

DIGITAL TALENT SCHOLARSHIP 2019



Program Fresh Graduate Academy Digital Talent Scholarship 2019 | Machine Learning

Classification : Logistic Regression

Nama pembicara dengan gelar





Bagian Pertama

Pendahuluan Mengenai Logistic Regression

Apa itu Logistic Regression

- Logistic Regression merupakan salah satu teknik machine learning untuk melakukan klasifikasi record dari dataset.
- Sebagai contoh, kita memiliki dataset telekomunikasi sebagai berikut.

| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
|---|--------|------|---------|--------|-----|--------|-------|----------|----------|-------|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | Yes |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | Yes |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | No |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | No |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | ? |

Pemahaman Data

| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
|---|--------|------|---------|--------|-----|--------|-------|----------|----------|-------|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | Yes |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | Yes |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | No |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | No |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | ? |

- Bayangkan Anda adalah seorang analis di perusahaan ini dan Anda harus mencari tahu siapa yang pergi dan mengapa.
- Anda harus menggunakan dataset untuk membangun model berdasarkan catatan-catatan sebelumnya dan menggunakannya untuk memprediksi “churn” di masa depan.
 - Churn = Apakah pelanggan meninggalkan perusahaan atau tidak bulan lalu.

Pemahaman Data

Independent Variable

Dependent Variable

| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
|---|--------|------|---------|--------|-----|--------|-------|----------|----------|-------|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | Yes |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | Yes |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | No |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | No |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | ? |

- Independent Variable = Variable / Fitur yang merupakan input dan akan dipakai untuk memprediksi sebuah output, *churn*.
- Dependent Variable = Nilainya bergantung pada nilai-nilai input
 - Pelanggan akan berhenti atau tidak bergantung dari data pelanggan tsb.

Linear vs. Logistic Regression

Linear Regression

- Melakukan Prediksi
- Prediksi nilai kontinyu dari sebuah variable, seperti:
 - Harga rumah berdasarkan ciri
 - Tekanan darah berdasarkan symptom
 - Konsumsi bensin berdasarkan kondisi mobil

Logistic Regression

- Melakukan Klasifikasi
- Klasifikasi nilai biner, seperti:
 - Kelompok A atau B
 - Sukses atau tidak sukses
 - Tetap berlangganan atau tidak.

| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
|---|--------|------|---------|--------|-----|--------|-------|----------|----------|-------|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | Yes |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | Yes |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | No |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | No |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | ? |

Catatan Khusus Logistic Regression

Independent Variable

Dependent Variable

| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
|---|--------|------|---------|--------|-----|--------|-------|----------|----------|-------|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | Yes |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | Yes |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | No |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | No |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | ? |

Numeric and Continuous Value

- Logistic Regression mewajibkan seluruh data dalam bentuk numerik
- Jika berkategori (Pria/Wanita, Ya/Tidak) harus diubah dalam bentuk angka.



Bagian Dua

Aplikasi Logistic Regression

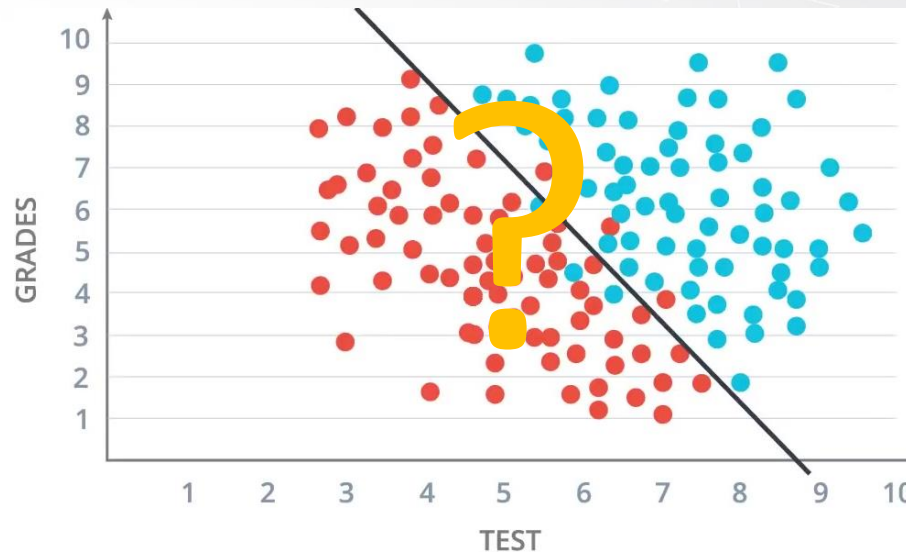
Beberapa Contoh Aplikasi

- Memprediksi probabilitas seseorang mengalami serangan jantung dalam satu periode tertentu
 - Berdasarkan: umur, sex, berat badan.
- Memprediksi apakah pasien memiliki penyakit yang dicurigai (seperti diabetes)
 - Berdasarkan: berat, tinggi, tekanan darah, dan beragam test darah lainnya.
- Memprediksi kemungkinan pelanggan akan membeli sebuah produk, atau berlangganan sebuah layanan (seperti contoh kita sebelumnya)
 - Berdasarkan: umur, sex, pekerjaan, lingkungan hidup.

Beberapa Contoh Aplikasi

- Memprediksi probabilitas kegagalan sebuah produk untuk menghindari kekecewaan pelanggan.
 - Berdasarkan: tingkat ketahanan produk, durabilitas, dll.
- Memprediksi apakah nasabah dapat menyanggupi pembayaran kredit.
 - Berdasarkan: umur, sex, pekerjaan, jumlah anak, gaji, dll.
- Berdasarkan beberapa contoh diatas, dapat disimpulkan bahwa: Logistic Regression digunakan untuk menghitung probabilitas sebuah data terkategori ke salah satu kelompok yang tersedia.

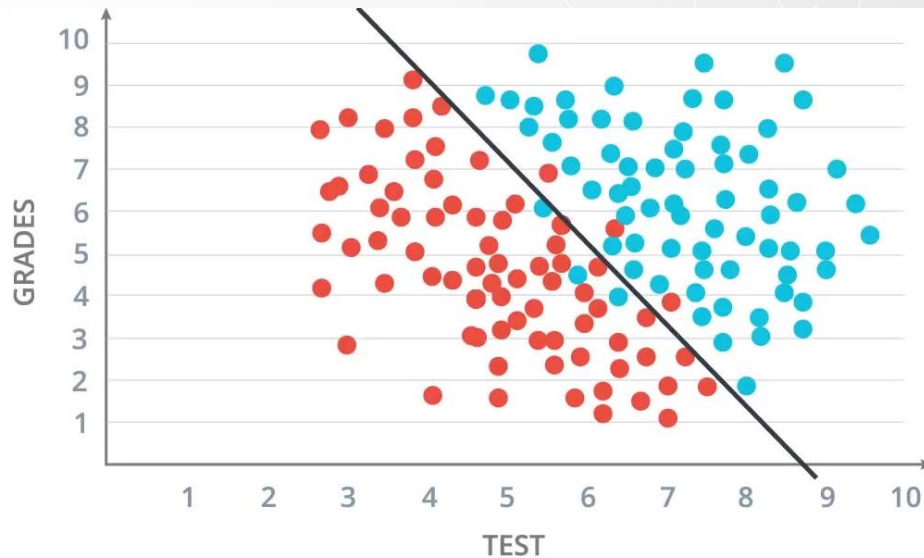
Kapan kita gunakan Logistic Regression?



