Современные методы машинного обучения (курс майнора)/ДЗ4

В этом задании вам предлагается проанализировать данные одной из американских телекоммуникационных компаний о пользователях, которые потенциально могут уйти.

[Данные]

Измерены следующие признаки:

state — штат США

account\_length — длительность использования аккаунта

area\_code — деление пользователей на псевдорегионы, использующееся в телекоме

intl\_plan — подключена ли у пользователя услуга международного общения

vmail\_plan — подключена ли у пользователя услуга голосовых сообщений

vmail\_message — количество голосых сообщений, который пользователь отправил / принял

day\_calls — сколько пользователь совершил дневных звонков

day\_mins — сколько пользователь проговорил минут в течение дня

day\_charge — сколько пользователь заплатил за свою дневную активность

eve\_calls, eve\_mins, eve\_charge — аналогичные метрики относительно вечерней активности

night\_calls, night\_mins, night\_charge — аналогичные метрики относительно ночной активности

intl\_calls, intl\_mins, intl\_charge — аналогичные метрики относительно международного общения

custserv\_calls — сколько раз пользователь позвонил в службу поддержки

treatment — номер стратегии, которая применялись для удержания абонентов (0, 2 = два разных типа воздействия, 1 = контрольная группа)

mes\_estim — оценка интенсивности пользования интернет мессенджерами

churn — результат оттока: перестал ли абонент пользоваться услугами оператора

Давайте рассмотрим всех пользователей из контрольной группы (treatment = 1). Для таких пользователей мы хотим проверить гипотезу о том, что штат абонента не влияет на то, перестанет ли абонент пользоваться услугами оператора. Постройте таблицы сопряженности между каждой из всех 1275 возможных неупорядоченных пар штатов и значением признака churn. Заметьте, что, например, (AZ, HI) и (HI, AZ) — это одна и та же пара. Какой критерий подходит для решения этой задачи? Сколько достигаемых уровней значимости оказались меньше, чем α=0.05?

Посчитайте корреляции Пирсона и Спирмена между day\_calls и mes\_estim на всех данных, оцените их значимость, дайте интерпретацию результата.

Посчитайте значение коэффицента корреляции Крамера между штатом (state) и оттоком пользователей (churn) для всех пользователей, которые находились в контрольной группе (treatment=1). Проверьте гипотезу об отсутствии связи между этими признаками.

Проведите анализ эффективности удержания (churn) с помощью раличных методов (treatment = 0, treatment = 2) относительно контрольной группы пользователей (treatment = 1). Что можно сказать об этих двух методах (treatment = 0, treatment = 2)? Одинаковы ли они с точки зрения эффективности? Каким бы методом вы бы посоветовали воспользоваться компании?