JOINT FLORIDA
Model Task Force & Transportation
Data and Analytics Workshop

FTO FORECASTING & TRENDS OFFICE

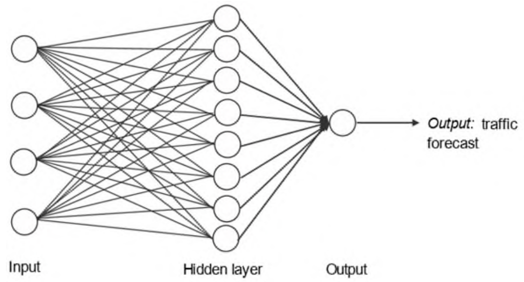
FDOT TDA TRANSPORTATION DATA & ANALYTICS

Machine Learning 101

Mohammed Hadi, Ph.D., PE

Background

- Detailed data from multiple sources has strong potential for advancing modeling, forecasting, and understanding traveler's behaviors.
- Advancements in data mining/machine learning and tools provide opportunity for such use of data.
  - Ability to deal with non-linear complex functions and noisy data
- Machine learning techniques are suitable to integrate different data sets that can supplement each others in providing answers.
  - Can be used in combinations with existing modeling techniques in an integrated analytical framework



Analysis Categories

- Descriptive analytics: describe current conditions
  - Statistical measures by category or clusters.
    - Performance measures, strategic behaviors, microscopic behavior
  - Patterns, trends, and relationships in the data.
- Diagnostic analytics: why things happen
  - Cause-and-effect relationships, conflicts, companion relationships, confounding factors, etc.
- Predictive analytics: predict/forecast future
  - Long-term, mid-term, short-term
- Prescriptive analytics: Recommend the best alternative and assess impacts

Analysis Approaches

- Descriptive statistics and visualization
- Analysis, modeling, and simulation (demand forecasting, macro, meso, micro, multi-agent, multi-resolution)
- Data analytics/machine learning
  - Statistical regression
  - Clustering
  - Associations and correlation rules
  - Decision trees and tree ensembles
  - Bayesian classifiers
  - Support vector machine (SVM)
  - Artificial neural networks (ANN)
  - K- nearest neighbor (KNN)
  - Expert Rules and Fuzzy Logic
- Return-on-Investment
- Multi-criteria decision analysis

Selection of Approach

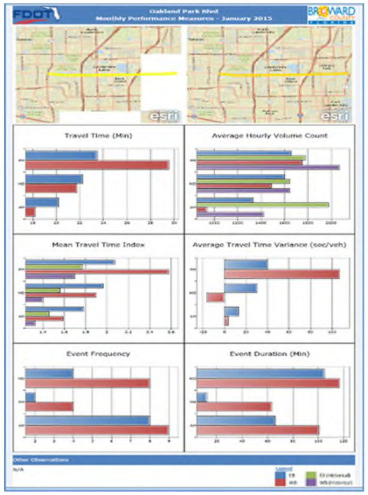
- Applicability to the problem (monitoring, diagnosis, forecasting, prescription)
- Ease and cost of use
- Ease of understanding and interpreting results
- Accuracy and required sample size
- Dealing with complex functions
- Handling high data dimensionality and handling large datasets
- Others: preprocessing requirements, data distribution requirements, dealing with collinear data, resistance to overfitting, complexity of parameters to tune.

Performance Metrics Identfiication

- Inputs are metrics of the utilized resources in the activity (e.g., dollars, person-hours, etc.).
- Activity or Process metrics are the actions of the agency and partners to meet a project’s objectives.
- Outputs are metrics of the activities produced with the inputs such as percent of facilities with sensors, etc.
- Outcomes are metrics that quantify the results of an activity in meeting the organization’s mission and vision.
- Impacts are metrics that reflect broader levels of change due to the activities such as health improvement and economic growth.

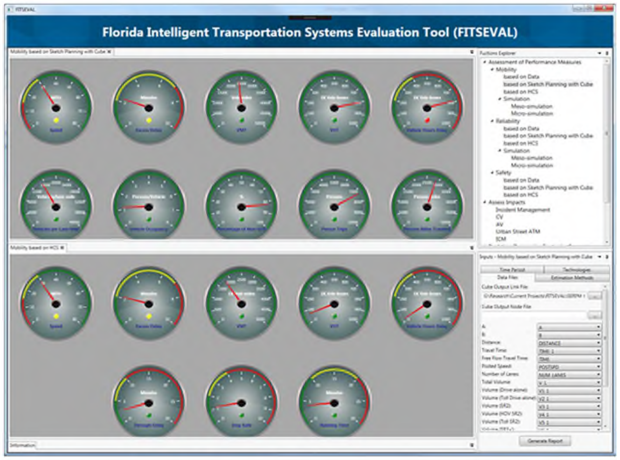
Reporting and Visualization

- Reports
  - Easy dynamic customization, ability to drilldown, rollup, slice, and dice.
  - Standard reports and ad-hoc reports
- Dashboards



Oakland Park Blvd
Monthly Performance Measures - January 2015

Other Observations
N/A



Florida Intelligent Transportation Systems Evaluation Tool (FITSEVAL)

Cube Output Link File: G:\Research\Current Projects\FITSEVAL\SERPM 1

Cube Output Node File:

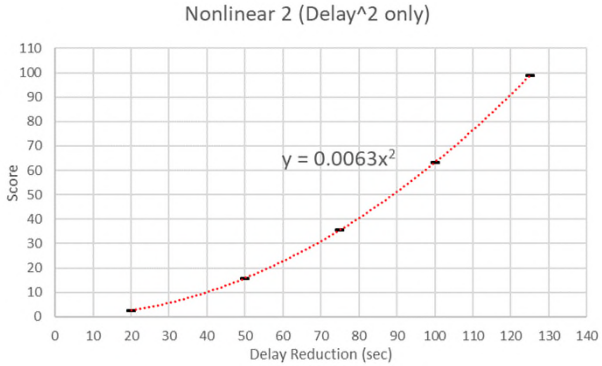
A: A
B: B
Distance: DISTANCE
Travel Time: TIME 1
Free Flow Travel Time: TIME
Posted Speed: POSTSPD
Number of Lanes: NUM LANES
Total Volume: V 1
Volume (Drive alone): V1 1
Volume (Toll Drive alone): V2 1
Volume (SR2): V3 1
Volume (HOV SR2): V4 1
Volume (Toll SR2): V5 1

Supervised and Unsupervised Learning

- ML can be supervised, unsupervised, and reinforcement learning.
- Supervised learning requires training data that include feeding paired inputs and outputs to the model. Examples of supervised learning are statistical regressions, K-nearest neighbors, SVM, ANN, decision trees, and tree ensembles.
- With unsupervised learning, the input data is not associate with outputs. Examples are clustering and association rules.
- Reinforcement learning can observe the conditions and select the best actions for a given situation.

Statistical Regression Models

- Linear Regression is the most widely known but has many assumptions that do not apply in all cases.
- Logistic Regression predicts a binary dependent variable.
- Multinomial Logit or Probit models predict the probability of discrete outcomes based on a set of factors.
- Poisson Regression and Negative Binomial Regression are used when the dependent variable is count data.

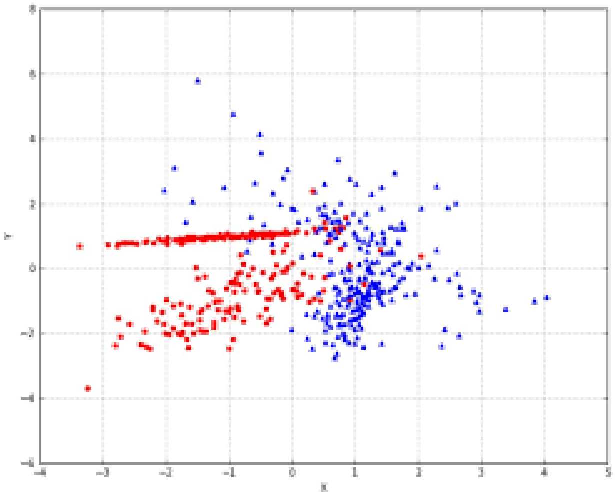


Decision Trees and Tree Ensembles

- Popular supervised machine learning tool that can be used for both classification and regression.
  - A decision tree can classify measurements and can also estimate the probability of an instant belonging to a particular class.
  - Number of decision tree algorithms available.
- Widely used in transportation engineering
- Easy to understood and develop and has a good accuracy.
- Scalability issues have been identified for very large data set.
- Tree ensembles combine the results from the development of multiple trees and generally outperform single trees. Examples are the Random Forests and Gradient Boosted Trees.

Bayesian Classification

- Bayesian classification uses the Bayes’ probability theorem to predict the class membership probabilities.
- It was reported that the performance of Bayesian classifiers can be comparable to classification decision trees and some neural networks.



Artificial Neural Networks

- Can deal with very complex and large classification, recognition, prediction, and recommendation of action task.
- Consists of nodes emulating neurons organized in layers and links that connect these nodes.
- The most common ANN is a supervised learning method referred to as the multi-layer perception (MLP).
    - Trained to determine the weight on each links using an optimization process referred to as the Backpropagation training algorithm.
- Deep networks (deep learning) have proven to be more efficient in modeling complex functions with less number of neurons.

Other Types of ANN

- Recurrent Neural Network (RNN) is a class of ANN that is very powerful in prediction.
- Long Short Term Memory (LSTM) has been successfully applied to traffic performance prediction.
- Convolutional neural network (CNN) emulates the brain’s visual cortex and used in visual applications such as image processing, automated vehicles, and automatic video classification
- Autoencoders can be used as feature detectors and removing noise in the data.

Support Vector Machine

- A powerful supervised machine learning tool allowing classification, regression, and outlier detection.
- Less susceptible to overfitting compared to decision trees.
- Linear SVM classifiers separate the instances into different classes by straight lines.
- In some cases, Linear SVM is not sufficient and nonlinear SVM classification has been used.
- However, the computation associated with SVM is slow and not efficient for large data.

Clustering

- Unsupervised learning techniques aim at segmenting of objects to subgroup.
  - Objects within a cluster are closely related compared to objects in other clusters.
  - utilize a dissimilarity measure to cluster the objects.
- Clustering have been recommended and used to identify operational patterns and modeling scenarios.
  - Recommended for use in the revised FHWA Traffic Analysis Toolbox Volume 3.
- K-means clustering has been widely used. Other examples are K-prototypes, K-medoids, Hierarchical clustering, clustering with dimension reduction using PCA, fuzzy clustering, Gaussian mixture models (GMM) clustering, clustering using Wavelet transformation.

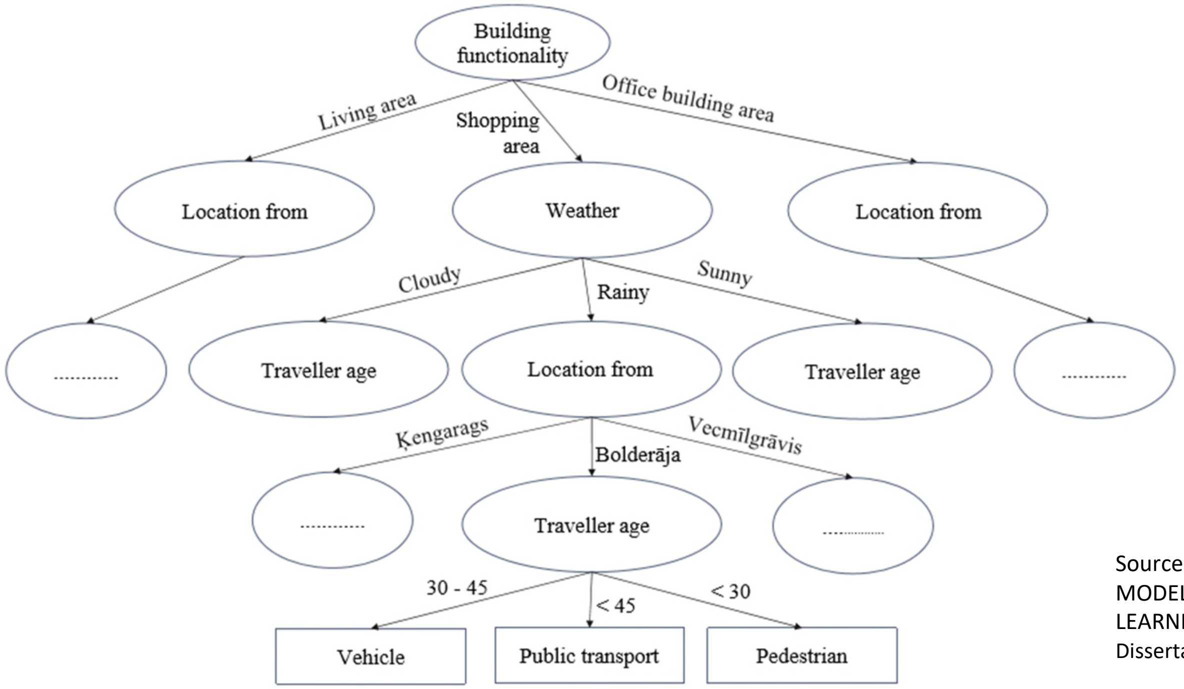
Expert Rules and Fuzzy Logic

- Oldest form of artificial intelligence is what was referred to as “Expert Systems” with expert rules constructed based on expert inputs.
- Often, the rules cannot be delimited by sharp boundaries and are associated with ambiguity and uncertainty.
- Fuzzy rule-based systems extend the problems of classification, prediction, and prescription to deal with vagueness and uncertainty.

Example Applications to Modeling

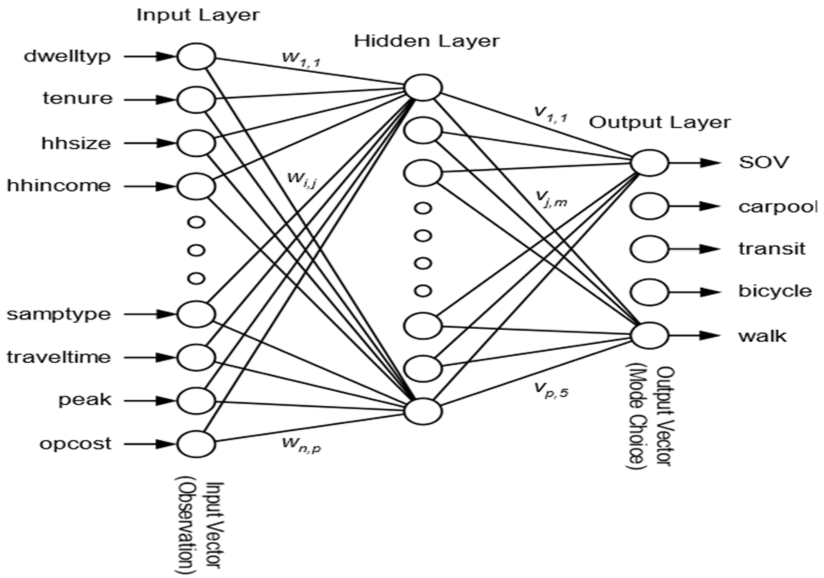
- Clustering for demand and traffic pattern modeling
- Several efforts on decision trees and neural networks applications to discrete choice behavior analysis
- Comparison of the performance of DT, ANN, and MNL for model split
- Rule-based machine learning method for trip generation model for eight different trip purposes
- Hierarchical rule-based model for trip generation and modal split
- Utilizing decision trees for trip generation, activity patterns, and mode of transportation
- Forecasting the impact of external factors on demands
- ABM model accounting for generation, distribution and mode split utilizing decision trees, random forest, and modified decision trees
- Neural network to forecast demands

Decision Tree Application



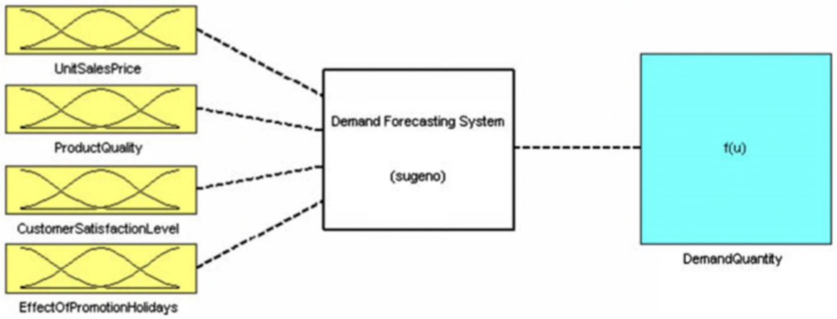
Source: ZEŅINA, N. TRANSPORT TRAVEL DEMAND MODEL DEVELOPMENT BASED ON MACHINE LEARNING AND SIMULATION METHOD. Ph.D. Dissertation RIGA TECHNICAL UNIVERSITY

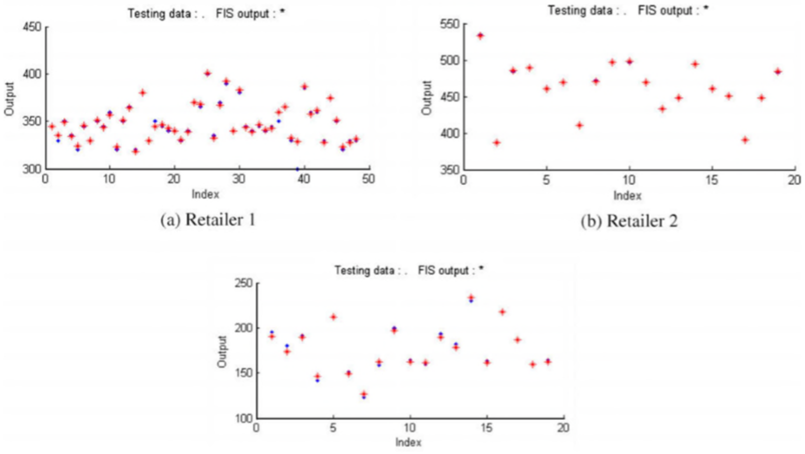
Mode Choice Using ANN



Source: Xie, C. WORK TRAVEL MODE CHOICE MODELING USING DATA MINING: DECISION TREES AND NEURAL NETWORKS. presentation at the 82nd Transportation Research Board Annual Meeting, January 2003, Washington, D.C.

Shipment Demand Forecasting Using ANN


UnitSalesPrice
ProductQuality
CustomerSatisfactionLevel
EffectOfPromotionHolidays
Demand Forecasting System
(sugeno)
f(u)
DemandQuantity



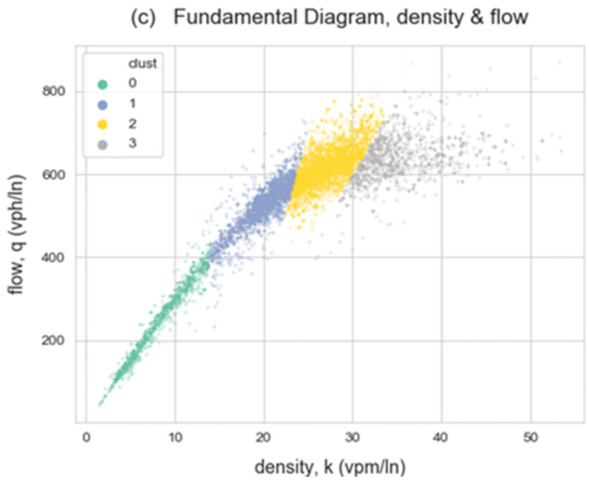
(a) Retailer 1

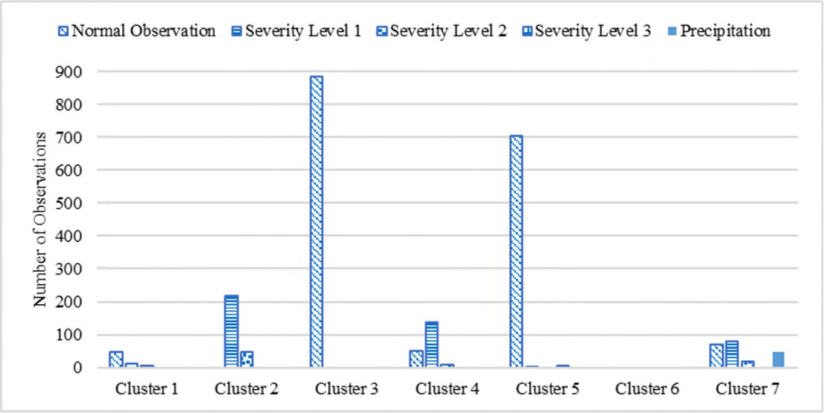
(b) Retailer 2

(c) Retailer 3

Source: Tuğba Efendigil Semih Önüt Cengiz Kahraman A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy models: A comparative analysis in Expert Systems with Applications: An International JournalApril 2009 https://doi.org/10.1016

Pattern Clustering





Source: Hadi et al. 2019