## ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИФЕДЕРАЛЬНОЕГОСУДАРСТВЕННОЕАВТОНОМНОЕ

**ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯНАЦИОНАЛЬНЫЙИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ**

## «ВЫСШАЯШКОЛАЭКОНОМИКИ»

Факультет компьютерных наукОбразовательнаяпрограмма«Программнаяинженерия»

СОГЛАСОВАНО

доцент департамента математики

факультета экономических наук,

кандидатфизико-математическихнаук

Е.Р. Горяинова

УТВЕРЖДАЮ

Академический руководительобразовательнойпрограммы

«Программная инженерия»старшийпреподавательдепартамента

программной инженерииН.А.Павлочев

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

|  |  |
| --- | --- |
| ***Подп.идата*** |  |
| ***Инв.№дубл.*** |  |
| ***Взам.инв.№*** |  |
| ***Подп.идата*** |  |
| ***Инв.№подл*** |  |

**Техническое заданиеЛИСТУТВЕРЖДЕНИЯ**

# RU.17701729.05.05-01ТЗ01-1-ЛУ

ИсполнителистудентгруппыБПИ221

/Знатнов Е.П./

студентгруппыБПИ221

/Панкратов С.Ю./

« » 2024 г.

УТВЕРЖДЕН

RU.17701729.05.05-01ТЗ01-1-ЛУ

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

**Техническое заданиеRU.17701729.05.05-01ТЗ01-1**

|  |  |
| --- | --- |
| ***Подп.идата*** |  |
| ***Инв.№дубл.*** |  |
| ***Взам.инв.№*** |  |
| ***Подп.идата*** |  |
| ***Инв.№подл*** |  |

# Листов 26

## АННОТАЦИЯ

Техническоезадание–основнойдокумент,оговаривающийнабортребованийипорядок создания программного продукта, в соответствии с которым производится разработкапрограммы,ее тестирование и приемка.

НастоящееТехническоезаданиенаразработку программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделяхсодержитследующиеразделы:«Введение»,«Основаниедляразработки»,«Назначениеразработки»,«Требованиякпрограмме»,«Требованиякпрограммнымдокументам»,«Технико-экономическиепоказатели»,«Стадиииэтапыразработки»,«Порядок контроляи приемки»и приложения.

Вразделе«Введение»указанонаименованиеикраткаяхарактеристикаобластиприменения программного комплекса.

Вразделе«Основаниядляразработки»указаныдокументы,наоснованиикоторыхведетсяразработка, а такженаименование темы разработки.

Вразделе«Назначениеразработки»указанофункциональноеиэксплуатационноеназначениепрограммного продукта.

Раздел «Требования к программе» содержит основные требования к функциональнымхарактеристикам,надежности,условиямэксплуатации,составуипараметрамтехническихсредств,информационнойипрограммнойсовместимости,маркировкеиупаковке,транспортированиюи хранению.

Раздел «Требования к программным документам» содержит предварительный составпрограммнойдокументации и специальные требования к ней.

Раздел«Технико-экономическиепоказатели»описываеториентировочнуюэкономическуюэффективность,предполагаемуюгодовуюпотребность,атакжеэкономическиепреимущества разработки посравнению с аналогами.

Раздел«Стадиииэтапыразработки»содержитстадиииэтапыразработки,ихсодержаниеисроки,а такжеуказываетлица,ответственные заихвыполнение.

Вразделе«Порядокконтроляиприемки»указаныобщиетребованиякприемкеработы,атакже зафиксированывсе допустимыепри этомвиды испытаний.

