Number: DSA-C02
Passing Score: 800
Time Limit: 120
File Version: 3.0

Exam Code: DSA-C02

Exam Name: SnowPro Advanced: Data Scientist

Exam A

QUESTION 1
Which of the following method is used for multiclass classification?

A. one vs rest
B. loocv
C. all vs one
D. one vs another

Correct Answer: A
Section:
Explanation:
Binary vs. Multi-Class Classification
Classification problems are common in machine learning. In most cases, developers prefer using a supervised machine-learning approach to predict class tables for a given dataset. Unlike regression, classification involves designing the classifier model and training it to input and categorize the test dataset. For that, you can divide the dataset into either binary or multi-class modules.
As the name suggests, binary classification involves solving a problem with only two class labels. This makes it easy to filter the data, apply classification algorithms, and train the model to predict outcomes. On the other hand, multi-class classification is applicable when there are more than two class labels in the input train data. The technique enables developers to categorize the test data into multiple binary class labels.
That said, while binary classification requires only one classifier model, the one used in the multi-class approach depends on the classification technique. Below are the two models of the multi-class classification algorithm.
One-Vs-Rest Classification Model for Multi-Class Classification
Also known as one-vs-all, the one-vs-rest model is a defined heuristic method that leverages a binary classification algorithm for multi-class classifications. The technique involves splitting a multi-class dataset into multiple sets of binary problems. Following this, a binary classifier is trained to handle each binary classification model with the most confident one making predictions.
For instance, with a multi-class classification problem with red, green, and blue datasets, binary classification can be categorized as follows:
Problem one: red vs. green/blue
Problem two: blue vs. green/red
Problem three: green vs. blue/red
The only challenge of using this model is that you should create a model for every class. The three classes require three models from the above datasets, which can be challenging for large sets of data with million rows, slow models, such as neural networks and datasets with a significant number of classes.
The one-vs-rest approach requires individual models to prognosticate the probability-like score. The class index with the largest score is then used to predict a class. As such, it is commonly used for classification algorithms that can naturally predict scores or numerical class membership such as perceptron and logistic regression.

QUESTION 2
Which ones are the key actions in the data collection phase of Machine learning included?

A. Label
B. Ingest and Aggregate
C. Probability
D. Measure

Correct Answer: A, B
Section:
Explanation:
The key actions in the data collection phase include:

Label: Labeled data is the raw data that was processed by adding one or more meaningful tags so that a model can learn from it. It will take some work to label it if such information is missing (manually or automatically).
Ingest and Aggregate: Incorporating and combining data from many data sources is part of data collection in AI.
Data collection
Collecting data for training the ML model is the basic step in the machine learning pipeline. The predictions made by ML systems can only be as good as the data on which they have been trained. Following are some of the problems that can arise in data collection:
Inaccurate data. The collected data could be unrelated to the problem statement.
Missing data. Sub-data could be missing. That could take the form of empty values in columns or missing images for some class of prediction.
Data imbalance. Some classes or categories in the data may have a disproportionately high or low number of corresponding samples. As a result, they risk being under-represented in the model.
Data bias. Depending on how the data, subjects and labels themselves are chosen, the model could propagate inherent biases on gender, politics, age or region, for example. Data bias is difficult to detect and remove.
Several techniques can be applied to address those problems:
Pre-cleaned, freely available datasets. If the problem statement (for example, image classification, object recognition) aligns with a clean, pre-existing, properly formulated dataset, then take ad-vantage of existing, open-source expertise.
Web crawling and scraping. Automated tools, bots and headless browsers can crawl and scrape websites for data.
Private data. ML engineers can create their own data. This is helpful when the amount of data required to train the model is small and the problem statement is too specific to generalize over an open-source dataset.
Custom data. Agencies can create or crowdsource the data for a fee.

QUESTION 3
Which ones are the type of visualization used for Data exploration in Data Science?

A. Heat Maps
B. Newton AI
C. Feature Distribution by Class
D. 2D-Density Plots
E. Sand Visualization

Correct Answer: A, D, E
Section:
Explanation:
Type of visualization used for exploration:
* Correlation heatmap
* Class distributions by feature
* Two-Dimensional density plots.
All the visualizations are interactive, as is standard for Plotly.
For More details, please refer the below link:
https://towardsdatascience.com/data-exploration-understanding-and-visualization-72657f5eac41

QUESTION 4
Which one is not the feature engineering techniques used in ML data science world?

A. Imputation
B. Binning
C. One hot encoding
D. Statistical

Correct Answer: D
Section:
Explanation:
Feature engineering is the pre-processing step of machine learning, which is used to transform raw data into features that can be used for creating a predictive model using Machine learning or statistical Modelling.
What is a feature?
Generally, all machine learning algorithms take input data to generate the output. The input data re-mains in a tabular form consisting of rows (instances or observations) and columns (variable or at-tributes), and these attributes are often known as features. For example, an image is an instance in computer vision, but a line in the image could be the feature. Similarly, in NLP, a document can be an observation, and the word count could be the feature. So, we can say a feature is an attribute that impacts a problem or is useful for the problem.
What is Feature Engineering?
Feature engineering is the pre-processing step of machine learning, which extracts features from raw data. It helps to represent an underlying problem to predictive models in a better way, which as a result, improve the accuracy of the model for unseen data. The predictive model contains predictor variables and an outcome variable, and while the feature engineering process selects the most useful predictor variables for the model.
Some of the popular feature engineering techniques include:
1. Imputation
Feature engineering deals with inappropriate data, missing values, human interruption, general errors, insufficient data sources, etc. Missing values within the dataset highly affect the performance of the algorithm, and to deal with them 'Imputation' technique is used. Imputation is responsible for handling irregularities within the dataset.
For example, removing the missing values from the complete row or complete column by a huge percentage of missing values. But at the same time, to maintain the data size, it is required to impute the missing data, which can be done as:
For numerical data imputation, a default value can be imputed in a column, and missing values can be filled with means or medians of the columns.
For categorical data imputation, missing values can be interchanged with the maximum occurred value in a column.
2. Handling Outliers
Outliers are the deviated values or data points that are observed too away from other data points in such a way that they badly affect the performance of the model. Outliers can be handled with this feature engineering technique. This technique first identifies the outliers and then remove them out.
Standard deviation can be used to identify the outliers. For example, each value within a space has a definite to an average distance, but if a value is greater distant than a certain value, it can be considered as an outlier. Z-score can also be used to detect outliers.
3. Log transform
Logarithm transformation or log transform is one of the commonly used mathematical techniques in machine learning. Log transform helps in handling the skewed data, and it makes the distribution more approximate to normal after transformation. It also reduces the effects of outliers on the data, as because of the normalization of magnitude differences, a model becomes much robust.
4. Binning
In machine learning, overfitting is one of the main issues that degrade the performance of the model and which occurs due to a greater number of parameters and noisy data. However, one of the popular techniques of feature engineering, 'binning', can be used to normalize the noisy data. This process involves segmenting different features into bins.
5. Feature Split
As the name suggests, feature split is the process of splitting features intimately into two or more parts and performing to make new features. This technique helps the algorithms to better understand and learn the patterns in the dataset.
The feature splitting process enables the new features to be clustered and binned, which results in extracting useful information and improving the performance of the data models.
6. One hot encoding
One hot encoding is the popular encoding technique in machine learning. It is a technique that converts the categorical data in a form so that they can be easily understood by machine learning algorithms and hence can make a good prediction. It enables group the of categorical data without losing any information.

