

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ  
РАБОТА  
ПО КУРСУ  
«АНАЛИЗ ДАННЫХ НА ЯЗЫКЕ PYTHON»  
ТЕМА:**

**ПРОГНОЗИРОВАНИЕ ИСХОДА КАРТЫ В  
КИБЕРСПОРТИВНОЙ ИГРЕ COUNTER-STRIKE: GLOBAL  
OFFENSIVE НА ОСНОВЕ ИСТОРИЧЕСКИХ ДАННЫХ О  
ПРЕДЫДУЩИХ ВСТРЕЧАХ КОМАНД**

Выполнила: Воровкина Анна Владимировна

Руководитель: Мещеряков Александр

# Актуальность

Month	Avg. Players	Gain	% Gain	Peak Players
Last 30 Days	1,006,064.0	+87,978.4	+9.58%	1,802,853
April 2023	918,085.6	+74,611.1	+8.85%	1,510,231
March 2023	843,474.5	+38,368.6	+4.77%	1,519,457
February 2023	805,105.9	+78,955.6	+10.87%	1,354,248
January 2023	726,150.3	+96,825.3	+15.39%	1,199,684
December 2022	629,325.0	+8,319.0	+1.34%	1,065,079
November 2022	621,006.0	+12,656.5	+2.08%	1,129,095
October 2022	608,349.5	-10,426.5	-1.69%	1,078,860
September 2022	618,776.0	-22,668.7	-3.53%	1,100,366
August 2022	641,444.8	+46,991.7	+7.91%	1,039,889
July 2022	594,453.1	+22,230.6	+3.88%	928,329
June 2022	572,222.5	+7,260.8	+1.29%	906,670
May 2022	564,961.7	-4,021.5	-0.71%	923,996
April 2022	568,983.2	-12,506.5	-2.15%	1,013,237
March 2022	581,489.7	-53,148.7	-8.37%	987,993
February 2022	634,638.4	+32,262.1	+5.36%	995,163
January 2022	602,376.3	+55,762.1	+10.20%	991,625
December 2021	546,614.2	-1,547.5	-0.28%	950,586
November 2021	548,161.7	+35,725.8	+6.97%	935,593



# Актуальность



Внешними пользователями данного исследования смогут являться:

– киберспортивная организация, тренеры команд, сами игроки: менеджеры команд, а также тренеры смогут иметь возможность прогнозирования результатов команды для корректировки стиля игры, состава команды. В свою очередь игроки, а в частности капитан команды, также будут обладать возможностью анализа собственных выступлений для выполнения спонсорских обязательств и обязательств по трудовому договору;

– потенциальные спонсоры и инвесторы: имея необходимый багаж данных при помощи данной модели компании смогут выбрать самый ценный объект для максимизации своих денежных средств;

- зрители, болельщики – смогут оценивать шансы своих любимых команд на победу в том или ином матче, принимать решения по участию в тотализаторе.



# Введение в основы, правила игры

- Две команды по пять человек попеременно играют за стороны террористов и контр-террористов. Задача первых установить бомбу в специально обозначенной локации. Задача вторых не дать им это сделать в установленное время или успеть обезвредить бомбу.
- Матч/Карта состоит максимум из 30 раундов, и делится на две половины. После 15 раунда команды меняются сторонами. Команда играющая за атаку переходит на сторону защиты и наоборот.
- Игра продолжается до тех пор пока одна из сторон не наберёт 16 побед.
- В случае если обе команды набирают по 15 раундов, играют 6 дополнительных раундов.
- Всего в соревновательном CS существует 7 возможных вариантов карт.

# Цель и задачи:

Цель: Предсказать исход карты между двумя командами в CS:GO

## Задачи:

- 1) Парсинг данных с открытых источников и создание датасета
- 2) Предварительная обработка данных, разведывательный анализ, оценка базовых статистик
- 3) Отбор признаков для построения модели
- 4) Разработка модели прогнозирования результатов игры на основе собранных данных.
- 5) Оценка качества модели и ее эффективность.
- 6) Проверка тестирования модели на новых данных.

# Парсинг данных и создание датасета

- Использованные библиотеки: BeautifulSoup, requests
- Источники данных: [escorenews.com](http://escorenews.com)



# Изучение и описание датасета

- Датасет состоит из 1822 строк и 19 столбцов, пропущенных значений нет.
- 5 столбцов с типом object, 1 столбец содержащий даты, а также 12 столбцов с числовыми значениями и 1 результативный (целевой столбец)

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1822 entries, 0 to 1821
Data columns (total 19 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Date                  1822 non-null  datetime64[ns]
1   Team1                 1822 non-null  object
2   team1_region          1822 non-null  object
3   rank_team1            1822 non-null  int64
4   Team2                 1822 non-null  object
5   team2_region          1822 non-null  object
6   rank_team2            1822 non-null  int64
7   map                   1822 non-null  object
8   map_winrate_t1        1822 non-null  float64
9   map_winrate_t2        1822 non-null  float64
10  score_t1              1822 non-null  int64
11  score_t2              1822 non-null  int64
12  start_ct              1822 non-null  int64
13  round_ct_t1           1822 non-null  int64
14  round_ct_t2           1822 non-null  int64
15  round_t_1             1822 non-null  int64
16  round_t_2             1822 non-null  int64
17  overtime              1822 non-null  int64
18  winner                1822 non-null  int64
dtypes: datetime64[ns](1), float64(2), int64(11), object(5)
ge: 270.6+ KB
```

	Date	Team1	team1_region	rank_team1	Team2	team2_region	rank_team2	map	map_winrate_t1	map_winrate_t2	score_t1	score_t2	start_ct	round_ct_t1	round_ct_t2	round_t_1	round_t_2	overtime	winner
0	2023-04-30 06:10:00	Bad News Eagles	EU	20	Movistar Riders	EU	25	de_anubis	28.6	75.0	10	16	1	7	8	8	3	0	
1	2023-04-30 02:55:00	Movistar Riders	NA	25	ECSTATIC	EU	43	de_nuke	42.6	48.5	16	11	1	10	6	5	6	0	
2	2023-04-29 23:00:00	Grayhound	Asia&oceania	28	Bad News Eagles	EU	20	de_mirage	42.9	66.7	1	16	1	1	2	14	0	0	
3	2023-04-29 23:00:00	Grayhound	Asia&oceania	28	Bad News Eagles	EU	20	de_ancient	78.6	59.0	16	19	1	11	11	4	4	1	
4	2023-04-29 23:00:00	Grayhound	Asia&oceania	28	Bad News Eagles	EU	20	de_inferno	67.7	60.0	16	6	0	6	7	9	0	0	

# Предварительная обработка данных

Добавление новых столбцов на основании имеющихся:

- месяц, день недели и время суток игры
- Код карты - название карт, закодированное в числовые значения

Преобразование столбцов Регион из типа object в числовой категориальный при помощи LabelEncoder

Команды сыгравшие менее 5 карт за данный период времени были удалены.

Таким образом датафрейм стал состоять из 1625 строк, 23 столбцов. Тип данных object – остался только у столбцов с названиями команд и карты, остальные столбцы числовые.

```
Index: 1625 entries, 0 to 1930
Data columns (total 23 columns):
#   Column                Non-Null Count  Dtype
---  ---                -
0   Date                  1625 non-null  datetime64[ns]
1   Team1                 1625 non-null  object
2   team1_region          1625 non-null  int32
3   rank_team1            1625 non-null  int64
4   Team2                 1625 non-null  object
5   team2_region          1625 non-null  int32
6   rank_team2            1625 non-null  int64
7   map                   1625 non-null  object
8   map_winrate_t1        1625 non-null  float64
9   map_winrate_t2        1625 non-null  float64
10  score_t1               1625 non-null  int64
11  score_t2               1625 non-null  int64
12  start_ct               1625 non-null  int64
13  round_ct_t1           1625 non-null  int64
14  round_t_1             1625 non-null  int64
15  round_ct_t2           1625 non-null  int64
16  round_t_2             1625 non-null  int64
17  overtime               1625 non-null  int64
18  winner                 1625 non-null  int64
19  weekday                1625 non-null  int32
20  map_code               1625 non-null  int8
21  month                  1625 non-null  int32
22  day_period             1625 non-null  int32
dtypes: datetime64[ns](1), float64(2), int32(5), int64(11), int8(1), object(3)
```

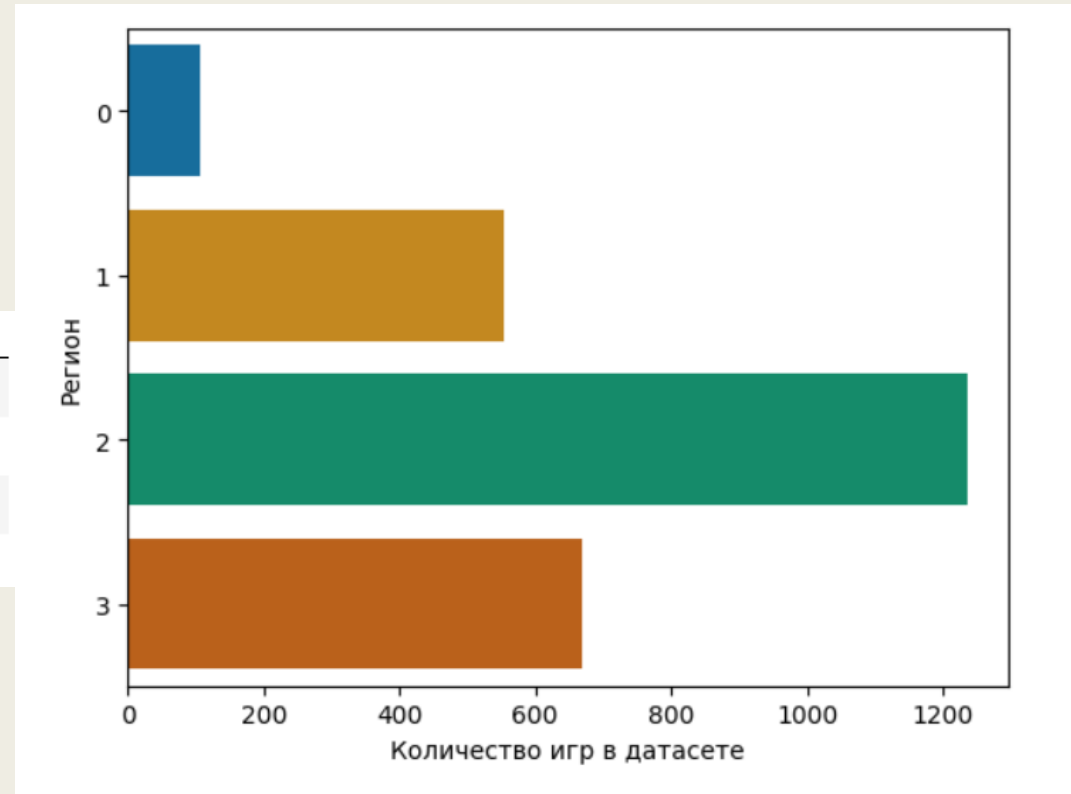
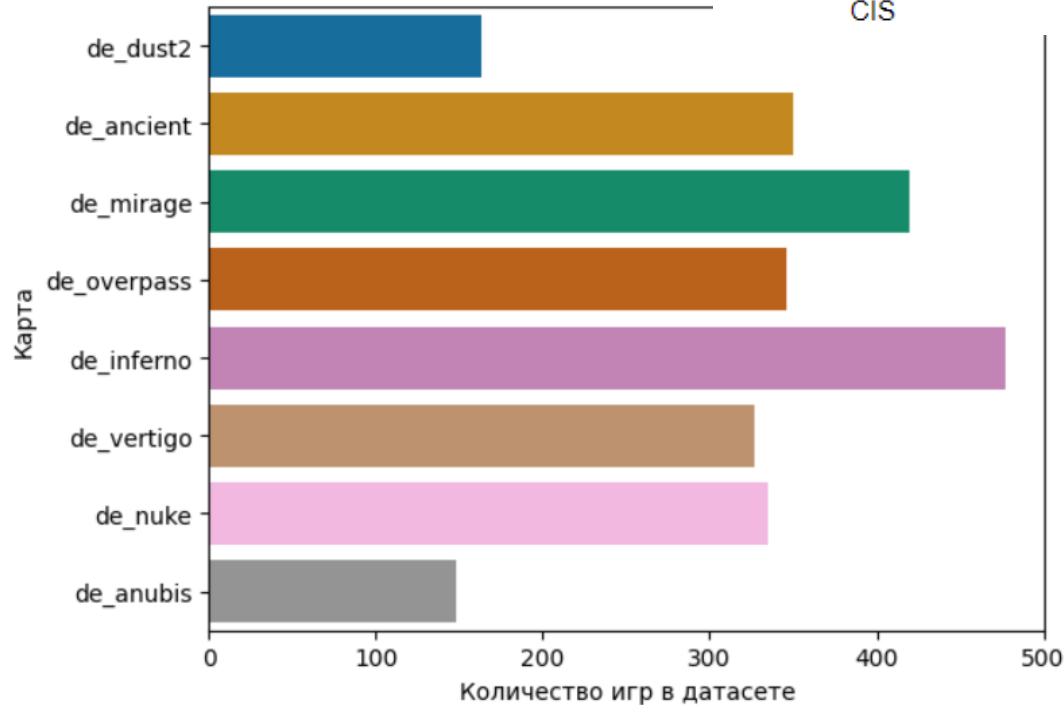
# Разведывательный анализ данных

Всего в датасете 113 уникальных команд.  
Каждая команда относится к одному из 4х регионов.

Преобладает регион EU.

Asia&Oceania заметно отстает.

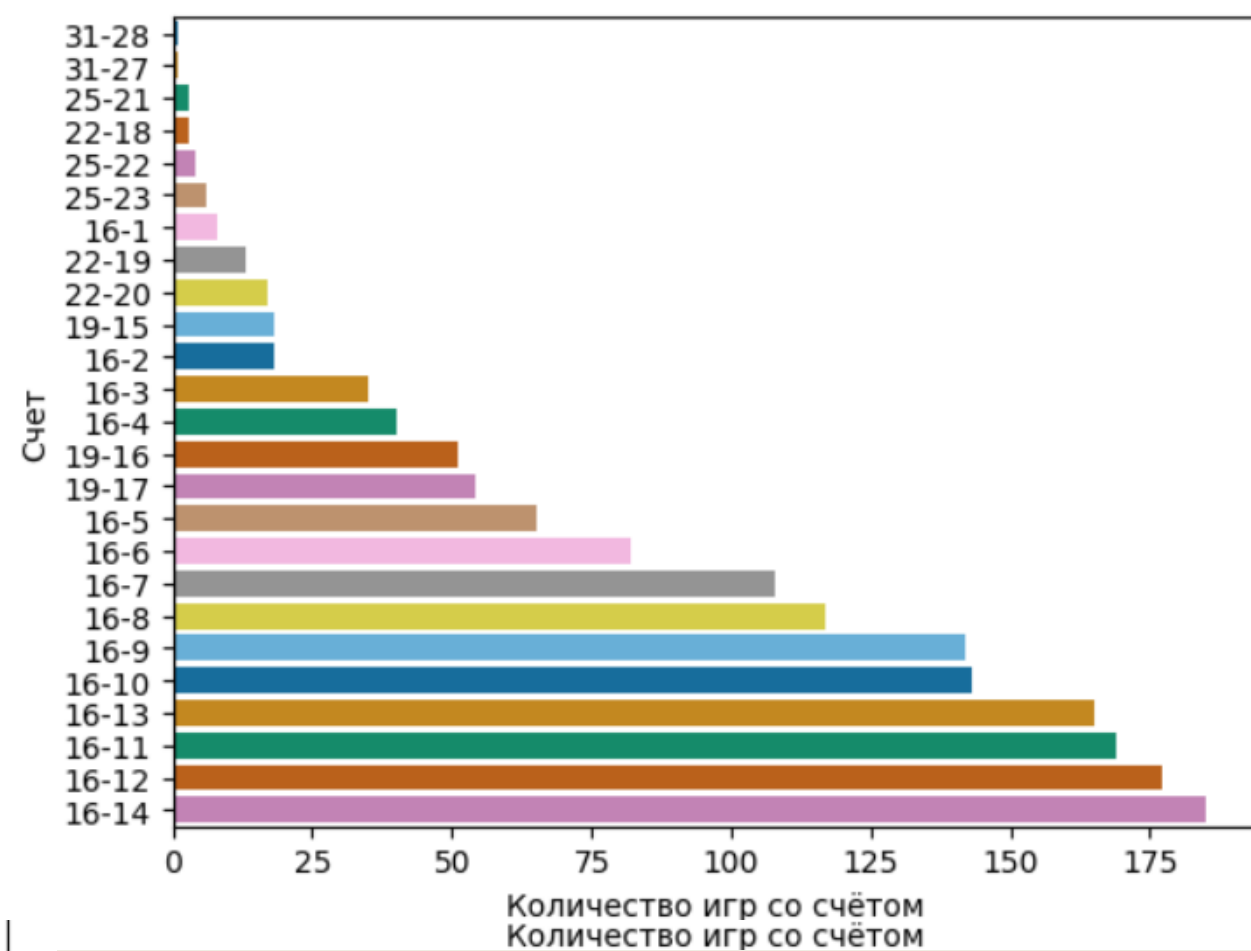
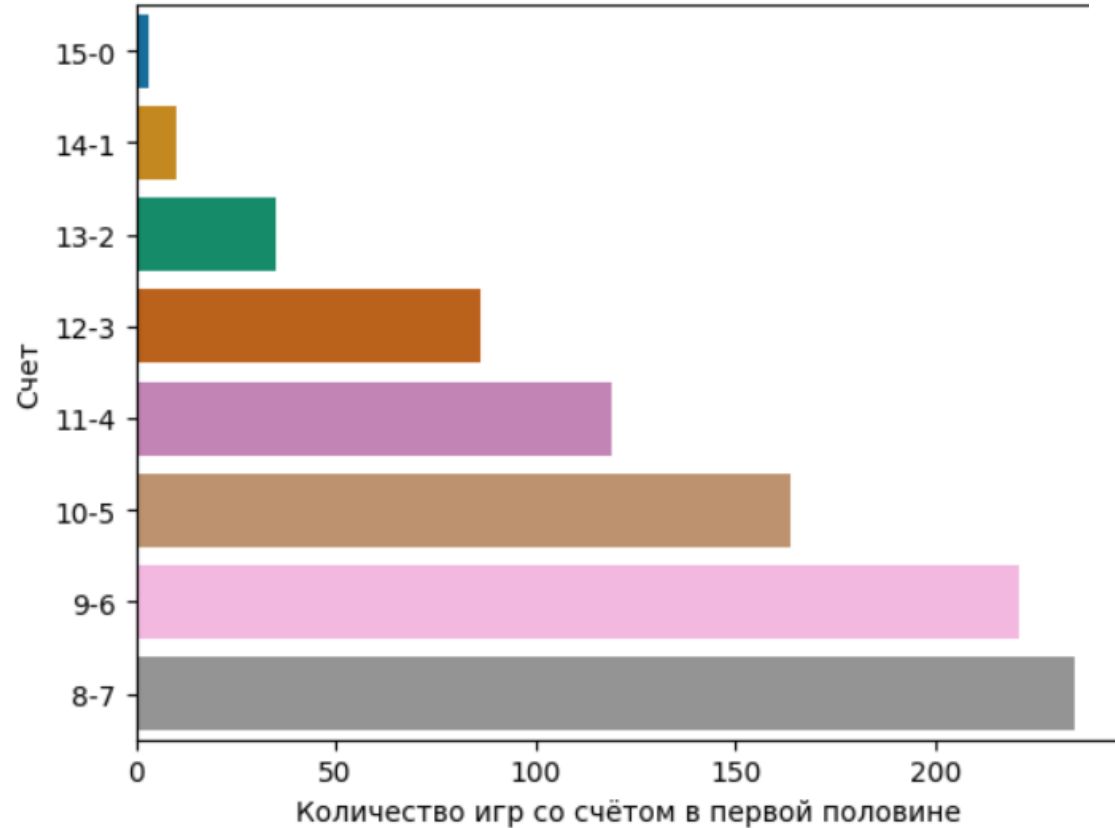
team1_region	team1_region_encoded
Asia&oceania	0
EU	2
NA	3
CIS	1



В соревновательный маппул входит 7 карт. В 2022 году произошла смена карты de\_dust2 на карту de\_Anubis, поэтому количество сыгранных игр на этих картах меньше, чем на остальных. Самой популярной картой является de\_inferno

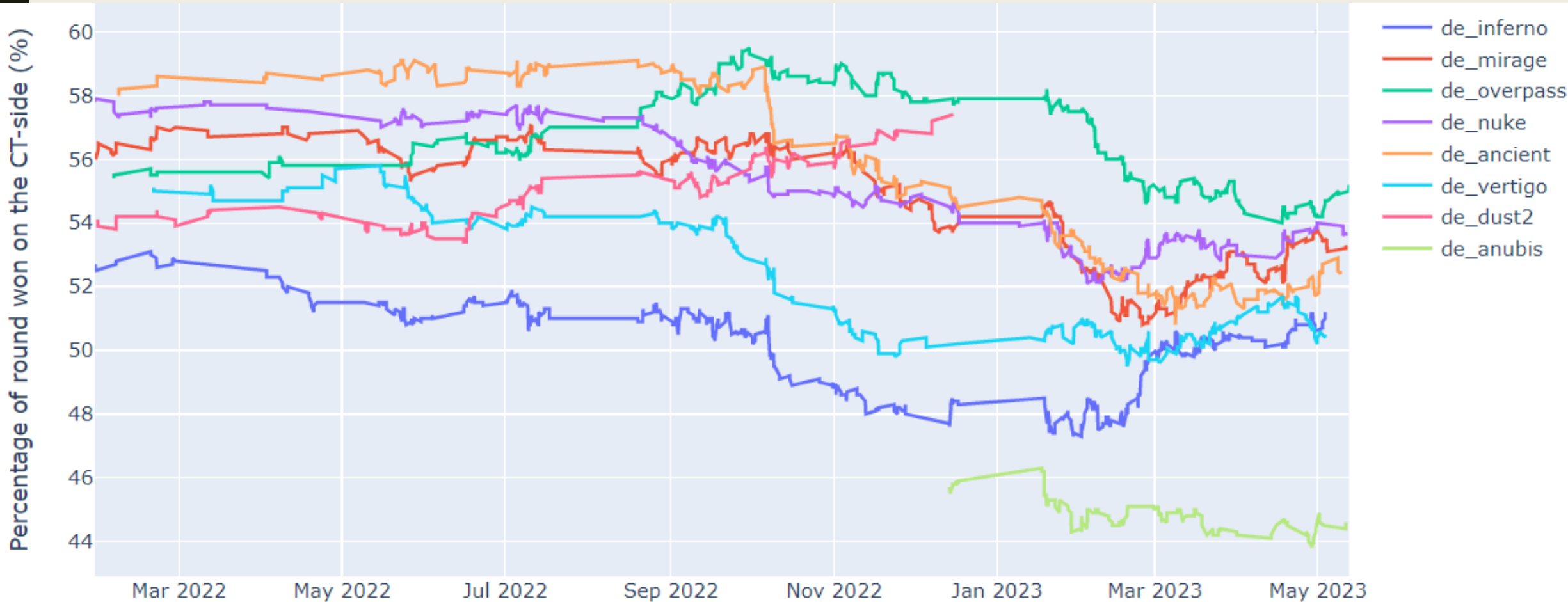
# Разведывательный анализ данных

Самый часто встречаемый счет на конец карты 16:14 (11,9%) и 16:12(10,3%)



Самый часто встречаемый счет на конец первой половины 8-7(26,9%) и 9-6(25,3%)

# Процент раундов выигранных за сторону контр-террористов по картам



# Baseline, построение первой модели

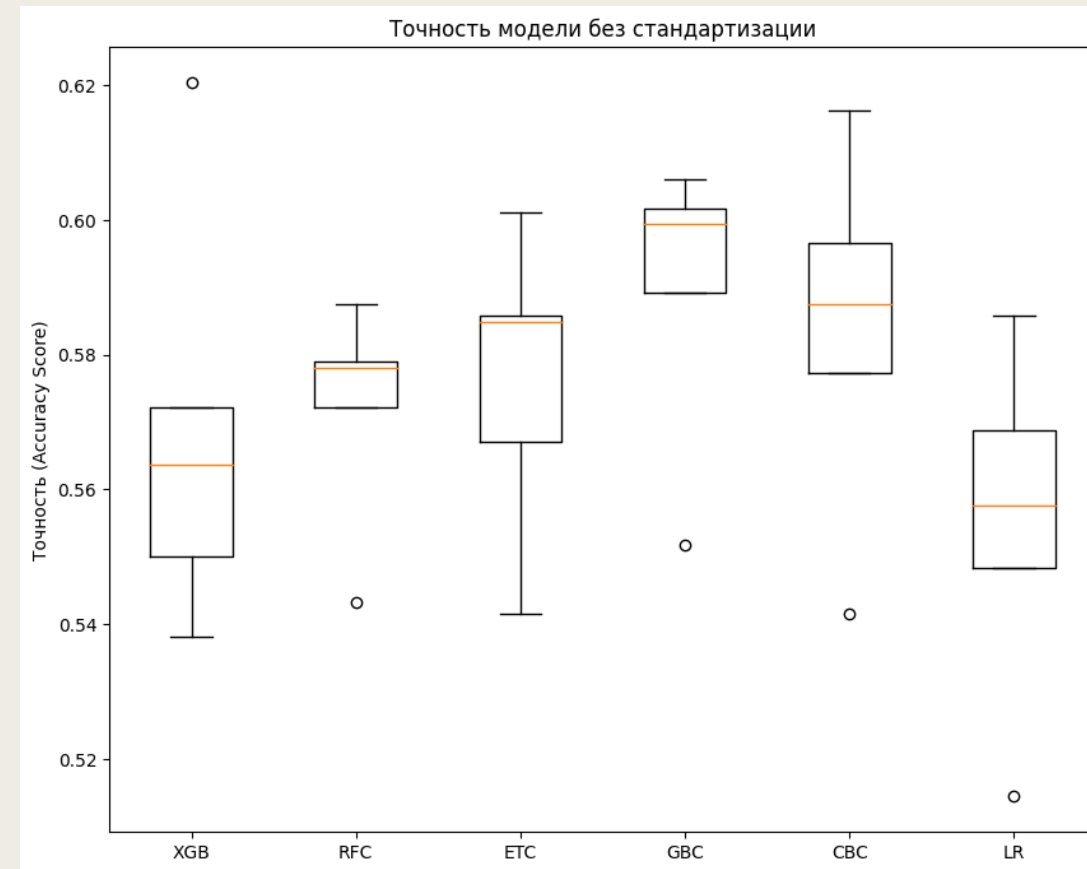
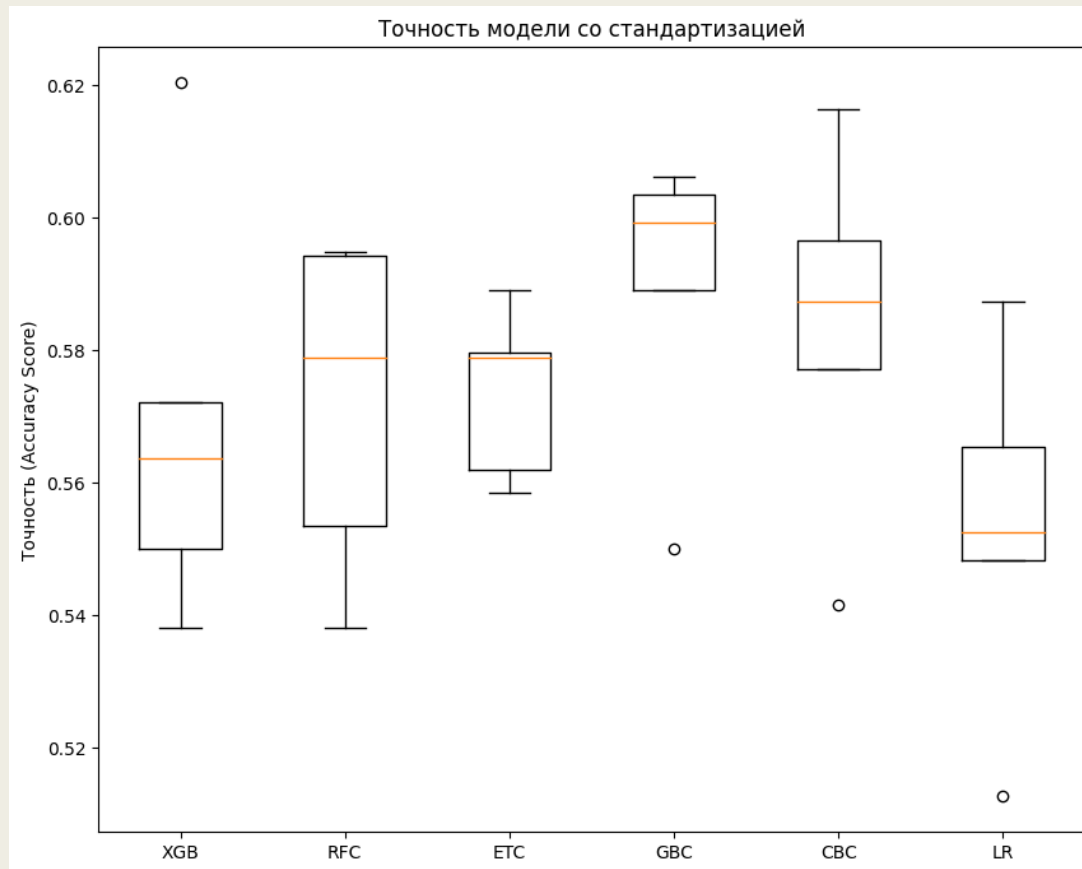
В качестве первой базовой модели был выбран классический метод LogisticRegression – точность составила 50,5%.

Для увеличения точности были добавлены новые расчётные столбцы на основании существующих.

Для увеличения количества данных, был создан новый датасет где каждая строка зеркально повторяет другую. То есть команда 1 стала командой 2 и наоборот.

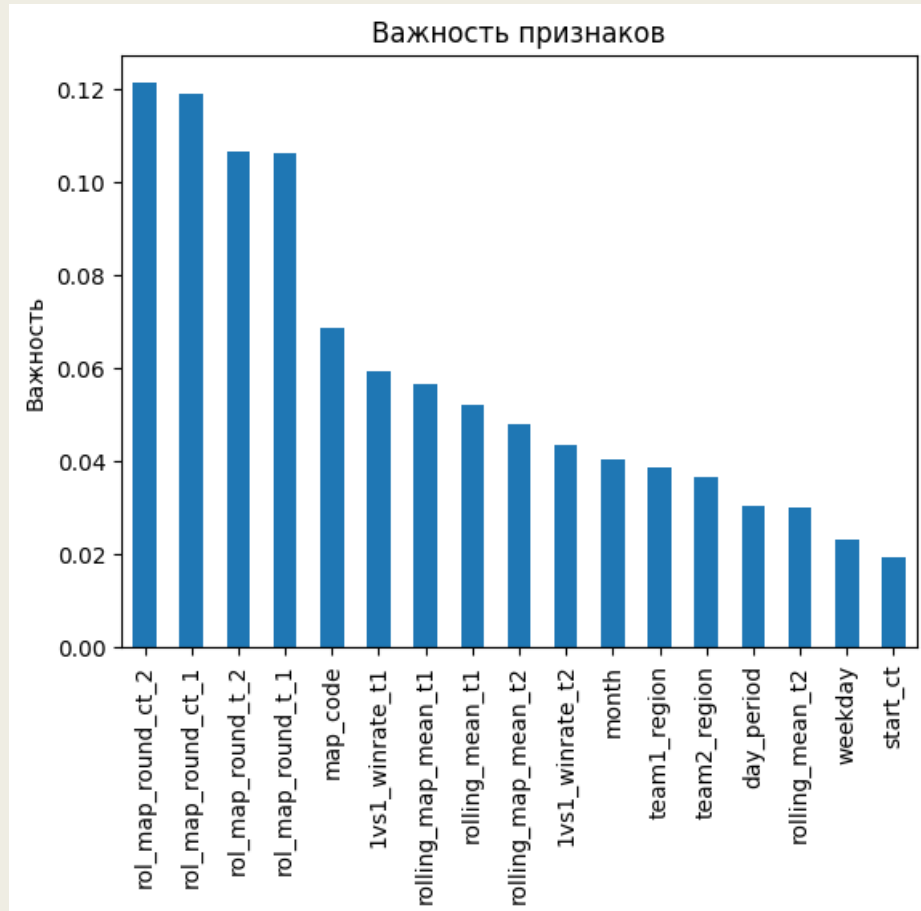
