Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ Информатика и системы управления

КАФЕДРА Программное обеспечение ЭВМ и информационные технологии

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ

НА ТЕМУ:

Метод прогнозирования исходов противоположных событий на основе регрессионного анализа.

Студент ИУ7-84		**Кизилов Д.В.**
(Группа)	(Подпись, дата)	(И.О.Фамилия)
Руководитель ВКР		**Корниенко В.В.**
	(Подпись, дата)	(И.О.Фамилия)
Консультант		
	(Подпись, дата)	(И.О.Фамилия)
Консультант		
	(Подпись, дата)	(И.О.Фамилия)
Нормоконтролер		**Строганов Ю.В.**
	(Подпись, дата)	(И.О.Фамилия)

2019 г.

РЕФЕРАТ

Отчет содержит 49 стр., 6 рис., 7 табл., 18 источн., 1 прил.

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ, СМЕШАННЫЕ ЕДИНОБОРСТВА, ПРОГНОЗИРОВАНИЕ, ЭКСТРАПОЛЯЦИЯ

Объектом исследования является статистические данные смешанных единоборств. Объект разработки – метод прогнозирования, используемый для вычисления вероятности победы первого спортсмена. Цель работы – разработка и реализация программного метода прогнозирования исходов проитвоположных событий на основе регрессионного анализа. Задачи, решаемые в работе:

— анализ статистические данные смешанных единоборств;

— анализ методов построения регрессионных моделей;

— подготовка входных данных;

— разрботка и исследование метода прогнозирования.

Область применения – системы прогнозирования. В первой части работы проводится анализ предметной области и методов, необходимых для достижения цели. Во второй части описываются алгоритмы и структуры данных. В третьей части описываются технологии, примененные при реализации алгоритмов. В четвертой части проведены экспериментальные исследования точностных характеристик метода.

Предлагаемые направления развития:

— применение методов машинного обучения;

— реализация web-приложения для большего охвата пользователей;

— повышение точности за счёт составления более детализированных выборок.

Поставленная цель была достигнута: метод прогнозирования исходов событий разработан, реализован и проверен на практике. Были рассмотрены существующие недостатки метода и предложены пути дальнейшего развития.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В последние годы, использование статистического анализа спортивных данных стало более популярно, чем когда-либо. Анализируют как командные виды спорта: футбол[13], хоккей[14], баскетбол[15], так и индивидуальные: автогонки[16], теннис[12] и другие.

Спорт смешанных единоборств, на сегодняшний день не имеет статистического анализа. Смешанные боевые искусства относительно молодой спорт. Первое событие UFC (с англ. Абсолютный бойцовский чемпионат), который отмечают началом MMA, как спорта высоких достижений, в Северной Америке провели в 1993 году, 26 лет назад. За это время, спорт прошёл путь от смешения разновидных стилей борьбы и минимальным количеством правил до всемирно известного спорта с преданными поклонниками[17].

Анализ событий и прогнозирования исходов при помощи методов статистического анализа слабо распространены, даже среди преданных поклонников, несмотря на наличие чётко оцениваемых показателей и характеристик в данном виде спорта. Прогнозоривание исходов событий в смешанных единоброствах может помочь спортсменам и тренерам увидеть статистические составлющие этого спорта, и улучшить свои результаты в нём.

1 Аналитический раздел

В данном разделе рассмотрен анализ существующих методов построения регрессионных моделей и экстраполяции, описана предметная область. Также дано определение противоположных событий.

Объектом исследования является статистические данные смешанных единоборств. Объект разработки – метод прогнозирования, используемый для вычисления вероятности победы первого спортсмена.

1.1 Анализ предметной области

1.1.1 Смешанные боевые искусства

Смешанные боевые искусства (также ММА – от англ. Mixed Martial Arts) – боевые искусства (часто неверно называемые "боями без правил"), представляющие собой сочетание множества техник, школ и направлений единоборств. ММА является полноконтактным боем с применением ударной техники и борьбы как в стойке (клинч), так и на полу (партер).

Данный вид спорта допускает использование технических приемов из арсенала различных единоборств, включая удары руками и ногами, броски, захваты, борьбу в стойке и в партере. Поединки проходят в полный контакт с использованием минимума защитных средств. В различных турнирах принята круговая система соревнований, система с выбыванием или "смешанная схема". Правила проведения боёв взяты из официального источника [3]. Во всем мире ежегодно проводится более 100 состязаний по боям без правил (под эгидой соответствующих организаций): такие, как UFC (в США) [1], "2 Hot 2 Handle"(в Голландии), "Pride"(в Японии), турниры по вале тюдо (в Бразилии), М-1 (в России) [2] и др. Упомянутые страны относятся к числу мировых лидеров в этом виде спорта. Помимо (принятого у нас в стране) термина "бои без правил"используются также названия "микс файт "фрифайт "смешанные единоборства "абсолютные поединки "панкратион"и др. – для обозначения вида спорта в целом и/или отдельных его разновидностей, что иногда приводит к терминологической путанице.

1.1.2 Весовые категории

Ввиду распространения знаний техники приемов, разница в весе стала значительным фактором. В сегодняшнем ММА девять весовых категорий, представленных в главной спортивной организации UFC [3]:

— от 120 кг, Супертяжёлый вес;

— 93 – 120 кг, Тяжёлый вес;

— 84 – 93 кг, Полутяжёлый вес;

— 77 – 84 кг, Средний вес;

— 70 – 77 кг, Полусредний вес;

— 66 – 70 кг, Лёгкий вес;

— 61 – 66 кг, Полулёгкий вес;

— 57 – 61 кг, Легчайший вес;

— до 57 кг, Наилегчайший вес;

Потребность в делении бойцов на весовые категории была обусловлена глобальным процессом универсализации атлетов, когда каждый из них начал совершенствовать себя во всех стилях. Весовые категории в UFC изначально имели совершенно не такой вид, как сегодня. Ещё в 1997 году был проведён турнир UFC 12, где были представлены только два весовых эшелона — тяжёлый и лёгкий. В первый входили все атлеты, вес которых превышал 200 фунтов, во второй — те спортсмены, которые весили меньше 199 фунтов.

1.1.3 Исход матча

Состязания согласно регламенту [4] могут заканчиваются следующим образом:

— Добровольная сдача (англ. Submission): боец отчётливо постукивает открытой ладонью или пальцами по мату или оппоненту. Также допустима вербальная (словесная) сдача.

— Нокаут (англ. KO): боец оказывается в бессознательном состоянии в результате разрешённого удара.

— Технический нокаут (англ. TKO): остановка боя третьим лицом в результате потери одним из бойцов способности продолжать бой. Технические нокауты могут подразделяться на три категории:

а) Остановка рефери (англ. Referee stoppage). Рефери решает, что один из бойцов не может осмысленно защищаться, и останавливает бой.

б) Остановка врачом (врач, присутствующий возле ринга, решает, что дальнейшее участие одного из бойцов ставит жизнь или здоровье этого участника под угрозу. Например, травмы или обильное кровотечение).

в) Остановка "углом"(англ. Corner stoppage). Угловой секундант бойца сигнализирует об остановке боя.

— Судейское решение (англ. Decision). В зависимости от подсчёта баллов, бой может закончиться:

а) Единогласным решением (англ. Unanimous decision). Все три судьи отдают предпочтение бойцу А.

б) Решением большинства (англ. Split decision). Двое судей присуждают победу бойцу А, один судья присуждает ничью.

в) Раздельным решением (англ. Majority decision). Двое судей присуждают победу бойцу А, один судья присуждает победу бойцу Б.

г) Единогласной ничьей (англ. Draw). Трое судей присуждают ничью.

д) Ничьей решением большинства (англ. Majority draw) Двое судей присуждают ничью, один – победу.

е) Раздельной ничьей (англ. Majority split). Один судья присуждает победу бойцу А, один судья присуждает победу бойцу Б, один судья присуждает ничью.

— Бой также может закончиться техническим решением, дисквалификацией, отменой, технической ничьей или признанием боя не состоявшимся (англ. no contest). Последние два варианта не имеют победителя.

Исходные данные о спорте: в одном состязании всегда участвуют только два спортсмена. Результатом боя является победа одного из спортсменов, либо ничья.

Все спортсмены выступают в одном виде спорта, являющимся по виду спортивного состязания единоборством.

Единоборство — вид спортивного состязания, в котором два участника физически противодействуют друг другу с целью выявить победителя в схватке, используя либо только физическую силу, либо также различные спортивное снаряжение и/или ручное холодное оружие.

1.2 Регрессионный анализ

Регрессионный анализ — метод моделирования измеряемых данных и исследования их свойств. Данные состоят из пар значений зависимой переменной (переменной отклика) и независимой переменной (объясняющей переменной). Регрессионная модель есть функция независимой переменной и параметров с добавленной случайной переменной.

Параметры модели настраиваются таким образом, что модель наилучшим образом приближает данные. Критерием качества приближения (целевой функцией) обычно является среднеквадратичная ошибка: сумма квадратов разности значений модели и зависимой переменной для всех значений независимой переменной в качестве аргумента.

Предполагается, что зависимая переменная есть сумма значений некоторой модели и случайной величины. Относительно характера распределения этой величины делаются предположения, называемые гипотезой порождения данных.

Для подтверждения или опровержения этой гипотезы выполняются статистические тесты, называемые анализом остатков. При этом предполага-

ется, что независимая переменная не содержит ошибок. Регрессионный анализ используется для прогноза, анализа временных рядов, тестирования гипотез и выявления скрытых взаимосвязей в данных.

