Launching into Machine Learning

Course Summary and Key Takeaways

Learning Objectives

- Describe how to improve data quality
- Perform exploratory data analysis
- Build and train AutoML Models using Vertex AI
- Build and train AutoML Models using BigQuery ML
- Optimize and evaluate models using loss functions and performance metrics
- Create repeatable and scalable training, evaluation, and test datasets

Module Breakdown

- Module 0: Introduction
- Module 1: Get to Know Your Data: Improve Data through Exploratory Data Analysis
- Module 2: Machine Learning in Practice
- Module 3: Training AutoML Models Using Vertex AI
- Module 4: BigQuery Machine Learning: Develop ML Models Where Your Data Lives
- Module 5: Optimization
- Module 6: Generalization and Sampling

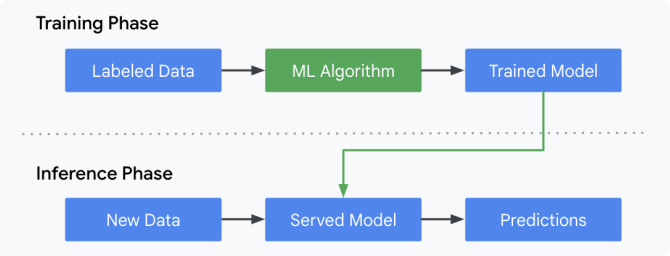
Summary

The course begins with a discussion about data: how to improve data quality and perform exploratory data analysis. We describe Vertex AI AutoML and how to build, train, and deploy an ML model without writing a single line of code. You will understand the benefits of Big Query ML. We then discuss how to optimize a machine learning model and how generalization and sampling can help assess the quality of ML models for custom training.

Key takeaways

Module 1: Get to Know Your Data: Improve Data through Exploratory Data Analysis

There are two phases in machine learning: a **training phase and an inference phase**.



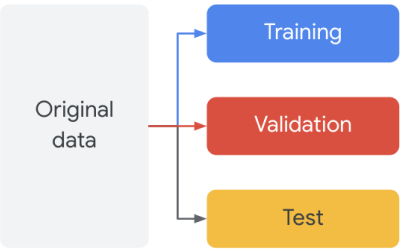
In any ML project, after you define the business use case and establish the success criteria, the process of delivering an ML model to production involves the following steps:

- Data extraction
- Data analysis
- Data preparation
- Model training
- Model evaluation
- Model validation

These steps can be completed manually or can be completed by an automated pipeline.

How do you **improve data quality**?

- Resolve missing values
- Convert the Date feature column to Datetime Format
- Parse date/time features
- Remove unwanted values
- Convert categorical columns to “one-hot encodings”

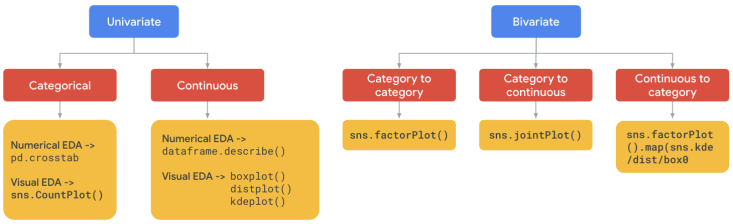


Your original source data will typically be split into a training, validation, and a test set.

The quality of your source data will influence the predictive value of your model.

In statistics, **exploratory data analysis (EDA)** is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. EDA is a set of techniques that allows analysts to quickly look at data for trends, outliers, and patterns.

Three popular data analysis approaches are classical, exploratory, and Bayesian.



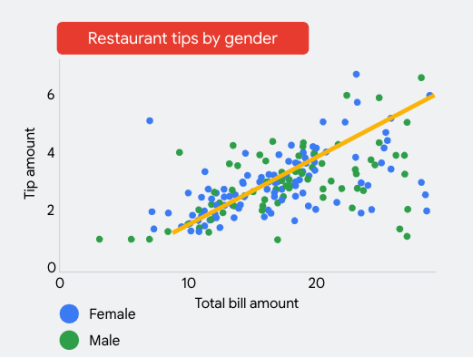
How is EDA used in machine learning? The EDA approach allows the data to suggest admissible models that best fit the data. It is typically performed using the following methods: **univariate and bivariate**.

Module 2: Machine Learning in Practice

Two of the most common classes of machine learning models are **supervised and unsupervised ML models**. Supervised models have labels, and unsupervised models do not have labels.

Within supervised ML there are two types of problems: **regression** and **classification**.

- In regression problems, the goal is to use mathematical functions of different combinations of our features to predict the continuous value of our label.

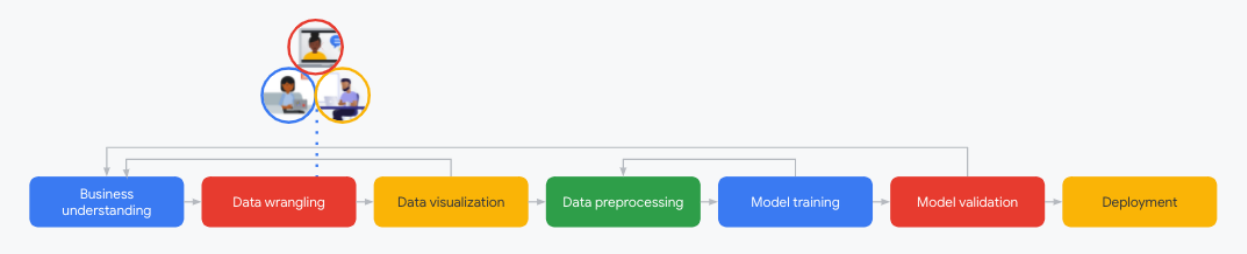


- In classification problems, instead of trying to predict a continuous variable, we are trying to create a decision boundary that separates the different classes.

With unsupervised ML, you will have to use **clustering** algorithms to discover interesting properties of the data.

Module 3: Training AutoML Models Using Vertex AI

Once you've established the business requirements and addressed your data requirements, you then find yourself continuing down the path of a machine learning pipeline or workflow, where you now need to train the model, evaluate it to see if the metrics are right, and then deploy it to production.



Machine learning versus statistics:

	Machine learning	Standard statistics (linear/logistic regressions)
Data preparation?	Doesn't require explicit commands to find patterns in data	Need to know variables and parameters beforehand
Hypothesis	No hypothesis needed	Need hypothesis to test

Type of data?	Multi-dimensional data that can be non-linear in nature	Linear data
Training?	Needs to be "trained"	No training
Goal?	Generally better for predictions	Generally better for inferences/hypothesis testing
Scientific question?	What will happen?	How/why did it happen?

What is **AutoML**?

Vertex AI automates the following components in the machine learning pipeline: Data readiness, feature engineering, training and hyper-parameter tuning, model serving, explainability and interpretability, and the ability to deploy to edge devices.

You can use Vertex AI to manage the following stages in the ML workflow:

- Create a dataset and upload data
- Train an ML model on your data
  - Train the model
  - Evaluate model accuracy
  - Tune hyperparameters (custom training only)
- Upload and store your model in Vertex AI.
- Deploy your trained model to an endpoint for serving predictions.
- Send prediction requests to your endpoint
- Specify a prediction traffic split in your endpoint
- Manage your models and endpoints.

Regression metrics:

Mean absolute error (MAE): MAE is the average absolute difference between the target and predicted values. It measures the average magnitude of the errors—the difference between a target and predicted value—in a set of predictions.

Mean absolute percentage error (MAPE): MAPE is the average absolute percentage difference between the labels and the predicted values.

Root mean square error (RMSE): RMSE is the square root of the average squared difference between the target and predicted values.

Root-mean-squared logarithmic error (RMSLE): RMSLE is similar to RMSE, except that it uses the natural logarithm of the predicted and actual values plus 1. RMSLE penalizes under-prediction more heavily than over-prediction.

R squared (R^2): R^2 is the square of the Pearson correlation coefficient between the observed and predicted values. It's also known as the coefficient of determination.

Module 4: BigQuery Machine Learning: Develop ML Models Where Your Data Lives

BigQuery ML is an easy-to-use way to invoke machine learning models on structured data using just SQL. It is a set of SQL extensions to support machine learning.

Steps for working with BigQuery ML:

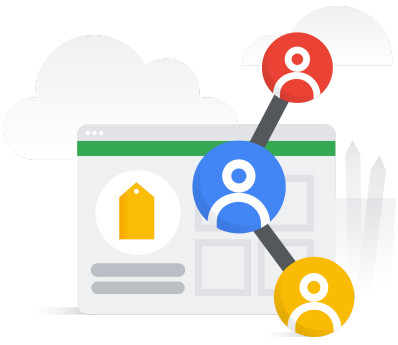
1. Write a SQL query to extract their training data from BigQuery
2. Create a model, specifying model type
3. Evaluate the model and verify that it meets requirements
4. Predict, using the model on data extracted from BigQuery

Hyperparameter tuning

Hyperparameter tuning identifies a set of optimal hyperparameters for a learning algorithm. A **hyperparameter** is a model argument whose value is set before the learning process begins. By contrast, the values of other parameters such as coefficients of a linear model are learned.

Hyperparameter tuning lets you spend less time manually iterating hyperparameters and more time focusing on exploring insights from data.

Recommendation system



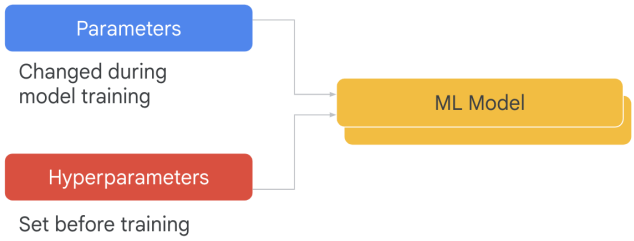
Recommendation systems are machine learning systems that help users discover new products and services.

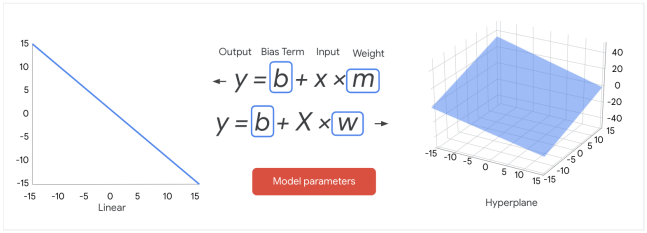
Common use case:

- Prepare your training data in BigQuery
- Train a recommendation system with BigQuery ML
- Use the predicted recommendations in production

Module 5: Optimization

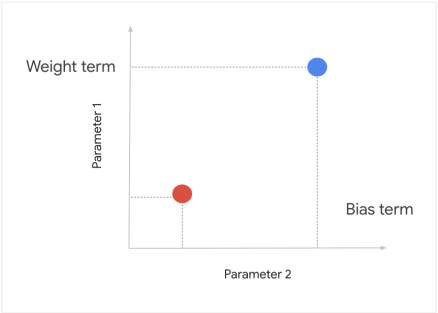
ML models are mathematical functions with parameters and hyperparameters.





Linear models have two types of parameters: **Bias and weight**

Think of optimizing your parameters as searching through parameter-space.

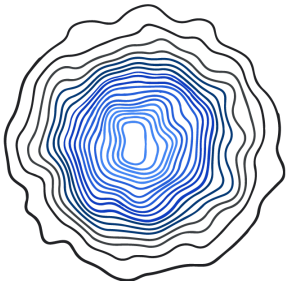


Loss functions are able to take the quality of predictions for a group of data points from your training set and compose them into a single number with which to estimate the quality of the model's current parameters.

Root Mean Squared Error (RMSE) is one loss function metric.

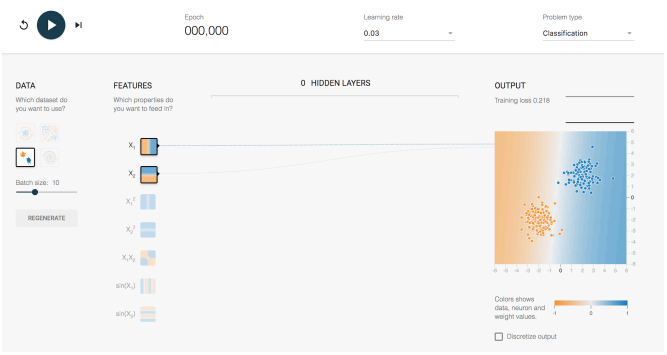
- Get the errors for the training examples
- Compute the squares of the error values
- Compute the mean of the squared error values
- Take a square root of the mean

A lower RMSE indicates a better performing model. But keep in mind that RMSE doesn't work as well for classification.



Gradient descent refers to the process of walking down the surface formed by using your loss function in parameter-space.

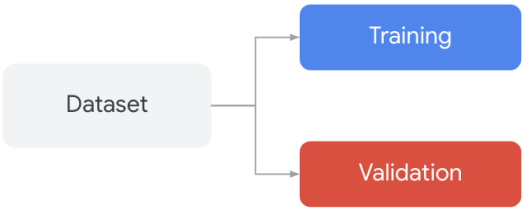
TensorFlow Playground is a tool for visualizing how neural networks work.

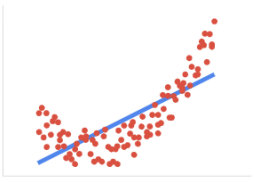


Module 6: Generalization and Sampling

Generalization helps show how a trained model will perform on unseen data, or in production.

Models that generalize well have similar error values across training and validation. As soon as you start seeing your models not performing well against your validation data (for example, if your loss metrics start to increase), it's time to stop.

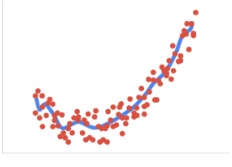




Underfit

An overly simplistic linear model that doesn't fit the relationships in the data is **underfitting**.

An overly complex model that fits the training dataset too well is **overfitting**.



Overfit