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 примеров: 93
2. Количество признаков: 26.

Обучающая выборка имеет маленький размер и внушительное для этого размера количество признаков.
Варианты решения проблемы:

- Уменьшение количества сильно коррелирующих между собой признаков в случае их наличия.
- Уменьшение размерности (dimensionality reduction)
- Аугментация данных (data augmentation)

3. Из 26 признаков 8 - качесвтенные (categorical), 18 - количественные (numerical). Визуально “Cylinders” - количественный признак. Однако, он принимает малое число значений, поэтому его можно оставить в списке качественных.
4. Количество классов: 3 ('Front', 'Rear', '4WD')
5. Выборка несбалансирована (imbalanced data): 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	Значение подобранного параметра
kNN (scikit-learn)	0.74 (+/- 0.07) Без PCA: 0.65 (+/- 0.07)	Количество соседей: 5, 11
Decision tree (scikit-learn)	0.73 (+/- 0.17)	
Random forest (scikit-learn)	0.73 (+/- 0.17)	
Gradient boosting trees (scikit-learn)	0.74 (+/- 0.22)	

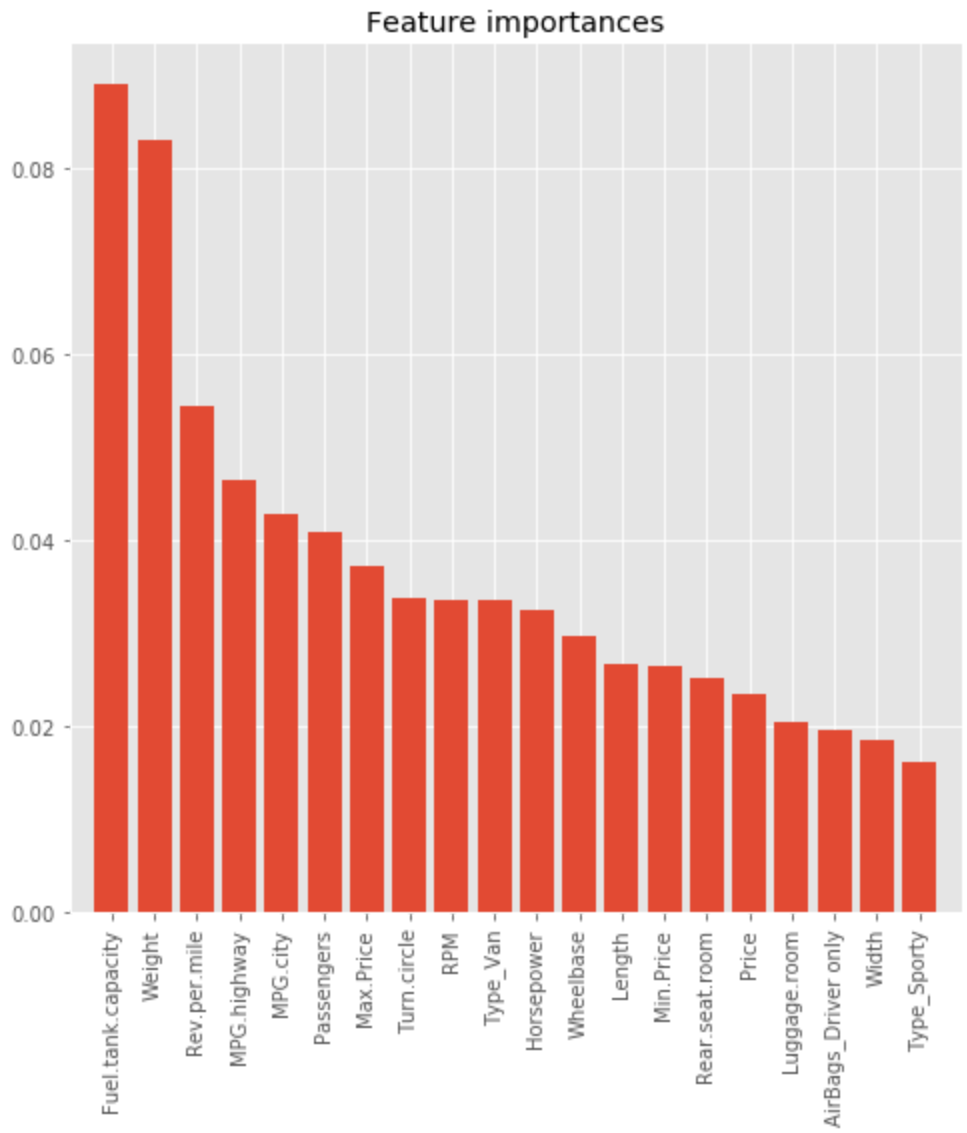
Отбор признаков (feature selection):

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.