МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В.Ломоносова

Механико-математический факультет
Кафедра вычислительной математики

КУРСОВАЯ РАБОТА

студента 504 группы
Анисимовой Марии Владимировны

Прогнозирование временных рядов на основе обобщённых японских свечей.

Forecasting of temporary ranks on a basis the generalized Japanese candles.

Научный руководитель,
д.ф.-м.н. Кумсков М.И.

Москва, 2020 год.

Аннотация.

Решение задачи прогнозирования играет важнейшую роль в процессах как стратегического планирования, так и оперативного управления в различных сферах науки и техники, финансах и медицине.

В данной работе рассматривается задача прогнозирования будущего тренда временного ряда по основным характеристикам японских свечей. Будет рассмотрена общая постановка задачи классификации и ее частный случай - классификация будущего тренда. А также возможные методы решения данной задачи.

Так же в работе приведена подробная реализация одного из описанных методов и ее результаты. Удалось добиться точности работы модели 0.86, а значит имеет смысл и дальше её развивать.

Содержание

0. Введение.

0.1 Обзор.

Одна из самых древних задач, в которых применимы методы анализа данных, — это задача прогнозирования. Прогнозировать можно что угодно: продажи товаров в магазинах, рейтинги телесериалов, пробки, погоду, землетрясения, котировки. Методы машинного обучения и анализа данных могут проанализировать историческую информацию, найти в ней какие-то закономерности и на основании этих закономерностей научиться предсказывать будущее.

Чаще всего процессы, перспективы которых необходимо предсказывать, описываются временными рядами, то есть последовательностью значений некоторых величин, полученных в определенные моменты времени.

В данной работе рассматривается задача прогнозирования временных рядов, а именно - котировок, показан достаточно старый, но в настоящий момент один из самых популярных , метод графического представления биржевой информации, называемый "японские свечи". Японские свечи - очень удобный метод подачи информации, он легко может быть использован для автоматизации работы биржевых аналитиков и трейдеров. Японские свечи универсальны — ими могут пользоваться как зрелые профессионалы, так и новички в техническом анализе. Причина состоит в том, что графики свечей можно использовать вместе с другими инструментами технического анализа. Возможность использования свечей вместе с другими инструментами, а не вместо них — их несомненное достоинство. Основная причина столь пристального внимания к свечам заключается в том, что использование их вместо или в дополнение к столбиковым графикам значительно увеличивает шансы на успех.

Интерпретация свечей, как, впрочем, и любой другой метод графического анализа, носит субъективный характер. Это можно считать недостатком. Но приобретая опыт использования японских свечей на конкретном рынке, можно отобрать те графические модели или их варианты, которые работают наиболее успешно.

0.2 Виды анализов.

0.2.1 Технический анализ.

Технический анализ – совокупность инструментов прогнозирования вероятного изменения цен на основе закономерностей изменений цен в прошлом в аналогичных обстоятельствах. Базовой основой является анализ графиков цен — «чартов» (от англ. chart — график, диаграмма).

Его основополагающий принцип заключается в том, что человек способен проанализировать динамику цен в исторически обозримом прошлом, и на такой основе определять текущие условия для торговли и потенциальную ценовую динамику. Очевидность использования технического анализа обоснована тем, что теоретически вся текущая информация о рынке отражена в ценах. И если цена отражает всю информацию, которая в ней содержится, то движение цен — это все, что действительно следует знать, чтобы успешно торговать. Технические аналитики ищут в текущей ситуации шаблоны, которые уже формировались когда-либо в прошлом и формируют свои торговые стратегии, полагая, что цена и в будущем будет действовать также, как она это делала прежде.

Основные постулаты:

1. Курс (цена) учитывает все. Любой фактор, влияющий на цену (экономический, политический или психологический), уже учтен рынком и включен в цену. Поэтому все, что требуется для прогнозирования, - изучать график цены.
2. Движение цен подчинено тенденциям (тенденция – это направление движения цены). Основная цель составления графиков динамики цены заключается в том, чтобы выявить эти тенденции на ранних стадиях их развития и торговать в соответствии с их направлением.
3. История повторяется. Аксиома базируется на неизменности основ человеческой психики, а отсюда и применимости анализа, работавшего в прошлом, к событиям настоящего.

О 5-ти преимуществах технического анализа рынка:
«Технический анализ имеет пять основных преимуществ.
Во-первых, хотя фундаментальный анализ и позволяет оценить соотношение спроса и предложения, показатели экономической статистики, определить отношение цены акции к прибыли компании и т.д., в нем отсутствует психологический компонент, столь необходимый для адекватной оценки рыночной ситуации. Грамм эмоций иногда равноценен килограмму фактов. Как заметил Джон Маньярд Кейнз: «Нет ничего более гибельного, чем рациональная инвестиционная политика в нашем иррациональном мире»2. Технический анализ — единственное средство измерить этот «иррациональный» (эмоциональный) компонент, обязательно присутствующий на каждом рынке.
Во-вторых, технический анализ является важным компонентом дисциплинированной торговли. Дисциплина позволяет усмирить основной бич трейдеров: эмоции. Как только вы вошли в рынок, место у руля занимают эмоции, а рациональный и объективный расчет превращаются в простых пассажиров.
В-третьих, даже если вы не верите в технический анализ, к рекомендациям аналитиков следует прислушиваться, поскольку иногда технические факторы являются основной движущей силой рынка.
В-четвертых, теория случайных событий утверждает, что сегодняшние рыночные цены никак не влияют на рыночные цены последующих дней. Но этот академический подход не учитывает важного компонента рынка: людей. Люди запоминают цены предшествующих дней и действуют соответственно. Тем самым действия людей влияют на цену, но справедливо и обратное — сама цена влияет на эти действия. Таким образом, сама цена является важным компонентом анализа рынка. Те, кто относится к техническому анализу с пренебрежением, забывают об этом.
В-пятых, цена является самым наглядным проявлением соотношения спроса и предложения. Широкая публика может и не знать каких-то фундаментальных обстоятельств, но они неизбежно отразятся в цене. Те, кто знает о том или ином важном событии на рынке заранее, скорее всего, купят или продадут до того, как эта информация повлияет на цену. Поэтому, когда происходит само событие, информация о нем зачастую уже бывает учтена рынком. Таким образом, в текущей цене отражается любая информация известная как широкой публике, так и ограниченному кругу посвященных.» - Steve Nison.

Большая часть того, что сегодня известно нам как Технический анализ - родилось из идей предложенных Чарльзом Доу (Charles Dow) и его партнером Эдвардом Джонсом (Edward Jones), работавших в компании Dow Jones & Company с 1882 года. Эти идеи были опубликованы в Уоллстрит джорнал (Wall Street Journal) и в наши дни принимаются подавляющим большинством практиков технического анализа, несмотря на то, что большинство из них не знакомы с их источником.

Теория Доу до сих пор доминирует в сегодняшнем, намного более сложном и хорошо вооруженном, подходе к техническому анализу.

Основные положения теории Доу:

- Индексы учитывают все. Согласно теории Доу любой фактор, способный, так или иначе, повлиять на спрос или предложение, неизменно найдет свое отражение в динамике индекса. Разумеется, эти события непредсказуемы, тем не менее, они мгновенно учитываются рынком и отражаются на динамике индексов.
- На рынке существуют три типа тенденций. При *повышательной тенденции* каждый последующий пик выше предыдущего и каждый последующий спад также выше предыдущего. При *нисходящей тенденции* каждый последующий пик ниже предыдущего и каждый последующее «донышко» ниже, чем предыдущее. При *горизонтальный тенденции (флэте)* каждый последующий пик (и спад) находится примерно на том же уровне, что и предыдущие (см. рис.).

рис. 1 Три типа тенденций.

- Основная тенденция имеет три фазы. Фаза первая, или фаза накопления, - когда наиболее дальновидные и информированные инвесторы начинают покупать, т.к. вся неблагоприятная экономическая информация уже учтена рынком. Вторая фаза наступает, когда в игру включаются те, кто использует технические методы следования за тенденциями. После того, как экономическая информация становится все более оптимистической, тенденция входит в свою третью, заключительную фазу, когда в действие вступает широкая публика, и на рынке начинается ажиотаж, подогреваемый средствами массовой информации. Экономические прогнозы в газетах и на телевидении полны оптимизма. Это первый признак окончания тенденции.
- Индексы должны подтверждать друг друга. Тут Чарльз Доу имел в виду промышленный и железнодорожный индексы. Он полагал, что любой важный сигнал к повышению или понижению курса на рынке должен пройти в значениях обоих индексов. Касательно современного технического анализа это утверждение означает, что сигнал, полученный от одного технического индикатора, должен быть подтвержден показаниями другого технического индикатора.
- Объем торговли должен подтверждать характер тенденции. Увеличение объема торговли должно происходить в моменты, когда цены двигаются в направлении основной тенденции, а уменьшение объема – в периоды отката.
- Тенденция действует до тех пор, пока не подаст явных сигналов о том, что она изменилась

0.2.2 Фундаментальный анализ.

Фундаментальный анализ — это способ наблюдения за рынком, при котором анализируются экономические, социальные и политические события, которые способны влиять на спрос и предложение.

И это имеет огромный смысл, так же, как и в экономической теории, где утверждается, что баланс спроса и предложения определяет цену. Использование спроса и предложения в качестве индикатора того, куда может «направиться» цена, несложно. Разумно анализировать все факторы, которые воздействуют на спрос и предложение.
Иначе говоря, следует учитывать различные факторы и статистические показатели, чтобы определить, какая из национальных экономик на подъеме, а какая входит в полосу падения. Необходимо также понять причины того, как определенные события, такие,

например, как увеличение безработицы, влияют на экономику конкретной страны, и, в конечном счете, на уровень спроса на соответствующую валюту.

Главная идея, положенная в основу фундаментального типа анализа, состоит в том, что если текущая или будущая экономическая перспектива для этой страны благоприятна, то ее валюта должна укрепляться. Чем лучше «работает» экономика какой-либо страны, тем больше инвестиций из-за рубежа будет привлечено в такую страну. Это приводит к возрастанию потребности в ее валюте с целью приобретения соответствующих активов. Именно этот принцип заложен в основу фундаментального анализа.

Обычно используются оба метода, но в разных пропорциях. При долгосрочной, инвестиционной , торговле больше внимания уделяется фундаментальному анализу. А для краткосрочной спекулятивной торговли в большей степени используется технический анализ.

1. Глава 1. Подход к постановке задачи.

1.1 Основные определения.

Рассмотрим основные понятия в биржевой торговле:

Финансовый инструмент– *это актив, который может выступать в качестве объекта торговли.*

Котировка (фр. Cote, англ. Financial quote) — *цена (курс, процентная ставка) товара, которую объявляет продавец или покупатель и по которой они готовы совершить покупку или продажу (предлагается оферта). Обычно подразумевается относительно быстро меняющаяся цена, например биржевая.*

Тик- *минимальное колебание курса на бирже, установленное биржевыми правилами.*

Примеры финансовых инструментов:

· акции компании
· наличная валюта
· золото, нефть и другие сырьевые товары
· опционы
· фьючерсы

Сделки совершаемые на бирже приводят к открытию или закрытию позиции трейдера. Так покупка финансового инструмента в расчете на рост его стоимости называется открытием длинной позиции. А суммарная стоимость купленного инструмента называется стоимостью позиции. Стоимость позиции будет расти или падать. Последующая продажа этого инструмента называется закрытием длинной позиции и приводит к фиксации накопленной прибыли или убытка от изменения стоимости позиции.

Позиции открываемые трейдером :

- **Длинная позиция (позиция лонг) (от англ. long position)**- *это позиция, которую трейдер открывает в надежде получить прибыль от роста рынка. Трейдер покупает акции дешево, ждет, когда цена вырастет, продает акции дорого, и таким образом получает прибыль от роста рынка.*
- **Короткая позиция (позиция шорт) (от англ. short position)**- *это позиция, которую трейдер открывает в надежде получить прибыль от падения рынка. Для этого трейдер берет акции взаймы у брокера в натуральной форме, продает акции на открытом рынке дорого, ждет, когда цена акций упадет, покупает акции на открытом рынке дешево, отдает займ брокеру в натуральной форме, а разница между дорогой продажей и дешевой покупкой остается у трейдера — это его прибыль. Таким образом трейдер получает прибыль от падения рынка.*

Трейдер анализирует ситуацию на рынке и принимает определенные решения. Основным параметром финансового инструмента является цена. Именно изменение цены инструмента приводит к прибыли или убытку трейдера. Правильная оценка динамики цены инструмента является базисом прибыльной торговли. Для принятия торговых решений трейдер тщательно изучает выбранный для торговли инструмент.

Держателей длинных позиций (лонгов), получающих прибыль от роста рынка, называют еще игроками на повышение, или быками. Этимология использования слова «бык» в этом контексте доподлинно неизвестна, но запомнить, что быки играют на повышение, очень легко: бык поддевает рынок на рога, подбрасывает вверх, и цены растут.

Держателей коротких позиций (шортов), получающих прибыль от падений рынка, называют еще игроками на понижение, или медведями. Этимология использования слова

«медведь» в этом контексте также неизвестна, но запомнить, что медведи играют на понижение, тоже очень легко с помощью ассоциации: медведь встает на задние лапы, наваливается на рынок, прижимает его к земле, и цены падают.

- **Брокер (англ., от broker— посредник, маклер)** – *это юридическое или физическое лицо, которое выступает в качестве посредника меж продавцами и покупателями (страховой компанией и страхователем, судовладельцем и фрахтователем) ценных бумаг, драгоценных металлов, валюты и других услуг и товаров. Брокер действует, придерживаясь поручений собственных клиентов, и выполняет сделки за их средства, получая вознаграждение в виде комиссионных выплат. Для ведения брокерской деятельности нужно получить соответствующую лицензию. Помимо посреднических услуг купли-продажи, брокер может также предоставлять консультационные услуги.*

Тренд(от английского слова **trend** - общее направление, тенденция) **-** явно выраженное движение стоимости финансового инструмента вверх или вниз.
Основные типы тренда:

- **Бычий тренд (повышающий тренд или восходящий тренд)** *представляет собой последовательность более высоких максимумов или минимумов. При повышающемся тренде движение рынка вверх всегда больше, чем движение вниз* (см. рис. 2).

рис.2

- **Медвежий тренд (нисходящий тренд, bear trend)** – *это такая стадия развития рынка, которая характеризуется непрерывным снижением цен. Является последовательностью более низких минимумов и более низких максимумов, что и вызывает движение рынка вниз.*

В противоположность повышающемуся (бычьему) тренду, медвежий считается не нарушенным до тех пор, пока не будет пробит относительный предыдущий максимум (см. рис. 2):

рис. 3

Пока не будет пробит относительный предыдущий минимум, бычий тренд считается не нарушенным. Пробой свидетельствует о возможном окончании тренда. Однако данное условие нужно рассматривать только как один из нескольких признаков разворота долговременной тенденции.

На графиках медвежий и бычий тренды выделяют, как правило, прямыми линиями. Линия медвежьего тренда соединяет последовательные максимумы, линия бычьего – последовательные минимумы. Данные линии являются линиями сопротивления и поддержки. Параллельные линии, которые с обеих сторон ограничивают тренд, носят название ценовых коридоров (трендовых каналов).

1.2 Постановка задачи.

Наша основная задача - определить какой дальше будет тренд: возрастающий, убывающий или плато.

Получаем на вход временной ряд котировок. Поскольку его график не является гладким и определить экстремумы не представляется возможным, сделаем сглаживание. После чего разобьем его на отрезки от i-ой точки экстремума до i+1. Набор таких отрезков будем называть выборкой. Разметим нашу выборку в зависимости от тренда, множество меток: {1, -1, 0}. Теперь наша задача - проанализировать полученную выборку и выбрать наиболее оптимальные параметры для алгоритма.

1.3 Вывод.

В данной главе были рассмотрены основные определения, связанные с биржевым рынком, а так же неформально поставлена задача.

2. Глава 2. Математическая постановка задачи, методы ее решения.

2.1 Математическая постановка задачи классификации.

Напомним определение задачи классификации. Пусть **X** - множество описаний объектов, Y - конечное множество меток классов, и существует неизвестная целевая зависимость - отображение y* : X→Y, значения которой известны только на конечном подмножестве объектов $x_1, \ldots, x_n \in$ X. Пары объектов и ответов (i, y_i) называют прецендентами. Совокупность таких пар называют обучающей выборкой.

Задача обучения по прецендентам заключается в восстановлении зависимости y* по заданной обучающей выборке, т.е. в построении решающей функции X→Y, которая бы приближала целевую функцию y*(x) причем не только на объектах обучающей выборки, но и на всем множестве X. Кроме того, решающая функция должна допускать эффективную реализацию на вычислительной системе.

Каждый объект x_i задается измерениями своих характеристик $f_j(x_i)$, которые называют признаками. Таким образом, признаковое описание задается набором функций f_j : X→ D_{f_j} - множество допустимых значений признака. Выделяют несколько типов признаков в зависимости от множества D_f :

- бинарный - D_f={0,1}
- номинальный - D_f - конечное множество
- порядковый - D_f - конечное упорядоченное множество
- количественный - D_f - множество действительных чисел

Пусть признаков m штук, тогда признаковым описанием каждого объекта $x_i \in$ X служит вектор $\left(i^1, ..., x_i^m\right)$. Таким образом обучающая выборка представляется в виде совокупности матрицы X $\in R^{n \times m}$ и вектора Y $\in \{1, ..., k\}^k$, где k - количество классов.

2.2 Математическая постановка задачи классификации типа тренда по характеристикам японских свечей.

Основной целью данной работы является прогнозирование будущего тренда котировки. Как было показано в предыдущем разделе, входными параметрами задачи классификации является признаковое пространство объектов, множество меток классов и обучающая выборка, составленная из пар элементов данных двух множеств. В качестве объектов рассматриваются временные ряды котировок, заданные минимальной, максимальной, входной и выходящей ценой. Множество меток классов - {0, 1, -1}.

2.3 Возможные методы решений.

В данном разделе рассмотрим некоторые возможные методы решений поставленной задачи.

Первый способ (при помощи матрицы объект-признак):

- Сглаживаем нашу исходную функцию, например при помощи скользящего среднего.
- Задаем параметр —длина одного элемента (т.е. количество тиков) и в соответствии с ним разбиваем нашу выборку.
- Составляем матрицу объект-признак для каждого элемента, где объект —интервал времени наблюдения, а признаки —max, min, out, in (и, возможно, их определенные соотношения, характерные для японских свечей).
- Обучение при помощи линейной регрессии или метода ближайших соседей (k-NN)

Второй способ:

- При помощи скользящего среднего сглаживаем нашу исходную функцию.
- Выявляем точки перегиба и разбиваем выборку на несколько частей, где каждая из них—временной промежуток от k-ой точки перегиба до k+1. Делим нашу выборку

(например при помощи функции split) в определенном соотношении на 2 части: train и test.

- Определяем метрику в терминах японских свечей.
- Обучение при помощи алгоритма кластеризации - метода ближайших соседей (k-NN).

2.4 Вывод.

В данной главе была рассмотрена общая постановка задачи классификации и ее частный случай - классификация будущего тренда. А также возможные методы решения данной задачи. В данной работе будет использован второй способ.

3. Глава 3. Программная реализация, результаты вычислительных экспериментов.

3.1. Программная реализация.

3.1.1. Обработка и анализ данных.

Для написания программы и дальнейших расчетов будет использовать язык программирования python 3.6, среду разработки - Anaconda, и несколько стандартных библиотек: pandas, numpy, sklearn, math.

Для начала скачаем данные с сайта http://stocks.investfunds.ru в формате xls. Функция pandas.read_exsel позволяет открыть и работать с ним, итерируясь по столбцам или строкам. Нарисуем график нашей выборки по закрытой цене (см. рис. 4).

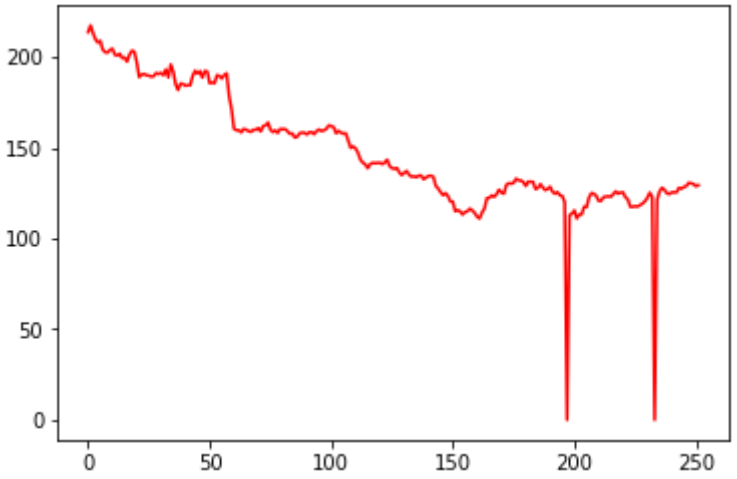


рис. 4 График котировок

Заметим, что у нас есть два выброса. Их надо убрать, поскольку иначе в дальнейшем они могут помешать обучению или тестированию. После удаления строк, у которых столбец «Close price»=0 получили следующий график:

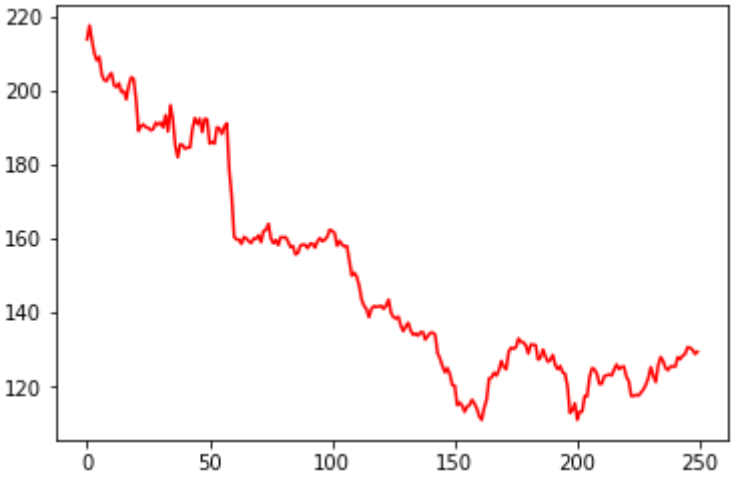


рис.5 График котировок после удаления выбросов

Далее при помощи библиотечной функции скользящего среднего (pandas.rolling_mean(input,window), где input — массив входящих данных,

a window — окно усреднения), сделаем сглаживание нашего графика. Изменяя параметр window можно получить большую или меньшую степень сглаживания данных (см. рис. 6):

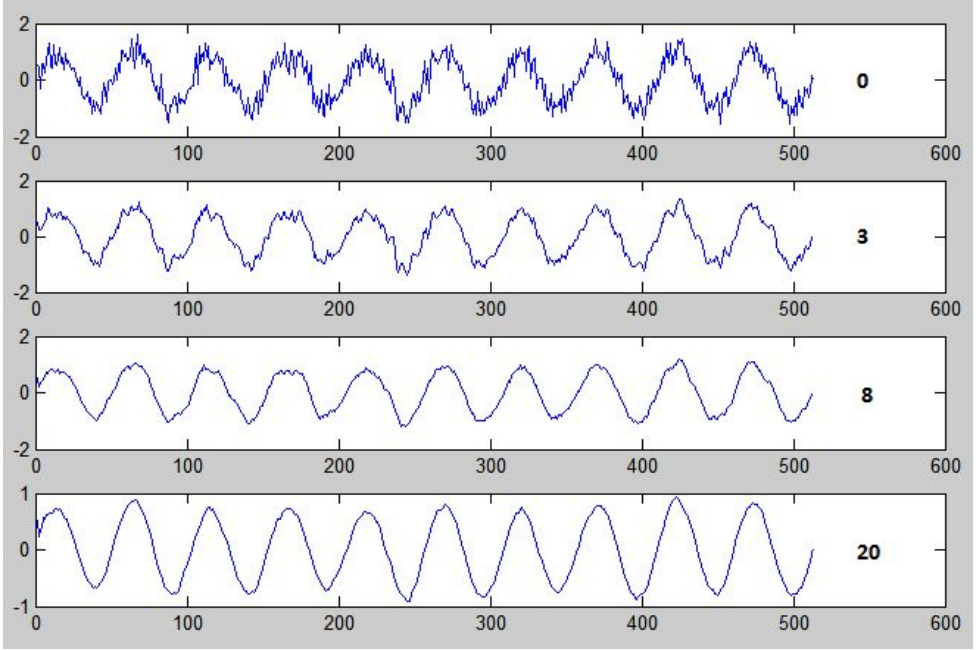


рис.6 Разные степени сглаживания на одних и тех же данных.

Применим скользящее среднее с параметром window равным 3,4 и сравним получившиеся графики (см. рис. 7 и рис. 8 соответственно).

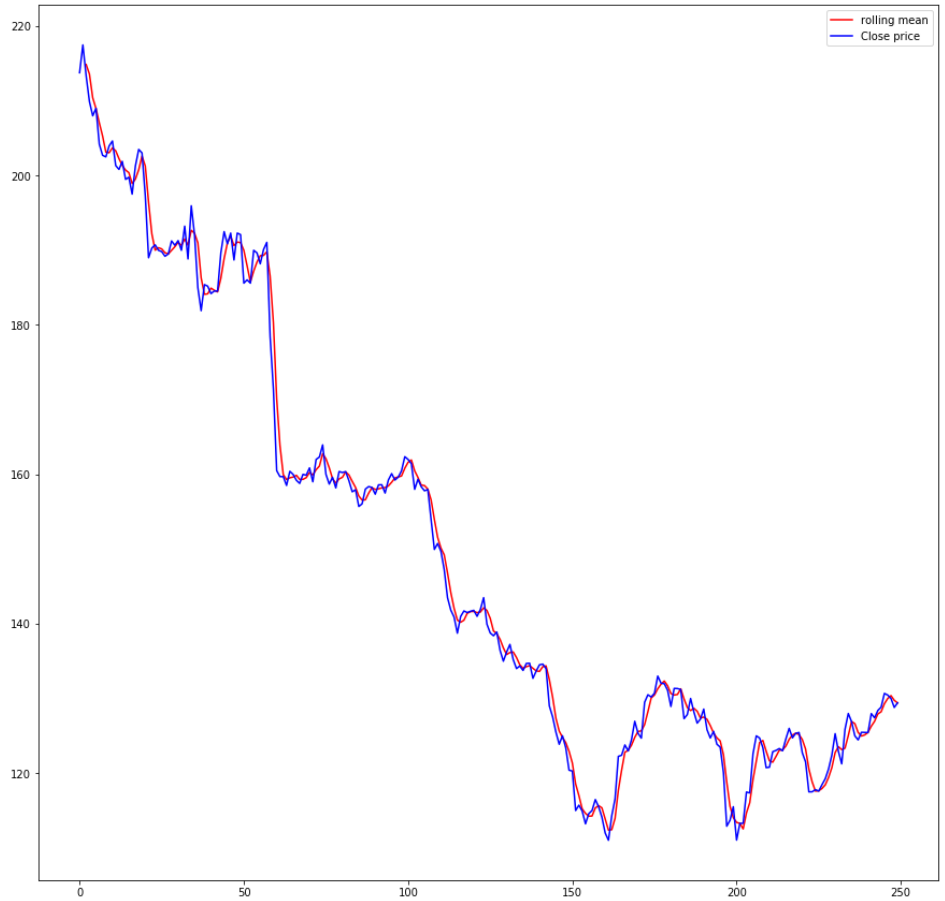


рис.7.Скользящее среднее со степенью сглаживания 3. Синия линия - исходный график, красная - после сглаживания.

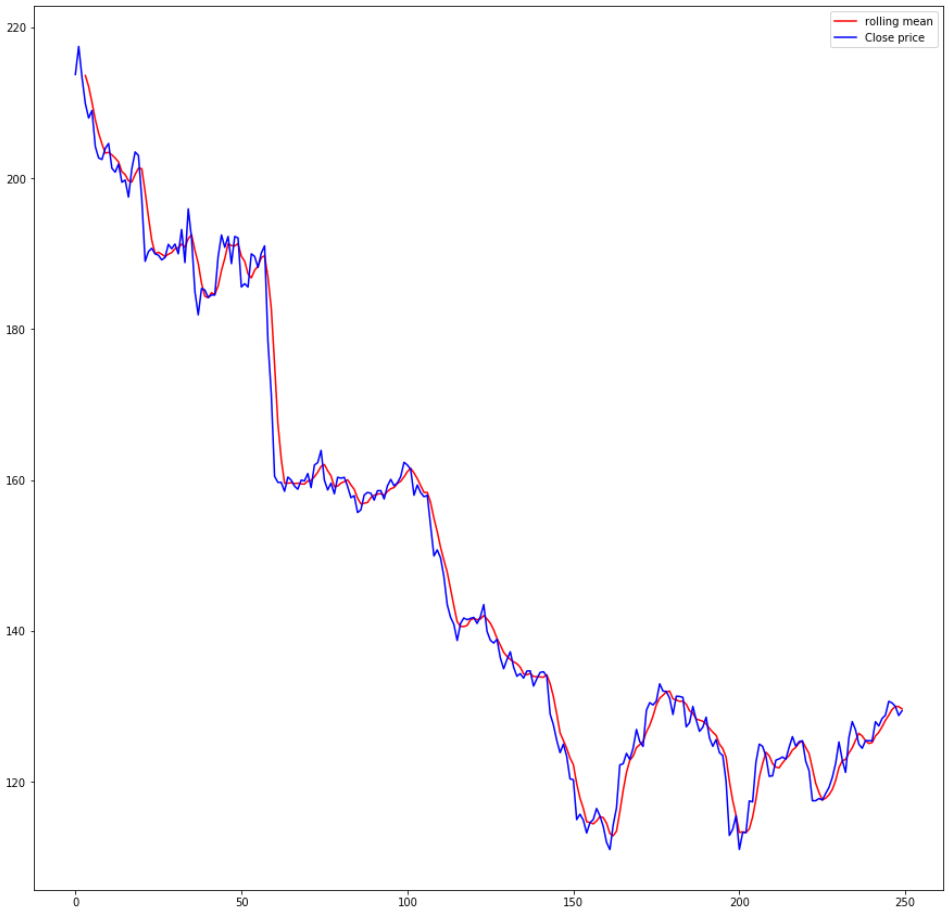


рис.8 Скользящее среднее со степенью сглаживания 4. Синия линия - исходный график, красная - после сглаживания.

