CS 216: Everything Data
Team Members: Jimin Cheon, Changmin Shin, Sara Park, Evan Kim, Justin Yang
NetIDs: jc910, cs521, sp472, wk50, jy242

Project Final Report: Cryptocurrencies and S&P 500

Part 1: Introduction and Research Question

Introduction: Since the onset of the pandemic-induced stock market drop in March 2020 (-20% in two months), the US stock market and the cryptocurrency market have consistently shown patterns of parallel movements and increasing correlation, maintaining only positive correlation since 2020 (see image). Nasdaq 100 index and Bitcoin reached an all-time high correlation of .6945 in April 2022. Traditional markets like stocks and bonds are closed from 4 PM Friday to 9:30 AM Monday (in the US; Eastern Time), and within traditional markets, it is a widely held belief that the "Monday effect" holds unless substantial economic or policy events occur over the weekend. The Monday effect is a theory that describes market trends in which "the trading pattern of Friday will continue at the opening of trade on Monday" (TheBusinessProfessor). Furthermore, as more and more retail investors participate in the financial markets (due to reasons like stimulus check, record low interest rates, quantitative easing, etc.), it is a working theory amongst traders that price movements in crypto over the weekend have an impact on the stock market's open at 9:30AM on Monday and vice versa. Aside from the "Monday effect", other notable theories are: the extension of stocks' Monday Effect, Sunday Effect (tendency for coin prices to drop on Sundays, triggered by bigger institutions), etc. For our final project, we wanted to see if there is a relationship between crypto's weekend price movements and the difference between S&P500's Friday and Monday prices. In other words, we wanted to see if various cryptocurrency's weekend "swings" had any effect or predictive power in how the US stock market will open Monday morning in comparison to how it closed on Friday.

Our updated research question: From 4PM ET on Friday to 9:30AM ET the following Monday, the US stock market is closed while the cryptocurrency market is still open and running. Using change in price during that time frame for major cryptocurrencies (i.e. Bitcoin/BTC, Ethereum/ETH, Ripple/XRP) and setting them as independent binary variables, using a logistic regression, can we predict the probability that the S&P 500 value on Monday will be higher than its value the previous Friday? In our initial proposal, we originally planned to use Friday's stock market movement as the sole independent variable to predict Bitocin's swing over the weekend. However, in our prototype and for our final report, we decided to diversify the inputs by setting them as three independent variables – major cryptocurrencies' weekend swing directions (namely, Bitcoin, Ethereum, and Ripple) – and use them in a logistic regression to predict the Monday opening value relative to its closing value on Friday for the S&P 500, one of the most commonly recognized market index representative of the US equities market.

Relevance: As explained in our topic reintroduction, there is great interest whether cryptocurrency markets and the S&P 500 are related, as retail investors and institutional investors are all looking to find the perfect algorithm that actualizes theories such as the Monday Effect and the Sunday Effect. If we are able to find a proper pattern and predict, traders and investors will be able to use new strategies to make profit. An accurate model that predicts S&P500's Monday value can potentially change the flow of money between the two markets. If people are able to predict whether the S&P 500 will be able to open higher or lower compared to Friday's closing price with our model, we see that an exponentially more volume and investor confidence will pour into event based futures and options contracts from retail investors, which is a market that is typically taken advantage by only institutions traditionally.

Part 2: Data Sources

1. Top 100 Cryptocurrencies Historical Dataset
   - https://www.kaggle.com/datasets/kaushiksuresh147/top-10-cryptocurrencies-historical-dataset
   - This dataset consists of 100 csv files that are the top 100 cryptocurrencies based on market cap. In each file, high/low/open/close prices and daily volume (in USD) from July 2019 to August 2022 are included. The dataset was collected using web scraping and various python packages such as investpy, Yahoo Finance, and pandas data reader by the author.
   - This dataset was used to create predictor variables we need for our model. The difference between high prices on Friday and Monday for Bitcoin, Ethereum, and Ripple was calculated to determine the weekend swing direction of cryptocurrencies, and a categorical variable that takes 1 when the change was positive and 0 when negative was created as potential predictors. These variables represent the change in cryptocurrency prices during the weekend which we need to examine their effects on stock values, and create a prediction model.
2. S&P 500 Stocks (daily updated)
   - https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks?select=sp500_index.csv
   - This dataset consists of a csv file of daily S&P 500 index(there are two other csv files, but we're not using them). The dataset was collected from FRED and yfinance, and there are two columns: Date and S&P 500. There are almost 10 years worth of data starting from daily S&P 500 values from Nov 12 2012 to present.
   - S&P 500 Stocks dataset was used to create response variables for our model. We examined the difference between index value on Monday and Friday to measure the change in stock market values, and a categorical variable that takes 1 when the change was positive and 0 when negative was also created as a potential response variable. This was also used as a target data for our prediction model.

Part 3: What Modules are We Using?

- **Module 4: Data Wrangling** – We used this module to build the dataframe we needed from the stock market dataset and the cryptocurrency dataset. We did this by extracting the Friday/Sunday rows from the datasets and creating new columns for the calculated differences between Friday and Sunday and the assigned binary value for the difference. Our justification is that we need this data as our train/test data and target as we are trying to find the relationship between the increase or decrease of cryptocurrencies and the stock market values. The concepts we used are data cleaning and formatting; working between csv and pandas, and creating new columns from previous data using a data wrangling method that processes all of our data. We used this in our beginning data analysis stage and plan to use it in further analysis, to work with different cryptocurrency data.
- **Module 6: Combining Data** – We used this module to combine datasets for three different cryptocurrencies (Bitcoin/BTC, Ethereum/ETH, and Ripple/XRP). We did this by extracting the columns we needed (differences, binary values) and stacking them together into a new dataframe. Our justification is that we wanted the cryptocurrencies and the S&P500 to be in the same dataframe so we can use prediction to accurately judge if the cryptocurrencies can be used to predict the change in S&P 500 value. The concepts we used are appending and joining tables. We used this in our beginning data analysis stage and may have to use it further to add columns from other cryptocurrency datasets.
- **Module 8: Visualization** – Given that we have imported datasets from different sources, it's important to be able to visualize individual data frames and explore how the data looks like. We used this module to generate a correlation plot of each feature using matplotlib and seaborn.
- **Module 9: Prediction & Supervised Machine Learning** – Since we are predicting whether the change in S&P 500 is positive or negative over the weekend, we are tackling a binary classification problem. Thus we would be using logistic regression, 1 being positive change over the weekend and 0 being negative change over the weekend. We used this module to create a prediction model

using logistic regression, train it and examine the accuracy. We first divided our dataset into training (80%) and testing data (20%) based on a 5-fold cross validation ratio, then conducted logistic regression with 2500 iterations of the dataset. Then, we visualized our data using the confusion matrix. We utilized pandas and scikit-learn libraries throughout the process.

Part 4: Results and Methods

First, we accessed Kaggle and exported each needed data source into a csv file. Then we opened it as a dataframe in a Google Colab workspace and wrangled the data. We filtered the Friday and Sunday data from the cryptocurrencies historical dataset and the stock market dataset, for Ripple/XRP, Bitcoin/BTC, Ethereum/ETH cryptocurrencies, and the S&P 500 values. We then subtracted the difference between the Monday and Friday data for these values and created a column that would store the value 1 for increase over the weekend and the value 0 for decrease. We named the columns "[cryptocurrency]_diff" and "snp500_binary".

In training models, some algorithms demand the absence of collinearity, as if more than two certain features have a high rate of collinearity, it would not yield a significant prediction rate. Therefore, determining whether features are highly correlated is essential to avoid potential pitfalls. We conducted a correlation analysis with heatmap visualization, which highlights pairs of features with high correlation coefficient, illustrated in Figure 1.

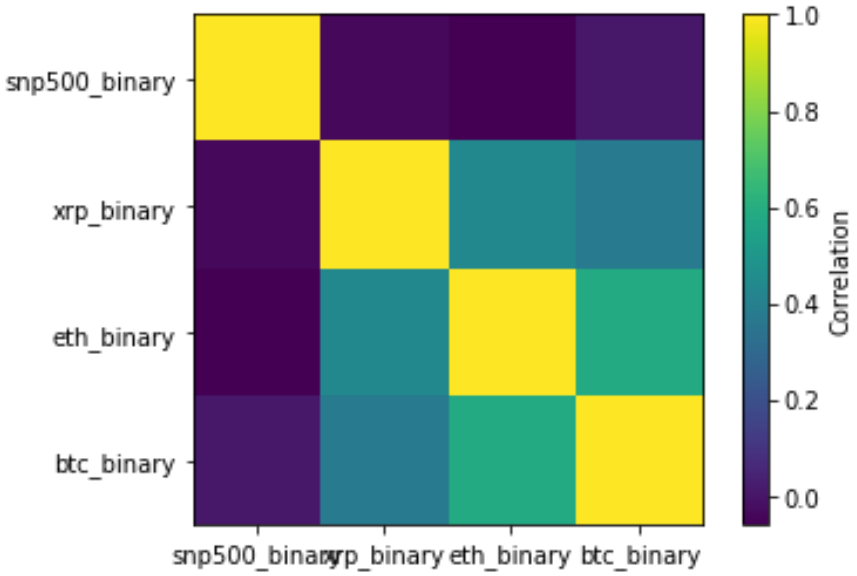

Figure 1: Correlation plot of individual features

The plot suggests that there is a very high level of correlation between the cryptocurrencies features, whereas there isn't any relationship with any of the cryptocurrencies features, which we use as the input dataset, and the stock market feature, which we use as the target value.

Necessary libraries were imported from scikit-learn and pandas. Setting the target to "snp500_binary" and the data to the cryptocurrency binaries, we first split the target and data into train and test purposes, splitting 80% as the train data and 20% as the test data based on the 5-fold cross validation principle. After creating a logistic regression classifier with iterations of 2500 and a random seed of 42, we made a prediction for the test dataset and generated a confusion matrix from the output. The initial classification results are shown in Figure 2, with a classification accuracy of 64.9%.

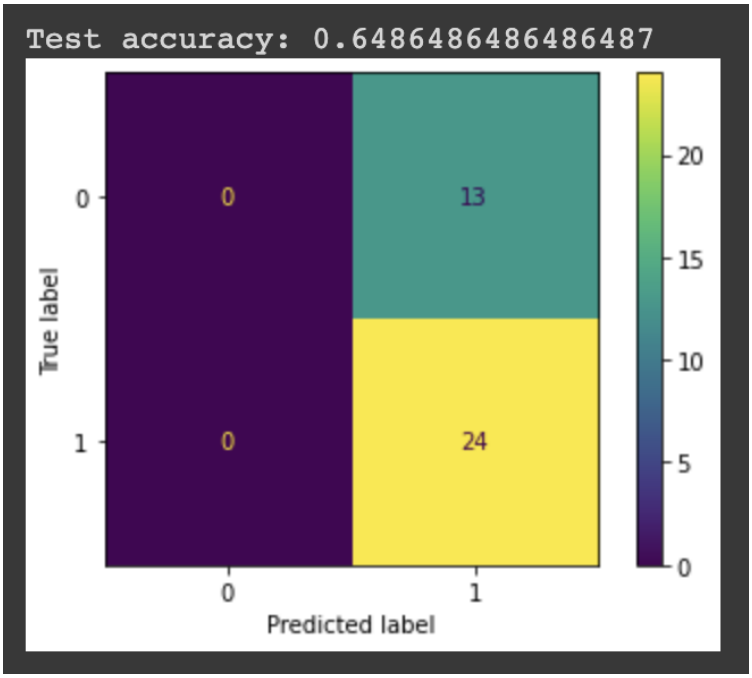

Figure 2: Initial classification results with a classification accuracy of 64.9%

To ameliorate the multicollinearity issue and eventually improve the model performance, we decided to implement principal component analysis (PCA) on the training dataset. PCA is a linear dimensionality reduction technique that transforms a set of correlated features into a smaller number of uncorrelated variables, while retaining as much of the variation in the original dataset as possible. It takes advantage of multicollinearity and combines the highly correlated variables into a set of uncorrelated variables -- eliminating high correlation between features at the end. The most important part in PCA is selecting the best number of components for a given dataset. To determine the extent of dimensionality reduction, a scree plot shown in Figure 3 has been plotted with respect to the explained variance percentages per number of components.

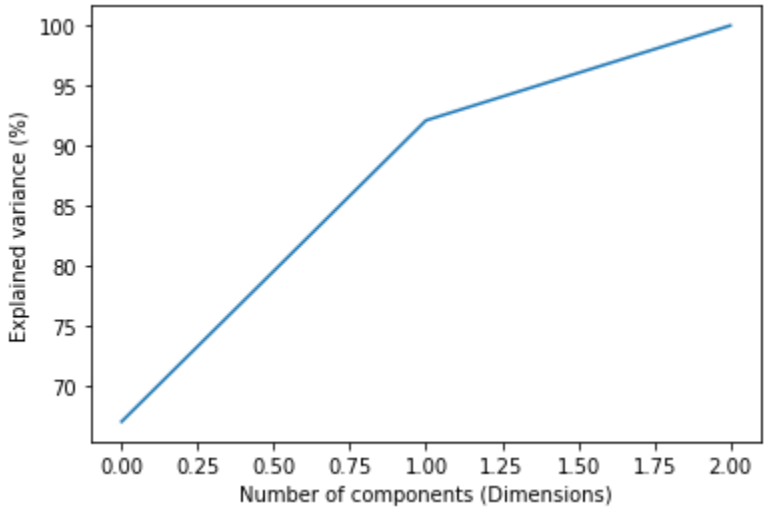

Figure 3: Scree plot for different values of dimensions in the input dataset

The baseline is applying PCA with keeping all components equal to the original number of dimensions, which is 3, and see how well PCA captures the variance of the original dataset. The first component alone captures about 65% variability. The first 2 components together capture about 95% variability in the data, and thus we will be running PCA by setting n_components to two, transforming the training dataset to one in 2 dimensions instead of 3. This changes the values in the dataset completely, and no variable in the transformed dataset is correlated with one or more of the other variables.

After running PCA on the training dataset, we retrained the same logistic regression model as the initial classifier with the same iterations and random seed value. The results are shown in Figure 4.

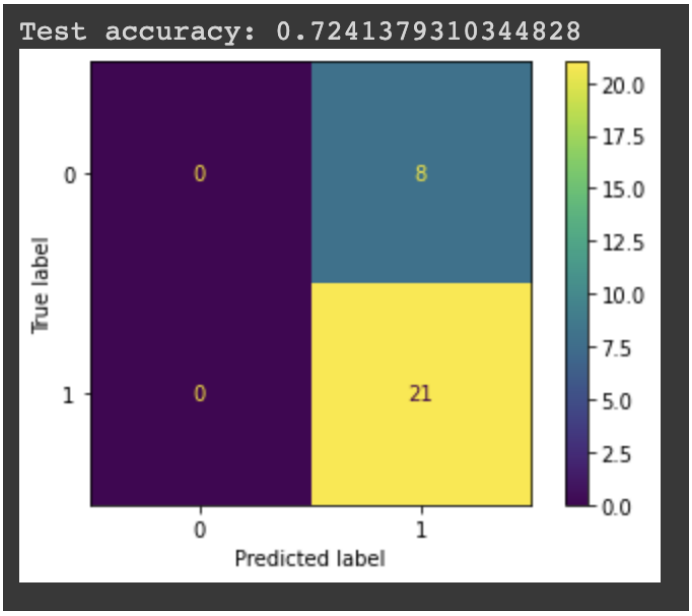

Figure 4: Classification results post-pca with a classification accuracy of 72.4%

Conducting logistic regression with a training dataset that has gone through dimension reduction yields a performance increase of 7.5%.

Here is a link to our Google Colab workspace:
https://colab.research.google.com/drive/1qinGB_veRZLr-bAckXIU_SlUCV8L6zBl?usp=sharing#scrollTo=1UTPZdLUk1A

Part 5: Limitations and Future Works

This project has its focus on predicting whether the S&P 500 price increased or decreased over the weekend by investigating the trends in certain cryptocurrencies, and attempted to best optimize the classifier using traditional machine learning techniques. After optimization, although a classification accuracy of 72.4% has been reached, there still is immense room for improvement as we identify a number of limitations in our model. First of all, because there are only 3 features available for training, there's only so much PCA could affect the performance after preprocessing the dataset. This suggests that a major issue in this learning model may have been the lack of variability in the training data. Moreover, given that most cryptocurrencies follow the price trends of BTC, a high correlation between each of the cryptocurrency is essentially inevitable. To tackle this problem, adding a couple more independent features to the training process would be necessary: to name a few, adding different types of cryptocurrencies that are seemingly independent of others, or add features other than price such as trading volume, visibility, etc.

Furthermore, using logistic regression as our primary classifier may have limited the model's performance, as multicollinearity directly affects linear or generalized linear models, which include logistic regression. When dealing with highly correlated data, decision tree models such as random forest or XGBoost are more widely used as they are by nature immune to multicollinearity. For example, when there are 2 features with high correlation, random forest would only choose only one of the features when deciding upon a split of the tree, whereas linear models would use both the features. Thus, moving forward, it may also be a good idea to train the dataset on multiple models and compare the classification results.

Additionally, cryptocurrencies have been hit with negative news that tanked its price while normal S&P 500 remained unscathed. For instance, FTX, a major cryptocurrency exchange went bankrupt which tanked the whole crypto market. It seems inevitable that due to the young age of cryptocurrency, it will be more unstable than the S&P 500 which will lead to more volatile and less accurate results.

Part 6: Conclusion

We aggregated the weekend datasets of three major cryptocurrencies, which were Bitcoin, Etereum, and Ripple and examined whether we can use the data to predict the direction of S&P value change over the weekend. Due to feature dependence on each other, initial classification results weren't ideal nor statistically significant. Conducting dimensionality reduction on the dataset to eliminate correlation marginally helped the logistic regression model at the end, however there is plenty of room for further improvement such as increasing the variability of data with more features or adding in more complex training models. We believe that our research question can provide a better predictor model with these improvements since the stock and cryptocurrency markets have been showing parallel trends for the past couple years; it is reasonable to think that with larger amounts of test and training data, we would be able to find a good predictor.