Introduction to Data Analytics

Edward (Jed) Frees
University of Wisconsin-Madison, Australian National University

Joint work with Hirokazu (Iwahiro) Iwasawa

16 November 2024

Outline

1. Elements of Data Analytics

In 1962, statistician John Tukey defined data analysis as:

> *procedures for analyzing data,*
> *techniques for interpreting the results of such procedures,*
> *ways of planning the gathering of data to make its analysis easier,*
> *more precise or more accurate, and*
> *all the machinery and results of (mathematical) statistics which apply to analyzing data.*

Key Data Analytic Concepts

Underpinning the elements of data analytics are:

- **Data Driven**. Conclusions and decisions made through a data analytic process depend heavily on data inputs.
  - In comparison, econometricians have long recognized the difference between a data-driven model and a structural model.
- **EDA** - exploratory data analysis - and **CDA** - confirmatory data analysis.
  - The purpose of EDA is to reveal aspects or patterns in the data without reference to any particular model.
  - CDA techniques use data to substantiate, or confirm, aspects or patterns in a model.
- **Estimation** and **Prediction**.
  - Medical statisticians test the efficacy of a new drug and econometricians estimate parameters of an economic relationship.
  - In insurance, predictions of yet to be realized random outcomes are critical for financial risk management (e.g., pricing) of existing risks in future periods.

Additional Key Data Analytic Concepts

- **Model Complexity, Parsimony,** and **Interpretability**.
  - Other things being equal, a model with fewer parameters is said to be *parsimonious* and hence less complex.
  - Complexity hinders our ability to understand the inner workings of a model and its interpretability.
- **Parametric** and **Nonparametric** models.
  - Parametric and nonparametric approaches have different strengths and limitations; neither is strictly better than the other.
- **Robustness** means that a model, test, or procedure is resistant to unanticipated deviations in model assumptions or the data used to calibrate the model.
- **Computational Statistics**.
  - Ideas of subsampling and resampling data (e.g., through cross-validation and bootstrapping) have introduced new methods for understanding statistical sampling errors and a model's predictive capabilities.

Additional Key Data Analytic Concepts

- **Big Data**.
- Examples of big data include text documents, videos, and audio files that are also known as *unstructured* data.

Analytic Trends

Data Sources	**Algorithms**
Mobile devices Auto telematics Home sensors (Internet of Things) Drones, micro satellites	Statistical learning Artificial intelligence Structural models
Data	**Software**
Big data (text, speech, image, video) Behavioral data (including social media) Credit, trading, financial data	Text analysis, semantics Voice recognition Image recognition Video recognition
Source : Stephen Mildenhall, Personal Communication	

2. Data Analysis Process

Data Analysis Process for Insurance Activities

I. Scoping Phase	II. Data Splitting	III. Model Development	IV. Validation	V. Determine Implications
Use background knowledge and theory to define goals	Split the data into training and testing portions	Select a candidate model	Repeat Phase III to determine several candidate models	Use knowledge gained from exploring the data, fitting and predicting the models to make data-informed statements about the project goals
Prepare, collect, and revise data		Select variables to be used with the candidate model	Assess each model using the testing portion of the data to determine its predictive capabilities	
EDA Explore the data		Evaluate model fit using training data Use deviations from model fit to improve suggested models		

I. Scoping Phase

I. Scoping Phase

Scoping, or problem formulation, can be divided into three components:

- **Use background knowledge and theory to define goals**.
  - For insurance, background knowledge includes market conditions and theory may include a person's attitude towards risk-taking.
- **Prepare, collect, and revise data**. Getting the right data that gives insights into questions at hand is typically the most time-consuming aspect of most projects.
- **EDA** - Exploring the data, without reference to any particular model, can reveal unsuspected aspects or patterns in the data.

These three components can be performed *iteratively*.

II. Data Splitting

- If the available dataset is sufficiently large, one can split the data into a portion used to calibrate one or more candidate models, the training portion, and another portion that can be used for testing, that is, evaluating the predictive capabilities of the model.
- The data splitting procedure guards against overfitting a model and emphasizes predictive aspects of a model.
- It is common to use data from an earlier time period to predict, or *forecast*, future behavior.

III. Model Development

As with the scoping phase, developing a model is an iterative procedure.

- **Select a candidate model.** One starts with a model that, from the analyst's perspective, is a likely "candidate" to be the recommended model.
- **Select variables to be used with the candidate model.** Many (if not most) situations deal with multivariate outcomes. Analysts give a great deal of thought as to which variables are considered inputs to a system and which variables can be treated as outcomes.
- **Evaluate model fit on training data.** Many measures of model fit are available.
- **Use deviations from the model fit to suggest improvements to the candidate model.** In regression analysis, this tactic is known as *diagnostic checking.*

IV. Validation

- **Repeat Phase III to determine several candidate models.** It is customary to narrow the field of candidates down to a handful based on their fit to the training data.
- **Assess each model using the testing portion of the data to determine its predictive capabilities.**
  - Each fitted model is used to make predictions with the predicted outcomes compared to the held-out test data.
  - This comparison may also be done using cross-validation.

V. Determine Implications

The relative importance of interpretability depends on the project goals.

- For example, a model devoted to enticing potential customers to view a webpage can be judged more on its predictive capabilities.
- In contrast, a model that provides the foundations for insurance prices typically undergoes scrutiny by regulators and consumer advocacy groups; here, interpretation plays an important role.

3. Single Variable Analytics

- We can describe much of the data analysis process in the context of just **one** variable
- Start by thinking about the type of variable
  - Qualitative or categorical?
    - Binary?
    - Ordinal or nominal (ordered or unordered)?
  - Continuous or Discrete?
    - Count or interval?
  - Loss data
    - Often, a combination of discrete and continuous components
    - Many zeros reflect no insured loss

Exploratory versus Confirmatory

Comparison of Exploratory Data Analysis and Confirmatory Data Analysis

	EDA	**CDA**
Data	Observational data	Experimental data
Goal	Pattern recognition, formulate hypotheses	Hypothesis testing, estimation, prediction
Techniques	Descriptive statistics, visualization, clustering	Traditional statistical tools of inference, significance, and confidence

Model Construction

- Parametric or nonparametric
  - Because nonparametric methods make fewer assumptions, they can be more flexible, more robust, and more applicable to non-quantitative data.
  - A drawback of nonparametric methods is that it is more difficult to extrapolate findings outside of the observed domain of the data, a key consideration in predictive modeling.

Model Construction - Explanation versus Prediction

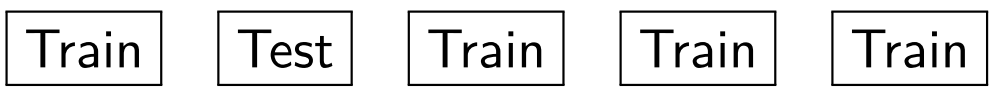
- In some scientific areas such as economics, the focus of data analysis is to explain the causal relationships between the input variables and the response variable.
- In other scientific areas such as natural language processing and actuarial science, the focus of data analysis is to predict what the responses are going to be given input variables
- **Predictive Modeling** - the goal is to predict new observations is the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations.
    - Predictions include point predictions, interval predictions, regions, distributions, and rankings of new observations.

Model Selection

- Traditionally, statistics has focused on *hypothesis testing* and *goodness of fit* for model selection.
  - Is a variable "statistically significant?"
  - Goodness of fit measures summarizes how well the model fits the data
    - Special cases include regression R^2 statistics and
    - Information criteria such as Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC)
- Modern approaches, based on data-splitting, attempt to emulate the classical "scientific method"
  - Formulate a theory
  - Gather data (evidence) to test the theory
  - Examine the theory in light of the data. Reject or modify the theory.
  - Repeat

Data Splitting

- Split a data set into
  - the *training* portion - used to calibrate a model, , and
  - the *test* portion - to quantify the predictive power of the model
- This approach is robust in the sense that it does not rely on any distributional assumptions and can be used to validate models.
- However, it does introduce additional variability into the process
  - which observations fall into the training and testing portions?
- To mitigate this problem, use *cross-validation.*
  - Randomly partitions dataset into five subsets



  - Use datasets 1, 3, 4, 5 to estimate the model. Assess predictive power of the model using dataset 2.
  - Repeat, using different selections of training versus test data.
  - The average of the comparisons results in a *cross-validation statistic*.

Example. Cross-Validation and Under- and Over-Fitting a Model

- We have 100 observations of a loss y where we have tracked outcomes using a factor with 6 levels. Denote outcomes as y_{ij} with mean μ_j.
- The true model is that $\mu_1 = \mu_2$ and $\mu_3 = \mu_4 = \mu_5 = \mu_6$.
- For the analysis, let us consider 3 models
  - Community-Rating: $y_{ij} = \mu + \epsilon_{ij}$
  - Six Levels: $y_{ij} = \mu_j + \epsilon_{ij}$, $j = 1, \ldots, 6$
  - Two Levels: $y_{ij} = \mu_A + \epsilon_{ij}$, $j = 1, 2$, $y_{ij} = \mu_B + \epsilon_{ij}$, $j = 3, 4, 5, 6$
- Process. Randomly split the data into five "folds". Use four folds to predict the fifth.

Table 1: Under- and Over-Fitting of Models

	Community Rating	Two Levels	Six Levels
Rmse - Fold 1	1.318	1.192	1.239
Rmse - Fold 2	1.034	0.972	1.023
Rmse - Fold 3	0.816	0.660	0.759
Rmse - Fold 4	0.807	0.796	0.824
Rmse - Fold 5	0.886	0.539	0.671
Rmse - Average	0.972	0.832	0.903
AIC - Average	227.171	206.769	211.333

Two Levels model a clear winner followed by the Six Levels model.

Code

```
{
# Generate the Data
rmse <- Metrics::rmse
n <- 100
set.seed(1234)
u <- sample(6, n, replace = TRUE)
x1 <- as.factor((u == 4) + (u ==5)   )
x2 <- as.factor(u)
y <- 1* (x1==1) + rnorm(n, sd = 1)
xyData <- data.frame(x1, x2, y)

n <- nrow(xyData)
set.seed(123)

# Number of folds
k <- 5
splt <- split(sample(n), 1:k)
Rmse.mat <- matrix(0, nrow=k, ncol=3) -> AIC.mat
for (i in 1:k) {
  test.id<- splt[[i]]
  test   <- xyData[test.id, ]
  train  <- xyData[-test.id, ]
  model0 <- lm(y ~ 1, data = train)
  model1 <- lm(y ~ x1, data = train)
  model2 <- lm(y ~ x2, data = train)

  Rmse.mat[i,1] <- rmse(test$y, predict(model0, test))
  Rmse.mat[i,2] <- rmse(test$y, predict(model1, test))
  Rmse.mat[i,3] <- rmse(test$y, predict(model2, test))
  AIC.mat[i,1]  <- AIC(model0)
  AIC.mat[i,2]  <- AIC(model1)
  AIC.mat[i,3]  <- AIC(model2)
}
OutMat <- rbind(round(Rmse.mat, digits=3),
                round(colMeans(Rmse.mat), digits=3),
```

4. Analytics with Many Variables

With many variables, the process is the same. But, the potential applications become much richer.

Supervised and Unsupervised Learning

- With many variables, we have the opportunity to think about some of them as "inputs" and others "outputs" of a system.
- Models based on input and output variables are known as **supervised learning methods** or as **regression methods.**
  - When the target variable is a categorical variable, supervised learning methods are called **classification methods**.
- **Unsupervised learning methods** - data are treated the same and there is no artificial divide between "inputs" and "outputs."

Common Names of Different Variables used in Regression Methods

Target Variable	Explanatory Variable
Dependent variable	Independent variable
Response	Treatment
Output	Input
Endogenous variable	Exogenous variable
Predicted variable	Predictor variable
Regressand	Regressor

Algorithmic Modeling

- Idea underpinning Algorithmic Modeling
  - : One variable, Y, is determined to be a target variable.
  - Other variables, $X_1, X_2, \ldots, X_p$, are used to understand or explain the target Y.
  - Goal - determine an appropriate function $f(\cdot)$ so that $f(X_1, X_2, \ldots, X_k)$ is a useful predictor of Y.
- Classic Special Case: Linear Regression with functions

$$f(x_{i1}, \ldots, x_{ik}) = \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = \mathbf{x}_i' \boldsymbol{\beta}.$$

- **Algorithmic Modeling Culture.**
- Emphasizes algorithmic fitting particularly in complex problems such as voice, image, and handwriting recognition.
- Algorithmic methods are especially useful when the goal is prediction.
- Does emphasize the distribution of outcomes
- In addition to linear regression, algorithmic fitting methods include ridge and lasso regression, as well as regularization methods.

Data Modeling

- One way to motivate an algorithmic development is through the use of a data model.
- With a "probability" or "likelihood" based model, in that our main goal is to understand the target (Y) distribution, typically in terms of the explanatory variables.
    - Data models are particularly useful for the goal of explanation.

Algorithmic Modeling Fitting Methods

Many of these algorithms take an approach similar to linear regression. As examples, other widely used algorithmic fitting methods include ridge and lasso regression, as well as regularization methods.

5. Data

See the online Chapter Two of Loss Data Analytics, Edition Two for a dicussion of data considerations in terms of

- data types,
- data structure and storage,
- data cleaning,
- big data issues, and
- ethical issues.