Predictive Analytics and Modelling of Data

CMSE11428 (2023/24)

Week 1

Dr Antonia Gieschen

The University of Edinburgh Business School

- Predictive Modelling
- Components of Predictive Models
- Supervised vs. Unsupervised Modelling
- Regression vs. Classification
- Types of variables
- Challenges in predictive modelling

Course overview

This course aims at training students in the field of predictive analytics to respond to the job market needs using a variety of methodologies. Students' journey shall be a quest to distinguish the "true" signal from a universe of "noise" through the lenses of predictive analytics. To be more specific, this course covers the typical **methodological steps** *of a prediction exercise, statistical modelling, and artificial intelligence methodologies for prediction of applications in business and economics. It also covers* **practical issues** *in predictive analytics and how to address them.*

Course overview

20 lecture hours, 10 tutorial hours over 11 weeks (incl. 1 reading week).
Guest lecture in week 5.
Week-by-week course structure can be found on Learn.
Theoretical concepts will be introduced in the lecture, then implemented during the computer labs.

Recommended readings are:

- Kuhn, M. and Johnson, K., 2013. Applied predictive modelling. New York: Springer.
- Witten, D. and James, G., 2013. An introduction to statistical learning with applications in Python. Springer.

Course overview

Assessment will be two-fold:

60% coursework (group)	40% exam (individual)

- Groupwork will consist of a report documenting the analysis of a provided dataset using techniques learned in class
- Further details will be shared in due course
- Dates are TBC, but exam will likely happen in December

Introducing myself

Dr Antonia Gieschen (she/her)
antonia.gieschen@ed.ac.uk
Office 3.24

- PhD from University of Edinburgh with a dissertation on spatio-temporal cluster analysis
- Postdoc at Carnegie Mellon University (Pittsburgh PA, USA) within Pittsburgh Supercomputing Center
- Now Lecturer in Predictive Analytics and teaching Principles of Data Analytics (CMSE11432) as well as Predictive Analytics and Modelling of Data (CMSE11428)

Introducing myself

My research interests are in the area of computational social science. I'm broadly interested in modelling:

- marketing, consumer behaviour and tourism
- financial wellbeing and mental/physical population health
- local food systems/access to fresh produce
- and especially spatial inequality in all of the above.

Today's lecture (Week 1)

Topics covered today:

- Introduction to Predictive Modelling and this course
- Overview of different types of models
- Structure of the modelling process

Recommended readings:

- Kuhn & Johnson: Chapter 1 and 2
- Witten & James: Chapter 1 and Section 2.1

Predictive Modelling

What is Predictive Modelling?

Everyday decisions...

- Which way should I walk to get to X fastest?
- When is the best time to buy fresh bread at the local bakery?
- Which car should I buy?

What is Predictive Modelling?

Everyday decisions...

... made based on information

- Which way should I walk to get to X fastest? Google Maps says that my usual road has been closed, I'll take this alternative which is always free.
- When is the best time to buy fresh bread at the local bakery? This Facebook review says they always bake fresh bread at 11am, I'll be there at 11:15.
- Which car should I buy? My mother recommends this brand based on her experience so I have checked the technical specifications, which suit my preferences.

What is Predictive Modelling?

But there are decisions which are much harder to make and sometimes information can be too much to deal with as a human.

- Which ad should we run for this new product?
- Should we invest in this stock?
- How will the housing prices develop in the next five years?

Enter: Predictive Modelling!

What is Predictive Modelling?

Predictive Modelling: the process of developing a mathematical tool or model that generates an accurate prediction.

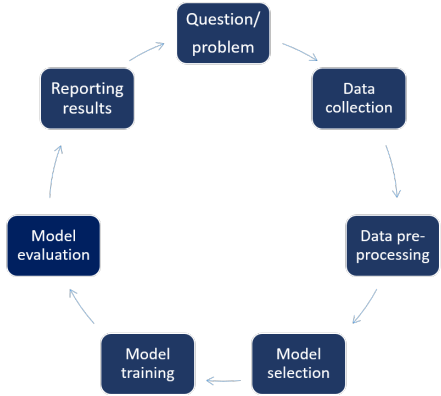
Kuhn & Johnson (2013) based on Geisser (1993)

The Predictive Modelling process

Idealised version

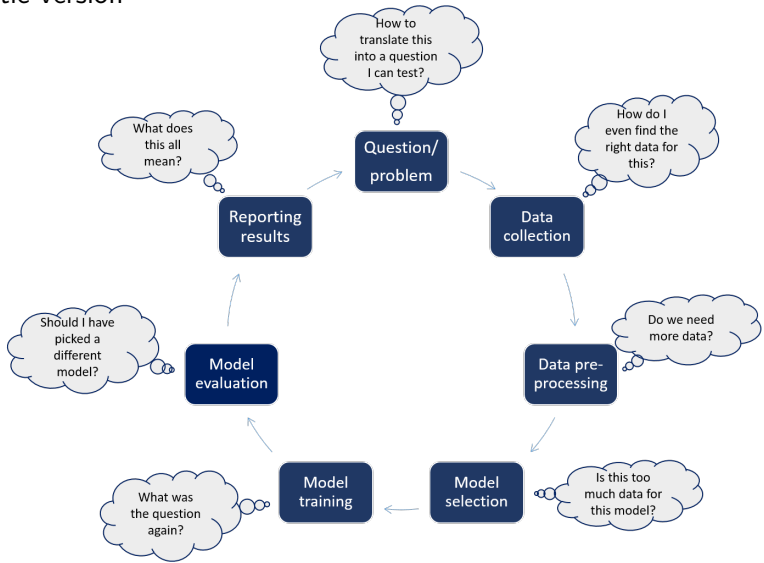
The Predictive Modelling process

Realistic version

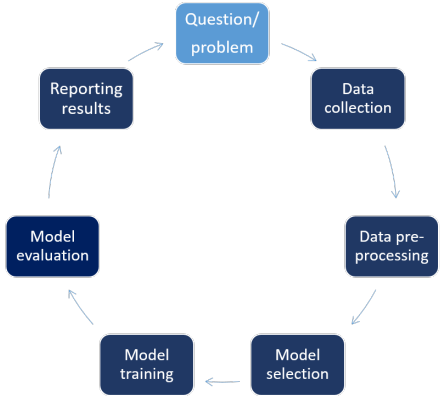


The Predictive Modelling process

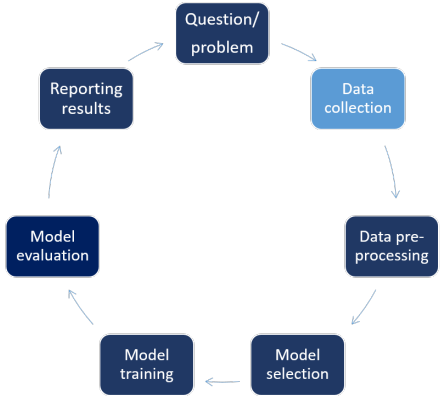
Realistic version



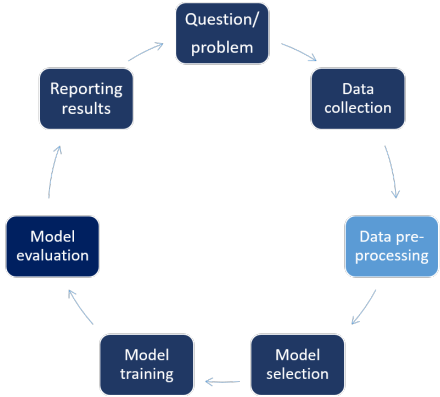
The Predictive Modelling process



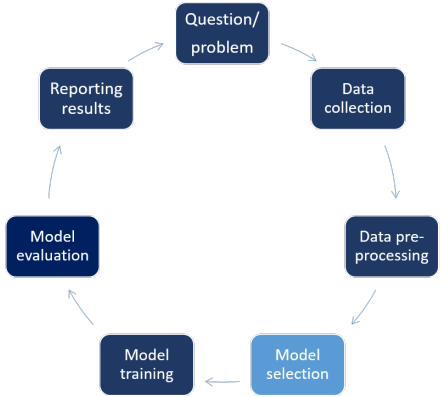
The Predictive Modelling process



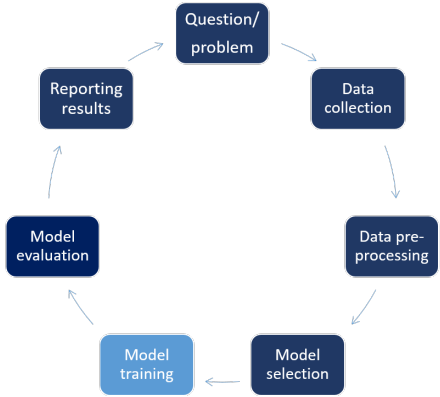
The Predictive Modelling process



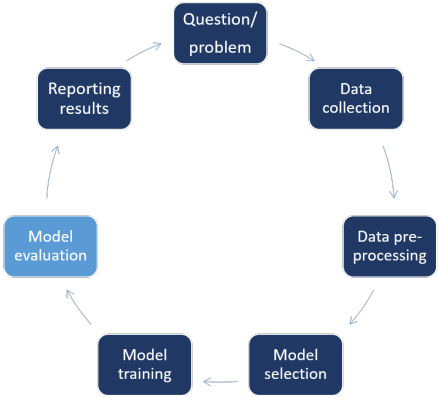
The Predictive Modelling process



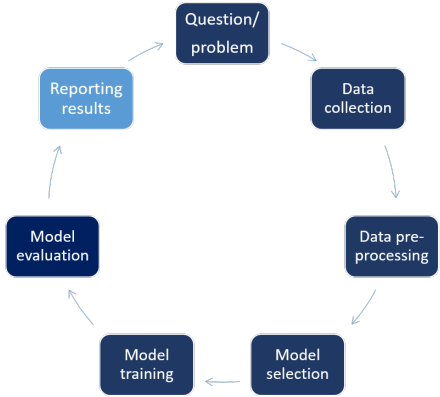
The Predictive Modelling process



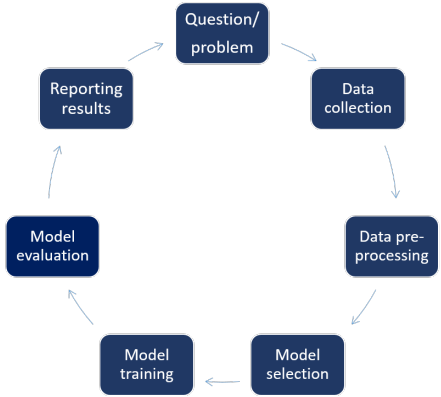
The Predictive Modelling process



The Predictive Modelling process



The Predictive Modelling process



Components of Predictive Models

Components of Predictive Modelling

Predictive Model:

$$y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

- $X_1, X_2, \ldots, X_p$: features (explanatory/independent variables), where p is the number of predictors
- y: target (label, response, or dependent variable)
- ε: error term

Vectors and matrices are usually written in boldface:

- $\boldsymbol{X}$: an $n \times p$ matrix with observations as rows, $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n})^{\mathrm{T}}$ the i-th observation of the j-th predictor $x_{ij}, i = 1 \ldots n, j = 1 \ldots p$
- $\boldsymbol{y}$ is an $n \times 1$ vector

- The function f is what we want to learn. Use available dataset $\{\boldsymbol{X}, \boldsymbol{y}\}$ to learn the relationship f then use f to obtain $\widehat{y}_i$ given $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$

Terminology of Predictive Modelling

Terminology can differ between scientific domains. A couple of examples for this are:

- data points are often called: sample (the singular being used for individual points, but can also be used for the statistical concept of a sample of multiple datapoints), observation, instance or measurement
- features X are often called: predictors, independent variables, input, attributes or descriptors
- dependent variables Y are often called: target class, outcome or response variables

Terminology of Predictive Modelling contd.

- data types can be referred to differently depending on domain, e.g. categorical data is also known as nominal, attribute or discrete, or in some instances qualitative variable; ordinal data is sometimes known as ranked data or Likert scale data in case of that scale (usually in social sciences)
- model training, building and parameter estimation can all refer to the same process, though, parameter estimation is usually used for a specific aspect while the general term building can refer to a wider number of tasks

Supervised vs. Unsupervised Modelling

Supervised vs. Unsupervised Modelling

Predictive models mostly fall into one of two categories: supervised or unsupervised.

