ML Final Project Proposal

By: Qianhui Dong, Daniel Mao, Rajlakshmi De,Yuming Liu

1. Problem: What is the exact business (finance) problem? What is the use scenario? What precisely is the data mining problem? Is it supervised or unsupervised?
    a. Business/Finance Problem - Our team will build algorithms, using both structured and unstructured data, to **predict whether the price of bitcoin 30 days into the future will be up or down.** The ability to predict bitcoin trend is of huge importance to cryptocurrency investors, which include banks as well as personal investors. Today, bitcoin is trading at $9,375, but the prices have fluctuated from $0.0008 all the way to $20,000 per bitcoin. Not only have the prices fluctuated greatly, but sentiment on bitcoin has also varied significantly, providing an opportunity to better incorporate unstructured text data into bitcoin models. By predicting trends, investors can improve their bitcoin trading returns. In addition to that, there is evidence that bitcoin is an uncorrelated asset with respect to stocks and bonds; leveraging bitcoin investing can be a powerful tool for diversification and improving returns per unit of risk.
    b. Use Scenario - As discussed in our course, today's trading relies on automation, which requires finding patterns in massive data. This massive data includes both structured data (financial/economic data, previous bitcoin prices, etc.) as well as unstructured data (news reports, investor calls, social media posts). One use scenario is for a bank to develop data pipelines for these various data sources, and use this data and our model to predict bitcoin trends, while retraining the model regularly as well.
    c. Data Mining Problem - There are two major components to our data mining problem. The first is to leverage unstructured social media text data, using NLP, to be able to include public opinion and sentiment of bitcoin as a feature in our model. The second large component is developing a supervised model that predicts whether or not bitcoin price will go up by mining a varied dataset that includes features from our social media text data in addition to structured financial data.
    d. Is it supervised or unsupervised? - Our data mining problem is supervised.

2. Solution: what is the solution proposed? (high level description), which forecasting algorithms do you think are appropriate for this problem domain and why?

    Will the stock of bitcoin go up in the next month?

    i. Classification - Is the price of bitcoin going to go up or down in the next 30 days? If so, what is the probability? Target class (price up from now or down). We will experiment with SVR, logistic regression, and Decision Trees, including ensemble methods
    ii. **We will not attempt to predict the price of bitcoin in the next month because regression is more difficult than classification as mentioned in the Algorithmic Trading lecture video.**

    Gathering insights from social media text data will involve supervised learning

    iii. LSTM modelings for analyzing sentiments of text data
    iv. Supervised learning to conduct Named Entity Recognition to match text data to defined entities, supervised sentiment analysis
3. Programming language.

    We are going to use python.

4. Performance evaluation: How would one test the performance of the algorithm to be used?

We will use training/testing splits as well as cross-validation when testing and tuning our algorithms. Testing accuracy will involve simple accuracy percentages, ROC curves, precision-recall, lift curves and cumulative response curves. To understand the performance with respect to data volumes, we will leverage learning curves. Another evaluation metric that we will use is the **Sharpe ratio.**

5. Data: What is the data to be used? What might be the target variable? What features would be useful?

The dataset is the daily ending price for the bitcoin market and other relevant stock markets (such as the SPY, price of oil, price of gold, price of other cryptocurrencies such as Ethereum). The target variable is the prediction of whether or not the price of Bitcoin is going up or going down in the next month.

For sentiment analysis part, twitter provides the api for downloading once the authentication credential is approved, so special Bitcoin-hashtagged tweets could be downloaded and splitted by date. Besides, for training the LSTM, the Niek Sanders' Corpus, a hand-classified tweets dataset is to be used.

Opinions about Bitcoin on social networks would have an impactful influence on holders actions, especially nowadays transactions are quick to make as long as internet is available, the shift of public attitudes could result in changes of the market within the same day. By incorporating this feature in our model, the prediction could be adaptive based on both the previous indicators and real-time common attitudes.

6. Impact: How exactly would it add business value?

The user can use this model to predict whether or not the price of bitcoin will go up in the **next month**, and use that information to **help** him/her **make a decision** to buy/sell bitcoin.

7. Names of all team members on the proposal.

Qianhui Dong, Daniel Mao, Rajlakshmi De, Yuming Liu