



Машинное обучение и
высоконагруженные системы

Галанов Максим
Ширяева Виктория

июнь, 2024

Программный проект: «Предсказательные модели прогнозирования результата матча Национальной Хоккейной Лиги»

Научный руководитель: Гаврилова Елизавета Владимировна



Содержание

1. Актуальность задачи
2. Обзор существующих решений
3. Цели и задачи работы
4. Сбор данных
5. Обработка и хранение данных
6. Архитектура DWH Data Vault 2.0
7. Yandex DataLens
8. Предобработка данных, EDA
9. Feature engineering
10. ML, Прогнозирование победителя
11. DL подход
12. Реализация сервиса
13. Telegram Bot





Актуальность задачи

Национальная Хоккейная Лига (NHL) – одна из самых популярных спортивных лиг в мире, привлекающая внимание миллионов болельщиков. Каждый сезон в NHL проводится множество матчей, что генерирует огромное количество данных о командах, играх и результатах игр. Анализ этих данных – непростая задача, требующая применения современных методов обработки данных и машинного обучения.

Любители NHL часто сталкиваются с трудностями при отслеживании и анализе статистики. Большинство существующих платформ предоставляют данные в виде статичных таблиц, что затрудняет их анализ и делает его менее удобным. Кроме того, возможности для прогнозирования исходов матчей на этих платформах ограничены в функциональности и не всегда удобны в использовании.



[Бот можно посмотреть тут :\)](#)



Обзор существующих решений

Statistics

Home Skaters Goalies **Teams** Glossary

By Season **By Game** **All-Time**

SEASONS SUM RESULTS GAME TYPE FRANCHISE REPORT

2023-24 2023-24 Regular Season All Franchises Summary

Export

Summary

Team	Season	GP	W	L	T	OT	P+	P%	RW	ROW	S/O	Win	GF	GA	GF/GP	GA/GP	PP%	PK%	Net PP%	Net PK%	Shots/GP	SA/GP	FOW%
1 New York Rangers	2023-24	82	55	23	--	4	114	.695	43	51	4	278	226	3.39	2.76	26.4	84.5	24.0	87.9	31.5	29.5	52.3	
2 Dallas Stars	2023-24	82	52	21	--	9	113	.689	40	48	4	294	232	3.59	2.83	24.2	82.0	22.5	87.0	31.6	28.7	54.0	
3 Carolina Hurricanes	2023-24	82	52	23	--	7	111	.677	43	50	2	277	211	3.38	2.57	26.9	86.4	23.7	90.7	33.3	25.6	52.6	
4 Winnipeg Jets	2023-24	82	52	24	--	6	110	.671	46	52	0	259	198	3.16	2.41	18.8	77.1	16.7	78.5	30.3	29.6	47.6	
5 Florida Panthers	2023-24	82	52	24	--	6	110	.671	42	49	3	265	198	3.23	2.41	23.5	82.5	20.2	85.2	33.7	27.8	51.3	
6 Vancouver Canucks	2023-24	82	50	23	--	9	109	.665	44	50	0	279	221	3.40	2.70	22.7	79.1	19.9	81.9	28.4	28.6	52.1	
7 Boston Bruins	2023-24	82	47	20	--	15	109	.665	36	43	4	263	221	3.21	2.70	22.2	82.5	19.8	85.1	29.3	30.5	49.5	
8 Colorado Avalanche	2023-24	82	50	25	--	7	107	.652	42	48	2	302	252	3.68	3.07	24.5	79.9	22.3	82.6	33.0	29.8	47.9	
9 Edmonton Oilers	2023-24	82	49	27	--	6	104	.634	39	47	2	292	236	3.56	2.88	26.3	79.5	24.3	82.2	33.8	28.1	53.2	
10 Toronto Maple Leafs	2023-24	82	46	26	--	10	102	.622	33	41	5	298	261	3.63	3.18	24.0	76.9	20.2	79.4	32.6	29.8	53.5	

nhl.com/stats/teams

flyquestas Банк 2850.0 Прибыль 0.00% Сего дня, 02:27

Хоккей. NHL. Плей-офф. 1/4 финала. До 4-х побед.

2024-05-11 05:00 **Ванкувер Кэнакс** **3 : 3** **Эдмонтон Ойлерз**

Прогноз **П2** Кф **2.28** Сумма **5.0% от банка** Проигрыш **-5% от банка**

Кристина Прохорова Банк 1916.0 Прибыль +4.31% Вчера, 20:03

Хоккей. NHL. Плей-офф. 1/4 финала. До 4-х побед.

