

МИНОБРНАУКИ РОССИИ
Федеральное государственное автономное образовательное
учреждение высшего образования
«Южный федеральный университет»
Институт высоких технологий и пьезотехники



**Кафедра прикладной информатики и
инноватики**

Направление подготовки: 09.03.03

«Прикладная информатика»

Отчет к проекту
“Сбор, предобработка и анализ данных о статистике игроков NBA”
По дисциплине
“Большие данные”

Выполнили:

студенты 3 курса 7 группы

подпись

Головин И.Н.

подпись

Соскин А.И.

Ростов-на-Дону 2024 г.

Цель работы

Цель данной работы заключается в сборе, предобработке и анализе данных о статистике игроков NBA. Основной задачей является прогнозирование игровых позиций баскетболистов с использованием методов линейной и логистической регрессий на основе их игровых статистик по сезонам.

Актуальность

Анализ статистики игроков NBA и прогнозирование их игровых позиций имеют большое значение для команд, тренеров и аналитиков. Это позволяет оптимизировать состав команды, планировать стратегии игры и развивать молодые таланты. В условиях высокого уровня конкуренции в NBA, точное предсказание позиций и ролей игроков на основе статистических данных помогает принимать более обоснованные решения, что повышает шансы на успех команды.

Гипотеза

Основная гипотеза работы заключается в том, что статистические показатели игроков NBA (например, количество очков, подборов, передач, перехватов и блок-шотов) могут быть использованы для точного прогнозирования их игровых позиций на площадке (например, разыгрывающий защитник, атакующий защитник, легкий форвард, тяжелый форвард или центровой) с помощью методов линейной и логистической регрессий.

В качестве решения предлагаем построить графики для визуализации данных, где основными полями являются:

"Year" - год,
"Age" – возраст игрока,
"G" - игры,
"TS%" – процент бросков команды,
"ORB%" – процент подборов мяча,
"DRB%" – дриблинг или владение,
"TRB%" – процент общего количества подборов мячей,
"AST%" – процент ассистов (помощи),
"STL%" – процент отбора мяча,
"BLK%" - процент заблокированных бросков,
"TOV%" – процент потерь,
"USG%" – процент командных приемов игрока,
"FG%" – процент бросков в корзину,
"3P%" – процент трехочковых бросков,
"2P%" – процент двухочковых бросков,
"eFG%" – процент результативных забитых очков на поле,
"FT%" – процент штрафных бросков,
"PF" – личные фолы (нарушения),
"PTS" – очки.

Обработка этих столбцов пригодится нам для прогнозирования позиций баскетболистов:

Point Guard – Разыгрывающий Защитник.

Shooting Guard – Атакующий Защитник.

Small Forward – Лёгкий форвард.

Power Forward – Тяжёлый/мощный форвард.

Center – Центровой.



Описание датасета

Для проектной работы был использован датасет из сайта kaggle.com. Данные были взяты из справочника по баскетболу basketball-reference.com

Датасет представляет собой совокупную индивидуальную статистику за 67 сезонов в НБА. От базовых показателей результативности, таких как очки, передачи, подборы и т.д., до более продвинутых функций.

Seasons_Stats.csv (5.12 MB)

