Программа курса: «Аналитика больших данных для бизнес-задач»

В курсе будут освещены следующие разделы[1]:

1. Машинное обучение.
2. Технологии больших данных.
3. Бизнес-анализ в контексте проектов больших данных.
4. Программирование на языке Python[2].

Целевая аудитория:

1. Управленцы, принимающие решения на основе большого числа данных.
2. Аналитики, работающие с массивами данных.
3. Программисты, работающие с аналитиками данных.

Время: 2 месяца по 3 занятия в неделю по 4 академических часа вечером в будни и с утра в субботу.

ПН	ВТ	СР	ЧТ	ПТ	СБ	ВС
					10:00	
	18:00		18:00		11:30	
	19:30		19:30			

В ходе курса слушатели выполняют проект на базе самостоятельно выбранного кейса (в своей компании или из предложенных преподавателями).

ЗАНЯТИЕ 1

Бизнес-анализ

1. Введение в большие данные. Общее понятие о больших данных. Основные вызовы больших данных. Отличия BI от Data Science.
2. Примеры реальных кейсов.
3. Цикл аналитики данных. Роль ученого по данным, другие роли в типичных проектах. Управление проектом аналитики данных.

ЗАНЯТИЕ 2

Машинное обучение

1. Обучение с учителем. Ставятся задачи классификации и регрессии. Показывается общая часть и различия в задачах. Модель, алгоритм и процесс обучения. Проблема переобучения и регуляризация. Критерии качества.
2. Линейная регрессия, как пример решения задачи регрессии. Метод минимизации эмпирического риска.
3. Разбор алгоритма kNN, как пример решения задачи классификации.

Технологии

1. Готовые решения анализа данных. Orange. Преимущества и недостатки.

[1] Программа ещё находится на стадии обсуждения и некоторые темы могут быть изменены, переставлены, добавлены.

[2] Python стремительно набирает популярность и обгоняет R, но если все заказчики выберут R, то можем преподать R.

Python

1. Введение в Python. Экскурс в историю. Особенности языка. Приоритетные задачи. Дзен Python.

ЗАНЯТИЕ 3

Машинное обучение

1. Вероятностная постановка задачи классификации.
2. Принцип наибольшего правдоподобия.
3. Логистическая регрессия.
4. Разбор алгоритма Наивный Байесовский Классификатор.

Python

1. Операторы и выражения.
2. Файлы.
3. Списки.
4. Словари.
5. Строки.

ЗАНЯТИЕ 4

Машинное обучение

1. Обучение без учителя. Ставится задача кластеризации. Отличия задачи кластеризации от задачи классификации. Критерии качества.
2. Метрики.
3. Разбор алгоритма k-means.

Python

1. Объектно-ориентированное программирование.

ЗАНЯТИЕ 5

Технологии

1. Технологии хранения больших данных. Сравнение RDBMS, NoSQL. CAP-теорема. Выбор СУБД под проект больших данных. Аналитические СУБД. Массово-параллельные СУБД. Hadoop, SAP HANA, EMC GreenPlum.

Python

1. Обработка ошибок.
2. Кодировки.
3. Работа с CSS, JSON.

ЗАНЯТИЕ 6

Технологии

1. Hadoop. HDFS. Map Reduce. Отличия от SQL-запросов.
2. Mahout. Spark.

Python

1. Итераторы.
2. Генераторы.
3. Продвинутая работа со списками и словарями.

ЗАНЯТИЕ 7

Машинное обучение

1. Разбор алгоритма подсчета кол-ва слов в документе в парадигме Map Reduce.
2. Разбор алгоритма подсчета TF-IDF в парадигме Map Reduce.
3. Разбор алгоритма k-means в парадигме Map Reduce.

Python

1. Python для анализа данных. IPython Notebook. NumPy. Pandas.

ЗАНЯТИЕ 8

Машинное обучение

1. Разбор принципов работы нейронных сетей. Метод обратного распространения ошибки.
2. Deep Learning.

Python

1. Python для анализа данных. Scikit-learn.

ЗАНЯТИЕ 9

Машинное обучение

1. Разбор алгоритма SVM.
2. Разбор алгоритма Decision Tree.

Технологии

1. Визуализация. Tableau. Mathplotlib.
2. Обзор фреймворков для визуализации на JS.

ЗАНЯТИЕ 10

Машинное обучение

1. Boosting.
2. Bagging.
3. Разбор алгоритма Random Forest.

Python

1. Функциональное программирование.
2. Еще раз про итераторы и генераторы.

ЗАНЯТИЕ 11

Машинное обучение

1. Разбор алгоритма EM.
2. Разбор алгоритма иерархической кластеризации.

Бизнес-анализ

1. Кейс из области финансовых моделей. Кредитный скоринг.

ЗАНЯТИЕ 12

Машинное обучение

1. Обучение с подкреплением. Постановка задачи. Методы решения. Критерии качества.

Бизнес-анализ

1. Кейс из области маркетинга. Движки рекомендаций, модели рекомендательных сервисов.

ЗАНЯТИЕ 13

Машинное обучение

1. Ассоциативные правила. Постановка задачи. Поддержка и достоверность правил.
2. Методы поиска ассоциативных правил.

Бизнес-анализ

1. Кейс из области прогнозирования. Прогнозирование нагрузки.

ЗАНЯТИЕ 14

Машинное обучение

1. Ранжирование и Learning to Rank.

Бизнес-анализ

1. Собственный кейс. Возможности больших данных в вашей компании.

ЗАНЯТИЕ 15

Машинное обучение

1. Прогнозирование временных рядов.
2. Offtop(15 мин). Куда двигаться и что учить, чтобы зарабатывать 300к $ в месяц.

Бизнес-анализ

1. Собственный кейс. Возможности больших данных в вашей компании.

ЗАНЯТИЕ 16

Бизнес-анализ

1. Юридические аспекты доступа к данным.
2. Безопасность данных. Соглашения о неразглашении.
3. Технические средства обеспечения безопасности данных.

ЗАНЯТИЕ 17

Бизнес-анализ

1. Собственный кейс. Возможности больших данных в вашей компании.

2. Представление результатов анализа данных. Перевод с языка аналитики на язык управленческих решений, просчитывание последствий непринятия решений.

ЗАНЯТИЕ 18

Технологии

1. Особенности анализа неструктурированных данных
2. Анализ текстов, анализ тональности
3. Анализ мнений.
4. Особенности обработки русского языка

ЗАНЯТИЕ 19

Технологии

1. Онтологическое моделирование. OWL.
2. Работа с онтологиями в Protégé.
3. Обработка текстов с использованием онтологий.

P.S.

К окончанию курса слушатели будут:

1. Понимать роль аналитика больших данных.
2. Понимать возможности больших данных для собственной компании.
3. Знать синтаксис языка Python и уметь писать программы на нем.
4. Знать все основные и важные области машинного обучения.
5. Знать весь джентельменский набор алгоритмов анализа данных.
6. Уметь работать в студии анализа данных Orange.
7. Уметь работать с Tableau (визуализация, представление данных).
8. Уметь ставить задачи в области больших данных.
9. Уметь представлять результаты анализа данных так, чтобы они были понятны начальству.