«Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет «Высшая школа экономики»»

НИУ ВШЭ – Нижний Новгород

Факультет математики, информатики и компьютерных наук

Интеллектуальный анализ данных

КУРСОВАЯ РАБОТА

Исследование особенностей поведения цен рыночных акций

(Research of the Specifics of the Market Stock Evolution)

Поляков В.И.

Руководитель

Шилов Андрей Сергеевич

Нижний Новгород, 2023

Оглавление

Глава 1

Введение

Имеется рынок ценных бумаг, на котором представлено некоторое количество акций. Торговля акциями происходит практически каждый рабочий день, при этом цена каждой акции многократно меняется в течение дня после каждой совершенной сделки купли-продажи. Необходимо исследовать поведение цены акций с целью получения максимального дохода на данном рынке.

Глава 2

Классические алгоритмы

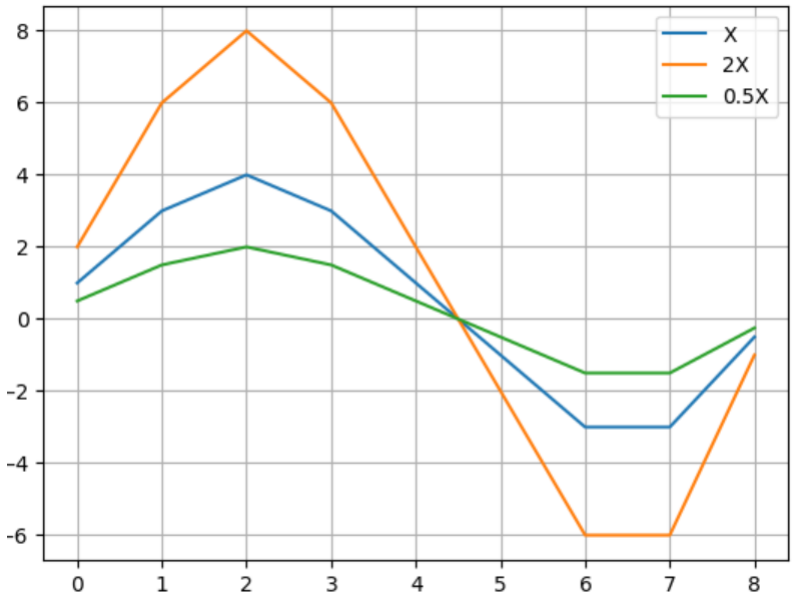
2.1 Исследовательский подход

Тип моделирования Среди представленных открытых исследований предсказание цен акций часто рассматривается как задача регрессивного предсказания цены на следующий день или на некоторый следующий период. В данной работе будет рассмотрены достоинства и недостатки несколько иного подхода. Для построения стратегии торговли на рынке акций в целом не очень важна конкретная предсказанная цена. Куда важнее, какой характер носит изменение цены. Поэтому представленные здесь модели будут построены на классификации поведения акций на следующий период. Дополнительным достоинством такого подхода является возможность

использования алгоритмов классификации и потенциальное упрощение процесса обучения для моделей.

Базовая идея Рынок акций до сих пор является достаточно сложной областью для методов анализа данных. Одной и причин этого является сложность в установлении точной зависимости между ценой акций и её характеристиками. Исследование в данной работе, в частности, базируется на гипотезе о существовании некоторых общих шаблонов в поведении цены акции, которые позволяют предсказать её поведение в следущий период. Таким образом, для используемых нейросетевых моделей поиск данных зависимостей будет положен основной стратегией предсказания.

При этом стоит внести некоторое уточнение про шаблоны. Под шаблоном понимается некоторое характерное поведение цены акции, то есть то, которое встречается достаточно часто в рамках одной или нескольких акций. При этом, так как цена каждого конкретного представителя рынка акций может на порядок отличаться, то вполне естественно положить корреляцию Пирсона как основную меру схожести и шаблонов, и акций. Таким образом, любой из рядов ниже является представителем одного и того же шаблона.



Базовые акции Для сравнения моделей необходимо зафиксировать небольшой набор акций, на котором будет происходить сравнение показателей моделей. Данный набор должен соответствовать ряду свойств: он должен быть достаточно небольшим, чтобы результаты были легко интерпретируемыми, и должен содержать акции, которые наиболее полно отражают рынок. Для данных целей можно использовать акции из пересечения максимальных клик в построенной по рынку графовой модели [2].

Для данного исследования был использован состав индекса NASDAQ на 24 мая 2023 года. В таблице ниже приведены результаты по количеству и пересечению максимальных клик для графовой модели,

построенной при разных значениях корреляционного порога.

	Threshold	Max clique count	Clique intersection
0	0.50	4	[RIVN, LCID, WBA, CSCO, MU, AMZN, ILMN, ADBE, TXN, MRVL, ISRG, IDXX, CHTR, EBAY, TEAM, ANSS, ASML, ALGN, DDOG, META, GOOG, WDAY, INTC, LRCX, ADSK, NFLX, WBD, ZS, ABNB, NXPI, CMCSA, GOOGL, AMAT, INTU]
1	0.55	2	[RIVN, WBA, CSCO, CRWD, MU, AMZN, ILMN, ADBE, ISRG, IDXX, CHTR, EBAY, TEAM, ANSS, ASML, ALGN, META, GOOG, WDAY, INTC, LRCX, ADSK, NFLX, WBD, ZS, ABNB, JD, NXPI, CMCSA, GOOGL, AMAT, INTU]
2	0.60	3	[WBA, CSCO, AMZN, ILMN, ADBE, ISRG, IDXX, CHTR, EBAY, TEAM, ANSS, ASML, ALGN, META, GOOG, WDAY, INTC, LRCX, ADSK, NFLX, WBD, ZS, ABNB, NXPI, CMCSA, GOOGL, AMAT, INTU]
3	0.65	4	[CSCO, AMZN, ILMN, ADBE, PYPL, ISRG, IDXX, CHTR, EBAY, ANSS, ALGN, META, GOOG, WDAY, INTC, ADSK, NFLX, WBD, ZS, ABNB, NXPI, CMCSA, GOOGL, AMAT, INTU]
4	0.70	3	[CSCO, AMZN, ILMN, ADBE, IDXX, CHTR, EBAY, ANSS, ALGN, META, GOOG, WDAY, INTC, ADSK, NFLX, WBD, ABNB, NXPI, CMCSA, GOOGL, INTU]
5	0.75	21	[META, NXPI, WDAY, EBAY, CMCSA, ILMN, ALGN, ADBE]
6	0.80	5	[META, AMAT, WDAY, LRCX, EBAY, ASML, ANSS, NFLX, ALGN, ADBE]
7	0.85	1	[META, CHTR, EBAY, CMCSA, ALGN, ADSK, ZM, ILMN, ADBE, PYPL]
8	0.90	4	[EBAY, ALGN, ILMN, CMCSA]
9	0.92	3	[EBAY, ALGN, CMCSA]

Таким образом, акции Comcast Corp (12) были выбраны основными для сравнения моделей, так как они присутствуют во всех пересечениях максимальных клик и являются наиболее весомыми в индексе NASDAQ из тройки акций при пороге 0.92. eBay Inc(91) и Align Technology Inc(92) так же могут быть использованы для дополнительного анализа.

Базовые данные Данная работа устроена таким образом, что данные и модели будут постепенно улучшаться и усложняться, что позволит в некоторой степени проследить влияние тех или иных дополнений на итоговый результат, а так же сравнить некоторые подходы к предсказанию акций. Базовый набор данных включает в себя исключительно логарифмическую доходность акции в конце каждого

дня в течение заданного периода. Логарифмическая доходность для акции рассчитывается по формуле:

$LnProfit_i = ln(C_i/C_{i-1})$, где C_i - цена акции в период i.

Метрики качества и способ тестирования Так как в работе выполняется классификация временного ряда, то естественно положить точность предсказания основной метрикой качества модели. Точность модели рассчитывается как отношение количества верных предсказаний к общему количеству дней, для которых предсказания необходимо произвести.

Другая естественная метрика, которая позволит оценить прибыльность модели, есть доходность за некоторый тестовый период. Данная метрика показывает, во сколько раз увеличится изначальный капитал, вложенный в данную акцию, если в течение этого периода торги будут происходить согласно предсказания модели. Следовательно, любое значение больше 1 означает, что модель увеличила вложенный капитал за этот период, а любое значение меньше означает потери.

Однако, метрика доходности не даёт нам никакой оценки того, насколько в действительности модель проявила себя за данный период времени. Например, если весь период наблюдается исключительно рост цены акции, то любая модель покажет положитель-

ный доход. Для определения качества модели в случае успешной торговли была введена метрика потенциала. Данная метрика рассчитывается как отношение дохода разработанной модели к доходу "идеальной"модели, которая предсказывает абсолютно точно. Таким образом, если модель была успешна, то потенциал покажет, какую часть от максимального дохода она смогла заработать. Отрицательные значения метрики не интерпретируются, так как не существует однозначной трактовки.

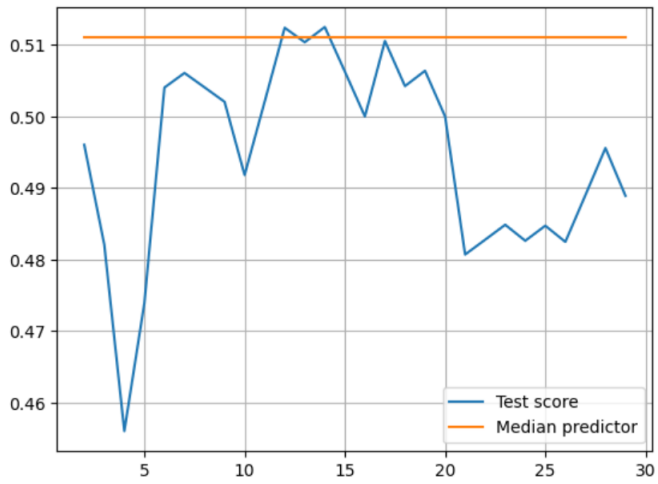
При этом важно понимать, что период оценки модели должен быть достаточно большой, ввиду наличия потенциально сильных колебаний цены в рамках короткого периода. Поэтому для оценки качества моделей используется разработанный тест, который выполняет предсказание для каждого месяца в тестовом году, при этом дообучая, обучая заново или используя предобученную модель. Обучающая выборка для всех случаев набирается для из данных до апреля 2022 года, чтобы модель не могла выучить заранее данные для теста.

Замечание. Для всех тестовых акций цена за тестируемый год снизилась или осталась на примерно таком же уровне. Такие условия делают невозможным применение простейшей стратегии заработка, при которой акция покупается в начале и продается в конце. Заработать можно исключительно на перепадах курса.

2.2 Stage 1: Moving Average

Алгоритм Данный метод применяется в основном для построения некоторой отправной точки при решении задачи предсказания временного ряда. Идея алгоритма достаточно простая и заключается в том, что предсказание на следующий день представляет из себя средневзвешенное всех значений за некоторый период, который называется окошком. Размер этого окошка (количество дней в периоде) - единственный гиперпараметр алгоритма, с помощью которого его можно адаптировать для данных.

Выбор размера окошка Для подбора оптимального гиперпараметра алгоритма было произведено дополнительное исследование. Был построен график зависимости точности алгоритма от размера окошка. Тут стоит уточнить, что данное исследование не должно проводиться на данных тестовой выборки (т.е. на данных с апреля 22 года по апрель 23 года), чтобы исключить любой фактор "запоминания"этих самых тестовых данных. Для выбранной основной акции оптимум точности достигается при размере окошка равному 12:

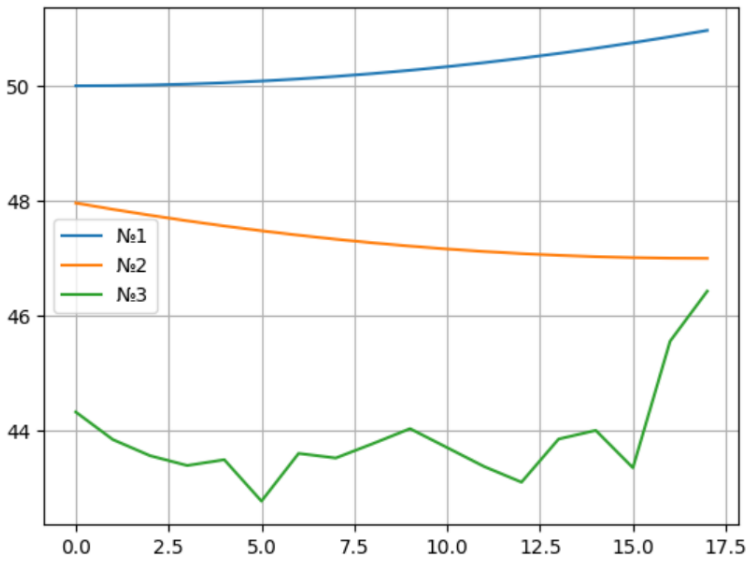


Важное уточнение! Размер окошка включает в себя и тестируемый день. Это означает, что реальное число дней, на котором достигается оптимальная точность алгоритма равно 11.

Результаты Замечание. Для каждого выигрышного исследованного подхода следующая таблица является отражением результата торгов с помощью модели за год. Она предоставляет все анализируемые метрики за тестируемый год.

	Month	TrainAcc	TestAcc	Income	Potential
0	May_2022	0.52	0.46	1.04	0.22
1	Jun_2022	0.59	0.46	0.96	-0.36
2	Jul_2022	0.63	0.51	1.03	0.24
3	Aug_2022	0.37	0.53	1.01	0.07
4	Sep_2022	0.50	0.51	0.98	-0.22
5	Oct_2022	0.48	0.56	1.02	0.20
6	Nov_2022	0.50	0.46	1.07	0.42
7	Dec_2022	0.47	0.61	1.01	0.07
8	Jan_2023	0.70	0.51	1.04	0.26
9	Feb_2023	0.47	0.46	0.94	-0.34
10	Mar_2023	0.41	0.44	0.99	-0.15
11	Apr_2023	0.45	0.59	1.08	0.58

Резульататы алгоритма оказались неоднозначными. С одной стороны, доходность такой модели в месяц в среднем составляет 1.3%. Это позволяет модели показать 16% годовых на упрощённом тесте торговли на рынке для выбранной акции за выбранный период. С другой, метрики имеют достаточно сильный разброс, что говорит о некоторой нестабильности в поведении модели в целом. Однако, модель данная базовая модель, несмотря на всю её простоту, обладает одним очень важным свойством, которое и позволяет ей демонстрировать подобную достаточно высокую доходность. Чтобы понять его, рассмотрим следующие графики цен на акции:



Первый график отражает стабильный растущий тренд. Но предсказание MA модели ограничено сверху максимальной ценой за период, опираясь на который модель выполняет презсказание. Следовательно, для данного случая модель будет постоянно предсказывать снижение цены на акцию, трейдинг, основанный на данном предсказании не будет закупать акции, что оставит капитал неизменным. Для стабильно убывающего тренда модель будет постоянно предсказывать повышение цены, что приведёт к тому, что трейдер купит акции по начальной цене и не будет их продавать, что приведёт к серьезным убыткам. Однако, для графика под номером три наблюдается немного иная ситуация. В начале, когда тренд стабильно убывающий, модель будет выдавать максимально неэффективные пред-

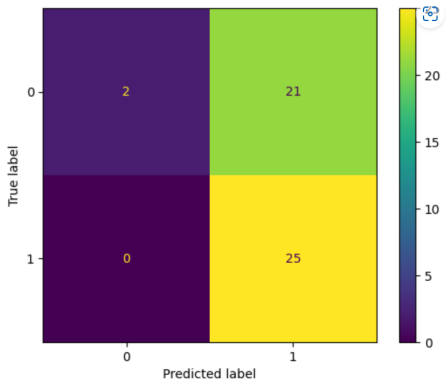
сказания. Но для дальнейшего периода модель ведёт себя иначе. Так как модель отражает тренд во временном ряде за некоторый период, то частые колебания вокруг этого тренда всегда точно определяются. Соответственно, для графика номер 3 данная модель должна показать достаточно неплохие результаты.

График номер 3 - это реальная цена на акции компании EBAY за апрель 2022 года. В этот период модель показала наилучшее качество предсказания. При этом, так общий тренд у акций растущий, MA модель в редко приводит к серьезным потерям, что и позволяет получить достаточно высокую доходность за год.

2.3 Stage 2: LogReg

Алгоритм Данная модель является в некотором смысле обобщением предыдущего метода. Она так же делает предсказание, основываясь взвешенном значении за последние несколько дней. Однако, веса значений подбираются так, чтобы наилучшим образом соответствовать обучающей выборке. Размер окошка для данного метода положим тем же, что был найден в предыдущем пункте.

Результаты и сравнение



Данный алгоритм в основном предсказывает, что цена акции на следующий день будет выше. Такое поведение приводит к неспособности модели показывать эффективные результаты торговли: за год доходность составила -3%, что означает полную бесполезность данной модели. Однако, подобный результат происходит из-за эффекта "загрязненности"текущих данных, о котором будет рассказано позже. Сейчас же попробуем облегчить действие данного эффекта, увеличив количество полезной информации в наших данных.

2.4 Stage 3: Relative Strength Index + LogReg

Описание RSI - это инструмент технического анализа акций, который показывает соотношение положительных и отрицательных изменений цены. Он обладает простой интерпретацией и зависит от небольшого окошка свежих данных. Данный индекс очень важен при

шаблонном анализе, так как с его помощью можно оценить наличие и силу тренда в цене акции.

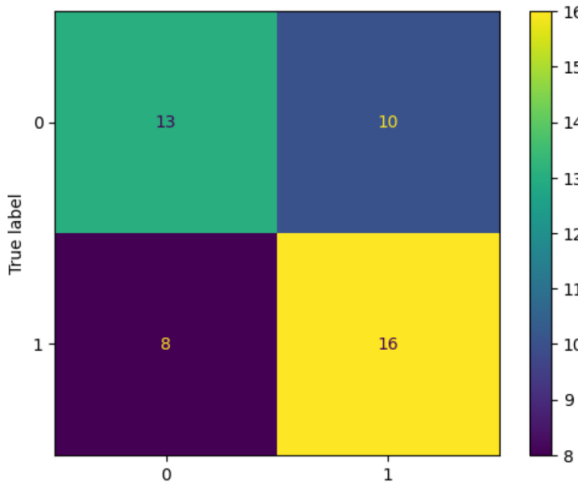
Индекс рассчитывается по следующей формуле:

$RSI = 100 - 100/(1 + Up/Down)$, где Up - количество дней с возрастающей ценой на акцию, а $Down$ - количество дней с убывающей ценой на акцию. Если $Down$ равен нулю, то показатель равен 100.

RSI измеряется в процентах и позволяет получить дополнительную информацию из графика цены на акцию. Например, достаточно высокие показатели данного индекса свидетельствуют о "перекупленности"акций (ситуации, когда людей, готовых продать больше, чем людей готовых купить актив) и о вероятном скором снижении цены, а достаточно низкие значения говорят о "перепроданности"(когда цена на актив слишком занижена, что ведет к повышенному интересу к покупке) и о скором увеличении цены.

Также дополнительную информацию несут в себе 50% переход и дивергенция. Первое можно наблюдать в случае, когда график RSI из области низких значений 20-40% уверенно переходит в область 50-70%. Второе происходит, когда несмотря на сильный убывающий тренд в цене на актив, наблюдается возрастающий тренд индекса RSI. В обоих случаях данные говорят о скорой смене тренда сонаправленно RSI.

Результаты и сравнение



	Month	TrainAcc	TestAcc	Income	Potential
0	May_2022	0.64	0.33	1.01	0.11
1	Jun_2022	0.60	0.60	1.04	0.34
2	Jul_2022	0.59	0.54	1.04	0.26
3	Aug_2022	0.60	0.51	0.96	-0.57
4	Sep_2022	0.61	0.50	0.95	-0.45
5	Oct_2022	0.62	0.49	1.03	0.30
6	Nov_2022	0.60	0.50	1.00	-0.03
7	Dec_2022	0.59	0.57	1.04	0.32
8	Jan_2023	0.59	0.57	1.04	0.48
9	Feb_2023	0.60	0.49	1.00	0.01
10	Mar_2023	0.60	0.52	0.99	-0.14
11	Apr_2023	0.60	0.47	1.04	0.33

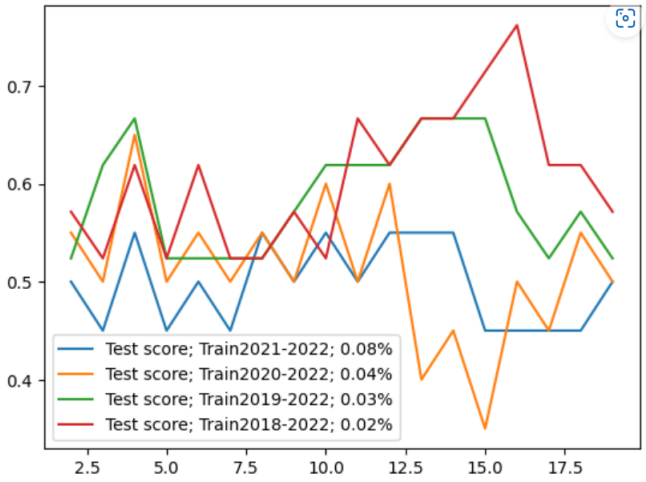
Как можно наблюдать, добавление данного индекса в наши данные позволило частично избавиться от проблемы "загрязненности"данных.

Модель показывает достаточно неплохие значения метрик, что позволяет ей увеличить вложенный капитал за год на 14%, что ниже чем доходность Moving Average модели. Потенциал и точность предсказаний данной модели также достаточно низки, что логично, ведь модель выучивалась распознавать один самый популярный шаблон.

2.5 Stage 4: ML classification

Описание Выбранный метод предсказания позволяет использовать группы классических классификационных алгоритмов для решения данной задачи. Так как данная работа базируется на предположении наличия в данных некоторых шаблонов, которые позволяют определить цену на будущие периоды, то естественно положить алгоритм KNearestNeighbours базовым в данной секции.

Наиболее важными параметрами для алгоритма являются метрика расстояния, количество соседей и количество тренировочных данных. Для начала положим метрикой сходства Эвклидово расстояние с коэффициентом 2 (базовое значение для алгоритма из библиотеки). Для выбора наилучших параметров рассмотрим поведение точности на одном месяце при 1, 2, 3 и 4 годах обучающих данных и разных значениях количества соседей:



Общий пик подъема точности происходит при 4 соседях, при этом наилучшее значение достигается при трёх годах данных. Пара 16 соседей и 4 года данных не может быть использована так как является некоторым аналогом переобучения, но для процесса подбора параметров: эта пара имеет лучшие показатели на валидационном наборе, но в её окрестности значение точности резко убывают, что означает её особенную приспособленность для валидационной выборки и слабую обобщающую способность для других данных.

Дополнительно была исследована возможная длительность предсказания модели. Иными словами, было рассмотрено, как долго можно использовать данную модель без необходимости дообучения новыми данными. Результаты для 1, 2, 3 и 4 месяцев тестирования можно увидеть на следующем графике точности от количества соседей:

Таким образом, результаты предсказания сильно ухудшаются при увеличении тестируемого периода, что требует дообучения модели в разработанном тесте годовой доходности.

Результаты и сравнение

	Month	TrainAcc	TestAcc	Income	Potential
0	May_2022	0.69	0.56	1.03	0.24
1	Jun_2022	0.69	0.49	1.01	0.10
2	Jul_2022	0.69	0.57	1.06	0.45
3	Aug_2022	0.69	0.46	0.97	-0.32
4	Sep_2022	0.69	0.51	0.96	-0.35
5	Oct_2022	0.69	0.48	1.03	0.25
6	Nov_2022	0.69	0.51	1.01	0.10
7	Dec_2022	0.69	0.54	1.01	0.13
8	Jan_2023	0.69	0.36	1.00	-0.04
9	Feb_2023	0.69	0.54	0.98	-0.30
10	Mar_2023	0.69	0.58	0.99	-0.09
11	Apr_2023	0.69	0.50	1.02	0.15

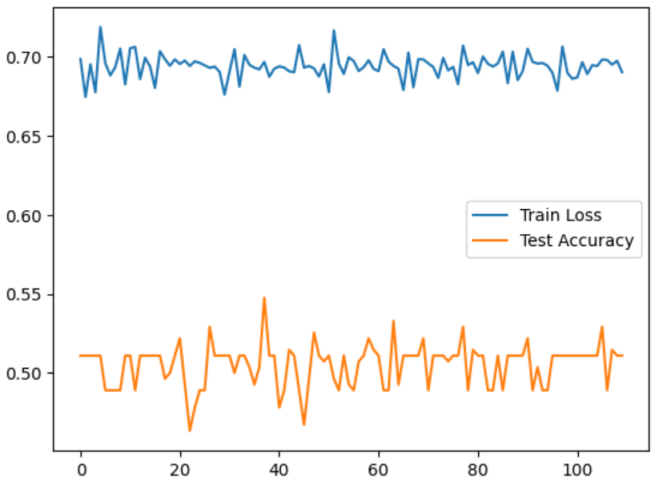
Модель не показала сильных потерь за тестируемый период, однако высоких значений доходности также не было продемонстрировано, что привело к относительно невысокой доходности данного подхода: 8.6%.

Глава 3

CNN подход

3.1 Описание выбранной модели

Для построения базовой модели в данной части воспользуемся классическим для CNN набором и последовательностью операторов. Блок свертки состоит из Conv1d, MaxPool1d и Relu активации, число блоков 3, при этом каждый блок понижает длину входящего ряда, но увеличивает количество каналов. В качестве периода для обучения возьмем данные за четыре года и год для тестирования. График обучения данной модели следующий:



3.2 Проблема загрязненности данных

Как можно видеть по графику обучения, модель не выучивается находить и запоминать зависимости в данных. При увеличении частоты отрисовки графика можно заметить, что значения функции потерь периодически колеблются вокруг некоторого значения. Одной из причин подобного поведения может быть „загрязненность“ данных.

При решении классификации временного ряда стоит помнить, что значение целевой переменной теперь строго фиксировано. При этом, если в задаче регрессии близкие временные ряды могли показывать достаточно близкие значения доходности, например, 0.98 и 1.01, то в задаче классификации подобные значения относятся к

противоположным категориям. Это приводит к тому, что модель получает противоречивые данные, когда один и тот же шаблон может означать что угодно. Потенциально данную проблему можно решить увеличением размера разового обучающего пакета, однако, есть куда более простое решение.

Чтобы понимать, как обстоят дела с данным эффектом сейчас и насколько хорошим будет любой избранный метод, необходимо ввести метрику подсчета влияния данного эффекта в наших текущих данных. Таким может быть коэффициент загрязненности данных.

В основе данного коэффициента лежит достаточно простая идея: мы хотим определить, какой процент пар записей, которые сильно коррелируют, в данных в действительности указывают на разные события: один на повышение цены, другой на понижение. Таким образом, чтобы рассчитать данный коэффициент необходимо провести следующие шаги:

1. Посчитать матрицу корреляций для тренировочного набора данных

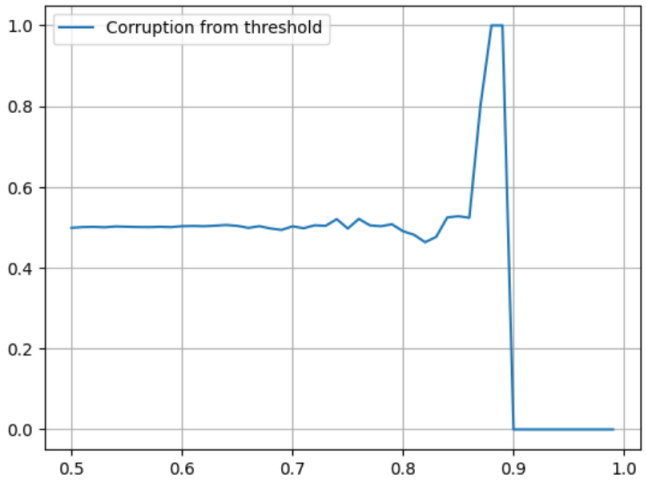
2. Найти все пары элементов из матрицы, для которых значение корреляции выше заданного порога

3. Посчитать, для скольких из этих пар значения целевой перемен-

ной различны (иными словами, количество false correlated пар)

4. Поделить найденное число на общее количество пар для приведения значения к отрезку [0, 1]

Интерпретация данного коэффициента очень проста: чем выше значение, тем более загрязнены наши данные, и тем сложнее модели выучить общие закономерности. При этом, мы так же хотим, чтобы при увеличении порога значение данного коэффициента снижалось, ведь обратная ситуация также ведет к усложнению процесса обучения. Для текущих данных график значения параметра выглядит следующим образом:



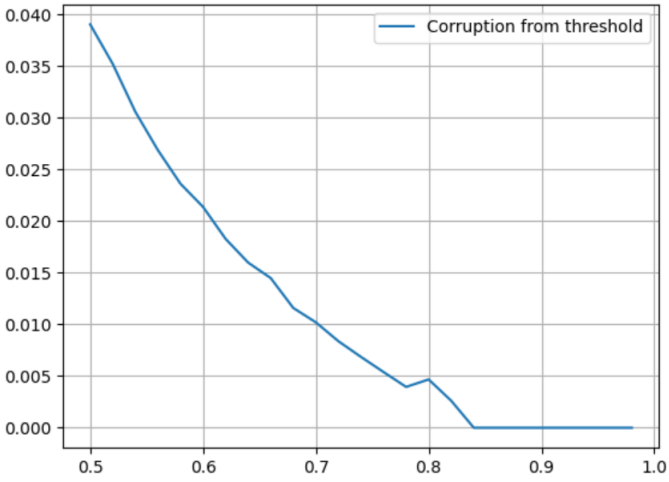
Загрязненность данных на уровне 50% для низких значений порога и крайне большое для высоких значений, что и приводит к по-

добным колебаниям значения функции потерь в процессе обучения модели.

3.3 Улучшение качества данных

Чтобы усилить отличия в данных необходимо добавить в них больше информативности. Для достижения этой цели можно добавить ранее неиспользованные исторические данные. Конкретно, были использованы цена открытия, также максимум и минимум цены за день, приведенные к относительной цене делением на цену закрытия.

Для многомерного случая рассчет коэффициента дополняется дополнительным шагом. Теперь вместо корреляции для каждой пары считается сумма корреляций по всем введенным данным, а выбор пар в шаге 2 происходит по некоторому заранее определенному порогу. Данный порог выбирается исходя из ответа на вопрос: какое количество переменных должно коррелировать, чтобы считать пару данных коррелирующей. То есть, чем больше данный порог, тем сильнее должна быть связь внутри пары. Дня следующего графика значение порога равно трём (иными словами, чтобы полагать пару коррелирующей, необходимо наличие корреляции по трём из пяти переменным):

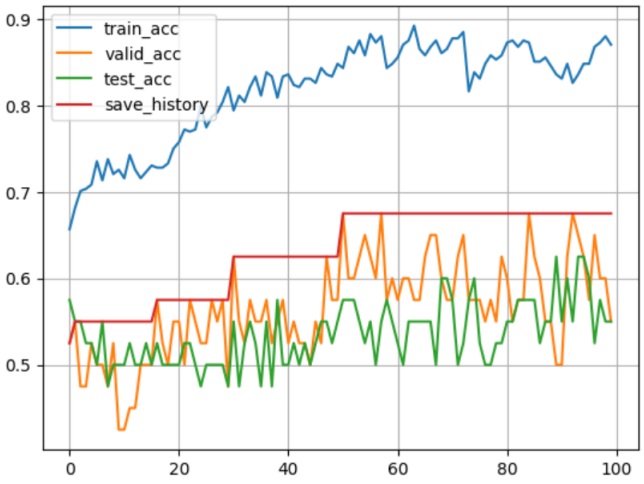


Таким образом, коэффициент загрязненности теперь ведет себя необходимым образом: значение в отдельных случаях относительно мало и тем меньше, чем больше порог корреляции.

Для достижения лучших результатов граф вычислений нейронной сети также был изменен. Для ускорения обучения был оставлен только один сверточный слой, и обычный MaxPool1d был заменен на ConcatPool.

3.4 Базовые результаты

После изменений внесенных в данные и в модель график обучения стал выглядеть следующим образом:



В процессе обучения использовались данные за два года без пересечения с глобальной тестовой выборкой. Для валидации и локального теста использовались последние четыре месяца выбранного периода. Модель явно начала выучивать закономерности в данных. Отбор лучшей просходил по точности на валидационной выборке.

	Month	TrainAcc	TestAcc	Income	Potential
0	Jun_2022	0.621849	0.705882	0.985177	-0.224187
1	Jul_2022	0.554622	0.588235	1.014576	0.167600
2	Aug_2022	0.607595	0.388889	0.986844	-0.216034
3	Sep_2022	0.579832	0.631579	0.941759	-0.493834
4	Oct_2022	0.546218	0.500000	1.040197	0.625789
5	Nov_2022	0.561181	0.352941	1.003814	0.036806
6	Dec_2022	0.550420	0.611111	1.006303	0.078177
7	Jan_2023	0.571429	0.466667	1.031548	0.442787
8	Feb_2023	0.662447	0.687500	1.016432	0.411261
9	Mar_2023	0.531646	0.578947	0.999928	-0.001155
10	Apr_2023	0.561181	0.722222	1.125997	0.904208
11	May_2023	0.605932	0.411765	1.007474	0.209106

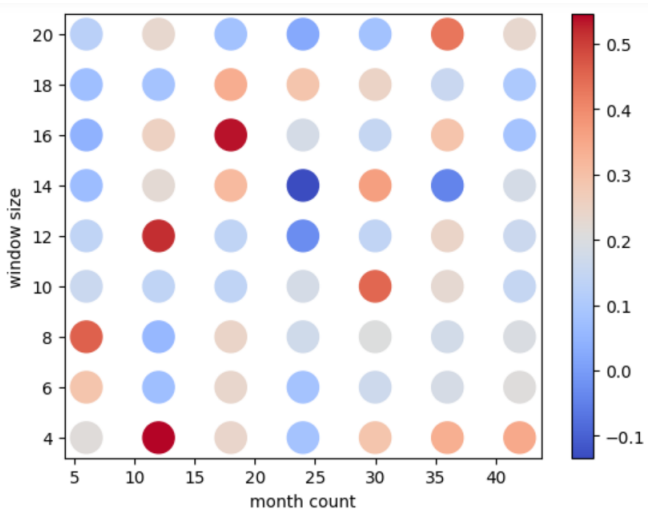
Модель часто показывает высокие значения точности предсказания и в апреле 2023 года показывает рекордную доходность в 12.6%, что позволяет получить в итоге доходность в 16% годовых, что сильно лучше результата регрессии и классификации и совпадает с результатом MA модели.

3.5 Исследование оптимальных параметров

В процессе обучения бейзлайна было обнаружено, что качество модели очень сильно зависит от выбранных параметров обучения та-

ких как размер обучающей выборки и размер окошка, поэтому было проведено дополнительное исследование зависимости от этих параметров. Для валидации использовался год данных, предшествующий глобальному тестовому году.

Для поиска оптимальных параметров был имплементирован алгоритм GridSearch. Диапазон значений выбирался из выбирался из соображений предыдущих экспериментов. Итоги алгоритма представлены ниже:



Значение годовой доходности колеблется между -10% и 55%. Наилучшую доходность показала пара размер окошка 4 и число тренировочных месяцев 12. Однако, мы можем наблюдать, что в окресности данной точки наблюдается резкое снижение доходности, а мы бы хотели стабилизировать наши результаты. Поэтому основной точкой

была выбрана пара 16, 16, около которой снижение доходности происходит не так резко. В итоге, для обучения модели использовалось 20 месяцев данных (не пересекающихся с глобальным тестом), 4 из которых были отданы на валидацию.

3.6 Результаты

	Month	TrainAcc	TestAcc	Income	Potential
0	Jun_2022	0.514768	0.625000	1.035848	0.572870
1	Jul_2022	0.481013	0.470588	0.974312	-0.295367
2	Aug_2022	0.525424	0.500000	0.977617	-0.367546
3	Sep_2022	0.489451	0.611111	1.008681	0.073607
4	Oct_2022	0.497890	0.555556	1.067916	1.057333
5	Nov_2022	0.538136	0.705882	1.062132	0.599592
6	Dec_2022	0.481013	0.666667	1.027451	0.340493
7	Jan_2023	0.502110	0.600000	1.008533	0.119757
8	Feb_2023	0.500000	0.500000	0.998371	-0.040782
9	Mar_2023	0.584746	0.722222	1.019508	0.311574
10	Apr_2023	0.542373	0.388889	1.000185	0.001324
11	May_2023	0.510638	0.588235	0.988921	-0.309973

Модель показала 17,7% годовой доходности, что является наилучшим результатом среди всех рассмотренных алгоритмов.

3.7 Предсказание на кликах

Как было показано в начале работы, на рынке существуют некоторые зависимые образования (клики) из акций, цена на которые изменяется схожим образом. В такой ситуации также достаточно инетесно проверить, сможет ли модель, обученная на данных одной акции, уверенно предсказывать и показывать прибыль на других зависимых акциях и насколько эффективным будет подобное предсказание.

Чтобы проверить это, для начала возпользуемся акциями, избранными в начале работы. Так как данные акции находятся на пересечении максимальных клик при высоком пороге, то зависимость между ними должна быть крайне высокой. При запуске претренированной модели на тикере EBAY получаем годовую доходность в 28%, при обучении с нуля для данного тикера получаем доходность в 29%. В данном случае можно считать, что претренированная модель предсказывает поведение акции столь же хорошо. Однако, для тикера ALGN доходность предобученной модели составляет 6%, в то время как доходность обученной с нуля - 65%. В чем причина подобного поведения?

Дело в том, что тикеры CMCSA и EBAY вели себя последний год следующим образом:

Графики очень похожи, EBAY отличается более резкими переходами, но тренды одинаковые. Теперь сравним это с графиком цены ALGN:

Поведение за последний год значительно отличается, что и не позволяет нашей предобученной модели столь же хорошо чувствовать себя на данной модели.

Другой вопрос заключается в столь серьезном успехе модели, которая была обучена с нуля. 65% - это невероятно высокая годовая доходность. Если посмотреть на табличку метрик,

	Month	TrainAcc	TestAcc	Income	Potential
0	Jun_2022	0.611814	0.562500	1.019814	0.198566
1	Jul_2022	0.637131	0.529412	1.105016	0.570812
2	Aug_2022	0.588983	0.500000	0.998116	-0.015154
3	Sep_2022	0.624473	0.777778	1.072966	0.636249
4	Oct_2022	0.582278	0.444444	1.046345	0.223800
5	Nov_2022	0.593220	0.411765	1.061227	0.332927
6	Dec_2022	0.632911	0.611111	1.040863	0.267133
7	Jan_2023	0.582278	0.600000	1.100757	0.744067
8	Feb_2023	0.584746	0.562500	0.970365	-0.430332
9	Mar_2023	0.601695	0.666667	1.007454	0.099961
10	Apr_2023	0.631356	0.666667	1.056909	0.350927
11	May_2023	0.634043	0.647059	1.043356	0.524279

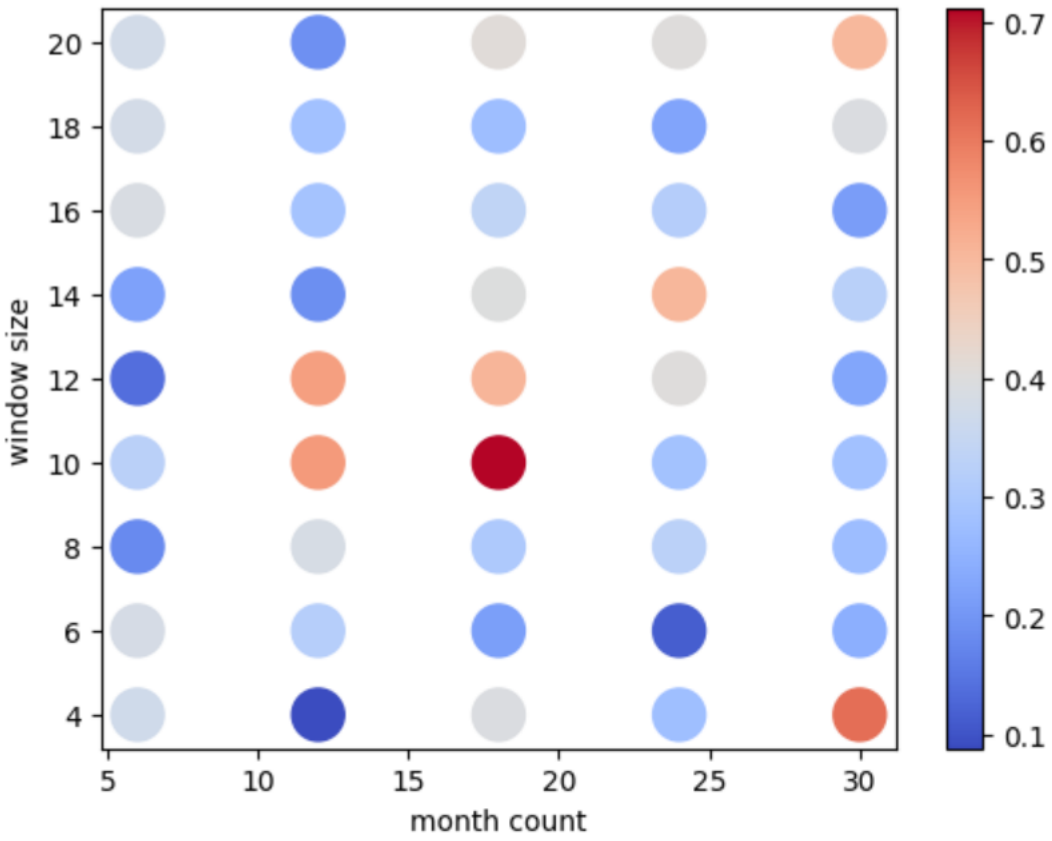
то можно заметить, что основная часть доходности приходится на Июль 22 года и Январь 23 года. В эти месяцы цена на актив стабильно росла, причем крайне высокими темпами, что и позволило модели так выгодно торговать в эти периоды и получить такую доходность.

3.8 Multi-stocks обучение

Подход описанный ранее позволяет нам обучить достаточно хорошую модель для предсказания цены для любой акции. Найденные гиперпараметры относительно стабильны и могут быть использованы как базовые для других акций. Однако, данный подход требует повторения всего процесса обработки для получения отдельной модели, предсказывающей конкретную акцию, что в некотором роде избыточно.

Для акций достаточно близких модель можно попробовать лишь немного дообучить, чтобы она запомнила особенности конкретной акции или можно сразу обучать модель на большом наборе данных, в котором содержатся данные нескольких активов сразу. В частности интересно то, какие последствия будут у подобного подхода: станет ли доходность для каждой акции меньше из-за того, что модель будет больше фокусироваться на общих признаках, или наоборот вырастет, так как у модели будет гораздо больше данных для обучения.

Для обучения были взяты основные тикеры CMCSA, EBAY и ALGN. Обучение происходит на совмещенном наборе, а торговля отдельно на каждой акции из списка. Полученная карта параметров выглядит следующим образом:



Исходя из целей стабилизации результатов, оптимальные параметры были немного сдвинуты. Таким образом, итоговые решение строилось на 16 месяцах данных с окном в 11 дней. При данных гиперпараметрах доходность модели составила 23% на ALGN, 12% на EBAY и -2% на CMCSA. Данные значения сильно уступают значениям от моделей, обученных отдельно на каждую акцию. Что подтверждает гипотезу о превосходстве раздельного обучения моделей.

Глава 4

Заключение

Таким образом, было проведено подробное исследование по применимости подхода классификации для предсказания акций. Подход имеет свои недостатки, однако, позволяет получить интересные результаты. Также в работе был исследован ряд дополнительных вопросов касающихся графового анализа и шаблонного анализа рынка акций. Для данного исследования был разработан специальный пакет функций, который упрощает работу с данными акций. В итоге, проведенное исследование является хорошей базой для дальнейших исследований в данном направлении по улучшению качества данных, моделей и подходов.

Глава 5

Источники и ресурсы

1. Код и материалы: https://github.com/PoliakovVI/SPP

2. Применение рыночных графов к анализу фондового рынка России https://publications.hse.ru/pubs/share/folder/t4pelon6ld/66469958.pdf

3. Wei Li, Ruihan Bao, Keiko Harimoto, Deli Chen, Jingjing Xu and Qi Su [2020] Modeling the Stock Relation with Graph Network for Overnight Stock Movement Prediction

4. C Anand [2021] Comparison of Stock Price Prediction Models using Pre-trained Neural Networks

5. Sidra Mehtab, Jaydip Sen and Abhishek Dutta [2020] Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning

Models

6. Saloni Mohan, Sahitya Mullapudi, Sudheer Sammeta, Parag Vijayvergia and David C. Anastasiu [2019] Stock Price Prediction Using News Sentiment Analysis

7. Payal Soni [2022] Machine Learning Approaches in Stock Price Prediction: A Systematic Review

8. https://machinelearningmastery.com/how-to-develop-convolutional-neural-network-models-for-time-series-forecasting/