TDWI Data Literacy Bootcamp: A Hands-on Workshop for Business

Course Outline

Day 1

Hands-on: Business Analysis for Everyone with Excel—Becoming a Data Detective Made Easy

Module 1 – Process-Centric Thinking

- Continual Improvement
  - Beyond Enterprise Data
  - Beyond Transactions
  - Understanding Cause and Effect
  - Business Impact
- Business Processes
  - Every Business Has Processes
  - The Process of Website Visits
- Process Variation
  - All Processes Have Variation
  - Two Types of Variation
  - Causes and Cures of Variation
  - Multiple Uses for Data
  - Traditional Focus on Transactions
  - Relational Perspective

Module 2 – Numeric Literacy and Data Visualization

- Numeric Summaries
  - Summaries Communicate Efficiently
  - Data Professionals Are Translators
- Measures of Location
  - Average (Arithmetic Mean)
  - Median
- Measures of Dispersion
  - Range
  - Standard Deviation
- Numerical Summaries & Data Types
  - Nominal vs Ordinal vs Numeric Data
  - Numeric Summaries Based on Data Type
- Ratios

  - Ratios
  - Percentages
  - Proportions
- Data Visualization
  - Running Records
  - Bar Charts
  - Histograms
- Hands-on Lab #1

Module 3 – Analyzing Business Data

- Analysis & Data Fundamentals
  - 4 Principles of Business Data Analysis
  - 4 Data Principles
- Business Data is Different
  - Aggregation
  - Data Sparsity
  - Autocorrelation

Module 4 – The Process Behavior Chart

- Plotting Your Process
  - Process Measures Over Time
  - Process Location
  - Process Variation
- Characterizing Process Variation
  - Exceptional Process Measures
  - Process Limits
  - The “Happy Path” Corridor
  - The Complete Chart
- Hands-on Lab #2

Module 5 – Using the Process Behavior Chart

- What Can You Analyze?
  - Types of Process Measures
  - Logically Comparable Data
- Predictable Processes
  - Predictable Processes Defined

    - Recognizing a Predictable Process
- Detecting Process Changes
    - Detection Rule #1
    - Detection Rule #2
    - Detection Rule #3
    - Detection Rule Prioritization
- Hands-on Lab #3

Module 6 – Becoming a Data Detective

- Buried Signals
    - How Process Behavior Charts Work
    - Aggregated Data
    - When Noise Might Not Be Noise
- Uncovering Insight
    - Effect of Process Hierarchies
    - Process Disaggregation
- Hands-on Lab #4

Module 7 – Data Detective Mastery

- Data Sleuthing
    - Case Study – Has e-commerce Improved?
    - Examining the Data
    - The Null Hypothesis
- Group Differences
    - Comparing the Data
    - Are Differences Real?
- Hands-on Lab #5

Module 8 – What's Next?

- Additional Resources

Day 2

Hands-on: Predictive Analytics with Excel—Linear Regression Made Easy

Module 1 – Linear Regression Introduction

- Families of Predictive Analytics

    - Classification
    - Regression
- An Example Model
    - Linear Regression Models as Equations
    - A House Price Linear Regression Model

Module 2 – Numeric Literacy

- Types of Data
    - Categorical vs Numeric
    - Interval vs Ratio
- Measures of Location
    - Average (Arithmetic Mean)
    - Median
- The Normal Distribution
    - Distribution Parameters
    - Standard Deviation
    - Probability of Values

Module 3 – Predictions Using the Mean

- The Mean is a Predictive Model
    - When the Mean is a Reasonable Predictive Model
    - Why the Mean is a Predictive Model
    - Issues with Prediction Using the Mean
- Our First Model
    - Predicting the Iris Dataset with the Mean
    - Is the Model Any Good?
    - Measuring Model Goodness – Total Sums of Squares
- Hands-on Lab #1

Module 4 – Correlation

- What is Correlation?
    - Data Synchronicity
    - Types of Correlation
    - Correlation Strength
    - Correlation in Excel
- Lies and Statistics

  - Correlation is Useful, but Beware
  - Anscombe's Quartet
  - The Importance of Data Visualization
- Visualizing the Iris Dataset
- Hands-on Lab #2

Module 5 – Simple Linear Regression

- Beating the Mean
  - The Mean as a Baseline Predictive Model
  - Linear Regression vs the Mean for Better Predictions
- Simple Linear Regression in Excel
  - The Excel Analysis ToolPak Output
  - Interpreting ToolPak Output as an Equation
- Understanding the Model
  - Uncertainty in Linear Regression
  - Confidence Intervals
  - Explanatory Power
  - Correlation and Simple Linear Regression
- Hands-on Lab #3

Module 6 – Multiple Linear Regression

- When Simple Won't Do
- Categorical Data
  - Dummy Encoding
  - Interpreting Categorical Features
- The Power of Multiple Features
  - Multiple Features and Predictions
  - Better Predictions?
- The Rewards of Complexity
  - You Need More Complexity
  - Beware Complexity
- Interaction Effects
  - Multiple Features – Chocolate and Peanut Butter
  - Interaction Modeling
  - Our "Final Model"
- Hands-on Lab #4

Module 7 – Is Your Model Any Good?

- Linear Regression Data Assumptions
    - Ordinary Least Squares (OLS) Linear Regression
    - OLS Makes Data Assumptions
    - The Six Data Assumptions
- Assumption 1 – Models Are Fully Specified
    - Intuition
    - Data Example
    - How to Validate Assumption
- Assumption 2 – Features Uncorrelated with Errors
    - Intuition
    - Data Example
    - How to Validate Assumption
- Assumption 3 – Errors Uncorrelated
    - Intuition
    - Data Example
    - How to Validate Assumption
- Assumption 4 – Errors Have Constant Variance
    - Intuition
    - Data Example
    - How to Validate Assumption
- Assumption 5 – No Multicollinearity
    - Intuition
    - Data Example
    - How to Validate Assumption
- Assumption 6 – Errors Normally Distributed
    - Intuition
    - Data Example
    - How to Validate Assumption
- Hands-on Lab #5

Module 8 – What We Did Not Cover

- Modeling Curvature
- Transforming Data
- Extrapolation

- Standardizing Data
- Models Without Intercepts

Module 9 – Additional Resources

Day 3

Hands-on Machine Learning Made Easy - No, Really!

Module 1 – Machine Learning Fundamentals

- What is Machine Learning?
    - Working Definition
    - Relationship to Predictive Analytics
    - What is an Algorithm?
    - Types of Data
- Types of Machine Learning
    - Supervised Learning
    - Unsupervised Learning
    - Semi-supervised Learning
    - Reinforcement Learning

Module 2 – Decision Trees

- Introduction
    - What is a Decision Tree Model?
    - How Decision Trees Learn
- The Decision Tree Algorithm
    - Determining the Best Split
    - Binning Numeric Data
    - Decision Tree Case Study
- Pros of Decision Trees
    - Interpretability
    - Feature Selection
    - Prediction & Data Flexibility
    - Speed of Predictions
- Hands-on Lab #1

Module 3 – The Mighty Random Forest

- The Con of a Single Decision Tree
- Random Forest "Fixes" a Single Decision Tree
    - Ensembling
    - Bagging
    - Feature Randomization
- Random Forest 201
    - Bagging Revisited
    - Feature Randomization Revisited
- The Random Forest is Mighty
    - Data Flexibility
    - Easy of Training
    - Broad Applicability
    - Error Estimates
- Preview of Coming Attractions
    - Understanding Random Forest Output
    - Understanding Confusion Matrices
- Hands-on Lab #2

Module 4 – Data Analysis & Feature Engineering

- Data Analysis
    - Feature Importance
    - Feature Engineering
- Preview of Coming Attractions
    - Logic for Engineering a New Feature
    - Implementing the New Feature in R Code
- Hands-on Lab #3

Module 5 – Intro to Model Testing

- The Bane of Machine Learning
    - Beware Overfitting
    - Feature Engineering Leads to Overfitting
    - You Must Engineer Features
- Date for Model Training
    - The Train/Test Split
    - The Random Forest Improves the Train/Test Split

- The Model Training Process
    - Split Data
    - Confidence Intervals
    - Explanatory Power
    - Correlation and Simple Linear Regression
- Model Complexity
    - Model Complexity in Decision Trees
    - Model Complexity-Overfitting Tradeoff
- Hands-on Lab #4

Module 6 – Improving Our Model

- Overfitting 201
    - Model Complexity-Overfitting Tradeoff Revisited
    - Finding the Complexity "Sweet Spot"
- Feature Randomization 301
    - Identifying Low-value Features
    - When/How to Remove Features
- Hands-on Lab #5

Module 7 – Additional Resources

