Predictive Modeling

Introduction and Fundamentals

Introduction

Predictive Modeling – Building of a statistical model (set of rules) for predicting or estimating an outcome's value based on the value of one or more input variables.

- Predictive analytics is a useful skill for tackling the following questions-

  ✓ How can I predict a continuous or categorical outcome using a set of predictors such that the difference between our predictions and answer is as small as possible?

  ✓ How can I make correct predictions about an outcome without knowing the underlying process that produced the outcome?

  ✓ How can I predict an outcome with a lot of potential predictors without knowing which ones are important?

Steps of building a predictive model

1. Choose your outcome of interest
2. Think about variables related to the outcome of interest
3. Gather data before and after the outcome occurs
4. Enter/Clean/Transform data as required
5. Create new variables from existing data
6. Fit a predictive model algorithm
7. Examine a models performance
8. Make adjustments to data/model
9. Keep model

Data and Scaling

- Scales of measurement (Nominal, Ordinal, Interval, Ratio)

- Qualitative vs Quantitative

- Importance of scaling-
  - ✓ The prediction method we choose differs depending on the scale type
  - ✓ Different scales provide different amount of information
  - ✓ Different scales help examine relationships accurately
  - ✓ Using different scaling might make the data invalid

Data and Scaling contd.

Nominal Scaling: Unordered Categories/Labels (Categories with no definite order). Example- Gender, Zipcode, Political affiliation

Ordinal Scaling: Ordered Values/Ranks (Values that have a natural order). Example- Ranking, School year

Interval Scaling: Equally spaced numbers (Ordered measurement in equally spaced units. Example- Temperature, IQ, SAT scores

Ratio Scaling: Equally spaced numbers with a true zero. Example- Height, Work Experience

How to choose the scale type?

- **Use the scale type that:**
  - Provides the most information possible
  - that can be collected at a reasonable range on (use fewer categories if you can only a single instance of one of the categories)
  - Is an accurate representation of the underlying construct (if it is a predictor variable)
  - Matches your outcome best (if it is an outcome variable)

Variable Types

Quantitative vs. **Qualitative**

Quantitative: Described as numbers

- Example: counts and measurements (-1, 2, 3.5,...)

Qualitative: measured as labels and names

- Example: Species (“Dog”, “Cat”, “Bird”)

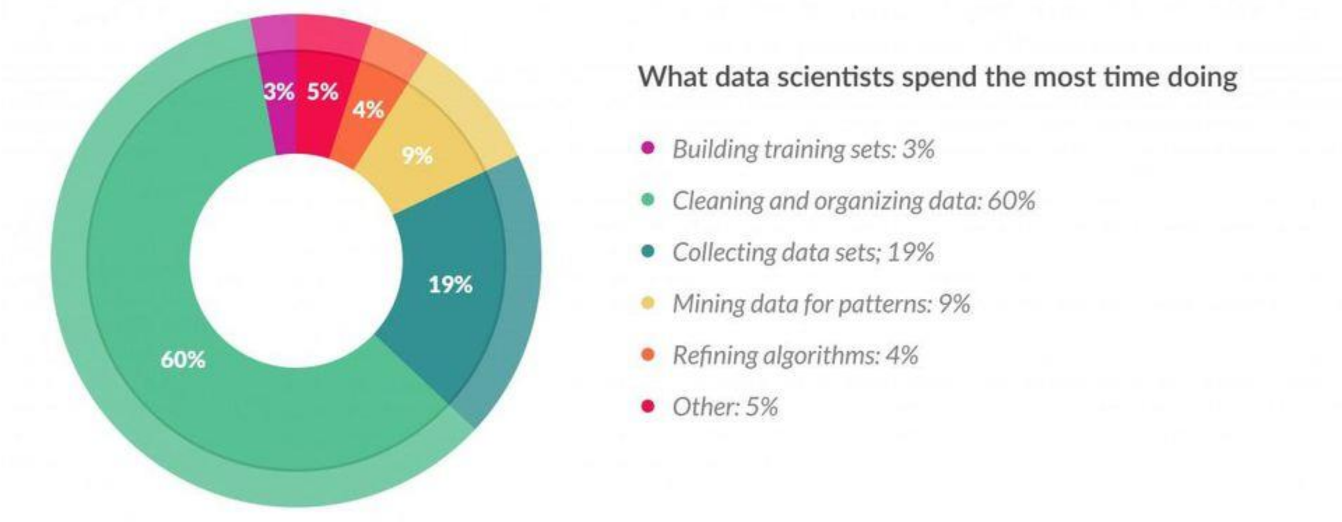
Continuous vs. **Discrete**

Continuous: Measured along a continuum at any decimal level precision. You can always be more precise

Discrete: Measured in whole units (integers) or categories

Data Cleaning

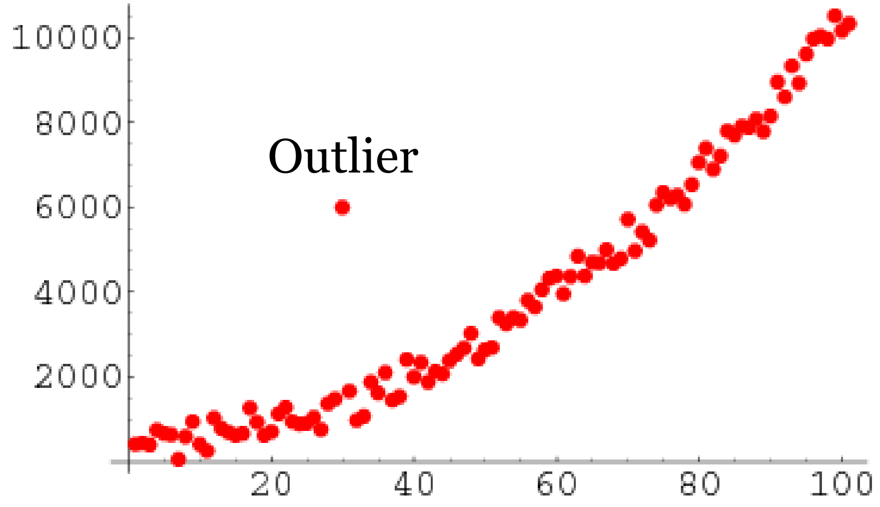
- The process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database
  - Describes a collection of actions taken to improve the quality of data
  - Performed by a human or a computer, or a combination of both



Survey of 80 data scientists was conducted by CrowdFlower

Outliers

An outlier is an observation that lies outside the overall pattern of a distribution. It is a value that is unexpectedly extreme relative to the value of the other respondents



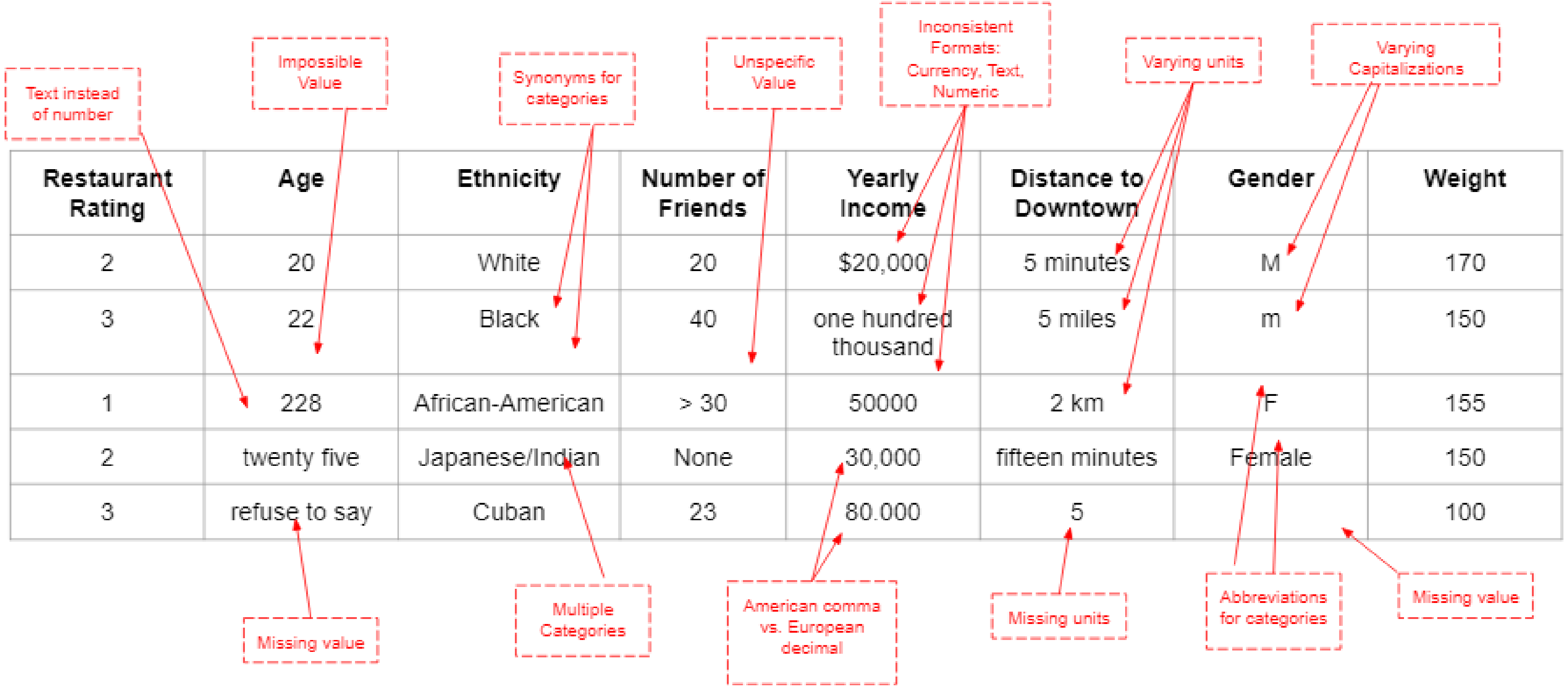
Ways to check for outliers:

- **More than 4 standard deviations from mean**
  - Calculate the mean and standard deviation of the column
  - Subtract the mean from each number, and divide the difference by the standard deviation
  - See if any value is greater than +4 or less than -4
- **Perform analysis with and without the data**
  - Examine how much the results change
  - If results change dramatically, then single point has undue influence, and you may want to discard it
  - Cook’s D is an example of this for regression

Data Cleaning

Common steps for data cleaning-

- Check for formatting errors
- Check for impossible values
- Check for outliers
- Check for missing values
- Check for inattention



Restaurant Rating	Age	Ethnicity	Number of Friends	Yearly Income	Distance to Downtown	Gender	Weight
2	20	White	20	$20,000	5 minutes	M	170
3	22	Black	40	one hundred thousand	5 miles	m	150
1	228	African-American	> 30	50000	2 km	F	155
2	twenty five	Japanese/Indian	None	30,000	fifteen minutes	Female	150
3	refuse to say	Cuban	23	80.000	5		100

Text instead of number
Impossible Value
Synonyms for categories
Unspecific Value
Inconsistent Formats: Currency, Text, Numeric
Varying units
Varying Capitalizations
Missing value
Multiple Categories
American comma vs. European decimal
Missing units
Abbreviations for categories
Missing value

Handling Missing Data

- Types of missing data –
  - ✓ Missing Completely at Random
  - ✓ Missing at Random
  - ✓ Missing not at Random

- Common methods to handle missing data –
  - ✓ Listwise Deletion
  - ✓ Pairwise Deletion
  - ✓ Single Imputation
  - ✓ Multiple Imputation

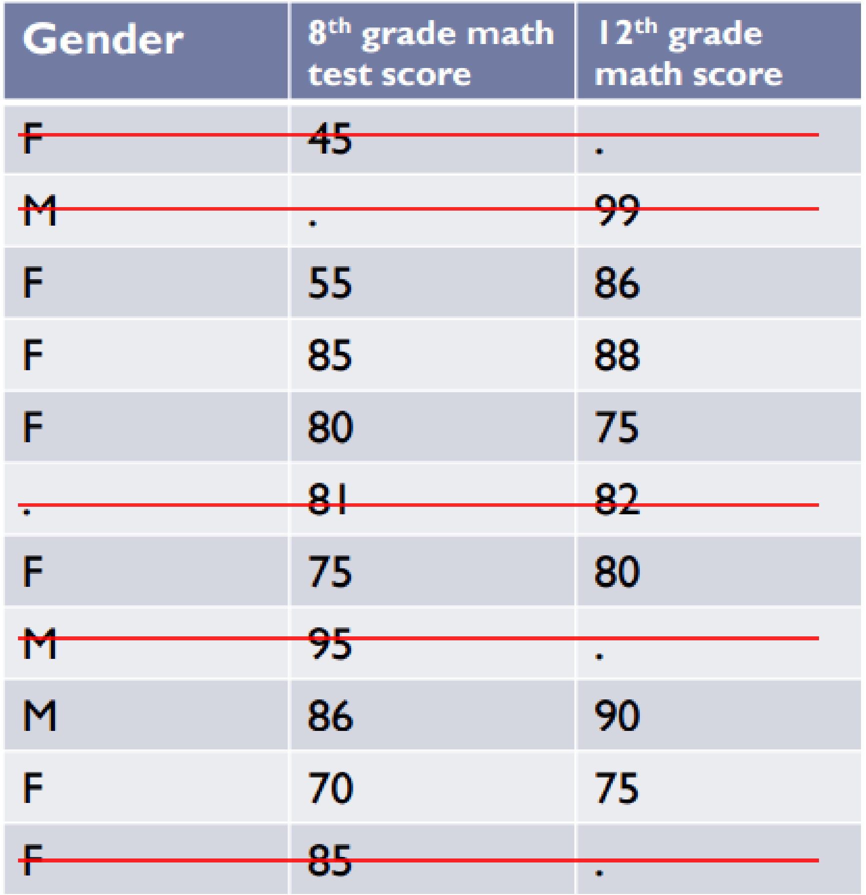
How can we address missing data?

- **Deletion**: treat the data like that person or non-response never existed
    - List-wise
    - Pair-wise

- **Imputation**: fill in the missing value with what it likely was
    - Mean/Median
    - Hot deck
    - Multiple

- **Maximum Likelihood**: select the parameter estimates that maximize the likelihood function based on the available data
    - Full Information Maximum Likelihood (FIML)
    - Expectation Maximization (EM)

Listwise Deletion



Gender	8th grade math test score	12th grade math score
~~F~~	~~45~~	~~.~~
~~M~~	~~.~~	~~99~~
F	55	86
F	85	88
F	80	75
~~.~~	~~81~~	~~82~~
F	75	80
~~M~~	~~95~~	~~.~~
M	86	90
F	70	75
~~F~~	~~85~~	~~.~~

In listwise deletion, any participant who did not completely answer all relevant questions is removed from the analysis

Consequences of Listwise Deletion

- Lowers the statistical power
  - Sample size is reduced
  - Less likely to predict the outcome accurately.
  - Deletes the relationship between the outcome and variable for that observation
  - Yields biased predictions

- Limits generalizability

- Discarding data can sometimes be unethical

Mean Imputation

- **Do not** impute data with the mean or median and then proceed as though you have complete data

Common Examples/Variations:

  - **Mean imputation**—replacing each missing observation with the mean for the corresponding variable based on the responses that exist
  - **Median imputation**—replacing each missing observation with the value that half of people were below and half were above

- **Mode imputation**—replacing each missing observation with the most common value for that variable

Model Based Imputation

- **Model Based Imputation:** Build a predictive model to determine what the missing value most likely was

Common Examples

- **Nearest Neighbor Approaches**: Fill in the value with the person who is closest on all the other variables
- **Random Forest**: Develop a set of if-then rules to predict a value from the other values, and then fill in the missing value with the predicted value

Predictive Mean Matching: Develop a predictive model for how the missing variable relates to the other variables, and then fill in the value with another person’s value who is closest to the predicted value

Multiple Imputations

- **Multiple Imputation methods**:
  - Impute missing values to create an imputed dataset.
  - Repeat, multiple (e.g., 40) times to have 40 separate imputed datasets.
  - Run the analysis on each imputed dataset.
  - Combine the 40 results to get parameter estimates and standard errors.

- Multiple imputation performs an unbiased single imputation routine over and over again (i.e., it makes multiple, different guesses at what the missing data might have been).

Consequences of Multiple Imputations

- **Benefits of Multiple Imputation methods**
  - **Unbiased** under MCAR and MAR
  - Improves as you add more variables to the imputation model
  - **Accurate SEs** and valid model fit information Performs well in **small samples** (as low as N=50), even with very large multiple regression models and with as much as 50% missing data in the DT.
  - **Does not require normally distributed data**

- **Limitations of Multiple Imputation methods**
  - **Biased under MNAR**
  - **Prone to overfit**: Number of variables in prediction model should be < 100 When the sample size is large (over N > 1,000), the # of variables should be kept smaller when the sample sizes are smaller
  - Gives **slightly different estimates** each time: No single dataset provided
  - When used with SEM, suffers more **non-convergences**

Qualitative Variables -> Nominal Variables

Qualitative Nominal Predictors need to be transformed into numbers that represent which category the observation belongs to

- You know what kind of housing people live in (dorm, apartment, house, townhouse)
- You need to make a variable that represents numerically which one a person lives in

Do NOT convert nominal data into numbers by giving each value a unique number

- Example of what NOT to do:
  - Dorm= 1
  - Apartment= 2
  - House= 3
  - Townhouse =4
- If you give each variable a unique number, the analysis thinks the variable is ratio/interval

Data Restructuring

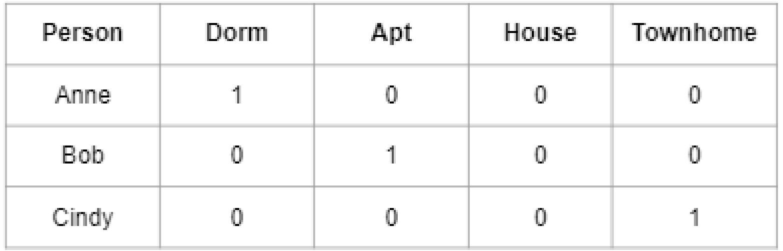
Converting qualitative nominal data using either **one-hot-encoding** or **dummy coding**

One-hot-encoding:

Each category is a separate variable/column

The category the person belongs to has a 1 in the corresponding column

The categories the person doesn’t belong to has a 0 in the corresponding column



Person	Dorm	Apt	House	Townhome
Anne	1	0	0	0
Bob	0	1	0	0
Cindy	0	0	0	1

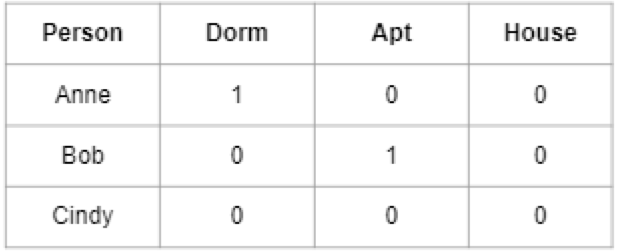
Dummy coding:

Each category has a separate variable/column, except for one: the reference category

The reference category does not receive a predictor column

The category the person belongs to has a 1 in the corresponding column, if the category has a column

If the person does not belong to the category, the corresponding column has a 0 in it



Person	Dorm	Apt	House
Anne	1	0	0
Bob	0	1	0
Cindy	0	0	0

Benefits and Limitations

One-hot-encoding

- **Benefits:**
  - Clear what category a person belongs to
  - Each variable's effect is described in absolute terms
- **Limitations:**
  - Redundant (some analysis won't run if there is high redundancy)
  - Uses more space (uses more memory for larger datasets)

Dummy coding

- **Benefits:**
  - Avoids redundancy (allows some analysis to run)
  - Space efficient
- **Limitations:**
  - Unclear what the unspecified category represents
  - Each variable's effect is described relative to the unused variable

Standardizing and Normalizing Variables

Many Predictive modeling algorithms incorporate optimization routines or distance measures

- Optimization routines: A computational way to discover the parameter values for the best model that describes your data. The amount of improvement in fit between two models helps decide which model's parameter values to keep
- Distance measures: The difference between variables for two people determines how similar they are

We sometimes need to put all the variables on a common scale

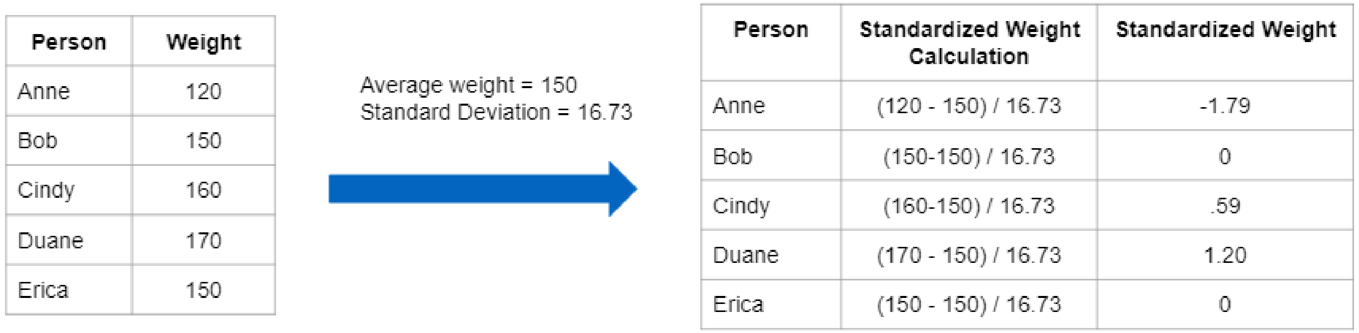
Two common ways to put the variables on a common scale

- **Standardization**: Each score represents the number of standard deviations away from the variable's mean it is
  - Typically ranges from -4 and +4
  - A standardized score of +2 is 2 standard deviations higher than the mean of the scores
- **Normalization**: Each score represents the percentile the score is in compared to the other scores
  - Ranges from 0.0 to 1.0
  - A normalized score of .6 is larger than 60% of the other data points

Data Standardization & Normalization

Standardization-

- Find the mean of all the recorded values of the variable
- Find the standard deviation of all the recorded values of the variable
- For each value in the column, subtract the mean from it, and divide the difference by the standard deviation



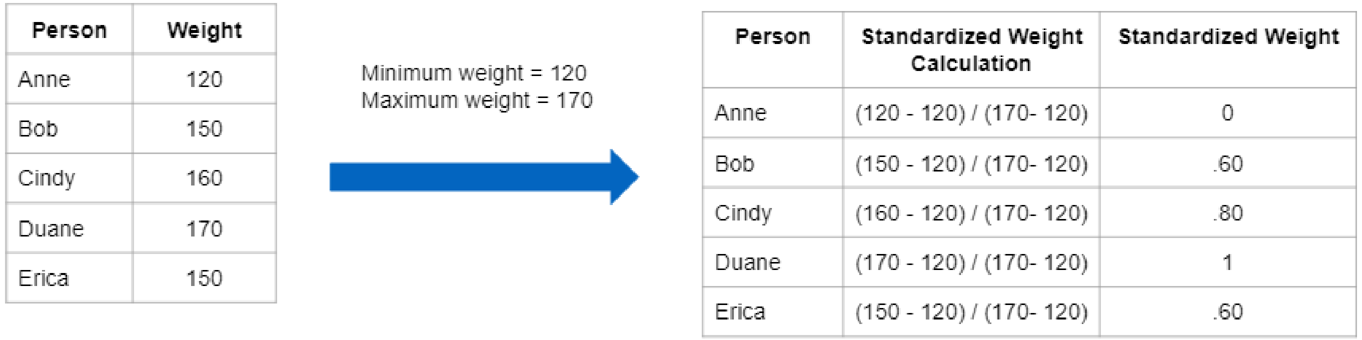
Person	Weight
Anne	120
Bob	150
Cindy	160
Duane	170
Erica	150

Average weight = 150
Standard Deviation = 16.73

Person	Standardized Weight Calculation	Standardized Weight
Anne	(120 - 150) / 16.73	-1.79
Bob	(150-150) / 16.73	0
Cindy	(160-150) / 16.73	.59
Duane	(170 - 150) / 16.73	1.20
Erica	(150 - 150) / 16.73	0

Normalization-

- Find the minimum and maximum value of all the recorded values of the variable
- Subtract the minimum value from the maximum value. This is your denominator
- For each value in the column, subtract the minimum value from it, and divide it by the denominator you just computed



Person	Weight
Anne	120
Bob	150
Cindy	160
Duane	170
Erica	150

Minimum weight = 120
Maximum weight = 170

Person	Standardized Weight Calculation	Standardized Weight
Anne	(120 - 120) / (170- 120)	0
Bob	(150 - 120) / (170- 120)	.60
Cindy	(160 - 120) / (170- 120)	.80
Duane	(170 - 120) / (170- 120)	1
Erica	(150 - 120) / (170- 120)	.60

Classification Vs Regression models

- Introducing the two types of predictive models with examples
- Importance of choosing the correct model
- Goals of classification and regression algorithms

Classification Algorithms

✓ Logistic Regression
✓ Support Vector Machines
✓ Classification Trees
✓ Naïve Bayes

Regression Algorithms

✓ Linear Regression
✓ Multiple Regression
✓ Regression Trees
✓ Support Vector Regressors