Настоящийдокументразработанвсоответствиистребованиями:

1. ГОСТ19.101-77[[1]](#_bookmark29):Видыпрограммипрограммныхдокументов.
2. ГОСТ19.102-77[[2]](#_bookmark30):Стадииразработки.
3. ГОСТ19.103-77[[3]](#_bookmark31):Обозначенияпрограммипрограммныхдокументов.
4. ГОСТ19.104-78[[4]](#_bookmark32):Основныенадписи.
5. ГОСТ19.105-78[[5]](#_bookmark33):Общиетребованиякпрограммнымдокументам.
6. ГОСТ19.106-78[[6]](#_bookmark34):Требованиякпрограммнымдокументам,выполненнымпечатнымспособом.
7. ГОСТ19.201-78[[7]](#_bookmark35):Техническоезадание.Требованияксодержаниюиоформлению.
8. ГОСТ 19.602-78[[8]](#_bookmark36): Правила дублирования, учетаихраненияпрограммныхдокументов,выполненных печатнымспособом.

ИзменениякнастоящемутехническомузаданиюдолжныбытьоформленысогласноГОСТ19.603-78[[9]](#_bookmark37)иГОСТ 19.604-78[[10]](#_bookmark38).

# Содержание

[1. ВВЕДЕНИЕ 5](#_Toc158543924)

[1.1. Наименованиепрограммы 5](#_Toc158543925)

[1.2. Краткаяхарактеристикаобластиприменения 5](#_Toc158543926)

[2. ОСНОВАНИЯДЛЯРАЗРАБОТКИ 6](#_Toc158543927)

[2.1.Документ(ы),наоснованиикоторого(ых)ведётсяразработка 6](#_Toc158543928)

[2.2Наименованиетемыразработки 6](#_Toc158543929)

[3. НАЗНАЧЕНИЕРАЗРАБОТКИ 7](#_Toc158543930)

[3.1. Функциональноеназначение 7](#_Toc158543931)

[3.2. Эксплуатационноеназначение 7](#_Toc158543932)

[4. ТРЕБОВАНИЯКПРОГРАММЕ 8](#_Toc158543933)

[4.1. Требованиякфункциональнымхарактеристикам 8](#_Toc158543934)

[4.1.2. Требованиякорганизациивходныхданных 11](#_Toc158543935)

[4.1.3Требованиякорганизациивыходныхданных 11](#_Toc158543936)

[4.1.4.Требованиякинтерфейсу 11](#_Toc158543937)

[4.2. Требованиякнадежности 13](#_Toc158543938)

[4.3. Условияэксплуатации 14](#_Toc158543939)

[4.4. Требованияксоставуипараметрамтехническихсредств 14](#_Toc158543940)

[4.5. Требованиякинформационнойипрограммнойсовместимости 15](#_Toc158543941)

[4.6. Требованиякмаркировкеиупаковке 15](#_Toc158543942)

[4.7. Требованияктранспортировке 15](#_Toc158543943)

[5. ТРЕБОВАНИЯКПРОГРАММНОЙДОКУМЕНТАЦИИ 16](#_Toc158543944)

[5.1. Предварительныйсоставпрограммнойдокументации 16](#_Toc158543945)

[5.2. Специальныетребованиякпрограммнойдокументации 16](#_Toc158543946)

[6. ТЕХНИКО-ЭКОНОМИЧЕСКИЕПОКАЗАТЕЛИ 17](#_Toc158543947)

[6.1 Ориентировочнаяэкономическаяэффективность 17](#_Toc158543948)

[6.2 Предполагаемаяпотребность 17](#_Toc158543949)

[6.3 Экономическиепреимуществаразработкипосравнениюсотечественнымиизарубежнымиобразцами или аналогами 17](#_Toc158543950)

[7. СТАДИИИЭТАПЫРАЗРАБОТКИ 18](#_Toc158543951)

[7.1. Стадииразработки,этапыисодержаниеработ 18](#_Toc158543952)

[7.2. Срокиразработкииисполнители 20](#_Toc158543953)

[8. ПОРЯДОККОНТРОЛЯИПРИЁМКИ 21](#_Toc158543954)

[СПИСОКИСПОЛЬЗУЕМОЙЛИТЕРАТУРЫ. 22](#_Toc158543955)

[ТЕРМИНЫИСОКРАЩЕНИЯ 25](#_Toc158543956)

## 1. ВВЕДЕНИЕ

## Наименованиепрограммы

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

Наименованиетемыразработкинаанглийскомязыке: "Development of a Software Package to Study the Influence of Anomalous Observations Outliers on the Prediction Accuracy in Regression Models".

## Краткаяхарактеристикаобластиприменения

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.

## ОСНОВАНИЯДЛЯРАЗРАБОТКИ

## 2.1.Документ(ы),наоснованиикоторого(ых)ведётсяразработка

Основаниемдляразработкиявляетсяучебныйпланподготовкибакалавров  
понаправлению09.03.04«Программнаяинженерия»иутвержденнаяакадемическимруководителемтема курсовогопроекта.

## 2.2Наименованиетемыразработки

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

## НАЗНАЧЕНИЕРАЗРАБОТКИ

## Функциональноеназначение

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.

## Эксплуатационноеназначение

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

## ТРЕБОВАНИЯКПРОГРАММЕ

## Требованиякфункциональнымхарактеристикам

* + 1. Составвыполняемыхфункций
       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. Все это делается через окно генерации. Здесь хотелось бы, чтобы была возможность моделировать выбросы не только как распределения с «тяжёлыми хвостами». Например, несколько наблюдений умножить на 100
      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 Yот Xв 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 Mи R-оценок. Эти методы направлены в основном не на поиск выбросов, а на построение «хороших» оценок при их наличии

Пользователю предлагается выбор из Mи 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 Mи R-оценок и МНК тоже.

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/Rоценок.

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 – которыезатем сравниваются.

Если в наборе присутствует разметка, обозначающая величину, котораяиспользуется 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.

Надо предусмотреть многократное моделирование данных с однотипными выбросами, чтобы затем сравнить метрики качества построенных моделей при применении разных методов оценивания

## Требованиякорганизациивходныхданных

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.

## 

## Требованиякорганизациивыходныхданных

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.

## Требованиякинтерфейсу

## 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; метод градиентного спуска решает задачу поиска минимума целевой функции, но иногда лучше использовать другие методы. Например, при построении ранговых оценок надо минимизировать кусочно-линейную функцию потерь. Здесь минимум ищется с помощью методов линейного программирования.

«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.

## Требованиякнадежности

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 впамяти дляповторнойпопытки.

## Условияэксплуатации

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.

## Требованияксоставуипараметрамтехническихсредств

* + 1. **Требованиякклиентскомуоборудованию**

Рекомендуемыетребованиякклиентскомуоборудованиюдлякорректнойработыприложения:

1. Процессор не ниже IntelCorei5-8400 или AMDRyzen 5 2600;
2. 8 ГБ оперативной памяти или больше;
3. 2ГБвидеопамятиили больше;
4. Видеокартаненижене ниже NvidiaGeForceGTX 1050 или AMDRadeonRX 560
5. Мониторразрешениемнеменее1920x1080.
6. Наличие SSD-диска с не менее 10 ГБ свободного пространства.

Минимальныетребованиякклиентскомуоборудованиюдляработыприложения:

1. Процессор Intel Core i3-6100 или AMD FX-4100;
2. 4ГБоперативнойпамяти;
3. ВидеокартаIntel HD Graphics 530 или AMD Radeon R5 230;
4. Мониторразрешением1280x720.
5. Жесткий диск с не менее 5 ГБ свободного пространства.

Общиетребованиякклиентскомуоборудованиюдляработыприложения:

1. Мышьилисовместноеуказывающееустройство;
2. Клавиатура;
3. ОС: Windows (7 и выше), Linux (Ubuntu 20.04 и выше) или macOS;

## Требованиякинформационнойипрограммнойсовместимости

Приложение должно быть написано в соответствии со стандартом языка С++17. Для сборки проекта используется система сборки CMakeверсии не ниже 3.24.1. Должен использоваться компилятор GCCверсии не ниже 10.5. Для управления версиями приложения используется система контроля версий Git. Версия языка Pythonдолжна быть не ниже 3.8.

## Требованиякмаркировкеиупаковке

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

## Требованияктранспортировке

Особыхтребованийктранспортировкеихранениюнепредъявляются.

## ТРЕБОВАНИЯКПРОГРАММНОЙДОКУМЕНТАЦИИ

## Предварительныйсоставпрограммнойдокументации

* + 1. Приложение для визуализации структуры данных «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях».Техническоезадание(ГОСТ19.201-78 [[7]](#_bookmark35)).
    2. Приложение для визуализации структуры данных «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях».Пояснительнаязаписка(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».Программаиметодикаиспытаний (ГОСТ19.301-79[[13]](#_bookmark40)).
    4. Приложение для визуализации структуры данных «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях».Текстпрограммы(ГОСТ19.401-78[[14]](#_bookmark41)).
    5. Приложение для визуализации структуры данных «Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях».Руководствооператора(ГОСТ 19.505-79[[15]](#_bookmark42)).

## Специальныетребованиякпрограммнойдокументации

Программная документация должна быть выполнена в соответствиисГОСТ19.106-78[[6]](#_bookmark34)и ГОСТамик каждомувидудокумента.

## ТЕХНИКО-ЭКОНОМИЧЕСКИЕПОКАЗАТЕЛИ

## Ориентировочнаяэкономическаяэффективность

В рамках проекта расчет экономической эффективности программного продукта непроизводился.

## Предполагаемаяпотребность

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).

## Экономическиепреимуществаразработкипосравнениюсотечественнымиизарубежнымиобразцами 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.

## СТАДИИИЭТАПЫРАЗРАБОТКИ

## Стадииразработки,этапыисодержаниеработ

СтадиииэтапыразработкибыливыявленысучётомГОСТ19.102-77[[2]](#_bookmark30).

|  |  |  |  |
| --- | --- | --- | --- |
| Стадияразработки | Этапработ | Содержаниеработ | Сроки  выполнения |
| Техническоезадание | Обоснованиенеобходимостиразработкипрограммы | Постановказадачи  Сборисходныхматериалов.  Выбориобоснованиекритериевэффективности и качества разрабатываемойпрограммы. | 04.11.2024 |
|  | Научно-исследовательскиеработы | Обоснование возможности решенияпоставленнойзадачи. | 21.11.2024 |
|  |  | Предварительныйвыборметодоврешениязадач. |  |
|  |  | Определениетребованийктехническимипрограммнымсредствам |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Разработка иутверждение  техническогозадания | Определениетребованийкпрограмме.  Определениестадий,этаповисроковразработкипрограммыидокументации нанеё.  Согласованиеиутверждениетехническогозадания.  Загрузкасогласованноготехническогозадания в SmartLMS | 04.12.2024 |
| Рабочийпроект | Разработкапрограммы | Предварительная разработка структурыпрограммы  Разработкаучебныхматериалов | 27.02.2025 |
|  | Разработка | Разработка программных документов в | 03.03.2025 |
| программной | соответствии с требованиями ГОСТ |  |
| документации | 19.101-77[[1]](#_bookmark29). |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Внедрение | Подготовкапередачапрограммы | и | Подготовкапрограммыипрограммнойдокументациидля презентации и защиты  Представление разработанногопрограммного продукта научномуруководителюиполучениеотзыва | 15.03.2025 |
|  |  |  | Представление разработанногопрограммного продукта научномуруководителюиполучениеотзыва. |  |
|  |  |  | ЗагрузкаПояснительнойзапискивсистемуАнтиплагиатчерез ЛМС НИУВШЭ. |  |
|  |  |  | ЗагрузкаматериаловкурсовогопроектавЛМС |  |
|  |  |  | Защитапрограммногопродуктакомиссии. |  |

## Срокиразработкииисполнители

Разработкапрограммногопродуктадолжнабытьзавершена к XX.XX.2024.(датазащитыкурсовойработы).Исполнители–Знатнов Егор и Панкратов Степан, студенты группыБПИ221факультетакомпьютерныхнаук НИУВШЭ

## ПОРЯДОККОНТРОЛЯИПРИЁМКИ

Контрольиприемкаразработкиосуществляютсявсоответствиисдокументом

«Разработка программного комплекса для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях».Программаи методикаиспытаний» и пунктом5.2 технического задания.

