## ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ

**ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ**

## «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук Образовательная программа «Программная инженерия»

СОГЛАСОВАНО

доцент департамента математики

факультета экономических наук,

кандидат физико-математических наук

Е.Р. Горяинова

УТВЕРЖДАЮ

Академический руководитель образовательной программы

«Программная инженерия» старший преподаватель департамента

программной инженерии Н.А. Павлочев

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

|  |  |
| --- | --- |
| ***Подп. и дата*** |  |
| ***Инв. № дубл.*** |  |
| ***Взам. инв. №*** |  |
| ***Подп. и дата*** |  |
| ***Инв. № подл*** |  |

**Техническое задание ЛИСТ УТВЕРЖДЕНИЯ**

# RU.17701729.05.05 -01 ТЗ 01-1-ЛУ

Исполнители студент группы БПИ221

/ Знатнов Е.П./

студент группы БПИ221

/ Панкратов С.Ю./

« » 2024 г.

УТВЕРЖДЕН

RU.17701729.05.05-01 ТЗ 01-1-ЛУ

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

**Техническое задание RU.17701729.05.05 -01 ТЗ 01-1**

|  |  |
| --- | --- |
| ***Подп. и дата*** |  |
| ***Инв. № дубл.*** |  |
| ***Взам. инв. №*** |  |
| ***Подп. и дата*** |  |
| ***Инв. № подл*** |  |

# Листов 26

# 

## 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.

Настоящий документ разработан в соответствии с требованиями:

1. ГОСТ 19.101-77[[1]](#_bookmark29): Виды программ и программных документов.
2. ГОСТ 19.102-77[[2]](#_bookmark30): Стадии разработки.
3. ГОСТ 19.103-77[[3]](#_bookmark31): Обозначения программ и программных документов.
4. ГОСТ 19.104-78[[4]](#_bookmark32): Основные надписи.
5. ГОСТ 19.105-78[[5]](#_bookmark33): Общие требования к программным документам.
6. ГОСТ 19.106-78[[6]](#_bookmark34): Требования к программным документам, выполненным печатным способом.
7. ГОСТ 19.201-78[[7]](#_bookmark35): Техническое задание. Требования к содержанию и оформлению.
8. ГОСТ 19.602-78[[8]](#_bookmark36): Правила дублирования, учета и хранения программных документов, выполненных печатным способом.

Изменения к настоящему техническому заданию должны быть оформлены согласно ГОСТ 19.603-78[[9]](#_bookmark37)и ГОСТ 19.604-78[[10]](#_bookmark38).

# Содержание

[1. ВВЕДЕНИЕ 5](#_Toc158543924)

[1.1. Наименование программы 5](#_Toc158543925)

[1.2. Краткая характеристика области применения 5](#_Toc158543926)

[2. ОСНОВАНИЯ ДЛЯ РАЗРАБОТКИ 6](#_Toc158543927)

[2.1. Документ(ы), на основании которого(ых) ведётся разработка 6](#_Toc158543928)

[2.2 Наименование темы разработки 6](#_Toc158543929)

[3. НАЗНАЧЕНИЕ РАЗРАБОТКИ 7](#_Toc158543930)

[3.1. Функциональное назначение 7](#_Toc158543931)

[3.2. Эксплуатационное назначение 7](#_Toc158543932)

[4. ТРЕБОВАНИЯ К ПРОГРАММЕ 8](#_Toc158543933)

[4.1. Требования к функциональным характеристикам 8](#_Toc158543934)

[4.1.2. Требования к организации входных данных 11](#_Toc158543935)

[4.1.3 Требования к организации выходных данных 11](#_Toc158543936)

[4.1.4. Требования к интерфейсу 11](#_Toc158543937)

[4.2. Требования к надежности 13](#_Toc158543938)

[4.3. Условия эксплуатации 14](#_Toc158543939)

[4.4. Требования к составу и параметрам технических средств 14](#_Toc158543940)

[4.5. Требования к информационной и программной совместимости 15](#_Toc158543941)

[4.6. Требования к маркировке и упаковке 15](#_Toc158543942)

[4.7. Требования к транспортировке 15](#_Toc158543943)

[5. ТРЕБОВАНИЯ К ПРОГРАММНОЙ ДОКУМЕНТАЦИИ 16](#_Toc158543944)

[5.1. Предварительный состав программной документации 16](#_Toc158543945)

[5.2. Специальные требования к программной документации 16](#_Toc158543946)

[6. ТЕХНИКО-ЭКОНОМИЧЕСКИЕ ПОКАЗАТЕЛИ 17](#_Toc158543947)

[6.1 Ориентировочная экономическая эффективность 17](#_Toc158543948)

[6.2 Предполагаемая потребность 17](#_Toc158543949)

[6.3 Экономические преимущества разработки по сравнению с отечественными и зарубежными образцами или аналогами 17](#_Toc158543950)

[7. СТАДИИ И ЭТАПЫ РАЗРАБОТКИ 18](#_Toc158543951)

[7.1. Стадии разработки, этапы и содержание работ 18](#_Toc158543952)

[7.2. Сроки разработки и исполнители 20](#_Toc158543953)

[8. ПОРЯДОК КОНТРОЛЯ И ПРИЁМКИ 21](#_Toc158543954)

[СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ. 22](#_Toc158543955)

[ТЕРМИНЫ И СОКРАЩЕНИЯ 25](#_Toc158543956)

## 1. ВВЕДЕНИЕ

## 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".

## 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.

## 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".

## НАЗНАЧЕНИЕ РАЗРАБОТКИ

## Функциональное назначение

Приложение «MsnOutliers» предоставляет возможность изучить влияние аномальных наблюдений (данных, отличающихся по абсолютному значению на порядок от среднего по всей остальной выборке) на качество работы различных регрессионных методов математической статистики. Предполагается, что основное использование приложения будет происходить в образовательных целях.

## 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 С++ с возможным использованием модулей на языке Python для визуализации разработанных в рамках проекта средств борьбы с влиянием выбросов и методов регрессии.

## ТРЕБОВАНИЯ К ПРОГРАММЕ

## Требования к функциональным характеристикам

* + 1. Состав выполняемых функций
       1. Компонент запуска анализа и моделирования шума
          1. Функциональность запуска анализа данных

Проверка и валидация конфигурации моделей, загружаемой из JSON-файла, включая проверку корректности указанных типов моделей, параметров шума и методов обнаружения аномалий.

Многопоточное выполнение экспериментов с использованием заданных конфигураций. Для каждого уровня шума и для каждой конфигурации модели выполняется серия испытаний.

Процесса обработки данных, включающего добавление контролируемого шума в данные в соответствии с выбранным распределением, применение выбранного метода машинного обучения для обнаружения и фильтрации аномальных наблюдений и построение регрессионной модели на очищенных данных

Агрегирование результатов экспериментов путем усреднения полученных метрик для одинаковых уровней шума и формирование единого набора данных для визуализации.

Группировка моделей для удобного отображения на графиках, с созданием до 5 моделей на одном графике для обеспечения читаемости и наглядности результатов.

Автоматическое формирование информативных заголовков и легенд для графиков, включающих информацию о типе регрессионной модели и используемом методе обнаружения аномалий.

* + - * 1. Моделирование шума с использованием различных распределений

Поддержка четырех типов статистических распределений для генерации шума, настройка параметров которых задается через пользовательский интерфейс.

Нормальное распределение для моделирования случайных отклонений умеренной величины. Параметр 1 соответствует среднему значению, параметр 2 - стандартному отклонению.

