**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА**

Факультет комп'ютерних наук та кібернетики

Кафедра дослідження операцій

# Кваліфікаційна робота

**на здобуття ступеня бакалавра**

за спеціальністю 113 Прикладна математика

на тему:

## Комбінування машинного навчання та класичних алгоритмів для торгівлі на біржі

Виконав студент 4-го курсу
Русін В'ячеслав Олександрович

_____
(підпис)

Науковий керівник:
асистент Денисов Сергій Вікторович

_____
(підпис)

Засвідчую, що в цій роботі немає запозичень з праць інших авторів без відповідних посилань.

Студент _____
(підпис)

Роботу розглянуто й допущено до захисту на засіданні кафедри дослідження операцій
«____» _____ 2021 р.,
протокол № ___

Завідувач кафедри
Іксанов О. М. _____
(підпис)

Київ – 2021

**ЗМІСТ**

ВСТУП ..... 3

РОЗДІЛ 1 ТЕОРЕТИЧНІ ОСНОВИ УПРАВЛІННЯ ПОРТФЕЛЬНИМИ РИЗИКАМИ ..... 5

РОЗДІЛ 2 ЕЛЕМЕНТИ ТЕХНІЧНОГО АНАЛІЗУ І СТРАТЕГІЇ НА ЇХ ОСНОВІ ..... 9

2.1 Допоміжні індикатори ..... 9

2.2 Свічковий аналіз ..... 9

2.3 Канал Дончіана ..... 10

2.4 Індикатор відносної сили ..... 11

2.5 Перекупленість та перепроданість ..... 12

2.6 Індикатор рівноваги ..... 14

2.7 Фрактальний індикатор ..... 15

2.8 The Ultimate Oscillator ..... 16

РОЗДІЛ 3 Машинне навчання ..... 18

3.1 Логістична регресія ..... 18

3.2 Випадкові ліси ..... 23

ВИСНОВКИ ..... 26

СПИСОК ЛІТЕРАТУРИ ..... 27

# ВСТУП

Люди здавна займались торгівлею. Найрозвинутіші в економічному і культурному плані країни існували на торговельних шляхах та їх перетинах, чи брали активну участь в процесі торгівлі. Проблема ціноутворення, купівлі та продажу, прибутку та втрати завжди була актуальна для людства. На цьому робились великі статки, а хтось втрачав все. Кожен із нас приймає участь в торгівлі, інвестує і сподівається на те, що не тільки не втратить, але й заробить.

Інвестиції сприяють економіці, розвиненю бізнеса, створенню стартапів, науковому прогресу. Але й є приклади ринкових та економічних криз, які сталися через те, що певні інвестиції були сильно переоцінені в свідомості людей, що створювало бульбашки на ринку, після краху яких нам був потрібен певний час на відновлення.

Є різні способи аналізу для інвестування, але до недавнього часу займалась цим і приймала рішення лише людина на основі свого досвіду і суб'єктивного передчуття. Але зараз все почало змінюватись із розвиненням обчислювальної техніки,  зі створенням торговельний , автоматизованих стратегій. В наш час найбільший прибуток з торгів дійсно отримують люди, а не алгоритми, проте вже більше половини всіх торгових операцій проводять торгівельні роботи. Алгоритм навряд зможе передбачити потенціал компанії в довгостроковій перспективі, але в ймовірнісній формі може передбачувати короткостроковий напрям ринку за рахунок аналізу настрою та положення ціни в певний час, що добре підходить для активної роботи з капіталом.  Такі алгоритми можна використовувати як підказки чи підкріплення своїх очікувань при роботі з цінними активами.

Дослідження полягає в знаходженні закономірностей ринка і їх використанні через торгівельні алгоритми.Є певна кількість теорій, які описують ринок, його поведінку та напрям. Але як показує практика ефективною оцінкою кожної теорії чи торговельного , підходу може бути,

якщо не прибуток, то хоча б збереження свого капіталу, через це можна дозволити собі певну упередженість чи суб’єктивність.

В роботі будуть проаналізовані різні індикатори, роботу яких буде покладено в фундамент торговельних алгоритмі, а найефективніші будуть об’єднані і використані для прогнозування ринку через машинне навчання.

# РОЗДІЛ 1. ТЕОРЕТИЧНІ ОСНОВИ УПРАВЛІННЯ ПОРТФЕЛЬНИМИ РИЗИКАМИ

## 1.1 Огляд біржі і засоби аналізу

Біржа, біржова торгівля або місця торгівлі є організованим ринком, де торгуються цінні папери, товари, валютні, ф'ючерси та опціони контракти продаються і купуються у вигляді стандартизованих біржових угод. На біржі укладаються угоди по біржових товарах, в результаті чого утворюється динаміка ціни тільки під впливом ринкового попиту та пропозиції, що дає змогу орієнтуватися учасникам ринку та прогнозувати хід торгів в майбутньому. Біржі беруть початок від купецьких сходок в італійських містах 13—14 ст., а пізніше в торговельних містах інших країн Західної Європи. В світовій торгівлі 17 ст. велику роль відіграла Амстердамська біржа, де сформувались основні типи біржових операцій, відомих у сучасних капіталістичних країнах. Вона вважається найстарішим "сучасним" ринком цінних паперів у світі. Біржі можна розділити на:

-За проданими предметами:

Фондова біржа або біржа цінних паперів

Товарна біржа

Валютний ринок

-За видом торгівлі:

Класичний обмін - для спотових торгів

Біржа ф’ючерсів або біржа ф’ючерсів та опціонів

Біржа зводить людей, які хочуть щось придбати і продати за щось інше. Візьмемо аналогію із золотом: покупці говорять, яку ціну в перерахунку на долари вони готові купувати золото, розмістивши "замовлення на купівлю". Якщо є "замовлення на продаж" за тією самою ціною, то біржа зіставляє ці два замовлення, відбувається торгівля і золото

та долари змінюють власників. Якщо немає замовлення за тією ж або кращою ціною, тоді це замовлення залишається "відкритим", доки майбутнє замовлення не співпаде йому (або скасовується особою, яка його розмістила).

Трейдер — учасник біржової торгівлі, людина, яка проводить торгово-інвестиційні операції з цінними паперами (облігаціями, акціями та іншими фінансовими інструментами) для отримання доходу. Робота полягає у покупці/продажу пакетів акцій, валюти та інших активів за однією ціною та перепродаж їх за іншою ціною. За рахунок різниці цін трейдер збагачується. Не все так просто, як може здаєтись – для справді прибуткової роботи необхідно досконало вивчити ринок і фактори, які на нього впливають, навчитися розуміти графіки цін на валюту та акції, правильно інтерпретувати все вивчене і тільки тоді будуть максимальні шанси на укладання вигідних угод купівлі/продажу.