1.2.1 Определение регрессионного анализа

Регрессия — зависимость математического ожидания (например, среднего значения) случайной величины от одной или нескольких других случайных величин (свободных переменных), то есть

$$M(y \mid x) = f(x). \tag{1.1}$$

Регрессионным анализом называется поиск такой функции f которая описывает эту зависимость. Регрессия может быть представлена в виде суммы неслучайной и случайной составляющих.

$$y = f(x) + \epsilon \tag{1.2}$$

где f — функция регрессионной зависимости, а ϵ - это аддитивная случайная величина с нулевым матожиданием.

Предположение о характере распределения этой величины называется гипотезой порождения данных. Обычно предполагается, что величина ϵ имеет гауссово распределение с нулевым средним и дисперсией σ_ϵ^2.

Задача нахождения регрессионной модели нескольких свободных переменных ставится следующим образом. Задана выборка — множество $\{x_1,...,x_N \mid x \in R\}$ значений свободных переменных и множество $\{y_1,...,y_N \mid x \in R\}$ соответствующих им значений зависимой переменной. Эти множества обозначаются как D, множество исходных данных $\{(x,y)_i\}$.

1.2.2 Линейная регрессия

Линейная регрессия — используемая в статистике регрессионная модель зависимости одной (объясняемой, зависимой) переменной y у от другой или нескольких других переменных (факторов, регрессоров, независимых переменных) x с линейной функцией зависимости.

Регрессионная модель

$$y = f(x,b) + \epsilon, M(\epsilon) = 0 \tag{1.3}$$

где b — параметры модели, ϵ — случайная ошибка модели; называется линейной регрессией, если функция регрессии $f(x,b)$ имеет вид

$$f(x,b) = b_0 + b_1x_1 + b_2x_2 + ... + b_kx_k, \tag{1.4}$$

где b_j — параметры (коэффициенты) регрессии, x_j — регрессоры (факторы модели), k — количество факторов модели. Параметр b_0, при котором нет факторов, называют часто константой. Формально — это значение функции при нулевом значении всех факторов. Для аналитических целей удобно считать, что константа — это параметр при «факторе», равном 1 (или другой произвольной постоянной, поэтому константой называют также и этот «фактор»). В таком случае, если перенумеровать факторы и параметры исходной модели с учетом этого (оставив обозначение общего количества факторов — k), то линейную функцию регрессии можно записать в следующем виде, формально не содержащем константу:

$$f(x,b) = b_0 + b_1x_1 + b_2x_2 + ... + b_kx_k = \sum_{0}^{k} b_jx_j = x^Tb, \tag{1.5}$$

где $x^T = (x_1,x_2,...,x_k)$ — вектор регрессоров, $b = (b_1,b_2,\ldots,b_k)^T)$ — вектор-столбец параметров (коэффициентов).

Линейная модель может быть как с константой, так и без константы. Тогда в этом представлении первый фактор либо равен единице, либо является обычным фактором соответственно.

1.2.3 Модели с дискретным выбором

Линейная регрессия обычно используется, когда переменная отклика является непрерывной и имеет неограниченный диапазон. Часто переменная ответа может быть не непрерывной, а скорее дискретной. Если зависимая переменная является дискретной, то применимы следующие методы: логистическая регрессия, полиномиальная логитная и пробитная модели. Логистическая

регрессия и пробитные модели используются, когда зависимая переменная является двоичной.

1.2.3.1 Логистическая регрессия

В условиях классификации назначение вероятностей исхода для наблюдений может быть достигнуто с помощью логистической модели, которая в основном представляет собой метод, который преобразует информацию о двоичной зависимой переменной в неограниченную непрерывную переменную и оценивает регулярную многомерную модель.

Делается предположение о том, что вероятность наступления события $y = 1$ равна:

$$P\{y = 1 \mid x\} = f(z), \tag{1.6}$$

где $z = \beta^T x = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$, x и β — векторы-столбцы значений независимых переменных $1, x_1, \ldots, x_n$ и параметров (коэффициентов регрессии) — вещественных чисел β_0,...,β_n, соответственно, а $f(z)$ — так называемая логистическая функция (иногда также называемая сигмоидом или логит-функцией):

$$f = \frac{1}{(1 + e^{-z})}$$

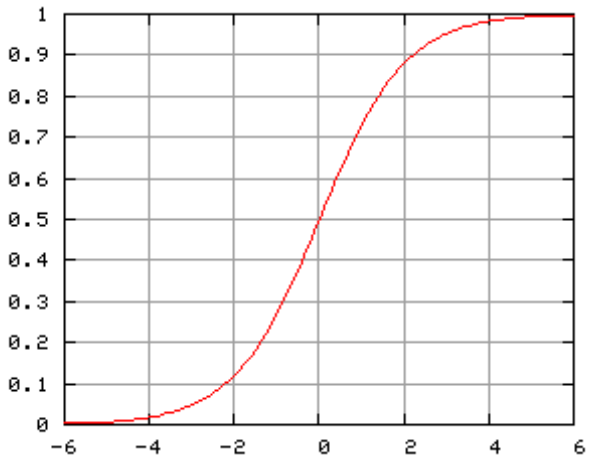

Рисунок 1.1 — Логистическая функция

Так как y у принимает лишь значения 0 и 1, то вероятность принять значение 0 равна:

$$P\{y = 0 \mid x\} = 1 = f(z) = 1 - (\beta^T x). \tag{1.7}$$

Для краткости функцию распределения y при заданном x можно записать в таком виде:

$$P\{y \mid x\} = f(\beta^T x)^y * (1 - f(\beta^T x))^{1-y}. \tag{1.8}$$

Фактически, это есть распределение Бернулли с параметром, равным $f(\theta^T x)$.

1.2.3.2 Полиномиальная логистическая регресия

Расширением бинарной логит-модели в случаях, когда зависимая переменная имеет более 2 категорий, является полиномиальная логит-модель. В таких случаях объединение данных в две категории может не иметь смысла или может привести к потере данных. Полиномиальная модель логита является подходящей техникой в этих случаях, особенно когда категории зависимых переменных не упорядочены (например, цвета, такие как красный, сий, зеленый).

Предполагается, что у нас есть серия из N наблюдаемых точек данных. Каждая точка данных i (в диапазоне от 1 до N) состоит из набора M объясняющих переменных x_1, ... x_M, (также известных как независимые переменные, переменные предиктора). И связанных с ними категориальных результатов Y_i (также известных как зависимая переменная, переменная ответа), которая может принимать одно из K возможных значений. Эти возможные значения представляют логически отдельные категории (например, разные политические партии, группы крови и т. д.) И часто описываются математически путем произвольного присвоения каждому числу от 1 до K.

Цель полиномиальной логистической регрессии состоит в том, чтобы построить модель, которая объясняет связь между объясняющими переменными и результатом, так, чтобы результат нового «эксперимента» мог быть правильно предсказан для новой точки данных, для которой объясняющие

переменные доступны. В процессе модель пытается объяснить относительное влияние различных объясняющих переменных на результат.

Некоторые примеры:

а) Наблюдаемые результаты - это различные варианты заболевания, такие как гепатит (возможно, включая «отсутствие заболевания» и / или другие сопутствующие заболевания) у группы пациентов, а объяснительные переменные могут быть характеристиками пациентов, которых считают подходящими (пол, раса) возраст, артериальное давление, результаты различных тестов функции печени и т. д.). Цель состоит в том, чтобы предсказать, какое заболевание вызывает наблюдаемые симптомы, связанные с печенью, у нового пациента.

б) Наблюдаемые результаты - это партия, выбранная группой людей на выборах, а объясняющие переменные - это демографические характеристики каждого человека (например, пол, раса, возраст, доход и т. Д.). Затем цель состоит в том, чтобы предсказать вероятное голосование нового избирателя с заданными характеристиками.

Полиномиальная логистическая регрессия использует функцию линейного предиктора $f(k, i)$, чтобы предсказать вероятность того, что наблюдение i имеет результат k следующего вида:

$$f(k,i) = \beta_{0,k} + \beta_{1,k}x_{1,i} + \beta_{2,k}x_{2,i} + ... + \beta_{M,k}x_{M,i} \tag{1.9}$$

где $\beta_{m,k}$ - коэффициент регрессии, связанный с m-й пояснительной переменной и k-м результатом. Коэффициенты регрессии и объясняющие переменные обычно группируются в векторы размером M + 1, так что функцию предиктора можно записать более компактно:

$$f(k,i) = \beta_k * x_i \tag{1.10}$$

где β_k - это набор коэффициентов регрессии, связанных с результатом k, и x_i (вектор строки) - это набор объясняющих переменных, связанных с наблюдением i. Прийти к полиномиальной модели логита можно представлением для K возможных результатов запуск независимых моделей бинарной

логистической регрессии $K-1$, в которых один результат выбирается как «опорный», а затем другие результаты $K-1$ отдельно регрессировал против разворота. Это будет действовать следующим образом, если результат K (последний результат) выбрана в качестве оси поворота:

$$\begin{gathered} ln\frac{P\{Y_i=1\}}{P\{Y_i=K\}}=\beta_1 X_i, \\ ln\frac{P\{Y_i=2\}}{P\{Y_i=K\}}=\beta_2 X_i, \\ \ldots\ldots \\ ln\frac{P\{Y_i=K-1\}}{P\{Y_i=K\}}=\beta_{K-1} X_i. \end{gathered} \tag{1.11}$$

Обратим внимание, что мы ввели отдельные наборы коэффициентов регрессии, по одному для каждого возможного результата. Если мы возведем в степень обе стороны и решим для вероятностей, мы получим:

$$\begin{gathered} P\{Y_i=1\}=P\{Y_i=K\}e^{\beta_1 X_i}, \\ P\{Y_i=2\}=P\{Y_i=K\}e^{\beta_2 X_i}, \\ \ldots\ldots \\ P\{Y_i=K-1\}=P\{Y_i=K\}e^{\beta_{K-1} X_i}. \end{gathered} \tag{1.12}$$

Используя тот факт, что все K вероятностей должны быть равны единице, мы находим:

$$\begin{gathered} P\{Y_i=K\}=1-\sum_{k=1}^{K-1}P\{Y_i=k\}=1-\sum_{k=1}^{K-1}P\{Y_i=K\}e^{\beta_k X_i}\Rightarrow \\ \Rightarrow P\{Y_i=K\}=\frac{1}{1+\sum_{k=1}^{K-1}P\{Y_i=K\}e^{\beta_k X_i}} \end{gathered} \tag{1.13}$$

Теперь можно использовать это, чтобы найти другие вероятности:

$$
\begin{aligned}
P\{Y_i = 1\} &= \frac{e^{\beta_1 X_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k X_i}} \\
P\{Y_i = 2\} &= \frac{e^{\beta_2 X_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k X_i}} \\
&\ldots\ldots \\
P\{Y_i = K-1\} &= \frac{e^{\beta_{K-1} X_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k X_i}}
\end{aligned}
\tag{1.14}
$$

1.2.4 Модели временного ряда

Модели временных рядов используются для прогнозирования или прогнозирования будущего поведения переменных. Эти модели учитывают тот факт, что точки данных, полученные со временем, могут иметь внутреннюю структуру (такую как автокорреляция, тренд или сезонные колебания), которую следует учитывать. В результате стандартные методы регрессии не могут быть применены к данным временных рядов, и была разработана методология для разложения трендового, сезонного и циклического компонентов ряда. Моделирование динамического пути переменной может улучшить прогнозы, поскольку предсказуемый компонент ряда может быть спроецирован на будущее.

Модели временных рядов оценивают разностные уравнения, содержащие стохастические компоненты. Двумя обычно используемыми формами этих моделей являются модели авторегрессии (AR) и модели скользящего среднего (MA). Методология Бокса-Дженкинса (1976), разработанная Джорджем Боксом и Г.М. Дженкинс объединяет модели AR и MA для создания модели ARMA (авторегрессионное скользящее среднее), которая является краеугольным камнем анализа стационарных временных рядов. ARIMA (модели авторегрессионных интегрированных скользящих средних), с другой стороны, используются для описания нестационарных временных рядов. Бокс

и Дженкинс предлагают дифференцировать нестационарные временные ряды, чтобы получить стационарный ряд, к которому может быть применена модель ARMA. Нестационарные временные ряды имеют ярко выраженную тенденцию и не имеют постоянного долгосрочного среднего значения или дисперсии.

1.2.5 Деревья классификации

Деревья классификации - это метод, позволяющий предсказывать принадлежность наблюдений или объектов к тому или иному классу категориальной зависимой переменной в зависимости от соответствующих значений одной или нескольких предикторных переменных.

Цель построения деревьев классификации заключается в предсказании (или объяснении) значений категориальной зависимой переменной, и поэтому используемые методы тесно связаны с более традиционными методами Дискриминантного анализа, Кластерного анализа, Непараметрической статистики и Нелинейного оценивания. Широкая сфера применимости деревьев классификации делает их весьма привлекательным инструментом анализа данных, но не следует поэтому полагать, что его рекомендуется использовать вместо традиционных методов статистики. Напротив, если выполнены более строгие теоретические предположения, налагаемые традиционными методами, и выборочное распределение обладает некоторыми специальными свойствами, то более результативным будет использование именно традиционных методов.

Однако, как метод разведочного анализа, или как последнее средство, когда отказывают все традиционные методы, деревья классификации, по мнению многих исследователей, не знают себе равных.

Представим, что нужно придумать устройство, которое отсортирует коллекцию монет по их достоинству (например, 1, 2, 3 и 5 копеек). Предположим, что какое-то из измерений монет, например - диаметр, известен и, поэтому, может быть использован для построения иерархического устройства сортировки монет. Заставим монеты катиться по узкому желобу, в котором прорезана щель размером с однокопеечную монету. Если монета провалилась в щель, то это 1 копейка; в противном случае она продолжает катиться дальше

по желобу и натыкается на щель для двухкопеечной монеты; если она туда провалится, то это 2 копейки, если нет (значит это 3 или 5 копеек) - покатится дальше, и так далее.

Таким образом, мы построили дерево классификации. Решающее правило, реализованное в этом дереве классификации , позволяет эффективно рассортировать горсть монет, а в общем случае применимо к широкому спектру задач классификации.

Изучение деревьев классификации не слишком распространено в вероятностно-статистическом распознавании образов [5], однако они широко используются в таких прикладных областях, как медицина (диагностика), программирование (анализ структуры данных), ботаника (классификация) и психология (теория принятия решений).

Деревья классификации идеально приспособлены для графического представления, и поэтому сделанные на их основе выводы гораздо легче интерпретировать, чем если бы они были представлены только в числовой форме.

1.3 Противоположные события

Два события называются совместными, если появление одного из них не исключает появление другого в одном и том же испытании.

Два события называются противоположными, если в данном испытании они несовместны и одно из них обязательно происходит. Вероятности противоположных событий в сумме дают 1.

Например, если при стрельбе по мишени попадание — это событие A, то событие $\bar{A}$ - это промах; сумма их вероятностей равна единице — при выстреле обязательно будет либо попадание, либо промах. То же самое и при подбрасывании монеты: обязательно выпадет либо орел, либо решка.

1.4 Методы экстраполяции

Экстраполяция - это метод научного исследования, который основан на распространении прошлых и настоящих тенденций, закономерностей, связей на будущее развитие объекта прогнозирования.

Цель методов экстраполяции – показать, к какому состоянию в будущем может прийти объект, если его развитие будет осуществляться с той же скоростью или ускорением, что и в прошлом. Методы экстраполяции достаточно широко применяются на практике, так как они просты, дешевы, и не требуют для расчетов большой статистической базы. Использование методов экстраполяции предполагает два допущения:

— основные факторы, тенденции прошлого сохранят свое проявление в будущем;

— исследуемое явление развивается по плавной траектории, которую можно выразить, описать математически.

Использование экстраполяции имеет в своей основе предположение о том, что рассматриваемый процесс представляет собой сочетание двух составляющих: регулярной составляющей X_t и случайной переменной ϵ. Временной ряд может условно представлен в виде:

$$Y_t = X_t + \epsilon_t. \tag{1.15}$$

1.4.1 Метод скользящей средней

Метод скользящей средней дает возможность выравнивать динамический ряд на основе его средних характеристик. При экстраполяции с помощью среднего уровня ряда используется принцип, при котором прогнозируемый уровень принимается равным среднему значению уровней ряда в прошлом.

Данный метод используется при краткосрочном прогнозировании. Его рабочая формула:

$$y_{t+1} = m_{t-1} + \frac{1}{n}(y_t - y_{t-1}), \tag{1.16}$$

где $t + 1$ – прогнозный период;

t – период, предшествующий прогнозному периоду (год, месяц и т.д.);

y_{t+1} – прогнозируемый показатель;

m_{t-1} – скользящая средняя за $n-1$ периода до прогнозного;

n – число уровней, входящих в интервал сглаживания;

y_t – фактическое значение исследуемого явления за предшествующий период;

y_{t-1} – фактическое значение исследуемого явления за $n-1$ периода, предшествующих прогнозному.

При больших значениях n колеблемость сглаженного ряда значительно снижается. Одновременно заметно сокращается количество наблюдений, что создает трудности.

Выбор интервала сглаживания зависит от целей исследования. При этом следует руководствоваться тем, в какой период времени происходит действие, а следовательно, и устранение влияния случайных факторов.

1.4.2 "Наивный"метод

Наивный метод основан на предположении, что будущее лучше всего характеризуется последними изменениями. Метод основывается на предположении о том, что прогнозируемые показатели в будущем периоде равно показателям предшествующего периода.

$$y_{t+1} = y_t \tag{1.17}$$

Наивный прогноз позволяет работать при отсутствии исторических данных. Наивный прогноз понятен, прост в подготовке, быстр в реализации, не требует, фактически, никаких затрат.

Основным недостатком наивного прогнозирования является вероятная низкая точность прогноза.

1.4.3 Метод средних

В данном подходе прогнозирования, все будущие значения принимаются равными средним значениям исторических данных. Этот подход может быть использован для любых исторических данных.

$$y_{t+1} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{1.18}$$

Подобная экстраполяция дает только точечную оценку. Однако, поскольку подобные прогнозы основываются на информации о поведении объекта в прошлом, то они всегда будут иметь ошибку.

Данный метод базируется на предположении о том, что средний уровень ряда не имеет тенденцию к изменению или если это изменение незначительно.

1.4.4 Метод экспоненциального сглаживания

Метод экспоненциального сглаживания дает возможность выявить тенденцию, сложившуюся к моменту последнего наблюдения. Метод экспоненциального сглаживания наиболее эффективен при разработке кратко- и среднесрочных прогнозов [6].