- In supervised modelling, for each observation of the predictor measurement(s) $x_i, i = 1, \ldots, n$ there is an associated response measurement y_i.
  - The modelling process fits a model to relate the response to the predictors with the aim of accurately predicting the response for future observations.
  - For this we **train** the model on a set of labelled data with response variable and then **test** it on a set of unlabelled data without the response variable, to assess the performance.
- In unsupervised modelling, for every observation $i = 1, \ldots, n$, we observe a vector of measurements x_i but no associated response y_i.
  - The goal is not to fit a model to predict the response for future observations.
  - Instead the objective is usually to explore and describe the data, for example for segmentation purposes.

Supervised vs. Unsupervised Modelling

Some supervised and unsupervised models we will cover in this course:
Supervised:

- Simple and multiple linear and logistic regression
- K-nearest neighbours
- Decision trees and random forests
- Support vector machines
- Artificial neural networks

Unsupervised:

- Principal component analysis[1]
- Cluster analysis

[1] Whether this "counts" depends on your definition of machine learning, as PCA is a dimensionality reduction technique

Example for supervised model: Predicting visitor count

Imagine you want to predict the likelihood of a tourist visiting a visitor attraction in Scotland. You collect the following data:

Index	Age	Nationality	Visited?
1	24	Scottish	1
2	56	English	0
3	42	Scottish	1
4	43	Welsh	1
5	77	English	0

What could be the rules that our model f learns?

Example for supervised model: Predicting visitor count

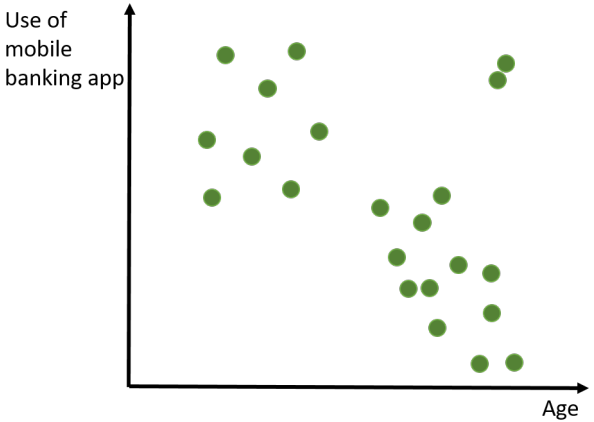
You are now given the following two new records and are trying to predict whether they will visit. What is your guess?

Index	Age	Nationality	Visited?
1	24	Scottish	1
2	56	English	0
3	42	Scottish	1
4	43	Welsh	1
5	77	English	0

Index	Age	Nationality	Visited?
1	70	Scottish	?
2	20	English	?

Example for unsupervised model: Segmenting customers

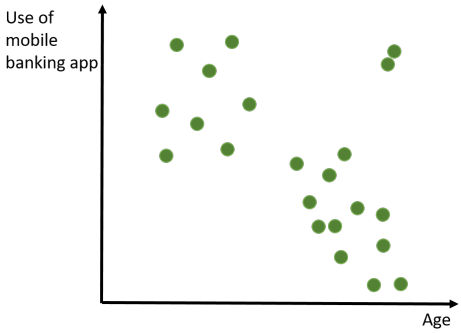
A bank is collecting data on their customers and wants to better understand the structure of their customer base. Look at the figure below. What kind of patterns could an **unsupervised** model find?

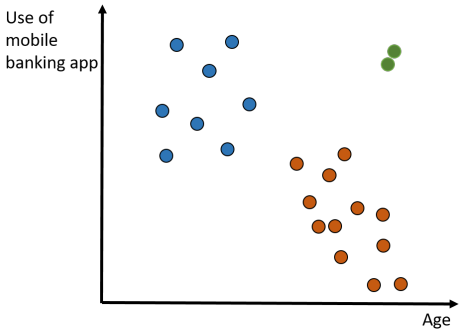


Example for unsupervised model: Segmenting customers

An example can be found below. Different colours denote different segments the model might find.

Note that the general trend you might observe is not found through an unsupervised model. But interpreting the resulting groups might lead to that conclusion.





Regression vs. Classification

Regression vs. Classification

- Regression and classification are categorized under the same umbrella of supervised modelling. Both share the same concept of utilizing known datasets to make predictions.
- In classification we are interested in the task of predicting a discrete class label.
- In regression we are interested in the task of predicting a continuous quantity.
- There is some overlap between the algorithms for classification and regression. Moreover, in some cases it is possible to convert a regression problem to a classification problem.

Regression vs. Classification

Regression or classification?

1. What will the stock price for this company be next month?
2. Will this customer of my company churn?
3. Which genre does this movie belong to based on its script?
4. How many visitors will this museum attract with its new exhibition?

Regression vs. Classification

Regression or classification?

1. What will the stock price for this company be next month? → Regression
2. Will this customer of this company churn (move to a competitor)? → Classification
3. Which genre does this movie belong to based on its script? → Classification
4. How many visitors will this museum attract with its new exhibition? → Regression

Types of variables

Different types of variables

Generally, we first distinguish between quantitative and qualitative data.

- Quantitative data is numeric, for example, a count, ratio, real number, anything that is recorded through numeric values.
- Qualitative data is non-numeric and in qualitative analysis it is treated as such. For example, text and transcripts, images, video recordings, sounds.

Different types of variables

Generally, we first distinguish between quantitative and qualitative data.

- Quantitative data is numeric, for example, a count, ratio, real number, anything that is recorded through numeric values.
- Qualitative data is non-numeric and in qualitative analysis it is treated as such. For example, text and transcripts, images, video recordings, sounds.

Note the "treated as such": Much qualitative data nowadays could be transformed to quantitative data and be analysed with the usual techniques. For categorical data, we do that. However, in many cases that leads to a loss of information and accuracy for the sake of being able to analyse large amounts of it.

For example: in-depth analysis of a small number of customer interviews in context vs. automatic text recognition to learn the general themes
Neither is better or worse than the other - it depends on the question you're asking!

Different types of variables

Within quantitative data, we can distinguish generally between:

- discrete
  - binary (0/1)
  - categorical / nominal (categories)[2]
  - ordinal (ranking)
  - numeric / integer / count (countable numbers)
- continuous
  - interval (zero has no meaning / no true zero)
  - ratio (true zero)

[2] This is usually considered qualitative data and transformed for analysis

Different types of variables

What is the most likely data type?

1. Temperature measured in degree Celsius
2. Number of visitors to a theme park
3. Education level achieved as recorded by the census
4. Favourite flavour of soda of a sample of restaurant visitors
5. Price for an item over time
6. Whether a patient has a disease or not

Different types of variables

What is the most likely data type?

1. Temperature measured in degree Celsius $\rightarrow$ interval
2. Number of visitors to a theme park $\rightarrow$ numeric
3. Education level achieved as recorded by the census $\rightarrow$ ordinal
4. Favourite flavour of soda of a sample of restaurant visitors $\rightarrow$ categorical
5. Price for an item over time $\rightarrow$ ratio
6. Whether a patient has a disease or not $\rightarrow$ binary

Different types of variables

In many application cases in the social sciences, you will encounter a mixture of data types for the same research question.

- A categorical response with a mixture of numeric and ordinal input
- A regression problem will continuous output, but both numeric and binary input

This brings a lot of challenges: How to treat variables of different type during pre-processing? How much information can get gained from categorical or ordinal variables? What do ordinal variables tell us? Which models are suitable for mixed data?

Different types of variables

It's **extremely important** to always check the assumptions of your model before starting your analysis.

Model choice should always be both question/problem **and** data driven.

Some important things to keep in mind and use when handling mixed data:

- Pre-processing (e.g. "one-hot encoding" for categorical data)
- Data selection (e.g. which variables should be chosen / which give us the best chances $\rightarrow$ expert knowledge?)
- Model selection (e.g. decision trees can handle categorical data better than some other models)
- Interpretation (e.g. care should be taken when interpreting numbers when ordinal variables are involved)

Challenges in predictive modelling

Challenges in predictive modelling

According to Kuhn & Johnson (2013):

- inadequate data pre-processing
- inadequate model validation
- unjustified extrapolation
- over-fitting the model to the existing data.

Challenges in predictive modelling

And a few more from me:

- Too little or too much data for the chosen model (picking the model before the data)
- p-value hacking (looking for something that isn't there)
- over-reliance on theory or "general knowledge" (not believing the data)

What comes next?

- Computer lab on Wednesday in the Hugh Robson Building
- Next week's lecture will be on data cleaning and pre-processing
- Assessment information will be published on Learn at the end of the week

References

- Geisser, S., 1993. Predictive Inference: An Introduction. Chapman and Hall.
- Kuhn, M. and Johnson, K., 2013. Applied predictive modeling. New York: Springer.