LOGINOM
ХАКАТОН 2020

Секция «Оригинальный проект»

Тюменский государственный университет

- Институт математики и программных наук, кафедра программной и системной инженерии
- Направления подготовки: Прикладная информатика, Информационные системы и технологии
- Участник академической программы с 2017 г.
- Команда ТюмГУ заняла I место в секции «Логистика и управление запасами» в Loginom Хакатон-2019
- Руководитель – доцент Цыганова М.С.

Команда

Студенты направления «Прикладная информатика»

Тропин Алексей

Построение ML-модели

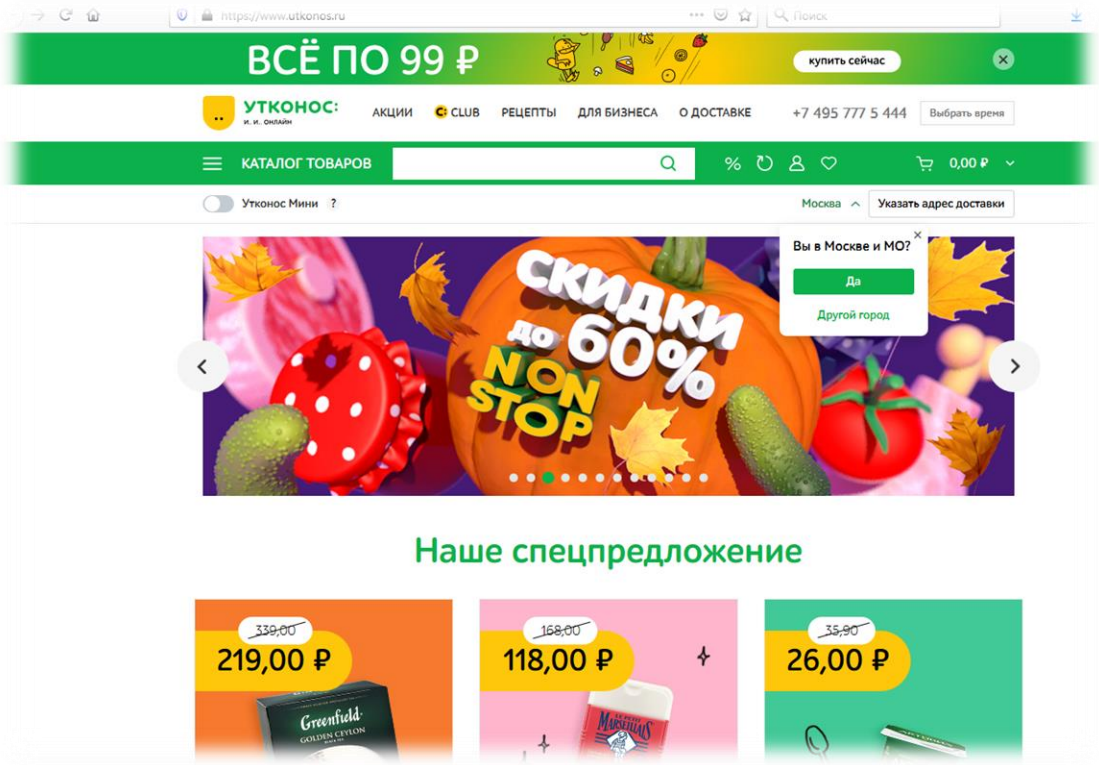
Алексеев Роман

Романов Илья

Анализ и предобработка исходных данных

Задача

- Прогнозирование вероятности отказа клиента от заказа для интернет-магазина Utkonos.ru.
- Задача с открытого Хакатона SAS Data Hack Platypus (осень 2019).
- Метрика качества- ROC-AUC.



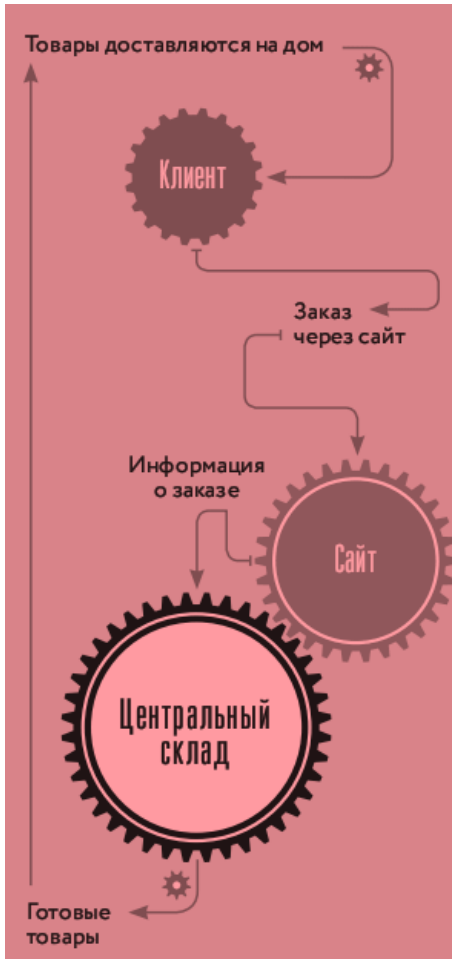
УТКОНОС:
И.. И.. ONLINE

Интернет-гипермаркет «Утконос»

- Сегмент E-Grocery – товары повседневного спроса через Интернет.
- Год основания – 2000.
- Оборот – 10 млрд. руб. (2019).

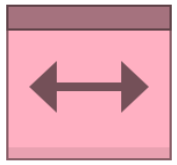
Отказ клиента от заказа во время доставки – актуальная проблема для каждой компании в электронной торговле:

- убытки логистики;
- убытки из-за скоропортящихся товаров;
- прочие издержки.



Обучающая выборка

Транзакций:
9 млн.



Период:
30/12/2017–
31/12/2018

Заказов:
354,8 тыс.

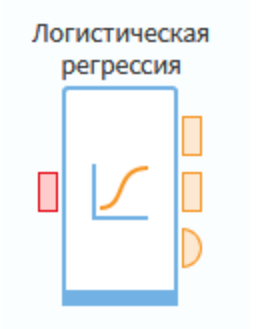
Клиентов:
31,2 тыс.

Целевое поле Флаг отмены заказа для тестовой выборки участникам конкурса доступно не было, поэтому мы сформировали ее из обучающей (пропорция 70/30)

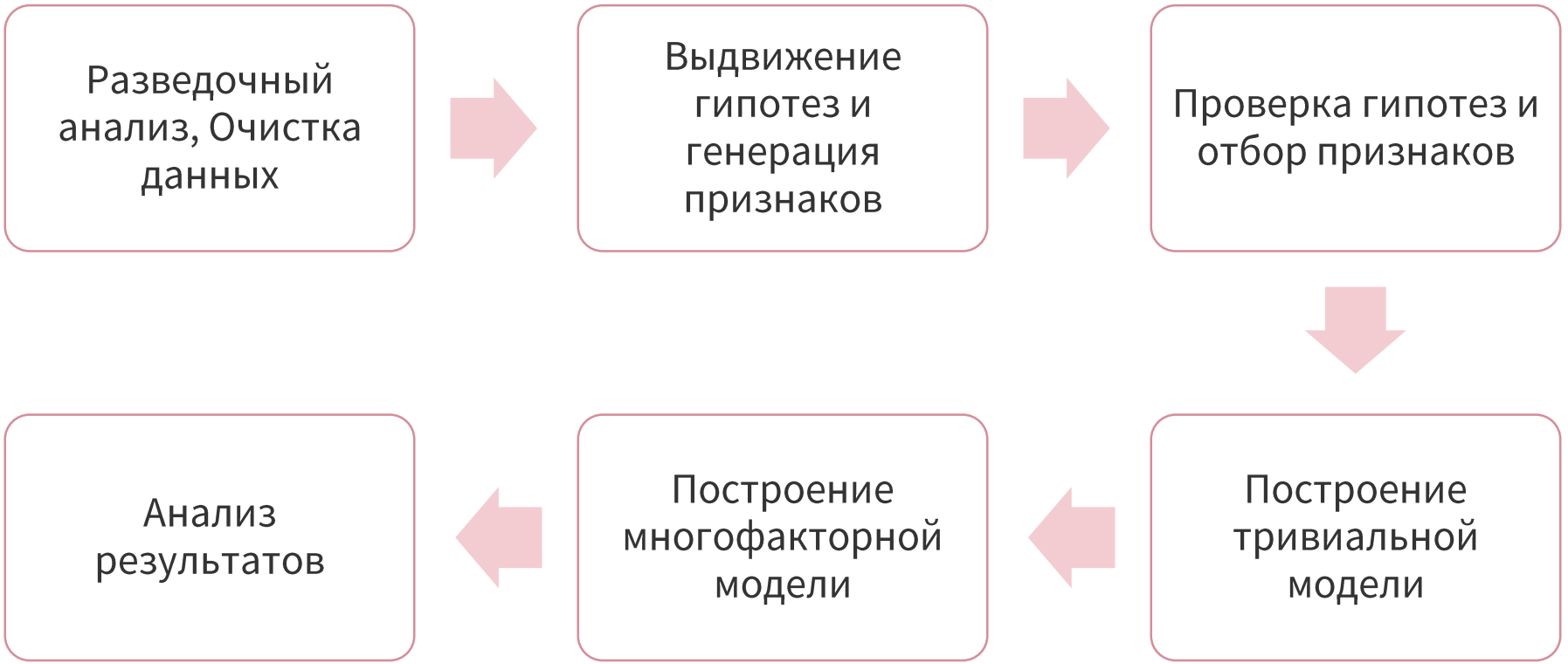
- Дата заказа, дата доставки
- ID клиента
- ID канала
- ID заказа
- ID товара
- ID товарной группы
- Тип доставки
- Кластер доставки
- Временной интервал
- **Флаг отмены заказа**
- Кол-во в заказе
- Предоплата
- Кол-во редактирований

Компоненты Loginom

- Компоненты для ETL-операций
- **Заполнение пропусков** и **Редактирование выбросов**
- **Конечные классы** и **Логистическая регрессия**
- **JavaScript** и **Python**



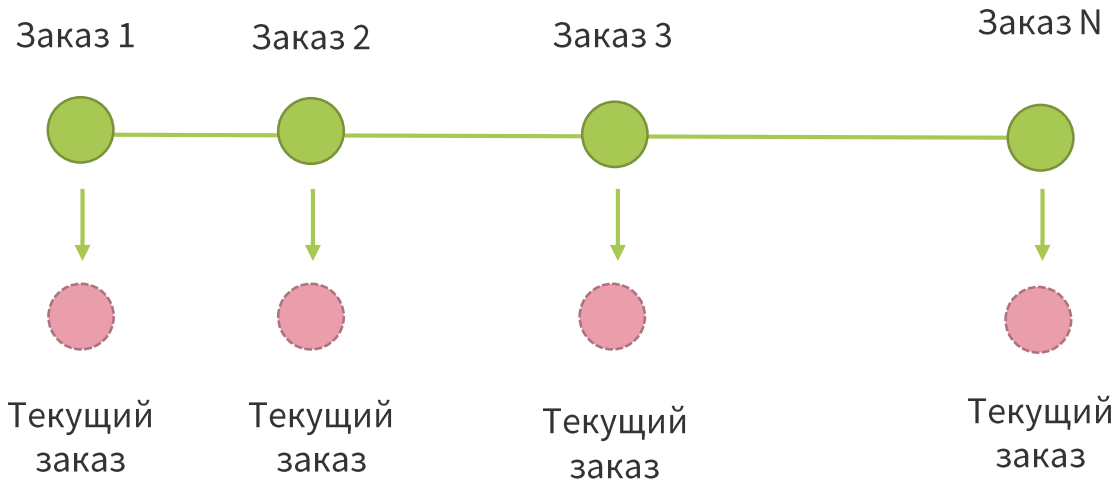
Этапы решения задачи



Формирование выборки для обучения модели

Признаки из текущего заказа

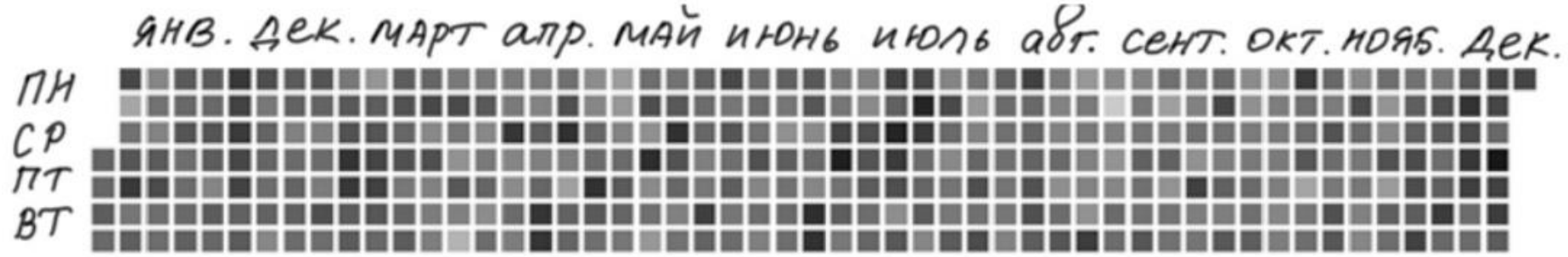
Признаки из истории прошлых заказов



- В выборку попадет N записей по количеству заказов.
- При расчетах признаков нельзя заглядывать «в будущее».

Разведочный анализ: поля **Дата заказа**

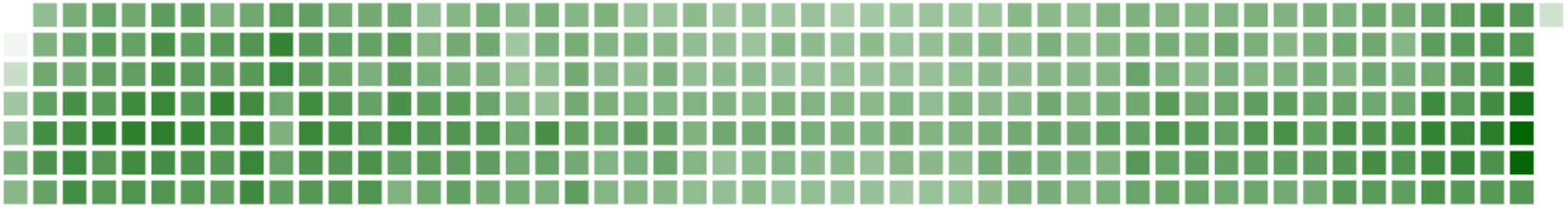
Тепловая карта по значениям «Число заказов в определенный день»



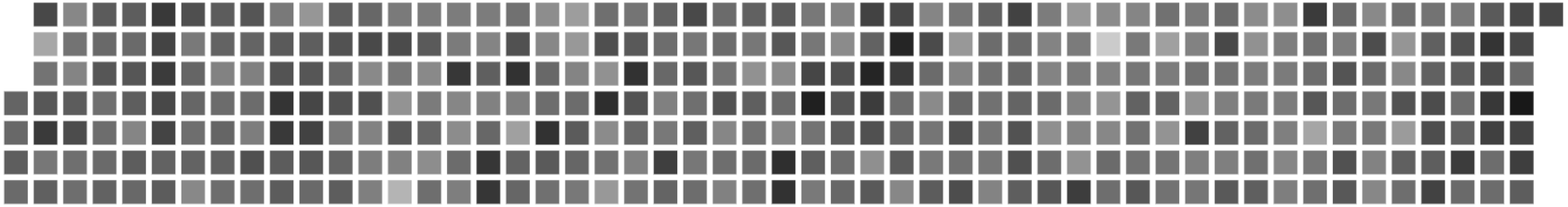
В определенные дни продажи резко растут или падают. Дополнительным исследованием были выявлены причины флуктуаций: в эти дни проходили масштабные мероприятия (ЧМ по футболу) и небольшие праздники.

Разведочный анализ: поля **Дата заказа**, **Флаг отмены**

Тепловая карта по значениям «Процент отмен в определенный день»



Тепловая карта по значениям «Количество отмен в определенный день»



Разведочный анализ: поля **Дата заказа**, **Дата доставки**

Процент отмен	Интервал между датами
100	20
100	24
100	23
87,5	18
75	19
75	21
53,85	9
50	16
46,15	14
40	12
40	17
37,5	22
35,29	15
35,29	13
32,58	6
28,57	11
20	7
18,13	5
17,42	4
12,5	10
9,47	8
9,37	2
8,56	3
4,87	1
4,72	0

- Чем меньше разница между **Дата заказа** и **Дата отмены**, тем меньше частота отмены заказов.
- \+ признак **Дней между заказом и доставкой**.

Разведочный анализ: поле **Интервал**

Процент отмен	Интервал
67,75	**10-18.**
28,1	**8-18.**
26,54	**9-16.**
21,16	**16-22.**
11,89	**14-17.**
10,9	22-2.
10,04	**0-2.**
9,72	8-13.
9,64	**13-18.**
8,34	6-14.
8,04	22-23.
7,67	22-0.
7,41	19-23.
6,42	10-16.
6,19	14-16.
5,72	6-8.
5,7	20-22.
5,68	10-14.
5,17	18-20.
5,17	15-17.
5,02	8-10.
5,02	12-14.
4,99	10-12.
4,92	16-18.
4,73	**21-22.**

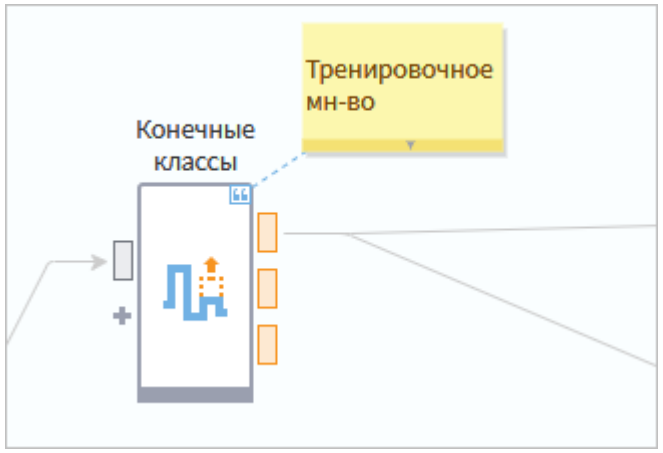
- Анализ проверяет, как влияет размер интервала доставки на отмену заказа.
- \+ признаки **Позднее время доставки** и **Размер интервала**.

Проверка гипотез

Проверка гипотез осуществлялась на основе WoE- и IV-анализа.

IV - величина, определяющая значимость анализируемого признака:

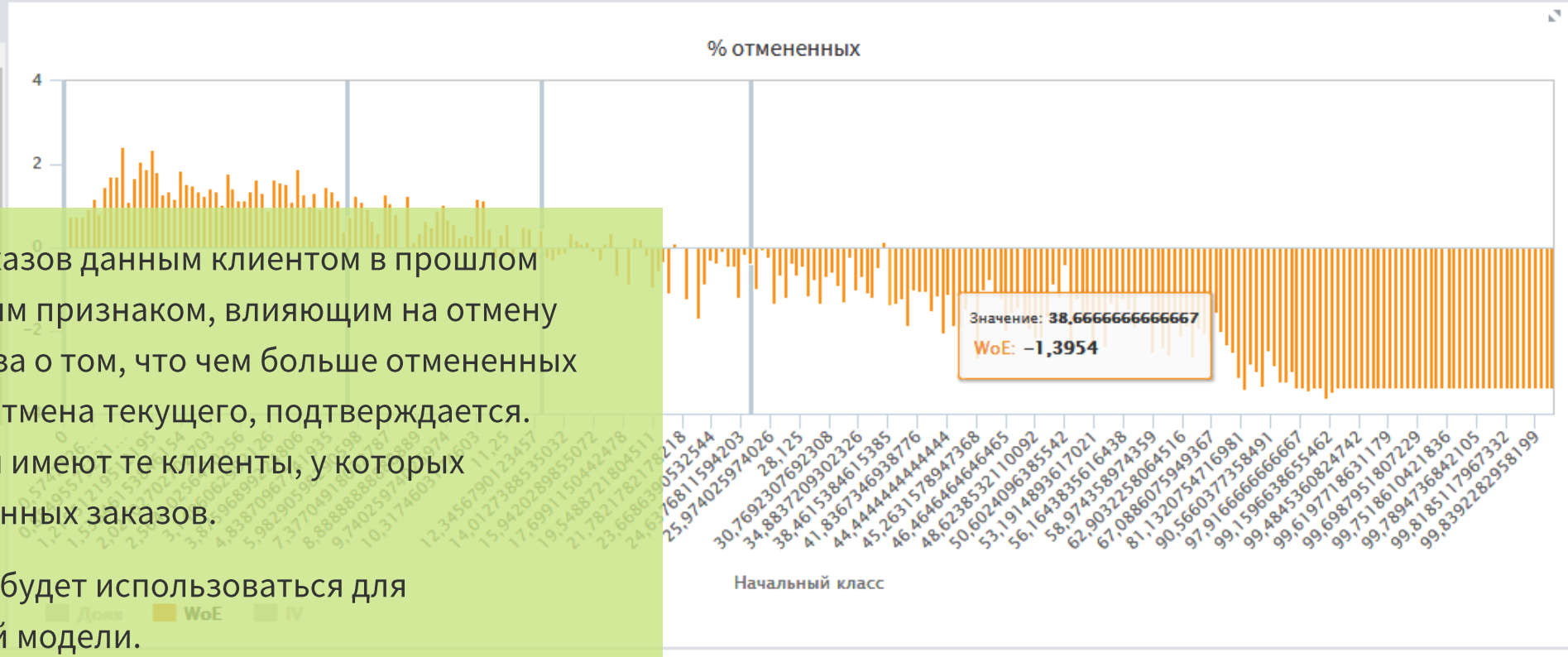
$$IV = \sum_{i=1}^{k} \left\{ \left(\frac{N_i}{N} - \frac{P_i}{P} \right) \times WoE_i \right\}$$



Настройка | Сценарий | Конечные классы

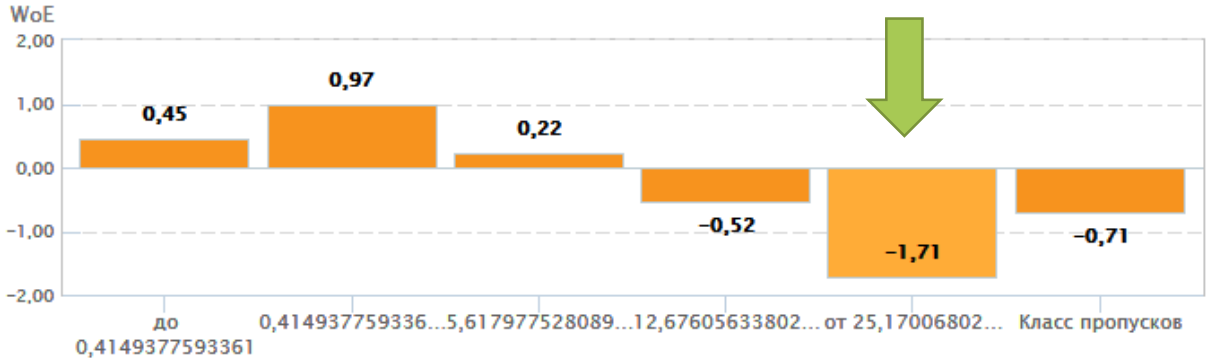
Пакеты | framework | Модуль1 | Сценарий | Конечные классы | Визуализаторы | Конечные классы

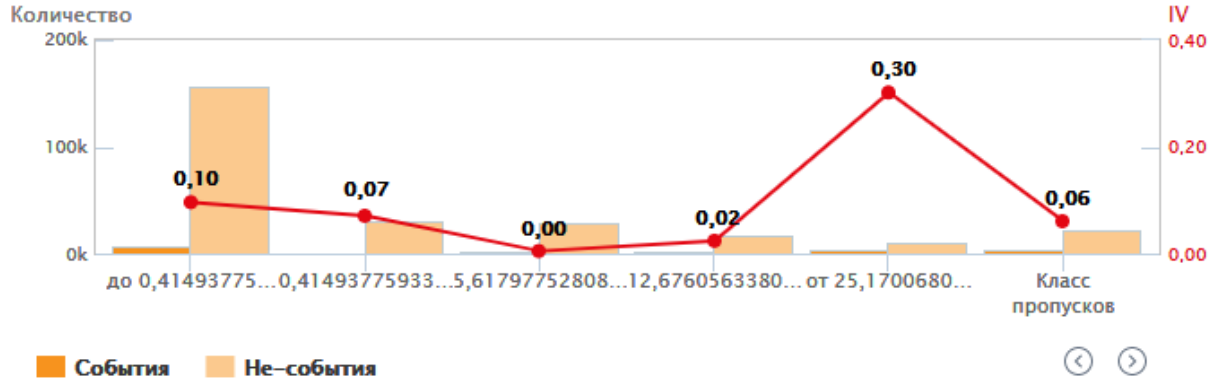
Столбец	IV
% отмененных	0,5
Количество отмененны…	0,3
Предыдущий заказ отм…	0,2
Дней с предыдущего за…	0,2
Средний интервал меж…	0,1
Объем заказа, шт.	0,1
Предоплата	0,1
Среднее кол-во позици…	0,1
Количество позиций	0,1
Количество групп	0,1
Заказ в предп. интервал	0,0
Интервал доставки	0,0



Процент отмененных заказов данным клиентом в прошлом является самым значимым признаком, влияющим на отмену текущего заказа. Гипотеза о том, что чем больше отмененных заказов, тем вероятнее отмена текущего, подтверждается. Особенно высокие риски имеют те клиенты, у которых больше четверти отмененных заказов.

Именно этот показатель будет использоваться для построения тривиальной модели.

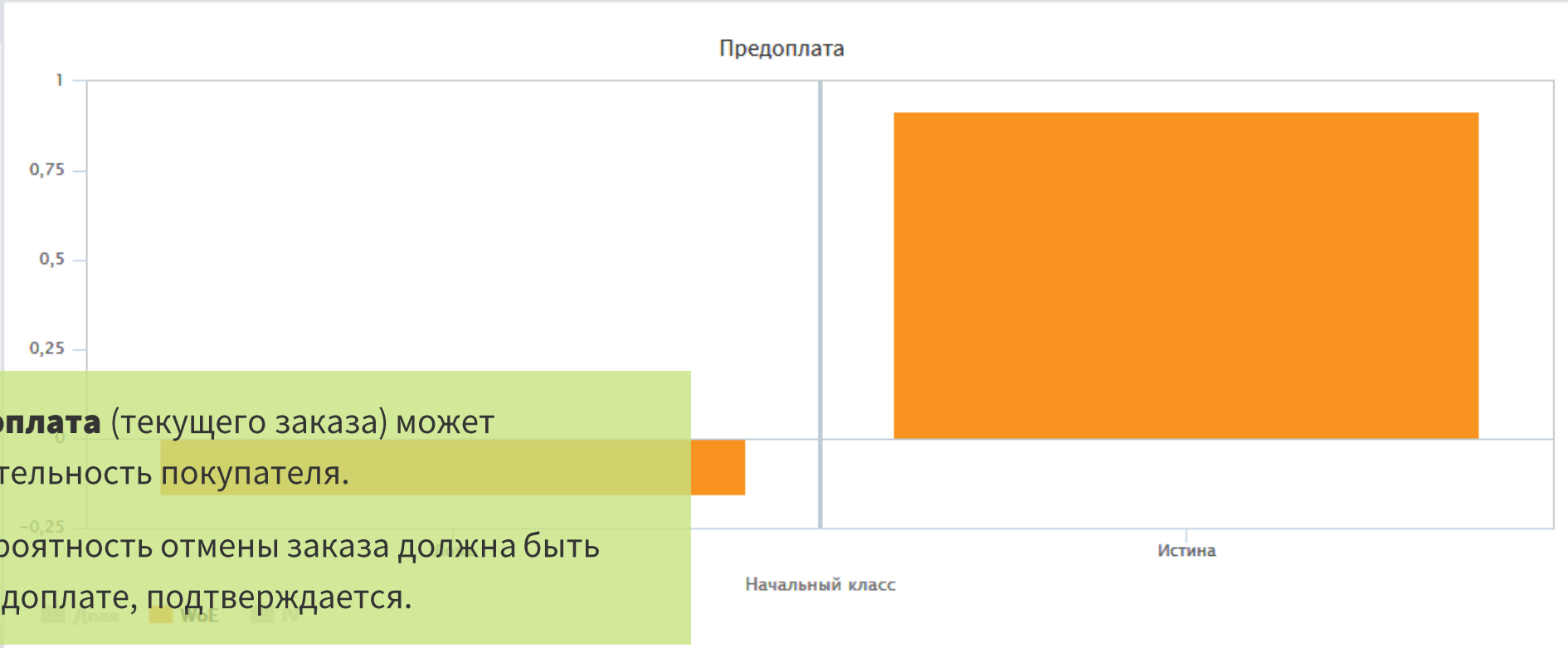




Настройка | Сценарий | Конечные классы

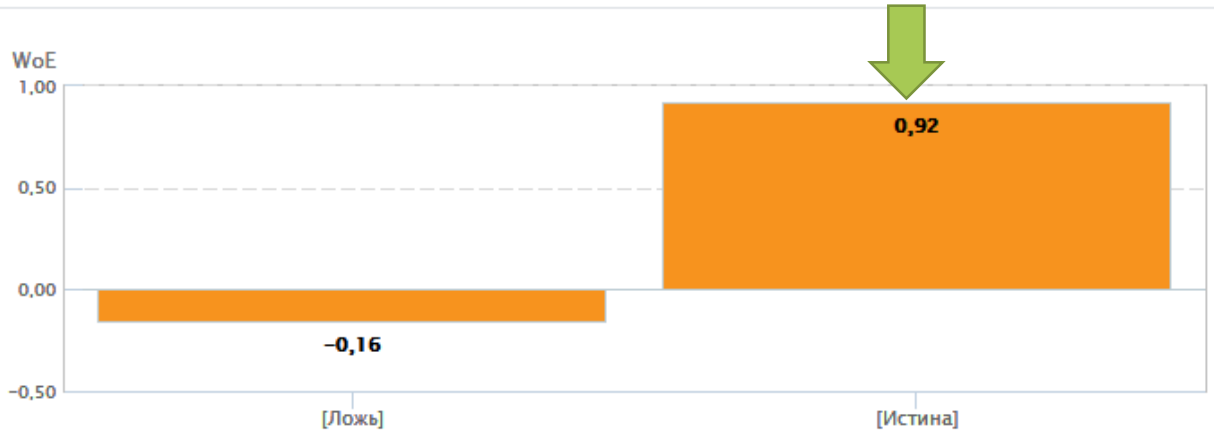
Пакеты | framework | Модуль1 | Сценарий | Конечные классы | Визуализаторы | Конечные классы

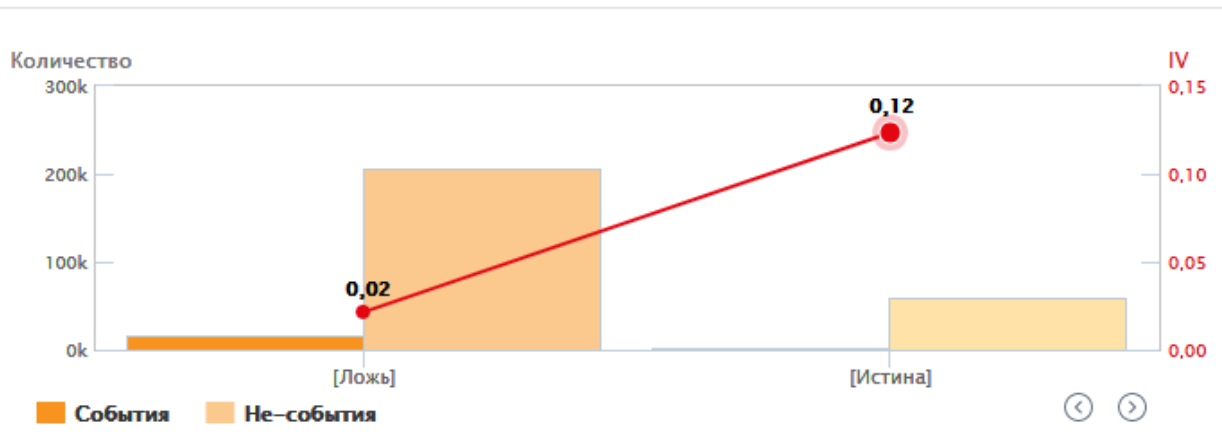
Столбец	IV
% отмененных	0,5
Количество отмененны…	0,3
Предыдущий заказ отм…	0,2
Дней с предыдущего за…	0,2
Средний интервал меж…	0,1
Объем заказа, шт.	0,1
Средний объем заказа	0,1
Предоплата	0,1
Среднее кол-во позици…	0,1



Значение в поле **Предоплата** (текущего заказа) может характеризовать решительность покупателя.

Гипотеза о том, что вероятность отмены заказа должна быть ниже для заказа по предоплате, подтверждается.

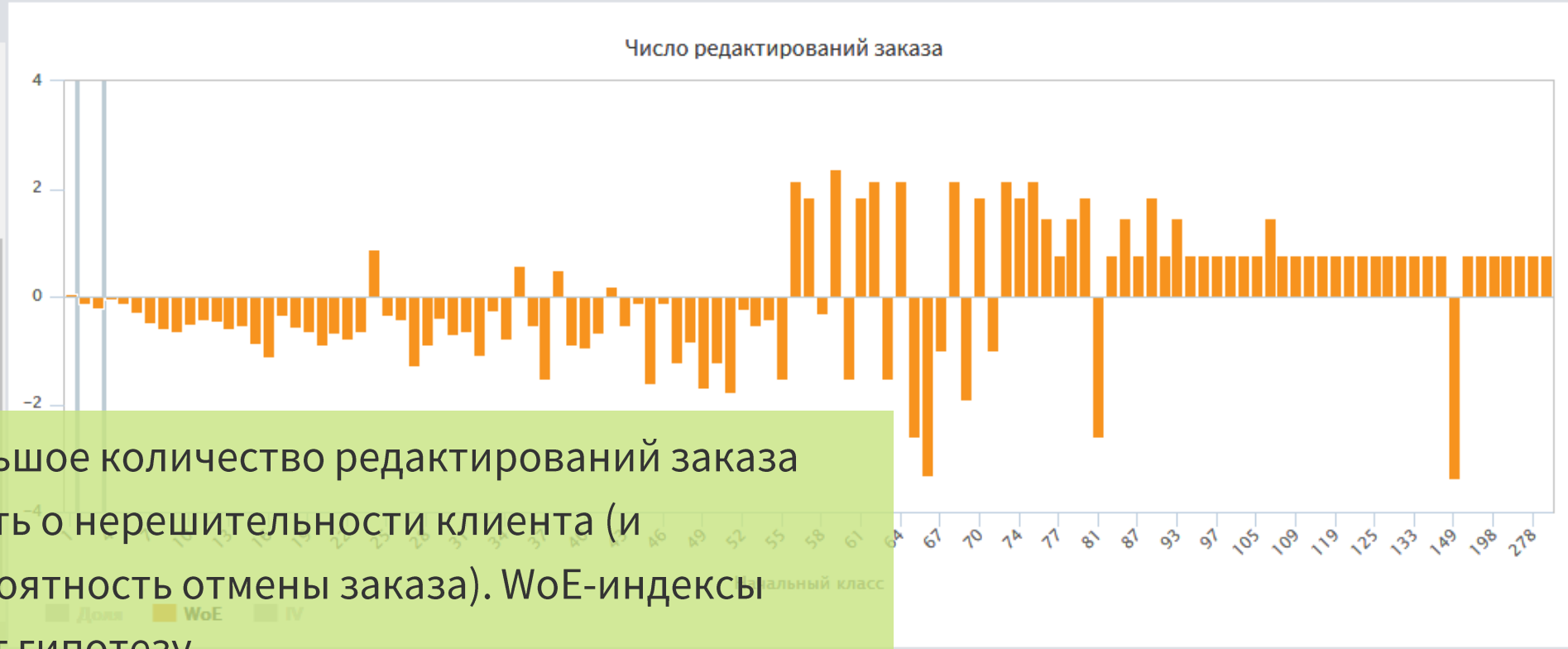




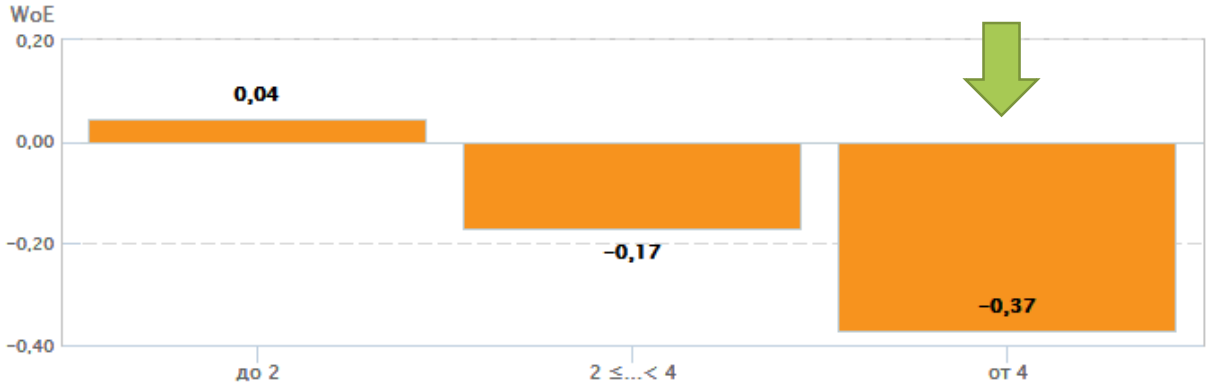
Настройка | Сценарий | Конечные классы

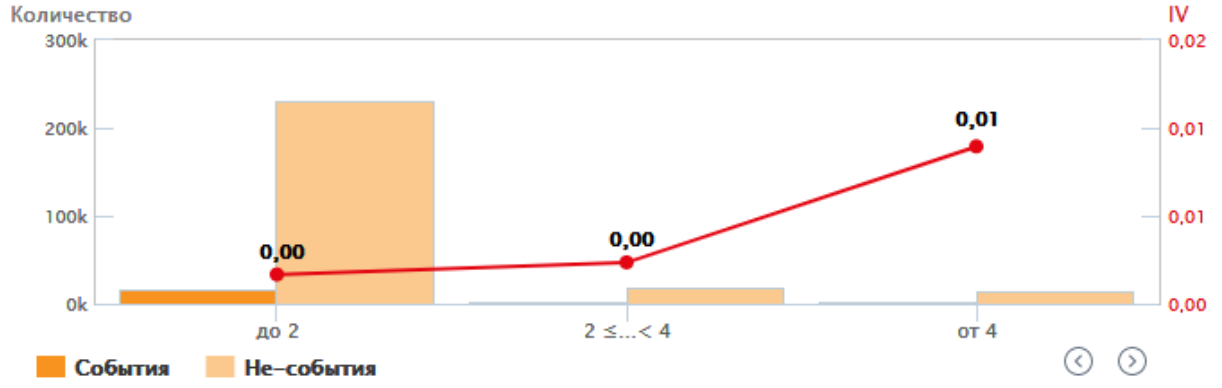
Пакеты | framework | Модуль1 | Сценарий | Конечные классы | Визуализаторы | Конечные классы

Столбец	IV
9.0 Среднее кол-во позици…	0,1
ab Предпочтительный инт…	0,1
12 Кол-во заказов на дату …	0,1
12 Количество позиций	0,1
0/1 Заказан товар из предп…	0,1
12 Дней между заказом и д…	0,1
12 Количество групп	0,1
0/1 Заказ в предп. интервал	0,0
ab Интервал доставки	0,0
12 Число редактирований …	0,0
ab Месяц заказа	0,0
ab Кластер доставки	0,0
0/1 Заказ в выходной	
0/1 Доставка в выходной	0,0
0/1 Одноразовый клиент	
ab ID канала	
ab Тип доставки	



Гипотеза: большое количество редактирований заказа может говорить о нерешительности клиента (и повышать вероятность отмены заказа). WoE-индексы подтверждают гипотезу.





Результаты проверки гипотез

Приняты:

- Большие интервалы предполагаемой доставки влияют на отмену
- Дата, выпадающая на выходные/будние дни, может влиять на отмену
- Количество редактирований способствует высокой вероятности отмены
- Тип доставки влияет на отмену
- Наличие предоплаты способствует низкой вероятности отмены заказа

Отвергнуты:

- Связь между группами товаров и отменой заказа
- Зависимость отмены заказов от больших праздников
- Связь между типом доставки и группами товаров (скоропортящихся) товаров
- Предполагаемое время доставки выпадает на ночное время суток или раннее утро, что влияет на вероятность отмены
- Некоторые отдельные товары могут иметь большую вероятность отмены заказа

Входные признаки

Из истории прошлых заказов:

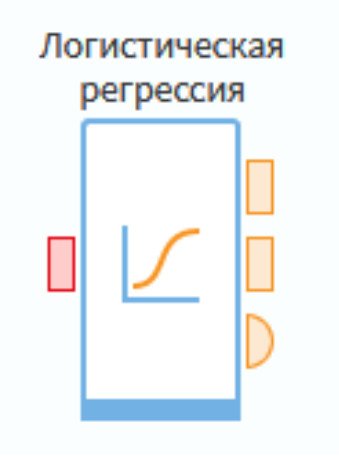
- % отмененных заказов
- Первый заказ

Из текущего заказа:

- Предоплата
- Сумма заказа
- Число корректировок заказа
- Интервал заказ-доставка
- Количество часов в интервале
- Кластер доставки
- Канал поставки
- Группа товара с максимальной суммой
- Месяц доставки
- Число групп в заказе

Построение модели

- Логистическая регрессия
- Ридж-регрессия с L2-регуляризацией
- Предварительный биннинг значений полей (компонент **Конечные классы**)
- Визуализатор **Качество бинарной классификации**



Так выглядит в Loginom наш сценарий решения задачи.

Сценарий | Сценарий | Качество бинарной классификации

Пакеты | framework | Модуль1 | Сценарий | Тривиальная модель | Визуализаторы | Качество бинарной классификации

Выбор диаграммы

- (•) ROC-кривая
- ( ) PR-кривая
- ( ) Базовые показатели
- ( ) Диаграмма точности
- ( ) Диаграмма равновесия
- ( ) % распознанных событий
- ( ) Диаграмма роста
- ( ) Диаграмма отклика
- ( ) Диаграмма выигрыша

[x] Кумулятивная

10 диапазонов

Множества

[x] Обучающее [ ] Тестовое

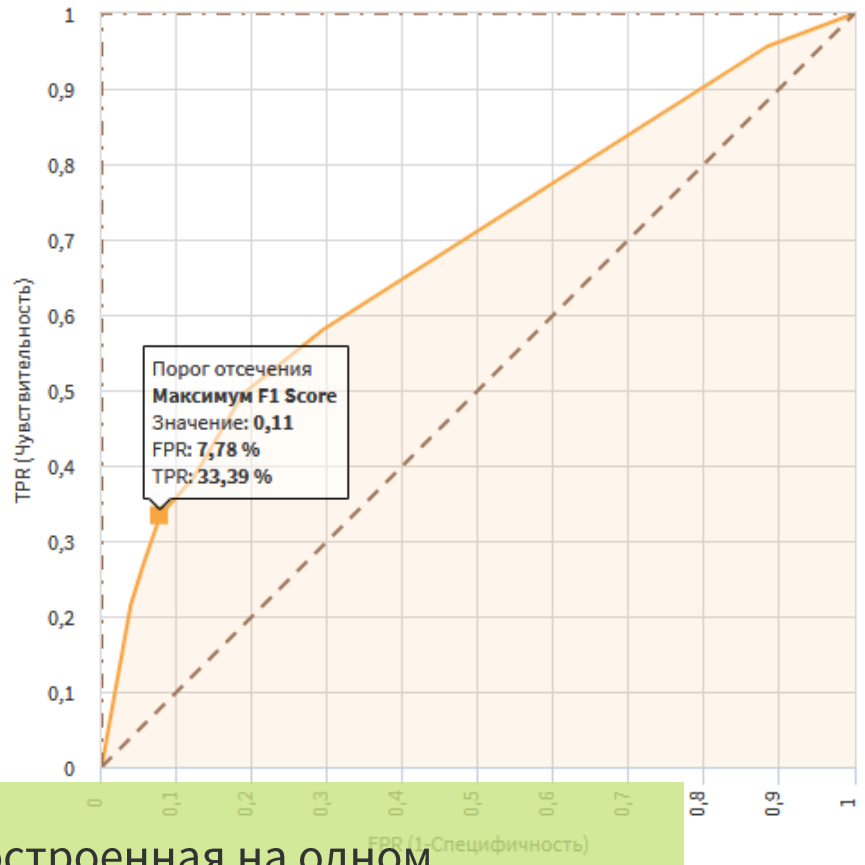
Порог отсечения

Максимум F1 Score

Значение порога:

ROC-кривая

Событие: Флаг отмены = Истина



Оценки классификации

Показатель	Множества	
	Обучающее	Тестовое
Оценки классификатора		
AUC ROC	0,6826	
AUC PR	0,1723	
Коэффициент Джини	0,3653	
KS	30,7959	
Порог отсечения: Максимум F1 Score		
Значение	0,1121	
TPR (Чувствительн…	0,3339	
TNR (Специфичность)	0,9222	
FPR (1-Специфично…	0,0778	
PPV	0,2356	
F1 Score	0,2763	
MCC	0,2185	

Матрицы ошибок

Классифицировано	Фактически		Итого
	Событие	Не-событие	
Обучающее	19 021	264 860	
Событие	6 352	20 604	26 956
Не-событие	12 669	244 256	256 925
Тестовое			
Событие			
Не-событие			

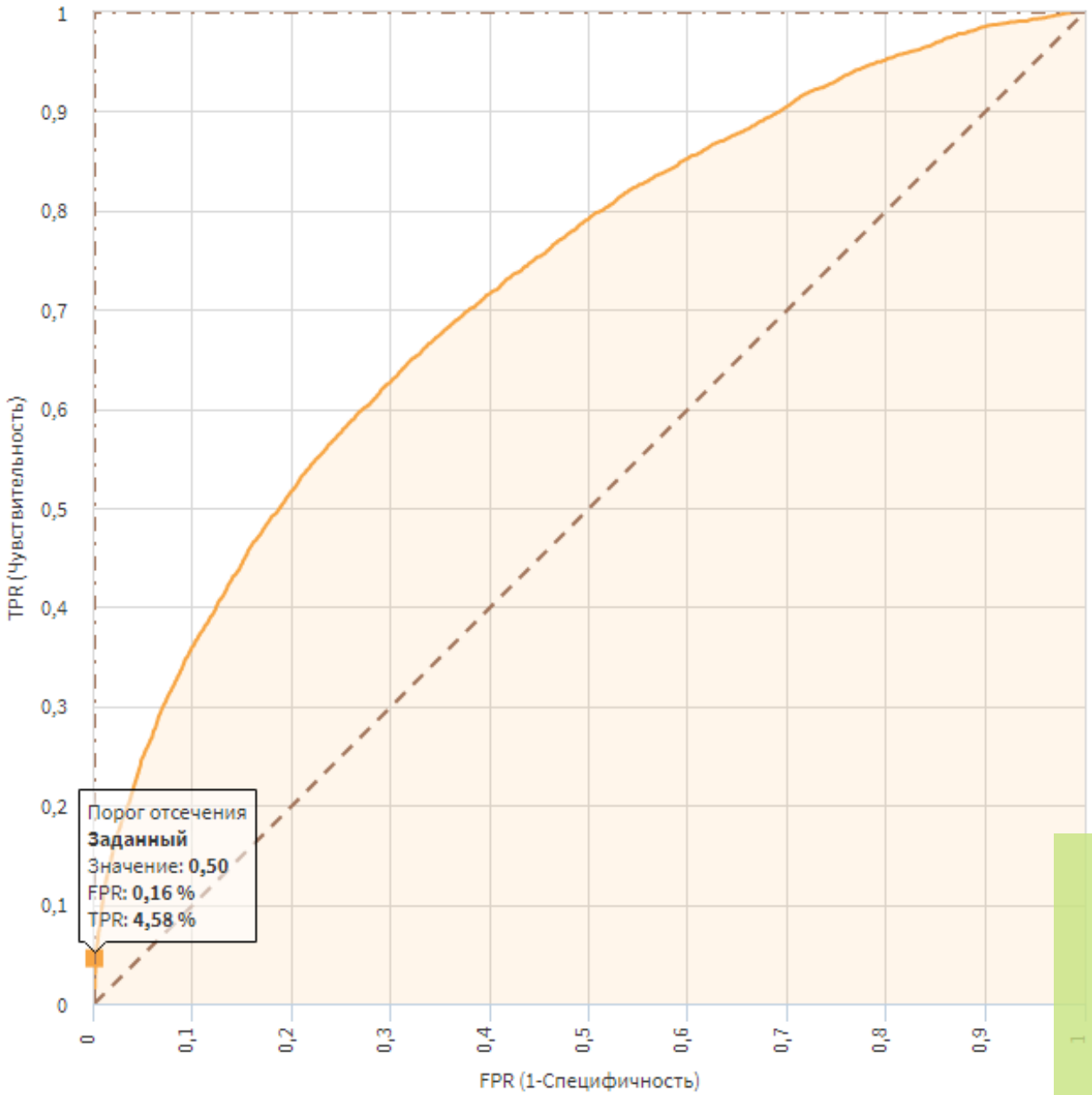
Распознано

Обучающее	250 608/283 881
Тестовое	

Тривиальная модель, построенная на одном признаке - % отмененных заказов в прошлом – имеет индекс ROC-AUC **0,6826**. Это базовый уровень предсказательной силы в нашем случае.

ROC-кривая

Событие: CancelFlag = Истина



Оценки классификации

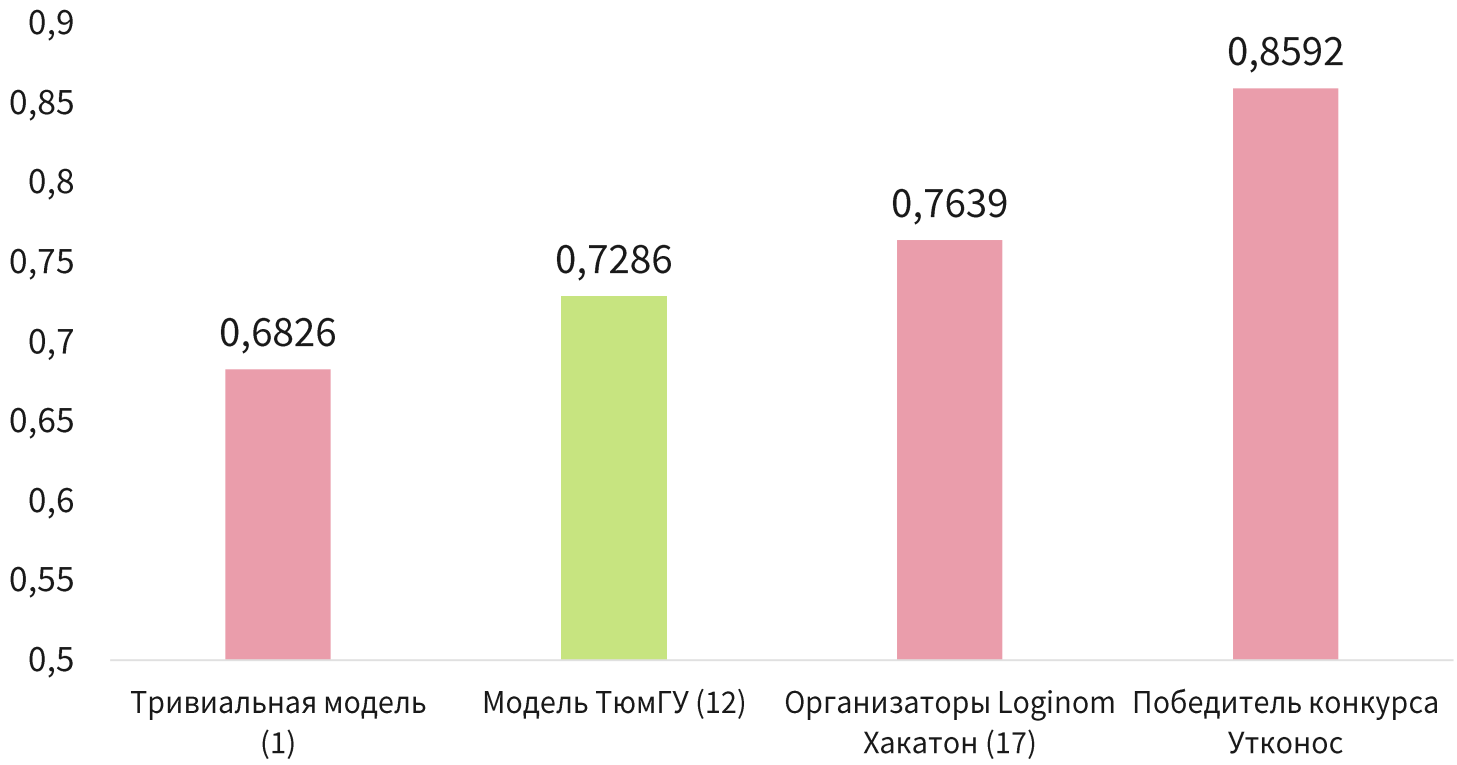
Показатель	Множества	
	Обучающее	Тестовое
Оценки классификатора		
AUC ROC	0,7282	
AUC PR	0,2238	
Коэффициент Джини	0,4564	
KS	32,9859	
Порог отсечения: Заданный		
Значение	0,5000	
TPR (Чувствительность)	0,0458	
TNR (Специфичность)	0,9984	
FPR (1-Специфичность)	0,0016	
PPV	0,6656	
F1 Score	0,0856	
MCC	0,1629	

Матрицы ошибок

Классифицировано	Фактически		Итого
	Событие	Не-событие	
Обучающее	4 743	66 195	
Событие	217	109	326
Не-событие	4 526	66 086	70 612
Событие			
Не-событие			

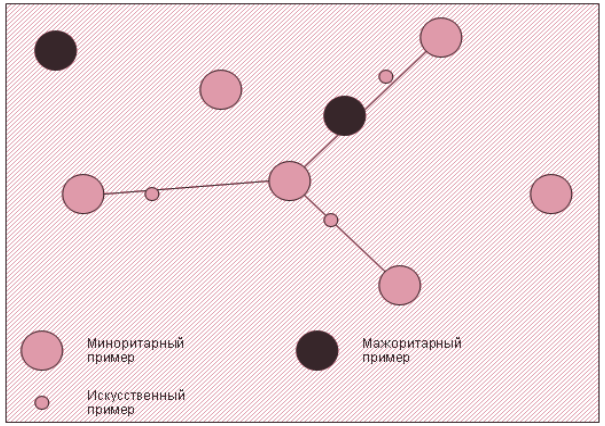
Многофакторная модель, построенная на 12 показателях, имеет индекс ROC-AUC **0,7282**. Прирост составил **6,7 %**.

Сравнение результатов

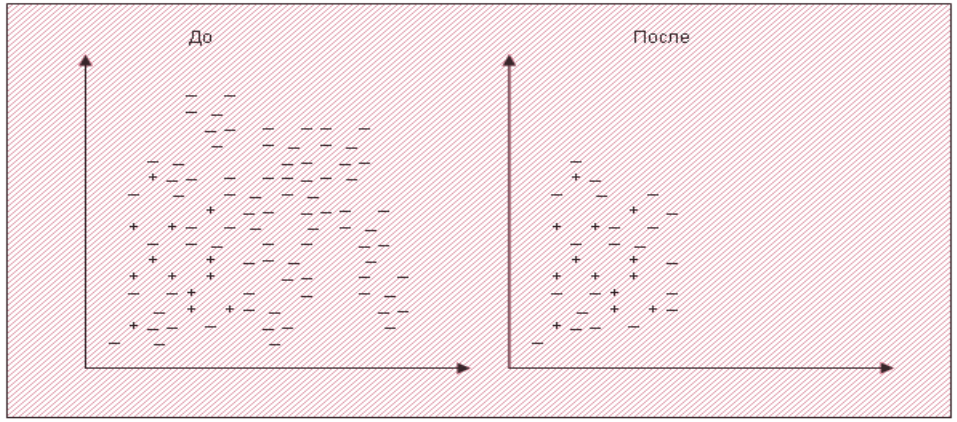


Невысокий результат объясняется тем, что в модели не использовались данные из истории прошлых заказов, кроме двух признаков – доли отмененных заказов и флага того, что заказ у клиента первый.

Учет дисбаланса классов



SMOTE
Генерация искусственных примеров, которые были бы «похожи» на имеющиеся в редком классе



NearMiss
Прореживание выборки путем удаления некоторых примеров мажоритарного класса

Результат: улучшения ROC-AUC не произошло.

Выводы

- Аналитическая платформа Loginom позволяет успешно решать задачи, связанные с построением моделей машинного обучения, без необходимости программировать.
- Исходные данные задачи нуждаются в предобработке перед использованием их для моделирования.
- Показатель ROC-AUC построенной нами модели по сравнению с тривиальной моделью выше на 6,7 %.
- Существенно поднять качество модели можно за счет процедур генерации признаков (Feature Engineering) на основе истории прошлых заказов клиентов.

Спасибо за внимание!