- Pada dasarnya, ada beragam teknik machine learning yang dapat digunakan untuk melakukan kategorisasi suatu data.
- Pertanyaan mendasar muncul: Kapan kita harus menggunakan Logistic Regression?

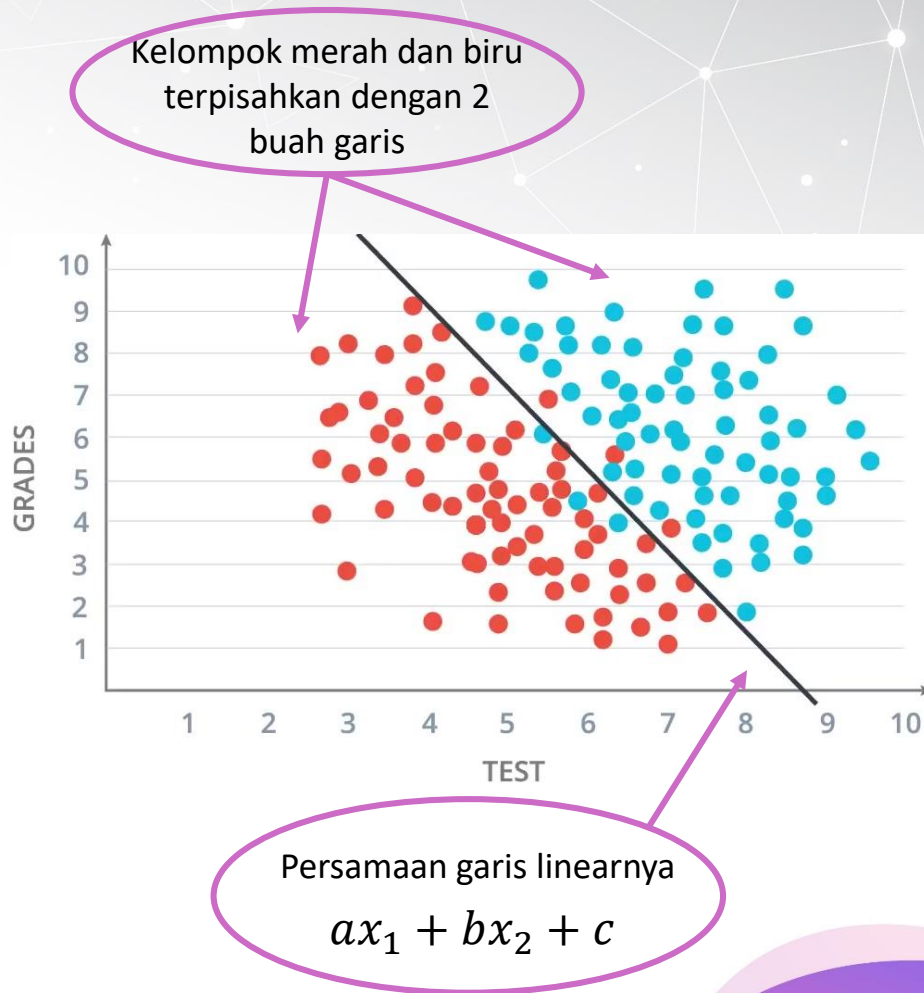
Kapan kita gunakan Logistic Regression?

- Jika data berupa binary, seperti:
 - Kelompok A atau B
 - Lulus atau Tidak
 - Berlangganan atau Tidak
- Jika kita membutuhkan pengelompokan dalam bentuk probabilitas
- Data bersifat “*linearly separable*”



Kapan kita gunakan Logistic Regression

- Linearly Separable
- Dapat dipisahkan secara linear
 - Jika data 2D, dipisahkan garis
 - Jika data 3D, dipisahkan plane
 - Jika data >3D, dipisahkan hyper-plane.
- Secara teori, Logistic Regression sebenarnya juga dapat digunakan untuk data yang bersifat “non-linearly separable”
 - Namun diluar dari pembahasan ini.



Memodelkan Logistic Regression

| X | | | | | | | | | | y |
|----------|--------|------|---------|--------|-----|--------|-------|----------|----------|----------|
| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | Yes |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | Yes |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | No |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | No |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | ? |

Memodelkan Logistic Regression

| X | | | | | | | | | | y |
|----------|--------|------|---------|--------|-----|--------|-------|----------|----------|----------|
| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | 1.0 |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | 0.0 |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | ? |

Data berkategori menjadi data numerik

$$\mathbf{x} \in \mathbb{R}^{m \times n}$$

$$\mathbf{y} \in \{0,1\}$$

$$\hat{\mathbf{y}} = P(y = 1|x)$$



Bagian Satu

Mengingat Kembali Formulasi

Meninjau Kembali Data

X **y**

| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
|----------|--------|------|---------|--------|-----|--------|-------|----------|----------|-------|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | Yes |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | Yes |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | No |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | No |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | ? |

- Tujuan dari Logistic Regression adalah untuk membangun sebuah model yang akan melakukan klasifikasi class setiap pelanggan.
 - Menentukan probabilitas pelanggan apakah masuk dalam kategori berlangganan atau tidak.

Meninjau Kembali Formulasi

X **y**

| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
|----------|--------|------|---------|--------|-----|--------|-------|----------|----------|-------|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | Yes |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | Yes |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | No |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | No |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | ? |

$$\hat{y} = P(y = 1 | x)$$

Membuat
sebuah model

Yang dapat mengestimasi
probabilitas classnya, \hat{y}

apakah masuk
dalam class 1

jika diberikan
observasi data x

Meninjau Kembali Formulasi

X **y**

| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
|----------|--------|------|---------|--------|-----|--------|-------|----------|----------|-------|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | Yes |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | Yes |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | No |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | No |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | ? |

$$\hat{y} = P(y = 1 | x)$$

\hat{y} , nilai yang diprediksi

y , nilai yang sebenarnya

Bagian Dua



Mencoba Menyelesaikan Permasalahan Kategorisasi dengan Linear Regression

Prediksi Income Berdasarkan Age

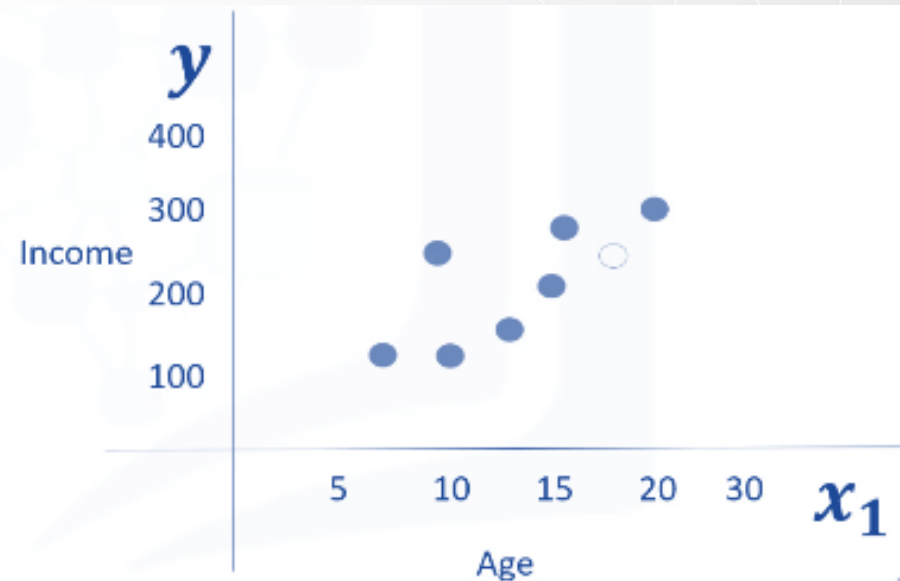
| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
|---|--------|------|---------|--------|-----|--------|-------|----------|----------|-------|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | 1 |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1 |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0 |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | 0 |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | 0 |

- Kita lupakan sejenak mengenai prediksi kategori *churn*, dan asumsikan tujuan kita adalah melakukan prediksi pendapatannya pelanggan.
- Untuk simplisitas, kita hanya ambil *age* (umur) sebagai variable yang akan mempengaruhi *income*
- Independent variable (x) = Age
- Dependent variable (y) = Income.

Prediksi Income Berdasarkan Age

| | tenure | age | address | income | ed | employ | equip | calldata | wireless | churn |
|---|--------|------|---------|--------|-----|--------|-------|----------|----------|-------|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | 1 |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1 |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0 |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | 0 |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | 0 |

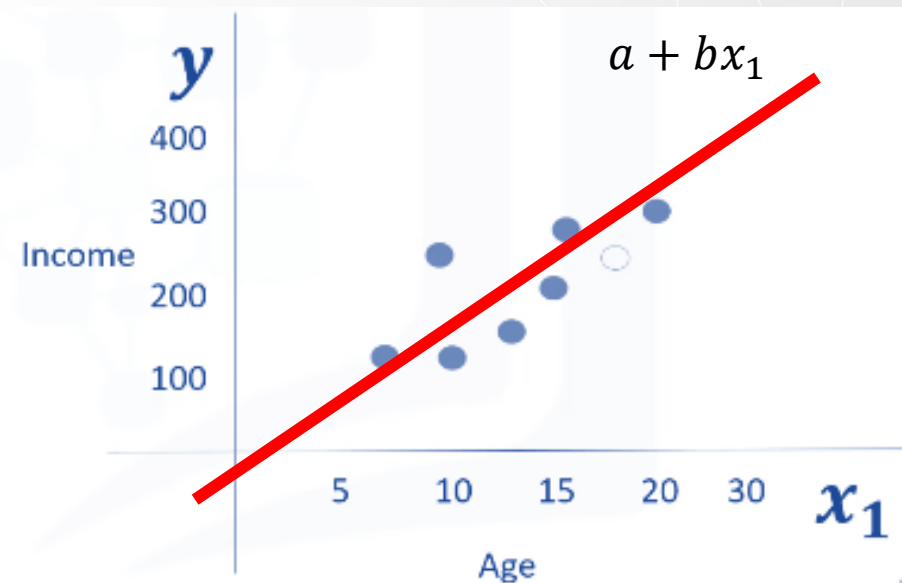
- Untuk memahami data, kita bisa lakukan plotting terlebih dahulu.
- Age, independent variable, sebagai sumbu x
- Income, dependent variable, sebagai sumbu y



Prediksi Income Berdasarkan Age

| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
|---|--------|------|---------|--------|-----|--------|-------|----------|----------|-------|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | 1 |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1 |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0 |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | 0 |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | 0 |

- Dengan Linear Regression kita dapat menyesuaikan sebuah garis yang merepresentasikan tren data.
- Kita dapat menemukan garis ini melalui training, atau menghitungnya secara matematis.
- Garis dapat diexpresikan dengan $a + bx_1$



Prediksi Churn Berdasarkan Age

| | tenure | age | address | income | ed | employ | equip | calldcard | wireless | churn |
|---|--------|------|---------|--------|-----|--------|-------|-----------|----------|-------|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | 1 |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1 |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0 |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | 0 |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | 0 |

- Sekarang mari kita ganti permasalahannya.
- Dengan teknik yang sama (linear regression), apakah kita bisa memprediksi kategori dari “*churn*”?
- Independent variable (x) = Age
- Dependent variable (y) = Churn.
- Mari kita diskusi bersama.

Prediksi Churn Berdasarkan Age

| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
|---|--------|------|---------|--------|-----|--------|-------|----------|----------|-------|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | 1 |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1 |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0 |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | 0 |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | 0 |

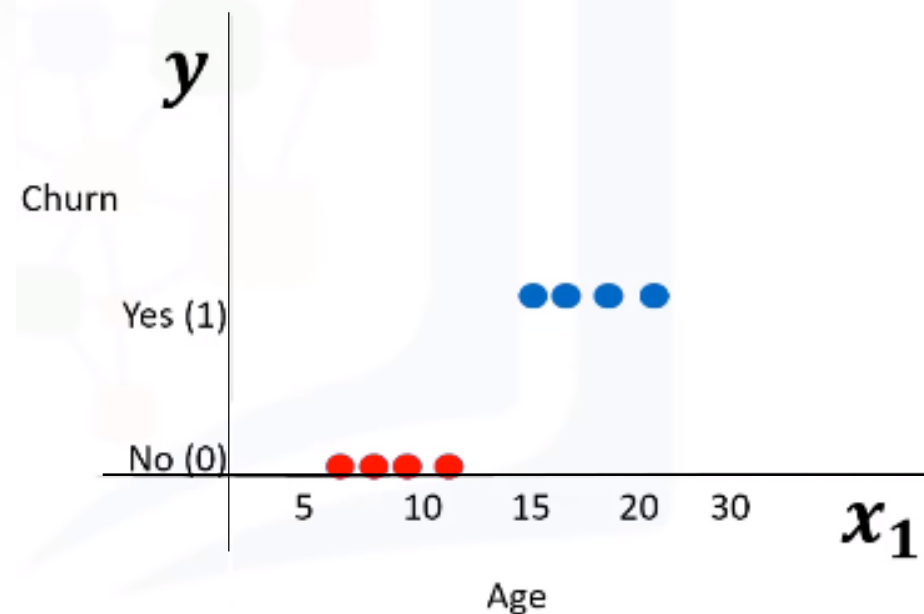
- Untuk memahami data, kita bisa lakukan plotting terlebih dahulu.
- *Age*, independent variable, sebagai sumbu x
- *Churn*, dependent variable, sebagai sumbu y



Prediksi Churn Berdasarkan Age

| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
|---|--------|------|---------|--------|-----|--------|-------|----------|----------|-------|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | 1 |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1 |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0 |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | 0 |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | 0 |

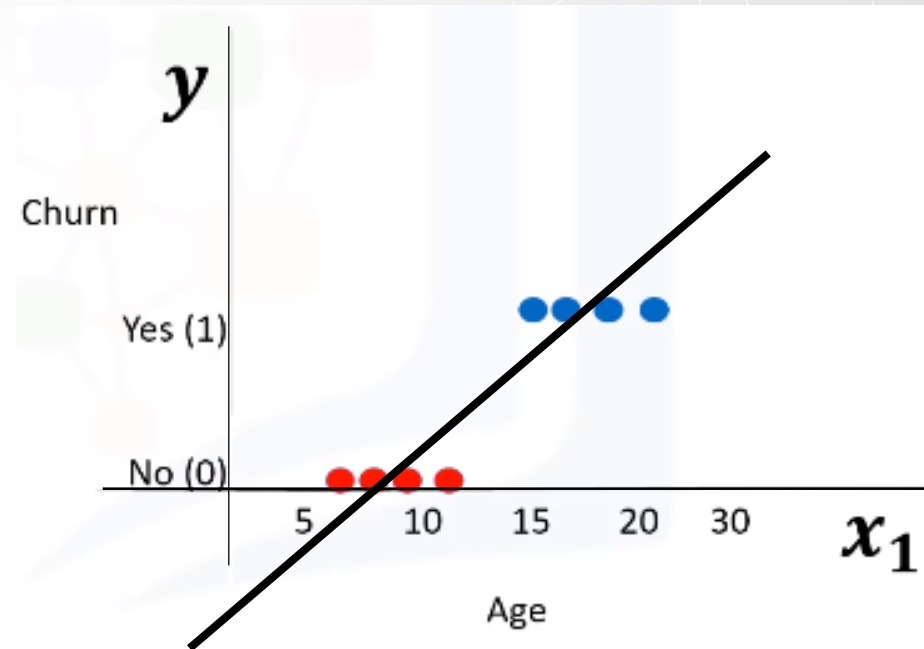
- Dapat kita lihat ada perbedaan mencolok dari plot yang kita buat.
- Sumbu y , *churn*, merupakan data berkategori.
 - Ya dan Tidak.
- Data tidak menjadi kontinyu



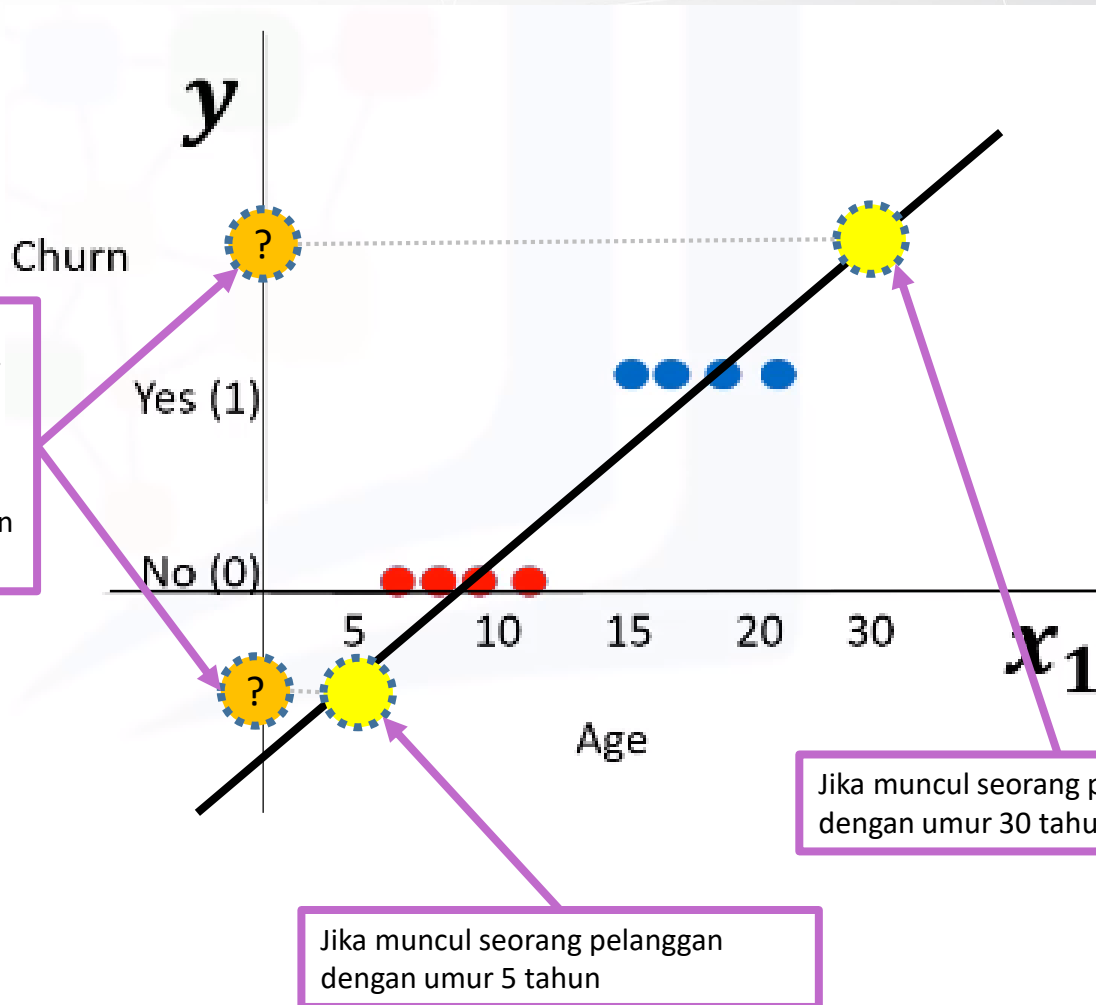
Prediksi Churn Berdasarkan Age

| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
|---|--------|------|---------|--------|-----|--------|-------|----------|----------|-------|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | 1 |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1 |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0 |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | 0 |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | 0 |

- Demi mengunggah rasa penasaran kita, mari kita coba (secara paksa) untuk melakukan estimasi persamaan garis.
- Kita dapat melihat bahwa persamaan garis menjadi tidak relevan.



Linear Regression Menjadi Tidak Relevan



Linear Regression Menjadi Tidak Relevan

- Dari percobaan sebelumnya, kita memahami bahwa meskipun kita memaksakan diri melakukan linear regression pada permasalahan kategorisasi, hasil yang didapat akan menjadi tidak relevan.
- Linear Regresion hanya untuk memprediksi! Bukan melakukan kategorisasi.
- Logistic Regresion sebaliknya, ditujukan untuk melakukan kategorisasi!



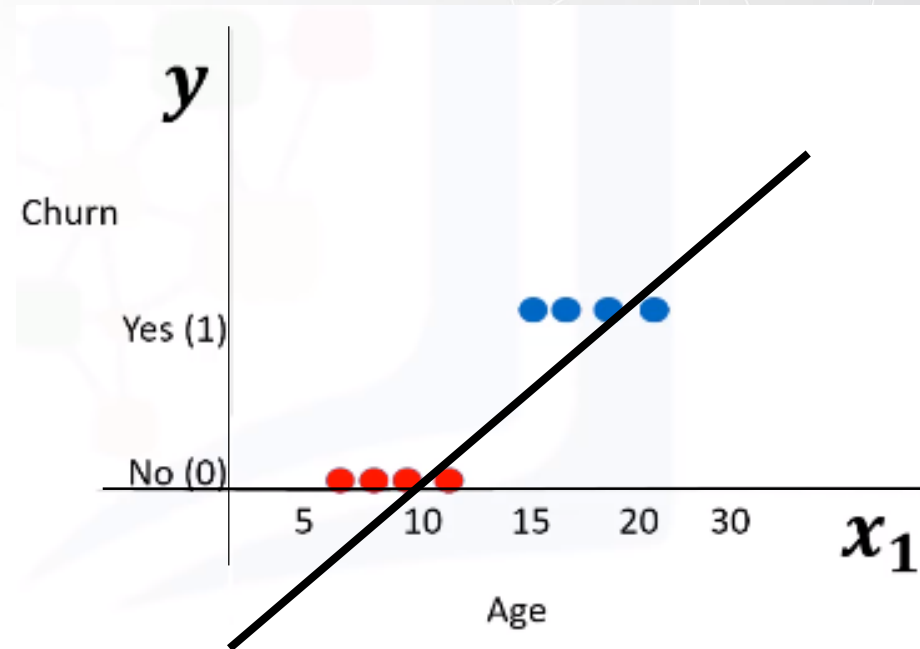
Bagian Tiga

Linear Regression menuju Logistic Regression

Prediksi Churn Berdasarkan Age

| | tenure | age | address | income | ed | employ | equip | callcard | wireless | churn |
|---|--------|------|---------|--------|-----|--------|-------|----------|----------|-------|
| 0 | 11.0 | 33.0 | 7.0 | 136.0 | 5.0 | 5.0 | 0.0 | 1.0 | 1.0 | 1 |
| 1 | 33.0 | 33.0 | 12.0 | 33.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1 |
| 2 | 23.0 | 30.0 | 9.0 | 30.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0 |
| 3 | 38.0 | 35.0 | 5.0 | 76.0 | 2.0 | 10.0 | 1.0 | 1.0 | 1.0 | 0 |
| 4 | 7.0 | 35.0 | 14.0 | 80.0 | 2.0 | 15.0 | 0.0 | 1.0 | 0.0 | 0 |

