МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное
учреждение высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Институт естественных и точных наук
Кафедра математического и компьютерного моделирования

РАБОТА ПРОВЕРЕНА
Рецензент, доцент каф. СП,
канд. физ.-мат. наук
_____/ А.Т. Латипова
«____»_____2019 г.

ДОПУСТИТЬ К ЗАЩИТЕ
Заведующий кафедрой МиКМ,
д-р физ.-мат. наук, доцент
_____/ С.А. Загребина
«____»_____2019 г.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ НЕСКОЛЬКИХ ПОДХОДОВ К РЕШЕНИЮ ЗАДАЧИ ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ НА ПРИМЕРЕ АНАЛИЗА ИНТЕНСИВНОСТИ ИНТЕРНЕТ-ТРАФИКА

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
ЮУрГУ– 01.03.02.2019.009.ВКР

Руководитель работы,
старший преподаватель каф. МиКМ,
_____/ М.С. Фокина
«____»_____2019 г.

Автор работы,
студент группы ЕТ-416
_____/ А.А. Чайко
«____»_____2019 г.

Нормоконтролер,
доцент каф. МиКМ,
канд. физ.-мат. наук
_____/ Т.А. Макаровских
«____»_____2019 г.

Челябинск 2019

Выпускная квалификационная работа выполнена мной совершенно самостоятельно. Все использованные в работе материалы и концепции из опубликованной научной литературы и других источников имеют ссылки на них.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Институт естественных и точных наук
Кафедра математического и компьютерного моделирования
Направление подготовки 01.03.02 Прикладная математика и информатика

УТВЕРЖДАЮ
Заведующий кафедрой МиКМ,
д-р физ.-мат. наук, доцент
_____/ С.А. Загребина
«____»_____2019 г.

З А Д А Н И Е

на выпускную квалификационную работу студента

Чайко Александра Александровича

Группа ЕТ-416

1. **Тема работы** «Сравнительный анализ нескольких подходов к решению задачи прогнозирования временных рядов на примере анализа интенсивности интернет-трафика»
   утверждена приказом по университету от « 25 » апреля 2019 г. № 899.

2. **Срок сдачи студентом законченной работы** « 2 » июля 2019 г.

3. **Исходные данные к работе**
   Данные по использованию интернет-трафика (в битах) от частного интернет-провайдера с центрами в 11 европейских городах. Данные были собраны с 6:57 часов 7 июня до 11:17 часов 31 июля 2015 года.

4. **Перечень вопросов, подлежащих разработке**
   4.1. Обзор задач по прогнозированию интенсивности интернет-трафика и применяемых методов для их решения
   4.2. Обзор современных моделей, применяемых для прогнозирования интенсивности интернет трафика.

4.3. Описание моделей прогнозирования временных рядов, обоснование использования рассмотренных моделей.

4.4. Проверка адекватности выбранных моделей прогнозирования временных рядов.

4.5. Анализ эффективности и оценка качества прогнозирования используемых моделей.

5. Графические материалы

5.1. Титульный лист (1 сл.)

5.2. Цели, задачи, актуальность и предмет исследования (3 сл.)

5.3. Характеристика исходных данных (2 сл.)

5.4. Описание и результаты применения методов (9 сл.)

5.5. Проверка адекватности построенных моделей (2 сл.)

5.6. Заключение (1 сл.)

6. Календарный план

Наименование этапов выпускной квалификационной работы	Срок выполнения этапов работы	Отметка о выполнении руководителя
1. Изучение литературы по тематике выпускной квалификационной работы	30.01.2019-23.02.2019	
2. Анализ применяющихся методов решения для анализа интернет трафика	18.03.2019-09.04.2019	
5. Реализация моделей авторегрессии и интегрированного скользящего среднего (ARIMA)	19.04.2019-29.04.2019	
6. Реализация многофакторной регрессионной модели, деревьев решений	30.04.2019-14.05.2019	
7. Проверка адекватности созданных моделей	15.05.2019-20.05.2019	
8. Оценка качества созданных моделей	21.05.2019-22.06.2019	
9. Нормоконтроль	25.06.2019	
10. Защита выпускной квалификационной работы	02.07.2019	

7. Дата выдачи задания « 1 » декабря 2018 г.

Руководитель работы _____/М.С. Фокина
(подпись)

Студент _____/А.А. Чайко
(подпись)

АННОТАЦИЯ

Чайко, А.А. Сравнительный анализ нескольких подходов к решению задачи прогнозирования временных рядов на примере анализа интенсивности интернет-трафика. / А.А. Чайко. – Челябинск: ЮУрГУ, ЕТ-416, 68 с., 23 ил., 6 табл., библиогр. список – 27 наим., 3 прил.

Выпускная квалификационная работа выполнена с целью проведения сравнительного анализа эффективности применения различных методов прогнозирования к временным рядам интенсивности сетевого трафика телекоммуникационных систем.

Проанализирована возможность использования деревьев решения и многофакторных регрессионных моделей наравне с интегрированной моделью авторегрессии – скользящего среднего (ARIMA).

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Системы мониторинга трафика играют ключевую роль в эффективном управлении сетью. Они являются источником информации о работе корпоративных систем, такая информация учитывается при распределении средств, основываясь на ней производится планирование вычислительных мощностей, определение и локализации отказов, решение различных вопросов безопасности.

В настоящее время, для прогнозирования всевозможных параметров телекоммуникационных систем используется большое количество различных методов, в том числе, широкое распространение получили методы, основанные на анализе временных рядов. Существуют работы по данной тематике, основанные на авторегрессионных моделях [1]. В настоящее время появились работы, которые описывают возможность применения моделей Хольта-Винтерса и нейронных сетей [2]. В описанных работах используются временные ряды с интервалами наблюдения от 5 и более минут.

В данной выпускной квалификационной работе будет рассмотрено несколько методов и подходов к анализу и прогнозированию интенсивности сетевого трафика на примере данных по интернет-трафику (в битах) от частного интернет-провайдера с центрами в 11 европейских городах. Данные были собраны с 6:57 часов 7 июня до 11:17 часов 31 июля 2015 года. Данные собирались с пятиминутными интервалами.

Актуальность решения такой задачи обусловлена, так как для поддержания корректной работы сети необходимо как можно быстрее детектировать аномальную активность и предпринимать меры по устранению проблем. В выпускной квалификационной с помощью методов анализа временных рядов проводится исследование, целью которого является проверка эффективности рассматриваемых методов анализа временных рядов для прогнозирования с некоторой допустимой точностью объема использованного трафика в определенный момент времени.

Используя полученные в итоге модели для прогнозирования интенсивности трафика в сети, можно построить прогноз на некоторый период времени в будущем. Если принять прогнозные значения как ожидаемый нормальный уровень интенсивности трафика, то в случае значительного уменьшения объема интенсивности сетевого трафика (когда значение выходит за пределы доверительного интервала), с большой долей вероятности можно говорить о возникновении неполадок в работе сети.

Целью исследования выпускной квалификационной работы является проверка эффективности рассматриваемых методов анализа временных рядов для прогнозирования с некоторой допустимой точностью объема использованного трафика в определенный момент времени.

Достижение поставленной цели потребовало решения следующих **задач.**

1. Создание и обоснование применения наиболее подходящей модели прогнозирования для данного конкретного случая.

2. Сравнение нескольких моделей прогнозирования между собой, описание их преимуществ и недостатков, применимо к анализу временного ряда интенсивности сетевого трафика.

3. Тестирование наиболее эффективной модели прогнозирования для моделирования изменения сетевого трафика, анализ и структуризация полученных результатов.

Объект исследования – любое предприятие, для работы которого так или иначе требуется взаимодействие с передачей данных, в результате чего появляется возможность отслеживать и анализировать интенсивность входящего сетевого трафика.

Предметом исследования являются подходы к решению задачи прогнозирования временных рядов, на примере проблемы анализа интенсивности сетевого трафика, возникающей у какого-либо предприятия в процессе работы с потоками данных.

Информационная база включает в себя учебники по статистическим методом прогнозирования и анализу временных рядов за авторством (Дубровой Т.А., Афанасьева В.Н., Олифера В.Г., Бокса Д., Дженкинса Г.М. и др.), фундаментальные научные статьи по теме за авторством (Покровской М.А. и Лысяка А.С.).

Итоговые научные результаты исследования получены с использованием методов математического моделирования, сравнительного и функционального анализов, методов обработки и анализа данных. За основу для проведения математического моделирования и анализа данных был взят функционал языка программирования Python, в частности Python-библиотек, ориентированных на выполнение представленных задач (NumPy, Pandas, Statsmodels, SciPy, Sklearn), для визуализации результатов работы использовалась библиотека Matplotlib и Plotly. Дополнительно для проведения анализа данных использовался функционал программы MS Excel.

Работа состоит из четырех глав, введения, заключения, библиографического списка (27 наименований), трёх приложений.

Во введении обосновывается актуальность работы, приводятся цели и задачи работы, указываются объект и предмет исследований, обосновывается информационная база исследований и приводится краткое содержание работы.

В главе 1 описываются особенности прогнозирования временных рядов интенсивности сетевого трафика, проводится обзор существующих методов прогнозирования трафика и приводятся выводы по поводу частоты применимости тех или иных моделей для создания краткосрочных и долгосрочных прогнозов сетевого трафика.

В главе 2 представлены модели, используемые в работе для анализа временного ряда, содержащего данные по интенсивности сетевого трафика. В данной главе описаны особенности каждой модели и представлена необходимая математическая база.

В главе 3 описывается предварительный анализ данных, показано наличие дневной и недельной сезонности в исходном временном ряду, с помощью критерия Дики-Фуллера проведена проверка исходного временного ряда на стационарность, в ходе которой гипотеза о том, что исходный временной ряд стационарен подтвердилась.

В главе 4 проводится создание, описание и сравнение рассматриваемых в работе моделей, делается вывод о точности предсказания для различных прогнозных периодов, после чего, проводится анализ адекватности построенных моделей, основанный на проверке ряда остатков.

В заключении приводятся основные результаты работы, формулируются выводы и рекомендации, описываются направления дальнейших исследований.

1 ИСПОЛЬЗОВАНИЕ МЕТОДОВ АНАЛИЗА ВРЕМЕННЫХ РЯДОВ ДЛЯ ПРОГНОЗИРОВАНИЯ ИНТЕНСИВНОСТИ СЕТЕВОГО ТРАФИКА

1.1 Особенности прогнозирования интенсивности сетевого трафика

Прогнозирование характеристик трафика является одной из важных задач при построении автоматизированных систем управления телекоммуникационными системами (ТКС). Различные виды коммуникационных услуг и различные конфигурации сетей порождают существенно различающиеся виды трафика [3]. Анализируя временные ряды параметров сетевого трафика, следует учитывать, что, как правило, данные временные ряды нестационарны. Это позволяет говорить о высокой сложности моделирования трафика в телекоммуникационных сетях.

Предполагая нестационарность в процессе построения моделей, необходимо проводить дополнительные исследования и обработку как исходных данных, так и данных, полученных как результат прогноза построенных моделей.

Особую сложность для обеспечения качества обслуживания вызывают приложения, в основе которых лежит самоподобный процесс. На качественном уровне самоподобие проявляется в появлении медленно-убывающей зависимости между величинами трафика в различные периоды времени, и в том, что трафик имеет выбросы, которые выглядят статистически подобными при различных масштабах времени. Использование классических пуассоновских моделей в этом случае приводит к недооценке реальной нагрузки. Самоподобные процессы характеризуются свойством долговременной памяти и, как следствие, наличием зависимости интенсивностей трафика даже между далеко стоящими друг от друга значениями. Это свойство позволяет на основе накопленной статистики трафика прогнозировать его будущее состояние, так как длительная память предполагает

наличие значительной автокорреляции между сравнительно удаленными друг от друга значениями трафика [4].

Принимая во внимание то, что существующие технологии прогнозирования временных рядов, как правило, подразумевают под собой создание модели данных рядов, после которого идет расчет необходимого значения с помощью построенной модели, как следствие, выбор модели является одним из центральных вопросов при построении прогноза.

В связи с тем, что в случае прогнозирования нестационарных временных рядов нельзя говорить о преимуществе какой-то из моделей без экспериментального подтверждения, следует постоянно обновлять результаты исследований [3].

1.2 Описание общих тенденций сетевого трафика в области телекоммуникационных сетей

В начале XXI века быстрыми темпами начала развиваться область инфокоммуникаций. На это повлияли многие факторы, в числе которых, рост производительности, резкое уменьшение стоимости и повсеместное использование микропроцессоров, а также технологические успехи, стимулирующие развитие систем передачи данных.

С каждым годом неотвратимо идет тенденция к увеличению объемов передаваемого и принимаемого трафика. Быстрый рост инфраструктуры уже существующих сетей и потребность в унификации услуг, все это привело к необходимости создания новых концепций, направленных на универсализацию, с помощью которых будет происходить развитие в будущем.

В настоящее время изменение подхода к построению сетей лежит в основе новой глобальной сети базирующейся на принципах самоорганизации. Применимо к данному случаю, самоорганизация является необходимостью в условиях гетерогенной среды, динамично меняющей свое поведение. Скопления мобильных устройств, сенсоров, разнообразных датчиков генерируют плохо предсказуемый

трафик, делая нецелесообразным, а зачастую и невозможным, планирование и централизованное управление такими сетями [5].

Развитие сетей связи в рамках обозначенных концепций неизбежно отражается на объёмах и структуре трафика. Согласно опубликованным результатам исследований суммарный трафик глобальной сети показывает устойчивый экспоненциальный рост со второй половины 90-х годов XX века [5]. В настоящее время эта тенденция сохраняется (рисунок 1).

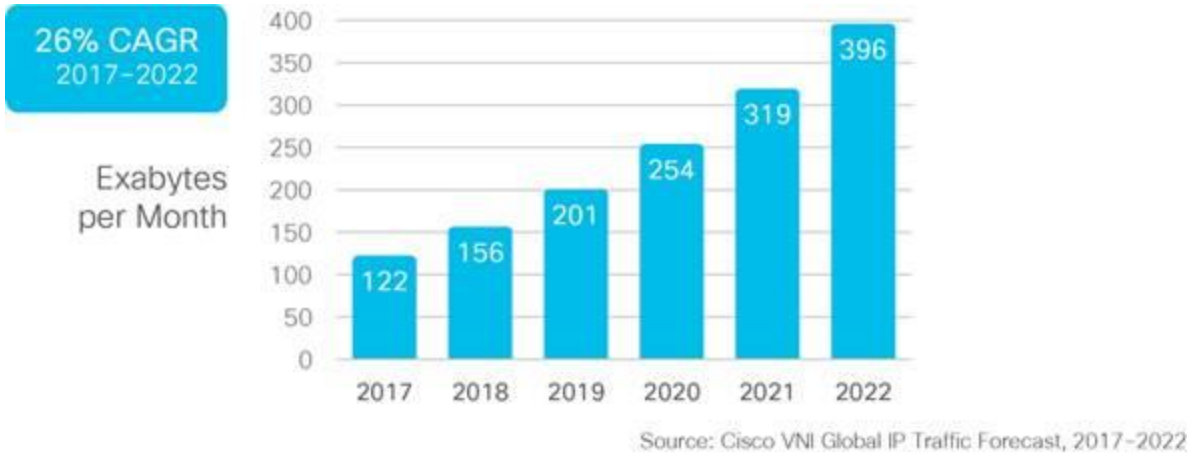

Рисунок 1 – Прогноз совокупного IP-трафика с 2017 по 2022 год

Во всем мире количество устройств способных генерировать IP трафик растет быстрее (10% CAGR – Compound Annual Growth Rate (Совокупный среднегодовой темп роста)), чем население (1,0% CAGR) и интернет-пользователи (7% CAGR). Эта тенденция отражает убыстряющийся рост среднего числа устройств и интернет подключений на домохозяйство и на душу населения [6] и, как следствие, в долгосрочной перспективе необходимо учитывать восходящий тренд в сферах, взаимодействующих с трафиком сети.

Следует также учитывать, что каждый год на рынке появляются и внедряются новые устройства в различных форм-факторах с расширенными возможностями и функциями. Растущее число приложений M2M – Machine-to-Machine (машинное взаимодействие), таких как интеллектуальные счетчики, Wi-Fi камеры видеонаблюдения и др. вносят основной вклад в рост количества устройств и ин-

тернет подключений. По прогнозам к 2022 году соединения M2M будут составлять 51 процент от общего количества устройств и соединений (рисунок 2) [6].

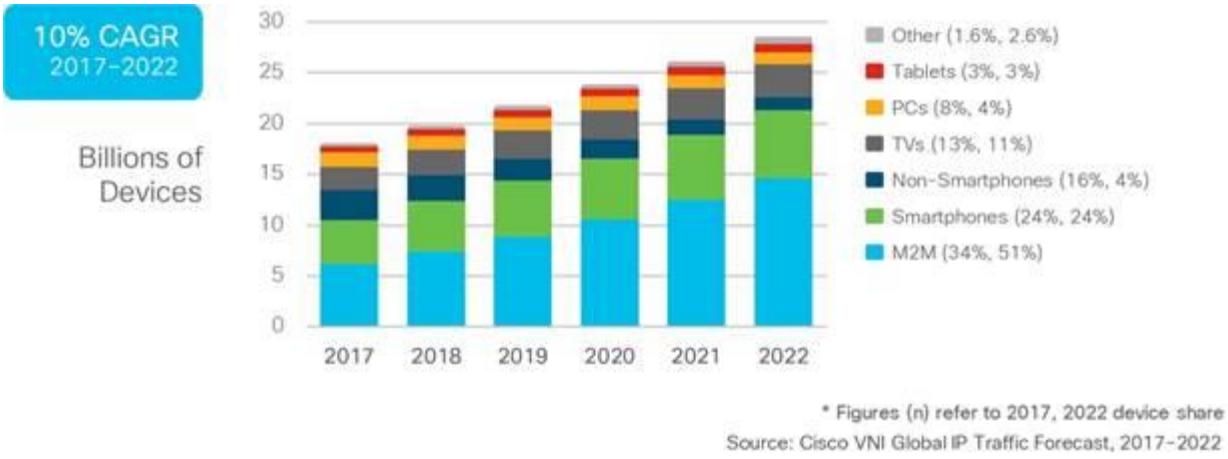

Рисунок 2 – Совокупный рост устройств и интернет-подключений по типу устройств

Вместе с рассмотренными выше тенденциями, все существенней становится нестационарность передаваемой нагрузки. Одним из примеров такой нестационарности может служить известный эффект flash-crowd, возникающий при вирусном распространении информации о некотором популярном ресурсе/контенте. Результатом этого эффекта является временный и практически непредсказуемый всплеск интенсивности передаваемого сетевого трафика [5]. Подобная координация вызывает резкий рост интенсивности передаваемого трафика.

Логично предположить, что спонтанные всплески интенсивности трафика практически невозможно предсказать. Однако можно создать систему детектирования таких аномальных значений интенсивности трафика. После чего, при возникновении неожиданного всплеска трафика возможно автоматическое выполнение заранее созданных сценариев, которые в свою очередь призваны снизить последствия всплеска трафика.

1.3 Методы анализа временных рядов для прогнозирования интенсивности сетевого трафика

В результате исследований в области анализа временных рядов телекоммуникационных сетей были выявлены такие свойства трафика в ТКС как группировка во времени значительных изменений значений («скачков»), фрактальность, хаотичность, нелинейность, циклические колебания, присущие в разной мере рядам его характеристик в зависимости от контекста [7, 8]. Отмечено, что наибольшее количество публикаций в этой области пришлось на период 1990-х – начала 2000-х годов [2].

Примерами ранних методов прогнозирования интенсивности сетевого трафика являются модели FARIMA (авторегрессионная дробно интегрированная скользящая средняя) для долгосрочного прогнозирования интенсивности сетевого трафика [9]. Показано, что правильно построенная FARIMA-модель позволяет получить довольно высокую точность прогнозирования на длительном периоде упреждения. Однако наибольшей популярностью для прогнозирования в данной области пользуется модель авторегрессии-проинтегрированного скользящего среднего (ARIMA) [10].

Существуют также работы, где оптимальным методом автономного краткосрочного сетевого прогнозирования является экспоненциальное сглаживание, оно сочетает в себе универсальность и в то же время надежность и точность среди всех рассмотренных методов, рассмотренных в работе (полиномиальная аппроксимация, линейное предсказание) [11].

В общем случае, если в работах по прогнозированию трафика в телекоммуникационных сетях рассматриваются методы прогнозирования вкупе с различными горизонтами прогноза, то для краткосрочного прогнозирования в работах отдается предпочтение более простым, быстрым и не требующим много ресурсов моделям, таким как стандартные линейные модели и модели экспоненциального сглаживания [7, 9, 11].

В случае долгосрочного прогноза в данных работах отдают предпочтение моделям авторегрессии и проинтегрированного скользящего среднего (ARIMA).

К достоинствам упомянутых методов можно отнести весьма высокую точность, а также дальний горизонт прогнозирования, т.е. максимальную длительность периода упреждения, позволяющую получить долгосрочный прогноз с хорошей степенью точности. В то же время такие решения обладают рядом недостатков с точки зрения автоматизированного применения. Так модели ARIMA/ SARIMA/FARIMA требуют предварительного анализа воспроизводимого временного ряда экспертом с целью выбора адекватных параметров модели. Кроме того, модели типа ARIMA довольно требовательным к вычислительным ресурсам, что также может послужить ограничением при их использовании [5].

Выводы по главе один

1. Особенностью прогнозирования временных рядов сетевого трафика является то, что чаще всего такие временные ряды являются нестационарными вследствие зашумленности данных, вызванных самой структурой работы сети. Как итог, временные ряды в исследуемой области являются достаточно сложными для прогнозирования.

2. Актуальность исследования выпускной квалификационной работы подтверждается при обзоре научных работ по данной теме. Несмотря на то, что большинство публикаций в данной области сделано более 20 лет назад, в настоящее время все еще существует потребность в прогнозировании сетевого трафика. Следует также учитывать, что структура трафика, как и его интенсивность за прошедшее время существенно изменилась, что необходимо учитывать при построении моделей прогнозирования.

2 ОПИСАНИЕ КЛЮЧЕВЫХ ОСОБЕННОСТЕЙ МЕТОДОВ ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ, ИСПОЛЬЗУЕМЫХ ДЛЯ АНАЛИЗА СЕТЕВОГО ТРАФИКА

Статистическая информация интенсивности сетевого трафика, находящаяся в свободном доступе, представлена в виде временных рядов, которые являются наиболее удобной формой представления данный для прогнозирования.

Временным рядом (динамическим рядом, английский термин «Time series») называется ряд расположенных в хронологической последовательности значений статистического показателя, характеризующего изменение некоторого явления во времени. В нем процесс развития изображается в виде совокупности прерывных случаях непрерывного, позволяющих детально проанализировать особенности развития при помощи характеристик, отображающих изменение параметров системы во времени. Фактор времени здесь приобретает решающее значение [12].

Перед описанием конкретных методов стоит разграничить понятие прогнозирования и предсказания. Прогнозирование - это процесс оценивания будущего события (событий), который в той или иной мере использует накопленные предыдущие данные и объединяет их предопределенным путем, чтобы получить необходимую оценку. Предсказание, помимо этого, оперирует субъективными соображениями.

В общем случае определение прогноза формулируется так: **прогноз** – это количественное, вероятностное утверждение в будущем о состоянии объекта или явления с относительно высокой степенью достоверности, на основе анализа тенденций и закономерностей прошлого и настоящего, **предсказание** – это предвидение таких событий, количественная характеристика которых невозможна или затруднена [12].

Для осуществления прогноза, то есть определения понятий, как будут осуществляться и развиваться прогнозируемые явления в будущем, необходимо знать тенденции и закономерности прошлого и настоящего. При этом, следует

помнить, что будущее зависит от многих случайных факторов, сложное переплетение и сочетание которых учесть практически невозможно. Следовательно, все прогнозы носят вероятностный характер.

Период упреждения прогноза – это отрезок времени от момента, для которого имеются последние фактические данные об изучаемом объекте, до момента, к которому относится прогноз. Период упреждения прогноза зависит от специфики и особенностей изучаемого объекта исследования, от интенсивности изменения показателей, от продолжительности действия выявленных тенденций и закономерностей, от длины временного ряда и от многих других факторов.

Для случая прогнозирования интенсивности сетевого трафика принимаются за основу следующие значения периодов упреждения прогноза:

- текущий прогноз – следующие 5 мин;
- краткосрочный прогноз – от пяти минут до часа;
- долгосрочный прогноз – от часа до одного дня.

В практике прогнозирования принято считать, что значения уровней временных рядов экономических показателей состоят из следующих компонент: тренда, сезонной, циклической и случайной составляющих [13].

Под трендом понимают изменение, определяющее общее направление развития, основную тенденцию временного ряда. Это систематическая составляющая долговременного действия.

Наряду с долговременными тенденциями во временных рядах экономических процессов часто имеют место более или менее регулярные колебания - периодические составляющие рядов динамики. Если период колебаний не превышает одного года, то их называют сезонными [13].

Если из временного ряда удалить тренд и периодические составляющие, то останется нерегулярная компонента.

Факторы, под действием которых формируется нерегулярная компонента, разделяют на два вида:

– факторы резкого, внезапного действия;

– текущие факторы.

В итоге аддитивная модель временного ряда имеет вид:

$$Y_t = u_t + s_t + v_t + e_t ,$$

где Y_t – уровни временного ряда, u_t – трендовая компонента, s_t – сезонная компонента, v_t – циклическая компонента, e_t – случайная компонента.

Мультипликативная модель имеет вид:

$$Y_t = u_t \cdot s_t \cdot v_t \cdot e_t .$$

Также существуют смешанные модели, включающие как аддитивную, так и мультипликативную составляющую. В случае нормальности распределения абсолютных отклонений связь является аддитивной, а относительных – мультипликативной [12, 14].

В случае, когда временные ряды, в которых можно установить тенденцию и случайный компонент, особенно при использовании годовых данных, где влияние сезонности не отражается. Аналитически данное положение можно выразить уравнением вида:

$$\widehat{Y_t} = f(t) + \varepsilon(t),$$

где $f(t)$ – некоторая функция времени, описывающая тенденцию исходного временного ряда, называемая трендом, $\varepsilon(t)$ – случайная величина (случайный компонент).

Функция $f(t)$ определяет общую тенденцию развития изучаемого явления. Поэтому прежде чем приступать к моделированию и прогнозированию необходимо проверить гипотезу о наличии тенденции в исходном временном ряду.

После этого, в зависимости от того, какие принципы и исходные данные положены в основу прогноза возможно применение различных методов прогнозирования с определенными преимуществами в конкретных случаях. Далее будут рассмотрены методы прогнозирования, применяющиеся в работе.

2.1 Авторегрессионные модели

Рассмотрим представление общего линейного процесса, на вход которого поступает белый шум a_t, т.е.

$$\tilde{z}_t = a_t + \Psi_1 a_{t-1} + \Psi_2 a_{t-2} + \ ... = \ a_t + \sum_{l=1}^{\infty} \Psi_l a_{t-l}, \qquad (1)$$

где $\tilde{z}_t$ – взвешенная сумма настоящего и прошлых значений белого шума a_t. Белый шум можно рассматривать как последовательность импульсов, приводящих систему в движение, Ψ_l – некоторый вес при коэффициенте a_{t-l} [15].

Белый шум a_t состоит из последовательности некоррелированных случайных переменных с нулевым средним значением

$$E[a_t] = 0,$$

и постоянной дисперсией

$$var[a_t] = \sigma_a^2.$$

Процесс (1) в называется линейным, если a_t независимы, одинаково распределены, имеют конечную дисперсию и удовлетворяют следующему условию:

$$\sum_{l=1}^{\infty} \Psi_l^2 < \infty.$$

Рассмотрим частный случай модели (1), когда только первые p значений весов ненулевые. В итоге можно записать такую модель:

$$\tilde{z}_t = \phi_t \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + \ ... + \ \phi_p \tilde{z}_{t-p} + a_t \ , \qquad (2)$$

где ϕ_1, ϕ_2 ... ϕ_p – конечный набор весовых параметров.

Процесс (2) называют процессом авторегрессии p-го порядка, или сокращенно процессом $AR(p)$. В частности большое практическое значение имеют процессы авторегрессии первого $(p = 1)$ и второго $(p = 2)$ порядка [15].

Уравнения данных процессов выглядят так:

$$\tilde{z}_t = \phi_t \tilde{z}_{t-1} + a_t \ ,$$

$$\tilde{z}_t = \phi_t \tilde{z}_{t-1} + \phi_t \tilde{z}_{t-2} + a_t \ .$$

Обычно итоговую формулу приводят в виде:

$$\tilde{z}_t = \sum_{i=1}^{p} \phi_t \tilde{z}_{t-i} + \varepsilon_i + C \ ,$$

где ε_i – белый шум, C – некоторая константа (часто принимается $C = 0$), а вместо $\tilde{z}_t$ обычно используют сокращение $AR(p)$, что по своей сути – авторегрессионная модель p-го порядка.

В общем случае **авторегрессионная модель** – это модель стационарного процесса, выражающего значение показателя в виде линейной комбинации конечного числа предшествующих значений этого показателя и аддитивной случайной составляющей.

Авторегрессионные модели по своей сути не предназначены для описания процессов с тенденцией, однако они хорошо описывают колебания, что весьма важно для отображения развития неустойчивых показателей.

2.2 Метод скользящего среднего

Распространенным приемом при выявлении тенденции развития является сглаживание временного ряда. Суть различных приемов сглаживания сводится к замене фактических уровней временного ряда расчетными уровнями, которые подвержены колебаниям в меньшей степени. Это способствует более четкому проявлению тенденции развития. Иногда сглаживание применяют как предварительный этап перед использованием других методов выделения тенденции.

Алгоритм сглаживания по простой скользящей средней может быть представлен в виде следующей последовательности шагов:

1. Определяют длину интервала сглаживания g, включающего в себя g последовательных уровней ряда ($g < n$). При этом надо иметь в виду, что чем шире интервал сглаживания, тем в большей степени взаимопогашаются колебания, и тенденция развития носит более плавный, сглаженный характер. Чем сильнее колебания, тем шире должен быть интервал сглаживания.

2. Разбивают весь период наблюдений на участки, при этом интервал сглаживания как бы скользит по ряду с шагом, равным 1.

3. Рассчитывают арифметические средние из уровней ряда, образующих каждый участок.

4. Заменяют фактические значения ряда, стоящие в центре каждого участка, на соответствующие средние значения [13].

Простое скользящее среднее вычисляется по формуле:

$$p_{t+1} = \frac{1}{n}\sum_{i=0}^{n-1} p_{t-i} = \frac{p_t + p_{t-1} + \ \ldots + p_{t-i} + \ \ldots + p_{t-n+2} + p_{t-n+1}}{n},$$

где n – количество значений временного ряда для расчета скользящего среднего, p_t – значение временного ряда в точке t.

Процедура сглаживания приводит к полному устранению периодических колебаний во временном ряду, если длина интервала сглаживания берется равной или кратной циклу, периоду колебаний.

Метод простой скользящей средней применим, если графическое изображение динамического ряда напоминает прямую. Когда тренд выравниваемого ряда имеет изгибы, и для исследователя желательно сохранить мелкие волны, применение простой скользящей средней нецелесообразно.

Если для процесса характерно нелинейное развитие, то простая скользящая средняя может привести к существенным искажениям. В этих случаях более надежным является использование взвешенной скользящей средней [13].

При сглаживании по взвешенной скользящей средней на каждом участке выравнивание осуществляется по полиномам невысоких порядков. Чаще всего используются полиномы 2-го и 3-го порядка. Так как при простой скользящей средней выравнивание на каждом активном участке производится по прямой (полиному первого порядка), то метод простой скользящей средней может рассматриваться как частный случай метода взвешенной скользящей средней.

Простая скользящая средняя учитывает все уровни ряда, входящие в активный участок сглаживания, с равными весами, а взвешенная средняя приписывает каждому уровню вес, зависящий от удаления данного уровня до уровня, стоящего в середине активного участка.

Выравнивание с помощью взвешенной скользящей средней осуществляется следующим образом. Для каждого активного участка подбирается полином вида

$$\widehat{y_t} = a_0 + a_1 t + a_2 t^2 + \ \dots\ ,$$

где t – порядковый номер уровня в интервале сглаживания. Полином первого порядка – есть уравнение прямой, следовательно, метод простой скользящей средней является частным случаем метода взвешенной скользящей средней. Коэффициенты находятся методом наименьших квадратов.

Метод скользящих средних используется в том случае, когда необходимо представить общую картину развития, основанную на механическом повторении одних и тех же действий по увеличению интервала времени. Либо же как еще один метод для сглаживания сезонных колебаний.

2.3 Модель авторегрессии – проинтегрированного скользящего среднего (ARIMA)

Для достижения большей гибкости в подгонке моделей к наблюдаемым временным рядам иногда целесообразно объединить в одной модели и авторегрессию, и скользящее среднее [15]. Это приводит к комбинированной модели авторегрессии-скользящего среднего – $ARMA(p, q)$ при $d = 0$:

$$\Delta^d X_t = \sum_{i=1}^{p} \phi_i \left(\Delta^d X_{t-1}\right) + \varepsilon_i + \sum_{j=1}^{q} \theta_j \left(\Delta^d \varepsilon_{t-j}\right),$$

$$\varepsilon_i \sim N(0, \sigma_t^2),$$

где ε_i – стационарный временной ряд, ϕ_i и θ_j – полиномы степени p и q, Δ^d - оператор разности временного ряда порядка d (последовательное взятие d раз разностей d-го порядка), например, разность первого порядка – $\Delta^1 X_t = X_{t-1} - X_t$.

Нестационарные временные ряды, которые повсеместно встречаются во многих сферах деятельности человека (например, в промышленности или торговле) и, в частности не имеют свойства колебания относительно фиксированного среднего, также могут быть в некотором смысле однородными. В таком случае, исследуемый однородный стационарный процесс может быть описан моделью, которая требует, чтобы d-я разность процесса была стационарной. На практике d обычно равно 0, 1 или максимум 2.

Более общий случай процесса ARMA(p, q), работающий как с стационарными, так и нестационарными временными рядами называется процессом авторегрессии – проинтегрированного скользящего среднего – ARIMA(p, d, q). В моделях типа ARIMA(p, d, q) параметр d – вариативен.

Впервые систематический подход к построению модели ARIMA был изложен Боксом и Дженкинсом в 1976 году. Методология построения ARIMA-модели (рисунок 3) для исследуемого временного ряда включает следующие основные этапы [13, 22]:

– идентификацию пробной модели;

– оценивание параметров модели и диагностическую проверку адекватности модели;

– использование модели для прогнозирования.

Рассмотрим поподробнее этапы построения модели ARIMA для области анализа временных рядов трафика (рисунок 3).

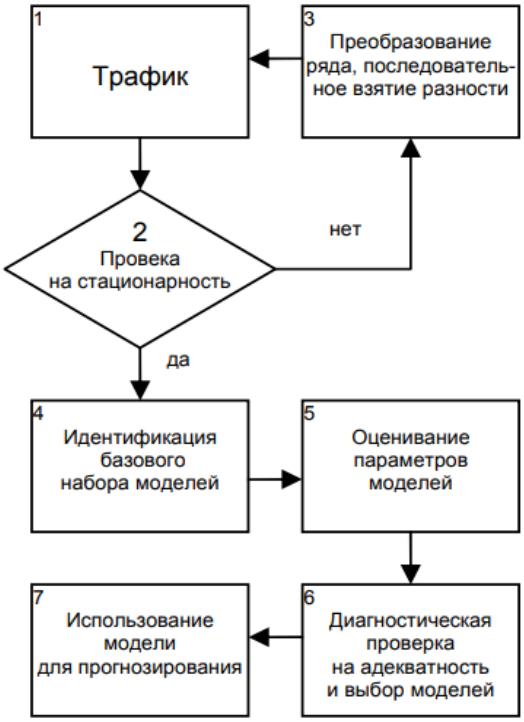

Рисунок 3 – Укрупненная структурная схема подбора модели ARIMA

Таким образом, сначала (в блоке 1 – 3) необходимо получить стационарный ряд. На этом этапе рекомендуется проводить анализ автокорреляционной функции (АКФ) и частной автокорреляционной функции (ЧАКФ). Быстрое затухание значений АКФ – простой тест на стационарность. На этом этапе используются также статистические тесты на наличие единичного корня (расширенный тест Дики-Фуллера или ADF-тест) [13, 22].

Если в соответствии со статистикой Дики-Фуллера или оценок АКФ ряд является нестационарным, то для перехода к стационарному ряду традиционно применяют оператор взятия последовательных разностей, тем самым определяется значение параметра d (порядка разности). Таким образом, значение одного параметра в модели ARIMA(p, d, q) уже известно.

В блоке 4 после получения стационарного ряда исследуется характер поведения выборочных АКФ и ЧАКФ и выдвигаются гипотезы о значениях параметров p (порядок авторегрессии) и q (порядок скользящего среднего). На входе блока 4

может формироваться базовый набор, включающий одну, две или даже большее число моделей, другими словами, портфель моделей.

В блоке 5 после осуществления идентификации модели необходимо оценить их параметры. Для этих целей используется метод максимального правдоподобия (ММП).

В блоке 6 для проверки каждой пробной модели на адекватность анализируется ее ряд остатков. У адекватной модели ряд остатков должен быть похож на белый шум, т.е. их выборочные АКФ не должны отличаться от нуля. Для проверки гипотезы о том, что наблюдаемые данные являются реализацией «белого шума», используется также Q-статистика. Q-статистика Льюинга-Бокса определяется как:

$$Q^* = n(n+2)\sum_{k=1}^{m}\frac{r_k^2}{n-k},$$

где n – объем выборки, m – максимальное количество лагов, r_k – коэффициенты автокорреляционной функции.

Если в результате проверки несколько моделей оказываются адекватны исходным данным, то при окончательном выборе следует учесть два фактора:

- повышение точности (качество подгонки модели);
- уменьшение числа параметров модели.

Воедино эти требования сведены в информационные критерии Акайке и Шварца. В данной статье выбран информационный критерий Акаике (AIC):

$$AIC = 2\frac{p+q}{n} + \ln\left(\frac{1}{n}\sum_{t=1}^{n} e_t^2\right),$$

где p – параметр порядка авторегрессионной модели, q – параметр окна модели скользящей средней, n – объем выборки по которой строилась модель.

Для оценки точности прогноза используется ряд стандартных показателей. Средняя абсолютная процентная ошибка (MAPE):

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{X_t - \widehat{X}_t}{X_t}\right|,$$

где X_t – реальное значение, $\hat{X}_t$ – прогнозное значение, n – интервал прогноза. Если $MAPE < 10\%$, то прогноз имеет высокую точность, $20\% < MAPE < 50\%$ – прогноз удовлетворительный, $MAPE > 50\%$ – прогноз имеет низкую точность [22].

2.4 Многофакторные регрессионные модели

Предположим, что зависимость результативного признака некоторого явления от ряда факторных признаков может быть записана уравнением:

$$\hat{y}_t = a_0 + a_1 x_{1t} + a_2 x_{2t} + \ ... + a_k x_k,$$

$(t = 1, 2, \ldots, k)$ и коэффициенты регрессии изменяются во времени по линейной функции так, что их можно записать уравнениями:

$$a_0 = b_{00} + b_{01} t,$$

$$a_1 = b_{10} + b_{11} t,$$

...

$$a_k = b_k + b_k t.$$

В этом случае уравнение регрессии имеет другой вид:

$$\widehat{y_t} = (b_{00} + b_{01} t) + (b_{10} + b_{11} t) x_{11} + (b_{20} + b_{21} t) x_{21} + \ ... + (b_k + b_k t) x_{k1}.$$

Параметры этого уравнения находятся по способу наименьших квадратов и показывают, как меняется во времени действие отдельных факторов на результативный признак рассматриваемого явления.

Метод наименьших квадратов, используемый в регрессионном анализе для определения коэффициентов регрессии, основывается на предпосылке независимости друг от друга отдельных наблюдений одной и той же переменной. В динамических рядах существует явление автокорреляции. Поэтому величина коэффи-

циентов регрессии, полученных по способу наименьших квадратов, не имеет нужных статистических свойств.

Наличие автокорреляции приводит к искажению средних квадратических ошибок коэффициентов регрессии, что в свою очередь затрудняет построение доверительных интервалов по ним и проверку их значимости по соответствующим критериям. Автокорреляция также может привести к сокращению числа наблюдений ввиду невозможности потерять показатели одного и того же объекта за ряд лет, поскольку наблюдение одного объекта за десять лет качественно отличается от наблюдений десяти объектов за одно и то же время. Возникает автокорреляция и в отклонениях от трендов, а также в случайных остатках уравнений регрессии, построенных по многомерным рядам динамики [21].

Автокорреляция — это наличие сильной корреляционной зависимости между последовательными уровнями временного ряда.

В настоящее время разработано несколько способов исключения автокорреляции, одним из них является метод, так называемых, последовательных или конечных разностей. Модель данным методом имеет вид:

$$\Delta y_{t+1} = a_0 + a_1\Delta x_1 + \ a_2\Delta x_2 + \cdots a_k\Delta x_k.$$

Сущность метода заключается в последовательном исключении величины предшествующих уровней из последующих:

$\Delta y_1 = y_t - y_{t-1}$, $\Delta x_1 = x_t - x_{t-1}$,

$\Delta y_2 = y_{t-1} - y_{t-2}$, $\Delta x_2 = x_{t-1} - x_{t-2}$,

…, …,

$\Delta y_k = y_{t-k+1} - y_{t-k}$. $\Delta x_k = x_{t-k+1} - x_{t-k}$.

При коррелировании разностей измеряется теснота связи между разностями последовательных величин уровней в каждом динамическом ряду. Показателем тесноты связей между изучаемыми рядами является **коэффициент корреляции разностей $r_{\Delta x\Delta y}$:**

$$r_{\Delta x \Delta y} = \frac{\sum \Delta x \Delta y}{\sqrt{\Delta x^2 \cdot \Delta y^2}}.$$

После получения уровней ряда динамики без автокорреляции остается еще сбалансировать эти уровни по времени. Для этого необходимо рассмотреть вопрос о временном лаге.

Временной лаг – запаздывание (или опережение) процесса развития, представленного одним временным рядом, по сравнению с развитием, предоставленным другим рядом. Временной лаг определяется при помощи перебора парных коэффициентов корреляции между абсолютными уровнями двух рядов динамики.

Приведение данных к сопоставимому виду с точки зрения автокорреляции, коллинеарности и временного лага является предварительным условием построения многофакторной модели динамики.

Построенная с соблюдением этих условий многофакторная регрессионная модель:

$$\widehat{y_t^{(n)}} = f^{(n)}\left(x_1^{(n)}, x_2^{(n)} \dots x_2^{(n)}\right),$$

где знак (n) показывает номер этапа, будет характеризовать среднее влияние факторных признаков на результативный признак за рассматриваемый интервал времени. Величина этого влияния, выраженная коэффициентами регрессии, частными коэффициентами эластичности будет изменяться в зависимости от времени.

Относительно принципов построения регрессии между временными рядами пока еще не сложилось единого мнения. Одни ученые считают, что регрессия непосредственно между временными рядами с трендом возможна, если между ними существуют причинные отношения. Но в этом случае неясно, почему из рядов динамики, а тем самым из регрессии, должен исключаться тренд. Разумеется, в противном случае нарушается предпосылка независимости значений. С этим нарушением, однако, приходится не считаться, если регрессия на основе тренда достаточно определена и по ней получают хорошие прогнозные значения.

Другие считают, что регрессию между временными рядами следует находить после устранения тренда. Во-первых, этим достигается независимость наблюдений, во-вторых, уменьшаются или совсем исключаются возмущения. Если тренд не устраняется, то можно показать, что функция регрессии временных рядов может быть заменена функцией тренда. Тем самым учитывается только долгосрочное движение, обусловленное однообразно действующими приблизительно в одном и том же направлении силами. А сезонная и другие компоненты, вызывающие более или менее регулярные колебания относительно тренда, практически не принимаются во внимание. Очевидно, что оба мнения вполне обоснованы. Если задачей исследования является получение по возможности хороших прогнозных значений, то при построении регрессии тренд не исключается. Если же хотят установить лишь общую закономерность между зависимой и объясняющими переменными, то тренд исключается [21].

2.5 Деревья решений

2.5.1 Общие принципы работы деревьев решения

Деревья решений – один из методов анализа данных. Первые идеи создания деревьев решений восходят к работам Ховленда (Hoveland) и Ханта(Hunt) конца 50-х годов XX века. Однако, основополагающей работой, давшей импульс для развития этого направления, явилась книга Ханта (Hunt, E.B.), Мэрина (Marin J.) и Стоуна (Stone, P.J) «Experiments in Induction», увидевшая свет в 1966 г.

Деревья решений – это способ представления правил в иерархической, последовательной структуре, где каждому объекту соответствует единственный узел, дающий решение.

Под правилом понимается логическая конструкция, представленная в виде «если ... то ...» (рисунок 4).

Рисунок 4 – Пример дерева решений

Область применения деревья решений в настоящее время широка, но все задачи, решаемые этим аппаратом, могут быть объединены в следующие три класса:

– **Описание данных:** Деревья решений позволяют хранить информацию о данных в компактной форме, вместо них мы можем хранить дерево решений, которое содержит точное описание объектов.

– **Классификация:** Деревья решений отлично справляются с задачами классификации, т.е. отнесения объектов к одному из заранее известных классов. Целевая переменная должна иметь дискретные значения.

– **Регрессия:** Если целевая переменная имеет непрерывные значения, деревья решений позволяют установить зависимость целевой переменной от независимых(входных) переменных. Например, к этому классу относятся задачи численного прогнозирования (предсказания значений целевой переменной).

Пусть нам задано некоторое обучающее множество T, содержащее объекты (примеры), каждый из которых характеризуется m атрибутами (атрибутами), причем один из них указывает на принадлежность объекта к определенному классу.

Пусть через $\{C_1, C_2, \dots C_k\}$ обозначены классы(значения метки класса), тогда существуют три ситуации.

1. Множество T содержит один или более примеров, относящихся к одному классу C_k. Тогда дерево решений для T – это лист, определяющий класс C_k.

2. Множество T не содержит ни одного примера, т.е. пустое множество. Тогда это снова лист, и класс, ассоциированный с листом, выбирается из другого множества отличного от T, скажем, из множества, ассоциированного с родителем.

3. Множество T содержит примеры, относящиеся к разным классам. В этом случае следует разбить множество T на некоторые подмножества. Для этого выбирается один из признаков, имеющий два и более отличных друг от друга значений $O_1, O_2, \dots O_n$. T разбивается на подмножества $T_1, T_2, \dots T_n$, где каждое подмножество T_i содержит все примеры, имеющие значение O_i для выбранного признака. Это процедура будет рекурсивно продолжаться до тех пор, пока конечное множество не будет состоять из примеров, относящихся к одному и тому же классу.

На сегодняшний день существует значительное число алгоритмов, реализующих деревья решений CART, C4.5, NewId, ITrule, CHAID, CN2 и т.д. Но наиболее подходящий для случая работы с временными рядами – алгоритм CART [23].

CART (Classification and Regression Tree) – это алгоритм построения бинарного дерева решений – дихотомической классификационной модели. Каждый узел дерева при разбиении имеет только двух потомков. Как видно из названия алгоритма, решает задачи классификации и регрессии.

Алгоритм CART использует так называемый индекс Gini (в честь итальянского экономиста Corrado Gini), который оценивает "расстояние" между распределениями классов.

$$Gini(c) = 1 - \sum_j p_j^2,$$

где c – текущий узел, а p_j – вероятность класса j в узле с.

Если набор T разбивается на две части T_1 и T_2 с числом примеров в каждом N_1 и N_2 соответственно, тогда показатель качества разбиения будет равен:

$$Gini_{split}(T) = \frac{N_1}{N} \cdot Gini(T_1) + \frac{N_2}{N} \cdot Gini(T_2),$$

наилучшим считается то разбиение, для которого $Gini_{split}(T)$ минимально.

Обозначим N – число примеров в узле – предке, L, R – число примеров соответственно в левом и правом потомке, l_i и r_i – число экземпляров i-го класса в левом/правом потомке, получим формулу оценки качества разбиения.

$$Gini_{split} = L - \frac{1}{L} \cdot \sum_{i=1}^{n} l_i^2 + R - \frac{1}{R} \sum_{i=1}^{n} r_i^2 \to min\,,$$

$$Gini_{split} = N - \left(\frac{1}{L} \cdot \sum_{i=1}^{n} l_i^2 + \frac{1}{R} \sum_{i=1}^{n} r_i^2 \right) \to min\,.$$

Но так как минимизируется операция вычитания из N, можно преобразовать формулу и максимизировать вычитаемое:

$$Gini_{split} = \frac{1}{L} \cdot \sum_{i=1}^{n} l_i^2 + \frac{1}{R} \sum_{i=1}^{n} r_i^2 \to max.$$

В итоге, лучшим будет то разбиение, для которого величина **максимальна**.

Вектор предикторных переменных, подаваемый на вход дерева может содержать как числовые (порядковые) так и категориальные переменные. В любом случае в каждом узле разбиение идет только по одной переменной.

На каждом шаге построения дерева алгоритм последовательно сравнивает все возможные разбиения для всех атрибутов и выбирает наилучший атрибут и наилучшее разбиение для него [23].

2.5.2 Применение деревьев решений в задачах прогнозирования многомерных временных рядов

Пусть для описания некоторого объекта исследования используется набор случайных характеристик $X(t) = \big(X_1(t), \dots, X_n(t)\big)$, значения которых меняются с течением времени. Следует добавить, что характеристики могут быть как количественными, так и качественными.

Предположим, что характеристики измеряются в последовательные моменты времени $t_1, t_2, \dots, t_k$. Для определенности будем предполагать, что измерения проводятся через равные интервалы времени. Обозначим через $x_j(t_k) = X_j(t_k)$ значение характеристики X_j в момент времени t_k. Таким образом, имеем n-мерный разнотипный временной ряд $x_j(t_k), j = 1, \dots, n, k = 1,2, \dots$.

Также необходима некоторая прогнозируемая характеристика $X_{j_0}, 1 \le j_0 \le n$. Обозначим, для удобства, эту характеристику через Y.

Рассмотрим произвольный момент времени t_k, а также набор предыдущих моментов времени $t_{k-1}, t_{k-2}, \dots, t_{k-l}$, где l – величина лага.

Кроме того, предположим, что эта зависимость одна и та же для любых k. Данное предположение означает, что статистические свойства ряда, определяющие зависимость, неизменны во времени [24].

Требуется построить модель зависимости характеристики Y от ее предыстории для произвольного момента времени. Модель позволяет прогнозировать значение характеристики Y в будущий момент времени по значениям характеристик за l прошлых моментов. Иначе говоря, данная модель представляет собой решающую функцию для прогнозирования по предыстории в виде дерева решений.

В дереве решений, построенном по таким входным параметрам проверяются высказывания относительно некоторых характеристик X_j в некоторый j-й предыдущий отсчет времени.

Итак, пусть имеется набор измерений характеристик $X(t) = \left(X_1(t), \dots, X_n(t)\right)$ в моменты времени $t_1, t_2, \dots, t_k$ и задано значение l. Таким образом, имеем многомерный разнотипный временной ряд длины N. Сформируем множество всех предысторий длины l для моментов времени $.t_{l+1}, t_{k+2}, \dots, t_N$.

Для любого заданного дерева решений в задаче прогнозирования по предыстории $t_{l+1}, t_{k+2}, \dots, t_N$ можно определить его качество [24]. Критерий качества:

$$Q = \frac{1}{N-l} \cdot \sum_{k=l+1}^{N} h(k),$$

где

$$h(k) = \left(Y(t^k) - \hat{Y}(t^k)\right)^2.$$

Для оценки точности прогноза подходит формула MAPE.

2.6 Анализ рассмотренных подходов

ARIMA-модели являются одним из основных методов при работе с временными рядами, в частности, для более глубокого понимания данных или предсказания будущих точек ряда. Этому способствуют главные положительные качества моделей данного типа:

– Такие модели имеют подробное математико-статистическое обоснование, из чего следует, что модели данного типа одни из наиболее научно обоснованных из всего множества моделей прогнозирования тенденций во временных рядах.

– Формализованная и достаточно четко описанная методика, отталкиваясь от которой можно путем изменения параметров модели ARIMA(p, d, q) подобрать модель, наиболее подходящую к каждому конкретному временному ряду,

– точечные и интервальные прогнозы следуют из самой модели и не требуют отдельного оценивания.

С другой стороны, к недостаткам данного типа моделей можно отнести требовательность к вычислительным ресурсам, времени и количеству данных в исследуемом временном ряде, а также то, что при получении новых данных модель

нужно время от времени переоценивать, а в некоторых случаях – заново проводить расчет параметров модели.

В данной выпускной квалификационной работе проводится сравнение с одной стороны ARIMA-моделей, с другой, моделей множественной регрессии и деревьев решений, которые также имеют как свои преимущества, так и недостатки.

Одним из главных плюсов моделей множественной регрессии и деревьев решений является их быстрый процесс обучения и сравнительно небольшая потребность в вычислительных ресурсах, в отличие от ARIMA-моделей. Другим преимуществом данных моделей является высокая точность прогноза, сопоставимая с другими методами при правильно подобранных факторах.

Однако тема применения деревьев решения и множественной регрессии для задач анализа и прогнозирования многомерных временных рядов является дискуссионной [21]. Данное направление мало описана в литературе. Рассматриваемая тема находит как противников, так как этот подход не является точно описанным, так и сторонников, так как существуют примеры применения этого подхода на практике, которые являются не только рабочими, но и иногда более точными по сравнению с более традиционными методами анализа временных рядов [18].

Выводы по главе два

1. ARIMA-модели, применяющиеся в работе имеют подробное математико-статистическое обоснование, а также гибкость и универсальность в работе с временными рядами, модели данного вида являются одним из основных методов при работе с временными рядами. Однако требовательность как к вычислительным ресурсам, так и к объему выборки, вкупе с дополнительной работой по периодической переоценке модели при получении новых данных не позволяют говорить об незаменимости такого подхода к анализу временных рядов.

2. Модели множественной регрессии и решающие деревья в области анализа временных рядов являются достаточно спорными методами с недостаточно хорошо описанной базой. Данные модели более требовательны к параметрам анали-

зируемых рядов динамики: необходимо приведение данных к сопоставимому виду с точки зрения автокорреляции, коллинеарности и временного лага. С другой стороны, при подходящих условиях данные модели показывают себя как менее требовательные к вычислительным ресурсам и времени альтернативы ARIMA-моделей.

3 ОПИСАНИЕ И ПРЕДВАРИТЕЛЬНЫЙ АНАЛИЗ ДАННЫХ

Данные в работе представлены интенсивностью сетевого трафика (в битах) от частного интернет-провайдера с центрами в 11 европейских городах. Данные были собраны с 6:57 часов 7 июня до 11:17 часов 31 июля 2015 года. Наблюдения проводились с интервалом в пять минут. Итоговая выборка содержит в себе 14548 наблюдения (рисунок 6).

Исходный временной ряд содержит абсолютные величины интенсивности трафика, данный ряд является моментным с равностоящими уровнями по времени.

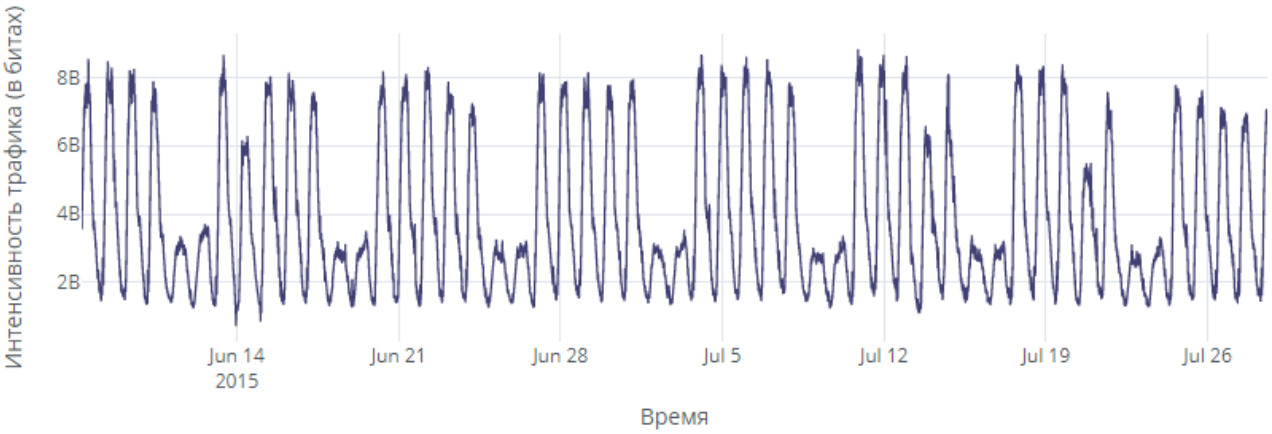

Рисунок 6 – График исходных данных

3.1 Анализ характеристик исходного ряда динамики

3.1.1 Сезонность

Исходные данные имеют ярко выраженную дневную и недельную сезонность, что характерно для временного ряда интенсивности трафика. Изучение сезонных колебаний имеет самостоятельное значение как исследование особого типа динамики. Выявление сезонной составляющей может быть произведено на основе нескольких методов. Одним из подходов к выявлению сезонности является

расчет отношений недельных (дневных) средних ($\bar{y}_t$) к средней за весь период. Такое отношение выражается в виде индекса сезонности:

$$I_s = \frac{\bar{y}_t}{\bar{y}} \cdot 100,$$

где $\bar{y}_t$ – средняя для каждого дня недели (дня) t, $\bar{y}$ – общий недельный уровень за весь период. Применив данный метод для анализа недельной сезонности, при $\bar{y} = 3972605053$ получим:

$$I_1 = 117\%,$$

$$I_3 = 113\%,$$

$$I_3 = 110\%,$$

$$I_4 = 109\%,$$

$$I_5 = 93\%,$$

$$I_6 = 67\%,$$

$$I_7 = 89\%.$$

Основываясь на данном показателе, можно делать вывод касательно общей структуры сезонности ряда динамики. Например, в 5-ый день недели, в общем случае, наблюдается снижение интенсивности трафика относительно пиковых величин интенсивности трафика в данный конкретный день недели.

3.1.2 Стационарность

Случайные процессы, протекающие во времени приблизительно однородно и имеющие вид непрерывных случайных колебаний вокруг некоторого среднего значения (причем ни средняя амплитуда, ни характеристика этих колебаний не обнаруживают существенных изменений с течением времени) в математической статистике называются стационарными [25].

Одним из часто употребляемых критериев для проверки на стационарность временных рядов является тест Дики-Фуллера и его расширенный аналог (ADF-тест), которые по своей сути являются тестами на единичные корни.

Смысл DF-теста заключается в поиске единичного корня исходя из уравнения:

$$\Delta y_t = (a - 1) \cdot y_{t-1} + \varepsilon_t б,$$

где Δ - оператор разности первого порядка $\Delta y_t = y_t - y_{t-1}$, a – проверочный коэффициент, если $a = 1$, то процесс имеет единичный корень, как следствие ряд не стационарен.

Результаты проверки на стационарность критерием Дики-Фуллера представлены в таблице 1:

Таблица 1 - Значение Критерия Дики-Фуллера

Значение Критерия Дики-Фуллера:	-12.948323
p-value:	0.000000

При критическом значении в $1\% = -3{,}4$ можно сделать вывод о том, что гипотеза о стационарности данного временного ряда принимается с 99% вероятностью.

3.1.3 Автокорреляция (АКФ)

Коэффициенты автокорреляции отражают степень тесноты связи между уровнями исходного временного ряда и уровнями ряда, сдвинутыми на один или несколько временных промежутков назад [26]

На основе автокорреляционной функции также можно делать выводы о наличии сезонной составляющей ряда динамики.

$$r_j = \frac{\sum_{t=j+1}^{n} \left(y_t - \bar{y}_{1j}\right) \cdot \left(y_{t-j} - \bar{y}_{2j}\right)}{\sqrt{\sum_{t=j+1}^{n} \left(y_t - \bar{y}_{1j}\right)^2 \cdot \sum_{t=j+1}^{n} \left(y_{t-j} - \bar{y}_{2j}\right)^2}},$$

где τ – порядок коэффициента корреляции или лаг (величина сдвига величина сдвига уровней ряда во времени),

$$\bar{y}_{1j} = \frac{\sum_{t=j+1}^{n} y_t}{n-j}, \qquad \bar{y}_{2j} = \frac{\sum_{t=j+1}^{n} y_{t-j}}{n-j}.$$

Построив график автокорреляционной функции (кореллограммы) можно визуально проанализировать временной ряд на наличие отдельных компонент функции автокорреляции (рисунок 7-8).

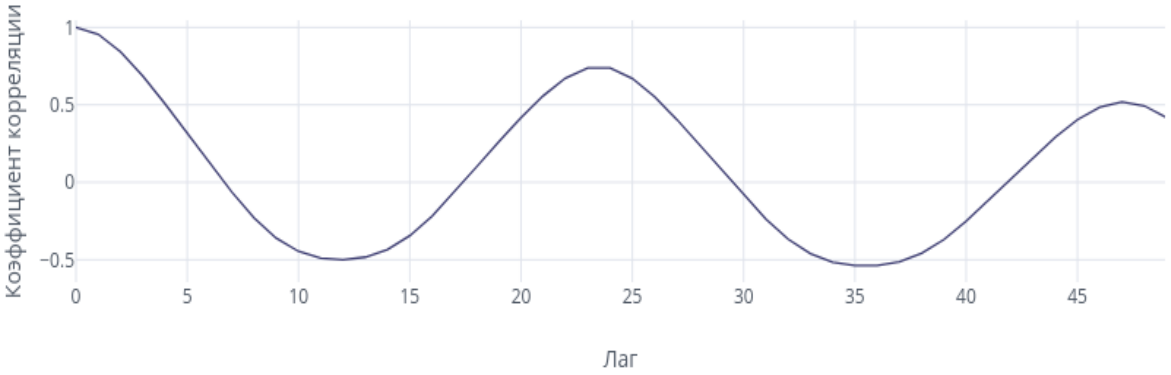

Рисунок 7 – График автокорреляционной функции для временного ряда интенсивности интернет-трафика с периодичностью значений в 1 час (диапазон значений – 2 дня)

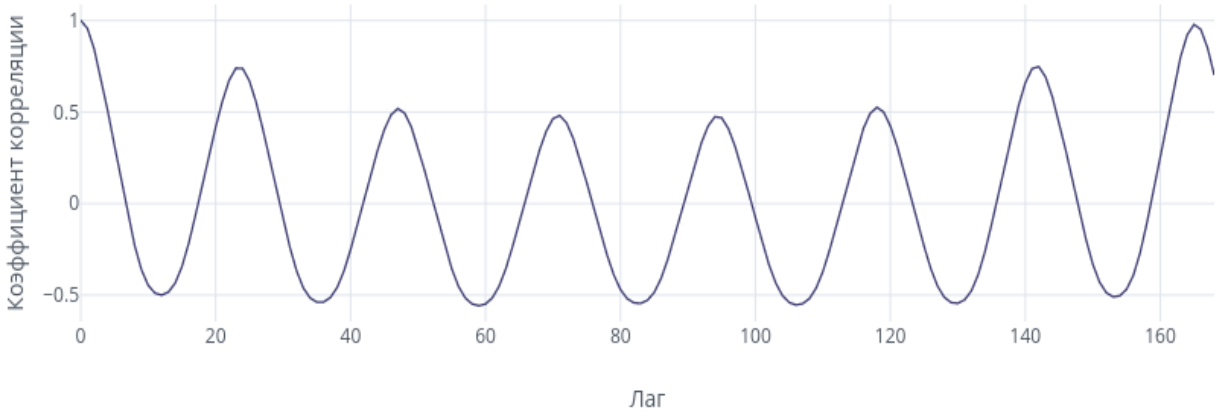

Рисунок 8 – График автокорреляционной функции для временного ряда интенсивности интернет-трафика с периодичностью значений в 1 час (диапазон значений – 1 неделя)

Анализируя графики автокорреляционной функции можно сделать вывод о наличии в больших количествах значимых лагов, соответствующих сезонным ко-

лебаниям, в следствие этого прежде чем полученное уравнение регрессии можно будет использовать для прогноза, необходимо устранить автокорреляцию.

Для того чтобы устранить серийную корреляцию сильно автокоррелирующих данных, можно также использовать в расчетах не сами значения ряда, а их разности (глава 2).

$$\Delta y_k = y_{t-k+1} - y_{t-k}, \qquad \Delta x_k = x_{t-k+1} - x_{t-k}.$$

Используя такой подход, строится еще один график автокорреляционной функции, но уже для ряда первых разностей (рисунок 9).

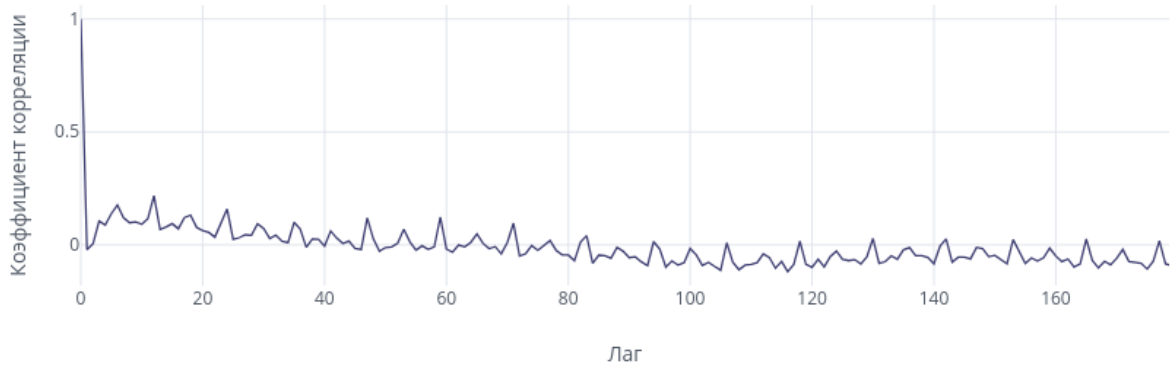

Рисунок 9 – График автокорреляционной функции для временного ряда первых разностей интенсивности интернет-трафика с периодичностью значений в 1 час (диапазон значений – 1 неделя)

В результате взятия первых разностей, автокорреляционная функция полностью избавилась от значимых лагов, что позволяет говорить о допустимости использования временного ряда первых разностей интенсивности интернет-трафика для построения многофакторных регрессионных моделей.

Также для устранения влияния времени на результат и факторы при изучении взаимосвязанных рядов динамики используется прием включения времени t в качестве независимой переменной в модель регрессии, что позволяет зафиксировать воздействие фактора t. Достоинством такого подхода является использование всей имеющейся выборки в отличие от метода последовательных разностей, который приводит к потере некоторого числа наблюдений.

Согласно проведенному автокорреляционному исследованию можно сделать следующие выводы:

1. Наиболее высоким оказался коэффициент автокорреляции первого порядка, что можно трактовать как возможное наличие тенденции во временном ряду спроса. Однако, существуют другие лаги с близким коэффициентом, поэтому, можно сделать вывод о том, что, если тенденция существует, то она практически не оказывает никакого влияния.

2. Существуют последовательные всплески в автокорреляционной функции с периодичностью 24 и 168 часов, следовательно, временной ряд интенсивности интернет-трафика содержит циклические колебания с периодом в 1 день и 1 неделю.

Выводы по главе три

1. Анализ исходного временного ряда интенсивности интернет-трафика позволяет говорить о наличии явной дневной и недельной сезонности, что в дополнение подтверждается графиком автокорреляционной функции. Для недельной сезонности также был применен расчет отношений недельных (дневных) средних ($\bar{y}_t$) к средней за весь период, на основе которого можно говорить о том, что в пятницу (y_5), в общем случае, наблюдается тенденция к снижению интенсивности интернет-трафика.

2. На основе проведенной проверки на стационарность с помощью критерия Дики-Фуллера, можно охарактеризовать ряд как стационарный.

3. Для построения моделей многофакторной регрессии подходит временной ряд первых разностей. Альтернативой служит включение в регрессионную модель дополнительных факторов времени, для устранения влияния времени на результат и используемые в построении регрессионной модели факторы.

4 ПОСТРОЕНИЕ МОДЕЛЕЙ ПРОГНОЗИРОВАНИЯ

Перед тем, как переходить к построению моделей прогнозирования временных рядов следует уточнить, что при использовании **статистических моделей**, в результате получается зависимость будущего значения от прошлого задается в виде некоторого уравнения.

В данной работе к статистическим моделям относятся модели ARIMA во всех ее вариациях и модели множественной регрессии.

В работе также используются модели на базе регрессионных деревьев, которые относятся к классу **структурных моделей**, в которых зависимость будущего значения от прошлого задается в виде некоторой структуры и правил перехода по ней как выглядит итоговая.

4.1 Описание параметров моделей

4.1.1 Параметры ARIMA модели

Как уже было описано, для модели ARIMA кроме входных данных, необходимо дополнительно задать параметры p, d, q. В модели Бокса-Дженкинса за авторегрессионную компоненту ϕ_i отвечает полином степени p. За компоненту процесса скользящего среднего θ_j – полином q. Δ^d - оператор разности временного ряда порядка d (последовательное взятие d раз разностей d-го порядка).

