УДК 330.143.1

Жегалин Александр Евгеньевич

магистрант направления подготовки «бизнес-информатика в высокотехнологичных отраслях экономики»
Национальный исследовательский ядерный университет «МИФИ»
Россия, Москва

zhegalin96@gmail.com

Alexander E. Zhegalin

Master's student in the direction of training "Business Informatics in High-Tech Sectors of the Economy"
National Research Nuclear University MEPhI
Russia, Moscow

Мыключенко Наталья Александровна

магистрант направления подготовки «бизнес-информатика в высокотехнологичных отраслях экономики»
Национальный исследовательский ядерный университет «МИФИ»
Россия, Москва

Natalia A. Myklyuchenko

Master's student in the direction of training "Business Informatics in High-Tech Sectors of the Economy"
National Research Nuclear University MEPhI
Russia, Moscow

Котов Евгений Юрьевич

магистрант направления подготовки «бизнес-информатика в высокотехнологичных отраслях экономики»
Национальный исследовательский ядерный университет «МИФИ»
Россия, Москва

zhek.a.k@yandex.ru

Evgeniy Ur. Kotov

Master's student in the direction of training "Business Informatics in High-Tech Sectors of the Economy"
National Research Nuclear University MEPhI
Russia, Moscow

АНАЛИЗ ФОНДОВОГО РЫНКА С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

STOCK MARKET ANALYSIS USING MACHINE LEARNING

Аннотация

Фондовый рынок или рынок акций - один из самых сложных и изощренных способов ведения бизнеса. Мелкие собственники, брокерские корпорации, банковский сектор - все зависит от этого самого органа в плане получения доходов и разделения рисков; очень сложная модель. Однако в этой статье предлагается использовать алгоритм машинного обучения для прогнозирования будущей цены акций для обмена с использованием библиотек с открытым исходным кодом и уже существующих алгоритмов, которые помогут сделать этот непредсказуемый формат бизнеса немного более предсказуемым. Мы увидим, как эта простая реализация принесет приемлемые результаты. Результат полностью основан на числах и предполагает множество аксиом, которые могут или не могут следовать в реальном мире, как и время предсказания.

Ключевые слова:

Анализ данных, линейная регрессия, фондовый рынок, машинное обучение

Abstract

The stock market or stock market is one of the most difficult and sophisticated ways of doing business. Smallholders, brokerage corporations, the banking sector - everything depends on this very body in terms of generating income and sharing risks; very complex model. However, this article proposes to use a machine learning algorithm to predict the future stock price for an exchange, using open source libraries and pre-existing algorithms to help make this unpredictable business format a little more predictable. We will see how this simple implementation yields acceptable results. The result is entirely based on numbers and assumes many axioms that may or may not follow in the real world, as does the timing of the prediction.

Keywords:

Data Analysis, Linear Regression, Stock Market, Machine Learning

Введение

Рынок акций – один из старейших методов, при котором нормальный человек будет торговать акциями, делать инвестиции и зарабатывать деньги на компаниях, которые продают часть себя на этой платформе. Эта система оказывается потенциальной инвестиционной схемой, если все сделано с умом. Однако цены и ликвидность этой платформы очень непредсказуемы, и именно здесь мы предлагаем технологии, чтобы помочь нам. Машинное обучение – один из таких инструментов, который помогает нам достичь того, чего мы хотим. Следующие три абзаца кратко объясняют ключевые компоненты этого документа.

Фондовый рынок, как мы знаем, является очень важной торговой платформой, которая влияет на каждого на индивидуальном и национальном уровне [2]. Основной принцип довольно прост: компании будут размещать свои акции в компаниях как небольшие товары, называемые акциями. Они делают это, чтобы собрать деньги для фирмы. Компания размещает свои акции по цене, называемой IPO или первичным публичным размещением. Это цена предложения, по которой компания продает акции и привлекает деньги. После этого эти акции становятся собственностью владельца, и он может продать их по любой цене покупателю на бирже, такой как БФБ или Бомбейская фондовая биржа. Трейдеры и покупатели продолжают продавать эти акции по своей цене, но компания сохраняет только деньги, заработанные вовремя IPO. Продолжение надежды на зайцев от одной стороны к другой с целью получения большей прибыли приводит к увеличению цены конкретной акции после каждой прибыльной сделки. Однако, если компания выпускает больше акций при более низком IPO, рыночная цена на бирже снижается, и трейдеры терпят убытки. Именно это явление является причиной страха людей инвестировать в фондовые рынки и, в двух словах, причиной падения и роста цен на акции.