- Faktanya, untuk mengimplementasikan Logistic Regression, kita hanya perlu menambahkan beberapa tahap tambahan.
- Untuk itu, mari kita notasikan secara formal beberapa variable yang kita butuhkan.



Menuju Logistic Regression

Persamaan Garis

$$y = ax + c$$

$$ax + by + c = 0$$

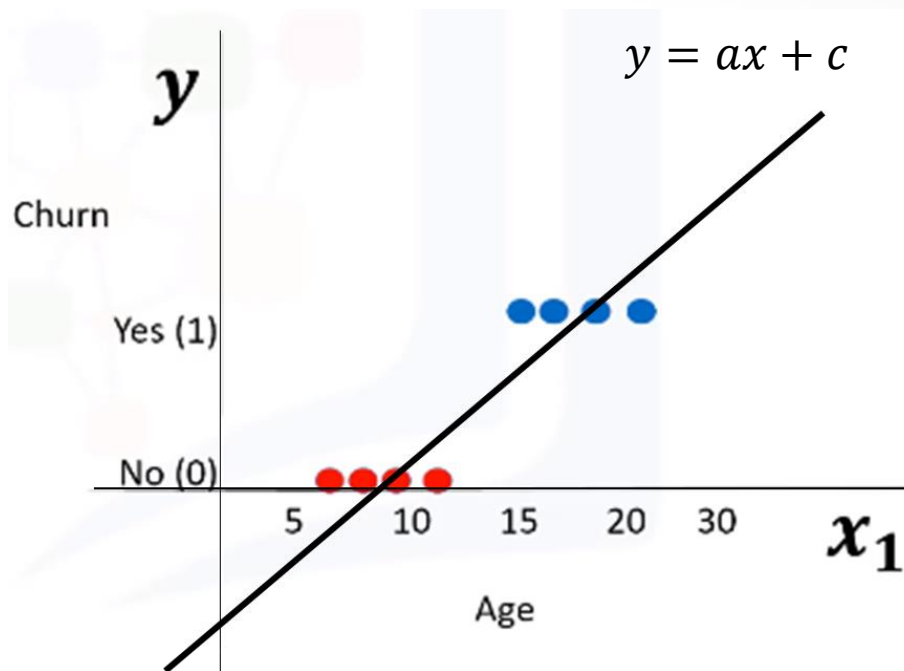
$$w_1x_1 + w_2x_2 + b = 0$$

$$w_1x_1 + w_2x_2 + w_31 = 0$$

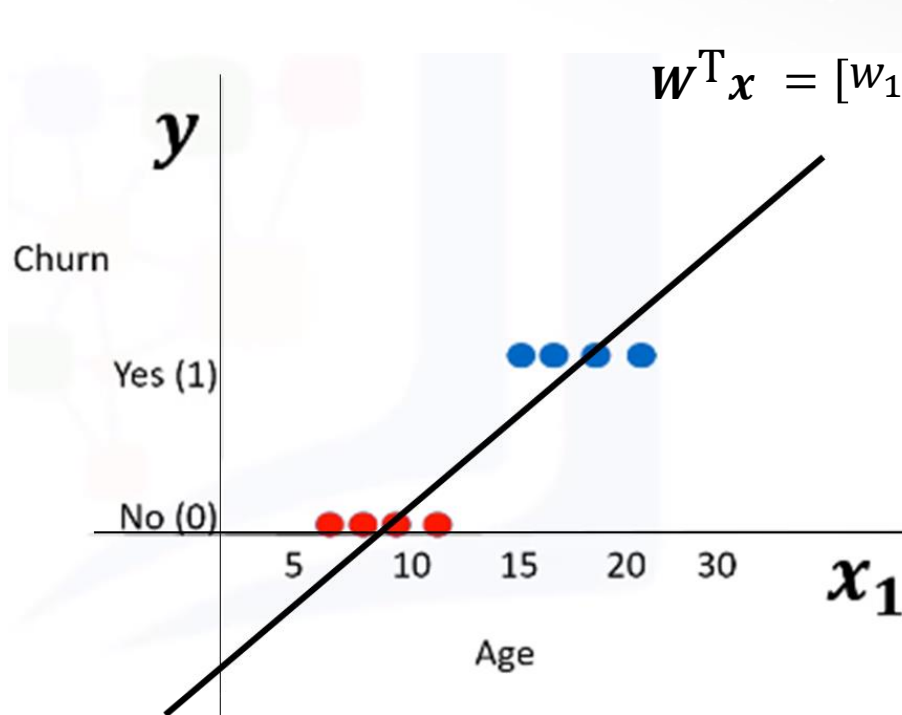
$$W^T x = 0$$

dimana : $W = [w_1 \quad w_2 \quad w_3]$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix}$$



Menuju Logistic Regression



Contoh, diketahui parameter garis:

$$W^T = [0.1 \quad -1 \quad -1]$$

$$W^T x = 0.1x_1 - 1x_2 - 1$$

Untuk mempermudah pemahaman

$$x_2 = 0.1x_1 - 1$$

$$y = 0.1x - 1$$

Menuju Logistic Regression

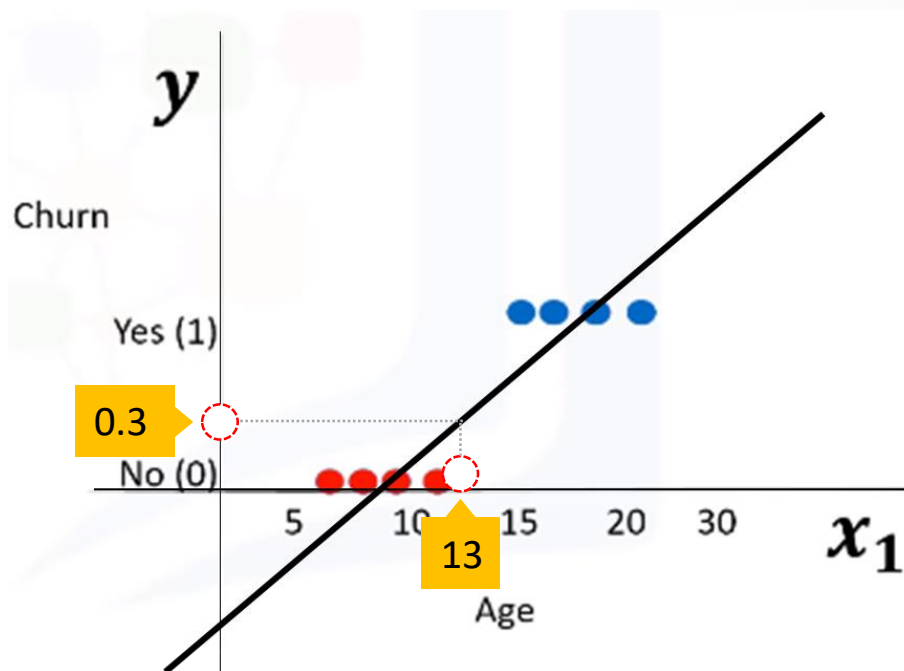
Sekarang, kita bisa menggunakan persamaan garis linear regression tersebut untuk melakukan kategorisasi churn berdasarkan umur.

$$y = 0.1x - 1$$

Sebagai contoh, ada pelanggan berumur 13 tahun, maka:

$$y = 0.1(13) - 1$$

$$y = 0.3$$



Menuju Logistic Regression

Sekarang, mari kita buat sebuah aturan – thresholding:

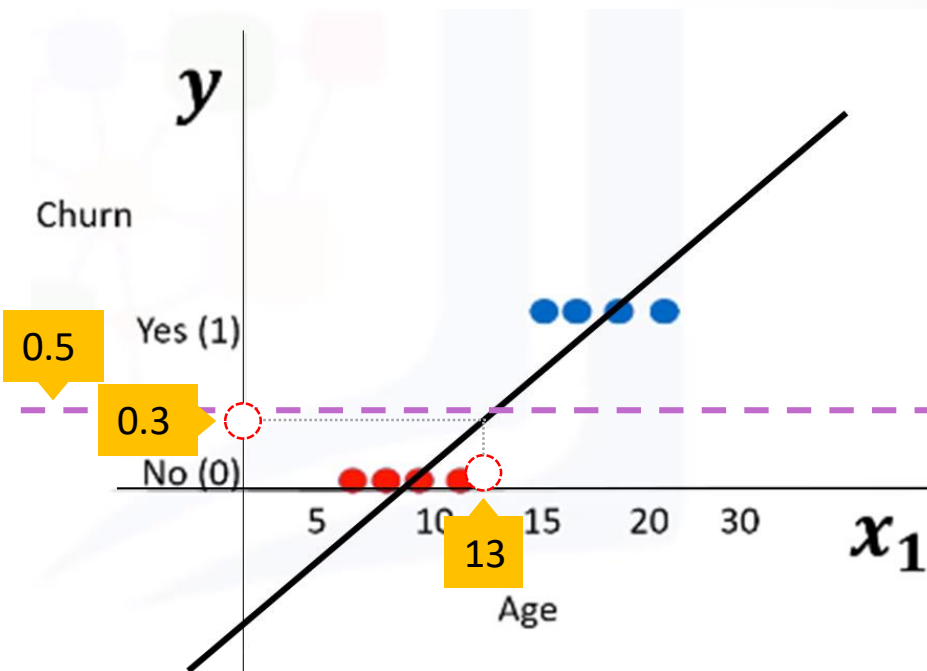
$$\hat{y} = \begin{cases} 1, & Wx \geq 0.5 \\ 0, & Wx < 0.5 \end{cases}$$

Karena

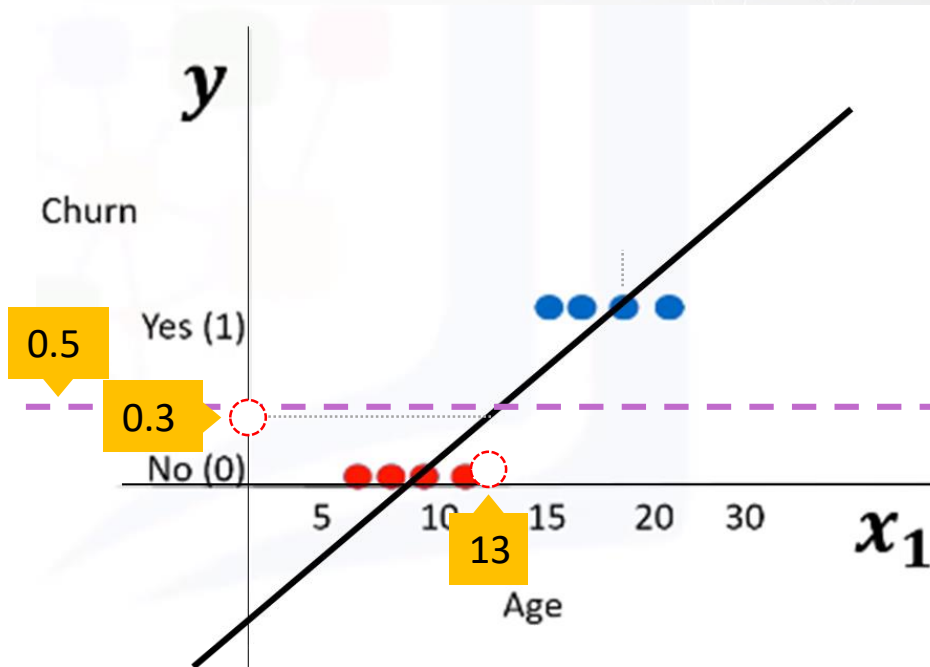
$$y = 0.3, \text{ dan } y < 0.5$$

Maka, \hat{y} terkategori ke class 0, tidak berlangganan.

$$\hat{y} = 0$$



Menuju Logistic Regression

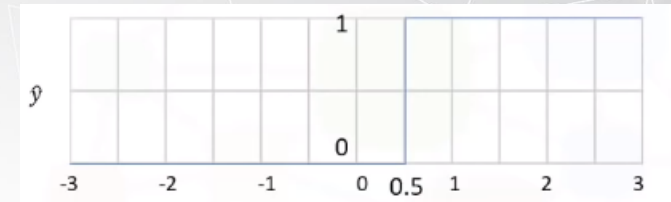


$$\hat{y} = \begin{cases} 1, & W^T x \geq 0.5 \\ 0, & W^T x < 0.5 \end{cases}$$

- Terdapat satu permasalahan yang masih belum disinggung.
- Rangkaian tahap ini tidak memberikan output berupa probabilitas.
- Berapa probabilitas pelanggan berumur 13 tahun berhenti berlangganan?

Yang telah kita lakukan sejauh ini

$$W^T x = w_1 x_1 + w_2 x_2 + w_3 1 \quad \hat{y} = \begin{cases} 1, & W^T x \geq 0.5 \\ 0, & W^T x < 0.5 \end{cases}$$



Gunakan persamaan garis linear regression untuk mengkalkulasi **score**

Setelah score didapatkan, lakukan thresholding untuk hasil score tersebut.

Proses thresholding tersebut merupakan Step Function.

