Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Business Analytics using Data Mining and Forecasting

BU7143 & BU7144

Dr. Nicholas P. Danks
Business Analytics
nicholas.danks@tcd.ie

How much coding experience do you have?

1. I am a code Jedi!
2. Quite a lot – I can create an iterative structure
3. Some, but I am not confident
4. Quite a little – I don’t know what a function is!
5. None

How much statistics experience do you have?

1. I can do complex derivations by hand!
2. Quite a lot – I can explain the mechanics of a linear regression
3. Some, but I am not confident
4. Quite a little –what is a mean?
5. None

Tools we will use

Coding language

Install R:
http://www.r-project.org/

Integrated Development Environment

Install RStudio:
http://www.rstudio.com/

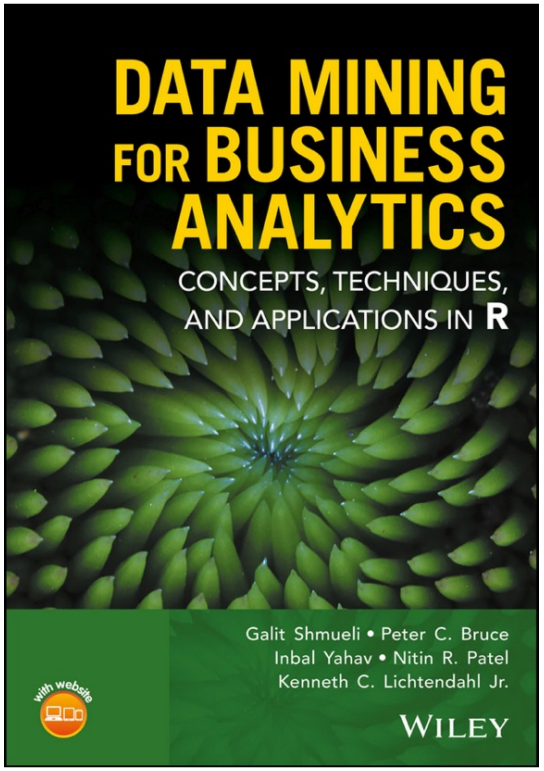
Version control

Join GitHub:
https://github.com/

Textbook

Data Mining for Business Analytics in R

Shmueli, Bruce, Yahav, Patel & Lichtendahl



(Merged) Outline

Session	Date & Venue	Lecture & readings
1		Introduction: Business and Statistical Challenges Classification, Prediction, Forecasting, Clustering, Supervised, etc. Reading: Chap. 2
2		**Dimension reduction, & Performance evaluation** Reading: Chap. 4 and 5
3		**General Regression**: Explanation and Prediction, Stationarity, Variable types Reading: Chap. 6
4		**Time Series Data**: Linear Regression with ext. predictors; Lags; Trend, Seasonality, Level, Noise Reading: Chap. 16 (and feedback on projects)
5		**Smoothing**: Simple, & Exponential smoothing Reading: Chap 17 and 18
7		**Group projects - Forecasting** presentation and feedback Reading: TBD

Grading

Presentation (20%)
15 mins
±10 Slides
Written Report (40%)
4 – 6 pages
1500 words
Homework (40%)
Weekly (40/6 = 6.7% per lesson)

ASSESSMENT
Group Assignment (60%)
The group assignment will take the form of a detailed business challenge translated into a statistical forecasting problem. It will detail the application of several possible methods for generating forecasts, their relative suitability and performance. Students will be evaluated on the business insights and conclusions, predictive performance, and ability to communicate effectively. It will include a group presentation and written. The deadline to submit your assignment is included in the schedule.

Weekly Homework (40%)
Weekly homework assignment will track the progress and learning of students. To help participants prepare for the homework, weekly tutorials will be held to discuss the homework problems.

Presentation & Report

Executive Summary

Problem description

Business goal:

Analytics goal:

Data description

Brief data preparation / cleaning details

Datamining solution

Comparison of performance

Conclusions

Advantages and Limitations

Operational Recommendations

Refer to the demo report and presentation (Blackboard)

Business Problem -> Statistical Problem

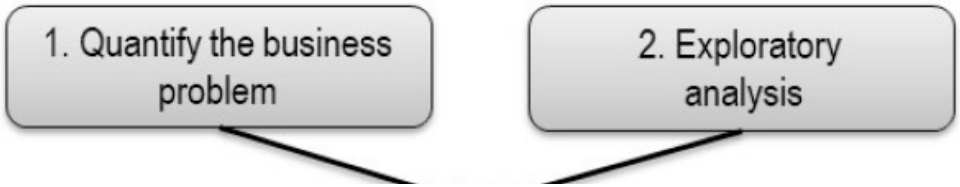
1. Understand & Define the problem
- *Frame the business problem*
- *Prepare for a decision*

2. Set analytic goals and scope your solution
- *Set objectives and define milestones*
- *Design minimum viable product*
- *Identify target metrics*

3. Plan the analysis
- *Plan your datasets*
- *Plan your methods*

https://ayotomiwasalau.medium.com/translate-business-problems-to-data-science-problems-b1693d8f313a

Quantifying the Business Problem and Exploratory Analysis



Conducted in tandem

- The business problem defines what you want to do
- Exploratory analysis provides constraints on what you can do

A business goal is often defined in an abstract manner with implicit meaning:
"We want to target our best customers."
Clarify and quantify:

- WHAT makes a best customer (lifetime value, purchases, $ or unit, profit?)
- What criteria make them BEST (over $50,000, tenure?)
- What Databases are available (sales, manufacturing, marketing?
- What data is stored in the database (individual sales, reports, costs?)
- Is data available real-time or periodically?
- How is the role currently served? What processes and data?

"Identify customers with a potential annual gross profit of over $25,000"

Pandemic Example

What **data** do we have?

How can it be **converted**?

What can be **predicted**?

What is the **business value**?

https://youtu.be/TGahNuPH9LY

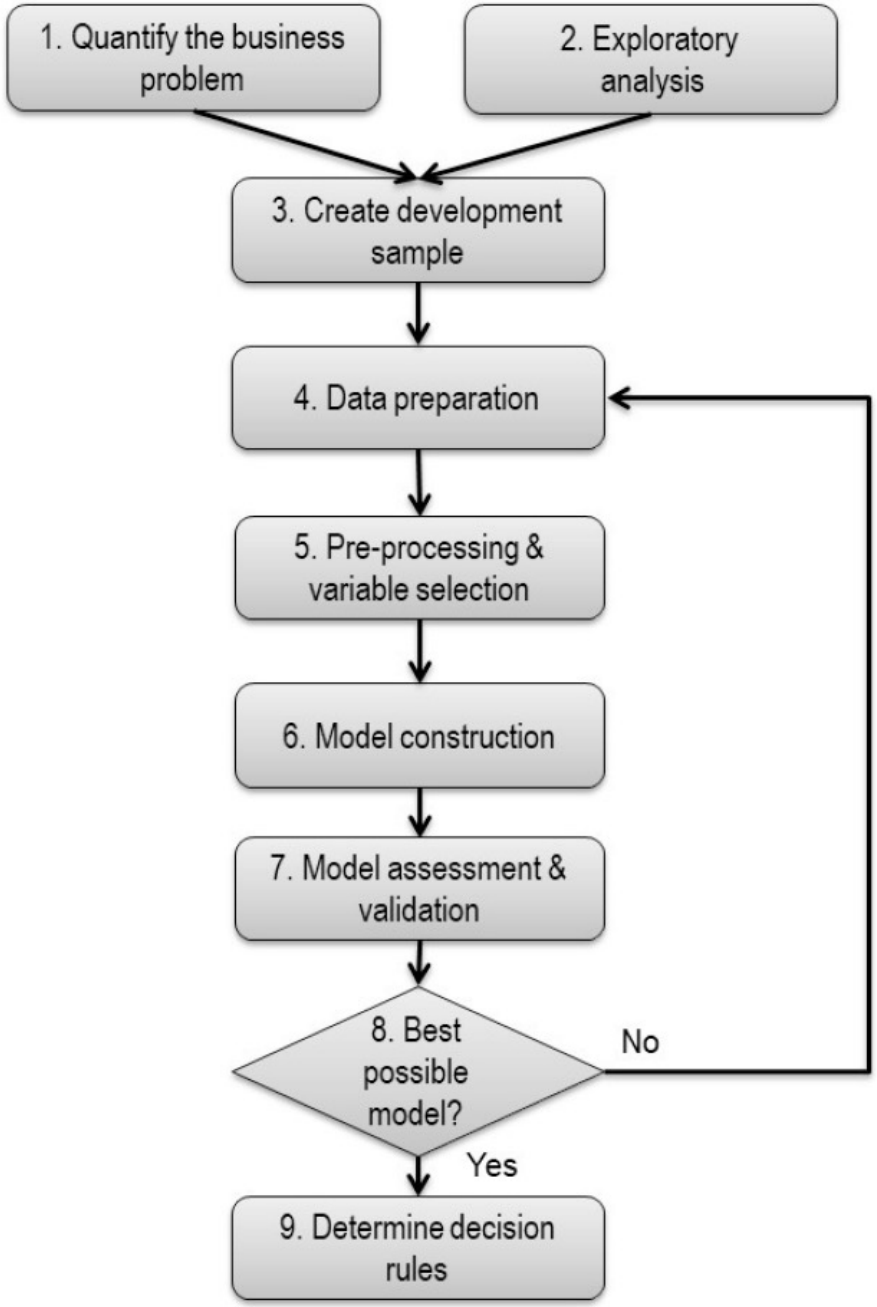
Vendor serial number	User phone number	Timestamp
111-111-111-111	0851991999	10:45:22-21:05:2021

Building a Forecasting or Predictive Model

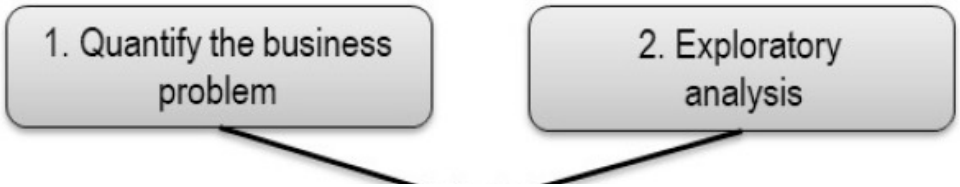
Most Predictive / AI processes can be broken down into a series of logical steps

