

Зеленым обозначено условие для первого варианта, желтым для второго. Для набора данных CARS вводится дополнительная бинарная переменная Expensive (Cheap), принимающая значение истина, если стоимость машины выше 35000 (ниже 20000).

- 1) Предложите свой вариант комбинации основного и стратифицирующего предикторов, чтобы выполнялось условие, что по тесту хи-квадрат зависимость отклика от основного предиктора была статистически значима (уровень значимости 0.05), а при стратифицированном тесте CMH с участием второго предиктора зависимость от исходного была уже не значима. В качестве предикторов можно предлагать дискретизированные версии непрерывных предикторов (рекомендуется не делать много интервалов дискретизации, в большинстве случаев достаточно будет бинарного разбиения непрерывного предиктора).
- 2) Постройте логистическую регрессию для Cheap (Expensive) с использованием функции step, отбирающую переменные прямым (обратным) методом и перебирающую все «сложности» модели, от одной переменной до всех (или наоборот). Для каждой модели, полученной в ходе перебора, оцените ее качество с помощью кросс-валидации (с 5 блоками) по критерию площадь под ROC кривой (она же статистика согласованности), постройте график зависимости оценки качества (CV ROC AUC) от сложности (количества предикторов) модели и выберете лучшую модель (по сути зафиксируйте список переменных лучшей модели).
- 3) Обучите лучшую модель на всей выборке и на oversampled выборке (пропорция 1 к 1), постройте бутстреп ROC кривые по всей выборке (можно реализовать свой код или воспользоваться функцией boot.roc) для обеих моделей. Существенно ли изменилось качество oversampled модели?