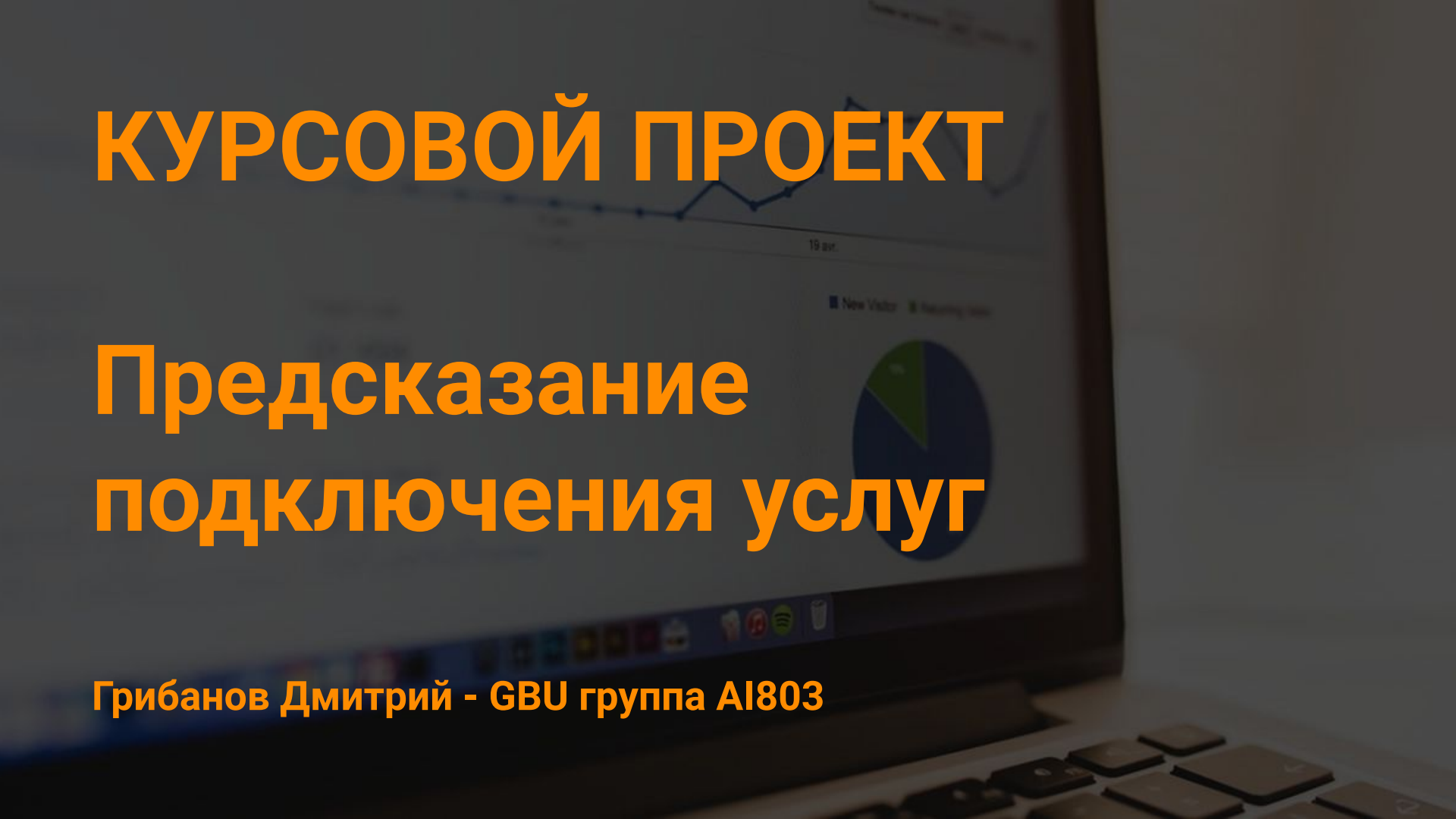


КУРСОВОЙ ПРОЕКТ

Предсказание подключения услуг

Грибанов Дмитрий - GBU группа AI803

Цель проекта

Построить алгоритм, который для каждой пары пользователь-услуга определяет вероятность подключения услуги

Исходные данные

Данные представлены 2 файлами (data_train и data_test), содержащими основные признаки (buy_time, id, id_vas), а также файлом features, содержащим детальные данные по каждому пользователю с анонимизированными признаками - 256 штук (метод PCA понижает размерность до 45 с сохранением 99,9% информации) для ускорения обучения

Присутствует сильный дисбаланс классов в пользу 0 класса

Пропуски данных отсутствуют

Датасеты объединяются с помощью метода merge_asof по id (жестко) и максимально близко по дате предложения услуги с датой “слепка” пользователя

Метрика - f1_score(macro)

Модели

Logistic Regression (baseline) - 0.72
Logistic Regression (L2, SVD) - 0.42
GradientBoostingClassifier - 0.67
CatBoostClassifier - 0.73

Алгоритм поиска гиперпараметров для CatBoostClassifier

1. Trainset - > Train (0,7) и Valid (0,3)

2. Подбор параметров: GridSearchCV (cv = 3), метрика: F1 (average=macro)

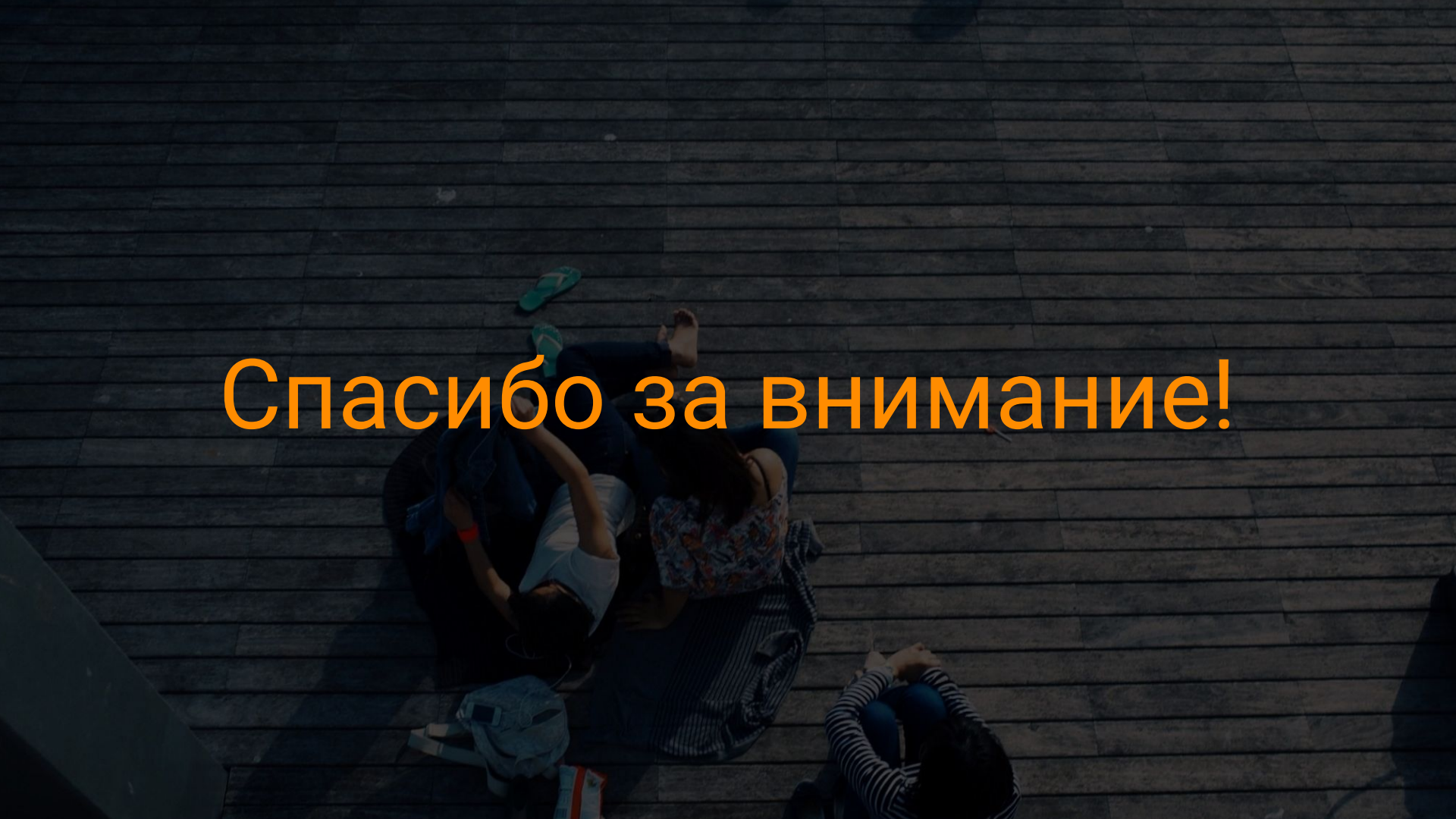
3. Обучение модели на train с оптимальными параметрами

Параметры финальной модели

CatBoostClassifier

n_estimators=200
max_depth=3
l2_leaf_reg=15
learning_rate=0.1

Метрика - f1_score(macro) - 0.73


Спасибо за внимание!