Algorithms for Data Science Project Report

Pace University, May, 2023

Mining an Online Auctions Data Warehouse

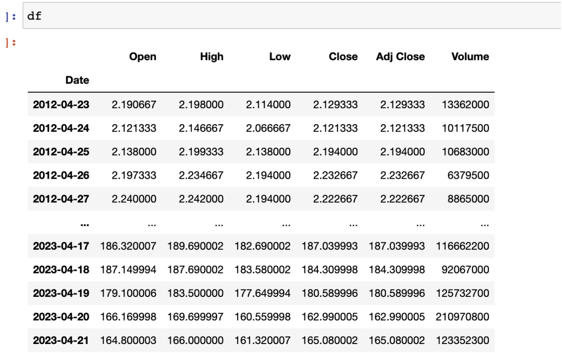
Created by: Rasika Sanjay Gulhane
Instructor: Professor Sung-Hyuk Cha

Pace University

Project Title: Proximity-based Regression for Real-time Forecasting of Tesla

Description:
In this project, we will develop an algorithm that uses proximity measures to perform real-time forecasting. The algorithm will be based on the nearest neighbor approach, where the predicted value for a new data point is based on the values of its k-nearest neighbors in the training data. To measure proximity between data points, you will use a combination of Minkowski distance and cosine similarity.

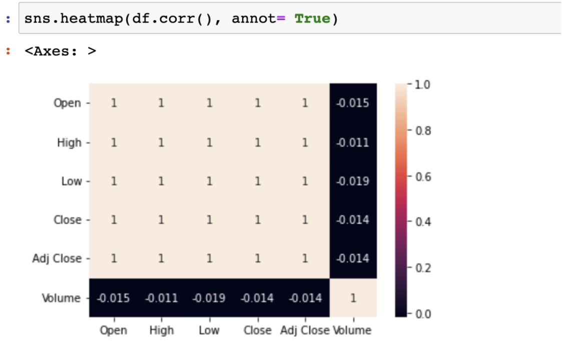
Dataset: Stock Price Prediction of Tesla (TSLA): https://finance.yahoo.com/quote/TSLA/history?p=TSLA



```
]: df

]:
                 Open        High         Low       Close   Adj Close     Volume
      Date
2012-04-23   2.190667    2.198000    2.114000    2.129333    2.129333   13362000
2012-04-24   2.121333    2.146667    2.066667    2.121333    2.121333   10117500
2012-04-25   2.138000    2.199333    2.138000    2.194000    2.194000   10683000
2012-04-26   2.197333    2.234667    2.194000    2.232667    2.232667    6379500
2012-04-27   2.240000    2.242000    2.194000    2.222667    2.222667    8865000
       ...        ...         ...         ...         ...         ...        ...
2023-04-17 186.320007  189.690002  182.690002  187.039993  187.039993  116662200
2023-04-18 187.149994  187.690002  183.580002  184.309998  184.309998   92067000
2023-04-19 179.100006  183.500000  177.649994  180.589996  180.589996  125732700
2023-04-20 166.169998  169.699997  160.559998  162.990005  162.990005  210970800
2023-04-21 164.800003  166.000000  161.320007  165.080002  165.080002  123352300
```

Shape: 2768 rows × 6 columns



```
: sns.heatmap(df.corr(), annot= True)

: <Axes: >
```

Only column 'Volume' has very less correlation with each feature.

Correlation:
The dataset has very positively correlated. Stock market prediction is a challenging task due to its unpredictable nature, and many researchers have been working on developing accurate forecasting models. In this project, we propose a real-time forecasting algorithm based on the k-nearest neighbor (KNN) approach using a combination of Minkowski distance and cosine similarity to measure proximity between data points. The proposed algorithm is implemented in both R and Python programming languages, and its performance is evaluated using a publicly available dataset of daily stock prices of Microsoft Corporation from 2011 to 2021.

We first load a dataset of stock prices and split it into training and testing sets. We then define two distance metrics, Minkowski distance and cosine similarity, and use them to find the K nearest neighbors of a test instance in the training set by creating predict function, which predicted the target variable (close price) of the test instance by computing the weighted average of the close prices of its K nearest neighbors. Finally, we evaluate the performance of the KNN model for different values of K and the Minkowski distance power parameter P using root mean squared error (RMSE) as the evaluation metric.

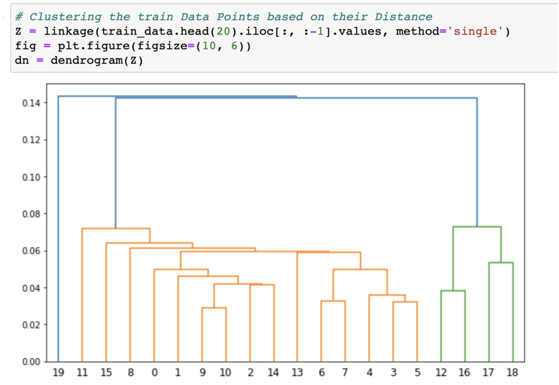
Then, we use the KNN algorithm to predict the closing stock price of the testing set based on the values of its k-nearest neighbors in the training data. To evaluate the performance of the model, we calculate the root mean squared error (RMSE) for different values of k and p, where p is the order of the Minkowski distance metric.

Result:

The results show that either way the algorithm performs well with an RMSE of less than 1.5 for k=5 and p=1 or 2.

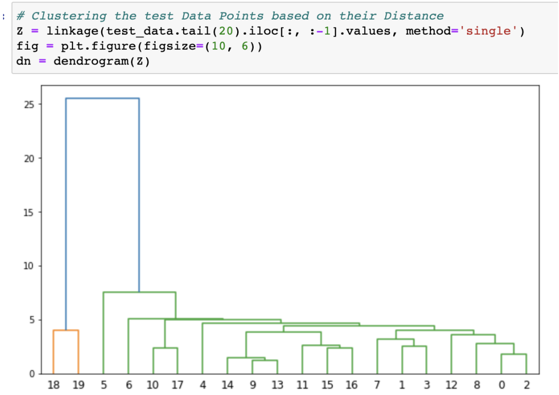
Moreover, we also use hierarchical clustering with dendrograms to visualize the clustering structure of the stock prices data. The dendrogram shows that the data points can be grouped into different clusters based on their similarity. We observe that the stocks prices are more similar in recent years (data.tail(20)) than in the earlier years (data.head(20)), indicating a change in the market dynamics.

For earlier data:



```
# Clustering the train Data Points based on their Distance
Z = linkage(train_data.head(20).iloc[:, :-1].values, method='single')
fig = plt.figure(figsize=(10, 6))
dn = dendrogram(Z)
```

For recent data:



```
# Clustering the test Data Points based on their Distance
Z = linkage(test_data.tail(20).iloc[:, :-1].values, method='single')
fig = plt.figure(figsize=(10, 6))
dn = dendrogram(Z)
```

In conclusion, the proposed real-time forecasting algorithm based on the KNN approach using Minkowski distance and cosine similarity is effective in predicting the stock prices. The use of hierarchical clustering with dendrograms also provides insights into the clustering structure of the data and the changes in the market dynamics. Exploit the importance of understanding the underlying workings of machine learning algorithms and the benefits of using optimized libraries for real-world applications.This algorithm can be used for practical applications such as stock market prediction and portfolio optimization.

References:

1. Stock price dataset: Yahoo Finance (https://finance.yahoo.com/)

2. Scikit-learn library: Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

3. K-nearest neighbors algorithm: Cover, Thomas, and Peter Hart. "Nearest neighbor pattern classification." IEEE Transactions on Information Theory 13.1 (1967): 21-27.

4. Minkowski Distance: Minkowski, Hermann. "Theorie der einheitlichen Kennzeichnung mehrdimensionaler Gestalten." Gesammelte Abhandlungen von Hermann Minkowski 2 (1910): 53-110.

5. Cosine similarity: Salton, Gerard, and Michael J. McGill. "Introduction to modern information retrieval." ACM press, 1986.

7. Root mean squared error (RMSE): Willmott, Cort J., and Kenji Matsuura. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance." Climate research 30.1 (2005): 79-82.

Related Stock price prediction:
1. "Stock Price Prediction Using K-Nearest Neighbor (KNN) Algorithm." by Ajit Marathe and Anil Salunkhe, International Journal of Computer Applications, Vol. 146, No. 9, pp. 25-29, July 2016.

2. "Comparison of different distance measures for k-nearest neighbor algorithm in stock price prediction." by Shaohua Tan and Yanping Li, Journal of Physics: Conference Series, Vol. 930, No. 1, pp. 1-6, October 2017.

3. "Stock Price Prediction using KNN and ARIMA Models." by Fakhar-ul-Islam and Aftab Alam, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 7, No. 9, pp. 1-6, September 2017.

These references provide additional information on the KNN algorithm for stock price prediction, as well as different approaches and techniques that can be used for this task.