ПРИЛОЖЕНИЕ1

## СПИСОКИСПОЛЬЗУЕМОЙЛИТЕРАТУРЫ.

* + 1. ГОСТ19.101-77:Видыпрограммипрограммныхдокументов.//Единаясистемапрограммнойдокументации.–М.: ИПКИздательствостандартов,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 системапрограммнойдокументации.–М.: ИПКИздательствостандартов,2001.
    4. ГОСТ 19.104-78: Основные надписи. // Единая система программной документации. –М.:ИПК Издательство стандартов, 2001.
    5. ГОСТ19.105-78:Общие требования к программным документам. // Единая системапрограммнойдокументации.–М.: ИПКИздательствостандартов,2001.
    6. ГОСТ19.106-78:Требованиякпрограммным документам, выполненным печатнымспособом.//Единаясистемапрограммнойдокументации.–М.:ИПКИздательствостандартов,2001.
    7. ГОСТ 19.201-78: Техническое задание. Требования к содержанию и оформлению. //Единаясистемапрограммнойдокументации.–М.:ИПКИздательствостандартов,2001.
    8. ГОСТ 19.602-78: Правила дублирования, учета и хранения программных документов,выполненных печатным способом. //Единая система программной документации. – М.:ИПКИздательство стандартов, 2001.
    9. ГОСТ 19.603-78: Общие правила внесения изменений. // Единая система программнойдокументации.– М.:ИПК Издательствостандартов, 2001.
    10. ГОСТ19.604-78:Правилавнесенияизмененийвпрограммныедокументы,выполненные печатным способом. // Единая система программной документации. – М.:ИПКИздательство стандартов, 2001.
    11. ГОСТ 19.404-79: Пояснительная записка. Требования к содержанию и оформлению. //Единаясистемапрограммнойдокументации.–М.:ИПКИздательствостандартов,2001.
    12. ГОСТ19.301-79:Программаиметодикаиспытаний.Требованияксодержаниюиоформлению. // Единая система программной документации. – М.: ИПК Издательствостандартов,2001.
    13. ГОСТ19.401-78:Текстпрограммы.Требованияксодержаниюиоформлению.//Единаясистемапрограммнойдокументации.–М.:ИПКИздательствостандартов,2001.
    14. ГОСТ 19.505-79: Руководство оператора. Требования к содержанию и оформлению. //Единаясистемапрограммнойдокументации.–М.:ИПКИздательствостандартов,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" : // целоечисло,

"features\_len" : // целоечисло,

"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" : // десятичноечисло,

"precision" : // десятичное число,

"recall" : // десятичное число,

}

"isolation\_forest" : {

"f1\_score" : // десятичноечисло,

"precision" : // десятичное число,

"recall" : // десятичное число,

}

// могут быть другие методы

},

"statistics\_estimates" : {

"f1\_score" : // десятичное число,

"precision" : // десятичное число,

"recall" : // десятичное число,

}

}

}

## ТЕРМИНЫИСОКРАЩЕНИЯ

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 математической статистики (эти задачи были сформулированы до того, как появилось машинное обучение), 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 лучше выброс или аномальное наблюдение**– объект в наборе данных, признаки которого многократно превосходят средние показатели других объектов.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Листрегистрацииизменений | | | | | | | | | |
| Номералистов(страниц) | | | | | Всеголистов(стр.) вдокум. | №докум. | Входящ.  №сопров.докум. идата | Подп. | Дата |
| Изм | изменен. | заменен | новых | аннулир. |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |