INTL 472 ADVANCED DATA ANALYSIS IN PYTHON
ARİF AKKAN - DENIZ AYCAN
FINAL PROJECT REPORT

1) Objective

For our final project, we decided to explore the social roots of cryptocurrency trends by incorporating machine learning models. We chose 2 currencies to examine, which are SHIBA and NEAR PROTOCOL. While Shiba is known for its volatility and fast fluctuations, Near Protocol promotes itself as being a climate-neutral currency. We were particularly interested in examining the influence of social media on currencies' price and market cap fluctuations. Both of these currencies are very popular on Twitter. Therefore, we added the tweet volume of particular hashtags attributed to these currencies. We also wanted to add Google trends data to the analysis as it is insightful to measure popularity and curiosity. Prior to the analysis, we claimed that SHIBA's higher popularity on social media would make the predictions related to SHIBA more accurate as its fluctuation trends are rooted in the social media discussions. Therefore, we will compare the results based on the accuracy of the predictions.

2) Background

There are various studies capturing the relationship between social media and cryptocurrency price fluctuations. While many academic papers cover the methods and datasets to build efficient models and provide meaningful results, there are many cryptocurrency users wanting to forecast prices accurately for financial purposes. However, most of the papers focus on either Bitcoin trends or the overall market cap of cryptocurrencies. For example, Aggarwal et al. (2019) focused on "Social Factors Affecting The Cryptocurrency Market" by conducting sentiment analysis based on the data collected from Twitter. They conducted the analysis based on overall market trends. Abraham et al. (2018), on the other hand, combined sentimental analysis with Google Trends data and focused on Bitcoin and Ethereum, the 2 currencies with the biggest market share. Their results indicate that sentiment analysis is not as effective as expected for price changes when prices decline.

There seem to be less of a focus either on the most volatile coins of the market, such as Shiba or Dogecoin, or the "purposeful" and "sustainable" coins, such as Near Protocol, SolarCoin or KLIMA. Therefore, we find it more interesting to work with currencies that have particular features.

There are several papers that use neural network models, particularly LSTM and RNN, supervised machine learning models, such as Decision Tree Regressor and time-series approaches, such as Bayesian Structural Time Series. Smuts (2019) develops the LSTM- RNN to analyze Google Trends and Telegram data and

compares the predictions regarding Bitcoin and Ethereum. He shows that Google Trends data is more successful in pointing Ethereum prices accurately, and the reverse is true for Bitcoin. Rathan et al. (2019) use the decision tree model and linear regression to predict and forecast, where they found that linear regression results outperform the accuracy score that the decision tree model provides. Lastly, Autoregressive stats model forms such as AR, MA, ARIMA, SARIMA and SARIMAX are also heavily incorporated in cryptocurrency price forecasting studies. Iqbal et al. (2021) use ARIMA, FBProphet, XG Boosting for time series analysis. They found ARIMA the most useful and efficient model in forecasting.

3) Methodology
    a) Data Collection

We have 3 sources to collect our data: Twitter, Google Trends, and CryptoCompare databases. In all domains, our analysis is limited to the time interval of Nov 10, 2021- Dec 10,2021. We picked this time interval for the relative stability of the markets since we did not identify major outside shocks causing heterogeneous fluctuations.

SNtwitter.py: We used Snscrape and Pandas modules to extract tweets that contain the desired keywords {#SHIB OR $SHIB, #NEAR OR $NEAR}in the analysis time interval. These keywords capture in nearly all cases tweets about SHIBA and NEAR coins. The number of tweets in each hour is counted. We do not perform a sentiment analysis because most tweets about cryptocurrencies do not have a direct argument. Rather, we rely on the people pushing the coin prices up and down by creating tweet volume. Therefore, we aim to measure the social dynamism and signals of cryptocurrency investors

trends.py: We used pytrends module to get Google Trends data about searches that contain related keywords {Shiba} and {Near Coin}. The keywords are decided after several checks to optimize the accuracy and the usage of people. The results are in the form of relative frequencies. Google trends capture people searching for specific coin names, so theoretically it captures a different tendency of getting information compared to twitter's communication..

Coindata.py: We mainly wanted to compare the marketcap differences on an hourly basis. However, marketcap data is not found freely on the internet separately for the currencies. Therefore, instead, we found hourly price data useful to capture the hourly trends regarding currency's popularity. We used CryptoCompare API with a user token. After we requested data from the API, we stored it in pandas dataframe and CSV.

b) Data processing

After we have collected our data for two coins, we have used pandas and NumPy modules to merge data for each coin. In the process, we have checked for NaN values and dropped unrelated columns.

Then, data is prepared for analysis. We used MinMaxScaler from the sklearn module to scale data. Since our data is volatile and do not have a normal distribution, we assumed scaling our data to a certain interval would inform models better.

Since we are using time series, we have faced the question of stationary or non-stationary data. To deal with this, we used the Augmented Dickey-Fuller method to check for stationary. Our results showed that the data is not stationary, as our null hypothesis was the data is non-stationary and the p-value is bigger than 0.05. Therefore, we cannot reject the null. However, cryptocurrencies are affected by people's perceptions, so to not limit analysis, we trained our models first with non-stationary data. Afterwards, we followed the common practice and used a differencing algorithm for turning the data into a stationary form. Our results showed that the data turned to a stationary form, as our null hypothesis was the data is non-stationary and the p-value has become smaller than 0.05. Therefore, we rejected the null hypothesis. Then, we have trained the data with stationary data.

ADF statistics are as follows:

	Pre-Difference Shiba	Post-Difference Shiba	Pre-Difference Near	Post-Difference Near
ADF Statistic	-1.747299	-14.807427	-1.749232	-28.851535
P-value	**0.406922**	**0.000000**	**0.405942**	**0.000000**
Critical Values (1%)	-3.439	-3.440	-3.439	-3.439
(5%)	-2.866	-2.866	-2.866	-2.866
(10%)	-2.569	-2.569	-2.569	-2.569

c) Model

The data that we have about price and social media is sequential. Therefore, we could not use ML models that we practised before, such as DecisionTreeClassifier, GaussianBayesianClassifier, LogisticRegression, KNearestClassifier. We chose to

explore other models that are more suitable for sequential data/time series. After investigating different approaches and examining previous attempts to work with similar datasets, we have settled on two models: MLP Regressor and SARIMAX models. By choosing these two models, we tried one Neural Network and one statistical model for time series analysis. Both of these models allow us to incorporate multivariate exogenous data as input that are sequential, as opposed to other statmodels such as AutoRegression or MovingAverage. In SARIMAX, as our data is not seasonal, we did not use this feature of the model. We also tried using VARMAX as our third model, however, we faced problems in transforming our variables into the form of VARMAX, which requires multiple dependent variables.

4) Results

We tested our claims by examining prediction graphs and R2 and RMSE (rooted mean squared error) scores. We split the results for Non-Stationary (Pre-Difference) Data and Stationary Data (Post-Difference).

Overall, our results do not provide a solid proof and suitable model for our claims. Yet, we were able to capture some trends in prices.

RMSE scores are smaller for stationary data, however, since RMSE is an absolute measure, we compare it in the same settings only. It shows that SARIMAX models, for every configuration, has less RMSE than MLP models. This may imply the predictive capabilities of SARIMAX compared to MLP.

We have observed negative R^2 values, meaning that our models are worse at predicting the prices compared to the average of the prices. They can imply a general inefficiency of our explanatory variables. Different settings may increase our prediction levels since we have low RMSE values pointing to some degree of model accuracy.

The graphs show that, for Non-stationary Shiba data, both of our models capture the movement patterns of SHIBA prices, in a consistent error term. Consequently, we predict SHIBA prices more accurately based on social media variables compared to NEAR, which is in line with our claims. However, the prediction success and R^2 need to be increased by working more on reducing error terms.

The graphs show that the hourly data for crypto coin prices get volatile waves when they become stationary. We see that our models have less prediction of deviations in prices. We also do not observe an increase in R^2 scores. Therefore, stationary data are not more informative in this case. One exception may be $NEAR with the MLP model. The graph shows it captures the trend better compared to non-stationary MLP.

