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 дискретезированные версии непрерывных предикторов (рекомендуется не делать много интервалов дискретизации, в большинстве случаев достаточно будет бинарного разбиения непрерывного предиктора).
2) Проведя подстановку пропусков с помощью mice (любым методом из mice) и постройте логистическую регрессию для Expensive (Cheap)с использованием функции step, отбирающую переменные прямым (обратным) методом и перебирающую все «сложности» модели, от одной переменной до всех (или наоборот). Для кодирования категориальных переменных используйте reference (effect) схему. 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) Перекодируйте в исходном наборе данных категориальные переменные в WOE представление и повторите шаг 2. Выберите лучший вариант.
4) Обучите лучшую модель на всей выборке и на undersampled выборке (пропорция 1 к 1), постройте бутсреп ROC кривые по всей выборке (можно реализовать свой код или воспользоваться функцией boot.roc) для обеих моделей. Существенно ли изменилось качество undersampled модели?
5) Для лучшей исходной модели (без undersampling) постройте график доверительных интервалов Odds Ratio для отобранных переменных. Для них же рассчитайте IV, числовые предикторы дискредитируйте по квантилям. Постройте столбчатую диаграмму для IV по отобранным переменным. Как оценки важности переменных по IV и по Odds Ratio отличаются и почему?