1. **Choice of dataset**: Explain the reasons why you choose this dataset. If you are going to create your own custom dataset, explain what kind of data you will be scraping.

https://www.kaggle.com/tarunpaparaju/apple-aapl-historical-stock-data

2. **Methodology**: Describe how you plan on approaching the project. This should be a high level overview of your plans, and this will allow us to judge the feasibility of your project. Be as thorough as you can, so we can give you critical feedback if necessary.

a) Data Preprocessing: Is the dataset you chose feasible? What information provided is/are the most useful? How are you planning on preprocessing the dataset to extract this information? You can take a look at these F2019 slides on data preprocessing.

The dataset is feasible because it has AAPL stock data going back 10 years which is enough to make predictions for the near future. The Volume column can be removed as the focus is on the open, close, high, and low for each day, and for these values we will need to remove the '$' and cast it into a float. The dates will also have to be parsed into dType datetime64.

Closing price will be the most useful resource. We will use a standard scaler to make mean = 0 and scale the data to unit variance. The reason why we use the *StandardScaler*, rather than the *MinMaxScaler* as you might have seen before. The reason is that stock prices are ever-changing, and there are no true min or max values. It doesn't make sense to use the *MinMaxScaler*, although this choice probably won't lead to disastrous results at the end of the day.

We will add exponential moving average predictions as input vectors and for the rows without the moving averages will be discarded, which means NAN will not be replaced with mean.

stock price data in its raw format can't be used in an LSTM model directly; we need to transform it using our pre-defined `*extract_seqX_outcomeY*` function. For instance, to predict the 51st price, this function creates input vectors of 50 data points prior and uses the 51st price as the outcome value.

b) Machine learning model: What do you want to predict/estimate from this dataset? Propose a machine learning model/algorithm for it, and explain your reasoning. Have you considered other alternative models? What are the pros and cons?

I want to predict how the future stock price is going to be. LSTM (long short term memory) will be suitable for this problem because it can capture historical trend patterns, and predict future values with high accuracy. The type would be many inputs to many outputs. (*Predicting Stock Prices Using Machine Learning*, 2022) LSTM is RNN (Recurrent Neural Network) based. A disadvantage is that LSTM based RNNs are difficult to interpret and it is challenging to gain intuition into their behavior. Also, careful hyperparameter tuning is required in order to achieve good results. (Shah, 2020)

The Prophet model also works and its advantage is that it requires less hyperparameter tuning as it is specifically designed to detect patterns in business time series. Its disadvantage is that it can fail spectacularly on time series datasets from other domains.

ARIMA models can also be used however it needs careful hyperparameter tuning and a good understanding of the data.

Regression models can be used too, however the residuals of stock prices are not normally distributed and there is heteroskedasticity. Therefore, it requires an approach to the data. Residuals is the difference between an observed value of the response variable and the value of the response variable predicted from the regression line. (*Residuals*, n.d.)

FYI: ESN (Echo State Network) is RNN based and it accounts for chaotic dynamics of the stock market by utilizing a hidden layer with several neurons flowing and loosely interconnected, which is designed to capture the non-linear history of input data. A linear activation function is applied to calculate the final predictions.

Hybrid model: we add MA (moving average) predictions as input vectors to the LSTM model. MA has two types: SMA(simple) & EMA (exponential).

https://data-flair.training/blogs/stock-price-prediction-machine-learning-project-in-python/

https://neptune.ai/blog/predicting-stock-prices-using-machine-learning

Note: We are aware that at the point of Deliverable 1, many machine learning models have yet to be covered in lectures. We expect you to find out what models are generally used by the ML/AI community to tackle the problem / make use of the dataset you chose, and to get a simple, intuitive understanding of how those models work.

Suggestion: Each Kaggle dataset has a “Code” tab that contains previous work done using the dataset. It can be an excellent source of inspiration for this section!

c) Evaluation Metric: Analysis requirements differ in every field, but some things to consider reporting include but should not be limited to:

i. Confusion matrix and accuracy/precision-recall/logistic loss (classification problems).

ii. Mean squared error (regression problems)

iii. Rand index (unsupervised models)

iv. BLEU score with brevity penalty (text generation)

v. Variance of the dimension reduced set vs variance of the initial dataset (dimensionality reduction/PCA)

If you are not sure, ask your assigned TPM.

Furthermore, you should be able explain the specific problem's accepted metrics. Keep track of the average baseline results which you hope to beat (For example, predict X with at least Y % accuracy).

The model will be evaluated based on the mean squared error and can expect a relatively high MSE because stock price is very difficult to accurately predict and depends on many other external factors.

3. **Application:** We want you to integrate your model in a simple landing-page webapp. For those of you who have experience, you are welcome to integrate your model in more sophisticated technologies (eg. mobile, hardware, webapps).

In this section, give the general idea of your application:

- - What does the user input? How does the user provide inputs? (Is there a webcam?

Specific date of the predicted stock price the user wants to know by typing.

A way for users to submit images? text?)

- - What does the user receive as output, how will the output be displayed?

The predicted stock price will be shown as a text and if image of the graph containing predicted stock price & past stock price.