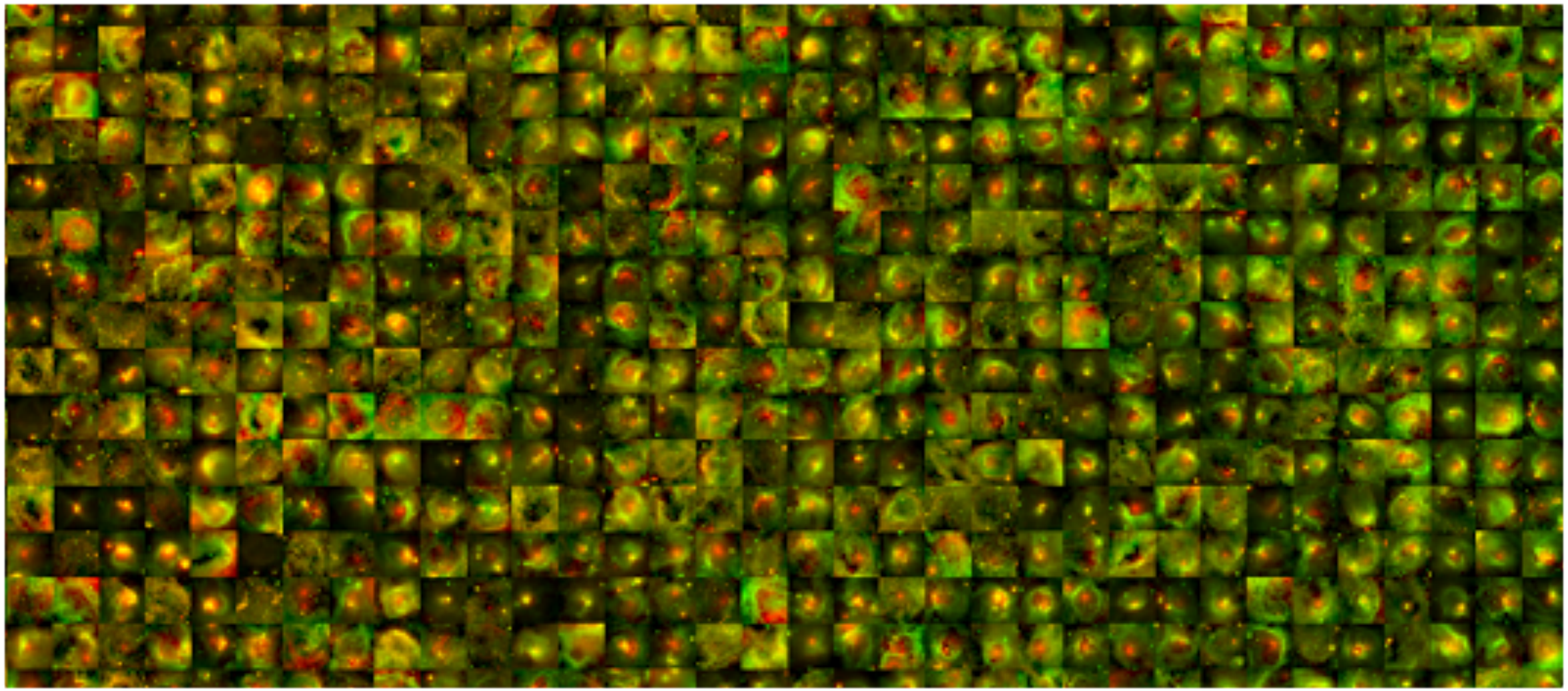
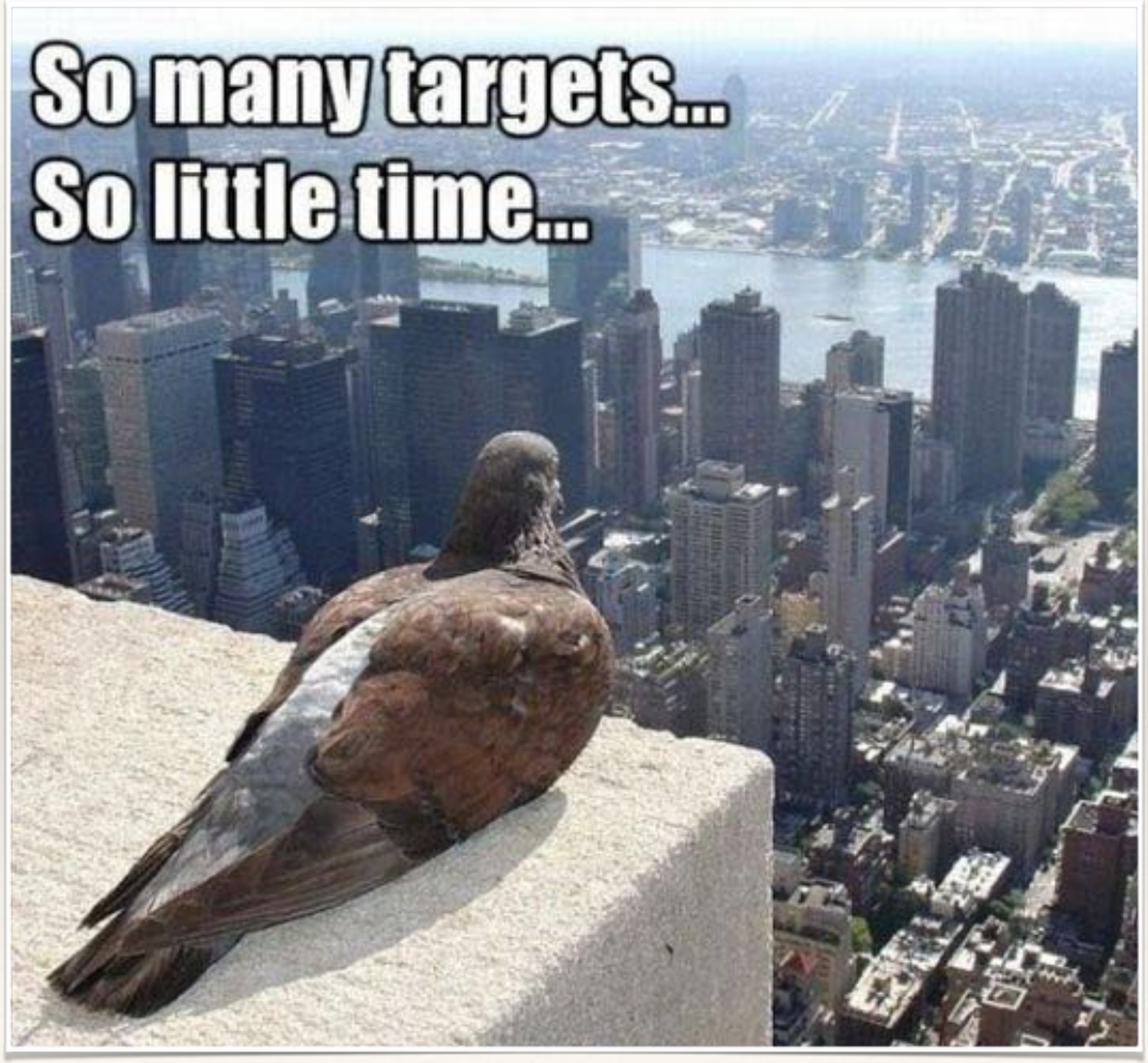


Факультет Компьютерных Наук ВШЭ

Практический анализ данных и машинное обучение

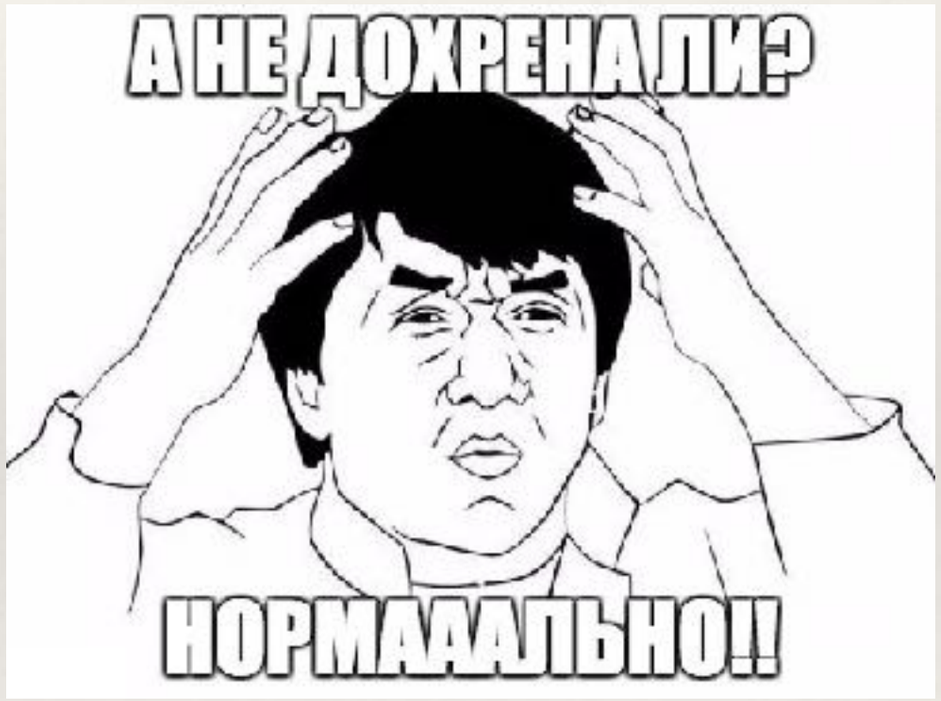
Кашницкий Юрий

Что нас ждет



Обзор курса

- 10 занятий
- Основные алгоритмы и их использование
- Домашние задания и практики
- Соревнование Kaggle Inclass
- Индивидуальные проекты

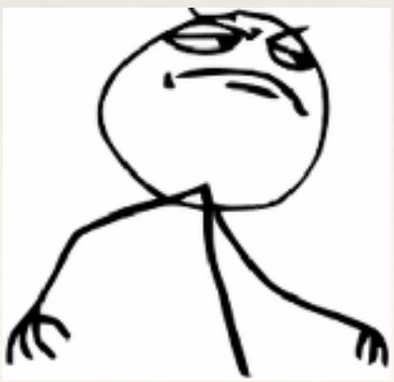


Особенности курса

- Обилие практики - задания на каждом занятии и после него
- Четкие инструкции к заданиям (макет в виде тетрадки Jupyter)
- Понимание теоретических основ алгоритмов
- Знакомство с платформой Kaggle
- В основе всего - свой собственный проект

Логистика

- Все вопросы, все общение - в форуме Piazza
- За домашние задания max 10 баллов
- За проект и соревнование - max 30 баллов
- Текущий рейтинг будет тут
- Нужны аккаунты GitHub и Kaggle
- Все материалы курса - в проекте на GitHub
- Победителю - бонус в карму!



Prerequisites

- Минимальное владение Python
- Основы аналитической геометрии (вектора, матрицы, скалярное произведение и т.д.)
- Основы математического анализа (производные, пределы, интегралы)
- Теория вероятностей и статистика (этого будет не много)

Инструменты

- Язык Python
- Jupyter notebooks
- Сборка библиотек Anaconda
- GitHub
- Kaggle
- Виртуальная машина Vagrant (опц.) с Xgboost, Vowpal Wabbit и т.д.

Anaconda

kaggle

Занятие 1

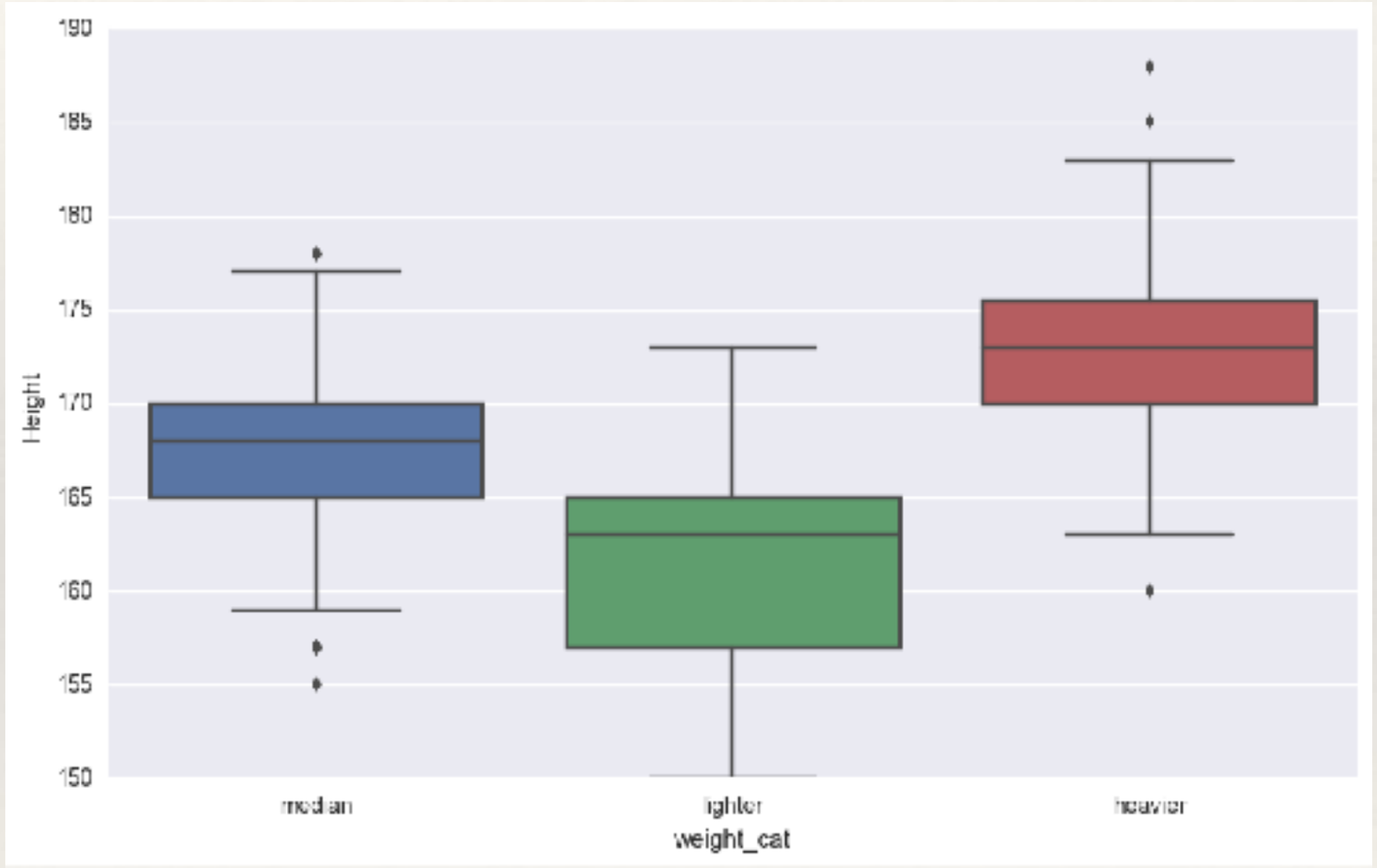
- Анализ данных с Pandas
- Практика на знакомство с данными
- ДЗ № 1. Анализ демографических данных по жителям США

```
df.head(4)
```

	wage	exper	union	goodhlth	black	female	married
0	5.73	30	0	1	0	1	1
1	4.28	28	0	1	0	1	1
2	7.96	35	0	1	0	1	0
3	11.57	38	0	1	0	0	1

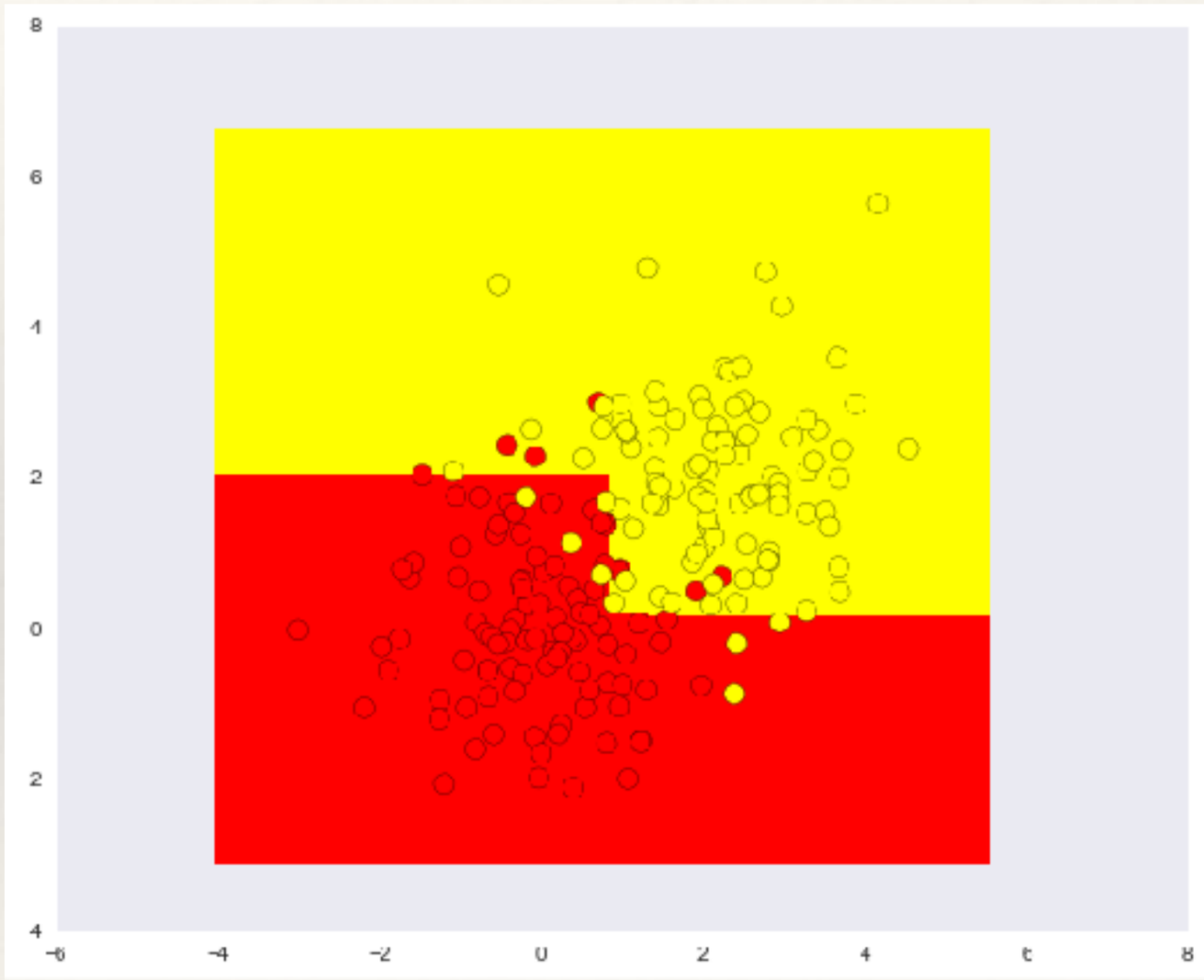
Занятие 2

- Визуальный анализ данных с Pandas и Seaborn
- Практика на «рисование»
- Мастер-класс: 1 часть проекта
- ДЗ № 2. Анализ данных по перелетам между городами США в 2008 году



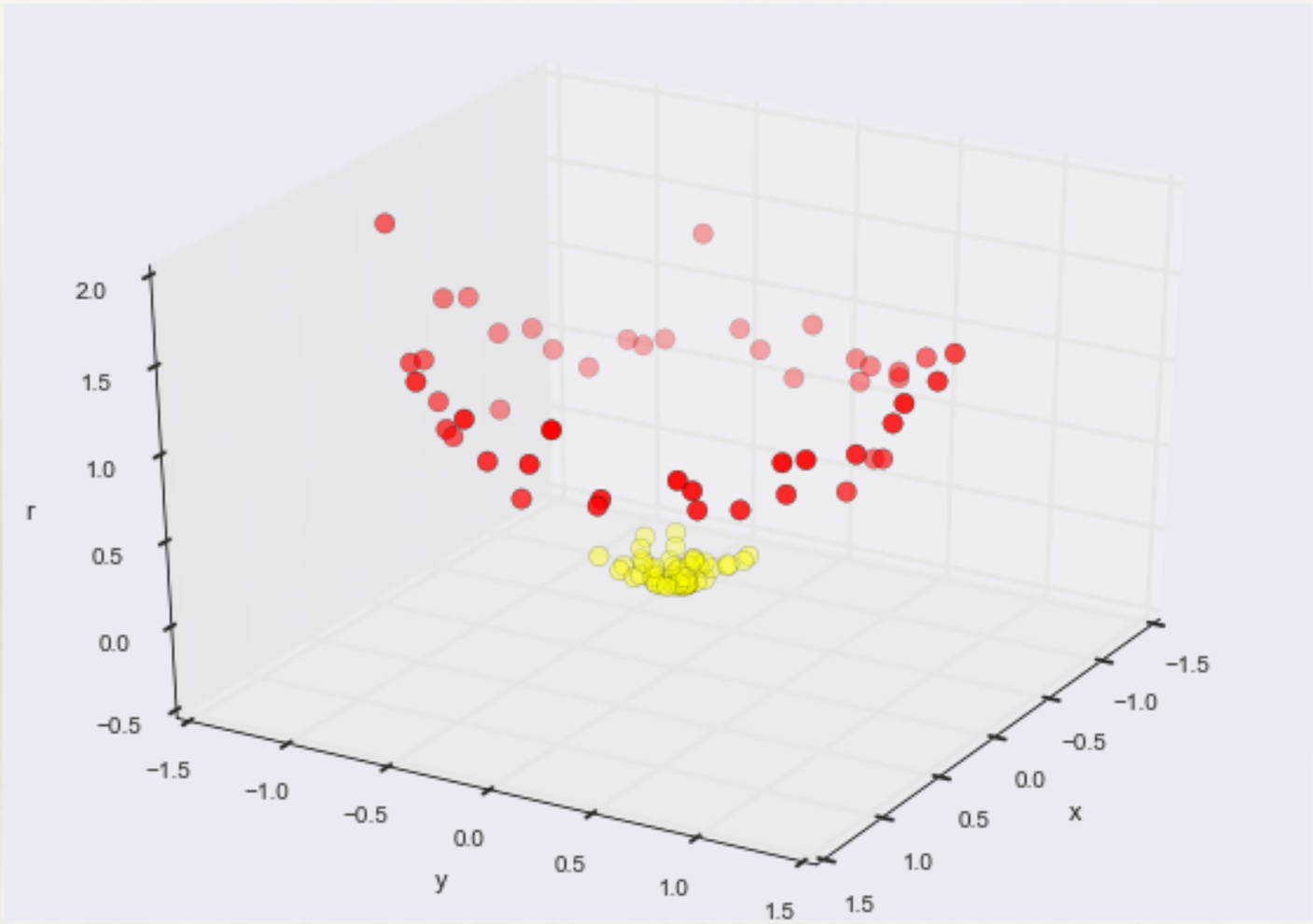
Занятие 3

- Основы машинного обучения
- Деревья решений
- Практика на знакомство с библиотекой Scikit-learn
- ДЗ № 3. Деревья решений и кредитный скоринг



Занятие 4

- *Линейные модели:* логистическая регрессия, метод опорных векторов
- Регуляризация
- Практика на применение logit
- ДЗ № 4. Сравнение нескольких алгоритмов классификации

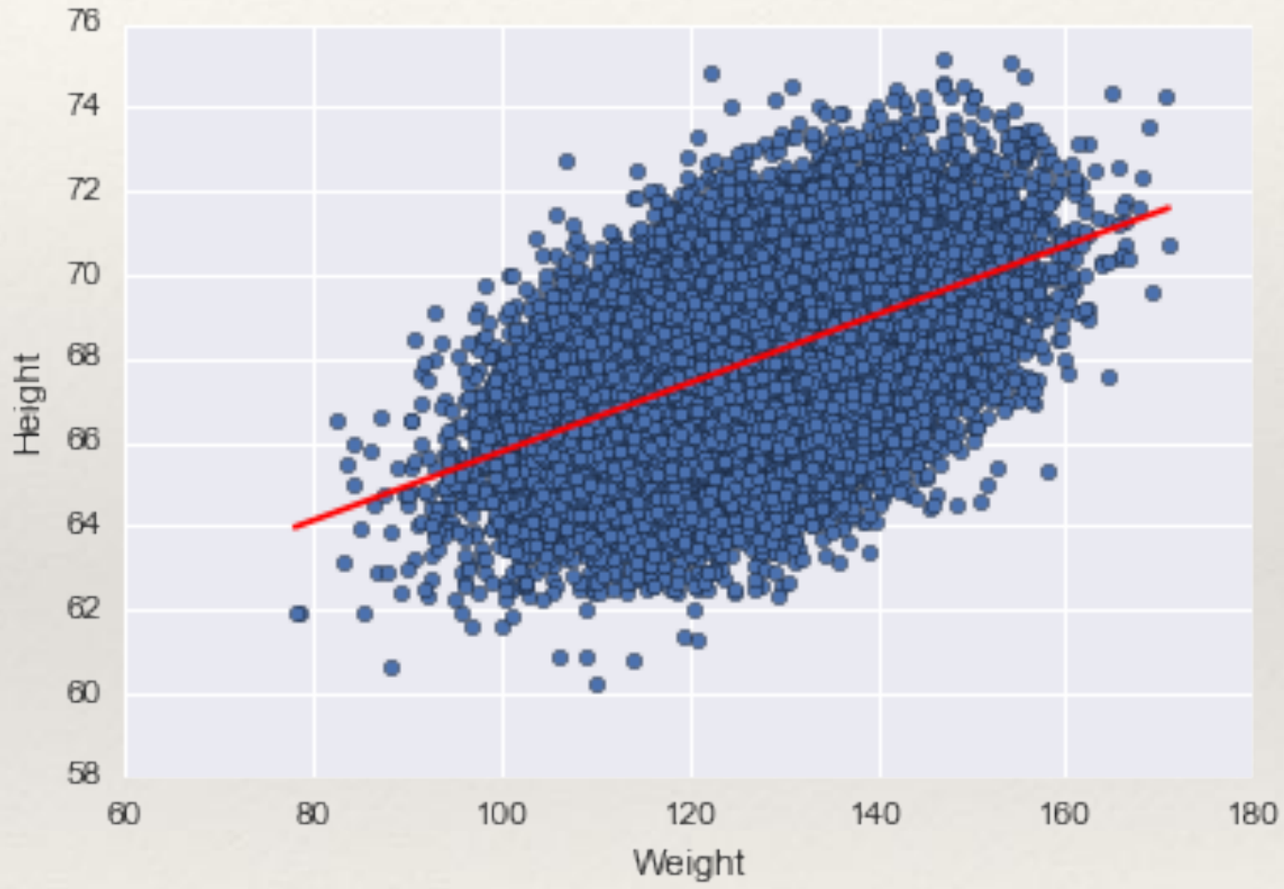


Занятие 5

- Композиции алгоритмов, случайный лес
- Мастер-класс: 2 часть проекта
- Практика на применение случайного леса и оценке важности признаков
- ДЗ № 5. 1 часть проекта

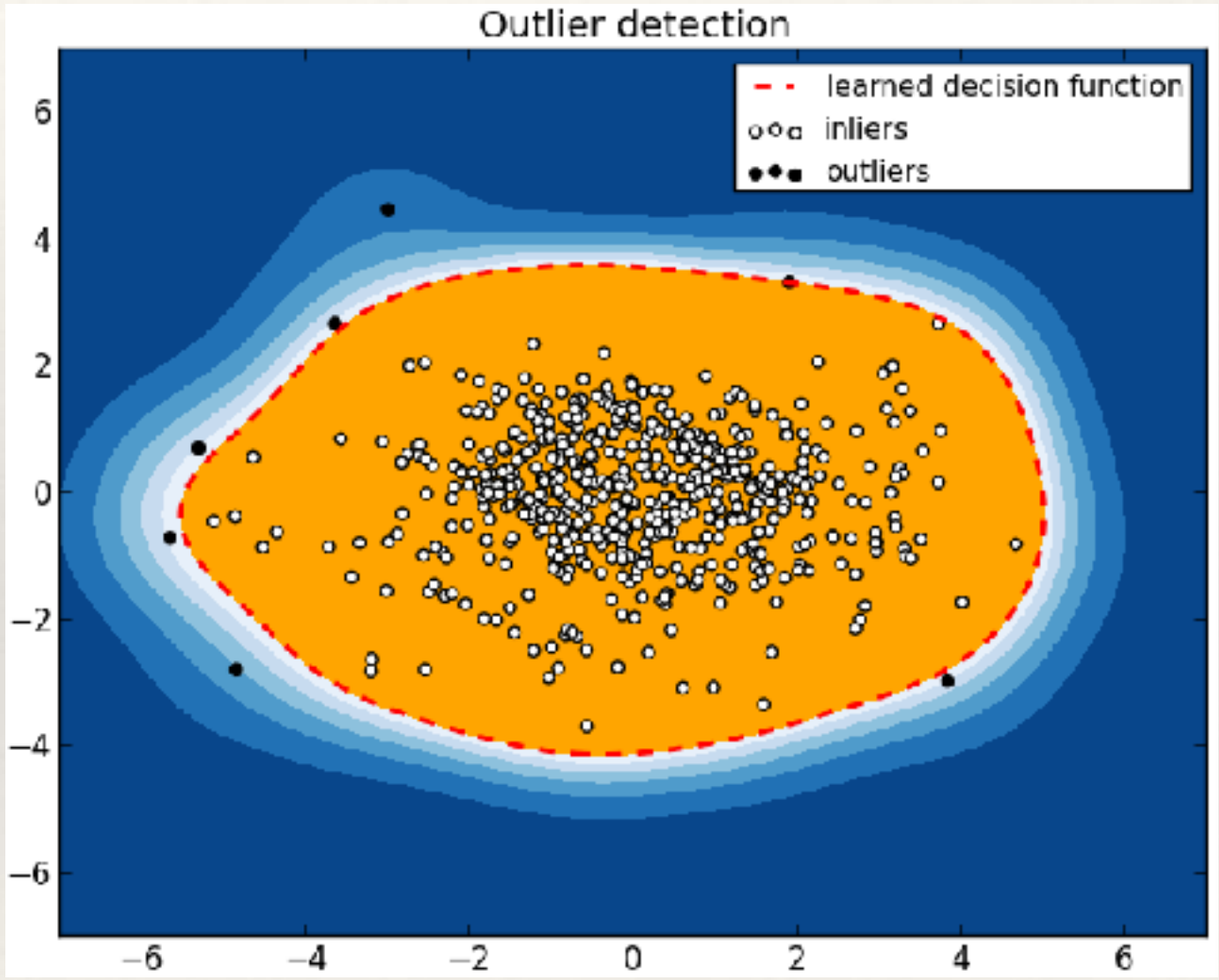
Занятие 6

- Задача регрессии, Lasso, Ridge, случайный лес
- Практика на понимание основ линейной регрессии
- ДЗ № 6. Отбор признаков с Lasso-регрессией



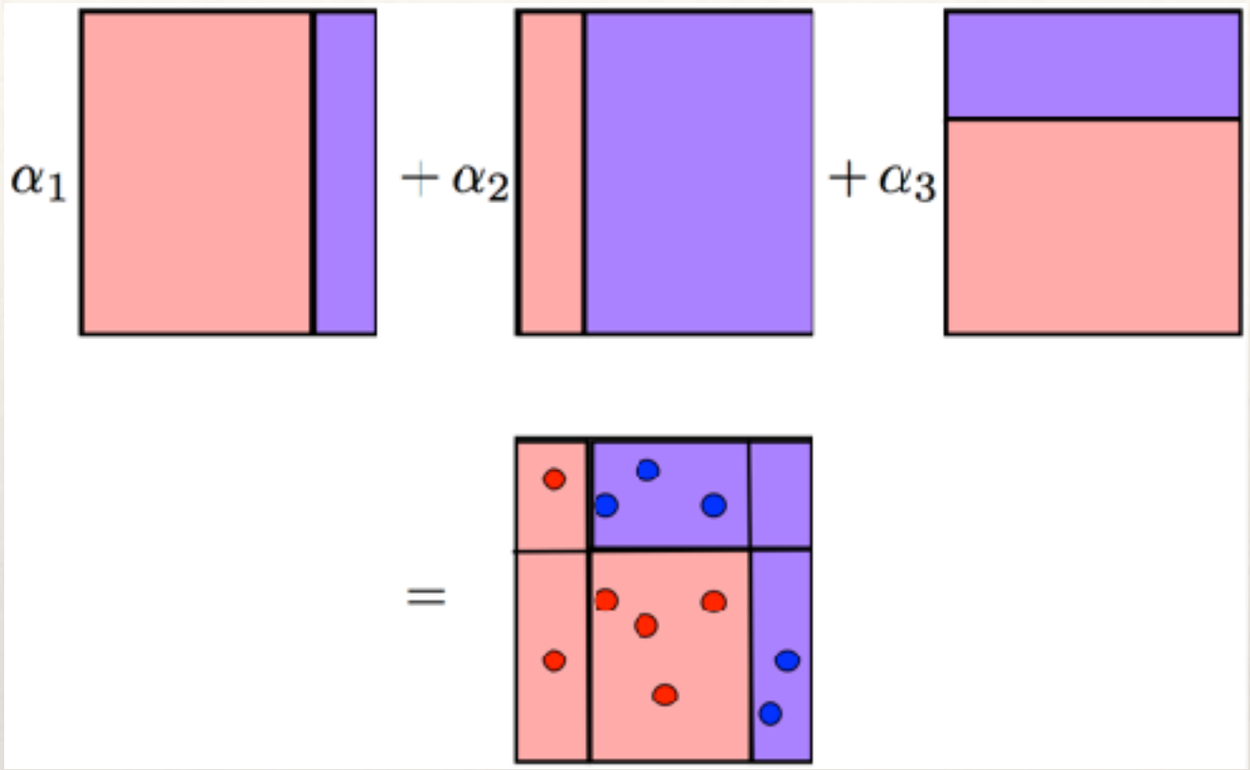
Занятие 7

- Обучение без учителя: PCA, кластеризация, поиск аномалий
- Практика на кластеризацию данных с Samsung Galaxy S3
- ДЗ № 7. 2 часть проекта



Занятие 8

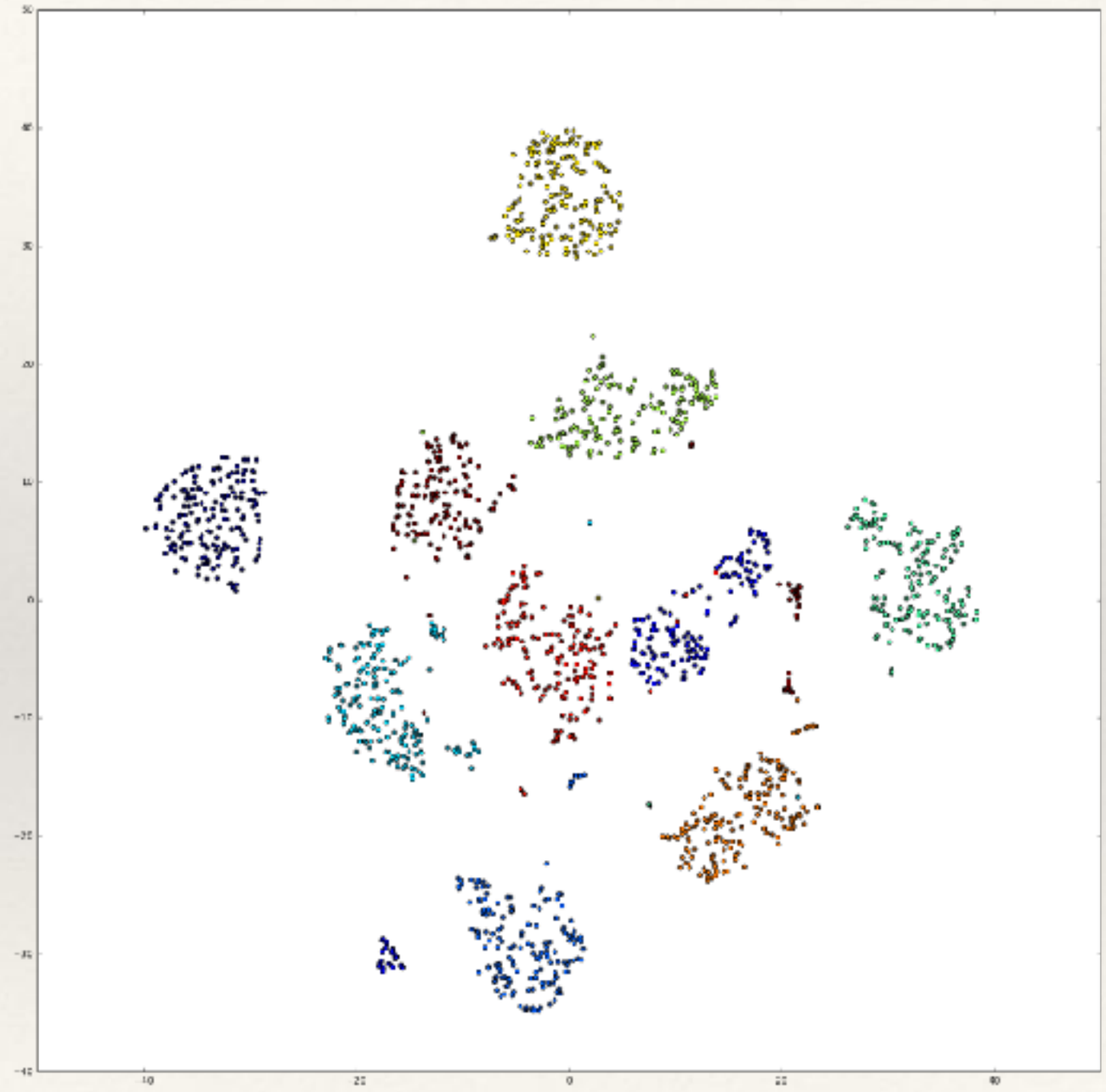
- Бустинг, градиентный бустинг, Xgboost
- Практика.
  Случайный лес и бустинг в задаче кредитного скоринга
- ДЗ № 8. Градиентный бустинг и переобучение



dmlc XGBoost

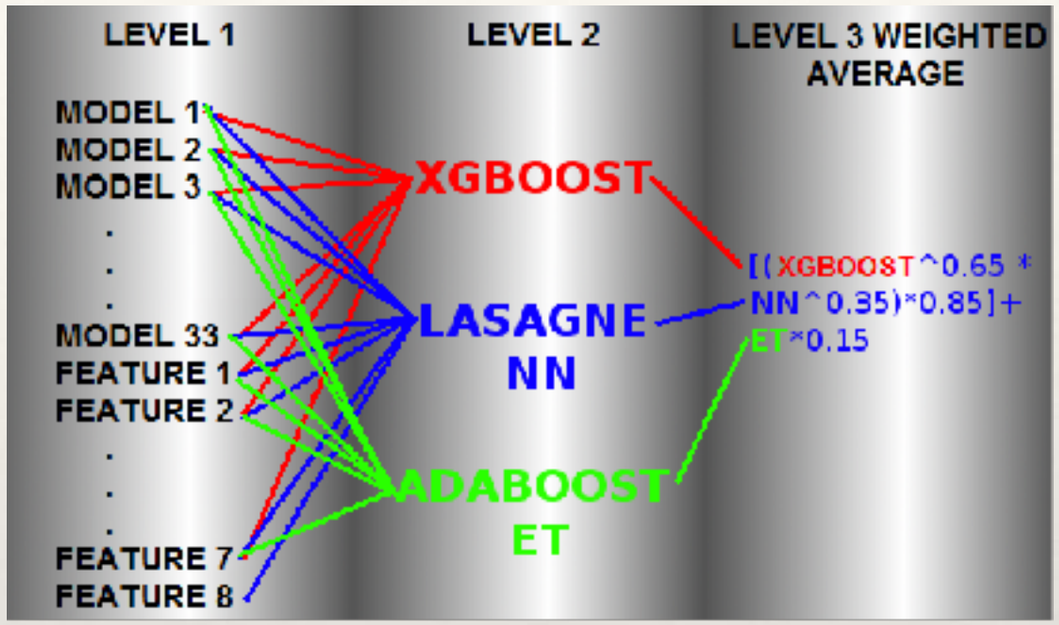
Занятие 9

- Vowpal Wabbit и основы анализа текстов, t-SNE
- Практика на классификацию текстов по темам
- ДЗ № 9. Vowpal Wabbit в одном из соревнований Kaggle



Занятие 10

- Стекинг и блендинг моделей классификации и регрессии
- Практика на блендинг случайного леса и Xgboost
- ДЗ № 10. Смешивание моделей в одном из соревнований Kaggle



Индивидуальный проект

- В течение всего курса
- *Лучше свои данные*
- Четкий план
- Пример работы над проектом в формате мастер-класса
- Подробный отзыв по проекту
- Презентации в конце