QUESTION 5
Skewness of Normal distribution is ___________

A. Negative
B. Positive
C. 0
D. Undefined

Correct Answer: C
Section:
Explanation:
Since the normal curve is symmetric about its mean, its skewness is zero. This is a theoretical explanation for mathematical proofs, you can refer to books or websites that speak on the same in detail.

QUESTION 6
What is the formula for measuring skewness in a dataset?

A. MEAN - MEDIAN
B. MODE - MEDIAN
C. (3(MEAN - MEDIAN))/ STANDARD DEVIATION
D. (MEAN - MODE)/ STANDARD DEVIATION

Correct Answer: C
Section:
Explanation:
Since the normal curve is symmetric about its mean, its skewness is zero. This is a theoretical expla-nation for mathematical proofs, you can refer to books or websites that speak on the same in detail.

QUESTION 7
What Can Snowflake Data Scientist do in the Snowflake Marketplace as Provider?

A. Publish listings for free-to-use datasets to generate interest and new opportunities among the Snowflake customer base.
B. Publish listings for datasets that can be customized for the consumer.
C. Share live datasets securely and in real-time without creating copies of the data or im-posing data integration tasks on the consumer.
D. Eliminate the costs of building and maintaining APIs and data pipelines to deliver data to customers.

Correct Answer: A, B, C, D
Section:
Explanation:
All are correct!
About the Snowflake Marketplace
You can use the Snowflake Marketplace to discover and access third-party data and services, as well as market your own data products across the Snowflake Data Cloud.
As a data provider, you can use listings on the Snowflake Marketplace to share curated data offer-ings with many consumers simultaneously, rather than maintain sharing relationships with each indi-vidual consumer. With Paid Listings, you can also charge for your data products.
As a consumer, you might use the data provided on the Snowflake Marketplace to explore and ac-cess the following:
Historical data for research, forecasting, and machine learning.
Up-to-date streaming data, such as current weather and traffic conditions.
Specialized identity data for understanding subscribers and audience targets.
New insights from unexpected sources of data.
The Snowflake Marketplace is available globally to all non-VPS Snowflake accounts hosted on Amazon Web Services, Google Cloud Platform, and Microsoft Azure, with the exception of Mi-crosoft Azure Government. Support for Microsoft Azure Government is planned.

QUESTION 8
What Can Snowflake Data Scientist do in the Snowflake Marketplace as Consumer?

A. Discover and test third-party data sources.

B. Receive frictionless access to raw data products from vendors.

C. Combine new datasets with your existing data in Snowflake to derive new business in-sights.

D. Use the business intelligence (BI)/ML/Deep learning tools of her choice.

Correct Answer: A, B, C, D
Section:
Explanation:
As a consumer, you can do the following:
* Discover and test third-party data sources.
* Receive frictionless access to raw data products from vendors.
* Combine new datasets with your existing data in Snowflake to derive new business insights.
* Have datasets available instantly and updated continually for users.
* Eliminate the costs of building and maintaining various APIs and data pipelines to load and up-date data.
* Use the business intelligence (BI) tools of your choice.

QUESTION 9
Which one is the incorrect option to share data in Snowflake?

A. a Listing, in which you offer a share and additional metadata as a data product to one or more accounts.

B. a Direct Marketplace, in which you directly share specific database objects (a share) to another account in your region using Snowflake Marketplace.

C. a Direct Share, in which you directly share specific database objects (a share) to anoth-er account in your region.

D. a Data Exchange, in which you set up and manage a group of accounts and offer a share to that group.

Correct Answer: B
Section:
Explanation:
Options for Sharing in Snowflake
You can share data in Snowflake using one of the following options:
* a Listing, in which you offer a share and additional metadata as a data product to one or more ac-counts,
* a Direct Share, in which you directly share specific database objects (a share) to another account in your region,
* a Data Exchange, in which you set up and manage a group of accounts and offer a share to that group.

QUESTION 10
Data providers add Snowflake objects (databases, schemas, tables, secure views, etc.) to a share us-ing Which of the following options?

A. Grant privileges on objects to a share via Account role.

B. Grant privileges on objects directly to a share.

C. Grant privileges on objects to a share via a database role.

D. Grant privileges on objects to a share via a third-party role.

Correct Answer: B, C
Section:
Explanation:
What is a Share?
Shares are named Snowflake objects that encapsulate all of the information required to share a database.
Data providers add Snowflake objects (databases, schemas, tables, secure views, etc.) to a share using either or both of the following options:

Option 1: Grant privileges on objects to a share via a database role.
Option 2: Grant privileges on objects directly to a share.
You choose which accounts can consume data from the share by adding the accounts to the share.
After a database is created (in a consumer account) from a share, all the shared objects are accessible to users in the consumer account.
Shares are secure, configurable, and controlled completely by the provider account:
* New objects added to a share become immediately available to all consumers, providing real-time access to shared data.
Access to a share (or any of the objects in a share) can be revoked at any time.

QUESTION 11
Secure Data Sharing do not let you share which of the following selected objects in a database in your account with other Snowflake accounts?

A. Sequences
B. Tables
C. External tables
D. Secure UDFs

Correct Answer: A
Section:
Explanation:
Secure Data Sharing lets you share selected objects in a database in your account with other Snow-flake accounts. You can share the following Snowflake database objects:
Tables
External tables
Secure views
Secure materialized views
Secure UDFs
Snowflake enables the sharing of databases through shares, which are created by data providers and "imported" by data consumers.

QUESTION 12
Which one is incorrect understanding about Providers of Direct share?

A. A data provider is any Snowflake account that creates shares and makes them available to other Snowflake accounts to consume.
B. As a data provider, you share a database with one or more Snowflake accounts.
C. You can create as many shares as you want, and add as many accounts to a share as you want.
D. If you want to provide a share to many accounts, you can do the same via Direct Share.

Correct Answer: D
Section:
Explanation:
If you want to provide a share to many accounts, you might want to use a listing or a data ex-change.

QUESTION 13
As Data Scientist looking out to use Reader account, Which ones are the correct considerations about Reader Accounts for Third-Party Access?

A. Reader accounts (formerly known as "read-only accounts") provide a quick, easy, and cost-effective way to share data without requiring the consumer to become a Snowflake customer.
B. Each reader account belongs to the provider account that created it.
C. Users in a reader account can query data that has been shared with the reader account, but cannot perform any of the DML tasks that are allowed in a full account, such as data loading, insert, update, and

similar data manipulation operations.

D. Data sharing is only possible between Snowflake accounts.

Correct Answer: D
Section:
Explanation:
Data sharing is only supported between Snowflake accounts. As a data provider, you might want to share data with a consumer who does not already have a Snowflake account or is not ready to be-come a licensed Snowflake customer.
To facilitate sharing data with these consumers, you can create reader accounts. Reader accounts (formerly known as ''read-only accounts'') provide a quick, easy, and cost-effective way to share data without requiring the consumer to become a Snowflake customer.
Each reader account belongs to the provider account that created it. As a provider, you use shares to share databases with reader accounts; however, a reader account can only consume data from the provider account that created it.
So, Data Sharing is possible between Snowflake & Non-snowflake accounts via Reader Account.

