Practical Data Science
&
Machine Learning
Overview

By
Miguel Hidalgo

Training Objectives

- Increase understanding of essentials Data Science concepts.
- Understand how to use basic tools for Data Science.
- Understand how to use basic Machine Learning algorithms.
- Understand the DS/ML model building process.
- Understand that teamwork is essential for Data Science.
- Understand how to apply easy DS techniques that will increase the success of your projects.
- Understand that what was presented is the tip of the iceberg with respect to Data Science/ML, but is a good beginning.

What is Analytics?

Analytics is the discovery, interpretation, and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance. (Wikipedia)

BI is NOT Analytics, instead is the foundation from where Analytics can take off.

BI answers the following questions:

- How did we do? & What has happened?

Analytics answers the following questions

- What should we do? & What can happen?

What is Analytics?

Four data analysis methods-

1. Forecasting techniques (Time series)
2. Descriptive analytics (clustering, association rules, etc.)
3. Predictive analytics (classification, regression, and text mining)
4. Decision optimization techniques

What is a Data Science?

- Data science, also known as data-driven science, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.

- Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, uncertainty quantification, computational science, data mining, databases, and visualization. (Wikipedia)

What is Machine Learning?

- Machine learning and statistics are part of data science. The word learning in machine learning means that the algorithms depend on some data, used as a training set, to fine-tune some model or algorithm parameters. This encompasses many techniques such as regression, naive Bayes or clustering. (Wikipedia)

Why are the Organization Objectives?

- Understand the Organization Objectives
  - Return On Investment (ROI)
  - Better understanding of the data
  - Accurate predictions (Forecasts, Equipment Failure, Credit Security, etc.)
  - Data Driven Decisions
  - Better Services for the customers
  - Better products
  - Etc.
- Improve the Organization
  - {Measuring Processes + Analyzing the Processes} yields Counter Measures=Improvement
- What Methods can we use to implement Data Science?
  - Project Management
  - Lean Six Sigma
  - Quality Management
  - Etc.

If you can’t measure it, you can’t control it

Machine Learning Problems

- Classification problems
  - Binary (Predicts a Yes or No, Survived or Died, two possible outcomes)
  - Multiclass (Predicts/classifies data into different classes)
- Regression problems
  - Predicts a number
- Clustering problems
  - Group data into similar clusters
- Association problems
  - Associate groups of data based on specific rules/behaviors
- Automation problems
  - Self Driving cars, Drones, etc.
- Anomaly Detection problems
  - Detecting something out of the norm

Machine Learning Categories

- Supervised Learning
  - Requires a Data Scientist to prepare the data with a target variable
- Unsupervised Learning
  - The ML algorithm discover by itself patterns hidden in the data. Cluster Algorithm like k-means, Hierarchical-Clustering, etc.
- Reinforcement Learning
  - Playing a game with rewards and payoffs. Self driving vehicles, drones, etc.
- Deep Learning
  - Deep learning architectures such as deep neural networks, deep belief networks and recurrent neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics and drug design, where they have produced results comparable to and in some cases superior to human experts.[Wikipedia]

Machine Learning Methods

CLASSIFICATION:

- Lazy Learning – Classification Using Nearest Neighbors K-NN Algorithm
- Probabilistic Learning – Classification Using Naïve Bayes
- Divide and Conquer – Classification Using Decision Trees and Rules – C5.0 Decision Tree Algorithm and 1R, RIPPER Rules algorithm and combination of Trees/Rules
- Classifying data with logistic regression

Forecasting Numeric Data:

- Regression Methods – Simple Linear Regression, Penalized Regression (Lasso, Ridge, etc.)
- Regression Trees and Model Trees

Machine Learning Methods/

Black Box Methods:

- Artificial Neural Networks(ANN) - Modeling Concrete strength
- Support Vector Machines (SVM) – Classification with Hyperplanes, Performing OCR with SVMs

Finding Patterns:

- Market Basket Analysis Using Association Rules- Identifying frequently purchased groceries with association rules
- Finding Groups of Data – Clustering with K-Means – Finding Teen market segments using k-means clustering

Time Series Modeling: (Not an ML model)

- Exponential, ARIMA, TBATS, etc.

