# Описание кейса «Прогнозирование оттока клиентов»

**Суть бизнес задачи:**

Каждый год компания пролонгирует полисы Каско клиентов - физических лиц. Для оптимизации работы со списками на пролонгацию необходимо прогнозировать с какой вероятностью каждый из клиентов пролонгируется и какие факторы на это влияют. В зависимости от этого расставляются приоритеты операторам колл-центра, которые обрабатывают список (осуществляют обзвон клиентов), а также принимаются решения по дополнительной мотивации клиентов к пролонгации.

**Описание массива:**

Дана выборка полисов на пролонгацию в формате CSV-файла с набором полей, характеризующих сам полис, клиента (и его историю страхования) и транспортное средство. Полис на пролонгацию – это полис, период действия заканчивается и который нужно пролонгировать.

Прогнозируемая переменная – факт пролонгации полиса «POLICY\_IS\_RENEWED», где «1» – клиент пролонгировался, «0» - клиент не пролонгировался.

Массив случайным образом разбит на 2 части: 80% данных – тренировочная выборка, 20% данных – тестовая выборка.

**Проверка и оценка результата:**

Участникам для целей моделирования и проверки данных предоставляется тренировочная и тестовая выборки. В тестовой выборке значения прогнозируемой переменной обнулены.

Ответ предоставляется в виде CSV файла, в котором содержатся только записи из тестовой выборки, вида:

Формат таблицы передачи результата:

|  |  |  |
| --- | --- | --- |
| POLICY\_ID | POLICY\_IS\_RENEWED | POLICY\_IS\_RENEWED\_PROBABILITY |

POLICY\_IS\_RENEWED\_PROBABILITY – вероятность пролонгации (если метод прогнозирования предполагает её расчёт), параметр не обязателен для заполнения.

**Методология определения победителя:**

По факту пересечения прогноза по факту пролонгации, полученного от участников, с фактом в тестовой выборке, строится матрица вида:

|  |  |  |  |
| --- | --- | --- | --- |
| Матрица проверки результата | | прогноз | |
| 0 | 1 |
| факт | 0 | A% | B% |
| 1 | C% | D% |

A% - количество верно предсказанных отказов от пролонгации

B% - количество неверно предсказанных пролонгаций (модель сказала, что клиент пролонгируется, по факту пролонгация не было)

С% - количество неверно предсказанных отказов от пролонгации (модель сказала, что клиент не пролонгируется, по факту пролонгация была)

D% - количество верно предсказанных пролонгированных полисов

Каждой модели проставляется рейтинг, рассчитанный по формуле A%+D% (с округлением до десятых долей). Победителем считается команда, чья модель получит наибольший рейтинг.