Лабораторная работа №2: Линейные модели. Кросс-валидация.

Упражнение 2

1. Данные своего варианта (см. таблицу ниже) разделить на выборку для построения моделей (80%) и отложенные наблюдения (20%). Оставить в таблице только указанные в варианте переменные. Отложенные наблюдения использовать только в задании 6.

2. Провести предварительный анализ данных с помощью описательных статистик и графиков, оценить взаимосвязь.

3. Проверить Y на нормальность. Если он распределён не по нормальному закону, прологарифмировать и снова провести анализ взаимосвязей переменных.

4. Составить список возможных спецификаций моделей множественной регрессии (на исходной Y и на логарифме Y).

5. Оценить параметры моделей из списка. Оценить точность моделей методом перекрёстной проверки, указанным в варианте. Найти самую точную из моделей для Y. Найти самую точную из моделей для $\log(Y)$.

6. Сделать прогноз с помощью самых точных моделей на отложенные наблюдения. Рассчитать MSE_{test} вручную и выбрать одну наиболее точную модель. Проинтерпретировать её параметры.

Варианты

Номер варианта – номер студента в списке. Студент под номером 21 берёт вариант 1, под номером 22 – 2, и т.д.

В качестве ядра генератора случайных чисел (в частности, для разделения данных на выборку для построения моделей и отложенные наблюдения) используйте номер своего варианта.

Наборы данных и справочники к ним выложены по адресу:
https://github.com/ania607/ML/tree/main/data .

Вариант	Проверка	Набор данных	Зависимая переменная	Объясняющие переменные	
				непрерывные	фиктивные
1	LOOCV	Boston_for_lab	medv / медианная стоимость домов, тыс. долл. США	zn / доля земли под жилую застройку indus / доля акров, не связанных с	tax_over_400 / 1 если полная ставка налога на имущество на $10000 превышает 400

				розничной торговлей, на город	
2	K-VAL(10)	Boston_for_lab	medv / медианная стоимость домов, тыс. долл. США	rm / среднее количество комнат в доме nox / концентрация оксидов азота (частей на 10 миллионов)	tax_over_400 / 1 если полная ставка налога на имущество на $10000 превышает 400
3	K-VAL(5)	Boston_for_lab	medv / медианная стоимость домов, тыс. долл. США	rm / среднее количество комнат в доме dis / средневзвешенное расстояние до пяти бостонских центров занятости	tax_over_400 / 1 если полная ставка налога на имущество на $10000 превышает 400
4	LOOCV	Boston_for_lab	medv / медианная стоимость домов, тыс. долл. США	rm / среднее количество комнат в доме indus / доля акров, не связанных с розничной торговлей, на город	tax_over_400 / 1 если полная ставка налога на имущество на $10000 превышает 400
5	K-VAL(10)	Boston_for_lab	medv / медианная стоимость домов, тыс. долл. США	indus / доля акров, не связанных с розничной торговлей, на город crim / уровень преступности на душу населения по району	tax_over_400 / 1 если полная ставка налога на имущество на $10000 превышает 400
6	K-VAL(5)	Auto_for_lab	mpg / пробег автомобиля на галлоне топлива	displacement / объемн двигателя (в кубических дюймах) acceleration / время разгона с 0 до 60 миль в час (в секундах)	cyl_over_4/ 1 если число цилиндров больше 4
7	LOOCV	Auto_for_lab	mpg / пробег автомобиля на галлоне топлива	horsepower / мощность (в лошадиных силах) weight / масса (в фунтах)	cyl_over_4/ 1 если число цилиндров больше 4
8	K-VAL(10)	Auto_for_lab	mpg / пробег автомобиля на галлоне топлива	displacement / объем двигателя (в кубических дюймах)	cyl_over_4/ 1 если число цилиндров больше 4

				weight / масса (в фунтах)	
9	K-VAL(5)	Auto_for_lab	mpg / пробег автомобиля на галлоне топлива	acceleration / время разгона с 0 до 60 миль в час (в секундах) horsepower / мощность (в лошадиных силах)	cyl_over_4/ 1 если число цилиндров больше 4
10	LOOCV	Carseats	Sales / продажа (в тысячах штук) в каждом магазине	Price / цены компании на автокресла в каждом магазине Advertising / бюджет затрат на рекламу в каждом магазине (в тысячах долларов)	ShelveLoc/ качество стеллажа для размещения автокресел в каждом магазине: Плохое, Хорошее и Среднее
11	K-VAL(10)	Carseats	Sales / продажа (в тысячах штук) в каждом магазине	Price / цены компании на автокресла в каждом магазине Population / плотность населения	ShelveLoc/ качество стеллажа для размещения автокресел в каждом магазине: Плохое, Хорошее и Среднее
12	K-VAL(5)	Carseats	Sales / продажа (в тысячах штук) в каждом магазине	Price / цены компании на автокресла в каждом магазине Income / уровень дохода сообщества (в тысячах долларов)	ShelveLoc/ качество стеллажа для размещения автокресел в каждом магазине: Плохое, Хорошее и Среднее
13	LOOCV	Carseats	Sales / продажа (в тысячах штук) в каждом магазине	Price / цены компании на автокресла в каждом магазине CompPrice / цена конкурента в каждом магазине	ShelveLoc/ качество стеллажа для размещения автокресел в каждом магазине: Плохое, Хорошее и Среднее
14	K-VAL(10)	College_for_lab	Grad_Rate / выпускной балл	Top10perc / процент зачисленных студентов, которые в старшей школе относились к топ-10% в диапазоне по освоению	Private / частный или государственный университет: Да - частный, Нет - государственный

				F_Undergrad / количество студенческой формы обучения	
15	K-VAL(5)	College_for_lab	Grad_Rate / выпускной балл	F_ Top25perc / процент зачисленных студентов, которые в старшей школе относились к топ-25% в диапазоне по освоению F_Undergrad / количество студенческой формы обучения	Private / частный или государственный университет: Да - частный, Нет - государственный
16	LOOCV	College_for_lab	Grad_Rate / выпускной балл	Accept / количество реализованных заявок на поступление Expend / расходы на обучение на одного студента	Private / частный или государственный университет: Да - частный, Нет - государственный
17	K-VAL(10)	College_for_lab	Grad_Rate / выпускной балл	Accept / количество реализованных заявок на поступление Top10perc / процент зачисленных студентов, которые в старшей школе относились к топ-10% в диапазоне по освоению	Private / частный или государственный университет: Да - частный, Нет - государственный
18	K-VAL(5)	College_for_lab	Grad_Rate / выпускной балл	Expend / расходы на обучение на одного студента Top25perc / процент зачисленных студентов, которые в старшей школе относились к топ-25% в диапазоне по освоению	Private / частный или государственный университет: Да - частный, Нет - государственный
19	LOOCV	College_for_lab	Grad_Rate / выпускной балл	Expend / расходы на обучение на одного студента P_Undergrad /	Private / частный или государственный университет: Да - частный, Нет -

				количество студентов, обучающихся по совместительству	государственный
20	K-VAL(10)	College_for_lab	Grad_Rate / выпускной балл	Accept / количество реализованных заявок на поступление F_Undergrad / количество студенческой формы обучения	Private / частный или государственный университет: Да - частный, Нет - государственный