**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. | СОГЛАСОВАНО  Научный руководитель, сотрудник факультета компьютерных наук, ассистент департамента анализа данных и искусственного интеллекта  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ А. О. Боревский «\_\_\_\_»­­ \_\_\_\_\_\_\_\_\_\_ 2024 г. | УТВЕРЖДАЮ  Академический руководитель образовательной программы «Программная инженерия»  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Н. А. Павлочев «\_\_\_\_»­­ \_\_\_\_\_\_\_\_\_\_ 2024 г. |



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

**Программа и методика испытаний**

**ЛИСТ УТВЕРЖДЕНИЯ**

**RU.17701729.11.04-01 ПМИ 03-1-ЛУ**

**Исполнитель:**  
Студент БПИ-221  
\_\_\_\_\_\_\_\_\_\_ / Панкратов С. Ю. /  
«01» апреля 2024 г

УТВЕРЖДЕН

RU.17701729.11.04-01 ПМИ 03-1-ЛУ

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

**Программа и методика испытаний**

**RU.17701729.11.04-01 ПМИ 03-1**

**Листов ??**

# АННОТАЦИЯ

Программа и методика испытаний — это документ, в котором содержится информация о программном продукте, а также полное описание приемочных испытаний для данного программного продукта.

Настоящая Программа и методика испытаний для «Статистика футболистов и трансферный рынок: математический анализ » содержит следующие разделы: «Объект испытаний», «Цель испытаний», «Требования к программе», «Требования к программной документации», «Средства и порядок испытаний», «Методы испытаний», «Приложения».

В разделе «Объект испытаний» указано наименование, краткая характеристика и назначение программы.

В разделе «Цель испытаний» указана цель проведения испытаний.

Раздел «Требования к программе» содержит основные требования к программе, которые подлежат проверке во время испытаний (требования к функционалу).

Раздел «Требования к программным документам» содержит состав программной документации, которая представляется на испытания.

Раздел «Средства и порядок испытаний» содержит информацию о технических и программных средствах, которые следует использовать во время испытаний, а также порядок этих испытаний.

Раздел «Методы испытаний» содержит информацию об используемых методах испытаний.

Настоящий документ разработан в соответствии с требованиями:

1. ГОСТ 19.101-77 Виды программ и программных документов [1];
2. ГОСТ 19.102-77 Стадии разработки [2];
3. ГОСТ 19.103-77 Обозначения программ и программных документов [3];
4. ГОСТ 19.104-78 Основные надписи [4];
5. ГОСТ 19.105-78 Общие требования к программным документам [5];
6. ГОСТ 19.106-78 Требования к программным документам, выполненным печатным способом [6];
7. ГОСТ 19.201-78 Программа и методика испытаний. Требования к содержанию и оформлению [10].

Изменения к данному Техническому заданию оформляются согласно   
ГОСТ 19.603-78 [8], ГОСТ 19.604-78 [9].

# 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**

[**АННОТАЦИЯ 2**](#_АННОТАЦИЯ)

**[ГЛОССАРИЙ 4](#_ГЛОССАРИЙ)**

**[1. ОБЪЕКТ ИСПЫТАНИЙ 6](#_ОБЪЕКТ_ИСПЫТАНИЙ)**

[1.1. Наименование программы](#_Наименование_программы) **[6](#_Наименование_программы)**

[1.2. Краткая характеристика области применения](#_Краткая_характеристика_области) **[6](#_Краткая_характеристика_области)**

[**2. ЦЕЛЬ ИСПЫТАНИЙ 7**](#_ЦЕЛЬ_ИСПЫТАНИЙ)

**[3. ТРЕБОВАНИЯ К ПРОГРАММЕ 8](#_ТРЕБОВАНИЯ_К_ПРОГРАММЕ)**

[3.1. Требования к функциональным характеристикам **8**](#_Требования_к_функциональным)

[3.2. Требования к прочим характеристикам](#_Требования_к_прочим) **[9](#_Требования_к_прочим)**

**[4. ТРЕБОВАНИЯ К ПРОГРАММНОЙ ДОКУМЕНТАЦИИ 10](#_ТРЕБОВАНИЯ_К_ПРОГРАММНОЙ)**

[4.1. Состав программной документации](#_ТРЕБОВАНИЯ_К_ПРОГРАММНОЙ) **[10](#_ТРЕБОВАНИЯ_К_ПРОГРАММНОЙ)**

[4.2. Специальные требования к программной документации](#_Специальные_требования_к) **[10](#_Специальные_требования_к)**

**[5. СРЕДСТВА И ПОРЯДОК ИСПЫТАНИЙ 11](#_СРЕДСТВА_И_ПОРЯДОК)**

[5.1. Технические средства, используемые во время испытаний **11**](#_Технические_средства,_используемые)

[5.2. Программные средства, используемые во время испытаний **11**](#_Программные_средства,_используемые)

[5.3. Порядок проведения испытаний](#_Порядок_проведения_испытаний) **[11](#_Порядок_проведения_испытаний)**

**[6. МЕТОДЫ ИСПЫТАНИЙ 12](#_МЕТОДЫ_ИСПЫТАНИЙ)**

[6.1. Проверка требований к программной документации](#_Проверка_требований_к) **[12](#_Проверка_требований_к)**

[6.2. Проверка требований к функциональным характеристикам](#_Проверка_требований_к_1) **[12](#_Проверка_требований_к_1)**

[6.3. Проверка требований к надёжности](#_Проверка_требований_к_2) **[15](#_Проверка_требований_к_2)**

**[СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ 16](#_СПИСОК_ИСПОЛЬЗУЕМОЙ_ЛИТЕРАТУРЫ)**

[**ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ 18**](#_ЛИСТ_РЕГИСТРАЦИИ_ИЗМЕНЕНИЙ)

# ОБЪЕКТ ИСПЫТАНИЙ

## Наименование программы

Наименование программы – «Статистика футболистов и трансферный рынок: математический анализ».

Наименование программы на английском языке – «Football player statistics and the transfer market: a mathematical analysis».

Краткое наименование программы – «FootPricing».

## Краткая характеристика области применения

Аудитория продукта, который анализирует цены на трансфер футболистов, включает в себя спортивных менеджеров и агентов, работающих в футбольных клубах, специалистов по обработке данных, а также всех, кто заинтересован в этой области - от футбольных академий и скаутов до студентов и исследователей в области спортивной аналитики и простых футбольных фанатов, ищущих информацию о ценообразовании на рынке трансферов и оценке игроков на основе статистических данных.

Этот продукт позволяет менеджерам получить прогнозную цену на трансфер определенного игрока, используя его статистику и физические данные, собранные во время его выступлений за клуб и национальную команду. Это помогает управлять политикой трансферов клуба. Программа предоставляет обоснование прогнозируемой стоимости на основе используемых показателей, что упрощает проведение аналитических исследований. Например, это может помочь выявить наиболее важные факторы, влияющие на стоимость трансферов игроков, и облегчить выбор потенциальных кандидатов для клуба.

# ЦЕЛЬ ИСПЫТАНИЙ

Целью проведения испытаний является проверка на соответствие требованиям, изложенным в настоящем Техническим заданием из комплекта документации в соответствии с ЕСПД (Единой системой программной документации).

# ТРЕБОВАНИЯ К ПРОГРАММЕ

## Требования к функциональным характеристикам

## Сборка датасета

Описание: Получение информации о футболистах с веб-сайтов или при помощи API.

Ключевые аспекты:

1. Выбор источников данных и методов их сбора.
2. Обеспечение актуальности и достоверности полученных данных.
3. Преобразование данных в формат, удобный для дальнейшего использования.

Формат данных:

Данные будут сохраняться в «.xlsx» формате для последующего удобства обработки и анализа.

