



# **Predictive Analytics Certificate Program Seminar**

**DECEMBER 15-16, 2021**

## Predictive Analytics Certificate Program

### Seminar Schedule

*All times Central (CT)*

TIME	SESSION	Speaker
Wednesday December 15, 2021		
9:00–9:15 a.m.	Welcome, Introductions and Review of Seminar Agenda	Stuart Klugman
9:15–10:30 a.m.	General Session: Review of Modules 1-4	Liam McGrath
10:30 -10:45 a.m.	Break	
10:45-11:30 p.m.	General Session: Review of Module 5	Liam McGrath
11:30-12:15 p.m.	Break-Out Groups – Module 5 Practice	
12:15 -1:00 p.m.	Break	
1:00-1:50 p.m.	General Session: Review of Module 6A	Boyang Meng
1:50–3:20 p.m.	Break-Out Groups – Module 6A Practice	
3:20-3:30 p.m.	Break	
3:30–3:45 p.m.	General Session: Communication & Storytelling	Kim Steiner
3:45–4:15 p.m.	Break-out groups	
4:15–4:45 p.m.	Wrap-up and Close  Logistics and questions for the assessment the next day.	Stuart Klugman

TIME	SESSION	Speaker
Thursday December 16, 2021		
9:00–9:30 a.m.	Assessment Introduction and Instructions	Stuart Klugman
9:30am-4:30pm	Assessment	

## Predictive Analytics Certificate Program

### Seminar Speakers and Facilitators



#### **Justin Fountain, ASA, MAAA**

##### **Associate Director – Risk**

Justin Fountain is an Associate Director in the Atlanta office of Willis Towers Watson. Justin joined the firm in 2011. His primary expertise is that of experience analysis and assumption setting using classical and predictive analytic techniques, including data management methods and best practices. Justin has extensive experience building models and end-to-end processes using R, Emblem™, Python, relational databases and various other platforms. Before joining Willis Towers Watson, Justin spent 9 years as a software application developer specializing in web-based financial applications and automation. Justin received a Bachelor of Science in Mathematics with concentration in Actuarial Science from Kennesaw State University. He is an Associate of the Society of Actuaries and a Member of the American Academy of Actuaries.



#### **Stuart Klugman, FSA, CERA, PhD**

##### **Senior Staff Education Fellow**

Stuart is a Senior Staff Education Fellow in the Society of Actuaries Education Department, a position he has held since 2009. For the previous 35 years he was a professor of actuarial at Drake University and The University of Iowa. He is a co-author of Loss Models, used on Exam STAM, and has contributed to two other SOA texts and written numerous papers. He helped develop Course 7 and was an instructor at 18 seminars, was elected to the SOA Board (1997-2000) and as a Vice-President (2001-2003). Stuart co-chaired the committee that re-designed the Education system in 2005 and has played a significant role in the 2018 ASA revisions and the introduction of predictive analytics into all levels of SOA education. He is also a two-time recipient of the SOA Presidential Award.



#### **Liam McGrath, ACAS**

##### **Data Scientist**

Liam is a data scientist/senior consultant based in the Philadelphia office. He has seven years of experience in insurance with a focus on predictive modeling and advanced analytics. Liam has extensive experience building models and tools in R and Python. These include supervised models, such as GLMs and GBMs, as well as unsupervised methods, such as Principal Components Analysis (PCA), clustering, and spatial smoothing. In addition, he has helped develop and maintain a natural language processing pipeline. Liam received a BS degree in

mathematics from Carnegie Mellon University. He is an Associate of the Casualty Actuarial Society (ACAS) and a Chartered Property Casualty Underwriter (CPCU).



## **Boyang Meng, ASA, MAAA**

### **Senior Consultant**

Boyang Meng is a Consultant in the Atlanta office of Willis Towers Watson. He joined the firm in 2013 and plays a key role in the Experience Analysis and Predictive Analytics Initiatives of the Americas Life Practice. He has worked on a variety of predictive analytics projects for clients and has authored articles in industry publications related to analytics. Boyang is proficient in a number of data and analytics platforms including: SQL, R, Radar, and Emblem. Boyang graduated from the University of Florida with BS and BA degrees in Statistics and Economics. He is an Associate of the Society of Actuaries and a Member of the American Academy of Actuaries.



## **Kim Steiner, FSA, MAAA**

### **Senior Director**

Kim Steiner is a Senior Director in the Dallas office of Willis Towers Watson and leader of the Analytics and Experience Analysis Initiatives for the Americas Life Practice. Kim joined the firm in 2008 and is one of the firm's experts in several areas including predictive analytics, mortality, term insurance, reserve financing transactions, principles-based reserving, and COLI / BOLI products. She has worked with clients in using predictive models for use in experience analysis and underwriting and has worked with the SOA on the Predictive Analytics Certificate Program since its inception. Kim earned her BBA in Actuarial Science and Risk Management and Insurance from the University of Wisconsin in 2004.



## **Annie Wang, FSA**

### **Consultant**

Annie is a Consultant with Willis Towers Watson's Insurance Consulting and Technology line of business, working in the Chicago Life Practice. Annie has extensive experience in building experience studies and automating other projects in R. In her spare time, she is studying for an Online Master of Science in Analytics at Georgia Tech. Before joining Willis Towers Watson, Annie worked for Lincoln Benefit Life. She is a Fellow of the Society of Actuaries and received her BA from the University of Chicago.

# SOA Predictive Analytics Certificate Program Seminar

**WILLIS TOWERS WATSON**

December 15-16, 2021



# Welcome

- Introduction
- Goals of the seminar
- Agenda

# Agenda

Time	Session
9:00 – 9:15 a.m.	Welcome and Review of Seminar Agenda
9:15 – 10:30 a.m.	Re-cap of Modules 1-4 and conducting a predictive analytics project
10:30 -10:45 a.m.	Break
10:45-11:30 a.m.	Re-cap of module 5
11:30-12:15 p.m.	Break-out groups – discuss practice from module 5
12:15-1:00 p.m.	Lunch Break



# Agenda

Time	Session
1:00-1:50 p.m.	Re-cap of module 6 and choosing the right solution for your problem. Interactions between the modeling stage and earlier stages
1:50-3:20 p.m.	Break out groups – discuss practice from module 6A
3:20-3:30 p.m.	Break
3:30-3:45 p.m.	Importance of communication and storytelling. Guidance on reporting results from predictive analytics projects
3:45-4:15 p.m.	Break-out groups – work through an example report, discuss strengths, weaknesses and audience.
4:15-4:45 p.m.	Wrap-up and Close Logistics and questions for the assessment the next day.



SOCIETY OF  
ACTUARIES®

# Predictive Analytics Certificate Program

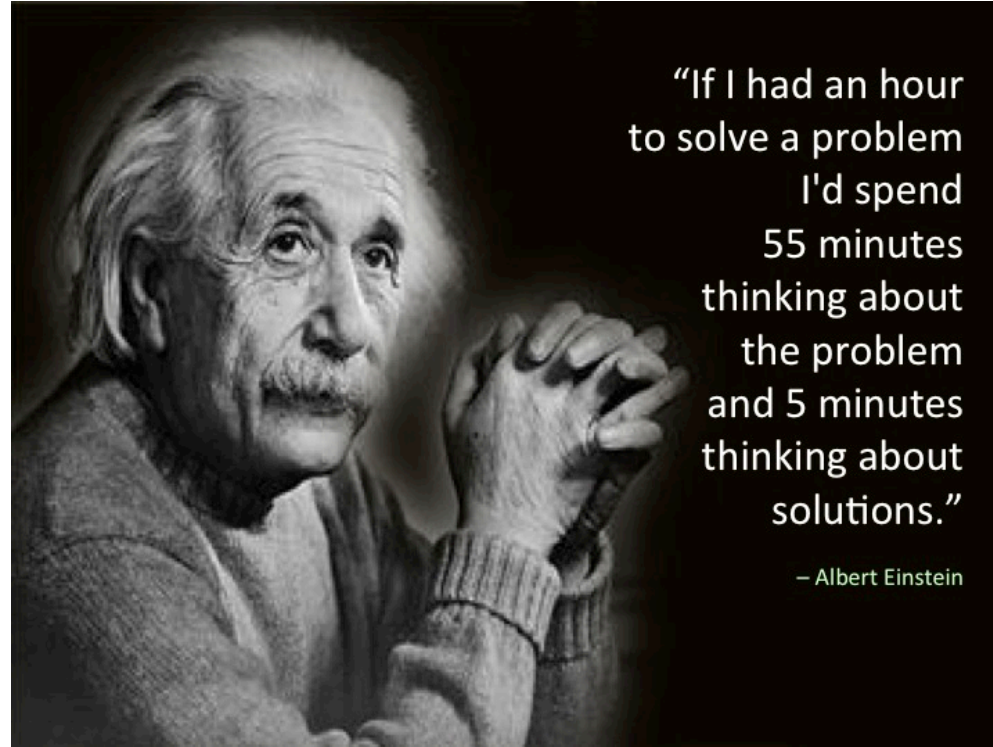
---

## Review: Modules 1-4



# Understanding...

- Your problem
- How to go about solving the problem
- Who needs to be involved
- Limitations, constraints and other considerations for implementation
- Your data



# A predictive analytics project...

Modeling

# A predictive analytics project...



# A predictive analytics project...



Understanding the problem

# Module 1 – What is predictive analytics?

- Predictive analytics is...

Traditional Analysis	Predictive Analytics
<ul style="list-style-type: none"><li>• Aggregate</li><li>• One-way analysis</li><li>• Pre-defined process</li></ul>	<ul style="list-style-type: none"><li>• Individualized</li><li>• Multiple variable analysis</li><li>• Hypothesis-driven</li></ul>

- Predictive analytics tools

# Module 2 – Effective problem definition

- Problem definition
  - Measurable
  - Actionable
  - Ability to assess the outcome
- Considerations for successful project
  - Alignment with business strategy
  - Risk
  - Technology
  - Project management



# Module 3 – Data Preparation

- Understanding your data structure
  - Structured vs. unstructured
  - Categorical vs. numeric vs. Boolean vs. date/time/geospatial
  - Granularity
- Data preparation
  - Missing values
  - Merging data from multiple sources
- Basic data transformations
  - Binarize
  - Numeric → categorical (factor)

# Module 4 – Data Exploration

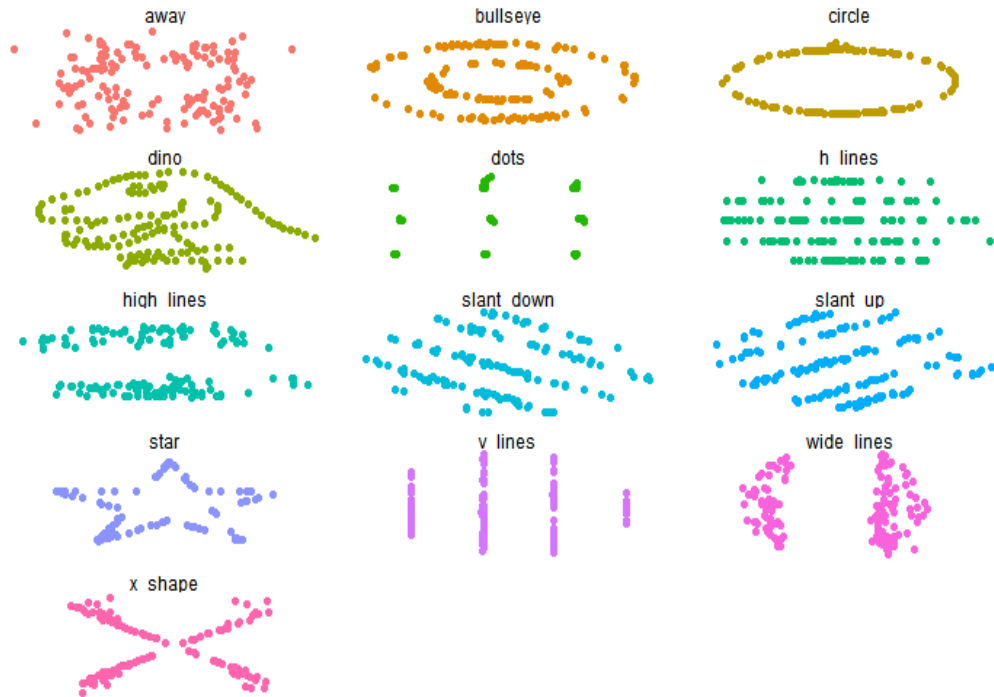
- Purpose
  - Check against common sense, knowledge, intuition
  - High level statistical features and relationship between variables
- Data to explore
  - Full data vs. train data vs. randomly selected data

# Module 4 – Data Exploration

- Understand distribution
  - One-way: Basic statistics, dispersion, outlier
  - Two-way: Relationship with target variable
- Numeric
  - Histogram, boxplot
  - Min, max, mean, median, variance, percentiles, outliers
  - Correlation, mutual information (after discretization)
- Categorical
  - Frequency table, bar chart
  - Number of levels, balance of levels, outliers
  - Mutual information
- Advanced techniques
  - PCA and clustering

# Module 4 – Data Exploration

- Correlation vs. Mutual information



Correlation: about -7% for all cases

Mutual information

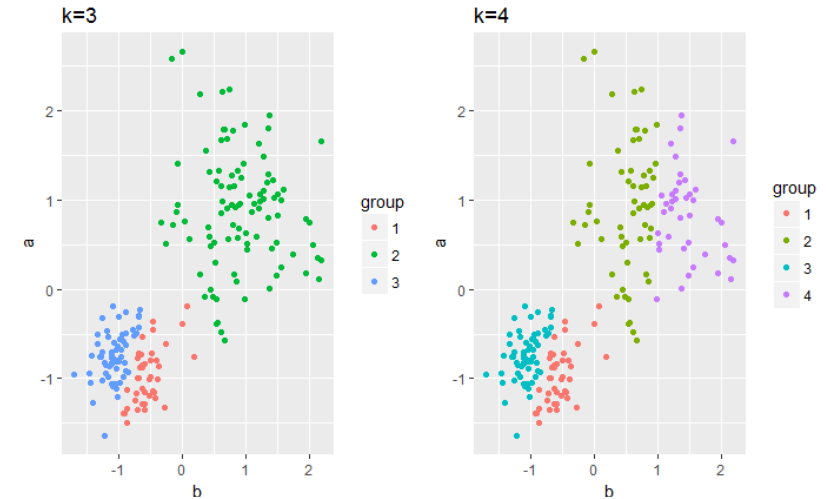
Away	Bulls Eye	Circle
0.13	0.24	0.51
Dino	Dots	H lines
0.13	0.27	0.11
High lines	Slant down	Slant up
0.10	0.21	0.26
Star	V lines	Wide lines
0.55	0.11	0.17
X shape		
0.58		

Source: Datasaurus

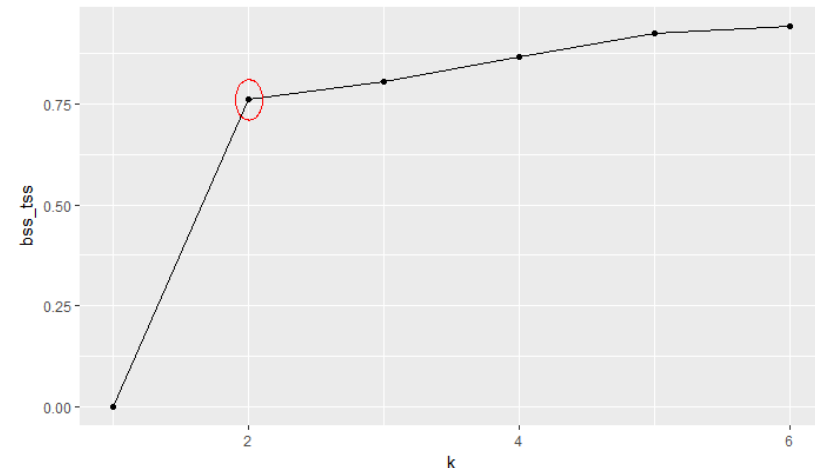
# Module 4 – Clustering and PCA

- Unsupervised
- Can be used to identify structure within data
- Principal component analysis (PCA)
  - Linear transformation of numeric variables and reorder from highest variance
  - Useful for data with multicollinearity
- Clustering
  - Group records based on definition of “similarity”
- Further useful for feature generation

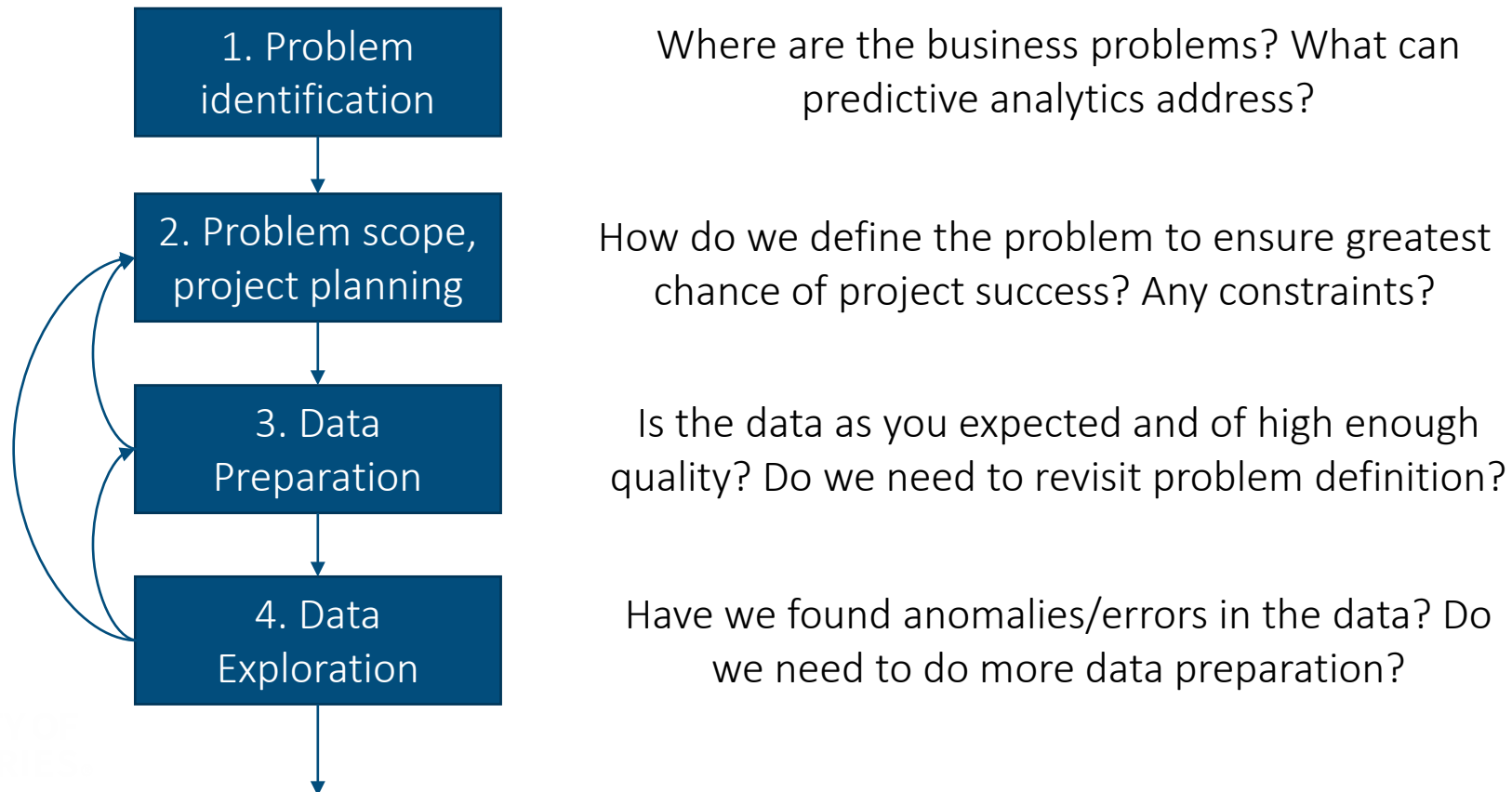
Clustering example



Elbow plot



# Summary - an iterative process





SOCIETY OF  
ACTUARIES®

# Predictive Analytics Certificate Program

---

## Break





SOCIETY OF  
ACTUARIES®

# Predictive Analytics Certificate Program

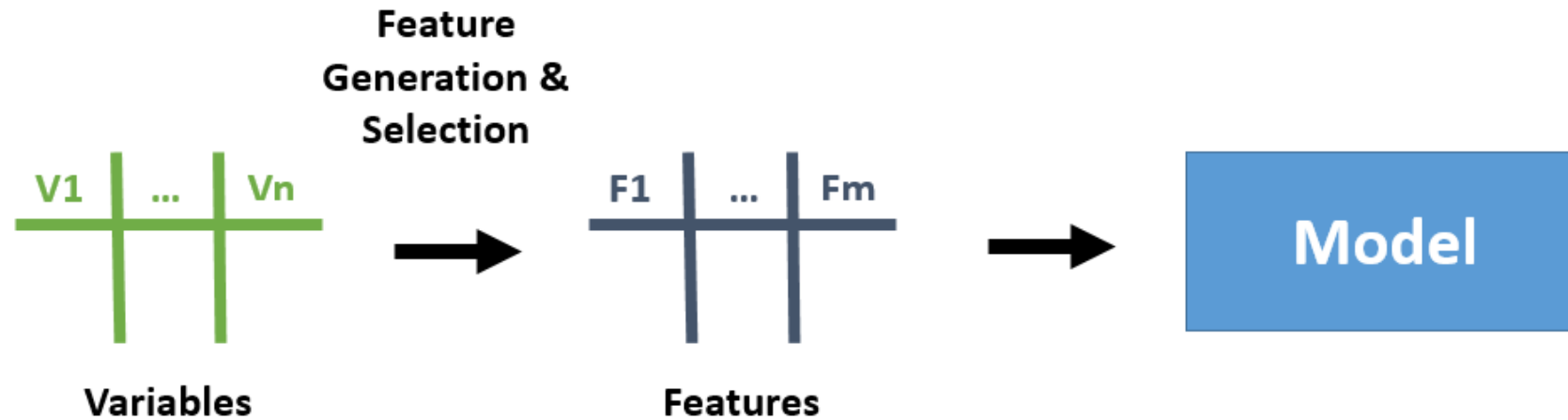
---

## Review: Module 5





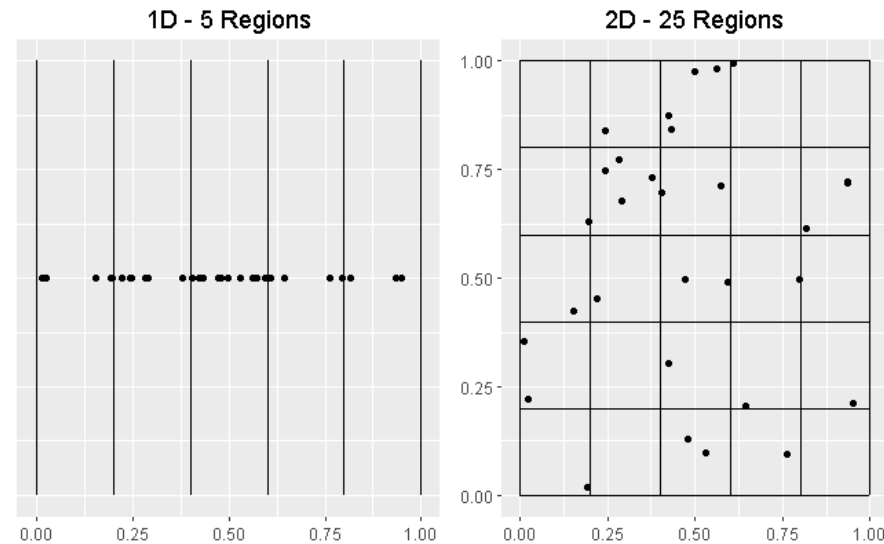
# Module 5 – Feature Generation & Selection



