05/21

05/10/2021

Please use this template. Retain the gray text. Your new materials—in black 12-point Times New Roman—should not exceed 5 additional pages excluding references and figures. Use the Appendix for bulky material that will be read on an as-needed basis only. Note the evaluation criteria, and leave plenty of time for editing.

# ASSIGNMENT 2P: PROJECT PROPOSAL PLUS

Aidan Duffy

## 2.0 WHAT’S CHANGED (if applicable)

Provide no more than a page of 12-point type explaining what has been changed or added since assignment 1P (if you selected it). Include in this whether and how the material in module 2 influenced this, or refer to reading that you did in working on this assignment (#2).

N/A – did Assignment 1T.

## 2.1 SUMMARY DESCRIPTION, VERSION 2 (as applicable)

One- or two-paragraph overall description of your proposed term project.

I want to develop a machine learning program, called CRYPTOFORECAST, that will help develop price predictions for the cryptocurrency market. Many programs are already tailored to predicting fluctuations in the stock market, and I did not want to reinvent the wheel. Additionally, given the rise of the popularity of investing cryptocurrencies like Bitcoin and Ethereum (and yes, even Dogecoin!), CRYPTOFORECAST will use the same fundamental financial indicators and technical analysis but rely on learning from a dataset that analyzes the largest tokens currently in circulation. This program may help future investors in DCA, or dollar cost averaging, to lower their cost bases, as well as realizing when the end of a “bull run” may be approaching.

## 2.2 I/O EXAMPLES, VERSION 2

At least two specific examples of projected output for designated input. You will not be held to this exactly.

Here is the I/O demonstrating the price history of a given input token:

Graphical user interface, text, application

Description automatically generated

## 2.3 FUNCTIONAL REQUIREMENTS, VERSION 2

Separate your requirements into two approximately even categories (select modest “definite” requirements, otherwise “nice-to-do”). This organization allows you to first attain readily do-able goals without getting bogged down, and then move on to other goals as you are able. State requirements in declarative language as advised in assignment 1. Giving each requirement a label (e.g., “(Recognize 0-9):”) helps with clarity and readability.

Your response replaces this.

### 2.3.1 Definite Functional Requirements (first priority)

### 2.3.1.1 Definite #1: Read Datasets

CRYPTOFORECAST shall read in and format all of the datasets.

### 2.3.1.2 Definite #2: Filter and Normalize Datasets

CRYPTOFORECAST shall sort data, apply necessary filters, and normalize the filtered datasets.

### 2.3.2 Nice-to-do Functional Requirements (second priority)

### 2.3.2.1 Nice-to-do #1: Integrate Visualizer for Datasets

CRYPTOFORECAST shall provide visualizations and graphs for the price history of cryptocurrencies in the datasets.

## 2.4 V2: HOW SUCCESS WILL BE ASSESSED

Explain, as specifically as possible (quantification is ideal) how success of the project should be assessed.

Luckily, given the short feedback cycle of financial markets, success can be assessed almost immediately. I will be able to cross reference my model’s price predictions daily with actual market prices.

## 2.5 V2 TECHNOLOGY EXPLANATION

Explain what two specific machine learning technologies you intend to use--and exactly why you feel they apply to your particular project. One of the two may be emphasized as the implementation and the other as an alternative or as a complement—discussed but not implemented if need be. Include a discussion of neural nets, whether you select them or not.

prices. After doing research on similar projects that have already been completed, I will attempt to utilize a recurrent neural network, or RNN. More specifically, I will attempt to implement an LSTM, or long short term memory, setup. This type of neural network was selected most often because of its ability to store past information for sequence prediction problems, making it a most powerful tool in a machine learning developer’s toolbox. Since the dataset will be sequential, the data from one hour or day is directly related to the data point from the previous unit of time, again reaffirming the importance of this style of neural network. As a proposed alternative, in my research I found an algorithm called “XGBoost”(Extreme Gradient Boosting), which a tree based supervised learning algorithm. It is often used as well as it is very efficient. The main reason it is an alternative is because all of the data would first need to be preprocessed and transformed into a supervised learning problem first.

## 2.6 CODE EXAMPLES

Show fragments of code execution. For example, if you are using TensorFlow, show that you have run some relevant code. The more relevant to your project, the better. This does not need to complete at this stage.

This is the code addressing functional requirement number one:

Table

Description automatically generated

Here is code snippet 2 that addresses the graphing functionality of price histories:

Graphical user interface, text, application, email

Description automatically generated

## 2.7 DATA SOURCES V2

Explain whether or not your project requires data. If so, describe were you will obtain it. Be careful about this because you won’t have a project if it needs data and you have to spend too much time hunting and gathering it.

This project most definitely requires a dataset, and as of writing this, I am using one acquired from Kaggle [here](https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory). If I determine this data is insufficient, there are ample other sources of data available on Kaggle that I will be able to utilize.

## 2.8 REFERENCES FOR PROPOSAL V2

Fill in, and also cite each of the following (e.g., “[2]”) within the text. References can include specific places in the notes and textbook. You are free to include references used in the prior assignment version. Keep in mind that this “use of resources” is a whole evaluation criterion.

[1] <https://towardsdatascience.com/lstm-time-series-forecasting-predicting-stock-prices-using-an-lstm-model-6223e9644a2f>

[2] <https://cs230.stanford.edu/projects_winter_2020/reports/32066186.pdf>

[3] <https://medium.com/swlh/a-technical-guide-on-rnn-lstm-gru-for-stock-price-prediction-bce2f7f30346>

## 2.9 Evaluation of Assignment 2



2.10 Appendix

# Assignment 4: Project Design + Implementation Preliminaries

Please limit this to 6 pages of 12-point text excluding figures, references, and appendices. This revision is your final view of the design prior to implementation (though you may still change it when you implement).

## 4.0 What’s New

Summarize what is new about your project since the last submission. This should reflect 15-20% of the total project effort.

Since my previous submission, I have continued designing the overall structure of the product and implementation changes as well as some new steps towards implementation. With regards to changes to the previous implementation, I refined the data normalization completed after everything is read-in and processed. In terms of new implementation, I integrated some averaging models for the prediction forecasting. This includes calculating the MSE, or mean squared error, for the standard average and exponential moving average, or EMA, which are both common metrics used in financial asset prediction models. I will discuss the actual design decisions in the 4.2 section.

## 4.1 Relevance of RNNs and GANs

Explain whether or not RNN’s and GAN’s could or should apply to your problem, and why/not.

RNNs apply to my project, as I will be implementing an LSTM, or long short-term memory, model. I found LSTM to be a useful form of the RNN in my research about past implementations of stock price prediction programs. It is primarily useful because of the ability to retain information about data over long stretches of time.

## 4.2 Design and Theory: Final Version

Describe the final pre-implementation version of the design of your proposed system. Use annotated diagrams. Explain the theory behind your design. Explain how the two technologies will interface or compare. The reader should understand how you plan to fit the pieces together. Show this at a high level, as well as providing as much relevant detail as you can. Include at least one (meaningful) figure.

Once the data is uploaded, the program will work as follows. Pandas will store all of the csv files’ data into pandas data frames. From there, some preliminary graphs will demonstrate the price history of selected cryptocurrencies. After this, pandas will split data into training and testing datasets. Once the datasets are created, CRYPTOFORECAST will normalize the data using a Min Max Scaler and utilize TensorFlow to implement an LSTM RNN.

**Note**: As I fell a bit behind in past weeks, I have not solidified all of the details for the specific parameters below (i.e. how many layers the neural network will have, but will be sure to include that information in the coming submissions).

The LSTM RNN will use several hidden layers. The RNN will need hyperparameters, like the number of hidden neurons for each cell or information related to the backpropagation through time for optimization. The BPTT hyperparameter will assist in determining how many continuous steps CRYPTOFORECAST will take into account for one optimization step. Additionally, CRYPTOFORECAST’s RNN will require standard inputs and outputs as well as parameters for the layers of the LSTM and a regression layer which would calculate the final prediction for the program.

Further, the program will need a tool to determine loss and an optimizer for the RNN. I will use the MSE for the calculation of loss, and I researched that “AdamOptimizer” is a recently developed optimizer that generally has been deemed as performing well. It is a stochastic gradient descent optimizer.

Once all of the data is reshaped by numpy, the model will create predictions using the testing dataset. After this, CRYPTOFORECAST will output a visualization to illustrate the predictions generated.

## 4.3 Tools: Final Version

Describe the tool(s) (e.g., TensorFlow) you will definitely use, or explain why you will build from scratch. Support the fact that you have reasonably investigated and tried out tools. Explain your choice. Show samples that make you and your reader reasonably confident of your choices.

TensorFlow will be one of the most important, along with pandas, numpy, and sklearn, tools that will be imported to complete the project. Pandas, numpy, and sklearn will be used for the standard data management, storage, and manipulation. I will be using pandas data frames to store the data, numpy for relevant operations, and sklearn for normalizing and preprocessing all of the cryptocurrency pricing data.

TensorFlow, specifically, is important has an API called “RNN API” that CRYPTOFORECAST will utilize to implement a time series model. As mentioned in my HW2 submission, one of the other tools that I discovered could be used was XGBoost. While I noticed that it processed the test and training datasets more efficiently than a TensorFlow implementation, upon more extensive research, I realized that XGBoost performs better when given additional features. While I mention in later sections below my intent to integrate a possible sentiment analysis aspect for the model, given my relative lack of experience in the field, I want to focus primarily on historical data points first and have a functioning program. Perhaps if I were to have done a project such as this one a few years down the line with more knowledge and experience, I would have leaned more towards XGBoost. However, the TensorFlow implementation of the RNN for CRYPTOFORECAST was easier for me to comprehend and ultimately take steps towards implementation. Additionally, given that the size of my current dataset is smaller, I wanted to prioritize accuracy and performance over efficiency. Further down the line in the project, however, I plan to revisit this decision for further analysis.

## 4.4 Challenges

Describe the challenges that your particular project faces. Avoid generic statements—focus on *your* project objectives. Explain how you intend to overcome them.

One of the largest challenges that faces this project, one that is separate from a typical stock price prediction machine learning program, is that the cryptocurrency market is both significantly more volatile than the stock market and can be swayed very easily by sentiment. As an example, the price of 1 BTC, which is meant to be the overall market indicator, fluctuates several hundred percent in a given year whereas the typical year for the S&P 500 rarely reaches double digits. I believe that since the whole market is volatile, that should be factored in properly from the LSTM RNN as the market has always behaved this way. I will need to ensure that even the direction of the price is correct in terms of increases and decreases, the scaling of the data is correct. If not, CRYPTOFORECAST will need to complete some post processing.

Further, single influential actors, such as Elon Musk, have the power to greatly influence prices of not only individual tokens but the whole market. For example, while Musk’s influence on Dogecoin is similar to his influence over his own Tesla stock, when his company stopped accepting BTC as payment for cars, the whole market fell 10% over the course of a few hours. These factors will need to factored in to make a more robust and accurate program. Therefore, adding a sentiment analysis stage would help improve the overall accuracy of the program. I will use Twitter as my main source of sentiment data as well as Google searches. CRYPTOFORECAST should be able to overcome the influencers with that.

One other issue that is more so specific to the cryptocurrency space is that individual asset owners, known as “whales” can own tens to hundreds of millions of USD worth of cryptocurrencies, and when these specific folks decide to sell (without the fanfare of someone like Elon Musk posting about it on Twitter) a massive amount of their assets, this can instigate a large dump or “flash crash” in the asset’s value. I hope to be able to factor in volume of transactions as well as the size of large purchases or sales, which can instigate panic sales. Also, given that sometimes these flash crashes can lead to a huge drop in price than immediately recover, CRYPTOFORECAST will eventually need to utilize a larger dataset as these massive fluctuations can occur over the course of ten minutes, whereas the current dataset gives daily price points.

## 4.5 Trade-offs

Describe trade-offs that you are making or may need to make, and their consequences. These can be divided into trade-offs given the duration of the term vs. those that would apply to a longer-term development period.

Short-term Trade-offs:

The most glaring trade-off for CRYPTOFORECAST is that, given the processing power of a single laptop and the shorter time given for the project, the dataset that illustrates pricing must be smaller so that all of the data can properly be processed and inform a meaningful forecast in a reasonable amount of time. If each new test run took hours to run, it would be next to impossible to make meaningful changes over a month of development. This means that the accuracy of pricing models may be smaller than what is ideal. Luckily, in the long-term, this is an easier fix provided a larger dataset can be found. If the smaller dataset is less accurate than random (<0.50), than I will need to search for larger datasets.

Additionally, given the sheer magnitude of the cryptocurrency market and ease of access to create new tokens, I have decided to use datasets that limit to larger, more popular tokens. This will, however, hurt the programs ability to determine if a token’s “moonshot” potential. A moonshot is a small market cap token that shoots up by thousands or tens of thousands of percent.

Long-term Trade-offs:

Once CRYPTOFORECAST enters into a larger, longer development lifecycle, when the dataset size increases, the program will be sacrificing efficiency of a quick decision for performance and accuracy. At times, however, the program will need to sacrifice some performance to deliver a decision in a timely manner.

## 4.5 Implementation Fragments

Show enough *significant parts* of your implementation code—or a simplified form of it—to convince the reader that you will have the implementation of the definite requirements completed on time. Precede each part with a sentence or two explaining its context within the whole application.

## 4.6 References

Add to your references. Instructions as above.

[1] https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/ [2] <https://towardsdatascience.com/lstm-time-series-forecasting-predicting-stock-prices-using-an-lstm-model-6223e9644a2f>

[3] https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory

## 4.7 Evaluation of Assignment 4

