Machine Learning Short Term Stock Price Movements - A2M.AX

Sam N.

Abstract

Using a daily price and volume dataset for the stock A2M.AX (listed on the Australian Stock Exchange), I wanted to find answers to the question - With the help of daily stock data and basic machine learning tools, are we able to predict short term future price movements of stock?

I used Decision Tree Classifiers and a multi-train-test-split approach to see if I could answer this question using my dataset.

My findings concluded that given my dataset and approach, I could not predict future stock price movements for A2M.AX with any reliable accuracy.

Motivation

The reason for this investigation is that I've always been curious about whether future stock prices can be predicted with any accuracy whatsoever using historical time series data and basic machine learning methods.

This insight could be valuable for stock traders who are interested in using machine learning or quantitative methods to trade stock profitably.

Dataset

The dataset used is from Yahoo Finance. It contains 1,180 data samples of daily stock data for the stock A2M.AX, listed on the Australian Stock Exchange. The data range is from 31 March 2015 to 22 November 2019.

I added my own features to this dataset in excel. Additional features include features related to range, value changes, and averages of existing features. I thought these additional features could be useful in the classification models.

Similar datasets can be downloaded from https://finance.yahoo.com/

Data Preparation and Cleaning

Null data samples were dropped. The null data samples were null due to some of their features, notably features that were averages for preceding days, not having enough previous days of data to average, thus returning null.

Features that were movements (e.g. ‘Up’, ‘Down’, ‘Flat’) were changed to 1, -1, 0 respecitvely so that they could be modelled appropriately.

Dates needed their datatype changed to datetime format so time related actions could be performed on the data.

Research Question

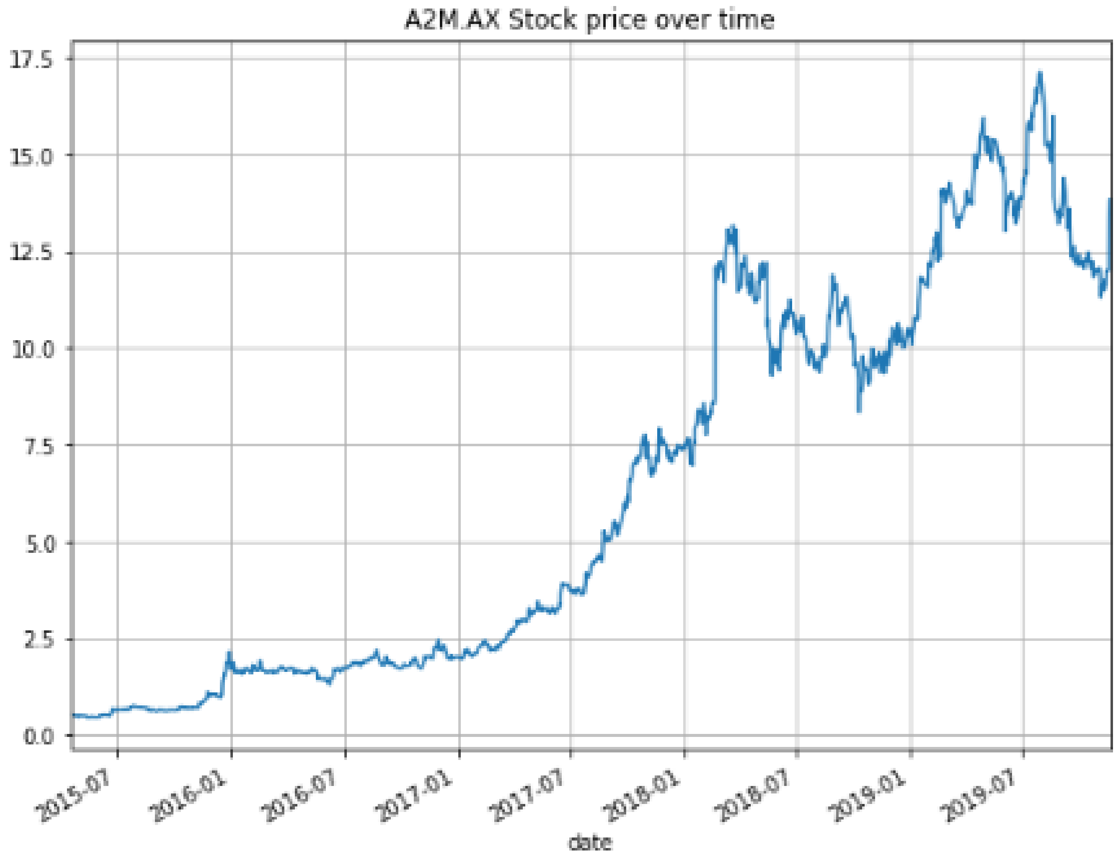
With the help of daily stock data and basic machine learning tools, are we able to predict whether the next day’s price for a stock will move up or down?

