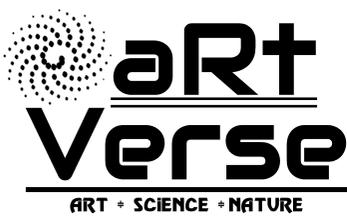## ML to detect patterns in historical data to make new predictions

- Train predictive models using historical data, statistics and mathematics to make predictions about unknown values.
- Function *f* operates on features x to calculate the predicted label (y)
  - *f*(x) = y (function fits features to label, function calculates y from x)
- Data requires pre-processing and specific operations to use - this depends on the model and training algorithm
  - Data Ingestion (gather data)
  - Data pre-processing (feature selection/identify x, clean data, impute missing, feature engineering/derive new features, normalization to same scale),
  - Train model and fit to data,
  - Predict unknown values.

Concepts

aRtVerse
ART • SCIENCE • NATURE

Setup

Prerequisites

- Azure subscription
- Azure ML workspace

Data Requirements

- Tabular format
- Value (i.e., label or target column) that you want to predict must be in dataset

Basic models

ML Algorithms

- **Supervised Learning:** Historical data points with known labels
  - **Regression:** y is numerical
  - **Time Series:** regression with time element, predict feature at a point in time
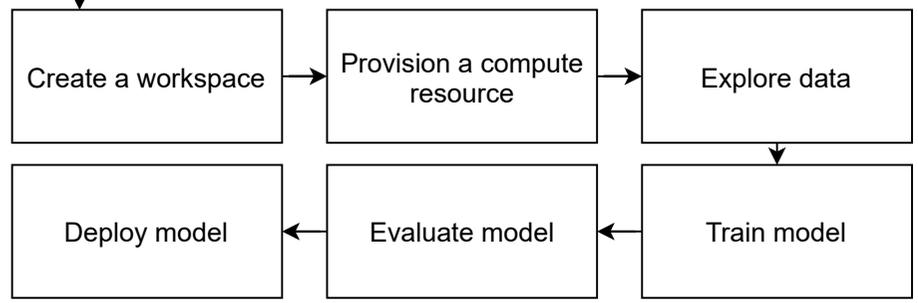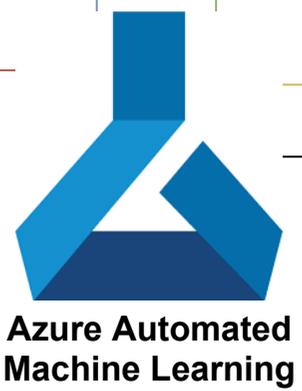 - **Classification:** y is categorical

Services

Azure Automated Machine Learning

_Modelling Steps

MS AZ AutoML

- No-code ML as a service
- AutoML to build and operate ML solutions in the cloud
- Prepare data, train models, publish predictive services, and monitor their usage.
- Increase data scientist efficiency by automating time consuming processes
- Scale solutions for large volumes of data with pay-as-you-use costs
- Automate time insensitive tasks
- Can add custom code when required (Azure Machine Learning Python SDK)

Create a workspace → Provision a compute resource → Explore data → Train model → Evaluate model → Deploy model

Step 1: Create a workspace

- Via the Azure portal
  - Provide subscription, resource group, workspace name, region, storage account, key vault, application insight and container registry info
- Sign into Azure Machine Learning Studio using Azure directory, subscription & workspace
  - Manage assets and resources via workspace

Step 2: Create a compute resource

- **Compute targets:** cloud-based resources on which to run models and explore data
  - **Compute Instances:** Development workstations that data scientists can use to work with data and models
  - **Computer Clusters:** Scalable clusters of virtual machines for on-demand processing of experiment code
  - **Inference Clusters:** Deployment targets for predictive services on trained models
  - **Attached Compute:** Link to existing resources (VMs and DBs)
- Compute Instances tab
  - Add new compute: provide compute name, VM type and VM size
  - Create compute cluster (while compute instance is pending): provide location, VM priority, VM type, VM size, compute name, min & max nodes, idle seconds, SSH access

Step 3: Explore the Data

- Data is encapsulated in a dataset object - for model training & operations
- Datasets page
  - Create new dataset: provide web url/location, name, dataset type, description, skip validation, file format, delimiter, encoding, column headers, skip rows, multi-line data, schema and confirm details

Step 4: Train the model

- Automated machine learning leverages cloud compute scaling to pre-process and train multiple algorithms in parallel for best performing supervised ML model
  - Classification
  - Regression
  - Time series forecasting
- Automated ML page
  - Create new: provide dataset, name, target column, compute cluster, task type (model), primary (evaluation) metric, explain best model, blocked algorithms, exit criterion (time, score), featurization.
- After completion, Models can be viewed and their scores

Step 5: Evaluate the model

- Review the best model on Details tab & select Algorithm name
- Metric calculated using cross-validation to iteratively tests the trained model with data not present in training data & compare the predicted and actual values
  - Difference between predicted and actual value is the residual and indicates the amount of error in the model (smaller value indicates better model accuracy).
  - Normalize root mean squared error (from residuals)
  - Predicted vs True chart with diagonal trend to show correlation between values
  - Residual Histogram is frequency of residual value ranges, indicates the variance that cannot be explained by the model. Most errors should be clustered around 0.
- Explanations tab > Explanation ID
  - View dashboard of feature importance (Global importance)

Step 6: Deploy model as a service

- As an Azure Container Instance (ACI) or Azure Kubernetes Service (AKS)
  - Production should have AKS with inference cluster compute targets
- On Model tab > Deploy and provide name, description, compute type and enable authentication
- On Endpoint page > select model > consume tab which shows REST Endpoint and Primary Key for your services
  - Notebook page will have model code: file location, file name, file type and overwrite settings.
  - Replace REST endpoint & Primary for service in notebook space provided.
  - Run notebook to complete testing.