Распределение Стьюдента для моделирования данных с "тяжелыми хвостами", характеризующихся более частым появлением экстремальных значений. Параметр 1 соответствует числу степеней свободы, определяющему "тяжесть хвостов"

Распределение Коши для моделирования сильно выраженных выбросов без определенного математического ожидания. Параметр 1 соответствует параметру расположения, параметр 2 - параметру масштаба

Распределение Лапласа для моделирования двусторонних экспоненциальных выбросов. Параметр 1 соответствует параметру расположения, параметр 2 - параметру масштаба.

Алгоритм добавления шума:

Случайное выбор подмножества наблюдений для внесения шума

Генерация шумовых значений в соответствии с выбранным распределением и его параметрами

Добавление сгенерированных шумовых значений к целевой переменной в выбранных наблюдениях

Поддержка контролируемого количества аномалий от 0 до 50 наблюдений, что позволяет исследовать влияние различного уровня зашумленности данных на качество регрессионных моделей.

Визуализация зависимости ошибки модели от уровня шума, что позволяет наглядно оценить устойчивость различных регрессионных методов к аномальным наблюдениям разной природы и интенсивности.

* + - 1. Компонент методов обнаружения аномалий
         1. Методы на основе плотности распределения

Метод оценки плотности с использованием ядерных функций KDE

Вычисление ядерной оценки плотности для каждого наблюдения в наборе данных с использованием радиально-базисных гауссовых функций.

Параметра сглаживания гамма по умолчанию определяется на основе размерности данных с возможностью его настройки через параметр, обратно пропорциональный размерности входных данных.

Пороговое значение плотности задается усреднением плотностей всех наблюдений в наборе данных.

Поддержка многомерных данных с произвольным количеством признаков, что позволяет анализировать сложные многомерные распределения.

Классификация наблюдений как нормальных или аномальных на основе сравнения их плотности вероятности с пороговым значением и исключение аномальных наблюдений из очищенного набора данных.

Метод кластеризации на основе плотности DBSCAN

Радиус окрестности eps по умолчанию с учетом размерности данных и их распределения определяется по формуле:

eps = k \* sqrt(размерность) \* sqrt(средний\_квадрат\_значений / количество\_наблюдений), где k - коэффициент масштабирования, по умолчанию равный 0.1.

Минимальный размер кластера minimumClusterSize по умолчанию определяется на основе логарифма от количества наблюдений в наборе данных.

Итеративное расширение кластеров с использованием алгоритма поиска в ширину для идентификации связных областей высокой плотности.

Поиск соседей с использованием евклидовой метрики для определения близости наблюдений.

Маркировка наблюдений, не вошедших ни в один кластер, как аномальных и их исключение из очищенного набора данных.

* + - * 1. Метод k-ближайших соседей на основе расстояний

Вычисление расстояния от каждого наблюдения до его k-го ближайшего соседа с использованием евклидовой метрики.

Параметр k (по умолчанию равный 10), определяющий количество ближайших соседей для рассмотрения.

Параметр contamination (по умолчанию равный 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.

Классификация наблюдений с расстоянием до k-го ближайшего соседа выше порогового значения как аномальных и их исключение из очищенного набора данных.

* + - * 1. Метод изолирующего леса на основе случайных разделений пространства признаков

Построение ансамбля деревьев изоляции (по умолчанию 100 деревьев).

Поддержка ограничения глубины деревьев (по умолчанию 12 уровней).

Случайный выбор признаков и точек разделения при построении каждого дерева для обеспечения разнообразия в ансамбле.

Оценка аномальности наблюдения на основе нормализованной средней глубины его изоляции во всех деревьях ансамбля.

Автоматическое выделение наиболее аномальных наблюдений (верхние 15% по шкале аномальности) и их исключение из очищенного набора данных.

* + - 1. Компонент визуализации и сохранения результатов

Графическая визуализация результатов анализа

Компонент включает разработанную подсистему визуализации, обеспечивающую информативное графическое представление результатов анализа. Основные функциональные возможности:

Автоматическое создание и настройка двумерных графиков с пользовательской конфигурацией:

Установка заголовков графиков с указанием используемого метода регрессионного анализа

Настройка подписей осей (X: "Noise Level", Y: "Error")

Настройка параметров отображения точек (размер, цвет, прозрачность)

Добавление легенды с описанием каждой серии данных

Интеллектуальная группировка результатов для обеспечения читаемости визуализации:

Автоматическое разделение результатов на группы по 5 кривых на одном графике

Формирование заголовков с указанием номера группы при наличии множества кривых

Использование различных цветов для серий данных в пределах одного графика для улучшения различимости

Реализация многообразных графических примитивов как точечные диаграммы для отображения зависимости ошибки от уровня шума, линейные графики, вертикальные, горизонтальные линии и функциональные зависимости.

Предоставление окна просмотра результатов с возможностью прокрутки при наличии множества графиков

Структурированная организация графиков с группировкой по методам регрессии

Интерактивные элементы для масштабирования и перемещения по графику

Интеграция с Python-скриптами для генерации высококачественных графиков:

Формирование JSON-конфигурации для графиков

Взаимодействие с внешним Python-интерпретатором

Автоматическое сохранение графиков в формате PNG с разрешением, обеспечивающим читаемость всех элементов

Расчет и сохранение метрик качества регрессии

Реализация пяти различных метрик качества регрессии для всесторонней оценки моделей:

Средняя абсолютная ошибка

Среднеквадратическая ошибка

Корень из среднеквадратической ошибки

Средняя абсолютная процентная ошибка

Симметричная средняя абсолютная процентная ошибка

Механизм агрегации результатов многократных экспериментов для получения статистически значимых оценок:

Вычисление средних значений метрик по всем проведенным экспериментам для каждого уровня шума

Регистрация количества экспериментов для каждого уровня шума для оценки статистической значимости

Структурированное сохранение результатов в формате JSON:

Организация результатов в иерархическую структуру с группировкой по моделям регрессии, методам обнаружения аномалий и уровням шума

Включение метаданных о параметрах экспериментов (тип распределения шума, параметры шума, путь к данным, количество признаков)

Сохранение всех рассчитанных метрик с указанием количества проведенных экспериментов для каждого уровня шума c соблюдением структуры данных, описанной в п. 4.1.2.2.

Расчет и сохранение метрик качества обнаружения аномальных наблюдений

Вычисление трех ключевых метрик качества классификации аномалий:

Точность (Precision): отношение количества правильно идентифицированных аномалий к общему количеству объектов, отмеченных алгоритмом как аномальные

Полнота (Recall): отношение количества правильно идентифицированных аномалий к общему количеству фактических аномалий

F1-мера: гармоническое среднее между точностью и полнотой

Агрегация метрик по результатам многократных экспериментов:

Вычисление средних значений метрик для каждого уровня шума

Учет специфики каждого метода обнаружения аномалий при интерпретации результатов

Анализ зависимости качества обнаружения аномалий от уровня шума и его типа

Интеграция с основным форматом сохранения результатов:

Добавление метрик качества обнаружения аномалий в структуру JSON-файла с результатами

Сохранение метрик для каждого метода обнаружения аномалий, типа распределения шума и уровня шума

Соблюдение структуры данных, описанной в Приложении 2 и п. 4.1.2.2, для обеспечения совместимости с аналитическими инструментами

## Требования к организации выходных данных

* + - 1. Форматы выходных файлов

Использование стандартного формата JSON для хранения метрик методов для платформенной независимости и удобства последующего анализа

Растровый формат PNG с поддержкой прозрачности для графической визуализации результатов, разрешение изображений 1200×720 пикселей, поддерживаются метаданные в заголовке PNG для хранения дополнительной информации о параметрах визуализации

* + - 1. Структура выходных данных
         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

* + - * 1. Графики с результатами методов регрессии (PNG-формат)

Графики содержат информацию о типе регрессионной модели, группе методов обнаружения аномалий, представленных на графике

