Khmer Riel Forecasting

Instructor: CHAN Sophal
Member: OUN Vikreth, MAO Kimlang, PEAN Chhinger, NANG Sreynich
Institute of Technology of Cambodia", PO Box 86, Russian Conf. Blvd. Phnom Penh, Cambodia.,

Abstract

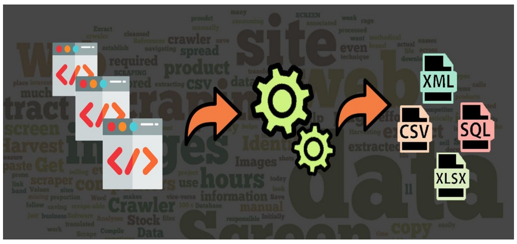
Forecasting using time series analysis presents a significant challenge. Nevertheless, several methods have been developed by mathematicians, statisticians and computer scientists to address this problem. In this paper, our team will introduce various web scraping techniques and demonstrate how insights can be derived from time series data. We will use three methodologies for forecasting: ARIMA and ETS, which are popular statistical models, and LSTM, a recurrent neural network. Despite the availability of numerous libraries for these methods, their complexity, particularly that of the LSTM model, poses a challenge. In conclusion, we will apply these methods and evaluate their accuracy to determine the most suitable approach for our project.

Keywords: Web Scraping, ARIMA, ETS, LSTM, Flask, Dockerfile

I. Our Mission

This project aims to make the future price prediction based on the 3 chosen models to fit the dataset that we scrap from the Yahoo Finance website. The Result part of this paper will tell you about the comparison between the 3 chosen models which performed very well for the given dataset that we had cleaned.

II. Web-Scraping



Web scraping is a technique used to extract and collect the data from websites. This data can be used for a variety of purposes, such as analyzing consumer behaviour, tracking market trends and improving search engine algorithms.

The best advantage of web scraping in data science is the ability to collect large amounts of data quickly and easily. With web scraping, data scientists can gather data from websites in an automated and efficient manner, saving time and effort compared to manually collecting data.

A. Methodology of Scraping

The method that we used to scrap or retrieve financial data is using Python. There are a lot of websites that provide historical data of Riel exchange, but we chose Yahoo Finance for scraping. The first step involves importing the necessary libraries such as pandas for data manipulation and analysis, BeautifulSoup for parsing HTML and XML documents, and urlopen and Request for sending HTTP requests and handling responses.We will discuss one by one on our code what are methodologies or technologies that we have used. The next step is to define a list of Unix timestamps that represent the start and end dates for the data to be retrieved. Unix timestamps are a way to represent time as the number of seconds that have elapsed since January 1, 1970. An empty list is then initialized to store the data and headers for the HTTP request are defined. These headers include a User-Agent string that identifies the client making the request. The code then enters a loop that iterates through the date ranges defined by the Unix timestamps. For each date range, a URL is constructed that specifies the parameters for retrieving data from Yahoo Finance. An HTTP request is sent to the constructed URL with the defined headers using urlopen and Request. The response is read and stored in a variable. The response is parsed using BeautifulSoup to extract the relevant data from the HTML table on the Yahoo Finance page. This involves finding the table element with a specific class attribute and iterating through its rows to extract the data. The extracted data is appended to the initialized list as a list of values. After all data has been retrieved, a Pandas DataFrame is created with the data and appropriate column names. The 'Date' column is converted to a datetime object using pd.to datetime() and the DataFrame is sorted by date using sort values(). Finally, the DataFrame is saved to a CSV file using csv(). This code provides an efficient way to retrieve financial data from Yahoo Finance for further analysis.

B. Data Collection

Web scraping is an automated process of collecting and parsing raw data from the website which contains the data. It collects and converts unstructured data in hypertext markup language which in short is HTML format into structured data and also be stored as csv file and many more. In our project, we decided to store our dataset as a CSV file that included several features which are date, open, high, low, close, adj close. We chose Yahoo Finance website in order to scrape and collect the dataset. The reason we chose this website was because it is a popular website in the world and it provided all that we needed. We also try to explore other websites as well, but we see that all the dataset of those websites are the same as Yahoo Finance. There were 2 times that we all tried to scrape the dataset. Firstly, we collected only around 100 pieces of data. The reason was because we didn't find a trick of this website in order to scrape those dataset. The last time of our scraping is when we found that in order to get the data all year with every month and all days, we had to change the integer values of each year of each website. Finally, we succeed in scraping all that data for all each day. The dataset that we got is around 5000.

III. Data Processing

A. Characteristics of Historical Dataset

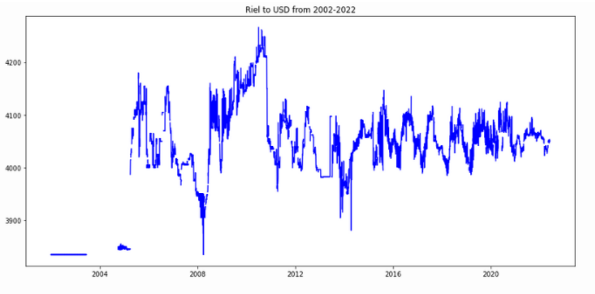


Figure1: Before Cleaned

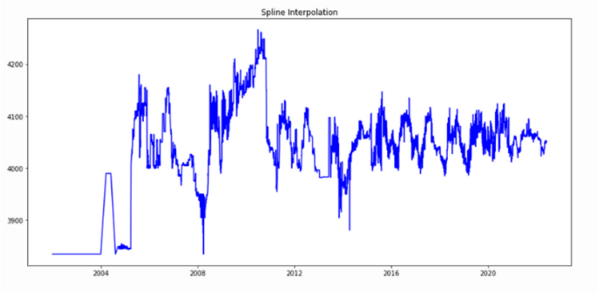


Figure2: After Cleaned

B. Step to Cleaned the Dataset

Even though we got a ton of data from scraping, there are a lot of missing values. Firstly, we used Excel to fill in missing data that is very important for time series data. Then another problem is missing Price data. There are several methods to fill the missing values. In this case our team decided to choose the Interpolation method. Interpolation method is a powerful method to fill in missing values in time-series data because it provides a smoothed response based on the characteristics of the surrounding data and the known structure of the errors. It is mostly used while working with time-series data because, in time-series data, we like to fill missing values with the previous one or two values. For example, suppose stock price, now we would always prefer to fill today's price with the mean of the last 2 days, not with the mean of the month.

IV. ARIMA Model

A. Introduction

ARIMA model stands for Autoregressive Integrated Moving Average. Two models are Auto Regression(AR) and Moving Average(MA). One method is differencing(I). These three work together when the time series we use is non-stationary. In simple words, we can call a model ARIMA model if we apply differencing (I) at least once to make the data stationary and combine autoregressive and moving averages to make some forecasting based on old time-series data. The equation of this model can be explained by the following expressions:

$$y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} \epsilon_t + \Phi_1 \epsilon_{t-1} + \Phi_2 \epsilon_{t-2} + \ldots + \Phi_q \epsilon_{t-q}$$

Prediction = constant + linear combination lags of Y + linear combination of lagged forecast errors.

To make a better explanation of ARIMA we can also write it as (AR, I, MA) and by this, we can assume that in the ARIMA, p is AR, d is I and q is MA. here our assumption is right. These parameters can be explained as follows

- p is the number of autoregressive terms.
- d is the number of nonseasonal differences.
- q is the number of lagged forecast errors in the prediction equation. Hence, the value of ARIMA(p,d,q) was defined by code python in order to choose the best model to fit the data that we have cleaned. The result of fitting the model is ARIMA(3,1,3).

B. Result

The ARIMA model's mean square error(MSE) value of 2.8995 signifies that, on the average, the model's predictions deviate by approximately from the actual values. On the other hand, the R-square error(RSE) value of 0.89 means that approximately 89% of this time series data can be explained by the ARIMA model.

The prediction of Khmer(Riel) Prediction Price in the next 10 days based on the following date which is the last date is 2022-06-06.

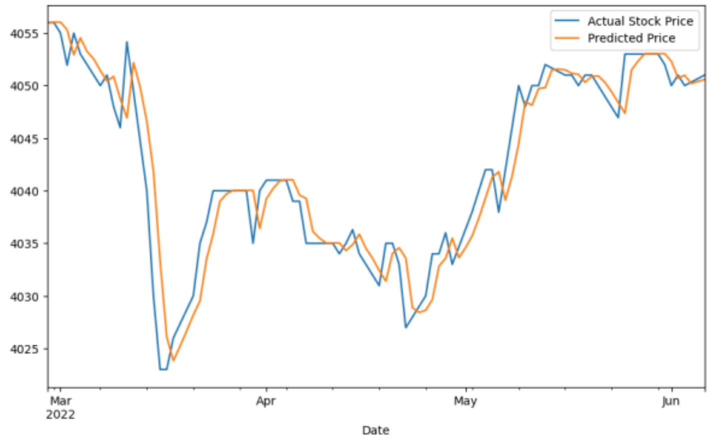


Figure3: ARIMA Model

Predicted Value	
2022-06-06	4050.577506
2022-06-07	4050.910896
2022-06-08	4050.937570
2022-06-09	4050.980524
2022-06-10	4051.005857
2022-06-11	4051.034169
2022-06-12	4051.063668
2022-06-13	4051.092503
2022-06-14	4051.121385
2022-06-15	4051.150329
2022-06-16	4051.179251

V. ETS Model

A. Introduction

Exponential smoothing is a broadly accurate forecasting method for short-term forecasts. The technique assigns larger weights to more recent observations while assigning exponentially decreasing weights as the observations get increasingly distant. This method produces slightly unreliable long-term forecasts. There are 3 types of ES models.

- Simple or Single Exponential Smoothing
- Double Exponential Smoothing
- Triple Exponential Smoothing

In triple exponential smoothing, seasonality can be multiplicative or additive. Multiplicative seasonality has a pattern where the magnitude increases when the data increase. Additive seasonality reflects a seasonal pattern that has a constant scale even as the observations change.

- *Multiplicative seasonality:* The seasonal component is multiplied by the trend component. This means that the seasonal fluctuations are proportional to the level of the series. In short, we can say that Triple Exponential Smoothing has a linear seasonality. For example, if the series is increasing, the seasonal fluctuations will also increase.

- *Additive seasonality:* The seasonal component is added to the trend component. This means that the seasonal fluctuations are independent of the level of the series. In short, we can say that Triple Exponential Smoothing has an exponential seasonality. For example, the series may be increasing, but the seasonal fluctuations may be constant.

