# Problem Statement

Cryptocurrencies are one of the most emerging technological advancements in the financial sector of the past decade, and the leading cryptocurrency is Bitcoin. Cryptocurrencies, by nature, are extremely volatile, and Bitcoin is no different. The ability to predict the future price of bitcoin and forecast it is an extremely valuable asset, as it could guide investors on when to invest into said cryptocurrency. It also is beneficial to gather information about the world economy as a whole; since bitcoin is the biggest cryptocurrency, it shows and dictates patterns to other cryptocurrencies and other economical factors.

The aim of this project is to develop a data mining model utilizing the KDD process in RapidMiner Studio to preprocess, mine, and extract valuable information from its historical price data, and compare which model and data mining technique yields the most accurate results, ultimately leading to a model that is able to accurately predict and forecast the future price of Bitcoin.

This project addresses the challenges of predicting the future price of such a volatile concept through historical data in relation to it. Applying data preprocessing, model training and evaluation techniques in RapidMiner Studio to come up with the most accurate model to predict the Bitcoin price in the future.

By the end of the project, we expect to find the most accurate model throughout our testing, that can effectively predict future prices of Bitcoin, and gain more knowledge on what are the factors that affect its pricing overall.

# Selection of Data Set

The dataset that we’ve chosen is the Bitcoin dataset from Kaggle from the article “Cryptocurrency Historical Prices.” Which contains historical Bitcoin statistics collected over several years, namely from 2009 to 2017, which encompasses numerous values such as market performance and blockchain activity.

We have chosen this dataset for our project for a multitude of reasons, including but not limited to:

* It’s directly related to our topic in the financial technology sector
* It has more than sufficient data and is structured well enough to come up with predictive models and forecast bitcoin prices
* It has great depth for modeling as it covers many concepts such as blockchain and economic features

The dataset consists of 24 attributes, 23 of them being quantitative (float64) and 1 date attribute. There are over 1500 instances/records, more specifically, 1584 instances.

The dataset also has 478 missing values in the btc\_trade\_volume attribute, which is a highly important attribute for our analysis and modeling, so we would have to fill those in during preprocessing in Task 3.

The dataset also contains some redundant attributes, with some attributes having a correlation value of >0.99, alongside multiple transaction related attributes such as btc\_n\_transactions, btc\_n\_transactions\_excluding\_popular, and btc\_n\_transactions\_total, this will also be addressed in task 3 with either feature selection or applying PCA.

The following are some visualizations regarding the dataset to further understand the distribution and summary of the data.

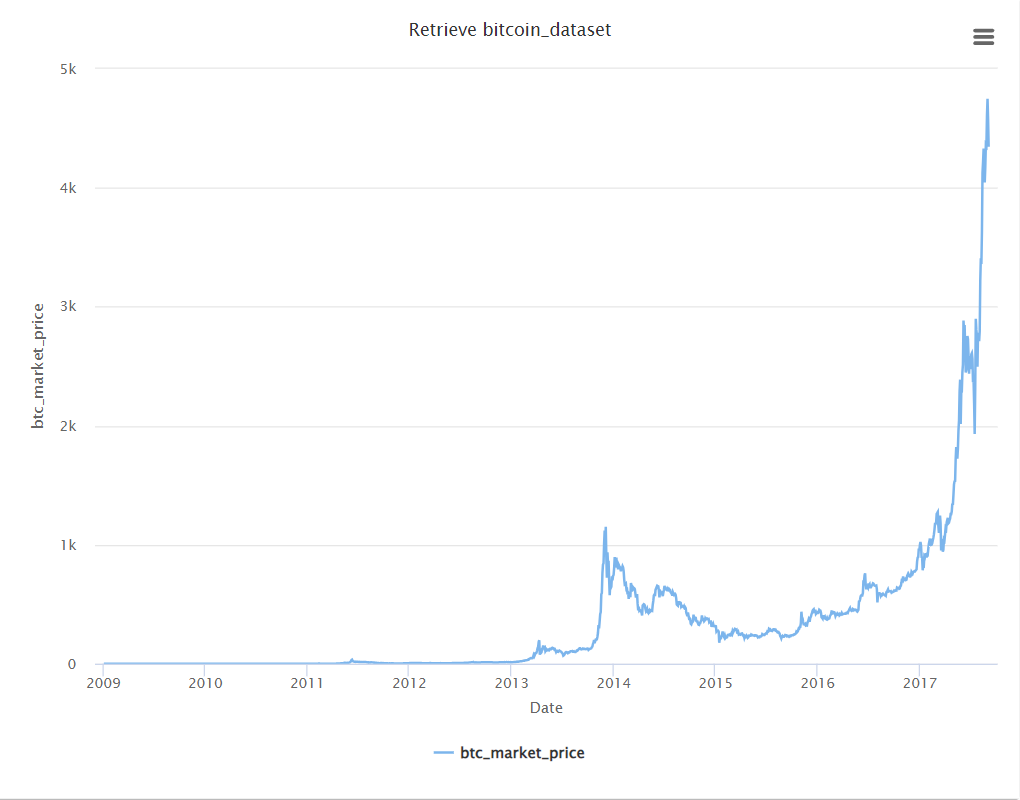


Figure - The market price of bitcoin throughout the years

Figure 1 shows how volatile the price of bitcoin can be, and how it has had exponential growth in the later years.

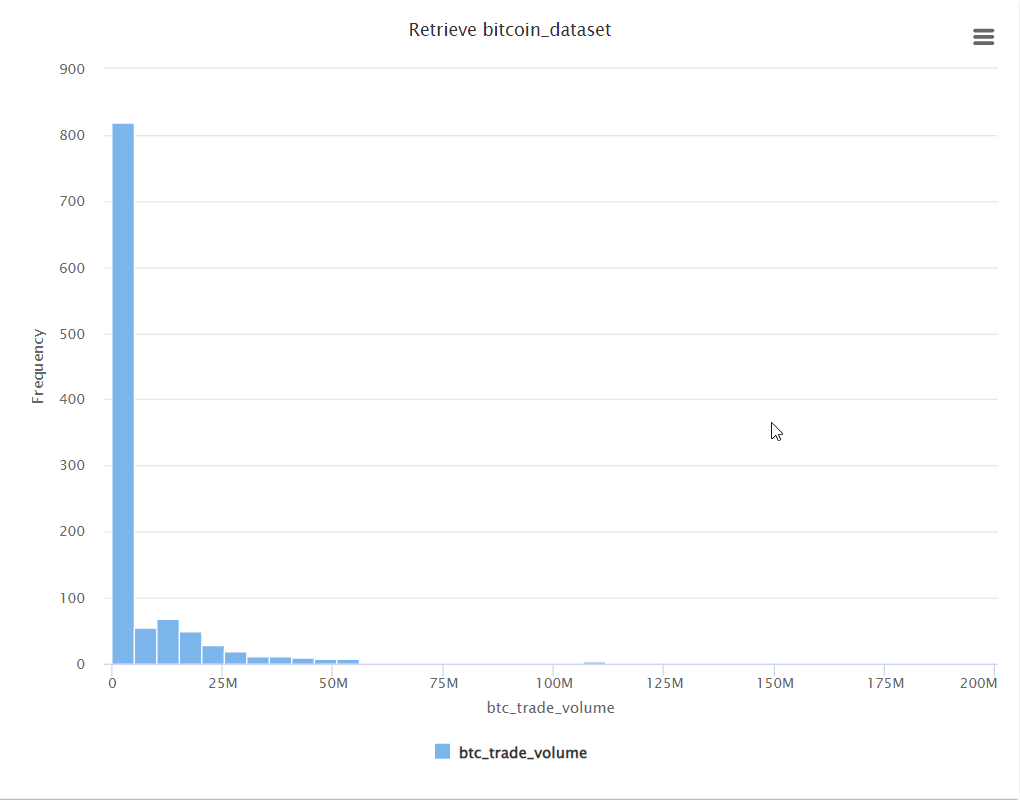


Figure - A Histogram showcasing btc\_trade\_volume

As we can see from the histogram in figure 2, the btc\_trade\_volume attribute is extremely rightly skewed, and there are significant outliers in the data, showcasing the need for proper outlier treatment and normalizing the values.

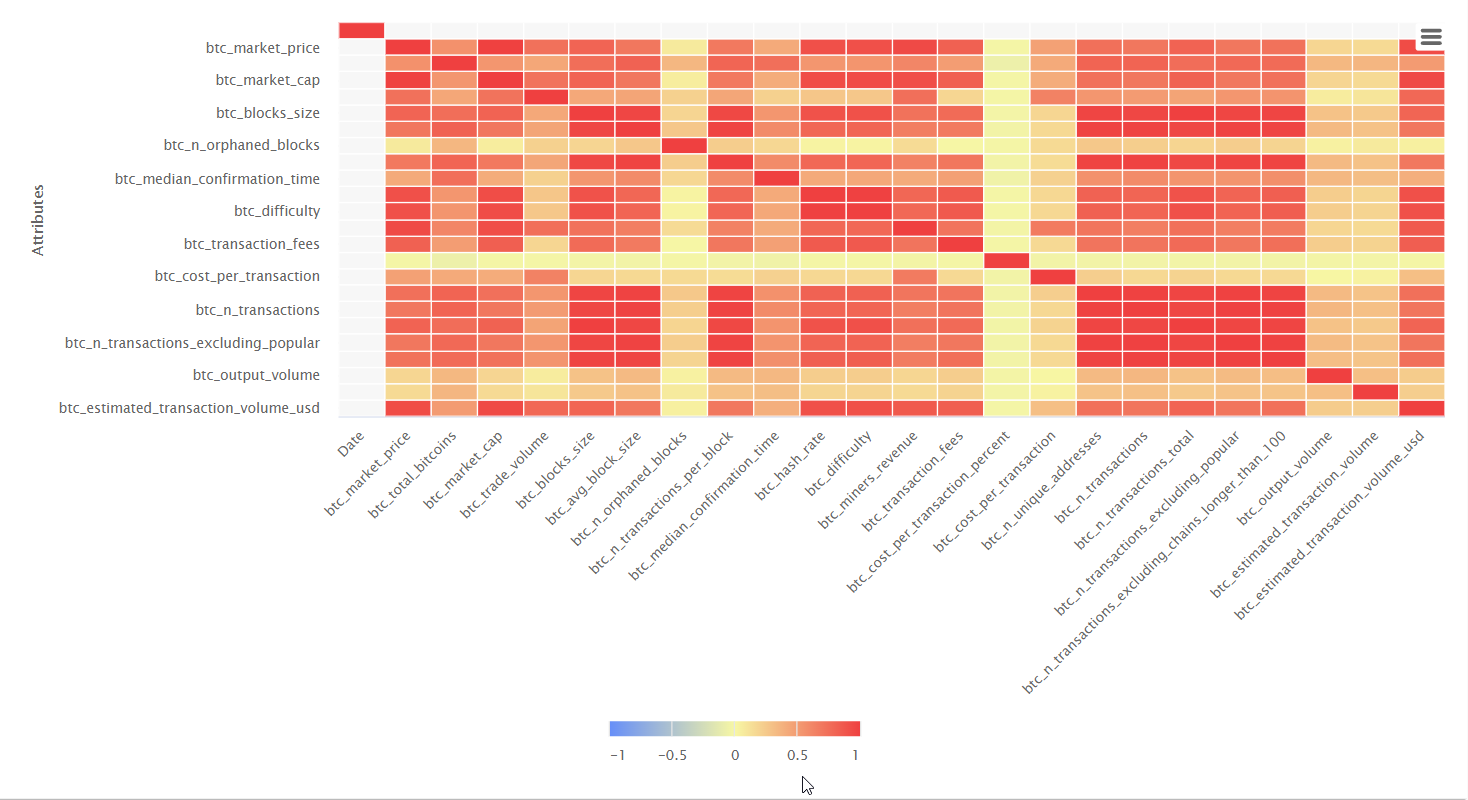


Figure 3 - Heatmap correlation Matrix

Figure 3 displays just how extremely highly correlated some attributes are, reinforcing the point we discussed earlier about redundant attributes, and the need for a dimenstionality reduction technique such as PCA.

A graph with blue lines

AI-generated content may be incorrect.

Figure - Normalized Box plot of difficulty, hash rate, and market price

Figure 4 shows a normalized box plot of the 3 attributes to showcase the outliers and how the large scale attributes can severely skew the data.

This dataset is a time-series dataset, showcasing the evolution and volatility of the bitcoin cryptocurrency over time. Numerous attributes (such as market\_price and difficulty) display exponential growth in the later years, which goes hand in hand with the evolution in the financial technology sector. Within the attributes, there are large scale attributes, which are magnitudes bigger than others, portraying the need to normalize the data for better results.

# Preparation and pre-processing of Data

A diagram of data processing

AI-generated content may be incorrect.

Figure - KDD process diagram

The Knowledge Discovery in Databases process comprises of numerous stages. It starts off with Data Selection, then goes to Data Cleaning and transformation and reduction, the step we are currently at.

Missing values, inconsistent values, redundant data and outliers are all challenges to be faced during the process of KDD, and all are obstacles that must be addressed to get proper outcomes and being able to extract desired data efficiently and effectively.

## Data Cleaning

The first part of it is data cleaning, the process of dealing with inaccurate, inconsistent, noisy, or incomplete data. The first step was to deal with the missing values, the dataset had significant gaps in the btc\_trade\_volume attribute, with seven significant gaps, ranging from 3 day gaps all the way to a gap that is 367 days. Based on this, a 2 step plan has been implemented to deal with the missing values, while keeping respect to the time-series nature of the data.

For the smaller gaps of less than 22 days, we applied a moving average using the moving average operator in RapidMiner, with a window width of 22 (for 22 rows, which equates to 22 days in our dataset)

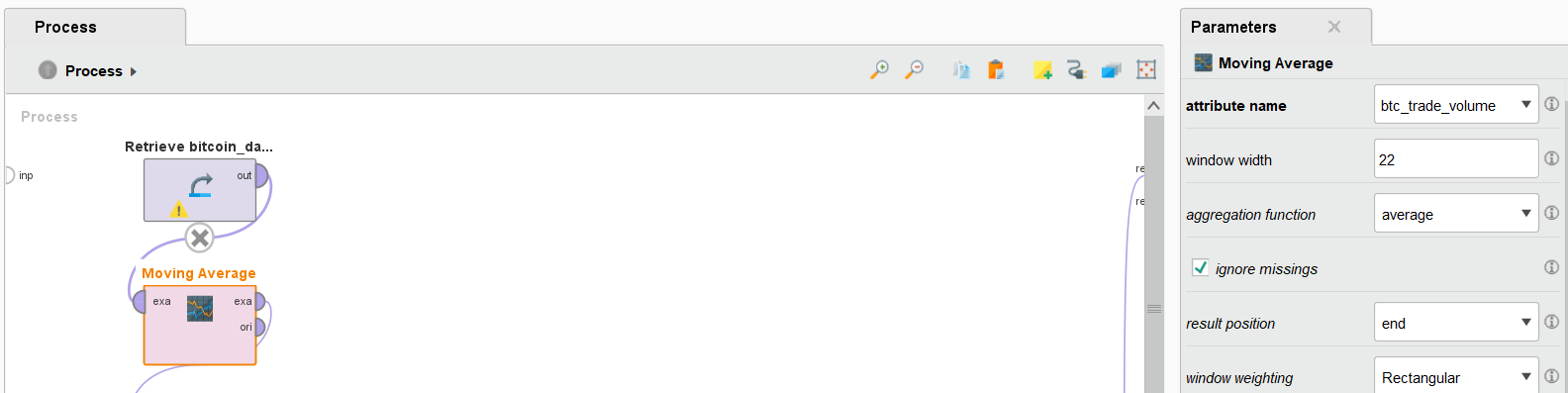


Figure - Parameters of the moving average operator

The moving average operator takes in the values consisting of the current value and x amount of neighboring values (in our case 22) and takes in the average of it, providing a way to smoothen the gaps without sacrificing the integrity of the data.

Consecutively applied is the Generate Attributes operator, used to create a new column called btc\_trade\_volume\_RMV that is conditioned to fill in the missing values in the original attribute of gaps<22 days with the average, skipping over non-missing rows.

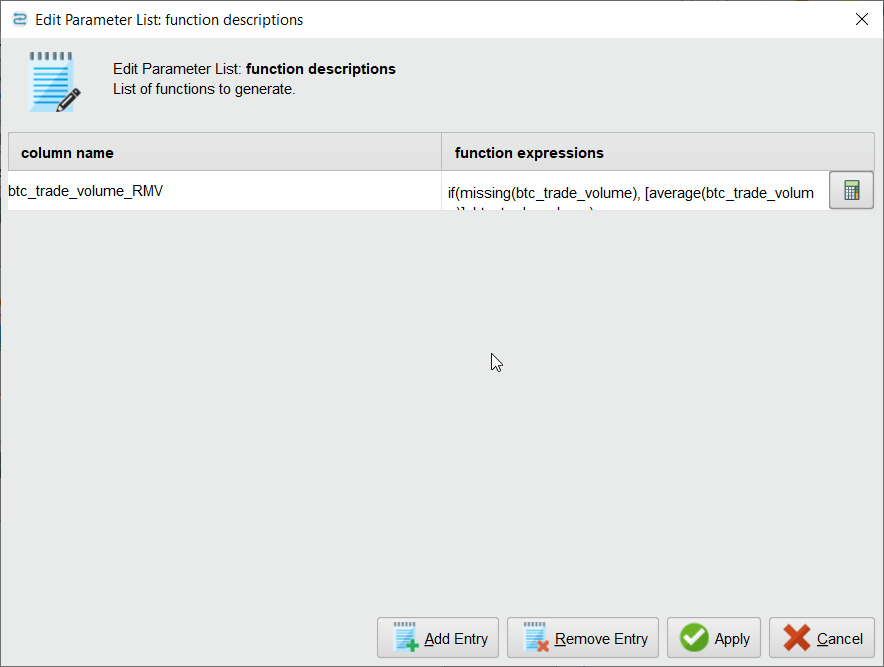


Figure - using the if(missing(btc\_trade\_volume), [average(btc\_trade\_volume)], btc\_trade\_volume) expression in the generate attributes operator

Secondly, for our large gap of 367 rows (i.e 367 days), the forward fill method has been used as averages of a gap this big could not be easily calculated and would not be feasible.

Utilizing the Replace Missing Values (Series) operator set with “previous value” to achieve forward fill to logically fill the aforementioned gap in respect to the time series nature of the dataset.

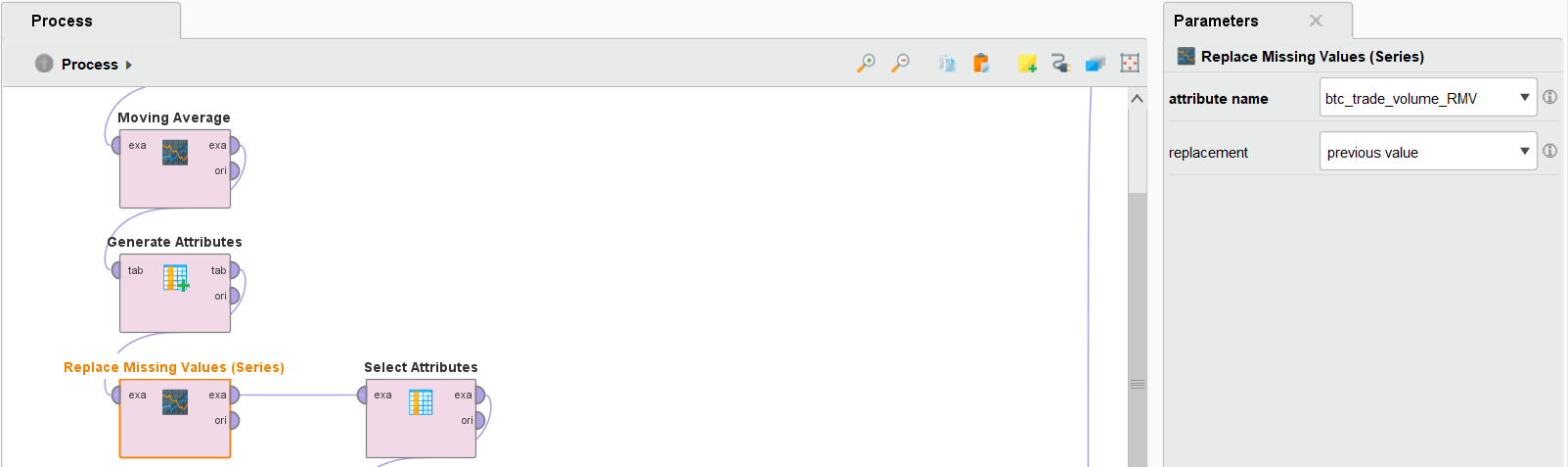


Figure - The set up of the operator to forward fill the gap

The combination of the moving average and the forward fill methods allows us to fill all the missing values in the dataset ensures readiness for analysis and mining in the next steps, while maintaining the integrity, realism and completion of all attributes within the dataset.

## Selection and Normalization of Attributes

After ensuring that the dataset is now complete, we’ve selected the most relevant attributes to keep for the next steps of preprocessing and mining, as those attributes encompass the most needed and most relevant parts of the dataset. We’ve utilized the select attributes operator to achieve this.

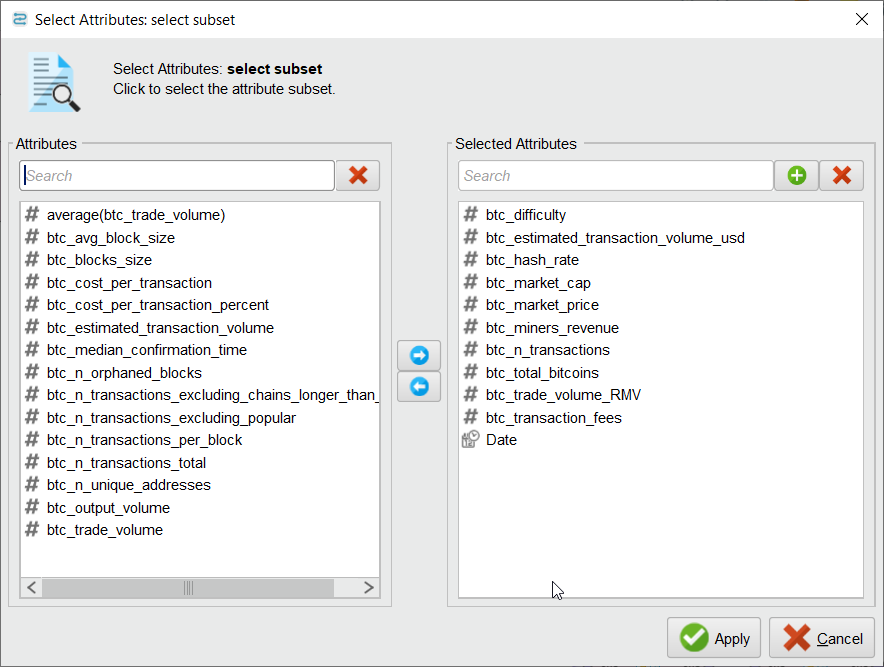


Figure - Using the select attributes operator to only include a subset of attributes

Alongside that, we’ve also used the rename operator to change the btc\_trade\_volume\_RMV attribute to its original name, as that is the column with the complete data.

We’ve added the Generate ID attribute to the process to have a unique identifier and keep respect to the time series regardless of which models we use.

Next comes normalization, an extremely important step in data pre-processing to ensure all values are taken into account equally, eliminating outliers, preventing numerical instability, and massively improve performance for a multitude of algorithms such as nueral networks and distance based algorithms such as KNN.

We have implemented the Normalize operator using Z-Transformation (or statistical normalization) to the selected attributes, it bases each attribute around a mean of zero, achieving consistent feature scaling across the whole dataset.

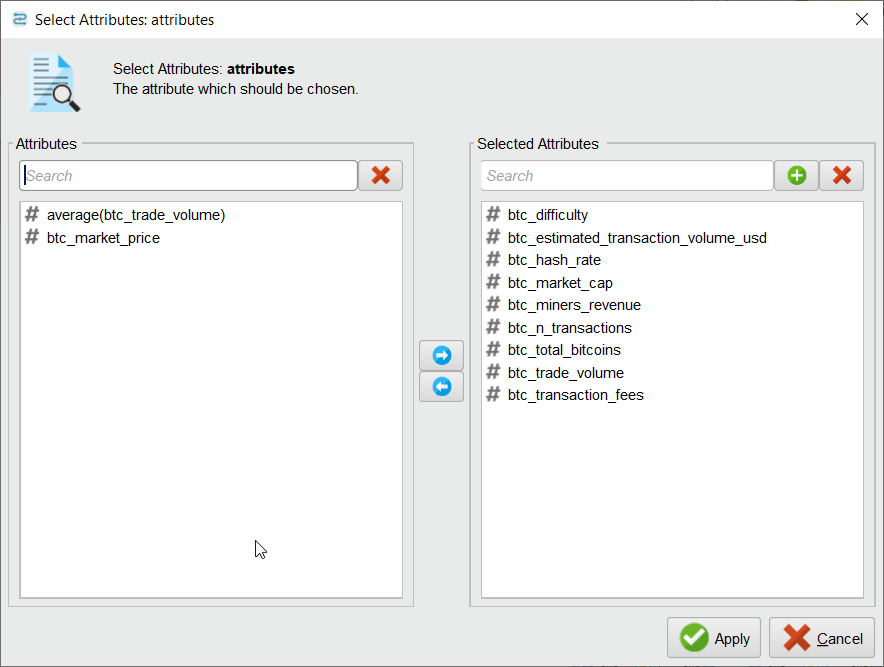


Figure - The configuration of the normalize operator

The btc\_market\_price was intentionally left out as preserving its original value is highly important for the analysis and the ability to predict future values.

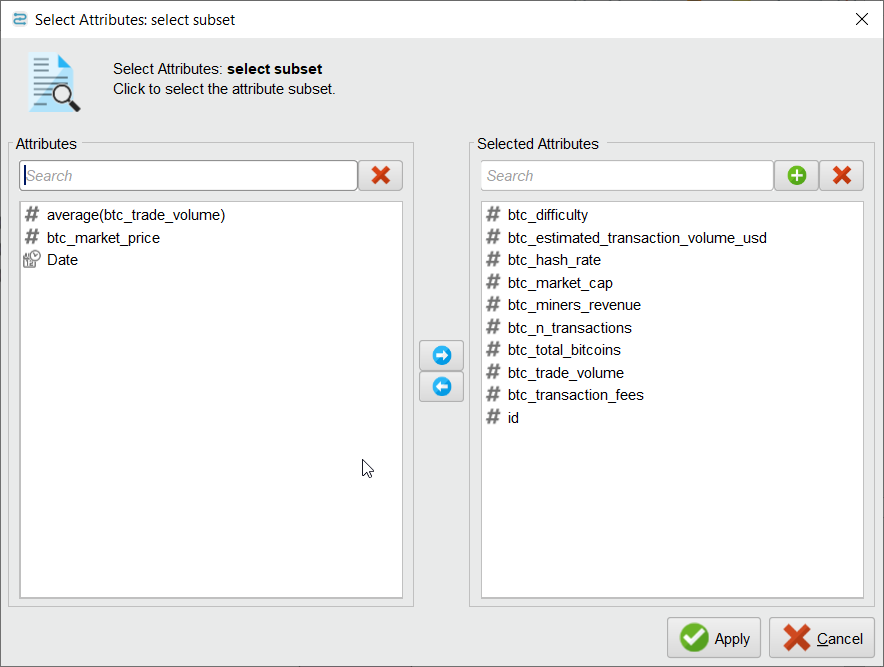
We have then added the multiply operator to duplicate our data, split into 2 different select attributes.   


Figure - Selection of PCA Attributes

The first select attributes feeding into the PCA operator, leaving out the trade volume and market price attributes to keep them as is.

A screenshot of a computer

AI-generated content may be incorrect.

Figure - Flow of Process and configuration of PCA

We have made the variance threshold at 0.95 to retain accurate data, but not to go too aggressive into dimensionality reduction and lose too much of the data.

The second split of the multiply leads to another select attribute wherein we take the btc market price, date and ID, to then join them with the attributes that have went through PCA.

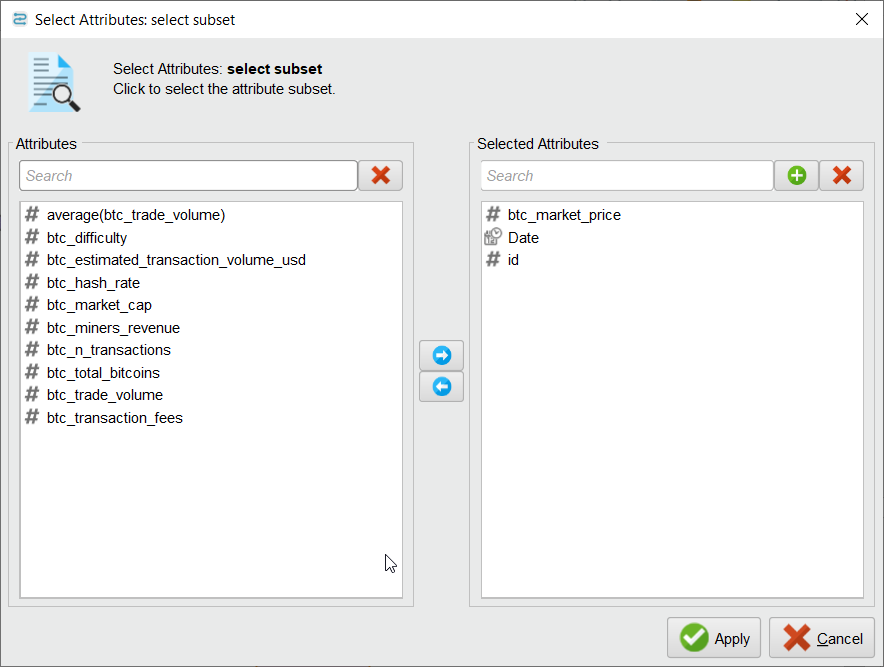


Figure - Second select attributes to join the date, market price, and ID into the PCA'd dataset

## Visualization differences between raw dataset and preprocessed dataset

Task 5, Model Evaluation

The different models have been evaluated been on 4 different performance criterion, since all of our models are classified under regression models, so the 4 criterion are Root Mean Squared Error(RMSE), Absolute Error(AE), Correlation(R) and Squared Correlation(R2).

A graph showing a line

AI-generated content may be incorrect.A black text with black text

AI-generated content may be incorrect.The first model’s prediction that was tested on the dataset is Linear Regression, it was the best performing model, having the lowest RMSE at 153.348, absolute error at 94.657, R of 0.992 and R2 of 0.983. It is expected for LR to be the best performing model as the linear, time-series nature of bitcoin prices works best with a simple model like linear regression, especially when other attributes such as btc\_transaction\_volume or btc\_hash\_rate move up or down with the btc\_market\_price attribute, our target attribute.

Figure - Visualization between btc\_market\_price and LRs predicted values

Figure - Linear Regression's Performance Vector

As it can be seen here, due to the aforementioned reasons, LR is within very close reach of the actual price of bitcoin throughout the dataset.

A graph with green and blue lines

AI-generated content may be incorrect.A black text with numbers and a black text

AI-generated content may be incorrect.Next up to be evaluated is the Random Forest model, with a RMSE of 889.614, absolute error of 428.586, R of 0.783 and R2 of 0.612. It was the least accurate model out of the 3 models we have used due to the nature of the bitcoin dataset. RF excels at handling noisy, nonlinear, high featured datasets since it creates a lot of decision trees which splits data into separated regions, which is not ideal for a (mostly) smooth, continuous attribute such as bitcoin’s market price. RF predictions are based on averages of leaves, which is counterintuitive for continuous data.

Figure - Random Forest's prediction against actual market price

Figure - Random Forest's performance vector

Due to the nature of random forest as a classification algorithm, it was not able to keep up with bitcoin price’s occasional volatility and spikes.

A screenshot of a computer code

AI-generated content may be incorrect.The last model to be tested is the neural net, which performed right in between LR and RF, showing RMSE of 623.387, AE of 310.208, R of 0.953 and R2 of 0.909. In general, NN performed adequately as it can capture interactions between attributes, it has very flexible learning capability due to the complexity of the algorithm and can approximate continuous functions, thus leading to good predictions. On the other hand, neural networks perform very well for huge datasets, larger than our dataset of 1584 examples, and due to the linearity of bitcoin’s market price, linear regression simply performs better.

Figure - Performance Vector for Neural Net

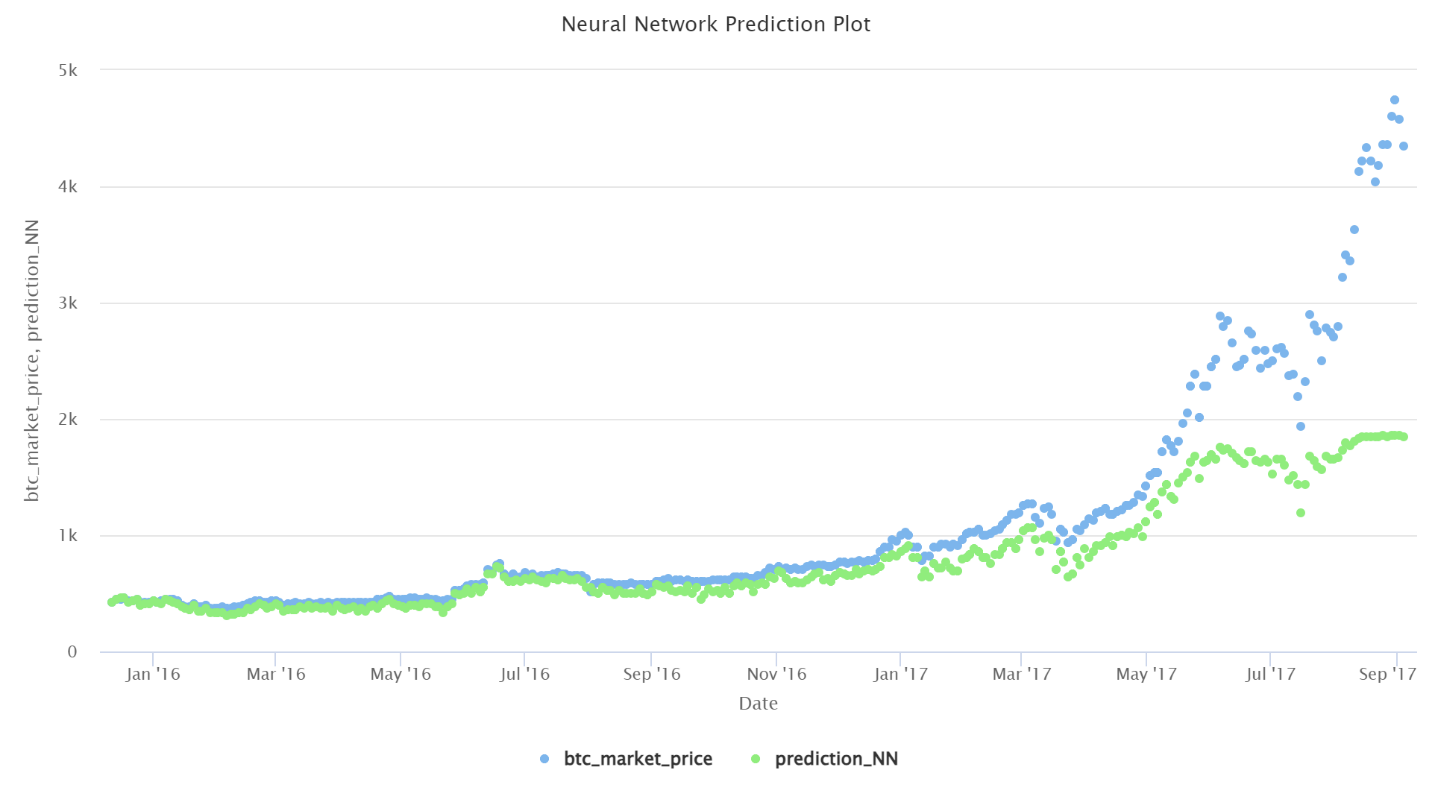


Figure - Neural Net's prediction against actual market price

We’ve tried numerous ensemble models, including a combination of all 3, LR&RF, and a LR,NN combination.

The one to be highlighted is the combination of Linear Regression and Neural Network, unsurpsingly, it was our best performing ensemble model due to the fact that their individual respective models were the best performing out of the 3.

A black text with white text

AI-generated content may be incorrect.It produced an RMSE of 305.313, an absolute error of 144.503, an R of 0.983 and R2 of 0.967, performing better than our neural network model alone, but slightly worse than the linear regression model. Due to the LR model being very highly suited for the data, the complexity added by the slightly less accurate neural net, even after benefitting from the strengths of the neural net, the linear regression alone remained superior for predicting bitcoin’s price in our dataset.

Figure - Performance Vector for LR NN Ensemble model

A graph showing a line of growth

AI-generated content may be incorrect.

Figure - Scatter plot of the best performing ensemble model

Overview of each model’s performance:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **RMSE** | **MAE (Absolute Error)** | **Correlation** | **R² (Squared Correlation)** |
| **Linear Regression** | 153.348 | 94.657 | 0.992 | 0.983 |
| **Random Forest** | 889.614 | 428.586 | 0.783 | 0.612 |
| **Neural Network** | 623.387 | 310.208 | 0.953 | 0.909 |
| **LR + NN** | 305.313 | 144.503 | 0.983 | 0.967 |
| **LR + RF** | 429.867 | 217.240 | 0.989 | 0.979 |
| **LR + RF + NN** | 487.046 | 235.055 | 0.983 | 0.967 |

Some other visualizations encompassing all 3 individual models fared against each other, and all 3 ensembles fared against each other. A graph of a graph

AI-generated content may be incorrect.

Figure - Scatter plot of all models individually against the market price

This scatter plot shows precisely how much better linear regression is compared to the other models individually.

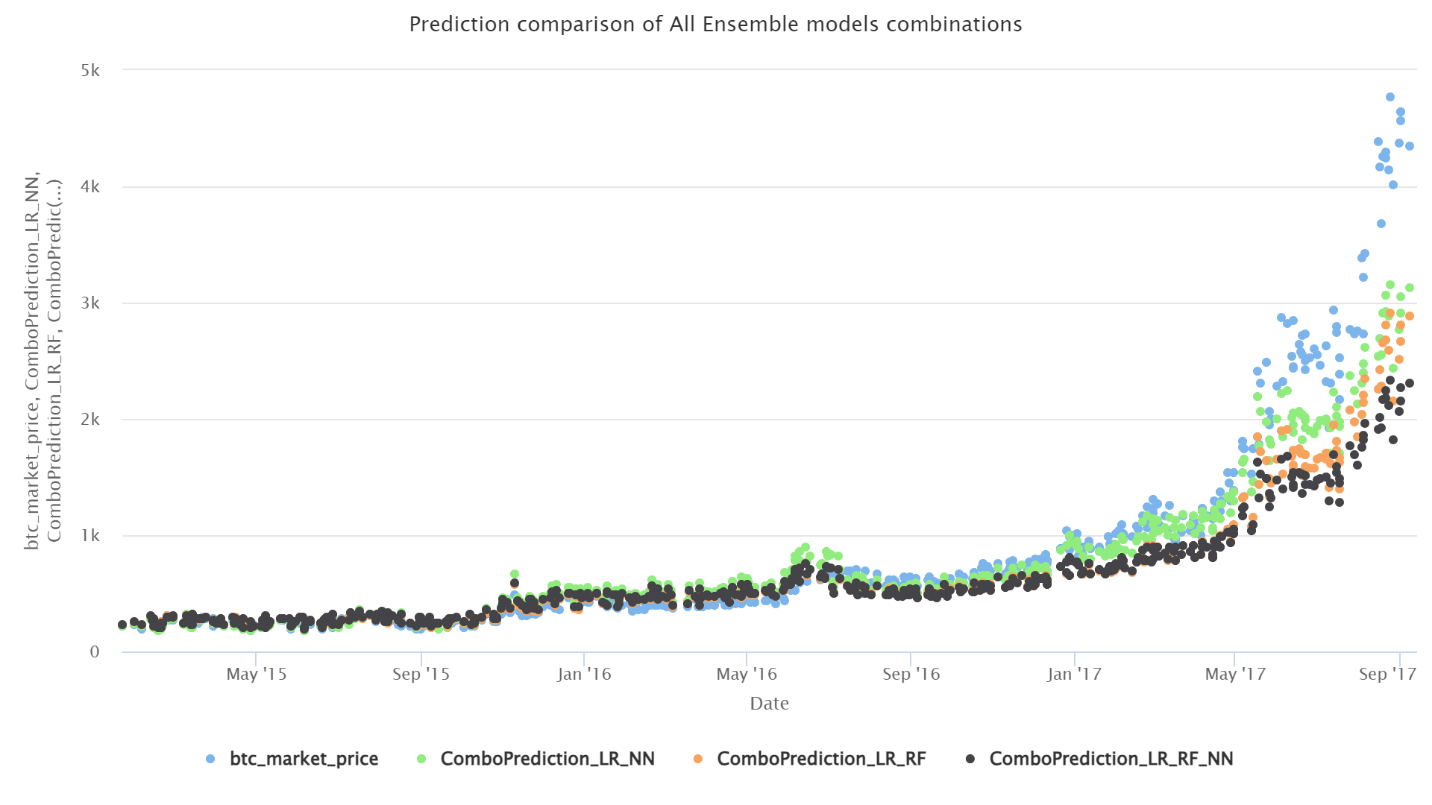


Figure - Scatter plot of all ensemble models against the bitcoin market price