Область графика включает координатные оси с подписями "Noise Level" ось абсцисс и "Error" ось ординат и до 5 кривых различных цветов, представляющих разные комбинации методов обнаружения аномалий и типов шума

Легенда расположена в нижней правой части графика и содержит цветовые индикаторы для каждой кривой и текстовые описания в формате "метод обнаружения аномалий / тип шума"

* + - 1. Правила именования выходных файлов
         1. Метрики качества

Формат имени: metrics\_{модель}\_{метод\_обнаружения}\_{тип\_шума}.json

Компоненты имени разделяются символом подчеркивания

Все сокращения моделей согласуются с их внутренними идентификаторами (LSM, HUB, TUK, LAD, THS)

Методы обнаружения аномалий указываются полным именем (IForest, DBSCAN, KDE, KNN, None)

Типы распределений указываются полным именем (Normal, Student, Cauchy, Laplace)

* + - * 1. Визуализации результатов:

Формат имени: out\_{модель}\_{группа}.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 config.h

## Требования к интерфейсу

* + - 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 аномалий. После запуска анализа результаты должны отображаться в отдельном окне с возможностью вертикальной прокрутки при наличии нескольких графиков. Каждый график должен быть снабжен информативным заголовком, включающим название метода регрессии и номер группы (при наличии нескольких графиков для одного метода), размещенным в верхней части графика. Под заголовком должен располагаться график размером 680×490 пикселей с четко обозначенными осями: по оси X - уровень шума (количество аномальных наблюдений), по оси Y - величина ошибки в виде среднеквадратической ошибки. На каждом графике должны отображаться до 5 различных кривых, представляющих различные комбинации методов обнаружения аномалий и типов распределения шума, каждая кривая должна иметь уникальный цвет и сопровождаться легендой, указывающей метод обнаружения аномалий и тип распределения шума. При отсутствии результатов для выбранных моделей должно отображаться соответствующее информационное сообщение.

* + - 1. Требования к интерфейсу анализа и сохранения результатов

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, расчет метрик качества регрессии и метрик качества обнаружения аномалий, сохранение полученных метрик в структурированном формате JSON, генерацию и сохранение графиков визуализации в формате PNG, и отображение результатов пользователю. Для каждой модели регрессии и каждого метода обнаружения аномалий 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.

## Требования к надежности

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

Компоненты должны обеспечивать детектирование и 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.

## 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.

## Требования к составу и параметрам технических средств

* + 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 видеопамяти 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 видеопамяти;
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";

## Требования к информационной и программной совместимости

Приложение должно быть написано в соответствии со стандартом языка С++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.

## Требования к маркировке и упаковке

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

## Требования к транспортировке

Особых требований к транспортировке и хранению не предъявляются.

## ТРЕБОВАНИЯ К ПРОГРАММНОЙ ДОКУМЕНТАЦИИ

## 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]](#_bookmark35)).
    2. «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях». Пояснительная записка (ГОСТ 19.404-79[[11]](#_bookmark39)).
    3. «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях». Программа и методика испытаний (ГОСТ 19.301-79[[12]](#_bookmark40)).
    4. «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях». Текст программы (ГОСТ 19.401-78[[13]](#_bookmark41)).
    5. «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях». Руководство оператора (ГОСТ 19.505-79[[14]](#_bookmark42)).
    6. «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях». Руководство программиста (ГОСТ 19.504-79[[15]](#_bookmark42)).

## Специальные требования к программной документации

Программная документация должна быть выполнена в соответствии с ГОСТ 19.106-78[[6]](#_bookmark34)и ГОСТами к каждому виду документа.

## ТЕХНИКО-ЭКОНОМИЧЕСКИЕ ПОКАЗАТЕЛИ

## Ориентировочная экономическая эффективность

В рамках проекта расчет экономической эффективности программного продукта не производился.

## 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).

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

Первое достоинство комплекса заключается в предоставлении средств для анализа влияния выбросов на качество моделей. Пользователь сможет легко проверить, как именно изменится качество его модели при включении части выбросов или их всех. Таким образом, пользователь получает механизм гибкой настройки модели и данных, позволяющий добиться именно того результата, к которому он, пользователь, идет. Второе достоинство заключается в предоставлении комплексом средств, позволяющих бороться с выбросами. Присутствуют как средства, разработанные на основе методов математической статистики и качество которых имеет доказательство, так и методы машинного обучения, направленные на обнаружение аномальных наблюдений.

Помимо пользы в работе с применением технологией машинного обучения, можно выделить и образовательную пользу. Программный комплекс может использоваться в качестве обучающего материал для демонстрации влияния выбросов на качество регрессионных моделей и средств создания устойчивых к выбросам моделей.

## СТАДИИ И ЭТАПЫ РАЗРАБОТКИ

## Стадии разработки, этапы и содержание работ

Стадии и этапы разработки были выявлены с учётом ГОСТ 19.102-77[[2]](#_bookmark30).

|  |  |  |  |
| --- | --- | --- | --- |
| 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-исследова тельские работы | Обоснование возможности решения поставленной задачи. | 21.11.2024 |
|  |  | Предварительный выбор методов решения задач. |  |
|  |  | Определение требований к техническим и программным средствам |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Разработка и утверждение  технического задания | Определение требований к программе.  Определение стадий, этапов и сроков разработки программы и документации на неё.  Согласование и утверждение технического задания.  Загрузка согласованного технического задания в SmartLMS | 04.12.2024 |
| Рабочий проект | Разработка программы | Предварительная разработка структуры программы  Разработка учебных материалов | 27.02.2025 |
|  | Разработка | Разработка программных документов в | 03.03.2025 |
| программной | соответствии с требованиями ГОСТ |  |
| документации | 19.101-77[[1]](#_bookmark29). |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Внедрение | Подготовка передача программы | и | Подготовка программы и программной документации для презентации и защиты  Представление разработанного программного продукта научному руководителю и получение отзыва | 15.03.2025 |
|  |  |  | 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. |  |

## 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

## 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.

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.
    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.

ПРИЛОЖЕНИЕ 2

{

"results": [

{

"model": "LSM",

"mlmodel": "None",

"noise\_type": "Normal",

"noise\_params": {

"param1": 0.0,

"param2": 1.0

},

"data\_path": "sample.csv",

"num\_features": 3,

"metrics\_by\_noise\_level": [

{

"noise\_level": 0,

"metrics": {

…

},

"detection\_metrics": {

"f1\_score" : 0.0,

"precision" : 0.0,

"recall" : 0.0,

},

"num\_experiments": 10

},

//...

]

},

//...

]

}

## ТЕРМИНЫ И СОКРАЩЕНИЯ

1. **Модель –** объект, представляющий собой набор весов, показывающих степень влияния некоторого признака в наборе данных. Модель используется для обнаружения зависимостей в наборе данных.
2. **Метрика** – некоторая мера, позволяющая измерить качество некоторой модели путем сравнения результатов применения модели с данными, описанными в наборе данных.
3. **Регрессия** – задача математической статистики, суть которой заключается в предсказании некоторой числовой величины по известным данным-признакам.
4. **Признак** – поле в наборе данных, которое используется как для предсказания величины-цели.
5. **Целевая переменная (цель)** – поле в наборе данных, значение которого модель стремится предсказать на основе признаков.
6. **Аномальное наблюдение** – объект в наборе данных, признаки которого многократно превосходят средние показатели других объектов.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Лист регистрации изменений | | | | | | | | | |
| Номера листов (страниц) | | | | | Всего листов (стр.) в докум. | № докум. | Входящ.  № сопров. докум. и дата | Подп. | Дата |
| Изм | изменен. | заменен | новых | аннулир. |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |