

Data mining assignment 1

by Vikram Kalta

Submission date: 09-Mar-2023 08:26PM (UTC+0000)

Submission ID: 198924081

File name: 42274_Vikram_Kalta_Data_mining_assignment_1_482892_1701812030.pdf (63.28K)

Word count: 707

Character count: 3715

Vikram Kalta (2209281)

March 9, 2023

1 Summary

This report aimed to analyse a dataset taken from UCI Machine Learning Repository. The dataset namely "CNNpred: CNN-based stock market prediction using a diverse set of variables Data Set" is selected by the author for this report. As the name suggests the dataset contains information about the stock market indices.

1.1 Origin and content

The dataset stores information about financial markets. The dataset consists of 5 files in CSV format for each of the indices such as S&P 500, Dow Jones Industrial Average, NASDAQ Composite, RUSSELL 2000, and NYSE Composite. The main purpose of this data collection is to use the historical data and make prediction of stock market prices [1]. The dataset is collected from various sources as specified in this paper [2]. Some major sources are yahoo finance, investing.com and FRED. The dataset contains time-series data from 2010 to 2017.

1.2 Dataset suitability

The dataset is considered suitable after running preprocessing tasks mentioned in the report in detail.

- Data format: As evident almost all the variables are numeric type but a couple of variables. This made some of the pre-processing tasks simple as the tasks such as handling missing string or categorical data etc were overruled.
- Reliability: Although some variables contained missing values but with strategies such as rolling mean used and explained in the report were used to handle such conditions with appropriate techniques such that the dataset can be prepared well for the final task.
- Consistency: Being a time series data, the data is collected over a time period from 2010 to 2017. This indicates strong consistency in the collection of the data.

- Relevancy: Relevancy of features other than the primitive features were proved by finding correlations in the dataset amongst the closing price of S&P500 index and closing prices of various other indices, technical indicators, commodities etc.
- Dataset size: The size of this dataset which is small worked in the favour of the suitability of the dataset.

1.3 Objective of the analysis

The objective of this analysis was to find the meaningful features which are correlated with the primitive feature such as closing price of S&P 500. This was done using various pre-processing techniques. Such establishments of hidden patterns and correlations within the dataset will help in training of the selected machine learning model which will then be used to predict the future stock market prices and aid the investors in achieving their goals. To achieve this objective the report highlights the use of various tools such as JASP and TABLEAU to get better insights of the data. The report also uses a python script for handling missing values with the help of rolling mean strategy. A technique popularly used for dimensionality reduction known as Principal Component Analysis [3] is also applied to find the features which accounts for maximum variance.

1.4 Visualization

This report highlighted some visuals of the dataset using Tableau. The report showed closing prices between years 2010 and 2017. The report also tried to look for various patterns when compared with other features in an attempt to look for correlations with the primitive feature. As was evident the week wise chart for year 2017 showed the formation of the trends. Some of which could be essential in configuring the final trained model and deciding the buy/sell signal [4]. The report also highlighted the high correlation found amongst closing price and DAAA, DBAA variables.

1.5 Research background

The author intends to research the stock market such that the knowledge gained in this module could be practiced and applied in this domain. Having said so the author has been collecting data and knowledge from various sources to support his study of the stock market.

1.6 Video link

[Video explanation for the report](#)

References

- [1] Ehsan Hoseinzade, Saman Haratizadeh, ‘CNNpred: CNN-based stock market prediction using a diverse set of variables.’ 2018
- [2] Ehsan Hoseinzade, Saman Haratizadeha, Arash Khoeini, “U-CNNpred: A Universal CNN-based Predictor for Stock Markets“, 2019
- [3] Herve Abdi and Lynne J. Williams “Principal component analysis“,
- [4] Shalini Talwar, Utkarsh Shah, Pranav Shah ”PICKING BUY-SELL SIGNALS: A PRACTITIONER’S PERSPECTIVE ON KEY TECHNICAL INDICATORS FOR SELECTED INDIAN FIRMS”, 2019

Data mining assignment 1

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

Instructor

63/100

PAGE 1

PAGE 2

PAGE 3

GRADING FORM: DATA MINING/ VISUALISATION

VIKRAM KALTA

63

KNOWLEDGE AND UNDERSTANDING

Dataset Suitability (What features of the dataset lend themselves to data mining? Has it been mined previously? Why does the student feel that they have the contextual background to work with this data?)



Financial market based set of data was examined and is valid for further analysis. However, prior analysis of said data is never made explicit in

22

submission. Firm rationale for dataset choice is given and some information is given concerning prior knowledge of stock markets generally.

COGNITIVE AND INTELLECTUAL

Objectives (What does the student plan to investigate/expect to find? Are these SMART?)



Very limited expectations are listed in submission. In other words, having evaluated your chosen dataset, what are you hoping to discover?

6

PRACTICAL APPLICATION SKILLS

Dataset Description (Where did the data come from? What does the data describe? Is the sample large enough? What data formats and types are involved?)



Limited description of the incorporated dataset. Its location is depicted but description of actual variables are very minimal. Data format within confines of showcased dataset is explained well.

28

TRANSFERABLE SKILLS FOR LIFE

Presentation/Slides (Clarity, readability, flow, accuracy, completeness, performance)



Well formatted document with clearly listed descriptors of key themes. Submission is also written in an articulate manner. Criteria met.

7