

Specialist Programme on Artificial Intelligence for IT & ITES Industry

AI Applications: Process & Best Practice

Dr Barry Shepherd
barryshepherd@nus.edu.sg

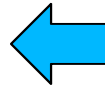
Singapore e-Government Leadership Centre
National University of Singapore

© 2020 NUS. The contents contained in this document may not be reproduced in any form or by any means, without the written permission of ISS, NUS other than for the purpose for which it has been supplied.

Inspire *Lead* *Transform*

Agenda

- AI Applications
- Process and Best Practices
- Challenges & Issues



Predictive Modeling Process



Predictive Analytical Model Process Flow



Source: Forrester Research, Inc.

Setting Business Goals

- Usually a **two way** process between the Data Scientists / AI experts and the domain experts
 - The Data Scientist / AI expert usually needs **some** domain knowledge for this conversation to succeed



- Business Goal Guidelines
 - Use only business terms – make no mention of analytics methods!
 - There must be an actionable outcome
 - Must be able to measure success (quantifiable metrics)

Setting Business Goals / Metrics

- Possible examples are ...
 - Improve the response rate for a direct marketing campaign by 50%
 - Increase the average order size by 2 items
 - Determine what drives customer acquisition identify top 2 factors
 - Forecast the size of the customer base in the future 80% accuracy on test set?
 - Retain profitable customers reduce churn of top 20% customers by 10%
 - Recommend the next, best product for existing customers get 20% buy rate lift?
 - Choose the right message for the right groups of customers too vague?

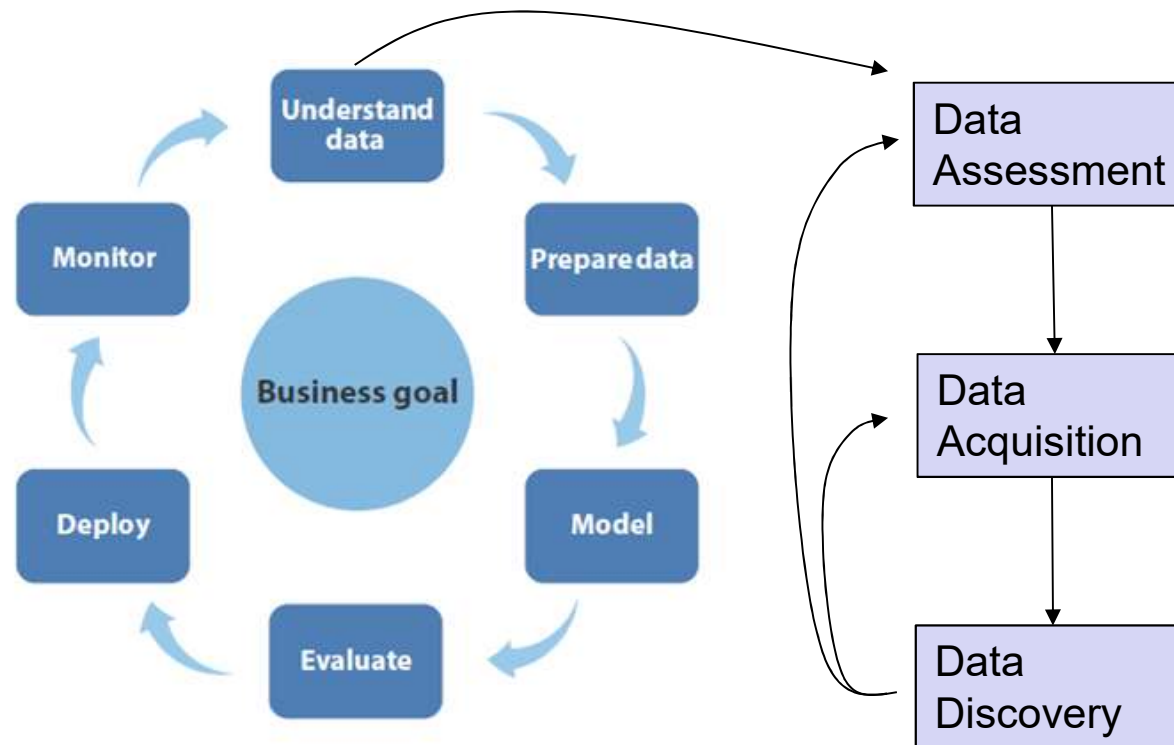
By how much?
Need clear success criteria

Business Goals are not Technical Goals

- A business goal states objectives in business terminology
 - E.g. *“Increase sales to existing customers.”*

- A technical goal states project objectives in technical terms.
 - E.g. *“Predict how many widgets a customer will buy, given their purchases over the past three years, demographic information (age, salary, city, etc.), and the price of the item.”*

Model Building Process



Source: Forrester Research, Inc.

Assess what data you have and what you need to start to collect / acquire

Basic analysis & visualisation to gain insights

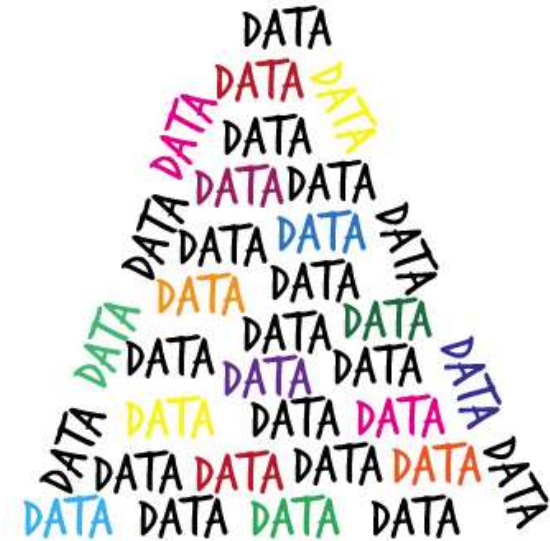
Identifying Data Requirements

■ Questions to Answer:

- What data is available?
- What must the data contain?
- What would be useful? (whether available or not) – **innovate!**
- What is the right level of granularity?
- What volume of data is needed?
- How much history is required?
how far back in time should the data go?

■ What data is required for comparison?

- What is currently being done?
 - E.g. what is the existing churn rate, response rate, failure rate?
- Obtain a control group ~ data describing the status quo
 - E.g. what happened to patients who did not receive the treatment?
 - E.g. what did customers buy who did not see the ad?



Identify any Data Gap

- Consolidate all of your data requirements
- Determine what (if any) essential data is missing
- How to bridge the gap?
 - Put in place mechanisms to start collecting the missing data (delay the analytics)
 - Get the data from elsewhere (e.g. 3rd party, the web)
 - Innovate to obtain missing data or data you think may be useful



Do I have Enough Data?

- How much data is required for Machine Learning / Deep Learning?
- Simple answer = LOTS
- Common rule of thumb is 10 examples per model parameter (e.g. per NN weight)

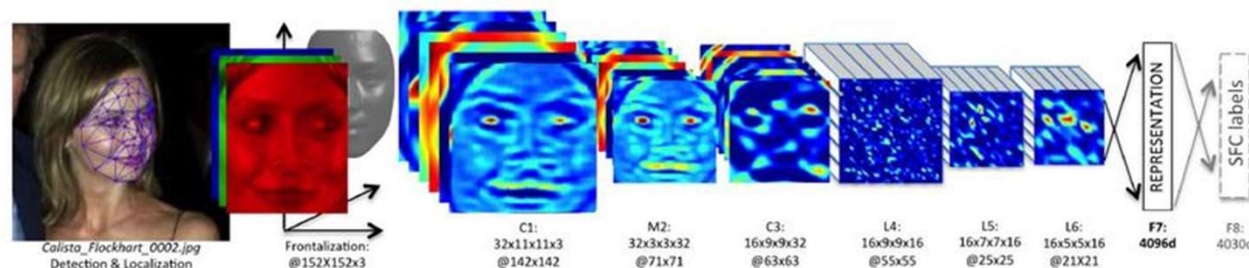
DeepFace

Alignment: 2D, 3D

Input: RGB image 152x152

Output feature size: 4096

Parameters: ~120 million



Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In CVPR, 2014

Data Augmentation - Images

- **rotation:** random with angle between 0° and 360° (uniform)
- **translation:** random with shift between -10 and 10 pixels (uniform)
- **rescaling:** random with scale factor between $1/1.6$ and 1.6 (log-uniform)
- **flipping:** yes or no (bernoulli)
- **shearing:** random with angle between -20° and 20° (uniform)
- **stretching:** random with stretch factor between $1/1.3$ and 1.3 (log-uniform)

Classifying plankton with deep neural networks

MARCH 17, 2015

<https://benanne.github.io/2015/03/17/plankton.html>



Gaussian noise
and cropping are
two common
methods

Many python libraries exist for image augmentation, e.g. imgaug (over 60 methods), see

<https://towardsdatascience.com/data-augmentation-for-deep-learning-4fe21d1a4eb9>

Data Augmentation - Business Data?

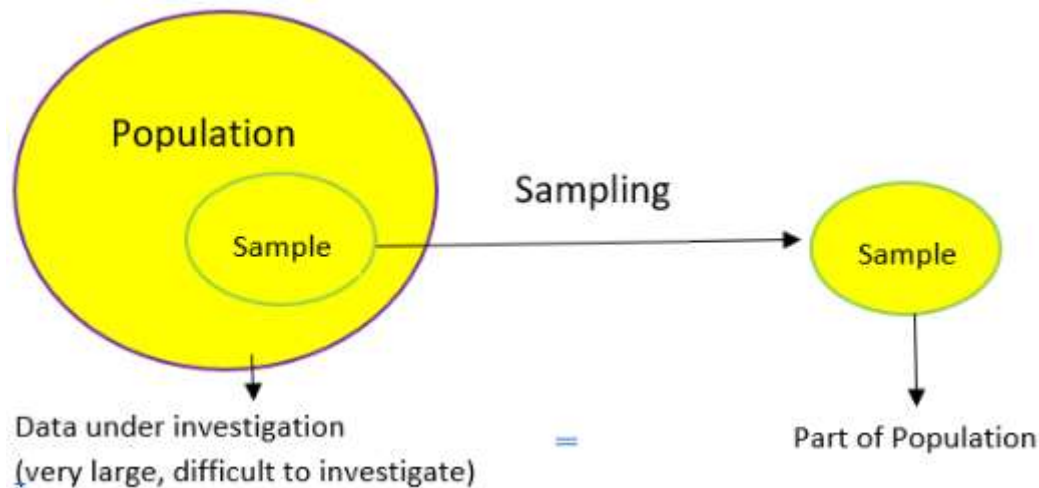
- Required for Data Balancing (e.g. when one class is very rare)
- Adding noise
- Creating duplications
- Synthesising “similar” cases (e.g. SMOTE)
- Using statistical distributions to generate simulated data (harder)

Too Much Data?

Many Big Data proponents say use it all....

“Collecting large amount of data & then processing it to make Inferences from them involves different costs associated with it”

Raw “big data” is often very noisy, low signal-to-noise ratio



E.g. randomly select 10% of records

but what if the training signal is very rare? – use stratified sampling or similar

You need Labelled Data!

- Predictive Models typically use Supervised Learning
 - The training dataset must contain examples of **all of the decision classes**
 - This can often be a challenge: tedious & time-consuming

AI Platform Data Labeling Service

AI Platform Data Labeling Service lets you work with human labelers to generate highly accurate labels for a collection of data that you can use in machine learning models.

Labeling your training data is the first step in the machine learning development cycle. To train a machine learning model, provide representative data samples that you want to classify or analyze, along with the machine learning algorithm to handle each sample. For example, to train a model that can identify flowers in images, you must label objects like sunflowers, roses, and tulips in the image dataset. To train a model that can identify the names of diseases in medical documents, you must highlight disease-related words in the document dataset.

To start data labeling in AI Platform Data Labeling Service, create three resources for the human labelers:

- A **dataset** containing the representative data samples to label
- A **label set** listing all possible labels in the dataset
- A set of **instructions** guiding human labelers through labeling tasks

Once you've created these resources, you submit them as part of a **labeling request**. The human labelers start annotating the items in the dataset according to your instructions. After human labelers finish the labeling, you can **export** well labeled datasets and use the datasets in the machine learning development.

