Middle Tennessee State University
Department of Mathematical Sciences

DATA 3550–
Applied Predictive Modeling

Instructor: Dr. Ramchandra Rimal

Chapter 1: General Strategies

Chapter 1: Overview and General Strategies for Data Preprocessing

Objectives:

- **Introduction**
- **Notations and Terminologies in Predictive Modeling**
- **Understanding the Predictive Modeling Process**
- **Study the Data Preprocessing Techniques**
  - **Learn data Transformation Techniques**
  - **Learn the strategies to deal with Missing Data**
  - **Identify the strategies on Adding or Removing Predictors**

Introduction

What is Statistics?

Statistics teaches us how to make intelligent judgments and informed decisions in the presence of uncertainty and variation

Statistics is a science that deals with the collection, presentation, analysis, and interpretation of data.

In statistics, we generally want to study a population.

What is Data Science?

Data science is the study of data to extract meaningful insights for business. It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data. This analysis helps data scientists to ask and answer questions like what happened, why it happened, what will happen, and what can be done with the results. -Amazon

Data science continues to evolve as one of the most promising and in-demand career paths for skilled professionals. Today, successful data professionals understand that they must advance past the traditional skills of analyzing large amounts of data, data mining, and programming skills. In order to uncover useful intelligence for their organizations, data scientists must master the full spectrum of the data science life cycle and possess a level of flexibility and understanding to maximize returns at each phase of the process. -UC Berkley Data Science

Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization's data. These insights can be used to guide decision making and strategic planning. -IBM

Notations and Terminologies

Population: A collection of persons, items, or objects under study
Sample: A subset of the population
Census: The desired information for all objects in the population
Sampling: The way to select a portion (or subset) of the population to study that portion (the sample) to gain information about the population
Data: The outcome of sampling from a population

A good sample should have the **same characteristics as the population** it is representing.

Types of sampling:
Simple random sampling: selecting a sample from the population in such a way that each sample of the same size has an equal chance of being selected

Stratified sampling: separating the population units into nonoverlapping groups and taking a random sample from each one

Convenience sampling: A type of sampling that is non-random and involves using results that are readily available

Variable: Any characteristic whose value may change from one object to another in the population

Variable Types:

Numerical: Take a numerical value

(i) A numerical variable is **discrete** if its set of possible values either is finite or else can be listed in an infinite sequence (one in which there is a first number, a second number, and so on).

(ii) A numerical variable is **continuous** if its possible values consist of an entire interval on the number line.

Categorical: Takes a categorical value: Categorical data, otherwise known as nominal, attribute, or discrete data, take on specific values that have no scale

A **univariate** data set consists of observations on a single variable.

Examples:

A **bivariate** data set consists of observations made on each of two variables.

Examples:

Multivariate data set consists of observations made on more than one variable.

Examples:

Branch of Statistics:

a) **Descriptive Statistics**

(1) Visual display
(2) Numerical summary

b) **Inferential Statistics**: Statistical Inference is the process of making conclusions using data that is subject to random variation.

Inferential statistics, **deals with making inferences and predictions about a population from a sample**. This is done by using statistical models and methods to estimate population parameters from sample statistics. The goal of inferential statistics is to make generalizations about a population based on a sample of data, and to estimate the level of uncertainty associated with these generalizations.

Predictive Modeling: the process of developing a mathematical tool or model that generates an accurate prediction

Predictive modeling, is used to **make predictions about future events or outcomes using available information**. It involves using statistical techniques to develop models that can be used to predict future outcomes based on historical data.

Example: **predict the likelihood of a customer defaulting on a loan, or forecasting future sales**.

Importance: Predictive modeling empowers organizations to leverage their data for **strategic decision-making, risk management, and operational efficiency** across diverse domains. Its applications are broad and varied, making it a valuable tool in the modern data-driven era.

Limitations: Our ability to make decisions is constrained by our present and past knowledge

Criticism: Rodriguez(2011) says, Predictive modeling, the process by which a model is created or chosen to try to best predict the probability of an outcome has lost credibility as a forecasting tool.

Why Predictive models fail?

The common reasons are the following:
1) **Inadequate prepossessing of data**
2) **Inadequate model validation**
3) **Unjustified extrapolation**
4) **Over-fitting the model to the existing data**

Predictive modeling (or empirically driven modeling) is not a substitute for an expert's intuition(or experience-driven modeling), but rather a complement.

- neither data-driven models nor the expert relying solely on intuition will do better than a combination of the two

Trade-Off between Prediction and Inference

The **primary interest** of predictive modeling is to **generate accurate predictions**, a **secondary interest** may be to **interpret the model and understand why it works**.

As we push towards **higher accuracy**, models become more **complex** and their interpretability becomes more difficult.

Parsimony (or simplicity) is a key consideration. **Simple models are generally preferable** to complex models, especially when **inference** is the goal.

However, **accuracy should not be seriously sacrificed for the sake of simplicity**.

Terminology and Notations used in Textbook

- Sample, data point, observation, or instance refer to a single, independent unit of data, such as a customer, patient, or compound
- Training set consists of the data used to develop models
- Validation sets are used solely for evaluating the performance of a final set of candidate models
- Predictors, independent variables, attributes, or descriptors are the data used as input for the prediction equation
- Outcome, dependent variable, target, or response refer to the outcome event or quantity that is being predicted
- Model building, model training, and parameter estimation all refer

to the process of using data to determine values of model equations

n = the number of data points
P = the number of predictors
y_i= the ith observed value of the outcome, $i = 1, \cdots, n$
$\hat{y_i}$= the predicted outcome of the ith data point, $i = 1, \cdots, n$

$\bar{y_i}$= the average or sample mean of the n observed values of the outcome
$\mathbf{y}$= a vector of all n outcome values
x_{ij}= the value of the jth predictor for the ith data point, $i = 1, \cdots, n$ and $j = 1, \cdots, P$
$\bar{x}_j$= the average or sample mean of n data points for the jth predictor, $j = 1, \cdots, P$
x_i= a collection (i.e., vector) of the P predictors for the ith data point, $i = 1, \cdots, n$
$\mathbf{X}$ = a matrix of P predictors for all data points; this matrix has n rows and P columns
$\mathbf{X}'$ = the transpose of $\mathbf{X}$; this matrix has P rows and n columns
$f(\cdot)$= a function of
β= an unknown or theoretical model coefficient
b = an estimated model coefficient based on a sample of data points

Understanding the Predictive Modeling Process

Lets have a short tour to the predictive modeling process:

1) **Data Splitting:** how we allocate data to certain tasks? For example: train, validation and test data

- Predict the value for the same population case - may use simple random sampling
- Predict the value for the different population case - Example: Predicting fuel economy of new vehicles
How well the model extrapolates to a different population?

- Size of the data: small data size - resampling techniques may be better

2) **Predictor Data: How many predictors to use?**
- Feature selection: the process of determining the minimum set of relevant predictors needed by the model
- When modeling data, there is **almost never a single model fit or feature set that will immediately solve the problem**. The process is more likely to be a campaign of **trial and error** to achieve the best results.
- The effect of feature sets can be much larger than the effect of different models.
- The interplay between models and features is complex and somewhat unpredictable.
- With the right set of predictors, it is common that many different

types of models can achieve the same level of performance.

3) **Performance Estimation:**

- Quantitative assessments of statistics (i.e., $RMSE, MAPE, R^2$) using resampling help the user understand how each technique would perform on new data
- Visualizations: creating a prediction plot

4) **Model Comparison: What are the models to consider?**

- The "No Free Lunch" Theorem (Wolpert 1996) argues that, **without having substantive information about the modeling problem**, there is **no single model** that will always do better than any other model
- Try a wide variety of techniques, then determine which model to focus on

5) **Model Selection: Which specific model to choose?**
Usually two types of model selection:
- We choose some models over others: For example linear regression vs quadratic regression model. i.e we **chose between models**

- We **choose within models**: For example: If LASSO was chosen, perform hyper-parameter tuning and choose the LASSO model with best value of hyper-parameters

Data Pre-processing Techniques

Data pre-processing techniques generally refer to the **addition**, **deletion**, or **transformation** of training set data.
- Data preparation can make or break a model's predictive ability
- Approaches to unsupervised data processing: the outcome variable is not considered by the pre-processing techniques

Feature engineering: Feature engineering is the **process of creating representations of data that increase the effectiveness of a model.**
The goal of feature engineering is to create a set of relevant and informative features that can improve the performance of the model.

Q: Which feature engineering methods are the best?
A: It depends. Specifically, **it depends** on the model being used and the true relationship with the outcome

- It is very important to realize that **there are a multitude of types of models and that each has its own sensitivities and needs**. Correct feature engineering depends on several factors:

- Some encodings may be optimal for some models and poor for others

- Relationship between the predictor and the outcome is a second factor

Example:

- Some models(e.g.: linear regression, lasso, ridge regression) cannot tolerate predictors that measure the same underlying quantity (i.e., multicollinearity or correlation between predictors)
- Many models cannot use samples with any missing values (e.g.: linear regression, lasso, ridge regression)
- Some models are severely compromised when irrelevant predictors are in the data (e.g. Linear Regression, Decision trees)

Data Transformations Techniques

Data Transformations for Individual Predictors

1) **Centering and Scaling:**

- To center a predictor variable, the **mean value of the predictor is subtracted from all the values** so that the new predictor has a zero mean
- To scale the data, **each value of the predictor variable is divided by its standard deviation** so that the new predictor has a standard deviation one
Upside: generally used to **improve the numerical stability** of some calculations
Downside: **loss of interpretability** of the individual values since the data are no longer in the original units

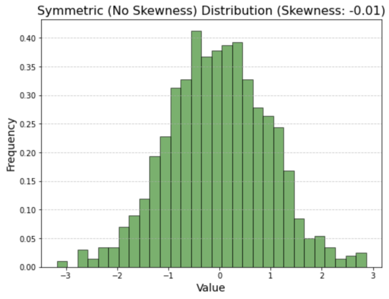
2) **Transformations to Resolve Skewness: How to reduce the skewness of the data?**
- An un-skewed distribution is one that is roughly symmetric, i.e the

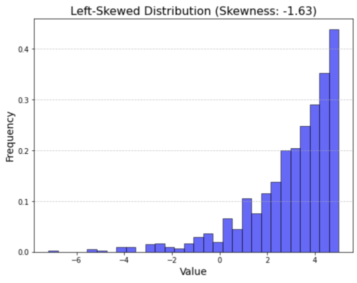
probability of falling on either side of the distribution's mean is roughly equal
- **Rule of thumb to detect Skewness: the data whose ratio of the highest value to the lowest value is greater than 20 have significant skewness**
- Sample skewness can be calculated by

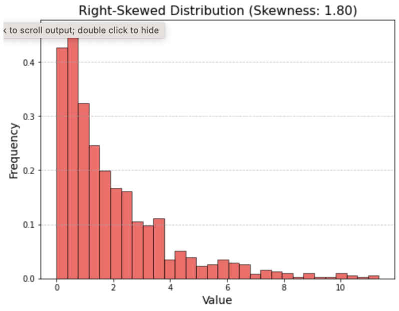
$$\text{skewness} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)v^{3/2}}, \text{ where } v = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

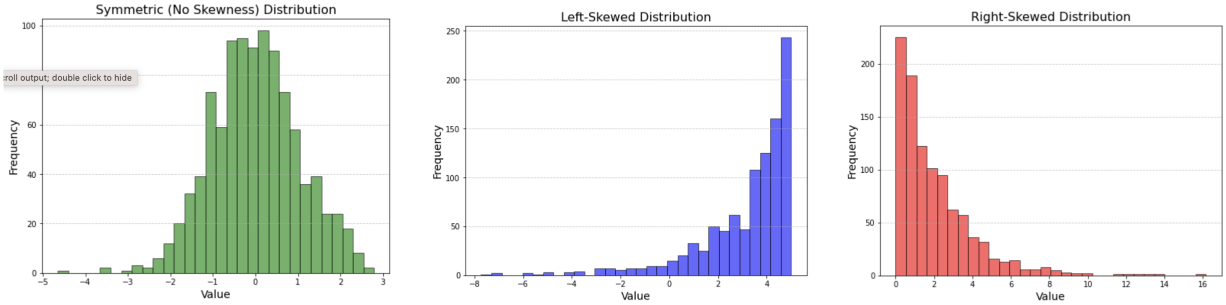
where x is the predictor variable, n is the number of values, and $\bar{x}$ is the sample mean of the predictor.

If the predictor distribution is:
Roughly **symmetric**: the skewness values will be **close to zero**
More **right skewed**: the skewness statistic becomes **larger**
More **left skewed**: the value becomes **negative**









Replacing the data with the **log, square root, or inverse transformation** may help to remove the skew.
Furthermore, If the skewness of the original predictors was the issue affecting the (for ex: logistic regression) model, other models (such as neural networks) exist that do not have the same sensitivity to this characteristic

Can we empirically identify an appropriate transformation?

Box and Cox (1964) propose a family of transformations that are indexed by a parameter, denoted as λ:

$$x^* = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log x & \text{if } \lambda = 0 \end{cases}$$

- λ can be estimated using the training data
- They described the procedure on using maximum likelihood estimation to determine the transformation parameter

Data Transformations for Multiple Predictors

- Transformations act on groups of predictors, typically the entire set under consideration

1) Transformations to Resolve Outliers:

Models insensitive to outliers:

(i) **Tree- based classification models** create splits of the training data and the prediction equation is a set of logical statements such as “if predictor A is greater than X, predict the class to be Y ,” so the outlier does not usually have an exceptional influence on the model

(ii) **Support vector machines for classification** generally disregard a portion of the training set samples when creating a prediction equation. The excluded samples may be far away from the decision boundary and outside of the data mainstream

Models sensitive to outliers:

(i) Use the data transformation called **spatial sign** (Serneels et al. 2006).

- Projects the predictor values onto a multidimensional sphere to make all the samples at the same distance from the center of the sphere

- Mathematically, each sample is divided by its squared norm:

$$x_{ij}^* = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{P} x_{ij}^2}}$$

-Since the denominator is intended to measure the squared distance to the center of the predictor's distribution, it is important to **center and scale the predictor data prior to using this transformation**

2) **Data Reduction and Feature Extraction:**

- Reduce the data by generating a smaller set of predictors that seek to capture a majority of the information in the original variables
- For most data reduction techniques, the new predictors are functions of the original predictors; therefore, all the original predictors are still needed to create the surrogate variables

a) **Principle Component Analysis (PCA)**: seeks to find linear combinations of the predictors, known as principal components (PCs), which capture the most possible variance

- The first PC is defined as the linear combination of the predictors that captures the most variability of all possible linear combinations
- Then, subsequent PCs are derived such that these linear combinations capture the most remaining variability while also being uncorrelated with all previous PCs

Mathematically, the jth PC can be written as:

$$PC_j = (a_{j1} \times \text{Predictor 1}) + (a_{j2} \times \text{Predictor 2}) + \cdots + (a_{jP} \times \text{Predictor } P),$$

P is the number of predictors
- The coefficients $a_{j1}, a_{j2}, \cdots, a_{jP}$ are called components weights and help us understand which predictors are most important to each PC

Advantages:
- It creates components that are uncorrelated
- Some predictive models prefer predictors to be uncorrelated (or at least low correlation) in order to find solutions and to improve the model's numerical stability. PCA pre- processing creates new predictors with desirable characteristics for these kinds of models
- Exploratory use of PCA is characterizing which predictors are associated with each component. Recall that each component is a linear combination of the predictors and the coefficient for each predictor is called the loading. Loadings close to zero indicate that the predictor variable did not contribute much to that component.

Limitations:
(i) PCA can generate components that summarize characteristics of the data that are irrelevant to the underlying structure of the data and also to the ultimate modeling objective
To help PCA avoid summarizing distributional differences and predictor scale information, it is best to first transform skewed predictors and then center and scale the predictors prior to performing PCA

(ii) It does not consider the modeling objective or response variable when summarizing variability since PCA is blind to the response, it is an unsupervised technique

Dealing with Missing Values: why the values are missing?

- If the pattern of missing data is related to the outcome. This is called “informative missingness” since the missing data pattern is instructional on its own. Informative missingness can induce significant bias in the model
- Missing data should not be confused with censored data where the exact value is missing but something is known about its value.
- For predictive models, it is more common to treat censored data as simple missing data or use the censored value as the observed value
- For large data sets, removal of samples based on missing values is not a problem, assuming that the missingness is not informative. In smaller data sets, there is a steep price in removing samples; some of the alternative approaches described below may be more appropriate.

How we deal with the missing data?

Two general approaches:
(1) Use the predictive models, especially tree-based techniques, can specifically account for missing data
(2) Impute(estimate values of the predictor variables based on other predictor variables) the missing data

(i) ***K*-nearest neighbor model :**

- A new sample is imputed by finding the samples in the training set "closest" to it and averages these nearby points to fill in the value

Troyanskaya et al. (2001) examine this approach for high-dimensional data with small sample sizes and made the following observation:
Advantage: imputed data are confined to be within the range of the training set values
Disadvantage: the entire training set is required every time a missing value needs to be imputed. Also, the number of neighbors is a tuning parameter, as is the method for determining "closeness" of two points.

They also found the nearest neighbor approach to be fairly robust to the tuning parameters, as well as the amount of missing data.

Removing Predictors:

- Fewer predictors means decreased computational time and complexity
- If two predictors are highly correlated, this implies that they are measuring the same underlying information
- Some models can be crippled by predictors with **degenerate distributions** (a probability distribution in a space with support only on a space of lower dimension. If the degenerate distribution is univariate (involving only a single random variable) it is a deterministic distribution and takes only a single value. Examples: a two-headed coin, rolling a die whose sides all show the same number).

-In these cases, there can be a significant improvement in model performance and/or stability without the problematic variables

A **rule of thumb for detecting near-zero variance predictors** is:
- The fraction of unique values over the sample size is low (say 10 %)
- The ratio of the frequency of the most prevalent value to the frequency of the second most prevalent value is large (say around 20)
- If both of these criteria are true and the model in question is susceptible to this type of predictor, it may be advantageous to remove the variable from the model

Collinearity: the situation where a pair of predictor variables have a substantial correlation with each other
In which case, one variable can be written as a (approximate) linear combination of another variable

Multicollinearity: relationships between multiple predictors at once
In this case, one variable can be written as a (approximate) linear combination of other variables

Reasons to avoid data with highly correlated predictors:

- Redundant predictors frequently add more complexity to the model than information they provide to the model
- Using highly correlated predictors in techniques like linear regression can result in highly unstable models, numerical errors, and degraded predictive performance

A less theoretical, more heuristic approach to identify and remove identify collinear predictors is as follows:

1) Calculate the correlation matrix of the predictors
2) Determine the two predictors associated with the largest absolute pairwise correlation (call them predictors A and B)
3) Determine the average correlation between A and the other variables. Do the same for predictor B
4) If A has a larger average correlation, remove it; otherwise, remove predictor B
5) Repeat Steps 2–4 until no absolute correlations are above the threshold

Adding Predictors:

- When a predictor is categorical, such as gender or race, it is common to decompose the predictor into a set of more specific variables
- The categories are re-encoded into smaller bits of information called "dummy variables"
- Usually, each category get its own dummy variable that is a zero/one indicator for that group
-The **decision to include all of the dummy variables can depend on the choice of the model**. Models that include an **intercept term**, such as simple linear regression would have numerical issues if each dummy variable was included in the model
- The reason is that, for each sample, these variables all add up to one and this would provide the same information as the intercept
- If the model is insensitive to this type of issue, using the complete set of dummy variables would help improve interpretation of the model

- A new technique for augmenting the prediction data **for classification models**, developed by Forina et al. (2009):

  - Calculate the class centroids: centers of the predictor data for each class. Then calculate the distance to each class centroid for each predictor, and finally add these distances to the model as a new variable

References

[1] Kuhn, Max., and Kjell Johnson. Applied Predictive Modeling. New York: Springer, 2013

[2] Kuhn, Max, and Kjell Johnson. Feature engineering and selection: A practical approach for predictive models. CRC Press, 2019.

[3] http://machinelearningintro.uwesterr.de

[4] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013

[5] Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome. The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc., 2001.

[6] Rodriguez, M. The failure of predictive modeling and why we follow the herd. Technical report, 2011.