Essential Data Science for Business: Top 10 Analytics Topics

What are the key topics that are used in the business?

NISS Webinar

Victor S.Y. Lo

July 29, 2020
Content

► Three Major Types of Analytics

► What Skills Data Scientists Should Have

► Top 10 Analytics Topics – Important and Practical

Disclaimer: The views expressed here are solely those of the speaker and do not in any way represent the views of Fidelity Investments
“The best thing about being a statistician is that you get to play in everyone's backyard.”

- John Tukey, decades ago

“We no longer simply enjoy the privilege of playing in or cleaning up everyone's backyard. We are now being invited into everyone's study or living room, and trusted with the task of being their offspring's first quantitative nanny.”

- Xiao-li Meng (2009), Harvard University
Three Types of Analytics

- Prescriptive
  - What should we do? What is the Best Decision?
    - Support decision making and proactive actions

- Predictive
  - What will happen?
    - Predict future forward-looking behavior, events, probabilities, or trends

- Descriptive
  - What happened?
    - Reports and profiling
    - Data visualization

Three Types of Analytics

- **Descriptive**
- **Predictive**
- **Prescriptive**

- **Causality**
Data Science is a Diversified field with professionals from a variety of disciplines, see Lo (2019)
Data Science Project Process

**Step 1: Defining Project Scope & Analytic Consulting**
- **Understand** problems;
- Form working team and define scope

**Step 2: Data Gathering**
- Complex process assisted by engineers

**Step 3: Data Processing**
- Data merging (from different files or systems), data cleaning

**Step 4: Data Mining & Algorithm Development**
- Hypothesis development, exploratory data analysis, **model development**, validation; report interim results and receive feedback

**Step 5: Insight Extraction & Final Report**
- Extract **insights**, develop **usage plan**, report findings, gather feedback & refine

**Step 6: Model Implementation & Business Usage**
- Work with IT on model implementation; work with business users on business usage/test & learn
Top 10 Important and Practical Topics that **May NOT Be Covered in Your Education Program**…
1. Analytical Consulting, Communication and Soft Skills
1. Analytical Consulting, Communication and Soft Skills

Analytical Consulting

1. Business Communication

2. Verbal & Written Presentation

3. Visualization & Data Storytelling

4. Communication with Data & Tech Professionals
Analytic Consulting Process

1) Initiate Project
   - Project opportunity identified by business owner or data scientist
   - Ask right questions and understand business problem
   - Agree on objective & scope

2) Define Scope
   - Recommend initial approach and draft proposal

3) Propose Approach
   - Establish a feasibility study, if appropriate
   - Identify other dependencies, e.g., business & data expertise
   - AGILE
   - Determine data scientist, data engineer, and other resources required

4) Identify Resources and Dependencies

5) Leadership during Project Development
   - Leadership discussion with business and technology stakeholders
   - Determine engagement process
   - Report Interim analyses to seek feedback and buy-ins
   - Performance tracking and refinement

6) Leadership during Deployment
   - Present to multiple levels of the business
2. Computer Science, Programming, and Tools
2. Computer Science, Programming, and Tools

Demand for computational power has dramatically increased and will need to expand much further:

- Growth in Structured and Unstructured Data
- Internet of Things (IoT)
- Practical Success of Deep Learning

IDC predicted that the global data size would increase by ~3X from 2019 to 2025 (175 zettabytes), see Reinsel et al (2020)
2. Computer Science, Programming, and Tools

1. AI/ML and Statistical Programming
   Python, R, SAS, GPU Programming

2. AI/ML Tools
   PyTorch, Keras, Tensorflow, MXNET, Caffe, CNTK, SageMaker

3. Data Knowledge

4. ETL (Extract, Transform, and Load) skills for Big Data

5. Model Deployment
   Kubernetes, Docker

Foundational Programming Skills and Computer Science
3. Descriptive Analytics, Exploratory Data Analysis, and Data Visualization

- Descriptive
- Predictive
- Prescriptive

- Causality
3. Descriptive Analytics, Exploratory Data Analysis (EDA), and Data Visualization

