

**ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**

Факультет компьютерных наук
Департамент программной инженерии

СОГЛАСОВАНО

Доцент департамента математики
факультета экономических наук, кандидат
физико-математических наук

_____ Е. Р. Горяинова
« ____ » _____ 2025 г.

УТВЕРЖДАЮ

Академический руководитель
образовательной программы
«Программная инженерия»
профессор департамента программной
инженерии, канд. техн. наук

_____ Н. А. Павлочев
« ____ » _____ 2025 г.

**Разработка программного комплекса для исследования
влияния аномальных наблюдений на точность
прогнозирования в регрессионных моделях**

Техническое задание

ЛИСТ УТВЕРЖДЕНИЯ

RU.17701729.11.04-01 ТЗ 01-1

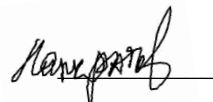
Подп. и дата	
Инв. № дубл.	
Взам. инв. №	
Подп. и дата	
Инв. № подл	

Исполнители:

Студент БПИ-221

/ Панкратов С. Ю. /

«12» мая 2025 г



Москва 2025

УТВЕРЖДЕН

RU.17701729.05.05-01 ТЗ 01-1

**Разработка программного комплекса для исследования
влияния аномальных наблюдений на точность
прогнозирования в регрессионных моделях**

Техническое задание

RU.17701729.11.04-01 ТЗ 01-1

Листов 28

<i>Подп. и дата</i>	
<i>Инв. № дубл.</i>	
<i>Взам. инв. №</i>	
<i>Подп. и дата</i>	
<i>Инв. № подл</i>	

Москва 2025

АННОТАЦИЯ

Техническое задание – основной документ, оговаривающий набор требований и порядок создания программного продукта, в соответствии с которым производится разработка программы, ее тестирование и приемка.

Настоящее Техническое задание на разработку программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях содержит следующие разделы: «Введение», «Основание для разработки», «Назначение разработки», «Требования к программе», «Требования к программным документам», «Технико-экономические показатели», «Стадии и этапы разработки», «Порядок контроля и приемки» и приложения.

В разделе «Введение» указано наименование и краткая характеристика области применения программного комплекса.

В разделе «Основания для разработки» указаны документы, на основании которых ведется разработка, а также наименование темы разработки.

В разделе «Назначение разработки» указано функциональное и эксплуатационное назначение программного продукта.

Раздел «Требования к программе» содержит основные требования к функциональным характеристикам, надежности, условиям эксплуатации, составу и параметрам технических средств, информационной и программной совместимости, маркировке и упаковке, транспортированию и хранению.

Раздел «Требования к программным документам» содержит предварительный состав программной документации и специальные требования к ней.

Раздел «Технико-экономические показатели» описывает ориентировочную экономическую эффективность, предполагаемую годовую потребность, а также экономические преимущества разработки по сравнению с аналогами.

Раздел «Стадии и этапы разработки» содержит стадии и этапы разработки, их содержание и сроки, а также указывает лица, ответственные за их выполнение.

В разделе «Порядок контроля и приемки» указаны общие требования к приемке работы, а также зафиксированы все допустимые при этом виды испытаний.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Настоящий документ разработан в соответствии с требованиями:

1. ГОСТ 19.101-77^[1]: Виды программ и программных документов.
2. ГОСТ 19.102-77^[2]: Стадии разработки.
3. ГОСТ 19.103-77^[3]: Обозначения программ и программных документов.
4. ГОСТ 19.104-78^[4]: Основные надписи.
5. ГОСТ 19.105-78^[5]: Общие требования к программным документам.
6. ГОСТ 19.106-78^[6]: Требования к программным документам, выполненным печатным способом.
7. ГОСТ 19.201-78^[7]: Техническое задание. Требования к содержанию и оформлению.
8. ГОСТ 19.602-78^[8]: Правила дублирования, учета и хранения программных документов, выполненных печатным способом.

Изменения к настоящему техническому заданию должны быть оформлены согласно ГОСТ 19.603-78^[9] и ГОСТ 19.604-78^[10].

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

СОДЕРЖАНИЕ

АННОТАЦИЯ	2
СОДЕРЖАНИЕ.....	4
1. ВВЕДЕНИЕ	5
1.2. Наименование программы	5
1.3. Краткая характеристика области применения.....	5
2. ОСНОВАНИЯ ДЛЯ РАЗРАБОТКИ	6
2.1. Документ(ы), на основании которого(ых) ведётся разработка.....	6
2.2. Наименование темы разработки	6
3. НАЗНАЧЕНИЕ РАЗРАБОТКИ.....	7
3.1 Функциональное назначение	7
3.2 Эксплуатационное назначение	7
4. ТРЕБОВАНИЯ К ПРОГРАММЕ.....	8
4.1. Требования к функциональным характеристикам	8
4.1.2. Требования к организации выходных данных	11
4.1.3. Требования к интерфейсу	12
4.2. Требования к надежности.....	13
4.3. Условия эксплуатации	14
4.4. Требования к составу и параметрам технических средств.....	15
4.5. Требования к информационной и программной совместимости	16
4.6. Требования к маркировке и упаковке	16
4.7. Требования к транспортировке	16
5. ТРЕБОВАНИЯ К ПРОГРАММНОЙ ДОКУМЕНТАЦИИ.....	17
5.1. Предварительный состав программной документации	17
5.2. Специальные требования к программной документации	17
6. ТЕХНИКО-ЭКОНОМИЧЕСКИЕ ПОКАЗАТЕЛИ.....	18
6.1 Ориентировочная экономическая эффективность.....	18
6.2 Предполагаемая потребность	18
6.3 Экономические преимущества разработки по сравнению с отечественными и зарубежными образцами или аналогами	18
7. СТАДИИ И ЭТАПЫ РАЗРАБОТКИ.....	20
7.1. Стадии разработки, этапы и содержание работ.....	20
7.2. Сроки разработки и исполнители.....	22
8. ПОРЯДОК КОНТРОЛЯ И ПРИЁМКИ	23
СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ.	24
ТЕРМИНЫ И СОКРАЩЕНИЯ	27
ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ.....	28

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

1. ВВЕДЕНИЕ

1.2. Наименование программы

Наименование темы разработки: «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях».

Наименование темы разработки на английском языке: «Development of a Software Package to Study the Influence of Outliers on the Prediction Accuracy in Regression Models».

Краткое наименование – «MSnOutliers».

1.3. Краткая характеристика области применения

«MSnOutliers» – приложение для исследования качества различных статистических методов на выборках данных с большим числом зашумленных (т. е. содержащих в себе помимо полезной нагрузки некоторый шум известного распределения) данных.

Комплекс также интегрирует различные алгоритмы машинного обучения для обнаружения и устранения аномальных наблюдений, что позволяет исследовать их эффективность в повышении точности прогнозирования регрессионных моделей.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

2. ОСНОВАНИЯ ДЛЯ РАЗРАБОТКИ

2.1. Документ(ы), на основании которого(ых) ведётся разработка

Основанием для разработки является учебный план подготовки бакалавров по направлению 09.03.04 «Программная инженерия» и утвержденная академическим руководителем тема курсового проекта.

2.2. Наименование темы разработки

Наименование темы разработки: «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях».

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

3. НАЗНАЧЕНИЕ РАЗРАБОТКИ

3.1 Функциональное назначение

Программный инструмент «MSnOutliers» предназначен для анализа и исследования робастности различных методов регрессионного анализа в условиях присутствия в данных аномальных наблюдений. Приложение предоставляет следующие возможности:

- 1) Моделирование набора данных с контролируемыми параметрами шума различных распределений.
- 2) Оценка эффективности различных статистических методов регрессии при наличии аномальных наблюдений разного количества и характера.
- 3) Применение алгоритмов машинного обучения для обнаружения аномальных наблюдений, последующее удаление аномальных наблюдений и оценка качества детектирования аномальных наблюдений.
- 4) Расчет метрик качества регрессионных моделей при их применении на очищенных моделью машинного обучения данных.
- 5) Итоговая визуализация полученных результатов в виде графиков зависимости ошибки от уровня шума.

