1. For a given input list: 1, 2, 3, 4
Cube (element to power 3) each element
Return the results as a list

list(map(lambda x: x**3, lst))

2. For a given input list: 1,2,3,4,5,6,7
Inspect each number in the input list and determine if it is even
Next square the even values
Finally return the squared evens in a list

list(map(lambda x: x**2, list(filter(lambda x: x%2==0, lst))))

3. Which of the following is/are NOT native or built-in data types in Python?
A: boolean
B: integer
C: float
D: heap
E: string
F: varchar

DF

4. For given input lists: a,b,c and 1,2,3
Create a dictionary from two input lists

{x:y for x,y in zip(lst1, lst2)}

5. Mutable data types/collections in Python can be changed in place. Immutable ones can not change in place. Which of the following are mutable?
A: bool
B: int
C: float
D: set
E: list
F: string

DE

G: tuple
H: complex

6. Which of the following is NOT true about Python?
A: Python code can run in IPython and Jupyter notebooks
B: Python allows for the inclusion of comments and pseudocode to better organize code
C: Users can save .py files with an editor then subsequently execute them from the command line
D: Base Python automatically parallelizes processing across cores when multiple cores are available
E: Python allows users to save multiple functions in a .py file then import those functions in a different file

D

7. For a given input list: abbcccddddeeeeefffffggggggghhhhhhhh
Return a dictionary of character counts
Count the number of times each character appears in the string
Characters with a count of 0 should not be included in the output dictionary

dict(Counter(string))

8. For the vector v = [2.0, -3.5, 5.1]:
Find the L1 norm of v
Return the result as a float

sum(list(map(lambda x: abs(x), v)))

9. NumPy array practice
Create a vector that starts at 1 and increases until 150
Turn the vector into a matrix with 10 rows and 15 columns
Return the sums for the 10 rows as a list. HINT: there should be ten values for the printed sum

np.arange(vectorLower, vectorUpper).reshape((10,15)).sum(axis=1).toli

Use the following input vector values: vectorLower = 1; vectorUpper = 151

10. Which of the following pairs of events are mutually exclusive. There can be more than one answer.
A: Odd numbers and the number 3
B: Even numbers and numbers greater than 10
C: Negative numbers and positive numbers less than 25
D: Numbers between 100-200 and numbers between 201-300
E: None of the above

CD

11. Geometric distribution
The geometric distribution is a useful tool for modeling time to event data. A successful street vendor says that on average 1 out of every 10 people who walk by on the street stop to buy a taco.
Represent these data with a geometric distribution
What is the probability that the vendor has to wait until 20 people walk buy before someone buys a taco?

stats.geom.pmf(k=k, p=p)

12. Poisson distribution
The Poisson distribution is a useful tool for modeling count data given discrete intervals. Based on historical data the expected number of accidents at a busy intersection is 4 per month.
Represent these data with a Poisson distribution
What is the probability of more than 7 accidents at that intersection next month?

1 - stats.poisson.cdf(k=k, mu=mu)

13. Gaussian distribution
The Gaussian or Normal distribution is use heavily

1--
stats.norm.cdf(x=cdf_val,loc=loc

throughout statistics and data science. Lets assume scores for this assessment have a mean of 50% and a standard deviation of 20%.
Represent these data with a Normal distribution
What is the probability of observing a score >= 80?
Use 50.0, 20.0, and 80 for your input values

14. Perform matrix multiplication on a square matrix
HINT: A 2X2 matrix times a 2x2 matrix should yield a 2x2 matrix

```
for i in range(len(matrix1)):
 for j in range(len(matrix2[0])):
 for k in range(len(matrix2)):
# resulted matrix
 res[i][j] += matrix1[i][k] * matrix2[k][j]
```

15. Which types of programming tasks best describes what you are expected to already have some familiarity with before beginning this course?
A: dashboarding, high performance computing, and code profiling
B: numeric computing, data munging, data visualization and data modeling
C: convex optimization, python programming, statistical programming
D: continuous integration, linear programming, and data exploration

B

16. Though the emphasis may change, which two elements are both essential and common to all three process models we talked about?

