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

Факультет компьютерных наук  
Департамент программной инженерии

|  |  |
| --- | --- |
| СОГЛАСОВАНО  Доцент департамента математики факультета экономических наук, кандидат физико-математических наук  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Е. Р. Горяинова «\_\_\_\_»­­ \_\_\_\_\_\_\_\_\_\_ 2025 г. | УТВЕРЖДАЮ  Академический руководитель образовательной программы «Программная инженерия» профессор департамента программной инженерии, канд. техн. наук  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Н. А. Павлочев «\_\_\_\_»­­ \_\_\_\_\_\_\_\_\_\_ 2025 г. |

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

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

**Руководство оператора**

**ЛИСТ УТВЕРЖДЕНИЯ**

**RU.17701729.11.04-01 РО 01-1-ЛУ**

**Исполнители:**  
Студент БПИ-221  
\_\_\_\_\_\_\_\_\_\_ / Панкратов С. Ю. /  
Студент БПИ-221  
\_\_\_\_\_\_\_\_\_\_ / Знатнов Е. П. /  
«12» мая 2025 г

УТВЕРЖДЕН

RU.17701729.11.04-01 РО 01-1-ЛУ

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

**Руководство программиста**

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

**RU.17701729.11.04-01 РП 01-1**

**Листов ??**

## АННОТАЦИЯ

Настоящий документ «Программный комплекс для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях. Руководство оператора» предназначен для ознакомления пользователей с графическим интерфейсом и функциональными возможностями программного комплекса "MSnOutliers".

Руководство оператора разработано в соответствии с ГОСТ 19.505-79 «Единая система программной документации. Руководство оператора» и содержит следующие разделы:

В разделе «Назначение программы» приведена информация о функциональном назначении программного комплекса, его возможностях и особенностях использования в области исследования влияния аномальных наблюдений на регрессионные модели.

В разделе «Условия выполнения программы» представлены требования к аппаратному и программному обеспечению, необходимому для корректной работы комплекса.

Раздел «Выполнение программы» содержит последовательное описание действий оператора при работе с программным комплексом, включая создание моделей, генерацию данных, запуск анализа и интерпретацию результатов.

В разделе «Сообщения оператору» приведены тексты информационных и предупреждающих сообщений, которые могут возникать в процессе работы с программой, а также рекомендации по действиям оператора в каждой из описанных ситуаций.

Настоящий документ предназначен для пользователей программного комплекса, имеющих базовые навыки работы с компьютером и общее представление о регрессионном анализе данных и методах обнаружения аномалий.

**СОДЕРЖАНИЕ**

**[ГЛОССАРИЙ 2](#_ГЛОССАРИЙ)**

[**1. НАЗНАЧЕНИЕ ПРОГРАММЫ 4**](#_НАЗНАЧЕНИЕ_ПРОГРАММЫ)

[1.1. Функциональное назначение](#_Наименование_программы) **[4](#_Наименование_программы)**

[1.2. Эксплуатационное назначение](#_Эксплуатационное_назначение_1) **[4](#_Эксплуатационное_назначение_1)**

**[2. УСЛОВИЯ ВЫПОЛНЕНИЯ ПРОГРАММЫ 6](#_УСЛОВИЯ_ВЫПОЛНЕНИЯ_ПРОГРАММЫ)**

[2.1. Требования к составу и параметрам технических средств](#_Требования_к_составу) **[6](#_Требования_к_составу)**

[2.2. Требования к пользователю (программисту)](#_Требования_к_пользователю) **[6](#_Требования_к_пользователю)**

**[3. ВЫПОЛНЕНИЕ ПРОГРАММЫ 8](#_ВЫПОЛНЕНИЕ_ПРОГРАММЫ)**

[3.1. Установка продукта](#_Установка_продукта) **[8](#_Установка_продукта)**

[3.2. Процесс сбора данных **8**](#_Процесс_сбора_данных)

[3.3. “Препроцессинг” данных](#_Процесс_предобработки_данных) **[9](#_Процесс_предобработки_данных)**

[3.4. Процесс обучения моделей регрессии и визуализации результатов](#_Процесс_обучения_моделей) **[9](#_Процесс_обучения_моделей)**

[3.5. Процесс обучения моделей кластеризации и визуализации результатов](#_Процесс_обучения_моделей_1) **[10](#_Процесс_обучения_моделей_1)**

**[СПИСОК ИСПОЛЬЗУЕМЫЙ ЛИТЕРАТУРЫ 11](#_СПИСОК_ИСПОЛЬЗУЕМОЙ_ЛИТЕРАТУРЫ)**

[**ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ 12**](#_ЛИСТ_РЕГИСТРАЦИИ_ИЗМЕНЕНИЙ)

# НАЗНАЧЕНИЕ ПРОГРАММЫ

## Функциональное назначение

Программный комплекс для исследования влияния аномальных наблюдений на точность прогнозирования в регрессионных моделях (коротко, "MSnOutliers") предназначен для исследования влияния аномальных наблюдений на точность прогнозирования в различных регрессионных моделях. Комплекс позволяет проводить сравнительный анализ устойчивости регрессионных методов к наличию аномалий в данных, а также оценивать эффективность различных алгоритмов машинного обучения для обнаружения и удаления выбросов.

