МИНОБРНАУКИ РОССИИ

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

Кафедра теории функций и стохастического анализа

ИСПОЛЬЗОВАНИЕ НЕЧЕТКИХ РЕГРЕССИОННЫХ МОДЕЛЕЙ ДЛЯ АНАЛИЗА ФИНАНСОВЫХ ДАННЫХ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 412 группы
направления 01.03.02 — Прикладная математика и информатика

механико-математического факультета
Анохина Евгения Николаевича

Научный руководитель
доцент, к. ф.-м. н. _____ Д. В. Мельничук

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

Саратов 2024

Введение.

Анализ финансовых данных является важным инструментом для принятия обоснованных решений в сфере инвестиций, управления рисками и стратегического планирования в бизнесе. В последние десятилетия машинное обучение и статистические методы стали широко применяться в финансовой аналитике для обработки и анализа огромного объема данных, что привело к развитию различных моделей прогнозирования и классификации.

Одним из мощных инструментов, применяемых в анализе финансовых данных, являются нечеткие регрессионные модели. Нечеткая логика позволяет учитывать неопределенность и нечеткость в данных, что особенно важно при работе с финансовыми временными рядами, где могут присутствовать различные источники неопределенности.

В рамках работы предполагается оценка параметров линейной регрессии и построение модели нечеткой регрессии, а также сравнение ее эффективности с традиционными методами моделирования временных рядов, такими как ARIMA, и современными методами машинного обучения, включая LSTM (Long Short-Term Memory) и GRU (Gated Recurrent Unit).

Актуальность бакалаврской работы.

Нечеткие регрессионные модели, основанные на принципах нечеткой логики, предоставляют возможность более точно моделировать и интерпретировать данные, содержащие неопределенности. Это особенно актуально для финансового анализа, где необходимо принимать в расчет размытые данные и экспертные оценки, которые трудно формализовать с помощью традиционных подходов. Применение нечетких регрессионных моделей позволяет не только повысить точность прогнозов финансовых показателей, но и обеспечить большую гибкость в управлении рисками.

Цель бакалаврской работы.

Цель данного исследования заключается в изучении нечеткой регрессионной модели и сравнение эффективности с другими моделями.

Задачи бакалаврской работы являются:

1. Построение регрессии для четких данных
2. Изучить сведения о нечетких числах

3. Построение регрессии для нечетких данных
4. Сравнение эффективности нечеткой регрессии с другими моделями

Структура бакалаврской работы

Работа структурирована следующим образом: введение, три основных раздела, заключение и приложение.

Первый раздел посвящен обзору основных понятий построения регрессии для четких данных. Здесь также рассматриваются теоретические аспекты нечетких чисел и методики построения регрессии для нечетких данных.

Во втором разделе представлен алгоритм регрессионного анализа, разработанный на основе торговых данных. Здесь демонстрируется процесс построения регрессионной модели. Также в этом разделе приведена практическая часть, содержащая графики, которые иллюстрируют применение моделей для случайных величин.

Третий раздел посвящён оценке построенных моделей и анализу полученных результатов. В этой части проводится анализ реальных данных, оценивается эффективность моделей и анализируются результаты их применения

Основное содержание работы

Первый раздел

Построение регрессии для четких данных

Многомерная регрессионная модель(multiple regression model), или *модель множественной регрессии*:

$$y_t = \beta_1 + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + \varepsilon_t, \quad t = 1, \ldots, n,$$

или

$$y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + \varepsilon_t, \quad t = 1, \ldots, n, \tag{1}$$

где x_{tp} - значения регрессора x_p в наблюдении t, а $x_{t1} = 1, \quad t = 1, \ldots, n$. С учетом этого замечания не будем далее различать модели вида (1) со свободным членом или без свободного члена.

Теорема Гаусса-Маркова Предположим, что:

1. $y = X\beta + \varepsilon$;
2. X - детерминированная $n \times k$ матрица, имеющая максимальный ранг

k;

3. $E(\varepsilon) = 0; V(\varepsilon) = E(\varepsilon\varepsilon') = \sigma^2 I_n$.

