Instructor Materials Chapter 4: Advanced Data Analytics and Machine Learning

Big Data & Analytics

Cisco | Networking Academy®
Mind Wide Open™

Chapter 4: Advanced Data Analytics and Machine Learning

Big Data & Analytics

Cisco | Networking Academy®
Mind Wide Open™

Chapter 4 - Sections & Objectives

- 4.1 Predictive Analytics
  - Identify the likelihood of future outcomes through the use of data, statistical algorithms and machine learning techniques, based on historical data.
- 4.2 Model Evaluation
  - Examine the various evaluation metrics used in predictive analytics.
- 4.3 Preparation for Chapter 4 Labs

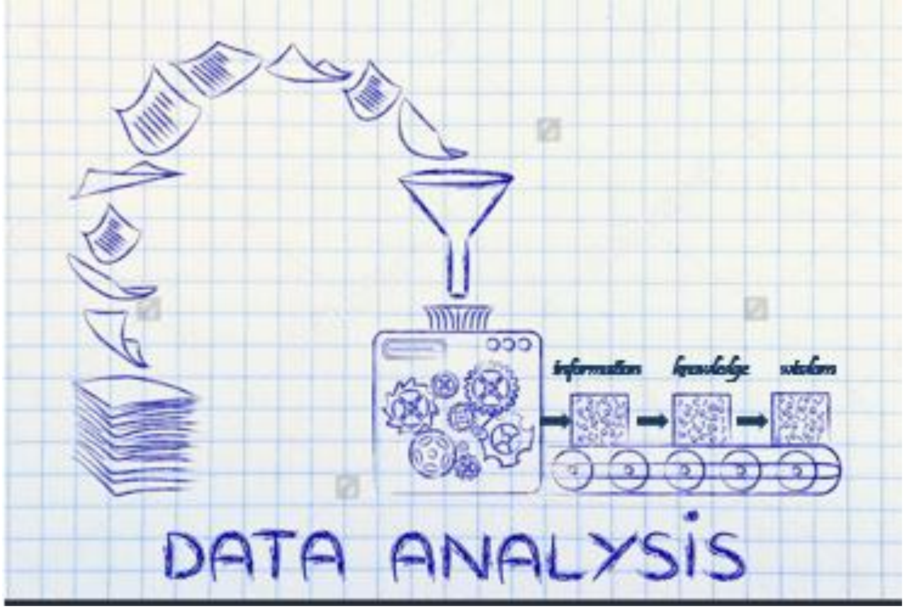
4.1 Predictive Analytics

Cisco | Networking Academy®
Mind Wide Open™

Machine Learning

Looking Ahead

- Characteristics that distinguish Big Data from data:
  - Volume
  - Velocity
  - Variety
  - Veracity
- Big Data is used to create predictive models that answer:
  - What will happen?
  - How should we act?



Machine Learning

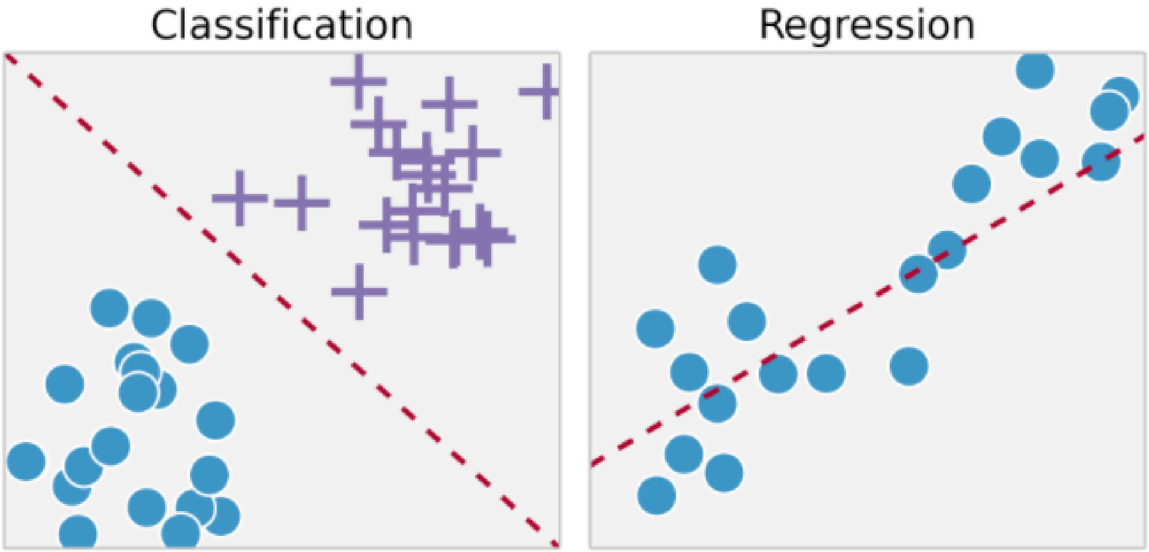
What is Machine Learning?

- Kevin Patrick Murphey defines machine learning as “…a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty.”
  - Machine learning algorithms improve their performance on specific tasks based on repeated performance of those tasks. Machine learning methods are applied to a wide range of applications including speech recognition, medical diagnostics, self-driving cars, sales recommendation engine, and many others.

