Business Applications of Machine Learning

JUNE 1, 2017

Paige Bailey

Roadmap


Logistics
What is machine learning?
Business Applications
Case Study
Additional Resources

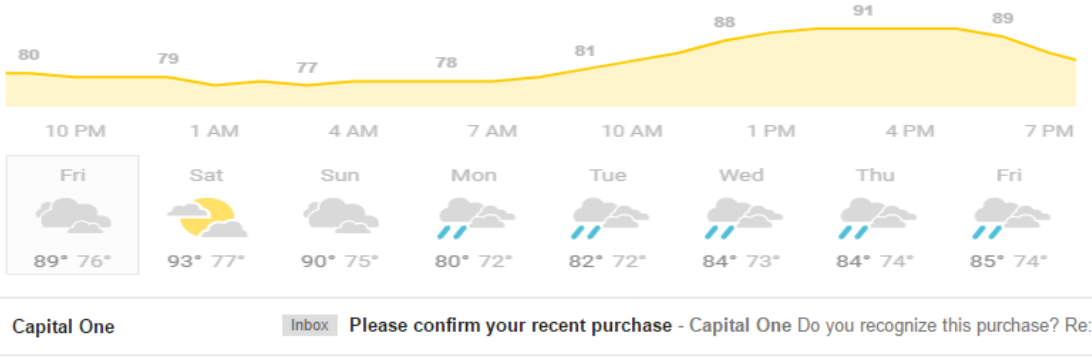
What is Machine Learning?

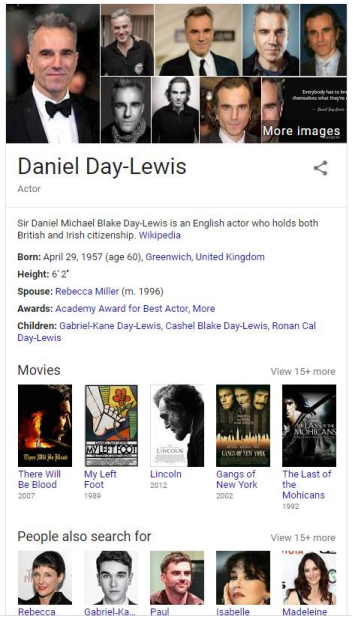
DATA EXPLORATION // TYPES OF LEARNING // MODEL TRAINING AND EVALUATION

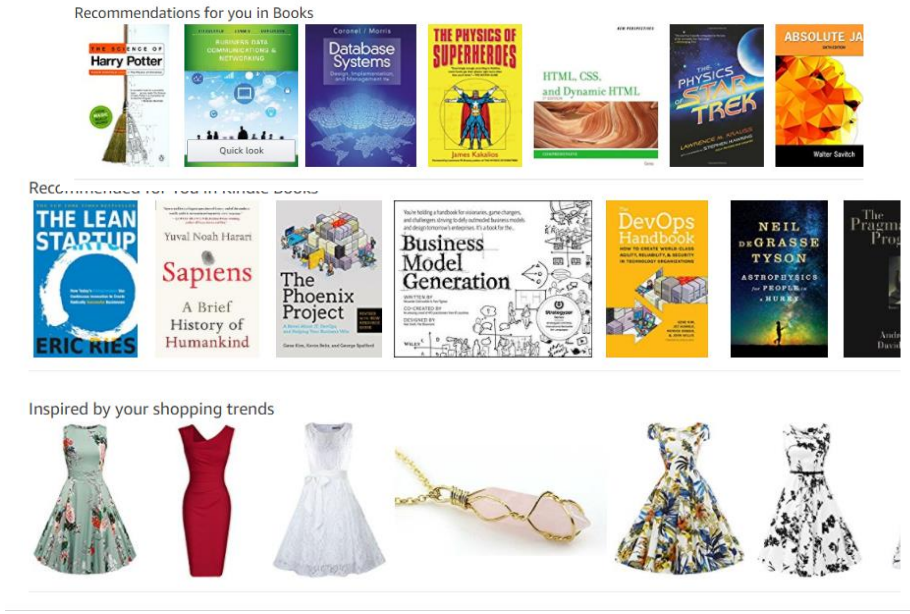
What is machine learning?

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

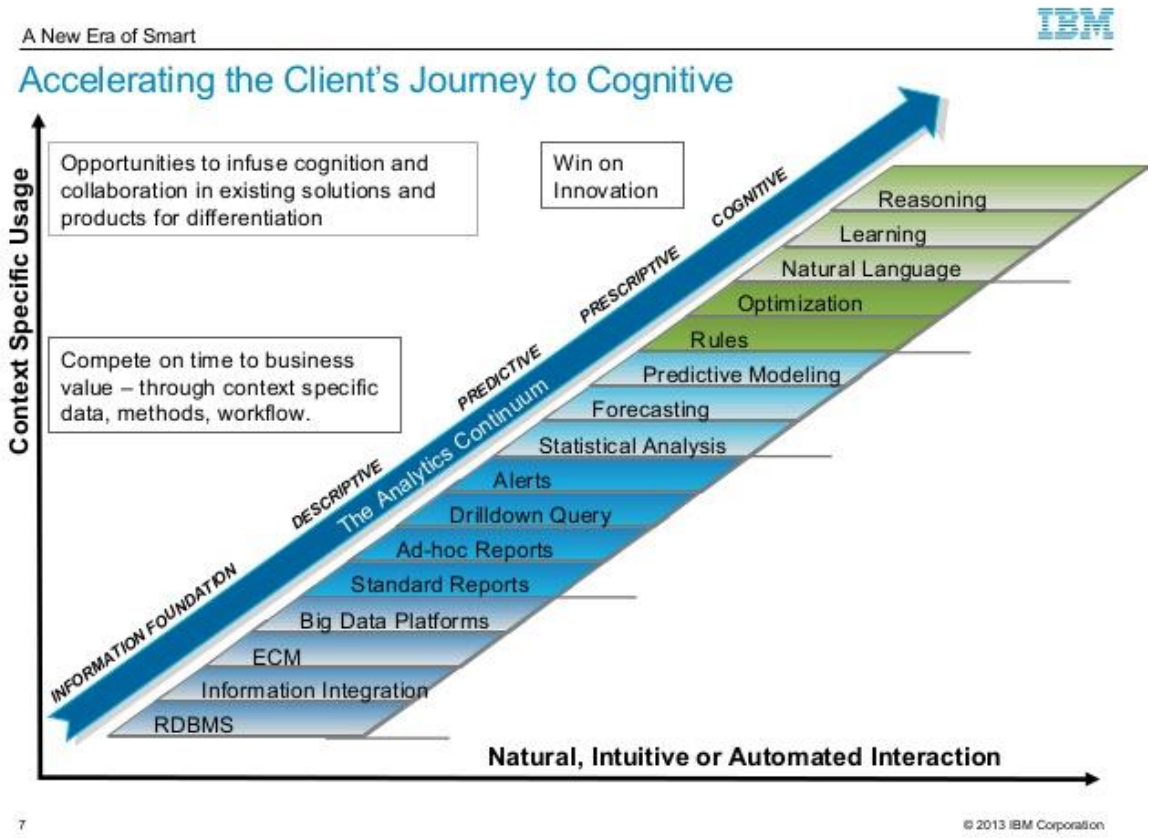
- Examples
  - Amazon and Netflix recommendations
  - Fraud detection for credit card transactions
  - Roomba deciding when a room is clean
  - Self-driving cars
  - Student performance predictions





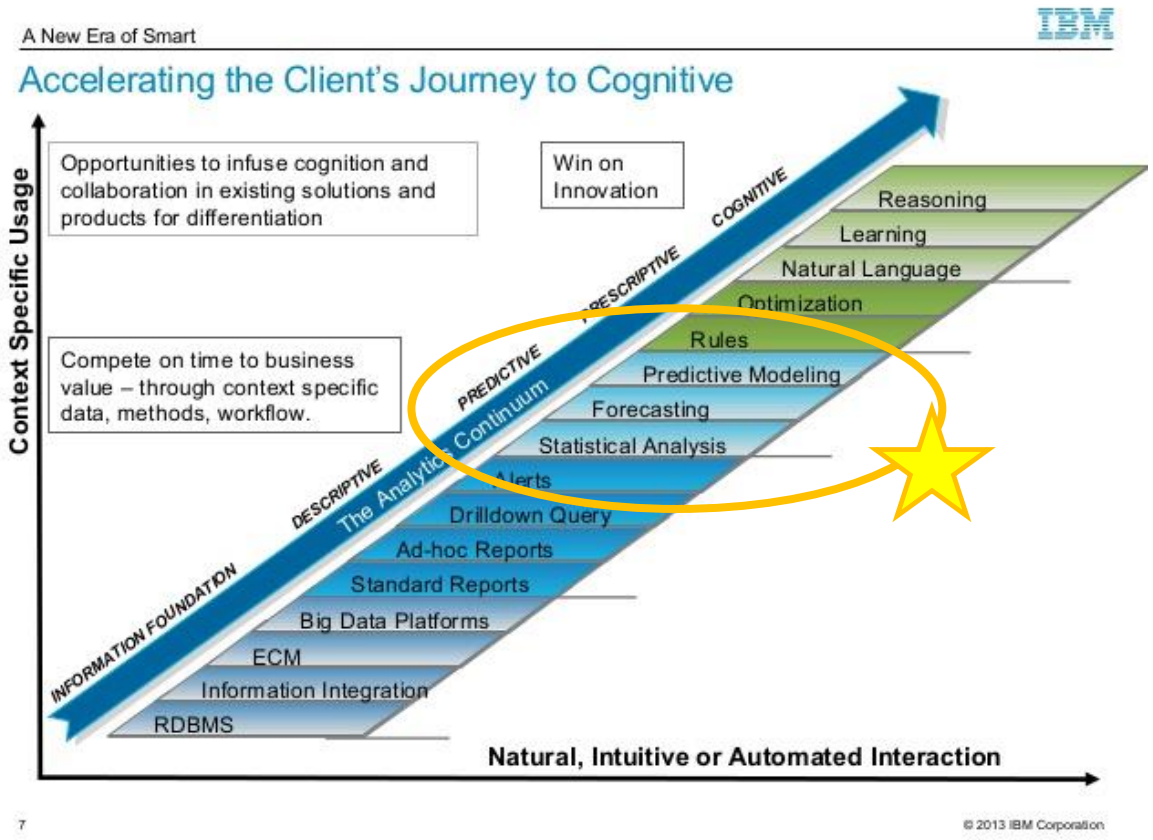


Data exploration, descriptive analytics, and predictive analytics



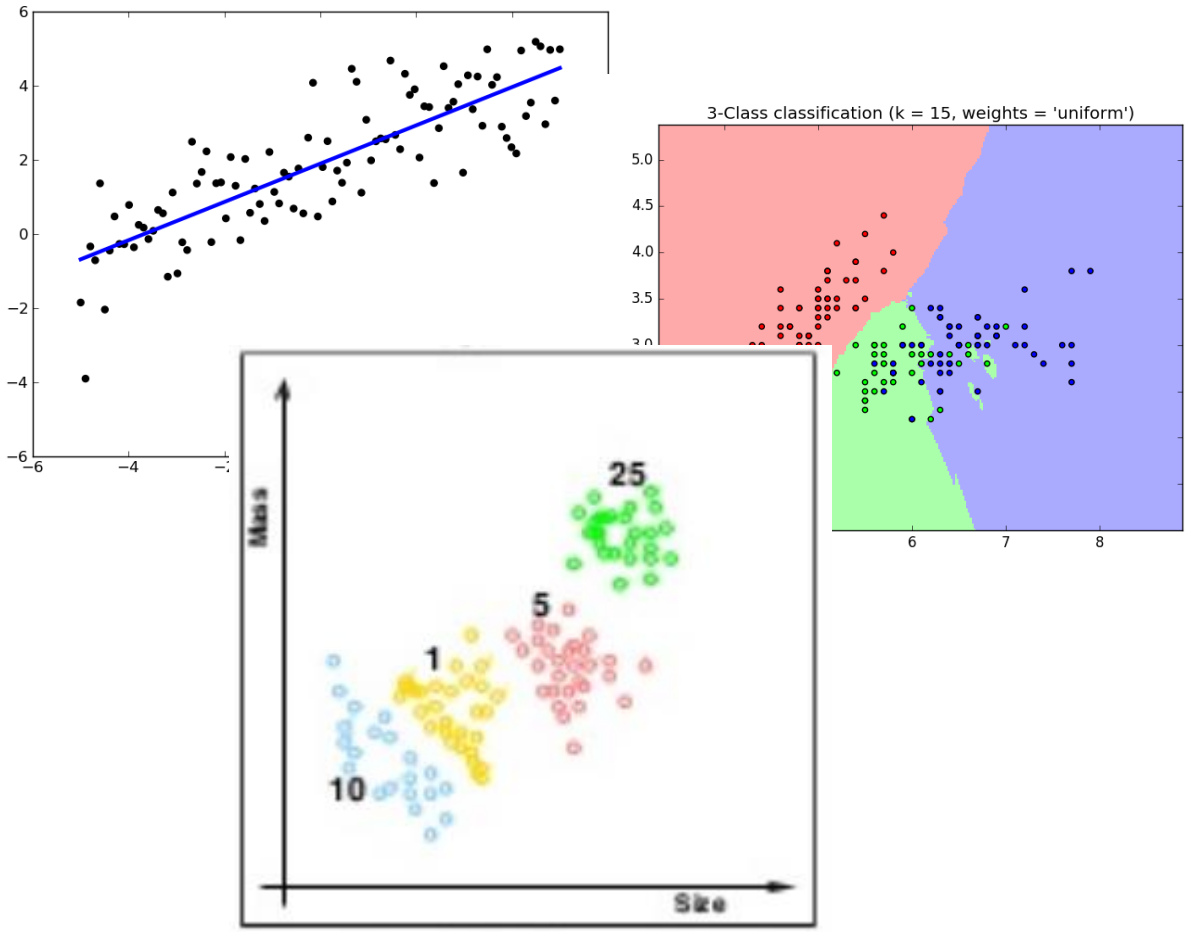
- Data Exploration: process of gathering information about a large and often unstructured data set in order to find characteristics for focused analysis.
- Data Mining: automated data exploration.
- Descriptive Analytics: process of analyzing a data set in order to summarize what happened.
  - Most **business 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

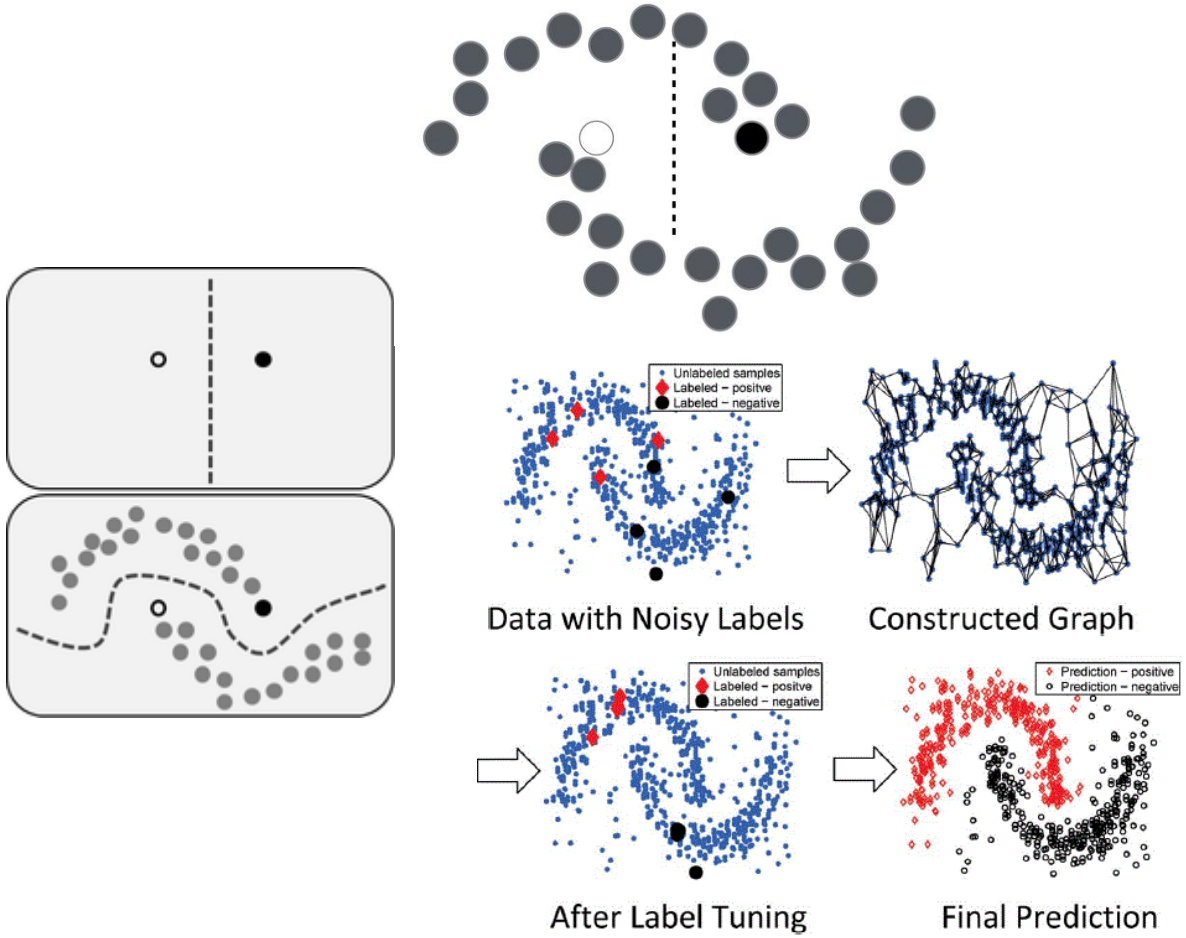


- Supervised Learning: algorithms are trained with labeled data.
  - Classification: used when the output variable is a category, such as “red” / “blue” or “disease” / “no disease”.
  - Regression: often used when the output variable is a real, numeric value – like dollars, weights.