Существует метод, при котором данные параметры не задаются в ручную, а определяются путем последовательного построения моделей ARIMA с различными комбинациями p, d, q, после чего выбирается наилучшая модель на основе некоторой оценки. В данной работе будет использован именно рассмотренный метод подбора параметров для построения оптимальной модели. За оценку качества при последовательном построении и выборе лучшей модели примем критерий Акаике.

4.1.2 Параметры деревьев решения

Основные параметры деревьев решения включают в себя:

– максимальную глубина дерева;

– максимальное число признаков, по которым ищется лучшее разбиение в дереве (это необходимо так как, при большом количестве признаков следует повышение сложности в поиске лучшего (по критерию типа прироста информации) разбиения среди всех признаков);

– минимальное число объектов в листе (лист не будет создан, если количество объектов в нем меньше, чем заданное).

4.1.3 Генерация признаков для многофакторных моделей

Для многофакторных регрессионных моделей и деревьев решения, входной массив данных должен содержать несколько факторных переменных. Принимая во внимание зависимость значений временного ряда от времени, можно добавить несколько отражающих эту зависимость факторов в виде лаговых переменных. Также существует возможность добавления фиктивных переменных, к примеру, отражающих выходные в недельной периодичности.

4.2 Построение моделей и анализ результатов

4.2.1 Применение метода скользящего среднего для первичного анализа исходного ряда динамики

В предыдущих главах было описано, что исходный временной ряд интенсивности интернет-трафика имеет дневную и недельную сезонность. На основе этого логично применить метод скользящего среднего с окном скольжения равным сезонности. В таком случае, используем метод скользящего среднего с окном равным $12 \cdot 24$, что равно 1 дню (интервал значений в временном ряду равен пяти минутам) и $12 \cdot 24 \cdot 7$, что равно 1 неделе. Результаты представлены на рисунках 10 и 11.

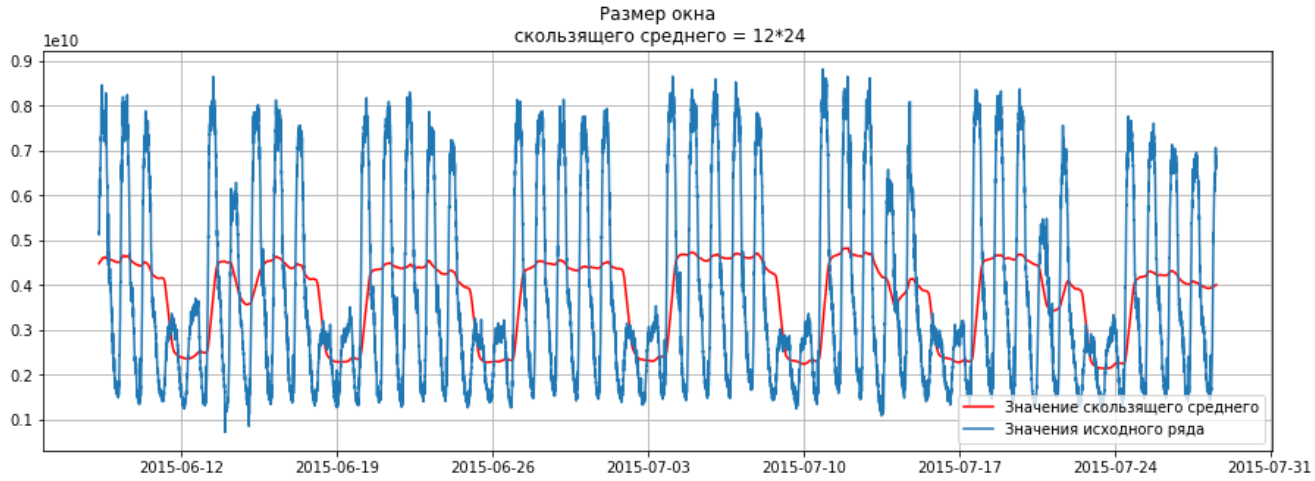

Рисунок 10 – Применение метода скользящего среднего с окном скольжения равным одному дню

Рисунок 11 – Применение метода скользящего среднего с окном скольжения равным одной неделе

На представленных рисунках отражены общие тенденции исходного временного ряда. К примеру, на рисунке 11 скользящее среднее с окном скольжения равным одной неделе усреднила все сезонные колебания и теперь представляет собой средний уровень интернет-трафика. В свою очередь, можно сделать вывод, что средний уровень интернет-трафика к концу временного ряда несколько ниже, чем в его середине.

4.2.2 Построение модели ARIMA

Так как проверка исходного ряда на стационарность показала, что ряд стационарен, то построим как модель ARMA(p, q), так и модель ARIMA(p, d, q), после чего сравним модели по критерию Акаике, а прогноз моделей по MAPE и MSE.

Для подбора оптимальных параметров для модели ARMA используем аналогичный метод перебора, однако параметр d примем равным 0, в отличие от модели ARIMA, где параметр d свободно изменяется в пределах от 0 до 1.

Использование в работе по прогнозированию стационарного ряда модели ARIMA с $d > 0$ обусловлено тем, что на графике автокорреляционной функции при $d = 0$ не наблюдается быстрого затухания (стационарности) АКФ-функции для $d = 0$, что в соответствии с работой Дж.Бокс и Г.Дженкинс является визуальным критерием нарушения стационарности. При $d = 1$ АКФ-функция быстро затухает, что позволяет говорить о возможности использования данного значения для параметра d. Следует также учитывать, что использование завышенного порядка разности d приводит к росту дисперсии ошибок и к заметному росту дисперсии прогноза.

Применение метода последовательного перебора параметров модели ARMA показало, что подходящей моделью ARMA является модель с параметрами $p = 5, q = 5$ и значением критерия $AIC = 594281$. В таблице ниже представлены параметры p и q, при которых модель ARMA имеет схожие с лидирующей комбинацией параметров характеристики качества (таблица 2).

Таблица 2 – Близкие по качеству модели ARMA

ARMA(p, q)	***Значение критерия Акаике (AIC)***
ARMA(6, 1)	*594296*
ARMA(5, 4)	*594304*
ARMA(5, 3)	*594586*

Для модели ARIMA применение метода последовательного перебора параметров показало, что оптимальной моделью ARIMA является модель с параметрами $p = 2, d = 1, q = 3$ и значением критерия $AIC = 594138$. В таблице ниже представлены параметры p и q, при которых модель ARIMA имеет схожие с лидирующей комбинацией параметров характеристики качества (таблица 3).

Таблица 3 – Близкие по качеству модели ARIMA

ARIMA(p, d, q)	***Значение критерия Акаике (AIC)***
ARIMA(2, 1, 2)	*594178*
ARIMA(2, 1, 1)	*594246*
ARIMA(5, 0, 5)	*594281*

Подав исходные данные интенсивности интернет-трафика на вход лучшим моделям ARMA и ARIMA можно использовать данные модели для предсказания интенсивности интернет-трафика.

Однако для того чтобы оценить точность предсказания модели, например, по формуле MSE (mean squared error) или MAPE (mean absolute percentage error) необходимо иметь в наличии как модельные значения, так и действительные. В таком случае следует отделить часть последних значений исходного временного ряда, равную количеству предсказываемых значений. Для текущего прогноза – 5 минут, т.е. одно предсказанное значение, для краткосрочного прогноза – 30 минут, т.е. шесть предсказанных значений, для долгосрочного прогноза – 3 часа (36 значений).

Та часть исходных данных, которая отделяется в следствие этого, не используется для обучения модели, ее называют тестовой частью и используют для вычисления MSE и MAPE прогноза, что является показателем качества всей модели.

Таким образом формула итоговой модели ARMA имеет вид:

$$X_t = 3.82X_{t-1} - 6.48X_{t-2} + 6.35X_{t-3} - 3.60X_{t-4} + 0.92X_{t-5} - 2.06\varepsilon_{t-1} + 1.91\varepsilon_{t-2} - 0.87\varepsilon_{t-3} - 0.24\varepsilon_{t-4} + 0.32\varepsilon_{t-5} + \varepsilon_t,$$

а формула итоговой модели ARIMA:

$$\Delta^1 X_t = 1.47\Delta^1 X_{t-1} - 1.52\Delta^1 X_{t-2} - 1.24\Delta^1 \varepsilon_{t-1} - 0.56\Delta^1 \varepsilon_{t-2} - 0.35\Delta^1 \varepsilon_{t-3} + \varepsilon_t.$$

После получения итоговых формул моделей, можно получить прогнозные значения на период тестовой части исходных значений, после чего можно оценить ошибку модели и сделать вывод о качестве прогноза:

Оценка качества прогноза представлена в таблице 4.

Таблица 4 – Качество прогноза лучших моделей ARIMA и ARMA

	Прогноз на 1 шаг вперед		***Прогноз на 6 шагов вперед***		***Прогноз на 36 шагов вперед***	
Вид модели	*MSE*	*MAPE%*	*MSE*	*MAPE%*	*MSE*	*MAPE%*
ARMA(5, 5)	91207003	4,41	158828678	8,15	372676739	13,46
ARIMA(2, 1, 3)	91827458	4,44	158531538	7,94	389551861	13,67

Построим график исходных данных и прогноза, с использованием лучшей модели (рисунок 12 и 13):

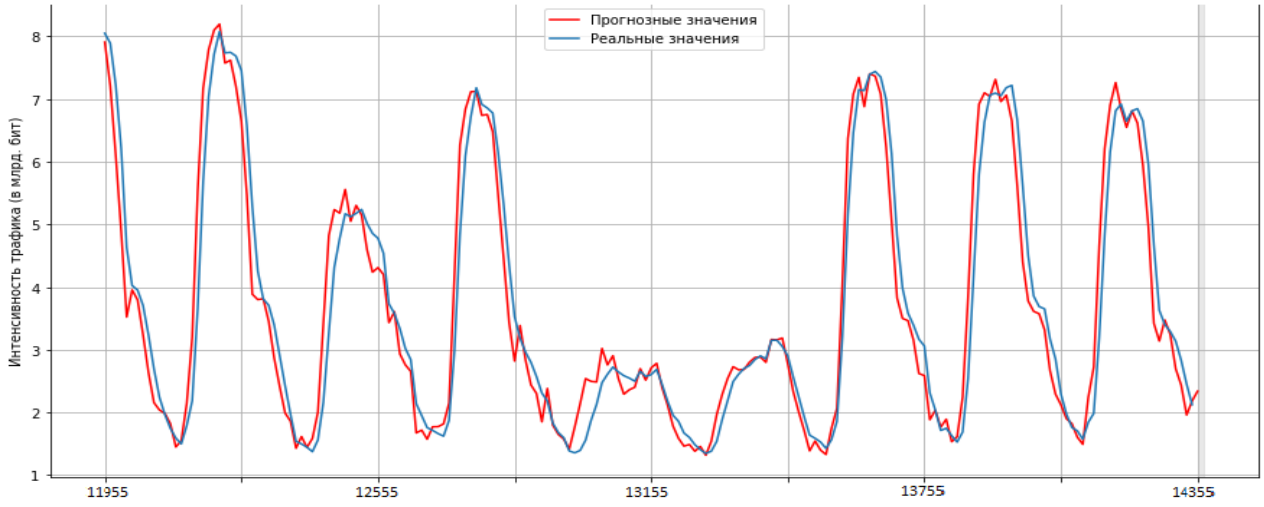

Рисунок 12 – График исходных и прогнозных значений интенсивности интернет-трафика с использованием модели ARMA(5,5) и прогнозом на 6 шагов вперед

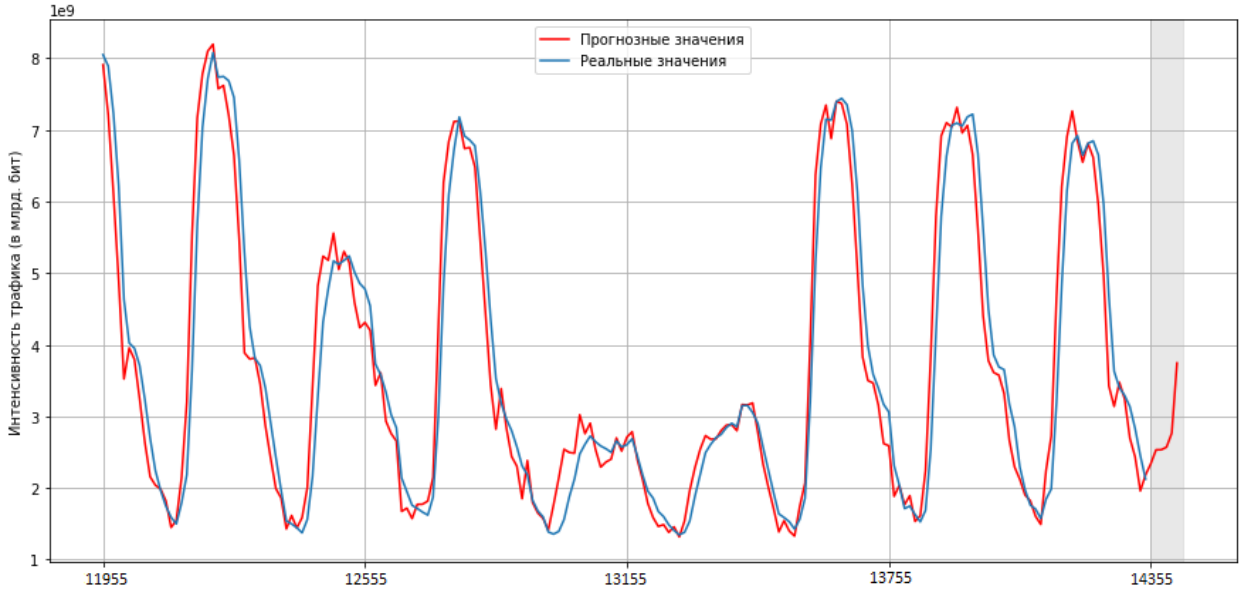

Рисунок 13 – График исходных и прогнозных значений интенсивности интернет-трафика с использованием модели ARMA(5,5) и прогнозом на 36 шагов вперед

4.2.3 Построение модели множественной регрессии

Возьмем к примеру 12 лагов начиная с 36 лага и далее (для долгосрочного прогноза), т.е. 36-47 лаги, это позволит нам предсказать следующие 36 значений трафика, имея в наличии 36 известных. Также добавим фиктивные факторы, отвечающие за выходной день.

Во временных рядах автокорреляция создает более серьезные трудности для применения обычного метода наименьших квадратов.

Результат проверки временного ряда первых разностей на автокорреляцию приведены на рисунке 9. Поскольку полученный ряд не содержит значимых значений автокорреляционной функции, его использование вместо исходных данных оправдано.

Однако другой метод устранения влияния времени на результат, заключающийся в включении в регрессионную модель дополнительных факторов времени является более предпочтительным из-за того, что используется все значения выборки в отличие от метода последовательных разностей, который приводит к по-

тере некоторого числа наблюдений. В данной работе будет использоваться именно данный метод.

Как и для модели ARIMA отложим несколько последних реальных значений временного ряда интенсивности интернет-трафика для тестирования качества получившейся модели множественной регрессии.

Применяя модель множественной регрессии к многомерному ряду первых разностей выбранных факторов в результате получаем формулу по расчету модельных значений для долгосрочного прогноза:

$$\hat{y} = -1.3156t_{36} + 1.4716t_{37} - 0.5531t_{38} - 0.1157t_{39} - 0.1790t_{40} + 0.1002t_{41} + 0.1930t_{42} - 0.1707t_{43} - 0.0853t_{44} - 0.4927t_{45} - 0.7628t_{46} + 0.1513t_{47} + 487676451z,$$

где $t_i, i = \overline{12,23}$ – лаговые переменные времени, z – фиктивная переменная, отвечающая за выходные дни (1 – выходной, иначе 0).

По приведенной выше формуле построим прогноз модели, в котором количество прогнозных значений равно количеству значений в отложенной тестовой части реальных данных.

Оценку качества модели будем проводить с использованием формул MSE и MAPE (таблица 5).

Таблица 5 – Качество прогноза модели множественной регрессии

Прогноз на 1 шаг вперед		***Прогноз на 6 шагов вперед***		***Прогноз на 36 шагов вперед***	
MSE	*MAPE%*	*MSE*	*MAPE%*	*MSE*	*MAPE%*
486315019	*16.46*	*828153705*	*28.03*	*843285628*	*25.62*

Построим график исходных данных и прогноза, с использованием многофакторной регрессионной модели (рисунок 14 и 15):

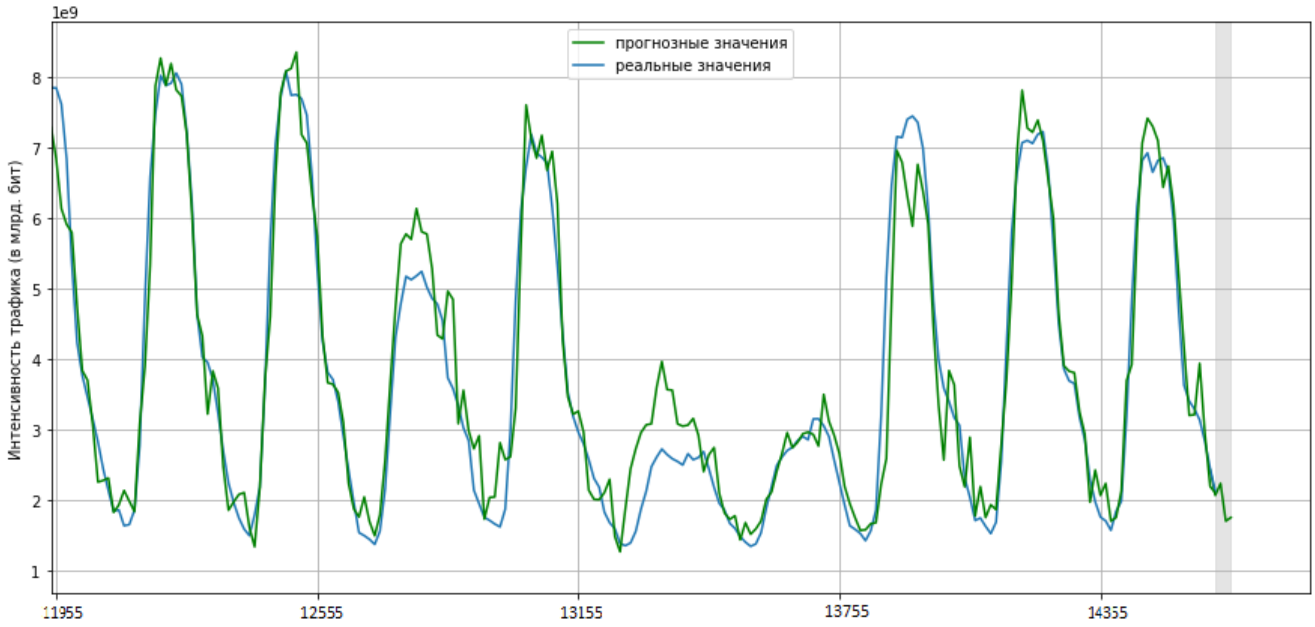

Рисунок 14 – График исходных и прогнозных значений интенсивности интернет-трафика с использованием многофакторной регрессионной модели и прогнозом на 6 шагов вперед

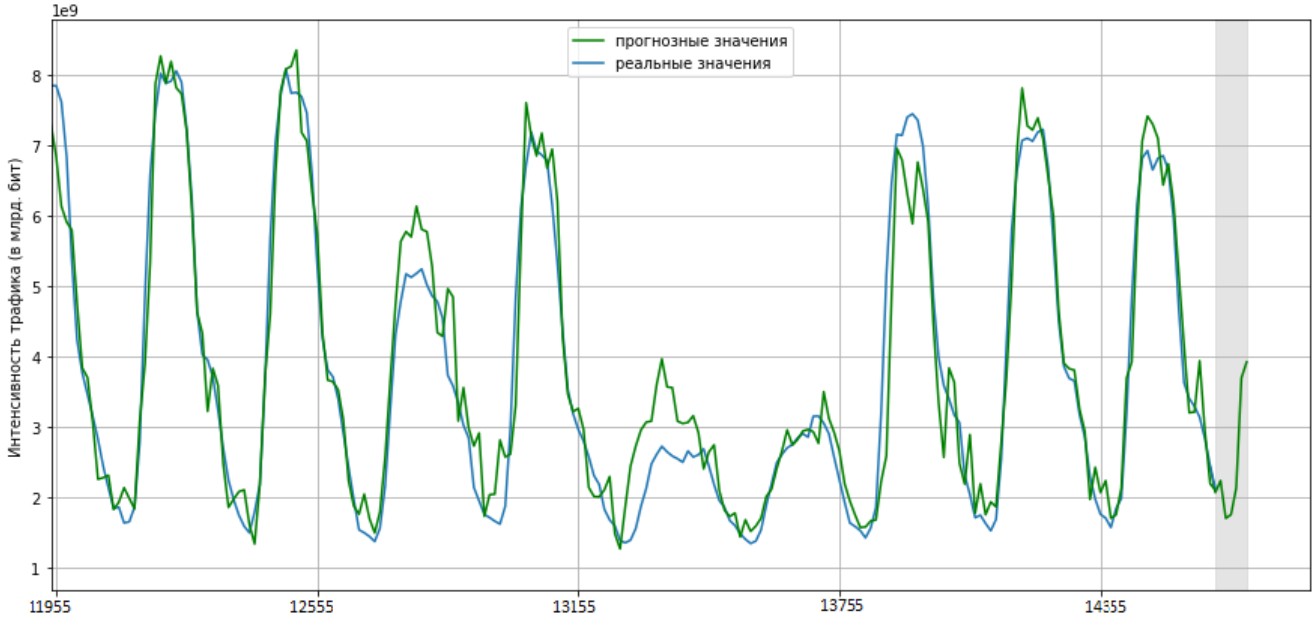

Рисунок 15 – График исходных и прогнозных значений интенсивности интернет-трафика с использованием многофакторной регрессионной модели и прогнозом на 36 шагов вперед

4.2.4 Построение дерева решений

Для построения дерева решений возьмем аналогичные факторы, что и для модели множественной регрессии: с 12 по 24 лаговые переменные и фиктивные факторы, отвечающие за выходной день.

На основе результатов нескольких тестовых деревьев решений примем, что:

– оптимальный параметр максимальной глубины дерева равен 10,

– максимальное число признаков, по которым ищется лучшее разбиение в дереве равно 14,

– минимальное число объектов в листе равно 3.

Обучение дерева решений будем проводить аналогично с ранее рассмотренными моделями: разбивая общий массив исходных данных на тренировочную и тестовую выборки. Оценка качества модели будет производиться по уже использовавшимся формулам MSE и MAPE.

После обучения дерева решений с такими параметрами имеем следующие результаты оценки качества предсказания модели (таблица 6):

Таблица 6 – Качество прогноза дерева решений

Прогноз на 1 шаг вперед		***Прогноз на 6 шагов вперед***		***Прогноз на 36 шагов вперед***	
MSE	*MAPE%*	*MSE*	*MAPE%*	*MSE*	*MAPE%*
10211945	*5,31*	*202508074*	*10,53*	*320480309*	*13,74*

Построим график исходных данных и прогноза, с использованием дерева решений (рисунок 16 и 17):

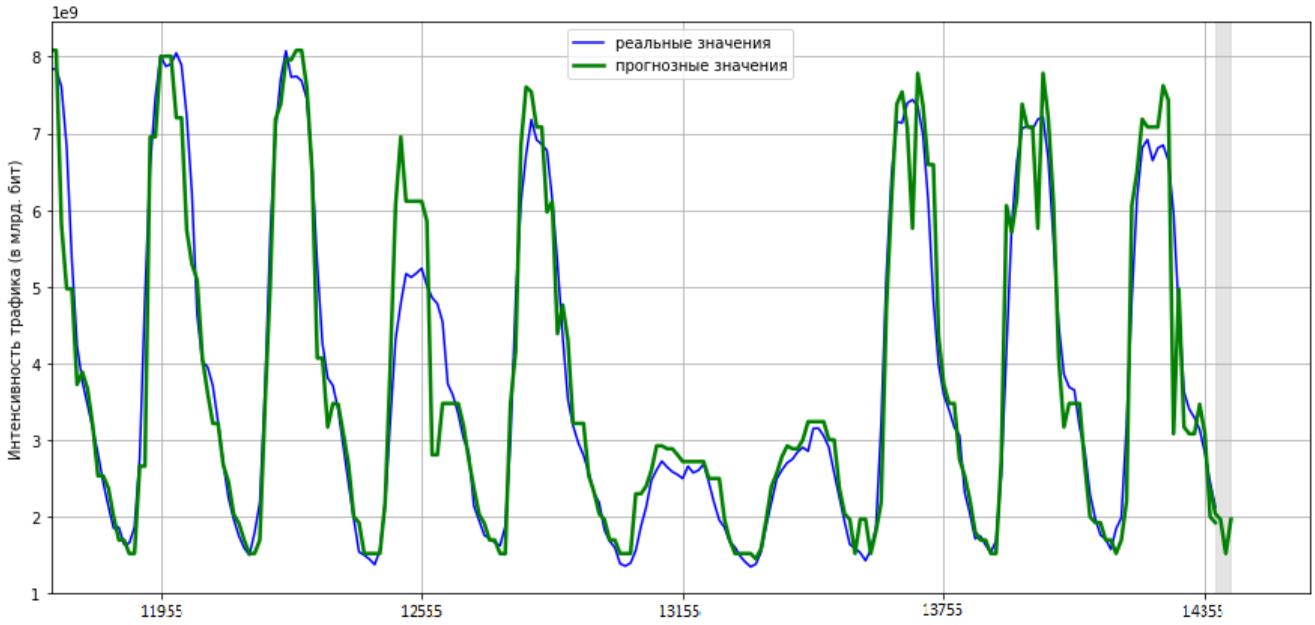

Рисунок 16 – График исходных и прогнозных значений интенсивности интернет-трафика с использованием дерева решений и прогнозом на 6 шагов вперед

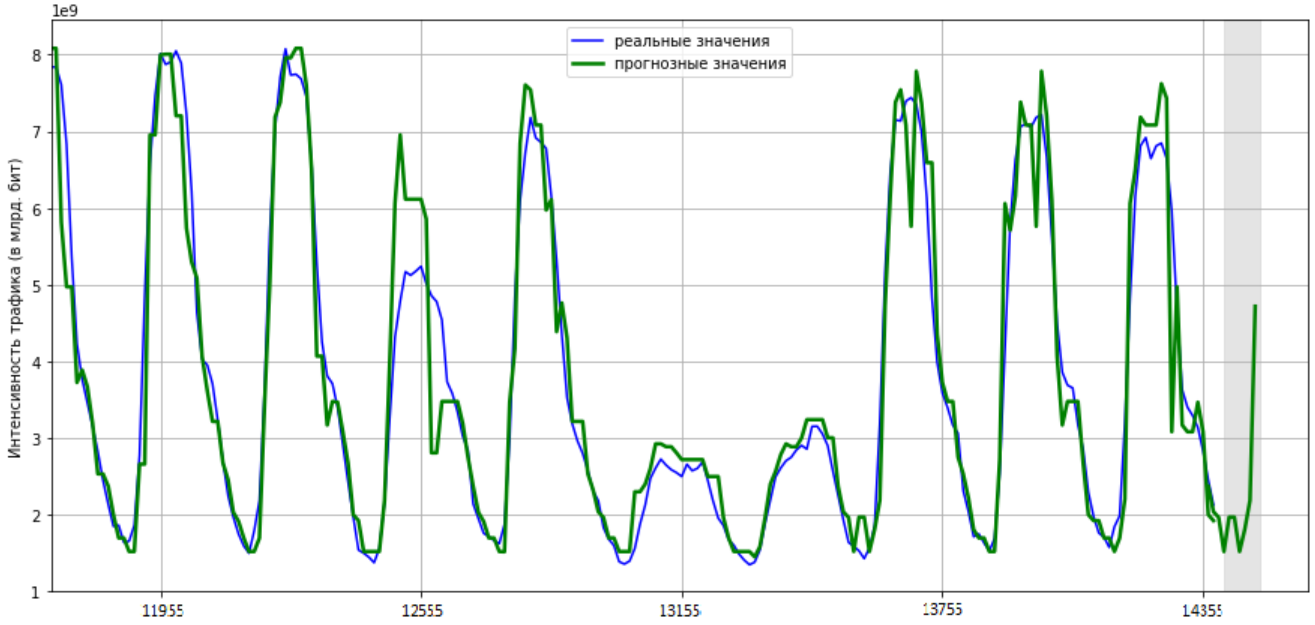

Рисунок 17 – График исходных и прогнозных значений интенсивности интернет-трафика с использованием дерева решений и прогнозом на 36 шагов вперед

4.3 Проверка адекватности моделей

Если процесс построения выбранных моделей успешно осуществлен, возникает проблема оценки качества построенной модели. Для правильно построенной модели остатки должны быть случайными, распределенными по нормальному закону, быть независимыми между собой и иметь равное нулю среднее значение. Другими словами, остатки должны быть «белым шумом», т.е. их выборочные автокорреляции не должны значимо отклоняться от нуля.

Остатки – это разности между наблюдаемыми значениями и значениями, предсказанными изучаемой моделью.

Кроме того, модель не должна содержать лишних параметров, т.е. нельзя уменьшить число параметров без появления значимой автокорреляции остатков.

Рассмотрим подробный график характеристик остатков для модели ARMA(5,5) (рисунок 18).

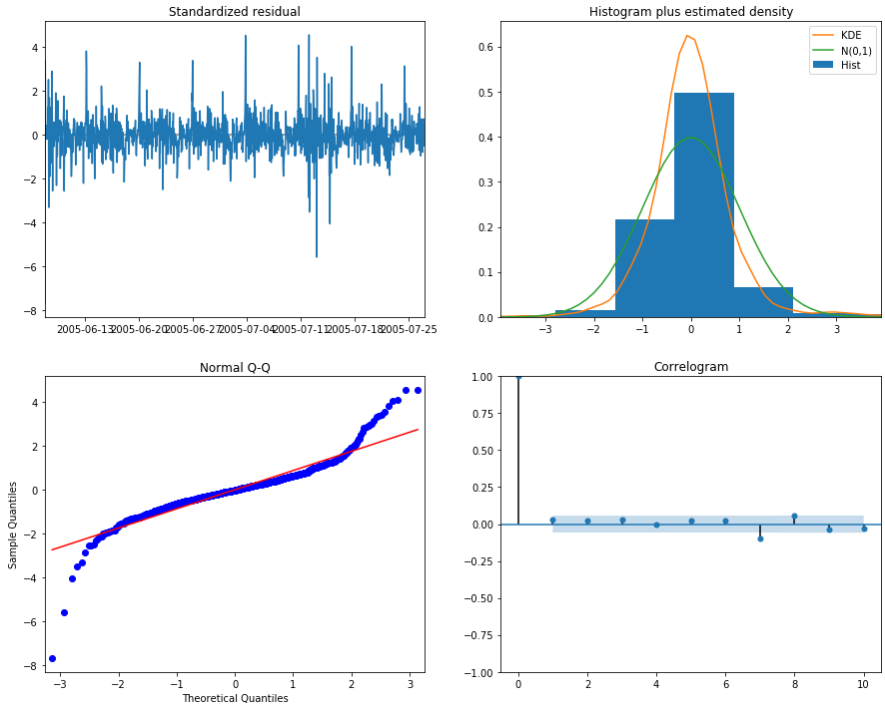

Рисунок 18 – Характеристика остатков модели ARMA(5,5)

По графику можно сказать, что остатки модели, в общем случае представляют собой белый шум, а выборочные автокорреляции ошибок значимо не отличаются от 0. Хоть распределение остатков отклоняется от нормального закона, это

отклонение является не значительным. Данная модель является достаточно адекватной и её можно применять для дальнейшего прогнозирования значений интенсивности интернет-трафика.

Построим аналогичный график характеристик остатков для модели ARIMA (2,1,3) (рисунок 19).

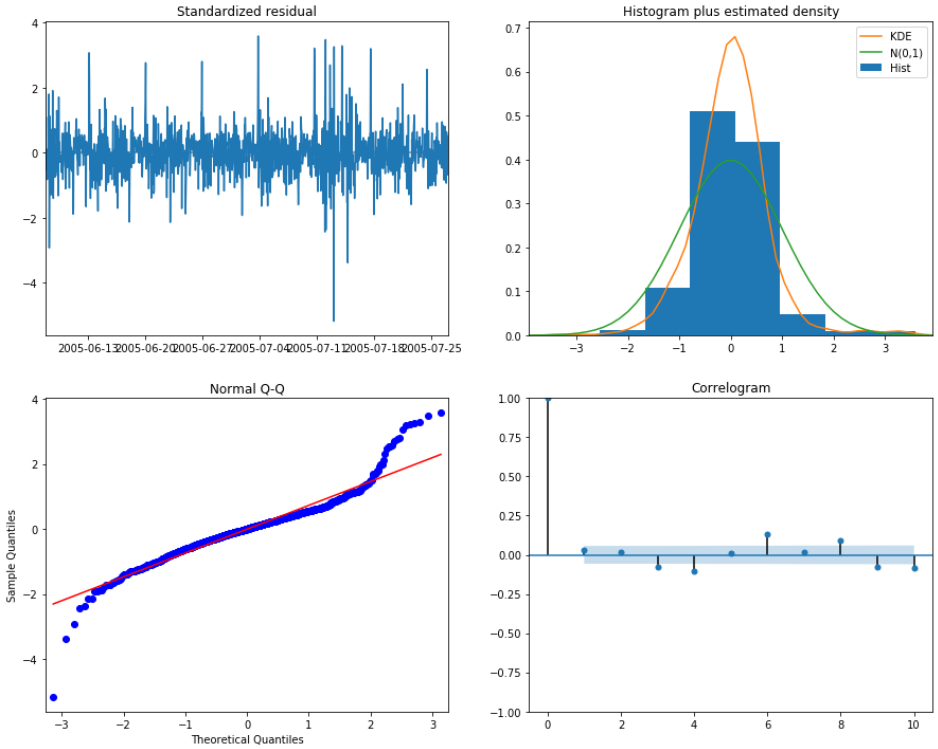

Рисунок 19 – Характеристика остатков модели ARIMA(2,1,3)

Сравнивая характеристики остатков модели ARIMA(2,1,3) с ARMA(5,5), можно сказать, что модель ARIMA(2,1,3) является в этом плане менее предпочтительной, так как распределение остатков сильнее отклоняется от нормального распределения, а на коррелограмме наблюдаются более значимые значения автокорреляции, чем в модели ARMA(5,5).

Более подробно следует разобрать характеристики остатков модели множественной регрессии и дерева решений, так как в прошлых главах были описаны ситуации, при которых теоретически можно моделировать временные ряды с помощью данных моделей.

Следует также учитывать, что временная составляющая, как уже было описано, автоматически вносит значимые значения в коррелограмме, поэтому модель множественной регрессии можно считать адекватной, только в том случае, когда значимых значений автокорреляционной функции остатков нет, либо их количество достаточно мало, в этом случае можно сказать, что выбранная методика устранения влияния времени на результат является достаточно эффективной, а модель является достаточно адекватной.

Также все вышеперечисленное не отменяет того, что остатки модели должны быть стационарными (рисунок 20).

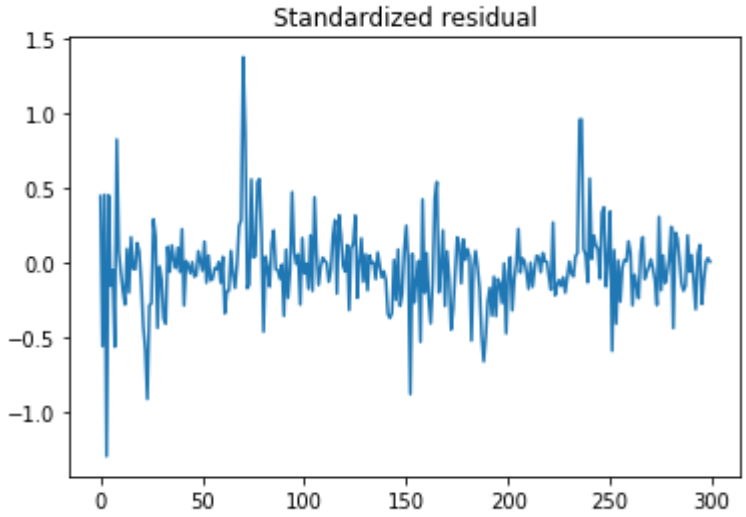

Рисунок 20 – Стандартизированные остатки модели многофакторной регрессии

По графику сложно сказать являются ли остатки стационарными, поэтому логично применить уже использованный ранее ADF-тест. В результате теста было получено значение критерия Дики-Фулера = -15.71. P-значение этого критерия = 0.000, что говорит о том, что гипотеза о стационарности ряда подтверждается.

p-значение - это вероятность наблюдения t-статистики, которая больше или больше по величине при нулевой гипотезе о том, что истинное значение коэффициента равно нулю. Если значение p больше 0,05 - что происходит примерно тогда, когда t-статистика меньше 2 в абсолютном значении - это означает, что коэффициент может быть только «случайно» значимым [27].

Следующим этапом проверки модели на адекватность служит построение коррелограммы остатков (рисунок 21).

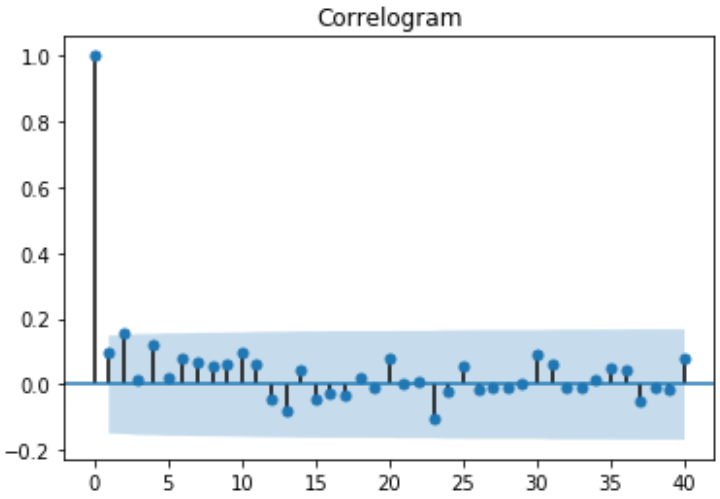

Рисунок 21 – Коррелограмма остатков модели многофакторной регрессии

В ходе визуального анализа графика можно сделать вывод, что практически все значения, представленные на коррелограмме незначимы с вероятностью 95%, что свидетельствует о том, что остатки являются «белым шумом», т.е. их выборочные автокорреляции значимо не отклоняться от нуля, а модель, построенная с такими показателями оценки можно назвать адекватной и применять в дальнейшем.

Графики стандартизированных остатков построенного дерева решений показывает аналогичную с многофакторными регрессионными моделями картину (рисунок 22 и 23).

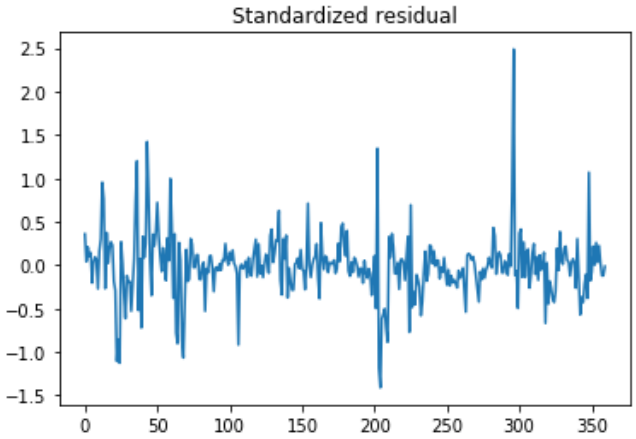

Рисунок 22 – Стандартизированные остатки дерева решений

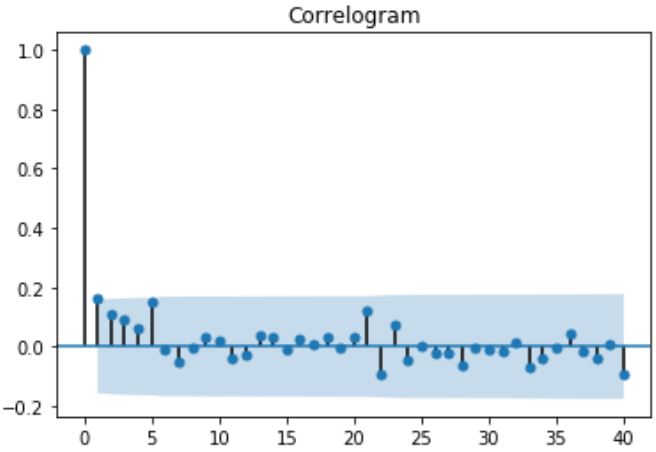

Рисунок 23 – Коррелограмма остатков дерева решений

Значение теста Дики-Фуллера, применённого к стандартизированным остаткам равно -14.7111, при критическом значении критерия для подтверждения гипотезы о стационарности с вероятностью 99% равно -3.436, при таких показателях, P-значение равно 0.000, что с более чем 99% вероятностью, данный временной ряд является стационарным.

В ходе визуального анализа коррелограммы также можно сделать вывод о незначимости практически всех значений с вероятностью 95%, и, как следствие, о том, что остатки являются «белым шумом», а модель в общем случае можно назвать адекватной и применять в дальнейшем.

Следует добавить, что проверка адекватности модели многофакторной регрессии и дерева решений таким образом является лишь условной, так как тема моделирования временных рядов таким образом является предметом обсуждения во многих работах и в методике построения также существуют различные разногласия, которые применимы также и к оценке адекватности таких моделей в целом.

Выводы по главе четыре

1. После получения результатов оценки качества прогнозов представленных в работе моделей, можно сказать, что лучшей по качеству текущего прогноза (на одно значение вперед) и долгосрочного прогноза (36 значений вперед) является

модель ARMA(5,5), текущий прогноз которой отклоняется от реальных показателей интенсивности интернет-трафика на 4,41%, а долгосрочный на 13,46%.

2. Модель ARIMA(2,1,3) показала схожую точность с моделью ARMA(5,5), в случае с краткосрочным прогнозом (6 значений вперед) модель ARIMA(2,1,3) показала даже большую точность, однако в общем случае и точность прогноза, и адекватность модели уступают аналогичным показателям модели ARMA(5,5).

3. Следующим по качеству прогнозирования ряда динамики интенсивности интернет-трафика является метод, основанный на деревьях решений. Самые худшие показатели качества остаются за моделью многофакторной регрессии. Характеристика полученных остатков, позволяет говорить о том, что по критериям проверки адекватности, построенное дерево решений и многофакторная регрессия, в данном случае можно назвать адекватными и применять в дальнейшем.

4. Учитывая относительно небольшую разницу в точности прогноза модели ARMA(5,5) и дерева решений при возникновении ситуации, когда требуется быстро спрогнозировать значение интенсивности интернет-трафика, возможно использование решающих деревьев, так как одно из главных их достоинств – скорость обучения.

ЗАКЛЮЧЕНИЕ

В настоящей выпускной квалификационной работе было произведён сравнительный анализ нескольких подходов к задаче прогнозирования временных рядов на примере анализа интенсивности интернет-трафика.

Данные в работе представлены интенсивностью сетевого трафика (в битах) от частного интернет-провайдера с центрами в 11 европейских городах. Данные были собраны с 6:57 часов 7 июня до 11:17 часов 31 июля 2015 года. Наблюдения проводились с интервалом в пять минут.

В выпускной квалификационной с помощью методов анализа временных рядов проводилось исследование, целью которого является проверка эффективности рассматриваемых методов анализа временных рядов для прогнозирования объема использованного трафика в определенный момент времени. Актуальность данного исследования обусловлена возможностью расчета допустимых отклонений для прогнозных значений интернет-трафика. Если реальное значение трафика вышло за доверительный интервал, можно говорить о возникновении аномалии, если точность прогноза является достаточной.

На основании информации из публикаций в данной области, можно сказать, что в настоящее время все еще существует потребность в прогнозировании сетевого трафика. Следует также учитывать, что структура трафика, как и его интенсивность за прошедшее время существенно изменилась, что необходимо учитывать при построении моделей прогнозирования.

Для прогнозирования интенсивности интернет-трафика использовались ARIMA-модели, имеющие подробное математико-статистическое обоснование, а также гибкость и универсальность в работе с временными рядами, модели данного вида являются одним из основных методов при работе с временными рядами, однако требовательность как к вычислительным ресурсам, так и к объему выборки, вкупе с дополнительной работой по периодической переоценке модели при

получении новых данных не позволяют говорить об незаменимости такого подхода к анализу временных рядов.

В противоположность ARIMA-моделям в работе решено было использовать модели множественной регрессии и решающие деревья, которые в области анализа временных рядов являются достаточно спорными методами с недостаточно хорошо описанной базой. Данные модели более требовательны к параметрам анализируемых рядов динамики: необходимо приведение данных к сопоставимому виду с точки зрения автокорреляции, коллинеарности и временного лага. С другой стороны, при подходящих условиях данные модели показывают себя как менее требовательные к вычислительным ресурсам и времени альтернативы ARIMA-моделей.

После получения результатов оценки качества прогнозов представленных в работе моделей, можно сказать, что лучшей по качеству текущего прогноза в целом, является модель ARMA(5,5), текущий прогноз которой отклоняется от реальных показателей интенсивности интернет-трафика на 4,41%, а долгосрочный на 13,46%.

Следующим по качеству прогнозирования ряда динамики интенсивности интернет-трафика является метод, основанный на деревьях решений. Самые худшие показатели качества остаются за моделью многофакторной регрессии. Характеристика полученных остатков, позволяет говорить о том, что по критериям проверки адекватности, построенное дерево решений и многофакторная регрессия, в данном случае можно назвать адекватными и применять в дальнейшем.

Дальнейшие разработки по данной теме рекомендовано вести в направлении изучения рядов динамики с более коротким интервалом времени, так как структура такого временного ряда будет существенно отличатся от используемого в работе. Возможно использование других методов, основанных в том числе и на деревьях решений, например, применение случайного леса, а также, оправдано использование нейронных сетей. В добавок, к моделям, работающим с многомер-

ными временными рядами (в данной работе такие модели были представлены многомерной регрессией и деревьями решений) возможно включение дополнительных факторов, отвечающих за дневную сезонность, погодные условия для конкретной географической зоны и.т.д.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1 Афанасьев, В.Н. Анализ временных рядов и прогнози-рование: Учебник / В.Н. Афанасьев, М.М. Юзбашев. – М.: Финансы и статистика, 2001. – 228 с.

2 Рукас, К.М. Предложения по выбору метода прогнозирования сетевого трафика / К.М. Рукас, С.Н. Теплицкая, К.А. Овчинников, А. А. Горюнов // Электронное научное специализированное издание - журнал «Проблемы телекоммуникаций». – 2014. – № 13. – С. 84 – 99.

3 Андерсен, Т. Статистический анализ временных рядов: учебник / Т. Андерсен; пер. с англ. И.Г. Жубенко, В.П. Носко. – М.: Мир, 1976. – 754 с.

4 Покровская, М.А. Сравнительный анализ методов прогнозирования мультимедийных приложений / М.А. Покровская // T-Comm. – 2013. – № 7. – С. 97–98.

5 Дорт-Гольц, А.А. Разработка и исследование метода балансировки трафика в пакетных сетях связи.: дис. канд. техн. наук / А.А. Дорт-Гольц. – СПб., 2014. – 166 с.

6 Jose, S. Cisco visual networking index: Forecast and methodology / S.Jose// Cisco systems – www.cisco.com.

7 Фрактальные процессы в телекоммуникациях: Монография / О.И. Шелухин, А.М. Тенякшев, А.В. Осин;под ред. О.И. Шелухина. – М.: Радиотехника, 2003. – 480 с.

8 Leland, W.E. On the self-similar nature of ethernet traffic / W.E. Leland, M.S. Taqqu, W. Willinger, D.V. Wilson // ACM Transactions of Networking. – 1994. – Vol. 2, № 1. – P. 1 – 15.

9 Shu, Y. Traffic prediction using FARIMA models / Y. Shu // IEEE International Conference. – 1999. – P. 891 – 895.

10 Гребенников, А. Моделирование сетевого трафика и прогнозирование с помощью модели ARIMA / А. Гребенников, Ю. Крюков, Д. Чернягин // Системный анализ в науке и образовании. – 2011. – №1. – С. 7–17.

11 Запорожец, Д.Б. Сравнительный анализ методов краткосрочного прогнозирования сетевого трафика / Д.Б. Запорожец, М.А. Скулиш // Збірник матеріалів Міжнародної науково-технічної конференції «Перспективи телекомунікацій» (проблеми телекомунікацій). – 2018. – № 12. – С 10 – 14.

12 Садовникова, Н.А. Анализ временных рядов и прогнозирование. Вып. 3: Учебно-методический комлекс / Н.А. Садовникова, Р.А. Шмойлова. – М.: Изд-во центр ЕАОИ, 2009. – 264 с.

13 Дуброва Т.А. Статистические методы прогнозирования в экономике: учебное пособие / Дуброва Т.А. - М.: Московская финансово-промышленная академия, 2004. – 60 с.

14 Садовникова, Н.А. Анализ временных рядов и прогнозирование. Учебное пособие. / Н.А. Садовникова, Р.А. Шмойлова. – М.: Московский государственный университет экономики, статистики и информатики, 2001. – 67 с.

15 Box, G.E.P. Time Series Analysis: Forecasting and Control / G.E.P. Box, G.M. Jenkins, G.C. Reinsel. – 4th edition. – Prentice Hall, 2008. – 810 p.

16 Олифер, В.Г. Компьютерные сети. Принципы, технологии, протоколы: Учебник для вузов / В.Г Олифер, Н.А. Олифер. – 3-е изд. – СПб.: Питер, 2006. – 246 с.

17 Лысяк, А.С. Методы прогнозирования временных рядов с большим алфавитом / А.С.Лысяк, Б.Я.Рябко// Вычисл. технологии. – 2014. – Вып. 19. – № 2. – С. 76–93.

18 Kane, M.J. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks./ M.J. Kane, P. Natalie, S. Matthew, R. Peter // BMC Bioinformatics. – 2014. – №15. – P. 270 – 276.

19 Фрактальные процессы в телекоммуникациях: Монография /О.И. Шелухин, А.М. Тенякшев, А.В. Осин; под ред. О.И. Шелухина. – М.: Радиотехника, 2003. – 480 с.

20 Рукас, К.М. Сравнительный анализ методов прогнозирования трафика в телекоммуникационных системах / К.М. Рукас, Ю.В. Соляник, К.А. Овчинников, О.Д. Олуватосин // Электронное научное специализированное издание - журнал «Проблемы телекоммуникаций». – 2014. – № 1 (13). – С. 95 – 99.

21 Фёрстер, Э. Методы корреляционного и регрессионного анализа. Руководство для экономистов. / Э. Фёрстер, Б. Рёнц; пер с нем. В. М. Ивановой. – М.: "Финансы и статистика", 1983. – 304 с.

22 Крюков, Ю.А. ARIMA – модель прогнозирования значений трафика / Ю.А. Крюков, Д.В. Чернягин // Информационные технологии и вычислительные системы. – 2011. – Вып. 2. – С. 41–49.

23 Шахиди, А. Деревья решений — общие принципы работы / А. Шахиди // BaseGroup Labs. – https://basegroup.ru/community/articles/description.

24 Волосенков, А. В. Применение деревьев решений в задачах прогнозирования многомерных временных рядов / А.В. Волосенков // Электронный математический и медико-биологический журнал. – http://www.sci.rostelecom67.ru/user/sgma/MMORPH/N-30-html/volosenkov/ volosenkov.htm.

25 Вентцель, Е.С. Теория вероятностей: Учеб. для вузов. / Е.С. Вентцель. – 6-е изд. стер. – М.: Высшая школа, 1999. – 576 с.

26 Демьяненко, Т.С. Совершенствование управления затратами на электрическую энергию промышленного предприятия по критерию энергоэффективности.: дис. … канд. экон. наук / Т.С. Демьяненко. – Челябинск., 2018. – 144 с.

27 Сергеев, Д. Анализ временных рядов с помощью Python. – https://habr.com/ru/company/ods/blog/327242/.

ПРИЛОЖЕНИЯ

ПРИЛОЖЕНИЕ А

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное
учреждение высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Институт естественных и точных наук
Кафедра математического и компьютерного моделирования

СРАВНИТЕЛЬНЫЙ АНАЛИЗ НЕСКОЛЬКИХ ПОДХОДОВ К РЕШЕНИЮ ЗАДАЧИ ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ НА ПРИМЕРЕ АНАЛИЗА ИНТЕНСИВНОСТИ ИНТЕРНЕТ-ТРАФИКА

ТЕХНИЧЕСКОЕ ЗАДАНИЕ
К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ
ЮУрГУ– 01.03.02.2019.009.ВКР

Руководитель работы,
старший преподаватель каф. МиКМ,
_____/ М.С. Фокина
«____»_____2019 г.

Автор работы
студент группы ЕТ-416
_____/ А.А. Чайко
«____»_____2019 г.

Нормоконтролер,
доцент каф. МиКМ,
канд. физ.-мат. наук
_____/ Т.А. Макаровских
«____»_____2019 г.

Челябинск 2019

ВВЕДЕНИЕ

1.1. Наименование программного изделия

Полное наименование программы – «MatModel».

1.2. Область применения

Построение моделей прогнозирования временных рядов (ARIMA, множественная регрессия, деревья решений) в области анализа интенсивности интернет-трафика.

2. Основание для разработки

2.1. Документ, на основании которого ведется разработка

Разработка ведется на основании задания выпускной квалификационной работы.

2.2. Организация, утвердившая этот документ

Задание на выпускную квалификационную работу утверждено руководителем работы, старшим преподавателем М.С. Фокиной.

2.3. Наименование темы разработки

Наименование темы разработки – Реализация математических моделей анализа временных рядов на языке программирования Python.

3. Назначение разработки

Разработка является частью выпускной квалификационной работы.

4. Требования к программе

4.1. Требования к функциональным характеристикам

4.1.1. Состав выполняемых функций

4.1.1.1. Программа должна построить все рассмотренные в выпускной квалификационной работе методы и модели.

4.1.1.2. Программа должна принимать в качестве входных данных файл формата .csv или .xlsx, содержащий два столбца, соответствующих представлению временного ряда (столбец с временем, столбец с значениями признака).

4.1.1.3. Программа должна вызывать используемые методы анализа временных рядов для построения моделей, получения прогнозных значений и оценки точности данных методов.

4.1.1.4. Программа должна выводить значения ошибки MAPE и MSE, графики исходного временного ряда и модельных значений, коррелограммы и стандартизированные ошибки модельных остатков.

4.1.2. Организация входных и выходных данных

Организация входных и выходных данных должна соответствовать ПРИЛОЖЕНИЮ Б, ПРИЛОЖЕНИЮ В. Входной информацией для программы должны являться данные, считываемые из csv таблиц. Выходная информация представляет собой графики исходных данных и модельных значений в виде png изображений, коэффициенты статистических моделей, прогнозные значения интернет-трафика в виде списка, содержащего данные значения.

4.2. Требования к надежности

4.2.1. Требования к надежному функционированию

Программа должна нормально функционировать при бесперебойной работе ЭВМ. При возникновении сбоя в работе аппаратуры восстановление нормальной работы программы должно производиться после:

1) перезагрузки операционной системы;
2) запуска исполняемого файла программы;

Уровень надежности программы должен соответствовать технологии программирования, предусматривающей:

1) инспекцию исходных текстов программы;
2) автономное тестирование модулей программы;
3) тестирование сопряжений модулей программы;
4) комплексное тестирование программы.

4.2.2. Время восстановления после отказа

Время восстановления после отказа должно состоять из:

1) времени запуска пользователем исполняемого файла программы
2) времени повторной загрузки исходных данных.

4.3. Условия эксплуатации

Программа должна храниться в виде двух маркированных копий: эталонной и рабочей. Периодическая перезапись информации должна осуществляться согласно нанесённой маркировке.

4.4. Требования к составу и параметрам технических средств

Программа должна корректно работать на следующем или совместном с ним оборудовании:

ЭВМ с операционными системами Windows 7, 8.1, 10.

4.5. Требования к информационной и программной совместимости

4.5.1. Требования к информационным структурам на входе и выходе

Требования к информационным структурам на входе и выходе определены в п. 4.1.2.

4.5.2. Требования к методам решения

Метод решения задач заранее определен разработчиком.

4.5.3. Требования к языкам программирования

Используется исключительно язык программирования Python.

4.5.4. Требования к программным средствам, используемым программой

Для корректной работы программы по построению моделей анализа временных рядов необходимы: дистрибутив языков программирования python – Anaconda Navigator версии 1.9.6 со средой разработки Jupyter Notebook версии 5.9.4, с языком программирования Python версии 3.6.1, модули: Pandas, NumPy, Scipy, Matplotlib, Plotly, Statmodels, Sklearn.

4.6. Требования к маркировке и упаковке

Диски с рабочими экземплярами программы должны иметь маркировку, состоящую из надписи: «Выпускная квалификационная работа. Чайко А.А. ЕТ-416, 2019», надписи «рабочий экземпляр», даты последней перезаписи программы.

Упаковка должна соответствовать условиям хранения диска. На упаковке должны быть указаны условия транспортирования и хранения диска.

4.7 Требования к транспортированию и хранению.

Условия транспортировки и хранения должны соответствовать п. 4.6.

5. Требования к программной документации

5.1. Документация к программе построения моделей анализа временных рядов на языке python должна содержать следующую информацию.

5.1.1. Технический проект программы по ГОСТ 19.404-79 в машинописном исполнении.

5.1.2. Описание программы по ГОСТ 19.402-78 на компакт-диске.

5.1.3. Текст программы по ГОСТ 19.401-78 на компакт-диске.

5.2. Пояснительная записка «Технический проект программы» должна содержать следующие разделы.

5.2.1. Раздел «Входные данные» (характер, организация входных данных).

5.2.2. Раздел «Выходные данные» (характер и организация выходных данных).

5.2.3. Перечень модулей программы и их характеристика (таблица с перечнем наименований модулей с указанием выполняемой каждым модулем функции);

6. Технико-экономические показатели

Технико-экономические показатели должны определяться заказчиком без участия исполнителя.

7. Стадии и этапы разработки

Разработка программы должна выполняться по следующим этапам:

1) разработка, согласование и утверждение технического проекта программы с пояснительной запиской – 3 недели;

2) разработка рабочего проекта программы с комплексным тестированием – 7 недель;

3) приемка-сдача с исправлением обнаруженных недостатков в программе и программной документации – 1.5 недели;

8. Порядок контроля и приемки

8.1. Виды испытаний

Испытания программы и верификация документации должны производиться в организации заказчика с привлечением сторонних экспертов. Проверочные тесты должны готовиться заказчиком.

8.2. Общие требования к приемке

Приемка программы осуществляется заказчиком. Программа должна считаться годной, если она удовлетворяет всем пунктам технического задания.

ПРИЛОЖЕНИЕ Б

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное
учреждение высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Институт естественных и точных наук
Кафедра математического и компьютерного моделирования

СРАВНИТЕЛЬНЫЙ АНАЛИЗ НЕСКОЛЬКИХ ПОДХОДОВ К РЕШЕНИЮ ЗАДАЧИ ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ НА ПРИМЕРЕ АНАЛИЗА ИНТЕНСИВНОСТИ ИНТЕРНЕТ-ТРАФИКА

РУКОВОДСТВО ПРОГРАММИСТА
К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ
ЮУрГУ– 01.03.02.2019.009.ВКР

Руководитель работы,
старший преподаватель каф. МиКМ,
_____/ М.С. Фокина
«____»_____2019 г.

Автор работы
студент группы ЕТ-416
_____/ А.А. Чайко
«____»_____2019 г.

Нормоконтролер,
доцент каф. МиКМ,
канд. физ.-мат.наук
_____/ Т.А. Макаровских
«____»_____2019 г.

Челябинск 2019

1. Назначение и условия применения программы

1.1 Назначение программы

Программа выполняет построение моделей прогнозирования временных рядов (ARIMA, множественная регрессия, деревья решений) в области анализа интенсивности интернет-трафика.

1.2 Требования к техническим средствам и программному обеспечению

Для работы программы необходима ЭВМ, характеристики которой поддерживают операционные системы: Windows XP, Windows Vista, Windows 7. Требования к программному обеспечению сводятся к одной из операционных систем Windows XP, Windows Vista, Windows 7, а также дистрибутива языков программирования python – Anaconda Navigator версии 1.9.6 со средой разработки Jupyter Notebook версии 5.9.4, с языком программирования Python версии 3.6.1 и python-библиотеками: Pandas, NumPy, Scipy, Matplotlib, Plotly, Statmodels, Sklearn.

2. Обращение к программе

Для работы программы необходимо в строке считывания таблицы указать путь к файлу с расширением .xlsx, после чего вся информация из хранится в памяти и может быть использована для построения математических моделей.

В функции построения коррелограммы, кроме указания объекта, содержащего данные временного ряда типа *dataframe* необходимо указать количество лагов, аргумент *lags.*

В функции построения графика аргументы:

– *n_pred1* – количество прогнозных значений;

– *sdvig* – количество значений временного ряда (с конца), которое необходимо отобразить;

– *data* – данные временного ряда в формате *dataframe.*

В функции, реализующей создание многомерного временного ряда с признаками (лаги, выходные) для построения модели множественной регрессии и дерева решений аргументы:

– *data* - данные временного ряда в формате *dataframe.*

– *lag_start* – начальный лаг;

– *lag_end* – конечный лаг;

– *test_size* – размер тестового отрезка (в данной работе равен количеству предсказываемых значений).

3. Входные и выходные данные

Программа должна принимать в качестве входных данных файл формата .csv или .xlsx, содержащий два столбца, соответствующих представлению временного ряда.

Заголовки столбцов соответствуют времени и значению признака в этот момент времени. Таблица заносится в объект *dataset* типа *dataframe* с помощью команды *pd.read_excel*, где *pd* – сокращенное название импортированной библиотеки Pandas.

Выходная информация представляет собой графики исходных данных и модельных значений в виде png изображений, коэффициенты статистических моделей, прогнозные значения интернет-трафика в виде списка, содержащего данные значения.

ПРИЛОЖЕНИЕ В

КОД НА ЯЗЫКЕ PYTHON, ИСПОЛЬЗОВАННЫЙ В РАБОТЕ

```
#Подключение необходимых библиотек
import sys
import warnings
warnings.filterwarnings('ignore')
from tqdm import tqdm

import pandas as pd
import numpy as np
from sklearn.metrics import mean_absolute_error, mean_squared_error

import statsmodels.formula.api as smf
import statsmodels.tsa.api as smt
import statsmodels.api as sm
import scipy.stats as scs
from scipy.optimize import minimize
from arch import arch_model

import matplotlib.pyplot as plt
import matplotlib as mpl
import pylab
%matplotlib inline

import plotly.plotly as py
from plotly import __version__
from plotly.offline import download_plotlyjs, init_notebook_mode, plot, iplot
from plotly import graph_objs as go
init_notebook_mode(connected = True)

import cufflinks as cf
import cufflinks

# Функция расчета MAPE
def mean_absolute_percentage_error(y_true, y_pred):
    y_true, y_pred = np.array(y_true), np.array(y_pred)
    return np.mean(np.abs((y_true - y_pred) / y_true)) * 100

# Считывание информации из Excel-файла и вывод информации о полученном датафрейме
dataset = pd.read_excel('ITDIB_1.xlsx', index_col=0)
dataset.info()

# Функция построения коррелограммы
def plot_autocorr(returns, lags):
    autocorrelation = []
    for lag in range(lags+1):
        corr_lag = returns.corr(returns.shift(-lag))
        autocorrelation.append(corr_lag)
    return np.array(autocorrelation)

x=plot_autocorr(dataset,24*7*12)
columns = {'Count':x}
x_df = pd.DataFrame(columns)
```

```
# Установка минимальных и максимальных значений параметров ARIMA
# для нахождения модели с лучшими параметрами путем перебора
ps = range(0, 6)
d= range(0, 2)
qs = range(0, 6)

from itertools import product

parameters = product(ps,d,qs)
parameters_list = list(parameters)
len(parameters_list)

# Нахождение лучшей модели в пределах заданных параметров с помощью критерия AIC
%%time
results = []
best_aic = float("inf")
warnings.filterwarnings('ignore')

for param in parameters_list:
    #try except необходим, так как на некоторых наборах параметров модель не обучается
    try:
        model=sm.tsa.statespace.SARIMAX(dataset, order=(param[0], param[1],
param[2])).fit(disp=-1)
    #вывод параметров, на которых модель не обучается и переход к следующему набору
    except ValueError:
        print('wrong parameters:', param)
        continue
    aic = model.aic
    #сохранить лучшую модель, aic, параметры
    if aic < best_aic:
        best_model = model
        best_aic = aic
        best_param = param
    results.append([param, model.aic])

warnings.filterwarnings('default')

result_table = pd.DataFrame(results)
result_table.columns = ['parameters', 'aic']

print(result_table.sort_values(by = 'aic', ascending=True).head(5))

# Обучение лучшей модели и вывод ее характеристик
%%time
best_model=sm.tsa.statespace.SARIMAX(dataset, order=(5, 0, 5)).fit(disp=-1)
print(best_model.summary())
best_model.plot_diagnostics(figsize=(15, 12))
plt.show()

# Функция получения прогнозных значений и построения графика прогноза
def PLOTIMA2(n_pred1,sdvig,data):
    data_day_df=data
    data_day_df["arima_model"] = best_model.fittedvalues
```

```
    forecast = best_model.predict(start = data_day_df.shape[0], end = da-
ta_day_df.shape[0]+n_pred1)
    forecast = data_day_df.arima_model.append(forecast).values[(-sdvig-n_pred1):]
    actual = data_day_df.Count.values[-sdvig:]
    mape=mean_absolute_percentage_error(data_day_df.dropna().Count, da-
ta_day_df.dropna().arima_model)
    mae=np.sqrt(mean_squared_error(data_day_df.dropna().Count, da-
ta_day_df.dropna().arima_model))
    plt.figure(figsize=(15, 7))
    plt.plot(forecast, color='r', label="Прогнозные значения")
    plt.plot(actual, label="Реальные значения")
    plt.legend()
    plt.ylabel("Интенсивность трафика (в млрд. бит)")
    plt.axvspan(len(actual), len(forecast), alpha=0.5, color='lightgrey')
    plt.grid(True)
    print(mape,mae)
PLOTIMA2(6,200,dataset)

# Функция, реализующая создание многомерного временного ряда с признаками (лаги, выход-
# ные) для построения модели множественной регрессии и дерева решений
def prepareData(data, lag_start, lag_end, test_size):

    data = pd.DataFrame(data.copy())
    data.columns = ["y"]

    # Расчет индекса в датафрейме, после которого начинается тестовый отрезок
    test_index = int(len(data-test_size))

    # Добавление лагов исходного ряда в качестве признаков
    for i in range(lag_start, lag_end):
        data["lag_{}".format(i)] = data.y.shift(i)

    data.index = pd.to_datetime(data.index,infer_datetime_format=True)
    data["hour"] = data.index.hour
    data["weekday"] = data.index.weekday
    data['is_weekend'] = data.weekday.isin([5,6])*1

    data = data.dropna()
    data = data.reset_index(drop=True)

    # разделение весь датасет на тренировочную и тестовую выборку
    X_train = data.loc[:test_index].drop(["y"], axis=1)
    y_train = data.loc[:test_index]["y"]
    X_test = data.loc[test_index:].drop(["y"], axis=1)
    y_test = data.loc[test_index:]["y"]

    return X_train, X_test, y_train, y_test

# Иморт библиотек, необходимых для регрессионной модели, построение данной модели с
заданными параметрами,
# полчение коэффициентов модели и оценка точности прогноза
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error
```

```
X_train, X_test, y_train, y_test = prepareData(data_day, test_size=36, lag_start=36,
lag_end=48)
lr = LinearRegression()
lr=lr.fit(X_train, y_train)
prediction =pd.DataFrame(data=lr.predict(X_test),index=None,
                  columns=["Count"])

lr.coef_
mape=mean_absolute_percentage_error(prediction.Count, y_test)
mae=np.sqrt(mean_squared_error(prediction.Count, y_test))
print(mape,mae)

# Иморт библиотек, необходимых для дерева решений, построение решающего дерева с задан-
ными параметрами,
# полчение коэффициентов модели и оценка точности прогноза
from sklearn.tree import DecisionTreeRegressor
from sklearn.tree import export_graphviz
tree = DecisionTreeRegressor(random_state=17)

from sklearn.model_selection import GridSearchCV, cross_val_score
tree_params = {'max_depth': range(1,12),
               'max_features': range(1,15),
               'min_samples_leaf':range(1,10)}
tree_grid = GridSearchCV(tree, tree_params,
                         cv=5, n_jobs=-1,
                         verbose=True)
%%time
tree_grid.fit(X_train, y_train)
tree_grid.best_params_

rtp=pd.DataFrame(data=tree_grid.predict(X_test),
                  columns=["Count"])
mape=mean_absolute_percentage_error(y_test,rtp.Count)
mae=np.sqrt(mean_squared_error(y_test,rtp.Count))
print(mape,mae)
```