- rolling\_mean\_t1, rolling\_mean\_t2 – процент побед каждой команды за последние 5 матчей от текущего.
- rolling\_map\_mean\_t1, rolling\_map\_mean\_t2 - процент побед каждой команды на текущей карте за последние 5 матчей от текущего.
- rol\_map\_round\_ct\_1, rol\_map\_round\_ct\_2 - среднее количество выигранных раундов за сторону контр-террористов каждой команды на текущей карте за последние 5 матчей от текущего.
- rol\_map\_round\_t\_1, rol\_map\_round\_t\_2 - среднее количество выигранных раундов за сторону террористов каждой команды на текущей карте за последние 5 матчей от текущего.
- 1vs1\_winrate\_t1, 1vs1\_winrate\_t2 - процент побед команды над конкретным противником последние 5 матчей от текущего.

# Выбор моделей классификаторов

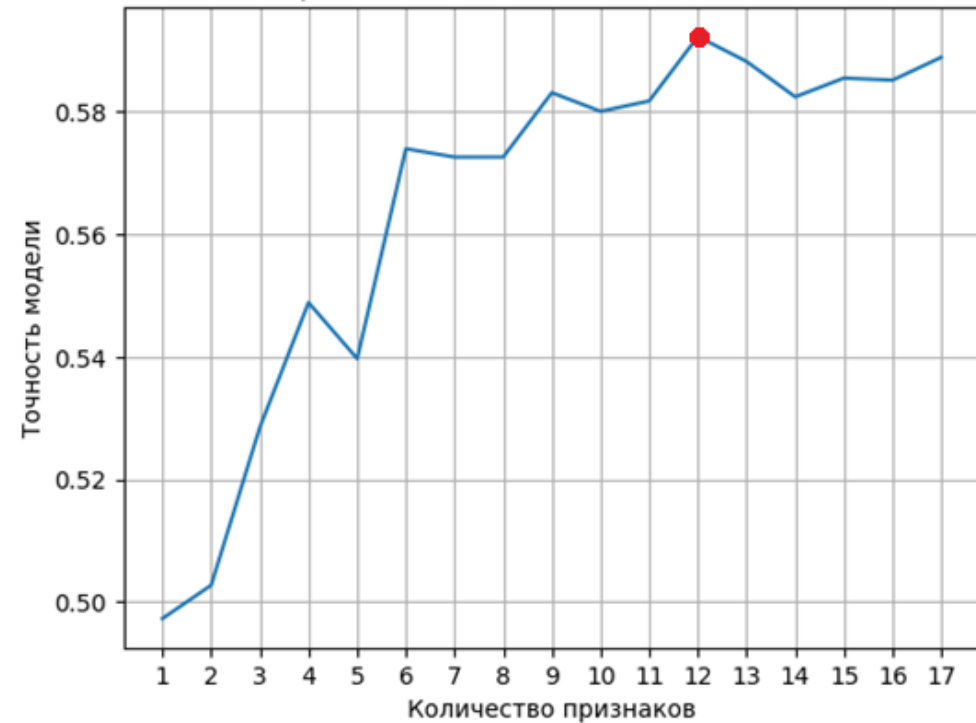


Лучшую точность (accuracy score) показывает модель «GradientBoostingClassifier» - метод классификации, который строит предсказание в виде ансамбля слабых предсказывающих моделей, которыми в основном являются деревья решений. Общая идея алгоритма – последовательное применение предиктора (предсказателя) таким образом, что каждая последующая модель сводит ошибку предыдущей к минимуму.

# Отбор признаков



Влияние количества признаков на точность модели GradientBoostingClassifier



Наибольшее влияние на результат прогнозирования оказывает среднее количество выигранных раундов за сторону контр-террористов, наименьшее - начало карты за сторону контр-террористов.

Лучшая точность достигается при использовании 12 признаков, поэтому остальные можно не использовать.

При использовании всех признаков точность модели незначительно снижается.

# Оценка качества модели и подбор гиперпараметров

Модель:

GradientBoostingClassifier, со стандартными параметрами

Размер тестовой выборки: 20%

Точность (accuracy score) 59,7%

Confusion matrix (Матрица ошибок):

