Financial Data Professional Institute

FDP CANDIDATE ORIENTATION
March 16 – April 4, 2020 FDP Exam

Hossein Kazemi, Senior Advisor, CAIA Association
Keith Black, Managing Director Curriculum & Exams, CAIA Association & recent FDP Charterholders
Mirjam Dekker, Project Manager, FDP Institute

www.fdpinstitute.org
January 15, 2020
• Welcome
• Introductions – today’s presenters

Agenda

• Part I - Exam Topics
  • Curriculum
  • Examination Format
  • Topic 1 – 8 (Required Readings, Sample Keywords, Sample learning Objectives, Sample Question & Answer Key)

• Part II - Candidate Experience

→ Goal: provide you the tools & knowledge that will lead to a successful exam day
FDP Curriculum

PMR E-Book
8 required readings

$99
Exam Registrants

Topics in Financial Data Science
12 required readings

Free
Exam Registrants

3 required books

- Data Science for Business
- Big Data and Machine Learning in Quantitative Investment
- An Introduction to Statistical Learning

Exam Registrants can look for best available price.
Examination Format

- 75 multiple choice questions (70%)
- 3-4 essay questions (30%)

Session 1: 70 MC questions
100 minutes

Optional break

Session 2: 5 MC questions
3-4 Essay questions
100 minutes
FDP Curriculum

Topics, Learning Objectives & Keywords

**TOPIC 1**
Introduction to Data Science & Big Data

**READING 1**

**READING 2**

**READING 3**

**READING 4**

**KEYWORDS**
- Data mining (p. 2)
- Churn (p. 4)
- Data engineering (p. 5, 7)
- Target (p. 24)
- Unsupervised data mining (p. 24)
- Data science (p. 4)
- Data-driven decision making (p. 5)
- Data-analytic thinking (p. 12)
- Label (p. 24)
- Supervised data mining (p. 25)

**LEARNING OBJECTIVE 1**

**Learning Objectives**
Demonstrate proficiency in the following areas:

1) Data analytic thinking (Ch. 1)
   *For example:*
   - Discuss the ubiquity of data opportunities.
   - Define data science, engineering, and data-driven decision making.
   - Explain data and data science capability as a strategic asset.
   - Describe data-analytic thinking.
   - Compare data science and the work of the data scientist.

2) Business problems and data science solutions (Ch. 2)
   *For example:*
   - Describe how one transitions from business problems to data mining tasks.
   - Compare supervised methods to unsupervised methods.
   - Describe the difference between data mining and using the results of data mining.
   - Describe key aspects of the data mining process, including business understanding, data understanding, data preparation, modeling, and evaluation.
# FDP Curriculum

## 1. Introduction to Data Science & Big Data

## 2. DM & ML: Introduction

## 3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models

## 4. DM & ML: Classification & Clustering

## 5. DM & ML: Performance Evaluation, Backtesting & False Discoveries

## 6. DM & ML: Representing & Mining Text

## 7. Big Data, DM & ML: Ethical & Privacy Issues

## 8. Big Data and Machine Learning in the Financial Industry

### Reading(s):  


### Sample Keywords:

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Data mining (p. 2)</th>
<th>Data science (p. 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Churn (p. 4)</td>
<td>Data-driven decision making (p. 5)</td>
</tr>
<tr>
<td></td>
<td>Data engineering (p. 5, 7)</td>
<td>Data-analytic thinking (p. 12)</td>
</tr>
<tr>
<td></td>
<td>Target (p. 24)</td>
<td>Label (p. 24)</td>
</tr>
<tr>
<td></td>
<td>Unsupervised data mining (p. 24)</td>
<td>Supervised data mining (p. 25)</td>
</tr>
</tbody>
</table>
Sample Learning Objective:

1.2.2 The benefits and limitations of big data for investment decisions

For example:

A. Explain how big data provides a new source of fundamental insights, relative to discretionary and quantitative strategies, by using the four data quadrants of any investment framework (financial data, alternative data, market data and internal data).