Тогда оценка метода наименьших квадратов $\hat{\beta}_{OLS} = (X'X)^{-1}X'y$ является наиболее эффективной (в смысле наименьшей дисперсии) оценкой в классе линейных (по y) несмещенных оценок (Best Linear Unbiased Estimator, BLUE).

Оценка дисперсии ошибок σ^2. Распределение s^2

Сумма квадратов остатков $\sum e_t^2 = e'e$ является естественным кандидатом на оценку дисперсии ошибок σ^2 (конечно, с некоторым поправочным коэффициентом, зависящим от числа степеней свободы):

$$E(e'e) = tr(V(e)) = \sigma^2 tr(I_n - N) = (n-k)\sigma^2. \tag{2}$$

При выводе (2) использовали свойства следа матрицы, а также соотношение

$$\begin{aligned} tr(N) &= tr(X(X'X)^{-1}X') \\ &= tr(X'X(X'X)^{-1}) = tr(I_k) = k. \end{aligned} \tag{3}$$

При выводе последнего равенства используется свойство следа матрицы: $tr(AB) = tr(BA)$

Из (2) следует, что

$$s^2 = \hat{\sigma}^2 = \frac{e'e}{n-k} = \frac{\sum e_t^2}{n-k} \tag{4}$$

является несмещенной оценкой дисперсии ошибок σ^2, т.е. $Es^2 = \sigma^2$.

Независимость оценок $\hat{\beta}$ и s^2. В предположении нормальной линейной множественной регрессионной модели удается доказать независимость оценок $\hat{\beta}$ и s^2.

Анализ вариации зависимой переменной в регрессии. Коэффициенты R^2 и скорректированный R^2_{adj}.

Оёпределим коэффициент детерминации R^2 как

$$R^2 = 1 - \frac{ESS}{TSS} = 1 - \frac{e'e}{\hat{y}'_*\hat{y}_*} = \frac{\bar{y}'_*\bar{y}_*}{\hat{y}'_*\hat{y}_*} = \frac{RSS}{TSS}. \tag{5}$$

Отметим, что коэффициент R^2 корректно определен только в том случае, если константа, т.е. вектор $\iota = (1, \ldots, 1)'$, принадлежит линейной оболочке векторов $x_1, \ldots, x_k$. В этом случае R^2 принимает значения из интервала [0,1].

Коэффициент R^2 показывает качество подгонки регрессионной модели к наблюденным значениям y_t.

Если $R^2 = 0$, то регрессия y на $x_1, \ldots, x_k$ не улучшает качество предсказания y_t по сравнению с тривиальным предсказанием $\hat{y}_t = \bar{y}$.

Другой крайний случай $R^2 = 1$ означает точную подгонку: все $e_t = 0$, т.е. все точки наблюдений удовлетворяют уравнению регрессии.

Скорректированным (adjusted) R^2 называется

$$R^2_{adj} = 1 - \frac{e'e/(n-k)}{\hat{y}'_*\hat{y}_*/(n-1)}. \tag{6}$$

Заметим, что нет никакого существенного оправдания именно такого способа коррекции.

Свойства скоректированного R^2:

1. $R^2_{adj} = 1 - (1 - R^2)\frac{(n-1)}{(n-k)}$.
2. $R^2 \geq R^2_{adj}, \quad k > 1$.
3. $R^2_{adj} \leq 1$, но может принимать значения < 0.

Необходимые сведения о нечетких числах

Нечетким множеством A в некотором (непустом) базовом пространстве X называется множество пар

$$A = \{(x, \mu_A(x); x \in X)\}, \tag{7}$$

где $\mu_A : X \to [0, 1]$ - функция принадлежности нечеткого множества A (или иначе характеристическая функция).

Нечеткое число - частный случай нечеткого множества, когда базовым пространством является множество действительных чисел R.

Построение регрессии для нечетких данных

Понятие нечеткости связано с ситуацией, когда возможны разные степени принадлежности (промежуточные между полной принадлежностью и непри-

надлежностью) объекта к некоторому классу

$$Y = A_0 + A_1 x_1 + A_n x_n \tag{8}$$

где $Y = (y_1, \ldots, y_n)^T$, $A_j, j = 0, \ldots, n$ - нечеткие множества, $x_j = (x_{1j}, \ldots, x_{nj})$, $j = 0, \ldots, n$ представлены в четкой форме, $x_0 = 1$.

Формула (8) описывает функцию, являющуюся основной нечеткой линейной регрессионной модели

Задача построения нечеткой линейной регрессионной модели состоит в подборе нечетких параметров $\hat{A}_j, j = 0, \ldots, n$. Критерий выбора этих параметров должен удовлетворять двум условиям:

1. Получаемые на основе нечетких параметров нечеткие множества $\hat{y}_i = \hat{A}_0 + \hat{A}_1 x_{i1} + \ldots + \hat{A}_n x_{in}$ должны содержать четкие наблюдаемые значения y_1^0 со степенью достоверности не меньшей, чем некоторая заданная степень h.
2. Общая нечеткость модели должна быть минимальна.

Некоторые дополнительные модели прогнозирования

ARIMA (Авторегрессионная интегрированная модель скользящего среднего)

ARIMA — это модель временных рядов, используемая для анализа данных и прогнозирования в статистике, экономическом и финансовом моделировании.ARIMA моделирует несколько аспектов временных рядов, учитывая тенденции, циклы, сезонность и шум. Модель описывается тремя параметрами: (p, d, q):

1. p — порядок компоненты авторегрессии, отражает количество лагов переменной, используемых как предикторы.
2. d — порядок интеграции, представляет собой количество раз, которое временной ряд должен быть дифференцирован, чтобы стать стационарным.
3. q — порядок скользящего среднего, определяет количество лагов ошибки прогноза, используемых в модели.

Модель ARIMA хорошо подходит для анализа рядов, которые показывают широкие колебания и не имеют строгих сезонных паттернов. ARIMA

также может включать сезонные компоненты, превращаясь в SARIMA (сезонная ARIMA).

LSTM (Долгосрочная краткосрочная память)

LSTM — это тип рекуррентной нейронной сети (RNN), специально разработанный для решения проблемы исчезающего градиента, с которой сталкиваются традиционные RNN. LSTM особенно подходит для задач, где важно улавливать долгосрочные зависимости в данных. Они широко используются для прогнозирования временных рядов, обработки естественного языка, распознавания речи и других задач.

Структурно LSTM состоит из ячеек с тремя "вратами":

1. Ворота забывания контролируют, какая информация будет удалена из ячейки.
2. Ворота входа определяют, какая новая информация добавляется в состояние ячейки.
3. Ворота вывода решают, какая информация будет передана в следующий слой сети.

Эти вратные структуры позволяют LSTM более эффективно управлять потоком информации, что делает их идеальными для задач с комплексными и долгосрочными временными зависимостями.

GRU (Управляемые рекуррентные блоки)

GRU (Gated Recurrent Unit) — это другой тип рекуррентной нейронной сети, похожий на LSTM, но более простой в структуре и быстрее в обучении. GRU был предложен, чтобы сделать каждый рекуррентный блок более эффективным, объединяя ворота забывания и входа в одно "обновляющее ворото".

GRU состоит из двух врат:

1. Обновляющие ворота определяют, какая прошлая информация должна сохраняться.
2. Сбросные ворота решают, как много предыдущей информации забыть.

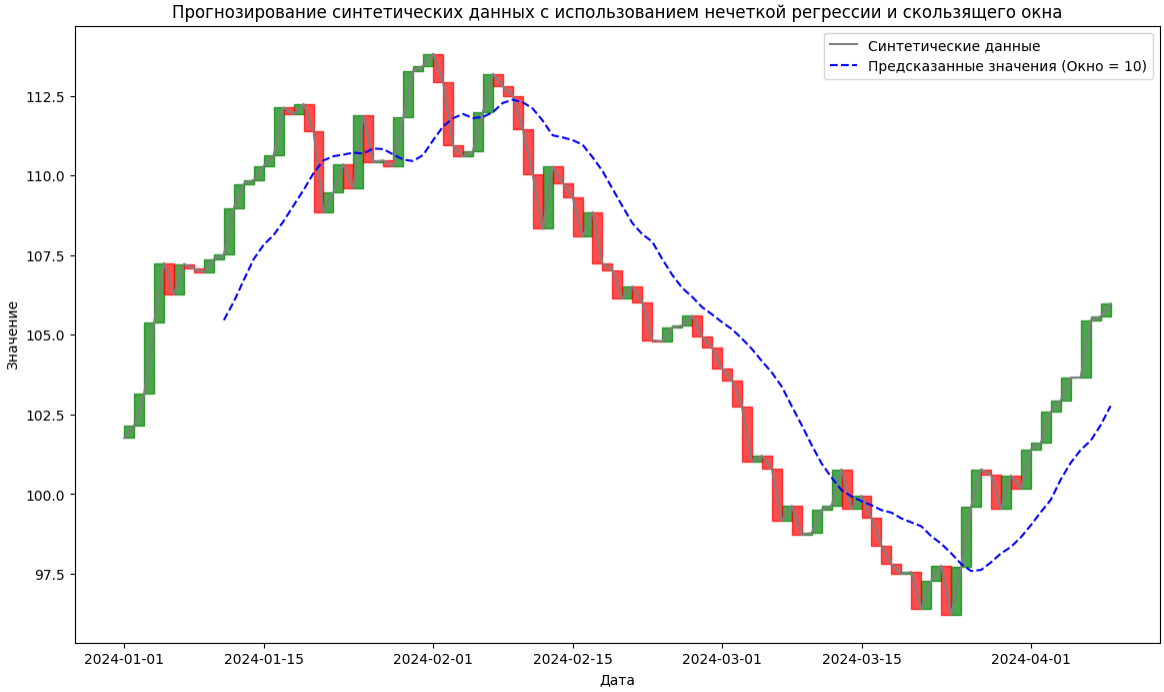
Такая структура позволяет GRU более эффективно управлять информацией и делает его особенно подходящим для тех случаев, когда модель LSTM может оказаться излишне сложной.

Второй раздел

Применение моделей для прогназирования случайных величин

Прогнозирование синтетических данных временного ряда с использованием нечеткой регрессии и скользящего окна. Синтетические данные представляют собой последовательность случайных значений, что делает их идеальными для тестирования методов прогнозирования в условиях неопределенности.

Основным методом прогнозирования в данной работе является нечеткая регрессия, применяемая к данным с использованием скользящего окна. Размер окна определяет количество предыдущих значений, используемых для прогнозирования следующего значения временного ряда. Mean Squared Error (MSE): 4.27

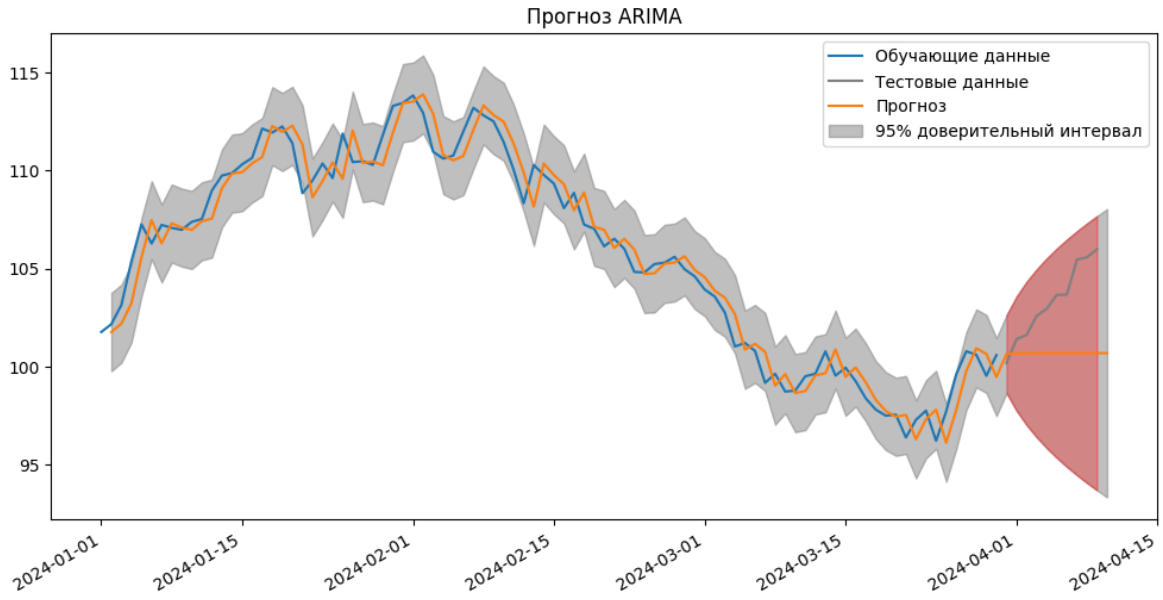


Применение модели ARIMA для прогнозирования случайных величин. В данной работе использован синтетический временной ряд, сгенерированный как накопленная сумма случайных изменений с начальным смещением, имитирующим тренд. Данные охватывают 100 дней с начала 2024 года. Временной ряд разделён на обучающую выборку (90 дней) и тестовую (10 дней), что позволяет оценить способность модели к прогнозированию.

Модель ARIMA с параметрами (1, 1, 1) была обучена на основе анализа стационарности и автокорреляции данных. Параметры указывают на исполь-

зование одного авторегрессионного лага, однократное дифференцирование и один лаг скользящего среднего.

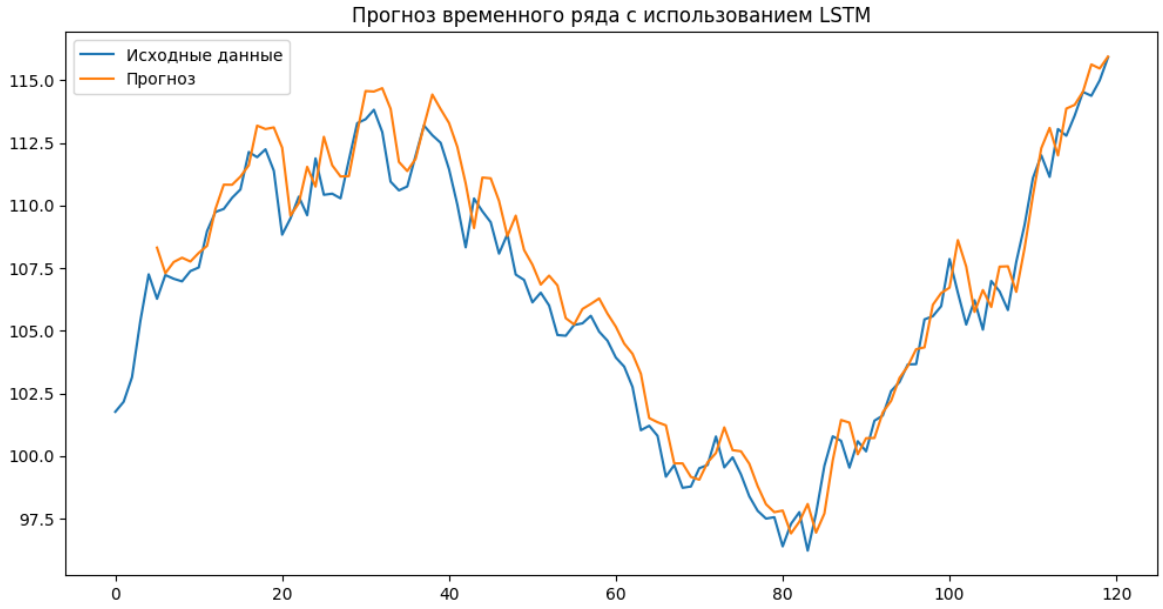
С помощью модели был сделан прогноз на 10 дней вперёд с расчётом среднеквадратичной ошибки (MSE), что позволило оценить точность прогнозов. График показывает обучающие и тестовые данные, прогнозы и 95% доверительные интервалы, демонстрируя эффективность модели.
Mean Squared Error (MSE): 10.27



Применение модели LSTM для прогнозирования временных рядов. Исходные данные были нормализованы в диапазоне от 0 до 1 с использованием MinMaxScaler для улучшения сходимости процесса обучения нейронной сети. Модель состоит из одного слоя LSTM с 50 нейронами и одного плотного слоя на выходе. В качестве функции потерь использовалась среднеквадратичная ошибка (mean squared error), оптимизатор — Adam. Модель обучалась на данных в течение 100 эпох с размером батча 1.

Модель использовалась для генерации прогнозов на основе обучающего набора данных. Предсказанные значения были преобразованы обратно в исходный масштаб с помощью inverse_transform для сравнения с реальными данными. На графике показаны исходные данные и результаты прогноза, начиная с индекса, равного look_back. Среднеквадратичная ошибка (MSE) между реальными данными (начиная с пятой точки, т.е. после первого look_back периода) и предсказаниями составила 1.4998, что дает оценку

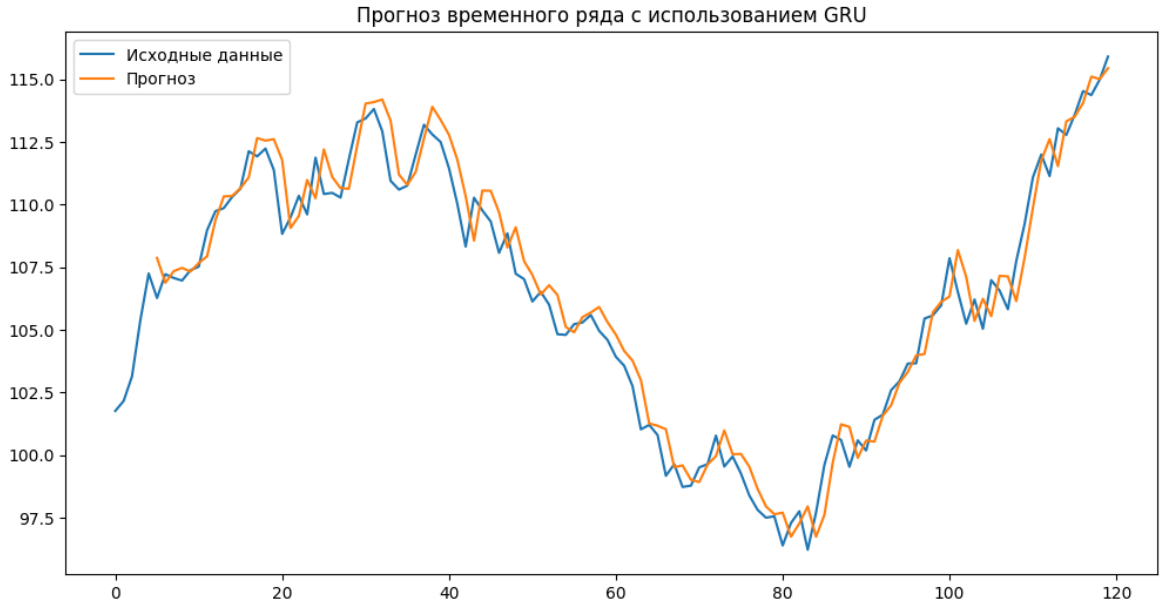
точности модели.



Применение модели GRU для прогнозирования временных рядов.

Модель GRU была построена с одним слоем из 50 единиц и одним плотным выходным слоем. GRU (Gated Recurrent Unit) является разновидностью рекуррентных нейронных сетей, которая эффективно обрабатывает временные зависимости благодаря механизмам обновления и сброса состояний. Модель обучалась на данных в течение 100 эпох с размером пакета 1, что обеспечивает подробное обновление весов на каждом шаге.

После обучения модель использовалась для генерации прогнозов. Прогнозные значения были преобразованы обратно к исходному масштабу для наглядного сравнения с реальными данными. На графике отображены как исходные, так и прогнозируемые значения, показывающие эффективность модели в воспроизведении и прогнозировании временных рядов. Для оценки точности модели использовалась среднеквадратичная ошибка (MSE), которая в данном случае составила 1.1018.

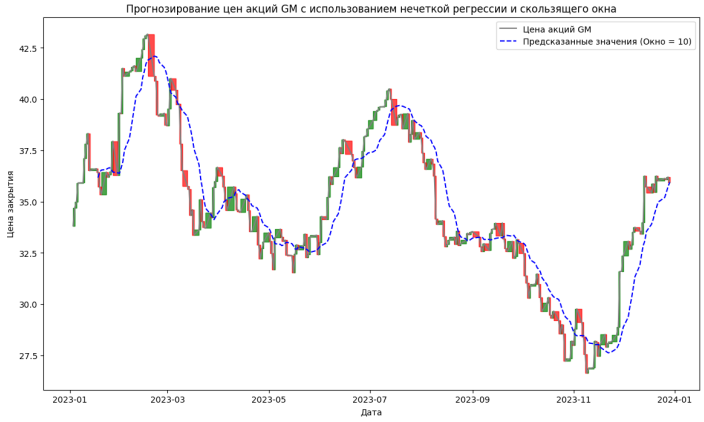


Третий раздел

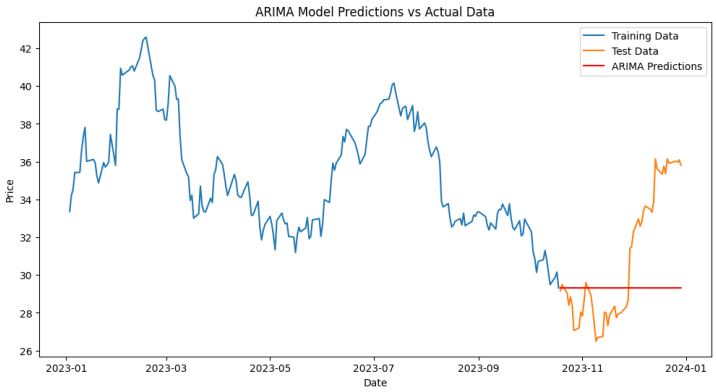
Применение моделей на реальных данных

Применение нечёткой регрессии для прогнозирования цен на акции General Motors. В анализе использованы данные об акциях General Motors (GM), загруженные с помощью библиотеки yfinance за период с 1 января 2023 года по 31 декабря 2023 года. Основой для анализа служат скорректированные цены закрытия. Прогнозные значения визуализированы на графике вместе с исходными данными о ценах акций, позволяя визуально оценить адекватность и точность нечёткой регрессии и метод скользящего окна. Также была рассчитана среднеквадратичная ошибка (MSE), которая показывает величину ошибок между фактическими и прогнозными значениями.

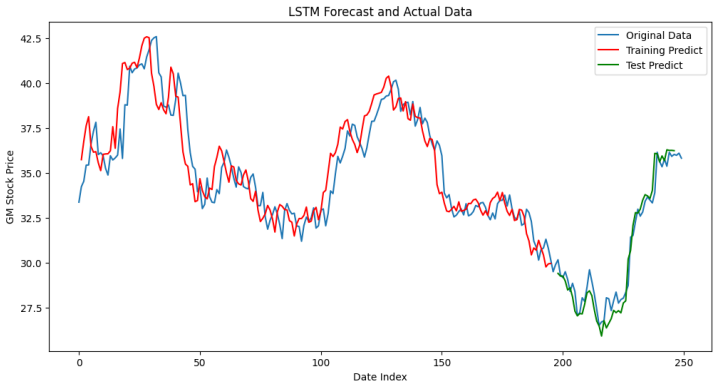
Mean Squared Error (MSE) of the Fuzzy Regression: 2.3833



