www.bigdataschool.ru

ШКОЛА БОЛЬШИХ ДАННЫХ

Программа учебного курса

DPREP: Подготовка данных для Data Mining

Длительность: 32 ак. часа

Цель: формирование базовых компетенций слушателей при подготовке исходных данных в области больших объемов данных**.**

О курсе: Процесс сбора и подготовки исходных данных, является одним из самых трудоемких и сложных этапов в анализе больших объемов данных, который порой занимает до 80% всего времени. Использование статистических методик и современного программного обеспечения позволяет значительно сократить временные и финансовые затраты на данном этапе и повысить эффективность и качество конечных результатов.

Аудитория: Архитекторы Data lake, Аналитики данных, дата инженеры отвечающие за процессы сбора, подготовки и очистки данных.

Предварительные требования: базовые знания в области программирования Python, высшей математики, статистики.

Программа курса

1. Введение в Data Mining

- ✓Процесс Data Mining и его стандартизация (на примере CRISP-DM)
- ✓Участники процесса (Data Scientist и Data Engineer) и их роли
- ✓Этапы процесса подготовки данных
- ✓ Подготовка данных и Data Lake

2. Инструментарий подготовки данных

- ✓Проблематика больших данных (Bid Data)
- ✓Подготовка данных с помощью pandas
- ✓Промышленная подготовка данных с помощью Apache Spark

sales@bigdataschool.ru
+7 (495) 41-41-121

3. Идеальный dataset.

- ✓ Требования к данным в Machine Learning
- ✓ Типичные проблемы (отсутствующие значения, дубликаты и выбросы, нормализация, категориальные значения и т.п.)
- ✓ Выборки (обучающая, тестовая, валидационная)

4. Отсутствующие значения.

- ✓ Понятие отсутствующего значения (missing value)
- ✓ Способы борьбы (генерация или удаление).

5. Дубликаты и выбросы.

- ✓Анализ выбросов (outliers)
- ✓Борьба с дубликатами

6. Нормализация данных.

- ✓ Нормализация и нормировка – что есть что
- ✓ Технические аспекты.

7. Категориальные значения

- ✓Строки, даты и другие источники категориальных значений
- ✓Способы представления и техника генерации

8. Отсутствующие значения в исходных данных

- ✓Отсутствующих данные.
- ✓Правила замены(генерации) отсутствующих данных или опущения(omit) .

9. Заключительный проект

- ✓ Выполнение полного цикла очистки и подготовки данных на примере выбранного dataset.
- ✓Знакомство с данными и предметной областью
- ✓Подготовка данных от начала до конца
- ✓Формирование выборок.

sales@bigdataschool.ru
+7 (495) 41-41-121