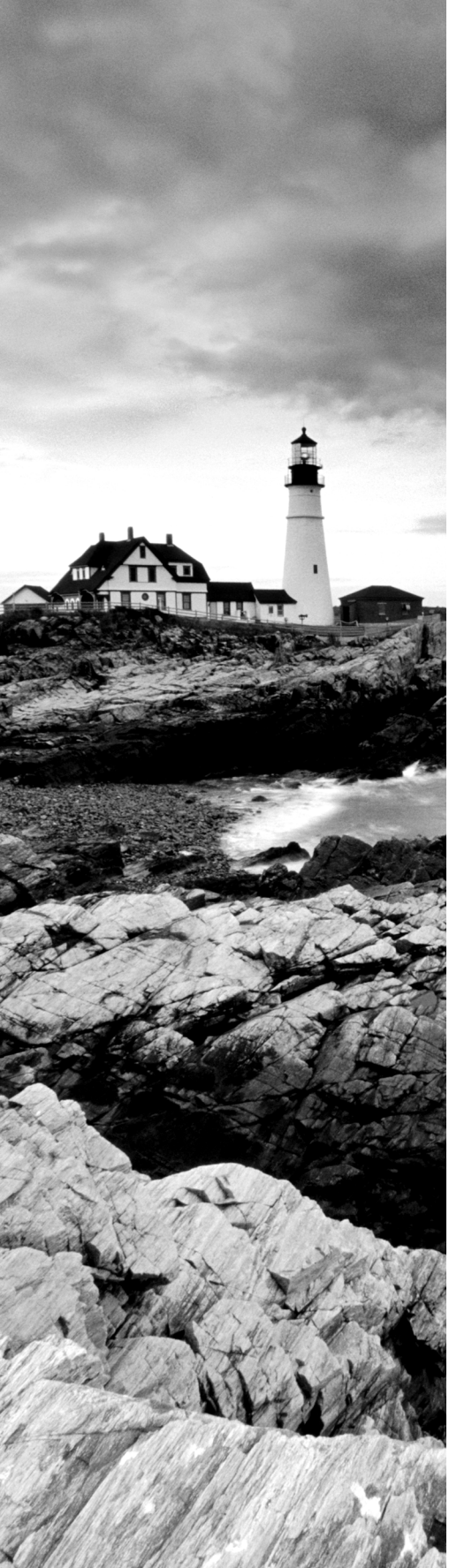

Glossary

accuracy The number of correct classification predictions divided by the total number of predictions. For example, a model that made 50 correct predictions and 10 incorrect predictions would have an accuracy of

$$\text{Accuracy} = 50 / 50 + 10 = 83.3\%$$

activation function Refers to a function that enables neural networks to learn nonlinear (complex) relationships between features and the label. Popular activation functions are ReLU and Sigmoid.

area under the precision-recall curve (AUC PR curve) In case of an imbalanced class, precision-recall curves (PR curves) are recommended for highly skewed domains. AUC PR gives more attention to the minority class (can be used in conjunction with downsampling or upsampling).

area under the ROC (AUC ROC) curve Refers to a number between 0.0 and 1.0 representing a binary classification model's ability to separate positive classes from negative classes. The closer the AUC is to 1.0, the better the model's ability to separate classes from each other. AUC ROC curves are used when the class is balanced or when you want to give equal weight to both positive and negative class while training a multiclass model.

ARIMA_PLUS ARIMA is the acronym for "autoregressive integrated moving average," a type of model for forecasting future points in a time series. ARIMA_PLUS is a type of model supported by BigQuery ML which has added features on top of the older ARIMA mode.

AutoML The process of *automating* the tasks of applying *machine learning* to real-world problems. AutoML potentially includes every stage from beginning with a raw dataset to building a machine learning model ready for deployment. The high degree of automation in AutoML aims to allow nonexperts to make use of machine learning models and techniques without requiring them to become experts in machine learning.

Avro This term refers to Avro files, which store a data definition in JSON format, making it easy to read and interpret. It also includes markers that can be used to split large datasets into subsets suitable for Apache MapReduce processing.

backpropagation The algorithm that implements gradient descent in neural networks. Neural networks contain many neurons across many hidden layers. Each neuron contributes to the overall loss in different ways. Backpropagation determines whether to increase or decrease the weights or minimize the loss applied to particular neurons.

bagging This refers to bootstrap aggregating. Bagging is a method to train an ensemble where each model trains on a random subset of training examples sampled with replacement. An example of bagging is a random forest algorithm; it is a collection of decision trees trained with bagging.

batch normalization Normalizing the input or output of the activation functions in a hidden layer. It can provide benefits such as making neural networks stable because batch normalization protects against outlier weights. It can enable a high learning rate, which can result in increasing training speed. Moreover, it can also reduce overfitting.

batch size This term refers to the number of examples in a batch. For instance, if the batch size is 1,000, then the model processes 1,000 examples per iteration. Originally, SGD was used on a batch of 1, but that is not required. In practice, SGD is used on minibatches of examples, usually between 32 and 256.

Bayesian neural network A network that relies on Bayes' theorem to calculate uncertainties in weights and predictions. A Bayesian neural network is useful when it is important to quantify uncertainty. Bayesian neural networks can also help prevent overfitting.

Bayesian optimization Optimizing expensive-to-evaluate tasks that have a small number of parameters, such as selecting hyperparameters.

bias A phenomenon that skews the result of an algorithm in favor of or against an idea. Bias is considered a systematic error that occurs in the machine learning model itself due to incorrect assumptions in the ML process.

bias (ethics/fairness) Stereotyping, prejudice, or favoritism toward some things, people, or groups over others. The terms *biases* and *fairness* are often used with Responsible AI practice and to make sure ML models are free from such bias.

binary cross-entropy Binary cross entropy compares each of the predicted probabilities to actual class output, which can be either 0 or 1.

binning Converting a single feature into multiple discrete features called buckets or bins, typically based on a value range. The chopped feature is typically a continuous feature. Let's look at an example of representing temperature. You could segregate the ranges of temperatures into discrete buckets, such as temperatures less than –10 degrees Celsius would be the "extreme cold" bucket and temperatures between –10 to 50 would be normal. The model will treat every value in the same bucket identically. For example, the values –10 and –40 are both in the cold temperature bucket, so the model treats the two values identically.

boosting A machine learning technique that iteratively combines a set of simple and not very accurate classifiers (referred to as "weak" classifiers) into a classifier with high accuracy (a "strong" classifier) by upweighting the examples that the model is currently misclassifying.

categorical cross-entropy Categorical cross-entropy is used when true labels are one-hot encoded. It is recommended to use categorical cross-entropy for multiclass (classes are mutually exclusive) problems but binary cross-entropy for multi-label problems.

categorical data Features having a specific set of possible values. For example, consider a categorical feature named traffic-light-state, which can only have one of the following three possible values: red, yellow, and green. Categorical features are sometimes called discrete features.

categorical hinge loss In machine learning, hinge loss is a loss function used for training classifiers. This function is popular with support vector machines (SVMs) and used for training the classifiers. Refer to Wikipedia to learn more (`https://en.wikipedia.org/wiki/Hinge_loss`).

classification This term refers to the process in machine learning of predicting "labels" or "classes" or "categories." Given a picture of a pet, classifying dogs vs. cats is a classification problem. If there are just two labels, it is called binary classification, and if there are more labels, it is called multiclass classification.

classification threshold In a binary classification, a number between 0 and 1 that converts the raw output of a logistic regression model into a prediction of either the positive class or the negative class. It is a value that a human chooses, not a value chosen by model training. A logistic regression model outputs a value between 0 and 1. Suppose the classification threshold is 0.6 set by the human and the value predicted by the model is 0.7; then the model predicts the positive class. This choice of classification threshold strongly influences the number of false positives and false negatives.

class-imbalanced dataset The dataset used for a classification problem in which the total number of labels of each class differs significantly. For example, consider a binary classification dataset whose two labels are divided as follows: 5,000,000 negative labels and 10 positive labels. The ratio of negative to positive labels is 500,000 to 1, so this is a class-imbalanced dataset. Multiclass datasets can also be class-imbalanced because one label has far more examples than the other two classes.

clipping A technique for handling outliers by reducing feature values that are greater than a maximum threshold or increasing feature values that are less than a minimum threshold. For example, suppose that <0.6% of values for a particular feature fall outside the range 40–60. In this case, you could do the following:

Clip all values over 60 (the maximum threshold) to be exactly 60.

Clip all values under 40 (the minimum threshold) to be exactly 40.

Cloud Storage or Google Cloud Storage (GCS) Cloud Storage is a service for storing your objects in Google Cloud. You store objects in buckets, which are similar to folders in your computer. All buckets are associated with a project, and a project can be under an organization. After you create a project, you can create Cloud Storage buckets, upload objects to your buckets, and download objects from your buckets. See `https://cloud.google.com/storage/docs/introduction`.

clustering The task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

concept drift A change in the relationship between the input and the predicted value. In general, there is a relationship between the input variables and predicted variables that we try to approximate using a machine learning model. When this relationship is not static and changes, the old data is considered stale, and new data is required to train the model.

continuous feature A floating-point feature with an infinite range of possible values, such as temperature or weight.

continuous integration and continuous delivery/continuous deployment(CI/CD) CI and CD stand for continuous integration and continuous delivery/continuous deployment. CI is a modern software development practice in which incremental code changes are made frequently and reliably. A CI/CD pipeline automates the process of software delivery. It builds code, runs tests, and helps you to safely deploy a new version of the software. A CI/CD pipeline reduces manual errors, provides feedback to developers, and allows fast product iterations.

convergence In deep learning, loss values sometimes stay constant or nearly so for many iterations before finally descending. Convergence is the state reached when loss values change very little or not at all with each iteration. A model converges when additional training will not improve the model.

convolutional layer The layer of a deep neural network in which a convolutional filter passes along an input matrix. A convolutional layer is the main building block of a CNN and contains a set of filters (or kernels), parameters that are to be learned throughout the training.

convolutional neural network (CNN) A neural network in which at least one layer is a convolutional layer. A typical convolutional neural network consists of some combination of convolution layer, pooling layers, and a dense layer. CNNs are mostly used in image recognition.

cross-validation A mechanism for estimating how well a model would generalize to new data. This is done by testing the model against one or more nonoverlapping data subsets withheld from the training set.

data analysis Refers to having an understanding of data by considering samples, measurement, and visualization. Data analysis is useful when a dataset is collected before one builds the first model. It helps to check for outliers in the data and to check how the features are correlated in the data.

data drift A change in the input data that is fed to the machine learning model after deployment. This could happen slowly over time or rapidly. This can be detected by monitoring the input values to your model. One kind is where the type of data changes due to change in the schema. Another kind of data drift is when there is a statistical distributional change, such as, for example, the change in the height of trees in one region compared to another region.

data extraction The process of collecting or retrieving disparate types of data from a variety of sources that can be structured or unstructured. This process helps prepare the data for analysis or business intelligence (BI) and machine learning.

data parallelism A way of scaling training or inference that replicates an entire model onto multiple devices and then passes a subset of the input data to each device. Data parallelism can enable training and inference on very large batch sizes; however, data parallelism requires that the model be small enough to fit on all devices.

data transformation Data transformation is the process of converting data from one format or structure to another format or structure.

data validation The process of verifying and validating data that is collected before it is used.

DataFrame A popular Pandas data type for representing datasets in memory. A DataFrame is analogous to a table or a spreadsheet. Each column of a DataFrame is a 2D array that has a name (a header), and each row is identified by a unique number.

Dataset API (tf.data) A high-level TensorFlow API for reading data and transforming it into a form that a machine learning algorithm requires. A tf.data.dataset object represents a sequence of elements, in which each element contains one or more tensors. A tf.data.iterator object provides access to the elements of a dataset. See `www.tensorflow.org/api_docs/python/tf/data/Dataset`.

dataset or data set A collection of raw data, commonly (but not exclusively) organized in a format such as a spreadsheet or a CSV (comma-separated values) file.

decision boundary The separation between classes learned by a model in binary class or multiclass classification problems.

decision trees Decision trees are used to model complex patterns in data. Decision trees are based on a hierarchical structure, and they can be used to solve a wide variety of problems, including regression and classification tasks.

deep model A neural network containing more than one hidden layer. Also called a deep neural network.

depth The sum of the number of hidden layers, number of output layers, and number of any embedding layers in a neural network. For example, a neural network with 6 hidden layers and 1 output layer has a depth of 7. Notice that the input layer does not influence depth.

discrete feature A feature with a finite set of possible values. For example, a feature whose values may only be animal, vegetable, or mineral is a discrete (or categorical) feature.

Dockerfile A text document that contains all the commands a user could call on the command line to assemble an image.

downsampling In a class-imbalanced dataset, models tend to learn a lot about the majority class and not enough about the minority class. Downsampling helps balance the amount of training on the majority and minority classes.

drift detection To detect if there is a statistical distribution change in the input data during prediction. This can be detected by monitoring the input values to your model.

early stopping A method for regularization that involves ending training before training loss finishes decreasing. In early stopping, you usually stop training the model when the loss on a validation dataset starts to increase.

embedding layer This term refers to a special hidden layer that trains on a high-dimensional categorical feature to gradually learn a lower dimension embedding vector. An embedding layer enables a neural network to train far more efficiently.

ensemble A collection of models trained independently whose predictions are averaged or aggregated. In many cases, an ensemble produces better predictions than a single model. For example, a random forest is an ensemble built from multiple decision trees. Note that not all decision forests are ensembles.

epoch A full training pass over the entire training set such that each example has been processed once. An epoch represents the total number of examples divided by batch size training iterations.

Estimator It is a deprecated TensorFlow API. Use tf.keras instead of Estimator.

exploding gradient problem The tendency for gradients in deep neural networks (especially recurrent neural networks) to become surprisingly steep (high). Steep gradients often cause very large updates to the weights of each node in a deep neural network. Models affected by exploding gradient problem become difficult to train.

exploratory data analysis (EDA) An approach to analyze the data using visual techniques. Exploratory data analysis can be used to discover trends and patterns and to check assumptions with the help of statistical and graphical representations.

extract, transform, load (ETL) A process in data warehousing. It is a three-phase process where data is extracted, transformed, and loaded into an output data container or a data warehouse. The data can be collected from one or more sources and it can also be sent to one or more destinations, usually a data warehouse, after transformation.

F1 A metric used to evaluate models trained on an unbalanced dataset. The F1 score is the mean of precision and recall. This metric is typically used when you are interested in reducing both false positives and false negatives.

false negative (FN) An example in which the model mistakenly predicts the negative class. Consider, for example, a false negative use case. Suppose the model predicts that a particular email message is not spam (negative class), but that email message actually is spam (positive class).

false positive (FP) An example in which the model mistakenly predicts the positive class. For example, consider a false positive use case. Suppose the model predicts that a particular email message is spam (positive class), but that email message is actually not spam (negative class).

feature An input variable to a machine learning model. Each feature, or column, represents a measurable piece of data that can be used for analysis, such as, for example, name, date of birth, sex, and salary.

feature cross This term refers to a synthetic feature formed by crossing categorical or bucketed features. This concept is explained in detail in Chapter 3, "Feature Engineering."

feature engineering Refers to feature extraction, a process that determines which features might be useful in training a model. This process involves selecting features from a raw dataset and transforming those features to be consumed by a machine learning model.

feature importance The set of scores that indicates the relative importance of each feature to the model. Feature importance is a techniques that calculates a score for all the input features for a given model. It determines which feature is contributing more to the model prediction.

Feature Store A data store where the features are stored and organized to either train models or make predictions (by applications that have a trained model). Feature Store is a centralized place where you can either create or update groups of features created from multiple different data sources. You can also create and update new datasets from those feature groups for training models or for use in applications that do not want to compute the features but just retrieve them when they need them to make predictions.

forecasting The process of making predictions based on past and present data. As a type of machine learning problem, the model is trained on time-series data and the prediction is future values.

full batch The number of examples in the entire training set. For example, if the training set contains a hundred examples, then the batch size would be a hundred examples. Using a full batch or complete batch is usually an inefficient strategy.

gcloud CLI (command-line interface) A set of tools to create and manage Google Cloud resources using a simple command line in the Google Cloud console. You can use these CLI commands to perform many common Google Cloud tasks such as creating storage or setting access permission from the command line or through scripts and other automation.

Google Cloud console The Google Cloud console is a UI-based management console that helps you launch and manage Google Cloud resources.

GPU (Graphics Processing Unit) A specialized processor originally designed to accelerate graphics rendering. This is now being used to accelerate model training. Models used for natural language, images, and videos need compute-intensive operations like matrix multiplications that can benefit by running on massively parallel architectures like GPUs.

hashed feature A feature that has been transformed using a popular technique called hashing. In hashing, a high cardinal variable is transformed into a low cardinal domain.

inference The process of using a trained model to predict a result from some input data. This involves taking the trained model and deploying it on a device or a compute endpoint. This deployed endpoint will then process incoming data and identify whatever it has been trained to recognize.

k-means A clustering algorithm that aims to partition n data points into k clusters in which each observation belongs to the cluster, with the nearest mean (cluster centroid) serving as a prototype of the cluster.

k-nearest neighbors (k-NN) algorithm A supervised learning algorithm that uses closest k nearest data points to decide on the output. For classification, the output is the majority class of its k neighbors. For regression, the output is the average value of the k neighbors.

label In supervised machine learning, each labeled example consists of one or more features and a label. For example, in a spam detection dataset, the label would probably be either "spam" or "not spam." In a rainfall dataset, the label might be the amount of rain that fell during a certain period.

lineage This term refers to the history of data, what happens to it, and where it moves over time. This can also be applied to ML models; what dataset was used to train the model, how the predictions are used later on, etc. When you are trying to investigate (for debugging purposes or audit purposes) a particular prediction, it would be essential to know which model produced a prediction, and then what data was used to train the model, and so on.

linear regression Used to predict the value of a target variable based on a set of input variables. It is often used for continuous-value modeling tasks, such as predicting the sales volume of a product based on historical sales data.

logistic regression Logistic regression is used to predict the value of a target variable based on a set of input variables. It is often used for discrete-value modeling tasks, such as predicting whether or not a customer will purchase a product.

machine learning (ML) pipeline A machine learning pipeline is automation of the ML workflow. ML pipelines are a sequence of steps from data extraction and preprocessing to model training and deployment.

machine learning orchestration Machine learning orchestration tools are used to automate and manage workflows and pipeline infrastructure with a simple, collaborative interface. Along with management and creation of custom workflows and their pipelines, these tools help us track and monitor models for further analysis.

majority class The class in a dataset that a majority of the datasets belong to. Suppose a given dataset has 98 percent negative labels and 2 percent positive labels; the negative labels are the majority class.

mean absolute error (MAE) A regression metric that is equal to the average absolute difference between the actual values and the predicted values.

mean absolute percentage error (MAPE) A regression metric that is the average absolute percentage difference between the labels and the predicted values. You would choose MAPE when you care about proportional difference between actual and predicted value.

mini-batch The fixed number of training examples that is fewer than those in the actual dataset. For example, in each iteration, we train the network on a different group of small mini-batch samples until all samples of the dataset are used.

minority class This term refers to less common labels in a class-imbalanced dataset. If a dataset contains 98 percent negative labels and 2 percent positive labels, the positive labels are the minority class.

MLOps This term represents the application of DevOps principles to machine learning operations. This provides architectural and policy guidance to ML engineers managing the operations and deployment.

model evaluation The process of using different evaluation metrics to understand a machine learning model's performance. It also means knowing model metrics such as, for example, how much compute resources a model is taking to converge.

model serving Hosting machine learning models on the cloud or on-premises and making their functions available via API so that applications can incorporate Artificial Intelligence into their systems.

model training The process in which on a training dataset you fit the best combination of weights and bias to a machine learning algorithm to minimize a loss function over the prediction range. Training a model means determining good values for all the weights and the bias from labeled examples in the dataset.

natural language Any language that has evolved naturally in humans through use and repetition without conscious planning or premeditation. Examples are English, Chinese, and Hindi.

neural network A type of machine learning algorithm that is used to model complex patterns in data. Neural networks can be used to solve a wide variety of problems, including regression and classification tasks, and are inspired by how neurons work in the human brain.

nonparametric algorithm Nonparametric algorithms are models that do not have a predefined form but predict based on information derived from data. Examples are decision trees and support vector machines.

normalization A technique used as a part of data preparation for machine learning. Normalization aims to change the range of numeric columns to a common scale without losing information.

parametric algorithms Parametric algorithms are best suited for problems where the input data is well-defined and predictable. Some examples of parametric algorithms are linear regression, logistic regression, and neural network.

personally identifiable information (PII) Information that is used to distinguish an individual's identity, such as, for example, their name, Social Security number, biometric records, etc. Also personal or identifying information that is linked or linkable to a specific individual, such as date and place of birth.

precision A metric used to evaluate models trained on an unbalanced dataset. Precision is the fraction of true positives to all the positive predictions. This is typically used when you want to reduce the false positives.

prediction The output of a machine learning algorithm. Once a model is trained on data, it gives an output or a result, which is called a prediction.

principal component analysis (PCA) A popular unsupervised learning technique for reducing the dimensionality of data. It helps to find the most significant features in a dataset.

protected health information (PHI) A person's health-related information. This is also defined in HIPAA, which grants federal protections and rights to patients regarding any PII or sensitive information of patients stored or collected by entities.

R^2 A regression metric that is the square of the Pearson's correlation coefficient (r) between the labels and predicted values. This range is from zero to one; a higher value indicates a higher quality model.

random forest A classification algorithm consisting of many decision trees.

recall A metric used to evaluate models trained on an unbalanced dataset. Recall is the measure of what percentage of the positive data points the model was able to predict correctly. This is typically used when you want to reduce the false negatives.

rectified linear unit (ReLU) A rectified linear unit (ReLU) is an activation function used in deep neural network modeling that introduces nonlinearity to a deep learning model. It also solves the vanishing gradients issue.

regression The process of predicting a number (continuous variable) given a set of inputs. Example: predicting the age of a tree given its height.

root-mean-squared error (RMSE) A regression metric that is equal to the square root of the average squared difference between the target and predicted values. Ranges from 0 to infinity.

root-mean-squared logarithmic error (RMSLE) A regression metric that is similar to RMSE except that it uses the natural logarithm of the predicted and actual values +1. This is an asymmetric metric, which penalizes under prediction (value predicted is lower than actual) rather than over prediction.

SavedModel A SavedModel is stored as a directory on disk and is created when you call `tf.saved_model.save()`. It is a protocol buffer describing the functional tf.Graph.

shard When you have multiple workers, and each worker should have its own subset of the dataset, a single dataset can be distributed as shards across multiple databases. These shards can then be stored on multiple machines. Sharding helps in splitting large datasets into smaller chunks and storing these chunks in multiple data nodes, thus increasing the total storage capacity of the system.

sigmoid activation unit Refers to an activation unit in a neural network. When you use a sigmoid activation function in deep neural network layers , the output will always be between 0 and 1.

SignatureDef This term defines the signature of a computation in the context of a TensorFlow graph. The aim of SignatureDef is to provide support to identify inputs and outputs of a function. Also, it can be specified when building a SavedModel in TensorFlow.

softmax activation function Softmax activation function output probabilities range between 0 to 1(the sum of all the probabilities will be equal to 1). When the softmax function is used for a multi-classification model, it will return the probabilities of each class and the target class will have the high probability.

sparse categorical cross-entropy Sparse categorical cross-entropy computes the cross-entropy loss between the labels and predictions. Use this cross-entropy loss function when there are two or more integer-encoded label classes while using TensorFlow for classification. For further reading check this link `www.tensorflow.org/api_docs/python/tf/keras/losses/SparseCategoricalCrossentropy`

squared hinge loss A loss function used in machine learning especially for support vector machines. It is the square of the hinge loss. It is intended to smooth the surface of the error function so that is differentiable. In practice, whenever hinge loss is needed, squared hinge loss is used.

Structured Query Language (SQL) A domain-specific language used in programming and designed for querying and managing data stored in a relational database management system.

supervised learning A type of machine learning algorithm problem where the training data has labeled examples, which means that each data point contains features and an associated label.

support vector machine (SVM) A type of machine learning algorithm that is used for classification problems. SVMs are suited for tasks where the input data can be separated into distinct classes or groups using a linear boundary or linearly separable.

Synthetic Minority Oversampling Technique (SMOTE) A statistical technique for increasing the number of cases in your dataset in a balanced way.

tabular data Information presented in the form of a table with rows and columns.

t-distributed stochastic neighborhood embedding (t-SNE) A technique to visualize high-dimensional data on a two- or three-dimensional map. The statistical algorithm is used extensively to visualize word embeddings, which help us to understand word clusters and gain useful insights.

tensor A number of a specific shape and a certain rank (dimensionality). Tensors are used in machine learning with TensorFlow to represent input data and output data in machine learning models.

TensorFlow Data Validation (TFDV) library Used to identify any anomalies in the input data by comparing data statistics against a schema.

TensorFlow Extended (TFX) TFX is a TensorFlow-based platform used to host end-to-end machine learning pipelines. Using the TFX configuration framework, you can prepare pipelines to clean data and train and serve production-ready ML systems.

TensorFlow Transform The tf.Transform library is part of TensorFlow Extended (TFX) and allows you to perform transformations prior to training the model and to emit a TensorFlow graph that reproduces these transformations during training. Using tf.Transform avoids the training-serving skew.

Tensor Processing Unit (TPU) Specialized hardware accelerator designed by Google specifically for machine learning workloads. Instead of having an arithmetic logic unit (ALU) that performs one operation at a time in a CPU, each TPU has multiple matrix multiply units (MXUs). TPUs can perform huge operations on huge matrices, which are the core of the training loop in neural networks.

TFRecord The TFRecord format is a simple format for storing a sequence of binary records. The advantages of the TFRecord format are that it helps store datasets efficiently and gets faster input/output (I/O) speed than reading raw data from disk. The TFRecord format of the data is useful for training deep neural networks with TPUs.

training The process of determining the ideal parameters (weights and biases) a model comprises. During training, a system reads in examples and gradually adjusts parameters. Training uses each example anywhere from a few times to billions of times.

training-serving skew A statistically significant difference between the training dataset and production dataset.

TRANSFORM A BigQuery ML clause that is part of the `CREATE_MODEL` function that is used to modify with the intent to enhance the data used to train a model.

unsupervised learning A machine learning paradigm where the data is unlabeled or does not have classes. The hope is to understand useful patterns even without labels or create methods to self-learn. This is also different from reinforcement learning where the agent (algorithm) learns about the world by actively exploring an environment.

versioning A method to give version numbers to trained models to manage them. This helps in identifying the sequence of models in historical order.

weight Refers to the parameter within a neural network that transforms input data within the network's hidden layers. A neural network is a series of neurons, and within each neuron is a set of inputs, weight, and a bias value. Training is the process of determining a model's ideal weights.

wide model A linear model that typically has many sparse input features. It is considered "wide" because the neural network has a large number of inputs that connect directly to the output node. Wide models are easier to debug and inspect than deep models. Although they cannot leverage nonlinearities through hidden layers, they can use transformations such as bucketization and feature crosses to model nonlinearities in different ways.

Z-score normalization A scaling technique that replaces a raw feature value with a floating-point value representing the number of standard deviations from that feature's mean. Z-Score value is used to understand how far the data point is from the mean. Technically, it measures the standard deviations below or above the mean. It ranges from –3 standard deviation up to +3 standard deviation.