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.

Отбор признаков (feature selection):

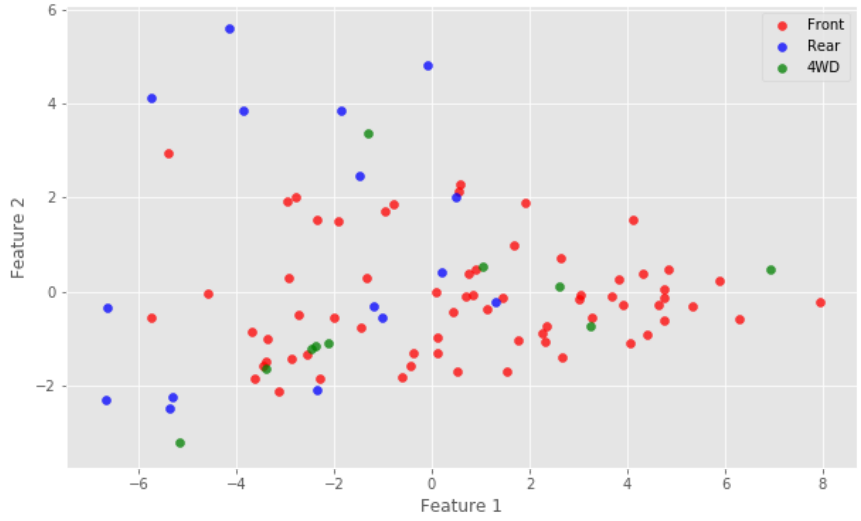
После PCA значимость основных признаков в Random forest следующая:

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

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

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

После применения процедуры двухкомпонентного PCA количество признаков уменьшилось до двух. Ниже изображено распределение по классам в зависимости от значений признаков:



Визуально заметно, что данные частично кластеризуются по двум данным признакам.

Оценка точности:

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

Построение классификатора. Результаты и выводы:

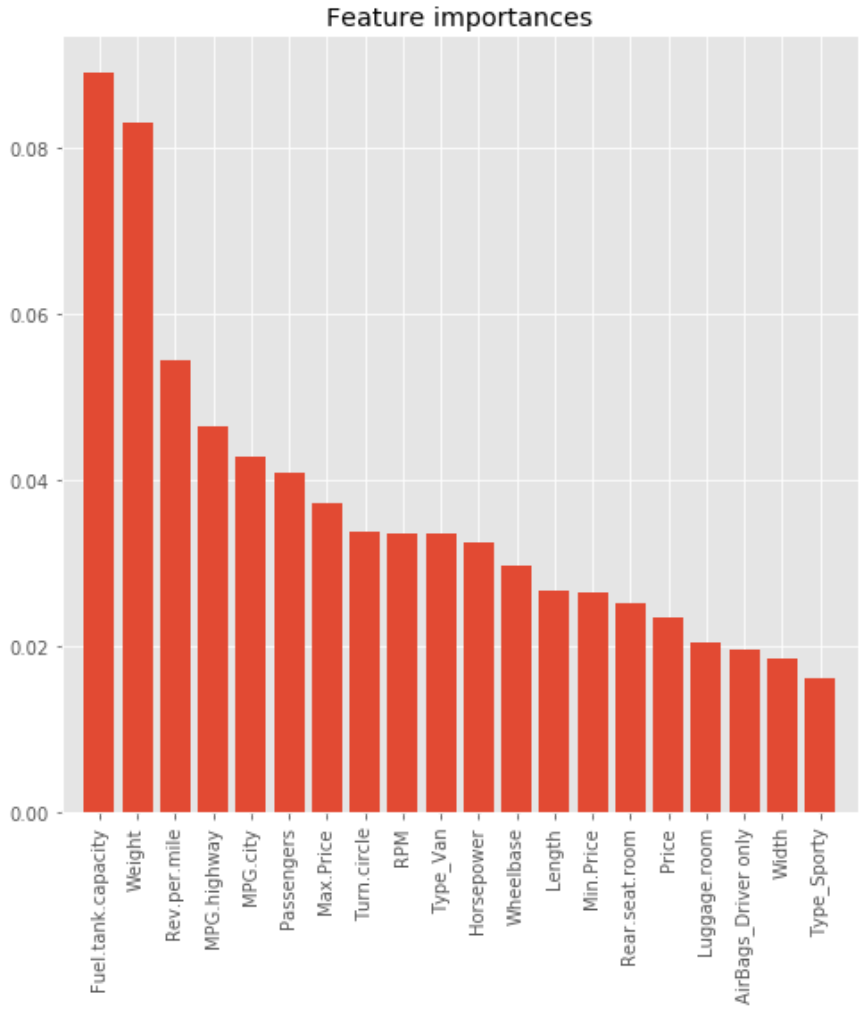
Классификатор	**Точность,** без отбора признаков	**Точность,** PCA	**Точность,** удаление сильно коррелирующих признаков	**Значение подобранного параметра**
kNN (scikit-learn)	0.72 (+/- 0.05)	0.74 (+/- 0.07)	0.72 (+/- 0.05)	Количество соседей: 5, 11
Decision tree (scikit-learn)	0.80 (+/- 0.19)	0.72 (+/- 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)	
SVM (scikit-learn)	0.72 (+/- 0.05)	0.72 (+/- 0.05)	0.72 (+/- 0.05)	
Extreme learning machines (scikit-learn extension)	0.72 (+/- 0.05)	0.72 (+/- 0.05)	0.68 (+/- 0.05)	
MLP (Tensorflow)	0.86	0.78	0.71	200 + 50, tanh, softmax epochs: 5, L2-reg.

Лучший результат: 80% для Decision tree и Gradient boosting trees без отбора признаков.

Confusion matrix для Decision tree:

	Front	*Rear*	*4WD*
Front	33,3%	66,6%	0%
Rear	9,5%	71,4%	19%
4WD	0%	0%	100%

- Confusion matrix для остальных классификаторов почти не отличаются.
- Высокая точность для класса 4WD обусловлена тем, что количество тестовых примеров для этого класса (собственно, как и общее количество примеров), очень невелико.
- Как видно, класс Front выделить гораздо сложнее.



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

Data Augmentation

Т.к. количество данных небольшое и пропорция примеров типа “4WD” критически мала, были сгенерированы новые данные с помощью imbalanced-learn API.
Точность классификации улучшилась следующим образом:

	Naive random oversampling	*Over-sample minority class (‘4WD’) by SMOTE*	*Over-sample minority class (‘4WD’) by ADASYN*
GBT	0.90 (+/- 0.16)	0.90 (+/- 0.06)	0.81 (+/- 0.19)
MLP	0.92	0.93	0.89

Confusion matrix для GBT. Количество примеров для каждого класса - 67.

33	3	0
0	27	0
0	3	35

Итого, **максимальная** достигнутая точность - 93.4%