Средства, использованные в сборе данных:

Библиотеки “requests” и “Beautiful Soup”.

## Предобработка данных

Описание: Подготовка данных к анализу и обучению моделей.

Ключевые аспекты:

1. Очистка данных от ошибок.
2. Преобразование данных для обеспечения их однородности и соответствия требованиям моделей.
3. Обработка пропущенных значений.

Средства, использованные в предобработке данных:

Библотека “pandas”.

Необходимые условия к полученным данным:

1. Корректность: Данные должны быть корректно обработаны и готовы для использования в моделях машинного обучения.
2. Сохранение информативности: Предобработка не должна искажать информацию и должна сохранить важные характеристики данных.

## Реализация моделей машинного обучения

Описание: Создание и обучение моделей для анализа и прогнозирования характеристик футболистов.

Ключевые аспекты:

1. Выбор и настройка моделей для конкретных задач анализа данных.
2. Обучение моделей на подготовленных данных и оценка их производительности.
3. Использование обученных моделей для прогнозирования и анализа данных о футболистах.

Средства, использованные в предобработке данных:

Библотеки “scikit-learn”, “matplotlib”, “numpy”, “optuna”, “yellowbrick”, “plotly”.

Методы реализации: Использование различных алгоритмов машинного обучения, таких как линейная регрессия и модели кластеризации.

## Взаимодействие с пользователями

Описание: Обеспечение удобства работы с программой для пользователей.

Ключевые аспекты:

1. Разработка пользовательского интерфейса для ввода параметров и получения результатов анализа.
2. Предоставление информации о процессе работы программы и возможных ошибках.

Требования к интерфейсу:

1. Интуитивное взаимодействие: Интерфейс должен быть простым и интуитивно понятным для пользователей.
2. Результаты анализа должны быть представлены в доступной форме для понимания пользователей.

## Требования к прочим характеристикам

При проведении испытаний другие характеристики (возможности) программы подлежат проверке на соответствие требованиям изложенным в п.4 «Требования к программе» настоящего Технического задания.

# ТРЕБОВАНИЯ К ПРОГРАММНОЙ ДОКУМЕНТАЦИИ

* 1. **Состав программной документации**

1. «Статистика футболистов и трансферный рынок: математический анализ». Техническое задание (ГОСТ 19.201-78) [7];
2. «Статистика футболистов и трансферный рынок: математический анализ». Программа и методика испытаний (ГОСТ 19.301-78) [10];
3. «Статистика футболистов и трансферный рынок: математический анализ». Пояснительная записка (ГОСТ 19.404-79) [11];
4. «Статистика футболистов и трансферный рынок: математический анализ». Руководство программиста (ГОСТ 19.504-79) [12];
5. «Статистика футболистов и трансферный рынок: математический анализ». Текст программы (ГОСТ 19.401-78) [13];

## Специальные требования к программной документации

1. Документы к программе должны быть выполнены в соответствии с ГОСТ 19.106-78 и ГОСТами к каждому виду документа (см. п. 5.1).
2. Пояснительная записка должна быть загружена в систему Антиплагиат через LMS «НИУ ВШЭ». Лист, подтверждающий загрузку пояснительной записки, сдается в учебный офис вместе со всеми материалами не позже, чем за день до защиты курсовой работы.
3. Вся документация также воспроизводится в печатном виде, она должна быть подписана академическим руководителем образовательной программы 09.03.04 «Программная инженерия», руководителем разработки и исполнителем перед сдачей курсовой работы в учебный офис, не позже одного дня до защиты.
4. Документация также сдается в электронном виде в формате .pdf или .docx, а программа – в архиве формата .zip или .rar.
5. Все документы перед защитой курсовой работы должны быть загружены в информационно-образовательную среду НИУ ВШЭ LMS (Learning Management System) в личном кабинете, дисциплина – «Курсовой проект, 2 курс ПИ», одним архивом.

