

## Applications & Use Cases Professional Certification

AG

You

Topic: Visualization

To inspect pairwise correlations between many attributes, you might look at a (Select one)

- A.  
Sankey chart
  - B.  
Scatter Matrix
  - C.  
Boxplot
  - D.  
None of these
- ChatGPT  
ChatGPT

The correct choice for inspecting pairwise correlations between many attributes is:

B. Scatter Matrix

A scatter matrix (also known as a pairs plot or pairwise plot) is a grid of scatterplots where each variable is plotted against every other variable. This visualization allows you to quickly identify relationships and correlations between multiple pairs of attributes in your dataset. Each scatterplot in the matrix represents the relationship between two variables, making it a useful tool for exploring multivariate data and identifying patterns or trends.

AG

You

Topic: Visualization

Your colleague shows you a scatterplot and explains that she added some 'jitter' to the plot. She did this because (Select one)

- A.  
she was unsure if anyone would understand the plot.
- B.  
she had data that were superimposed on (on top of) other data.
- C.

she wished to present the data using a variety of colors.

D.

she wished to show error bars on the plot.

ChatGPT

ChatGPT

The correct choice for why your colleague added 'jitter' to the scatterplot is:

B. she had data that were superimposed on (on top of) other data.

Adding 'jitter' to a scatterplot means introducing a small amount of random noise to the position of the data points along the x-axis or y-axis. This technique helps to prevent overplotting, which occurs when multiple data points have the same or very similar coordinates and would otherwise visually stack on top of each other in the plot. By adding jitter, each data point is slightly shifted from its actual position, making it easier to see the density and distribution of the data more clearly, especially when there are many overlapping points. This is particularly useful when visualizing dense datasets to avoid obscuring patterns or relationships in the data.

AG

You

Topic: Intro to ML and DS

In which one of the use cases below would you apply unsupervised learning techniques? (Select one)

A.

Predicting whether or not a customer is likely to churn in the next 90 days

B.

Grouping customers based on their gender and geographic region

C.

Forecasting the price of copper based on historical copper prices

D.

Classifying texts by genre based on a known corpus of previously-classified texts

ChatGPT

ChatGPT

The use case where you would apply unsupervised learning techniques is:

B. Grouping customers based on their gender and geographic region

Unsupervised learning techniques are used when you want to discover patterns, group similar items together, or reduce the dimensionality of data without explicit labels or predefined outcomes. In the case of grouping customers based on gender and geographic region, you would use unsupervised learning algorithms like clustering (e.g., K-means clustering) to segment customers into distinct groups based on similarities in their gender and geographic attributes. This approach helps in identifying natural groupings within the data and can be useful for targeted marketing strategies or personalized customer experiences.

Let's break down the other options:

A. Predicting whether or not a customer is likely to churn in the next 90 days - This is a supervised learning task where historical data with labeled outcomes (churn or no churn) would be used to train a predictive model.

C. Forecasting the price of copper based on historical copper prices - This is typically a time-series forecasting problem and can be addressed using supervised learning techniques such as autoregressive models or recurrent neural networks.

D. Classifying texts by genre based on a known corpus of previously-classified texts - This is also a supervised learning task where a classifier is trained on labeled text data (genres) to predict the genre of new texts based on their features.

Therefore, based on the options provided, B is the use case that aligns with applying unsupervised learning techniques.

AG

You

Topic: Visualization

To compare distributions of many attributes with the distributions of the other attributes, you might view a (Select one)

A.

Sankey chart

B.

Pie Chart

C.

Boxplot

D.

None of these

ChatGPT

ChatGPT

The correct choice for comparing distributions of many attributes with the distributions of other attributes is:

C. Boxplot

A boxplot, also known as a box-and-whisker plot, is a powerful visualization tool for comparing distributions of numerical data across different groups or attributes. It displays key statistical measures such as median, quartiles, and outliers in a concise and effective manner.

When comparing distributions of multiple attributes, you can create side-by-side boxplots for each attribute or group of interest. This allows you to visually assess the central tendency, spread, and variability of each attribute's distribution and compare them against each other easily. Boxplots are particularly useful for identifying differences in distribution shapes, ranges, and potential outliers across various attributes in your dataset.

Let's briefly discuss the other options:

A. Sankey chart - A Sankey chart is typically used to visualize flows or relationships between different categories. It is not suitable for comparing distributions of attributes.

B. Pie Chart - Pie charts are used to show proportions or percentages of a whole. They are not ideal for comparing distributions of multiple attributes.

C. Boxplot - As explained above, boxplots are well-suited for comparing distributions of numerical data across different attributes or groups.

D. None of these - While the Sankey chart and Pie Chart are not suitable for this purpose, the Boxplot is indeed a suitable choice for comparing distributions of attributes.

Therefore, the correct answer for comparing distributions of many attributes with the distributions of other attributes is C. Boxplot.

AG

You

Topic: Use Cases for ML

Which statements are true about selecting a predictive model type? (Select ALL correct answers)

Hint: Choose 3 correct answers

- A. You should pick both formulaic (e.g. regression) and non-formulaic (e.g. decision tree) models to compare.
- B. You should pick both simple and complex models to compare.
- C. Some models have data type limitations that cannot be used in all situations.
- D. You should pick a model that a Kaggle Grandmaster has used before.

ChatGPT

ChatGPT

The correct statements about selecting a predictive model type are:

- A. You should pick both formulaic (e.g., regression) and non-formulaic (e.g., decision tree) models to compare.
- B. You should pick both simple and complex models to compare.
- C. Some models have data type limitations that cannot be used in all situations.

Explanation for each correct statement:

A. It's important to consider a variety of model types when selecting a predictive model. Formulaic models like linear regression are based on mathematical formulas and assumptions, while non-formulaic models like decision trees rely on hierarchical decision rules. Comparing different types of models allows you to assess their strengths and weaknesses in relation to your specific dataset and problem.

B. Selecting both simple and complex models for comparison is advisable. Simple models (e.g., linear regression) are easier to interpret and often have lower computational requirements, while complex models (e.g., deep neural networks) can capture intricate patterns but may be harder to interpret and require more computational resources. Evaluating a range of model complexities helps in finding the right balance between model performance and complexity.

C. Different models are suited for different data types and structures. For example, decision trees work well with categorical data, while linear regression is suitable for continuous numerical data. Some models may have limitations in handling certain types of data (e.g., text data for image-based models). Understanding these data type limitations is crucial when selecting a predictive model to ensure compatibility with your dataset.

The incorrect statement is:

D. You should pick a model that a Kaggle Grandmaster has used before.

This statement is not necessarily true. While it can be insightful to learn from experienced practitioners like Kaggle Grandmasters, model selection should primarily be based on the specific requirements of your problem, the nature of your data, and empirical evaluation rather than solely relying on popularity or past usage by individuals.

AG

You

Topic: Use Cases for ML

You work for a manufacturing company and are assigned to build a model that will try to predict the number of weeks until a product stops working . You have a large historical data set of IDs, sensor data (att1-att7), and a numerical attribute representing how many weeks until the product failed (weeksToFailure):

In data science, we consider this what type of problem? (Select one)

A.

regression

B.

classification

C.

clustering

D.

text mining

ChatGPT

ChatGPT

In data science, the problem described, where you are tasked with predicting the number of weeks until a product stops working based on sensor data and historical failure times, is considered:

A. Regression

This is a regression problem because the goal is to predict a continuous numerical value (specifically, the number of weeks until failure) based on input features (sensor data attributes).

Regression models are used when the target variable (in this case, weeks until failure) is a continuous, quantitative value. The objective here is to build a predictive model that can estimate the expected number of weeks until product failure based on the sensor measurements captured over time.

Let's briefly discuss the other options:

B. Classification - Classification is used when the goal is to predict a categorical label or class based on input features. In this scenario, predicting the exact number of weeks until failure (a continuous value) does not fall under classification.

C. Clustering - Clustering is an unsupervised learning technique used to group similar data points together based on their characteristics. It is not applicable in this context where the goal is to predict a specific numerical outcome.

D. Text Mining - Text mining involves extracting useful information and patterns from text data. Since the problem description involves sensor data and numerical attributes related to product failure, it does not fall into the category of text mining.

Therefore, the correct type of problem described in this manufacturing scenario is A. Regression.

You

Topic: Intro to ML and DS

A common reason to embed scripts or code (e.g. Python or R) inside a RapidMiner process is when (Select one)

- A.  
the process is running too slow without scripts or code.
- B.  
you wish to take advantage of GPU processors on your VM or local machine.
- C.  
you want to share RapidMiner processes with other data scientists.
- D.  
you need unique functionality or libraries that is not found in RapidMiner operators.

ChatGPT

ChatGPT

The common reason to embed scripts or code (e.g., Python or R) inside a RapidMiner process is when:

- D. you need unique functionality or libraries that are not found in RapidMiner operators.

RapidMiner provides a comprehensive set of built-in operators and functionalities for data preprocessing, modeling, and analysis. However, there are scenarios where you may require specific functionalities or libraries that are not directly available within RapidMiner operators. In such cases, embedding scripts or code (e.g., Python or R scripts) allows you to leverage the capabilities of these programming languages and their respective libraries to extend the functionality of your RapidMiner processes.

By embedding scripts or code within RapidMiner processes, you can access specialized libraries for tasks such as advanced statistical analysis, machine learning algorithms, custom data transformations, or integration with external systems. This flexibility enables data scientists to combine the strengths of RapidMiner's visual workflow environment with the power and flexibility of scripting languages, enhancing the overall capabilities of data analysis and modeling workflows.

Let's briefly discuss the other options:



A. If the process is running too slow without scripts or code, this might indicate performance issues that could potentially be addressed through optimization of RapidMiner workflows or configurations, rather than solely relying on embedding scripts.

B. Taking advantage of GPU processors on your VM or local machine is typically not a direct reason for embedding scripts or code within RapidMiner processes. GPU utilization can be beneficial for certain computations but does not directly relate to embedding scripts.

C. Sharing RapidMiner processes with other data scientists can be achieved without necessarily relying on embedded scripts or code. RapidMiner processes can be shared and collaborated on within the RapidMiner platform itself or by exporting workflows in various formats.

Therefore, based on the provided options, D. you need unique functionality or libraries that are not found in RapidMiner operators, is the most appropriate reason for embedding scripts or code inside a RapidMiner process.

AG

You

Topic: Intro to ML and DS

Prior to deploying a model for a business use case, it must (Select one)

- A.  
be written in Python or R.
- B.  
be published to a public repository such as GitHub.
- C.  
be examined by a PhD data scientist.
- D.  
follow data science best practices such as ensuring no data leakage and proper model validation.

ChatGPT

ChatGPT

Prior to deploying a model for a business use case, it must:

- D. follow data science best practices such as ensuring no data leakage and proper model validation.

Explanation:

Deploying a machine learning model into a production environment for a business use case involves several critical steps to ensure the model's effectiveness, reliability, and suitability for its intended purpose. Some key considerations and best practices include:

1. **No Data Leakage:** Ensuring that the model is trained and validated without any unintentional leakage of information from the test set into the training process. Data leakage can lead to overly optimistic performance estimates and inaccurate model evaluations.
2. **Proper Model Validation:** Conducting robust validation procedures to assess the performance of the model accurately. This involves splitting the data into training, validation, and test sets, applying appropriate evaluation metrics, and tuning model hyperparameters to optimize performance.
3. **Evaluation of Business Objectives:** Validating that the model's performance aligns with the specific business objectives and requirements. This involves assessing metrics relevant to the business case, such as accuracy, precision, recall, or other domain-specific metrics.
4. **Scalability and Efficiency:** Ensuring that the model is scalable and efficient for deployment in a production environment, considering factors such as computational resources, response time, and real-time processing requirements.
5. **Monitoring and Maintenance:** Establishing procedures for ongoing monitoring and maintenance of the deployed model, including monitoring performance drift, updating the model with new data, and addressing potential issues or changes in the business environment.

Options A, B, and C do not directly address the critical steps required prior to deploying a model for a business use case. While writing the model in Python or R (Option A) is common for implementing machine learning algorithms, it is not a standalone requirement for deployment. Similarly, publishing the model to a public repository (Option B) or having it examined by a PhD data scientist (Option C) are not mandatory steps for deployment and do not ensure the model's readiness for business use.

Therefore, the most important step prior to deploying a model for a business use case is to follow data science best practices such as ensuring no data leakage and proper model validation (Option D). This ensures that the model is reliable, effective, and aligned with the business requirements before moving into production.

AG

You

Topic: Intro to ML and DS

Your company asks you to build a model that will predict existing customer churn based on historical churn data. In data science this is known as (Select one)

- A.  
unsupervised learning
  - B.  
supervised learning
  - C.  
artificial intelligence (AI)
  - D.  
deep learning
- ChatGPT  
ChatGPT

In data science, building a model to predict existing customer churn based on historical churn data is known as:

B. supervised learning

Explanation:

- **Supervised Learning:** This refers to a type of machine learning where the model is trained on labeled data, meaning the historical data used for training contains both input features (e.g., customer attributes, behavior, etc.) and corresponding output labels or target variable (e.g., churn or no churn). In the case of predicting customer churn, the historical data would include features of customers (e.g., demographics, usage patterns, transaction history) along with a label indicating whether each customer churned or not. The goal of supervised learning is to learn a mapping from input features to the target variable (churn prediction) based on the labeled examples provided during training. Common supervised learning algorithms for this task include logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks.
- **Unsupervised Learning:** This type of machine learning is used when the data is unlabeled, and the goal is to discover patterns or structure within the data without explicit target labels. Unsupervised learning techniques are not suitable for predicting churn directly since they focus on clustering, anomaly detection, or dimensionality reduction tasks where the output is not predefined.
- **Artificial Intelligence (AI):** AI is a broad field encompassing various disciplines and technologies aimed at creating intelligent systems capable of performing tasks that typically require human intelligence. Machine learning is a subset of AI that focuses on training algorithms to learn patterns and make predictions from data.
- **Deep Learning:** Deep learning is a specialized area of machine learning that utilizes neural networks with multiple layers (deep neural networks) to automatically learn hierarchical

representations of data. While deep learning can be applied to supervised learning tasks like churn prediction, it is not synonymous with supervised learning.

Therefore, in the context of building a model to predict customer churn based on historical data, the appropriate term is supervised learning (Option B), where the model learns from labeled examples to make predictions about future instances.

AG

You

Topic: Using Models

The easiest way to do interactive 'what-if' experiments where you enter different hypothetical examples and get predictions is to (Select one)

- A.  
run the input data through two or more built-in RapidMiner Studio templates.
- B.  
run Auto Model and use the simulator.
- C.  
run Turbo Prep to build two or more ETL processes.
- D.  
use random data generators on your input data.

ChatGPT

ChatGPT

The easiest way to do interactive 'what-if' experiments where you enter different hypothetical examples and get predictions is to:

- B. run Auto Model and use the simulator.

Explanation:

Auto Model in RapidMiner Studio is a powerful tool that automates the process of building and evaluating machine learning models. The Auto Model feature allows you to quickly train multiple models on your dataset, optimize hyperparameters, and assess model performance using cross-validation or holdout validation.

The simulator feature in Auto Model enables you to perform 'what-if' experiments by inputting different hypothetical examples (i.e., changing input variables or feature values) and obtaining

predictions from the trained models. This interactive capability allows you to explore how changes in input data affect the model's predictions without manually adjusting code or rerunning the entire modeling process.

Let's briefly discuss the other options:

A. Running the input data through two or more built-in RapidMiner Studio templates: While RapidMiner Studio provides various templates for common data processing tasks, such as data preparation and modeling, using templates alone may not offer the direct interactive 'what-if' experimentation capability provided by Auto Model's simulator.

C. Running Turbo Prep to build two or more ETL processes: Turbo Prep in RapidMiner Studio is a data preparation tool that helps streamline and automate the process of cleaning, transforming, and enriching data. However, it is focused on data preparation rather than interactive model experimentation.

D. Using random data generators on your input data: Generating random data does not facilitate specific 'what-if' experiments where you can control and modify input variables to observe changes in model predictions.

Therefore, the most suitable option for conducting interactive 'what-if' experiments with different hypothetical examples and obtaining predictions is B. run Auto Model and use the simulator within RapidMiner Studio. This allows for intuitive exploration and understanding of model behavior based on input variations.

AG

You

Topic: Methodologies

Which are true statements about CRISP-DM? (Select ALL correct answers)

- A. It is iterative and each phase is revisited many times.
- B. It is applicable in all industries.
- C. It begins with a phase called Business Understanding.
- D. It generally takes 1-3 weeks to complete.

ChatGPT

ChatGPT

The true statements about CRISP-DM (Cross-Industry Standard Process for Data Mining) are:

A. It is iterative and each phase is revisited many times.

C. It begins with a phase called Business Understanding.

Explanation:

A. CRISP-DM is an iterative process where each phase (Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment) can be revisited and iterated upon as needed throughout the project lifecycle. Iteration allows for refinement and improvement of the data mining process based on insights gained and feedback received.

C. CRISP-DM begins with the phase of Business Understanding, where the project objectives are defined, and the requirements from a business perspective are identified. Understanding the business context and goals is crucial for guiding subsequent phases of the data mining process.

Let's discuss why the other options are not correct:

B. CRISP-DM is a flexible and widely used methodology, but it is not strictly applicable in all industries. While it is designed to be adaptable and can be used across various sectors, the specific implementation and applicability may vary based on the nature of the project and industry context.

D. The duration of completing a CRISP-DM project can vary significantly depending on factors such as project scope, complexity, available resources, and data characteristics. While some projects might be completed within 1-3 weeks, others could take months or longer to execute fully. CRISP-DM is designed to be scalable and adaptable to different project timelines and requirements.

Therefore, the correct statements about CRISP-DM are A and C: it is iterative, and each phase is revisited many times, and it begins with a phase called Business Understanding. These

characteristics highlight the flexibility and adaptability of CRISP-DM as a structured methodology for effective data mining projects.

AG

You

Topic: Use Cases for ML

You are assigned to build a predictive classification model for customer feedback comments (free-form text). You have a large historical data set of customer IDs, texts, and feedback scores that show how much a customer liked the product:

In data science, generating scores such as these via text is called (Select one)

- A.  
Price Risk Clustering
  - B.  
Churn Prevention
  - C.  
Price Optimization
  - D.  
Sentiment Analysis
- ChatGPT  
ChatGPT

In data science, generating scores from free-form text data to assess sentiments or attitudes is called:

D. Sentiment Analysis

Explanation:

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique used to determine the sentiment expressed in text data. It involves analyzing text to identify and extract subjective information such as opinions, emotions, attitudes, and sentiments conveyed by individuals. In the context of customer feedback comments (free-form text), sentiment analysis would be used to automatically classify the sentiment expressed in each comment as positive, negative, or neutral based on the content.

For the given use case described (building a predictive classification model for customer feedback comments and feedback scores), sentiment analysis would be a crucial component. By applying sentiment analysis techniques, you can extract insights from customer comments and use these sentiments as features for predicting feedback scores or categorizing customer sentiment towards the product.

Let's briefly discuss the other options:

A. Price Risk Clustering: This refers to grouping or clustering entities based on the level of risk associated with pricing strategies or market conditions. It is unrelated to the task of analyzing text data for sentiment.

B. Churn Prevention: Churn prevention involves identifying and reducing customer attrition or churn. While customer feedback analysis can be related to customer retention efforts, it is distinct from sentiment analysis.

C. Price Optimization: Price optimization refers to the process of determining the optimal pricing strategy for products or services based on various factors. It does not directly relate to analyzing text data for sentiment.

Therefore, in the context of analyzing customer feedback comments to generate sentiment scores or classifications, the appropriate term in data science is D. Sentiment Analysis. This technique is essential for extracting valuable insights from unstructured text data and understanding customer sentiments towards products or services.

AG

You

Topic: Use Cases for ML

The best question that is answered via prescriptive analytics is (Select one)

A.

What happened in the past?

B.

Why did something happen in the past?



C.

What will happen in the future?

D.

How can we make something happen in the future?

ChatGPT

ChatGPT

The best question that is answered via prescriptive analytics is:

D. How can we make something happen in the future?

Explanation:

Prescriptive analytics focuses on providing recommendations or actionable insights to optimize decisions and actions. It goes beyond descriptive and predictive analytics by not only identifying what is likely to happen (predictive analytics) but also suggesting the best course of action to achieve desired outcomes (prescriptive analytics).

The key objective of prescriptive analytics is to answer the question "How can we make something happen in the future?" This involves leveraging data-driven models and algorithms to simulate different scenarios, evaluate potential actions, and recommend the most effective strategies or decisions to achieve specific business objectives or outcomes.

Let's briefly discuss the other options:

A. What happened in the past? - This question pertains to descriptive analytics, which focuses on summarizing historical data and understanding past events or trends.

B. Why did something happen in the past? - This question relates to diagnostic analytics, which aims to identify the root causes or factors that led to specific outcomes or events in the past.

C. What will happen in the future? - This question is addressed by predictive analytics, where statistical models and machine learning algorithms are used to forecast future outcomes based on historical data and patterns.

Therefore, among the given options, D. How can we make something happen in the future? aligns with the scope and purpose of prescriptive analytics, which is to provide actionable insights and recommendations to guide decision-making and achieve desired future outcomes.

AG

You

Topic: Methodologies

Your company asks you to take on a data science project and then sends you a large data set. As a data analyst following the CRISP-DM model, you should first (Select one)

- A.  
clean the data to get it ready for analysis.
- B.  
examine the data using data visualization.
- C.  
put the data into Auto Model to get quick model prototypes.
- D.  
ask to understand the business use case prior to any analysis.

ChatGPT

ChatGPT

As a data analyst following the CRISP-DM (Cross-Industry Standard Process for Data Mining) model, when you receive a large dataset for a data science project, the first step you should take is:

- D. ask to understand the business use case prior to any analysis.

Explanation:

CRISP-DM begins with the Business Understanding phase, which emphasizes the importance of clarifying the business objectives, goals, and requirements of the data science project before diving into data analysis or modeling. Understanding the business context and use case is critical for guiding subsequent phases of the data mining process and ensuring that the analysis and modeling efforts are aligned with the overall business goals.

By asking to understand the business use case first, you can gather essential information such as:

- The specific problem or question that needs to be addressed through data analysis.
- The intended business outcomes or objectives that the project aims to achieve.

- The stakeholders involved and their requirements or expectations from the project.
- The scope and constraints of the project, including data availability, timelines, and resource constraints.

Once you have a clear understanding of the business use case, you can proceed with other phases of the CRISP-DM model, such as Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment, in a structured and goal-oriented manner.

Let's briefly discuss why the other options are not the first step in the CRISP-DM model:

A. Cleaning the data to get it ready for analysis: Data cleaning and preparation typically follow the Business Understanding phase and occur during the Data Preparation phase of CRISP-DM. Before cleaning the data, it is important to understand the business context and requirements to ensure that the data cleaning process is aligned with the project objectives.

B. Examining the data using data visualization: Data visualization is a valuable technique for exploring and understanding the characteristics and patterns in the data. However, it is more effective after gaining a clear understanding of the business use case and objectives to focus the visualization efforts on relevant aspects of the data.

C. Putting the data into Auto Model to get quick model prototypes: Model building using automated tools like Auto Model is an iterative process that comes later in the CRISP-DM process, after completing the initial phases of Business Understanding and Data Preparation. Starting with Auto Model without understanding the business context can lead to misalignment with business goals and requirements.

Therefore, based on the CRISP-DM methodology, the first step upon receiving a large dataset for a data science project is to ask to understand the business use case prior to any analysis (Option D). This ensures that the subsequent data analysis and modeling efforts are purposeful and directly contribute to solving the business problem at hand.

AG

You

Topic: Methodologies

Which are phases of CRISP-DM? (Select one)

- A.  
Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment
  - B.  
Data Collection, Data Munging, Data Cleansing, Data Modeling, Data Validation, Data Deployment
  - C.  
Data Access, Blending, Cleansing, Modeling, Scoring, Validation
  - D.  
Concept & Initiation, Planning, Execution, Performance Monitoring, Project Close
- ChatGPT  
ChatGPT

The phases of CRISP-DM (Cross-Industry Standard Process for Data Mining) are:

- A. Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment

Explanation:

CRISP-DM is a widely used methodology for conducting data mining and data science projects. It consists of the following phases:

1. Business Understanding: In this phase, the project objectives and requirements are defined from a business perspective. This involves understanding the project goals, determining what success looks like, and identifying the business context and constraints.
2. Data Understanding: This phase focuses on collecting and exploring the data to gain insights into its characteristics, quality, and potential issues. It involves activities such as data collection, data description, initial data exploration, and identifying data quality issues.
3. Data Preparation: In this phase, the data is cleaned, transformed, and prepared for modeling. This includes tasks such as handling missing values, dealing with outliers, feature selection, feature engineering, and formatting the data for modeling.
4. Modeling: The modeling phase involves selecting appropriate modeling techniques, building and training machine learning models using the prepared data, and evaluating model performance against business objectives.
5. Evaluation: In this phase, the model's performance is assessed thoroughly to ensure it meets the business objectives and requirements. Different evaluation metrics are used to measure the model's effectiveness and robustness.
6. Deployment: The final phase involves deploying the developed model into production or operational use. This includes integrating the model into existing systems, establishing monitoring mechanisms, and ensuring ongoing maintenance and updates as needed.

Let's briefly discuss why the other options are not correct:

- B. This option presents a sequence of data-related activities (Data Collection, Data Munging, Data Cleansing, Data Modeling, Data Validation, Data Deployment) that do not align with the structured phases of CRISP-DM.
- C. This option includes steps such as Data Access, Blending, Cleansing, Modeling, Scoring, and Validation, which are not part of the standard CRISP-DM methodology.
- D. This option describes phases of a project management lifecycle (Concept & Initiation, Planning, Execution, Performance Monitoring, Project Close) rather than the specific phases of the CRISP-DM process for data mining and data science projects.

Therefore, the correct answer is A. Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment, representing the well-defined and structured phases of the CRISP-DM methodology.

AG

You

Topic: Using Models

You have a two-class predictive model with 80% accuracy and 5% standard deviation. Is this considered an 'underfit' model? (Select one)

- A.  
Yes
- B.  
No
- C.  
Yes if the recall is also  $\geq 80\%$
- D.  
There is not enough information to determine whether or not the model is underfit.

ChatGPT

ChatGPT

To determine whether a two-class predictive model with 80% accuracy and 5% standard deviation is considered 'underfit', we need to understand the concept of underfitting in the context of machine learning models.

Underfitting occurs when a model is too simple to capture the underlying patterns or relationships in the data, resulting in poor performance even on the training dataset. This often leads to low accuracy and high bias, where the model fails to learn the complexities of the data.

Given the provided information (80% accuracy and 5% standard deviation), we can assess whether the model is underfit:

1. **Accuracy:** An accuracy of 80% alone does not directly indicate whether the model is underfit. Accuracy is a measure of overall correctness of the model's predictions across both classes. While 80% accuracy may seem reasonable, it depends on the specific context and requirements of the task.
2. **Standard Deviation:** The 5% standard deviation mentioned here is less commonly used in the context of model evaluation. Standard deviation typically refers to the variability or spread of data points around the mean. Without additional context or information on how standard deviation is calculated in this scenario, it's challenging to interpret its significance in determining underfitting.
3. **Consideration of Model Complexity:** To assess underfitting, it's crucial to consider the complexity of the model relative to the complexity of the underlying data. If the model is too simple (e.g., linear model for highly nonlinear data), it may underfit the training data and generalize poorly to unseen data.

Given the options provided:

A. Yes: This suggests that the model is underfit. However, underfitting is typically associated with low accuracy and high bias, which may not align with the given 80% accuracy.

B. No: This suggests that the model is not underfit. This could be a reasonable interpretation based on the provided accuracy metric, but further analysis is needed to confirm.

C. Yes if the recall is also  $\geq 80\%$ : Recall (sensitivity) is a measure of the model's ability to correctly identify positive instances out of all actual positive instances. While recall is important for assessing model performance, it alone does not determine whether the model is underfit.