Each step has its own considerations and opportunities for error

No step is more/less important



Quantifying the Business Problem and Exploratory Analysis



Conducted in tandem

- The business problem defines what you want to do
- Exploratory analysis provides constraints on what you can do

A business goal is often defined in an abstract manner with implicit meaning:
"We want to target our best customers."
Clarify and quantify:

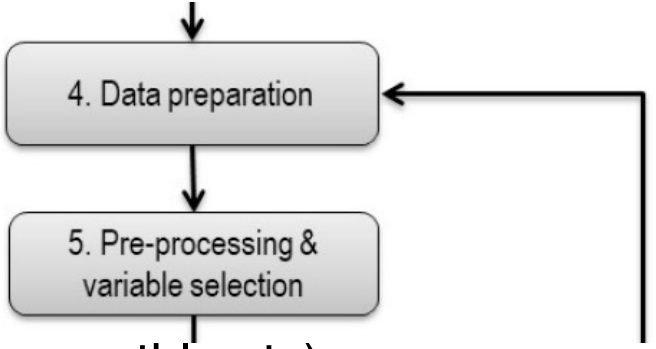
- WHAT makes a best customer (lifetime value, purchases, $ or unit, profit?)
- What criteria make them BEST (over $50,000, tenure?)
- What Databases are available (sales, manufacturing, marketing?
- What data is stored in the database (individual sales, reports, costs?)
- Is data available real-time or periodically?
- How is the role currently served? What processes and data?

"Identify customers with a potential annual gross profit of over $25,000"

Data considerations

- Out-of-date data
- Not representative of the target population
- Stability of data
- Legal and ethical reasons
- Deterministic cases
- Inexplicability

Processing data



Creating new data

- Age vs DOB
- Granularity of data (converting daily purchases to monthly etc)
- IP address -> city name

Data cleaning

- missing or incorrect?
- Remove or recode missing (missingness contains info?)
- Remove duplications

Consolidation

- Ensure consistency in data

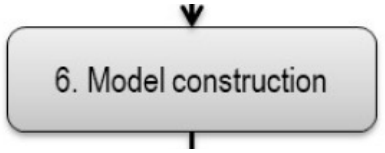
Conversion to numeric

- “Yes”/”No”/”Maybe” -> 0/1/2 dummy variables
- Twitter feeds converted to key words or summarized for intent

Dimension reduction

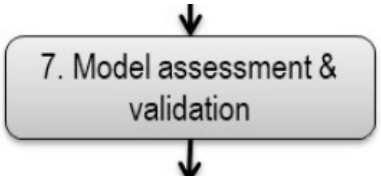
Standardization

Model Construction



- ML Algorithm is trained on the **development** data
- **Parameters** and **hyper-parameters** are set
- Predictive model is produced at the end
- Set of rules and logic statements
- Application to data generates predictions

Model Evaluation



Evaluate the **costs** and **performance**
Apply the model to a “testing” or ”validation set”
Calculate accuracy on this set

The method for evaluating **accuracy** is important
And the **costs** of inaccuracy are important

I.e. covid test vs marketing dollars

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,

$\hat{y}$ – *predicted value of* y
$\bar{y}$ – *mean value of* y

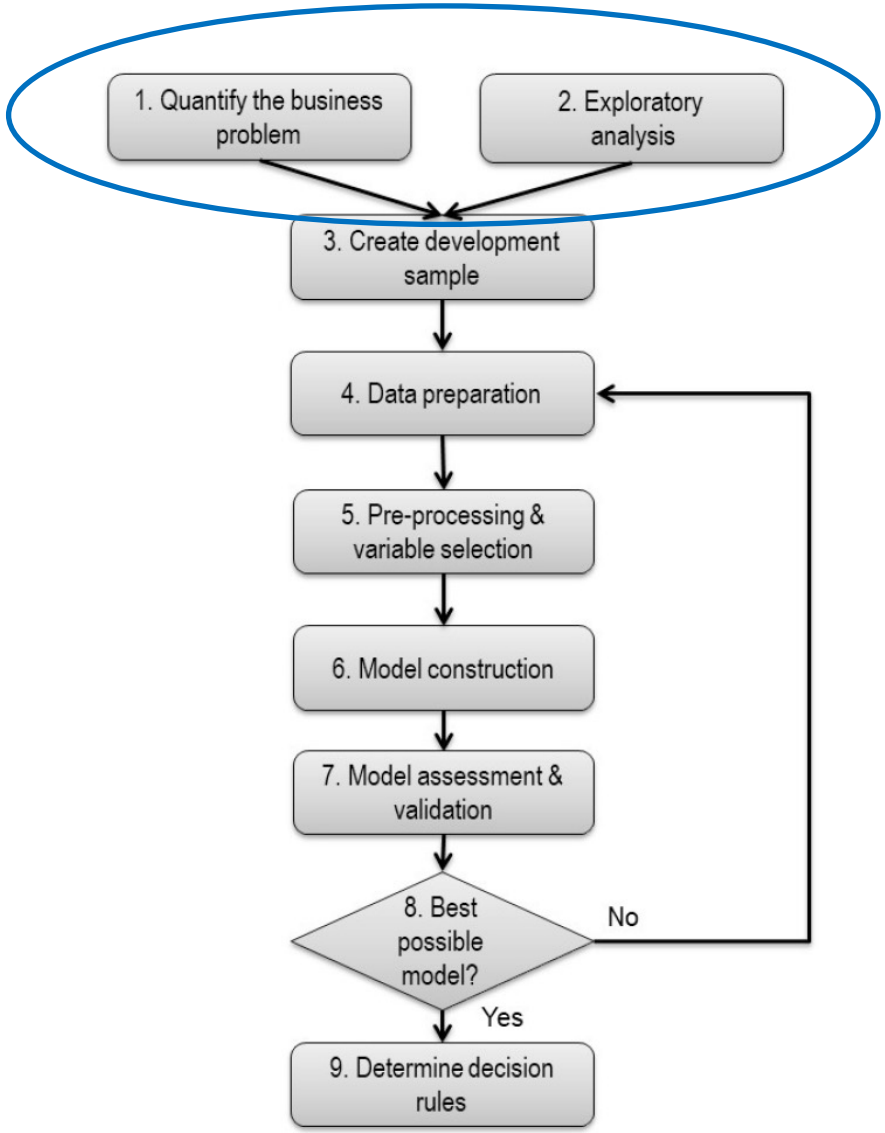
True Response \ Predicted Response	$\hat{y} = 1$	$\hat{y} = 0$	
$y = 1$	True Positive	False Negative	Recall (Sensitivity) TP/(y=1)
$y = 0$	False Positive	True Negative	Specificity TN/(y=0)
	Precision TP/(ŷ=1)		Accuracy (TP+TN)/total

Business Problem

VP of Marketing: “We’re losing customers. We need to identify customers before they leave so that we can target them with marketing offers.”
Data Scientist: “The only data we have is:

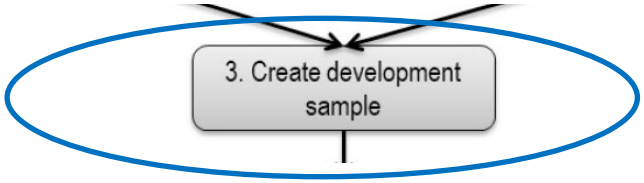
- Simple demographic data
- Products subscribed to
- Contract details
- Charges incurred

kaggle



Data

kaggle



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneServic	MultipleLine	InternetServ	OnlineSecuri	OnlineBacku	DeviceProte	TechSupport	StreamingTV	StreamingM	Contract	PaperlessBill	PaymentMet	MonthlyChar	TotalCharges	Churn
2	7590-VHVEG	Female	0	Yes	No	1	No	No phone se	DSL	No	Yes	No	No	No	No	Month-to-m	Yes	Electronic ch	29.85	29.85	No
3	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
4	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-m	Yes	Mailed check	53.85	108.15	Yes
5	7795-CFOCW	Male	0	No	No	45	No	No phone se	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank transfe	42.3	1840.75	No
6	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-m	Yes	Electronic ch	70.7	151.65	Yes
7	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to-m	Yes	Electronic ch	99.65	820.5	Yes
8	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	No	Month-to-m	Yes	Credit card (	89.1	1949.4	No
9	6713-OKOM	Female	0	No	No	10	No	No phone se	DSL	Yes	No	No	No	No	No	Month-to-m	No	Mailed check	29.75	301.9	No
10	7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	Month-to-m	Yes	Electronic ch	104.8	3046.05	Yes
11	6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank transfe	56.15	3487.95	No
12	9763-GRSKD	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No	Month-to-m	Yes	Mailed check	49.95	587.45	No
13	7469-LKBCI	Male	0	No	No	16	Yes	No	No	No internet s	No internet s	No internet s	No internet s	No internet s	No internet s	Two year	No	Credit card (	18.95	326.8	No
14	8091-TTVAX	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	One year	No	Credit card (	100.35	5681.1	No
15	0280-XJGEX	Male	0	No	No	49	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	Month-to-m	Yes	Bank transfe	103.7	5036.3	Yes

The data scientist creates the development sample for us.
By accessing the database for all customers in the past month.
She then tags all customers who unsubscribed with **Churn: Yes**.

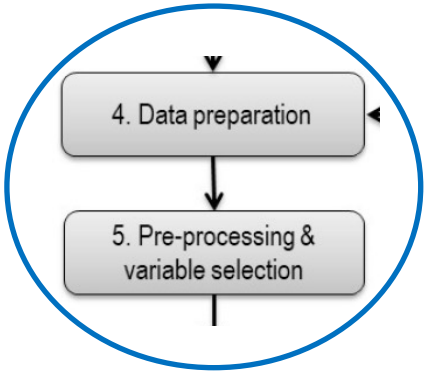
Rows: 7043

Columns: 21

Target: Churn

Open it up in spreadsheet software and have a look.

