Исследовательский проект по курсу «Эконометрика-2» на тему:

Оценка вероятности быть выбранным на драфте НБА для игрока студенческой лиги (NCAA Division I)

Над проектом работали

Арзуманян Давид, 302

Левин Марк, 302

Рамазанов Аслан, 302

Тараканов Федор, 302

Штыканова Анна, 308

Оглавление

[Введение 2](#_Toc99306426)

[Обзор литературы 2](#_Toc99306427)

[Данные 4](#_Toc99306428)

[Модели 7](#_Toc99306429)

[Логит-модель 7](#_Toc99306430)

[Пробит-модель 9](#_Toc99306431)

[Линейно-вероятностная модель 11](#_Toc99306432)

[Сравнение моделей 11](#_Toc99306433)

[Итоговый выбор модели 11](#_Toc99306434)

[Заключение 11](#_Toc99306435)

[Список литературы 12](#_Toc99306436)

# Введение

Спорт и связанные с ним мероприятия интересуют многих людей. В центре внимания проектной работы находиться Национальная баскетбольная ассоциация. Почему для нашей работы мы выбрали именно этот вид спорта? Кроме личной симпатии к баскетболу, наблюдается активный поступательный рост его популярности. Баскетбол занимает твердое второе место по популярности среди всех видов спорта. НБА составляет 6% мирового спортивного рынка с точки зрения доходов, уступая лишь футболу с колоссальными 43%. Может быть, сложно подсчитать годовой доход такой непубличной компании, как НБА, однако Forbes сообщает, что в прошлом сезоне НБА заработала 8 миллиардов долларов, при этом каждая команда получила 255 миллионов долларов национального дохода от лиги. Каждая команда лиги сейчас стоит более 1 миллиарда долларов, а «Нью-Йорк Никс» имеет наибольшую стоимость в 4 миллиарда долларов.[[1]](#footnote-2)

Баскетбол имеет большую ценность и как телевизионный продукт. За последние несколько лет НБА пережила падение своих телевизионных рейтингов. Цифры восстановились после своего абсолютного минимума во время пандемии, но они все еще ниже примерно на 25% по сравнению с цифрами до COVID-19. Первая игра сезона показала рост в просмотрах на 15% по сравнению с финальным матчем «Лейкерс-Хит» в 2020 году, но на 35% меньше, чем в 2019 году. Тем не менее, НБА занимает лидирующие позиции в интернет-пространстве. Данные по подписчикам аккаунтов национальных спортивных лиг в ведущих социальных сетях (Twitter, YouTube и тд.) демонстрируют крайнюю популярность и доминирующую позицию Национальной Баскетбольной Ассоциации.[[2]](#footnote-3) Стремительный рост актуальности баскетбола в Китае также говорит лишь о позитивных прогнозах.

Достаточное количество людей, занимающихся тем или иным видом спорта, стремятся связать свое будущее с профессиональной карьерой спортсмена. Для достижения поставленной цели ежегодно происходят отборы в спортивные клубы. Самым популярным «транзитом» в большой спорт можно считать драфт. Драфт НБА — ежегодная процедура выбора клубами Национальной баскетбольной ассоциации молодых баскетболистов, желающих вступить в лигу. Обычно участники драфта – это студенты американских колледжей, но в последнее время увеличилось число иностранных игроков. Драфт состоит из двух раундов, в каждом из которых клубы выбирают по одному игроку в порядке, определяемой специальной лотереей, проводимой перед драфтом. При лотерее приоритет имеют клубы, показавшие в предшествующем сезоне худшие результаты и не попавшие в плей-офф. Это делается с целью обеспечения баланса сил в ассоциации и усиления слабых команд.

В преддверии драфта выходит множество статей, в которых спортивные эксперты и тренеры пытаются предугадать исход события[[3]](#footnote-4). Поэтому и мы решили на основе открытых данных выявить факторы, которые влияют на шансы игрока быть выбранным на драфте НБА. Именно этому и будет посвящена наша работа.

В начале работы над проектом мы выдвинули ряд гипотез, которые, на наш взгляд, должны подтвердиться в итоге:

**1) Влияет ли процент попадания штрафных бросков на вероятность попадания в НБА?**

Мы предполагаем, что да, так как этот показатель используется скаутами для прогнозирования броскового потенциала игрока.

**2) Зависит ли от команды, в которой играл баскетболист в студенческой лиге, прохождение драфта?**

Иначе говоря, мы предполагаем, что у игроков из более сильных команд и из больших конференций есть больше шансов перехода в профессиональную лигу.

**3) Если ли зависимость между эффективностью игрока в атаке/защите и результатами драфта?**

Мы считаем, что спортивные навыки и опыт игроков в определенных амплуа могут иметь большое влияние на попадание участника драфта в профессиональную лигу.

# Обзор литературы

Существует обширный блок литературы, посвященной исследованию показателей, влияющих на отбор игроков в лигу НБА.

Дэвид Дж. Берри, Стейси Л. Брук и Аджу Дж. Фенн (David J. Berri, Stacey L. Brook & Aju J. Fenn, 2010) рассмотрели в своей работе такое явление, как драфт, где худшие команды выбирают игроков в первую очередь. Этот институт является общим для всех основных профессиональных лиг в Северной Америке. Такая система разработана для того, чтобы предоставить более слабым командам доступ к будущим звездам спорта. В центре исследования авторов находится то, как информация о выступлениях игроков-любителей используется лицами, принимающими решения в командах Национальной баскетбольной ассоциации. Анализ показал, что будущие игроки НБА, которые набирают много очков в колледже, могут ожидать более высокого пика на драфте. Однако такой акцент на скоринге ухудшает качество принятия решений, так как лица, выбирающие игрока, часто учитывают факторы (например, последние четыре выступления непосредственно перед драфтом; иными словами, подвержены так называемому recency bias), которые не имеют отношения к будущим результатам.

В похожей статье Дэвид Дж. Берри и Роб Симмонс (Berri, D.J., Simmons, R.) выбрали в качестве объекта исследования драфты Национальной Футбольной Лиги (Американский футбол). Они искали зависимости между позицией игрока на драфте и последующими успехами в профессиональной карьере. Результаты статистического анализа показали очень слабую зависимость между итогами драфта и карьерными успехами. Большую роль в данной “неудаче” сыграли показатели роста, скорости рывка и академические успехи, которые главным образом учитываются экспертным жюри во время отбора в НФЛ, но, по мнению авторов статьи, мало влияют на игровые показатели. Проблематика данной статьи и объекты исследования сильно отличается от представленных нами, однако, некоторые методы анализа данных нам показались очень полезными и для нашей проектной работы. Авторы данной работы концентрируются исключительно на одной игровой позиции, квотербэке.

Моделирование драфта НБА не ограничивается данной работой. Джерад Х. Моксли и Тайлер Дж. Таун (Jerad H. Moxley, Tyler J. Towne,2014) в своей работе стремились проверить, в какой степени карьера в НБА может быть предсказана переменными, отражающими прошлые результаты, и переменными, демонстрирующими неиспользованный потенциал. Авторы создали строгую модель успеха в начале карьеры игроков Национальной Баскетбольной ассоциации (НБА). В ходе анализа было выяснено, что ни одна из переменных не подтвердила гипотезу о «неиспользованном потенциале». Исследование доказало важность роли тренировочной среды вместе со стабильностью навыков и показало явные преимущества «талантливого» игрока. Эффект Мэтью, описанный в работе, объясняет, как статус «таланта» обеспечивает неравномерное распределение ценности игроков с самого детства и часто лишь усиливается со временем.

Идея об исследовании вероятности попадания в Национальную баскетбольную ассоциацию путем драфта является результатом комбинирования различных гипотез и результатов и других научных работ. И это «вдохновение» имеет свое место не только на идейном уровне, но и на уровне моделирования самого проекта. В статье Питера А. Гротуиса, Джеймса Ричарда Хилла и Тимоти Перри (Groothuis, P. A., Hill, J. R., & Perri, T.) непосредственно применяется вероятностная модель выбора клубом игрока, участвующего в драфте. В оригинальной модели авторы использовали агрегированные показатели, кроме того, сами данные охватывают довольно ранний временной промежуток (1987 – 2004). В проектной работе же не используются агрегированные данные с целью выявления значимой зависимости между вероятностью заключения контракта с баскетбольной командой и каждым чистым показателем отдельно. Кроме того, ранняя дата публикации рассматриваемого исследования имеет риск потери актуальности результатов в 2022 году, что говорит о необходимости применения «свежих» данных.

Статья Колина Ф. Камерера и Роберто А. Вебера (Camerer, C. F., & Weber, R. A.) раскрывает тему попадания игроков в НБА с иной стороны. В центре внимания все также остаются результаты драфтов, однако авторы применяли результаты отбора в НБА и последующие игровые показатели с целью определения причин иррациональных действий руководства клуба (в статье это называется эскалацией), когда игрок, не защитивший свои драфтовые результаты и не оправдавший звездные ожидания клуба, по-прежнему активно выступает за команду и имеет длительное время на площадке. Поведенческие методы анализа выявили ряд оснований появлению эскалации в клубе, имеющих экономичную и психологическую природу. Данная статья была первой изучена для проекта по эконометрике и стала главным факторов выбора тематики студенческого исследования, однако было принято решение отойти от поведенческой проблематики с целью обеспечения реализуемости проекта.

Делая упор на ранее выпущенные работы, мы собираемся провести свое исследование.

# Данные

Для построения модели нами был проведен необходимый анализ данных, собранным по игрокам студенческих лиг. Итоговый датасет был отобран по следующему принципу:

• Изначально были собраны данные игроков за 5 сезонов (2017–2021) вместе с индивидуальной статистикой каждого участника. В игровой статистике была учтена позиция, рост, вес и другие основные игровые показатели[[4]](#footnote-5).

• Количество позиций при анализе было сокращено с 5 (point guard, shouting guard, small forward, power forward, center)[[5]](#footnote-6) до 3 (guard, forward, center)[[6]](#footnote-7). Это было сделано, потому что при отборе амплуа участников не носят закрепленного характера.

• Игроки для начальной выборки отбирались с учетом конференции[[7]](#footnote-8). На основе отбора была введена фиктивная переменная «big conference», в которой за 1 обозначены команды, состоящие в одной из крупных конференций, за 0 – все остальные команды.

• Были также учтены соотношения побед и поражений команд в игровом сезоне.

• В конце мы собрали финальные данные драфтов 2017–2021 годов. В данной выборке осталось всего 250 игроков.

