Interpretable Machine Learning

The Basics

Hendra Bunyamin

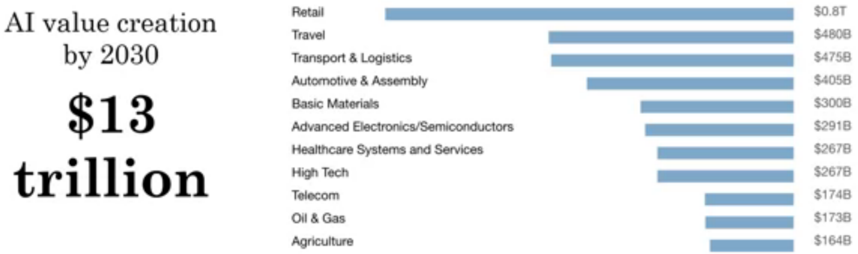
Informatics Engineering
Faculty of Information Technology
Maranatha Christian University

9 April 2021
NUNI IT Online Seminar

Outline

Outline

Introduction



Source: McKinsey Global Institute (**?**)

Introduction

Source: McKinsey Global Institute (**?**)

$13 trillion = $13 $\times$ 10^{12} $=$ Rp183.000.000.000.000.000, $-$

Demystifying AI

Artificial Intelligence or **AI** can be divided into 2 as follows (**?**):

Artificial Intelligence or **AI** can be divided into 2 as follows (**?**):

- **ANI** $\Rightarrow$ Artificial Narrow Intelligence.

Demystifying AI

Artificial Intelligence or **AI** can be divided into 2 as follows (**?**):

- **ANI** $\Rightarrow$ Artificial Narrow Intelligence.
  Examples: smart speaker, self-driving car, web search, AI in farming and factories.

Demystifying AI

Artificial Intelligence or **AI** can be divided into 2 as follows (**?**):

- **ANI** $\Rightarrow$ Artificial Narrow Intelligence.
  Examples: smart speaker, self-driving car, web search, AI in farming and factories.
- **AGI** $\Rightarrow$ Artificial General Intelligence.

Demystifying AI

Artificial Intelligence or **AI** can be divided into 2 as follows (**?**):

- **ANI** $\Rightarrow$ Artificial Narrow Intelligence.
  Examples: smart speaker, self-driving car, web search, AI in farming and factories.

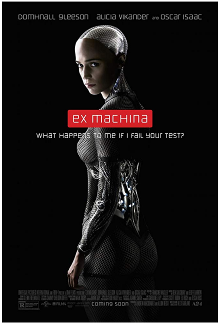
- **AGI** $\Rightarrow$ Artificial General Intelligence.
  Examples: Do anything a human can do or even more than human capability.

Demystifying AI

Artificial Intelligence or **AI** can be divided into 2 as follows (**?**):

- **ANI** $\Rightarrow$ Artificial Narrow Intelligence.
  Examples: smart speaker, self-driving car, web search, AI in farming and factories.

- **AGI** $\Rightarrow$ Artificial General Intelligence.
  Examples: Do anything a human can do or even more than human capability.

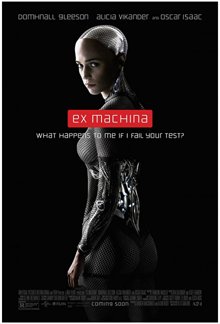


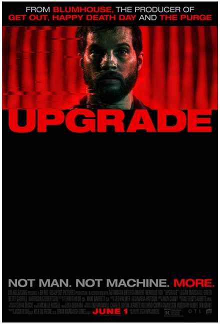
Demystifying AI

Artificial Intelligence or **AI** can be divided into 2 as follows (**?**):

- **ANI** $\Rightarrow$ Artificial Narrow Intelligence.
  Examples: smart speaker, self-driving car, web search, AI in farming and factories.

- **AGI** $\Rightarrow$ Artificial General Intelligence.
  Examples: Do anything a human can do or even more than human capability.

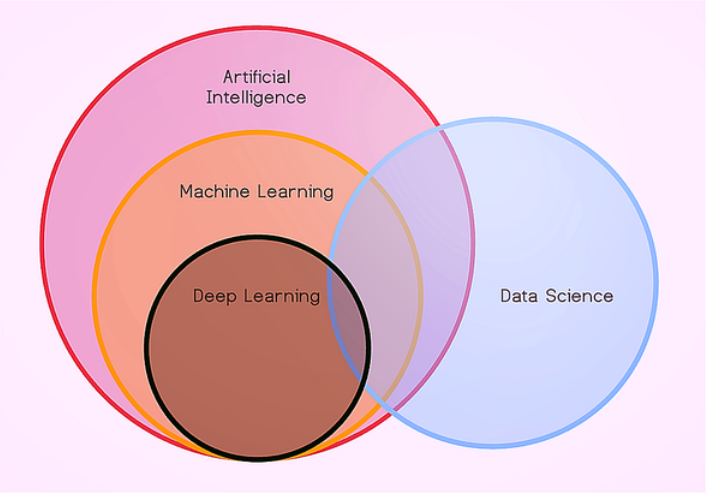




Outline

Diagram Venn tentang AI, ML, DL, Data Science

Diagram Venn tentang AI, ML, DL, Data Science



Relationship among AI, ML, DL, and DS (**?**)

Machine Learning

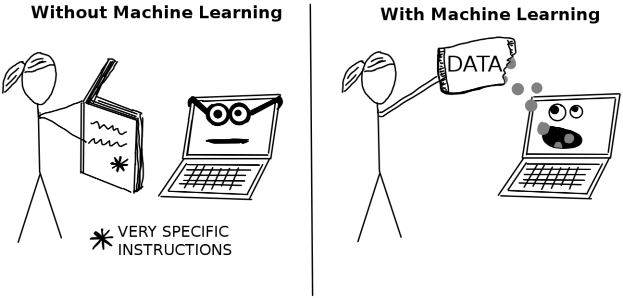
Machine Learning

- One of the tools that drive the significant progress of AI is **Machine Learning** (ML).

Machine Learning

- One of the tools that drive the significant progress of AI is **Machine Learning** (ML).
- **Machine Learning** is a set of methods that allow computers to *learn from data* to *make and improve predictions*, e.g., cancer, weekly sales, credit default.

- One of the tools that drive the significant progress of AI is **Machine Learning** (ML).
- **Machine Learning** is a set of methods that allow computers to *learn from data* to *make and improve predictions*, e.g., cancer, weekly sales, credit default.



A paradigm shift from "normal programming" to "indirect programming"

Machine Learning: Supervised Learning (1/2)

Machine Learning: Supervised Learning (1/2)

- A common type of of **Machine Learning** is a type of AI that learns from X to Y or is often called ***Supervised Learning***.

$$\underset{\text{input}}{X} \longrightarrow \underset{\text{output}}{Y}$$

Machine Learning: Supervised Learning (1/2)

- A common type of of **Machine Learning** is a type of AI that learns from X to Y or is often called ***Supervised Learning***.

$$\underset{\text{input}}{X} \longrightarrow \underset{\text{output}}{Y}$$

- A **Machine Learning Model** is *the learned program* that maps inputs into outputs/predictions.

Machine Learning: Supervised Learning (1/2)

- A common type of of **Machine Learning** is a type of AI that learns from X to Y or is often called ***Supervised Learning***.

$$\underset{\text{input}}{X} \longrightarrow \underset{\text{output}}{Y}$$

- A **Machine Learning Model** is *the learned program* that maps inputs into outputs/predictions.
- A **Machine Learning Algorithm** is *the program* used to learn a ML model from data (**?**).

Machine Learning: Supervised Learning (2/2)

Input (X)	Output (Y)	ML Model

Machine Learning: Supervised Learning (2/2)

Input (X)		Output (Y)	ML Model
email	$\longrightarrow$	spam? (0/1)	spam filtering

Machine Learning: Supervised Learning (2/2)

Input (X)		Output (Y)	ML Model
email	$\longrightarrow$	spam? (0/1)	spam filtering
audio	$\longrightarrow$	text transcript	speech recognition

Machine Learning: Supervised Learning (2/2)

Input (X)		Output (Y)	ML Model
email	$\longrightarrow$	spam? (0/1)	spam filtering
audio	$\longrightarrow$	text transcript	speech recognition
English	$\longrightarrow$	Indonesia	machine translation

Machine Learning: Supervised Learning (2/2)

Input (X)		Output (Y)	ML Model
email	$\longrightarrow$	spam? (0/1)	spam filtering
audio	$\longrightarrow$	text transcript	speech recognition
English	$\longrightarrow$	Indonesia	machine translation
ad, user info	$\longrightarrow$	click? (0/1)	online advertising

Machine Learning: Supervised Learning (2/2)

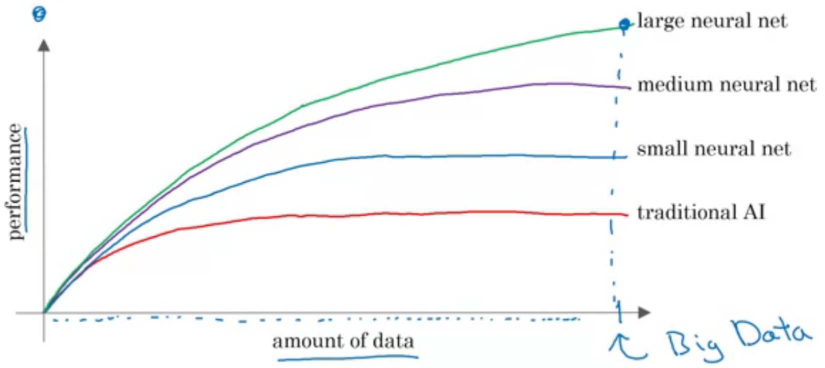
Input (X)		Output (Y)	ML Model
email	$\longrightarrow$	spam? (0/1)	spam filtering
audio	$\longrightarrow$	text transcript	speech recognition
English	$\longrightarrow$	Indonesia	machine translation
ad, user info	$\longrightarrow$	click? (0/1)	online advertising
image, radar info	$\longrightarrow$	position of other cars	self-driving car

Machine Learning: Supervised Learning (2/2)

Input (X)		**Output (Y)**	**ML Model**
email	$\longrightarrow$	spam? (0/1)	spam filtering
audio	$\longrightarrow$	text transcript	speech recognition
English	$\longrightarrow$	Indonesia	machine translation
ad, user info	$\longrightarrow$	click? (0/1)	online advertising
image, radar info	$\longrightarrow$	position of other cars	self-driving car
image of phone	$\longrightarrow$	defect? (0/1)	visual inspection

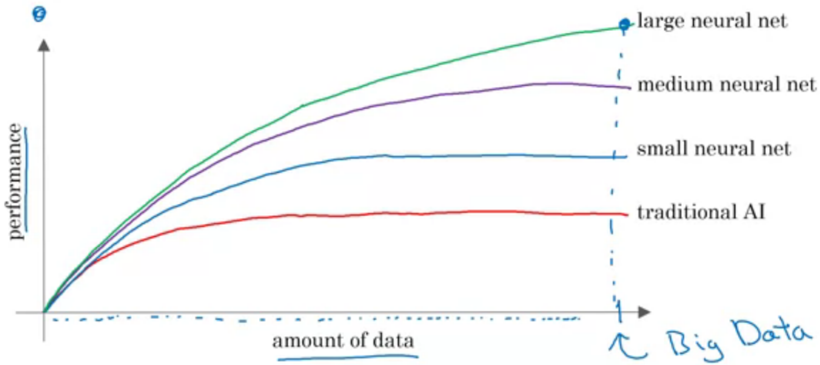
Why Now?

Why Now?



large neural net
medium neural net
small neural net
traditional AI
performance
amount of data
Big Data

Why Now?



Large neural net + Big Data = High Performance (**?**)

Common Machine Learning Components

Typically, a Machine Learning Algorithm consists of

Common Machine Learning Components

Typically, a Machine Learning Algorithm consists of

1. Data $\rightarrow X = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ dan $Y = \{y^{(1)}, y^{(2)}, \ldots, y^{(m)}\}$

Common Machine Learning Components

Typically, a Machine Learning Algorithm consists of

1. Data $\rightarrow X = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ dan $Y = \{y^{(1)}, y^{(2)}, \ldots, y^{(m)}\}$
2. Model/Hypothesis $\rightarrow h$

Common Machine Learning Components

Typically, a Machine Learning Algorithm consists of

1. Data $\rightarrow X = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ dan $Y = \{y^{(1)}, y^{(2)}, \ldots, y^{(m)}\}$
2. Model/Hypothesis $\rightarrow h$
3. Cost/Loss Function $\rightarrow J$

Common Machine Learning Components

Typically, a Machine Learning Algorithm consists of

1. Data $\rightarrow X = \left\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\right\}$ dan $Y = \left\{y^{(1)}, y^{(2)}, \ldots, y^{(m)}\right\}$
2. Model/Hypothesis $\rightarrow h$
3. Cost/Loss Function $\rightarrow J$
4. Optimization Algorithm $\rightarrow$ *gradient descent*

Common Machine Learning Components

Typically, a Machine Learning Algorithm consists of

1. Data $\rightarrow X = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ dan $Y = \{y^{(1)}, y^{(2)}, \ldots, y^{(m)}\}$
2. Model/Hypothesis $\rightarrow h$
3. Cost/Loss Function $\rightarrow J$
4. Optimization Algorithm $\rightarrow$ *gradient descent*

Let's walk through all these components in a concrete example!

Outline

Predicting Profits for Food Trucks

Predicting Profits for Food Trucks

Suppose you are the *CEO of a restaurant franchise* and are considering different cities for opening a new outlet.

Predicting Profits for Food Trucks

Suppose you are the *CEO of a restaurant franchise* and are considering different cities for opening a new outlet.
The franchise already *has trucks in various cities* and you have *data for profits and populations from the cities*.

Predicting Profits for Food Trucks

Suppose you are the *CEO of a restaurant franchise* and are considering different cities for opening a new outlet.
The franchise already *has trucks in various cities* and you have *data for profits and populations from the cities*.

A food truck serving chinese food(**?**)

Outline

Machine Learning Component: Data (1/2)

Population (***X***)	**Profit** (***Y***)
6.1101	17.592
5.5277	9.130
8.5186	13.662
⋮	⋮
5.4369	0.617

Population of city is in 10,000s while **Profit** is in $10,000s

Machine Learning Component: Data (1/2)

Population (X)	**Profit** (Y)
6.1101	17.592
5.5277	9.130
8.5186	13.662
$\vdots$	$\vdots$
5.4369	0.617

Population of city is in 10,000s while **Profit** is in $10,000s

We can write:

Machine Learning Component: Data (1/2)

Population (X)	Profit (Y)
6.1101	17.592
5.5277	9.130
8.5186	13.662
$\vdots$	$\vdots$
5.4369	0.617

Population of city is in 10,000s while **Profit** is in $10,000s

We can write:

$$x_1^{(1)} = 6.1101 \text{ and } y^{(1)} = 17.592$$

Machine Learning Component: Data (1/2)

Population (X)	**Profit** (Y)
6.1101	17.592
5.5277	9.130
8.5186	13.662
⋮	⋮
5.4369	0.617

Population of city is in 10,000s while **Profit** is in $10,000s

We can write:

$$x_1^{(1)} = 6.1101 \text{ and } y^{(1)} = 17.592$$
$$x_1^{(2)} = 5.5277 \text{ and } y^{(2)} = 9.130$$

Machine Learning Component: Data (1/2)

Population (X)	Profit (Y)
6.1101	17.592
5.5277	9.130
8.5186	13.662
$\vdots$	$\vdots$
5.4369	0.617

Population of city is in 10,000s while **Profit** is in $10,000s

We can write:

$$
\begin{aligned}
x_1^{(1)} &= 6.1101 \text{ and } y^{(1)} = 17.592 \\
x_1^{(2)} &= 5.5277 \text{ and } y^{(2)} = 9.130 \\
x_1^{(3)} &= 8.5186 \text{ and } y^{(3)} = 13.662, \text{ and}
\end{aligned}
$$

Machine Learning Component: Data (1/2)

Population (X)	Profit (Y)
6.1101	17.592
5.5277	9.130
8.5186	13.662
⋮	⋮
5.4369	0.617

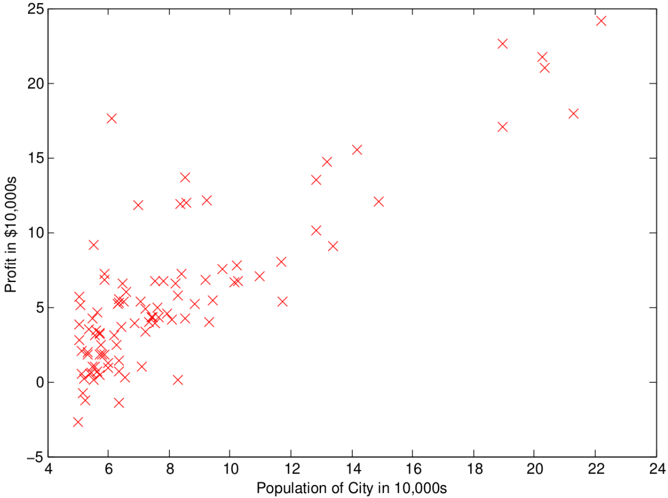
Population of city is in 10,000s while **Profit** is in $10,000s

We can write:

$$x_1^{(1)} = 6.1101 \text{ and } y^{(1)} = 17.592$$
$$x_1^{(2)} = 5.5277 \text{ and } y^{(2)} = 9.130$$
$$x_1^{(3)} = 8.5186 \text{ and } y^{(3)} = 13.662, \text{ and}$$
$$x_1^{(97)} = 5.4369 \text{ and } y^{(97)} = 0.617$$

Machine Learning Component: Data (2/2)

Machine Learning Component: Data (2/2)



Scatter plot of food trucks data

Outline

Machine Learning Component: Model (1/3)

We utilize a linear regression model as our model/hypothesis:

Machine Learning Component: Model (1/3)

We utilize a linear regression model as our model/hypothesis:

$$h_\theta(x) = \theta_0 + \theta_1 x_1$$

Machine Learning Component: Model (1/3)

We utilize a linear regression model as our model/hypothesis:

$$
\begin{aligned}
h_\theta(x) &= \theta_0 + \theta_1 x_1 \\
&= \begin{bmatrix} 1 & x_1 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}
\end{aligned}
$$

Machine Learning Component: Model (1/3)

We utilize a linear regression model as our model/hypothesis:

$$\begin{aligned} h_\theta(x) &= \theta_0 + \theta_1 x_1 \\ &= \begin{bmatrix} 1 & x_1 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} \\ &= x\theta \end{aligned}$$

Machine Learning Component: Model (1/3)

We utilize a linear regression model as our model/hypothesis:

$$\begin{aligned} h_\theta(x) &= \theta_0 + \theta_1 x_1 \\ &= \begin{bmatrix} 1 & x_1 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} \\ &= x\theta \end{aligned}$$

Since we have 97 instances, we write X as

Machine Learning Component: Model (1/3)

We utilize a linear regression model as our model/hypothesis:

$$\begin{aligned} h_\theta(x) &= \theta_0 + \theta_1 x_1 \\ &= \begin{bmatrix} 1 & x_1 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} \\ &= x\theta \end{aligned}$$

Since we have 97 instances, we write X as

$$X = \begin{bmatrix} 1 & x_1^{(1)} \\ 1 & x_1^{(2)} \\ \vdots & \vdots \\ 1 & x_1^{(97)} \end{bmatrix},$$

We utilize a linear regression model as our model/hypothesis:

$$
\begin{aligned}
h_\theta(x) &= \theta_0 + \theta_1 x_1 \\
&= \begin{bmatrix} 1 & x_1 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} \\
&= x\theta
\end{aligned}
$$

Since we have 97 instances, we write X as

$$
X = \begin{bmatrix} 1 & x_1^{(1)} \\ 1 & x_1^{(2)} \\ \vdots & \vdots \\ 1 & x_1^{(97)} \end{bmatrix},
$$

X is also called a *design matrix*.

Machine Learning Component: Model (2/3)

therefore, we can construct our model for each instance as follows:

Machine Learning Component: Model (2/3)

therefore, we can construct our model for each instance as follows:

$$\begin{bmatrix} h_\theta(x^{(1)}) \\ h_\theta(x^{(2)}) \\ \vdots \\ h_\theta(x^{(97)}) \end{bmatrix} =$$

therefore, we can construct our model for each instance as follows:

$$\begin{bmatrix} h_\theta(x^{(1)}) \\ h_\theta(x^{(2)}) \\ \vdots \\ h_\theta(x^{(97)}) \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} \\ 1 & x_1^{(2)} \\ \vdots & \vdots \\ 1 & x_1^{97} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

therefore, we can construct our model for each instance as follows:

$$\begin{bmatrix} h_\theta(x^{(1)}) \\ h_\theta(x^{(2)}) \\ \vdots \\ h_\theta(x^{(97)}) \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} \\ 1 & x_1^{(2)} \\ \vdots & \vdots \\ 1 & x_1^{97} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$
$$= X\theta$$

therefore, we can construct our model for each instance as follows:

$$\begin{bmatrix} h_\theta(x^{(1)}) \\ h_\theta(x^{(2)}) \\ \vdots \\ h_\theta(x^{(97)}) \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} \\ 1 & x_1^{(2)} \\ \vdots & \vdots \\ 1 & x_1^{97} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$
$$= X\theta$$

$X\theta$ is our *vectorized hyphotesis/model*.

Machine Learning Component: Model (2/3)

therefore, we can construct our model for each instance as follows:

$$\begin{bmatrix} h_\theta(x^{(1)}) \\ h_\theta(x^{(2)}) \\ \vdots \\ h_\theta(x^{(97)}) \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} \\ 1 & x_1^{(2)} \\ \vdots & \vdots \\ 1 & x_1^{97} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$
$$= X\theta$$

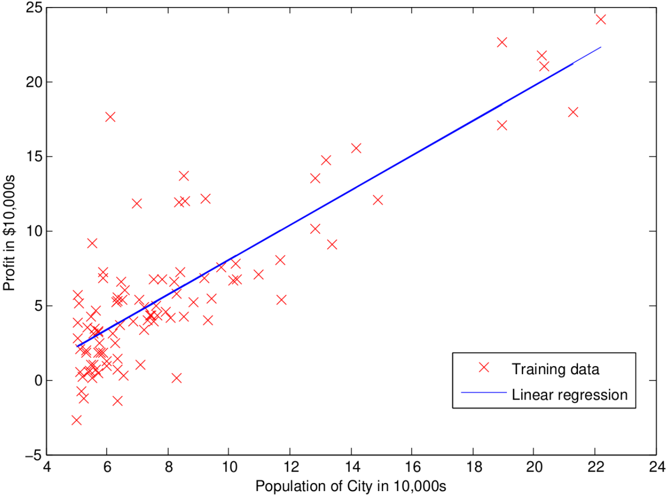
$X\theta$ is our *vectorized hyphotesis/model*.