Machine Learning

Types of Machine Learning Analysis

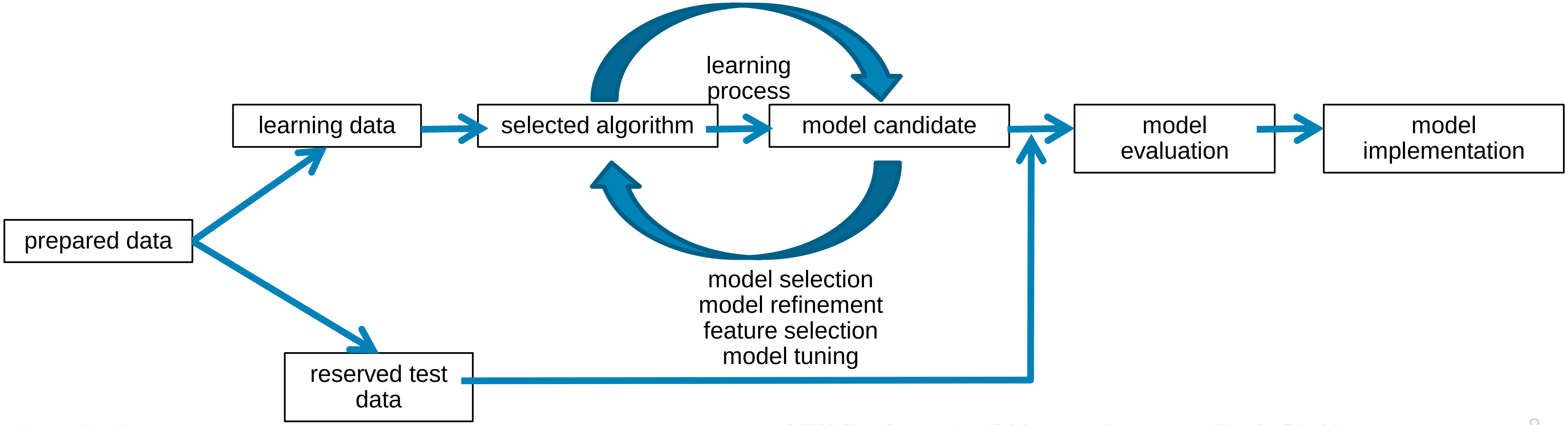
- Two main categories of machine learning algorithms:
  - Supervised – commonly used for predictive analytics. The are used to solve regression and classification problems.
  - Unsupervised – they autonomously discover patterns in data. Examples of problems solved with unsupervised methods are clustering and association.



Machine Learning

A Machine Learning Process

- Developing machine learning solutions can be simplified into the following steps:
  - Step 1 – Prepare the data
  - Step 2 – Create a learning set
  - Step 3 – Create a test set
  - Step 4 – Create a loop
  - Step 5 – test the solution
  - Step 6 – Implement the solution

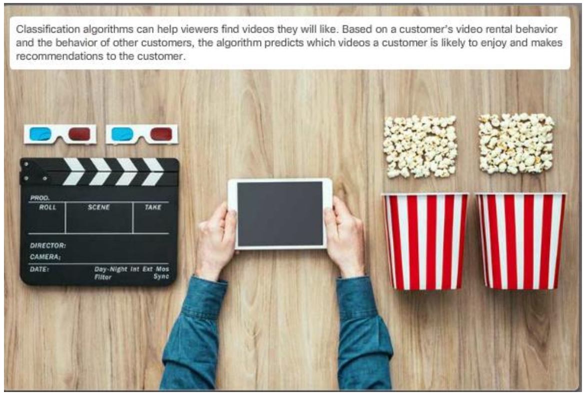


Machine Learning

Common Applications of Machine Learning

- Predictive analytics algorithms have a wide range of applications, including the use of analytics technology in the fields of entertainment, agriculture, medicine, and retail sales.

Large retail chain stores use IoT sensors to identify the location of shoppers within their stores. The predictive analytics system then sends targeted sales offers to the shopper's cell phone in real time.



Classification algorithms can help viewers find videos they will like. Based on a customer's video rental behavior and the behavior of other customers, the algorithm predicts which videos a customer is likely to enjoy and makes recommendations to the customer.

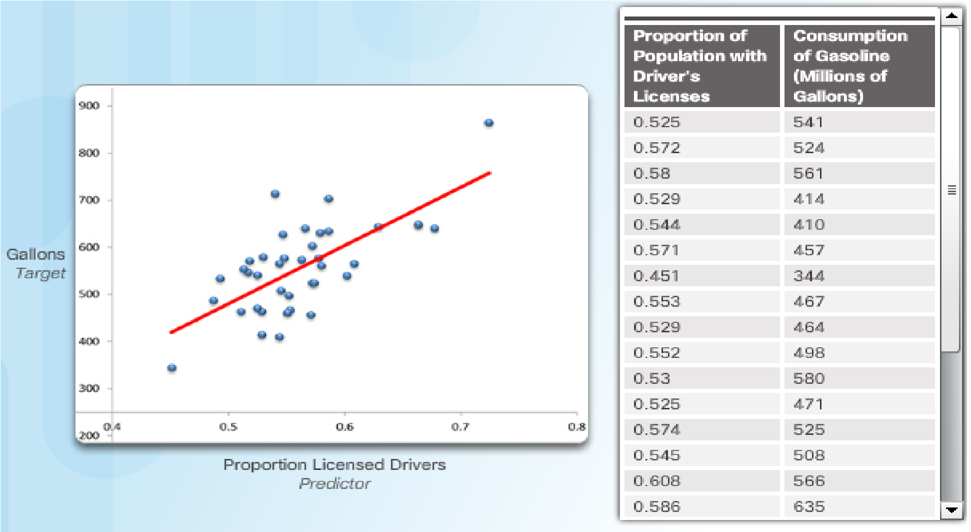
Farmers use cellphones to provide researchers with images of plant diseases. These images are used in image recognition systems to diagnose plant diseases. Combined with environmental data regression algorithms could then predict future outbreaks of disease.

A machine learning classification algorithm uses 20 input variables to predict the possibility of breast cancer. This approach can accurately identify patients who should carefully be watched for early detection of the disease.

Regression

Regression Analysis

- Regression Analysis is one of the oldest and most commonly used statistical methods for analyzing data.
- The main goal of regression is to qualify the mathematical relationship between one or more independent variables (predictor variable(s)), and a dependent one (target variable).

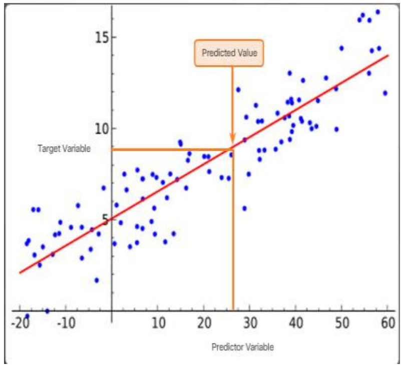


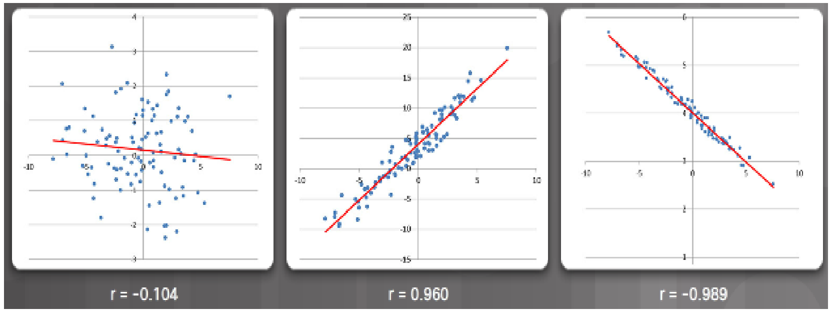
Proportion of Population with Driver's Licenses	Consumption of Gasoline (Millions of Gallons)
0.525	541
0.572	524
0.58	561
0.529	414
0.544	410
0.571	457
0.451	344
0.553	467
0.529	464
0.552	498
0.53	580
0.525	471
0.574	525
0.545	508
0.608	566
0.586	635

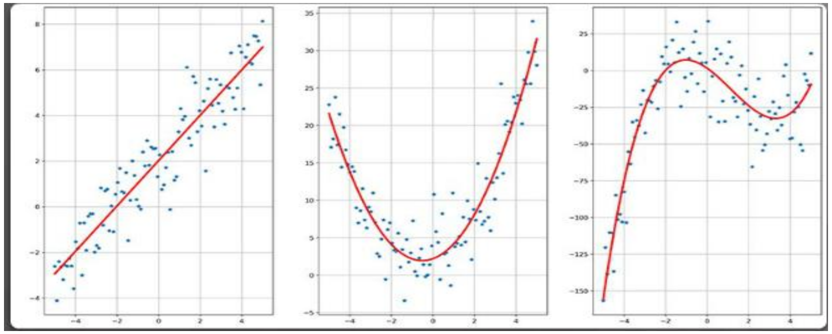
Regression

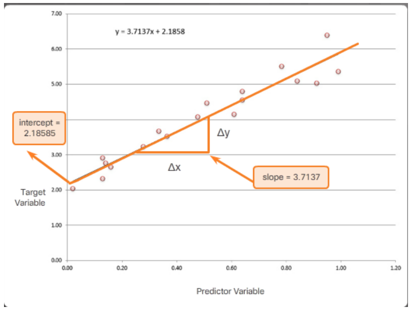
Linear Regression

- Linear regressions are the simplest from both a computational and mathematical point of view.
  - The term linear implies that the regression function will always try to fit the data using a weighted average of other functions, whether those functions are linear or not.









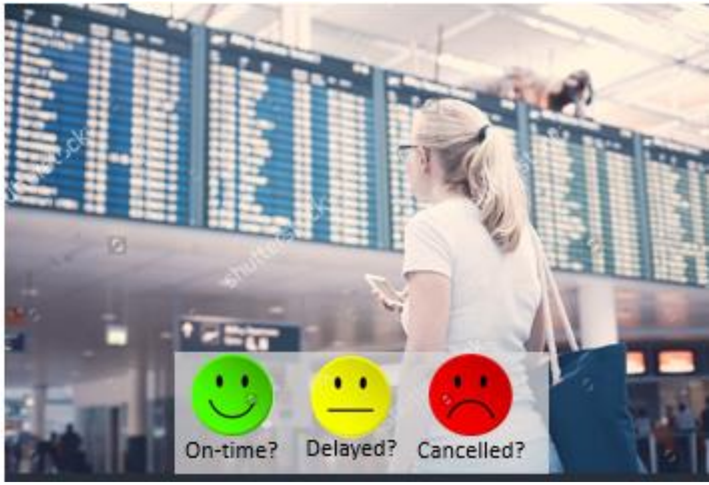
Applications of Regression Analysis

- Regression Analysis has many applications. It is frequently used in business and financial analysis with historical data to inform strategies for future action.

- It can be used to predict trends in economics and can inform political action to guide economic growth.
- Customer behavior can also be predicted to determine normal from possibly fraudulent behavior in fields of insurance and consumer credit.

Classification Problems

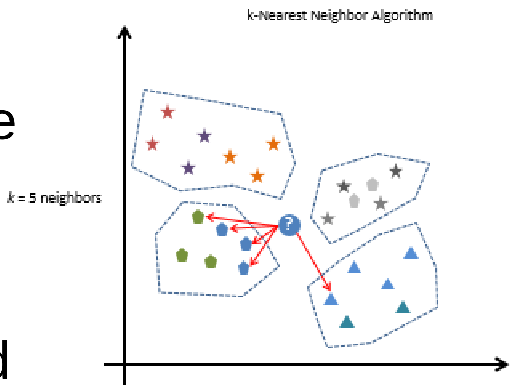
- Classification can be seen as a regression problem where the target variable is **discrete,** and represents a class in which a human expert has classified the data sample.
  - For example, a web-based travel company is interested in providing a reliability rating for the flights that it finds for customers. Via trial error of different models, it has been determined which variables among all the ones in the dataset are the most relevant for the classifications. This is also known as the variables with the highest discriminant power. Only these relevant features are extracted from the data and used to train the classifier.

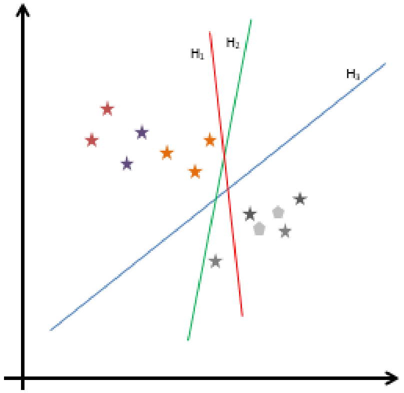


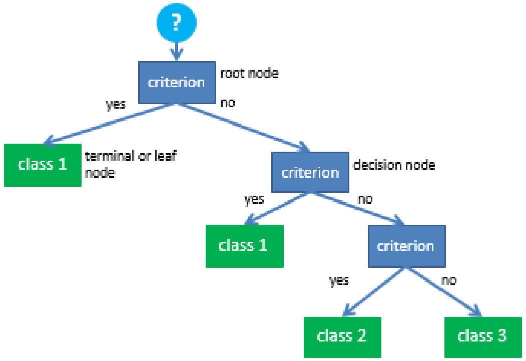
Statistical Analysis

Classification Algorithms

- **k-nearest neighbor (k-NN)** - k-NN is possibly the simplest classifier, which uses the distance between training examples as a measure of similarity. To visualize how a k-NN classifier works, imagine that each sample has two features, for which the values can be represented in a 2D plot.
- **Support vector machines (SVM)** - Support vector machines (SVM) are examples of supervised machine learning classifiers. Rather than basing the assignment of category membership on distances from other points, support vector machines compute the border, or hyperplane, that better separates groups.
- **Decision trees** - Decision trees represent a classification problem as a set of decisions based on the values of the features. Each node of the tree represents a threshold over the value of a feature, and splits the training samples in two smaller sets.







Statistical Analysis

Applications of Classifications

- Classification algorithms have many applications. For example:
  - **Risk Assessment** - Classification systems can be used to determine which of many factors contribute to the likelihood of various risks.
  - **Medical Diagnostics** - Classification systems can use guided questions to build a decision tree that can help diagnose various diseases and risks of disease.
  - **Image Recognition** - In handwriting recognition, a system may be working at the task of identifying handwritten numerals.

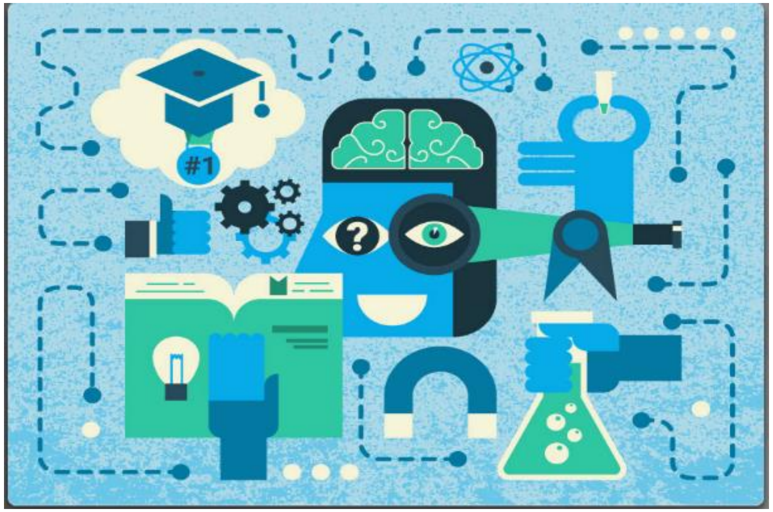
4.2 Model Evaluation

Cisco | Networking Academy®
Mind Wide Open™

Validity and Reliability

Issues in Using analysis

- The six step process for scientific discovery are:
  - Ask a question about an observation
  - Perform research
  - Form a hypothesis
  - Test the hypothesis
  - Analyze the data from the experiments to draw a conclusion
  - Communicate the results



Validity and Reliability

Validity

- While there are many terms used to describe types of validity, researchers typically distinguish between four types of validity:
  - **Construct validity** - Does the study actually measure what it claims to measure?
  - **Internal validity** - Was the experiment designed correctly? Does it include all the steps of the scientific method?
  - **External validity** - Can the conclusions apply to other situations or other people in other places at other times? Are there any other causal relationships in the study that might account for the results?
  - **Conclusion validity** - Based on the relationships in the data, are the conclusions of the study reasonable?

Validity and Reliability

Reliability

- A Reliable experiment or study means that someone else can repeat it and achieve the same results. Researchers distinguish between four types or reliability:
  - **Inter-rater reliability** - How similarly do different people score on the same test?
  - **Test-Retest Reliability** - How much variation is there between scores for the same person taking a test multiple times?
  - **Parallel-Forms Reliability** - How similar are the results of two different tests that are constructed from the same content?
  - **Internal Consistency Reliability** - What is the variation of results for different items in the same test?

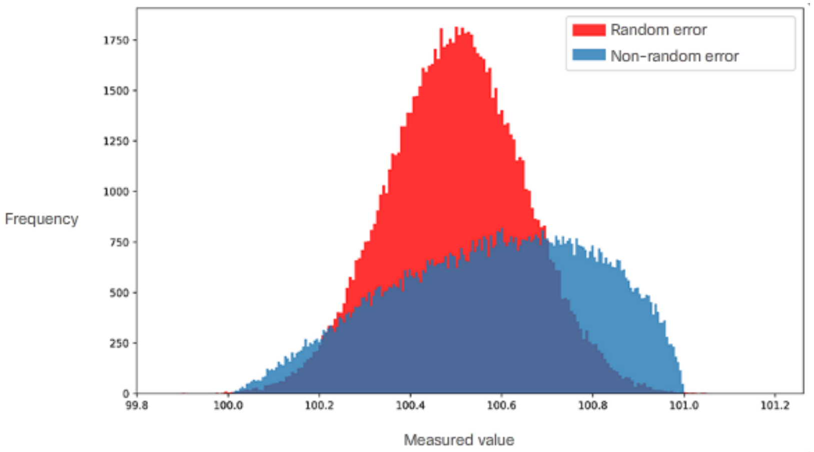
Error in Data Analytics

- Errors, and more in general, uncertainty, affect the data analytics process at different levels:
  - The first type of error is the **measurement error**. Any device for taking measurements is limited in its precision. Therefore, all measurements have a built-in error component.
  - Another type of error is the **prediction error**. In supervised learning, the prediction error is quantified as the difference between the value predicted by the model and the observed value.

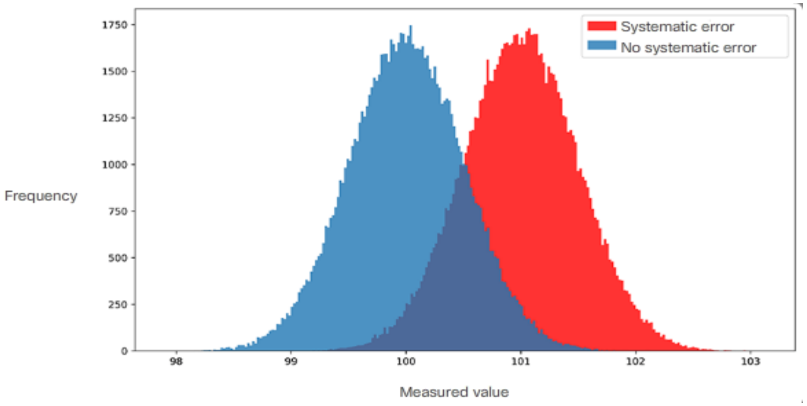
Error in Analysis

Types and Sources of Measurement Error

- Measurement errors can be categorized into these three groups:
  - **Gross errors** - These are caused by a mistake in the instrument being used to take the measurement, or in recording the result of the measurement.
  - **Random errors** – These are caused by factors that randomly impact the measurement over a sample of data.
  - **Systematic errors** – These are caused by instrumental or environmental factors that impact all measurements taken over a given period of time.



Random errors



Systematic errors

Error in Analysis

Random Error Distribution

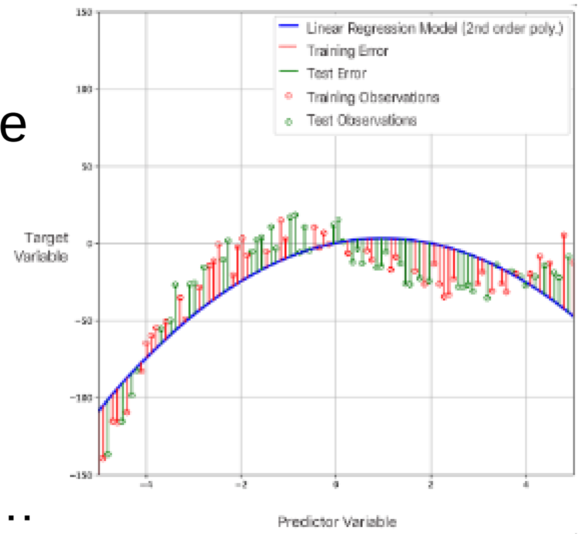
- **Random errors** tend to create a normal distribution around the mean of the observation. It is possible to build a statistical model of the error, in which case regression and classification algorithms can easily take it into account.

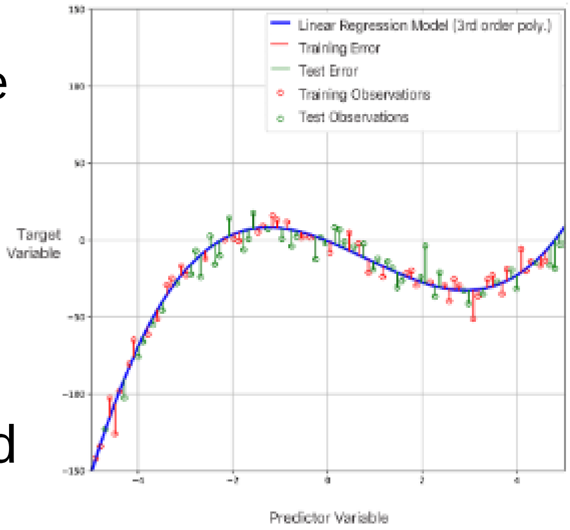
- **Systematic errors** tend to shift the distribution of the observations (right side of the figure) in one direction or another. A systematic error is therefore harder to deal with, because the true value is not known, so the only way to detect a systematic error is to use another measurement system that we deem more reliable.

Error in Analysis

Errors in Predictive Analytics

- **Prediction error is a difference** between the value predicted by the regression or classification model, and the measured value.
- **Prediction error is the distance** between the regression function, and the data points. The prediction error has **two components**
  - The first component is caused by the choice of model… we make an assumption on how the data is distributed, which is inevitably an approximation.
  - Even when the chosen model perfectly reflects the true distribution, there will still be differences between predicted and actual values because of the measurement error.
- In machine learning, the first cause of prediction error is often called **bias** of a model, while the second is **variance**. One cannot minimize both, and this situation is often called the **bias-variance tradeoff.**

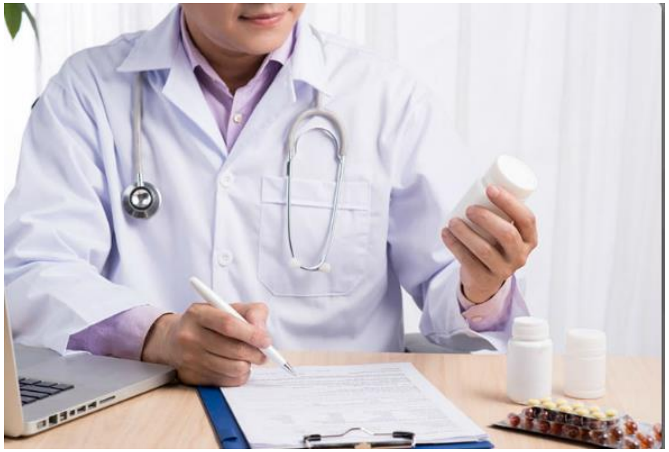




Model Evaluation

Misleading Research

- Understanding the impact of validity, reliability, and errors in a pattern of data is an important first step to ensuring that your conclusions are based on a solid research design.
- Misleading, bad, or erroneous research is more common than you may think. In fact, John P.A. Ioannidis states that most research findings are false.



Model Evaluation

Guidelines for Evaluating Results

- There are several guidelines you can following when evaluating the results reported by a research study or a data analysis report:
  - **Statistics** - Does the study have a large enough sample size to support the findings?
  - **Research design** - Did the architects of the study follow generally accepted methods of research design?
  - **Duration** - Does the research appropriately account for the impact on time?
  - **Correlation and causation** - Just because two variables are correlated does not mean that one caused the other.
  - **Alignment to other studies** - Do the results confirm or align with other studies in the field?
  - **Peer review** - Has the study been reviewed by experts in the same field?

4.3 Preparation for Chapter 4 Labs

Cisco | Networking Academy®
Mind Wide Open™

Preparation for Chapter 4 Labs

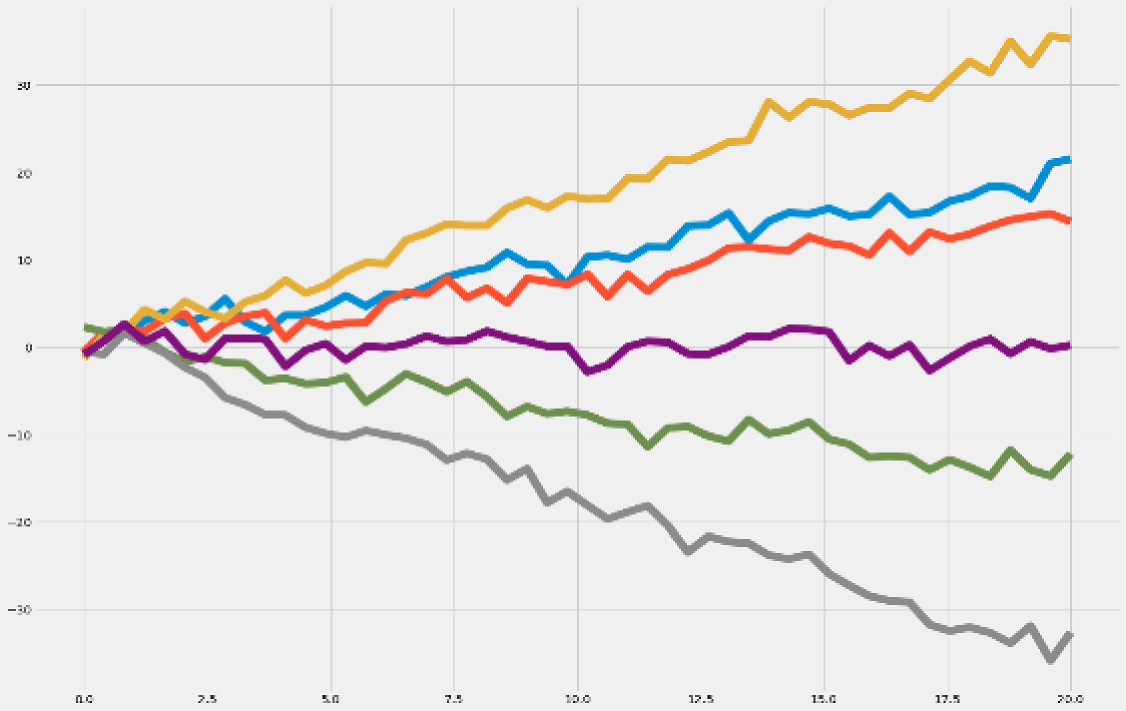
Using scikit-learn for Regression Analysis

- **scikit-learn is a machine learning library for Python** built on NumPy, SciPy, and matplotlib
- In the first lab, you will use regression analysis to view historical data about the growth of Internet traffic. You will quantify the relationship between the year and the measurement of Internet traffic.

Preparation for Chapter 4 Labs

Style Sheets for Plots

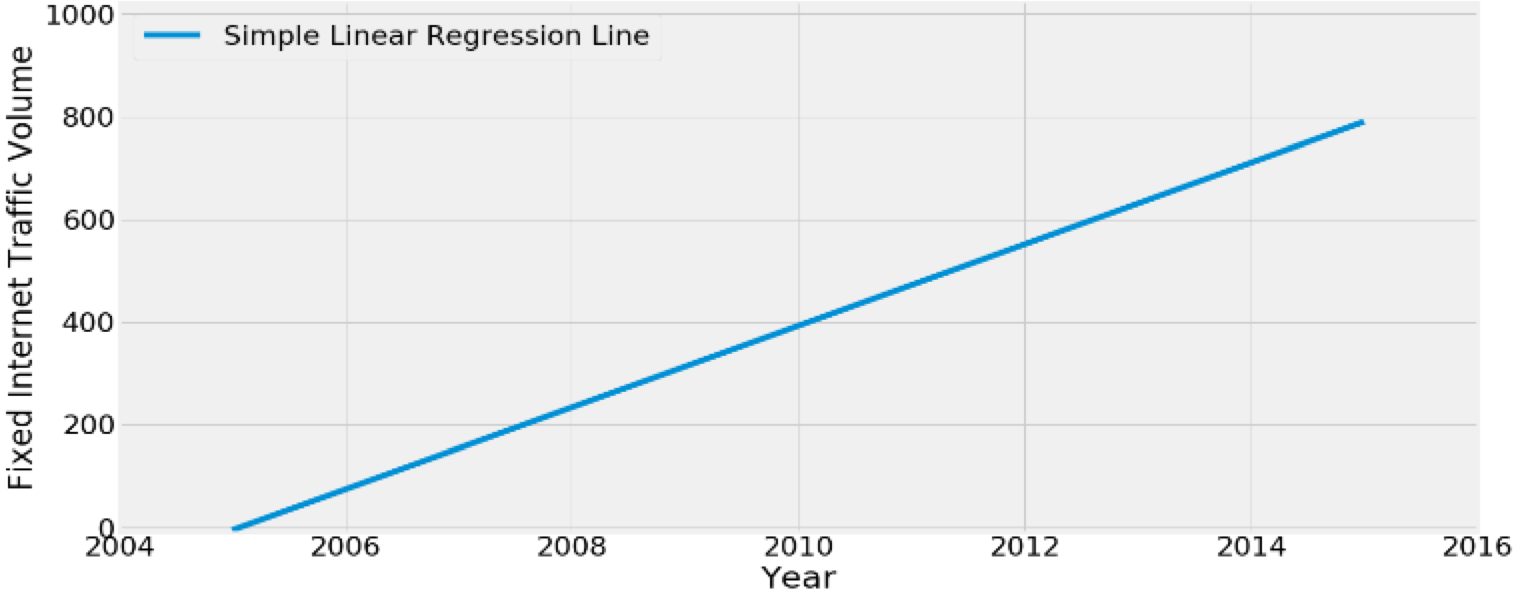
- You will install pandas, numpy, and matplotlib. The matplotlib library includes different styles for showing your plots.



Preparation for Chapter 4 Labs

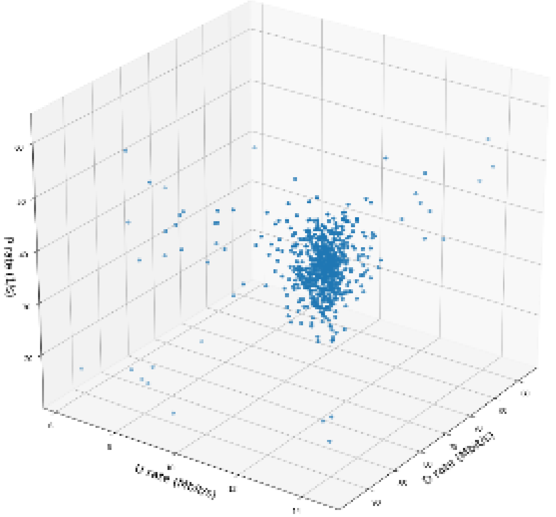
Fitting the Data

- To do linear regression in Python, you will call on the Numpy class, polyfit. Although polyfit has many arguments, you will only define the values for x, y, and deg. The value for x and y will be used for the x and y axis. Using polyfit will allow you to plot the simple linear regression shown in the figure. The value for deg will define the degree of fit..



Plotting in 3D

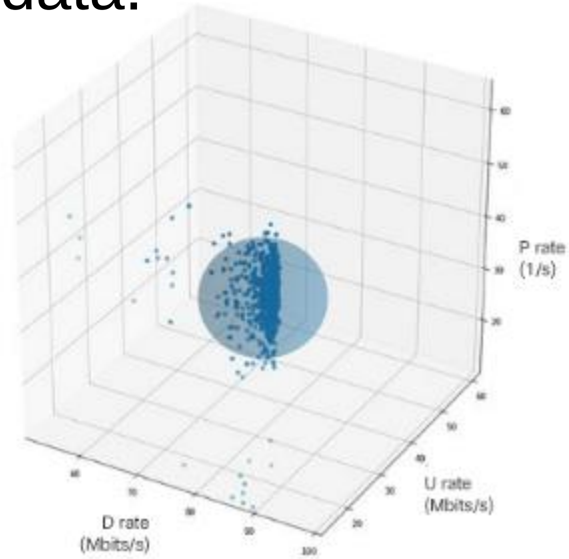
- You will visualize data in three dimensions. To do so, you will extend the matplotlib library by installing the mpl_toolkits class from the mplot3d library. You will then use the Internet meter data to create a 3D plot to display three axis: download rate (x axis); upload rate (y axis); and ping rate (z axis). This visualization will display where the rates for most of the pings cluster



Preparation for Chapter 4 Labs

Visualizing the Boundary for Anomalies

- Data Anomalies can be caused by corruptions or distortions during measurement, transmission, or storage. These values are considered outliers. They deviate so far from expected values that they could distort the results of the analysis.
- Anomalies are frequently removed from the data set after careful consideration.
- The sphere shows the decision boundary between normal data and anomalous data.



4.4 Summary

Cisco | Networking Academy®
Mind Wide Open™

Chapter Summary

Summary

- Big Data is characterized by volume, velocity, variety and veracity.
- Examples of supervised machine learning approaches, ie: Regression and Classification.
  - Regression uses historical relationship between one or more independent variables and a dependent variable to predict future values of dependent variables.
  - Classification models are knows as classifiers. There are numerous classifier algorithms. Example: k-nearest neighbor, Support vector machine and Decision tree.
- The chapter discusses the six step process used by the scientific method for validating the evaluation model.
- The four types of validity are: construct, internal, external, and conclusion.
- The four types of reliability are: inter-rater, test-retest, parallel-forms, and internal consistency.
- Error is the difference between the actual value and the measured value of an observation.

Cisco | Networking Academy®

Mind Wide Open™