Отчет о неуспешно проделанной работе над итоговым проектом «Применение методов машинного обучения для раннего выявления эпизодов аномального и/или резкого движения цен на рынках процентных финансовых инструментов развивающихся стран».

Выполнил Егоров А. В.

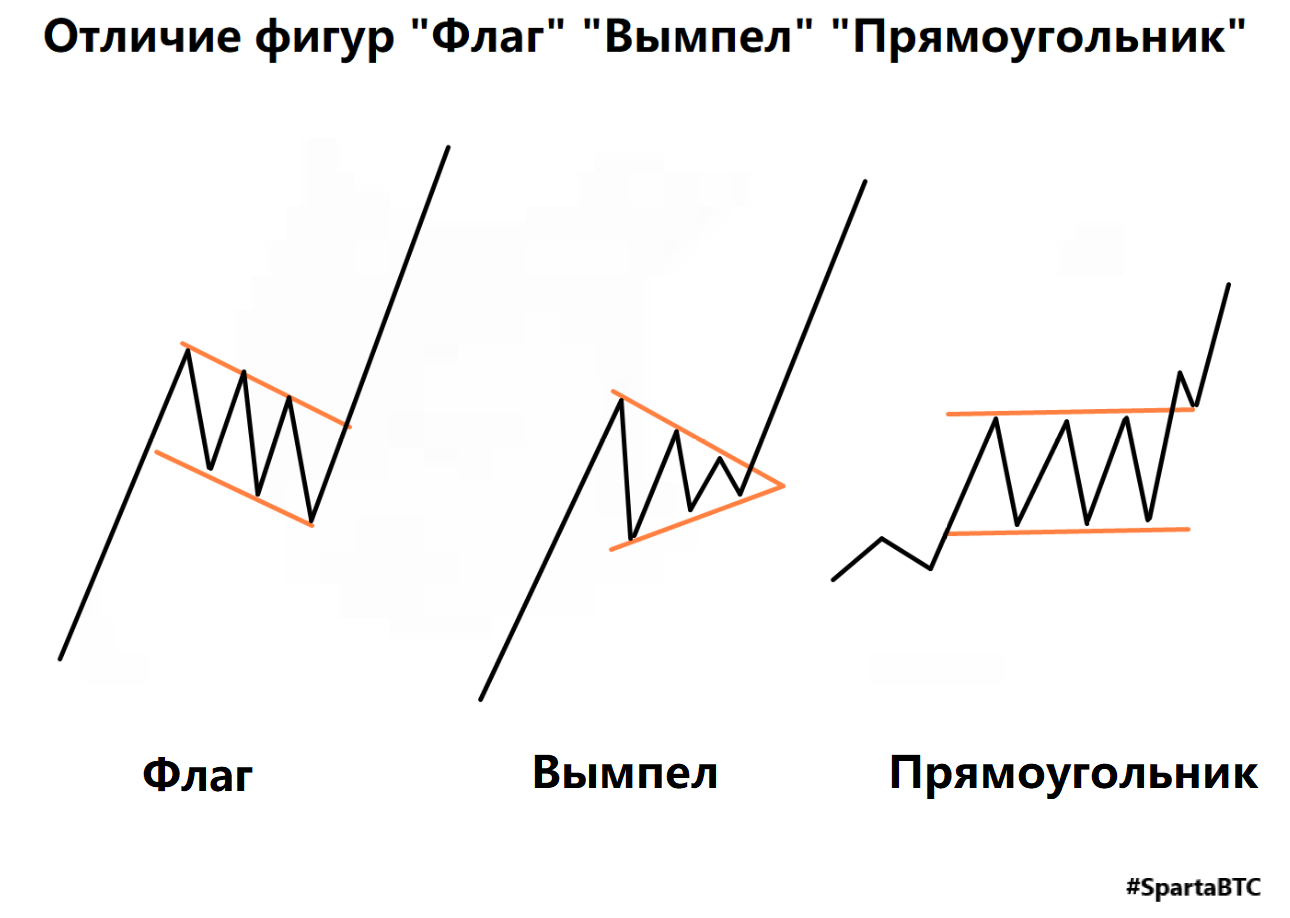
Введение:

В данной работе была сделана неуспешная попытка применить методы машинного обучения для раннего выявления резких движений цены. На рынке производных финансовых инструментов развивающихся стран наибольшую проблему представляют из себя резкие движения ставок наверх (называемые «взрывом» ставок). Решение данной задачи, хотя бы частичное, несло бы в себе значительную ценность: например, инвесторы, оперирующие данными финансовыми инструментами, могли бы заблаговременно сокращать уровень риска своих позиций.

Изначальные проблемы и переформулировка задачи:

При первоначальном взгляде задачу можно сравнить с упреждающим распознаванием поломок оборудования. Однако, если с оборудованием нам как правило доступен для наблюдения исчерпывающий перечень влияющих на его состояние переменных, ситуация на рынке прямо противоположная. Допустим, нам доступна цена инструмента, цены связанных инструментов, общеэкономические показатели и, возможно, новости. Несмотря на то, что при поверхностном рассмотрении начинает казаться, что все ситуации «взрыва» ставок имеют что-то неуловимо общее и описываемое данным набором переменных (что изначально и пробудило интерес к решению задачи), при более детальном изучении каждая ситуация оказалась в значительной степени отлична от всех остальных. Я выгрузил необходимые данные (кроме новостей) по 7 странам (папка related instruments на git), и пришел к выводу, что, по всей видимости, зачастую на ситуации «взрыва» оказывают влияние скрытые переменные, не входящие в данный перечень. Таким образом, у меня сложилось стойкое ощущение, что функции, которая более-менее корректно отображает из пространства матриц, соответствующим протяженным во времени наборам перечисленных переменных, в множество ситуаций {«взрыв», «нет взрыва»}, попросту не существует.

Я решил попробовать найти достаточную аппроксимацию этой функции, и определил ее в виде функции, которая отображала бы векторное пространство протяженной во времени цены самого процентного инструмента (в данном случае, был выбран годовой процентный своп) в множество суждений эксперта о наличии и отсутствии в ближайшем будущем «взрыва» цены (она должна существовать по определению, если эксперт принимает решения не случайным образом). Исходное пространство было выбрано таким образом ввиду того, что эксперты при принятии этих решений могут руководствоваться эвристическими правилами нахождения и интерпретации определенных паттернов на графике цены. В частности, одним из таких так называемых «технических» паттернов, являющимся с некоторой статистической значимостью предвестником резкого роста цены, является фигура «флаг». В стилизованном виде она представлена на рис. 1; в реальных ситуациях она гораздо зашумленнее, чем создает проблемы для hard-coded описания и выявления.



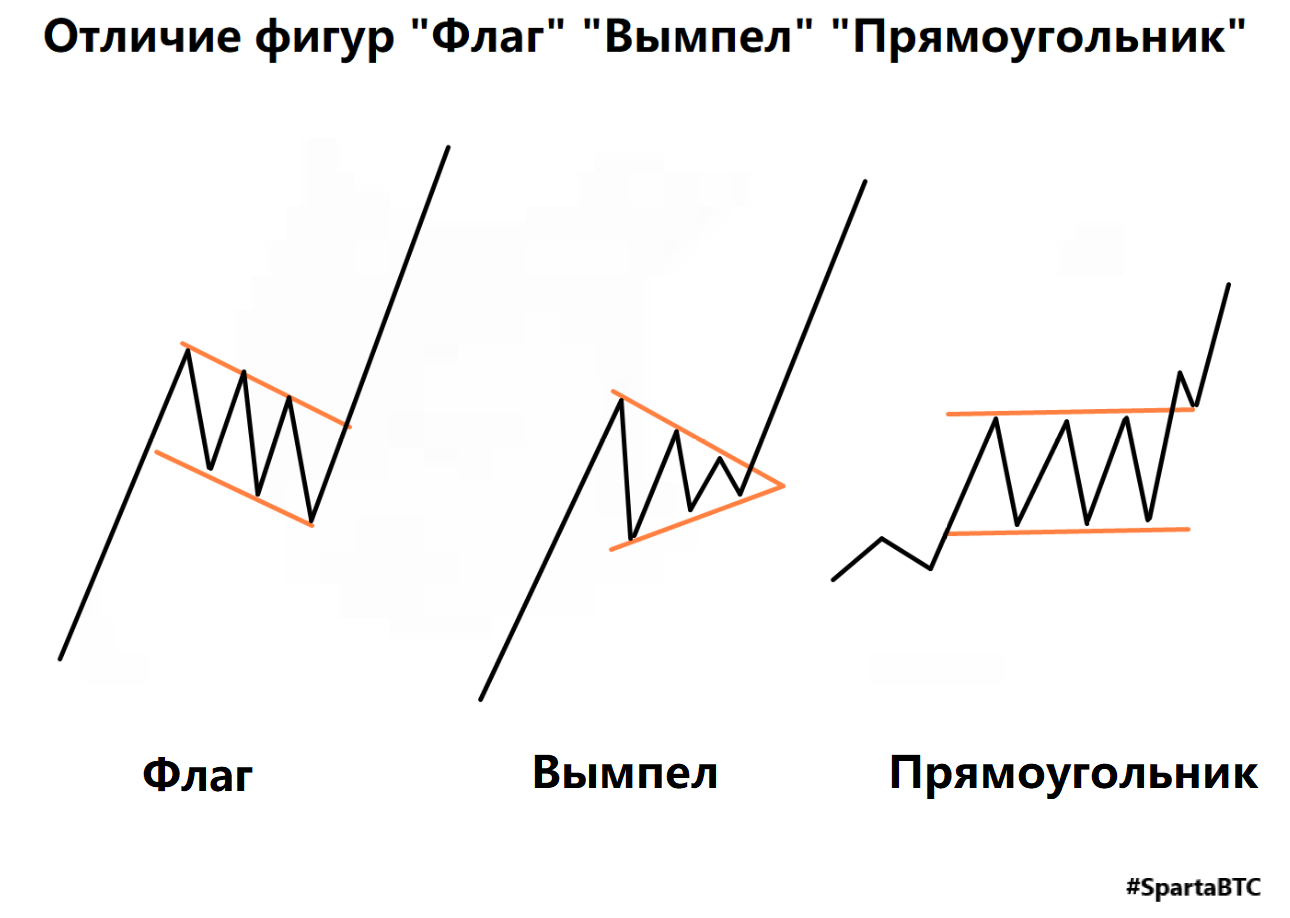


Рис. 1 Рис. 2

Новой задачей таким образом является выявление фигуры «флаг» в «упреждающей фазе», или же в фазе формирования, то есть когда фигуру можно определить с достаточной вероятностью, но второй фазы резкого движения еще не произошло (рис. 2).

Модель:

Схожесть с упреждающим распознаванием поломок задала направление поиска в сторону одномерных сверток. Ознакомившись со статьей[1], я решил свести задачу к еще более узкой, успешно разрешенной проблеме – классификации ударов сердца на нормальные и аномальные на кардиограмме. В качестве основного образца сети мной была взята модель из следующего репозитория: <https://github.com/jeremyruss/ECG-Signal-Classification-by-1D-Convolutional-Neural-Network/blob/master/model.py> При помощи нее автор успешно решал задачу по классификации классического датасета mit\_bih.

Поскольку я обладаю некоторым экспертным опытом, по примеру аннотаций, произведенных кардиологами, я вручную разметил более 4000 рыночных ситуаций (выбранных случайным образом без возвращения) на дневных графиках котировок с окном в 100 торговых дней. Я определял, есть ли данный паттерн (указывающий наверх «флаг» в фазе формирования) в конце графика, а также размечал точку его начала; по аналогии с ударами сердца, а также чтобы не передавать в модель излишний шум, ряд до начала фигуры был занулен. Там, где фигуры не было (таких ситуаций, разумеется, было большинство), в начале ряда было выставлено количество нулей, взятое для каждого объекта случайным образом из распределения количества нулей в тренировочной выборке для класса 1. Последнее потребовало дополнительного преобразования исходных размеченных файлов train.csv и test.csv. Для удобства при передаче в сеть данные разворачиваются (нули остаются в конце); также сделана нормировка в диапазон [0; 1]. Для предотвращения возможных временных перехлестов (и как следствия контроля «утечек») в тренировочных и тестовых данных используются процентные инструменты разных стран.

Shape

Description automatically generatedA picture containing shape

Description automatically generated

Как выяснилось, в статье [2] (которую, к сожалению, я обнаружил слишком поздно) была проделана во многом схожая работа. В ней для изначального выявления паттернов используется алгоритм DTW; по сути, авторы сравнивают модели по тому, насколько хорошо они могут приблизить алгоритм DTW, в то время как моей работе акцент был сделан на размеченных вручную данных. Авторы статьи пришли к выводу о том, что двумерные свертки работают лучше, чем одномерные (факт, который мне представляется как противоречивый с теоретической точку зрения, так как порождает по существу искусственное «раздувание» одномерных по своей сути и не самых сложных по структуре данных в двумерные). Но лучший результат среди всех показала рекуррентная модель. Таким образом, кажется, что как параллельное направление исследования, возможно, перспективно было бы попробовать применить модели из другой области со схожей постановкой задачи – области детекции звуковых событий (SELD, например Key Word Spotting, как в статье [3]). В частности, TCN (temporal convolution network).

Проблемы использованной модели:

К сожалению, мне не удалось добиться хоть сколько-нибудь внятного обучения модели.

Background pattern

Description automatically generated with low confidence

В отношении метрик, надо заметить, что, конечно же, вследствие большого дисбаланса классов, accuracy не является правильной метрикой. Однако, попытки написать метрики precision, recall и balanced accuracy вручную не удались – они, по всей видимости, содержат ошибку, выявить которую мне не удалось выявить. Дело в том, что, к примеру, расчетное значение recall близко к единице, при том, что при проверке результата модель не находит ни одного значения из класса 1 из тестовой выборки. Вероятно, ошибка связана с обработкой ситуаций деления на ноль.

Главные ошибки и выводы:

В целом, работа над данной проблематикой убедила меня в том, что ряды биржевых котировок вследствие крайней зашумленности довольно плохо подходят для обработки методами машинного обучения с целью прогнозирования (по крайней мере, в рамках ограниченных ресурсов, которыми располагает учебный проект).

Также, необходимо было более тщательно осуществить поиск схожих работ, чтобы ознакомиться со статьей [2] на более раннем этапе и, возможно, скорректировать направление исследования.

Большой ошибкой в части организации процесса было плохое информирование мной научного руководителя о ситуации/проблемах с работой несмотря на то, что научный руководитель в любой момент был готов предоставить мне все необходимые консультации, а также недостаточное использование ресурсов и материалов, которые им были предоставлены. Хотел бы выразить благодарность научному руководителю за стремление всячески содействовать выполнению работы, которым я не воспользовался должным образом.

Также я отчетливо осознал, что на данный момент, несмотря на, как мне кажется, удовлетворительные теоретические знания, мой опыт в плане практической реализации моделей, к сожалению, недостаточен для качественного выполнения квалификационной работы с собственной разметкой оригинальных данных. Чтобы устранить этот недостаток, я планирую вернуться к отработке навыков обучения стандартных архитектур на базовых публичных датасетах.

Список литературы:

1. 1D convolutional neural networks and applications: A survey. Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, Daniel J. Inman. Mechanical Systems and Signal Processing 151 (2021) 107398
2. Stock Chart Pattern recognition with Deep Learning. Marc Velay and Fabrice Daniel. arXiv: 1808.00418 April 2018
3. Temporal Convolution for Real-time Keyword Spotting on Mobile Devices. Seungwoo Choi, Seokjun Seo, Beomjun Shi, Hyeongmin Byun, Martin Kersner, Beomsu Kim, Dongyoung Kim, Sungjoo Ha
4. Predictive Maintenance – Bridging Artificial Intelligence and IoT. Gerasimos G. Samatas, Seraphim S. Moumgiakmas, George A. Papakostas. arXiv:2103.11148 April 2021