## 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 ".

## Краткая характеристика области применения

Различные технологии из области машинного обучения все чаще используются в самых разных компаниях и проектах. На основе прогнозов, которые выносятся моделями машинного обучения могут приниматься решения о состоянии здоровья человека, его благосостоянии, могут выноситься прогнозы роста или падения тех или иных показателей успешности предприятий и проч.

Самыми распространёнными и простыми моделями в машинном обучении являются регрессионные модели. Принципы работы этого класса моделей основаны на теоретических подходах, сформированных еще в рамках такой дисциплины, как математическая статистика, что не мешает ему быть эффективным инструментом в решении задач, требующих привлечения технологий машинного обучения. Однако у регрессионных моделей существует проблема, связанная с аномальными данными – алгоритмы, использующиеся для их обучения, часто не могут обрабатывать подобные данные, не уменьшая качество итоговой модели. Таким образом, встает вопрос о том, чтобы понять, насколько влияние выбросов сильно и как можно с ним бороться.

В рамках курсового проекта мы разработаем программный комплекс, позволяющий оценивать степень влияния выбросов на качество моделей, а также предоставляющий функционал для работы с данными, где есть выбросы. Такой функционал позволит обучать регрессионные модели на выбросах, не теряя в качестве результирующей модели.

Программный комплекс будет полезен как в образовательных целях на курсах по машинному обучению на программах, производящих подготовку по соответствующим направлениям, так и в коммерческих целях при использовании на данных с большим числом аномалий.

## 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

Программный комплекс, разрабатываемый в рамках проекта, может использоваться в качестве обучающего материала на курсах по машинному обучению на образовательных программах, реализующих подготовку студентов по таким направлениям, как «Прикладная математика и информатика», «Программная инженерия» и т.п. Также, комплекс может применяться при решении прикладных задач, связанных с обработкой данных с большим числом выбросов.

## Эксплуатационное назначение

Программный комплекс может использоваться на ПК с операционной системой Windows или Linux. Предполагается, что основная часть комплекса будет написана на языке С++ с возможным использованием модулей на языке Python для анализа качества моделей и оценки качества разработанных в рамках проекта средств борьбы с влиянием выбросов. Предполагается, что пользователь обладает знанием синтаксических особенностей этих языков.

## ТРЕБОВАНИЯ К ПРОГРАММЕ

## Требования к функциональным характеристикам

* + 1. Состав выполняемых функций
       1. Загрузка данных пользователя.
          1. При запуске комплекса пользователь попадает на начальный экран, где ему предлагается загрузить свои данные в память комплекса.
          2. Пользователь может загрузить свои данные в формате csv. Разделителем в файле с данными должна быть запятая, при наличии другого разделителя корректная работа комплекса не гарантируется.

Если в файле пользователя есть разметка, обозначающая, является ли объект аномалией, и пользователь захочет, чтобы эта разметка использовалась в расчёте метрик качества детектирования аномалий, ему необходимо указать название столбца, содержащего эту разметку.

Если в файле пользователя есть разметка, обозначающая величину, которая в дальнейшем будет использоваться как величина-цель в линейной регрессии или МНК, ему необходимо указать название столбца, содержащего эту разметку.

* + - * 1. Если пользователь не хочет работать со своими данными или же не имеет данных для анализа комплексом, ему предлагается выбрать набор данных из списка предохранённых в файлах комплекса. Данные комплекса сохранены так же в формате csv (разделитель - запятая).
        2. Выбор данных осуществляется через соответствующее окно, которое отображается при нажатии кнопки «Выбрать готовые данные».
        3. В случае, если пользователь не хочет использовать готовые данные, ему предлагается воспользоваться сгенерированными данными. Для этого пользователю необходимо нажать на кнопку «Сгенерировать данные». При генерации имеется возможность указать распределение желаемых данных, их типы и количество признаков для каждого объекта, в том числе и величины-цели. Комплекс должен предоставлять различные способы моделирования выбросов, включая генерацию данных из распределений с "тяжелыми хвостами" и внесение контролируемых искажений в некоторые наблюдения (например, возможность умножить определенное количество случайно выбранных наблюдений на заданный коэффициент). Эти настройки доступны через окно генерации данных.
      1. Первичная визуализация данных
         1. ***После загрузки данных в комплекс пользователю предлагается один из вариантов визуализации данных.***

***Визуализация в виде таблицы. Визуализация представляет собой pandas.DataFrame-подобное отображение данных в виде таблицы NxM, где N – количество свойств каждого объекта, а M – количество объектов.***

