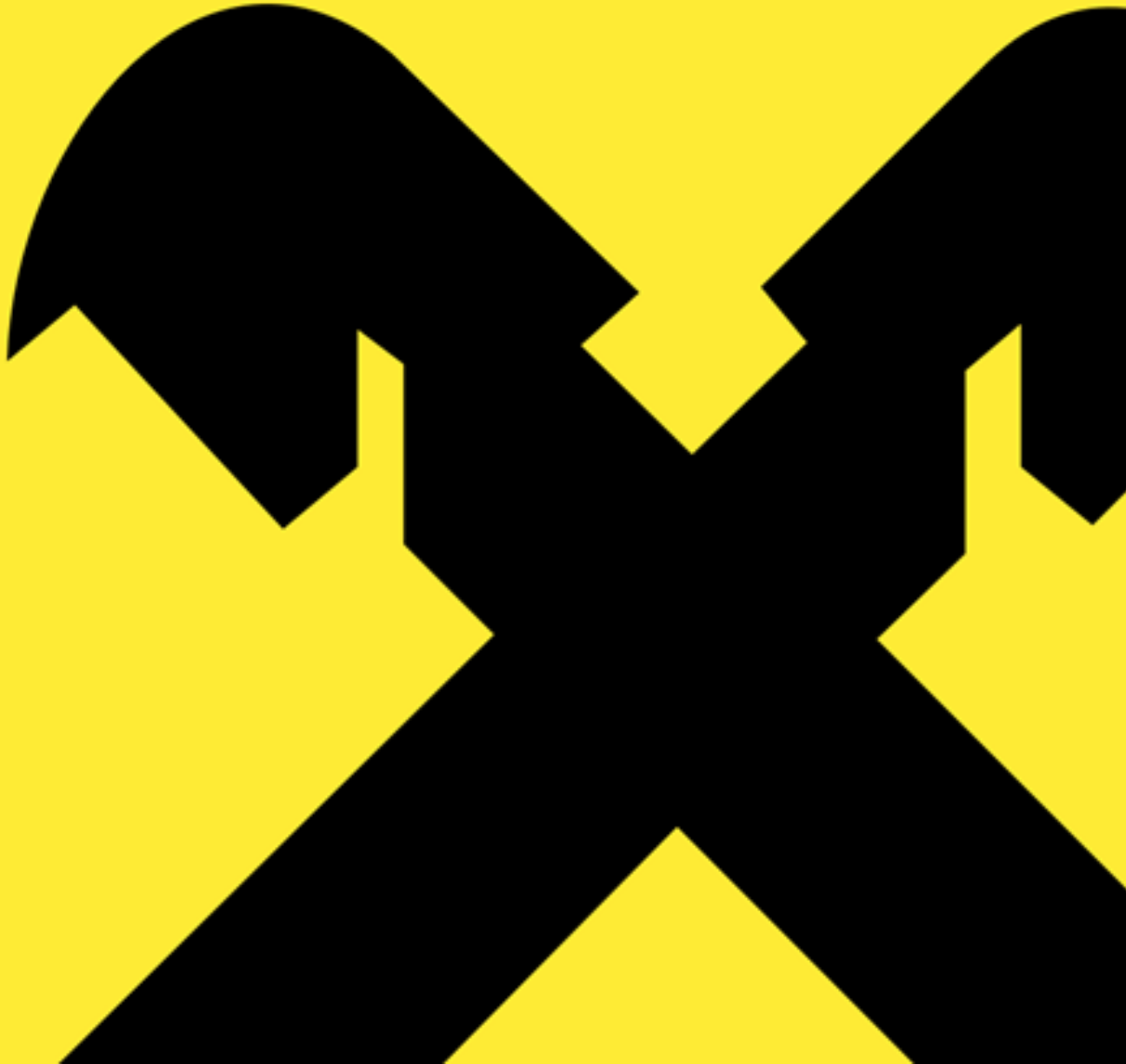
Прогноз значений CLTV клиентов «Райффайзен Банка»



Команда MYNI

Михаил Сарафанов
Юлия Борисова
Наталья Власова
Ирина Макаркина

Команда

Михаил Сарафанов

mik_sar@mail.ru

Студент 1 курса магистратуры ИТМО
Направление: Big data and Machine Learning

Юлия Борисова

yulashka.htm@yandex.ru

Студентка 1 курса магистратуры ИНоЗ СПбГУ
Направление: Картография и геоинформатика