За стратегіями торгівлі трейдерів розподіляють на:

- Денних трейдерів — трейдер, що протягом доби проводить не менше однієї транзакції;
- Позиційних трейдерів — як правило 2-4 транзакції протягом місяця;
- Середньострокових трейдерів — декілька транзакцій в рік;
- Довгограйючих трейдерів — 1 транзакція в пів року, а то й в декілька років.

Залежно від стратегії поведінки трейдерів на фондовій біржі поділяють на дві основні групи:

- «Бики» — трейдери, які заздалегідь скуповують цінні папери за низькою ціною, сподіваючись на те що їхня ціна з часом зросте, щоб потім продати їх за вигідною для них ціною. Така стратегія піднімає ціну цінних паперів, подібно бику, що піднімає рогами свого ворога.
- «Ведмеді» — трейдери, які передбачають, що цінні папери будуть дешевшати в ціні, для їх подальшого придбання. Своїми діями вони

можуть направити ціни вниз, подібно ведмедю, що б'є ворога лапою зверху вниз.

Зараз для торгівлі на біржі необхідно лише укласти договір з брокером про надання послуг пов’язаних з торгівленю. Загальновідомим фактом є те, що значна частина трейдерів зазнає невдачі. Різні джерела стверджують, що 70%, 80% і навіть більше 90% трейдерів втрачають гроші і в кінцевому підсумку йдуть.

Основні причини через які люди втрачають гроші в трейдингу:

-Позиції проти тренду

-Малий стартовий капітал

-Невдале управління ризиком

-Жадібність

-Нерішуча торгівля

-Відмова помилятись

Суб’єктивність людей грає велику роль в трейдингу – вона дає змогу одним заробляти собі на життя, а іншим втрачати кошти. Тому спроби створити автоматичну торгівельну систему є цілком зрозумілими. Поняття автоматизованої торгової системи вперше було введено Річардом Дончіаном у 1949 році, коли він використовував набір правил для купівлі та продажу коштів.

Технічний аналіз є аналіз методології прогнозування напрямку цін на основі вивчення минулих ринкових даних, в першу чергу ціни і обсягу. Технічний аналіз використовує моделі та правила торгівлі, засновані на трансформації ціни та обсягу, такі як : індекс відносної сили, рухоме середнє, регресія, кореляція міжринкових та внутрішньоринкових цін, ділові цикли, цикли фондового ринку або, класично, шляхом розпізнавання графічних візерунків. Будучи аспектом активного управління капіталом, суперечить більшій частині сучасної теорії портфеля. Ефективність технічного аналізу заперечується гіпотезою ефективного ринку, де зазначається, що ціни на фондовому ринку по суті непередбачувані. Непередбачуваність

цін – це логічно, проте сама суть роботи по технічному аналізу базується не на загальному прогнозуванні ринка, а на ймовірнісній оцінці руху ціни в моменти, коли через певні індикатори ми отримуємо сигнали про настрій учасників рикну. Активний трейдинг також неможливий без правильного розуміння та оцінці ризиків, які можуть спіткати трейдера. Саме тому суб’єктивність людини грає найважливішу роль в цій справі. Індиктори ж використовують для отримання об’єктивної оцінки ринку.

# РОЗДІЛ 2.ЕЛЕМЕНТИ ТЕХНІЧНОГО АНАЛІЗУ І СТРАТЕГІЇ НА ЇХ ОСНОВІ

## 2.1 Допоміжні індикатори

Моментум показує різницю між сьогоднішньою ціною закриття та закриттям N днів тому.

Для згладжування ціни використовують різні середні значення:

Просте ковзке середнє $\frac{1}{k}\sum_{i=n-k+2}^{n} p_i$, де $p_i$-ціна.

Експоненційне ковзке середнє:

$EMA(t) = EMA(t-1) + 2P(t) - EMA(t-1)$, $P(t)$-ціна закриття.

$MACD = EMA_5(P) - EMA_{26}(P)$

Просте середнє:

SMA = $\frac{A_1+A_2+\cdots+A_n}{n}$

## 2.2 Свічковий аналіз

Свічкова аналіз - це вид прогнозування цінових коливань за допомогою спеціального виду графічного відображення ціни - японських свічок, якіпоказують волатильність, або «розмах цін». Цей метод аналізу був розроблений японським трейдером Мунехіса Хомма, який жив в середньовіччі. Він створив систему, яка дозволяє прогнозувати цінові коливання без будь-яких додаткових інструментів технічного аналізу. В цій роботі я буду використовувати лише свічкові моделі, які сигналізують про корекцію або закінчення тренду: молот, зірка, хаммер, волчок, доджи, три ворони та три солдати. Вони мають подібний вид: ціна відкриття та закриття свічі знаходяться на невеликій вістані, у порівнянні з ціновим діапазоном протягом всієї свічі.

Кожна з цих моделей передає певний настрій внутрішньоденних трейдерів, який вказує, що ані «бики», ані «ведмеді» вже не мають

переваги над ринком, що може стати підставою для розвороту ціни в краткостроковій або довгостроковій перспективі.

Нажаль, жодна з цих свічкових моделей не є ефективною самом по собі.

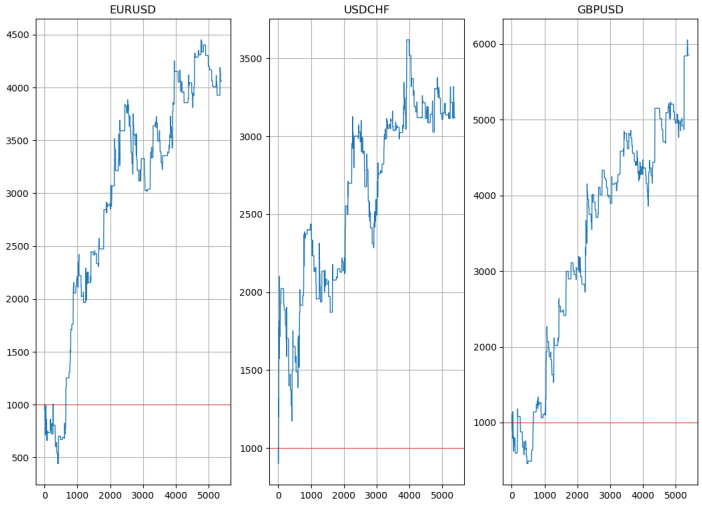
### 2.3 Канал Дончіана

Канал Дончіана - технічний індикатор, розроблений Річардом Дончяном, є варіацією правила пробою торгового інтервалу. Будується взяттям найвищої і найнижчої ціни за попередні *n* періодів з подальшим позначенням області між ними на графіку.

Дончян рекомендував використовувати свій індикатор для денних таймфреймів з інтервалом *n = 20.* Основна ідея полягає в тому, щоб зробити діапазони максимально об'єктивними (тобто вимірюваними), а потім торгувати на прориві (тобто на початку тренду). Отже, мета тесту - з'ясувати, чи може цей показник додати значення в нашу загальну торгову систему чи ні. Чи дає це хороші сигнали? Чи слід сприймати серйозно сигнали цього індикатора?

Стратегія: Купівля, коли ринок перевершує останній верхній канал та продаж, коли ринок прориває останній нижній канал.

Обрав довгострокові параметри для бек-тесту : використовую щоденні графіки з періодом Дончі 60 днів. Це означає, що якщо ринок перевершує верхній дончійський канал за останні 60 днів, тоді ми вкладаємося в тренд. Я включив простий стоп-лосс у розмірі 100 піпсів для трьох валютних пар: EURUSD, USDCHF та GBPUSD.




### 2.4 Індикатор відносної сили

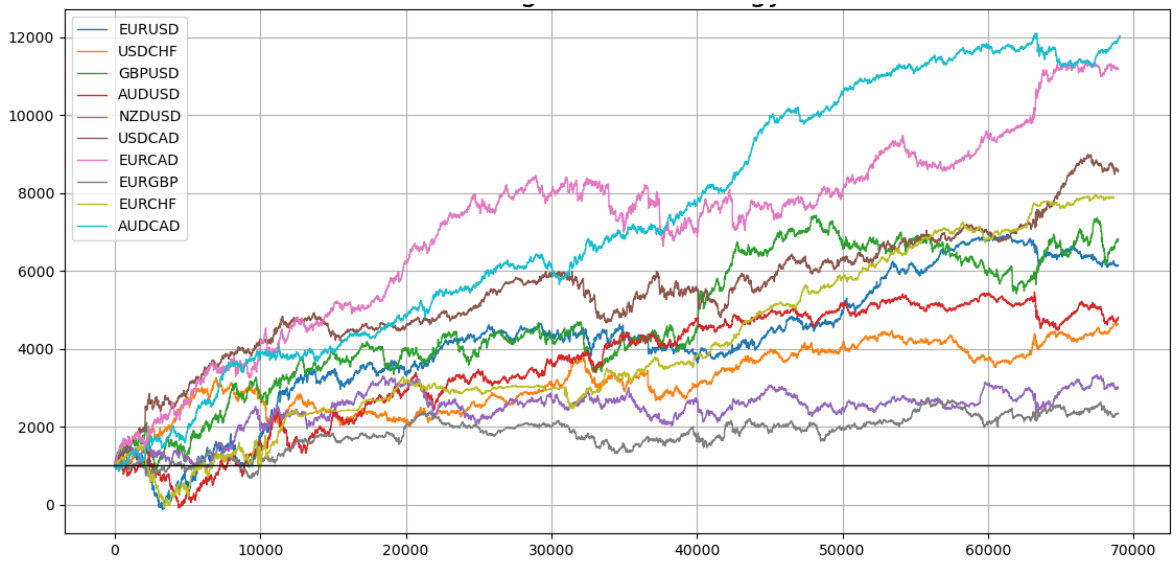
Індекс відносної сили RSI - обчислюється досить простим способом. Спочатку ми починаємо з того, що беремо різницю в цінах за один період. Це означає, що ми повинні відняти кожну ціну закриття з тієї, що перед нею. Потім ми обчислимо згладжене середнє позитивних різниць і поділимо його на згладжене середнє негативних різниць. Останній розрахунок дає нам відносну силу, яка потім використовується у формулі RSI для перетворення в міру від 0 до 100. RSI подає сигнали, які вказують інвесторам купувати, коли цінні папери або валюта перепродані, і продавати, коли вони перекуплені.

RS = середній приріст / середній збиток в ціні

$$\text{RSI} = 100 - \frac{100}{(1+RS)}$$

Порівняно менш відомий показник, який називається трикутною ковзною середньою, використовується як форма надмірного згладжування. Це ковзна середня, що застосовується до ковзної середньої.

Стратегія: покупка, коли трикутний RSI досягає 30 з двома попередніми значеннями вище 30, продаж, коли трикутний RSI досягає 70 з двома попередніми значеннями нижче 70. Утримую цю позицію до отримання нового сигналу або зупинки системою управління ризиками.




### 2.5 Перекупленість та перепроданість

Виявлення статистичних рівнів перекупленості та перепроданості за технічними показниками. Для кращого результату потрібно знайти більш об'єктивний та динамічний спосіб виявлення перепроданих та перекуплених рівнів за технічними показниками, тому що коли ринки рухаються, змінюється їх структура та статистичні властивості. Рухомі середні допомагають підтвердити і керувати тенденцією. Вони є найвідомішим технічним показником, і це через їх простоту та перевірений досвід додавання вартості до аналізів. Можно використовувати їх, щоб знайти рівні підтримки та опору, зупинки та цілі та зрозуміти основну тенденцію. Ця універсальність робить їх незамінним інструментом у нашому торговому арсеналі. Як випливає з назви, це ваше просте середнє значення, яке використовується скрізь у статистиці. Це просто загальні значення спостережень, розділені на кількість спостережень.

Для  виявлення об'єктивних рівнів перекупленості та перепроданості припустимо, що рівні динамічні і рухаються відповідно до останніх розрахунків RSI. Це означає, що ми не будемо говорити, що 30/70 - це перепродані рівні, а скоріше, ми обчислимо їх на основі довгострокової ковзної середньої на RSI. У нашому випадку це буде 500-періодна ковзна

середня, що застосовується до 20-періодної RSI. Також для цього застосуємо до значень індикатора діапазон Болліндже́ра.

Одним із стовпів описової статистики або будь-якого базового методу аналізу є концепція середніх показників. Середні показники дають нам уявлення про наступне очікуване значення, враховуючи історичну тенденцію. Вони також можуть бути репрезентативним числом більшого набору даних, який допомагає нам швидко зрозуміти дані. Ще однією опорою є концепція волатильності. Волатильність - це середнє відхилення значень від їх середнього значення. Стандартне відхилення просто вимірює середню відстань від середнього, переглядаючи окремі значення та порівнюючи їх відстань із середнім.

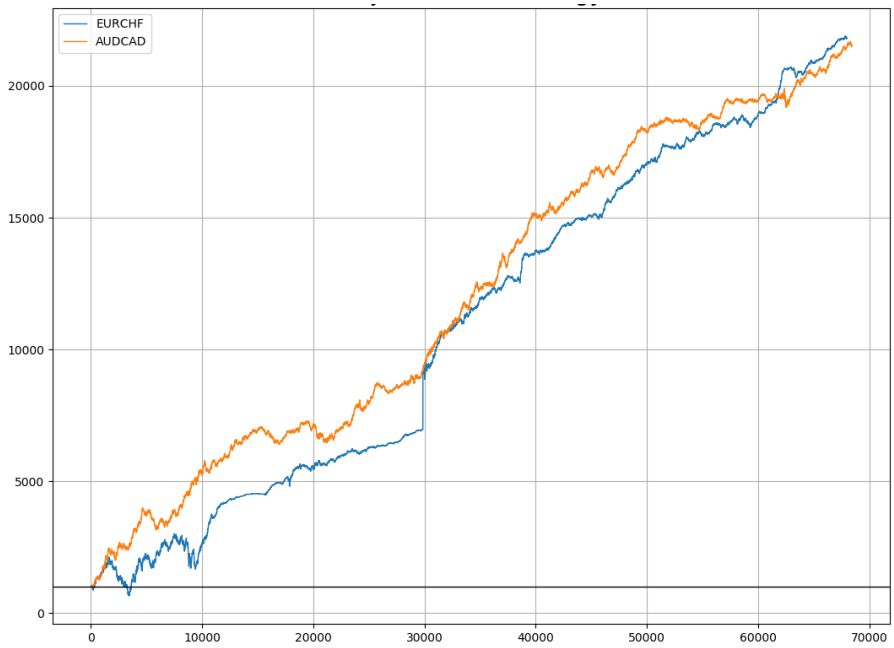
Діапазони Болліндже́ра - одне з перших речей, яким ми повинні навчитися, аналізуючи часові ряди. Це пов’язано з їх вагомими статистичними міркуваннями, їх широким розповсюдженням серед учасників ринку та їхнім успіхом при застосуванні в торгових стратегіях.  Коли ціни рухаються, ми можемо обчислити рухоме середнє навколо них, щоб краще зрозуміти їхнє становище щодо їх середнього значення. Ідея смуг Болліндже́ра полягає у формуванні двох бар'єрів, розрахованих на консанту, помножену на стандартне відхилення, що ковзає. Вони, по суті, є бар'єрами, які дають ймовірність того, що ринкова ціна повинна міститися в них. Нижня смуга Болліндже́ра може розглядатися як динамічна опора, тоді як верхня смуга Болліндже́ра може розглядатися як динамічний опір.

$$\text{верхній бар}'\text{єр} = MA + const\sqrt{\frac{\sum_{i=1}^{n}(y_i - MA)^2}{n}}$$

$$\text{нижній бар}'\text{єр} = MA - const\sqrt{\frac{\sum_{i=1}^{n}(y_i - MA)^2}{n}}$$

За замовчуванням індикатор обчислює 20-періодову просту ковзну середню і два стандартних відхилення від ціни, а потім складає їх разом, щоб краще зрозуміти будь-які статистичні крайнощі. Це означає, що в будь-який час ми можемо обчислити середнє та стандартне відхилення останніх 20 спостережень, які ми маємо, а потім помножити стандартне відхилення на постійну. Нарешті, ми можемо додати і відняти його від середнього, щоб знайти верхню та нижню смуги.

Стратегія: купівля, коли RSI досягне нижнього бар'єра з двома попередніми показниками над нижнім бар'єром; продаж, коли RSI досягне верхнього бар’єра, коли два попередні показники будуть нижчими за верхній бар’єр. Позиція утримується до отримання іншого сигналу або зупинки алгоритмом управління ризиками.




### 2.6 Індикатор рівноваги

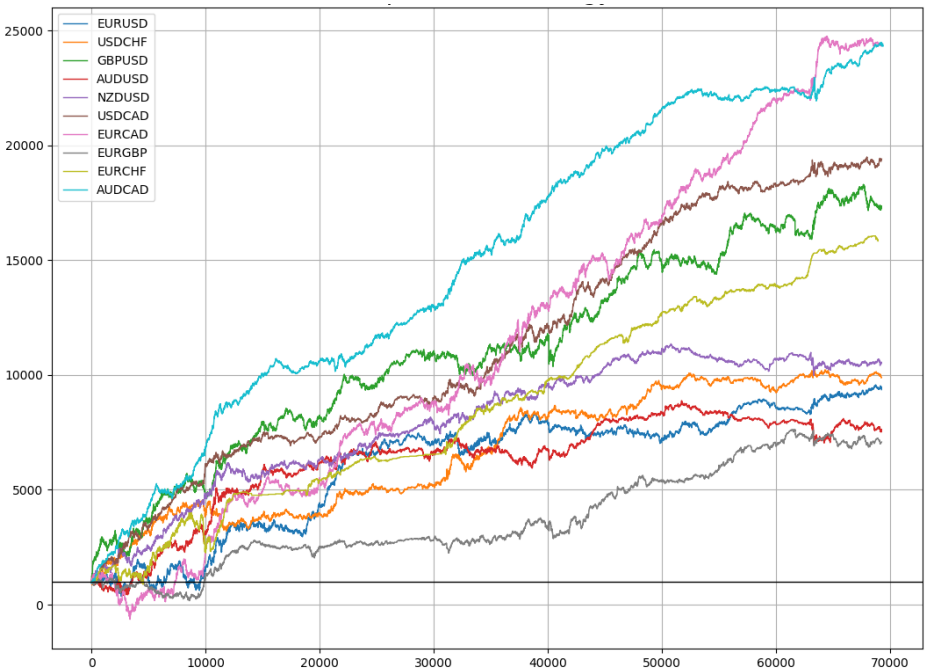
Індикатор рівноваги - прибуток від реверсії середньої величини.

Стан рівноваги можна виміряти багатьма складними способами, але в цій роботі це зроблю, як експоненціальне згладжене середнє значення відстані між ринковою ціною та її ковзною середньою.

Алгоритм: Обчислити просте рухоме середнє, відняти ринкову ціну від її рухомого середнього, обчислити експоненціальне рухоме середнє  за

відніманими даними. Результатом є 5-періодний індикатор рівноваги, який ми будемо використовувати для формування сигналів, що повертають середнє значення.

Стратегія: покупка, коли індикатор рівноваги досягає -0,001 з двома попередніми значеннями, більшими за -0,001; продаж, коли показник діапазону процентних відсотків досягає 0,001, а два попередні значення нижче 0,001. Позиція утримується до отримання нового сигналу або до зупинки системою управління ризиками.




## 2.7 Фрактальний індикатор

Гіпотеза ефективного ринку не враховує безліч аномалій і повторюваних експлуатованих закономірностей в фінансових активах. Саме тому активне управління портфелем як і раніше є домінантною стороною в порівнянні з пасивним інвестуванням.

Фінансові ринки не є абсолютно випадковими, вони є випадково-подібними, тобто мають низьке співвідношення сигнал / шум. В цьому індикаторі нам доведеться використати теорію хаосу.

Теорія хаосу - це дуже складне математичне поле, яке має завдання пояснити вплив дуже малих факторів. Хаотична система - це середовище, в якому чергуються передбачуваність та випадковість, і це найближче

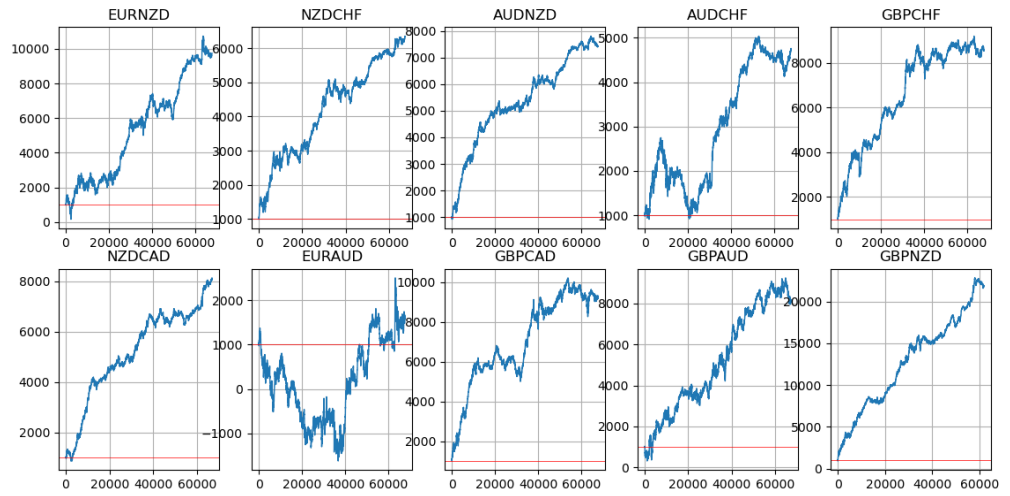
пояснення, яке ми маємо на сьогодні для фінансових ринків. Ефективна ринкова гіпотеза не може докладно пояснити динаміку ринку, і зараз краще дотримуватися реальних результатів торгів та історичних показників, оцінюючи, чи передбачувані ринки. У теорії Хаосу припущення щодо фінансових ринків передбачає, що ціна - це останнє, що змінюється, і що поточна ціна є найважливішою інформацією.

Британський гідролог Гарольд Едвін Херст представив міру мінливості часових рядів протягом аналізованого періоду часу. Цей показник називається Rescaled Range Analysis, який є основою нашого фрактального показника.

$$RS(n) = \frac{1}{S_n}[Max\sum_{j=1}^{k}(X_j - \overline{X_n}) - Min\sum_{j=1}^{k}(X_j - \overline{X_n})]$$

Формула Rescaled Range дуже цікава, оскільки вона враховує волатильність (S), середнє значення (X-bar) і діапазон даних для аналізу їх властивостей. Наведена вище формула свідчить, що ми повинні розрахувати діапазон між міні діапазонами максимального і мінімального значень, а потім розділити їх на стандартне відхилення, яке в такому випадку є непрямим показником волатильності.

Стратегія: покупка, коли індикатор фракталу досягає межі 1.00, коли ринкова ціна має тенденцію до зниження; продаж, коли індикатор фракталу досягає межі 1.00, коли ринкова ціна рухається вгору.




## 2.8 The Ultimate Oscillator

Творець індикатору мав на меті виявити розтягнуті рухи імпульсу в різні періоди огляду. Це, в свою чергу, зменшує надмірну волатильність, спричинену меншими періодами огляду. Щоб побудувати Ultimate Oscillator, треба виконати такі кроки:

- Вибрати періоди ретроспективного огляду трьох ковзних середніх: 5, 13 і 21.
- Розрахувати купівельний тиск для кожного періоду.

$$\text{Тиск} = \text{ціна закриття} - \text{m}in(\text{мінімум, минуле закриття})$$

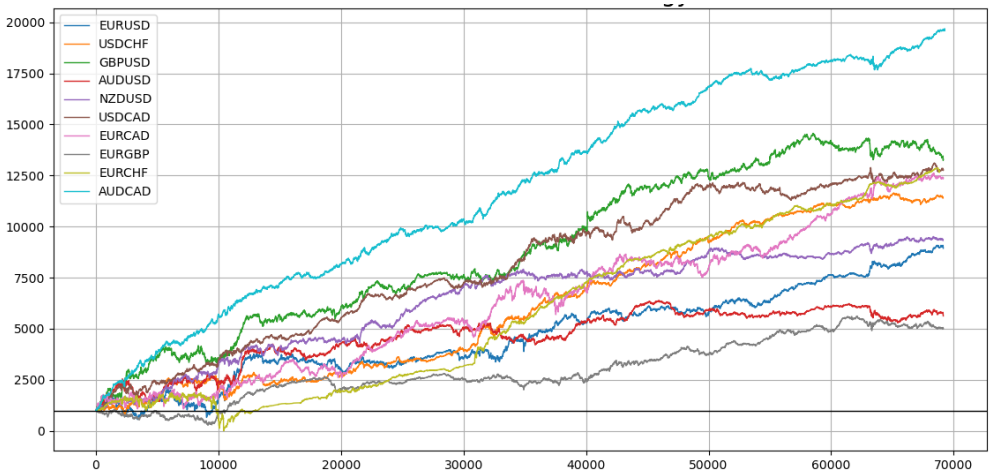
- Обчисліть дійсний діапазон для кожного періоду.

$$\begin{aligned}\text{Дійсний діапазон} \\ &= \max(\text{вища ціна, минуле закриття}) \\ &- \min\ (\text{нижча ціна, минуле закриття})\end{aligned}$$

- Розділіть купівельний тиск на дійсний діапазон для кожного періоду.
- Обчисліть експоненціальну ковзну середню серед попереднього результату тричі, використовуючи три різні періоди.
- Обчисліть середньозважене з трьох експоненціальних ковзних середніх і помножте на 100.

$$Oscilator = \left[\frac{\left((A_5 * 4) + (A_{13} * 2) + A_{21}\right)}{7}\right] * 100$$

Стратегія: покупка, коли $Oscilator$ досягає значення 40 з двома попередніми значеннями вище 40; продаж, коли $Oscilator$ досягає значення 60 з двома попередніми значеннями нижче 60.


EURUSD
USDCHF
GBPUSD
AUDUSD
NZDUSD
USDCAD
EURCAD
EURGBP
EURCHF
AUDCAD
20000
17500
15000
12500
10000
7500
5000
2500
0
0
10000
20000
30000
40000
50000
60000
70000

# 3. Машинне навчання

## 3.1 Логістична регресія

Логістична Регресія – є лінійною моделлю класифікації. Використовується щоб моделювати ймовірність певного класу події, може бути розширена для моделювання кількох класів подій. Вона призначена для наборів даних, що мають числові вхідні змінні, та категоріальну цільову змінну, яка має два значення або класи. Проблеми цього типу називаються проблемами бінарної класифікації. Логістична регресія призначена для двокласних задач, моделюючи ціль за допомогою біноміальної функції розподілу ймовірностей. Мітки класів відображаються на 1 для позитивного класу чи результату та 0 для негативного класу чи результату. Модель придатності передбачає ймовірність того, що приклад належить до класу 1.

Імовірності, що описують можливі результати одного випробовування, у цій моделі моделюються через логістичну функцію. Логістична регресія вимірює взаємозв'язок між категоріальною залежною змінною та однією чи більше незалежними змінними через оцінку ймовірностей за допомогою логістичної функції, яка є кумулятивною фунцією розподілу логістичного розпоподілу.

Логістична функція являє собою сигмововидну функцію, яка отримує на вхід будь-яке дійсне t, і видає значення, що дорівнюють 0 чи 1 та має рівняння:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

$x, x_0$ – значення середньої точки сигмоподібної

L – максимальне значення кривої

k – швидкість логістичного зростання

Ця функція була розроблена П'єром Франсуа Верхульстом як модель для передбачення зростання населення через коригування моделі експоненціального зростання. Початкове зростання функції є приблизно

геометричним; після певного насичення ріст сповільнюється до арифметичного; в завершенні ріст зупиняється.

Стандартна логістична функція:

$$f(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$$

Логістична регресія – важливий алгоритм машинного навчання, ціллю якого є промоделювати ймовірнітсть того, що випадкова величина Y буде дорівнювати 0 чи 1, з врахуванням експериментальних даних.

Узагальнена функція лінійної моделі, параметризовану Θ:

$h_\Theta(\mathrm{X}) = \frac{1}{1+e^{-\Theta^T X}} = \Pr(Y = 1|X; \Theta)$, з чого $\Pr(Y = 0|X; \Theta) = 1 - h_\Theta(X)$

$\Pr(y|X; \Theta) = h_\Theta(X)^y (1 - h_\Theta(X))^{(1-y)}$

Обчислення функції правдоподібності з припущенням, що все спостереження незалежно розподілені по Бернуллі:

$$L(\Theta|\mathrm{y}; \mathrm{x}) = \Pr(Y|X; \Theta) = \prod_i \Pr(y_i|x_i; \Theta) = \prod_i h_\Theta(x_i)^{y_i}(1 - h_\Theta(x_i))^{(1-y_i)}$$

Зазвичай максимізується логарифмічна правдоподібність:

$N^{-1} \log L(\theta|y; x) = N^{-1} \sum_{i=1}^{N} logPr(y_i|x_i; \Theta)$, максимізується завдяки оптимізації градієнтним спуском.

Як задача оптимізації, двійковий клас $l_2$ штрафна логістична регресія мінімізує наступну функцію витрат:

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log\left(\left(\exp\left(-y_i\left(X_i^T w + c\right)\right) + 1\right)\right)$$

Так само, $l_1$ регуляризована логістична регресія вирішує таку задачу оптимізації:

$$\min_{w,c} ||\, w||_1 + C \sum_{i=1}^{n} \log\left(\exp\left(-y_i\left(X_i^T w + c\right)\right) + 1\right)$$

Еластична мережева регуляризація - це поєднання $l_1$ і $l_2$, і мінімізація наступну функцію витрат:

$$\min_{w,c} \frac{1-\rho}{2} w^T w + \rho|w|_1 + C \sum_{i=1}^{n} \log\left(\exp\left(-y_i\left(X_i^T w + c\right)\right) + 1\right),$$

де $\rho$ контролює силу регулярізації $l_1$ з регулярізацією $l_2$.

Основна ідея логічної регрессії у використанні механізму, вже розробленого для лінійної регресії, через моделювання ймовірності $\rho_i$ за допомогою функції лінійного предиктора. Лінійний предиктор являє собою лінійну комбінацію пояснювальних змінних та коефіцієнтів регресії. Функція предиктора для певної точки має вигляд:

$f(i) = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_1 x_{m,i}$,

де $\beta_0, \ldots, \beta_m$ – коефіцієнти регресії, які вказують на відносний вплив певної пояснювальної змінної на результат. Їх групують в єдиний вектор β розміром m+1. Для кожної точки даних *i* додається додаткова пояснювальна псевдо змінна $x_{0,i}$ з фіксованим значенням 1, що відповідає коефіцієнту перехоплення $\beta_0$. Отримані пояснювальні змінні групуються в єдиний вектор X розміром m+1.

Таким чином в кінці функція лінійного предиктора має вигляд:

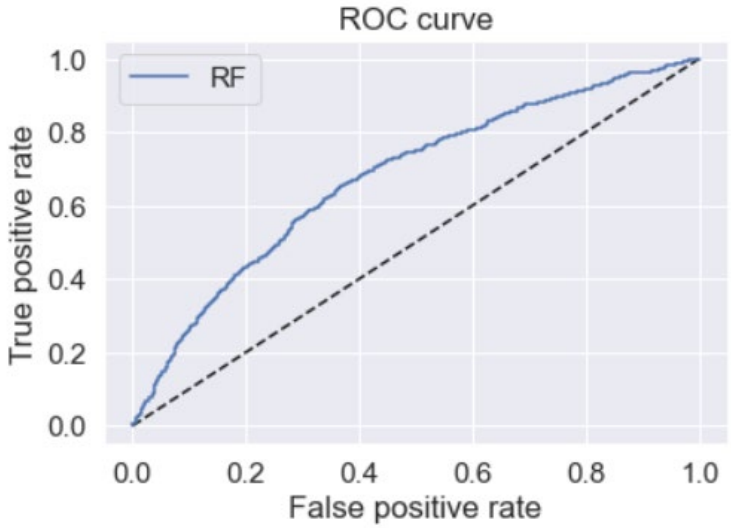
$$f(i) = \beta X_i$$

Для роботи з цим алгоритмом використаю історичні індексу S&P 500, кошик якого скрадається з 500 акціонерних компаній США і мають найбільшу капіталізацію. Для прогнозування візьму ціни індексу, приріст ціни з попереднього дня у відсотках,відставання від прогнозованої алгоритмом цілі за періоди від одного до п’яти днів, розраховані полоси болінджера, які представляють діапазон, в якому рухається ціна, додам індикатор MACD, моментум, волатильність та відстань ціни від середнього значення за 50 попередніх днів.

| Date | SPY | return | direction | lag_1 | lag_2 | lag_3 | lag_4 | lag_5 | bb_bbm | bb_bbh | bb_bbl | macd | macd_diff | macd_signal | momentum | volatility | distance |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1993-04-28** | 25.998814 | -0.002139 | 0 | 0.010742 | -0.007889 | -0.004276 | -0.012721 | -0.000703 | 26.386663 | 26.984872 | 25.788454 | -0.134615 | -0.075668 | -0.058947 | -0.002969 | 0.007520 | -0.393977 |
| **1993-04-29** | 26.110176 | 0.004274 | 1 | -0.002139 | 0.010742 | -0.007889 | -0.004276 | -0.012721 | 26.350477 | 26.923103 | 25.777851 | -0.133862 | -0.059932 | -0.073930 | -0.003257 | 0.007458 | -0.455796 |
| **1993-04-30** | 26.147274 | 0.001420 | 1 | 0.004274 | -0.002139 | 0.010742 | -0.007889 | -0.004276 | 26.320786 | 26.870436 | 25.771136 | -0.128788 | -0.043886 | -0.084902 | 0.000142 | 0.007572 | -0.353168 |
| **1993-05-03** | 26.314287 | 0.006367 | 1 | 0.001420 | 0.004274 | -0.002139 | 0.010742 | -0.007889 | 26.327280 | 26.873389 | 25.781172 | -0.110022 | -0.020096 | -0.089926 | 0.001282 | 0.007580 | -0.325917 |
| **1993-05-04** | 26.407076 | 0.003520 | 1 | 0.006367 | 0.001420 | 0.004274 | -0.002139 | 0.010742 | 26.331920 | 26.879083 | 25.784756 | -0.086663 | 0.002610 | -0.089273 | 0.004133 | 0.006128 | -0.170243 |