2024-05-11 02:00 **Бостон Брюинз** **2 : 6** **Флорида Пантерз**

Прогноз **П1** Кф **2.85** Сумма **0.5% от банка** Проигрыш **-0,5% от банка**

vprognoze.ru/forecast/hockey/nhl

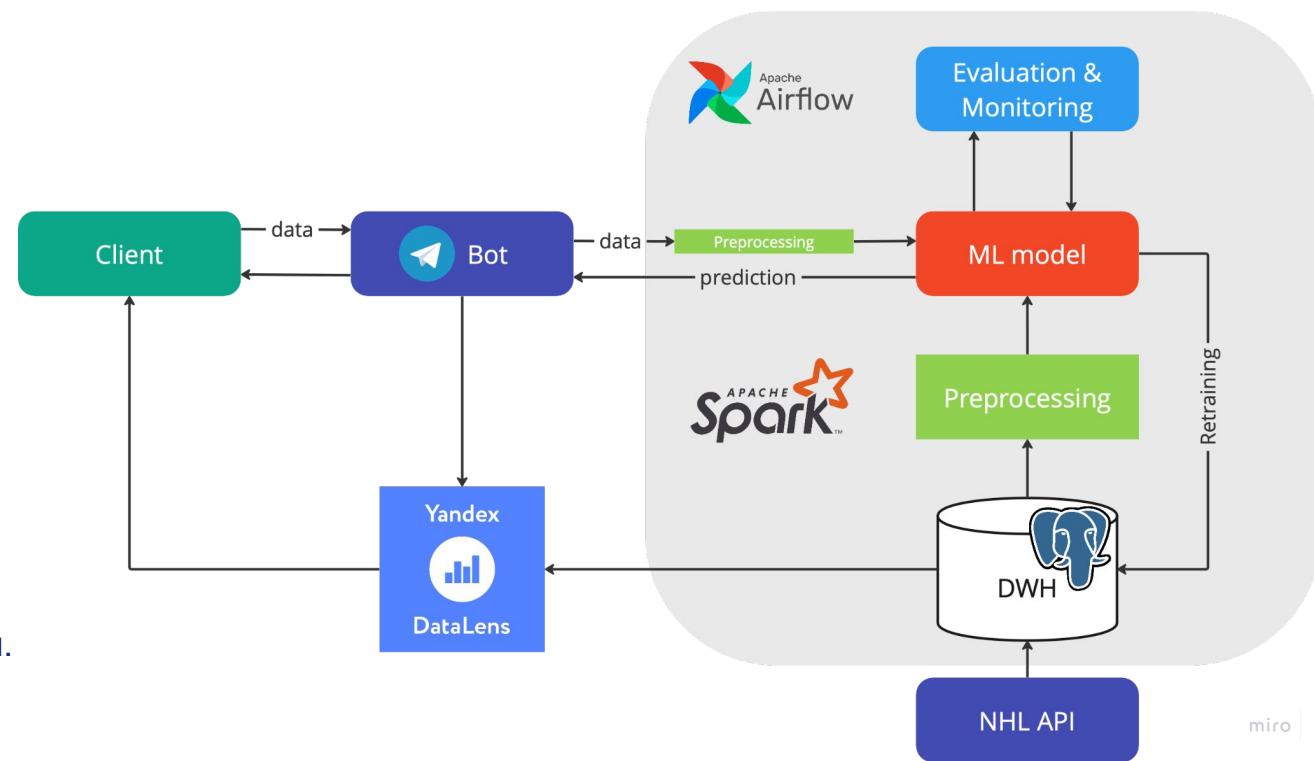


Цели и задачи

Разработать сервис, который предоставляет обширные статистические данные по играм, игрокам и командам, а также прогнозы результатов матчей NHL, интегрированные через Telegram Bot для удобного доступа.

Ключевые функции:

- Автоматический сбор данных: Используем NHL API для получения актуальной статистики.
- Хранение данных: Применение архитектуры Data Vault 2.0 для гибкого хранения данных.
- Аналитическая обработка: Обработка и анализ данных с помощью Apache Spark и Apache Airflow.
- Машинное обучение: Применение ML моделей для генерации прогнозов.
- Интерактивные дашборды: Визуализация данных через Yandex DataLens для наглядного представления статистики.
- Удобный доступ: Telegram Bot как основной интерфейс взаимодействия с пользователем.





Сбор данных

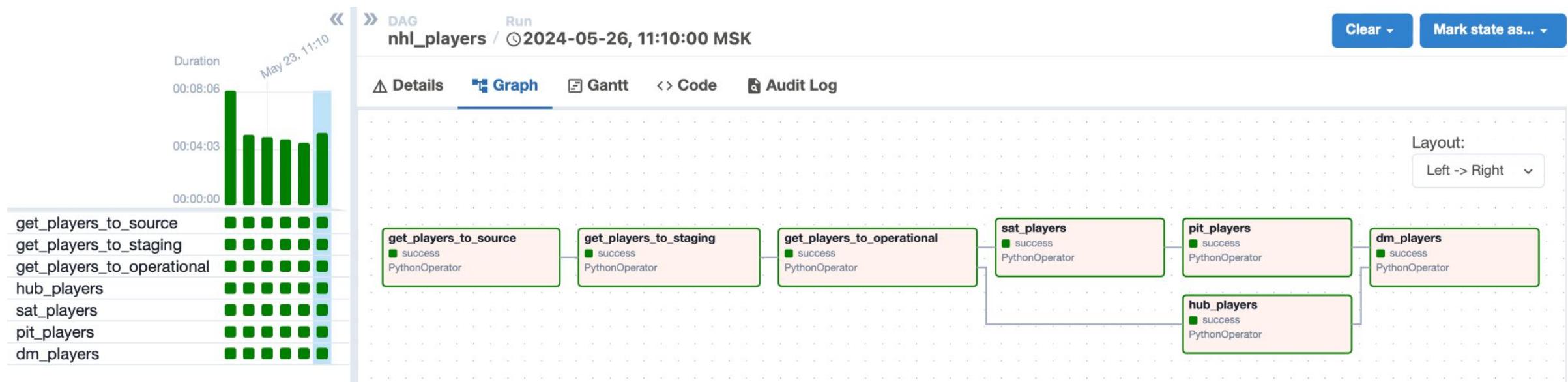


NHL API: Основной источник данных о матчах, командах и играх NHL.

- Предоставляет актуальные статистические данные и результаты игр.
- Разнообразие информации о играх, включая статистику производительности и исторические данные.

Процесс сбора данных:

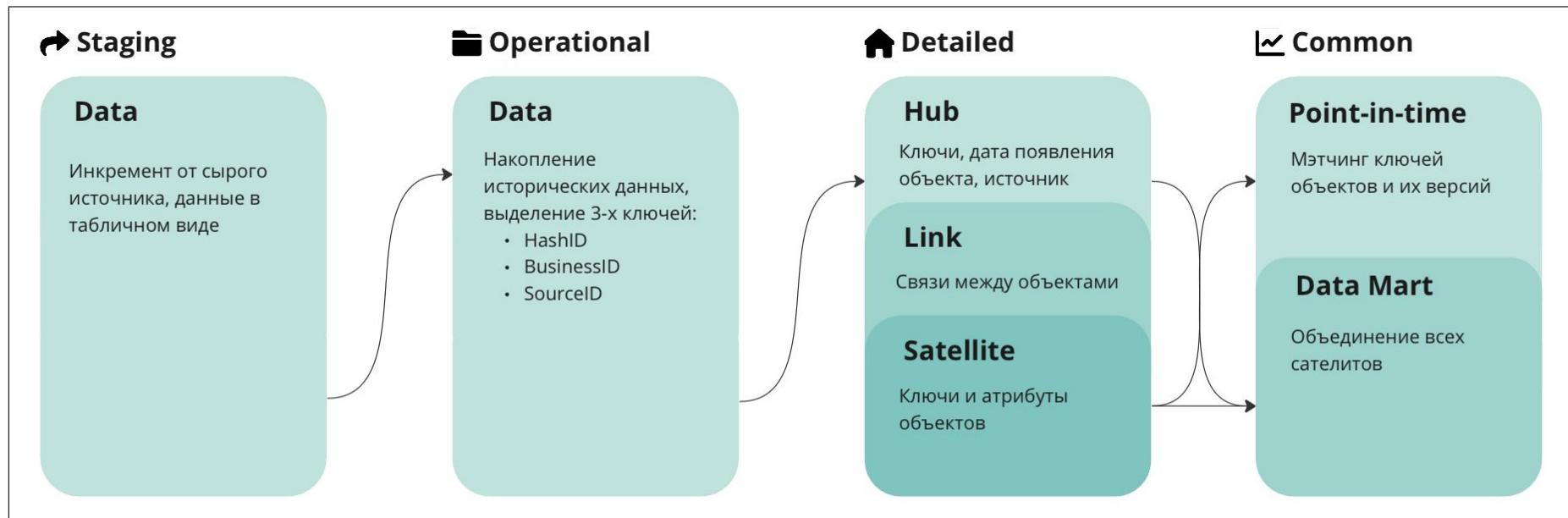
- Сбор данных осуществляется посредством регулярных и автоматизированных ELT процессов, которые оркестрируются с помощью Apache Airflow.





Обработка и хранение данных

Данные проходят предварительную обработку с помощью Apache Spark и загружаются ETL
процессами в DWH СУБД PostgreSQL согласно архитектуре Data Vault 2.0 для обеспечения гибкости и
масштабируемости.



Захват данных с
источника

Накопление
исторических
данных

Разложение
данных на
сущности

Формирование
аналитических
витрин

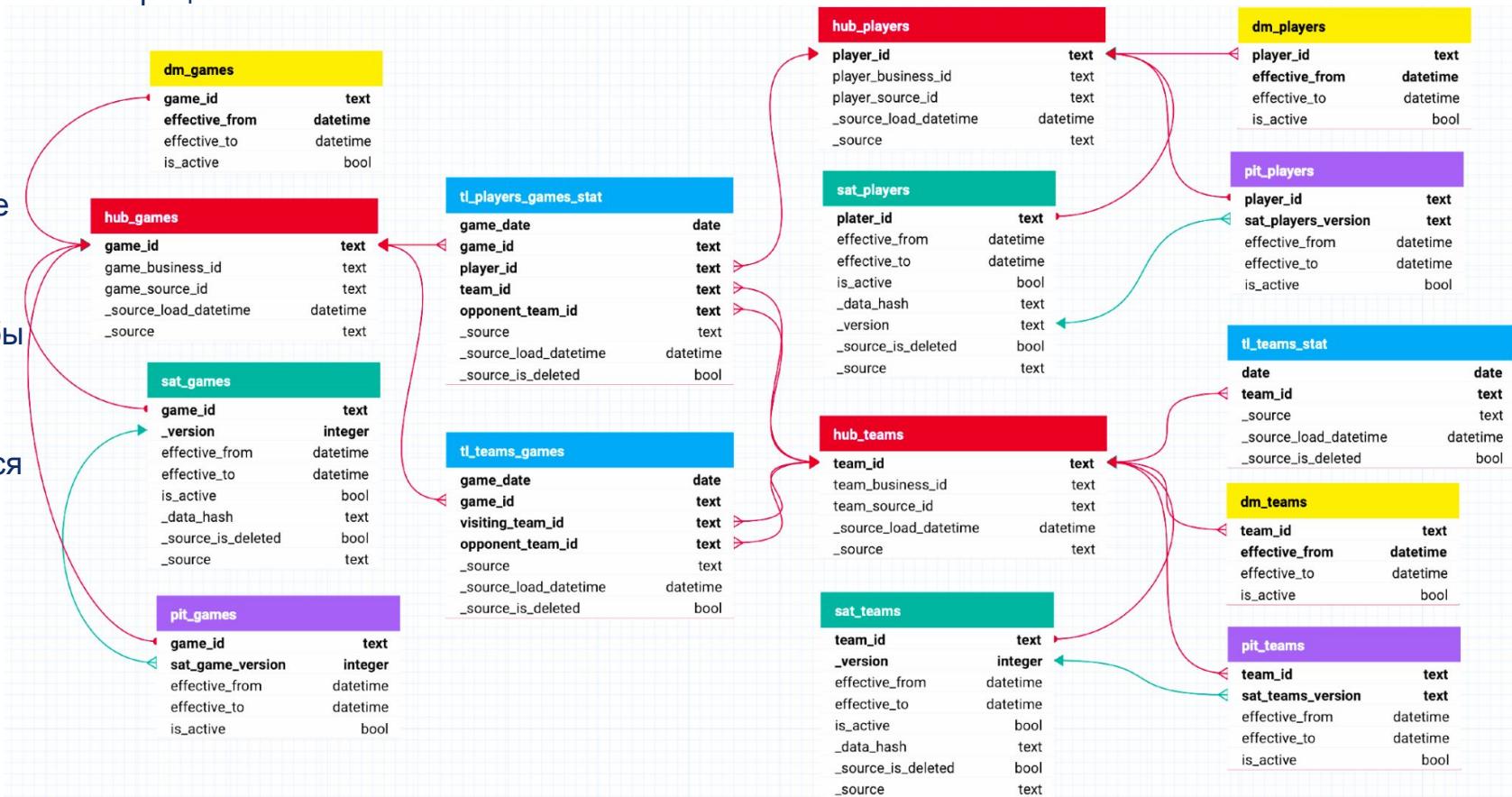


Архитектура DWH Data Vault 2.0

DWH (data warehouse, "склад данных") — структурированный набор хронологических данных, подробно описывающих бизнес-процессы и оптимизированных для аналитики этих бизнес-процессов.

Компоненты:

- **Hubs:** Хранение уникальных ключей бизнес-объектов
- **Satellites:** Содержат описательные атрибуты, связанные с каждым хабом (SCD2).
- **Transaction Links:** Соединяют хабы для описания взаимоотношений между объектами.
- **Point-in-Time Tables:** Используются для исторического отслеживания изменений и упрощения запросов по времени.
- **Data Marts:** Объединение всех сателитов (SCD2)





Архитектура DWH Data Vault 2.0

Выбор в пользу данной архитектуры был сделан из-за её превосходства в обработке и хранении больших объемов данных из различных источников. В отличие от традиционных подходов "звезда" или "снежинка", которые хорошо подходят для стабильных, неизменяющихся данных, Data Vault 2.0 позволяет гибко реагировать на изменения в данных, что крайне важно для постоянно обновляющейся статистики и результатов матчей NHL.

Преимущества:

1. Гибкость в изменениях структуры данных без перестройки всей базы данных
2. Высокий уровень отказоустойчивости и восстанавливаемости данных благодаря разделению данных на хабы, линки и сателлиты
3. Легкая масштабируемость
4. Поддержка историчности данных

Недостатки:

1. Сложность реализации и администрирования
2. Проектирование и настройка системы на базе Data Vault может занимать больше времени из-за её сложности и необходимости точного определения бизнес-ключей и отношений



Yandex DataLens

[Ссылка на дашборд](#)

На дашбордах собрана статистика по командам, игрокам и матчам, включая исторические данные, для аналитики.

Лидербордин

Дата	Сезон	Команда	Положение в лиге	Сыграно матчей	Набранные очки	Количество побед	Процент побед	Процент побед дома	Процент побед на выезде	Конференция	Положение в конференции
2024-04-18	20232024	New York Rangers 🏆	1 🥇	82	114	55	67,1%	73,2%	61,0%	Eastern	1
2024-04-18	20232024	Dallas Stars 🏆	2 🥈	82	113	52	63,4%	63,4%	63,4%	Western	1
2024-04-18	20232024	Carolina Hurricanes 🏆	3 🥉	82	111	52	63,4%	65,9%	61,0%	Eastern	2
2024-04-18	20232024	Winnipeg Jets 🏆	4	82	110	52	63,4%	65,9%	61,0%	Western	2
2024-04-18	20232024	Florida Panthers 🏆	5	82	110	52	63,4%	63,4%	63,4%	Eastern	3
2024-04-18	20232024	Vancouver Canucks 🏆	6	82	109	50	61,0%	65,9%	56,1%	Western	3
2024-04-18	20232024	Boston Bruins 🏆	7	82	109	47	57,3%	58,5%	56,1%	Eastern	4
2024-04-18	20232024	Colorado Avalanche 🏆	8	82	107	50	61,0%	75,6%	46,3%	Western	4
2024-04-18	20232024	Edmonton Oilers 🏆	9	82	104	49	59,8%	68,3%	51,2%	Western	5
2024-04-18	20232024	Toronto Maple Leafs 🏆	10	82	102	46	56,1%	53,7%	58,5%	Eastern	5
2024-04-18	20232024	Nashville Predators 🏆	11	82	99	47	57,3%	56,1%	58,5%	Western	6



Yandex DataLens

Ссылка на дашборд



Мин. возраст

18

Средн. возраст

29

Макс. возраст

40

Меткие парни				
Игрок	Команда	Страна	Амплуа	Эффективность бросков
Seth Jarvis	Carolina Hurricanes	CAN	C	23,58%
Sam Reinhart	Florida Panthers	CAN	C	21,81%
Brock Boeser	Vancouver Canucks	USA	R	20,35%
Brayden Point	Tampa Bay Lightning	CAN	C	20,08%
Leon Draisaitl	Edmonton Oilers	DEU	C	18,61%
Zach Hyman	Edmonton Oilers	CAN	L	18,58%
Auston Matthews	Toronto Maple Leafs	USA	C	17,99%
J.T. Miller	Vancouver Canucks	USA	C	17,94%
Kirill Kaprizov	Minnesota Wild	RUS	L	16,61%

Не менее 100 бросков за сезон



Yandex DataLens

Среднее время на льду за игру, мин

16,6

Лидеры по времени на льду

Игрок	Команда	Страна	Амплуа	Среднее время на льду
John Carlson	Washington Capitals	USA	D	26,05
Drew Doughty	Los Angeles Kings	CAN	D	25,88
Mike Matheson	Montréal Canadiens	CAN	D	25,55
Seth Jones	Chicago Blackhawks	USA	D	25,48
Rasmus Dahlin	Buffalo Sabres	SWE	D	25,42
Miro Heiskanen	Dallas Stars	FIN	D	25,31
Cale Makar	Colorado Avalanche	CAN	D	25,05
Brock Faber	Minnesota Wild	USA	D	24,97
Charlie McAvoy	Boston Bruins	USA	D	24,96
Victor Hedman	Tampa Bay Lightning	SWE	D	24,86
Roman Josi	Nashville Predators	CHE	D	24,71
Kris Letang	Pittsburgh Penguins	CAN	D	24,69
Quinn Hughes	Vancouver Canucks	USA	D	24,65
Noah Dobson	New York Islanders	CAN	D	24,49
Zach Werenski	Columbus Blue Jackets	USA	D	24,45

<

1

>

Строки: 1-15

[Ссылка на дашборд](#)

Среднее время штрафа за игру, мин

0,52

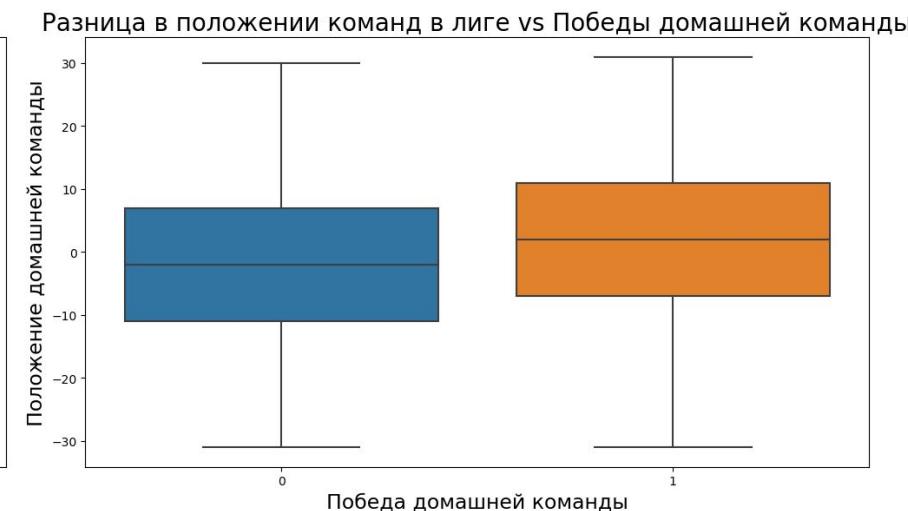
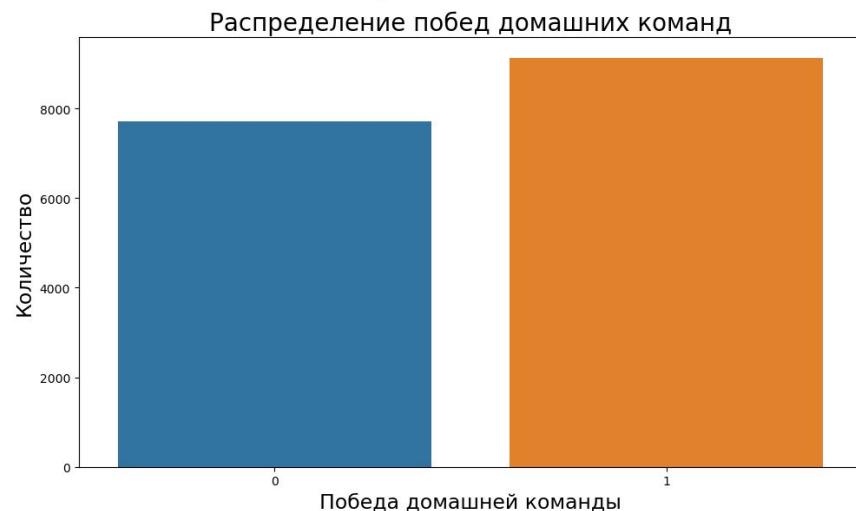
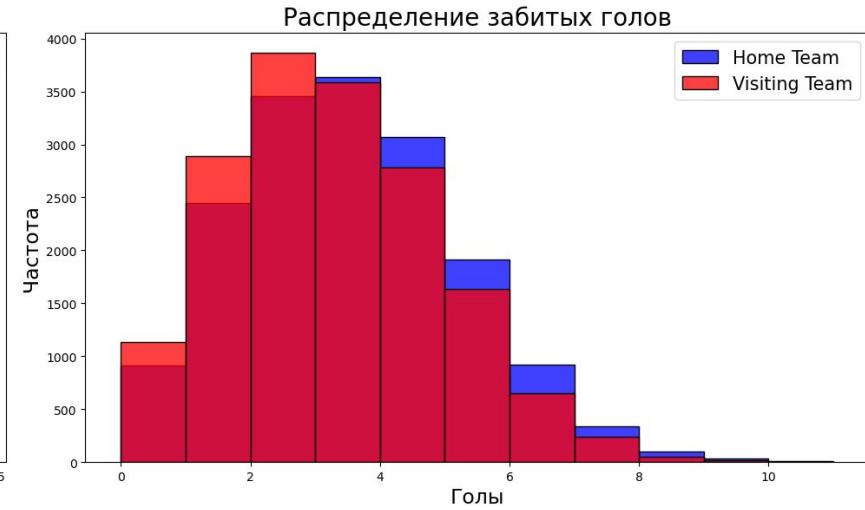
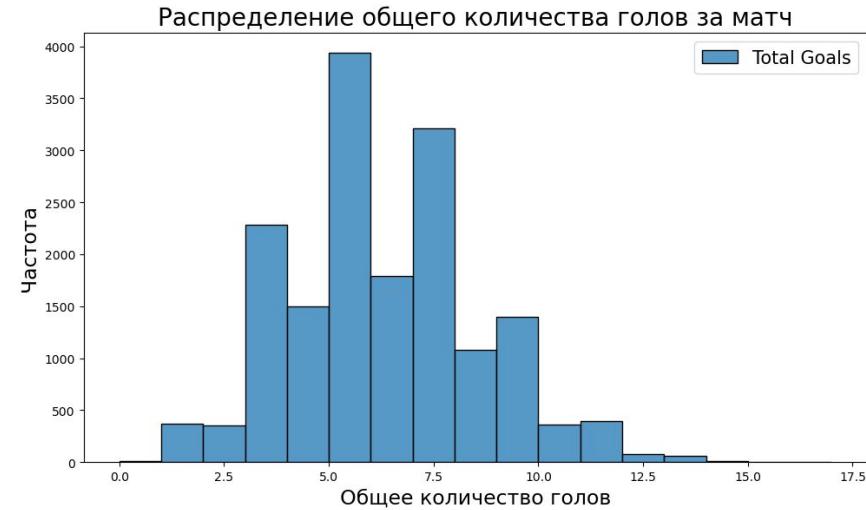
Лидеры по штрафным минутам



Не менее 40 минут штрафа

Предобработка данных, EDA

- Большинство игр заканчивается с общим количеством голов в диапазоне от 5 до 7.
 - Данные показывают, что домашние команды (синий цвет) забивают голы чаще, чем гостевые команды (красный цвет), что может указывать на домашнее преимущество.
 - Есть тенденция домашних команд к победе в матчах.





Feature engineering

Данные игры:

- Дата игры
- Домашняя команда
- Гостевая команда

Статистики команды на день игры за сезон:

- Позиция в лиге
- Конференция
- Процент побед
- Процент забитых голов и пропущенных голов

Статистики команды за последние 10 игр:

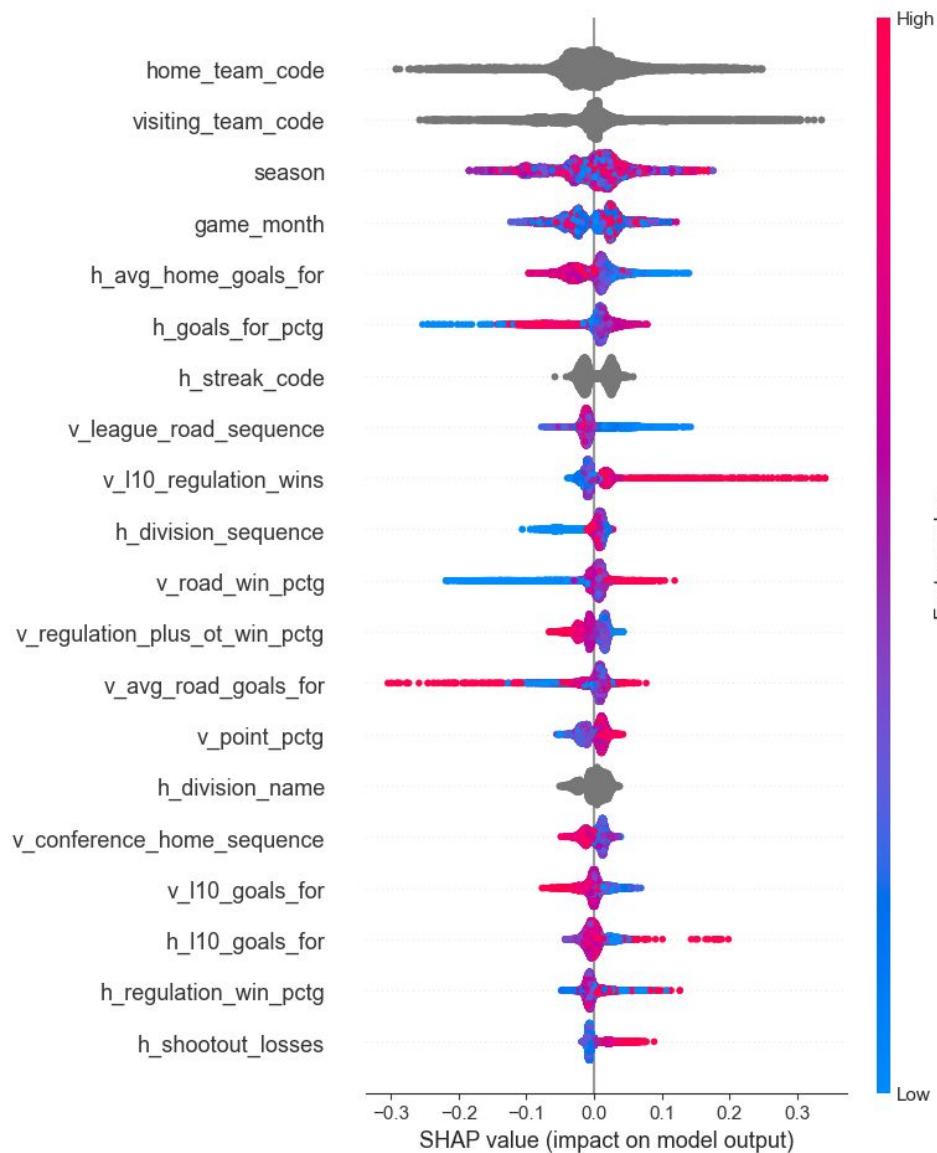
- Процент побед для матчей дома и матчей на выезде

Дополнительные признаки:

- Серия побед или поражений
- Разница в позиции между командами
- Среднее число забитых голов за игру

Целевой признак:

- Победа домашней команды





ML Прогнозирование победителя

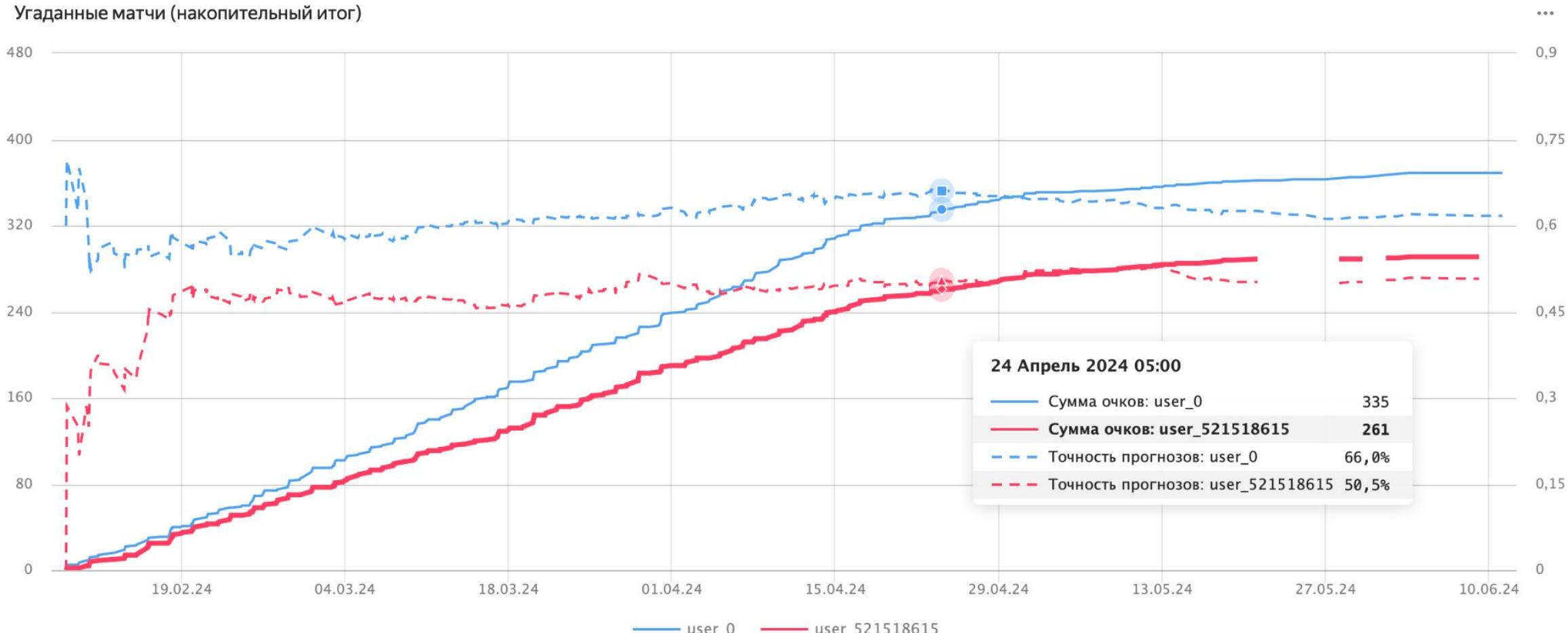
- Основные метрики оценки качества моделей – F1 мера и ROC AUC
- Нам удалось превзойти первый baseline 1 – Random
- После отбора признаков лучшее качество показала модель – CatBoost Classifier
- Нам удалось достичь уровня, сопоставимого с коэффициентами букмекеров baseline 2 – Букмекеры

	model	accuracy	precision	recall	f1_score	roc_auc	threshold
7	Baseline - Bookmaker Odds	0.612488	0.627010	0.710383	0.666097	0.603017	0.50
6	CatBoost Classifier Top	0.605550	0.617785	0.721311	0.665546	0.594351	0.48
4	Logistic Regression Top	0.607532	0.627288	0.686703	0.655652	0.599873	0.48
1	Logistic Regression	0.610505	0.643382	0.637523	0.640439	0.607892	0.50
3	CatBoost Classifier	0.596630	0.624126	0.650273	0.636931	0.591441	0.50
5	SVC Top	0.591675	0.617900	0.653916	0.635398	0.585654	0.52
2	SVC	0.588702	0.618794	0.635701	0.627134	0.584155	0.52
0	Baseline - Random	0.483647	0.526316	0.510018	0.518039	0.481096	NaN



ML Геймификация прогноза в Yandex Data Lens

[Ссылка на дашборд](#)





ML Геймификация прогноза в Yandex Data Lens

[Ссылка на дашборд](#)

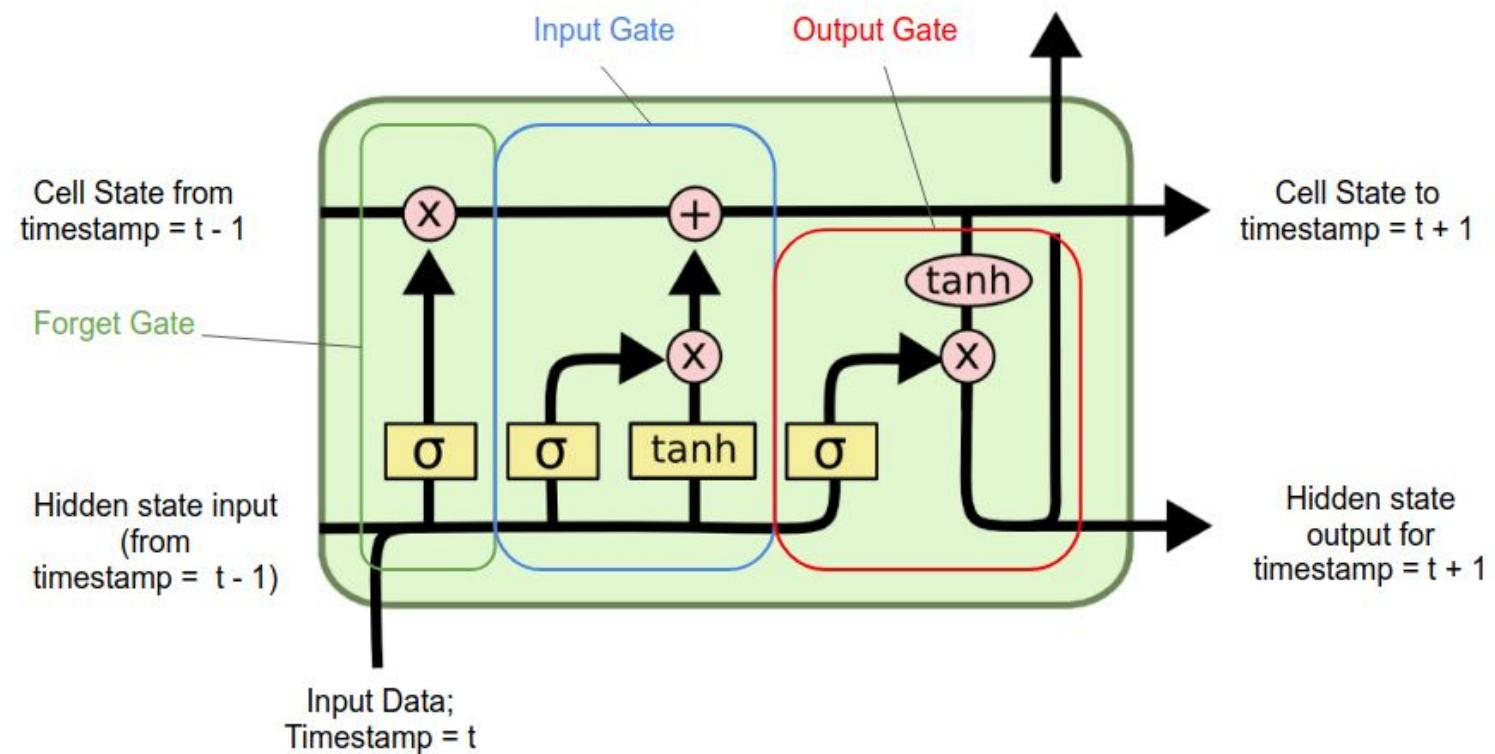
Таблица с прогнозами и результатами

ID пользователя	Дата игры (UTC+3)	Домашняя команда	Счет	Команда на выезде	Предсказанный счет	Прогноз победителя	Набранные очки
user_0	11.06.2024 03:00:00	Florida Panthers	0 : 0	Edmonton Oilers	null	Florida Panthers	0
user_521518615	09.06.2024 03:00:00	Florida Panthers	0 : 0	Edmonton Oilers	4 : 3	Florida Panthers	0
user_0	09.06.2024 03:00:00	Florida Panthers	0 : 0	Edmonton Oilers	null	Florida Panthers	0
user_521518615	03.06.2024 03:00:00	Edmonton Oilers	2 : 1	Dallas Stars	4 : 3	Edmonton Oilers	1
user_0	03.06.2024 03:00:00	Edmonton Oilers	2 : 1	Dallas Stars	null	Edmonton Oilers	1
user_521518615	02.06.2024 03:00:00	Florida Panthers	2 : 1	New York Rangers	3 : 4	New York Rangers	0
user_0	02.06.2024 03:00:00	Florida Panthers	2 : 1	New York Rangers	null	Florida Panthers	1
user_521518615	01.06.2024 03:30:00	Dallas Stars	1 : 3	Edmonton Oilers	2 : 4	Edmonton Oilers	1
user_0	01.06.2024 03:30:00	Dallas Stars	1 : 3	Edmonton Oilers	null	Edmonton Oilers	1
user_0	31.05.2024 03:00:00	New York Rangers	2 : 3	Florida Panthers	null	Florida Panthers	1
user_521518615	30.05.2024 03:30:00	Edmonton Oilers	5 : 2	Dallas Stars	2 : 5	Dallas Stars	0
user_0	30.05.2024 03:30:00	Edmonton Oilers	5 : 2	Dallas Stars	null	Dallas Stars	0



DL подход

LSTM – это тип рекуррентной нейронной сети (RNN), которая способна учитывать долгосрочные зависимости в последовательных данных, что делает её полезной для анализа временных рядов и последовательностей.





Реализация сервиса

Сервис реализован на облачной платформе Yandex Cloud.

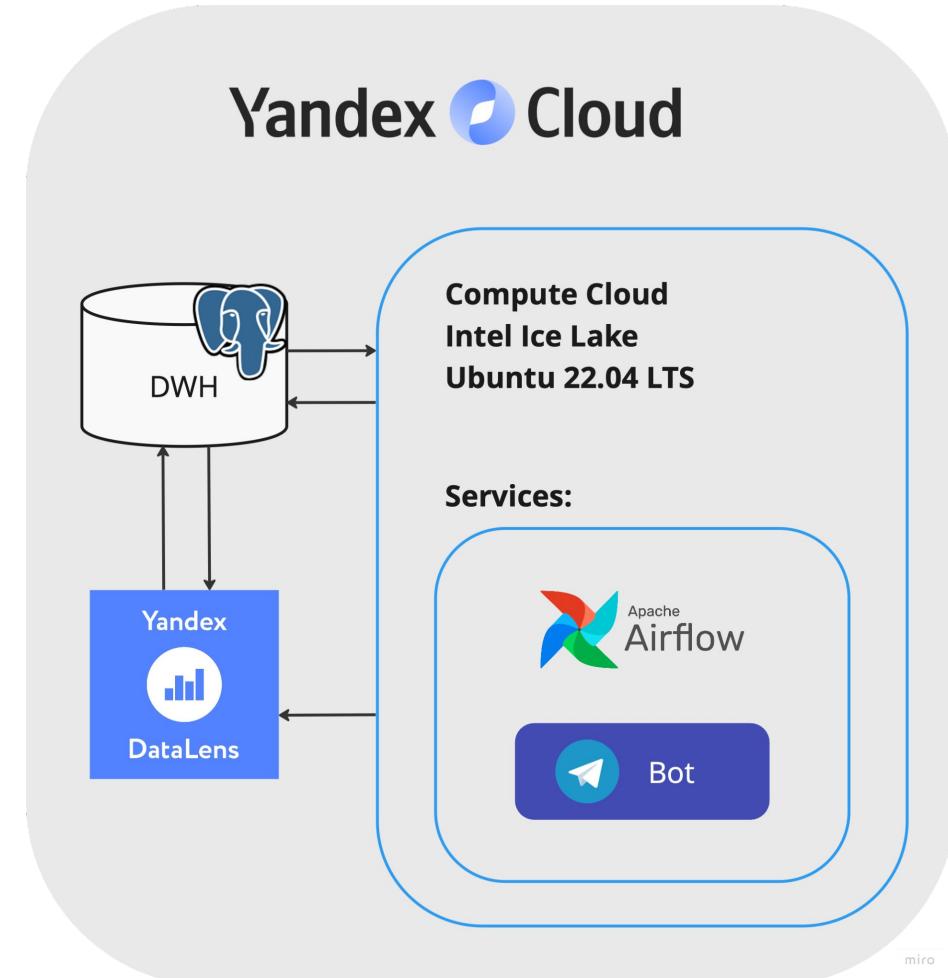
Включает в себя:

- Managed Service for PostgreSQL
- DataLens
- Compute Cloud (сервер)

На сервере в качестве сервисов запущены Apache Airflow и Telegram Bot

```
airflow-scheduler.service
airflow-webserver.service
nhl_bot.service
```

```
loaded active running Airflow scheduler daemon
loaded active running Airflow webserver daemon
loaded active running NHL Telegram Bot
```





Telegram Bot



Создан с помощью библиотеки TeleBot, является основным UI интерфейсом к данным. Благодаря асинхронной обработке, бот способен эффективно управлять множественными запросами и обеспечивать быстрое взаимодействие с минимальными задержками.

Maxim Galanov
[/start](#)

NHL
Привет!
Вот, что я умею:
Показать результаты игр за последние 7 дней: [/results](#)
Статистика игроков: [/player_stats](#)
Статистика команд: [/team_stats](#)
Сделать прогноз на исход ближайших матчей: [/make_bet](#)
Посмотреть прогноз ML модели на ближайшие матчи: [/show_preds](#)
Дашборды в DataLens: [/datalens](#)

Maxim Galanov
[/results](#)

NHL
Результаты за последние 7 дней:
2024-05-09 🏆 Dallas Stars 5 : 3 Colorado Avalanche
2024-05-09 Carolina Hurricanes 2 : 3 New York Rangers 🏆

2024-05-08 🏆 Vancouver Canucks 5 : 4 Edmonton Oilers
2024-05-08 🏆 Florida Panthers 6 : 1 Boston Bruins

2024-05-07 Dallas Stars 3 : 4 Colorado Avalanche 🏆
2024-05-07 🏆 New York Rangers 4 : 3 Carolina Hurricanes

2024-05-06 Florida Panthers 1 : 5 Boston Bruins 🏆

2024-05-05 🏆 Dallas Stars 2 : 1 Vegas Golden Knights
2024-05-05 🏆 New York Rangers 4 : 3 Carolina Hurricanes

2024-05-04 🏆 Boston Bruins 2 : 1 Toronto Maple Leafs

Maxim Galanov
[/show_preds](#)

NHL
Вот предсказания на ближайшие 2 дня:
05-12 21:30 🏆 Boston Bruins vs Florida Panthers
05-13 00:30 Edmonton Oilers vs Vancouver Canucks 🏆
05-11 22:00 Carolina Hurricanes vs New York Rangers 🏆
05-12 01:00 Colorado Avalanche vs Dallas Stars 🏆
* Сортировка по убыванию уверенности модели.

Maxim Galanov
[/make_bet](#)

NHL
Выберите матч для прогноза:
 05-11 22:00 Carolina Hur vs New York Ran
 05-12 01:00 Colorado Ava vs Dallas Sta
 05-12 21:30 Boston Bru vs Florida Pan
 05-13 00:30 Edmonton Oil vs Vancouver Can

Выберите победителя:
[Edmonton Oilers](#) [Vancouver Canucks](#)

Хотите предсказать точный счет?
[Да](#) [Нет](#)

Введите счет в формате 'home:away'
Maxim Galanov
3:2

NHL
Ваш прогноз со счетом сохранен!

[Посмотреть свои прогнозы](#)

Maxim Galanov
[/player_stats](#)

NHL
Вратарь или полевой игрок?

Maxim Galanov
Полевой игрок

NHL
Введите имя, например, Alex Ovechkin

Maxim Galanov
Alex Ovechkin

NHL



Дата рождения: 1985-09-17
Возраст: 38
Страна: RUS
Город: Moscow
Команда: Washington Capitals (WSH)
Игр сыграно: 83
Очки: 65
Голов: 31
Показатель полезности: -24

Более подробную информацию можете посмотреть в нашем [дашборде по игрокам](#)

Maxim Galanov
[/team_stats](#)

NHL
Введите название команды или трикод, например 'Washington Capitals' или 'WSH'

Maxim Galanov
NYR

NHL



New York Rangers
Конференция: Eastern
Дивизион: Metropolitan
Сыграно матчей: 82
Место в лиге: 1
Очки: 114
Побед: 55 (67.07%)
Забито голов: 282, Пропущено голов: 229

Более подробную информацию можете посмотреть в нашем [дашборде по командам](#)



