Machine intelligence

Prediction of cryptocurrency

|  |  |
| --- | --- |
| Names | IDs |
| Farah Tarek Ahmed | 19100674 |
| Radwa Wael | 19105956 |
| Omar Shaban | 19105397 |
| Omar Abdelrazik | 19106149 |
| Mohamed El Sayed | 19105049 |

Dr: Ghada Khoriba

Table of Contents

[Problem statement: 3](#_Toc124011793)

[Related work: 3](#_Toc124011794)

[Models Architecture: 4](#_Toc124011795)

[KNN “key nearest neighbor” 4](#_Toc124011796)

[What is the KNN? 4](#_Toc124011797)

[The advantages of KNN 4](#_Toc124011798)

[The disadvantages of KNN 5](#_Toc124011799)

[Random Forest Regression 6](#_Toc124011800)

[What is Random Forest Regression? 6](#_Toc124011801)

[What are the advantages of random forest for real production applications? 6](#_Toc124011802)

[SVR model 7](#_Toc124011803)

[Advantages of SVR model 8](#_Toc124011805)

[Disadvantages of SVR model 8](#_Toc124011806)

[Linear Regression 8](#_Toc124011807)

[Simple implementation 9](#_Toc124011808)

[1) Performance on linearly separable datasets 9](#_Toc124011809)

[2) Over fitting can be reduced by regularization 9](#_Toc124011810)

Lasso Model……………………………………………………………………………….………………………10

Why lasso………………………………………………………………………………………….……………….10

The Advantages of Lasso………………………………………………………….………………..….…..11

The Disadvantages of Lasso………………………………………………………………………………..11

[Model evaluation and result 11](#_Toc124011811)

[Resources: 12](#_Toc124011812)

# Problem statement:

Cryptocurrencies have become a major element in enterprises and financial market potential during the past ten years. Predictions that are accurate can help bitcoin investors make the best decisions and possibly enhance their earnings. They can also assist financial analysts and policy makers in analyzing the behavior of the bitcoin markets. However, because of its chaotic and extremely complicated nature, bitcoin price prediction is regarded as a highly difficult undertaking. In this study, we compare some of the most popular and effective deep learning forecasting methods for cryptocurrencies. The outcomes show conclusive proof that deep learning models are unable to resolve this issue properly and efficiently. After extensive experimentation and analysis of the data, we conclude that it is crucial to develop and use new methodologies, other methodologies, and approaches, such as more advanced ensemble methods, feature engineering techniques, and other validation measures, sophisticated prediction algorithms.

# Related work:

This section describes some existing works in cryptocurrency price prediction. There are many deep learning models that are adopted for price prediction. In early work, Sin et al. (2017) proposed a forecasting algorithm that can be applied to different financial, engineering, and medical tasks. The algorithm integrated an artificial neural network (ANN) and a multilayer perceptron (MLP). From the experiments, the incorporation of MLP into ANN increased the Bitcoin price prediction accuracy from about 58% to 63%.

Yenidogan et al. (2018) utilized a deep learning system known as the Facebook Prophet for predicting the price of bitcoin. To obtain the best ratios for training, testing, and validation sets, a three-fold splitting strategy was used. The experimental outcomes demonstrated that the PROPHET algorithm performed better than the ARIMA method, with Prophet obtaining a lower root mean square error (RMSE) of 652.18 as opposed to 817.01 in the ARIMA algorithm.

Jiang (2020) suggested using deep learning techniques to acquire and reorganize data on the Bitcoin price each minute to hour to anticipate future prices for the cryptocurrency. Before feeding the dataset into the regression models, the dataset underwent pre-processing, mini-batch normalization, and min-max normalization. To forecast the price of Bitcoin, the paper suggested using a few deep learning networks, including MLP, RNN with the extension of LSTM, and GRU. The experimental findings demonstrated that the MLP model, which utilized two layers of GRU, had the greatest outcome and had the lowest RMSE of 19.020.

# Models Architecture:

## KNN “key nearest neighbor”

What is the KNN?

One of the fundamental machine learning algorithms is K Nearest Neighbor (KNN). The k-nearest neighbors’ algorithm, sometimes referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to produce classifications or predictions about the grouping of a single data point. A set of input values are used by machine learning models to forecast output values. One of the simplest machine learning algorithms, KNN is primarily employed for categorization. The data point is categorized based on how its neighbor is categorized.

## The advantages of KNN

1. There Is No Training Period: KNN Is Known as A Lazy Learner (Instance based learning). During the training phase, it doesn't learn anything. The training data are not used to derive any discriminative function. In other words, it doesn't require any training. Only when making real-time predictions does it draw on the training dataset it has stored. Due to this, the KNN algorithm is much faster than other training-required algorithms, such as SVM, Linear Regression, etc.
2. Since the KNN algorithm doesn't need to be trained before producing predictions, new data can be supplied without disrupting the system's accuracy.
3. KNN is simple to use. KNN implementation just needs two parameters: the value of K and the distance function (e.g., Euclidean or Manhattan etc.)

## The disadvantages of KNN

1. Does not perform well with large datasets because it is expensive to calculate the distance between each current point and a new point in a large dataset, which hurts the algorithm's efficiency.

2. Does not perform well with high dimensional data: The KNN algorithm performs poorly with high dimensional data because it becomes challenging for the algorithm to calculate the distance in each dimension as the number of dimensions increases.

3. Need for feature scaling: Prior to using the KNN method on any dataset, feature scaling (also known as standardization and normalization) is required. If we don't, KNN might produce inaccurate forecasts.

4. KNN is susceptible to noise in the dataset. It is also sensitive to missing values and outliers. Manual imputation of missing values is required and remove outliers

## Random Forest Regression

## What is Random Forest Regression?

## 

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. Random forest regression is used to solve a variety of business problems where the company needs to predict a continuous value: Predict future prices/costs.

## What are the advantages of random forest for real production applications?

Random forest regression is a popular algorithm due to its many benefits in production settings:

1. Extremely high accuracy. Thanks to its ‘wisdom of the crowds’ approach, random forest regression achieves extremely high accuracies. It usually produces better results than other linear models, including [linear regression](https://www.keboola.com/blog/linear-regression-machine-learning) and [logistic regression](https://www.keboola.com/blog/linear-regression-machine-learning).
2. Scales well. Computationally, the algorithm scales well when new features or samples are added to the dataset.
3. Interpretable. Although it is not as easily explainable as its underlying algorithm decision tree regression, random forests can be inspected to output the decision trees which were used in the final decision. The individual trees can be used to understand what the important decision nodes were, as well as prompt questions around what led to the final prediction.
4. Easy to use. Random forest works with both categorical and numerical input variables, so you spend less time one-hot encoding or labeling data. It’s not sensitive to missing data, and it can handle outliers to a certain extent. Overall, it saves you time that would otherwise be spent [cleaning data](https://www.keboola.com/blog/logistic-regression-machine-learning), which is usually the biggest step in any data science pipeline. This doesn’t mean that you should skip the cleaning stage entirely: you will often obtain better performance by working the data into an appropriate shape. But random forest does make it easier to use and faster to deploy to reach the base model.

## SVR model

Support Vector Machine (SVM) classification and Support Vector Regression (SVR) are both non-parametric. SVR's ability to build a non-linear model without altering the explanatory variables is another benefit. To prevent over-fitting, the regression can also be penalized using a cost parameter.

## What is the SVR model?

Support Vector Regression is a supervised learning algorithm that is used to predict discrete values.

Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the **hyperplane** that has “the maximum number of points”.

The Gaussian-RBF kernel is the best method because it produces a stable MAPE value

Support Vector Regression has the advantage of making accurate cryptocurrency price predictions and can overcome the problem of overfitting by itself, because standard SVM tries to separate all positive and negative examples (i.e., two different classes) and does not allow any points to be misclassified.

## Advantages of the SVR model

The SVR model is a support vector machine-based model specifically designed for regression problems. It has several advantages and benefits over other models, such as:

1. It is highly accurate compared to other models.
2. It is fast, accurate and reliable in predicting complex and nonlinear data.
3. It is robust to outliers and noise.
4. It can model non-linear functions easily, without the need to approximate them.
5. It is highly stable and has good generalization performance.
6. It can handle multi-dimensional data easily and can also provide non-linear solutions.

Because of these advantages, SVR model is widely used in prediction of cryptocurrency projects. It is powerful and flexible and provides better accuracy than other models for predicting the cryptocurrency market

## Disadvantages of SVR model

## Despite of good theoretic foundations and high classification accuracy of support vector machines (SVM), normal SVM is not suitable for classification of large data sets, because the training complexity of SVM is very high.

1. Advanced Settings
2. Suitable for Small Dataset
3. Costly Computation, compared to some algorithms such as random forests, decision trees are a lighter alternative.
4. Feature Vectors Required
5. Low Interpretability
6. Overfitting Risk

So based on our project this model was the most suitable.

## Linear Regression

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x).

# Simple implementation

Linear Regression is a very simple algorithm that can be implemented very easily to give satisfactory results. Furthermore, these models can be trained easily and efficiently even on systems with relatively low computational power when compared to other complex algorithms. Linear regression has a considerably lower time complexity when compared to some of the other machine learning algorithms. The mathematical equations of Linear regression are also fairly easy to understand and interpret. Hence Linear regression is very easy to master.

### Performance on linearly separable datasets

Linear regression fits linearly separable datasets almost perfectly and is often used to find the nature of the relationship between variables.

### Over fitting can be reduced by regularization

**Over fitting** is a situation that arises when a machine learning model fits a dataset very closely and hence captures the **noisy data** as well. This negatively impacts the performance of model and reduces its accuracy on the test set.  
**Regularization** is a technique that can be easily implemented and is capable of effectively reducing the complexity of a function so as to reduce the risk of overfitting.

## Lasso Model

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

## Why lasso

**Lasso tends to give sparse weights (most zeros),** because the l1 regularization cares equally about driving down big weights to small weights, or driving small weights to zeros. If you have a lot of predictors (features), and you suspect that not all of them are that important, Lasso may be really good idea to start with.

## The Advantages of Lasso

The biggest pro of LASSO is that it is better than the usual methods of automatic variable selection such as forward, backward and stepwise - all of which can be shown to give wrong results. The results from LASSO are much better.

## The Disadvantages of Lasso

The biggest con of LASSO is that it is automatic; therefore, it has problems. The biggest problem is that it lets you (the data analyst) avoid thinking. Other, lesser problems:

It can also produce models that make no sense.

It ignores nonsignificant variables that may, nevertheless, be interesting or important

# Model evaluation and result

Models Accuracy

Random Forest Regression 94.374 %

Linear Regression 0.9368561604691311

Linear model. Lasso 0.9370164461642735

Support Vector Regression (SVR). 0.8593135496753768

K-Neighbors Regressor. 0.9485867559032419

# Resources:

1) Sin, E.; Wang, L. Bitcoin price prediction using ensembles of neural networks. In Proceedings of the 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Guilin, China, 29–31 July 2017; pp. 666–671. [Google Scholar]

2) Yenidoğan, I.; Çayir, A.; Kozan, O.; Dağ, T.; Arslan, Ç. Bitcoin forecasting using ARIMA and PROPHET. In Proceedings of the 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, Bosnia and Herzegovina, 20–23 September 2018; pp. 621–624. [Google Scholar]

3) Jiang, X. Bitcoin price prediction based on deep learning methods. J. Math. Financ. 2019, 10, 132–139. [Google Scholar] [CrossRef]

4)<http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-knn.html>

5)https://www.kdnuggets.com/2017/03/building-regression-models-support-vector-regression.html