Methods

To analyze the data I used Decision Tree Classifier models. All of the data was passed in as features to the model, and the target was tomorrow's price move (Up or Down). As the target is discrete (Up or Down), I saw using a classification model as appropriate.

To facilitate the analysis, as the data is time series, I used a multi-train-test-split model to train and test the model. During my analysis I discovered that a random train-test-split model is not appropriate for time series data, as future information should not be used for predicting historical values.

For context, here is A2M.AX's stock price over the time period examined



Findings

By using a multi-train-test-split model to test and evaluate our Decision Tree Classifier models, we can see how the Decision Tree Classifier models changed their feature weights per iteration.

In the next slide we can see this. Interestingly, only the “range_change” feature appears in all of our splits. “change_in_5d_avg_range” was another dominant feature, being weighted in all models except Split 2.

Also interesting to note is that the different decision tree classifiers had as low as 6 features and as high as 9 features per iteration.

	Split 1	Split 2	Split 3	Split 4	Split 5	Split 6	Split 7
open	NaN	NaN	NaN	NaN	NaN	0.104523	NaN
high	NaN	0.130463	0.220674	NaN	NaN	0.104252	NaN
close	NaN	NaN	NaN	NaN	0.111289	0.106151	0.079706
adj_close	0.122858	NaN	NaN	NaN	NaN	NaN	NaN
range_over_open	NaN	0.088336	NaN	NaN	NaN	NaN	0.159397
range_change	0.068106	0.229291	0.212057	0.126350	0.123912	0.113098	0.134874
price_change	NaN	0.236483	NaN	0.097371	NaN	NaN	NaN
volume	0.125167	NaN	NaN	NaN	NaN	0.117920	NaN
qty_change	0.121482	0.228583	NaN	0.313288	0.115478	NaN	NaN
5d_avg_range	0.207370	NaN	0.116110	NaN	NaN	NaN	0.117246
change_in_5d_avg_range	0.178979	NaN	0.114623	0.234229	0.233727	0.208796	0.080880
5d_avg_px	NaN	NaN	NaN	0.118798	NaN	0.138495	0.124164
change_in_5d_avg_px	0.176038	0.086843	NaN	NaN	0.210149	NaN	NaN
5d_avg_qty	NaN	NaN	NaN	0.109964	NaN	NaN	0.107643
change_in_5d_avg_qty	NaN	NaN	0.113278	NaN	0.106649	NaN	0.098381
10d_avg_px	NaN	NaN	0.128376	NaN	NaN	NaN	NaN
change_in_10d_avg_px	NaN	NaN	0.094883	NaN	0.098795	0.106765	0.097708

