Анализ и применение алгоритмов и методов машинного обучения в трейдинге

Московский физико-технический институт

Студент – Емцев Илья

Научный руководитель – Орлова Е. Р.

Москва 2023

Предпосылки

• Финансовые рынки становятся все более сложными и эффективными (уменьшение транзакционных издержек)

• Традиционные методы работают уже не так хорошо

• Большое количество доступных данных для анализа

• Новые технологии (ML, DL), которые могут улучшить предсказание и выявить какие-то закономерности

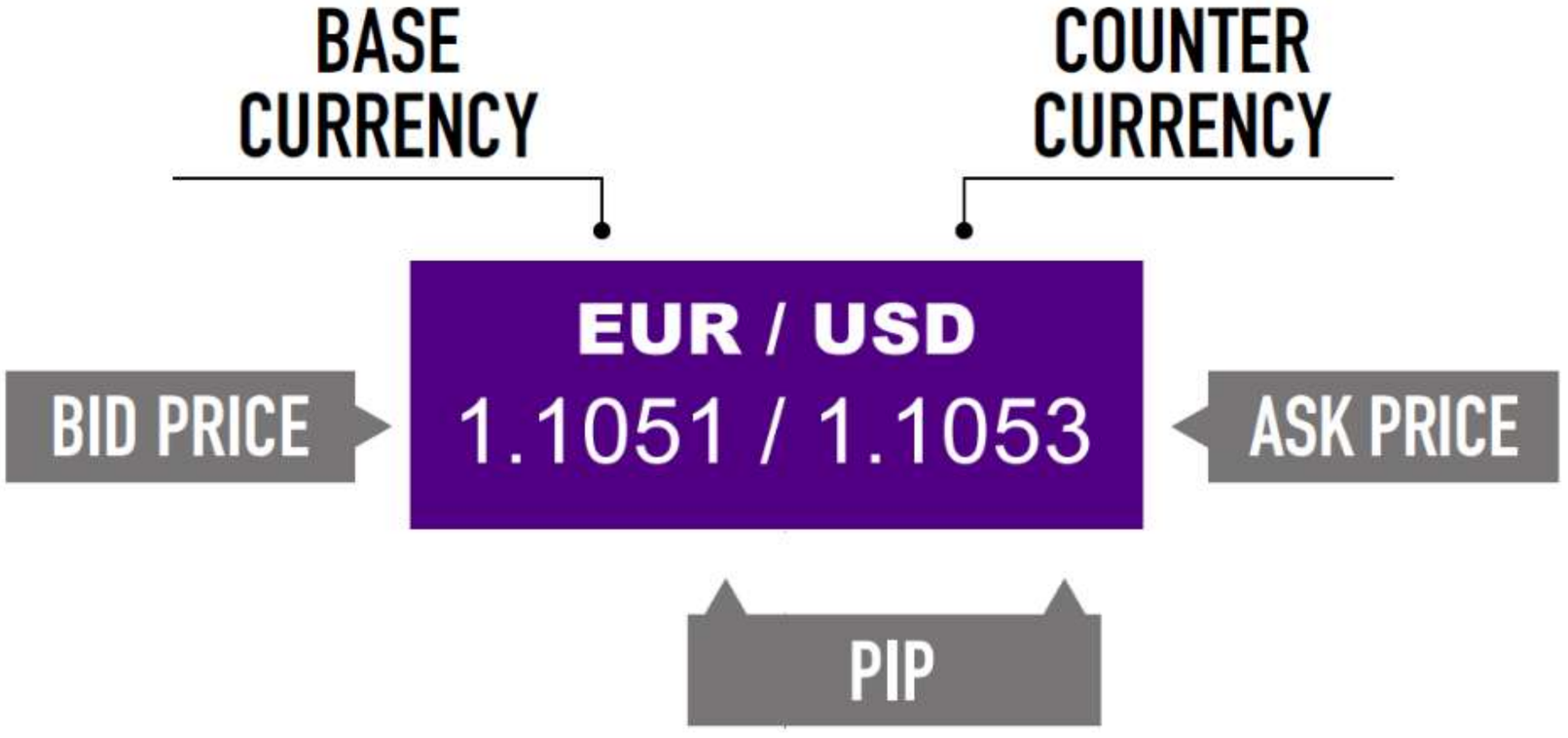
P.S. Новые технологии - не являются магическим решением и требуют осторожного и осознанного подхода при принятии инвестиционных решений!

Из чего состоит презентация:

- Краткое введение в трейдинг
- Доходность (определение, виды)
- Краткое рассмотрение данных
- Распределение доходности и GMM (Gaussian Mixture Models)
- Стратегии в трейдинге
- Линейная / Логистическая регрессия
- Нейронная сеть (DNN)

Основные понятия в трейдинге

Основные показатели



SPREAD = 1.1053 - 1.1051 = 2 pips

Доходность

Логарифмическая доходность

- **Аддитивна**
- **Более устойчива к выбросам**
- **Более склонна к нормальному распределению**
- **Среднее арифметическое информативно (работает как среднее геометрическое для обычной доходности)**

$$\text{Log } \mathrm{R}_t = \log\left(\frac{\mathrm{P}_t}{\mathrm{P}_{t-1}}\right) = \log(1 + R_t)$$

$$\begin{cases} \mathrm{FV} = \mathrm{PV} \cdot (1 + \frac{r}{m})^{n \cdot m}, \text{ дискретная капитализация} \\ \mathrm{FV} = \mathrm{PV} \cdot e^{r \cdot n}, \text{непрерывная капитализация} \end{cases}$$

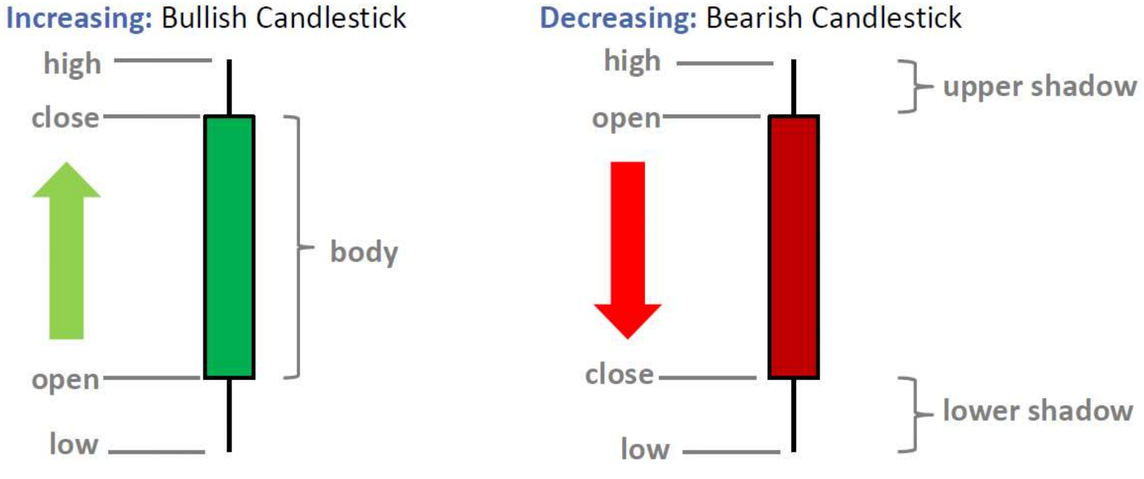
$$\log(1 + R_t) \approx R_t.$$

Рассмотрение данных

Вид данных

- Библиотека yfinance, выгрузка данных google за три года:

Date	Open	High	Low	Close	Adj Close	Volume	Name
2020-05-01	66.425003	67.603500	65.550003	66.030502	66.030502	41450000	GOOG
2020-05-04	65.411499	66.383003	64.949997	66.339996	66.339996	30080000	GOOG
2020-05-05	66.896004	68.696999	66.873001	67.555496	67.555496	33030000	GOOG
2020-05-06	68.084503	68.556000	67.364502	67.364998	67.364998	24308000	GOOG
2020-05-07	68.296997	68.879997	67.763496	68.627998	68.627998	27952000	GOOG



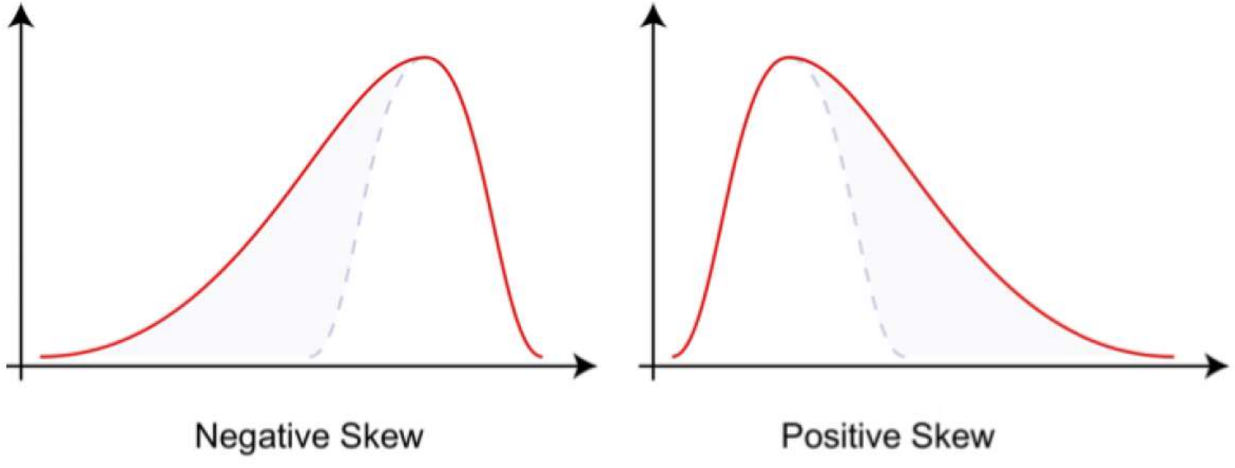
Распределение доходности и GMM

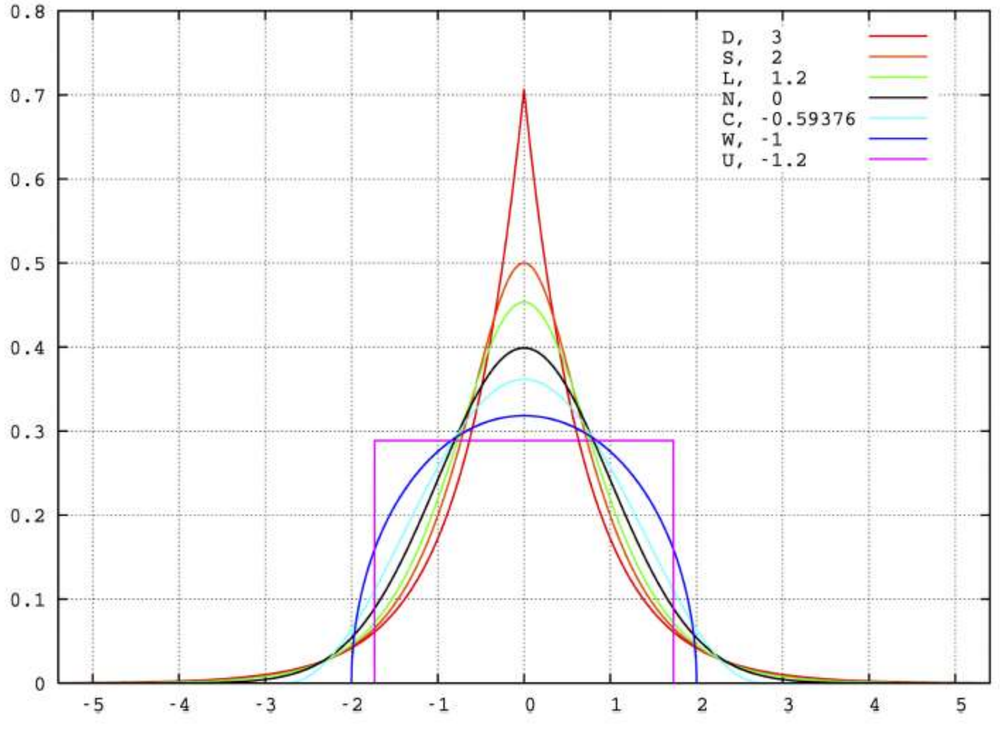
Коэффициенты асимметрии/эксцесса

- Пытаемся избегать смещения влево
- Пытаемся избегать тяжёлых хвостов

$$Skewness = \frac{E[(X-EX)^3]}{DX^3}$$

$$Kurtosis = \frac{E[(X-EX)^4]}{DX^4} - 3$$



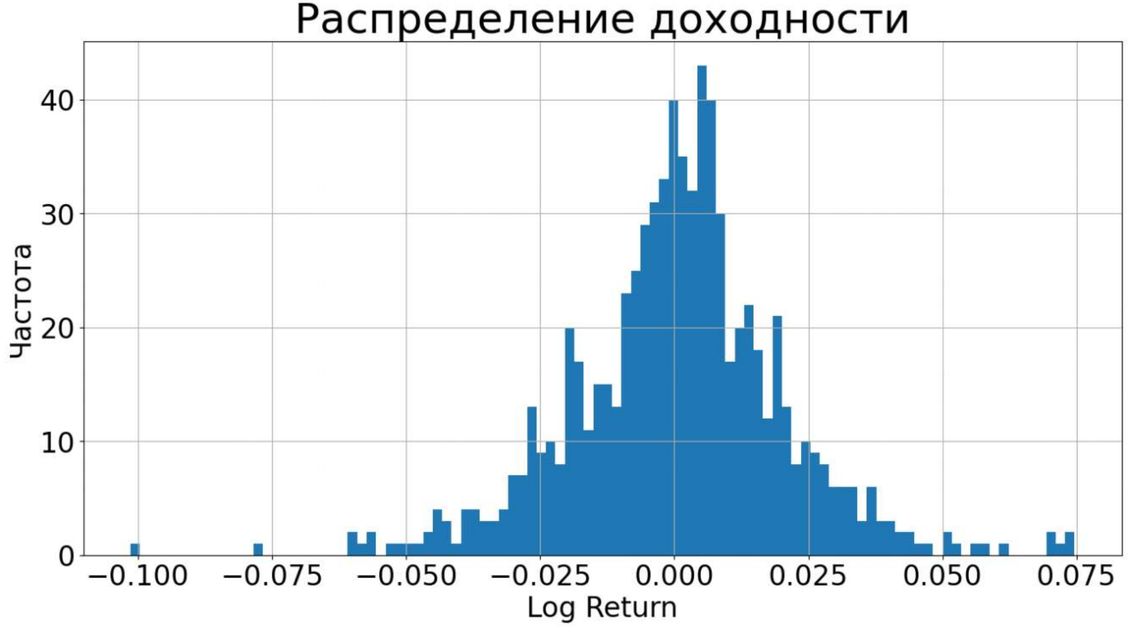


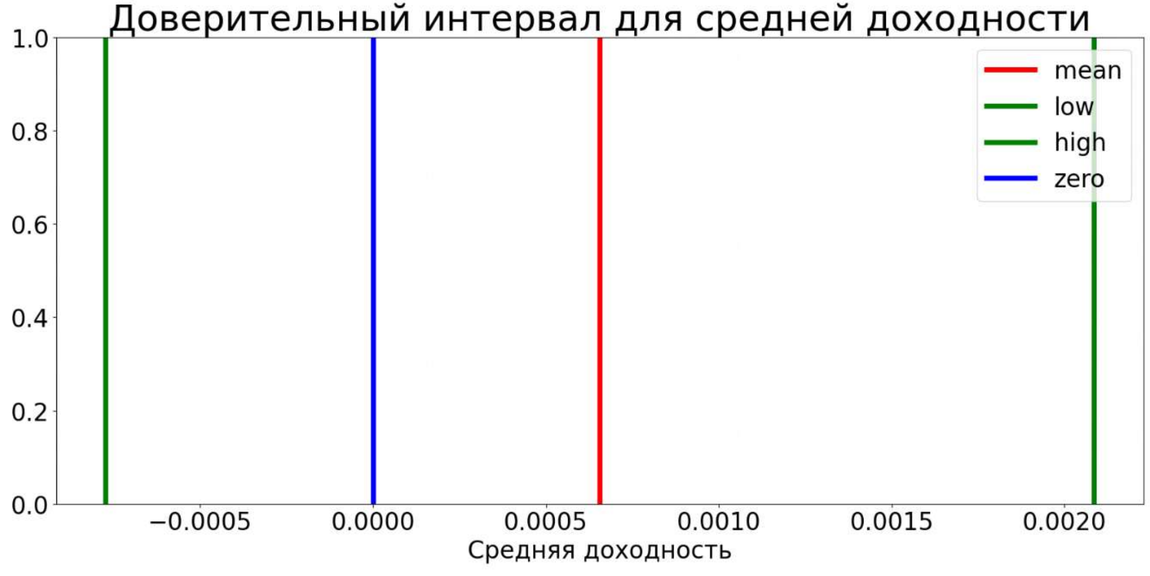
Доверительный интервал для среднего

- Предположим, что нормально распределено с неизвестной дисперсией, тогда средняя доходность лежит в следующем 95% доверительном интервале

$$R = \bar{x} \pm 1.96 \cdot \frac{s}{\sqrt{n}}$$

где $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ - выборочная несмещённая дисперсия, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$



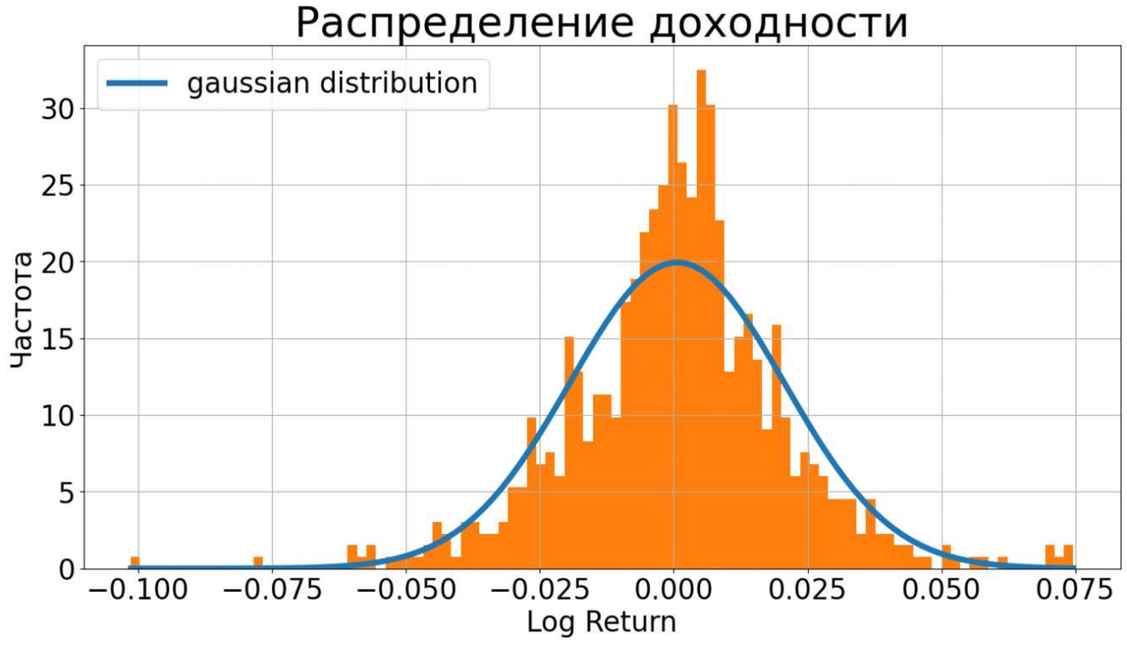


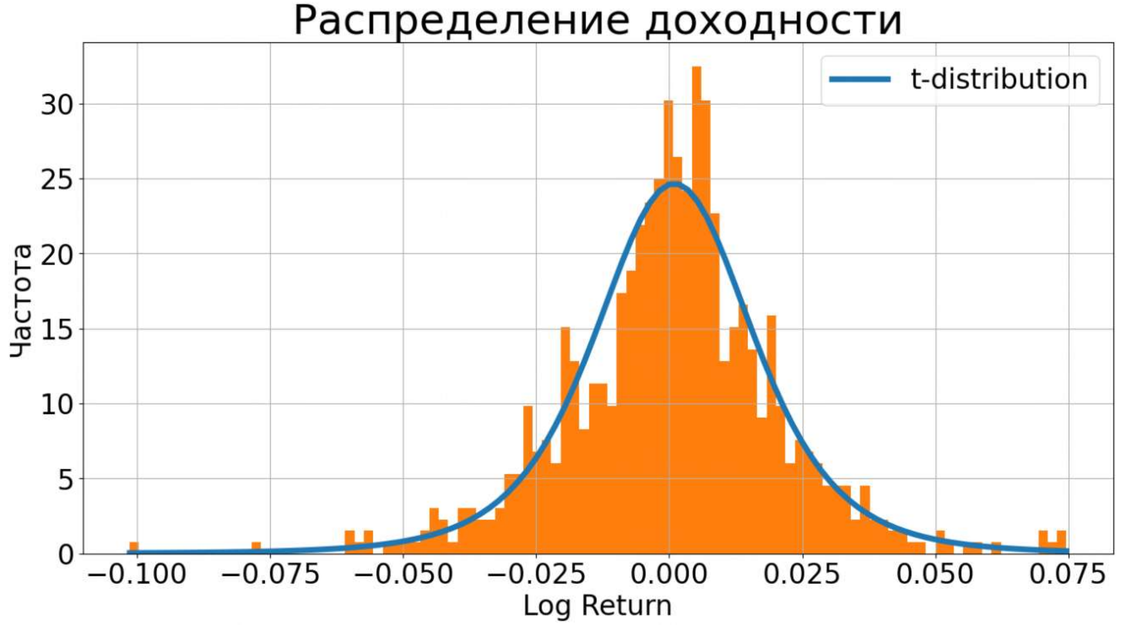
Распределение Стьюдента

- Значения коэффициента асимметрии и эксцесса наших данных:

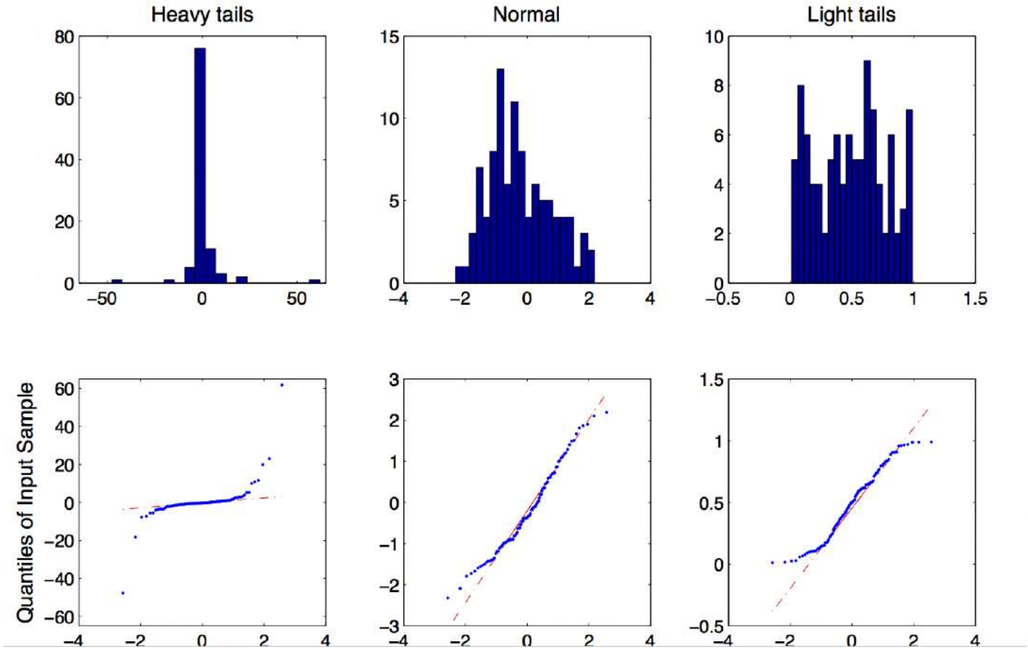
$$Skewness \quad -0.115$$

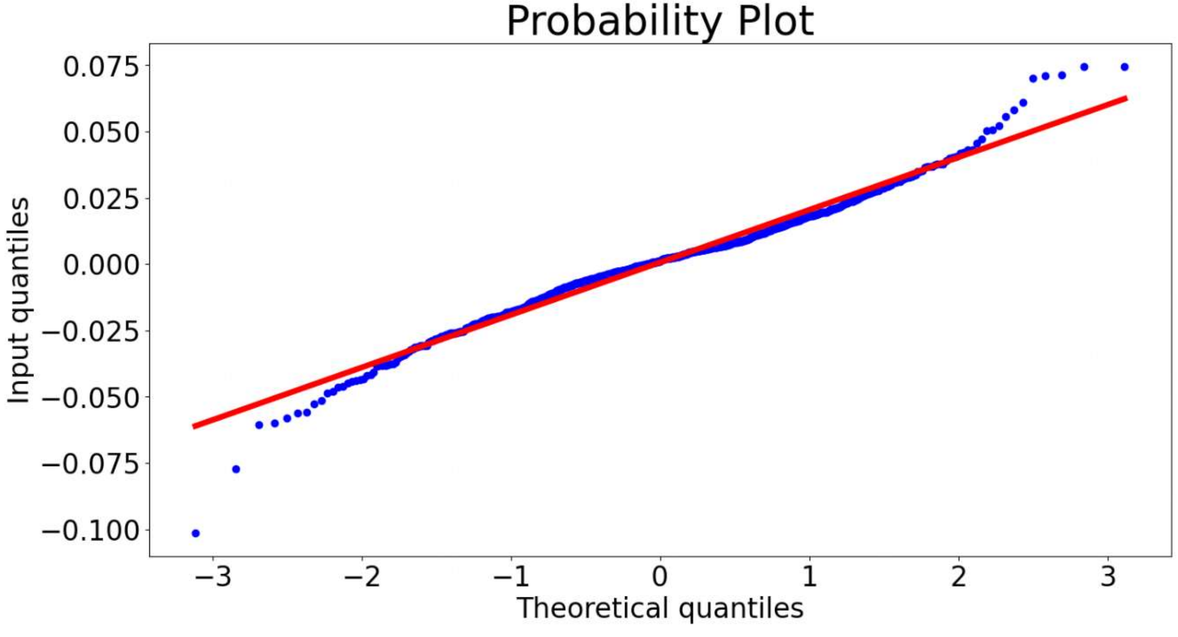
$$Kurtosis \quad 2.083$$



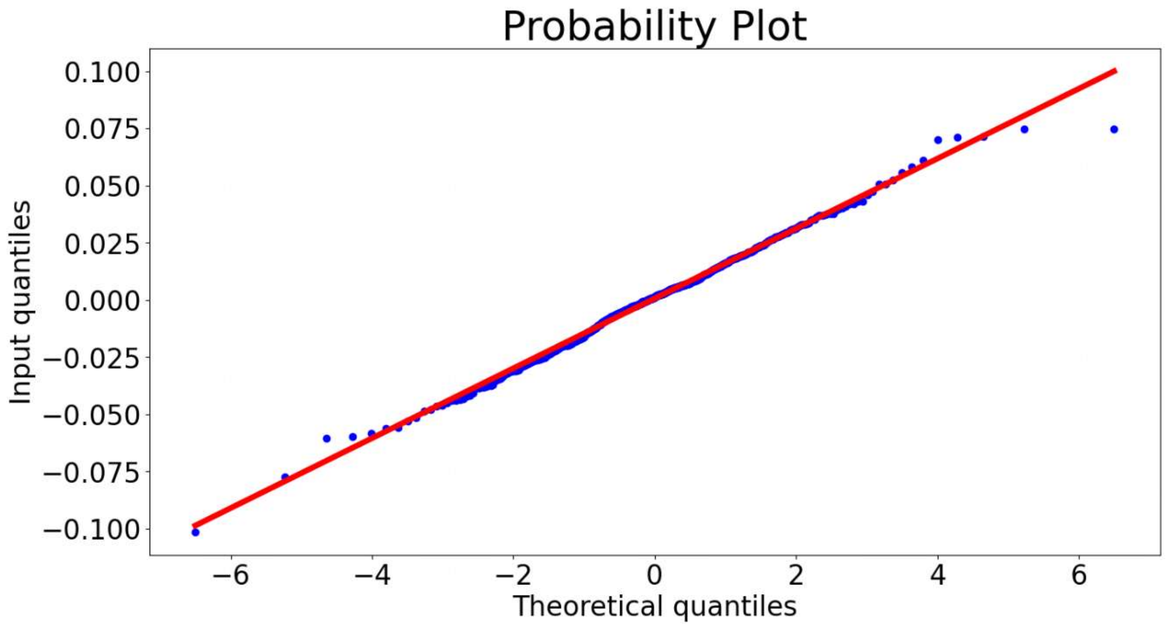


QQ-plot





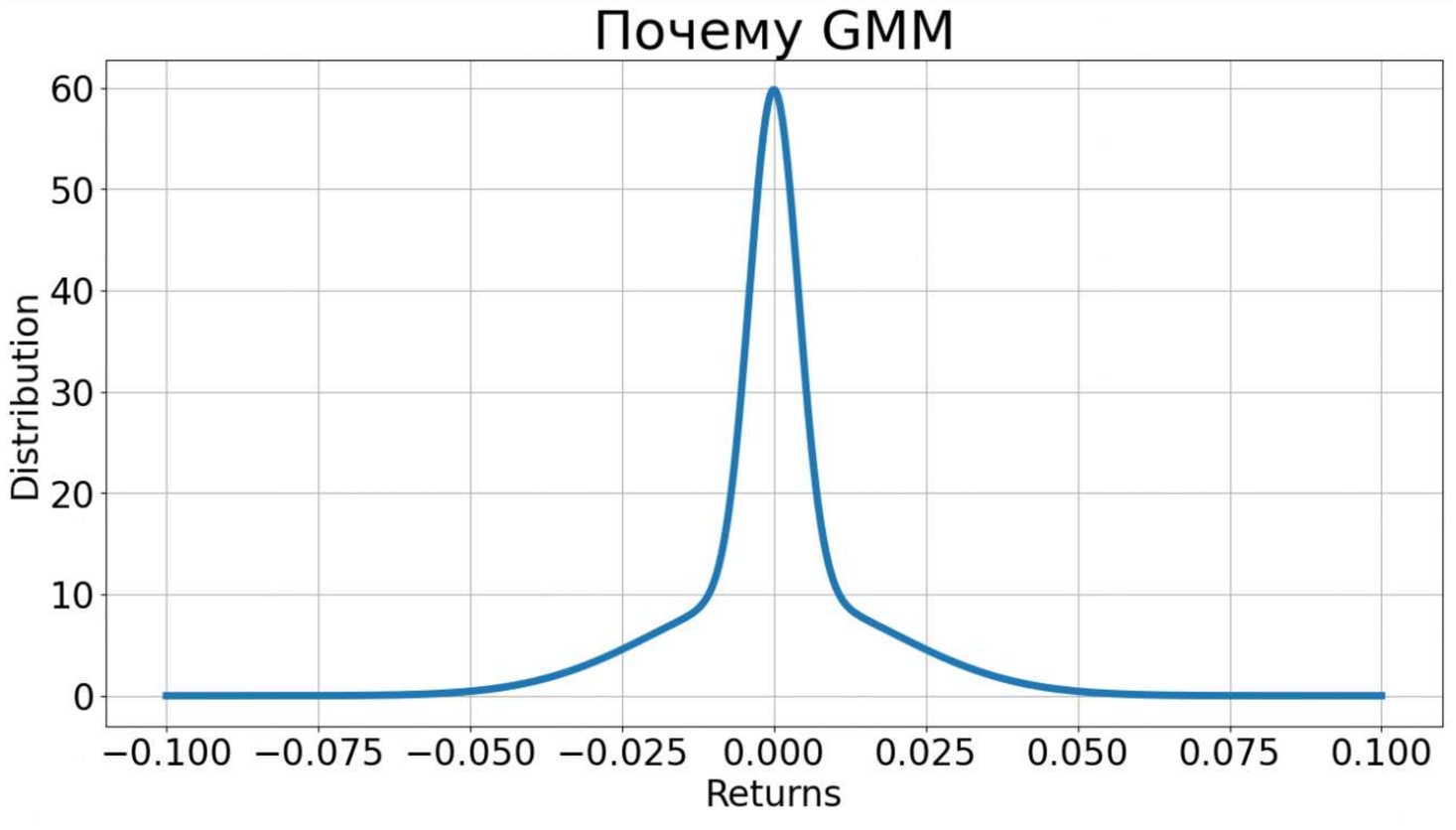
Нормальное



Стьюдент

GMM (Gaussian Mixture Models)

$$p(\mathbf{x}) = \pi \cdot N(x \mid \mu, \sigma_1^2) + (1 - \pi) \cdot N(x \mid \mu, \sigma_2^2)$$

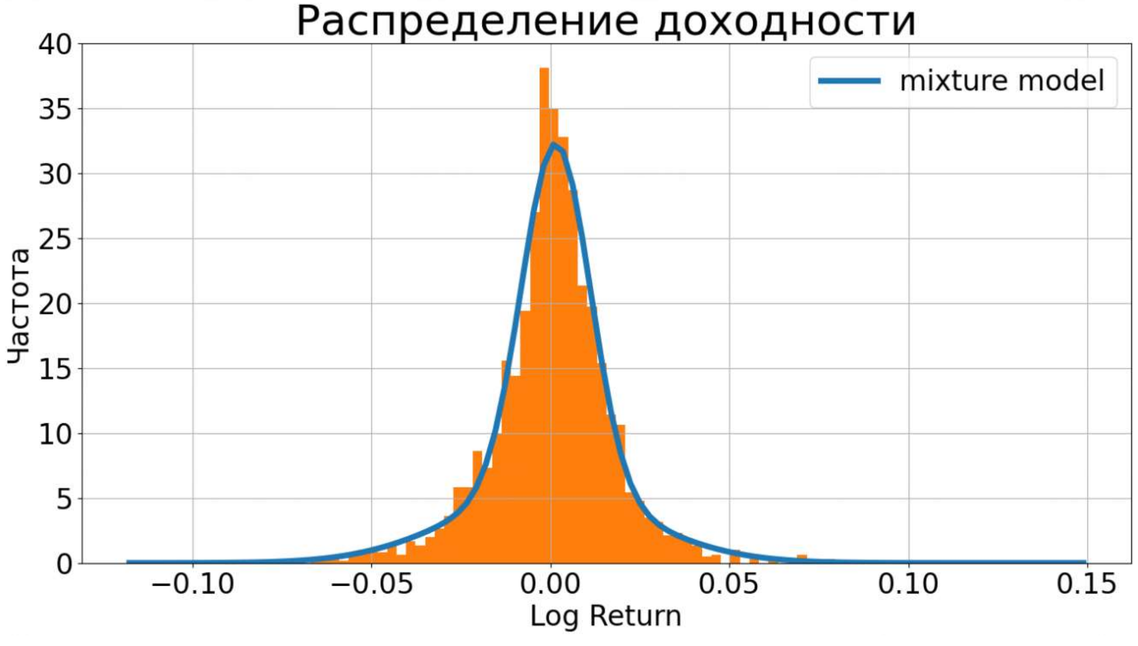


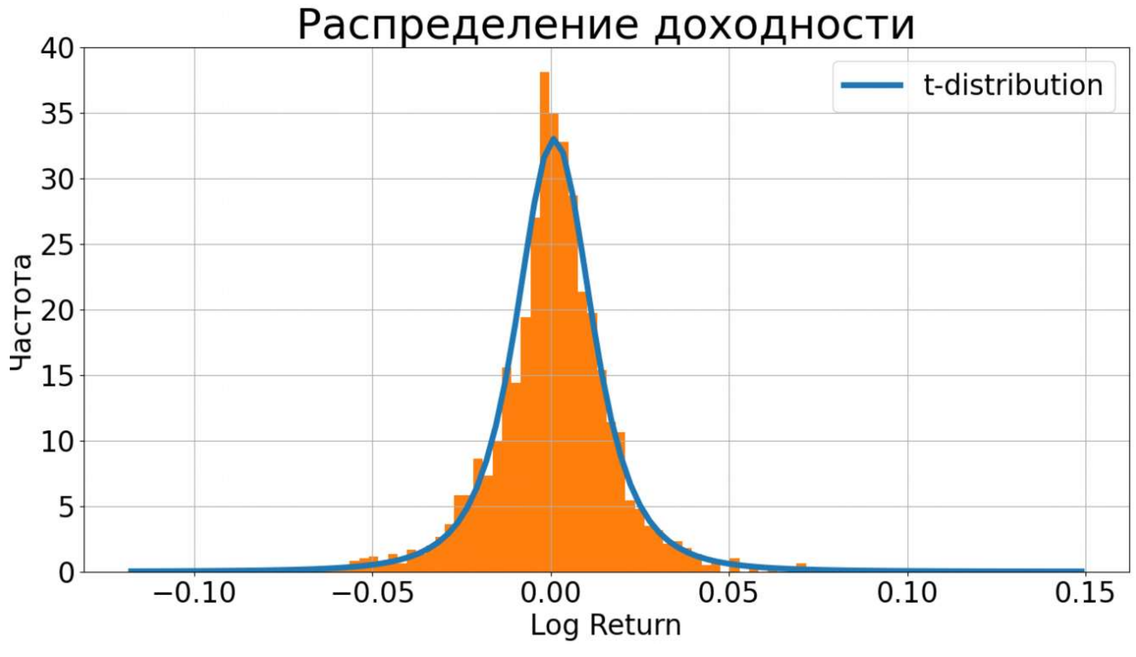
GMM (Gaussian Mixture Models)

