Черненко А.Е.

Проект

«Предсказание отклика абонента на подключение услуги»

Описание проекта

В качестве исходных данных представлена информация об отклике абонентов на предложение подключения одной из услуг. Каждому пользователю может быть сделано несколько предложений в разное время, каждое из которых он может или принять, или отклонить.

Тренировочный набор: **data_train.csv**

Тестовый набор: **data_test.csv**

Отдельным набором данных представлен нормализованный анонимизированный набор признаков, характеризующий профиль потребления абонента. Эти данные привязаны к определенному времени, поскольку профиль абонента может меняться с течением времени.

Набор признаков: **features.csv**

Задача:

Построить модель предсказания отклика абонента на предлагаемую к подключению услугу.

Создание датасета

Создание датасета осуществляется с применением библиотеки для распараллеливания Dask.

Задача:

Дополнить все наблюдения из data_train.csv и data_test.csv признаками из `features.csv`.

Алгоритм решения:

1. Слияние входной выборки и выборки признаков методом пересечения (how='inner') по столбцу id.
2. Создание признака 'time_delta‘, который отражает абсолютную разницу во времени между поступлением предложения и фиксированием признаков в профиле потребления.
3. Удаление всех наблюдений с дублированным индексом входной выборки, остаются только по одному наблюдению, которое имеют минимальное значение `time_delta` среди своих дубликатов.
4. Сохранение полученных тренировочного и тестового наборов для дальнейшей работы. Для удобства хранения и использования наборы сохраняются в .pkl с типом данных float32.

EDA

- **buy_time** - временной штамп отклика клиента на услугу. Анализ признака показал, что он несет информацию только о неделе и месяце. Более того, наблюдается короткий период с аномальным скачком частоты положительных откликов. Принято решение не использовать этот признак.
- Использование **id** в модели нецелесообразно. **id** - уникальный идентификатор клиента: как число использовать нельзя, а как категорию практически невозможно, так как появляются новые клиенты.
- **time_delta** - временной интервал в секундах, разница между временным штампом отклика клиента на услугу и временным штампом записи о профиле потребления клиента, по смыслу может отражать актуальность данных профиля потребления на момент отклика клиента на услугу.
- **vas_id** - вид услуги, будем использовать как категориальный тип.
- **Target -** Целевая переменная, определяет отклик клиента на услугу.
  0 - клиент отказался от услуги
  1 - клиент подключил услугу

  Наблюдается сильная несбалансированность классов в target.

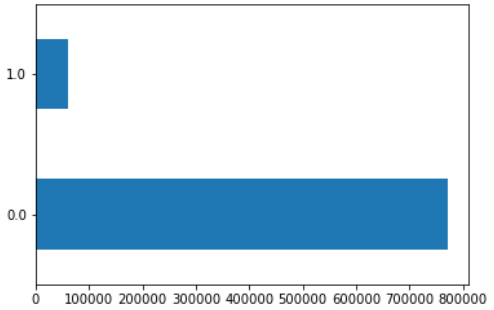


Рис. – распределение target

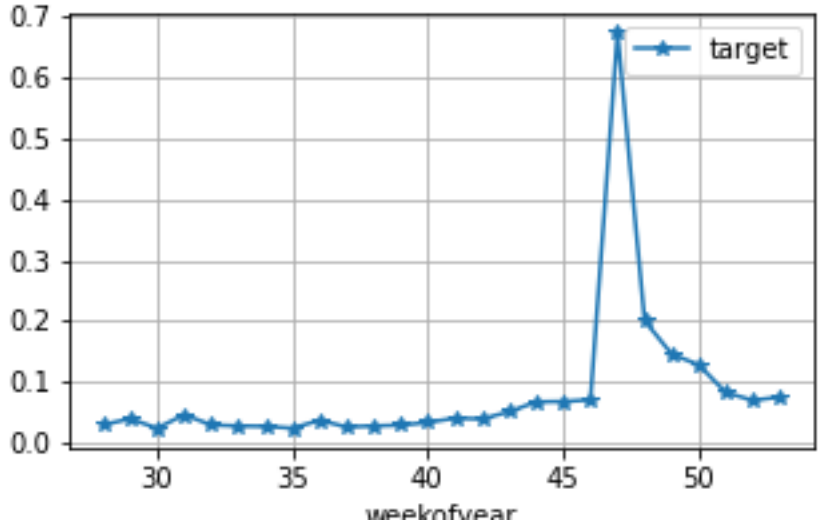


Рис. - Зависимость частоты положительного отклика от недели

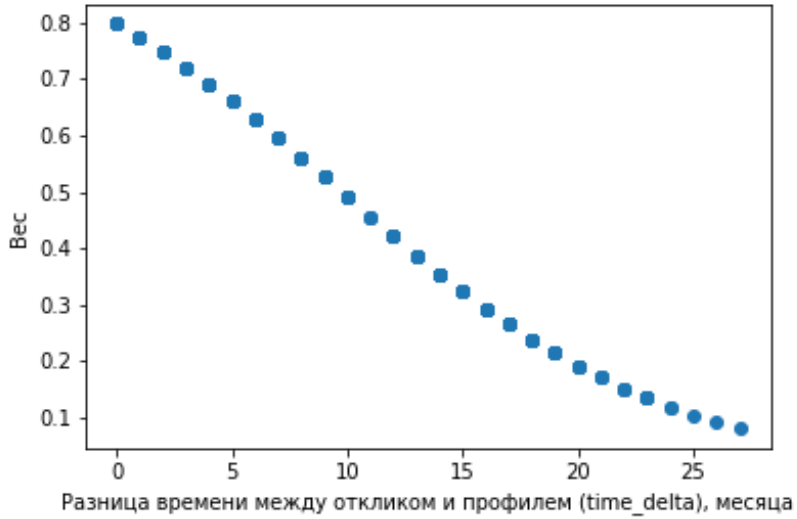


Рис. – Переход от time_delta к весовому значению

Определение типов признаков

Распределение признаков по типу:

Всего : 255

Константные : 5 (*к удалению*)

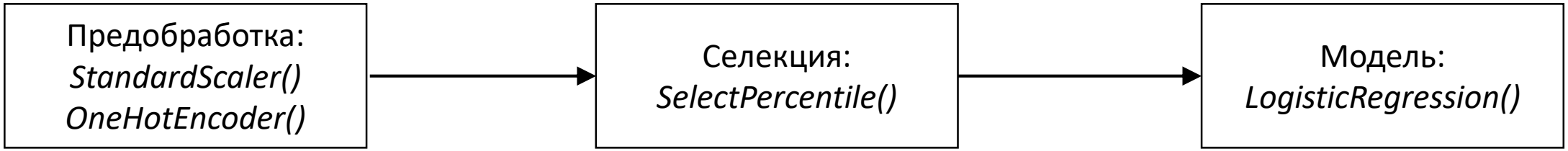
Категориальные : 2 (*vas_id и признак 252*)

Числовые : 247

Весовой признак : 1 (*трансформированный time_delta*)

Логистическая регрессия

Построение модели логистической регрессии выполнялось с применением пайплайнов.



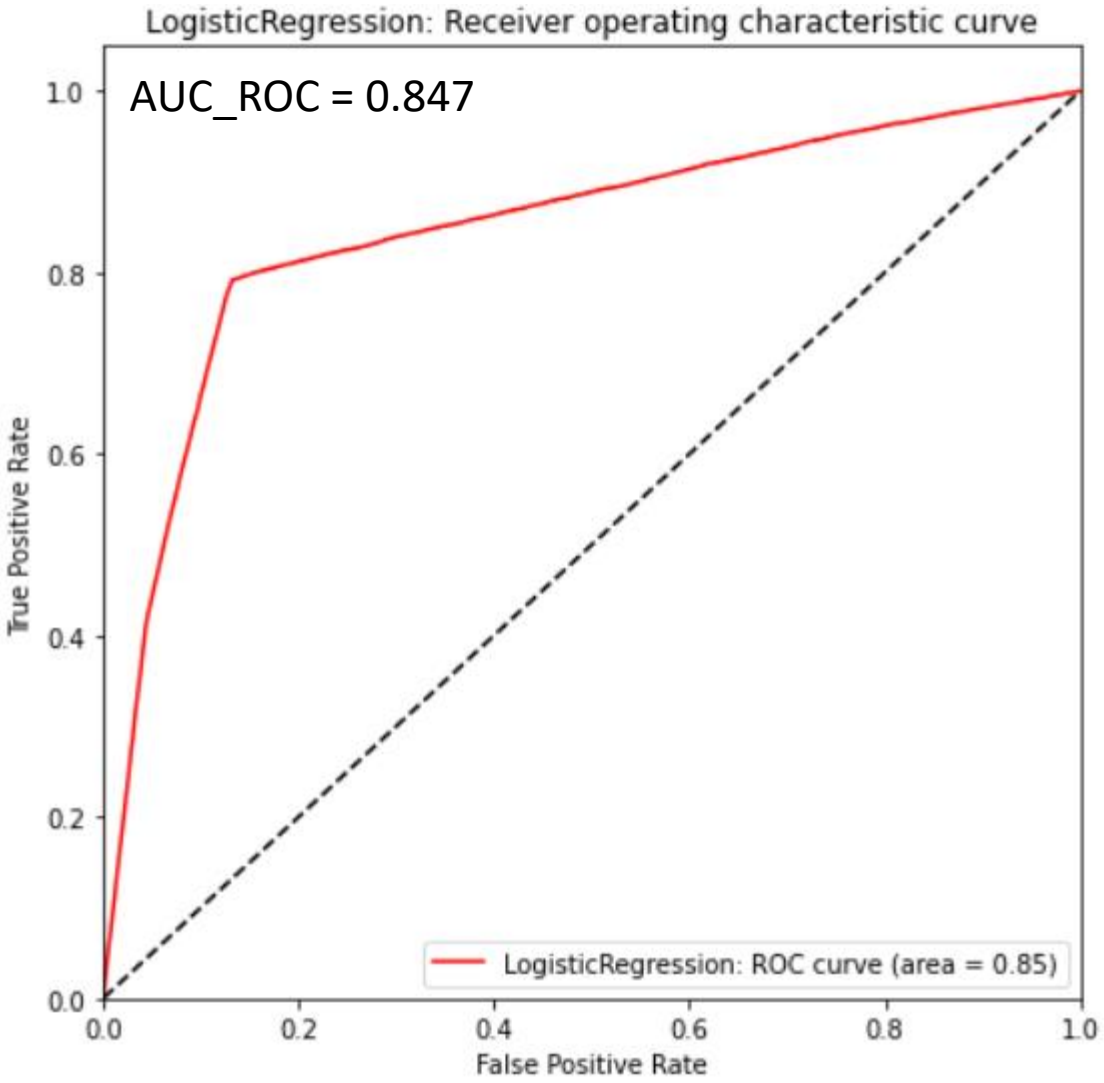
Подбор гиперпараметров модели осуществлялся при помощи **поиска по сетке** с использованием **кросс-валидации.**

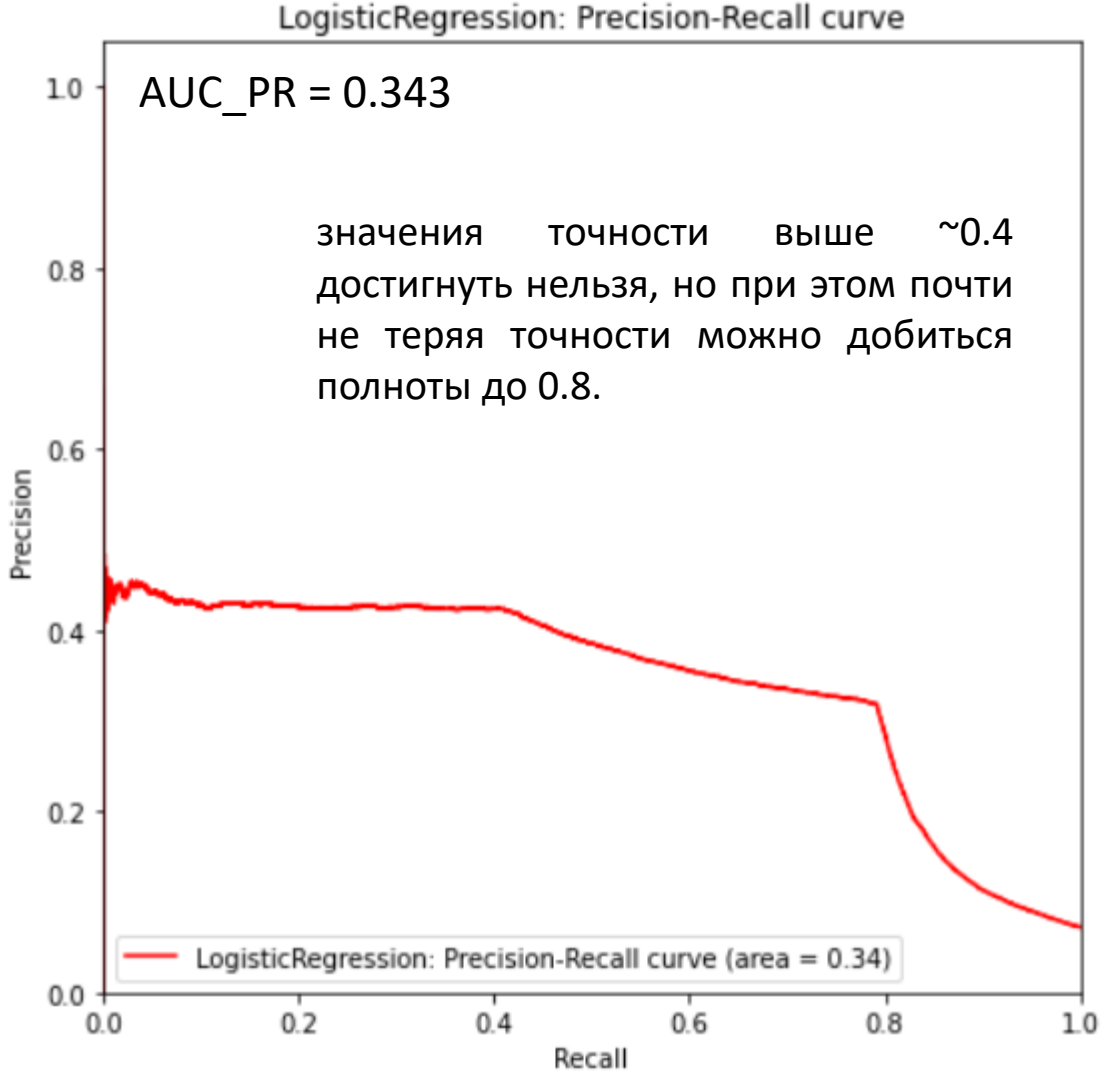
Параметры выбранной модели:

- class_weight='balanced‘ – автоматическая балансировка классов
- C=10 – обратная сила регуляризации

Рис. – Переход от time_delta к весовому значению

Логистическая регрессия





значения точности выше ~0.4 достигнуть нельзя, но при этом почти не теряя точности можно добиться полноты до 0.8.

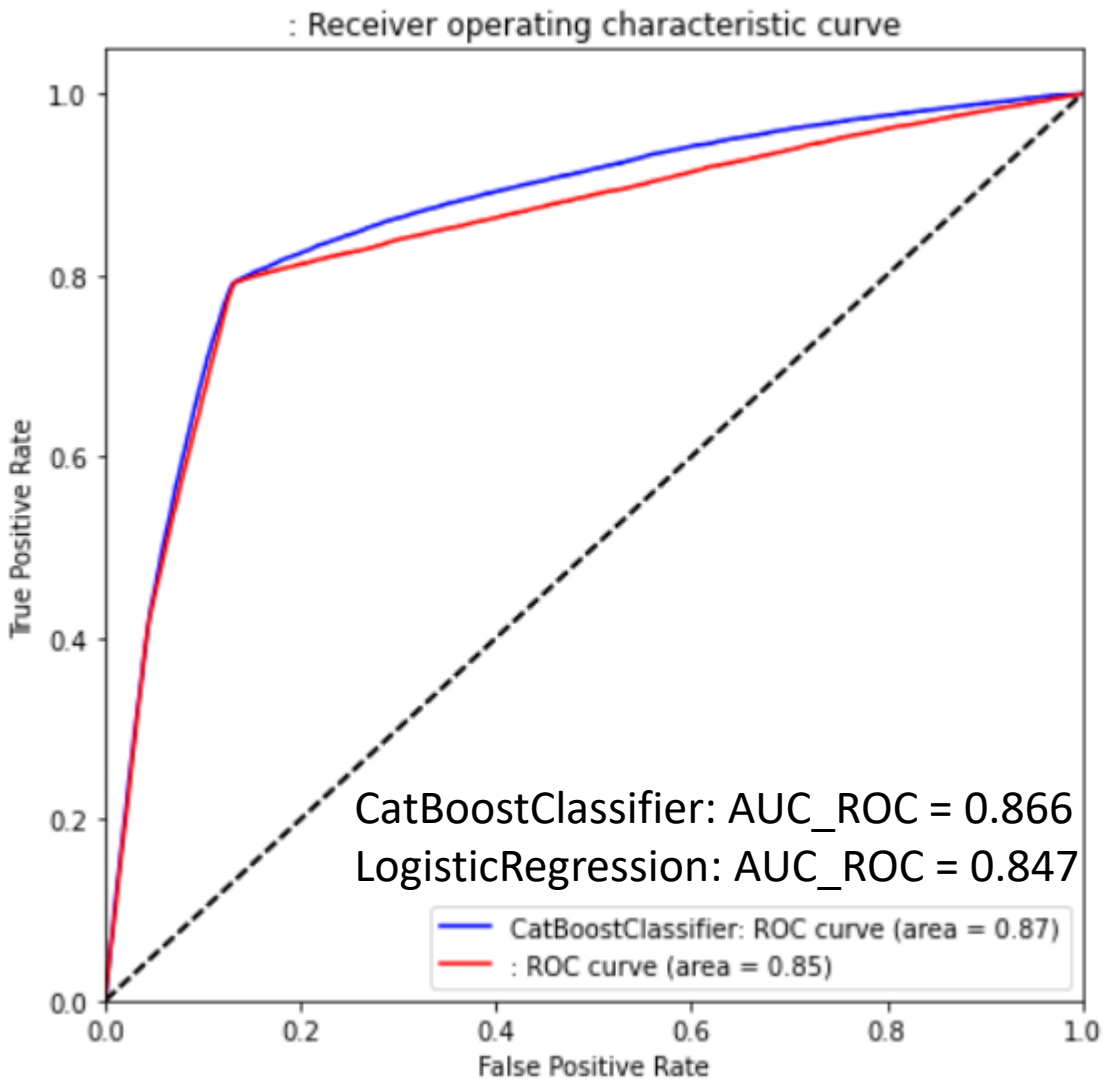
Библиотека Catboost

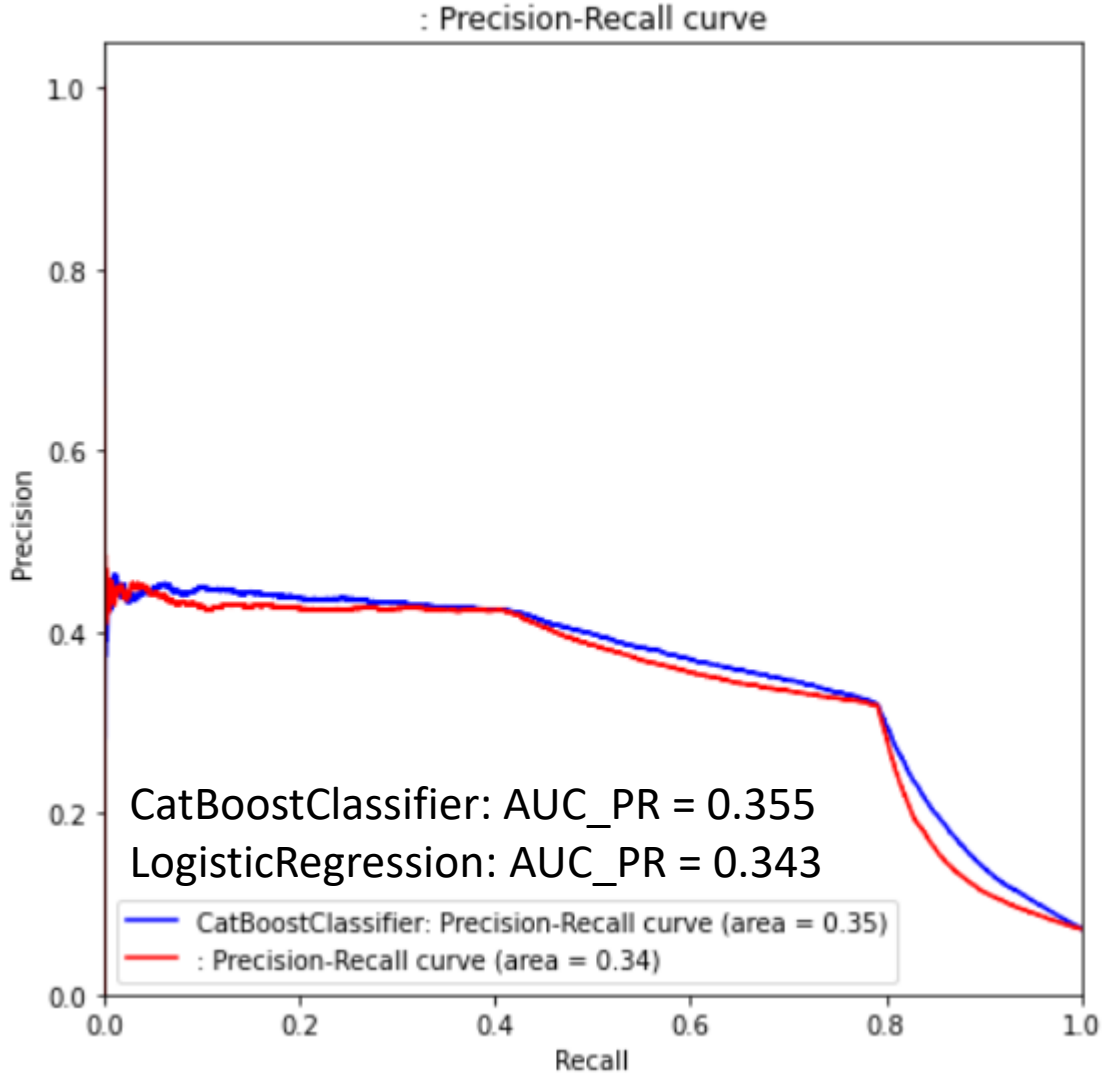
Подбор гиперпараметров модели осуществляется при помощи **рандомизированного поиска по сетке** с использованием **кросс-валидации,** проверяется 30 наборов гиперпарам**етров.**

Параметры выбранной модели:

- функция потерь : Logloss
- автобалансировка классов
- количество деревьев : 500
- скорость обучения : 0.1
- глубина дерева : 8
- регуляризация : 5

Сравнение моделей





Модели практически не отличаются, Catboost немного лучше

Confusion matrix и практическое применение результата

Нулевая гипотеза - положительный отклик клиента на услугу (класс 1)
Ошибки первого рода (False Negative – FN) - сколько клиентов не получили предложения, хотя потенциально готовы совершить подключение;
Ошибки второго рода (False Positive – FP) - сколько клиентов получили предложения, хотя они не собираются совершать подключение.

Catboos valid prediction

TN : 221258	**FP** : 33369
FN : 4138	**TP :** 15681

Прибыль оператора от положительного отклика клиента на услугу равна разнице между **Доходом** от клиента и **Затратой** на рассылку предложения этому клиенту.

$$\text{Прибыль} = \text{Доход} - \text{Затрата}$$

Таким образом, ошибка первого рода будет отражать какой **Доход** оператор потерял не отправив предложение. А ошибка второго рода - какие **Затраты** оператора на рассылку оказались напрасными.
Итого **реальная прибыль** от рассылки предложений будет определяться формулой:

$$REAL = N_{TP} \cdot (\text{Доход} - \text{Затрата}) - N_{FP} \cdot \text{Затрата}$$

а **упущенная прибыль**:

$$\text{L}OSS = N_{FN} \cdot (\text{Доход} - \text{Затрата})$$

Максимально возможная прибыль определяется формулой

$$\text{M}AX = (N_{TP} + N_{FP}) \cdot (\text{Доход} - \text{Затрата})$$

где :
N_{TP} - количество положительных откликов на отправленное предложение
N_{FN} - количество упущенных клиентов, готовых подключить услугу (ош. 1 рода, FN)
N_{FP} - количество напрасно отправленных предложений (ош. 2 рода, FN)

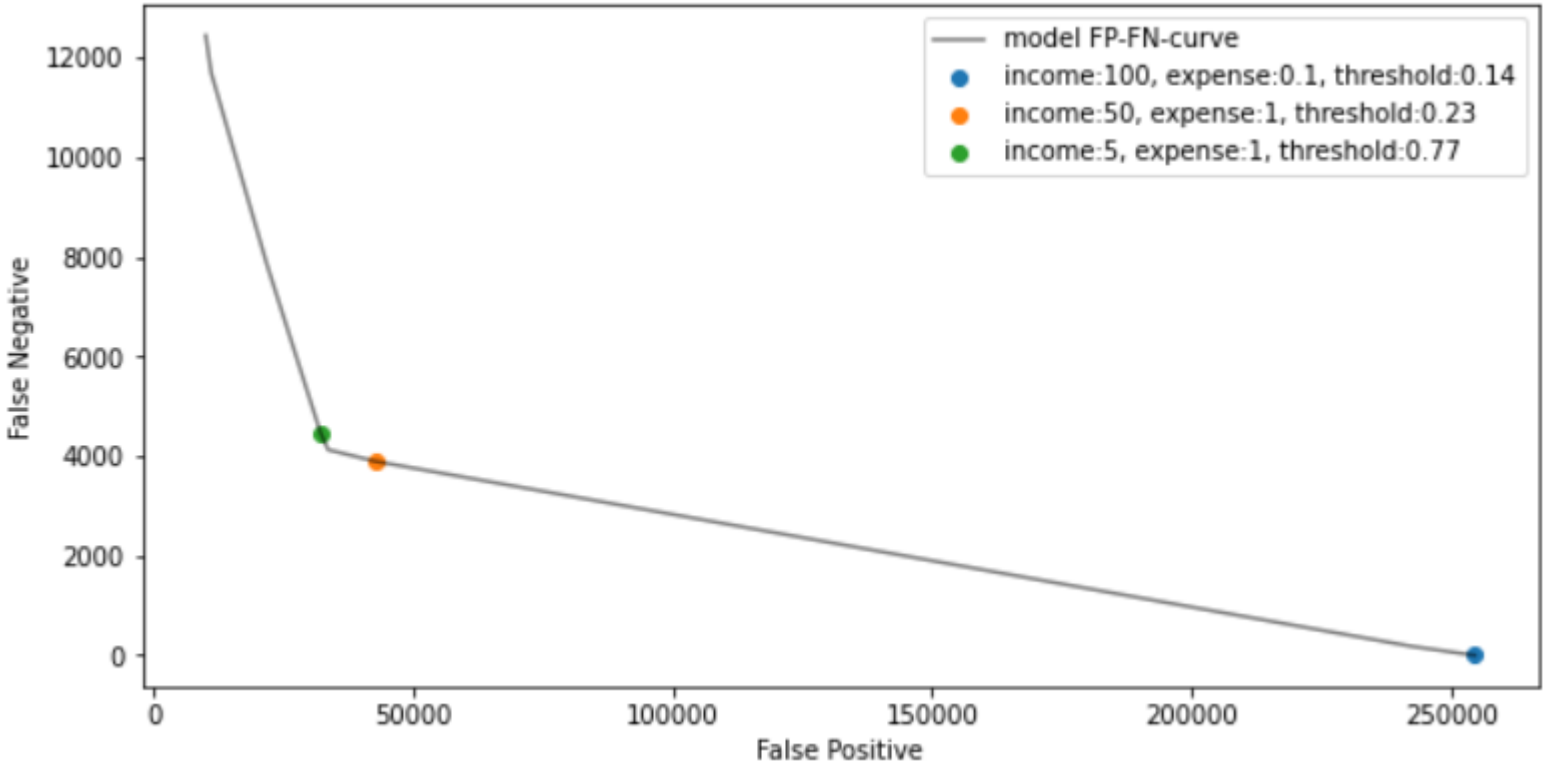
Для получения **максимальной выгоды** необходимо минимизировать разницу **MAX - REAL**, то есть упрощая выражения

$$\frac{\text{Доход} - \text{Затрата}}{\text{Затрата}} \cdot N_{FN} + N_{FP} \rightarrow min$$

Так как **Доход** от подключения услуги, как правило, на несколько порядков превышает **Затрату** на рассылку предложения (например, услуга со стоимостью подключения 100 р., затрата на смс-рассылку 10 коп.), то N_{FN} гораздо сильнее влияет, чем N_{FP}, на изменение выгоды. **Максимально возможная выгода** достигается лишь, когда $N_{FN} = N_{FP} = 0$, то есть оператор абсолютно безошибочно разослал все предложения. Это практически невозможно и является идеальным случаем.

Выбор стратегии

Построим график количества N_{FN} и N_{FP} при разных порогах на предсказаниях модели Catboost. Используя формулу минимизации и задавая конкретные значения **Дохода (income)** от услуги и **Затрат (expense)** на предложение найдем порог классификации, при котором мы добьёмся **максимальной выгоды.**



Пример расчета

Доход, руб.	Затрата, коп.	Оптимальный порог
100	10	0.14
50	100	0.23
5	100	0.77

Таким образом, показана стратегия выбора порога классификации для максимизации получаемой прибыли при заданных значениях дохода от подключенной услуги и затраты на рассылку предложения.