1. Придумайте и опишите наибольшее число вариантов использования этих данных. Для решения каких практических задач их можно применить?
2. Прогнозирование цены автомобиля в зависимости от комплектации. Полезно при выводе нового продукта на рынок, помогает с определением цены.
3. Выдача информации по автомобилю, API. Можно фильтровать подходящие пользователю автомобили по выбранным им параметрам.
4. Разведочный анализ данных (выведение общих понятных закономерностей) полезен для понимания общего устройства спецификации автомобилей. Например, человек при выборе автомобиля будет знать как важные для него параметры влияют друг на друга. К примеру, человек, видя большой расход топлива, который ему не подходит можно начать искать автомобиль полегче. Он сделал такой вывод на основании прочитанной ранее информационной заметке.
5. Возможно создание рекомендательной системы. Например, сайт по продаже автомобилей сможет рекомендовать автомобили, которые посетителю также стоит посмотреть, основываясь на его предыдущих просмотрах.
6. Score – риск-скор, сформированный для задач страхования. Соответственно с помощью этих данных страховые компании оценивают риск управления автомобилем.
7. Можно также предсказывать риски управления автомобилем для водителя. Этот же скор позволит помочь человеку прикинуть сумму страховки и определить совокупную стоимость владения машиной.
8. Использование для других моделей. Сгенерированный риск-скор можно отдать банку, чтобы тот сформировал для своего клиента предложение по автокредиту.
9. Кластеризация машин позволит выделить группы схожих автомобилей, изучить их особенности и построить для каждой группы отдельную модель. Это полезно при разработке отдельных маркетинговых стратегий: можно устроить акцию для определенной группы автомобилей.
10. С помощью имеющихся данных можно достроить недостающие. Например, примерно восстановить bore и stroke для некоторых машин, опираясь на значения максимально похожих на них автомобилей.
11. Выберите одну из описанных вами задач и реализуйте ее решение на Python/R при помощи оптимальной на ваш взгляд модели. Опишите, почему вы выбрали именно эту модель.

Я выбрала 8 задачу по кластеризации и сделала ее на Pyhton. Пожалуйста смотрите в прикрепленных файлах Cars\_doc+.ipynb. Я выбрала k-means как наиболее простой метод кластеризации. Несмотря на то, что число кластеров нужно знать заранее, я подобрала его методом плеча. Так как k-means чувствителен к начальному значению, то я использовала модернизированный алгоритм k-means++.

1. Опишите ваш подход к предварительному анализу (в т.ч. визуализации) и обработке данных, работе с признаками, кросс валидацией, настройкой модели и ее оценкой. Объясните выбор того или иного решения и, по возможности, сравните с альтернативами.

В самом начале я заменила ячейки данных с «?» на среднее значение по числовой переменной. Я не стала эти данные выкидывать, так как их и так мало. Так как мы не может заменить на среднее категориальные признаки, то я выкинула несколько строк из данных по переменной «количество дверей». Вcе числовые признаки, где это возможно, приведены к типу int для экономии памяти.

Далее я визуализировала числовые переменные диаграммой ящик с усами для того, чтобы были видны выбросы. Так как k-means достаточно чувствителен к выбросам, то их стоило бы почистить. Так как данных достаточно мало, я решила сделать это только для переменной оборотов, где небольшое их количество было ярко выражено.

Далее я пыталась оценить предположения, которые должны выполняться, для того чтобы k-means хорошо работал: кластеры сферические, примерно одинаковое количество наблюдений в каждом кластере.

Я также построила корреляционную матрицу, чтобы посмотреть взаимосвязь между переменными, как и еще несколько других графиков.

После я перевела факторные переменные в дамми-переменные, так как алгоритм работает с числами. Соответственно, изначальные переменные, которые преобразованы, были удалены.

После чего я построила k-means, варьируя количество кластеров. Финальный набор кластеров 3 был выбран исходя из результатов метода плеча: баланс SSE и количества кластеров. После чего я визуализовала кластеры для наглядности по переменным длины, ширины и высоты автомобиля.

1. Опишите, как бы вы улучшили модель из пункта 2, если бы у вас было больше времени и существенно больший датасет.

Я бы попробовала использовать KNN для замены «?», что дало бы большую правдоподобность данным. Я бы более смело избавлялась от выбросов. Потом я бы сократила размерность (PCA), чтобы ускорить вычисление кластеров. Я бы также попробовала иерархическую кластеризацию и GMM (на переменных, подчиняющихся нормальному распределению).