Данные были взяты на трех сайтах, на которых была представлена официальная информация относительно ежегодных драфтов:

1. [NCAA DI Basketball Players - RealGM](https://basketball.realgm.com/ncaa/players/)

2. [2021-22 School Stats | College Basketball at Sports-Reference.com](https://www.sports-reference.com/cbb/seasons/2022-school-stats.html)

3. [NCAA Basketball Stats - NCAA BB Team Win % - All Games | TeamRankings.com](https://www.teamrankings.com/ncaa-basketball/stat/win-pct-all-games?date=2021-04-06)

Итоговые переменные, которые будут использоваться для анализа, отражены в таблице 1 ниже.

**Таблица 1: Показатели для анализа**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Регрессор | Показатель | Предполагаемый характер связи | Коэффициент корреляции между регрессором и переменной Drafted |
| 1 | d\_2017 | Фиктивна переменная для года окончания сезона (1 – для 2019 года, 0 – для остальных | +/- | -0,0019 |
| 2 | d\_2018 | Фиктивна переменная для года окончания сезона (1 – для 2019 года, 0 – для остальных | +/- | -0,0098 |
| 3 | d\_2019 | Фиктивна переменная для года окончания сезона (1 – для 2019 года, 0 – для остальных | +/- | 0,0024 |
| 4 | d\_2020 | Фиктивна переменная для года окончания сезона (1 – для 2019 года, 0 – для остальных | +/- | -0,0036 |
| 5 | Guard | Фиктивная переменная амплуа игрока на площадке  (1 - для игроков защиты, 0 - для остальных) | +/- | 0,0022 |
| 6 | Forward | Фиктивная переменная амплуа игрока на площадке  (1 – для игроков нападения, 0 – для остальных) | +/- | 8,1092e-03 |
| 7 | Center | Фиктивная переменная амплуа игрока на площадке  (1 – для игроков центра, 0 – для остальных) | +/- | -0,0147 |
| 8 | Big\_conference | Фиктивная переменная принадлежности команды к большой конференции  (1 – для игрока из команды большой конференции, 0 – для остальных) | + | 1,229989e-01 |
| 9 | Age | Возраст, в годах | + | 0,0646 |
| 10 | Win\_perc | Процент выигранных матчей, в процентах | + | -4,808645e-03 |
| 11 | Weight | Вес игрока, в килограммах | +/- | 0,0263 |
| 12 | Height | Рост игрока, в сантиметрах | + | 0,0286 |
| *Игровые показатели представлены в среднем значении, рассчитанном на 40 минут игры* | | | | |
| 13 | GP | Число сыгранных матчей | + | 0,0789 |
| 14 | GS | Число матчей, в которых игрок вышел со старта | + | 0,1348 |
| 15 | MIN | Число сыгранных минут | + | 0,1396 |
| 16 | PTS | Число набранных очков | + | 0,1323 |
| 17 | FGM | Количество заброшенных мячей | + | 0,1216 |
| 18 | FGA | Количество бросков | + | 0,0986 |
| 19 | FG% | Процент попаданий | + | 0,0502 |
| 20 | 3PM | Количество успешных трехочковых бросков | + | 0,0469 |
| 21 | 3PA | Количество трехочковых бросков | +/- | 0,0286 |
| 22 | 3P% | Процент попадания трехочковых бросков | + | 0,0599 |
| 23 | FTM | Количество заброшенных штрафных | + | 0,0946 |
| 24 | FTA | Количество штрафных бросков | +/- | 0,0671 |
| 25 | FT% | Процент попадания штрафных бросков | + | 0,0675 |
| 26 | OFF | Количество подборов в атаке | + | 0,00097 |
| 27 | DEF | Количество подборов в защите | + | 0,0509 |
| 28 | TRB | Общее количество подборов мяча | + | 0,3,6486e-02 |
| 29 | AST | Количество результативных передач | + | 0,0656 |
| 30 | STL | Количество перехватов мяча | + | 0,0239 |
| 31 | BLK | Число заблокированных ударов | + | 0,0350 |
| 32 | TOV | Число совершенных возвратов мяча | + | 0,0133 |
| 33 | PF | Персональные предупреждения | - | -0,0471 |

# Модели

Теперь перейдем к проверке выдвинутых гипотез на основе моделей.

## Логит-модель

Для начала попробуем построить логит-модель. По нашему предположению, а также на основании результатов работ многих исследователей, именно данная спецификация является наиболее подходящей.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Модель1. Логит, использованы наблюдения 1–17221  Зависимая переменная: Drafted  Стандартные ошибки QML | | | | |
| Регрессор | Коэффициент | Стандартная ошибка | z | P-значение |
| Const | -42,8922 | 4,0828 | -10.5057 | <2.2e-16\*\*\* |
| Age | -0,5163 | 0,0662 | -7.7959 | 6.398e-15\*\*\* |
| Win\_perc | -1,5250 | 0,4857 | -3.1397 | 0.0016910\*\* |
| Guard | 1,4828 | 0,3705 | 4.0021 | 6.278e-05\*\*\* |
| Forward | 0,8070 | 0,2629 | 3.0694 | 0.0021448\*\* |
| Big\_conference | 1,8095 | 0,1777 | 10.1819 | <2.2e-16\*\*\* |
| Height | 0,1265 | 0,0167 | 7.5816 | 3.412e-14\*\*\* |
| GP | 0,1479 | 0,0286 | 5.1694 | 2.348e-07\*\*\* |
| GS | -0,0482 | 0,0134 | -3.5963 | 0.0003228\*\*\* |
| MIN | 0,2185 | 0,0236 | 9.2604 | <2.2e-16\*\*\* |
| FGA | 0,1098 | 0,0325 | 3.3742 | 0.0007404\*\*\* |
| `FG%` | 11,8311 | 1,4585 | 8.1116 | 4.997e-16\*\*\* |
| `3PA` | 0,1602 | 0,0378 | 4.2404 | 2.231e-05\*\*\* |
| `3P%` | 1,2129 | 0,5410 | 2.2418 | 0.0249755\* |
| FTA | 0,1098 | 0,0385 | 2.8480 | 0.0043997\*\* |
| `FT%` | 5,6620 | 0,9399 | 6.0241 | 1.700e-09\*\*\* |
| TRB | 0,1108 | 0,0406 | 2.7283 | 0.0063658\*\* |
| AST | 0,3385 | 0,0484 | 6.9903 | 2.743e-12\*\*\* |
| STL | 0,3650 | 0,0498 | 7.3242 | 2.403e-13\*\*\* |
| BLK | 0,4091 | 0,0479 | 8.5397 | <2.2e-16\*\*\* |
| PF | -0,2271 | 0,1086 | -2.0925 | 0.0363982\* |
| d\_2017 | -1,0887 | 0,3210 | -3.3917 | 0.0006945\*\*\* |
| d\_2018 | -1,5920 | 0,3406 | -4.6742 | 2.951e-06\*\*\* |
| d\_2019 | -1,1370 | 0,3117 | -3.6477 | 0.0002646\*\*\* |
| d\_2020 | -0,8648 | 0,2612 | -3.3116 | 0.0009276\*\*\* |
| Макфадена | | 0,446 | | |
| Критерий Акаике | | 1474,8 | | |

Примечания. Зависимая переменная - выбор игрока на драфте. В скобках указаны стандартные ошибки. \* обозначает значимость на 10%-м уровне; \*\* обозначает значимость на 5%-м уровне; \*\*\* обозначает значимость на 1%-м уровне.

Изначально была построена регрессия на все 31 переменную. После теста на критерий Акаике осталось 25 регрессоров. Далее мы использовали робастные стандартные ошибки (QML). В результате получилась наилучшая модель. Все переменные значимы, большая часть на 1% уровне.

Далее проверим, имеет ли смысл добавлять фиктивные переменные для амплуа на площадке и года игрового сезона. Для этого построим две вспомогательных модели – одну без фиктивных переменных для позиции и одну без фиктивных переменных для сезонов. Далее проведем два теста отношения правдоподобия.

Нулевой гипотезой здесь будет одновременное равенство нулю коэффициентов при избыточных переменных в «длинной» модели, альтернативной – то, что эти переменные значимы, то есть то, что «длинная» модель лучше. Р-значения получились 3,7\*10^(-4) и 8\*10^(-6) соответственно. В обоих случаях дополнительные переменные оказались значимы и «длинная» модель имела лучшие описательные свойства.

Чтобы оценить качество прогнозирования нашей модели, мы разделили прогнозные значения на два класса: 0 – те игроки, кому модель не прогнозирует успеха на драфте, 1 – будущие профессиональные спортсмены. В качестве порога вероятности было выбрано значение 2,5% - при нем точность прогноза выше 90% как для значений зависимой переменной, равным нулю, так и для единицы. Это значение выбиралось вручную, т. к. функции поиска оптимального порога максимизируют общую точность классификации. В том случае, когда количество наблюдений, в которых значение зависимой переменной равно единице, значительно меньше количества наблюдений, где зависимая переменная равна нулю, точность прогнозирования значительно ухудшится.

Таблица сопряженности выглядит следующим образом:

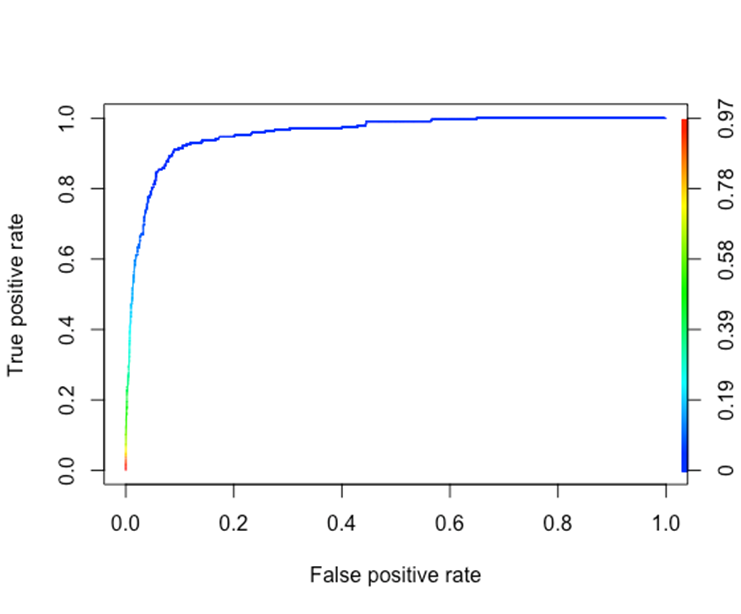
|  |  |  |
| --- | --- | --- |
|  | Drafted = 0 | Drafted = 1 |
|  | 15 501 | 24 |
|  | 1 475 | 221 |

Доля корректно спрогнозированных наблюдений всего: 91,3%

Доля корректно спрогнозированных наблюдений для Drafted = 0: 91,31%

Доля корректно спрогнозированных наблюдений для Drafted = 1: 90,2%

Такое низкое значение вероятности, выбранное в качестве порога, может свидетельствовать о высоком уровне конкуренции в студенческих лигах и о том, что редко встречаются атлеты, которым гарантировано начало профессиональной карьеры.

Представить визуально качество бинарной классификации можно с помощью ROC – кривой. Этот график отображает соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущие признак (True Positive rate), и долей объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущие признак (False Positive rate) при изменении порога вероятности (задан цветом кривой). Значение AUC равно 0,958 и близко к единице, что также говорит о высоком качестве классификации.

## Пробит-модель

Для оценки вероятности прохождения также построим пробит-модель.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Модель 2. Пробит, использованы наблюдения 1–17221  Зависимая переменная: Drafted  Стандартные ошибки QML | | | | |
| Регрессор | Коэффициент | Стандартная ошибка | z | P-значение |
| Const | -18,8099 | 2,1024 | -8.9469 | <2.2e-16\*\*\* |
| Age | -0,2306 | 0,0365 | -6.3100 | 2.790e-10\*\*\* |
| Win\_perc | -0,7838 | 0,2274 | -3.4470 | 0.0005669\*\*\* |
| Guard | 0,6779 | 0,1761 | 3.8490 | 0.0001186\*\*\* |
| Forward | 0,4161 | 0,1234 | 3.3734 | 0.0007424\*\*\* |
| Big\_conference | 0,8529 | 0,0831 | 10.2604 | <2.2e-16\*\*\* |
| Height | 0,0546 | 0,0081 | 6.7368 | 1.619e-11\*\*\* |
| GP | 0,0617 | 0,0133 | 4.6479 | 3.353e-06\*\*\* |
| GS | -0,0225 | 0,0060 | -3.7545 | 0.0001737\*\*\* |
| MIN | 0,0947 | 0,0130 | 7.2769 | 3.416e-13\*\*\* |
| FGA | 0,0418 | 0,0152 | 2.7466 | 0.0060212\*\* |
| `FG%` | 4,9614 | 0,7924 | 6.2610 | 3.826e-10\*\*\* |
| `3PA` | 0,0771 | 0,0172 | 4.4767 | 7.580e-06\*\*\* |
| `3P%` | 0,5773 | 0,2433 | 2.3727 | 0.0176564\* |
| FTA | 0,0548 | 0,0177 | 3.0999 | 0.0019360\*\* |
| `FT%` | 2,5966 | 0,4269 | 6.0823 | 1.185e-09\*\*\* |
| TRB | 0,0640 | 0,0208 | 3.0792 | 0.0020754\*\* |
| AST | 0,1593 | 0,0226 | 7.0578 | 1.692e-12\*\*\* |
| STL | 0,1533 | 0,0256 | 5.9901 | 2.097e-09\*\*\* |
| BLK | 0,1935 | 0,0268 | 7.2238 | 5.055e-13\*\*\* |
| PF | -0,0911 | 0,0530 | -1.7192 | 0.0855730\* |
| d\_2017 | -0,4758 | 0,1540 | -3.0898 | 0.0020028\*\* |
| d\_2018 | -0,7077 | 0,1683 | -4.2057 | 2.603e-05\*\*\* |
| d\_2019 | -0,5277 | 0,1504 | -3.5094 | 0.0004492\*\*\* |
| d\_2020 | -0,4120 | 0,1253 | -3.2890 | 0.0010054\*\* |
| Макфадена | | 0,435 | | |
| Критерий Акаике | | 1512,7 | | |

Примечания. Зависимая переменная - выбор игрока на драфте. В скобках указаны стандартные ошибки. \* обозначает значимость на 10%-м уровне; \*\* обозначает значимость на 5%-м уровне; \*\*\* обозначает значимость на 1%-м уровне.

После теста на критерий Акаике в пробит-модели осталось 25 регрессоров. Далее мы использовали робастные стандартные ошибки (QML). Все переменные значимы, большая часть на 1% уровне. Макфадена ниже, чем у логит-модели.

Как и с логистической регрессией, мы провели два аналогичных LR – теста, чтобы выяснить значимость фиктивных переменных. P-значения оказались равны и соответственно, то есть включение в модель фиктивных переменных амплуа на площадке и времени улучшило ее качество.

В качестве порога вероятности для классификации было выбрано значение 2,75%. При нем точность прогнозирования оказалась выше 90% как для значений зависимой переменной, равным нулю, так и для единицы.

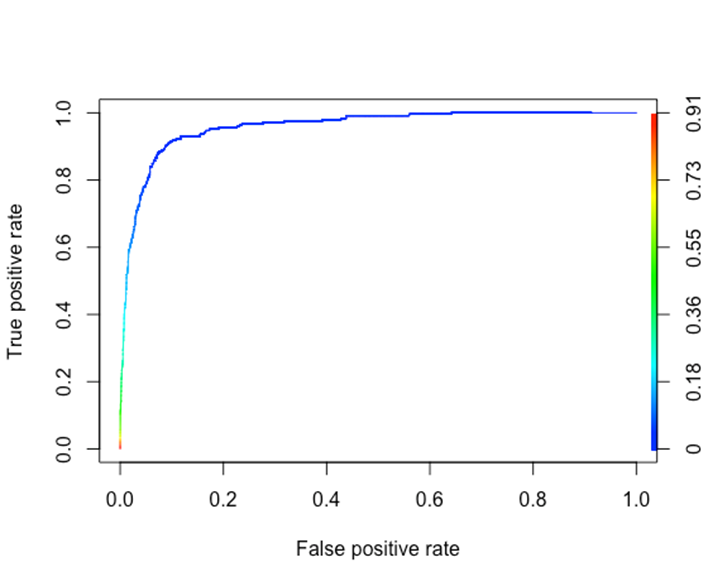
Таблица сопряженности выглядит следующим образом:

|  |  |  |
| --- | --- | --- |
|  | Drafted = 0 | Drafted = 1 |
|  | 15 405 | 23 |
|  | 1 571 | 222 |

Доля корректно спрогнозированных наблюдений всего: 90,74%

Доля корректно спрогнозированных наблюдений для *Drafted = 0*: 90,75%

Доля корректно спрогнозированных наблюдений для *Drafted = 1*: 90,61%

ROC – кривая:

AUC равно 0,959.

# Линейно-вероятностная модель

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Модель 3.Линейно-вероятностная модель, использованы наблюдения 1–17221  Зависимая переменная: Drafted  Стандартные ошибкиHC1 | | | | |
| Регрессор | Коэффициент | Стандартная ошибка | z | P-значение |
| **Const** | -0,1242 | 0,0332 | -3.7375 | 0.0001865\*\*\* |
| Age | -0,0058 | 0,0007 | -8.6394 | <2.2e-16\*\*\* |
| Win\_perc | -0,0144 | 0,0054 | -2.6724 | 0.0075375\*\* |
| Guard | 0,0085 | 0,0030 | 2.8480 | 0.0044041\*\* |
| Big\_conference | 0,0270 | 0,0023 | 11.6395 | <2.2e-16\*\*\* |
| Height | 0,0010 | 0,0002 | 6.1059 | 1.044e-09\*\*\* |
| GS | 0,0004 | 0,0001 | 2.6254 | 0.0086612\*\* |
| MIN | 0,0009 | 0,0002 | 5.2951 | 1.204e-07\*\*\* |
| FGA | 0,0015 | 0,0002 | 8.7620 | <2.2e-16\*\*\* |
| FTA | 0,0012 | 0,0002 | 5.2914 | 1.229e-07\*\*\* |
| TRB | 0,0005 | 0,0002 | 2.4031 | 0.0162694\* |
| AST | 0,0033 | 0,0006 | 5.6063 | 2.099e-08\*\*\* |
| BLK | 0,0029 | 0,0008 | 3.4935 | 0.0004780\*\*\* |
| PF | -0,0009 | 0,0003 | -3.0132 | 0.0025888\*\* |

Примечания. Зависимая переменная - выбор игрока на драфте. В скобках указаны стандартные ошибки. \* обозначает значимость на 10%-м уровне; \*\* обозначает значимость на 5%-м уровне; \*\*\* обозначает значимость на 1%-м уровне.

В линейно-вероятностную модель первоначально мы также добавили все имеющиеся переменные. На первом этапе мы убрали незначимые переменные, ещё часть отбросили благодаря критерию Акаике. Использовали стандартные ошибки НС1, так как тест Уайта показал наличие гетероскедостичности. В итоге осталось 14 регрессоров.

R^2исправленный мал, тест Рамсея показал неправильную спецификацию. Мультиколлиниарности нет, показатель VIF меньше 10 для всех переменных.

Модель неудачная.

## Сравнение моделей

|  |  |  |  |
| --- | --- | --- | --- |
| Сводная таблица трёх моделей. Зависимая переменная: Drafted | | | |
|  | Логит | Пробит | Линейно-вероятностная |
| Age | -0.516\*\*\* | -0.231\*\*\* | -0.006\*\*\* |
|  | (0.061) | (0.028) | (0.001) |
| Win\_perc | -1.525\*\*\* | -0.784\*\*\* | -0.014\*\*\* |
|  | (0.486) | (0.230) | (0.005) |
| Guard | 1.483\*\*\* | 0.678\*\*\* | 0.008\*\*\* |
|  | (0.391) | (0.185) | (0.003) |
| Forward | 0.807\*\*\* | 0.416\*\*\* |  |
|  | (0.296) | (0.142) |  |
| Big\_conference | 1.810\*\*\* | 0.853\*\*\* | 0.027\*\*\* |
|  | (0.181) | (0.082) | (0.002) |
| Height | 0.126\*\*\* | 0.055\*\*\* | 0.001\*\*\* |
|  | (0.017) | (0.008) | (0.0002) |
| GP | 0.148\*\*\* | 0.062\*\*\* |  |
|  | (0.023) | (0.010) |  |
| GS | -0.048\*\*\* | -0.022\*\*\* | 0.0004\*\*\* |
|  | (0.014) | (0.006) | (0.0001) |
| MIN | 0.218\*\*\* | 0.095\*\*\* | 0.001\*\*\* |
|  | (0.025) | (0.011) | (0.0002) |
| FGA | 0.110\*\*\* | 0.042\*\*\* | 0.002\*\*\* |
|  | (0.035) | (0.016) | (0.0002) |
| `FG%` | 11.831\*\*\* | 4.961\*\*\* |  |
|  | (1.394) | (0.647) |  |
| `3PA` | 0.160\*\*\* | 0.077\*\*\* |  |
|  | (0.041) | (0.019) |  |
| `3P%` | 1.213\* | 0.577\* |  |
|  | (0.645) | (0.302) |  |
| FTA | 0.110\*\*\* | 0.055\*\*\* | 0.001\*\*\* |
|  | (0.042) | (0.020) | (0.0003) |
| `FT%` | 5.662\*\*\* | 2.597\*\*\* |  |
|  | (0.991) | (0.454) |  |
| TRB | 0.111\*\*\* | 0.064\*\*\* | 0.001\* |
|  | (0.037) | (0.018) | (0.0003) |
| AST | 0.339\*\*\* | 0.159\*\*\* | 0.003\*\*\* |
|  | (0.051) | (0.024) | (0.001) |
| STL | 0.365\*\*\* | 0.153\*\*\* |  |
|  | (0.063) | (0.030) |  |
| BLK | 0.409\*\*\* | 0.193\*\*\* | 0.003\*\*\* |
|  | (0.055) | (0.028) | (0.001) |
| PF | -0.227\*\* | -0.091\* | -0.001\*\* |
|  | (0.106) | (0.048) | (0.0003) |
| d\_2017 | -1.089\*\*\* | -0.476\*\*\* |  |
|  | (0.290) | (0.136) |  |
| d\_2018 | -1.592\*\*\* | -0.708\*\*\* |  |
|  | (0.302) | (0.141) |  |
| d\_2019 | -1.137\*\*\* | -0.528\*\*\* |  |
|  | (0.288) | (0.137) |  |
| d\_2020 | -0.865\*\*\* | -0.412\*\*\* |  |
|  | (0.248) | (0.119) |  |
| Const | -42.892\*\*\* | -18.810\*\*\* | -0.124\*\*\* |
|  | (4.130) | (1.939) | (0.037) |
| Критерий Акаике | 1,474.835 | 1,501.799 |  |
| исправленный |  |  | 0,048 |
| Число наблюдений | 17,221 | 17,221 | 17,221 |
| Макфадена | 0,446 | 0,435 |  |

Примечания. Зависимая переменная - выбор игрока на драфте. В скобках указаны стандартные ошибки. \* обозначает значимость на 10%-м уровне; \*\* обозначает значимость на 5%-м уровне; \*\*\* обозначает значимость на 1%-м уровне.

Худшей из трех моделей оказалась линейная модель: тест Рамсея указал на ее неверную спецификацию. Также значения прогнозируемой вероятности могут «вылетать» за допустимые границы 0 и 1, из-за чего применение этой модели на практике становится невозможным.

Логит и пробит модели дают близкие результаты и оказываются в целом схожими друг с другом. Тем не менее, мы все-таки отдаем предпочтение логит модели. R^2 Макфаддена оказывается немного выше, а значение критерия Акаике, наоборот, ниже.

# Заключение

Подводя итог нашей работы, посмотрим на результаты проведенного нами анализа. Мы сделали следующие выводы, опираясь на построенные в проекте модели:

1) Нами изначально было выдвинуто предположение, что процент попадания штрафных бросков на вероятность попадания в НБА. Как общее количество забитых трехочковых бросков, так и доля успешных трехочковых значимо влияют. Это можно увидеть по коэффициентам в итоговой модели при переменных «3PM» и «3P%». Данная гипотеза подтвердилась.

2)Также нас интересовал вопрос о влиянии команды, в которой играл баскетболист в студенческой лиге, на прохождение драфта. Показатели Win\_perc и Big\_conference продемонстрировали положительное влияние на шансы участника драфта стать игроком профессиональной лиги. Высокий процент выигранный матчей судит об успешности команды, в которой играет участник драфта. Кроме того, большая конференция также судит о высоком статусе команды. Эти два показателя вместе непосредственно оказывают влияние на отбор игрока в профессиональные клубы. Гипотеза оказалась верной.

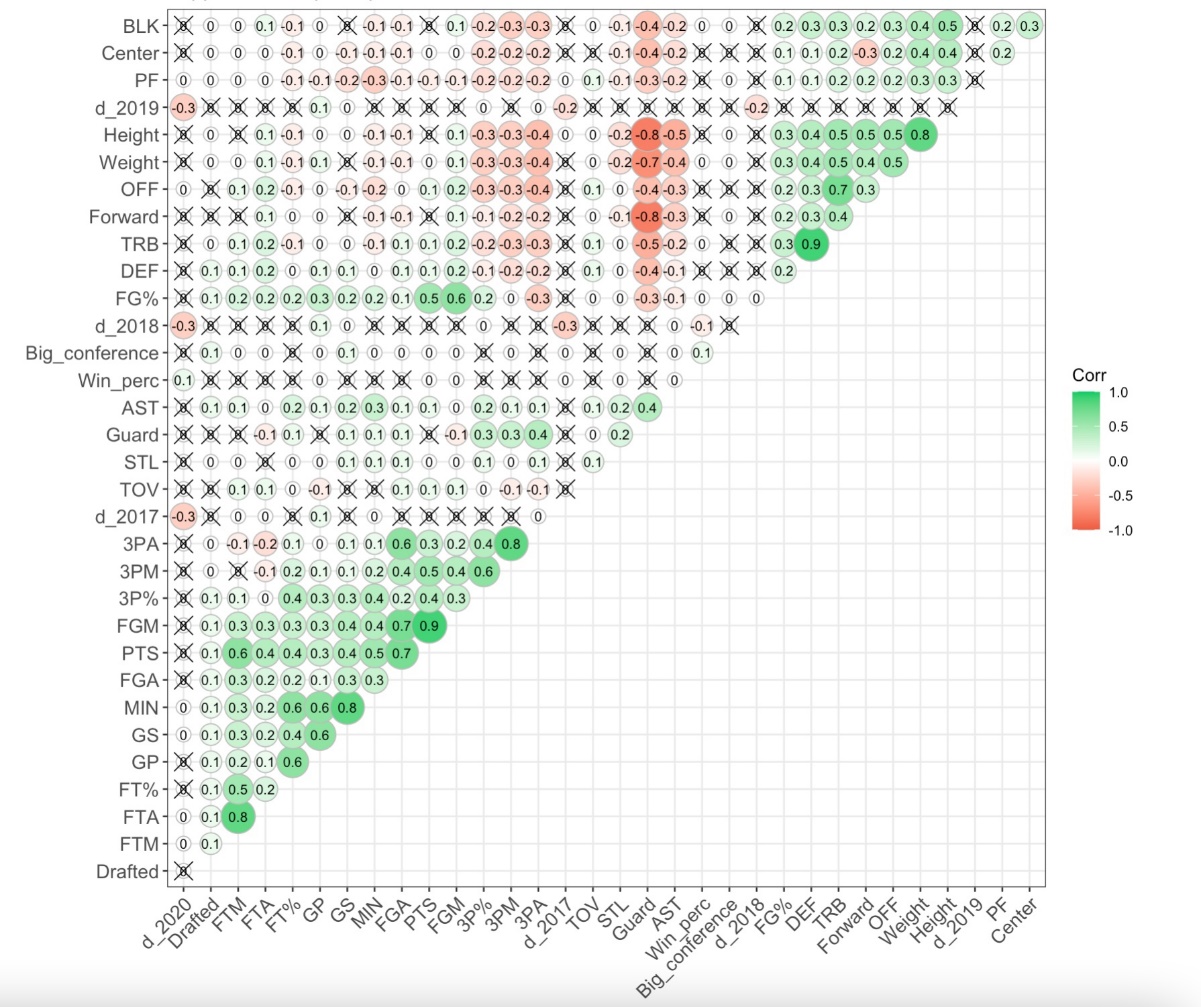
3) В начале нашего исследования мы предположили, что есть зависимость между эффективностью игрока и результатами драфта. В течение работы мы заметили, что форвардам и центровым игрокам на студенческом уровне легче доминировать на площадке за счёт физических данных. Однако, при отборе в НБА предъявляются высокие требования к атлетизму всех игроков. Таким образом, данное предположение отвергается.

# Список литературы

* Berri, D.J., Brook, S.L. & Fenn, A.J. From college to the pros: predicting the NBA amateur player draft. *Journal of Productivity Analysis* **35**, 25–35 (2011). <https://doi.org/10.1007/s11123-010-0187-x>
* Berri, D.J., Simmons, R. Catching a draft: on the process of selecting quarterbacks in the National Football League amateur draft. *Journal of Productivity Analysis* **35**, 37–49 (2011). <https://doi.org/10.1007/s11123-009-0154-6>
* Moxley, J. H., & Towne, T. J. Predicting success in the National Basketball Association: Stability & potential. *Psychology of Sport and Exercise* **16**, 128–136 (2015). <https://doi.org/10.1016/j.psychsport.2014.07.003>
* Groothuis, P. A., Hill, J. R., & Perri, T. The dilemma of choosing talent: Michael Jordans are hard to find. *Applied Economics* **41(25)**, 3193–3198 (2009). <https://doi.org/10.1080/00036840701564459>
* Camerer, C. F., & Weber, R. A. The econometrics and behavioral economics of escalation of commitment: A re-examination of Staw and Hoang’s NBA data. *Journal of Economic Behavior & Organization* **39(1)**, 59–82 (1999). https://doi.org/10.1016/S0167-2681(99)00026-8
* Coates, D., & Oguntimein, B. The length and success of nba careers: Does college production predict professional outcomes? (Issue 0806; Working Papers). *International Association of Sports Economists & North American Association of Sports Economists* (2008)*.* <https://ideas.repec.org/p/spe/wpaper/0806.html>

# Приложения

**Приложение 1:Корреляционная матрица**



1. [The Multiple Revenue Streams Of The NBA by Harvey Carr](https://businessingmag.com/9825/money/nba/) [↑](#footnote-ref-2)
2. [TV Ratings Are Down, but the NBA’s Popularity Isn’t by Johnathan Tillman](https://boardroom.tv/nba-popularity-tv-ratings/) [↑](#footnote-ref-3)
3. [2022 NBA Mock Draft: Predicting both rounds after March Madness start (usatoday.com)](https://ftw.usatoday.com/lists/nba-mock-draft-march-madness-chet-holmgren-paolo-banchero-jaden-ivey) [↑](#footnote-ref-4)
4. все они представлены в **Таблица 1: Показатели для анализа** [↑](#footnote-ref-5)
5. разыгрывающий защитник, атакующий защитник, лёгкий форвард, тяжёлый (или мощный) форвард и центровой [↑](#footnote-ref-6)
6. защитник, форвард и центровой [↑](#footnote-ref-7)
7. выбирались все игроки, которые есть в первом дивизионе (всего там 32 конференции, из них 8 считаются крупными) [↑](#footnote-ref-8)