ПСАД. ВМК. Практическое задание №3. Линейная и обобщенная линейная регрессия

Задание

Необходимо сдать jupyter notebook с подробным отчётом по проведённому исследованию, содержащий визуализацию исходных данных, описания и выводы каждого этапа анализа — используемые методы, обоснование их применимости, графики.

Все данные доступны по ссылке.

Часть 1

Воспользуйтесь jupyter notebook с семинара 8. Для моделей 2, 4 и 6 постройте ROC-кривые. С помощью метода бутстреп постройте доверительный интервал для значения AUC каждой из построенных ROC кривых?. Влияет ли качество вашей модели на размер интервала?

Часть 2

1.1 Фёдоров Илья

Для 8416 грибов задано признаковое описание согласно справочнику The Audubon Society Field Guide to North American Mushrooms.

Построить модель вероятности ядовитости гриба, оценить вклад факторов.

Данные: mushroom.csv

1.2 Марков Владислав

1055 химических молекул описаны с помощью 41 признака (число атомов кислорода, нитратных групп, донорных связей с водородом, потенциал ионизации и т.д.); 355 из них биоразложимы.

Какие свойства молекул влияют на их биоразлагаемость?

Данные: biodeg.xlsx

1.3 Осин Дмитрий

Собраны данные мониторинга сейсмической активности в польских угольных шахтах столбовой системы разработки. При сейсмической опасности существует серьёзный риск обрушения; в этом случае необходимо отозвать рабочих или использовать направленные взрывы для нейтрализации напряжения породы. Для каждого измерения известен бинарный индикатор сейсмической опасности — наличия в следующую восьмичасовую смену сейсмических толчков с энергией выше 10^4 Джоулей.

Построить модель сейсмической опасности, дать интерпретацию вклада показателей сейсмической активности.

Данные: seismic.xlsx

1.4 Анна Суходоева

Для 500 участниц исследования Global Longitudinal Study of Osteoporosis in Women (Center for Outcomes Research, the University of Massachusetts/Worcester) измерены возраст, вес, рост, ИМТ, бинарные признаки: курение, индикатор наступления менопаузы до 45 лет, индикатор необходимости помощи при подъёме из сидячего положения, перелом шейки бедра в прошлом (был/не было), перелом шейки бедра у матери (был/не было), а также самостоятельная субъективная оценка вероятности перелома (меньше/такая же/больше, чем у сверстниц). Известно, у кого из участниц в первый год исследования произошёл перелом шейки бедра.

Построить модель вероятности перелома с учётом имеющихся признаков, дать интерпретацию.

Данные: GL0W500.txt

1.5 Бакланова Анастасия

Данные собраны из переписи населения США 1990 года, отчёта ФБР о преступности за 1995 год и опроса сотрудников полиции LEMAS за 1990 год. По 2215 округам собрана статистика преступлений и 125 демографических показателей.

Построить функцию, оценивающую абсолютное число автомобильных краж по демографическим показателям, дать

интерпретацию коэффициентов модели.

Данные: crimes.xlsx

1.6 Веселов Арсений

Полихлорированные дифенилы — органические соединения, активно использовавшиеся в промышленности до 1970 годов, когда была показана их токсичность. Накопление ПХБ в организме приводит к подавлению иммунитета, провоцирует развитие рака, поражений печени, почек, нервной системы, кожи, способствуют развитию детской патологии. Из-за накопления ПХБ в озёрах США некоторые виды рыб в некоторых областях запрещены к употреблению в пищу. Для своевременного обновления таких запретов необходимо периодически проводить мониторинг ПХБ. К сожалению, существует 209 различных разновидностей ПХБ, концентрация каждой из которых измеряется отдельным тестом. Для 69 видов рыбы известны концентрации семи соединений ПХБ (в миллионных долях), а также суммарная концентрация всех разновидностей ПХБ, их токсическая эквивалентность (TEQ) и суммарная токсическая эквивалентность образца, определяемая также вкладом диоксинов и фуранов.

Насколько точно токсичность рыбы можно предсказывать по концентрации только нескольких ПХБ? Концентрации какого минимального количества соединений ПХБ нужно измерить, чтобы достаточно точно предсказать суммарную токсичность, или хотя бы токсичность только совокупности ПХБ?

Данные: pcb.txt

1.7 Девбунова Вилиана

Для 649 учеников старших классов двух португальских школ известны ряд демографических показателей и показателей успеваемости; для каждого студента известны также уровень потребления алкоголя по выходным и будним дням в пятибалльной шкале от очень низкого до очень высокого и финальная оценка по португальскому языку.

Смоделировать финальную оценку как функцию от всех показателей, кроме итоговых оценок по промежуточным семестрам; оценить влияние уровня потребления алкоголя на неё.

Данные: student-por.xlsx

1.8 Деев Александр

Для 247 мужчин и 260 женщин измерены две группы антропометрических показателей - легко измеримые характеристики скелета и обхваты, всего 21 признак. Указаны возраст, пол, вес и рост.

Построить функцию, эффективно оценивающую вес по наименьшему набору признаков; сравнить точность оценки веса при отсутствии информации по обхватам и отсутствии информации по характеристикам скелета.

Данные: body.xlsx

1.9 Корнилов Максим

Для 30000 клиентов тайваньского банка известны сумма кредита, демографические показатели и история платежей по кредитам за последние пять месяцев (факт просрочки, сумма неоходимой выплаты, сумма платежа) .

Построить модель, предсказывающую вероятность просрочки следующего платежа, оценить вклад факторов.

Данные: default.xls

1.10 Кузьмин Никита

Госпиталь города Карайкуди, Тамилнад, Индия, собрал данные анализов 250 пациентов с хронической болезнью почек и 150 пациентов без неё.

Построить диагностическую модель хронической болезни почек, оценить вклад факторов.

Данные: chronic_kidney_disease .xlsx

1.11 Мишустина Маргарита

Мерой надёжности шарикоподшипников служит величина Lw — максимальное число оборотов, которое выдерживает 90% одинаковых подшипников. Имеются данные измерений надёжности по шарикоподшипникам трёх производителей (для одного из производителей исследовано три вида подшипников), для каждого испытания указаны диаметр и число шаров в подшипнике, нагрузка и величина Lw -

Построить функцию, оценивающую Lw по имеющимся признакам, оценить вклад признаков.

Данные: bearing.xlsx

1.12 Ниничук Марина

Имеются результаты обзвона 4119 клиентов португальского банка, которым предлагалось завести депозит. Известны социально-демографические характеристики клиентов, история предыдущих коммуникаций, социально- экономические показатели на момент совершения звонка.

Какие признаки определяют готовность клиента открыть депозит по результатам обзвона?

Данные: deposit.xlsx

1.13 Новосёлов Вадим

Изучалось влияние внешних характеристик самок морских ракообразных мечехвостов (1) на их привлекательность для самцов. Выборка состоит из данных о наблюдениях над 173 особями и содержит закодированные данные о размере самок, их весе, цвете, состоянии панциря, а также о количестве спутников.

Построить функцию, по внешним параметрам самки предсказывающую количество спутников у самки, оценить значимость каждого фактора.

Данные: horseshoe crab.txt

1.14 Пойманов Дмитрий

Массовая доля жира, важная характеристика здоровья, рассчитывается через плотность тела, измеряемую при помощи взвешивания в воде. Для 252 мужчин проведены такие расчёты. Имеются также данные антропометрии (возраст, рост, вес, обхват грудной клетки и т.д.)

Построить функцию, оценивающую массовую долю жира по легко измеряемым антропометрическим признакам.

Данные: fat.xls

1.15 Сенин Александр

Имеется 1066 наблюдений над различными участками поверхности Солнца. Известны: класс участка, размер максимального пятна на участке, распределение пятен, относительная активность, тип эволюции участка, код активности в предыдущие 24 часа, площадь участка. Известны также сложность участка в наблюдавшемся прошлом и при последнем повороте вокруг Солнца. Известно также число вспышек на каждом участке в течение 24 часов после начала наблюдения, причём вспышки разделены на три категории по мощности.

Построить модель, по свойствам участка предсказывающую суммарное число вспышек в последующие 24 часа, дать интерпретацию коэффициентов.

Данные: solar flares.xls

1.16 Травникова Арина

Для 60021 постов в блогах, опубликованных не более, чем за 72 часа до базового времени, собрана информация о количестве комментариев, времени публикации, длине и количестве каждого из 200 часто встречающихся слов.

Построить модель, предсказывающую количество новых комментариев за следующие 24 часа.

Данные: blog_feedback.xlsx

1.17 Трошков Артем

Имеются результаты обработки 1147 изображений сетчаток. По изображениям рассчитаны значения 17 признаков; записаны также результаты предварительного скрининга на наличие диабетической ретинопатии и окончательный диагноз.

Построить модель, оценивающую вероятность наличия диабетической ретинопатии, дать интерпретацию коэффициентов.

Данные: retinopathy .xlsx

1.18 Ушаков Даниил

Данные собраны из переписи населения США 1990 года, отчёта ФБР о преступности за 1995 год и опроса сотрудников полиции LEMAS за 1990 год. По 2215 округам собрана статистика преступлений и 125 демографических показателей.

Построить функцию, оценивающую число ненасильственных преступлений на сто тысяч населения по демографическим показателям, дать интерпретацию коэффициентов модели.

Данные: crimes.xlsx

1.19 Шарыпов Руслан

Имеются данные по продажам 479 платьев на сайте aliexpress.com за полтора месяца осени 2013 года. Для каждого из платьев

известны также стиль, ценовая категория, рейтинг, размер, сезон, ряд характеристик внешнего вида и индикатор участия в программе рекомендаций.

Оценить влияние рейтинга товаров на продажи с учётом остальных факторов.

Данные: aliexpress_dress_data. csv

1.20 Швец ПАВЕЛ

Имеется выборка из 1009 детей, родившихся в Северной Каролине в 2004 году; известны пол ребёнка, вес при рождении, период вынашивания, возрастная группа матери, а также курила ли мать во время беременности и употребляла ли алкоголь

Как вес ребёнка зависит от курения и употребления алкоголя (после поправки на остальные признаки)?

Данные: birthweight. csv

1.21 Шибанов Алексей

Имеются данные использования городского велопроката Вашингтона за каждый день 2011-2012 годов; известны также данные о погоде и ряд календарных признаков

Построить модель использования велопроката в зависимости от имеющихся признаков. Достаточно ли использовать дату с точностью до сезона, или месяц позволяет предсказывать значение признака значимо лучше? Есть ли смысл в использовании полной информации о днях недели, или достаточно разделять выходные и рабочие дни?

Данные: bike shares. xls