Motivations for Predictive Analytics & Machine Learning

HI 743

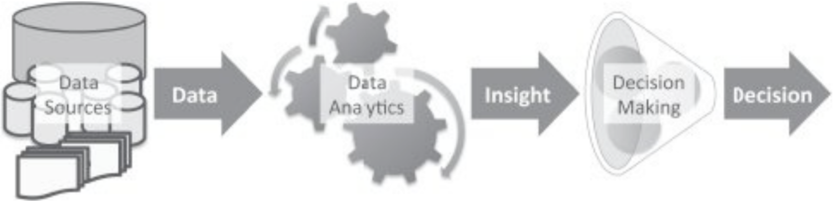
Ryan Gallagher

Department of Health Informatics and Administration
Zilber College of Public Health
University of Wisconsin - Milwaukee

January 30, 2025

Machine Learning for Predictive Analytics

Modern organizations collect massive amounts of data. For data to be of value to an organization, they must be analyzed to **extract insights that can be used to make better decisions**. Extracting insights from data is the job of data analytics.



What is Predictive Analytics?

Predictive data analytics is the art of building and using models that make predictions based on patterns extracted from historical data. Applications of predictive data analytics include:

- **Dosage Prediction**: Doctors and scientists frequently decide how much of a medicine or other chemical to include in a treatment. Predictive analytics models can be used to assist this decision making by predicting optimal dosages based on data about past dosages and associated outcomes.
- **Diagnosis**: Doctors, engineers, and scientists regularly make diagnoses as part of their work. Typically, these diagnoses are based on their extensive training, expertise, and experience. Predictive analytics models can help professionals make better diagnoses by leveraging large collections of historical examples at a scale beyond anything one individual would see over his or her career. The diagnoses made by predictive analytics models usually become an input into the professional's existing diagnosis process.

What is Predictive Analytics?

- **Document Classification**: Predictive data analytics can be used to automatically classify documents into different categories. Examples include email spam filtering, news sentiment analysis, customer complaint redirection, and medical decision making. In fact, the definition of a document can be expanded to include images, sounds, and videos, all of which can be classified using predictive data analytics models.

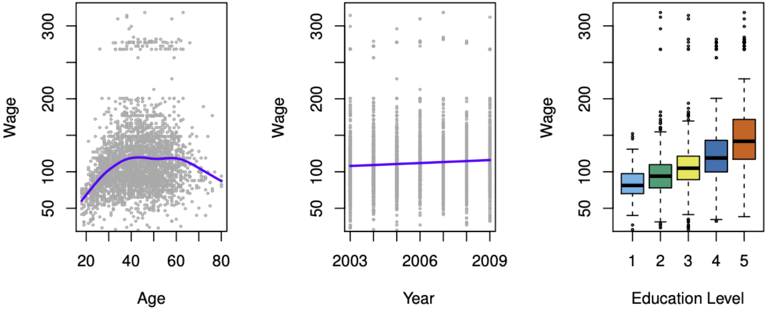
In each case a model is used to make a prediction to help a person or organization make a decision. In predictive data analytics **we use a broad definition of the word prediction**. In everyday usage, the word prediction has a temporal (time) aspect—we predict what will happen in the future. However, **in data analytics a prediction is the assignment of a value to any unknown variable**. The examples listed above are trained to make predictions based on a set of historical examples. We use machine learning to train these models.

What is Machine Learning?

- Machine learning is defined as an **automated process that extracts patterns from data**. To build the models used in predictive data analytics applications, we use supervised machine learning.
- Supervised machine learning involves building a statistical model for predicting, or estimating, an *output* based on one or more *inputs*.
- With unsupervised machine learning, there are inputs but *no supervising output*; nevertheless we can learn relationships and structure within the data.

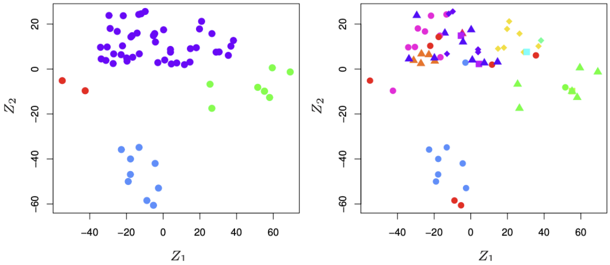
Application Example: Wage Data

In our text, we're given the Wage dataset which examines a number of factors that relate to wages for a group of men from the Atlantic region of the United States. We wish to understand the association between an employee's age and education, as well as the calendar year, on his wage.



Application Example: Gene Expression

Also in our text, we're given the NCI60 data set, which consists of 6,830 gene expression measurements for each of 64 cancer cell lines. Instead of predicting a particular output variable, we are interested in determining whether there are groups, or clusters, among the cell lines based on their gene expression measurements.



How Does Machine Learning Work?

- Machine learning algorithms work by searching through a set of possible prediction models for the model that best captures the relationship between the *descriptive features* and *target feature* in a dataset.
- A prediction model that makes the correct predictions for these queries captures the underlying relationship between the descriptive and target features and is said to **generalize** well.
- The goal of machine learning is to find the predictive model that generalizes best by using some criteria for choosing among the candidate models it considers during its search.

How Does Machine Learning Work?

- The criterion to select the best prediction model is why there are a lot of different machine learning algorithms. When we choose to use one machine learning algorithm, we are choosing to use one model selection criterion over another.

- All the different model selection criteria consist of a set of assumptions about the characteristics of the model that we would like the algorithm to induce. The set of assumptions that defines the model selection criteria of a machine learning algorithm is known as the **inductive bias** of the machine learning algorithm.

How Does Machine Learning Work?

- There are two types of **Inductive Bias**:
    - **Restriction Bias**: constrains the set of models allowed for the algorithm to consider.
    - **Preference Bias**: guides the algorithm to prefer certain models over others.

In summary, machine learning works by searching through a set of potential models to find the prediction model that best generalizes beyond the dataset. Machine learning algorithms use two sources of information to guide this search, the training dataset and the inductive bias assumed by the algorithm.

How Does Machine Learning Work?

- **Prediction:** In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained.

  $$\hat{Y} = \hat{f}(X)$$

  where $\hat{f}$ represents our estimate for f, and $\hat{Y}$ represents the resulting prediction for Y.

- The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities, which we will call the *reducible error* and the *irreducible error* (or *noise*).

How Does Machine Learning Work?

- Consider a given estimate $\hat{f}$ and a set of predictors X, which yields the prediction $\hat{Y} = \hat{f}(X)$. Assume for a moment that both $\hat{f}$ and X are fixed, so that the only variability comes from ϵ.

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{Var(\epsilon)}_{\text{Irredicible}} \end{aligned}$$

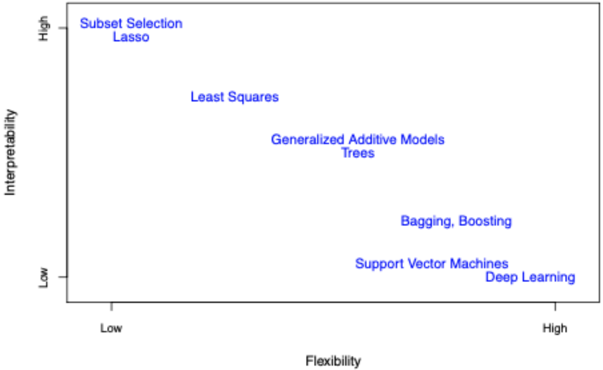
- Machine Learning focuses on techniques for estimating f with the aim of **minimizing the reducible error**. It is important to remember that the irreducible error will always maintain an upper bound on the accuracy of our prediction for Y. This bound is almost always unknown in practice.

How Does Machine Learning Work?

- **Inference**: We still wish to estimate f, but are not necessarily trying to make predictions for Y. In this setting, we might be interested in answering the following questions:
    - *Which predictors are associated with the response?*
      Identifying the few **important** predictors among a large set can be extremely useful.
    - *What is the relationship between the response and each predictor?*
      Some predictors may have positive/negative relationships with Y.
    - *Can the relationship between Y and each predictors be adequately summarized using a linear equation, or is it more complex?*

The Accuracy & Interpretability Trade-Off

- When inference is the goal we would prefer using simple and relatively inflexible machine learning methods. In some settings, however, we may only be interested in prediction, and the interpretability of the predictive model is simply not of interest.



Assessing Model Accuracy

- As previously mentioned, model selection consists of a set of assumptions where we seek to find the best fit for the data & our question. Selecting the best approach can be one of the most challenging parts of performing statistical learning in practice.
- A classic quantitative approach to measuring model-fit in the **regression** setting is *mean squared error* (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

In regression settings, we can fit a model to the *training data* and assess then test the model by obtaining the the *test MSE*. This will be explored in our Error Based Machine Learning module.

Assessing Model Accuracy

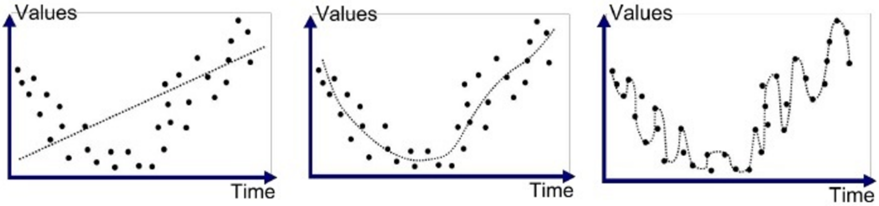
- A quantitative approach to measuring model-fit in the **classification** setting is the *error rate*:

$$ER = \frac{1}{n}\sum_{i=1}^{n}(y_i \neq \hat{y}_i))^2$$

In the classification settings, this is a simple proportion statistic that measures how observations where were incorrectly classified using the proposed model. This will be explored further in our Classification module.

What Can Go Wrong with Machine Learning?

- There are two kinds of mistakes that an inappropriate inductive bias can lead to:
  - **Underfitting**: occurs when the prediction model selected by the algorithm is too simplistic to represent the underlying relationship in the dataset between the descriptive features and the target feature.
  - **Overfitting**: occurs when the prediction model selected by the algorithm is so complex that the model fits to the dataset too closely and becomes sensitive to noise in the data.



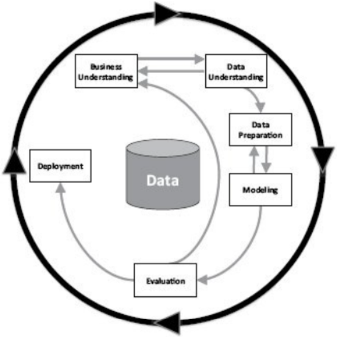
Underfitted

Good Fit/Robust

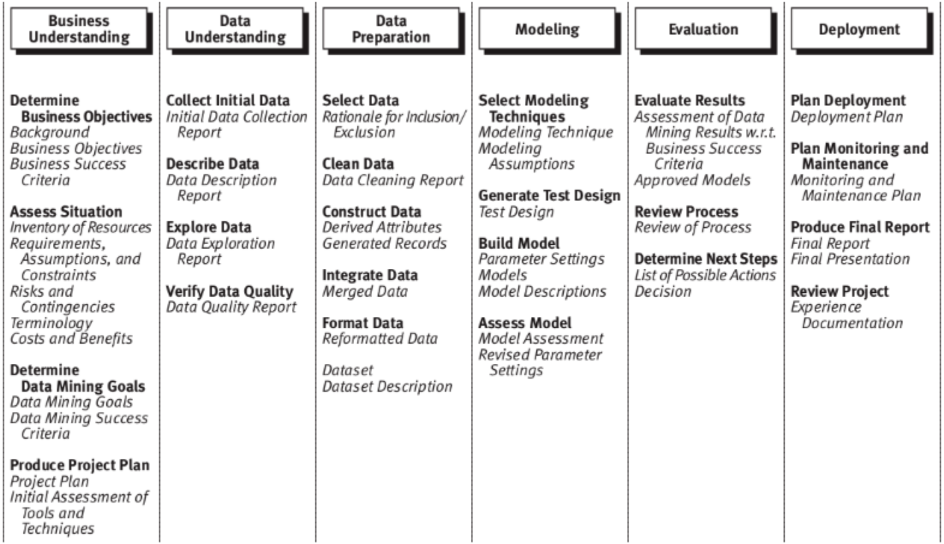
Overfitted

Predictive Analytics Project Lifecycle: CRISP-DM

- Our text describes a widely adopted predictive analytics project framework, **CRIPS-DM**: CRoss Industry Standard Process for Data Mining.
- This framework is non-proprietary and industry/application neutral. It's a good way to look at a project from an objectively analytical point of view. (Methodology Handbook Link)



Predictive Analytics Project Lifecycle: CRISP-DM



Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Assess Situation
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Select Data
Rationale for Inclusion/ Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Dataset
Dataset Description

Modeling

Select Modeling Techniques
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Descriptions

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

Predictive Analytics Tools

There are multiple applications which are capable of importing, preparing, modeling, and interpreting data.

- **Application-based Tools**: IBM SPSS, Knime Analytics, SAS Enterprise, Weka, Excel
- **Programming Languages**: *R*, Python, SAS, SQL

This course will be partial towards programming languages, particularly *R*. *R* is robust, well maintained, and highly adopted within the data analytics industry.

For Next Time: Download & Install R

- **Install R**: Follow the installation instructions at https://cran.rstudio.com/
  - **Windows Users:** Follow Download R for Windows link → Download base → Follow installation wizard.
  - **macOS Users:** Follow Download R for macOS → Select .pkg for your processor type (Apple Silicon vs. Intel) → Follow installation wizard.
- **Install RStudio**: Instruction at https://posit.co/download/rstudio-desktop/
  - This website should detect your operating system. Big blue button should say `Download Rstudio Desktop for {Your OS}`.

(Windows Install Tutorial) (macOS Install Tutorial)