

Цели и задачи построения модели дохода

Построить и обучить ML-
модель для точного прогноза
доходов (метрика WMAE → min).

Создать рабочий
прототип (веб-интерфейс
для сотрудника банка)

Обеспечить
мониторинг качества модели и
успешности рекомендаций в
реальном времени.

Показать бизнес-
ценность решения для
повышения эффективности
кросс-продаж и клиентского
сервиса.

Используемые библиотеки

1 Обработка и анализ данных

Pandas – загрузка, обработка и анализ табличных данных.

Numpy – математические вычисления и работа с массивами

2 Визуализация

matplotlib.pyplot – построение базовых графиков и диаграмм

Seaborn – создание статистических визуализаций

3 Машинное обучение

Sklearn – предобработка, валидация, метрики и модели
(randomforest)

Xgboost, lightgbm – продвинутые алгоритмы градиентного бустинга

Данные и их предобработка

1

Загрузка и первичный анализ данных для оценки их структуры и полноты.

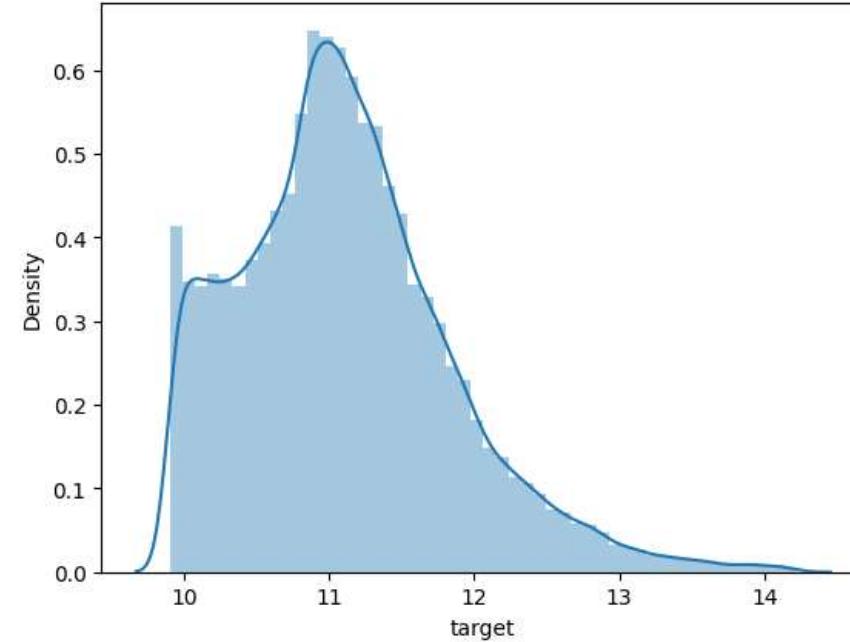
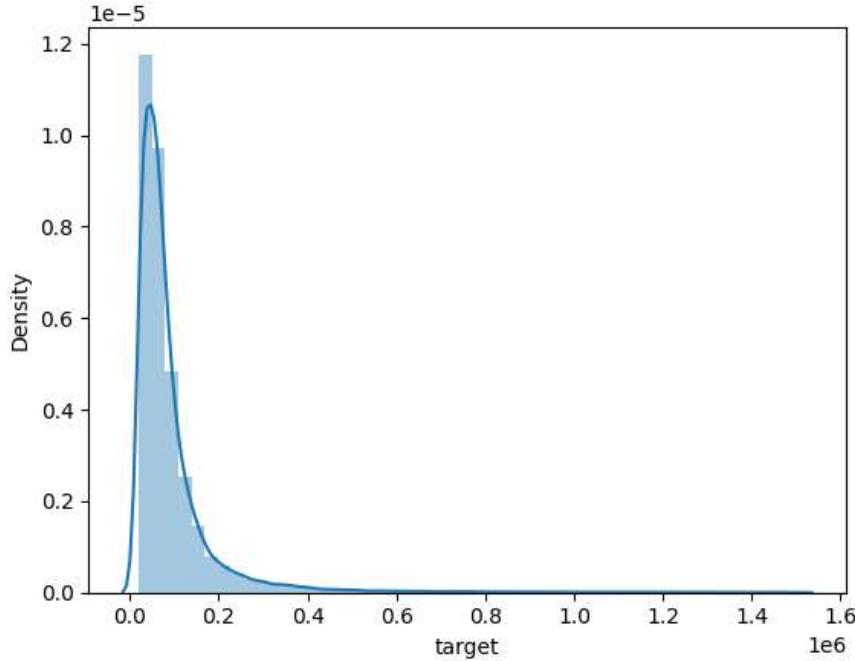
2

Обработка пропусков: Удалены признаки с уровнем пропусков >90%, остальные — обработаны.

3

Борьба с мультиколлинеарностью
Изключены коррелирующие переменные с наименьшей полнотой данных.

Распределение целевой переменной target



Правосторонняя асимметрия.
Выполним логарифмическое
преобразование.

Числовые и категориальные признаки и их пропущенные значения

Числовые признаки

Определены как количественные данные,
требуют обработки

Категориальные признаки

Используют качественные данные, важна
классификация

Обработка пропущенных значений

Заполнение или удаление для анализа

Влияние на анализ

Пропуски могут искажать результаты и выводы

Наиболее сильная корреляция target

	target	Total	Percent
first_salary_income	0.928217	68118	0.887115
salary_6to12m_avg	0.927699	61911	0.806280
dp_payoutincomedata_payout_sum_3_month	0.644542	65855	0.857643
dp_payoutincomedata_payout_avg_prev_year	0.630530	68915	0.897494
turn_cur_cr_sum_v2	0.630285	17354	0.226005
dp_ilss_paymentssum_avg_12m	0.613925	60283	0.785078
avg_6m_all	0.570469	21656	0.282031
dp_payoutincomedata_payout_max_6_month	0.564312	65855	0.857643
dp_payoutincomedata_payout_max_3_month	0.561221	65855	0.857643
dp_ilss_paymentssum_month_avg	0.558366	60283	0.785078
avg_cur_db_turn	0.556654	16572	0.215821
avg_cur_cr_turn	0.556571	16572	0.215821
dp_ilss_avg_salary_3y	0.536375	60387	0.786432

Построение модели и выбор между LightGBM и XGBoost

1

Сравнение
производительности

2

Особенности LightGBM
Преимущества: высокая
скорость и малое
потребление памяти

3

Особенности Xboost
Преимущества:
стабильность и
поддержка различных
функций

LightGBM: WMAE = 43097.9945
XGBoost: WMAE = 44070.3783