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

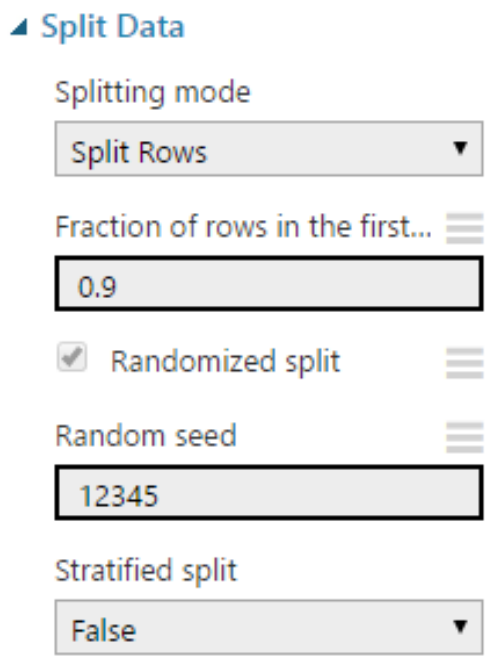


- Used when you have a large amount of input data (X) and **only some of the data is labeled** (Y).
- **Example**: photo archive where only some of the images are labeled, and the majority are unlabeled.
- Many real-world machine learning problems fall into this area
  - *Expensive and time-consuming* 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



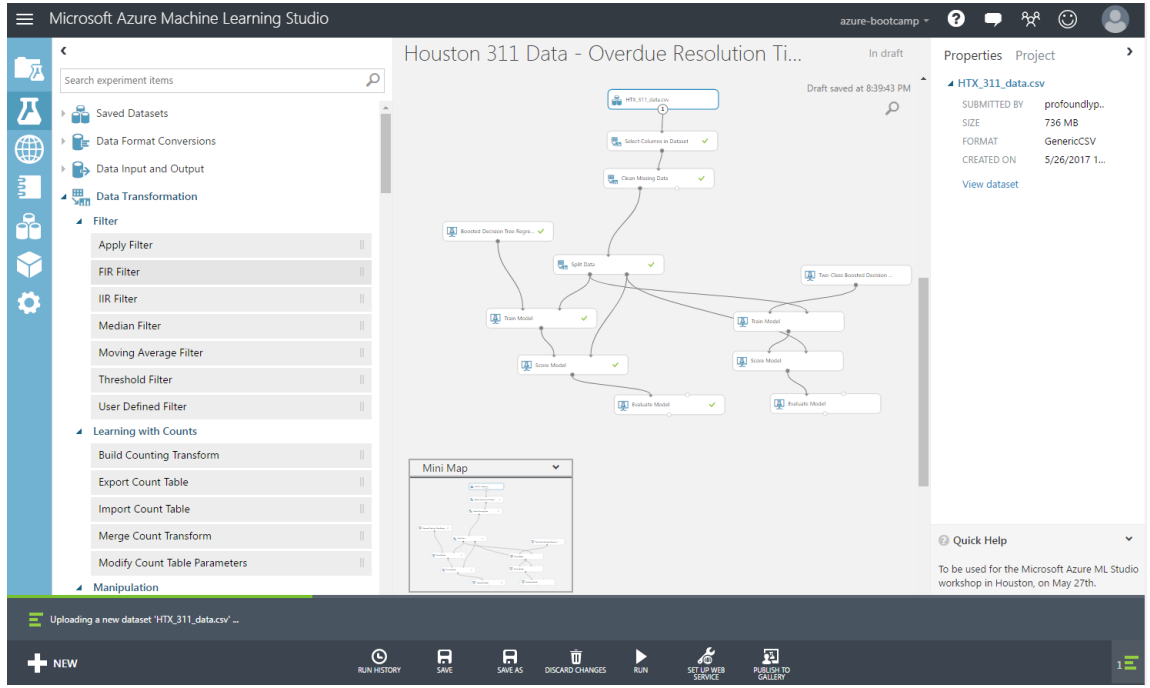
Additional vocabulary

- 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?

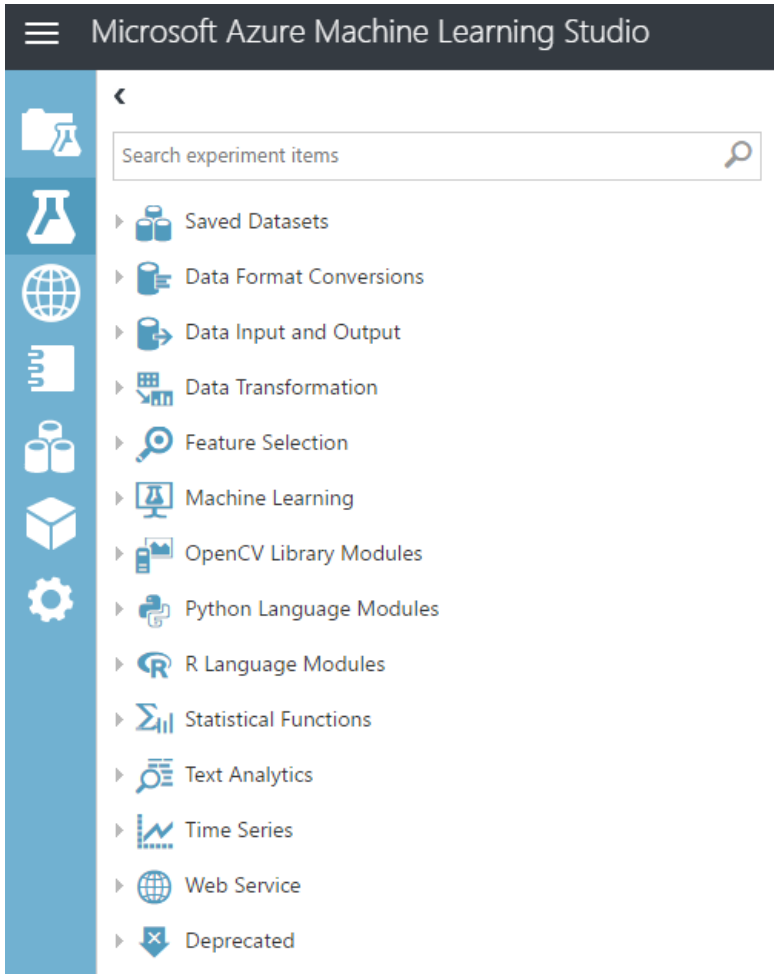
What is Azure Machine Learning Studio?

- 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.

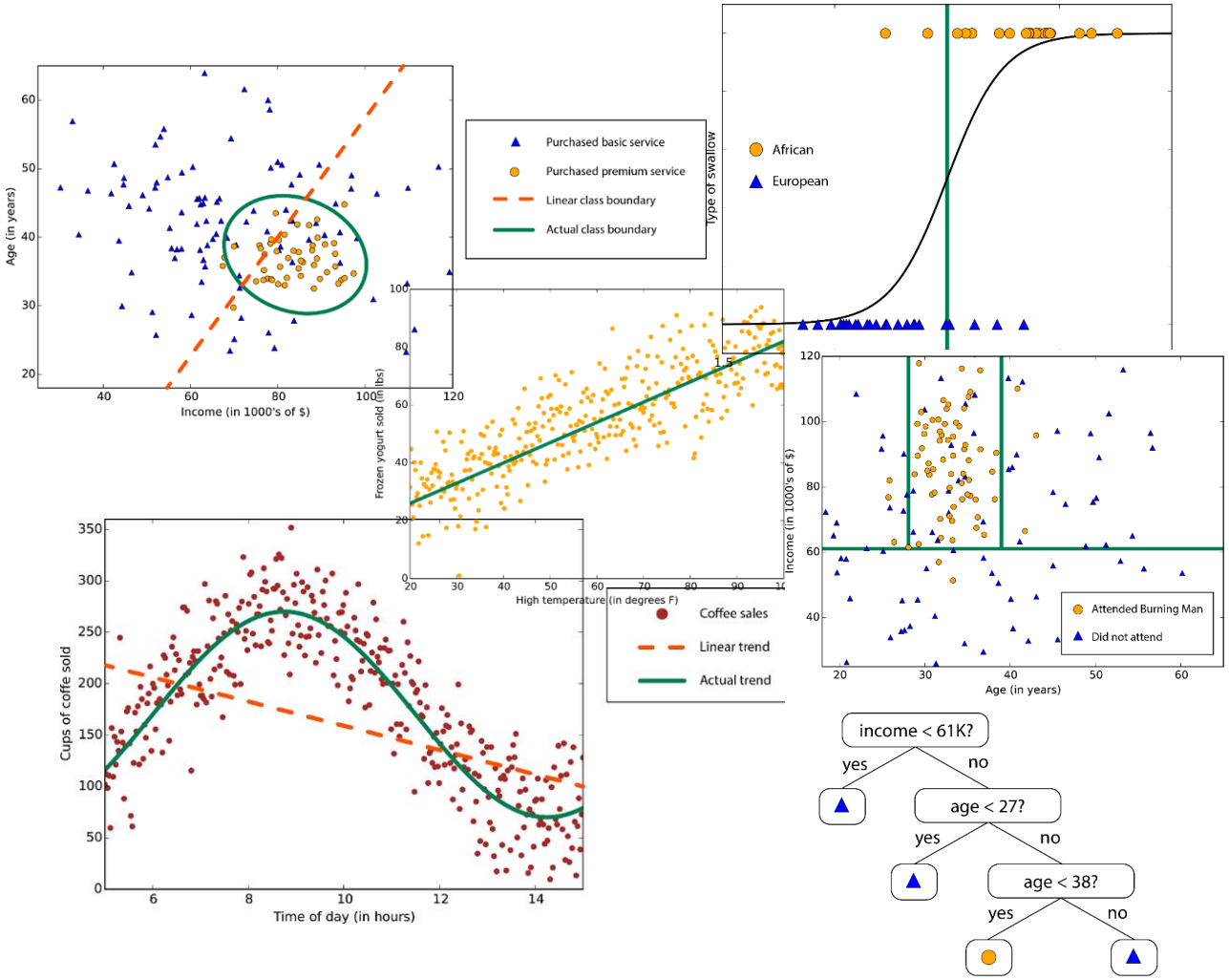


Modules – Machine Learning

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

Machine Learning in the Wild

INITIALIZE // TRAIN // SCORE // EVALUATE

How do companies leverage machine learning?

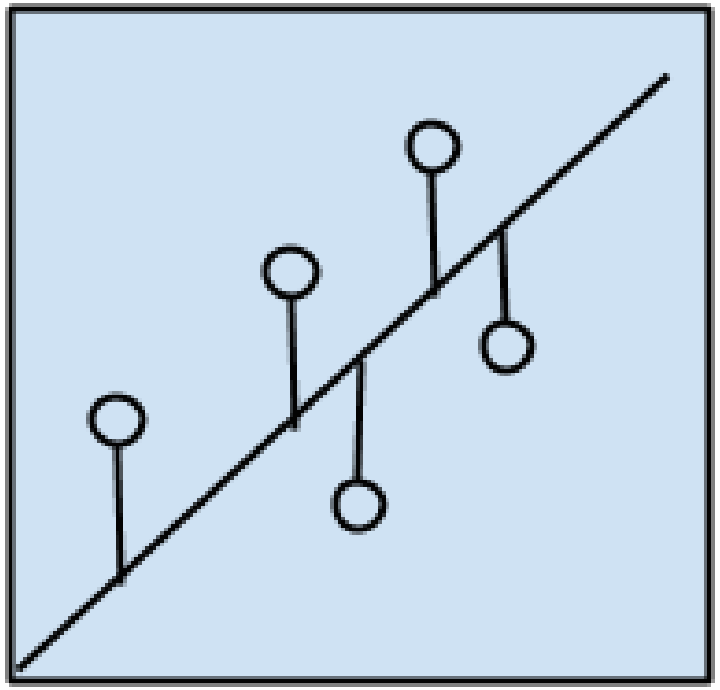
- **Over the next 5- 10 years, the biggest business gains will likely stem from getting the right information to the right people at the right time.**

- Machine learning turbocharges finding patterns and automates value extraction.
  - Better allocation and distribution of resources
  - Data-driven, real-time economy
  - On-demand production of goods and services
  - Lower failure rates
  - Better predictability

Regression

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

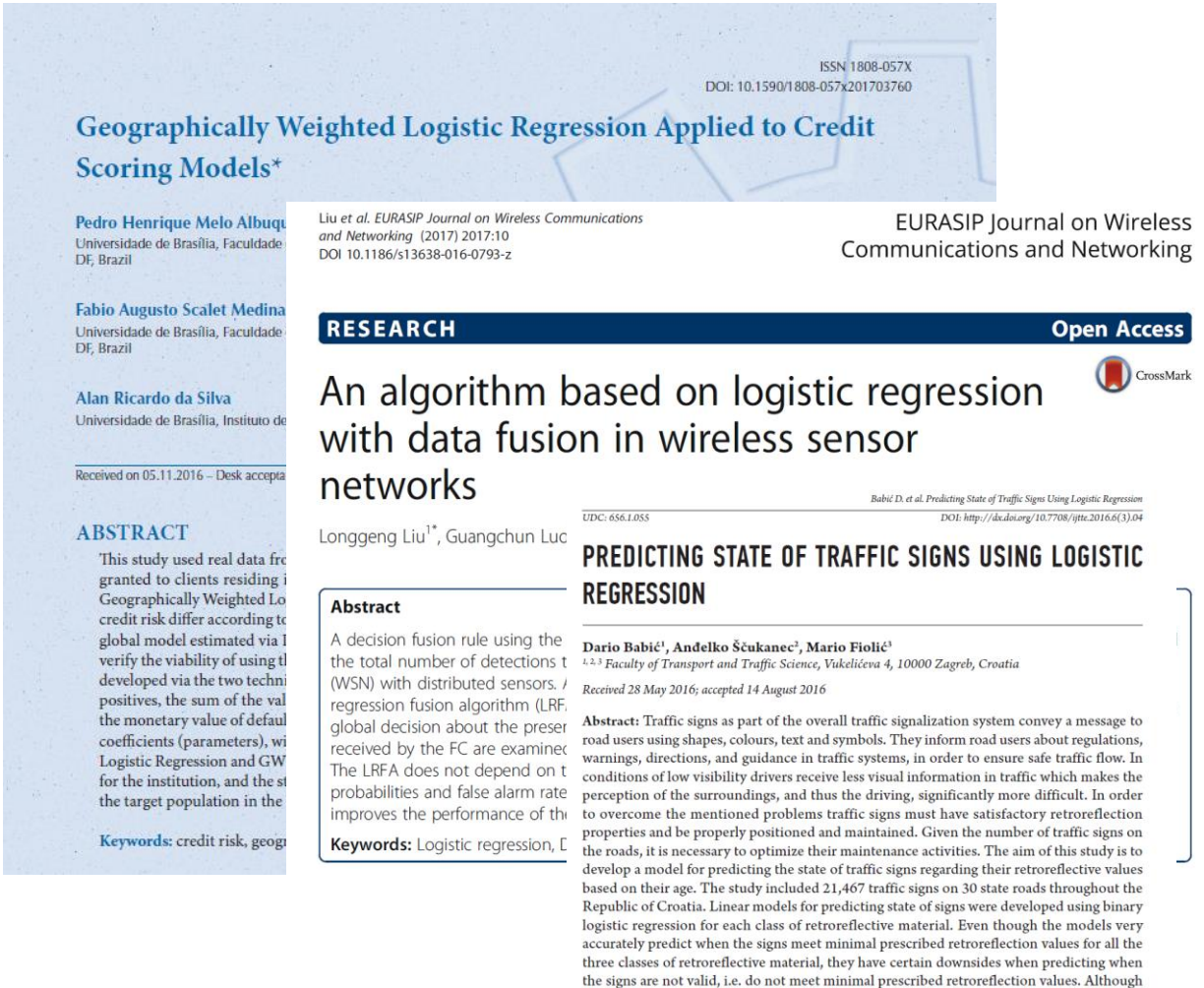
- **Most popular regression algorithms:**
  - Ordinary Least Squares
  - Linear Regression
  - Logistic Regression
  - Stepwise Regression
  - Multivariate Adaptive Regression Splines (MARS)
  - Locally Estimated Scatterplot Smoothing (LOESS)



Regression Algorithms

Regression - Examples

- How does the number of hours spent studying for an exam impact the probability that a student will pass?
- How can you predict financial losses for your credit card loaning business?
- How often should traffic signs be maintained?
- How can you predict the number of widgets produced in a factory, or the stock that you should keep in your facility?



ISSN 1808-057X
DOI: 10.1590/1808-057x201703760

Geographically Weighted Logistic Regression Applied to Credit Scoring Models*

Pedro Henrique Melo Albuqu
Universidade de Brasília, Faculdade
DF, Brazil

Fabio Augusto Scalet Medina
Universidade de Brasília, Faculdade
DF, Brazil

Alan Ricardo da Silva
Universidade de Brasília, Instituto de

Received on 05.11.2016 – Desk accepta

ABSTRACT

This study used real data fr
granted to clients residing
Geographically Weighted Lo
credit risk differ according t
global model estimated via
verify the viability of using t
developed via the two techni
positives, the sum of the val
the monetary value of defaul
coefficients (parameters), wi
Logistic Regression and GW
for the institution, and the s
the target population in the

Keywords: credit risk, geogr

Liu et al. EURASIP Journal on Wireless Communications and Networking (2017) 2017:10
DOI 10.1186/s13638-016-0793-z

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

CrossMark

An algorithm based on logistic regression with data fusion in wireless sensor networks

Longgeng Liu[1*], Guangchun Luo

Abstract

A decision fusion rule using the
the total number of detections t
(WSN) with distributed sensors.
regression fusion algorithm (LRF
global decision about the preser
received by the FC are examine
The LRFA does not depend on t
probabilities and false alarm rate
improves the performance of th

Keywords: Logistic regression, D

Babić D. et al. Predicting State of Traffic Signs Using Logistic Regression

UDC: 656.1.055 DOI: http://dx.doi.org/10.7708/ijtte.2016.6(3).04

PREDICTING STATE OF TRAFFIC SIGNS USING LOGISTIC REGRESSION

Dario Babić[1], Anđelko Ščukanec[2], Mario Fiolić[3]
[1, 2, 3] Faculty of Transport and Traffic Science, Vukelićeva 4, 10000 Zagreb, Croatia

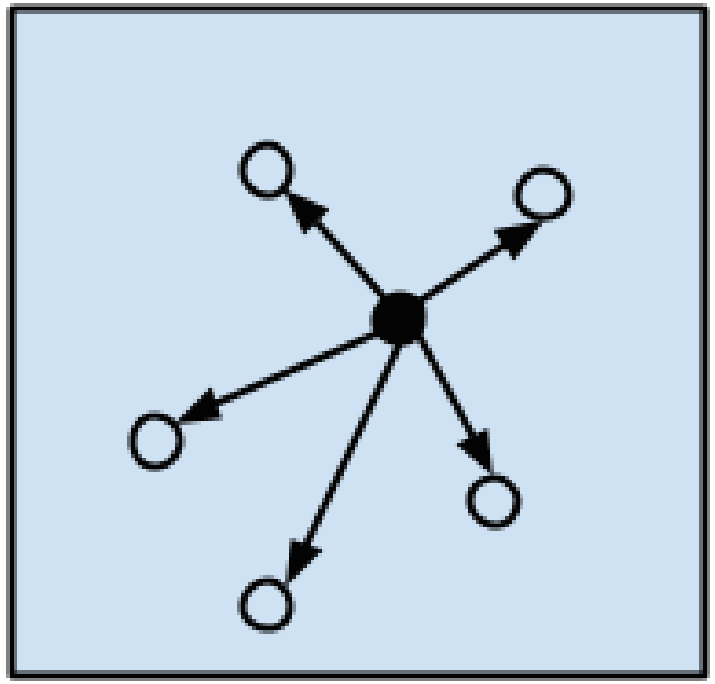
Received 28 May 2016; accepted 14 August 2016

Abstract: Traffic signs as part of the overall traffic signalization system convey a message to road users using shapes, colours, text and symbols. They inform road users about regulations, warnings, directions, and guidance in traffic systems, in order to ensure safe traffic flow. In conditions of low visibility drivers receive less visual information in traffic which makes the perception of the surroundings, and thus the driving, significantly more difficult. In order to overcome the mentioned problems traffic signs must have satisfactory retroreflection properties and be properly positioned and maintained. Given the number of traffic signs on the roads, it is necessary to optimize their maintenance activities. The aim of this study is to develop a model for predicting the state of traffic signs regarding their retroreflective values based on their age. The study included 21,467 traffic signs on 30 state roads throughout the Republic of Croatia. Linear models for predicting state of signs were developed using binary logistic regression for each class of retroreflective material. Even though the models very accurately predict when the signs meet minimal prescribed retroreflection values for all the three classes of retroreflective material, they have certain downsides when predicting when the signs are not valid, i.e. do not meet minimal prescribed retroreflection values. Although

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.

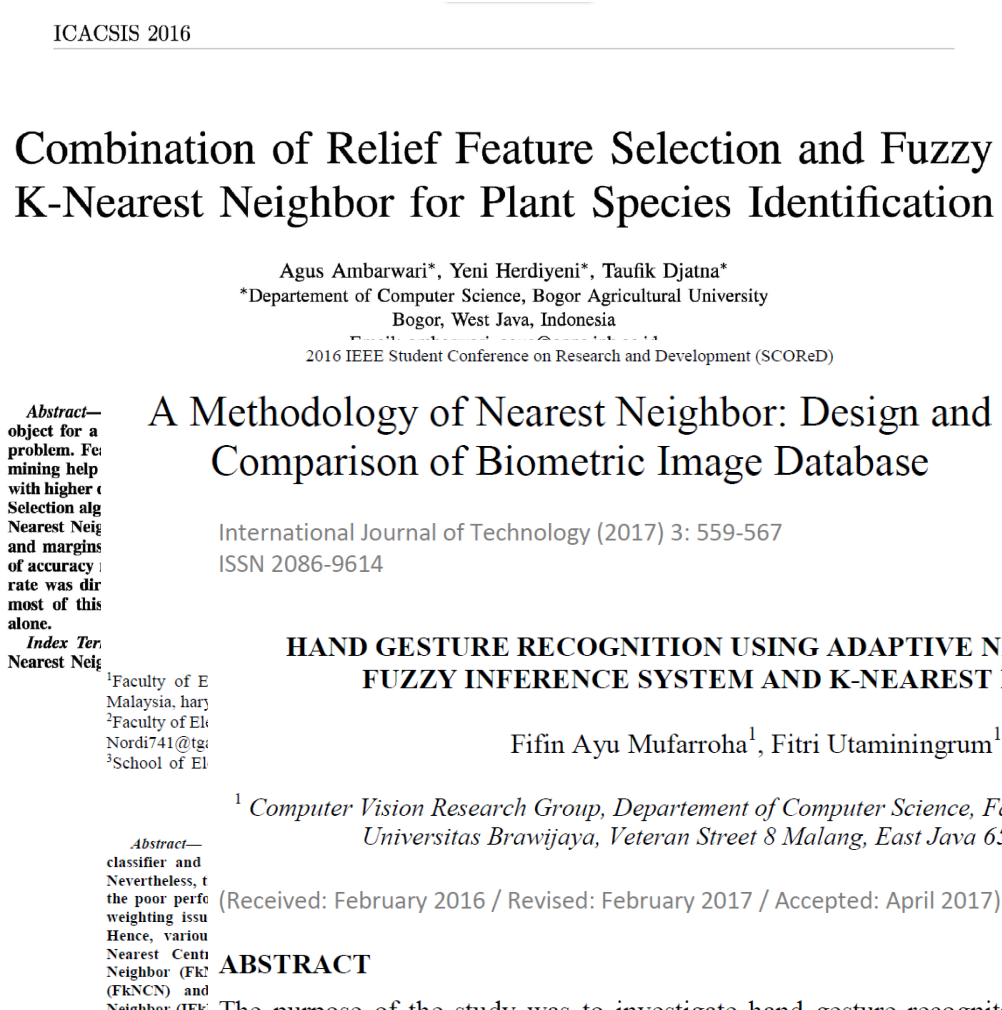
- **Most popular instance-based algorithms:**
  - K-Nearest Neighbor (kNN)
  - Learning Vector Quantization (LVQ)
  - Self-Organizing Map (SOM)
  - Locally Weighted Learning (LWL)



Instance-based Algorithms

Instance-based - Examples

- How can you detect handwritten characters?
- How can you detect which cells are potentially cancerous, and which are benign?
- How can your virtual reality headset detect which movements indicate hand gestures?
- How can you identify plant species based on shape, texture, and margin of the leaves?



ICACSIS 2016

Combination of Relief Feature Selection and Fuzzy K-Nearest Neighbor for Plant Species Identification

Agus Ambarwari*, Yeni Herdiyeni*, Taufik Djatna*
*Departement of Computer Science, Bogor Agricultural University
Bogor, West Java, Indonesia
2016 IEEE Student Conference on Research and Development (SCOReD)

A Methodology of Nearest Neighbor: Design and Comparison of Biometric Image Database

International Journal of Technology (2017) 3: 559-567
ISSN 2086-9614

© IJTech 2017

HAND GESTURE RECOGNITION USING ADAPTIVE NETWORK BASED FUZZY INFERENCE SYSTEM AND K-NEAREST NEIGHBOR

Fifin Ayu Mufarroha[1], Fitri Utaminingrum[1*]

[1] *Computer Vision Research Group, Departement of Computer Science, Faculty of Computer Science, Universitas Brawijaya, Veteran Street 8 Malang, East Java 65145, Indonesia*

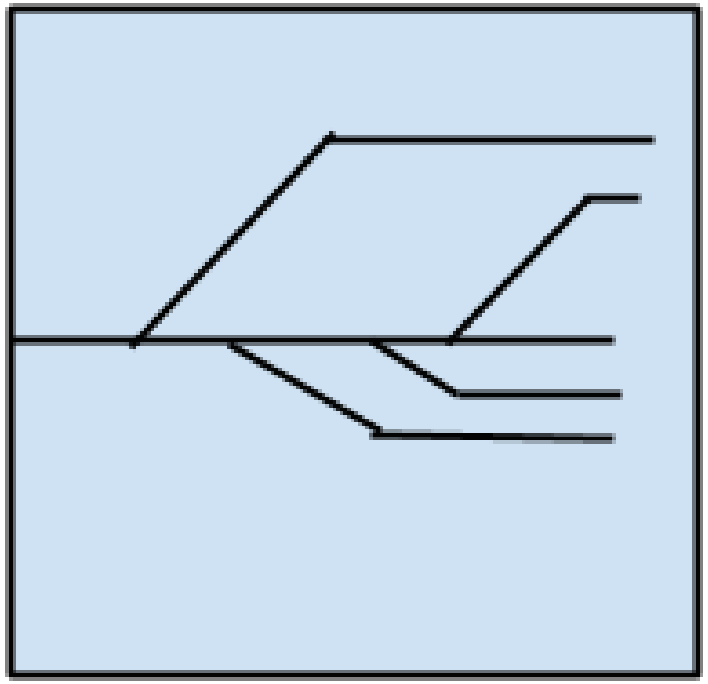
(Received: February 2016 / Revised: February 2017 / Accepted: April 2017)

ABSTRACT

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)

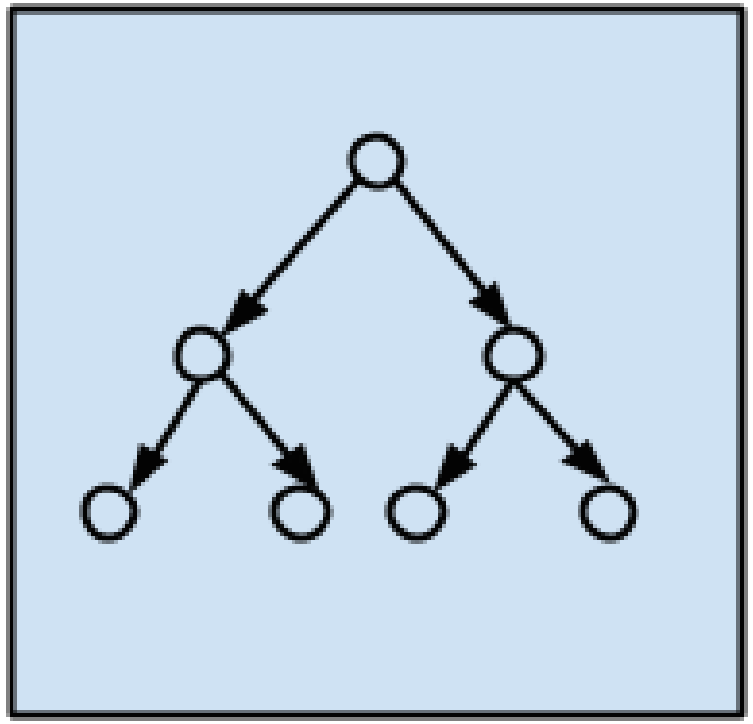


Regularization Algorithms

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.

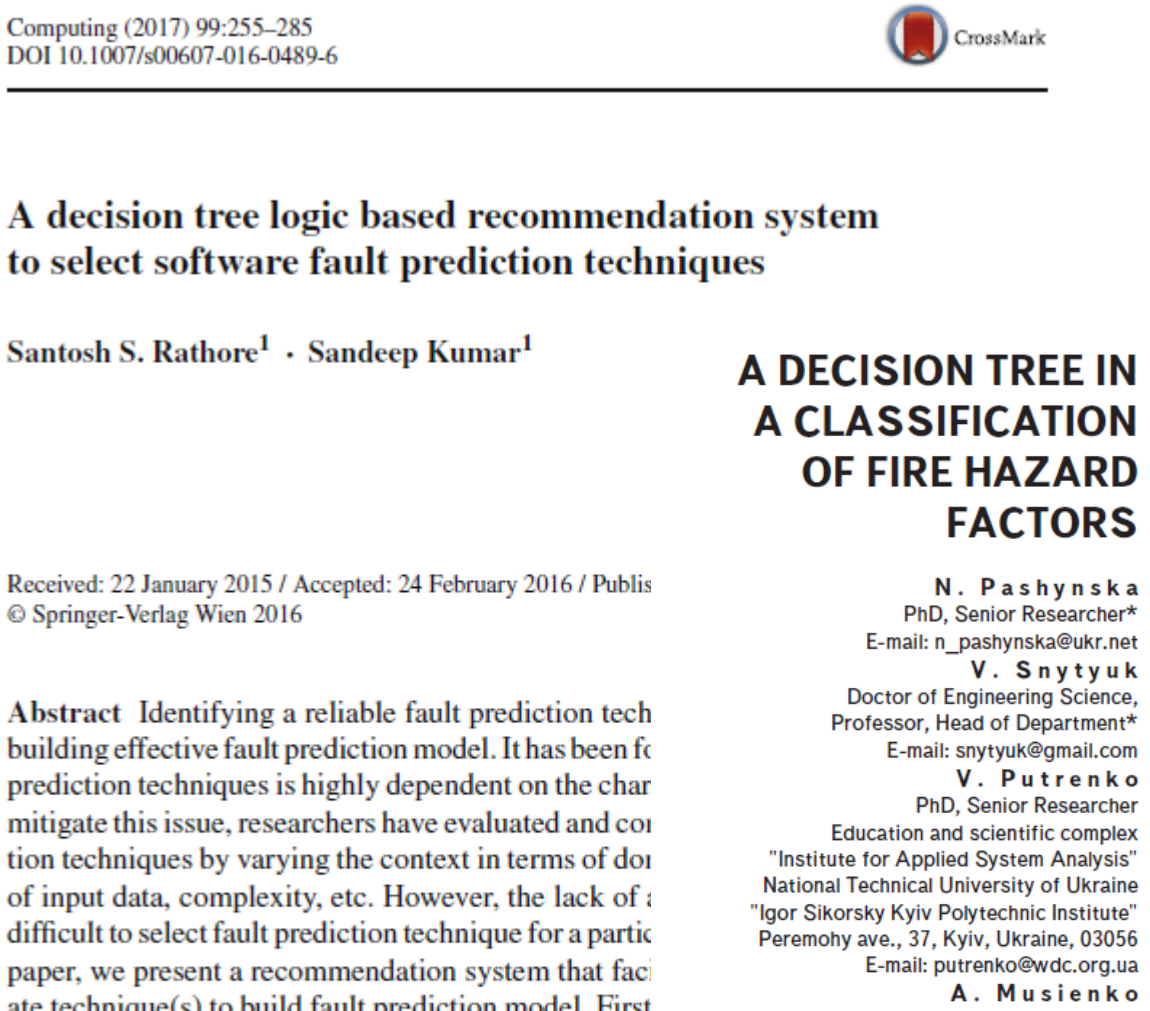
- **Most popular decision tree algorithms:**
  - Classification and Regression Tree (CART)
  - Iterative Dichotomiser 3 (ID3)
  - C4.5 and C5.0 (different versions of a powerful approach)
  - Chi-squared Automatic Interaction Detection (CHAID)
  - Decision Stump
  - M5
  - Conditional Decision Trees



Decision Tree Algorithms

Decision Trees - Examples

- Used primarily for models that will be implemented by human beings, and are potentially susceptible to legal ramifications
- Based on a potential customer's characteristics, should our company give them a loan?
- Based on a patient's symptoms, how should they be diagnosed?
- Based on time of day, day of month, season, and temperature, should we expect to see a surge in Texas electricity prices in Houston?
- Based on observed criteria, is it likely that a mushroom is poisonous?



Computing (2017) 99:255–285
DOI 10.1007/s00607-016-0489-6

CrossMark

A decision tree logic based recommendation system to select software fault prediction techniques

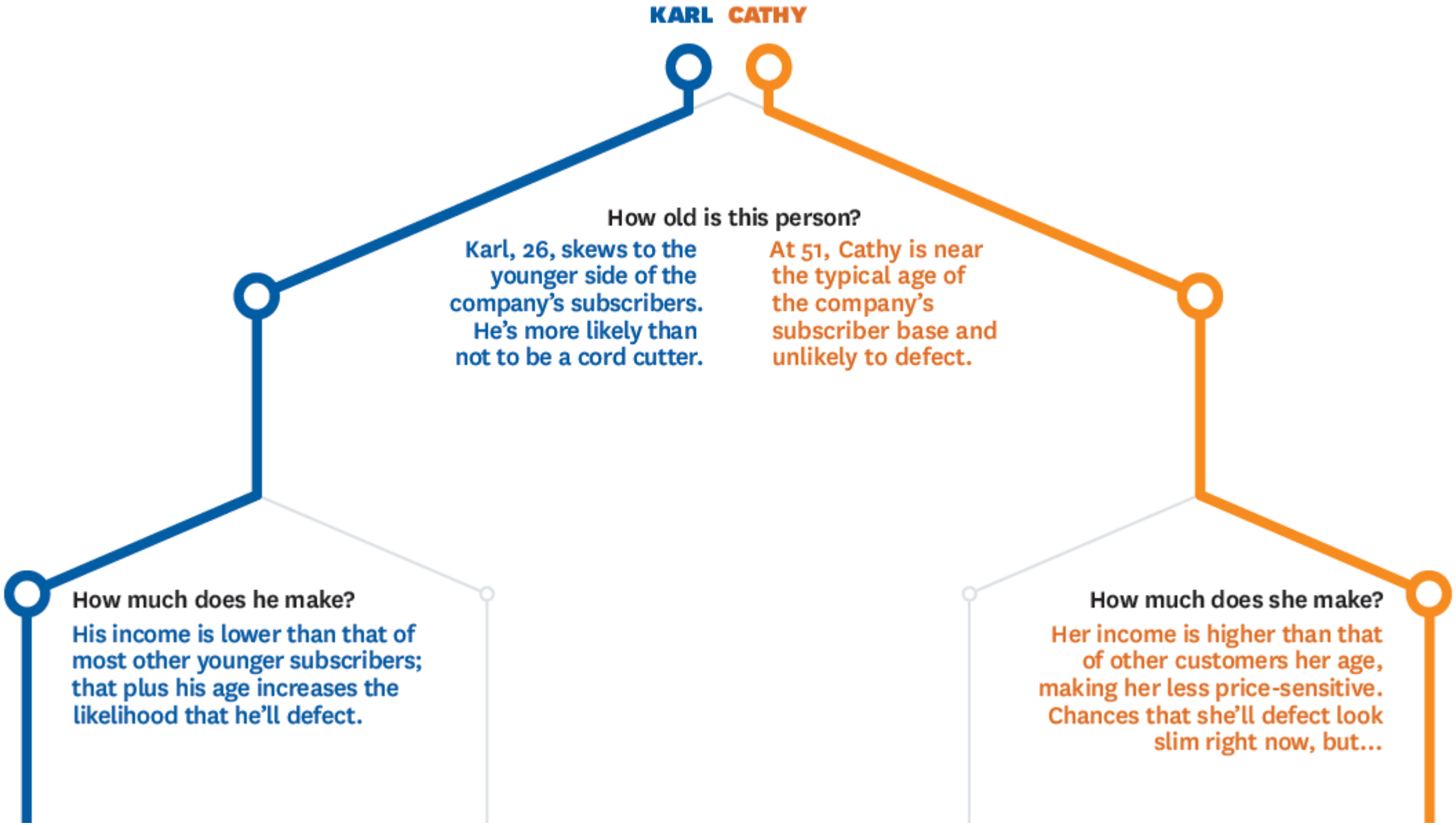
Santosh S. Rathore[1] · Sandeep Kumar[1]

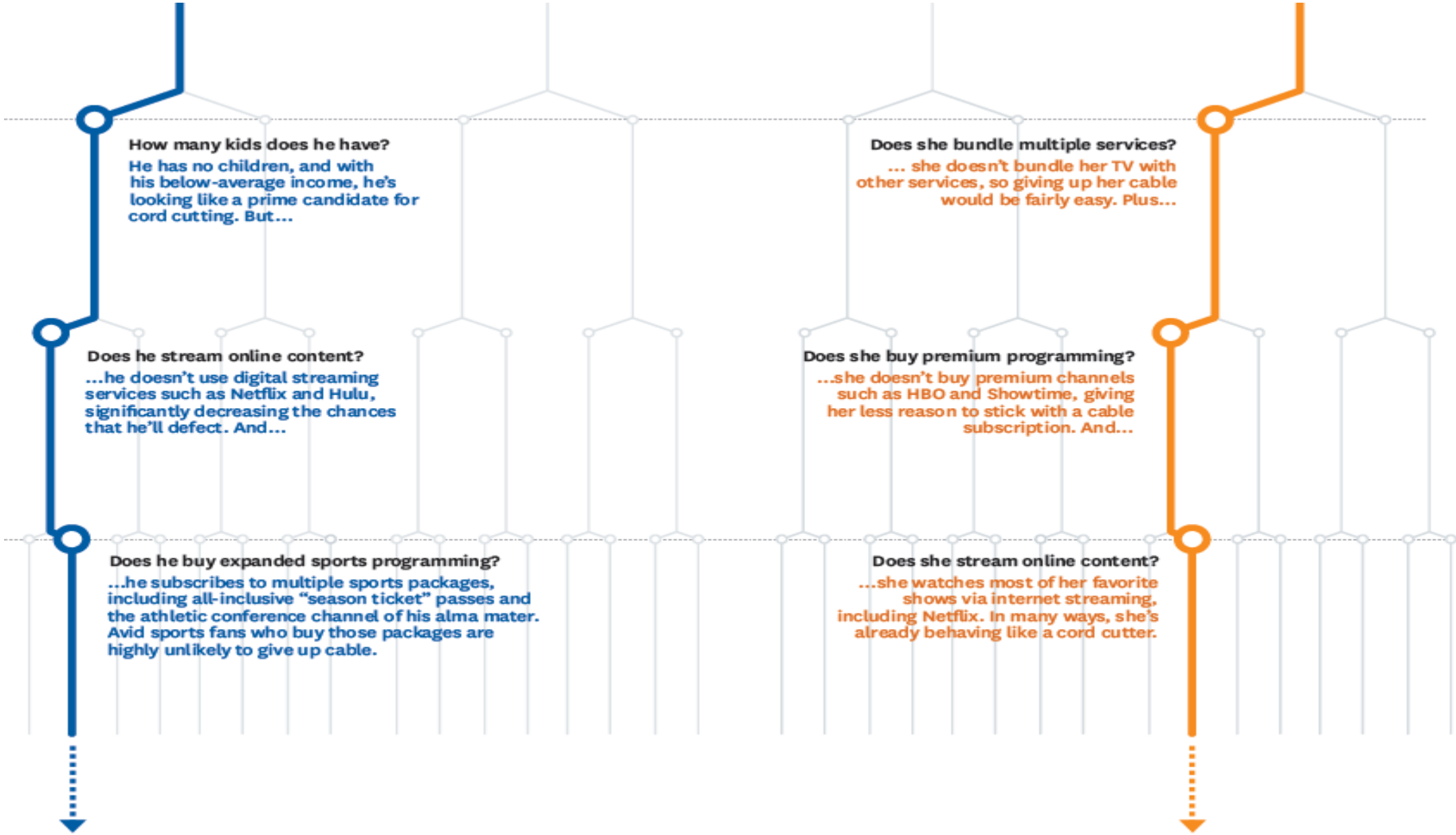
Received: 22 January 2015 / Accepted: 24 February 2016 / Publis

Abstract Identifying a reliable fault prediction tech
building effective fault prediction model. It has been fo
prediction techniques is highly dependent on the char
mitigate this issue, researchers have evaluated and co
tion techniques by varying the context in terms of do
of input data, complexity, etc. However, the lack of
difficult to select fault prediction technique for a partic
paper, we present a recommendation system that fac
ate technique(s) to build fault prediction model. First

A DECISION TREE IN A CLASSIFICATION OF FIRE HAZARD FACTORS

N. Pashynska
PhD, Senior Researcher*
E-mail: n_pashynska@ukr.net
V. Snytyuk
Doctor of Engineering Science,
Professor, Head of Department*
E-mail: snytyuk@gmail.com
V. Putrenko
PhD, Senior Researcher
Education and scientific complex
"Institute for Applied System Analysis"
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"
Peremohy ave., 37, Kyiv, Ukraine, 03056
E-mail: putrenko@wdc.org.ua
A. Musienko


KARL
CATHY
How old is this person?
Karl, 26, skews to the younger side of the company's subscribers. He's more likely than not to be a cord cutter.
At 51, Cathy is near the typical age of the company's subscriber base and unlikely to defect.
How much does he make?
His income is lower than that of most other younger subscribers; that plus his age increases the likelihood that he'll defect.
How much does she make?
Her income is higher than that of other customers her age, making her less price-sensitive. Chances that she'll defect look slim right now, but...


How many kids does he have?
He has no children, and with his below-average income, he's looking like a prime candidate for cord cutting. But...
Does he stream online content?
...he doesn't use digital streaming services such as Netflix and Hulu, significantly decreasing the chances that he'll defect. And...
Does he buy expanded sports programming?
...he subscribes to multiple sports packages, including all-inclusive "season ticket" passes and the athletic conference channel of his alma mater. Avid sports fans who buy those packages are highly unlikely to give up cable.
Does she bundle multiple services?
... she doesn't bundle her TV with other services, so giving up her cable would be fairly easy. Plus...
Does she buy premium programming?
...she doesn't buy premium channels such as HBO and Showtime, giving her less reason to stick with a cable subscription. And...
Does she stream online content?
...she watches most of her favorite shows via internet streaming, including Netflix. In many ways, she's already behaving like a cord cutter.

KARL IS NOT LIKELY
to cut the cord in the near future.

He's something of an outlier. Although he has demographic indicators typical of a cord cutter, the company can feel confident that he isn't a likely flight risk.

Why? Karl's avoidance of streaming services is a factor, but the most important one is his sports channel subscriptions. The nature of athletic programming—it's usually broadcast live and via regional networks—means that it's difficult to replicate online. And the company's statistical models show that sports enthusiasts tend to be cable loyalists.

CATHY IS LIKELY
to cut the cord in the near future.

Although she's older and more well-off than the typical cord cutter, she's relatively tech-savvy. Unlike Karl, Cathy already gets her favorite programming from a noncable source, and there's no compelling reason for her not to cut the cord.

The cable company should include her in its planned promotion to try to keep her as a happy customer.

BUILD A MODEL

To learn, the machine must access rich customer data: demographics, purchasing behavior, other services customers pay for, whether they've called to complain or cancel a service, and much more. Using these variables, data scientists build a statistical model that determines how predictive each variable is in terms of the answer the machine is trying to learn.

ACCUMULATE ANSWERS

For each question, the probability is refined in one direction or the other on the basis of answers up to that point. The machine looks for combinations of attributes that create a high level of certainty about the answer it's seeking.

MAKE A DECISION

Eventually, the probability is heavily weighted one way or the other. That may come after just a few questions or dozens of them. The data scientists decide how confident they want the machine to be in its answer. They may say that once it's 95% confident, it can stop.

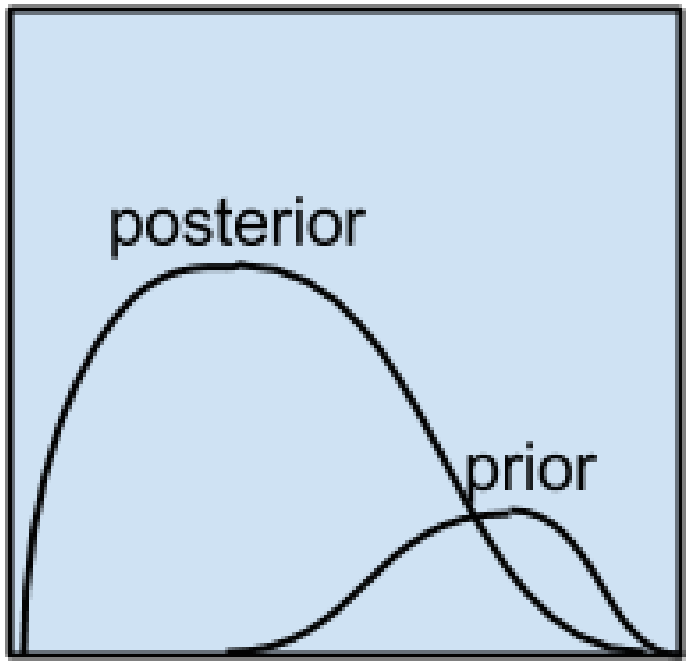
ITERATE

Now the computer can carry out the same evaluation thousands or even millions of times on other entries in its data set, confidently sorting each entry into group A or group B. It has learned.

Bayesian

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

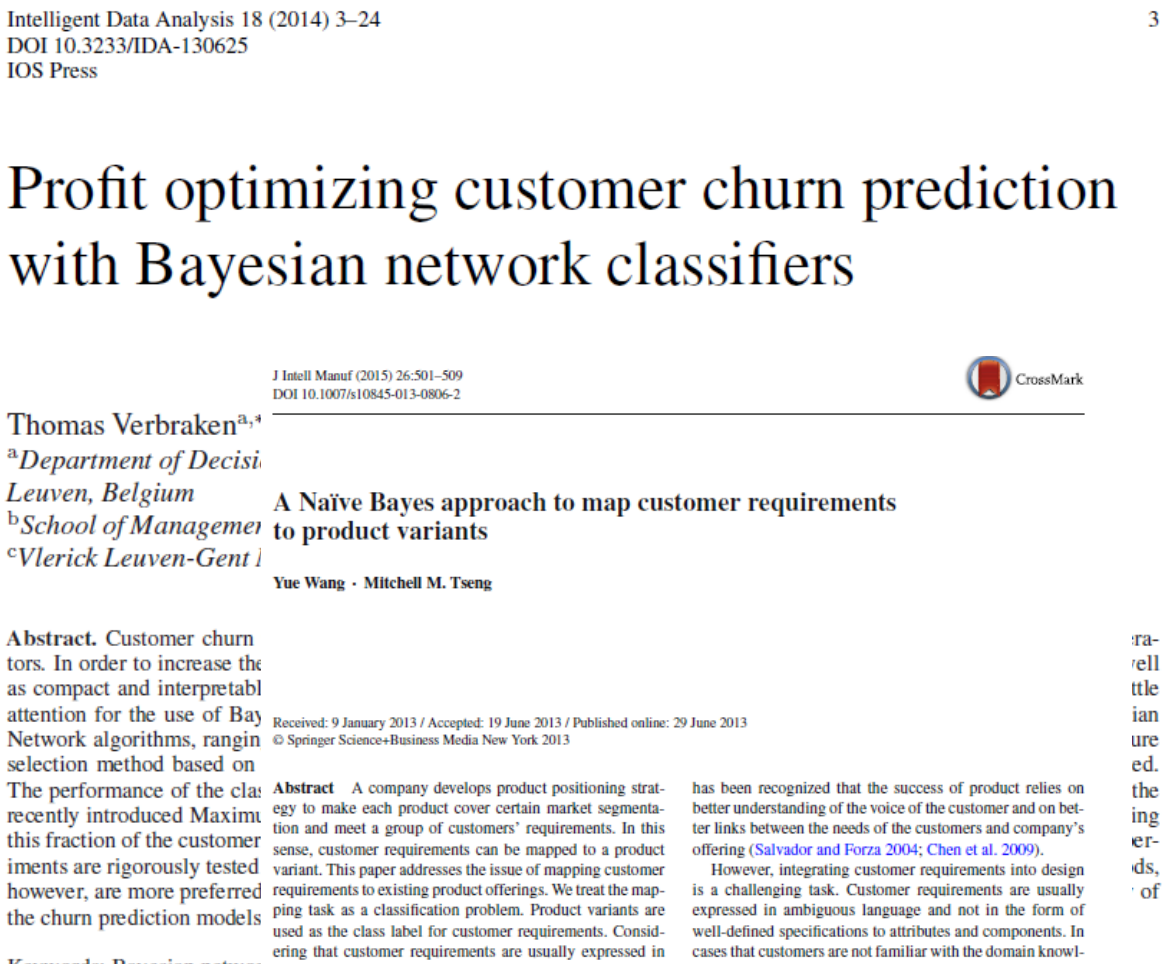
- Most popular Bayesian algorithms:
  - Naive Bayes
  - Gaussian Naive Bayes
  - Multinomial Naive Bayes
  - Averaged One-Dependence Estimators (AODE)
  - Bayesian Belief Network (BBN)
  - Bayesian Network (BN)



Bayesian Algorithms

Bayesian - Examples

- How can you predict customer churn?
- How can you predict whether a customer will purchase a product, based on desired functionality and requirements?
- How can you classify Twitter sentiment analysis?
- How can you predict which email messages are spam, and which are legitimate?
- How can you detect anomalous credit card transactions?



Intelligent Data Analysis 18 (2014) 3–24
DOI 10.3233/IDA-130625
IOS Press

3

Profit optimizing customer churn prediction with Bayesian network classifiers

Thomas Verbraken[a,*]
[a]*Department of Decisi*
Leuven, Belgium
[b]*School of Manageme*
[c]*Vlerick Leuven-Gent*

Abstract. Customer churn
tors. In order to increase the
as compact and interpretabl
attention for the use of Bay
Network algorithms, rangin
selection method based on
The performance of the clas
recently introduced Maximu
this fraction of the customer
iments are rigorously tested
however, are more preferred
the churn prediction models

J Intell Manuf (2015) 26:501–509
DOI 10.1007/s10845-013-0806-2

CrossMark

A Naïve Bayes approach to map customer requirements to product variants

Yue Wang · Mitchell M. Tseng

Received: 9 January 2013 / Accepted: 19 June 2013 / Published online: 29 June 2013
© Springer Science+Business Media New York 2013

Abstract A company develops product positioning strategy to make each product cover certain market segmentation and meet a group of customers' requirements. In this sense, customer requirements can be mapped to a product variant. This paper addresses the issue of mapping customer requirements to existing product offerings. We treat the mapping task as a classification problem. Product variants are used as the class label for customer requirements. Considering that customer requirements are usually expressed in

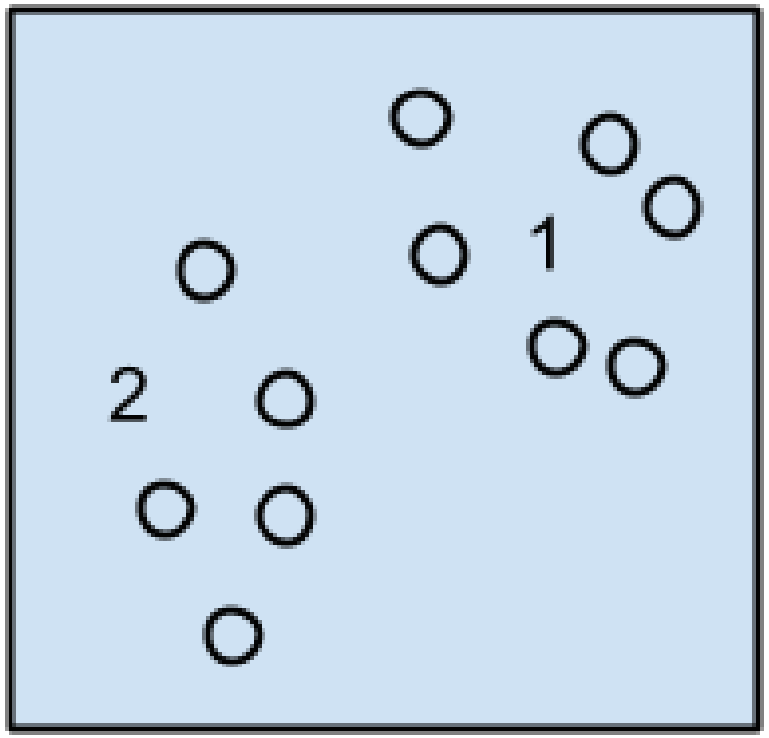
has been recognized that the success of product relies on better understanding of the voice of the customer and on better links between the needs of the customers and company's offering (Salvador and Forza 2004; Chen et al. 2009).

However, integrating customer requirements into design is a challenging task. Customer requirements are usually expressed in ambiguous language and not in the form of well-defined specifications to attributes and components. In cases that customers are not familiar with the domain knowl-

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

Clustering - Examples

- How can you predict the path of a cyclone?
- Given a list of characteristics, how can you determine which customers fall into which segments (jocks, nerds, weirdos, princesses)?
- How can you determine which health insurance claims are fraudulent and which are legitimate?
- Given a customer's viewing history, how can you predict the next film they should watch?

Vol.22 No.2 JOURNAL OF TROPICAL METEOROLOGY June 2016

Article ID: 1006-8775(2016) 02-0127-09

K-MEANS CLUSTERING FOR CLASSIFICATION OF THE NORTHWESTERN PACIFIC TROPICAL CYCLONE TRACKS

YU Jin-hua (余锦华)[1], ZHENG Ying-qing (郑颖青)[1,2], WU Qi-shu (吴启树)[2], LIN Jin-gan (林金淦)[2], GONG Zhen-bin (龚振彬)[2]

Anale. Seria Informatică. Vol. XIV fasc. 1 – 2016 51
Annals. Computer Science Series. 14th Tome 1st Fasc. – 2016

DEVELOPMENT OF IMPROVED K-MEANS CLUSTERING TO PARTITION HEALTH INSURANCE CLAIMS

Stephen G. Fashoto[1], Adekunle Adekoya[2], Jacob A. Gbadeyan[3], J. S. Sadiku[3], W.B. Yahya[3]

[1]Kampala International University, Kampala, Uganda
[2]Redeemer's University, Ede Osun State, Nigeria
[3]University of Ilorin, Ilorin, Nigeria

Corresponding author: Stephen Gbenga Fashoto, Stephen.gbenga@kiu.ac.ug

ABSTRACT: Healthcare insurance, delivered via the National Health Insurance Scheme (NHIS) is a veritable tool for making quality healthcare available to the majority of Nigerian citizens, irrespective of income status. Segmentation of dataset comprising of several features is a drawback of many applications. K-

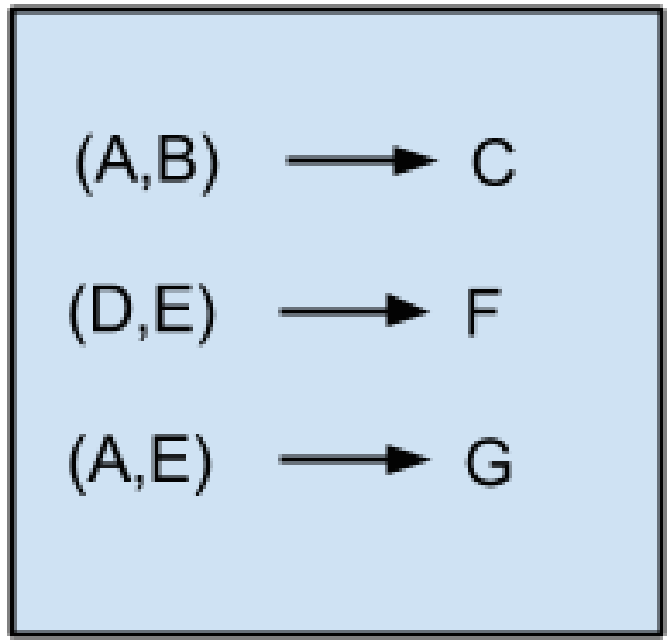
healthcare services or products without the need to pay for such services and products at the point of care [Car11].

The health insurance industry has historically been a growing industry. The role of the industry in the

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

Association Rule Learning - Examples

- How can you predict if a customer will buy a product, given their purchase history?
- How can you forecast which startups will be successful?
- How can you forecast whether a company will buy or sell stocks or commodities?
- Which courses should a student take, based on their transcript history?

3. Application of an improved **Apriori algorithm** in a mobile e-commerce recommendation system.

By: Guo, Yan; Wang, Minxi; Li, Xin. Industrial Management & Data Systems. 2017, Vol. 117 Issue 2, p287-303. 17p. DOI: 10.1108/IMDS-03-2016-0094. , Database: Business Source Complete

Subjects: Electronic commerce; Data mining; Recommender systems (Information filtering); Taobao.com Inc.; Electronic shopping and mail-order houses; Electronic Shopping; **Apriori algorithm**

Academic Journal

Full Text Finder

ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Model using Improved Apriori Algorithm to generate Association Rules for Future Contracts of Multi Commodity Exchange (MCX)

Chirag A. Mewada
Assistant Professor
Naran Lala College of Professional and Applied Sciences
Navsari

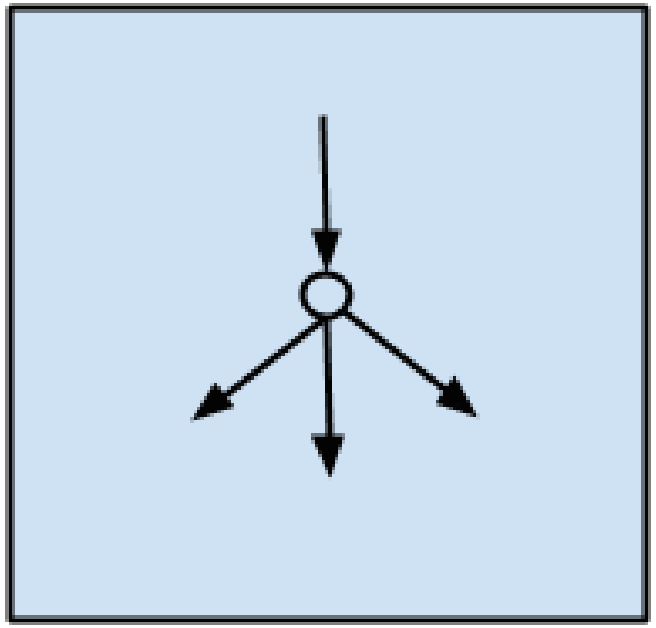
Rustom D. Morena
Professor
Department of Computer Science
Veer Narmad South Gujarat University, Surat

Abstract: Now-a-days commodity markets are getting equal importance as equity market. Moreover, just like stock trend prediction, researchers are taking active interest in forecasting commodity trend. Trend prediction is really a challenging task especially for commodity market. There are so many factors, which affects commodity market returns. This empirical study generates association rules form the data of Multi Commodity Exchange (MCX) of India from November 2003 to December 2016. Data contains End of Day (EOD) prices of each commodity future contract. We have developed a model to analyze MCX data and derive association rules. Using our model, we have found co-movement between fundamentally different commodity future contracts. In our model, we have used improved apriori algorithm. Our result portrays that the contracts of fundamentally different commodities are correlated even if they neither substitutes nor complements each other.

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.

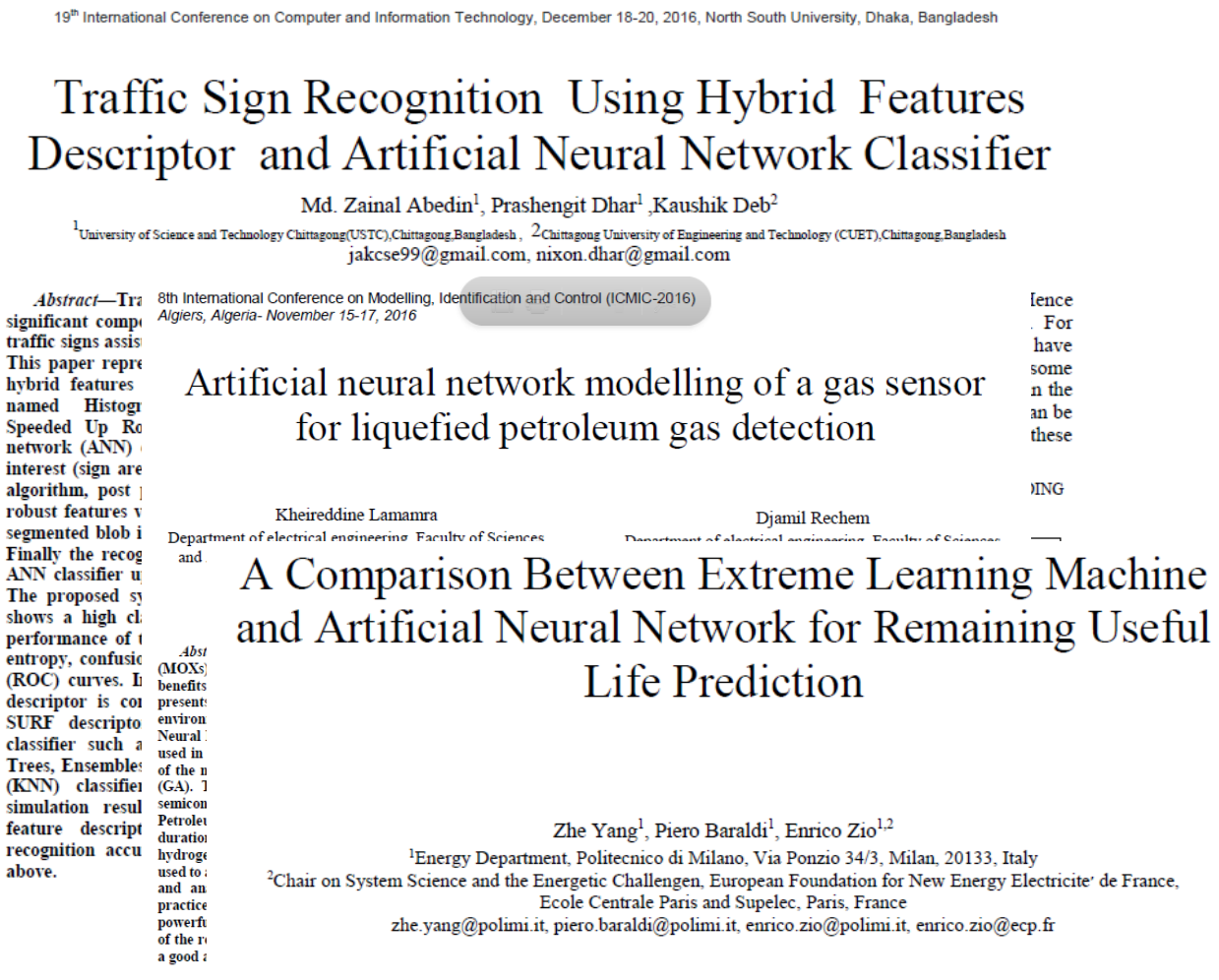
- **Most popular artificial neural network algorithms:**
  - Perceptron
  - Back-Propagation
  - Hopfield Network
  - Radial Basis Functional Network



Artificial Neural Network Algorithms

Artificial Neural Network (ANN) - Examples

- Early detection of lung cancer based on images
- How long can we anticipate satellites to last in space, or equipment in an industrial setting?
- How can we detect traffic signs and determine their meanings?
- How can we use image sensor data to detect hazardous gases?



19th International Conference on Computer and Information Technology, December 18-20, 2016, North South University, Dhaka, Bangladesh

Traffic Sign Recognition Using Hybrid Features Descriptor and Artificial Neural Network Classifier

Md. Zainal Abedin[1], Prashengit Dhar[1] ,Kaushik Deb[2]

[1]University of Science and Technology Chittagong(USTC),Chittagong,Bangladesh , [2]Chittagong University of Engineering and Technology (CUET),Chittagong,Bangladesh

jakcse99@gmail.com, nixon.dhar@gmail.com

8th International Conference on Modelling, Identification and Control (ICMIC-2016)

Algiers, Algeria- November 15-17, 2016

Artificial neural network modelling of a gas sensor for liquefied petroleum gas detection

Kheireddine Lamamra

Djamil Rechem

A Comparison Between Extreme Learning Machine and Artificial Neural Network for Remaining Useful Life Prediction

Zhe Yang[1], Piero Baraldi[1], Enrico Zio[1,2]

[1]Energy Department, Politecnico di Milano, Via Ponzio 34/3, Milan, 20133, Italy

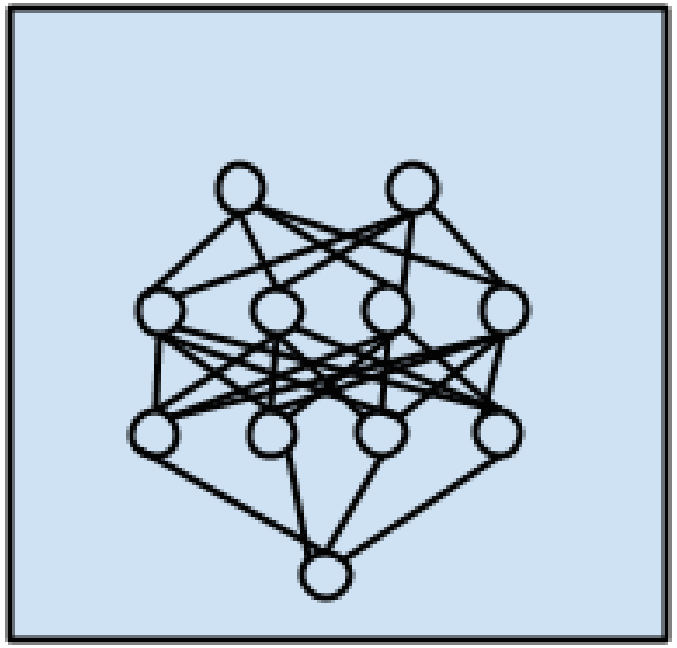
[2]Chair on System Science and the Energetic Challengen, European Foundation for New Energy Electricite' de France, Ecole Centrale Paris and Supelec, Paris, France

zhe.yang@polimi.it, piero.baraldi@polimi.it, enrico.zio@polimi.it, enrico.zio@ecp.fr

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.

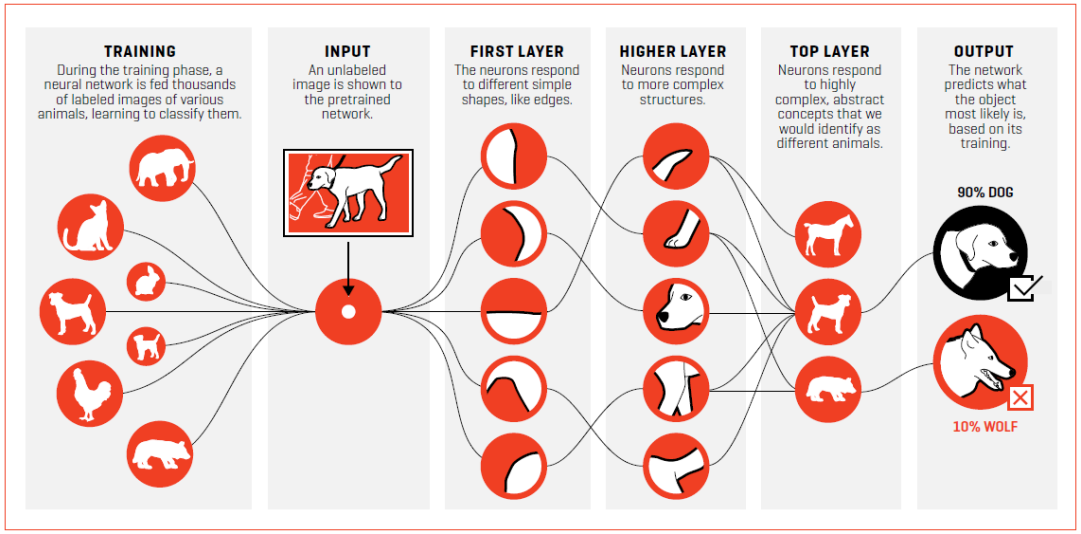
- **Most popular deep learning algorithms:**
  - Deep Boltzmann Machine (DBM)
  - Deep Belief Networks (DBN)
  - Convolutional Neural Network (CNN)
  - Stacked Auto-Encoders



Deep Learning Algorithms

Deep Learning - Examples

- Given a large corpus of images, how can you determine which are wolves and which are dogs?
- How can self-driving cars detect (and hopefully avoid!) pedestrians?
- Speech recognition, translation, photo searches, self-driving cars and self-flying drones…



GOOGLE

GOOGLE LAUNCHED the deep-learning-focused Google Brain project in 2011, introduced neural nets into its speech-recognition products in mid-2012, and retained neural nets pioneer Geoffrey Hinton in March 2013. It now has more than 1,000 deep-learning projects underway, it says, extending across search, Android, Gmail, photo, maps, translate, YouTube, and self-driving cars. In 2014 it bought DeepMind, whose deep reinforcement learning project, **AlphaGo, defeated the world's go champion, Lee Sedol (above)**, in March, achieving an artificial intelligence landmark.

MICROSOFT

MICROSOFT INTRODUCED deep learning into its commercial speech-recognition products, including Bing voice search and X-Box voice commands, during the first half of 2011. The company now uses neural nets for its search rankings, photo search, translation systems, and more. "It's hard to convey the pervasive impact this has had," says Lee. Last year it won the key image-recognition contest, and in September it scored a record low error rate on a speech-recognition benchmark: 6.3%.

FACEBOOK

IN DECEMBER 2013, Facebook hired French neural nets innovator Yann LeCun to direct its new AI research lab. Facebook uses neural nets to translate about 2 billion user posts per day in more than 40 languages, and says its translations are seen by 800 million users a day. (About half its community does not speak English.) Facebook also uses neural nets for photo search and photo organization, and it's working on a feature that would generate spoken captions for untagged photos that could be used by the visually impaired.

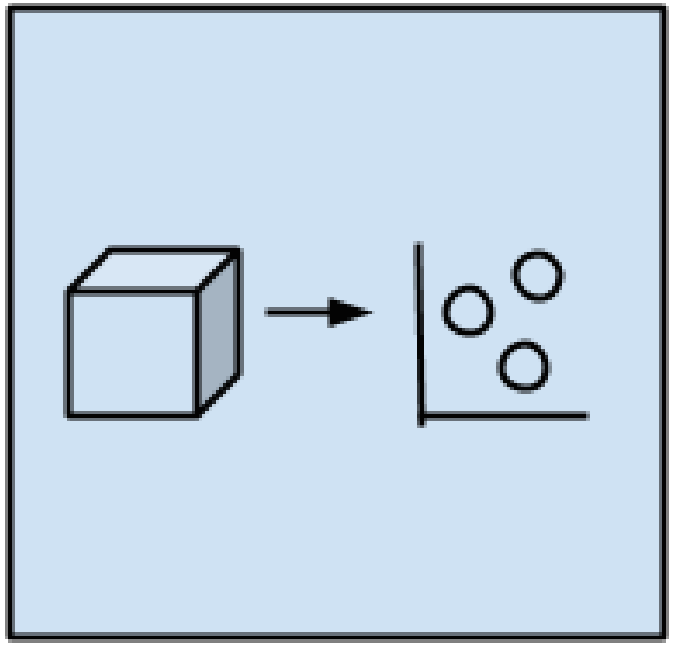
BAIDU

IN MAY 2014, Baidu hired Andrew Ng, who had earlier helped launch and lead the Google Brain project, to lead its research lab. China's leading search and web services site, Baidu uses neural nets for speech recognition, translation, photo search, and a self-driving car project, among others. Speech recognition is key in China, a mobile-first society whose main language, Mandarin, is difficult to type into a device. The number of customers interfacing by speech has tripled in the past 18 months, Baidu says.

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)

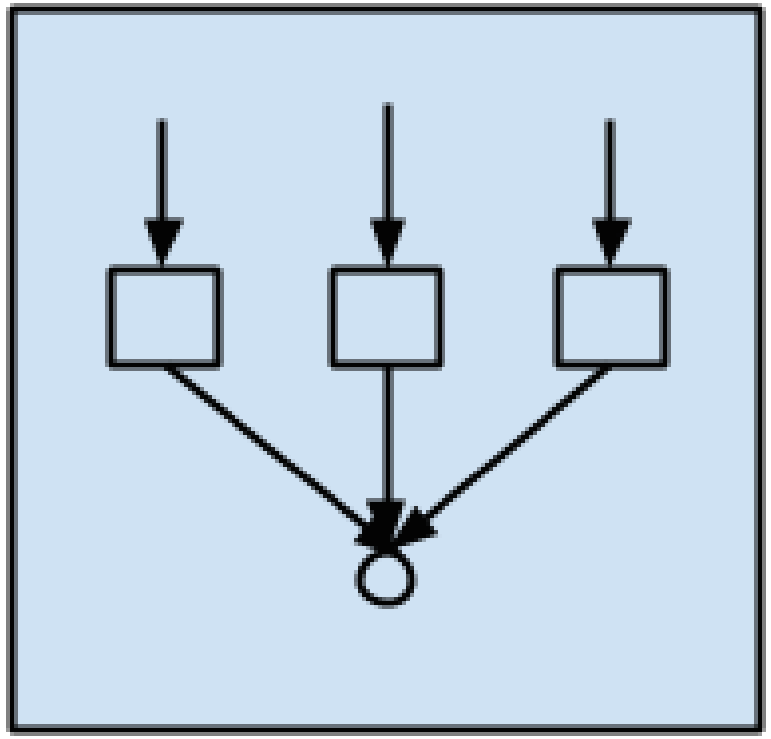


Dimensional Reduction Algorithms

Ensemble Methods

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

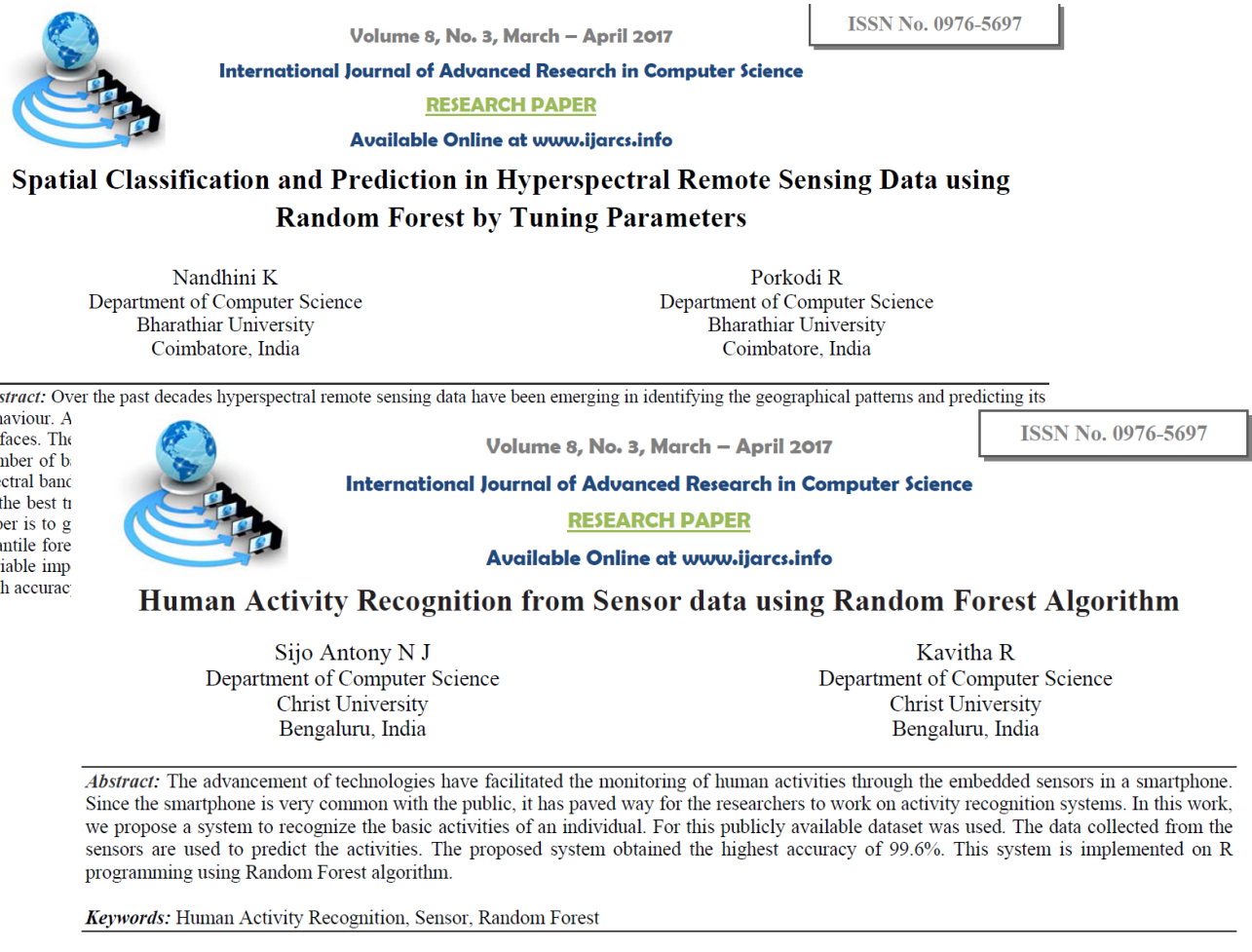
- **Most popular ensemble algorithms:**
  - Boosting
  - Bootstrapped Aggregation (Bagging)
  - AdaBoost
  - Stacked Generalization (blending)
  - Gradient Boosting Machines (GBM)
  - Gradient Boosted Regression Trees (GBRT)
  - Random Forest



Ensemble Algorithms

Ensemble Methods - Examples

- In remote sensing (satellite imagery) how can you detect which pixels are related to forests, which are related to streets, which are related to buildings?
- Given embedded sensor data in a smart phone or a Fitbit, how can you predict the activity of an end user?
- Given completions data and production of a well, how can you optimize parameters of interest (proppants, fluids, numbers of stages and cluster spacing, etc.)?



ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Spatial Classification and Prediction in Hyperspectral Remote Sensing Data using Random Forest by Tuning Parameters

Nandhini K
Department of Computer Science
Bharathiar University
Coimbatore, India

Porkodi R
Department of Computer Science
Bharathiar University
Coimbatore, India

Abstract: Over the past decades hyperspectral remote sensing data have been emerging in identifying the geographical patterns and predicting its behaviour. A
surfaces. The
number of b
spectral banc
of the best t
paper is to g
quantile fore
variable imp
high accurac

ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Human Activity Recognition from Sensor data using Random Forest Algorithm

Sijo Antony N J
Department of Computer Science
Christ University
Bengaluru, India

Kavitha R
Department of Computer Science
Christ University
Bengaluru, India

Abstract: The advancement of technologies have facilitated the monitoring of human activities through the embedded sensors in a smartphone. Since the smartphone is very common with the public, it has paved way for the researchers to work on activity recognition systems. In this work, we propose a system to recognize the basic activities of an individual. For this publicly available dataset was used. The data collected from the sensors are used to predict the activities. The proposed system obtained the highest accuracy of 99.6%. This system is implemented on R programming using Random Forest algorithm.

Keywords: Human Activity Recognition, Sensor, Random Forest

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

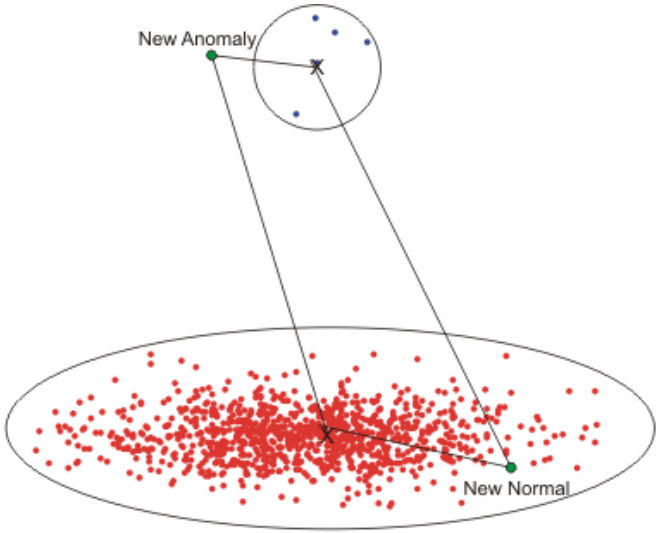
- **Examples**
  - Identifying transactions that are potentially fraudulent
  - Learning patterns that indicate a network intrusion has occurred
  - Finding abnormal clusters of patients
  - Checking values input to a system

- **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

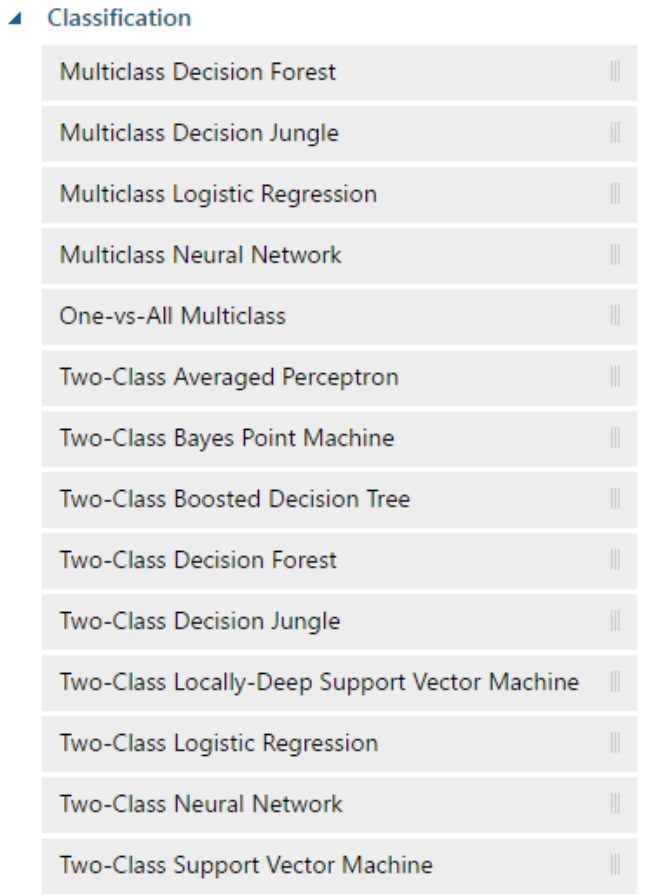
PCA-Based Anomaly Detection

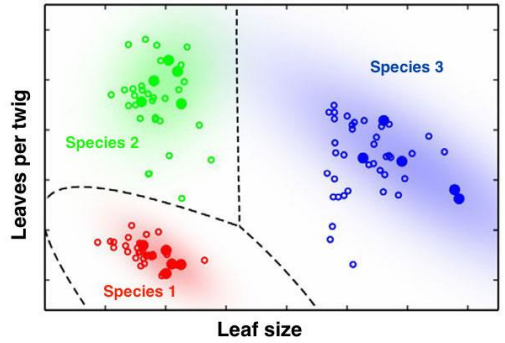


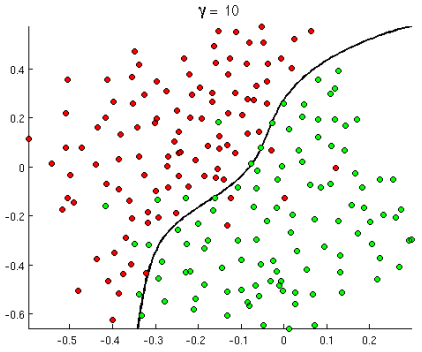
Classification

- 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.



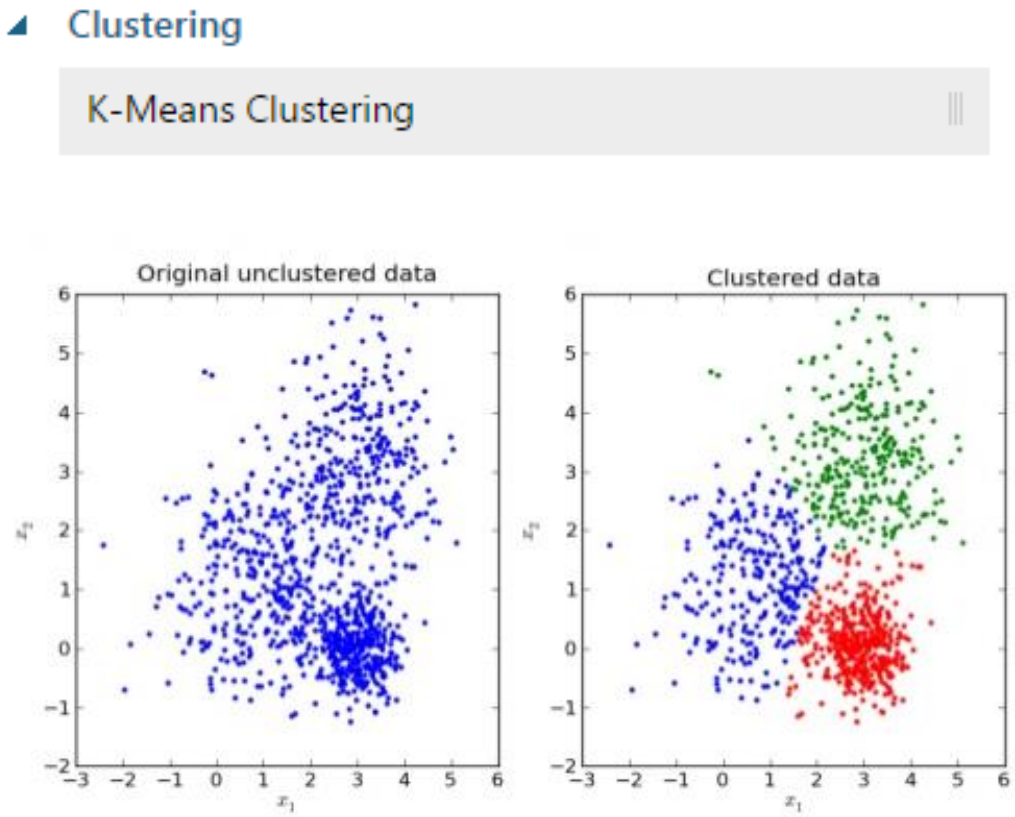




Clustering

- 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.



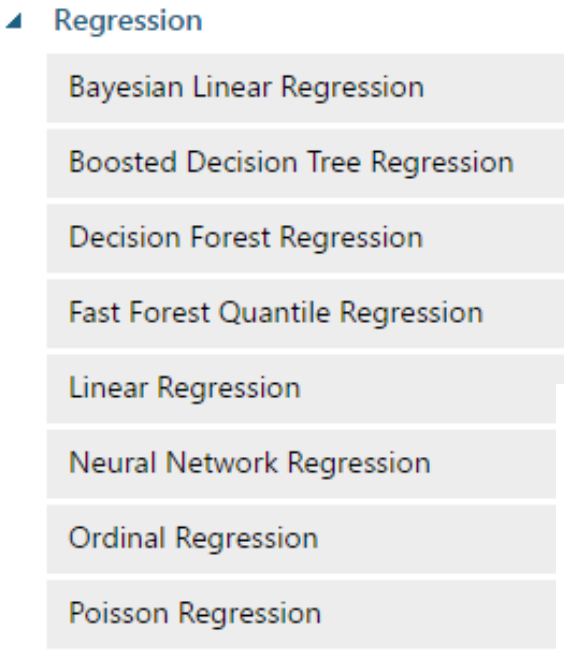
http://pypr.sourceforge.net/kmeans.html

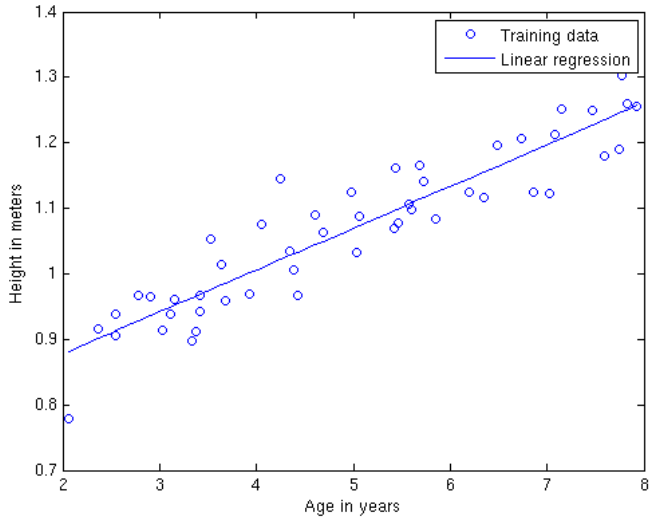
Regression

- 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 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

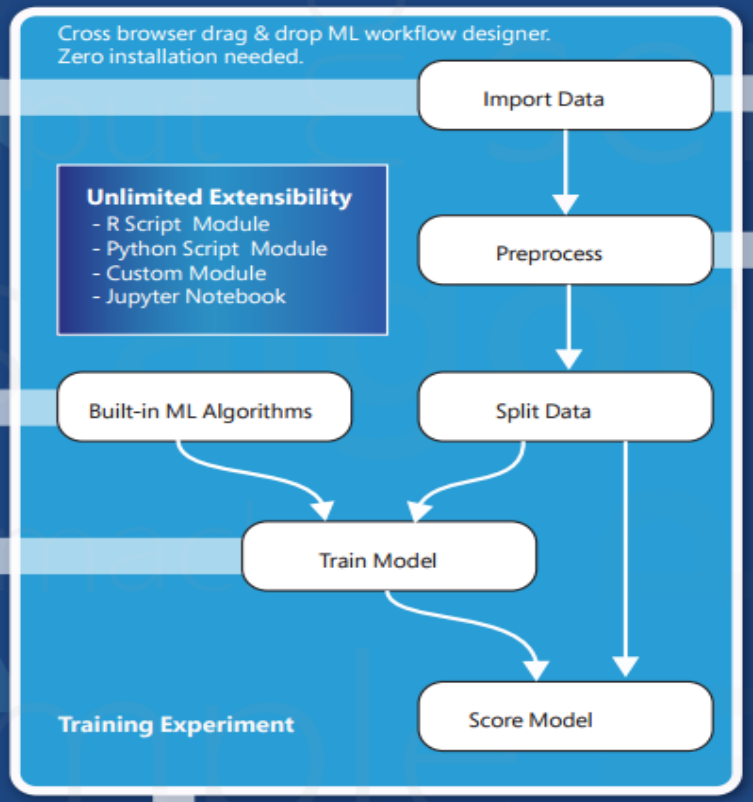
Computer Vision
- OpenCV Library

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

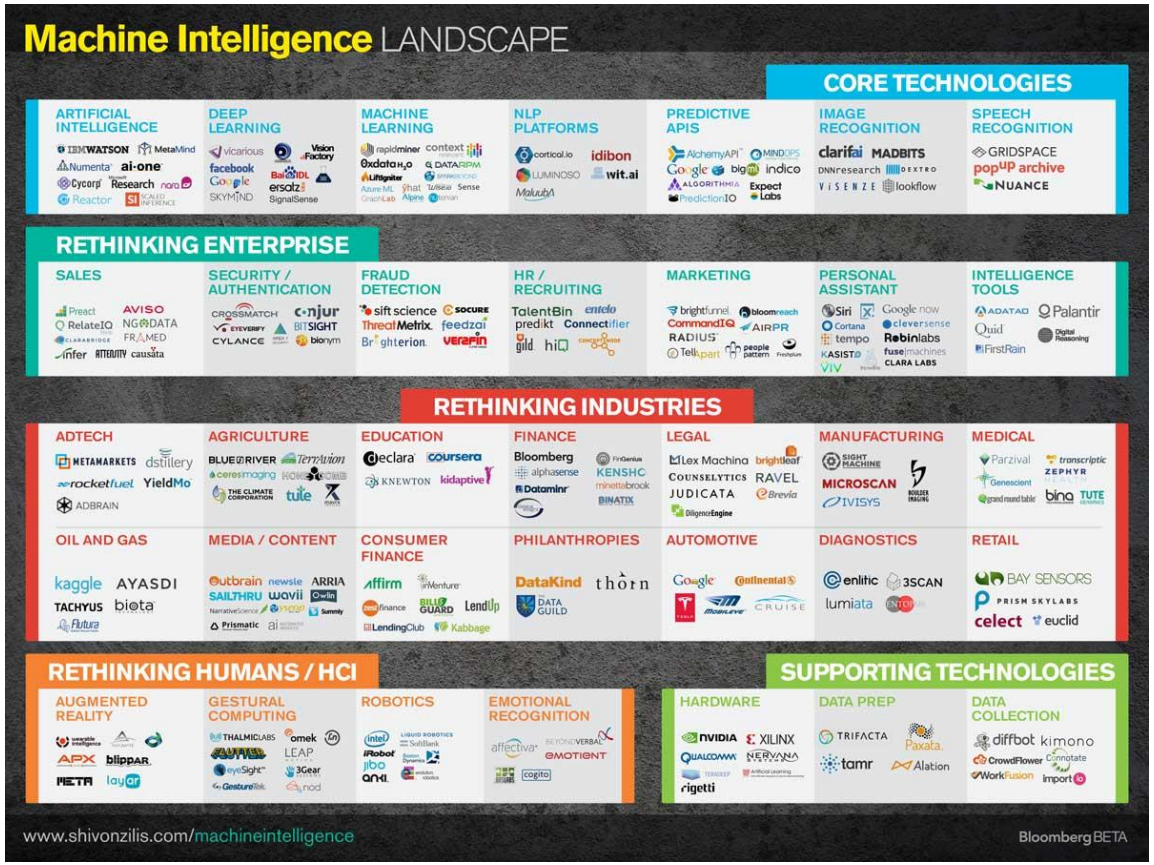
Azure Machine Learning Studio Capabilities Overview

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

Microsoft

Machine Learning – Industry Impacts

- **Services:** get better at forecasting demand, and learn how to provide the right product on a hyper-individualized basis (the Netflix approach).
- **Retail**: more sophisticated supply chains, deeper understanding of consumer preferences, ability to customize products and purchase experiences online and offline. Trend creation / brand building.
- **Manufacturing**: real-time, complete system monitoring (anomaly detection); issues mitigated before they occur; asset lifecycle optimization; reducing the need for human intervention.
- **Agriculture**: determine which crops to grow, in what quantities, and in what locations.



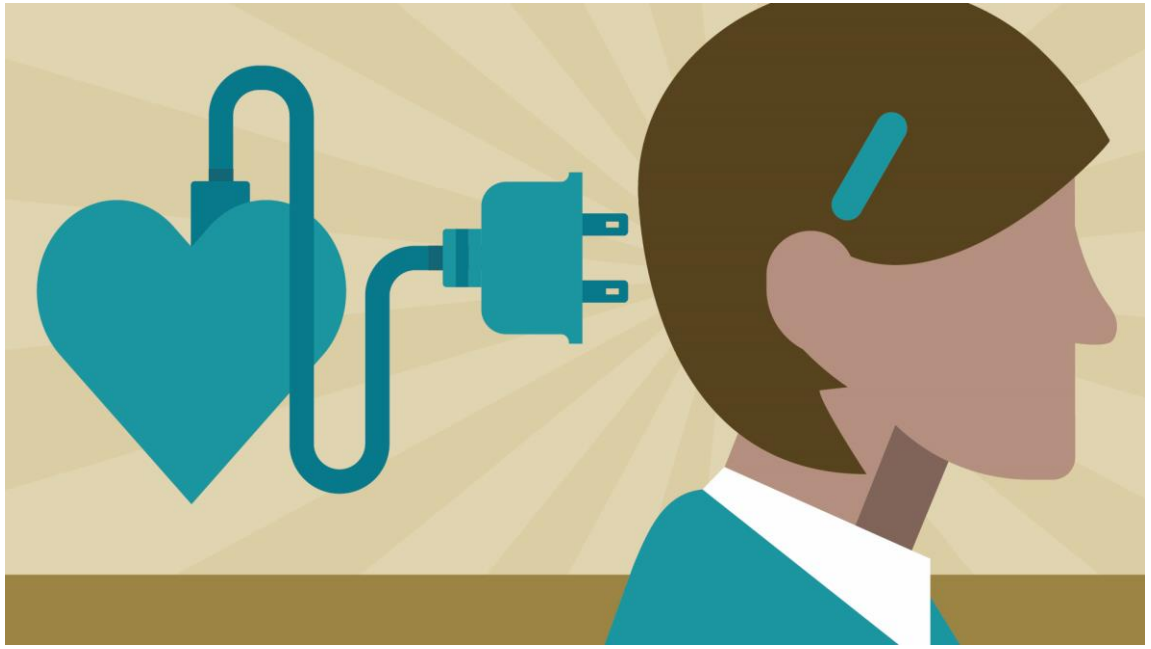
Social Impact - Ethical Algorithms

- Which candidates will be good workers?
- How should college students decide which courses they're prepared to tackle?
- Can you forecast which families will wind up on the streets (reactive rather than preventative)?
- Can you predict whether a defendant will skip court, if they're released on bail? Or if they'll commit further crimes?
- **How biased are these forecasts?**

How to stay relevant?

- Many skilled jobs follow the same general workflow.
- **Constant, continuing education.**

- Focus on skills and capabilities that artificial intelligence has trouble replicating.
  - **Examples**: understanding, motivating, interacting with human beings.
  - A smart machine may be able to diagnose an illness and even recommend treatment better than a doctor, but **it takes a person to sit with the patient, understand their life situation (finances, family, quality of life) and determine what treatment plan is optimal.**



Case Study

Predictive Analytics – Houston 311 Data

Houston 311 Data

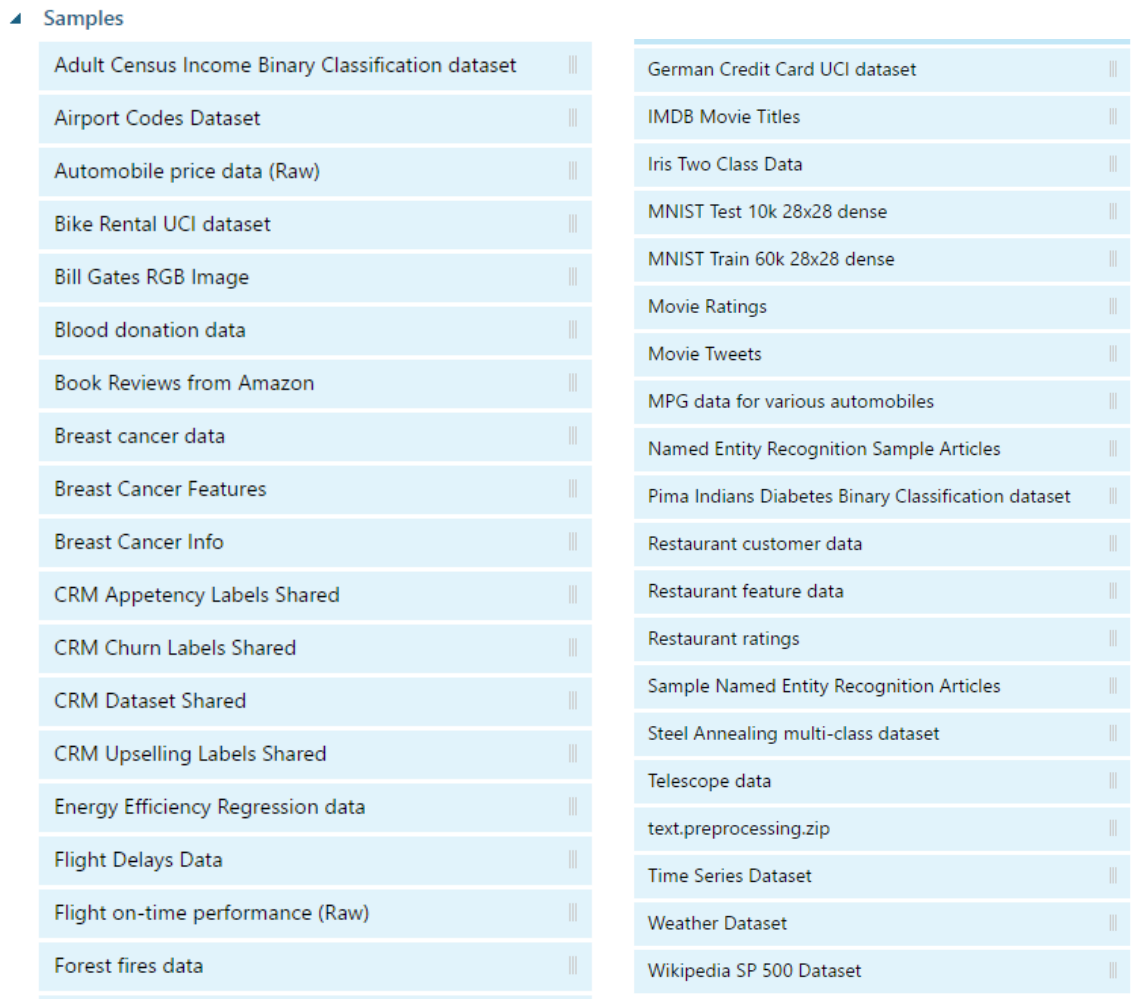
- **Houston 311** is the City of Houston's non-emergency service to report troubles and to get answers on city services.
- 311 files in the Houston data portal are updated nightly, and you can explore here: http://data.ohouston.org/dataset/city-of-houston-311-service-requests
- The data set we'll be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

5-step process for ML

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

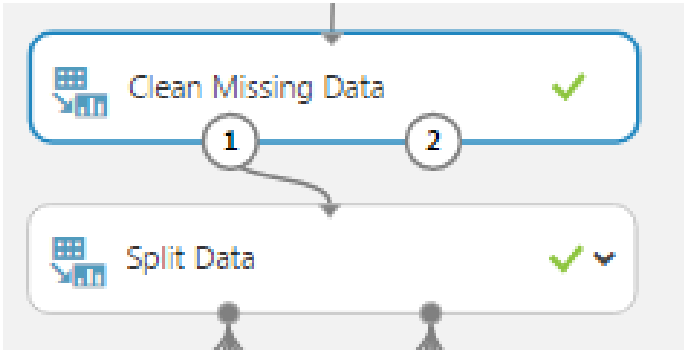
Step 1: Get the data

- 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/

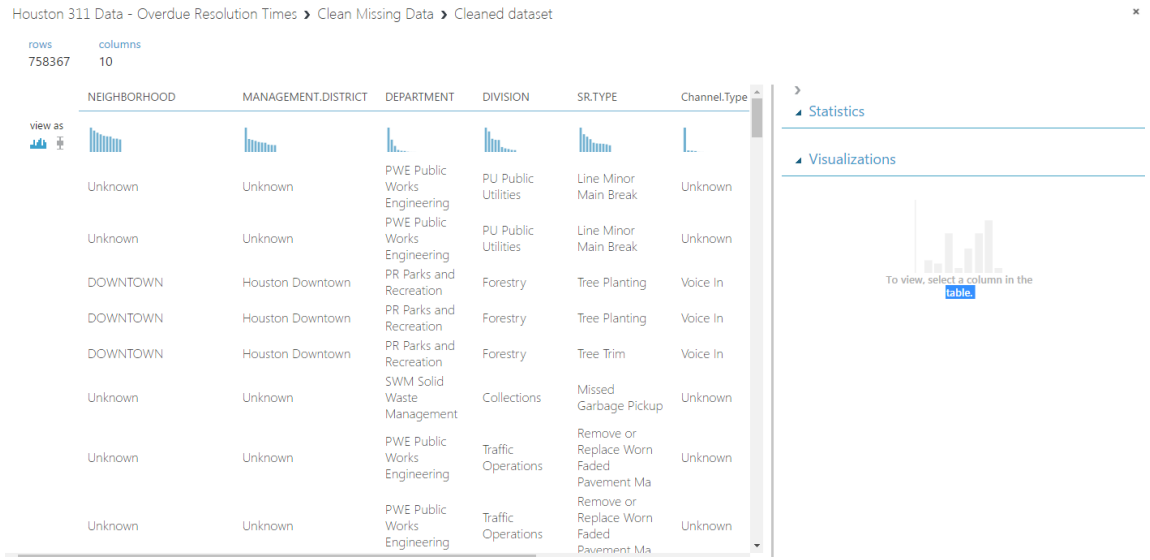


Step 2: Prepare the data

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



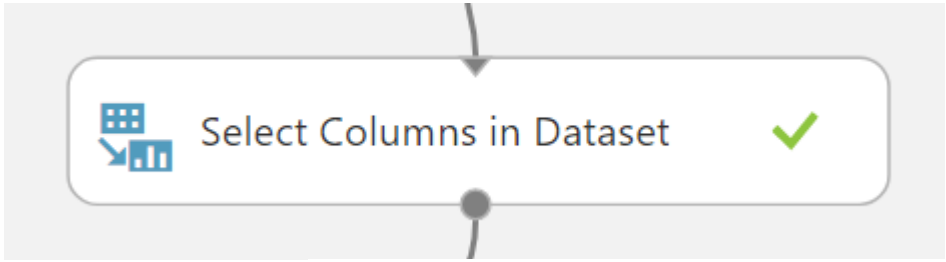
- Examples:
  - Missing values (NA, -999.25, blanks, etc.)
  - Nonsensical values (10000 instead of 1000, etc.)
  - Values that have been mistakenly imported as a string when they should have been considered numeric



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

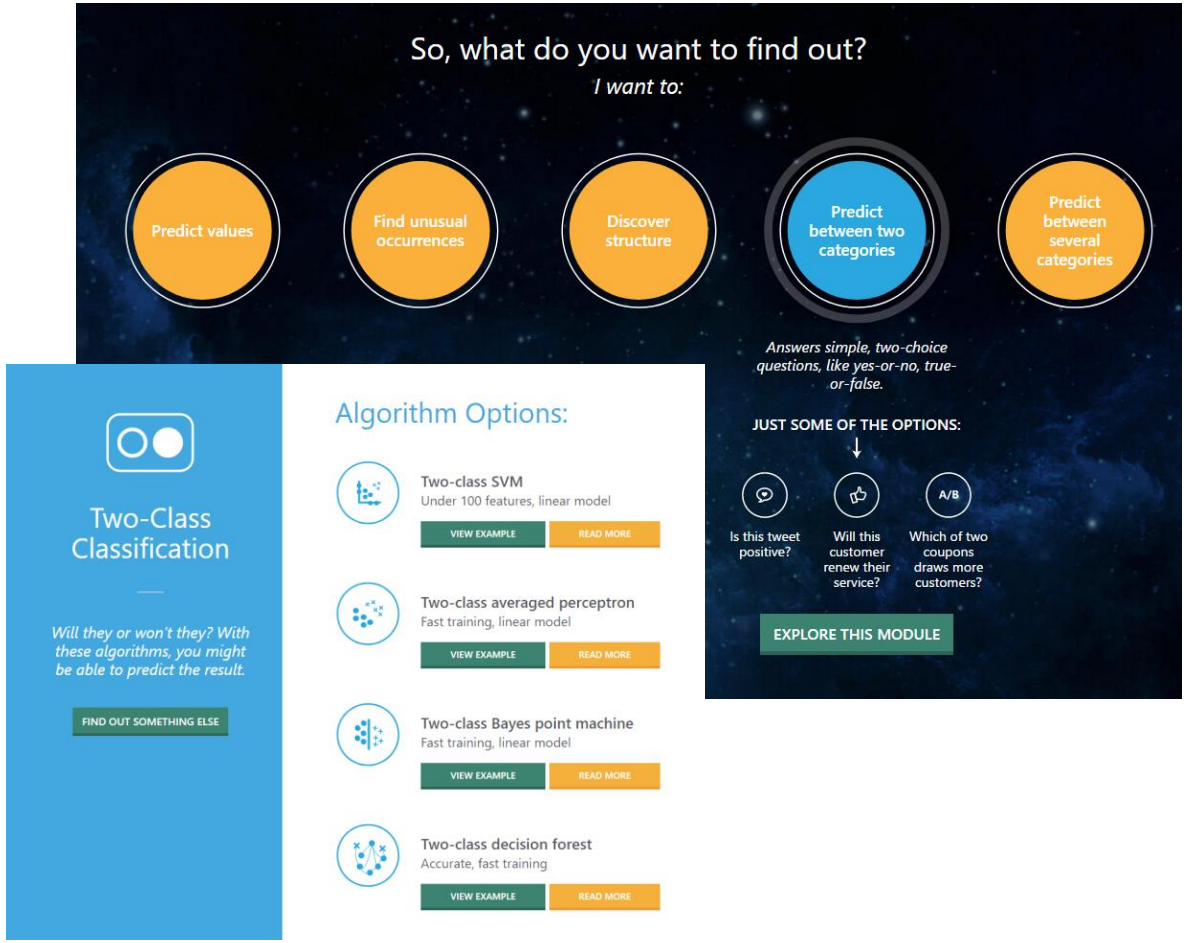
rows 1679336 columns 10

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

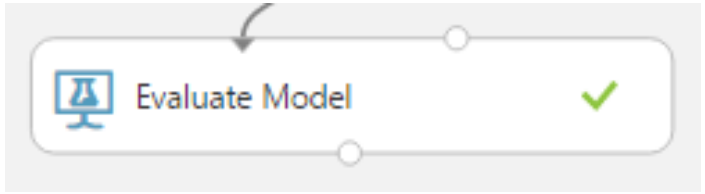
- Constructing a predictive model consists of training and testing. We'll use our data to train the model, and then we'll test the model to see how closely it's able to predict whether a given issue will be completed on time.
- Is this a **classification** task or a **regression** task?
- Help for choosing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

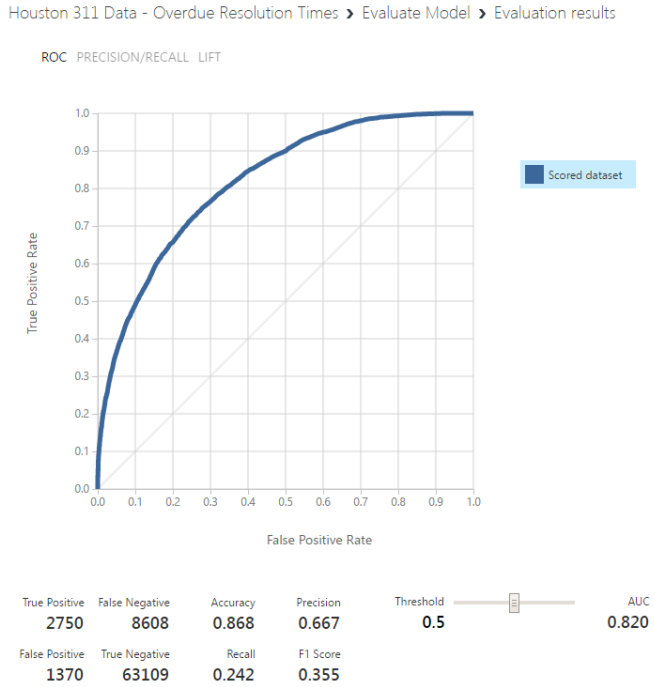


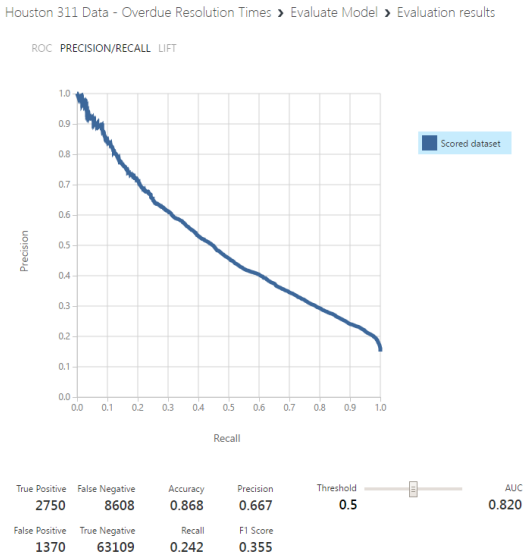
Step 5: Determine predictive power

- 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-answer

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

Thank you!

Q&A