AWS Certified AI Practitioner Study Guide

Quiz

Instructions: Answer the following questions in 2-3 sentences each, demonstrating your understanding of the concepts.

1. Explain the key difference between Traditional Programming and Artificial Intelligence (AI) in terms of their operational approach.
2. Describe the concept of "overfitting" in machine learning models and suggest one method to mitigate it.
3. You are tasked with building a system that groups similar customer reviews without any predefined categories. Which type of machine learning would you use and why?
4. What is the purpose of Feature Engineering in the Machine Learning Pipeline, and provide one example of a technique used in it.
5. Explain the difference between model parameters and hyperparameters during model training. Provide an example for each.
6. A company wants to evaluate a classification model used for detecting fraudulent transactions, where it is critical to identify as many actual fraudulent transactions as possible, even if it means flagging some legitimate ones. Which metric should they prioritize and why?
7. What is "hallucination" in the context of Generative AI, and how can Retrieval Augmented Generation (RAG) help address this issue?
8. You are designing a chatbot that needs to respond quickly to user queries. Which two model design considerations would be most crucial for this requirement?
9. Explain the concept of "sampling bias" and how it can impact the fairness of an AI model.
10. Describe the primary function of Amazon Bedrock Guardrails in a Generative AI application.

Answer Key

1. **Traditional Programming vs. AI:** Traditional programming relies on explicit, rule-based logic and is deterministic, meaning the output is predictable given the input. AI, conversely, is data-driven, adaptable, and learns patterns from data to make decisions or predictions without being explicitly programmed for every scenario, making its internal logic often less transparent.
2. **Overfitting:** Overfitting occurs when a model learns the training data too well, capturing noise and specific patterns that do not generalize to new, unseen data. This leads to high performance on the training set but poor performance on new data. To mitigate overfitting, one can use more diverse data, increase the dataset size through data augmentation, or select fewer, more important features.
3. **Unsupervised Learning for Customer Reviews:** I would use Unsupervised Learning, specifically clustering, for this task. Since there are no predefined categories or labels for the customer reviews, clustering algorithms can automatically identify inherent groupings or themes within the text data based on similarities in word usage and sentiment.
4. **Feature Engineering:** Feature Engineering is the process of using domain knowledge to create, select, and transform features from raw data that enhance the performance of ML

algorithms. An example technique is Categorical Encoding, where categorical (text) values are converted into numerical representations that the model can process.

5. **Parameters vs. Hyperparameters:** Parameters are values learned automatically by the model during training, such as weights and biases, which directly influence the model's predictions. Hyperparameters, on the other hand, are external configurations set by the user *before* training, controlling the behavior and learning process of the algorithm, like batch size or learning rate.
6. **Metric for Fraud Detection:** The company should prioritize **Recall (True Positive Rate)**. Recall measures the proportion of actual positive cases (fraudulent transactions) that were correctly identified by the model. In fraud detection, missing a fraudulent transaction (a false negative) can be very costly, making high recall crucial to minimize these missed cases.
7. **Hallucination in Generative AI and RAG:** Hallucination refers to a generative AI model producing plausible-sounding but factually incorrect or fabricated information that is not grounded in its training data or the provided input. Retrieval Augmented Generation (RAG) helps mitigate this by retrieving relevant and up-to-date information from an external knowledge base and incorporating it into the model's response, thereby grounding the output in accurate data.
8. **Chatbot Design Considerations:** For a chatbot requiring quick responses, **Latency** and **Model Size/Complexity** would be crucial design considerations. Lower latency ensures faster response times, while selecting a model that is not overly complex for the task can reduce computational overhead and improve inference speed.
9. **Sampling Bias:** Sampling bias occurs when the training data used for an AI model is not truly representative of the entire population the model is intended to serve. This can lead to the model learning unfair patterns or making skewed predictions for underrepresented groups, ultimately impacting the fairness and equitable performance of the AI system.
10. **Amazon Bedrock Guardrails:** Amazon Bedrock Guardrails serve as a safety and content moderation layer for foundation models. Their primary function is to enforce policies and block against prompt attacks and unwanted, inappropriate, or harmful outputs, ensuring that the AI-generated content remains safe, relevant, and compliant with organizational guidelines.

Essay Format Questions (No Answers Provided)

1. Compare and contrast the key characteristics, use cases, and underlying methodologies of Supervised Learning, Unsupervised Learning, and Reinforcement Learning. Discuss how an AWS solution architect would choose between these paradigms for different business problems.
2. Describe the end-to-end Machine Learning Pipeline, detailing each major step from identifying a business goal to monitoring a deployed model. For each step, suggest at least one relevant AWS service that can be utilized.
3. Discuss the critical considerations for building responsible AI systems, including fairness, explainability, robustness, and controllability. How do AWS services like SageMaker Clarify and Bedrock Guardrails contribute to addressing these principles?
4. Explain the concept of "fine-tuning" a foundation model, including its purpose, the types of data required, and its advantages over training a model from scratch. Furthermore, describe different fine-tuning methods like domain adaptation and instruction tuning.

5. Analyze the security and governance requirements for deploying a generative AI application handling sensitive customer data on AWS. Discuss how services like IAM, KMS, CloudTrail, AWS Config, and Bedrock Guardrails contribute to ensuring compliance and mitigating risks such as prompt injection and data exposure.

Glossary of Key Terms

- **Accuracy:** The proportion of correct predictions (both true positives and true negatives) out of all predictions; used when positive and negative predictions are equally important.
- **Anomaly Detection:** A type of unsupervised learning that identifies rare items or events that deviate significantly from the majority of the data.
- **Asynchronous Inference:** A SageMaker deployment option for long-running inference requests with large payloads, processed without requiring immediate responses.
- **AUC-ROC (Area Under the Receiver Operating Characteristic Curve):** A metric that represents the overall performance of a classification model across all possible classification thresholds; often visualized as a graph of the True Positive Rate vs. False Positive Rate.
- **AWS Augmented AI (A2I):** An AWS service that enables human review of machine learning predictions, either randomly or based on confidence scores, to ensure accuracy.
- **AWS Bedrock Guardrails:** A feature in Amazon Bedrock that allows developers to implement safety policies, block against prompt attacks, and filter unwanted or harmful outputs from foundation models.
- **AWS CloudTrail:** An AWS service that logs API calls and user activity across AWS services, providing a historical record for auditing, security analysis, and compliance.
- **AWS Config:** An AWS service that provides continuous monitoring and recording of AWS resource configurations, helping with compliance auditing, security analysis, and change management.
- **AWS Data Exchange:** An AWS service that allows users to find, subscribe to, and securely use third-party data products in the AWS Cloud.
- **AWS Glue:** A fully managed extract, transform, and load (ETL) service that makes it easy to prepare and load data for analytics and machine learning.
- **AWS Glue DataBrew:** A visual data preparation tool that helps data analysts and scientists clean and normalize data using an interactive, no-code interface.
- **AWS IAM (Identity and Access Management):** An AWS service that enables you to securely manage access to AWS services and resources, controlling who is authenticated and authorized to use resources.
- **AWS Inspector:** An automated security assessment service that helps improve the security and compliance of applications deployed on AWS by identifying vulnerabilities.
- **AWS KMS (Key Management Service):** An AWS service that makes it easy to create and manage cryptographic keys and control their use across a wide range of AWS services and in your applications.
- **AWS PrivateLink:** An AWS networking service that enables private connectivity between VPCs and AWS services without exposing data to the public internet.
- **AWS Rekognition:** An AWS service that provides image and video analysis, including object and scene detection, facial analysis, and content moderation.

