Free Questions for MLS-C01

Shared by Battle on 06-06-2022

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

An engraving company wants to automate its quality control process for plaques. The company performs the process before mailing each customized plaque to a customer. The company has created an Amazon S3 bucket that contains images of defects that should cause a plaque to be rejected. Low-confidence predictions must be sent to an internal team of reviewers who are using Amazon Augmented Al (Amazon A2I).

Which solution will meet these requirements?

Options:

A- Use Amazon Textract for automatic processing. Use Amazon A2I with Amazon Mechanical Turk for manual review.
B- Use Amazon Rekognition for automatic processing. Use Amazon A2I with a private workforce option for manual review.
C- Use Amazon Transcribe for automatic processing. Use Amazon A2I with a private workforce option for manual review.
D- Use AWS Panorama for automatic processing Use Amazon A2I with Amazon Mechanical Turk for manual review

Answer:

B

Explanation:

Amazon Rekognition is a service that provides computer vision capabilities for image and video analysis, such as object, scene, and activity detection, face and text recognition, and custom label detection. Amazon Rekognition can be used to automate the quality control process for plaques by comparing the images of the plaques with the images of defects in the Amazon S3 bucket and returning a confidence score for each defect. Amazon A2I is a service that enables human review of machine learning predictions, such as low-confidence predictions from Amazon Rekognition. Amazon A2I can be integrated with a private workforce option, which allows the engraving company to use its own internal team of reviewers to manually inspect the plaques that are flagged by Amazon Rekognition. This solution meets the requirements of automating the quality control process, sending low-confidence predictions to an internal team of reviewers, and using Amazon A2I for manual review.References:

1: Amazon Rekognition documentation

2: Amazon A2I documentation

3: Amazon Rekognition Custom Labels documentation

4: Amazon A2I Private Workforce documentation

Question 2

Question Type: MultipleChoice

An ecommerce company sends a weekly email newsletter to all of its customers. Management has hired a team of writers to create additional targeted content. A data scientist needs to identify five customer segments based on age, income, and location. The customers' current segmentation is unknown. The data scientist previously built an XGBoost model to predict the likelihood of a customer responding to an email based on age, income, and location.

Why does the XGBoost model NOT meet the current requirements, and how can this be fixed?

Options:

A- The XGBoost model provides a true/false binary output. Apply principal component analysis (PCA) with five feature dimensions to predict a segment.
B- The XGBoost model provides a true/false binary output. Increase the number of classes the XGBoost model predicts to five classes to predict a segment.
C- The XGBoost model is a supervised machine learning algorithm. Train a k-Nearest-Neighbors (kNN) model with K = 5 on the same dataset to predict a segment.
D- The XGBoost model is a supervised machine learning algorithm. Train a k-means model with K = 5 on the same dataset to predict a segment.

Answer:

D

Explanation:

The XGBoost model is a supervised machine learning algorithm, which means it requires labeled data to learn from. The customers' current segmentation is unknown, so there is no label to train the XGBoost model on. Moreover, the XGBoost model is designed for classification or regression tasks, not for clustering. Clustering is a type of unsupervised machine learning, which means it does not require labeled data. Clustering algorithms try to find natural groups or clusters in the data based on their similarity or distance. A common clustering algorithm is k-means, which partitions the data into K clusters, where each data point belongs to the cluster with the nearest mean. To meet the current requirements, the data scientist should train a k-means model with K

= 5 on the same dataset to predict a segment for each customer. This way, the data scientist can identify five customer segments based on age, income, and location, without needing any labels.References:

What is XGBoost? - Amazon SageMaker

What is Clustering? - Amazon SageMaker

K-Means Algorithm - Amazon SageMaker

Question 3

Question Type: MultipleChoice

A company is building a new version of a recommendation engine. Machine learning (ML) specialists need to keep adding new data from users to improve personalized recommendations. The ML specialists gather data from the users' interactions on the platform and from sources such as external websites and social media.