Из графиков видно, что при степени сглаживания равной 4 учитываются только значимые скачки, а значит для нашей модели будем использовать параметр window=4, поскольку при больших значениях будет большая погрешность.

Сделаем разбиение. Наша выборка будет состоять из отрезков от i-ой точки перегиба до i+1 точки перегиба. Но не смотря на сглаживание при помощи скользящего среднего, наша функция может иметь локальные экстремумы (и в нашем случае такие тоже есть). Поэтому наша задача состоит еще и в том, чтобы не принять данные точки за точки перегиба.

Разметим полученный интервалы: 1 - тренд возрастает, -1 - убывает, 0 - плато. Определим, какие интервалы мы будем размечать как 0. Так как у некоторых акций колебания незначительны, а у кого-то наоборот очень резкие, а так в зависимости от выбора тика дельта колебаний может меняться. В связи с чем необходимо для каждой выборки отдельно выбирать тот самый порог, до которого мы считаем что на данном интервале тренд остается неизменным. Определим eps, тогда интервал размечаем как 0, если | max(close_price) - min(close_price) | < eps, где close_price - множество всех цен закрытия, присутствующих в данном интервале.

3.1.1. Выбор метрики, кластеризация.

Для начала определимся с метрикой. Введём следующие обозначения: «In» - цена открытия, «Out» - цена закрытия, «Min» - минимальная цена, «Max» - максимальная цена. Все параметры берутся за определённый тик, в нашем случае — за день. В качестве признаков возьмём следующие соотношения из японских свечей: (Out — In) — размер тела свечи, умноженный на знак; (Max — Min) — длина свечи; sign(Out — In) — цвет

тела; и некоторые соотношения: (Max / Min), (Out / In) . Взяв такие признаки, мы сразу получаем нормализованную выборку.

Заметим, что элементы нашей выборки могут иметь разные длинные, в таком случае есть 2 подхода к решению данной проблемы:

1) Накладывать друг на друга 2 интервала и “обрезать” более длинную. Таким образом мы сравниваем только начала наших элементов.
2) Смотреть “окном”. Пусть у нас есть 2 элемента нашей выборки x = $(x_1, ..., x_n)$ и y = $(y_1, ..., y_m)$, причем m > n, тогда расстоянием между x и y будем считать min из расстояний между x и $(y_i, ..., y_{i+n})$, где i+n <= m.

За основу возьмем евклидово расстояние. В итоге имеем: x_i

$$d(x,'y') = \sqrt{\sum_{i=0}^{m \in (len(x),len(y))} (x_i - y_i)^2}$$

Далее воспользуемся методом кластеризации - к ближайших соседей. Метод К ближайших соседей - один из наиболее простых алгоритмов классификации, относящийся к группе структурных методов.

В качестве обучающей выборки используется набор объектов, каждый из которых принадлежит к одному из двух или более классов. Каждый объект может быть представлен точкой в n-мерном пространстве, где n – число аналитических признаков, используемых для классификации.

Неизвестный объект относится к одному из классов по следующему принципу: находится K ближайших объектов из обучающей выборки в пространстве образов. Затем определяется, к какому классу принадлежит большинство ближайших объектов обучающей выборки – к этому классу относится и неизвестный объект. Оптимальное число K, как правило, подбирают экспериментальным путем. Увеличение K приводит к уменьшению влияния случайных погрешностей в данных, но при этом разделение на классы становится менее четким.

3.2. Результаты эксперимента.

3.2.1. При обрезании элемента наибольшей длины при нахождении расстояния между интервалами.

Далее везде будем рассматривать выборку из акций за 3 года: с 30.04.2015 по 30.04.2018. В данном алгоритме у нас есть несколько параметров: к - количество ближайших соседей, p - степень сглаживания для скользящего среднего, min — минимальный размер элемента из выборки, eps - параметр для разметки (см. пункт 3.3.1). Возьмем eps = 0.001. Проведём подборку оптимальных параметров для акций Сбербанка (см. таблицу 1). Подборка проводит методом кросс валидации.

Степень сглаживания	**Количество ближайших соседей**	**Минимальный размер элемента**	**Процент правильных ответов**
2	3	2	0,595959595959596
2	3	3	0,529411764705882
2	3	4	0,581081081081081
2	3	5	0,6
2	4	2	0,595959595959596

2	4	3	0,521008403361345
2	4	4	0,621621621621622
2	4	5	0,666666666666667
2	5	2	0,611111111111111
2	5	3	0,521008403361345
2	5	4	0,554054054054054
2	5	5	0,622222222222222
2	6	2	0,580808080808081
2	6	3	0,53781512605042
2	6	4	0,554054054054054
2	6	5	0,622222222222222
2	7	2	0,626262626262626
2	7	3	0,546218487394958
2	7	4	0,5
2	7	5	0,555555555555556
2	8	2	0,621212121212121
2	8	3	0,53781512605042
2	8	4	0,540540540540541
2	8	5	0,666666666666667
3	3	2	0,551282051282051
3	3	3	0,653543307086614
3	3	4	0,585365853658537
3	3	5	0,528301886792453
3	4	2	0,564102564102564
3	4	3	0,622047244094488
3	4	4	0,634146341463415
3	4	5	0,622641509433962
3	5	2	0,596153846153846
3	5	3	0,653543307086614
3	5	4	0,597560975609756
3	5	5	0,584905660377358
3	6	2	0,608974358974359
3	6	3	0,677165354330709
3	6	4	0,609756097560976
3	6	5	0,622641509433962
3	7	2	0,628205128205128
3	7	3	0,677165354330709
3	7	4	0,585365853658537

3	7	5	0,622641509433962
3	8	2	0,634615384615385
3	8	3	0,653543307086614
3	8	4	0,560975609756098
3	8	5	0,641509433962264
4	3	2	0,454545454545454
4	3	3	0,564814814814815
4	3	4	0,576086956521739
4	3	5	0,575757575757576
4	4	2	0,492424242424242
4	4	3	0,574074074074074
4	4	4	0,608695652173913
4	4	5	0,621212121212121
4	5	2	0,5
4	5	3	0,601851851851852
4	5	4	0,619565217391304
4	5	5	0,636363636363636
4	6	2	0,5
4	6	3	0,592592592592593
4	6	4	0,608695652173913
4	6	5	0,636363636363636
4	7	2	0,522727272727273
4	7	3	0,62962962962963
4	7	4	0,641304347826087
4	7	5	0,621212121212121
4	8	2	0,507575757575758
4	8	3	0,592592592592593
4	8	4	0,608695652173913
4	8	5	0,606060606060606

таблица 1.

Заметим, что при увеличении минимального размера почти всегда результаты улучшаются. Также видно, что лучше всего себя показала степень сглаживания – 3. Лучшее количество ближайших соседей сильно зависит от того, насколько меньше или больше у нас получилась выборка. В итоге лучший результат 67,7% был достигнут при следующих параметрах: 3,7,3.

Так же такие подборки были проведены на ряде акций других российских компаний.

Название компании	**Степень сглаживания**	**Количество ближайших соседей**	**Минимальный размер элемента**	**Процент правильных ответов**
Газпром	2	6	5	73.33
Ростелеком	3	5	4	65.38
Татнефть	2	7	2	65.4

Таблица 2.

Посмотрим другие метрики для лучших результатов полученных для предыдущих кампаний, которые широко распространены для оценивания подобных моделей:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y^{pred}{}_i - y^{true}{}_i)^2$$

$$\text{NMSE} = \frac{MSE}{\frac{1}{n}\sum_{i=1}^{n}(y^{pred}{}_i - \bar{y})^2}$$, где $\bar{y}$ - среднее из y_i^{pred}

$$\text{RMSE} = \sqrt{MSE}$$

$$\text{MAE} = \frac{1}{n} * \sum_{i=1}^{n}|y^{pred}{}_i - y_i^{true}|$$

Название компании	**MSE**	**NMSE**	**RMSE**	**MAE**	**Процент правильных ответов**
Газпром	1.06	1.08	1.03	0.53	73.33
Ростелеком	1.38	1.38	1.17	0.69	65.38
Татнефть	1.36	1.38	1.17	0.68	65.4
Сбербанк	1.29	1.38	1.13	0.64	67.71

Итого видим явную зависимость между процентом правильных ответов и остальными метриками: чем больше процент, тем лучше остальные метрики.

3.2.2. При просмотре “окном” для нахождения расстояния между интервалами.

Посмотрим как изменились результаты на тех же данных (как в пункте 3.2.1) при добавлении нового функционала в виде просмотра окном расстояния между интервалами.

Название компании	**MSE**	**NMSE**	**RMSE**	**MAE**	**Процент правильных ответов**
Газпром	1.19	1.22	1.09	0.59	70.15 (-3%)
Ростелеком	1.21	1.23	1.10	0.60	69.56 (+4%)
Татнефть	1.32	1.33	1.15	0.66	66.48 (+1%)
Сбербанк	1.23	1.28	1.09	0.61	69.23 (+1.5%)

В итоге “просмотр окном” в среднем дает +1% правильных ответов, выравниваются остальные метрики, итого мы получаем независимо от данных достаточно близкие друг к другу результаты. Но если мы преследуем цель учитывать особенности каждой из выборок, а не предсказывать хорошо в среднем на всем, то для таких примеров как “Газпром” лучше не использовать этот метод.

3.3. Вывод.

В данной главе была подробно рассмотрена реализация метода, изложенного ранее. В итоге мы видим, что даже простейшая реализация работает достаточно неплохо, а значит можно будет еще и улучшить ее.

4. Инструкция по использованию написанной программы.

Ссылка на код: https://github.com/Mashka8512/Finance-s_pain .

Название класса: Learner

Параметры:

- k_neighbors – параметр алгоритм KNN, количество ближайших соседей.
- rolling_mean_window – степень сглаживание исходного временного ряда.
- min_size – минимальный размер элемента в выборке.
- eps – предел для определения состояния «плато».
- metrics:
    - 'euclidean' - евклидово расстояние
    - ‘euc_distance_with_window’ - евклидово расстояние с учетом окна

Имеются следующие функции:

- prepare_data(X) - где X является объектом типа pandas.table. Она преобразовывает исходный временной ряд в независимые элементы выборки. Возвращает data, target (list object).
- fit_predict(X_train, y_train, X_test) – все параметры являются numpy array. Так как KNN алгоритм не требует предварительного обучения, то мы можем сразу сделать предсказание. Возвращает numpy array предсказанных значений.
- KNN_cross_validation(X, y) - где X, y - numpy array or list преобразованных данных. Возвращает (количество верных предсказаний) / (общее количество элементов).

5. Заключение.

В данной работе была поставлена и решена с вероятностью в среднем 0.69 задача классификации будущего тренда временного ряда котировок. Рассмотрим, какие еще изменения можно добавить в реализацию, чтобы повысить точность прогноза:

- Фильтр. В данной реализации мы считали выбросами только те элементы, у которых цена закрытия была равна 0. Но ведь может случится и такое, что цена резка упала или возросла, а до и после скачка была монотонной. Такие элементы тоже надо считать выбросами.
- Отказ от прогнозирования. Данная реализация выдает ответ в любом случае, даже если наша точка расположена достаточно далеко от всех остальных в обучающей выборке. В связи с чем, надо ввести "радиус поиска", и если мы выходим за него - то делать отказ.
- Библиотеки под GeForce. Здесь мы использовали стандартные библиотеки языка python. Но так же имеются специально адаптированные под GeForce, и скорость обучения на них должна возрасти.
- Нейронные сети. Обучить нейронную сеть нашему алгоритму.

Приложение.

1. **Сравнение результатов при разных тиках и степенях сглаживания на данных Яндекса.**

Результаты по задаче прогнозирование тренда временного ряда котировок для акций «Яндекс» за 1 год с 2018-10-20 по 2019-10-10.

Метрики:

Statistics = $\frac{\text{количество правильных ответов}}{\text{общее число ответов}}$

MSE = $\frac{1}{n}\sum_{i=1}^{n}(y^{pred}{}_i - y^{true}{}_i)^2$

NMSE = $\frac{MSE}{\frac{1}{n}\sum_{i=1}^{n}(y^{pred}{}_i - \bar{y})^2}$, где $\bar{y}$ - среднее из y_i^{pred}

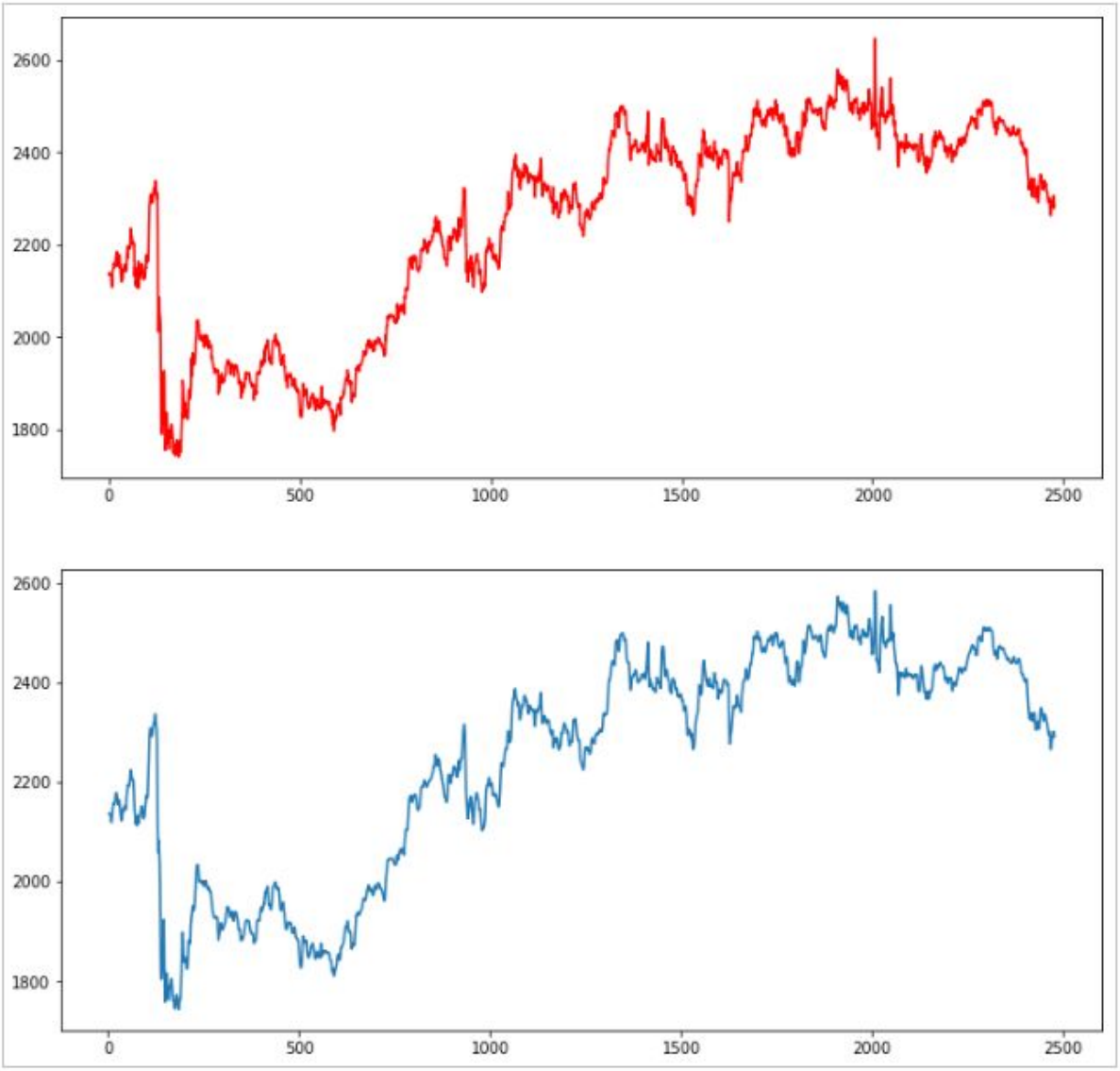
RMSE = $\sqrt{MSE}$

MAE = $\frac{1}{n} * \sum_{i=1}^{n}|y^{pred}{}_i - y_i^{true}|$

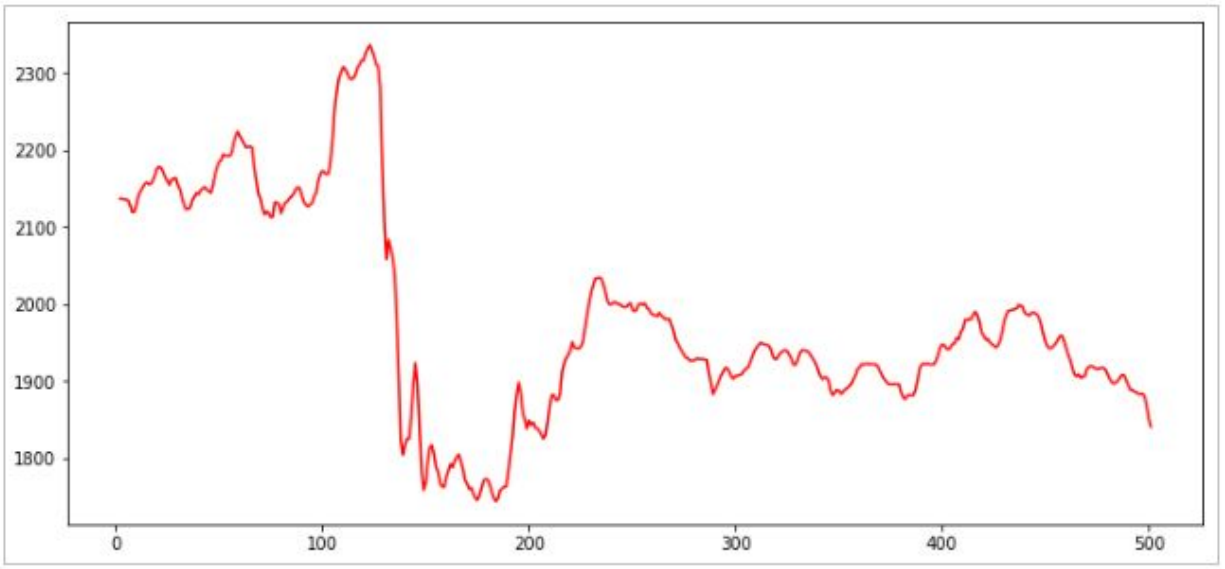
1) Тик = час.

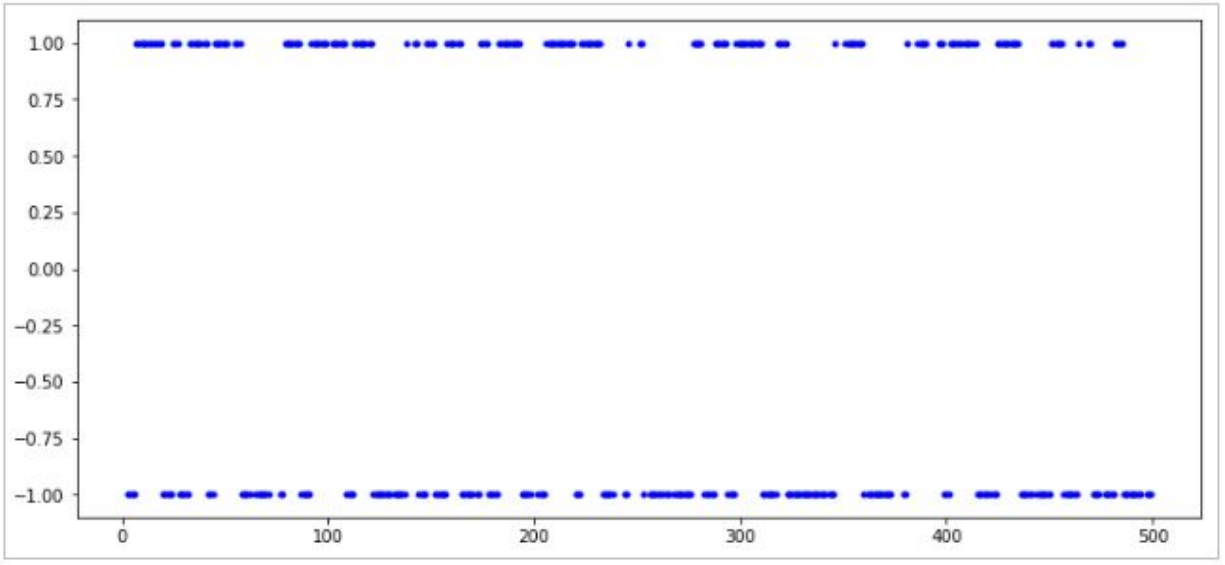
- **Rolling mean window = 3 (степень сглжаивания)**

Верхний график — исходные данные, именно цена закрытия.
Нижний график — сглаженные исходные данные.



Верхний график — сглаженные исходные данные.
Нижний график — метки каждого из интервалов на графике выше.



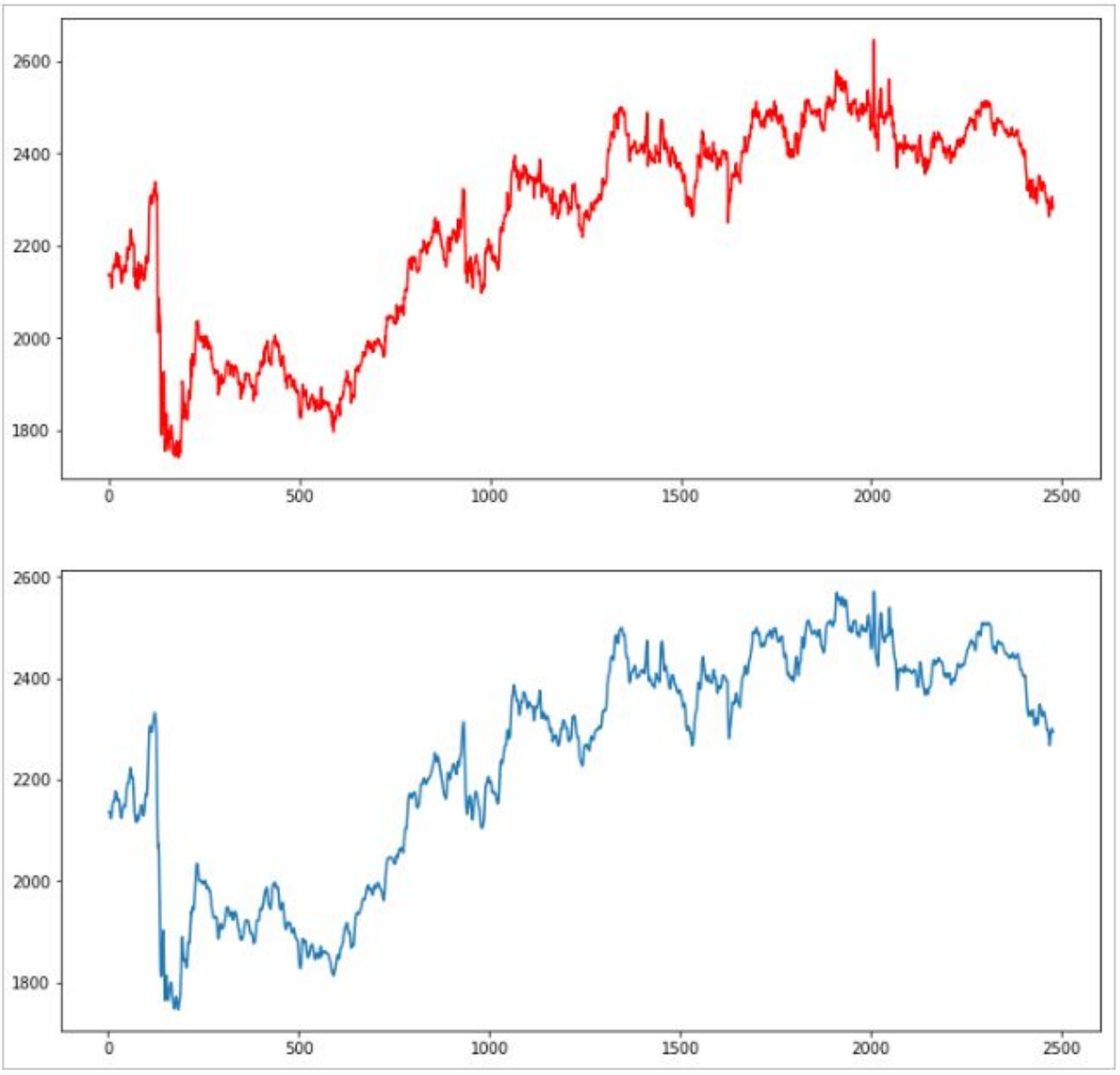


Лучшие результаты без «просмотра окном»:

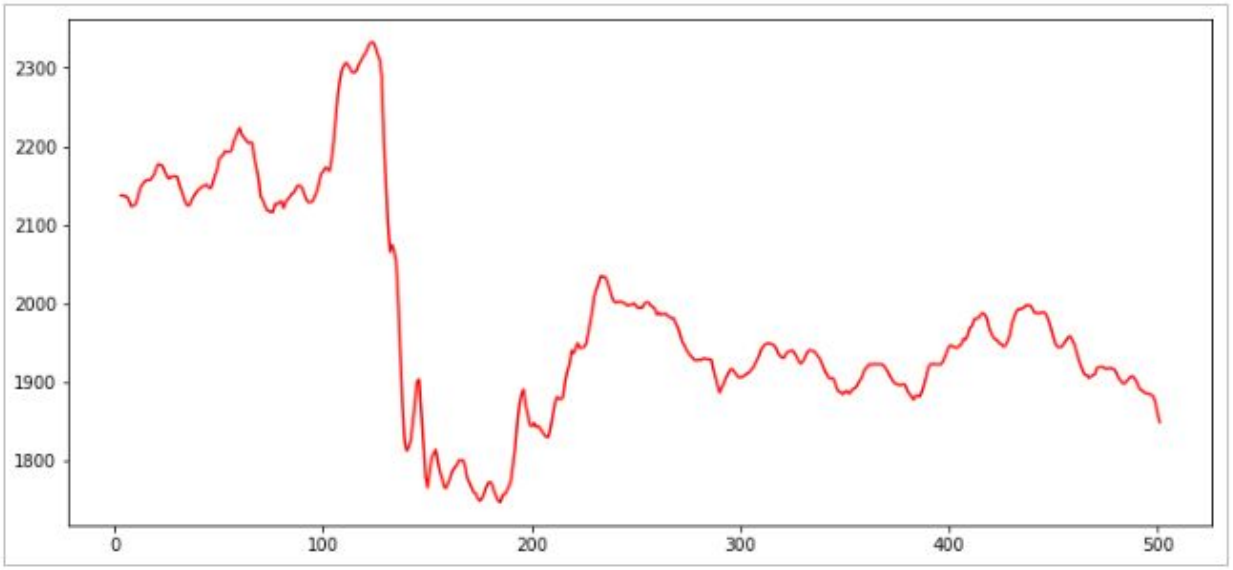
rolling mean window	k neighbors	min size	statistics	MSE	NMSE	RMSE	MAE
3	8	2	0,787321	0,85	0,85	0,922343	0,425358
3	7	2	0,783231	0,86	0,86	0,931169	0,433538
3	6	2	0,762781	0,94	0,97	0,974102	0,474438

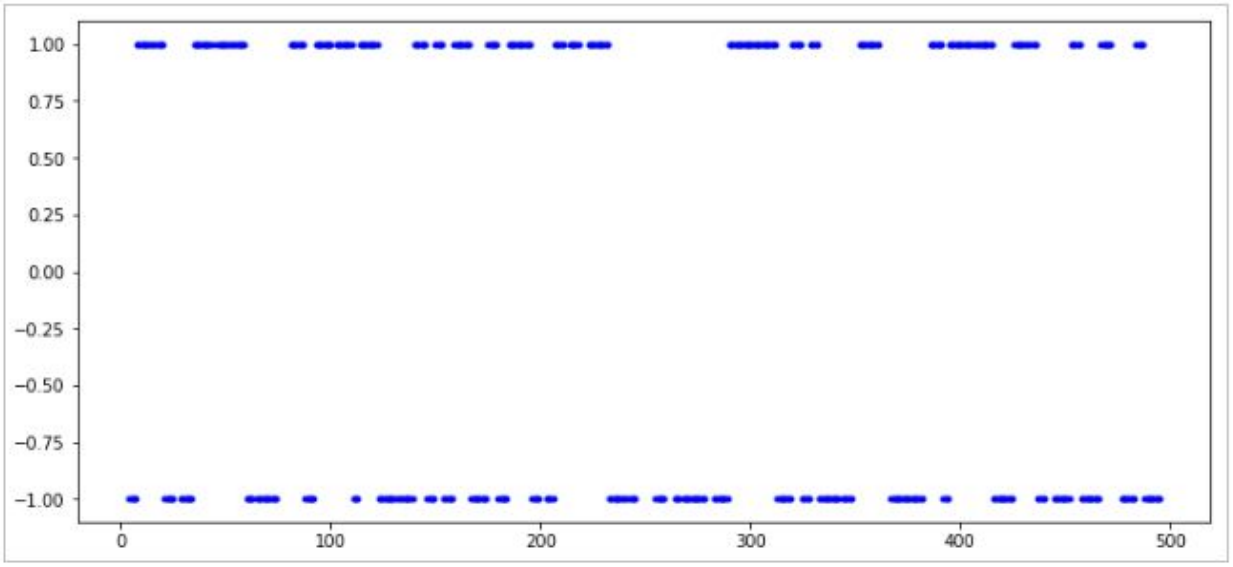
- **Rolling mean window = 4**

Верхний график — исходные данные, именно цена закрытия.
Нижний график — сглаженные исходные данные.



Верхний график — сглаженные исходные данные.
Нижний график — метки каждого из интервалов на графике выше.



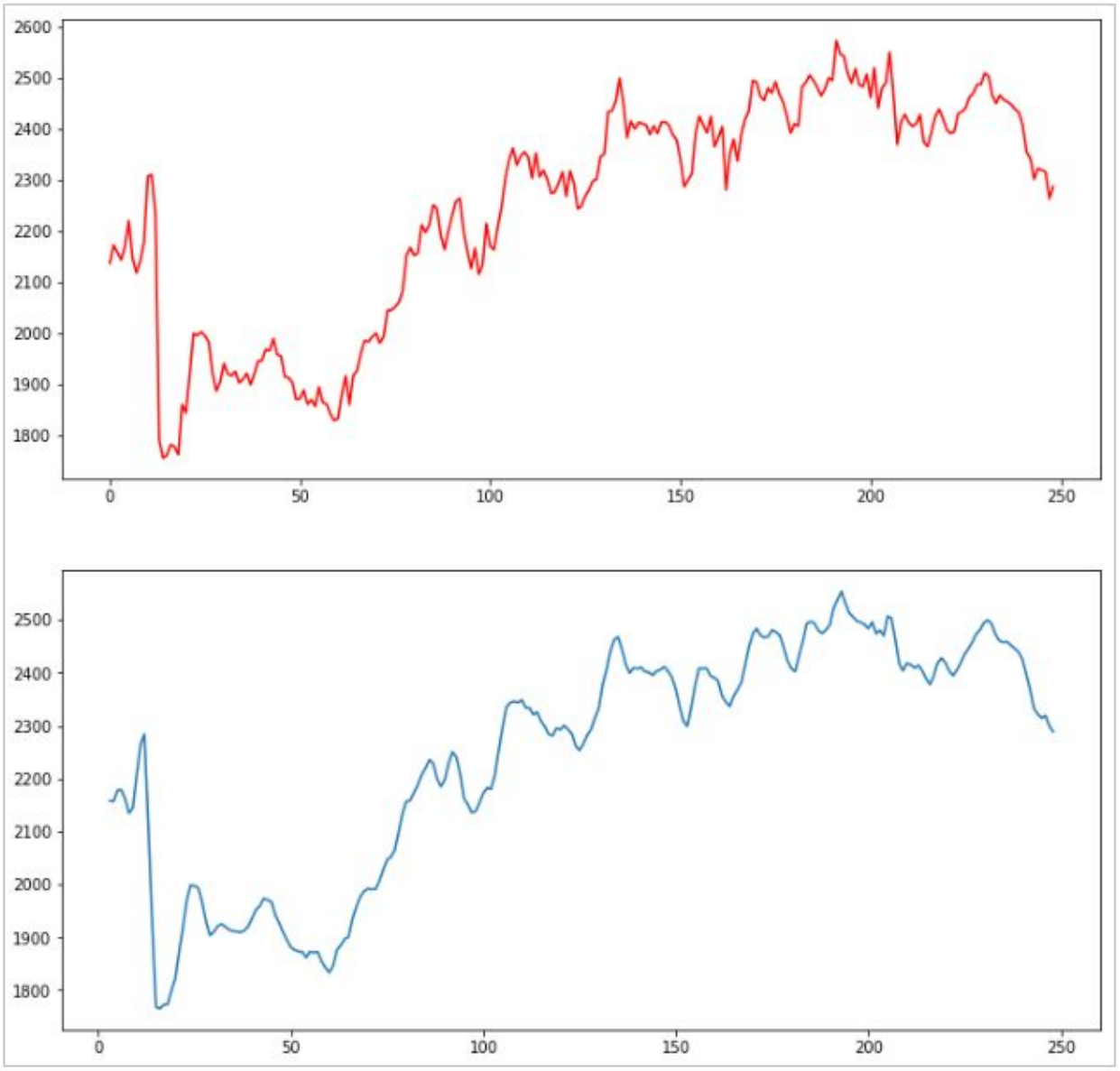


Лучшие результаты без «просмотра окном»:

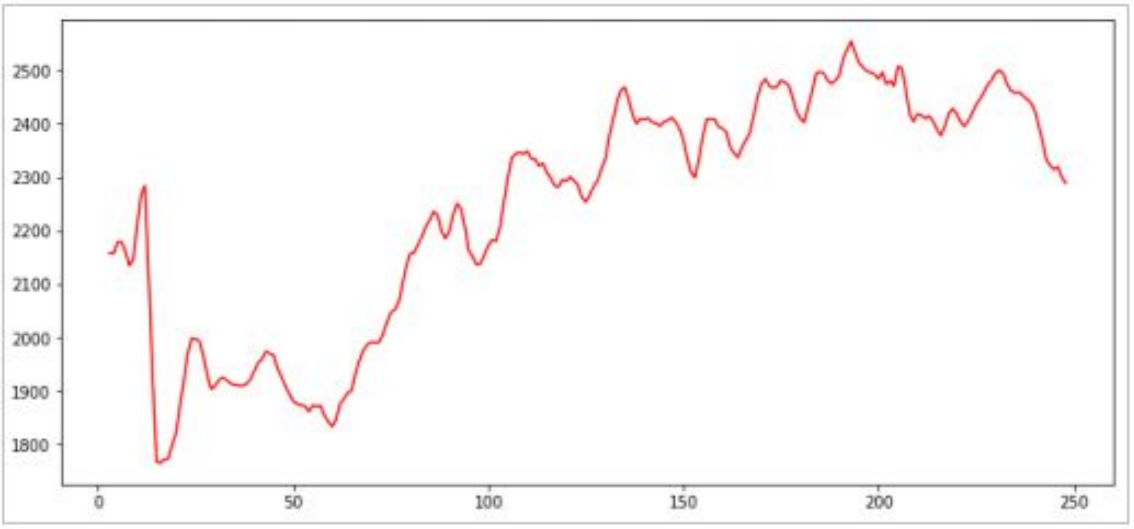
rolling mean window	k neighbors	min size	statistics	MSE	NMSE	RMSE	MAE
4	8	3	0,676301	1,29	1,33	1,137892	0,647399
4	5	3	0,67341	1,3	1,3	1,14296	0,653179
4	3	3	0,66474	1,34	1,34	1,158033	0,67052

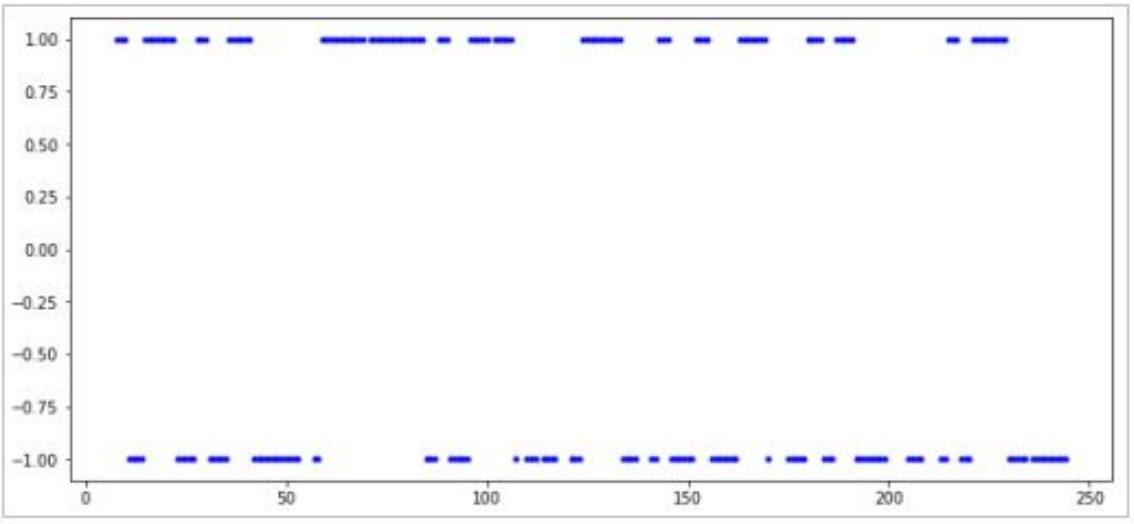
2) Тик = день.
Rolling mean window = 3

Верхний график — исходные данные, именно цена закрытия.
Нижний график — сглаженные исходные данные.



Верхний график — сглаженные исходные данные.
Нижний график — метки каждого из интервалов на графике выше.



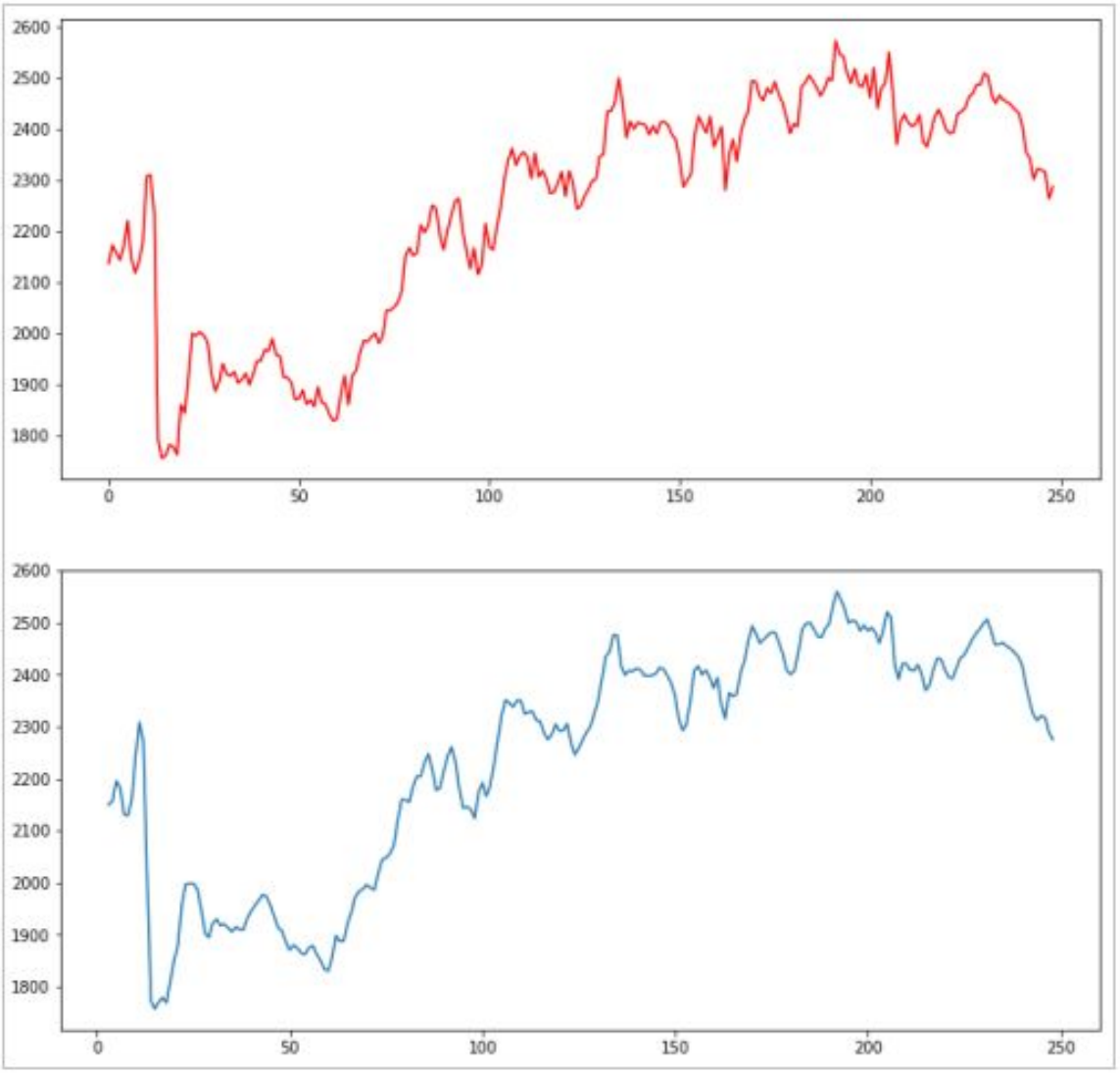


Лучшие результаты:

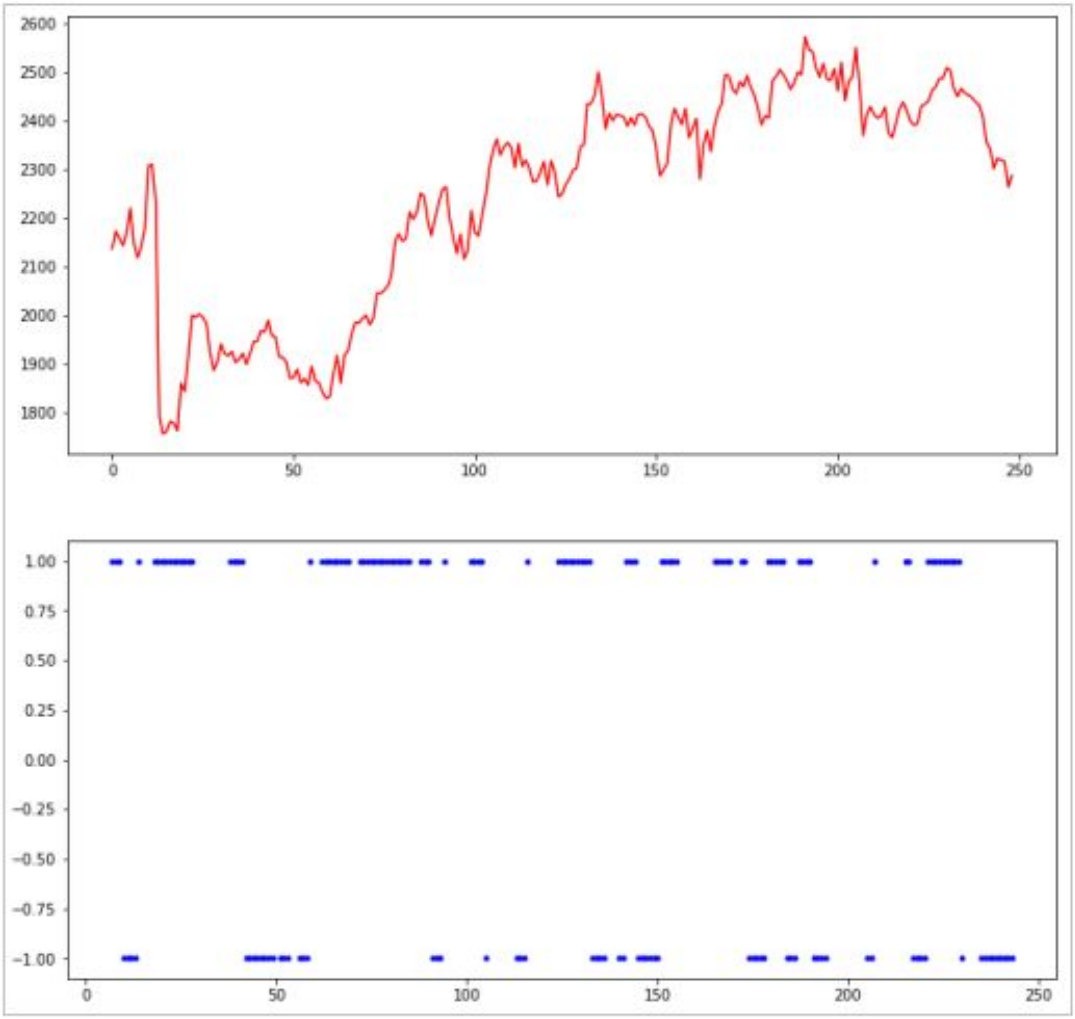
rolling mean window	k neighbors	min size	statistics	MSE	NMSE	RMSE	MAE
3	4	4	0,666667	1,33	3,37	1,154701	0,666667
3	4	5	0,666667	1,33	3,37	1,154701	0,666667
3	5	5	0,666667	1,33	1,92	1,154701	0,666667
3	7	3	0,65	1,4	1,49	1,183216	0,7
3	4	2	0,644444	1,42	2,22	1,19257	0,711111

Rolling mean window = 2

Верхний график — исходные данные, именно цена закрытия.
Нижний график — сглаженные исходные данные.



Верхний график — сглаженные исходные данные.
Нижний график — метки каждого из интервалов на графике выше.



Лучшие результаты:

rolling mean window	k neighbors	min size	statistics	MSE	NMSE	RMSE	MAE
2	3	4	0,619048	1,52	1,86	1,234427	0,761905
2	4	2	0,580645	1,67	1,97	1,295152	0,83871
2	8	3	0,575758	1,69	2,53	1,302678	0,848485
2	7	5	0,571429	1,71	6,46	1,309307	0,857143
2	7	4	0,571429	1,71	3,5	1,309307	0,857143
2	5	4	0,571429	1,71	3,5	1,309307	0,857143

6. Литература.

https://www.ifcmarkets.ru/ntx-indicators/dow-theory

https://habrahabr.ru/post/312450/

https://habrahabr.ru/post/313216/

Нисон Стив. Японские свечи: графический анализ финансовых рынков.

Дэви Силен, Арно Мейсман, Мохамед Али «Основы Data Science и Big Data»

http://www.mbureau.ru/articles/dissertaciya-model-prognozirovaniya-vremennyh-ryadov-glava-1

http://statsoft.ru/home/textbook/modules/sttimser.html#spectrum

https://habrahabr.ru/post/134375/

https://habrahabr.ru/company/ods/blog/327242/

Грегори Моррис – Японские свечи. Метод анализа акций и фьючерсов, проверенный временем

Афанасьев, В. Н. Анализ временных рядов и прогнозирование / В.Н. Афанасьев, М.М.

Юзбашев. - М.: Финансы и статистика, Инфра-М, 2010. - 320 с.

Бокс, Дж. Анализ временных рядов прогноз и управление (часть 2) / Дж. Бокс, Г. Дженкинс. - М., **1995**. - **667** с.

Бокс, Дж. Анализ временных рядов прогноз и управление. Выпуск 1 / Дж. Бокс, Г. Дженкинс. - М.: Мир, **2010**. - 408 с.

Бриллинджер, Д. Временные ряды. Обработка данных и теория / Д. Бриллинджер. - М., 1980. - **383** с.

Лукашин, Ю. П. Адаптивные методы краткосрочного прогнозирования временных рядов / Ю.П. Лукашин. - М.: Финансы и статистика, 2003. - 416 с.

James D. Hamilton. Time Series Analysis. Princeton University Press, 1994, 820 стр.

Philip H. Franses & Dick van Dijk. Nonlinear Time Series Models in Empirical Finance. Cambridge University Press, 2000, 296 стр.

Walter Enders. Applied Econometric Time Series. Wiley, 2-е издание, 2004, 460 стр.

Peter J. Brockwell & Richard A. Davis. Time Series Theory and Methods. Springer-Verlag, 2-е издание, 1991, 577 стр.

Christian Gourieroux & Alain Monfort. Time Series and Dynamic Models. Cambridge University Press, 1997, 668 стр.

Philip H. Franses. Time Series Models for Business and Economic Forecasting. Cambridg University Press, 1998, 280 стр.