- **AWS SageMaker JumpStart:** A feature in Amazon SageMaker that provides built-in algorithms, pre-trained models, and customized solutions, allowing users to quickly deploy ML solutions.
- **AWS Shield:** An AWS service that provides distributed denial of service (DDoS) protection for applications running on AWS.
- **AWS Textract:** An AWS service that automatically extracts text, handwriting, and data from scanned documents using machine learning.
- **AWS WAF (Web Application Firewall):** An AWS service that helps protect web applications or APIs against common web exploits that may affect availability, compromise security, or consume excessive resources.
- **Balanced Fit:** A model performance state where the model performs well on both the training data and new, unseen data, indicating good generalization.
- **Batch Size:** A hyperparameter that defines the number of training examples processed at one time by the model before its weights are updated.
- **Batch Transform (Batch Inference):** A SageMaker deployment option for asynchronously processing large volumes of data for predictions.
- **BERTScore:** An evaluation metric for text generation that measures semantic similarity between generated and reference texts using contextual embeddings from pre-trained BERT models.
- **Bias (Model):** The difference between the predictive values and the actual values, representing the level of error. High bias (underfitting) means the model consistently misses the mark.
- **Bias (Data/Algorithmic):** Systemic and unfair discrimination in a model's outcomes due to unrepresentative or skewed training data, or flaws in the algorithm.
- **BLEU (Bilingual Evaluation Understudy):** An evaluation metric primarily used for machine translation, measuring the similarity between a generated text and one or more reference texts.
- **Categorical Encoding:** The process of converting categorical (textual) data into numerical representations so that machine learning algorithms can process it.
- **Chain-of-Thought Prompting:** A prompt engineering technique that asks the model to explain its reasoning process step-by-step, improving the accuracy of its output for complex tasks.
- **Chunking:** The process of breaking larger datasets or documents into smaller, more manageable pieces, often used to fit within a model's context window.
- **Classification:** A type of supervised learning task where the model predicts a discrete category or class for a given input.
- **Clustering:** A type of unsupervised learning task that groups similar data points together based on their inherent characteristics.
- **Confirmation Bias:** A type of bias where a model or person focuses on data that confirms existing beliefs while ignoring contradictory data.
- **Confusion Matrix:** A table used to evaluate the performance of a classification model, summarizing the number of true positives, true negatives, false positives, and false negatives.
- **Context Window:** The maximum number of tokens a Large Language Model (LLM) can process as input and generate as output in a single interaction.

- **Continuous Pre-Training:** The process of continually updating a pre-trained model with new, unlabeled data to expand its general knowledge base without specializing it for a specific task.
- **Controllability:** The ability to guide and manage an AI system's behavior to ensure it operates within defined boundaries and achieves desired outcomes.
- **Correlation Matrix:** A table that displays the correlation coefficients between many variables, used in Exploratory Data Analysis (EDA) to identify relationships.
- **Customer Managed Key (CMK):** An encryption key where the user has full control over its management, including creation, rotation, and access policies.
- **Data Augmentation:** Techniques used to increase the amount of data by adding slightly modified copies of existing data or newly created synthetic data from existing data.
- **Data Drift:** A phenomenon where the statistical properties of the input data to a model change over time, causing model performance to degrade.
- **Data Lineage:** The lifecycle of data, tracking its origin, transformations, and movement over time, crucial for governance and reproducibility.
- **Data Poisoning:** A security risk where false or malicious data is injected into an AI model's training dataset to compromise its integrity or performance.
- **Deep Learning (DL):** A subset of machine learning that uses multi-layered neural networks (deep neural networks) to learn complex patterns from large amounts of data.
- **Dimensionality Reduction:** The process of reducing the number of input variables (features) in a dataset, often by combining them, to simplify the model and prevent overfitting.
- **Diffusion Models:** A type of generative model that works by progressively adding noise to data (forward process) and then learning to reverse that process to generate new data (reverse process).
- **Domain Adaptation Fine-tuning:** A fine-tuning method that customizes a pre-trained foundation model for a specific field (e.g., medical, legal) using a small amount of domain-specific data.
- **Embeddings:** Numerical, vectorized representations of values or objects (like text, images, or audio) that capture their semantic meaning in a high-dimensional space, useful for tasks like similarity search.
- **Epochs:** One complete pass of the entire training dataset through the learning algorithm, during which the model's weights are updated.
- **Ethical AI:** AI systems designed and operated in a manner that adheres to moral principles, respects human rights, and promotes societal well-being.
- **Explainability:** The ability to understand and interpret how an AI model arrived at a particular decision or prediction, making its logic transparent to humans.
- **Exploratory Data Analysis (EDA):** The process of analyzing data sets to summarize their main characteristics, often with visual methods, before model training.
- **F1 Score:** The harmonic mean of precision and recall, providing a balanced measure of a model's accuracy, especially useful in cases of class imbalance.
- **False Negative (FN):** An outcome where the model incorrectly predicts a negative class when the actual class is positive (missed prediction).
- **False Positive (FP):** An outcome where the model incorrectly predicts a positive class when the actual class is negative (false alarm).

- **False Positive Rate (FPR):** The ratio of false positives to the total number of actual negative instances; measures how often incorrect positive predictions are made.
- **Fairness and Bias Mitigation:** The principle of ensuring that AI systems produce equitable outcomes across different demographic groups and that biases present in training data are identified and reduced.
- **Feature Engineering:** The process of using domain knowledge to create new features or transform existing ones from raw data to improve the performance of machine learning algorithms.
- **Feature Extraction:** The process of creating new columns or features from existing data to enhance the information available for model training.
- **Feature Store:** A centralized repository for storing, managing, and sharing machine learning features, ensuring consistency and reusability across models.
- **Few-Shot Prompting:** A prompt engineering technique where a few examples are provided in the prompt to guide the model's behavior and improve its output.
- **Fine-tuning:** The process of taking a pre-trained model and further training it on a smaller, labeled, domain-specific dataset to adapt it for a new task or improve its performance on a specific problem.
- **Foundation Models:** Large general-purpose pre-trained models that are expensive to create but can be adapted for a wide range of tasks and applications.
- **Generative AI:** A subset of deep learning focused on creating new content (e.g., text, images, audio) from learned data patterns.
- **Hallucination (Generative AI):** When a generative AI model produces plausible-sounding but factually incorrect, nonsensical, or fabricated information that is not grounded in reality or the provided input.
- **Hyperparameters:** Configuration variables that are external to the model and whose values are set by the user *before* the training process begins (e.g., learning rate, batch size, number of epochs).
- **In-Context Learning:** A method where a generative AI model learns a new task simply by being provided with examples within the prompt itself, without explicit weight updates.
- **Inference Parameters:** Settings that control the behavior of a generative AI model during text generation, such as temperature, top-k, and top-p.
- **Instruction-based Fine-tuning:** A fine-tuning method that trains a model to follow specific instructions or perform better on particular tasks by providing labeled examples (e.g., summarization, classification).
- **Interpretability:** The degree to which a human can understand the cause and effect of a model's decision-making process.
- **Latent Space:** An encoded, compressed, and meaningful representation of data learned by a model, where relationships between data points are captured.
- **Latency:** The time it takes for a model to produce an output after receiving an input.
- **Learning Rate:** A hyperparameter that controls how much the model's weights are adjusted with respect to the loss gradient during training.
- **Machine Learning (ML):** A subset of AI that enables systems to learn from data, identify patterns, and make predictions or decisions without being explicitly programmed.
- **Mean Absolute Error (MAE):** A metric for regression models that measures the average magnitude of errors between predicted and actual values, without considering their direction.

- **Mean Squared Error (MSE):** A metric for regression models that calculates the average of the squared differences between predicted and actual values, penalizing larger errors more heavily.
- **Measurement Bias:** A type of bias caused by inaccurate or inconsistent measurement during data collection, leading to systematic errors in the data.
- **Model Exposure:** A security risk where confidential information or sensitive details about the model's architecture or training data are inadvertently revealed.
- **Model Monitor (SageMaker Model Monitor):** A SageMaker feature that continuously monitors deployed models for data drift, concept drift, and other performance issues, sending alerts and automating corrective actions.
- **Model Parameters:** Internal variables of a model that are learned automatically from the training data (e.g., weights and biases in a neural network).
- **Multiclass Classification:** A classification task where the model predicts one of more than two possible discrete categories.
- **Multi-modal Models:** AI models that can process and generate content across multiple data types, such as images, text, and audio.
- **MLOps:** A set of practices that combines machine learning, DevOps, and data engineering to standardize and streamline the entire machine learning lifecycle, from development to deployment and monitoring.
- **Negative Prompts:** Instructions or components within a prompt that specify what the generative AI model should avoid or exclude in its output.
- **Normalization:** A feature scaling technique that rescales data to a fixed range, typically between 0 and 1.
- **Observer Bias:** A type of bias that occurs when the person collecting or interpreting data inadvertently influences the results to align with their expectations.
- **One-Shot Prompting:** A prompt engineering technique where exactly one example is provided in the prompt to guide the model's behavior for a specific task.
- **Overfitting:** A model performance issue where the model learns the training data too well, including its noise, resulting in excellent performance on training data but poor performance on new, unseen data.
- **Parameters:** See Model Parameters.
- **Perplexity:** A metric used to evaluate language models, measuring how well a model predicts a sequence of tokens. Lower perplexity indicates better predictive performance.
- **Personally Identifiable Information (PII):** Any data that can be used to identify a specific individual.
- **Precision:** A metric for classification models that measures the proportion of predicted positive instances that were actually correct (True Positives / (True Positives + False Positives)). Useful when minimizing false positives is important.
- **Prompt Engineering:** The process of designing, refining, and optimizing inputs (prompts) to get desired and effective outputs from an AI model, especially generative AI models.
- **Prompt Hijacking/Jailbreaking:** Manipulating a prompt to bypass the safety mechanisms or intended behavior of a generative AI model, often to generate unsafe or unintended outputs.
- **Prompt Leaking:** When a generative AI model inadvertently reveals too much context, internal instructions, or sensitive information from previous interactions or its training data.
- **Prompt Templates:** Pre-defined structures or formats for prompts that ensure consistency and can include placeholders for dynamic content.

- **Provisioned Throughput (Amazon Bedrock):** A pricing and deployment option in Amazon Bedrock that provides guaranteed capacity for consistent, high-throughput inference, often required for custom or fine-tuned models.
- **R^2 (R-squared):** A statistical measure in regression models that represents the proportion of the variance in the dependent variable that can be explained by the independent variables. Higher R-squared indicates a better fit.
- **Recall (True Positive Rate - TPR):** A metric for classification models that measures the proportion of actual positive instances that were correctly identified (True Positives / (True Positives + False Negatives)). Useful when minimizing false negatives is critical.
- **Regression:** A type of supervised learning task where the model predicts a continuous numerical value.
- **Reinforcement Learning (RL):** A type of machine learning where an agent learns to make decisions by interacting with an environment, receiving rewards or penalties to achieve a specific goal.
- **Reinforcement Learning from Human Feedback (RLHF):** A training technique where a model adapts its behavior based on human preferences and evaluations, aligning its outputs more closely with human values.
- **Retrieval Augmented Generation (RAG):** An architectural pattern for generative AI that enhances model outputs by retrieving relevant information from an external knowledge base or data source before generating a response, helping to reduce hallucinations.
- **RMSE (Root Mean Squared Error):** A metric for regression models that is the square root of the Mean Squared Error, providing an error measure in the same unit as the predicted values, making it more interpretable.
- **Robustness and Veracity:** Principles of Responsible AI ensuring that AI systems are resilient to challenging conditions, provide reliable outputs, and are truthful in their responses.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** An evaluation metric for text summarization and machine translation that measures the overlap (recall) between a generated text and one or more reference texts.
- **Sampling Bias:** A type of bias that occurs when the training data is not representative of the entire population, leading to the model performing poorly or unfairly on underrepresented groups.
- **SageMaker Canvas:** A visual, no-code tool within Amazon SageMaker that allows business users to build, train, and deploy machine learning models without extensive technical expertise.
- **SageMaker Clarify:** A SageMaker feature that helps detect bias in machine learning models and data, and provides explanations for model predictions to increase transparency.
- **SageMaker Data Wrangler:** A SageMaker feature that simplifies the process of data preparation and transformation for machine learning using a visual interface.
- **SageMaker Ground Truth:** A SageMaker feature that provides data labeling capabilities, allowing human annotators to label data for training machine learning models.
- **SageMaker Model Cards:** A SageMaker feature for creating documentation for trained machine learning models, detailing their purpose, risk ratings, limitations, ethical considerations, and performance metrics.

- **SageMaker Pipelines:** A SageMaker feature that provides a workflow orchestration service for building, training, and deploying machine learning models with repeatable and automated workflows.
- **Self-supervised Learning:** A type of machine learning where the model learns patterns and features from data by generating its own labels from the data itself, without explicit human-provided labels.
- **Semi-supervised Learning:** A type of machine learning that uses a mix of a small amount of labeled data and a large amount of unlabeled data for training.
- **Sentiment Analysis (Amazon Comprehend):** An NLP technique used to determine the emotional tone or opinion expressed in text, such as positive, negative, or neutral.
- **Serverless Inference:** A SageMaker deployment option for models with intermittent traffic, where the infrastructure automatically scales up and down, minimizing cost and management overhead.
- **Slots (Amazon Lex):** Variables that capture specific details or pieces of information from a user's input within a chatbot interaction.
- **Specificity (True Negative Rate - TNR):** A metric for classification models that measures the proportion of actual negative instances that were correctly identified (True Negatives / (True Negatives + False Positives)). Useful when correctly identifying negatives is important.
- **Standardization:** A feature scaling technique that rescales data to have a mean of 0 and a standard deviation of 1.
- **Structured Data:** Data that is highly organized and follows a predefined schema, typically found in tabular formats like spreadsheets or relational databases.
- **Supervised Learning:** A type of machine learning that uses labeled data (input variables and corresponding target variables/correct labels) to train a model to make predictions.
- **Temperature (Inference Parameter):** An inference parameter that controls the randomness or creativity of a generative AI model's output. Higher values lead to more diverse outputs, while lower values result in more deterministic outputs.
- **Token:** The smallest unit of text (e.g., a word, subword, or character) that a language model processes.
- **Tokenization:** The process of breaking down text into smaller units called tokens.
- **Top-k (Inference Parameter):** An inference parameter that narrows the selection of the next token to only the 'k' most likely candidates predicted by the model.
- **Top-p (Inference Parameter):** An inference parameter that narrows the selection of the next token to a cumulative percentage 'p' of the most likely candidates, balancing randomness and accuracy.
- **Traditional Programming:** A programming paradigm where developers explicitly define rules and logic for a system to follow, resulting in deterministic and transparent behavior.
- **Training Data:** The dataset used to train a machine learning model, consisting of input variables and, for supervised learning, corresponding target variables (correct labels).
- **Transfer Learning:** A machine learning technique where a model trained for one task is reused as a starting point for a model on a second, related task, speeding up training and improving performance.
- **Transparency (Responsible AI):** The principle of making AI systems' operations and decision-making processes understandable and observable to stakeholders.
- **True Negative (TN):** An outcome where the model correctly predicts a negative class when the actual class is also negative.

- **True Positive (TP):** An outcome where the model correctly predicts a positive class when the actual class is also positive.
- **Underfitting:** A model performance issue where the model fails to learn the underlying patterns in the training data, performing poorly on both training and new data (high bias).
- **Unsupervised Learning:** A type of machine learning that works with unlabeled data to discover hidden patterns, structures, or relationships within the data.
- **Unstructured Data:** Data that does not have a predefined data model or organization, such as text, images, audio, or video files.
- **Variance (Model):** The extent to which a model's predictions change when trained on different subsets of the training data. High variance (overfitting) means the model is too sensitive to the training data.
- **Vector Databases:** Specialized databases designed to store and query high-dimensional vectors (embeddings) efficiently, enabling similarity searches.
- **Vectors:** Numerical representations of embeddings that indicate an object's location in a high-dimensional space, capturing semantic relationships.
- **Zero-Shot Prompting:** A prompt engineering technique where no examples are provided in the prompt, and the model is expected to perform the task based solely on its pre-trained knowledge.