Coca Cola (KO) Stock Price Prediction Modelling Project

10th February 2022

1 Introduction

In this report I elaborate on my problem formulation for predicting the Coca Cola's next day's closing stock price of the Coca Cola Stock (stock market ticker: 'KO').

The Coca Cola Company is a multinational beverage corporation involved in manufacturing, retailing, and marketing of the aforementioned. It has a history of more than a century in which it has managed to become a household brand in many regions of the world.

However, the specific niche of beverages The Coca Cola Company is involved in is heavily reliant on two ingredients: sugar and caffeine (derived from coffee beans and coffee's stock market ticker: 'KC=F'). From the lens of Finance, there exists a hint at the relation between commodity stock price of Sugar (stock market ticker: 'SB=F') and KO as stated in an article by Stevenson (2017), however, little literature is found on the impact of commodity (sugar and coffee) prices on the KO stock.

2 Problem Formulation

The problem I intend to advance in is: given historical stock data of the KO stock, can the next day's closing price be predicted accurately? If so, do average price of sugar and coffee improve the stock price prediction of KO?

The **datapoints** of the problem are the stock market data on a single day. The stock market data will be having the following instances/characteristics (data type's are mentioned in square brackets): KO's Opening Price ($) [float], KO's Highest Price ($) [float], KO's Lowest Price ($) [float], KO's Closing Price ($) [float], KO's Volume (No.) [integer], KC=F's Opening Price ($) [float], KC=F's Closing Price ($) [float], KC=F's Average Price ($) [float], SB=F's Opening Price ($) [float], SB=F's Closing Price($) [float], SB=F's Average Price ($) [float]. One datapoint has the characteristics of the stock market data as its candidate **features**. The KC=F's Average Price ($) and SB=F's Average Price ($) feature will be calculated using their historical stock market data by implementing the formula: (opening price + closing price)/2. The resulting average price will be used in as a feature.

The target value is KO's Next Day's Closing Price ($) [float], which is the **label** of a datapoint.

3 Methods

3.1 Methods: Dataset

The range of the set of datapoints we will use is from 1st of January 2010 to 1st of January 2020 (exactly 2513 datapoints), excluding the datapoints with missing features. Our sources of data are as follows:

- KO's Historical Stock Market Data: Yahoo! Finance (Yahoo, 2022c)
- SB=F's Historical Stock Market Data: Yahoo! Finance (Yahoo, 2022b)
- KC=F's Historical Stock Market Data: Yahoo! Finance (Yahoo, 2022a)

3.1 Methods: Feature Selection

After **visualizing** the data with **scatterplots**, as displayed in Figure 2, a few of the features will be dropped from the dataset. We additionally, verified this using **correlation** statistically through Pearson's correlation using the *f_regression()* function and *SelectKBest* class, as shown in Figure 1. While, KO's opening, highest, lowest, and closing price does correlate highly with the label column, KO's volume is highly uncorrelated to it and we are not doing a weighted financial stock prediction model where volume traded's impact needs to be taken into account,

```
F-Scores of Features:
Feature 0: 343324.660144
Feature 1: 433355.425698
Feature 2: 424892.758577
Feature 3: 537690.818394
Feature 4: 332.282862
Feature 5: 1625.977702
Feature 6: 1628.953765
Feature 7: 1631.258063
Feature 8: 1046.145036
Feature 9: 1051.016127
Feature 10: 1050.276037
```

Figure 1: F-Score's of Features (in the order mentioned in section 2)

which is why we will not be selecting this. Moreover, if we have taken into account KC=F's and SB=F's average prices, we will be dropping their opening and closing prices because they are highly correlated, and intuitively, it is a repetition of columns because we have calculated average prices using the opening and closing prices.

Additionally, we also notice that KC=F's Average Price and SB=F's Average Price has less stronger correlation with our label but we will be investigating this further since it is part of our problem formulation to find out whether they impact KO's prices or not.

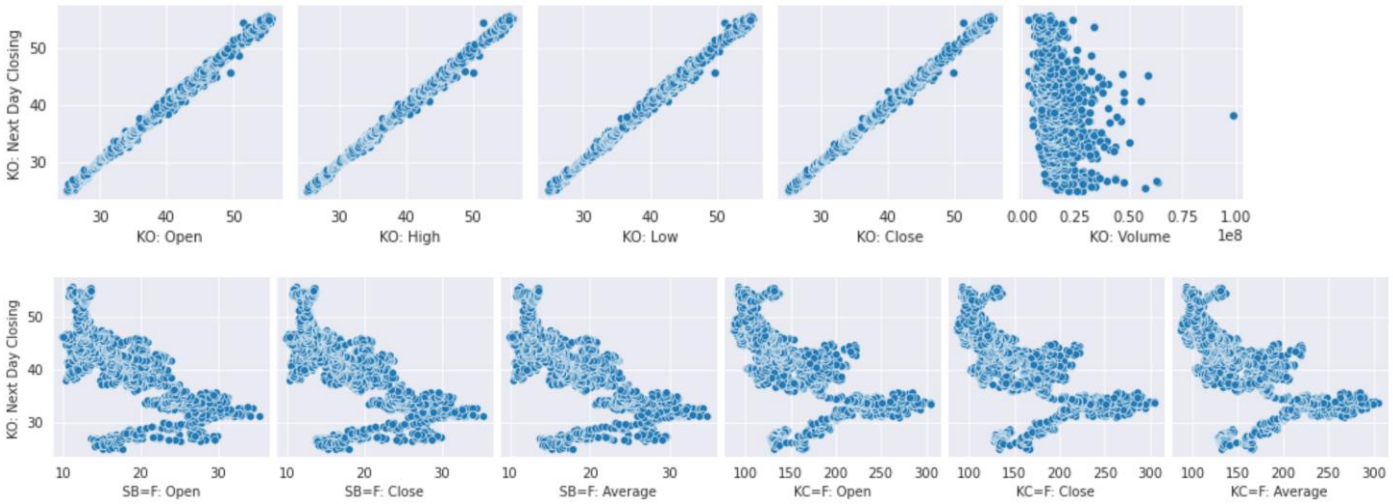

Figure 2: Visual Scatterplots with respect to the Label

3.2 Methods: Visual Summary of Datapoint (Example)

Datapoints	**Features**						**Label**
Date	KO: Open	KO: High	KO: Low	KO: Close	KC=F: Average	SB=F: Average	KO: Next Day's Closing Price
05/01/10	60.00	61.12	59.88	60.60	250.74	18.45	59.97

3.3 Methods: Model (Hypothesis Space) & Motivation

Linear Regression model is used because a mostly linear relationship can be observed between features and label. Moreover, this model is used because we want to predict a continuous dependent variable (label) from a number of independent continuous variables (features). Additionally, it also allows polynomial terms to model curvature.

3.4 Methods: Loss Function Used & Motivation

The mean squared error (MSE) loss is chosen as it allowed the use of a ready-made library for linear regression. The loss is the mean of the squared differences between the true and predicted values.

3.5 Methods: Design Choice

The design choice for the model had an 80/20 split between the training and validation set. I chose such a split ratio because it is common for such ML models related to stock prediction to use such relatively large set for training and a smaller set for validation. Another reason for such a split are the reasonings given in the scientific report by Gholamy, Kreinovich and Kosheleva (2018).

References

Gholamy, A., Kreinovich, V. and Kosheleva, O. (2018). Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. [online] ScholarWorks@UTEP. Available at: https://scholarworks.utep.edu/cs_techrep/1209/ [Accessed 10 Mar. 2022].

Stevenson, A. (2017). *Cramer's charts predict Coca-Cola on the rise thanks to the price of sugar*. [online] CNBC. Available at: https://www.cnbc.com/2017/01/31/cramers-charts-predict-coca-cola-on-the-rise-thanks-to-the-price-of-sugar.html [Accessed 10 Feb. 2022].

Yahoo (2022a). Coffee May 22 (KC=F) Stock Historical Prices & Data - Yahoo Finance. [online] Yahoo! Finance. Available at: https://finance.yahoo.com/quote/KC%3DF/history?period1=1262304000&period2=1577836800&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true [Accessed 1 Mar. 2022].

Yahoo (2022b). Sugar #11 May 22 (SB=F) Stock Historical Prices & Data - Yahoo Finance. [online] Yahoo! Finance. Available at: https://finance.yahoo.com/quote/SB%3DF/history?period1=1262304000&period2=1577836800&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true [Accessed 1 Mar. 2022].

Yahoo (2022c). The Coca-Cola Company (KO) Stock Historical Prices & Data - Yahoo Finance. [online] Yahoo! Finance. Available at: https://finance.yahoo.com/quote/KO/history?period1=1262304000&period2=1577836800&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true [Accessed 1 Mar. 2022].

Appendix