B. Describe a process of harnessing data-driven insights including:
   - Sourcing
   - Backtesting
   - Tagging
   - Mapping, and
   - Visualization

Sample Question:

In the article “Big Data is a Big Deal,” there are three main types of manager orientations (legacy fundamental, legacy quantitative, and big data) which use various combinations of four types of data sources. Examples are given for each of the four categories. Which type of manager orientation is most likely to rely on implied volatility data and factor analysis data?

- Fundamental
- Quantitative*
- Big data

Answer key ➔ LO 1.2.2. p.9
FDP Curriculum

1. Introduction to Data Science & Big Data

2. DM & ML: Introduction

3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models

4. DM & ML: Classification & Clustering

5. DM & ML: Performance Evaluation, Backtesting & False Discoveries

6. DM & ML: Representing & Mining Text

7. Big Data, DM & ML: Ethical & Privacy Issues

8. Big Data and Machine Learning in the Financial Industry

Reading(s):


Sample Keywords:

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical learning (p. 1)</td>
<td>Cross validation (p. 33)</td>
</tr>
<tr>
<td>Classification problems (p.28)</td>
<td>Expected test MSE (p. 34)</td>
</tr>
<tr>
<td>Semi-supervised learning (p. 28)</td>
<td>Bias (p. 35)</td>
</tr>
<tr>
<td>Quantitative variables (p. 28)</td>
<td>Bias-variance trade-off (p. 36)</td>
</tr>
<tr>
<td>Qualitative response (p. 28)</td>
<td>Error rate (p.37)</td>
</tr>
<tr>
<td>Binary response (p.28)</td>
<td>Indicator variable (p. 37)</td>
</tr>
<tr>
<td>Regression (p.28)</td>
<td>Training error (p. 37)</td>
</tr>
<tr>
<td>Predictors (p. 29)</td>
<td>Test error (p.37)</td>
</tr>
<tr>
<td>Mean squared error (MSE) (p. 29)</td>
<td>Bayes classifier (p. 37)</td>
</tr>
<tr>
<td>Training MSE (p. 30)</td>
<td>Conditional probability (p. 37)</td>
</tr>
<tr>
<td>Test data (p. 30)</td>
<td>Bayes decision boundary (p. 38)</td>
</tr>
<tr>
<td>Test MSE (p. 30)</td>
<td>Bayes error rate (p. 38)</td>
</tr>
<tr>
<td>Flexibility (p. 31)</td>
<td>K-nearest neighbors (p. 39)</td>
</tr>
<tr>
<td>Degrees of freedom (p. 32)</td>
<td></td>
</tr>
</tbody>
</table>
Sample Learning Objective:

2.2.6 Learning with gradient descent

For example:

A. Recognize a quadratic cost function of weights and biases and alternative terminology for the cost function.

B. Explain why minimizing a quadratic cost function is preferable to working with other types of cost functions.

C. Recognize an equation for an update rule that defines the gradient descent algorithm and explain the purpose of each component in the equation.

D. Explain how quickly stochastic gradient descent can speed up learning given a training set size n and a mini-batch size, m.

Sample Question:

What is the best description of the gradient descent algorithm as it applies to neural networks?

- It iteratively and incrementally changes biases of a neuron.
- It iteratively and incrementally changes weights of a neuron.
- It iteratively and incrementally changes weights and biases of a neuron.*

Answer key ⇒ LO 2.2.6. p. 26
FDP Curriculum

1. Introduction to Data Science & Big Data

2. DM & ML: Introduction

3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models

4. DM & ML: Classification & Clustering

5. DM & ML: Performance Evaluation, Backtesting & False Discoveries

6. DM & ML: Representing & Mining Text

7. Big Data, DM & ML: Ethical & Privacy Issues

8. Big Data and Machine Learning in the Financial Industry

Reading(s):


Sample Keywords:

- Information (p. 43)
- Tree induction (p. 44)
- Predictive model (p. 45)
- Descriptive modeling (p. 46)
- Target variable (p. 46)
- Attributes or features (p. 46)
- Model induction (p. 47)
- Deduction (p. 47)
- Training data (p. 47)
- Labeled data (p. 47)
- Supervised segmentation (p. 48)
- Information gain (p. 51)
- Entropy (p. 51)
- Parent set (p. 52)
- Child set (p. 52)
- Variance (p. 56)
- Entropy graph/chart (p. 58)
- Classification tree (p. 63)
- Decision nodes (p. 63)
- Laplace correction (p. 73)
- Linear classifier (p. 85)
- Linear discriminant (p. 86)
- Margin (p. 92)
- Support vector machine (p. 92)
- Hinge-loss (p. 94)
- Zero-one loss (p. 95)
- Squared error (p. 95)
- Odds ratio (p. 98)
- Log-odds (p. 99)
- Logistic function (p. 101)
- Generalization (p. 112)
- Overfitting (p. 113)
- Fitting graph (p. 113)
- Holdout data (p. 113)
- Base rate (p. 115)
- Sweet spot (p. 117)
- Cross-validation (p. 126)
- Folds (p. 127)
FDP Curriculum

1. Introduction to Data Science & Big Data

2. DM & ML: Introduction

3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models

4. DM & ML: Classification & Clustering

5. DM & ML: Performance Evaluation, Backtesting & False Discoveries

6. DM & ML: Representing & Mining Text

7. Big Data, DM & ML: Ethical & Privacy Issues

8. Big Data and Machine Learning in the Financial Industry

Sample Learning Objective:

3.2.4 Subset selection

For example:

A. Describe the best subset selection.
B. List the steps used in the best subset selection.
C. Define deviance.
D. Describe the forward stepwise selection and the backward stepwise selection.
E. List the steps used in forward stepwise selection and backward stepwise selection.
F. Recognize and apply the equations for $C_p$, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and adjusted $R^2$.
G. Describe the model selection criterion using $C_p$, AIC, BIC, and adjusted $R^2$.

Sample Questions:

Which of the following approaches to model selection provides a direct estimate of the test error and makes the least number of assumptions about the underlying model?

- $C_p$
- Cross validation*
- Akaike information criterion (AIC)

Answer key ➔ LO: 3.2.4. p.213
1. Introduction to Data Science & Big Data

2. DM & ML: Introduction

3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models

4. DM & ML: Classification & Clustering

5. DM & ML: Performance Evaluation, Backtesting & False Discoveries

6. DM & ML: Representing & Mining Text

7. Big Data, DM & ML: Ethical & Privacy Issues

8. Big Data and Machine Learning in the Financial Industry

**Reading(s):**


**Sample Keywords:**

**Keywords**

- Euclidean distance (p. 144)
- Nearest neighbors (p. 144)
- Combining function (p. 147)
- Weighted voting (p. 150)
- Curse of dimensionality (p. 156)
- Manhattan distance (p. 159)
- Jaccard distance (p. 159)
- Cosine distance (p. 160)

- Edit distance or Levenshtein metric (p. 161)
- Clustering (p. 164)
- Hierarchical clustering (p. 165)
- Dendrogram (p. 165)
- Linkage function (p. 167)
- Cluster center or centroid (p. 170)
- k-means clustering (p. 170)
- Distortion (p. 173)

- Logistic function (p. 132)
- Odds (p. 132)
- Log odds (p. 132)
- Likelihood function (p. 133)
- Principal component analysis (p. 375)

- Loadings (p. 375)
- Bottom-up agglomerative clustering (p. 390)
- Linkage (p. 394)
- Inversion (p. 395)
### Sample Learning Objective:

#### 4.1.1 Similarity and distance

*For example:*
- A. Calculate the Euclidean distance.
- B. Define nearest neighbors and combining function.
- C. Explain how combining function can be used for classification.
- D. Define weighted voting or similarity moderated voting.
- E. Calculate contributions for weighted voting or similarity moderated voting.
- F. Explain how k in k-NN can be used to address overfitting.
- G. Discuss issues with nearest-neighbor methods with focus on
  - i. Intelligibility,
  - ii. Dimensionality and domain knowledge, and
  - iii. Computational efficiency.

### Sample Question:

In the k nearest neighbors (NN) classification system, let n represents the sample size. For what value of k is a k-NN classifier going to make each prediction be the average value of the data set?

- 0
- 1
- n*

**Answer Key ➔** LO: 4.1. 1. p. 153
**FDP Curriculum**

1. Introduction to Data Science & Big Data

2. DM & ML: Introduction

3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models

4. DM & ML: Classification & Clustering

5. DM & ML: Performance Evaluation, Backtesting & False Discoveries

6. DM & ML: Representing & Mining Text

7. Big Data, DM & ML: Ethical & Privacy Issues

8. Big Data and Machine Learning in the Financial Industry

**Reading(s):**


**Sample Keywords:**

<table>
<thead>
<tr>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (p. 189)</td>
</tr>
<tr>
<td>Confusion matrix (p. 189)</td>
</tr>
<tr>
<td>False positive (p. 190)</td>
</tr>
<tr>
<td>False negative (p. 190)</td>
</tr>
<tr>
<td>Expected value (p. 194)</td>
</tr>
<tr>
<td>Class prior (p. 201)</td>
</tr>
<tr>
<td>Precision (p. 204)</td>
</tr>
<tr>
<td>Recall (p. 204)</td>
</tr>
<tr>
<td>F-measure (p. 204)</td>
</tr>
<tr>
<td>Profit curve (p. 212)</td>
</tr>
<tr>
<td>Base rate (p.214)</td>
</tr>
<tr>
<td>ROC graph (p. 215)</td>
</tr>
<tr>
<td>Hit rate (p. 216)</td>
</tr>
<tr>
<td>False alarm rate (p.216)</td>
</tr>
<tr>
<td>AUC (p. 219)</td>
</tr>
<tr>
<td>Cumulative response curve (p. 219)</td>
</tr>
</tbody>
</table>
Sample Learning Objective:

5.3.1 An investigation of the false discovery rate and the misinterpretation of p-values
   For example:
   A. Define specificity and sensitivity.
   B. Describe the false discovery rate with the help of a tree diagram.
   C. Calculate the probability of real effect given a result is significant.
   D. Define power of a test.
   E. Describe the false discovery rate in simulated t-tests.
   F. Calculate false discovery rate.
   G. Describe underpowered study.
   H. Describe the inflation effect in the context of false discovery.
   I. Describe what happens when we consider p=0.05 rather than p<=0.05.
   J. Describe Berger’s approach.

Sample Question:
The performance of 1000 different hedge fund managers is tested. In reality, 10% of the managers actually have abnormal performance. The power of the test is 80% and a p-value of 5% is used. How many managers does the test indicate had abnormal performance?

- 100
- 125*
- 200

Answer Key → LO 5.3.1. p. 10
FDP Curriculum

1. Introduction to Data Science & Big Data

2. DM & ML: Introduction

3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models

4. DM & ML: Classification & Clustering

5. DM & ML: Performance Evaluation, Backtesting & False Discoveries

6. DM & ML: Representing & Mining Text

7. Big Data, DM & ML: Ethical & Privacy Issues

8. Big Data and Machine Learning in the Financial Industry

Reading(s):

Sample Keywords:

**Keywords**

- Linguistic structure (p. 252)
- Dirty data (p. 252)
- Document (p. 253)
- Token (p. 253)
- Terms (p. 253)
- Corpus (p. 253)
- Bag of words (p. 254)
- Term frequency (p. 254)
- Stemmed (p. 255)
- Stopwords (p. 255)
- Inverse document frequency (p. 256)
- n-grams (p. 265)
- Latent information model (p. 268)
- Information triage (p. 276)

**Keywords**

- Sentiment analysis (p. 1)
- Probabilistic classifier (p. 2)
- Generative classifier (p. 2)
- Discriminative classifier (p. 2)
- Linear classifier (p. 5)
- Sentiment lexicon (p. 9)
- Gold labels (p. 11)
- Precision (p. 12)
- Recall (p. 12)
- F-measure (p. 13)
- Macroaveraging (p. 13)
- Microaveraging (p. 13)
# FDP Curriculum

| 1. Introduction to Data Science & Big Data |
| 2. DM & ML: Introduction |
| 3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models |
| 4. DM & ML: Classification & Clustering |
| 5. DM & ML: Performance Evaluation, Backtesting & False Discoveries |
| 6. DM & ML: Representing & Mining Text |
| 7. Big Data, DM & ML: Ethical & Privacy Issues |
| 8. Big Data and Machine Learning in the Financial Industry |

## Sample Learning Objective:

**6.1.2 Text representation**

*For example:*

A. Understand the meaning of terms when used in the field of information retrieval.
B. Describe the “bag of words” approach including the following steps:
   - Measuring term frequency (TF)
   - Measuring sparseness: inverse document frequency (IDF)
   - Combining them: TFIDF
C. Apply appropriate methods to search an example set of documents.
D. Express entropy in terms of the IDF.

## Sample Question:

When representing and mining text using the “bag of words” approach, which of the following is a common name for a measure of term sparseness?

- Term infrequency
- Inverse document frequency*
- Inverse corpus document frequency

*Answer Key ➔ LO 6.1.2. p.256
FDP Curriculum

1. Introduction to Data Science & Big Data

2. DM & ML: Introduction

3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models

4. DM & ML: Classification & Clustering

5. DM & ML: Performance Evaluation, Backtesting & False Discoveries

6. DM & ML: Representing & Mining Text

7. Big Data, DM & ML: Ethical & Privacy Issues

8. Big Data and Machine Learning in the Financial Industry

---

**Reading(s):**


**Sample Keywords:**

<table>
<thead>
<tr>
<th>Key Terms</th>
<th>Veracity (p. 6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data trust deficit (p. 2)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Code of ethics (p. 6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial intelligence (p. 1)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Keywords</th>
<th>People risk (p. 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Data Protection Regulation (p. 1)</td>
<td></td>
</tr>
</tbody>
</table>
FDP Curriculum

1. Introduction to Data Science & Big Data
2. DM & ML: Introduction
3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models
4. DM & ML: Classification & Clustering
5. DM & ML: Performance Evaluation, Backtesting & False Discoveries
6. DM & ML: Representing & Mining Text
7. Big Data, DM & ML: Ethical & Privacy Issues
8. Big Data and Machine Learning in the Financial Industry

Sample Learning Objective:

7.3.1 General Data Protection Regulation (GDPR)
For example:
A. Describe the primary purpose of the GDPR.
B. Describe the key changes in data protection regulation including the meaning of
   • rights of the individual
   • informed consent
   • notification
   • data portability
   • supervision and enforcement, and
   • liability

Sample Question:

According to the General Data Protection Regulation (GDPR) what is the longest period that a company can take to report a security breach to blank without penalty?

- 24 hours
- 48 hours
- 72 hours*

Answer Key ➔ LO 7.3.1, p. 2
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Introduction to Data Science &amp; Big Data</td>
</tr>
<tr>
<td>2.</td>
<td>DM &amp; ML: Introduction</td>
</tr>
<tr>
<td>3.</td>
<td>DM &amp; ML: Regression, LASSO, Predictive Models, Time Series &amp; Tree Models</td>
</tr>
<tr>
<td>4.</td>
<td>DM &amp; ML: Classification &amp; Clustering</td>
</tr>
<tr>
<td>5.</td>
<td>DM &amp; ML: Performance Evaluation, Backtesting &amp; False Discoveries</td>
</tr>
<tr>
<td>6.</td>
<td>DM &amp; ML: Representing &amp; Mining Text</td>
</tr>
<tr>
<td>7.</td>
<td>Big Data, DM &amp; ML: Ethical &amp; Privacy Issues</td>
</tr>
<tr>
<td>8.</td>
<td>Big Data and Machine Learning in the Financial Industry</td>
</tr>
</tbody>
</table>

**Reading(s):**


FDP Curriculum

Sample Keywords:

1. Introduction to Data Science & Big Data

2. DM & ML: Introduction

3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models

4. DM & ML: Classification & Clustering

5. DM & ML: Performance Evaluation, Backtesting & False Discoveries

6. DM & ML: Representing & Mining Text

7. Big Data, DM & ML: Ethical & Privacy Issues

8. Big Data and Machine Learning in the Financial Industry

Keywords

- Big data (p. 4)
- Artificial intelligence (p. 4)
- Natural language processing (p. 5)
- Machine learning (ML) (p. 4)
- Supervised learning (p. 5)
- Unsupervised learning (p. 5)
- Deep learning (p. 5)
- Reinforcement learning (p. 5)
- Sentiment indicators (p. 10)
- Trading signals (p. 11)
- Fraud detection (p. 11)
- RegTech (p. 11)
- InsurTech (p. 13)
- Chatbots (p. 14)
- Know your customer (KYC) (p. 20)
- SupTech (p. 21)
- Auditability (p. 33)
- Fintech (p. 35)
- Robo-advisors (p. 35)
- Tonality analysis (p. 36)

Keywords

- Fintech (p. 79)
- Robo-advisor (p. 80)
- Work-flow (p. 83)
- D2C platforms (p. 86)
- Hybrid (p. 86)
- B2B platforms (p. 86)

Keywords

- Alternative data (p. 14)
- Social media (p. 14)
- Microdata (p. 14)
- Data exhaust (p. 14)
- Rivalry (p. 16)
- Excludability (p. 16)
- Defensive strategies (p. 17)
- Defensible strategies (p. 18)
- Operational alpha (p. 19)
- Aggregation (p. 19)
- Disaggregation (p. 19)
- Volume (p. 21)
- Velocity (p. 21)
- Variety (p. 21)
- Veracity (p. 21)
- Granularity (p. 21)
- Relationality (p. 21)
- Flexibility (p. 21)
- Actionability (p. 22)
- Excludable (p. 28)
- Data hoarding (p. 29)
Sample Learning Objective:
8.10.1 Using linguistic analysis to perform risk analysis of investments.

For example:
A. Explain the difficulties associated with manual parsing of unstructured text.
B. Describe the concept of RegTech.
C. Describe how content and structure of emails could be used for risk analysis.
D. Explain the effectiveness of textual versus hard numbers in corporate risk analysis.
E. Define RegTech.
F. Apply the net sentiment metric to calculate the polarity of a text using its Pos and Neg figures.
G. Calculate the disagreement measure of a text using Pos and Neg figures of a text.
H. Describe the findings of the paper regarding the effectiveness of email length as predictor of risk analysis of Enron.

Sample Question:
According to the article “Zero-Revelation RegTech: Detecting Risk through Linguistic Analysis of Corporate Emails and News,” what does the decomposition of the ‘document term matrix’ facilitate?

- Topic sentiment*
- Network activity
- Vocabulary trends

Answer Key → LO 8.10.1. p. 28-29
Marilyn Taylor is a quantitative analyst responsible for finding new investment ideas in the equity space. In the past, she has used unsupervised learning techniques to filter firms from a larger list, and this technique seems to have worked well for her. She mainly considers firm characteristics, such as P/E ratio, P/B ratio, and size, for her analysis. Recently, she has come across Volta Electric Company, which looks interesting to her. Currently, it has a P/E ratio of 16.5, a P/B ratio of 2.3, and a size value of 4.6 billion. To compare Volta with some other companies, she pulled out information on 3 companies that she had analyzed in the past. Out of the 3 companies, 2 were recommended for investment and 1 was not recommended. The following table provides information on the three companies.
Essay Questions

Sample Question (cont.):

A. Marilyn uses Euclidean distance to measure the difference between a new company and the companies she has analyzed in the past. What are the Euclidean distances between Volta and the 3 companies listed in the table above?

<table>
<thead>
<tr>
<th>Name</th>
<th>P/E Ratio</th>
<th>P/B Ratio</th>
<th>Size ($ billions)</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern Healthcare</td>
<td>14</td>
<td>2.1</td>
<td>5.9</td>
<td>Invest</td>
</tr>
<tr>
<td>Wholesome Foods</td>
<td>17</td>
<td>1.5</td>
<td>6.7</td>
<td>Invest</td>
</tr>
<tr>
<td>Real Tech</td>
<td>21</td>
<td>1.8</td>
<td>7.8</td>
<td>Do not invest</td>
</tr>
</tbody>
</table>

If we take a majority vote of the distances, what would be the recommendation by Marilyn for Volta? Explain how you decided on this.
The recommendation would be to invest in Volta. If we consider the distances between Volta and the three companies analyzed earlier, we see that two of the companies are quite close in terms of the Euclidean distance and the recommendations for both of these companies is to invest in them. Majority voting requires us to look at all the points that are close to point we are trying to classify and assigns the majority class from the closest points to the new point.

<table>
<thead>
<tr>
<th>Euclidean distance between Volta and:</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern Healthcare</td>
<td>(\sqrt{(16.5 - 14)^2 + (2.3 - 2.1)^2 + (4.6 - 5.9)^2} = 2.82)</td>
</tr>
<tr>
<td>Wholesome Foods</td>
<td>(\sqrt{(16.5 - 17)^2 + (2.3 - 1.5)^2 + (4.6 - 6.7)^2} = 2.30)</td>
</tr>
<tr>
<td>Real Tech</td>
<td>(\sqrt{(16.5 - 21)^2 + (2.3 - 1.8)^2 + (4.6 - 7.8)^2} = 5.54)</td>
</tr>
</tbody>
</table>
Sample Question:

B. Marilyn also calculates weighted voting or similarity moderated voting for any new company using the companies she has analyzed in the past.

What are the similarity weights and contribution to the probability of investing or not investing for the 3 companies?

What would be the estimated probability of investing and not investing for Volta?
### Constructive Response Questions

**Sample Answer:**

<table>
<thead>
<tr>
<th>Name</th>
<th>Euclidean Distance</th>
<th>Similarity Weight</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern Healthcare</td>
<td>2.82</td>
<td>( \frac{1}{2.82^2} = 0.125748 )</td>
<td>( \frac{0.125748}{0.347366} = 0.362 )</td>
</tr>
<tr>
<td>Wholesome Foods</td>
<td>2.30</td>
<td>( \frac{1}{2.30^2} = 0.189036 )</td>
<td>( \frac{0.189036}{0.347366} = 0.544 )</td>
</tr>
<tr>
<td>Real Tech</td>
<td>5.54</td>
<td>( \frac{1}{5.54^2} = 0.032582 )</td>
<td>( \frac{0.032582}{0.347366} = 0.094 )</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td></td>
<td><strong>0.347366</strong></td>
<td></td>
</tr>
</tbody>
</table>

Estimated probability of investing = 0.362 + 0.544 = 0.906

Estimated probability of not investing = 0.094
THE FINANCIAL DATA PROFESSIONAL PROGRAM

MARCH 16 - APRIL 4, 2020

CANDIDATE EXPERIENCE
FDP Exam Candidates

FDP Program
- Exam Registration
- Exam Curriculum
- FAQ’s
- Webinars & Events

If you are registered for the exam, we urge you to reserve your seat as soon as possible:

https://www.prometric.com/test-takers/search/fdpi
Online Resources (for registered exam candidates)

- Candidate Handbook
- Study guide

*Both are exam specific*

*From the handbook: Prometric tutorials*
Online Resources (for registered exam candidates)

- Curriculum and Study Tools
- Order the PMR E-Book
- Download the (12) readings from Topics in Financial Data Science
- Sample Questions (registered exam candidates have access to a webpage with directions)
Study Time

- Average time for the FDP exam is 200 – 250 hours
- Plan your study time
Exam Day

• **ID Policy**
  - Two valid (not expired) forms of ID
  - Make sure your name in your FDP profile MATCHES that on your ID
  - Both must have a signature
  - One must have a photo
  - Passport preferred
  - The name on your IDs must exactly match the name with which you schedule your exam.

• **Calculator Policy**
  - Texas Instruments BA II Plus (including the TI BA Plus Professional)
  - Exam proctors will inspect your calculator prior to the start of the exam.
Exam Session

- You will receive a dry-erase white board

<table>
<thead>
<tr>
<th>Activity</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-disclosure agreement (must be completed within 5 minutes)</td>
<td>5 minutes</td>
</tr>
<tr>
<td>Exam instructions</td>
<td>5 minutes</td>
</tr>
<tr>
<td>FDP Multiple Choice (MC) Questions</td>
<td>100 minutes</td>
</tr>
<tr>
<td>Break (optional)</td>
<td>20 minutes</td>
</tr>
<tr>
<td>FDP MC Questions (approx. 5) and 2-4 Constructive Response Questions</td>
<td>100 minutes</td>
</tr>
<tr>
<td>Comment Period (optional)</td>
<td>10 minutes</td>
</tr>
<tr>
<td><strong>Total Exam Session Time</strong></td>
<td><strong>4 hours</strong></td>
</tr>
</tbody>
</table>
Pass Rate

November Exam ➔ 77%

March – April Exam ➔ On our way to double the number of participants
Candidate Performance Report

- The FDP Candidate Performance Report is intended to aid in self-assessment by indicating your areas of relative strength and weakness among the topics in the exam, compared to a reference group.

- The reference group is comprised of all candidates whose total test scores were within the bottom quartile of those who passed the exam during the testing window in which you sat for the exam.

<table>
<thead>
<tr>
<th>FDP Examination Topics</th>
<th>Relative Performance Level*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction to Data Science &amp; Big Data</td>
<td>Above</td>
</tr>
<tr>
<td>2. Data Mining &amp; Machine Learning: Introduction</td>
<td>Above</td>
</tr>
<tr>
<td>4. Data Mining &amp; Machine Learning: Classification &amp; Clustering</td>
<td>Above</td>
</tr>
<tr>
<td>5. Data Mining &amp; Machine Learning: Performance Evaluation, testing &amp; False Discoveries</td>
<td>Above</td>
</tr>
<tr>
<td>6. Data Mining &amp; Machine Learning: Representing &amp; Mining Text</td>
<td>Above</td>
</tr>
</tbody>
</table>
Kind reminders of upcoming webinars as we go through the Q & A

Guen Dondé
Head Of Research at Institute of Business Ethics.
Co-researcher "IBE Ethics at Work" and "Implications of AI on business ethics
January 22, 2020 - 11am EST

A CONVERSATION WITH
Guen Dondé
Head Of Research
Institute of Business Ethics
January 22, 2020
11am EST

Gene Getman
Co-Author of "Big Data is a Big Deal: An investor's guide to the applications and challenges of alternative data"
January 29, 2020 - 2pm EST

A CONVERSATION WITH
Gene Getman
Co-Author of "Big Data is a Big Deal"
January 29, 2020
2pm EST

Seoyoung Kim
Co-Author of "Zero-Revelation RegTech: Detecting Risk through Linguistic Analysis of Corporate Emails and News!"
February 19, 2020 - 1pm EST

A CONVERSATION WITH
Dr. Seoyoung Kim
Professor of Business Analytics at Santa Clara University
February 19, 2020
1 pm EST

Some of the topics that will be discussed:

- What is the impact of the GDPR from an ethical standpoint?
- Can codes of ethics help?
- How involved should regulators be?

Note: Three readings on the upcoming exam are published by the Institute for Business Ethics.
Our conversation will last approximately 35 minutes with the remaining time for Q&A.

Gene Getman is a Client Portfolio Manager for the 1798 Alternatives group, focused on Investor Relations for the Hedge Fund business. He is the Product Specialist for the 1798 Q Strategy and other innovative or capacity constrained alternative investment strategies based in New York. Before joining LOIM in April 2013, Gene was an Analyst on the Barclays Capital Solutions team, where he worked with Pensions, Endowments, Foundations and their consultants to make capital introductions to hedge funds. He began his career working for the NYSE Market Makers group within Lehman Brothers and Barclays Capital.

Note: The article "Big Data is a Big Deal" is one of the readings for the upcoming exam.

Natural language processing is a fast-growing area of data science for the finance industry. Recent advances in financial technology (FinTech) have dramatically transformed the financial landscape with respect to the way we access, invest, and transfer financial capital. In this article, the authors explore a promising avenue for the use of natural-language processing in an effective yet non-invasive method by which to monitor the health and integrity of financial institutions and corporations in general by analyzing corporate emails and news.

Note: The article "Zero-Revelation RegTech: Detecting Risk through Linguistic Analysis of Corporate Emails and News!" is one of the readings for the upcoming exam.
We wish you continued success with your exam preparations!

If you have additional questions, please contact us at info@fdpinsttitute.org