Detail Compact Column 20 of 53 columns

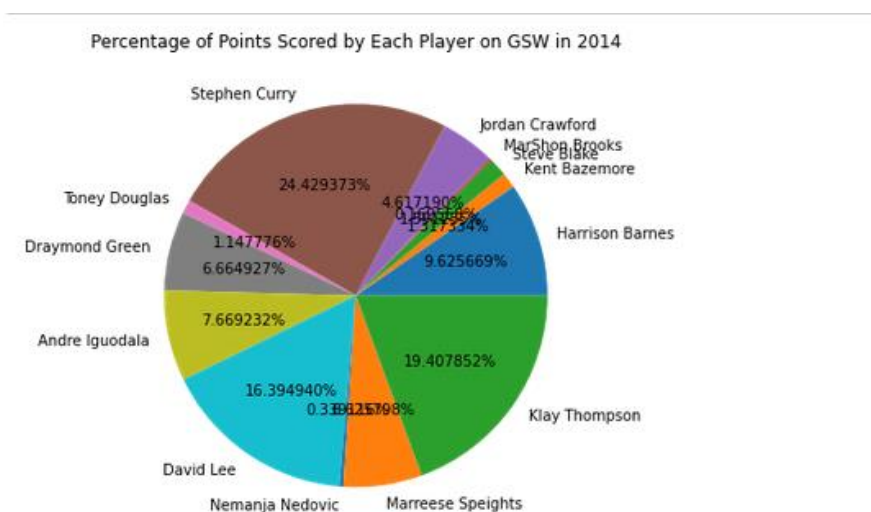
Year Season	Player name	Pos Position	Age Age	Tm Team	G Games	TS% True Shooting %	ORB% Offensive Rebound Percentage	DRB% Defensive Rebound Percentage
	3922 unique values	PF SG Other (14914)	20% 19% 60%	TOT NYK Other (21525)	9% 4% 87%		[null] 0 Other (19934)	16% 3% 81%
2017	Alex Abrines	SG	23	OKC	68	0.56	1.9	7.1
2017	Quincy Acy	PF	26	TOT	38	0.565	3.9	18
2017	Quincy Acy	PF	26	DAL	6	0.355	4.6	15.2
2017	Quincy Acy	PF	26	BRK	32	0.587	3.8	18.2
2017	Steven Adams	C	23	OKC	88	0.589	13	15.5
2017	Arron Afflalo	SG	31	SAC	61	0.559	0.7	8.4
2017	Alexis Ajinca	C	28	NOP	39	0.529	8.3	23.8
2017	Cole Aldrich	C	28	MIN	62	0.549	11	23.9
2017	LaMarcus Aldridge	PF	31	SAS	72	0.532	8.6	16.6
2017	Lavoy Allen	PF	27	IND	61	0.485	13.7	14.6
2017	Tony Allen	SG	35	MEM	71	0.493	9.6	13.8
2017	Al-Farouq Aminu	SF	26	POR	61	0.586	4.8	23.5

Ход работы

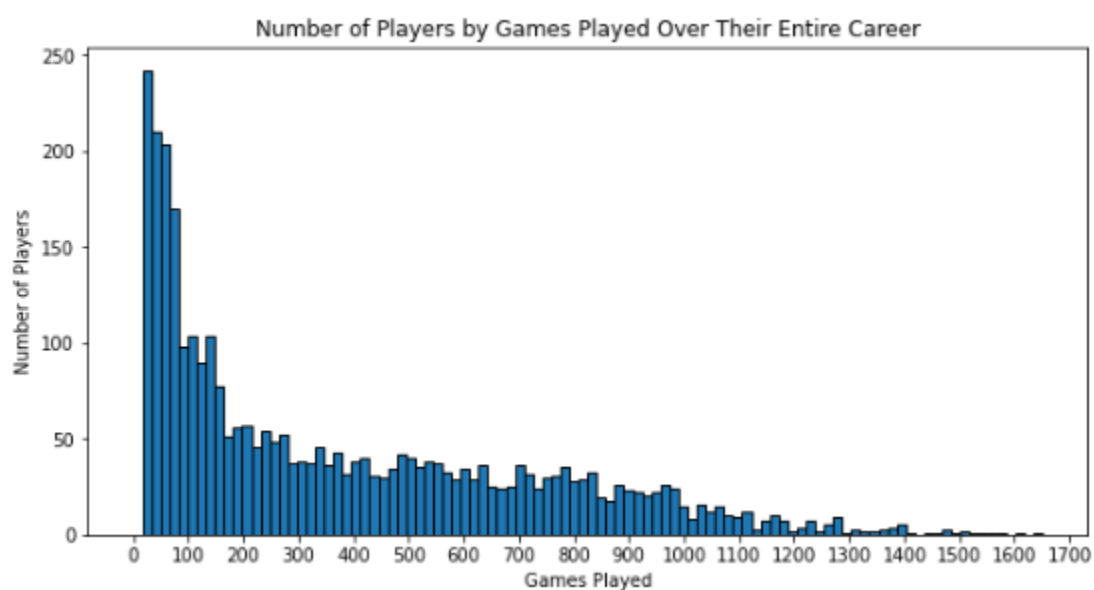
В ходе работы мы использовали следующие инструменты для выполнения поставленных задач:

- Python.
- PySpark для более эффективной обработки больших объемов данных.
- Виртуальная среда Jupyter Notebook.
- Matplotlib для визуализации данных.

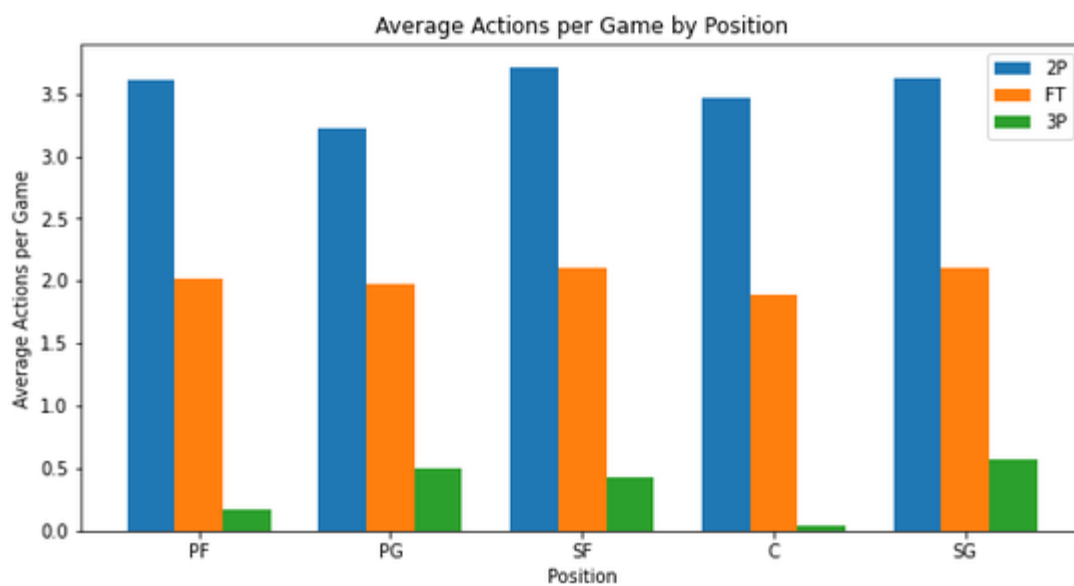
Визуализация процента очков каждого игрока в команде ГСВ в 2014 году



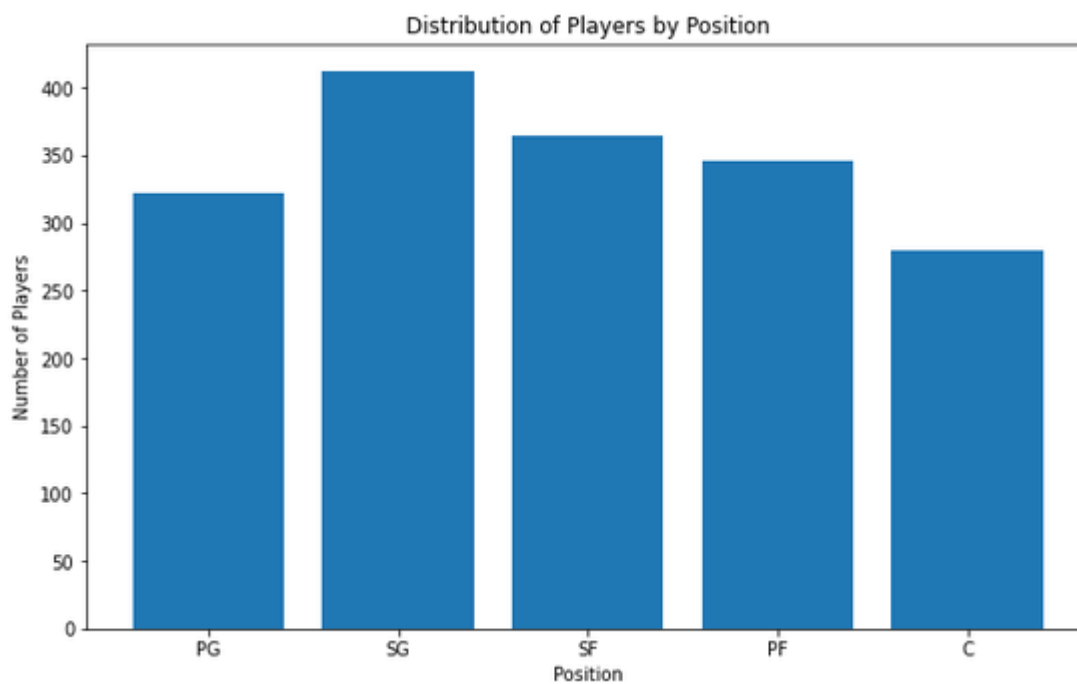
Количество игроков по сумме сыгранных матчей за всю карьеру



