Engineering, Test & Technology
Boeing Research & Technology

Machine Learning Approach to Cost Analysis

Karen Mourikas, Joe King, Denise Nelson

ICEAA SoCal Workshop, September 2017

Machine Learning Approach to Cost Analysis

Machine Learning
& Cost Estimating

Random Forest Prediction

Applications

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 …

A. teach themselves to grow (i.e. learn) from data

Algorithms …

- analyze, generalize and learn from relationships & trends
- derive a complex model that produces data-driven predictions

B. learn without being explicitly programmed

Machine Learning …

- does not require prior knowledge of relationships between inputs & outputs
- implies “dummy-proof system” but it’s not

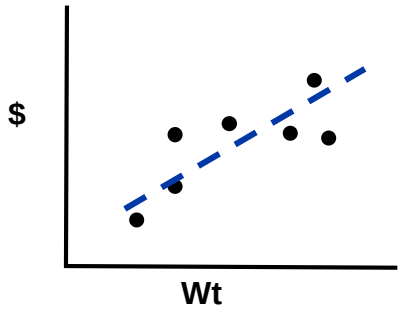
* Russell, Stuart J.; Norvig, Peter ; Artificial Intelligence: A Modern Approach, 2003 & 2009

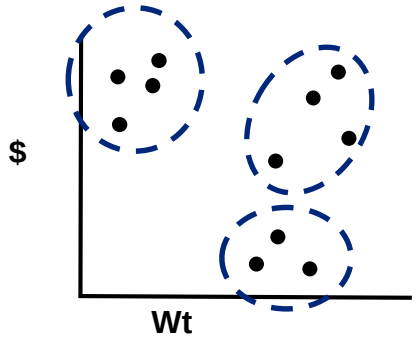
Machine Learning is a type of Artificial Intelligence

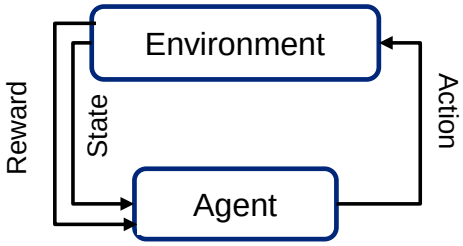
Three types of Machine Learning

Machine Learning

- **Supervised**
  Task Driven:
  Regression,
  Classification
- **Unsupervised**
  Data Driven:
  Clustering,
  Neural Nets
- **Reinforced**
  Reaction to
  environment:
  "WarGames"







Our focus is on Supervised Regression

Machine Learning for Cost Estimating & Analysis

Typical Cost Estimating

- 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 easy to implement

Could we use Machine Learning techniques for cost estimating?

Machine Learning Approach to Cost Analysis

Machine Learning & Cost Estimating

Random Forest Prediction

Applications

Machine Learning: Random Forest Prediction

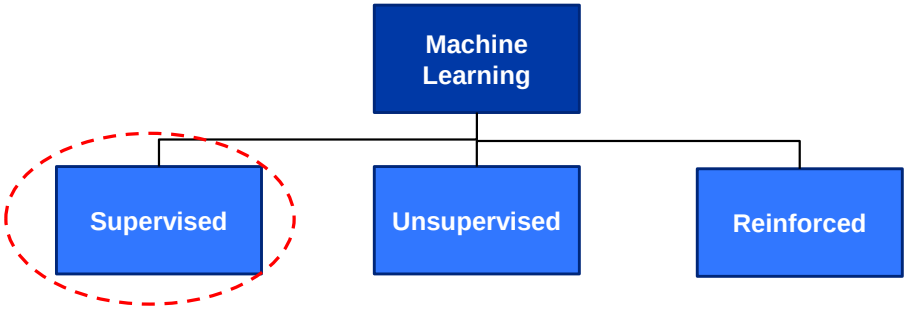
Supervised Learning

- Regression

Ensemble Approach

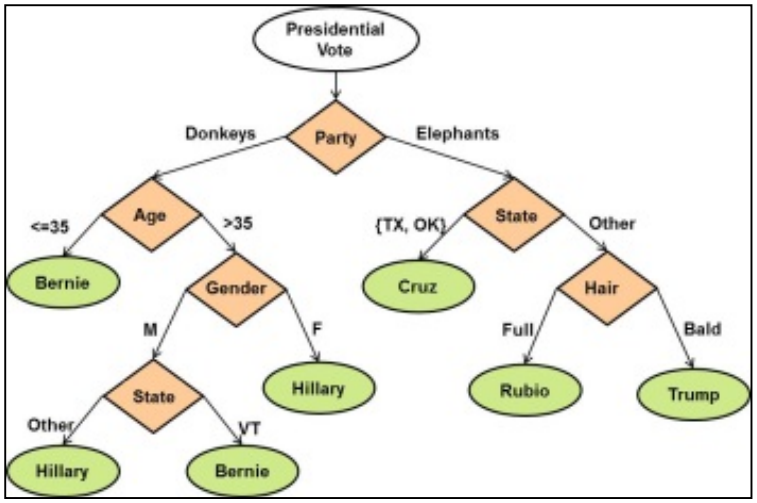
Decision Tree Theory

Randomness



Decision Tree

- Easily interpretable model
- Represents a set of decisions & outcomes
- Similar to “20 Questions”

