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

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

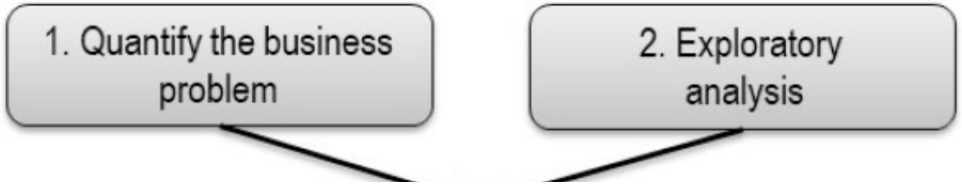
1. Understand & Define the problem
- *Frame the business problem*
- *Prepare for a decision*

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"

Bob's Burgers Example

What **data** do we have?

How can it be **converted**?

What can be **predicted**?

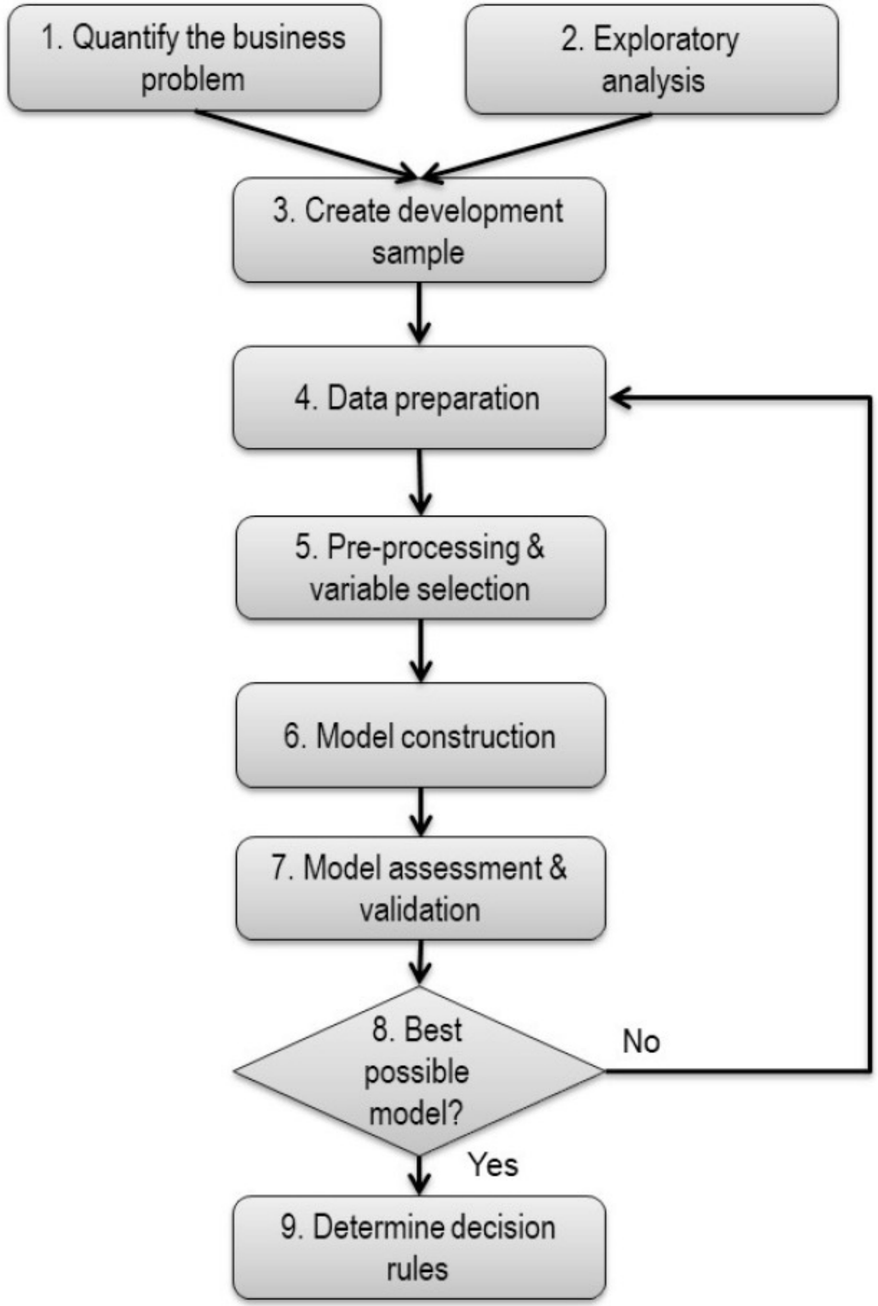
What is the **business value**?

	invoice_uuid	item_category_name	item_name	item_uuid	people	type	dining
1	invoice_uuid	item_category_name	item_name	item_uuid	people	type	dining
2	000026EA-B2E6-41C8-9A99-6C52E825FE4F	送一	拿-經典瑪格	5ebb6673-0086-4b8a-a052-cdb3424ee3c3	1	combo	takeout
3	000026EA-B2E6-41C8-9A99-6C52E825FE4F	送一	拿-堤諾先生	57525b6c-4086-4bc1-afa4-aa3def92eab4	1	combo	takeout
4	000026EA-B2E6-41C8-9A99-6C52E825FE4F	電話	電話	6785d383-5a26-4133-ae11-1e9c1bd4462b	1	item	takeout
5	000026EA-B2E6-41C8-9A99-6C52E825FE4F	送一	拿-經典辣味燻雞	3090bcb2-16e9-4971-8087-b3f92ee6343d	1	combo	takeout
6	000026EA-B2E6-41C8-9A99-6C52E825FE4F	外帶羅馬	羅馬-義式果香燻雞	9ba372f0-038f-4b3b-9afa-833abe6bdfe0	1	combo	takeout

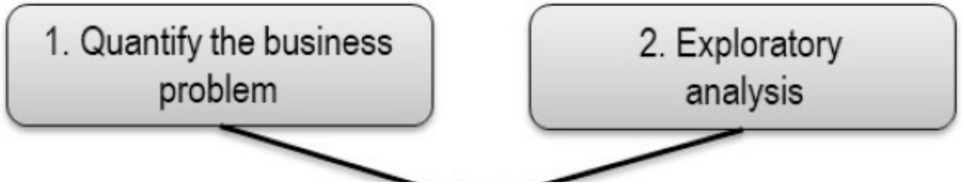
price	timestamp	restaurant_uuid	combo_name	menu_category_name	checkout_type	sales_amount
180	2016/6/15 09:27:30	6d0ebab3-edf8-4e04-a947-1973e76ab11f	歡慶義大利	活動	cash	1140
200	2016/6/15 09:27:30	6d0ebab3-edf8-4e04-a947-1973e76ab11f	歡慶義大利	活動	cash	1140
0	2016/6/15 09:27:30	6d0ebab3-edf8-4e04-a947-1973e76ab11f	NA	NA	cash	1140
220	2016/6/15 09:27:30	6d0ebab3-edf8-4e04-a947-1973e76ab11f	歡慶義大利	活動	cash	1140
360	2016/6/15 09:27:30	6d0ebab3-edf8-4e04-a947-1973e76ab11f	歡慶義大利	活動	cash	1140

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:

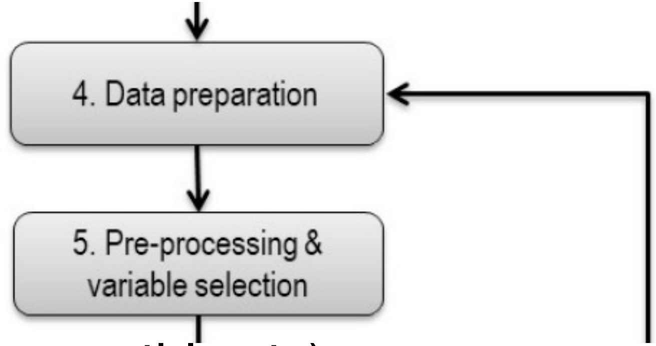
- WHAT makes a best customer (lifetime value, purchases, $ or unit, profit?)
- What criteria make them BEST (over $50,000, tenure?)
- What Databases are available (sales, manufacturing, marketing?
- What data is stored in the database (individual sales, reports, costs?)
- Is data available real-time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

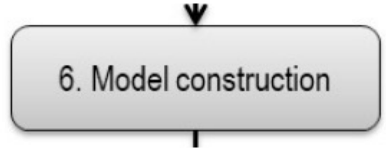
Consolidation
- Ensure consistency in data

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

7. Model assessment & validation

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*

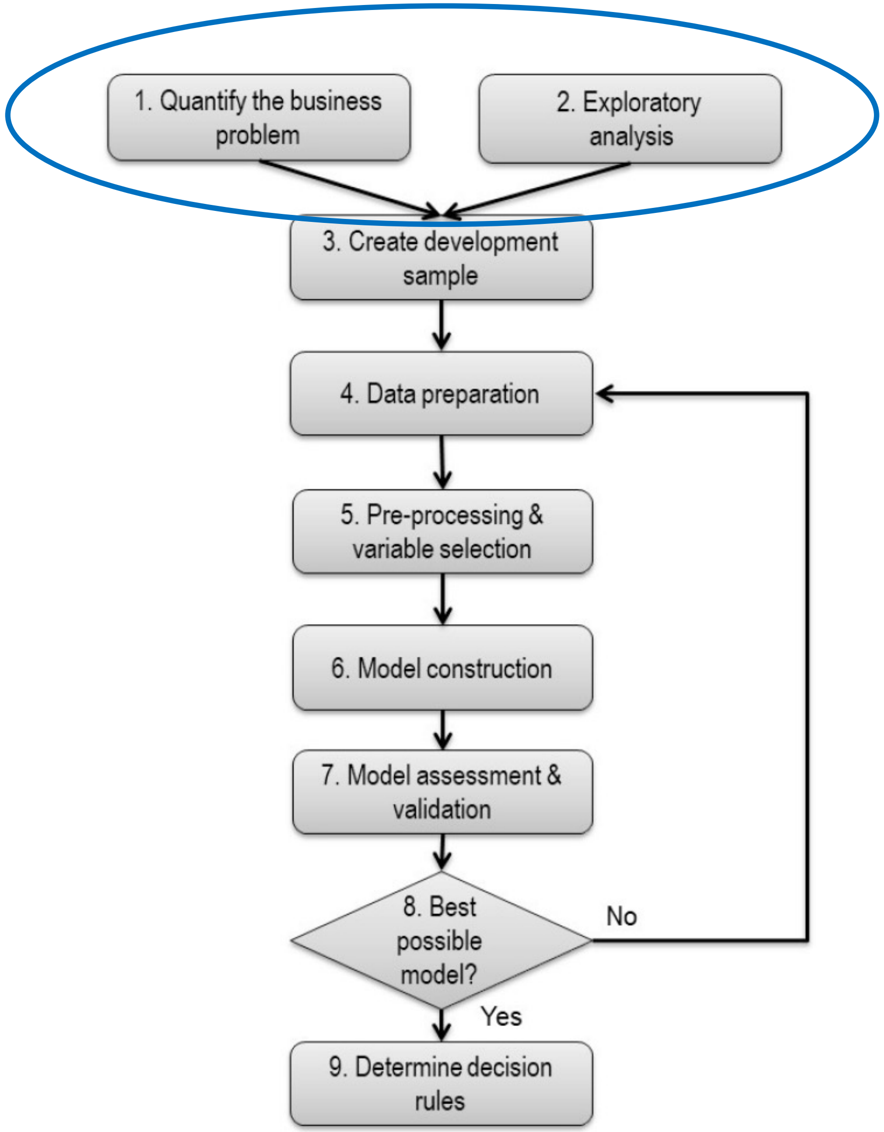
		Predicted Response		
		$\hat{y}$ = 1	$\hat{y}$ = 0	
True Response	y = 1	True Positive	False Negative	Recall (Sensitivity) TP/(y=1)
	y = 0	False Positive	True Negative	Specificity TN/(y=0)
		Precision TP/($\hat{y}$=1)		Accuracy (TP+TN)/total

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.”

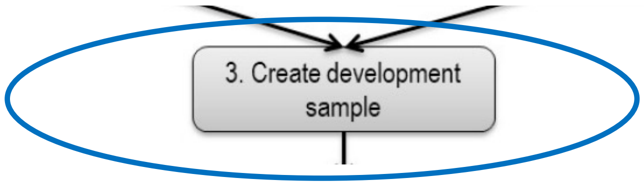
Data Scientist: “The only data we have is:

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



kaggle

Data



	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 (a	89.1	1949.4	No
9	6713-OKOMC	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 (a	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 (a	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**.

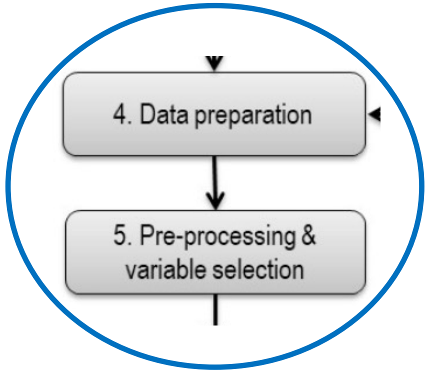
Rows: 7043

Columns: 21

Target: Churn

Open it up in spreadsheet software and have a look.

Data Preparation



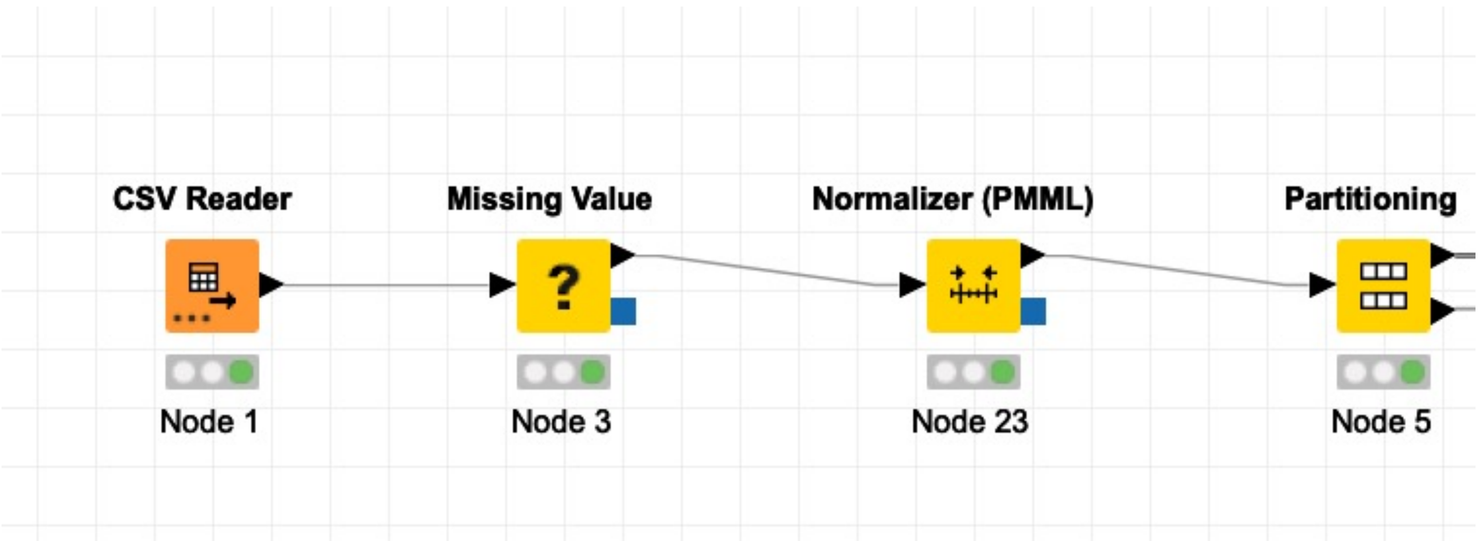
KNIME readily handle categorical vars

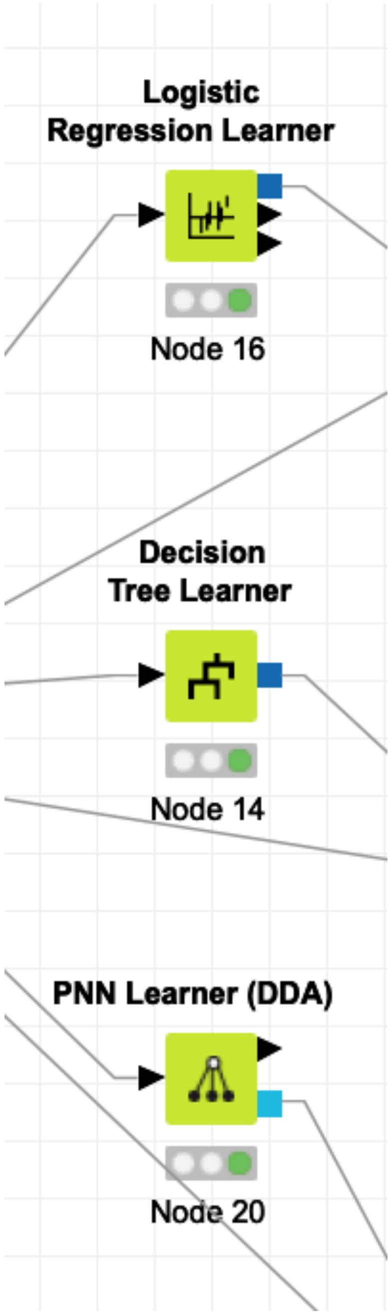
Standardize the **numeric** vars

Remove rows with **missing values**