D. There is not enough information to determine whether or not the model is underfit: This option acknowledges that more information is needed to make a definitive assessment of whether the model is underfit.

Based on the typical understanding of underfitting and the provided information, the most appropriate answer is:

B. No: Without additional information suggesting high bias or poor generalization performance, it's not immediately evident that the model is underfit based solely on the given accuracy and standard deviation metrics. Further analysis of the model's complexity, performance on training and validation data, and comparison with baseline models would be necessary to determine if underfitting is occurring.

AG

You

Topic: Intro to ML and DS

Select the TWO main types of machine learning. (Select BOTH correct answers)

- A. Deep Learning
- B. Supervised
- C. Programmed
- D. Unsupervised

ChatGPT

ChatGPT

The TWO main types of machine learning are:

B. Supervised

D. Unsupervised

Explanation:

1. Supervised Learning (Option B): In supervised learning, the model is trained on a labeled dataset where the input data is paired with corresponding target labels or outcomes. The goal is to learn a mapping or relationship between input features and target labels. Examples of supervised learning tasks include classification (predicting discrete labels) and regression (predicting continuous values).
2. Unsupervised Learning (Option D): In unsupervised learning, the model is trained on an unlabeled dataset where the input data has no corresponding target labels. The objective is to discover patterns, structures, or relationships within the data without explicit guidance from labeled outcomes. Unsupervised learning tasks include clustering (grouping similar

data points together) and dimensionality reduction (reducing the number of input features while preserving important information).

Let's briefly discuss why the other options are not correct:

A. Deep Learning: Deep learning refers to a subset of machine learning techniques that use deep neural networks with multiple layers to learn representations of data. Deep learning can be applied within both supervised and unsupervised learning paradigms but is not a standalone type of machine learning.

C. Programmed: "Programmed" is not a recognized type of machine learning. Machine learning involves training models to learn from data rather than being explicitly programmed to perform specific tasks.

Therefore, the correct answers are B. Supervised and D. Unsupervised, representing the two main categories of machine learning based on the nature of training data and learning objectives.

AG

You

Topic: Using Models

You built a machine learning model in RapidMiner for your company, and there are many opportunities to apply the model. It has been validated for those opportunities with an overall accuracy of 90%. Any correct prediction generates a \$10 profit, and any wrong prediction generates a \$50 loss. Generating no prediction is a lost opportunity and generates neither profit nor loss (\$0). What is the overall most profitable option for your company? (Select one)

Hints - In this simple example:

- all correct predictions have the same profit, and all incorrect predictions have the same cost.
- if no model is in production when the opportunity presents itself, then the opportunity is lost (\$0).
- there are no other significant costs to put the model into production.

A.

First, put the model into production, then consider refining the model.

B.

First, refine the model until you have over 95% accuracy, then put the model into production.

C.

First, refine the model until you have over 99.5% accuracy, then put the model into production.

D.



Do not ever put the model into production; work on a different project instead.

ChatGPT

ChatGPT

To determine the overall most profitable option for your company based on the machine learning model's performance and associated costs and benefits, we need to consider the expected profit or loss under different scenarios. Let's evaluate the options:

Given:

- Correct prediction profit: \$10
- Incorrect prediction loss: \$50
- No prediction opportunity loss: \$0

The overall profitability of putting the model into production depends on the model's accuracy and the associated costs and benefits of correct and incorrect predictions.

Option A. First, put the model into production, then consider refining the model:

- With a model accuracy of 90%, the expected profit per prediction can be calculated as follows:
  - Probability of correct prediction ( $P(\text{correct})$ ) = 90% = 0.9
  - Probability of incorrect prediction ( $P(\text{incorrect})$ ) = 10% = 0.1

Expected profit per prediction = ( $P(\text{correct}) * \text{Profit per correct prediction}$ ) - ( $P(\text{incorrect}) * \text{Loss per incorrect prediction}$ )

$$= (0.9 * \$10) - (0.1 * \$50)$$

$$= \$9 - \$5$$

- = \$4 (expected profit per prediction)

Option B. First, refine the model until you have over 95% accuracy, then put the model into production:

- With a higher model accuracy (e.g., 95%), the expected profit per prediction would increase. However, refining the model to achieve higher accuracy may require additional time, resources, and investment.

Option C. First, refine the model until you have over 99.5% accuracy, then put the model into production:

- Further refining the model to achieve very high accuracy (e.g., 99.5%) would likely result in diminishing returns, as achieving such high accuracy may be challenging and may not significantly improve profitability given the diminishing error rate.

Option D. Do not ever put the model into production; work on a different project instead:

- Choosing not to put the model into production would result in lost opportunities for generating profit from correct predictions, potentially outweighing the losses from incorrect predictions.

Based on this analysis:

- Option A (First, put the model into production, then consider refining the model) appears to be the most practical and potentially profitable option given the current model accuracy of 90%. Implementing the model into production would start generating profit from correct predictions while considering further refinements based on practical considerations and observed performance.
- Options B and C (Refining the model to achieve higher accuracies) may not be as justified if the additional effort and resources required to achieve higher accuracies do not significantly increase expected profitability.
- Option D (Not putting the model into production) would result in missed opportunities and potential losses from not leveraging the model's predictive capabilities.

Therefore, based on the provided scenario and considerations, the overall most profitable option for your company would be:

A. First, put the model into production, then consider refining the model.

AG

You

Topic: Visualization

What is an advantage of a Parallel Coordinates chart? (Select one)

- A.  
It is intuitive and easy to read.
- B.  
It allows the user to visualize many columns at the same time.
- C.  
It condenses the information into a small amount of space.
- D.  
None of these is a unique advantage of a Parallel Coordinates chart.

ChatGPT

ChatGPT

The advantage of a Parallel Coordinates chart is:

B. It allows the user to visualize many columns at the same time.

Explanation:

A Parallel Coordinates chart is a type of visualization that displays multivariate data in a structured manner, particularly useful for datasets with multiple numerical variables (columns). The key advantage of a Parallel Coordinates chart is its ability to represent and visualize relationships among multiple variables simultaneously.

By plotting each data point as a polyline across multiple parallel axes (one axis for each variable), a Parallel Coordinates chart enables users to observe patterns, trends, and correlations across different variables in the dataset. This visualization technique allows for the simultaneous examination of multiple columns (variables) within the same chart, making it effective for exploratory data analysis and understanding complex datasets.

Let's briefly discuss why the other options are not correct:

A. "It is intuitive and easy to read" may not universally apply to Parallel Coordinates charts. While they can be effective for certain datasets, the interpretation of these charts can be challenging for datasets with many variables or complex patterns.

C. "It condenses the information into a small amount of space" is not a unique advantage of Parallel Coordinates charts. While they can display multiple variables in a compact format, the complexity of the chart can increase with the number of variables, potentially affecting readability.

D. "None of these is a unique advantage of a Parallel Coordinates chart" is incorrect because the ability to visualize many columns simultaneously (Option B) is indeed a key advantage of using a Parallel Coordinates chart, distinguishing it from other types of visualizations that focus on individual or pairwise relationships between variables.

Therefore, the correct advantage of a Parallel Coordinates chart is B. It allows the user to visualize many columns at the same time, making it a valuable tool for gaining insights into multivariate datasets and exploring relationships across multiple variables within a single visualization.

AG

You

Topic: Methodologies

After completing the Evaluation phase of a CRISP-DM project, you determine that the next step is to go back to the Business Understanding phase. A reason why this could happen is (Select ALL correct answers)

- A. the model did not provide information relevant to the business use problem.
- B. you have learned new information that changes the business problem.
- C. the model is not well set-up to find patterns.
- D. you have run out of time for your project.

ChatGPT

ChatGPT

After completing the Evaluation phase of a CRISP-DM project and deciding to return to the Business Understanding phase, the possible reasons for this decision could include:

- A. The model did not provide information relevant to the business use problem.
- B. You have learned new information that changes the business problem.

Explanation:

A. The model did not provide information relevant to the business use problem:

- This situation occurs when the evaluation of the model's performance reveals that the insights or predictions generated by the model are not aligned with the business objectives or requirements identified in the Business Understanding phase. It may indicate that the initial problem framing or understanding of business needs was incomplete or inaccurate.

B. You have learned new information that changes the business problem:

- Over the course of the project, new information or insights may emerge that impact the original business problem definition. This could involve changes in stakeholder requirements, market conditions, regulatory factors, or other external influences that necessitate a reassessment of the business problem and objectives.

C. The model is not well set-up to find patterns:

- While this statement reflects a potential issue with the modeling phase (Modeling), it may not directly lead to a decision to return to the Business Understanding phase unless the limitations of the model highlight gaps or discrepancies in the initial understanding of the business problem or data requirements.

D. You have run out of time for your project:

- Time constraints and project deadlines can certainly impact the progression and completion of a CRISP-DM project. However, the decision to return to the Business Understanding phase would typically be driven by the need to address specific issues related to business alignment, problem definition, or new insights rather than solely due to project time constraints.

Therefore, based on the CRISP-DM methodology, the correct reasons for returning to the Business Understanding phase after completing the Evaluation phase include:

- The model did not provide information relevant to the business use problem (Option A).
- You have learned new information that changes the business problem (Option B).

These scenarios highlight the importance of iterative and adaptive approaches in data science projects, where insights gained from evaluation and analysis inform subsequent iterations of problem formulation, data preparation, modeling, and deployment within the CRISP-DM framework.