CFA Institute

Data Science for Investment Professionals Certificate

Understand the application of data science in the investment process and act as the expert at the intersection of data scientists, investment professionals and clients

Are you ready for the machine learning revolution?

The Certificate in Data Science for Investment Professionals provides you with practical knowledge of machine learning fundamentals and how they are used in the investment process.

This certificate is designed to enable you to apply machine learning concepts to real-world investment problems and explain them clearly to a non-expert audience and clients.

Master the language of data science to better serve your clients

The certificate is composed of five courses with practical application exercises and one final exam. You will be able to:

- Describe and evaluate machine learning techniques
- Select and develop data visualizations using Python
- Apply machine learning to address investment problems
- Explain machine learning techniques to a non-expert audience
- Use natural language processing to make investment decisions
- Evaluate and mitigate machine learning biases

Course topics

Course 1	Data and Statistics Foundation
Course 2	Statistics for Machine Learning
Course 3	Machine Learning
Course 4	Natural Language Processing
Course 5	Mitigating Biases in the Data Science Pipeline
Final Exam	90-minute multiple-choice exam (online)

Key facts

- Launched in Spring 2023
- Online, self-paced learning
- Time to complete: 90 – 100 hours
- Content access: 12 months
- Recognition: Certificate and Digital badge

Course 1 – Data and Statistics Foundation

Acceptance Region
In a hypothesis test, the acceptance region is a range of test statistic results (sample outcomes) in which the null hypothesis is accepted (i.e., we cannot rule out chance as an explanation for the observed results).

Alternative Hypothesis
In a hypothesis test, the alternative hypothesis is constructed so that it represents the conclusion you reach if the null hypothesis is rejected. Together, the null hypothesis and alternative hypothesis must account for all possible outcomes to the study.

Arithmetic Mean
The sum of a set of values, divided by the number of values in the set (also known as the average).

Array Data
Array data is a collection of similar data types stored contiguously. A vector is an of array data, and a two-dimensional matrix is also an array of data. A spreadsheet range is an array.

Bar Chart
A bar chart represents items or categories on one axis and counts or values for those items on the other axis.

Bias in Estimation
Estimation typically involves samples taken to measure some quantity or fit a model in a larger population or process, and bias is error in a consistent direction across multiple samples.

Categorical Data
Data that indicate which category a record belongs to (e.g., for a resident's home: condo, single-family dwelling, rental apartment). Ordered categorical data represents categories that can be ranked by magnitude (e.g., no pain, mild pain, moderate pain, severe pain).

Central Limit Theorem
The central limit theorem states that the distribution of the sample mean approaches normality in its shape as the sample size increases, even if the underlying data are not normally distributed.

Coefficient of Variation
The coefficient of variation for a dataset is the ratio of the standard deviation to the mean.

Confidence Interval
For a statistic calculated from a sample, a confidence interval is a range that would be expected to enclose the true population value most of the time (e.g., 90% of the time for a 90% confidence interval).

Contingency Table
A contingency table is a table with columns representing one categorical variable and rows representing another. The cells reflect counts of data falling into the two categories. For example, for investor data, the rows might represent male/female and the columns might represent owns stocks/doesn't own stocks. The cell in the "male" row and the "doesn't own stocks" column is the count of male investors in the data who do not own stocks.

Continuous Data
Continuous data are numeric data that can be measured on an infinite scale (i.e., between any two points there are an infinite number of values a data point can assume).

Critical Value
In a hypothesis test, the critical value is a sample statistic that is just barely extreme enough (compared with what you would expect under the null model) to be deemed improbable enough to reject the null hypothesis.

Cross-Sectional Data
Cross-sectional data are data from many units or records at one point in time. It is like taking a snapshot of all details at a point in time.

Cumulative Frequency Distribution
A cumulative frequency distribution can be visualized as a two-dimensional plot, with possible data values (or ranges) arrayed in ascending order on the x-axis, with the y-axis representing the proportion of the data falling at or below the x-value (or range).

Deciles
Deciles are segments of the data that are delineated by ranking the observations and then segmenting them into tenths.

Discrete Data
Discrete data assume values on a scale that is finite, or countably infinite. Integer scales are a common example, but the data need not be in integer form.

Estimate
In statistics, an estimate is a statistic calculated from, or a model derived from, a sample of data drawn from a larger population.

False Positive
With binary data in which one category is of special interest (typically designated as a "1" or positive), a false positive is a record that is estimated by a model or diagnostic procedure to be a "1" but is, in reality, a "0" (or negative).

Family-Wise Error Rate (FWER)
In a single hypothesis test of one question, the probability of mistakenly concluding a sample result is significant, given the null model, is set at some low level, often 0.05. When multiple questions are asked and tested, all at that 0.05 level, the probability of such a mistake rises with the number of tests. This multiple test error probability is the FWER.

Frequency Table
For a dataset, a frequency table tabulates the counts of how many observations have the various values, or how many observations fall within the various ranges (bins). The values or bins are arranged contiguously to cover all the data, and zero-count values or bins are included.

Geometric Mean
The geometric mean of a set of n values is found by multiplying the values together and then taking the nth root of the product. The geometric mean is often used to describe proportional growth.

Harmonic Mean
The harmonic mean of a set of values is calculated by dividing the number of values by the sum of their reciprocals. The geometric mean is often used to average rates (e.g., price multiples, speeds).

Histogram
A histogram is a plot of a frequency table. One axis (usually the x-axis) has the data's possible values or bin ranges, and the other has the counts or proportions of observations that take a given value or fall in a given bin range.

Hypothesis Testing
A hypothesis test is a statistical procedure applied to sample outcomes in a study or experiment to test how far an outcome departs from what might occur in a chance model (the "null hypothesis" embodying the assumption of no treatment effect, or nothing unusual). If the outcome is sufficiently extreme compared with the chance model, the null hypothesis is rejected and the outcome is deemed "statistically significant." The purpose of a hypothesis test is to protect against being fooled by random chance.

Interval Estimate
An interval estimate is a range bracketing a measurement from a sample. You see interval estimates often in opinion poll results: "35% favor the proposition, with a margin of error of 5% in either direction."

Joint Probability
In considering two events, the joint probability is the probability that both will happen.

Kurtosis
Kurtosis is a measure of the extent to which a data distribution is more or less spread out than a normal distribution (i.e., has more or fewer values in its tails).

Kurtosis, Leptokurtic or Fat-Tailed
More extreme or distant observations occur than with a normal distribution.

Kurtosis, Mesokurtic
Extreme or distant observations occur to the same extent as in a normal distribution.

Kurtosis, Platykurtic or Thin-Tailed
Fewer extreme or distant observations occur than with a normal distribution.

Look-Ahead Bias
Bias that results when a simulation study or a model uses information that will not be available at the time of real-word deployment.

Mean Absolute Deviation (MAD)
Mean absolute deviation is the average absolute deviation from the mean.

Median
After a dataset has been ranked according to the magnitude of one variable, the median is the record in the middle (i.e., half of the records lie above and half below).

Mode
The mode is the most frequent value in a dataset.

Monte Carlo Simulation
A Monte Carlo simulation is the repeated execution of an algorithm that incorporates a random component; each time the trial is repeated, this component is rerandomized.

Mutually Exclusive Events
Events (or, more precisely, outcomes to an event) are mutually exclusive if only one of them can occur.

Nominal Data
Nominal data are data, usually non-numeric, that provide no quantitative information. For example, the terms "transportation," "communications," and "tech" to describe business sectors are nominal.

Null Hypothesis
In a statistical hypothesis test, the null hypothesis embodies the proposition you are hoping to disprove: the treatment has no effect on blood pressure, the new web page does not increase click throughs. It is represented by a probability model reflecting this proposition, such as all the blood pressure readings grouped together in a single hat, or a single click-through rate for both web pages.

Numerical Data
Numeric data are data that are expressed in numbers that convey quantitative information; numbers that simply represent categories are not included.

Ordinal Data
Ordinal data are rank data in which the values (usually integers) represent rank order, but no further quantitative information (a financial stability value of 4 does not imply 33% more stability than 3.)

Panel Data
Panel data are data for a group of subjects over time (e.g., vaccine recipients monitored monthly).

Pareto Chart
A Pareto chart for a dataset is a bar chart in which the bars represent categories of some variable of interest, and the height of a bar represents the percent of records in the category for that bar. The bars are arranged in descending order of height and are often supplemented by a line plot above them representing cumulative percentages. Pareto charts are often used to attribute a condition to likely primary contributors (e.g., 40% of aircraft incidents are due to pilot error, 25% to air traffic control, 20% to maintenance issues).

Percentiles
After a dataset has been ranked according to the magnitude of one variable, a percentile, k (from 0 to 100), indicates what percentage of the records lie below k. For example, below the 15th percentile lie the lowest 15% of the records (according to the ranked variable).

Pie Chart
A pie chart is a circle, with pie-shaped segments whose sizes indicate the portion of the data that falls in the category represented by that segment.

Point Estimate
A point estimate is a measurement based on a sample. It is a single value, which distinguishes it from an interval estimate.

Population
In a study or modeling problem, a population is the entire data of interest. Often we deal only with a specified sample of the data, particularly when the population as a whole is difficult to define or reach in specific practical terms (e.g., "sports fans").

Population Mean
The average value of a variable of interest for the population; this is often a theoretical construct if all members of the population are not available.

Population Parameter
The value of some measure of a population (mean, standard deviation, percentile); this is often a theoretical construct if all members of the population are not available.

Probability
Most people have an intuitive sense of what probability means, such as "the chance of rain is 40%." To a gambler, the probability of winning a game, or of correctly specifying the outcome of an event such as tossing a coin, is the proportion of times the game is won, or the coin outcome happens. This is the "frequentist" definition of probability: the outcome of interest as a proportion of all outcomes to an event, if the event could be repeated indefinitely.

Probability Density Function
A probability density function is a function that can be used to evaluate the probability that a value will fall in a certain range.

Probability Distribution
A probability distribution is a tabular or graphical display of the probability of values in a population; it is like an extension of the frequency distribution. Common theoretical probability distributions are the uniform (all outcomes equally likely), the binomial (think of it as representing the number of heads in *n* tosses of a coin that has a *p* probability of landing heads), and the bell-shaped normal (think of it as a smoothed version of the binomial distribution where *n* is large).

P-Value
A hypothesis test estimates the probability that a chance "null hypothesis" model would produce an outcome as extreme as an observed value (e.g., a value such as the difference between a treatment sample and a control sample). This probability is the *p*-value.

Quantiles
Quantiles are equal-size segments of a ranked dataset, common quantiles being quartiles (fourths), quintiles (fifths), and deciles (tenths). The "second quintile," for example, would span from the 20th to the 40th percentile.

Random Variable
A random variable is a function or rule that assigns a real number to each element in a sample space. For example, in flipping a coin, the sample space elements are heads and tails, and the rule could be "1 if heads, 0 if tails." The number assigned could be discrete (from a finite or countably infinite set of numbers) or continuous (any real value).

Range
In describing a dataset, the range is the largest value minus the smallest value.

Rejection Region
In a hypothesis test, the rejection region is the range of sample statistics deemed too improbable under the null hypothesis to be attributable to chance. Typically, this means the 5% or 1% of values in the tail(s) of a sample statistic's distribution.

Sample
A sample is a set of observations or records drawn from a larger population or process. Samples can be drawn by a variety of methods to well-represent the population; samples selected randomly (simple random samples) are statistically useful because they reduce bias. Stratified sampling operates hierarchically (e.g., select schools randomly from among schools, select classes from the selected schools, select students from the selected classes). In systematic sampling, a rule governs selection (e.g., every 10th record).

Sample Selection Bias
Selection bias occurs when the sampling process yields samples that are not representative of the population you want to study. For example, in 1936, the *Literary Digest* poll predicted Alf Landon would beat Franklin Roosevelt for president, largely because the *Digest* relied on, among other things, lists of owners of telephones (a luxury good at the time).

Sample Statistic
A sample statistic is a metric (e.g., mean, standard deviation) calculated from a sample.

Sampling Error
Samples drawn from a population typically differ from one another, as do the metrics (e.g., means, proportions) calculated. Sampling error is a general term referring to the variability in these metrics.

Significance Level
In a hypothesis test, the significance level is a low percentage (typically 5% or 1%) that sets a cutoff in the distribution of the test statistic under the null hypothesis. Beyond this cutoff point, the result is deemed sufficiently extreme so as not to be attributed to chance and is, therefore, statistically significant.

Skewness
Skewness is departure from symmetry in the distribution of data, or in the distribution of a statistic. If the tail of the distribution extends more to the right, the skew is positive. If the tail extends more to the left, the skew is negative.

Standard Deviation
The standard deviation of a dataset is the square root of the variance.

Standard Error
The standard error of a sample statistic (say, the mean of a sample) is the standard deviation of that statistic, which is calculated from many sample results. If the larger population is not available to draw additional samples, bootstrap sampling of the original sample can be used to estimate the standard error.

Standard Normal Distribution
The standard normal distribution is a normally shaped data distribution with a mean of 0 and a standard deviation of 1. Data are often normalized or standardized by subtracting the mean and dividing it by the standard deviation; this gives the data a mean of 0 and a standard deviation of 1, but it does not make the data normally distributed if they were not normally distributed to begin with.

Structured Data
Structured data are data that are tabular and quantitative (or categorical) in nature. Structured data are contrasted with unstructured data (e.g., a set of Tweets, or doctor's notes).

Survivor Bias
Survivor bias occurs when the entire population is of interest but only the surviving members of a population are examined. For example, in studying an industry, looking only at firms that are currently in operation can cause survivor bias.

T-Distribution
The t-distribution is a bell-shaped distribution that is used to describe, among other things, the distribution of a sample mean, or the difference between two sample means. It is a family of distributions, whose shapes (peaked versus flat) depend on the sizes of the samples.

Test Statistic
In a statistical study, the test statistic is what you measure (e.g., mean, correlation, difference in means, ratio) in a sample or samples to answer the research question of the study.

Time-Period Bias
Time-period bias occurs when the time period for an analysis is not representative of a longer period. The opportunity for bias is higher when the time period is relatively short and when it is chosen specifically to illustrate a point.

Time-Series Data
Time-series data are measurements of one or more variables over time, with the measurements equally spaced in time.

Tree Map
A tree map is a rectangular representation of categories and subcategories in a dataset. Interior blocks of different sizes represent, by their size, the relative magnitudes of the categories and subcategories they represent. Colors may also be used to group subcategory blocks into categories.

Type I Error
In a hypothesis test, a type I error is mistakenly concluding an effect (e.g., the effect of an intervention in an experiment) is real, as opposed to being the result of random chance. A type I error is also known as a false positive, which happens when a null hypothesis is rejected when it is actually true.

Type II Error
This type of error, also known as false negative, occurs when a null hypothesis is accepted when it is actually not true.

Unstructured Data
Unstructured data are data that are not tabular and quantitative (e.g., a set of Tweets, or doctor's notes).

Variance
For a set of values, the variance is the average of the squared deviations from the mean.

Weighted Mean
For a set of values, a weighted mean is calculated by multiplying each value by a weight, and then by dividing by the sum of the weights. It is used when some values are deemed more important than others.

Z-Score
The z-score for a value from a dataset is calculated by subtracting the mean of the dataset, and dividing it by the standard deviation of the dataset. This process is sometimes called standardization or normalization.

Course 2 – Statistics for Machine Learning

Adjusted R^2

R^2 is a measure of the proportion of variation in data that is explained by a linear model. Adjusted R^2 is a modified version that penalizes added predictors, to encourage parsimony. It is:

$1 - (1 - R^2)((n - 1)/(n - p - 1))$,

where n is the number of records, and p is the number of predictors.

Analysis of Variance

Analysis of variance (ANOVA) is a statistical method for assessing the differences among multiple groups, where each group is summarized numerically (typically with the mean). In particular, the method evaluates whether the differences are statistically significant.

Autoregressive Time Series Model

An autoregressive time series model is one that forecasts future values in a time series on the basis of prior values.

Bayes Theorem

Bayes theorem is a formula for calculating the probability of an event, A, given that some other event, B has occurred, P(A|B), when you know the reverse: P(B|A). The formula is:

$P(A|B) = P(B|A) * P(A) / P(B)$.

For example, if you know the prevalence of a disease, and the probability of a positive diagnostic test if you have the disease, and the general rate of positive tests, you would use Bayes formula to calculate P(disease|positive test).

Bernoulli Distribution

The Bernoulli distribution is a special case of the binomial in which there is just one trial. For each event, there is only two possible outcomes.

Bias–Variance Trade-off

Statistical and machine learning prediction algorithms can underperform in two ways: (1) stable predictions that fall short in accuracy (bias); or (2) accurate predictions that are unstable, depending on the data sample (high variance). Typically, there is a trade-off: improve accuracy (minimize bias) and decrease stability (increase variance).

Binomial Probability Distribution

The binomial probability distribution is the frequency distribution of the "success" outcomes of a set of *n* probabilistic trials (e.g., flipping a coin) each having constant success probability of *p*. Success, or 1 (with 0/1 designations), is an arbitrary designation and is usually the class of interest or the more unusual class.

Black-Box Models

Black-box machine learning models generate predictions without yielding usable information about the relationships between predictor values and outcomes. Neural networks and ensembles of trees are examples of black-box models.

Central Limit Theorem

The central limit theorem states that the distribution of a sample mean gradually becomes normally shaped as the size of the samples (all drawn from a common population) increases, even if the observations from the common population is not normally distributed. It was very important in the early days of statistics when it enabled formula alternatives to the resampling and permutation procedures that made up the foundation of inference.

Character Data

Character data do not convey any quantitative information, such as letters, punctuation, or even numbers that are part of a text.

Coefficient of Determination

The coefficient of determination, or R^2, is a measure of the proportion of variation in data that is explained by a linear model.

Collectively Exhaustive Events

In probability theory, events are collectively exhaustive if at least one of them must occur. In ice hockey, for example, a team's outcome must be one of three events: win, loss, or tie.

Conditional Probability

Conditional probability is the probability that some event occurs given that some other event or condition has occurred. For example, the probability that a web visitor purchases a product during a session, given that they have purchased a product in the past.

Continuous Values
Where measurements can take any numerical value in a relevant range, those values are termed continuous values. Depending on the precision of a thermometer, for example, a temperature might be measured at 12 degrees, or 12.5, or 12.39, or 12.387.

Correlation
Correlation between two variables exists when they move in tandem: one goes up and the other goes up, and vice versa. Often, "correlation" is used synonymously with "correlation coefficient," which is a measure of the degree of linear correlation that ranges from +1 (perfect positive correlation) to −1 (perfect negative correlation).

Cost Function
See **Loss Function**.

Counterfactual
A counterfactual in machine learning is a modification to data that is used to explore model behavior, for example "suppose the input had been x2 instead of x1: how would the output change?" Counterfactuals are used in interpretability methods for black-box models.

Counterfactual Thinking
In social psychology, counterfactual thinking is the human tendency to imagine or explore alternative paths of possible prior events.

Covariance Stationary
A time series is covariance stationary if it does not change over time. Specifically, its mean, variance, and autocorrelation remain constant.

Data Drift
Machine learning models are trained on a given set of data. After the model is deployed in a pipeline with new data, the characteristics (e.g., the mean or variance) of the input data may change, in which case, the model may no longer be appropriate and require retraining.

Dependent Variable
In a machine learning algorithm, the dependent variable is the variable to be predicted. Also termed the response, target, or outcome.

Deterministic Models
A deterministic model is one that has all the information needed to explain or predict an outcome with certainty, and it does not explicitly incorporate random behavior. Determining the volume of a sphere if you know its radius is a deterministic model. A model might be effectively deterministic—for example, determining stopping distance using full braking power for a car traveling a certain speed, even though minor random elements might be involved. The application of a deterministic model to a situation that clearly involves random behavior (e.g., using a linear model to predict sales) imparts a false degree of confidence.

Discrete Values
Where quantitative measurements can take specific, usually evenly spaced values, those values are termed discrete values. The number of occupied seats on an airplane is a discrete value (e.g., 76 seats might be occupied, or 77, but not 76.5).

Dummy Variables
Dummy variables are 0/1 variables created to record the presence or absence of a category. For example, a stock might be small cap, mid cap, or large cap. Most machine learning algorithms require numeric inputs, so this single categorical variable might be recast as three binary dummies to indicate whether that category applies: small cap (0/1), mid cap (0/1), and large cap (0/1).

Empirical Probability
An empirical probability is a probability estimate based on prior data. For example, the probability that a company's overall market cap will rise more than 1% following a stock split is best estimated by what has happened in similar circumstances in the past. This is distinguished from a theoretical probability based on physical knowledge and assumptions: the probability that a coin flip will land heads can be estimated based on knowledge of the coin and the flipping process, without having to flip thousands of coins.

Entropy
In information theory, entropy is randomness or, particularly with respect to the transmission of information, confusion, or disorder. In machine learning, entropy means heterogeneity or impurity in outcomes. It is a popular cost function (or loss function) in decision trees, where the more homogeneous the classes in a node, the better (the tree's rules are doing a good job separating the classes). In a classification tree, entropy ranges from 0 (completely homogeneous) to 1. It is calculated as the sum of $-p_k \ln(p_k)$ across the outcome classes k, where p is the proportion of records belonging to class k.

Explained Variation
The effectiveness of statistical and machine learning modeling can be measured by the extent to which a model, with its predictions, can match the variation in the independent (outcome) variable. Metrics, such as R^2, measure this effectiveness: $R^2 = 0$ means that none of the variation in the outcome variable is explained by the model, while $R^2 = 1$ means that all of the variation is explained by the model.

F-Distribution
The F-distribution is a probability distribution of the ratio of variances of two samples drawn from the same (normal) population. It is used in testing null hypotheses about whether variances are equal. It is also used in analysis of variance (ANOVA) procedures in which the goal is to determine whether two or more samples of numerical data are statistically different from one another.

Features
Features are measurements of characteristics in a dataset that vary from record to record and are useful in prediction. Synonyms for "features" are as follows: predictor variables (from statistics), fields (from database management), columns, and attributes. Some examples of numerical features are as follows: sales, number of customers, and assets. Some examples of categorical features (which are often binary) are as follows: purchase/don't purchase, survive/die, level of education, and low/mid/large cap.

Gradient Descent
Machine learning algorithms typically seek to minimize predictive error or a related loss function, and neural networks do so through some form of iterative trial and error. Gradient descent is one popular form of this trial-and-error process. It can be likened to hiking over hills and valleys in search of the lowest point, following a rule that says to proceed in whatever direction takes you downhill from your current position. This does not guarantee a global minimum, but the algorithm can be adapted so that it is not necessarily stopped by a local minimum.

Heteroskedasticity
Heteroskedasticity is when the variance of a variable changes over different ranges of the data. For example, wages may vary only a little for people with a high school education but may vary considerably more so for people with a graduate education.

Homoskedasticity
Homoskedasticity is when the variance of a variable changes little over different ranges of the data.

Independent Events
Events are independent if knowing the outcome of one event does not affect the probability of the other. Quarterly advertising revenue for Meta is probably independent of quarterly deaths from infectious diseases. But it is not independent of quarterly change in gross domestic product.

Independent Variable
An independent variable is a variable (also called a feature or attribute) that is included as a predictor in a statistical or machine learning model. The term "independent variable" is used in statistics, and terms like "feature" or "attribute" tend to be used more in machine learning.

Labels
In machine learning, labels are another term for "outcome variable," particularly in cases in which the value must be assigned by human review (e.g., pictures of cats or dogs in training data).

Linear Regression
Linear regression is a statistical procedure that estimates a linear relationship between predictor variables (also called features or independent variables) and a numerical outcome. It typically does so by minimizing the squared errors or residuals (i.e., the differences between predicted and actual values).

Logistic Regression
The goal of a logistic regression is to predict the probability of being a 1 in binary data. A linear model is used, but the output must be scaled so that it lies between 0 and 1. This is done by using the logs of all terms and fitting a standard linear model. After the model is fit, the coefficients can then be exponentiated so that relationships between predictors and the output can be interpreted. The output is now in terms of odds, which can be converted to probability.

Lognormal Distribution
The lognormal distribution is a distribution of data that is long-tailed to the right. When the logs of the data in a lognormal distribution are taken, the resulting transformed distribution becomes normally distributed.

Loss Function
A loss (or cost) function is a measure of the degree of error in a machine learning algorithm's prediction, compared with actual values. Simple loss functions are sum of squared errors (for numeric data) and percent misclassified (for categorical data).

Multicollinearity
Multicollinearity is when two or more predictors (features, attributes) are correlated with one another. Multicollinearity in linear models can cause the model-fitting process to fail; one or more of the highly correlated predictors should be dropped. Because of the high correlation, not much information will be lost.

Mutually Exclusive Event
Mutually exclusive events are events that cannot coexist or happen together, by definition. "Quarterly profit rises" and "quarterly profit falls" are mutually exclusive events.

Nonlinear Relationship
A nonlinear relationship is one that cannot be described by a linear function. Some nonlinear relationships become linear after a transformation of the data (e.g., by taking the log).

Nonparametric Hypothesis Test
A nonparametric hypothesis test does not rely on the data being normally-distributed. Most resampling procedures are nonparametric. The assumptions about normally distributed data that accompany standard formula-based hypothesis tests help approximate the distribution of a test statistic under resampling.

Normal Distribution
The normal distribution is a symmetric bell-shaped curve. Its formula is not intuitive, but you can think of the normal distribution as the approximate result of a huge number of marbles being dropped down a pinball board where, at each stage, the marble must bounce left or right, before finally coming to rest in a set of columns. The columns at the extremes have only one path to them (repeated bounces left or right), and hence are short. The columns at the center have many arrival paths, and hence are taller.

Normalization
See **Scaling**.

Odds Ratio
Odds are familiar to gamblers, being the comparison of one probabilistic event to another. For example, a roulette wheel has 38 slots: 18 red, 18 black, and 2 green (which cannot be selected as one of the betting options). The odds of winning a red bet are 18 to 20 (or 18:20), meaning that if you were to play indefinitely you would lose 20 plays for every 18 that you win. This translates to a probability of 18/38 or 0.474. The odds can also be expressed as a ratio of probabilities, or p/(1 – p), or, for the roulette example, 0.474/0.526 = 0.901. An odds ratio is simply a ratio of two odds—for example, the odds of getting cancer given exposure to chemical X, divided by the odds of getting cancer without that exposure.

Optimization
Optimization is the process of "making the best possible." The "possible" part of the definition is important, as optimization almost always involves working within some limits. In business, it is typically maximizing something (profit) or minimizing something (cost), subject to constraints. In machine learning, one typically optimizes the performance metric (loss or cost function).

Ordinal Scale
Ordinal data can be ordered or ranked by magnitude but are not necessarily arithmetically operable. In ordering maximum attained education level as 1, 2, 3, and 4, corresponding to elementary, secondary, college, and graduate, the most you can say is that each level is more than the preceding. You cannot say, for example, that level 4 is twice level 2.

Outliers
Outliers are values that are distant from the bulk of the data in a dataset. The definition of "distant" is arbitrary, although statistical rules of thumb are used. The handling of outliers depends on the circumstances. Sometimes they signal a need for review and possibly signal removal as a distraction or confirmation as valid. Sometimes they are just what you are looking for (e.g., in anomaly detection).

Overfitting
Overfitting in machine learning results when a complex model is trained on data in such a way that it performs "too well" on the training data by virtue of learning not just the signal but the noise as well. That same random noise will be different in new data, so the model's learnings will not only be useless but counterproductive when deployed: interpreting random noise as real patterns will introduce error. Data-centric models such as decision trees and neural networks are especially susceptible to overfitting unless those models are simplified or curtailed.

Pearson Correlation Coefficient
The Pearson correlation coefficient measures the degree of linear association between two numeric variables, and ranges between −1 (perfect negative correlation) and +1 (perfect positive correlation). It is calculated by averaging the product of the deviations from the mean for each variable, and then dividing by the product of their standard deviations:

corr = [mean of $(x_i - mu_x)(y_i - mu_y)$]/(stdev x)(stdev y).

Poisson Distribution
The Poisson distribution is the frequency distribution of the number of events per interval of time or area of space, when there is a constant rate of random events per time or space units. For example, the number of people arriving at a queue in a specified time span when the flow of customers is an average of 20 per hour, or the number of insects per square meter of beach when there are, on average, three per square meter.

R^2
R^2, or the coefficient of determination, is a measure of the proportion of variation in data that is explained by a linear model.

Random Walks
A random walk is a time series in which movements from one period to the next are random. If a time series is not distinguishable from a random walk, it is not likely to be predictable.

Sample
A sample is a set of observations or records from a larger dataset. Often the sample is drawn randomly. In the machine learning community, it sometimes has a different meaning, especially when it is associated more with computer science than statistics. In this context, the term "sample" is often used to refer to one record or observation.

Scaling
Scaling is a data transformation that facilitates placing multiple variables on the same scale, so that their values fall roughly within the same range. Some machine learning algorithms fail if, for example, some variables are in the millions and others range between 0 and 1. There are several scaling options, including standardization (subtracting the mean and dividing by the standard deviation) and 0–1 scaling (i.e., "shrinking" the data so that the maximum value is 1 and the minimum is 0).

Serial Correlation
Serial correlation is the presence of correlation between successive values in a time series.

Simple Linear Regression
Simple linear regression is a model that expresses one variable (the dependent variable) as a linear function of another variable (the independent variable).

Skewness
Skewness is the extension of one tail of a probability distribution farther out than the other. If the right tail is longer, it is referred to as "right skew," and if the left tail is longer, it is referred to as "left skew."

Spearman Rank Correlation Coefficient
The Spearman rank correlation coefficient for two variables is found by converting the values in each variable to ranks, and then calculating the Pearson correlation coefficient.

Standard Normal Distribution
The standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1.

Standardization
See **Scaling**.

Statistical Regularity
Statistical regularity describes randomness in a process that becomes predictable if the process continues long enough. The proportion of heads in 5 tosses of a coin is not predictable, but the proportion of heads in 5,000 tosses is quite predictable.

T-Distribution
The t-distribution is a probability distribution that is symmetric, bell-shaped, and like the normal distribution but with fatter tails. It is used in formulas to perform the t-test for A/B and other hypothesis tests. It is really a family of distributions whose shape and parameters depend on the sizes of the samples in the tests.

Test Statistic
A test statistic is a statistic (metric) applied to samples in a hypothesis test. For example, in an A/B test of two web page designs, a relevant test statistic might be "proportion of clicks."

Time Series
A time series is a plot or table of successive values of a variable over time.

Time-Series Model
A time-series model seeks to explain the pattern in a time series and/or predict future values.

Training Set
Training data are data used to fit (train) a machine learning model. They are separate from holdout or validation data, where the model is used to make predictions and the accuracy of those predictions can be measured.

Transformation
Transformation of data is consistent modification of all the values of a variable to render them more tractable for analysis. For example, data that seem to show an exponential trend can be transformed by taking the log of all values, after which a linear analysis can be performed.

Unconditional Probability
Unconditional probability is a probability (usually an estimate) that an event will occur, without considering any other factors or conditions. For example, the probability that any public firm will declare bankruptcy, without taking financial history into account, is an unconditional probability. The probability that a firm will declare bankruptcy, given that they have run losses for three years is a conditional probability.

Underfitting
Underfitting in a machine learning model is when the model underperforms because there are some patterns or relationships that it is failing to capture effectively.

Unexplained Variation
Unexplained variation in a target or outcome variable refers to that variation that is not explained by a model. For example, in a revenue prediction model, the differences between predicted revenue and actual revenue constitute unexplained variation.

Validation Data
In a machine learning process, a model typically learns from training data, and is then applied to validation or holdout data, to see how well the model predicts.

Variance Inflation Factor
The variance inflation factor (VIF) is used to detect multicollinearity among predictors in a (usually linear) model. It is calculated for an individual predictor by regressing that predictor against all others and finding the R^2 for that model. If that R^2 is high, then that predictor is largely explained by the other predictors, and there is likely to be multicollinearity. The VIF formula for a given predictor i uses the R^2 from this model:

$$VIF_i = 1/(1 - R^2_i).$$

Venn Diagram
A Venn diagram is a two-dimensional visual depiction of possible outcomes of an event or categories of a feature, in which a rectangle or circle represents a category or outcome, and the interior of the shape represents the observations that belong to that outcome or category. Typically, the shapes intersect to some degree and the space in the intersection area represents observations that belong to all the categories or outcomes that the shapes enclosing that intersection space represent.

Course 3 – Machine Learning

Accuracy
In a machine learning model to predict categorical data, accuracy is the proportion predicted correctly.

Agglomerative Clustering
Agglomerative clustering is a clustering algorithm that proceeds sequentially by grouping together nearby clusters. At the initial stage, each record is its own singleton cluster, and then the two closest records are joined into a cluster. At each successive stage, the two closest clusters are joined together.

Bagging
Bagging, or bootstrap aggregating, is the process of repeatedly taking bootstrap samples of records and repeatedly applying a machine learning algorithm to each sample. The results are then averaged, or a majority vote is taken (for categorical outcomes), yielding predictions that are more stable than those of individual models.

Bayesian Classifier (or naive Bayesian classifier)
In principle, an exact Bayesian classifier would classify a record as follows: (1) find the records whose predictors have the same values as the record to be classified, and (2) find out what class is dominant (or, if the outcome is numerical, find the average). Of course, with most data, it is unlikely that you will be able to find exact matches. The "naive Bayes" shortcut drops the requirement for an exact match and works with the independent probabilities (by class) of the individual predictor categories.

Bias Error
Bias error is consistent predictive inaccuracy in one direction or the other and is not the result of random variation.

Binary Classifier
A binary classifier is a machine learning algorithm that predicts the class of a record, where the record can belong to one of two classes (e.g., a web visitor could click or not, an account could be current or not).

Bootstrapping
Bootstrapping in statistics is the process of repeatedly taking samples with replacement from a dataset and making an estimate from, or fitting a model to, each bootstrap sample. Bootstrapping is used to estimate variability, and it also facilitate methods to avoid overfitting in machine learning models.

CART
Classification and regression trees (CART) are also called decision trees. The CART algorithm repeatedly divides the dataset into partitions that are increasingly homogenous in the outcome variable, resulting in a set of decision rules that predict a class or numerical value. For example, "If a web shopper has made a prior purchase AND the shopper spends more than two minutes on the site AND the shopper clicks to view a product, THEN predict 'purchase'."

Categorical Variable
A categorical variable can assume one of several values that denote a category (e.g., a loan could be current, delinquent, or in default).

Centroid
The centroid of a dataset is the vector of the mean values of all the variables.

Classification Tree
See **CART**.

Clustering
Clustering is the process of joining together similar records into a smaller set of clusters. The optimal number of clusters is determined by the analyst. In technical terms, you seek clusters that are distinct (distant) from other clusters, and whose members are close to one another. The organizational use requirements are also considered: for example, for customer clusters (segments), the sales process might deal well with 3 or 4 clusters, but not 20 or 30.

Covariance Matrix
For multivariate data, a covariance matrix lists all the variables both as row and column headers. Each cell is the covariance of the row and column variables.

Cross-Validation
Cross-validation is the repeated splitting of data into model-fitting and model-assessment subsets. For example, a model might be fit to 80% of the data, and then assessed on the remaining 20%. This process is iterated, with a different 20% being held out for assessment each time (or "fold"). The 80/20 split would constitute five-fold cross-validation.

Cutoff Value
A classification model generates an estimated probability (propensity) that a record belongs to a particular class. The model algorithm, or the analyst, can set a cutoff value (threshold value) for that probability to distinguish records that

are classified as a 1 (i.e., those records with a probability above the cutoff) from those classified as a 0.

Decision Node
In a decision tree, a decision node is a split in a branch. The rule governing the split allocates records to one subbranch or another for further splitting.

Deep Learning
Deep learning is machine learning using deep, or complex, neural networks. Deep neural nets (with many nodes and multiple layers), coupled with lengthy training, are especially good at making sense of unstructured data (images and text) that lack the high-level features found in tabular data.

Dendrogram
A dendrogram is a visual depiction of a hierarchical clustering process. Distance between clusters is represented on the y-axis, and clusters along the x-axis. When two clusters are joined, they are shown on the plot with two vertical lines linked by a horizontal line at their between-cluster distance.

Dimension Reduction
Too many features (dimensions) impair the performance of a machine learning algorithm by adding random noise associated with features that lack predictive value. Dimension reduction is the process of reducing the number of features in a model, eliminating those that are not useful.

Divisive Clustering
Divisive clustering is the process of creating clusters (groups of records similar to one another) by repeatedly dividing the data into subgroups, increasing within-cluster homogeneity as you go.

Ensemble
An ensemble in machine learning is a collection of machine learning algorithms. Predictions from a set of multiple machine learning algorithms, using the average prediction (for numerical outcomes) or the majority-vote prediction (for classification), are usually more accurate than those based on the individual algorithms themselves. This is sometimes referred to as "the wisdom of the crowd," and it is particularly useful when you are aggregating "weak learners," or algorithms with only modest predictive power for a given task.

Error Rate
The error rate is the proportion of records that are misclassified in a machine learning model.

Euclidean Distance
The Euclidean distance between two records is the square root of the sum of the squared differences between them.

Forecasting
Forecasting is the prediction of future values in a time series.

Hidden Layer
In a neural network, a hidden layer contains nodes whose output is then fed into other nodes in a subsequent layer.

Hierarchical Clustering
Hierarchical clustering is the process of sequentially creating clusters, either by repeatedly joining records and then clusters together (agglomerative methods), or by repeatedly dividing the data (divisive methods). The end result is a set of nested clusters, from a low level at which there are many small clusters (each record its own cluster in the extreme case) to the top level at which there is just one cluster (i.e., the entire dataset).

Information Gain
Many machine learning performance metrics are based on prediction errors (losses or costs). Another way of measuring performance is to look at information gain, which is the decrease in some metric of loss or error (e.g., entropy).

In-Sample Error
In-sample error is a machine learning algorithm's error (i.e., the difference between predicted and actual values) in predicting training data.

Iterative Procedure
An iterative procedure is one that is repeated over and over or "iterated," often incorporating the results of prior rounds.

K-Nearest Neighbor
k-Nearest neighbor is a prediction algorithm that assigns to a record the same class (or average value) as a certain number (k) of records that are the closest to it (neighbors). K is set by the user, and Euclidean distance is a popular metric for measuring closeness.

K-Means Clustering
In k-means clustering, the analyst specifies in advance the number, k, of desired clusters. After an initial (usually random) division into k clusters, records are iteratively reassigned to the closest cluster. The process stops when the only available reassignments increase within-cluster dispersion.

Labeled Dataset
A labeled dataset is one that has records whose outcomes are known, usually as a result of a human review that assigns a label (e.g., an insurance claim could be fraudulent or normal).

LASSO
LASSO (least absolute shrinkage and selection operator) is a regression method that penalizes regression coefficients in a way that causes noninfluential predictors to drop from the model. This decreases variance in the coefficient estimates and yields a model that is more parsimonious.

Market Impact Analysis
In finance, a market impact analysis is the assessment of how much a purchase or sale of securities affects the security price.

Mean Squared Error
In machine learning, mean squared error (MSE) is the average of the sum of squared errors (where errors are the differences between actual and predicted values).

Multilayer Convolutional Neural Network
A convolutional neural network (CNN) is a deep neural net that adds convolutions to its learning process. A convolution is an operation applied to multiple observations at the same time, and it helps uncover higher level features that are not discoverable at the granular level. In processing images, for example, operations can be applied to small matrices of pixels, moving sequentially across the image. These operations help reveal basic higher level features (e.g., edges or lines) that are not discoverable at the individual pixel level but appear when you look at multiple pixels.

Natural Language Processing
Natural language processing (NLP) is the application of machine learning methods to natural languages, like English. The goal may be classification, extraction of meaning, or generative (producing content). The older term "text mining" is sometimes used synonymously.

Neural Network
A neural network is a machine learning algorithm that performs numerous mathematical operations (both parallel and sequential) on input data, generating output that is interpreted as predictions. Initially, the predictions are essentially random, but they are compared to actual values, and the weights that govern the mathematical operations are modified in additional passes through the data to improve the predictions.

Node
In a neural network, a node is a location where weights are applied to inputs to generate an output: either an intermediate output that becomes an input to another node, or a final output that becomes a prediction, which, in turn, is either final, or provisional (if the latter, it is compared to the actual known value).

Noise
Noise in data and models is variation that is not explainable by a statistical or machine learning model.

Nonparametric
Nonparametric statistical methods or estimates do not incorporate assumptions about a normal (or other) data distribution. Such assumptions reduce input data to parameters (e.g., mean, variance) that are often required for it to be used in many mathematical formulas.

Out-of-Sample Error
In implementing machine learning models, a subset of the data is typically held out (holdout or validation data) to allow the model to be applied to data that were not used to train the model. The model's error on these out-of-sample data is a more unbiased estimate of how the model will perform than the error on the training data.

Output Layer
In a neural network, the output layer contains a node whose output is the prediction (either final or, if there are to be further passes through the data, provisional).

Parsimonious Models
Parsimonious machine learning models are those that use only essential and useful predictors and records. Including variables and records that do not provide information that is useful for making predictions adds noise and degrades performance.

Partition
In machine learning, data are typically split into two or three subsets. The "training" partition is used to train or fit the model. A second partition is used to assess the model and tune (adjust) its parameters. A third partition may be used at the end to estimate how well the model will do in the real world (no further model tweaking is allowed after this). The second partition is sometimes called the validation partition and may also be called a "holdout" partition. The third partition is also sometimes called the "holdout" partition and is also called the "test" partition.

Penalized Regression
Standard linear regression may include predictors that introduce instability, reflected in high variance for the coefficients. This can result from multicollinearity, from the presence of noise, or when there are simply too many variables. Regression models can be improved by penalizing coefficients if their variables do not contribute stable and useful information.

Principal Component Analysis
Principal component analysis (PCA) transforms multivariate data into a limited set of variables, each of which is a linear combination of others. The first such variable or component seeks to capture as much of the variability in the data as possible. This makes it possible to create a parsimonious machine learning model that uses just a handful of these newly created variables.

Pruning
If allowed to grow indefinitely, decision trees overfit the training data and do not perform well with new data. An optimal tree is one that is not fully grown. One way to achieve smaller trees is to allow them to grow fully, and then to prune them back to a point at which they have minimal error on validation data.

Random Forest Classifier
A form of bagging, a random forest is a collection of decision trees that are grown using a set of bootstrap samples. The bootstrapping, or bagging, is done by taking random samples, with replacement from the records; random selection (without replacement) is also applied to the predictors at each stage. The results are then averaged, or a majority vote is taken (for categorical outcomes), yielding predictions that are more stable than those of individual models.

Regression Tree
A regression tree is a decision tree where the outcome is numerical. See **CART**.

Reinforcement Learning
Reinforcement learning is a machine learning method in which the algorithm interacts with people (e.g., shoppers at a website) and, from their reactions, learns the optimal assignment of options.

Semi supervised Technique
In some machine learning situations, labeled data are scarce or expensive (e.g., insurance claims, which require a human review to determine validity), but there is a lot of unlabeled data. In semi supervised learning, an initial model is trained on a limited set of trained data, and then is applied to the unlabeled data to generate predictions. The predictions that have the most confidence (i.e., have the highest probability attached to their classification by the model) are then added to the original labeled data as "pseudo-labeled data," and a new model is trained. This process is then repeated.

Sentiment Analysis
Sentiment analysis is the analysis of human-generated text to classify it to a sentiment (e.g., positive/negative, enthusiastic/indifferent).

Supervised Learning
Many machine learning models have as their goal the prediction of something, such as whether a web visitor will click on a link, how long a patient will need to be hospitalized, or whether a firm presents a financial risk to stakeholders. Models are "trained" on data for which the outcome is known, and then are applied to data for which the outcome is not known. This process is termed "supervision."

Tensor
A tensor is a multidimensional matrix map. A scalar is a tensor of rank 0, a vector is a tensor of rank 1, a two-dimensional matrix is a tensor of rank 2, and so on. Tensors are central to the ability of deep neural nets to cope with complex data and features that are at a higher level than the input values (e.g., shapes instead of pixel values); a popular deep learning environment is termed "TensorFlow."

Terminal Node
In a decision tree, a terminal node is a final split in a branch. The records resulting from the terminal node split are assigned classes or values, and they are not split further.

Threshold Value
See **Cutoff Value**.

Unsupervised Learning
Unsupervised learning methods are machine learning methods that do not involve the prediction of an outcome variable. Clustering and recommendation engines are examples of unsupervised learning.

Course 4 – Natural Language Processing

Alternative Hypothesis
In a statistical hypothesis test, the alternative hypothesis is the proposition you hope to demonstrate, for example, that a treatment (e.g., a new drug, or a different web page) had an effect beyond what chance might produce.

Area Under the Curve (AUC)
The area under the curve (AUC) is a metric that measures how effectively a predictive algorithm separates two classes. The curve in question is the ROC curve, and an area of 1 indicates perfect separation, while an area of 0.5 indicates that the model does no better than chance. See **Receiver Operator Characteristics Curve**.

Bag of Words
In natural language processing, the bag of words technique treats a document as simply a collection of terms, without regard to order.

Binary Classification
Binary classification algorithms predict classifications for records that have two (binary) outcomes. Typically, the class of interest is in the minority (e.g., a sale at a website, the presence of a disease, a fraudulent tax return), and is labeled 1, while the other records are labeled 0.

Chi-Square Test for Independence
The chi-square test of independence tests whether two or more samples differ in the proportion of 1's to a greater extent than chance would allow, if the samples were drawn from the same population with a common proportion of 1's. The test derives from a chance model in which all samples are pooled, are dealt out again in random order, and the 1's in each sample are counted.

Complexity in a Model
In machine learning, a complex model is contrasted with a parsimonious one. A complex model will have many variables and hyperparameters. It may fit the data better, but it is vulnerable to overfitting and misinterpreting noise as signal.

Confusion Matrix
A confusion matrix for a classification model is a 2 x 2 table of predicted versus actual outcomes. The columns typically represent predictions (1 or 0), and the rows actual classes (1 or 0). The upper left cell, for example, represents the number or records predicted as 1's that are, in fact, 1's. Numerous metrics can be calculated from the confusion matrix, including accuracy rate, error rate, sensitivity (recall), specificity, false discovery rate, and false omission rate.

Constraint
In optimization, constraints (e.g., limited supplies of labor and capital) are almost always present and are an integral part of the optimization algorithm.

Cross-Sectional Data
Cross-sectional data are measurements of many different entities (e.g., people in a medical study, web visitors, emails) at the same time, or in contexts in which time is not a factor.

Curating Text Data
Text data encompasses a huge variety of topics and types, such as scholarly articles, tweets, news items, maintenance reports, medical notes, and dictionaries. A curated corpus (body) of text is text from one topic or type that has been selected, reviewed, and processed to a sufficient degree that it can be subjected to analysis.

Data Frame
A data frame is a 2 x 2 tabular set of data, like a spreadsheet range, with records as rows and columns as variables.

Data Mining
Data mining is the science of gaining insight from data and finding patterns. It encompasses techniques like predictive modeling, clustering, principal component analysis, and more. The term has largely been superseded by the term "machine learning."

Data Wrangling
Data wrangling is the process of obtaining raw data and preparing it for analysis and modeling. Obtaining the data, dealing with missing values, understanding variables, reconciling different sources, and achieving consistent formatting are all part of data wrangling.

Document Frequency
In natural language processing (NLP), document frequency is the number of documents in which a term appears.

Document Term Matrix
In natural language processing (NLP), the document term matrix is a matrix in which rows are documents, columns are terms, and the entries are the counts of a given term in the document for that row.

Entity Recognition
In natural language processing (NLP), entity recognition is a technique for identifying named entities (e.g., the names of individuals or organizations) in a set of documents.

Entity Resolution
Entity recognition (or entity resolution) is a method for identifying matches in a set of records of entities. For example, if a person purchases something at a web retailer, the company would like to know whether this is an existing customer or contact. Sometimes the customer will provide exactly the same information as the company has on file and the match is easy, but more often, there may be some differences (e.g., missing information, spelling variations, the person moved to a new address). Entity resolution can provide an estimate of the probability of a match.

Exploratory Data Analysis
Exploratory data analysis consists of various techniques for human review and understanding of a dataset. Descriptive statistics (e.g., mean, median, variance, outliers) are included, as are visualizations, such as scatterplots, histograms, and boxplots. Exploratory analysis typically precedes automated modeling.

False Negative
A false negative is a record that is classified as a 0 but is actually a 1. See also **Binary Classification** and **Confusion Matrix**.

False Positive
A false positive is a record that is classified as a 1 but is actually a 0. See also **Binary Classification** and **Confusion Matrix**.

Feature Engineering
Preparatory to fitting a statistical or machine learning model, some work is typically done on the features (predictor variables). This includes understanding their meaning, standardization (normalization), binning of numerical variables if appropriate, consolidation of categories if there are too many, deriving or finding new features, imputing missing values, and more.

Feature Selection
Feature selection is the process of selecting the features (variables) to be used in a machine learning model. Often this involves distinguishing between the variables to be included in a model and those to be left out.

Grid Search
Most machine learning models have hyperparameters (settings) controlled by the user, such as the number of hidden layers in a neural network. One way of finding out which setting is best is to try multiple combinations and assess model performance, and then use the hyperparameter combination that works best. This process is a grid search.

Model Tuning
Model tuning is the process of setting model hyperparameters—for example, the number of layers in a neural network or how far to grow the decision trees. Tuning can be done manually by the user or can be done automatically.

N-Gram
An N-gram is a sequence of N terms (or other units of text) in a document. For example, "and seven years ago" is a four-gram (one of many four-grams) from President Lincoln's Gettysburg Address. N-grams are a useful unit of analysis in natural language processing (NLP). You can see them in action when you type and your software suggests corrections: for example, if you type "the output of a *dep* neural net" in a paper on machine learning, the software will reference five-grams or six-grams containing similar correctly spelled phrases and suggest "deep" as the correction.

Parsimony in a Model
In machine learning, a complex model is contrasted with a parsimonious one. A parsimonious model will have relatively few variables and hyperparameters. Other things being equal, a parsimonious model is preferred to a complex one.

Penalized Regression
In penalized regression, penalty terms are attached to regression coefficients, causing the least squares algorithm to "shrink" (regularize) them. While this introduces bias into the coefficients, this bias is offset by a reduction in variance of estimators and an increase in model simplicity and interpretability.

Porter Stemmer Algorithm
The Porter stemmer algorithm in natural language processing (NLP) is a set of rules for removing suffixes and reducing words to a common stem. See also **Stemming**.

Precision
Precision in a machine learning classifier is the proportion of predicted 1's that turn out to actually be 1's. See also **Confusion Matrix**.

Random Sampling
Simple random sampling is the process of selecting an observation from a dataset in such a way that each element in the dataset has an equal probability of being selected. Other (non simple) forms of sampling incorporate randomness but lack this strict condition (e.g., stratified sampling).

Recall (Sensitivity)
Recall in a machine learning classifier is the proportion of actual 1's that are correctly classified as 1's. See also **Confusion Matrix**.

Receiver Operator Characteristics (ROC) Curve
The receiver operator characteristics curve is the plot of a machine learning algorithm's predictions, specifically recall (the y-axis) and specificity (the x-axis, with 1 on the left and 0 on the right) as you change the cutoff for classifying a record as a 1. The name arose in World War II to describe the performance of radar receiving stations, whose job was to determine whether radar blips were incoming aircraft.

Regularization
See **Penalized Regression**.

Scraping
Scraping is the automated collection of data from websites.

Sensitivity
See **Recall**.

Shrinkage Models
See **Penalized Regression**.

Specificity
Specificity in a machine learning algorithm is the proportion of actual 0's that are correctly classified as 0. See also **Confusion Matrix**.

Stemming
The huge scale of natural language vocabularies poses a challenge for machine learning methods of natural language processing. Stemming (also called lemmatization) reduces multiple variants of a word to just one word for analysis purposes (e.g., "departure, departed, departing" all become "depart").

Stop Words
In natural language processing (NLP), stop words are terms in a document that are dropped before analysis. NLP programs typically offer a generic list of stop words like "the," "an," and "to" that do little to help classification and whose removal simplifies analysis. Usually, modifications to the default generic list, or substitution of a custom list, are permitted.

Term-Frequency Inverse Document Frequency (TF-IDF)
TF-IDF helps identify document-term pairs that are unusual, specifically where a term that is relatively rare overall appears in a document. TF-IDF yields relatively high values in such a case, and low values when the term is either absent from the document or relatively common overall. Definitions of TF-IDF can vary slightly, but according to a common definition, it is the product of two things:

- term frequency (TF) = # times term t appears in document d, and
- inverse document frequency (IDF) = log(total # docs)/(# docs with term t) + 1.

Term-Frequency Matrix
The term-frequency matrix for a set of documents summarizes how often different terms appear. Columns are terms, rows are documents, and the entry for a given row and column indicates how often the term for the column appears in the document for the row.

Tokenization
Tokenization is the process of taking a text and, in automated fashion, dividing it into separate terms or "tokens," which become the basic unit of analysis. A word separated by spaces is a token, "2 + 3" would divide into three tokens, and "23" would be one token. NLP software comes with default tokenizers, whose rules can be altered.

True Negative
In binary classification, a true negative is a 0 that was correctly classified as a 0. See also **Binary Classification** and **Confusion Matrix**.

True Positive
In binary classification, a true positive is a 1 that was correctly classified as a 1. See also **Binary Classification** and **Confusion Matrix**.

Type I Error
In statistical hypothesis testing, a type I error is mistakenly accepting the alternative hypothesis that an effect or phenomenon (e.g., a new therapy improves health, or a stock price is not a random walk) is real. In other words, a type I error is being fooled by chance patterns into thinking that something interesting is going on.

Type II Error

In statistical hypothesis testing, a type II error is mistakenly accepting the null hypothesis that an effect or phenomenon (e.g., a new therapy improves health, or a stock price is not a random walk) is not real and is just the product of chance. A type II error (which can occur only when the effect is real) usually results from an inadequate sample size.

Unstructured Data

Unstructured data are data that do not occur naturally in the form of a table with rows and columns. The text in documents, images, and "messy" data (e.g., medical notes that have a mixture of numeric and text data) are all unstructured data.

Course 5 – Mitigating Biases in the Data Science Pipeline

Accountability
The principle suggesting that a set of rules and responsibilities must be established to identify who is accountable in case the data science tool goes wrong, who is able to explain the model assumptions and results, and who controls the decisions based on the quantitative models' output.

Accuracy
The ratio of all correctly predicted classes to the total number of predictions.

Acquiescence Bias
A bias that arises when participants respond to a survey question in a manner that is not consistent with their position.

Adaptive Windowing (ADWIN)
An algorithm that uses moving time-period windows, comparing the means in the moving windows, and dropping older windows when their means are significantly different compared with those in the more recent windows.

Alpha
Manager skill or the premium return over market averages that cannot be explained by common factor betas, especially that of market beta.

Back-Propagation
The process of tuning the weights in a neural network process to reduce error.

Baseline Model
The model produced using the training dataset.

Bias
A measurable, systematic difference between actual results and the correct results in statistics, *or* a belief or assumption that is implicit or explicit.

Bias Error
The degree to which the model fits the training data, where the lower the bias, the better the model's performance in the training dataset.

Behavioral Bias
Typically, cognitive, or emotional bias in decision making that leads to errors in judgment. Personality difference can exert a relatively huge influence on its overall impact.

Binning
A grouping of continuous observations into ranges.

Black-Box Model
Adjective to describe the opacity of a model framework.

Boosting
The sequential learning of models, where each new model learns from the errors of the previous model.

Bootstrap Aggregating
A technique whereby the original training dataset is used to generate *n* new training datasets or bags of data (also called bagging).

Bootstrapping
The repeated resampling of data with replacement from a dataset.

Cause–Effect Bias
Also known as correlation bias, occurs when the correlation is mistaken for causation.

Characteristic Stability Index (CSI)
A metric for comparison of the distributions of the features in the training data with that of the deployment data.

Citation Bias
A potential bias when there is a reliance on specific research rather than the entirety of the extant published research.

Classification and Regression Tree (CART) Algorithm
A supervised machine learning algorithm based on decision trees.

Cohen-D
A statistic that is used to measure the effect of size in power analysis.

Coefficient of Determination
The percentage of the variation of the dependent variable in a regression explained by the independent or explanatory variables. Also known as R^2.

Complexity Bias
Refers to a behavioral bias that the complex can be more appealing than the simple. In reality, a simple model that is not underfitted is usually better than a complex model.

Composite Variable
A variable that combines two or more variables that are statistically strongly related.

Concept Drift
A change in the target or feature variable, which is often abrupt but also may be incremental or recurring.

Conformation Bias
Also known as observer bias. Occurs during model interpretation, where the data scientist interprets or looks for information consistent with their beliefs.

Confounding Variables
Variables that are correlated with each other because they share a common (possibly hidden) relationship.

Confusion Matrix
A 2 x 2 matrix that is used to classify a model's results into four cells (true positives, false positives, false negatives, and true negatives) and is used to measure accuracy, prediction, and recall.

Conjectural Bias
A bias that may arise when we form a conclusion or opinion based on preliminary analytics results or based on incomplete or suboptimal data.

Correlation
The degree of co-movement in an asset's or investment strategy's return relative to the return of the relevant market or benchmark.

Covariate Drift
A change in the relationship between feature and target variables.

Cross-Validation
A technique for estimating out-of-sample (OOS) error directly by determining the error in validation samples.

Data Availability Bias
The bias that arises when using readily available data, which results from limited data sources or the researcher using other researchers' shared data.

Data Drift
A situation in which the data used in model deployment are different from that used in model building and training, sufficient to affect the model's performance.

Data Leakage Bias
A situation in which the training dataset in model development uses data that are outside of the training data to be used in model building.

Data Lifecycle
People select, clean, analyze, and interpret data to make it useful for various purposes. This is an iterative, ongoing process, not a simple or linear one. How people work with data affects what they can learn from those data. Also called data journey or data biography.

Data Provenance
The detailed record of the source of the data, the data collection process, the data cleaning process, and the data exploration process.

Data Reduction Bias
The situation in which a smaller dataset is used in the analysis, such that the smaller dataset is not representative of the population.

Dimension Reduction
The process of eliminating variables that are not intended to be used in the data analysis and modeling.

Data Snooping Bias
A bias that occurs when a model is developed not from an economic or financial rationale and domain-specific knowledge, but rather by searching through data to find a statistically significant model.

Deployment Data
Data used in the implementation of the model.

Effect Size
A value measuring the strength of the relationship between two variables in a population or the difference between groups.

Efficient Frontier
Output of modern portfolio theory that generates a set of optimal portfolios with the highest return for a given level of risk (standard deviation of return).

Eigenvalue
The proportion of the total variance in a dataset that is explained by a composite variable (the eigenvector) that results from principal components analysis.

Eigenvector
A new, mutually uncorrelated composite variable that is a linear combination of original features in principal component analysis.

Endogeneity Bias
A bias that occurs when an outcome of a model influences the effect of the model's inputs.

Ensemble Bagging
The use of a group of models that outperforms any one individual model.

Ensemble Learning
A technique that involves combining the predictions from a collection of models.

Ensemble Method
A combination of multiple learning algorithms.

Entropy
The randomness of a variable.

Environmental Social Governance (ESG)
A broad framework to score an investment based on its expected contribution to climate risk, societal improvement, and prudent governance. Also called sustainability.

Exclusion Bias
The situation in which the data are not representative of the population because of the elimination of subgroups in the population.

Explainable Artificial Intelligence (XAI)
Refers to various approaches to explain complex machine learning algorithms that cannot be directly interpreted.

Fairness
The principle suggesting that results of data science tools should give fair results, without any bias and discrimination toward certain ethnic, gender, or minority groups of people. Also called objective accuracy.

F1 Score
A measure of classification model fit that is the harmonic mean of the precision and recall metrics.

Feature Shuffling
A method of testing a model in which a feature's values are randomly reassigned among observations and the resultant model is compared with the baseline model, that is, the model without shuffling.

Funding Bias
This bias occurs when the results of a model are biased to support the financial sponsor of a project.

Generalization Bias
Unintentional tendency to generalize model results to a larger population because of the cognitive bias of a researcher.

Gini Index
A measure of the likelihood of misclassifying an observation if the observation is labeled randomly.

Grid Search
A method of systematically training a machine learning model by using various combinations of hyperparameter values, cross validating each model, and determining which combination of hyperparameter values ensures the best model performance.

High Leverage Point
An extreme data point in an explanatory variable.

Hyperparameter
A parameter whose value must be set by the researcher before machine learning begins.

Influence Analysis
The process that identifies both outliers and high leverage points.

Influential Data Bias
The situation in which an outlier or a high-leverage data point that is not representative of the population affects the statistical results of a data analysis.

Informative Censoring
This bias occurs when the flow of data abruptly drops off, for a variety of reasons, during the extraction process.

Instrumental Variable Bias
A bias that occurs when a variable unintentionally proxies for another feature.

Interpretability
A model is interpretable when it is fully transparent; the user can understand why and how the model makes its predictions.

Interview Bias
A bias in the data collection that may occur if questions are confusing or misleading.

Irrelevant Variable
A variable that contributes nothing to the solution and interferes by adding random noise to the model that can fool predictive algorithms or by making their searches less effective. Also called superfluous variables.

K-Fold Cross-Validation
A method in which the data (excluding the test data) are shuffled randomly and then divided into k equal subsamples, with k – 1 samples used as training samples and one sample, the kth, used as a validation sample.

K-Nearest Neighbor (k-NN)
A supervised machine learning method that works by finding the k known training cases closest to a new case and then combining (e.g., by averaging) those answers to estimate its value or class.

Kolmogorov–Smirnov (K-S) Test
A nonparametric test statistic that can be used to compare the distributions of two datasets.

Label Bias
A data collection bias where the variable is labeled or identified inconsistently.

Least Absolute Shrinkage and Selection Operator (LASSO) Regression
A type of regression that uses regularization to penalize regression coefficients, resulting in a parsimonious model.

Local Interpretable Model-Agnostic Explanation (LIME)
A second or surrogate prediction model designed for supervised learning models. Provides local linear interpretability to complex models, such as deep learning, by testing the variability of the prediction by iterating small changes to individual input data.

Look-Ahead Bias
A bias that results from using variables that contain information that could be known only at a future date.

Majority-Vote Classifier
A method of using the results of ensemble methods that assigns the predicted label with the most votes to a new data point.

Market Beta
The measure of the magnitude of price sensitivity of a stock or investment strategy relative to the relevant market or benchmark over time.

Measurement Bias
A bias that may occur in data collection when collecting or calculating features resulting in inaccurate or erroneous data.

Model Monitoring
The use of techniques to determine whether there are issues that may arise, such as data drift, that need to be identified and addressed.

Model Tuning
The process of adjusting model hyperparameters to find the point of minimum total error.

Modern Portfolio Theory (Mean-Variance Optimization)
A portfolio optimization framework that produces a diversified portfolio with the highest expected or mean return given the level of expected variance in return.

Nonrepresentative Bias
The situation in which the sample is not representative of the population.

Nonresponse Bias
A bias from surveys that occurs when participants fail to answer some of the questions.

Neural Networks (NNs)
A flexible, nonlinear machine learning algorithm that learns from the data using three types of layers: input, hidden, and output.

Omitted Variable Bias
A bias that occurs when one or more relevant variables are not included in the dataset and in model development, resulting in a model that is not reliable.

Open-Source Risk
Relates to, among others, risks of intellectual property infringement claims, risks stemming from biases and errors in data obtained from public sources, and security risks from vulnerabilities embedded in code or software obtained from public sources.

Out-of-Sample Testing
Testing using data that was not used in the training of the model. Also called hold-out sample testing.

Outlier
An extreme data point for the explained or dependent variable.

Overfitting
The situation in which a model fits the training data well but does not fit the test data well.

P-Hacking
The use of multiple testing to search for a statistically significant results, resulting in false discoveries (i.e., false positives).

P-Value
The smallest level of significance at which the null hypothesis can be rejected.

Page-Hinkley Method
Identifies concept drift by comparing deployment data means with the training data means and indicates abrupt changes based on a user-defined tolerance.

Penalized Regression
A computationally efficient technique useful for reducing a large number of features to a manageable set and for making good predictions, thus avoiding overfitting.

Permutation Importance
A measure of variable importance that compares the prediction accuracy of a variable with that of the same variable that has had its observations shuffled.

Population Stability Index (PSI)
A method of detecting data drift in the target variable that compares the distributions of the target variable in two different datasets.

Power Analysis
The analysis of statistical procedures to determine whether they generate sufficient explanatory power at a specific significance level. Power analysis also estimates the minimum sample size required given the required significance level and effect size.

Precision
The ratio of correctly predicted positive classes to all predicted positive classes.

Prediction Bias
This situation implies that the prediction for the total dataset is inferior to the predictions for either of the subgroups.

Principal Components Analysis (PCA)
A method used to transform highly correlated features into a few main, uncorrelated composite variables referred to as eigenvectors.

Protected Classes, Protected Characteristics
The Equal Credit Opportunity Act of 1974 aimed at ensuring access to credit on a nondiscriminatory basis by introducing protected classes of individuals, such as age, race, gender, national origin, and other parameters that could not be used in a credit decision.

Publication Bias
A bias that may occur when relying on published research that likely has a bias from the higher likelihood that studies with significant results are published compared with studies with insignificant results.

Random Forest
An ensemble of decision trees.

Recall
The ratio of correctly predicted positive classes to all actual positive classes. Also called sensitivity or true positive rate.

Recall Bias
A bias that arises when survey participants are asked for information about events or experiences in the past, with the possibility that some participants have better recall than others.

Receiver Operating Characteristic (ROC) Curve
An approach for comparing models that involves a graphical plot of a curve showing the trade-off between the false positive rate (on the x-axis) and the true positive rate or recall (on the y-axis).

Recursive Feature Elimination (RFE)
A process related to backward selection is sequentially removing features based on a ranking of explanatory power.

Redundant Variables
Highly correlated variables that do not contribute to the solution.

Regularization
A method that simplifies the model by penalizing high-valued regression coefficients with the goal of reducing the risk of overfitting.

Regulatory Risk
The risk of regulatory noncompliance that can result in steep fines and/or loss of reputation.

Relational Database
A database that consists of rows and columns.

Robust Statistic
A statistic that performs well for a wide-range of probability distributions.

Root Mean Squared Error (RMSE)
An error metric that is the square root of the mean of the squared errors; the differences between the predicted and actual values of the target variable.

Sample Selection Bias
The non representativeness of a sample due to improper criteria or processes in choosing the sample.

Sampling Bias
A bias that may arise when the sample is not representative of the population.

Shapley Additive Explanations (SHAP)
A second or surrogate model used to create an interpretability framework for complex models. It does not assume that the local model is linear. Based on cooperative game theory, the model determines the marginal contribution of a given feature to the prediction. The sum of individual Shapley variables is equal to the difference between the average prediction and the total prediction.

Simpson's Paradox
The situation in which the relationship between two variables that is observed for subgroups sampled from the population is different than that of the entire sample.

Social Desirability Bias
A bias that arises when participants answer a question in a way to conform with acceptable norms or to appear more socially attractive.

Statistical Power
The likelihood that a model can detect an effect when there is one.

Stepwise Regression
A method in both linear and nonlinear regression to simplify the set of features by either adding features sequentially (forward stepping) or deleting features sequentially (backward selection or backward stepping).

Survivorship Bias
The use of data in modeling that includes only surviving entities or securities.

Target Shuffling
Random shuffling of the target variable, so the connection between it and the model's features is broken. The model's fit with the target shuffled dataset is compared with its fit using the unshuffled target dataset. Performance with the unshuffled dataset is expected to dominate, assuming the model is uncovering a significant result and not a random result.

Time-Lag Bias
The possible bias when relying on published research for which the study was completed well before publication.

Time Period Bias
A bias that may arise when selecting time-series data when the data distribution changes over time.

Transparency
The principal suggesting that data analysts and decision makers have to understand and be able to explain the model assumptions and data used for the algorithmic learning and also how the algorithm has reached the final results.

Trimming
Method of dealing with extreme observations by deleting extreme observations in the database, often based on observations outside a percentile range (e.g., deleting the lowest and highest 1%).

Type I Error
Rejection of a true null hypothesis. Also called a false positive.

Type II Error
Nonrejection of a false null hypothesis. Also called a false negative.

Underfitting
The situation in which the model does not fit the training data well because the model has not learned the patterns in the training data and, therefore, the model does not perform well using the test data.

Validation Sample
Subsample of the training data that is used to validate or test the model.

Variance Error
A measure of the degree to which the model results change from the training dataset to the validation and testing datasets.

Variance Inflation Factor (VIF)
A metric used to identify and quantify multicollinearity among features in a dataset.

Variable Importance Analysis (VIA)
Methods that identify and rank the importance of features in a model.

Winsorizing
Method of dealing with extreme data that involves substituting values for extreme observations, generally based on converting observations below a specified percentile to the value at that percentile and repeating this for those above another specified percentile to the value at that percentile.