Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Engineering, Test & Technology
Boeing Research & Technology

Machine Learning & Non-Parametric Methods for Cost Analysis

Karen Mourikas, Nile Hanov, Joe King, Denise Nelson

ICEAA Workshop, June 2018

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Machine Learning Approach to Cost Analysis

Machine Learning in General

ML* Algorithms for Cost Analysis

ML Applications related to Cost

- Random Forest Prediction
- Latent Semantic Analysis

Challenges

*** ML = Machine Learning**

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Machine Learning Buzz Words

- Big Data
- Smart Manufacturing
- Deep Learning
- Predictive Analytics
- Neural Networks
- Autoencoders
- NLP (Natural Language Processing)
- IOT (Internet of Things)
- Feature Extraction

Machine Learning Vocabulary

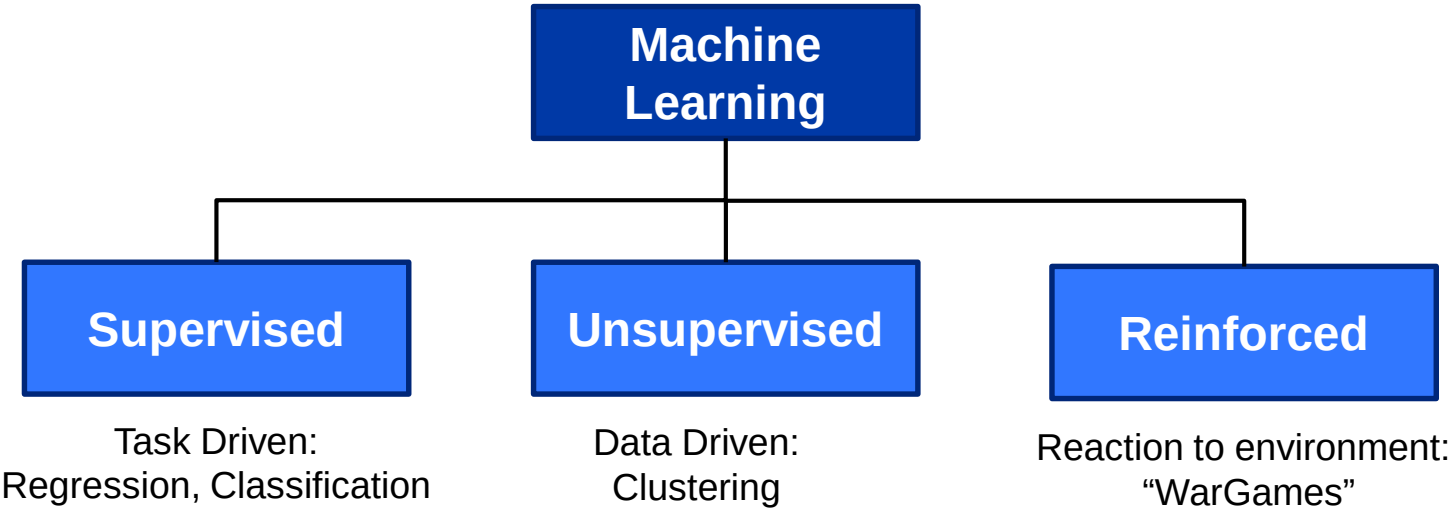
Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

What is Machine Learning?

Simply, when a machine mimics "cognitive" functions such as "learning" and "problem solving" *

Machine Learning (ML) is a method in which algorithms …

- teach themselves to grow (i.e. learn) from data
- learn without being explicitly programmed



Machine Learning is a type of Artificial Intelligence

* Russell, Stuart J.; Norvig, Peter ; Artificial Intelligence: A Modern Approach, 2003 & 2009

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

What can Machine Learning do?

- Speech recognition
- Autonomous scheduling
- Financial forecasting
- Spam filtering
- Logistics planning
- VLSI layout
- Automatic assembly
- Information extraction
- Market Share Analysis
- Route finding
- Robotics
  - household, surgery, navigation
- Failure prediction
- Fraud detection
- Web search engines
- Autonomous cars
- Energy optimization
- Question answering systems
- Social network analysis
- Medical diagnosis, imaging
- Document summarization

Many applications for Machine Learning

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Why is Machine Learning so popular now?

Machine Learning has been around for a long time

- Has become more popular recently

Data Explosion

- Much more data available for complex analyses

Machine Power

- Moore's Law: faster and cheaper computers

Accuracy of Algorithms

- Reliable enough for usable products

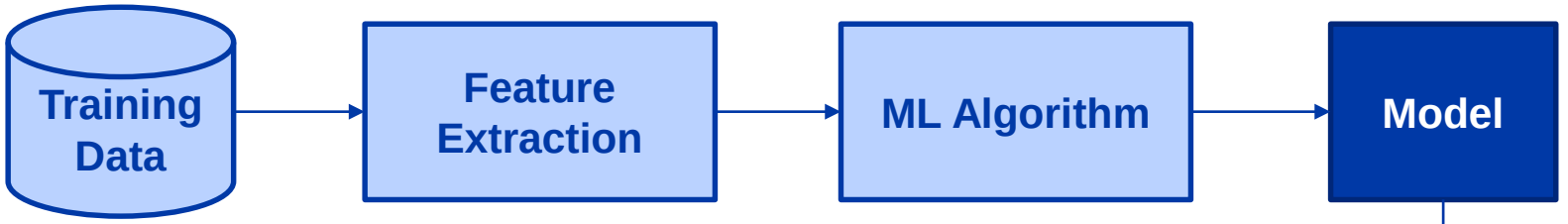
The Future is Here

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

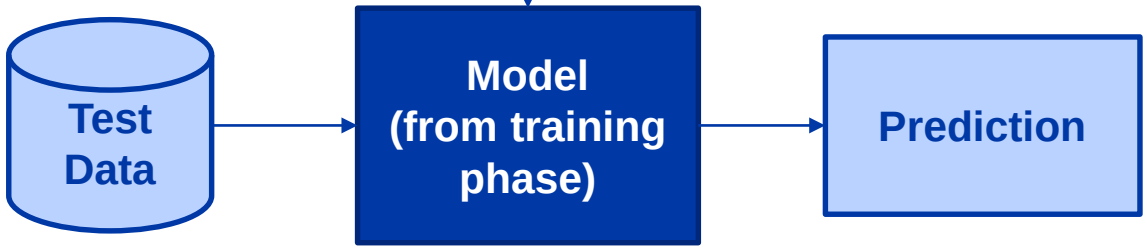
How does Machine Learning Work?

Typically consists of two stages

- Training phase



- Testing Phase



General Process

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Machine Learning Approach to Cost Analysis

Machine Learning in General

ML* Algorithms for Cost Analysis

ML Applications related to Cost

- Random Forest Prediction
- Latent Semantic Analysis

Challenges

*** ML = Machine Learning**

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Machine Learning for Cost Prediction & Analysis

Typical Cost Prediction Methods

- Analogies
- Engineering / Bottoms up
- Parametric Equations / Top down

Machine Learning

- Alternative to traditional cost estimating
- Age of Big Data & Messy Data
- Interactions and non-linear behavior
- Relationship not well understood nor apparent
- Relatively quick & easy to implement

Could we use Machine Learning techniques for cost prediction?

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

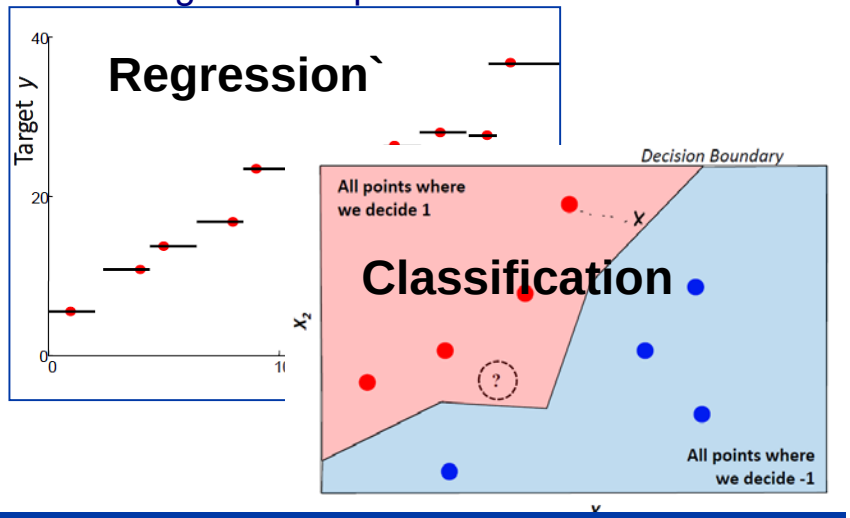
Supervised Algorithms

K-Nearest-Neighbors (KNN)

- Clustering approach
- Given new features, finds nearest example and return its value

Key features

- Regression and Classification



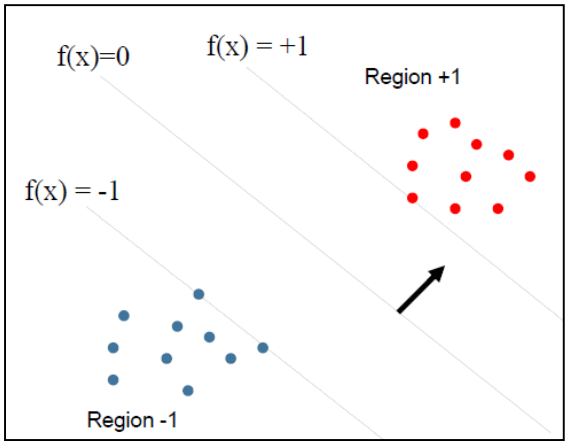
Fast Classification, Similarity Detection

Support Vector Machines (SVM)

- Clustering approach
- Finds the widest margin between classes (boundary decisions)

Key features

- Able to separate non-linearly- separable regions

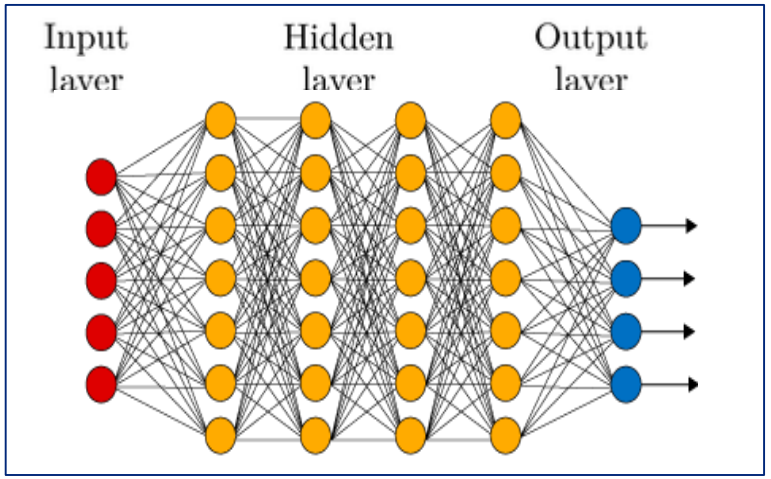


Able to find Optimal Solutions

Supervised Algorithms

Neural Networks (NN)

- Multi-layer perceptron model
- Finds weights for inputs that optimize the cost function



Key features

- Very complex shapes/decision boundaries
- Needs a lot of data

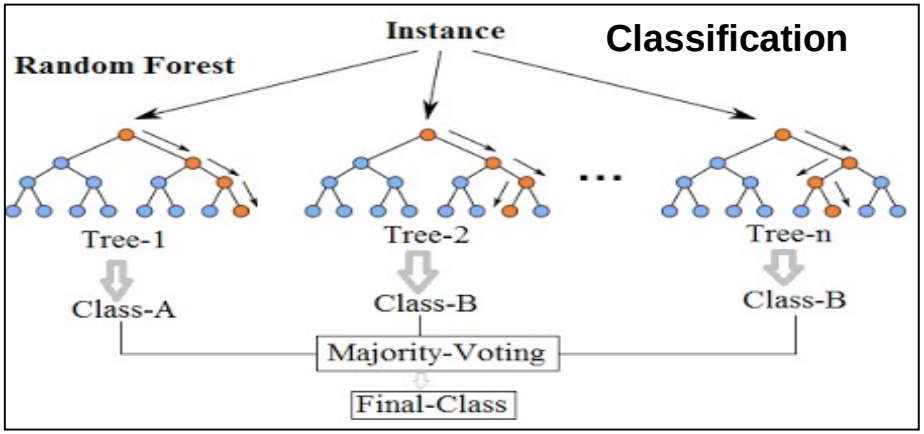
Finds patterns in large amounts of data

Random Forest Prediction

- Decision Tree Ensemble
- Each tree is built from a sample (random) set of features

Key features

- Training set can be small
- Regression & Classification



Handles small n, large p problems

Unsupervised Algorithms

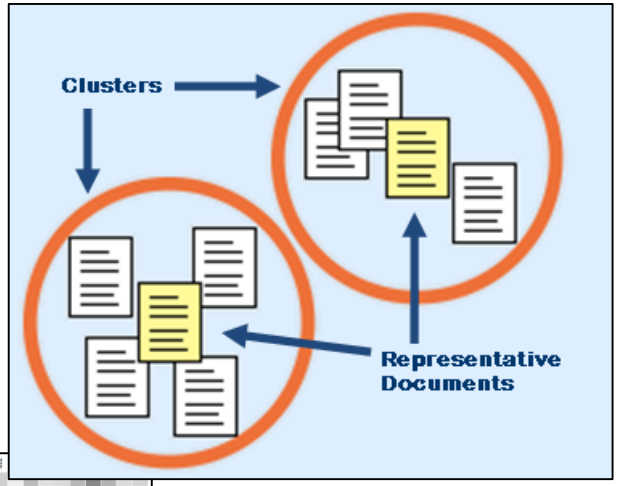
Natural Language Processing -

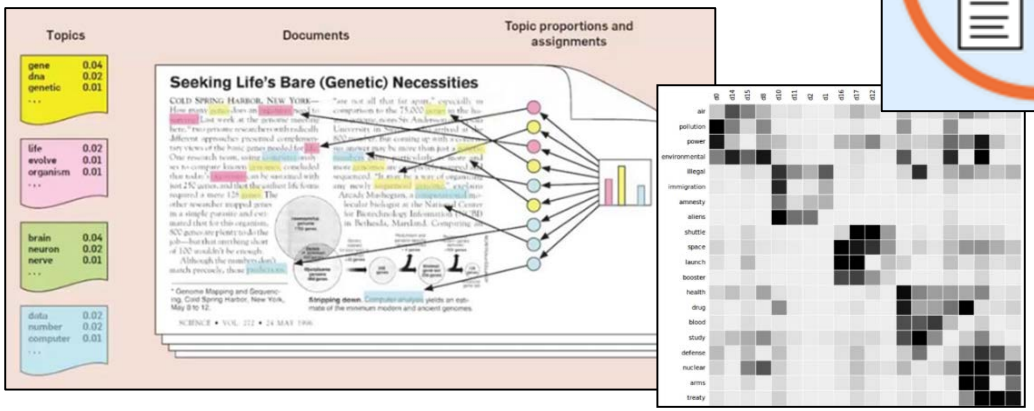
Latent Semantic Analysis (LSA) / Latent Dirichlet Allocation (LDA)

- Document Clustering
- Information retrieval in document groups

Key features

- Automatic topic detection
- Key term discovery
- Word Clustering





Automatic Document Grouping

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Machine Learning Approach to Cost Analysis

Machine Learning in General

ML Algorithms for Cost Analysis

ML Applications related to Cost

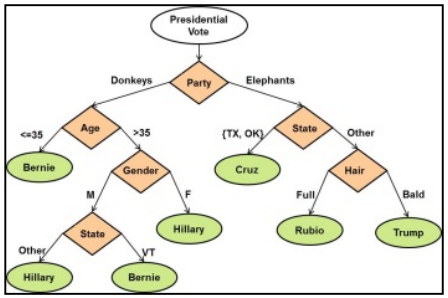
- **Random Forest Prediction**
- Latent Semantic Analysis

Challenges

*** ML = Machine Learning**

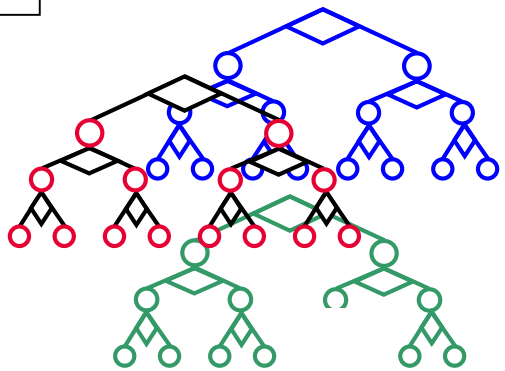
Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Trees and Forests



A Single Decision Tree

- Represents a set of decisions
- Easily interpretable, but …
- Not a great predictor



An Ensemble of trees

- Many trees (100s)
- Not as easy to interpret, but …
- Provides greater prediction accuracy & more stability

Random Forests

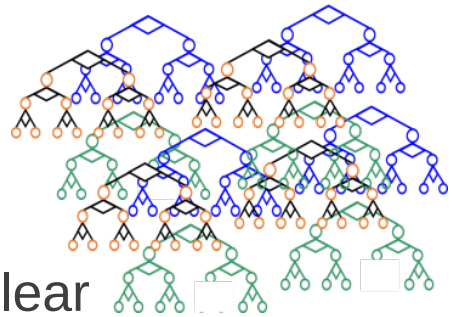
- Ensemble of decision trees “randomly” constructed
- More accurate predictions and reduced error

Source: Alexas_Fotos/Pixabay

Random Forests Prediction based on Decision Tree Theory

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Why use Random Forest Prediction?



Advantages

- Excellent predictors
- Useful if relationship between inputs and outputs is unclear
- Captures non-linear and interaction behavior
- Handles qualitative data as well as missing values
- Relatively stable due to diversity in trees
- Can handle small population size with large number of predictors
- Lower generalization error than other methods
- Runtime very fast, commercial/open source software available

Disadvantages

- Not so easily interpreted
- Predicts a numeric value (cost) - Not a parametric equation (CER)

Versatile Black-box Approach

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

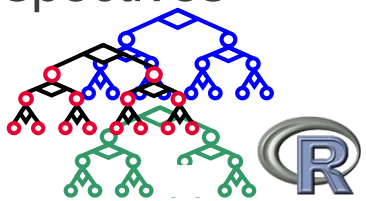
Application: Logistics Transport Cost Prediction

Objective

- Predict the shipping cost of products to help determine the best locations to manufacture them

Analysis Approach

- 1000's of data points, messy, missing values, many potential predictors
- Initial Plan: Multivariate Regression
  - Very cumbersome; required manual partitioning into suitable subsets
- Chosen method: Random Forest Prediction
  - Limited data prep; automatic partitioning / different perspectives
  - Very easy to implement, execute, and analyze



Random Forest Prediction facilitates logistics transport cost analysis

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Logistics Transport Cost Prediction Model

Data Description

- Consists of 150K data points
- Automatically separated into two distinct data sets
  - Domestic with ~ 100K data points
  - International with ~ 50K data points

Potential Predictors

- Started with 20 potential predictors
- Reduced to 3 key predictors
  - Mode of transportation
  - Origin &/or Destination (country/state)
  - Bill weight

Getty images credits: Mario Gutiérrez – delivery truck; Anucha Sirivisansuwan: barge; hollydc: mailbox; oat autta: cargo truck; JPM: train

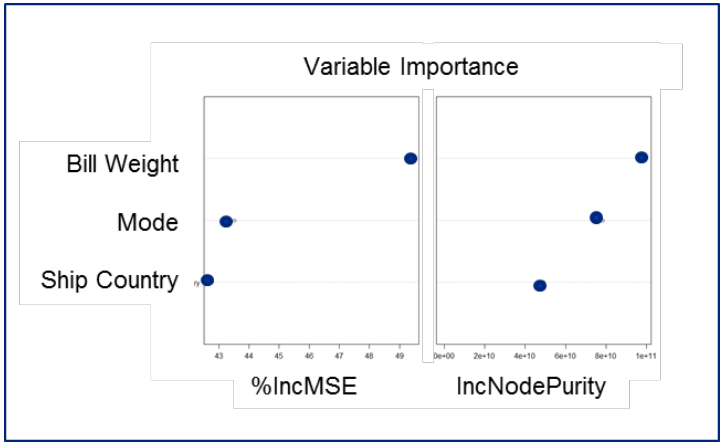
Random Forest Prediction for Big, Messy Data

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Analytical Results

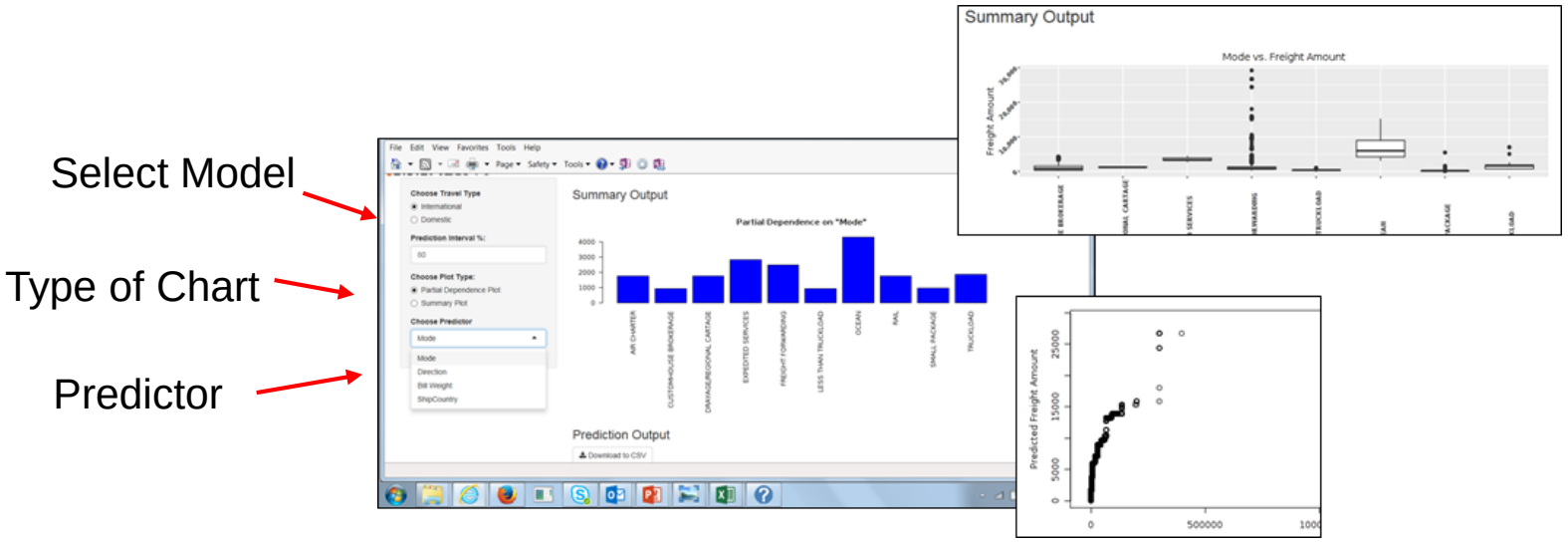
Goodness of fit – Predicted R^2

- International: 0.83
- Domestic: 0.88



Graphical Interpretations

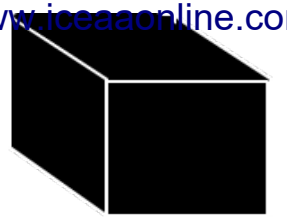
- Quickly produce various charts via R Shiny web-based application



Analysis made easy with R Shiny Package

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Next Steps: What to do about the …

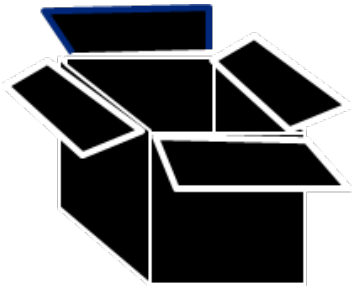


Decision makers want to know what's inside

- What can we do?

Compare results to actuals …

- Using excel? **Be Careful!**

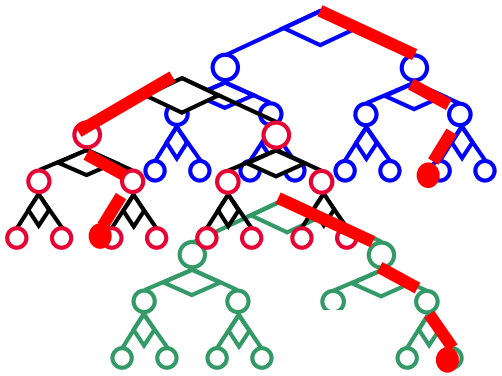


Develop Interpretation GUI

- R-Shiny to peek inside the black box
- Visualize / Automate standard statistical analyses
- Ability to "play" with the model

Build algorithm to "create" a CER

- From all the trees, branches, values
- Cost prediction ≈ f(tree$_i$) $i = (1..n)$



Provide ability to "peek" into black box

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Machine Learning Approach to Cost Analysis

Machine Learning in General

ML Algorithms for Cost Analysis

Applications related to Cost

- Random Forest Prediction
- **Latent Semantic Analysis**

Challenges

*** ML = Machine Learning**

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Application: Analysis of Cost Saving Ideas

Objective

- Identify best cost savings ideas to apply to other products

Analysis Approach

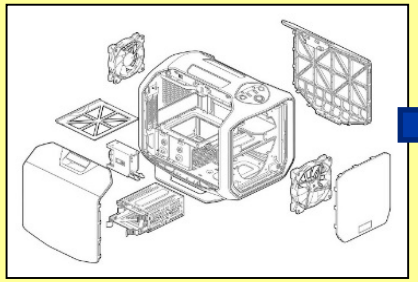
- Collaborative workshops to generate ideas to optimize the product
- 1000's of ideas in free form text from 100's of workshops
  - Could any of these ideas be applicable to other products?
- Natural Language Processing to identify cost-savings ideas for reuse
- Chosen Methods: Latent Semantic Analysis, Latent Dirichlet Allocation
  - Powerful, well-proven, task-invariant algorithms
  - Framework already in place – Open source algorithms

Natural Language Processing Analyses highlight ideas for reuse

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Generalize Cost Savings Ideas via Text Analytics

Collaborative Idea Generation



Review Product Detail

Generate Ideas: 10s

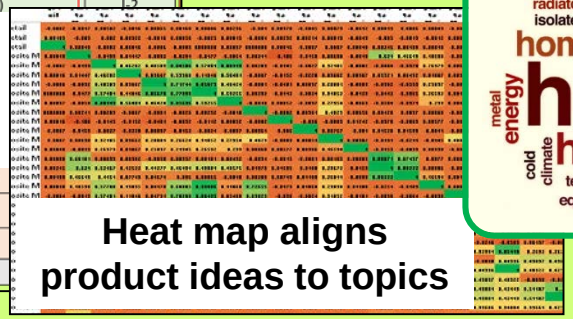
Aggregate ideas from 100s of products

1000s Unique Ideas

Machine Learning Analysis

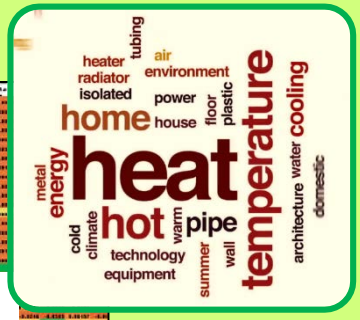
Campaign	Idea	Idea Id
Flight Deck Seats	Alternative Fabric	I-1
	Use off the shelf actuators (rotary actuator)	I-2
Fuel Act		
Power Distribution	Use of Updated and Lower Cost Transistors	
	Change Material of Solenoid Housing	
	..	
...	...	

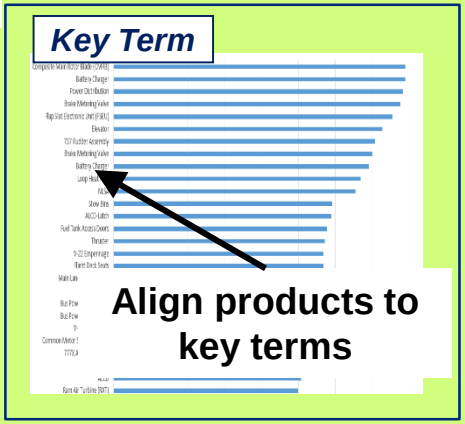
Group ideas into topics to generalize results



Heat map aligns product ideas to topics

Identify key terms





Can we identify & apply Ideas from one product to others?

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Similarity Matrices to Align Ideas

Unstructured Text

- 100s documents
- 1000s freeform "texts"

Product	Idea	Idea Id
Flight Deck Seats	Alternative Fabric ...	I-1
	Use off the shelf actuat	I-2
	...	I-3
Fuel Valve Actuator	Invest in a design prior to request bids ...	I-4
	Make gears l	I-5
	...	I-6
Power Distributi on	Use of Updated and Lower Cost Transistors	I-7
	Switch	I-8
		I-9
...	...	...

Product X

Product Y

Product Z

	X			Y			Z			
	I-1	I-2	I-3	I-4	I-5	I-6	I-7	I-8	I-9	...
I-1	1	0.8	0.2	0.5	0.2	0.4	0.3	0.6	0.8	...
I-2	0.8	1	0.5	0.1	0.3	0.6	0.2	0.8	0.2	...
I-3	0.2	0.5	1	0.9	0.2	0.8	0.5	0.2	0.2	...
I-4	0.5	0.1	0.9	1	0.6	0.4	0.5	0.2	0.3	...
I-5	0.2	0.3	0.2	0.6	1	0.6	0.1	0.3	0.6	...
I-6	0.4	0.6	0.8	0.4	0.6	1	0.9	0.2	0.8	...
I-7	0.3	0.2	0.5	0.5	0.1	0.9	1	0.1	0.5	...
I-8	0.6	0.8	0.2	0.2	0.3	0.2	0.1	1	0.9	...
I-9	0.8	0.2	0.2	0.3	0.6	0.8	0.5	0.9	1	...
...	...	...	...	...	...	...	...	...	...	1

Idea #1 from Product X highly similar to Idea #9 from Product Z

Cluster similar ideas from unique products via similarity matrices

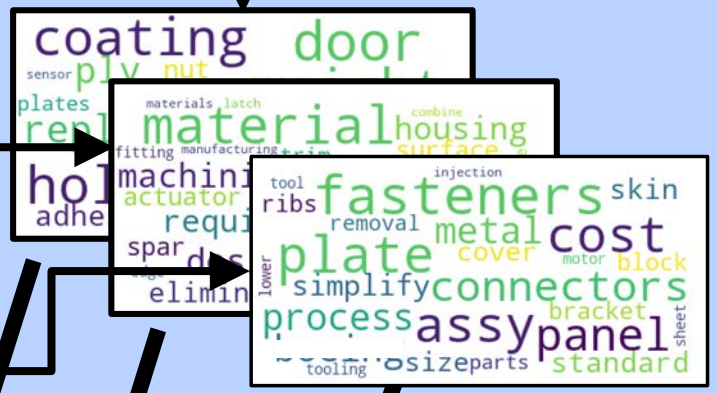
Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Text Analytics to Identify Reusable Ideas (1 of 2)

Topic cluster

	I-1	I-2	I-3	I-4	I-5	I-6	I-7	I-
I-1	1	0.8	0.2	0.5	0.2	0.4	0.3	0
I-2	0.8	1	0.5	0.1	0.3	0.6	0.2	0
I-3	0.2	0.5	1	0.9	0.2	0.8	0.5	0
I-4	0.5	0.1	0.9	1	0.6	0.4	0.5	0
I-5	0.2	0.3	0.2	0.6	1	0.6	0.1	0
I-6	0.4	0.6	0.8	0.4	0.6	1	0.9	0
I-7	0.3	0.2	0.5	0.5	0.1	0.9	1	0
I-8	0.6	0.8	0.2	0.2	0.3	0.2	0.1	1

Key Terms



Latent Semantic Analyses

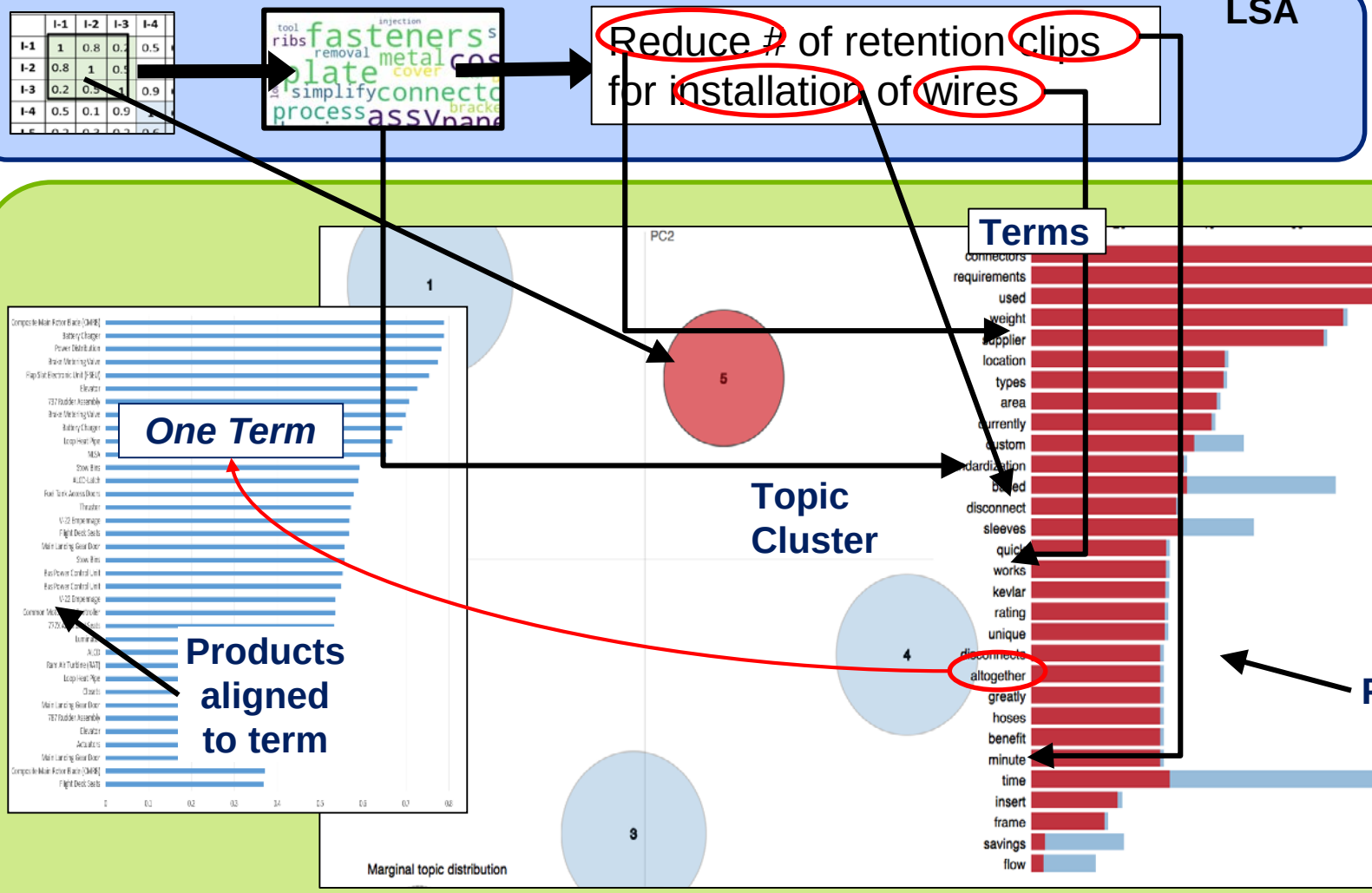
Reduce # of retention clips for installation of wires

Main Cost-Savings Idea

Cluster similar ideas & identify key terms and main concept

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Text Analytics to Identify Reusable Ideas (2 of 2)



Latent Dirichlet Allocation & Term Frequency – Inverse Document Frequency

Term frequency ~ importance ~ of idea aligned with product

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Next Steps

Validate model and verify results

- Modify & Implement existing GUI Framework
- "Evaluate" results – requires thinking!

Scale to larger population

- Hundreds more workshops & products
- Thousands more ideas

Capture and incorporate actuals

Implement cost-saving ideas on other products

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Machine Learning Approach to Cost Analysis

Machine Learning in General

ML Algorithms for Cost Analysis

Applications related to Cost

- Random Forest Prediction
- Latent Semantic Analysis

Challenges

*** ML = Machine Learning**

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Challenges for Cost Analysis Community

Machine Learning for cost analysis & estimating

- Different … from traditional methods
  - Will take time to catch on
- Black box method
  - Not so easy to interpret or follow input-to-output logic
- Regression Algorithms
  - Predict a numeric value (cost) - not a parametric equation (CER)
- ML Algorithms
  - Require pre and post processing for reasonable results

Do Benefits outweigh Challenges?

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Authors

Karen Mourikas is an Associate Technical Fellow at The Boeing Company specializing in Operations Analysis, Affordability, and Systems Optimization. Her current work includes Product Teardown & Should-cost analyses, and Production Systems modeling. Karen has MS degrees in Applied Math and in Operations Research Engineering from the University of Southern California. Karen is a life-time member of ICEAA and has presented at several ICEAA & ISPA/SCEA conferences over the years.

Nile Hanov is a Data Scientist at Boeing Research & Technology where he develops novel next gen solutions for commercial and military platforms. In this role, he applies machine learning to event driven data to help organizations better understand and predict failures on board of an aircraft. Nile has four patents under review by the U.S. Patent Office all of which focus on event forecasting and system improvement. He is also currently pursuing a Ph.D. in Computer Science (with a focus on Artificial Intelligence and Machine Learning) at University of California - Irvine.

Joseph King is a data scientist at The Boeing Company with Boeing Commercial Airplane Analytics, utilizing data to build predictive models and provide analytical solutions. Joseph has contributed to areas such as sensor data analysis, text mining maintenance messages, and customer behavior modeling. Joseph's education background includes a MS in Business Analytics from the University of Tennessee and a background in mathematics and operations research.

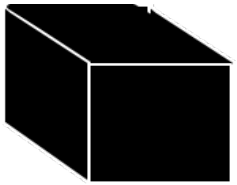
Denise Nelson is a Systems Analyst at The Boeing Company specializing in software estimating, cost-risk analysis and parametric modeling. Currently, Denise supports Boeing Commercial Airlines Product Development activities. Previous efforts include life-cycle cost analysis; reliability and maintainability analysis; and project management of immersive simulation modeling. Denise graduated from Cal Poly Pomona with an MS in Pure Math and BS in Statistics.

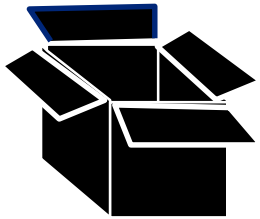
karen.mourikas@boeing.com Nile.Hanov@boeing.com joseph.a.king3@boeing.com Denise.J.Nelson@boeing.com

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Machine Learning & Non-Parametric Methods for Cost Analysis

The world of big data opens up new opportunities for ICEAA, such as machine learning and non-parametric methods. These methods are more flexible since they do not require explicit assumptions about the structure of the model. However, a large number of observations is needed in order to obtain accurate results. Hence, big data to the rescue! This presentation examines several non-parametric methods, with examples related to our community, and discusses opportunities and limitations going forward.





Abstract

BOEING

Engineering, Test & Technology
Boeing Research & Technology

Questions?