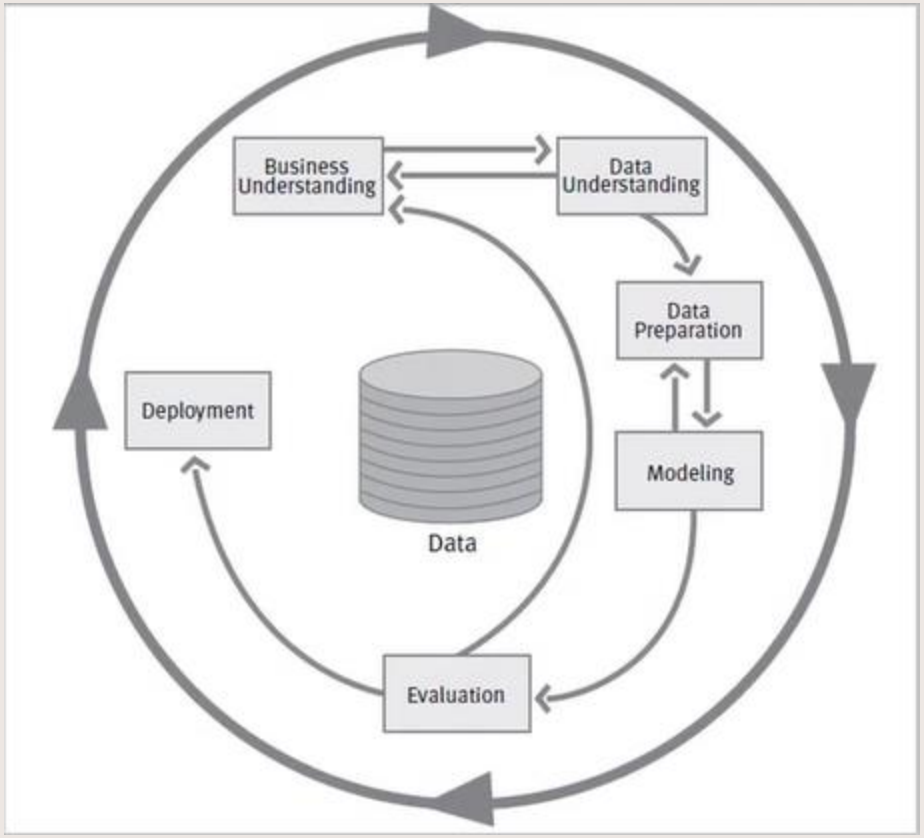
This are just some samples, there are over 300 ML algorithms and growing!

What is CRISP-DM?

CRoss-Industry Standard Process for Data Mining

==========

Structured approach to planning data mining projects. A high level plan or framework.



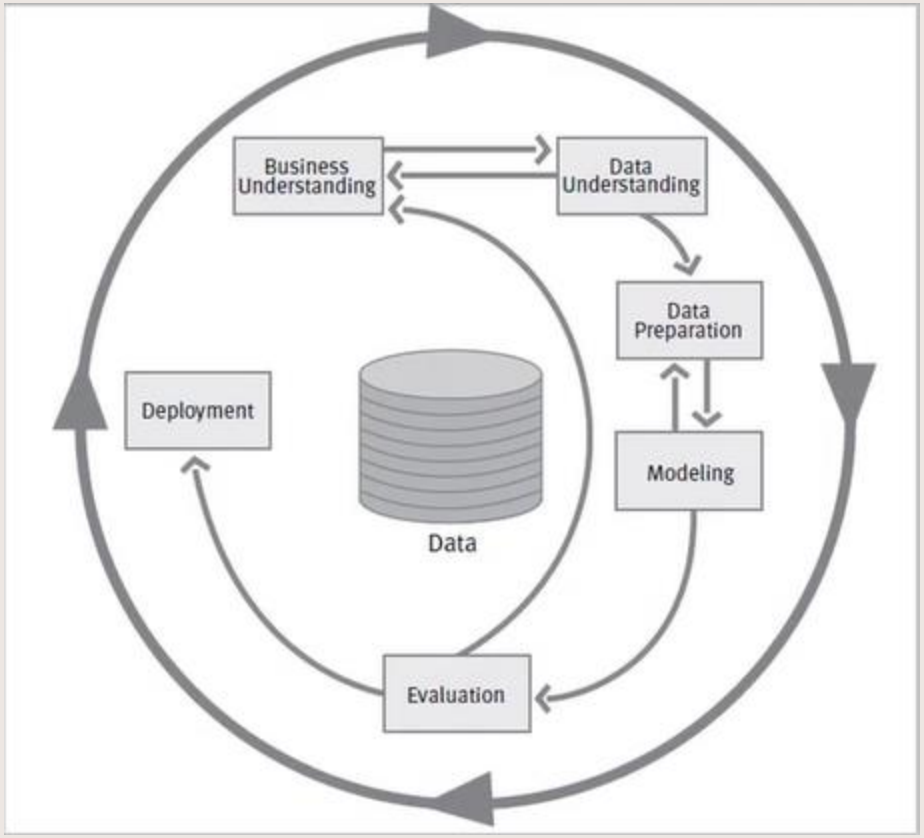
Those who don't plan, plan to fail!

by MHidalgo

CRISP-DM Process

Process Cycle:

I. Business Understanding
II. Data Understanding
III. Data Preparation
IV. Modeling
V. Evaluation
VI. Deployment

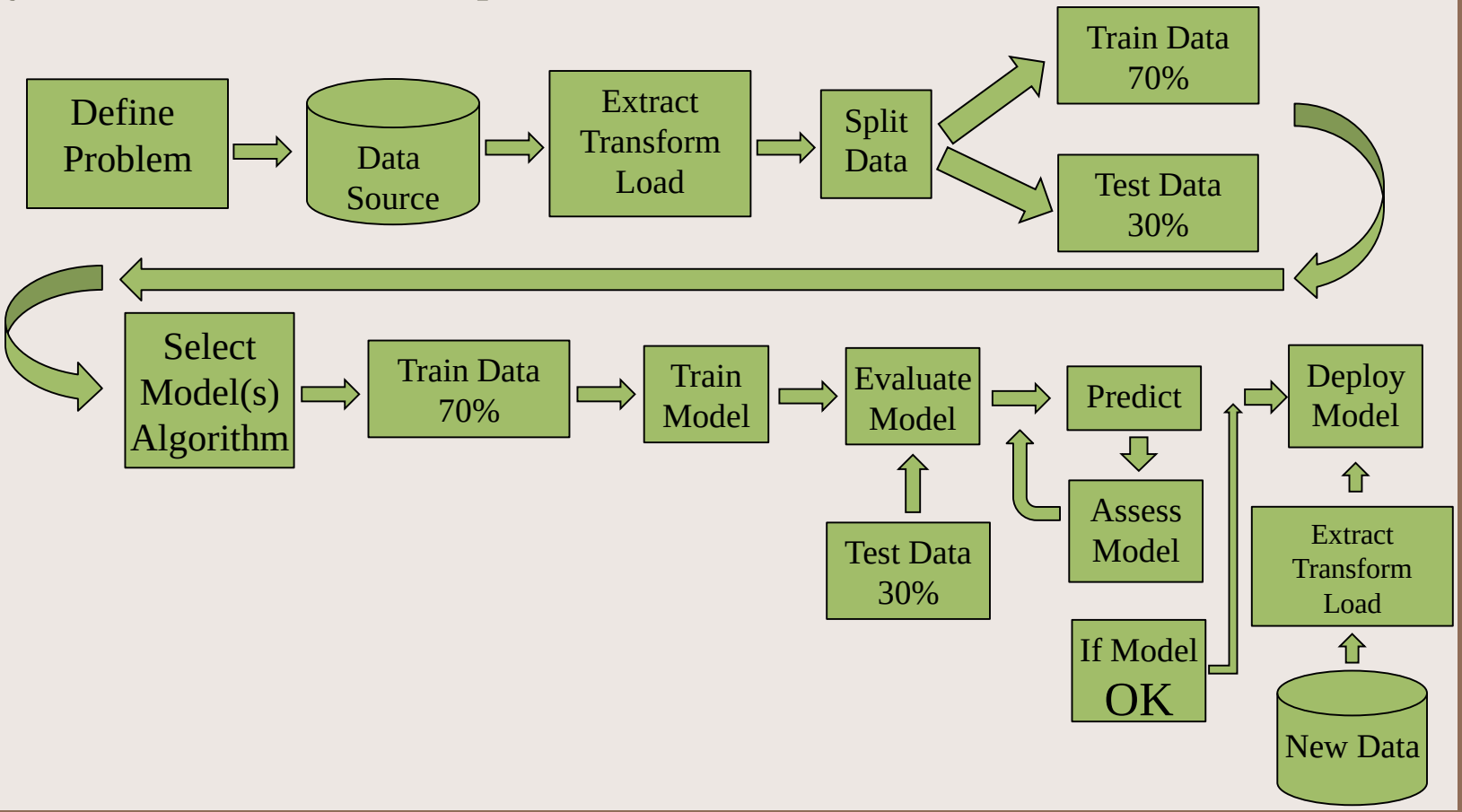


Those who don't plan, plan to fail!

by MHidalgo

Model Building Process

In order to build a predictive machine learning model in data science, you must follow a basic process as follows:



Model Building Process

In order to build a predictive machine learning model in data science, you must follow a basic process as follows:

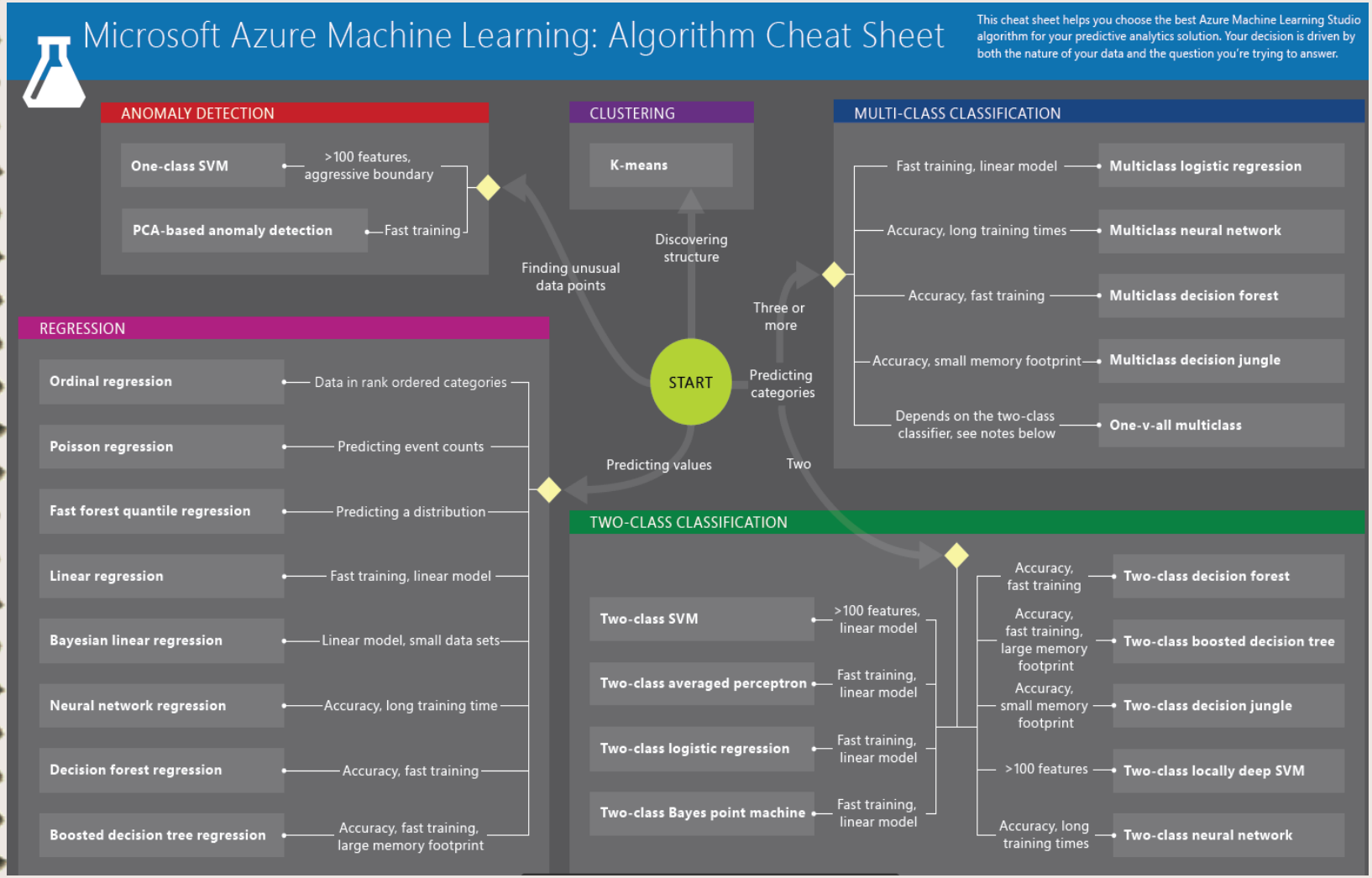
- Extract data from a source (Database, files, etc.)
- Load data into your analytical platform (Python, R, Knime, etc.)
- Explore, Visualize & cast your data (Understand your data & domain)
- Transform data (Clean, complete, derive, simplify, etc.) [Feature Engineering]
- Determine what type of problem you are solving (Classification, Regression, Clustering, Anomaly detection, Recommendation Engines, Deep Learning, etc.)

Machine Learning Categories

In general, there are 4 Machine Learning Categories:

- **Classification**
  - Binary (Predicts a Yes or No, Survived or Died, two possible outcomes)
  - Multiclass (Predicts/classifies data into different classes)
- **Regression**
  - Predicts a number
- **Clustering**
  - Group data into similar clusters
- **Anomaly Detection**
  - Detecting something out of the norm

Machine Learning Algorithms

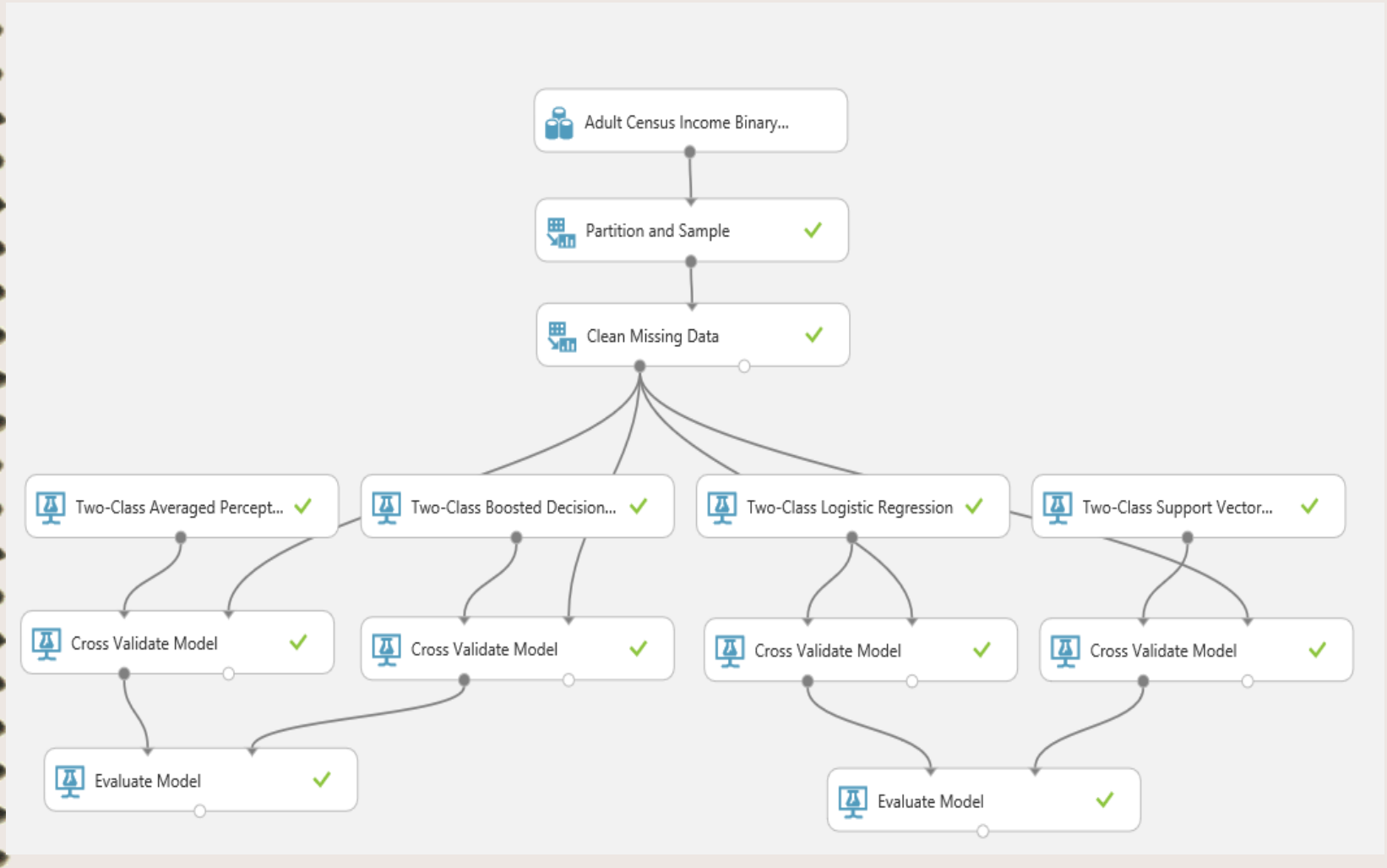


Model Building Process

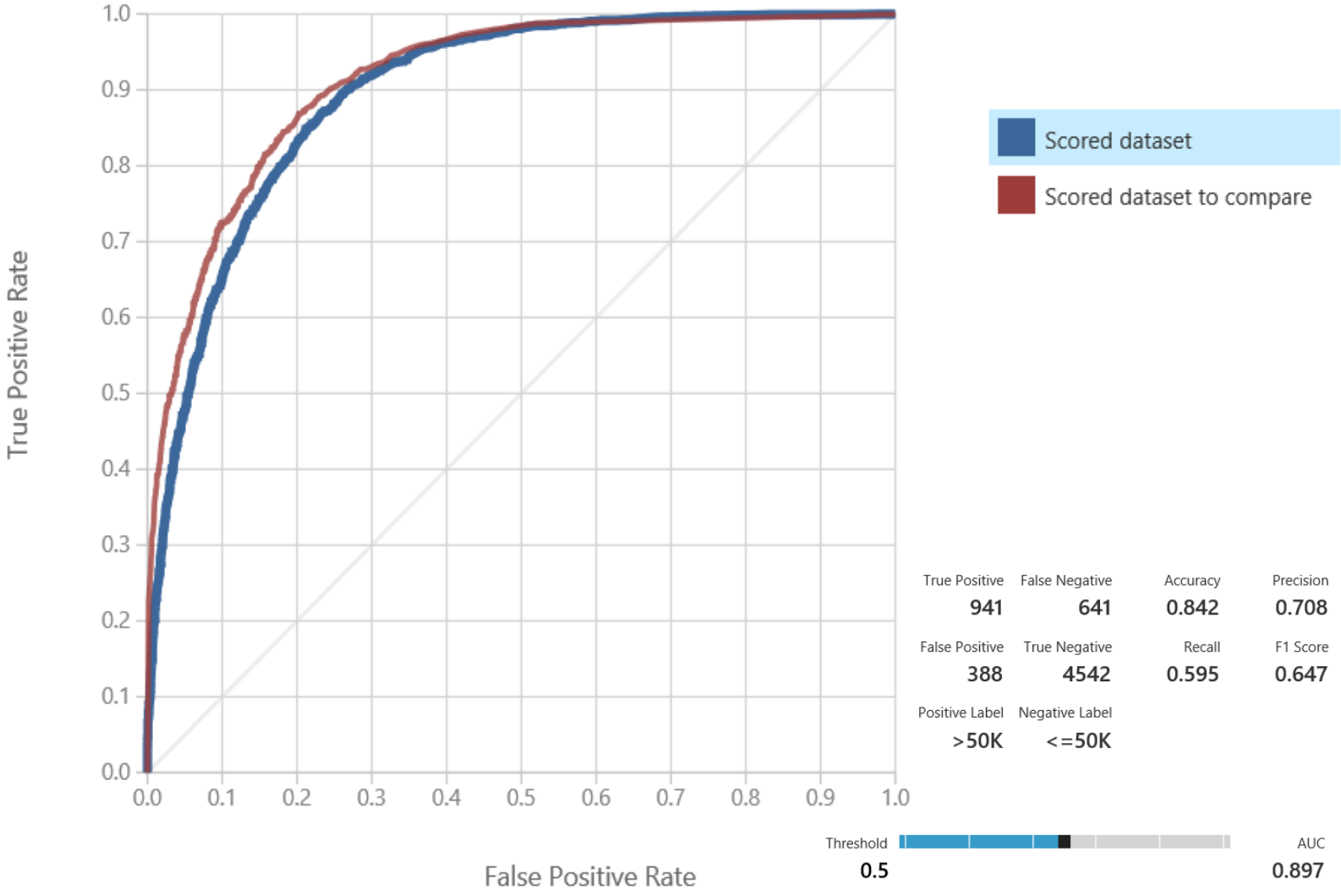
Continues…

- Split randomly your data in Train Set & Test Set (i.e. 70/30, Cross-validation)
- Select, configure & Fine Tune the type of ML algorithm (i.e. For classification –Random Forest, Logistic Regression, Neural Networks, Naïve Bayes, etc.)
- Understand the metrics to evaluate your model (Confusion matrix, RMSE, MSE, etc.)
- Train/Configure and review the results of the ML model (Outcome-Trained Model)
- Test and Score/Predict the results of the ML model (Predict with the Trained Model and new Test data)
- Deploy your ML model (on-line, in a server, or your laptop)
- Continue feeding new data, retrain ML model, Predict (Continuous Loop)

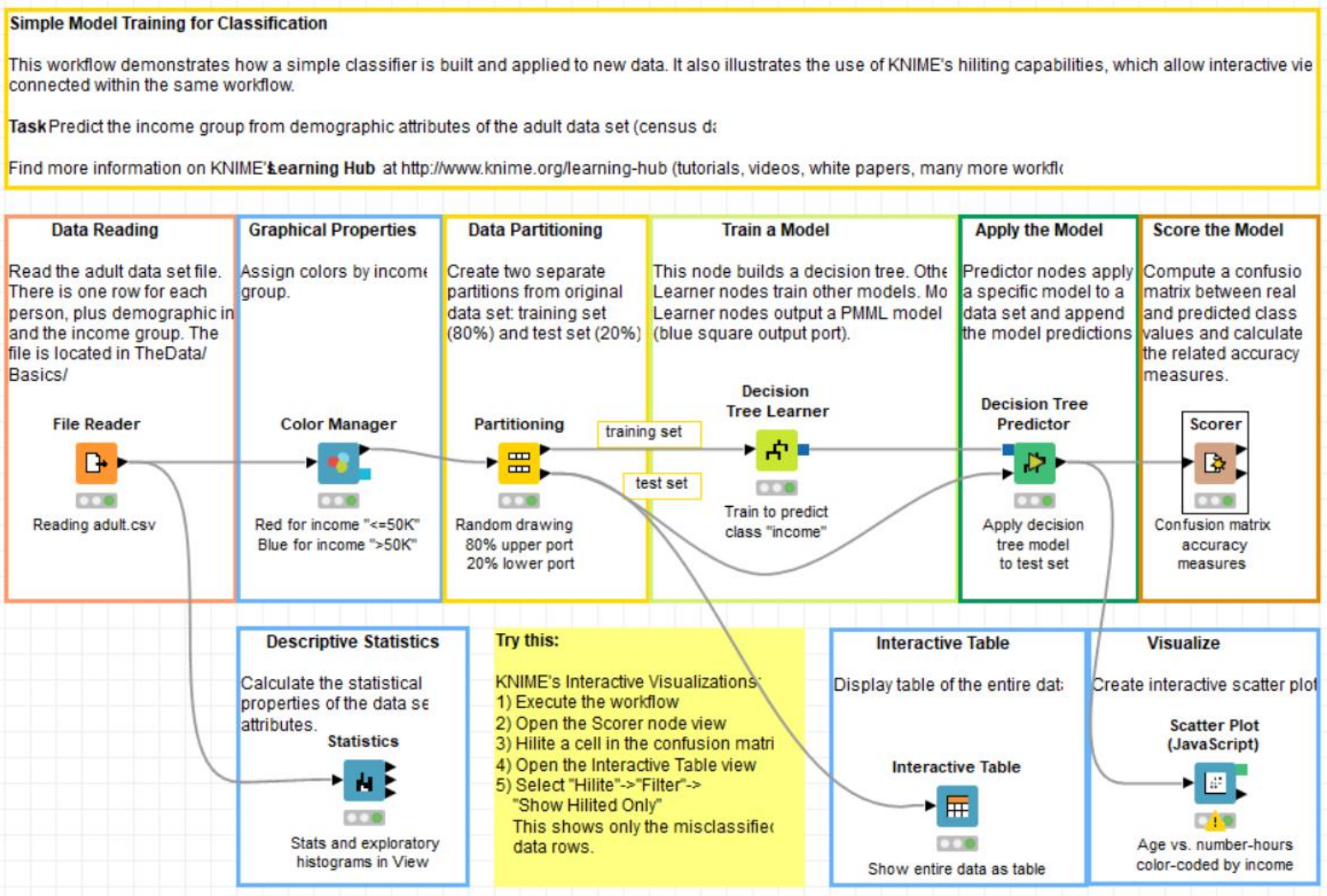
Analytical Platforms
Simple Classifier with ML Azure



Simple Classifier with ML Azure Model Comparison/Evaluation



Analytical Platforms
Simple Classifier with KNIME



Simple Classifier with KNIME Model Evaluation

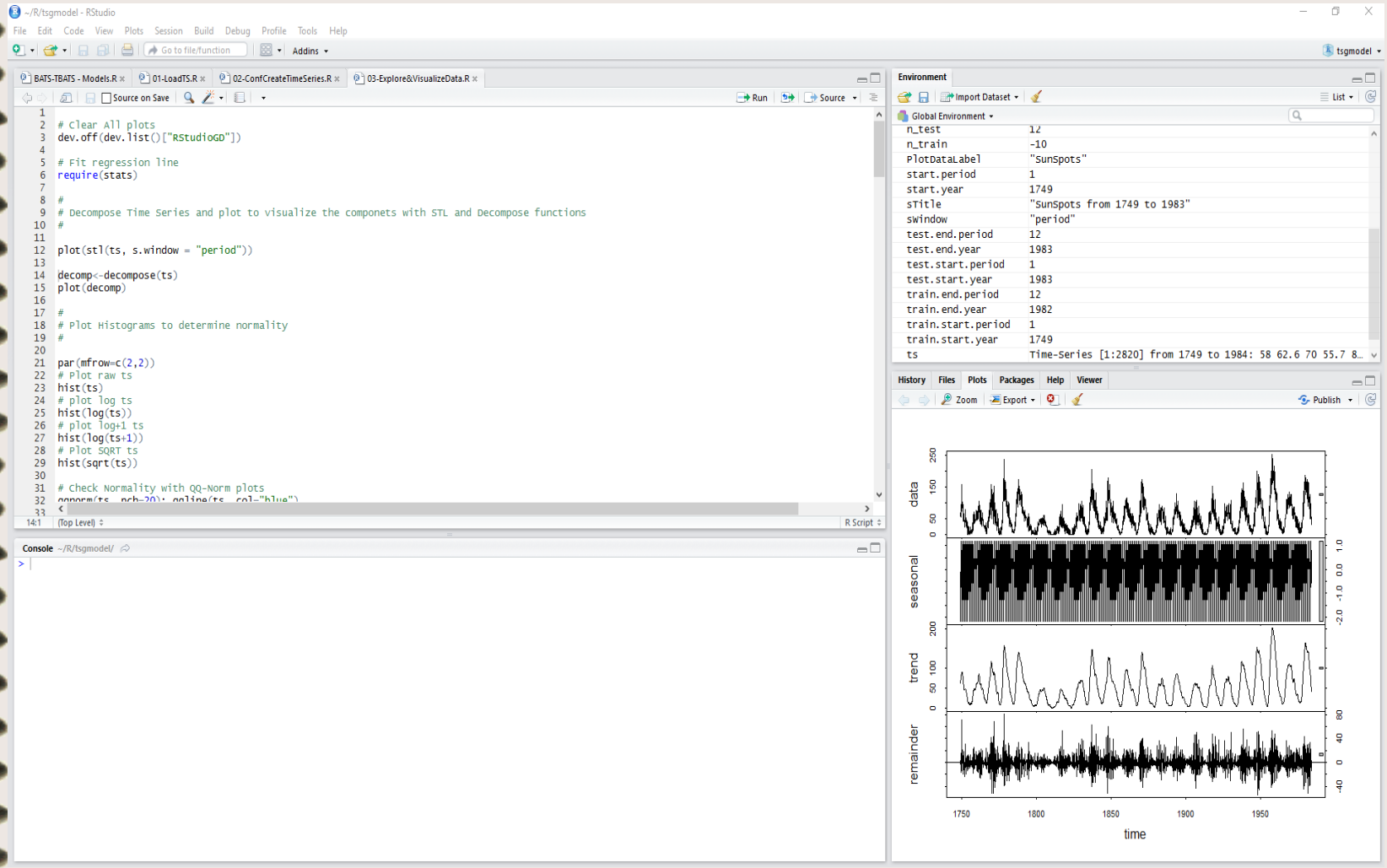
Accuracy Statistics

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specifity	F-meas...	Accuracy	Cohen'...
<=50K	4603	592	930	388	0.922	0.886	0.922	0.611	0.904	?	?
>50K	930	388	4603	592	0.611	0.706	0.611	0.922	0.655	?	?
Overall	?	?	?	?	?	?	?	?	?	0.85	0.559

Confusion Matrix

Row ID	<=50K	>50K
<=50K	4603	388
>50K	592	930

Time Series Analysis
R Studio



References

- *R-Cran Repository on-line (https://cran.r-project.org/)*
- *Rstudio (https://www.rstudio.com/)*
- *Jupyter Notebook/Python 3/Anaconda*
  - *(https://www.anaconda.com/download/)*
  - *(http://jupyter.org/install)*
- *Kaggle online site (https://www.kaggle.com/)*
- *Microsoft Machine Learning Azure (on-line)*

(https://azure.microsoft.com/en-us/services/machine-learning-studio/)
(https://studio.azureml.net/)

- *Microsoft Azure Machine Learning: Algorithm Cheat Sheet*
- *Knime on-line Site (https://www.knime.com/)*
- *Packt Publishing (https://www.packtpub.com/)*
- *Analytics Vidhya (https://www.analyticsvidhya.com)*
- *Business Science (http://www.business-science.io/)*
- *CRISP-DM (https://www.the-modeling-agency.com/crisp-dm.pdf)*
- *Implementing Analytics by Nauman Sheikh*
- *Many Internet sites!*

Credentials

- *J. Miguel Hidalgo:*
  - *Master in Business Administration in E-Commerce*
  - *B.S. in Computer Science*
  - *B.S. Professional Aeronautics*
  - *Certified DSDM Agile Project Management Foundations by APM Group (UK)*
  - *ITIL Foundations certified*
  - *Certified Project Management Professional (PMP) by PMI*
  - *Certified Lean Sigma Black Belt (Ohio State University)*
  - *Certified ISO9001:2015 Lead Auditor (LRQA)*
  - *Data Science Dojo Bootcamp*
  - *Coursera – John Hopkins Univ. Data Science courses*
  - *edX- Microsoft Data Science Certificate Courses*
  - *U.S. Army Aviation Electrical Technician*
  - ***Life Long Learner***

Practical Data Science & Machine Learning

"When you have exhausted all possibilities, remember this - you haven't."

- Thomas Edison

Practical Data Science & Machine Learning

Q & A