1. Data Visualization

2. Reports, Summary Statistics, Significance Testing, and Profiling

3. Feature Selection, as input to Predictive Modeling

Descriptive Analytics
3. Descriptive Analytics, Exploratory Data Analysis, and Data Visualization

Birth Rate

Death Rate

LITERACY

Graphics by Leland Wilkinson with permission
4. Predictive Analytics and Machine Learning

- Descriptive
- Predictive
- Prescriptive

Causality
4. Predictive Analytics and Machine Learning

Predictive Analytics and Machine Learning

Supervised Learning

“Traditional” Regression based Methods; Statistical Inference vs Prediction

Modern ML Methods, e.g., CART, Random Forest, Gradient Boosted Tree, SVM, Neural Network

Model Selection and Validation

Training, Validation, Test

K-fold Cross-Validation

Measurement Metrics

Sensitivity and Specificity

Recall and Precision

Accuracy, F1, AUROC, Gini

MAPE, MAE

Decile and Gains charts

Unsupervised Learning

Dimension Reduction: PCA, Factor Analysis, t-SNE

Cluster Analysis

Topic Modeling

See also a new thought-provoking paper by Efron (2020) and a classic one by Breiman (2001)
Decision Tree

How well they are doing financially – Illustrative Only

Overall

Savings

Protection

- 50% well
- 80% well
- 25% well
- 90% well
- 60% well
- 50% well
- 15% well
5. Deep Learning

A.I.

Rule-Based: Tell a Machine What to Do

Machine Learning

Let the machine *learn*
- Feed data and set a goal

DEEP LEARNING
Introduction to Neural Network

Inside a Multi-Layer Perceptron (MLP) neural network, it is a set of nonlinear functions\(^1\).

The special composite function leads to a **Universal Approximator** to ANY functions.

\[
Y = w_{2,0} + \sum w_{2,j} I_j, \text{ where } I_j = \frac{1}{1 + \exp(-Z_j)}, \quad Z_j = \sum w_{1,kj} x_k
\]

---

\(^1\) Other common activation functions include ReLu (most popular) and Tanh
Types of Deep Learning

- **Multi-layer Perceptron (MLP)**
  - Designed for **standard IID data**
  - No sharing of weights, i.e. **fully-connected**

- **Convolutional Neural Network (CNN)**
  - **Image data** or spatial-temporal data
  - Allow sharing of weights
  - Many possible architectures

- **Recurrent Neural Network (RNN)**
  - **Sequential data** (time series, word sequence)
  - Allow sharing of weights
  - Many possible architectures
  - Special forms: LSTM & GRU
Deep Learning for Medical Use Case

**Goal:** Predict Total Joint Replacement (TJR)

**Data:** De-identified claims data with detailed individual level time series of medical codes (diagnosis, procedure, etc.)

**Approach:**
Compare various deep learning architectures (CNN, RNN/GRU) with Lasso Logistic and RF

*Source: Qiu et al (2019), with permission*
6. Causal Inference and Uplift Modeling

- Descriptive
- Predictive
- Prescriptive

Causality
6. Causal Inference and Uplift Modeling

Common Causality Related Questions in Business

► **Price:** Would a price reduction generate high demand?

► **Promotion:** What are the impact of direct marketing and advertising?

► **Place:** What are the effects of store location and appearance on business outcomes?

► **Product:** Would an improvement in product feature be valuable to customers?

Similar questions can apply to other fields.
Blocking the “Back-Door” Path

Goal: Measure Effect of Sales Campaign, using Historical Sales Data

Confounders
E.g. age, income, geography

Sales reps cherry-picked by demographics

Treatment / Control Assignment
E.g. Sales Call

Causal ??