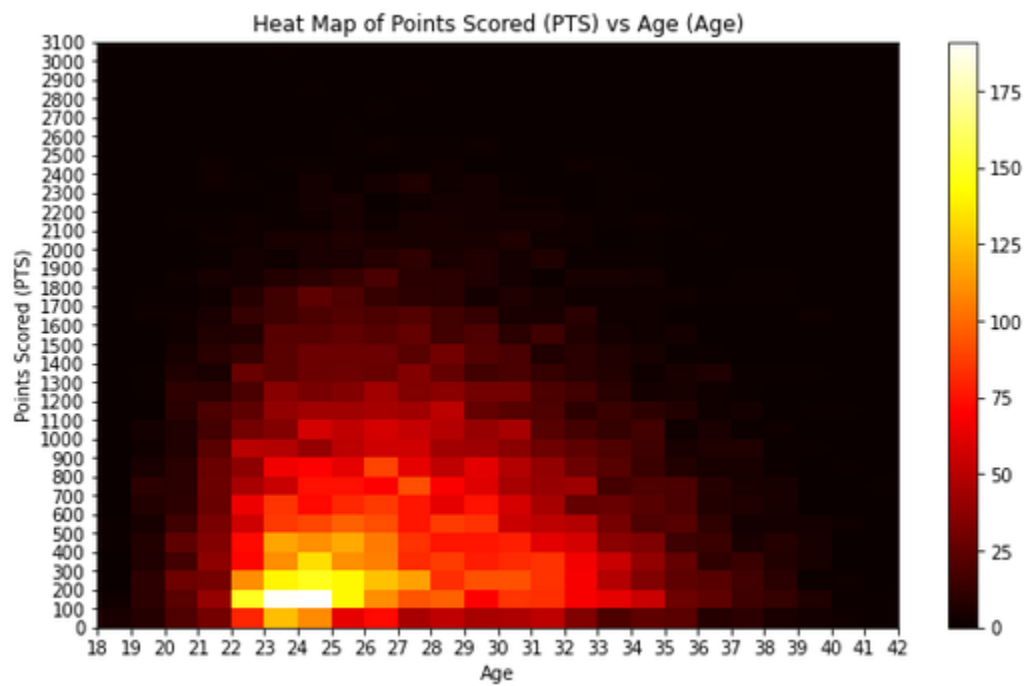
Среднее количество действий за игру в разбивке по позициям



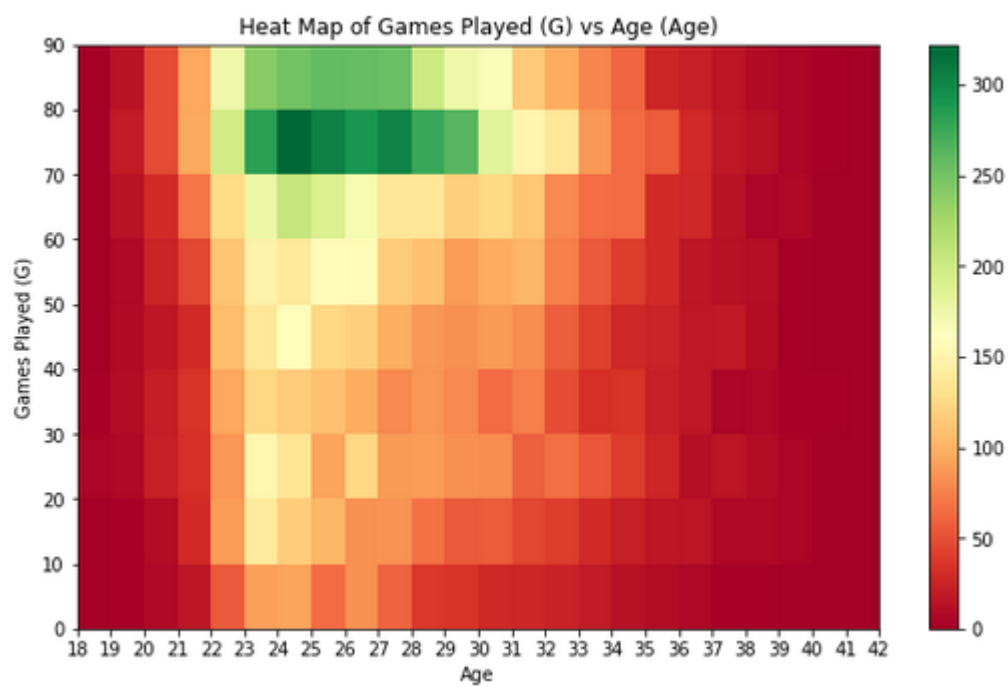
Распределение игроков по позициям



Тепловая карта заброшенных очков в зависимости от возраста



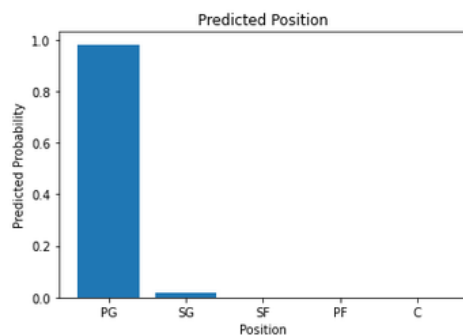
Тепловая карта сыгранных матчей в зависимости от возраста



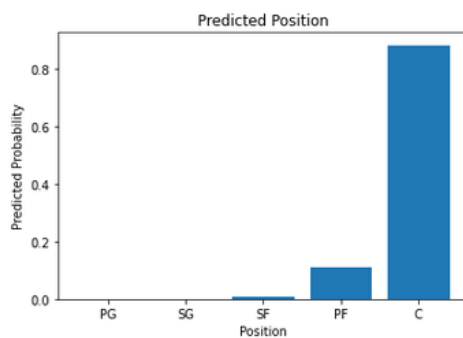
Логистическая регрессия

Предполагаемая позиция игрока в зависимости от его статистики

```
+-----+-----+-----+-----+-----+
|Year|Player      |Pos|PosNumber|probability
+-----+-----+-----+-----+
|2002|Rafer Alston|PG |0        |[0.9824821526892772,0.0168751609041145,6.326006985393863E-4,9.576301200522727E-6,
5.094068684280418E-7]|
+-----+-----+-----+-----+
only showing top 1 row
```



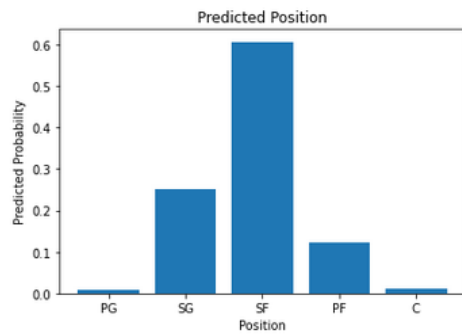
```
+-----+-----+-----+-----+-----+
|Year|Player      |Pos|PosNumber|probability
+-----+-----+-----+-----+
|2016|Chris Andersen|C  |4        |[6.95528587073927E-8,2.4313171282365043E-5,0.006112446713176814,0.10954720020663
605,0.884315970356046]|
+-----+-----+-----+-----+
only showing top 1 row
```




```

+-----+-----+-----+-----+
|Year|Player      |Pos|PosNumber|probability|
+-----+-----+-----+-----+
|2011|Trevor Ariza|SF |2       |[0.00885815374987844,0.25184920424302826,0.6072706634573916,0.12199246817459859,0.010029510375103263]|
+-----+-----+-----+-----+
only showing top 1 row

