МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

ПРОГНОЗ ЦЕН НА ЗОЛОТО НА РАЗНЫХ ВРЕМЕННЫХ ПРОМЕЖУТКАХ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 451 группы
направления 38.03.05 — Бизнес-информатика

механико-математического факультета
Агаповой Елизаветы Евгеньевны

Научный руководитель
к.ф.-м.н. _____ Р. Н. Фадеев

Заведующий кафедрой
д. ф.-м. н., профессор _____ С. П. Сидоров

Саратов 2023

ВВЕДЕНИЕ

Актуальность темы. Тема прогнозирования цен на золото на разных временных промежутках является актуальной в связи с тем, что золото является одним из наиболее важных инструментов инвестирования и хранения ценности. В условиях быстро меняющейся экономической ситуации и глобальных финансовых рисков, способность предсказывать изменения цен на золото может быть важным инструментом для принятия решений об инвестировании.

Использование языка Python для анализа и прогнозирования цен на золото имеет ряд преимуществ, так как этот язык программирования имеет богатый набор библиотек для анализа данных и машинного обучения, которые могут использоваться для создания точных и надежных прогнозов.

Дипломная работа на эту тему включает в себя разработку и применение различных алгоритмов машинного обучения для прогнозирования цен на золото на различных временных интервалах, включая краткосрочные, среднесрочные и долгосрочные периоды.

Целью бакалаврской работы является создание модели прогнозирования цен на золото, которая будет иметь высокую точность и будет способна предсказывать цены на различных временных промежутках.

Обьект исследования- исторические данные цен на золото и их изменения на различных временных интервалах, полученные из открытых источников. В моей дипломной работе взят dataset с платформы kaggle.

Предмет исследования- разработка алгоритмов машинного обучения и методов анализа данных для прогнозирования цен на золото на различных временных интервалах.

Для достижения поставленных целей в работе необходимо решить следующие **задачи**:

— сбор и предобработка данных: проанализировать источники данных для получения исторических цен на золото, определить подходящие временные интервалы для прогнозирования цен, и провести предварительную обработку данных;

— разработка моделей машинного обучения: выбрать наиболее подходя-

щие алгоритмы машинного обучения для прогнозирования цен на золото, определить гиперпараметры моделей и обучить их на исторических данных;
— сравнение производительности различных алгоритмов и выбор наиболее эффективного подхода для прогнозирования цен на золото;
— оценка точности прогнозов и определение фактороввлияющих на цены на золото, таких как экономические и политические события, изменения спроса и предложения на рынке;
— построение графиков и визуализация результатов прогнозирования цен на золото.

Прогноз цен на золото имеет **практическую значимость** для инвесторов, трейдеров и других участников финансовых рынков. На основе прогнозов цен на золото можно принимать решения о покупке и продаже золота, определять оптимальные моменты для входа и выхода с рынка, а также управлять рисками. Кроме того, данная программа может быть использована в качестве инструмента для анализа и прогнозирования цен на золото в различных экономических и политических ситуациях, что поможет более точно предсказывать изменения цен на золото в будущем.

Структура и содержание бакалаврской работы. Работа состоит из введения, четырех разделов, заключения, списка использованных источников, содержащего 20 наименований и 3 приложений. Общий объем работы составляет 50 страниц.

Основное содержание работы

Во **введении** обосновывается актуальность темы работы, формулируется цель работы и решаемые задачи, отмечается практическая значимость получаемых результатов.

В **первом** разделе проводится анализ предметной области, предварительный анализ данных.

Золото - драгоценный металл, который использовался для монетизации, ювелирных изделий и других искусств на протяжении всей истории. В прошлом золотой стандарт часто использовался как монетарная политика. Золотые монеты перестали выпускаться в качестве обращающейся валюты в

1930-х годах, и мировой золотой стандарт был заменен фиатной денежной системой после мер Никсона в 1971 году.

В 2021 году крупнейшим производителем золота в мире был Китай, за которым следовали Австралия и Россия. Всего над землей существует около 190 тысяч тонн на данный момент. Это равно кубу с каждой стороной примерно 221 см.. Потребление нового произведенного золота в мире составляет примерно 50% на ювелирных изделиях, 40% на инвестициях и 10% на промышленности. Высокая ковкость, пластичность, устойчивость к коррозии и большинству других химических реакций, а также проводимость электричества позволили продолжать использование золота в коррозионностойких электрических соединителях во всех типах компьютеризированных устройств. Золото также используется в инфракрасном экранировании, производстве цветного стекла, золочении и восстановлении зубов. Некоторые соли золота до сих пор используются в медицине в качестве противовоспалительных средств. Поэтому так важно уметь анализировать рынок золота и понимать его тенденции.

Данные, которые используются в работе были найдены на платформе **kaggle**.

Датасет содержит исторические данные о цене на золото с 4 января 2000 года по 2 сентября 2022 года. Датасет состоит из 7 колонок:

1. Date (дата в формате год-месяц-день)
2. Open (цена открытия)
3. High (наивысшая цена)
4. Low (наименьшая цена)
5. Close (цена закрытия)
6. Volume (объем торгов)
7. Currency (валюта, в которой отображается цена)

Всего в датасете содержится 5703 строк. Из них две колонки (Date и Currency) являются категориальными, одна (Volume) - целочисленной, а остальные четыре (Open, High, Low, Close) - числовыми. Также было отмечено, что стандартное отклонение цен на золото велико по сравнению с их средним значением.

Анализ корреляции между переменными позволяет оценить связь меж-

ду ними. В данном датасете была построена корреляционная матрица методом **.corr()** и визуализирована с помощью градиентной цветовой схемы.

Каждая ячейка в матрице показывает коэффициент корреляции Пирсона между двумя переменными. Коэффициент корреляции находится в диапазоне от -1 до 1, где -1 означает полную обратную корреляцию, 0 - отсутствие корреляции и 1 - полную прямую корреляцию.

В данном случае видно, что есть высокая корреляция между переменными "open "high "low"и "close что можно объяснить тем, что это цены на золото в разное время дня. Также видно, что между переменными "open" "high" "low" "close"и "volume"корреляция невысока, что говорит о том, что они не сильно зависят друг от друга.

В **втором** разделе производится визуализация данных различными методами, а также по ним сделаны выводы, необходимые для дальнейшей работы.

Тепловая карта (**Heatmap**) - это метод визуализации данных, который позволяет отображать величину переменных в двумерном пространстве с помощью цвета. На данном наборе данных была построена тепловая карта, которая показывает корреляцию между всеми переменными, используемыми для предсказания цены золота. Результат показал, что все переменные имеют достаточно сильную корреляцию с ценой закрытия золота, кроме объема продаж.

Pairplot - это метод визуализации, который позволяет построить графики зависимостей между всеми парами переменных в наборе данных. На данном наборе данных был построен **pairplot**, который показал, что некоторые переменные имеют достаточно сильную корреляцию между собой.

Distplot - это метод визуализации, который позволяет построить гистограмму распределения значений переменной. На данном наборе данных были построены гистограммы распределения цены золота и объема продаж. Результат показал, что распределение цены золота имеет форму нормального распределения, а распределение объема продаж имеет форму, близкую к экспоненциальному распределению. **Lineplot** - это метод визуализации, который позволяет построить график изменения значения переменной во времени. На данном наборе данных был построен **lineplot**, который показал изменение

цены золота с течением времени. Результат показал, что цена золота имеет тенденцию к увеличению со временем.

Jointplot - это метод визуализации, который позволяет построить график зависимости между двумя переменными, включая гистограммы распределений каждой переменной. На данном наборе данных был построен jointplot, который показал корреляцию между ценой закрытия и максимальной ценой на золото.

Визуализация истории цен на золото открытия и закрытия, показывает, как цены на золото колебались за все время. График показывает, что цены на золото имели значительные колебания за последние 22 года, но общая тенденция была восходящей.

Диаграмма гистограммы с помощью библиотеки Plotly Express показывает распределение переменной "Close". Диаграмма гистограммы демонстрирует, что распределение переменной "Close"близко к нормальному, со средним значением в районе 1300 долларов за унцию.

Заметны колебания цены на золото на разных временныз промежутках, для того, чтобы с этим разобраться необходимо рассмотреть исторические сведения.

Некоторые конкретные события, которые могли повлиять на цены золота в период с 2000 по 2022 годы:

1. 2001: террористические атаки 11 сентября в США
2. 2008: мировой финансовый кризис
3. 2011: кризис еврозоны, начало гражданской войны в Сирии
4. 2013: закрытие правительственных организаций в США из-за финансового спора в Конгрессе
5. 2014: аннексия Крыма Россией, падение цен на нефть
6. 2016: выборы президента США, Brexit
7. 2020: пандемия COVID-19, нестабильность на мировых рынках, рост безработицы.

Если посмотреть на график изменения цены, то видно, что каждое из событий повлияло на цену, поэтому частью задания будет, реализовать прогноз после важного события в мире и сравнить с реальными значениями и только после этого строить прогноз на будущее.

В **третьем** разделе реализована предварительная обработка данных для разделения на тестовую и обучающую выборку.

Предварительная обработка набора данных включает в себя несколько шагов:

1. Удаление столбцов, которые не будут использоваться для анализа (High, Low, Open, Volume, Date).
2. Удаление строк с отсутствующими значениями.
3. Нормализация данных для обеспечения стабильности обучения модели. MinMaxScaler - это метод масштабирования данных, который приводит значения признаков к заданному диапазону, обычно от 0 до 1. Он преобразует данные, масштабируя признаки до определенного диапазона значений. Метод сохраняет форму исходного распределения данных и не изменяет существенно информацию, заключенную в исходных данных. Однако следует отметить, что метод не уменьшает важность выбросов. По умолчанию диапазон значений для каждого признака после применения метода MinMaxScaler составляет от 0 до 1.
4. Разбиение набора данных на обучающую и тестовую выборки.
5. Преобразование выборок в массив.
6. Изменение формы данных.

Эти шаги позволяют обеспечить корректную работу модели и получить точные результаты прогнозирования.

В **четвертом** разделе происходит применение метода **LSTM**, рассчет среднеквадратической ошибки(**RMSE**), оценка точности модели, нанесение данных на график и сравнение спрогнозированных цен с действительными. После оценки качества модели будет сделан вывод о введении нового метода машинного обучения для прогноза.

Long Short-Term Memory (LSTM) - это тип рекуррентных нейронных сетей (RNN), который может запоминать предыдущие значения входных данных в течение длительного периода времени. LSTM представляет собой более сложную и продвинутую версию простой RNN и используется для решения проблем, связанных с зависимостью от контекста в данных, таких как последовательности или временные ряды.

Архитектура LSTM состоит из четырех основных слоев: слой входных

данных, слой забывания, слой обновления и слой вывода. Каждый слой состоит из блоков памяти, нейроны которых могут хранить и обрабатывать информацию. Блок памяти LSTM оснащен механизмом, который позволяет ему решать проблему затухающего или взрывающего градиента.

Применение LSTM для прогнозирования цен на золото включает подготовку данных и обучение модели. После предварительной обработки данных они подаются на вход модели, которая обучается на примерах их временной зависимости, чтобы выявить скрытые закономерности и создать прогнозы на основе этих закономерностей. После обучения модель может использоваться для прогнозирования будущих цен на золото.

Спрогнозированный результат:

```
array([[1775.2266],
       [1772.3136],
       [1767.8931],
       [1767.6296],
       [1766.8391],
       [1765.4539],
       [1765.4197],
       [1765.7999],
       [1787.0837],
       [1789.0782]], dtype=float32)
```

В данном случае, среднеквадратическая ошибка была рассчитана для оценки качества предсказаний модели LSTM на тестовом наборе данных.

RMSE (англ. Root Mean Squared Error) – это метрика, которая широко используется для измерения различий между фактическими значениями и прогнозируемыми значениями в статистических и машинных моделях.

RMSE вычисляется как квадратный корень из среднего значения квадратов разностей между прогнозируемыми и фактическими значениями. Это полезная метрика для оценки точности прогнозов модели, так как она дает представление о том, насколько близки прогнозируемые значения к фактическим.

Чем меньше значение RMSE, тем лучше модель прогнозирует фактиче-

ские значения. Однако стоит отметить, что RMSE подвержен влиянию выбросов, поэтому его не следует использовать как единственную метрику оценки производительности модели.

Получился следующий результат:

RMSE score of the model: 39.42306093881927

Результат RMSE 39.42 означает, что средняя ошибка прогноза модели составляет около 39.42 доллара на унцию золота. Чем ближе значение RMSE к нулю, тем лучше модель способна предсказывать значения. Это говорит о том, что модель находится на среднем расстоянии от реальных значений на этом наборе данных.

Последний этап: сравнение спрогнозированных цен и действительных.

```
close Predictions
5418 1811.4 1775.226562
5419 1803.4 1772.313599
5420 1805.4 1767.893066
5421 1801.8 1767.629639
5422 1799.2 1766.839111
5423 1799.8 1765.453857
5424 1799.7 1765.419678
5425 1835.8 1765.799927
5426 1817.2 1787.083740
5427 1822.2 1789.078247
```

Из данного примера видно, что для каждой из дат приведены исходные значения цены на золото **close** и значения, которые предсказала модель **Predictions**. Можно заметить, что в целом значения, предсказанные моделью, следуют тенденции реальных значений, но есть и расхождения. Так, например, для строки 5425 модель предсказала значение 1765.799927, в то время как реальное значение равно 1835.8.

Также будет сделан новый прогноз на основе другого метода машинного обучения- линейной регресси, в связи с недостаточным качеством, полученнрой модели.

В **заключении** приведены результаты бакалаврской работы.

Основные результаты

1. Источники данных для получения исторических цен на золото были проанализированы, подходящие временные интервалы для прогнозирования цен были определены, и данные были предварительно обработаны.
2. Были выбраны наиболее подходящие алгоритмы машинного обучения для прогнозирования цен на золото, были определены гиперпараметры моделей и они были обучены на данных.
3. Производительность различных алгоритмов была сравнена, и наиболее эффективный подход для прогнозирования цен на золото был выбран.
4. Точность прогнозов была оценена, и были определены факторы, влияющие на цены на золото.
5. Были построены графики и визуализированы результаты прогнозирования цен на золото для дальнейшего анализа и принятия решений.