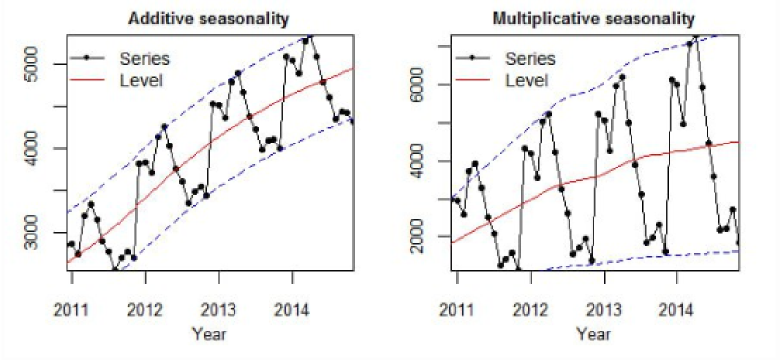


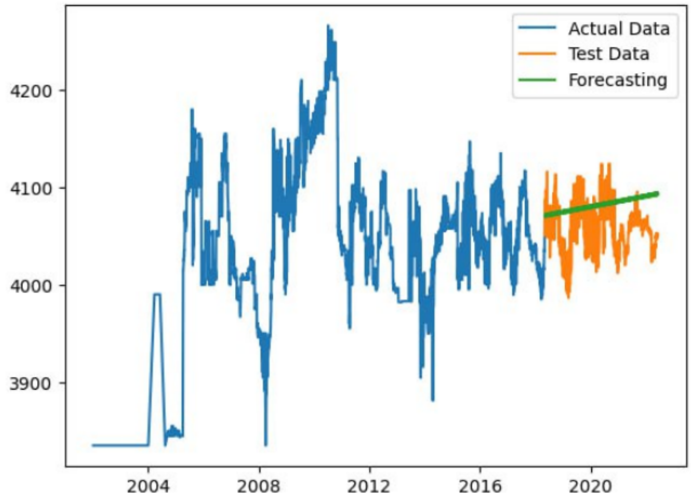
Figure4: Additive Seasonality vs Multiplicative Seasonality

B. Result

A mean squared error (MSE 1200.3512) for Khmer Prediction Price is relatively high, so it is not very accurate. Therefore the Khmer Prediction Price is a very volatile market, and it can be difficult to predict the future prices. Moreover an R-squared value of -1.3211 suggests that the forecast is actually worse than random guessing. This is likely due to the fact that the Khmer prediction price is a very volatile market, and it is difficult to predict future prices with any accuracy.

- ARIMA mean square error = 2.8953142610314884
- ARIMA R-square error = 0.891858652421149

The prediction of the last 10 days using the ETS model.



Predicted Value
2022-06-07 4110.112999
2022-06-08 4109.687805
2022-06-09 4110.352751
2022-06-10 4110.030020
2022-06-11 4109.456395
2022-06-12 4109.238843
2022-06-13 4109.276502
2022-06-14 4111.527079
2022-06-15 4110.031802
2022-06-16 4109.669123

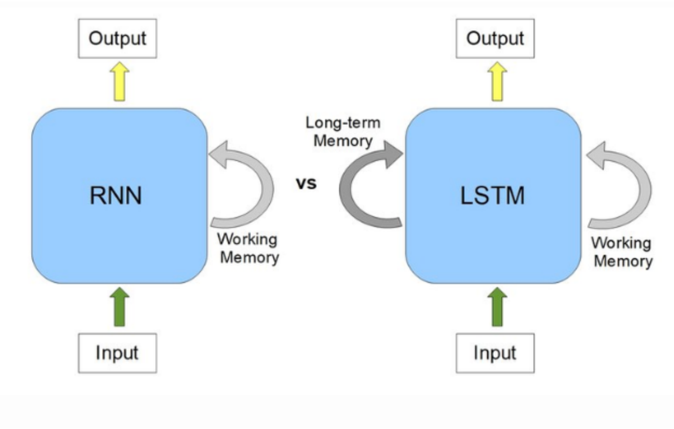
Figure5: ETS Model

VI. LSTM Model

A. Introduction

LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) architecture that is widely used in machine learning to analyse and make predictions based on time series data. LSTMs are designed to overcome the limitations of traditional RNNs, which struggle to remember long-term dependencies in sequential data due to the vanishing gradient problem.

B. Processing of LSTM Model



The key difference between an LSTM and a traditional RNN is the addition of memory cells, which allow the model to selectively retain or forget information over time. These memory cells are controlled by gates, which are like filters that decide how much information enters and exits the cell at each time step.

Figure6: RNN vs LSTM

An LSTM model typically consists of three types of gates: input gates, forget gates, and output gates.