The pipeline cleans, transforms, enriches, and compresses terabytes of data daily, and this data is stored in Amazon S3. A set of Python scripts was coded to do the job and is stored in a large Amazon EC2 instance. The whole process takes more than 20 hours to finish, with each script taking at least an hour. The company wants to move the scripts out of Amazon EC2 into a more managed solution that will eliminate the need to maintain servers.

Which approach will address all of these requirements with the LEAST development effort?

Options:

A- Load the data into an Amazon Redshift cluster. Execute the pipeline by using SQL. Store the results in Amazon S3.
B- Load the data into Amazon DynamoDB. Convert the scripts to an AWS Lambda function. Execute the pipeline by triggering Lambda executions. Store the results in Amazon S3.
C- Create an AWS Glue job. Convert the scripts to PySpark. Execute the pipeline. Store the results in Amazon S3.
D- Create a set of individual AWS Lambda functions to execute each of the scripts. Build a step function by using the AWS Step Functions Data Science SDK. Store the results in Amazon S3.

Answer:

C

Explanation:

The best approach to address all of the requirements with the least development effort is to create an AWS Glue job, convert the scripts to PySpark, execute the pipeline, and store the results in Amazon S3. This is because:

AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it easy to prepare and load data for analytics1.AWS Glue can run Python and Scala scripts to process data from various sources, such as Amazon S3, Amazon DynamoDB, Amazon Redshift, and more2.AWS Glue also provides a serverless Apache Spark environment to run ETL jobs, eliminating the need to provision and manage servers3.

PySpark is the Python API for Apache Spark, a unified analytics engine for large-scale data processing4.PySpark can perform various data transformations and manipulations on structured and unstructured data, such as cleaning, enriching, and compressing5.PySpark can also leverage the distributed computing power of Spark to handle terabytes of data efficiently and scalably6.

By creating an AWS Glue job and converting the scripts to PySpark, the company can move the scripts out of Amazon EC2 into a more managed solution that will eliminate the need to maintain servers.The company can also reduce the development effort by using the AWS Glue console, AWS SDK, or AWS CLI to create and run the job7.Moreover, the company can use the AWS Glue Data Catalog to store and manage the metadata of the data sources and targets8.

The other options are not as suitable as option C for the following reasons:

Option A is not optimal because loading the data into an Amazon Redshift cluster and executing the pipeline by using SQL will incur additional costs and complexity for the company. Amazon Redshift is a fully managed data warehouse service that enables fast and scalable analysis of structured data . However, it is not designed for ETL purposes, such as cleaning, transforming, enriching, and compressing data. Moreover, using SQL to perform these tasks may not be as expressive and flexible as using Python scripts. Furthermore, the company will have to provision and configure the Amazon Redshift cluster, and load and unload the data from Amazon S3, which will increase the development effort and time.

Option B is not feasible because loading the data into Amazon DynamoDB and converting the scripts to an AWS Lambda function will not work for the company's use case. Amazon DynamoDB is a fully managed key-value and document database service that provides fast and consistent performance at any scale . However, it is not suitable for storing and processing terabytes of data daily, as it has limits on the size and throughput of each table and item . Moreover, using AWS Lambda to execute the pipeline will not be efficient or cost-effective, as Lambda has limits on the memory, CPU, and execution time of each function . Therefore, using Amazon DynamoDB and AWS Lambda will not meet the company's requirements for processing large amounts of data quickly and reliably.

Option D is not relevant because creating a set of individual AWS Lambda functions to execute each of the scripts and building a step function by using the AWS Step Functions Data Science SDK will not address the main issue of moving the scripts out of Amazon EC2. AWS Step

Functions is a fully managed service that lets you coordinate multiple AWS services into serverless workflows . The AWS Step Functions Data Science SDK is an open source library that allows data scientists to easily create workflows that process and publish machine learning models using Amazon SageMaker and AWS Step Functions . However, these services and tools are not designed for ETL purposes, such as cleaning, transforming, enriching, and compressing data. Moreover, as mentioned in option B, using AWS Lambda to execute the scripts will not be efficient or cost-effective for the company's use case.

