

Факультет Кибернетики и информационной
безопасности

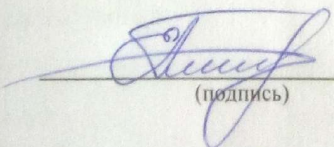
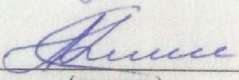
Кафедра кибернетики (№ 22)

Направление подготовки 09.03.04 Программная инженерия

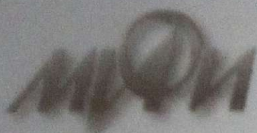
Пояснительная записка

к учебно-исследовательской работе студента на тему:

Разработка модуля определения и обработки факторов,
характеризующих спорадический спрос на денежные средства в сети
банкоматов.

Группа	Б14-506	
Студент	 (подпись)	Пономарёв Е.А. (ФИО)
Руководитель	 (подпись)	Немешаев С.А. (ФИО)
Научный консультант	 (подпись)	 (ФИО)
Оценка руководителя	12 (0-15 баллов)	Оценка консультанта (0-15 баллов)

Москва 2017



Задание на УИР

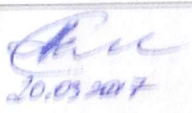

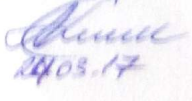
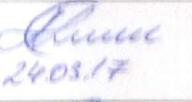

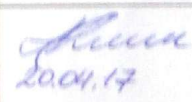

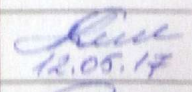
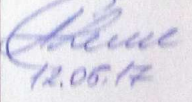
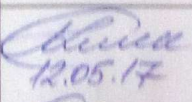
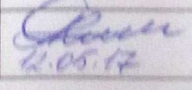
Студенту гр. Б14-506
(группа)

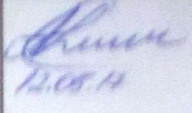
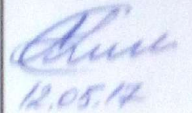
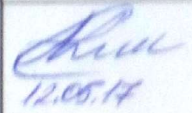
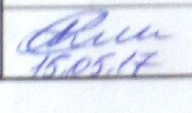
Пономарёву Евгению Александровичу
(ф.и.о.)

ТЕМА УИР

Разработка модуля определения и обработки факторов, характеризующих
 спорадический спрос на денежные средства в сети банкоматов.

ЗАДАНИЕ

№ п/п	Содержание работы	Форма отчетности	Срок исполне- ния	Отметка о выполнении Дата, подпись
1.	Аналитическая часть			
1.1.	Изучение теоретической основы поставленной задачи в рамках статистического анализа с целью определения аномального спроса денежных средств в сети банкоматов.	Аналитический отчет	20.03.2017	 20.03.2017
1.2.	Сравнение и выбор подходящих алгоритмов прогнозирования финансовых рядов, характеризующихся спорадическим типом спроса.	Аналитический отчет	20.03.2017	 20.03.2017
1.3.	Анализ статистических данных на наличие факторов, при которых возникает аномальный спрос денежных средств в сети банкоматов с целью их учета в алгоритме прогнозирования.	Аналитический отчет	24.03.2017	 24.03.17
1.4.	Анализ и оценка эффективности методов прогнозирования финансовых рядов уже существующих программных средств.	Аналитический отчет	24.03.2017	 24.03.17
1.5.	Оформление расширенного содержания пояснительной записки (РСПЗ)	Текст РСПЗ	27.03.2017	
2.	Теоретическая часть			
2.1.	Разработка подходящей модели прогнозирования финансовых рядов, чувствительной к факторам аномального спроса.	Алгоритм	20.04.2017	 20.04.17
2.2.	Модификация полученной модели для перестроения прогноза в результате учёта новых факторов.	Текст раздела ПЗ	05.05.2017	 05.05.17
3.	Инженерная часть			
3.1.	Проектирование модуля определения и обработки факторов аномального спроса денежных средств.	Диаграмма, описание	12.05.2017	 12.05.17
3.2.	Разработать архитектуру модуля, определить необходимые методы программной реализации, выбрать необходимые библиотеки для разработки.	Диаграмма, описание	12.05.2017	 12.05.17
3.3.	Результаты проектирования оформить с помощью графического представления на языке UML.	Диаграмма, описание	12.05.2017	 12.05.17
4.	Технологическая и практическая часть		12.05.2017	 12.05.17

4.1.	Реализовать модуль определения и обработки факторов аномального спроса в системе прогнозирования инкассаций.	Исполняемые файлы, исходный текст	12.05.2017	 12.05.17
4.2.	Провести тестирование разработанного модуля на реальных данных.	Исполняемые файлы, исходные тексты тестов и тестовых примеров	12.05.2017	 12.05.17
4.3.	Реализация должна принимать на вход статистические ряды снятий денежных средств в банкоматах, определять факторы, обуславливающие аномальные значения и корректировать итоговый прогноз.	Исходный код	12.05.2017	 12.05.17
5.	Оформление пояснительной записки (ПЗ) и иллюстративного материала для доклада.	Текст ПЗ, презентация	15.05.2017	 15.05.17

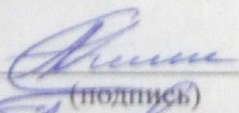
ЛИТЕРАТУРА

[1]	Суслов В.И., Ибрагимов Н.М., Талышева Л.П., Цыплаков А.А. Эконометрия. Часть III Эконометрия - I: Анализ временных рядов Учебное пособие Новосибирск: Изд-во СО РАН, 2005. 744 с
[2]	Ханк Д.Э., Уичерн Д.У., Райте А.Дж. Бизнес-прогнозирование – М.: Издательский дом "Вильямс", 2003. – 656 с.
[3]	Сигел Э. Практическая бизнес-статистика – М.: Издательский дом "Вильямс", 2008. – 1056 с.
[4]	M.D. Aseev, S.A. Nemshaev, A.P. Nesterov. Forecasting Cash Withdrawals In The ATM Network Using A Combined Model Based On The Holt-Winters Method And Markov Chains// International Journal of Applied Engineering Research (IJAER). — 2016. —Т. 11. — №. 11. — С. 7573-7578
[5]	А. А. Марков, О представлении рекурсивных функций. М.: Изв. АН СССР. Сер. матем., 1949, том 13, выпуск 5, 417– 424.
[6]	Алан Купер. Алан Купер об интерфейсе. Основы проектирования взаимодействия/ А. Купер, Р. Реймани, Д. Кронин – М.: Издательство «Символ-плюс», 2009, – 688 с.

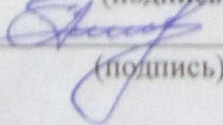
Дата выдачи задания:

12.05 2017г.

Руководитель

 (подпись) (Немешаев С.А.) (фио)

Студент

 (подпись) (Пономарёв Е.А.) (фио)

Уважаемый пользователь! Обращаем ваше внимание, что система «Антиплагиат» отвечает на вопрос, является ли тот или иной фрагмент текста заимствованным или нет. Ответ на вопрос, является ли заимствованный фрагмент именно плагиатом, а не законной цитатой, система оставляет на ваше усмотрение.

Отчет о проверке № 1

дата выгрузки: 22.06.2017 18:15:37
пользователь: ponomaryov_ea@mail.ru / ID: 4803991
отчет предоставлен сервисом «Антиплагиат»
на сайте <http://www.antiplagiat.ru>

Информация о документе

№ документа: 2
Имя исходного файла: PZ_Ponomaryov-vesna_2017_3_1_.docx
Размер текста: 744 кБ
Тип документа: Не указано
Символов в тексте: 63203
Слов в тексте: 7749
Число предложений: 403

Информация об отчете

Дата: Отчет от 22.06.2017 18:15:37 - Последний готовый отчет
Комментарии: не указано
Оценка оригинальности: 88.99%
Заимствования: 11.01%
Цитирование: 0%



Оригинальность: 88.99%
Заимствования: 11.01%
Цитирование: 0%

Источники

Доля в тексте	Источник	Ссылка	Дата	Найдено в
3.07%	[1] ОБРАБОТКА АНОМАЛЬНЫХ ЗНАЧЕНИЙ УРОВНЕЙ ВРЕМЕННОГО РЯДА КАК ЭТАП КОМПЛЕКСНОЙ ОЦЕНКИ ИНФОРМАЦИИ В ПОДСИСТЕМЕ ПРОГНОЗИРОВАНИЯ ДЛЯ СИТУАЦИОННОГО ЦЕНТРА	http://cyberleninka.ru	08.10.2015	Модуль поиска Интернет
2.31%	[2] Учебное пособие по дисциплине "Эконометрика" для магистрантов 1-го и 2-го года обучения (под ред. проф. Горбаткова С.А.)	http://website.vzfei.ru	раньше 2011 года	Модуль поиска Интернет
2.17%	[3] скачать	http://vball5.ru	30.11.2016	Модуль поиска Интернет

Реферат

Пояснительная записка содержит 44 страницы, 10 рисунков, 3 таблицы, 20 ссылок на источники.

Ключевые слова: модель прогнозирования, анализ временных рядов, прогнозирование инкассаций, аномальный спрос на денежные средства, аномальный уровень.

Целью данной учебно-исследовательской работы является разработка модуля обнаружения и корректировки аномальных значений спроса денежных средств в сети банкоматов самообслуживания для системы прогнозирования инкассаций.

В первом разделе описывается изучение предметной области поставленной задачи в рамках статистического анализа. Рассмотрены виды аномальных значений и методы их обнаружения в статистических данных.

Во втором разделе описывается построение подходящей модели прогнозирования финансовых рядов, чувствительной к факторам аномального спроса помощью современных теоретических и программных средств.

В третьем разделе описывается проектирование модуля определения и обработки факторов аномального спроса на денежные средства в сети банкоматов.

В четвертом разделе приведено описание разработанного программного модуля выделения факторов аномального спроса на наличность в сети банкоматов, используемых библиотек программных средств, а также приведены результаты работы программного комплекса при обработке входных статистических данных.

Содержание

1. Анализ проблемы обнаружения аномального спроса в статистических данных	6
1.1 Классификация аномальных значений и методы их обработки	6
1.2. Анализ существующих методов определения аномального спроса	9
1.2.1. Метод Граббса	10
1.2.2. Метод Диксона (Dixon's Q test)	11
1.3. Анализ данной системы прогнозирования.	12
1.4. Анализ методов и средств прогнозирования уже существующих программных реализаций.	13
1.5. Выводы	17
1.6 Цели и задачи учебно-исследовательской работы	18
2.Разработка подходящей модели прогнозирования финансовых рядов, чувствительной к факторам аномального спроса.	20
2.1. Разработка метода определения аномального спроса на денежные средства	21
2.2. Предложенный алгоритм определения аномального спроса на денежные средства.	22
3. Проектирование системы	27
3.1 Требования к проектируемому модулю	27
3.2. Проектирование архитектуры разрабатываемого модуля	27
3.3 Технологии и инструментальные средства проектирования	28
3.4 Результаты разработки	29
4. Программная реализация модуля определения аномального спроса на наличность	32
4.1. Реализация модуля обнаружения и корректировки аномальных значений в системе прогнозирования инкассаций	33
4.2 Тестирование на реальных данных	35
Заключение	41
Литература	42

Введение

На сегодняшний день довольно актуальной проблемой является оптимизация процессов банковского самообслуживания. Количество банкоматов в России на начало 2017 года насчитывает 207 тыс. ед. [1]. Так как каждый банкомат нуждается в техническом обслуживании и постоянным контролем над доступными денежными средствами, поднимается вопрос об актуальности разработки системы прогнозирования инкассаций.