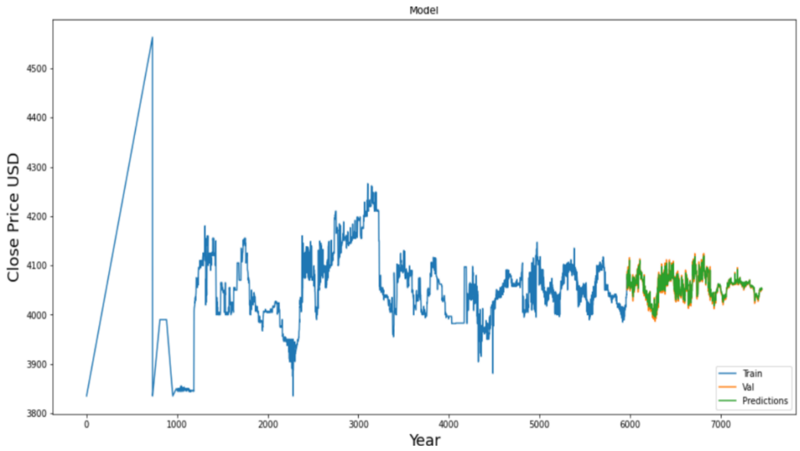
The input gate determines how much new information should be added to the memory cell, while the forget gate decides how much old information should be discarded. Finally, the output gate controls how much information from the cell should be passed on to the next layer in the network.

When applying an LSTM model to time series data, the input data is typically formatted as a sequence of values over time. For example, if we were trying to predict the price of a stock, we might use historical price data as input to the model, with each time step representing a single trading day.

To train the LSTM model, we would provide it with a set of input sequences and corresponding output sequences (i.e., the values we want the model to predict). The model would then adjust its parameters using backpropagation through time, which involves computing the gradients of the loss function with respect to the model's weights at each time step.

Once the model has been trained, we can use it to make predictions on new data by feeding in a sequence of input values and letting the model generate a sequence of predicted output values. This can be useful for a wide range of applications, such as weather forecasting, stock price prediction, and speech recognition.

C. Result



Predicted Value
2022-06-06 4052.697
2022-06-07 4054.064
2022-06-08 4055.2515
2022-06-09 4056.296
2022-06-10 4057.248
2022-06-11 4058.1445
2022-06-12 4059.003
2022-06-13 4059.830
2022-06-14 4060.6272
2022-06-15 4060.6272

Figure7: LSTM Model

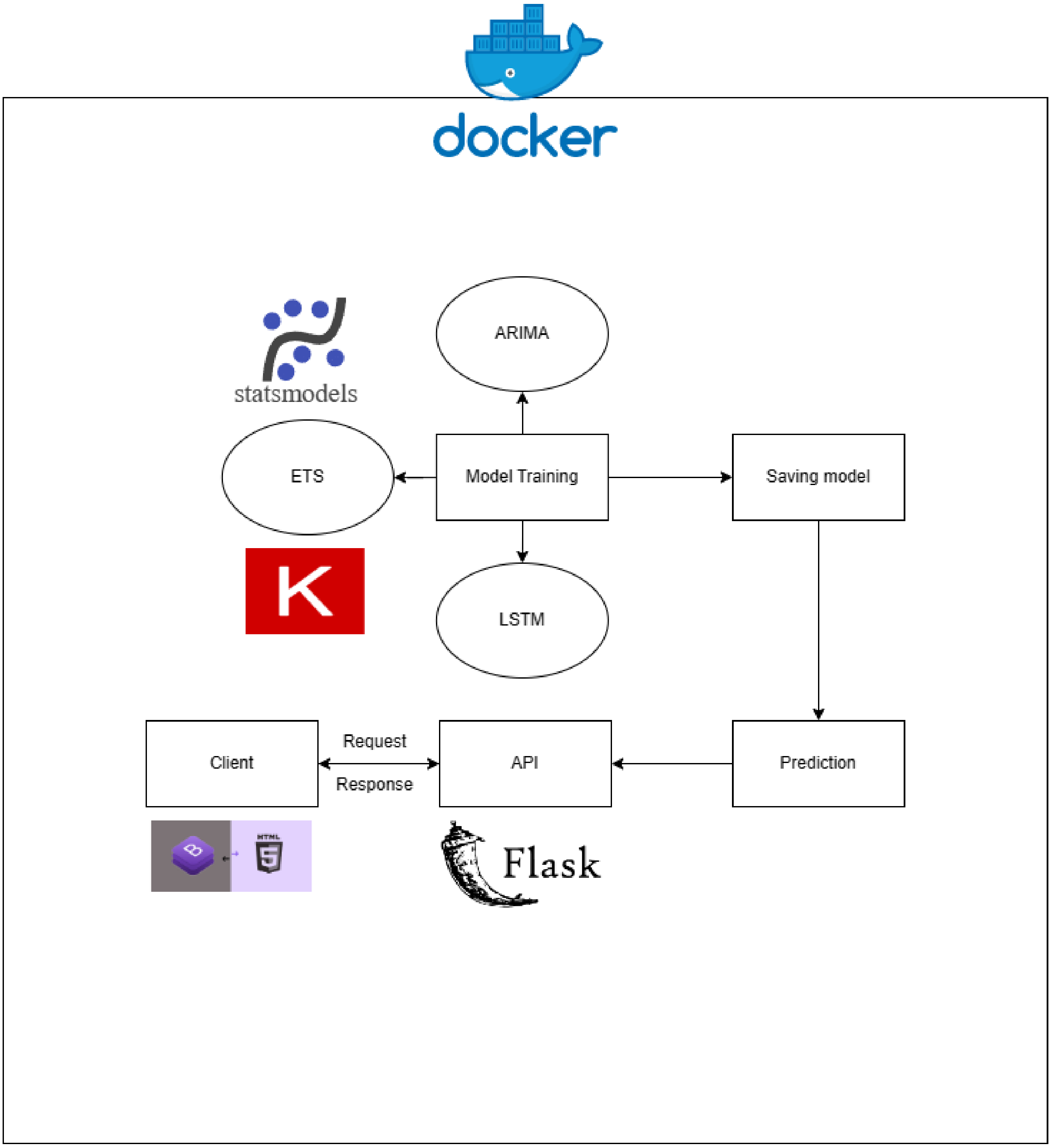
Lastly, the root mean squared error is 10.064. In summary, an RMSE of 10.064 may not be considered ‘good enough’ for the stock price prediction in some contexts, but it’s essential to consider the specific domain and problem when evaluating model performance.

VII. Flask

Flask is a Python web framework that allows for the quick and easy development of web applications and APIs. It provides a lightweight and flexible approach to handling HTTP requests and responses, making it ideal for integrating machine learning models into client applications. With Flask, we can create routes that define the URLs our application will respond to, and use decorators to associate these routes with functions that execute when accessed. Flask also provides request and response objects for accessing client data and sending responses. Integrating machine learning models involves loading the model into memory and using it to process input data and generate predictions. Error handling and validation are important aspects, and Flask offers mechanisms for catching exceptions and validating user input. When deploying a Flask application, you have various options, such as running it locally or deploying it to a production server using hosting platforms like Heroku or AWS. Overall, Flask is a powerful framework for connecting machine learning models to web applications, offering simplicity, efficiency, and flexibility.

VIII. Docker

Docker is a platform designed to help developers build, share, and run modern applications. We handle the tedious setup, so you can focus on the code. Running an ML model on the computer is an easy task. But when we want to use that model at the production stage in other systems, it's a complex task. Docker makes this task easier, faster, and more reliable. It also helps to install python modules and other dependencies that our project needs to run. In short, Docker is a productivity platform that could make the deployment much more easy especially, for ML model deployment that have api, client request or response that we have to execute a lot of file but, wit docker we just do one click and it is done and it also makes sure the project is runnable in all platform or other pc.


docker
statsmodels
ARIMA
ETS
Model Training
Saving model
K
LSTM
Client
Request
Response
API
Prediction
HTML 5
Flask