# Report

December 2022

## 1 Introduction

The research question is how accurately can a regression model predict the closing price of a stock exchange (SE) based on data from another SE. The attempt to answer this question is done by seeing whether it is possible to predict the closing price of the shenzhen stock exchange (SZSE) based on the closing price of the shanghai stock exchange(SSE) using regression. The shanghai stock exchange (SE) closes only 3 minutes after the SSE stock exchange(1). If predictions can be accurately made on the closing price of the SSE then stocks and options can be traded to make a profit. While 3 minutes doesn't seem like a lot of time traders are able to use algorithms to trade within fractions of a second where each microseconds counts (2). I will use machine learning (ML) to train regression models in an attempt to find the most accurate one.

Machine learning techniques are used to predict and analyse stock markets. Using deep-learning models the daily closing price of the New York Stock Exchange (NYSE) was used to predict the daily closing price of the Indian National Stock Exchange(NSE)(3). Despite being located far away from one another they worked in a similar manner and were affected similarly by external factors. This was made proven because the models succeeded at predicting the price.

I also want to see how the model performs with time. A study reports that increasingly split second decisions are increasingly important as access to the internet and automation increases(4). This allows users to trade instantly when a set of predetermined criteria are met or an algorithm can trade thousands of times a second if an advantage is found. However over the longer term these algorithms are less effective and need to be retooled and 'retaught'. The data I am examining stretches back into the 90s so the model can be taught using this data and it can be seen whether the model is relevant years later.

## 2 Methodology and Dataset

The dataset I chose has very few variables but a large number of datapoints. It contains information about various SE that is updated every day they are open. The data contains the highest lowest, close and close adjusted price of the SE. The close adjusted price is the close price after taking into account dividends or

splits that may affect the stock price when the market next opens. The prices are in the local currency. The daily trading volume is also present however there are large patches of this data missing so I will not be using that variable.

I chose SZSE and SSE for their similarities. They rarely open on different days and use the same currency so the data processing necessary when comparing these 2 is minimized. They are also located in the same currency and subject to the same laws and regulation. They also have the strongest correlation across a range of variables. I found this by comparing pairplots between different SE. While only two exchanges were chosen there is a large amount of data for each exchange since it is almost daily data that dates back to 1997.

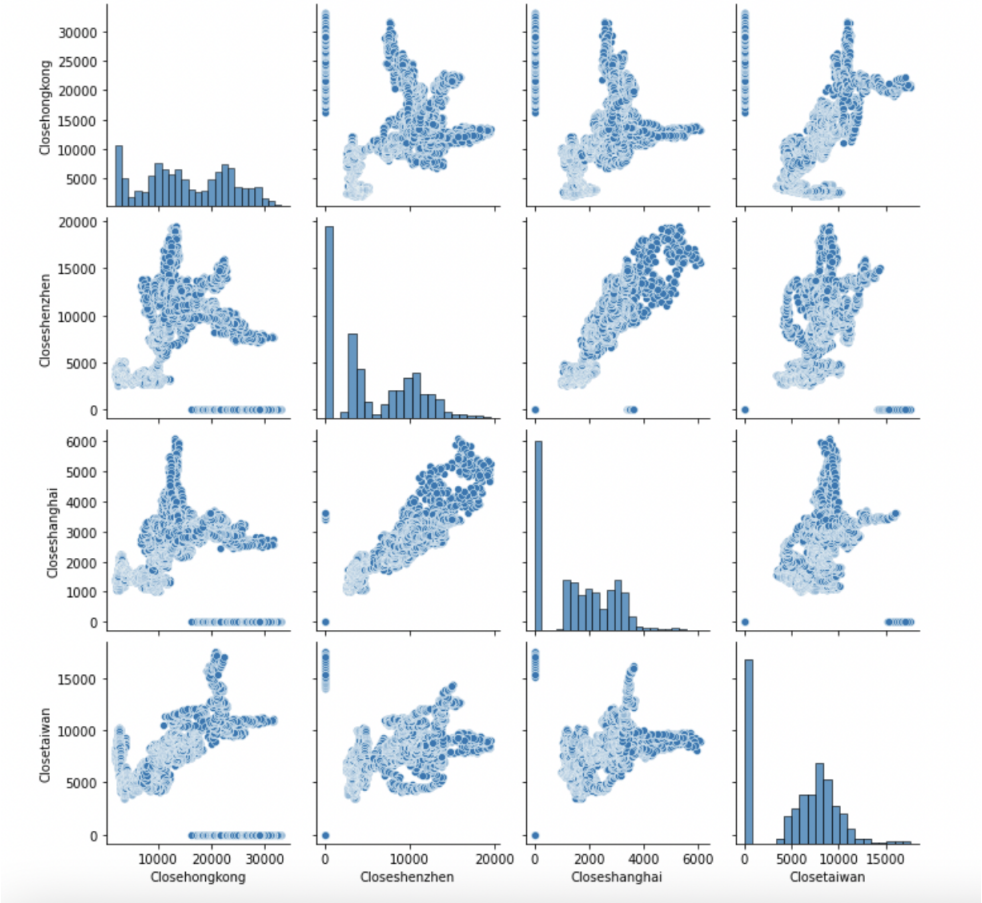


Figure 1: Pairplot of the closing price of SE in Asia

I cleaned the data by removing the volume column as it did not have enough data to reliably analyse. I removed all unique dates so that each datapoint can be compared. There were not a large amount of unique dates so it did not

noticeably affect the data. I added two additional features "Daily Change" and "High Low Difference". This was because the change from the opening to closing price in SSE was correlated with the closing price of SZSE. The difference betwen the high and low price of the day also had a correlation with the closing price of SZSE. I will experiment with using different sizes for training and testing the model to see if it has an effect.

I conducted 4 types of regression modelling and modified the features for each one to try and find the model that had the best fit. I will analyse the accuracy by seeing how accurate each model is on the test. I will analyse how applicable the model is by trying to predict and predict SSEs price using the SZSE data. I will analyse how long the model is relevant for by building a model on old data and testing the predictions out on modern data.

# 3 Results

The training for LASSO regression and ridge regression were quite simple. I started off randomly changing features however there was almost no difference in the accuracy of the model. This was the same for the polynomial and multiple linear regression models. Most of the time the regression score was at .88-.89 with the best being 1.0 and the worst 0 regardless of the feature selection used. This held true across all 4 types of models. Changing the size of the training and test data set didn't have a consistent effect on the accuracy of the model.

The linear regression model shown below shows the linearity of the relationship between the two datasets and that it can therefore easily be modelled by a linear regression model. It can also be mapped by a polynomial regression model however I noticed that when the polynomials reached a higher order the model became less accurate or the higher order polynomials were given significantly less weight than the lower order polynomials.

```
In [19]: plt.scatter(y_test, predictions)
Out[19]: <matplotlib.collections.PathCollection at 0x7f7f8c789ca0>
```

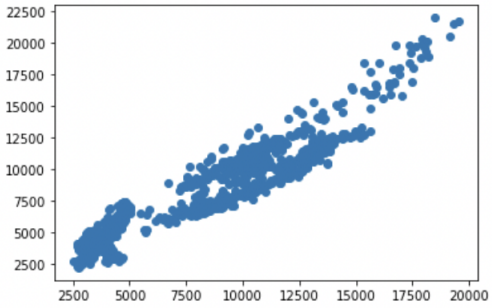


Figure 2: Scatter Plot of the Linear regression predictions

I do think that the model may be overfitted. This is because I when trying

to use the SZSE data to predict the SSE price with the same model it had a low score and was unable to predict the price accurately.

The Polynomial regression model was the one that was most overfit with the overfitting increasing as the order of the polynomial increased. I however found that linear regression didn't work as well with a larger number of features and was best when using 1-2 features to predict the closing price.

The model also does not hold over time. I created a LASSO and ridge regression model trained on the first half of the data so from 1997/08/15 to 2009/07/20. I then tried to use it to predict the second half of the data so from 2009/07/21 to 2021/05/31. The models worked well on the old data with a regression score more accurate than the model trained on the entirety of the data of .93-.94. When the model was used on the modern data the score was negative, meaning it was unsuccessful at accurately predicting prices.

# 4 Discussion

It was possible to predict the closing price of two stock exchanges using a machine learning model. In doing so it is possible to make trades with future knowledge and make a profit. The LASSO regression was the most accurate and least variable when testing with different hyperparameters so I would use this one.

The highest correlation is also between the difference between the highest and lowest prices of the day of the SSE and the closing price of the SZSE. This was surprising was consistently given more weight in both linear and polynomial regression. For linear regression the coefficient of this variable was an absolute value of 6.5. The Daily change and closing price had a similar coefficient of 3.3 and 3.8 respectively and were therefore less important. This level of correlation between the data types stayed conssitent over different models

It should be noted that it is likely best to create a seperate model using similar methodology for each different feature and SE. This may be because the data can be very closely correlated between 2 specific SE and not translate over as effectively to other SE.

The model being unable to accurately predict prices using old data means the model should be regularly updated with more recent data. It also means that it may be more advantageous to have a model that uses exclusively recent data rather than all available data. It could however also simply mean that the model was overfitted to the old data. This could be exacerbated because the model was trained on a smaller data set than the model trained on the entirety of the data.

The modelling was limited in that it only analysed two SE and only had access to a very limited number of features to draw from. These features were also very correlated with one another, all being focused on price. I would likely add more data features that do not have anything to do with the price. This could for example make sure that the trading volume is there for every data point. This would allow for more variables to predict the price of a SE without

being linked to the price of another SE. This means that any fluctuations in price such as a depreciation of a currency wouldn't affect the model as much.

Further steps would be to analyse how data from multiple SE can be used to predict the price of one SE as more data is likely to make the model more robust. It also reduces the variance if for some reason one SE is disproportionately affected by something that no other SE is then the model's accuracy is not as severely impacted as if it relied on one SE.

# References


[1] "List of stock markets tradinghours.com." [Online]. Available: https://www.tradinghours.com/markets/

[2] S. Asthana, "Making money in microseconds: This trader reveals his success recipe for algo trading," Dec 2018. [Online]. Available: https://www.moneycontrol.com/news/business/making-money-in-microseconds-this-trader-reveals-his-success-recipe-for-algo-trading-3293791.html

[3] H. M, G. E.A., V. K. Menon, and S. K.P., "Nse stock market prediction using deep-learning models," *Procedia Computer Science*, vol. 132, p. 1351–1362, 2018.

[4] D. Shah, H. Isah, and F. Zulkernine, "Stock market analysis: A review and taxonomy of prediction techniques," *International Journal of Financial Studies*, vol. 7, no. 2, p. 26, 2019.