**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 программиста**

**ЛИСТ УТВЕРЖДЕНИЯ**

**RU.17701729.11.04-01 РП 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 РП 03-1-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 ??**

# ГЛОССАРИЙ

1. БД - база данных, система, которая хранит и обрабатывает информацию.
2. Набор данных - объем информации, собранный и подготовленный для анализа в рамках приложения, содержащий данные о футбольных игроках и их игровых характеристиках.
3. Параметры - факторы или атрибуты, используемые в модели для прогнозирования цены трансфера футболиста.
4. Алгоритм машинного обучения - процедура, которую компьютер применяет для создания модели на основе данных. В данном контексте алгоритм машинного обучения используется для прогнозирования цены трансфера футболиста.
5. Значимость параметров - мера важности каждого параметра для прогнозов модели.
6. Трансферный рынок - область в футбольной индустрии, где клубы могут приобретать или продавать права на игроков.
7. Цена трансфера - предполагаемая стоимость игрока, за которую клубы могут приобрести его права у текущего клуба.
8. Статистические показатели - количественные данные, отражающие производительность футболиста на поле, такие как количество голов, передач, пробежек и другие метрики.

**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

"Статистика футболистов и трансферный рынок: математический анализ" предоставляет возможность предсказать оценку стоимости трансфера футболиста, используя информацию о его физическом состоянии и статистике, собранной во время последних выступлений.

Этот сервис обеспечивает следующие возможности:

1. Сбор всей необходимой информации о профессиональных футболистах из различных источников (в большинстве случаев, “transfermarkt”) для ее дальнейшего анализа.
2. Предварительную обработку данных, коррекцию ошибок, заполнение пропущенных значений и нормализацию данных.
3. Обучение ML-модели на основе собранных данных с использованием подходящего алгоритма, тюнинг гиперпараметров для достижения наилучших результатов и оценки качеств построенных моделей с использованием различных метрик.
4. Визуализацию данных и результатов с использованием современных инструментов (“matplotlib”, “plotly”).
5. Разработка отчетов, основанных на проведенном анализе, которые содержат детальное описание выбранной методологии и полученных результатов.

Этот инструмент представляет собой сервис, предназначенный для ведения статистики, изучения и предсказывания цен на трансферы футболистов. Он предоставляет пользователям все инструменты, необходимые для принятия обоснованных футбольных бизнес-решений.

## Эксплуатационное назначение

Эксплуатационное назначение программы "Статистика футболистов и трансферный рынок: математический анализ" заключается в предоставлении спортивным менеджерам, сотрудникам футбольных клубов и аналитикам важных данных и аналитических инструментов для принятия обоснованных решений в сфере трансферной политики.

Этот программный инструмент облегчает процесс оценки справедливой цены игрока, основываясь на его статистических показателях, что ускоряет процесс проведения трансферов. Кроме того, благодаря прогнозам и данным о футболистах, комплексный анализ их статистики, физических параметров и игровых характеристик помогает клубам привлекать талантливых игроков, соответствующих игровой философии и стратегии клуба, и избегать неудачных и затратных приобретений.

Программное средство предоставляет не только прогнозы и данные для исследования игроков, но и глубокий анализ их игровой деятельности, включая влияние на результаты матчей. Этот анализ помогает выявить не только сильные и слабые стороны игроков, но и определить их потенциальную ценность для команды. Такой аналитический подход позволяет разрабатывать эффективные стратегии для улучшения результатов всей команды и повышения ее конкурентоспособности.

# УСЛОВИЯ ВЫПОЛНЕНИЯ ПРОГРАММЫ

## Требования к составу и параметрам технических средств

Для надёжной и корректной работы продукта требуется устройство (персональный компьютер), обладающий следующими характеристиками:

1. Процессор с 4 или более ядрами (AMD Ryzen 5 / Intel Core i5 или другие).
2. Оперативная память – 8 ГБ или более.
3. Операционная система – Windows (7, 8, 10, 11), любой дистрибутив Linux, MacOS.
4. Свободное дисковое пространство – 8ГБ или более.
5. Установленный язык программирования Python версии 3.11 и выше.
6. Установленные для языка программирования Python пакеты (библиотеки):  
   “beautiful soup” – версии 4.12 и выше; “requests” – версии 2.30 и выше; “joblib” – версии 1.3 и выше; “matplotlib” – версии 3.7 и выше; “numpy” – версии 1.25 и выше ; “optuna” – версии 3.6 и выше; “pandas” – версии 2.0 и выше; “plotly” – версии 5.18 и выше; “scikit-learn” – версии 1.3 и выше; “yellowbrick” – версии 1.5 и выше.
7. Среда разработки для запуска программ – поддерживающая работу с файлами “Python notebook” формата “.ipynb”: Visual Studio Code, PyCharm, Jupyter Notebook и т.п.
8. В процессе установки языка программирования, необходимых пакетов, среды разработки или сбора данных (статистик футбольных игроков) устройство должно иметь доступ к сети Интернет.
9. Приложение для работы с электронными таблицами: Google Sheets, Microsoft Excel, LibreOffice Calc и т.п.

## Требования к пользователю (программисту)

Для эффективной работы над проектом требуется программист, владеющий языком программирования Python версии 3.0 или более поздней. Рекомендуется наличие опыта работы с указанными библиотеками. Знание основ машинного обучения, таких как алгоритмы классификации и регрессии, а также понимание ключевых концепций вроде переобучения и кросс-валидации, будет являться преимуществом. Кроме того, необходим опыт работы с средами разработки, поддерживающими файлы формата ".ipynb", а также умение работать с таблицами и базовыми функциями работы с ячейками или аналогичными программами. Пользователи, с которыми будет взаимодействовать программист, должны обладать базовыми навыками работы с компьютером и уметь использовать специализированные приложения для просмотра данных и графиков. Также им требуется базовое понимание данных и результатов, полученных при работе с моделями машинного обучения в рамках проекта.

# 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.

## СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

1. Библиотека "pandas" [Электронный ресурс] / Python Software Foundation – Режим доступа: https://pandas.pydata.org, свободный.
2. Библиотека "scikit-learn" [Электронный ресурс] / Scikit-learn developers – Режим доступа: https://scikit-learn.org/stable, свободный.
3. Библиотека "beautiful soup" [Электронный ресурс] / Crummy – Режим доступа: https://www.crummy.com/software/BeautifulSoup, свободный.
4. Библиотека "optuna" [Электронный ресурс] / Optuna Developers – Режим доступа: https://optuna.org, свободный.
5. Библиотека "matplotlib" [Электронный ресурс] / Matplotlib Development Team – Режим доступа: https://matplotlib.org, свободный.
6. Библиотека "plotly" [Электронный ресурс] / Plotly Technologies Inc. – Режим доступа: https://plotly.com/python, свободный.
7. Библиотека "joblib" [Электронный ресурс] / Joblib developers – Режим доступа: https://joblib.readthedocs.io/en/stable, свободный.
8. Библиотека "yellowbrick" [Электронный ресурс] / The scikit-yb developers – Режим доступа: https://www.scikit-yb.org/en/latest, свободный.
9. Библиотека "numpy" [Электронный ресурс] / Numpy community – Режим доступа: https://numpy.org/doc/stable, свободный.
10. Библиотека "requests" [Электронный ресурс] / Cory Benfield, Ian Stapleton Cordasco, Nate Prewitt – Режим доступа: https://requests.readthedocs.io/en/stable/, 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 |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |