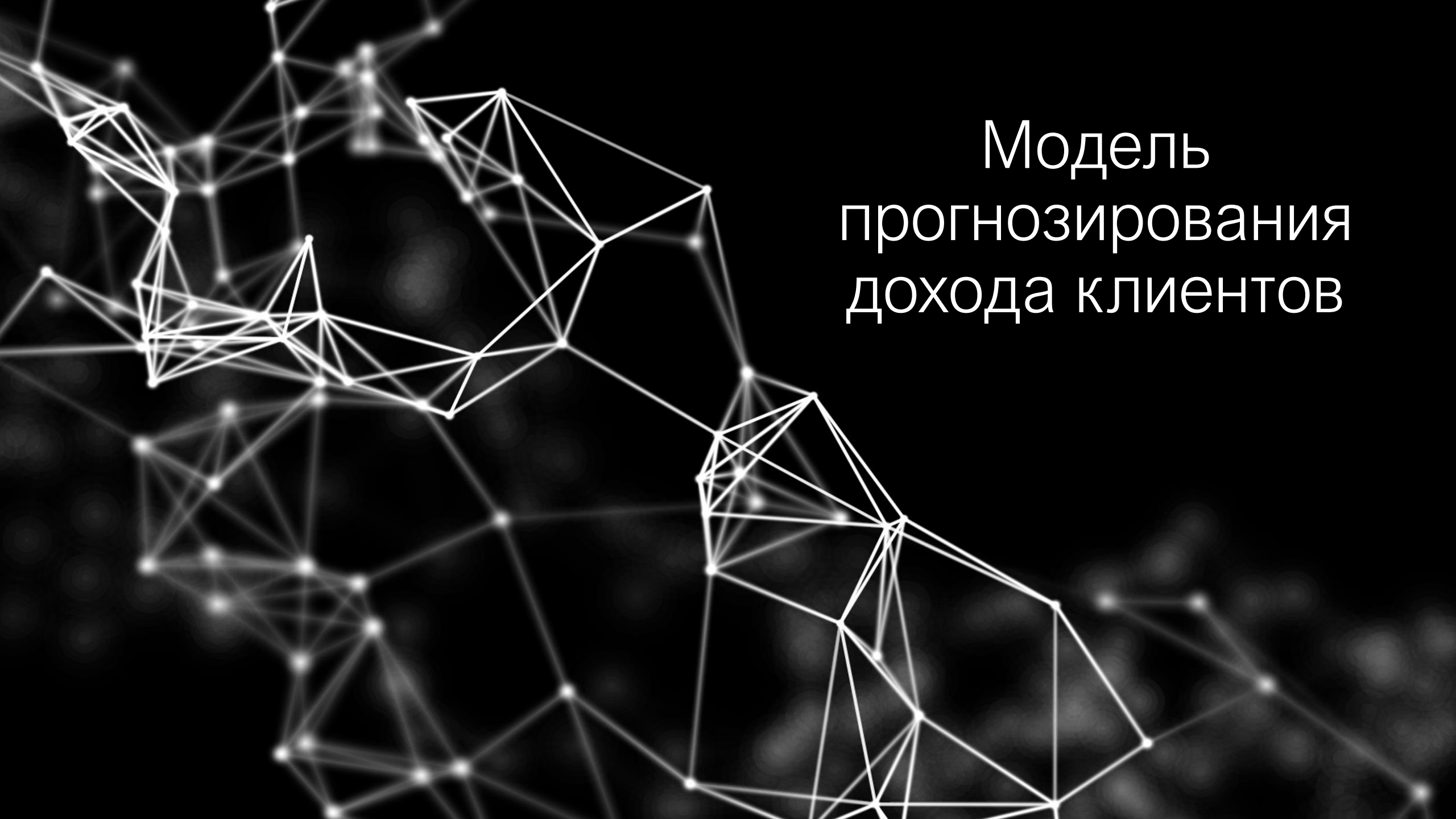


Модель прогнозирования дохода клиентов

Алгоритм и структура решения задачи

Предобработка данных

- Первичная агрегация данных (Каждому клиенту соответствует только одна строка таблицы)
- Создание колонки CLTV в тестовой выборке
- Генерация дополнительных признаков
- Поиск наиболее важных признаков

Тестирование гипотез

- Обучение одиночных регрессионных моделей
- Обучение ансамблей регрессионных моделей
- Выбор лучших моделей и подбор гиперпараметров
- Выбор лучшей модели по точности и простоте
- Составление ответов тестовой выборки

