Predictive Modeling – R Code Documentation

1. PREPARE THE DATASET
   1. Import the dataset from 1/1/2010 to 31/12/2022

* Sp500
* JPMorgan Chase
* US 3-month Treasury Bill
* Economic Indicators
  1. Calculate market quarterly rate of return
* Quarterly\_return () function that loops through the dataframe to calculate quarterly return, in which current quarter return = ((current price - previous price)/previous price)\*100

Graphical user interface, text

Description automatically generated

* Calculate S&P500 quarterly return

Table

Description automatically generated

* Calculate JPMorgan Chase quarterly return

Table

Description automatically generated

* 1. Merge the quarterly return datasets into 1 table
* Merge sp500\_qr & jpm\_qr together

Graphical user interface, text, table

Description automatically generated

* 1. Find Beta by taking (sp500\_quarterlyReturn – risk-free-rate) / (jpm\_quarterlyReturn – risk-free-rate)

Table

Description automatically generated

1. Development of statistical model
   1. Check the normality of Quarterly Closing Price (QCP)

* Draw Q-Q plot to check normality of JMP QCP

Chart

Description automatically generated

* Perform Shapiro-Wilk test for normality

Graphical user interface, text, application

Description automatically generated

The Q-Q plot shows that there is a deviation from normality and the QCP does not entirely follow a Normal Distribution. The Shapiro-Wilk normality test yields a p-value = 0.001204 (< 0.05 at a 5% level of significance) confirming that the response variable QCP does not follow the Normal Probability Distribution. Therefore, we must apply a Johnson non-linear transformation to QCP and see if the transformation can adjust the scale of the response to follow the Normal Probability Distribution.

After conducting the Johnson Transformation is conducted using Minitab software, the csv file contains the transformed JPMorgan Chase QCP will be imported in R.

* 1. After johnson transformation on minitab statistical softwar
* Import transformed JPM QCP

Table

Description automatically generated

* Draw Q-Q plot to check normality of transformed JPM QCP

Chart

Description automatically generated

* Perform Shapiro-Wilk test for normality after Johnson transformation

Graphical user interface, text, application

Description automatically generated

We then checked the normality of the QCP for goodness-of-fit using the Shapiro-Wilk test. The normality test has p-value = 0.9337 confirming that the new transformed response indicator QCP\_T does follow the Normal Probability Distribution.

* 1. Check correlation using data transformation
* Generate the dataset containing all indicators and transformed JPM QCP

Table

Description automatically generated

* Check correlation among the indicators

Chart

Description automatically generated with medium confidence

There should be non-linear relationship among our attributable indicators to run the multiple regression model. Most of our attributable indicators are linearly associated with QCP of the stock which is a good sign to develop a multiple regresssion model for the given dataset. However, there is strong linear association among some indicators showing the posibility of multicollinearity in our dataset. Given the cut off value of correlation coefficient between 0.8 to -0.8, QCP is found to be strongly correlated with P/B Ratio, and GDP. Also, P/B Ratio is strongly correlated with GDP.

* 1. Variance Inflation Factor (VIF)
* Check Variance Inflation Factor (VIF) to verify the existence of multicollinearity among attributable indicators

Graphical user interface, text, email

Description automatically generated

Table

Description automatically generated with low confidence

* Drop indicators with VIF values > 10

Graphical user interface, text

Description automatically generated

Text

Description automatically generated with medium confidence

Dropping the variables with VIF values greater than 10 has resulted in changes to the coefficients, standard errors, and p-values of some variables in the model. The adjusted R-squared value has also decreased from 0.9493 to 0.6361, indicating a decrease in the model's explanatory power.

All indicators are conceptually important to study the relations of indicators and their two-way interactions to JPM stock price. Therefore, it s important to further assess whether the variables with high VIF values are truly collinear or if there are other reasons for their high VIF values, such as sample size or measurement errors.

Dropping variables with high VIF values from the initial model can result in a loss of information and potential bias in the model. Therefore, I decide to retain them in the model and consider the LASSO regression model as an approach to mitigate multicollinearities.

* 1. LASSO REGRESSION MODEL
* Prepare data for LASSO regression

Text

Description automatically generated

* Find optimal lambda value

Graphical user interface, text, application, email

Description automatically generated

* Train LASSO model

Text, email

Description automatically generated

Text

Description automatically generated with medium confidence Table

Description automatically generated with low confidence