0.1 Introduction

For this CA I will be investigating the AT&T stock prices dataset. I will be aiming to anwser the following question: Can I predict the future price of a stock based on a time series of previous stock data. This time series will not contain all of the data from the previous time periods and the specific data that will be used will be chosen as part of my hyperparameter selection. The data for the stock prices has come from kaggle and contains stock data for AT&T between the years 2000 and 2020.

0.2 Methodology and Dataset

Some of the key data in this dataset is: The date (DATE) The volume of the stock (VOL), this is the amount of shares that were traded on that day. If a stock has a high volume then it means that shares for that stock changed hands a lot. Opening price (OPENPRC), this is the value that the stock was first traded at on that day. The asking high (ASKHI), this is the highest asking price for the stock on that day. The asking price of a seller is the lowest price that they will sell their share for. This makes the ASKHI the theoretical maximum that the stock was worth that day. The lowest bid on the stock (BIDLO), this is the lowest bid that was put on the stock in that day. A bid is the highest price that someone is willing to pay for the stock. This makes the BIDLO the theoretical minimum a stock was worth that day. the closing price of the stock (PRC). this is the final price that the stock was traded at during the day. An interesting combined datapoint that could be included is the liquidity of the stock, this could be calculated using the ASKHI and the BIDLO and is a representation of how volatile the price of the stock is.

these are the data points that will be the most useful when we are trying to predict a stocks future values. When we are creating a model for predicting the future of the stock we dont have to worry about things in this dataset like: PERMNO which is just a unique identification number for the shares that are being traded, this will be the same over all the datapoints and so is useless to include in our model. The same justification can be used to remove the company name (COMNAM), TICKER, CUSIP, NCUSIP, PERMCO, SHRCD, ISSUNO, EXCHCD, SICCD, HSICCD, PRIMEXCH, TRDSTAT, SECSTAT

the following datapoints are not complete for all entries but when they are present they are identical across all entries and will therefore be discarded for this reseach question: SHRCLS, HSICMG, HSICIG, NAMEENDT, TSYMBOL NAICS

In this dataset there are 2 companies that have their stock data listed so the research question can be asked for both independantly. Another hyperparameter that we can adjust in our model is how many time steps back we feed it to get our predictions. it will be interesting to see how it changes given more data and possibly at what point more data stops improving the mdoel.

i will be using the following pieces of data... for the following reasons...

0.3 Results

For this project I have used Linear regression and Lasso regression and Ridge regression to investigate the impact of stronger regularisation on my research question. For my initial model I had chosen to include all the data that i have deemed important in the section above for 5 previous days on the time series. This resulted in accurate predictions for just 1 day into the future. This was also the case with bot ridge and lasso regression with linear and ridge regression having almost identical mean squared error and lasso falling just a little behind

Table 1: Means Squared Error for predictions 1 day into the future using all important data

Algorithm	MSE
Linear Regression	0.3135
Lasso Regression	0.3994
Ridge Regression	0.3135

For the next set of models I removed some of the data and only left what I thought would be the most impactful for predicting the future price of the stock. This was PRC VOL and LIQ (liquidity). Doing this did not improve my results for any of the models when compared to using all of the available data.

Table 2: Means Squared Error for predictions 1 day into the future using all important data

Algorithm	MSE
Linear Regression	0.3143
Lasso Regression	0.4102
Ridge Regression	0.3143

Predicting one day into the future isnt very useful to us as the timeframe is so short that we can resonably predict one day ourselves. This means that we have to extrapolate further if we want to have our model be more useful. With this further extrapolation comes more error in our model. When increasing the time frame we still see that linear and ridge regression perform very similarly with lasso regression falling slightly behind however, this time there is an order of magnitude more error than from our one day predictions

Table 3: Means Squared Error for predictions 1 day into the future using all important data

Algorithm	MSE
Linear Regression	2.6592
Lasso Regression	2.7783
Ridge Regression	2.6591

0.4 Discussion