$$t\text{-}distribution\ kurtosis = 5.481$$

$$GMM\ kurtosis = 5.500$$

$$Real\ kurtosis = 6.422$$

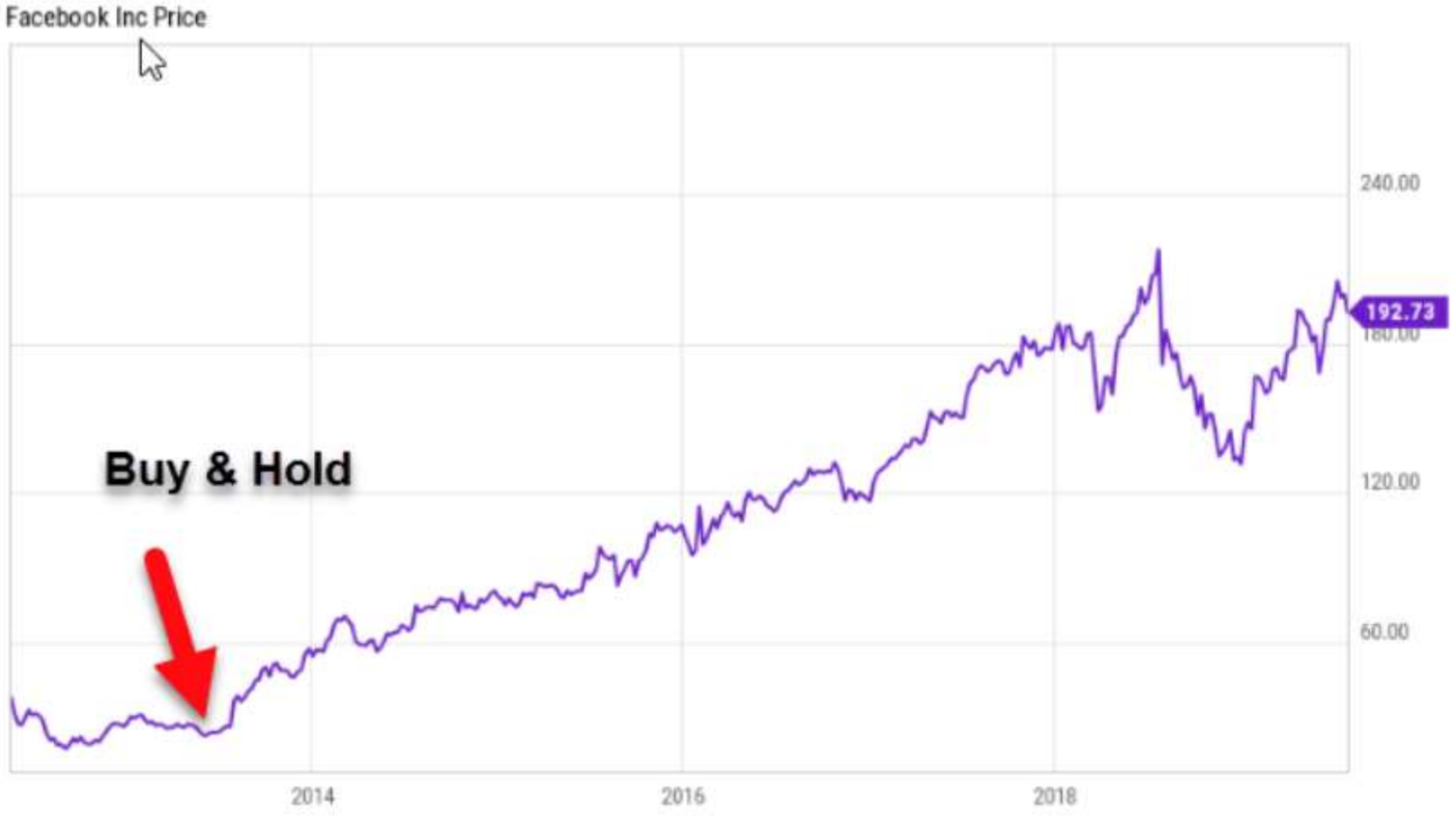




Стратегии и методы оценки их работы

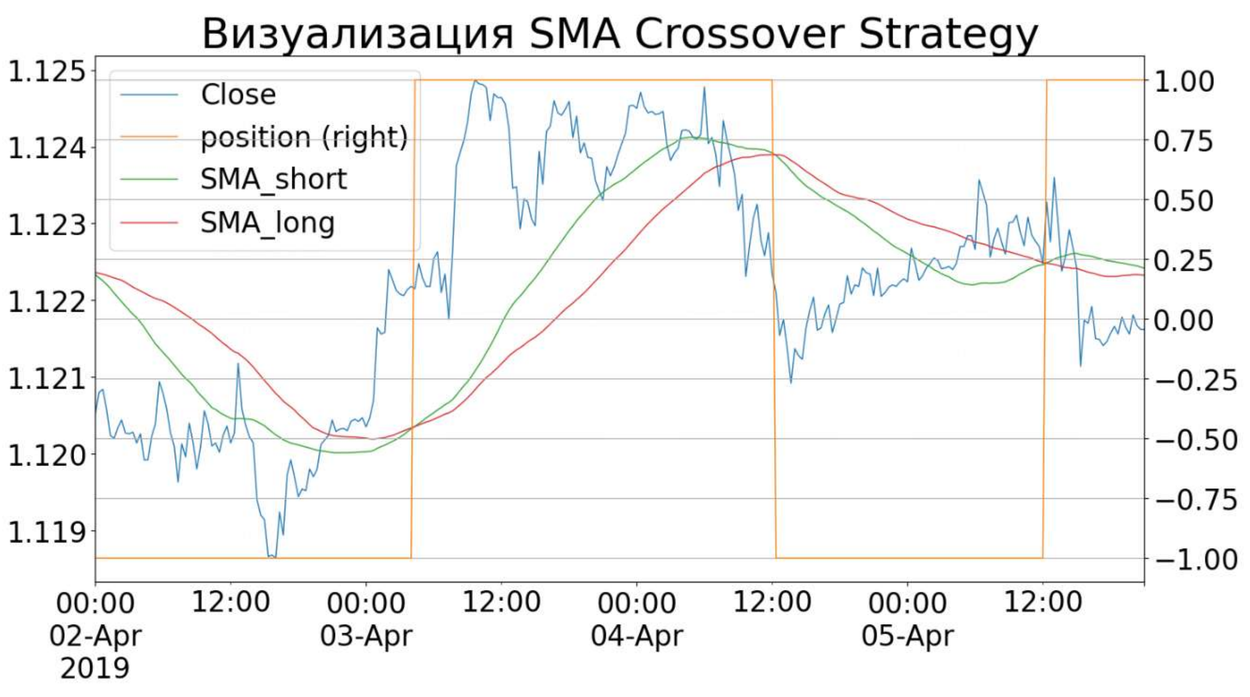
Buy and Hold

- Наш benchmark



Simple SMA Crossover

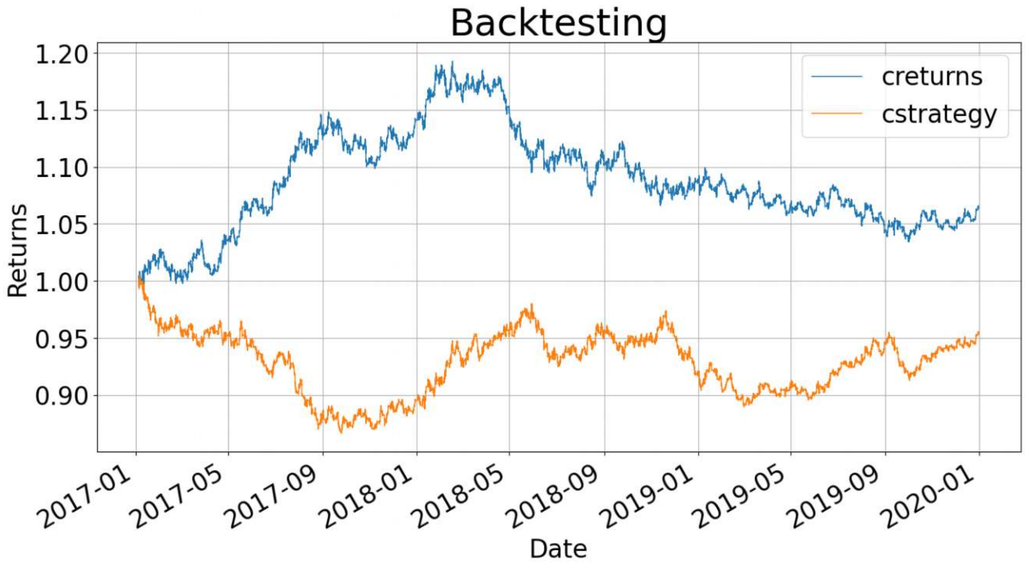
- Данные - инструмент EUR/USD с частотой 20 минут (2017-го года по 2020-ый год)

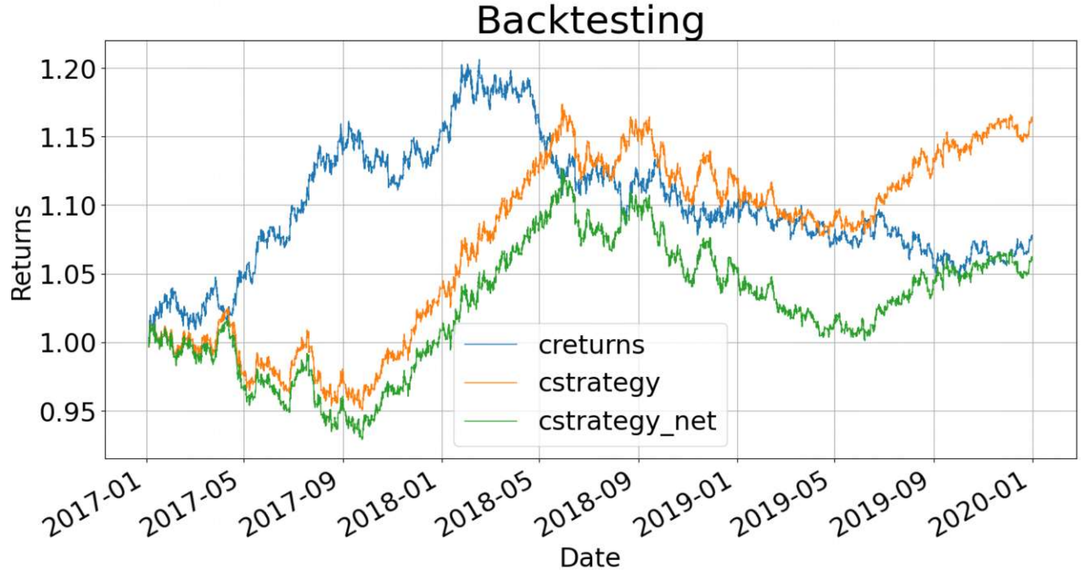


$$short = 61,\ long = 88$$

Backtesting SMA Crossover

- Сравним с базовым SMA : 50 и 150 (количество транзакций 1%)