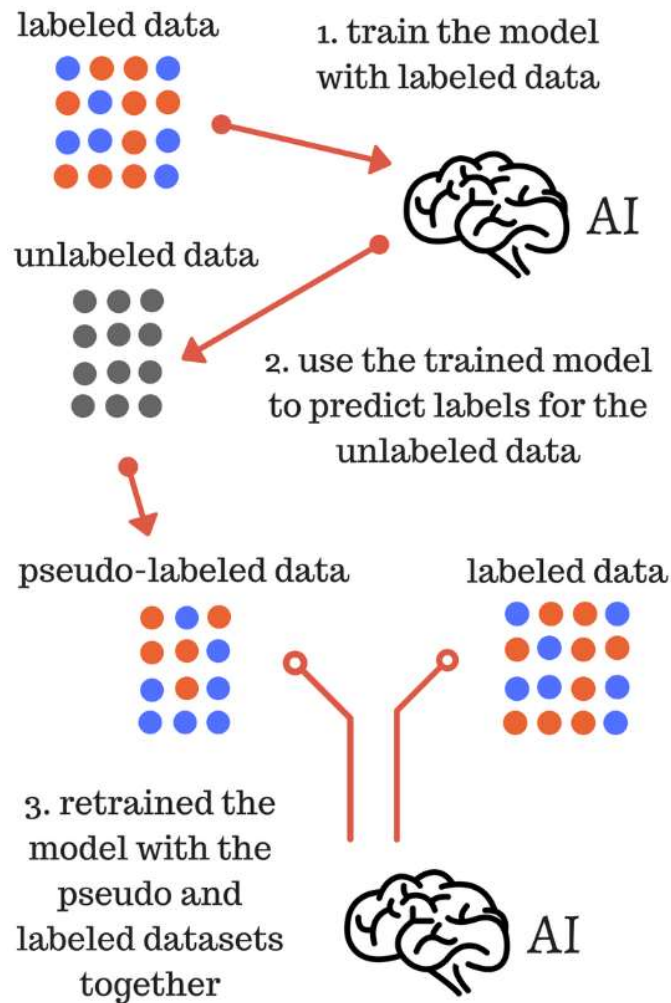
Get someone
else to do it

(e.g. get your
students to do it!)



AI & Machine Learning Products

Avoiding the Data Labelling chore



Pseudo Labelling

- One form of **Semi-Supervised Learning**
- Good for if you have a small amount of labelled data and lots of unlabelled

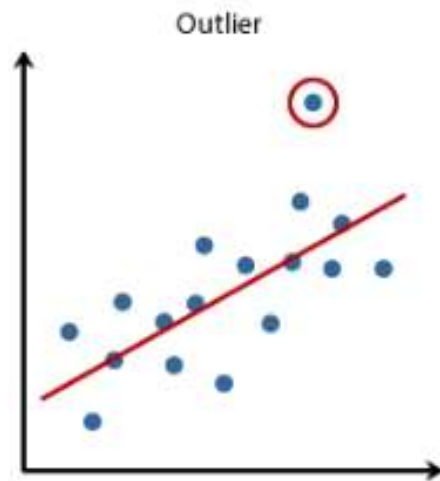
Can do iteratively, e.g.

- Add only the most confident predicted labels into the training data
- Build new model
- repeat

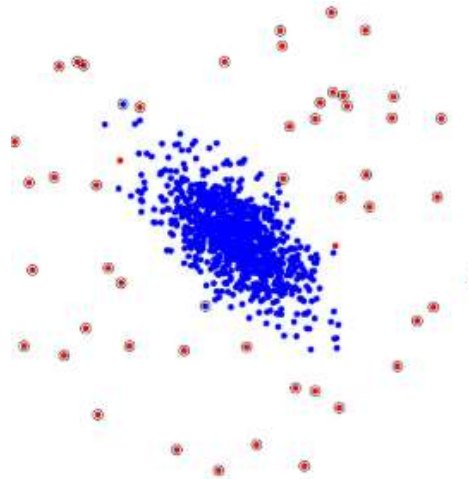
Avoiding the Data Labelling chore

■ Unsupervised Learning – build a model to “describe” the data

- Good if the target pattern is relatively rare, e.g. people without disease, machines with no faults, people who don't buy....
- Identify the minority class as outliers – those that don't fit the usual situation
- If labels are available then model only the baseline (normal) situation,

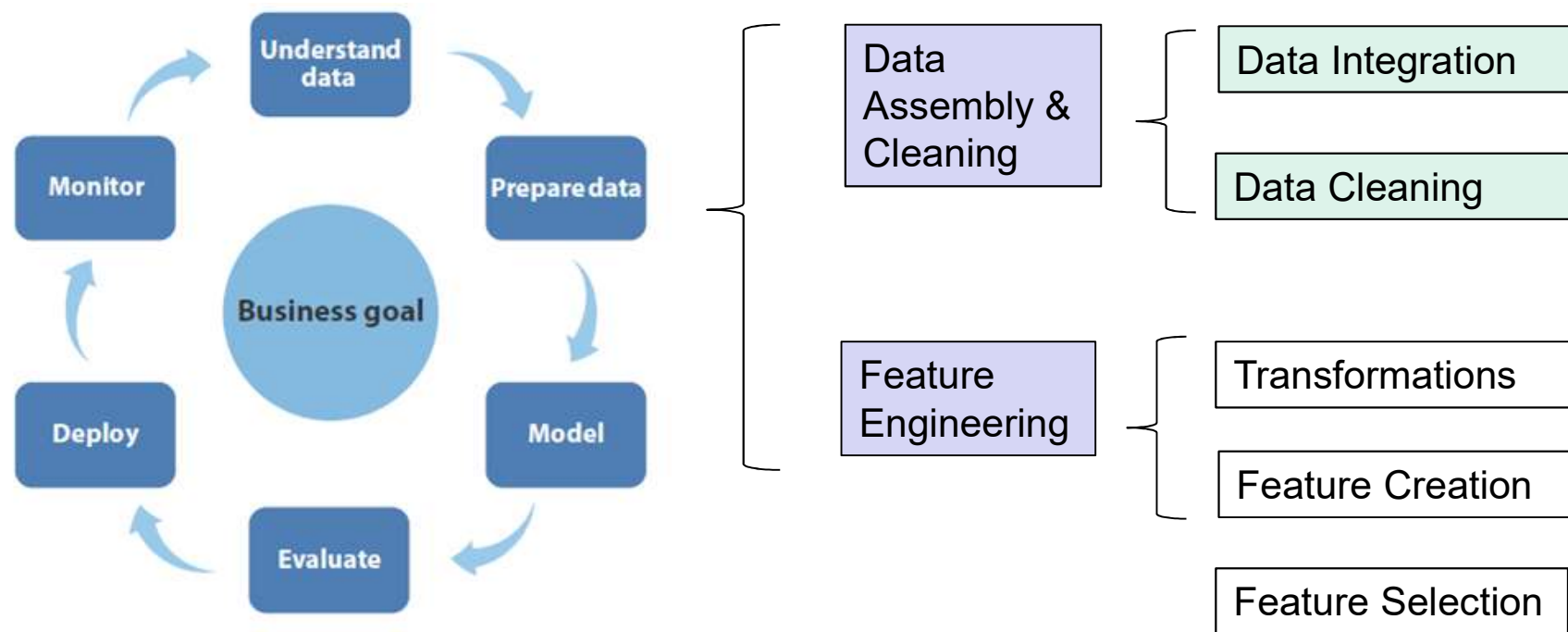


E.g. Fit lines or curves to the baseline data, look at distance to the best fit line or curve



E.g. Cluster the baseline data, anything far away from the cluster centres is an outlier

Predictive Modeling Process



Source: Forrester Research, Inc.

Data Audit

A useful starting point is the **Data Audit**.

- Is the data adequate?
- Is it what you expect?, Does it look sensible?
- What are the data quality issues?

Data Audit of [26 fields]

File Edit Generate

Audit Quality Annotations

Complete fields (%): 42.31% Complete records (%): 0%

| Field | Measurement | Outliers | Extremes | % Complete | Valid Records | Null Value | Empty String | White Space |
|------------------------|-------------|----------|----------|------------|---------------|------------|--------------|-------------|
| Label | Continuous | 0 | 1 | 100 | 4000 | 0 | 0 | 0 |
| Contract_Id | Continuous | 0 | 0 | 100 | 4000 | 0 | 0 | 0 |
| Payment_Method | Continuous | 0 | 136 | 100 | 4000 | 0 | 0 | 0 |
| Promotion_Description | Categorical | -- | -- | 92.475 | 3699 | 0 | 301 | 301 |
| Civility | Continuous | 189 | 5 | 93.1 | 3724 | 276 | 0 | 0 |
| Job | Categorical | -- | -- | 96.975 | 3879 | 0 | 121 | 121 |
| Nationality | Continuous | 0 | 3 | 99.525 | 3981 | 19 | 0 | 0 |
| DOB | Continuous | 121 | 2 | 99.525 | 3981 | 19 | 0 | 0 |
| Age | Continuous | 133 | 0 | 100 | 4000 | 0 | 0 | 0 |
| Age_Band | Continuous | 0 | 18 | 97.05 | 3882 | 118 | 0 | 0 |
| Gender | Continuous | 0 | 18 | 96 | 3840 | 160 | 0 | 0 |
| Credit_Score | Categorical | -- | -- | 61.225 | 2449 | 0 | 1551 | 1551 |
| Tariff_Plan | Continuous | 0 | 0 | 99.95 | 3998 | 2 | 0 | 0 |
| Num_Active_VAS | Continuous | 0 | 11 | 99.8 | 3992 | 8 | 0 | 0 |
| Num_Inactive_VAS | Continuous | 2 | 7 | 100 | 4000 | 0 | 0 | 0 |
| Service_DataFax | Continuous | 0 | 11 | 96.65 | 3866 | 134 | 0 | 0 |
| Service_Voicemail | Continuous | 1 | 9 | 55 | 2200 | 1800 | 0 | 0 |
| Service_SMS | Categorical | -- | -- | 0.45 | 18 | 0 | 3982 | 3982 |
| Cust_Activation_Date | Continuous | 0 | 15 | 99.825 | 3993 | 7 | 0 | 0 |
| Cust_Contact | Continuous | 83 | 17 | 99.525 | 3981 | 19 | 0 | 0 |
| Cust_Contact_Compl... | Continuous | 0 | 19 | 100 | 4000 | 0 | 0 | 0 |
| Num_active_VAS_pre... | Continuous | 0 | 0 | 100 | 4000 | 0 | 0 | 0 |
| Num_inactive_VAS_pr... | Continuous | 0 | 0 | 100 | 4000 | 0 | 0 | 0 |
| Cust_Contact_prev... | Continuous | 5 | 13 | 100 | 4000 | 0 | 0 | 0 |
| Cust_Contact_Compl... | Continuous | 1 | 10 | 100 | 4000 | 0 | 0 | 0 |
| Churn_Flag | Continuous | 0 | 0 | 100 | 4000 | 0 | 0 | 0 |

Data Audit Tool from SPSS Modeller

Data Cleaning



GIGO



- Handle missing values
- Handle noisy / erroneous data
- Handle outliers
- Correct inconsistent data
- Resolve redundancy caused by data integration



X

Handling Missing Values

- Data Imputation - fill in the missing values automatically
 - Guiding Principle: Avoid adding bias and distortion to the data
 - Understand why the data is missing can help guide the imputation
 - Often a missing value means zero or the default value. E.g. for 'rainfall' variable, a missing value may mean no rain on that day → 0

- Common Options

- A global **constant** : e.g., “unknown” or 0 (zero)

Easy, but modeling algorithms may mistakingly treat “unknown” as a concept

- The **attribute mean** (or median, mode)

Simple and quick though not always satisfactory

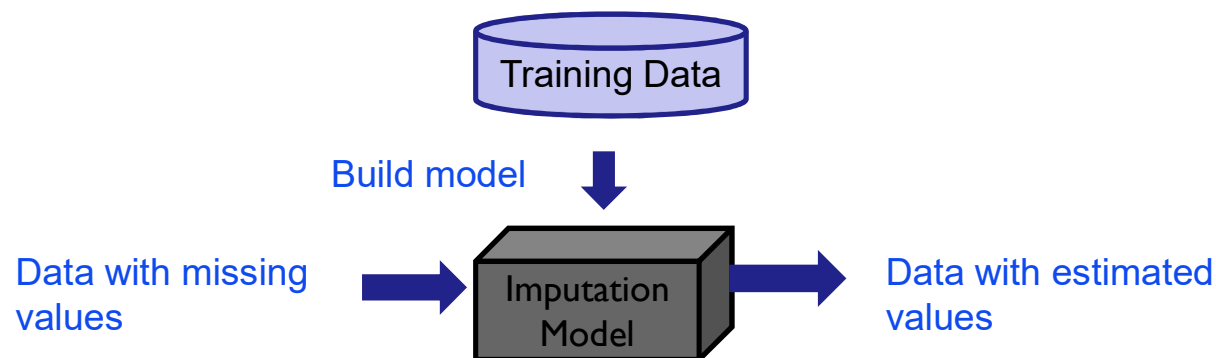
- The **attribute mean** for all samples belonging to the **same class**

Often a better estimate than attribute mean

| Gender | Children | Salary | Bought PEP |
|--------|----------|--------|------------|
| M | 1 | 29,000 | Y |
| M | 0 | 65,000 | Y |
| F | 2 | - | Y |
| M | 0 | 47,000 | Y |
| F | - | 15,000 | N |
| - | 1 | 23,000 | N |
| F | 1 | 36,000 | N |

Data Imputation

- Train a prediction model (e.g. regression model, decision tree) to predict the most probable value
 - Use variables containing values to estimate the variable with missing values
 - Can produce good estimates.
 - Need training data and additional modeling



Outliers

- Observations that “*deviate so much from other observations as to arouse suspicion that it was generated by a different mechanism*”. (Hawkins, 1980)
- Appearing at the maximum or minimum end of a variable, skewing or distorting the distribution
 - E.g. extreme weather conditions on a particular day, a very wealthy person financially very different from the rest of the population, etc.

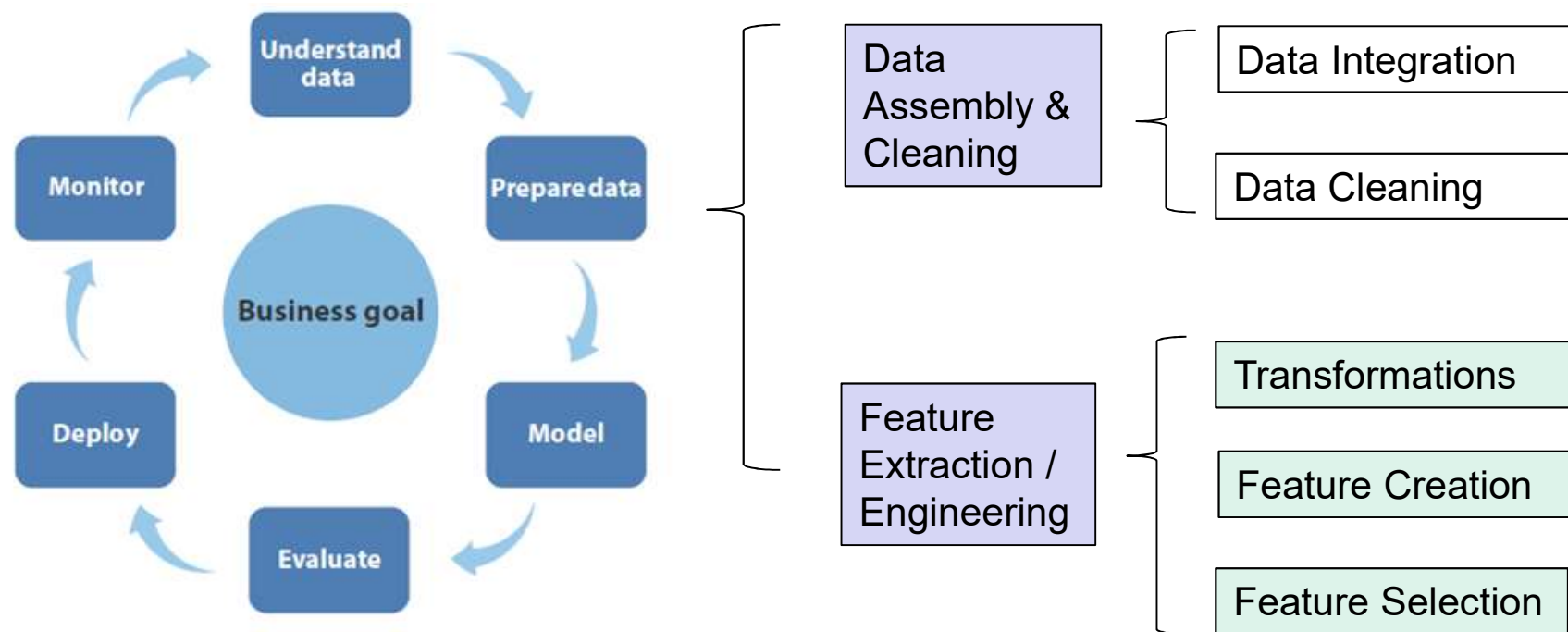


Outliers may be errors
(remove)
or
they may be valid data
(keep, or handle another way)

Identifying Outliers

- Statistical tests for variance
- Clustering
- Human inspection

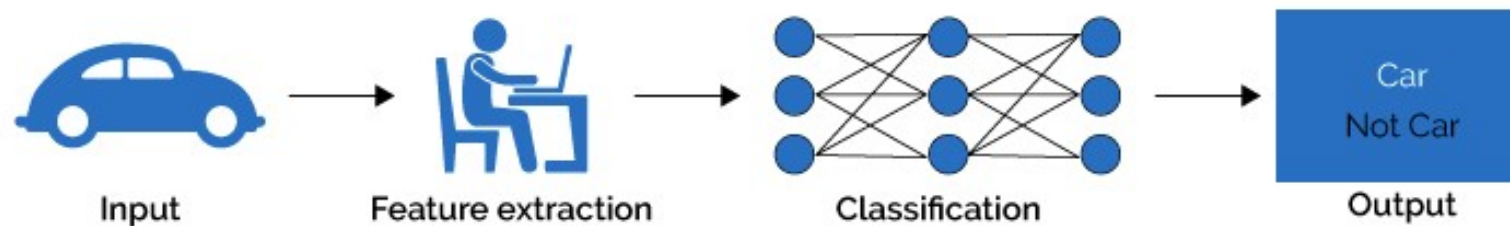
Predictive Modeling Process



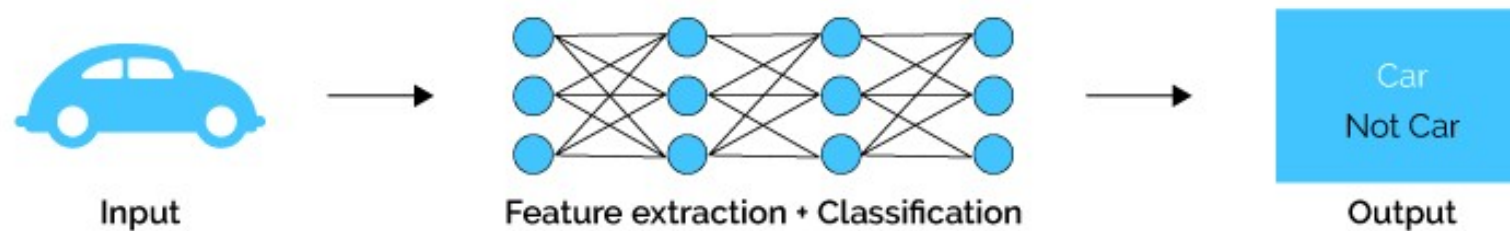
Source: Forrester Research, Inc.

Feature Engineering for DL?

Machine Learning



Deep Learning



Data Transformations - Normalisation

- Reduces outlier distortion and enhances linear predictability
- Ensure all variables have approximately the same scale
 - E.g. variable *Age* vs *Income*: a distance of 10 “years” may be more significant than a distance of \$1000, yet \$1000 swamps 10 numerically
- Common Methods.....

$$v' = \frac{v - \min_A}{\max_A - \min_A}$$

Min-max scaling

$$v' = \frac{v - \text{mean}_A}{\text{stand_dev}_A}$$

Z-score scaling

$$v' = \frac{v}{10^j}$$

Decimal scaling

Where j is the smallest integer such that $\text{Max}(|v'|) < 1$



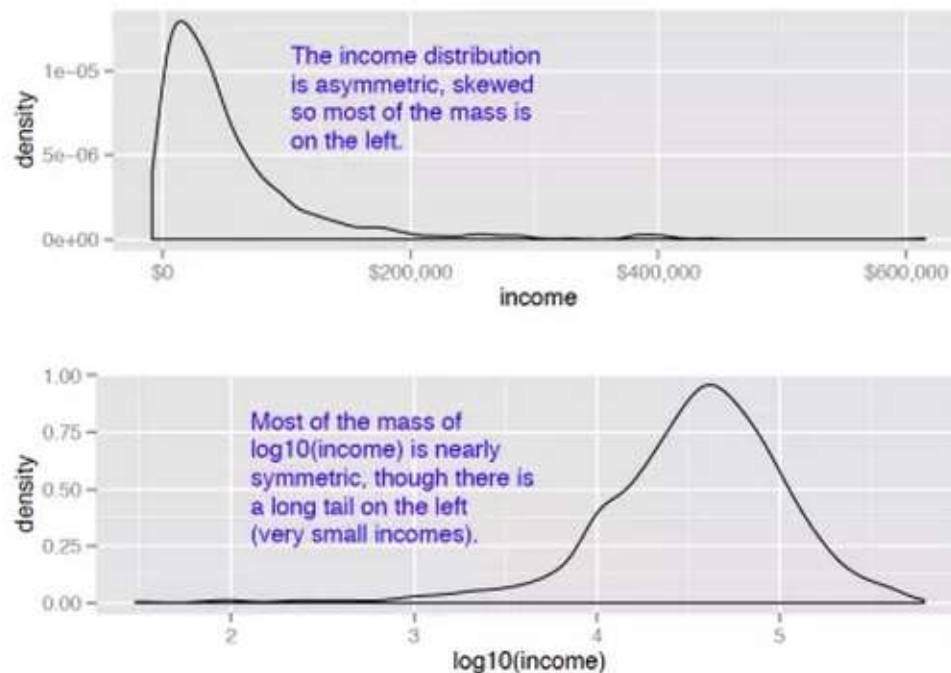
zero mean unit variance (ZMUV) normalization

X

For images this will normalise over different brightness and contrast ranges

Data Transformations - Logs

- Makes a skewed attribute more symmetric
- Reduces the magnitudes
- Common bases 10, 2, e (*which base to use is often not important*)



- Incomes, customer value, account or purchase sizes—are commonly encountered sources of skewed distributions in data science applications.
- Often they are log-normally distributed: the log of the data is normally distributed

Transforming Categorical Data

- How to handle...
 - Marital status = single, married, divorced, widowed?

- Could convert to a single new integer variable...
 - Marital status = 0,1,2,3 where
0 = single, 1=married, 2=divorced, 3=widowed

Integer Encoding

- Better to create four new T/F variables
 - Single = 0,1
 - Married = 0,1
 - Divorced = 0,1
 - Widowed = 0,1

One-Hot Encoding



Feature Construction

- Deriving a value that is more useful / making something more explicit, e.g.


| ID | Cost per unit | Units purchased |
|----|---------------|-----------------|
| 1. | 10 | 10 |
| 2. | 15 | 5 |
| 3. | 8 | 8 |
| 4. | 10 | 5 |



| ID | Cost per unit | Units purchased | Total \$ Revenue |
|----|---------------|-----------------|------------------|
| 1. | 10 | 10 | 100 |
| 2. | 15 | 5 | 75 |
| 3. | 8 | 8 | 64 |
| 4. | 10 | 5 | 50 |

Feature Construction : PCA

- PCA = Principle Component Analysis
- Generates new features as linear combinations of the old ones
- New features are created to explain the variance most efficiently
- Often used as a mechanism to do feature reduction and/or data visualisation



| | | | | | | | | | | | | |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 215.02 | 218.38 | 218.71 | 217.72 | 218.29 | 218.16 | 216.54 | 214.8 | 213.84 | 213.28 | 214.8 | 215.98 | 215.98 |
| 215.43 | 218.87 | 217.08 | 218.02 | 220.32 | 219.59 | 219.18 | 215.81 | 214.16 | 213.89 | 215 | 216.19 | 216.19 |
| 219.92 | 217.7 | 218.54 | 219.48 | 220.88 | 219.77 | 218.57 | 216.78 | 215.8 | 215.59 | 217.01 | 217.84 | 216.51 |
| 217.84 | 218.58 | 219.68 | 220.82 | 221.51 | 220.58 | 219.58 | 217.9 | 216.88 | 216.88 | 217.7 | 218.55 | 219.58 |
| 218.74 | 219.08 | 219.88 | 221.39 | 222.24 | 221.47 | 219.74 | 217.77 | 216.21 | 215.99 | 217.12 | 218.21 | 218.99 |
| 219.57 | -98.99 | -98.99 | -99.99 | 222.24 | 221.89 | 220.44 | 218.7 | 216.7 | 216.79 | 217.79 | 218.71 | 218.99 |
| 219.44 | 220.84 | 218.89 | 221.3 | 222.38 | 221.89 | 221.39 | 219.8 | 217.81 | 217.9 | 218.89 | 219.42 | 220.64 |
| 220.82 | 221.59 | 220.29 | 222.87 | 224.31 | 223.75 | 222.39 | 220.37 | 218.84 | 218.1 | 219.79 | 221.08 | 221.38 |
| 222.96 | 222.5 | 223.64 | 224.42 | 225 | 224.89 | 223.93 | 222.82 | 219.21 | 220.71 | 221.86 | 222.16 | 222.35 |
| 222.57 | 223.15 | 223.89 | 225.02 | 225.37 | 225.36 | 224.14 | 222.03 | 220.41 | 220.25 | 221.31 | 222.84 | 223.05 |
| 224 | 224.42 | 225.44 | 226.46 | 227.14 | 226.76 | 225.88 | 223.67 | 222.46 | 221.76 | 222.89 | 224.12 | 224.63 |
| 225.03 | 225.89 | 226.87 | 228.14 | 228.07 | 227.68 | 226.35 | 224.66 | 223.1 | 222.18 | 223.98 | 225.13 | 225.88 |
| 226.17 | 226.68 | 227.18 | 227.78 | 228.32 | 228.07 | 227.39 | 225.48 | 223.46 | 222.87 | 224.8 | 226.01 | 226.32 |
| 226.77 | 227.43 | 227.75 | 228.72 | 229.07 | 228.69 | 228.05 | 226.32 | 224.93 | 223.68 | 225.8 | 227.05 | 227.45 |
| 228.55 | 228.56 | 230.1 | 231.5 | 232.46 | 232.07 | 230.87 | 228.31 | 227.51 | 227.18 | 228.18 | 229.44 | 229.68 |
| 229.18 | 230.71 | 231.48 | 232.45 | 233.36 | 232.24 | 231.18 | 229.4 | 227.43 | 227.37 | 228.48 | 229.37 | 230.25 |
| 230.4 | 231.41 | 232.64 | 233.31 | 233.86 | 233.4 | 231.51 | 230.06 | 228.36 | 227.43 | 228.76 | 231.15 | 231.15 |
| 231.74 | 232.58 | 233.5 | 234.58 | 235.47 | 234.25 | 233.05 | 231.44 | 229.5 | 228.64 | 230.11 | 231.48 | 231.15 |
| 232.03 | 233.42 | 234.37 | 235.47 | 236.37 | 235.15 | 233.95 | 232.15 | 230.25 | 229.35 | 230.45 | 231.85 | 231.15 |
| 233.17 | 234.08 | 235.04 | 236.1 | 236.91 | 235.69 | 234.48 | 232.68 | 230.78 | 229.88 | 230.98 | 232.38 | 231.18 |
| 234.23 | 235.14 | 236.1 | 237.16 | 237.91 | 236.69 | 235.48 | 233.68 | 231.78 | 230.88 | 231.98 | 233.38 | 232.18 |
| 235.31 | 236.26 | 237.28 | 238.31 | 239.24 | 237.97 | 236.76 | 234.96 | 233.06 | 232.16 | 233.26 | 234.66 | 233.46 |
| 236.41 | 237.54 | 238.57 | 239.59 | 240.52 | 239.25 | 238.04 | 236.24 | 234.34 | 233.44 | 234.54 | 235.94 | 234.74 |
| 237.51 | 238.58 | 239.61 | 240.64 | 241.57 | 240.3 | 239.09 | 237.29 | 235.39 | 234.49 | 235.59 | 236.99 | 235.79 |
| 238.61 | 239.64 | 240.67 | 241.7 | 242.62 | 241.35 | 240.14 | 238.34 | 236.44 | 235.54 | 236.64 | 238.04 | 236.84 |
| 239.71 | 240.7 | 241.73 | 242.76 | 243.69 | 242.42 | 241.21 | 239.41 | 237.51 | 236.61 | 237.71 | 239.11 | 237.91 |
| 240.81 | 241.8 | 242.81 | 243.84 | 244.77 | 243.5 | 242.29 | 240.49 | 238.59 | 237.69 | 238.79 | 240.19 | 238.99 |
| 241.91 | 242.94 | 243.97 | 244.99 | 245.92 | 244.65 | 243.44 | 241.64 | 239.74 | 238.84 | 239.94 | 241.34 | 240.14 |
| 243.01 | 244.04 | 245.07 | 246.09 | 247.02 | 245.75 | 244.54 | 242.74 | 240.84 | 239.94 | 241.04 | 242.44 | 241.24 |
| 244.11 | 245.14 | 246.17 | 247.19 | 248.12 | 246.85 | 245.64 | 243.84 | 241.94 | 241.04 | 242.14 | 243.54 | 242.34 |
| 245.21 | 246.24 | 247.27 | 248.29 | 249.22 | 247.95 | 246.74 | 244.94 | 243.04 | 242.14 | 243.24 | 244.64 | 243.44 |
| 246.31 | 247.34 | 248.37 | 249.39 | 250.32 | 249.05 | 247.84 | 245.94 | 244.04 | 243.14 | 244.24 | 245.64 | 244.44 |
| 247.41 | 248.44 | 249.47 | 250.49 | 251.42 | 250.15 | 248.94 | 247.04 | 245.14 | 244.24 | 245.34 | 246.74 | 245.54 |
| 248.51 | 249.54 | 250.57 | 251.59 | 252.52 | 251.25 | 250.04 | 248.14 | 246.24 | 245.34 | 246.44 | 247.84 | 246.64 |
| 249.61 | 250.64 | 251.67 | 252.69 | 253.62 | 252.35 | 251.14 | 249.24 | 247.34 | 246.44 | 247.54 | 248.94 | 247.74 |
| 250.71 | 251.74 | 252.77 | 253.79 | 254.72 | 253.45 | 252.24 | 250.34 | 248.44 | 247.54 | 248.64 | 250.04 | 248.84 |
| 251.81 | 252.84 | 253.87 | 254.89 | 255.82 | 254.55 | 253.34 | 251.44 | 249.54 | 248.64 | 249.74 | 251.14 | 249.94 |
| 252.91 | 253.94 | 254.97 | 255.99 | 256.92 | 255.65 | 254.44 | 252.54 | 250.64 | 249.74 | 250.84 | 252.24 | 251.04 |
| 254.01 | 255.04 | 256.07 | 257.09 | 258.02 | 256.75 | 255.54 | 253.64 | 251.74 | 250.84 | 251.94 | 253.34 | 252.14 |
| 255.11 | 256.14 | 257.17 | 258.19 | 259.12 | 257.85 | 256.64 | 254.74 | 252.84 | 251.94 | 253.04 | 254.44 | 253.24 |
| 256.21 | 257.24 | 258.27 | 259.29 | 260.22 | 258.95 | 257.74 | 255.84 | 253.94 | 253.04 | 254.14 | 255.54 | 254.34 |
| 257.31 | 258.34 | 259.37 | 260.39 | 261.32 | 260.05 | 258.84 | 256.94 | 255.04 | 254.14 | 255.24 | 256.64 | 255.44 |
| 258.41 | 259.44 | 260.47 | 261.49 | 262.42 | 261.15 | 260.04 | 258.14 | 256.24 | 255.34 | 256.44 | 257.84 | 256.64 |
| 259.51 | 260.54 | 261.57 | 262.59 | 263.52 | 262.25 | 261.04 | 259.14 | 257.24 | 256.34 | 257.44 | 258.84 | 257.64 |
| 260.61 | 261.64 | 262.67 | 263.69 | 264.62 | 263.35 | 262.14 | 260.24 | 258.34 | 257.44 | 258.54 | 259.94 | 258.74 |
| 261.71 | 262.74 | 263.77 | 264.79 | 265.72 | 264.45 | 263.24 | 261.34 | 259.44 | 258.54 | 259.64 | 261.04 | 259.84 |
| 262.81 | 263.84 | 264.87 | 265.89 | 266.82 | 265.55 | 264.34 | 262.44 | 260.54 | 259.64 | 260.74 | 262.14 | 260.94 |
| 263.91 | 264.94 | 265.97 | 266.99 | 267.92 | 266.65 | 265.44 | 263.54 | 261.64 | 260.74 | 261.84 | 263.24 | 262.04 |

Original Data

Reduced Data

Feature Decomposition

- Decomposing compound features into simpler components, e.g....

| ID | Product Holdings | Purchased Service |
|-----|------------------|-------------------|
| 1. | "ProdA, ProdB" | Y |
| 2. | "ProdA, ProdB" | N |
| 3. | "ProdA, ProdB" | N |
| 4. | "ProdA, ProdB" | Y |
| ... | | |



| ID | ProdA | ProdB | ProdB | ProdB | Purchased Service |
|-----|-------|-------|-------|-------|-------------------|
| 1. | 1 | 0 | 1 | 0 | Y |
| 2. | 0 | 1 | 1 | 0 | N |
| 3. | 1 | 0 | 0 | 1 | N |
| 4. | 0 | 1 | 0 | 1 | Y |
| ... | | | | | |

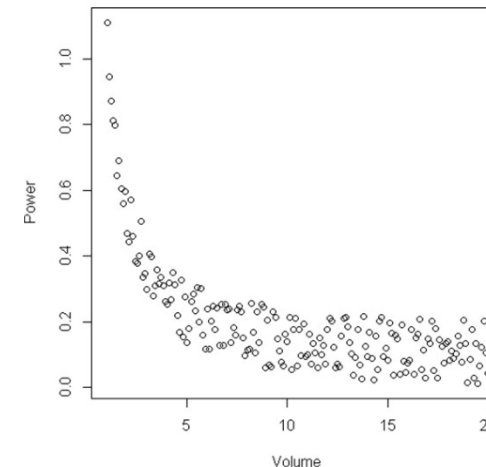
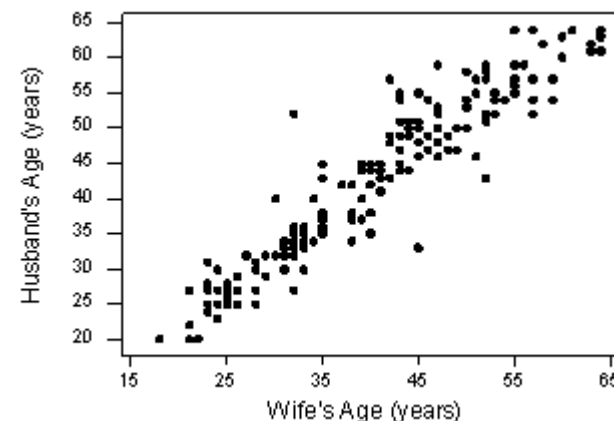
Feature Selection

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|---------|
| 215.62 | 216.71 | 217.72 | 218.73 | 219.74 | 220.75 | 221.76 | 222.77 | 223.78 | 224.79 | 225.80 | 226.81 | 227.82 | 228.83 | 229.84 | 230.85 | 231.86 | 232.87 | 233.88 | 234.89 | 235.90 | 236.91 | 237.92 | 238.93 | 239.94 | 240.95 | 241.96 | 242.97 | 243.98 | 244.99 | 245.00 | 246.01 | 247.02 | 248.03 | 249.04 | 250.05 | 251.06 | 252.07 | 253.08 | 254.09 | 255.10 | 256.11 | 257.12 | 258.13 | 259.14 | 260.15 | 261.16 | 262.17 | 263.18 | 264.19 | 265.20 | 266.21 | 267.22 | 268.23 | 269.24 | 270.25 | 271.26 | 272.27 | 273.28 | 274.29 | 275.30 | 276.31 | 277.32 | 278.33 | 279.34 | 280.35 | 281.36 | 282.37 | 283.38 | 284.39 | 285.40 | 286.41 | 287.42 | 288.43 | 289.44 | 290.45 | 291.46 | 292.47 | 293.48 | 294.49 | 295.50 | 296.51 | 297.52 | 298.53 | 299.54 | 300.55 | 301.56 | 302.57 | 303.58 | 304.59 | 305.60 | 306.61 | 307.62 | 308.63 | 309.64 | 310.65 | 311.66 | 312.67 | 313.68 | 314.69 | 315.70 | 316.71 | 317.72 | 318.73 | 319.74 | 320.75 | 321.76 | 322.77 | 323.78 | 324.79 | 325.80 | 326.81 | 327.82 | 328.83 | 329.84 | 330.85 | 331.86 | 332.87 | 333.88 | 334.89 | 335.90 | 336.91 | 337.92 | 338.93 | 339.94 | 340.95 | 341.96 | 342.97 | 343.98 | 344.99 | 345.00 | 346.01 | 347.02 | 348.03 | 349.04 | 350.05 | 351.06 | 352.07 | 353.08 | 354.09 | 355.10 | 356.11 | 357.12 | 358.13 | 359.14 | 360.15 | 361.16 | 362.17 | 363.18 | 364.19 | 365.20 | 366.21 | 367.22 | 368.23 | 369.24 | 370.25 | 371.26 | 372.27 | 373.28 | 374.29 | 375.30 | 376.31 | 377.32 | 378.33 | 379.34 | 380.35 | 381.36 | 382.37 | 383.38 | 384.39 | 385.40 | 386.41 | 387.42 | 388.43 | 389.44 | 390.45 | 391.46 | 392.47 | 393.48 | 394.49 | 395.50 | 396.51 | 397.52 | 398.53 | 399.54 | 400.55 | 401.56 | 402.57 | 403.58 | 404.59 | 405.60 | 406.61 | 407.62 | 408.63 | 409.64 | 410.65 | 411.66 | 412.67 | 413.68 | 414.69 | 415.70 | 416.71 | 417.72 | 418.73 | 419.74 | 420.75 | 421.76 | 422.77 | 423.78 | 424.79 | 425.80 | 426.81 | 427.82 | 428.83 | 429.84 | 430.85 | 431.86 | 432.87 | 433.88 | 434.89 | 435.90 | 436.91 | 437.92 | 438.93 | 439.94 | 440.95 | 441.96 | 442.97 | 443.98 | 444.99 | 445.00 | 446.01 | 447.02 | 448.03 | 449.04 | 450.05 | 451.06 | 452.07 | 453.08 | 454.09 | 455.10 | 456.11 | 457.12 | 458.13 | 459.14 | 460.15 | 461.16 | 462.17 | 463.18 | 464.19 | 465.20 | 466.21 | 467.22 | 468.23 | 469.24 | 470.25 | 471.26 | 472.27 | 473.28 | 474.29 | 475.30 | 476.31 | 477.32 | 478.33 | 479.34 | 480.35 | 481.36 | 482.37 | 483.38 | 484.39 | 485.40 | 486.41 | 487.42 | 488.43 | 489.44 | 490.45 | 491.46 | 492.47 | 493.48 | 494.49 | 495.50 | 496.51 | 497.52 | 498.53 | 499.54 | 500.55 | 501.56 | 502.57 | 503.58 | 504.59 | 505.60 | 506.61 | 507.62 | 508.63 | 509.64 | 510.65 | 511.66 | 512.67 | 513.68 | 514.69 | 515.70 | 516.71 | 517.72 | 518.73 | 519.74 | 520.75 | 521.76 | 522.77 | 523.78 | 524.79 | 525.80 | 526.81 | 527.82 | 528.83 | 529.84 | 530.85 | 531.86 | 532.87 | 533.88 | 534.89 | 535.90 | 536.91 | 537.92 | 538.93 | 539.94 | 540.95 | 541.96 | 542.97 | 543.98 | 544.99 | 545.00 | 546.01 | 547.02 | 548.03 | 549.04 | 550.05 | 551.06 | 552.07 | 553.08 | 554.09 | 555.10 | 556.11 | 557.12 | 558.13 | 559.14 | 560.15 | 561.16 | 562.17 | 563.18 | 564.19 | 565.20 | 566.21 | 567.22 | 568.23 | 569.24 | 570.25 | 571.26 | 572.27 | 573.28 | 574.29 | 575.30 | 576.31 | 577.32 | 578.33 | 579.34 | 580.35 | 581.36 | 582.37 | 583.38 | 584.39 | 585.40 | 586.41 | 587.42 | 588.43 | 589.44 | 590.45 | 591.46 | 592.47 | 593.48 | 594.49 | 595.50 | 596.51 | 597.52 | 598.53 | 599.54 | 600.55 | 601.56 | 602.57 | 603.58 | 604.59 | 605.60 | 606.61 | 607.62 | 608.63 | 609.64 | 610.65 | 611.66 | 612.67 | 613.68 | 614.69 | 615.70 | 616.71 | 617.72 | 618.73 | 619.74 | 620.75 | 621.76 | 622.77 | 623.78 | 624.79 | 625.80 | 626.81 | 627.82 | 628.83 | 629.84 | 630.85 | 631.86 | 632.87 | 633.88 | 634.89 | 635.90 | 636.91 | 637.92 | 638.93 | 639.94 | 640.95 | 641.96 | 642.97 | 643.98 | 644.99 | 645.00 | 646.01 | 647.02 | 648.03 | 649.04 | 650.05 | 651.06 | 652.07 | 653.08 | 654.09 | 655.10 | 656.11 | 657.12 | 658.13 | 659.14 | 660.15 | 661.16 | 662.17 | 663.18 | 664.19 | 665.20 | 666.21 | 667.22 | 668.23 | 669.24 | 670.25 | 671.26 | 672.27 | 673.28 | 674.29 | 675.30 | 676.31 | 677.32 | 678.33 | 679.34 | 680.35 | 681.36 | 682.37 | 683.38 | 684.39 | 685.40 | 686.41 | 687.42 | 688.43 | 689.44 | 690.45 | 691.46 | 692.47 | 693.48 | 694.49 | 695.50 | 696.51 | 697.52 | 698.53 | 699.54 | 700.55 | 701.56 | 702.57 | 703.58 | 704.59 | 705.60 | 706.61 | 707.62 | 708.63 | 709.64 | 710.65 | 711.66 | 712.67 | 713.68 | 714.69 | 715.70 | 716.71 | 717.72 | 718.73 | 719.74 | 720.75 | 721.76 | 722.77 | 723.78 | 724.79 | 725.80 | 726.81 | 727.82 | 728.83 | 729.84 | 730.85 | 731.86 | 732.87 | 733.88 | 734.89 | 735.90 | 736.91 | 737.92 | 738.93 | 739.94 | 740.95 | 741.96 | 742.97 | 743.98 | 744.99 | 745.00 | 746.01 | 747.02 | 748.03 | 749.04 | 750.05 | 751.06 | 752.07 | 753.08 | 754.09 | 755.10 | 756.11 | 757.12 | 758.13 | 759.14 | 760.15 | 761.16 | 762.17 | 763.18 | 764.19 | 765.20 | 766.21 | 767.22 | 768.23 | 769.24 | 770.25 | 771.26 | 772.27 | 773.28 | 774.29 | 775.30 | 776.31 | 777.32 | 778.33 | 779.34 | 780.35 | 781.36 | 782.37 | 783.38 | 784.39 | 785.40 | 786.41 | 787.42 | 788.43 | 789.44 | 790.45 | 791.46 | 792.47 | 793.48 | 794.49 | 795.50 | 796.51 | 797.52 | 798.53 | 799.54 | 800.55 | 801.56 | 802.57 | 803.58 | 804.59 | 805.60 | 806.61 | 807.62 | 808.63 | 809.64 | 810.65 | 811.66 | 812.67 | 813.68 | 814.69 | 815.70 | 816.71 | 817.72 | 818.73 | 819.74 | 820.75 | 821.76 | 822.77 | 823.78 | 824.79 | 825.80 | 826.81 | 827.82 | 828.83 | 829.84 | 830.85 | 831.86 | 832.87 | 833.88 | 834.89 | 835.90 | 836.91 | 837.92 | 838.93 | 839.94 | 840.95 | 841.96 | 842.97 | 843.98 | 844.99 | 845.00 | 846.01 | 847.02 | 848.03 | 849.04 | 850.05 | 851.06 | 852.07 | 853.08 | 854.09 | 855.10 | 856.11 | 857.12 | 858.13 | 859.14 | 860.15 | 861.16 | 862.17 | 863.18 | 864.19 | 865.20 | 866.21 | 867.22 | 868.23 | 869.24 | 870.25 | 871.26 | 872.27 | 873.28 | 874.29 | 875.30 | 876.31 | 877.32 | 878.33 | 879.34 | 880.35 | 881.36 | 882.37 | 883.38 | 884.39 | 885.40 | 886.41 | 887.42 | 888.43 | 889.44 | 890.45 | 891.46 | 892.47 | 893.48 | 894.49 | 895.50 | 896.51 | 897.52 | 898.53 | 899.54 | 900.55 | 901.56 | 902.57 | 903.58 | 904.59 | 905.60 | 906.61 | 907.62 | 908.63 | 909.64 | 910.65 | 911.66 | 912.67 | 913.68 | 914.69 | 915.70 | 916.71 | 917.72 | 918.73 | 919.74 | 920.75 | 921.76 | 922.77 | 923.78 | 924.79 | 925.80 | 926.81 | 927.82 | 928.83 | 929.84 | 930.85 | 931.86 | 932.87 | 933.88 | 934.89 | 935.90 | 936.91 | 937.92 | 938.93 | 939.94 | 940.95 | 941.96 | 942.97 | 943.98 | 944.99 | 945.00 | 946.01 | 947.02 | 948.03 | 949.04 | 950.05 | 951.06 | 952.07 | 953.08 | 954.09 | 955.10 | 956.11 | 957.12 | 958.13 | 959.14 | 960.15 | 961.16 | 962.17 | 963.18 | 964.19 | 965.20 | 966.21 | 967.22 | 968.23 | 969.24 | 970.25 | 971.26 | 972.27 | 973.28 | 974.29 | 975.30 | 976.31 | 977.32 | 978.33 | 979.34 | 980.35 | 981.36 | 982.37 | 983.38 | 984.39 | 985.40 | 986.41 | 987.42 | 988.43 | 989.44 | 990.45 | 991.46 | 992.47 | 993.48 | 994.49 | 995.50 | 996.51 | 997.52 | 998.53 | 999.54 | 1000.55 |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|---------|



Selecting the most relevant attributes

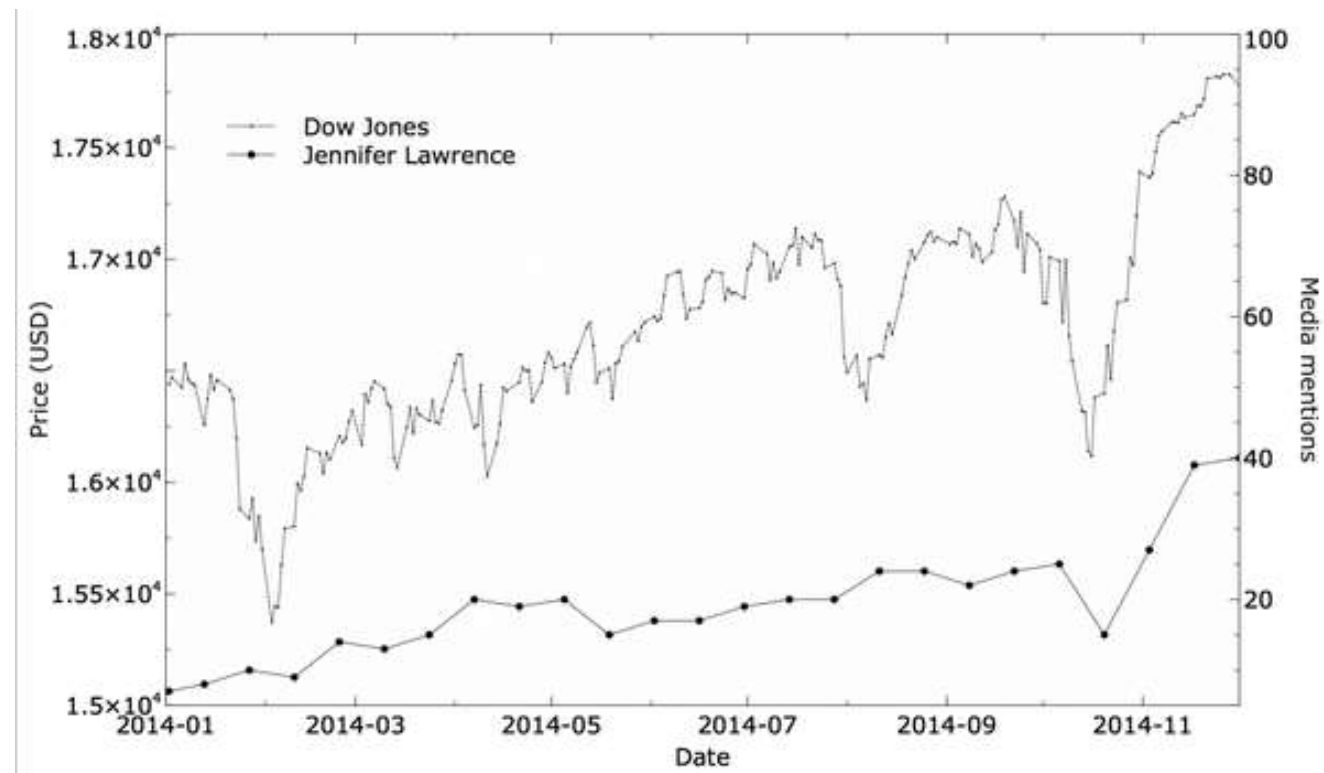
- Good input features are those that correlate with the target
 - Beware - the inputs and target may be correlated non-linearly, or linear only in a certain range
 - Sometimes two inputs that have poor correlation with the target when considered alone may correlate well when combined, e.g. predicting spending power from a family income and family size



Correlation vs Causation

- Correlation does not imply causation.
 - If A and B are correlated it does not mean A causes B or vice versa
- Model inputs need not be causal factors
 - BUT there should be some plausible link with the target

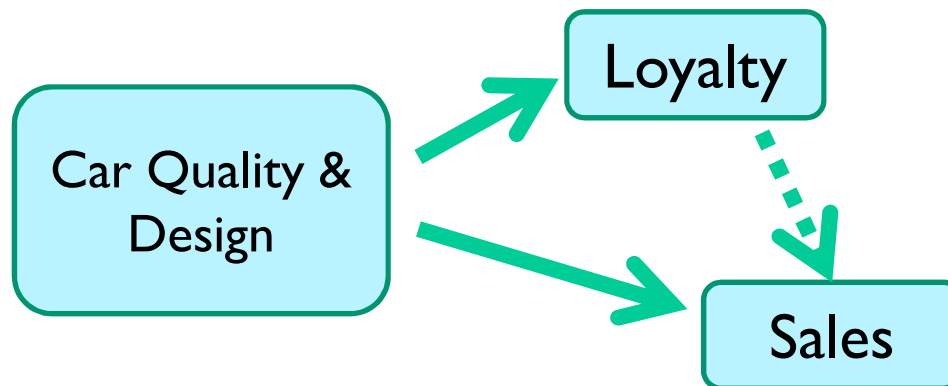
Both increase over the same period (correlation = 0.86) but there is no connection! We cannot predict the Dow Jones using Jennifer Lawrence



Correlation can be good!

- Correlation is often a result of hidden Causal Factors

Loyalty and sales may be correlated but loyalty does not (necessarily) cause sales. Instead a hidden (latent) causal factor may be at play....



...but loyalty may still be a useful predictor in the absence of suitable model input variables describing car design and car quality

Correlation can be bad!

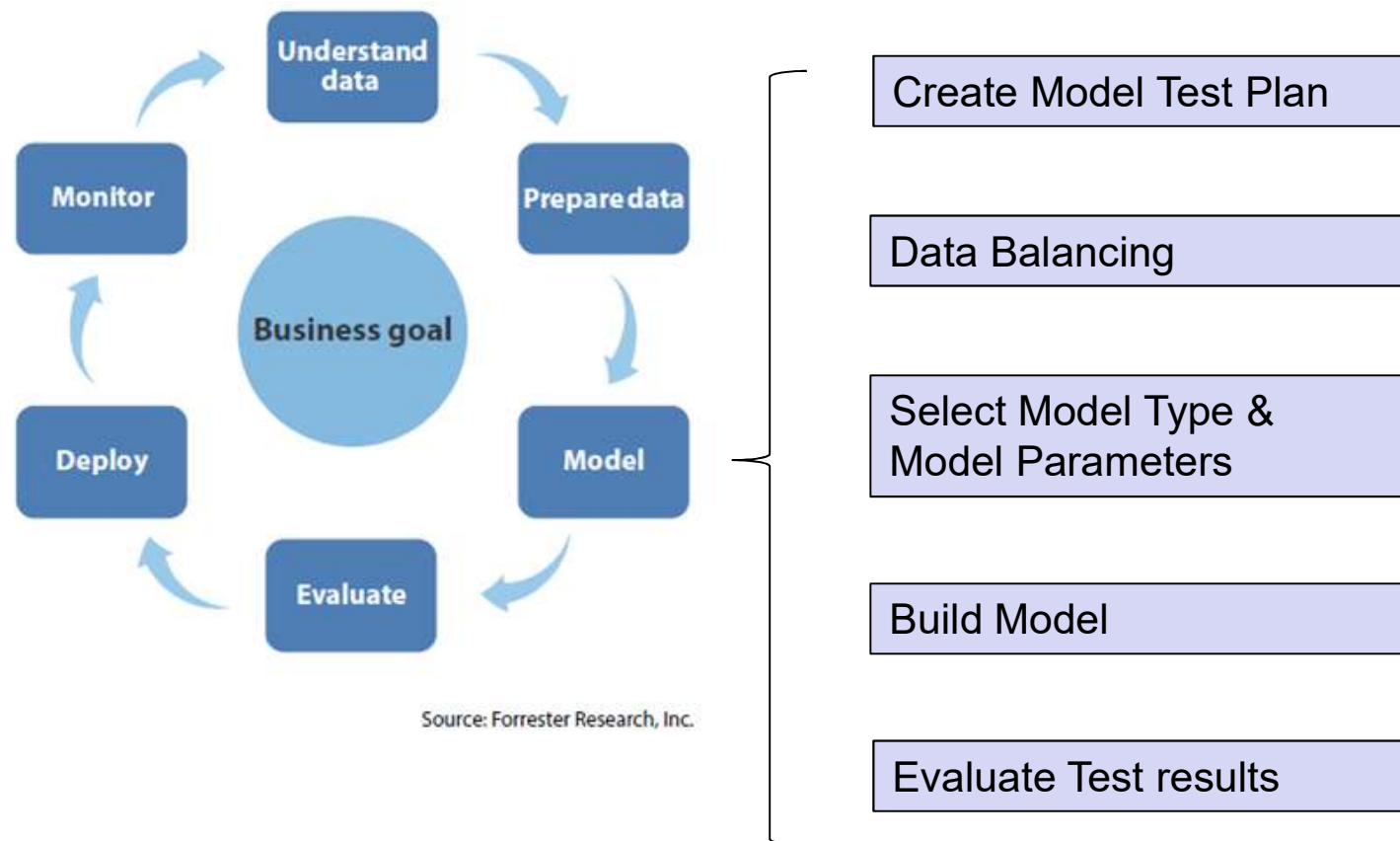
- Input features that are highly correlated may be redundant – can be omitted
- Input features that correlate with each other are not independent
 - Linear regression assumes there is no multi-collinearity between the inputs
- BUT many modelling methods can perform well even if inputs are correlated, so check the model performance using all input variables first as a **baseline** (if practical)
- Beware of **leakers** ~ inclusion of the target (in some disguised form) into the input variables
 - E.g. If we are predicting **Age** then we can't put **Date of Birth** as an input
 - Easy to do if you are not familiar with the problem domain and/or the variable names are not informative (e.g. "F23B")
 - CLUE: Very high correlation between an input variable and the target



Feature Engineering for IOT & Sensor Data

- Signal Processing cheat sheet?
- Time Series signal processing
- Contrast with tech analysis for stock data etc?
- <https://www.datasciencecentral.com/profiles/blogs/training-with-historical-data-surely-you-re-joking-says-the-iot>

Predictive Modeling Process



Selecting Training & Test sets

- In order to test the model, it is necessary to divide the data into training and test sets, e.g.



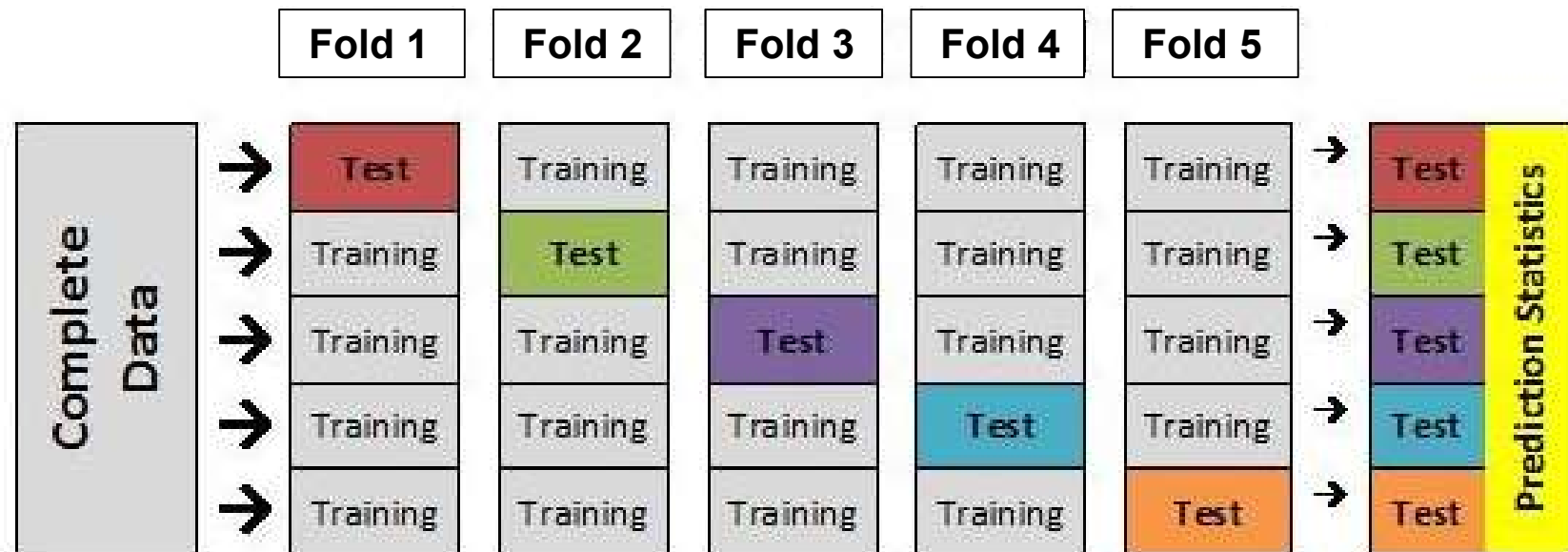
- Select the training data randomly (unless it's a time series problem).
- Usually select the test set to be everything not in the training set
- Simple split is good for large data sets

X

Selecting Training & Test Sets

■ K-fold Cross-validation

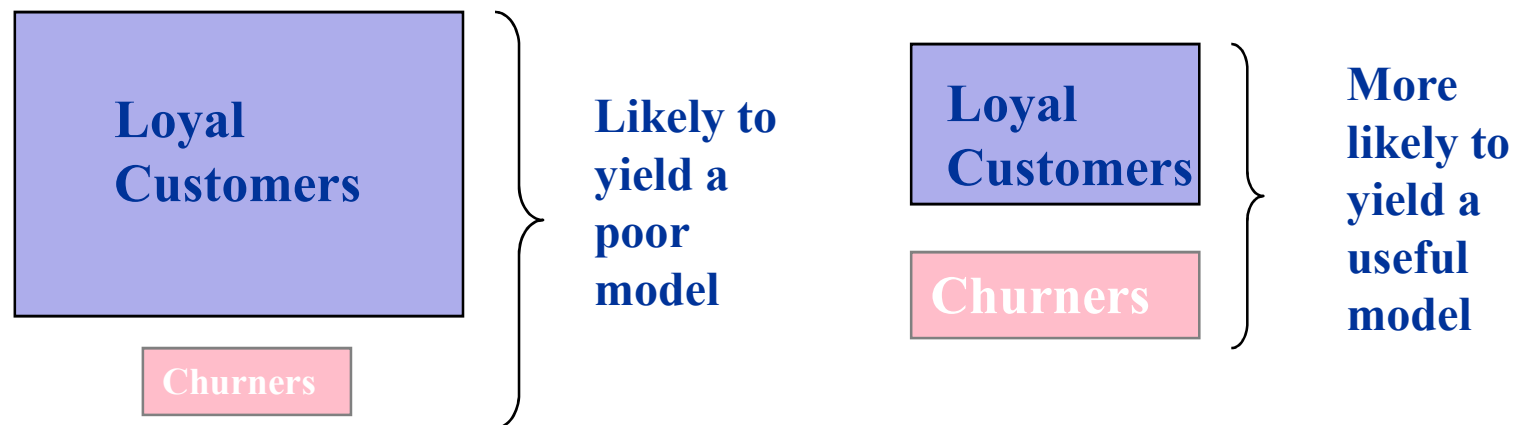
- Divide the data into k subsamples. Use $k-1$ as training data and one as test data
- Use for data set with moderate size



X

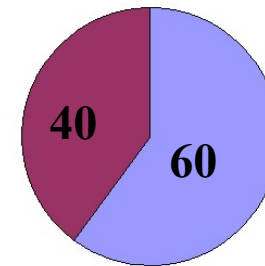
Data Balancing

- Unbalanced training data can result in poor models
- E.g. 9900 records where decision = loyal
100 records where decision = churning
 - The mined model is likely to be decision = loyal (99% correct!), very accurate but useless for predicting churn!
- Solution = balance the data as much as practical



Data Balancing

- It's not necessary to achieve equal size categories
 - Between 20/80 and 40/60 is OK
- Common methods use “oversampling”
 - Delete records from the most frequent category
 - Duplicate records in the less frequent category

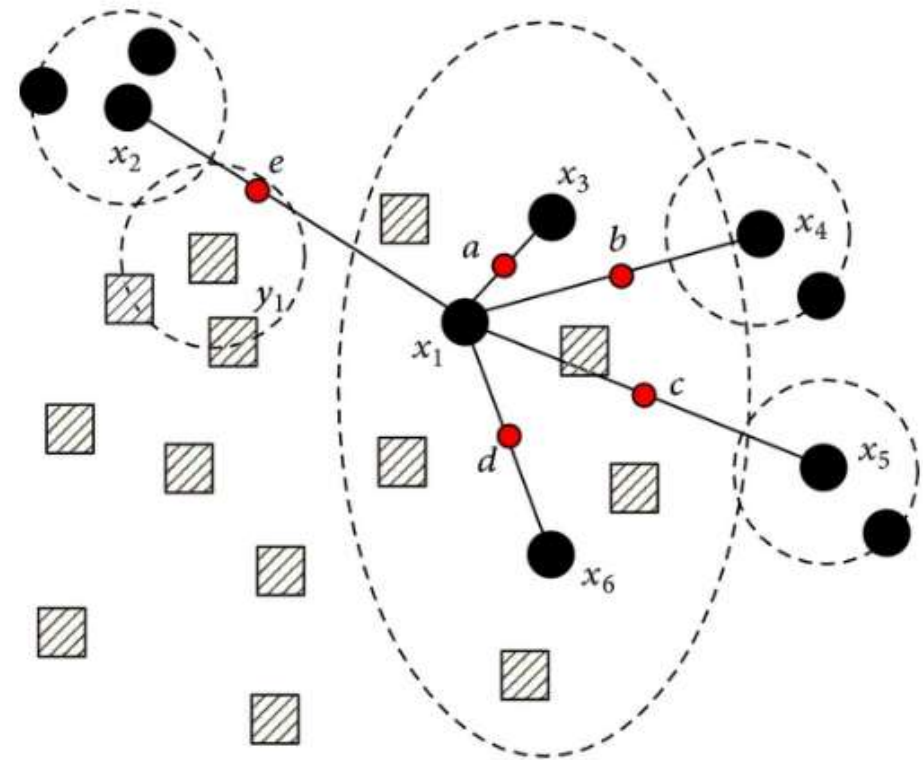


SMOTE (Synthetic Minority Oversampling)

- An over-sampling approach in which the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with replacement.

Add new minority class instances by:

- For each minority class instance c
 - neighbours = Get KNN(5)
 - n = Random pick one from neighbours
 - Create a new minority class r instance using c 's feature vector and the feature vector's difference of n and c multiplied by a random number
 - » i.e. $r.\text{feats} = c.\text{feats} + (c.\text{feats} - n.\text{feats}) * \text{rand}(0,1)$



- ▨ Majority class samples
- Minority class samples
- Synthetic samples

Classification Evaluation Metrics

| Actual \ Predicted | Predicted True | Predicted False | Sum |
|--------------------|----------------------|----------------------|-----|
| Actual True | True Positives (TP) | False Negatives (FN) | P |
| Actual False | False Positives (FP) | True Negatives (TN) | N |
| Sum | P' | N' | |

- Accuracy = $(TP + TN) / All$
- Error rate: $1 - \text{accuracy}$, or
 - Error rate = $(FP + FN) / All$
- Sensitivity: True Positive Rate
 - Sensitivity = TP / P
- Specificity: True Negative Rate
 - Specificity = TN / N

- Precision: what % of predictions are correct?
 - Precision = TP / P'
- Recall: what % of actual responders were predicted?
 - Recall = TP / P

X

The Confusion Matrix – Model A

- ◆ Be sure to use the right metric for the business problem at hand

| | Predicted buyer | Predicted Non-buyer | Total |
|------------------|-------------------------------|-------------------------------|-------|
| Actual Buyer | 200 <i>True Positives</i> | 100 <i>False Negatives</i> | 300 |
| Actual Non-Buyer | 800 <i>False Positives</i> | 1900 <i>True Negatives</i> | 2700 |
| Total | 1000 | 2000 | 3000 |

This model gets 70% correct (2100/3000, overall accuracy) but only 66% of the actual buyers (200/300) are captured by the model.

The Confusion Matrix – Model B

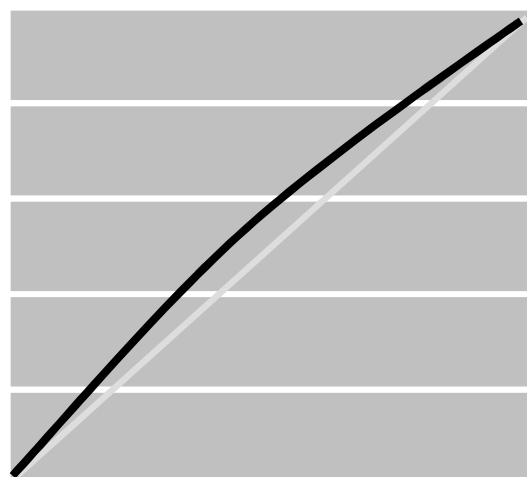
- Test results for a second model
- Is this a better model than the previous one?

| | Predicted buyer | Predicted Non-buyer | Total |
|------------------|--------------------------------|-------------------------------|-------|
| Actual Buyer | 280 <i>True Positives</i> | 20 <i>False Negatives</i> | 300 |
| Actual Non-Buyer | 1000 <i>False Positives</i> | 1700 <i>True Negatives</i> | 2700 |
| Total | 1280 | 1720 | 3000 |

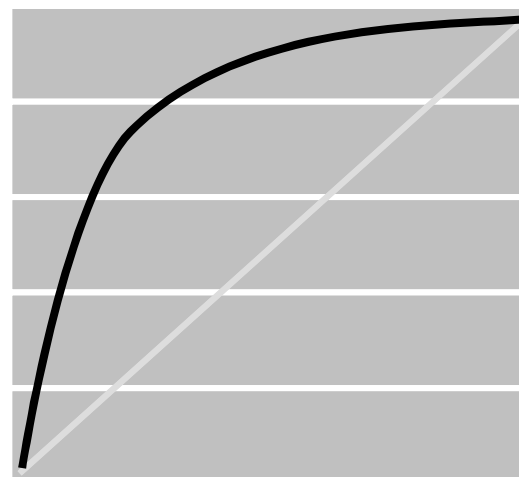
This model gets only 66% correct (1980/3000) but 93.3% of the actual buyers (280/300) are captured by the model.

Model Assessment with ROC Chart

weak model



strong model



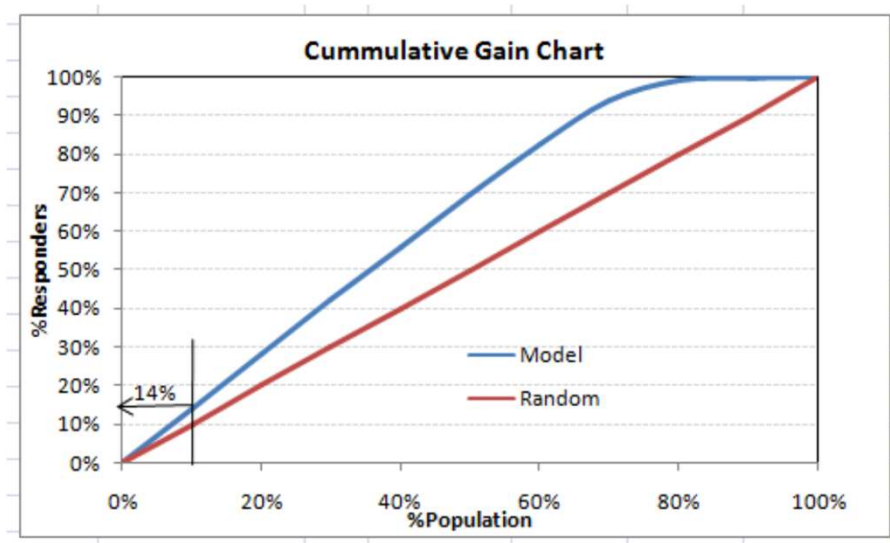
*True positive rate
(sensitivity)*

*False positive fraction
(1-specificity)*

*False positive fraction
(1-specificity)*

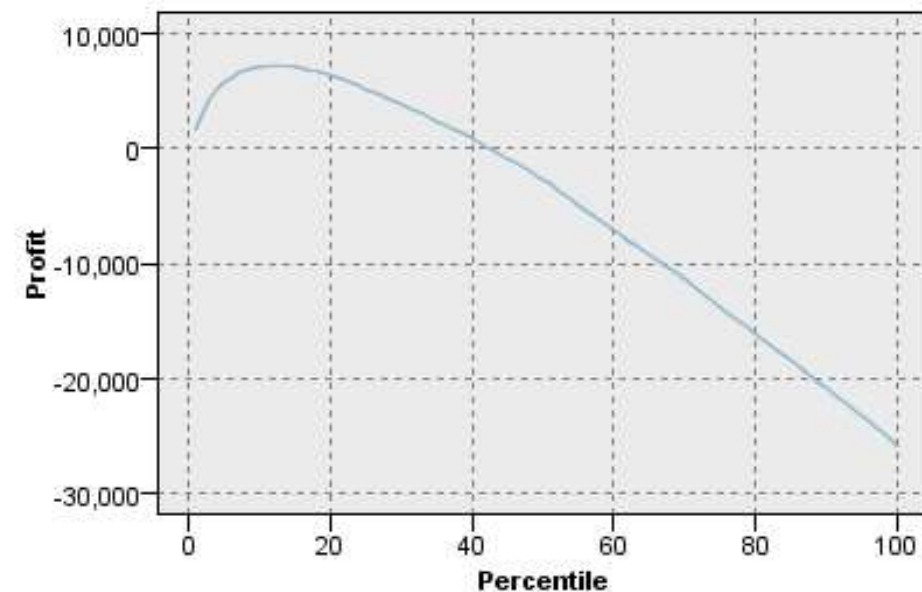
X

Model Assessment with Gains Chart



In many situations a cumulative profit (or ROI) chart can add more insight:

E.g. If you know the cost of sending a mailing and the revenue obtained when a positive response occurs



Cumulative Profit

X

Comparison with Baselines

- The accuracy required by a model can only be determined by examining the (business) context in which it will be used
- E.g. For ad click models,
 - Overall **display ad CTR** is 0.05%, rich media **ad CTR** is 0.1%, and Facebook **Ad CTR** ranges from 0.5% to 1.6%. Non-targeted CTR ~ 0.05% (DoubleClick, 2018)
- Compare Model results against a Baseline model
 - E.g. what was the accuracy if random guessing was used?
- Compare Model results against a quick First-Try model

Always start with a stupid model, no exceptions.

How to efficiently build Machine Learning powered products.



Emmanuel Ameisen

Follow

Mar 7, 2018 · 9 min read



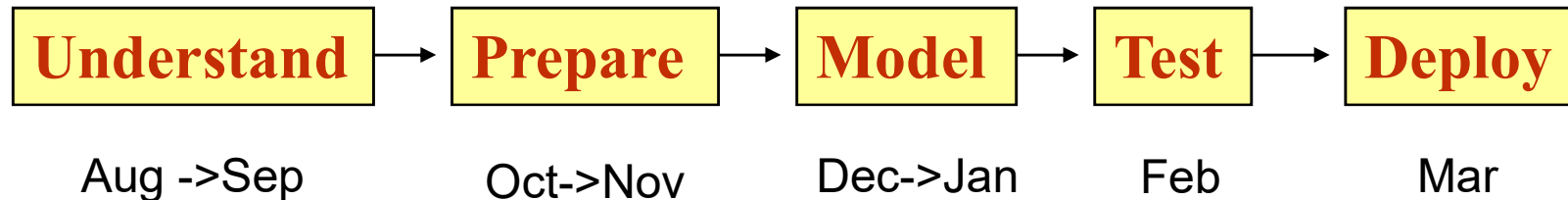
Fundamentally, a baseline is a model that is both simple to set up and has a reasonable chance of providing decent results. Experimenting with them is usually quick and low cost

A baseline will take you less than 1/10th of the time, and could provide up to 90% of the results

<https://blog.insightdatascience.com/always-start-with-a-stupid-model-no-exceptions-3a22314b9aaa>

Overall Execution: Incorporating Agile

Its not linear pass through the various stages....



Instead... Try to work in short iterations. Each iteration generates a prototype working system which is tested and then refined and improved over time – using test results and feedback from stakeholders

Caution: Data Preparation can often occupy the majority of the project time – often up to 80%!

Being Agile – Frequent iterations

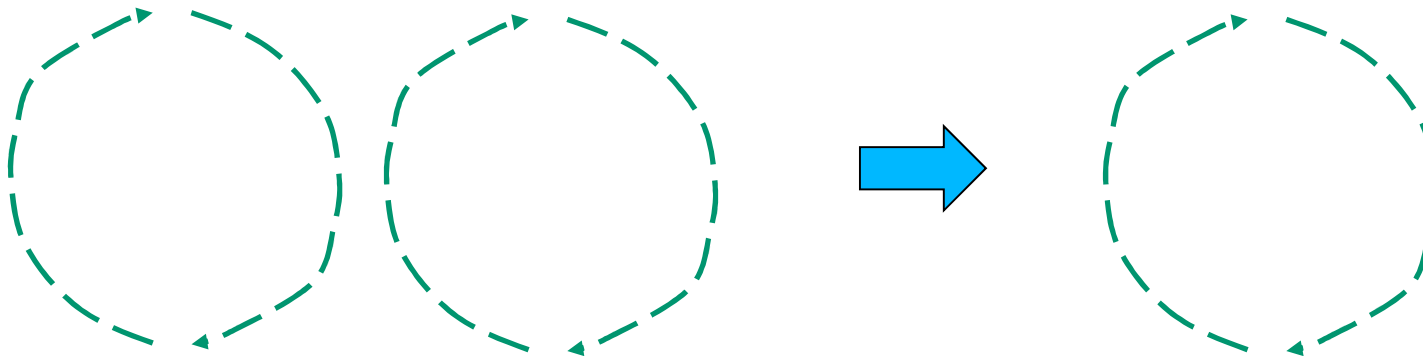
Increasing Business Understanding ->

Increasing Data Preparation ->

Improving Models ->

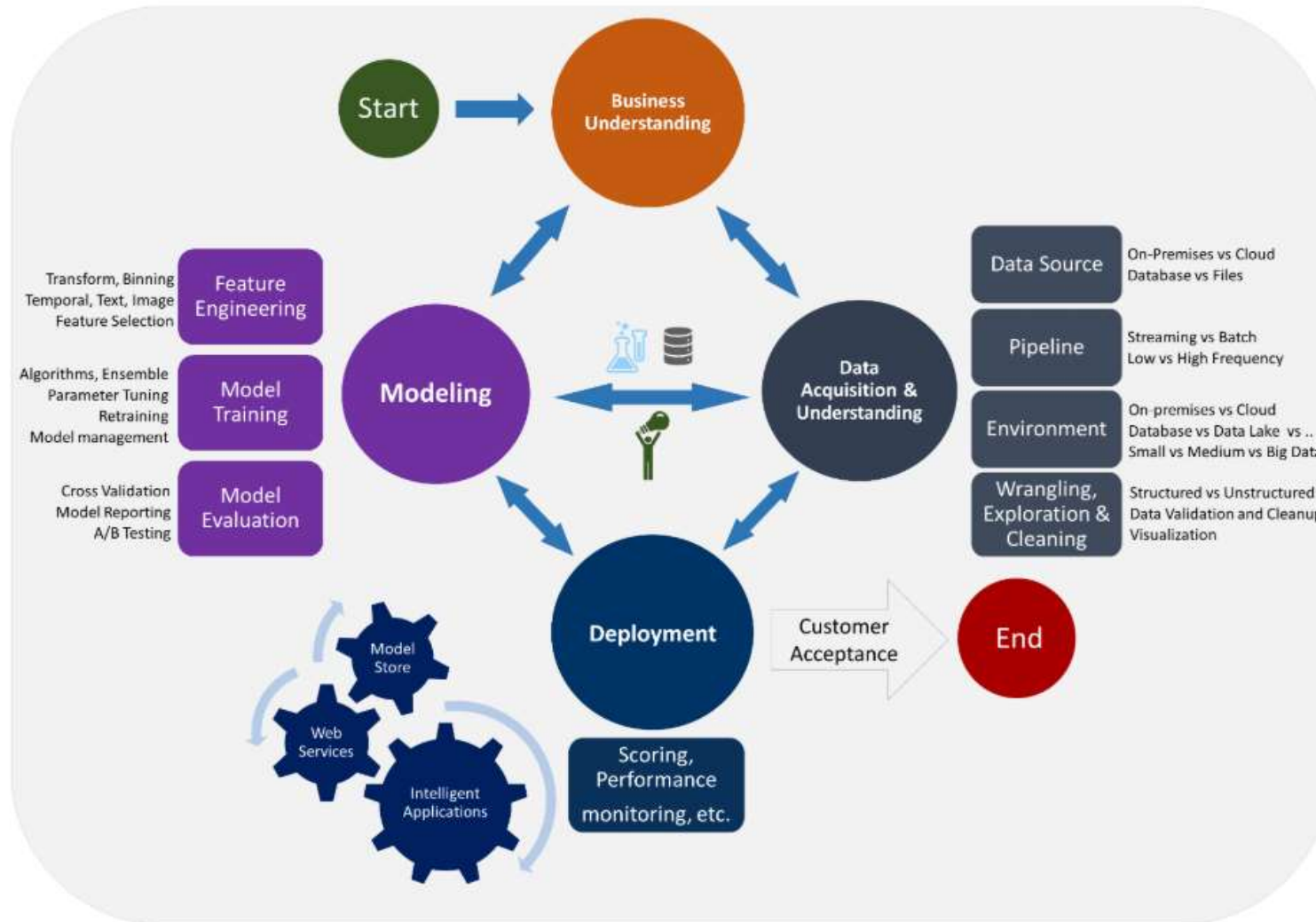
Lab Testing ->

Production Testing ->



Microsoft TDSP (Team Data Science Process)

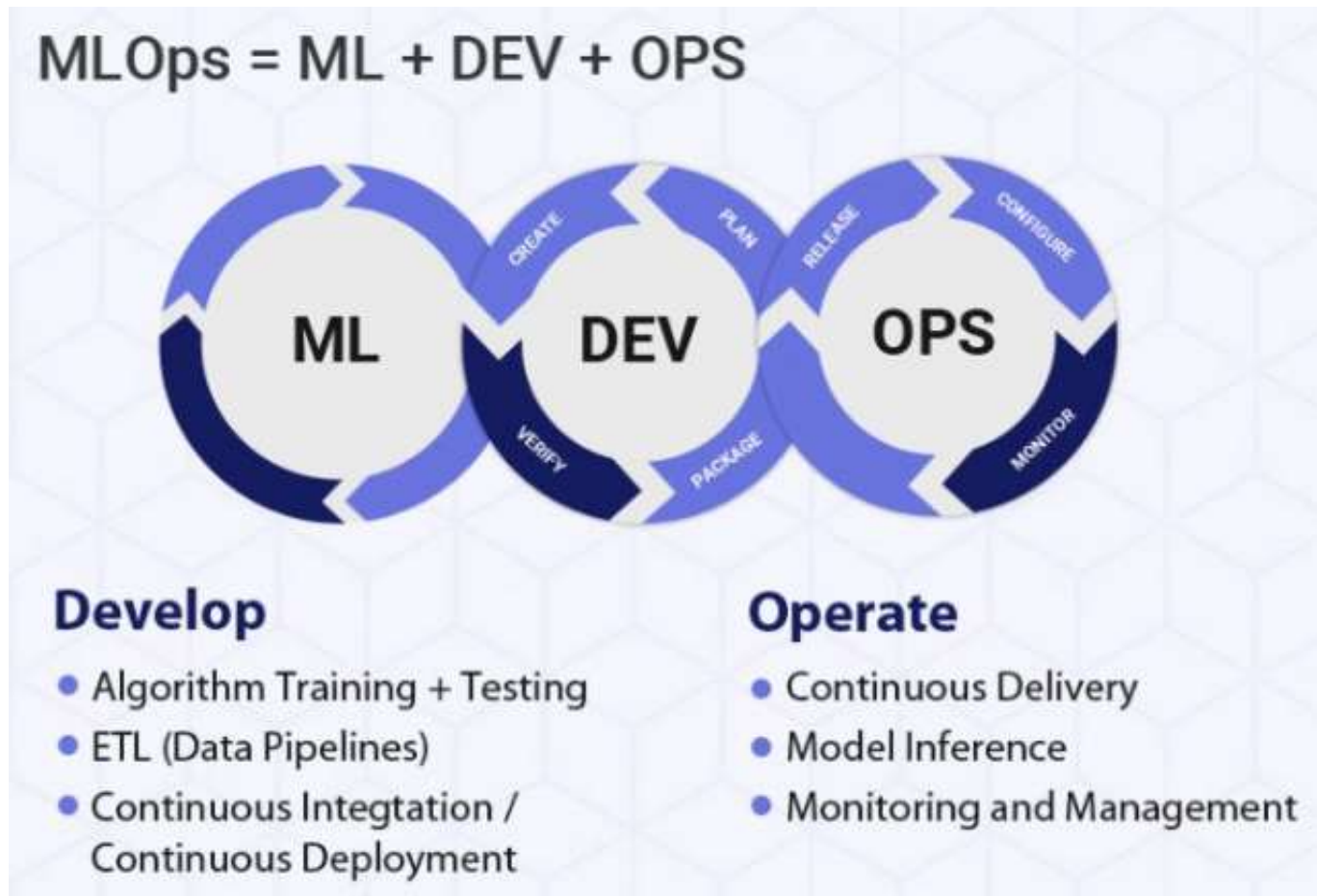
Data Science Lifecycle



<https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview>

From DevOps to MLOps

- Developing and deploying ML systems can be relatively fast and cheap, but maintaining them over time is often difficult and expensive.



ML Lifecycle Management Issues (e.g.'s)

■ Entanglement

- ❖ Changing Anything Changes Everything (CACE). Applies to input signals, hyper-parameters, learning settings, sampling methods, convergence thresholds, data selection etc

■ Correction Cascades

- ❖ A model $m1$ for a problem may exist, but a model for a slightly different problem is needed
- ❖ A fast solution is to learn a model $m2$ that takes $m1$ as input and learns a small correction
- ❖ But this creates a system dependency on $m1$, making it expensive to change $m1$ in future
- ❖ The cost increases when correction models are cascaded

■ Undeclared Consumers

- ❖ Often, a prediction from a model $m1$ is made widely accessible, either at runtime or by writing to files or logs that may later be consumed by other systems.
- ❖ Without access controls, some of these consumers may be undeclared, silently using $m1$ output as an input to another system.
- ❖ Changes to $m1$ will very likely impact these other parts, potentially in ways that are unintended, poorly understood, and detrimental. Undeclared consumers may also create hidden feedback loops

<https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf>