Indhold

1 Dataindsamling

Planen er at indhente data fra 2018-01-01 til 2022-04-30. Vi vil træne på data fra 2018-01-01 til 2020-31-12, for senere at kunne teste vores modeller på data fra 2021-01-01 til dags dato.

Dukascopy (Schweizisk bank)

Vi indsamler historisk minut-data fra dukascopy. Vi ønsker at indsamle data fra

Indexer:

- DAX-indexet (Germany 40)
- FTSE (UK 100)
- Hong Kong (Hong Kong 40)
- NASDAQ (USA 100 Technical)
- S&P 500 (USA 500)

Obligationer:

- US T BOND

Valuta:

- EUR/USD

Råvarer:

- Kaffe
- Olie
- Gas

ETF'er:

- Emerging market ETF
- Growth og value ETF
- Real estate ETF

Cryptovaluta:

- Bitcoin/USD
- Ether/USD
- Cardano/USD

Forbrugerprisindex (Troels' opfordring)

Vi har lavet et datasæt for forbrugerprisindekset og alle de under kategorier, som anvendes til at bestemme det. Vi har omdannet `Time` til `Year` og `Month`, så vi kan matche med datoerne i de ovenstående investerings muligheder. Data kan findes i mappen:

`Google_Drev_data/Anden data`

Storebælts trafik

Vi har lavet et datasæt for Storebælts trafik, altså hvor mange og hvilke type biler, som kører over Storebælt. Vi har omdannet `Time` til `Year` og `Month`, så vi kan matche med datoerne i de ovenstående investerings muligheder. Data kan findes i mappen:

`Google_Drev_data/Anden data`

Kaggle competition (Japansk markedsdata)

2 Databehandling

Index'er

Fra dataudtrækket fås følgende kolonner:

`CET` `Hour` `Day` `Month` `Year` `Open` `High` `Low` `Close` `Volume`

Vi ønsker for ting handlet på børser (index, enkeltaktier, ETF'er) at flage åbningstider og omskrive data til dollar-bars.

For at flage åbningstider laves en ny kolonne, som markerer hvorvidt dette tidsrum er inden børsen åbner (pre-market), efter børsen er lukket (after-market) eller i børsens normale åbningsvindue.

Omskrivning af timeframe-data til dollarbar-data:
For at omskrive til dollarbars oprettes en ny kolonne `Mean_HL` som er den (estimerede) gennemsnitlige handelspris i tidrummet (minuttet/timen):

$$\texttt{Mean\_HL} = \frac{\texttt{High} + \texttt{Low}}{2}$$

Der laves en ny kolonne `Mean_OC`, som er gennemsnittet mellem `Open` og `Close`:

$$\texttt{Mean\_OC} = \frac{\texttt{Open} + \texttt{Close}}{2}$$

Derefter laves en ny kolonne `Total_transaction`, som angiver det beløb der er handlet for i det givne tidsrum (minut/time):

$$\texttt{Total\_transaction} = \texttt{Volume} \cdot \texttt{Mean\_HL}$$

Der skal nu bestemmes et **dollarbar-cap**, som skal være det beløb en enkelt dollarbar bliver dannet udfra. Ved at køre pandas kommandoen .describe() kan max{`Volume`} og max{`High`} ses. **dollarbar-cap** bliver da bestemt som produktet af de to tal afrundet til nærmeste *pæne* tal (for DAX'en fås 78 mia EUR, og der rundes af til 100 mia. EUR). Dette sikrer at der ikke kan være flere dollarbars pr. minut.

Der laves nu en ny række for hver gang den aggregerede `Total_transaction` når det givne **dollarbar-cap**, og værdierne (`Local time`, `Open`, `High`, `Low`, `Close` og `Volume` = **dollarbar-cap**) indsættes som den nye dollarbar-række:

- Alle tids variabler sættes til værdien for den række hvor **dollarbar-cap** nås (altså den sidste række):

- `Open` er for den første dollarbar `Open` fra den første timeframe-bar, og for de resterende `Mean_OC` fra den timeframe-bar hvor seneste **dollarbar-cap** blev nået.

- `Close` sættes til `Mean_OC` fra den timeframe-bar hvor **dollarbar-cap** nås.

- `High` sættes til max{`High`} fra de timeframe-bars der er brugt til at lave dollarbaren.

- `Low` sættes til min{`Low`} fra de timeframe-bars der er brugt til at lave dollarbaren.

- `Volume` bliver sat til summen over de observationer som bliver brugt til at lave dollarbaren. Dog tager vi kun den andel af den sidste observations `Volume` som ligger inden for `Dollarbar_cap`. Givet n er antallet af observation, som bliver brugt til at danne vores dollarbar, så vil `Volume` blive udregnet på følgende måde

$$p = \frac{\texttt{Dollarbar\_cap} - \sum_{i=1}^{n-1} (\texttt{Volume}_i) + \texttt{Mean\_HL}_n \cdot \texttt{Volume}_n}{\texttt{Mean\_HL}_n \cdot \texttt{Volume}_n}$$

$$\texttt{Volume\_dollarbar} = \sum_{i=1}^{n-1} (\texttt{Volume}_i) + p \cdot \texttt{Volume}_n$$

  hvor $(1-p) \cdot \texttt{Volume}_n$ kommer med i den næste dollarbar.

- `Monetary_Volume` sættes fast til **dollarbar-cap**.

- `Period` er tiden der er gået for at lave dollarbaren. Så hvis observation '10/02-2020 10.00', '10/02-2020 11.00', '10/02-2020 12.00', så er perioden 3 timer. Det er ikke helt korrekt, men det er bare for at lave en tidsfornemmelse over dollarbaren.

OBS: Når **dollarbar-cap** er nået videreføres det resterende beløb fra den timeframe-bar til den næste dollarbar.

Flere covariater

Udfra price-dataet eller dollarbar-dataet danner vi nye covariater.

- **Change of rate**, `ROC_x`, som er procent ændringen af lukningsprisen fra x antal minutter/timer/dage. Det regnes på følgende måde

$$\frac{\texttt{Close}_{today} - \texttt{Close}_{x\_old}}{\texttt{Close}_{x\_old}} \cdot 100.$$

  Vi gør det for 5/10/15/20 minutter/timer

- **Exponential moving average**, `EMA_x`, som er en type 'moving average', som vægter nuværende data points højere end ældre. Her svare x til dage.

$$k = \frac{\text{smoothing}}{\text{days} + 1} \texttt{Close}_{today} \cdot k + \texttt{EMA}_{yesterday} \cdot (1 - k)$$

  Vi gør det for 10/50/200 minutter/timer.

Price/Earning:
Vi tilføjer også månedlig P/E for vores indeks. Dataen har vi fundet lidt forskellige steder henne, men vi har sat datasættet op således, at vi har `Year` og `Month`, som vores tidsfaktorer og så også `Name`, så vi nemt kan merge P/E på vores indeks data.