Task : Clustering based identification

Data: Date and time, OCHL, volume , open interest

Deciding Features (X):

- ☐ Normalization of position of the OHCL candle:
  - ☐ Find the moving mean of past n candles (close, open , 0.5(C+O)) - used close, close is generally chosen
  - ☑ ~~(O, C, H, L) - moving mean~~
  - ☑ ~~Vedika: use exponential moving average instead of simple moving average. EMA is better for short term markets, suitable for the per-minute prediction.~~
  - ☑ ~~Normalisation works as expected, verified using visualisation~~
- ☑ ~~Moving mean~~
- ☑ ~~Gradients 1st, 2nd …. Of moving mean~~
- ☑ ~~Volume …. Their gradients~~
- ☑ ~~Differences between OHLC <> OHCL~~
- ☑ ~~Time of candle (hour + min)~~
- ☑ ~~Week day~~
- ☑ ~~Think of more features: gradients, frequencies (maybe 30)~~
  - ☑ ~~Moving average convergence/divergence (MACD) = 26 period EMA - 12 period EMA~~
  - ☑ ~~SMA can also be used in addition to the EMA, 2-3 SMAs and EMAs can be used, over different periods~~
  - ☑ ~~On balance volume (OBV)~~
  - ☑ ~~accumulation/distribution indicator (A/D)~~
  - ☐ Momentum indicators
    - ☑ ~~RSI (relative strength index)~~
    - ☐ Stochastic oscillator
    - ☑ ~~Average directional index (ADX)~~
    - ☑ ~~Dynamic momentum index~~
    - ☐ Directional movement index
    - ☑ ~~Ultimate oscillator~~
    - ☑ ~~Disparity index~~
  - ☐ Economic indicators
    - ☐ More useful for long term investments
    - ☐ Economic cost index (ECI)
    - ☐ GDP
    - ☐ Purchasing managers’ index (PMI)

- [ ] Personal consumption expenditures (PCE)
- [ ] Also include seasonality, and inflation indicators, like PPI, CPI…
- [ ] Coincident indicators
  - [ ] Similar to economic indicators, not useful in this context
  - [ ] GDP and employment figures
- [ ] Lagging indicators
  - [ ] Similar to economic indicators
  - [ ] GDP
  - [ ] CPI
  - [ ] Balance of Trade (BOT)
  - [x] ~~Moving average crossovers~~
    - [x] ~~200 period and 50 period SMA crossover~~
- [ ] Technical indicators have been observed to not work for us

- [ ] ESN (echo state networks) can be used for prediction
- [ ] LSTM is commonly used for stock price prediction
- [ ] Look for features that can help
  - [x] ~~Perform more intensive EDA, use boxplots, (preferably) pdf plots and heatmaps~~
    - [x] ~~Plotted distribution plots for starter data, more variables need to be introduced~~
- [x] ~~Vedika: insights from EDA~~
  - [x] ~~Start and the end of the work week witness high volatility (spikes and drops in C-O)~~
  - [x] ~~First two hours of the market are volatile, followed by increasing tranquility~~
  - [x] ~~Midweek trading volume is high~~
  - [x] ~~Midday trading volume is low~~
  - [x] ~~Relationship among volume, C-O, and open interest needs to be looked at - not useful~~
  - [x] ~~EMA vs open interest is a complicated graph - unlikely to be useful~~
  - [x] ~~High H and low L point to volatility~~

- [ ] Clustering part: k-mean :
  - [x] ~~Clustered on EMA, C-O, volume, and open interest~~
    - [x] ~~Scatter plot didn't provide clarity~~
    - [x] ~~Help needed for numerical analysis of the clusters~~
- [ ] H-clustering
- [ ] Latent space VAEs
- [ ] Look for similar projects to cluster candlesticks, try implementing them
  - [ ] Look for insights from a single candlestick, or a group of candlesticks
  - [ ]

- ☐ 6 clusters:
    - ☐ 1st cluster has candles from 2018 and 2019
    - ☐ 2nd cluster also has candles from 2023 and 2024
    - ☐ 3rd cluster has candles from 2020 end and 2021
    - ☐ 4th cluster has candles from 2017 and 2018 start
    - ☐ 5th cluster has candles from 2022 and 2023 start
    - ☐ 6th cluster has candles from 2019 and 2020
- ☐ The economic conditions during those periods are leading to these clusters
    - ☐ 1: fluctuation in mid range - hold
    - ☐ 2: set up trend, high prices - good y value
    - ☐ 3: increasing trend - optimistic
    - ☐ 4: low prices, upward trend - not the right time to invest, keep a look
    - ☐ 5: very volatile in the medium high range - better to hold
    - ☐ 6: sharp drop from medium to low prices - avoid
- ☐ Comparing the trends to the time frame
    - ☐ 2018-19: Rafale controversy, LTCG tax
    - ☐ 2023-24: election season, euphoria of strong campaigning
    - ☐ 2020-21: vaccines, first wave had ebbed
    - ☐ 2017-18: economic decisions, GST, market had started gaining strength
    - ☐ 2022-23: wars
    - ☐ 2019-20: pandemic - sharp drop
    - ☐ Events are in conformity with the trends
- ☐ Engineered features can also be used for clustering

Let's say we get k no of clusters
KPI: identify the clusters with better Y var.

Deciding the Y var. (close - open)

- ☐ Occurrence of Green/ Red in n future candles (categorical type)
    - How many time it is green and red, neutral
    - Also decide the green/red based on threshold
- ☐ Amount of change in the open and close - for n future candles (numerical)

Required a table:

- Cluster number - Y_cat - Y_num

Docs:

https://www.quantstart.com/articles/k-means-clustering-of-daily-ohlc-bar-data/
https://github.com/samuelclk/ethcandleclusters/blob/master/ethuat%20kmeans%20candlesticks%20250817.ipynb

Code NBs:

OptAlpha Candle Prediction.ipynb
OptAlpha Candle Prediction v2.ipynb

Work Logs

4/June/2024
- Found moving mean avg - simple and Exponential
- Difference between OHCL candles

5/June/2024
- Visualization of candles
- Started K-mean clustering

6/June/2024
- Tried on K-mean clustering
- Plotted variations of features

7/June/2024
- Looked for technical indicators
- Found potential ideas for prediction model

10/June/2024
- Plotted distribution plots
- Started looking for clustering ideas

11/June/2024

- Calculated gradients of features
- Looked for more features

12/June/2024

- Added more features

13/June/2024

- Looked for clustering ideas
- Searched for discontinuities in data

14/June/2024

- Removed discontinuous days
- Updated dataframes

17/June/2024

- Prepared clustering database (as per notebook)
- Plotted scatter plot

18/June/2024

- Fixed some df features found to contain NANs

19/June/2024

- Plotted mpf candlesticks
- Plotted go candlesticks

20/June/2024

- Plotted normalised candlesticks
- Preliminary insights were drawn from normalised and unconditioned ohlc plots

21/June/2024

- Deeper insights were sought from the snaps
- Additional clustering ideas can be looked for