C

A: prediction, recommendation
B: data mining, data cleaning
C: resolve the business question, feedback loops
D: testing, model deployment

17. Is the following statement True/False? To succeed in this course you are expected to be proficient in any one of the following: R, Python or Java.
A: TRUE
B: FALSE

B

18. Which of the following is the least accurate statement about the advantages of using process models in data science? Process models generally help by...
A: avoiding unnecessary tangents
B: speeding up the process of getting through the workflow
C: minimizing the model selection process
D: guiding effective time allocation

B

19. Is the following statement True/False?
Design thinking is applied in other domains which helps make the task of communicating the AI workflow to those outside of data science easier.
A: TRUE
B: FALSE

A

20. It is day one on the job and you need to come up with a plan—how do you begin?
A: Gather what data you can quickly and perform some EDA to understand the problem better
B: Plan to interview or study reviews of both satisfied and dissatisfied subscribers as soon as possible

C

C: Get the perspective from management and follow the leads they might provide
D: Something else entirely

21. In order to come up with the back-of-the-envelope ROI calculation for this project, how might you approach it?
A: Number of active users X Yearly payment
B: Number of active users X monthly payment X % increase of users (assumption)
C: Yearly costs X (number of users at month 2 - number of currently active users)
D: Number of active users X Yearly payment - estimate for cost of project time

AD

22. Thinking with the lens of the scientific process, what would your next steps be if you wanted to decide where to open the next store for your sled business?
A: Start pulling sales and other data to create a business viability assessment for Vermont
B: Gather more data and repeat the snowfall experiment
C: Gather different data say snowfall by county and repeat the experiment
D: Start a business viability assessment for all three states

ABCD

23. When embarking on a data science project, why do you ultimately want to format your data so that it can be housed in something like a Pandas DataFrame or NumPy Array?
A: DataFrames/Arrays most closely resemble tables in

D

relational databases.

B: DataFrames/Arrays are the only structures in Python capable of holding significant amounts of data.

C: Nearly all modeling algorithms take input data in a tabular format analogous to format of DataFrames/Arrays.

D: All of the above

24. Lets imagine there is a start-up that has a speech-to-text service that incorporates gestures and body language into its output. Which of the following products represents the most defensible business opportunity.

A: Offer a service that hooks into streaming video and predicts the emotional state of people in the videos

B: Create an app that allows job interviewers to get additional information about candidates

C: Create a new and improved conferencing app

D: Create a service that improves on existing audio recognition systems as a richer interface to mobile devices

D

25. Lets imagine there is a start-up that has a speech-to-text service that incorporates gestures and body language into its output. They offer annotated meeting reports as a product and customers are generally very satisfied, but sales to new customers tend to be very slow to acquire. Which of the following business opportunities should be the highest priority?

A: Develop and delivery new products to existing customers

C

B: Develop new products and target new customers
C: Use customer segmentation and/or market analysis to help marketing with new customers
D: Use customer segmentation and/or market analysis to move into a different market

26. Your company is convinced it is time to change the nature of your companies core product and management has come to ask your advise. Which question DOES NOT exemplify scientific thinking in this situation?
A: Do we have any data like a corpus of customer feedback to support this decision?
B: Can we run an experiment like A/B testing to see if it helps support this decision?
C: Which members of leadership support this decision?
D: Have any other companies been successful making a comparable change?

C

27. Is the following statement true or false?
CSV files are one of the most commonly used file formats for data science because file input/output is easy they are plain-text, and they work well with commonly used spreadsheet tools.
A: TRUE
B: FALSE

A

28. Which of the following DOES NOT represent a valid relational database to connector relationship?
A: MySQL --> MySQL-python
B: PostgreSQL --> psycopg/psycopg2

D

C: SQLite --> sqlite3
D: Berkeley DB --> bsddb

29. Which tasks should be included in a data ingestion pipeline? (Choose one or more)
A: Account for missing data, faulty data, repeated observations and other data integrity issues
B: Ensure that expected data is returned given a specific set of parameters
C: Ensure that an expected format is returned
D: Ensure that models produce expected results

ABC

30. Sparse matrices can be useful as a target destination for ETL, but what are the main caveats (choose one or more)?
A: You cannot convert directly from a numpy.array to any of the scipy.sparse matrices
B: NumPy linear algebra functions generally cannot be called directly
C: Saving to disk is not possible directly from a scipy.sparse format
D: The train test splits need to be performed by hand with scipy.sparse matrices
E: It is difficult to print to screen scipy.sparse matrices directly