Агрегация данных

Название признака	Агрегационные функции	Название признака	Агрегационные функции
cif_id	Группировка	sa_volume	Среднее значение, max
dlk_cob_date	Количество, max,min	mf_volume	Среднее значение, max
gi_smooth_3m	Среднее значение, сумма	dc_cash_spend_v	Среднее значение, max
big_city	Последнее значение	dc_cash_spend_c	Среднее значение
cu_gender	Последнее значение	cc_cash_spend_v	Среднее значение, max
cu_education_level	Последнее значение	cc_cash_spend_c	Среднее значение
cu_empl_area	Последнее значение	dc_pos_spend_v	Среднее значение, max
cu_empl_level	Последнее значение	dc_pos_spend_c	Среднее значение
payroll_f	Последнее значение	cc_pos_spend_v	Среднее значение, max
cur_quantity_pl	Последнее значение	cc_pos_spend_c	Среднее значение
cur_quantity_mort	Последнее значение, среднее	ca_f	Последнее значение
cur_quantity_cc	Последнее значение, среднее	rc_session_qnt_cur_mon	Последнее значение, среднее
cur_quantity_deposits	Последнее значение, среднее	cur_qnt_sms	Среднее значение
cur_quantity_dc	Последнее значение, среднее	active	Последнее значение, среднее
cur_quantity_accounts	Последнее значение, среднее	standalone_dc_f	Последнее значение
cur_quantity_saccounts	Последнее значение, среднее	standalone_payroll_dc_f	Последнее значение
cur_quantity_mf	Последнее значение, среднее	standalone_nonpayroll_dc_f	Последнее значение
cc_balance	Среднее значение, min, max	salary	Среднее значение, min, max
cl_balance	Последнее значение, max	cu_age	Последнее значение
ml_balance	Последнее значение, max	cu_mob	Последнее значение
pl_balance	Последнее значение, max	cu_empl_cur_dur_m	Последнее значение
td_volume	Среднее значение, max	is_married	Последнее значение
ca_volume	Среднее значение, max		

Генерация новых признаков

Признак	Описание
gi_smooth_3m_agg_linear_trend	Линейный тренд доходов от клиента (направление изменения)
gi_smooth_3m__fft_aggregated	Быстрое преобразование Фурье, прореживание по частоте
gi_smooth_3m__mean_second	Численная производная второго порядка (скорость изменения)
salary__absolute_sum_of_changes	Сумма модулей изменений зарплаты
salary__sample_entropy	Энтропия зарплаты
salary_fft_aggregated_aggtype	Быстрое преобразование Фурье з/п
all_spends__absolute_sum	Сумма модулей изменений трат
all_spends__mean_change	Среднее изменение трат
all_credits__absolute_sum_of_changes	Сумма модулей изменений кредитов
all_credits__mean_second_derivative	Численная производная второго порядка (скорость изменения)
all_debits__mean_change	Среднее изменение дебетовых счетов

Из 76 сгенерированных признаков только 10 оказались значимы.

Для извлечения трендов и характеристики изменения сигнала была использована библиотека tsfresh

Тестирование регрессионных моделей

Модель	MAPE	MAE	Комментарий
Линейная регрессия	9400	0.82	Простая модель. Даёт хороший результат на всей выборке.
SVR	8000	2.9	Сложная модель, предсказание занимает слишком много времени. Ошибка большая.
Random Forest	10000	1.44	Слишком быстро переобучается. Чем сильнее переобучение, тем хуже MAPE.
Gradient Boosting	20000 –> 2500	3.2 –> 0.93	Варьируя фичи удалось достичь отличного результата. Выбрана за основу.

В качестве лучшей модели был выбран градиентный бустинг с параметрами:

- 150 эстиматоров
- Максимальной глубиной дерева 7
- Кросс валидация на 10 фолдах

Градиентный бустинг является оптимальным решением относительно сложности и точности

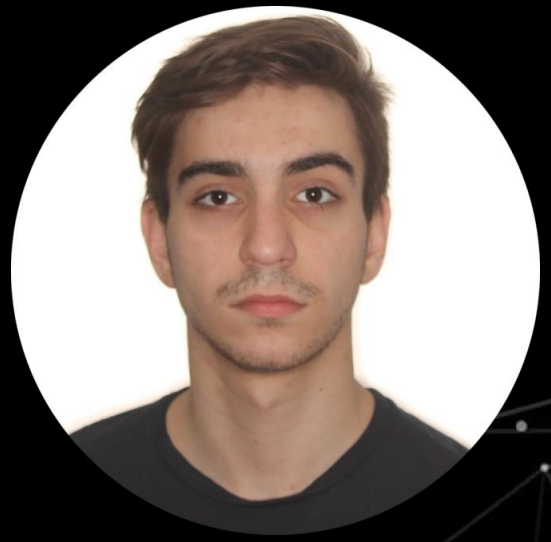
Ансамбль моделей работает менее точно, чем нейронные сети, но лучше, чем одиночные модели.

Исходя из исследования данных и тестирования ряда моделей, следует, что наиболее точным оказался ансамбль моделей – градиентный бустинг, при 150 эстиматорах и максимальной глубиной дерева равной 7.

CLTV – сложно прогнозируемый показатель так как зависит от многочисленных факторов. Тем не менее данная модель является масштабируемой. При изменении столбца CLTV, на сумму прибыли банка по заданному количеству месяцев.

Наша команда

Синяев Максим



Камаев Богдан

Кущ Артем

Алексей Мышлянов