Программный комплекс обеспечивает выполнение следующих функций:

1. Работа с регрессионными моделями:
   1. Создание и настройка классических методов регрессии (метод наименьших квадратов - LSM).
   2. Создание и настройка робастных методов регрессии (метод Хьюбера - HUB, метод Тьюки - TUK).
   3. Создание и настройка метода наименьших абсолютных отклонений (LAD).
   4. Создание и настройка метода медианных наклонов Тейла-Сена (THS).
2. Работа с данными:
   1. Генерация синтетических данных с заданным количеством признаков и наблюдений.
   2. Добавление контролируемого уровня шума различных распределений и контролируемых искажений.
   3. Загрузка внешних наборов данных в формате CSV.
3. Очистка данных от аномалий:
   1. Применение алгоритма Isolation Forest (IForest).
   2. Применение алгоритма плотностной кластеризации (DBSCAN).
   3. Применение алгоритма оценки плотности ядра (KDE).
   4. Применение алгоритма k-ближайших соседей (KNN).
4. Анализ результатов:
   1. Проведение экспериментов с различными уровнями шума.
   2. Визуализация зависимости ошибки от уровня шума для каждой модели.
   3. Сравнение эффективности различных методов очистки данных.

## Возможности программы

Программный комплекс "MSnOutliers" реализует следующие возможности:

1. Создание и настройка различных типов регрессионных моделей через удобный графический интерфейс.
2. Генерация синтетических данных с контролируемым уровнем и типом аномалий.
3. Применение алгоритмов машинного обучения для обнаружения и удаления аномальных наблюдений.
4. Параллельное выполнение экспериментов для ускорения процесса анализа.
5. Визуализация результатов в виде информативных графиков.
6. Сохранение конфигураций моделей и результатов анализа для последующего использования.

## Целевая аудитория и область применения

Программный комплекс "MSnOutliers" предназначен для школьников и студентов, проходящих подготовку по дисциплине “Математическая статистика” или каким-либо смежным с ней, а также преподавателей, читающих вышеупомянутые курсы.

Комплекс может использоваться как для образовательных целей, так и для практического применения при выборе наиболее подходящих методов регрессии и очистки данных для конкретных задач анализа.

# УСЛОВИЯ ВЫПОЛНЕНИЯ ПРОГРАММЫ

## Аппаратные средства, необходимые для функционирования программы

Для эксплуатации программного комплекса "MSnOutliers" должны использоваться следующие средства вычислительной техники (СВТ):

1. СВТ индивидуального пользования - АРМ оператора, представляющее собой ПЭВМ со следующими минимальными характеристиками:
   1. Процессор: Intel Core i5 6-го поколения / AMD Ryzen 5 1-го поколения / Apple Silicon (M1/M2/M3) или аналогичный.
   2. Оперативная память: не менее 8 ГБ.
   3. Свободное дисковое пространство: не менее 500 МБ.
   4. Видеоадаптер: с поддержкой разрешения не менее 1280x720.
2. Рекомендуемые характеристики для комфортной работы при обработке больших наборов данных:
   1. Процессор с поддержкой многопоточности (4+ ядер).
   2. Оперативная память: 16 ГБ и выше.
   3. SSD-накопитель для ускорения операций чтения/записи файлов.

## Программное обеспечение, необходимое для функционирования программы

Для эксплуатации программного комплекса "MSnOutliers" необходимо следующее программное обеспечение, устанавливаемое на АРМ оператора:

1. Операционные системы (один из вариантов)
   1. Microsoft Windows версии не ниже 10.
   2. Apple macOS версии не ниже 11 "Big Sur".
   3. Linux (дистрибутивы Ubuntu 20.04 LTS и выше, Fedora 34 и выше, или другие с поддержкой Qt 6.x).
2. Библиотеки и зависимости:
   1. Qt версии 6.2.3 или выше.
   2. Библиотека Eigen версии 3.4.0 или выше.
   3. Библиотека nlohmann/json версии 3.11.2 или выше.
   4. Python версии 3.9 или выше с библиотеками matplotlib версии 3.5.1 или выше и numpy версии 1.22.3 или выше.
3. Инструменты сборки (для установки из исходного кода):
   1. CMake версии 3.14 или выше.
   2. Компилятор с поддержкой C++17 (в зависимости от операционной системы)
      1. GCC 9.0 или выше (для Linux).
      2. Clang 13.0.0 или выше (для macOS).
      3. MSVC 2019 или выше (для Windows).

## Требования к персоналу

К эксплуатации программного комплекса "MSnOutliers" допускаются операторы:

1. Имеющие навыки работы с ПЭВМ и полностью освоившие графический пользовательский интерфейс операционной системы Microsoft Windows, macOS или Linux.
2. Обладающие базовыми знаниями в области регрессионного анализа данных и понимающие принципы работы с различными регрессионными моделями.
3. Ознакомленные с документацией программного комплекса "MSnOutliers" и принципами его функционирования.
4. Понимающие базовые концепции обнаружения аномалий в данных и методы устойчивой регрессии.

# ВЫПОЛНЕНИЕ ПРОГРАММЫ

## Установка продукта

Для запуска разработанных программ, необходимо их загрузить по ссылке на репозиторий: <https://github.com/Octopupu5/Fpsattmama>.

Для получения всех результатов, необходимо запустить программы в следующем порядке:

1. Parse.ipynb
2. Preproccesing.ipynb
3. ModelsLoader.ipynb для обучения моделей регрессии

или

Clusterization.ipynb для обучения моделей кластеризации

При выполнении этих ноутбуков будут создаваться файлы формата ".xlsx" и ".cbm". Убедитесь, что они сохраняются в той же папке, что и файлы ".ipynb". Если нужно, можно воспользоваться готовыми файлами с данными из репозитория, чтобы ускорить процесс сбора данных.

Для более подробного понимания кода рекомендуется изучить документацию по используемым библиотекам: requests, BeautifulSoup, pandas.

## Процесс сбора данных

Чтобы запустить процесс сбора данных, откройте ноутбук "Parse.ipynb" и выполните следующие шаги:

1. Импортируйте необходимые библиотеки: requests, BeautifulSoup, re, pandas, numpy, tqdm, warnings.
2. Начните процесс парсинга данных об игроках с веб-сайта transfermarkt, используя библиотеки requests и BeautifulSoup.
3. Объедините данные об игроках в один датафрейм с помощью библиотеки pandas.
4. Сохраните полученный датафрейм в формате “.xlsx” файла.

После запуска процесса, подождите, пока будут собраны все данные. После завершения, датафрейм будет автоматически сохранен в формате ".xlsx".

## “Препроцессинг” данных

Шаги запуска процесса предобработки данных описаны в ноутбуке "Preprocessing.ipynb" следующим образом:

1. Начинается с импорта необходимых библиотек: pandas, datetime, relativedelta, numpy, ast, re, seaborn.
2. Происходит предварительная обработка данных, полученных на предыдущем этапе. Включает в себя замену пропущенных значений и преобразование признаков.
3. По завершении всех предобработок, данные автоматически сохраняются в формате .xlsx.

Инструкции для запуска процесса:

1. Нажмите кнопку "Run All Cells" в верхней панели для запуска процесса.
2. Дождитесь завершения предварительной обработки данных. Время выполнения зависит от объема данных.
3. По завершении предобработки, файл данных будет автоматически сохранен в формате ".xlsx".
4. Проверьте наличие сохраненного файла в директории вашего проекта.

Для лучшего понимания кода и его компонентов рекомендуется ознакомиться с документацией по соответствующим библиотекам: pandas, datetime, numpy, seaborn.

## Процесс обучения моделей регрессии и визуализации результатов

Шаги запуска процесса обучения моделей можно пронаблюдать в ноутбуке "ModelsLoader.ipynb" следующим образом:

1. Начинается с импорта необходимой библиотеки: models. В ней содержатся функции для обучения моделей и их записи в директорию.
2. При необходимости обучении моделей необходимо раскоментировать указанный код и запустить первую ячейку с кодом.
3. Следующие ячейки используются с полученными обученными моделями и демонстрируют их все метрики качества и визуализируют полученные результаты.

Для лучшего понимания кода и его компонентов рекомендуется ознакомиться со следующей библиотекой: models.

## Процесс обучения моделей кластеризации и визуализации результатов

Шаги запуска процесса обучения моделей кластеризации можно пронаблюдать в ноутбуке "Clusterization.ipynb" следующим образом:

1. Начинается с импорта необходимых библиотек: pandas, yellowbrick, sklearn, matplotlib, plotly.
2. Следующие ячейки используются для обучений моделей кластеризации и визуализации полученных результатов.

Для лучшего понимания кода и его компонентов рекомендуется ознакомиться со следующей библиотекой: pandas, yellowbrick, sklearn, matplotlib, plotly.

# СООБЩЕНИЯ ОПЕРАТОРУ

## СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

1. Библиотека "Qt" [Электронный ресурс] / The Qt Company – Режим доступа: <https://www.qt.io/>, свободный.
2. Библиотека "Eigen" [Электронный ресурс] / Eigen Development Team – Режим доступа: <https://eigen.tuxfamily.org/>, свободный.
3. Библиотека "nlohmann/json" [Электронный ресурс] / Niels Lohmann – Режим доступа: <https://github.com/nlohmann/json>, свободный.
4. Библиотека "matplotlib" [Электронный ресурс] / Matplotlib Development Team – Режим доступа: [https://matplotlib.org](https://matplotlib.org/), свободный.
5. Библиотека "numpy" [Электронный ресурс] / NumPy Development Team – Режим доступа: <https://numpy.org/>, свободный.
6. Система сборки "CMake" [Электронный ресурс] / Kitware Inc. – Режим доступа: <https://cmake.org/>, свободный.

# ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ

| Лист регистрации изменений | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Номера листов (страниц) | | | | | Всего листов (страниц в докум.) | № документа | Входящий № сопроводительного докум. и дата | Подп. | Дата |
| Изм. | Измененных | Замененных | Новых | Аннулированных |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |