MSCI152: Introduction to Business Intelligence and Analytics

Lecture 15: Modelling in practice

Dr Anna Sroginis

Lancaster University Management School

Agenda

1. Recap
2. Project understanding
3. Data understanding
   - Missing values
4. Data preparation
5. Modelling
   - Under-/overfitting
6. Evaluation
   - Prediction accuracy versus model interpretability

Highly recommend to read Berthold et al. (2010) Guide to Intelligent Data Analysis

More details can be found Camm et al. Chapters 5.1, 5.2

Recap

What we looked at:

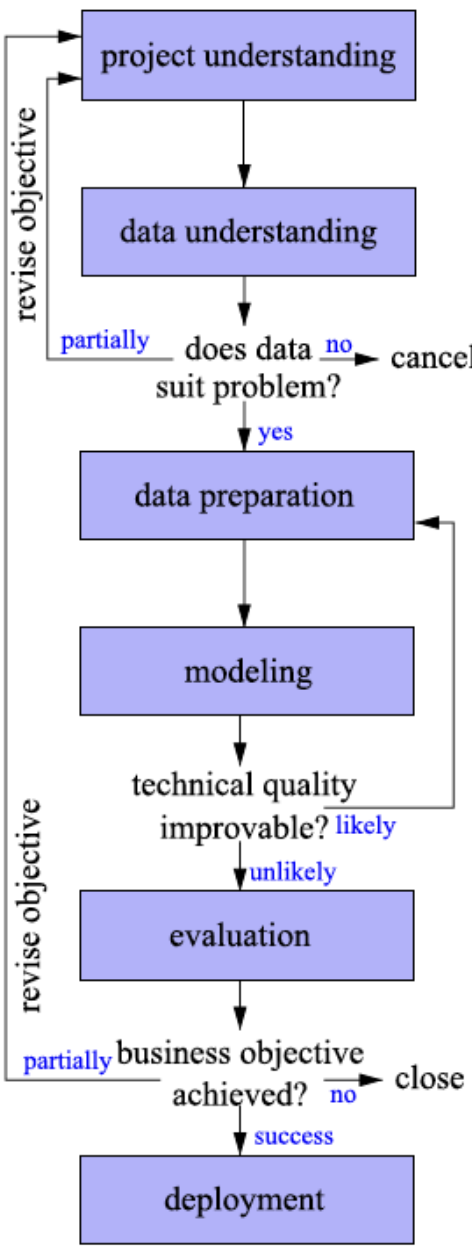
- Types of data
- Sampling methods
- Descriptive statistics
- Correlation and regression
- Introduction to forecasting

What other problems might we have?

Agenda

1. Recap
2. Project understanding
3. Data understanding
    - Missing values
4. Data preparation
5. Modelling
    - Under-/overfitting
6. Evaluation
    - Prediction accuracy versus model interpretability

The data analysis process



What exactly is the problem, the expected benefit?
What should a solution look like?
What is known about the domain?

What data do we have available?
Is the data relevant to the problem?
Is it valid? Does it reflect our expectations?
Is the data quality, quantity, recency sufficient?

Which data should we concentrate on?
How is the data best transformed for modeling?
How may we increase the data quality?

What kind of model architecture suits the problem best?
What is the best technique/method to get the model?
How well does the model perform technically?

How good is the model in terms of project requirements?
What have we learned from the project?

How is the model best deployed?
How do we know that the model is still valid?

Project understanding

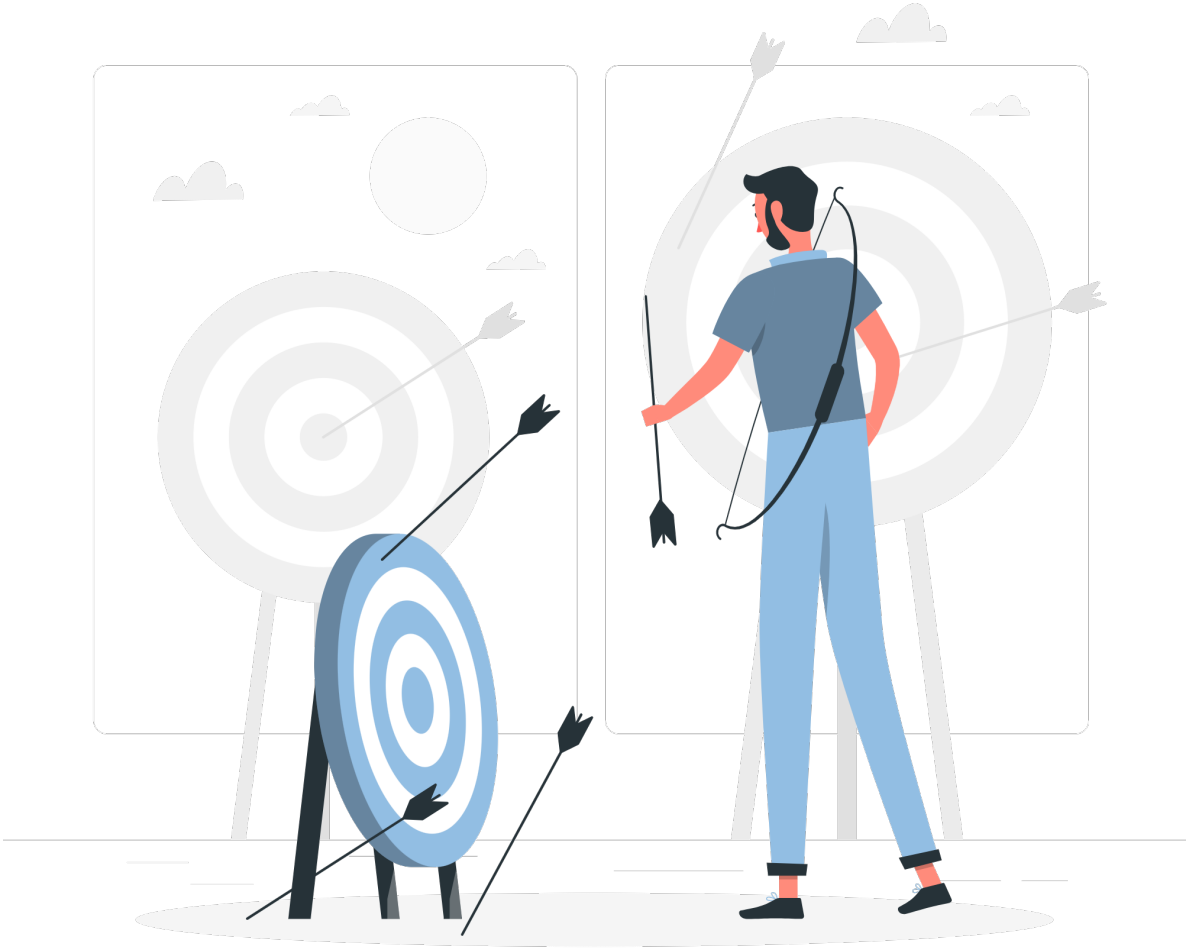
Any modelling must start with a problem:

