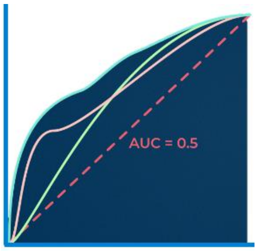
DATA SCIENCE TERMS

Accuracy – the number of correct predictions divided by the total number of predictions. Typically associated with classification models.

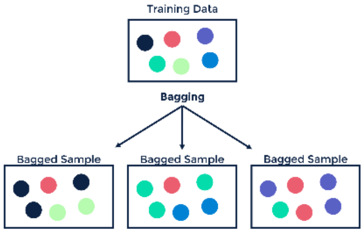
Akaike Information Criterion – a metric used to estimate quality of models for a given dataset. The values are primarily used to compare like for like versions of models, with a lower score representing a more parsimonious model. aka AIC

Algorithm – steps to accomplish a task, usually solving a problem. These steps can be intended for a computer to solve a problem, or a human. Each model has its own algorithm.

Area Under the Curve (AUC) – a measure of the area below/to the right of a ROC curve used to compare model performance. Values range from 0 – 1 with a larger AUC indicating a better performing model. Typically associated with classification models.



Bagging – a process associated with Random Forest models in which the training data is randomly divided into subgroups with replacement (duplication). This technique is useful for creating trees that differ from each other, despite being created from a finite dataset. aka bootstrap aggregating.



Balanced Accuracy – an average of the accuracy within each classification value. E.g. the average of red category accuracy (80%) and blue category accuracy (60%) would be 70%. Ideally, this number is close to the overall accuracy of the model, indicating that it performs reasonably well for all categories.

Bias-Variance Tradeoff – refers to a property of prediction errors made by a model. Biased results consistently show a similar pattern in the error of predictions while variance refers to predictions that are inconsistent or do not follow a pattern in their errors. This type of variance can indicate overfitting while bias can be indicative of underfitting. These values have an inverse relationship, thus balancing the bias-variance tradeoff results in the best possible model. Read More

Black Box – refers to a model which has and input and output, but no insight into its internal workings.

Boosting – a sequential process associated with boosting algorithms in which each iteration of training data emphasizes the errors from the previous iteration. This process creates “specialized” learners which are combined to perform better than any single iteration could.

Categorical – a classification of variables indicative of values ranging across categories. (usually a “quality” of a subject).

Centering – subtracting the mean from all values in a variable. This moves the mean to 0. This is useful when examining the variance of a dataset because it prevents the mean value from being included as an important factor in analysis.

Chi squared – a statistic calculated to determine if the distributions of two categorical variables differ from each other. The output will be a numeric value, with smaller values representing a better performing model. Represented by χ^2.

Collinearity – overlap of predictor variables. Essentially, when two predictor variables are closely related, the amount of new/useful information contained in a row of data used for making the model is decreased. This reduces the precision when determining the effect of the predictor, which creates bias in the model. (i.e. you can not state with certainty which of the related predictor variables created the effect). While the coefficient may not be effected, the standard error will increase, resulting in less confidence. This means your results are more likely to be insignificant. Read More

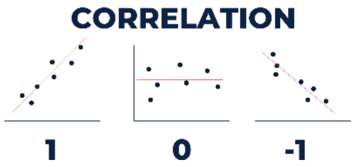
■ = general
🟩 = classification
🟦 = regression

Confusion Matrix – a table showing the performance of classification models by comparing model predictions to known values. All possible values are listed as rows and as columns. Columns represent actual values and rows represent predicted values. Correctly classified values (true) will appear at the intersection of identically named rows and columns.

True Positive	**False Positive**	Predicted Positive
False Negative	**True Negative**	Predicted Negative
Actual Positive	Actual Negative	

Continuous – a classification of variables indicative of an infinite number of possible values. (decimal that just keeps going)

Correlation – calculation of the strength and direction of two variables when the event was measured. (how likely is it that x is present when y is present) range from -1 to 1 when plotted on x-y. 0 = no relationship. Values closer to 1 or -1 imply stronger relationship. Usually represented by r.



Covariance – a measure of the relationship between the variance of two variables. If the variance of one variable goes up as the variance of the other goes down, the covariance is negative. Covariance can look similar to correlation when graphed, but the two are different measures.

Cross Validation – a method of data segmentation used to split the historical dataset into training data for creating a model and test data for evaluating the model. Cross validation randomly splits the data into the desired number of groups and uses all but one group to create the model. The last group is used to evaluate the model. The groups are switched and the process repeated until each grouping is used as the evaluate the model. The results are then combined to give an estimate of the model's performance. This method utilizes all available data for model creation and testing, without overfitting. Read More

Curse of Dimensionality – as you add more variables to your model, the generalizability of the predictions is decreased. This is a result of the added dimensional space required to "map" the variables as variables are added. Typically associated with regression techniques.

Data Mining – determining patterns and knowledge from datasets. The analysis phase of data discovery. (uncovering hidden patterns in large volumes of data). Mining in the sense of extraction of value, not the "crunching" of the numbers.

Dimensionality – the number of variables included in your dataset. Each variable has a distribution along a single axis. As your dataset adds variables, the dimensionality increases. This is an important consideration as the relationships between datapoint tend to spread as more dimensions are added. In "high dimensional" space, this can cause issues in training algorithms (known as the curse of dimensionality).

Discrete – a classification of variables indicative of a finite (countable) number of possible values. (you might not want to count to a million, but you could). E.g. trips to the store, likert-scale scores

Extraneous variable – any variable that has no impact or a negligible effect on the model's ability to predict the target variable. Can increase noise if included.

Feature Engineering – using existing predictor variables to "create/derive" a more useful predictor variable (e.g. using a date to create days since start)

Feature Selection – choosing which variables to include as predictors in a modeling algorithm. This process includes deciding which variables not to include and whether or not feature engineering is required.

Fitting a Model – refers to the process of training a model to match (fit) the dataset. Each model will fit the data differently. E.g. in a linear regression, the slope of the line will effect the fit of the model to the data. A good fit reflects the data accurately, a bad fit does not.

Gamma Distribution – a family of continuous probability distributions which are always positive. Exponential distribution is an example of a gamma distribution. Typically, the values in these distributions cluster toward the lower range with few values in the upper range. E.g. Income distribution The Gamma Regression tool assumes a gamma distribution for the target variable.

Gaussian Distribution – a probability distribution which is commonly used in statistics and machine learning. Many modeling algorithms make assumptions about the data based on this distribution. aka normal distribution

Gini index – a measure of performance associated with classification models. This measure indicates the information gain achieved by the model (reduction of uncertainty; entropy). Information before splitting minus information after splitting is the information gain.

Goodness of fit – a common term which refers to how well a model describes a given dataset. There are many measures for goodness of fit, but they are all intended as a way to evaluate model performance. Some common examples include chi-squared, coefficient of determination (R^2), and AIC.

Hyperparameters – refer to settings for models which control the process of training a model. These settings dictate things like when a model is considered finished learning and the number of records processed at a time. Adjustment of hyperparameters takes place before training a model. Read More

Hyperplane – a boundary which separates groups in a Support Vector Machine classification or indicates the trend of datapoints in a SVM regression. The hyperplane will have one less dimension than the dataspace (e.g. a 3 dimensional dataset will have a 2 dimensional hyperplane).

Impute – replacing missing values in a dataset based on values for other records in the dataset. Typically, the median or mode for a variable is used.

Log loss – a measure of performance which factors in predicted probability and the deviation from the correct classification. Lower scores indicate better performance.

Mean Absolute Error – a metric which represents the difference between the actual values and those predicted by the model. The returned value indicates the average difference between expected and actual values. E.g. if the MAE is 3, the average predicted value is 3 units off from the actual value. A lower value is generally considered better. aka MAE

$$\frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Model – a “map” of the relationships between the variables selected. This can be an output from a tool, macro, or workflow. The model serves as a tool to approximate the relationships between variables, not the code/workflow that got you to that output. Idea is that if you can map the relationships between variables, you can change variables to “predict” an outcome, fill in values that weren’t captured, and better understand the nature of those relationships.

Model Object – one of the outputs from a modelling tool in Designer (typically the O output). This refers to the R object containing the model. The Model Object is a single item which can be saved as a .yxdb file for later use. Think of it as a package containing the R code for the model.

Noise – refers to the amount of precision/decipherability/predictability of the data. Data collection can introduce error in datasets (reaction time, sensitivity, etc.). When modelling, extraneous variables and/or missing predictor variables can add noise to the outcome, making the pattern less defined. This results in more uncertainty and a larger range of probability. Think about hearing a conversation in a crowded auditorium. More noise makes it harder to distinguish the important auditory information.

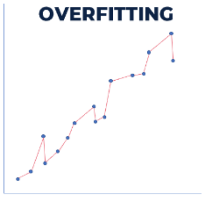
Normalizing – refers to techniques used to standardize the scale of variables for use in a modeling algorithm. Normalizing values can negate the impact that units can have on the algorithm. Normalizing can also increase the speed of processing. Read More

One-hot encoding – A method of transforming values in which a new column is created for each possible value of the original column. Each record is designated as 0 (does not fit that column) or 1 (does fit that column). Each record should fall into only one column. When used for regression problems, one column is dropped as we can infer that a record fits the missing value if it does not fit any of the others. Read More

Out of Bag Error – an error measurement associated with some classification algorithms. For algorithms that use bagging, the out of bag error indicates the average model performance for data that was not included in the sample used to create a given tree. I.e. how does a given tree perform on data that was not used in creating that tree. This measure can be important when determining how many trees a model should create. aka OOB Error; bootstrap estimate of standard error

Outlier – a value that is significantly different from other values. These values usually represent an anomaly, a niche case, or an error in the dataset. Defining what constitutes an outlier is discretionary but common measures are >2 standard deviations from the mean or 1.5 times the Inner Quartile Range. Some modeling algorithms are more sensitive to outliers than others, but identification of outliers is key step in all modeling projects.

Overfitting – it is possible for a model to fit a subset of data too well. Overfitting creates a very accurate model for the data used to create the data, but may not perform well on new data. If the data sample used to create a model is biased, overfitting to that data will create a poor model when applied to the real world. Read More



Oversampling – a technique used to alter the ratio of class distributions in datasets. This is useful when trying to predict an outcome that is underrepresented in the dataset (e.g. defaults in loan data). Oversampling creates a more balanced dataset which can help to create a more accurate model. Read More

p-value – a value associated with the t-value after a t-test is performed. The p-value indicates the probability that a t-value as large or larger than the returned value would be returned if the two groups were not significantly different. A smaller p-value indicates that the result was less likely to have occurred by chance. Traditional thresholds for accepting results as significant are a p-value of .05 or .01.

Poisson Distribution – a discrete probability distribution which is always >= 0 and results in whole numbers. This distribution is associated with problems around the number of times an event occurs. The Count Regression tool assumes a Poisson distribution for the target variable.

Precision – of the items labelled as positive, what percentage were correct.

$$\frac{True\ Positives}{True\ Positives + False\ Positives}$$

Predictor Variable – any variable that helps to calculate the target variable. Usually several predictor variables to find one target variable. Must be significant and contribute to the model's accuracy. Aka Independent variable

Principal Component Analysis (PCA) – a technique which evaluates the numeric variables available and transforms them into a compressed version without losing valuable information. This is possible by evaluating the patterns with the data between each combination of variables and selecting the top contributors. This is extremely useful in reducing dimensionality, as the amount of information lost is minimized. PCA is sensitive to scale (you'll need to normalize) and only works for numeric variables. Read More

Python – a programming language with a variety of applications. The Assisted Modeling tools are based on the Sci-Kit Learn library in Python.

R – a programming language developed specifically for statistical use cases. Many Data Investigation and Predictive tools are based in R.

R² – a statistic calculated to determine what amount of the variation in the data is described by the model. It is commonly used as a method of evaluating model performance. Values range from 0 to 1 with higher values indicating a better model. aka Coefficient of Determination. The R² value will always increase as more variables are added to a model, so the Adjusted-R² can offer a more accurate indication of performance.

Recall – ability to identify all occurrences (no false negatives, of the things you were looking for, what percentage did you get right, did you find every instance of the thing you were looking for). True positives divided by the sum of the true positives and false negatives.

$$\frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Receiver Operating Characteristic (ROC) Curve – a tool used to compare model performance. The False Positive rate is plotted on the x-axis and the Recall on the y-axis. The curve is generated by finding the FPR and Recall for different threshold values (the value at which the categories are split). ROC curves that approach the upper left corner of the graph are considered good performers, while those that approach a linear bisection of the graph are considered poor. Typically associated with classification models.

Regularization – refers to several techniques of penalizing coefficients in order to prevent overfitting. These methods appear as an optional hyperparameter in applicable modeling tools. Examples include Ridge, LASSO, and Elastic Net. Read More

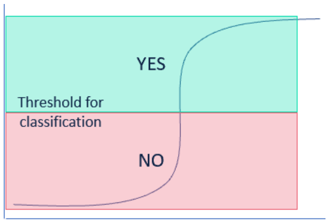
Root Mean Squared Error – a metric which represents the difference between the actual values and those predicted by the model. The value returned is in the standard deviation of the residuals. The difference between prediction and actual values is squared, then the square root of the sum of those squared values is calculated. The result is similar to Mean Absolute Error but is more sensitive to larger errors made by the model. A lower value is generally considered better. aka RMSE

$$\sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Scaling – refers to techniques used to adjust the scale of values associated with a variable while maintaining the relative difference between those values. Commonly used to prevent a variable with larger values from dominating a model. E.g. when comparing income vs age, you may scale income on a 0-100 scale.

Scoring – using a model to predict on data you've collected. Essentially, the relationships mapped out in the model are applied to the incoming data. The results are appended to the dataset. This can occur in the Training phase or implementation phase.

Sigmoid Curve – a function associated with the probability distribution of a logistic regression. The plot resembles an "S", with the y values indicating the probability of belonging to a particular category.



Sigmoid Curve

Stratified Sampling – a method of sampling in which each group within a population is included in every subset of data. I.e. each group is proportionately represented in a sampling. This can improve the precision of a model, especially when the sample data is imbalanced.

t-test – This test is used to determine if the difference between a control group and other groups are significant. This test results in a t-value which is associated with a p-value. The combination is used to determine the significance of the finding. aka test of means

Target Variable – is the variable that will be predicted. I.E. what you are trying to find. Aka Dependent Variable.

Training a Model – refers to the creation and improvement of a model. Training is not the same as “training” an employee. In this context, the process of selecting a model and the subsequent process of “tuning” that model are both considered training the model.

Validating a Model – Validation data (data previously collected, but not used in the creation of the model) is fed into the model. The results of the model are compared to the collected values and which provides a means of evaluating the “fit” of the model to our dataset.

Value – a measurement of a variable at a particular moment in time. (a cell within a column or row).

Variable – quantities, qualities, or properties that can be measured. (column of data). Also known as an attribute, feature, or dimension.

Variance – sometimes referred to as the “spread” of numbers within a variable. Variance is calculated as the sum of (the differences between each value and the mean for that variable) squared. (the volatility of a variable). A variable with no variance is a constant which has no predictive value. Too much variance within a variable can indicate noise and error, making it unreliable for predictions.