Прогнозирование исхода “команда забьет в матче” для матчей РПЛ

Кирщин Иван, БЭАД223

Данные

Данные

- Результаты матчей РПЛ с сезона 2020/21 по сезон 2024/25 включительно.
- Коэффициенты на линию “Индивидуальный тотал больше” для каждого из матчей.
- Всего 1188 матчей * 2 = 2376 наблюдений.

Обработка данных

- Так как ставки на линию “индивидуальный тотал больше” для одной команды в конкретном матче открываются лишь на одно значение тотала, вместо столбца с коэффициентами имеем разреженную матрицу коэффициентов с ровно одним не NaN значением в каждой строке.
- Для тоталов больше 2, 2.5, 3, 3.5, 4 присутствуют буквально единичные наблюдения. Для матчей, на которые открывался один из этих тоталов, и матчей, для которых коэффициенты отсутствуют, введем дамми-переменную на то, что отсутствуют коэффициенты.
- Таким образом, для каждого матча информация о коэффициентах представляет собой 4 столбца: ИТБ 0.5, ИТБ 1, ИТБ 1.5, дамми на то, что коэффициенты отсутствуют. Пропущенные значения закодируем нулями.

Обработка данных

- Разобьем данные на тренировочную и тестовую выборки. Тренировочная выборка: матчи РПЛ с начала сезона 2020/21 до первой половины сезона 2024/25 включительно. Тестовая выборка: матчи второй половины сезона 2024/25.
- В тренировочной выборке содержится 2206 наблюдений, в тестовой — 170.
- На тренировочной выборке будем подбирать признаковое описание, гиперпараметры и оценивать саму модель, а на тестовой выборке — тестировать ставочную стратегию.

Обработка данных

Тренировочная выборка:

	Date	Team	Opponent	scored	missed	Venue	Result	Season	MD	Result Margin	target	total_over_0_5	total_over_1	total_over_1_5	no_coefs
0	2020-08-08	FC Khimki	CSKA Moscow	0	2	H	L	2020/21	1	-2	0	1.65	0.0	0.00	0
1	2020-08-08	CSKA Moscow	FC Khimki	2	0	A	W	2020/21	1	2	1	0.00	0.0	2.05	0
2	2020-08-08	Tambov	Rostov	0	1	H	L	2020/21	1	-1	0	1.46	0.0	0.00	0
3	2020-08-08	Rostov	Tambov	1	0	A	W	2020/21	1	1	1	0.00	0.0	2.16	0
4	2020-09-08	Ufa	FK Krasnodar	0	3	H	L	2020/21	1	-3	0	0.00	0.0	0.00	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2201	2024-08-12	CSKA Moscow	Fakel Voronezh	1	0	A	W	2024/25	18	1	1	0.00	0.0	0.00	1
2202	2024-08-12	Akhmat Grozny	Orenburg	1	0	H	W	2024/25	18	1	1	0.00	0.0	0.00	1
2203	2024-08-12	Orenburg	Akhmat Grozny	0	1	A	L	2024/25	18	-1	0	0.00	0.0	0.00	1
2204	2024-08-12	FK Krasnodar	Lokomotiv Moscow	0	0	H	D	2024/25	18	0	0	0.00	0.0	0.00	1
2205	2024-08-12	Lokomotiv Moscow	FK Krasnodar	0	0	A	D	2024/25	18	0	0	0.00	0.0	0.00	1

2206 rows × 15 columns

Обработка данных

Тестовая выборка:

	Date	Team	Opponent	scored	missed	Venue	Result	Season	MD	Result Margin	target	total_over_0_5	total_over_1	total_over_1_5	no_coefs
2206	2025-02-28	Dynamo Makhachkala	Lokomotiv Moscow	1	1	H	D	2024/25	19	0	1	0.0	2.23	0.00	0
2207	2025-02-28	Lokomotiv Moscow	Dynamo Makhachkala	1	1	A	D	2024/25	19	0	1	0.0	1.64	0.00	0
2208	2025-01-03	Nizhny Novgorod	Akron	2	1	H	W	2024/25	19	1	1	0.0	0.00	0.00	1
2209	2025-01-03	Akron	Nizhny Novgorod	1	2	A	L	2024/25	19	-1	1	0.0	0.00	0.00	1
2210	2025-01-03	Akhmat Grozny	Rubin Kazan	2	1	H	W	2024/25	19	1	1	0.0	0.00	0.00	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2371	2025-05-18	Spartak Moscow	Krylia Sovetov	2	0	A	W	2024/25	29	2	1	0.0	0.00	2.17	0
2372	2025-05-18	Akhmat Grozny	Dynamo Makhachkala	1	1	H	D	2024/25	29	0	1	0.0	0.00	2.05	0
2373	2025-05-18	Dynamo Makhachkala	Akhmat Grozny	1	1	A	D	2024/25	29	0	1	0.0	1.90	0.00	0
2374	2025-05-19	Lokomotiv Moscow	CSKA Moscow	2	2	H	D	2024/25	29	0	1	0.0	1.68	0.00	0
2375	2025-05-19	CSKA Moscow	Lokomotiv Moscow	2	2	A	D	2024/25	29	0	1	0.0	0.00	2.14	0

170 rows × 15 columns

Feature engineering

Признаки

1) Среднее количество забитых голов за последние n_avg матчей командой.

2) Дамми-переменная на то, что команда является топ-клубом (CSKA Moscow, Lokomotiv Moscow, Krasnodar, Spartak Moscow, Zenit St Petersburg, Dinamo Moscow).

3) Дамми-переменная на то, что команда играет дома.

4) Среднее количество пропущенных голов за последние n_avg матчей соперником.

5) Дамми-переменная на то, что соперник является топ-клубом.

6) "Транзитивные" победы: количество команд, которые команда обыграла, из числа тех, что обыгрывали соперника за последние k матчей.

7) Взвешенная разница позиций: (номер тура / 30) * (текущая позиция соперника - текущая позиция команды) + ((30 - номер тура) / 30) * (позиция соперника в прошлом сезоне - позиция команды в прошлом сезоне).

Признаки

8) Последний релевантный результат: количество забитых командой голов в последнем матче с одной из n ближайших по разнице позиций команд к сопернику.

9) Дамми-переменная на вторую половину сезона.

10) Коэффициент на индивидуальный тотал больше 0.5 для команды, 0, если отсутствует.

11) Коэффициент на индивидуальный тотал больше 1 для команды, 0, если отсутствует.

12) Коэффициент на индивидуальный тотал больше 1.5 для команды, 0, если отсутствует.

13) Дамми-переменная на отсутствие коэффициентов.

n_avg, k, n, n_margin (с какого тура в сезоне начинаем обучение, требуется для корректного подсчета признаков, смотрящих назад) являются гиперпараметрами и подбираются при помощи кросс-валидации

Кросс-валидация

Кросс-валидация будет устроена следующим образом:

- Разбиваем выборку на 8 фолдов с последовательными матчами (разбиение по времени) по 11 туров.
- На очередной итерации делаем oversampling тренировочной выборки ввиду дисбаланса классов (в 72.5% наблюдений рассматриваемая команда забивала в матче).
- Оцениваем модель на 7 фолдах, на оставшемся рассчитываем score = 0.5 * TP - FP для симуляции реальных ставок.
- Усредняем score по 8 итерациям.

Подбор гиперпараметров

- По итогам подбора упомянутых ранее гиперпараметров при помощи кросс-валидации наилучшим набором оказался следующий:

```
best score: 22.5
best params: {'n_avg': 3, 'n': 2, 'k': 6, 'n_margin': 3}
```

- Сохраняем 5 наборов гиперпараметров, дающих наилучшее качество на валидационной выборке.

Отбор признаков

- В имеющейся задаче критически важна обобщающая способность итоговой модели.
- Так как у нас имеется всего 2000 наблюдений, а признаков 13, хотелось бы сократить признаковое описание в связи с потенциальным переобучением.