# СРЕДСТВА И ПОРЯДОК ИСПЫТАНИЙ

## Технические средства, используемые во время испытаний

Для испытаний программы используется устройство (персональный компьютер) со следующими средствами:

1. Процессор – AMD Ryzen 5 (6 ядер).
2. Оперативная память – 16 ГБ.
3. Операционная система – Windows 11.
4. Жёсткий диск: 200 ГБ свободной памяти, гарантирующий быстрый доступ к данным и программному обеспечению, ускоряя процесс тестирования и анализа.
5. Интернет-соединени со скоростью загрузки 20МБ в секунду для сбор данных с источников.

## Программные средства, используемые во время испытаний

Для испытаний программы используется среда разработки Visual Studio Code.

## Порядок проведения испытаний

Испытания должны проводиться в следующем порядке:

1. Проверка требований к программной документации.
2. Проверка требований к составу выполняемых функций.
3. Проверка требований к надежности.

Испытания проводятся на основе раздела «Методика испытаний».

# МЕТОДЫ ИСПЫТАНИЙ

## Проверка требований к технической документации

Проверка состава программной документации включает в себя анализ наличия всех необходимых документов в системе SmartLMS, а также уверенность в наличии всех требуемых подписей. Кроме того, осуществляется контроль соответствия документации стандартам ГОСТ.

В результате проверки подтверждено, что комплект документов полный и все представленные документы соответствуют установленным требованиям.

## Проверка требований к функциональным характеристикам

* + 1. **“Препроцессинг” данных**

1. Корректность конвертации возраста и высоты к числовому виду:
   * 1. **“Препроцессинг” данных**
2. Корректность конвертации возраста и высоты к числовому виду:

Убедиться, что функция для конвертации возраста корректно извлекает числовые значения возраста из строки с использованием регулярного выражения.  
Проверить, что функция для конвертации высоты корректно преобразует строки в числа, учитывая замену запятых на точки и удаление единиц измерения "м".

1. Корректность конвертации столбцов с датами:

Проверить, что функция для конвертации столбцов с датами в числовой вид правильно вычисляет количество месяцев, прошедших с указанной даты.  
Убедиться, что функция обрабатывает различные форматы дат и корректно преобразует их в числовой вид.

1. Проверка переименования столбцов и удаление ненужных столбцов:

Убедиться, что переименование столбцов в датафрейме происходит корректно и не приводит к дублированию столбцов или потере данных.  
Проверить, что столбцы, которые были удалены из датафрейма, действительно не содержат важной информации для анализа данных.

1. Корректность конвертации столбцов с ценами:

Убедиться, что функция для конвертации цен в числовой вид корректно обрабатывает различные форматы цен, включая использование суффиксов "k", "m" и "n" для обозначения тысяч, миллионов и миллиардов соответственно.  
Проверить, что функция правильно извлекает числовые значения из строк с ценами и корректно умножает их на соответствующие коэффициенты.

1. Проверка разбиения столбца с клубными статистиками:

Убедиться, что разбиение столбца с клубными статистиками на отдельные столбцы происходит корректно и не приводит к потере данных или искажению структуры датафрейма.

Проверить, что функция правильно извлекает числовые значения из словарей, представленных в виде строк, и сохраняет их в отдельные столбцы.

1. Проверка генерации признака из столбца с местами команды в лиге:

Убедиться, что генерация признака из столбца с местами команды в лиге происходит корректно и не приводит к искажению данных или потере информации.

Проверить, что функция правильно вычисляет среднее, максимальное и минимальное значение из списка мест команды в лиге и сохраняет их в отдельные столбцы.

1. Обработка пропущенных значений:

Убедиться, что пропущенные значения в датафрейме заполняются корректно, например, средними значениями или нулями, в зависимости от контекста.

Проверить, что функции обработки пропущенных значений не приводят к потере информации или искажению результатов анализа.

* + 1. **Реализация моделей машинного обучения**

1. Корректность предобработки данных:

Убедиться, что данные были корректно предобработаны перед обучением моделей. Это включает в себя проверку наличия всех необходимых признаков и целевой переменной, а также обработку пропущенных значений.

1. Проверка разбиения данных на обучающий и тестовый наборы:

Убедиться, что данные были правильно разделены на обучающий и тестовый наборы для оценки производительности моделей. Это может быть выполнено с помощью метода train\_test\_split из библиотеки scikit-learn. В файле ModelsLoader в качестве обучающий и тестовой выборок используется весь датасет.

1. Обучение моделей:

Проверить, что функции обучения моделей запускаются без ошибок и завершаются успешно. Убедиться, что параметры моделей оптимизированы для достижения наилучшей производительности.

1. Оценка производительности моделей:

Выполнить оценку производительности обеих моделей с использованием различных метрик, таких как среднеквадратическая ошибка (MSE), средняя абсолютная ошибка (MAE), коэффициент детерминации (R^2) и другие метрики.

1. Сравнение моделей:

Сравнить производительность моделей между собой, чтобы определить, какая из них лучше подходит. Это может быть выполнено путем сравнения значений метрик производительности для каждой модели.

1. Проверка сохранения моделей:

Убедиться, что обученные модели сохраняются в файлы без ошибок и могут быть успешно загружены для дальнейшего использования или развертывания.

1. Дополнительная проверка производительности:

Проверить дополнительные метрики производительности моделей, такие как средняя абсолютная процентная ошибка (MAPE) или средняя квадратическая ошибка (RMSE).

* + 1. **Реализация моделей машинного обучения**

1. Загрузка данных:

Убедиться, что данные были успешно загружены из файла Excel или другого источника данных.

1. Проверка структуры данных:

Убедиться, что данные имеют правильную структуру и не содержат пропущенных значений или других аномалий, которые могут повлиять на процесс кластеризации.

1. Определение оптимального числа кластеров с помощью метода "локтя" (Elbow method):

Визуализировать зависимость значения инерции (сумма квадратов расстояний от каждой точки до центроида ее кластера) от числа кластеров.

Найти точку на графике, где увеличение числа кластеров приводит к уменьшению инерции снижается, что позволяет выбрать оптимальное количество кластеров.

1. Оценка качества кластеризации с помощью визуализации силуэта:

Построить визуализацию силуэта для каждого значения числа кластеров.

Убедиться, что коэффициент силуэта близок к 1, что указывает на хорошую кластеризацию, и что нет значительных перекрытий между кластерами.

1. Проверка кластеризации для различных подгрупп данных:

Провести кластеризацию для различных подгрупп данных, например, для игроков с разными ценами (больше или меньше 5 миллионов), чтобы увидеть, какие структуры кластеров могут быть выявлены для различных категорий игроков.

1. Визуализация результатов кластеризации:

Построить визуализации результатов кластеризации, например, с использованием трехмерных графиков, чтобы визуально оценить разделение на кластеры.

Убедиться, что кластеры имеют отличимые характеристики и могут быть интерпретированы для дальнейшего анализа.

## Проверка требований к надёжности

После проведения анализа функциональных требований в разделе 6.2 можно заключить, что требования к надежности успешно выполнены.

## 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 19.603-78 Общие правила внесения изменений. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
9. ГОСТ 19.604-78 Правила внесения изменений в программные документы, выполненные печатным способом. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
10. ГОСТ 19.505-79 Руководство программиста. Требования к содержанию и оформлению. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
11. ГОСТ 19.404-79 Программа и методика испытаний. Требования к содержанию и оформлению. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
12. ГОСТ 19.404-79 Пояснительная записка. 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 19.301-79 Программа и методика испытаний. Требования к содержанию и оформлений. // Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
15. "Transfermarkt" [Электронный ресурс] / Transfermarkt GmbH & Co. KG – Режим доступа: https://www.transfermarkt.com, 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 |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |