IBM

Certification

Data & AI

IBM Cloud Professional Certification Program

Study Guide Series

Exam C1000-144: IBM Machine Learning Data Scientist v1

Contents

Purpose of Exam Objectives 2
High-level Exam Objectives 3
Detailed Exam Objectives 4
Section 1 - Evaluate business problem including ethical implications 4
Section 2 – Exploratory Data Analysis including data preparation 6
Section 3 – Implement the proper model 8
Section 4 – Refine and deploy the model 11
Section 5 – Monitor models in production 14
Next Steps 18

Purpose of Exam Objectives

When an exam is developed, Subject Matter Experts work together to define the role the certified individual will fill. They define the tasks and knowledge that an individual would need to successfully perform this job role doe the product or solution. This creates the foundation for the objectives and measurement criteria, which form the basis of the certification exam. Question writers then use these objectives to develop exam questions.

It is recommended that you review these objectives and ask yourself the following questions:

- Do you know how to complete the task in the objective?
- Do you know why that task needs to be done?
- Do you know what will happen if you do it incorrectly?

If you are not familiar with a task, go through the objective, perform that task in your own environment and read more information on the task. If there is an objective on a task, there is a high likelihood that you WILL see a question about it on the actual exam. Review the recommended learning designed to prepare you to take the certification exam.

After reviewing the objectives in this guide and completing your own research, take the assessment exam. While the assessment exam does not indicate which specific questions were answered incorrectly, it does indicate overall performance by section. This is a good indicator of preparedness or if further preparation is warranted.

High-level Exam Objectives

Section 1 - Evaluate business problem including ethical implications	
1.1	Understand business requirements
1.2	Understand what data is available
1.3	Understand ethical challenges in the business problem
1.4	Perform AI design thinking
1.5	Assess progress on the AI Ladder
Section 2 - Exploratory Data Analysis including data preparation	
2.1	Identify the methods used to clean, label, and anonymize data
2.2	Visualize data
2.3	Balance and partition data
Section 3 - Implement the proper model	
3.1	Implement Supervised Learning: Regression
3.2	Implement Supervised Learning: Classification
3.3	Implement Unsupervised Learning: Clustering
3.4	Implement Unsupervised Learning: Dimensional Reduction
Section 4 -Refine and deploy the model	
4.1	Identify operations and transformations taken to select and engineer features
4.2	Select the proper tools
4.3	Configure the appropriate environment specifications for training the model
4.4	Train the model and optimize hyperparameters
4.5	Implement the ability for the model to explain itself
4.6	Deploy the model
Section 5 -Monitor models in production	
5.1	Assess the model
5.2	Monitor the model in production
5.3	Determine if there is unfair bias in the model

Detailed Exam Objectives

Section 1 - Evaluate business problem including ethical implications

1.1. Understand business requirements

SUBTASKS:

1.1.1. Explain how IBM Garage Methodology works
1.1.2. Understand the CRISP-DM process
1.1.3. Identify which business opportunity to prioritize and define success metrics for an MVP

REFERENCES:

https://www.ibm.com/garage
https://thinkinsights.net/digital/crisp-dm/

1.2. Understand what data is available

SUBTASKS:

1.2.1. Use SQL to access data
 1.2.1.1. Extracting specific columns
 1.2.1.2. Filtering data
 1.2.1.3. Combining Tables

1.2.2. Use Python APIs to access data
1.2.3. Scrape information from a website
1.2.4. Read data into a Pandas Dataframe
 1.2.4.1. Reading different types of data assets
 1.2.4.2. Manipulating column names
 1.2.4.3. Obtaining specific rows

REFERENCES:

https://www.w3schools.com/sql/
https://realpython.com/python-api/
https://www.crummy.com/software/BeautifulSoup/bs4/doc/
https://pandas.pydata.org/docs/getting_started/index.html

1.3. Understand ethical challenges in the business problem

SUBTASKS:

1.3.1. List potential sources of unfair bias

1.3.2. List potential sources of privacy violations
1.3.3. List potential secondary and tertiary effects of the application
1.3.4. Plan to prevent or mitigate negative consequences

REFERENCES:
https://learn.ibm.com/course/view.php?id=8390
https://learning.oreilly.com/library/view/ai-fairness/9781492077664/ Introduction
https://www.ibm.com/design/thinking/page/courses/AI_Essentials > Clearbridge case study > Reflect on your AI's capabilities https://www.designethically.com/layers
https://www.ibm.com/design/ai/ethics/

1.4. **Perform AI design thinking**
SUBTASKS:
1.4.1. Align on user intents for a solution
1.4.2. Document the available data
1.4.3. Determine what training will be required
1.4.4. Create hypotheses about what the behavior of the system will be
1.4.5. Assess feasibility and refine if needed
1.4.6. Consider direct and indirect effects of the solution

REFERENCES:
https://www.ibm.com/design/thinking/page/courses/AI_Essentials
https://www.ibm.com/design/thinking/page/toolkit/activity/ai-essentials-intent
https://www.ibm.com/design/thinking/static/team-essentials-for-ai-workbook-8dc9aadb2cc2dc6343cc5e420b522ca2.pdf
https://learning.oreilly.com/library/view/operationalizing-ai/9781098101329/ --- Chapter 3

1.5. Assess progress on the AI Ladder

SUBTASKS:
1.5.1. Assess progress in collecting data
1.5.2. Assess progress in organizing data
1.5.3. Assess progress in analyzing data
1.5.4. Assess progress in infusing AI into the organization

REFERENCES:

https://www.ibm.com/downloads/cas/O1VADKY2
https://learn.ibm.com/course/view.php?id=8496

Section 2 – Exploratory Data Analysis including data preparation

2.1. Identify the methods used to clean, label, and anonymize data SUBTASKS:

2.1.1. Clean data
- 2.1.1.1. Fill or drop missing values
- 2.1.1.2. Remove duplicate rows
- 2.1.1.3. Remove outliers
- 2.1.1.4. Converting data types
- 2.1.1.5. Data normalization

2.1.2. Label data
- 2.1.2.1. Understand the benefits and challenges to labeling data
- 2.1.2.2. Explain data labeling approaches

2.1.3. Anonymize data

REFERENCES:
https://www.ibm.com/garage/method/practices/reason/prepare-data-for-machinelearning/
https://www.ibm.com/garage/method/practices/code/data-preparation-ai-data-science/
https://www.ibm.com/cloud/learn/data-labeling
https://dataplatform.cloud.ibm.com/docs/content/wsj/governance/dmg22.html

2.2. Visualize data
SUBTASKS:

2.2.1. Choose the column(s) from your dataset to be visualized

2.2.2. Identify what the visualization should describe about the column(s)
- 2.2.2.1. Distribution
- 2.2.2.2. Correlation
- 2.2.2.3. Comparison
- 2.2.2.4. Time Series

2.2.3. Select a type of chart based on the descriptive need
- 2.2.3.1. Histogram/Box plot/Violin plot
- 2.2.3.2. Scatterplot/Heatmap
- 2.2.3.3. Bar chart
- 2.2.3.4. Line plot

2.2.4. Select a library or tool for visualization
- 2.2.4.1. Matplotlib
- 2.2.4.2. Seaborn

2.2.4.3. Bokeh
2.2.4.4. Plotly
2.2.5. Plot the visualization

REFERENCES:
https://seaborn.pydata.org/introduction.html
https://matplotlib.org/stable/tutorials/introductory/usage.html#sphx-glr-tutorialsintroductory-usage-py
https://docs.bokeh.org/en/latest/docs/first_steps.html https://plotly.com/python/
https://learn.ibm.com/course/view.php?id=8794
https://learning.oreilly.com/library/view/statistics-in-a/9781449361129/ Chapter 4

2.3. Balance and partition data
SUBTASKS:
2.3.1. Partition data
2.3.1.1. Create train/test/validation splits

REFERENCES:
https://learn.ibm.com/mod/video/view.php?id=165773 (data leakage mentioned in passing)

2.3.1.2. Understand and implement cross validation

REFERENCES:
https://learn.ibm.com/mod/video/view.php?id=166655
https://learn.ibm.com/mod/page/view.php?id=170328&forceview=1
2.3.1.3. Prevent data leakage

REFERENCES:
https://en.wikipedia.org/wiki/Leakage_(machine_learning)
https://reproducible.cs.princeton.edu/ (this is a common problem)
2.3.1.4. Create data splits that are reproducible

REFERENCES:
https://cs230.stanford.edu/blog/split/
https://learn.ibm.com/mod/video/view.php?id=166646&forceview=1
https://learning.oreilly.com/library/view/machine-learningdesign/9781098115777/ch06.html#problem-id00022

2.3.2. Balance data

2.3.2.1. Understand why imbalanced data is problematic

REFERENCES:
https://learn.ibm.com/mod/video/view.php?id=167242
https://learn.ibm.com/mod/video/view.php?id=168614
https://learn.ibm.com/mod/page/view.php?id=170229&forceview=1

2.3.2.2. Understand and implement pros, cons, and how to of up-, down-, and re-sampling

REFERENCES:
https://learn.ibm.com/mod/video/view.php?id=167243
https://learn.ibm.com/mod/video/view.php?id=167247
https://learn.ibm.com/mod/video/view.php?id=167246
https://learn.ibm.com/mod/page/view.php?id=170230&forceview=1
https://learn.ibm.com/mod/video/view.php?id=168610&forceview=1

2.3.2.3. Understand and implement other methods to handle imbalanced data, such as weighting and stratified sampling

REFERENCES:
https://learn.ibm.com/mod/video/view.php?id=167245 https://imbalanced-learn.org/stable/index.html (included in videos)
https://learn.ibm.com/mod/page/view.php?id=170231

Section 3 – Implement the proper model

3.1. Implement Supervised Learning: Regression
SUBTASK(S):
3.1.1. Describe Regression

REFERENCES:
https://towardsdatascience.com/supervised-learning-basics-of-linear-regression1cbab48d0eba

3.1.2. Understand the benefits of Regression

REFERENCES:
https://towardsdatascience.com/supervised-learning-the-what-when-why-good-and-badpart-1-f90e6fe2a606

3.1.3. Understand some of the most popular Regression algorithms
3.1.3.1. Gradient Boosting Tree
3.1.3.2. Neural Network
3.1.3.3. Random Forest
3.1.3.4. Linear Regression
3.1.3.5. Decision Tree

REFERENCES:
https://scikit-learn.org/stable/supervised_learning.html

3.2. Implement Supervised Learning: Classification
SUBTASK(S):
3.2.1. Describe Classification

REFERENCES:
https://towardsdatascience.com/supervised-learning-the-what-when-why-good-and-badpart-1-f90e6fe2a606

3.2.2. Understand the benefits of Classification

REFERENCES:
https://www.javatpoint.com/regression-vs-classification-in-machine-learning

3.2.3. Understand some of the most popular Classification algorithms
3.2.3.1. Naïve Bayes
3.2.3.2. Linear SVM
3.2.3.3. Logistic Regression
3.2.3.4. K-Nearest Neighbors
3.2.3.5. Stochastic Gradient Descent
3.2.3.6. Neural Network
3.2.3.7. Decision Trees & Random Forest
3.2.3.8. Boosting Classifiers

REFERENCES:
https://analyticsindiamag.com/7-types-classification-algorithms/

3.3. Implement Unsupervised Learning: Clustering
SUBTASK(S):
3.3.1. Describe Clustering

REFERENCES:
https://machinelearningmastery.com/clustering-algorithms-with-python/

3.3.2. Understand the benefits of Clustering

REFERENCES:
https://www.explorium.ai/blog/clustering-when-you-should-use-it-and-avoid-it/

3.3.3. Understand some of the most popular Clustering algorithms
3.3.3.1. K-means
3.3.3.2. Gaussian Mixture Model
3.3.3.3. DBSSCAN

REFERENCES:
https://scikit-learn.org/stable/modules/clustering.html

Additional REFERENCES:

https://www.statlearning.com/ http://www.mmds.org/
https://scikit-learn.org/stable/modules/mixture.html#gmm
https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf
https://www.dbs.ifi.lmu.de/Publikationen/Papers/OPTICS.pdf

3.4. Implement Unsupervised Learning: Dimensional Reduction
SUBTASK(S):

3.4.1. Describe Dimensional Reduction

REFERENCES:
https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/

3.4.2. Understand the benefits of Dimensional Reduction

REFERENCES:
https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/

3.4.3. Understand some of the most popular Dimensional Reduction Algorithms
3.4.3.1. Singular Value Decomposition
3.4.3.2. Latent Dirichlet Analysis
3.4.3.3. Principal Component Analysis

REFERENCES:

https://machinelearningmastery.com/dimensionality-reduction-algorithms-with-python/

General Reference for differentiation:
https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

Additional References
https://www.statlearning.com/
http://www.mmds.org/
https://geometria.math.bme.hu/sites/geometria.math.bme.hu/files/users/csgeza/howardanton-chris-rorres-elementary-linear-algebra-applications-version-11th-edition.pdf
https://jmlr.org/papers/volume18/14-546/14-546.pdf
https://en.wikipedia.org/wiki/Curse_of_dimensionality

Section 4 – Refine and deploy the model

4.1. Identify operations and transformations taken to select and engineer features
SUBTASK(S):
4.1.1. Obtain raw data
4.1.2. Engineer features using attributes of the raw data
4.1.3. Use automated techniques to augment and/or select features for use in learning

REFERENCES:
https://machinelearningmastery.com/discover-feature-engineering-how-to-engineerfeatures-and-how-to-get-good-at-it/
https://developer.ibm.com/articles/automated-feature-engineering-for-relational-datawith-ibm-autoai/
https://developer.ibm.com/patterns/model-mgmt-on-watson-studio-local/

4.2. Select the proper tools
SUBTASK(S):
4.2.1. Identify the tools required based on the:
    4.2.1.1. Model type
    4.2.1.2. Type of data
    4.2.1.3. Feature engineering requirements
    4.2.1.4. Amount of automation desired
    4.2.1.5. Production environment requirements

REFERENCES:

https://www.ibm.com/garage/method/practices/reason/evaluate-and-select-machinelearning-algorithm/
https://developer.ibm.com/articles/cc-models-machine-learning/
https://www.ibm.com/support/producthub/icpdata/docs/content/SSQNUZ_latest/wsj/analyze-data/ml-overview_local.html
https://www.ibm.com/support/producthub/icpdata/docs/content/SSQNUZ_latest/wsj/getting-started/tools.html

4.3. Configure the appropriate environment specifications for training the model
SUBTASK(S):
4.3.1. Identify the frameworks supported by Watson Machine Learning
4.3.2. Explain the GPU-accelerated computing
REFERENCES:
https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/ml-overview.html
https://dataplatform.cloud.ibm.com/docs/content/wsj/analyzedata/pm_service_supported_frameworks.html

4.4. Train the model and optimize hyperparameters
SUBTASK(S):
4.4.1. Choose and justify the type of algorithm
 4.4.1.1. Regression
 4.4.1.2. Classification
 4.4.1.3. Clustering
 4.4.1.4. Recommendation engines
 4.4.1.5. Anomaly detection
4.4.2. Describe the trade-offs between underfitting and overfitting a model
 4.4.2.1. Avoid underfitting or overfitting by splitting the data into training, testing, and validation sets
4.4.3. Compare model parameters and hyperparameters
4.4.4. Explain hyperparameters and hyperparameter tuning
 4.4.4.1. Tuning is a trial-and-error process
 4.4.4.2. Tuning is based on the training output loss value
 4.4.4.3. Learning rate, number of epochs, hidden layers, hidden units, activation functions
4.4.5. Summarize search algorithms
 4.4.5.1. Grid Search
 4.4.5.2. Random Search
 4.4.5.3. Bayesian Optimization
4.4.6. Ensemble multiple models
4.4.7. Choose and justify the type of algorithm
 4.4.7.1. Regression

4.4.7.2. Classification
4.4.7.3. Clustering
4.4.7.4. Recommendation engines
4.4.7.5. Anomaly detection

REFERENCES:
https://www.ibm.com/garage/method/practices/reason/optimize-train-ai-model/
https://www.ibm.com/docs/en/wmla/2.2.0?topic=optimization-hyperparameter-searchalgorithms
https://www.ibm.com/garage/method/practices/reason/evaluate-and-select-machinelearning-algorithm/
https://developer.ibm.com/articles/cc-models-machine-learning/
https://www.ibm.com/garage/method/practices/reason/evaluate-and-select-machinelearning-algorithm/
https://developer.ibm.com/articles/cc-models-machine-learning/

4.5. Implement the ability for the model to explain itself
SUBTASK(S):
4.5.1. Determine what user profiles need explanations
4.5.2. Determine what sort of explanations will make sense to those users
4.5.3. Select and apply algorithms to generate model explanations
4.5.3.1. Boolean Decision Rule
4.5.3.2. Generalized Linear Rule Model
4.5.3.3. ProfWeight
4.5.3.4. Teaching Explanations for Decisions (TED)
4.5.3.5. Contrastive Explanations
4.5.3.6. Disentangled Inferred Prior VAE
4.5.3.7. ProtoDash
4.5.4. Present expalantions in a form that will make sense to the target users

REFERENCES:
https://learn.ibm.com/course/view.php?id=8717
https://learn.ibm.com/course/view.php?id=8718

https://aix360.mybluemix.net/

4.6. Deploy the model

SUBTASK(S):
4.6.1. Containerize the model with Docker
4.6.2. Embed the model into Spark

4.6.3. Deploy the model with Watson Machine Learning

REFERENCES:
https://learn.ibm.com/course/view.php?id=8797

Section 5 – Monitor models in production

5.1. Assess the model
SUBTASK(S):
5.1.1. Distinguish metrics for Classification Models
    5.1.1.1. Explain how a confusion matrix works
    5.1.1.2. Explain what AUC measures
    5.1.1.3. ROC curve

Reference for plots with ROC curves:
https://people.inf.elte.hu/kiss/11dwhdm/roc.pdf
https://synapse.koreamed.org/articles/1027596

    5.1.1.4. Difference in distance measurements
        5.1.1.4.1. Manhatta
        5.1.1.4.2. Euclidean
        5.1.1.4.3. Cosine similarity
REFERENCES:
https://learning.oreilly.com/library/view/thoughtful-machine-learning/9781491924129/
Chapter 3, Distances
https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

    5.1.1.5. Understand how tree-based models determine features to split on

5.1.2. Distinguish metrics for Regression Models
    5.1.2.1. How do L1 and L2 Regularization impact the model features
    5.1.2.2. Understand distinction between Bias and Variance
    5.1.2.3. What do MSE and R-Squared measure
    5.1.2.4. Understand common error metrics to evaluate regression models

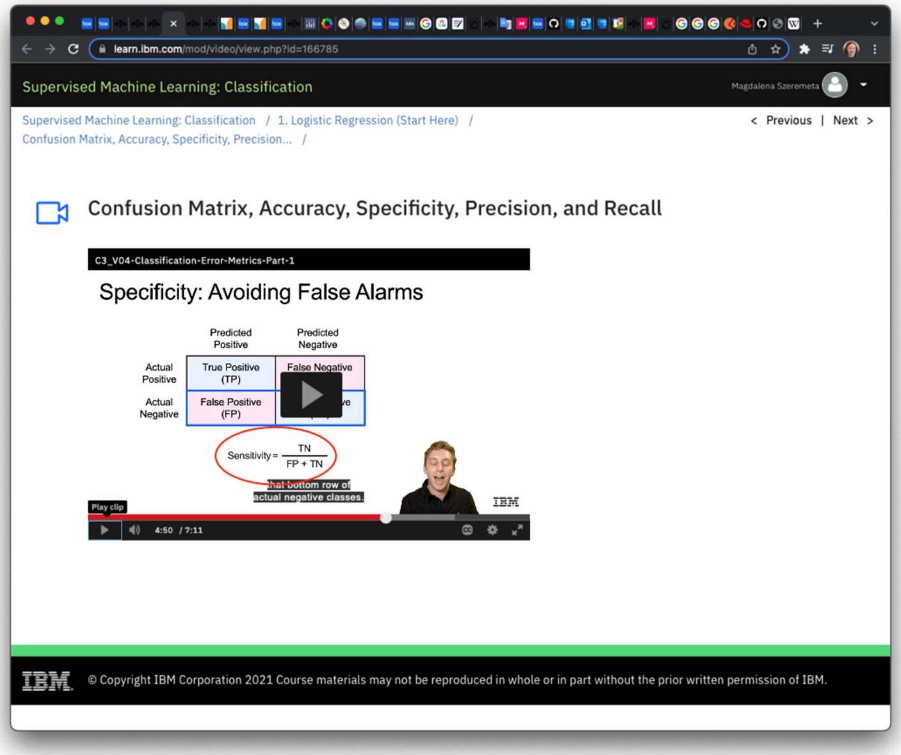
5.1.3. Distinguish metrics for Unsupervised Models
    5.1.3.1. How do you determine optimal number of K for K-Means Algorithm
    5.1.3.2. Explain how the inertia metric is calculated
    5.1.3.3. Explain how the Distortion metric is calculated

5.1.3.4. How can you avoid your centroids getting stuck in bad local optima

5.1.4. Identify trade-offs between model performance and computational cost
5.1.5. Choose the best metric for the model and business problem

REFERENCES:
https://scikit-learn.org/stable/modules/model_evaluation.html
https://learn.ibm.com/mod/video/view.php?id=166640
https://learn.ibm.com/mod/page/view.php?id=170322&forceview=1
https://learn.ibm.com/mod/video/view.php?id=166668
https://learn.ibm.com/mod/video/view.php?id=166669
https://learn.ibm.com/mod/video/view.php?id=166785
https://learn.ibm.com/mod/page/view.php?id=170325&forceview=1
https://learn.ibm.com/mod/video/view.php?id=166786
https://learn.ibm.com/mod/page/view.php?id=170329&forceview=1
https://learn.ibm.com/mod/video/view.php?id=169061&forceview=1

There is an error in learning material minute 5:10:
https://learn.ibm.com/mod/video/view.php?id=166785 The presenter is talking about specificity and the formula for specificity is displayed, but it is incorrectly signed as “Sensitivity”. The next slide has corrected version of formula.



learn.ibm.com/mod/video/view.php?id=166785

Supervised Machine Learning: Classification

Supervised Machine Learning: Classification / 1. Logistic Regression (Start Here) / Confusion Matrix, Accuracy, Specificity, Precision... /

Confusion Matrix, Accuracy, Specificity, Precision, and Recall

C3_V04-Classification-Error-Metrics-Part-1

Specificity: Avoiding False Alarms

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative
Actual Negative	False Positive (FP)	ve

$$\text{Sensitivity} = \frac{TN}{FP + TN}$$

hat bottom row of actual negative classes.

5.2. Monitor the model in production
SUBTASK(S):
5.2.1. Understand what MLOps is

REFERENCES:
https://www.ibm.com/blogs/journey-to-ai/2021/04/paving-the-paths-to-ai-engineeringand-modelops/
https://ibm-cloud-architecture.github.io/refarch-data-ai-analytics/methodology/MLops/
https://learning.oreilly.com/library/view/introducing-mlops/9781492083283/ch01.html

5.2.1.1. Understand types of data drift and their impact

5.2.2. Monitor model performance metrics using logging

https://learn.ibm.com/mod/video/view.php?id=169287
https://learn.ibm.com/mod/page/view.php?id=169579&forceview=1
https://learn.ibm.com/mod/page/view.php?id=169581&forceview=1
https://learn.ibm.com/mod/page/view.php?id=169598&forceview=1

5.2.3. Monitor model business KPIs

REFERENCES:
https://learn.ibm.com/mod/page/view.php?id=170330&forceview=1
https://learn.ibm.com/mod/video/view.php?id=169575

5.2.4. Decide when to retrain model

REFERENCES:
https://learning.oreilly.com/library/view/introducing-mlops/9781492083283/ch07.html#online_evaluation – Champion/Challenger section
https://learning.oreilly.com/library/view/ml-ops-operationalizing/9781492074663/ch01.html#retraining_and_remodeling - Retraining and remodeling section
https://learn.ibm.com/mod/video/view.php?id=169604

5.2.5. Use IBM OpenPages to govern models

REFERENCES:
https://www.ibm.com/docs/en/cloud-paks/cp-data/4.0?topic=governance-set-up-modelopenpages-mrg

5.3. Determine if there is unfair bias in the model
SUBTASK(S):

5.3.1. Understand how model bias can creep in

REFERENCES:
https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-bestpractices-and-policies-to-reduce-consumer-harms/
https://developer.ibm.com/articles/machine-learning-and-bias/

5.3.2. Understand the role of transparency in mitigating bias

REFERENCES:
https://www.forbes.com/sites/cognitiveworld/2020/05/23/towards-a-more-transparentai/?sh=b928073d9371

5.3.3. Create an AI FactSheet
REFERENCES:
https://www.ibm.com/blogs/research/2020/07/aifactsheets/

5.3.4. Detect bias in models using IBM AI Fairness 360 Toolkit and Watson OpenScale

REFERENCES:
https://developer.ibm.com/blogs/ai-fairness-360-raise-ai-right/
https://github.com/IBM/bias-mitigation-of-machine-learning-models-usingaif360/blob/main/README.md
https://learn.ibm.com/mod/video/view.php?id=168628
https://www.ibm.com/docs/en/cloud-paks/cp-data/4.0?topic=governance-managemodel-risk

Next Steps

1. Take the assessment test for IBM Machine Learning Data Scientist v1.
2. If you pass the assessment exam, visit pearsonvue.com/ibm to schedule your testing sessions.
3. If you failed the assessment exam, review how you did by section. Focus attention on the sections where you need improvement. Keep in mind that you can take the assessment exam as many times as you would like ($30 per exam); however, you will still receive the same questions only in a different order.