- Problem can be captured by some data sets (?)
- Appropriate modelling techniques can learn relationships (?)
- Findings or models can be transferred back to real problem and applied successfully (?)

While time spent on project and data understanding is small compared to data preparation and modelling (20% : 80%) the importance to success is the opposite

Setting objectives

1. Objective?
2. Deliverable?
3. Success criteria?



Agenda

1. Recap
2. Project understanding
3. Data understanding
   - Missing values
4. Data preparation
5. Modelling
   - Under-/overfitting
6. Evaluation
   - Prediction accuracy versus model interpretability

Data understanding

- What kind of data do we have?
- Is the data relevant to the problem? Is it sufficient?
- Is it valid? Showing what we would expect?
- Any missing values? Errors?

Missing values

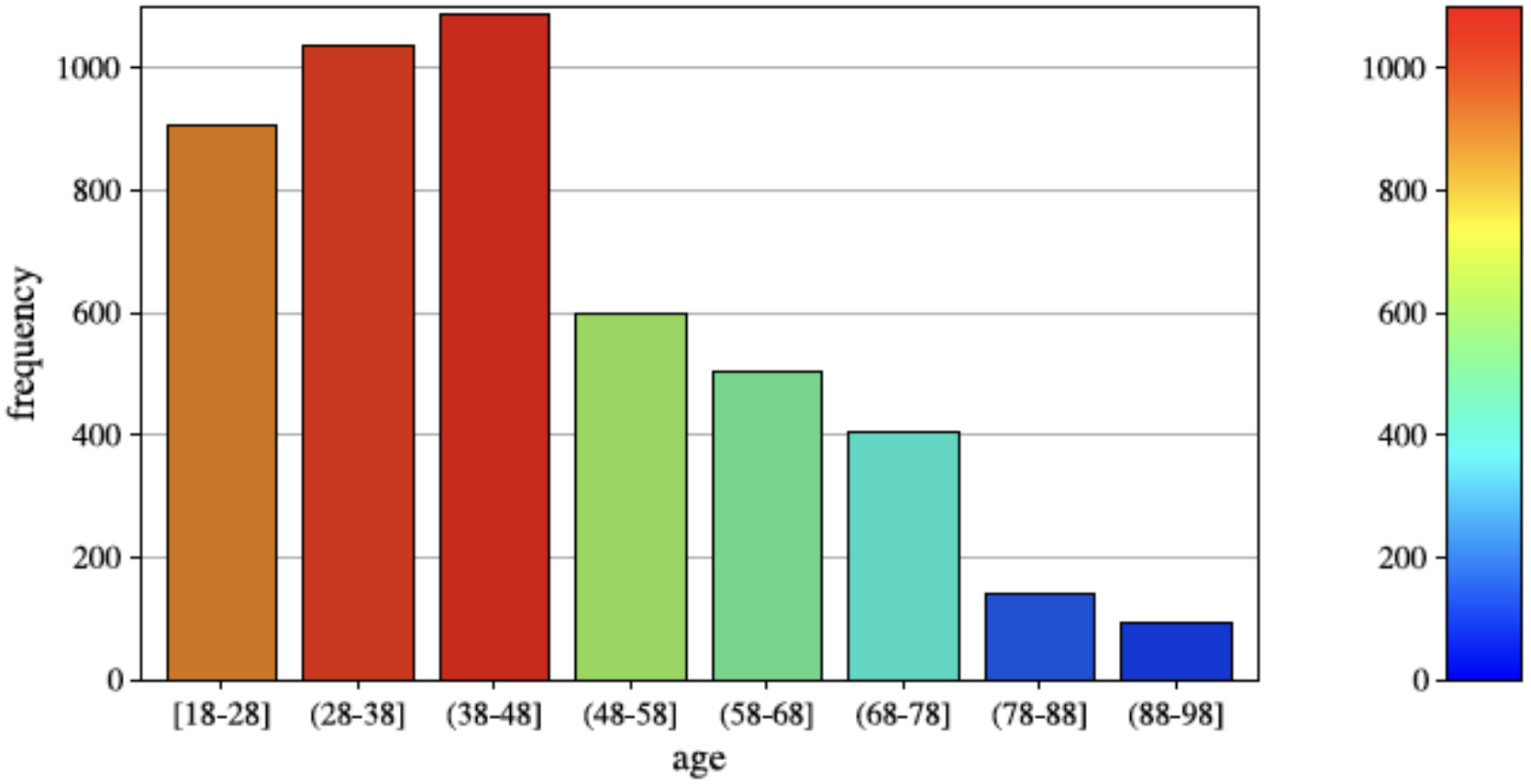
The occurrence of missing values can have different causes, e.g.,

- A sensor might be broken;
- People might have refused or forgotten to answer a question;
- Not applicable value

It is very important to identify any missing/incorrect values!
Otherwise, the further analysis can be completely mislead by such erroneous values.

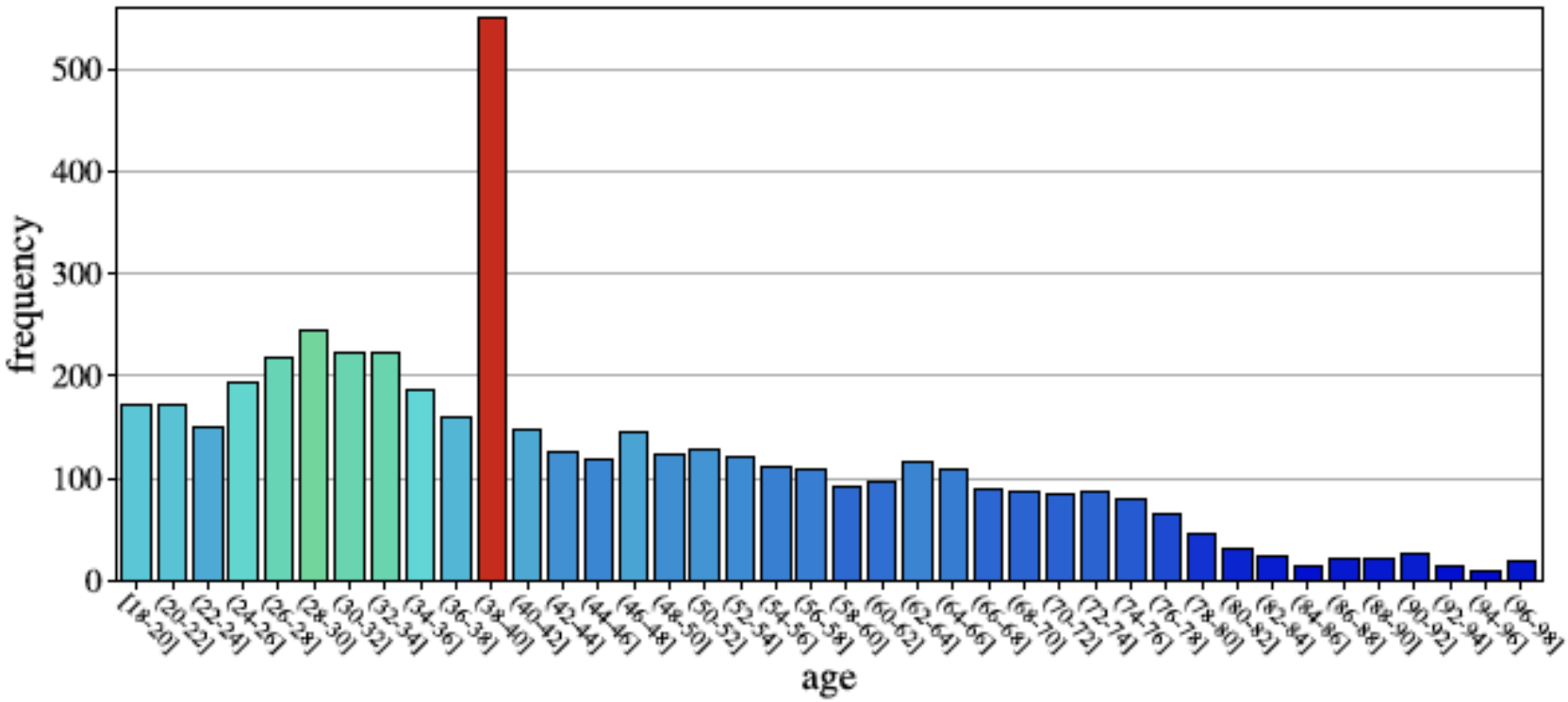
Missing values: example

Is there anything wrong with this data?



Missing values: example

And now? Is there anything wrong with this data?



Agenda

1. Recap
2. Project understanding
3. Data understanding
    - Missing values
4. Data preparation
5. Modelling
    - Under-/overfitting
6. Evaluation
    - Prediction accuracy versus model interpretability

Data preparation steps

1. Select data
2. Clean data
   - treating missing data,
   - identifying erroneous data and outliers, and
   - defining the appropriate way to represent variables
3. Construct data
   - Split into training/test sets if needed
   - Consider taking a smaller sample (when dealing with big datasets)
   - Transform variables
     - categorical variables $\Rightarrow$ dummies);
     - normalise/standardise variables
     - non-linear transformations

Agenda

1. Recap
2. Project understanding
3. Data understanding
   - Missing values
4. Data preparation
5. Modelling
   - Under-/overfitting
6. Evaluation
   - Prediction accuracy versus model interpretability

The bias-variance tradeoff

A key problem in modelling is to choose a model that accurately captures both:

- in-sample structure
- behaviour of unseen data (generalise/predict)

We want a model to neither **underfit** (high bias) nor **overfit** (high variance)

- it applies to both regression and forecasting,
- but also to data mining techniques (clustering and classification) that we will cover next week

The bias-variance tradeoff

To visualise this tradeoff, think of a dart board:

Low bias
Low variance

Low bias
High variance

Appropriate model structure, but inaccurate.

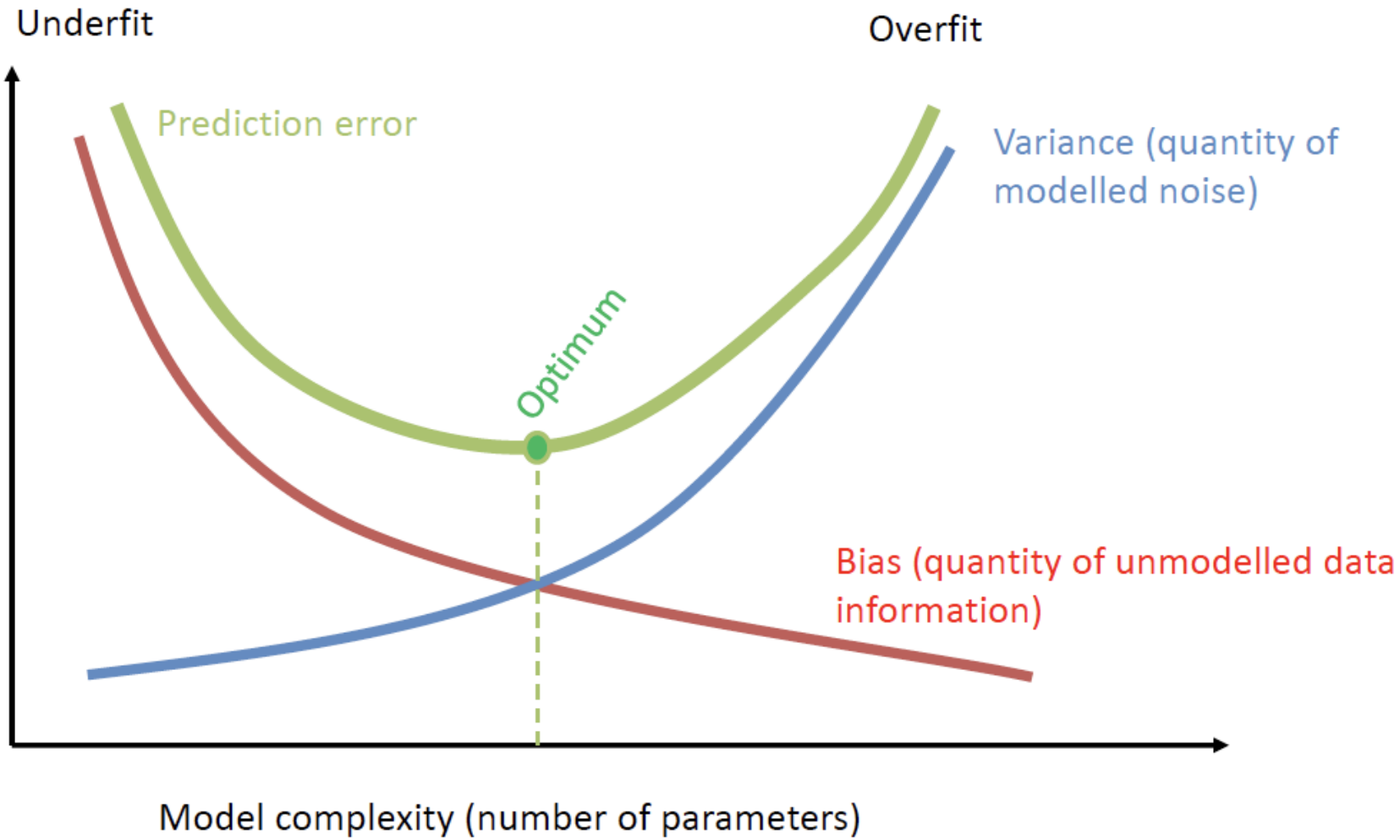
High bias
Low variance

Systematic model error.

High bias
High variance

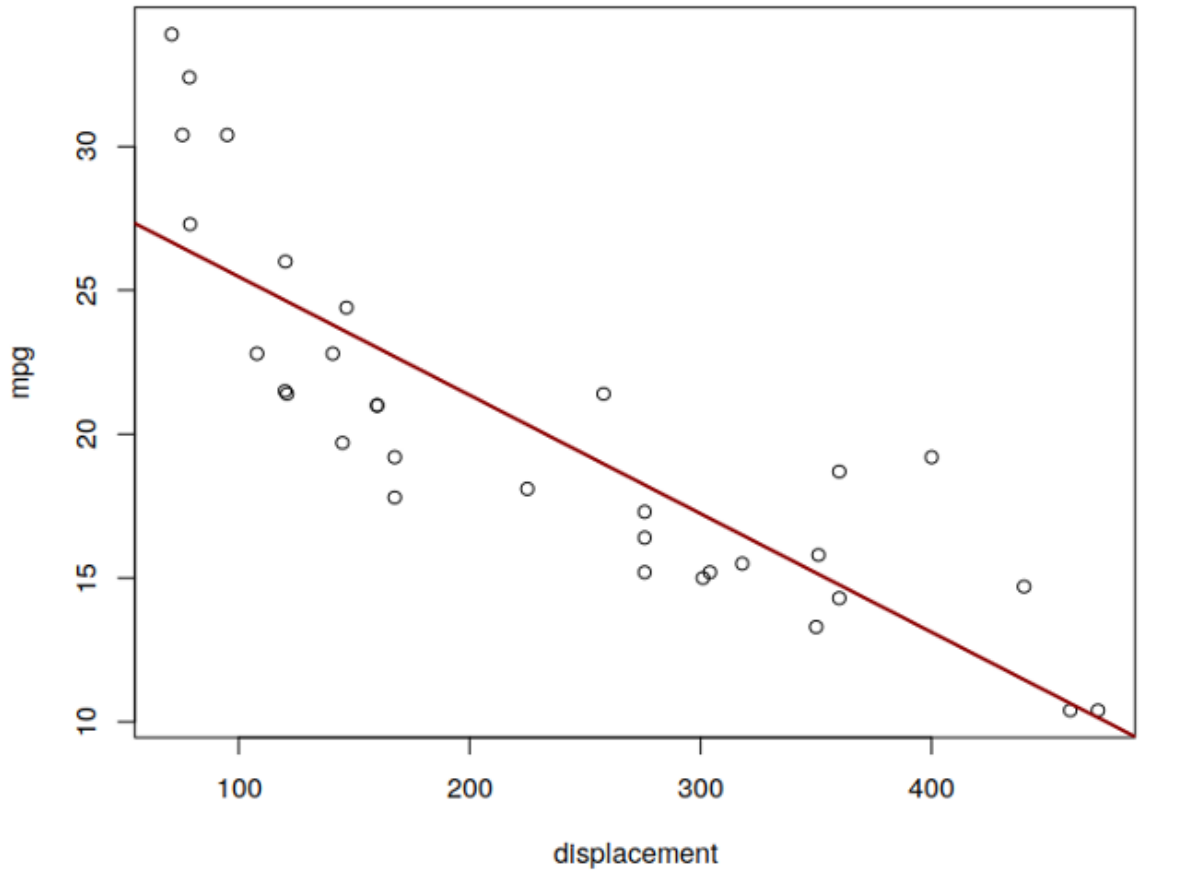
Systematic model error & inaccurate.

The bias-variance tradeoff



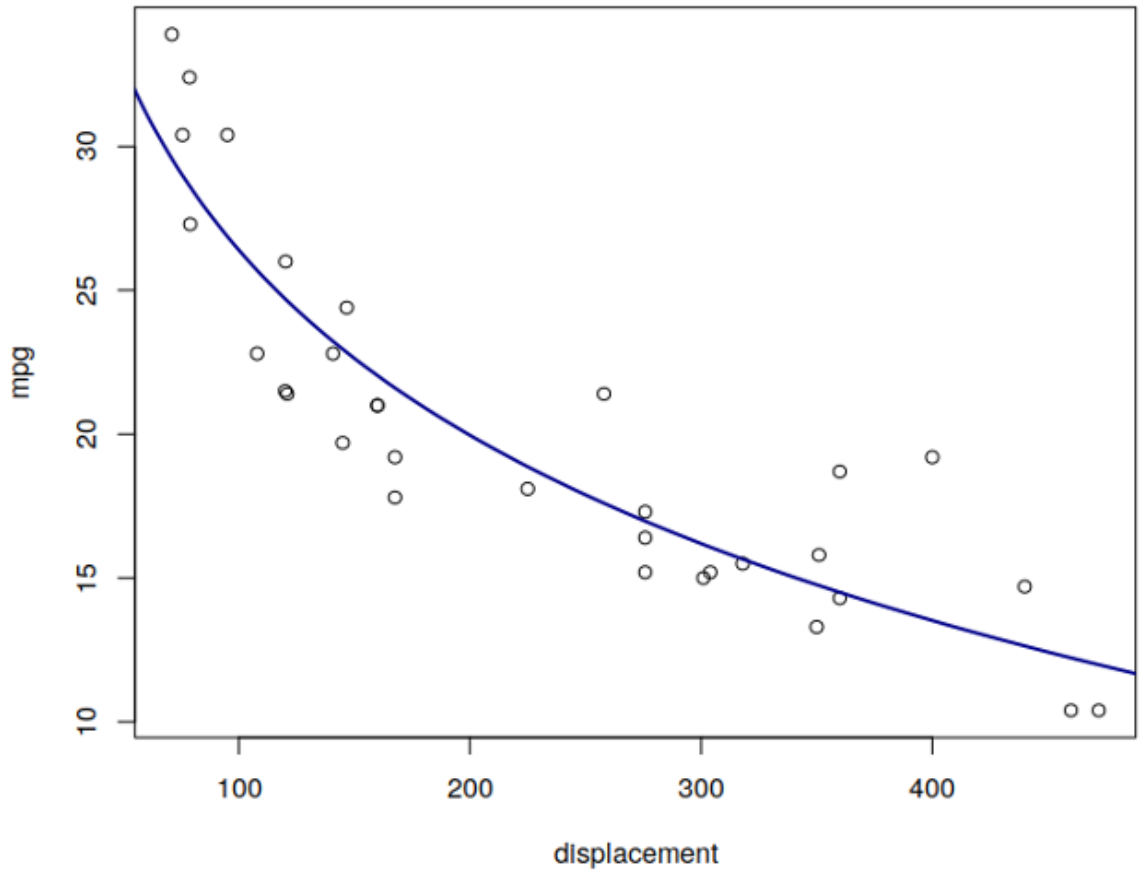
Under/overfitting: example

Let's consider this example:



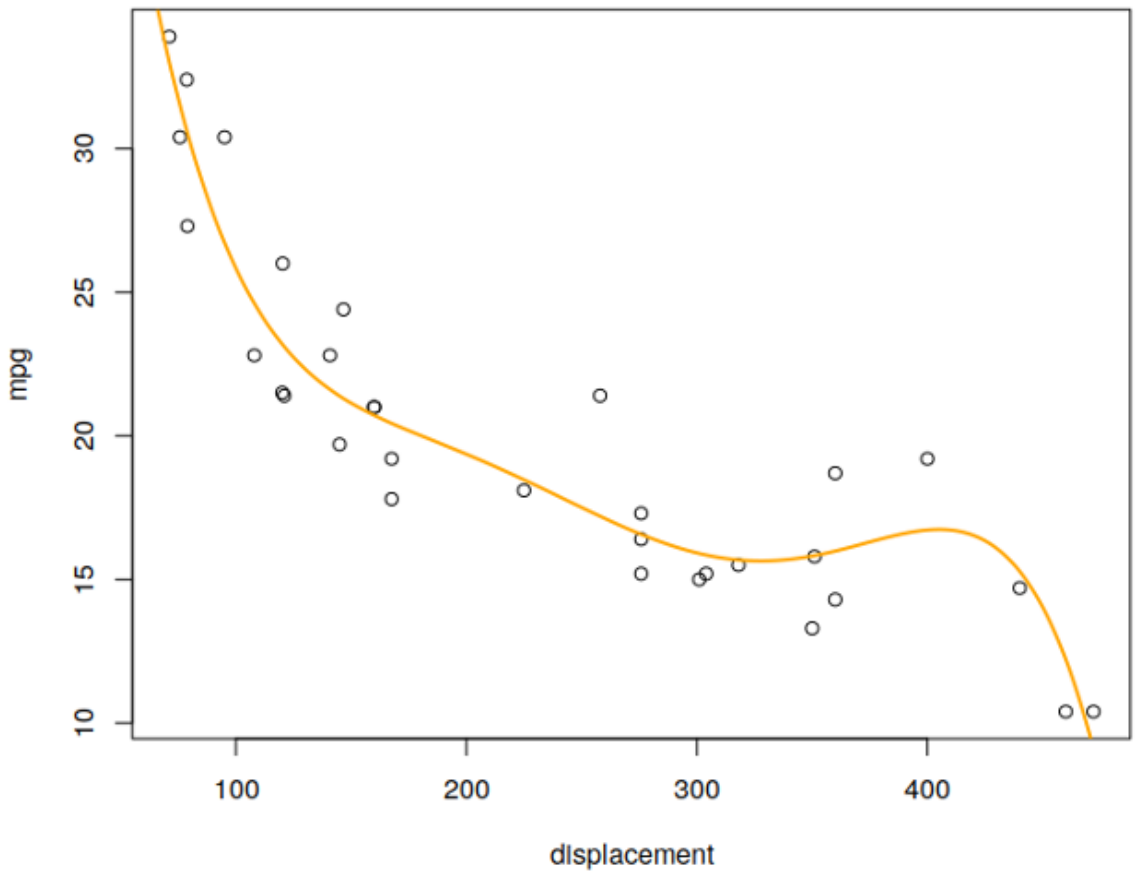
Under/over fitting: example

What do you think about this model?



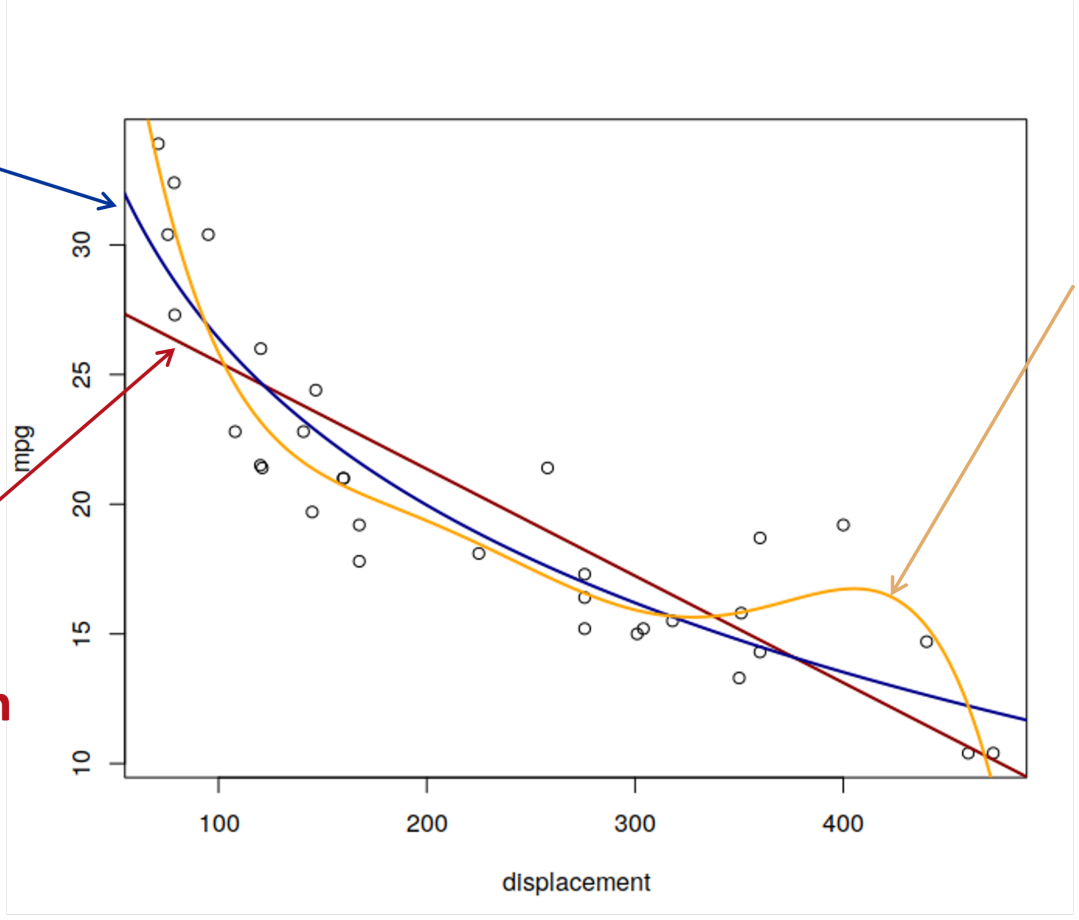
Under/over fitting: example

And now? What do you think about this model?

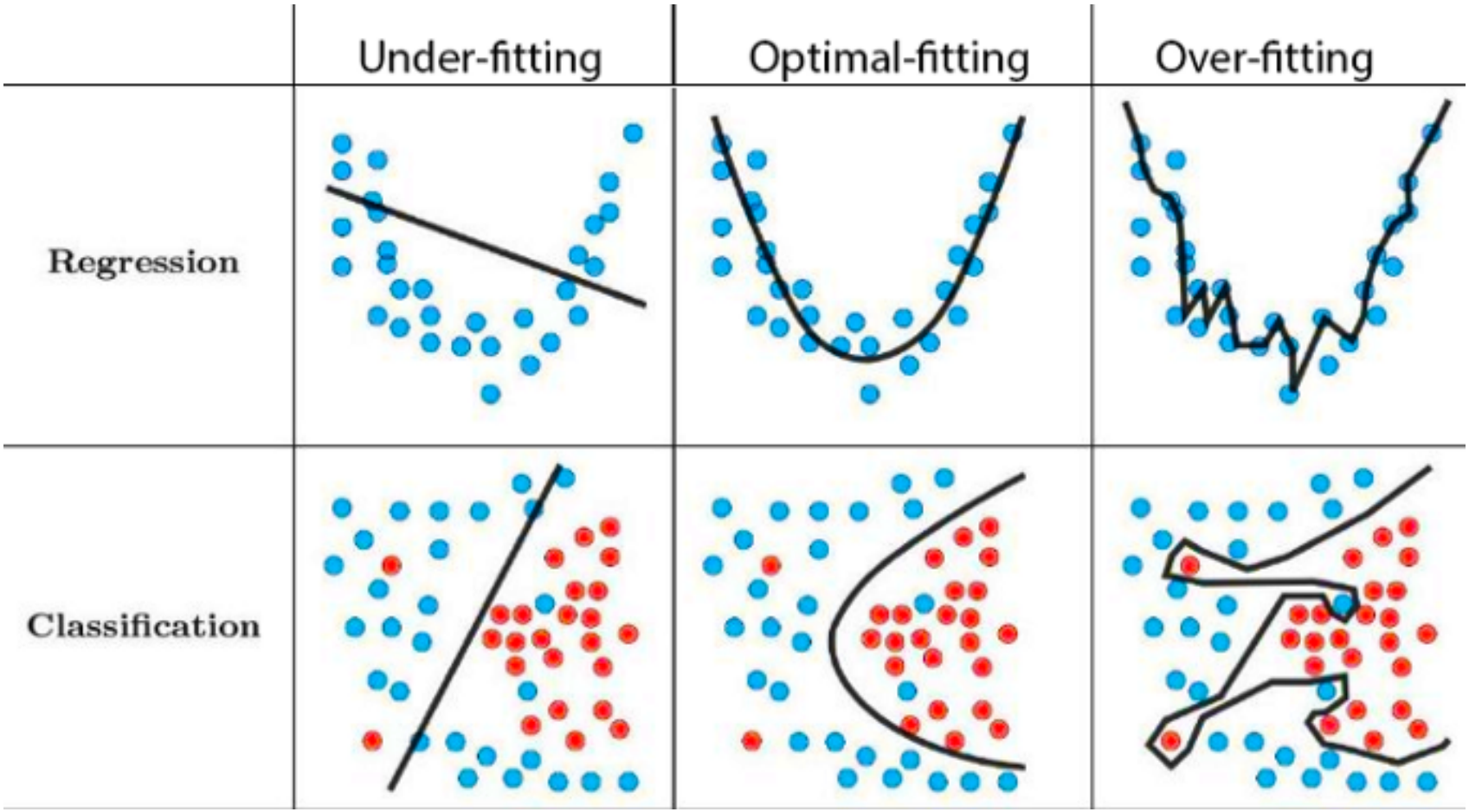


Under/over fitting: example

In this case:



Under/over fitting: regression & classification



Agenda

1. Recap
2. Project understanding
3. Data understanding
   - Missing values
4. Data preparation
5. Modelling
   - Under-/overfitting
6. Evaluation
   - Prediction accuracy versus model interpretability

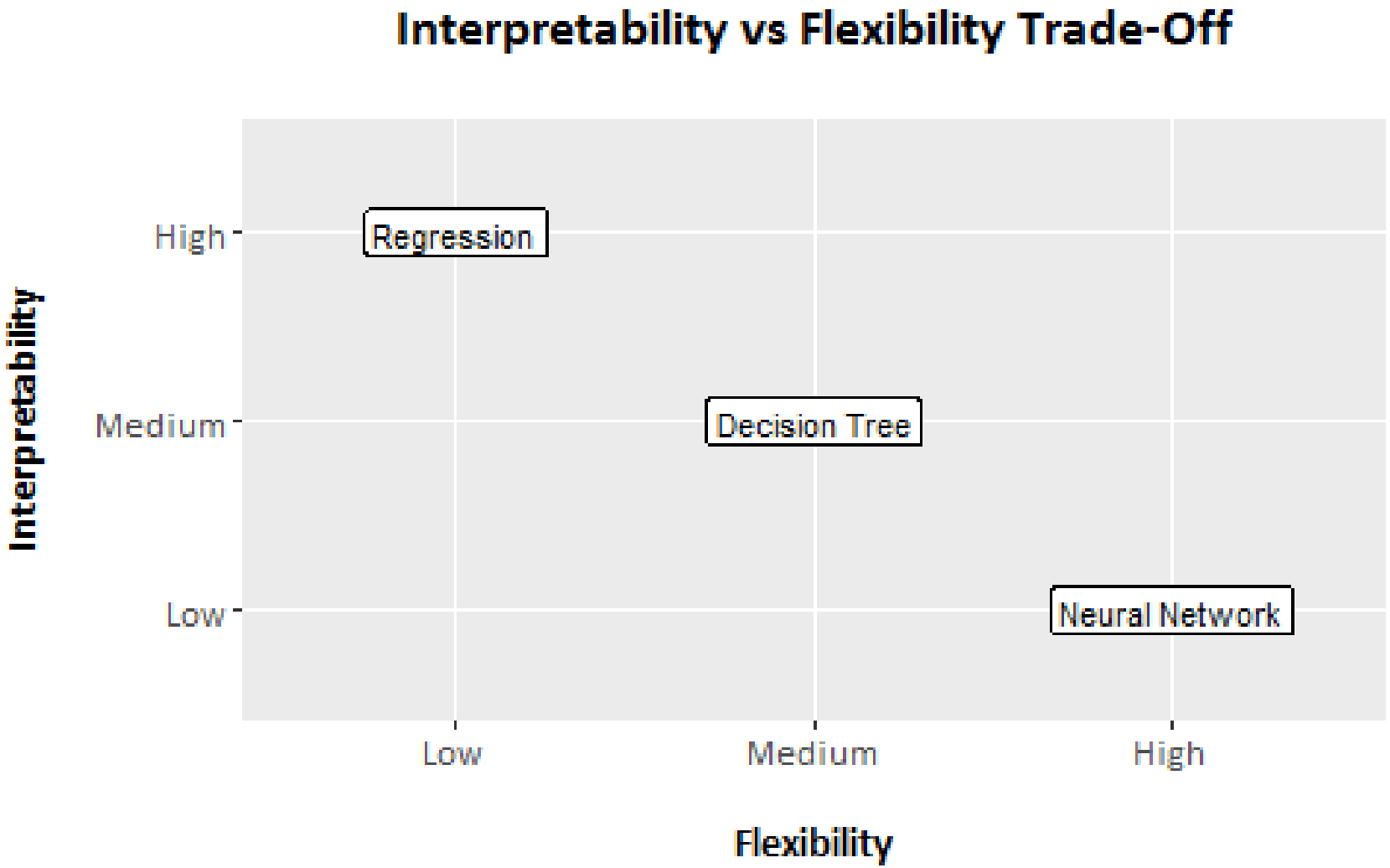
The tradeoff between prediction accuracy and model interpretability

Model interpretability

A model is better interpretable than another model if its decisions are easier for a human to comprehend than decisions from the other model.

- For example, linear regression is a relatively interpretable method, but also quire inflexible (restrictive), because it can only generate linear functions.
- While many data mining approaches are very flexible, but are difficult to understand (e.g., how any individual predictor is associated with the response), essentially they are **black boxes**.
- **Occam's razor principle**: when faced with several methods that give roughly equivalent performance, pick the simplest.

Model flexibility vs interpretability



Wrap up

Today we:

- Looked at modelling in practice
- Some pitfalls of business analytics

Next time:

- Data visualisation & dashboards