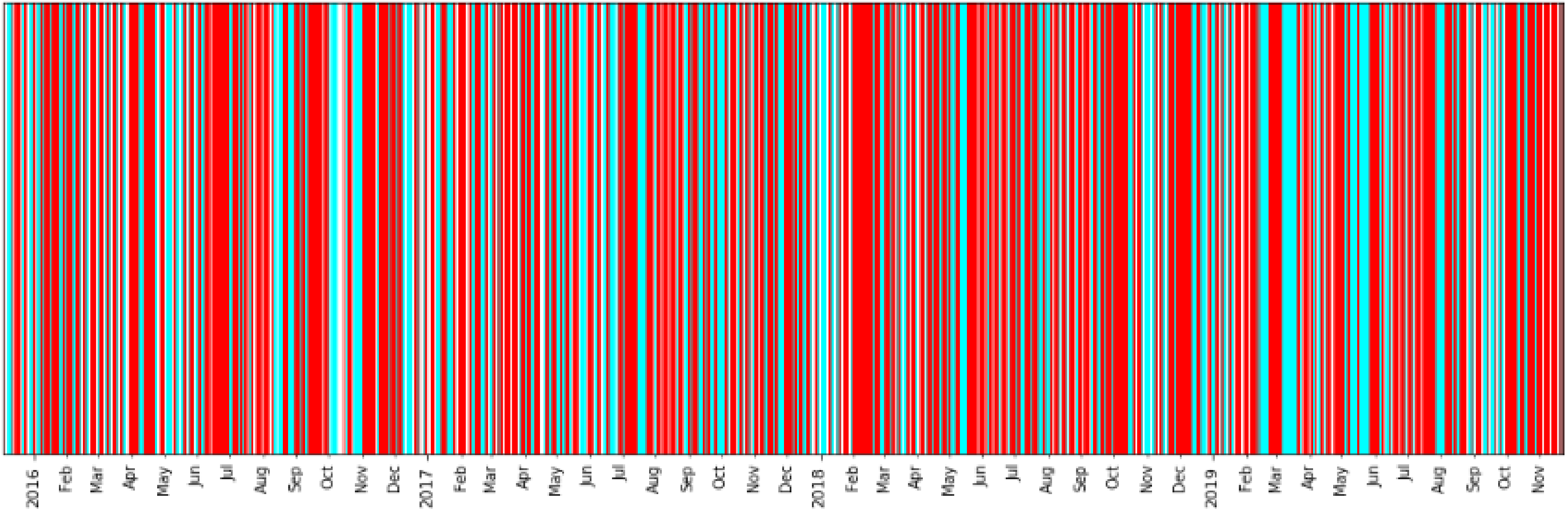
Findings

In checking our multi train test split model’s accuracy, I found the accuracy score to be 0.5151. This accuracy seems to be positive, although it is more or less the same as guessing and could be above 0.5 due to chance.

When visualizing our predictions we can check to see if there are any consistencies or seasonality in our predictions.

The results are plotted on to a binary time series bar chart. Correct predictions are cyan and incorrect predictions are red.

Decision Tree Classifier prediction accuracy over time



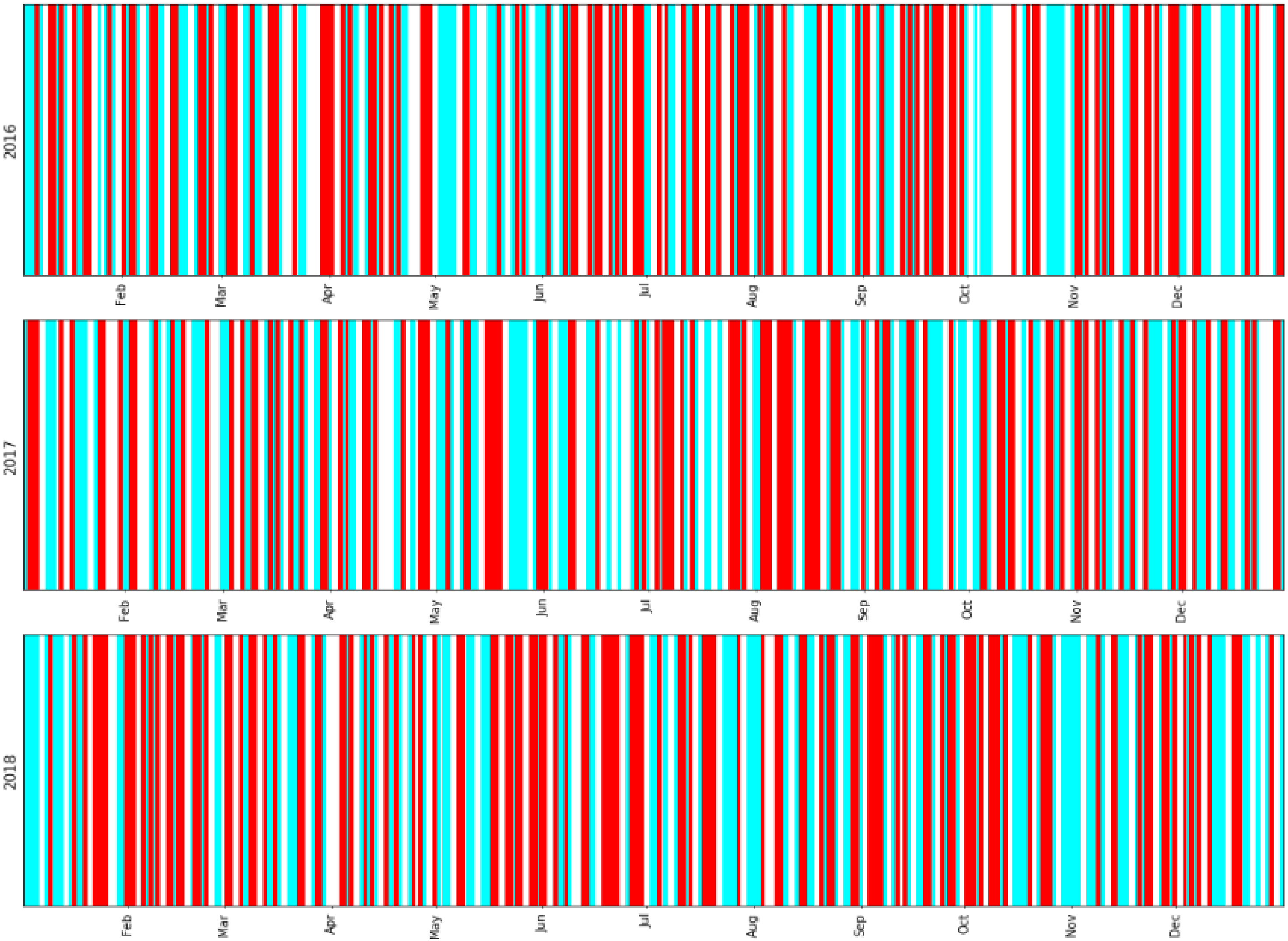
Findings

There are white spaces in our data. This is due to daily data preceding flat price days being removed, and for there being no stock data on weekends or holidays.

From this plot it is difficult to see any consistencies in the model's predictions.

The next slide plots 3 full years of predictions for 2016-2018 stacked on top of each other. It should help make finding any consistencies or seasonality easier.

Can you find any consistencies in predictions over these 3 years?


2016
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2017
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2018
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

Findings

Even stacked on top of each other, it is still difficult to find notable consistencies in the data.

Perhaps if we zoomed in further to monthly or weekly timeframes it would be more obvious, or perhaps there isn't any consistency or seasonality in the data at all!

For 2016-2018 I collated a list of the count of correct and incorrect predictions per year. Let's look at this in the next slide.

Findings

Year	Prediction	Count
2016	Correct	122
2016	Incorrect	113
2017	Correct	127
2017	Incorrect	110
2018	Correct	122
2018	Incorrect	125

There are a few things to note about this data. 2016 and 2017 appear fairly consistent, while 2018 has far greater incorrect predictions.

The count of total predictions in 2018 is 247, which is greater than that of 235 in 2016 and 237 in 2017.

The reason for this is most likely due to there being less flat day price moves occurring due to the price of A2M.AX getting higher, and thus there being less data points removed from the data in 2018.

Limitations

It is possible that the predictions in 2018 are less accurate compared to prior years because older training data from as early as mid 2015 is still being used to train the model to predict price movements in 2018.

The stock characteristics went through massive changes in this 3 year period. This was visually evidenced through the stock price time series in slide 8. The statistics between these periods are widely different.

Limitations

A2M.AX in 2018 (and beyond) appears to be in a different state to where it was a few years prior, as can be seen mostly by the increased volatility.

Due to this, perhaps we could retrain and test our model using a maximum size per training split. In effect, this would mean our models per split are being trained only on the most recent x data points, rather than continuing to be trained on more historical and perhaps less relevant data.

Conclusion

It seems clear that we have not founded a crystal ball for stock price movements.

The accuracy scores for the model, albeit above 50 percent for the overall analysis, perform poorly on more recent data (122/247 = 0.494 accuracy score in 2018).

This can be attributed to the unpredictable nature of the stock market, as can be seen in the year 2018 for A2M.AX, which was significantly more volatile than prior years.

Conclusion

To answer the initial research question - With the help of daily stock data and basic machine learning tools, are we able to predict whether the next day’s price for a stock will move up or down?

My answer has to be no. At least not in this analysis!

The potential for further analysis is really endless - the features themselves can be explored further by looking at mean, mode, median, standard deviation and autocorrelation among other statistical measures.

With deeper analysis I do believe statistically significant results could be found to help predict tomorrow’s stock price movement.

Acknowledgements

Data acquired from https://finance.yahoo.com/

You can obtain similar data yourself by creating a profile on Yahoo Finance, and adding whatever stock or stocks you're interested to your portfolio, then exporting the data using csv format.

Thank you Grady for the feedback on which train-test model to use for time series analysis and other tidbits.

References

Jason Brownlee (December 19, 2016), ‘How To Backtest Machine Learning Models for Time Series Forecasting’. https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting/