**WGU C951**

**Task 3**

**MACHINE LEARNING PROJECT PROPOSAL:**

**Scrappy Investor’s Incorporated**

**James Revello**

**SID# 010649181**

**6/19/2024**

**A. Project Overview**

It is the mission of Scrappy Investors Incorporated (hereafter referred to as SII) to leverage the volatility of cryptocurrencies like Bitcoin and Ethereum to generate profit for its stakeholders through strategic day-trading. While the company’s main profit driver requires one to take advantage of cryptocurrency volatility, this same volatility generates substantial risk. The following proposal suggests the implementation of a recurrent neural network for predicting crypto prices based on long term patterns, in the hopes of maximizing profits and mitigating risks.

**A.1. Organizational Need**

As with any investing strategy, SII’s goal is to buy low and sell high. Unfortunately, the high reward potential of cryptocurrencies also lends itself to extremely high risks. Many traditional investors do not understand blockchain technology and see it as a risky investment to be avoided. Indeed, cryptocurrency valuations are swayed not only by due diligence research and charting, but by ever-swaying social media favor and ever-changing current events. Under stress, SII traders can be influenced by these variables. In this volatile environment, traders see prices rise and jump in head-first for fear of missing out, only to see prices drop. On another day, traders think prices have hit rock bottom and buy high volumes only to catch a falling knife and lose even more. We propose SII utilize a predictive machine learning model to forecast price movements and equip traders to make more informed decisions while minimizing losses. Even if the model is marginally better than a human trader’s best guess, it is well worth the time and money investment required to implement.

**A.2. Context and Background**

Cryptocurrencies are so volatile that fortunes can be made and lost over relatively short periods of time. This presents both great opportunity and even greater risk. Our startup investment company has largely managed to capitalize on rapid price fluctuations to maximize returns. However, high volatility environments can push even our most experienced traders into making inappropriate decisions. We require a reliable machine learning model which can predict price trends and better inform our traders.

**A.3. Outside Works Review**

To select the most appropriate artificial intelligence model for this proposal, I have reviewed three specific works which explore machine learning as it relates to cryptocurrency price prediction.

The first, “Analysis of Bitcoin Price Prediction Using Machine Learning” by Chen Junwei (2023) investigates Long Short-Term Memory (LSTM) networks and Random Forest Regression. Where Random Forest Regression increases reliability by examining multiple explanatory variables (Bitcoin features, market indices, public attention metrics, etc.), LTSM networks can capture time-related dependencies in each dataset. The study is relevant for our purposes because it narrows the focus of which models we should consider. Based on its findings, random forest regression has much more prediction accuracy than LSTM if either model is used on its own; however, using the two in conjunction makes for an even more accurate model (Junwei, 2023). For this reason, we will use both AI Models for our own project.

The second study reviewed was “Forecasting and Trading Cryptocurrencies with Machine Learning under Changing Market Conditions” by Sebastiao et al (2021). This study examined linear models, random forests, and support vector machines to predict the prices of Bitcoin, Ethereum, and Litecoin. It further explored how public recognition, social media activity, and online searches might impact cryptocurrency prices in an extremely thorough. This study provided additional models which we might consider, adding linear models and support vector machines to the docket. However, my largest take away, and the most relevant portion to our problem statement, is the study’s analysis of public recognition’s influence on pricing. Sebastiao et al. (2021) made a case for traditional market factors playing a more significant role on valuations than first thought, arguing that US money supply, gross domestic product, inflation, and interest rates were followed closely by crypto price dynamics. For this reason, we will pick only relevant columns of data to feed our model, with particular focus on traditional market factors like public perception and inflation.

The third and final study was “Bitcoin Price Prediction: A Machine Learning Sample Dimension Approach” by Sumit Ranjan, Parthajit Kayal, and Malvika Saraf (2018). This study provided an extraordinary analysis of machine learning models from logistic regression to linear discriminant analysis to random forests to support vector machines (this list is not exhaustive). The study finds, perhaps intuitively, that simple models like logistic regression and linear discriminant analysis perform reasonably well, while more complex models like random forests and XGBoost offer moderate improvements in accuracy (Ranjan et al, 2018). The takeaway, for our purposes, is that random forest regression and LSTM networks, while more complicated, also provide for moderate increases in accuracy. This is well worth the added effort in situations where other people’s money is involved.

**A.4. Solution Summary**

From what we have learned from the research garnered in the above studies, we would propose using an LSTM based model with random forest regression incorporated for additional accuracy. Based on the insights gained from these studies, the proposed machine learning solution involves developing an LSTM-based model augmented with random forest regression for enhanced prediction accuracy. The former will capture temporal dependencies in prices while the latter will capture more explanatory variables which might correlate with price and improve accuracy.

**A.5. Machine Learning Benefits**

Our machine learning model will provide traders with actionable predictions based on a wide variety of variables that they otherwise be overwhelmed with. By providing traders with tools to make informed, data driven trades, we give them a psychological anchor in an otherwise highly speculative arena. By combining LSTM with our Random Forest Regression model, we will be able to mitigate risk and capture explanatory variables such as gross domestic product and inflation, while also capturing public sentiment and general outlook. Ultimately, we will be able to mitigate risk and become adaptable in uncertain market conditions. Perhaps most of all, our machine learning model will allow our startup to make more money with less risk.

**B. Machine Learning Project Design**

**B.1. Scope**

Within our project scope, we plan to:

* Build an LSTM based model that utilizes Random Forest Regression to predict Bitcoin prices specifically. As time goes on, we will extend our model for other currencies.
* Train the model using historical Bitcoin prices (Kaggle dataset) and relevant market variables such as market sentiment, volume, etcetera.
* Deploy the model to a live trading environment so traders can make informed decisions, integrating it with our current trading platform.
* Ensure the tool has an intuitive and useable UI that doesn’t require in-depth technical knowledge to operate.

Outside of our project’s scope:

* Analyzing other cryptocurrencies such as Ethereum and Dogecoin. Our model’s first iteration will be focused on doing one thing well.
* Performing last-minute automatic trades. This model is not intended to automatically place buy and sell orders. It is strictly for on-demand prediction based on past data.

**B.2. Goals, Objectives, and Deliverables**

Goals

* Automate the analysis of historical Bitcoin data and the relationship between price and diverse explanatory variables, thereby enhancing productivity. Measure the number of automated analyses used per trading day and compare quarter to quarter.
* Lowering costs by decreasing trader tendencies for emotional or reactive decisions. Giving them faith and psychological resiliency by providing them more tools to do their job. Ensure a reduction in average losses per quarter through audits.
* Improve overall trade performance by reducing large losses and marginally increasing profits through successful trades over time. Ensure average profit per quarter is increasing through audits.

Objectives

* Increase quarterly profits by 10 percent with the implementation of the new model.
* Achieve prediction accuracy of 55-60 percent. This is a tall order, as there is no such thing as a “perfect” trader. (High prediction percentages like 95 percent are evidence of overly-fitted data and limited applicability to diverse real-world scenarios).
* Ensure the model and GUI can provide predictions/results within three seconds of prompting.
* Ensure the program maintains an uptime of 99.99 percent during trading hours. Perform maintenance on days during which the market is closed and no analysis needs to take place (Uptime considerations).

Deliverables

* The LSTM/Random Forest Regression model.
* A user friendly and technically simple graphical user interface.
* Data sets and libraries for data analysis and preparation.

**B.3. Standard Methodology**

Because we are implementing a plan with well-established business objectives, we will use the cross-industry process for data mining (CRISP-DM). CRISP-DM is more structured and therefore better suited to our purposes than more exploratory methodologies like SEMMA.

* Business Understanding: Define business objectives and requirements for our model based on SII’s needs.
  + Example: Required quarterly profit increase of 10% after model implementation.
* Data Understanding: Gather the Bitcoin dataset for all years and determine which explanatory variables will be used.
  + Example: Examine the Bitcoin Kaggle data set with associated dates and prices. Identify any trends or patterns which are readily observable.
* Data Preparation: Prepare the data for modeling.
  + Example: Process the data by checking for null values, normalizing data, separating data into training/validation/test buckets.
* Modeling: Create the model and train it.
  + Example: Build our LSTM/Random Forest model. The former will capture dependencies based on time and price. The latter will consider explanatory variables such as inflation and public sentiment.
* Evaluation: Assess our model’s performance based on our business requirements.
  + Example: Compare the model’s predictions for previous quarters as compared to our own firm’s results. Test if accurate enough for use with real stakes.
* Deployment: Deploy the model in real training environment with aforementioned-graphical user interface.
  + Example: Deploy the model. Provide training to all traders on how to use the model and how to confirm results. Setup automated monitoring of model behavior to capture odd behavior should it arise. Stress upon employees the pros and cons of the tool, with emphasis on its limitations and the trader’s responsibility to take care in trading.

**B.4. Projected Timeline**/ **Sprint Schedule**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sprint** | **Start** | **End** | **Tasks** |
| 1 | 07/01/24 | 07/07/24 | Establish stakeholders, team members, and solidify business requirements |
| 2 | 07/08/24 | 07//14/24 | Acquire relevant data and begin analysis |
| 3 | 07/15/24 | 07/28/24 | Prepare the data, normalize prices, handle null values, determine training/test sets for later evaluation |
| 4 | 07/29/24 | 08/28/24 | Build LSTM and Random Forest Regression models |
| 5 | 8/29/24 | 09/14/24 | Train models and tune |
| 6 | 09/15/24 | 09/21/24 | Evaluate the model against criteria from business requirements meeting |
| 7 | 09/22/24 | 10/14/24 | Test models in simulated trading environment |
| 8 | 10/15/24 | 10/28/24 | Integrate models into our live trading platform |
| 9 | 10/29/24 | 11/21/24 | GUI is developed with feedback from developers |
| 10 | 11/22/24 | 12/14/24 | Final product is tested and signed off |
| 11 | 12/15/24 | No true end date | Monitoring begins |

**B.5. Resources and Costs**

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost in US Dollars** |
| Computers | Computers with future-proofed hardware that can accommodate model training and predictions | 20,000 |
| Cloud Servers | IaaS Cloud infrastructure where the model will be hosted | 5,500 monthly recurring |
| Software licenses | Open-source tools will be used (Python and related libraries, Kaggle datasets) | 0.00 |
| Training | In-house training | 0.00 |
|  | **Total** | (86,000 for first year) |

**B.6. Evaluation Criteria**

Describe the criteria used to evaluate and measure the success of the completed project.

|  |  |
| --- | --- |
| **Objective** | **Success Criteria** |
| Financial | Increase quarterly profits 10% |
| Accuracy | Achieve 55-60% accuracy |
| Throughput | Achieve predictions within 3 minutes of prompt |
| Stability | Maintain uptime of 99.99 percent on trading days |
| Trader Adoption | Traders rate the GUI as user friendly |
| Risk Mitigation | Risk Managers report reduction in losses |

**C. Machine Learning Solution Design**

**C.1. Hypothesis**

My hypothesis is that our predictive machine learning model will leverage previous trading values and market indicators to forecast Bitcoin price predictions which are marginally better than the average trader at our company. Moreover, the model will make traders more resilient to market volatility and less likely to sustain unnecessary losses due to emotional response.

**C.2. Selected Algorithm**

I will use a supervised learning model with Long Short-Term Memory Networks and Random Forest Regression in conjunction.

**C.2.a Algorithm Justification**

LSTM Networks are excellent for time-related data sets because they can retain long-term dependencies in memory. Random Forest Regression can handle diverse explanatory variables such as inflation and gross domestic product and public sentiment to add resiliency to our predictions.

**C.2.a.i. Algorithm Advantage**

I am confident that both models used in conjunction will accomplish our stated business needs, as the LSTM model will capture trends over time and the Random Forest model will provide more context to otherwise highly volatile markets (the large reason for the sometimes-questionable trading decision made by traders).

**C.2.a.ii. Algorithm Limitation**

There are limitations to the models we plan to implement. LSTMs are compute-heavy and will be harder to train due to their high complexity and memory requirements. As for Random Forest, the variables we choose to incorporate at the outset of the training will determine the effectiveness of the model. Bad data in will result in bad data out.

**C.3. Tools and Environment**

For this project, we will utilize the Python language as its largest strengths lie in machine learning and data analysis. There are many pre-baked libraries which can perform both functions.

We will use Jupyter Notebooks for our environment because they allow for interactive development and provide clean visualization when working with large and clunky datasets.

Our operating system will be Windows based, as that is what is currently licensed at our organization and will mitigate costs.

We will use TensorFlow and Keras for building and training the LSTM, while using the Scikit-learn library for implementing the Random Forest Regression. When analyzing data and cleaning it for training, we will use Pandas and NumPy. We will use Matplotlib and Seaborn for data visualization. These tools will suit our limited purposes and are freely available as open Python libraries.

**C.4. Performance Measurement**

* Quantity Measured: Our performance will be measured on how much quarterly profit will be increased compared to previous quarters. Additionally, we will assess this profit increase against the costs associated with implementing the models in the first place. We will also assess efficiency increases (reduction in losses and increases in good decision-making speed). As for correctness, we will compare the accuracy of our model’s predictions compared to real world market movement. Finally, we will assess the satisfaction of traders with the tool’s performance and general useability (rating the GUI, etcetera).
* We will use Kaggle datasets for Bitcoin from inception to 2023, to include price data, trading volume, market sentiment, and general markers of economic health.
  + The training set will be comprised of 80 percent of cleaned and processed data.
  + The testing set will make up the remaining 20 percent and will be used to validate the model’s performance.
* We will measure our model’s accuracy using Mean Absolute Error (the average of absolute errors between the prediction and the testing set). We will also use Root Mean Squared Error which is the square root of the average squared differences between the predicted and actual values to make the testing more sensitive to large mistakes and discourage big losses.
* The model should predict with a Mean Absolute Error of less than five percent. This will ensure it is reliable in real life situations. Predicting prices which are *significantly* different from reality means larger margins for losing money.

**D. Description of Data Sets**

**D.1. Data Source**

The data source will be extracted from a Kaggle CSV file at the following link: <https://www.kaggle.com/datasets/jkraak/bitcoin-price-dataset>. The file provides Bitcoin price history with minute by minute changes accounted for via timestamp. It provides a timestamp, opening price, highest price, lowest price, closing price, volume of trading, number of trades, etc.

**D.2. Data Collection Method**

The data collection will simply entail downloading the CSV file directly from Kaggle, accessed and downloaded using a Kaggle Python API for automatic retrieval.

**D.2.a.i. Data Collection Method Advantage**

This dataset is accessible, comprehensive (has a minute by minute play by play of stock history) and free to acquire. Since we don’t need to painstakingly acquire our own data, we can allocate more resources to training and development.

**D.2.a.ii. Data Collection Method Limitation**

The dataset is not vetted against more official data sets from trusted brokerages. There may be errors which are impossible to root out from merely comparing. Likewise, the dataset does not provide any of the variables needed for our Random Forest Model, only the LSTM. We will need to acquire a separate data set which captures other market indicators.

**D.3. Quality and Completeness of Data**

In order to prepare our data, we will first import the data from Kaggle as explained above. Then, we will pass the data into a Data Frame and ensuring the formatting is consistent between fields. Then, we will scan for missing data, any outliers which may exist, compare random points in time with trusted sources to vet the Kaggle file for credibility, check for any null values, and so on. If there are any fields which are not relevant for our purposes, we will remove them at this juncture.

Next, we will load the now processed data into our model environment and begin training/evaluation.

**D.4. Precautions for Sensitive Data**

For the datasets we intend on using, all data is non-personal and there is no personal identifying information as it is all publicly accessible.

**References**

Chen, Junwei. (2021). Analysis of Bitcoin price prediction using machine learning. Journal of Risk and Financial Management, 14(1), 10. <https://www.mdpi.com/1911-8074/16/1/51>

Sebastião, H., Godinho, P. (2021) Forecasting and trading cryptocurrencies with machine learning under changing market conditions. Journal of Financial Innovation. <https://doi.org/10.1186/s40854-020-00217-x>

Saleem, M. N., Doumenis, Y., Katsikas, E., Izadi, J., & Koufopoulos, D. (2024). Decrypting cryptocurrencies: An exploration of the impact on financial stability. Journal of Risk and Financial Management, 17(5), 186. <https://doi.org/10.3390/jrfm17050186>