```



Линейная регрессия

Предполагаемая позиция игрока в зависимости от его статистики

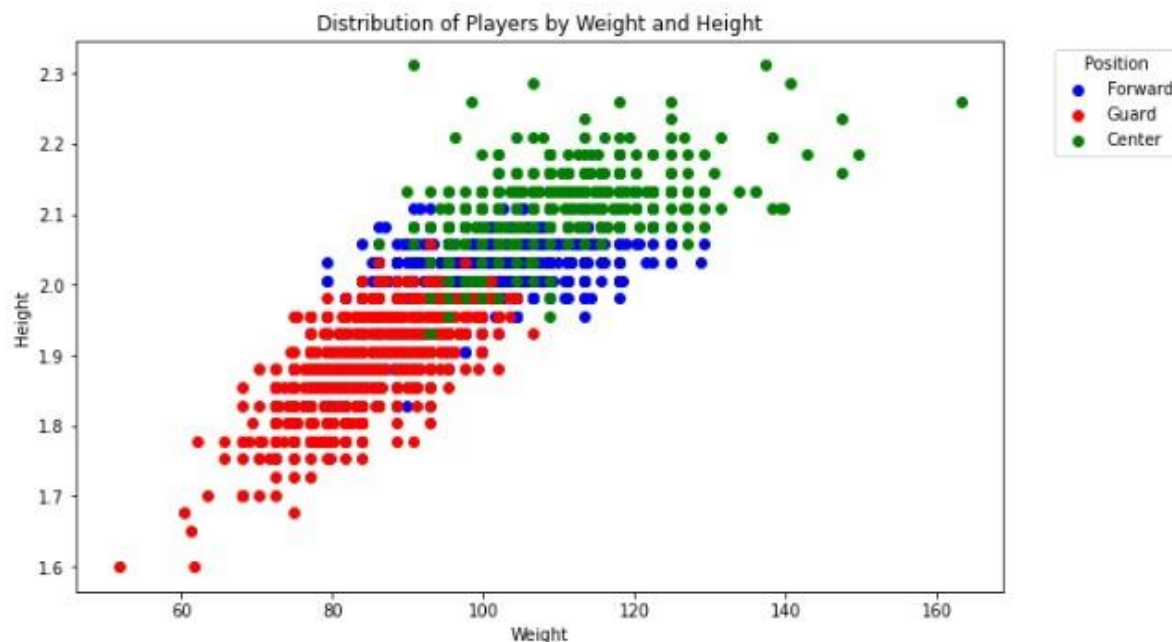
```

+-----+-----+-----+-----+-----+
|Year|          Player|Pos|PosNumber|prediction|
+-----+-----+-----+-----+-----+
|2001|Shandon Anderson|SG|1|1.681528437692074|
|2001|David Benoit|SF|2|2.4586120448796343|
|2001|Chauncey Billups|SG|1|0.5602411964041893|
|2001|Corie Blount|PF|3|3.3390710210588592|
|2001|Shawn Bradley|C|4|4.8174797753288825|
|2001|Elton Brand|PF|3|3.110663087566863|
|2001|Randy Brown|PG|0|0.7761110208762552|
|2001|Kobe Bryant|SG|1|1.280774365343175|
|2001|Jud Buechler|SF|2|1.6460158554977722|
|2001|Vince Carter|SF|2|1.5806965837715572|
+-----+-----+-----+-----+-----+
only showing top 10 rows

```

Root Mean Squared Error (RMSE): 0.6361117764817057

Распределение позиций игроков по весу и росту. Для решения данной задачи бралась статистика по росту и весу с другой таблицы: Player_Data.csv



Вывод

В ходе выполнения работы были собраны и предобработаны данные о статистике игроков NBA, проведен анализ и построены модели линейной и логистической регрессий для прогнозирования игровых позиций.

Гипотеза о том, что игровые статистики могут быть использованы для точного прогнозирования позиций баскетболистов, была подтверждена. Модели хоть и показали высокую точность предсказаний, однако сложно предсказать конкретную позицию между разными нападающими (легкий и тяжелый) или защитниками (разыгрывающий и атакующий).