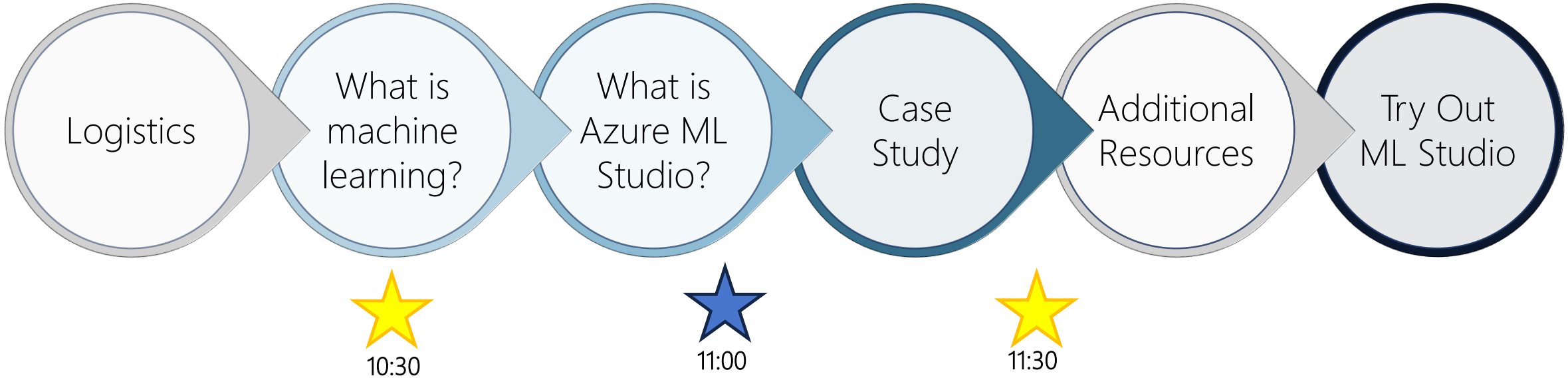
Getting Started with Azure ML Studio

MAY 26, 2017

Paige Bailey
Paul DeCarlo

Roadmap


Logistics
What is machine learning?
What is Azure ML Studio?
Case Study
Additional Resources
Try Out ML Studio
10:30
11:00
11:30
5-minute stretch break
10-minute stretch break

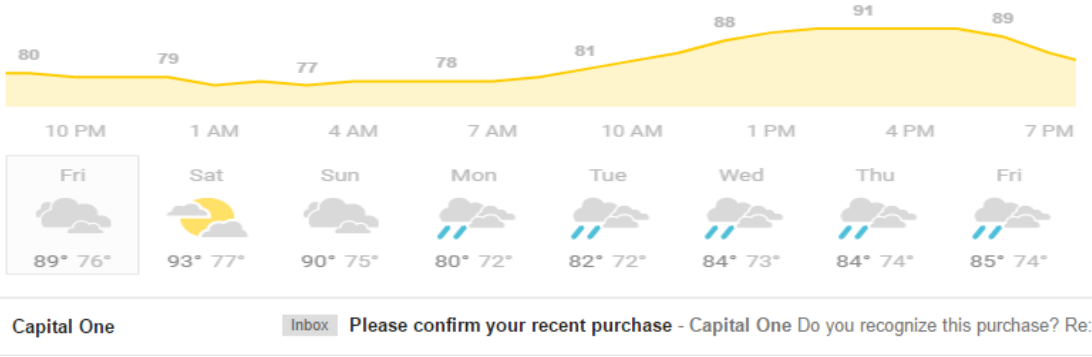
What is Machine Learning?

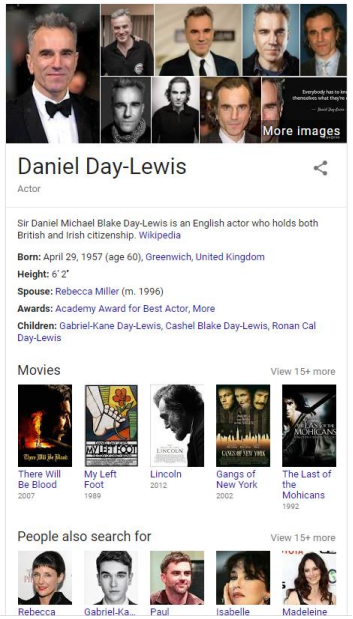
DATA EXPLORATION // TYPES OF LEARNING // MODEL TRAINING AND EVALUATION

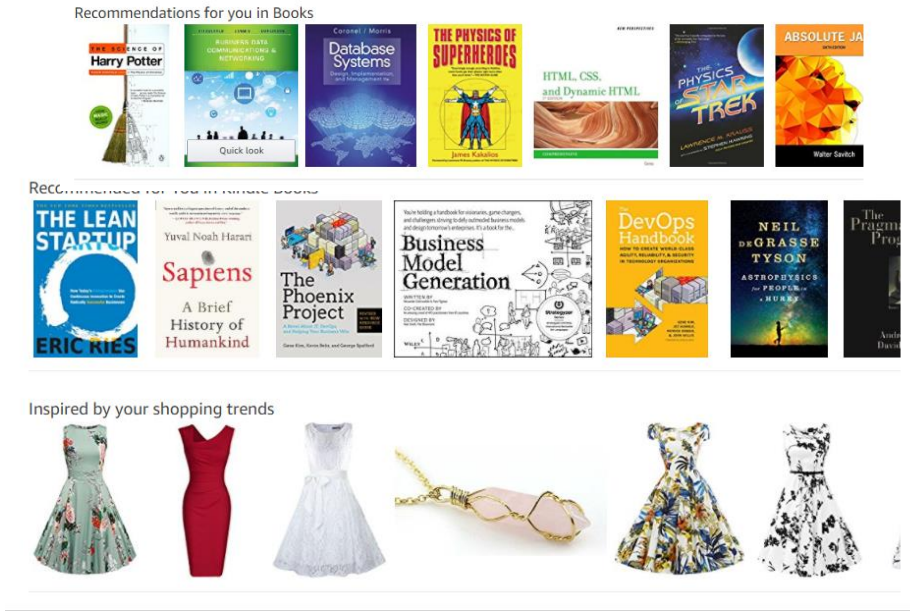
What is machine learning?

- Data science technique that helps computers learn from existing data in order to **forecast** future behaviors, outcomes, and trends.

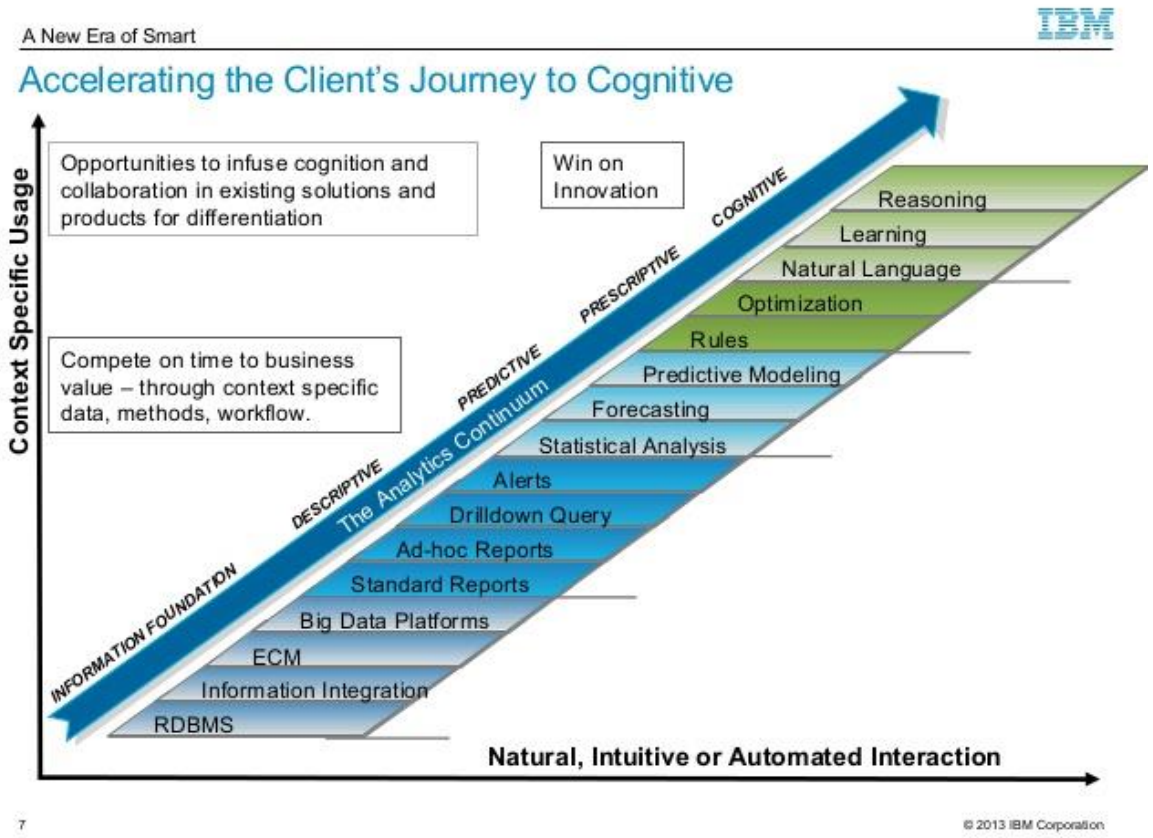
- Examples
  - Amazon and Netflix recommendations
  - Fraud detection for credit card transactions
  - Roomba deciding when a room is clean
  - Self-driving cars
  - Student performance predictions





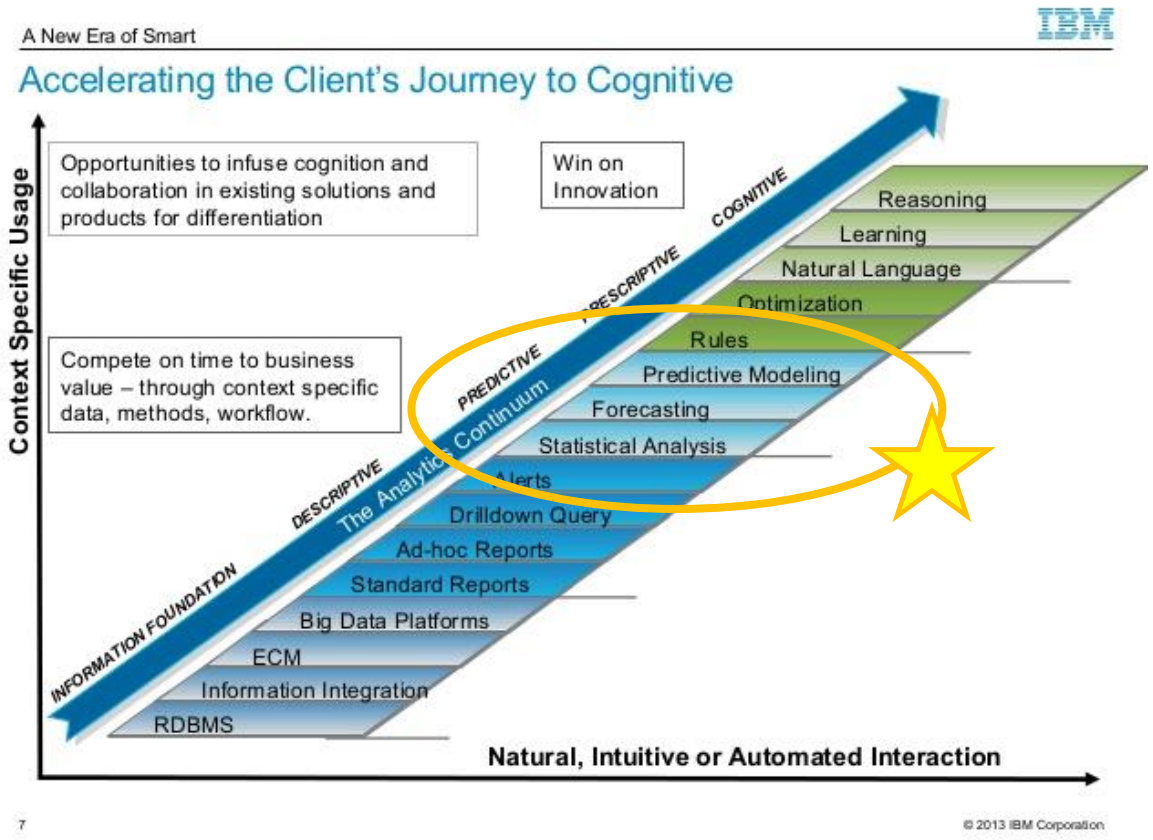


Data exploration, descriptive analytics, and predictive analytics



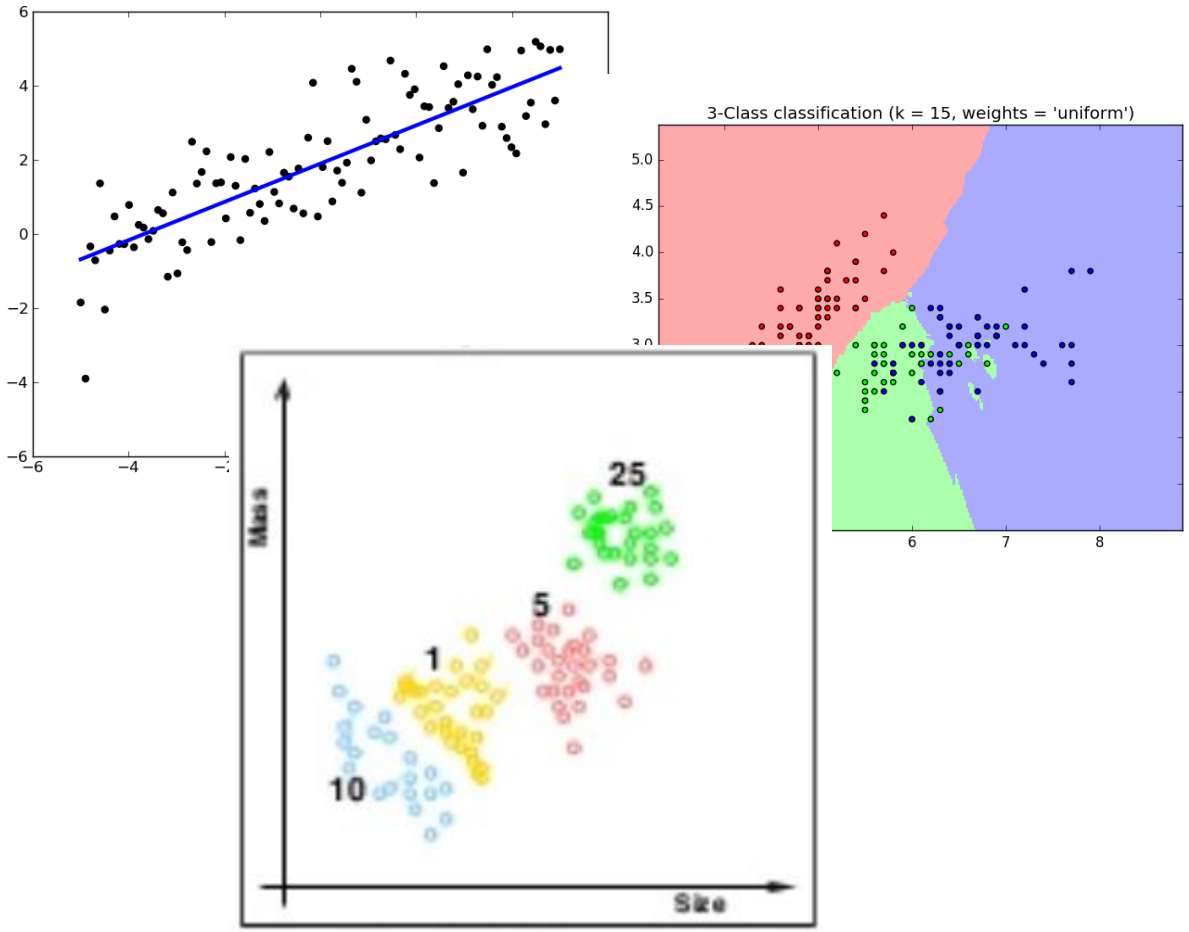
- Data Exploration: process of gathering information about a large and often unstructured data set in order to find characteristics for focused analysis.
- Data Mining: automated data exploration.
- Descriptive Analytics: process of analyzing a data set in order to summarize what happened.
  - Most **business analytics** falls into this category
  - **Examples**: sales reports, web metrics, social network analysis
- Predictive Analytics: process of building models from historical or current data in order to forecast future outcomes.

Data exploration, descriptive analytics, and predictive analytics



- Data Exploration: process of gathering information about a large and often unstructured data set in order to find characteristics for focused analysis.
- Data Mining: automated data exploration.
- Descriptive Analytics: process of analyzing a data set in order to summarize what happened.
  - Most **business analytics** falls into this category
  - **Examples**: sales reports, web metrics, social network analysis
- Predictive Analytics: process of building models from historical or current data in order to forecast future outcomes.

Supervised and Unsupervised Learning

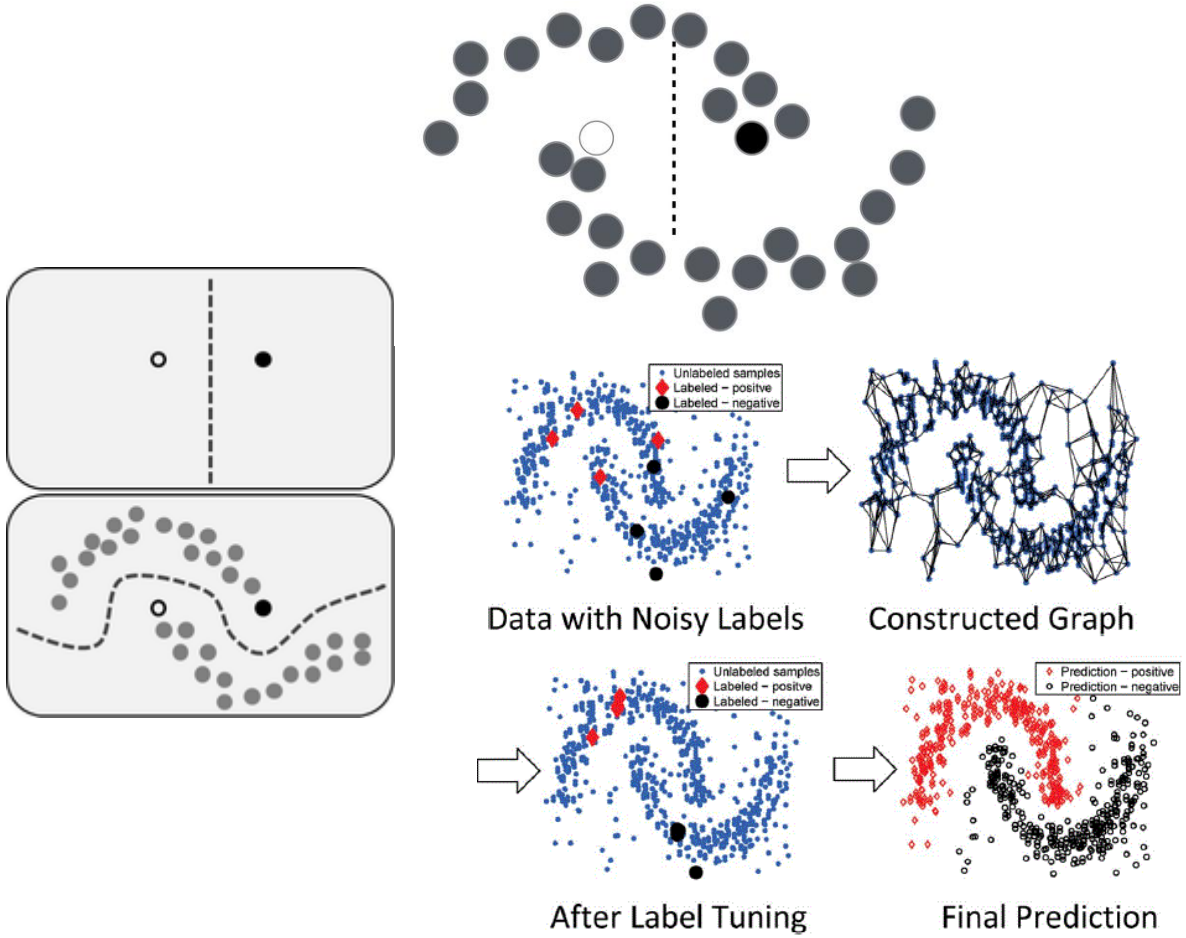


- Supervised Learning: algorithms are trained with labeled data.
  - Classification: used when the output variable is a category, such as “red” / “blue” or “disease” / “no disease”.
  - Regression: often used when the output variable is a real, numeric value – like dollars, weights.

Unsupervised Learning: used on data with no labels, with the goal of finding relationships in the data.

- Clustering: used to discover groupings in the data, such as grouping by purchasing behavior.
- Association: rules that describe large portions of your data, such as “people who buy X also tend to buy Y”.

Semi-Supervised Learning

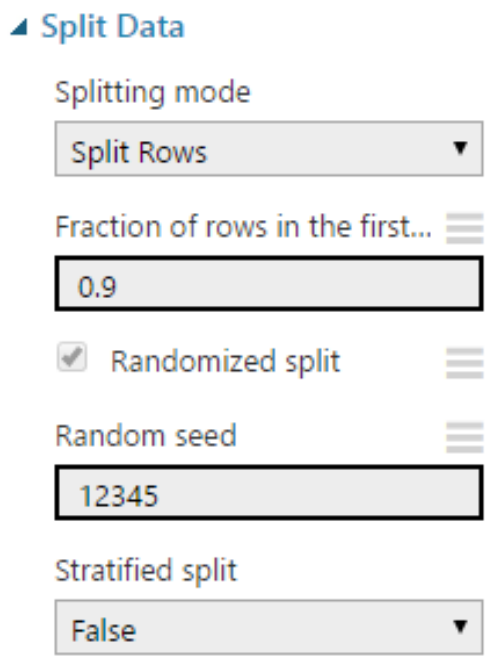


- Used when you have a large amount of input data (X) and **only some of the data is labeled** (Y).
- **Example**: photo archive where only some of the images are labeled, and the majority are unlabeled.
- Many real-world machine learning problems fall into this area
  - *Expensive and time-consuming* to label data
  - Unlabeled data is *cheap and easy* to collect and store
- Can also use unsupervised learning techniques to make **best-guess predictions** for the unlabeled data, then feed that data back into the supervised learning algorithm as training data.

Model training and evaluation

- A machine learning model is an abstraction of the question you are trying to answer, or the outcome you want to predict.
- Models are trained and evaluated from existing data.

- **Training:** data used to fit the model (typically falls into the range of 70 – 90%)
- **Evaluation (Test):** data used only at the end of the model building and selection process to assess how well the final model might perform on future data
- **Holdout (Validation):** not used to fit or initially test a model, used to asses the performance of that model



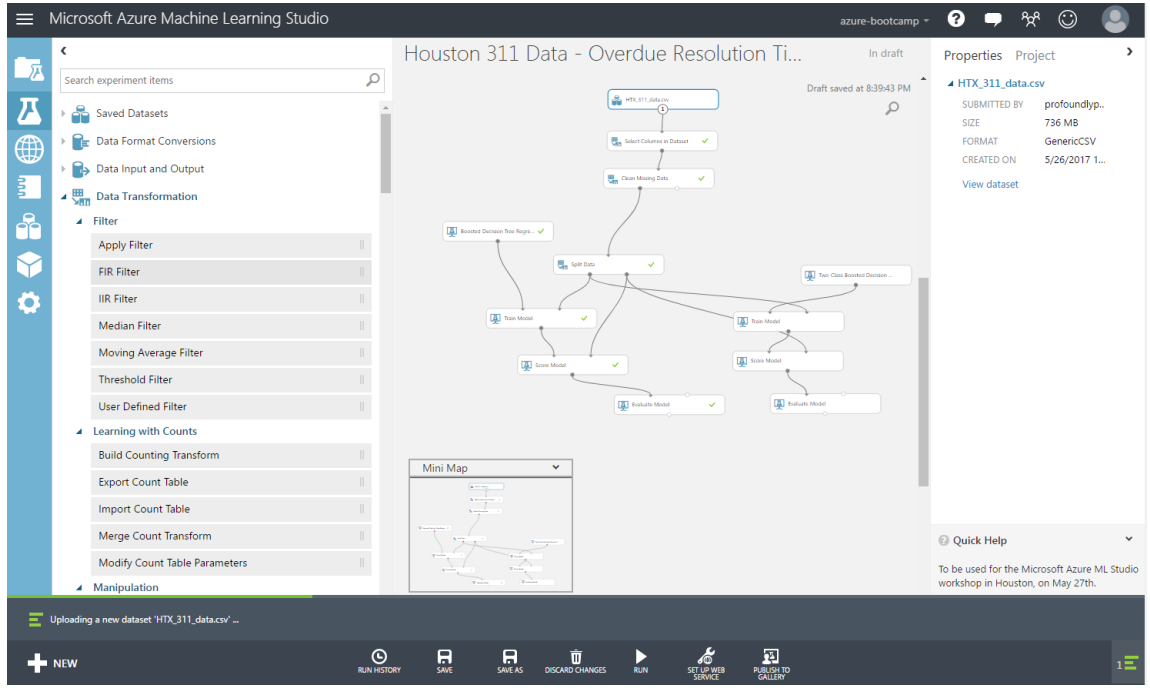
Additional vocabulary

- Algorithm: self-contained set of rules used to solve problems through data processing, math, or automated reasoning.
- Anomaly detection: model that flags unusual events or values and helps you discover problems.
- Categorical data: data that is organized by categories, and that can be divided into groups.
- Classification: model for organizing data points into categories based on a data set for which categorical groupings are already known.
- Feature engineering: process of extracting or selecting features related to a data set in order to enhance the data set and improve outcomes.
- Model: supervised learning model is the product of a machine learning experiment comprised of training data and an algorithm.
- Numerical / quantitative: data that has meaning as measurements (continuous data) or counts (discrete data).
- Partition: method by which you divide data into samples.
- Prediction: forecast of a value or values from a machine learning model.
- Regression: model for predicting a value based on independent variables.
- Score: predicted value generated from a trained classification or regression model.
- Sample: part of a data set intended to be representative of the whole. Can be selected randomly, or based on specific features of the data set.

What is Azure Machine Learning Studio?

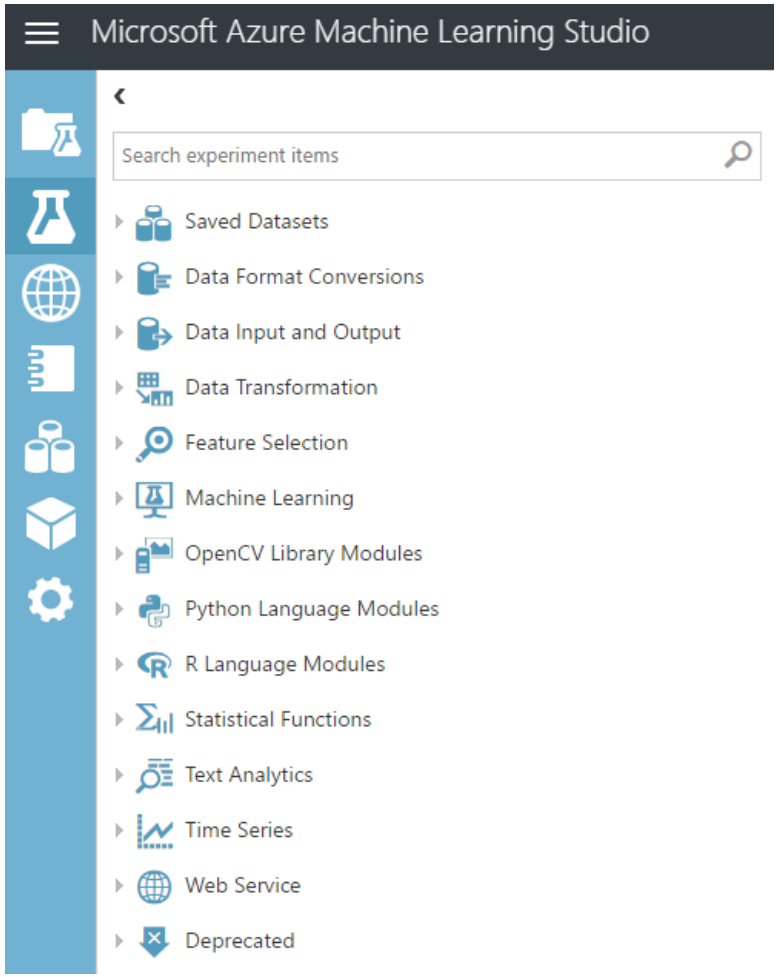
What is Azure Machine Learning Studio?

- Collaborative, drag-and-drop tool you can use to **build, test, iterate, and deploy** predictive analytics solutions on your data
- Publishes models as **web services** that can be easily consumed by custom apps or BI tools such as Excel
- Interactive, visual workspace with a library of ready-to-use algorithms
- **No programming required**
- Why would that be useful?
  - RStudio example



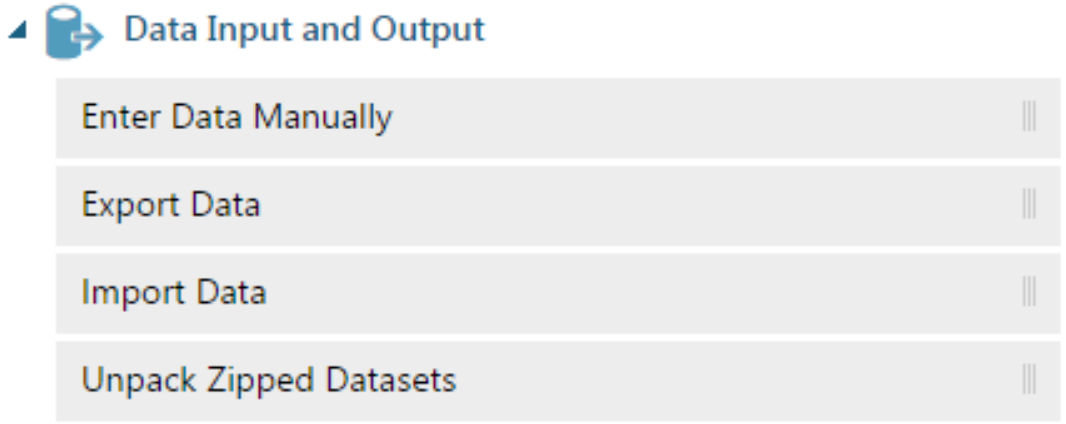
What is a module?

- Each **module** represents a set of code that can run independently and perform a machine learning task, given the required inputs.
- Viewable in the left-hand pane of ML Studio.
- A module may contain a particular **algorithm**, or perform a task that is important in machine learning, such as **missing value replacement** or **statistical analysis**.
- Modules in ML Studio are organized by **functionality** – and more are being added (and deprecated!) by the month.



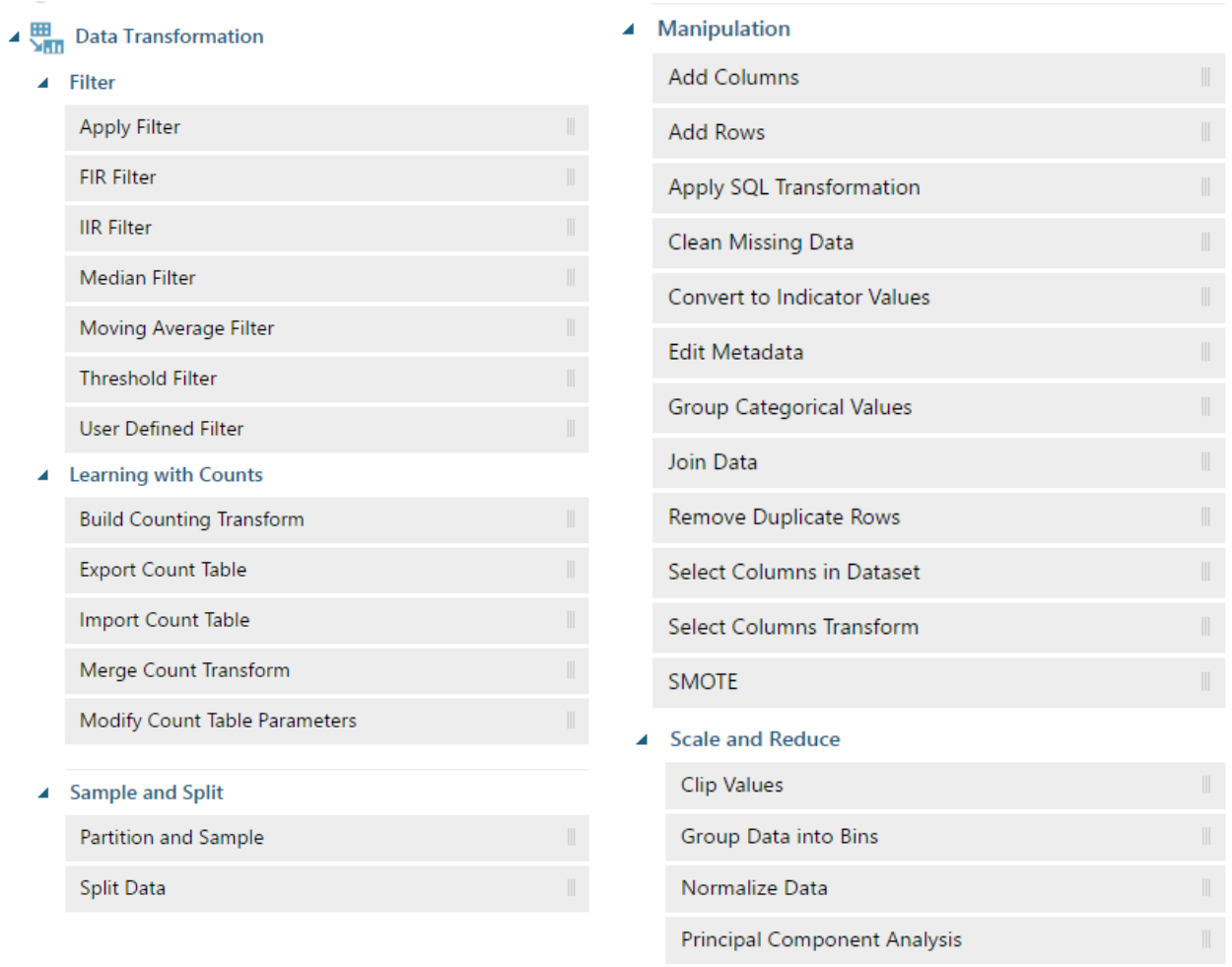
Modules – Data Input and Output

- Enter data manually: small, single-column data set generated by typing values.
- Export Data: used to save results, intermediate data, and working data from experiments into cloud storage destinations.
  - Hive Query, Azure SQL Database, Azure Table, Azure Blob Storage
- Import Data: used to load data into an experiment from existing cloud services outside of ML Studio.
  - Web URL, Hive Query, Azure SQL Database, Azure Table, Azure Blob Storage, Data Feed Providers, On-Premises SQL Server Database, DocumentDB
- Unpack zipped data sets: get compressed files and unzip them for use in an experiment.



Modules – Data Transformation

- Filter: can be applied to numeric data to support machine learning tasks such as image recognition, voice recognition, and waveform analysis.
- Learning with Counts: develop compact features for use in machine learning
- Manipulate: includes merging data sets, cleaning missing values, grouping and summarizing data, changing column names and data types, and indicating which column is a label or a feature.
- Sample and Split: divides data into training and test sets, splits data by percentage or by a filter condition, and performs sampling.
- Scale and Reduce: prepares numerical data for analysis by applying normalization or by scaling. Bin data into groups, remove or replace outliers, perform principal component analysis (PCA).

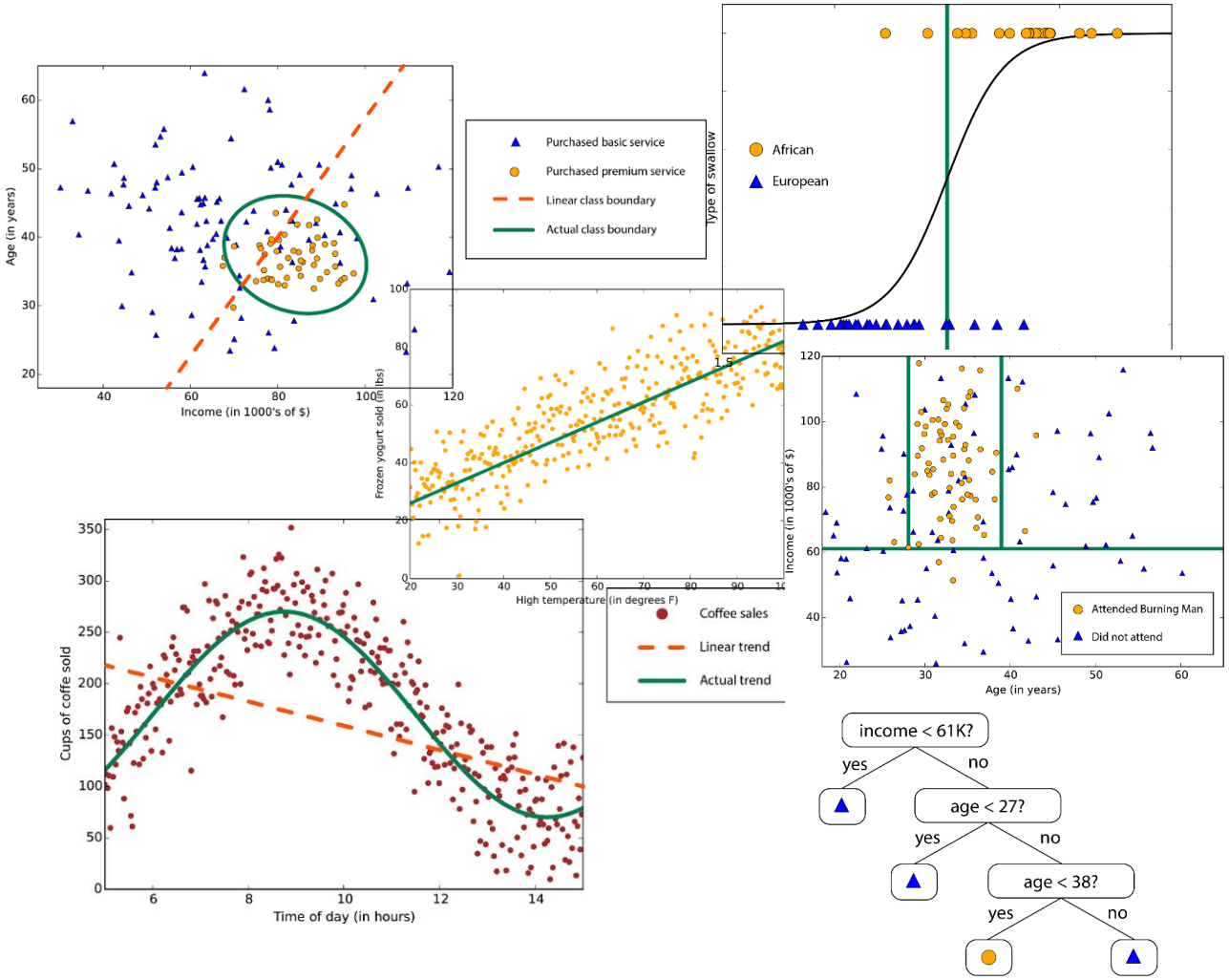


Modules – Machine Learning

Typical workflow for machine learning:

- Identifying a problem to solve and a metric for measuring results
- Finding, cleaning, and preparing appropriate data
- Identifying features and engineering new features
- Building, evaluating, and tuning models
- Using models to generate predictions, recommendations, and other results

- For ML Studio, these processes fall into the broad buckets of **Initialize, Train, Score, and Evaluate**.



Data Preprocessing – Missing Values

- Why is this important?
- For **missing values**, you'll either want to replace them with statistical methods or remove them from the data set completely.
  - Fill in the missing value manually (<10%)
  - Use a global constant (example: NA)
  - Attribute mean or most probable value
- For **noisy data** (incorrect values), you'll want to identify those outliers and smooth out.
  - Binning
  - Regression
  - Clustering

Less Data

Higher Accuracy

Simplify Results

Fewer Attributes

Data Preprocessing – Consolidation

- Smoothing
  - Remove noise from the data
  - Binning, regression, and clustering
- Aggregation
- Generalization
- Normalization
  - Min-max normalization
  - Z-Score Normalization
  - Normaliization by decimal scaling
- Attribute Construction
  - Constructed from given attributes and added in order to help improve accuracy, understand structure of high-dimensional data

Min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

z-score normalization

$$v' = \frac{v - \mu_A}{\sigma_A}$$

Normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

where j is the smallest integer such that $Max(|v'|) < 1$

OVERDUE	OVERDUE_STATUS
10.80	TRUE
10.76	TRUE
-14.99	FALSE
-14.99	FALSE
334.20	TRUE
-158.39	FALSE

Data Preprocessing – Reduction

- Can be applied to obtain a reduced representation of the data set that is much smaller in volume, but closely maintains the integrity of the original data.

Options:

- Data Cube Aggregation
- Attribute (Subset) Selection
- Dimensionality Reduction
- Numerosity Reduction
- Data Discretization
- Concept Hierarchy Generation

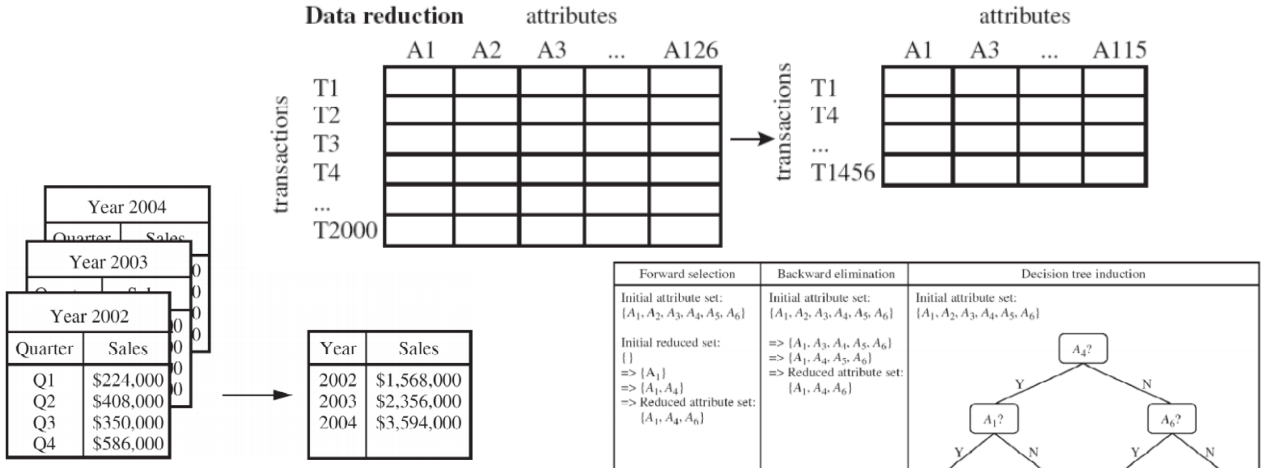


Figure 2.13 Sales data for a given branch of *AllElectronics* for the years 2002 to 2004. On the left, the sales are shown per quarter. On the right, the data are aggregated to provide the annual sales

Figure 2.15. Greedy (heuristic) methods for attribute subset selection

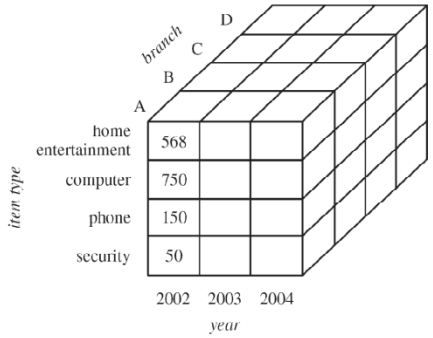
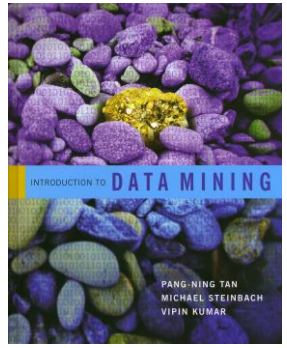
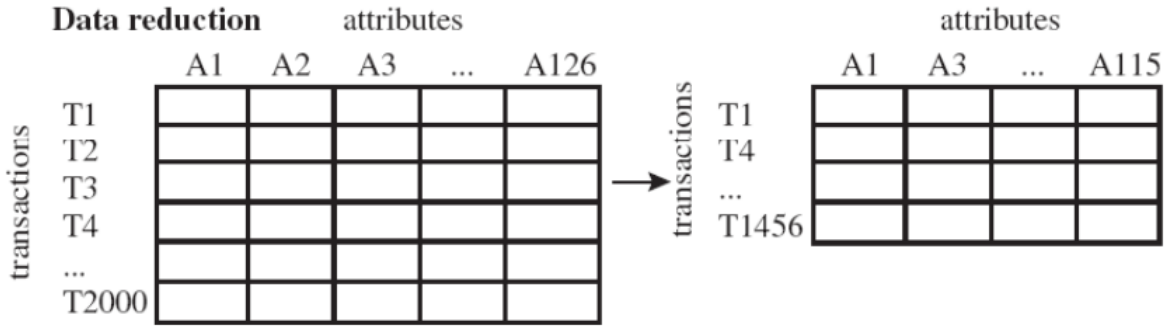


Figure 2.14 A data cube for sales at *AllElectronics*

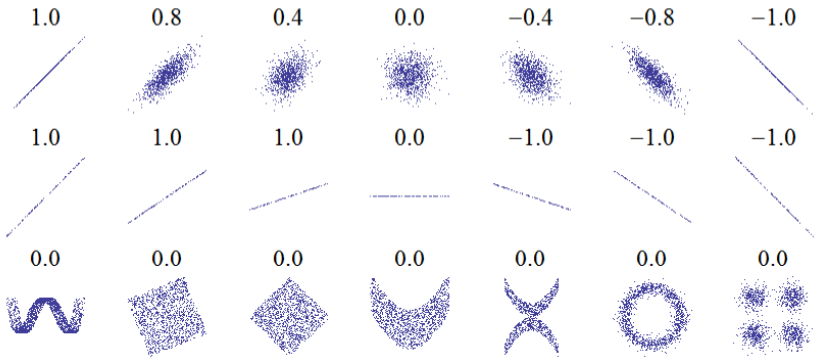


Data Preprocessing – Dimensionality

- "The Curse of Dimensionality"
  - *Size, Radius, Distance, Outlier*
- Data Cube Aggregation
  - Summarize data based on dimensions
  - Resulting data set is smaller in volume, without loss of information necessary for analysis task
- Attribute Selection
  - Removes dimensionality, reduces noise
  - Improves performance (speed of learning, predictive accuracy, simplicity of rules)
- Data Discretization
  - Transforms quantitative data into qualitative data
  - Interval labels are used to replace attributes



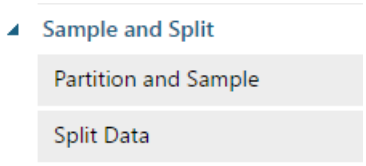
Data transformation $-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

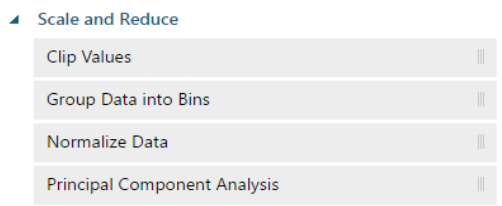


Data Preprocessing – 1- slide Overview

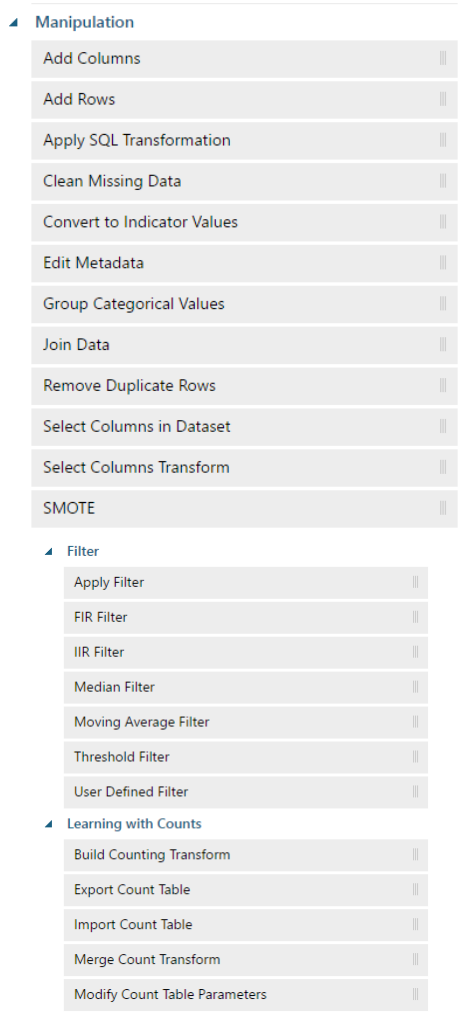
- Data Cleaning
  - Missing Values: remove or replace
  - Noisy Data: binning, regression, clustering
- Data Integration
  - Entity ID Problem: incorporate metadata
  - Redundancy: correlation analysis

    (correlation coefficient, chi-square test)
- Data Transformation
  - Smoothing: see data cleaning
  - Aggregation: see data reduction
  - Generalization: see data reduction
  - Normalization: min-max; z-score; decimal scaling
  - Attribute Construction





- Data Reduction
  - Data cube aggregation
    - Multidimensional aggregated information
  - Attribute subset selection
    - Stepwise forward selection; stepwise backward selection; combination; decision tree induction
  - Dimensionality reduction
    - Discrete wavelet transforms (DWT); principal component analysis (PCA)
  - Numerosity reduction
    - Regression and log-linear models; histograms; clustering; sampling
  - Data discretization
    - Binning; histogram analysis; entropy-based discretization; interval merging by chi-square analysis; cluster analysis; intuitive partitioning
  - Concept hierarchy generation



Machine Learning Modules

INITIALIZE // TRAIN // SCORE // EVALUATE

Machine Learning - Initialize

- Azure ML Studio provides many different algorithms to help build analytical models.
- However, algorithm selection and parameter tuning is the responsibility of the data scientist.
- https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-algorithm-choice

- Categories of Learning Algorithms
  - Anomaly detection
  - Classification
  - Clustering
  - Regression
  - Instance-Based
  - Regularization
  - Decision Tree
  - Bayesian
  - Association Rule Learning
  - Artificial Neural Network
  - Deep Learning
  - Dimensionality Reduction
  - Ensemble

Algorithm	Accuracy	Training time	Linearity	Parameters	Notes
Two-class classification					
logistic regression		•	•	5	
decision forest	•	○		6	
decision jungle	•	○		6	Low memory footprint
boosted decision tree	•	○		6	Large memory footprint
neural network	•			9	Additional customization is possible
averaged perceptron	○	○	•	4	
support vector machine		○	•	5	Good for large feature sets
locally deep support vector machine	○			8	Good for large feature sets
Bayes' point machine		○	•	3	
Multi-class classification					
logistic regression		•	•	5	
decision forest	•	○		6	
decision jungle	•	○		6	Low memory footprint
neural network	•			9	Additional customization is possible
one-v-all	-	-	-	-	See properties of the two-class method selected
Regression					
linear		•	•	4	
Bayesian linear		○	•	2	
decision forest	•	○		6	
boosted decision tree	•	○		5	Large memory footprint
fast forest quantile	•	○		9	Distributions rather than point predictions
neural network	•			9	Additional customization is possible
Poisson			•	5	Technically log-linear. For predicting counts
ordinal				0	For predicting rank-ordering
Anomaly detection					
support vector machine	○	○		2	Especially good for large feature sets
PCA-based anomaly detection		○	•	3	
K-means		○	•	4	A clustering algorithm

Algorithm properties:

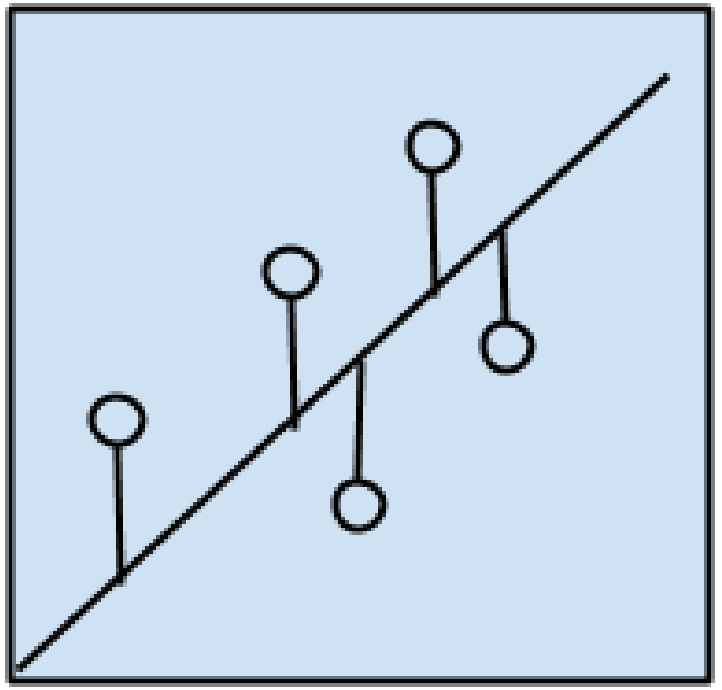
• - shows excellent accuracy, fast training times, and the use of linearity

○ - shows good accuracy and moderate training times

Regression

- Concerned with modeling the relationship between variables that is iteratively refined using a measures of error in the predictions made by the model.

- Most popular regression algorithms:
  - Ordinary Least Squares
  - Linear Regression
  - Logistic Regression
  - Stepwise Regression
  - Multivariate Adaptive Regression Splines (MARS)
  - Locally Estimated Scatterplot Smoothing (LOESS)

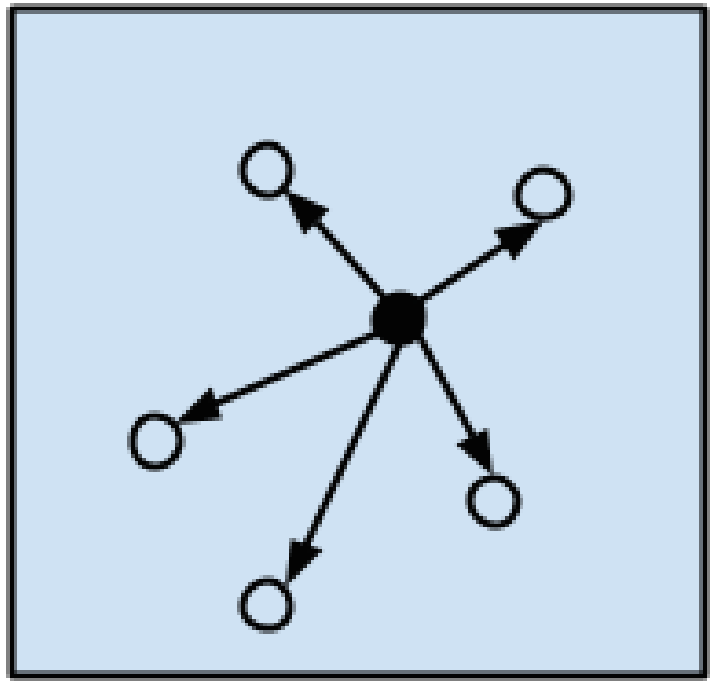


Regression Algorithms

Instance-based

- Decision problem with instances or examples of training data that are deemed important or required to the model.
- Typically build up a database of example data, and then compare new data to that database.

- Most popular instance-based algorithms:
  - K-Nearest Neighbor (kNN)
  - Learning Vector Quantization (LVQ)
  - Self-Organizing Map (SOM)
  - Locally Weighted Learning (LWL)

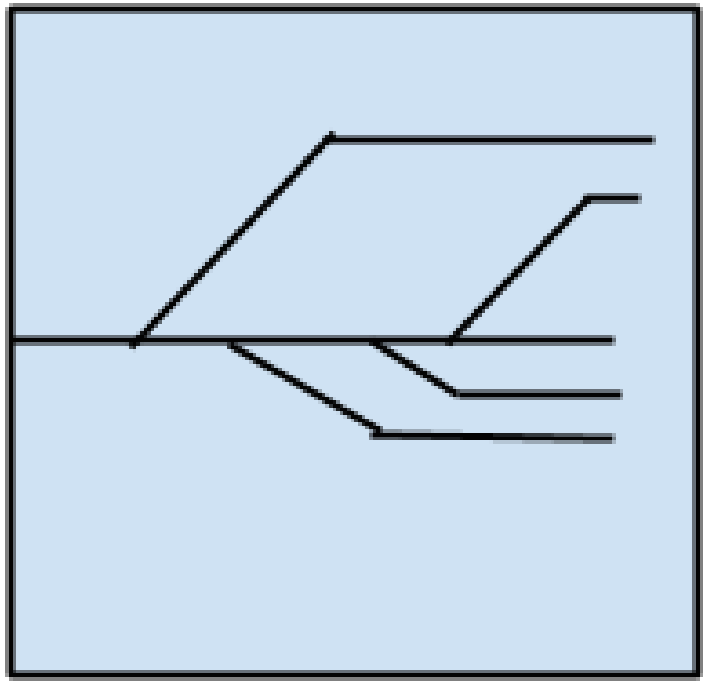


Instance-based Algorithms

Regularization

- Extension made to another method (typically regression methods) that penalizes models based on their complexity, favoring simpler models that are more generalizable.

- Most popular regularization algorithms:
  - Ridge Regression
  - Least Absolute Shrinkage and Selection Operator (LASSO)
  - Elastic Net
  - Least-Angle Regression (LARS)

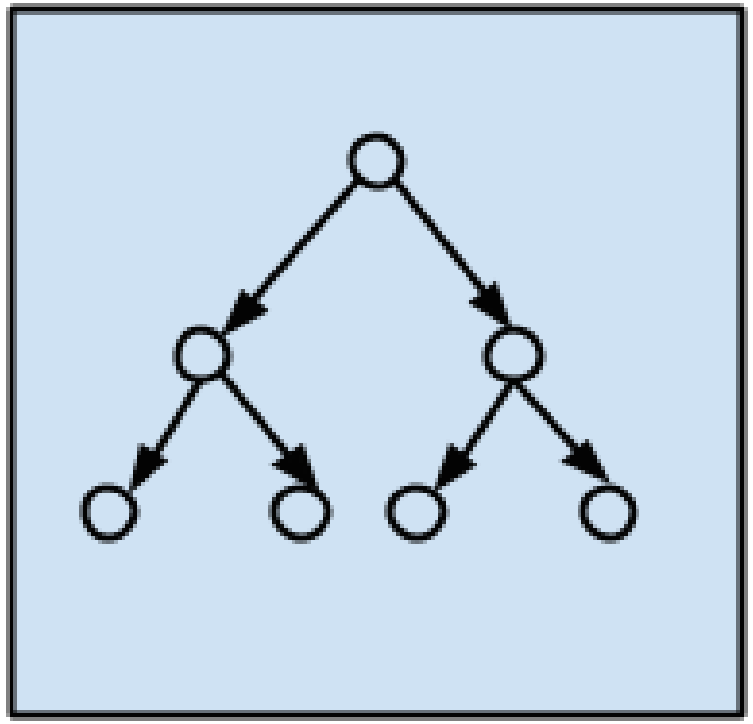


Regularization Algorithms

Decision Trees

- Construct a model of decisions based on actual values of attributes in the data.
- Decisions fork in tree structures until a prediction decision is made for a given record. Decision trees are trained on data for classification and regression problems.

- Most popular decision tree algorithms:
  - Classification and Regression Tree (CART)
  - Iterative Dichotomiser 3 (ID3)
  - C4.5 and C5.0 (different versions of a powerful approach)
  - Chi-squared Automatic Interaction Detection (CHAID)
  - Decision Stump
  - M5
  - Conditional Decision Trees

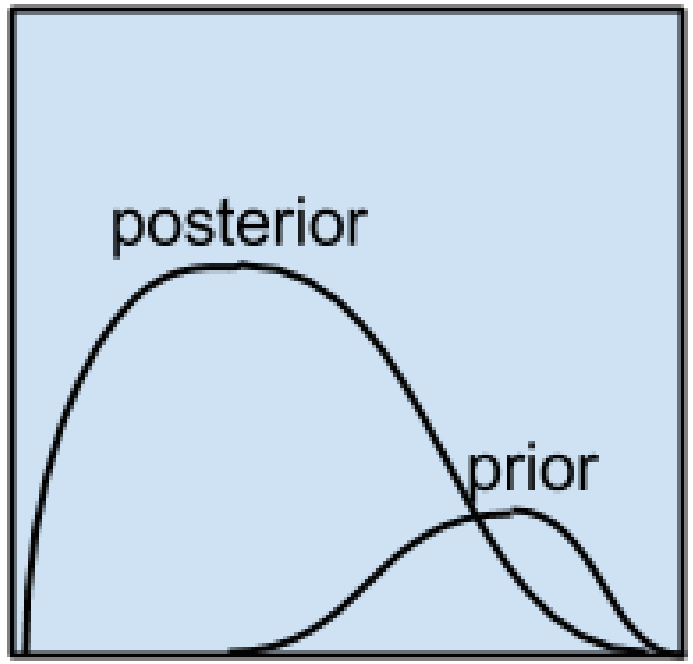


Decision Tree Algorithms

Bayesian

- Used for methods that explicitly apply Bayes' Theorem for problems such as classification and regression.

- Most popular Bayesian algorithms:
  - Naive Bayes
  - Gaussian Naive Bayes
  - Multinomial Naive Bayes
  - Averaged One-Dependence Estimators (AODE)
  - Bayesian Belief Network (BBN)
  - Bayesian Network (BN)

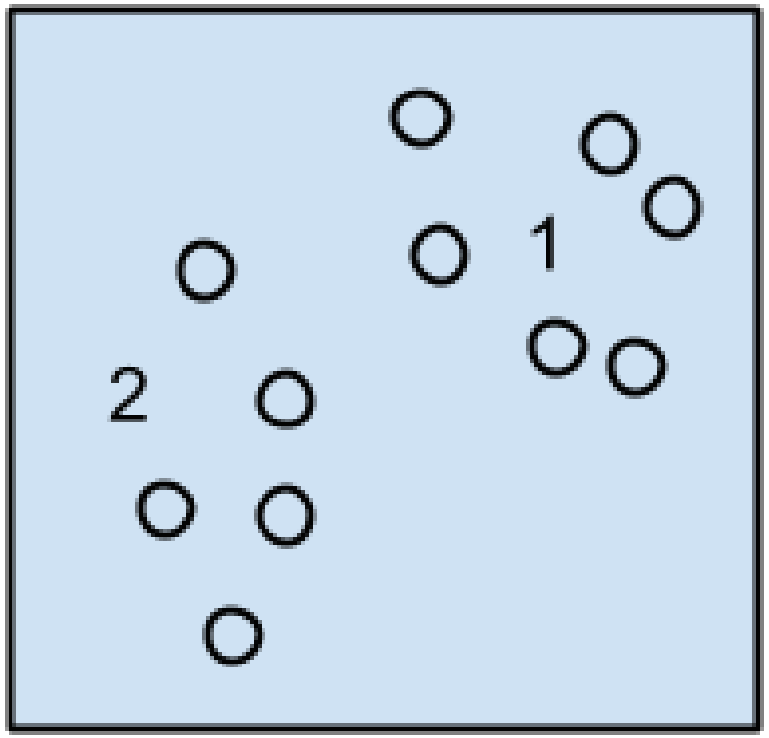


Bayesian Algorithms

Clustering

- "Clustering" describes the class of problems and the class of methods.
- Typically organized by the modeling approaches used (e.g., centroid-based and hierarchical). Concerned with using the inherent structures in the data to best organize the groups of maximum commonality.

- Most popular clustering algorithms:
  - K-Means
  - K-Medians
  - Expectation Maximization (EM)
  - Hierarchical Clustering

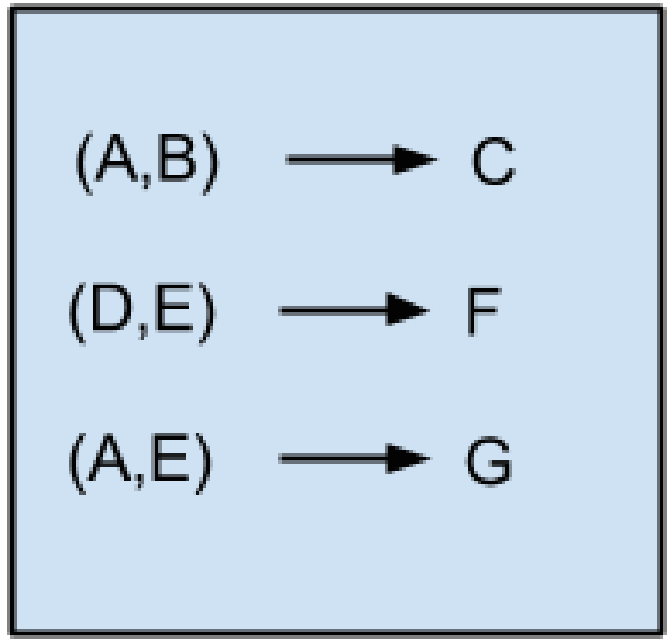


Clustering Algorithms

Association Rule Learning

- Extract rules that best explain observed relationships between variables and data.
- Can discover commercially useful associations in large, multi-dimensional datasets.

- Most popular association rule learning algorithms:
  - Apriori algorithm
  - Eclat algorithm

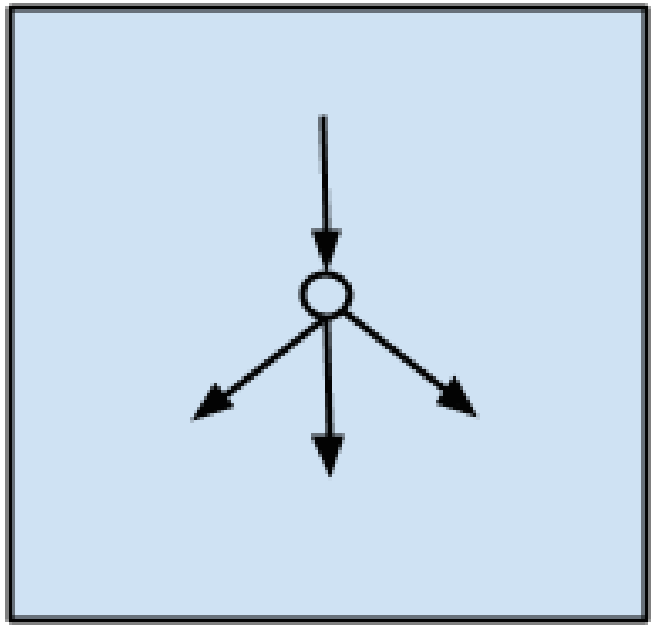


Association Rule Learning Algorithms

Artificial Neural Network (ANN)

- Inspired by the structure and function of biological neural networks.
- Commonly used for regression and classification problems, but are an enormous subfield.

- Most popular artificial neural network algorithms:
  - Perceptron
  - Back-Propagation
  - Hopfield Network
  - Radial Basis Functional Network

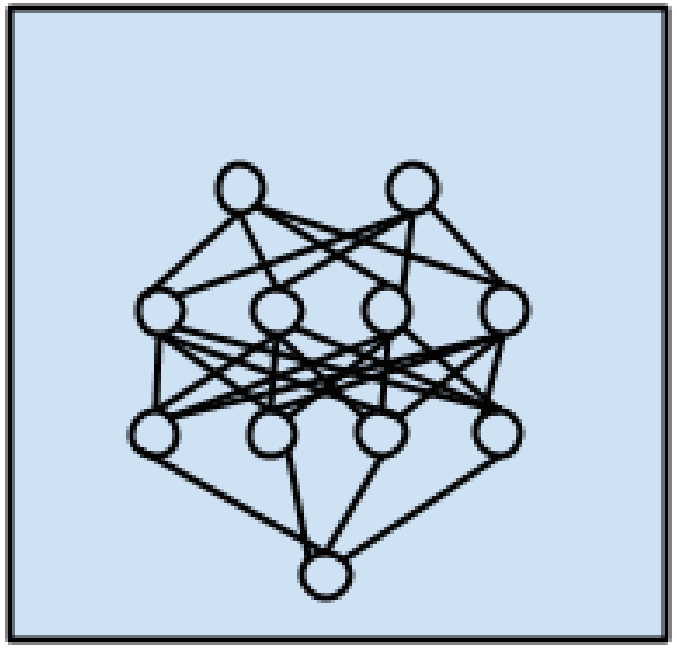


Artificial Neural Network Algorithms

Deep Learning

- Modern update to Artificial Neural Networks that exploit abundant cheap computation.
- Build larger and more complex neural networks; focused on semi-supervised learning with very little labeled data.

- Most popular deep learning algorithms:
  - Deep Boltzmann Machine (DBM)
  - Deep Belief Networks (DBN)
  - Convolutional Neural Network (CNN)
  - Stacked Auto-Encoders

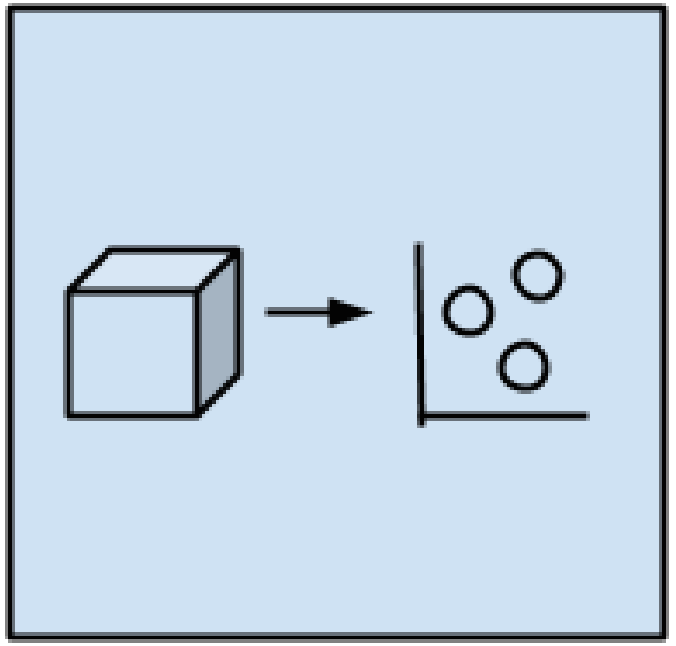


Deep Learning Algorithms

Dimensionality Reduction

- Similar to clustering
- Seek and exploit the inherent structures in the data, but are focused on summarizing or describing data using less information

- Most popular dimensionality reduction algorithms:
  - Principal Component Analysis (PCA)
  - Principal Component Regression (PCR)
  - Partial Least Squares Regression (PLSR)
  - Sammon Mapping
  - Multidimensional Scaling (MDS)
  - Projection Pursuit
  - Linear Discriminant Analysis (LDA)
  - Mixture Discriminant Analysis (MDA)
  - Quadratic Discriminant Analysis (QDA)
  - Flexible Discriminant Analysis (FDA)

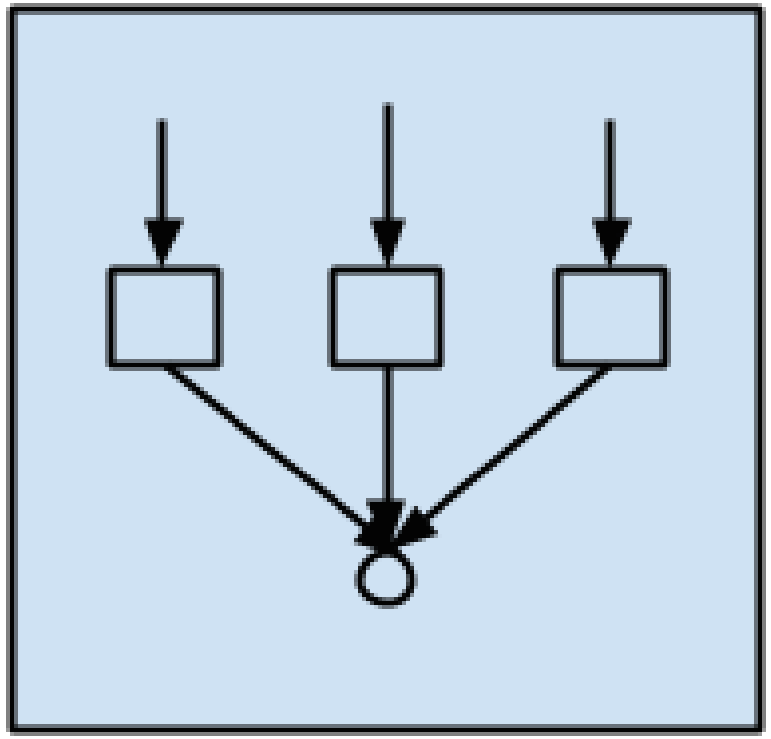


Dimensional Reduction Algorithms

Ensemble Methods

- Take multiple weak models that are independently trained and combine them to make an overall prediction.

- Most popular ensemble algorithms:
  - Boosting
  - Bootstrapped Aggregation (Bagging)
  - AdaBoost
  - Stacked Generalization (blending)
  - Gradient Boosting Machines (GBM)
  - Gradient Boosted Regression Trees (GBRT)
  - Random Forest



Ensemble Algorithms

How should I choose?

Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.

START

Predicting categories

Discovering structure

Finding unusual data points

Predicting values

Three or more

Two

ANOMALY DETECTION

- One-class SVM ← >100 features, aggressive boundary
- PCA-based anomaly detection ← Fast training

CLUSTERING

- K-means

MULTICLASS CLASSIFICATION

- Fast training, linear model → Multiclass logistic regression
- Accuracy, long training times → Multiclass neural network
- Accuracy, fast training → Multiclass decision forest
- Accuracy, small memory footprint → Multiclass decision jungle
- Depends on the two-class classifier, see notes below → One-v-all multiclass

REGRESSION

- Ordinal regression ← Data in rank ordered categories
- Poisson regression ← Predicting event counts
- Fast forest quantile regression ← Predicting a distribution
- Linear regression ← Fast training, linear model
- Bayesian linear regression ← Linear model, small data sets
- Neural network regression ← Accuracy, long training time
- Decision forest regression ← Accuracy, fast training
- Boosted decision tree regression ← Accuracy, fast training

TWO-CLASS CLASSIFICATION

- Two-class SVM ← >100 features, linear model
- Two-class averaged perceptron ← Fast training, linear model
- Two-class logistic regression ← Fast training, linear model
- Two-class Bayes point machine ← Fast training, linear model

- Accuracy, fast training → Two-class decision forest
- Accuracy, fast training → Two-class boosted decision tree
- Accuracy, small memory footprint → Two-class decision jungle
- >100 features → Two-class locally deep SVM
- Accuracy, long training times → Two-class neural network

Created by the Azure Machine Learning Team Email: AzurePoster@microsoft.com Download this poster: http://aka.ms/MLCheatSheet

Microsoft

Anomaly Detection

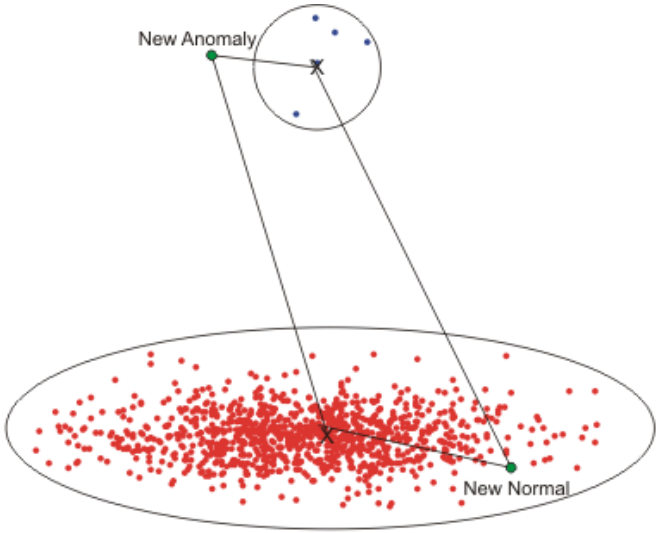
- Examples
  - Identifying transactions that are potentially fraudulent
  - Learning patterns that indicate a network intrusion has occurred
  - Finding abnormal clusters of patients
  - Checking values input to a system

- One-Class Support Vector Machine: creates a one-class support vector machine model for anomaly detection.
- PCA-Based Anomaly Detection: creates an anomaly detection model using Principal Component Analysis.

Anomaly Detection

One-Class Support Vector Machine

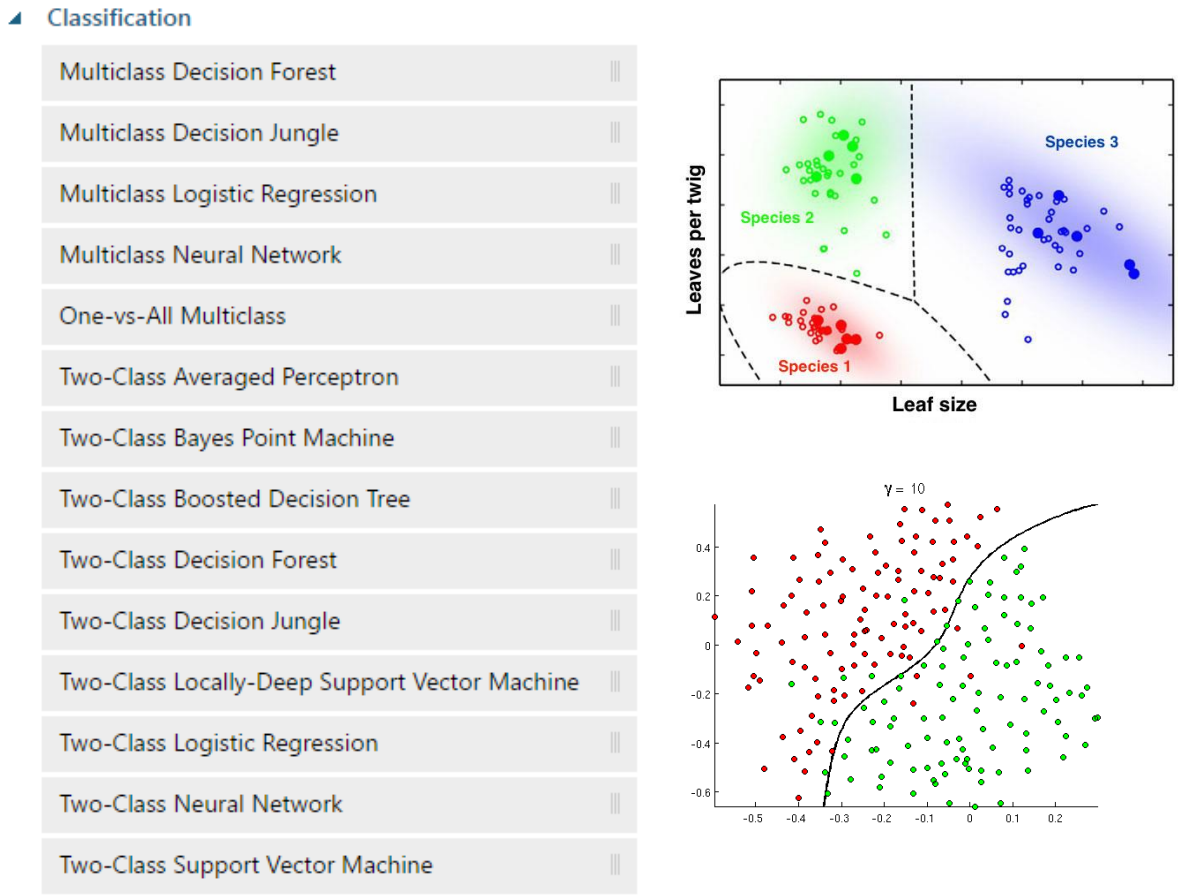
PCA-Based Anomaly Detection



Classification

- Predicts class or category for a single instance of data.
- **Example:** email filters use binary classification to determine if an email is spam.

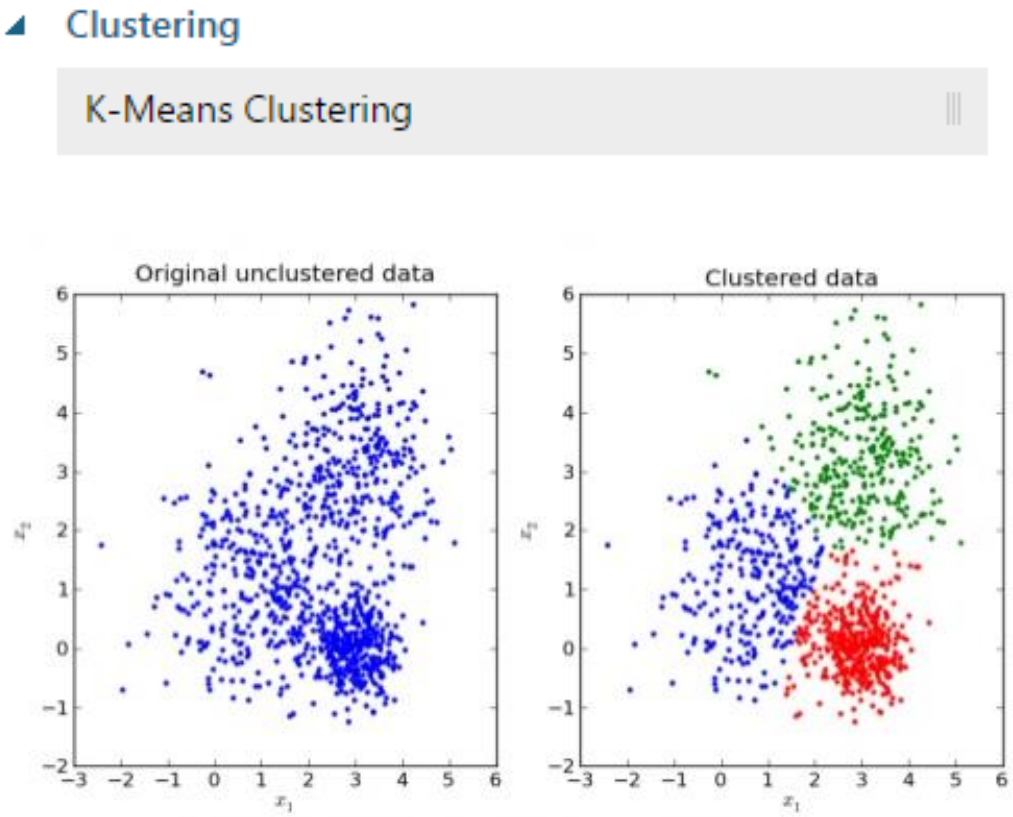
- Two forms of classification tasks:
  - **Binary Classification:** predict one of two outcomes
  - **Multiclass Classification:** predict one of many outcomes
- Output of a classification algorithm is called a **classifier**, which can be used to predict the label of a new (unlabeled) instance.



Clustering

- Algorithms that learn to group a set of items together based on a set of features.
- Often used in analysis to group pieces of text that contain common words together.
- Can be used to group unlabeled data by figuring out which data points are closest together, and then determining the centroid (or central point) of each grouping.
- Once the algorithm has been trained, it can be used to predict which cluster an instance of data belongs to.

- K-Means Clustering: configures and initializes a k-means clustering model.



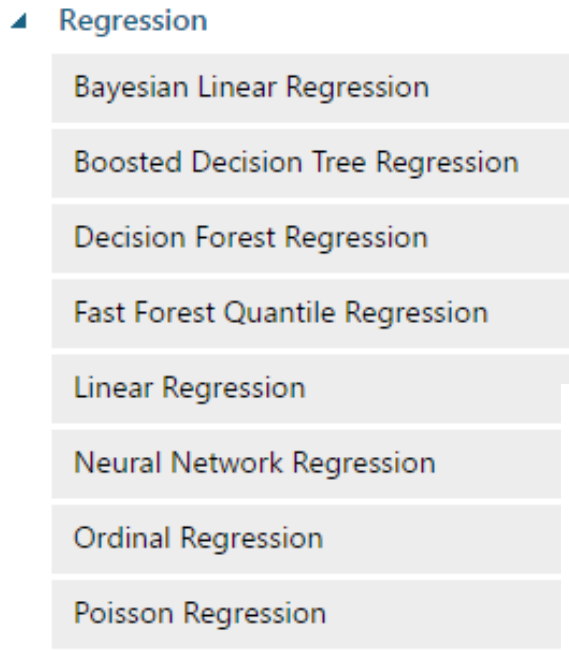
http://pypr.sourceforge.net/kmeans.html

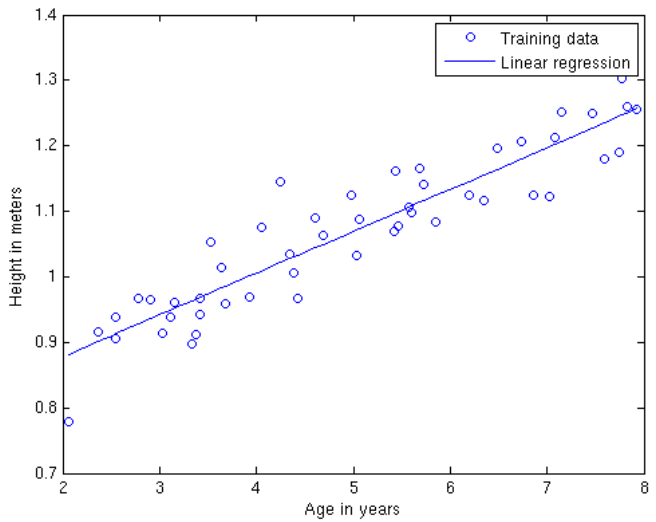
Regression

- Learns to predict the value of a real function for a single instance of data.
- Can incorporate input from multiple features, by determining the contribution of each feature of the data to the regression function.

Modules

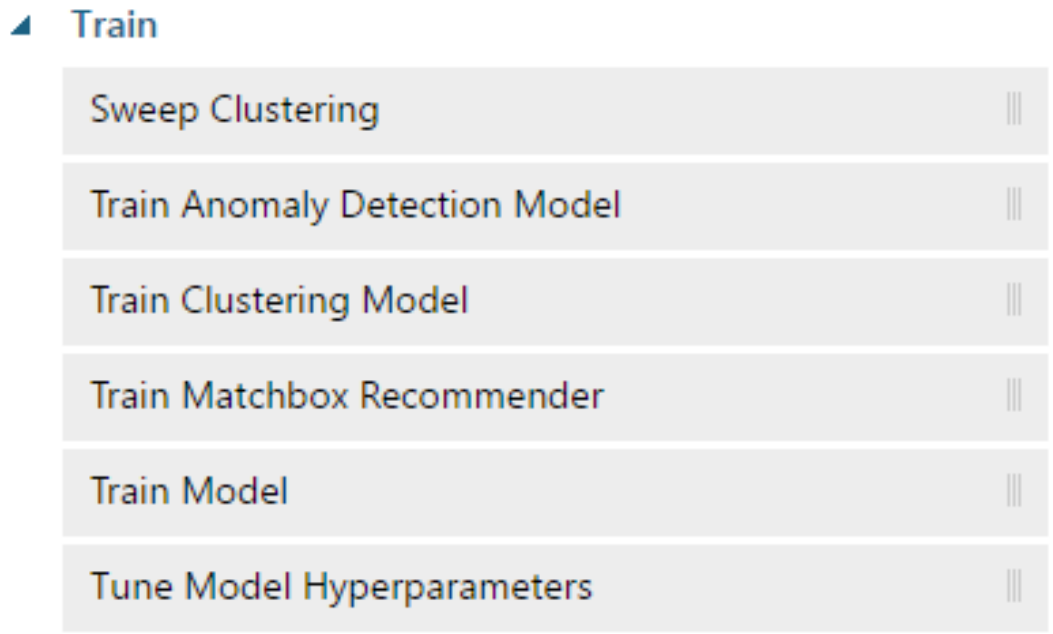
- Bayesian Linear Regression
- Boosted Decision Tree Regression
- Decision Forest Regression
- Fast Forest Quantile Regression
- Linear Regression
- Neural Network Regression
- Ordinal Regression
- Poisson Regression





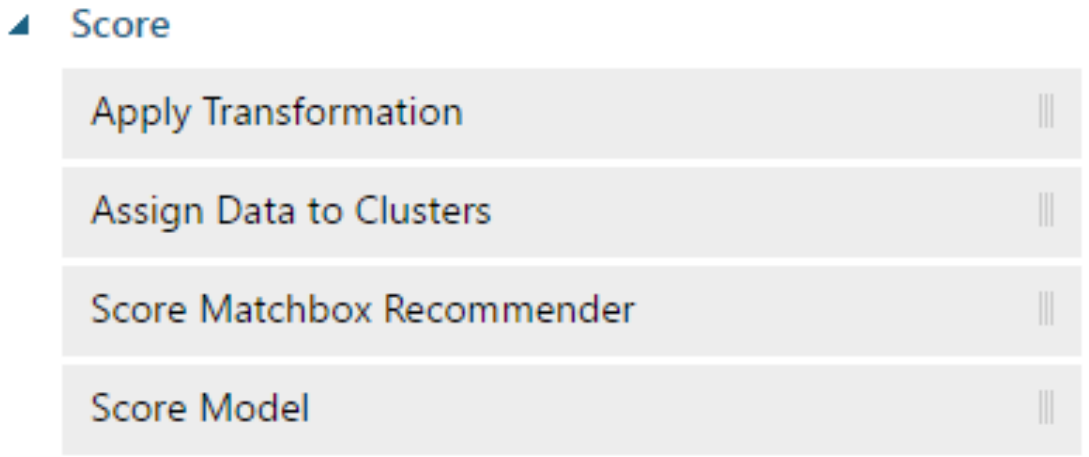
Machine Learning - Train

- Sweep Clustering: performs a parameter sweep on a clustering model to determine the optimum parameter settings and trains the best model.
- Train Anomaly Detection Model: trains an anomaly detector model and labels data from a training set.
- Train Clustering Model: trains a clustering model and assigns data from the training set to clusters.
- Train Matchbox Recommender: trains a Bayesian recommender using the Matchbox algorithm.
- Train Model: trains a classification or regression model from a training data set.
- Tune Model Hyperparameters: performs a parameter sweep on a regression or classification model to determine the optimum parameter settings and trains the best model.



Machine Learning - Score

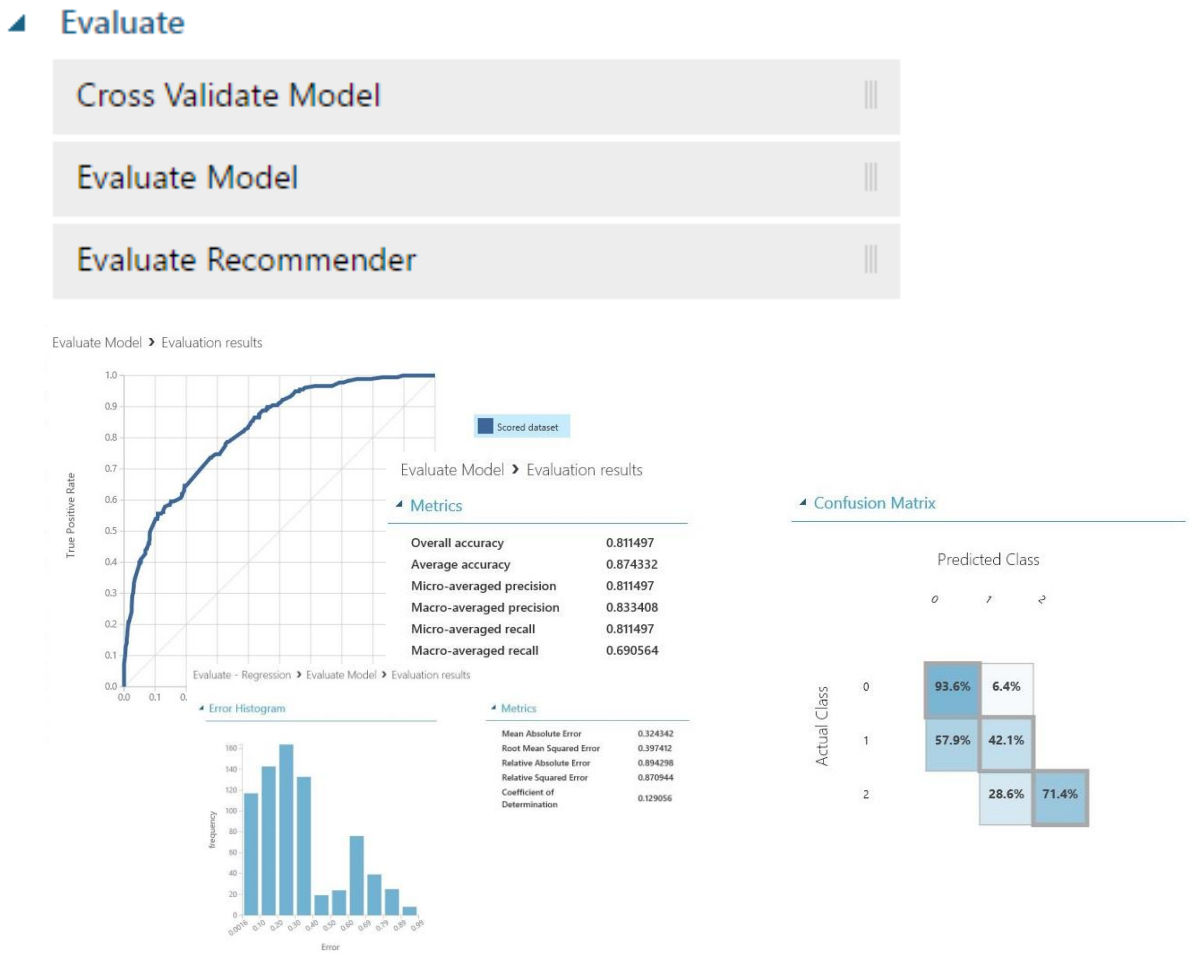
- Apply Transformation: applies a well-specified data transformation to a dataset.
- Assign Data to Clusters: Assigns data to clusters using an existing trained clustering model.
- Score Matchbox Recommender: scores predictions for a dataset using the Matchbox recommender.
- Score Model: scores predictions for a trained classification or regression model.



- Example Use Cases:
  - Developing a prediction solution
  - Survival analysis
  - Anomaly detection
  - Lexicon-based sentiment analysis

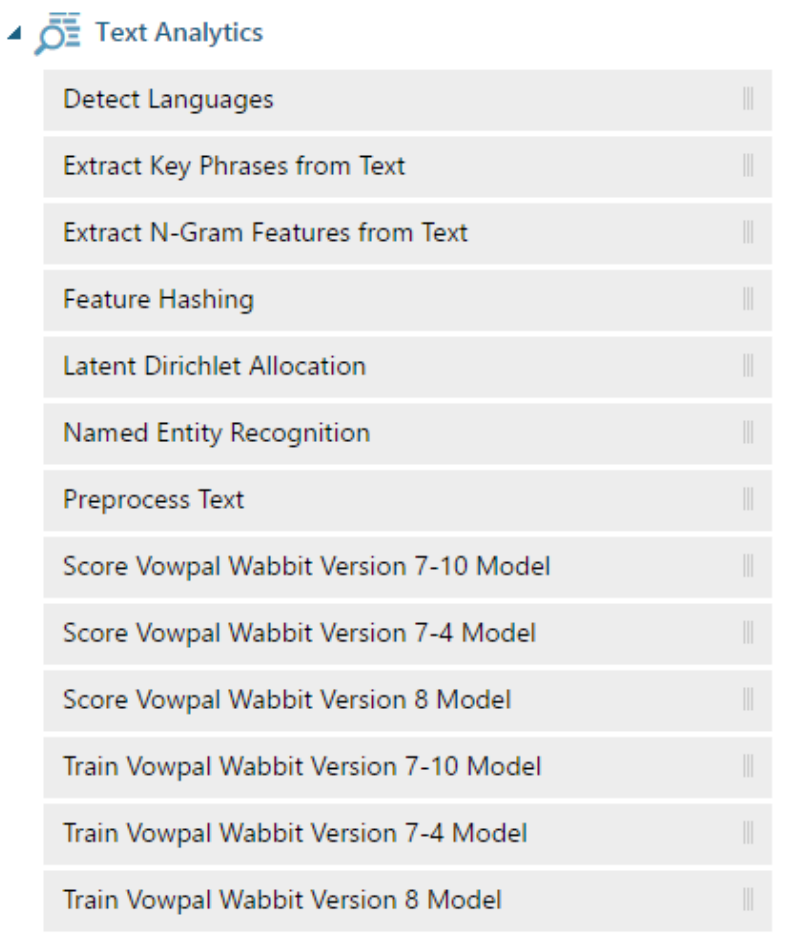
Machine Learning - Evaluate

- Evaluate Model: used if model is based on supported classification or regression algorithms.
- Evaluate Recommender: used for recommendation models.
- Assign Data to Clusters: used for clustering models; you can use the visualizations in the model to see evaluation results.
- Cross Validate Model: used to test the validity of the training set and the model.
  - Partitions the data into some number of folds (user specified) and then tests multiple models on the combinations of folds.



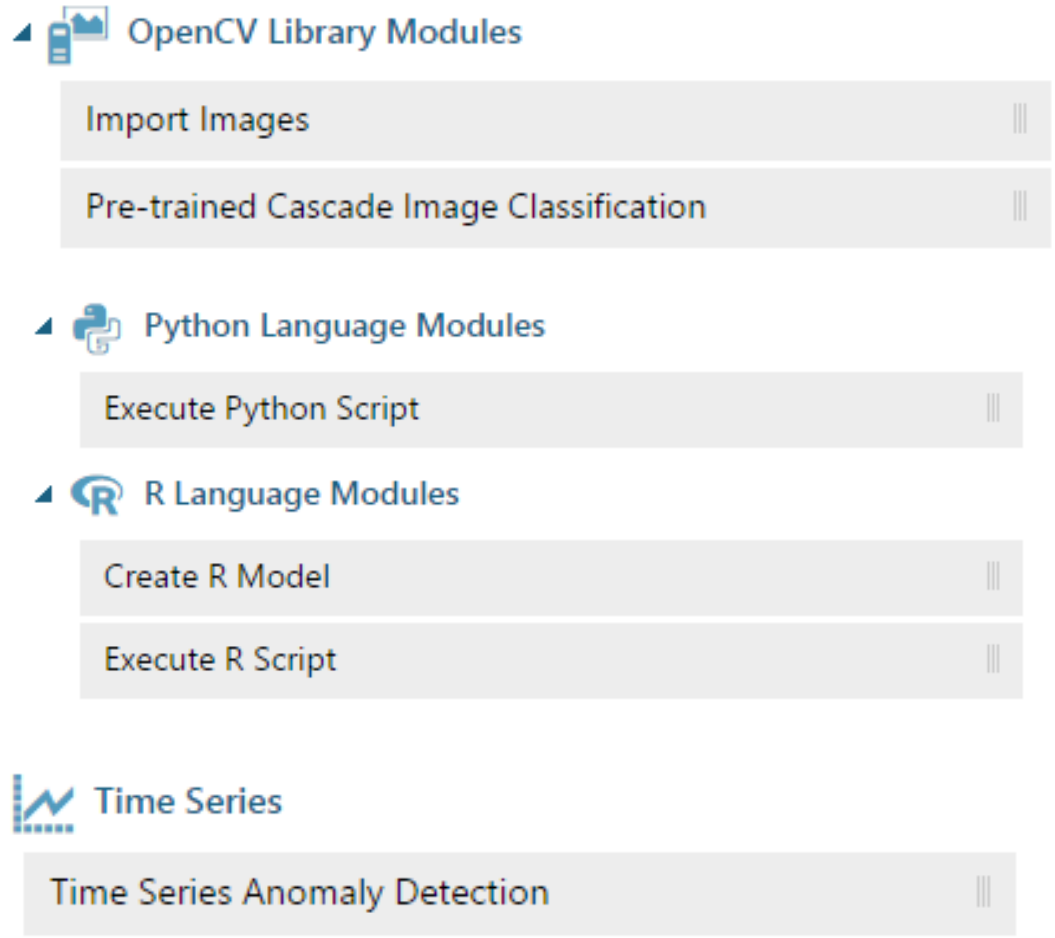
Modules – Text Analytics

- Used for working with both structured and unstructured text.
- Functionality Examples
  - **Detect the language** of input text
  - Create **N-Gram dictionary features** and do feature selection on them
  - **Feature hashing**: convert text data to integer-encoded features using the Vowpal Wabbit library
  - **Latent Dirichlet Allocation**: perform topic modeling
  - Recognize **named entities** in a text column
  - Extensive options for **text preprocessing**
  - ….and new options added monthly!
- Example Use Cases
  - Use feature hashing to classify articles into a predefined list of categories
  - Use the text of Wikipedia articles to categorize companies
  - Use text from Twitter messages to perform sentiment analysis



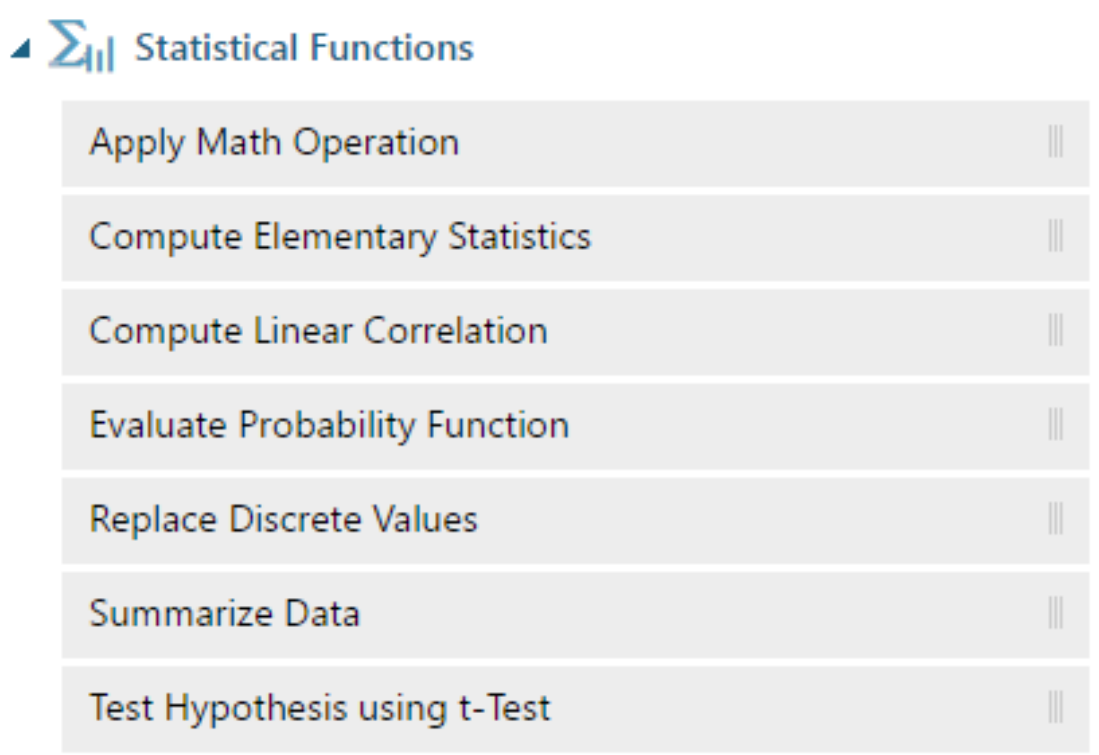
Modules – Other Languages, Tasks

- OpenCV Library Modules
  - Face API: detects faces and analyzes critical facial attributes, including emotion.
  - Computer Vision API: supports domain detection, identification of adult content, image tagging, and image type or color analysis
  - Bing Image Search: get images for machine learning projects by searching by type, color, region, and other attributes
- Python Language Modules
- R Language Modules
- Time Series Anomaly Detection



Modules – Statistical Functions

- Apply Math Operation: applies a mathematical operation to column values.
- Compute Elementary Statistics: calculates specified summary statistics for selected dataset columns.
- Compute Linear Correlation: calculates the linear correlation between column values in a dataset.
- Evaluate Probability Function: fits a specified probability distribution function to a dataset.
- Replace Discrete Values: replaces discrete values from one column with numeric values based on another column.
- Summarize Data: generates a basic descriptive statistics report for the columns in a dataset.
- Test Hypothesis Using t-Test: compares means from two datasets using a t-test.



Machine Learning in ML Studio

Anomaly Detection
- One-class Support Vector Machine
- Principal Component Analysis-based Anomaly Detection
- Time Series Anomaly Detection*

Classification

Two-class Classification
- Averaged Perceptron
- Bayes Point Machine
- Boosted Decision Tree
- Decision Forest
- Decision Jungle
- Logistic Regression
- Neural Network
- Support Vector Machine

Multi-class Classification
- Decision Forest
- Decision Jungle
- Logistic Regression
- Neural Network
- One-vs-all

Clustering
- K-means Clustering

Recommendation
- Matchbox Recommender

Regression
- Bayesian Linear Regression
- Boosted Decision Tree
- Decision Forest
- Fast Forest Quantile Regression
- Linear Regression
- Neural Network Regression
- Ordinal Regression
- Poisson Regression

Statistical Functions
- Descriptive Statistics
- Hypothesis Testing T-Test
- Linear Correlation
- Probability Function Evaluation

Text Analytics
- Feature Hashing
- Named Entity Recognition
- Vowpal Wabbit

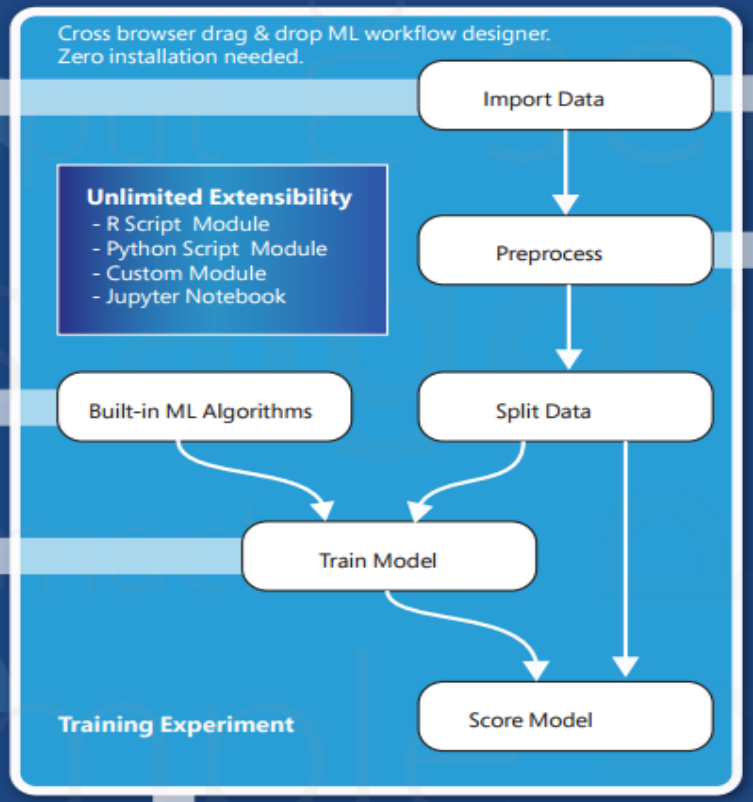
Computer Vision
- OpenCV Library

https://studio.azureml.net

Guest Access Workspace: Free trial access without logging in.
Free Workspace: Free persisted access, no Azure subscription needed.
Standard Workspace: Full access with SLA under an Azure subscription.

Data/Model Visualization
- Scatterplots
- Bar Charts
- Box plots
- Histogram
- R and Python Plotting Libraries
- REPL with Jupyter Notebook
- ROC, Precision/Recall, Lift
- Confusion Matrix
- Decision Tree*

Training
- Cross Validation
- Retraining
- Parameter Sweep



Unlimited Extensibility
- R Script Module
- Python Script Module
- Custom Module
- Jupyter Notebook

Data Source
- Azure Blob Storage
- Azure SQL DB
- Azure SQL DW*
- Azure Table
- Desktop Direct Upload
- Hadoop Hive Query
- Manual Data Entry
- OData Feed
- On-prem SQL Server*
- Web URL (HTTP)

Data Format
- ARFF
- CSV
- SVMLight
- TSV
- Excel
- ZIP

Data Preparation
- Clean Missing Data
- Clip Outliers
- Edit Metadata
- Feature Selection
- Filter
- Learning with Counts
- Normalize Data
- Partition and Sample
- Principal Component Analysis
- Quantize Data
- SQLite Transformation
- Synthetic Minority Oversampling Technique

Enterprise Grade Cloud Service
- SLA: 99.95% Guaranteed Up-time
- Azure AD Authentication
- Compute at Large Scale
- Multi-geo Availability
- Regulatory Compliance*

One-click Operationalization

Predictive Experiment

Make Prediction with Elastic APIs
- Request-Response Service (RRS)
- Batch Execution Service (BES)
- Retraining API

Community
- Gallery (http://gallery.azureml.net)
- Samples & Templates
- Workspace Sharing and Collaboration
- Live Chat & MSDN Forum Support

* Feature Coming Soon

Azure Machine Learning Studio Capabilities Overview

Created by the Azure Machine Learning Team Email: AzurePoster@microsoft.com
Download this poster: http://aka.ms/MLStudioOverview

Microsoft

Case Study

Predictive Analytics – Houston 311 Data

Houston 311 Data

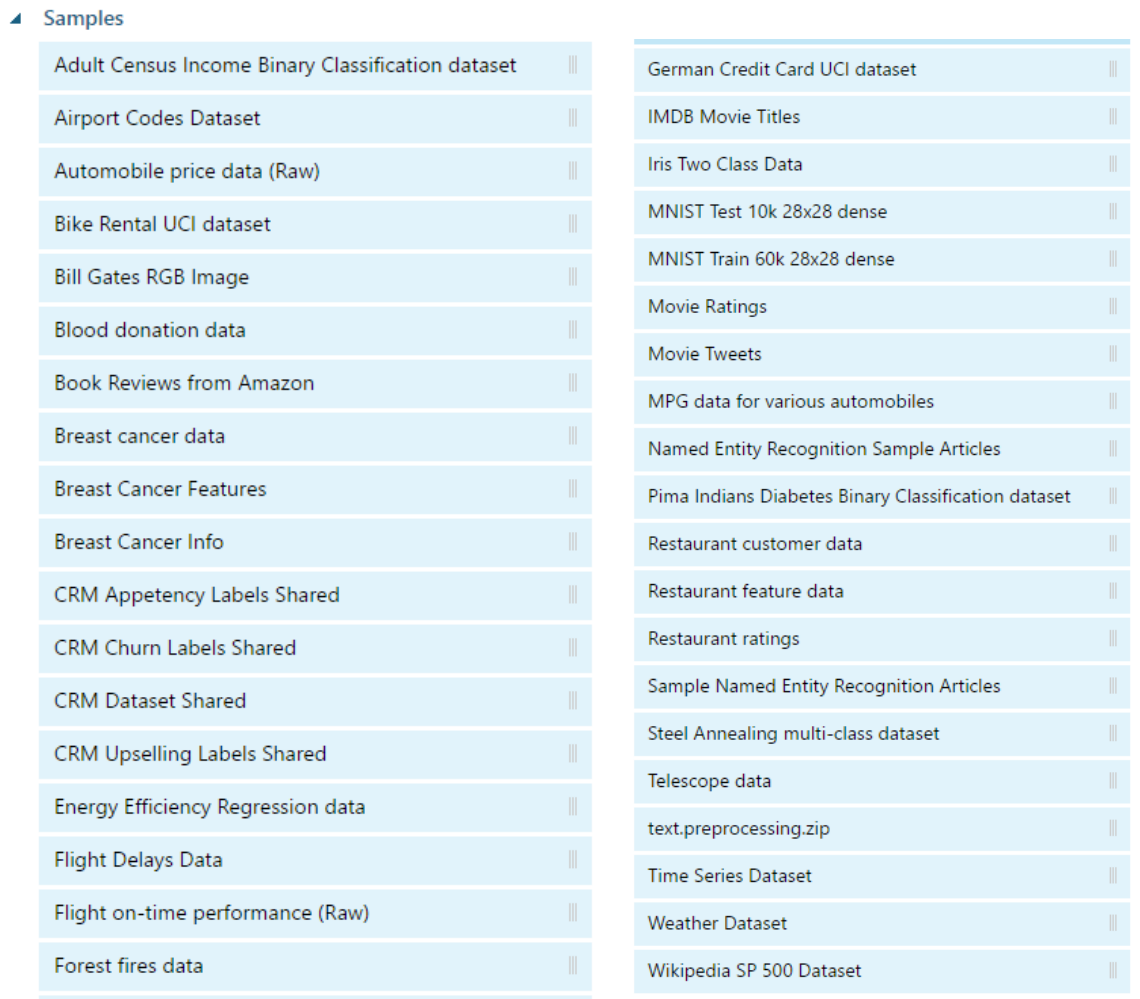
- **Houston 311** is the City of Houston's non-emergency service to report troubles and to get answers on city services.
- 311 files in the Houston data portal are updated nightly, and you can explore here: http://data.ohouston.org/dataset/city-of-houston-311-service-requests
- The data set we'll be using spans from November 2011 to present.

5-step process for ML

ACQUIRE // PREPARE // DEFINE FEATURES // SELECT & TEST ALGORITHM // PREDICT & SCORE

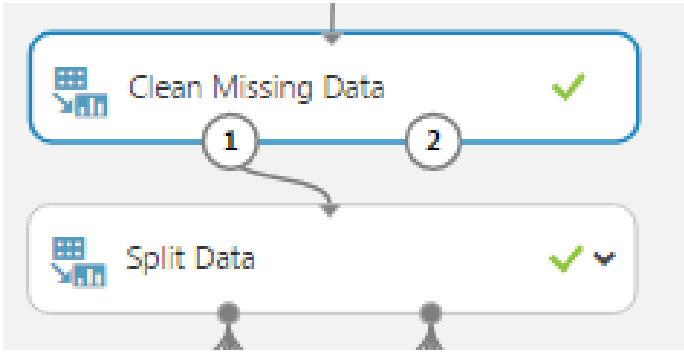
Step 1: Get the data

- Many sample data sets are available on the ML Studio website; you can also find others from NASA, Kaggle, UC Irvine, Stanford, open government data, etc.
- https://www.kaggle.com/datasets?gclid=CKb02c6mr9ACFYyXvQodwaQGvg
- https://archive.ics.uci.edu/ml/datasets.html
- https://data.nasa.gov/
- https://data.usgs.gov/datacatalog/
- http://data.ohouston.org/
- https://www.data.gov/open-gov/

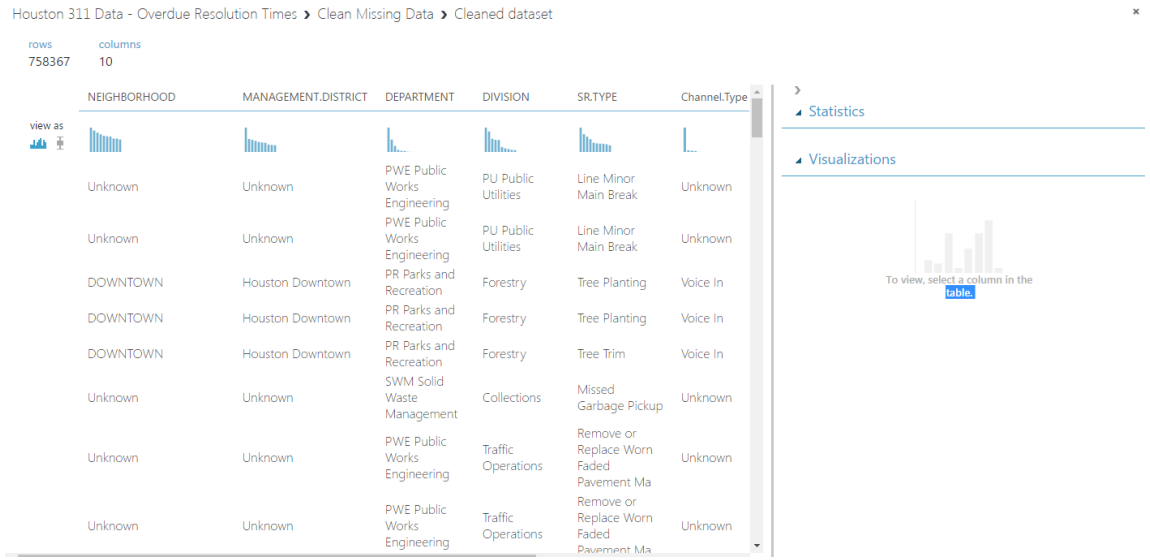


Step 2: Prepare the data

- Data sets usually required some pre-processing before they can be analyzed.
- Must also be split into **TEST, TRAIN**, and (optionally) **HOLDOUT** data sets.



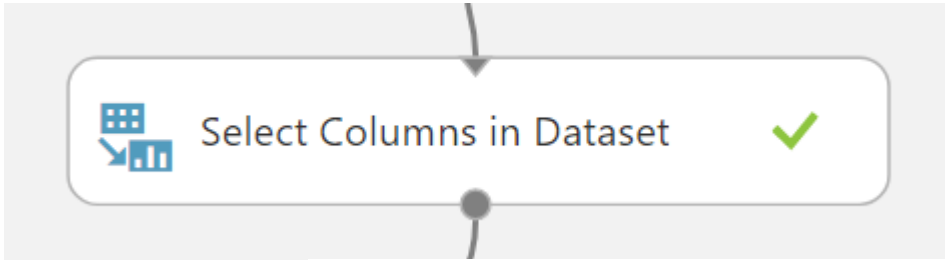
- Examples:
  - Missing values (NA, -999.25, blanks, etc.)
  - Nonsensical values (10000 instead of 1000, etc.)
  - Values that have been mistakenly imported as a string when they should have been considered numeric



Step 3: Define the features

- In machine learning, **features** are individual measurable properties of something you're interested in.

- Finding a good set of features for creating a predictive model requires **experimentation** and **knowledge** about the problem you want to solve.
  - Some features are better for predicting the target than others.
  - Some features have a strong correlation with other features, and can be removed.
  - The strong correlations can be revealed by showing linear relationships and cross validation.



Houston 311 Data - Overdue Resolution Times › Select Columns in Dataset › Results dataset

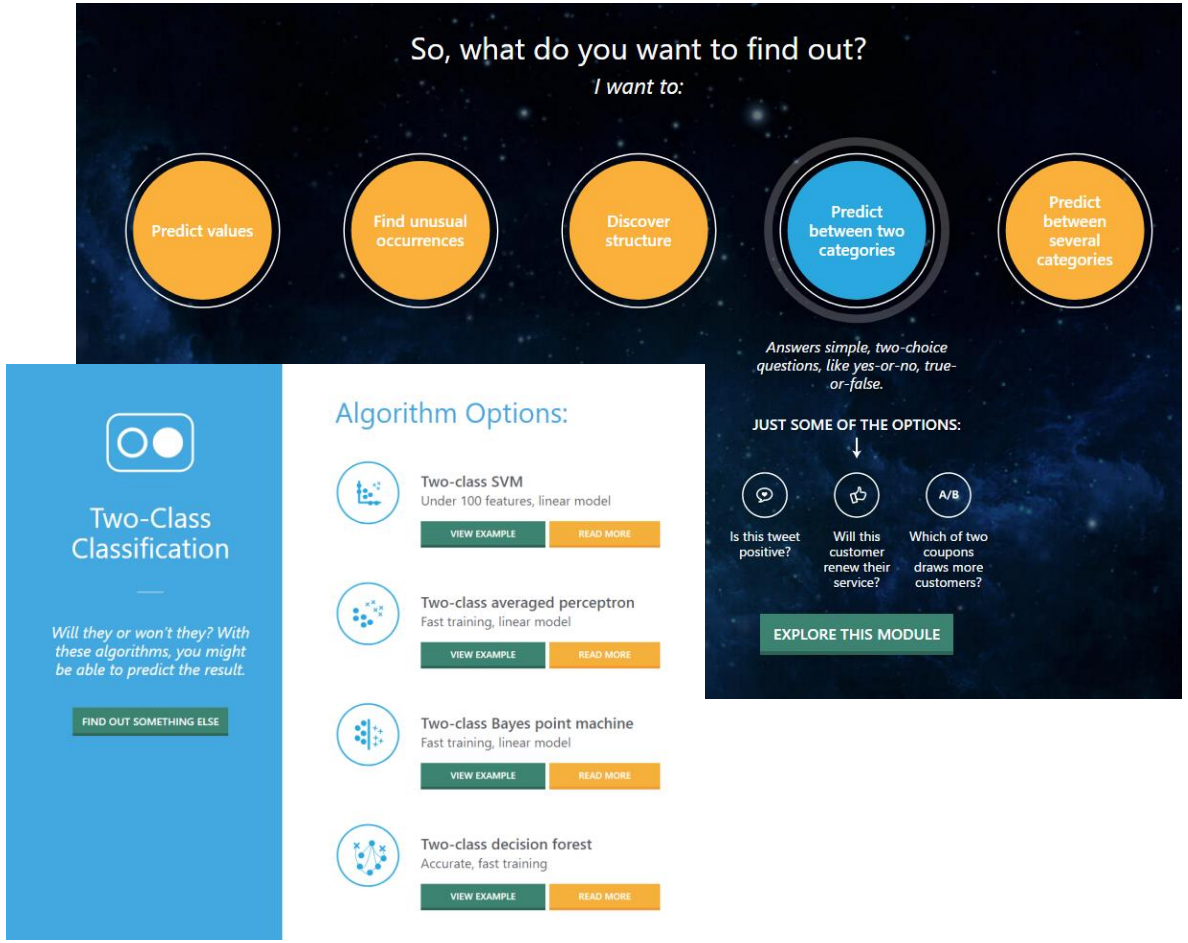
rows 1679336 columns 10

view as

NEIGHBORHOOD	MANAGEMENT.DISTRICT	DEPARTMENT	DIVISION	SR.TYPE	Channel.Type
Unknown	Unknown	PWE Public Works Engineering	PU Public Utilities	Line Minor Main Break	Unknown
Unknown	Unknown	PWE Public Works Engineering	PU Public Utilities	Line Minor Main Break	Unknown
DOWNTOWN	Houston Downtown	PR Parks and Recreation	Forestry	Tree Planting	Voice In
DOWNTOWN	Houston Downtown	PR Parks and Recreation	Forestry	Tree Planting	Voice In
		PR Parks and Recreation	Forestry	Tree Planting	Voice In
DOWNTOWN	Houston Downtown	PR Parks and Recreation	Forestry	Tree Trim	Voice In
Unknown	Unknown	SWM Solid Waste Management	Collections	Missed Garbage Pickup	Unknown
		PR Parks and Recreation	Forestry	Tree Planting	Voice In
Unknown	Unknown	PWE Public Works Engineering	Traffic Operations	Remove or Replace Worn Faded Pavement Ma	Unknown
		PWE Public		Remove or	

Step 4: Choose and apply an algorithm

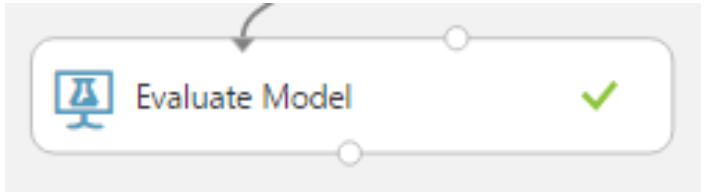
- Constructing a predictive model consists of training and testing. We'll use our data to train the model, and then we'll test the model to see how closely it's able to predict whether a given issue will be completed on time.
- Is this a **classification** task or a **regression** task?
- Help for choosing the correct algorithm: http://azuremlsimpleds.azurewebsites.net/simpleds/

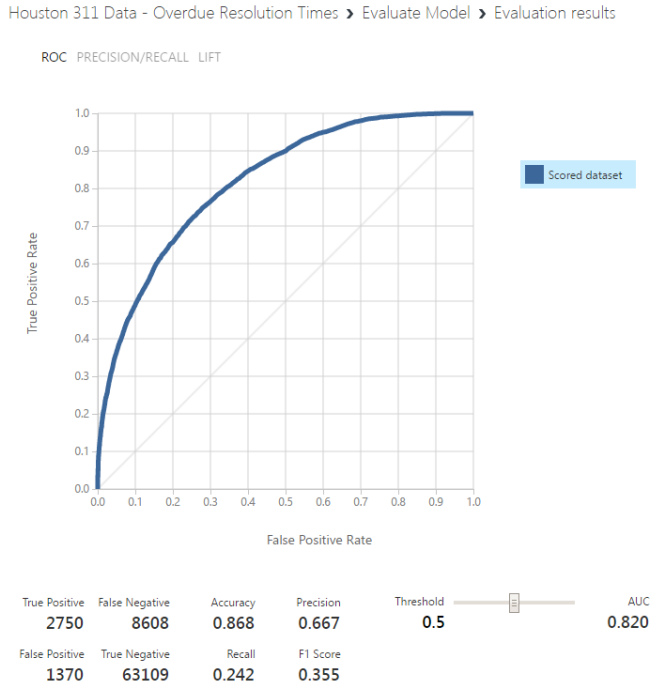


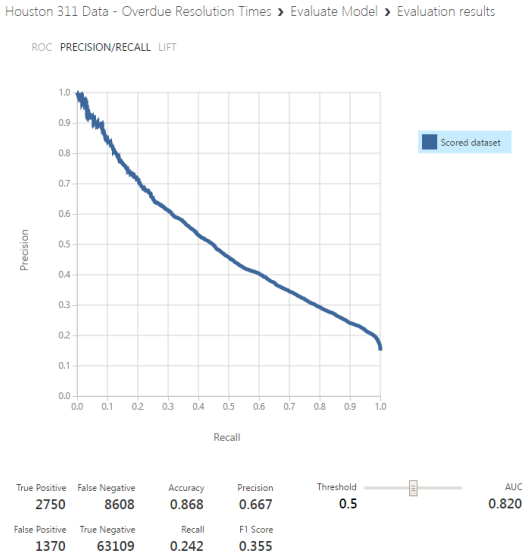
Step 5: Determine predictive power

- Now that the models have been trained using 90% of the data, we can use it to score the remaining 10% and see how well the models function.

- Five models used for sample project:
  - Boosted Decision Tree Regression
  - Two-Class Boosted Decision Tree
  - Two-Class Bayes Point Machine
  - Two-Class Decision Jungle
  - Two-Class Locally-Deep Support Vector Machine







Additional Resources

Paige.Bailey@alumni.rice.edu // @DynamicWebPaige

Additional Resources – Machine Learning

Online Training

- DataCamp ($49/month)
- Coursera (free, or small charge)
- EdX (free, or small charge)
- Microsoft's Azure Resources (free)
- ArXiV-Sanity (free)
- Kaggle (free)
- http://gallery.azureml.net/

Books

- Artificial Intelligence: A Modern Approach
  http://aima.cs.berkeley.edu/

Machine learning basics with algorithm examples
https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-basics-infographic-with-algorithm-examples

How to choose algorithms for Microsoft Azure Machine Learning
https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-algorithm-choice

List of Machine Learning Studio Modules
https://msdn.microsoft.com/en-us/library/azure/dn906033.aspx

Getting started with Azure ML Studio video
https://azure.microsoft.com/en-us/resources/videos/getting-started-with-ml-studio/

Process for getting started
https://azure.microsoft.com/en-us/trial/get-started-machine-learning/

Documentation
https://docs.microsoft.com/en-us/azure/machine-learning/
https://azure.microsoft.com/en-us/services/machine-learning/

Data Science for beginners
https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-for-beginners-the-5-questions-data-science-answers

Help files
https://msdn.microsoft.com/library/azure/dn905974.aspx

Try Out Azure Machine Learning Studio

REGISTER FOR AN ACCOUNT // EXPLORE AVAILABLE EXPERIMENTS // ASK QUESTIONS

Thank you!

Q&A