BaseGroup Labs
ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ

# Data Mining [K.04]

Сертификационная программа (АРХИВ)

# Аннотация программы

Сертификационная программа ***Data Mining*** действовала с июня 2008 по июнь 2013 года и представляла собой электронный курс и контрольные испытания по моделям Data Mining на базе аналитической платформы Deductor 5.1 и 5.2 объемом 140 ак. часов. Теоретические сведения, изучаемые в курсе, практически полностью соответствуют соответствующим главам книги *Бизнес-аналитика: от данных к знаниям* (издательство «Питер», 2009, 2010, 2012 годы).

Сертификация по данной программе проводилась только в процессе электронного обучения, в течение которого слушатель сдавал 10 аттестационных тестов и выполнял 8 индивидуальных задач (режим ручной проверки с тьютором).

# Состав сертификационной программы

| N | Изучаемые разделы тренинга | Всего часов (академических) | Форма контроля |
|---|---|---|---|
| 1 | Технологии анализа данных | 21 | Аттестационный тест (2 теста) |
| 2 | Трансформация данных | 14 | Аттестационный тест, индивидуальная бизнес-задача |
| 3 | Data Mining – задача ассоциации | 14 | Аттестационный тест, индивидуальная бизнес-задача |
| 4 | Data Mining – задача кластеризации | 17 | Аттестационный тест, индивидуальная бизнес-задача |
| 5 | Data Mining – классификация и регрессия | 56 | Аттестационный тест (2 теста), индивидуальная бизнес-задача (2 задачи) |
| 6 | Сравнение моделей | 14 | Аттестационный тест, индивидуальная бизнес-задача |
| 7 | Ансамбли моделей | 8 | Аттестационный тест |
| **ИТОГО** | | 140 | |

# Программа сертификации

## Раздел «Технологии анализа данных»

### Теория

Современные подходы к анализу данных. Базовая терминология анализа данных, понятие модели и моделирования. Эксперты, аналитики и конечные пользователи. Виды и способы моделирования, роль экспертов в построении моделей. Последовательность шагов по анализу данных. Структурированные данные и понятия, связанные с ними: типы и виды данных, упорядоченные и неупорядоченные данные, транзакционные данные. Этапы процесса KDD для извлечения знаний из массивов данных. Машинное обучение и классы задач Data Mining. Причины популярности KDD и Data Mining и история развития технологий. Классификация программных продуктов для создания аналитических решений. Корпоративные аналитические платформы. Характеристики аналитических платформ. Языки графического моделирования в аналитических платформах.

### Практика

Практикум «Базовые навыки работы в Deductor Studio 5.x». Практикум «Возможности Deductor Studio Professional 5.x».

## Раздел «Трансформация данных»

### Теория

Понятие трансформации данных. Цели трансформации и ее роль в процессе ETL. Основные методы трансформации. Трансформация временных рядов: скользящее окно, интервал и горизонт прогноза, глубина погружения. Преобразование даты и времени. Группировка и разгруппировка данных. Объединение данных. Внутреннее и внешнее соединение. Цели квантования. Выбор числа интервалов квантования. Методы квантования. Основные методы нормализации. Нормализация с помощью поэлементных преобразований. Кодирование категориальных данных.

### Практика

Практикум «Трансформация данных в Deductor Studio 5.x».

## Раздел «Data Mining – задача ассоциации»

### Теория

Введение в аналитические модели. Обучающая выборка. Обучение «с учителем» и «без учителя». Обучающее и тестовое множества. Эффект переобучения. Сложность аналитических алгоритмов как критерий их сравнения. Понятие масштабируемых алгоритмов. Введение в ассоциативные правила. Значимость ассоциативных правил. Поиск ассоциативных правил. Алгоритм a priori. Генерация ассоциативных правил. Иерархические ассоциативные правила и методы их поиска.

### Практика

Практикум «Ассоциативные правила в Deductor Studio 5.x»

# Раздел «Data Mining – задача кластеризации»

### Теория

Введение в кластеризацию. Обзор методов кластеризации. Алгоритм k-means. Меры расстояний. Сети Кохонена. Обучение сети Кохонена. Карты Кохонена. Методика построения карты. Выбор числа нейронов карты. Недостатки и ограничения карт Кохонена.

### Практика

Практикум «Карты Кохонена в Deductor Studio 5.x».

# Раздел «Data Mining – классификация и регрессия»

Введение в классификацию и регрессию. Применение классификации и регрессии. Обзор методов классификации и регрессии. Простая линейная регрессия. Оценка соответствия регрессии реальным данным. Простая регрессионная модель. Гипотезы в регрессии. Критерии оценки значимости регрессионной модели: t-критерий и H-критерий. Множественная линейная регрессия. Модель множественной линейной регрессии. Оценка значимости множественной регрессионной модели. Регрессия с категориальными входными переменными. Проблема мультиколлинеарности. Отбор переменных в регрессионные модели. Ограничения применимости регрессионных моделей. Основы логистической регрессии. Интерпретация модели логистической регрессии. Множественная логистическая регрессия. Практикум «Логистическая регрессия в Deductor Studio»

Введения в деревья решений. Основные алгоритмы и методы построения деревьев решений. Алгоритмы ID3 и C4.5. Алгоритм CART. Упрощение деревьев решений. Нейронные сети. Принципы построения нейронных сетей. Алгоритмы обучения нейронных сетей. Многослойный персептрон. Алгоритм BackProp. Выбор параметров и архитектуры нейронных сетей. Обучение в условиях несбалансированных классов.

### Практика

Практикум «Деревья решений в Deductor Studio 5.x». Практикум «Многослойный персептрон в Deductor Studio 5.x».

# Раздел «Сравнение моделей»

Сравнение моделей с бинарной целевой переменной. ROC-анализ, чувствительность и специфичность, площадь под ROC-кривой. Правило Байеса вычисления оптимального порога. Задача оптимизации почтовой рассылки. Lift-кривая, коэффициент лифта, Lift-диаграмма и «бесполезная» модель. Площадь под кривой и индекс AUC. Gain-диаграмма. Profit-кривая. ROC-анализ, ошибки I и II

рода. Диаграммы 'Точность-полнота' (PR-Curve). Показатели для оценки качества работы регресионных моделей.

# Раздел «Ансамбли моделей»

## Теория

Понятие ансамбля моделей и идея комбинирования решений. Виды ансамблей. Методы и алгоритмы формирования ансамблей. Бэггинг: основные идеи. Технология "Возмущение и комбинирование". Почему бэггинг работает? Бустинг. Алгоритм AdaBoost.M1. Генерация классификаторов. Недостатки бустинга. Альтернативные методы построения ансамблей: аддитивная регрессия, аддитивная логистическая регрессия. Деревья выбора, стэкинг и метаобучение.