Его основные достоинства заключаются в простоте вычисления и учете весов исходной информации, т. е. новые данные или данные за последние периоды имеют больший вес, чем данные более отдаленных периодов.

$$U_{t+1} = \alpha y_t + (1 - \alpha) U_t, \tag{1.19}$$

где t – период, предшествующий прогнозному;

$t + 1$ – прогнозный период;

U_{t+1} - прогнозируемый показатель;

α - параметр сглаживания;

y_t - фактическое значение исследуемого показателя за период, предшествующий прогнозному;

U_t - экспоненциально взвешенная средняя для периода, предшествующего прогнозному.

При использовании для прогнозирования данного метода возникают следующие затруднения:

— выбор значения параметра сглаживания;

— определение начального значения экспоненциально взвешенной средней.

От величины α зависит, как быстро снижается вес влияния предшествующих наблюдений. Чем больше α, тем меньше сказывается влияние предшествующих лет. Если значение α близко к единице, то это приводит к учету при прогнозе в основном влияния лишь последних наблюдений. Если значение α близко к нулю, то веса, по которым взвешиваются уровни временного ряда, убывают медленно, т.е. при прогнозе учитываются все (или почти все) прошлые наблюдения.

Таким образом, если есть уверенность, что начальные условия, на основании которых разрабатывается прогноз, достоверны, следует использовать небольшую величину параметра сглаживания ($\alpha \to 0$). Когда параметр сглаживания мал, то исследуемая функция ведет себя как средняя из большого числа прошлых уровней. Если нет достаточной уверенности в начальных условиях прогнозирования, то следует использовать большую величину , что приведет к учету при прогнозе в основном влияния последних наблюдений.

Точного метода для выбора оптимальной величины параметра сглаживания α нет. В отдельных случаях автор данного метода профессор Браун предлагал определять величину α, исходя из длины интервала сглаживания[7].

При этом α вычисляется по формуле:

$$\alpha = \frac{2}{n+1}, \tag{1.20}$$

где n – число наблюдений, входящих в интервал сглаживания.

Задача выбора U_o (экспоненциально взвешенного среднего начального) решается следующими способами:

— если есть данные о развитии явления в прошлом, то можно воспользоваться средней арифметической и приравнять к ней U_o;

— если таких сведений нет, то в качестве U_o используют исходное первое значение базы прогноза y_1.

Также можно воспользоваться экспертными оценками.

Стоит отметить, что при изучении экономических временных рядов и прогнозировании экономических процессов метод экспоненциального сглаживания не всегда «срабатывает». Это обусловлено тем, что экономические временные ряды бывают слишком короткими (15-20 наблюдений), и в случае, когда темпы роста и прироста велики, данный метод не «успевает» отразить все изменения.

1.5 Вывод

В ходе анализа предметной области были определены положительные и отрицательные стороны уже существующих решений.

Метод	**Формат зависимой**	**Особенность**
Линейная регрессия	Непрерывная	Значение зависимой переменная может выйти за диапазон [0,1].
Полиномиальная регрессия	Непрерывная	Значение зависимой переменная может выйти за диапазон [0,1].
Логистическая регрессия	Дискретная (0, 1)	Значение зависимой переменной всегда в диапазоне [0,1]. Есть возможность получения коэффициентов регрессионной модели.
Полиномиальная логистическая регрессия	Дискретная (0, 1, 2, ...)	Применяется в задачах, где выделены более 2 классов.
Модели временного ряда	Непрерывная	Требуется определить рост, сезонность и другие показатели временного ряда.
Деревья классификации	Дискретная (0, 1, 2, . . .)	Нет возможности получения коэффициентов регрессионной модели.

Таблица 1.1 — Сравнение существующих методов построения регрессионной модели.

На основе таблицы 1.1 в качестве метода построения регрессионной модели была выбрана логистическая регрессия, так как необходимо рассчитать вероятность исходов события, т.е. значение зависимой переменной всегда

находится в диапазоне [0,1], также для последующего прогнозирования новых событий потребуются коэффициенты регрессионной модели.

Между всеми рассмотренными выше методами экстраполяции будет проведено сравнение, так как на данном этапе нельзя явно сказать, какой метод покажет лучший результат на данной выборке. Сравнение приведено на табл. 1.2.

Метод	**Особенность**
Метод средних	Позволяет разработать прогноз, основываясь на среднем значении прошлых наблюдений.
“Наивный” подход	Метод основывается на предположении о том, что прогнозируемые показатели в будущем периоде равно показателям предшествующего периода.
Метод скользящих средних	Метод основан на вычислении среднего среди n узлов. Затем период сдвигается на один узел. При этом периоды определения средней берутся все время одинаковыми. Таким образом, в каждом рассматриваемом случае средняя центрирована, т.е. отнесена к серединной точке интервала сглаживания и представляет собой уровень для этой точки.
Метод экспоненциального сглаживания	Данный метод приемлем при прогнозировании только на один период вперед. Его основные достоинства простота процедуры вычислений и возможность учета весов исходной информации.

Таблица 1.2 — Сравнение методов экстраполяции.

2 Конструкторский раздел

Для решения поставленной задачи предлагается использовать логистическую регрессию. Событие является победа, либо поражение первого спортсмена в поединке. В качестве модели построения регрессионной модели была выбрана логистическая регрессия. В этом разделе приводится описание архитектуры метода, а так же описываются основные этапы метода построения логистической регрессии. Также в данном разделе описаны критерии для оценки адекватности регрессионной модели и точности метода экстраполяции.

2.1 Описание выборки данных

Для построения регрессионной модели необходимо определить данные, которые состоят из пар значений зависимой переменной (переменной отклика) и независимой переменной (объясняющей переменной).

Один элемент выборки представляет собой бой, где в качестве зависимой переменной, будет бинарное значение 0, при поражении первого бойца, и 1 в случае победы первого бойца. Также в элемент выборки входят 20 регрессоров, по 10 показателей каждого из двух спортсменов:

а) Количество потрясений противника;

б) Количество значимых ударов;

в) Общее количество ударов;

г) Процент значимых ударов;

д) Количество переводов на землю;

е) Количество попыток перевода на землю;

ж) Процент успешных переводов;

и) Количество попыток удушений/болевых приёмов;

к) Количество переходов в опасные позиции;

л) Количество уходов из опасных позиций.

2.2 Описание этапов работы разрабатываемого метода

Применительно к задаче прогнозирования противоположных событий, регрессионная модель используется для выявления зависимости между входными значениями (показатели спортсмена) и выходным значением (вероятностью победу).

В рамках данной работы, с учетом показателей спортсменов, необходимо использовать двадцать входных переменных по десять на каждого спортсмена, участвующего в бою и одной выходной переменной, которая является веротяностью победы первого спортсмена в бою.

Таким образом, этапы разработки метода прогнозирования противоположных событий на основе регрессионного анализа можно представить в виде следующей последовательности действий на рис.2.1:

а) Построение регрессионной модели - расчёт коэффициентов, проверка корректности параметров модели, оценка её адекватности. На этом этапе расчёт коэффициентов производится на полной выборке боёв.

б) Выбор двух спортсменов из одной весовой категории и соответсвующих им показателей за все проведённые бои.

в) Так как прогнозируется бой, который ещё не состоялся, необходимы новые показатели спортменов, основанные на их исторических данных. На этом этапе выполняется экстраполяция данных.

г) Затем для получения веротяностей исходов событий требуется расчитать уравнение регрессионной модели с новыми экстраполированными данными.

2.2.1 Построение регрессионной модели

Поскольку для определения вероятностей исходов событий используется логистическая регрессия, обучающая выборка выглядит следующим образом:

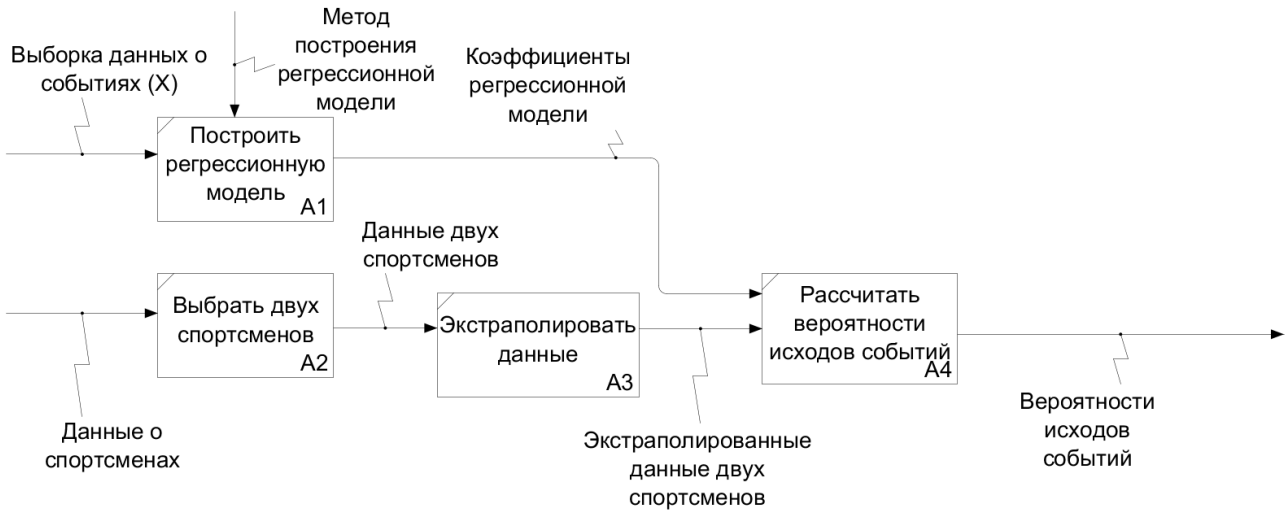

Рисунок 2.1 — Функциональная модель разработанного метода

$$\vec{y} = [y_0, y_1, \dots y_n], \tag{2.1}$$

$$X = \begin{bmatrix} x_{1,1} & x_{2,1} & \dots & x_{20,1} \\ x_{1,2} & x_{1,2} & \dots & x_{20,2} \\ \dots\dots & & & \\ x_{1,n} & x_{2,n} & \dots & x_{20,n} \end{bmatrix},$$

где $\vec{y}$ - вектор исходов событий для первого спортсмена,

n - количество элементов выборки,

X - матрица, каждая строка в которой является набором показателей первого и второго спортсменов.

Делается предположение о том, что вероятность наступления события $y = 1$ равна:

$$P\{y = 1 \mid x\} = f(z). \tag{2.2}$$

$$f = \frac{1}{(1 + e^{-z})}, \tag{2.3}$$

где $z = \beta^T \vec{x} = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$,

β - вектор коэффициентов(весов) при показателях спортсменов,

$f(z)$ - логистическая функция, значения которой для нашей задачи являются веротностями победы первого спортсмена.

$$P\{y = 0 \mid x\} = 1 - f(z) = 1 - (\beta^T x). \tag{2.4}$$

Согласно уравнению 2.21 вычисляется вероятность проигрыша первого спортсмена.

2.2.1.1 Подбор коэффициентов регрессионной модели

Для нахождения коэффициентов регрессионной модели используется метод максимального правдоподобия, согласно которому вбираются параметры β, максимизирующие значение функции правдоподобия на обучающей выборке:

$$\hat{\beta} = argmax_\beta L(\beta) = argmax_\beta \prod_{i=1}^{m} P\{y = y_i \mid x = x_i\}, \tag{2.5}$$

Максимизация функции правдоподобия эквивалентна максимизации её логарифма:

$$lnL(\beta) = \sum_{i=1}^{m} lnP\{y = y^{(i)} \mid x = x^{(i)}\} = \\ = \sum_{i=1}^{m} \left[y^{(i)} ln(f(\beta^T x^{(i)})) + (1 - y^{(i)}) ln(1 - f(\beta^T x^{(i)})) \right] \tag{2.6}$$

Для максимизации этой функции может быть применён, например, алгоритм Бройдена — Флетчера — Гольдфарба — Шанно. Итерационный метод численной оптимизации, предназначенный для нахождения локального максимума, либо минимума нелинейного функционала без ограничений[9].

Результатом данного этапа является вектор, рассчитанных коэффициентов:

$$\vec{\beta} = [\beta_0, \beta_1, \beta_2, \ldots, \beta_{k-1}], \tag{2.7}$$

где k = 20, количеству регрессоров в модели.

2.2.1.2 Оценка адекватности модели

Для оценки модели будут использоваться метрики, оценивающие качество классификации.

ROC-кривая (англ. Receiver Operator Characteristic) – кривая, которая часто [8] используется для представления результатов бинарной классификации. Название пришло из систем обработки сигналов.

Поскольку классов два, один из них называется классом с положительными исходами, второй – с отрицательными исходами. ROC-кривая показывает зависимость количества верно классифицированных положительных примеров от количества неверно классифицированных отрицательных примеров. В терминологии ROC-анализа первые называются истинно положительным, вторые – ложно отрицательным множеством.

Для понимания сути ошибок I и II рода рассмотрим четырехпольную таблицу сопряженности (англ. confusion matrix), которая строится на основе результатов классификации моделью и фактической принадлежностью примеров к классам.

	Фактически	
Модель	положительно	отрицательно
положительно	TP	FP
отрицательно	FN	TN

Таблица 2.1 — Таблица сопряжённости.

TP (True Positives) – верно классифицированные положительные примеры (так называемые истинно положительные случаи);

TN (True Negatives) – верно классифицированные отрицательные примеры (истинно отрицательные случаи);

FN (False Negatives) – положительные примеры, классифицированные как отрицательные (ошибка I рода). Это так называемый "ложный пропуск"– когда интересующее нас событие ошибочно не обнаруживается (ложно отрицательные примеры);

FP (False Positives) – отрицательные примеры, классифицированные как положительные (ошибка II рода); Это ложное обнаружение, т.к. при отсутствии события ошибочно выносится решение о его присутствии (ложно положительные случаи).

Что является положительным событием, а что – отрицательным, зависит от конкретной задачи. Например, если мы прогнозируем вероятность наличия заболевания, то положительным исходом будет класс "Больной пациент отрицательным – "Здоровый пациент". И наоборот, если мы ходим определить вероятность того, что человек здоров, то положительным исходом будет класс "Здоровый пациент и так далее.

При анализе чаще оперируют не абсолютными показателями, а относительными – долями (англ. rates), выраженными в процентах:

Доля истинно положительных примеров (англ. True Positives Rate):

$$TPR = \frac{TP}{TP + FN} * 100\% \tag{2.8}$$

Доля ложно положительных примеров (англ. False Positives Rate):

$$FPR = \frac{FP}{TN + FP} * 100\% \tag{2.9}$$

Введем еще два определения: чувствительность и специфичность модели. Ими определяется объективная ценность любого бинарного классификатора.

Чувствительность (англ. Sensitivity) – это и есть доля истинно положительных случаев:

$$S_e = TPR = \frac{TP}{TP + FN} \cdot 100\% \tag{2.10}$$

Специфичность (англ. Specificity) – доля истинно отрицательных случаев, которые были правильно идентифицированы моделью:

$$S_p = \frac{TN}{TN + FP} \cdot 100\% \tag{2.11}$$

Модель с высокой чувствительностью часто дает истинный результат при наличии положительного исхода (обнаруживает положительные примеры).

Наоборот, модель с высокой специфичностью чаще дает истинный результат при наличии отрицательного исхода (обнаруживает отрицательные примеры). Если рассуждать в терминах медицины – задачи диагностики заболевания, где модель классификации пациентов на больных и здоровых называется диагностическим тестом, то получится следующее:

— Чувствительный диагностический тест проявляется в гипердиагностике – максимальном предотвращении пропуска больных;

— Специфичный диагностический тест диагностирует только доподлинно больных. Это важно в случае, когда, например, лечение больного связано с серьезными побочными эффектами и гипердиагностика пациентов не желательна.

ROC-кривая выглядит следующим образом:

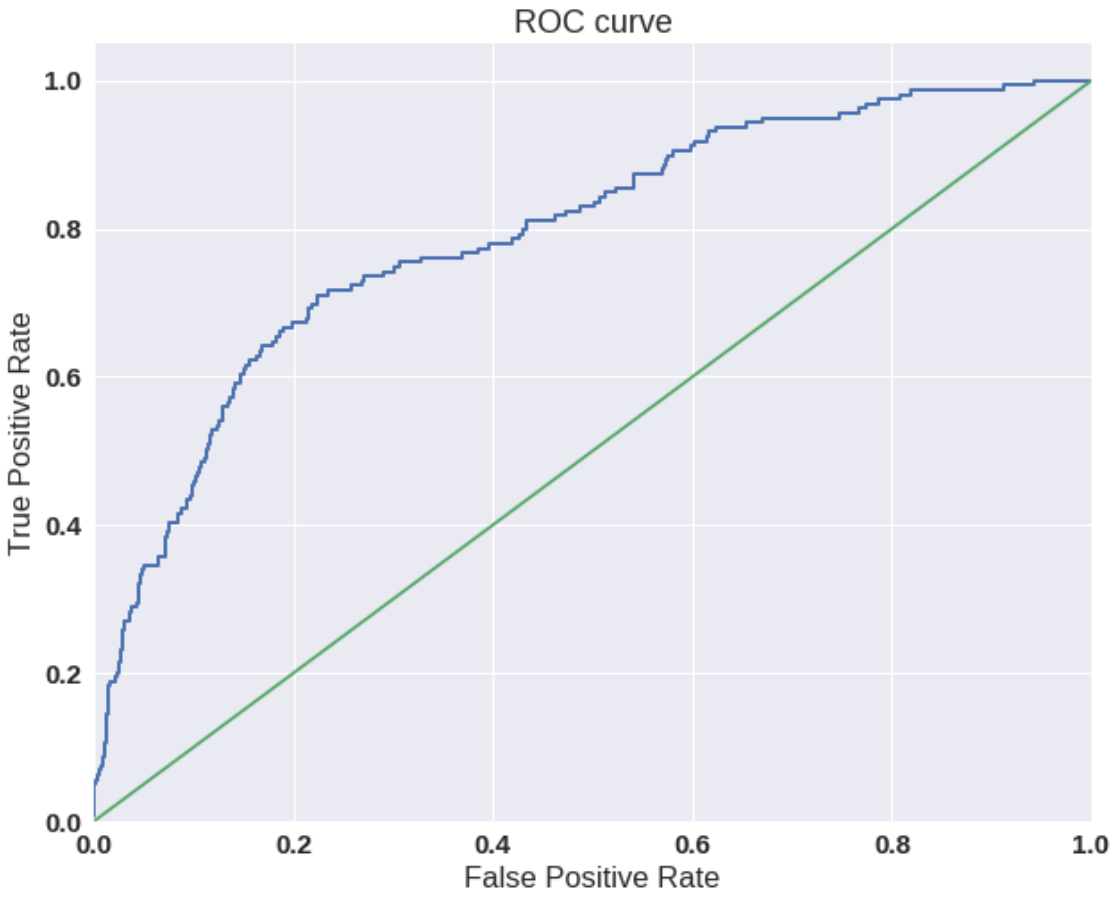

Рисунок 2.2 — Кривая ROC

Для каждого значения порога отсечения, которое меняется от 0 до 1 с шагом dx (например, 0.01) рассчитываются значения чувствительности S_e и специфичности S_p. В качестве альтернативы порогом может являться каждое последующее значение примера в выборке. Строится график зависимости: по оси Y откладывается чувствительность S_e, по оси X – 100% – S_p (сто процентов минус специфичность), или, что то же самое, FPR – доля ложно положительных случаев.

Методом сравнения ROC-кривых является оценка площади под кривыми. Теоретически она изменяется от 0 до 1.0, но, поскольку модель всегда характеризуются кривой, расположенной выше положительной диагонали, то обычно говорят об изменениях от 0.5 ("бесполезный"классификатор) до 1.0 ("идеальная"модель). Эта оценка может быть получена непосредственно вычислением площади под многогранником, ограниченным справа и снизу осями координат и слева вверху – экспериментально полученными точками. Численный показатель площади под кривой называется от англ. AUC (Area Under Curve). Вычислить его можно, например, с помощью численного метода трапеций:

$$AUC = \int f(x)dx = \sum_i \frac{X_{i+1} + X_i}{2}(Y_{i+1} - Y_i) \tag{2.12}$$

В литературе [8] приводится следующая экспертная шкала для значений AUC, по которой можно судить о качестве модели:

Интервал AUC	**Качество модели**
0.9-1.0	отличное
0.8-0.9	очень хорошее
0.7-0.8	хорошее
0.6-0.7	среднее
0.5-0.6	неудовлетворительное

Таблица 2.2 — Экспертная шкала для значений AUC.

Идеальная модель обладает 100% чувствительностью и специфичностью. Однако на практике добиться этого невозможно, более того, невозможно одновременно повысить и чувствительность, и специфичность модели.

2.2.2 Выбор двух спортсменов

Пользователю предлагается выбрать весовую категорию, описанную в разделе 1.1.2, спортсменов. Затем выбрать двух спортсменов из данной категории. Прогнозироваться будет вероятность победы первого и второго бойцов в новом, ещё не проводившемся поединке.

Результатом данного этапа являются два набора данных о спортсменах в виде:

$$X_{\text{первый}} = \begin{bmatrix} x_{1,1} & x_{2,1} & \dots & x_{10,1} \\ x_{1,2} & x_{2,2} & \dots & x_{10,2} \\ \dots & \dots & \dots & \dots \\ x_{1,n} & x_{2,n} & \dots & x_{10,n} \end{bmatrix};$$

$$X_{\text{второй}} = \begin{bmatrix} x_{1,1} & x_{2,1} & \dots & x_{10,1} \\ x_{1,2} & x_{2,2} & \dots & x_{10,2} \\ \dots & \dots & \dots & \dots \\ x_{1,n} & x_{2,n} & \dots & x_{10,n} \end{bmatrix},$$

где n количество боёв спортсмена.

Количество элементов в строке равно десяти, так как в предметной области были определены десять регрессоров на каждого спортсмена для объяснения зависимой переменной.

2.2.3 Реализация метода экстраполяции

После вычисления коэффициентов регрессионной модели, для вычисления вероятностей исходов событий необходимо экстраполировать данные о двух спортсменах.

Используется комбинированный метод экстраполяции, основанный на методах экспоненциального сглаживания и скользящей средней.

Формула для вычисления экстраполированных показателей спортсменов 2.13.

$$U_{t+1} = \alpha y_t + (1 - \alpha)U_0 \tag{2.13}$$

Перед расчётом новых показателей спортсменов необходимо определить значение U_0. Заменим на y_{t+1} из формулы для расчёта скользящей средней для всех периодов.

$$y_{t+1} = m_{t-1} + \frac{y_t - y_{t-1}}{n}, \tag{2.14}$$

где $t + 1$ – прогнозный период;

t – период, предшествующий прогнозному периоду;

y_{t+1} – прогнозируемый показатель;

$m - 1$ – скользящая средняя за $n - 1$ периода до прогнозного;

n – число уровней, входящих в интервал сглаживания;

y_t – фактическое значение исследуемого явления за предшествующий период;

y_{t-1}– фактическое значение исследуемого явления за $n - 1$ период, предшествующих прогнозному.

После этого этапа получен следующий набор значений:

$$\vec{x}_{\text{экстр}} = \left[x_{1,\text{первый}}, x_{2,\text{первый}}, \ldots, x_{10,\text{первый}}, x_{1,\text{второй}}, x_{2,\text{второй}}, \ldots, x_{10,\text{второй}}\right], \tag{2.15}$$

где $\vec{x}_{\text{экстр}}$ - экстраполированные данные спортсменов.

2.2.3.1 Оценка точности методов экстраполяции

Оценка точности прогноза, построенного методом экстраполяции Всякий прогноз должен иметь высокую точность, которая является важнейшей его характеристикой. Существует несколько способов оценки точности прогноза:

а) Средняя абсолютная оценка:

$$t = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n} \tag{2.16}$$

где y_i – фактическое значение исследуемого явления, $\hat{y}_i$ – расчетное значение исследуемого явления, n – число уровней временного ряда;

б) Средняя квадратическая оценка:

$$\delta = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{2.17}$$

Чем ближе к нулю первый и второй показатели, тем выше точность экстраполяции.

в) Средняя относительная ошибка:

$$\epsilon = \frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - \hat{y}_i}{y_i} * 100) \tag{2.18}$$

2.2.4 Расчёт вероятностей исходов событий

После вычисления коэффициентов логистической регрессии и экстраполяции данных двух спортсменов перейдём к расчёту вероятности победу первого спортсмена.

Так как использовалась логистическая регрессия, то вероятность победы первого спортсмена вычисляется по следующим формулам:

$$P\{y = 1 \mid x\} = f(z). \tag{2.19}$$

$$f = \frac{1}{(1 + e^{-z})}, \tag{2.20}$$

где $z = \beta^T \vec{x}_{\text{экстр}} = \beta_0 + \beta_1 x_{1,\text{первый}} + \beta_2 x_{2,\text{первый}} + ... + \beta_{10} x_{1,\text{второй}} + \beta_{11} x_{2,\text{второй}} + \cdots + \beta_n x_n$,

β - вектор коэффициентов(весов) при показателях спортсменов. Согласно следующему уравнению вычисляется вероятность проигрыша первого спортсмена:

$$P\{y = 0 \mid x\} = 1 - f(z) = 1 - (\beta^T x). \tag{2.21}$$

2.3 Структура программного обеспечения

Схема ПО, реализующего описанные выше методы:

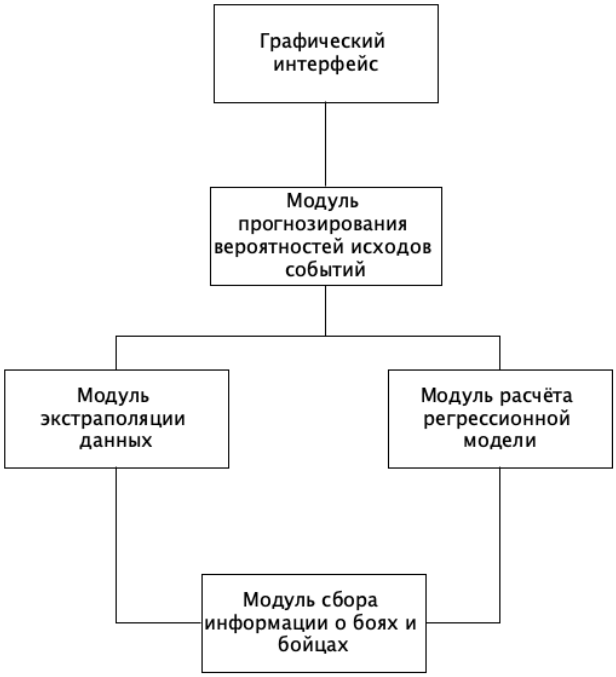

Рисунок 2.3 — Схема взаимосвязей модулей программного обеспечения

Данная схема представляет собой вариант реализации паттерна MVP (Model – View – Presenter). Компонент графического интерфейса должен обеспечивать взаимодействия пользователя с приложением. Модуль прогнозирования должен отвечать за расчёт веротяностей исходов событий, в зависимости от данных потспувших от модулей экстраполяции и расчёта регрессионной модели. Модуль сбора информации формирует наборы данных необходимые модулям экстраполяции и расчёта регрессионной модели.

3 Технологический раздел

В этом разделе представлены основные моменты реализации разрабатываемого программного обеспечения, выбор и обоснование выбора технических средств необходимых для разработки программного продукта. Устанавливаются входные данные и их формат, осуществляется выбор языка программирования, с помощью которого выполнялось кодирование программного обеспечения. Представлена структура и интерфеqс разработанноq программы, а также описаны способы работы с реализованным программным обеспечением.

Также в данном разделе был описан модуль сбора выборки данных для построения регрессионной модели и экстраполяции.

3.1 Выбор языка программирования

Реализуемое программное обеспечение состоит из двух основных частей – блока прогнозирования вероятностей исходов событий и блока графического отображения. Блок прогнозирования вероятностей исходов событий основывается на использовании логистической модели и метода экстраполяции. Реализация логистической модели присутствует в библиотеке skikit-learn[10], реализованной на Python3. Было решено использовать skikit-learn, поскольку она предоставляет удобный функционал для работы с математическими алгоритмами, а также содержит реализацию необходимых функций для обучения и использования логистической регрессии.

В рамках данной работы предполагается, что результат работы метода прогнозирования можно увидеть не только в виде графиков, но и на примере работы с программой. В связи с этим, было решено использовать графический интерфейс, реализованный в библиотеке PyQt5[11]. В качестве языка программирования был выбран язык Python3, в качестве среды разработки – Microsoft Visual Studio Code 2018. Выбор Python3 был обусловлен тем, что библиотека, в которой представлена реализация логистической регрессии, написана на этом языке программирования.

Выбор Visual Studio Code 2018 в качестве среды разработки был обусловлен следующими факторами:

— поддержка современных стандартов языка Python2 и Python3;

— интеграция с системой контроля версия Git;

— наличие удобных инструментов для отладки программы и рефакторинга кода; наличие встроенного компонента для модульного тестирования.

3.2 Модуль сбора данных

Обязательным требованием к использованию логистической регрессии является наличие выборки данных. Было реализовано ПО, осуществляющее сбор данных о боях и спортсменах. В качестве языка программирования был выбран Python3, так как имеется опыт его использования для подобных задач. В ПО использовались библиотеки для сохранения и анализа html страниц. Все данные брались с официального сайта Абсолютного бойцовского чемпионата [18].

Модуль сохраняет выборку боёв в формате CSV в следующем виде:

```
1 ,Fight_Result,First_KnockDowns,Second_KnockDowns,
2 First_Significant_Strikes,First_Total_Strikes,
3 Second_Significant_Strikes,Second_Total_Strikes,
4 First_Significant_Strikes_%,Second_Significant_Strikes_%,
5 First_TakeDowns,First_TakeDowns_Attempts,
6 Second_TakeDowns,Second_TakeDowns_Attempts,First_TD_%,
7 Second_TD_%,First_Subs,Second_Subs,First_Passes,
8 Second_Passes,First_Rev.,Second_Rev.
9 0,0,0,0,77,148,63,117,52,53,1,1,1,3,100,33,0,2,2,1,0,0
10 1,1,0,0,99,219,67,189,45,35,1,1,0,0,100,0,0,0,0,0,0,0
11 2,1,0,0,11,18,3,5,61,60,1,1,0,3,100,0,2,0,1,0,0,0
12 3,1,2,0,79,138,71,111,57,63,3,3,1,2,100,50,1,0,1,1,0,0
```

На листинге выше показан пример выборки боёв, где под первые 8 строк отводится названия столбцов, перечисленные через запятую. В оставшихся

строках, каждая строка представляет отдельный бой. Показатели сохраняются в порядке названий столбцов, и разделяются запятыми.

Также в модуле реализовано сохранение выборки боёв отдельного бойца. На вход подаётся url-ссылка спортсмена. На выходе получим следующую выборку всех боёв спортсмена.

```
 1 ,Fight_Result,First_KnockDowns,Second_KnockDowns,
 2 First_Significant_Strikes,First_Total_Strikes,
 3 Second_Significant_Strikes,Second_Total_Strikes,
 4 First_Significant_Strikes_%,Second_Significant_Strikes_%,
 5 First_TakeDowns,First_TakeDowns_Attempts,
 6 Second_TakeDowns,Second_TakeDowns_Attempts,First_TD_%,
 7 Second_TD_%,First_Subs,Second_Subs,First_Passes,
 8 Second_Passes,First_Rev.,Second_Rev.
 9 0,1,2,138,361,38,0,0,0,0,0,0,0,90,223,40,0,5,0,0,0,0,1
10 1,1,1,80,162,49,1,1,100,0,0,0,0,49,105,46,0,2,0,0,0,0,1
11 2,1,0,9,13,69,0,1,0,1,2,1,0,11,20,55,1,1,100,0,1,0,0
12 3,0,0,60,131,45,1,9,11,0,1,0,0,84,166,50,0,4,0,0,0,0,0
13 4,1,3,28,57,49,0,1,0,0,0,0,0,25,79,31,0,0,0,0,0,0,0
14 5,0,0,14,24,58,1,2,50,0,0,0,1,24,55,43,0,0,0,0,0,0,0
15 6,0,0,82,166,49,1,10,10,0,2,0,0,77,156,49,0,1,0,0,0,0,1
16 7,0,0,30,65,46,0,1,0,0,0,0,2,58,102,56,0,0,0,0,0,0,0
```

Принцип разделения показателей соответственно столбцам такой же, как и в сохранении боёв, однако с учётом личностей спортсменов.

3.3 Интерфейс

Пользователь производит взаимодействие с графическим интерфейсом для реализованного метода прогнозирования исходов противоположных событий только курсором мыши.

Для прогнозирования пользователю необходимо выбрать весовую категорию спортсменов из выпадающего меню.

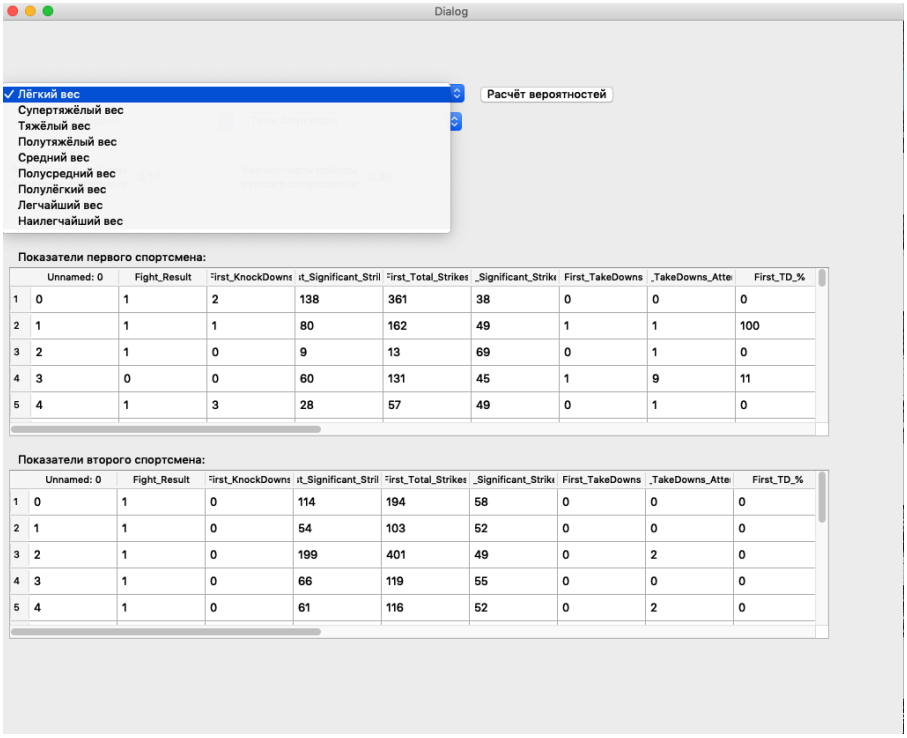

Рисунок 3.1 — Выбор весовой категории спортсменов

После выбора весовой категории спортсменов происходит выбор двух спортсменов из выпадющих меню.

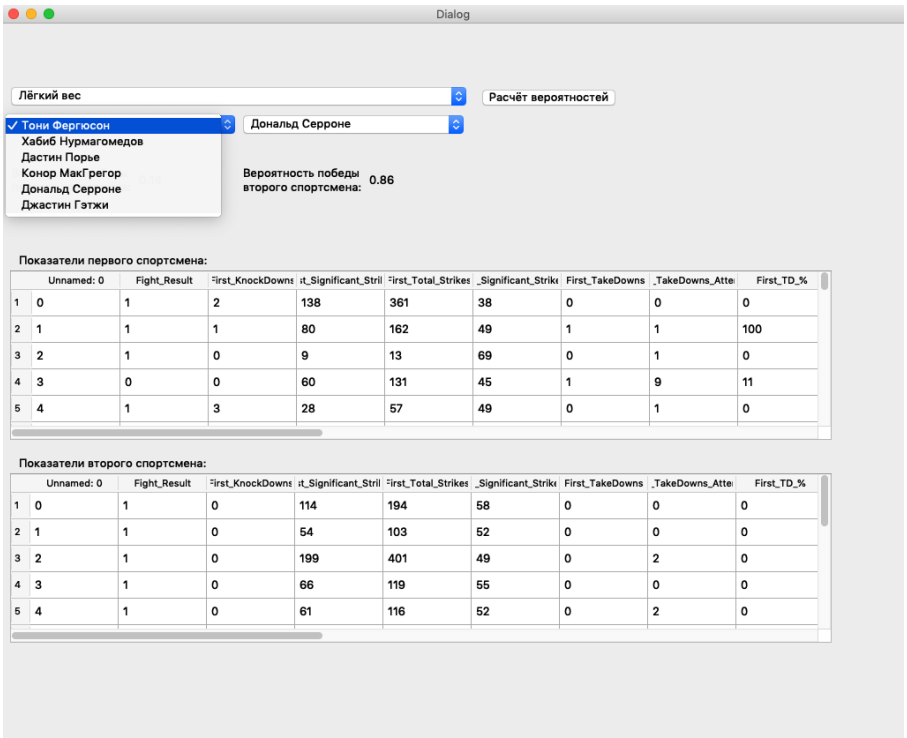

Рисунок 3.2 — Выбор спортсменов

Далее для расчёта вероятностей исходов требуется нажать на кнопку "Рассчёт вероятностей". Также приведены в таблицах ниже данные спортсменов за все их бои. Порядок показа с самых новых до самых старых.

