**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ**  
**федеральное государственное автономное образовательное учреждение высшего**

**образования «Балтийский федеральный университет имени Иммануила Канта»**

**ОНК «Институт высоких технологий»**

КУРСОВАЯ РАБОТАПО ДИСЦИПЛИНЕ

«Интерпретируемое машинное обучение»

НА ТЕМУ «Прогнозирование даты достижения поставленной  
 цели накопления средств на основе нейронных сетей»

Выполнил Волков Илья Евгеньевич \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

студент очной формы обучения 1-го курса

направления подготовки 02.04.03   
«Математическое обеспечение и администрирование   
информационных систем»

профиль обучения: «Банковские информационные технологии»

Руководитель Ткаченко Сергей Николаевич \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

к.т.н. доцент ОНК «ИВТ» БФУ им. И. Канта

Калининград 2023

Содержание

[Введение 3](#_Toc134544369)

[Глава 1. Изучить способы и инструменты для решения задачи 5](#_Toc134544370)

[1.1 Формулировка задачи 5](#_Toc134544371)

[1.2 Временные ряды и методы их анализа. 6](#_Toc134544372)

[1.3 Методы машинного обучения для предсказания временного ряда. 7](#_Toc134544373)

[1.3.1. Линейная регрессия (Linear Regression). 8](#_Toc134544374)

[1.3.2. XGBRegressor 9](#_Toc134544375)

[1.4. Нейронные сети LSTM 9](#_Toc134544376)

[1.4.1. Основные определения LSTM. 9](#_Toc134544377)

[1.4.2. Основная идея LSTM. 11](#_Toc134544378)

[Глава 2 Реализация 12](#_Toc134544379)

[2.1 Подготовка и обработка данных для обучения моделей 12](#_Toc134544380)

[2.2 Алгоритмы машинного обучения (LinearRegression, XGBRegressor, SARIMAX) 13](#_Toc134544381)

[2.2.1 LinearRegression и XGBRegressor. 14](#_Toc134544382)

[2.2.2 SARIMAX 17](#_Toc134544383)

[2.3 Нейронная сеть LSTM 18](#_Toc134544384)

[2.4 Обертка модели 21](#_Toc134544385)

[Заключение 23](#_Toc134544386)

[Список источников 24](#_Toc134544387)

Введение

Рано или поздно в жизни каждого человека возникает такая ситуация. Когда ему что-то приглянулось, заинтересовало или попросту надо взять, но денег на данный момент у него немного или просто не хватает. Откуда же ему их взять? Первое что приходит в голову, это кредит или взять взаймы недостающую сумму, но её в последствии нужно будет вернуть и возможно с процентами. Вторая мысль же - это накопить. Плюсом данного подхода является, то что деньги принадлежат этому же человеку и возвращать их никому не нужно. Однако минус перекрывает плюс, поскольку нужно будет ждать пока сумма не накопится. А вот сколько, не известно. Поэтому ему нужно посчитать сколько дней понадобиться для достижения поставленной цели.

В рамках данной курсовой работы рассматривается метод решения данной проблемы. Однако какая проблема в простом подсчете суммы, которую нужно откладывать каждый месяц, день, или неделю. Данную сумму нужно просто разделить на сумму накопления и получить нужное количество времени. Но это линейный способ подсчета, и он не учитывает сторонние ситуации. Скажем, поход в больницу, уплата штрафов и так далее. Поэтому в качестве решения данной проблемы будет использование искусственного интеллекта, с помощью которого можно будет предугадать подобные ситуации.

**Целью** курсовой работы является разработка искусственного интеллекта, при помощи которого можно будет предсказать приблизительную дату достижения поставленной цели накопления средств.

Для достижения цели были поставлены следующие **задачи**:

1. Изучить способы решения задачи
2. Подобрать инструменты для ее реализации
3. Найти подходящие данные для обучения модели
4. Реализовать модель
5. Написать обертку модели

Глава 1. Изучить способы и инструменты для решения задачи

1.1 Формулировка задачи

Рассмотрим задачу более подробно. Как упоминалось ранее целью работы является предсказание приблизительной даты достижения цели накопления. Однако дополнительным условием является, что модель должна предполагать появление непредвиденных расходов.

Поставим задачу непредвиденных расходов немного иначе. Предположим, что через некоторое время данная ситуация может повториться. Тогда необходимо рассмотреть условия возникновения предыдущего случая. Таким образом можно представить данные как временной ряд, на котором с равным промежутком времени, скажем в один день расположены данные. Основываясь на этих исторических данных, мы сможем предсказать подобные ситуации.

В качестве данных временного ряда будут использоваться данные о доходах и расходах человека за месяц, расписанные по дням. Однако в качестве предсказанного значение по временному ряду будет выступать одно значение – разница между доходом и расходом.

Поскольку человек не обязательно будет класть всю сумму, а только её часть, ведь жить ему тоже на что-то надо, то введем еще один параметр – процент, который человек будет отдавать с положительной разницы между доходом и расходом. Если же эта разница будет отрицательной, то человек не будет класть деньги на счет.

В результате получается следующая задача. Прогнозирование даты достижения заданной цели накопления с учетом возможных непредвиденных расходов с использованием модели предсказания по временному ряду, основанному на исторических данных о доходах и расходах пользователя за месяц, расписанных по дням. Модель предсказывает разницу между доходом и расходом на день вперед, и на основе этой разницы, и процента, который будет откладываться пользователем на достижение цели, будет предсказываться исходная дата накопления.

1.2 Временные ряды и методы их анализа.

В качестве входных данных для модели был взят временной ряд.

Временной ряд - это последовательность наблюдений за определенным параметром в разные моменты времени. Таким образом, временной ряд содержит информацию о том, как изменяется параметр со временем [1].

В качестве примеров временного ряда могут выступать данные о продажах продукции в определенный день или данные о температуре на определенной территории в различные временные промежутки. Цель анализа временных рядов - определить закономерности в изменениях параметра во времени и сделать прогноз на будущее.

Анализ временных рядов является важным инструментом для многих областей, в том числе бизнеса, науки, технологий и экономики. Методы анализа временных рядов помогают понять поведение и изменение временных данных и выявить скрытые тенденции и закономерности. Существует три основных метода анализа временных рядов: стационарность, автокорреляция и спектральный анализ.

Стационарность - это свойство временного ряда, которое означает, что его средние и стандартные отклонения не меняются со временем. Если временной ряд является стационарным, то его можно легко анализировать и прогнозировать. Нестационарный временной ряд может иметь тренд (постоянный рост или падение), цикличность (повторение циклов) или сезонность (повторение определенных событий в разное время года).

Автокорреляция - это мера корреляции между значениями ряда с разницей во времени. Если временной ряд имеет высокую автокорреляцию, это означает, что значения ряда в разные периоды времени имеют сильную связь между собой.

Спектральный анализ - позволяет исследовать частотную составляющую ряда. Он используется для выявления скрытых цикличных паттернов во временных рядах. Для этого временной ряд разбивается на сигналы разных частот, а затем анализируется спектр этих частот.

Таким образом, временные ряды являются одним из важных компонентов современной аналитики данных и имеет большие практические применения в различных областях. В рамках курсовой работы временные ряды пригодятся для выявления закономерностей поведения разности между доходом и расходом человека.

1.3 Методы машинного обучения для предсказания временного ряда.

Для предсказания и выявления закономерностей временного ряда будут использоваться следующие методы машинного обучения [1]:

1. Регрессия - это метод машинного обучения, который моделирует зависимость между независимыми (входными) и зависимыми (выходными) переменными. В случае прогнозирования временных рядов, независимыми переменными будут значения ряда в прошлых периодах, а зависимой переменной - будущие значения ряда.
2. Нейронные сети - это более сложные методы машинного обучения, которые используют алгоритмы обработки информации, подобные тем, которые используют мозг. Нейронные сети могут использоваться для прогнозирования временных рядов и обычно имеют более высокую точность, чем методы экстраполяции.
3. SARIMA - это метод прогнозирования, который использует модель SARIMA (Seasonal AutoRegressive Integrated Moving Average). Эта модель объединяет в себе методы авторегрессии, интегрированного скользящего среднего и сезонности, что позволяет прогнозировать будущие значения ряда на основе его предыдущих значений с учетом сезонных паттернов.

Авторегрессия означает зависимость последующего значения параметра от предыдущих. Зависимость в случае авторегрессии предполагается линейная, то есть прогноз представляет собой сумму от параметра за предыдущие дни с некоторыми коэффициентами, которые являются постоянными и определяют параметры модели авторегрессии [2].

Скользящая средняя показывает среднее значение параметра за указанный период времени и является запаздывающим индикатором. Простая скользящая средняя — это просто среднее значение параметра за определенный период времени [3].

В данной работе рассматриваются следующие варианты регрессии:

1.3.1. Линейная регрессия (Linear Regression) [4].

Линейная регрессия — это метод анализа данных, который предсказывает ценность неизвестных данных с помощью другого связанного и известного значения данных. Он математически моделирует неизвестную или зависимую переменную и известную или независимую переменную в виде линейного уравнения.

Например, предположим, что у вас есть данные о ваших расходах и доходах за прошлый год. Методы линейной регрессии анализируют эти данные и определяют, что ваши расходы составляют половину вашего дохода. Затем они рассчитывают неизвестные будущие расходы, сокращая вдвое будущий известный доход.

1.3.2. XGBRegressor [5]

XGBRegressor - это алгоритм машинного обучения, основанный на градиентном бустинге деревьев решений. Он является реализацией градиентного бустинга на основе деревьев решений с использованием библиотеки XGBoost.

Градиентный бустинг - это метод построения ансамблей моделей, когда каждая последующая модель в ансамбле учится исправлять ошибки предыдущей модели. В случае XGBRegressor, каждая модель является деревом решений, которое строится по обучающей выборке. Градиентный бустинг позволяет получать более точные прогнозы, чем отдельные деревья решений.

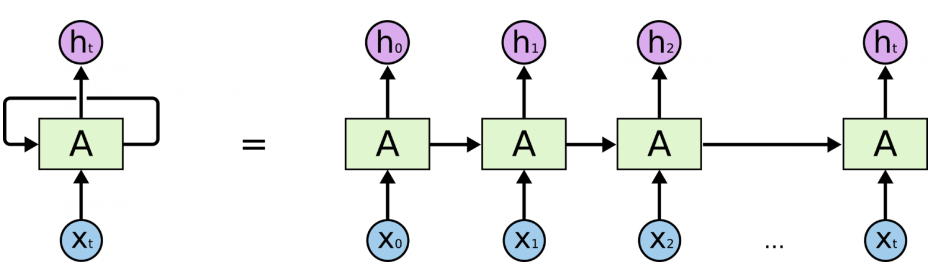
1.4. Нейронные сети LSTM

1.4.1. Основные определения LSTM.

В качестве модели нейронной сети была выбрана сеть LSTM.

Долгая краткосрочная память (Long short-term memory; LSTM) – особая разновидность архитектуры рекуррентных нейронных сетей, способная к обучению долговременным зависимостям [6].

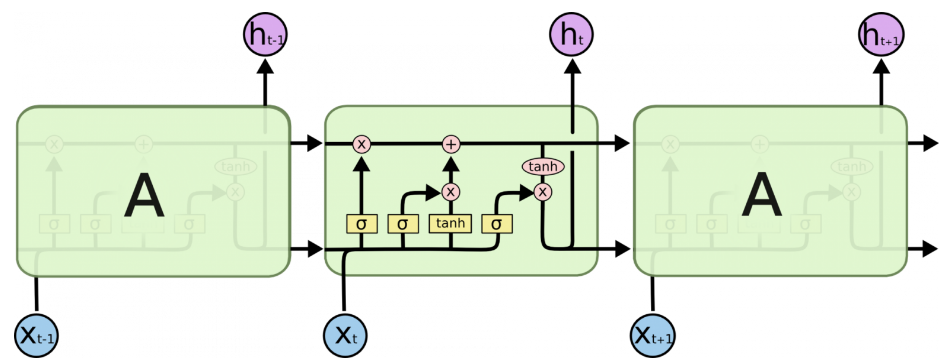
Рекуррентые нейронные сети (Recurrent Neural Networks, RNN). Это сети, содержащие обратные связи и позволяющие сохранять информацию [6].



**Рис 1.** Рекуррентная нейронная сеть (RNN) в развертке

LSTM разработаны специально, чтобы избежать проблемы долговременной зависимости. Данная проблема заключается в следующем: разрыв между актуальной информацией и точкой ее применения может стать очень большим. Поэтому по мере роста этого расстояния рекуррентные сети теряют способность связывать информацию.

Структура LSTM напоминает цепочку, где модули выглядят иначе чем у RNN. Вместо одного слоя нейронной сети они содержат целых четыре, и эти слои взаимодействуют особенным образом.



**Рис 2.** Повторяющийся модель в LSTM сети состоит из четырех взаимодействующих слоев.

1.4.2. Основная идея LSTM.

Ключевой компонент LSTM – это состояние ячейки (cell state) – горизонтальная линия, проходящая по верхней части схемы.

LSTM может удалять информацию из состояния ячейки; этот процесс регулируется структурами, называемыми фильтрами (gates).

Сигмоидальный слой возвращает числа от нуля до единицы, которые обозначают, какую долю каждого блока информации следует пропустить дальше по сети. Ноль в данном случае означает “не пропускать ничего”, единица – “пропустить все”.

В LSTM три таких фильтра, позволяющих защищать и контролировать состояние ячейки.

Конкретно в этой работе сети LSTM используются из-за того, что необходимо помнить информацию, предсказанную ранее, и на основе поступающих и предыдущих данных выявлять закономерности временного ряда.

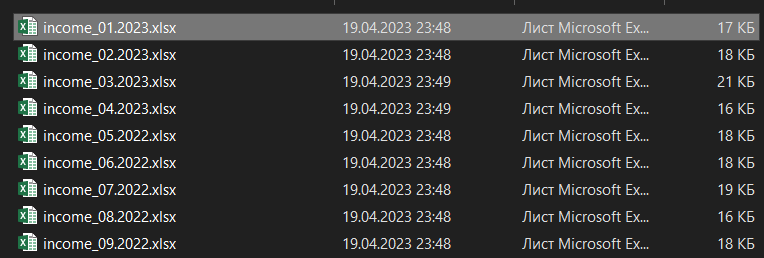
Таким образом, в данной главе была подробно разобрана исходная задача, рассмотрены основные подходы к прогнозированию временных рядов, а также исследованы методы разработки искусственного интеллекта на основе нейронных сетей LSTM.

Глава 2 Реализация

2.1 Подготовка и обработка данных для обучения моделей

В качестве данных были взяты свои показатели доходов и расходов за прошедший год, а конкретно с 2022-05-01 по 2023-04-19.

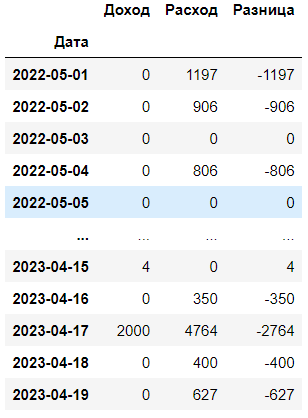
Показатели взяты из отчетов, предоставленных Сбербанком из личного кабинета. Они изначально представлены следующим образом. Первая колонка — это дата, вторая — тип операции, третья — категории выплат и так далее. Кроме того, данные изначально находятся в разных файлах, которые распределены по месяцам.



**Рис 3**. Иерархия данных

В последствии данные были обработаны следующим образом. Произведено слияние данных в один период, убраны излишние данные по типу описание, состояние, валюта и так далее, а оставлены только сумма доходов и расходов. Дополнительно была высчитана разница между доходом и расходом, далее будет упоминаться как просто разница, и записана в соответствующую колонку “Разница”. Результат обработки представлен на рисунке 4.

Изобразим искомую разницу в виде графика по датам, смотреть рисунок 5. Проанализировав этот график можно прийти к выводу что данный ряд является крайне нестабильным, однако на графике виднеется некоторая сезонность, тренда как такового не наблюдается на всем промежутке, однако данное утверждение может быть поспешным, поэтому далее будет рассмотрен этот ряд более подробно.



**Рис 4**. Результат обработки данных



**Рис 5** График разницы между доходом и расходом

2.2 Алгоритмы машинного обучения (LinearRegression, XGBRegressor, SARIMAX)

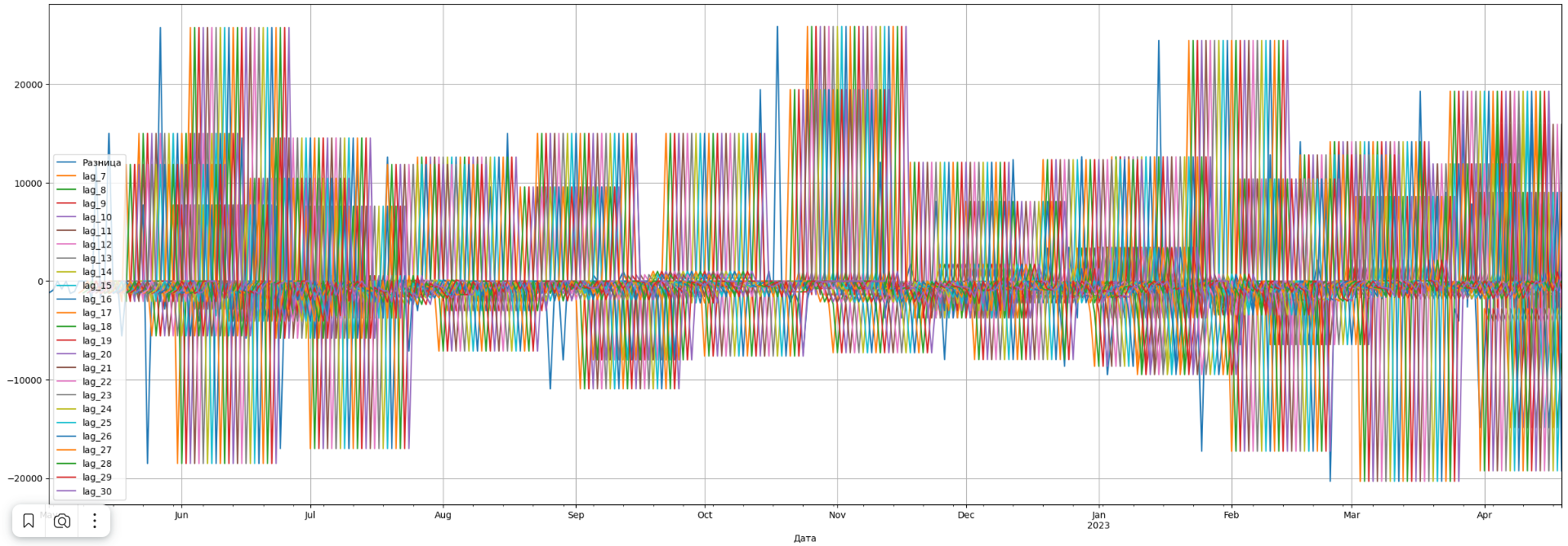
Прежде чем приступить к обучению нейронной сети, нужно понять, а нужно ли ее использовать, может с задачей справятся алгоритмы машинного обучения, такие как LinearRegression, XGBRegressor, SARIMAX, и нейронная сеть не понадобится.

По имеющимся предположениям, рассмотренные выше, попробуем посмотреть и проанализировать ряд при помощи нескольких алгоритмов машинного обучения.

2.2.1 LinearRegression и XGBRegressor.

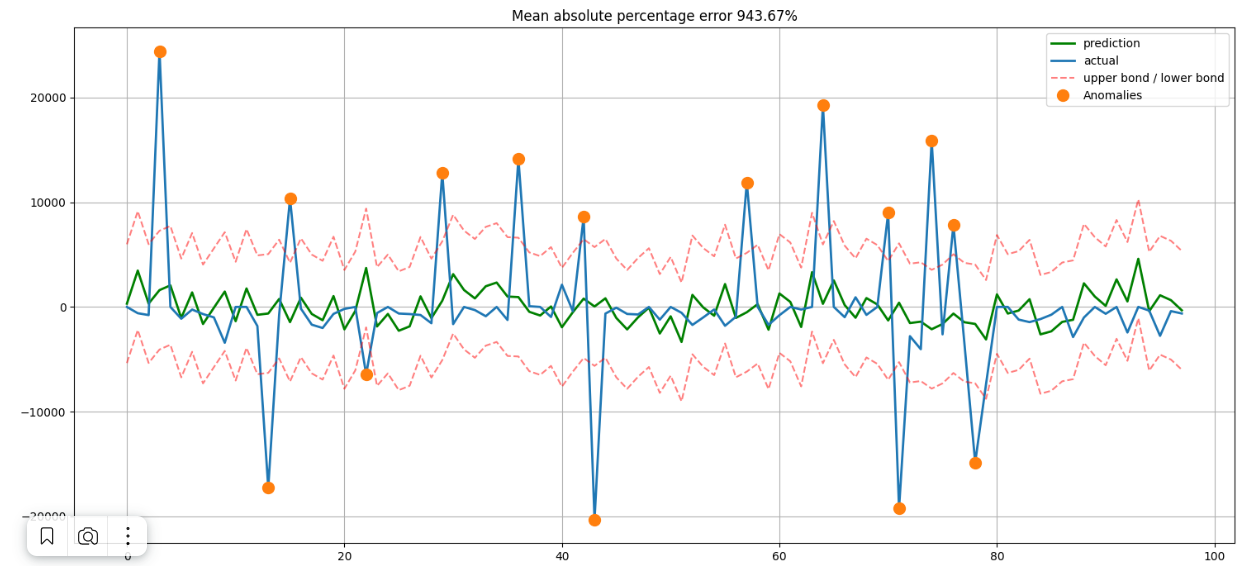
Для начала оставим данными нетронутыми и проведем анализ при помощи алгоритмов LinearRegression и XGBRegressor.

Поскольку эти алгоритмы опираются на предыдущие данные, то нужно добавить сдвиги или лаги на равномерное количество измерений. В нашем случае был проведен сдвиг от 7 до 30 дней в виде лага.

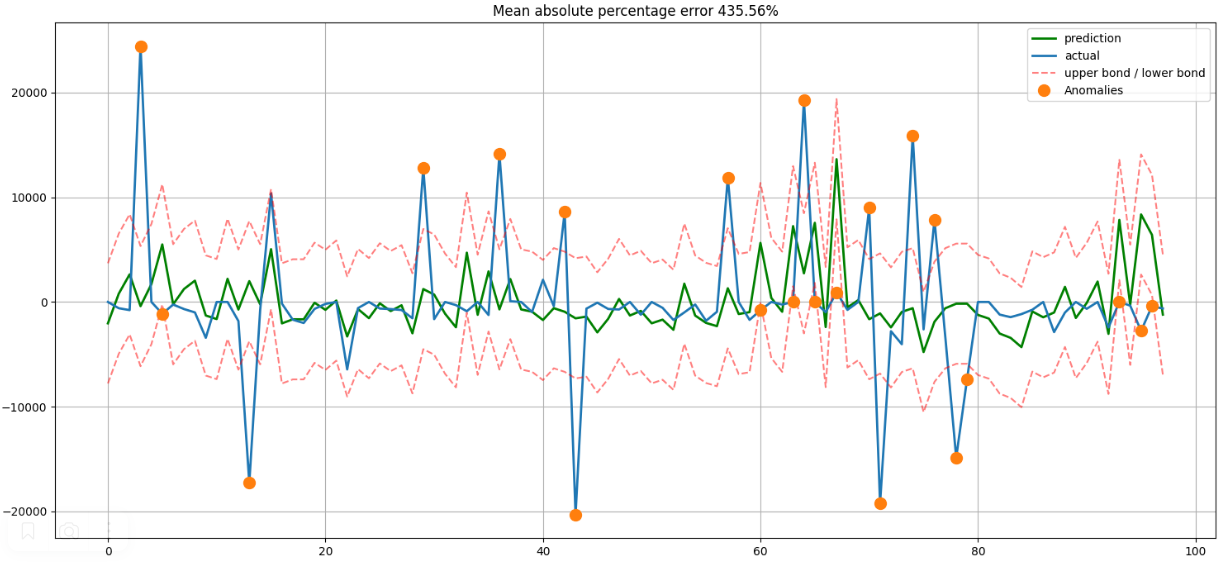


**Рис 6**. График лагов

После проведения обучения LinearRegression и XGBRegressor с их параметрами по умолчанию, получаеются следующие результаты по предсказанию будущего значении разницы, представленные на рисунке 7 и рисунке 8.



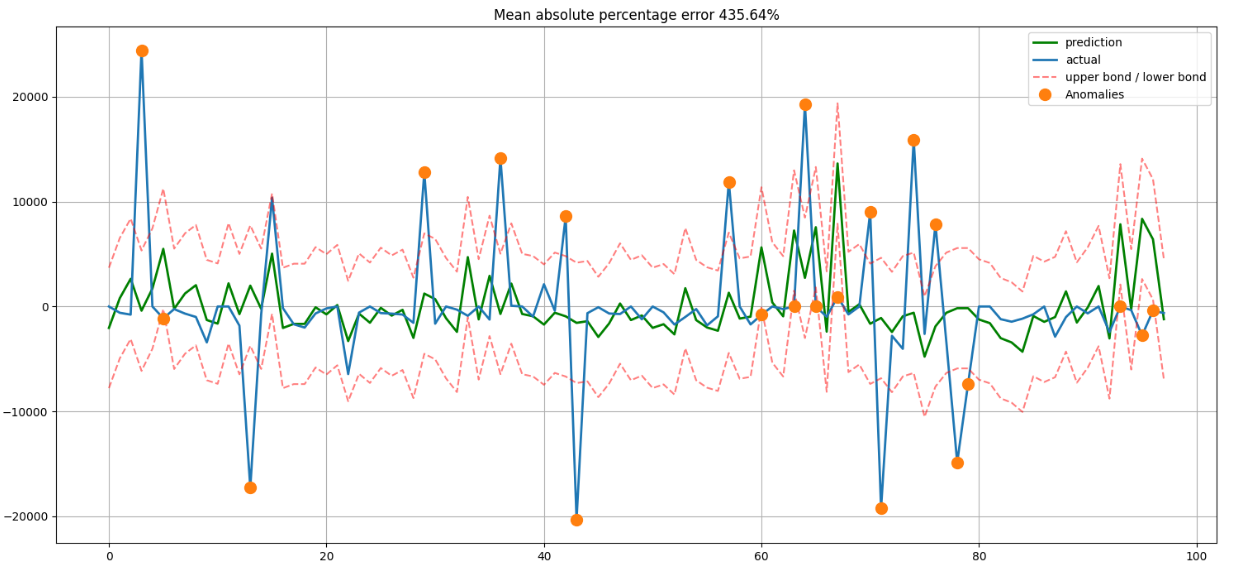
**Рис 7** Разница, предсказанная LinearRegression



**Рис 8**. Разница, предсказанная XGBRegressor

Как видно из графиков на рисунке 7 и рисунке 8, такие данные являются крайне не валидными, так как наблюдается много аномалий в виде незапланированных трат или дохода в виде подарка на день рождение. Однако можно сделать вывод что модель XGBRegressor предсказывает более точно нежели LinearRegression.

Проведем стандартизацию данных при помощи метода StandartScaler. Таким образом уменьшим разрывы между значениями в данных и приведем к более общему виду. В результате чего алгоритмам должно стать легче предсказать будущие значения.



**Рис 9**. Разница, предсказанная XGBRegressor с использованием StandartScaler

В результате обобщения данных получили следующий результат, точность предсказания снизилась на сотый процент, то есть модели стали предсказывать чуть хуже, чем было, что можно наблюдать по лучшей модели XGBRegressor на рисунке 9.

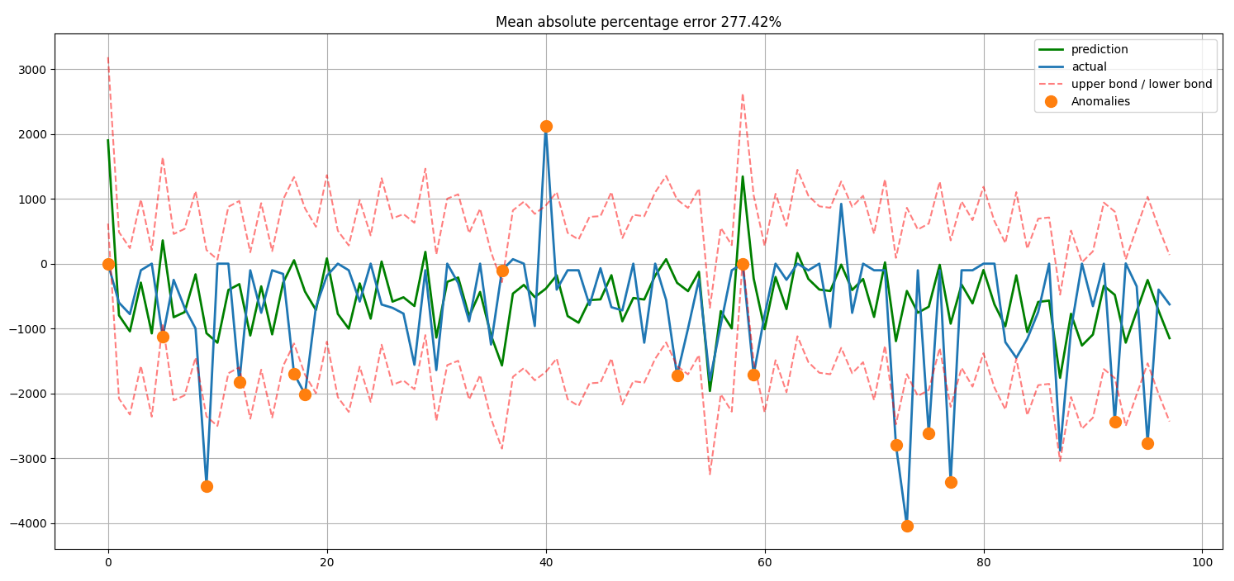
Попробуем убрать аномалии из данных, а именно заменить их на среднее значение в колонке. В результате получается, что избавились от части данных которые могут быть важны, и поэтому уже идет относительный прогноз.

Выведем общие статистические показатели для исходных данных, смотреть рисунок 10. На основе них будем производить поиск и замену значений найденных аномалий. Поиск осуществляем следующим образом, если значение больше или меньше среднего значения сложенного или вычтенного со средним квадратичным отклонением, то найдена аномалия, которая заменяется средним значением.



**Рис 10.** Общие статистические показатели для исходных данных

В результате отсеивания аномалий получили что XGBRegressor стал лучше предсказывать значения, однако это все еще далеко от допустимых значений, смотреть рисунок 11.

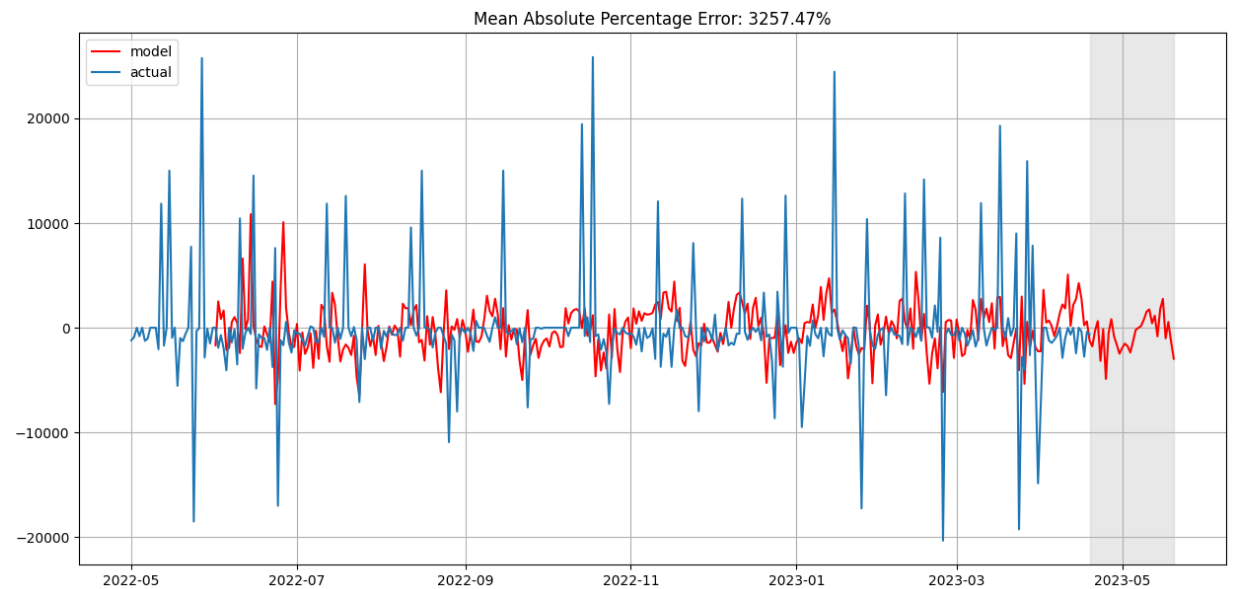


**Рис 11**. Разница, предсказанная XGBRegressor после отсеивания аномалий.

2.2.2 SARIMAX

Проведем еще один анализ при помощи модели SARIMAX. Исследовав разные выше описанные закономерности в данных, были выявлены следующие признаки сезонность повторяется каждые 30 дней, порядок модели авторегрессии равен 20, порядок скользящего среднего равен 1, порядок сезонной составляющей авторегрессии равен 0, порядок сезонной составляющей скользящего среднего равен 1.

В результате обучения SARIMAX, получался результат, представленный на рисунке 12.



**Рис 12** Разница, предсказанная SARIMAX.

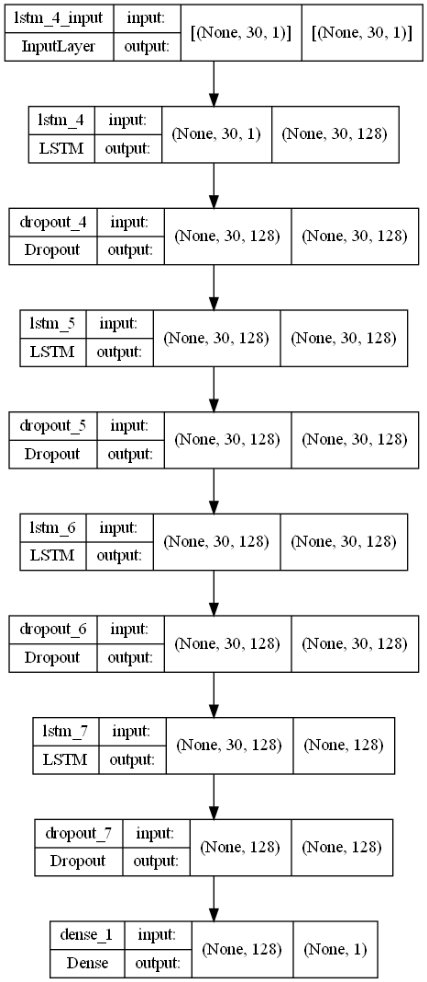
В результате обучения всех моделей получается, что модели могут повторить поведение данного ряда, однако аномалии они не могут предсказать, в нашем же случае аномалиями является дополнительное условие задачи, а именно незапланированные расходы пользователя. Таким образом лучшее значение, которого смогли добиться модели, около 300 процентов точности.

2.3 Нейронная сеть LSTM

Так как алгоритмы машинного обучения не смогли предсказать будущие значения ряда до приемлемого значения точности, то воспользуемся нейронной сетью LSTM.

Предположим, что каждый слой должен обрабатывать определенную часть ряда. Скажем два из них должны определить куда будет отклоняться значение: в положительную или отрицательную стороны, остальные два слоя определят величину отклонения: либо приемлемый расход, либо непредвиденный, в качестве аномалии. В результате получили по 4 слоя, на этом и остановимся.

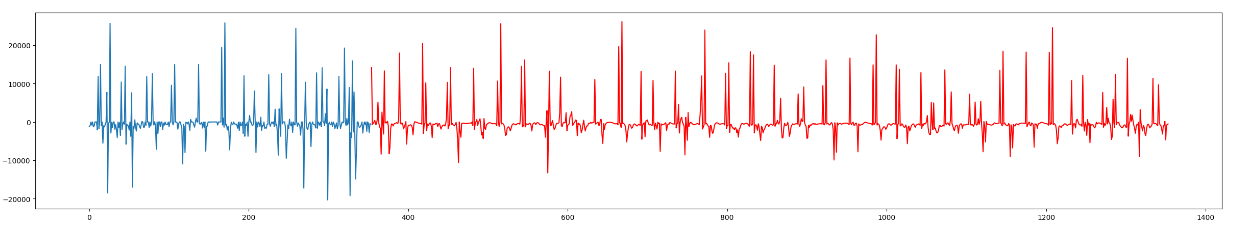
В результате экспериментальных наблюдений и методом проб и ошибок удалось разработать следующую модель нейронной сети LSTM, смотреть рисунок 13.



**Рис 13**. Граф нейронной сети LSTM

На вход сети подаются значения разности за прошедшие 30 дней, а результатом сети выводится значение разности в 1 день. Таким образом модель предсказывает будущее значение разности на 1 день вперед. Для того чтобы модель выводила значение разности больше 1, используется линейная функция активации. Для контролирования переобучения и повышения точности предсказания модели используются слои Dropout.

Модель обучается на 300 эпохах и с оптимизатором АДАМ, в качестве функции ошибки используется среднеквадратичная логарифмическая ошибка, а в качестве метрики средняя абсолютная погрешность. В результате получили следующий результат, представленный на рисунке 14.



**Рис 14**. Предсказанные на 1000 дней вперед значения нейронной сетью

Видно, что предсказанные значения моделью, обозначенные красными линиями, а синими исходные данные, повторяют поведение ряда разности, и видно, что модель способна предвидеть резкие всплески, то есть непредвиденные расходы.

В результате исследования моделей при помощи алгоритмов машинного обучения и полученной модели нейронной сети. Можно прийти к выводу, что выбор нейронной сети будет лучшим вариантом для прогнозирования ряда с большими всплесками значений, поскольку она лучше определила закономерности поведения. Поэтому в качестве основы остается нейронная сеть.

2.4 Обертка модели

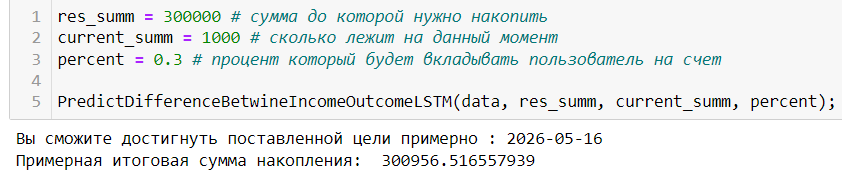
Модель предсказывает значение разности на 1 день вперед, поэтому добавляя предсказанное значение в исходные данные и выбирая новые 30 значений, можно производить прогнозы на несколько дней вперед. Однако модель предсказывает значения разности, а не дни, поэтому пока что она не решает поставленную задачу. Для того чтобы выполнить предсказание приблизительной даты обернем полученную модель в функцию.

На вход данной функции будут подаваться следующие значения:

1. data - значения доходов и расходов, объединенных в один датасет, в течение 30 дней, тто есть месяца,
2. res\_summ – итоговая сумма которую нужно накопить,
3. current\_summ – текущая сумма, которая уже накоплена,
4. percent – процент, который будет вкладывать пользователь на счет.

Внутри обертки происходит предсказание значения разности на 100 дней вперед, после же положительные предсказанные значения прибавляются к текущей накопленной сумме. Таким образом подсчитывается количество дней. И это продолжается до тех пор, пока текущая сумма не достигнет или не перепрыгнет итоговую сумму, или не пройдет лимит в 40 лет, то есть приблизительно 14 600 дней.

Результатом обертки является приблизительная дата, к которой наберется нужная сумма и итоговая полученная сумма накопления. Например, при текущей сумме в 1 000 руб. нужно накопить 300 000 руб., если пользователь от дохода будет откладывать 30 процентов на накопительный счет. Тогда модель выдает следующий результат, смотреть рисунок 15.



**Рис 15**. Результат работы обвертки

Получается, что с исходными данными, выйдет накопить 300 000 руб., через 3 года или к 16 мая 2026 года.

Заключение

В ходе выполнения курсовой работы получилось решить все поставленные задачи.

Была сформулирована развернутая задача, по поставленной цели работы. А именно предсказать приблизительную дату накопления средств, с учетом непредвиденных расходов пользователя, при условии, что он будет откладывать некий процент от дохода на накопительный счет.

Для решения данной задачи было рассмотрено множество инструментов. В качестве входных данных было принято решение использовать временной ряд. Поэтому для анализа и предсказания этого временного ряда были рассмотрены алгоритмы машинного обучения: LinearRegression, XGBRegressor, SARIMAX. А также рекуррентная нейронная сеть с долгой краткосрочной памятью LSTM.

В реализации данной задачи, были проанализированы выше описанные алгоритмы, и на основе показанных ими результатов, была выбрана модель нейронной сети LSTM.

Поскольку модель предсказывала только значение разности наперед в 1 день, то была реализована обертка данной модели, которая выдает требуемый результат задачи, а именно дату приблизительного накопления средств и итоговую сумму накопления.

Таким образом поставленная цель была достигнута. Однако точность модели не удалось приблизить к 85 процентам, поэтому в будущем хорошим решением будет улучшить или пересмотреть модель нейронной сети. Так же оптимизировать работу обертки, для того чтобы ускорить время ее работы.

Список источников

1. Анализ временных рядов // MaxRokatansky // Хабр [Электронный ресурс] URL: <https://habr.com/en/companies/otus/articles/732080/> (Дата обращения 2023-04-29)
2. Авторегрессия // Грицай Александр Александрович // fnow.ru [Электронный ресурс] URL: <https://fnow.ru/algorithm-comparison/avtoregressia> (Дата обращения 2023-04-29)
3. 9 ключевых алгоритмов машинного обучения простым языком Vszlo93 // Хабр [Электронный ресурс] URL: <https://habr.com/en/articles/509472/> (Дата обращения 2023-04-29)
4. Что такое линейная регрессия? // amazon // amazon [Электронный ресурс] URL: <https://aws.amazon.com/ru/what-is/linear-regression/> (Дата обращения 2023-04-29)
5. Решаем задачи машинного обучения с помощью алгоритма градиентного бустинга // Alex Maszański // proglib.io [Электронный ресурс] URL: <https://proglib.io/p/reshaem-zadachi-mashinnogo-obucheniya-s-pomoshchyu-algoritma-gradientnogo-bustinga-2021-11-25> (Дата обращения 2023-04-29)
6. LSTM – сети долгой краткосрочной памяти // Wunder Fund // Хабр [Электронный ресурс] URL: <https://habr.com/en/companies/wunderfund/articles/331310/> (Дата обращения 2023-04-29)
7. Keras API reference // keras // keras.io [Электронный ресурс] <https://keras.io/api/> (Дата обращения 2023-04-29)
8. API Reference // scikit-learn // scikit-learn [Электронный ресурс] <https://scikit-learn.org/stable/modules/classes.html> (Дата обращения 2023-04-29)