Capstone Project

Martin Piják

Bratislava, Slovakia

Abstract

This is project for using Machine Learning (ML) methods in futures trading.

Keywords: Machine Learning, Trading, Udacity, Nanodegree

1. Domain

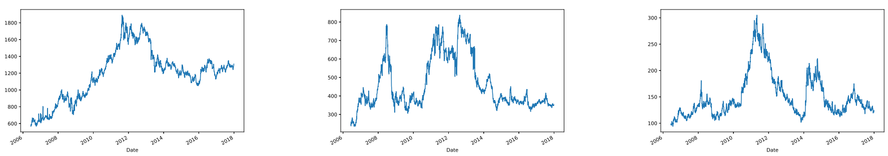
This project investigates possible futures trading strategies on Chicago Mercantile Exchange (CME) and Intercontinental Exchange (ICE) markets with Machine Learning methods. The goal is to find the trading strategy mostly based on the price, Commitment of Traders report (COT) and seasonality pattern. We will compare this trading strategy to commonly used investing approaches as returns of Nasdaq and fixed returns of interest rate 4%.

Following commodities were investigated:

- Gold
- Corn
- Coffee

2. Datasets and Inputs

I used data from Quandl. Data contains Open, High, Low, Close and volume (OHLCV) and commitment of traders (COT). Continuous data was generated by taking contract with the highest volume for the trading day. For details please see **data-preparation.ipynb**.



(a) Gold (b) Corn (c) Coffee

Figure 1: Close price graphs

3. ML

3.1. Tools

Following tools were used:

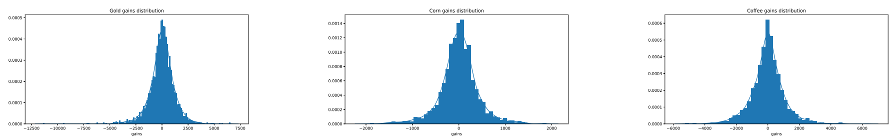
- General ML tools
  - Scikit
- Visualization
  - matplotlib
  - seaborn
- Model building tools
  - Keras
  - TensorFlow (keras backend)
  - H2O
  - LightGBM

3.2. Where is the code?

All the data and code that was used to generate this report can be found in the **trading_project.ipynb** notebook. When training neural network with keras you can arrive at slightly different results than those in this report because of different seed.

4. Data analysis

In the proposal I wanted to investigate classification of volatility. Below is 95% of daily volatility. 95% was selected as a simulation of slippage. General trading idea is to keep trade for a day.



(a) Gold (b) Corn (c) Coffee

Figure 2: Gains distribution

4.1. Label Generation for classifier

Labels were generated based on the trading target. Label -1 is for short trade 0 for no trade and 1 for long trade.

Threshold was chosen for different commodities differently based on the overall gains of the next day considering whole contract.

Commodity	Gains target (USD)
Gold	500
Corn	150
Coffee	375

$$fee = 1.5 \tag{1}$$

$$v_{volatility} = (\text{close} - \text{open}) * \delta \text{ where } \delta = 0.95 \tag{2}$$

$$labels \begin{cases} |v_{volatility}| > t_{threshold} + fee \implies \begin{cases} v_{volatility} > 0 \implies 1 \text{ (long)} \\ v_{volatility} < 0 \implies -1 \text{ (short)} \end{cases} \\ |v_{volatility}| < t_{threshold} + fee \implies 0 \text{ (no trade)} \end{cases} \tag{3}$$

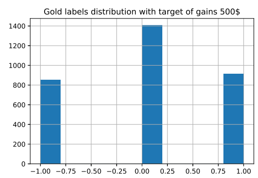
δ constant is used for simulating slippage.

I decided to have a look at regressor as well.

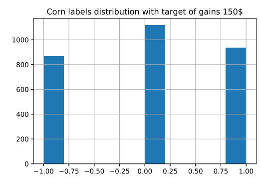
Based on the past 2 years of trading data (OHLCV and COT) classifier is deciding whether to trade or not.

Approach to training classifier/regressor and evaluation does not take into account that stop-loss can still make trade unsuccessful. From Machine Learning perspective the best approach is to have result of classification regression as close to the desired outcome as possible.

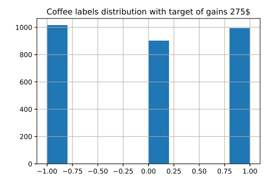
Stop-loss was selected so that 90% of trades will be successfully executed (exited on close).



(a) Gold

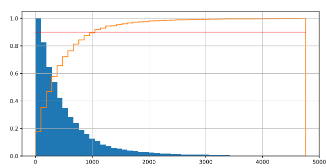


(b) Corn

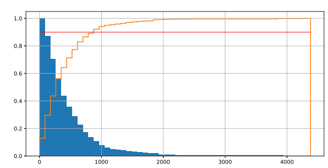


(c) Coffee

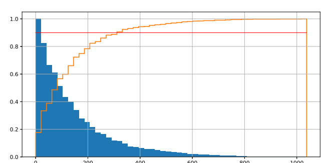
Figure 3: Labels distribution



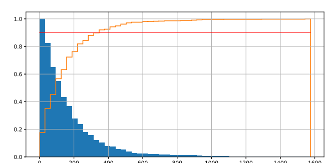
(a) Gold long



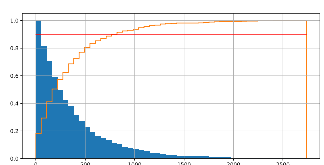
(b) Gold short



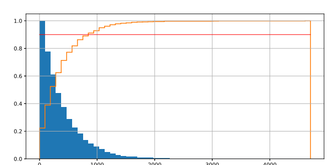
(c) Corn long



(d) Corn short



(e) Coffee long



(f) Coffee short

Figure 4: Stop Loss analysis counter position move

Commodity	Long Stop Loss	Short Stop Loss
Gold	1000	800
Corn	300	300
Coffee	800	800

5. Feature Engineering

We transformed last $2 * 252$ trading days into vector with 2520 scalars.

In order to capture cyclicality, we have transformed features as trading day of month, day of week and quarter of year into *sin* and *cos* values. I am not sure whether this transformation grants subsequent PCA usage. After adding data features vector now contains 2528 scalars.

We transformed COT to reflect extreme in the index. COT 1 of industrials corresponds to the maximum of traders' positions throughout the last 2 years. Respectively 0 to minimum. We have included COT for commercial and industrial users from the last 8 weeks which is 16 values. Training vector now contains $2528 + 16 = 2544$ values.

5.1. PCA

We try to capture about 80% of variance of the data. For corn this corresponds to about 160 components. For gold it is about 180 components. Case of coffee is very strange.

It is interesting Corn is best explained by PCA transformation. It is probably due to clear seasonal patterns in trading.

I am surprised that gold is better explained by PCA transformation than Coffee. I would expect that coffee has stronger seasonal trading patterns than gold because of the growth cycle. Maybe gold mining is subject to the weather in similar way as, agricultural commodities. Gold is still mostly recycled and new production has limited impact on total amount of traded gold.

Possible explanations for lack of variance explanation by PCA for coffee:

- corn traded on CME is mostly US produced with stable harvest season
- production of coffee is very unpredictable depending on the conditions of a given year
- there are multiple producers around the world (coffee is more of a global market with limited US production) with different harvest periods Coffee Harvest
- important difference between coffee and corn is price per unit corn is much less efficient to transport
  - 1 kg of corn is worth about 15¢
  - 1 kg of coffee is worth about 230¢

Based on the PCA variance graph I think that PCA transformation is not suitable for coffee. Information in components is growing linearly. If we don't see sharp increase of cumulative explained variance with few first components, then PCA transformation is not suitable. Therefore, coffee should not be considered for trading with PCA transformation. I will continue with coffee as well but based on this transformation I would not go ahead with trading unless I would find different transformation.

