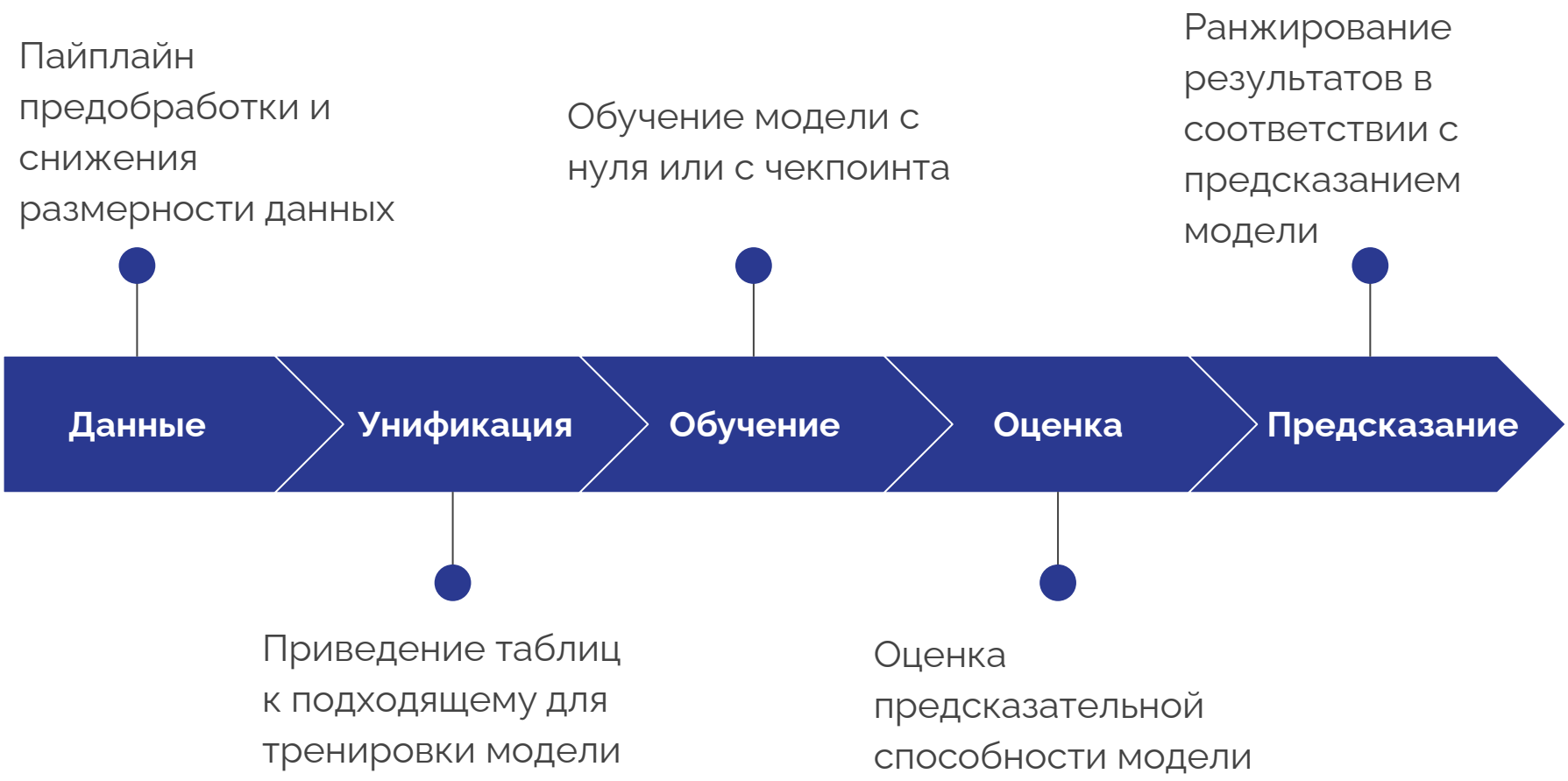
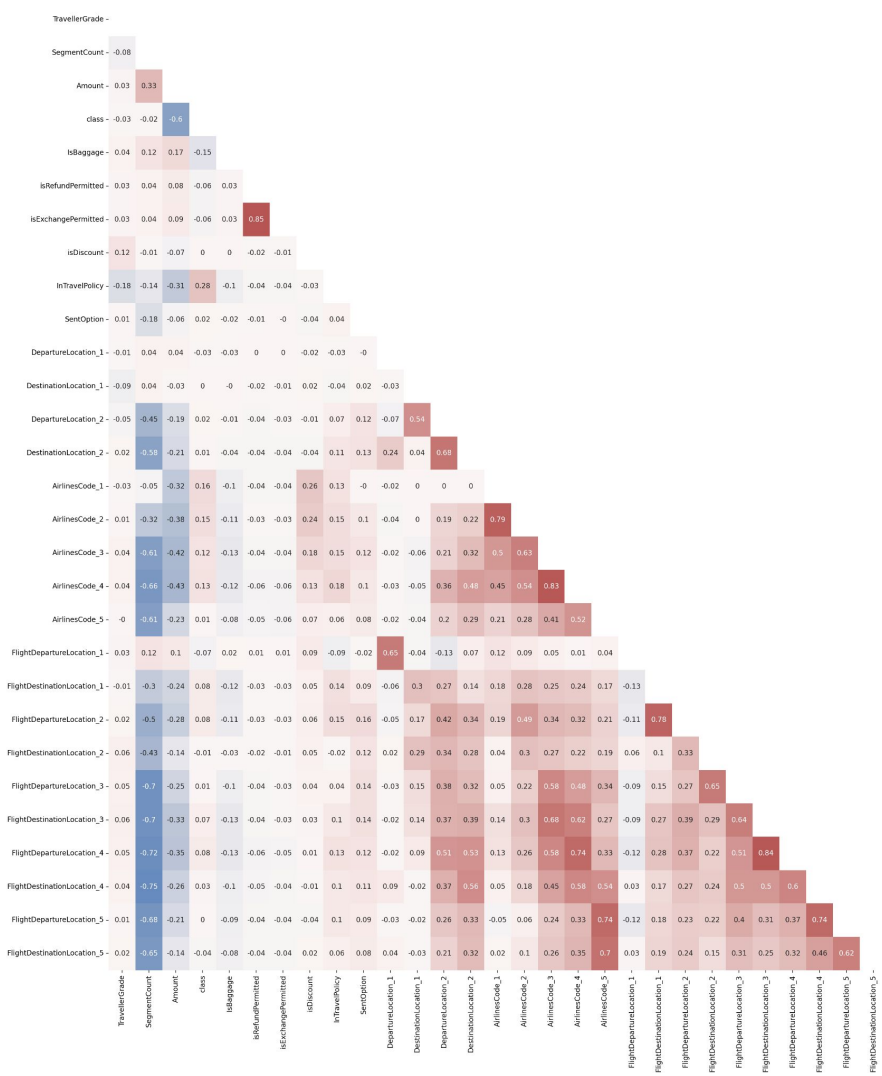
# Решение команды CRD

*Aeroclub Challenge 2023*


Пайплайн предобработки и снижения размерности данных
Обучение модели с нуля или с чекпоинта
Ранжирование результатов в соответствии с предсказанием модели
Данные
Унификация
Обучение
Оценка
Предсказание
Приведение таблиц к подходящему для тренировки модели виду
Оценка предсказательной способности модели

# Обработка данных

Результаты



# Решение

CatBoost

Градиентный бустинг деревьев решений с пайплайном предобработки данных, валидацией модели и дообучением с чекпоинта

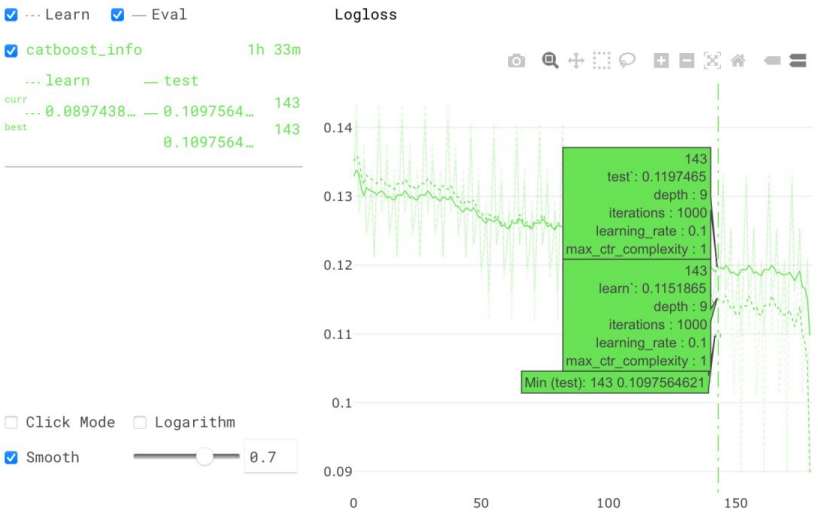
# Оценка важности фичей

В результате анализа данных было выявлено, что наибольшее влияние на итоговый выбор модели имеют такие переменные как:

**Amount** - стоимость выбранного варианта
**AirlinesCode (½)** - авиакомпания, организующая перелет
**isDiscount** - наличие скидки

| | Feature Id | Importances |
|---|---|---|
| 0 | Amount | 14.306613 |
| 1 | AirlinesCode_1 | 12.680958 |
| 2 | isDiscount | 9.810465 |
| 3 | AirlinesCode_2 | 9.630254 |
| 4 | InTravelPolicy | 9.111123 |
| 5 | isRefundPermitted | 8.450338 |
| 6 | IsBaggage | 8.338336 |
| 7 | isExchangePermitted | 4.545903 |
| 8 | DestinationLocation_1 | 4.509348 |
| 9 | FlightDestinationLocation_1 | 3.052727 |
| 10 | FlightDestinationLocation_2 | 2.972441 |
| 11 | FlightDepartureLocation_1 | 2.846693 |
| 12 | FlightDepartureLocation_2 | 2.221847 |
| 13 | DepartureLocation_2 | 1.625754 |
| 14 | DestinationLocation_2 | 1.502072 |
| 15 | FlightDestinationLocation_3 | 1.272931 |
| 16 | SegmentCount | 0.835218 |
| 17 | class | 0.631148 |
| 18 | FlightDepartureLocation_3 | 0.521605 |
| 19 | AirlinesCode_3 | 0.314990 |
| 20 | DepartureLocation_1 | 0.289502 |
| 21 | FlightDestinationLocation_4 | 0.191061 |
| 22 | FlightDepartureLocation_4 | 0.155808 |
| 23 | AirlinesCode_4 | 0.109973 |
| 24 | FlightDestinationLocation_5 | 0.031830 |
| 25 | FlightDepartureLocation_5 | 0.023945 |
| 26 | AirlinesCode_5 | 0.017116 |

# Оптимизация функции потерь




# Подобранные оптимальные параметры

Depth: 16
Iterations: 2000
Learning Rate: 0.1

# Основные метрики

test_size = 0.2

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Not choosed | 0.99 | 1.00 | 0.99 | 319947 |
| Choosed | 0.53 | 0.24 | 0.33 | 5932 |
| accuracy | | | 0.98 | 325879 |
| macro avg | 0.76 | 0.62 | 0.66 | 325879 |
| weighted avg | 0.98 | 0.98 | 0.98 | 325879 |

# Дополнительные метрики

Specificity

0.996005588425583

ROC AUC score

0.9785627169004467

# Возможные доработки

## Данные

Перенос обработки данных в *Airflow* и унификация процесса обработки данных, а также дальнейшего сохранения в *feature store*

## Модель

Вынос работы модели  в отдельный сервис, при помощи *MLFlow* и настройка регулярного дообучения и обновления артефактов на основе *S3*

## Оценка

Внедрение карты метрик и динамический сбор пользовательского фидбека с последующей обработкой и передачей в модель