В соответствии с описанными в конструкторском разделе алгоритмами, на языке Python3 в среде разработки Microsoft Visual Studio Code под управлением MacOS был реализован программный метод прогнозирования противоположных событий. Было произведено тестирование и отладка программного продукта. Описаны форматы используемых при работе ПО файлов и приведен пример интерфейса программы.

4 Экспериментальный раздел

Целью данной работы является прогнозирование противоположных событий, то есть вычисление вероятностей победы, либо поражения.

Как уже упоминалось в аналитическом разделе, данная задача состоит из трёх частей: построения регрессионной модели, экстраполяция данных спортсменов и вычисление вероятностей. В данном разделе проводится экспериментальное исследование эффективности разработанного метода с использованием реализующего его программного обеспечения. В рамках экспериментального раздела также проводится сравнение результатов работы методов экстраполяции, для выбора подоходящего, так как на этапе анализа не представлялось возможным выявить наиболее подходящий метод.

4.1 Значимость коэффициентов регрессии

Так как функция по которой производится расчёт вероятности победы имеет вид:

$$f = \frac{1}{(1 + e^{-z})}, \tag{4.1}$$

то для оценки значимости коэффициентов β, необходимо будет вычислить e^{β}.

Согласно данным представленным в таблице 4.1 наиболее значимым регрессором для победы стало "Количество потрясений первым спортсменом". Соответственно максимально понижает шансы на победу первого спортсмена значение регрессора "Количество потрясений вторым спортсменом".

Также достаточно важными для победы являются такие показатели как "Количество переводов на землю первым спортсменом "Количество попытокудушений/болевых приёмов первым спортсменом "Количество уходов из опасных позиций первым спортсменом"и "Количество переходов в опасные позиции первым спортсменом".

Регрессор	β	e^{β}
Количество потрясений первым спортсменом	1.39	4.05
Количество значимых ударов от первого спортсмена	0.06	1.06
Количество переводов на землю первым спортсменом	0.36	1.44
Количество попыток удушений/болевых приёмов первым спортсменом	0.75	2.12
Количество уходов из опасных позиций первым спортсменом	0.15	1.16
Количество переходов в опасные позиции первым спортсменом	0.49	1.63
Количество потрясений вторым спортсменом	-1.37	0.25
Количество значимых ударов от второго спортсмена	-0.05	0.94
Количество переводов на землю вторым спортсменом	-0.25	0.77
Количество попыток удушений/болевых приёмов вторым спортсменом	-0.29	0.74
Количество уходов из опасных позиций вторым спортсменом	-0.31	0.72
Количество переходов в опасные позиции вторым спортсменом	-0.48	0.61

Таблица 4.1 — Таблица значимости коэффициентов регрессионной модели.

4.2 Оценка адекватности модели

Необходимо провести исследование бинарного классификатора, коим является логистическая модель.

В первую очередь важна доля ошибок 1 и 2 рода в нормализованной таблице сопряжённости.

Положения значений в таблице сопряжённости:

— В первом столбце, первой строке процент правильно классифицированных поражений.

	Истинное поражение	**Истинная победа**
Предполагаемое поражение	0.80	0.19
Предполагаемая победа	0.06	0.93

Таблица 4.2 — Нормализованная таблица сопряжённости

— В первом столбце, второй строке процент неправильно классифицированных поражений, принятых за победу.

— Во втором столбце, первой строке процент неправильно классифицированных поражений, принятых за поражение.

— Во втором столбце, второй строке процент правильно классифицированных побед.

Из таблицы 4.2 видно, что проценты ошибок 1 и 2 рода не превышают 20%. Также по данной таблице можно оценить, что победы классифицируются лучше поражений на 13%.

4.3 Оценка метода прогнозирования

После вычисления коэффициентов регрессионной модели и экстраполяции данных спортсменов, необходимо показать точность метода прогнозирования. Для этого будем рассматривать каждый бой из выборки, экстраполировать показатели спортсменов, на основании всех его данных до выбранного боя. Затем вычислять вероятность победы бойца в этом бою. На следующей таблице представлены метрики оценки метода прогнозирования:

	Точность	**Полнота**	**Количество элементов выборки**
Поражение	0.87	0.68	1526
Победа	0.87	0.95	3277

Таблица 4.3 — Метрики оценки метода прогнозирования

Из табл. 4.3 видно, что средняя точность прогнозирования как побед, так и поражений равна 87%. Высокая чувствительность 95% показывает, что разработанный метод хорошо прогнозирует вероятность побед.

ЗАКЛЮЧЕНИЕ

В результате выполнения выпускной квалификационной работы был спроектирован, реализован, протестирован и отлажен метод прогнозирования исходов противоположных событий на основе регрессионного анализа.

В процессе создания программной системы были достигнуты следующие результаты:

а) Разработан метод прогнозирования исходов противоположных событийна основе регрессионного анализа.

б) Спроектирована структура программного обеспечения для моделирования метода прогнозирования противоположных событий.

в) Создано программное обеспечение, реализующее разработанный метод. Проведено исследование работоспособности и точностных характеристик разработанного программного обеспечения. ПО полностью соответствует требованиям технического задания.

Достоинства разработанного программного обеспечения:

— высокая точность прогноза;

— данное программное обеспечение является полностью бесплатным и открытым для модификации;

Основным недостатком разработанной системы является зависимость от библиотек skikit-learn.

Можно выделить следующие пути дальнейшего развития разработанного метода:

— применение методов машинного обучения;

— реализация web-приложения для большего охвата пользователей;

— повышение точности за счёт составления более детализированных выборок.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Официальный сайт организации UFC. Электронный ресурс. - Режим доступа: http://se.ufc.com/ - (Дата обращения: 18.05.2019)

2. Официальный сайт организации M-1 Global. Электронный ресурс. - Режим доступа: http://mixfight.ru/ - (Дата обращения: 18.05.2019)

3. Свод правил и нормативных документов смешанных единоборств. Электронный ресурс. - Режим доступа: http://www.ufc.com/discover/sport/rules-and-regulations - (Дата обращения: 18.05.2019)

4. Нормативный документ о судействе смешанных единоборств. Электронный ресурс. - Режим доступа: https://www.state.nj.us/lps/sacb/docs/martial.html - (Дата обращения: 18.05.2019)

5. Брайан Д. Рипли. Распознавание образов и нейронные сети: – Кембридж: Изд-во Cambridge University Press, 1996. - 215 с.

6. Владимирова Л.П. Прогнозирование и планирование в условиях рынка: Учеб. пособие. М.: Издательский Дом «Дашков и Ко», 2001. - 34 с.

7. Новикова Н.В., Поздеева О.Г. Прогнозирование национальной экономики: Учебно-методическое пособие. Екатеринбург: Изд-во Урал. гос. экон. ун-та, 2007 - 138 с.

8. Zweig, Mark H.; Campbell, Gregory. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical Chemistry. Нью-Йорк: Изд-во Oxford, 577 с.

9. Цыплаков А.А. Некоторые эконометрические методы. Метод максимального правдоподобия в эконометрии. Методическое пособие. - Новосибирск: НГУ, 1997. - 129 с.

10. Документация к библиотеке skikit-learn. Электронный ресурс. - Режим доступа: https://scikit-learn.org/stable/documentation.html - (Дата

обращения: 18.05.2019)

11. Документация к библиотеке PyQt5. Электронный ресурс. - Режим доступа: https://www.riverbankcomputing.com/static/Docs/PyQt5/ - (Дата обращения: 19.05.2019)

12. Complex tennis analysis solution for coaches, players and mass media. Электронный ресурс. - Режим доступа: http://tenniscomstat.com - (Дата обращения: 19.05.2019)

13. Using Statistical Analysis in Football Betting. Электронный ресурс. - Режим доступа: https://totalfootballanalysis.com/thought-analysis/using-statistical-analysis-football-betting - (Дата обращения: 19.05.2019)

14. Hockey Analytics. Devoted to the Scientific Exploration of the Game of Hockey. Электронный ресурс. - Режим доступа: http://hockeyanalytics.com/ - (Дата обращения: 19.05.2019)

15. Basketball analytics solutions. Электронный ресурс. - Режим доступа: https://www.stats.com/basketball/ - (Дата обращения: 19.05.2019)

16. Mathematical and statistical insights into Formula 1. - Режим доступа: https://f1metrics.wordpress.com/ - (Дата обращения: 19.05.2019)

17. The growth of the UFC ahead of the historic UFC200.- Режим доступа: https://www.thestatszone.com/the-growth-of-the-ufc -ahead-of-the-historic-ufc200 - (Дата обращения: 19.05.2019)

18. Официальный сайт статистик UFC. - Режим доступа: http://ufcstats.com/statistics/events/completed - (Дата обращения: 20.05.2019)

ПРИЛОЖЕНИЕ А

ИЗОБРАЖЕНИЯ