References:

What Is AWS Glue?

AWS Glue Components

AWS Glue Serverless Spark ETL

PySpark - Overview

PySpark - RDD

PySpark - SparkContext

Adding Jobs in AWS Glue

Populating the AWS Glue Data Catalog

[What Is Amazon Redshift?]

[What Is Amazon DynamoDB?]

[Service, Account, and Table Quotas in DynamoDB]

[AWS Lambda quotas]

[What Is AWS Step Functions?]

[AWS Step Functions Data Science SDK for Python]

Question 4

Question Type: MultipleChoice

A company wants to predict stock market price trends. The company stores stock market data each business day in Amazon S3 in Apache Parquet format. The company stores 20 GB of data each day for each stock code.

A data engineer must use Apache Spark to perform batch preprocessing data transformations quickly so the company can complete prediction jobs before the stock market opens the next

day. The company plans to track more stock market codes and needs a way to scale the preprocessing data transformations.

Which AWS service or feature will meet these requirements with the LEAST development effort over time?

Options:

A- AWS Glue jobs
B- Amazon EMR cluster
C- Amazon Athena
D- AWS Lambda

Answer:

A

Explanation:

AWS Glue jobs is the AWS service or feature that will meet the requirements with the least development effort over time. AWS Glue jobs is a fully managed service that enables data engineers to run Apache Spark applications on a serverless Spark environment. AWS Glue jobs can perform batch preprocessing data transformations on large datasets stored in Amazon S3, such as converting data formats, filtering data, joining data, and aggregating dat

a. AWS Glue jobs can also scale the Spark environment automatically based on the data volume and processing needs, without requiring any infrastructure provisioning or management. AWS Glue jobs can reduce the development effort and time by providing a graphical interface to create and monitor Spark applications, as well as a code generation feature that can generate Scala or Python code based on the data sources and targets.AWS Glue jobs can also integrate with other AWS services, such as Amazon Athena, Amazon EMR, and Amazon SageMaker, to enable further data analysis and machine learning tasks1.

The other options are either more complex or less scalable than AWS Glue jobs. Amazon EMR cluster is a managed service that enables data engineers to run Apache Spark applications on a cluster of Amazon EC2 instances. However, Amazon EMR cluster requires more development effort and time than AWS Glue jobs, as it involves setting up, configuring, and managing the cluster, as well as writing and deploying the Spark code.Amazon EMR cluster also does not scale automatically, but requires manual or scheduled resizing of the cluster based on the data volume and processing needs2. Amazon Athena is a serverless interactive query service that enables data engineers to analyze data stored in Amazon S3 using standard SQL. However, Amazon Athena is not suitable for performing complex data transformations, such as joining data from multiple sources, aggregating data, or applying custom logic.Amazon Athena is also not designed for running Spark applications, but only supports SQL queries3. AWS Lambda is a serverless

compute service that enables data engineers to run code without provisioning or managing servers. However, AWS Lambda is not optimized for running Spark applications, as it has limitations on the execution time, memory size, and concurrency of the functions. AWS Lambda is also not integrated with Amazon S3, and requires additional steps to read and write data from S3 buckets.

References:

1: AWS Glue - Fully Managed ETL Service - Amazon Web Services

2: Amazon EMR - Amazon Web Services

3: Amazon Athena -- Interactive SQL Queries for Data in Amazon S3

[4]: AWS Lambda -- Serverless Compute - Amazon Web Services

Question 5

Question Type: MultipleChoice

A global bank requires a solution to predict whether customers will leave the bank and choose another bank. The bank is using a dataset to train a model to predict customer loss. The training dataset has 1,000 rows. The training dataset includes 100 instances of customers who left the bank.

A machine learning (ML) specialist is using Amazon SageMaker Data Wrangler to train a churn prediction model by using a SageMaker training job. After training, the ML specialist notices that the model returns only false results. The ML specialist must correct the model so that it returns more accurate predictions.

Which solution will meet these requirements?

Options:

A- Apply anomaly detection to remove outliers from the training dataset before training.
B- Apply Synthetic Minority Oversampling Technique (SMOTE) to the training dataset before training.
C- Apply normalization to the features of the training dataset before training.
D- Apply undersampling to the training dataset before training.

Answer:

B

Explanation:

The best solution to meet the requirements is to apply Synthetic Minority Oversampling Technique (SMOTE) to the training dataset before training. SMOTE is a technique that generates synthetic samples for the minority class by interpolating between existing samples. This can help balance the class distribution and provide more information to the model. SMOTE can improve the performance of the model on the minority class, which is the class of interest in churn prediction. SMOTE can be applied using the SageMaker Data Wrangler, which provides a built-in analysis for oversampling the minority class1.

The other options are not effective solutions for the problem. Applying anomaly detection to remove outliers from the training dataset before training may not improve the model's accuracy, as outliers may not be the main cause of the false results. Moreover, removing outliers may reduce the diversity of the data and make the model less robust. Applying normalization to the features of the training dataset before training may improve the model's convergence and stability, but it does not address the class imbalance issue. Normalization can also be applied using the SageMaker Data Wrangler, which provides a built-in transformation for scaling the features2. Applying undersampling to the training dataset before training may reduce the class imbalance, but it also discards potentially useful information from the majority class. Undersampling can also result in underfitting and high bias for the model.

References:

* Analyze and Visualize

* Transform and Export

* SMOTE for Imbalanced Classification with Python

* Churn prediction using Amazon SageMaker built-in tabular algorithms LightGBM, CatBoost, TabTransformer, and AutoGluon-Tabular

Question 6

Question Type: MultipleChoice

A retail company wants to update its customer support system. The company wants to implement automatic routing of customer claims to different queues to prioritize the claims by category.

Currently, an operator manually performs the category assignment and routing. After the operator classifies and routes the claim, the company stores the claim's record in a central database. The claim's record includes the claim's category.

The company has no data science team or experience in the field of machine learning (ML). The

company's small development team needs a solution that requires no ML expertise.

Which solution meets these requirements?

Options:

A- Export the database to a .csv file with two columns: claim_label and claim_text. Use the Amazon SageMaker Object2Vec algorithm and the .csv file to train a model. Use SageMaker to deploy the model to an inference endpoint. Develop a service in the application to use the inference endpoint to process incoming claims, predict the labels, and route the claims to the appropriate queue.
B- Export the database to a .csv file with one column: claim_text. Use the Amazon SageMaker Latent Dirichlet Allocation (LDA) algorithm and the .csv file to train a model. Use the LDA algorithm to detect labels automatically. Use SageMaker to deploy the model to an inference endpoint. Develop a service in the application to use the inference endpoint to process incoming claims, predict the labels, and route the claims to the appropriate queue.
C- Use Amazon Textract to process the database and automatically detect two columns: claim_label and claim_text. Use Amazon Comprehend custom classification and the extracted information to train the custom classifier. Develop a service in the application to use the Amazon Comprehend API to process incoming claims, predict the labels, and route the claims to the appropriate queue.
D- Export the database to a .csv file with two columns: claim_label and claim_text. Use Amazon Comprehend custom classification and the .csv file to train the custom classifier. Develop a service in the application to use the Amazon Comprehend API to process incoming claims, predict the labels, and route the claims to the appropriate queue.

Answer:

D

Explanation:

Amazon Comprehend is a natural language processing (NLP) service that can analyze text and extract insights such as sentiment, entities, topics, and language. Amazon Comprehend also provides custom classification and custom entity recognition features that allow users to train their own models using their own data and labels. For the scenario of routing customer claims to different queues based on categories, Amazon Comprehend custom classification is a suitable solution. The custom classifier can be trained using a .csv file that contains the claim text and the claim label as columns. The custom classifier can then be used to process incoming claims and predict the labels using the Amazon Comprehend API. The predicted labels can be used to route the claims to the appropriate queue. This solution does not require any machine learning expertise or model deployment, and it can be easily integrated with the existing application.

The other options are not suitable because:

Option A: Amazon SageMaker Object2Vec is an algorithm that can learn embeddings of objects such as words, sentences, or documents. It can be used for tasks such as text classification, sentiment analysis, or recommendation systems. However, using this algorithm requires machine learning expertise and model deployment using SageMaker, which are not available for the company.

Option B: Amazon SageMaker Latent Dirichlet Allocation (LDA) is an algorithm that can discover the topics or themes in a collection of documents. It can be used for tasks such as topic modeling, document clustering, or text summarization. However, using this algorithm requires machine learning expertise and model deployment using SageMaker, which are not available for the company. Moreover, LDA does not provide labels for the topics, but rather a distribution of words for each topic, which may not match the existing categories of the claims.

Option C: Amazon Textract is a service that can extract text and data from scanned documents or images. It can be used for tasks such as document analysis, data extraction, or form processing. However, using this service is unnecessary and inefficient for the scenario, since the company already has the claim text and label in a database. Moreover, Amazon Textract does not provide custom classification features, so it cannot be used to train a custom classifier using the existing data and labels.

References:

Amazon Comprehend Custom Classification

Amazon SageMaker Object2Vec

Amazon SageMaker Latent Dirichlet Allocation

Amazon Textract

Question 7

Question Type: MultipleChoice

A manufacturing company wants to use machine learning (ML) to automate quality control in its facilities. The facilities are in remote locations and have limited internet connectivity. The company has 20 of training data that consists of labeled images of defective product parts. The training data is in the corporate on-premises data center.

The company will use this data to train a model for real-time defect detection in new parts as the parts move on a conveyor belt in the facilities. The company needs a solution that minimizes costs for compute infrastructure and that maximizes the scalability of resources for training. The solution also must facilitate the company's use of an ML model in the low-connectivity environments.

Which solution will meet these requirements?

Options:

A- Move the training data to an Amazon S3 bucket. Train and evaluate the model by using Amazon SageMaker. Optimize the model by using SageMaker Neo. Deploy the model on a SageMaker hosting services endpoint.
B- Train and evaluate the model on premises. Upload the model to an Amazon S3 bucket. Deploy the model on an Amazon SageMaker hosting services endpoint.
C- Move the training data to an Amazon S3 bucket. Train and evaluate the model by using Amazon SageMaker. Optimize the model by using SageMaker Neo. Set up an edge device in the manufacturing facilities with AWS IoT Greengrass. Deploy the model on the edge device.
D- Train the model on premises. Upload the model to an Amazon S3 bucket. Set up an edge device in the manufacturing facilities with AWS IoT Greengrass. Deploy the model on the edge device.

Answer:

C

Explanation:

The solution C meets the requirements because it minimizes costs for compute infrastructure, maximizes the scalability of resources for training, and facilitates the use of an ML model in low-connectivity environments. The solution C involves the following steps:

Move the training data to an Amazon S3 bucket. This will enable the company to store the large amount of data in a durable, scalable, and cost-effective way.It will also allow the company to access the data from the cloud for training and evaluation purposes1.

Train and evaluate the model by using Amazon SageMaker. This will enable the company to use a fully managed service that provides various features and tools for building, training, tuning, and deploying ML models.Amazon SageMaker can handle large-scale data processing and distributed training, and it can leverage the power of AWS compute resources such as Amazon EC2, Amazon EKS, and AWS Fargate2.

Optimize the model by using SageMaker Neo. This will enable the company to reduce the size of the model and improve its performance and efficiency.SageMaker Neo can compile the model into an executable that can run on various hardware platforms, such as CPUs, GPUs, and edge devices3.

Set up an edge device in the manufacturing facilities with AWS IoT Greengrass. This will enable the company to deploy the model on a local device that can run inference in real time, even in low-connectivity environments.AWS IoT Greengrass can extend AWS cloud capabilities to the edge, and it can securely communicate with the cloud for updates and synchronization4.

Deploy the model on the edge device. This will enable the company to automate quality control in its facilities by using the model to detect defects in new parts as they move on a conveyor belt.The model can run inference locally on the edge device without requiring internet connectivity, and it can send the results to the cloud when the connection is available4.

The other options are not suitable because:

Option A: Deploying the model on a SageMaker hosting services endpoint will not facilitate the use of the model in low-connectivity environments, as it will require internet access to perform inference. Moreover, it may incur higher costs for hosting and data transfer than deploying the model on an edge device.

Option B: Training and evaluating the model on premises will not minimize costs for compute infrastructure, as it will require the company to maintain and upgrade its own hardware and software. Moreover, it will not maximize the scalability of resources for training, as it will limit the company's ability to leverage the cloud's elasticity and flexibility.

Option D: Training the model on premises will not minimize costs for compute infrastructure, nor maximize the scalability of resources for training, for the same reasons as option B.

References:

1: Amazon S3

2: Amazon SageMaker

3: SageMaker Neo

4: AWS IoT Greengrass

Question 8

Question Type: MultipleChoice

A retail company uses a machine learning (ML) model for daily sales forecasting. The company's brand manager reports that the model has provided inaccurate results for the past 3 weeks.

At the end of each day, an AWS Glue job consolidates the input data that is used for the forecasting with the actual daily sales data and the predictions of the model. The AWS Glue job stores the data in Amazon S3. The company's ML team is using an Amazon SageMaker Studio notebook to gain an understanding about the source of the model's inaccuracies.

What should the ML team do on the SageMaker Studio notebook to visualize the model's degradation MOST accurately?

Options:

A- Create a histogram of the daily sales over the last 3 weeks. In addition, create a histogram of the daily sales from before that period.
B- Create a histogram of the model errors over the last 3 weeks. In addition, create a histogram of the model errors from before that period.
C- Create a line chart with the weekly mean absolute error (MAE) of the model.
D- Create a scatter plot of daily sales versus model error for the last 3 weeks. In addition, create a scatter plot of daily sales versus model error from before that period.

Answer:

B

Explanation:

The best way to visualize the model's degradation is to create a histogram of the model errors over the last 3 weeks and compare it with a histogram of the model errors from before that period. A histogram is a graphical representation of the distribution of numerical data. It shows how often each value or range of values occurs in the data. A model error is the difference between the actual value and the predicted value. A high model error indicates a poor fit of the model to the data. By comparing the histograms of the model errors, the ML team can see if there is a significant change in the shape, spread, or center of the distribution. This can indicate if the model is underfitting, overfitting, or drifting from the data. A line chart or a scatter plot would not be as effective as a histogram for this purpose, because they do not show the distribution of the errors. A line chart would only show the trend of the errors over time, which may not capture the variability or outliers. A scatter plot would only show the relationship between the errors and another variable, such as daily sales, which may not be relevant or informative for the model's performance.References:

Histogram - Wikipedia

Model error - Wikipedia

SageMaker Model Monitor - visualizing monitoring results

Question 9

Question Type: MultipleChoice

A pharmaceutical company performs periodic audits of clinical trial sites to quickly resolve critical findings. The company stores audit documents in text format. Auditors have requested help from a data science team to quickly analyze the documents. The auditors need to discover the 10

main topics within the documents to prioritize and distribute the review work among the auditing team members. Documents that describe adverse events must receive the highest priority.

A data scientist will use statistical modeling to discover abstract topics and to provide a list of the top words for each category to help the auditors assess the relevance of the topic.

Which algorithms are best suited to this scenario? (Choose two.)

Options:

A- Latent Dirichlet allocation (LDA)
B- Random Forest classifier
C- Neural topic modeling (NTM)
D- Linear support vector machine
E- Linear regression

Answer:

A, C

Explanation:

The algorithms that are best suited to this scenario are latent Dirichlet allocation (LDA) and neural topic modeling (NTM), as they are both unsupervised learning methods that can discover abstract topics from a collection of text documents.LDA and NTM can provide a list of the top words for each topic, as well as the topic distribution for each document, which can help the auditors assess the relevance and priority of the topic12.

The other options are not suitable because:

Option B: A random forest classifier is a supervised learning method that can perform classification or regression tasks by using an ensemble of decision trees.A random forest classifier is not suitable for discovering abstract topics from text documents, as it requires labeled data and predefined classes3.

Option D: A linear support vector machine is a supervised learning method that can perform classification or regression tasks by using a linear function that separates the data into different classes.A linear support vector machine is not suitable for discovering abstract topics from text documents, as it requires labeled data and predefined classes4.

Option E: A linear regression is a supervised learning method that can perform regression tasks by using a linear function that models the relationship between a dependent variable and one or more independent variables.A linear regression is not suitable for discovering abstract topics from text documents, as it requires labeled data and a continuous output variable5.

References:

1: Latent Dirichlet Allocation

2: Neural Topic Modeling

3: Random Forest Classifier

4: Linear Support Vector Machine

5: Linear Regression

Question 10

Question Type: MultipleChoice

A machine learning specialist is developing a regression model to predict rental rates from rental listings. A variable named Wall_Color represents the most prominent exterior wall color of the property. The following is the sample data, excluding all other variables:

Property_ID	Wall_Color
1000	Red
1001	White
1002	Green

The specialist chose a model that needs numerical input data.

Which feature engineering approaches should the specialist use to allow the regression model to learn from the Wall_Color data? (Choose two.)

Options:

A- Apply integer transformation and set Red = 1, White = 5, and Green = 10.
B- Add new columns that store one-hot representation of colors.
C- Replace the color name string by its length.
D- Create three columns to encode the color in RGB format.
E- Replace each color name by its training set frequency.

Answer:

B, D

Explanation:

In this scenario, the specialist should use one-hot encoding and RGB encoding to allow the regression model to learn from the Wall_Color data. One-hot encoding is a technique used to convert categorical data into numerical data. It creates new columns that store one-hot representation of colors. For example, a variable named color has three categories: red, green, and blue. After one-hot encoding, the new variables should be like this:

color_red	color_green	color_blue
1	0	0
0	1	0
0	0	1

One-hot encoding can capture the presence or absence of a color, but it cannot capture the intensity or hue of a color. RGB encoding is a technique used to represent colors in a digital image. It creates three columns to encode the color in RGB format. For example, a variable named color has three categories: red, green, and blue. After RGB encoding, the new variables should be like this:

color_R	color_G	color_B
255	0	0
0	255	0
0	0	255

RGB encoding can capture the intensity and hue of a color, but it may also introduce correlation among the three columns. Therefore, using both one-hot encoding and RGB encoding can provide more information to the regression model than using either one alone.

References:

Feature Engineering for Categorical Data

How to Perform Feature Selection with Categorical Data

Question 11

Question Type: MultipleChoice

A power company wants to forecast future energy consumption for its customers in residential properties and commercial business properties. Historical power consumption data for the last 10 years is available. A team of data scientists who performed the initial data analysis and feature

selection will include the historical power consumption data and data such as weather, number of individuals on the property, and public holidays.

The data scientists are using Amazon Forecast to generate the forecasts.

Which algorithm in Forecast should the data scientists use to meet these requirements?

Options:

A- Autoregressive Integrated Moving Average (AIRMA)
B- Exponential Smoothing (ETS)
C- Convolutional Neural Network - Quantile Regression (CNN-QR)
D- Prophet

Answer:

C

Explanation:

CNN-QR is a proprietary machine learning algorithm for forecasting time series using causal convolutional neural networks (CNNs). CNN-QR works best with large datasets containing hundreds of time series. It accepts item metadata, and is the only Forecast algorithm that accepts related time series data without future values. In this case, the power company has historical power consumption data for the last 10 years, which is a large dataset with multiple time series. The data also includes related data such as weather, number of individuals on the property, and public holidays, which can be used as item metadata or related time series data. Therefore, CNN-QR is the most suitable algorithm for this scenario.References:Amazon Forecast Algorithms,Amazon Forecast CNN-QR