The Evolution of Regression Modeling

Cheryl Pammer
Statistician, Technical Trainer/User Experience Designer
Minitab Inc.

May 24, 2018

Minitab®

Learning Objectives

- What Is Machine Learning?
- Basic Machine Learning Algorithms
- Moving From Regression to Regression Trees
- Why Teach Regression Trees

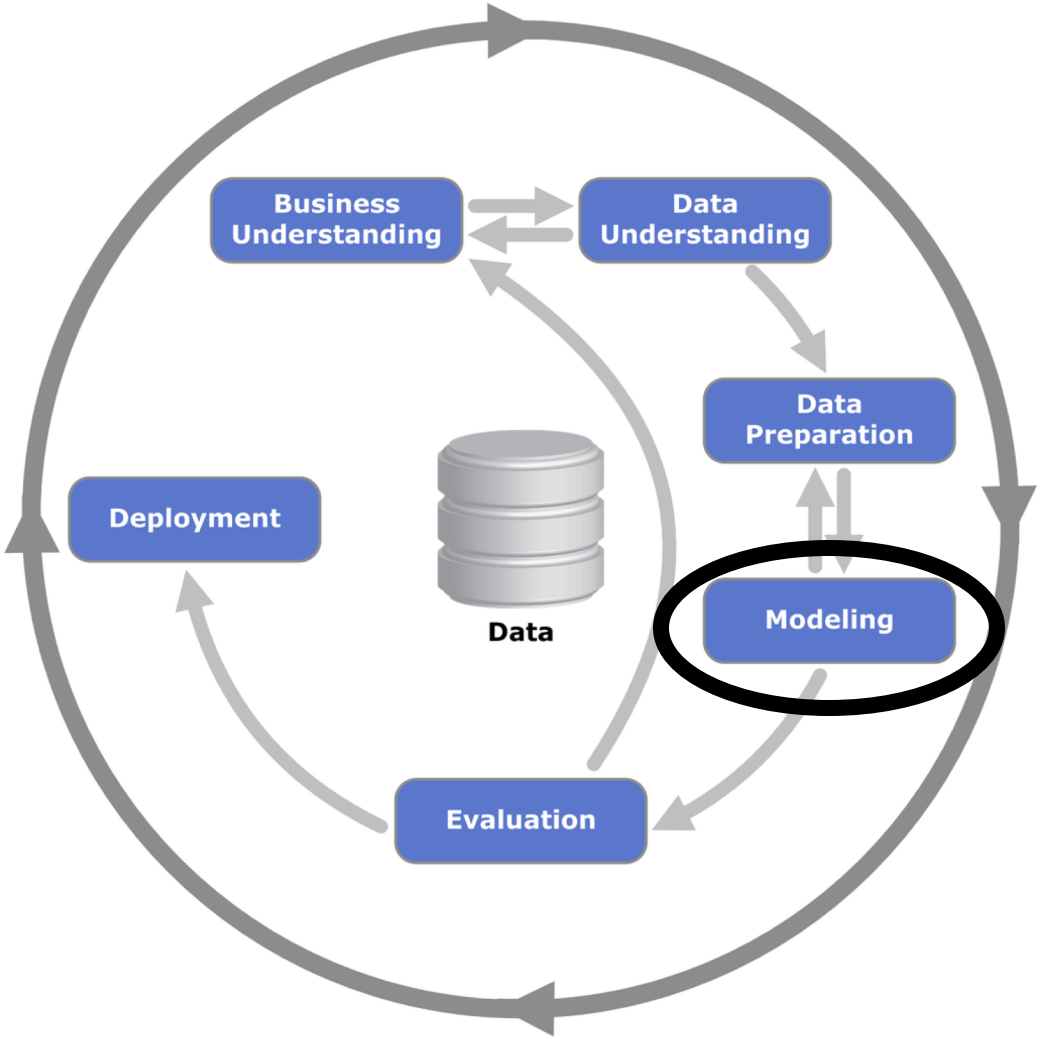
Introduction to Machine Learning

A machine learning algorithm “teaches” a computer to recognize patterns using available data.

Data is usually split into a training set and a test set:

- Training (or learn) = creates model
- Test = assesses model performance

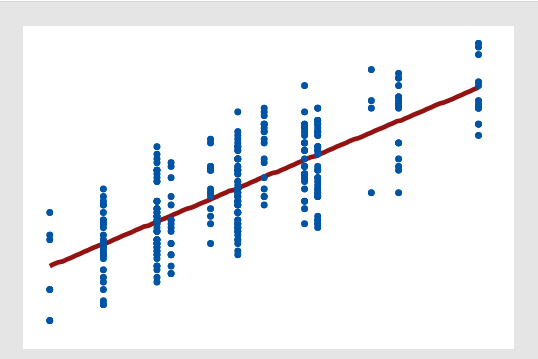
CRISP-DM

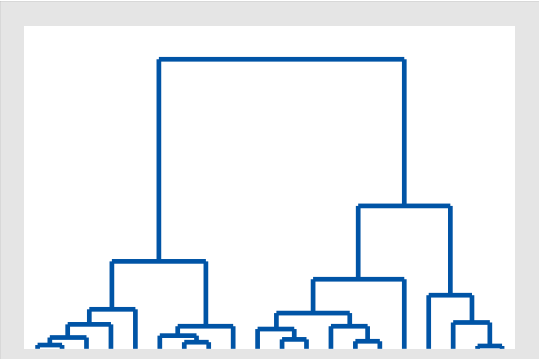


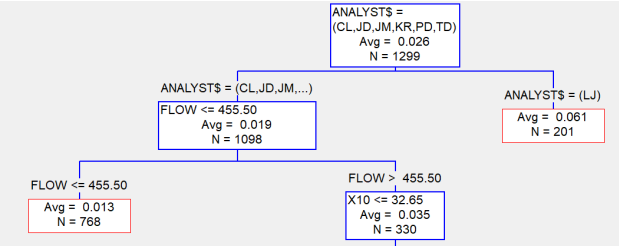
Cross-**i**ndustry **s**tandard **p**rocess for **d**ata **m**ining

Basic Machine Learning Algorithms

Type	**Tools**
Supervised (Y and X's)	Regression, Logistic Regression CART, Random Forests
Unsupervised (Only X's)	K-Means Clustering, Hierarchical Clustering Principal Components Analysis







Who Survived the Titanic?

Target Variable: Survival (0 = No, 1 = Yes)

Predictor Variables:

- Pclass (1st, 2nd, or 3rd class)
- Sex
- Age
- Sibsp (# of siblings/spouses aboard)
- Parch (# of parents/children aboard)
- Fare
- Embarked (C=Cherbourg, Q=Queenstown, S=Southhampton)

Binary Logistic Regression

Model the relationship between predictors and a response with two outcomes (Survived/Died).

Model (logit link function)

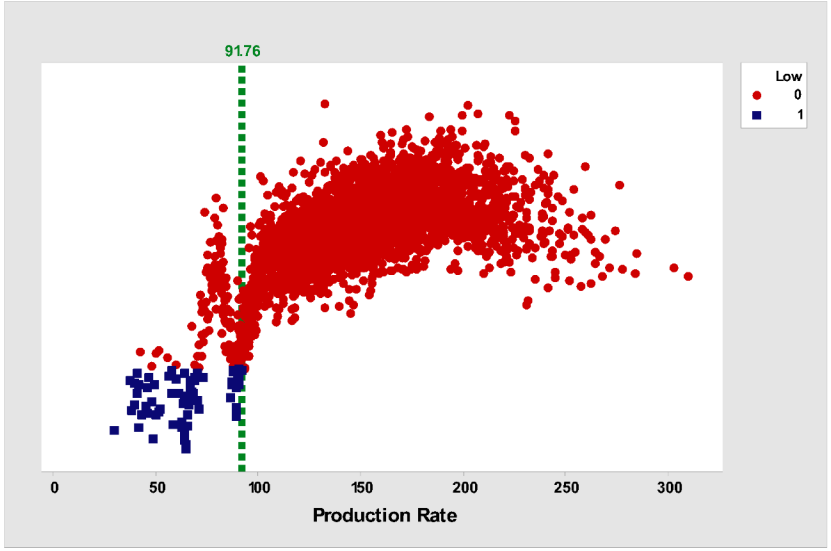
$$\text{Log}_e[p/(1-p)] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

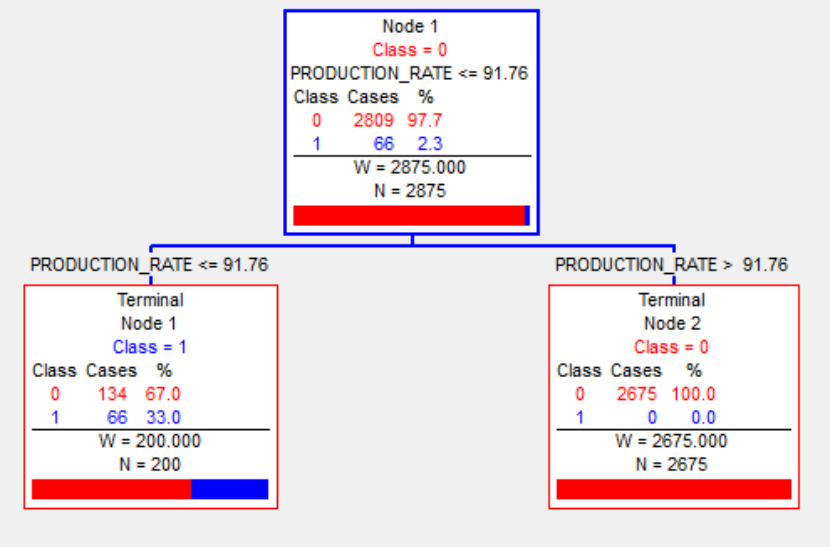
Regression Challenges

Regression and logistic regression often don't work well, particularly with larger observational data sets:

- Everything is significant.
- Determination of predictors to include in model is challenging.
- Relationships are nonlinear.
- Complex interactions exist.
- Extreme outliers exist.
- Many missing values.

Classification and Regression Tree (CART)





- Decision trees quickly find the X's that best partition the data into distinct groups relative to Y.
- Y can be continuous or categorical.

Classification and Regression Trees

- Nonparametric machine learning for classification and regression.
- Stepwise procedure in which predictors enter the model one at a time.
- Procedure works by carving a high-dimensional data space into small to moderate set of regions.
- A prediction is produced for each region.
- Early data mining tool, developed between 1974-1983.

Leo Breiman, Jerome Friedman, Richard Olshen, Charles Stone. (1984) Classification and Regression Trees.

Regression Trees in Basic Stat Classes

- Classes in machine learning algorithms typically start with regression and build from there.
- Traditional statistics classes can easily extend from regression/logistic regression into regression trees.
- As data sets grow larger, regression trees can be an important part of a data analyst's tool kit.
- Many professionals entering the workforce will encounter data scientists. They will benefit by having some knowledge around machine learning algorithms and terminology.

Thank you!

► Feel free to contact me:

Cheryl Pammer
cpammer@Minitab.com

► To try Salford Predictive Modeler or Minitab, contact Christine Bayly, Academic Sales Manager:

Cbayly@minitab.com
800-448-3555 x 3304