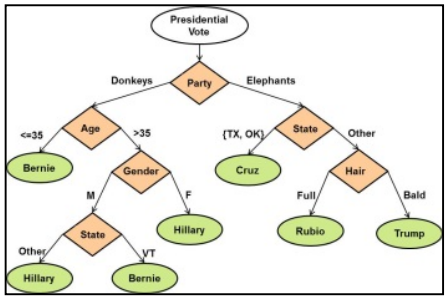


Random Forests popular with machine learning community

Trees and Forests

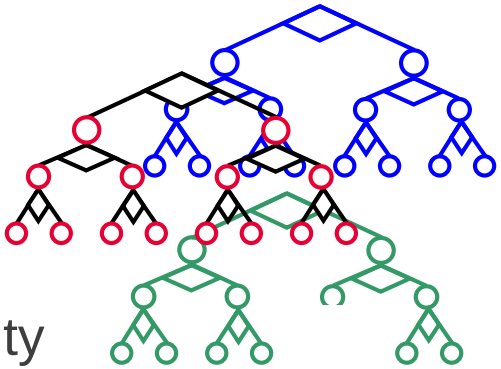
A Single Decision Tree

- Interpretable, but …
- Not a great predictor



An Ensemble of trees

- Not as easy to interpret, but …
- Provides greater prediction accuracy & more stability



Random Forests

- Ensemble of decision trees "randomly" constructed
  - Advantages of an ensemble, plus …
  - More accurate predictions and reduced error

Source: Alexas_Fotos/Pixabay

Random Forests Prediction based on Decision Tree Theory

Random Forest Background

History of Random Forests

- Fairly recent – introduced in late 90s
- Leo Breiman – father of Random Forests
- Combined “bagging” with random selection methods*
- Incorporated variable importance via permutation and out-of-bag error

Two phases of Random Forest modeling

- Training – building the model
- Prediction – estimating response of new input

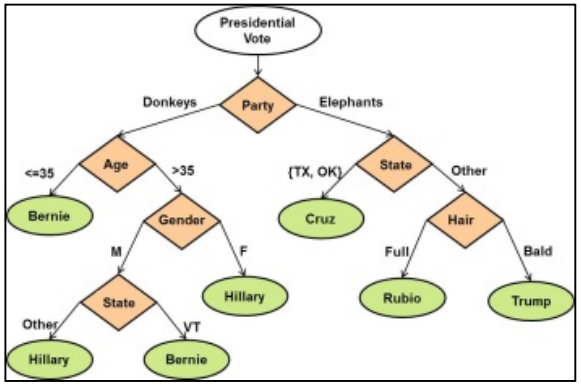
*Introduced independently by Ho, and by Amit and Geman

Bridging gap between Statistics and Computer Science

Building a Random Forest model

To grow Random Forest trees (Training Phase)

- Select a ***random subset*** of data
- Select with replacement from subset (predictors and responses)
  - Bootstrapping Aggregation (Bagging)
- Search over ***random subset*** of the values to select split point
- Continue until number of observations in node small
  - Response is average of all observations in terminal node
- Repeat to grow more trees



To Predict a new response (Prediction Phase)

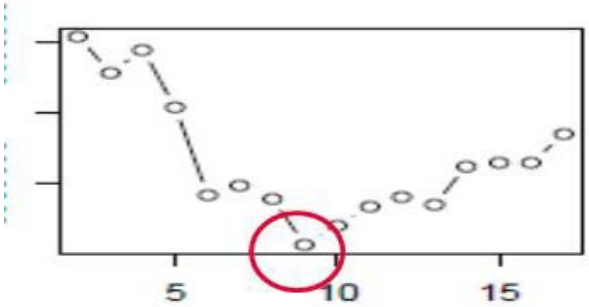
- Run new input down each tree and average result from terminal nodes

Similar to building Decision Trees but with variation among trees

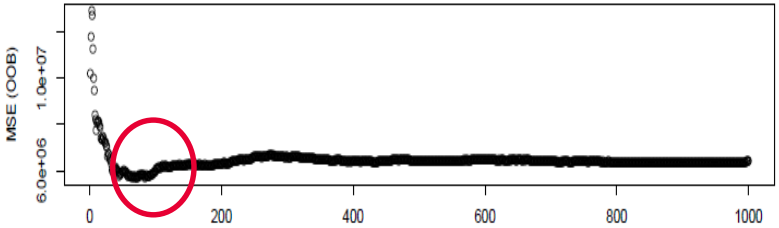
"Random" Advantages

OOB data to minimize errors and determine optimal predictors

- Using different (random) training sets reduces correlation between trees
- Reducing size of random subset reduces correlation
  - But also strength



- Optimal number of randomly selected predictors at each split node
- Optimal number of trees
  - From a few hundred to several thousand

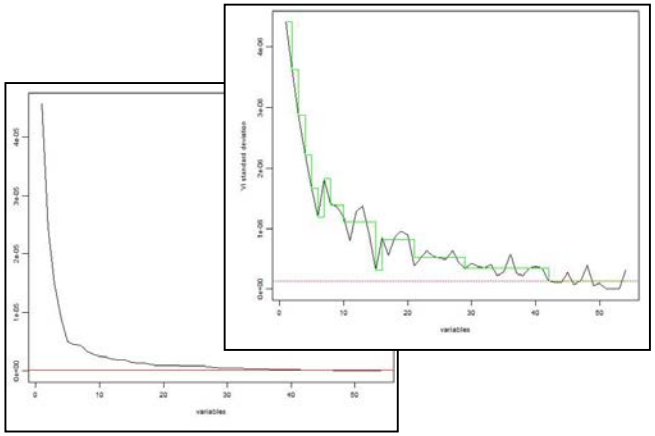


Algorithm incorporates randomness for better prediction

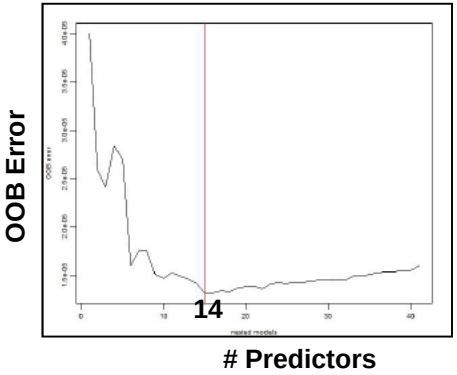
Variable Selection Process*

Three Step Process

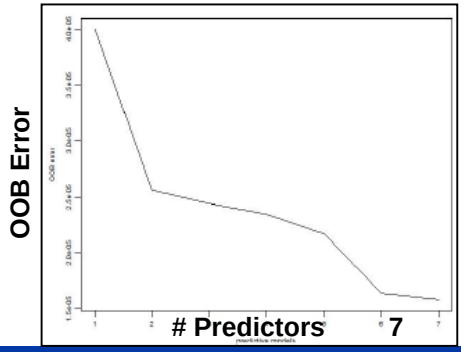
1) Thresholding
   - Eliminate irrelevant variables from the dataset
     - Original dataset: >50 potential predictors



2) Interpretation
   - Select all variables related to the response for interpretation purpose
   - 14 “most important” predictors



3) Prediction
   - Refine the selection by eliminating redundancy in the set of variables from step 2
   - Final set contains 7 predictors

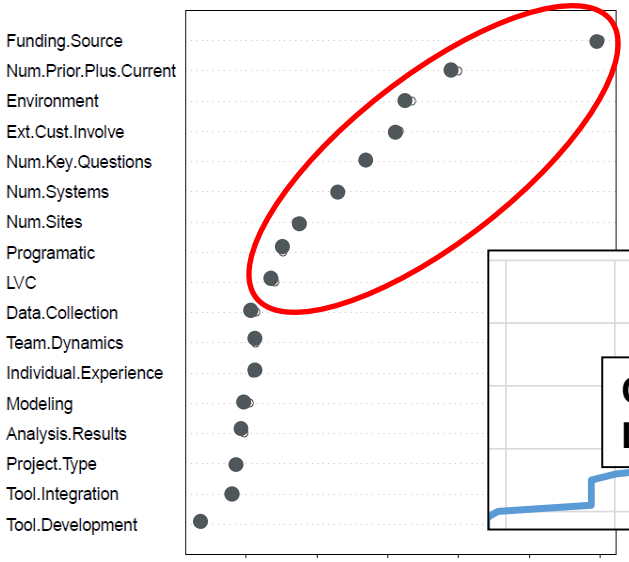


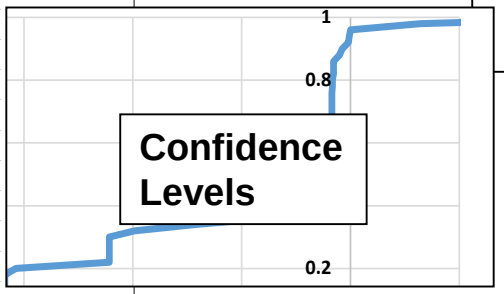
*VSURF package in R

Random Forest model reduced from ~70 to 7 variables

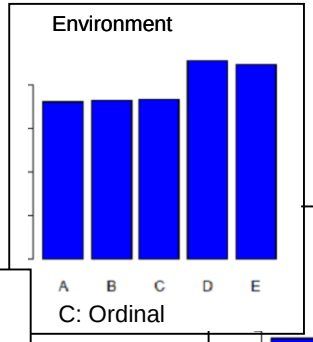
Tools to Understand the Model & Results

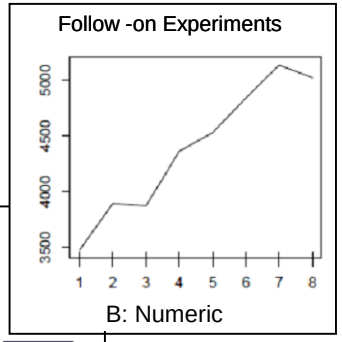
Variable Importance

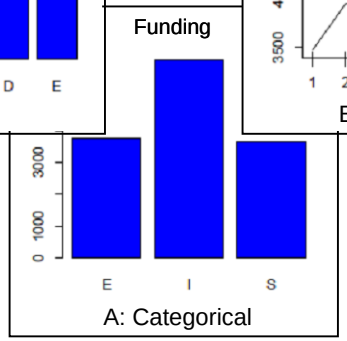




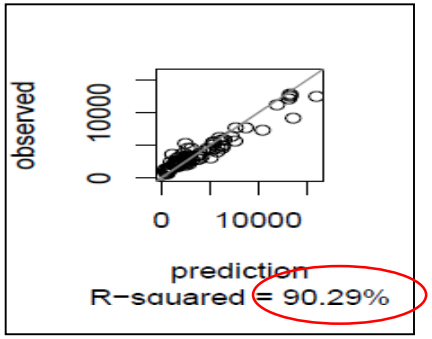
Impact per Predictor



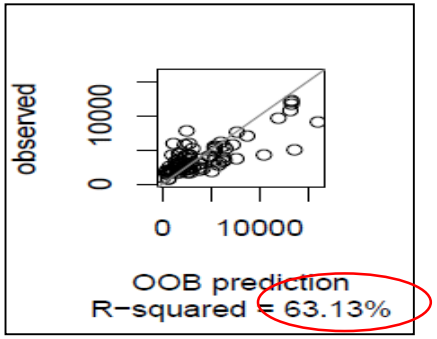




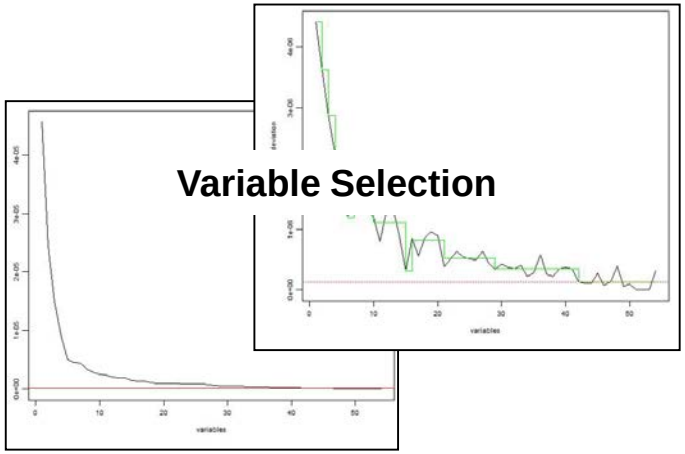
Goodness of Fit



With Training data



Using OOB data



Typical Statistical Analysis to Interpret Model Fit & Results

General Applications & Uses

Agriculture / Forestry

Pollutants

Astronomy

Medical and Health fields

Manufacturing

Marketing

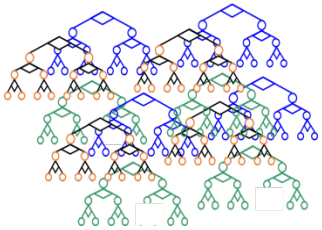
Buying patterns

Remote sensing

Object recognition

Random Forests widely used in other fields

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
- Estimates variable importance
- Handles problems with small population size and large number of predictors
- Lower generalization error than other methods
- Runtime very fast

Disadvantages

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

Versatile Approach

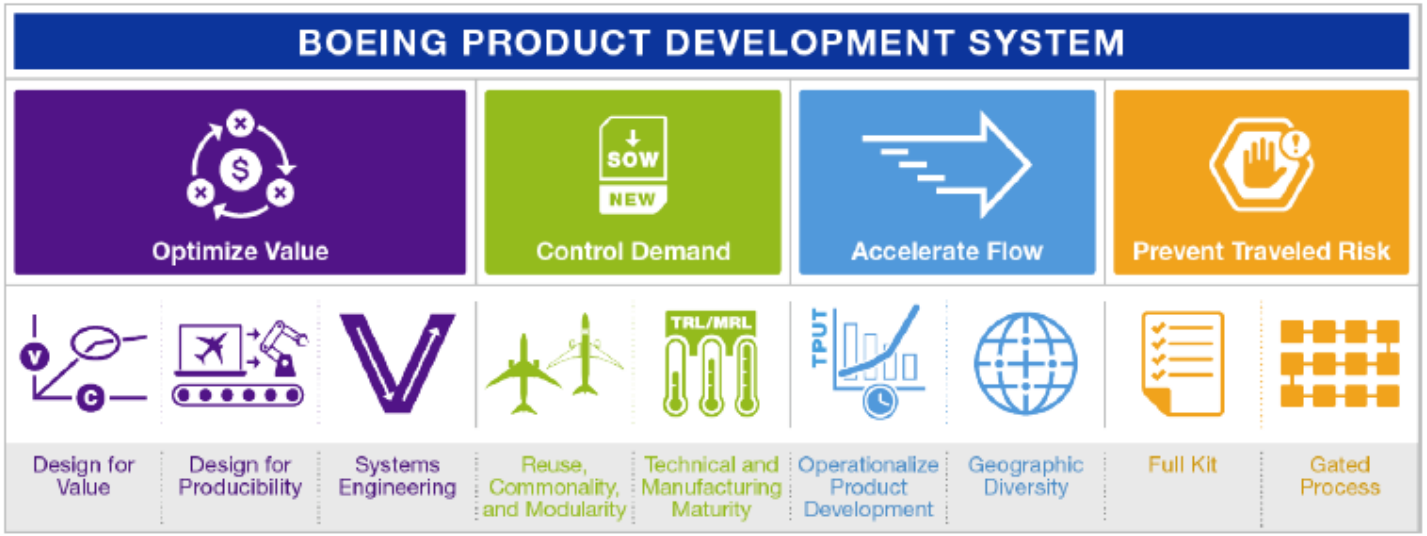
Machine Learning Approach to Cost Analysis

Machine Learning & Cost Estimating

Random Forest Prediction

Applications

RFP Application #1: BPDS Cost Savings



Objective of *BPD$aves*

- Develop a simple-to-use estimation model to identify cost savings that programs could achieve with BPDS methods
- Develop a comprehensive program-focused model for predictive valuing of future BPDS projects

BPDS methods used to reduce costs

BPD$aves Model and Data

Characteristics of Model

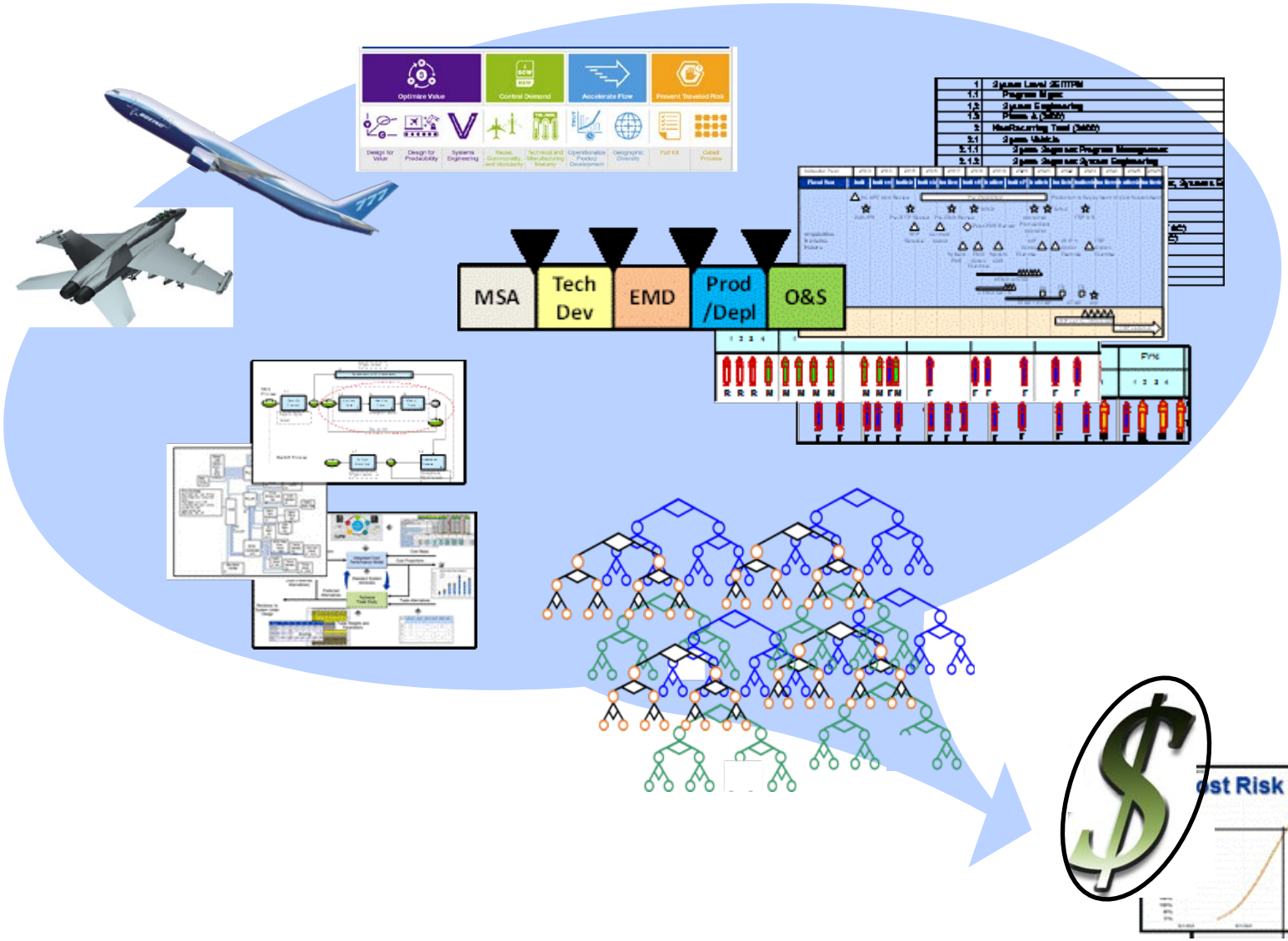
- NRE/RE, all phases
- Limited number of inputs
- Quick turn-around & credible ROM results

Limited, disparate data set

- Common and unique variables for each BPDS method
- Relatively small dataset
- Numerous potential predictors
- Qualitative data as well as missing values
- Multiple data sources

Dataset characteristics well suited for Random Forest

BPD$aves: Characteristics of Data Set



Limited inputs identify savings opportunities

BPD$aves Model Statistics

Goodness of Fit

- Predicted R^2 = 0.82
  - Measure of variability in the data sample explained by the model
  - Cross-validation approach to measure model fit to predict new values
- Mean Square Error = 0.15
  - Measure of variance of residuals (non-fit) in the population

Prediction Test

- 7 new projects, multiple domains
- Predicted savings within 20% of program actual savings
  - Some high, some low

Average Results	
Average High	17.3%
Average Low	-20.1%

Model predicts savings within 20% of actuals

BPD$aves Input Sheet & Results Sample

Input Sheet

Put R Application Directory Here:	C:\Users\cf731e\Documents\R\R-3.2.3\bin
Put Directory Folder Location Here:	C:\Users\cf731e\Documents\Projects\RF Project
Would You Like to Deploy the Model?	TRUE
Would you like to Re-Build the Model?	FALSE
% Intervals	70

Run

1) Select BPDS Method

3) Run model

2) Input 14 parameters
Characteristics of the program/campaign

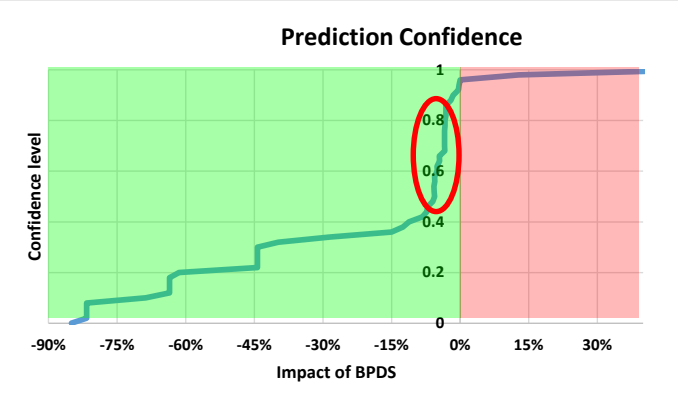
INPUTS:

Method	*ProjectID*	*Business unit*	*First time used / Previously*	*Team Experience*	*Gated process rigor*	*Lean+ rigor*	*WBS level*	*Integration Complexity*	*FRP*	*# Parts relating to system*	*OPD Maturity*	*Leadership Experience with OPD*	*Proj Duration*	*Gate Start*	*Pgm Size*
dfv	a1	BDS	F	M	M	N	4	M	M	S	L	L	S	M	S
opd	a72	BDS	F	L	M	W	2	M	M	L	M	M	Q	M	S
rcm	a24	BCA	F	E	W	W	4	M	H	S	NULL	NULL	Q	M	L

4) Outputs

Prediction interval based on selected input (70%)

ProjectID	Prediction: % Impact*	Lower Prediction Interval	Upper Prediction Interval
n1	-3.4%	-3.1%	-63.5%



Simple Excel-based interface

Example R code in *BPD$aves* Model

```
#### Function to find predictors that are of most importance
if(VI_bool==TRUE){
rFmodel2 <- VSURF(xTRAIN,yTRAIN,ntree=100)
jpeg('Plots and Performance/Variable Selection
Plots.jpg',width=1200,height=1000)
plot(rFmodel2)
dev.off()
keepthese <<- names(xTRAIN)[rFmodel2$varselect.pred]
buildModel(finalFrame,VI_bool=FALSE)
}
rFmodel <<- randomForest(x=xTRAIN,
                         y=yTRAIN,
                         ntree=100,
                         importance=TRUE)
qrfModel <<- quantregForest(xTRAIN,yTRAIN, keep.inbag=FALSE)
save(list = c('qrfModel','rFmodel'),file="rFmodel.RData")
}
```

Variable Importance Analysis

Building the model

Built-in functions in R for Random Forest modeling

Analyzing the Model's fit with R

Mean Square Error

Predicted R Squared Val

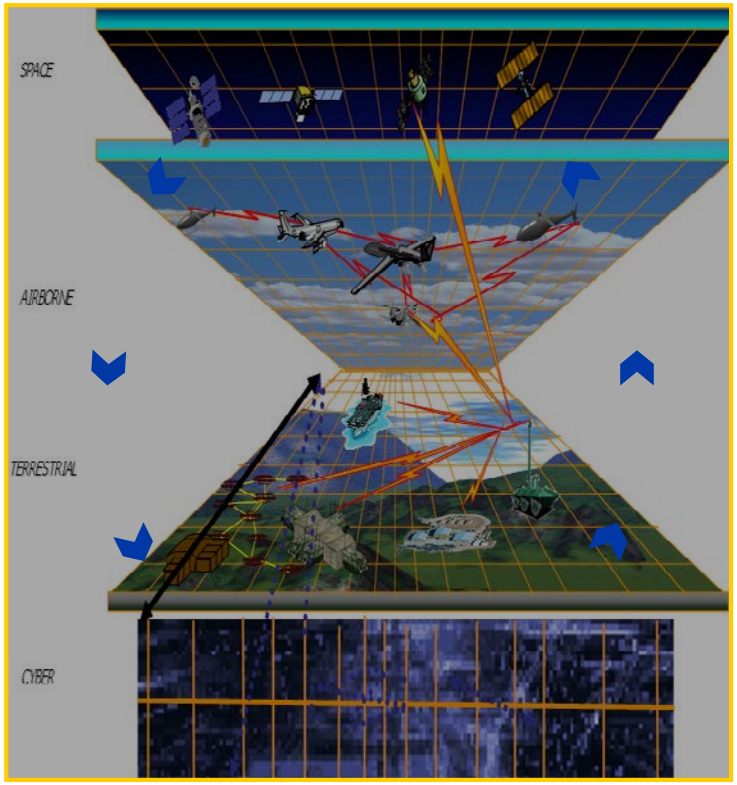
Goodness of fit chart

```
#MSE
(theMSE <<- sum((CVstat$yhat-CVstat$yobs)^2)/length(CVstat$yobs)*10^6)

#### make R Squared value
r2_check <- sum((CVstat$yobs - CVstat$yhat)^2)/sum((CVstat$yobs -
mean(CVstat$yobs))^2)
r2_check <- 1 - r2_check

###### save predicted vs. actual plots
jpeg('Plots and Performance/Predicted vs. Actual.jpg', width=800, height=600)
print(ggplot(CVstat, aes(x = yobs, y = yhat)) +
  geom_point() +
  geom_abline(color = "red") +
  ggtitle(paste("OOB Goodness of Fitness for: RandomForest Regression in R r^2=",
r2_check, sep="")) +
  labs(x="ACTUAL OBSERVATION", y="PREDICTED OBSERVATION"))
dev.off()
```

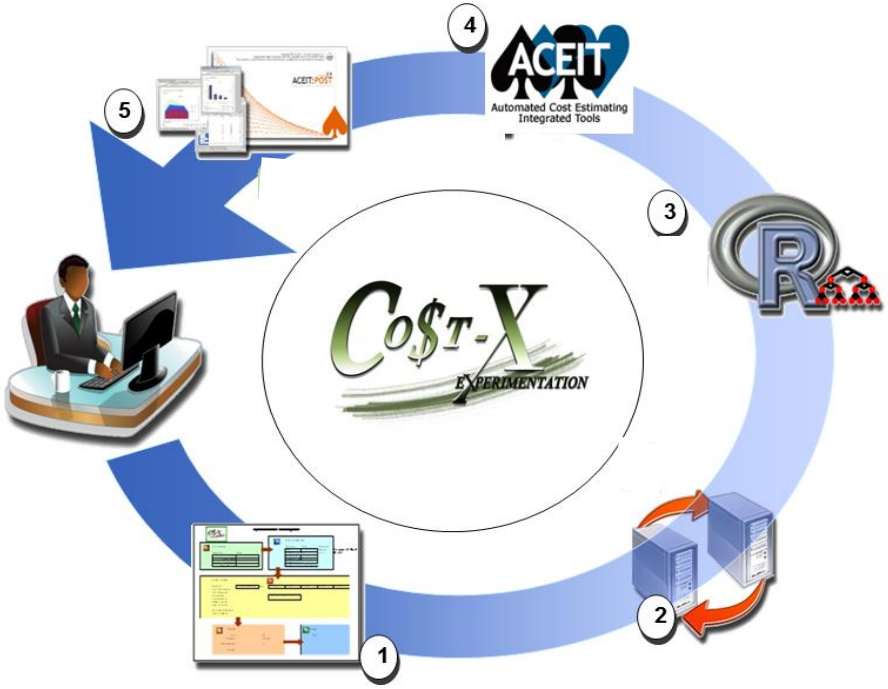
Typical statistical functions in R

RFP Application #2: Project Budgeting

Experimentation



- **Experimentation Estimation**



Co$t-X model estimates cost of experimentation projects

Co$t-X based on Random Forest Prediction

Co$t-X Model

Assumptions

- Non-linear model
- Expect interaction

Project Data

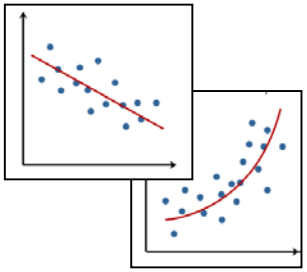
- 70 Completed Projects
  - Various characteristics
  - Incorporated reuse & learning
- Data Issues
  - Relatively small (70 data points)
  - Numerous potential predictors (18 attributes)
  - Lots of qualitative (2 categorical and 10 ordinal with various levels) data

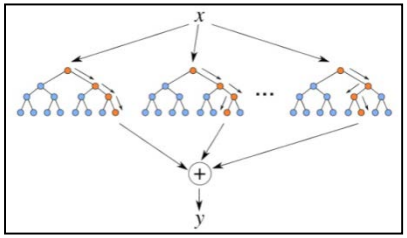
	Reusability							Team	
Project ID	Modeling	Tool Developmt	Tool Integration	Environment	Programatic	Data Collection	Analysis Results	Individual Experience	Team Dynamics
a1	B	B	E+	A	A	B	A	E	E
a2	C	A	B-	B	B	C	B-	B+	B

	Proj Info	Complexity			Execution		Cost/Sched			
Project ID	POC	Type	Num Sys	VCL	Num MOEs	Num Sites	Mths	Cost ($K)	POP	D
a1	John Smith	Modeling	5	C	0	1	4	$ 90	4Q 2008	E
a2	Pocahontas	Analysis	12	C	4	2	10	$ 450	mar 07-dec 07	B
a3	Hilary Clinton	Analysis	3	C	2	5	4	$ 61	Jun - Sep 08	B
a4	Meriwether Lewi	Modeling	20	C	4	1	8	$ 870	Oct 08-may 09	
a5	William Clark	Experiment	4	V	6	4	16	$ 1,350	Sept 07 Dec 08	
a6	Clara Barton	Analysis	7	C	5	1	12	$ 400	Jan 05-Dec0 5	

Model Method Options

- Linear Regression
- Random Forest Prediction

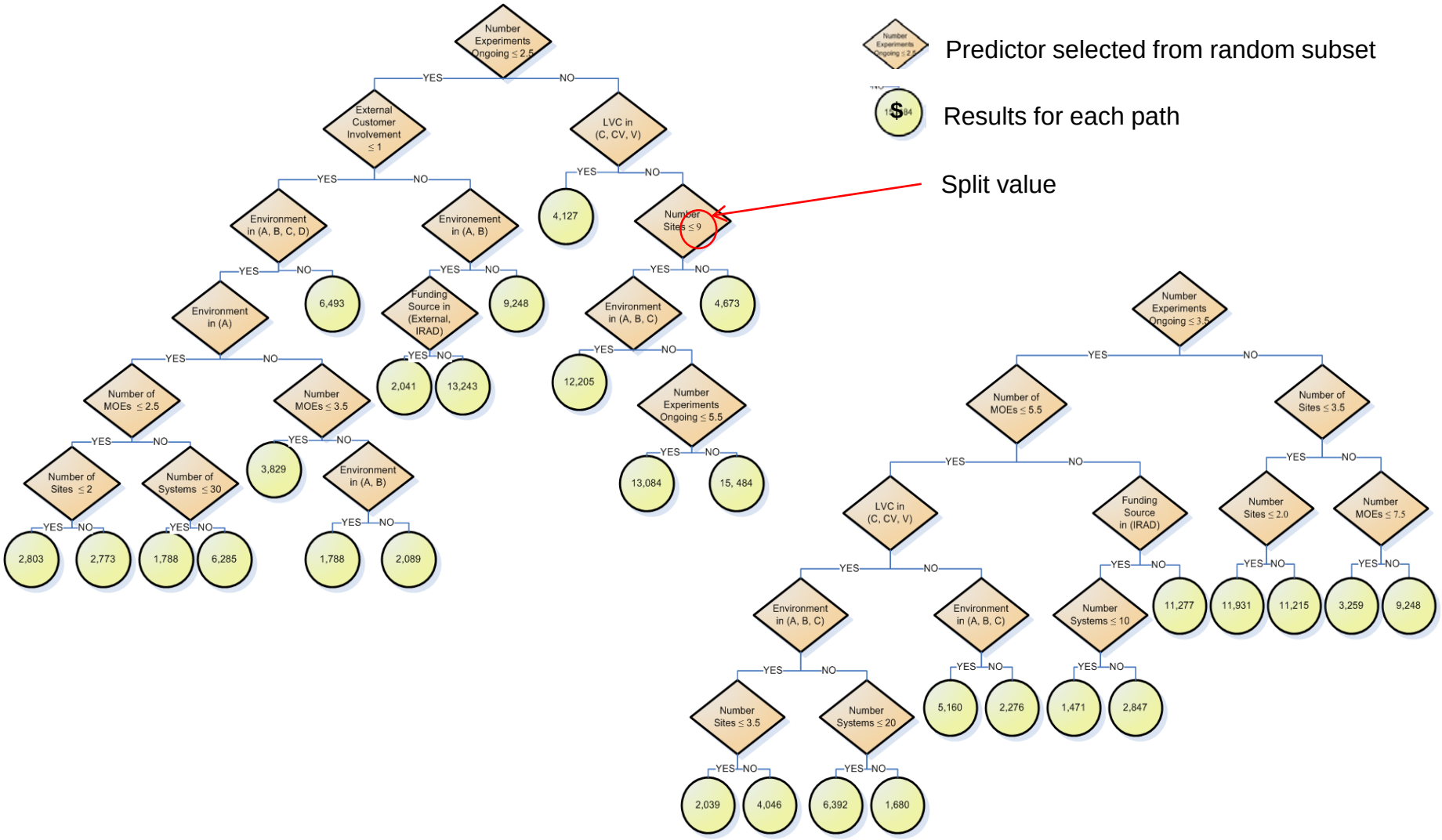




*Not required

Random Forest selected for Co$t-X model

Two Experimentation Trees Produced



Randomness in Predictor Nodes and Split Values

Experimentation Prediction Analysis Results

Random Forest Model developed based on 70 data points

- Adjusted R^2 (90%) based on training data (optimistic)
- Predicted R^2 (63%) based on OOB data (more realistic)

Compared Random Forest Results to Actual costs

- For 34 newly completed projects
- Goal: predict within 20% of actual costs

Results of Random Forest Predictions …

Lower than Actual Costs

Well within Goal!

RFP Application #3: Logistics Transport Model

Objective of Logistics Transport Cost Model

- Determine best locations to manufacture products and parts based on shipping considerations

Analysis Approach

- Data
  - Thousands and thousands of data points
  - Messy data: missing values, lots of potential predictors
- Initial Plan: Multivariate Regression
  - Very cumbersome and required manual partitioning into suitable subsets
- Chosen method: Random Forest Prediction
  - No need to clean up the data - Automatic partitioning / different perspectives
  - Very easy to implement, execute, and analyze

Random Forest Prediction implemented via R Programming Language

Logistics Transport Cost Model

Data Description

- Consists of 133K data points
- Automatically separated into two distinct data sets
  - Domestic with ~ 87K data points
  - International with ~ 25K data points

Potential Predictors

- Started with 16 potential predictors
- Reduced to 3 key predictors
  - Mode of transportation
  - Origin (country or state)
  - Bill weight

Random Forest Prediction facilitates big data analysis

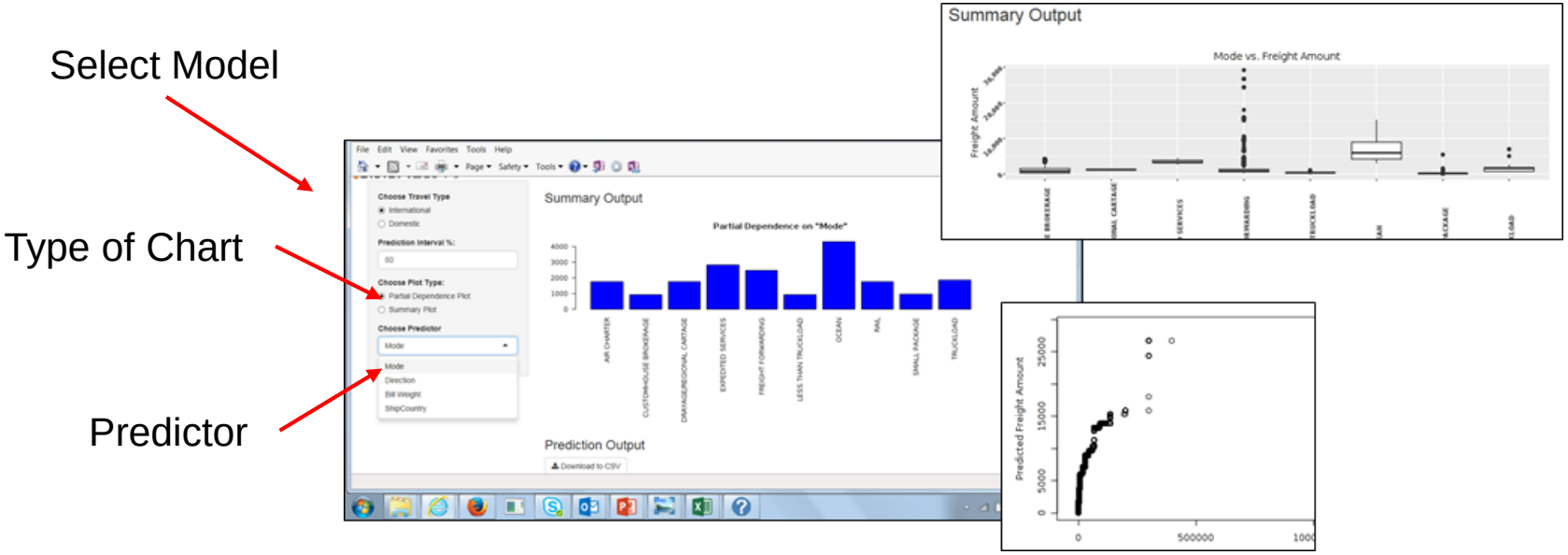
Analytical Results

Goodness of fit – Predicted R^2

- International: 0.76 → 0.83
- Domestic: 0.75 → 0.86

Graphical Interpretations

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



Analysis made easy with R Shiny Package

Challenges for Cost Analysis Community

Machine Learning for cost analysis & estimating

- Different from traditional methods
- Black box method
- Not so easy to interpret
- Not so easy to follow input-to-output logic
- Predicts a numeric value (cost)
- Not a parametric equation (CER)

Do Benefits outweigh Challenges?

Questions?