November 22, 2022
Ai Hua Li
Saiful Islam
Ziyi Huang

Buy Low, Sell High - Preliminary Analysis

Data Cleaning Code

Code for cleaning and processing our data is found here:
https://github.com/AiHuaLi-CS/CSC460_Final_Project/blob/main/data-cleaning.py

Data dictionary for our transformed dataset:

Column	Data Type	Description	Example
Volume	int	The amount of shares or contracts traded in an asset or security over a period of time, usually over the course of a trading day.	115539
Open	float	The price at which a security first trades upon the opening of an exchange on a trading day	3899
High	float	The highest intraday price of a stock in the most recent (or current) trading session	1219.60
Low	float	The lowest intraday price of a stock in the most recent (or current) trading session	1205.00
Close	float	The price at which a security first trades upon the closing of an exchange on a	1211.45

November 22, 2022
Ai Hua Li
Saiful Islam
Ziyi Huang

		trading day	
Datetime	date	Current Datetime in yyyy-mm-dd-hh	2022-08-25 10:00:00
Hour	int	Current hour (10 am to 16 pm)	10
Profit	float	Calculated by Close - Open	0.4500

Exploratory Analysis

Code for exploratory analysis is found here:
https://github.com/AiHuaLi-CS/CSC460_Final_Project/blob/main/exploratory-analysis.ipynb

So far, we were able to read, process, add, update, and drop data. We read our data from an API and processed our data by getting it into a tabular, structured format. We dropped columns of data that we don't need for the anticipated model and dropped data that is outside of regular market hours as our focus is for intervals during regular market hours only. Our data doesn't contain NaN values, so there was no work needed to be done in terms of dealing with rows or columns that have NaN values.

We created a profit column that indicates the profit during each interval, renamed columns to make the data more understandable, and updated the Timestamp column to a more usable format (datetime).

Descriptive statistics are found at the end of the notebook. We created a graph that shows the profit per interval. It is what we expected but there were some intervals that gave dramatic profits or losses, which could be due to certain news. For the most part, the profit fluctuates, as expected. We created a heatmap that shows the correlations between the variables. Moreover, we can predict one variable from the other using correlation. Therefore, we created a correlation heat map to help us visualize which columns were correlated to each other. However, we found out that there's a -0.013 correlation between hour and profit.

Challenges

We solved the challenges we encountered thus far. While we do have an idea of what type of model to use, our next challenge is to find the appropriate one. The current challenge that we

have is that there's -0.013 correlation between hour and profit gain which it's technically a zero correlation. We need to identify whether there are additional features that affect the profit.

Future Work

For the final analysis, we plan to create, train, and test the model. We will split the data by 30% and 70% initially and make adjustments based on the accuracy, precision, recall and F1 score of the classification model. We will start with Logistic Regression, Decision Tree and Random Forest models to find the appropriate model.

Contributions

Ai Hua Li

Ai Hua added challenges and future work sections for the preliminary analysis.

Saiful Islam

Saiful made the code more reproducible so that relevant, updated data is used every time the notebook is run (data from past 90 days). He also created and dropped data where relevant. Finally, he put together this document.

Ziyi Huang

Ziyi performed the initial data loading and cleaning, made the .py file.