BE

31. Which types of data generally work well with sparse matrices?
A: word counts, time-series data
B: audio files, images
C: word counts, user-item matrix for recommenda-

C

	tions D: text data, audio files	
32.	Is the following statement True or False? Sparse matrices from SciPy need to be transformed into a dense matrix before using scikit-learns train-test-split function? A: TRUE B: FALSE	B
33.	Which fundamental part of the data ingestion process is concerned with to the phrase "bad data in equals bad data out"? A: Gather all relevant data from the sources of provided data B: Implement several checks for quality assurance C: Take the initial steps towards automation of the ingestion pipeline	B
34.	Which of the following is most concerned with ensuring deployed models scale well with added users? A: data scientist B: data analysts C: data engineer D: product manager	C
35.	Which of the following is statements is the least correct in the context of the EDA process A: EDA is used to provide summary level insight into a dataset B: EDA consists of both exploratory and confirmatory data analysis C: EDA can be used to discover missing data, outliers	B

and class inbalance issues
D: The EDA process can be used to help predict time to completion for a project
E: The EDA process is an ideal time to explore the connection between the data and the business opportunity

36. Which of the following is an example of a data manipulation that is NOT considered reproducible research? — E

A: Saving classes and functions in a Python file to be called by Jupyter
B: Code blocks in Jupyter notebooks
C: The use of proprietary tools to carry out research
D: Graphics, plots and other visualizations
E: Copy and paste actions in a spreadsheet

37. True/False. The seaborn pairplot and other seaborn plotting functions exist as distinct tools from the plots available through matplotlib. — B
A: True
B: False

38. In the continuing AAVAIL streaming case study example, one of the data features that can be useful in answering questions about customer churn is the total number of streams that a customer has watched. Imagine that you are working with a dataset where 10% of customers are missing this feature. A good place to start would be to go back and see if it's possible to gather this information from the user logs, but assuming that this initiative is unsuccessful, you will have to decide what to do about this missing data. — B

Which course of action is LEAST likely to be helpful in modeling churn?
A: Replace the missing stream count with the mean stream count among users where this information is available.
B: Replace the missing stream count with a -1 to flag that it is unknown for a given user.
C: Use the other features in the dataset in a model to predict the missing stream counts.

39. What is the main reason for using multiple imputation?
A: Multiple imputation is necessary when more than one feature in the training data has missing values.
B: Multiple imputation is a way to increase the size of your training dataset.
C: Multiple imputation helps to better characterize the error introduced by replacing missing/unknown data with some chosen values.

C

40. Which of the following is NOT normally a part of the EDA process
A: Visual summaries of the data
B: Connecting the data to the business opportunity
C: Communication to stakeholders
D: Predictive linear or logistic regression

D

41. True/False. The EDA process is decoupled from modeling and cannot be used to help esitmate the time it will take to complete a modeling procedure
A: True
B: False

B

42. The three types of missingss discussed during this module were:
A: MRAR, MAR, MCAR
B: MNAR, MRAR, MCAR
C: MNAR, MAR, MARC
D: MAR, MRAR, MCAR
E: MCAR, MNAR, MAR

E

43. Which statement is the least true about using Jupyter notebooks in the context of EDA
A: They naturally lend themselves to version control systems
B: They can be ported from one environment to another easily
C: They are helpful because a mixture of code and markdown enables storytelling
D: They are integrated with the plotting library matplotlib
E: They are integrated with the data manipulation library pandas

A

44. Which of the following is NOT an example of assumption that you work with when making probability statements about
a sample of data?
A: That there is an underlying population that your sample comes from
B: That the population follows an assumed probability distribution
C: That the observations in your sample are independent and identically distributed
D: That random variables represent the possible val-

E

ues that the data can take
E: That the probability statement applies to one observation at a time in a data set

45. There are many ways to carry out statistical inference. Which one method of the following is NOT used to compute estimates in the context of statistical inference.
A: Null Hypothesis Significance Testing (NHST)
B: Maximum Likelihood Estimation (MLE)
C: Markov Chain Monte Carlo (MCMC)
D: Expectation Maximization (EM)
E: Simulation via Permutations

A

46. Company Z sent out a user satisfaction survey to its customers that included some demographic questions. They want to determine if there is a difference in the age among users of Product 1 versus users of Product 2 (at least among the survey respondents). Which of the following is an appropriate null hypothesis for this study?
A: Users of Product 1 are on average older than users of Product 2.
B: Users of Product 1 are on average younger than users of Product 2.
C: Users of Product 1 and Product 2 are on average the same age.

C

47. Which of the following is the least valid statement when it comes to dashboards?
A: Dashboards are an easy way to share summaries and findings

D

B: Dashboards have interactive functionality that helps create a rich experience for the user
C: Dashboards are generally used after serveral iterations of the AI workflow
D: Dashboards are quick way to create portable simple plots
E: Dashboards can be used to tell the story of investigative visualizations

48. A data scientist at Company Z sorted the survey responses by whether the respondents used Product 1 or Product 2 and then compiled their ages:
Of the hypothesis test discussed in these contents what one is the most appropriate for testing the following hypothesis?
There is no age difference, on average, between the users of product 1 and the users of product 2
A: A 1-sample t-test
B: A 2-sample t-test assuming equal variance
C: Z-Test with continuity correction
D: A 2-sample unequal variances t-test
E: Binomial Test

D

49. Suppose that on average 2.5% of visitors to your website sign up for your newsletter. In a recent week, 2701 visitors out of a total of 108879 signed up.
Using a binomial distribution. What is the probability that number of visitors who signed up is 2701 or fewer?
A: 0.125
B: 0.346
C: 0.414

B

D: 0.007
E: 0.015

50. True/False. If there customer churn were quantified using a Poisson distribution, then a bootstrap could be used to quantify the uncertainty associated with the estimate.
A: True
B: False

A

51. Which of the following is NOT and example of a valid strategy to deal with the multiple comparisons problem?
A: Benjamini/Hochberg correction based on False discovery Rates
B: Create a null distribution using permutations to help provide context
C: Perform all comparisons then only keep the single test that performs the best
D: If appropriate use an alternative modeling framework like generalized linear models
E: Bonferroni Correction

C

52. Which scikit-learn API interface would be used to carry out feature engineering with a domain expert?
A: Transformer
B: Estimator
C: Fit
D: Predict
E: Pipeline

A

53. Which variant of SMOTE is most appropriate when you have a mixture of categorical and continuous vari-

D

ables?
A: KMeansSMOTE
B: BorderlineSMOTE
C: SVMSMOTE
D: SMOTENC
E: SMOTE

54. Which of the following is not an example of a technique used for dimensionality reduction technique?
A: Latent Dirichlet allocation
B: Non-negative matrix factorization
C: Singular value decomposition
D: Eigenvalue decomposition
E: K-nearest neighbors
F: Principal Components Analysis

E

55. When printing the most representative words from each topic what best describes the insight we gain?
A: The top words in each topic correspond to the most frequently used words in the corpus
B: The topics are defined by their representative words and the document is a mixture of these topics
C: Documents have topics and the words describe the corpus
D: The words make up the document and the topics describe the words
E: Topics are latent features and the top words describe the average document

B

56. The .fit_transform method corresponds to which scikit-learn interface(s)? Choose one answer.
A: Transformer, Estimator, Predictor

B

B: Transformer, Estimator
C: Estimator, Predictor
D: Transformer
E: Transformer, Predictor

57. True/False. A principal reason for emphasizing the use of pipelines in the AI workflow is to have a consistent platform that enables comparison of many variants of the workflow.
A: True
B: False

A

58. Which of the following statements describes the best strategy to address class imbalance?
A: If there is a lot of data just use under-sampling otherwise use outlier detection algorithms.

B: Determine the best variant of SMOTE by comparisons and use it.
C: Continue to collect data until you have balanced classes.
D: Use an outlier detection algorithm or SVM instead of a re-sampling technique.
E: Compare re-sampling approaches to a baseline and to detection algorithms.