Data Preparation



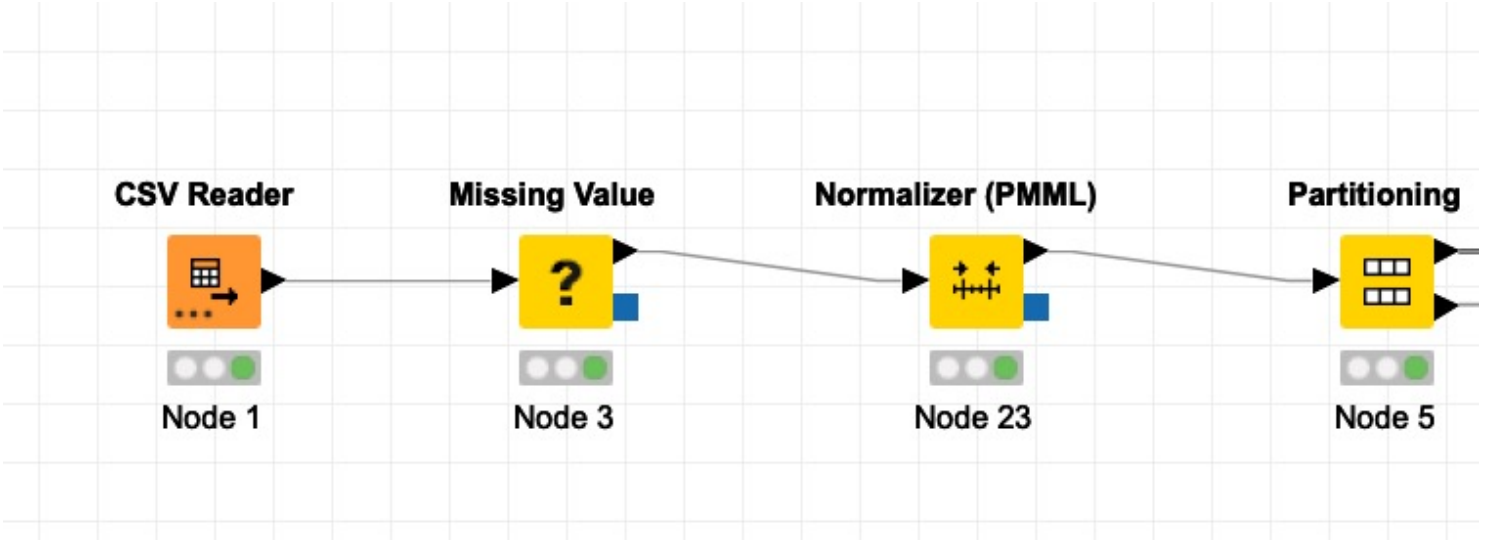
KNIME readily handle categorical vars

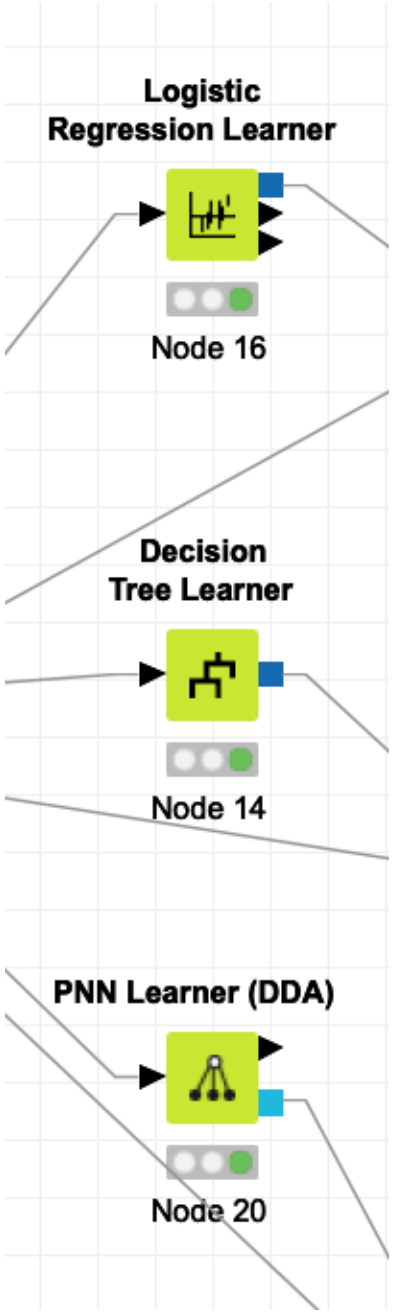
Standardize the **numeric** vars

Remove rows with **missing values**

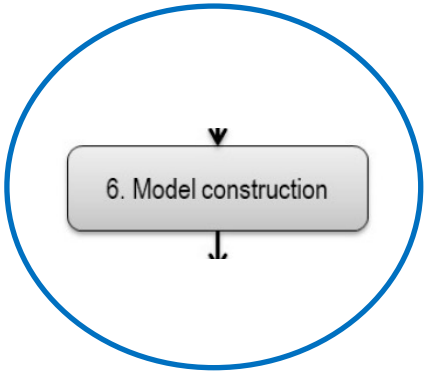
We want to **partition** our data:

- **Training**
- **Validation set**





Model construction

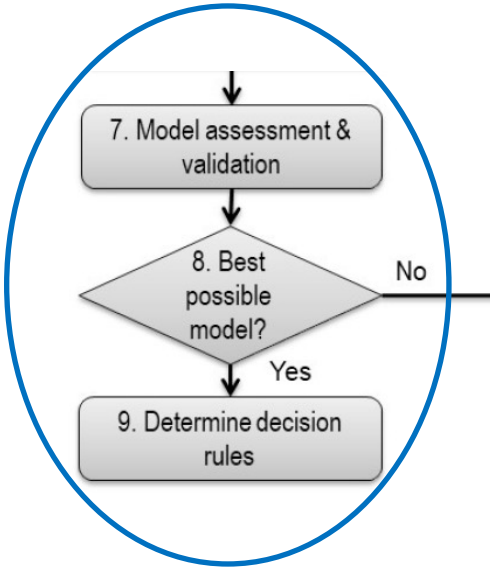


We have a classification task
With a binary outcome
We will apply:

- **Logistic Regression**
- **Decision Tree**
- **Neural Network**

And **compare** the **performance**

Model selection



We are specifically tasked with identifying the customers that will leave

Binary – confusion matrix

Focus – Churn:Yes

Maximizing Recall > Minimizing Precision

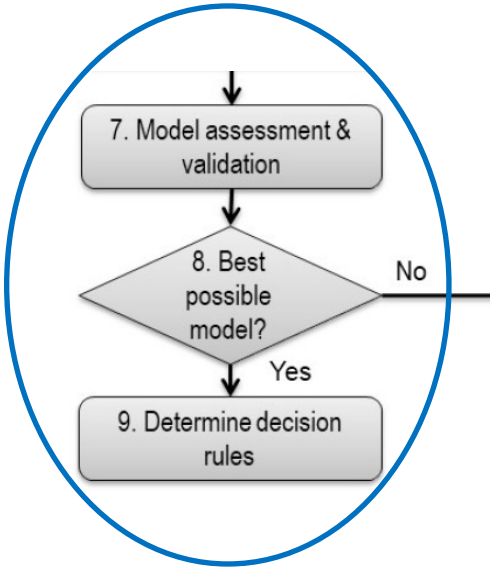
		Predicted Response: $\hat{y} = 1$	Predicted Response: $\hat{y} = 0$	
True Response	$y = 1$	True Positive	False Negative	Recall (Sensitivity) TP/(y=1)
True Response	$y = 0$	False Positive	True Negative	Specificity TN/(y=0)
		Precision TP/(ŷ=1)		Accuracy (TP+TN)/total

Model selection

Logistic Regression

	Pred = Yes	Pred = No
Churn = Yes	322	426
Churn = No	170	1900

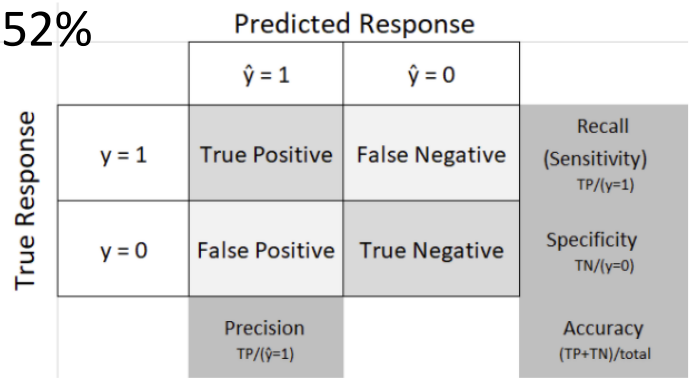
Recall: 322/748 = 43%
Precision: 322/492 = 65%