3.2 Эксплуатационное назначение

Приложение «MsnOutliers» позволяет отслеживать влияние выбросов в данных на качество работы регрессионных методов. Для анализа могут использоваться данные, подготовленные пользователем либо же сгенерированные на стороне приложения.

Целевая аудитория – школьники и студенты, проходящие подготовку по дисциплине «Математическая статистика» или каким-либо смежным с ней, а также преподаватели, читающие вышеупомянутые курсы.

Программный комплекс может использоваться на ПК с операционной системой Windows, Linux или MacOS. Предполагается, что основная часть комплекса будет написана на языке C++ с возможным использованием модулей на языке Python для визуализации разработанных в рамках проекта средств борьбы с влиянием выбросов и методов регрессии. Приложение «MsnOutliers» позволяет отслеживать влияние выбросов в данных на качество работы регрессионных методов. Для анализа могут использоваться данные, подготовленные пользователем либо же сгенерированные на стороне приложения.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

4. ТРЕБОВАНИЯ К ПРОГРАММЕ

4.1. Требования к функциональным характеристикам

4.1.1. Состав выполняемых функций

4.1.1.1 Запуск анализа и моделирования шума

4.1.1.1.1 Функциональность запуска анализа данных

- 1) Валидация конфигурации моделей из JSON-файла, состоящей из проверки корректности указанных типов моделей, параметров шума и методов обнаружения аномальных наблюдений.
- 2) Многопоточное выполнение экспериментов с использованием заданных конфигураций. Для каждого уровня шума и для каждой конфигурации модели выполняется серия испытаний.
- 3) Добавление контролируемого шума в данные в соответствии с выбранным распределением, применение выбранного метода машинного обучения для обнаружения, фильтрация аномальных наблюдений и построение регрессионной модели на очищенных данных.
- 4) Агрегирование результатов экспериментов путем усреднения полученных метрик для одинаковых уровней шума и формирование единого набора данных для визуализации.
- 5) Группировка моделей для удобного отображения на графиках, с созданием до 5 моделей на одном графике для обеспечения читаемости и наглядности результатов.

4.1.1.1.2 Моделирование шума с использованием различных распределений

Доступны различные виды генерации шума, настройка параметров которых задается через пользовательский интерфейс. Непосредственный процесс генерации должен состоять из следующих этапов:

- 1) Выбирается количество наблюдений, у которых будет зашумлена целевая переменная.
- 2) Выбирается случайное подмножество наблюдений для внесения шума.
- 3) Генерация шумовых значений в соответствии с выбранным распределением и его параметрами.
- 4) Добавление сгенерированных шумовых значений к целевой переменной в выбранных наблюдениях.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Доступны следующие типы распределений:

- 1) Нормальное распределение для моделирования случайных отклонений умеренной величины. Первый параметр распределения соответствует среднему значению, второй - стандартному отклонению.
- 2) Распределение Стьюдента для моделирования данных с “тяжелыми хвостами”, характеризующихся более частым появлением экстремальных значений. Первый параметр распределения соответствует числу степеней свободы, определяющему “тяжесть хвостов”
- 3) Распределение Коши для моделирования сильно выраженных выбросов без определенного математического ожидания. Первый параметр распределения соответствует параметру расположения, второй - параметру масштаба.
- 4) Распределение Лапласа для моделирования двусторонних экспоненциальных выбросов. Первый параметр распределения соответствует параметру расположения, второй - параметру масштаба.

Также доступна возможность зашумления целевой переменной в виде её увеличения в k раз, где k - константа.

4.1.1.2 Методы обнаружения аномальных наблюдений

4.1.1.2.1 Методы на основе плотности распределения

4.1.1.2.1.1 Метод оценки плотности с использованием ядерных функций KDE

Ядерная оценка плотности для каждого наблюдения в наборе данных вычисляется с использованием радиально-базисных гауссовых функций.

Параметр сглаживания гамма по умолчанию определяется как обратно пропорциональный размерности входных данных. Пороговое значение плотности задается средним плотностей всех наблюдений в наборе данных.

Наблюдения классифицируются аномальными на основе сравнения их плотности вероятности с пороговым значением и исключаются из набора данных.

4.1.1.2.1.2 Метод кластеризации на основе плотности DBSCAN

Радиус окрестности r по умолчанию равен 0.05.

Минимальный размер кластера `minimumClusterSize` по умолчанию определяется как натуральный логарифм от количества наблюдений в наборе данных.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Кластеры расширяются с использованием алгоритма поиска в ширину, близость объектов определяется евклидовой метрикой.

Наблюдения, не вошедшие ни в один кластер, классифицируются как аномальные и исключаются из набора данных.

4.1.1.2.2 Метод k-ближайших соседей на основе расстояний

Расстояние от каждого наблюдения до его k-го ближайшего соседа вычисляется с использованием евклидовой метрики.

Устанавливаются параметры k (по умолчанию равен 10), определяющий количество ближайших соседей для рассмотрения и contamination (по умолчанию равен 0.15), определяющий ожидаемую долю аномальных наблюдений в наборе данных.

Сортировка наблюдений по расстоянию до их k-го ближайшего соседа и определение порогового значения, отделяющего нормальные наблюдения от аномальных.

Наблюдения с расстоянием до k-го ближайшего соседа выше порогового значения определяются аномальными и исключаются из набора данных.

4.1.1.2.3 Метод изолирующего леса на основе случайных разделений пространства признаков

Строится ансамбль деревьев (по умолчанию количество деревьев равно 100) изоляции и поддерживается ограничение глубины деревьев (по умолчанию параметр равен 12).

При построении каждого дерева признаки и точки разделения выбираются случайно.

Оценка аномальности наблюдения считается как нормализованное среднее глубины его изоляции во всех деревьях ансамбля. Наиболее аномальные наблюдения исключаются из набора данных.

4.1.1.3 Визуализация и сохранения результатов

4.1.1.3.1.1 Графическая визуализация результатов анализа

Предоставляется информативное графическое представление результатов анализа в виде двумерных графиков с установкой заголовков графиков с указанием используемого метода регрессионного анализа, подписью осей координат, параметров отображения точек и легенды с описанием каждой серии данных.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Заголовки формируются с указанием номера группы (в группе до 5 различных графиков) при наличии множества кривых.

В пределах одного графика используются различные цвета для серий данных. Доступны многообразные графические примитивы как точечные диаграммы для отображения зависимости ошибки от уровня шума, линейные графики, вертикальные, горизонтальные линии и функциональные зависимости.

Процесс построения графика выполняется средствами взаимодействия с внешним Python-интерпретатором и библиотекой Python matplotlib.

4.1.1.3.1.2 Расчет и сохранение метрик качества регрессии и обнаружения аномальных наблюдений

Рассчитываются следующие метрики качества:

- 1) Средняя абсолютная ошибка
- 2) Среднеквадратическая ошибка
- 3) Корень из среднеквадратической ошибки
- 4) Средняя абсолютная процентная ошибка
- 5) Симметричная средняя абсолютная процентная ошибка

Также рассчитываются следующие метрики качества детекции:

- 1) Точность: отношение количества правильно идентифицированных аномальных наблюдений к общему количеству объектов, отмеченных алгоритмом как аномальные
- 2) Полнота: отношение количества правильно идентифицированных аномальных наблюдений к общему количеству фактических аномальных наблюдений
- 3) F1-мера: гармоническое среднее между точностью и полнотой

Значения метрик усредняются для каждого уровня шума.

Результаты сохраняются в иерархическую структуру с группировкой по моделям регрессии, методам обнаружения аномальных наблюдений и уровням шума.

Рассчитанные метрики с указанием количества проведенных экспериментов для каждого уровня шума сохраняются с соблюдением структуры данных, описанной в Приложении 2 и п. 4.1.2.2.

4.1.2. Требования к организации выходных данных

4.1.2.1. Форматы выходных файлов

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Используется стандартный формат JSON для хранения метрик методов для платформенной независимости и удобства последующего анализа.

Для графической визуализации результатов используется растровый формат PNG с поддержкой прозрачности для графической визуализации результатов, разрешение изображений приблизительно 1200x720 пикселей, в заголовке PNG поддерживаются метаданные статистического анализа.

4.1.2.2. Структура выходных данных

4.1.2.2.1 Метрики качества (JSON-формат)

Корневой объект содержит массив “results”, включающий результаты всех проведенных экспериментов, каждый элемент массива “results” представляет собой объект с метаданными о конкретном эксперименте и его результатах, массив “metrics_by_noise_level” содержит результаты для каждого уровня шума.

4.1.2.2.2 Графики с результатами методов регрессии (PNG-формат)

Графики содержат информацию о типе регрессионной модели, группе методов обнаружения аномальных наблюдений, представленных на графике. Область графика включает ось абсцисс с подписью “Noise Level” и ось ординат с подписью “Error” и до 5 кривых различных цветов, представляющих комбинации методов обнаружения аномальных наблюдений и типов шума. Легенда содержит цветовые индикаторы для каждой кривой и текстовые описания в формате “метод обнаружения аномальных наблюдения / тип шума”.

4.1.2.3. Правила именования выходных файлов

4.1.2.3.1 Метрики качества

Метрики качества регрессии и обнаружения аномальных наблюдений сохраняется в единственный файл metrics.json.

Все файлы сохраняются в директории, указанной в конфигурационном файле.

4.1.2.3.2 Визуализации результатов

Формат: out_<название статистического метода>_<номер группы>.png

Префикс “out_” для всех файлов визуализации, при группировке нескольких методов на одном графике используется суффикс “group” с указанием номера группы, начиная с 0.

Все файлы сохраняются в директории, указанной в конфигурационном файле.

4.1.3. Требования к интерфейсу

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

4.1.3.1. Требования к интерфейсу визуализации результатов

Интерфейс визуализации результатов должен обеспечивать наглядное представление зависимости ошибки регрессионных моделей от уровня шума при использовании различных методов обнаружения аномальных наблюдений. После запуска анализа результаты должны отображаться в отдельном окне с возможностью вертикальной прокрутки. Каждый график должен быть снабжен информативным заголовком, включающим название метода регрессии и номер группы, размещенным в верхней части графика. Под заголовком должен располагаться график с четко обозначенными осями: по оси абсцисс уровень шума, по оси ординат величина ошибки в виде среднеквадратической ошибки. На каждом графике должны отображаться до 5 различных кривых, представляющих различные комбинации методов обнаружения аномальных и типов распределения шума, каждая кривая должна иметь уникальный цвет и сопровождаться легендой, указывающей метод обнаружения аномальных наблюдений и тип распределения шума.

4.1.3.2. Требования к интерфейсу анализа и сохранения

Интерфейс анализа и сохранения результатов должен обеспечивать полностью автоматизированный процесс выполнения экспериментов, агрегации результатов и формирования выходных файлов. При нажатии кнопки “Run on models” должна запускаться последовательность операций, включающая валидацию конфигурационного файла моделей, многопоточное выполнение экспериментов с различными уровнями шума для всех указанных конфигураций моделей, сохранение полученных результатов в структурированном формате JSON, генерацию и сохранение графиков визуализации в формате PNG. Для каждой модели регрессии и каждого метода обнаружения аномальных наблюдений должны сохраняться метрики для всех уровней шума с указанием усредненных значений по всем проведенным экспериментам и количества экспериментов, а файлы результатов должны быть структурированы в соответствии с требованиями к организации выходных данных.

4.2. Требования к надежности

Обработка данных и обнаружение аномальных наблюдений должны обеспечивать корректную работу с наборами данных различного объема и структуры. При обработке данных должны корректно учитываться возможные численные

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

особенности, такие как деление на ноль или расчет расстояний в многомерном пространстве, с применением соответствующих механизмов предотвращения ошибок.

Система должна обеспечивать точный расчет всех предусмотренных метрик качества регрессии. Агрегация результатов множественных экспериментов должна выполняться корректно, с надлежащим учетом статистических характеристик и без потери данных. Многопоточное выполнение экспериментов должно быть организовано с учетом возможных взаимных блокировок, с корректной синхронизацией доступа к общим ресурсам. Система должна обеспечивать корректную работу с различными распределениями шума.

Визуализация должна корректно отображать графики зависимости ошибки от уровня шума для любых комбинаций моделей регрессии и методов обнаружения аномальных наблюдений, включая корректное масштабирование осей и отображение легенд. Система должна обеспечивать надежное сохранение результатов анализа в структурированном формате, с корректным форматированием и без потери данных. Визуальные элементы пользовательского интерфейса должны корректно реагировать на взаимодействие с пользователем, без некорректного отображения при изменении размеров окна или прокрутке контента.

Методы должны обеспечивать корректную обработку исключительных ситуаций, включая ошибки формата файлов, отсутствие необходимых данных или файлов, и вычислительные ошибки. При обнаружении ошибок пользователю должны предоставляться информативные сообщения с указанием характера проблемы и возможных способов ее устранения. Система должна сохранять свою работоспособность после возникновения ошибок, без необходимости перезапуска приложения. Все критические операции должны выполняться с проверкой условий и валидацией входных данных для предотвращения непредвиденного поведения программы.

4.3. Условия эксплуатации

Требования к условиям эксплуатации программного продукта совпадают с требованиями эксплуатации устройства, используемого для работы с программным обеспечением. Программа предназначена для работы на персональных компьютерах и серверах, соответствующих минимальным системным требованиям, указанным в документации. Специальных требований к условиям эксплуатации приложения, таких как температура, влажность или другие физические параметры, не предъявляется.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

4.4. Требования к составу и параметрам технических средств

4.4.1. Требования к клиентскому оборудованию

Рекомендуемые требования к клиентскому оборудованию для корректной работы приложения:

1. Процессор с 4 или более ядрами (AMD Ryzen 5 / Intel Core i5 или другие);
2. 8 ГБ оперативной памяти или больше;
3. 2 ГБ видеопамати или больше;
4. Монитор разрешением не менее 1920x1080.
5. Наличие SSD-диска с не менее 8 ГБ свободного пространства.

Минимальные требования к клиентскому оборудованию для работы приложения:

1. Процессор Intel Core i3-5100f;
2. 4 ГБ оперативной памяти;
3. 2 ГБ видеопамати;
4. Монитор разрешением 1280x720.
5. Жесткий диск с не менее 5 ГБ свободного пространства.

Общие требования к клиентскому оборудованию для работы приложения:

1. Мышь или совместное указывающее устройство;
2. Клавиатура;
3. ОС: Windows (10 и выше), Linux (Ubuntu 20.04 и выше) или macOS версии не ниже 11 “Big Sur”;

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

4.5. Требования к информационной и программной совместимости

Приложение должно быть написано в соответствии со стандартом языка C++17. Для сборки проекта используется система сборки CMake версии не ниже 3.14. Должен использоваться компилятор GCC версии не ниже 10.5 или Clang версии не ниже 13.0.0. Для управления версиями приложения используется система контроля версий Git. Версия языка Python должна быть не ниже 3.9.

4.6. Требования к маркировке и упаковке

Программа распространяется в виде электронного пакета, содержащего программную документацию, приложение (исполняемые файлы и прочие необходимые для работы файлы).

4.7. Требования к транспортировке

Особых требований к транспортировке и хранению не предъявляются.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

5. ТРЕБОВАНИЯ К ПРОГРАММНОЙ ДОКУМЕНТАЦИИ

5.1. Предварительный состав программной документации

1. «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях». Техническое задание (ГОСТ 19.201-78^[7]).
2. «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях». Пояснительная записка (ГОСТ 19.404-79^[11]).
3. «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях». Программа и методика испытаний (ГОСТ 19.301-79^[12]).
4. «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях». Текст программы (ГОСТ 19.401-78^[13]).
5. «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях». Руководство оператора (ГОСТ 19.505-79^[14]).
6. «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях». Руководство программиста (ГОСТ 19.504-79^[15]).

5.2. Специальные требования к программной документации

Программная документация должна быть выполнена в соответствии с ГОСТ 19.106-78[6] и ГОСТами к каждому виду документа.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

6. ТЕХНИКО-ЭКОНОМИЧЕСКИЕ ПОКАЗАТЕЛИ

6.1 Ориентировочная экономическая эффективность

В рамках проекта расчет экономической эффективности программного продукта не производился.

6.2 Предполагаемая потребность

Регрессионные модели машинного обучения широко используются в разных областях бизнеса и науки. Данные, на которых эти модели обучаются не всегда свободны от выбросов, что, в свою очередь, создает опасность ухудшения качества предсказаний и снижения эффективности использования регрессионных моделей. Разрабатываемый программный комплекс может использоваться как для оценки просадок качества при обучении на богатых выбросами выборках, так и средства борьбы с выбросами (если нельзя выбросить и/или заменить аномальные наблюдения из выборки, мы должны обеспечить возможность обеспечить устойчивость модели к этим аномальным наблюдениям).

6.3 Экономические преимущества разработки по сравнению с отечественными и зарубежными образцами или аналогами

Существующие решения для анализа методов регрессии и влияния аномальных наблюдений на них такие как ELKI, auditor, outliers или scikit-learn обладают рядом ограничений и недостатков для поставленной задачи:

- 1) Данные инструменты обычно сфокусированы на решение только одной из задач обнаружения аномальных наблюдений и построения методов регрессии.
- 2) Не моделируются различные виды систематических и случайных аномальных наблюдений.
- 3) У большинства методов отсутствуют подробная визуализация и/или подробные отчеты по каждому запущенному методу.
- 4) Не поддерживают многопоточные эксперименты с выбранными различными параметрами и гиперпараметрами.
- 5) Не интегрированы в единое решение.

Разработанный инструмент «MSnOutliers» предоставляет следующие возможности, демонстрирующие экономические преимущества разработки:

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

- 1) Комплексный подход к программе от генерации данных с шумом до метрик качества.
- 2) Возможность создания шума различного типа.
- 3) Эффективная многопоточная архитектура для запуска параллельных экспериментов.
- 4) Все метрики качества обнаружения аномальных наблюдений и методов регрессии, а также результаты визуализации методов укомплектованы в одной директории для удобного использования.

Таким образом, «MSnOutliers» дает более глубокое понимание влияния аномальных наблюдений на методы регрессии и принимать более обоснованные решения при выборе методов работы с реальными данными.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

7. СТАДИИ И ЭТАПЫ РАЗРАБОТКИ

7.1. Стадии разработки, этапы и содержание работ

Стадии и этапы разработки были выявлены с учётом ГОСТ 19.102-77^[2].

Стадия разработки	Этап работ	Содержание работ	Сроки выполнения
Техническое задание	Обоснование необходимости разработки программы	Постановка задачи Сбор исходных материалов. Выбор и обоснование критериев эффективности и качества разрабатываемой программы.	04.11.2024
	Научно-исследовательские работы	Обоснование возможности решения поставленной задачи. Предварительный выбор методов решения задач. Определение требований к техническим и программным средствам	21.11.2024

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

	Разработка и утверждение технического задания	<p>Определение требований к программе.</p> <p>Определение стадий, этапов и сроков разработки программы и документации на неё.</p> <p>Согласование и утверждение технического задания.</p> <p>Загрузка согласованного технического задания в SmartLMS</p>	04.12.2024
Рабочий проект	Разработка программы	<p>Предварительная разработка структуры программы</p> <p>Разработка учебных материалов</p>	27.02.2025
	Разработка программной документации	Разработка программных документов в соответствии с требованиями ГОСТ 19.101-77 ^[1] .	03.03.2025

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

Внедрение	Подготовка и передача программы	Подготовка программы и программной документации для презентации и защиты Представление разработанного программного продукта научному руководителю и получение отзыва Представление разработанного программного продукта научному руководителю и получение отзыва. Загрузка Пояснительной записки в систему Антиплагиат через ЛМС НИУ ВШЭ. Загрузка материалов курсового проекта в ЛМС Защита программного продукта комиссии.	15.03.2025
-----------	---------------------------------	---	------------

7.2. Сроки разработки и исполнители

Разработка программного продукта должна быть завершена к XX.XX.2025. (дата защиты курсовой работы). Исполнитель – Панкратов Степан, студент группы БПИ221 факультета компьютерных наук НИУ ВШЭ

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

8. ПОРЯДОК КОНТРОЛЯ И ПРИЁМКИ

Контроль и приемка разработки осуществляются в соответствии с документом «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях». Программа и методика испытаний» и пунктом 5.2 технического задания.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

ПРИЛОЖЕНИЕ 1

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ.

1. ГОСТ 19.101-77: Виды программ и программных документов. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
2. ГОСТ 19.102-77: Стадии разработки. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
3. ГОСТ 19.103-77: Обозначения программ и программных документов. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
4. ГОСТ 19.104-78: Основные надписи. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
5. ГОСТ 19.105-78: Общие требования к программным документам. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
6. ГОСТ 19.106-78: Требования к программным документам, выполненным печатным способом. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
7. ГОСТ 19.201-78: Техническое задание. Требования к содержанию и оформлению. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
8. ГОСТ 19.602-78: Правила дублирования, учета и хранения программных документов, выполненных печатным способом. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
9. ГОСТ 19.603-78: Общие правила внесения изменений. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
10. ГОСТ 19.604-78: Правила внесения изменений в программные документы, выполненные печатным способом. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
11. ГОСТ 19.404-79: Пояснительная записка. Требования к содержанию и оформлению. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
12. ГОСТ 19.301-79: Программа и методика испытаний. Требования к содержанию и оформлению. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

13. ГОСТ 19.401-78: Текст программы. Требования к содержанию и оформлению. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
14. ГОСТ 19.505-79: Руководство оператора. Требования к содержанию и оформлению. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
15. ГОСТ 19.504-79: Руководство программиста. Требования к содержанию и оформлению. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

ПРИЛОЖЕНИЕ 2

```
{
  "results": [
    {
      "data_path": "...",
      "max_noise": 1,
      "metrics_by_noise_level": [
        {
          "detection_metrics": {
            "f1_score": ...,
            "precision": ...,
            "recall": ...
          },
          "metrics": {
            "MAE": ...,
            "MAPE": ...,
            "MSE": ...,
            "RMSE": ...,
            "SMAPE": ...
          },
          "noise_level": 0
        },
        ...
      ],
      "min_noise": ...,
      "mlmodel": "...",
      "mlmodel_params": {
        "param1": ...,
        "param2": ...
      },
      "model": "LSM",
      "noise_params": {
        "param1": ...,
        "param2": ...
      },
      "noise_type": "...",
      "num_experiments": ...,
      "num_features": ...
    }
  ]
}
```

Изм.	Лист	№ докум.	Подп.	Дата
RU.17701729.05.10-01 ТЗ 01-1				
Инв. № подл.	Подп. и дата	Взам. инв. №	Инв. № дубл.	Подп. и дата

ТЕРМИНЫ И СОКРАЩЕНИЯ

1. Регрессионная модель – математический метод прогнозирования, устанавливающий зависимость между целевой переменной и одним или несколькими признаками.
2. Аномальные наблюдения – точки данных, которые значительно отклоняются от остальных наблюдений в наборе данных и могут негативно влиять на точность прогнозирования.
3. JSON – легкий формат обмена данными, используемый для хранения конфигураций моделей и параметров экспериментов.
4. Асинхронные вычисления - метод параллельного выполнения задач для повышения производительности, особенно при проведении множества экспериментов.
5. Целевая переменная - поле в наборе данных, значение которого модель стремится предсказать на основе признаков.
6. Признак - поле в наборе данных, которое используется как для предсказания величины-цели.
7. Метрики качества - показатели, используемые для оценки точности регрессионных моделей или обнаружения аномальных наблюдений.
8. Формат CSV – текстовый формат для представления табличных данных, где значения разделены специальным символом.
9. Шум – случайные отклонения в данных, которые не отражают истинную закономерность и могут возникать из-за ошибок измерения.
10. Хвост распределения – область распределения вероятностей, удаленная от его центральной части.
11. Гиперпараметр – параметр алгоритма машинного обучения, который устанавливается перед началом обучения и не изменяется в процессе его обучения.

[illegible]