SC1015 Mini-Project

Predicting Bitcoin Prices

Lab Group A139 Team 3
Juannisa Putri Sunarya - U2220227K
Marcelino Sutanto - U2222222F

Contents

I. Introduction

Cryptocurrency and its Rise and Fall

I. Introduction

Project Objective and its Significance

Objective

To better predict the price of Bitcoin based on its historical prices from 2016 to 2022

Significance

Accurately predicting its price is crucial for individuals to make informed decisions

II. Data Preprocessing

Raw Data

	Unnamed: 0	open	high	low	close	volume	marketCap	timestamp	crypto_name	date
0	0	112.900002	118.800003	107.142998	115.910004	0.0	1.288693e+09	2013-05-05T23:59:59.999Z	Bitcoin	2013-05-05
1	1	3.493130	3.692460	3.346060	3.590890	0.0	6.229819e+07	2013-05-05T23:59:59.999Z	Litecoin	2013-05-05
2	2	115.980003	124.663002	106.639999	112.300003	0.0	1.249023e+09	2013-05-06T23:59:59.999Z	Bitcoin	2013-05-06
3	3	3.594220	3.781020	3.116020	3.371250	0.0	5.859436e+07	2013-05-06T23:59:59.999Z	Litecoin	2013-05-06
4	4	112.250000	113.444000	97.699997	111.500000	0.0	1.240594e+09	2013-05-07T23:59:59.999Z	Bitcoin	2013-05-07

Cleaned Data

date	open	high	low	close	volume	marketCap	date	Tomorrow	Target
2016-01-01	430.721008	436.246002	427.515015	434.334015	36278900.0	6.529300e+09	2016-01-01	433.437988	0
2016-01-02	434.622009	436.062012	431.869995	433.437988	30096600.0	6.517390e+09	2016-01-02	430.010986	0
2016-01-03	433.578003	433.743011	424.705994	430.010986	39633800.0	6.467430e+09	2016-01-03	433.091003	1
2016-01-04	430.061005	434.516998	429.084015	433.091003	38477500.0	6.515713e+09	2016-01-04	431.959991	0
2016-01-05	433.069000	434.182007	429.675995	431.959991	34522600.0	6.500393e+09	2016-01-05	429.105011	0

II. Data Preprocessing

Cleaned Data

Drop the first 629 data

Filter newest data from 2016 to 2022

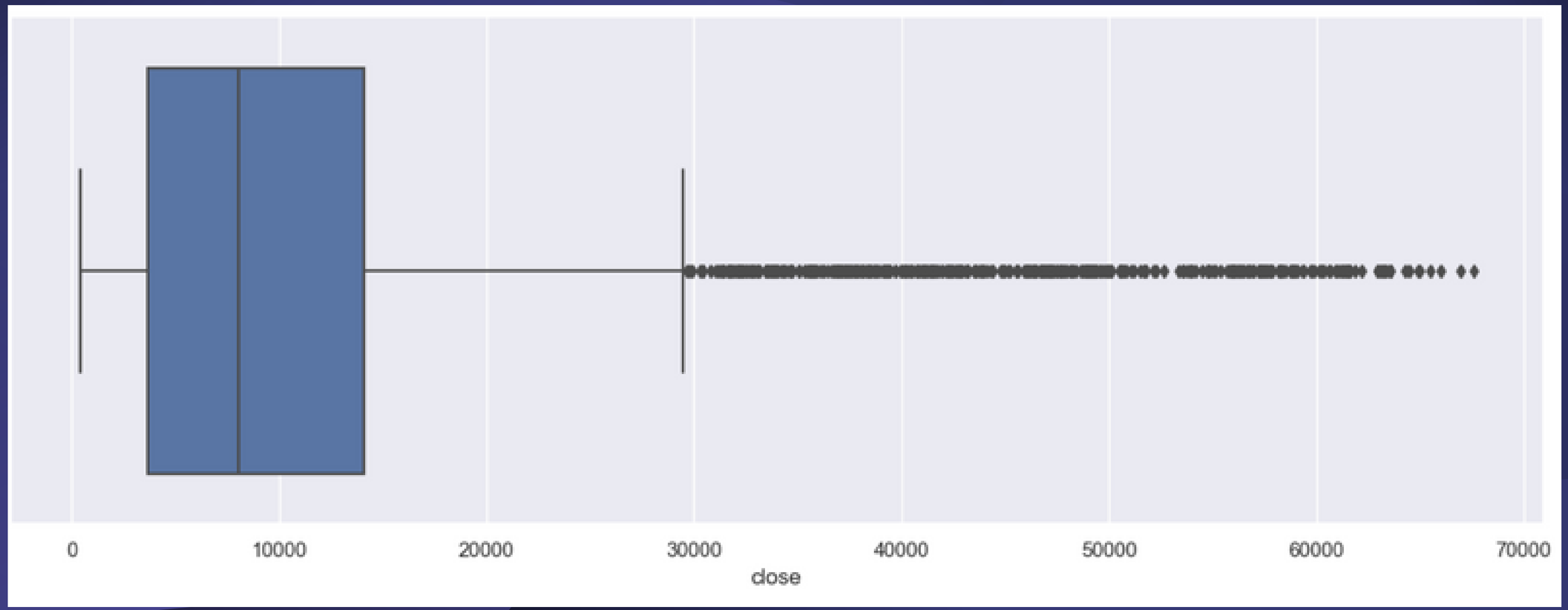
Extract only Bitcoin data

Exclude unnecessary columns

Create new variables

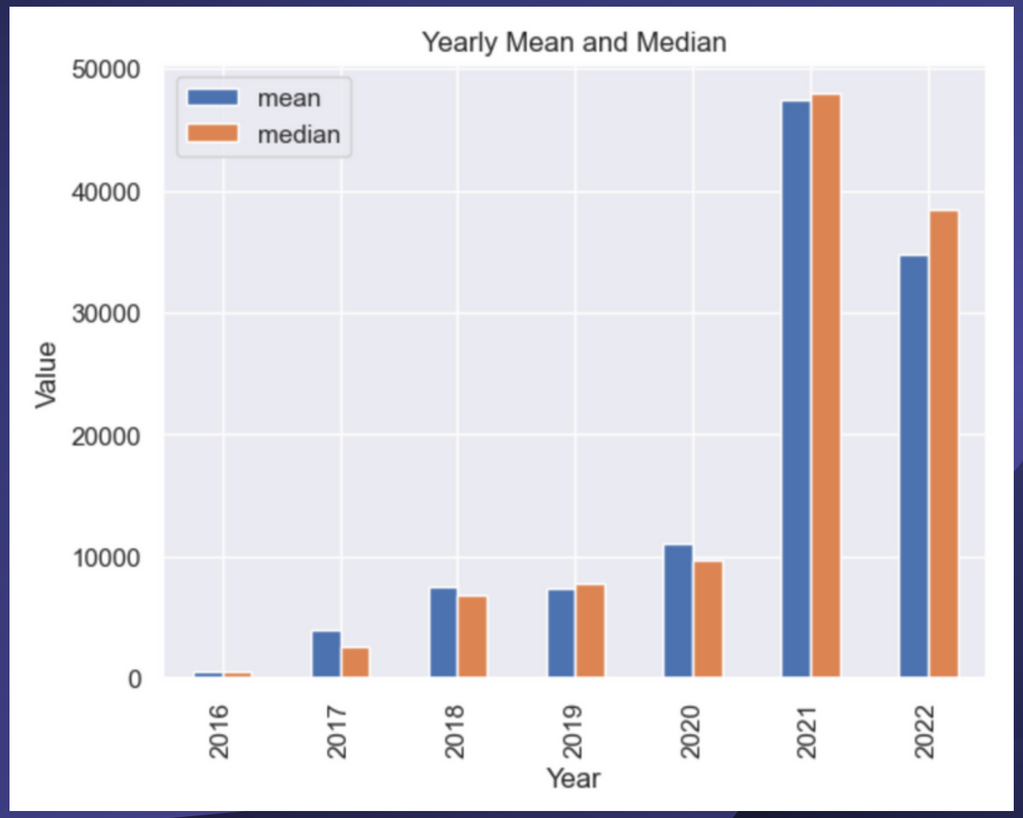
Data Visualisation

Boxplot

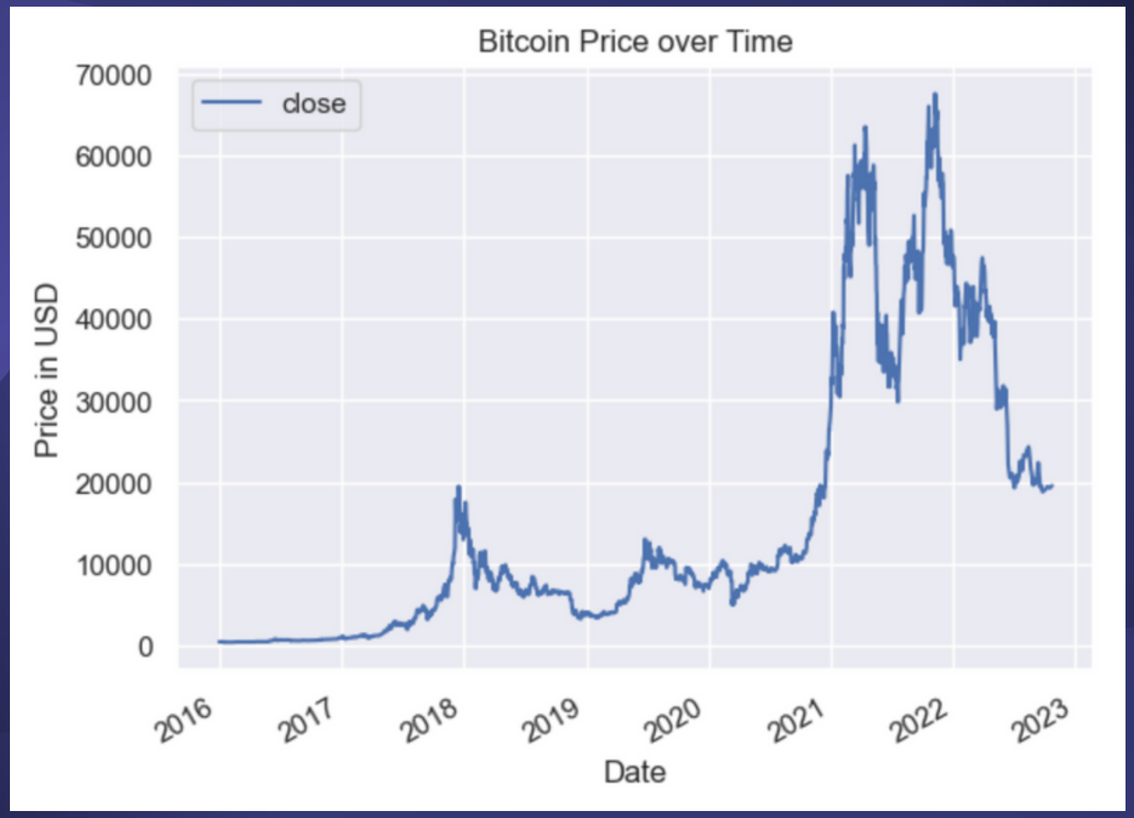


II. Data Preprocessing

Data Visualisation



Mean and median

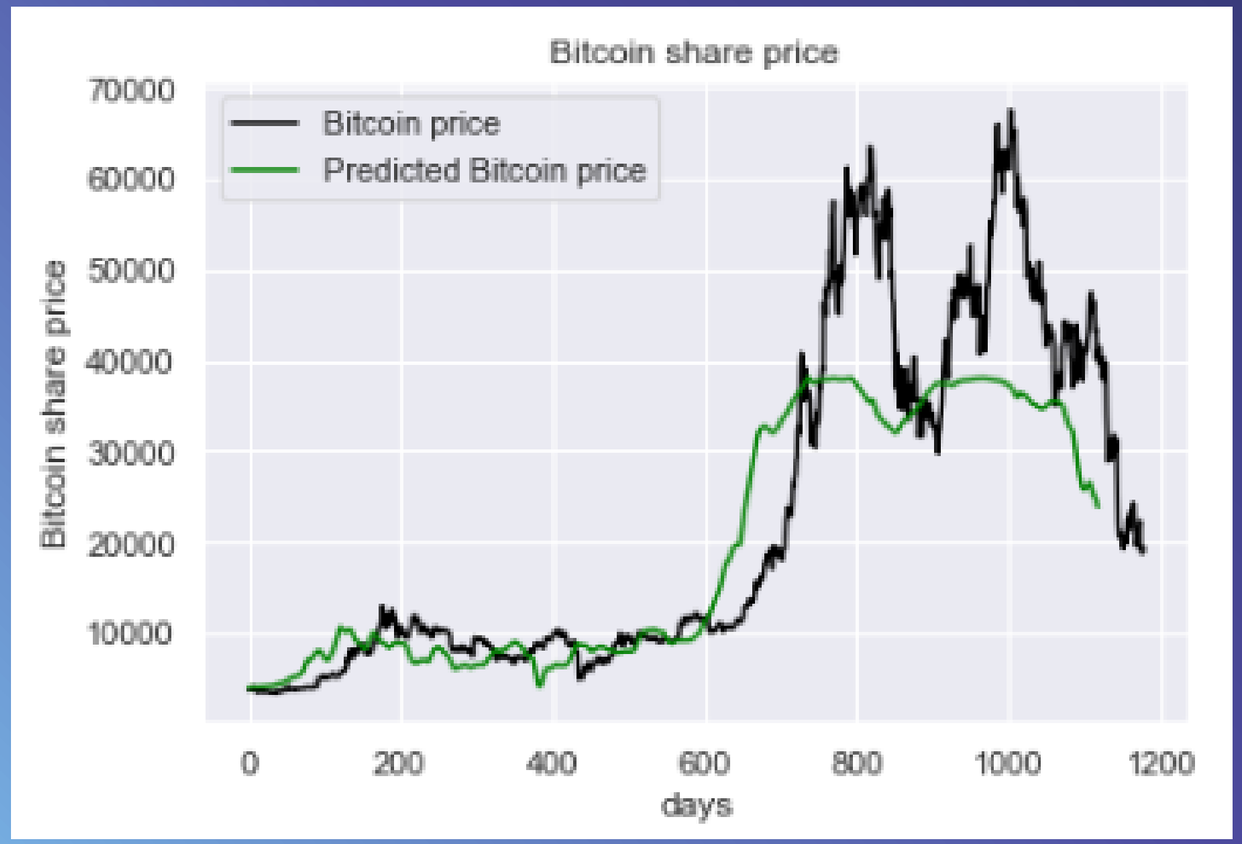


Line Plot

LSTM Model

- A variety of recurrent neural networks that are capable of learning long-term dependencies
- Hyperparameters used:
  - Number of LSTM units
  - Dropout layers
  - Number of epochs
  - Batch size
- Three LSTM layers with 50 units each, batch normalization to normalize the activations of the previous layer, and dropout to prevent overfitting.

LSTM Model




```python
# Generate price predictions using the LSTM model
predicted_prices = model.predict(x_test)

# Invert the scaling applied to the predicted prices to obtain actual values
predicted_prices = scaler.inverse_transform(predicted_prices)

# Split the test data into input and output variables
x_test = x_test.reshape((x_test.shape[0], x_test.shape[1]))
actual_prices = test_data['close'][prediction_days:].values.reshape(-1,1)

# Calculate mean squared error and explained variance score
mse = mean_squared_error(actual_prices, predicted_prices)
evs = explained_variance_score(actual_prices, predicted_prices)

print(f"Mean Squared Error: {mse:.4f}")
print(f"Explained Variance Score: {evs:.4f}")
```

```
35/35 [==============================] - 1s 21ms/step
Mean Squared Error: 64509829.6241
Explained Variance Score: 0.8612
```

Models Used

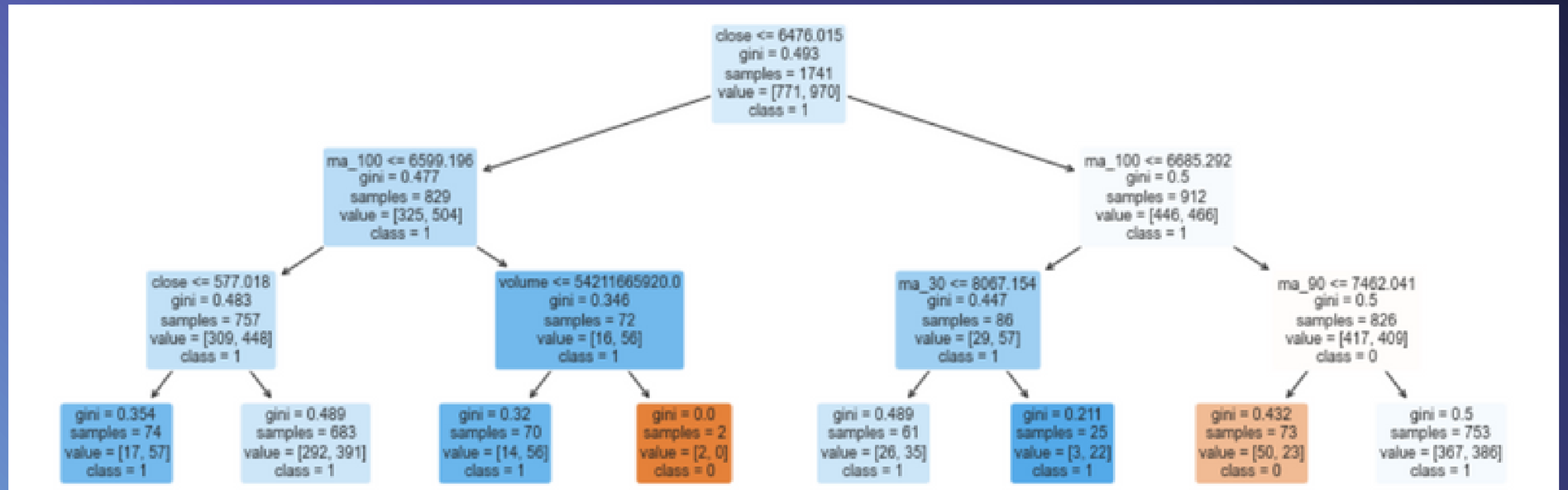
Linear Regression Model

- A statistical method used to model the **relationship** between a dependent variable and an independent variable.
- To find the line of best fit that represents the relationship between the variables, allowing for the **prediction** of the dependent variable given values for the independent variable.

Random Forest Classifier

- An ensemble learning technique that builds multiple decision trees and **combines their predictions** to improve the overall accuracy and stability of the model.
- Each tree casts a vote for the predicted class, and the class with the most votes is chosen as the final prediction.

Random Forest Classifier

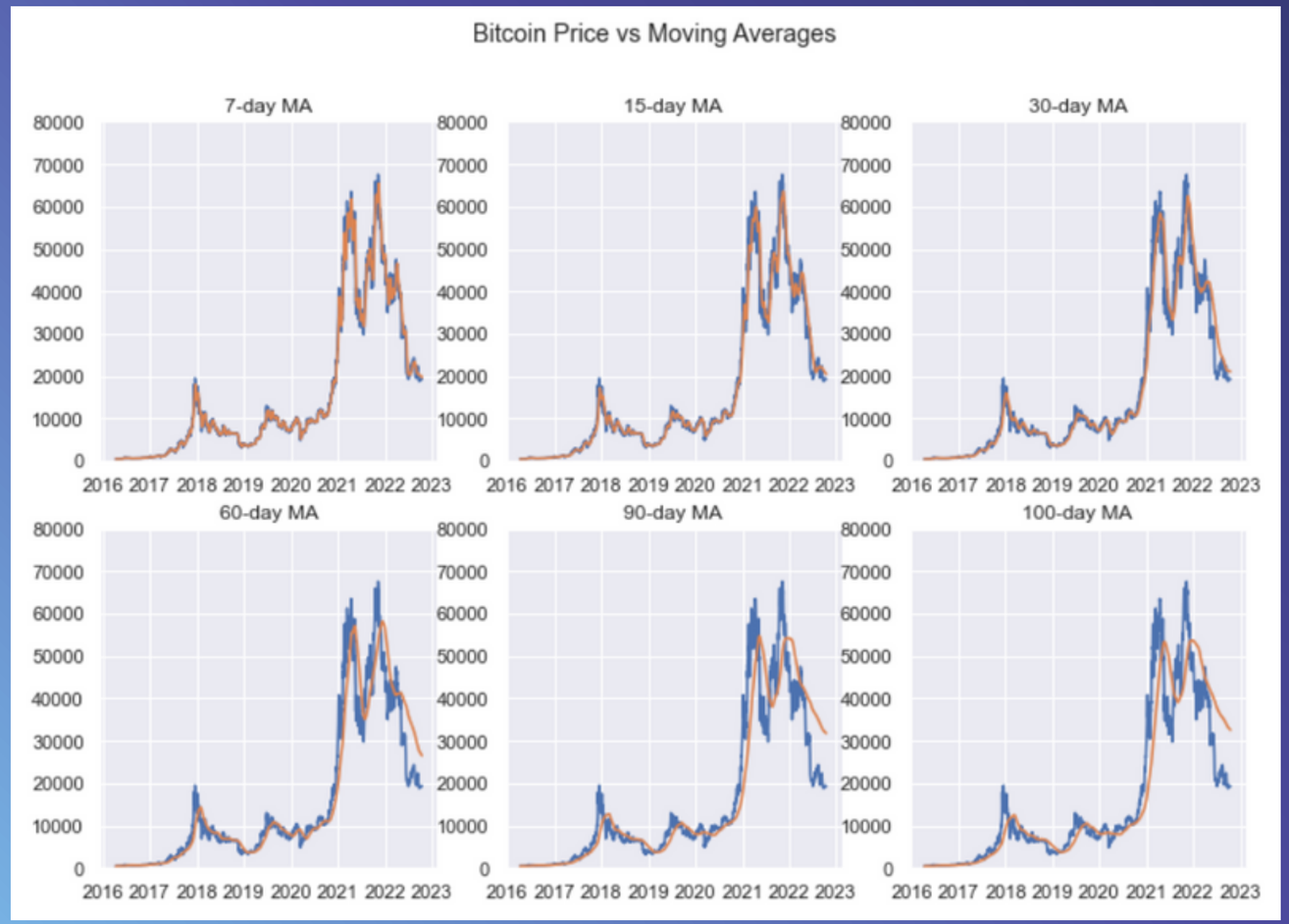



```
precision = precision_score(predictions[target_col], predictions["Predictions"])
print(f"Precision: {precision}")
```

```
Precision: 0.4947089947089947
```

IV. Experiments

Random Forest Classifier



	Features	Importance
4	ma_90	0.183692
2	ma_30	0.173378
5	ma_100	0.169235
0	ma_7	0.162515
1	ma_15	0.160854
3	ma_60	0.150326

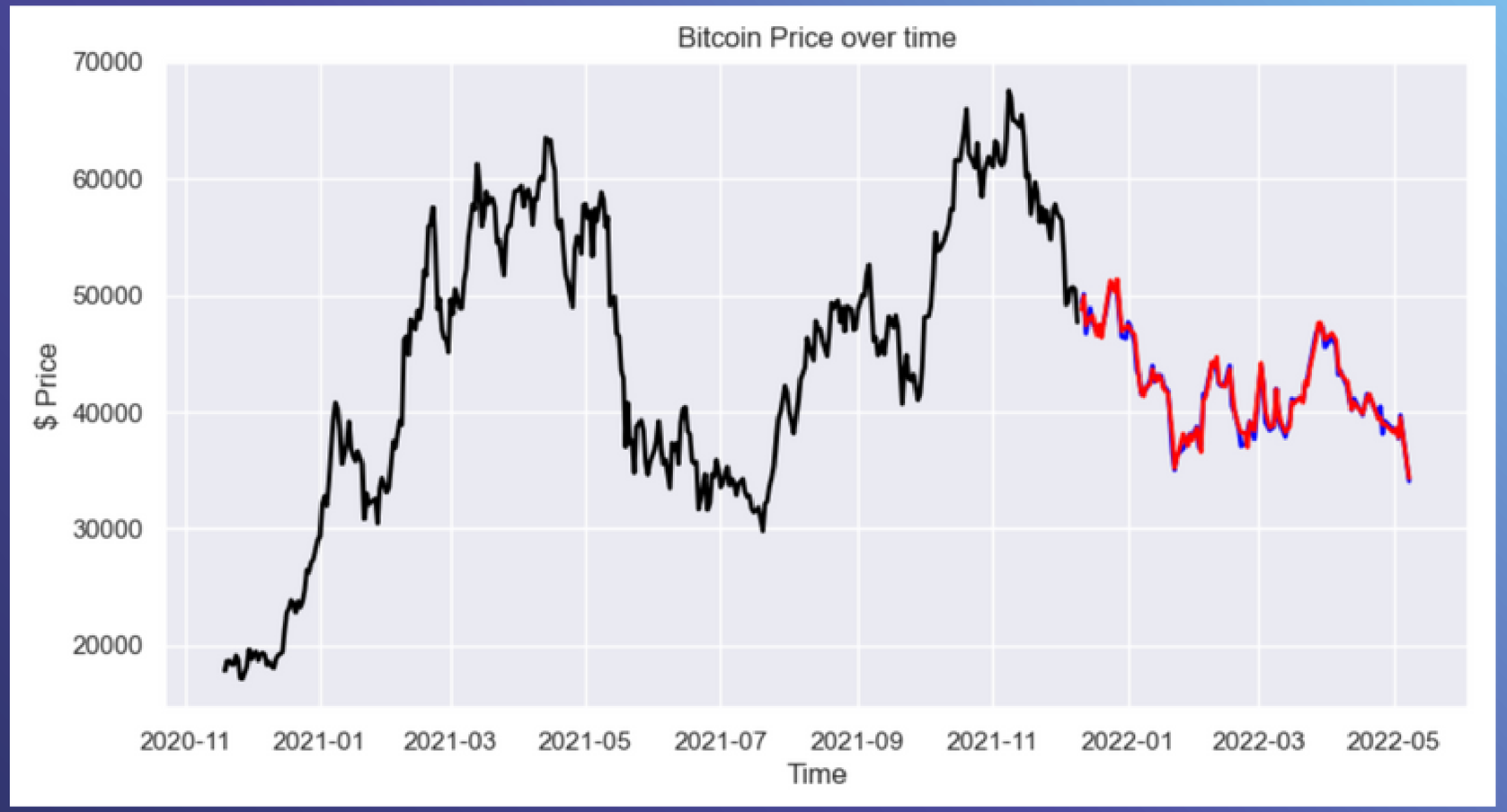
```
precision = precision_score(predictions[target_col], predictions["Predictions"])
print(f"Precision: {precision}")
```

```
Precision: 0.5159235668789809
```

Linear Regression Model

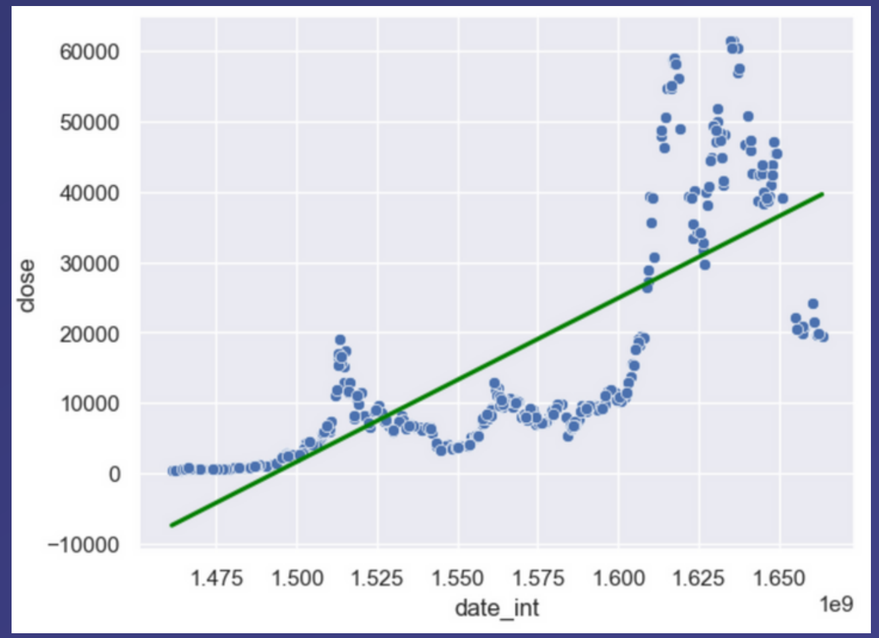
Predicting the value of Bitcoin for 100 days using linear regression model

- **Blue** line is the actual value
- **Red** is the predicted value



IV. Experiments

Linear Regression Model




```
# Compute the R-squared score to evaluate the performance
r2_score = reg.score(X_test, y_test)
print('R-squared score:', r2_score)

# Assuming y_test and y_pred are the actual and
#predicted values of the target variable respectively
mse = mean_squared_error(y_test, y_pred)
print('Mean Squared Error:', mse)
```

```
R-squared score: 0.5835251371977449
Mean Squared Error: 103586773.08521338
```

- Split the data into training and testing sets, and train the linear regression model on the training set
- R-Squared score of the linear regression model is still lower than the R-Squared score of the LSTM Model
- Conclusion: it is hard to predict the value of Bitcoin using best fit line method since the fluctuations of the value is too great, and it does not increase linearly.

Summary

- LSTM might be a better choice to predict the value of the Bitcoin more accurately.
- Linear regression model might not be a good choice as the fluctuations in the value of Bitcoin is too big.
- Due to the dynamic nature of the real world and unforeseen circumstances, the value of the Bitcoin might easily deviate from the predicted value.
- Regardless, the predicted value of Bitcoin can still help stakeholders to make important future decisions.

V. Conclusion

What we have learned

- Long Short-Term Memory Networks (LSTM Model)
- Tensorflow and Keras
- Predictors and backtests

Improvements

- Our model only takes into account historical price data and does not incorporate external factors
- Increase the amount of high quality training data
- Experiment with different hyperparameters

References

1. 10 popular types of cryptocurrency and how they work. N26. (n.d.). Retrieved April 22, 2023, from https://n26.com/en-eu/blog/types-of-cryptocurrency
2. Intellipaat. (2023, March 4). What is LSTM - introduction to long short term memory. Intellipaat Blog. Retrieved April 22, 2023, from https://intellipaat.com/blog/what-is-lstm/#:~:text='%20LSTM%20stands%20for%20long%20short,especially%20in%20sequence%20prediction%20problems.
3. MaharshiPandya. (2022, October 26). Cryptocurrency prices data. Kaggle. Retrieved April 22, 2023, from https://www.kaggle.com/datasets/maharshipandya/-cryptocurrency-historical-prices-dataset
4. What is Linear Regression? CORP-MIDS1 (MDS). (2022, August 5). Retrieved April 22, 2023, from https://www.mastersindatascience.org/learning/machine-learning-algorithms/linear-regression/#:~:text=What%20Is%20Linear%20Regression%3F,results%20based%20on%20past%20relationships.
5. What is Random Forest? IBM. (n.d.). Retrieved April 22, 2023, from https://www.ibm.com/in-en/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.