

Постановка задачи:

Решить задачу классификации автомобилей по типу привода

Dataset:

93 Cars on Sale in the USA in 1993

<http://www.amstat.org/publications/jse/v1n1/datasets.lock.html>

Первичный анализ данных:

1. Количество классов: 3 ('Front', 'Rear', '4WD')
2. Количество примеров: 93
3. Количество признаков до преобразования: 26.

Обучающая выборка имеет маленький размер и внушительное для этого размера количество признаков.

Возможные приёмы работы:

- Уменьшение количества сильно коррелирующих между собой признаков в случае их наличия.
 - Уменьшение размерности (dimensionality reduction)
 - Аугментация данных (data augmentation)
4. Из 26 признаков 8 - качественные (categorical), 18 - количественные (numerical). Визуально "Cylinders" - количественный признак. Однако, он принимает малое число значений, поэтому его можно оставить в списке качественных. Векторизуем качественные признаки.
 5. Выборка не сбалансирована (imbalanced): 67/16/10. Это будет необходимо учесть при оценке точности модели.
 6. Значения многих признаков сильно коррелируют между собой:

Коэффициенты корреляции Пирсона:

> 0.9	{Price, Min.Price, Max.Price} {Fuel.tank.capacity, Weight}
>0.8	{Price, Horsepower}, {Engine.size, Weight} {Fuel.tank.capacity, Wheelbase} {Weight, MPG.city}
>0.6	{Price, Fuel.tank.capacity} {Price, Weight} {Price, MPG.city} {Price, Engine.size} {Engine.size, Rev.per.time} {Engine.size, Length} and most of their combinations

7. Визуально, в выборке присутствуют много outliers по множеству направлений
8. Также визуальный анализ показывает, что наиболее информативные признаки:
 - Price
 - MPG.city
 - MPG.highway
 - Rev.per.mile
 - Length
 - Weight

Предобработка:

1. Заполнение пропущенных значений (количественные признаки восстановлены средними значения по выборке, качественные - самыми часто встречающимися)
2. Нормализация количественных признаков
3. Замена качественных признаков на бинарные
4. Выполнение PCA (уменьшения размерности данных методом главных компонент) с разным числом компонент и сравним точности.

Построение классификатора:

Классификатор	Точность, без отбора признаков	Точность, PCA	Точность, удаление сильно коррелирующих признаков	Значение выбранного параметра
kNN (scikit-learn)	0.72 (+/- 0.05)	0.74 (+/- 0.07)	0.72 (+/- 0.05)	Количество соседей: 5, 11
Decision tree (scikit-learn)	0.73 (+/- 0.17)	0.72 (+/- 0.14)	0.57 (+/- 0.25)	
Random forest (scikit-learn)	0.76 (+/- 0.07)	0.69 (+/- 0.13)	0.71 (+/- 0.16)	
Gradient boosting trees (scikit-learn)	0.80 (+/- 0.19)	0.69 (+/- 0.08)	0.64 (+/- 0.24)	

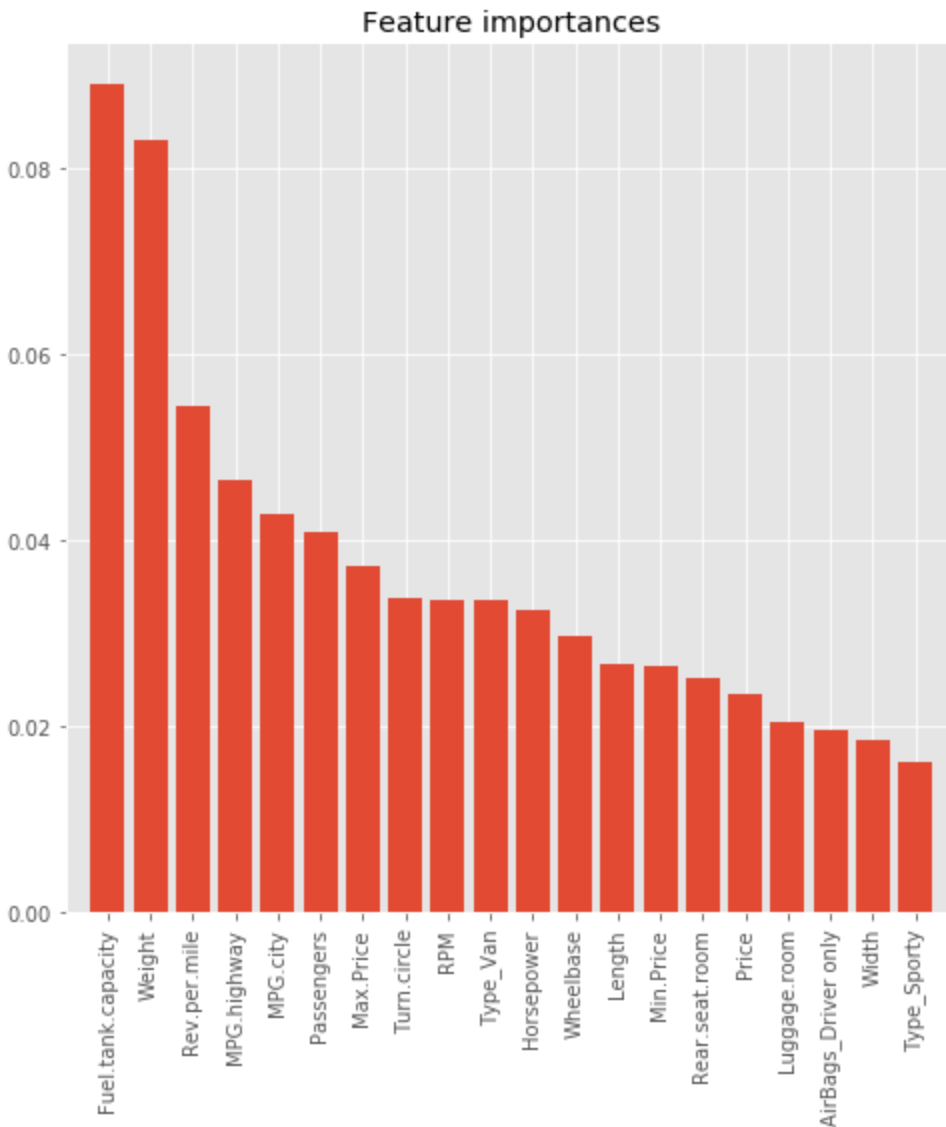
Отбор признаков (feature selection):

После PCA значимость основных признаков в Random forest следующая:

- 'Manufacturer' - 0.5195
- 'Model' - 0.4805

Наибольшую значимость среди всех признаков (без использования PCA) имеют:

- 'Fuel.tank.capacity' - 0.0890
- 'Weight' - 0.0831



Использование данных признаков без использования PCA дало сравнимые результаты (0.73 на Random forest)

Оценка точности:

Оценка точности производилась методом стратифицированной k-fold кросс-валидации с перемешиванием. Количество разбиений - 2.