**From Understanding to Preparation and Modeling to Evaluation**

In this module, you will learn what data scientists do when their tasks and goals are to understand, prepare, and clean the data. You’ll examine the purposes, characteristics, and goals of the data modeling process. You’ll also explore how to prepare a data set by handling missing, invalid, or misleading data. Then check out the hands-on labs where you can gain experience completing tasks relevant to the Data Understanding, Data Preparation, and Modeling and Evaluation stages. You’ll be able to apply the skills you learn to future data science problems.

**Learning Objectives**

* Prepare a data set by handling missing, invalid, or misleading data.
* Describe the purpose and characteristics of the data modeling process.
* Evaluate a decision tree model using a training and a test dataset.
* Build a decision to tree to determine the cuisine type for a data set of recipes.
* Summarize the processes of understanding data preparing data, modeling, and evaluation phases of the data science methodology.

# **From Data Understanding to Preparation**

## **Data Understanding**

Simple Explanation:

* **Data Understanding**: This is the first step in data science where you check if the data you collected is good enough to solve the problem you are working on.
* **Descriptive Statistics**: You look at basic statistics of your data, like:
  + **Mean**: Average value
  + **Median**: Middle value
  + **Minimum/Maximum**: Smallest and largest values
  + **Standard Deviation**: How spread out the values are
* **Pairwise Correlations**: This helps you see how different variables (data points) relate to each other. If two variables are very similar, you might only need to keep one of them for your analysis.
* **Histograms**: These are graphs that show how data values are distributed. They help you understand the shape of your data and decide if you need to change anything to make it better for analysis.
* **Data Quality**: You check for missing or incorrect values. For example, if a variable like "age" has a value of 999, it might mean the data is missing, and you need to fix it.
* **Refining Definitions**: Sometimes, the initial way you define a problem (like what counts as a "congestive heart failure admission") might not be complete. You may need to go back and adjust your definitions based on what you learn from the data.

Summary:

* The **Data Understanding** stage is crucial for ensuring that the data collected is relevant and useful for solving the problem at hand. It involves analyzing basic statistics, checking relationships between variables, visualizing data distributions, and ensuring data quality. This process may lead to refining the definitions of the problem based on insights gained from the data.

## **Data Preparation Concepts**

Simplified Explanation:

* **Data Preparation** is like washing vegetables before cooking. You remove unwanted parts (like dirt) to make sure the data is clean and usable.
* This phase takes a lot of time—up to **90%** of a data science project! Automating some tasks can help reduce this time.
* Just like chopping onions helps flavors mix better in cooking, transforming data makes it easier to work with.
* Key tasks in data preparation include:
  + **Handling missing or invalid values**
  + **Removing duplicates**
  + **Formatting data correctly**
  + **Feature engineering**, which means creating new data features that help in analysis.
* If data preparation is done well, it sets the stage for successful analysis. If skipped, it can lead to poor results.

Summary:

The video emphasizes the importance of **Data Preparation** in data science, highlighting that it is a time-consuming but crucial step. It involves cleaning and transforming data to ensure it is ready for analysis, which includes handling missing values, removing duplicates, and feature engineering. Proper data preparation leads to better outcomes in data science projects.

## **Data Preparation – Case Study**

, the focus is on **Data Preparation** in the context of a case study related to **congestive heart failure**. Here’s a simplified explanation and summary:

Simplified Explanation:

1. **Data Preparation**: This is like cleaning vegetables before cooking. You remove unwanted parts to make sure the data is ready for analysis.
2. **Defining Congestive Heart Failure**: The first step was to clearly define what congestive heart failure is, which involved identifying specific medical codes related to it.
3. **Readmission Criteria**: They set a rule to determine if a patient was readmitted for heart failure within 30 days after leaving the hospital.
4. **Aggregating Data**: They combined many records for each patient into one single record. This included all medical visits, diagnoses, and treatments.
5. **Creating New Information**: During this process, they added new columns to the data that showed important details like how often patients visited doctors and any other health conditions they had.
6. **Literary Review**: They checked existing literature to ensure they didn’t miss any important data about other health conditions that could affect heart failure.
7. **Final Data Table**: The result was a single table for each patient with all relevant information, which would be used for building a predictive model.

Summary:

The video discusses the **Data Preparation** stage in a case study about congestive heart failure. It highlights the importance of defining the condition, setting readmission criteria, and aggregating patient data into a single record. The preparation process ensures that all relevant information is included for effective analysis and modeling. Ultimately, this stage is crucial for making accurate predictions about patient readmissions.

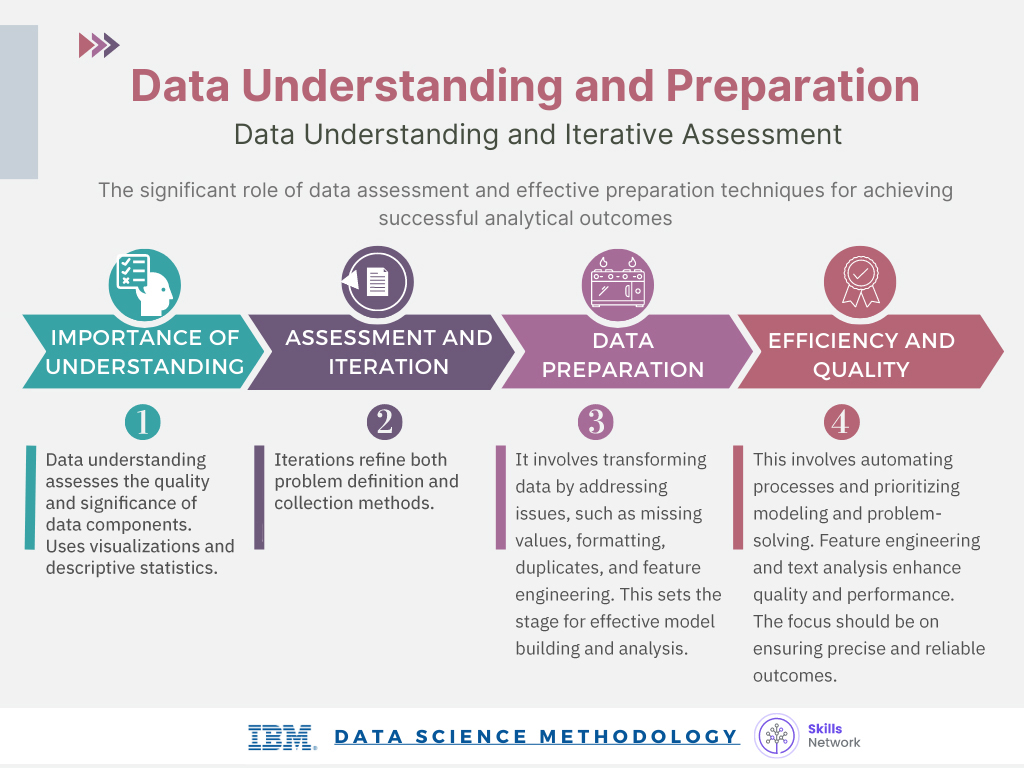
## **Summary: From Understanding to Preparation**

**Lesson summary**

**Module 2 Lesson 1: From Understanding to Preparation**

Congratulations! You have completed this lesson. At this point in the course, you know:

* The Data Understanding stage encompasses all activities related to constructing the data set and answers the question as to whether the data you collected represents the problem to be solved.
* During the Data Understanding stage, scientists might use descriptive statistics, predictive statistics, or both.
* Data scientists commonly apply Hurst, univariates, and other statistics on each variable, such as mean, median, minimum, maximum, standard deviation, pairwise correlation, and histograms.
* Data scientists also use univariates, statistics, and histograms to assess data quality.



* During the Data Preparation stage, data scientists must address missing or invalid values, remove duplicates, and validate that the data is properly formatted.
* Feature engineering, also part of the Data Preparation stage, uses domain knowledge of the data to create features that make the machine learning algorithms work.
* Text analysis during the Data Preparation stage is critical for validating that the proper groupings are set and that the programming is not overlooking hidden data.

## **Glossary**

**Glossary: From Understanding to Preparation**

Welcome! This alphabetized glossary contains many of the terms you'll find within this lesson. These terms are important for you to recognize when working in the industry, when participating in user groups, and when participating in other certificate programs.

|  |  |
| --- | --- |
| **Term** | **Definition** |
| **Automation** | Using tools and techniques to streamline data collection and preparation processes. |
| **Data Collection** | The phase of gathering and assembling data from various sources. |
| **Data Compilation** | The process of organizing and structuring data to create a comprehensive data set. |
| **Data Formatting** | The process of standardizing the data to ensure uniformity and ease of analysis. |
| **Data Manipulation** | The process of transforming data into a usable format. |
| **Data Preparation** | The phase where data is cleaned, transformed, and formatted for further analysis, including feature engineering and text analysis. |
| **Data Preparation** | The stage where data is transformed and organized to facilitate effective analysis and modeling. |
| **Data Quality** | Assessment of data integrity and completeness, addressing missing, invalid, or misleading values. |
| **Data Quality Assessment** | The evaluation of data integrity, accuracy, and completeness. |
| **Data Set** | A collection of data used for analysis and modeling. |
| **Data Understanding** | The stage in the data science methodology focused on exploring and analyzing the collected data to ensure that the data is representative of the problem to be solved. |
| **Descriptive Statistics** | Summary statistics that data scientists use to describe and understand the distribution of variables, such as mean, median, minimum, maximum, and standard deviation. |
| **Feature** | A characteristic or attribute within the data that helps in solving the problem. |
| **Feature Engineering** | The process of creating new features or variables based on domain knowledge to improve machine learning algorithms' performance. |
| **Feature Extraction** | Identifying and selecting relevant features or attributes from the data set. |
| **Interactive Processes** | Iterative and continuous refinement of the methodology based on insights and feedback from data analysis. |
| **Missing Values** | Values that are absent or unknown in the dataset, requiring careful handling during data preparation. |
| **Model Calibration** | Adjusting model parameters to improve accuracy and alignment with the initial design. |
| **Pairwise Correlations** | An analysis to determine the relationships and correlations between different variables. |
| **Text Analysis** | Steps to analyze and manipulate textual data, extracting meaningful information and patterns. |
| **Text Analysis Groupings** | Creating meaningful groupings and categories from textual data for analysis. |
| **Visualization techniques** | Methods and tools that data scientists use to create visual representations or graphics that enhance the accessibility and understanding of data patterns, relationships, and insights. |

# **From Modeling to Evaluation**

## **Modeling Concepts**

Simple Explanation:

* **Modeling** is like cooking; you mix ingredients (data) to create a dish (model).
* There are two main types of models:
  + **Descriptive Models**: These help us understand patterns. For example, if someone likes a certain movie, they might also like another one.
  + **Predictive Models**: These help us make predictions, like guessing if it will rain tomorrow (yes/no).
* To create a predictive model, data scientists use a **training set**, which is a collection of past data with known outcomes. This helps them check if their model is working correctly.
* The success of modeling depends on understanding the problem and using the right methods.
* Data scientists need to keep adjusting their models to ensure they are accurate and useful.

Summary:

The video covers the importance of modeling in data science, explaining the difference between descriptive and predictive models. It emphasizes the use of training sets for predictive modeling and the need for constant refinement to achieve accurate results. The ultimate goal is to build a model that effectively answers specific questions.

## **Modeling – Case Study**

Comprehensive Explanation:

1. **Modeling Stage**: This is the phase in the data science methodology where data scientists create a model to predict outcomes, such as whether a patient will be readmitted for heart failure.
2. **First Model**:
   * The initial model achieved an accuracy of **85%**. However, it only correctly identified **45%** of actual readmissions (the "yes" outcomes), indicating that it was not very effective in predicting readmissions.
3. **Types of Errors**:
   * **Type I Error (False Positive)**: This occurs when the model incorrectly predicts a readmission when there isn't one. This can lead to unnecessary treatments and interventions.
   * **Type II Error (False Negative)**: This happens when the model fails to predict a readmission when there is one. This can result in serious consequences for the patient.
4. **Adjusting Costs**:
   * The model's parameters can be adjusted to weigh the importance of these errors differently. For instance, setting a higher cost for missing a readmission (Type II error) can improve the model's sensitivity.
5. **Second Model**:
   * With a cost ratio of **9-to-1** (favoring the identification of readmissions), the model correctly identified **97%** of readmissions. However, it had a low overall accuracy of **49%**, indicating many false positives, which is not ideal.
6. **Third Model**:
   * A more balanced approach was taken with a **4-to-1** cost ratio. This resulted in **68% sensitivity** (correctly identifying readmissions) and **85% specificity** (correctly identifying non-readmissions), achieving an overall accuracy of **81%**. This was a better balance between identifying both types of outcomes.

Key Metrics:

* **Sensitivity** (True Positive Rate):
  + Measures the proportion of actual positive cases (readmissions) correctly identified by the model.
  + Formula: [ \text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} ]
* **Specificity** (True Negative Rate):
  + Measures the proportion of actual negative cases (non-readmissions) correctly identified by the model.
  + Formula: [ \text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} ]

Summary:

The video explains the modeling stage in data science, focusing on how to build and improve a predictive model for heart failure readmissions. It highlights the importance of adjusting the costs associated with different types of errors to enhance model performance. The key metrics discussed are **sensitivity**, which measures the model's ability to correctly identify readmissions, and **specificity**, which measures its ability to correctly identify non-readmissions. The video illustrates the process of refining the model through different iterations to achieve a better balance between sensitivity and specificity, ultimately leading to a more effective predictive model.

## **Evaluation**

Simplified Explanation:

* **Model Evaluation**: This is the process of checking if a model (a way to make predictions or understand data) is working correctly and if it answers the original question we had.
* **Two Phases of Evaluation**:
  1. **Diagnostic Measures**: This checks if the model is functioning as expected. For predictive models, we can use tools like decision trees to see if the predictions match what we designed. For descriptive models, we can test them against known outcomes to see if they need adjustments.
  2. **Statistical Significance Testing**: This ensures that the data is being interpreted correctly in the model, helping to avoid confusion when we get results.
* **Finding the Best Model**: The video uses a case study to show how to find the best model by adjusting the cost of making mistakes (misclassifying outcomes). It explains that different models can have different rates of correctly predicting outcomes (true positives) and incorrectly predicting them (false positives).
* **ROC Curve**: This is a tool used to visualize how well a model performs. It helps us see which model is the best by comparing the true positive rate to the false positive rate.

Summary:

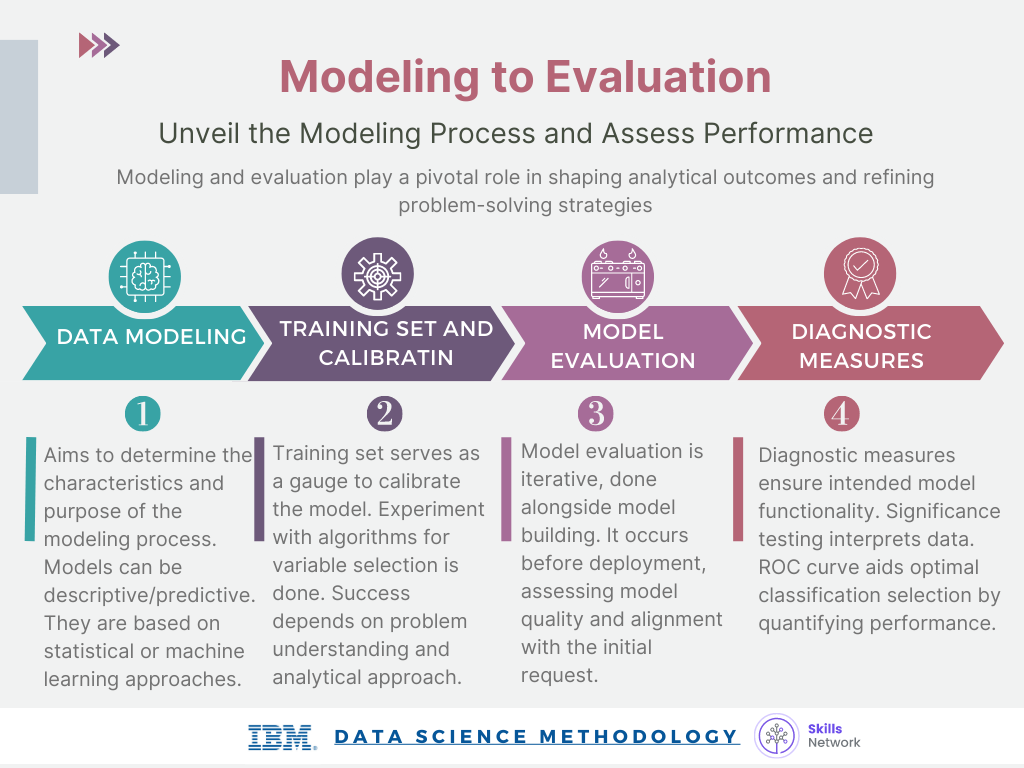
The video explains the importance of model evaluation in data science, highlighting two main phases: diagnostic measures and statistical significance testing. It illustrates how to find the optimal model by tuning misclassification costs and introduces the ROC curve as a tool for assessing model performance. The goal is to ensure that the model effectively answers the initial question posed.

## **Summary**

**Lesson Summary: Modeling to Evaluation**

**Congratulations! You have completed this lesson. At this point in the course, you know:**

* The end goal of the Modeling stage is that the data model answers the business question.
* The data modeling process uses a training data set. Data scientists test multiple algorithms on the training set data to determine whether the variables are required and whether the data supports answering the business question. The outcome of those models are either descriptive or predictive.



* The Evaluation phase consists of two stages, the diagnostic measures phase, and the statistical significance phase.
* During the Evaluation stage, data scientists and others assess the quality of the model and determine if the modelanswers the initial Business Understanding question or if the data model needs adjustment.
* The ROC curve, known as the receiver operating characteristic curve, is a useful diagnostic tool for determining the optimal classification model. This curve quantifies how well a binary classification model performs, declassifying the yes and no outcomes when some discrimination criterion is varied

## **Glossary**

**Glossary: From Modeling to Evaluation**

Welcome! This alphabetized glossary contains many of the terms you'll find within this lesson. These terms are important for you to recognize when working in the industry, when participating in user groups, and when participating in other certificate programs.

|  |  |
| --- | --- |
| **Term** | **Definition** |
| **Binary classification model** | A model that classifies data into two categories, such as yes/no or stop/go outcomes. |
| **Data compilation** | The process of gathering and organizing data required for modeling. |
| **Data modeling** | The stage in the data science methodology where data scientists develop models, either descriptive or predictive, to answer specific questions. |
| **Descriptive model** | A type of model that examines relationships between variables and makes inferences based on observed patterns. |
| **Diagnostic measure based tuning** | The process of fine-tuning the model by adjusting parameters based on diagnostic measures and performance indicators. |
| **Diagnostic measures** | The evaluation of a model's performance of a model to ensure that the model functions as intended. |
| **Discrimination criterion** | A measure used to evaluate the performance of the model in classifying different outcomes. |
| **False-positive rate** | The rate at which the model incorrectly identifies negative outcomes as positive. |
| **Histogram** | A graphical representation of the distribution of a dataset, where the data is divided into intervals or bins, and the height of each bar represents the frequency or count of data points falling within that interval. |
| **Maximum separation** | The point where the ROC curve provides the best discrimination between true-positive and false-positive rates, indicating the most effective model. |
| **Model evaluation** | The process of assessing the quality and relevance of the model before deployment. |
| **Optimal model** | The model that provides the maximum separation between the ROC curve and the baseline, indicating higher accuracy and effectiveness. |
| **Receiver Operating Characteristic (ROC)** | Originally developed for military radar, the military used this statistical curve to assess the performance of binary classification models. |
| **Relative misclassification cost** | This measurement is a parameter in model building used to tune the trade-off between true-positive and false-positive rates. |
| **ROC curve (Receiver Operating Characteristic curve)** | A diagnostic tool used to determine the optimal classification model's performance. |
| **Separation** | Separation is the degree of discrimination achieved by the model in correctly classifying outcomes. |
| **Statistical significance testing** | Evaluation technique to verify that data is appropriately handled and interpreted within the model. |
| **True-positive rate** | The rate at which the model correctly identifies positive outcomes. |