Outcome
E.g. Buy or not

Estimate **Average Treatment Effect** by breaking the Confounder-Treatment link: **Propensity Score Matching**

Personalized Medicine:
Stratify for more efficient treatment

<table>
<thead>
<tr>
<th>Clinical Benefit achieved if Receiving Placebo or no treatment</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>Wasteful [Over-Treat]</td>
<td>Beneficial [Should-Treat]</td>
</tr>
<tr>
<td>NO</td>
<td>Harmful [Do-Not-Treat]</td>
<td>Futile [Do-Not-Treat]</td>
</tr>
</tbody>
</table>

Source: Chapter 3 of Yong (2015), with permission
7. Prescriptive Analytics and Optimization

- Prescriptive
- Predictive
- Descriptive

Causality
7. Prescriptive Analytics and Optimization

**Causal Inference**

- **Price**: Would a price discount generate high demand?

- **Promotion**: What are the Impact of direct marketing and advertising?

- **Place**: What are the effects of store location and appearance on business outcomes?

- **Product**: Would an improvement in product feature be valuable to customers?

**Prescriptive Analytics/Optimization**

- **Price**: What is the optimal price?

- **Promotion**: How to optimally invest in direct marketing and advertising campaigns?

- **Place**: Where to open new stores? How should they look?

- **Product**: What are best product features?
7. Prescriptive Analytics and Optimization

- Mindset – Objective Function, Constraints
- Mathematical Programming (MP)
  - LP
  - ILP
  - MIP
  - QP
  - NP
  - DP & MDP
- Heuristics
- Multi-Objective Optimization (MOO)
- Optimization Under Uncertainty
  - Stochastic Programming
  - Robust Optimization
  - Mean-Variance Optimization, Nobel Econ 1990
- Reinforcement Learning – e.g., Alpha Go
- Stable Marriage and Kidney Exchange, Nobel Econ 2012
Application: Customer Relationship Management (CRM)

- **Treatments**
  - Channel
  - Message/Offer
  - Individuals

- **Individual Characteristics**

  - Best deal…
  - No risk free trial…
  - Lowest price ever…
  - Super benefits…
  - Unbeatable service…
  - Top performance…
  - Scientifically proven…

Lots of Possible Treatment Combinations
8. Unstructured Data Analysis
8. Unstructured Data Analysis

Natural Language Processing (NLP) / Text Analytics

- Document Processing
  - Contract, legal
  - Doctor’s notes
- Survey Verbatim
- Search Engine
- Chatbot

Image Recognition
- Radiology
- Check Scan
- Security & Biometrics
- Insurance Claims

Speech Analytics
- Call Center:
  - Sentiment Analysis
  - Topic Modeling
  - Features
8. Unstructured Data Analysis

Natural Language Processing (NLP) / Text Analytics

- Computational Linguistics

- Advanced – **word embedding**, deep learning based (esp. RNN), Attention, Transformers, ELMO, BERT, etc.

- Specific applications: search, chatbot (QA), topic modeling, sentiment analysis

Image Recognition

- Convolutional Neural Network (**CNN** or Convnet)
- Computer Vision (OCR, R-CNN)

Speech Analytics

- Language Model and Acoustic Model
- Hidden Markov Model (HMM)
- Deep Learning
9. Social Sciences and Data Science Ethics
Microeconomics: Price Determination

1. Assess Price Elasticity from data

\[ S = 2817 - 354P \]

2. Determine the Optimal Price

\[ f(P) = S(P - 2) = (2817 - 354P)(P - 2) \]
Behavioral Economics = Economics + Psychology

Prospect Theory

Value

Gains

Losses

Loss Aversion

Nudge Theory

- Opt-in vs Opt-out
- Choice architecture - # choices
- Language Framing

Can be tested and modeled statistically as input to behavioral change optimization

Daniel Kahneman, Nobel Econ 2002
See Kahneman (2011)

Richard Thaler, Nobel Econ 2017
See Thaler & Sunstein (2009)
Data Science Ethics involves Multiple Disciplines

1. Data Usage: Sampling Bias & Predictors
2. Data Privacy & Policy
3. Algorithmic: Bias Detection & Mitigation
4. Model Transparency
5. Governance

Advanced techniques include Causal Inference

10. Domain Knowledge and Application Areas
10. Domain Knowledge and Application Areas

Common Daily Usage of Data Science

1) Marketing & Sales
   • Database Marketing / CRM
   • Market Research
   • Marketing Mix
   • Marketing Strategy

2) Financial & Risk Management
   • Modern Portfolio Theory (MPT)
   • Risk Management: Market, Credit, Operational Risks
   • Actuarial Science & Insurance

3) Operations Management, Supply Chain, and Logistics
   • Call Center Analytics
   • Logistics & Transportation
   • Supply Chain Management
   • Intelligent Automation

4) Healthcare & Biomedicine
   • Health Informatics
   • Drug Discovery
   • Genomics
   • Clinical trials
   • Epidemiology
Future NISS Tutorials, see https://www.niss.org/

1) Analytical Consulting, Communication and Soft Skills

2) Computer Science, Programming, and Tools

3) Descriptive Analytics, Exploratory Data Analysis, and Data Visualization

4) Predictive Analytics and Machine Learning

5) Deep Learning

6) Causal Inference and Uplift Modeling

7) Predictive Analytics and Optimization

8) Unstructured Data Analysis

9) Social Sciences and Data Science Ethics

10) Domain Knowledge and Application Areas
### Translation Between Statistics and AI / ML: Same or Similar Terminology

<table>
<thead>
<tr>
<th>Statistics / Economics / Epidemiology / Math</th>
<th>Data Science / AI / Data Mining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical modeling</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>Dependent Variable / Response Variable</td>
<td>Target Variable / Label</td>
</tr>
<tr>
<td>Independent Variable</td>
<td>Feature¹</td>
</tr>
<tr>
<td>Parameters / coefficients</td>
<td>Weights</td>
</tr>
<tr>
<td>Intercept</td>
<td>Bias²</td>
</tr>
<tr>
<td>Estimation</td>
<td>Training</td>
</tr>
<tr>
<td>Out-of-Sample / Holdout Sample</td>
<td>Test Data</td>
</tr>
<tr>
<td>Regression / Classification</td>
<td>Supervised Learning</td>
</tr>
<tr>
<td>Cluster Analysis / PCA / Factor Analysis / SVD</td>
<td>Unsupervised Learning</td>
</tr>
<tr>
<td>Variable Selection</td>
<td>Feature Selection</td>
</tr>
<tr>
<td>Dimension Reduction</td>
<td>Feature Reduction</td>
</tr>
<tr>
<td>Data point / observation</td>
<td>Instance / Sample³ / Example</td>
</tr>
<tr>
<td>Outlier Detection</td>
<td>Anomaly Detection</td>
</tr>
<tr>
<td>Log likelihood function of a binary variable</td>
<td>Cross Entropy</td>
</tr>
<tr>
<td>Logistic function</td>
<td>Sigmoid function</td>
</tr>
<tr>
<td>Multinomial Logit</td>
<td>Softmax</td>
</tr>
<tr>
<td>Dummy Coding</td>
<td>One-hot Coding</td>
</tr>
<tr>
<td>Misclassification Table</td>
<td>Confusion Matrix</td>
</tr>
<tr>
<td>Bayesian Computation</td>
<td>Probabilistic Programming</td>
</tr>
<tr>
<td>Approximate Dynamic Programming / Markov Decision Process</td>
<td>Reinforcement Learning</td>
</tr>
<tr>
<td>Randomized Controlled Trial (RCT)</td>
<td>A/B Testing</td>
</tr>
<tr>
<td>Factorial Design</td>
<td>Multivariate Testing (MVT)</td>
</tr>
<tr>
<td>Time series data</td>
<td>Sequential data</td>
</tr>
<tr>
<td>Classification Matrix</td>
<td>Confusion Matrix</td>
</tr>
<tr>
<td>Power [P(\text{Reject H0</td>
<td>H1 is true}) \text{ or } 1-P(\text{Type II error})]</td>
</tr>
<tr>
<td>False Discovery Rate (FDR)</td>
<td>1 − Precision</td>
</tr>
<tr>
<td>Average Treatment Effect (ATE)</td>
<td>Lift (Marketing)</td>
</tr>
<tr>
<td>Heterogeneous Treatment Effect (Econ.)</td>
<td>Uplift Modeling</td>
</tr>
<tr>
<td>Or Conditional Average Treatment Effect (CATE; Econ.)</td>
<td>Uplift Modeling</td>
</tr>
<tr>
<td>Or, Effect Modification (Epidemiology)</td>
<td>Uplift Modeling</td>
</tr>
<tr>
<td>Or, Impactability Modeling (Health)</td>
<td>Uplift Modeling</td>
</tr>
<tr>
<td>Or, Subgroup Analysis (Biostat)</td>
<td>Uplift Modeling</td>
</tr>
</tbody>
</table>

¹ A feature can also be a function of original variables.

² The standard statistical definition of Bias is the discrepancy between the actual value of an unknown parameter and the expected value of its estimator. Such definition is also used in machine learning, which is totally different from the Intercept-equivalent meaning in neural networks.

³ The traditional definition of a sample refers to a subset of the population, which is a collection of observations. In some AI/ML literature, a single observation is sometimes called a sample.
References


Kahneman, Daniel (2011). Thinking, Fast and Slow, FSG.


Leamer, Edward E. Macroeconomic Patterns and Stories: A Guide for MBAs, Springer.


References (continued)

The Institute and Faculty of Actuaries (IFoA) and the Royal Statistical Society (RSS) (2019) “A Guide for Ethical Data Science.”
APPENDIX
History of Data Science

- Wu 1997, proposed:
  - Statistics → Data Science
  - Statistician → Data Scientist

- Cleveland 2001, proposed:
  - Enlarge the major areas of Statistics → Data Science

Source:
https://course.ccs.neu.edu/cs7280sp16/CS7280-Spring16_files/50YearsOfDataScience.pdf
What is the Right Way to Measure Lift?

A successful response model

A successful treatment (e.g. marketing) program

LIFT?

LIFT?

CUME Pct of Responders
Random

Pct of Treatment Responders
Pct of Treatment Group

Incidence of Treatment Responders

Treatment Response Rate
Control Response Rate

<table>
<thead>
<tr>
<th>Decile</th>
<th>Incidence of Treatment Responders</th>
<th>Pct of Treatment Group</th>
<th>CUME Pct of Responders</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>2</td>
<td>7%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>3</td>
<td>4%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>4</td>
<td>2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test 1
Test 2
Total

3.3%
2.3%
2.7%
3.0%
1.7%
2.0%
0%
0%
0%
0%
0%
0%
0%
0%
5. Deep Learning

History of A.I.: From Programming to Deep Learning

1. Birth of Programming
   - Ada Lovelace: computers can never be as intelligent as humans

2. Birth of A.I.
   - Alan Turing: a machine can possibly think for itself
   - Participants at Dartmouth workshop called it Artificial Intelligence

   - Rule-based: hard-code with human expertise
   - Work well on limited applications

4. Machine Learning (Modern A.I.)
   - Feed Big Data data to an algorithm and set a goal
   - Wide applications in medicine, marketing, finance, logistics, operations, and beyond

5. Deep Learning (Latest Machine Learning)
   - Deep Learning became widely used for image processing, natural language processing (NLP), and so on
   - Hinton, Bengio, LeCun won 2018 Turing Award

- Most A.I.’s are designed to do a single task: Narrow AI
- Can Machines Think? It depends…
Framework for Causal and Association Analysis

Causal Inference
(Lift Analysis, Average Treatment Effect)

Uplift Modeling
(Heterogeneous Treatment Effect, Conditional ATE, Effect Modification)

Reporting / Summary Statistics

Response Modeling / Propensity Modeling

Population / Sub-population

Personalized

For the do-operator above, see Pearl (2000) and Pearl and MacKenzie (2018).
Macroeconomics: Sensitivity to Macroecon Factors

Outcomes, e.g. Sales, Revenue, Cash Flow, Risk

- Interest Rate
- Financial Market
- GDP
- Seasonality
- Unemployment

Controllable Drivers (Treatments/Interventions)

See Oxelheim and Wihlborg (2008) and Leamer (2009)