Наталья Власова

natalya9vlasova@gmail.com

Студентка 1 курса магистратуры ВШМ СПбГУ
Направление: Business analytics and Big data

Ирина Макаркина

st079308@student.spbu.ru

Студентка 1 курса магистратуры ВШМ СПбГУ
Направление: Business analytics and Big data

Подготовка данных

Результаты первичного анализа

- Признаки *'cu_education_level'* и *'cu_eduaction_level'* закодированы по-разному, но несут одну и ту же информацию
- Для некоторых признаков отсутствует значительная часть данных (NaN), поэтому при построении модели они не использовались:

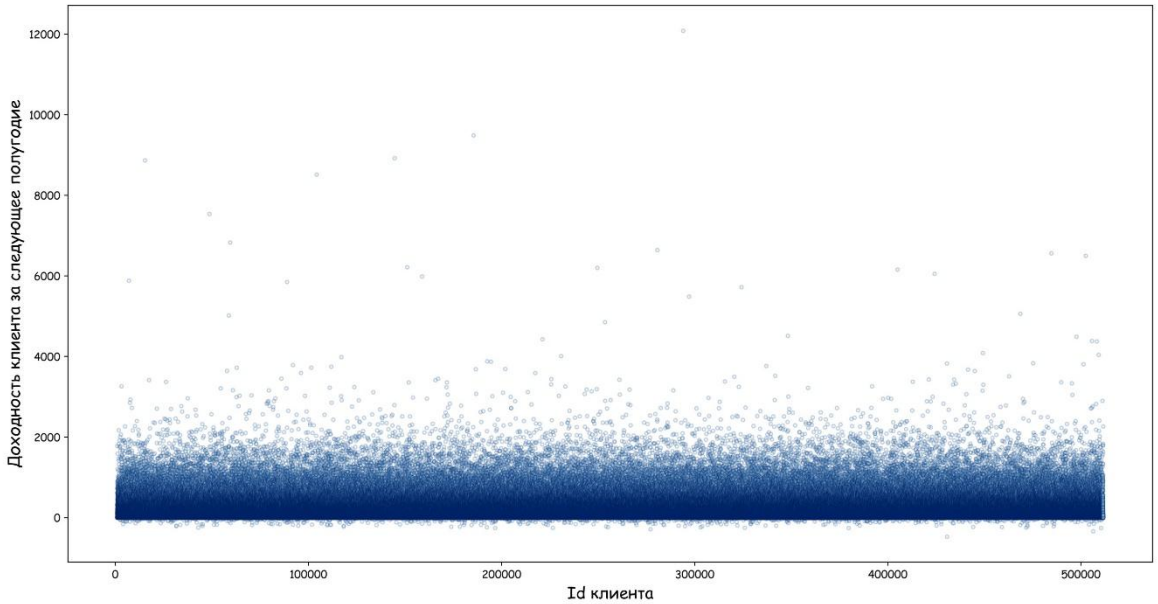
область работы клиента (cu_empl_area)	баланс кредитов (pl_balance)
уровень должности клиента (cu_empl_level)	баланс депозитов (td_volume)
баланс кредитных карт (cc_balance)	баланс счетов (ca_volume)
баланс автокредитов (cl_balance)	баланс накопительных счетов (sa_volume)
баланс ипотеки (ml_balance)	баланс инвестиций (mf_volume)

В результате построения диаграммы Кливленда было выявлено, что подавляющее большинство значений доходности лежит в диапазоне от 0 до 4000 единиц, однако присутствуют и значения более 8000. В рамках данной задачи мы не будем считать такие объекты выбросами

Целевая переменная

Поскольку для расчета метрики CLTV традиционным* способом данных недостаточно, показатель CLTV рассчитывался на основе признака gi_smooth_3m как сумма значений за второе полугодие

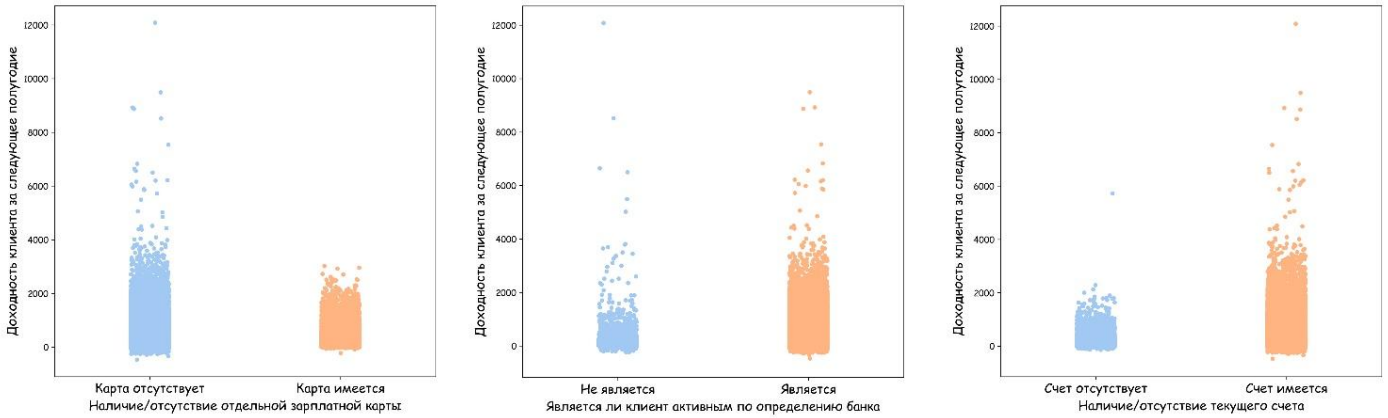
Диаграмма Кливленда для доходов от клиентов

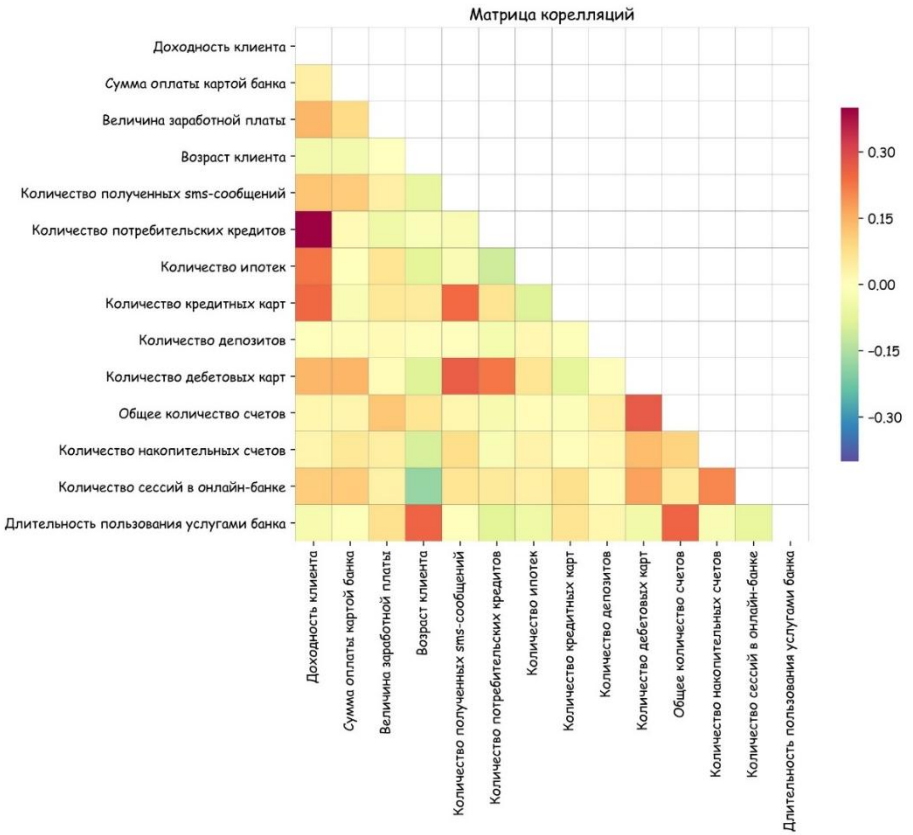


*Традиционно при расчете CLTV компании также используют значения маркетинговых затрат и чистой прибыли

Предложенные признаки

- Анализ категориальных признаков выявил, что пол и город клиента не разделяют целевую функцию в достаточной степени, однако немного лучше с этой задачей справляются следующие признаки:





- Был произведен синтез новых переменных, однако к существенному улучшению качества модели в последствии это не привело, поэтому в модель они включены не были

Для создания моделей были выбраны следующие признаки

Линейная регрессия		
	Категориальные	**Вещественные**
	active ca_f	salary gi_smooth_3m cur_quantity_pl

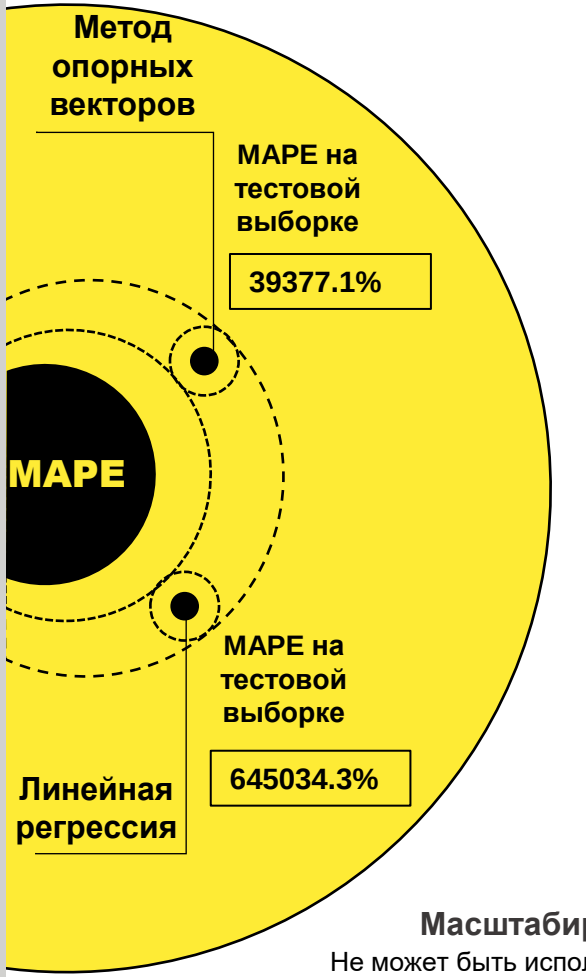
Метод опорных векторов		
	Категориальные	**Вещественные**
	active ca_f	1gi_smooth_3m 2gi_smooth_3m 3gi_smooth_3m 4gi_smooth_3m 5gi_smooth_3m 6gi_smooth_3m*

*Цифра перед признаком означает временную метку месяца

Модели

Линейная регрессия

Метод опорных векторов



Конфигурация

Использовалась LASSO-регрессия с параметром регуляризации альфа = 5

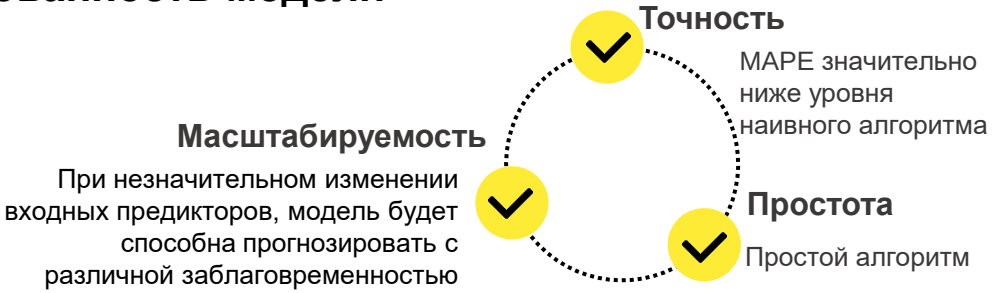
Использовалось радиальное базисное ядро, параметр C = 200

Качество модели

• Средняя абсолютная ошибка на тестовой выборке (MAE)	**93.8**
• Средняя медианная ошибка на тестовой выборке	**32.5**
• Корень из среднеквадратической ошибки на тестовой выборке (RMSE)	**156.1**

• Средняя абсолютная ошибка на тестовой выборке (MAE)	**61.6**
• Средняя медианная ошибка на тестовой выборке	**21.5**
• Корень из среднеквадратической ошибки на тестовой выборке (RMSE)	**137.9**

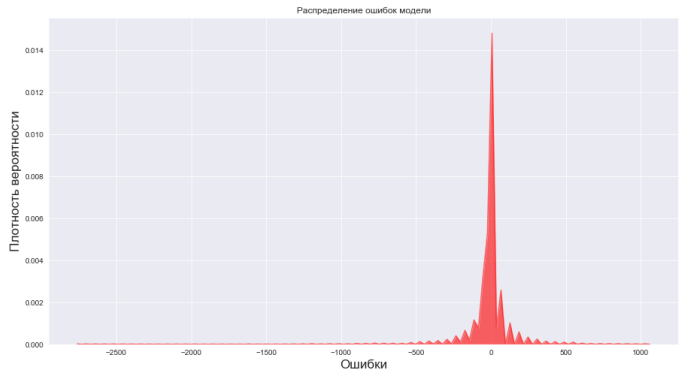
Сбалансированность модели



Результаты

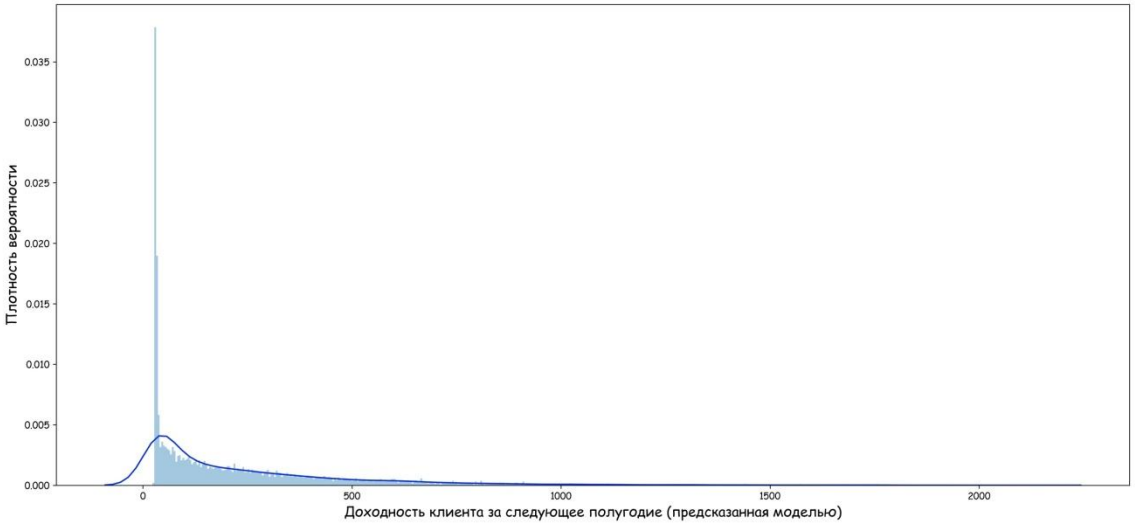
В качестве финальной модели был выбран метод опорных векторов

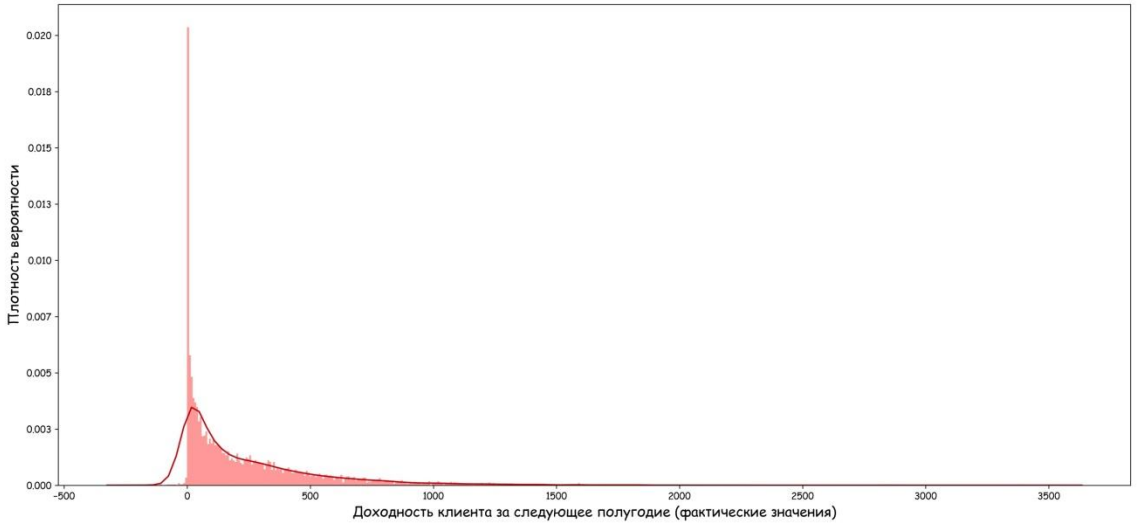
Распределение ошибок не смещено и симметрично, условие гомоскедастичности соблюдается при прогнозе на величину превышающую 500 единиц



В качестве основных предикторов в модели выступает временной ряд, составленный из среднемесячных значений доходности и при незначительном изменении длины временного ряда и гиперпараметров модели, модель будет способна принимать на вход векторы любой величины, и при необходимости предсказывать с любой заблаговременностью

Тестирование двух выборок (фактических и предсказанных моделью значений) на принадлежность одной генеральной совокупности с помощью критерия Колмогорова-Смирнова показало, что данные выборки можно считать «похожими», а значит, прогноз модели достаточно точен





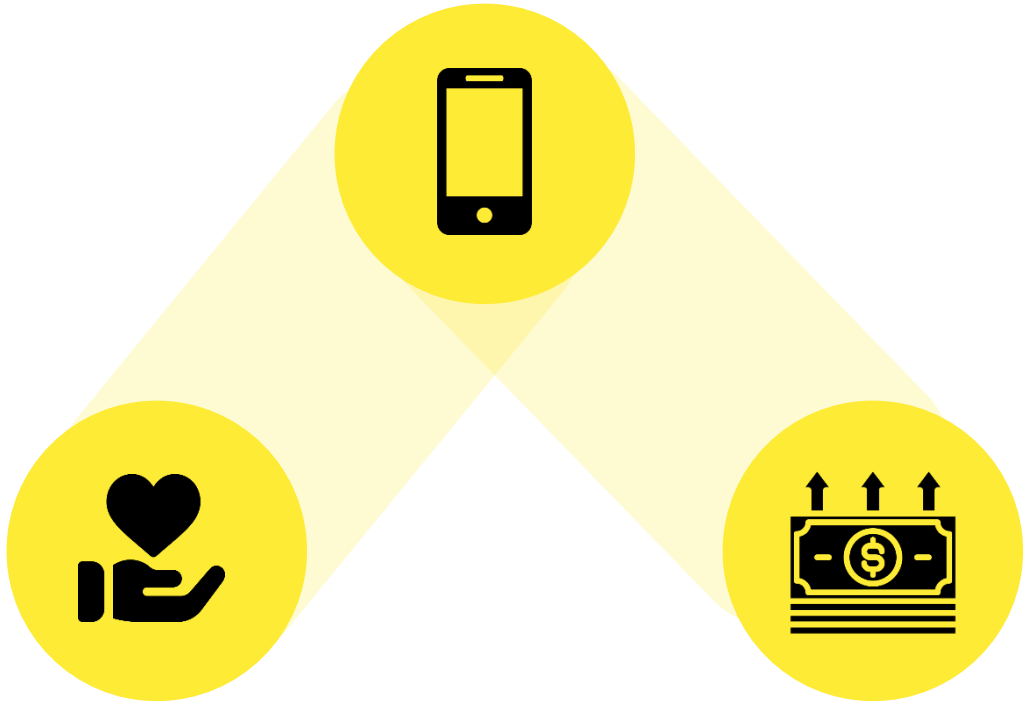
Использование данных и модели

Важнейшие факторы и области дальнейшего глубинного анализа

- Сегментация клиентов по модели использования продуктов банка («зарплатники», «вкладчики», «пользователи кредитных продуктов», «диджитал-ориентированные»)
- Сопоставление гипотез данной модели с моделью скоринга (например, влияние количества кредитных продуктов на скоринг и CLTV)
- Влияние наличия зарплатных карт на CLTV –> базис для стратегии развития партнерских отношений и зарплатных проектов
- Влияние использования онлайн-банкинга на CLTV –> базис для дальнейшего развития онлайн-приложения
- Оценка влияния смс-рассылок на CLTV: определение оптимального количества смс в месяц, оценка эффективности кампаний
- Выявление наиболее популярных продуктов банка + продуктов, наиболее распространенных среди самых лояльных потребителей

Дальнейшее использование созданной модели:

Определение и управление лояльностью клиентов



Настройка персональных предложений по СМС и push-уведомлениям

Повышение качества скоринговых процедур и предсказания платежеспособности клиента