**Project Proposal**

**DSA 5593 – Section 995**

**Summer 2021**

## 1.0 Project Title

CoinPIX – A Cryptocurrency Recommender Using Data Mining

## 2.0 Project Authors

Andre Robinson, andrecrob@ou.edu

Christopher Nguyen, christopher.p.nguyen-1@ou.edu

Steven Wilson, steven.wilson@ou.edu

## 3.0 Project Objectives

The project will use data mining to recommend cryptocurrency to a user based on an individual risk profile.

- A binary classification model will be built to screen the total population of cryptocurrency coins and find the top ten coins matching an individual risk profile.
- A regression model will be constructed to predict rates of return for each cryptocurrency in the users top ten recommended coins.
- A genetic algorithm will be used to optimize a portfolio for the top ten recommended coins.

*Scope and Assumptions*

- Analysis will be limited to daily prices and assumes only long term investors with time horizons longer than one year. This removes the effect of short term trading and day trading.
- Analysis will use aggregated data from all exchanges. This removes the effect of price difference between individual exchanges.
- Analysis will be limited to the top 100 coins based on market capitalization. This will ensure sufficient liquidity to assume normal distribution of random variables.

## 4.0 Project Significance

### 4.1 Project Description

Cryptocurrencies have become a huge market for the economy and finance. As more and more consumers interact online, people looked to transact smoother, bypassing financial controllers and decentralize control [Sovbetov, Yhlas, 2018]. Cryptocurrency did not start until recent years but its growth has been exponential with relatively stable growth. By 2018, around 1500 had existed to a value of $830 billion and investors have and continue to watch the market closely [Wu K, Wheatley S, Sornette D, 2018].

Numerous researchers have used data mining to predict price movement for the largest market capitalization coins such as Bitcoin and Ethereum. However, similar research has not been performed on the smaller market capitalization coins. As a result, traditional investment strategies for equities are not reliable when applied to cryptocurrency portfolios.

This project will apply data mining to the top 100 cryptocurrency coins based on market capitalization data tracked by coinmarketcap.com. The resulting models will be used to develop an application named CoinPIX, providing a recommender engine to screen coins against an individual's risk profile and suggest an optimum portfolio of coins maximizing expected return.

**4.2 Datasets**

Yahoo finance has a ton of downloadable data. For example, BTC goes back to 2014, and has daily data on OHLCV. Each dataset can be downloaded as a csv file.The project will use historical daily price data for up to 2 years for each of the top 100 largest market cap coins. Each record of a day is 33 bytes and total is 2.9 mb. Each dataset will include date, coin symbol, open price, close price, and volume which indicate price changes and how much is being traded throughout the day. There will be a minimum of 2 years of data for each coin (730 rows x 100 coins) or 73,000 rows.

| Data Element | Description | Data Type | Size (bytes) |
|---|---|---|---|
| Symbol | Coin symbol | varchar | 10 |
| Date | Date for price data | date | 3 |
| Open | Starting price for the day | money | 8 |
| Close | End price for the day | money | 8 |
| Volume | Number of coins traded | integer | 4 |

| Date | Open | High | Low | Close | Volume |
|---|---|---|---|---|---|
| 2014-09-17 | 465.864014 | 468.174011 | 452.421997 | 457.334015 | 21056800 |
| 2014-09-18 | 456.859985 | 456.859985 | 413.104004 | 424.440002 | 34483200 |
| 2014-09-19 | 424.102997 | 427.834991 | 384.532013 | 394.79599 | 37919700 |
| 2014-09-20 | 394.673004 | 423.29599 | 389.882996 | 408.903992 | 36863600 |
| 2014-09-21 | 408.084991 | 412.425995 | 393.181 | 398.821014 | 26580100 |
| 2014-09-22 | 399.100006 | 406.915985 | 397.130005 | 402.152008 | 24127600 |
| 2014-09-23 | 402.09201 | 441.557007 | 396.196991 | 435.790985 | 45099500 |
| 2014-09-24 | 435.751007 | 436.112 | 421.131989 | 423.204987 | 30627700 |
| 2014-09-25 | 423.156006 | 423.519989 | 409.467987 | 411.574005 | 26814400 |

### 4.3 Data Mining Methods

The following data mining tasks will be performed to support functionality in the CoinPIX application.

1. Classification – A binary classifier to determine if a specific coin satisfies a user's risk profile. This will be used by CoinPIX to screen the full dataset of coins and select the top ten coins meeting a user's risk profile.
2. Regression – Construct a regression model that predicts future annual returns based on historical returns. This model will be used by CoinPIX to predict the expected return of the top ten currencies.
3. Genetic Algorithm – This will be used to construct an optimum portfolio to maximize rate of return for the top ten currencies.

### 4.3 Project Justification

Classification can be worked through support vector machines to forecast direcion of prices and match risks better. SVM's have been used to great success in analyzing stock index prices although k-nearest neighbors is also viable for our data [Wang, Hengshan, Ou, Phichhang, 2009].

For regression, there are some available that have been used to analyze Bitcoin, Etherium, etc. including linear and lasso regression. Time intervals can be plotted against predicted values to create different models. For example, one of the articles notes that Lasso operated at 98.6% accuracy and linear at 98.7% [Kathyayini R S, D G Jyothi, 2019].

Genetic algorithm has been used in conjunction with artificial neural networks to optimize feature transformations [Rather, Akhter Mohiuddin, Arun Agarwal, V.N. Sastry, 2014] but in the context of CoinPIX, the portfolio should be optimized for rate of return, running multiple times.

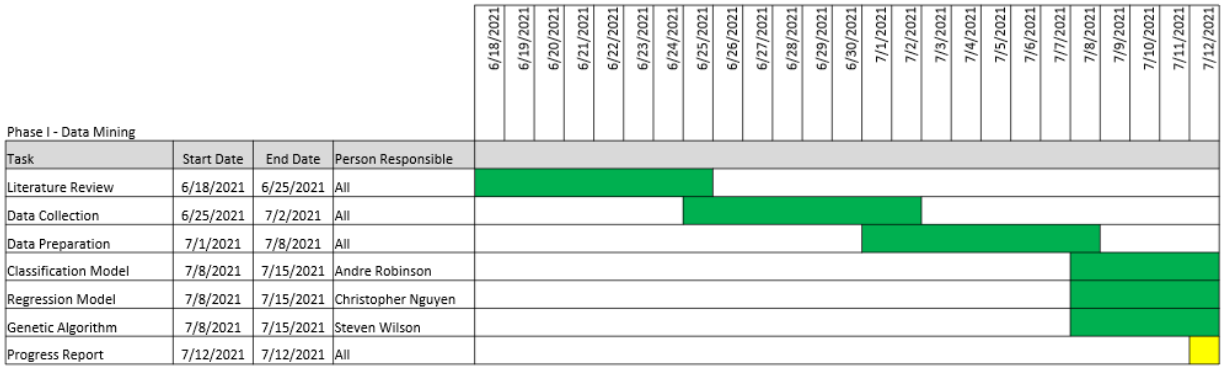
**5.0 Project Timeline and Deliverables**

The primary deliverables for the project center on the data mining models and the CoinPIX application. The data mining models have been previously described and the main components of the CoinPIX application are listed below.

- CoinPIX will be a web based application hosted in Azure.
- The user interface will be built using the Angular Javascript Framework.
- The data for the application will persist in an Azure SQL Server.
- Data access will be provided through a c# REST API .
- Data mining will be conducted in Jupyter Notebooks using Python and PySpark.
- All source code will be stored in a github repository that will be established later.

The project will be completed in two phases with the first phase focused on data mining and the second phase focused on CoinPIX development. The literature review is currently underway and will end with the submission of the project proposal. Following

the literature review, the data mining phase will take place from 6/25/2021 to 7/12/2021. The data mining phase will be near completion for the mid semester progress report. Phase II will begin following completion of the data mining phase on 7/25/2021. A high level project plan is shown in figures 1 and 2 below.




| Task | Start Date | End Date | Person Responsible |
|---|---|---|---|
| Literature Review | 6/18/2021 | 6/25/2021 | All |
| Data Collection | 6/25/2021 | 7/2/2021 | All |
| Data Preparation | 7/1/2021 | 7/8/2021 | All |
| Classification Model | 7/8/2021 | 7/15/2021 | Andre Robinson |
| Regression Model | 7/8/2021 | 7/15/2021 | Christopher Nguyen |
| Genetic Algorithm | 7/8/2021 | 7/15/2021 | Steven Wilson |
| Progress Report | 7/12/2021 | 7/12/2021 | All |

Figure 1. High Level Timeline for Phase I - Data Modeling


| Task | Start Date | End Date | Person Responsible |
|---|---|---|---|
| Classification Model | 7/8/2021 | 7/15/2021 | Andre Robinson |
| Regression Model | 7/8/2021 | 7/15/2021 | Christopher Nguyen |
| Genetic Algorithm | 7/8/2021 | 7/15/2021 | Steven Wilson |
| Progress Report | 7/12/2021 | 7/12/2021 | All |
| UI Development | 7/15/2021 | 7/25/2021 | Andre Robinson |
| Database Development | 7/15/2021 | 7/25/2021 | Christopher Nguyen |
| API Development | 7/15/2021 | 7/25/2021 | Steven Wilson |
| Final Report | 7/25/2021 | 8/4/2021 | All |
| Final Presentation | 7/26/2021 | 8/5/2021 | All |

Figure 2. High Level Timeline for Phase II - CoinPIX Development

## 6.0 References

1. Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra. Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis. SMU Data Science Review. SMU Scholar. 2018. Pg. 1-21
2. Kathyayini R S, D G Jyothi. Crypto-Currency Price Prediction using Machine Learning. International Journal of Advanced Research in Computer and Communication Engineering. IJARCCE. August 2019. Pg. 68-71
3. Liu, Yukun and Tsyvinski, Aleh. Risks and Returns of Cryptocurrency. The Review of Financial Studies. The Authors. August 2018. Pg. 2690-2727
4. Rather, Akhter Mohiuddin, Arun Agarwal, V.N. Sastry. Recurrent neural network and a hybrid model for prediction of stock returns. Expert Systems with Applications. Elsevier. December 2014. Pg. 3234-3241
5. Sovbetov, Yhlas. Factors Influencing Cryptocurrency Prices: Evidence from Bitcoin, Ethereum, Dash, Litcoin, and Monero. Journal of Economics and Financial Analysis. Tripal Publishing House. February 2018. Pg. 1-27
6. Fahmi Azim Muhammad, Noor Azah Samsudin, Aida Mustapha, Nazim Razali, Shamsul Kamal Ahmad Khalid. International Journal of Engineering & Technology. The Authors. December 2018. Pg. 1070-1073
7. Uras N, Marchesi L, Marchesi M, Tonelli R. Forecasting Bitcoin closing price series using linear regression and neural networks models. PeerJ Computer Science. Uras et al. July 2020. Pg. 1-25
8. Wang, Hengshan, Ou, Phichhang. Prediction of Stock Market Index Movement by Ten Data Mining Techniques. Modern Applied Science. Mathematical Models and Methods in Applied Sciences. December 2009, Pg. 28-42

9. Wu K, Wheatley S, Sornette D. Classification of cryptocurrency coins and tokens by the dynamics of their market capitalizations. Royal Society Open Science. The Royal Society Publshing. August 2018. Pg. 1-10

10. Yudong, Zhan and Lenan Wu. Stock market prediction of S&P 500 via combination of improved BCO approach and BP neural network. Expert Systems with Applications. Elsevier. July 2009. Pg. 8850-8854