Домашнее задание 4
Deadline: 23.59 10 марта 2020

Общая постановка задачи Задание выполняется на массиве CDAhw2.dta. Описание переменных представлено в этом файле. Вам необходимо смоделировать участие в голосовании на выборах 4 декабря 2011 года. Используйте из предложенного массива в качестве отклика вопрос со следующей формулировкой: q29 — Принимали ли вы участие в выборах в Государственную Думу России 4 декабря этого года?
В рамках домашнего задания 2 Вы уже оценивали логит-модель с самостоятельно выбранным набором предикторов. Используйте эту модель как основную для выполнения последующих заданий.

1. Проверьте гипотезу о согласии наблюдаемых значений (данных) и модели посредством Hosmer-Lemeshow goodness-of-fit test. Кратко поясните, каким образом устроена статистика теста, сделайте вывод на основании полученных результатов.

2. Рассчитайте значение baseline accuracy.

3. Сохраните предсказанные вероятности участия в выборах 4 декабря 2011 г., задайте сами порог отсечения и представьте в качестве результата confusion matrix. По представленной таблице классификации рассчитайте ошибку первого рода, ошибку второго рода и мощность (запишите в явном виде, как рассчитываются эти значения, объясните, что они содержательно показывают в контексте поставленной содержательной задачи в этом домашнем задании). Можно ли говорить о значимых различиях по сравнению с baseline accuracy? Прокомментируйте результаты.

4. Сравните решение (выбранный Вами порог отсечения) с классификацией наблюдений в соответствии с

   (a) оптимальным порогом, выбранным на основе минимизации ошибки классификации

   (b) оптимальным порогом, выбранным в соответствии с Youden index (максимизация $Sensitivity + Specificity - 1$)

   Примечание: сравните по ряду мер: accuracy, специфичность, чувствительность, precision (точность). Сделайте вывод: какая модель из представленных, на Ваш взгляд, наиболее удачная с точки зрения классификации?

5. Проверьте массив данных на наличие

   (a) outliers

   (b) leverage

   (c) влиятельных наблюдений (influential observations). Переоцените Вашу модель на массиве без пяти наблюдений с самим высоким значением меры Кука. Привело ли это к изменениям в оценках коэффициентов? Можно ли оставить эти наблюдения в массиве?

6. Предложите самостоятельно две спецификации невложенных моделей. Протестируйте посредством информационных критериев, какая модель является предпочтительной.