Utilizing a product of two matrices $\Longrightarrow$ improve computing efficiency.

Assuming that we have θ, we can compute and plot our regression line as follows:

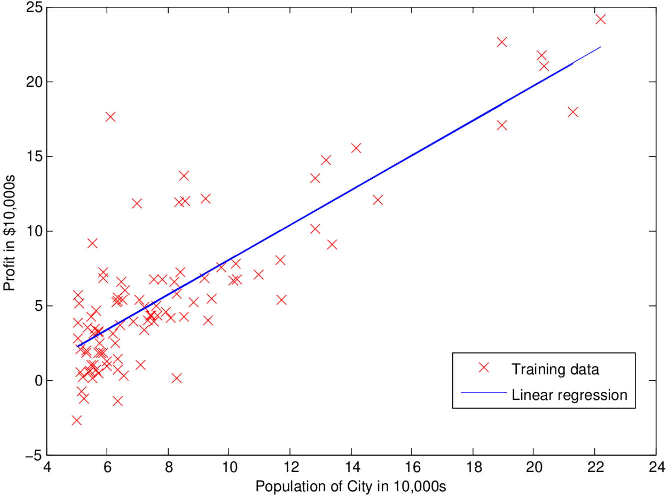
Machine Learning Component: Model (3/3)

Assuming that we have θ, we can compute and plot our regression line as follows:



Machine Learning Component: Model (3/3)

Assuming that we have θ, we can compute and plot our regression line as follows:



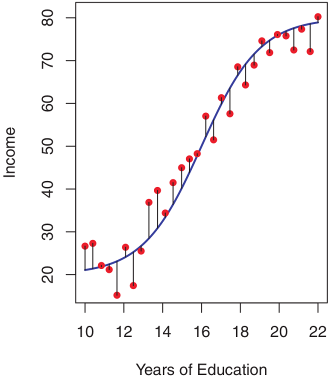
The regression line; however, how do we find θ_0 and θ_1?

Outline

A Cost/Loss Function calculates *all the errors made by the model*.

Machine Learning Component: Cost Function (1/3)

A Cost/Loss Function calculates *all the errors made by the model*.



An example of errors; Specifically *the black lines* represent the error associated with each instance (**?**)

Machine Learning Component: Cost Function (2/3)

The Cost/Loss Function is usually denoted as J; moreover, in case of linear regression:

$$
\begin{aligned}
J(\theta) &= \\
&= \\
&=
\end{aligned}
$$

The Cost/Loss Function is usually denoted as J; moreover, in case of linear regression:

$$
\begin{aligned}
J(\theta) &= \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \\
&= \\
&=
\end{aligned}
$$

Machine Learning Component: Cost Function (2/3)

The Cost/Loss Function is usually denoted as J; moreover, in case of linear regression:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

$$=$$

$$=$$

The Cost/Loss Function is usually denoted as J; moreover, in case of linear regression:

$$
\begin{aligned}
J(\theta) &= \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 \\
&= \frac{1}{2\times 97}\sum_{i=1}^{97}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 \\
&=
\end{aligned}
$$

Machine Learning Component: Cost Function (2/3)

The Cost/Loss Function is usually denoted as J; moreover, in case of linear regression:

$$\begin{aligned}
J(\theta) &= \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \\
&= \frac{1}{2 \times 97} \sum_{i=1}^{97} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \\
&= \frac{1}{2 \times 97} \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)} \right)^2
\end{aligned}$$

Machine Learning Component: Cost Function (2/3)

The Cost/Loss Function is usually denoted as J; moreover, in case of linear regression:

$$
\begin{aligned}
J(\theta) &= \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \\
&= \frac{1}{2 \times 97} \sum_{i=1}^{97} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \\
&= \frac{1}{2 \times 97} \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)} \right)^2
\end{aligned}
$$

The Goal:

The Cost/Loss Function is usually denoted as J; moreover, in case of linear regression:

$$
\begin{aligned}
J(\theta) &= \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \\
&= \frac{1}{2 \times 97} \sum_{i=1}^{97} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \\
&= \frac{1}{2 \times 97} \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)} \right)^2
\end{aligned}
$$

The Goal: How do we minimize **the cost function**?

Machine Learning Component: Cost Function (2/3)

The Cost/Loss Function is usually denoted as J; moreover, in case of linear regression:

$$
\begin{aligned}
J(\theta) &= \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \\
&= \frac{1}{2 \times 97} \sum_{i=1}^{97} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \\
&= \frac{1}{2 \times 97} \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)} \right)^2
\end{aligned}
$$

The Goal: How do we minimize **the cost function**?

We need **Calculus**.

We need to find θ_0 and θ_1 which minimize $J(\theta)$, that is

$$\arg\min_{\theta_0,\theta_1} \qquad .$$

Machine Learning Component: Cost Function (3/3)

We need to find θ_0 and θ_1 which minimize $J(\theta)$, that is

$$\underset{\theta_0,\theta_1}{\arg\min} \left\{ \frac{1}{2 \times 97} \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)} \right)^2 \right\}.$$

We need to find θ_0 and θ_1 which minimize $J(\theta)$, that is

$$\arg\min_{\theta_0,\theta_1} \left\{ \frac{1}{2 \times 97} \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)} \right)^2 \right\}.$$

How do we find θ_0 and θ_1 that minimize the **cost function**?

We need to find θ_0 and θ_1 which minimize $J(\theta)$, that is

$$\arg\min_{\theta_0,\theta_1} \left\{ \frac{1}{2 \times 97} \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)} \right)^2 \right\}.$$

How do we find θ_0 and θ_1 that minimize the **cost function**?

We need an **Optimization Algorithm.**

Outline

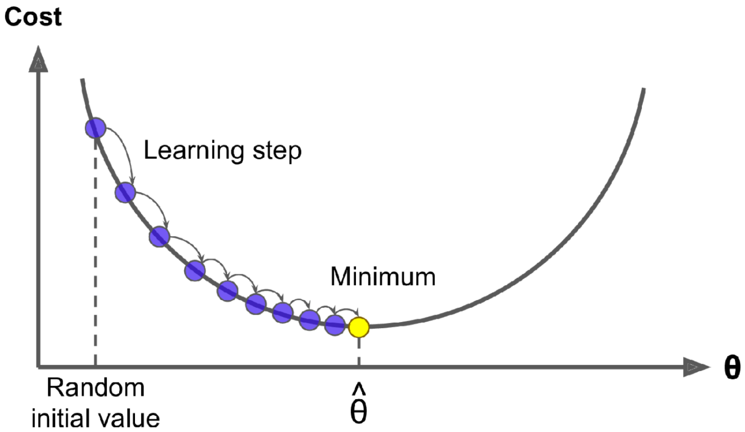
ML Component: Optimization Algorithm (1/7)

A famous optimization algorithm:

ML Component: Optimization Algorithm (1/7)

A famous optimization algorithm: **Gradient Descent**

A famous optimization algorithm: **Gradient Descent**



Gradient descent in action (**?**)

ML Component: Optimization Algorithm (2/7)

Gradient Descent is a simple algorithm:

ML Component: Optimization Algorithm (2/7)

Gradient Descent is a simple algorithm:

$$\theta_{\text{new}} = \theta_{\text{old}} - \alpha \times \frac{\partial J}{\partial \theta}$$

ML Component: Optimization Algorithm (2/7)

Gradient Descent is a simple algorithm:

$$\theta_{\text{new}} = \theta_{\text{old}} - \alpha \times \frac{\partial J}{\partial \theta}$$

with α is a *learning rate* and $\frac{\partial J}{\partial \theta}$ is called the *gradient*.

Gradient Descent is a simple algorithm:

$$\theta_{\text{new}} = \theta_{\text{old}} - \alpha \times \frac{\partial J}{\partial \theta}$$

with α is a *learning rate* and $\frac{\partial J}{\partial \theta}$ is called the *gradient*.

The *gradient* is the **partial derivative** of J
(We need **Calculus** to calculate this one).

ML Component: Optimization Algorithm (3/7)

Our cost/loss function for linear regression:

$$J(\theta) = \frac{1}{2 \times 97} \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)} \right)^2$$

With the help of **Calculus**, we have

$$\frac{\partial J}{\partial \theta_0} =$$

$$=$$

$$=$$

ML Component: Optimization Algorithm (3/7)

Our cost/loss function for linear regression:

$$J(\theta) = \frac{1}{2 \times 97} \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)^2$$

With the help of **Calculus**, we have

$$\frac{\partial J}{\partial \theta_0} = \left(\frac{1}{97 \times 2}\right)$$

$$=$$

$$=$$

ML Component: Optimization Algorithm (3/7)

Our cost/loss function for linear regression:

$$J(\theta) = \frac{1}{2 \times 97} \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)^2$$

With the help of **Calculus**, we have

$$\frac{\partial J}{\partial \theta_0} = \left(\frac{1}{97 \times 2}\right) \times 2$$

$$=$$

$$=$$

ML Component: Optimization Algorithm (3/7)

Our cost/loss function for linear regression:

$$J(\theta) = \frac{1}{2 \times 97} \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)^2$$

With the help of **Calculus**, we have

$$\frac{\partial J}{\partial \theta_0} = \left(\frac{1}{97 \times 2}\right) \times 2 \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)$$

$$=$$

$$=$$

Our cost/loss function for linear regression:

$$J(\theta) = \frac{1}{2 \times 97} \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)} \right)^2$$

With the help of **Calculus**, we have

$$\begin{aligned}
\frac{\partial J}{\partial \theta_0} &= \left(\frac{1}{97 \times 2} \right) \times 2 \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)} \right) \\
&= \left(\frac{1}{97 \times \cancel{2}} \right) \cancel{2} \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)} \right) \\
&=
\end{aligned}$$

ML Component: Optimization Algorithm (3/7)

Our cost/loss function for linear regression:

$$J(\theta) = \frac{1}{2 \times 97} \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)^2$$

With the help of **Calculus**, we have

$$\begin{aligned}
\frac{\partial J}{\partial \theta_0} &= \left(\frac{1}{97 \times 2}\right) \times 2 \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right) \\
&= \left(\frac{1}{97 \times \cancel{2}}\right) \cancel{2} \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right) \\
&= \frac{1}{97} \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right).
\end{aligned}$$

Our cost/loss function for linear regression:

$$J(\theta) = \frac{1}{2 \times 97} \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)^2$$

Again with the help of **Calculus**, we have

$$\frac{\partial J}{\partial \theta_1} =$$

$$=$$

$$=$$

ML Component: Optimization Algorithm (4/7)

Our cost/loss function for linear regression:

$$J(\theta) = \frac{1}{2 \times 97} \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)^2$$

Again with the help of **Calculus**, we have

$$\frac{\partial J}{\partial \theta_1} = \left(\frac{1}{97 \times 2}\right)$$

$$=$$

$$=$$

ML Component: Optimization Algorithm (4/7)

Our cost/loss function for linear regression:

$$J(\theta) = \frac{1}{2 \times 97} \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)} \right)^2$$

Again with the help of **Calculus**, we have

$$\begin{aligned}
\frac{\partial J}{\partial \theta_1} &= \left(\frac{1}{97 \times 2} \right) \times 2 \\
&= \\
&=
\end{aligned}$$

Our cost/loss function for linear regression:

$$J(\theta) = \frac{1}{2 \times 97} \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)^2$$

Again with the help of **Calculus**, we have

$$\begin{aligned}
\frac{\partial J}{\partial \theta_1} &= \left(\frac{1}{97 \times 2}\right) \times 2 \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right) \\
&= \\
&=
\end{aligned}$$

ML Component: Optimization Algorithm (4/7)

Our cost/loss function for linear regression:

$$J(\theta) = \frac{1}{2 \times 97} \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)^2$$

Again with the help of **Calculus**, we have

$$\begin{aligned}\frac{\partial J}{\partial \theta_1} &= \left(\frac{1}{97 \times 2}\right) \times 2 \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right) x_1^{(i)} \\ &= \\ &= \end{aligned}$$

Our cost/loss function for linear regression:

$$J(\theta) = \frac{1}{2 \times 97} \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)^2$$

Again with the help of **Calculus**, we have

$$\begin{aligned}
\frac{\partial J}{\partial \theta_1} &= \left(\frac{1}{97 \times 2}\right) \times 2 \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right) x_1^{(i)} \\
&= \left(\frac{1}{97 \times \cancel{2}}\right) \cancel{2} \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right) x_1^{(i)} \\
&=
\end{aligned}$$

ML Component: Optimization Algorithm (4/7)

Our cost/loss function for linear regression:

$$J(\theta) = \frac{1}{2 \times 97} \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)^2$$

Again with the help of **Calculus**, we have

$$\begin{aligned}
\frac{\partial J}{\partial \theta_1} &= \left(\frac{1}{97 \times 2}\right) \times 2 \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right) x_1^{(i)} \\
&= \left(\frac{1}{97 \times \cancel{2}}\right) \cancel{2} \times \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right) x_1^{(i)} \\
&= \frac{1}{97} \sum_{i=1}^{97} \left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right) x_1^{(i)}.
\end{aligned}$$

As a summary, we have

ML Component: Optimization Algorithm (5/7)

As a summary, we have

$$\theta_0 = \theta_0 - \alpha \times \frac{1}{97}\sum_{i=1}^{97}\left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right) \text{ and}$$

As a summary, we have

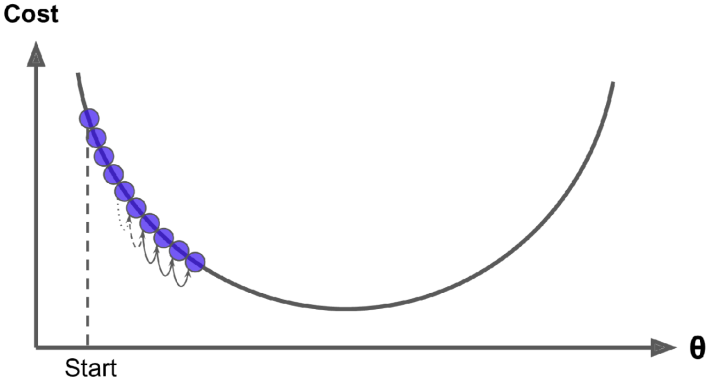
$$\theta_0 = \theta_0 - \alpha \times \frac{1}{97}\sum_{i=1}^{97}\left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right) \text{ and}$$

$$\theta_1 = \theta_1 - \alpha \times \frac{1}{97}\sum_{i=1}^{97}\left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right) x_1^{(i)}$$

ML Component: Optimization Algorithm (6/7)

Learning rate (α) is the size of the steps.

Learning rate (α) is the size of the steps.



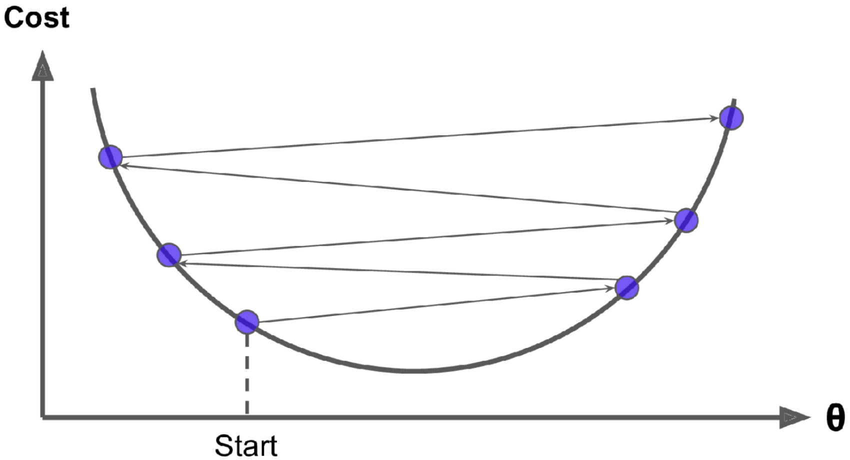
This will happen when the α is too small (**?**)

ML Component: Optimization Algorithm (7/7)

What happens when α is too large?

ML Component: Optimization Algorithm (7/7)

What happens when α is too large?



This will happen when α is too large (**?**)

Outline

Demo Time

We show a *Linear Regression* as our Machine Learning demo[1].

[1] https://www.coursera.org/learn/machine-learning

Outline

Importance of Interpretability

Why?

Why?

- A *single metric*, for example: *classification accuracy*, is an *incomplete description* of most real-world tasks.

Importance of Interpretability

Why?

- A *single metric*, for example: *classification accuracy*, is an *incomplete description* of most real-world tasks.
- The '*Why*' can help you learn more about *the problem*, *the data*, and *the reason* why a model might fail.

Importance of Interpretability

Why?

- A *single metric*, for example: *classification accuracy*, is an *incomplete description* of most real-world tasks.
- The '*Why*' can help you learn more about *the problem*, *the data*, and *the reason* why a model might fail.
- *Human curiosity* and *learning* $\Longrightarrow$ *interpretability* & *explanations*.

Importance of Interpretability

Why?

- A *single metric*, for example: *classification accuracy*, is an *incomplete description* of most real-world tasks.
- The '*Why*' can help you learn more about *the problem*, *the data*, and *the reason* why a model might fail.
- *Human curiosity* and *learning* $\Longrightarrow$ *interpretability* & *explanations*.

That's why we need the **interpretability of machine learning**.

When We Do NOT Need Interpretability

When We Do NOT Need Interpretability

- Interpretability is not required if the model has *no significant impact*.

When We Do NOT Need Interpretability

- Interpretability is not required if the model has *no significant impact*.
  Example: Predict where our friends will go for their next culinary tour based on Instagram/Facebook data.

When We Do NOT Need Interpretability

- Interpretability is not required if the model has *no significant impact*.
  Example: Predict where our friends will go for their next culinary tour based on Instagram/Facebook data.

- Interpretability is not required when *the problem is well-studied*.

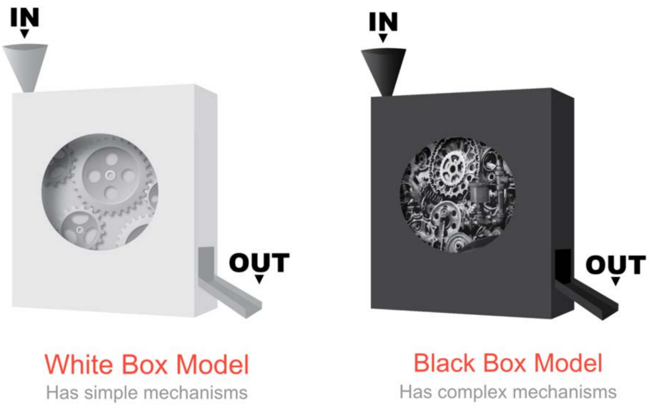
When We Do NOT Need Interpretability

- Interpretability is not required if the model has *no significant impact*.
  Example: Predict where our friends will go for their next culinary tour based on Instagram/Facebook data.

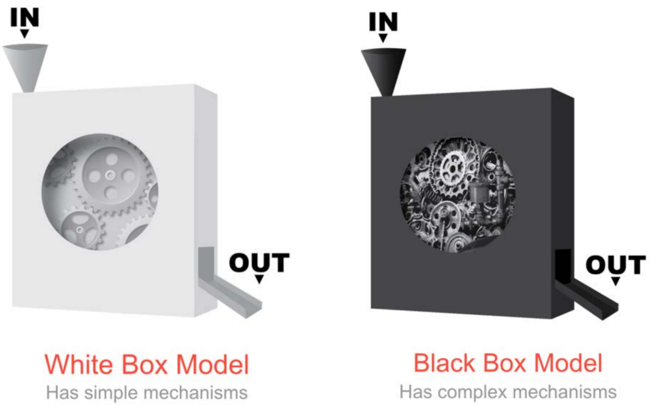
- Interpretability is not required when *the problem is well-studied*.
  Example: A machine learning model for OCR that processes images from envelopes and extracts addresses.

The Black Box of Machine Learning (?)

White Box vs Black Box (?)

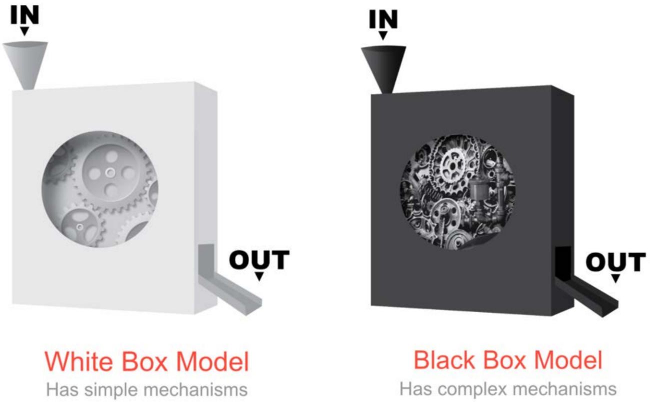


White Box vs Black Box (?)



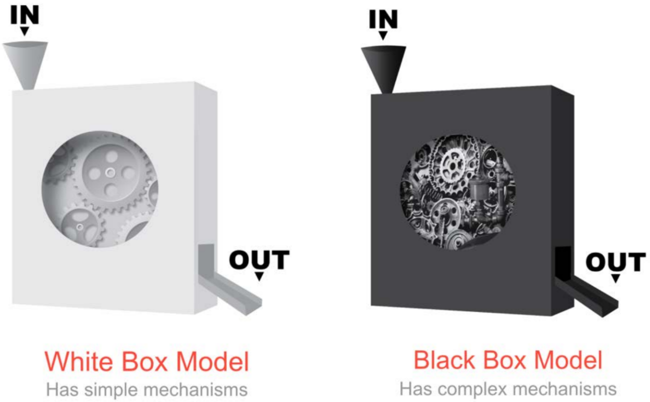
White box models are *transparent*.

White Box vs Black Box (?)



White box models are *transparent*.
They achieve *total* or *near-total interpretation transparency*

White Box vs Black Box (?)



White box models are *transparent*.
They achieve *total* or *near-total interpretation transparency*
$\Longrightarrow$ **interpretable.**

Outline

Interpretable Models (1/3)

- The easiest way to achieve interpretability is

Interpretable Models (1/3)

- The easiest way to achieve interpretability is to *use a subset of algorithms that create interpretable models*.

Interpretable Models (1/3)

- The easiest way to achieve interpretability is to *use a subset of algorithms that create interpretable models*.
- Commonly used interpretable models (**?**) are

Interpretable Models (1/3)

- The easiest way to achieve interpretability is to *use a subset of algorithms that create interpretable models*.
- Commonly used interpretable models (**?**) are
  - **linear regression** $\Longrightarrow$ We'll talk about this later,

Interpretable Models (1/3)

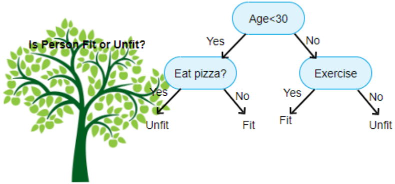
- The easiest way to achieve interpretability is to *use a subset of algorithms that create interpretable models*.
- Commonly used interpretable models (**?**) are
  - **linear regression** $\Longrightarrow$ We'll talk about this later,
  - logistic regression,

Interpretable Models (1/3)

- The easiest way to achieve interpretability is to *use a subset of algorithms that create interpretable models*.
- Commonly used interpretable models (**?**) are
  - **linear regression** $\Longrightarrow$ We'll talk about this later,
  - logistic regression,
  - other linear regression extensions, such as *Generalized Linear Models* (GLMs) and *Generalized Additive Models* (GAMs),

Interpretable Models (1/3)

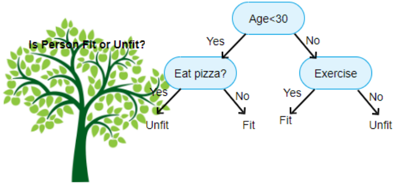
- The easiest way to achieve interpretability is to *use a subset of algorithms that create interpretable models*.
- Commonly used interpretable models (**?**) are
  - **linear regression** $\Longrightarrow$ We'll talk about this later,
  - logistic regression,
  - other linear regression extensions, such as *Generalized Linear Models* (GLMs) and *Generalized Additive Models* (GAMs),
  - decision tree,



An example of a decision tree (**?**)

Interpretable Models (1/3)

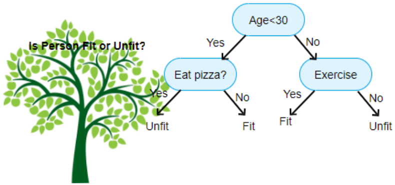
- The easiest way to achieve interpretability is to *use a subset of algorithms that create interpretable models*.
- Commonly used interpretable models (**?**) are
  - **linear regression** $\Longrightarrow$ We'll talk about this later,
  - logistic regression,
  - other linear regression extensions, such as *Generalized Linear Models* (GLMs) and *Generalized Additive Models* (GAMs),
  - decision tree,
  - decision rules, and



An example of a decision tree (**?**)

Interpretable Models (1/3)

- The easiest way to achieve interpretability is to *use a subset of algorithms that create interpretable models*.
- Commonly used interpretable models (**?**) are
  - **linear regression** $\Longrightarrow$ We'll talk about this later,
  - logistic regression,
  - other linear regression extensions, such as *Generalized Linear Models* (GLMs) and *Generalized Additive Models* (GAMs),
  - decision tree,
  - decision rules, and
  - the RuleFit algorithm.



An example of a decision tree (**?**)

The properties of interpretable models:

Interpretable Models (2/3)

The properties of interpretable models:

- A model is **linear** if the association between features and target is modelled linearly.

Interpretable Models (2/3)

The properties of interpretable models:

- A model is **linear** if the association between features and target is modelled linearly.
- A model with **monotonicity constraints** ensures that the relationship between a feature and the target outcome always goes in the same direction over the entire range of the feature $\implies$ easier to understand a relationship.

Interpretable Models (2/3)

The properties of interpretable models:

- A model is **linear** if the association between features and target is modelled linearly.
- A model with **monotonicity constraints** ensures that the relationship between a feature and the target outcome always goes in the same direction over the entire range of the feature $\Longrightarrow$ easier to understand a relationship.
- Some models can automatically include **interactions between features** to predict the target outcome $\Longrightarrow$ improve predictive performance.

Interpretable Models (2/3)

The properties of interpretable models:

- A model is **linear** if the association between features and target is modelled linearly.
- A model with **monotonicity constraints** ensures that the relationship between a feature and the target outcome always goes in the same direction over the entire range of the feature $\Longrightarrow$ easier to understand a relationship.
- Some models can automatically include **interactions between features** to predict the target outcome $\Longrightarrow$ improve predictive performance.
  A famous example comes from **?**.

Interpretable Models (3/3)

Algorithm	Linear	Monotone	Interaction	Task

Interpretable Models (3/3)

Algorithm	Linear	Monotone	Interaction	Task
Linear regression	✓	✓	✗	regr

Algorithm	**Linear**	**Monotone**	**Interaction**	**Task**
Linear regression	✓	✓	✗	regr
Logistic regression	✗	✓	✗	class

Interpretable Models (3/3)

Algorithm	**Linear**	**Monotone**	**Interaction**	**Task**
Linear regression	✓	✓	✗	regr
Logistic regression	✗	✓	✗	class
Decision trees	✗	**?**	✓	class, regr

? indicates that a decision tree can *sometimes* be monotone

Interpretable Models (3/3)

Algorithm	Linear	Monotone	Interaction	Task
Linear regression	✓	✓	✗	regr
Logistic regression	✗	✓	✗	class
Decision trees	✗	?	✓	class, regr
RuleFit	✓	✗	✓	class, regr

? indicates that a decision tree can *sometimes* be monotone

Interpretable Models (3/3)

Algorithm	Linear	Monotone	Interaction	Task
Linear regression	✓	✓	✗	regr
Logistic regression	✗	✓	✗	class
Decision trees	✗	?	✓	class, regr
RuleFit	✓	✗	✓	class, regr
Naïve-Bayes	✗	✓	✗	class

? indicates that a decision tree can *sometimes* be monotone

Interpretable Models (3/3)

Algorithm	Linear	Monotone	Interaction	Task
Linear regression	✓	✓	✗	regr
Logistic regression	✗	✓	✗	class
Decision trees	✗	?	✓	class, regr
RuleFit	✓	✗	✓	class, regr
Naïve-Bayes	✗	✓	✗	class
k-nearest neighbors	✗	✗	✗	class, regr

? indicates that a decision tree can *sometimes* be monotone

Interpretable Models (3/3)

Algorithm	Linear	Monotone	Interaction	Task
Linear regression	✓	✓	✗	regr
Logistic regression	✗	✓	✗	class
Decision trees	✗	**?**	✓	class, regr
RuleFit	✓	✗	✓	class, regr
Naïve-Bayes	✗	✓	✗	class
k-nearest neighbors	✗	✗	✗	class, regr

? indicates that a decision tree can *sometimes* be monotone

We shall explore a **linear regression algorithm** as an example of *interpretable model*.

Outline

Example of Interpretable Model: Dataset (1/3)

[2] https://www.capitalbikeshare.com

Example of Interpretable Model: Dataset (1/3)

- Dataset: **Bike Rentals** (Regression).

[2] https://www.capitalbikeshare.com

Example of Interpretable Model: Dataset (1/3)

- Dataset: **Bike Rentals** (Regression).
- This dataset contains *daily counts of rented bicycles* from the bicycle rental company Capital-Bikeshare[2] in Washington D.C.

[2] https://www.capitalbikeshare.com

Example of Interpretable Model: Dataset (1/3)

- Dataset: **Bike Rentals** (Regression).
- This dataset contains *daily counts of rented bicycles* from the bicycle rental company Capital-Bikeshare[2] in Washington D.C.
- The goal is to *predict how many bikes will be rented* depending on the weather and the day.

[2] https://www.capitalbikeshare.com

Example of an Interpretable Model: Dataset (2/3)

Here are the *nine features* that are used:

Example of an Interpretable Model: Dataset (2/3)

Here are the *nine features* that are used:

- Count of bicycles

Example of an Interpretable Model: Dataset (2/3)

Here are the *nine features* that are used:

- Count of bicycles $\Longrightarrow y$

Example of an Interpretable Model: Dataset (2/3)

Here are the *nine features* that are used:

- Count of bicycles $\Longrightarrow y$
- The season, either spring, summer, fall, or winter

Example of an Interpretable Model: Dataset (2/3)

Here are the *nine features* that are used:

- Count of bicycles $\Longrightarrow y$
- The season, either spring, summer, fall, or winter
- Indicator whether the day was a holiday or not

Example of an Interpretable Model: Dataset (2/3)

Here are the *nine features* that are used:

- Count of bicycles $\Longrightarrow y$
- The season, either spring, summer, fall, or winter
- Indicator whether the day was a holiday or not
- Number of days since the 01.01.2011 (the first day in the dataset)

Example of an Interpretable Model: Dataset (2/3)

Here are the *nine features* that are used:

- Count of bicycles $\Longrightarrow y$
- The season, either spring, summer, fall, or winter
- Indicator whether the day was a holiday or not
- Number of days since the 01.01.2011 (the first day in the dataset)
- Indicator whether the day was a working day or weekend

Example of an Interpretable Model: Dataset (2/3)

Here are the *nine features* that are used:

- Count of bicycles $\Longrightarrow y$
- The season, either spring, summer, fall, or winter
- Indicator whether the day was a holiday or not
- Number of days since the 01.01.2011 (the first day in the dataset)
- Indicator whether the day was a working day or weekend
- The weather situation on that day: good, misty, or rain/snow/storm

Example of an Interpretable Model: Dataset (2/3)

Here are the *nine features* that are used:

- Count of bicycles $\Longrightarrow y$
- The season, either spring, summer, fall, or winter
- Indicator whether the day was a holiday or not
- Number of days since the 01.01.2011 (the first day in the dataset)
- Indicator whether the day was a working day or weekend
- The weather situation on that day: good, misty, or rain/snow/storm
- Temperature in degrees Celcius

Example of an Interpretable Model: Dataset (2/3)

Here are the *nine features* that are used:

- Count of bicycles $\Longrightarrow y$
- The season, either spring, summer, fall, or winter
- Indicator whether the day was a holiday or not
- Number of days since the 01.01.2011 (the first day in the dataset)
- Indicator whether the day was a working day or weekend
- The weather situation on that day: good, misty, or rain/snow/storm
- Temperature in degrees Celcius
- Relative humidity in percent (0 to 100)

Example of an Interpretable Model: Dataset (2/3)

Here are the *nine features* that are used:

- Count of bicycles $\Longrightarrow y$
- The season, either spring, summer, fall, or winter
- Indicator whether the day was a holiday or not
- Number of days since the 01.01.2011 (the first day in the dataset)
- Indicator whether the day was a working day or weekend
- The weather situation on that day: good, misty, or rain/snow/storm
- Temperature in degrees Celcius
- Relative humidity in percent (0 to 100)
- Wind speed in km per hour

Example of Interpretable Model: Dataset (3/3)

```
summary(bike_to_interpreted)
```

```
##       cnt            season          holiday       days_since_2011
##  Min.   :  22   SPRING:181   NO HOLIDAY:710   Min.   :  0.0
##  1st Qu.:3152   SUMMER:184   HOLIDAY   : 21   1st Qu.:182.5
##  Median :4548   FALL  :188                    Median :365.0
##  Mean   :4504   WINTER:178                    Mean   :365.0
##  3rd Qu.:5956                                 3rd Qu.:547.5
##  Max.   :8714                                 Max.   :730.0
##          workingday            weathersit       temp             hum
##  NO WORKING DAY:231   GOOD              :463   Min.   :-5.221   Min.   : 0.00
##  WORKING DAY   :500   MISTY             :247   1st Qu.: 7.843   1st Qu.:52.00
##                       RAIN/SNOW/STORM: 21   Median :15.422   Median :62.67
##                                                Mean   :15.283   Mean   :62.79
##                                                3rd Qu.:22.805   3rd Qu.:73.02
##                                                Max.   :32.498   Max.   :97.25
##    windspeed
##  Min.   : 1.500
##  1st Qu.: 9.042
##  Median :12.125
##  Mean   :12.763
##  3rd Qu.:15.625
##  Max.   :34.000
```

A summary of `bike_to_interpreted` dataset

Example of Interpretable Model: Linear Regression

The learned relationships between X and y are **linear** and can be written for a single instance i as follows:

Example of Interpretable Model: Linear Regression

The learned relationships between X and y are **linear** and can be written for a single instance i as follows:

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p + \epsilon$$

Example of Interpretable Model: Linear Regression

The learned relationships between X and y are **linear** and can be written for a single instance i as follows:

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p + \epsilon$$

The epsilon (ϵ) is the **error** we still make, i.e. the difference between the prediction and the actual outcome.