E

59. Which of the following statements is not a feature of the package imbalanced-learn imbalance?
A: Has a suite of over and under-sampling methods implemented
B: Works with TensorFlow
C: Has a number of tutorials to work from
D: Has outlier detection algorithm packaged as part of

D

library
E: Works with scikit-learn pipelines

60. Which of the following statements does not describe valid use case for dimensionality reduction? E
A: Principal components analysis to process images used in classification.
B: Non-negative matrix factorization to resolve topics from a corpus of words.
C: t-distributed stochastic neighbor embedding to visualize the results of a clustering algorithm.
D: Using an ANOVA to select a subset of features
E: Down-sampling of the majority class

61. True/False. tSNE is a reasonable alternative to PCA because it describes a wider variety of structures. However, it is still recommended to use another dimensionality reduction method, like PCA if the number of features is very high. A
A: True
B: False

62. Which statement best describes why visualization of topics can have an impact on the business opportunity? A
A: Because sharing with domain experts might enable topic-specific feature engineering
B: Because visual inspection can help choose the number of topics
C: Because we are able to see the top words with each topic
D: Because we are able to see the relative importance

of each topic across the corpus
E: Because domain experts can visually inspect the validity of the topics

63. For credit applications, which of the following types of patterns would be the most undesirable for a supervised learning algorithm to use?
A: Given salary > 50K and dept < 5K they will repay the loan
B: Given salary > 80K they will repay the loan
C: Given salary > 50K and Age > 25 and they will repay the loan
D: Given salary > 50K and that they reside in a metropolitan (population >100K)
E: Given salary > 50K and the degree name

C

64. Which of the following is not an example of an outlier detection algorithms?
A: Adaptive Synthetic (ADASYN)
B: Elliptic Envelope
C: One Class SVM
D: Isolation Forest
E: Local Outlier Factor

A

65. Which of the following is NOT considered a reasonable metric to choose the number of clusters?
A: Calinski-Harbasz index
B: Inertia or the within cluster sum-of-squares
C: Silhouette Score
D: Davies-Bouldin index
E: Adjusted Rand index

B

66.

E

You are asked to build a recommendation engine for new products at an online retailer. Which of the following features is the least likely to be a protected attribute? Recall that a protected attribute is one that may contain privileged and unprivileged classes.
A: age
B: gender identity
C: race
D: religion
E: purchase history

67. True/False. The API for the reweighting algorithm has a number of custom methods that will require you to consult the documentation to ensure appropriate use.
A: True
B: False

B

68. Which outlier detection method is known to work well on high-dimensional data?
A: Random Forests
B: Elliptic Envelope
C: One Class SVM
D: Isolation Forest
E: Local Outlier Factor

E

69. True/False. In an imbalanced dataset if the minority class represents less than 5% of all of the samples then outlier detection algorithms must be used in place of other supervised learning algorithms.
A: True
B: False

B

70. Which clustering method can be readily applied to graphs?
A: Gaussian mixture models
B: Spectral clustering
C: Affinity Propagation
D: k-means
E: Dirichlet process Gaussian mixture models

B

71. Which of the following clustering methods does not need to set the number of clusters?
A: Gaussian mixture models
B: Spectral clustering
C: MiniBatch k-means
D: k-means
E: Dirichlet process Gaussian mixture models

E

72. True/False. In the clustering case study, the suggested re-sampling methods drove a major improvement in model performance.
A: True
B: False

B

73. True/False. In the context of customer profiling and the AAVAIL data set it makes sense to first perform a dimension reduction technique like PCA before running the model through a clustering estimator.
A: True
B: False

B

74. In which situation would you most strongly consider MAE over RMSE as a regression metric?
A: where we would like to interpret the error metric in terms of the original units

C

B: like predicting daily temperature where we expect a small range of values
C: like predicting time to failure for a machine where we expect a long tailed distribution of values
D: like predicting the category or topic associated with a document
E: where we would like to interpret the error metric as a squared version of the original units

75. If you have data with a large number of features and you are sure that it will take some time to train and tune the model, which approach is LEAST likely to result in a speed improvement during grid-searching?
A: In your pipeline use variance thresholding to limit the number of features
B: Use the Shuffle and split form of cross-validation
C: Use a randomized grid search form of cross validation
D: Randomly subset the data
E: Use PCA to reduce the dimensionality of the data before training

B

76. Which of the following is not an example of a variant/application of gradient decent that we have covered?
A: batch gradient descent
B: mini-batch gradient descent
C: regularized gradient decent
D: stochastic gradient descent
E: gradient descent applied to regression

C

77.

B

What important feature of the Watson NLU API do we have to ensure so that we can repeat an experiment in the future in a reproducible way?

A: GitHub integration

B: Passing a version argument with each request

C: Each future version of the API is guaranteed to keep the same arguments

D: The NLU can be run locally

E: The exchanged JSON is guaranteed to keep the same format

---

78. Processing the corpus with the provided lemmatize_document reduces the total number of tokens to what percentage of the original?

A: 10-15%

B: 20-35%

C: 45-50%

D: 70-75%

E: 85-95%

C

---

79. If we have a situation where false positive is not as potentially costly as a false negative say flagging comments for manual review based on suspected unlawful activity, which of the following is the best approach to consider?

A: Only look at the f-score of the negative class for evaluation

B: Use recall as the evaluation metric

C: Use precision as the evaluation metric

D: Set beta to 0.5 in the fscore

E: Set beta to 2.0 in the fscore

E

80. True/False. All classifiers in scikit-learn do multi-class classification out-of-the-box. These classifiers can differ in their approach though (e.g one-vs-all or one-vs-one).
A: True
B: False

A

81. Which of the following is not an example of a generalized linear model (GLM)?

A: ANOVA
B: Multinomial regression
C: Poisson regression
D: KNN regression
E: Logistic regression

D

82. True/False. Models in the generalized linear mixture model (GLMM) family like multilevel models are generally optimized using sophisticated techniques like MCMC sampling.
A: True
B: False

A

83. When you use Watson Services like Watson Natural Language Understanding via the Python SDK, what are the three items that need to be saved? These items are generally saved on a local machine and included in scripts and notebooks as imported variables.
A: service version, service API key, service JSON map
B: service URL, service JSON map, service API key
C: service API key, service version, service URL
D: service version, service IAMAuthenticator, service URL

C

E: service API key, service URL, service IAMAuthenticator

84. Which of the following does not describe a feature of the Watson Natural Language Understanding service?
A: Perform document classification tasks using a custom model built from text
B: Identify high-level concepts that aren't necessarily directly referenced in the text
C: Find people, places, events, and other types of entities mentioned in your content
D: Recognize when two entities are related, and identify the type of relation
E: Analyze the sentiment toward specific target phrases and the sentiment of the document as a whole

A

85. Which of the following is not an example of a relevant question when tuning a NLP classification pipeline?
A: Should I use bag-of-words or a vector embedding representation?
B: Which stop words do I include?
C: Which n-grams do I include?
D: Should I use a TF or a tf-idf transformation?
E: Should I use RMSE or MAE as an evaluation metric?

E

86. Which of the following model is not an example of an ensemble approach to learning?
A: Decision tree
B: Random forest
C: Boosting

A

D: Model stacking
E: Gradient boosting

87. Which of the following neural network architectures are most-commonly used for time-series analysis?
A: Multi-layer perceptron
B: Recurrent neural networks
C: Transfer learning
D: Convolutional neural network
E: Autoencoders

B

88. True/False. All images must be downloaded and saved locally before you can call classify from the connected IBM Watson Visual Recognition service.
A: True
B: False

B

89. A simple CNN that runs on all ten classes and uses all of the data obtains approximately what level of accuracy?

A: 62-77%
B: 78-83%
C: 84-89%
D: 90-94%
E: 95-99%

D

90. True/False. Bagging and boosting ensemble methods both use only decision trees as base classifiers. The difference is in the bias and variance of the individual trees.
A: True
B: False

B

91. True/False. A decision tree classifier is useful as a model for the AAVAIL subscriber churn data.
A: True
B: False

A

92. Which of the following was not discussed as a tunable parameter of a neural network?

A: Hardware availability
B: Activation functions: sigmoid, tanh, softmax, ReLU, leaky ReLU
C: Regularization techniques: weight decay, early stopping, dropout
D: Training method: Loss function, learning rate, batch size, number of epochs
E: Structure: the number of hidden layers, the number of nodes in each layer

A

93. True/False. Transfer learning is a recent advancement to come out of the field of reinforcement learning.
A: True
B: False

B

94. When training a custom classifier in Watson Visual Recognition the negative images should be:
A: As visually similar as possible to the positive images
B: Background images without the positive images
C: As random as possible to establish a background
D: Randomly generated from the positive images
E: Visually distinct from the positive images

A

95.	True/False. The Watson Visual Recognition service can only be accessed using an API key via Python or curl. A: True B: False	B
96.	Which of the following use cases is the least appropriate use case for a convolutional neural network? A: Image classification B: Image retrieval C: Image composition D: Object detection E: Image segmentation	C
97.	True/False. A typical convolutional neural network is constructed using a combination of convolutional, pooling and dense layers. A: True B: False	A
98.	True/False. If we continue to add GPUs or other computational resources, the time it takes to train a model will always continue to decrease. A: True B: False	B
99.	True/False. It is reasonable to think of the commands in a Dockerfile as a step-by-step recipe on how to build up a Docker image. A: True B: False	A
100.	What are the two steps that must be carried out if you want to iterate locally on a model then deploy it to the	D

	Watson Machine Learning (WML) service? A: Provision a WML service & Dockerize your model B: Create a Python virtual environment & Dockerize your model C: Provision cloud storage & Provision a WML service D: Provision a WML service & create a Python virtual environment E: Dockerize your model & Provision cloud storage	
101.	Which of the following lists contains one or more references to a technology that is not a specific python package used to speed up and improve the performance of python code? A: py-cuda, Cython B: multiprocessing, mpi4py C: subprocessing, symmetric-multiprocessing D: threading, ipyparallel	C
102.	A Spark cluster is generally managed using a Docker container and a YAML file. A: True B: False	B
103.	Which of the following is least likely to be a use case for Docker containers? A: Microservices: many loosely coupled and independently deployable services B: DevOps/Data Engineers use containers as a common platform for many teams C: Avoid install overhead with a Spark environment D: Hybrid, multi-cloud portability for a machine learn-	E

ing model
E: Replacement for a standard virtual machine

104. Docker images are the basis of containers. It is possible to pull an image from the registry and ask the Docker client to run a container based on that image. Some images are official while many others are user defined.
A: True
B: False

A

105. What is the principal reason to create a virtual environment before creating your model locally?
A: Because all models should have their own virtual environment
B: Because it will ensure that the most recent packages are used
C: Because virtual environments can be containerized easily
D: Because the Python client for Watson Machine Learning has specific dependencies
E: Because it ensures that locally created/trained models are compatible with Watson Machine Learning

E

106. True/False. You may only pass a pickle file to save your model in the your Watson Machine Learning library.
A: True
B: False

B

107. True/False. The Spark ML API uses DataFrames from Spark SQL. They can hold a variety of data types with different columns for storing text including: feature vectors, truth labels, and predictions.

A

A: True
B: False

108. Which type of recommender system is readily available in Spark Machine learning?
A: Utility-based recommender systems
B: Collaborative-based recommender systems
C: Content-based recommender systems
D: Hybrid recommender systems
E: Demographics-based recommender systems

B

109. Docker containers run a private file system that is isolated from the host and other containers. What is the suggested way to access notebooks and scripts from within the container?
A: tmpfs mount
B: use a named pipe
C: bind mounts
D: GitHub
E: volumes

E

110. Which of the following is not a valid way to pass parameter to a Spark MLlib model?
A: Pass a pickle file when the model is declared
B: Set the named parameters directly when the model is declared
C: Pass a ParamMap to the .fit() method of the model
D: Create a ParamMap, update it then pass it to a .fit() method of the model
E: Pass a ParamMap to the .fit() method of a pipeline

A

111. True/False. Spark MLlib has several estimators and transformers, but it lacks basic tooling for natural

B

language processing (NLP) like the ability to make term frequency-inverse document frequency (TF-IDF) transformations.
A: True
B: False

112. Which phrase most accurately reflects what is meant by the cold start problem in recommendation systems?
A: A. When a new user is introduced and the recommendations are made on similarities
B: B. When a new item is introduced and the recommendations are made on similarities
C: C. When a new user is introduced and the recommendations are made on item popularity
D: D. When a new user or new item is introduced and the recommendations depend on popularity
E: E. When a new user or new item is introduced and the recommendations depend on similarities

E

113. True/False. Matrix factorization is commonly associated with collaborative filtering recommender systems.
A: True
B: False

A

114. When you worked on model deployment case study, which modification to the ALS algorithm had the largest effect on model performance?

A: The explicit training vs implicit training comparison
B: The lambda or regularization parameter
C: The epsilon or scale parameter
D: The l1 vs l2 comparison

B

115. True/False. The Spark collaborative filtering implementation is built by default for explict feedback, but it can be used with implicit feedback just as easily.
A: True
B: False

A

116. Which of the following was not mentioned as a reason to bundle unit tests with a deployed model?
A: promotes code quality
B: regression tests
C: they perform the workflow feedback loops
D: automate performance monitoring tasks
E: can be readily integrated into CI/CD pipelines

C

117. True/False. When logging for the predict endpoint, runtime is considered an optional feature to be monitored.
A: True
B: False

B

118. To pass a query to an API endpoint that has been set up to run inside of a Docker container, which of the following best describes the formats that have been used in the presented deployment workflows?
A: numpy ndarray, pandas DataFrame
B: JSON
C: JSON, flatfiles
D: YAML
E: All of the above

B

119. In the context of the AI workflow presented in these materials which of the following is not an example of

D

a valid feedback loop?

A: Trying different data transformations on a given model
B: Returning to the data collection stage from transformations to reduce the number of transforms
C: Performing EDA on the data after a model has been deployed and data have been logged
D: Moving from the business opportunity and data collection to model iteration
E: Returning to discuss the business opportunity after a model has been deployed

120. True/False. It is critical in machine learning model deployment workflows to ensure close to 100% test coverage before production.
A: True
B: False

B

121. Common sources of performance drift include:

A: Changes in customer demographics or user behavior after a model is trained
B: Incomplete, incorrect or unbalanced training data
C: Changes in versions of libraries or modules used in models
D: All of the above

D

122. Which list contains one or more elements that were presented as non-essential when creating a logging system to monitor the performance of a machine learning model that has been deployed.
A: request_type, input data

A

B: model_version_number, timestamp
C: predictions/recommendations, timestamp
D: input_data_summary, runtime

123. Model accuracy metrics such as precision, recall and F1 scores provide objective evidence of fairness.
A: True
B: False

B

124. What is the purpose of Minikube in the context of Kubernetes?
A: It deploys a Kubernetes cluster to a remote service based on a specified YAML file
B: It launches a Kubernetes cluster locally within a virtual machine or directly on the host
C: It is used through a command line interface to interact with a Kubernetes cluster
D: It allows you to work with multiple Docker containers as if they were a single application
E: It is the management layer for a Kubernetes cluster

B

125. Which of the following are three services provided by IBM Watson OpenScale?
A: Automatic CSS-HTML Generation, Hyerparameter Tuning, Performance Metrics
B: Compliance, Cross Regional Availabilty, Automatic Backup
C: Drift Monitoring, Fairness, Explainability
D: General Artificial Intelligence, Time Travel, World Domination

C

126. True/False. IBM Watson OpenScale only works with models that have been built on, and deployed by, IBM

B

Watson Stuido.
A: True
B: False

127. What is Port Forwarding in Docker?

A: The process of creating a data file on the local computer that persists after the container is closed
B: Sending one container to multiple nodes on a multicloud environment
C: Allowing data to be passed in and out of a docker container and controlling which applications can do this
D: Building one Docker image using another docker image as a template

C

128. True/False. The date and time of a transaction should never be logged because this will only reflect past transactions and is not relevant to the performance of predictions on an on-going basis.
A: True
B: False

B

129. What is the purpose of kubectl in kubernetes?
A: Automatic logging of requests and responses
B: A tool that makes it easy to run a single-node cluster locally
C: The primary node agent on each node, responsible for the processes running on that machine
D: The CLI for communicating with the kubernetes cluster

D

130.

B

True/False. A Kubernetes pod can contain multiple kubernetes deployments
A: True
B: False

131. Your Flask application needs to error check the input when making a prediction. Which features need to be provided (at a minimum) as input given the context of this business opportunity?
A: The previous three months of data and the target month from the year before
B: The target date and country
C: At least one year of revenue data
D: The recent number of transactions and the number of views
E: Metadata associated with the streams

B

132. True/False. Unit testing and novelty detection provide an automated mechanism to monitor model performance.
A: True
B: False.

A

133. To move through the AI enterprise workflow quickly, we used the concept of workflow templates. Which of the following aspects of the template contributes the least to the re-usability of these templates?
A: When they are used with a version control system
B: Ensuring that the model exists as a separate Python module to be called by Flask
C: Have an HTML endpoint for the Flask app to create dashboards

C

D: Use a template for a unit test suite
E: Ensure that data ingestion and data visualization code exists as a modules or scripts