***Графическая визуализация. На основе полученных данных строятся графики, которые позволяют получить информацию о зависимостях величин или же о характеристиках некоторой величины внутри всего набора данных.***

***Визуализация зависимости. Пользователь выбирает величину-признак (величина X) и величину-цель (величина Y). Комплекс строит зависимость Y от X в виде набора точек на плоскости и аппроксимирующей кривой (кривой Лагранжа). Если величина-цель отсутствует в наборе данных, данная опция заблокирована.***

***Визуализация статистики величины. Пользователь выбирает величину-признак (величина Х). Комплекс отображает на графике распределение этой величины (в дискретном виде с аппроксимацией некоторой кривой). На этом же графике показываются такие статистические характеристики величины, как медиана, мода, среднее.***

***Выбор визуализации происходит в соответствующем окне, которое появляется при нажатии кнопки «Визуализировать данные». Визуализированные данные отображаются в отдельном окне. Визуализации можно будет сохранить, нажав на кнопку «Сохранить», указав путь до директории, где будет сохранены графики.***

* + - 1. Обработка данных.
         1. Поиск аномалий на основе M и R-оценок.

Пользователю предлагается выбор из M и R-оценок. Выбор оценки влияет на то, какие методы будут применятся комплексом для детекции аномалий.

Рядом с каждой из опций расположена кнопка «Узнать больше». По ее нажатии пользователь попадает на окно, в котором дается краткая характеристика выбранного метода.

Для каждой оценки имеется возможность выбора метода ее применения.

* + - * 1. Отображение результатов работы M и R-оценок.

После обнаружения аномалий данные визуализируются способом, схожим с тем, что описано в п. 4.1.1.2.1.2.2. На графике выделяются наблюдения, которые были признаны аномальными в процессе работы комплекса. После этого пользователь может создать новый файл, включающий входной набор данных и разметку аномальности, полученной из M/R оценок.

После визуализации с отображением аномальных наблюдений пользователь получает возможность получить дополнительную информацию о данных.

При выборе опции «Получить статистику» пользователю в графическом виде предоставляется информация об основных статистических характеристиках набора данных – моде, медиане и среднем.

При выборе опции «Регрессия» строится модель, которая отражает зависимость целевой переменной от остальных переменных в наборе данных. Построение модели производится с использованием метода градиентного спуска. Полученная модель визуализируется на графике. Если величина-цель отсутствует в наборе данных, данная опция заблокирована.

При выборе опции «МНК» строится модель, создание которой происходит с использованием метода наименьших квадратов. Полученная модель также визуализируется на графике. Если величина-цель отсутствует в наборе данных, данная опция заблокирована.

Поиск аномалий с использованием методов машинного обучения.

Поиск аномалий происходит на изначальном наборе данных с использованием классических методов машинного обучения для обнаружения аномалий – кластеризации и построения плотности. После обнаружения аномалий результат визуализируется по принципу, описанному в п. 4.1.1.3.2.2.

Сравнительный анализ метрик.

Если в наборе данных присутствует разметка, обозначающая, является ли объект аномалией, то для каждого из способов детекции аномалий строятся метрики классификации: полнота, точность и f1-мера – которые затем сравниваются.

Если в наборе присутствует разметка, обозначающая величину, которая используется как величина-цель в линейной регрессии или МНК, то вычисляются метрика качества регрессии: средняя квадратическая ошибка, средняя абсолютная ошибка, коэффициент детерминации R^2 и другие.

Результат сравнения может быть сохранен в формате .json, в котором содержатся тип данных: синтетические, пользовательские или из содержащихся в комплексе, количество объектов в наборе, размер признакового описания одного объекта, тип оценок (M/R) и метрики качества классификации и регрессии. Пример организации подобного файла приведён в Приложении 2.

## Требования к организации входных данных

Входные данные должны быть представлены в формате CSV с использованием запятой в качестве разделителя. Первой строкой файла должны быть названия столбцов (заголовки). Каждое имя столбца должно быть уникальным и не содержать специальных символов. Данные внутри столбцов должны быть однородными: для числовых переменных допустимы только значения типа float или int. Пропуски данных недопустимы, при обнаружении пропусков пользователю будет предложено удалить признаки с пропусками. Данные не должны содержать дублирующихся строк. При обнаружении дублирования пользователю будет предложено удалить их. Размер файла ограничен 1 ГБ. При превышении лимита пользователь получает предупреждение с предложением уменьшить объем данных.

## 

## Требования к организации выходных данных

Все результаты работы комплекса должны сохраняться в CSV или JSON форматах.

При сохранении данных с аномалиями к исходным данным добавляется столбец anomaly, где 1 означает аномалию, а 0 — нормальное наблюдение.

Результаты сравнения методов (метрики) сохраняются в JSON-формате, как описано в пункте 4.1.1.3.2.4.1. Графики, построенные в процессе анализа, могут быть сохранены в форматах PNG или SVG.

## Требования к интерфейсу

## Общие требования: интерфейс программного комплекса должен быть интуитивно понятным и удобным для пользователя, обеспечивая доступ ко всем функциям через понятные и логически организованные элементы управления. Графический интерфейс (GUI) должен поддерживать управление с использованием мыши и клавиатуры.

## Требования к интерфейсу загрузки данных: при запуске комплекса на начальном экране должна отображаться кнопка «Загрузить данные», которая позволяет пользователю выбрать файл формата CSV с разделителем-запятой. При загрузке данных с разметкой пользователь должен иметь возможность указать названия столбцов для: метки аномальности, целевой величины для линейной регрессии или МНК. В случае отсутствия пользовательских данных интерфейс должен отображать кнопки «Выбрать готовые данные» и «Сгенерировать данные». Окно генерации данных должно включать параметры для выбора типа распределения, количества признаков и указания величины-цели.

## Требования к интерфейсу визуализации данных: после загрузки данных пользователю должны предлагаться два типа визуализации: табличное представление (в виде таблицы, подобной pandas.DataFrame); графическая визуализация. Для графической визуализации пользователь должен иметь возможность выбрать: зависимость между двумя величинами (с отображением аппроксимирующей кривой); распределение одной величины с отображением статистических характеристик. Визуализация должна быть доступна через кнопку «Визуализировать данные», результат отображается в отдельном окне. Графики должны быть сопровождаемы кнопкой «Сохранить», позволяющей выбрать директорию для сохранения результатов в форматах PNG или JPG.

Требования к интерфейсу обработки данных: для выбора методов обнаружения аномалий (M/R-оценки) интерфейс должен предоставлять список с описанием каждого метода и кнопкой «Узнать больше», которая открывает окно с дополнительной информацией. Результаты работы методов (например, выделение аномальных наблюдений) должны визуализироваться графически, аналогично интерфейсу первичной визуализации, с выделением аномальных объектов. Пользователь должен иметь возможность получить дополнительные данные через кнопки:

«Получить статистику» – для вывода статистических характеристик;

«Регрессия» – для построения модели на основе градиентного спуска;

«МНК» – для построения модели методом наименьших квадратов.

Недоступные опции (например, при отсутствии целевой величины) должны быть визуально заблокированы.

## Требования к интерфейсу анализа результатов: в случае работы с размеченными данными интерфейс должен предоставлять функции сравнения метрик качества для различных методов детекции аномалий. Пользователь должен иметь возможность сохранить результат анализа в формате JSON через кнопку «Сохранить результаты». Файлы результатов должны включать метаданные: тип данных, размер выборки, тип оценки (M/R) и метрики качества.

## Требования к надежности

При обнаружении аномалий на основе M и R-оценок пользователь выбирает метод оценки из предложенных в интерфейсе. Если выбранный метод недоступен из-за отсутствия данных или некорректного формата загруженного файла, программа отображает сообщение об ошибке с указанием причины. В таких случаях состояние комплекса остается неизменным, и пользователь может повторить попытку выбора метода после устранения проблемы.

При загрузке данных пользователь выбирает файл в формате CSV через диалоговое окно. Если выбранный файл не соответствует требованиям (например, отсутствует разделитель-запятая, файл пустой или имеет неверную структуру), программа отображает сообщение об ошибке, не загружает данные и возвращает пользователя на начальный экран для повторного выбора или генерации данных.

При генерации данных пользователь указывает параметры распределения, количество признаков и другие настройки через специальное окно. Если введенные параметры невалидны (например, отрицательное количество признаков или некорректное распределение), программа выдает сообщение об ошибке, отклоняет введенные параметры и сохраняет предыдущие корректные настройки.

При визуализации данных программа строит графики на основе выбранных пользователем величин. Если выбранные данные не позволяют построить визуализацию (например, величина-цель отсутствует или содержит некорректные значения), интерфейс блокирует недоступные опции и сообщает пользователю о невозможности выполнить операцию. В таких случаях программа не изменяет текущее состояние. При сохранении результатов пользователь указывает директорию и имя файла для сохранения. Если указанный путь недоступен (например, отсутствует разрешение записи или путь не существует), программа отображает сообщение об ошибке и предлагает выбрать другой путь. Состояние программы и результаты анализа при этом сохраняются в памяти для повторной попытки.

## 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 не ниже Intel Core i5-8400 или AMD Ryzen 5 2600;
2. 8 ГБ оперативной памяти или больше;
3. 2 ГБ видеопамяти или больше;
4. Видеокарта не ниже не ниже Nvidia GeForce GTX 1050 или AMD Radeon RX 560
5. Монитор разрешением не менее 1920x1080.
6. Наличие SSD-диска с не менее 10 ГБ свободного пространства.

Минимальные требования к клиентскому оборудованию для работы приложения:

1. Процессор Intel Core i3-6100 или AMD FX-4100;
2. 4 ГБ оперативной памяти;
3. Видеокарта Intel HD Graphics 530 или AMD Radeon R5 230;
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 (7 и выше), Linux (Ubuntu 20.04 и выше) или macOS;

## 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 3.24.1. Должен использоваться компилятор GCC версии не ниже 10.5. Для управления версиями приложения используется система контроля версий Git. Версия языка Python должна быть не ниже 3.8.

## 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. Приложение для визуализации структуры данных «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях». Техническое задание (ГОСТ 19.201-78 [[7]](#_bookmark35)).
    2. Приложение для визуализации структуры данных «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях». Пояснительная записка (ГОСТ 19.404-79[[12]](#_bookmark39)).
    3. Приложение для визуализации структуры данных «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях». Программа и методика испытаний (ГОСТ 19.301-79[[13]](#_bookmark40)).
    4. Приложение для визуализации структуры данных «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях». Текст программы (ГОСТ 19.401-78[[14]](#_bookmark41)).
    5. Приложение для визуализации структуры данных «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях». Руководство оператора (ГОСТ 19.505-79[[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 аномалиям).

## 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. Инженер, который будет использовать комплекс, 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 именного того результат, к которому он, пользователь, идет. Второе достоинство заключается в предоставлении комплексом средств, позволяющих бороться с выбросами. Эти средства и механизмы будут разработаны на основе методов математической статистики, они надежны, их качество имеет доказательство.

Помимо пользы в работе с применением технологий 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.2024. (дата защиты курсовой работы). Исполнители – Знатнов Егор и Панкратов Степан, студенты 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. Maronna R. A. et al. Robust statistics: theory and methods (with R). – John Wiley & Sons, 2019. [Электронный ресурс] /Режим доступа: приватный (дата обращения: 26.11.2024).
    16. Chang W. H. et al. High-breakdown rank regression //Journal of the American Statistical Association. – 1999. – Т. 94. – №. 445. – С. 205-219. [Электронный ресурс] /Режим доступа: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10473836> Свободный (дата обращения: 26.11.2024)
    17. Hettmansperger T. P., McKean J. W. Robust nonparametric statistical methods. – CRC press, 2010. [Электронный ресурс] /Режим доступа: приватный (дата обращения: 26.11.2024)

ПРИЛОЖЕНИЕ 2

{

"input\_data" : "generated" или "custom" или "builtin",

"input\_size" : // целое число,

"features\_len" : // целое число,

"estimates\_type" : "R" или "M",

// этот раздел не пуст, если присутствует величина-цель

"target\_regression" : {

"machine\_learning\_regression" : {

"mse" : // десятичное число,

"mae" : // десятичное число,

// другие метрики

},

"least\_squares" : {

"mse" : // десятичное число,

"mae" : // десятичное число,

// другие метрики

}

},

// этот раздел не пуст, если присуствует разметка,

// характеризующая "аномальность" объекта

"anomaly\_detection" : {

"machine\_learning\_algorithms" : {

"dbscan" : {

"f1\_score" : // десятичное число,

"precision" : // десятичное число,

"recall" : // десятичное число,

}

"isolation\_forest" : {

"f1\_score" : // десятичное число,

"precision" : // десятичное число,

"recall" : // десятичное число,

}

// могут быть другие методы

},

"statistics\_estimates" : {

"f1\_score" : // десятичное число,

"precision" : // десятичное число,

"recall" : // десятичное число,

}

}

}

## 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 машинного обучения, 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. **Аномалия** – 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. |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |