Прогнозированние оттока клиентов телекоммуникационной компании

Финальный проект специализации Coursera "Машинное Обучение и Анализ Данных"

Костарев Дмитрий

Цель проекта:
научиться находить пользователей, склонных к оттоку. Если научиться находить таких пользователей с достаточной точностью заблаговременно, то можно эффективно управлять оттоком.

* * *

Задача проекта:
Построить модель машинного обучения позволяющую прогнозировать вероятность того, что пользователь покинет сервис.

Особенности данных:

- размер выборки 50 тыс. обьектов
- 230 признаков (из которых первые 190 числовые, остальные 40 категориальные)

- несбалансированная выборка 93% пользователей не склонных к оттоку

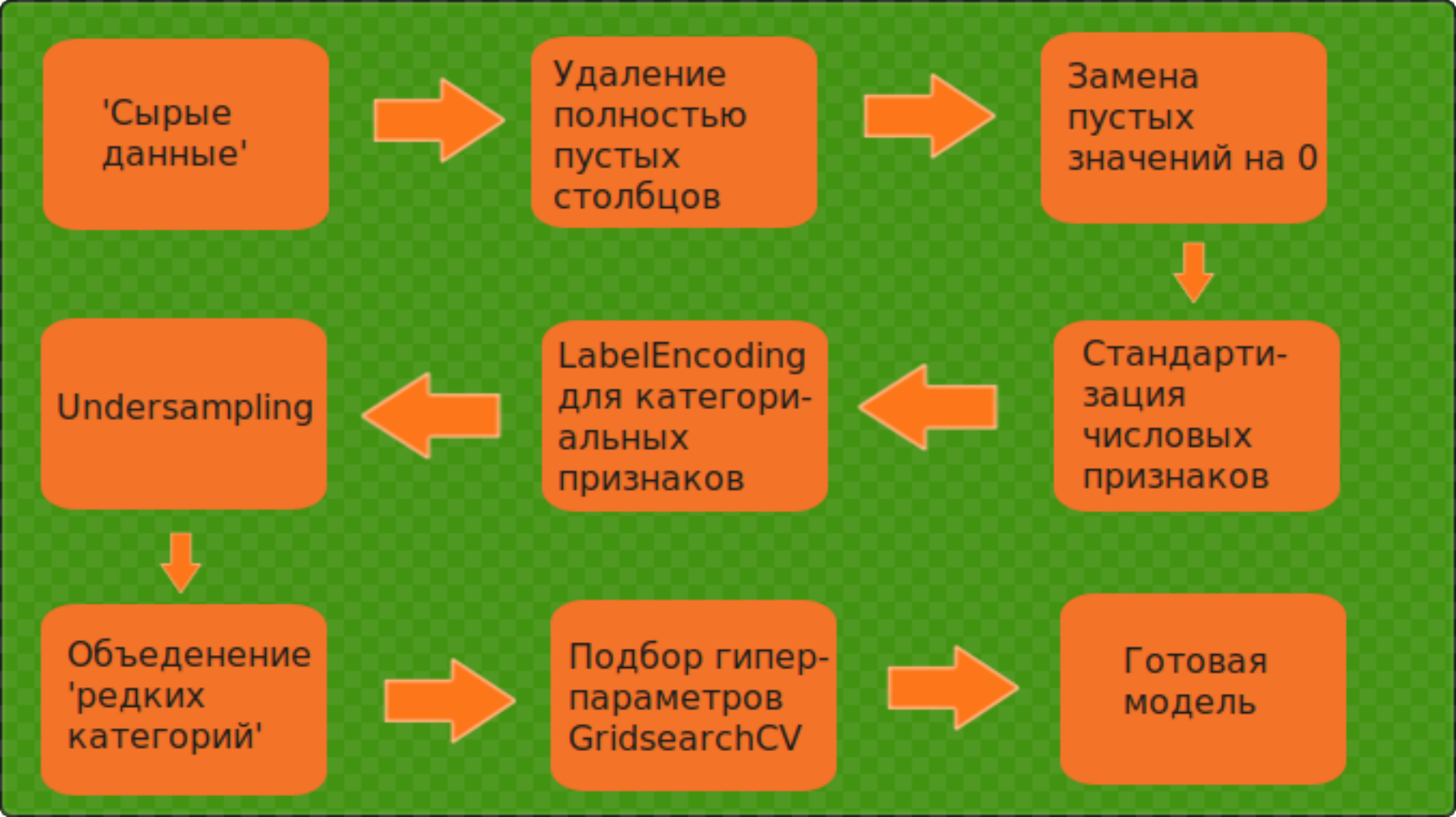
- в данных присутствует некоторое количество пропусков

* * *

Из-за дисбаланса классов и необходимости считать вероятность оттока пользователей из компании, ключевой метрикой будем использовать ROC-AUC. Успешным будем считать решение которое будет иметь положительный экономический эффект.

* * *

Пайплайн решения


'Сырые данные'
Удаление полностью пустых столбцов
Замена пустых значений на 0
Стандарти-зация числовых признаков
LabelEncoding для категори-альных признаков
Undersampling
Объеденение 'редких категорий'
Подбор гипер-параметров GridsearchCV
Готовая модель

В качестве классификатора использовали GradientBoostingClassifier из библиотеки sklearn.

Качество оценивали на отложенной выборке которая составляла 30% от всей выборки.

Финальное решение показало ROC-AUC = 0.73.

* * *

Экономический эффект от применения нашей модели:

$$Revenue = ARPU * TP * PoA - MHPU * (FP + TP)$$

ARPU - Average revenue per user (средняя выручка от одного пользователя).

MHPU - Money hold per user (кол-во денег, которое мы будем вкладывать в его удержание).

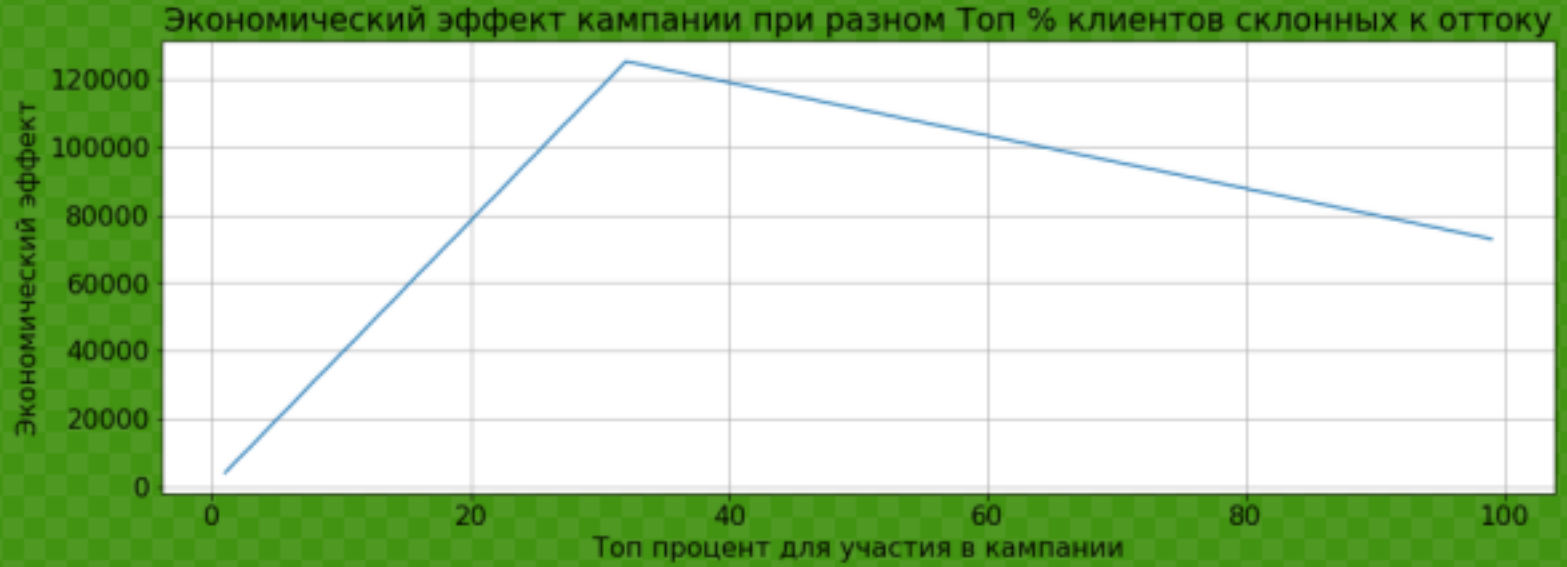
PoA - Probability of accept (вероятность, что предложение будет принято им).

TP - количество верно определенных пользователей склонных к оттоку

FP - количество неверно определенных пользователей склонных к оттоку

Расчет производился на следующих данных:

ARPU = 500 ед.
MHPU = 50 ед.
PoA = 0.6
TP, FP были взяты из нашей модели



Из графика видно, что наибольшую прибыль размером 125 тыс. ед. с нашими изначальными параметрами, можно получить используя топ 31% пользователей, которых наша прогнозная модель отнесла к классу отток.

* * *

Исходя из того, что нами был использован небольшой датасет с частью пользователей, применение данной модели может принести значительную прибыль компании.