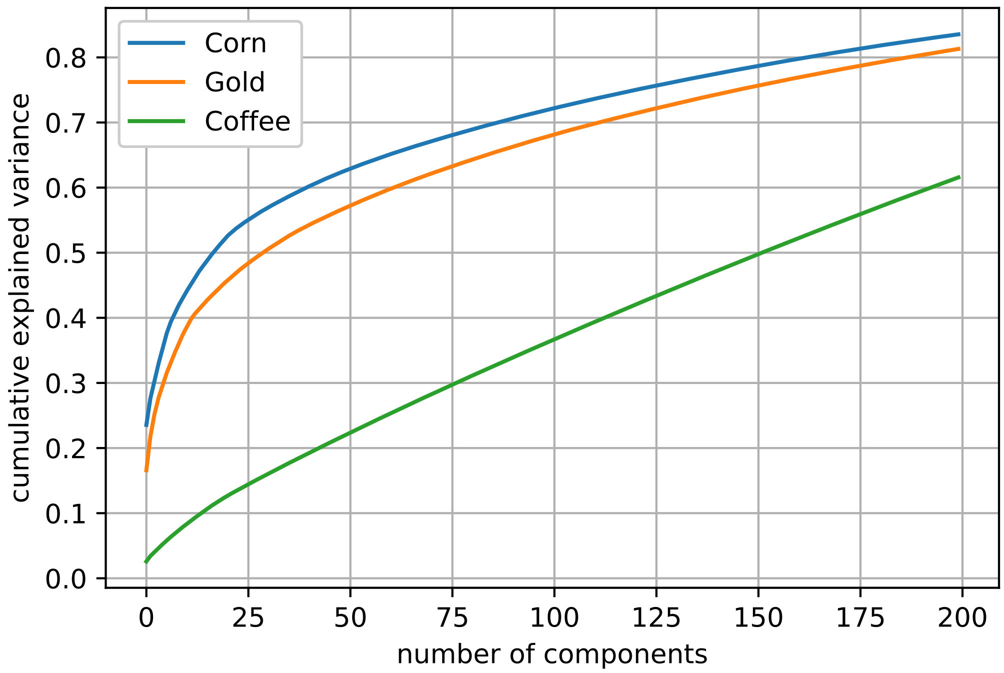
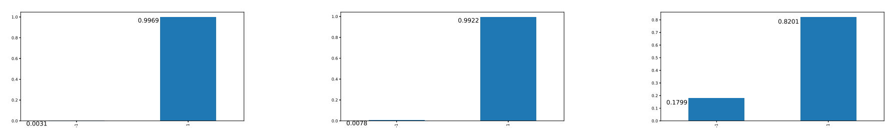


Figure 5: PCA explained variance

5.2. Outliers

We checked dataset for outliers. I used IsolationForest from scikit with automatic outlier detection after PCA transformation. Data was split into 3 sets training, validation and performance check. Training set contains more that 1930 samples. On this training set outlier detection was trained. Following number of outliers were detected.

Commodity	Number of Inliers	Number of Outliers
Gold	1927	6
Corn	1915	15
Coffee	1577	346



(a) Gold (b) Corn (c) Coffee

Figure 6: Outliers histogram

6. Model Training

6.1. General strategy for training

We tried the same model for all three commodities (gold, corn and coffee). We are looking for classifier with the same parameters for any commodity. This way we can be sure that we have found some general classification and results are not just a coincidence for a given model. Generally, we try to find one model that works for most markets.

Following exploration can be split into:

- classificator
- regressor

After model training all the models will be validated on the year 2018 trading data. Trading signals from model will be used for simulated trading. Open question is in case of regressor where the threshold for trading should lie? Empirically I chosen to trigger at most around 80 trades.

6.2. LightGBM regressor

6.2.1. Parameters

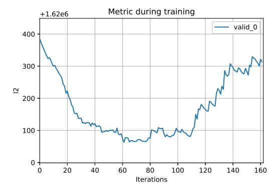
We used following parameters to train LightGBM regressor.

Parameter	Value
num_leaves	40
objective	regression
boosting	dart
metric	l2 (mean squared error)
estimators	1000
learning_rate	0.001
num_class	1
max_bin	30
reg_alpha	5
reg_lambda	10
num_round	1000
early_stopping_rounds	100

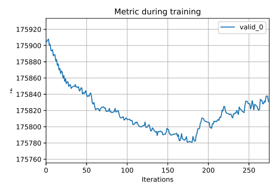
6.2.2. Training

When determining trading performance of the LightGBM regressor we have chosen following trading thresholds.

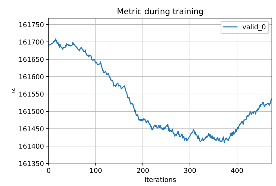
Commodity	Trading threshold	Classifier threshold	Ratio
Gold	10	500	$\frac{1}{50}$
Corn	10	150	$\frac{1}{15}$
Coffee	10	375	$\frac{1}{38}$



(a) Gold

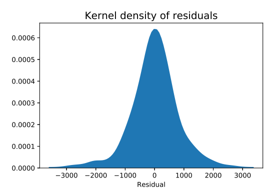


(b) Corn

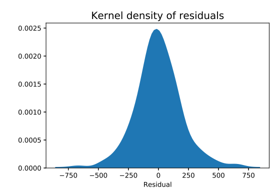


(c) Coffee

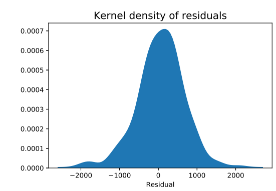
Figure 7: Training Error



(a) Gold

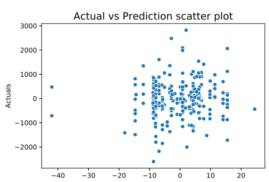


(b) Corn

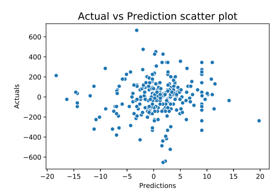


(c) Coffee

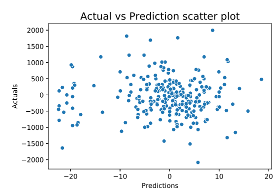
Figure 8: Residual Errors



(a) Gold



(b) Corn



(c) Coffee

Figure 9: Actual vs Prediction Scatter Plot

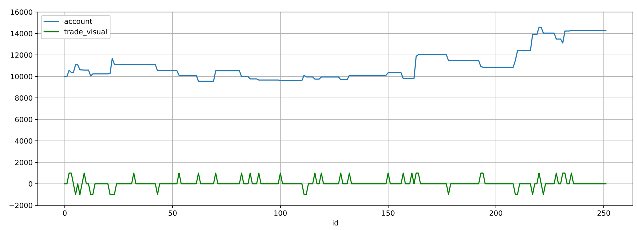
This part of trading process is somewhat unknown to me. Regressor is not predicting many days with big volatility. Predictions are much flatter than the actual volatility.

Based on the chosen threshold you can get some idea of how well is the regressor approximating actual volatility.

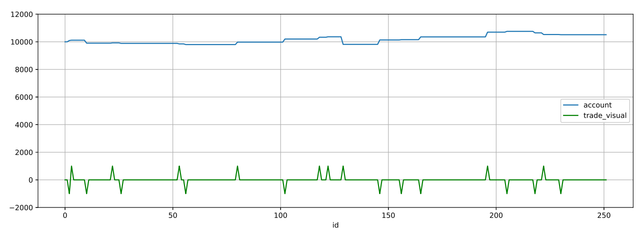
6.2.3. *Trading performance*

Commodity	Prediction/actual result correlation	Gains
Gold	0.079	42%
Corn	0.078	5%
Coffee	-0.001	62%

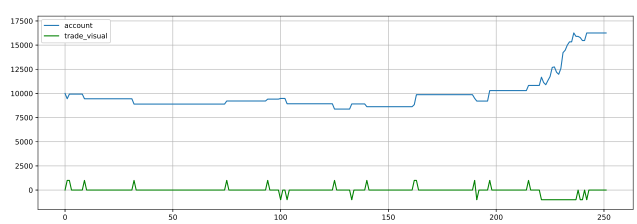
We evaluated trading performance on the year 2018. Below is account performance and histogram of trades.



(a) Gold

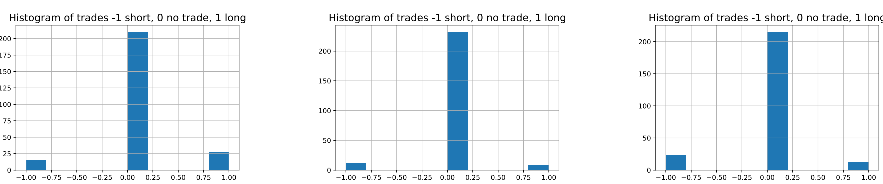


(b) Corn



(c) Coffee

Figure 10: Account performance



(a) Gold (b) Corn (c) Coffee

Figure 11: Trade histogram

6.3. *Neural Network Classication*

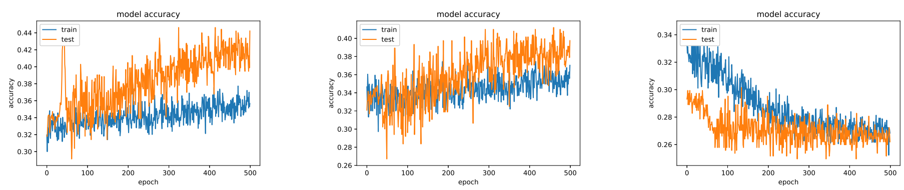
6.3.1. *Neural network design*

Neural network is fully connected with input layer, 2 hidden layers and output layer consisting of 3 neurons. Output layer of 3 layers represent 3 categories $\{-1, 0, 1\}$.

Input	input layer	hidden layers		output layer	output
200	32	16	8	3	3

6.3.2. *Classifier Training*

In the training setup we used training data with labels described above. If validation loss improved then new weights were saved. Total number of epochs was 500 and batch size 20.



(a) Gold (b) Corn (c) Coffee

Figure 12: Accuracy during training

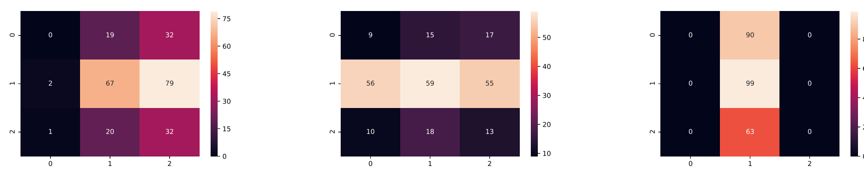
6.3.3. *Performance Evaluation*

Results below were obtained by running classifier on trading data from year 2018.

Commodity	Gains
Gold	68 %
Corn	-40 %
Coffee	0 % (no trade)

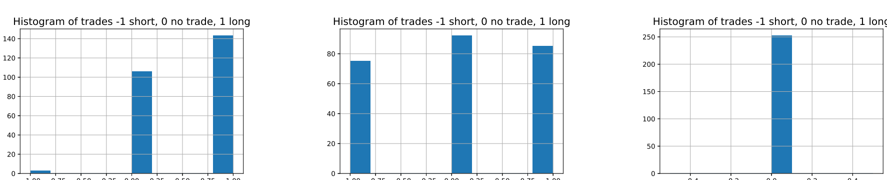
Precision Score:

Commodity	Short	No trade	Long
Gold	0.	0.63	0.22
Corn	0.12	0.64	0.15
Coffee	0.	0.39	0.



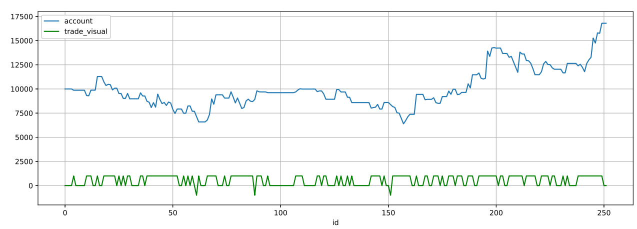
(a) Gold (b) Corn (c) Coffee

Figure 13: Confusion Matrix

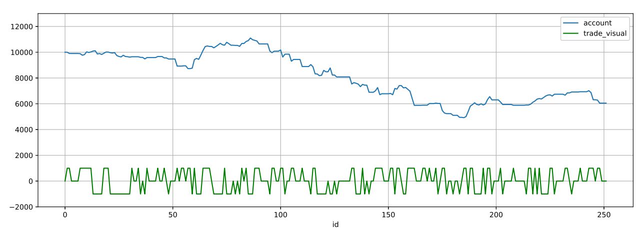


(a) Gold (b) Corn (c) Coffee

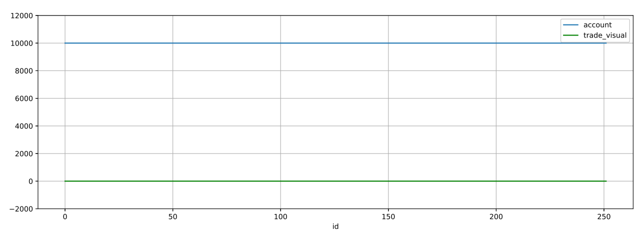
Figure 14: Trade histogram



(a) Gold



(b) Corn



(c) Coffee

Figure 15: Trade histogram

6.4. Neural Network Regressor

6.4.1. Neural Network Design

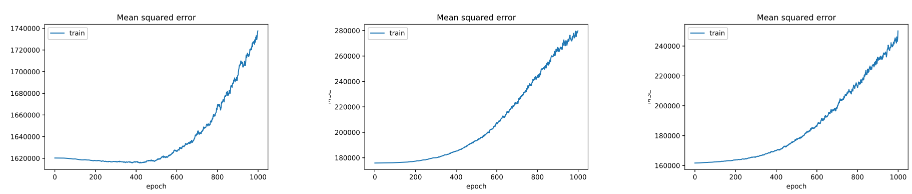
Neural network is fully connected with input layer, 3 hidden layers and output layer consisting of 1 neuron.

Input	input layer	hidden layers			output layer	output
200	500	128	64	32	1	1

6.4.2. Regressor Training

In the training setup we used training data with labels described above. If validation loss improved then new weights were saved. Total number of epochs was 1000 and batch size 20.

During training only in case of gold we can see improvement of loss on the validation set. For details see validation loss (MSE) below.



(a) Gold (b) Corn (c) Coffee

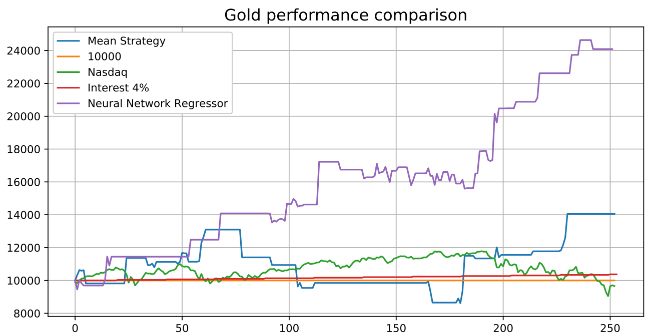
Figure 16: Mean Squared Error

7. Benchmarking Model

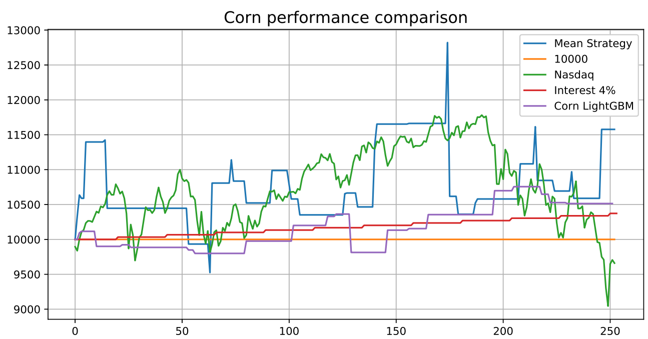
We used three comparisons to model performance.

- fixed percentage of 4% with monthly interest calculation (bank/bond deposit comparison)
- Nasdaq performance
- mean reversal trading on given commodity markets with mean of 20 days and mean of 10 days

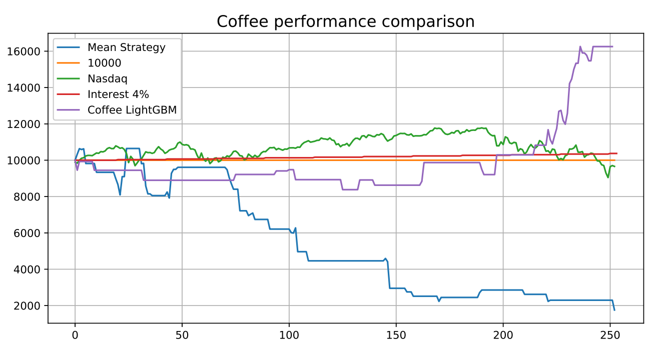
It is interesting to notice that mean reversal strategy works for gold and corn but does not produce gains for coffee market.



(a) Gold



(b) Corn



(c) Coffee

Figure 17: Benchmark Performance

8. Conclusion

Trading is a difficult ML problem. Out of three compared commodities gold, corn and coffee we were able to predict volatility/trade class with gold and corn. Coffee behaved randomly with approximately 0 correlation to the actual volatility.

In the beginning of project, I was thinking of a classifier (short, no trade, long) because it is closer to the usage of model. I tested regressor as well. I think that regressor works better because there is more information preserved for training. I tried different loss functions when training regressor. I decided to use weighted MSE. This could be further modified for better function omitting errors below threshold.

Gold trading is the most capital intensive with very big stop losses (1000 long, 800 short). This can be problem for trading with 10000$ account.

Corn moves are much smaller and seem more suitable for the start of trading. General risk is more acceptable (stop-loss 300$). In case of corn mean reversal strategy is defeating our trading strategy.

In case of coffee the data was almost impossible to classify. I suspect more data transformation is needed to get better results.