Отбор признаков

- Будем итеративно исключать по одному из оставшихся признаков и замерять скор на кросс-валидации без этого признака, после чего будем выбрасывать признак, исключение которого дает наибольший прирост качества.
- Чтобы избавиться от зависимости от конкретного выбора гиперпараметров, все скоры усредняются по 5 лучшим наборам, оставленным ранее.
- По итогам проведенного отбора оказалось, что исключение дамми-переменной на то, что соперник является топ-клубом, и последнего релевантного результата увеличивает скор на кросс-валидации на 7.5%

Итоговая модель

Набор гиперпараметров: `{'n_avg': 2, 'n': 3, 'k': 12, 'n_margin': 3}`

Оценка модели:

Logit Regression Results

Dep. Variable:	target	No. Observations:	2964
Model:	Logit	Df Residuals:	2952
Method:	MLE	Df Model:	11
Date:	Sun, 08 Jun 2025	Pseudo R-squ.:	0.05412
Time:	17:58:30	Log-Likelihood:	-1943.3
converged:	True	LL-Null:	-2054.5
Covariance Type:	nonrobust	LLR p-value:	1.656e-41

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.4024	0.379	1.062	0.288	-0.340	1.145
avg_scored	0.1359	0.045	2.993	0.003	0.047	0.225
top_club	0.3979	0.097	4.114	0.000	0.208	0.587
home_team	0.4360	0.078	5.619	0.000	0.284	0.588
opponent_avg_missed	0.0241	0.044	0.546	0.585	-0.062	0.111
transitive_wins	0.0542	0.037	1.475	0.140	-0.018	0.126
positions_diff	0.0161	0.008	2.050	0.040	0.001	0.032
season_second_part	0.0339	0.080	0.421	0.674	-0.124	0.192
total_over_0_5	-0.7971	0.233	-3.419	0.001	-1.254	-0.340
total_over_1	-0.5656	0.179	-3.166	0.002	-0.916	-0.216
total_over_1_5	-0.2890	0.162	-1.788	0.074	-0.606	0.028
no_coefs	-1.0865	0.359	-3.027	0.002	-1.790	-0.383

Предельные эффекты

- При прочих равных увеличение среднего количества голов, забитых в последних 2 матчах командой, на 1 увеличивает вероятность того, что команда забьет в матче, на 3.5%.
- При прочих равных тот факт, что команда является топ-клубом, увеличивает вероятность того, что команда забьет в матче, на 9.2%.
- При прочих равных тот факт, что команда играет дома, увеличивает вероятность того, что команда забьет в матче, на 8.9%.
- При прочих равных увеличение взвешенной позиции команды в таблице на 1 по сравнению с соперником увеличивает вероятность того, что команда забьет в матче, на 0.4%.

Разработка и тестирование стратегии

Ставочная стратегия

- Будем ставить лишь на те матчи, на которые открывалась линия "Индивидуальный тотал больше 1". При ставке на данную линию выигрыш рассчитывается следующим образом: 0 рублей, если команда не забила в матче; деньги возвращаются, если команда забила ровно 1 гол в матче; деньги * коэффициент, если команда забила более 1 гола.
- Из этих матчей будем ставить по 100 рублей на те, для которых предсказанная моделью вероятность того, что команда забьет, больше 0.5, и вероятность * коэффициент > 0.9. На остальные матчи не будем ставить ничего.

Baseline

При помощи бутстрапа оценивался 95% ДИ для итоговых прибыли и ROI.

- **Голосование по доле объектов:** для каждого матча будем предсказывать вероятность положительного класса равной доле объектов положительного класса в обучающей выборке

```
Profit on test set: -303.0
95% CI for profit: (-1319.53, 605.28)

ROI on test set: -0.06
95% CI for ROI: (-0.28, 0.14)
```

Baseline

- **Ставки на основе букмекерских вероятностей:** будем предсказывать вероятность положительного класса равной 1 / коэффициент букмекера, то есть брать букмекерские “вероятности”.

```
Profit on test set: 340.0
95% CI for profit: (-242.5, 849.44)

ROI on test set: 0.14
95% CI for ROI: (-0.13, 0.34)
```

Логит-модель

Результаты:

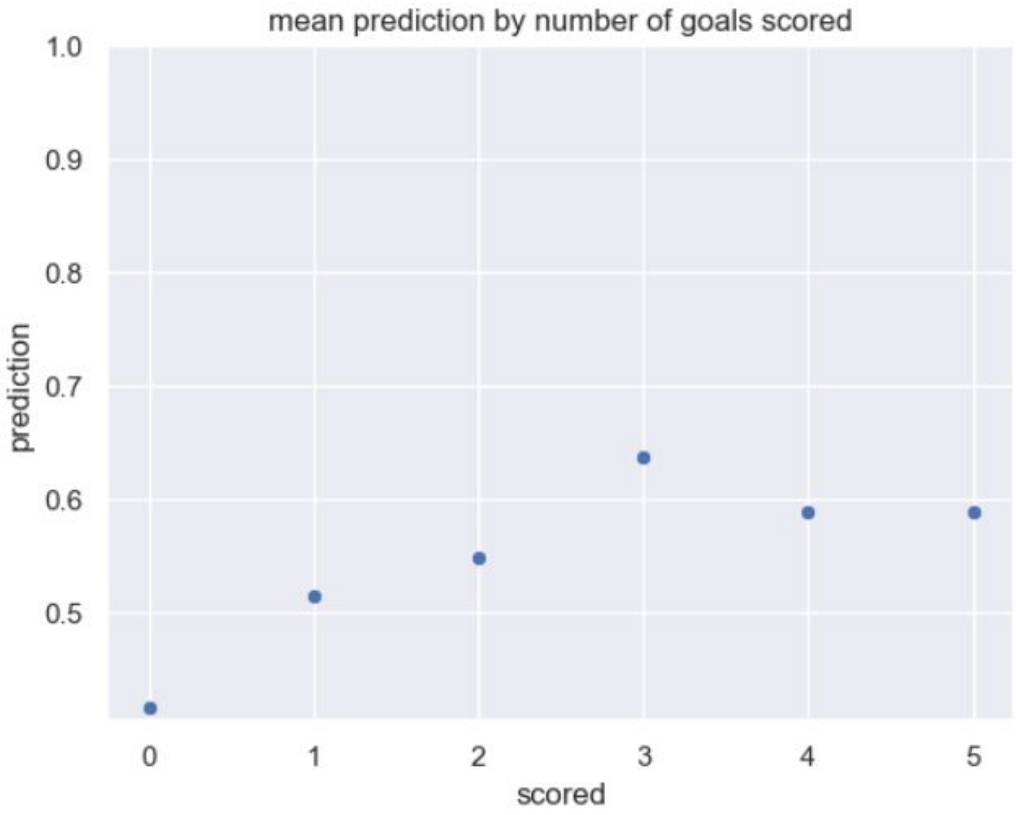
```
Profit on test set: 275.0
95% CI for profit: (67.46, 627.54)

ROI on test set: 0.25
95% CI for ROI: (0.08, 0.48)

ROC-AUC train: 0.6543733984603365
ROC-AUC test: 0.7553846153846153

AUC-PR train: 0.6438096921580939
AUC-PR test: 0.9086229740832751

TP test: 78
FP test: 6
TN test: 34
FN test: 52
```



Пуассоновская регрессия

Вероятность того, что команда забьет в матче:

$$\mathbb{P}(y > 0) = 1 - \mathbb{P}(y = 0)$$

Будем ставить на те матчи, для которых EV > 1, где

$$\mathbf{EV} = \mathbb{P}(y = 1) + \mathbf{coef} \cdot \mathbb{P}(y > 1)$$

т.е. матожидание выручки.

Пуассоновская регрессия

Результаты оценивания модели:

Poisson Regression Results

Dep. Variable:	scored	No. Observations:	2964
Model:	Poisson	Df Residuals:	2952
Method:	MLE	Df Model:	11
Date:	Sun, 08 Jun 2025	Pseudo R-squ.:	0.06078
Time:	18:04:53	Log-Likelihood:	-3913.2
converged:	True	LL-Null:	-4166.5
Covariance Type:	nonrobust	LLR p-value:	1.328e-101

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.0313	0.185	-0.169	0.866	-0.394	0.332
avg_scored	0.0487	0.021	2.316	0.021	0.007	0.090
top_club	0.2670	0.046	5.822	0.000	0.177	0.357
home_team	0.3429	0.039	8.812	0.000	0.267	0.419
opponent_avg_missed	0.0305	0.021	1.431	0.152	-0.011	0.072
transitive_wins	0.0418	0.016	2.599	0.009	0.010	0.073
positions_diff	0.0218	0.004	5.530	0.000	0.014	0.030
season_second_part	0.0404	0.039	1.026	0.305	-0.037	0.118
total_over_0_5	-0.5068	0.117	-4.327	0.000	-0.736	-0.277
total_over_1	-0.2912	0.093	-3.122	0.002	-0.474	-0.108
total_over_1_5	-0.1591	0.081	-1.970	0.049	-0.317	-0.001
no_coefs	-0.5012	0.171	-2.926	0.003	-0.837	-0.165

Пуассоновская регрессия

Результаты:

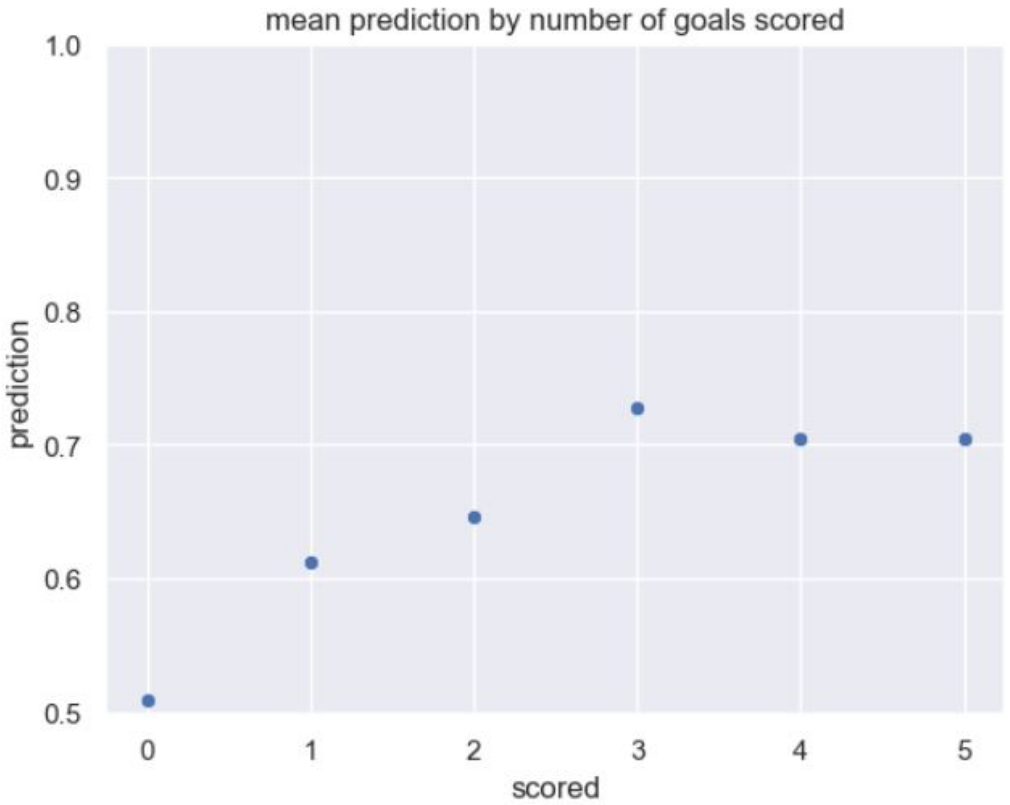
```
Profit on test set: 145.0
95% CI for profit: (0.0, 444.0)

ROI on test set: 0.48
95% CI for ROI: (0.0, 0.77)

ROC-AUC train: 0.6504737006015506
ROC-AUC test: 0.7586538461538461

AUC-PR train: 0.6433666281834598
AUC-PR test: 0.909221398667986

TP test: 110
FP test: 21
TN test: 19
FN test: 20
```



Потенциальные улучшения

- Увеличение количества данных за счет расширения временного диапазона либо увеличения числа чемпионатов
- Применение более сложных моделей (деревья, бустинг, GNN)
- Более качественные коэффициенты (заполнена вся матрица или имеется один столбец для всех матчей), так как полезность очень важной переменной размывается в текущем формате
- Добавление коэффициентов на другие исходы (победа, поражение, ничья, ОЗ и т.д.)
- Более сложные признаки (признаки на основе графовой структуры данных, статистика по ударам/владению/угловым/xG, моделирование силы футболистов)

Спасибо за внимание!