Data Preprocessing Notes

Major Operations

1. Handling Negative Values in Stock Data
    - Stocks, especially in the PRC and OPENPRC columns, sometimes have negative values.
    - The forcePositives.py function in .src/helperFunctions/dataPreprocessing/ is used to convert these values to positive. The full implications of this conversion are yet to be fully evaluated.
2. Addressing NaN Values in PRC and OPENPRC
    - OPENPRC missing values are filled with the previous row's PRC.
    - PRC missing values are filled with the current row's OPENPRC, or if unavailable, the previous row's PRC.
    - Rows at the beginning with NaNs are removed until a complete row without NaNs is encountered.
    - This is done in .src/helperFunctions/dataPreprocessing/replaceNaNs.py
3. Dataset Operation Log
    - The major operations you make on the dataset are stored in a log in the tests folder.
    - You do not necessarily need to apply both, and is customizable upon init.

Open High Low Close (OHLC) Data Processing

1. Currently, OHLC values are computed using 5-minute return intervals.

Selection Criteria for Analysis

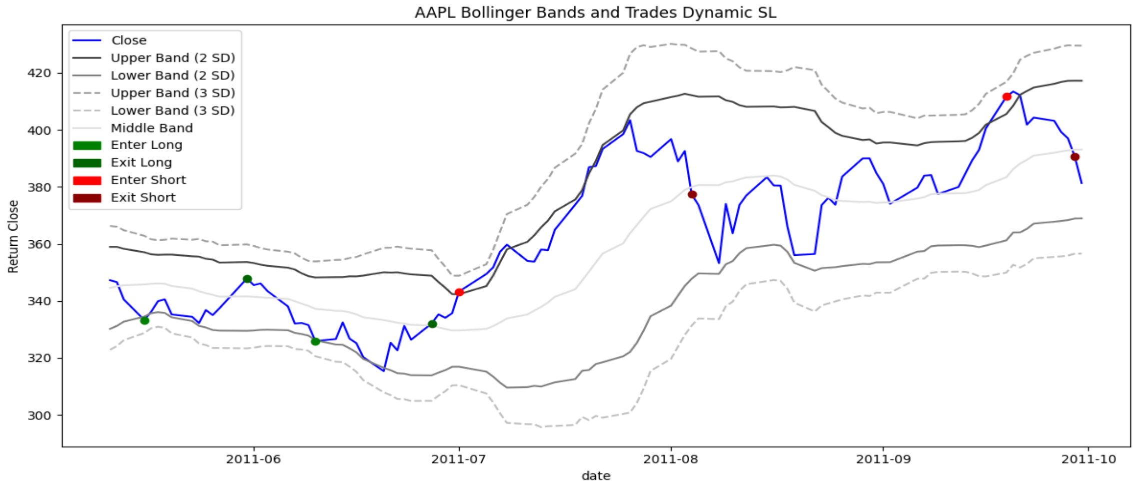
1. Currently, I first apply the NaN major operation, but not the positive, as I am unsure whether all negative values are collection errors or are a product of market illiquidity.
2. Stocks are chosen based on the availability of complete data from January 4, 2010, to December 31, 2020. This results in a total of 1,120 stocks for the study from the original 10,040. This range can be modified as well as the NaN/Positive requirement.

Non-Trades(Splits)

1. Non-trades are points of data in which a trade entry does not occur. They are calculated as follows.
2. Loop over every stock and generate an equal number of non-trades as there are trades for that stock.
    - Specify distances from trades and the splits in which to do. E.X. dist=(min=1, low=3, med=5, max=8), splits=(34, 33, 33)
    - This means a third of the non-trades will be between 1 and 3 days away inclusive, a third will be between 3 and 5 days away inclusive, and the last third will be between 5 and 8 days away inclusive.

Bollinger Naïve 1

Trade Cost = .2% per entry/exit



The Naive strategy enters on band highs/lows and exits on middle band / stop loss triggered at 3SD

Bollinger 1 is only able to trade 1 unit at a time.

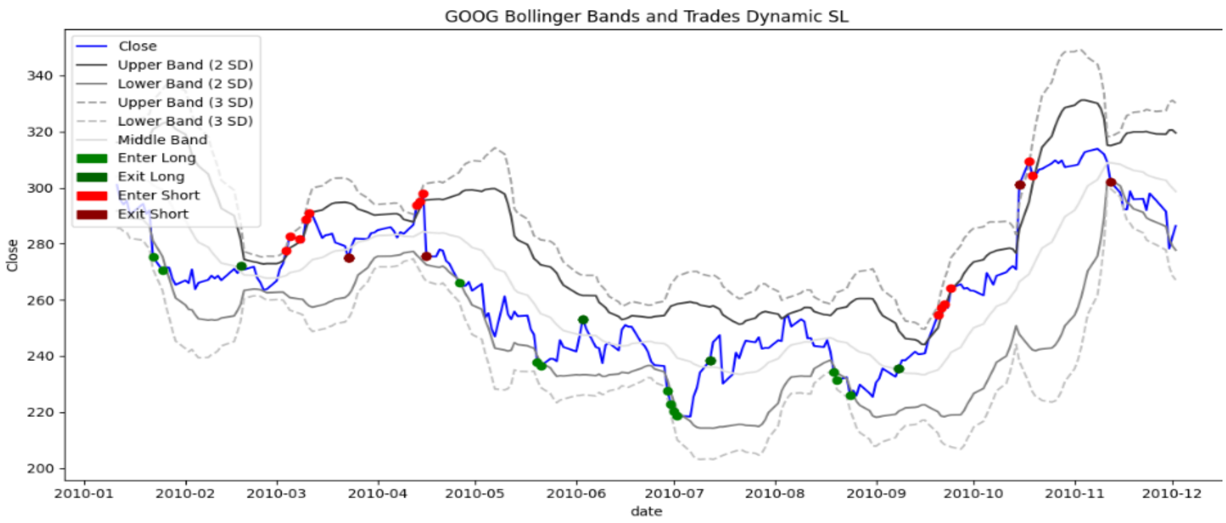
Bands are calculated as follows (Bperiods = 19 -> N = 20)

- Middle Band (MB): MB = (Sum of Close Prices over last N periods) / N)
- Standard Deviation (SD): SD = sqrt((Sum of (Close - MB)^2 over last N periods) / N)
- Upper Band (UB): UB = MB + (1.96 * SD)
- Lower Band (LB): LB = MB - (1.96 * SD)
- Upper Band 3 Standard Deviations (UB3SD): UB3SD = MB + (2.96 * SD)
- Lower Band 3 Standard Deviations (LB3SD): LB3SD = MB - (2.96 * SD)

	Factor	Total Trades	Different Stocks	Win Rate	Avg. Trade Return	Avg. Win on Trades	Avg. Loss on Trades	Max Trade Duration	Avg. Trade Duration	Total Return
Overall	Without Costs	125,554	1120	62.21%	0.595%	5.09%	-6.86%	170 days	18d 22h 55m	74743.49%
Overall	With Costs	125,554	1120	62.21%	0.20%	4.45%	-7.93%	170 days	18d 22h 55m	24,521.90%
LONG	Without Costs	58,449	1120	65.79%	0.86%	5.14%	-7.43%	127 days	17d 11h 12m	50,401.13%
LONG	With Costs	58,449	1120	65.79%	0.46%	4.52%	-8.60%	127 days	17d 11h 12m	27,021.53%
SHORT	Without Costs	67,105	1120	59.01%	0.36%	5.05%	-6.46%	170 days	20d 6h 2m	24,343.37%
SHORT	With Costs	67,105	1120	59.01%	-0.04%	4.34%	-7.44%	170 days	20d 6h 2m	-2,499.63%

Bollinger Naïve 2

Trade Cost = .2% per entry/exit



The Naive strategy enters on band highs/lows and exits on middle band / stop loss triggered at 3SD.

Bollinger 2 can trade multiple units at the same time.

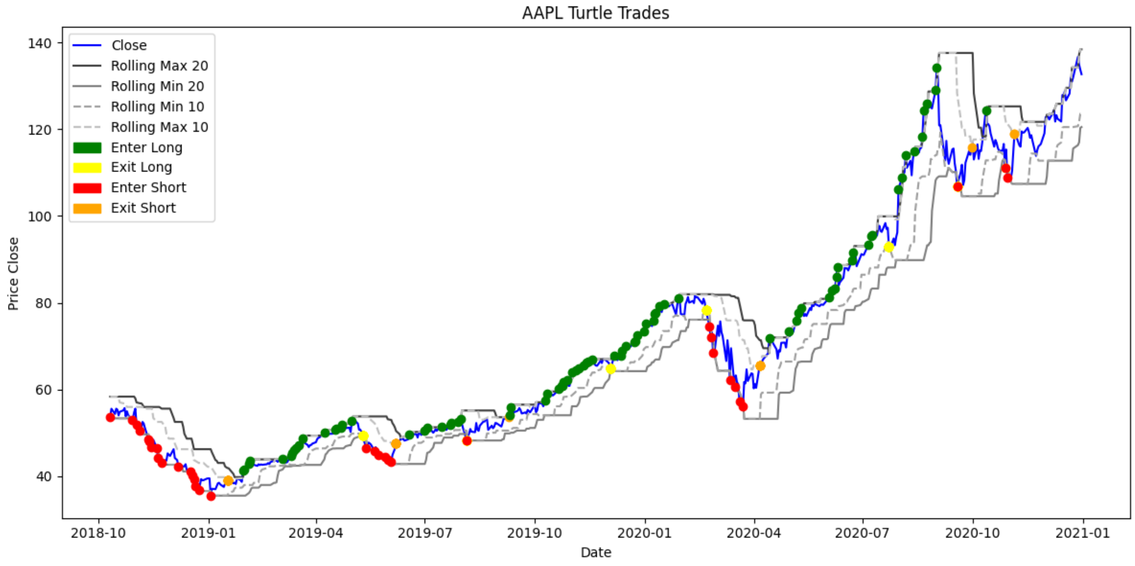
Bands are calculated as follows (Bperiods = 19 -> N = 20)

- Middle Band (MB): MB = (Sum of Close Prices over last N periods) / N)
- Standard Deviation (SD): `SD = sqrt((Sum of (Close - MB)^2 over last N periods) / N)
- Upper Band (UB): UB = MB + (1.96 * SD)
- Lower Band (LB): LB = MB - (1.96 * SD)
- Upper Band 3 Standard Deviations (UB3SD): UB3SD = MB + (2.96 * SD)
- Lower Band 3 Standard Deviations (LB3SD): LB3SD = MB - (2.96 * SD)

Trade Type	Factor	Total Units Traded	Different Stocks	Win Rate	Avg. Trade Return	Avg. Win on Trades	Avg. Loss on Trades	Max Trade Duration	Avg. Trade Duration	Total Return
Overall	Without Costs	350,546	1120	62.59%	0.71%	5.37%	-7.15%	170 days	20d 10h 25m	248,516.96%
Overall	With Costs	350,546	1120	62.59%	0.44%	4.94%	-7.88%	170 days	20d 10h 25m	152,885.69%
LONG	Without Costs	156,708	1120	65.70%	0.99%	5.74%	-8.21%	128 days	9d 0h 17m	155,170.38%
LONG	With Costs	156,708	1120	65.70%	0.71%	5.34%	-9.02%	128 days	9d 0h 17m	111,928.22%
SHORT	Without Costs	193,838	1120	60.07%	0.48%	5.04%	-6.42%	170 days	21d 14h 1m	93,346.58%
SHORT	With Costs	193,838	1120	60.07%	0.1%	4.59%	-7.10%	170 days	21d 14h 1m	40,957.47%

Turtle Naïve

Trade Cost = .2% per entry/exit



Entry based on 20-day highs or lows.

