## 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./

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 10](#_Toc158543935)

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

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

[4.2. Требования к надежности 12](#_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 Anomalous Observations 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 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. Работа с данными
          1. При запуске комплекса пользователь попадает на главный экран, где ему доступны основные функции для работы с данными:

1. Создание моделей регрессии
2. Сохранение конфигураций моделей
3. Запуск анализа
4. Генерация синтетических данных
5. Статистики целевой переменной
   * + - 1. Пользователь может загрузить свои данные в формате CSV при создании модели. CSV-файлы должны использовать точку с запятой (";") в качестве разделителя. Данные должны быть структурированы таким образом, чтобы признаки располагались в первых N столбцах, а целевая переменная - в последнем столбце. При создании модели пользователь указывает путь к файлу данных в поле "Path to data", а также количество признаков в поле "Num. features". Если файл не существует или имеет некорректный формат, пользователь получит соответствующее уведомление.
         2. Если пользователь не имеет своих данных для анализа, комплекс предоставляет возможность генерации синтетических данных. Для этого необходимо нажать кнопку "Generate data", после чего откроется диалоговое окно с настройками генерации.

В диалоговом окне пользователю доступны следующие параметры:

1. Нижняя граница генерации данных, по умолчанию – 1.
2. Верхняя граница генерации данных, по умолчанию – 100.
3. Количество признаков - число независимых переменных в генерируемом наборе данных
4. Количество наблюдений - общее число строк в генерируемом наборе данных
5. Поле для передачи коэффициентов. Коэффициенты для искомой зависимости передаеются строкой, каждый коэффициент – действительное число с разделителем точкой. Сами числа разделяются с помощью точки с запятой. Коэффициентов в этом окне должно быть на один больше, чем количество признаков.

После указания параметров, комплекс генерирует данные и сохраняет их в файл. Данные генерируются с использованием равномерного распределения.

* + - * 1. После генерации данных пользователю предоставляется возможность просмотреть выборку из первых 25 объектов из сгенерированного набора данных для оценки качества и структуры сгенерированных данных. Образец данных отображается в отдельном диалоговом окне.
        2. Комплекс должен предоставлять различные способы моделирования выбросов, включая генерацию случайного шума из распределений с "тяжелыми хвостами" и внесение контролируемых искажений в некоторые наблюдения. Эти настройки доступны через окно выбора метода регрессии.
        3. Комплекс предоставляет возможность запуска модели на предоставленных данных. Такая опция имеет кодовое название dryRun. Под запуском модели на некоторых данных подразумевается следующее – комплекс не производит анализ метрик качества модели, а просто запускает выбранный пользователем метод для поиска им зависимости на предоставленных данных. Результатом работы метода в таком режиме является информация о найденной зависимости в виде весов линейной регресии, веса сохраняются в текстовый файл.
        4. Комплекс предоставляет возможность выполнения статистического анализа загруженных или сгенерированных данных. При выборе кнопки для подсчета статистик выводится диалоговое окно, в котором пользователь может указать путь к CSV-файлу с данными.

После выбора файла комплекс вычисляет и отображает следующие статистические характеристики для целевой переменной:

1. Минимальное значение
2. Максимальное значение
3. Среднее значение
4. Медиана
5. Дисперсия
   * + 1. Методы обработки и анализа данных
          1. Методы регрессионого анализа

Классический метод регрессии

Пользователю предлагается воспользоваться традиционным методом регрессионного анализа – методом наименьших квадратов, который обычно чувствителен к наличию аномальных наблюдений в данных.

Робастные методы регрессии

Пользователю предоставляются различные устойчивые методы регрессии, которые менее чувствительны к аномальным наблюдениям:

1. Методы М-оценок – используют различные функции потерь для уменьшения влияния выбросов на модель.
2. Методы на основе медианы – используют медианные оценки для обеспечения устойчивости к аномалиям.
3. Методы на основе наименьших абсолютных отклонений – минимизируют сумму абсолютных значений отклонений.

При создании модели пользователь может выбрать тип регрессии и настроить ее параметры, такие как:

1. Параметр чувствительности для робастных методов
2. Количество итераций для градиентных методов обучения
3. Скорость обучения для методов на основе градиентного спуска
   * + - 1. Визуализация и сохранение результатов

## Требования к организации входных данных

Входные данные должны быть представлены в формате CSV с использованием точки с 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, JSON или PNG форматах. Результаты сравнения методов сохраняются в JSON-формате, как описано в пунктах 4.1.1.2.3.3. и 4.1.1.2.3.2. Графики, построенные в процессе анализа, будут сохранены в формате PNG. При запуске комплекса в режиме dryRun метод генерирует файл в формате TXT с описанием весов, полученных при проходе выбранных пользователем методов.

## 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 работы с данными При запуске комплекса на главном экране должны отображаться кнопки "Create model", при нажатии которой должно открываться диалоговое окно, позволяющее выбрать тип регрессионной модели и настроить ее параметры, "Dump models" для сохранения конфигураций созданных моделей, "Run on models", при нажатии которой должен запуститься анализ на основе созданных моделей, "Generate data" для генерации синтетических данных и "Calculate statistics" для расчета статистик целевой переменной.

## Требования к интерфейсу визуализации результатов После запуска анализа результаты должны автоматически визуализироваться в виде графиков, показывающих зависимость ошибки от уровня шума для каждой комбинации регрессионной модели и метода обнаружения аномалий. Графики должны отображаться в отдельном окне и включать заголовок с указанием типа регрессионной модели и метода обнаружения аномалий, подписи осей, точки, представляющие усредненные результаты экспериментов для каждого уровня шума и легенду, идентифицирующую каждую серию данных на графике с указанием используемого метода обнаружения аномалий и типа распределения шума.

## Требования к интерфейсу анализа и сохранения результатов Интерфейс должен обеспечивать возможность просмотра графиков зависимости ошибки от уровня шума, метрик качества регрессионных моделей и метрик качества обнаружения аномалий. Все результаты анализа должны автоматически сохраняться в структурированном формате JSON, содержащем информацию о модели регрессии, информацию о методе обнаружения аномалий, метрики качества регрессии для каждого уровня шума, метрики качества обнаружения аномалий для каждого уровня шума. Файлы результатов должны сохраняться с именами, содержащими информацию о типе регрессионной модели и методе обнаружения аномалий, для удобства последующего анализа.

## Требования к надежности

При создании модели, если указанный путь к файлу данных не существует, программа должна выводить информативное сообщение об ошибке и оставлять поле пути пустым. Если файл данных имеет неправильный формат, программа должна выводить соответствующее сообщение об ошибке. При невозможности сохранения результатов, программа должна информировать пользователя о проблеме.

При вводе недопустимых значений параметров, интерфейс должен блокировать ввод таких значений или выводить предупреждение. При попытке запуска анализа без созданных моделей, программа должна информировать пользователя о необходимости создания моделей. При возникновении ошибок в процессе вычислений, программа должна обрабатывать эти ошибки и продолжать работу с оставшимися моделями

При сбоях программного комплекса состояние программы и результаты анализа должны быть сохранены 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;

## 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 | 15.04.2025 |
|  |  |  | 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. Конспект ИТМО «Персистентный стек» [Электронный ресурс] /Режим доступа:

<https://neerc.ifmo.ru/wiki/index.php?title=Персистентный_стек>

свободный (дата обращения: 02.01.2024).

* + 1. Лекторий ФПМИ МФТИ «Алгоритмы и структуры данных (продвинутый поток). 2. Амортизированный анализ. Персистентная очередь» [Электронный ресурс] /Режим доступа: <https://www.youtube.com/watch?v=r6NnmynZMRs>, свободный (дата обращения: 04.01.2024).

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. |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |