STAATLICH
ANERKANNTE
HOCHSCHULE

Bitcoin price prediction using ARIMA and LSTM methods that incorporate Tweet sentiment and Google trend analytics

HOCHSCHULE
SRH HEIDELBERG
Intelligence in Learning

Presented by :
Manmeet Kumar Chaudhuri

DAS CORE PRINZIP

INTRODUCTION

METHOD/ARCHITECTURE

DATASET

ANALYSES & VISUALISATIONs

FORECASTING MODELS

EVALUATION & VISUALISATIONS

SHORTCOMINGS/FUTURE RESEARCH

INTRODUCTION

Use Case

STAATLICH
ANERKANNTE
HOCHSCHULE

- Investors (Retails/Hedge Funds) want to allocate a portfolio of their investment/trades to cryptocurrency
- They want a price prediction system that will support them in making investment/trading decision
- The price prediction system/application discussed in the project can be used along-with many other analyses/factors for supporting investment/trading decisions
- Suitable visualizations/dashboards for displaying analyses and prediction

Cryptocurrency

- Cryptocurrencies are a subset of virtual currencies that use cryptography for security
- Blockchain is the guiding technology behind cryptocurrencies
- One of the most valuable and decentralized cryptocurrencies is Bitcoin, which was introduced by Satoshi Nakamoto on October 31, 2008. It captured around 70% of the total market capitalization
- Currently, there are more than 2646 cryptocurrencies of varying types
- People are using cryptocurrencies to implement a new form of economy, because of its cheapness, online, and anonymous means of exchange
- These currencies are unregulated and highly volatile. As a result, it can quickly devalue/sky-rocket overnight. These currencies have aggressive swings in their prices, as prices are largely based on public perception

Problem Statement

Motivation:

- Combine the aspects of machine learning and data visualization techniques for addressing a current challenge/topic

Aim:

- To perform exploratory data analysis on cryptocurrencies and to develop a machine learning pricing algorithm that is capable of modelling cryptocurrencies to offer next-day pricing prediction

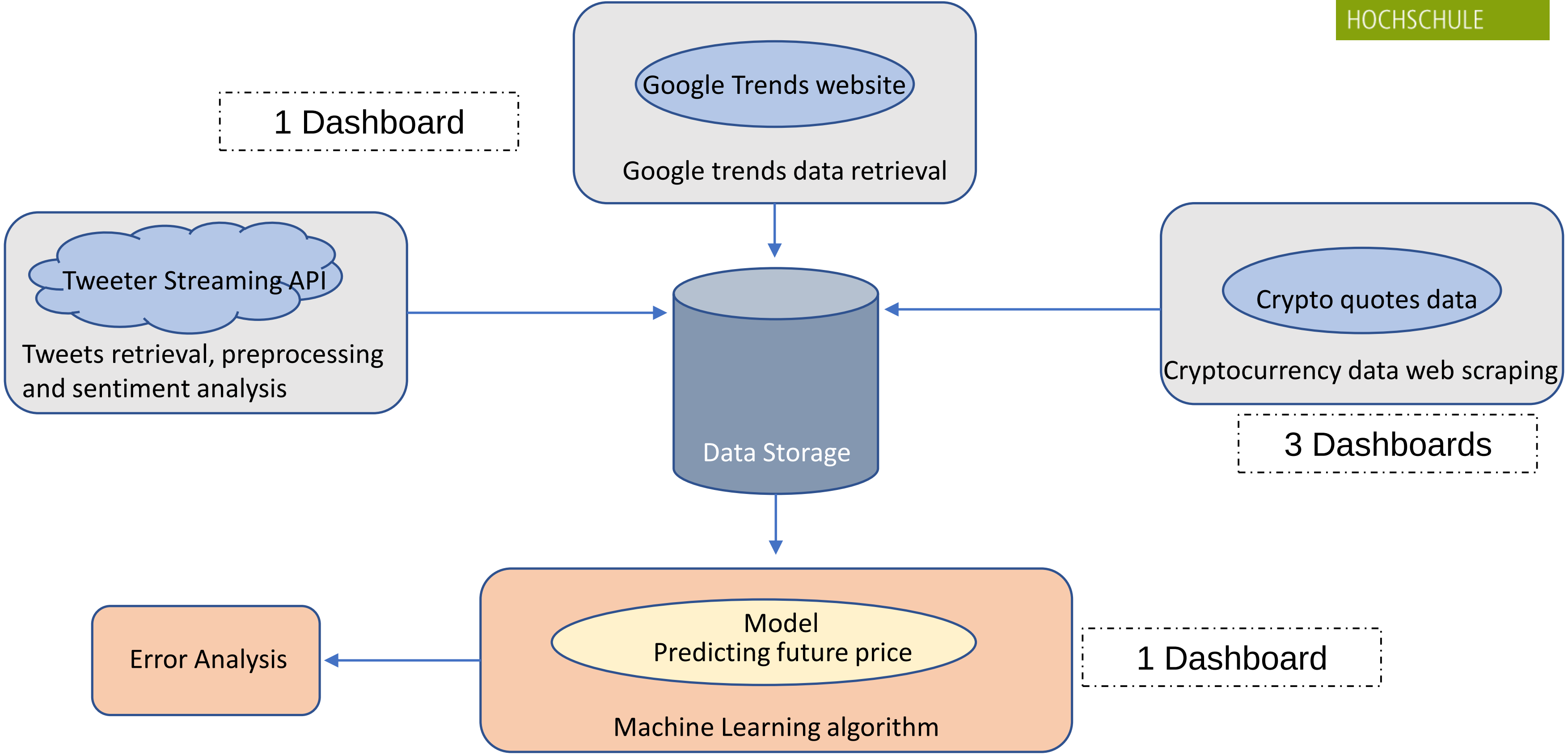
- To incorporate social media influence and google trend data in the pricing model

- To present the results of all the above-mentioned analysis with the help of effective data visualization techniques

METHOD/ARCHITECTURE

STAATLICH
ANERKANNTE
HOCHSCHULE

System Design



DATASET

Data

STAATLICH
ANERKANNTE
HOCHSCHULE

Cryptocurrency Data (Top 10 cryptocurrencies as per market capitalization)

S. No.	Cryptocurrency	Time Period
1	Bitcoin	28th April 2013 – until now
2	Ethereum	7th August 2015 – until now
3	Ripple	4th August 2013 – until now
4	Bitcoin Cash	23rd July 2017 – until now
5	Litecoin	28th April – until now
6	Tether	25th February 2015 – until now
7	Binance Coin	25th July 2017 – until now
8	EOS	1st July 2017 – until now
9	Bitcoin SV	9th November 2018 – until now
10	Monero	21st May 2014 – until now

Features for each of the cryptocurrency

S. No.	Features	Unit
1	Open	USD
2	High	USD
3	Low	USD
4	Close	USD
5	Volume	USD
6	Market Cap	USD

Data

Tweeter Data

- Verified Tweeter profiles
- Tweets from 24th August onwards
- Tweets with ‘Bitcoin’ or ‘btc’ as keywords

S. No.	Feature
1	Tweet ID
2	Date
3	Time
4	Tweet

Google trend data

- Keyword ‘Bitcoin’
- Data from 24th August 2018 onwards

ANALYSIS

Cryptocurrency Data: Price and Market Capitalization

- These are the raw features in the cryptocurrency dataset
- Market capitalization is a measure of the value of security/cryptocurrency
- A high or low market capitalization can reveal a coin that is resistant to volatility, or vulnerable to volatility
- Also, normally the currency with higher price is less volatile as compared to the currency with lower price
- A dashboard is prepared in Tableau containing the important pricing and market capitalization information for the users

From the perspective for an investor

- Choose the cryptocurrency with large Market Capitalization and high Price per unit

Cryptocurrency Data: Daily Volatility and Volume/Market Cap ratio

- These are the derived features in the cryptocurrency dataset

- Volatility describes the extent to which an asset’s price fluctuates over time. Daily volatility is the increase/decrease in price with respect to the previous day’s price

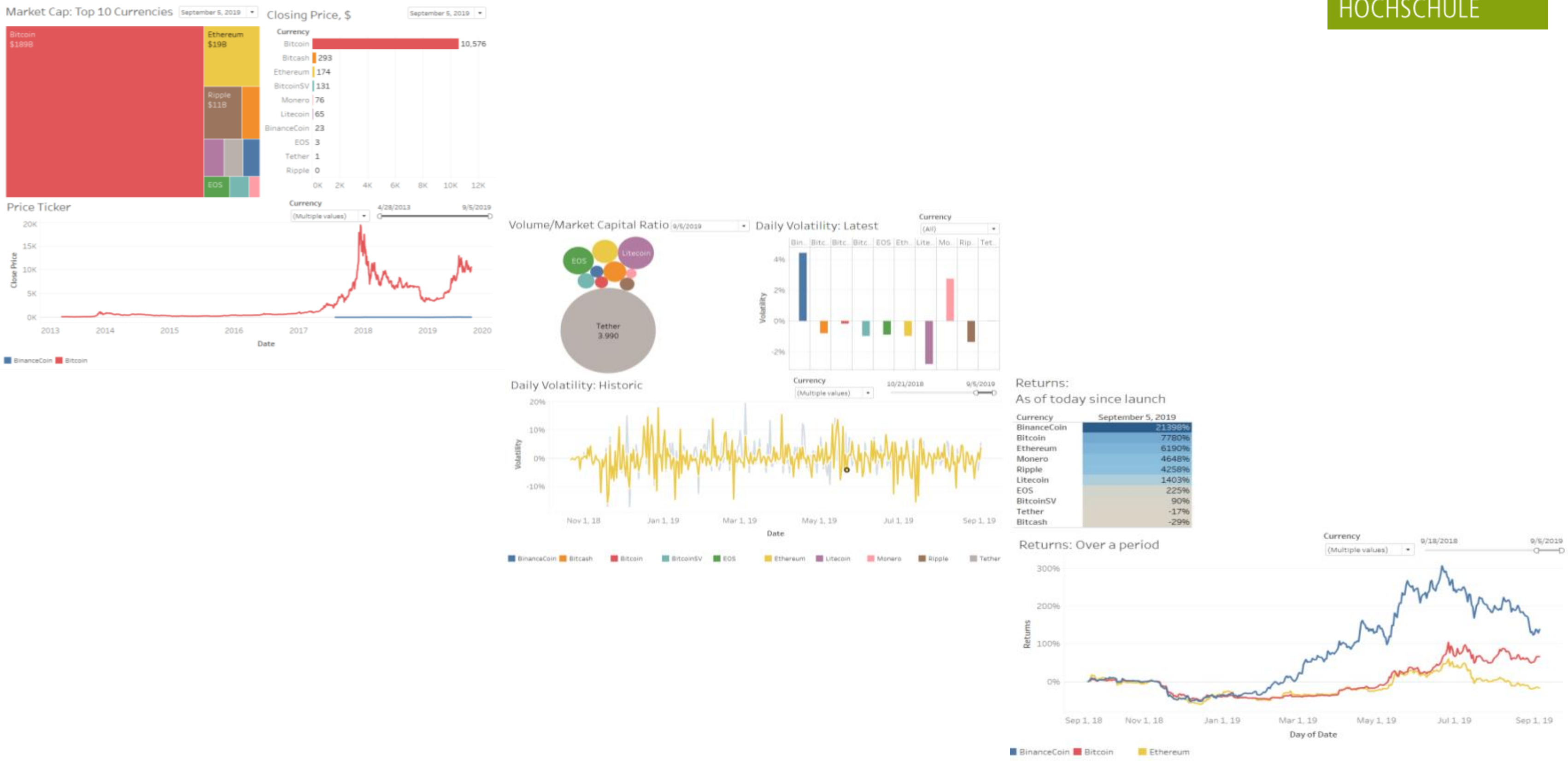
- Volume/Market cap is also an indirect measure of volatility as high volume/market ratio indicates more trading and more volatility

- A dashboard is prepared in Tableau containing the Daily Volatility and Volume/Market Cap

From the perspective for an investor

- For novice/beginner investors, it makes sense to choose less volatile cryptocurrency having low Volume/Market Cap ratio to begin with

Real Time Tableau Visualization



Returns:
As of today since launch

Currency	September 5, 2019
BinanceCoin	21398%
Bitcoin	7780%
Ethereum	6190%
Monero	4648%
Ripple	4258%
Litecoin	1403%
EOS	225%
BitcoinSV	90%
Tether	-17%
Bitcash	-29%

Cryptocurrency Data: Returns

- Returns (R_t) over a period (n days) is the relative increase or decrease in the price at the end (P_t) of the period over the price at the start (P_{t-n}) of the period. Mathematically,

$$R_t = (P_t - P_{t-n}) / P_{t-n}$$

- Returns are the direct measure of the performance of a cryptocurrency over a period. Higher the returns, higher is the probability of money chasing those assets
- A dashboard is prepared in Tableau containing return since launch of respective cryptocurrencies and the returns over a period of time

From the perspective for an investor

- Choose the cryptocurrency with large Market Capitalization and high Price per unit

Cryptocurrency Data: Currency shortlisting

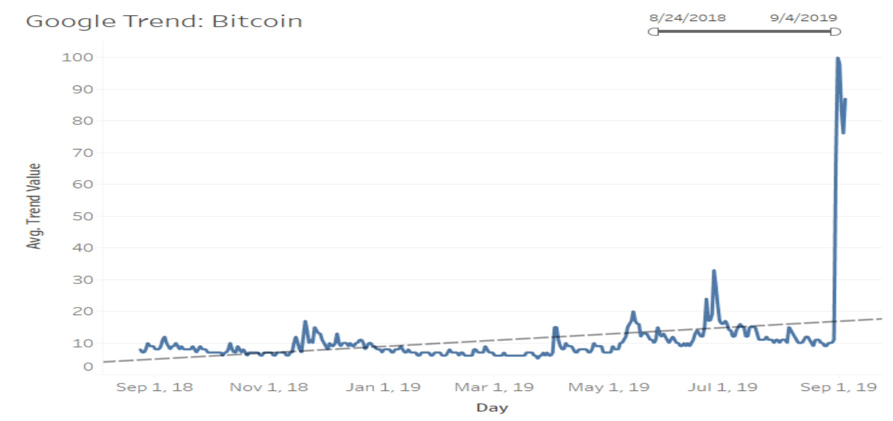
- The exploratory data analysis based upon the various raw and derived features can be summarized in the following way:

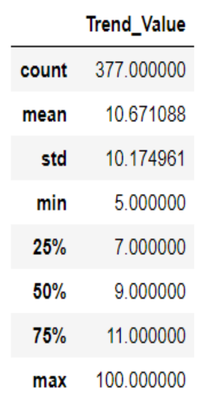
Ranking	Market Capitalization	Volatility	Returns
1	Bitcoin	Bitcoin	Binance Coin
2	Ethereum	Binance Coin	Bitcoin
3	Bitcash	Monero	Ethereum

- It can be seen that Bitcoin is the most suitable currency for investment/trading for beginners based upon the selected parameters

Google Trend Data

- This provides a proxy metric for the general interest there is in cryptocurrencies at any given time
- Google trends data is indexed to 100, where 100 is the maximum search interest for the time



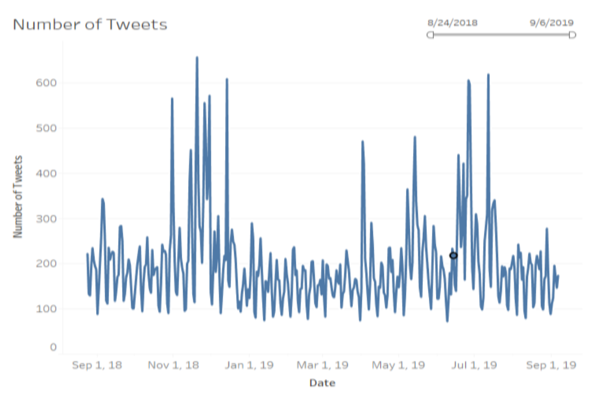


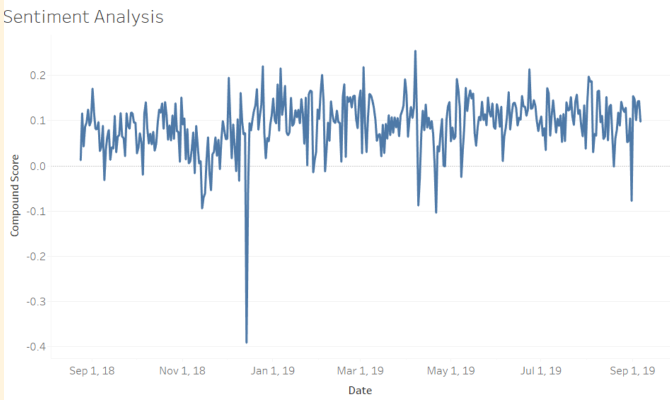
	Trend_Value
count	377.000000
mean	10.671088
std	10.174961
min	5.000000
25%	7.000000
50%	9.000000
75%	11.000000
max	100.000000

- It can be seen that the trend is of upward nature. However, an anomaly can be observed wherein the last couple of datapoints from 31st August are very high compared to the average trend

Tweeter data

- Tweets with keyword ‘Bitcoin’ and from verified account only were collected for this study
- VADER (Valence Aware Dictionary for sEntiment Reasoning) sentiment analysis was used to analyze the collected tweets

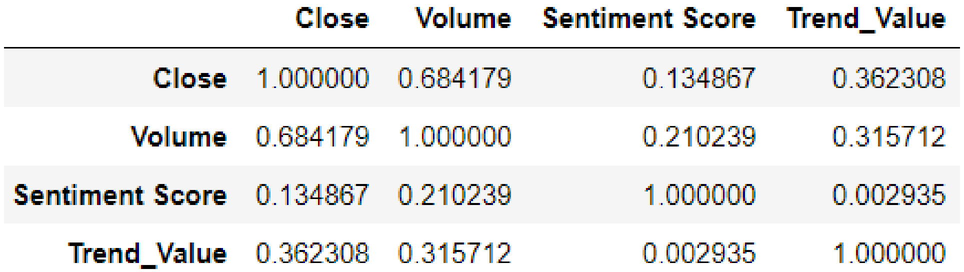




- It can be observed that the average daily count of verified tweets is c. 200
- It is observed that 94% of the tweets have positive/neutral sentiment with compound scores between 0.00 and 0.3. Only around 6% of the tweets have negative sentiment

Correlation

Correlation between Closing Price, Volume, Sentiment Score and Google trend (daily values)



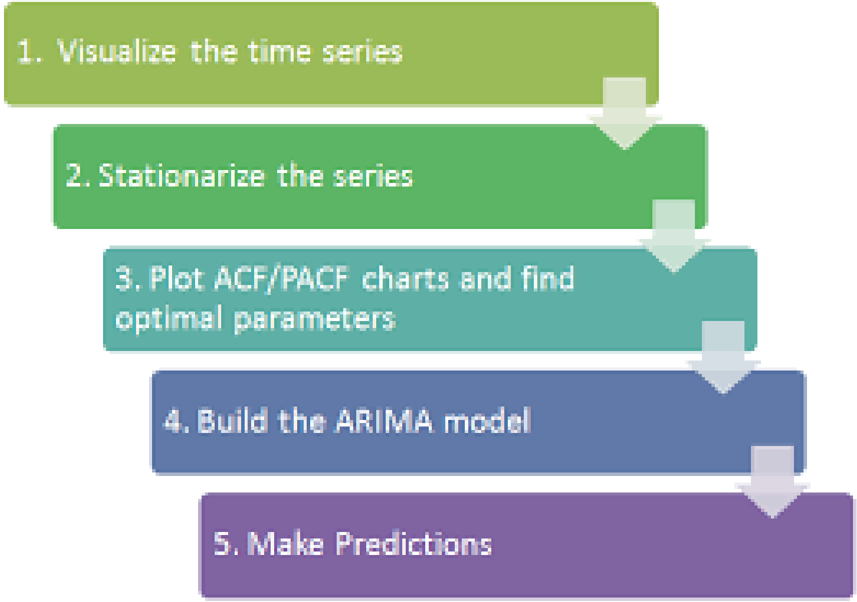
	Close	Volume	Sentiment Score	Trend_Value
Close	1.000000	0.684179	0.134867	0.362308
Volume	0.684179	1.000000	0.210239	0.315712
Sentiment Score	0.134867	0.210239	1.000000	0.002935
Trend_Value	0.362308	0.315712	0.002935	1.000000

- All the variable are positively correlated
- Closing Price and Volume has a highest correlation
- Sentiment score and the Google trend value has the least correlation

FORECASTING MODELS

ARIMA Model

STAATLICH
ANERKANNTE
HOCHSCHULE


1. Visualize the time series
2. Stationarize the series
3. Plot ACF/PACF charts and find optimal parameters
4. Build the ARIMA model
5. Make Predictions

- The underlying principle in ARIMA (Auto regression integrated moving average) model is to estimate the trend and seasonality in the series and remove those from the series to get a stationary series

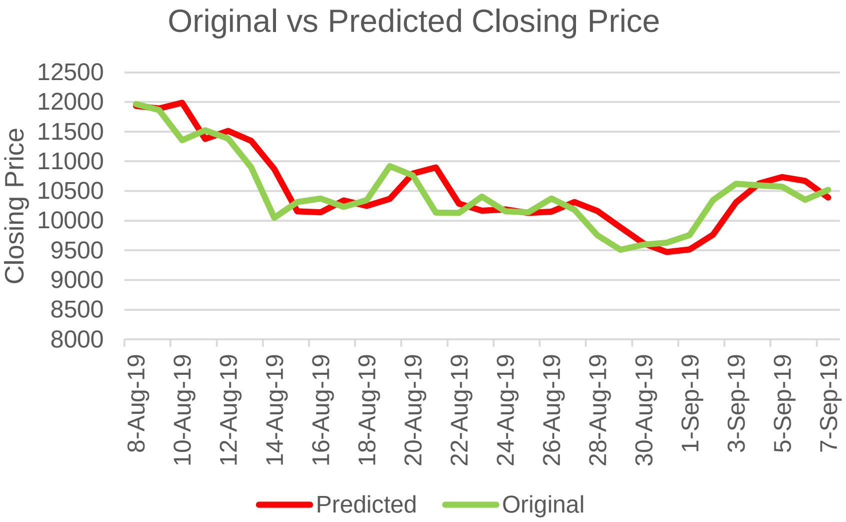
- Then statistical forecasting techniques can be implemented in this series

- The final step would be to convert the forecasted values into the original scale by applying trend and seasonality constraints back

- A standard notation is used for ARIMA(p,d,q)
  p is the lag order
  q is the degree of differencing
  d is the order of moving average

ARIMA Model: Prediction

- The Bitcoin data from 24th August 2018 onwards is used in the pricing model
- Only historical Closing Price data is used in the model
- Next-day Bitcoin price is predicted from 8th August 2019 until 6th September 2019
- The period between 8th August 2019 and 6th September 2019 is used to test the model
- The variation in Predicted vs Original Closing Price is visualized in Tableau. The visualization is maintained up-to-date by running a python script



RMSE	331.62
MAPE, %	2.38

LSTM Model

- A type of RNN is the Long Short-Term Memory (LSTM) model
- A RNN is capable of learning sequences and temporal processing and can therefore be applied to forecasting problems
- The prediction of financial assets using a LSTM is a rather new research topic
- The advantage of the LSTM over traditional RNN’s is that it is capable of learning long time lag problems and additionally generalizes well for smaller lag problems
- Using LSTM techniques, researchers have found that a LSTM provides superior results in prediction of stock market price movement when compared to traditional statistical methods (Nelson et al., 2017)

LSTM Model: Scenarios for analysis

We have used the following four scenarios in the next day price prediction:

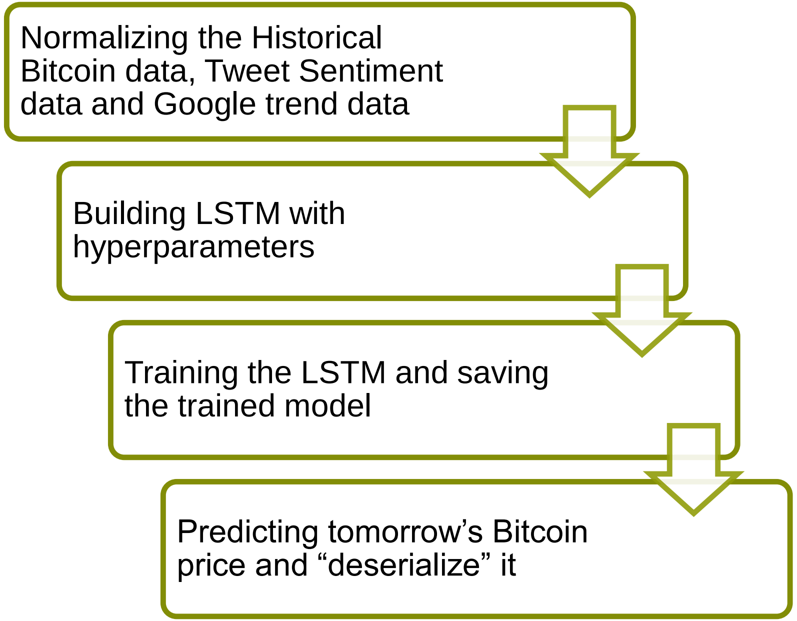
- Scenario 1: Effect of Tweet sentiment, Google trend data, volume and historic closing price on next day Bitcoin Closing price
- Scenario 2: Effect of Google trend data, volume and historic closing price on next day Bitcoin Closing price
- Scenario 3: Effect of volume and historic closing price on next day Bitcoin Closing price
- Scenario 4: Effect of historic closing price on the next day Bitcoin Closing price

Training and test data

The Bitcoin pricing data from 24th August 2018 is used in the pricing model. We have also used the Tweeter data and Google trends data from 24th August onwards in the LSTM Model.

Next-day Bitcoin price is predicted from 8th August 2019 until 6th September 2019. The period between 8th August and 6th September is used to test the model

LSTM Model: Steps and Parameters



Normalizing the Historical Bitcoin data, Tweet Sentiment data and Google trend data

Building LSTM with hyperparameters

Training the LSTM and saving the trained model

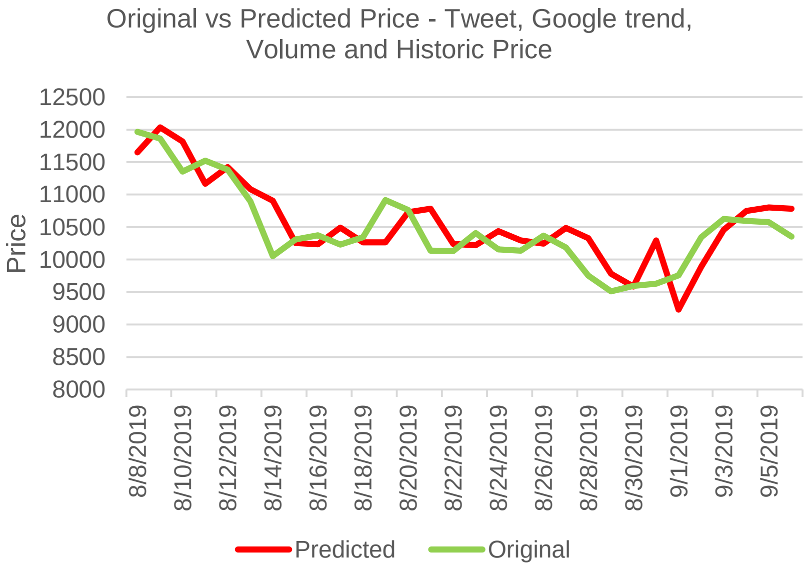
Predicting tomorrow’s Bitcoin price and “deserialize” it

- For Deep Learning backend system, we choose Tensor-flow, and Keras as the front-end layer of building neural networks fast. Pandas is used extensively for data related tasks, Numpy is utilized for matrix/vector operations and for storing training and test data sets, Scikit-learn (also known as: sklearn) is used for performing the min-max normalization
- Sequential model with one LSTM layer and one Dense layer
- Hyperparameters used in the model based upon literature review, and various trials and errors

Parameters	Value
Optimizer	adam
Loss function	MAE (mean absolute error)
Activation function	ReLu
Number of neurons in hidden layer	100
Epochs	1000
Batch size	100

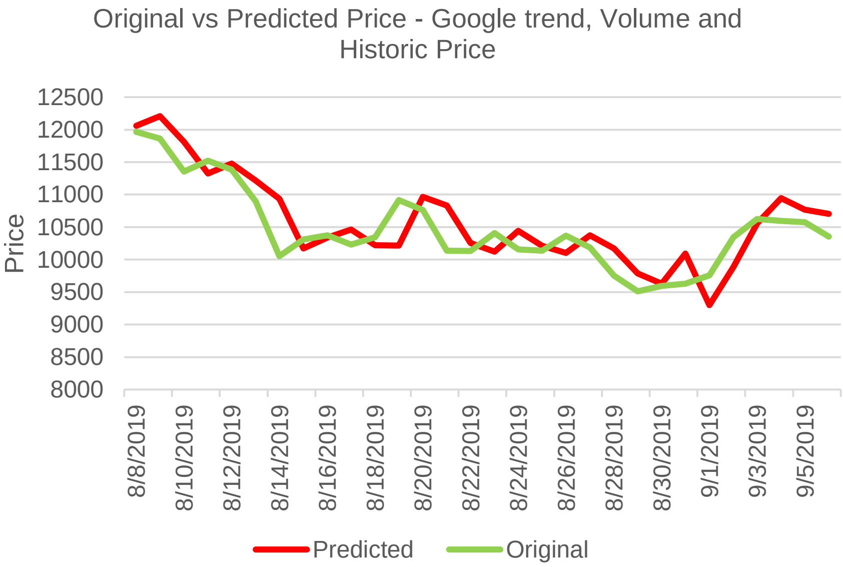
LSTM Model: Prediction

Scenario 1



RMSE	367.36
MAPE, %	2.86

Scenario 2

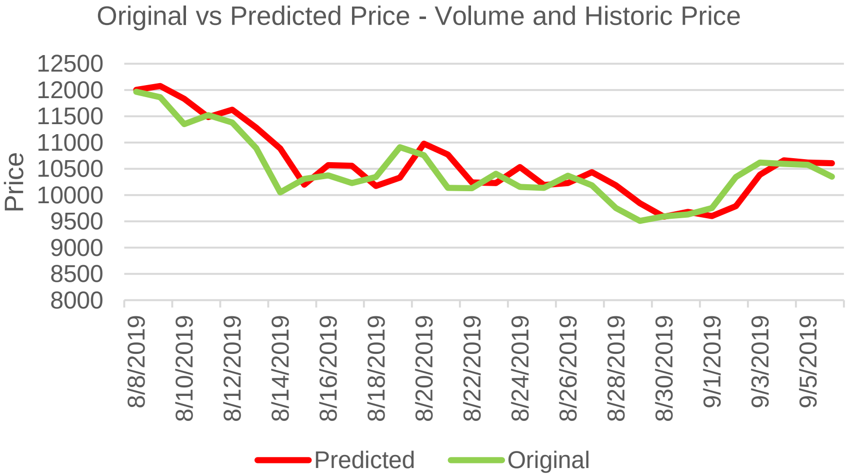


RMSE	357.52
MAPE, %	2.82

LSTM Model: Prediction

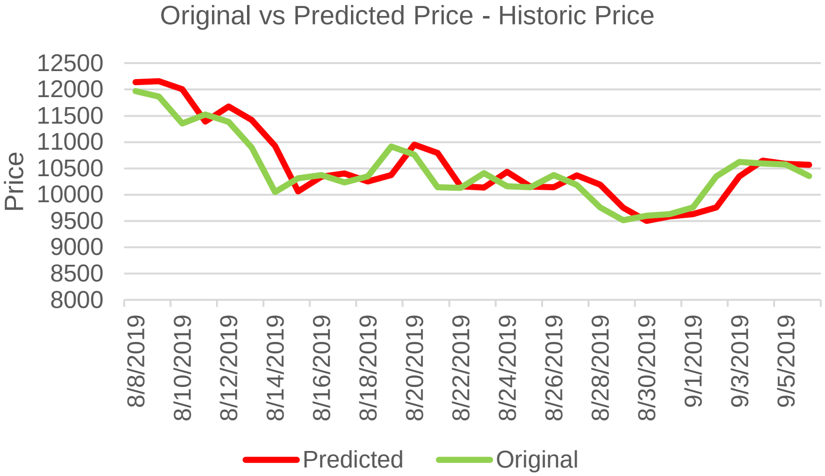
STAATLICH
ANERKANNTE
HOCHSCHULE

Scenario 3



RMSE	327.19
MAPE, %	2.48

Scenario 4



RMSE	342.92
MAPE, %	2.53

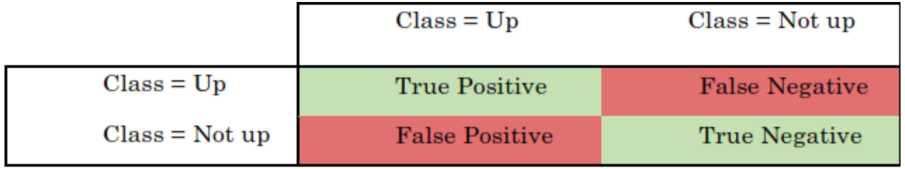
EVALUATION

Parameters

- The performance of both the models i.e. ARIMA and LSTM on the test set for the prediction of exact prices, was evaluated using the Root Mean Squared Error metric together with the Mean Absolute Percentage Error as error metrics

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y_i} - y_i)^2}{n}} \qquad MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|true - predicted|}{|true|}$$

- Subsequently, the direction (up, down) of the predicted variables were compared to the direction of the real Bitcoin price



		Class = Up	Class = Not up
True	Class = Up	True Positive	False Negative
	Class = Not up	False Positive	True Negative

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \qquad F1 = 2 * \frac{Recall * Precision}{Recall + Precision}$$

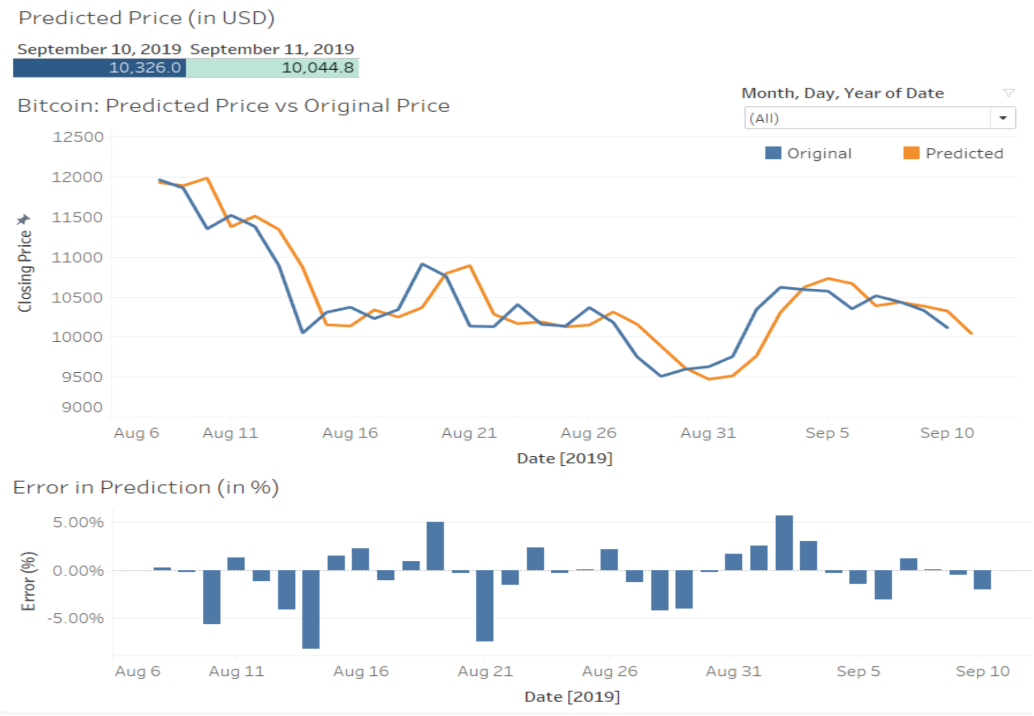
Evaluation and Discussion

The evaluation of ARIMA and LSTM model on the performance parameters is as follows:

Parameter	ARIMA	LSTM			
		Scenario 1	Scenario 2	Scenario 3	Scenario 4
RMSE	331.62	367.36	357.52	327.19	342.92
MAPE, %	2.38	2.86	2.82	2.48	2.53
Accuracy, %	45%	43.3	50	60	50
Precision	0.37	0.35	0.45	0.5	0.41
Recall	0.43	0.38	0.38	0.58	0.38
F1-Score	0.40	0.37	0.41	0.54	0.40

- ARIMA method is more robust than the LSTM Method for predicting the price of the Bitcoin
- LSTM model with historic volume and price data is more robust than other methods for predicting the up/down movement in Bitcoin price
- Tweet sentiments are not effective in predicting the next day Bitcoin price
- Google trend data on ‘Bitcoin’ keyword is also not useful in predicting the next day Bitcoin price forecast
- Historic price and historic volume of the Bitcoin are the best variable to predict the next day Bitcoin price

Dashboard: Real time next day Bitcoin price



Predicted Price (in USD)

September 10, 2019	September 11, 2019
10,326.0	10,044.8

Month, Day, Year of Date: (All)

SHORTCOMINGS AND FUTURE RESEARCH

Some pointers

STAATLICH
ANERKANNTE
HOCHSCHULE

- Cryptocurrency prices are influenced by numerous other factors like: microeconomics, macroeconomics, general news, sentiment on other internet forums or harmful events in the world of cryptocurrencies such as hacks

- Preference of dense data with smaller time intervals as input. Additionally, sentiment on internet/tweeter can change by the hour so the same shortcoming of density might apply to the sentiment input data

- Implementing hyperparameter tuning in LSTM, in order to get a more accurate network architecture

Thank You!