# Возможный подход к задаче прогнозирования спортивных результатов методами анализа данных

Журавлев А. Д.[1]

В данной статье построен алгоритм сведения задачи прогнозирования спортивных результатов к задаче бинарной классификации. При этом обоснована оптимальность этого алгоритма с точки зрения применения результатов прогнозирования в игре против букмекерских контор.

**Ключевые слова:** машинное обучение, прогнозирование спортивных результатов, бинарная классификация.

## 1. Введение

Современные методы машинного обучения (ассоциативные правила, деревья решений, модель гауссовых смесей, алгоритмы регрессии, нейронные сети, байесовские сети и т. д.) используются во многих областях для решения проблем ассоциации, классификации, сегментации, диагностики и прогнозирования. Вполне логично, что эти алгоритмы находят применение в такой экстремальной человеческой деятельности, как спорт уровня высоких достижений.

Спортивная аналитика – новый быстрорастущий рынок, объем которого превысит $4,7 млрд к 2021 г.(по прогнозам WinterGreen Research). Около $1 млрд из них придется на долю хоккея, считают представители стартапа Iceberg Sports Analytics [1].

Принимая во внимание огромное количество исторических данных по хоккею, можно предложить подход к прогнозированию хоккейных матчей – машинное обучение. Параметры игроков и матча вместе с результатом могут составить обучающую выборку. Алгоритм машинного

[1] *Журавлев Артем Дмитриевич* — аспирант каф. математической теории интеллектуальных систем мех.-мат. ф-та МГУ, e-mail: artemzhuravlev.msu@gmail.com.

Zhuravlev Artem Dmitrievich — graduate student, Lomonosov Moscow State University, Faculty of Mechanics and Mathematics, Chair of Mathematical Theory of Intellectual Systems.

обучения с учителем может использовать эту выборку для построения функции предсказания результатов новых матчей.

В данной статье рассмотрен один из возможных способов формализации задачи прогнозирования спортивных результатов и сведение ее к задаче бинарной классификации, для возможности дальнейшего применения классических алгоритмов машинного обучения.

## 2. Рынок спортивных ставок

Существуют две основные категории ставок на хоккей: предматчевые и live-ставки, различающиеся уровнем коэффициентов. Кроме того, сделать ставку можно не только на победителя матча, но и на множество других факторов, например, на счет в отдельных периодах, победителя с учетом форы и т. д. Прогностические модели в основном ориентированы на предматчевые ставки на победителя, так как именно на этот тип ставок доступно больше всего исторических данных по коэффициентам, что позволяет провести наиболее полную оценку эффективности прогностической модели.

Ставки на хоккейные матчи можно размещать либо в букмекерских конторах (онлайн и оффлайн), либо на биржах ставок. Традиционные букмекеры, например Pinnacle, устанавливают коэффициенты на различные исходы матча, а клиент играет против букмекера.

Коэффициент ставки означает прибыль, которую получит клиент, если верно угадает исход события. Например, если клиент верно спрогнозировал победу команды, коэффициент на которую составляет 3.00, то он получит 2 доллара на каждый поставленныи доллар (в добавок к сумме самой ставки, которая возвращается). Если прогноз клиента оказался неверен, он теряет только сумму своей ставки независимо от коэффициентов.

Коэффициенты выражают предполагаемую вероятность исхода матча, то есть оценку букмекером истинной вероятности. В описанном выше примере с коэффициентом 3,00 (1 к 3) предполагаемая вероятность p победы игрока в матче равна 33%.

### 2.1. Обзор литературы

ИИ все чаще стал применяться к областям, связанным с интеллектуальными играми, такими как, к примеру, игра в го. Программа AlphaGo, разработанная компанией DeepMind (одна из дочерних компаний Google), выиграла у профессионального игрока пять игр подряд. Подробнее результат описан в [2]. Помимо го, системам ИИ покорилась еще одна сверхсложная игра – покер. В марте прошлого года канадские

программисты из университета Альберты создали искусственный разум DeepStack, способный играть в одну из простейших версий покера. Ему удалось стать победителем на одном из турниров по покеру, который проводился под эгидой Международной федерации покера. Подробнее результат описан в [3].

Спорт высоких достижений является инновационной областью применения искусственного интеллекта. В работе [4] авторы описали подход к прогнозированию спортивных событий на примере футбольных матчей. В итоге исследователи создали распределение описывающее футбольный матч и с помощью него получили "теоретический"перевес над букемерскими конторами в 5.5%. К недостаткам можно отнести то множество допущений, при которых было проведено исследование, что в итоге не позволило применить в жизни созданный алгоритм. Так как данные исследования имеют огромную финансовую значимость для рынка спортивных ставок, то других известных результатов в последние годы не было представлено.

### 2.2. Сбор статистических данных

Исторические данные по хоккейным матчам широко доступны в интернете. Официальные сайты турниров, например, www.khl.ru предоставляют информацию об игроках и результатах матчей, а также результативность спортсмена по каждому матчу. Некоторые источники предоставляют исторические данные в структурированнои форме (CSV или Excel фаилы). Доступны и платные базы данных – более комплексные, на более длинные периоды и с лучшей точностью. Наиболее релевантные данные, которые можно взять из подобных баз данных, представлены на рисунке 1.

## 3. Постановка задачи классификации

**Определение**: Пусть X - множество описаний объекта, Y - множество номеров классов. Существует целевая зависимость - y*: X $\to$ Y, значения которой известны на объектах конечной обучающей выборки $X^m = \{(x_1, y_1), ..., (x_m, y_m)\}$. Требуется построить алгоритм $z : X \to Y$, способный сопоставить произвольному объекту x $\in$ X, объект y $\in$ Y.

**Определение**: Пусть $Y^n = (y_1, ..., y_n)$ - множество событий в матче такого типа, при которых $X_i = \{x_{i1}, x_{i2}\}$ - множество исходов, состоит из двух элементов, где $x_{i1}$ означает, что событие вида $y_i$ наступило, а $x_{i2}$, что событие вида $y_i$ не наступило. Тогда исходы $x_{i1}$ и $x_{i2}$ назовем противоположными.


Рис. 1. пример статистических данных о матче

**Пример**: Пусть множество Y - это суммарное количество шайб заброшенных двумя командами, тогда события «было заброшено больше 4.5 шайб» и «было заброшено меньше 4.5 шайб» являются противоположными, а множество Y конечно ввиду ограниченности времени матча.

**Определение**: Введем функцию

$$f(x_{i1}, x_{i2}) = |x_{i1} - x_{i2}|, \tag{3.1}$$

где $x_{i1}, x_{i2}$ - коэффициенты на противоположные исходы события $y_i$, тогда событие $y_i$ будем называть оптимальным для прогнозирования если

$$f_i(x_{i1}, x_{i2}) = \min_{y_j \in Y} f(x_{j1}, x_{j2}) \tag{3.2}$$

Что это будет значить с точки зрения оценки вероятностей? Это значит, что события имеют коэффициенты около 2 и эти исходы практически равновероятны. Почему данный вид ставки будет оптимален? Обратимся к теории случайных процессов.

**Определение**: Одномерное случайное блуждание - это случайный процесс $\{Y_n\}_{n\geq 0}$ с дискретным временем, имеющим вид

$$Y_n = Y_0 + \sum_{i=1}^{n} X_i \tag{3.3}$$

где $Y_o$ - начальное состояние, $X_i$ -независимые случайные величины равные 1 с вероятностью p и -1 с вероятностью 1-p, если $p = q = \frac{1}{2}$ то блуждание симметрично.

**Теорема 1** (О возвратности блуждания)**.** *Вероятность* $P(Y_n = 0, n > 1|Y_0 = 0)$ *того, что случайное симметричное блуждание вернется в точку старта в одномерном случае равна 1.*

Подробное доказательство приводится в [5]

**Определение**: ROI (return on investment) - $ROI = \frac{P_n * 100}{s * n}$, где $P_n$ это прибыль на дистанции в n матчей, s-сумма одной ставки, n - количество ставок. ROI – это основнои показатель успешности и, соответственно, целевои показатель эффективности прогностическои модели.

**Замечание**: Очевидно, при событии вида «в сумме забросили меньше 4 шайб» пространство исходов состоит из трех элементов: забросили меньше 4 шайб, забросили ровно 4 шайбы, забросили больше 4 шайбы. Но по правилам спортивных ставок, в случае если результат совпал с прогнозируемым событием, то идет возврат ставки, поэтому можно считать что исхода два, так как при третьем не идет никакой потери.

**Определение**: Игрой "наугад" назовем игру, где на каждом шаге мы равновероятно выбираем один из двух возможных видов ставки.

**Гипотеза:** Благодаря выбору оптимальных событий для прогнозирования, даже используя стратегию ставок "наугад", мы играем с минимально возможным отрицательным матожданием для себя.

**Проверка гипотезы:** Если предположить, что букмекеры обладали бы истинными вероятностями событий и определяли величину ставки исходя из них, то тогда на большой дистацнии наш выигрыш был бы равен нулю. Это утверждение легко следует из расчета математачиеского ожидания выигрыша.

$$\begin{aligned} &\frac{1}{2} * \frac{1}{k_1} * (k_1 - 1) + \frac{1}{2} * (1 - \frac{1}{k_1}) * (-1) + \\ &+\frac{1}{2} * \frac{1}{k_2} * (k_2 - 1) + \frac{1}{2} * (1 - \frac{1}{k_2}) * (-1) = 0 \end{aligned} \tag{3.4}$$

где $k_1, k_2$ - коэффициенты на то, что событие произойдет или не произойдет. Очевидно, что для того, чтобы букмекерским конторам это стало выгодно, необходимо занижать коэффициенты вводя комиссию. Предположение состоит в том, что в зависимости от величины функции (3.1) размер комиссии разный и необходимо найти события, где при игре "наугад" на длинной дистанции мы будем иметь наименьшее отрицательное математическое ожидание. Проведем моделирование на выборке из 5000 матчей, где решение о выборе ставки принимается наугад и посчитаем среднее ROI при повторении симуляции 10000 раз. На всех возможных событиях для прогнозирования получили наименьшее отрицатель-

ное ROI при событиях оптимальных для прогнозирования, и оно составило $-4,8\%$ против $-7,2\%$ и $-9.4\%$ для неоптимальных. Моделирование для оптимальных событий на рисунке ниже, по оси x ROI (рис 2).

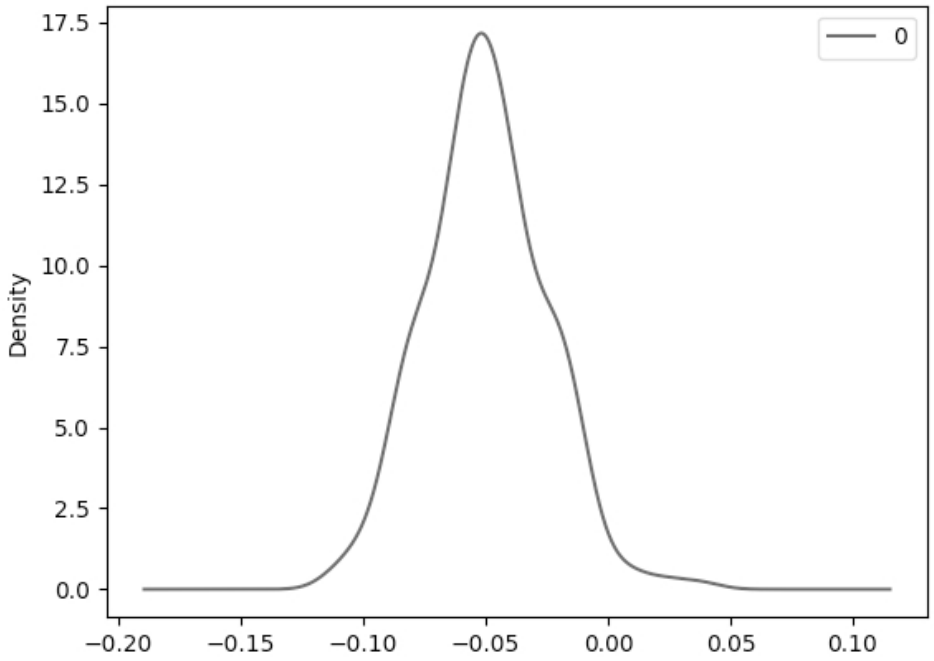


Рис. 2. Плотность распределения ROI.

## 4. Результаты и выводы

Опираясь на данные моделирования теперь мы можем сформулировать задачу бинарной классификации. В качестве множества событий матча выберем суммарное количество заброшенных шайб в основное время. Это множество конечно. Среди множества событий гарантировано есть событие, оптимальное для прогнозирования, которое позволяет нам использовать теорему 1, а значит минимизировать риски. В качестве целевой переменной будем использовать пространство исходов события, оптимального для прогнозирования. В качестве одной из метрик будем использовать ROI. Таким образом, мы сформулировали задачу бинарной классификации, которую можно теперь решать с помощью методов анализа данных.

## Список литературы


[1] http://azure.cnews.ru/articles/2017-02-14_hokkej_budushchego_intellektualnye_matchi_v_oblake, “Хоккей будущего: интеллектуальные матчи в "облаке"”.

[2] “Google AI algorithm masters ancient game of Go”, *Nature*, 2016, №529,, 445–446.

[3] “Superhuman AI for heads-up no-limit poker: Libratus beats top professionals”, *Science*, 2017.

[4] Lisandro Kaunitz, Shenjun Zhong , Javier Kreiner, “Beating the bookies with their own numbers - and how the online sports betting market is rigged”.

[5] Марк Кельберт, Юрий Сухов, “Вероятность и статистика в примерах и задачах. Том 2. Марковские цепи как отправная точка теории случайных процессов и их приложения”, 2017, 57.


## References


[1] http://azure.cnews.ru/articles/2017-02-14_hokkej_budushchego_intellektualnye_matchi_v_oblake, “Hokkey buduschego: intellektualnie matchi v oblake [Hockey of the future: intelligent matches in the "cloud] (in Russian)”.

[2] “Google AI algorithm masters ancient game of Go”, *Nature*, 2016, №529,, 445–446.

[3] “Superhuman AI for heads-up no-limit poker: Libratus beats top professionals”, *Science*, 2017.

[4] Lisandro Kaunitz, Shenjun Zhong , Javier Kreiner, “Beating the bookies with their own numbers - and how the online sports betting market is rigged”.

[5] Mark Kelbert, Yuri Suhov, “Veroyatnost i statistica v primerah i zadachyah: Tom 2. Markovskie cepi kak otpravnaya tochka teorii sluchainih processov i ih prilozhenia [Probability and statistics in examples and problems. Volume 2. Markov chains as a starting point of the theory of random processes and their applications] (in Russian)”, 2017, 57.


### Possible approach to the problem of predicting sports results using data analysis methods


**Zhuravlev A. D.**



This article is devoted to the construction of an algorithm for reducing the problem of predicting sports results to the problem of binary classification. At the same time, the optimality of this algorithm has been substantiated from the point of view of the application of forecasting results in the game against bookmakers.



*Keywords:* machine learning, predicting sports results, binary classification.