Decision Tree

	Pred = Yes	Pred = No
Churn = Yes	362	386
Churn = No	332	1738

Recall: 362/748 = 48%
Precision: 362.694 = 52%

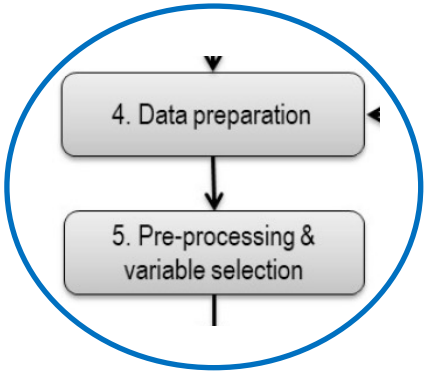


NNet

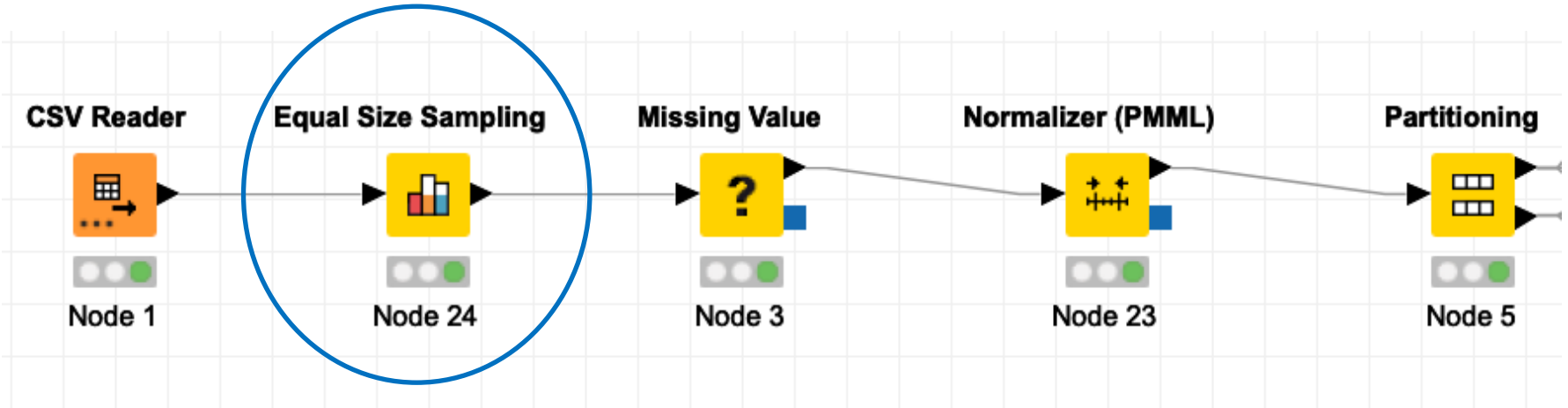
	Pred = Yes	Pred = No
Churn = Yes	318	430
Churn = No	200	1870

Recall: 318/748 = 43%
Precision: 318/518 = 61%

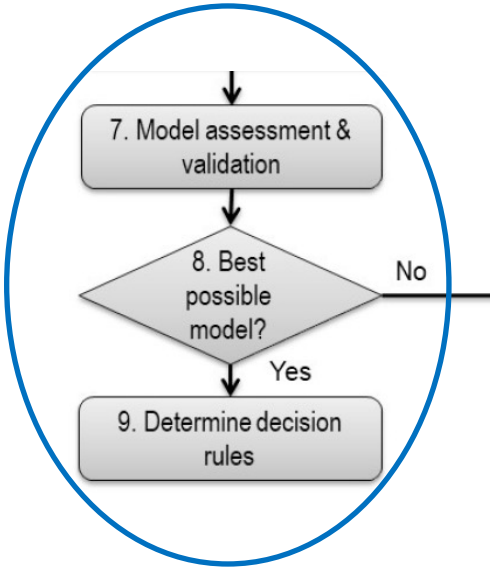
Model Tweaking



The outcome variable is **imbalanced**
Far more Churn:No than Churn:Yes
We can **balance our data** so that there are **equal observations**.



Model selection



Logistic Regression

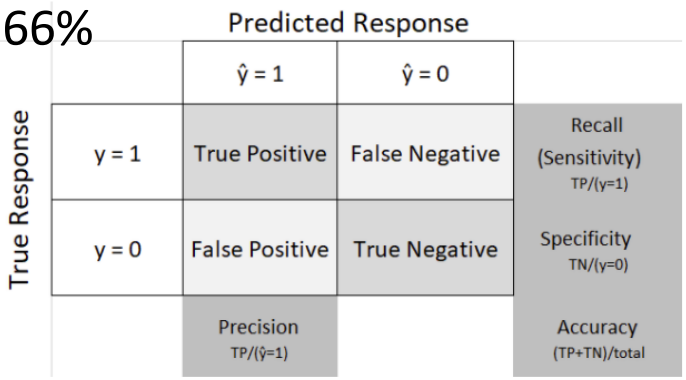
	Pred = Yes	Pred = No
Churn = Yes	524	224
Churn = No	213	535

Recall: 524/748 = 70%
Precision: 524/737 = 71%

Decision Tree

	Pred = Yes	Pred = No
Churn = Yes	534	214
Churn = No	271	475

Recall: 534/748 = 71%
Precision: 534/805 = 66%



NNet

	Pred = Yes	Pred = No
Churn = Yes	536	212
Churn = No	226	522

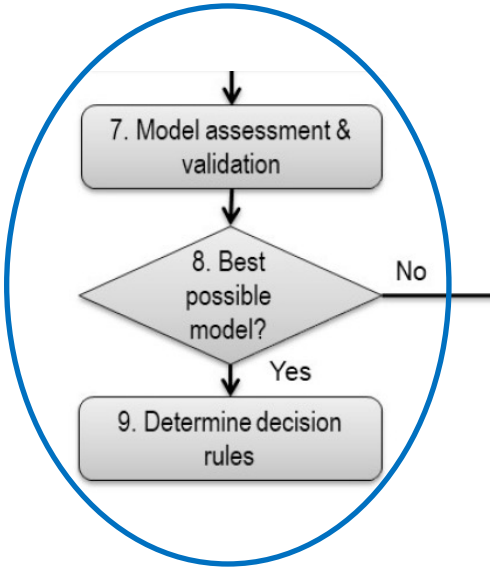
Recall: 536/748 = 72%
Precision: 536/762 = 70%

Model selection

Logistic Regression

	Pred = Yes	Pred = No
Churn = Yes	524	224
Churn = No	213	535

Recall: 524/748 = 70%
Precision: 524/737 = 71%



We choose the Logistic Regression as it is interpretable
Only a minor difference in performance to best model

Coefficients and Statistics - 0

File Edit Hilite Navigation View

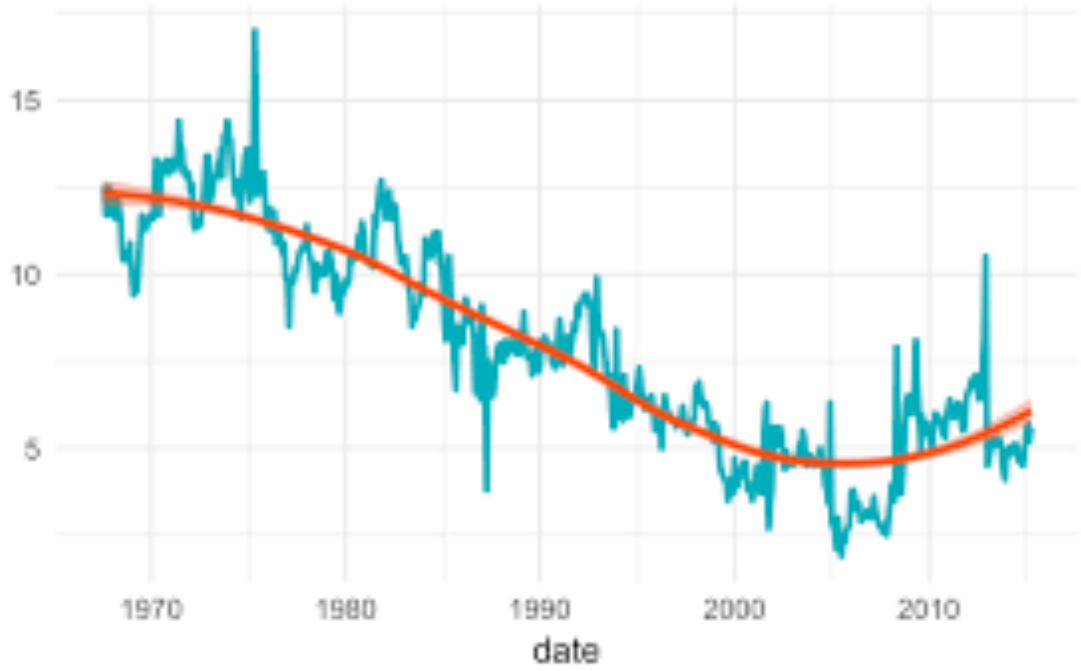
Table "Coefficients and Statistics" – Rows: 3

Row ID	S Logit	S Variable	D Coeff.	D Std. Err.	D z-score	D P>|z|
Row1	No	MonthlyCharges	-1.319	0.066	-19.941	0
Row2	No	TotalCharges	1.344	0.067	19.98	0
Row3	No	Constant	-0.03	0.049	-0.6	0.548

Monthly charges higher, then less likely to switch
Total charges higher, more likely to switch

Data: Time Series (Longitudinal)

	A	B
1	Month	Ridership
2	01/01/1991	1708.917
3	01/02/1991	1620.586
4	01/03/1991	1972.715
5	01/04/1991	1811.665
6	01/05/1991	1974.964
7	01/06/1991	1862.356
8	01/07/1991	1939.86
9	01/08/1991	2013.264
10	01/09/1991	1595.657
11	01/10/1991	1724.924
12	01/11/1991	1675.667
13	01/12/1991	1813.863
14	01/01/1992	1614.827
15	01/02/1992	1557.088
16	01/03/1992	1891.223
17	01/04/1992	1955.981
18	01/05/1992	1884.714
19	01/06/1992	1623.042
20	01/07/1992	1903.309



- Notation:

t = Index of time period

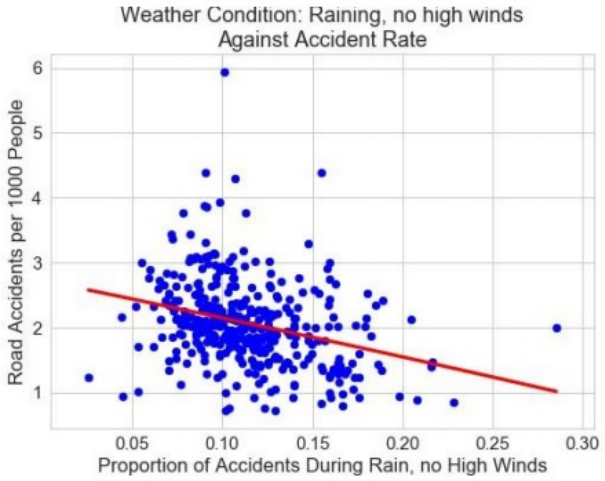
A_t = Actual demand in period t, t = 1, 2, ...

F_t = Forecast for period t, t = 1, 2, ...

w_t = Weight assigned to period t, t = 1, 2, …

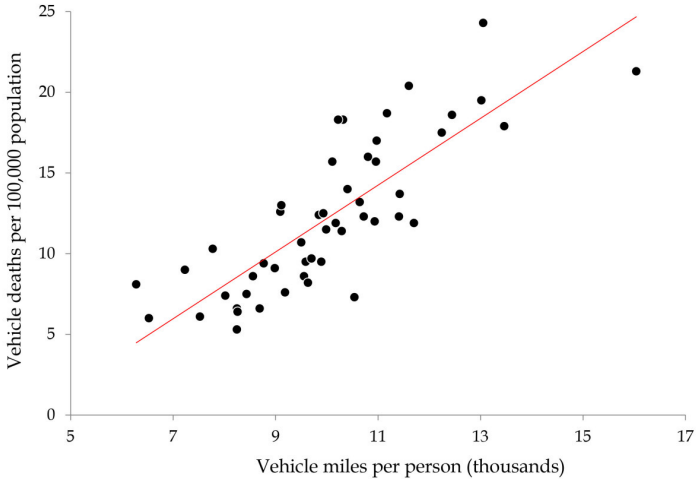
Data: Cross-sectional (Stationary)

	A	B	C	D	E	F	G	H	I	J	K
1	RushHour	WRK_ZONE	WKDY	INT_HWY	LGTCON_day	LEVEL	SPD_LIM	SUR_COND_	TRAF_two_v	WEATHER_a	MAX_SEV
2	1	0	1	1	0	1	70	0	0	1	no-injury
3	1	0	1	0	0	0	55	0	1	0	non-fatal
4	1	0	0	0	0	0	35	0	0	1	no-injury
5	1	0	1	0	0	1	35	0	0	1	no-injury
6	1	0	1	0	0	0	25	0	0	1	non-fatal
7	1	0	1	0	0	0	35	0	0	1	non-fatal
8	1	0	1	0	0	0	60	0	0	0	no-injury
9	1	0	1	0	0	1	45	1	1	0	non-fatal
10	0	0	1	1	0	0	55	1	0	0	no-injury
11	1	0	1	1	0	0	70	1	0	0	non-fatal
12	0	0	1	1	0	0	65	1	0	0	no-injury
13	1	0	1	0	0	0	40	1	0	0	non-fatal
14	1	0	1	0	0	0	45	1	0	0	non-fatal
15	1	0	0	0	0	0	45	1	1	0	non-fatal
16	1	0	1	0	0	0	45	1	1	0	no-injury
17	1	0	1	0	0	0	30	1	1	0	non-fatal
18	1	0	1	0	0	0	55	1	1	0	non-fatal
19	1	0	1	0	0	0	55	1	1	0	no-injury
20	1	0	1	0	0	0	25	1	1	0	no-injury
21	0	0	1	0	0	1	35	0	0	1	no-injury
22	0	0	1	0	0	1	35	0	1	1	no-injury
23	0	0	1	0	0	0	25	0	1	1	no-injury
24	0	0	1	0	0	1	45	0	0	1	no-injury
25	1	0	1	0	0	0	35	0	1	1	no-injury
26	0	0	1	0	0	1	55	0	0	1	non-fatal
27	1	0	1	0	0	1	40	0	0	1	no-injury
28	1	0	0	0	0	1	35	0	1	1	non-fatal
29	0	0	0	0	0	1	25	0	1	0	non-fatal
30	0	0	1	0	0	1	25	0	1	1	no-injury



Does driving cause traffic fatalities?

Miles driven and fatality rate: U.S. states, 2012



single value of dependent variable

slope

single value of independent variable

y-intercept

$$y = mx + b$$

$$Y = \beta_0 + \beta_1 X + \epsilon$$

all observed values for dependent variable

y-intercept a.k.a "bias"

slope a.k.a. "coefficient"

all observed values of independent variable

error*

* additional term

α

Explain vs. Predict

Explanation is the goal of "time series **analysis**"

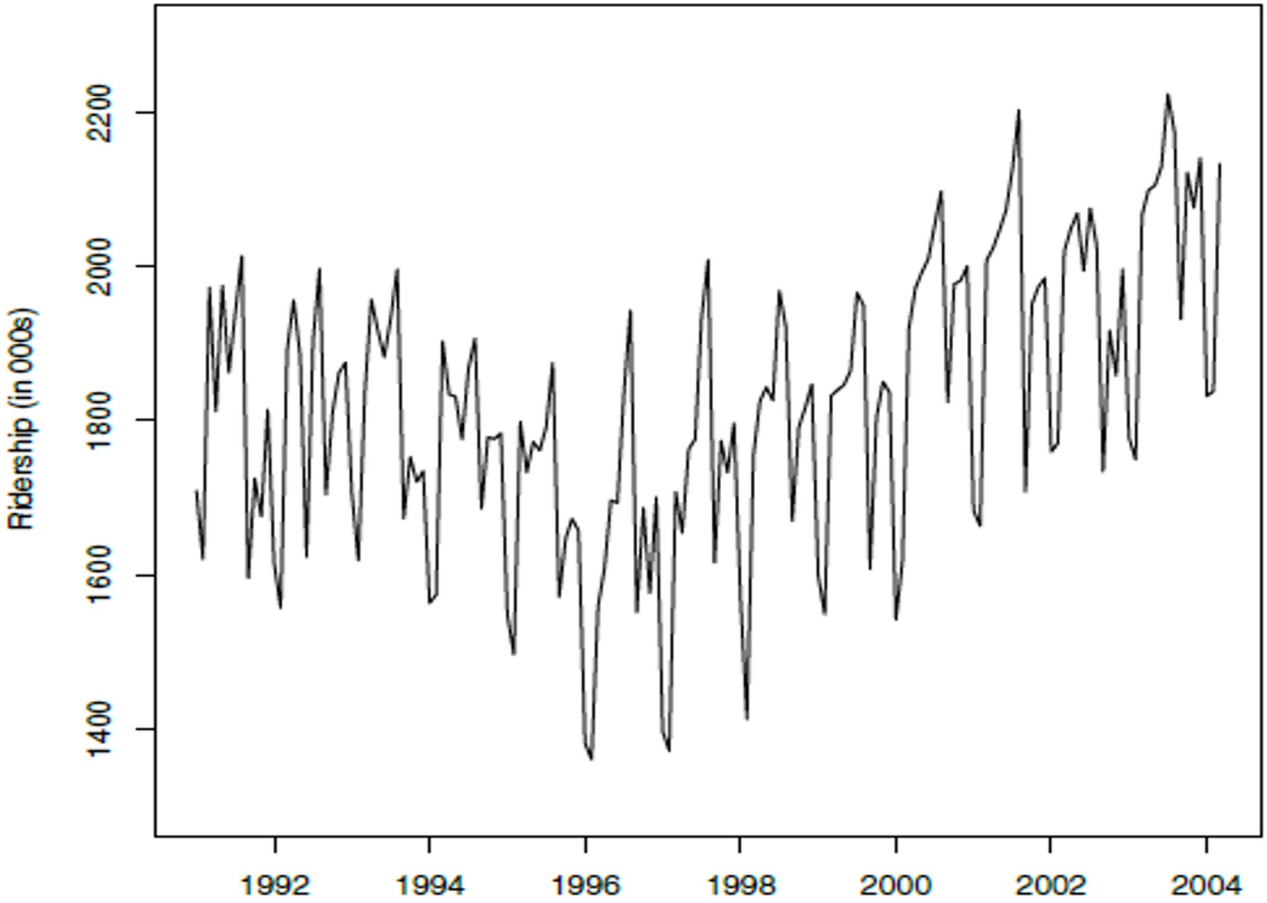
- Models are based on causal argument
- Models are not "black-box"

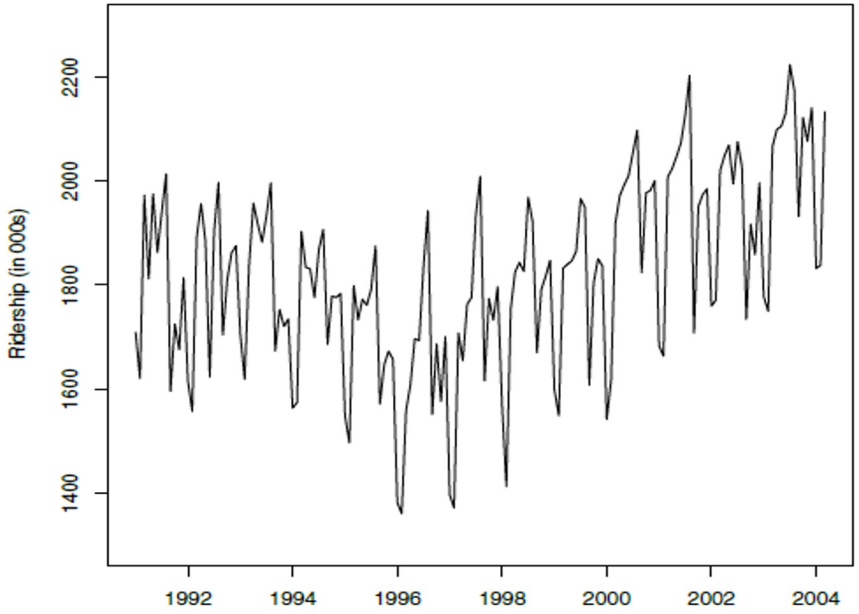
Forecasting (our focus) seeks to **predict** future values

Amtrak Ridership (monthly)

Level - about 1,800,000 passengers per month

Appears to have U-shaped trend





Amtrak Ridership (monthly)

Data Quality:

What error (noise, deviation,etc) is included in our dataset

Temporal Frequency

What is the time scale and frequency?

Series granularity

Coverage of the data

Domain Expertise

Know-how and experience of a highly skilled person in that field or area.

Time series components (next page)

Time Span

What length of time is important? Week, hour, century, millennium?

Outliers (Extreme values)

Are there any data points that seem to lie out of the normal range?

Are these points errors?

Are these not errors, but behaving strangely?

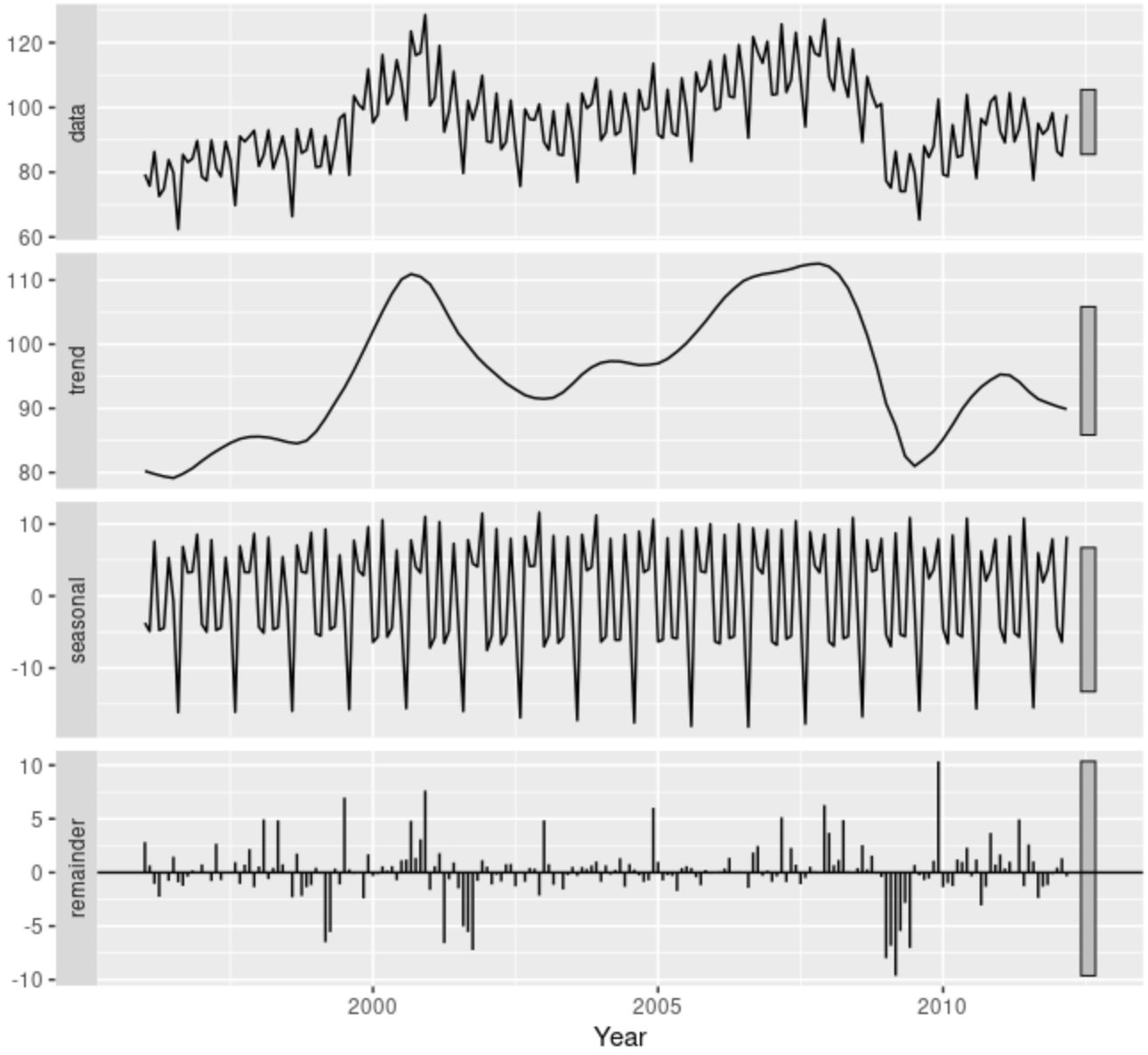
Time Series Components

(Level)

Trend

Seasonality

Noise



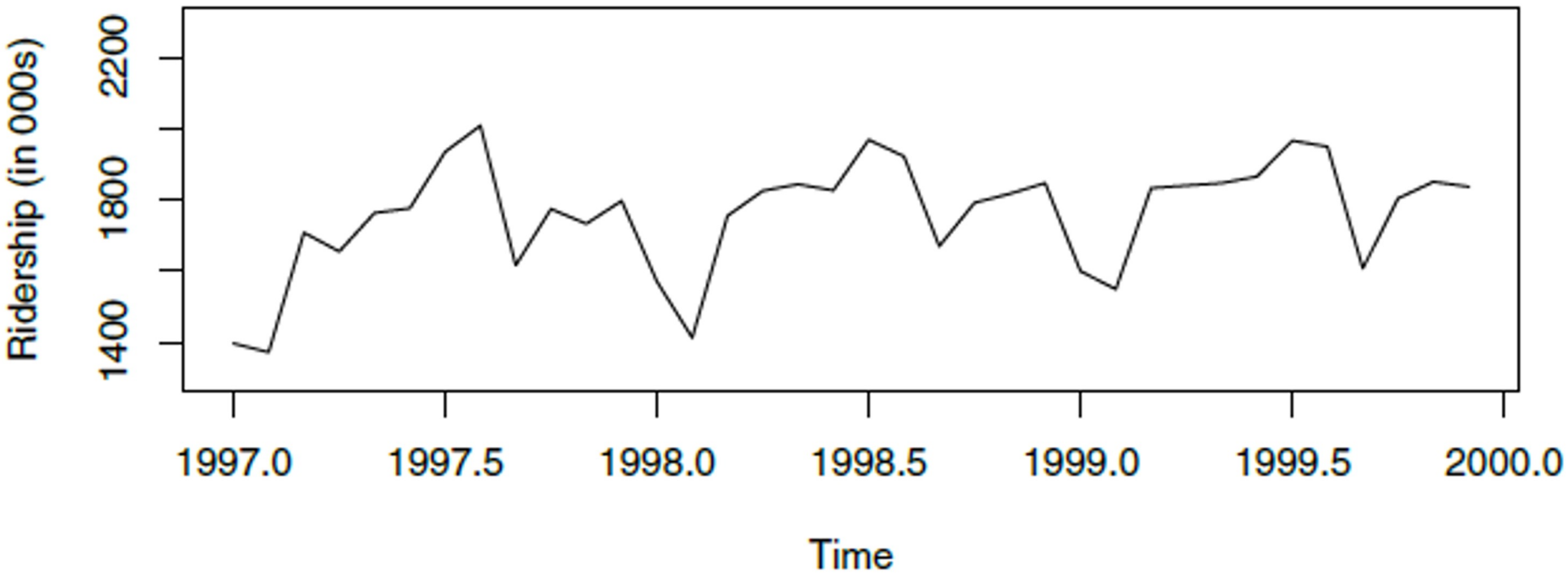
Zoom to 3 years (1997-1999)

Seasonality* appears:

Each year traffic peaks in summer

Noise:

Departure from the general level that is neither trend nor seasonality



Don’t confuse the time series term “season,” which is the period over which a cyclical pattern repeats (e.g. a year), with the standard English seasons of the year (fall, winter, etc.)

Partitioning

Divide data into training portion and validation portion

Test model on the validation portion

Random partitioning would leave holes in the data, which causes problems

Forecasting methods assume regular sequential data

Instead of random selection, divide data into two parts

Train on early data

Validate on later data

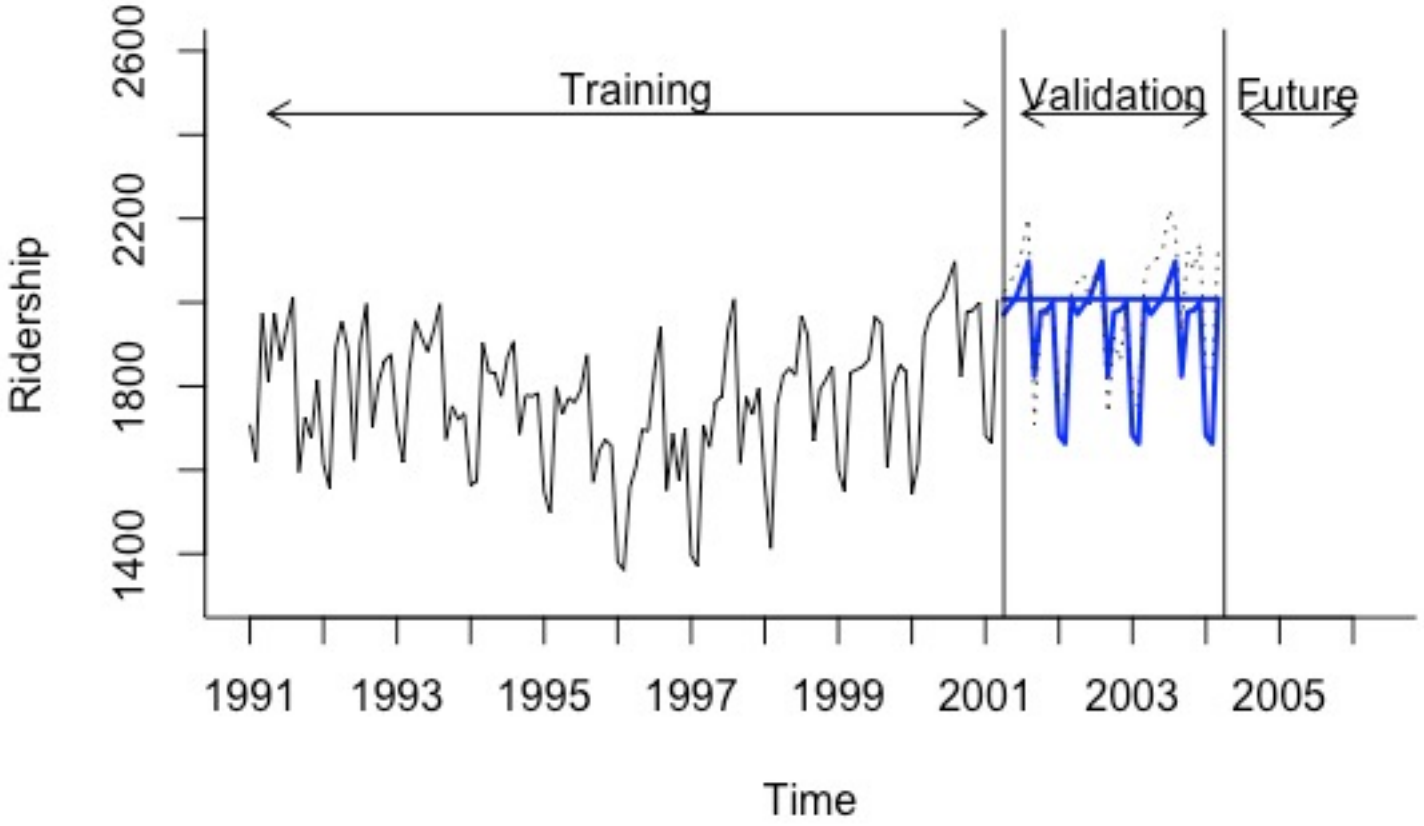
Performance can be assessed against the “naïve benchmark” – *naïve forecast* is simply the most recent value in the time series

Timeseries partitioning is not random!!

Benchmarks

Naïve benchmark is the trend, or average

Seasonal Naïve is the same value for prior season period (m,d,y)

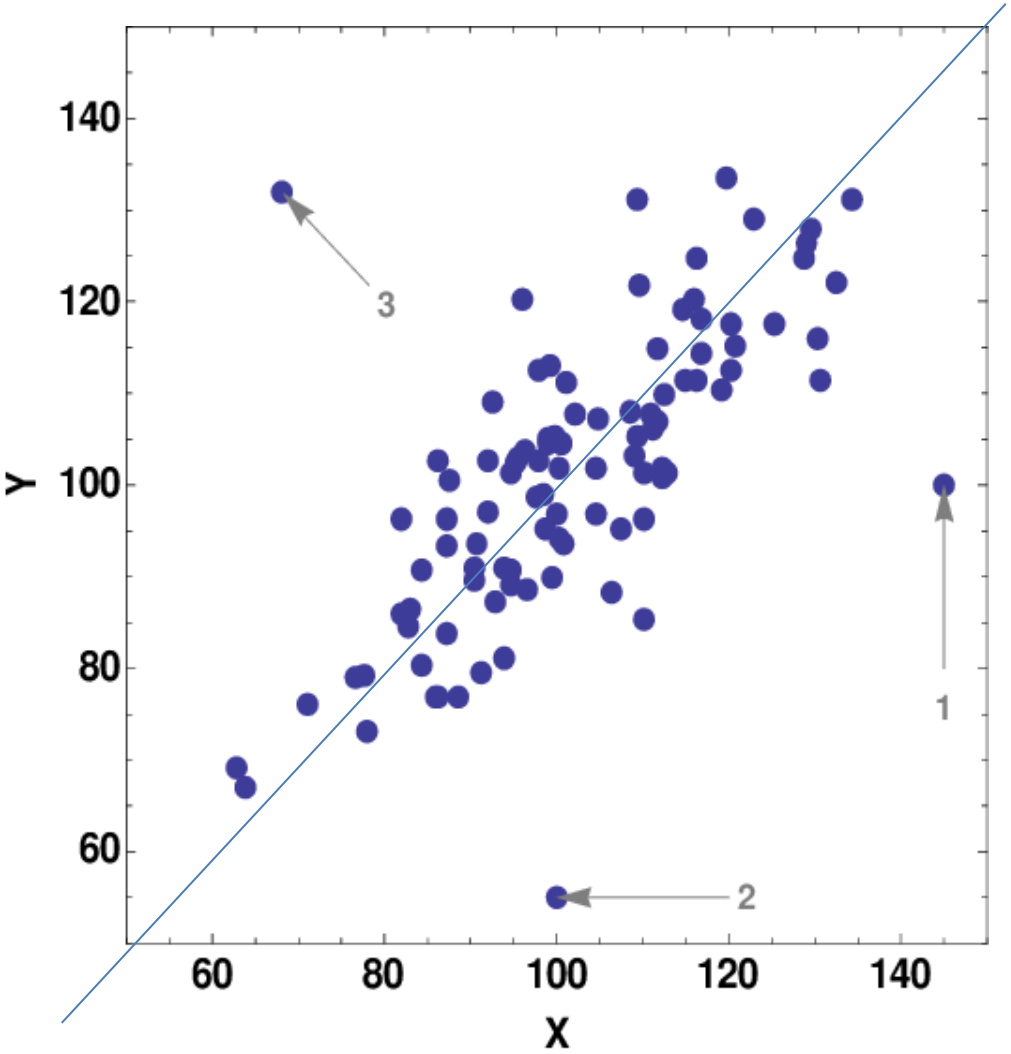


Types of Variables

- Determine the types of pre-processing needed, and algorithms used
- Main distinction: Categorical vs. numeric
- Numeric
  - Continuous
  - Integer
- Categorical
  - Ordered (low, medium, high)
  - Unordered (male, female)

Detecting Outliers

- An outlier is an observation that is "extreme", being distant from the rest of the data (definition of "distant" is deliberately vague)
- Outliers can have disproportionate influence on models (a problem if it is spurious)
- An important step in data pre-processing is detecting outliers
- Once detected, domain knowledge is required to determine if it is an error, or truly extreme.



In some contexts, finding outliers is the purpose of the DM exercise (airport security screening). This is called "anomaly detection".

Handling Missing Data

- Most algorithms will not process records with missing values. Default is to drop those records.
- Solution 1: Omission
  - If a small number of records have missing values, can omit them
  - If many records are missing values on a small set of variables, can drop those variables (or use proxies)
  - If many records have missing values, omission is not practical
- Solution 2: Imputation [see Table 2.7 for R code]
  - Replace missing values with reasonable substitutes
  - Let's you keep the record and use the rest of its (non-missing) information

NB: Determine if "missingness" has value!!

Normalizing (Standardizing) Data

- Used in some techniques when variables with the largest scales would dominate and skew results
- Puts all variables on same scale
- Normalizing function: Subtract mean and divide by standard deviation
- Alternative function: scale to 0-1 by subtracting minimum and dividing by the range
  - Useful when the data contain dummies and numeric

$$Z = \frac{x - \mu}{\sigma}$$

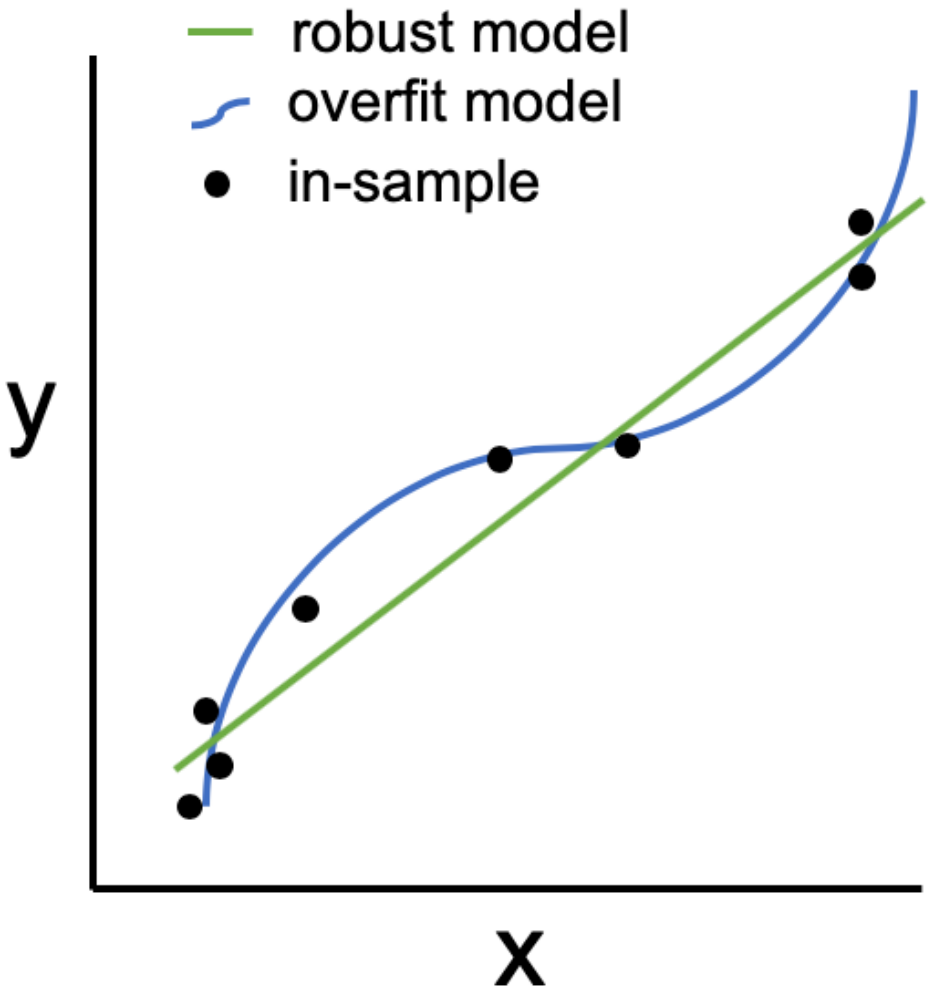
$$z_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

The Problem of Overfitting

- Statistical models can produce highly complex explanations of relationships between variables
- The “fit” may be excellent
- When used with new data, models of great complexity do not do so well.

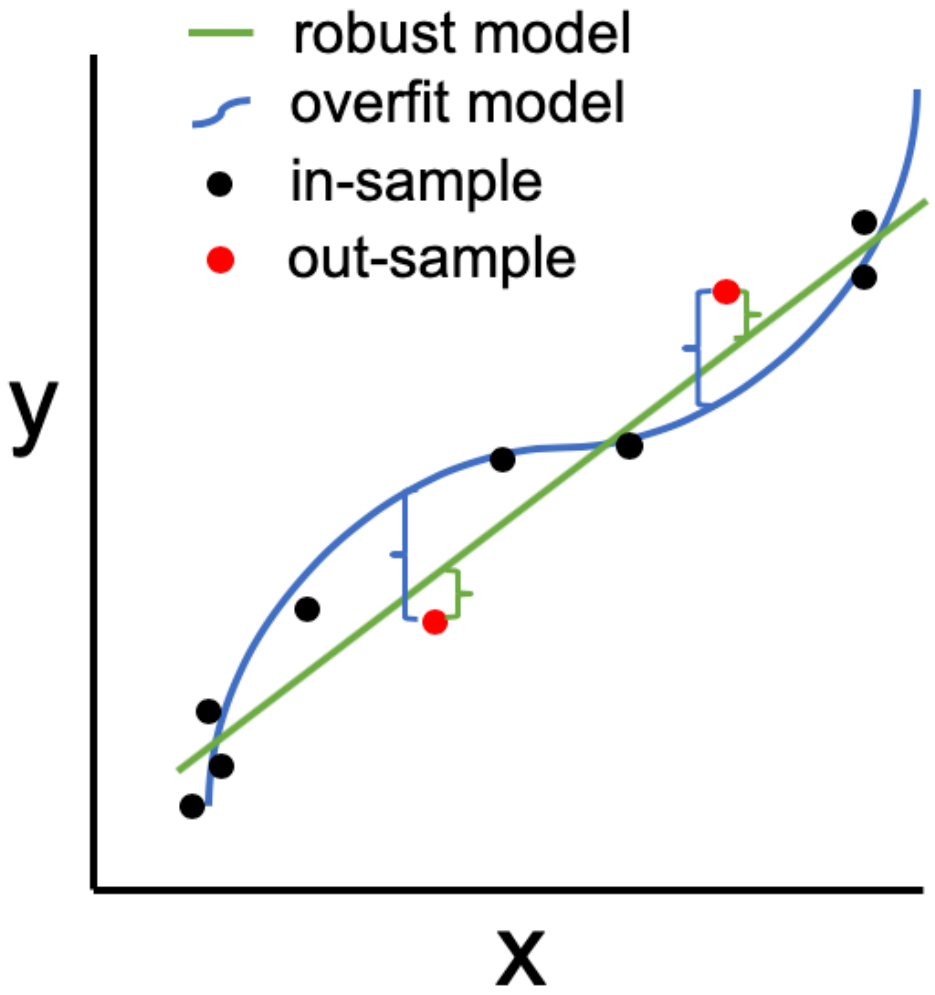
The Problem of Overfitting

100% fit – Excellent!!



The Problem of Overfitting

100% fit – not useful for new data



When used with new data, models of great complexity do not do so well.

Overfitting (cont.)

Causes:

- Too many predictors (too many p, or too few n)
- A model with too many parameters
- Trying many different models

(When p = n, we have perfect fit)

Consequence: Deployed model will not work as well as expected with completely new data.

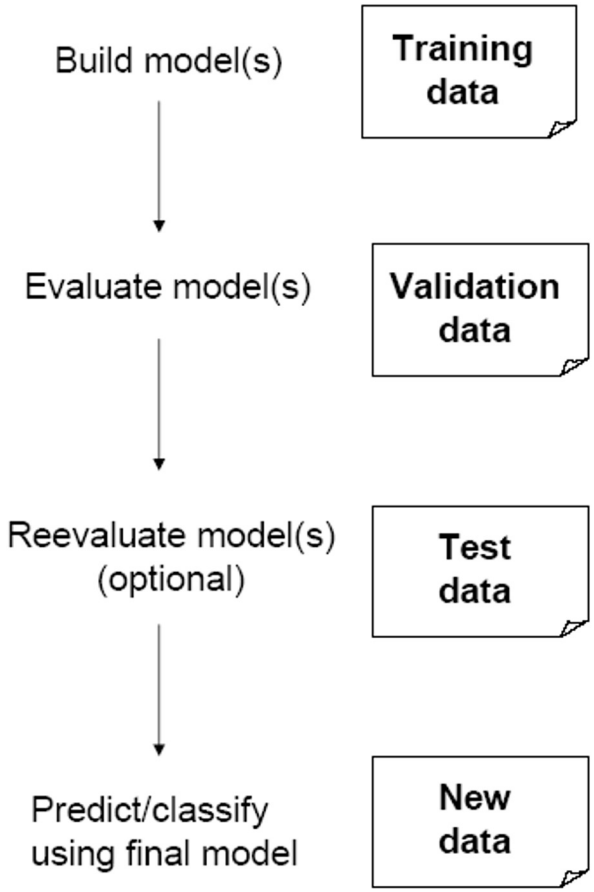
Partitioning the Data

Problem: How well will our model perform with new data?

Solution: Separate data into two parts

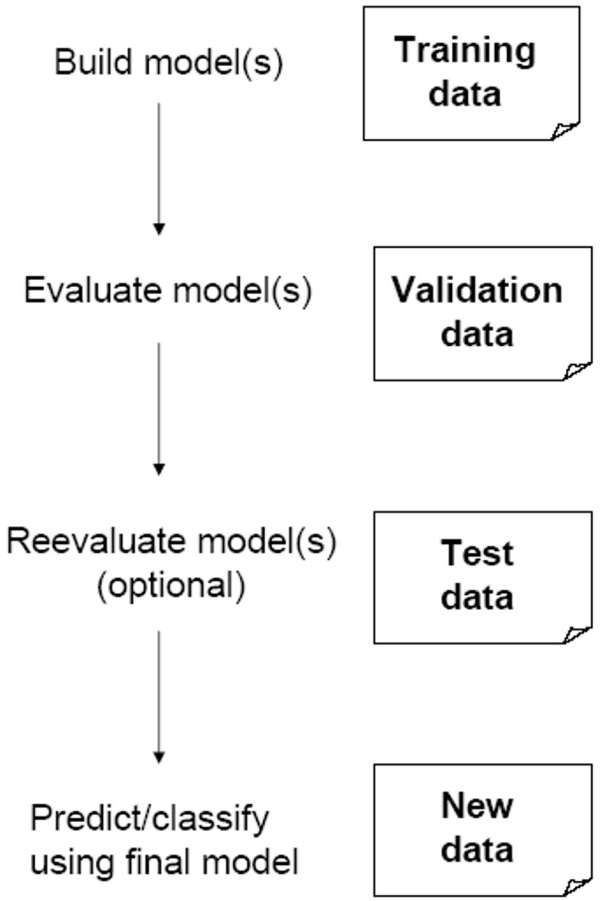
- Training partition to develop the model
- Validation partition to implement the model and evaluate its performance on "new" data

Addresses the issue of overfitting



Test Partition

- When a model is developed on **training data**, it can overfit the training data (hence need to assess on validation)
- Assessing multiple models on same **validation data** can overfit validation data
- Some methods use the validation data to choose a parameter. This too can lead to overfitting the validation data
- Solution: final selected model is applied to a **test partition** to give unbiased estimate of its performance on new data



Error metrics

Error = actual – predicted

ME = Mean error

RMSE = Root-mean-squared error (sd of error)

MSE = mean-squared error (var. of error)

MAE = Mean absolute error

MPE = Pean percentage error

MAPE = Mean absolute percentage error

$$e_i = y_i - \hat{y}_i$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |e_i|$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}}$$

Summary

- Data Mining consists of supervised methods (Classification & Prediction) and unsupervised methods (Association Rules, Data Reduction, Data Exploration & Visualization)
- Before algorithms can be applied, data must be explored and pre-processed
- To evaluate performance and to avoid overfitting, data partitioning is used
- Models are fit to the training partition and assessed on the validation and test partitions
- Data mining methods are usually applied to a sample from a large database, and then the best model is used to score the entire database

Exercises

Impact of September 11 on Air Travel in the United States: The Research and Innovative Technology Administration's Bureau of Transportation Statistics (BTS) conducted a study to evaluate the impact of the September 11, 2001, terrorist attack on U.S. transportation. The study report and the data can be found at `www.bts.gov/publications/estimated_impacts_of_9_11_on_us_travel`. The goal of the study was stated as follows:

> The purpose of this study is to provide a greater understanding of the passenger travel behavior patterns of persons making long distance trips before and after September 11.

(Image by africa / FreeDigitalPhotos.net)

The report analyzes monthly passenger movement data between January 1990 and April 2004. Data on three monthly time series are given in the file *Sept11Travel.xls* for this period: (1) actual airline revenue passenger miles (Air), (2) rail passenger miles (Rail), and (3) vehicle miles traveled (Auto).

In order to assess the impact of September 11, BTS took the following approach: Using data before September 11, it forecasted future data (under the assumption of no terrorist attack). Then, BTS compared the forecasted series with the actual data to assess the impact of the event.

1. Is the goal of this study descriptive or predictive?
2. What is the forecast horizon to consider in this task? Are next-month forecasts sufficient?
3. What level of automation does this forecasting task require? Consider the four questions related to automation.
4. What is the meaning of $t = 1,2,3$ in the Air series? Which time period does $t = 1$ refer to?
5. What are the values for y_1, y_2, and y_3 in the Air series?

16.2 Performance on Training and Validation Data. Two different models were fit to the same time series. The first 100 time periods were used for the training set and the last 12 periods were treated as a hold-out set. Assume that both models make sense practically and fit the data pretty well. Below are the RMSE values for each of the models:

	Training Set	Validation Set
Model A	543	690
Model B	669	675

a. Which model appears more useful for explaining the different components of this time series? Why?

b. Which model appears to be more useful for forecasting purposes? Why?

16.3 contains

HW Suggestions

CREATE well formatted reports

Briefly summarize the question

Format it to distinguish:
question / description / code / output / answers

Show code and relevant text output
use text, not screenshots

Show relevant visualizations
export graphics from Rstudio; not screenshots

CREDIT peers who helped!!

Mention their ID at the top of your assignment!

Peers who help will get extra-credit at end-of-semester