QUESTION 14
A Data Scientist as data providers require to allow consumers to access all databases and database objects in a share by granting a single privilege on shared databases. Which one is incorrect SnowSQL command used by her while doing this task?
Assuming:
A database named product_db exists with a schema named product_agg and a table named Item_agg.
The database, schema, and table will be shared with two accounts named xy12345 and yz23456.
1. USE ROLE accountadmin;
2. CREATE DIRECT SHARE product_s;
3. GRANT USAGE ON DATABASE product_db TO SHARE product_s;
4. GRANT USAGE ON SCHEMA product_db. product_agg TO SHARE product_s;
5. GRANT SELECT ON TABLE sales_db. product_agg.Item_agg TO SHARE product_s;
6. SHOW GRANTS TO SHARE product_s;
7. ALTER SHARE product_s ADD ACCOUNTS=xy12345, yz23456;
8. SHOW GRANTS OF SHARE product_s;

A. GRANT USAGE ON DATABASE product_db TO SHARE product_s;

B. CREATE DIRECT SHARE product_s;

C. GRANT SELECT ON TABLE sales_db. product_agg.Item_agg TO SHARE product_s;

D. ALTER SHARE product_s ADD ACCOUNTS=xy12345, yz23456;

Correct Answer: C
Section:
Explanation:
CREATE SHARE product_s is the correct Snowsql command to create Share object.
Rest are correct ones.
https://docs.snowflake.com/en/user-guide/data-sharing-provider#creating-a-share-using-sql

QUESTION 15
Which object records data manipulation language (DML) changes made to tables, including inserts, updates, and deletes, as well as metadata about each change, so that actions can be taken using the changed data of Data Science Pipelines?

A. Task

B. Dynamic tables

C. Stream
D. Tags
E. Delta
F. OFFSET

Correct Answer: C
Section:
Explanation:
A stream object records data manipulation language (DML) changes made to tables, including inserts, updates, and deletes, as well as metadata about each change, so that actions can be taken using the changed data. This process is referred to as change data capture (CDC). An individual table stream tracks the changes made to rows in a source table. A table stream (also referred to as simply a ''stream'') makes a ''change table'' available of what changed, at the row level, between two transactional points of time in a table. This allows querying and consuming a sequence of change records in a transactional fashion.
Streams can be created to query change data on the following objects:
* Standard tables, including shared tables.
* Views, including secure views
* Directory tables
* Event tables

QUESTION 16
Which are the following additional Metadata columns Stream contains that could be used for creating Efficient Data science Pipelines & helps in transforming only the New/Modified data only?

A. METADATA$ACTION
B. METADATA$FILE_ID
C. METADATA$ISUPDATE
D. METADATA$DELETE
E. METADATA$ROW_ID

Correct Answer: A, C, E
Section:
Explanation:
A stream stores an offset for the source object and not any actual table columns or data. When que-ried, a stream accesses and returns the historic data in the same shape as the source object (i.e. the same column names and ordering) with the following additional columns:
METADATA$ACTION
Indicates the DML operation (INSERT, DELETE) recorded.
METADATA$ISUPDATE
Indicates whether the operation was part of an UPDATE statement. Updates to rows in the source object are represented as a pair of DELETE and INSERT records in the stream with a metadata column METADATA$ISUPDATE values set to TRUE.
Note that streams record the differences between two offsets. If a row is added and then updated in the current offset, the delta change is a new row. The METADATA$ISUPDATE row records a FALSE value.
METADATA$ROW_ID
Specifies the unique and immutable ID for the row, which can be used to track changes to specific rows over time.

QUESTION 17
Mark the Incorrect understanding of Data Scientist about Streams?

A. Streams on views support both local views and views shared using Snowflake Secure Data Sharing, including secure views.
B. Streams can track changes in materialized views.
C. Streams itself does not contain any table data.

D. Streams do not support repeatable read isolation.

Correct Answer: B, D
Section:
Explanation:
Streams on views support both local views and views shared using Snowflake Secure Data Sharing, including secure views. Currently, streams cannot track changes in materialized views.
stream itself does not contain any table data. A stream only stores an offset for the source object and returns CDC records by leveraging the versioning history for the source object. When the first stream for a table is created, several hidden columns are added to the source table and begin storing change tracking metadata. These columns consume a small amount of storage. The CDC records returned when querying a stream rely on a combination of the offset stored in the stream and the change tracking metadata stored in the table. Note that for streams on views, change tracking must be enabled explicitly for the view and underlying tables to add the hidden columns to these tables.
Streams support repeatable read isolation. In repeatable read mode, multiple SQL statements within a transaction see the same set of records in a stream. This differs from the read committed mode supported for tables, in which statements see any changes made by previous statements executed within the same transaction, even though those changes are not yet committed.
The delta records returned by streams in a transaction is the range from the current position of the stream until the transaction start time. The stream position advances to the transaction start time if the transaction commits; otherwise it stays at the same position.

QUESTION 18
Data Scientist used streams in ELT (extract, load, transform) processes where new data inserted in-to a staging table is tracked by a stream. A set of SQL statements transform and insert the stream contents into a set of production tables. Raw data is coming in the JSON format, but for analysis he needs to transform it into relational columns in the production tables. which of the following Data transformation SQL function he can used to achieve the same?

A. He could not apply Transformation on Stream table data.
B. lateral flatten()
C. METADATA$ACTION ()
D. Transpose()

Correct Answer: B
Section:
Explanation:
To know about lateral flatten SQL Function, please refer:
https://docs.snowflake.com/en/sql-reference/constructs/join-lateral#example-of-using-lateral-with-flatten

QUESTION 19
Which command manually triggers a single run of a scheduled task (either a standalone task or the root task in a DAG) independent of the schedule defined for the task?

A. RUN TASK
B. CALL TASK
C. EXECUTE TASK
D. RUN ROOT TASK

Correct Answer: C
Section:
Explanation:
The EXECUTE TASK command manually triggers a single run of a scheduled task (either a standalone task or the root task in a DAG) independent of the schedule defined for the task. A successful run of a root task triggers a cascading run of child tasks in the DAG as their precedent task completes, as though the root task had run on its defined schedule.
This SQL command is useful for testing new or modified standalone tasks and DAGs before you enable them to execute SQL code in production.
Call this SQL command directly in scripts or in stored procedures. In addition, this command sup-ports integrating tasks in external data pipelines. Any third-party services that can authenticate into your Snowflake account and authorize SQL actions can execute the EXECUTE TASK command to run tasks.

QUESTION 20
Which of the following Snowflake parameter can be used to Automatically Suspend Tasks which are running Data science pipelines after specified Failed Runs?

A. SUSPEND_TASK
B. SUSPEND_TASK_AUTO_NUM_FAILURES
C. SUSPEND_TASK_AFTER_NUM_FAILURES
D. There is none as such available.

Correct Answer: C
Section:
Explanation:
Automatically Suspend Tasks After Failed Runs
Optionally suspend tasks automatically after a specified number of consecutive runs that either fail or time out. This feature can reduce costs by suspending tasks that consume Snowflake credits but fail to run to completion. Failed task runs include runs in which the SQL code in the task body either produces a user error or times out. Task runs that are skipped, canceled, or that fail due to a sys-tem error are considered indeterminate and are not included in the count of failed task runs.
Set the SUSPEND_TASK_AFTER_NUM_FAILURES = num parameter on a standalone task or the root task in a DAG. When the parameter is set to a value greater than 0, the following behavior applies to runs of the standalone task or DAG:
Standalone tasks are automatically suspended after the specified number of consecutive task runs either fail or time out.
The root task is automatically suspended after the run of any single task in a DAG fails or times out the specified number of times in consecutive runs.
The parameter can be set when creating a task (using CREATE TASK) or later (using ALTER TASK). The setting applies to tasks that rely on either Snowflake-managed compute resources (i.e. serverless compute model) or user-managed compute resources (i.e. a virtual warehouse).
The SUSPEND_TASK_AFTER_NUM_FAILURES parameter can also be set at the account, database, or schema level. The setting applies to all standalone or root tasks contained in the modified object. Note that explicitly setting the parameter at a lower (i.e. more granular) level overrides the parameter value set at a higher level.

QUESTION 21
Mark the incorrect statement regarding Python UDF?

A. Python UDFs can contain both new code and calls to existing packages
B. For each row passed to a UDF, the UDF returns either a scalar (i.e. single) value or, if defined as a table function, a set of rows.
C. A UDF also gives you a way to encapsulate functionality so that you can call it repeatedly from multiple places in code
D. A scalar function (UDF) returns a tabular value for each input row

Correct Answer: D
Section:
Explanation:
A scalar function (UDF) returns one output row for each input row. The returned row consists of a single column/value

QUESTION 22
Data Scientist can query, process, and transform data in a which of the following ways using Snowpark Python. [Select 2]

A. Query and process data with a DataFrame object.
B. Write a user-defined tabular function (UDTF) that processes data and returns data in a set of rows with one or more columns.
C. SnowPark currently do not support writing UDTF.
D. Transform Data using DataIKY tool with SnowPark API.

Correct Answer: A, C

Section:
Explanation:
Query and process data with a DataFrame object. Refer to Working with DataFrames in Snowpark Python.
Convert custom lambdas and functions to user-defined functions (UDFs) that you can call to process data.
Write a user-defined tabular function (UDTF) that processes data and returns data in a set of rows with one or more columns.
Write a stored procedure that you can call to process data, or automate with a task to build a data pipeline.

QUESTION 23
Which Python method can be used to Remove duplicates by Data scientist?

A. remove_duplicates()
B. duplicates()
C. drop_duplicates()
D. clean_duplicates()

Correct Answer: D
Section:
Explanation:
The drop_duplicates() method removes duplicate rows.
dataframe.drop_duplicates(subset, keep, inplace, ignore_index)
Remove duplicate rows from the DataFrame:

```
1. import pandas as pd
2. data = {
3. 'name': ['Peter', 'Mary', 'John', 'Mary'],
4. 'age': [50, 40, 30, 40],
5. 'qualified': [True, False, False, False]
6. }
7.
8. df = pd.DataFrame(data)
9. newdf = df.drop_duplicates()
```

QUESTION 24
Consider a data frame df with 10 rows and index ['r1', 'r2', 'r3', 'row4', 'row5', 'row6', 'r7', 'r8', 'r9', 'row10']. What does the aggregate method shown in below code do?
g = df.groupby(df.index.str.len())
g.aggregate({'A':len, 'B':np.sum})

A. Computes Sum of column A values
B. Computes length of column A
C. Computes length of column A and Sum of Column B values of each group
D. Computes length of column A and Sum of Column B values

Correct Answer: C
Section:
Explanation:
Computes length of column A and Sum of Column B values of each group

QUESTION 25

Consider a data frame df with columns ['A', 'B', 'C', 'D'] and rows ['r1', 'r2', 'r3']. What does the ex-pression df[lambda x : x.index.str.endswith('3')] do?

A. Returns the row name r3
B. Results in Error
C. Returns the third column
D. Filters the row labelled r3

Correct Answer: D
Section:
Explanation:
It will Filters the row labelled r3.

QUESTION 26
Consider a data frame df with 10 rows and index ['r1', 'r2', 'r3', 'row4', 'row5', 'row6', 'r7', 'r8', 'r9', 'row10']. What does the expression g = df.groupby(df.index.str.len()) do?

A. Groups df based on index values
B. Groups df based on length of each index value
C. Groups df based on index strings
D. Data frames cannot be grouped by index values. Hence it results in Error.

Correct Answer: D
Section:
Explanation:
Data frames cannot be grouped by index values. Hence it results in Error.

QUESTION 27
Which command is used to install Jupyter Notebook?

A. pip install jupyter
B. pip install notebook
C. pip install jupyter-notebook
D. pip install nbconvert

Correct Answer: A
Section:
Explanation:
Jupyter Notebook is a web-based interactive computational environment.
The command used to install Jupyter Notebook is pip install jupyter.
The command used to start Jupyter Notebook is jupyter notebook.

QUESTION 28
Which of the following process best covers all of the following characteristics?
* Collecting descriptive statistics like min, max, count and sum.
* Collecting data types, length and recurring patterns.
* Tagging data with keywords, descriptions or categories.
* Performing data quality assessment, risk of performing joins on the data.

* Discovering metadata and assessing its accuracy.
Identifying distributions, key candidates, foreign-key candidates, functional dependencies, embedded value dependencies, and performing inter-table analysis.

A. Data Visualization
B. Data Virtualization
C. Data Profiling
D. Data Collection

Correct Answer: C
Section:
Explanation:
Data processing and analysis cannot happen without data profiling---reviewing source data for con-tent and quality. As data gets bigger and infrastructure moves to the cloud, data profiling is increasingly important.
What is data profiling?
Data profiling is the process of reviewing source data, understanding structure, content and interrelationships, and identifying potential for data projects.
Data profiling is a crucial part of:
* Data warehouse and business intelligence (DW/BI) projects---data profiling can uncover data quality issues in data sources, and what needs to be corrected in ETL.
* Data conversion and migration projects---data profiling can identify data quality issues, which you can handle in scripts and data integration tools copying data from source to target. It can also un-cover new requirements for the target system.
* Source system data quality projects---data profiling can highlight data which suffers from serious or numerous quality issues, and the source of the issues (e.g. user inputs, errors in interfaces, data corruption).
Data profiling involves:
* Collecting descriptive statistics like min, max, count and sum.
* Collecting data types, length and recurring patterns.
* Tagging data with keywords, descriptions or categories.
* Performing data quality assessment, risk of performing joins on the data.
* Discovering metadata and assessing its accuracy.
* Identifying distributions, key candidates, foreign-key candidates, functional dependencies, embedded value dependencies, and performing inter-table analysis.

QUESTION 29
Which of the Following is not type of Windows function in Snowflake?

A. Rank-related functions.
B. Window frame functions.
C. Aggregation window functions.
D. Association functions.

Correct Answer: C, D
Section:
Explanation:
Window Functions
A window function operates on a group (''window'') of related rows.
Each time a window function is called, it is passed a row (the current row in the window) and the window of rows that contain the current row. The window function returns one output row for each input row. The output depends on the individual row passed to the function and the values of the other rows in the window passed to the function.
Some window functions are order-sensitive. There are two main types of order-sensitive window functions:
Rank-related functions.
Window frame functions.
Rank-related functions list information based on the ''rank'' of a row. For example, if you rank stores in descending order by profit per year, the store with the most profit will be ranked 1; the second-most

profitable store will be ranked 2, etc.
Window frame functions allow you to perform rolling operations, such as calculating a running total or a moving average, on a subset of the rows in the window.

QUESTION 30
Which of the following Functions do Support Windowing?

A. HASH_AGG
B. ENCRYPT
C. EXTRACT
D. LISTAGG

Correct Answer: D
Section:
Explanation:
What is a Window?
A window is a group of related rows. For example, a window might be defined based on timestamps, with all rows in the same month grouped in the same window. Or a window might be defined based on location, with all rows from a particular city grouped in the same window.
A window can consist of zero, one, or multiple rows. For simplicity, Snowflake documentation usually says that a window contains multiple rows.
What is a Window Function?
A window function is any function that operates over a window of rows.
A window function is generally passed two parameters:
A row. More precisely, a window function is passed 0 or more expressions. In almost all cases, at least one of those expressions references a column in that row. (Most window functions require at least one column or expression, but a few window functions, such as some rank-related functions, do not required an explicit column or expression.)
A window of related rows that includes that row. The window can be the entire table, or a subset of the rows in the table.
For non-window functions, all arguments are usually passed explicitly to the function, for example:
MY_FUNCTION(argument1, argument2, ...)
Window functions behave differently; although the current row is passed as an argument the normal way, the window is passed through a separate clause, called an OVER clause. The syntax of the OVER clause is documented later.
LISTAGG
Returns the concatenated input values, separated by the delimiter string.
Window function
1. LISTAGG([DISTINCT] <expr1> [, <delimiter>])
2. [WITHIN GROUP (<orderby_clause>)]
3. OVER ([PARTITION BY <expr2>])
HASH_AGG
Returns an aggregate signed 64-bit hash value over the (unordered) set of input rows. HASH_AGG never returns NULL, even if no input is provided. Empty input ''hashes'' to 0.
Window function
HASH_AGG([DISTINCT] <expr> [, <expr2> ...]) OVER ([PARTITION BY <expr3>])
HASH_AGG(*) OVER ([PARTITION BY <expr3>])

QUESTION 31
All aggregate functions except ______ ignore null values in their input collection

A. Count(attribute)
B. Count(*)
C. Avg
D. Sum

Correct Answer: B
Section:
Explanation:
Count(*)
* is used to select all values including null.

QUESTION 32
Mark the Incorrect statements regarding MIN / MAX Functions?

A. NULL values are skipped unless all the records are NULL
B. NULL values are ignored unless all the records are NULL, in which case a NULL value is returned
C. The data type of the returned value is the same as the data type of the input values
D. For compatibility with other systems, the DISTINCT keyword can be specified as an argument for MIN or MAX, but it does not have any effect

Correct Answer: B
Section:
Explanation:
NULL values are ignored unless all the records are NULL, in which case a NULL value is returned

QUESTION 33
Which one is not the types of Feature Engineering Transformation?

A. Scaling
B. Encoding
C. Aggregation
D. Normalization

Correct Answer: C
Section:
Explanation:
What is Feature Engineering?
Feature engineering is the process of transforming raw data into features that are suitable for ma-chine learning models. In other words, it is the process of selecting, extracting, and transforming the most relevant features from the available data to build more accurate and efficient machine learning models.
The success of machine learning models heavily depends on the quality of the features used to train them. Feature engineering involves a set of techniques that enable us to create new features by combining or transforming the existing ones. These techniques help to highlight the most important pat-terns and relationships in the data, which in turn helps the machine learning model to learn from the data more effectively.
What is a Feature?
In the context of machine learning, a feature (also known as a variable or attribute) is an individual measurable property or characteristic of a data point that is used as input for a machine learning al-gorithm. Features can be numerical, categorical, or text-based, and they represent different aspects of the data that are relevant to the problem at hand.
For example, in a dataset of housing prices, features could include the number of bedrooms, the square footage, the location, and the age of the property. In a dataset of customer demographics, features could include age, gender, income level, and occupation.
The choice and quality of features are critical in machine learning, as they can greatly impact the ac-curacy and performance of the model.
Why do we Engineer Features?
We engineer features to improve the performance of machine learning models by providing them with relevant and informative input data. Raw data may contain noise, irrelevant information, or missing values, which can lead to inaccurate or biased model predictions. By engineering features, we can extract meaningful information from the raw data, create new variables that capture important patterns and relationships, and transform the data into a more suitable format for machine learning algorithms.
Feature engineering can also help in addressing issues such as overfitting, underfitting, and high di-mensionality. For example, by reducing the number of features, we can prevent the model from be-coming too

complex or overfitting to the training data. By selecting the most relevant features, we can improve the model's accuracy and interpretability.
In addition, feature engineering is a crucial step in preparing data for analysis and decision-making in various fields, such as finance, healthcare, marketing, and social sciences. It can help uncover hidden insights, identify trends and patterns, and support data-driven decision-making.
We engineer features for various reasons, and some of the main reasons include:
Improve User Experience: The primary reason we engineer features is to enhance the user experience of a product or service. By adding new features, we can make the product more intuitive, efficient, and user-friendly, which can increase user satisfaction and engagement.
Competitive Advantage: Another reason we engineer features is to gain a competitive advantage in the marketplace. By offering unique and innovative features, we can differentiate our product from competitors and attract more customers.
Meet Customer Needs: We engineer features to meet the evolving needs of customers. By analyzing user feedback, market trends, and customer behavior, we can identify areas where new features could enhance the product's value and meet customer needs.
Increase Revenue: Features can also be engineered to generate more revenue. For example, a new feature that streamlines the checkout process can increase sales, or a feature that provides additional functionality could lead to more upsells or cross-sells.
Future-Proofing: Engineering features can also be done to future-proof a product or service. By an-ticipating future trends and potential customer needs, we can develop features that ensure the product remains relevant and useful in the long term.
Processes Involved in Feature Engineering
Feature engineering in Machine learning consists of mainly 5 processes: Feature Creation, Feature Transformation, Feature Extraction, Feature Selection, and Feature Scaling. It is an iterative process that requires experimentation and testing to find the best combination of features for a given problem. The success of a machine learning model largely depends on the quality of the features used in the model.
Feature Transformation
Feature Transformation is the process of transforming the features into a more suitable representation for the machine learning model. This is done to ensure that the model can effectively learn from the data.
Types of Feature Transformation:
Normalization: Rescaling the features to have a similar range, such as between 0 and 1, to prevent some features from dominating others.
Scaling: Rescaling the features to have a similar scale, such as having a standard deviation of 1, to make sure the model considers all features equally.
Encoding: Transforming categorical features into a numerical representation. Examples are one-hot encoding and label encoding.
Transformation: Transforming the features using mathematical operations to change the distribution or scale of the features. Examples are logarithmic, square root, and reciprocal transformations.