```
[[173 122]
 [116 179]]
Accuracy: 0.597
```

Classification report (Отчет о классификации):

	precision	recall	f1-score	support
0	0.60	0.59	0.59	295
1	0.59	0.61	0.60	295
accuracy			0.60	590
macro avg	0.60	0.60	0.60	590
weighted avg	0.60	0.60	0.60	590

- Для подбора и настройки гиперпараметров были использованы библиотеки GridSearchCV и Optuna
- Лучшими гиперпараметрами стали: learning\_rate=0.01, max\_depth=5, n\_estimators=500, min\_samples\_leaf=3, min\_samples\_split=0.1
- С помощью гиперпараметров удалось улучшить модель до точности 62,2%

Confusion matrix (Матрица ошибок):

```
[[176 119]
 [104 191]]
Accuracy: 0.622
```

Classification report (Отчет о классификации):

	precision	recall	f1-score	support
0	0.63	0.60	0.61	295
1	0.62	0.65	0.63	295
accuracy			0.62	590
macro avg	0.62	0.62	0.62	590
weighted avg	0.62	0.62	0.62	590

# Прогнозирование исхода новых игр

- Данные по 6 новым играм, проходящим на следующий игровой день, не входящие в обучающие или тестовые данные, были переданы в модель для получения предсказания.
- Модель корректно определила победителя на 6 из 10 играх.

	Team1	Team2	winner	pred_winner
0	Natus Vincere	GamerLegion	1	0
1	9INE	Liquid	1	1
2	fnatic	Ninjas in Pyjamas	0	0
3	Heroic	FaZe	0	0
4	Into the Breach\it	Apeks	0	1
5	Vitality	G2	0	0
6	Natus Vincere	Ninjas in Pyjamas	1	1
7	Monte	fnatic	0	1
8	Apeks	forZe	1	0
9	G2	Liquid	0	0

# Заключение и выводы

В данной работе были выполнены следующие задачи:

- 1) Был собран датасет с результатами прошедших матчей по CS:GO за последние 2 года с помощью парсинга данных с сайта [escorenews.com](http://escorenews.com)
- 2) На основании существующих данных были добавлены новые признаки, рассчитанные по результатам последних данных.
- 3) Была выбрана модель GradientBoostingClassifier, с начальной точностью в 59,7%
- 4) Были выбраны 12 лучших признаков, имеющих наибольший вес, остальные признаки были отброшены.
- 5) С помощью настройки гиперпараметров модели, удалось улучшить точность до 62,2%. Данная точность на 12,2% выше, чем простое предположение (вероятность выигрыша команды 50%-50%)

Конечно, данная точность достаточно низка, возможно это связано с недостатком данных или признаков. Исход карты зависит от очень многих факторов, которые не удалось получить с сайта [escorenews.com](http://escorenews.com)