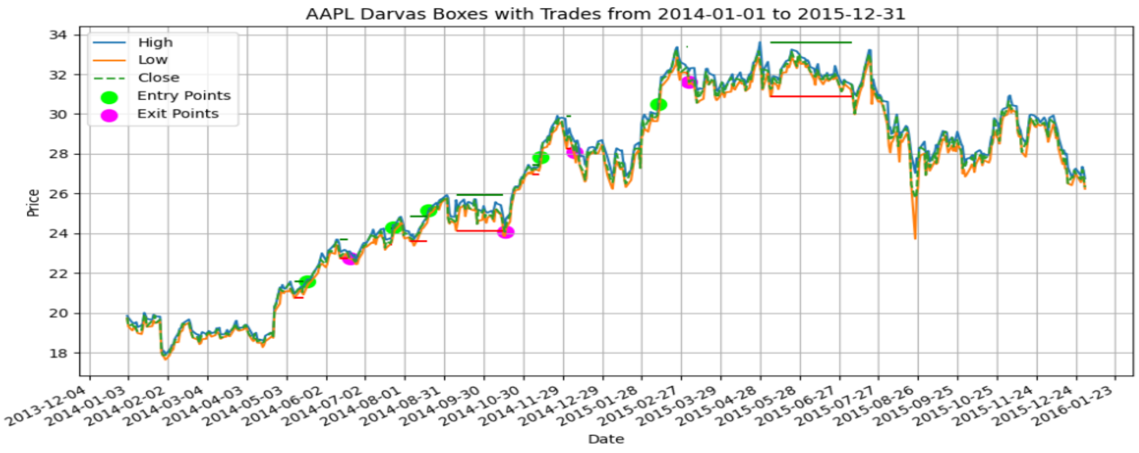
Exit on 10-day highs or lows, opposite to the entry condition.

Trade Type	Factor	Total Units Traded	Different Stocks	Win Rate	Avg. Trade Return	Avg. Win on Trades	Avg. Loss on Trades	Max Trade Duration	Avg. Trade Duration	Total Return
Overall	Without Costs	424,456	1120	34.56%	0.21%	11.7%	-5.87%	362 days	33d 55m	90246.35%
Overall	With Costs	424,456	1120	34.92%	-0.03%	10.79%	-6.24%	362 days	33d 55m	-13,469.67%
LONG	Without Costs	252,693	1120	38.28%	0.45%	9.71%	-5.32%	362 days	35d 13h 55m	113513.61%
LONG	With Costs	252,693	1120	38.28%	0.21%	9.08%	-5.71%	362 days	33d 55m	52,194.07%
SHORT	Without Costs	171,763	1120	29.97%	-0.13%	14.88%	-6.57%	323 days	29d 7h 11m	-21,674.03%
SHORT	With Costs	171,763	1120	29.97%	-0.38%	14.02%	-6.94%	323 days	29d 7h 11m	-65,663.74%

Box Naïve

No Short Trades

Trade Cost = .2% per entry/exit



Boxes are calculated as follows.

- Find a new 12-month high.
- Find the top of the box, which is the highest high for the next three days (4 days total).
- After finding the top, look for the bottom of the box. It's the lowest low for the next three days (4 days total).
- Once the box is complete, a close above the top of the box signals a buy.
- A close below the bottom of the box is the sell signal. Exit and then go back to step 1.

Factor	Total Trades	Different Stocks	Win Rate	Avg. Trade Return	Avg. Win on Trades	Avg. Loss on Trades	Max Trade Duration	Avg. Trade Duration	Total Return
Without Costs	22,129	1120	40.25%	0.25%	8.83%	-5.54%	1113 days	46d 20h 33m	5590.63%
With Costs	22,129	1120	40.25%	-0.04%	8.09%	-6.04%	1113 days	46d 20h 33m	-989.55%

Split/Model Notes

Split dictionary:

- Split 1: (10_10_80) non-trades far distance from trades.
- Split 2: (10_80_10) non-trades medium distance from trades.
- Split 3: (80_10_10) non-trades close distance from trades.
- Split 4: (34_33_33) non-trades uniform distance from trades.

Features:

- The current close price and the previous 20 close prices

Scaling/Trimming:

- A MinMaxScaler is applied to every row of the feature data independently. All data points missing all previous close prices are pruned.

Train/Test Split:

- The dataset is divided using an 80/20 train-test split, where the training set begins January 4, 2010. The testing set start date and value counts varies as follows:
  - Bollinger 1: October 1, 2018. || Train: {0: 100,576, 1: 99,186} Test: {0: 24,960, 1: 24,986}
  - Bollinger 2: October 11, 2018. || Train: {0: 281,545, 1: 276,604} Test: {0: 69,001, 1: 70,537}
  - Turtles: October 5, 2018. || Train: {0: 340,666, 1: 335,421} Test: {0: 83,790, 1: 85,232}
  - Box: December 22, 2017. || Train: {0: 17,707, 1: 17,569} Test: {0: 4,422, 1: 4,397}

Naive:

1. For Bollinger/Box
   - If current scaled price > .95 or < .05 that signals an entry
2. For Turtles
   - If current scaled price == 1 or == 0 that signals an entry

Naïve Bayes:

- Using Gaussian NB from sklearn.naive_bayes package

Polynomial Logistic Regression:

- Using 2nd degree polynomial features on logistic regression from sklearn.linear_model

KNN:

- Using 20 neighbors with KNN Classifier from sklearn.neighbors

RFC:

- Using 100 estimators with RF Classifier from sklearn.ensemble

NN:

- Using 5 hidden layers with 100 neurons for each layer
- Dropout of .1 between each layer

Bollinger Naïve 1

Accuracy

Model	Split 1	Split 2	Split 3	Split 4
Naive	83.13%	81.04%	77.05%	80.51%
Naive Bayes	76.22%	74.94%	72.22%	74.32%
Log Reg	94.38%	91.68%	87.10%	90.94%
KNN	91.86%	89.61%	85.95%	88.77%
RFC	95.81%	94.60%	92.01%	94.08%
NN	**96.35%**	**95.51%**	**94.14%**	**94.49%**

Precision

Model	Split 1	Split 2	Split 3	Split 4
Naive	77.18%	74.67%	70.24%	74.06%
Naive Bayes	74.52%	73.22%	70.57%	72.63%
Log Reg	92.28%	89.17%	83.68%	88.06%
KNN	86.94%	84.13%	79.61%	82.77%
RFC	93.78%	92.59%	89.41%	**91.77%**
NN	**94.04%**	**93.07%**	**91.01%**	91.53%

Specificity

Model	Split 1	Split 2	Split 3	Split 4
Naive	72.42%	68.17%	60.17%	67.11%
Naive Bayes	73.01%	71.29%	68.16%	70.61%
Log Reg	91.95%	88.49%	81.99%	87.16%
KNN	85.29%	81.60%	75.22%	79.63%
RFC	**93.54%**	**92.24%**	**88.70%**	**91.32%**
NN	91.96%	91.13%	88.50%	88.77%

Recall

Model	Split 1	Split 2	Split 3	Split 4
Naive	93.91%	93.92%	93.92%	93.92%
Naive Bayes	79.45%	78.59%	76.28%	78.04%
Log Reg	96.82%	94.87%	92.21%	94.72%
KNN	**98.47%**	97.64%	96.67%	**97.93%**
RFC	98.09%	96.96%	95.32%	96.83%
NN	98.45%	**97.94%**	**97.63%**	97.86%

Bollinger Naïve 2

Accuracy

Model	Split 1	Split 2	Split 3	Split 4
Naive	86.22%	84.75%	82.11%	84.33%
Naive Bayes	76.73%	76.27%	70.02%	73.64%
Log Reg	98.55%	**98.21%**	95.97%	97.40%
KNN	94.83%	93.14%	89.90%	92.14%
RFC	97.73%	97.01%	95.44%	96.64%
NN	**98.60%**	97.84%	**97.12%**	**97.65%**

Precision

Model	Split 1	Split 2	Split 3	Split 4
Naive	81.42%	79.06%	76.08%	78.73%
Naive Bayes	76.67%	75.53%	70.21%	73.24%
Log Reg	97.53%	**96.93%**	94.07%	95.82%
KNN	91.00%	88.26%	84.13%	86.81%
RFC	96.69%	95.66%	93.92%	95.28%
NN	**97.62%**	96.18%	**96.10%**	**97.88%**

Specificity

Model	Split 1	Split 2	Split 3	Split 4
Naive	77.50%	74.88%	69.19%	73.88%
Naive Bayes	75.19%	74.73%	68.63%	71.84%
Log Reg	**97.37%**	**96.84%**	93.58%	95.58%
KNN	89.71%	86.72%	80.70%	84.54%
RFC	96.47%	95.52%	93.47%	95.03%
NN	95.67%	94.64%	**94.28%**	**96.37%**

Recall

Model	Split 1	Split 2	Split 3	Split 4
Naive	94.58%	94.58%	94.57%	94.55%
Naive Bayes	78.20%	77.80%	71.35%	75.39%
Log Reg	99.68%	**99.58%**	98.28%	99.18%
KNN	**99.74%**	99.55%	**98.79%**	**99.57%**
RFC	98.95%	98.50%	97.33%	98.21%
NN	98.95%	98.98%	98.06%	96.97%

Turtle Naïve

Accuracy

Model	**Split 1**	**Split 2**	**Split 3**	**Split 4**
Naive	91.77%	90.07%	87.45%	89.77%
Naive Bayes	65.18%	64.41%	61.02%	62.92%
Log Reg	95.00%	93.39%	91.03%	92.94%
KNN	89.56%	86.80%	83.22%	85.95%
RFC	**96.23%**	**94.82%**	**92.74%**	**94.52%**
NN	95.65%	94.00%	92.29%	93.83%

Precision

Model	**Split 1**	**Split 2**	**Split 3**	**Split 4**
Naive	91.55%	88.49%	84.50%	88.13%
Naive Bayes	67.67%	65.93%	63.06%	64.93%
Log Reg	92.36%	90.08%	87.28%	89.52%
KNN	84.10%	80.76%	77.39%	79.80%
RFC	**94.16%**	**91.85%**	89.40%	**91.66%**
NN	92.93%	90.77%	**90.71%**	90.34%

Specificity

Model	**Split 1**	**Split 2**	**Split 3**	**Split 4**
Naive	91.13%	87.79%	82.26%	87.04%
Naive Bayes	69.81%	68.00%	64.79%	66.66%
Log Reg	91.54%	89.06%	85.27%	88.11%
KNN	80.71%	76.52%	70.98%	74.43%
RFC	**93.63%**	**91.12%**	**87.93%**	**90.70%**
NN	88.91%	86.89%	87.75%	86.15%

Recall

Model	**Split 1**	**Split 2**	**Split 3**	**Split 4**
Naive	92.38%	92.32%	92.40%	92.38%
Naive Bayes	60.73%	60.89%	57.41%	59.32%
Log Reg	98.32%	97.65%	96.53%	97.58%
KNN	98.06%	96.92%	94.91%	97.02%
RFC	**98.72%**	**98.46%**	**97.33%**	**98.19%**
NN	98.11%	96.83%	93.52%	97.81%

Box Naïve

Accuracy

Model	**Split 1**	**Split 2**	**Split 3**	**Split 4**
Naive	81.59%	79.82%	76.41%	80.10%
Naive Bayes	86.37%	84.75%	82.42%	84.31%
Log Reg	90.42%	88.89%	86.39%	88.43%
KNN	82.53%	80.84%	76.82%	79.80%
RFC	**91.85%**	**91.00%**	**88.68%**	**90.29%**
NN	86.46%	88.25%	85.54%	87.42%

Precision

Model	**Split 1**	**Split 2**	**Split 3**	**Split 4**
Naive	73.34%	71.26%	68.14%	71.51%
Naive Bayes	79.20%	77.06%	74.52%	76.52%
Log Reg	86.62%	84.64%	81.97%	84.01%
KNN	78.89%	76.66%	72.71%	75.06%
RFC	88.55%	**87.47%**	**84.77%**	**86.50%**
NN	**90.82%**	86.82%	84.46%	84.57%

Specificity

Model	**Split 1**	**Split 2**	**Split 3**	**Split 4**
Naive	62.90%	59.81%	52.63%	60.45%
Naive Bayes	73.40%	70.62%	65.69%	69.83%
Log Reg	84.87%	82.80%	79.12%	82.04%
KNN	75.56%	73.09%	67.11%	70.56%
RFC	87.27%	**86.32%**	**82.78%**	**85.19%**
NN	**90.64%**	84.97%	82.66%	81.99%

Recall

Model	**Split 1**	**Split 2**	**Split 3**	**Split 4**
Naive	**99.87%**	**99.86%**	**99.86%**	**99.86%**
Naive Bayes	99.06%	98.91%	98.92%	98.86%
Log Reg	95.85%	94.98%	93.56%	94.86%
KNN	89.34%	88.60%	86.40%	89.11%
RFC	96.32%	95.69%	94.51%	95.43%
NN	81.11%	89.56%	86.66%	91.11%