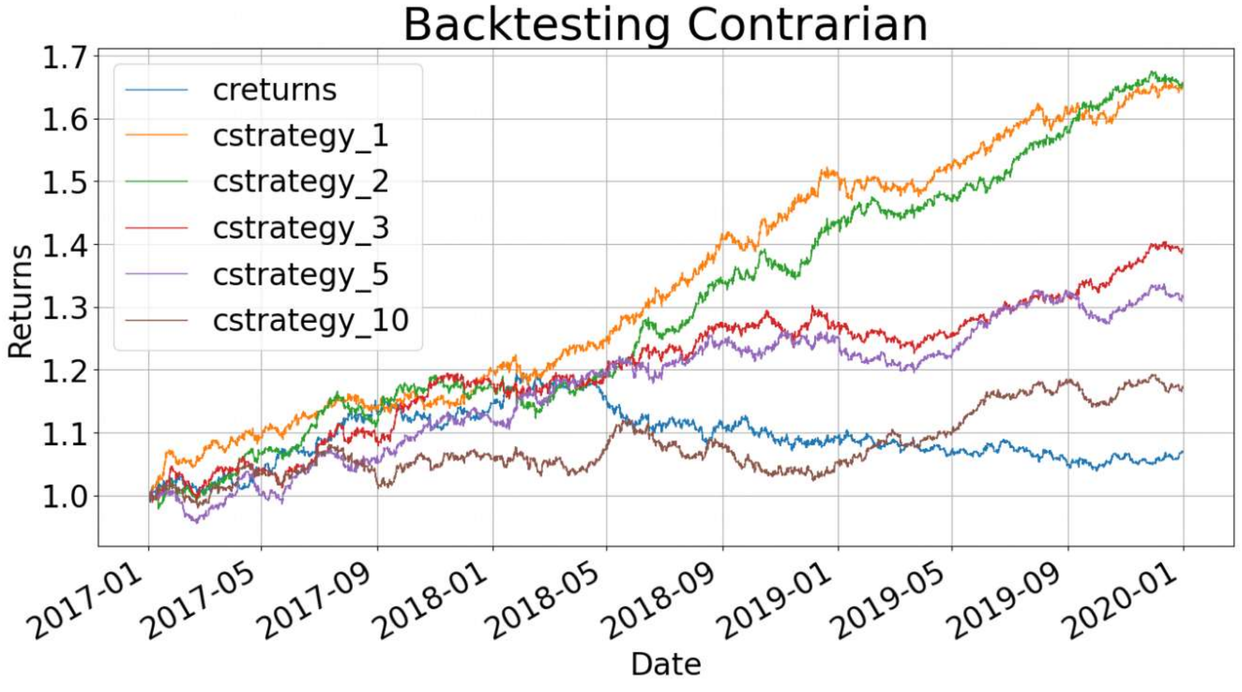


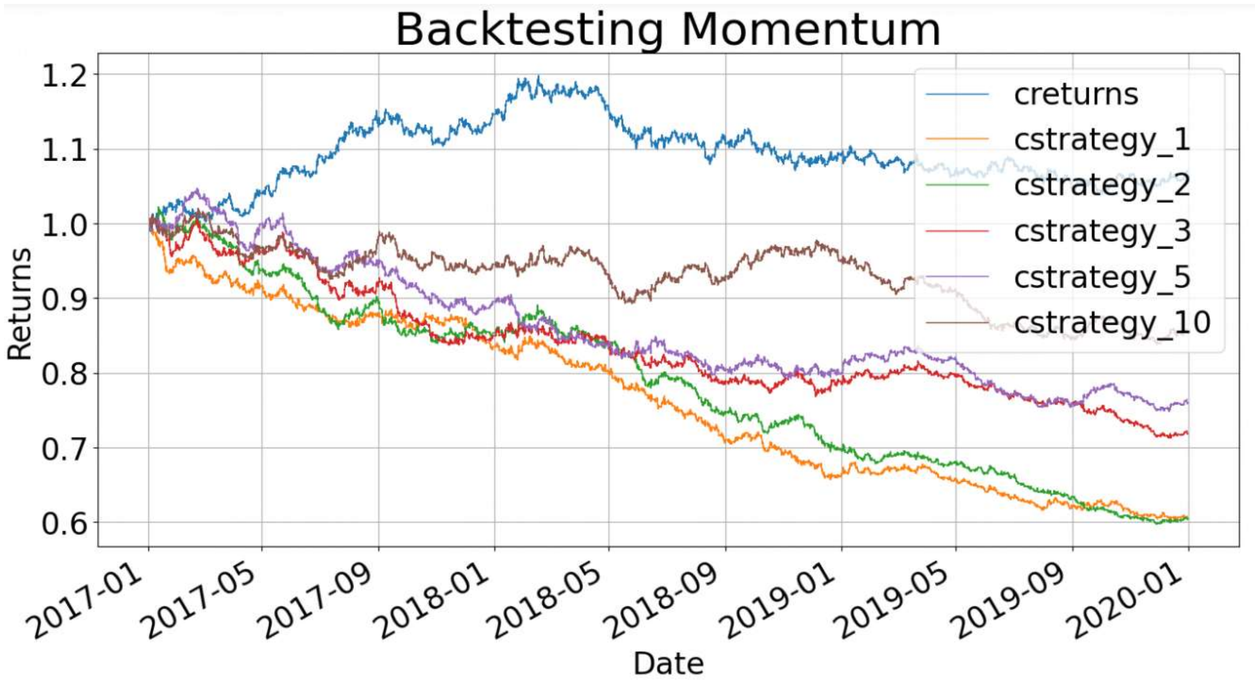
$$\text{cumulative log return} = \sum_t \text{log return}_t$$

$$half\ \ spread = \tfrac{1.5}{2}\ pip.$$

$$proportion = \frac{half\ \ spread}{price_{mean}}.$$

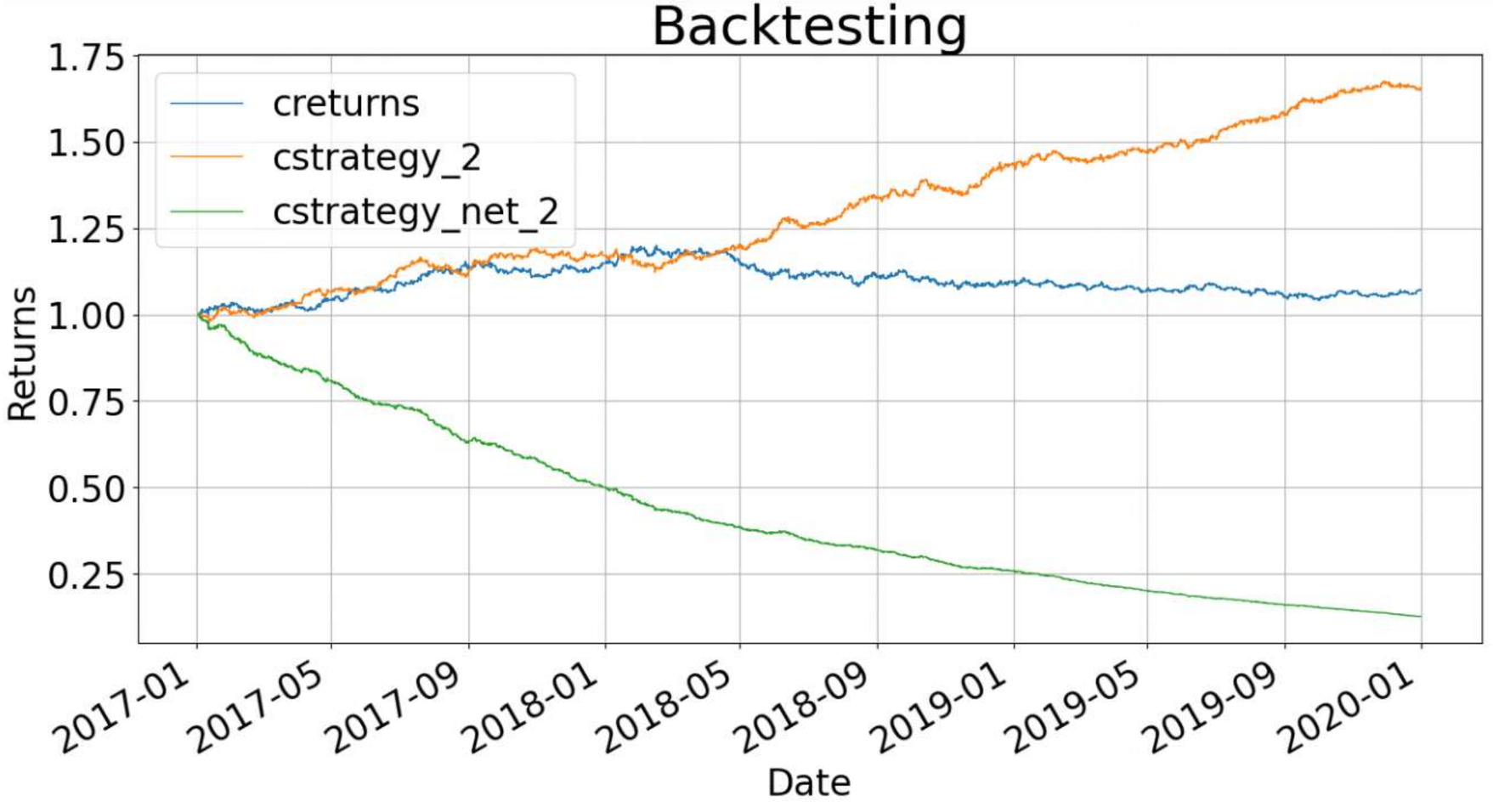
Simple Contrarian/Momentum Strategy





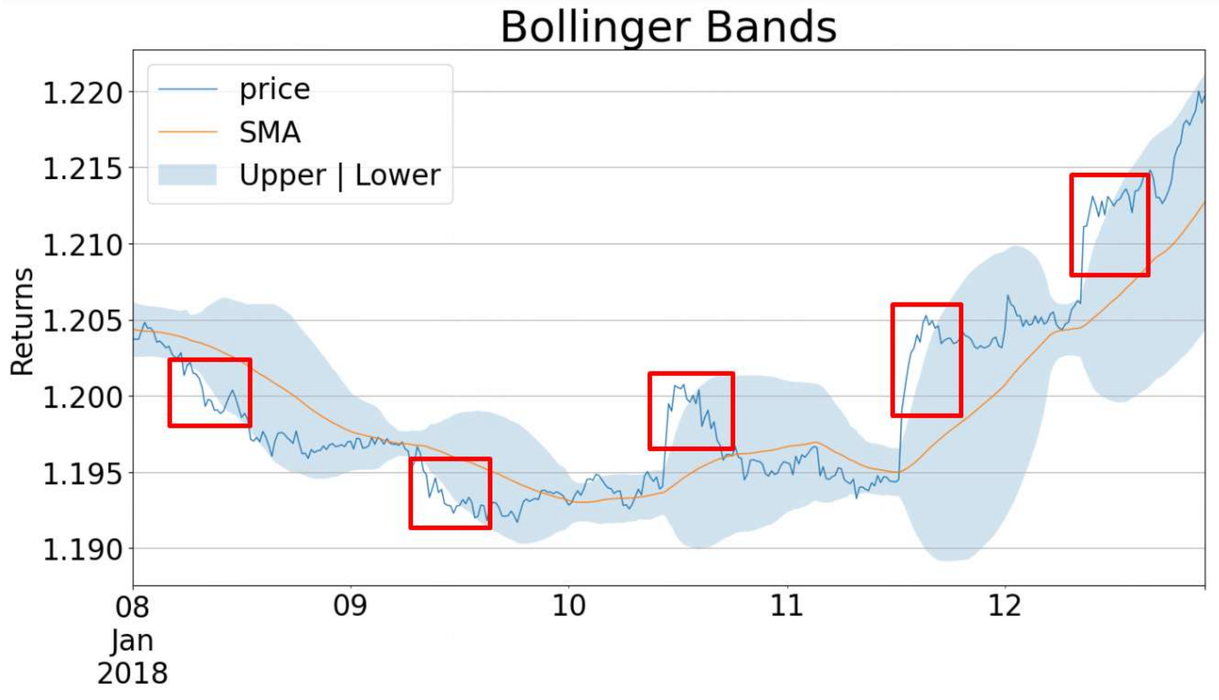
Simple Contrarian/Momentum Strategy

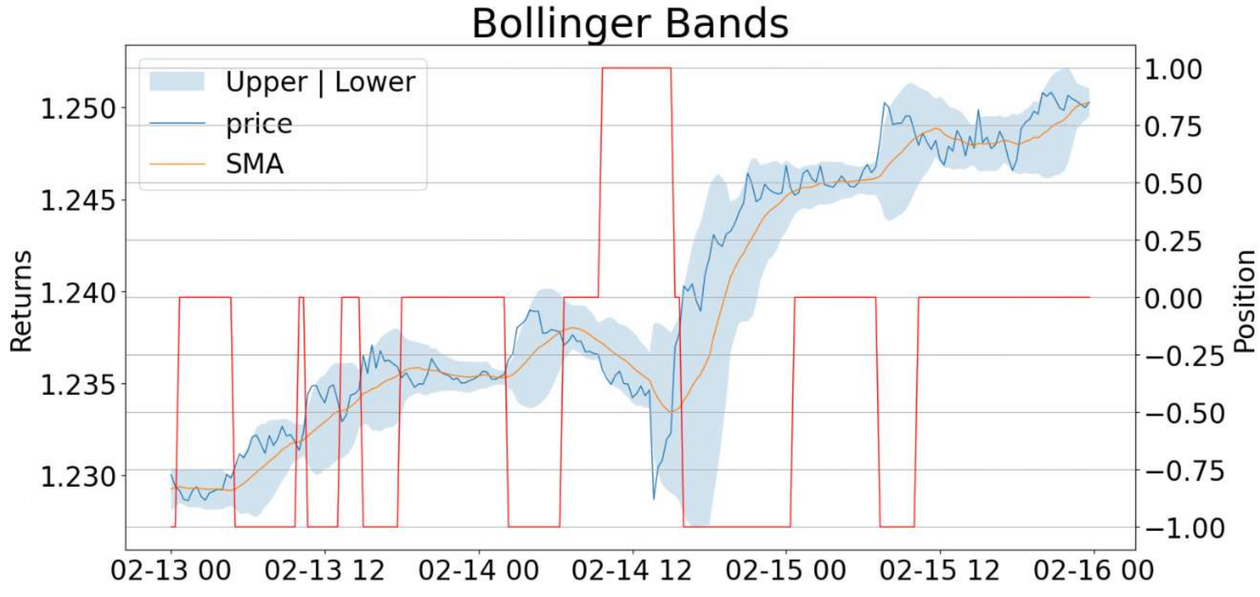
- Частота трейда - 50%



Mean-Reversion Strategy

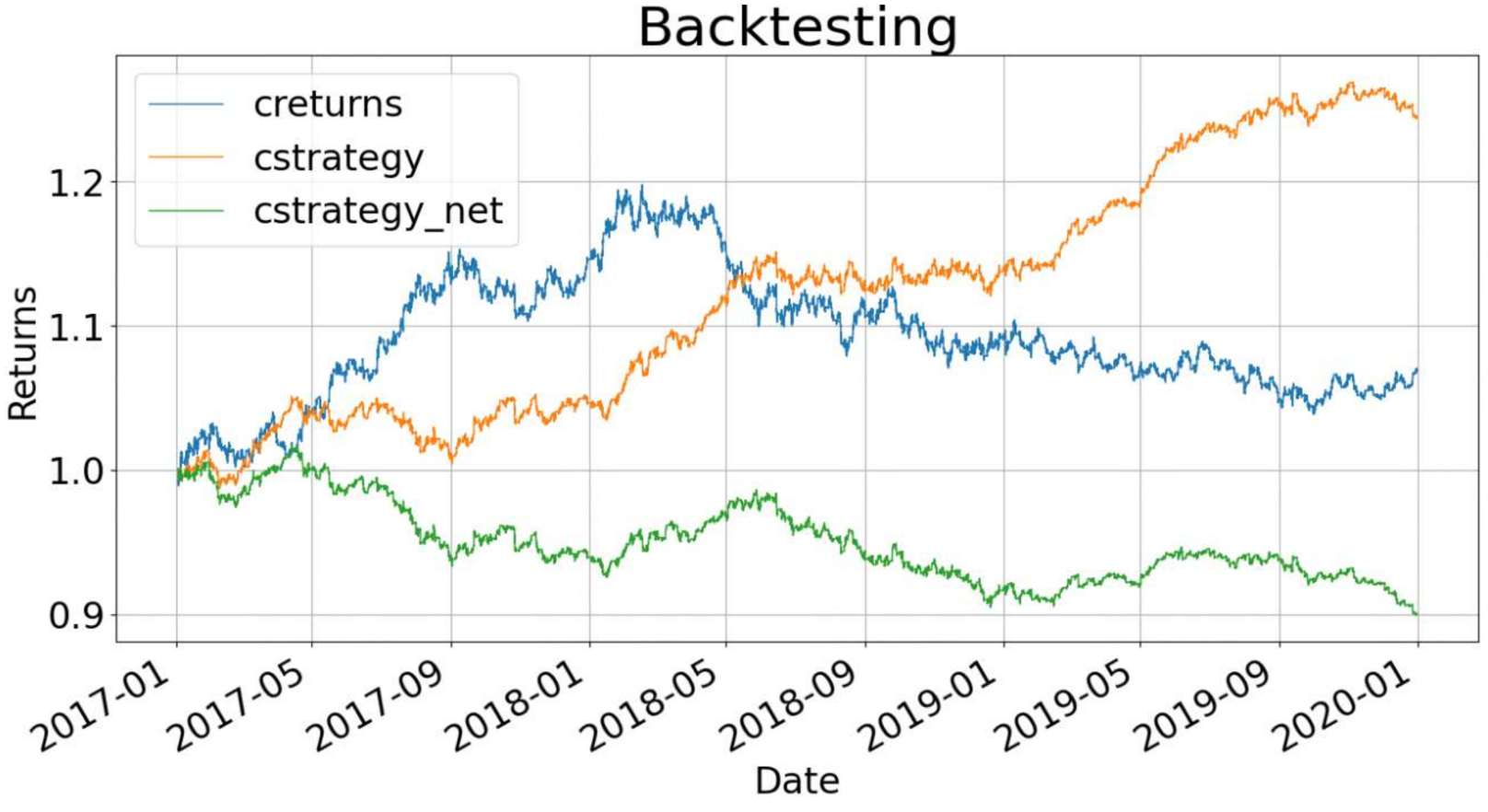
- SMA – 18 и два стандартных отклонения





Mean-Reversion Strategy

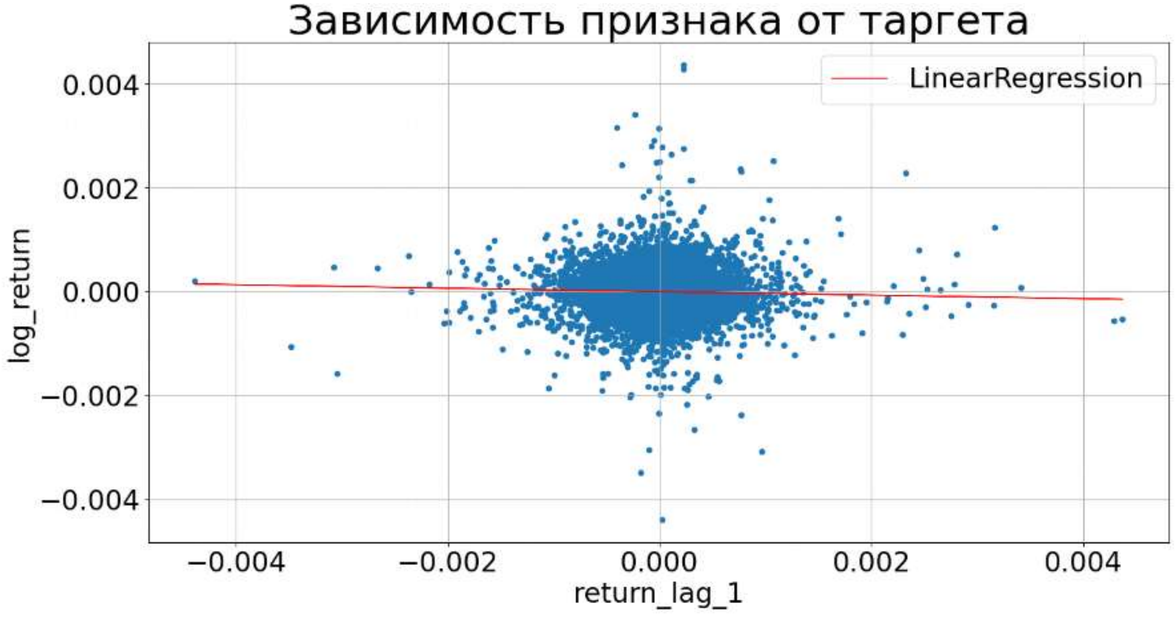
- Частота трейдинга - 4%

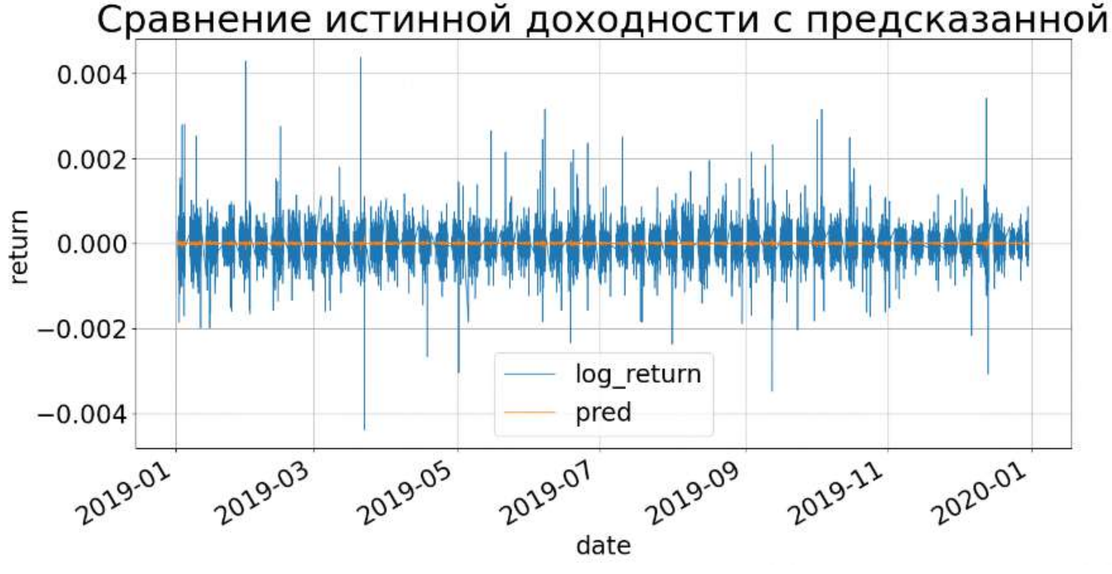


Линейная/Логистическая регрессия

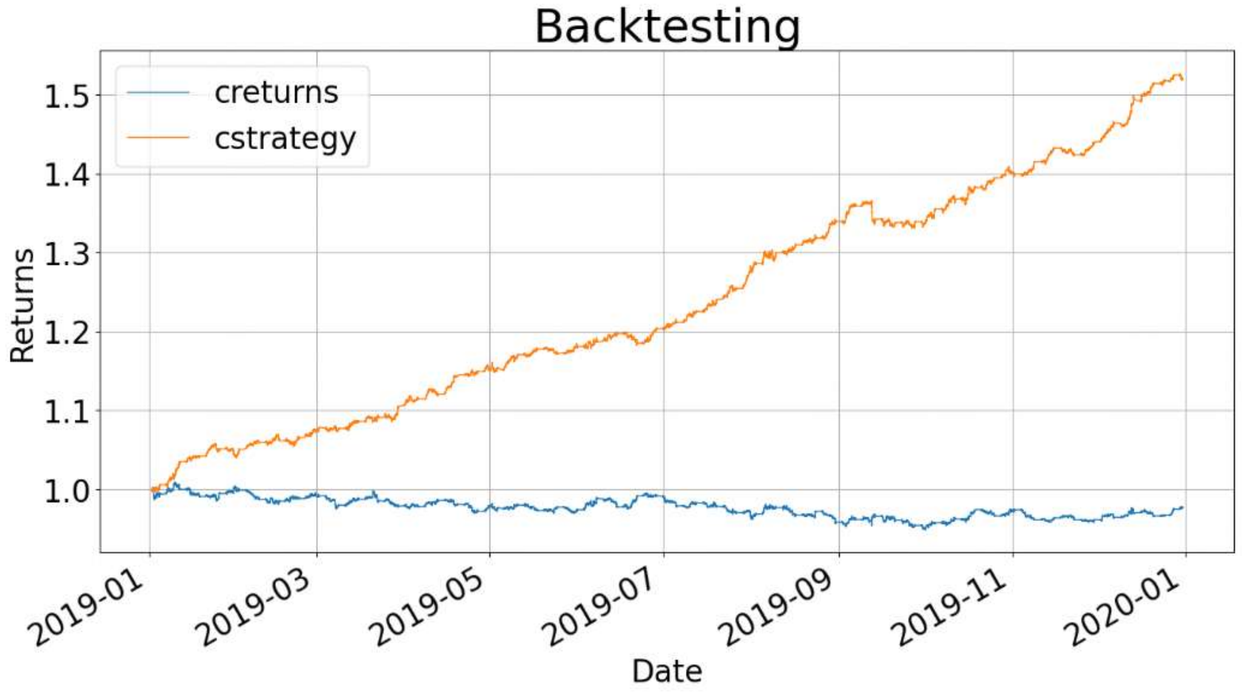
Линейная регрессия

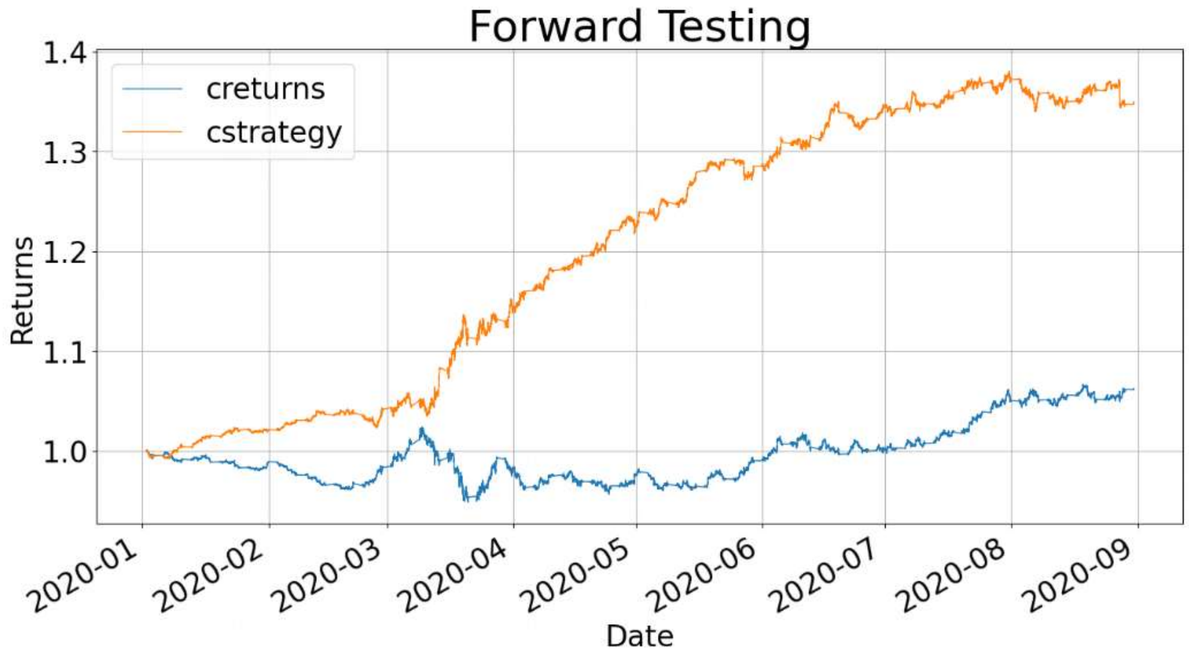
- Данные - инструмент EUR/USD с частотой 5 минут (2019-ый год по 2020-ый год - обучение, 2020-ый год по сентябрь 2020-го года - тест)
- Признаки - смещение доходности на 1-5 вперёд





Backtesting и Forward Testing
Линейная регрессия



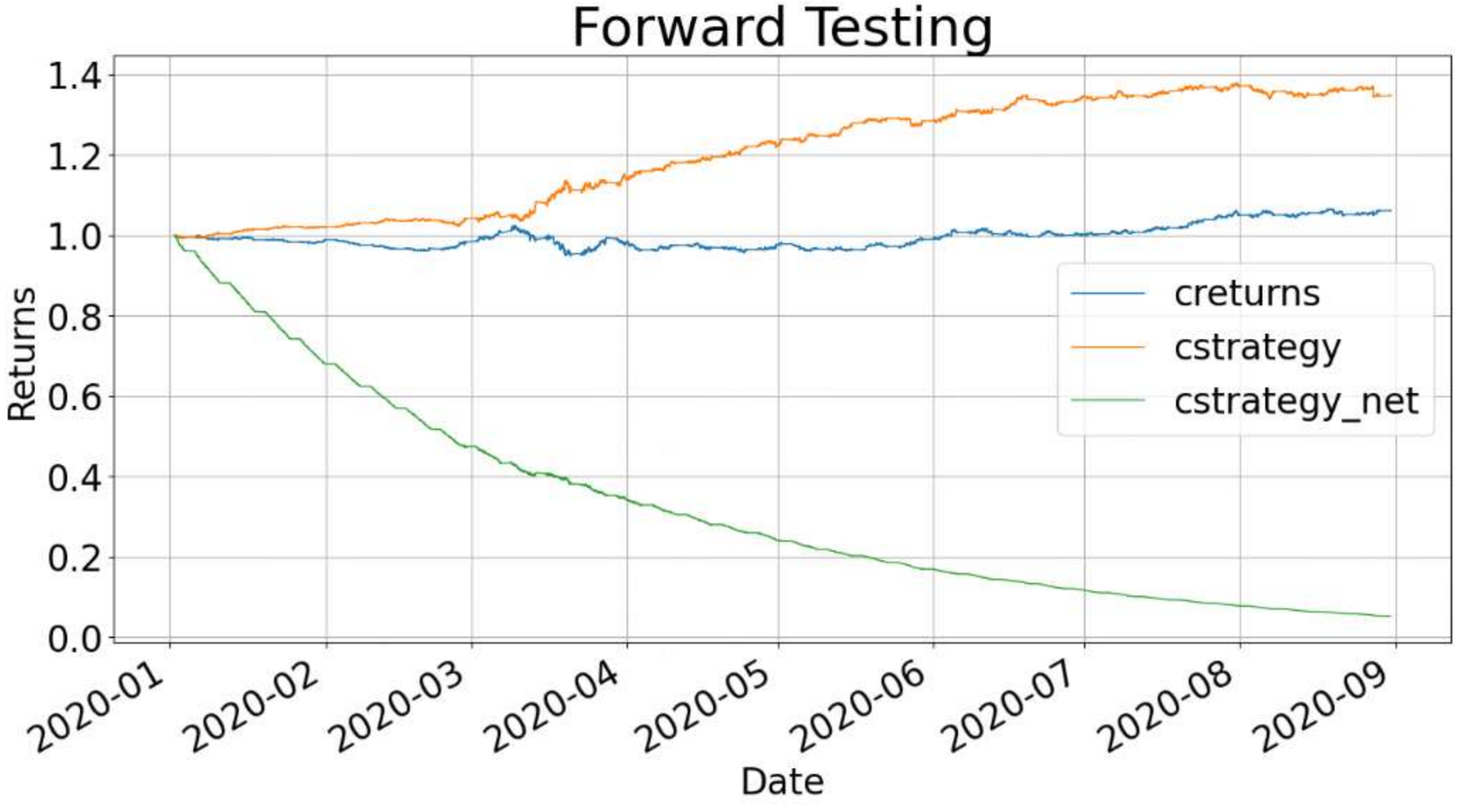


$$Hit\ Ratio\ (HT) = 50.865\%$$

$$Hit\ Ratio\ (HT) = 50.745\%$$

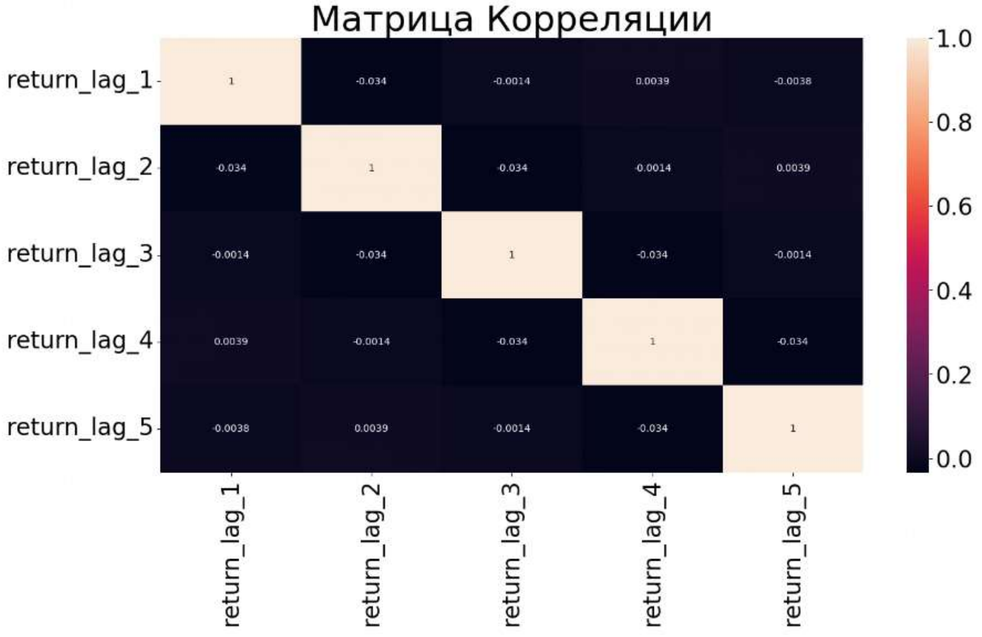
Цена трейдинга

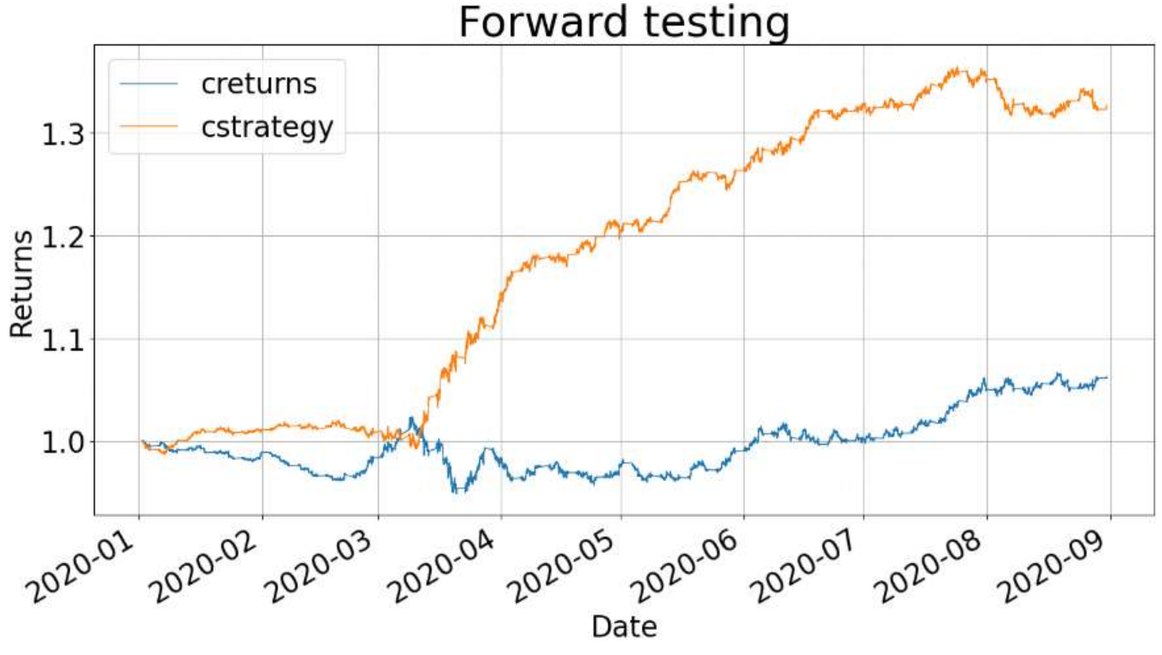
- Частота трейдинга - 50%



Логистическая регрессия

- Маленький коэффициент регуляризации (10^(-6))





$$Hit\ Ratio\ Backtesting\ (HT) = 51.00\%$$

$$Hit\ Ratio\ Forward\ Testing\ (HT) = 50.88\%$$

Полносвязная нейронная сеть (DNN)

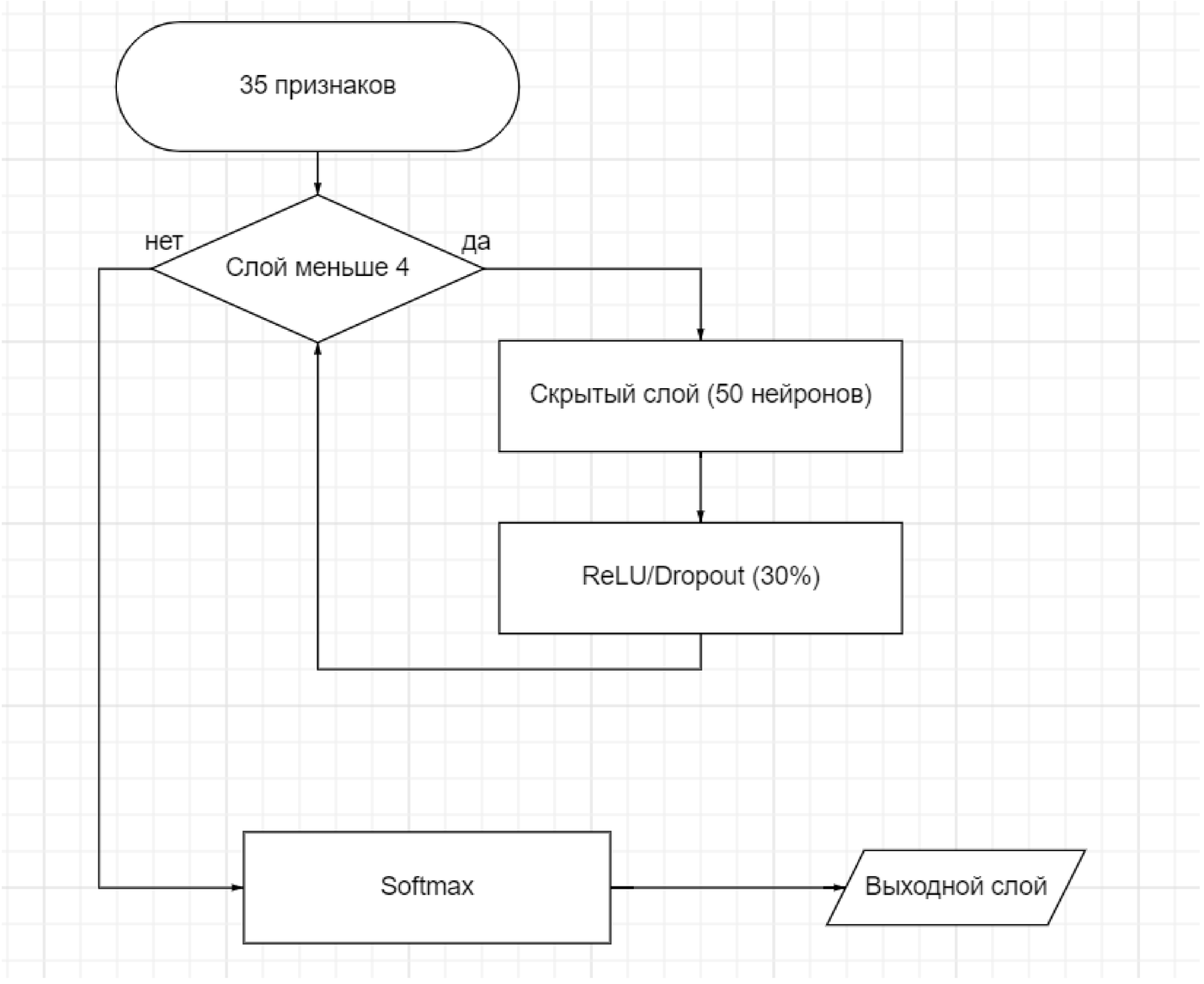
Признаки

- **Direction** - направление движения рынка (наш таргет), где мы ставим единицу, если доходность больше нуля и нуль иначе.
- **Simple SMA** - возьмём короткую SMA с окном 61 и длинную SMA с окном 88 и посчитаем стратегию, как в 3.2.3.
- **Bollinger** - возьмём скользящую среднюю с окном 14 и два стандартных отклонения с таким же окном и посчитаем стратегию, как в 3.2.5.
- **Momentum** - будем брать отрицательный *momentum* (*contrarian*) с окном 2 (как в 3.2.4).

И к генерации ещё дополнительных признаков (отсечки выбраны наугад):

- **Min** - расстояние между минимальной среди 50-ти предыдущих цен и текущей ценой в процентах.
- **Max** - расстояние между максимальной среди 50-ти предыдущих цен и текущей ценой в процентах.
- **Volatility** - разброс (стандартное отклонение) предыдущих 10-ти значений.
- **Функция ошибки** - бинарная кросс-энтропия (BCE)
- **Метрика оценки качества** - *accuracy* (она же *hit ratio*)

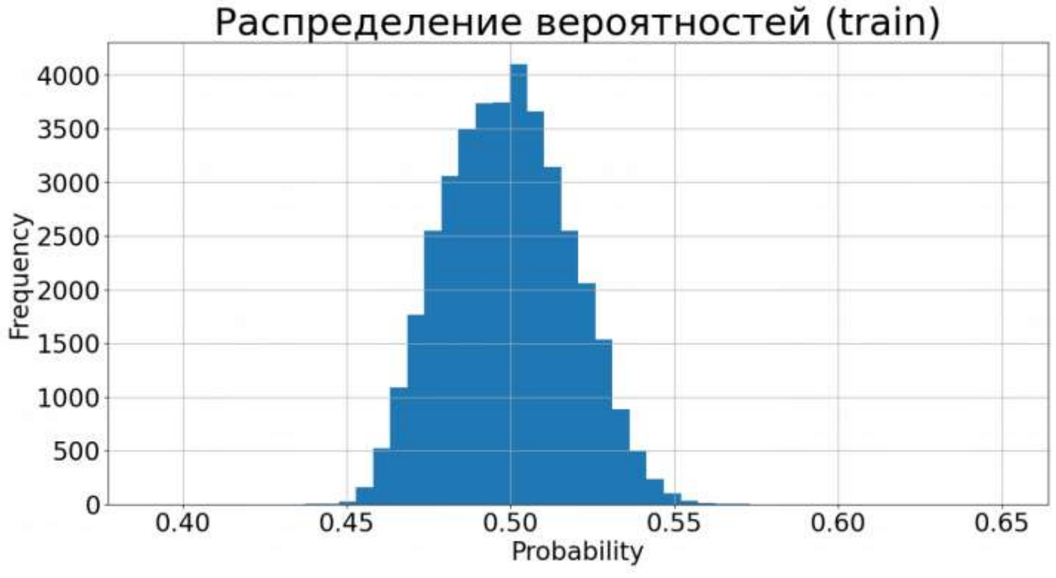
Визуализация DNN



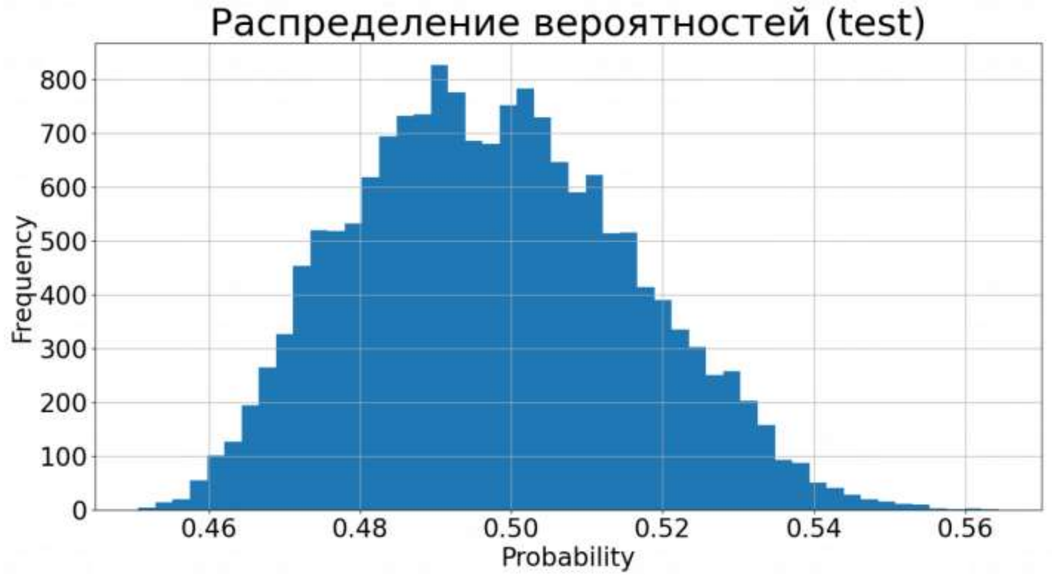
Функция ошибки и метрика качества

Вероятность и отсечка

- Отсечки - продаём, если вероятность ниже 47% и покупаем, если вероятность выше 52%. В остальных случаях придерживаемся прошлых стратегий



train



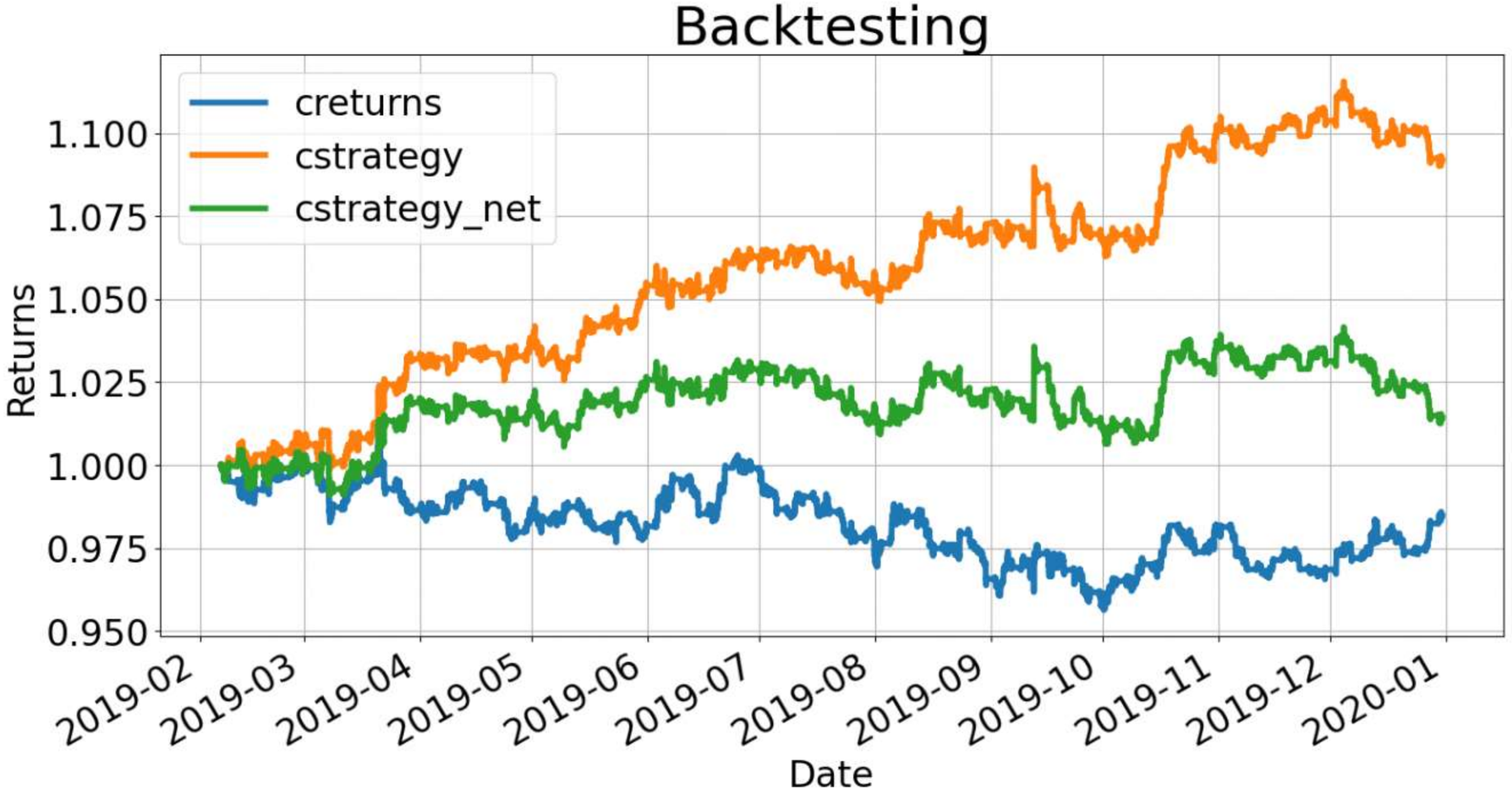
test

$$Accuracy\ (Test) = 52.11\%,\ \text{посчитано моделью}$$

$$Accuracy\ (Test) = 50.67\%,\ \text{посчитано мной с моей отсечкой}$$

Forward Testing

- Количество трейда - 3.3%



Заключение

- Рассматривать стоит логарифмическую доходность, т.к она более приближена к реальности и обладает информативными свойствами
- Доходность распределена не совсем нормально (смещена влево, имеет более тяжёлые хвосты)
- Распределение может быть приближено распределением Стьюдента и через GMM
- Традиционные стратегии по отдельности работают не очень хорошо
- Линейная/Логистическая регрессия не могут уловить сложные зависимости (таргет не зависит от признака линейной)
- Нейронная сеть способна принести прибыль, используя все стратегии, как признаки
- Если gross profit высокий, это не означает, что net profit будет таким же. Более того, он может быть хуже, чем benchmark

Спасибо за внимание!

Приложения

Основные термины и метрики

- Мы будем рассматривать вид трейдинга - **Day Trading** и тип трейдинга - **Derivative Trading** (контракты, цена которых зависит от базового актива **S&P 500**)
- **Волатильность** - степень изменчивости цен на финансовые инструменты в определенный период времени
- **Half Spread** - приблизительная комиссия за открытие и закрытие позиции
- **Trade (Pip) Value** - общая стоимость сделки

$$Half\ Spread\ Costs = -\frac{Spread \cdot Pip\ Value}{2}$$

$$Trade\ Value = Units * Price$$

Вознаграждение/риск

- **Вознаграждение** - потенциальная прибыль
- **Риск** - потенциальные потери (волатильность)

Оценка вознаграждения:

- **CAGR** - годовая средняя ставка роста инвестиций за период n

$$\text{Multiple} = \frac{\text{Last Price}_t}{\text{Initial Price}_t},\ t - \text{заданный промежуток времени}$$

$$\text{CAGR} = \text{multiple}^{\frac{1}{n}} - 1,\ n - \text{количество лет инвестирования(не обязательно целое)}$$

Типы доходности

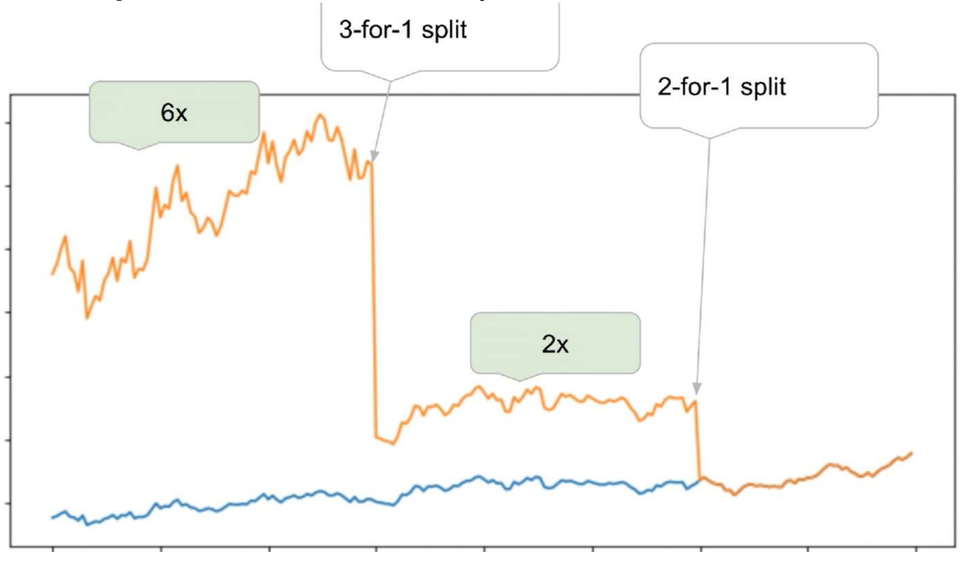
- **(Обычная) доходность** - какую прибыль/убыток получил трейдер - среднее арифметическое не информативно
- **Геометрическая доходность** - схожа с CAGR и меньше или равна среднему арифметическому обычной доходности

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

$$\text{geometric mean} = \text{multiple}^{\frac{1}{n}} - 1,\ n - \text{количество периодов времени (дней/часов/минут)}$$

Скорректированная цена закрытия

- **Stock Split** - процесс, при котором компания увеличивает общее количество своих акций, путём деления одной акции на несколько меньших
- **Дивиденды** - представляют собой денежные выплаты, которые компания делает своим акционерам из своей прибыл
- **Скорректированная цена закрытия -** это цена акции, учтённая после прохождения корпоративных действий (Stock Split/Дивиденды)



Визуализация и доходность

$R_t = \frac{P_t - P_{t-1} + D_t}{P_{t-1}}$, где P_t - цена акции в момент времени t, D_t - выплаченные дивиденды в момент времени t