Linear Regression Assumptions

Whether the linear regression model is the "correct" model depends on whether the relationships in the data meet certain assumptions (**?**), which are

Linear Regression Assumptions

Whether the linear regression model is the "correct" model depends on whether the relationships in the data meet certain assumptions (**?**), which are

- **Linearity of the data**

Linear Regression Assumptions

Whether the linear regression model is the "correct" model depends on whether the relationships in the data meet certain assumptions (**?**), which are

- **Linearity of the data** $\rightarrow$ the relationship between the predictor (X) and the outcome (y) is assumed to be linear.

Linear Regression Assumptions

Whether the linear regression model is the "correct" model depends on whether the relationships in the data meet certain assumptions (**?**), which are

- **Linearity of the data** $\rightarrow$ the relationship between the predictor (X) and the outcome (y) is assumed to be linear.
- **Normality of residuals**

Linear Regression Assumptions

Whether the linear regression model is the "correct" model depends on whether the relationships in the data meet certain assumptions (**?**), which are

- **Linearity of the data** $\rightarrow$ the relationship between the predictor (X) and the outcome (y) is assumed to be linear.
- **Normality of residuals** $\rightarrow$ the residual errors are assumed to be *normally distributed*.

Linear Regression Assumptions

Whether the linear regression model is the "correct" model depends on whether the relationships in the data meet certain assumptions (**?**), which are

- **Linearity of the data** $\rightarrow$ the relationship between the predictor (X) and the outcome (y) is assumed to be linear.
- **Normality of residuals** $\rightarrow$ the residual errors are assumed to be *normally distributed*.
- **Homoscedasticity**

Linear Regression Assumptions

Whether the linear regression model is the "correct" model depends on whether the relationships in the data meet certain assumptions (**?**), which are

- **Linearity of the data** $\rightarrow$ the relationship between the predictor (X) and the outcome (y) is assumed to be linear.
- **Normality of residuals** $\rightarrow$ the residual errors are assumed to be *normally distributed*.
- **Homoscedasticity** $\rightarrow$ the residuals are assumed to have a *constant variance*.

Linear Regression Assumptions

Whether the linear regression model is the "correct" model depends on whether the relationships in the data meet certain assumptions (**?**), which are

- **Linearity of the data** $\rightarrow$ the relationship between the predictor (X) and the outcome (y) is assumed to be linear.
- **Normality of residuals** $\rightarrow$ the residual errors are assumed to be *normally distributed*.
- **Homoscedasticity** $\rightarrow$ the residuals are assumed to have a *constant variance*.
- **Independence of residuals error terms**

Linear Regression Assumptions

Whether the linear regression model is the "correct" model depends on whether the relationships in the data meet certain assumptions (**?**), which are

- **Linearity of the data** $\rightarrow$ the relationship between the predictor (X) and the outcome (y) is assumed to be linear.
- **Normality of residuals** $\rightarrow$ the residual errors are assumed to be *normally distributed*.
- **Homoscedasticity** $\rightarrow$ the residuals are assumed to have a *constant variance*.
- **Independence of residuals error terms** $\rightarrow$ each instance is *independent* of any other instance.

Outline

Example of Interpretable Model: How to Interpret (1/5)

The linear regression model:

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p + \epsilon$$

Example of Interpretable Model: How to Interpret (1/5)

The linear regression model:

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p + \epsilon$$

The interpretation of the features in the linear regression model can be automated by using following text templates:

Example of Interpretable Model: How to Interpret (1/5)

The linear regression model:

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p + \epsilon$$

The interpretation of the features in the linear regression model can be automated by using following text templates:

- **Interpretation of a Numerical Feature**

The linear regression model:

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p + \epsilon$$

The interpretation of the features in the linear regression model can be automated by using following text templates:

- **Interpretation of a Numerical Feature**
  An increase of feature x_k by one unit increases the prediction for y by θ_k units when *all other feature values remain fixed*.

Example of Interpretable Model: How to Interpret (1/5)

The linear regression model:

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p + \epsilon$$

The interpretation of the features in the linear regression model can be automated by using following text templates:

- **Interpretation of a Numerical Feature**
  An increase of feature x_k by one unit increases the prediction for y by θ_k units when *all other feature values remain fixed*.
- **Interpretation of a Categorical Feature**

Example of Interpretable Model: How to Interpret (1/5)

The linear regression model:

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p + \epsilon$$

The interpretation of the features in the linear regression model can be automated by using following text templates:

- **Interpretation of a Numerical Feature**
  An increase of feature x_k by one unit increases the prediction for y by θ_k units when *all other feature values remain fixed*.

- **Interpretation of a Categorical Feature**
  Changing feature x_k from the reference category to the other category increases the prediction for y by θ_k when *all other features remain fixed*.

Example of Interpretable Model: How to Interpret (2/5)

Example of Interpretable Model: How to Interpret (2/5)

- Another important measurement for interpreting linear models is

Example of Interpretable Model: How to Interpret (2/5)

- Another important measurement for interpreting linear models is the ***R*-squared** measurement.

Example of Interpretable Model: How to Interpret (2/5)

- Another important measurement for interpreting linear models is the ***R*-squared** measurement.
- ***R*-squared** tells us how much of the total variance of your target outcome is explained by the model.

Example of Interpretable Model: How to Interpret (2/5)

- Another important measurement for interpreting linear models is the ***R*-squared** measurement.
- ***R*-squared** tells us how much of the total variance of your target outcome is explained by the model.
- The higher ***R*-squared**, the better our model explains the data.

The formula for calculating ***R*-squared** is:

Example of Interpretable Model: How to Interpret (3/5)

The formula for calculating ***R*-squared** is:

$$R^2 = \frac{1 - SSE}{SST}$$

Example of Interpretable Model: How to Interpret (3/5)

The formula for calculating ***R*-squared** is:

$$R^2 = \frac{1 - SSE}{SST}$$

with *SSE* is squared sum of the error terms:

Example of Interpretable Model: How to Interpret (3/5)

The formula for calculating ***R*-squared** is:

$$R^2 = \frac{1 - SSE}{SST}$$

with *SSE* is squared sum of the error terms:

$$SSE = \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

Example of Interpretable Model: How to Interpret (3/5)

The formula for calculating ***R*-squared** is:

$$R^2 = \frac{1 - SSE}{SST}$$

with SSE is squared sum of the error terms:

$$SSE = \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

and SST is the squared sum of the data variance:

Example of Interpretable Model: How to Interpret (3/5)

The formula for calculating ***R*-squared** is:

$$R^2 = \frac{1 - SSE}{SST}$$

with *SSE* is squared sum of the error terms:

$$SSE = \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

and *SST* is the squared sum of the data variance:

$$SST = \sum_{i=1}^{m} (\overline{y} - y^{(i)})^2 \text{ with } \overline{y} = \frac{\sum_{i=1}^{m} y^{(i)}}{m}$$

Example of Interpretable Model: How to Interpret (3/5)

The formula for calculating ***R*-squared** is:

$$R^2 = \frac{1 - SSE}{SST}$$

with SSE is squared sum of the error terms:

$$SSE = \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

and SST is the squared sum of the data variance:

$$SST = \sum_{i=1}^{m} (\overline{y} - y^{(i)})^2 \text{ with } \overline{y} = \frac{\sum_{i=1}^{m} y^{(i)}}{m}$$

***R*-squared** tells us how much of our variance can be explained by the linear model ($0 \leq R$-squared ≤ 1).

It is better to use the **adjusted *R*-squared**, which accounts for the number of features used in the model. Its calculation is:

It is better to use the **adjusted *R*-squared**, which accounts for the number of features used in the model. Its calculation is:

$$\overline{R}^2 = 1 - (1 - R^2)\frac{m-1}{m-p-1}$$

Example of Interpretable Model: How to Interpret (4/5)

It is better to use the **adjusted *R*-squared**, which accounts for the number of features used in the model. Its calculation is:

$$\overline{R}^2 = 1 - (1 - R^2)\frac{m-1}{m-p-1}$$

where p and m are the number of features and the number of instances respectively.

Example of Interpretable Model: How to Interpret (4/5)

It is better to use the **adjusted *R*-squared**, which accounts for the number of features used in the model. Its calculation is:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{m-1}{m-p-1}$$

where p and m are the number of features and the number of instances respectively.

It is not meaningful to interpret a model with very low (**adjusted**) ***R*-squared**,

Example of Interpretable Model: How to Interpret (4/5)

It is better to use the **adjusted *R*-squared**, which accounts for the number of features used in the model. Its calculation is:

$$\overline{R}^2 = 1 - (1 - R^2)\frac{m-1}{m-p-1}$$

where p and m are the number of features and the number of instances respectively.

It is not meaningful to interpret a model with very low (**adjusted**) ***R*-squared**, because such a model basically does not explain much of the variance

It is better to use the **adjusted *R*-squared**, which accounts for the number of features used in the model. Its calculation is:

$$\overline{R}^2 = 1 - (1 - R^2)\frac{m-1}{m-p-1}$$

where p and m are the number of features and the number of instances respectively.

It is not meaningful to interpret a model with very low (**adjusted**) ***R*-squared**, because such a model basically does not explain much of the variance $\Longrightarrow$ any interpretation of the weights would not be meaningful.

Example of Interpretable Model: How to Interpret (5/5)

The importance of a feature (**feature importance**) in a linear regression model can be measured by

Example of Interpretable Model: How to Interpret (5/5)

The importance of a feature (**feature importance**) in a linear regression model can be measured by the *absolute value* of its t-statistic.