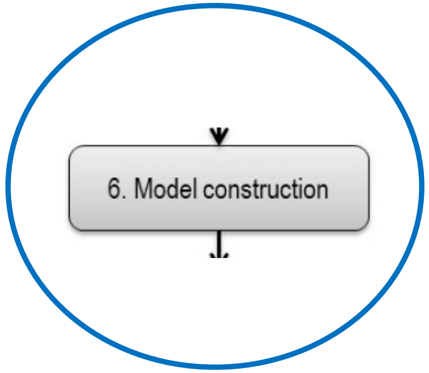
We want to **partition** our data:

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





Model construction

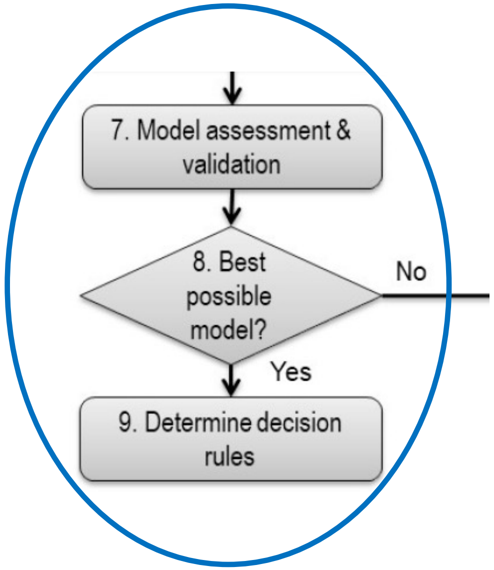


We have a classification task
With a binary outcome
We will apply:

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

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

Model selection



We are specifically tasked with identifying the customers that will leave

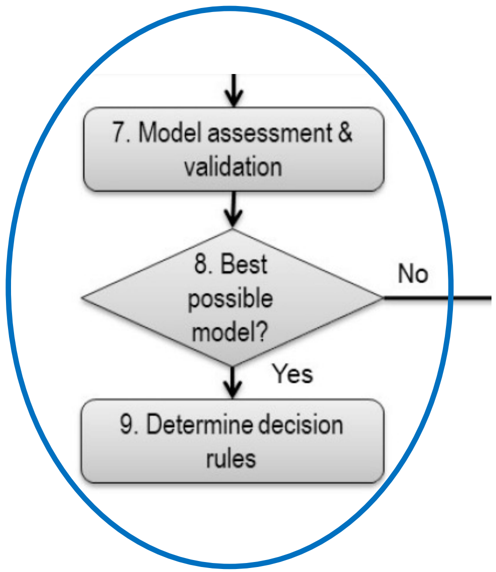
Binary – confusion matrix

Focus – Churn:Yes

Maximizing Recall > Minimizing Precision

		Predicted Response		
		ŷ = 1	ŷ = 0	
True Response	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

Model selection



Logistic Regression

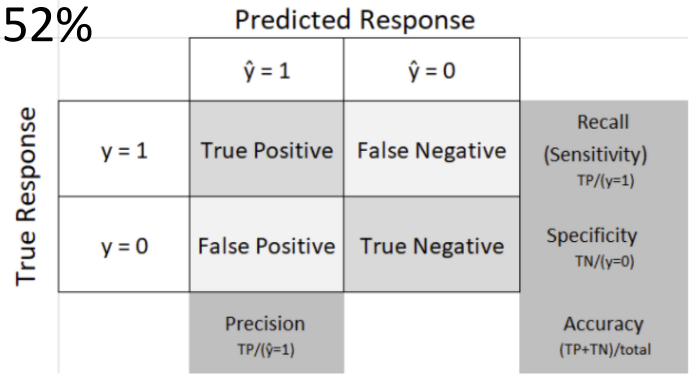
	Pred = Yes	Pred = No
Churn = Yes	322	426
Churn = No	170	1900

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%

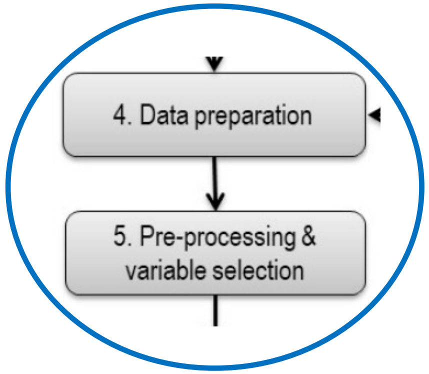


NNet

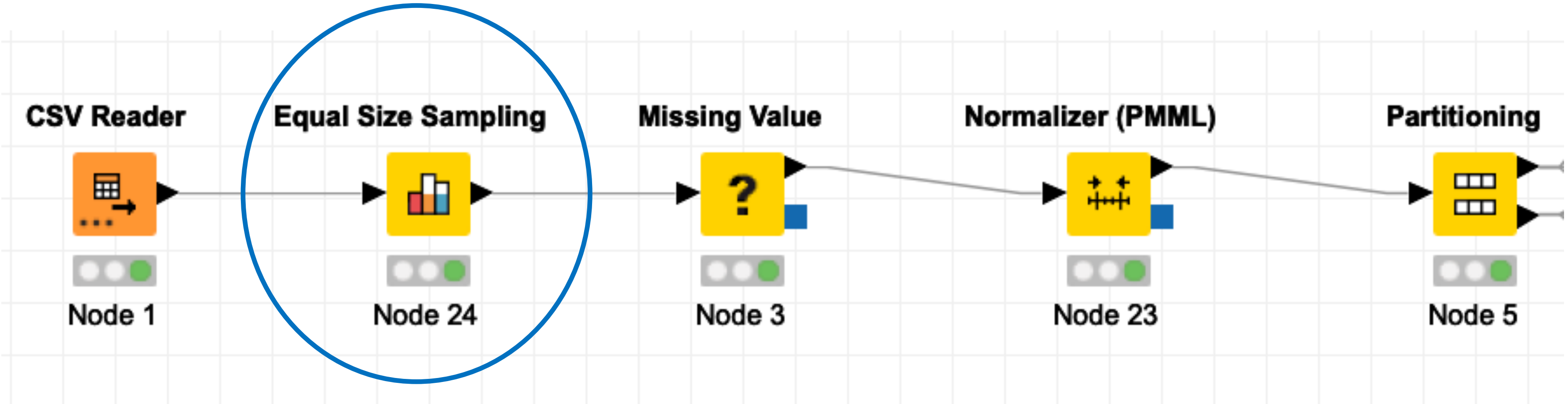
	Pred = Yes	Pred = No
Churn = Yes	318	430
Churn = No	200	1870

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**.

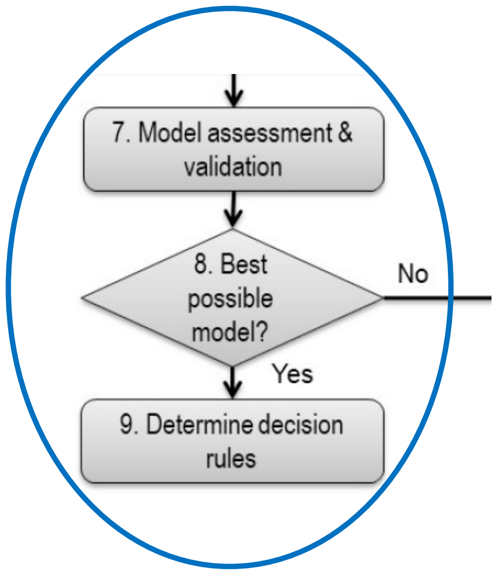


Model selection

Logistic Regression

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

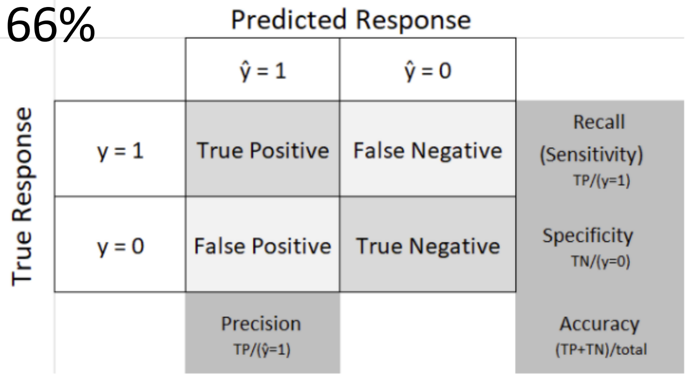
Recall: 524/748 = 70%
Precision: 524/737 = 71%



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

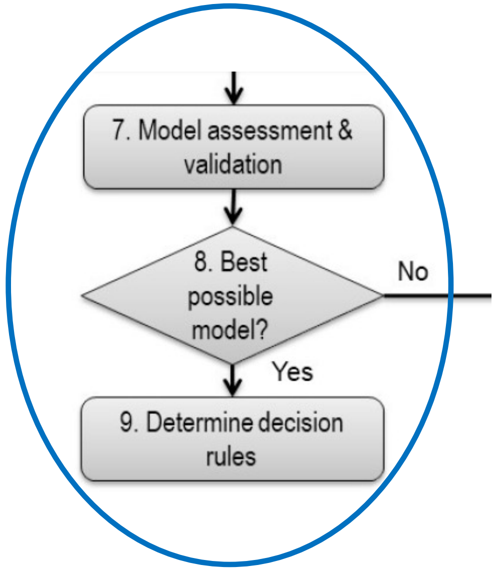
Recall: 536/748 = 72%
Precision: 536/762 = 70%

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

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

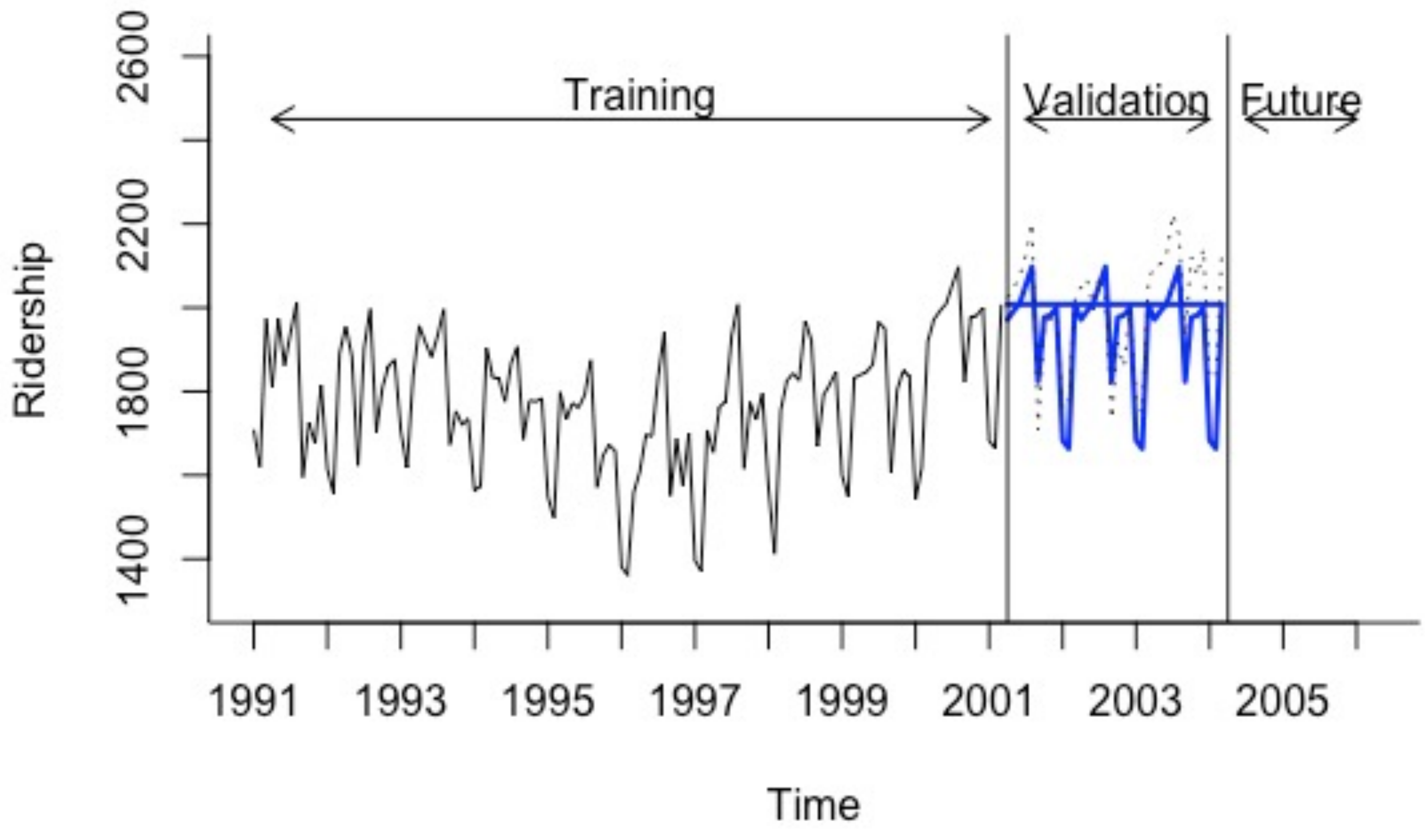
Performance can be assessed against the “naïve benchmark” – *naïve forecast* is simply the most recent value in the time series

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)

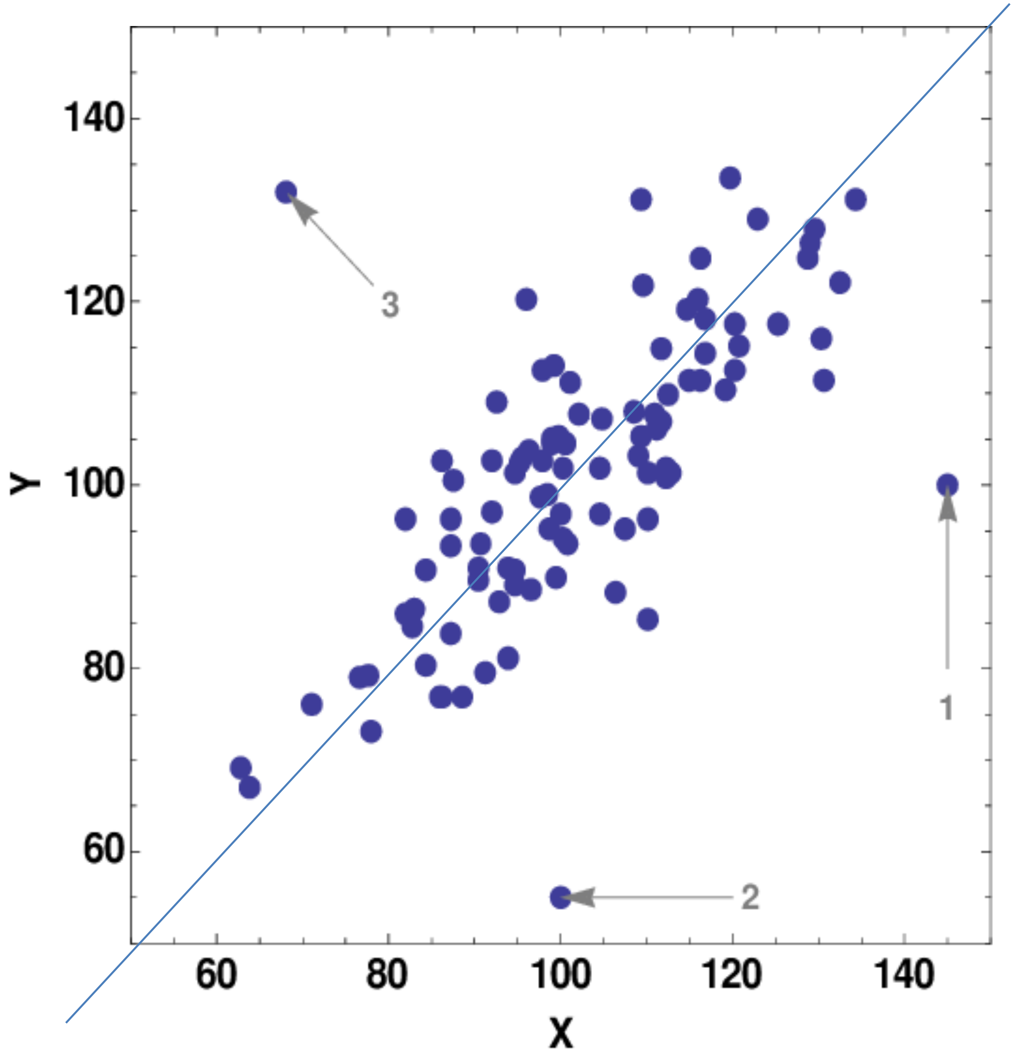


Types of Variables

- 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

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

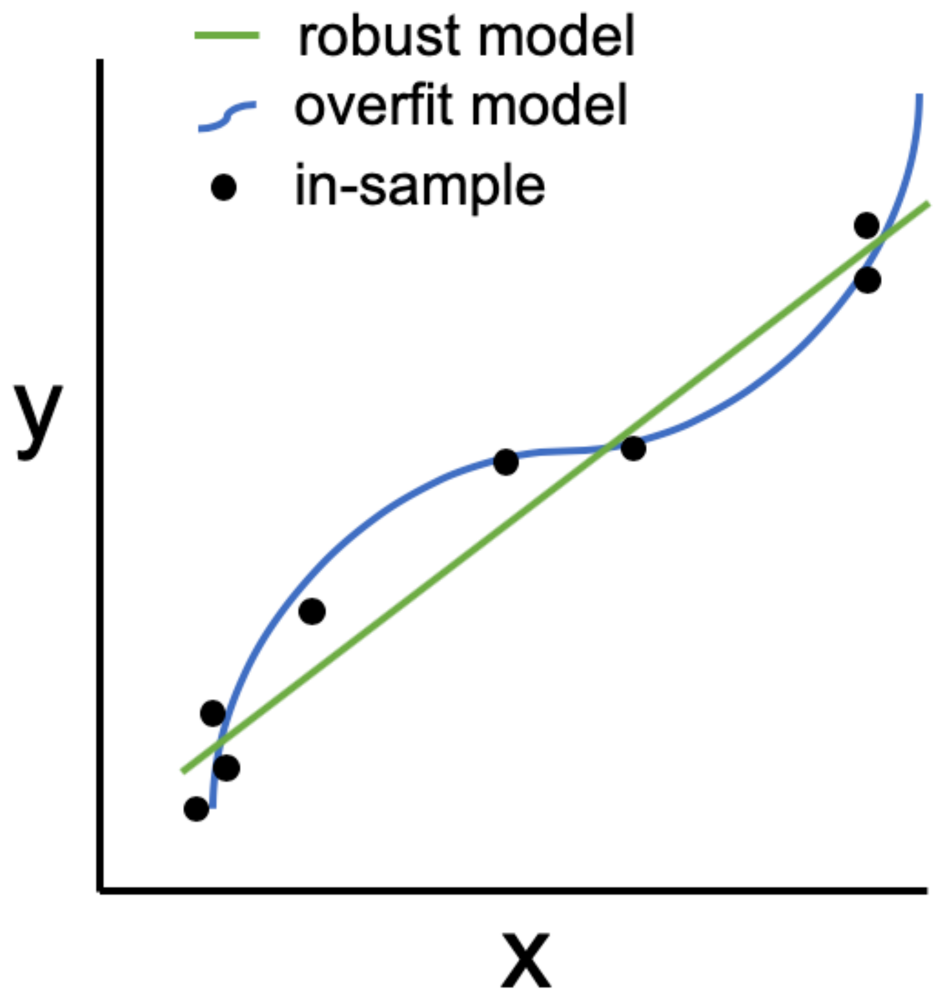
$$z_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

The Problem of Overfitting

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

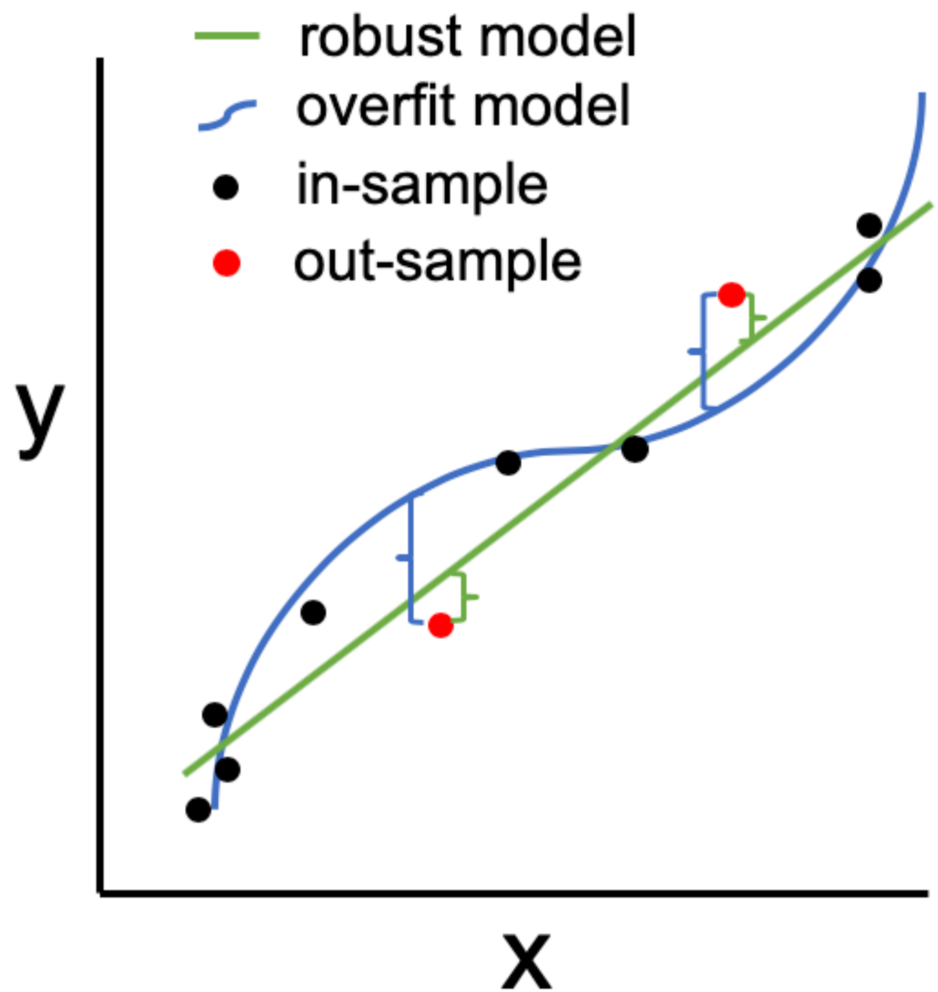
The Problem of Overfitting

100% fit – Excellent!!



The Problem of Overfitting

100% fit – not useful for <u>new</u> data



When used with <u>new</u> 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.

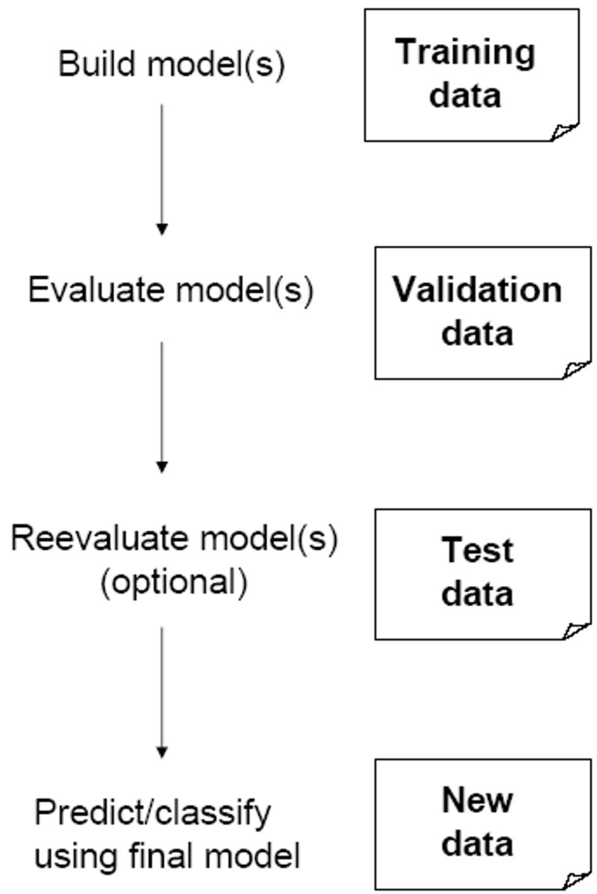
Partitioning the Data

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

Solution: Separate data into two parts

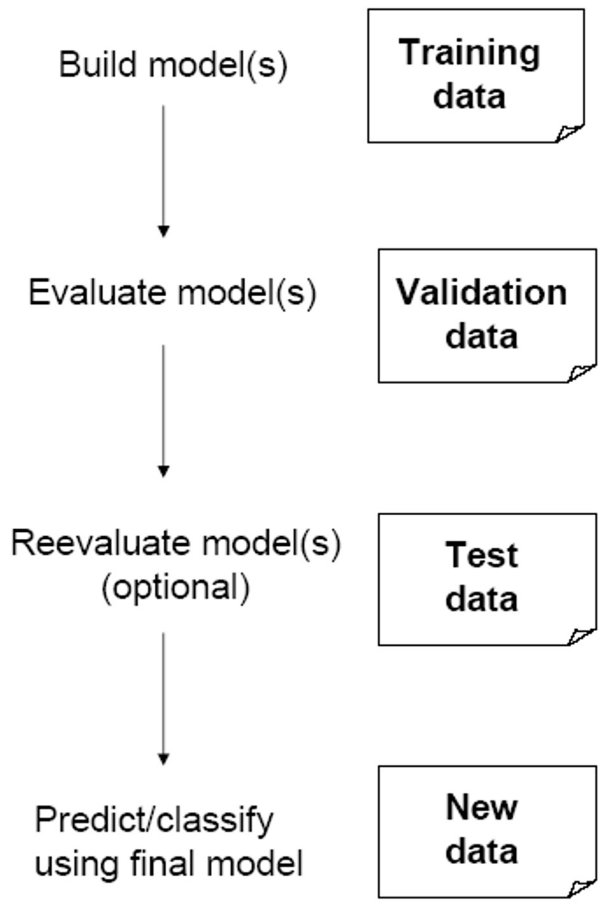
- Training partition to develop the model
- Validation partition to implement the model and evaluate its performance on "new" data

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

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

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

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

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

MAE = Mean absolute error

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

MPE = Pean percentage error

MAPE = Mean absolute percentage error

Summary

- 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

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