УДК 004.85

ПРИМЕНЕНИЕ АЛОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ПРИКЛАДНЫХ ЗАДАЧ

Рыспаев[1] А.О., Байгазаков[2] К.А., Имангазы[2] уулу Н., Сабитов[1] Б.Р., Анарбек[2] к Э.
[1]КНУ им.Ж.Баласагына, [2]КТУ им.И.Раззакова

В данной работе, используя методы машинного обучения, выявляется побочные процессы в банковской деятельности. Изучим применение нейронных сетей в процесс построения модели банковского кризиса по некоторым данным. Рассматривается задача как алгоритм машинного обучения, который может прогнозировать банковский криз. Применение технологий глубокого обучения могут дать хорошие результаты работы с базами данных.

Ключевые слова: кластеризация, машинное обучение, фондовый рынок, линейная регрессия

КОЛДОНУЛГАН МАСЕЛЕЛЕРДИ БОЛЖОО ҮЧҮН МАШИНА ҮЙРӨНҮҮ АЛГОРИТМДЕРИН КОЛДОНУУ

Рыспаев[1] А.О., Байгазаков[2] К.А., Имангазы[2] уулу Н., Сабитов[1] Б.Р., Анарбек[2] К.
[1]Ж.Баласагын атындагы КУУ, [2]И.Раззаков атындагы КТУ

Бул макалада, машиналык окутуу ыкмаларын колдонуу менен, банк ишинде кошумча процесстер аныкталган. Биз кээ бир маалыматтардын негизинде банктык кризис моделин куруу процессинде нейрон тармактарын колдонууну изилдейбиз. Көйгөй банктык кризисти алдын ала ала турган машина үйрөнүү алгоритми катары каралат. Маалымат базалары менен иштөөдө терең окутуу технологияларын колдонуу жакшы натыйжаларды бере алат.

Баштапкы сөздөр: кластерлөө, машина үйрөнүү, биржа, сызыктуу регрессия.

APPLICATION OF MACHINE LEARNING ALORHYTHMS FOR PREDICTION OF APPLIED PROBLEMS

Ryspaev[1] A.O., Baigazakov[2] K.A., Imangazy[2] uulu N.,
Sabitov[1] B.R., Anarbek[2] to E.
[1]KNU named after Zh.Balasagyn, [2]KTU named after I.Razzakov

In this paper, using machine learning methods, side processes in banking are identified. We will study the use of neural networks in the process of building a banking crisis model based on some data. The problem is considered as a machine learning algorithm that can predict a banking crisis. The use of deep learning technologies can give good results when working with databases.

Keywords: clustering, machine learning, stock market, linear regression

Введение. В данной статье мы рассмотрим, задачу как алгоритмы машинного обучения могут прогнозировать банковский криз. Финансовые кризисы, в банковской системе, начиная с дефолтов до массовых изъятий средств из банков и валютных кризисов. Алгоритмы машинного обучения, для улучшения оценки вероятности финансового кризиса. Во-первых, обучение без учителя отличается от обучения с учителем тем, что в нем нет переменной отклика. Кластеризация — это один из методов, который стоит выделить. Цель кластеризации — разумно сгруппировать точки данных. Эти группы данных будут связаны с центром масс, чтобы помочь определить структуру в наборах данных. Кластеризация может применяться как к зависимой, так и к независимой переменной. Например, вместо того, чтобы использовать фиксированный порог для определения валютного кризиса, мы можем разделить доходность валюты на разные кластеры и извлечь разумное значение из каждого кластера.

Таким образом, алгоритмы машинного обучения могут принести значительную пользу.

В данной статье мы также, изучим с применение нейронных сетей в процесс построения модели банковского кризиса по некоторым данным. Применение технологий глубокого обучения могут дать хорошие результаты работы с базами данных. Прогнозируя с помощью модели

глубокого обучения, мы увидим, что эта модель дает высокую точность в этой задаче.

Рассматривается задача, как будет развиваться фондовый рынок. Это одна из самых сложных задач в банковской сфере. Она является нелинейным процессом. В прогнозирование могут быть задействовано так много факторов - физические факторы против физиологического, рационального и иррационального поведения и т. д. Все эти аспекты в совокупности делают цены на акции нестабильными, и их очень трудно предсказать с высокой степенью точности. Будем использовать технологии машинного и глубокого обучения. Изучим для прогнозирования будущей цены акций этой компании, начиная с простых алгоритмов, таких как усреднение и линейная регрессия, а затем перейдем к продвинутому методу LSTM т.е. к нейронным сетям.

Основная идея этой работы это - продемонстрировать, как реализованы эти алгоритмы. Опишем технику и предоставим соответствующие коды для реализации задачи, а также, будем использовать ряд алгоритмов прогнозирования временных рядов.

Мы также рассмотрим применение машинного обучения до продвинутых концепций машинного обучения, глубокого обучения и временных рядов.

Методы исследования. Результаты прогнозирования.

Рассмотрим с анализа фондового рынка, которая делится на две части - фундаментальный анализ и технический анализ. Фундаментальный анализ в нашем случае, включает анализ будущей прибыльности компании на основе ее текущей деловой среды и финансовых показателей. Для анализа мы взяли открытую систему данных, в которой мы будем использовать csv файл для прогнозирования. Структура базы данных фондового рынка. Теперь

давайте загрузим набор данных и определим целевую переменную для проблемы с помощью pandas. Вот база данных.

	Date	Open	High	Low	Last	Close	Total Trade Quantity	Turnover (Lacs)
0	2018-10-08	208.00	222.25	206.85	216.00	215.15	4642146.0	10062.83
1	2018-10-05	217.00	218.60	205.90	210.25	209.20	3519515.0	7407.06
2	2018-10-04	223.50	227.80	216.15	217.25	218.20	1728786.0	3815.79
3	2018-10-03	230.00	237.50	225.75	226.45	227.60	1708590.0	3960.27
4	2018-10-01	234.55	234.60	221.05	230.30	230.90	1534749.0	3486.05

В наборе данных есть несколько переменных - дата, открытие, максимум, минимум, последнее, закрытие, total_trade_quantity и оборот. Столбцы «Открытие» и «Закрытие» представляют собой начальную и конечную цену, по которой акция торгуется в определенный день.
High, Low и Last представляют максимальную, минимальную и последнюю цену акции за день.

Общее количество сделок - это количество акций, купленных или проданных в день, а оборот (Lacs) - это оборот конкретной компании на заданную дату.

Еще одна важная вещь, которую следует отметить, это то, что рынок закрыт по выходным и праздничным дням. Обратим внимание на приведенную выше таблицу, некоторые значения даты отсутствуют - 10.02.2018, 10.06.2018, 10.07.2018. Давайте нарисуем целевую переменную, чтобы понять, как она формируется в наших данных. Вот динамика изменения данных

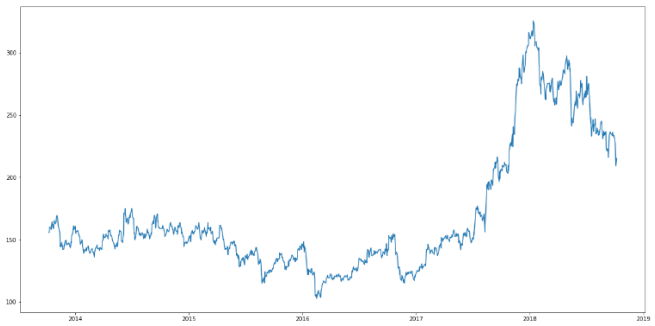

Рис.1. Динамика изменения данных.

Далее, исследуем эти переменные и будем использовать различные методы для прогнозирования дневной цены закрытия акций.

Скользящая средняя

«Средний» - одна из самых распространенных вещей, которые мы используем в повседневной жизни. Например, вычисление средних оценок для определения общей производительности или определение средней температуры за последние несколько дней, чтобы получить представление о сегодняшней температуре - все это рутинные задачи, которые мы выполняем на регулярной основе. Так что это хорошая отправная точка для использования в нашем наборе данных для составления прогнозов.

Прогнозируемая цена закрытия для каждого дня будет средним из набора ранее наблюдаемых значений. Вместо использования простого среднего мы будем использовать метод скользящего среднего, который использует последний набор значений для каждого прогноза. График прогнозируемых значений вместе с фактическими значениями.

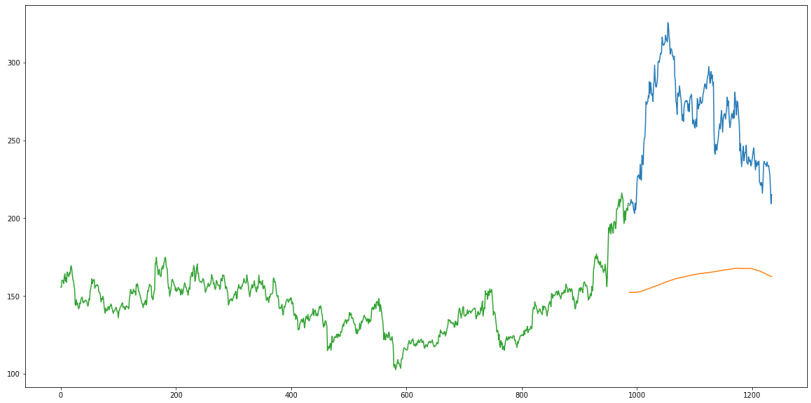

Рис.2. Применение скользящего среднего в задаче прогнозирования фондового рынка

Значение RMSE близко к 105, но результаты не очень многообещающие (как вы можете понять из графика). Прогнозируемые значения находятся в том же диапазоне, что и наблюдаемые значения в

наборе данных обучения (сначала наблюдается тенденция к увеличению, а затем - к медленному снижению).

Рассмотрим два часто используемых метода машинного обучения - линейную регрессию и kNN - и посмотрим, как они работают с данными нашего фондового рынка.

Линейная регрессия

Самый простой алгоритм машинного обучения, который можно реализовать на этих данных, - это линейная регрессия. Модель линейной регрессии возвращает уравнение, определяющее связь между независимыми переменными и зависимой переменной.

Уравнение линейной регрессии можно записать как:

$$Y = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n, \quad (1)$$

В регрессионной формуле x_1, x_2,… . x_n представляют собой независимые переменные, а коэффициенты θ_1 , θ_2,…. θ_nпредставляют собой веса. Используя (1) можно изучить линейную регрессию более для нашей рассматриваемой задачи. Для нашей постановки задачи у нас нет набора независимых переменных. Вместо этого у нас есть только даты. Давайте воспользуемся столбцом даты для извлечения таких функций, как день, месяц, год, пн / пт и т. д., А затем применим модель линейной регрессии.

Реализация примера

Сначала мы отсортируем набор данных в порядке возрастания, а затем создадим отдельный набор данных, чтобы любая новая созданная функция не влияла на исходные данные.

Это создает такие функции, как:

```
[ ]  'Year', 'Month', 'Week', 'Day', 'Dayofweek',
      'Dayofyear', 'Is_month_end', 'Is_month_start',
       'Is_quarter_end', 'Is_quarter_start',
        'Is_year_end', and  'Is_year_start'.
```

Помимо этого, мы можем добавить наш собственный набор функций, которые, по нашему мнению, будут иметь отношение к прогнозам. Например, гипотеза состоит в том, что первый и последний дни недели потенциально могут повлиять на цену закрытия акций гораздо больше, чем другие дни. Итак, мы создали функцию, которая определяет, является ли данный день понедельником / пятницей или вторником / средой / четвергом.

Если день недели равен 0 или 4, значение столбца будет равно 1, иначе 0. Точно так же вы можете создать несколько объектов. Если у вас есть идеи по функциям, которые могут быть полезны при прогнозировании курса акций, поделитесь ими в разделе комментариев. Теперь мы разделим данные на наборы для обучения и проверки, чтобы проверить производительность модели.

Вот результат. Среднеквадратичное значение = 121.16291596523156

Значение RMSE выше, чем в предыдущем методе, что ясно показывает, что линейная регрессия работает плохо. Давайте посмотрим на график и поймем, почему линейная регрессия не принесла успеха:

Рис.3. Линейная регрессия фондового рынка.

Линейная регрессия - это простой метод, который довольно легко интерпретировать, но у него есть несколько очевидных недостатков. Одна из проблем при использовании алгоритмов регрессии

заключается в том, что модель не соответствует столбцу даты и месяца. Вместо того, чтобы принимать во внимание предыдущие значения с точки зрения прогноза, модель будет рассматривать значение с той же даты месяц назад или с той же даты / месяца год назад.

k-Ближайшие соседи

Еще один интересный алгоритм машинного обучения, который можно использовать здесь, - это kNN (k ближайших соседей). На основе независимых переменных kNN находит сходство между новыми и старыми точками данных. На простом примере данный алгоритм выглядит следующим образом.

ID	Age	Height	Weight
1	45	5	77
2	26	5.11	47
3	30	5.6	55
4	34	5.9	59
5	40	4.8	72
6	36	5.8	60
7	19	5.3	40
8	28	5.8	60
9	23	5.5	45
10	32	5.6	58
11	38	5.5	?

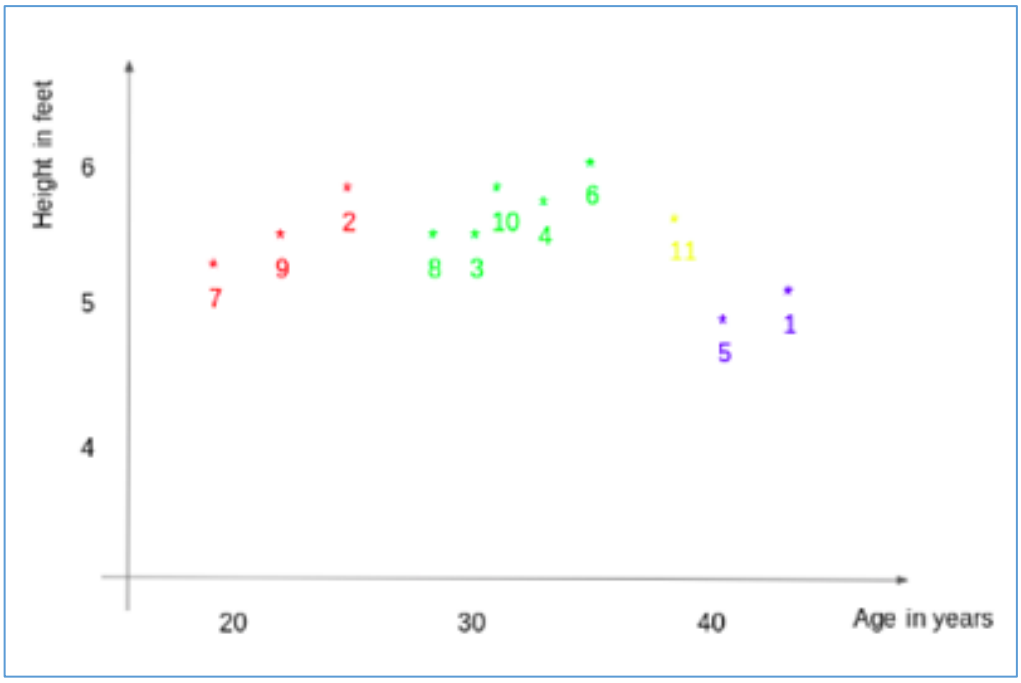

Рис.4. Расположение ближайших соседей.

Чтобы определить вес для ID # 11, kNN учитывает вес ближайших соседей этого ID. Предполагается, что вес ID # 11 будет средним для его соседей. Если мы сейчас рассмотрим трех соседей (k = 3), вес для ID # 11 будет = (77 + 72 + 60) / 3 = 69,66 кг.

ID	Height	Age	Weight
1	5	45	77
5	4.8	40	72
6	5.8	36	60

Среднеквадратичное значение =115.17086550026721

В значении RMSE нет большой разницы, но график для прогнозируемых и фактических значений должен обеспечить более четкое понимание.

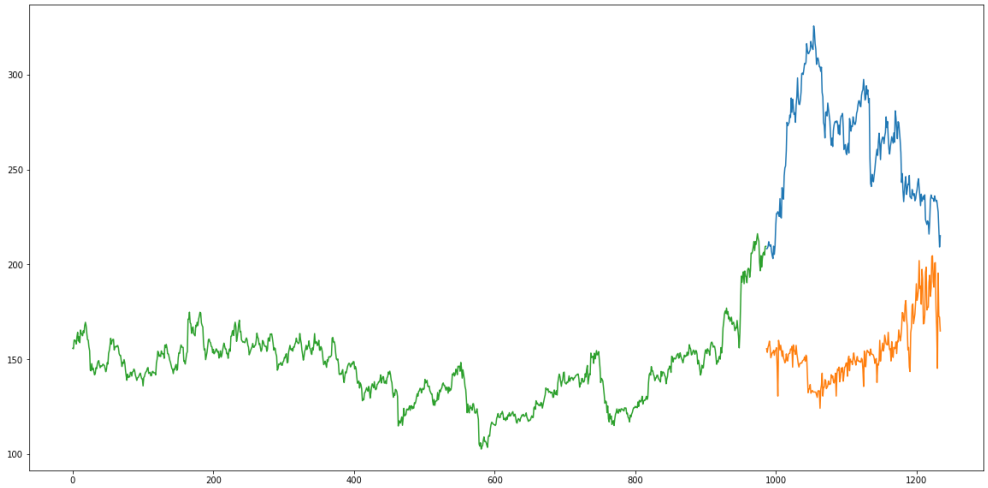

Рис.5. Применение метода ближайших соседей.

Значение RMSE почти аналогично модели линейной регрессии, и график показывает ту же картину. Как и линейная регрессия, kNN также выявила падение в январе 2018 года, поскольку это было закономерностью в течение последних лет. Мы можем с уверенностью сказать, что алгоритмы регрессии плохо работают с этим набором данных.

Технологии FBProphet

Существует ряд методов временных рядов, которые могут быть реализованы в наборе данных прогнозирования запасов, но большинство

из этих методов требуют предварительной обработки большого количества данных перед подгонкой модели. Prophet, разработанная и впервые использованная Facebook, представляет собой библиотеку прогнозирования временных рядов, которая не требует предварительной обработки данных и чрезвычайно проста в реализации. Входными данными для Prophet является фрейм данных с двумя столбцами: дата и цель (ds и y).

Prophet пытается уловить сезонность в прошлых данных и хорошо работает, когда набор данных большой. Вот интересная статья, которая объясняет Prophet простым и интуитивно понятным образом:

Вот полученный прогноз

Рис.6.Прогнозирование с помощью технологии FBProphet.

Prophet (как и большинство методов прогнозирования временных рядов) пытается уловить тренд и сезонность на основе прошлых данных. Эта модель обычно хорошо работает с наборами данных временных рядов, но в этом случае не соответствует своей репутации.

Давайте продолжим и попробуем другой продвинутый метод - Long Short Term Memory (LSTM) основанный на нейронных технологиях.

Долговременная краткосрочная память -LSTM

Метод LSTM широко используются для задач прогнозирования последовательности и оказались чрезвычайно эффективными.

Нейронная сеть LSTM может хранить прошлую информацию, которая важна, и забывать информацию, которая не является важной.

Рис.7.Применение нейронной сети к задаче прогнозирования фондового рынка.

ЗАКЛЮЧЕНИЕ

В статье мы изучили ряд методов машинного и глубокого обучения для нелинейного моделирования и прогнозирования фондового рынка. Но среди всех лидеромоказалось метод основанный на нейронных сетях. Для большинства прикладных задач, особенно медицины, экономики, климата, банковской деятельности модель LSTM можно настроить для различных параметров, таких как изменение количества слоев LSTM, добавление значения выпадения или увеличение количества эпох. Таким образом, мы изучили прогнозирование временных рядов, которые одновременно являются очень сложной областью нелинейного прогнозирования

ЛИТЕРАТУРА

1. Ясер Абу-Мостафа, Малик Магдон-Исмаил, Сюань-Тянь Линь − Learning From Data, 2012 г.
2. П. Брюс, Э. Брюс – Практическая статистика для специалистов Data Science,2020.

3. О'Нил, Шатт – Data Science. Инсайдерская информация для новичков,2020
4. Ын, Су – Теоретический минимум по Big Data. Всё что нужно знать о больших данных,2020.
5. Силен, Мейсман, Али – Основы Data Science и Big Data. Python и наука о данных,2020.
6. Дж. Вандер Плас – Python для сложных задач. Наука о данных и машинное обучение,2020