Data Understanding and Preparation for Data Science

by Deanne Larson

Module 0. About the Course (2 min)

Module 1. Introduction to Crisp DM (21 min)

- *The Data Science Process*
    - CRISP DM–Business Understanding
    - CRISP DM–Data Understanding
    - CRISP DM–Data Preparation
    - CRISP DM–Modeling
    - CRISP DM–Evaluation
    - CRISP DM–Deployment
- *Supervised Learning*
- *Unsupervised Learning*
- *Define the Problem or Opportunity*
- *Define the Data Sources*

Module 2. Data Sources Identification (20 min)

- *Data Understanding*
- *Data Sources and the Problem Statement*
- *Data Source Inventory*
    - Data Source Types
    - Different Data Structures
    - Big Data Considerations
    - Technology Platforms
- *Preparing for Exploratory Data Analysis*
- *Data Modeling and Data Science*
- *Data Pipelines and Data Stores*
- *Work with the End in Mind*

Module 3. Exploratory Data Analysis (37 min)

- *Data Understanding*
- *Exploratory Data Analysis*
- *Data Understanding – Data Profiling*
- *Sampling Size*
- *Data Profiling – EDA Methods*
- *Sample Quality*
- *Statistics Basics: Attributes*
- *Summary Statistics*
- *Distribution*
- *Data Relationships*
- *Data Relationships: Correlation Matrix*
- *Data Relationships: Outliers and Anomolies*
- *Results of Data Profiling*
- *Findings – Important Variables*
- *Outcomes and Interpretations*
- *EDA Checklist*

Module 4. Data Preparation for Modeling (30 min)

- *Data Preparation*

- *Feature Selection*
- *Data Quality Report*
- *Feature Scaling and Standardization*
- *Subset Selection*
- *Feature Selection Wrappers*
- *Feature Selection Filters*
- *Feature Selection Embedded Methods*
- *Transform for Data Modeling*
    - Data Integration
    - Data Cleansing
    - Data Formatting
- *Data Ready State*
- *Modeling – Create and Train a Model*
- *Cross Validation*

Module 5. Data Pipelines (10 min)

- *What are Data Pipelines?*
- *Why are Data Pipelines Important?*
- *Data Pipelines and Data Science*
- *Data Pipelines – Repeatable Assets*
- *Data Pipelines – Existing*

Module 6. Visualization Techniques (33 min)

- *Data Description*
- *Data Description Report*
- *Data Evaluation*
    - Data Evaluation Scope
    - Data Evaluation – Distribution: Part 1 & 2
    - Data Evaluation – Outliers
    - Data Evaluation – Correlation
    - Data Evaluation – Trends and Patterns
    - Data Evaluation – Categories
    - Data Evaluation – Geospatial
    - Data Evaluation – Relationships
    - Data Evaluation – Clusters
- *Data Quality Report*
- *Data Cleansing*
- *Data Quality Scorecards and Dashboards*
- *Feature Ranking*

Module 7. Data Quality and Integrity (36 min)

- *Data Quality and Machine Learning*
- *Accurate Data*
- *Consistent and Complete Data*
- *Algorithms and Data Quality*
- *Algorithm Requirements*
- *Categorical Data and Algorithms*
- *Bins and Ranges*
- *High Cardinality*

- *Reduce Cardinality*
- *Dealing with Outliers*
- *Missing Data*
- *Time of Event*