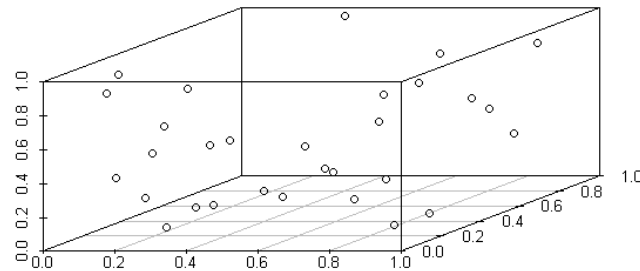
# Module 5 – Feature Generation & Selection



Occam's Razor

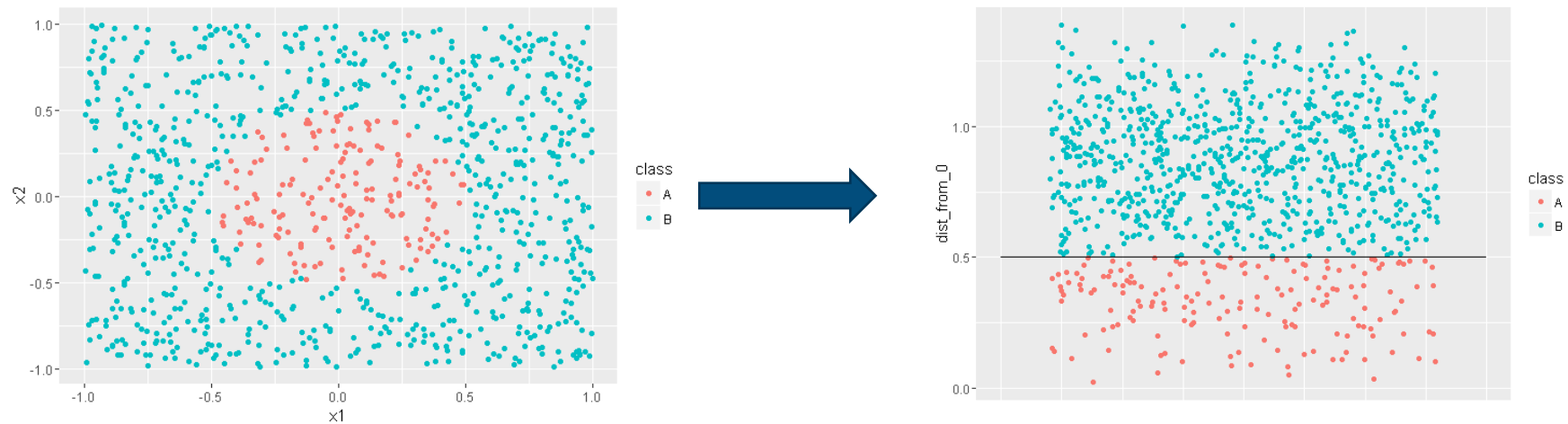


3D - 125 Regions



Curse of Dimensionality

# Module 5 – Feature Generation & Selection



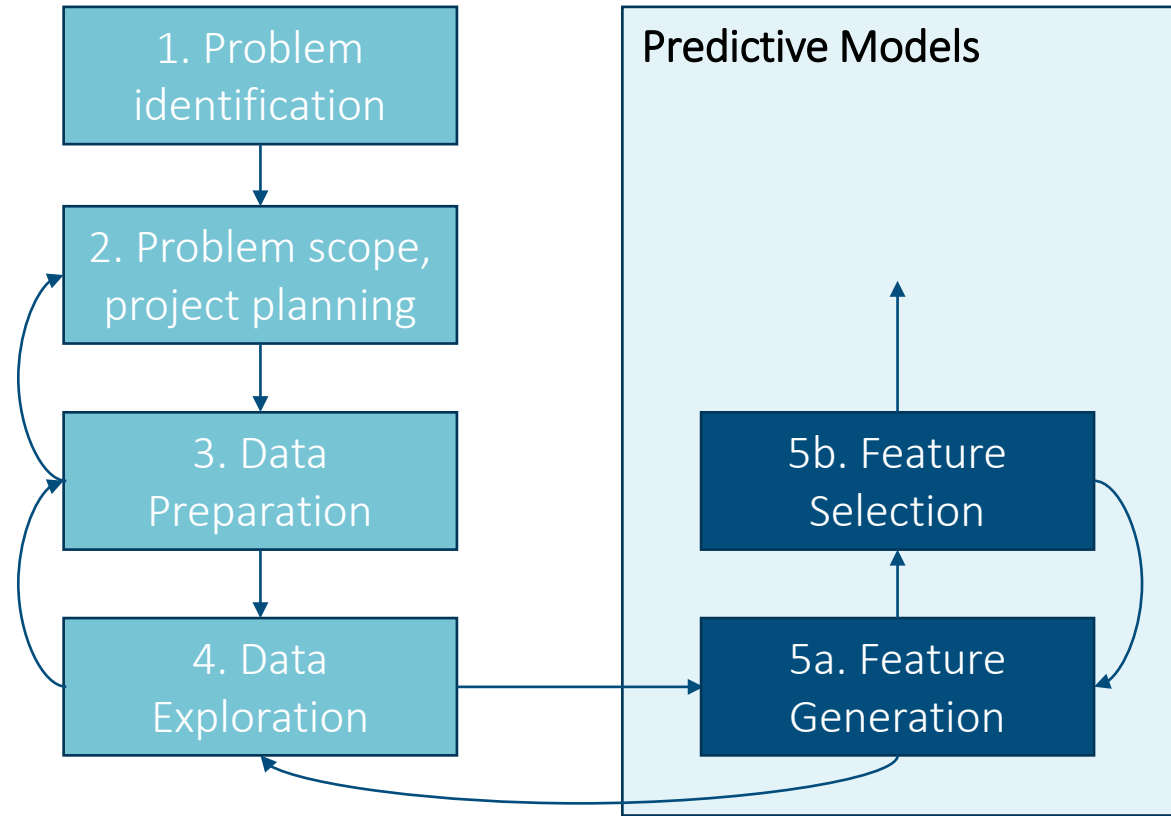
# Module 5 – Feature Generation

- Subject matter experts
  - We know that mortality risk might depend on the distance of the insured from a highly polluted area
- Feature transformations
  - Standardize or normalize
  - Log of skewed distributions
  - Group postcodes into regions
- Unsupervised modelling techniques
  - Clustering
  - PCA

# Module 5 – Feature Selection

- Filter based techniques
  - Measure correlations with the target variable
    - Pearson's, Spearman's, Kendall
    - Mutual information
- Algorithmic based techniques that are specific to a model
  - Feature selection is “built-in” to the optimization process of the algorithm
    - Lasso regression
    - Stepwise approaches based on AIC, BIC, adjusted R, etc.
  - Importance of features can be calculated from resulting model
    - Tree-based methods

# Summary – an iterative process





SOCIETY OF  
ACTUARIES®

# Predictive Analytics Certificate Program

---

## Practice: Module 5



# Module 5 – Practice

- Discuss different approaches to feature generation and selection
- Discuss R programming questions/best practice





SOCIETY OF  
ACTUARIES®

# Predictive Analytics Certificate Program

---

## Lunch





SOCIETY OF  
ACTUARIES®

# Predictive Analytics Certificate Program

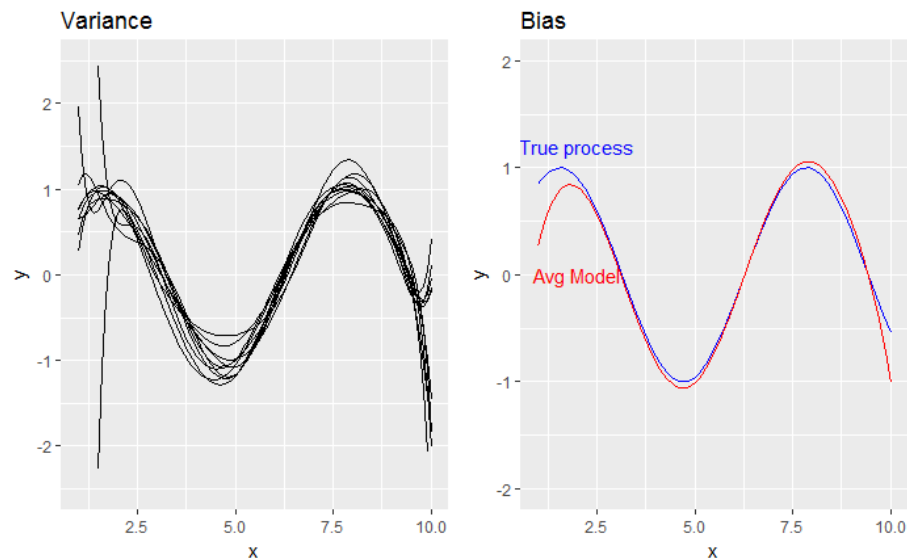
---

## Review: Module 6



# Module 6 – Modeling and Validation

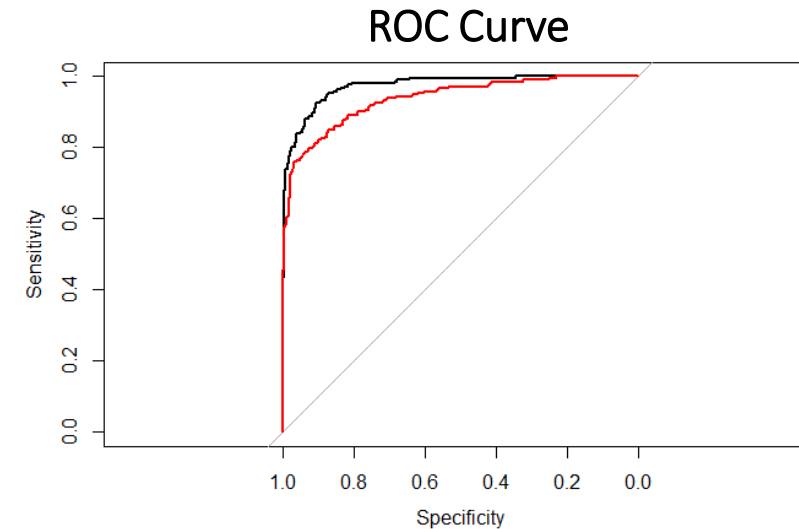
- Supervised vs unsupervised models
- Learning objectives – regression or classification?
- The bias-variance decomposition



# Module 6 – Model selection

## Assessing the goodness of fit

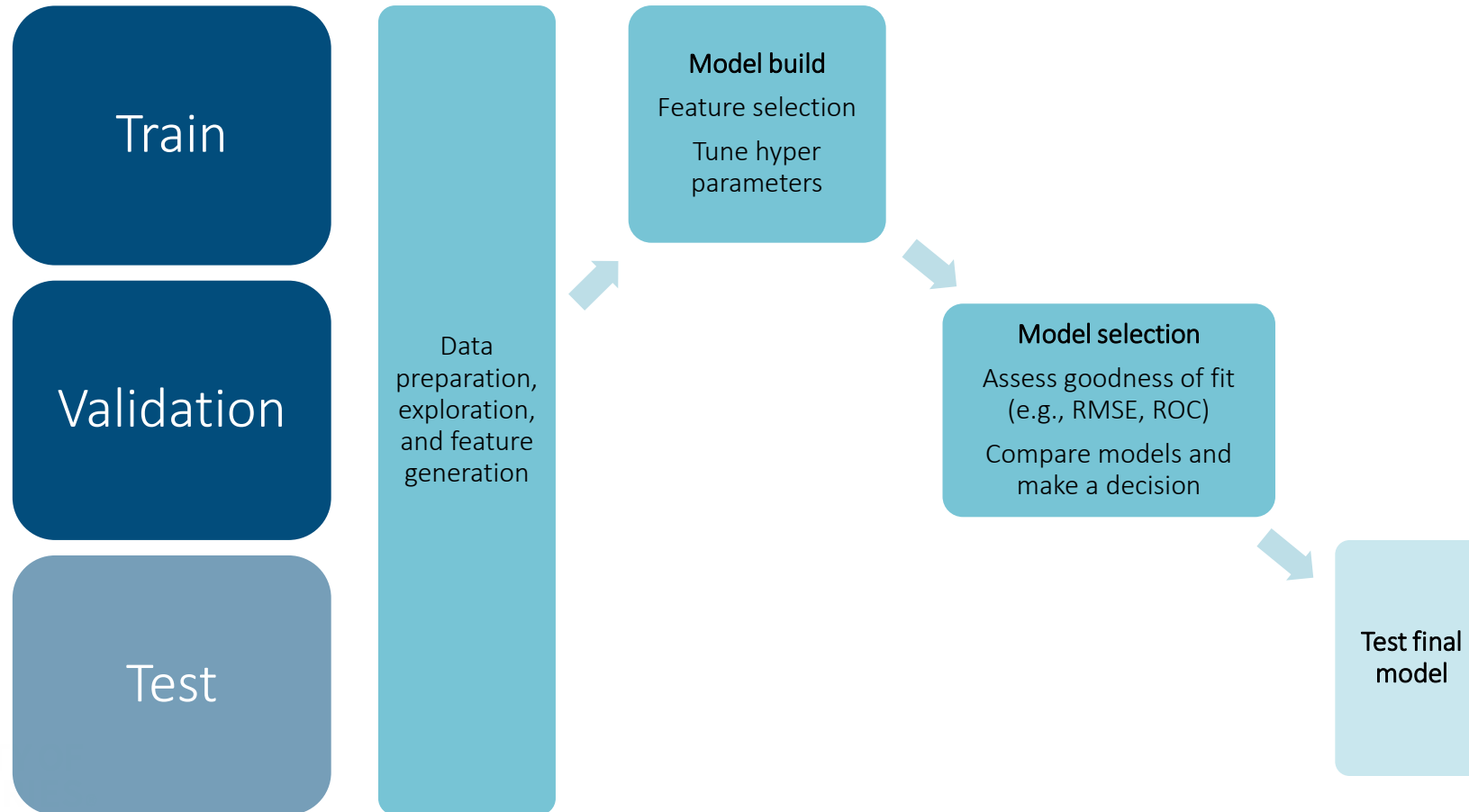
- Regression
  - RMSE/MAE
  - Residual Analysis
- Classification
  - Classification Error
  - ROC and AUC
- Information loss to select the best model
  - AIC/BIC



Ultimately we want to minimize the generalization error

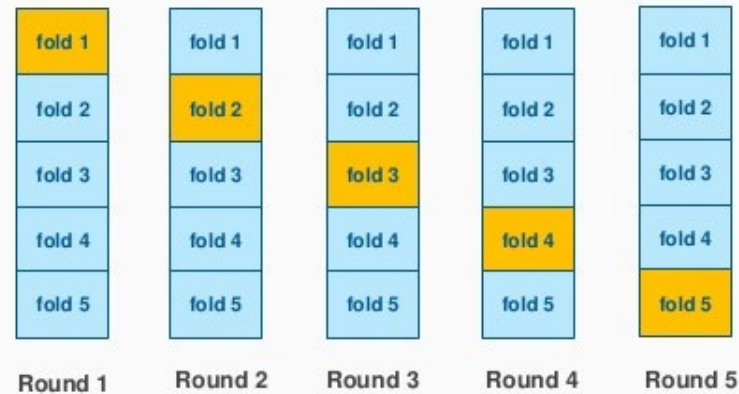
# Module 6 – The validation framework

- Train/Validation/Test data



# Module 6 – K-fold Cross Validation

## K-fold Cross Validation (K = 5)



*score(CV) = the average of evaluation scores from each fold*  
*You can also repeat the process many times!*

Training Data  
Validation Data


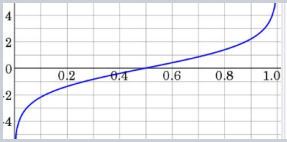
11

# Module 6 – GLMs

$$E(Y) = \mu = g^{-1}(X\beta)$$

Distribution	Support of distribution	Typical uses
Normal	real: $(-\infty, +\infty)$	Linear-response data
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters
Gamma		
Inverse Gaussian	real: $(0, +\infty)$	
Poisson	integer: $0, 1, 2, \dots$	count of occurrences in fixed amount of time/space
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence
Binomial	integer: $0, 1, \dots, N$	count of # of "yes" occurrences out of N yes/no occurrences
Categorical	integer: $[0, K)$	outcome of single K-way occurrence
	K-vector of integer: $[0, 1]$ , where exactly one element in the vector has the value 1	
Multinomial	K-vector of integer: $[0, N]$	count of occurrences of different types (1 .. K) out of N total K-way occurrences

Source: [https://en.wikipedia.org/wiki/Generalized\\_linear\\_model](https://en.wikipedia.org/wiki/Generalized_linear_model)

Link option	Link function
Identity	$\mathbf{X}\beta = \mu$
Log	$\mathbf{X}\beta = \ln(\mu)$ 
Logit	$\mathbf{X}\beta = \ln\left(\frac{\mu}{1 - \mu}\right)$ 

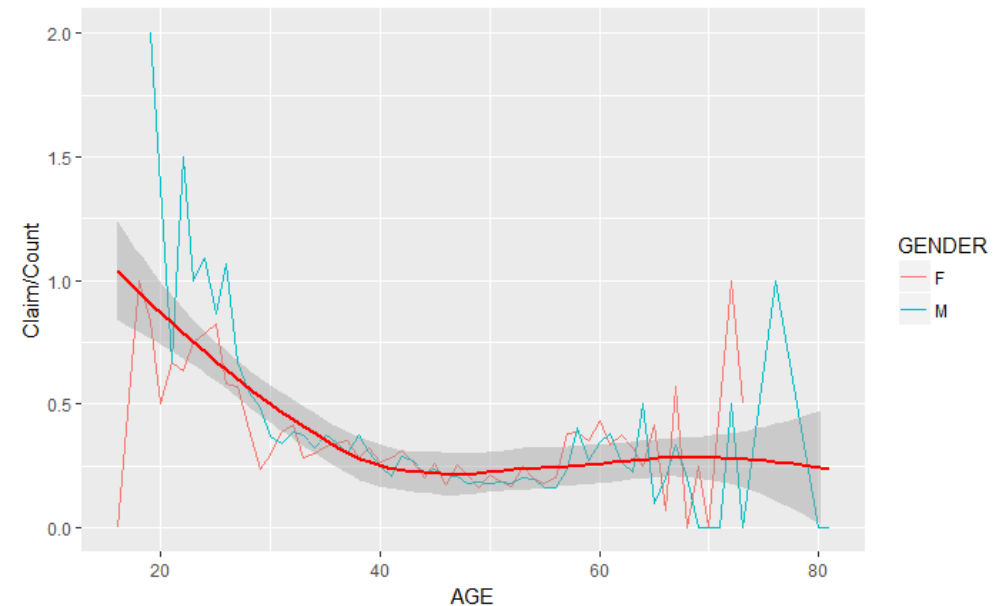
# Module 6 – GLMs

## Advantages:

- ✓ Efficient computation
- ✓ Interpretable/transparent
- ✓ Smooth prediction surface

## Disadvantages:

- ✗ Usually high bias (can under-fit)
- ✗ Linear in the parameters





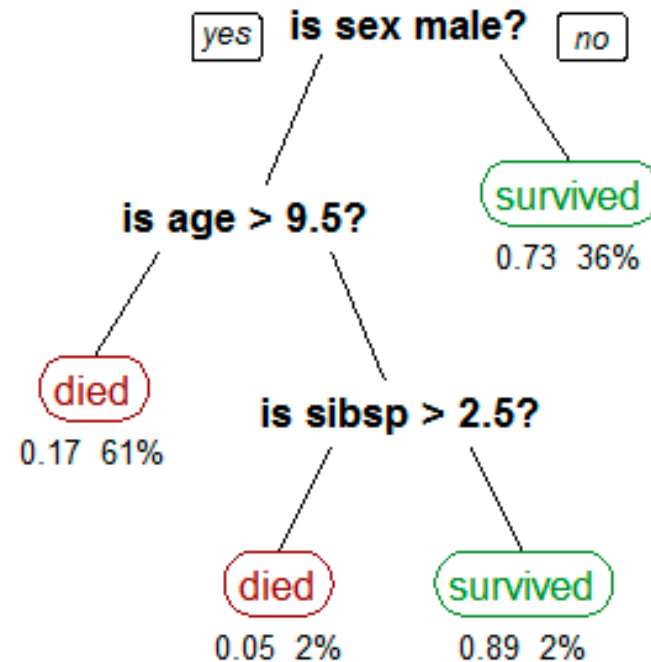
# Module 6 – Decision Trees

## Advantages:

- ✓ Simple
- ✓ Non-linear
- ✓ Interpretable/transparent
- ✓ Handles missing values

## Disadvantages:

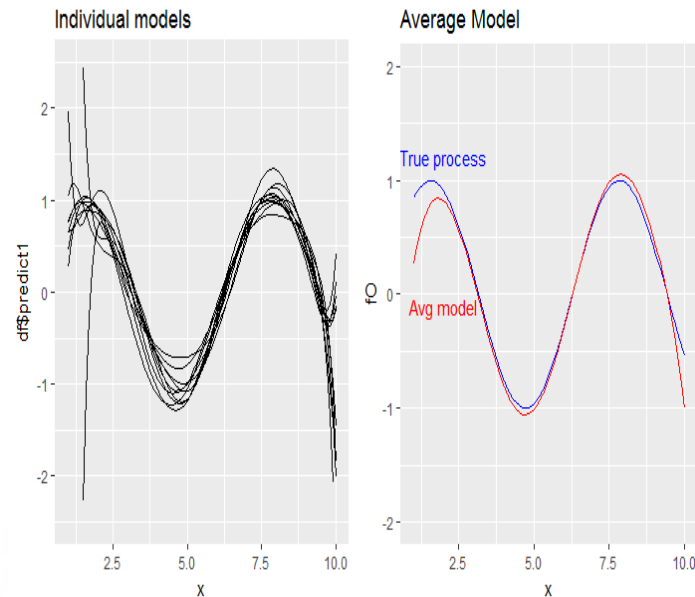
- ✗ Sensitive to noise/unstable
- ✗ Greedy (non-optimal)
- ✗ Non-smooth prediction surface
- ✗ Poor predictive power



# Module 6 – Ensembles

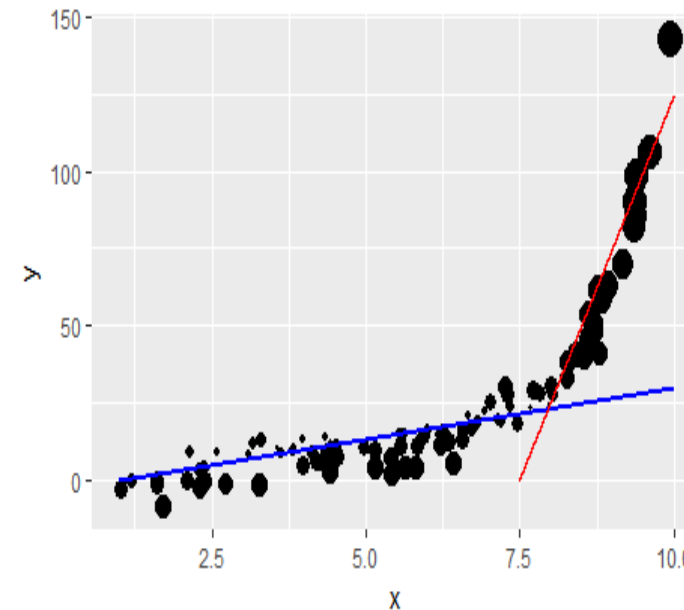
## Bagging

- Train many *independent* models in parallel
- Take average of outputs



## Boosting

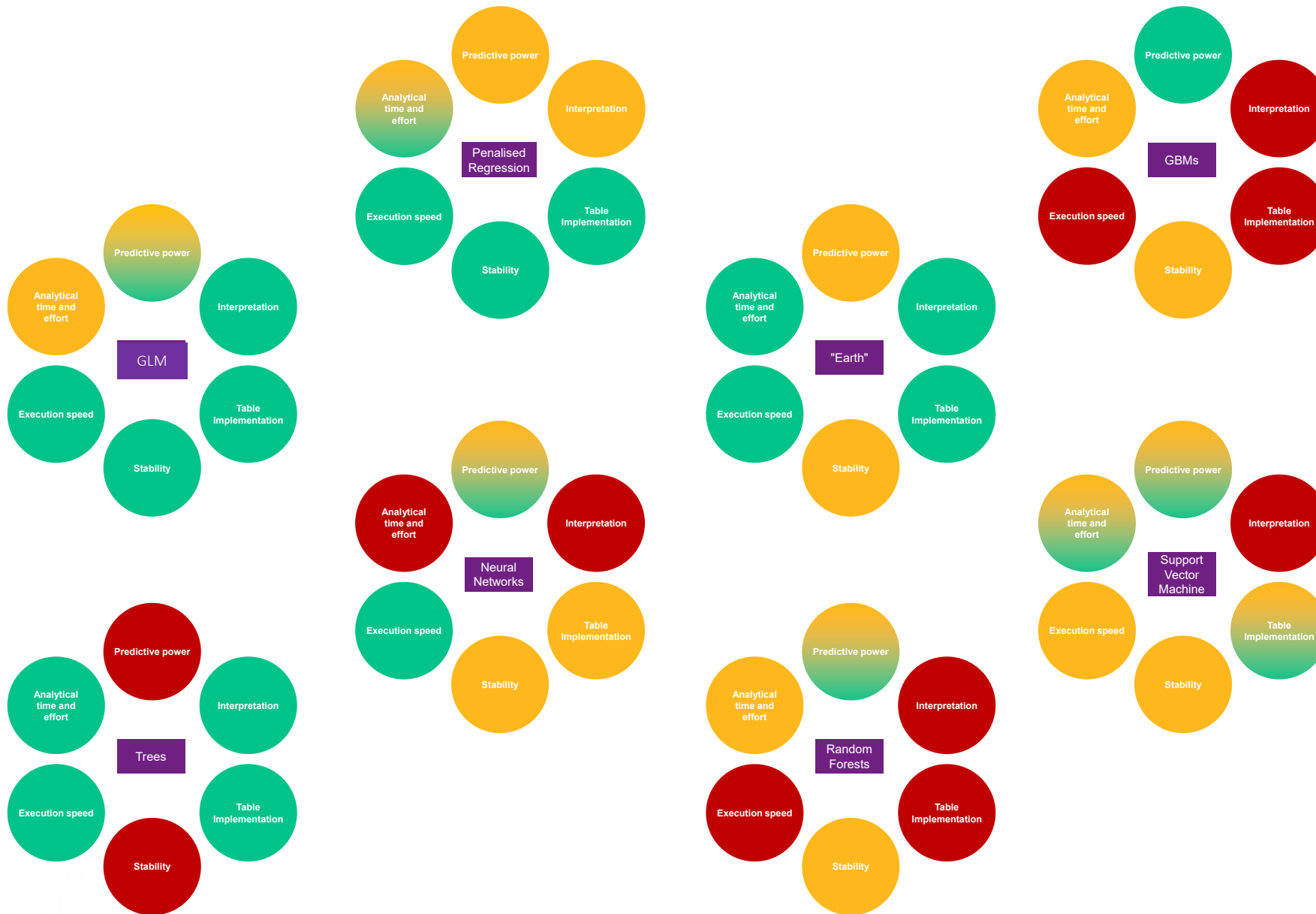
- Train many *dependent* models each on the scaled residuals of the previous.
- Final model is a weighted sum of all component models



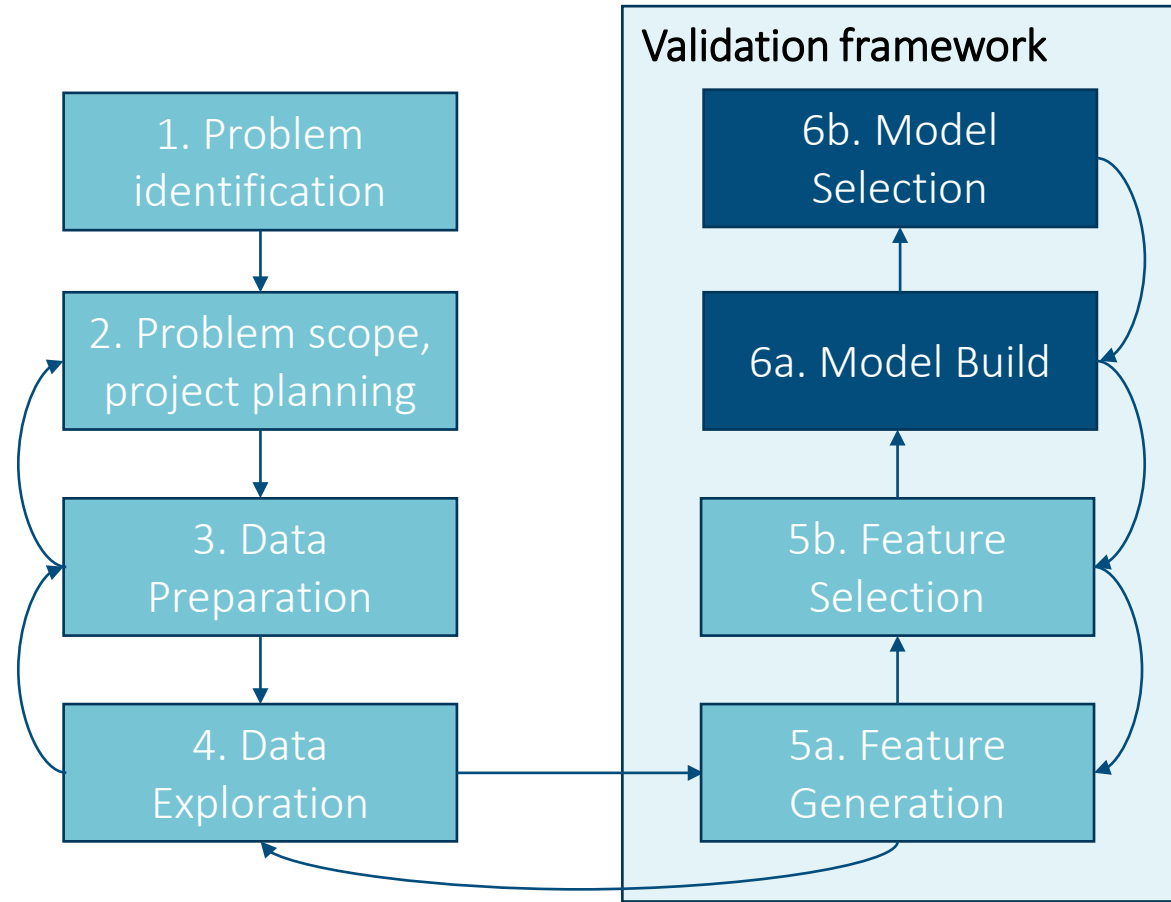
# How to select the right model

- What type of problem are you trying to solve?
  - Note you can often rephrase the same problem in different ways that will result in different models
- Implementation constraints
- Time constraints
- Secondary objectives
  - E.g. is it important to be able to interpret the models, at least initially





# Summary – an iterative process



# Modeling best practice

- Data must be in decent shape before spending too much time on model building (GIGO)
- Feature generation/selection often gives more improvement than complex models
- Remember Occam's Razor and don't underestimate the curse of dimensionality
- An iterative approach is best! Create a simple model with a subset of data **first** – then progressively refine and improve.



SOCIETY OF  
ACTUARIES®

# Predictive Analytics Certificate Program

---

## Discussion...







SOCIETY OF  
ACTUARIES®

# Predictive Analytics Certificate Program

---

## Practice: Module 6





SOCIETY OF  
ACTUARIES®

# Predictive Analytics Certificate Program

---

## Break





SOCIETY OF  
ACTUARIES®

# Predictive Analytics Certificate Program

---

## Communication



# Importance of Communication

*“I don't start with a design objective, I start with a communication objective. I feel my project is successful if it communicates what it is supposed to communicate.”*

- Mike Davidson

# Understand the Audience

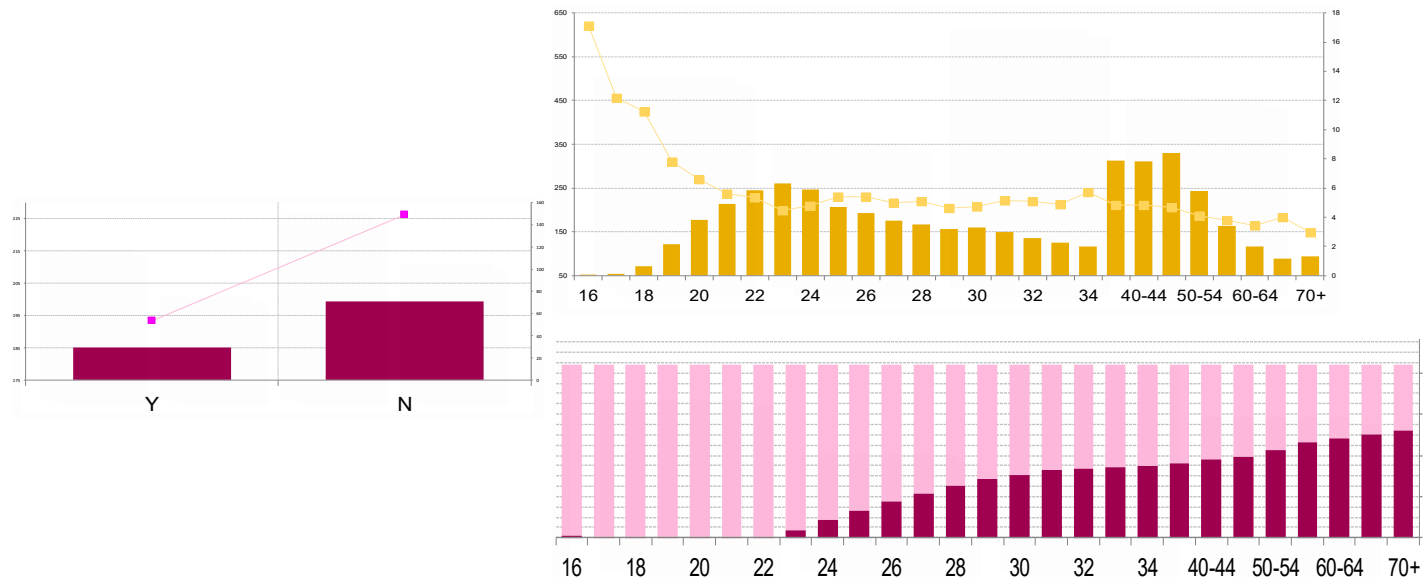
- Frame the problem in the proper business context
- Tailor to the audience
- Keep in mind accessibility (e.g. individuals who may be color-blind)

# What to Communicate

- Provide the answer
  - Provide a concise summary that links your results to the business problem
- Support your answer with your analysis
- Clear audit trail – reproducibility is key
- **Remember - an underwhelming result is just as important to communicate clearly as a game-changer is**

# Communication Style

- Use structure to guide the audience to what is important
- Be concise yet don't be afraid to be a bit repetitive
- Use visual aids





# Art of Communication

Be clear about the results and approach you have taken but...

... don't get bogged down in the technical detail!

While you are likely to be excited about the intricacies of modeling, your audience will probably lose sight of what is actually important.



# Example report format

Report audience: technical peers and management

- Executive summary
- Problem statement
- Data
- Approach/Method
- Results & conclusions
- Appendices (code etc.)



SOCIETY OF  
ACTUARIES®

# Predictive Analytics Certificate Program

---

## Exercise: Report



# Example Report Exercise

- Exercise
  - Outline what the report would look like from the module 6 practice (15 mins)
  - Discuss in the full group what you would put into each section and how you would communicate your results (15 mins)



SOCIETY OF  
ACTUARIES®

# Predictive Analytics Certificate Program

---

## Wrap Up



# Tomorrow's Project

- You may use/access any of the course materials including previously produced code
- You may use books and/or notes
- The dataset has been distributed to participants; problem statement will be distributed tomorrow
  - All passwords will be revealed tomorrow.
- Use the Assessment ID (see email message) as part of your file nomenclature
- Information on how to ask the faculty questions regarding coding will be shared on Thursday morning.

# Project Information

- Participants may use time allotted for completion of the project as they wish. Recommended allocation
  - Morning: Coding and data analysis work
  - Afternoon: Write up report and analysis
- The majority of the marks will be awarded for good modelling practice, which can be achieved with a relatively simple model. You should only move onto more complex models/techniques once you have demonstrated the approach with a simple one.
- You may work through lunch; all work must be submitted by 4:30pm (Central) to [PACertificate@soa.org](mailto:PACertificate@soa.org).

# Grading Rubric

COMPONENT		WEIGHTING	
REPORT		36%	
	Report Structure	1%	
	Tables, graphs, images	1%	
	Appendices	1%	
	Executive summary	5%	
	Problem Statement	5%	
	Data	5%	
	Method/Assumptions/Issues	10%	
	Results	8%	
CODE		12%	
	Code readability & structure	4%	
	Code functionality	8%	
ANALYSIS		52%	
	Data preparation/cleaning	4%	
	Data exploration	10%	
	Feature generation/selection	8%	
	Modelling	15%	
	Validation	15%	