Non- Stationary SHIBA results	RMSE	R2
MLP	0.089047529499478	-2.8921780922432663
SARIMAX	0.05756126939877096	-0.6263337678457792

Non- Stationary NEAR results	RMSE	R2
MLP	0.16252659658934165	-0.10927885899646173
SARIMAX	0.15698509900166932	-0.03492461495679322

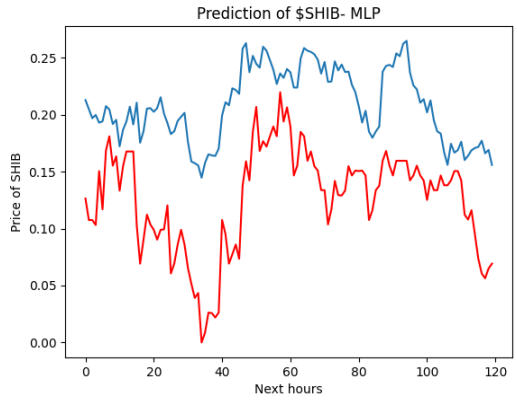
Stationary SHIBA results	RMSE	R2
MLP	0.024003123719341 92	-0.21367320891565322
SARIMAX	0.022374182710976697	-0.05453402881530911

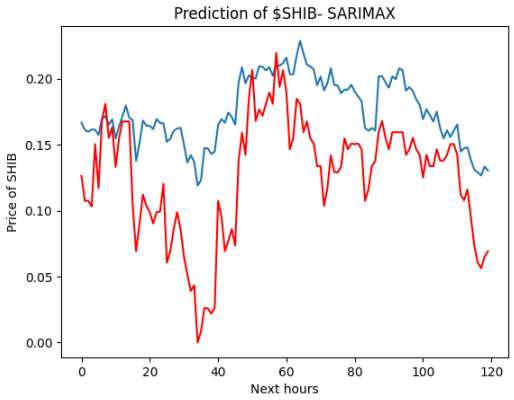
Stationary NEAR results	RMSE	R2
MLP	0.026892477829592754	-0.22802069517718904
SARIMAX	0.02485991296830637	-0.04940530435261192

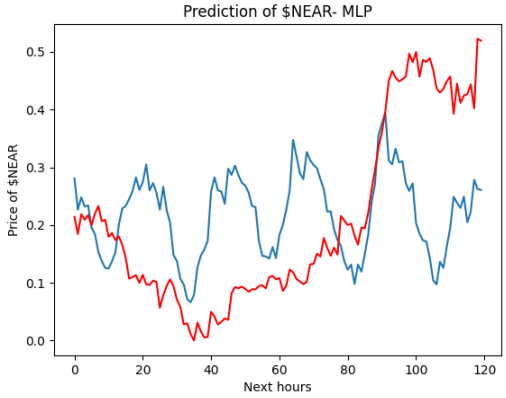
GRAPHS

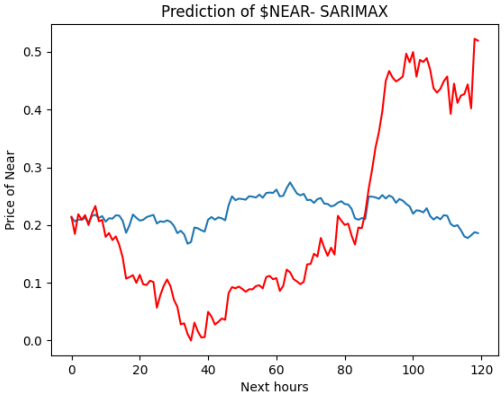
Red Line: Test Values for Y Blue Line: Prediction values for Y

Predictions with Non-Stationary Data

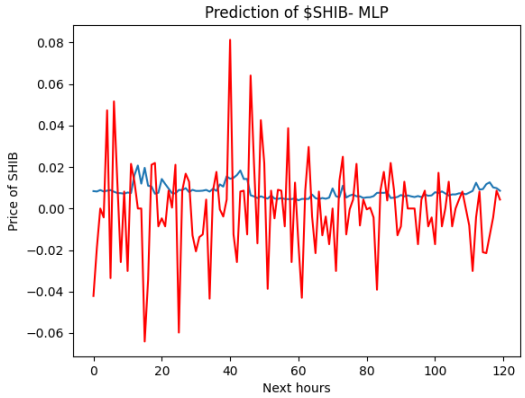


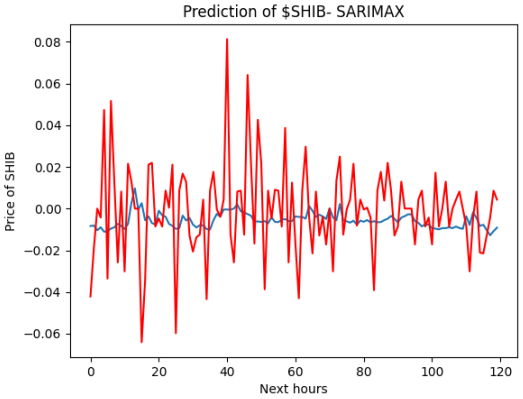


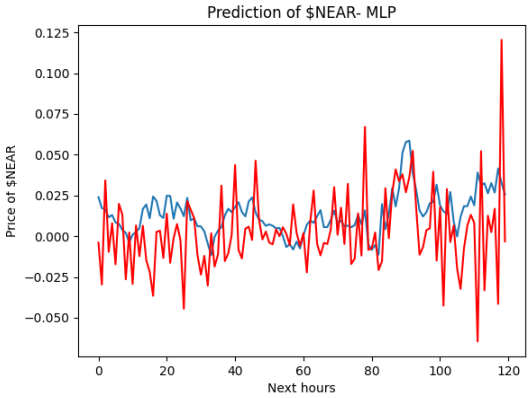


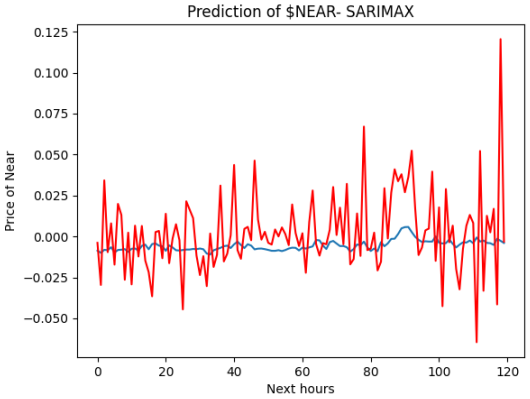


Predictions with Stationary Data









5) References

- Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan (2018) "Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis," *SMU Data Science Review*: Vol. 1 : No. 3 , Article 1.
- Aggarwal, Gourang, et al. "Understanding the social factors affecting the cryptocurrency market." *arXiv preprint arXiv:1901.06245* (2019).
- Iqbal, Mahir, et al. "Time-series prediction of cryptocurrency market using machine learning techniques." *EAI Endorsed Transactions on Creative Technologies* (2021): e4.
- Poyser, O. Exploring the dynamics of Bitcoin's price: a Bayesian structural time series approach. *Eurasian Econ Rev* **9,** 29–60 (2019). https://doi.org/10.1007/s40822-018-0108-2
- K. Rathan, S. V. Sai and T. S. Manikanta, "Crypto-Currency price prediction using Decision Tree and Regression techniques," *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2019, pp. 190-194, doi: 10.1109/ICOEI.2019.8862585.
- Nico Smuts. 2019. What Drives Cryptocurrency Prices? An Investigation of Google Trends and Telegram Sentiment. SIGMETRICS Perform. Eval. Rev. 46, 3 (December 2018), 131–134. DOI:https://doi.org/10.1145/3308897.3308955
- V. M. Hao, N. H. Huy, B. Dao, T. -T. Mai and K. Nguyen-An, "Predicting Cryptocurrency Price Movements Based on Social Media," *2019 International Conference on Advanced Computing and Applications (ACOMP)*, 2019, pp. 57-64, doi: 10.1109/ACOMP.2019.00016.