Подобная система способна вычислять сумму наличности, которую необходимо загрузить в банкомат, а также процентное соотношение видов загружаемых купюр. При этом система должна стараться по максимуму исключать факт пролёживания денежных средств в отдельном банкомате и выбирать необходимое время для производства инкассации.

Ключевым фактором в разработке подобной системы является обнаружение и обработка аномальных значений спроса на денежные средства в банкомате, которые могут быть обусловлены как случайными выбросами, так и некоторыми календарными эффектами (праздники, зарплатные периоды и т. д.). А значит, для построения качественного прогноза необходимо использовать алгоритмы и технологии, позволяющие учитывать наличие аномальных значений в статистических данных и в зависимости от их природы корректировать прогноз.

Задачей данного исследования является изучение и анализ данного вопроса, что в конечном итоге позволит спроектировать и построить подходящий модуль определения аномального спроса на денежные средства для данной системы прогнозирования инкассаций.

В первом разделе работы представлен анализ существующих подходов к решению проблемы определения факторов спорадического спроса при прогнозировании временных рядов. Рассмотрены некоторые методы определения аномальных значений в статистических данных, а также проведен сравнительный анализ данной системы прогнозирования с мировыми аналогами.

Во втором разделе описан процесс разработки подходящей модели прогнозирования, чувствительной к наличию аномальных значений спроса на денежные средства в сети банкоматов самообслуживания. Также представлен процесс модификации текущей системы прогнозирования инкассаций. Проведен анализ

эффективности полученной модели прогнозирования при помощи программных средств языка статистической обработки R.

В третьем разделе описывается инженерная часть учебно-исследовательской работы, а именно проектирование модуля определения и обработки факторов аномального спроса на наличность. Представлены требования к проектируемому модулю, описаны технические аспекты программной реализации построенного модуля.

В четвёртом разделе описывается программная реализация разработанного модуля определения аномального спроса на наличность. Описан принцип работы исходных текстов, а также используемых инструментальных средств разработки с последующим обоснованием их выбора. Представлены результаты тестирования разработанного модуля на реальных статистических данных.

1. Анализ проблемы обнаружения аномального спроса в статистических данных

В данном разделе приводятся теоретические аспекты природы аномальных значений, их подробная классификация и методы корректировки. На основе материала из учебно-технической литературы рассматриваются некоторые современные методы обнаружения аномальных уровней во временных рядах. В подразделе «Выводы» подведены итоги анализа предметной области. В конце раздела формулируется цель учебно-исследовательской работы, а также перечисляются задачи, которые необходимо выполнить для достижения поставленной цели.

Обнаружение аномалий - это проблема поиска шаблонов в данных, которые не соответствуют модели «нормального» поведения. Обнаружение таких отклонений от ожидаемого поведения временных данных важно для обеспечения нормальной работы систем в разных областях, таких как экономика, биология, вычислительная техника, финансы, экология и другие. Приложения в таких областях нуждаются в способности определять ненормальное поведение системы, которое может быть признаком отказа системы или злонамеренных действий. Такие системы должны быть способны инициировать соответствующие шаги для принятия корректирующих действий. Причем на каждом шаге важно охарактеризовать, что является нормальным, что является девиантным или аномальным, и насколько значительной является аномалия. Эта характеристика проста для систем, где поведение может быть задано с помощью простых математических моделей - например, выход гауссовского распределения с известным средним и стандартным отклонением. Тем не менее, большинство сложных систем прогнозирования имеют сложный принцип работы с временными рядами. Для таких систем становится задача необходимости характеризовать нормальное состояние системы, собирая данные о системе в течение этого периода времени, и использовать эту характеристику в качестве основы для обозначения аномального поведения[2-3].

1.1 Классификация аномальных значений и методы их обработки

В данном разделе приведена подробная классификация аномальных значений, описаны этапы их обработки. Представлены возможные способы корректировки найденных аномалий в исследуемом бизнес-процессе.

Согласно оценкам многих авторов [3-5] входящие данные могут содержать до 10-15% аномальных значений. Однако, даже при меньшем их количестве есть риск того, что такой выброс может существенно повлиять на характеристику анализируемого ряда.

Для того, чтобы охарактеризовать набор данных, аналитику необходимо провести две процедуры, а именно:

- Анализ и изучение общего вида графической интерпретации таких важных особенностей, как симметрия и отклонение от прогноза.
- Анализ аномальных значений, которые выделяются из общей совокупности.

Таким образом, задача обработки аномальных значений состоит из двух этапов.

Обнаружение. Производится поиск аномальных значений и, если возможно, проверяется, является аномалия искусственной или естественной. Если аномалия искусственная, то необходимо удалить соответствующую запись либо произвести корректировку значений одним из доступных методов. Если аномалия естественная (то есть отражает естественную изменчивость данных), то ее, возможно, следует оставить, но дальнейший анализ производить с соответствующими поправками.

Корректировка. Обнаруженные аномальные значения исключаются или корректируются в зависимости от логики и особенностей задачи анализа.

Особое внимание аналитик должен уделить корректировке найденных выбросов.

Если обнаруженное аномальное значение действительно искажает информацию об исследуемом бизнес-процессе и может повлиять на качество модели и достоверность результатов анализа, то необходимо выполнить его корректировку. Для этого в зависимости от логики и особенностей решаемой задачи можно использовать несколько способов.

- Удаление записи с аномальным значением. Если число записей в выборке данных существенно превышает минимум, требуемый для анализа, то записи, содержащие аномальные значения, можно просто удалить.
- Ручная замена аномальных значений. Применяется, если количество аномальных значений невелико и они могут быть обработаны вручную. При этом аналитик заменяет аномальные значения на другие, более соответствующие модели поведения данных.

- Сглаживание и фильтрация данных. Для обработки аномальных значений можно использовать методы частотной или пространственной фильтрации, применяемые при сглаживании данных и очистке от шумов(впрочем, это совершенно отдельный класс алгоритмов). При этом следует учитывать, что в результате обработки будут изменены не только аномальные значения, но и все значения ряда.
- Интерполяция данных. Аномальные значения заменяются другими, вычисленными на основе нескольких ближайших соседей.
- Замена на наиболее вероятное значение. Строится гистограмма распределения значений ряда, и по ней определяется значение, соответствующее моде гистограммы, которое и будет являться статистически наиболее вероятным.[4]

Если же говорить о природе аномальных значений, то все аномальные значения можно разделить на два класса:

искусственные — связанные с ошибками ввода данных, некорректной работой программа или систем регистрации данных(например, сканера штрихкода);

естественные — отражают факты и события, имевшие место в действительности, но вызванные исключительными обстоятельствами, которые встречаются очень редко или в единичных случаях.

Аномальным уровнем будем называть отдельное значение, которое не соответствует общей характеристике той или иной выборки, и являясь уровнем ряда, оказывает существенное влияние на значение общего результата, в том числе и на трендовую модель.

Нехарактерные уровни во временном ряду можно подразделить на три группы[5]:

- значения, отражающие объективное развитие процесса, но сильно отличающиеся от общей тенденции, так как они проявляют свои экстремальные воздействия крайне редко;
- значения, возникающие вследствие изменений методики расчета;
- значения, возникающие вследствие ошибок при измерении показателя, при записи и передаче информации, а также значения, связанные с различными катастрофическими явлениями, не влияющими на дальнейший ход развития явления, агрегировании и дезагрегировании показателей и т. д.

Аномальные значения из первой группы не всегда нуждаются в корректировке и исключении из временного ряда, а иногда даже представляют собой существенно важную информацию. Например, в процессе расследования мошенничества с банковскими счетами и кредитными картами, аномальные значения являются индикаторами мошеннической деятельности. И тогда эти самые выбросы будут являться не помехами, а целью анализа.

Неявные значения второй группы также не должны исключаться из общей совокупности, а скорее приниматься за «повторные» (или пороговые), начиная с которых необходимо пересчитывать все предыдущие значения временного ряда по новой методике.

Значения, принадлежащие третьей группе, должны быть исключены из рассмотрения в любом случае, так как они искажают общее представление о характере развития явления и несомненно оказывают существенное влияние на выводы, полученные в результате анализа ряда, содержащего искаженную информацию подобного типа.[5]

1.2. Анализ существующих методов определения аномального спроса

В данном разделе представлены некоторые известные методы обнаружения аномальных значений в статистических данных, в общем виде описаны их алгоритмы. Приводятся формулы для расчёта критических значений.

Существуют стандартные критерии определения выбросов в выборке при заданном уровне значимости/доверия. Примеры таких критериев: критерий Шовене, тест Граббса, критерий Пирса, Q-тест Диксона.

Упомянутые критерии (исключение: критерий Граббса) выстраивают выборку по возрастанию и проверяют крайние значения (min, max элемент выборки) на выброс, подключается таблица критических значений. Значения в таблицах зависят от количества элементов в выборке и уровня доверия/значимости. Критерии позволяют определить точно один выброс, в случае, когда их много, критерии могут не работать.

Выбор того или иного метода выявления и анализа аномальных наблюдений определяется объемом совокупности, характером исследуемых процессов и задач (одномерные и многомерные).[6]

1.2.1. Метод Граббса

В данном разделе рассматривается метод определения аномальных значений в выборке с помощью критерия Граббса. Приводится алгоритм и соответствующие формулы для вычисления.

Критерий Граббса проверки на один выброс. Пусть X_1, X_2, \dots, X_n – наблюдаемая выборка, $X_{(1)} \leq X_{(2)} \dots \leq X_{(n)}$ – построенный по ней вариационный ряд. Проверяемая гипотеза H_0 заключается в том, что все X_1, X_2, \dots, X_n принадлежат одной генеральной совокупности. При проверке на выброс наибольшего выборочного значения конкурирующая гипотеза H_1 заключается в том, что X_1, X_2, \dots, X_{n-1} принадлежат одному закону, а $X_{(n)}$ – некоторому другому, существенно сдвинутому вправо. При проверке на выброс $X_{(n)}$ статистика критерия Граббса имеет вид

$$G_n = (X_{(n)} - \bar{X})/S, \quad (1)$$

$$\text{где } \bar{X} = \frac{1}{n} \sum_{j=1}^n X_j \quad (2)$$

$$S^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2, \quad (3)$$

$$S = \sqrt{S^2}. \quad (4)$$

При проверке на выброс наименьшего выборочного значения конкурирующая гипотеза H_1 предполагает, что $X_{(1)}$ принадлежит некоторому другому закону, существенно сдвинутому влево. В данном случае вычисляемая статистика принимает вид

$$G_1 = (\bar{X} - X_{(1)})/S. \quad (5)$$

Максимальный или минимальный элемент выборки считается выбросом, если значение соответствующей статистики превысит критическое: $G_n \geq G_{n,1-\alpha}$ или $G_1 \geq G_{n,1-\alpha}$, где α – задаваемый уровень значимости.

Статистики (1) и (5) распределены одинаково. Вид условных распределений

$F(G_i \vee H_0)$ статистик (1) и (5) в зависимости от объема анализируемой выборки при нормальном законе наблюдаемых величин представлен на рис. 1. Распределения статистики существенно зависят от объема выборки n . Аналитический вид распределений статистики в стандарте [5] и первоисточниках [7-8] не приводится. Даются лишь верхние процентные точки для различных объемов выборок, так как

решение об аномальности проверяемого минимального или максимального выборочного значения принимается по правому “хвосту” распределения статистики. Если в стандарте процентные точки приведены для объемов выборок n лишь от 3 до 40, то в [4] процентные точки приведены в диапазоне n до 147.

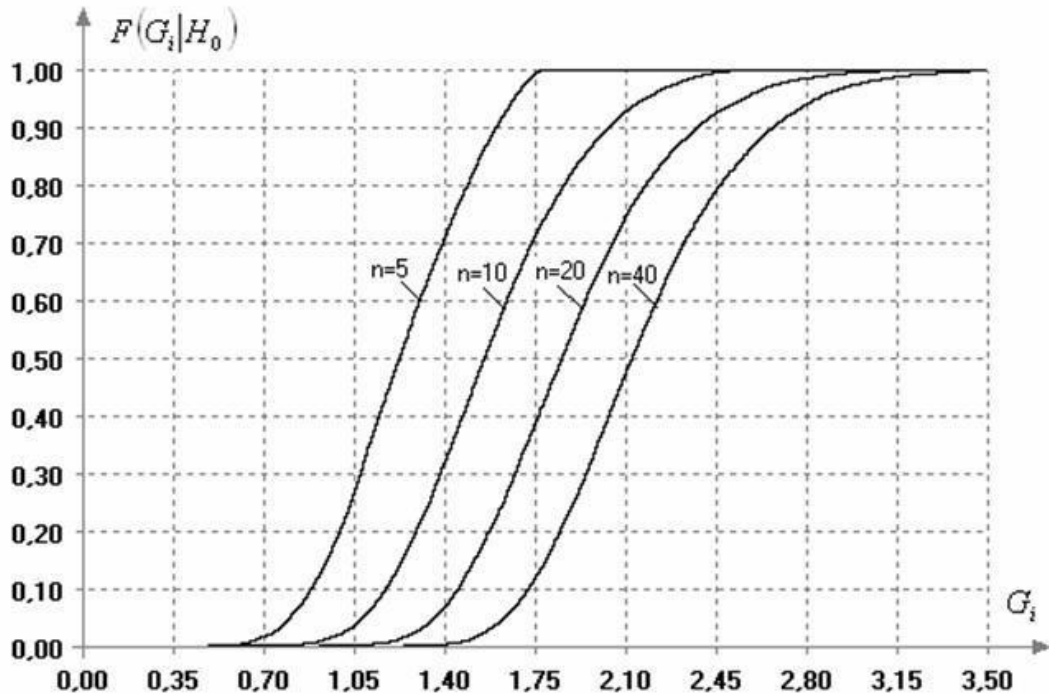


Рис. 1. Зависимость распределения статистик (1) и (5) критерия Граббса от объема выборок n (в случае нормального закона)

Критерий Граббса позволяет находить несколько выбросов (1, 2, 3 можно расширить до n), но основное предположение этого критерия, это нормальное распределение данных, что очень сильно ограничивает возможность применения данного критерия к автоматической обработки данных.[8]

1.2.2. Метод Диксона (Dixon's Q test)

В данном разделе описывается подход к решению проблемы определения аномального спроса на наличность в сети банкоматов самообслуживания с помощью Q -критерия Диксона. Приводятся формулы для расчёта критических точек.

В статистике, Q -критерий Диксона, или же просто Q -тест, используется для идентификации и исключения аномальных значений (выбросов). Предполагается, что выборка имеет нормальное распределение и тогда. Чтобы применить Q -тест для плохих

данных, необходимо расположить данные в порядке возрастания значений и вычислить Q следующим образом:

$$G = \frac{g}{r}$$

где g – модуль разности между предполагаемым выбросом и ближайшим к нему значением, а r – разность между максимальной и минимальной точкой. Если $Q > Q_{table}$, где Q_{table} является контрольным значением, соответствующим размеру выборки и уровнем достоверности, отклоните сомнительную точку. Обратите внимание, что из набора данных с помощью Q-теста может быть отвергнута только одна точка.

В зависимости от объёма выборки критерий (или коэффициент) Диксона обозначают, как показано на рис. 1. При наличии одновременно наименьшего и наибольшего выброса (двусторонних выбросов) считают, что односторонний выброс один.[9-11]

n	Число односторонних выбросов в вариационном ряду	
	Один	Два и больше
3..7	r_{10}	r_{20}
8..10	r_{11}	r_{20}
11..13	r_{21}	r_{21}
14..30	r_{22}	r_{22}

Рис.2. Число односторонних выбросов в вариационном ряду.

Рассчитывают коэффициент Диксона, как показано на рис.3.

Коэффициент Диксона для выброса	
Наименьшего	Наибольшего
$r_{10} = (x_2 - x_1) / (x_n - x_1)$	$r_{10} = (x_n - x_{n-1}) / (x_n - x_1)$
$r_{11} = (x_2 - x_1) / (x_{n-1} - x_1)$	$r_{11} = (x_n - x_{n-1}) / (x_n - x_2)$
$r_{21} = (x_3 - x_1) / (x_{n-1} - x_1)$	$r_{21} = (x_n - x_{n-2}) / (x_n - x_2)$
$r_{22} = (x_3 - x_1) / (x_{n-2} - x_1)$	$r_{22} = (x_n - x_{n-2}) / (x_n - x_3)$
$r_{20} = (x_3 - x_1) / (x_n - x_1)$	$r_{20} = (x_n - x_{n-2}) / (x_n - x_1)$

Рис.3. Коэффициент Диксона для выброса

1.3. Анализ данной системы прогнозирования.

В данном разделе рассматривается данная система прогнозирования инкассаций, приводятся её преимущества в сравнении с мировыми аналогами.

В виду актуальности указанной выше проблемы, на кафедре Кибернетики НИЯУ «МИФИ» была разработана соответствующая модель прогнозирования

инкассаций . На примере реальных данных за 2009 год одного из партнеров компании был построен прогноз, оптимизирующий работу банкоматов, позволивший увеличить прибыль с одного устройства до 60 000 рублей в год, а также уменьшить количество отказов в 5 раз.

В данной модели все банкоматы классифицируются на 4 типа:

- Зарплатные банкоматы
- Банкоматы с ограниченным доступом
- Проходные банкоматы
- Непроходные банкоматы

Для каждого из типов существуют вспомогательные модели прогнозирования (модель Хольта, модель Хольта-Уинтерса, модель множественной регрессии) [2], позволяющие добиться наибольшей точности прогноза.

Разрабатываемая система обладает следующими характеристиками:

- 1) Модуль составления маршрутов инкассации - на основе разбиения банкоматов по регионам и объединения их в специализированные кластеры, позволяет снизить расходы на инкассацию за счет подбора оптимального маршрута.
- 2) Мультивалютность системы - возможность прогнозирования и рекомендаций сразу по нескольким валютам. Если банкомат работает с несколькими валютами, то состояние по валютам может не совпадать: одна из них кончается, а другие еще есть. Предлагаемое решение связано с ранжированием валют: одна из валют определяется главной и именно по ней первоначально определяется необходимость инкассации.

3) Построение системы под 8-кассетные банкоматы. На данный момент в мире подавляющее число банкоматов имеют 4 кассеты, но уже начинают появляться и 8-кассетники. Реализация системы для 8-кассетных банкоматов является одним из приоритетных направлений для дальнейшего развития данной системы. [13]

1.4. Анализ методов и средств прогнозирования уже существующих программных реализаций.

Полезным этапом реализации модуля для системы прогнозирования инкассаций является рассмотрение уже существующих программных решений, предназначенных для прогнозирования временных рядов, с целью проведения их сравнительного анализа и определения требований, которым должен удовлетворять модуль, разрабатываемый в рамках данной работы.

Программа для прогнозирования продаж ForExSal

Программа ForExSal(Forecasting Expert Sales System) компании KonSi предназначена для постоянного построения прогнозов для многочисленных рядов различных товаров.

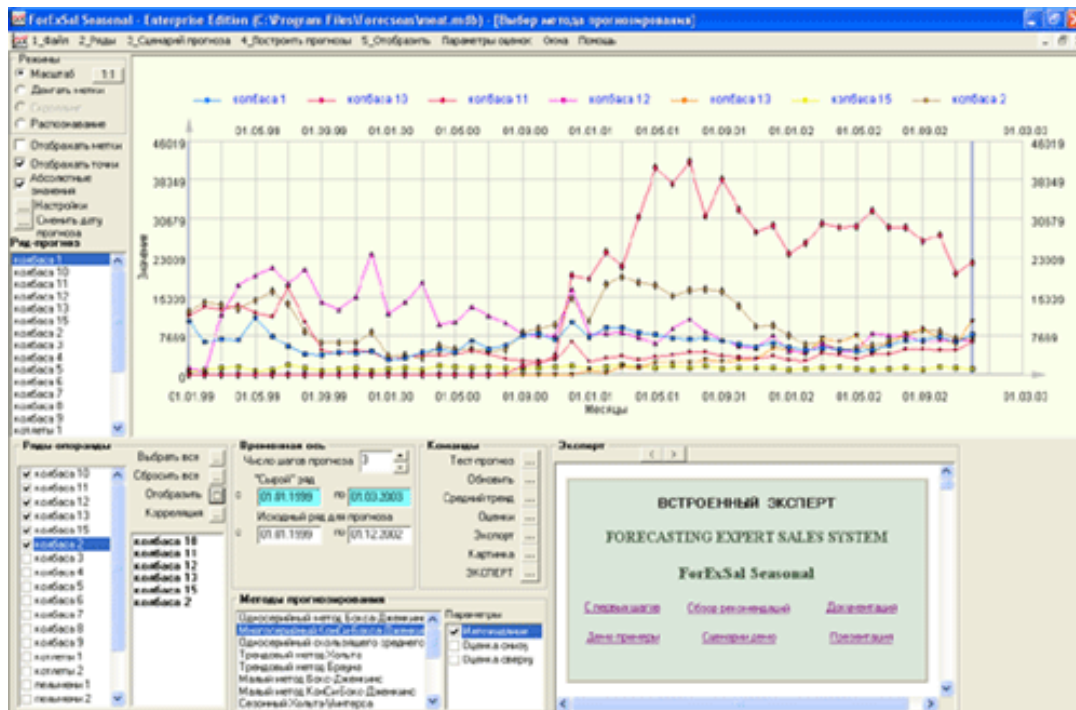


Рис. 4. Вид пользовательского интерфейса программы ForExSal

Система ForExSal позволяет быстро импортировать данные из таблиц Excel, файлов текстового формата .csv. Также вводить ряды непосредственно в программе. Данное решение обладает следующими функциональными возможностями:

- Построение прогнозов для односерийных и многосерийных рядов, между которыми существуют корреляционные зависимости;
- Одновременное прогнозирование многих независимых рядов;
- Возможность подготовки индивидуальных сценариев построения прогноза для каждого исследуемого ряда;
- Построение краткосрочных и долгосрочных прогнозов (долгосрочные прогнозы строятся обычно для сезонных рядов);
- Оценки точности прогноза как с помощью табличного сопоставления рассчитанных значений и реальных данных, так и с помощью графического наложения двух рядов;

- Предобработка данных: сглаживание аномалий и устранение статистически значимых случайностей.

В качестве основных методов прогнозирования система ForExSal использует:

- метод Бокса-Дженкинса;
- метод скользящего среднего;
- метод Хольта;
- метод Брауна;
- метод Хольта-Уинтерса.

Программа ForExSal предназначена для планирования продаж товаров, однако для оценки запасов и определения оптимальных размеров запаса она малоприменима.

Программа для прогнозирования спроса и управления запасами Forecast NOW!

Forecast NOW! - это система управления складскими запасами. Программа предназначена для розничных и оптовых торговых предприятий, а также аутсорсинговых компаний в сфере управления складскими запасами.

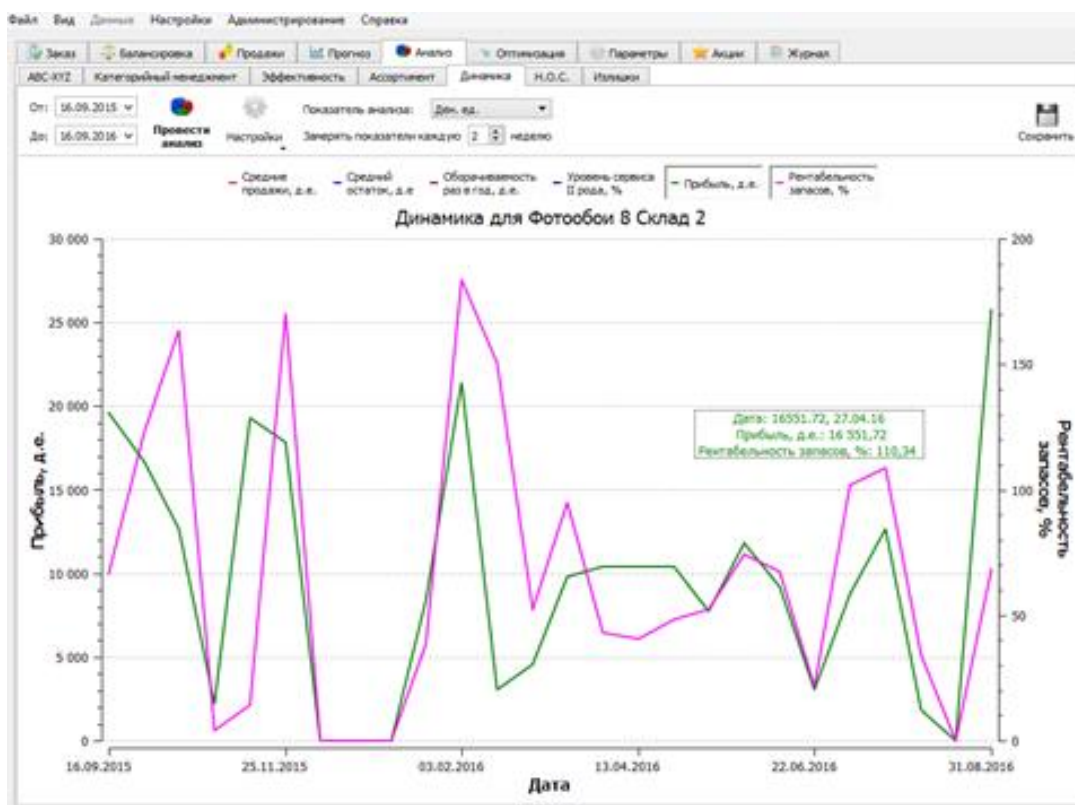


Рис. 5. Вид пользовательского интерфейса программы Forecast NOW!

Данный продукт обладает следующими функциональными возможностями:

- Построение краткосрочного (до 4 недель) и долгосрочного (до 6 месяцев) прогноза;
- Определение оптимального товарного запаса с заданным уровнем сервиса;
- Построение отчёта о необходимых закупках в ручном и автоматическом режимах с учётом внешних ограничений (кратность поставки, минимальный остаток) и расписания поставок;
- Применение фильтров, по любым свойствам товаров и с учетом заданной группировки товаров.

Для прогнозирования в программе используются следующие модели:

- нейронные сети;
- генетическая стабилизация;
- статистический бутстрэп (метод Эфрона);
- алгоритм Деккера;
- метод Виллемейна.

В системе Forecast NOW! также реализован алгоритм определения характера спроса и выбора оптимальной прогнозирующей модели. Данный подход позволяет быстро рассчитать прогноз без глубокого анализа предметной области.

Система автоматического прогнозирования Sales-Forecast

Система Sales-Forecast является последней разработкой компании StatSoft Russia в области автоматизации процессов прогнозирования. Ядром системы являются методы прогнозирования, реализованные в пакете STATISTICA, которые позволяют строить прогнозы временных рядов как на основе их собственной истории, так и с привлечением дополнительных переменными построением многомерных поясняющих моделей.

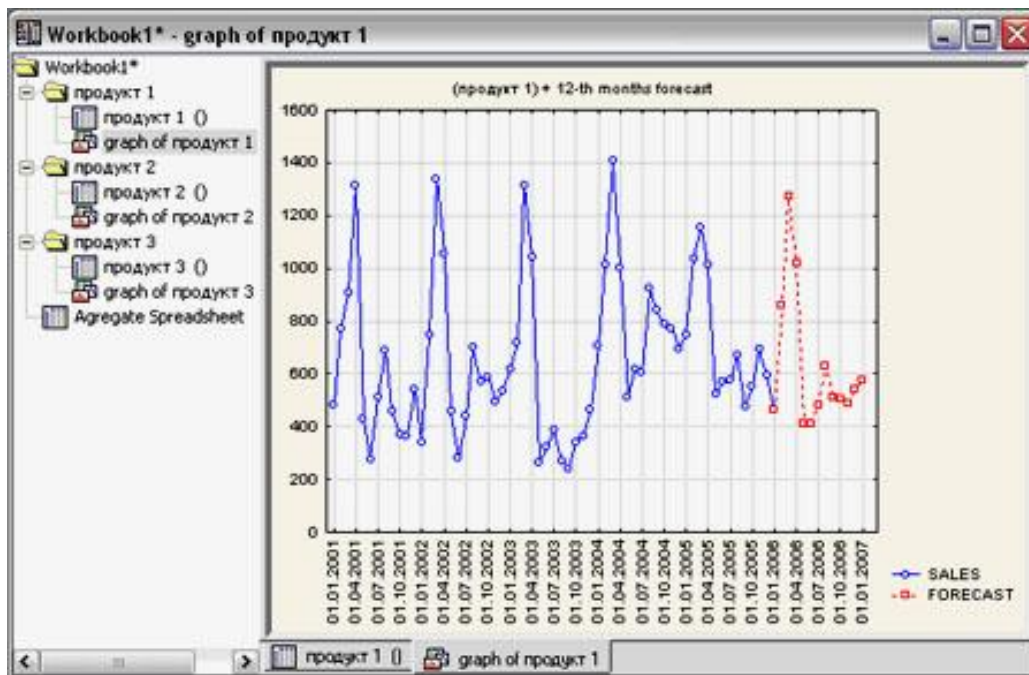


Рис. 6. Вид пользовательского интерфейса программы Sales-Forecast.

Данное решение использует следующий набор прогностических и аналитических методов:

- спектральный анализ;
- методы экспоненциального сглаживания;
- модели ARIMA и ARIMA с интервенциями;
- методы сезонной декомпозиции;
- нейронные сети.

Система Sales-Forecast допускает гибкие настройки формата исходных данных, требований к вычислительным ресурсам, а также сложности и количества используемых моделей. Основное достоинство системы автоматического прогнозирования состоит в том, что все вычисления, связанные с оптимизацией прогностических моделей перекладываются на компьютер. Это, впрочем, никак не исключает возможности внесения аналитиком экспертных поправок.

1.5. Выводы

В данном разделе приводятся заключения относительно анализа предметной области и существующих методов решения поставленной задачи.

Сформулированы основные выводы, сделанные после изучения природы аномальных значений в статистических рядах и анализа методов их обнаружения и корректировки.

Был проведен анализ и оценка методов прогнозирования для сети банкоматов самообслуживания программных реализаций мировых аналогов.

В ходе анализа теоретической части поставленной задачи выяснилось, что природу аномальных значений во временных рядах принято разделять на два класса, к которым могут относиться данные выбросы — искусственные и естественные. Искусственные выбросы чаще всего связаны с ошибками ввода данных, в то время как естественные отражают факты и события, имевшие место в действительности и имеющие некоторую закономерность (зарплатные периоды). Также из определенных источников стало известно, что принято разделять три группы нехарактерных аномальных уровней.

К первой группе относятся аномальные значения выборки, которые представляют собой существенно важную информацию. Например, в процессе расследования мошенничества с банковскими счетами и кредитными картами, аномальные значения являются индикаторами мошеннической деятельности.

Ко второй группе относятся аномальные значения, которые тоже не должны исключаться из общей совокупности, а скорее приниматься за «повторные» (или пороговые), начиная с которых необходимо пересчитывать все предыдущие значения временного ряда по новой методике.

К третьей группе относятся аномальные значения, которые должны быть исключены из рассмотрения в любом случае, так как они искажают общее представление о характере развития явления и несомненно оказывают существенное влияние на выводы, полученные в результате анализа ряда, содержащего искаженную информацию подобного типа.[4]

Стоит отметить, что в настоящее время существует множество методов определения аномальных значений в статистических данных, и их определение является одной из ключевых задач в прогнозировании временных рядов.

1.6 Цели и задачи учебно-исследовательской работы

В данном разделе на основе проведения анализа предметной области сформулирована цель, которая должна быть достигнута в рамках данной учебно-

исследовательской работы. Также в разделе перечислены основные задачи, которые необходимо выполнить для достижения поставленной цели.

Основной целью данной учебно-исследовательской работы является разработка программного модуля определения аномальных значений спроса на денежные средства в сети банкоматов для системы прогнозирования инкассаций. Для достижения поставленной цели необходимо решить следующие задачи:

- Провести анализ предметной области на тему обнаружения аномальных уровней временных рядов;
- Рассмотреть существующие подходы к решению задачи обнаружения аномальных уровней во временных рядах;
- Построить модель прогнозирования временных рядов, чувствительную к факторам аномального спроса;
- Спроектировать модуль обнаружения аномального спроса на наличность в сети банкоматов самообслуживания;
- Протестировать разработанный модуль на реальных данных и сделать вывод о полученных результатах;
- Оформить результаты учебно-исследовательской работы в виде пояснительной записки.

Таким образом, проведена постановка задачи на учебно-исследовательскую работу.

2.Разработка подходящей модели прогнозирования финансовых рядов, чувствительной к факторам аномального спроса.

Одним из преимуществ работы с временными рядами является возможность прогнозирования их дальнейших значений с помощью тех или иных моделей и алгоритмов прогнозирования. Так, например, на основании предыдущих значений временных рядов можно построить прогноз на тенденцию спроса на рынке либо на изменение погоды. Ввиду специфических свойств данных временных рядов для работы с ними применяются специализированные статистические методы и подходы.

В данной работе было принято решение использовать модель логистической регрессии для предсказания будущих значений временных рядов. Логистическая регрессия или логит-регрессия (англ. *logit model*) — это статистическая модель, используемая для предсказания вероятности возникновения некоторого события путём подгонки данных к логистической кривой.[13]

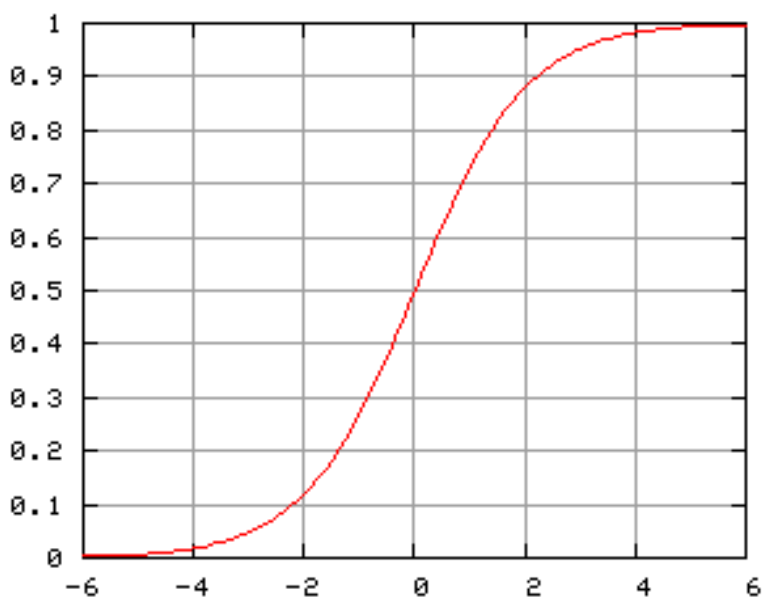


Рис. 7. Логистическая кривая

При выборе учитывалось то, что данный статистический метод довольно известный и его реализация используется в Microsoft как разновидность алгоритма работы нейронных сетей - Microsoft Logistic Regression Algorithm[14].

Как и для многих моделей, для подбор параметров модели логистической регрессии необходимо составить обучающую выборку, состоящую из наборов значений независимых переменных и соответствующих им значений зависимой переменной y .

После построения модели временного ряда можно приступать к прогнозированию данных. Для начала нужно сравнить прогнозируемые значения с реальными значениями временного ряда, что поможет нам понять точность прогнозов. Соответствующие методы позволяют получать значения и интервалы для прогнозов временных рядов.

В конце работы будет приведена точность прогнозов, как оценка эффективности работы используемых алгоритмов. В качестве данной оценки будет использоваться MSE (Mean Squared Error), что суммирует среднюю ошибку прогнозов. Для каждого прогнозируемого значения нужно вычислить его расстояние до истинного значения. Результаты нужно возводить в квадрат, чтобы различия не компенсировали друг друга при вычислении общего среднего.

2.1. Разработка метода определения аномального спроса на денежные средства

Реализация алгоритма определения аномальных значений временных рядов будет проходить в несколько этапов.

Перед импортом необходимых библиотек и определением функций для чтения данных стоит этап нормализации функций и оценка гауссовой дистрибуции. Модель Гаусса здесь будет использоваться для изучения базового шаблона набора данных с надеждой на то, что используемые в алгоритме функции следуют за гауссовским распределением. После этого находятся точки данных, которые с очень низкой вероятностью являются нормальными. Следовательно, такие точки можно считать выбросами. Для набора обучения алгоритма сначала проводится изучение гауссовского распределения каждой функции, для которой требуется среднее и дисперсия функций. В данном случае подключаемый модуль Numpy языка Python обеспечивает метод для вычисления как среднего значения, так и дисперсии (ковариационной матрицы). Аналогично, библиотека Scipy предоставляет метод оценки гауссовского распределения.

Затем определяется функция для нахождения оптимального значения порога (эпсилон), которое может использоваться для различения нормальных и аномальных

точек данных. Для изучения оптимального значения ϵ используются разные значения в ряде изученных вероятностей в наборе кросс-валидации. F-оценка рассчитывается для прогнозируемых аномалий, основанных на доступных исходных данных. Значение ϵ с наивысшим показателем f будет выбрано как пороговое значение, то есть вероятности, которые лежат ниже выбранного порога, будут считаться аномальными.[15]

После предобработки данных идёт непосредственная реализация алгоритма обнаружения и обработки аномальных значений. В данной работе было принято решение использовать критерий Граббса-Смирнова. Подробное описание алгоритма и теоретические аспекты изложены в п. 1.2.1.

Стоит отметить, что критерий Граббса-Смирнова нашел своё применение в такой программной реализации как система автоказов Microsoft Business Solution (MBS). Анализируя историю продаж и текущие остатки по подразделениям, Microsoft Assistant формирует прогноз и дает оптимальные рекомендации о том, какие товары необходимо закупать, и на какой магазин или склад их везти. А так же формирует план продаж и закупок в будущих периодах.[16]

После объявления всех вышеперечисленных методов идёт их непосредственное использование на тестовых данных с последующей визуализацией полученных результатов.

Для удобства визуального представления графиков, данные временного ряда будут агрегированы по дням.

В конце реализуется описанная выше модель прогнозирования и на основе сравнения прогноза с исходными данными оценивается эффективность работы алгоритма.

2.2. Предложенный алгоритм определения аномального спроса на денежные средства.

Как и говорилось ранее, разработанный алгоритм состоит из следующих этапов: Пусть $dataset$ - импортируемый набор данных о транзакциях в сети банкоматов самообслуживания.

1. Оценка гауссовского распределения.

Найдём коэффициенты μ , σ :

- μ найдём как среднее арифметическое значений снятия выбранного автомата;
- Σ - ковариационная матрица $dataset$

Пусть мы исследуем N-мерный набор данных $X = [x_1, x_2, \dots, x_N]^T$, тогда ковариационный матричный элемент C_{ij} является ковариацией x_i и x_j . При этом элемент C_{ii} является дисперсией элемента x_i .

2. Вычисление функции многомерной нормальной случайной величины.

Определим переменную p , как

$$f(x) = \frac{1}{\sqrt{(2\pi)^k \det \Sigma}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right),$$

где μ и Σ нам уже известно, k - размерность пространства, в котором значения массива исходных данных принимают значения.

3. Определение оптимального значения порога различения нормальных и аномальных точек данных.

Оптимальное значение порога ε будет находиться следующим образом:

Пусть $\varepsilon=0$, $\Sigma=0$. Построим последовательность элементов

$$[p_{min}; p_{max}] \quad (1)$$

таким образом, что:

$$x_{i+1} - x_i = \frac{p_{max} - p_{min}}{1000}.$$

По сути это и есть последовательность значений ε , среди которых необходимо выбрать оптимальные значения. Для этого воспользуемся F1-метрикой, суть которой состоит в том, что необходимо вычислить оценку F1, также известную как сбалансированная F-оценка или F-мерка. F1-метрику можно интерпретировать как средневзвешенное значение точности и отзыва, где оценка F1 достигает своего наилучшего значения при 1 и худшем балла в 0. Относительный вклад точности и отзыва в оценку F1 является равным. Формула для оценки F1:

$$F1 = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall}) \quad (2)$$

В данном случае переменная *precision* принимает значение 1, если $p < \varepsilon$, и 0 - в противном случае.

Определим 2 фиктивные переменные:

$$f_{best} = 0, \varepsilon_{best} = 0.$$

Теперь для каждого элемента последовательности (1) найдём переменную \square по F1-метрике. При этом, если $f > f_{best}$,

$$f_{best} = f,$$

$$\varepsilon_{best} = \varepsilon.$$

Найденное таким образом значение ε_{best} и является оптимальным значением порога между нормальными и аномальными значениями данных. Соответственно за выбросы будут приниматься те значения исходного массива данных, для которых $p < \varepsilon_{best}$.

4. Использование теста Граббса-Смирнова для более глубокого сканирования на наличие аномальных значений в данных.

Так как тест Граббса-Смирнова уязвим к большому количеству выбросов в выборке исходных данных, было принято решение использовать его после отсеивания явных кандидатов на аномальные значения в качестве более глубокого анализа.

4.1. Все элементы выборки выстраиваются по воз(от меньшего к большему)

$$x_1 \leq x_2 \leq \dots \leq x_n$$

4.2. Определение максимально возможный процент выбросов в выборке (указывается экспертом, например 5%)

Здесь определяется верхняя граница количества выбросов. Это не значит, что число выбросов будет составлять строго 5% от выборки, это число будет меньше либо равно 5%. Этот параметр обозначим за g .

4.3. Строим ряд разностей

$$\Delta_i = x_{i+1} - x_i,$$

таким образом получим $n-1$ положительных элементов в последовательности:

$\Delta_1, \Delta_2, \dots, \Delta_{n-1}$, (последовательность не обязательно возрастающая)

4.4. Находим максимальную разность

В последовательности разностей находим максимальную разность: $\max \Delta_i$.

4.5. Проверка, попадает ли выброс в заданную в п.2. область

В п.4. получили некоторый индекс i' — который является максимальным в последовательности разностей. Если этот индекс $i' \geq n * g$ и, в тоже время, $i' \leq n - n * g$ то в выборке нет выбросов, прекращаем проверку. В противном случае, элемент с индексом i' подозрителен на выброс. (если индекс попал в интервал

$[n * g, n - n * g]$ это значит, что максимальная разность достигнута в середине выборки, где предположили, что выбросов нет, задав параметр g)

4.6. Вырезаем подвыборку из начального набора данных

Здесь имеем два случая:

если $i' < n * g$, тогда рассматриваем только Δ_i у которых $i \in [i', n - 1]$ (т.е. обрезаем выборку с начала),

если $i' > n - n * g$, тогда рассматриваем только Δ_i , у которых $i \in [1, i']$ (т.е. обрезаем выборку в конце).

4.7. Среднее значение разностей

Находим среднее значение выборки разностей (п.3.) с учетом набора индексов (п.6.):

$$\Delta' = \frac{1}{n'} \sum_i^n n \Delta_i,$$

где n' — число элементов в вырезанной подвыборке.

4.8. Среднеквадратическое отклонение

Рассчитываем характеристику разброса элементов:

$$S^2 = \frac{1}{n'-1} \sum_i^n n (\Delta_i - \Delta')^2.$$

4.9. Считаем статистику Граббса

Используем формулу Граббса для расчета значения статистики:

$$G_{i'} = \frac{|\Delta' - \Delta_{i'}|}{S}$$

4.10. Определение уровня значимости

α — величина вероятности ошибки (обычно 1%, 5%, 10%).

4.11. Использование таблицы критических значений, сравнение величины статистики из п.9. с критическим значением.

По значению уровня значимости и числу элементов в подвыборке n' находим критическое значение в таблице Граббса, обозначим его за $G_{\alpha, n'}$.

Сравниваем полученное значение полученное в п.9. Возможны два случая:

$G_{i'} < G_{\alpha, n'}$ — исследуемое значение не является выбросом и в начальном наборе данных нет выбросов вообще,

$G_{i'} > G_{\alpha, n'}$ — исследуемое значение является выбросом. Если оно находится во второй половине выборки (п.1.), то все значения идущие после i' являются выбросами, если

оно находится в первой половине выборки (п.1.), то все значения идущие до i' являются выбросами.

4.12. Если в п.11. был обнаружен выброс, то из начальной выборки убираются выбросы и весь алгоритм повторяется с первого пункта (п.1.)

Стоит отметить, что данный подход в некотором роде учитывает природу аномалий. Так, например, зарплатные периоды, имеющие некоторую цикличность, не будут удалены из общей выборки данных.

По сути, предложенный алгоритм является гибридом описанных выше методов, правильное сочетание и применение которых может улучшить результат прогнозов.

3. Проектирование системы

Важным этапом разработки является проектирование, так как результаты этой стадии являются входной информацией для стадии программной реализации. В данной работе этап проектирования состоит из трёх шагов: определения требований к проектируемому модулю, выбор инструментальных средств и представления общей схемы работы модуля в виде диаграммы UML.

3.1 Требования к проектируемому модулю

Для того чтобы избежать ошибок или недочетов в ходе программной реализации, необходимо определить ряд требований к проектируемому модулю. В конкретном случае было принято решение, что модуль должен удовлетворять следующим требованиям:

- Модуль должен загружать файлы с данными в формате .csv, .xls и .xlsx;
- Модуль должен проверять корректность входных данных;
- Модуль должен при необходимости обрабатывать пропущенные значения в данных;
- Модуль должен содержать выбранную модель прогнозирования;
- Модуль должен предоставлять возможность установки необходимых параметров выбранной модели;
- Модуль должен корректно реализовать разработанный алгоритм по обнаружению и корректировке аномальных значений в данных;
- Модуль должен визуализировать исходные данные, исходные данные с восстановленными значениями и прогноз;
- Модуль должен приводить оценку точности прогноза до и после работы алгоритма по обнаружению аномальных значений;
- Модуль должен быть написан таким образом, чтобы обеспечить возможность его дальнейшей модификации;

3.2. Проектирование архитектуры разрабатываемого модуля

Так как исходные данные могут быть представлены в любом виде, необходимо проделать ряд операций для их обработки и приемлемого представления для реализации последующих алгоритмов работы с ними. По сути данная подзадача сводится к написанию методов, таких как:

- Метод чтения из файла определённого формата (csv, xls, xlsx, ...).

- Метод формирования временных рядов из соответствующих столбцов исходных данных. Это необходимо, потому как не всегда значения времени представлены соответствующим видом и типом; а также из-за того, что многие алгоритмы и модели прогнозирования работают именно с временными рядами, а не с произвольными массивами данных.
- Метод агрегирования данных для представления в нужном виде.
- Непосредственно метод, реализующий обнаружение аномальных значений спроса на денежные средства в данных о снятиях денежных средств.
- Метод для создания нужной модели прогнозирования с определёнными параметрами.
- Методы визуализации конкретных этапов работы модуля.
- Метод для оценки эффективности работы модуля определения аномального спроса.

Правильное построение структуры программы и соблюдение всех требований должен обеспечить корректное взаимодействие вышеприведенных методов и получение ожидаемого результата.

3.3 Технологии и инструментальные средства проектирования

В первую очередь, язык программной реализации должен иметь широкий спектр возможностей в области анализа данных и отличаться доступностью необходимых подключаемых пакетов. Так как придётся иметь дело с начальной обработкой исходных данных и работать с их статистическими характеристиками, разумно будет выбрать из специально адаптированных (ну или же просто универсальных) языков: R, Matlab, Python, Java и т.д. Также заранее известно, что работа будет проводиться с подключением некоторого количества пакетов для статистического анализа и математических операций.

Также стоит отметить, что многие алгоритмы опираются на предположение о нормальном законе распределения данных, вследствие чего необходимо создание методов, нормализующих исходную выборку, либо же модификация соответствующего алгоритма для возможности автоматической обработки данных.

3.4 Результаты разработки

Для описания результатов, полученных на предыдущих этапах разработки модуля, было принято решение использовать средства унифицированного языка моделирования (UML).

Для описания реакции объекта на то или иное событие используется диаграмма состояний. Данная диаграмма может описывать как поведение системы в целом, так и отдельных её компонентов[17]. На рисунке 7 представлена диаграмма состояний модуля.

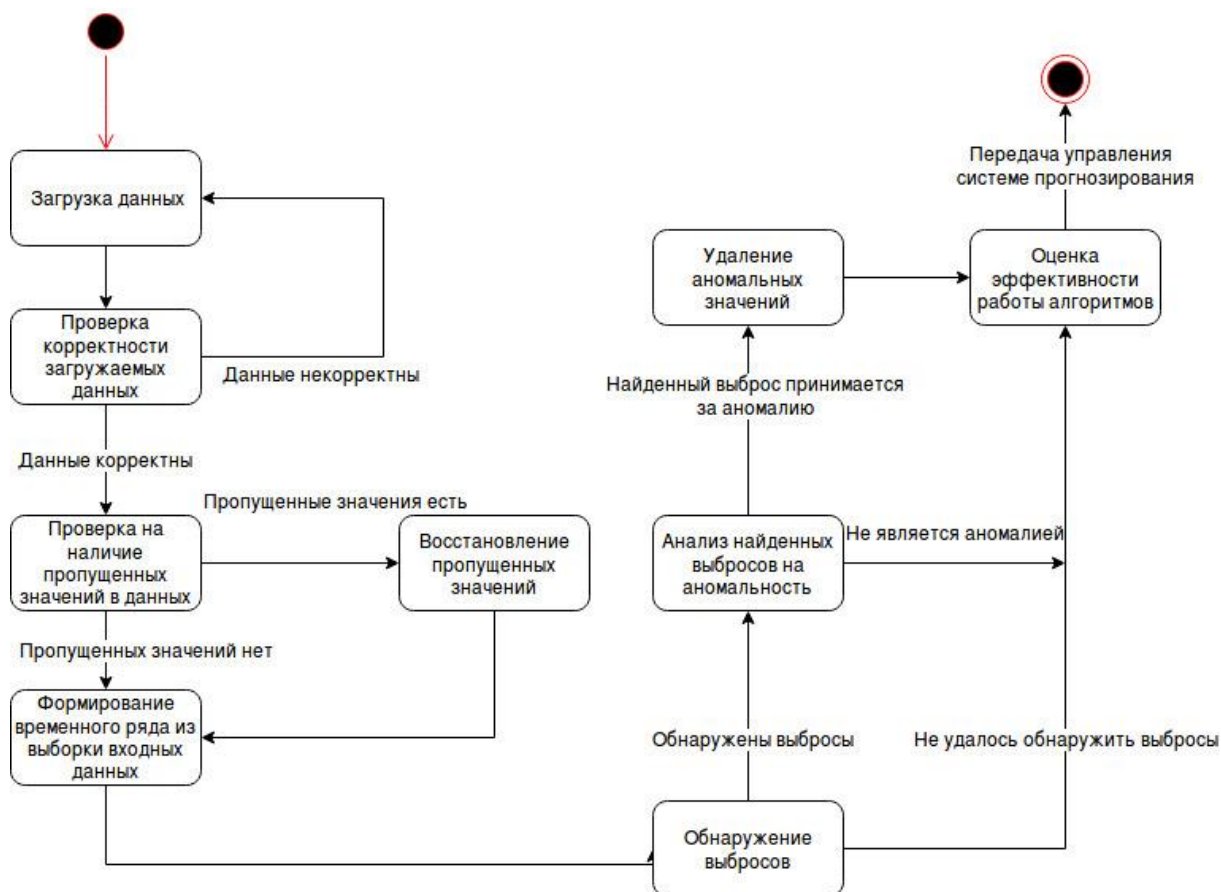


Рис. 7.UML-диаграмма состояний разрабатываемого модуля

Описание основных состояний модуля:

- **Загрузка данных.** На данном этапе пользователь выбирает исходный файл с данными, представляющие собой данные о денежных снятиях в сети банкоматов самообслуживания. Предполагается, что пользователь загружает файл формата .csv, .xls или же .xlsx. После загрузки данных алгоритм модуля переходит в состояние “Проверка корректности загружаемых данных”.

- Проверка корректности загружаемых данных. После того, как пользователь загрузил файл с исходными данными о транзакциях, система начинает проверку на соответствие данных установленным форматам. Если входные данные корректны, модуль переходит в состояние “Проверка на наличие пропущенных значений в данных”, иначе - возвращается в “Загрузка данных”.
- Проверка на наличие пропущенных значений в данных. После проверки на корректность входные данные анализируются на наличие в них пропущенных значений (NaN), усложняющих расчёт прогноза и нарушающих работу методов отдельных программных пакетов. В случае, если данные имеют пропуски, система переходит в состояние “Восстановление пропущенных значений”. В обратном случае система переходит в состояние “Формирование временного ряда из выборки входных данных”.
- Восстановление пропущенных значений. После того, как в загруженных данных были обнаружены пропущенные значения, запускается метод их восстановления. По этапу восстановления модуль переходит в состояние “Формирование временного ряда из выборки входных данных”.
- Формирование временного ряда из выборки входных данных. На данном этапе загруженные данные обрабатываются соответствующим образом для их последующего представления в виде временного ряда. Используются методы обработки значений даты транзакций, значений транзакций определённого АТМ, а также методы агрегации значений временных рядов для удобства работы с ними. По завершению данного этапа модуль переходит в состояние “Обнаружение выбросов”.
- Обнаружение выбросов. После того, как временной ряд из значений спроса на денежные средства сформирован, запускаются алгоритмы определения аномальных значений, описанные в п.2.1. Если в ходе работы алгоритмов были найдены статистические выбросы, модуль переходит в состояние “Оценка эффективности работы алгоритмов”. В противном случае модуль переходит в состояние “Анализ найденных выбросов на аномальность”.
- Анализ найденных выбросов на аномальность. После того, как соответствующие алгоритмы находят выбросы в статистических данных временных рядов, следующие алгоритмы проверяют данные значения на аномальность (необходимость их корректировки). Если в последствии работы алгоритмов на

данном этапе найденные значения принимаются за аномальные, система переходит в состояние “Удаление аномальных значений”. В обратном случае система переходит в состояние “Оценка эффективности работы алгоритмов”.

- Удаление аномальных значений. Если найденные ранее выбросы были приняты за аномальные значения, искажающие общую статистику временного ряда, принимается решение их удалить. После завершения данной операции модуль переходит в состояние “Оценка эффективности работы алгоритмов”.
- Оценка эффективности работы алгоритмов. В данном состоянии модуль запускает последний алгоритм оценки полученного результата и выводит отчёт о проделанной работе. В свою очередь отчёт включает в себя исходные данные с восстановленными значениями и исключенными аномалиями, оценку алгоритмов по некоторой метрике, а также графики для визуального анализа проделанной работы. Далее модуль передаёт управление системе прогнозирования вместе с данными, полученными по окончании его работы.

4. Программная реализация модуля определения аномального спроса на наличность

Основной задачей реализации программной системы является выбор языка программирования. В начале выполнения данной работы было принято использовать два языка программирования - один для сбора общих статистических характеристик предоставленных данных, другой для непосредственного построения модуля прогнозирования и обработки аномальных значений спроса на денежные средства в системе банкоматов самообслуживания.

Для выбора языка реализации модели были сформулированы следующие требования: инструмент программной реализации должен быть адаптирован для работы со статистическими данными и давать возможность достаточно легко и быстро провести анализ данных и реализовать необходимые алгоритмы (прогнозирование, обнаружения аномального спроса и т.д.). В ходе подробного анализа различных источников был сформулирован следующий список языков:

- **R.** Данный язык был разработан непосредственно для статистической обработки данных и работы с графикой, а также является свободной программной средой вычислений с открытым исходным кодом в рамках проекта GNU. Является куда менее универсальным, чем тот же Python, тем не менее широко используется как стандартное программное обеспечение для анализа данных и написания статистических программ. Имеет большое количество подключаемых библиотек для работы с различными видами данных.
- **Python.** Язык программирования Python — это мощный инструмент для создания программ самого разнообразного назначения, имеющий широкий спектр возможностей в области Data Science за счёт большого количества подключаемых библиотек. Как интерпретируемый язык Python также имеет ряд преимуществ в использовании: понятный и лаконичный синтаксис, динамическая типизация, кроссплатформенность и т.д. При этом Python поддерживает принципы ООП и с его помощью можно решать задачи различных типов.
- **Matlab.** Несмотря на широкую известность исключительно как пакета программных продуктов для математических вычислений, является также высокоуровневым интерпретируемым языком программирования. Реализует

куда более широкий спектр возможностей, нежели обозначенные выше языки с учетом библиотек расширений. Однако данный инструмент является проприетарным и при этом довольно дорогим.[18]

В ходе сравнительного анализа представленных языков, и в процессе реализации поставленной задачи моделирования и прогнозирования временных рядов был выбран язык Python, так как он является более функциональным и простым в использовании, а также проявил себя как более универсальный в решении задач анализа данных, вызываемых трудности (такие, как некорректное поведение подключаемых модулей, временные затраты и т.д.) при использовании вышеприведенных программных инструментов. Хотя и для сбора общих статистических характеристик предоставленных данных изначально использовался язык R, дальнейшая программная реализация, представленная в данной исследовательской работе, была полностью выполнена с использованием программных средств языка Python.

4.1. Реализация модуля обнаружения и корректировки аномальных значений в системе прогнозирования инкассаций

Для выполнения программной реализации разрабатываемого модуля было принято решение использовать версию Python 3, так как на сегодняшний день она является самой новой, и предоставляет самые широкие возможности для работы, несмотря на проблемы с адаптируемостью некоторых “старых” подключаемых библиотек.

Для написания исходного кода на языке Python 3 был установлен инструмент IPython Notebook. IPython это интерактивная оболочка для языка программирования Python, которая предоставляет расширенную интроспекцию, дополнительный командный синтаксис, а также подсветку и автоматическое дополнение кода.

Как уже отмечалось ранее, Python 3 предоставляет огромное количество модулей, как входящих в стандартную поставку Python, так и сторонних. В некоторых случаях для написания программы достаточно лишь найти подходящие модули и правильно их скомбинировать. Удобность и функциональность подключения необходимых модулей и стали главным критерием выбора языка реализации модели прогнозирования. В таблице 1 представлены используемые в данной работе пакеты этого языка:

Таблица 1. Используемые пакеты языка Python

Название пакета	Описание
NumPy	Библиотека языка Python, добавляющая поддержку больших многомерных массивов и матриц, вместе с большой библиотекой высокоуровневых (и очень быстрых) математических функций для операций с этими массивами.
Pandas	Библиотека с открытым исходным кодом, предоставляющая высокопроизводительные, простые в использовании структуры данных и инструменты анализа данных для языка программирования Python. Даёт возможность провести первичный анализ данных, а также строить сводные таблицы, выполнять группировки, предоставляет удобный доступ к табличным данным.
Scikit-Learn	Библиотека scikit-learn предоставляет реализацию целого ряда алгоритмов для обучения с учителем (Supervised Learning) и обучения без учителя (Unsupervised Learning) через интерфейс для языка программирования Python. Позволяет реализовать такие задачи как кластеризация, отбор признаков и т.д.
StatsModels	Модуль Python, который предоставляет классы и функции для оценки многих различных статистических моделей, а также для проведения статистических испытаний и исследования статистических данных. Для каждого оценщика доступен обширный список статистических данных результатов. Результаты тестируются на основе существующих статистических пакетов, чтобы убедиться, что они верны.
Matplotlib	Библиотека для визуализации данных.
Datetime	Пакет для представления данных в типе Date и дальнейшей работы с ними.

LogisticRegression	Пакет, предоставляющий методы для создания модели логистической регрессии и работы с ней.
--------------------	---

В таблице 2 приведены используемые функции из вышеупомянутых пакетов с их кратким описанием.

Таблица 2. Используемые пакетные функции языка Python

Название функции	Описание	Пакет
read_excel	Позволяет считать данные из файла формата xlsx в массив pandas	Pandas
Series	Позволяет создать ряд из данной выборки	Pandas
to_datetime	Конвертация в тип Date для построения временного ряда.	Datetime
plot()	Позволяет создать график стандартного типа	Matplotlib
LogisticRegression()	Позволяет создать модель логистической регрессии.	LogisticRegression

4.2 Тестирование на реальных данных

В качестве реальных данных используются данные о снятиях в сети банкоматов самообслуживания города Екатеринбург за 2012 год. После загрузки входных данных и их предварительной обработки (проверка на корректность данных, восстановления пропущенных значений, и т.д.) можно приступить к непосредственной работе с ними. В качестве исследуемого образца был выбран терминал АТМ-1120014. Для представления значений денежных снятий в виде временного ряда было принято решение провести агрегирование по дням. В итоге на Рис.8 можно увидеть динамику денежных снятий АТМ-1120014 в период с марта по июль.

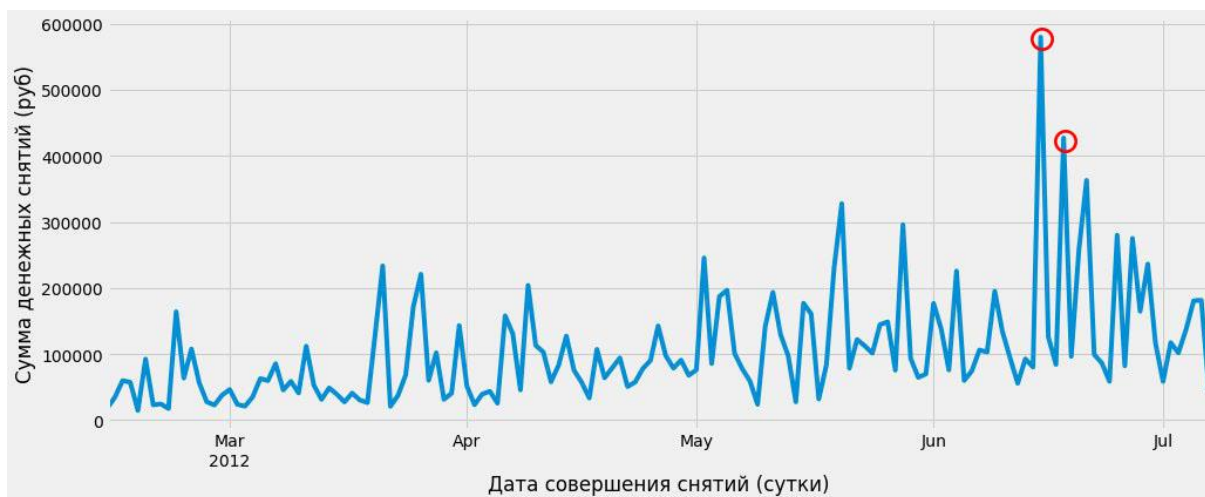


Рис. 8. Временной ряд до обнаружения аномальных значений

Красным цветом здесь изображены значения, которые на фоне общих могут быть приняты за аномальные и исключены из ряда (либо сглажены).

После иллюстрации временного ряда можно приступить к обнаружению аномальных значений спроса на денежные средства. Отправив исходные данные на вход алгоритму, описанному в п. 2.2, можно проанализировать полученный результат. Если повторно создать временной ряд из значений, полученных на выходе работы алгоритма, можно увидеть, что ситуация сложилась следующим образом:



Рис. 9. Временной ряд после обнаружения аномальных значений

Проанализировав данный график и сравнив его с предыдущим можно сделать некоторые выводы о работе модуля определения аномального спроса. Как и предполагалось, кандидаты на аномальные значения, представленные на Рис.8, по окончании работы алгоритма были сглажены, а некоторые и вовсе удалены из данного временного ряда. Значения, отмеченные на Рис. 9 красным цветом имеют некую

цикличность и вероятно могут быть зарплатными периодами, поэтому в ходе работы алгоритма не были затронуты. Однако для того, чтобы убедиться в правильности работы модуля, необходимо численно оценить эффективность используемых алгоритмов.

Следующим этапом программной реализации является построение модели логистической регрессии для прогнозирования значений денежных снятий. При помощи метода `LogisticRegression()` из пакета `sklearn.linear_model` была построена модель логистической регрессии. Для работы с моделью и составления прогнозов было принято решение разбить временной ряд на обучающие (70%) и тестовые (30%) выборки. Для обучения данной модели с помощью `model.fit(X_train, y_train)` на вход подаются массив значений транзакции выбранного автомата самообслуживания и массив дат снятия денежных средств. Для создания прогноза в данном случае необходимо использовать метод `.predict(X_test)`, который в качестве входных данных принимает массив дат транзакции, предназначенный для тестирования. В итоге исходный временной ряд разбивается на 3 массива: 2 для обучения модели (`X_train`, `y_train`) и 1 для создания прогноза (`X_test`). Таким образом, необходимо построить 2 модели: первую для временного ряда, полученного из исходных данных; вторую для данных, из которых исключены аномальные значение. Заключаящим этапом будет сравнение точности прогнозов этих двух моделей по выбранной ранее метрике.

В первую очередь была построена модель логистической регрессии для временного ряда, полученного из исходных данных о снятиях денежных средств. Для обучения модели временной ряд был разбит на два массива - значений транзакций (`y_train`) и соответствующих дат (`X_train`) - каждый из которых составляет 70% от общих массивов значений временного ряда.

Также стоит учитывать, что при подборе параметров модели необходимо проводить диагностику. Главная задача такой диагностики – убедиться, что остатки модели некоррелированы и распределяются с нулевым средним значением. Если данная модель не удовлетворяет этим свойствам, это значит, что ее еще можно улучшить. Метод `.plot_diagnostics()` позволяет посмотреть на результат построения модели прогнозирования.

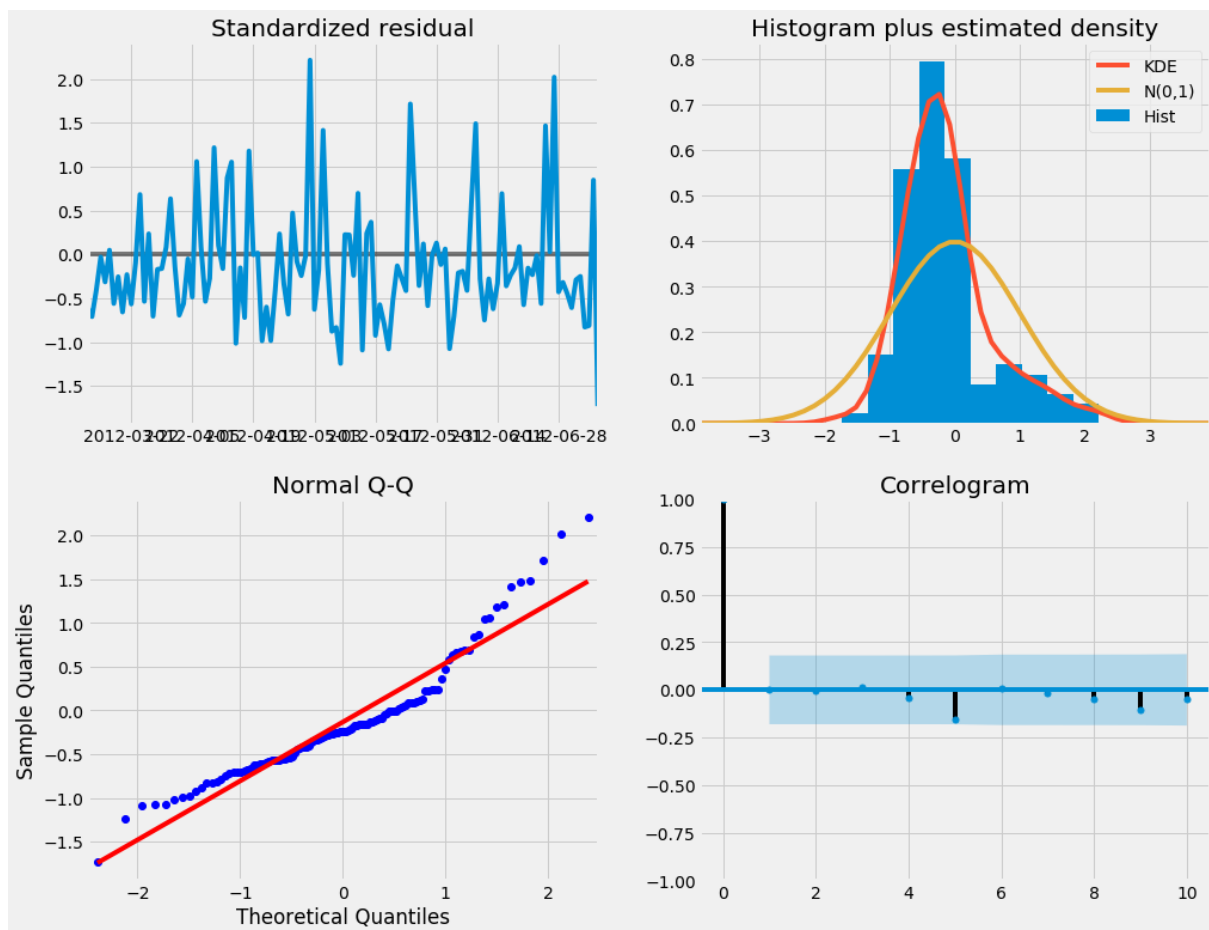


Рис. 10. Графики распределения остатков модели.

В этом случае диагностика показала, что остатки модели распределяются относительно правильно:

- На верхнем правом графике красная линия KDE находится недалеко от линии $N(0,1)$ (где $N(0,1)$ является стандартным обозначением нормального распределения со средним 0 и стандартным отклонением 1).
- График q-q в левом нижнем углу показывает, что упорядоченное распределение остатков (синие точки) следует линейному тренду выборок, взятых из стандартного распределения $N(0,1)$. Это признак того, что остатки нормально распределены.
- Остатки с течением времени (верхний левый график) не показывают явной сезонности и кажутся белыми шумами. Это подтверждается графиком автокорреляции (внизу справа), который показывает, что остатки временных рядов имеют низкую корреляцию с запаздывающими данными.

Эти графики позволяют сделать вывод о том, что выбранная модель (удовлетворительно) подходит для анализа и прогнозирования данных временных рядов. Также среди методов пакета LogisticRegression есть метод, позволяющий оценить точность обучения построенной модели. В данном случае точность обучения модели логистической регрессии для банкомата АТМ-1120014 составила 41%, что является удовлетворительным результатом в рамках данной учебной исследовательской работы.

Следующим этапом будет построение прогноза на последние 2 месяца, используя массив дат для теста - X_{test} . После того как прогноз был получен, мы можем сопоставить его с реальными данными о снятиях и оценить его качество. Так как для оценки эффективности работы алгоритма модуля был выбран Метод Наименьших Квадратов, именно его мы будем использовать как анализ точности прогноза. Погрешность работы алгоритма составляет 62.5%, то есть полученный таким образом прогноз будет на 37.5% достоверным, что является удовлетворительным результатом для данной модели, точность обучения которой равна 41%.

Далее строится аналогичная модель за исключением того, что на вход подаются обучающие и тестовые данные выборки, из которой были удалены аномальные значения впоследствии работы разработанного алгоритма. В результате была получена точность обучения модели - 41%, и точность прогноза модели, полученного по методу наименьших квадратов - 39%.

Ниже в таблице 3 представлена сравнительная характеристика построенных моделей логистической регрессии.

Обозначим:

Модель №1 - Модель прогнозирования, построенная на основе исходных данных.

Модель №2 - Модель прогнозирования, построенная на основе данных, полученных в результате работы модуля определения аномального спроса.

Таблица 3. Сравнительная характеристика построенных моделей прогнозирования.

Название модели	Точность обучения	Точность прогноза
Модель №1	~41%	37.5%
Модель №2	~41%	39%

Аналогичные испытания были проведены со следующими банкоматами самообслуживания:

АТМ - 1140039

1140064

1290001

1140043

1120002

1520002

Суммарный результат оказался следующим: точность прогноза, составленного после работы алгоритма по обнаружению аномальных данных, для большинства АТМ составила 28-39%, что в целом на $\sim 1,86\%$ больше, чем точность прогнозов, составленным по данным, которые не проходили проверку алгоритма - 26-37.5%. Стоит также отметить, что часть банкоматов показала неудовлетворительные результаты (АТМ-1520002, 1140039), точность прогноза которых была либо крайне мала, либо значительно расходилась со значением точности, полученным после работы алгоритма.

На основе данной информации можно сделать вывод, что разработанный в рамках данной работы модуль показывает удовлетворительный результат. Однако предложенный алгоритм недостаточно эффективен для внедрения в существующую систему прогнозирования инкассаций, либо же требует более детального рассмотрения и доработки.

Заключение

В данной учебно-исследовательской работе рассматривалась задача определения аномального спроса на денежные средства в сети банкоматов, с целью увеличения точности предсказаний системы прогнозирования инкассаций.

Сначала был проведен анализ ситуаций, при которых в статистических данных возникают аномальные значения. Рассмотрена классификация, а также методы обнаружения и корректировки аномальных уровней финансовых временных рядов.

Дальнейшие исследования были направлены на сравнение соответствующих моделей и методов прогнозирования, а также на сравнительный анализ методов обнаружения аномальных значений в статистических данных с целью создания программного модуля определения аномального спроса на денежные средства в сети банкоматов.

В итоге был выбран ряд методов для обработки данных и обнаружения аномальных значений спроса на наличность. Данный алгоритм был реализован при помощи инструментальных средств языка Python 3. После этого было проведено тестирование программного модуля на реальных статистических данных. В конце отчёта приводятся графики результатов работы, а также оценка эффективности используемых алгоритмов.

Литература

1. Сведения об устройствах, расположенных на территории России и предназначенных для осуществления операций с использованием и без использования платежных карт.URL: http://www.cbr.ru/statistics/print.aspx?file=p_sys/sheet016.htm (Дата обращения 12.03.2017).
2. Vijay K Narayanan, Alok Kirpal, Nikos Karampatziakis «Anomaly Detection – Using Machine Learning to Detect Abnormalities in Time Series Data».URL: <https://blogs.technet.microsoft.com/machinelearning/2014/11/05/anomaly-detection-using-machine-learning-to-detect-abnormalities-in-time-series-data/> (Дата обращения: 01.03.2017)
3. Sudheer G., Suseelatha A. Short term load forecasting using wavelet transform combined with Holt–Winters and weighted nearest neighbor models //International Journal of Electrical Power & Energy Systems. – 2015. – Т. 64. – С. 340-346.
4. Tukey, J.W. The future of data analysis [Text] / J.W. Tukey. – Annals of mathematical statistics, 1962. – 1-67.p. Выявление аномальных значений © 1995-2010 Компания BaseGroup™ Labs.URL: <http://docplayer.ru/36397870-Vyyavlenie-anomalnyh-znacheniy-m-077.html> (Дата обращения 27.02.2017).
5. Бучацкая В.В. Обработка аномальных значений уровней временного ряда как этап комплексной оценки информации в подсистеме прогнозирования для ситуационного центра // Вестник Адыгейского государственного университета. Серия 4: Естественно-математические и технические науки. 2013. No3 (122). URL: <http://cyberleninka.ru/article/n/obrabotka-anomalnyh-znacheniy-urovney-vremennogo-ryada-kak-etap-kompleksnoy-otsenki-informatsii-v-podsisteme-prognozirovaniya-dlya>(дата обращения: 2.03.2017).
6. Садовникова Н . А ., Шмойлова Р . А. Анализ временных рядов и прогнозирование : учеб . пособие . Вып . 2 / Моск . гос . ун-т экономики , статистики и информатики . М ., 2004. 184 с.
7. Микешина, Н.Г. Выявление и исключение аномальных значений [Текст]/ Н.Г. Микешина// Заводская лаборатория. – 1966,- Т.38 No 3. – С.310-318.

8. Лемешко Б.Ю., Лемешко С.Б. Расширение области применения критериев типа Граббса, используемых при отбраковке аномальных измерений // Измерительная техника. 2005. № 6. – С. 13-19.
9. Суслов В.И., Ибрагимов Н.М., Талышева Л.П., Цыплаков А.А. Эконометрия. Часть III Эконометрия - I: Анализ временных рядов Учебное пособие Новосибирск: Изд-во СО РАН, 2005. 744 с
10. Ханк Д.Э., Уичерн Д.У., Райте А.Дж. Бизнес-прогнозирование – М.: Издательский дом "Вильямс", 2003. – 656 с.
11. Сигел Э. Практическая бизнес-статистика – М.: Издательский дом "Вильямс", 2008. – 1056 с.
12. M.D. Aseev, S.A. Nemeshaev, A.P. Nesterov. Forecasting Cash Withdrawals In The ATM Network Using A Combined Model Based On The Holt-Winters Method And Markov Chains// International Journal of Applied Engineering Research (IJAER). — 2016. —Т. 11. — №. 11. — С. 7573-7578
13. Модель логистической регрессии. URL: <http://www.machinelearning.ru/wiki/>. (Дата обращения: 07.06.17)
14. Microsoft Logistic Regression Algorithm URL: <https://docs.microsoft.com/ru-ru/sql/analysis-services/data-mining/microsoft-logistic-regression-algorithm> (Дата обращения: 07.06.17)
15. Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." *ACM computing surveys (CSUR)* 41.3 (2009): 15.
16. Mycroft Business Solution // Определение выбросов по критерию Граббса URL: <http://mycroftbs.ru/grabbs/> (Дата обращения: 07.06.17).
17. Lilius J., Paltor I. P. The semantics of UML state machines. — 1999.
18. Borse G. J. Numerical methods with MATLAB: A resource for scientists and engineers. —International Thomson Publishing, 1996.
19. А. А. Марков, О представлении рекурсивных функций. М.: Изв. АН СССР. Сер. матем., 1949, том 13, выпуск 5, 417– 424.
20. Алан Купер. Алан Купер об интерфейсе. Основы проектирования взаимодействия/ А. Купер, Р. Рейманн, Д. Кронин – М.: Издательство «Символ-плюс», 2009, – 688 с.