Домашнее задание по Spark

Любой из двух вариантов задания на выбор. Можно использовать любой язык программирования и любой способ сборки, а так же jupyter.

Вариант 1.

Продолжить код из занятия и реализовать кросс-валидацию, подбор параметров модели, выбор модели. Улучшить точность предсказания по сравнению с точностью, полученной на занятии (~0.2). Для повышения точности можно проделать работу с признаками.

Присылать получившийся код с комментариями на каждом блоке кода, что он делает (для более быстрой проверки задания). Также необходимо прислать значение полученной точности и пример предсказания для 100 случайных элементов на заранее отложенной выборке.

Если по существующему коду возникают вопросы, необходимо разобраться и написать комментарии к блокам кода.

Существующий код может быть разбит на части в соответствии с логикой выполнения - например, разделить на отдельные задачи преобразование признаков и подбор параметров модели. Можно организовать директории на hdfs так, чтобы удобно было проводить эксперименты - отдельная директория с датасетами, отдельная директория с предсказаниями, и пр.

Вариант 2.

С помощью Spark и SparkML реализовать модель на любых своих данных. В проекте должны присутствовать отложенная выборка, кросс-валидация, подбор параметров модели. Модель может быть любой - предсказание характеристики, кластеризация, тематическое моделирование и др.