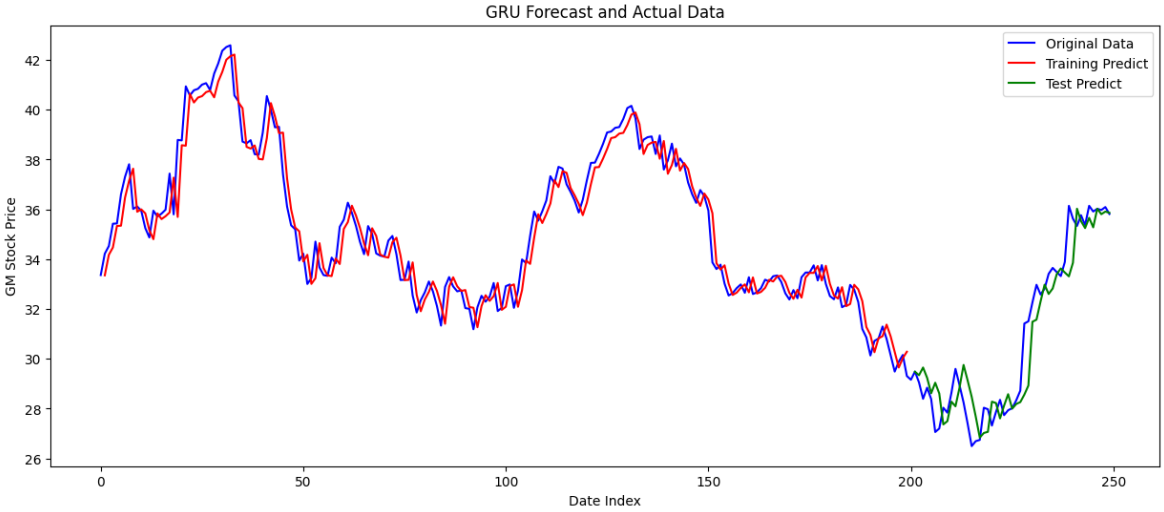
Применение модели ARIMA для прогнозирования цен акций General Motors. Данные были разделены на обучающий и тестовый наборы. Обучающий набор составил 80% от общего объема данных, что позволило оставить последние 20% данных для проверки прогноза модели. Прогнозы модели ARIMA были сравнены с фактическими значениями тестового набора данных. Прогнозы корректно отражают тенденции и колебания рынка, что видно на графике визуализации. Среднеквадратичная ошибка (MSE) была рассчитана для оценки точности модели и составила 13.4495. На графике отображены данные обучающего набора, тестового набора и прогнозы модели. Графическое представление позволяет визуально оценить качество прогнозов и их соответствие реальным данным рынка.



Применение модели LSTM для прогнозирования цен акций General Motors. Для обеспечения более эффективного обучения и предотвращения проблем с сходимостью модели цены акций были нормализованы в диапазон от 0 до 1 с помощью MinMaxScaler. Данные были преобразованы в формат, пригодный для анализа временных рядов, с использованием функции create_dataset, которая создаёт пары входных и выходных данных для обучения. Входные данные представляли собой последовательности из пяти предыдущих значений (look_back=5), что позволило улучшить качество прогнозов. Обучающая выборка составила 80% от всего набора данных, оставшиеся 20% были использованы для тестирования модели. На графике представлены исходные данные о ценах акций, а также прогнозы модели для обучающей и тестовой выборок. Визуальное представление позволяет оценить адекватность модели в воспроизведении и прогнозировании динамики цен. MSE на тестовой выборке: 1.01



Применение модели GRU для прогнозирования цен акций General Motors.Цены акций были нормализованы в диапазон от 0 до 1 с использованием MinMaxScaler, что необходимо для эффективного обучения модели GRU, улучшения сходимости и предотвращения проблем с чувствительностью к масштабу входных данных. Используя функцию create_dataset, из цен акций был создан датасет для временного ряда, где каждое последующее значение предсказывается на основе предыдущего значения (look_back=1). Это позволяет модели учиться на зависимости текущей цены от предыдущей. На графике представлены оригинальные данные о ценах акций и предсказания модели для обучающего и тестового наборов. Визуализация помогает наглядно оценить эффективность модели в задаче прогнозирования цен акций и показывает, как предсказания модели совпадают с фактическими изменениями в ценах. MSE на тестовом наборе: 0.54



Заключение В бакалаврской работе было проведено исследование нечетких регрессионных моделей, а также выполнено сравнение их эффективности с классическими моделями ARIMA и современными нейросетевыми методами, такими как LSTM и GRU.