Example of Interpretable Model: How to Interpret (5/5)

The importance of a feature (**feature importance**) in a linear regression model can be measured by the *absolute value* of its t-statistic.
The t-statistic is

Example of Interpretable Model: How to Interpret (5/5)

The importance of a feature (**feature importance**) in a linear regression model can be measured by the *absolute value* of its t-statistic.
The t-statistic is the *estimated weight scaled with its standard error*.

Example of Interpretable Model: How to Interpret (5/5)

The importance of a feature (**feature importance**) in a linear regression model can be measured by the *absolute value* of its *t*-statistic.
The *t*-statistic is the *estimated weight scaled with its standard error*.

$$t_{\theta_j} = \frac{\theta_j}{SE(\theta_j)}$$

Example of Interpretable Model: How to Interpret (5/5)

The importance of a feature (**feature importance**) in a linear regression model can be measured by the *absolute value* of its *t*-statistic.
The *t*-statistic is the *estimated weight scaled with its standard error*.

$$t_{\theta_j} = \frac{\theta_j}{SE(\theta_j)}$$

The importance of a feature $\uparrow$,

Example of Interpretable Model: How to Interpret (5/5)

The importance of a feature (**feature importance**) in a linear regression model can be measured by the *absolute value* of its *t*-statistic.
The *t*-statistic is the *estimated weight scaled with its standard error*.

$$t_{\theta_j} = \frac{\theta_j}{SE(\theta_j)}$$

The importance of a feature $\uparrow$, the weight also $\uparrow$.

Bike Rentals: Interpretation (1/2)

In this example, we use linear regression to predict the **number of rented bikes** (cnt) on a particular day, given weather and calendar information.

Bike Rentals: Interpretation (1/2)

In this example, we use linear regression to predict the **number of rented bikes** (cnt) on a particular day, given weather and calendar information.

	Weight	SE	t
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

The estimated weight, the standard error of the estimate (*SE*), and the absolute value of *t*-statistic ($|t|$)

Bike Rentals: Interpretation (1/2)

In this example, we use linear regression to predict the **number of rented bikes** (cnt) on a particular day, given weather and calendar information.

$$\text{cnt} =$$

	Weight	SE	t
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

The estimated weight, the standard error of the estimate (SE), and the absolute value of t-statistic ($|t|$)

Bike Rentals: Interpretation (1/2)

In this example, we use linear regression to predict the **number of rented bikes** (cnt) on a particular day, given weather and calendar information.

$$\text{cnt} = \theta_0 +$$

	Weight	SE	t
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

The estimated weight, the standard error of the estimate (*SE*), and the absolute value of *t*-statistic ($|t|$)

Bike Rentals: Interpretation (1/2)

In this example, we use linear regression to predict the **number of rented bikes** (cnt) on a particular day, given weather and calendar information.

$$\text{cnt} = \theta_0 + \theta_1 x_1 +$$

	Weight	SE	t
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

The estimated weight, the standard error of the estimate (*SE*), and the absolute value of *t*-statistic ($|t|$)

Bike Rentals: Interpretation (1/2)

In this example, we use linear regression to predict the **number of rented bikes** (cnt) on a particular day, given weather and calendar information.

$$\text{cnt} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 +$$

	Weight	SE	t
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

The estimated weight, the standard error of the estimate (*SE*), and the absolute value of *t*-statistic ($|t|$)

Bike Rentals: Interpretation (1/2)

In this example, we use linear regression to predict the **number of rented bikes** (cnt) on a particular day, given weather and calendar information.

$$\text{cnt} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 +$$

	Weight	SE	t
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

The estimated weight, the standard error of the estimate (SE), and the absolute value of t-statistic ($|t|$)

Bike Rentals: Interpretation (1/2)

In this example, we use linear regression to predict the **number of rented bikes** (cnt) on a particular day, given weather and calendar information.

$$\text{cnt} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 +$$

	Weight	SE	t
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

The estimated weight, the standard error of the estimate (*SE*), and the absolute value of *t*-statistic ($|t|$)

Bike Rentals: Interpretation (1/2)

In this example, we use linear regression to predict the **number of rented bikes** (cnt) on a particular day, given weather and calendar information.

$$\text{cnt} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5 +$$

	Weight	SE	t
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

The estimated weight, the standard error of the estimate (*SE*), and the absolute value of *t*-statistic ($|t|$)

Bike Rentals: Interpretation (1/2)

In this example, we use linear regression to predict the **number of rented bikes** (cnt) on a particular day, given weather and calendar information.

$$\text{cnt} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5 + \theta_6 x_6 +$$

	Weight	SE	t
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

The estimated weight, the standard error of the estimate (SE), and the absolute value of t-statistic ($|t|$)

In this example, we use linear regression to predict the **number of rented bikes** (cnt) on a particular day, given weather and calendar information.

$$\text{cnt} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5 + \theta_6 x_6 + \theta_7 x_7 +$$

	Weight	SE	t
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

The estimated weight, the standard error of the estimate (*SE*), and the absolute value of *t*-statistic ($|t|$)

Bike Rentals: Interpretation (1/2)

In this example, we use linear regression to predict the **number of rented bikes** (cnt) on a particular day, given weather and calendar information.

$$\text{cnt} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5 + \theta_6 x_6 + \theta_7 x_7 + \theta_8 x_8 + \epsilon$$

	Weight	SE	t
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

The estimated weight, the standard error of the estimate (*SE*), and the absolute value of *t*-statistic ($|t|$)

- **Interpretation of a numerical feature ("temperature")**:

,

Bike Rentals: Interpretation (2/2)

- **Interpretation of a numerical feature ("temperature")**: An increase of the temperature by 1 degree Celcius increases the predicted number of bicycles by 110.7,

Bike Rentals: Interpretation (2/2)

- **Interpretation of a numerical feature ("temperature")**: An increase of the temperature by 1 degree Celcius increases the predicted number of bicycles by 110.7, *when all other features remain fixed*.

Bike Rentals: Interpretation (2/2)

- **Interpretation of a numerical feature ("temperature")**: An increase of the temperature by 1 degree Celcius increases the predicted number of bicycles by 110.7, *when all other features remain fixed*.
- **Interpretation of a categorical feature ("weathersit")**:

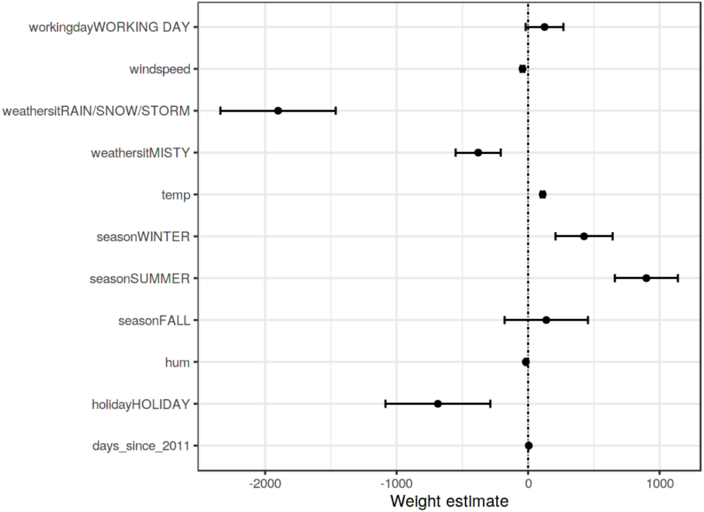
Bike Rentals: Interpretation (2/2)

- **Interpretation of a numerical feature ("temperature")**: An increase of the temperature by 1 degree Celcius increases the predicted number of bicycles by 110.7, *when all other features remain fixed*.
- **Interpretation of a categorical feature ("weathersit")**: The estimated number of bicycles is -1901.5 lower when it is raining, snowing or stormy, compared to good weather

Bike Rentals: Interpretation (2/2)

- **Interpretation of a numerical feature ("temperature")**: An increase of the temperature by 1 degree Celcius increases the predicted number of bicycles by 110.7, *when all other features remain fixed*.
- **Interpretation of a categorical feature ("weathersit")**: The estimated number of bicycles is -1901.5 lower when it is raining, snowing or stormy, compared to good weather—again *assuming that all other features do not change*.

Visual Interpretation: Weight Plot (1/2)



Weights are displayed as points and the 95% confidence intervals as lines

Visual Interpretation: Weight Plot (2/2)

Visual Interpretation: Weight Plot (2/2)

- The figure shows that rainy/snowy/stormy weather has a **strong negative effect** on the predicted number of bikes.

Visual Interpretation: Weight Plot (2/2)

- The figure shows that rainy/snowy/stormy weather has a **strong negative effect** on the predicted number of bikes.
- The weight of the working day feature is close to zero and zero is included in the 95% interval $\longrightarrow$ the effect is **not** statistically significant.

Visual Interpretation: Weight Plot (2/2)

- The figure shows that rainy/snowy/stormy weather has a **strong negative effect** on the predicted number of bikes.
- The weight of the working day feature is close to zero and zero is included in the 95% interval $\longrightarrow$ the effect is **not** statistically significant.
- Some *confidence intervals* are *very short* and the estimates are *close to zero*, yet the *feature effects were statistically significant* $\longrightarrow$ temperature.

Visual Interpretation: Weight Plot (2/2)

- The figure shows that rainy/snowy/stormy weather has a **strong negative effect** on the predicted number of bikes.
- The weight of the working day feature is close to zero and zero is included in the 95% interval $\longrightarrow$ the effect is **not** statistically significant.
- Some *confidence intervals* are *very short* and the estimates are *close to zero*, yet the *feature effects were statistically significant* $\longrightarrow$ temperature.
- The problem: the features are measured on different scales.

Visual Interpretation: Weight Plot (2/2)

- The figure shows that rainy/snowy/stormy weather has a **strong negative effect** on the predicted number of bikes.
- The weight of the working day feature is close to zero and zero is included in the 95% interval $\longrightarrow$ the effect is **not** statistically significant.
- Some *confidence intervals* are *very short* and the estimates are *close to zero*, yet the *feature effects were statistically significant* $\longrightarrow$ temperature.
- The problem: <u>the features are measured on different scales</u>.
- The solution: *scaling the features (zero mean and standard deviation of one) before fitting the linear model*.

Visual Interpretation: Effect Plot (1/3)

Visual Interpretation: Effect Plot (1/3)

- The **weights** of the linear regression model can be more meaningfully analyzed *when they are multiplied by the actual feature values*.

Visual Interpretation: Effect Plot (1/3)

- The **weights** of the linear regression model can be more meaningfully analyzed *when they are multiplied by the actual feature values*.
- The **effect plot** can help you understand how much *the combination of weight and feature contributes to the predictions* in your data.

Visual Interpretation: Effect Plot (1/3)

- The **weights** of the linear regression model can be more meaningfully analyzed *when they are multiplied by the actual feature values*.
- The **effect plot** can help you understand how much *the combination of weight and feature contributes to the predictions* in your data.
- The effect of feature x_j for instance i is computed as

Visual Interpretation: Effect Plot (1/3)

- The **weights** of the linear regression model can be more meaningfully analyzed *when they are multiplied by the actual feature values*.
- The **effect plot** can help you understand how much *the combination of weight and feature contributes to the predictions* in your data.
- The effect of feature x_j for instance i is computed as

$$\text{effect}_j^{(i)} = \theta_j \times x_j^{(i)}$$

Visual Interpretation: Effect Plot (1/3)

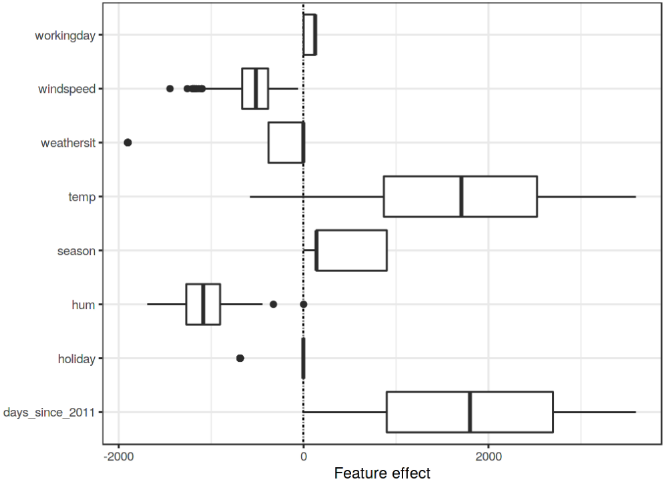
- The **weights** of the linear regression model can be more meaningfully analyzed *when they are multiplied by the actual feature values*.
- The **effect plot** can help you understand how much *the combination of weight and feature contributes to the predictions* in your data.
- The effect of feature x_j for instance i is computed as

$$\text{effect}_j^{(i)} = \theta_j \times x_j^{(i)}$$

- The effects can be visualized with **boxplots**.

Visual Interpretation: Effect Plot (2/3)

Visual Interpretation: Effect Plot (2/3)



The *feature effect* plot shows the distribution of effects (= feature value $\times$ feature weight) across the data per feature

- The **largest contributions** to the expected number of rented bicycles comes from the and , which captures the trend of bike rentals over time.

Visual Interpretation: Effect Plot (3/3)

- The **largest contributions** to the expected number of rented bicycles comes from the **temperature feature** and , which captures the trend of bike rentals over time.

Visual Interpretation: Effect Plot (3/3)

- The **largest contributions** to the expected number of rented bicycles comes from the **temperature feature** and **the days feature**, which captures the trend of bike rentals over time.

Visual Interpretation: Effect Plot (3/3)

- The **largest contributions** to the expected number of rented bicycles comes from the **temperature feature** and **the days feature**, which captures the trend of bike rentals over time.
- **The temperature** has *a broad range* of how much it contributes to the prediction.

Visual Interpretation: Effect Plot (3/3)

- The **largest contributions** to the expected number of rented bicycles comes from the **temperature feature** and **the days feature**, which captures the trend of bike rentals over time.
- **The temperature** has *a broad range* of how much it contributes to the prediction.
- The **day trend** feature goes from *zero* to *large positive contributions*, because the first day in the dataset (01.01.2011) has a *very small trend effect* and the estimated weight for this feature is *positive* (4.93).

Visual Interpretation: Effect Plot (3/3)

- The **largest contributions** to the expected number of rented bicycles comes from the **temperature feature** and **the days feature**, which captures the trend of bike rentals over time.
- **The temperature** has *a broad range* of how much it contributes to the prediction.
- The **day trend** feature goes from *zero* to *large positive contributions*, because the first day in the dataset (01.01.2011) has a *very small trend effect* and the estimated weight for this feature is *positive* (4.93).
  This means that the effect $\uparrow$ with each day and is highest for the *last day* in the dataset (31.12.2012).

Explain Individual Predictions (1/4)

Explain Individual Predictions (1/4)

How much has each feature of an instance contributed to the prediction?

Explain Individual Predictions (1/4)

How much has each feature of an instance contributed to the prediction?
This can be answered by computing the effects for this instance.

Explain Individual Predictions (1/4)

How much has each feature of an instance contributed to the prediction?
This can be answered by computing the effects for this instance.

Feature	Value
season	SPRING
yr	2011
mnth	JAN
holiday	NO HOLIDAY
weekday	THU
workingday	WORKING DAY
weathersit	GOOD
temp	1.604356
hum	51.8261
windspeed	6.000868
cnt	1606
days_since_2011	5

The 6th instance from the bicycle dataset

Explain Individual Predictions (2/4)

To obtain the feature effects of the 6th instance, we have to

To obtain the feature effects of the 6th instance, we have to *multiply its feature values by the corresponding weights* from the linear regression model.

To obtain the feature effects of the 6th instance, we have to *multiply its feature values by the corresponding weights* from the linear regression model.

- For the value "WORKING DAY" of feature "workingday", the effect is 124.9 (124.9×1).

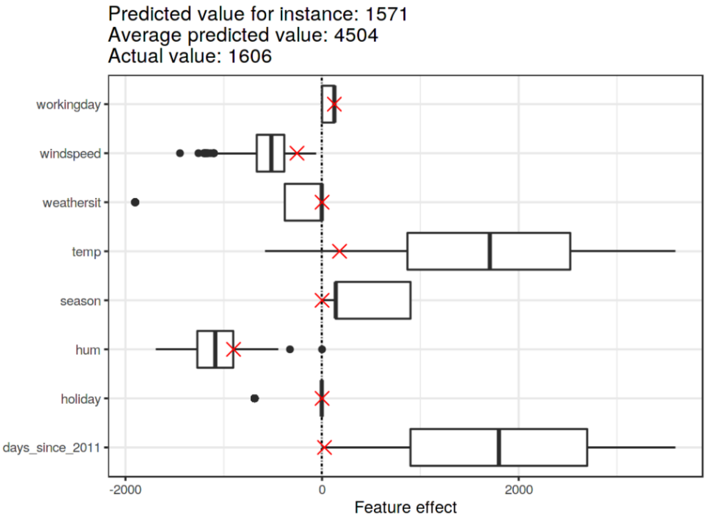
Explain Individual Predictions (2/4)

To obtain the feature effects of the 6th instance, we have to *multiply its feature values by the corresponding weights* from the linear regression model.

- For the value "WORKING DAY" of feature "workingday", the effect is 124.9 (124.9×1).
- For a temperature of 1.6 degrees Celcius, the effect is 177.6 (1.604356×110.7096).

Explain Individual Predictions (3/4)

Explain Individual Predictions (3/4)



The effect plot for the 6-th instance is labeled cross $\times$

Explain Individual Predictions (4/4)

Explain Individual Predictions (4/4)

- If we average the predictions for the training data instances, we get an average of 4504.

Explain Individual Predictions (4/4)

- If we average the predictions for the training data instances, we get an average of 4504.
- In comparison, the prediction of the 6th instance is <u>small</u>, since only 1571 bicycle rents are predicted.

- If we average the predictions for the training data instances, we get an average of 4504.
- In comparison, the prediction of the 6th instance is small, since only 1571 bicycle rents are predicted.

- Why?

Explain Individual Predictions (4/4)

- If we average the predictions for the training data instances, we get an average of 4504.
- In comparison, the prediction of the 6th instance is small, since only 1571 bicycle rents are predicted.
- Why?
  - The 6th instance has a *low temperature effect* because on this day the temperature was 2 degrees, which is low compared to most other days.

Explain Individual Predictions (4/4)

- If we average the predictions for the training data instances, we get an average of 4504.
- In comparison, the prediction of the 6th instance is small, since only 1571 bicycle rents are predicted.
- Why?
  - The 6th instance has a *low temperature effect* because on this day the temperature was 2 degrees, which is low compared to most other days.
  - The effect of the *trend feature* "days_since_2011" is small compared to the other data instances because this instance is from early 2011 (5 days) and the *trend feature* also has a positive weight.

Outline

Conclusion

Conclusion

- We review that a machine learning algorithm can be decomposed into *four components*.

Conclusion

- We review that a machine learning algorithm can be decomposed into *four components*.
- For the coming future, the focus will be on **model-agnostic interpretability tools**.

Conclusion

- We review that a machine learning algorithm can be decomposed into *four components*.
- For the coming future, the focus will be on **model-agnostic interpretability tools**.
  Please read **?**: ***Model-Agnostic Methods***.

Conclusion

- We review that a machine learning algorithm can be decomposed into *four components*.
- For the coming future, the focus will be on **model-agnostic interpretability tools**.
  Please read **?**: ***Model-Agnostic Methods***.
- Machine Learning will be **automated** and, with it, **interpretability**.

Conclusion

- We review that a machine learning algorithm can be decomposed into *four components*.
- For the coming future, the focus will be on **model-agnostic interpretability tools**.
  Please read **?**: ***Model-Agnostic Methods***.
- Machine Learning will be **automated** and, with it, **interpretability**.
- We do not analyze data, we **analyze models.**

Conclusion

- We review that a machine learning algorithm can be decomposed into *four components*.
- For the coming future, the focus will be on **model-agnostic interpretability tools**.
  Please read **?**: ***Model-Agnostic Methods***.
- Machine Learning will be **automated** and, with it, **interpretability**.
- We do not analyze data, we **analyze models.**
- Robots and programs will explain themselves. For example:

Conclusion

- We review that a machine learning algorithm can be decomposed into *four components*.
- For the coming future, the focus will be on **model-agnostic interpretability tools**.
  Please read **?**: ***Model-Agnostic Methods***.
- Machine Learning will be **automated** and, with it, **interpretability**.
- We do not analyze data, we **analyze models.**
- Robots and programs will explain themselves. For example:
  - A self-driving car that reports why it stopped abruptly ("70% probability that a kid will cross the road");

Conclusion

- We review that a machine learning algorithm can be decomposed into *four components*.
- For the coming future, the focus will be on **model-agnostic interpretability tools**.
  Please read **?**: ***Model-Agnostic Methods***.
- Machine Learning will be **automated** and, with it, **interpretability**.
- We do not analyze data, we **analyze models.**
- Robots and programs will explain themselves. For example:
  - A self-driving car that reports why it stopped abruptly ("70% probability that a kid will cross the road");
  - A credit default program that explains to a bank employee why a credit application was rejected ("Applicant has too many credit cards and is employed in an unstable job.");

Conclusion

- We review that a machine learning algorithm can be decomposed into *four components*.
- For the coming future, the focus will be on **model-agnostic interpretability tools**.
  Please read **?**: ***Model-Agnostic Methods***.
- Machine Learning will be **automated** and, with it, **interpretability**.
- We do not analyze data, we **analyze models.**
- Robots and programs will explain themselves. For example:
  - A self-driving car that reports why it stopped abruptly ("70% probability that a kid will cross the road");
  - A credit default program that explains to a bank employee why a credit application was rejected ("Applicant has too many credit cards and is employed in an unstable job.");
  - A robot arm that explains why it moved the item from the conveyor belt into a trash bin ("The item has a craze at the bottom.")

Conclusion

- We review that a machine learning algorithm can be decomposed into *four components*.
- For the coming future, the focus will be on **model-agnostic interpretability tools**.
  Please read **?**: ***Model-Agnostic Methods***.
- Machine Learning will be **automated** and, with it, **interpretability**.
- We do not analyze data, we **analyze models.**
- Robots and programs will explain themselves. For example:
  - A self-driving car that reports why it stopped abruptly ("70% probability that a kid will cross the road");
  - A credit default program that explains to a bank employee why a credit application was rejected ("Applicant has too many credit cards and is employed in an unstable job.");
  - A robot arm that explains why it moved the item from the conveyor belt into a trash bin ("The item has a craze at the bottom.")
- Lastly, Interpretability could boost **machine intelligence research**.

Daftar Pustaka I

hendra.bunyamin@it.maranatha.edu