МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра Математического и компьютерного моделирования

ОТЧЕТ ПО ПРОИЗВОДСТВЕННОЙ ПРАКТИКЕ «НАУЧНО-ИССЛЕДОВАТЕЛЬСКАЯ РАБОТА»

студентки 4 курса 441 группы

направления 09.03.03 — Прикладная информатика

механико-математического факультета

Ткачевой Арины Алексеевны

Место прохождения практики: Кафедра математического и компьютерного моделирования

Сроки прохождения практики: 06.02.23-02.05.23

Оценка

Руководитель практики от СГУ
доцент, к.т.н., доцент И.А. Панкратов

Саратов 2023

Введение. Анализирование фондового рынка на сегодняшний день считается одним из перспективных направлений экономических исследований. Вложение капитала в ценные бумаги все чаще рассматривается инвесторами как наиболее привлекательная возможность получения прибыли.

Сегодня, когда машинное обучение широко применяется во многих сферах жизни, в том числе отраслях экономики, системы искусственного интеллекта ежедневно взаимодействуют с человеком и его социальным окружением. Например, умная станция «Алиса» переключает музыку по голосовой команде человека, биометрия данных, используемая в качестве формы управления идентификаторами доступа и контроля доступа, и т.д. Если говорить об экономике, нейронные сети помогают решить такие задачи, как:

1. задача прогнозирования;
2. задача апроксимации;
3. задача оптимизации.

В данной работе решается задача прогнозирования. Объекты исследования: фондовый рынок, методы машинного обучения. Предмет исследования: котировки акций 5 компаний, принцип работы (обучения) нейронных сетей и методов машинного обучения.

Цель работы: разработка и сравнение моделей машинного обучения для прогнозирования на фондовом рынке. Они указывают не конкретные цены, а их тендецию (движение вверх или вниз).

Задачи, которые будут решены в ходе работы:

1. сбор данных о котировках;
2. расчет 10 технических показателй;
3. оформление данных в датасет;
4. ознакомление с построением модели для сравнения точности предсказанных значений методом опорных векторов (SVM), случайный лес (RF) и искусственная нейронная сеть (ANN);
5. построение моделей с помощью данных методов;
6. сравнение моделей.

Актуальность данной работы состоит в том, чтобы продемонстрировать возможность прогнозирования тренда стоимости ценных бумаг на

фондовом рынке посредством машинного обучения, для помощи трейдерам(инвесторам).

В первом разделе рассматриваются теоретические сведения о фондовом рынке и машинном обучении.

Фондовые биржи – это объединение покупателей и продавцов различных ценных бумаг. На фондовых биржах торгуют различными ценными бумагами. В первую очередь – акциями. Акция – это свидетельство права собственности на корпорацию. Торговля акциями – сделка между двумя физическими лицами, которая не оказывает прямого влияния на корпорацию, чьи акции были первоначально выпущены. Голубые фишки – акции наиболее крупных и известных среди инвесторов компаний с большой капитализацией и высокой ликвидностью: их легко купить и продать по рыночной цене в больших объемах. Голубые фишки – не строгий термин, а скорее сленговое выражение для неформального обозначения наиболее крупных и известных компаний. Другой распространенный инструмент – облигации. Инвестор дает компании в долг денежные средства, а она обязуется вернуть деньги с процентами через определенное время. Паи, ETF, фьючерсы и опционы также можно купить на фондовом рынке.

Московская биржа – биржевая группа, управляющая единственной в России многофункциональной биржевой площадкой по торговле акциями, облигациями, производными инструментами, валютой, инструментами денежного рынка, драгоценными металлами и товарами. Образована в декабре 2011 года в результате слияния двух основных российских биржевых групп ММВБ и РТС. Основные акционеры: Центральный банк РФ, Сбербанк России, Внешэкономбанк, ЕБРР.

Московская биржа создала отдельный индекс голубых фишек – MOEXBC. Он состоит из 15 самых крупных и ликвидных компаний российского рынка: АЛРОС, Газпром, Норильский Никель, Лукойл, Магнит, Новатэк, ФосАгро, Полюс, Полиметалл, Роснефть, Сбербанк, Сургутнефтегаз, Татнефть, TCS Group, Яндекс.

Для полноты данных, акции некоторых компаний взяты из голубых фишек.

Фундаментальный анализ – это анализ состояния и динамики рынка, основанный на фундаментальных экономических показателях. Целью фундаментального анализа является оценка реальной (внутренней, справедливой) стоимости компании для определения справедливой стоимости ее акций. Задачей фундаментального анализа является выявление недооцененных или переоцененных компаний и, соответственно, бумаг.

Под техническим анализом понимается метод прогнозирования рыночной стоимости ценной бумаги на основе информации о ее котировках, основанный на математических вычислениях и статистических данных. Котировка – установление курса валюты, ценных бумаг, цен на товары на биржах в соответствии с действующим законодательством и сложившейся ситуацией на рынке.

Машинное обучение – это метод анализа данных, который автоматически строит аналитическую модель для определенной задачи, учитывая набор данных, описывающих эту задачу. Алгоритмы машинного обучения могут выполнять задачи как регрессии, так и классификации. Отдельных алгоритмов много, и зачастую они сильно отличаются друг от друга. Однако все алгоритмы машинного обучения используют идею «обучающих данных» для автоматического создания аналитической модели для требуемой задачи. После обучения алгоритмам предоставляются новые образцы данных для выполнения задачи (например, регрессия или классификация). Алгоритмы и составные части, используемые в данной работе, описаны ниже.

Во втором разделе собирается и оформляется датасет для прогнозирования.

Данные, которые использовались для формирования датасета, представляют из себя набор данных о котировках акций 5 выбранных компаний: Сбербанк, Лукойл, Магнит, Полюс и Яндекс. Данные собраны с 01.01.22 по 01.04.23 с сайта Финам.

Набор данных содержит 3612 строк и 11 столбцов.

Данные DATE, TIME, OPEN, HIGH, LOW, CLOSE и VOL представляют из себя:

- DATE - дата;
- TIME - время (временной шаг был выбран 1 час);

- OPEN - цена открытия;
- HIGH - самая высокая цена до закрытия;
- LOW - самая низкая цена до закрытия;
- CLOSE - цена закрытия;
- VOL - количество сделок за период.

Произведена обработка данных: были рассчитаны 10 технических показателей в качестве новых характеристик: SMA, WMA, Momentum, K% и D%, RSI, MACD, Williams'%R, StoOsc, A/D, CCI.

SMA 10 – простая скользящая средняя (Simple Moving Average). Основная задача индикатора — выделить из рыночного «шума» наиболее вероятный тренд, а также оценить его силу. Линия скользящей средней при этом выступает как динамический уровень поддержки/сопротивления.

Рассчитывается как среднеарифметическое значение цены торгового актива за определенный период времени:

$$SMA = \sum_{i=1}^{n} P_i/n.$$

WMA 10 – взвешенная скользящая средняя (Weighted Moving Average). Это модификация простой скользящей средней со специальной подборкой весов таким образом, чтоб новые цены имели больший вес.

Формула:

$$WMA = \frac{\sum\limits_{i=1}^{n} P_i * W_i}{\sum\limits_{i=1}^{n} W_i}.$$

Здесь P_i – значение цены i-периодов назад, W_i — значение весов для цены i-периодов назад.

Momentum 14 – оценки величины изменения цены финансового инструмента за определенный промежуток времени. В первоначальном варианте формула индикатора имеет следующий вид:

$$Momentum = P_n(close) - P_{n-t}(close).$$

Если показания индикатора выше 100%, цена за t периодов растет, если ниже – падает. Здесь и далее этот период будем брать равным 14.

K% и D% – стохастический осциллятор. Это индикатор, показывающий положение текущей цены относительно диапазона цен за определенный период в прошлом. Измеряется в процентах.

Индикатор строится из двух линий:

$$K_t\% = \frac{C_t - L_n}{H_n - L_n} * 100$$

быстрый стохастик, где

C_t – цена закрытия текущего периода,

L_n – самая низкая цена за последние n периодов,

H_n – самая высокая цена за последние n периодов.

$D_t\%$ – медленный стохастик, является скользящей средней относительно $K_t\%$ с небольшим периодом усреднения. Могут использоваться различные механизмы усреднения (простая средняя, экспоненциальная, сглаженная, взвешенная).

RSI 14 – индекс относительной силы. Это отношение среднего прироста цены к среднему падению за период. Эта величина позволяет оценить, покупатели или продавцы сильнее влияли на цену в выбранном периоде и предположить дальнейшее развитие событий. Для расчета относительной силы выбираются все свечи выбранного промежутка времени, которые показали закрытие выше, чем предшествующая свеча, и определяется среднее значение прироста с помощью формулы экспоненциального скользящего средней. Аналогичная операция производится для свечей, показавших закрытие ниже предшествующей. Отношение этих двух величин и даст значение относительной силы (RS). Формула выглядит следующим образом:

$$RS = EMA_n(Up)/EMA_n(Down).$$

MACD – индикатор схождения-расхождения скользящих средних. Для расчета используются три экспоненциальные скользящие средние с разными периодами. Из быстрой скользящей средней с меньшим периодом (EMA_s)

вычитается медленная скользящая средняя с большим периодом (EMA_l). По полученным значениям строится линия MACD.

$$MACD = EMA_s(P) - EMA_l(P).$$

Williams'%R – удобный технический индикатор, помогающий определять состояния перекупленности или перепроданности рынка, а также дающий сигналы дивергенции, свидетельствующие о развороте рынка.

Математически Williams'%R выражает отношение между разницей максимума с ценой закрытия и диапазоном «максимум-минимум» за определенный период, умноженное на -100. Оптимальным считается период 14, однако этот параметр можно настраивать в зависимости от инструмента.

$$\%R = (max(High_n) - ClosingPrice)/(max(High_n) - min(Lown)) * (-100).$$

Таким образом, значения индикатора располагаются в отрицательной зоне, в диапазоне от 0 до -100. Зоной перекупленности считаются значения от -20 до 0, а перепроданности — значения от -100 до -80.

StoOsc – стохастический осциллятор, предназначен для определения импульса цены.

Математически стохастик выражает отношение между ценой закрытия и диапазоном «максимум-минимум» за определенный период в виде процентной величины от 0 до 100. Значение стохастического осциллятора равное 80 и выше показывает, что цена закрытия находится вблизи верхней границы диапазона; стохастик равный 20 и ниже означает, что цена закрытия находится вблизи нижней границы диапазона. Соответственно, если на рынке прослеживается тенденция к закрытию в верхней части дневного диапазона, то он – бычий, если в нижней, то он — медвежий.

Формула:

$$StoOsc = 100 * \frac{C_0 - min(L_n)}{max(H_n) - min(L_n)},$$

где $max(H_n)$ – максимум за N периодов, $min(L_n)$ – минимум за N периодов, C_0 – цена закрытия текущего периода.

A/D – индикатор накопления/распределения (accumulation/distribution), в основе которого лежит сравнение цены закрытия с серединой диапазона от минимума до максимума.

Расчет A/D начинается с вычисления разницы цены закрытия и дневного минимума, от которой отнимается разница максимума и цены закрытия. Полученный показатель делится на разницу максимума и минимума, после чего умножается на объем торгов. На завершающем этапе результат суммируется со значением A/D на предыдущей свече. Таким образом, каждое следующее значение индикатора показывает накопленный результат вычислений всех предыдущих периодов.

$$AD_m = \left[\frac{(Close - Low) - (High - Close)}{High - Low} * Volume \right] + AD_{n-1}$$

CCI – индекс товарного канала. Это осциллирующий индикатор, который помогает определять, когда актив является перекупленным или перепроданным. Индекс канала товара помогает заметить ослабление или конец тренда и изменение его направления. Данный индикатор помогает определять пики и спады в цене актива, и может служить сигналом об ослаблении или конце тренда и изменении его направления.

Формула расчета:

$$CCI_n = (TP - SMA_n * TP)/(0.015 * MD_n),$$

где

TP – типичная цена: сумма максимума, минимума и цены закрытия бара, деленная на 3:

$TP = (High + Low + Close)/3$,

$SMAnTP$ – простое скользящее среднее (Simple Moving Average) типичной цены за расчетный период,

MD_n – среднее отклонение (Mean Deviation) за расчетный период:

$MDn = (|TP1 - SMAnTP1| + \cdots + |TPn - SMAnTP_n|)/n$

Данные индикаторы станут основными входными данными для нейронной сети.

Третий раздел посвящен построению моделей машинного обучения.

Метод опорных векторов (Support Vector Machine — SVM) – это очень мощная и универсальная модель машинного обучения, способная выполнять линейную или нелинейную классификацию, регрессию и даже выявление выбросов. Она является одной из самых популярных моделей в машинном обучении.

Основная идея метода – перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с наибольшим зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. Разделяющей гиперплоскостью будет гиперплоскость, создающая наибольшее расстояние до двух параллельных гиперплоскостей. Алгоритм основан на допущении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

Случайный лес – это набор деревьев решений, где каждое дерево немного отличается от остальных. Идея случайного леса заключается в том, что каждое дерево может довольно хорошо прогнозировать, но скорее всего переобучается на части данных. Если построить много деревьев, которые хорошо работают и переобучаются с разной степенью, можно уменьшить переобучение путем усреднения их результатов.

Для реализации вышеизложенной стратегии необходимо построить большое количество деревьев решений. Каждое дерево должно на приемлемом уровне прогнозировать целевую переменную и должно отличаться от других деревьев. Случайные леса получили свое название из-за того, что в процесс построения деревьев была внесена случайность, призванная обеспечить уникальность каждого дерева.

Благодаря своей гибкости, Random Forest применяется для решения практически любых задач в области машинного обучения. Сюда относятся классификации (RandomForestClassifier) и регрессии (RandomForestRegressor), а также более сложные задачи, вроде отбора признаков, поиска выбросов/аномалий и кластеризации.

Нейронные сети – это направление машинного обучения, в котором аналитическая модель строится с использованием многослойной архитектуры,

где каждый слой преобразует входные данные и выдает одно или несколько выходных значений. Каждый слой состоит из определенного количества узлов (нейронов). Каждый нейрон принимает входные данные и выполняет линейное преобразование, затем применяет нелинейную функцию к преобразованному значению и выдает результат. Эта процедура выполняется на входе для каждого нейрона в слое, но с разными весами для каждого нейрона. После этого выходные данные могут быть использованы для классификации или регрессии, либо переданы в другой слой. Процесс настройки весов всей нейронов таким образом, чтобы они давали желаемый выход, является фазой обучения искусственных нейронных сетей.

Сначала к созданному датасету применен метод опорных векторов с полиномиальной гиперплоскостью. Данный способ построения показал наиболее точные результаты в сравнении с другими возможными конфигурациями:

```
from sklearn.svm import SVC
svclassifier = SVC(kernel='poly')
svclassifier.fit(X_train, y_train)
```

После обучения модель показывает точность 85,18%.

Далее применен алгоритм случайный лес со следующими конфигурациями:

```
from sklearn.ensemble import RandomForestClassifier
model = RandomForestClassifier(criterion='log_loss',
                               max_depth=5)
```

Случайный лес показал точность 85,33%.

Площадь под ROC-кривой AUC (Area Under Curve) является агрегированной характеристикой качества классификации, не зависящей от соотношения цен ошибок. Чем больше значение AUC, тем «лучше» модель классификации. Данный показатель часто используется для сравнительного анализа нескольких моделей классификации. В данном случае сравнивается работа деревьев решений:

```
roc_value = roc_auc_score(y_test, rf_probs)
print(roc_value)
```

Далее спроектированы 3 нейронные сети ANN.

Архитектура первой нейронной сети: 10 входных нейронов, 1 слой с 20 скрытыми и 1 выходной нейрон. Функция активация - сигмоидная:

```
import tensorflow as tf
import tensorflow
from tensorflow import keras
ann1 = tf.keras.models.Sequential()
ann1.add(tf.keras.layers.Dense(units=10, activation='relu',
        input_shape=X_train[0].shape))
ann1.add(tf.keras.layers.Dense(units=1, activation='sigmoid'))
```

Архитектура второй нейронной сети: 10 входных нейронов, 2 скрытых слоя с 10 и 5 нейронами и функциями активацией relu и 1 выходной нейрон с сигмоидной функцией активации:

```
import tensorflow as tf
import tensorflow
from tensorflow import keras
ann2 = tf.keras.models.Sequential()
ann2.add(tf.keras.layers.Dense(units=10, activation='relu',
        input_shape=X_train[0].shape))
ann2.add(tf.keras.layers.Dense(units=5, activation='relu'))
ann2.add(tf.keras.layers.Dense(units=1, activation='sigmoid')))
```

Архитектура третьей нейронной сети: 10 входных нейронов, 3 скрытых слоя, два из которых с 10 нейронами и один с 5 нейронами, функции активации relu и 1 выходной нейрон с сигмоидной функцией активации:

```
import tensorflow as tf
import tensorflow
from tensorflow import keras
ann3 = tf.keras.models.Sequential()
ann3.add(tf.keras.layers.Dense(units=10, activation='relu',
        input_shape=X_train[0].shape))
ann3.add(tf.keras.layers.Dense(units=5, activation='relu'))
```

```
ann3.add(tf.keras.layers.Dense(units=5, activation='relu'))
ann3.add(tf.keras.layers.Dense(units=1, activation='sigmoid'))
```

Первая нейронная сеть показала наилучший результат прогнозирования на тестовых данных. Точность данной модели составляет 84,23%.

В приложении представлены исходные программные коды реализации.

Заключение. При повышении точности моделей прогнозирования в будущем, предлагаемая система может быть развернута в режиме реального времени для прогнозирования тренда акций, что делает инвестиции более прибыльными и безопасными. Повышение точности с помощью этого подхода, основанного на распространенных методах инвестирования в акции, также продвигает идею предварительной обработки данных на основе предметной области, в которых алгоритмы машинного обучения используются. Эта идея может быть расширена не только в области акций за счет включения других человеческих подходов к инвестированию, но и в различных других областях, где используются экспертные системы и методы машинного обучения.

В данной работе для построения базы знаний используются десять технических показателей, однако другие макроэкономические переменные, такие как курсы обмена валют, инфляция, государственная политика, процентные ставки и т.д., которые влияют на фондовый рынок, также могут быть использованы в качестве исходных данных для моделей или при построении базы знаний эксперта система. Средний объем акций также является потенциальным кандидатом, который может быть полезен при определении тренда.

Архитектура и правила обучения нейронных сетей, как правило, носят ситуационный характер - разрабатываются для решения определенных (или группы родственных) задач, поэтому важно понимание принципов и особенностей их работы, что позволит профессионально подходить к вопросу выбора готовых или создания новых алгоритмов обучения.