Розділю данні для навчання на тестовий і навчальний набори в пропорції 70:30. Зворотна сила регуляризації = 1e7, застосовує штраф до збільшення величини значень параметрів, щоб зменшити ступінь перенавчання. Задам параметр 1000 для максимальної кількісті ітерацій, необхідних для зближення вирішувачів. З такими параметрами отримуємо точність близько 64%.

Отримуємо криву ROC, яка ілюструє діагностичну спроможність в бінарному класифікаторі системи, відображає співвідношення між часткою об'єктів від загальної кількості носіїв ознаки, вірно класифікованих як несучі ознака (TPR), і часток об'єктів від загальної кількості об'єктів, що не несуть ознаки, помилково класифікованих як несучі ознака (FPR) при варіюванні порога вирішального правила.




Залишилось проаналізувати ефективність моделі, для чого візьмемо такі параметри: коефіцієнт Шарпа, який є відношенням прибутковості портфеля до стандартного відхилення, максимальну просадку, Коефіцієнт Кальмара, який є співвідношенням між річною прибутковістю та максимальною просадкою, Стабільність, визначає R-квадрат лінійні відповідності кумулятивної логарифмічної прибутковості, Коефіцієнт Сортіно, який схожий на коефіцієнт Шарпа, але враховує лише нестабільність у зворотному напрямку, кефіцієнт Омега, який визначається як коефіцієнт прибутку та збитків, зважений за ймовірністю, для певної

межі повернення порогу, додамо формулу розподілу прибутку, куртоз, який вказує на піковість віддачі, співвідношення хвостів.

| | |
|---|---|
| **Start date** | 2013-01-02 |
| **End date** | 2021-03-02 |
| **Total months** | 97 |
| | **Backtest** |
| **Annual return** | 56.153% |
| **Cumulative returns** | 3687.506% |
| **Annual volatility** | 10.419% |
| **Sharpe ratio** | 4.33 |
| **Calmar ratio** | 7.97 |
| **Stability** | 0.99 |
| **Max drawdown** | -7.043% |
| **Omega ratio** | 3.07 |
| **Sortino ratio** | 9.51 |
| **Skew** | 2.36 |
| **Kurtosis** | 22.70 |
| **Tail ratio** | 2.25 |
| **Daily value at risk** | -1.134% |

### 3.2 Випадкові ліси

Алгоритм випадкового лісу є частиною широкого типу алгоритмів, які називаються ансамблевими методами. Кожне дерево в ансамблі будується на основі вибірки, взятої із заміною з навчального набору даних. Багато дерев рішень утворюють випадковий ліс, який є різновидом ансамблевих методів. Дерево рішень - це керований алгоритм навчання, який одночасно підходить для задач класифікації та регресії завдяки своїй лінійній та нелінійній придатності.

Дерева рішень розбивають набір даних на більш дрібні підмножини, використовуючи умови, щоб в результаті отримати оцінку ймовірності. Перше дерево називається кореневим вузлом, а наступні, що виходять з нього, називаються вузлами прийняття рішень. Алгоритм, який використовується для перебору дерев з метою обчислення ентропії, називається Iterative Dichotomiser 3 і був винайдений Россом Куінланом. Ідея полягає в тому, що при усередненні безлічі моделей ви отримаєте більш гладкий і точний прогноз. Саме тому алгоритми випадкового лісу добре підходять для вирішення завдань як класифікації, так і регресії.

Алгоритм випадкового лісу - це просто набір дерев рішень для зменшення надлишкової підгонки і усереднення результатів, що дає нам імовірно більш високу точність.

Ліси випадкових рішень коригують звичку дерев рішень в вигляді перенавчання (аналіз, який занадто точно відповідає певному набору даних, через що може не вдатись приєднати додаткові данні або надійно передбачити майбутні спостереження) їх навчальну вибірку. Випадкові ліси, як правило, перевершують дерева рішень, але їх точність нижча, ніж підсилені градієнтами дерева. Однак характеристики даних можуть впливати на їх ефективність.

Випадкові ліси - це спосіб усереднення декількох дерев глибоких рішень, що навчаються на різних частинах одного навчального набору, з метою зменшення дисперсії. Це відбувається за рахунок невеликого збільшення упередженості та певної втрати інтерпретації, але загалом значно підвищує ефективність в кінцевій моделі. Алгоритм навчання для випадкових лісів застосовує загальну техніку бутстреп-агрегації для навчання дерев. З огляду на навчальну множину $X = x_1,..., x_n$ з відповідями $Y = y_1,..., y_n$, бутстреп-агрегація багаторазово (B разів) вибирає випадкову вибірку з заміною з навчальної множини і підганяє дерева до цих вибіркок. Після навчання, прогнози для невидимих зразків $x'$ можуть бути зроблені шляхом усереднення прогнозів від всіх окремих дерев регресії на $x'$:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b\,(x')$$

Ця процедура бутстраппінга призводить до поліпшення роботи моделі, оскільки вона зменшує дисперсію моделі, не збільшуючи зміщення. Це означає, що якщо передбачення одного дерева дуже чутливі до шуму в навчальному наборі, то середнє значення багатьох дерев - ні, до тих пір, поки дерева не корелюють між собою. Просте навчання багатьох дерев на одному навчальному наборі дасть сильно корельовані дерева (або навіть

одине і те же дерево багато разів, якщо алгоритм навчання детермінований); бутстреп-вибірка - це спосіб декорреліровать дерева, показуючи їм різні навчальні набори.

Крім того, оцінка невизначеності прогнозу може бути зроблена як стандартне відхилення прогнозів всіх окремих дерев регресії на $x'$:

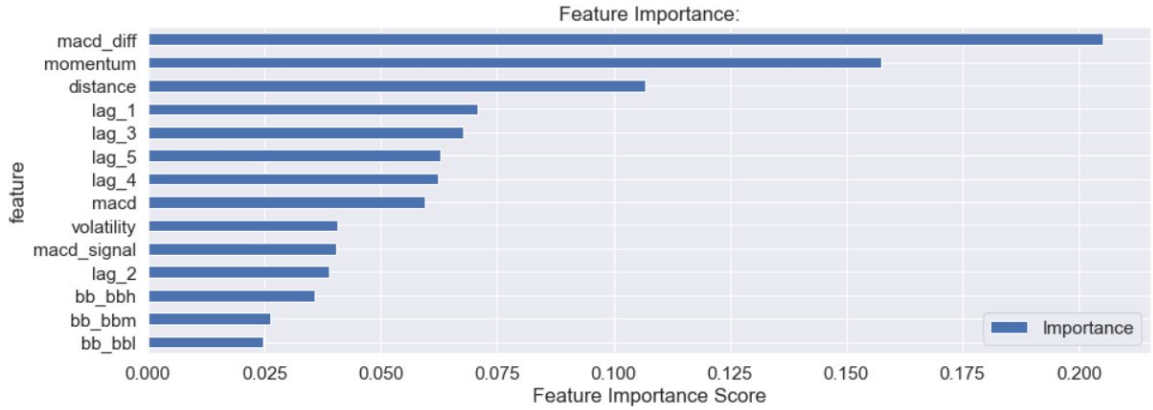
$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}}$$

Кількість зразків (B) – вільний параметр. Оптимальна кількість дерев B можна знайти за допомогою перехресної валідації або спостерігаючи за помилкою поза мішка: середня помилка передбачення на кожній навчальній вибірці $x_i$, використовуючи тільки ті дерева, які не мали $x_i$ в своїй бутстреп-вибіркою. Помилки навчання і тестування, як правило, вирівнюються після підбору деякої кількості дерев.
Перший крок у вимірюванні важливості змінних в наборі даних $D_n\{(X_i, X_i)\}_{i=1}^{n}$ полягає в підгонці випадкового лісу до даних. В процесі підгонки реєструється out-of-bag помилка для кожної точки даних і усереднюється по лісі (помилки на незалежному тестовому наборі можуть бути замінені, якщо бутстреп-агрегація не використовується під час навчання). Щоб виміряти важливість j-тої ознаки після навчання, значення j-ї ознаки переставляються серед навчальних даних, і помилка out-of-bag знову обчислюється на цьому збуреному наборі даних. Оцінка важливості для j-ї ознаки обчислюється шляхом усереднення різниці в помилці out-of-bag до і після перестановки по всім деревам. Оцінка нормується стандартним відхиленням цих різниць. Ознаки, які дають великі значення цього показника, ранжуються як більш важливі, ніж ознаки, які дають малі значення.

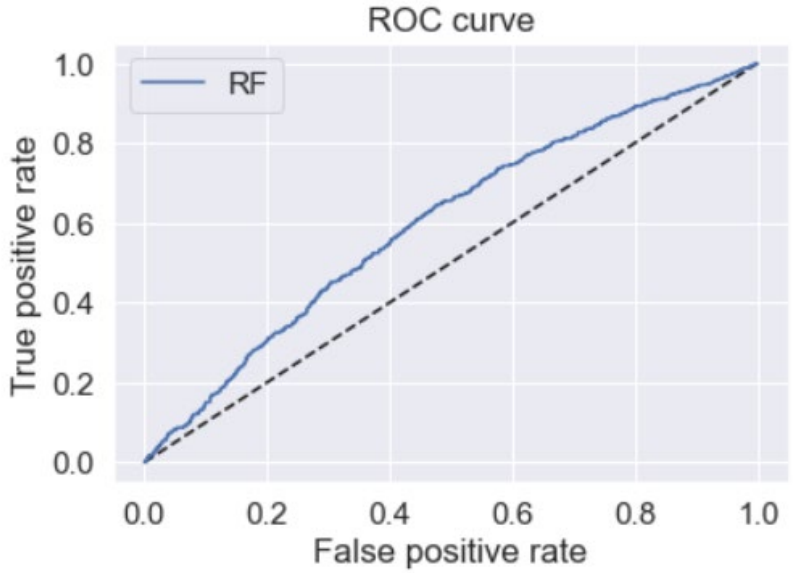
Для роботи з цим алгоритмом візьмему ті ж самі дані і індикатори, застосуємо то же принцип поділення даних на тренувальні і тестові.

Використаю метод GridSearchCV для оптимізації гіперпараметрів. При використанні цих параметрів точність близько 59%.

Для алгоритму випадкового лісу є можливість оцінити значимість вхідних даних для прогнозування:




Крива ROC




Фактичні показники стратегії:

| | |
|---|---|
| **Start date** | 2013-01-02 |
| **End date** | 2021-03-02 |
| **Total months** | 97 |
| | **Backtest** |
| **Annual return** | 29.324% |
| **Cumulative returns** | 714.184% |
| **Annual volatility** | 12.577% |
| **Sharpe ratio** | 2.11 |
| **Calmar ratio** | 1.85 |
| **Stability** | 0.99 |
| **Max drawdown** | -15.841% |
| **Omega ratio** | 1.74 |
| **Sortino ratio** | 3.08 |
| **Skew** | -1.51 |
| **Kurtosis** | 42.99 |
| **Tail ratio** | 1.55 |
| **Daily value at risk** | -1.479% |

## ВИСНОВКИ

Під час проведеної роботи були представлені різні технічні індикатори і математика на якій вони базуються, були створені торгові алгоритми і протестовані на історичних даних, був створений алгоритм машинного навчання на основі певних індикаторів. Завдяки дотриманню правил роботи з ризиком і алгоритмам результат такого трейдингу на історичних даних був додатнім.

В майбутньому подібні роботи будуть лише набирати потенціал, особливо з подальшим розвиненням машинного навчання і технологій. Кожен рік кількість торгових операцій, проведених торговельними роботами, зростає і вже на певних ринках складає більше половини від всіх. Людина ніколи не буде цілком посунута з трейдингу, проте такі торгові алгоритми не допускають більшість помилок, притаманних людині.

Використання таких роботів великими інвесторами та фондами зумовлено дивесифікацією ризиків.

## СПИСОК ЛІТЕРАТУРИ

1. Kaabar Sofien New technical indicators, 1st ed, 2020

2. Kaabar Sofien New book of back-tests: trading objectively, 1st ed, 2020

3. Kaabar Sofien The handbook of exotic trading strategies: Uncommon techniques to diversify your prediction methods, 2st ed, 2020

4. Wolfgang Paul, Jorg Baschnagel Stochastic Process from physics to finance, 2st ed, 2020

5. Kevin J. Davey Building Winning Algorithmic Trading Systems, 2st ed, 2014

6. Hands-On Machine Learning with Scikit-Learn & Tensorflow, 1st ed, A. Geron, 2018

7. Perry J. Kaufman A guide co creating a successful algorithmic trading strategy, 2st ed, 2016

8. Gwilym M. Jenkins Time Series Analysis: Forecasting and Control, 4th ed, 1970

9. T. Hastie, Robert Tibshirani The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2st ed, 2009

10 Ruey S. Tsay Analysyis of financial time series, 3th ed, 2010