QUESTION 34
Which one is not Types of Feature Scaling?

A. Economy Scaling
B. Min-Max Scaling
C. Standard Scaling
D. Robust Scaling

Correct Answer: B
Section:
Explanation:
Feature Scaling
Feature Scaling is the process of transforming the features so that they have a similar scale. This is important in machine learning because the scale of the features can affect the performance of the model.
Types of Feature Scaling:
Min-Max Scaling: Rescaling the features to a specific range, such as between 0 and 1, by subtracting the minimum value and dividing by the range.
Standard Scaling: Rescaling the features to have a mean of 0 and a standard deviation of 1 by subtracting the mean and dividing by the standard deviation.
Robust Scaling: Rescaling the features to be robust to outliers by dividing them by the interquartile range.
Benefits of Feature Scaling:
Improves Model Performance: By transforming the features to have a similar scale, the model can learn from all features equally and avoid being dominated by a few large features.
Increases Model Robustness: By transforming the features to be robust to outliers, the model can become more robust to anomalies.
Improves Computational Efficiency: Many machine learning algorithms, such as k-nearest neighbors, are sensitive to the scale of the features and perform better with scaled features.
Improves Model Interpretability: By transforming the features to have a similar scale, it can be easier to understand the model's predictions.

QUESTION 35

Select the Correct Statements regarding Normalization?

A. Normalization technique uses minimum and max values for scaling of model.

B. Normalization technique uses mean and standard deviation for scaling of model.

C. Scikit-Learn provides a transformer RecommendedScaler for Normalization.

D. Normalization got affected by outliers.

Correct Answer: A, D

Section:

Explanation:

Normalization is a scaling technique in Machine Learning applied during data preparation to change the values of numeric columns in the dataset to use a common scale. It is not necessary for all datasets in a model. It is required only when features of machine learning models have different ranges.

Scikit-Learn provides a transformer called MinMaxScaler for Normalization.

This technique uses minimum and max values for scaling of model.It is useful when feature distribution is unknown.It got affected by outliers.

QUESTION 36

To return the contents of a DataFrame as a Pandas DataFrame, Which of the following method can be used in SnowPark API?

A. REPLACE_TO_PANDAS

B. SNOWPARK_TO_PANDAS

C. CONVERT_TO_PANDAS

D. TO_PANDAS

Correct Answer: D

Section:

Explanation:

To return the contents of a DataFrame as a Pandas DataFrame, use the to_pandas method.

For example:

1. >>> python_df = session.create_dataframe(['a', 'b', 'c'])
2. >>> pandas_df = python_df.to_pandas()

QUESTION 37

Which of the following is a Python-based web application framework for visualizing data and analyzing results in a more efficient and flexible way?

A. StreamBI

B. Streamlit

C. Streamsets

D. Rapter

Correct Answer: B

Section:

Explanation:

Streamlit is a Python-based web application framework for visualizing data and analyzing results in a more efficient and flexible way. It is an open source library that assists data scientists and academics to develop Machine Learning (ML) visualization dashboards in a short period of time. We can build and deploy powerful data applications with just a few lines of code.

Why Streamlit?

Currently, real-world applications are in high demand and developers are developing new libraries and frameworks to make on-the-go dashboards easier to build and deploy. Streamlit is a library that reduces your dashboard development time from days to hours. Following are some reasons to choose the Streamlit:
It is a free and open-source library.
Installing Streamlit is as simple as installing any other python package
It is easy to learn because you won't need any web development experience, only a basic under-standing of Python is enough to build a data application.
It is compatible with almost all machine learning frameworks, including Tensorflow and Pytorch, Scikit-learn, and visualization libraries such as Seaborn, Altair, Plotly, and many others.

QUESTION 38
Which is the visual depiction of data through the use of graphs, plots, and informational graphics?

A. Data Interpretation
B. Data Virtualization
C. Data visualization
D. Data Mining

Correct Answer: D
Section:
Explanation:
Data visualization is the visual depiction of data through the use of graphs, plots, and informational graphics. Its practitioners use statistics and data science to convey the meaning behind data in ethical and accurate ways.

QUESTION 39
Which method is used for detecting data outliers in Machine learning?

A. Scaler
B. Z-Score
C. BOXI
D. CMIYC

Correct Answer: B
Section:
Explanation:
What are outliers?
Outliers are the values that look different from the other values in the data. Below is a plot high-lighting the outliers in 'red' and outliers can be seen in both the extremes of data.
Reasons for outliers in data
Errors during data entry or a faulty measuring device (a faulty sensor may result in extreme readings).
Natural occurrence (salaries of junior level employees vs C-level employees)
Problems caused by outliers
Outliers in the data may causes problems during model fitting (esp. linear models).
Outliers may inflate the error metrics which give higher weights to large errors (example, mean squared error, RMSE).
Z-score method is of the method for detecting outliers. This method is generally used when a variable' distribution looks close to Gaussian. Z-score is the number of standard deviations a value of a variable is away from the variable' mean.
Z-Score = (X-mean) / Standard deviation
IQR method , Box plots are some more example of methods used to detect data outliers in Data science.

QUESTION 40
Mark the correct steps for saving the contents of a DataFrame to a Snowflake table as part of Moving Data from Spark to Snowflake?

A. Step 1.Use the PUT() method of the DataFrame to construct a DataFrameWriter. Step 2.Specify SNOWFLAKE_SOURCE_NAME using the NAME() method. Step 3.Use the dbtable option to specify the table to which data is written. Step 4.Specify the connector options using either the option() or options() method. Step 5.Use the save() method to specify the save mode for the content.

B. Step 1.Use the PUT() method of the DataFrame to construct a DataFrameWriter. Step 2.Specify SNOWFLAKE_SOURCE_NAME using the format() method. Step 3.Specify the connector options using either the option() or options() method. Step 4.Use the dbtable option to specify the table to which data is written. Step 5.Use the save() method to specify the save mode for the content.

C. Step 1.Use the write() method of the DataFrame to construct a DataFrameWriter. Step 2.Specify SNOWFLAKE_SOURCE_NAME using the format() method. Step 3.Specify the connector options using either the option() or options() method. Step 4.Use the dbtable option to specify the table to which data is written. Step 5.Use the mode() method to specify the save mode for the content. (Correct)

D. Step 1.Use the writer() method of the DataFrame to construct a DataFrameWriter. Step 2.Specify SNOWFLAKE_SOURCE_NAME using the format() method. Step 3.Use the dbtable option to specify the table to which data is written. Step 4.Specify the connector options using either the option() or options() method. Step 5.Use the save() method to specify the save mode for the content.

Correct Answer: C
Section:
Explanation:
Moving Data from Spark to Snowflake
The steps for saving the contents of a DataFrame to a Snowflake table are similar to writing from Snowflake to Spark:
1. Use the write() method of the DataFrame to construct a DataFrameWriter.
2. Specify SNOWFLAKE_SOURCE_NAME using the format() method.
3. Specify the connector options using either the option() or options() method.
4. Use the dbtable option to specify the table to which data is written.
5. Use the mode() method to specify the save mode for the content.
Examples
1. df.write
2. .format(SNOWFLAKE_SOURCE_NAME)
3. .options(sfOptions)
4. .option('dbtable', 't2')
5. .mode(SaveMode.Overwrite)
6. .save()

QUESTION 41
Select the Data Science Tools which are known to provide native connectivity to Snowflake?

A. Denodo

B. DvSUM

C. DiYotta

D. HEX

Correct Answer: D
Section:
Explanation:
Hex --- collaborative data science and analytics platform
Denodo --- data virtualization and federation platform
DvSum --- data catalog and data intelligence platform
Diyotta --- data integration and migration

QUESTION 42
Which one of the following is not the key component while designing External functions within Snowflake?

A. Remote Service
B. API Integration
C. UDF Service
D. Proxy Service

Correct Answer: C
Section:
Explanation:
What is an External Function?
An external function calls code that is executed outside Snowflake.
The remotely executed code is known as a remote service.
Information sent to a remote service is usually relayed through a proxy service.
Snowflake stores security-related external function information in an API integration.
External Function:
An external function is a type of UDF. Unlike other UDFs, an external function does not contain its own code; instead, the external function calls code that is stored and executed outside Snowflake.
Inside Snowflake, the external function is stored as a database object that contains information that Snowflake uses to call the remote service. This stored information includes the URL of the proxy service that relays information to and from the remote service.
Remote Service:
The remotely executed code is known as a remote service.
The remote service must act like a function. For example, it must return a value.
Snowflake supports scalar external functions; the remote service must return exactly one row for each row received.
Proxy Service:
Snowflake does not call a remote service directly. Instead, Snowflake calls a proxy service, which relays the data to the remote service.
The proxy service can increase security by authenticating requests to the remote service.
The proxy service can support subscription-based billing for a remote service. For example, the proxy service can verify that a caller to the remote service is a paid subscriber.
The proxy service also relays the response from the remote service back to Snowflake.
Examples of proxy services include:
Amazon API Gateway.
Microsoft Azure API Management service.
API Integration:
An integration is a Snowflake object that provides an interface between Snowflake and third-party services. An API integration stores information, such as security information, that is needed to work with a proxy service or remote service.
An API integration is created with the CREATE API INTEGRATION command.
Users can write and call their own remote services, or call remote services written by third parties. These remote services can be written using any HTTP server stack, including cloud serverless compute services such as AWS Lambda.

QUESTION 43
Which ones are the known limitations of using External function?

A. Currently, external functions cannot be shared with data consumers via Secure Data Sharing.
B. Currently, external functions must be scalar functions. A scalar external function re-turns a single value for each input row.
C. External functions have more overhead than internal functions (both built-in functions and internal UDFs) and usually execute more slowly
D. An external function accessed through an AWS API Gateway private endpoint can be accessed only from a Snowflake VPC (Virtual Private Cloud) on AWS and in the same AWS region.

Correct Answer: A, B, C, D
Section:

QUESTION 44
What is the risk with tuning hyper-parameters using a test dataset?

A. Model will overfit the test set
B. Model will underfit the test set
C. Model will overfit the training set
D. Model will perform balanced

Correct Answer: A
Section:
Explanation:
The model will not generalize well to unseen data because it overfits the test set. Tuning model hyper-parameters to a test set means that the hyper-parameters may overfit to that test set. If the same test set is used to estimate performance, it will produce an overestimate. The test set should be used only for testing, not for parameter tuning.
Using a separate validation set for tuning and test set for measuring performance provides unbiased, realistic measurement of performance.
What are hyper-parameters?
Hyper-parameters are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning. We can't calculate their values from the data.
Example: Number of clusters in clustering, number of hidden layers in a neural network, and depth of a tree are some of the examples of hyper-parameters.
What is the hyper-parameter tuning?
Hyper-parameter tuning is the process of choosing the right combination of hyper-parameters that maximizes the model performance. It works by running multiple trials in a single training process. Each trial is a complete execution of your training application with values for your chosen hyper-parameters, set within the limits you specify. This process once finished will give you the set of hyper-parameter values that are best suited for the model to give optimal results.

QUESTION 45
Select the correct mappings:
I) W Weights or Coefficients of independent variables in the Linear regression model --> Model Pa-rameter
II) K in the K-Nearest Neighbour algorithm --> Model Hyperparameter
III) Learning rate for training a neural network --> Model Hyperparameter
IV) Batch Size --> Model Parameter

A. I,II
B. I,II,III
C. III,IV
D. II,III,IV

Correct Answer: B
Section:
Explanation:
Hyperparameters in Machine learning are those parameters that are explicitly defined by the user to control the learning process. These hyperparameters are used to improve the learning of the model, and their values are set before starting the learning process of the model.
What are hyperparameters?
In Machine Learning/Deep Learning, a model is represented by its parameters. In contrast, a training process involves selecting the best/optimal hyperparameters that are used by learning algorithms to provide the best result. So, what are these hyperparameters? The answer is, 'Hyperparameters are defined as the parameters that are explicitly defined by the user to control the learning process.'
Here the prefix 'hyper' suggests that the parameters are top-level parameters that are used in con-trolling the learning process. The value of the Hyperparameter is selected and set by the machine learning engineer before the learning algorithm begins training the model. Hence, these are external to the model, and their values cannot be changed during the training process.
Some examples of Hyperparameters in Machine Learning
* The k in kNN or K-Nearest Neighbour algorithm

* Learning rate for training a neural network
* Train-test split ratio
* Batch Size
* Number of Epochs
* Branches in Decision Tree
* Number of clusters in Clustering Algorithm

Model Parameters:

Model parameters are configuration variables that are internal to the model, and a model learns them on its own. For example, W Weights or Coefficients of independent variables in the Linear regression model. or Weights or Coefficients of independent variables in SVM, weight, and biases of a neural network, cluster centroid in clustering. Some key points for model parameters are as follows:

They are used by the model for making predictions.

* They are learned by the model from the data itself
* These are usually not set manually.
* These are the part of the model and key to a machine learning Algorithm.

Model Hyperparameters:

Hyperparameters are those parameters that are explicitly defined by the user to control the learning process. Some key points for model parameters are as follows:

These are usually defined manually by the machine learning engineer.

One cannot know the exact best value for hyperparameters for the given problem. The best value can be determined either by the rule of thumb or by trial and error.

Some examples of Hyperparameters are the learning rate for training a neural network, K in the KNN algorithm.

QUESTION 46

Performance metrics are a part of every machine learning pipeline, Which ones are not the performance metrics used in the Machine learning?

A. R (R-Squared)

B. Root Mean Squared Error (RMSE)

C. AU-ROC

D. AUM

Correct Answer: D

Section:

Explanation:

Every machine learning task can be broken down to either Regression or Classification, just like the performance metrics.

Metrics are used to monitor and measure the performance of a model (during training and testing), and do not need to be differentiable.

Regression metrics

Regression models have continuous output. So, we need a metric based on calculating some sort of distance between predicted and ground truth.

In order to evaluate Regression models, we'll discuss these metrics in detail:

* Mean Absolute Error (MAE),
* Mean Squared Error (MSE),
* Root Mean Squared Error (RMSE),
* R (R-Squared).

Mean Squared Error (MSE)

Mean squared error is perhaps the most popular metric used for regression problems. It essentially finds the average of the squared difference between the target value and the value predicted by the regression model.

Few key points related to MSE:

* It's differentiable, so it can be optimized better.
* It penalizes even small errors by squaring them, which essentially leads to an overestimation of how bad the model is.
* Error interpretation has to be done with squaring factor(scale) in mind. For example in our Boston Housing regression problem, we got MSE=21.89 which primarily corresponds to (Prices).
* Due to the squaring factor, it's fundamentally more prone to outliers than other metrics.

Mean Absolute Error (MAE)

Mean Absolute Error is the average of the difference between the ground truth and the predicted values.
Few key points for MAE
* It's more robust towards outliers than MAE, since it doesn't exaggerate errors.
* It gives us a measure of how far the predictions were from the actual output. However, since MAE uses absolute value of the residual, it doesn't give us an idea of the direction of the error, i.e. whether we're under-predicting or over-predicting the data.
* Error interpretation needs no second thoughts, as it perfectly aligns with the original degree of the variable.
* MAE is non-differentiable as opposed to MSE, which is differentiable.
Root Mean Squared Error (RMSE)
Root Mean Squared Error corresponds to the square root of the average of the squared difference between the target value and the value predicted by the regression model.
Few key points related to RMSE:
* It retains the differentiable property of MSE.
* It handles the penalization of smaller errors done by MSE by square rooting it.
* Error interpretation can be done smoothly, since the scale is now the same as the random variable.
* Since scale factors are essentially normalized, it's less prone to struggle in the case of outliers.
R Coefficient of determination
R Coefficient of determination actually works as a post metric, meaning it's a metric that's calcu-lated using other metrics.
The point of even calculating this coefficient is to answer the question ''How much (what %) of the total variation in Y(target) is explained by the variation in X(regression line)''
Few intuitions related to R results:
If the sum of Squared Error of the regression line is small => R will be close to 1 (Ideal), meaning the regression was able to capture 100% of the variance in the target variable.
Conversely, if the sum of squared error of the regression line is high => R will be close to 0, meaning the regression wasn't able to capture any variance in the target variable.
You might think that the range of R is (0,1) but it's actually (-,1) because the ratio of squared errors of the regression line and mean can surpass the value 1 if the squared error of regression line is too high (>squared error of the mean).
Classification metrics
Classification problems are one of the world's most widely researched areas. Use cases are present in almost all production and industrial environments. Speech recognition, face recognition, text classification -- the list is endless.
Classification models have discrete output, so we need a metric that compares discrete classes in some form. Classification Metrics evaluate a model's performance and tell you how good or bad the classification is, but each of them evaluates it in a different way.
So in order to evaluate Classification models, we'll discuss these metrics in detail:
Accuracy
Confusion Matrix (not a metric but fundamental to others)
Precision and Recall
F1-score
AU-ROC
Accuracy
Classification accuracy is perhaps the simplest metric to use and implement and is defined as the number of correct predictions divided by the total number of predictions, multiplied by 100.
We can implement this by comparing ground truth and predicted values in a loop or simply utilizing the scikit-learn module to do the heavy lifting for us (not so heavy in this case).
Confusion Matrix
Confusion Matrix is a tabular visualization of the ground-truth labels versus model predictions. Each row of the confusion matrix represents the instances in a predicted class and each column represents the instances in an actual class. Confusion Matrix is not exactly a performance metric but sort of a basis on which other metrics evaluate the results.
Each cell in the confusion matrix represents an evaluation factor. Let's understand these factors one by one:
* True Positive(TP) signifies how many positive class samples your model predicted correctly.
* True Negative(TN) signifies how many negative class samples your model predicted correctly.
* False Positive(FP) signifies how many negative class samples your model predicted incorrectly. This factor represents Type-I error in statistical nomenclature. This error positioning in the confusion matrix depends on the choice of the null hypothesis.
* False Negative(FN) signifies how many positive class samples your model predicted incorrectly. This factor represents Type-II error in statistical nomenclature. This error positioning in the confu-sion matrix also depends on the choice of the null hypothesis.
Precision
Precision is the ratio of true positives and total positives predicted

Recall/Sensitivity/Hit-Rate
A Recall is essentially the ratio of true positives to all the positives in ground truth.
Precision-Recall tradeoff
To improve your model, you can either improve precision or recall -- but not both! If you try to re-duce cases of non-cancerous patients being labeled as cancerous (FN/type-II), no direct effect will take place on cancerous patients being labeled as non-cancerous.
F1-score
The F1-score metric uses a combination of precision and recall. In fact, the F1 score is the harmonic mean of the two.
AUROC (Area under Receiver operating characteristics curve)
Better known as AUC-ROC score/curves. It makes use of true positive rates(TPR) and false posi-tive rates(FPR).

QUESTION 47
Which of the following cross validation versions may not be suitable for very large datasets with hundreds of thousands of samples?

A. k-fold cross-validation
B. Leave-one-out cross-validation
C. Holdout method
D. All of the above

Correct Answer: B
Section:
Explanation:
Leave-one-out cross-validation (LOO cross-validation) is not suitable for very large datasets due to the fact that this validation technique requires one model for every sample in the training set to be created and evaluated.
Cross validation
It is a technique to evaluate a machine learning model and it is the basis for whole class of model evaluation methods. The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it. It works by the idea of splitting dataset into number of subsets, keep a subset aside, train the model, and test the model on the holdout subset.
Leave-one-out cross validation
Leave-one-out cross validation is K-fold cross validation taken to its logical extreme, with K equal to N, the number of data points in the set. That means that N separate times, the function approximator is trained on all the data except for one point and a prediction is made for that point. As be-fore the average error is computed and used to evaluate the model. The evaluation given by leave-one-out cross validation is very expensive to compute at first pass.

QUESTION 48
Which of the following cross validation versions is suitable quicker cross-validation for very large datasets with hundreds of thousands of samples?

A. k-fold cross-validation
B. Leave-one-out cross-validation
C. Holdout method
D. All of the above

Correct Answer: C
Section:
Explanation:
Holdout cross-validation method is suitable for very large dataset because it is the simplest and quicker to compute version of cross-validation.
Holdout method
In this method, the dataset is divided into two sets namely the training and the test set with the basic property that the training set is bigger than the test set. Later, the model is trained on the training dataset and evaluated using the test dataset.

QUESTION 49
Which of the following is a common evaluation metric for binary classification?

A. Accuracy
B. F1 score
C. Mean squared error (MSE)
D. Area under the ROC curve (AUC)

Correct Answer: D
Section:
Explanation:
The area under the ROC curve (AUC) is a common evaluation metric for binary classification, which measures the performance of a classifier at different threshold values for the predicted probabilities. Other common metrics include accuracy, precision, recall, and F1 score, which are based on the confusion matrix of true positives, false positives, true negatives, and false negatives.

QUESTION 50
The most widely used metrics and tools to assess a classification model are:

A. Confusion matrix
B. Cost-sensitive accuracy
C. Area under the ROC curve
D. All of the above

Correct Answer: D
Section: