Разработка модели по поиску потенциальной аудитории для программы «Клуба полезных привычек»

Команда YEET

Елисеева Екатерина
Тимонина Мария
Тен Су Бок
Щербакова Екатерина

X5Group

- **Executive summary**
- Анализ набора данных
- Baseline-модель
- Анализ ошибок
- Финальная модель
- Инициативы
- Эффект

X5Group

Разработанная и улучшенная look-alike модель позволит сократить затраты на привлечение новых членов Клуба Полезных Привычек

Задача

Для нахождения потенциальных членов Клуба Полезных Привычек необходимо **построить look-alike модель**

Для этого рассмотрим ряд задач:

- Exploratory data analysis
- Построение baseline-модели
- Тестирование методов повышения качества модели

Анализ ошибок

Baseline-модель логистической регрессии со SMOTE и масштабированием данных имела характерные ошибки прогноза

Такие как:

- Проблема классификации
- Проблема oversampling
- Проблема прогноза вероятностей
- Проблема отбора признаков

Улучшенная модель

Алгоритм применения модели

Подготовка данных

↓

Отбор признаков

↓

Oversampling SMOTE

↓

Undersampling

↓

PU Learning с базовой

Калибровка вероятностей

Логистическая регрессия

Улучшение позволило повысить качество прогноза на **5 п.п.**

Метрики качества

F2-score **42,84%**

Precision **16,69%**

Recall **70,42%**

Источник: анализ проектной команды.

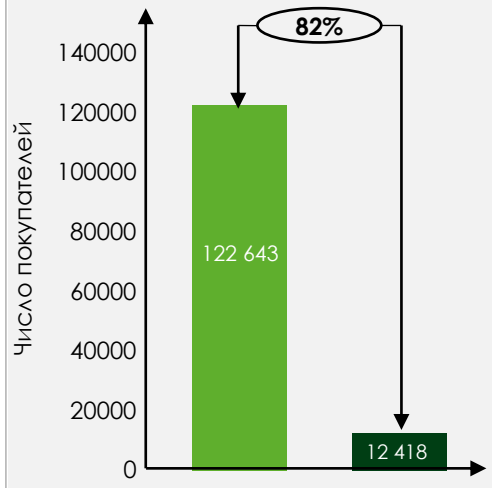
- Executive summary
- Анализ набора данных
- Baseline-модель
- Анализ ошибок
- Финальная модель
- Инициативы
- Эффект

X5Group

Исходя из анализа данных, мы выявили, что покупателей входящих в Клуб Полезных Привычек на 82% меньше чем покупателей не состоящих в клубе, а их товарооборот в среднем выше на 39%

1 Соотношение классов покупателей

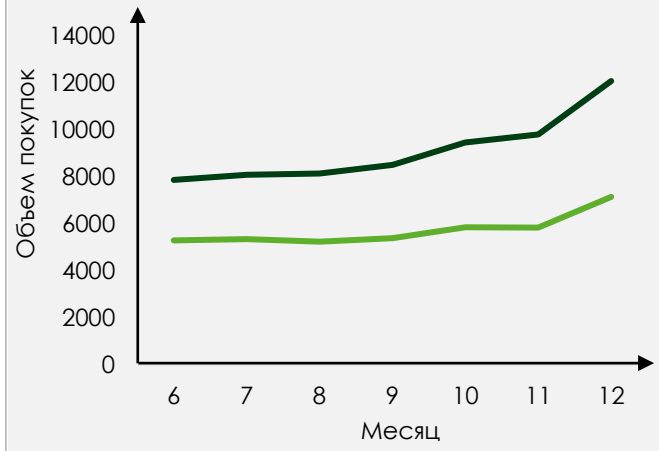
Соотношение покупателей в состоящих в программе*



Клиентов, не вступивших в клуб, **на 82% больше**, чем вступивших

В клубе
Не в клубе

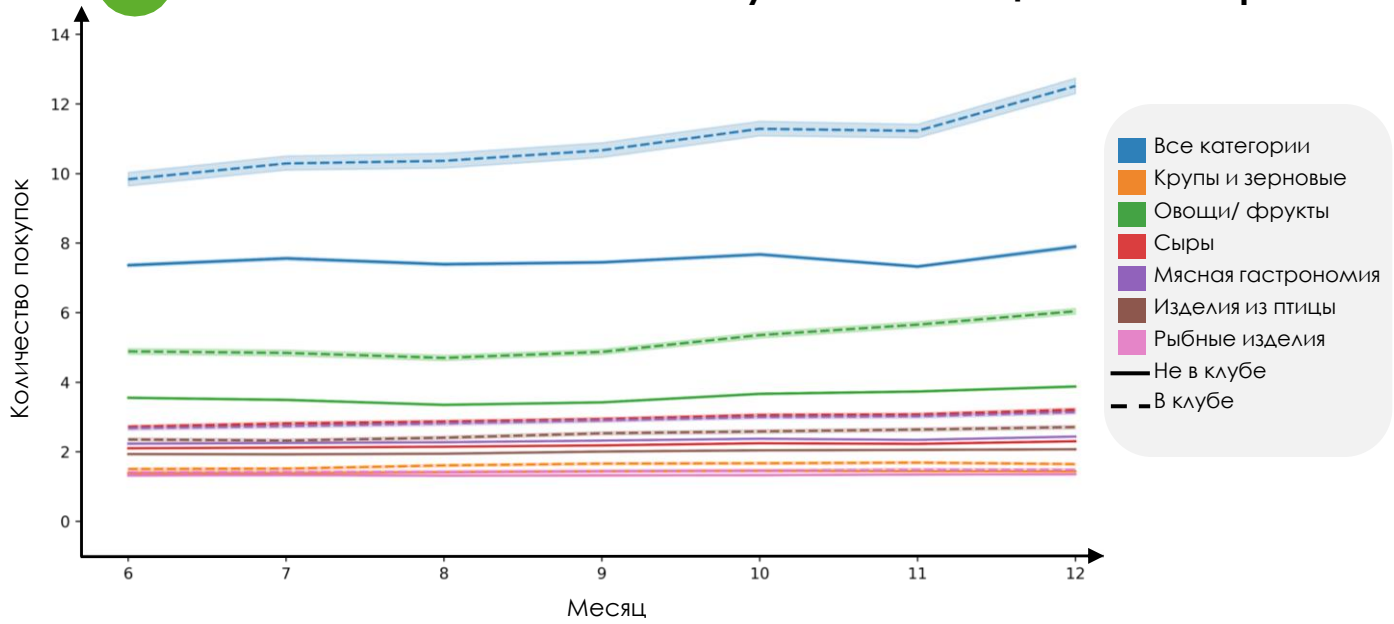
Средний товарооборот по месяцам



- У покупателей в клубе товарооборот **в среднем выше**, чем у покупателей вне клуба
- У обеих категорий покупателей **товарооборот увеличивается** к концу года

В клубе
Не в клубе

2 Зависимость количества покупок от месяца по категориям



Проанализировав данные о количествах покупок в разрезе категорий , мы выделили:

- **7 месяцев**: с июня по декабрь
  - самый прибыльный месяц: декабрь
  - самый неприбыльный месяц: июнь
- **6 категорий** товаров
  - самая популярная категория: **Овощи/Фрукты**
  - самая непопулярная категория: **Рыбные изделия**
- все признаки (кроме client_id и is_in_club) имеют **логнормальное распределение**

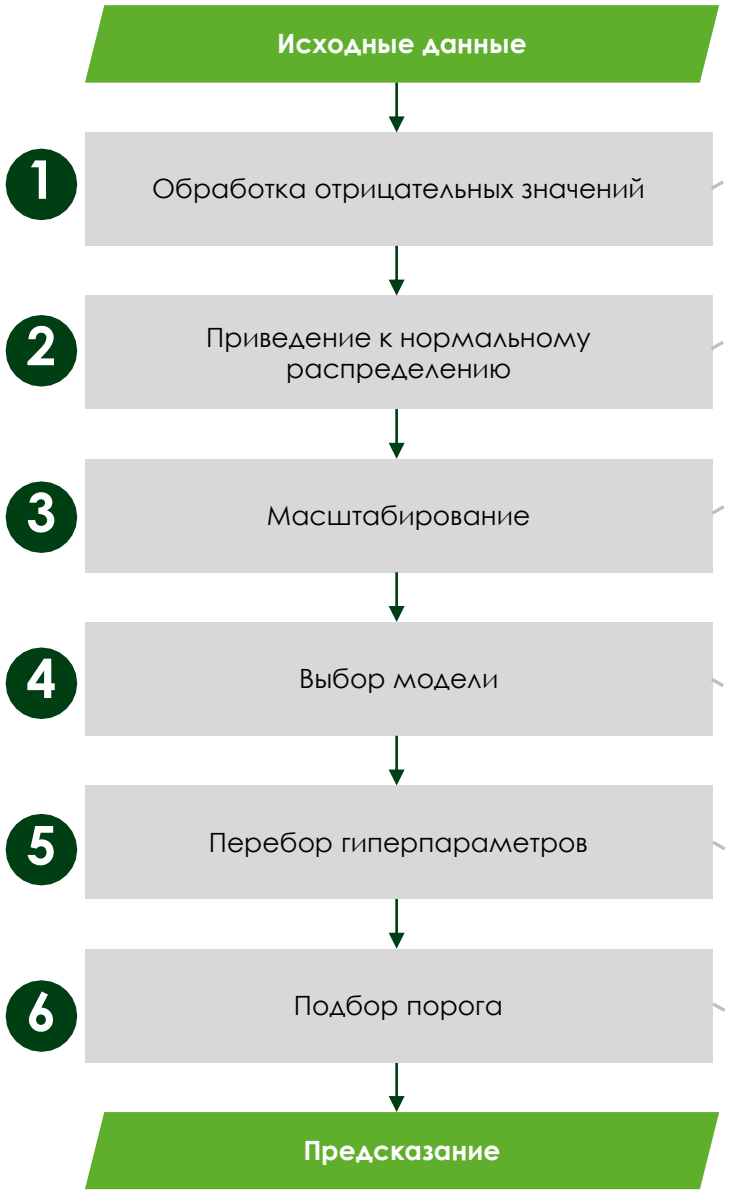
В результате проведенного анализа данных необходимо выбрать наилучшую модель и преобразовать Dataset для улучшения метрик прогнозирования

Источник: данные кейса, анализ проектной команды.

- Executive summary
- Анализ набора данных
- Baseline-модель
- Анализ ошибок
- Финальная модель
- Инициативы
- Эффект

X5Group

Лучшая модель - логистическая регрессия на прологарифмированных и отмасштабированных данных с использованием алгоритма oversampling' а SMOTE



Отрицательные значения std возводим в квадрат
Отрицательные значения товарооборота заменим на 0

Логарифмируем данные для логнормального распределения

Некоторые модели работают лучше и быстрее если данные находятся примерно в одном масштабе.

Лучшая модель выбиралась по наибольшему значению **метрики F2-score**.

Перебор гиперпараметров**:** алгоритма, метода и коэффициента регуляризации).

Позволяет **понять при каком значении вероятности мы будем считать клиента потенциальным членом клуба** и взаимодействовать с ним.

Метрики качества на тестовых данных:

F2-score
37,01%

Precision
14,06%

Recall
62,53%

Источник: анализ проектной команды, приложение – 1-2.

- Executive summary
- Анализ набора данных
- Baseline-модель
- Анализ ошибок
- Финальная модель
- Инициативы
- Эффект

X5Group

Улучшение baseline-модели проводится через настройку под специфику данных

Проблематика	Анализ ошибок	Гипотеза решения
Проблема классификации	Бинарная классификация **не позволяет учитывать потенциальных членов** Клуба Полезных Привычек	Использование алгоритма PU Learning
Проблема oversampling **Проблема прогноза вероятностей**	Количество объектов единичного класса возрастает **в 10 раз** в результате oversampling 'а, следовательно синтетическое добавление данных является некачественным Логистическая регрессия некорректно предсказывает вероятности	Применение Calibrated Classifier CV Использование одновременно undersampling и oversampling
Проблема отбора признаков	Широкий диапазон коэффициентов выбранной модели показывает наличие признаков незначительно влияющих на качество прогноза	Провести отбор параметров

Источник; анализ проектной команды, данные кейса.

Использование Positive-Unlabeled Learning позволит выделить группу потенциальных членов Клуба Полезных Привычек

Проблема

Бинарная классификация **не позволяет учитывать потенциальных членов** Клуба Полезных Привычек

Обоснование

Среди покупателей, не вступивших в клуб, есть потенциальные члены Клуба Полезных Привычек

Матрица ошибок

Прогнозирование

Реальность	+Positive	-Negative
+Positive	649	3 076
-Negative	1 942	34 852

Вывод: модель неплохо разделяет классы, но не умеет выделять потенциальных членов Клуба Полезных Привычек

Решение

Результат

X5Group Источник: анализ проектной команды.

Использование Positive-Unlabeled Learning позволит выделить группу потенциальных членов Клуба Полезных Привычек

Проблема

Бинарная классификация **не позволяет учитывать потенциальных членов** Клуба Полезных Привычек

Обоснование

Среди покупателей, не вступивших в клуб, есть потенциальные члены Клуба Полезных Привычек

Матрица ошибок

Прогнозирование

Реальность	+Positive	-Negative
+Positive	649	3 076
-Negative	1 942	34 852

Вывод: модель неплохо разделяет классы, но не умеет выделять потенциальных членов Клуба Полезных Привычек

Решение

PU Learning

1. Необходимо разделить объекты на позитивный класс и неразмеченные данные
2. Подбор параметра hold-out (размер отложенной выборки)
3. Подбор параметра порога

Результат

Источник: анализ проектной команды.

Использование Positive-Unlabeled Learning позволит выделить группу потенциальных членов Клуба Полезных Привычек

Проблема

Бинарная классификация **не позволяет учитывать потенциальных членов** Клуба Полезных Привычек

Обоснование

Среди покупателей, не вступивших в клуб, есть потенциальные члены Клуба Полезных Привычек

Матрица ошибок

Реальность \ Прогнозирование	+Positive	-Negative
+Positive	649	3 076
-Negative	1 942	34 852

Вывод: модель неплохо разделяет классы, но не умеет выделять потенциальных членов Клуба Полезных Привычек

Решение

PU Learning

1. Необходимо разделить объекты на позитивный класс и неразмеченные данные
2. Подбор параметра hold-out (размер отложенной выборки)
3. Подбор параметра порога

Результат

Ошибка PU Learning модели на объектах нулевого класса встречались **в ≈12 раз чаще**, чем на объектах единичного класса.

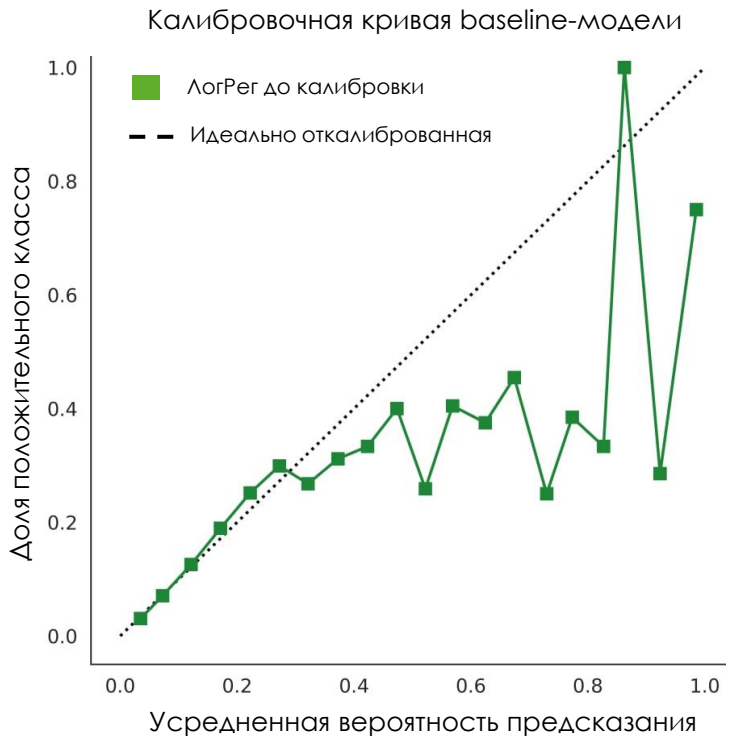
Количество ошибок модели, объекты

Источник: анализ проектной команды.

Базовая модель в виде логистической регрессии стала более корректно предсказывать вероятности

Проблема	Решение	Результат

1. Дисбаланс классов
2. В модели логистической регрессии распределение вероятностей смещено

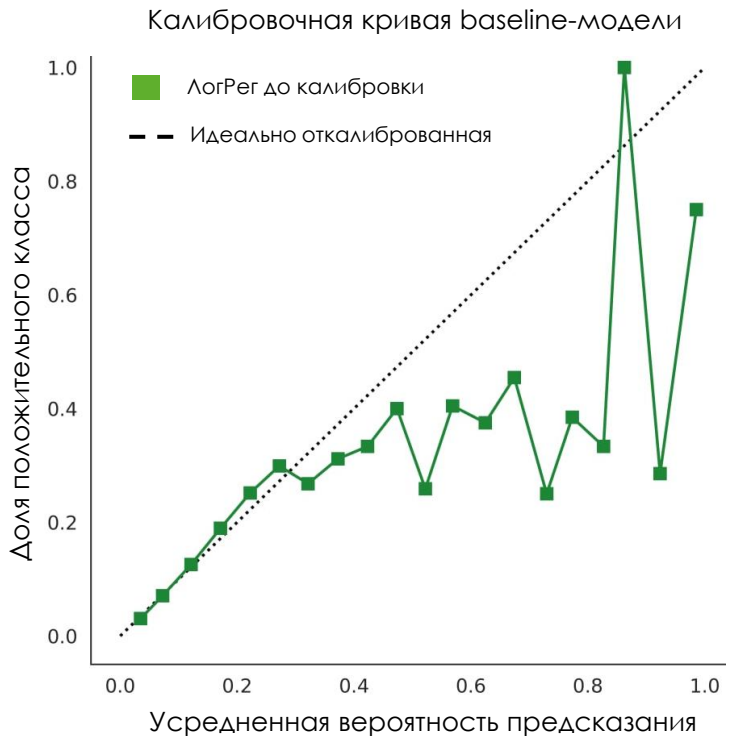


Источник: анализ проектной команды.

Базовая модель в виде логистической регрессии стала более корректно предсказывать вероятности

Проблема

1. Дисбаланс классов
2. В модели логистической регрессии распределение вероятностей смещено



Решение

SMOTE Oversampling
Random Undersampling

В улучшенной модели в обучающей выборке объектов каждого класса одинаковые доли

CalibratedClassifierCV
Калибровка вероятностей

Модель дообучается, корректируя параметры исходной модели для точного предсказания вероятности

Результат

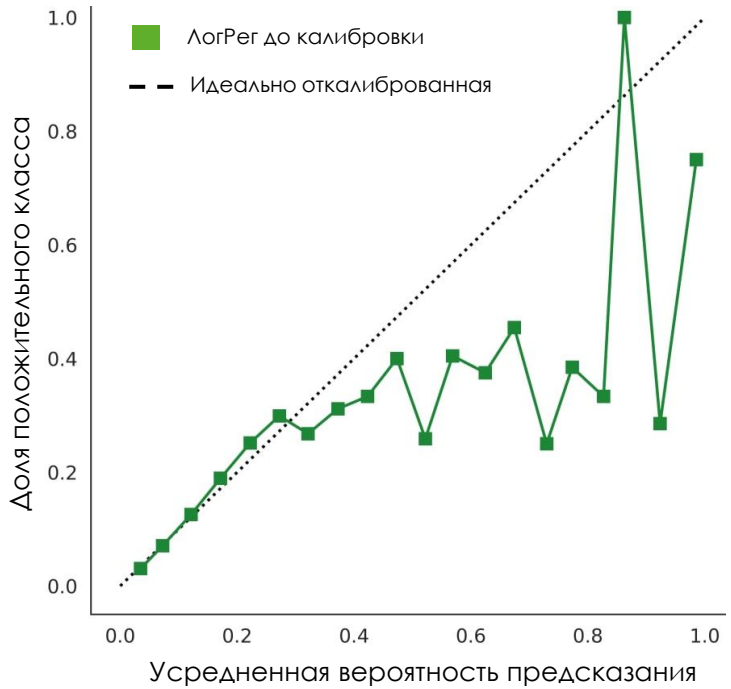
Источник: анализ проектной команды.

Базовая модель в виде логистической регрессии стала более корректно предсказывать вероятности

Проблема

1. Дисбаланс классов
2. В модели логистической регрессии распределение вероятностей смещено

Калибровочная кривая baseline-модели



Решение

SMOTE Oversampling
Random Undersampling

В улучшенной модели в обучающей выборке объектов каждого класса одинаковые доли

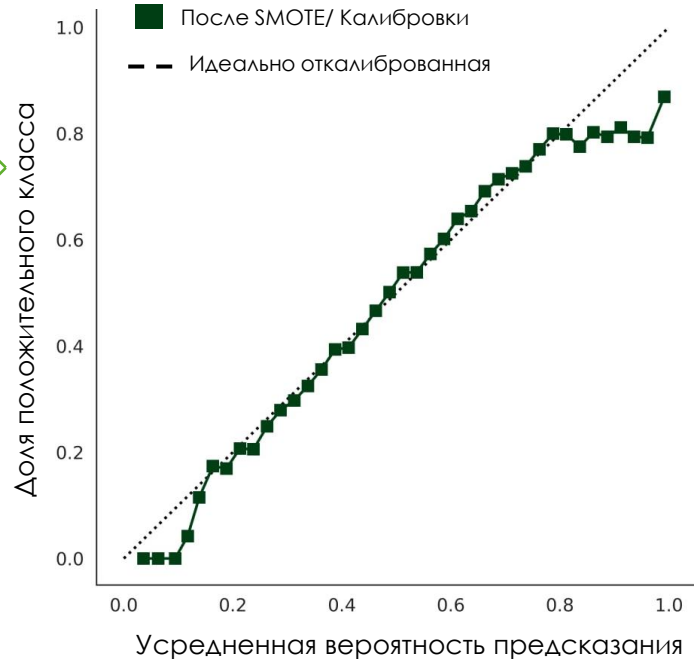
CalibratedClassifierCV
Калибровка вероятностей

Модель дообучается, корректируя параметры исходной модели для точного предсказания вероятности

Результат

Откалиброванная модель на масштабированной выборке проводит более точную классификацию

Калибровочная кривая улучшенной модели

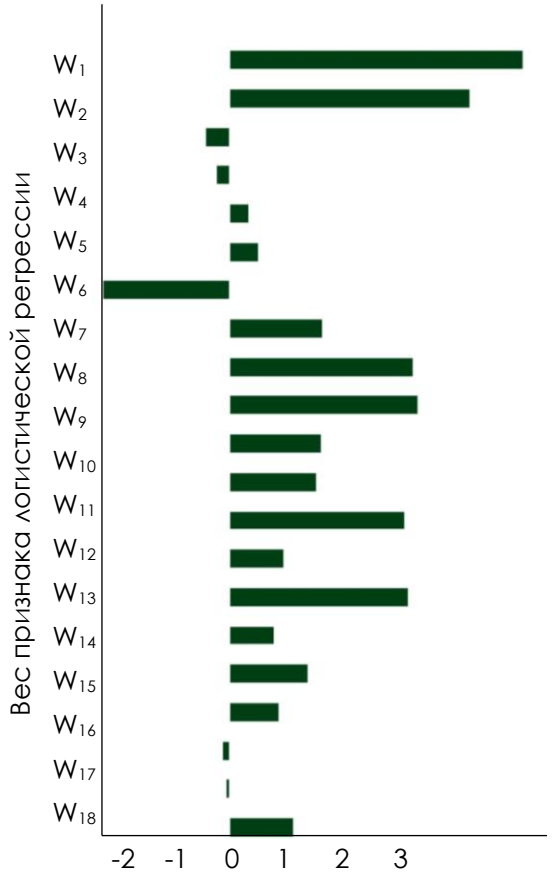


Источник: анализ проектной команды.

Отбор признаков повышает качество метрики на 1,4%

Проблема

Большое количество малозначимых признаков приводит к переобучению модели



Решение

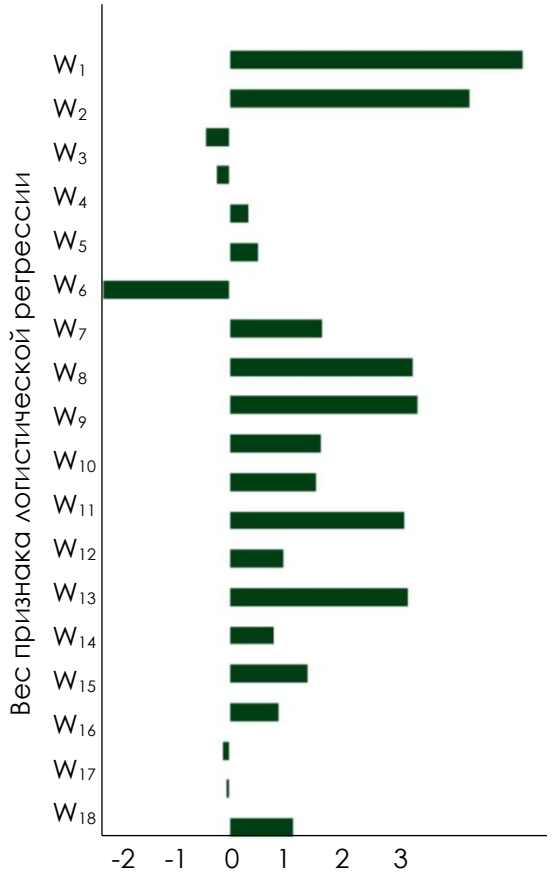
Результат

Источник: анализ проектной команды.

Отбор признаков повышает качество метрики на 1,4%

Проблема

Большое количество малозначимых признаков приводит к переобучению модели



Решение

Отбор признаков

1. Интуитивный подбор
2. Исследование коэффициентов модели
3. Отбор по статистическому параметру
4. Использование «жадного» алгоритма

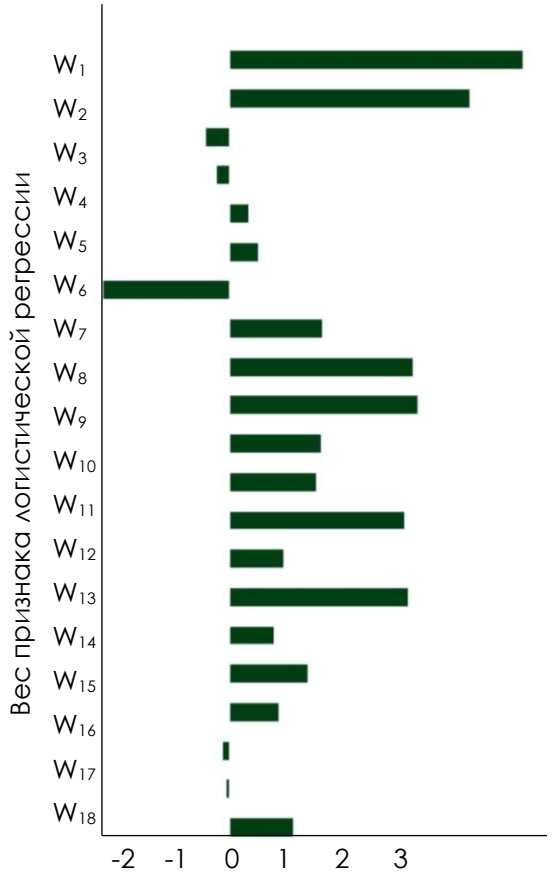
Результат

Источник: анализ проектной команды.

Отбор признаков повышает качество метрики на 1,4%

Проблема

Большое количество малозначимых признаков приводит к переобучению модели



Решение

Отбор признаков

1. Интуитивный подбор
2. Исследование коэффициентов модели
3. Отбор по статистическому параметру
4. Использование «жадного» алгоритма

Результат

Лучшее значение метрики дал метод отбора признаков по t-статистике

После отбора наиболее значимых признаков качество на тестовой выборке возросло

F2-Score

38,7% → 40,1%

Источник: анализ проектной команды.

- Executive summary
- Анализ набора данных
- Baseline-модель
- Анализ ошибок
- Финальная модель
- Инициативы
- Эффект

X5Group

В результате улучшения модели значение метрики возросло более чем на 5 п.п.

Выбранная модель до улучшения

Исходные данные

1. Подготовка данных
2. Oversampling SMOTE
3. Логистическая регрессия

Предсказание

Метрики качества до улучшения:

F2-score **37,01%**

Precision **14,06%**

Recall **62,53%**

Выбранная модель после улучшения

Исходные данные

1. Подготовка данных
2. Отбор признаков
3. Oversampling SMOTE
4. Undersampling
5. PU Learning с базовой моделью
   - Калибровка вероятностей
     - Логистическая регрессия

Предсказание

Метрики качества после улучшения:

F2-score **42,84%**

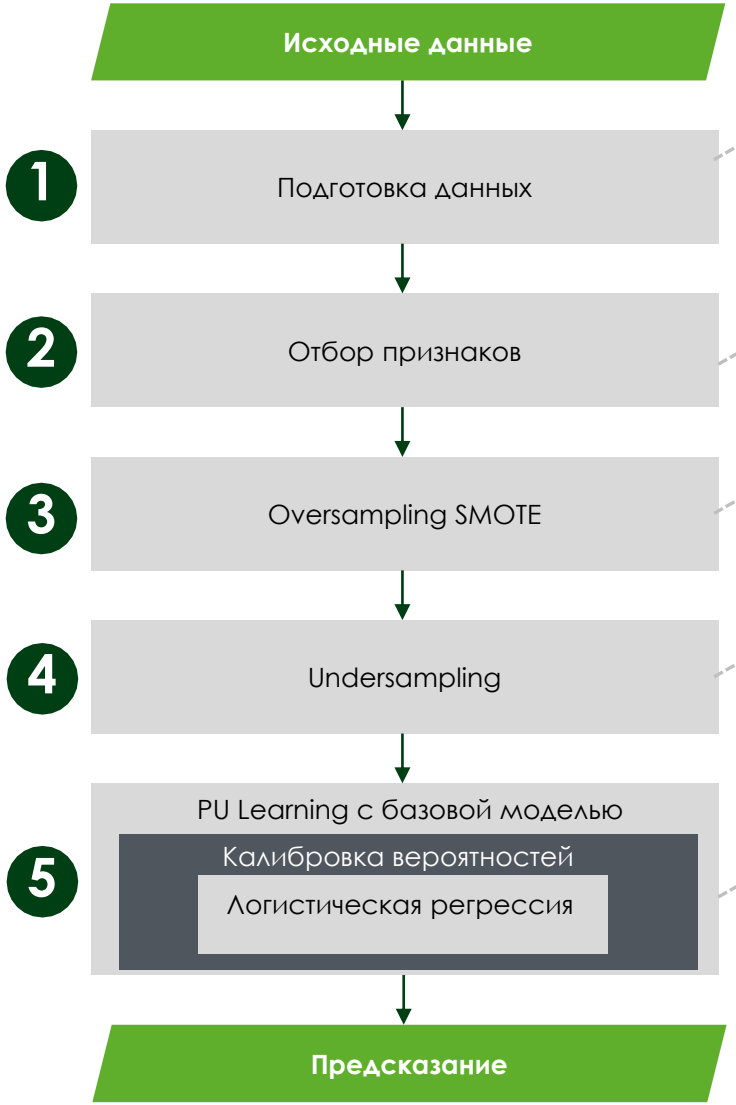
Precision **16,69%**

Recall **70,42%**

Источник: анализ проектной команды.

Улучшенная нами модель учитывает природу данных

Улучшенная модель



Описание

- заполнение пропусков;
- обработка отрицательных значений;
- логарифмирование отдельных признаков;
- масштабирование данных.

- добавление новых признаков;
- подсчет t-статистики;
- поиск оптимального количества признаков.

- генерация объектов минорного класса.

- уменьшение размера мажоритарного класса.

- базовая модель: логистическая регрессия;
- калибровка вероятностей позволяет настроить модель на корректное оценивание вероятностей;
- модель PU Learning: именно этот этап позволяет учитывать природу данных.

Источник: анализ проектной команды.

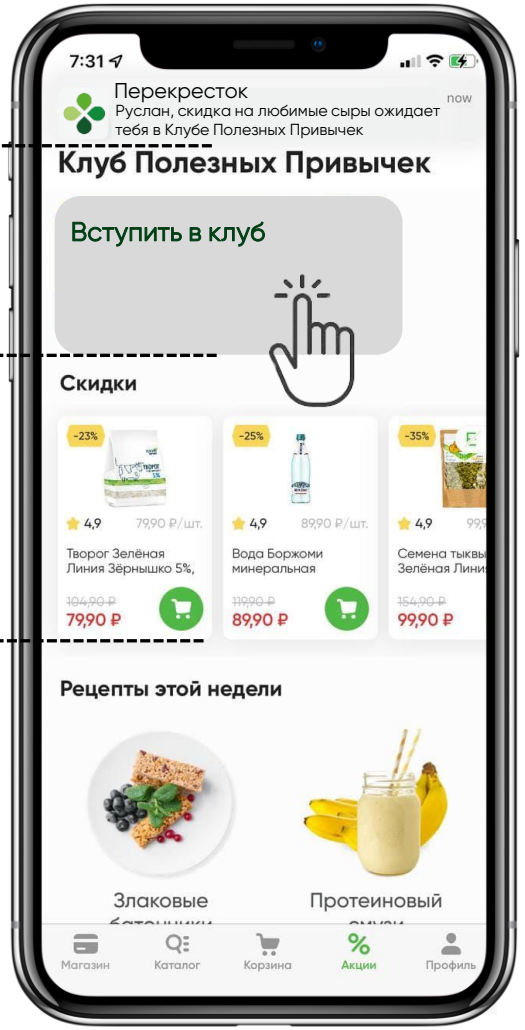
- Executive summary
- Анализ набора данных
- Baseline-модель
- Анализ ошибок
- Финальная модель
- Инициативы
- Эффект

X5Group

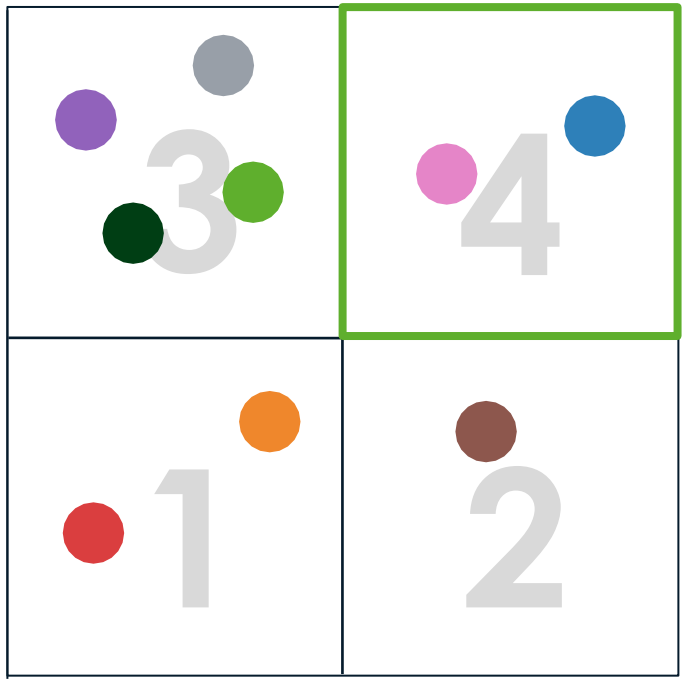
Для привлечения и удержание потенциальных членов Клуба Полезных Привычек необходимо использовать Push-уведомление и механики для дальнейшего взаимодействия

Механика взаимодействия с выделенной аудиторией происходит через Push-уведомления

1. Push-уведомление приходит на телефон клиента
2. Клиент принимает приглашение в Клуб Полезных Привычек
3. Клиент получает скидку на любимую категорию в Клубе Полезных Привычек



Для продолжения взаимодействия с клиентами «Клуба полезных привычек», мы выделили основные инициативы:



- Проведение опросов по удовлетворенности
- Розыгрыши сертификатов по ЗОЖ
- Получение доп. баллов за вступление в клуб/покупку товаров ЗОЖ
- Проведение закрытых дегустаций
- Персонализированные рекомендации продуктов
- Проведение семинаров по ЗОЖ
- Проведение мастер-классов «здоровых блюд»
- Геймификация
- Благотворительные акции

Источник: анализ проектной команды, моб. приложение «Перекресток», приложение – 3.

- Executive summary
- Анализ набора данных
- Baseline-модель
- Анализ ошибок
- Финальная модель
- Инициативы
- Эффект

X5Group

Эффект

Внедрение модели логистической регрессии позволяет снизить CAC на 45% и улучшить показатели эффективности маркетинговой системы в 2 раза

1 Эффективность таргетинговой системы

Соотношение эффективности моделей

Результат:

Для получения X числа конверсий до внедрения модели необходимо было охватить 10X потенциальных клиентов. Для достижения такого же результата числа конверсий **необходимый охват будет составлять 5X** потенциальных клиентов при прочих равных

2 Customer acquisition cost*

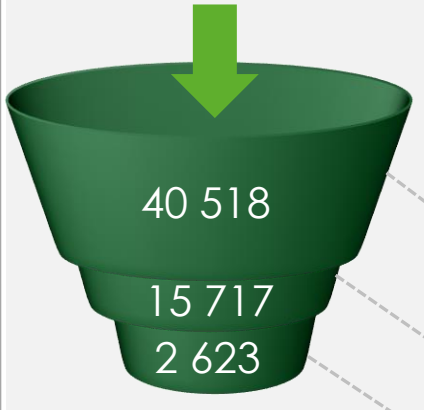
Маркетинговые расходы на человека = x

Число привлеченных клиентов = y

Точность модели = **0,167**

	Итоговые расходы на маркетинг	CAC*
После внедрения модели	74 360x	74 360x/y
До внедрения модели	135 061x	135 061x/y

После использования модели показатель CAC* понизился **на 49%**



Количество пользователей приложения

Количество пользователи которым направлены Push-уведомления

Ожидаемый приток клиентов в Клуб Полезных Привычек

3 Метрики:

F2-score **42,84%**

Precision **16,69%**

Recall **70,42%**

X5Group Источник: данные кейса, анализ проектной команды.

*CAC – Стоимость привлечения клиента

YEET

Елисеева Екатерина
Бакалавриат, НИУ ВШЭ, Факультет Высшая Школа Бизнеса ‘23

Тимонина Мария
Бакалавриат, НИУ ВШЭ, Факультет Компьютерных наук ‘23

Тен Су Бок
Магистратура, РЭУ им Г.В. Плеханова, Факультет Менеджмента ‘23

Щербакова Екатерина
Бакалавриат, НИУ ВШЭ, Факультет Высшая Школа Бизнеса ‘23

Приложения


567 баллов
В 60 раз больше баллов за покупку товаров L'Oreal и Garnier
Все
Купоны
x5 Баллов за кулинар
Все
Партнеры
Потратить баллы
Получить баллы
Семейный клуб
Карта
Магазин

Приложение 1. Логистическая регрессия на прологарифмированных и отмасштабированных данных с использованием алгоритма оверсэмплинга SMOTE

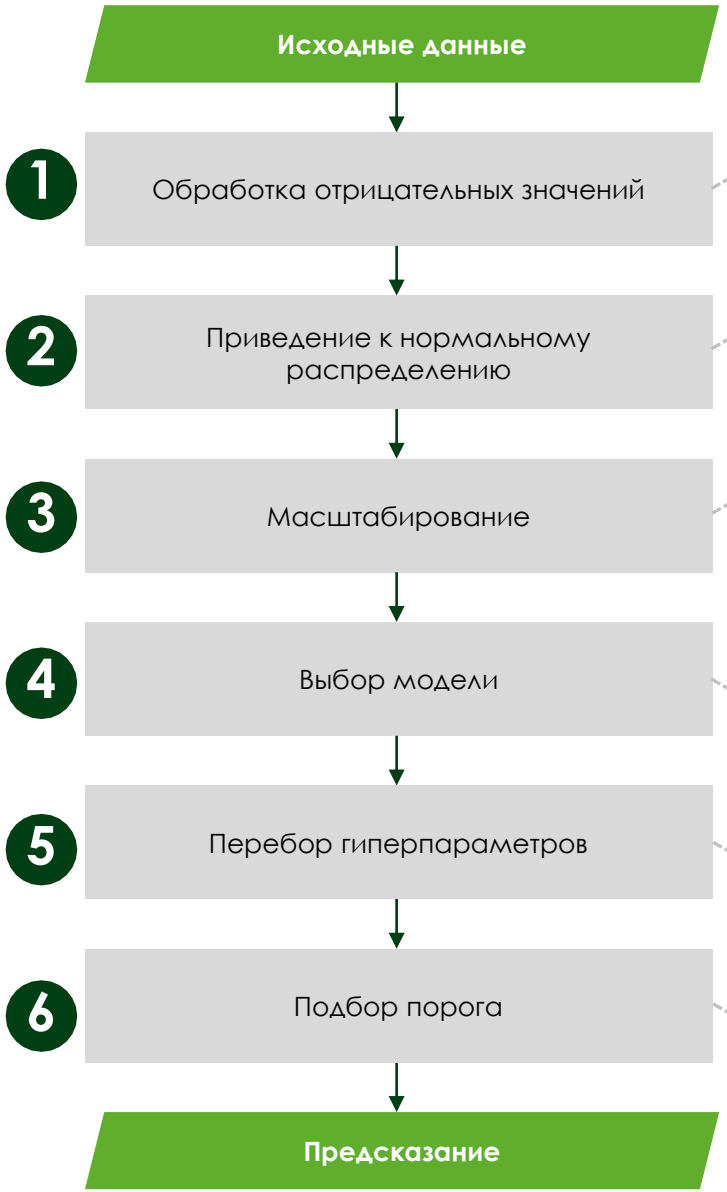
*Значения метрик были получены в результате кросс-валидации

Model Class	Model extension	F2-score	Precision	Recall
	No extension	5.98%	30.99%	5.00%
Logistic Regression	**SMOTE**	**41.94%**	**18.73%**	**61.33%**
	PU Classifier	14.01%	7.53%	99.80%
Random Forest	No extension	0.93%	39.99%	0.75%
	SMOTE	39.61%	20.67%	52.35%
	PU Classifier	36.93%	12.04%	76.43%
XGBoost	No extension	0.89%	55.90%	0.71%
	SMOTE	0.73%	51.02%	0.59%

Учет факторов при выборе модели:

- F-мера позволяет учитывать **Precision и Recall;**
- Коэффициент β = 2, показывает что Recall для нас важнее, так как лучше отправить несколько лишних уведомлений чем не отправить уведомление заинтересованному клиенту.

Приложение 2. Модель логистическая регрессия



Отрицательные данные при логарифмировании уйдут в бесконечность, модели не смогут с этим работать: необходимо их обработать. Среднеквадратическое отклонение по определению **должно быть неотрицательным**, но в наших данных есть отрицательные значения std, чтобы избавиться от них посчитаем дисперсию. Отрицательных значений товарооборота - их немного, поэтому заменим их на 0.

Некоторые модели работают лучше на нормально распределённых данных. Наши данные не соответствуют нормальному распределению. Но мы можем частично привести их к нему, например, **с помощью логарифмирования.**

Некоторые модели работают лучше и быстрее если данные находятся примерно в одном масштабе.

Лучшая модель выбиралась по наибольшему значению **метрики F2-score**. Наилучшее качество показала **логистическая регрессия** на прологарифмированных и отмасштабированных данных.

Для выбранной модели был **произведён перебор гиперпараметров:** алгоритма, метода и коэффициента регуляризации).

Выходом логистической регрессии является вероятность того, что данный клиент является членом клуба. Выбор порога позволяет **понять при каком значении вероятности мы будем считать клиента потенциальным членом клуба** и взаимодействовать с ним.

Метрики качества на тестовых данных:

F2-score
37,01%

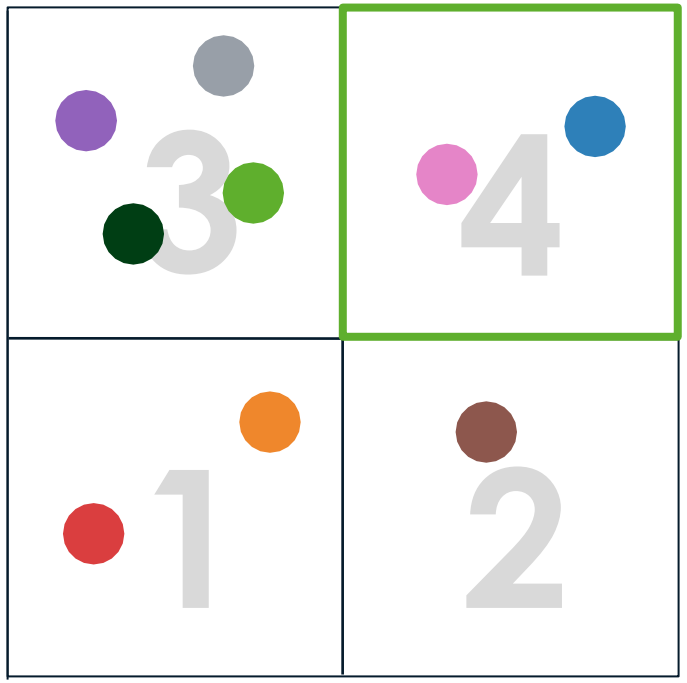
Precision
14,06%

Recall
62,53%

Источник: анализ проектной команды.

Приложение 3. Критерии для пиритизации инициатив

Для продолжения взаимодействия с клиентами «Клуба полезных привычек», мы выделили основные инициативы:



Критерии влияния на лояльность клиента	Балл
Выступает в качестве дополнения к основной стратегии, не дает финансовых стимулов, не сильно влияет на преданность к бренду, улучшает осведомленность	1
Увеличивает лояльность к бренду, не дает финансовых стимулов, не увеличивает уровень персонализации	2
Увеличивает лояльность к бренду, не дает финансовые стимулы, увеличивает уровень персонализации или дает финансовые стимулы, но не увеличивает уровень персонализации	3
Увеличивает лояльность к бренду, дает финансовые стимулы, увеличивает уровень персонализации	4

- Проведение опросов по удовлетворенности
- Розыгрыши сертификатов по ЗОЖ
- Получение доп. баллов за вступление в клуб/покупку товаров ЗОЖ
- Проведение закрытых дегустаций
- Персонализированные рекомендации продуктов
- Проведение семинаров по ЗОЖ
- Проведение мастер-классов «здоровых блюд»
- Геймификация
- Благотворительные акции

Источник: анализ проектной команды, моб. приложение «Перекресток».

Приложение 4. Расчет эффекта CAC

Data	Value
Cost of Sales & Marketing per person	x
Customers acquired	y
True positive	2623
Total customers	135061
Customers selected by model	15717
Club joiners	12418

Воронка для тестовой выборки	
Пользователи приложения	40518
Пользователи, получившие Push-уведомление	15717
Ожидаемый приток пользователей в Клуб Полезных Привычек	2623

CAC

= Total cost of Sales & Marketing / Customers acquired

Effectiveness of targeting system

=Conversion/Reach

	Total cost of Sales & Marketing	CAC	Decrease by
Before model implementation	135061x	135061x/y	45%
After model implementation	74360x	74360x/y	

	Effectiveness
Before model implementation	9,19%
After model implementation	16,7%