Теперь, если мы попытаемся построить график биржевой цены за период времени (скажем, 6 месяцев), действительно ли трудно предсказать следующий результат на графике?

Человеческий мозг очень способен расширить график на несколько координат, просто взглянув на него в течение нескольких минут [1]. И если мы проведем массовые вычисления, то есть заставим группу случайных людей попытаться расширить график на фиксированный промежуток времени (скажем, на неделю), мы получим очень разумный и приблизительный ответ на график реальной жизни.

Потому что многие мозги будут пытаться интерпретировать закономерность и делать предположения, и такая деятельность оказалась намного более успешной на практике, чем кажется в теории. [5] При этом прогнозирование истинной стоимости акций лучше всего оценивается методом краудсчетов.

Но поскольку это очевидно, что крауд-вычисления – это очень медленная деятельность, поэтому мы пытаемся использовать компьютер для моделирования такого примера с более научным и математическим подходом.

В статистике есть способ, с помощью которого мы смотрим на значения и атрибуты проблемы на графиках, определяем зависимые и независимые переменные и пытаемся установить или идентифицировать существующие отношения между ними [3, 4]. Этот метод известен как линейная регрессия в статистике и очень часто используется из-за его очень простого и эффективного подхода. В машинном обучении мы адаптировали тот же алгоритм, в котором мы используем функции для обучения классификатора, который затем предсказывает значение метки с определенной точностью, которую можно проверить во время обучения и тестирования классификатора. Чтобы классификатор был точным, вы должны выбрать правильные функции и иметь достаточно данных для обучения классификатора. Точность вашего классификатора прямо пропорциональна количеству данных, предоставленных классификатору и выбранным атрибутам.

Итак, с базовыми знаниями фондового рынка, графиков и анализа данных в сочетании с машинным обучением; Теперь мы готовы разработать программу.

Прогнозная модель

А. Этап анализа данных

На этом этапе мы рассмотрим доступные нам необработанные данные и изучим их, чтобы определить подходящие атрибуты для прогнозирования выбранной нами метки. Теперь данные, которые мы собираемся использовать для нашей программы, взято с www.quandl.com, ведущей платформы, предоставляющей наборы данных.

Набор данных предназначен для GOOGL от WIKI и может быть извлечен из quandl с помощью токена «WIKI / GOOGL». Мы извлекли и использовали данные примерно за 14 лет.

Атрибуты набора данных включают:

Открытие (цена открытия акций)

High (самая высокая цена, возможная в определенный момент времени) Low (самая низкая цена, возможная в определенный момент времени) Close (цена закрытия акции)

Объем (общее количество сделок в течение дня)

Коэффициент разделения

Прил. Открыть

Прил. Высокая

Прил. Низкий

Прил. Закрывать

Прил. Объем

(Скорректированные значения вышеуказанных атрибутов)

Мы выбираем атрибут «Close» в качестве нашей метки (переменная, которую мы будем прогнозировать) и используем «Adj. Открыть, Настр. Высокий, прил. Close, Adj. Низкий и Настр. Объем», чтобы выделить функции, которые помогут нам лучше предсказать результат.

Следует отметить, что мы используем скорректированные значения вместо необработанных, поскольку эти значения уже обработаны и не содержат общих ошибок при сборе данных.

Теперь мы знаем, что графики, созданные для анализа запасов, используют указанные выше атрибуты для их построения. Такие графики называются графиками OHLCV [11] и очень информативны о состоянии запасов. Теперь мы используем те же параметры графика, чтобы определить особенности классификатора.

Давайте определим набор функций, которые мы будем использовать:

Прил. Закрытие: это важный источник информации, поскольку он определяет цену открытия рынка на следующий день и ожидаемый объем в течение дня.

HL_PCT: это производная функция, которая определяется (1):

$$HL\_PCT = \frac{\text{Adj. High} - \text{Adj. Low}}{Adj.\,Close} \times 100 \quad (1)$$

Мы используем процентное изменение, поскольку это помогает нам уменьшить количество функций, но сохранить задействованную сетевую информацию. High-Low - важная функция, потому что она помогает нам сформулировать форму графика OHLCV.

PCT_change: это также производная функция, которая определяется (2):

$$PCT\_Change = \frac{Adj.Close - Adj.Open}{Adj.Open} \times 100 \quad (2)$$

Мы делаем то же самое с открытием и закрытием, как с максимумом и минимумом, поскольку они оба очень важны в нашей модели прогнозирования и помогают нам уменьшить количество избыточных функций.

Мы успешно проанализировали данные и извлекли полезную информацию, которая нам понадобится для классификатора. Это очень важный шаг, к которому следует относиться с особой осторожностью. Отсутствие информации или небольшая ошибка при получении полезной информации приведет к модели прогнозирования сбоев и очень неэффективному классификатору.

Кроме того, извлеченные функции очень специфичны для используемого предмета и определенно будут варьироваться от предмета к предмету. Обобщение возможно тогда и только тогда, когда данные другого субъекта собираются с той же последовательностью, что и предыдущий субъект.

Б. Этап обучения и тестирования

На этом этапе мы будем использовать то, что мы извлекли из наших данных и внедрить в нашу модель машинного обучения.

Мы будем использовать библиотеки SciPy, Scikit-learn и Matplolib в python, чтобы запрограммировать нашу модель, обучить их функциям и меткам, которые мы извлекли, а затем протестировать их с теми же данными.

Сначала мы предварительно обработаем данные, чтобы получить данные, которые включают:

1. Значения атрибута метки смещены на процент, который вы хотите спрогнозировать.

2. Формат Dataframe преобразуется в формат массива Numpy.

3. Все значения данных NaN удаляются перед подачей их в классификатор.

4. Данные масштабируются таким образом, что для любого значения X из отрезка [-1,1].

5. Данные разделены на тестовые данные и данные для обучения.

6. В соответствии с его типом, т. е. меткой и функцией.

Теперь данные готовы для ввода в классификатор. Мы будем использовать простейший классификатор, то есть линейную регрессию, которая определена в

библиотеке Sklearn пакета Scikit-learn. Мы выбрали этот классификатор из-за его простоты и потому, что он идеально подходит для наших целей. Линейная регрессия - очень часто используемый метод анализа и прогнозирования данных. По сути, он использует ключевые функции для прогнозирования отношений между переменными на основе их зависимости от других функций [9]. Эта форма прогнозирования известна как машинное обучение с учителем. Контролируемое обучение – это метод, при котором мы вводим помеченные данные, то есть функции сопоставляются с их метками. Здесь мы обучаем классификатор таким образом, чтобы он узнал шаблоны того, какая комбинация функций приводит к какой метке.

В нашем случае классификатор видит признаки, просто смотрит на их метку и запоминает. Он запоминает комбинацию функций и соответствующий ярлык, который в нашем случае является ценой акции через несколько дней. Затем он переходит к следующему этапу и изучает, по какому шаблону следуют функции, чтобы создать соответствующую метку. Так работает машинное обучение с учителем [10].

Для тестирования в управляемом машинном обучении мы вводим некоторую комбинацию функций в обученный классификатор и перекрестно проверяем выходные данные классификатора с реальной меткой. Это помогает нам определить точность нашего классификатора. Что очень важно для нашей модели. Классификатор с точностью менее 95% практически бесполезен.

Точность – очень важный фактор в модели машинного обучения. Вы должны понимать, что означает точность и как повысить точность следующей подтемы.

Полезные подсказки

А. Требования и спецификации

На самом первом этапе вы должны знать точные требования к проблеме, а также спецификацию машины и производительности. Не торопитесь с этим шагом, так как этот шаг очень важен при принятии решения об общем плане развития программы.

Внимательно изучите кейс, сделайте небольшую проверку биографических данных, соберите достаточные знания по предмету, определите, чего вы действительно хотите, и установите это в качестве своей цели.

Б. Тщательный анализ функций

Вы должны быть очень осторожны при извлечении признаков из данных, поскольку они играют непосредственную роль в модели прогнозирования. Все они должны иметь прямой смысл в сочетании с этикетками. Также настоятельно

рекомендуется минимизировать функции, подчиняющиеся ограничениям требований, насколько это возможно.

C. Осуществление

Вы должны выбрать подходящую модель, в которой вы будете реализовывать свои математические вычисления для получения результатов.

Выбранная или разработанная модель должна соответствовать входным данным. Неправильная модель, разработанная или выбранная для несоответствующих данных, или наоборот, приведет к модели мусора, которая совершенно бесполезна. Вы должны увидеть совместимую SVM или другие доступные методы обработки ваших данных. Также хорошей практикой является одновременное опробование разных моделей, чтобы проверить, какая из них работает наиболее эффективно.

Более того, реализация - это самый простой шаг из всех, и он должен занимать минимум времени, чтобы сэкономить нам время из общих временных затрат, которые можно было бы использовать на некоторых других важных шагах.

D. Обучение и тестирование

Обучение модели очень простое. Вам нужно только убедиться, что данные непротиворечивы, согласованы и доступны в большом количестве. Большой набор обучающих данных способствует созданию более сильного и точного классификатора, что в конечном итоге увеличивает общую точность.

Тестирование - это тоже очень простой процесс. Убедитесь, что ваши тестовые данные составляют не менее 20% от размера ваших тренировочных данных. Важно понимать, что тестирование - это проверка точности ваших классификаторов, и иногда наблюдается, что оно обратно пропорционально баллу классификаторов. Однако точность классификатора не зависит и не коррелирует с тестированием. Иногда так кажется, но тестирование никак не связано с классификатором.

E. Оптимизация

Практически невозможно создать универсальный классификатор за один раз, поэтому мы всегда должны продолжать оптимизацию. Всегда есть место для улучшений. При оптимизации учитывайте стандартные методы и основные требования.

Переход на SVM, опробование и тестирование различных моделей, поиск новых и улучшенных функций, изменение всей модели данных для полного соответствия

модели и т.д. – вот некоторые очень фундаментальные способы оптимизации вашего классификатора.

Вывод

Машинное обучение, как мы видели до сих пор, является очень мощным инструментом, и, как ни крути, у него есть отличное применение. До сих пор мы видели, что машинное обучение очень сильно зависит от данных. Таким образом, важно понимать, что данные бесценны, и, как бы просто это ни звучало, анализ данных – непростая задача.

Машинное обучение нашло огромное применение и эволюционировало в глубокое обучение и нейронные сети, но основная идея более или менее одинакова для всех них. В этой статье дается четкое представление о том, как реализовать машинное обучение. Существуют различные способы, методы и техники для решения и решения различных проблем в различных вообразимых ситуациях. Эта статья ограничивается только контролируемым машинным обучением и пытается объяснить только основы этого сложного процесса.

Список использованных источников

1. Andrew McCallum, Kamal Nigam, Jason Rennie, Kristie Seymore "A Machine learning approach to Building domain-specific Search engine", IJCAI, 1999 - Citeseer

2. Yadav, Sameer. (2017). STOCK MARKET VOLATILITY - A STUDY OF INDIAN STOCK MARKET. Global Journal for Research Analysis. 6. 629-632.

3. Montgomery, D.C., Peck, E.A. and Vining, G.G., 2012. Introduction to linear regression analysis (Vol. 821). John Wiley & Sons.

4. Draper, N.R.; Smith, H. (1998). Applied Regression Analysis (3rd ed.). John Wiley. ISBN 0-471-17082-8.

5. Robert S. Pindyck and Daniel L. Rubinfeld (1998, 4h ed.). Econometric Models and Economic Forecasts

6. "Linear Regression", 1997-1998, Y ale University http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

7. Agarwal (July 14, 2017). "Introduction to the Stock Market". Intelligent Economist. Retrieved December 18, 2017.

8. Jason Brownlee, March 2016, "Linear Regression for machine learning", Machine learning mastery, viewed on December 2018, https://machinelearningmastery.com/linear-regression-for-machine- learning/

9. Google Developers, Oct 2018, "Decending into ML: Linear Regression", Google LLC, https://developers.google.com/machine- learning/crash-course/descending-into-ml/linear-regression

10. Fiess, N.M. and MacDonald, R., 2002. Towards the fundamentals of technical analysis: analysing the information content of High, Low and Close prices. Economic Modelling, 19(3), pp.353-374.

11. Hurwitz,E. and Marwala,T.,2012.Commonmistakeswhenapplying computational intelligence and machine learning to stock market modelling. arXiv preprint arXiv:1208.4429.