Y_i - \left(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}\right)\right]^2 \quad (3)$$

Like OLS, Ridge Regression coefficients are estimated by minimizing:

$$R(\beta, \lambda) = RSS(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \quad (4)$$

where λ is known as a tuning parameter and $\lambda \geq 0$.

A penalty term *L2*-norm, which is the sum of the squared coefficients is used to shrink the coefficients. Selecting a good tuning parameter λ is crucial as it determines the impact on (5). For example, when $\lambda = 0$, RR estimates become OLS estimates, but when λ gets larger, the term $\lambda \sum_{j=1}^{p} \beta_j^2$ shrinks, hence the coefficient estimates of RR, $\hat{\beta}^R(\lambda)$ approach zero.

LASSO Regression

LASSO shrinks the regression coefficients near zero by penalizing the regression model with a penalty term *L1*-norm, which is the sum of the absolute coefficients. LASSO regression coefficients are estimated as follows:

$$L(\beta, \lambda) = RSS(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \quad (5)$$

The LASSO estimator, $\hat{\beta}^L(\lambda)$ minimizes (5) over β for a given λ. The tuning parameter λ determines whether $\hat{\beta}^L(\lambda)$ is sparse or not (setting some coefficients exactly to zero). When λ gets larger, $\hat{\beta}^L(\lambda)$ gets smaller, which results in a sparse solution.

Elastic Net

Elastic Net is a hybrid, regularisation, and variable selection technique that is a mixture of LASSO regression and Ridge regression. Thus, Elastic Net estimator is defined as:

$$E(\beta, \lambda) = RSS(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (6)$$

2.3 K-fold Cross Validation Procedure for Tuning Parameter

This study used the *glmnet* package in R software to perform Ridge, LASSO, and Elastic Net regression. *glmnet* package fits a generalized linear model via penalized maximum likelihood, and the algorithm in this package uses the cyclical coordinate descent method to find the optimal solution.

To find the best value for the shrinkage or tuning parameter, a cross-validation process is necessary. [13] describes in detail how cross-validation works in determining tuning parameter λ.

2.4 Performance Measures

This study employed two types of performance measures which are root mean square error (RMSE) and coefficient of determination, R^2 to compare the performance of each regression technique.

Root Mean Square Error (RMSE)

This measure is frequently used to compare the performance of different regression techniques. It is simple and easy to compute, and when employed outside the sample, it typically meets the criteria within sample. RMSE is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-p}} \quad (7)$$

where $\hat{Y}_i$ represents predicted Y_i. Small value of RMSE indicates a better model estimator.

Coefficient of Determination ($\boldsymbol{R^2}$)

R^2 is used as a measure for goodness of fit. Large R^2 implies a good model. R^2 is expressed as follows:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \quad (8)$$

with $\bar{Y}$ indicates the average of Y.

3. RESULTS AND ANALYSIS

The TNB market value dataset which contains 33 variables with 22 observations used in this study have gone through cleaning data processes prior to any type of analysis. Missing values in the dataset are well taken care of by employing the median imputation technique.

3.1 Cross-validation Plot

A cross-validation plot is useful to depict the mean squared error (MSE) against log λ. The red dots in Fig. 1, Fig. 2, and Fig. 3 represent the cross-validation error curve along with the lower and upper standard deviation curves. The plot contains list of numbers at the top which indicates the number of variables a particular model is using. The model with the lowest cross-validation MSE is shown by the left vertical dashed line (λ_{min}). The right vertical dashed line (λ_{1se}) indicates that the model is one standard deviation away from the lowest MSE. The mean square error of each model was calculated, and the λ values with the minimum mean square error was chosen to build the best model [15]. It can be observed from Fig. 1, Fig. 2, and Fig. 3 that the suggested non-zero regression coefficients for RR, LASSO and Elastic Net are 31, 4 and 13 respectively. It is obvious that for LASSO and Elastic Net, as the value of lambda, λ increases, the penalty became severe, and fewer predictors were chosen as more coefficients shrank near zero.

International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 6 Issue 9, September - 2022, Pages: 131-134

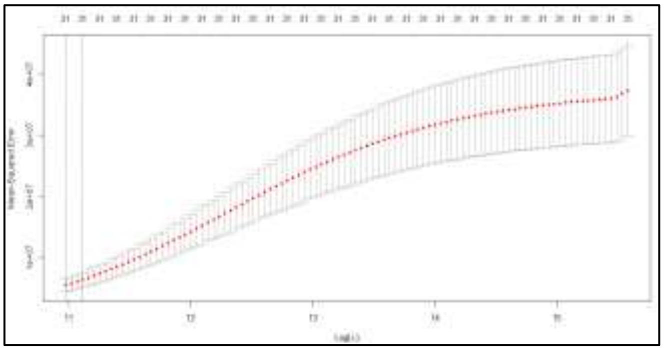

Fig. 1. *Cross-validation plot for RR*

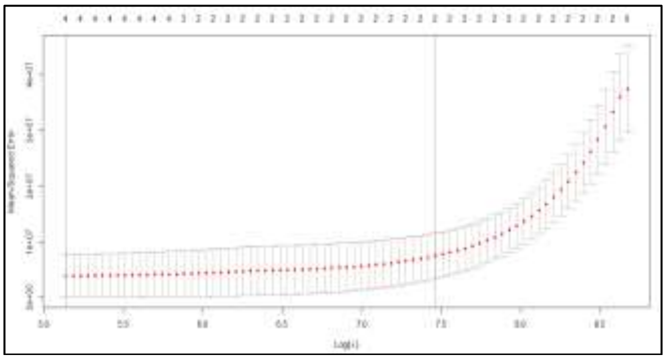

Fig. 2. *Cross-validation plot for LASSO*

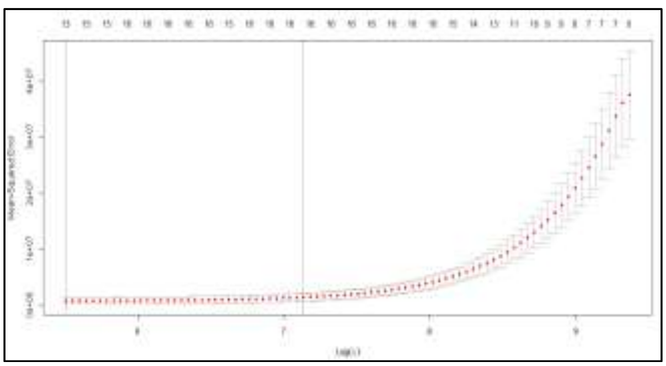

Fig. 3. *Cross-validation plot for Elastic Net*

3.2 Number of Variables Retained in Each Regression Technique

Penalized methods can bring down the model prediction error to zero by reducing the regression coefficients and lowering the estimates. Results obtained from RR analysis indicate that 31 variables have been selected from 33 covariate variables where 2 variables are excluded from the model. This finding is similar to the study done by [16] which stated that ridge regression tends to select all variables in the model. The blue line color displayed in Fig. 4 indicates the most influential variable towards the response variable. Hence, it becomes the first variable to enter the model, followed by variables with varying impacts. For the insignificant variables, their coefficient will become zero.

Fig. 5 shows the non-zero coefficients plot for TNB Market Value data using LASSO Regression. It shows there are only 4 variables are considered in LASSO model. The green line color in Fig. 4 indicates the most significant variable towards the response variable.

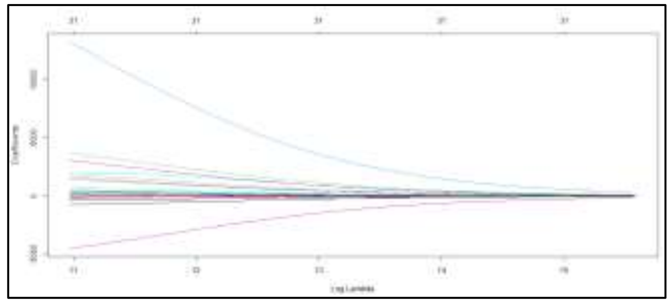

Fig. 4. *Non-zero Coefficients plot for RR*

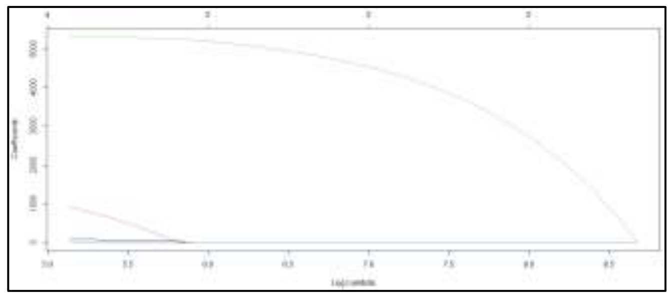

Fig. 5. *Cross-validation plot for LASSO*

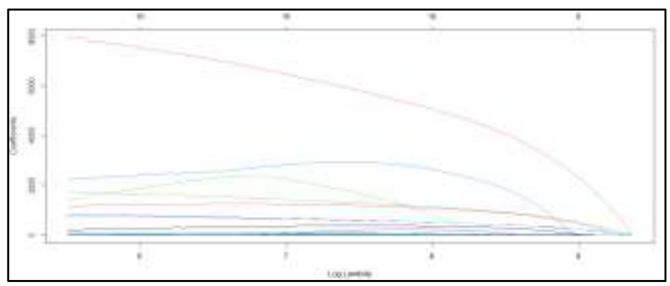

Fig. 6. *Cross-validation plot for LASSO*

As can be observed from Fig. 6, the pink line color variable is the most influential variable towards the response variable. It is the first variable to enter the model compared to other line color variables. For the ineffective variables in the model, their estimator coefficients will become zero. Same as LASSO, as the value of lambda, λ increases, the penalty became severe, and fewer predictors were chosen as their coefficients shrank to near zero and more of them were set to zero.

3.3 Model Performance

Table 1 shows the R^2 and RMSE values for each technique. LASSO produced the highest R^2 value followed by Elastic-net and Ridge regression. LASSO also has the smallest RMSE value compared to the others. This result indicates that LASSO outperforms Ridge and elastic-net in modelling the TNB market value data.

Table 1: RMSE and R^2 Values for Each Technique

Method	Performance Measure	
	RMSE	R^2
Ridge	2105. 559	0.8701656
LASSO	179.0563	0.9990611
Elastic-net	181.4869	0.9990354

4. CONCLUSION

Penalized regression techniques like ridge, LASSO, and elastic-net can be a good alternative to classical OLS regression when modelling high-dimensional data. In this study, the LASSO model with only four variables is found to be the best approach to modelling the TNB market value. The four variables are price trade, price to book value, price cash flow ratio, and unadjusted price. This finding shows that the most parsimonious model becomes the best model as it has the smallest errors and fits the data better than other models.

5. REFERENCES

[1] Seth, S. (2021, December 7). Book value vs. market value: What's the difference? Investopedia. Retrieved February 19, 2022, from https://www.investopedia.com/articles/investing/110613/market-value-versus-bookvalue.asp#:~:text=The%20market%20value%20represents%20the,price%20of%20the%20company's%20shares.

[2] Chen, J. (2021, December 30). Market value definition. Retrieved November 29, 2021, from https://www.investopedia.com/terms/m/marketvalue.asp

[3] Ko, K. (2009). Multiple Regression Model For Market Capitalization. Journal of Global Business Issues, 3(2).

[4] Pinsent, W. (2021, November 20). *Understanding stock prices and values*. Retrieved November 29, 2021, from https://www.investopedia.com/articles/stocks/08/stock-prices-fool.asp.

[5] Alshubiri, F. (2021). The stock market capitalization and financial growth nexus: an empirical study of western European countries. Future Business Journal, 7(1), 1-20.

[6] Al-Afeef, M. A. M. (2020). Factors affecting market capitalization: A practical study ase 1978-2019. International Journal of Scientific and Technology Research, 9(3), 7049-7053.

[7] Budiyono, & Santoso, S. B. (2020). Effect of EPS, ROE, PER, DPR, and Interest Rate on Stock Prices in the Jakarta Islamic Index Group from 2014 to 2017. 1229–1235. https://doi.org/10.5220/0009498812291235

[8] James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

[9] Bruce, Peter, and Andrew Bruce. 2017. *Practical Statistics for Data Scientists*. O'Reilly Media.

[10] Kassambara, A. (2017). Machine Learning Essentials: Practical Guide in R. STHDA. http://www.sthda.com

[11] Fan, J., Lv, J. and Qi, L. (2011) Sparse High Dimensional Models in Economics. Annual Review of Economics, 3, 291- 317.

[12] Kutner, M. H., Nachtsheim, C. J., and Neter, J. (2004). Applied Linear Regression Models. NY: McGraw-Hill.

[13] AlNasser, H. (2017). *On ridge regression and least absolute shrinkage and selection operator*, Research Master thesis, Department of Mathematics and Statistics, Victoria University.

[14] Lazim, M. A. (2011). Introductory Business Forecasting a practical approach. Shah Alam: UPENA Universiti Teknologi MARA.

[15] Hassan, M. Y., & Arman, H. (2021). Comparison of Six Machine-Learning Methods for Predicting the Tensile Strength (Brazilian) of Evaporitic Rocks. Applied Sciences, 11(11), 5207.

[16] Farhadi, Z. F. Z., Belaghi, R. A., & Alma, O. G. (2019). Analysis of penalized regression methods in a simple linear model on the High-Dimensional data. American Journal of Theoretical and Applied Statistics, 8(5), 185-192.