Introduction

Bojan Božić

TU Dublin

Summer 2021

What is Predictive Data Analytics?

- Predictive Data Analytics encompasses the business and data processes and computational models that enable a business to make **data-driven decisions**.

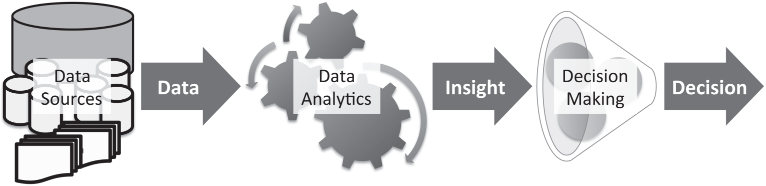


Figure: Predictive data analytics moving from **data** to **insights** to **decisions**.

Example Applications:

- Price Prediction
- Fraud Detection
- Dosage Prediction
- Risk Assessment
- Propensity modelling
- Diagnosis
- Document Classification
- . . .

What is Machine Learning?

- (Supervised) Machine Learning techniques automatically learn a model of the relationship between a set of **descriptive features** and a **target feature** from a set of historical examples.

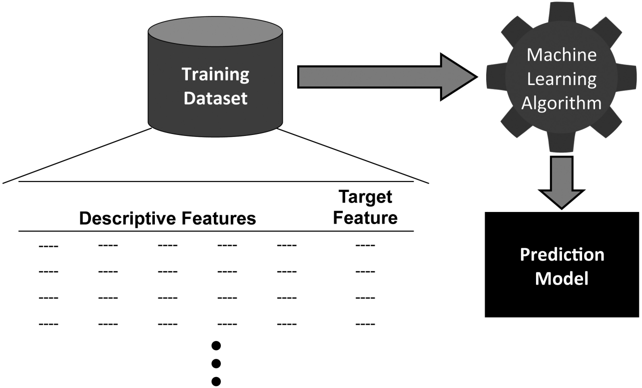


Figure: Using machine learning to induce a prediction model from a training dataset.

Figure: Using the model to make predictions for new query instances.

ID	OCCUPATION	AGE	LOAN-SALARY RATIO	OUTCOME
1	industrial	34	2.96	repaid
2	professional	41	4.64	default
3	professional	36	3.22	default
4	professional	41	3.11	default
5	industrial	48	3.80	default
6	industrial	61	2.52	repaid
7	professional	37	1.50	repaid
8	professional	40	1.93	repaid
9	industrial	33	5.25	default
10	industrial	32	4.15	default

- What is the relationship between the **descriptive features** (OCCUPATION, AGE, LOAN-SALARY RATIO) and the **target feature** (OUTCOME)?

if LOAN-SALARY RATIO > 3 **then**
  OUTCOME=*'default'*
else
  OUTCOME=*'repay'*
end if

if LOAN-SALARY RATIO > 3 **then**
  OUTCOME=*'default'*
else
  OUTCOME=*'repay'*
end if

- This is an example of a **prediction model**

if LOAN-SALARY RATIO > 3 **then**
  OUTCOME=*'default'*
else
  OUTCOME=*'repay'*
end if

- This is an example of a **prediction model**
- This is also an example of a **consistent** prediction model

if LOAN-SALARY RATIO > 3 **then**
  OUTCOME=*'default'*
else
  OUTCOME=*'repay'*
end if

- This is an example of a **prediction model**
- This is also an example of a **consistent** prediction model
- Notice that this model does not use all the features and the feature that it uses is a derived feature (in this case a ratio): **feature design** and **feature selection** are two important topics that we will return to again and again.

- What is the relationship between the **descriptive features** and the **target feature** (OUTCOME) in the following dataset?

ID	Amount	Salary	Loan-Salary Ratio	Age	Occupation	House	Type	Outcome
1	245,100	66,400	3.69	44	industrial	farm	stb	repaid
2	90,600	75,300	1.2	41	industrial	farm	stb	repaid
3	195,600	52,100	3.75	37	industrial	farm	ftb	default
4	157,800	67,600	2.33	44	industrial	apartment	ftb	repaid
5	150,800	35,800	4.21	39	professional	apartment	stb	default
6	133,000	45,300	2.94	29	industrial	farm	ftb	default
7	193,100	73,200	2.64	38	professional	house	ftb	repaid
8	215,000	77,600	2.77	17	professional	farm	ftb	repaid
9	83,000	62,500	1.33	30	professional	house	ftb	repaid
10	186,100	49,200	3.78	30	industrial	house	ftb	default
11	161,500	53,300	3.03	28	professional	apartment	stb	repaid
12	157,400	63,900	2.46	30	professional	farm	stb	repaid
13	210,000	54,200	3.87	43	professional	apartment	ftb	repaid
14	209,700	53,000	3.96	39	industrial	farm	ftb	default
15	143,200	65,300	2.19	32	industrial	apartment	ftb	default
16	203,000	64,400	3.15	44	industrial	farm	ftb	repaid
17	247,800	63,800	3.88	46	industrial	house	stb	repaid
18	162,700	77,400	2.1	37	professional	house	ftb	repaid
19	213,300	61,100	3.49	21	industrial	apartment	ftb	default
20	284,100	32,300	8.8	51	industrial	farm	ftb	default
21	154,000	48,900	3.15	49	professional	house	stb	repaid
22	112,800	79,700	1.42	41	professional	house	ftb	repaid
23	252,000	59,700	4.22	27	professional	house	stb	default
24	175,200	39,900	4.39	37	professional	apartment	stb	default
25	149,700	58,600	2.55	35	industrial	farm	stb	default

if LOAN-SALARY RATIO < 1.5 **then**
  OUTCOME=*'repay'*
else if LOAN-SALARY RATIO > 4 **then**
  OUTCOME=*'default'*
else if AGE < 40 **and** OCCUPATION $=$*'industrial'* **then**
  OUTCOME=*'default'*
else
  OUTCOME=*'repay'*
end if

if LOAN-SALARY RATIO < 1.5 **then**
    OUTCOME=*'repay'*
else if LOAN-SALARY RATIO > 4 **then**
    OUTCOME=*'default'*
else if AGE < 40 **and** OCCUPATION $=$*'industrial'* **then**
    OUTCOME=*'default'*
else
    OUTCOME=*'repay'*
end if

- The real value of machine learning becomes apparent in situations like this when we want to build prediction models from large datasets with multiple features.

How Does Machine Learning Work?

- Machine learning algorithms work by searching through a set of possible prediction models for the model that best captures the relationship between the descriptive features and the target feature.

- Machine learning algorithms work by searching through a set of possible prediction models for the model that best captures the relationship between the descriptive features and the target feature.
- An obvious search criteria to drive this search is to look for models that are consistent with the data.

- Machine learning algorithms work by searching through a set of possible prediction models for the model that best captures the relationship between the descriptive features and the target feature.
- An obvious search criteria to drive this search is to look for models that are consistent with the data.
- However, because a training dataset is only a sample ML is an ill-posed problem.

Table: A simple retail dataset

ID	BBY	ALC	ORG	GRP
1	no	no	no	couple
2	yes	no	yes	family
3	yes	yes	no	family
4	no	no	yes	couple
5	no	yes	yes	single

Table: A full set of potential prediction models before any training data becomes available.

BBY	ALC	ORG	GRP	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	$\mathbb{M}_4$	$\mathbb{M}_5$	...	$\mathbb{M}_{6\,561}$
no	no	no	?	couple	couple	single	couple	couple		couple
no	no	yes	?	single	couple	single	couple	couple		single
no	yes	no	?	family	family	single	single	single		family
no	yes	yes	?	single	single	single	single	single		couple
yes	no	no	?	couple	couple	family	family	family	...	family
yes	no	yes	?	couple	family	family	family	family		couple
yes	yes	no	?	single	family	family	family	family		single
yes	yes	yes	?	single	single	family	family	couple		family

Table: A sample of the models that are consistent with the training data

BBY	ALC	ORG	GRP	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	$\mathbb{M}_4$	$\mathbb{M}_5$	...	$\mathbb{M}_{6\,561}$
no	no	no	couple	couple	couple	single	couple	couple		couple
no	no	yes	couple	single	couple	single	couple	couple		single
no	yes	no	?	family	family	single	single	single		family
no	yes	yes	single	single	single	single	single	single		couple
yes	no	no	?	couple	couple	family	family	family	...	family
yes	no	yes	family	couple	family	family	family	family		couple
yes	yes	no	family	single	family	family	family	family		single
yes	yes	yes	?	single	single	family	family	couple		family

Table: A sample of the models that are consistent with the training data

BBY	ALC	ORG	GRP	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	$\mathbb{M}_4$	$\mathbb{M}_5$	...	$\mathbb{M}_{6\,561}$
no	no	no	couple	couple	couple	single	couple	couple		couple
no	no	yes	couple	single	couple	single	couple	couple		single
no	yes	no	?	family	family	single	single	single		family
no	yes	yes	single	single	single	single	single	single		couple
yes	no	no	?	couple	couple	family	family	family	...	family
yes	no	yes	family	couple	family	family	family	family		couple
yes	yes	no	family	single	family	family	family	family		single
yes	yes	yes	?	single	single	family	family	couple		family

- Notice that there is more than one candidate model left! It is because a single consistent model cannot be found based on a sample training dataset that ML is ill-posed.

- Consistency $\approx$ memorizing the dataset.
- Consistency with noise in the data isn't desirable.
- Goal: a model that **generalises** beyond the dataset and that isn't influenced by the noise in the dataset.
- So what criteria should we use for choosing between models?

- Inductive bias the set of assumptions that define the model selection criteria of an ML algorithm.
- There are two types of bias that we can use:
  1. restriction bias
  2. preference bias
- Inductive bias is necessary for learning (beyond the dataset).

How ML works (Summary)

- ML algorithms work by searching through sets of potential models.
- There are two sources of information that guide this search:
  1. the training data,
  2. the inductive bias of the algorithm.

Inductive Bias Versus Sample Bias

- Inductive bias is necessary for machine learning, and in a sense, a key goal of a data analyst is to find the correct inductive bias.
- Inductive bias is not the only type of bias that affects machine learning, a particularly important type of bias that we need to be aware of is **sampling bias**

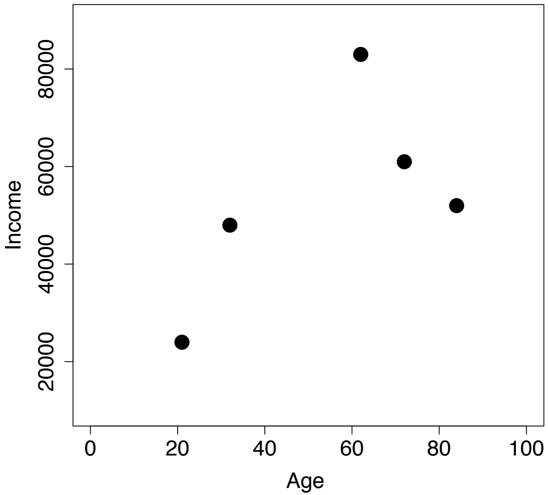
- Sampling bias arises when the sample of data used within a data-driven process is collected in such a way that the sample is not representative of the population the sample is used to represent.
- If a sample of data is not representative of a population, then inferences based on that sample will not generalize to the larger population.
- Sample bias is something that a data analyst should proactively work hard to remove from the data used in any data analytics project.

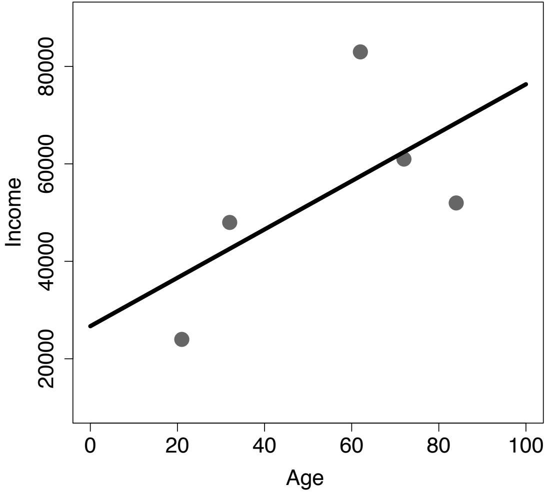
What Can Go Wrong With ML?

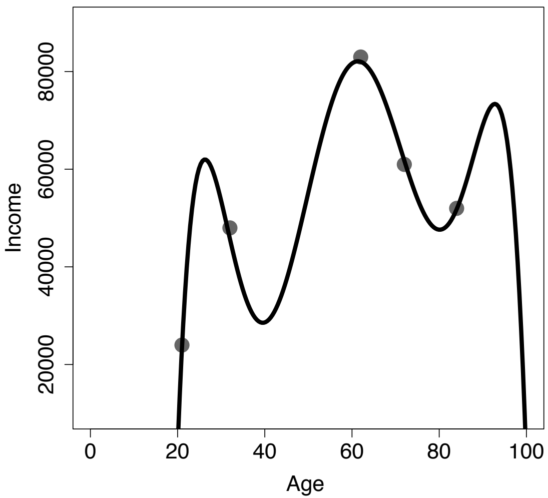
- No free lunch!
- What happens if we choose the wrong inductive bias:
  1. **underfitting**
  2. **overfitting**

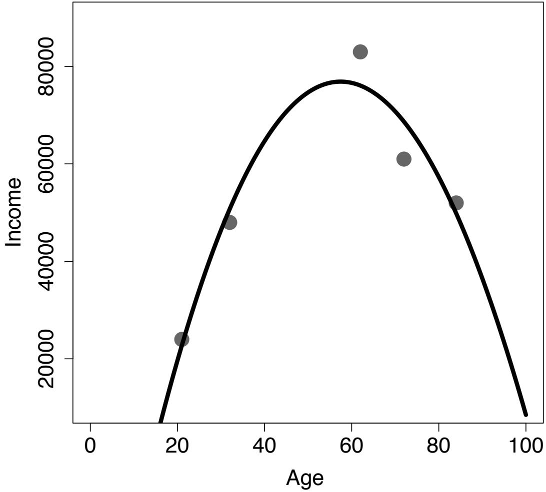
Table: The age-income dataset.

ID	AGE	INCOME
1	21	24,000
2	32	48,000
3	62	83,000
4	72	61,000
5	84	52,000

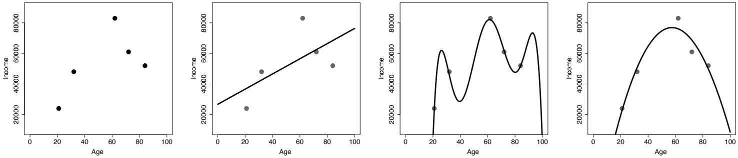

Income
80000
60000
40000
20000
0
20
40
60
80
100
Age


Income
Age
20000
40000
60000
80000
0
20
40
60
80
100


Income
Age
20000
40000
60000
80000
0
20
40
60
80
100



Income
Age
80000
60000
40000
20000
0
20
40
60
80
100



(a) Dataset (b) Underfitting (c) Overfitting (d) Just right

Figure: Striking a balance between overfitting and underfitting when trying to predict age from income.

The Predictive Data Analytics Project Lifecycle: Crisp-DM

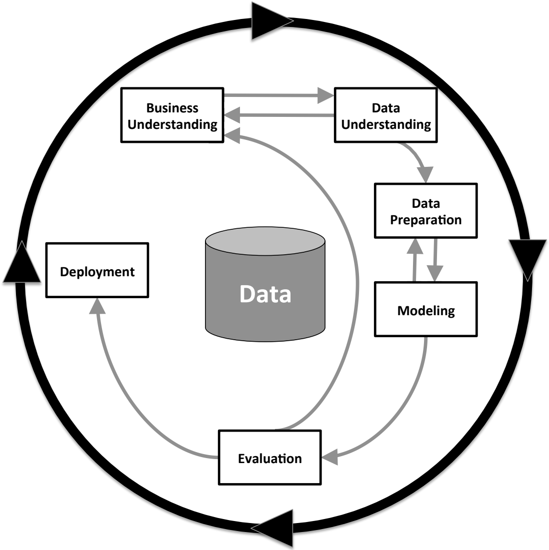


Figure: A diagram of the CRISP-DM process which shows the six key phases and indicates the important relationships between them. This figure is based on Figure 2 of [1].

The Road Ahead

- Part 1 of the course will cover the preparatory activity prior to model building.
  1. Business Understanding
  2. Data Understanding
  3. Data Preparation

- Part II focuses on five families of machine learning algorithms for predictive data analytics:
  1. **Information based learning**
  2. **Similarity based learning**
  3. **Probability based learning**
  4. **Error based learning**
  5. **Deep Learning**
- We also cover a range of approaches to evaluating prediction models.

- Part III also deals with modelling but looks at modelling approaches beyond prediction
    1. **Unsupervised Learning**
    2. **Reinforcement learning**
- Par IV covers questions relating deployment and includes case studies that illustrate how everything described in the preceding sections come together in a successful predictive data analytics project.

Summary

- Machine Learning techniques automatically learn the relationship between a set of **descriptive features** and a **target feature** from a set of historical examples.
- Machine Learning is an **ill-posed** problem:
  1. **generalize**,
  2. **inductive bias**,
  3. **underfitting**,
  4. **overfitting**.
- Striking the right balance between model complexity and simplicity (between underfitting and overfitting) is the hardest part of machine learning.

[1] R. Wirth and J. Hipp.
Crisp-dm: Towards a standard process model for data mining.
In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, pages 29–39. Citeseer, 2000.