Angka ini bisa saja diluar dari nilai yang diperbolehkan oleh dependent variable

Contoh:

Churn, hanya memiliki nilai 0 dan 1, namun hasil score bisa saja diluar dari angka ini

Contoh:

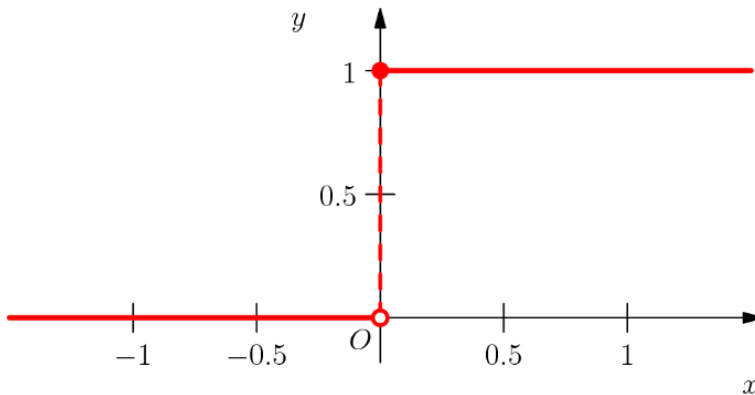
Threshold yang dipakai adalah 0.5, sehingga, jika score berada dibawah 0.5, maka data tersebut terkategori dalam class 0,

Jika ≥ 0.5 , maka data tersebut terkategori dalam class 1.

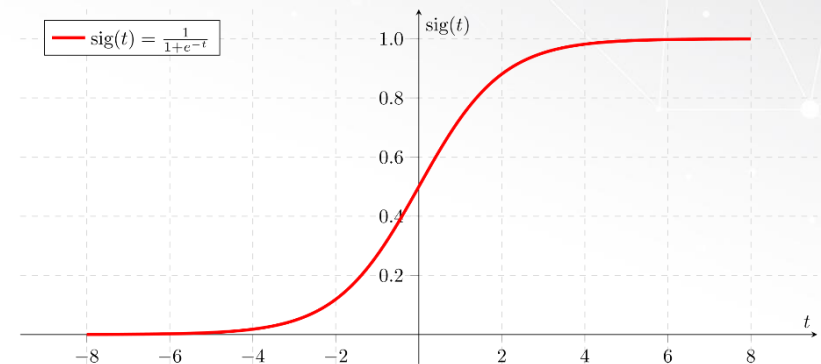
Step Function diilustrasikan dalam grafik diatas.

Menuju Logistic Regression

- Agar mengeluarkan output berupa probabilitas, Step Function harus diganti dengan sebuah fungsi yang lain, yang disebut dengan Logistic Function



Step Function



Logistic Function

$$\hat{y} = \begin{cases} 1, & \mathbf{W}^T \mathbf{x} \geq \text{Threshold} \\ 0, & \mathbf{W}^T \mathbf{x} < \text{Threshold} \end{cases}$$

$$\hat{y} = \sigma(\mathbf{W}^T \mathbf{x})$$

Logistic Function

- Logistic Function juga umum disebut dengan Sigmoid Function.
- Didefinisikan sebagai

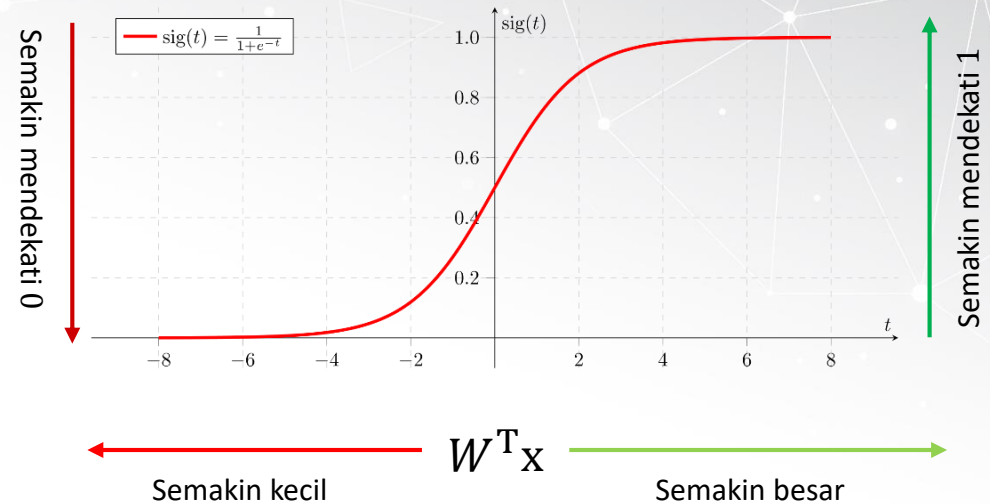
$$\sigma(\hat{y}) = \frac{1}{1 + e^{-\hat{y}}}$$

\hat{y} bernilai besar,
 $e^{-\hat{y}}$ mendekati 0
 $\sigma(\hat{y}) \cong 1$

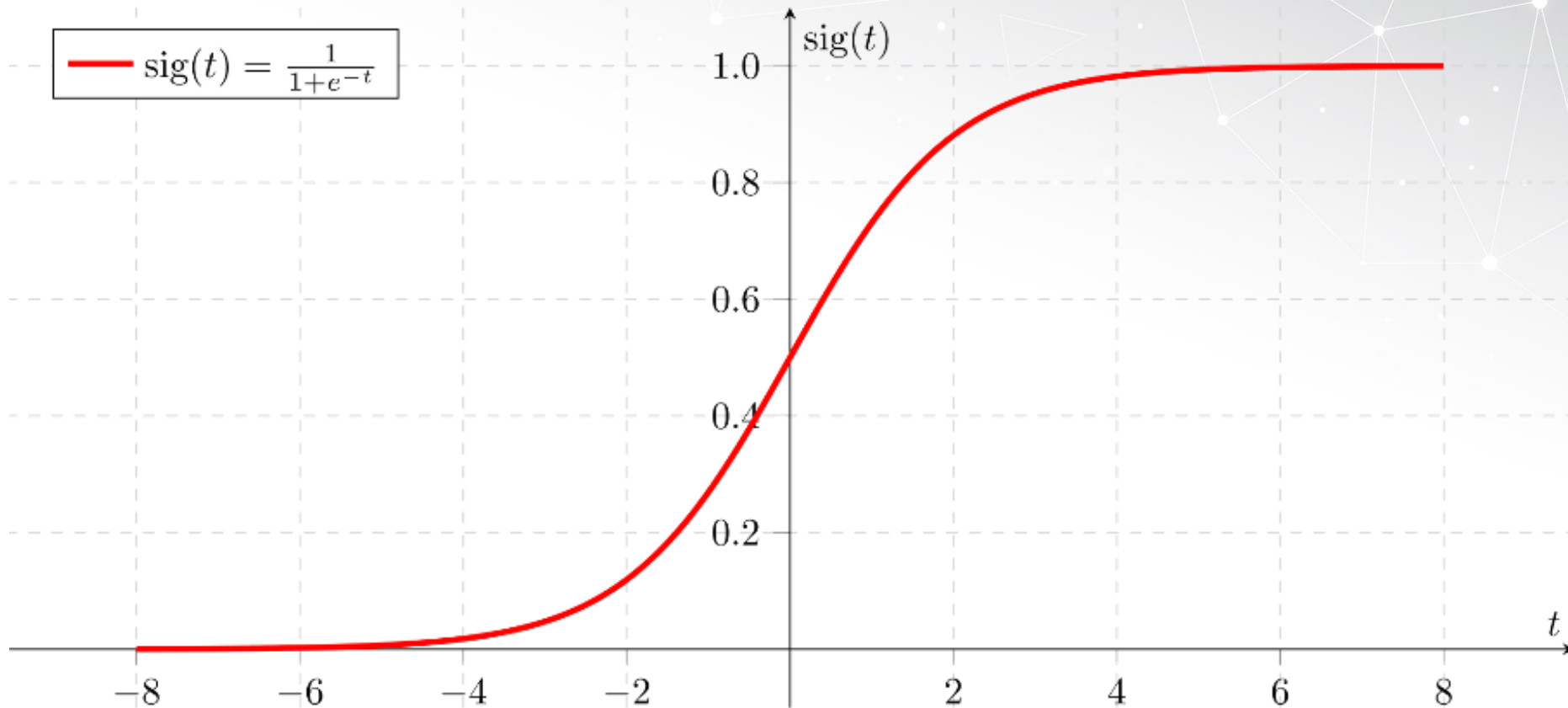
\hat{y} bernilai kecil,
 $e^{-\hat{y}}$ mendekati ∞
 $\sigma(\hat{y}) \cong 0$

- Dimana

$$\begin{aligned}\hat{y} &= W^T X \\ &= w_1 x_1 + w_2 x_2 + w_3 1\end{aligned}$$



Logistic Function



Output dari Contoh Kasus

- Persamaan garis

$$\hat{y} = 0.1x - 1$$

- Pelanggan berumur 13 Tahun

$$\hat{y} = 0.1(13) - 1 = 0.3$$

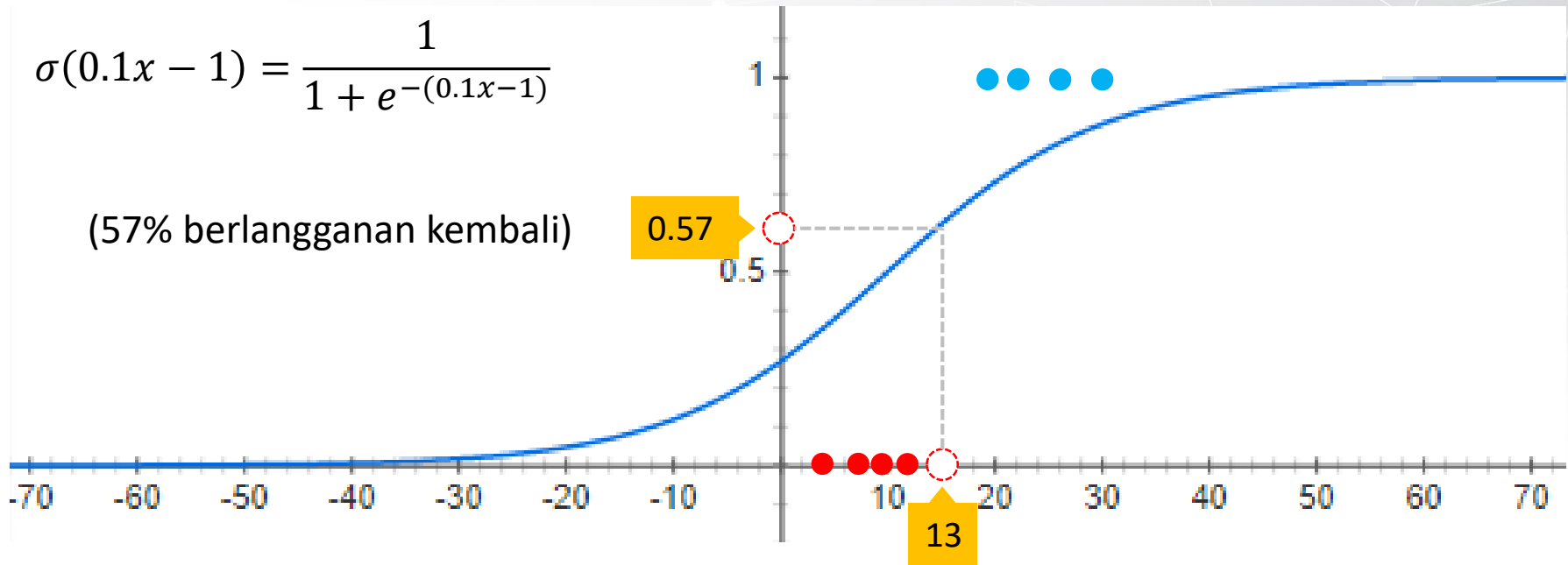
- Probabilitas menggunakan Logistic/Sigmoid Function

$$\sigma(y = 1|x) = \sigma(\hat{y}) = \sigma(0.3) = \frac{1}{1 + e^{-0.3}} = 0.57$$

Output dari Contoh Kasus

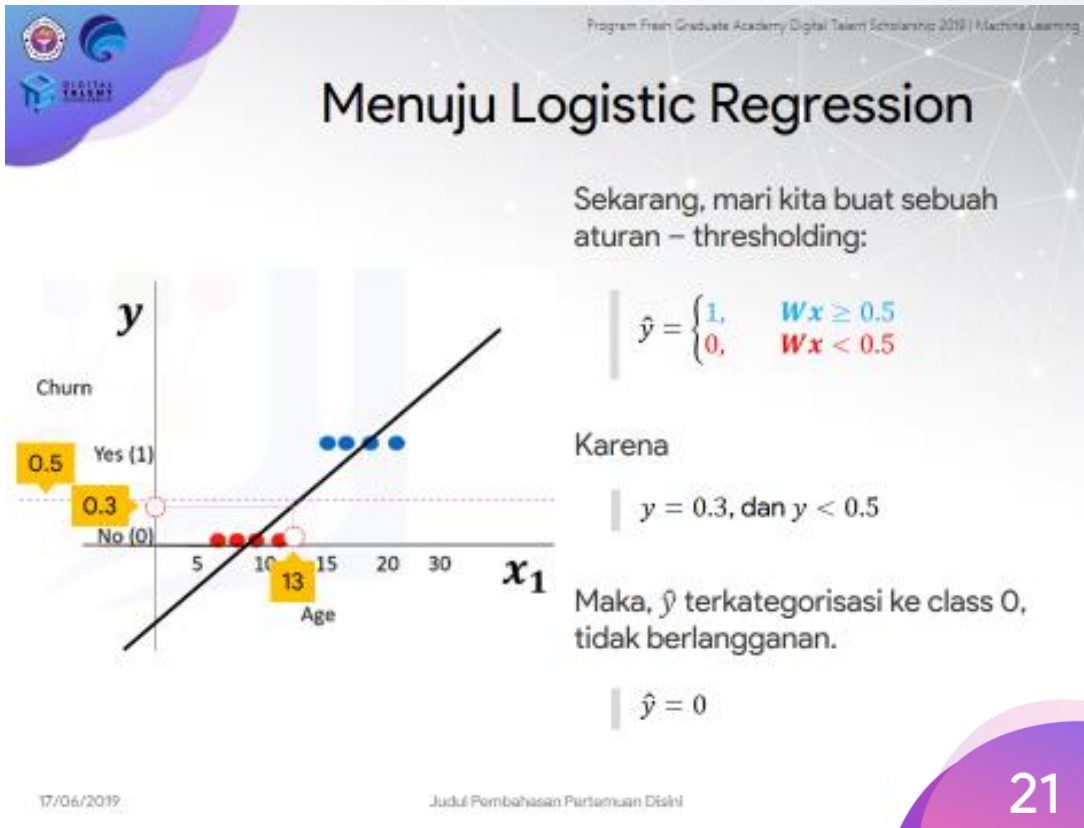
$$\sigma(0.1x - 1) = \frac{1}{1 + e^{-(0.1x - 1)}}$$

(57% berlangganan kembali)



- Sepertinya ada yang tidak benar.
- Jika kita ingat percobaan kita saat menggunakan Step Function, pelanggan berumur 13 tahun tersebut terkategori tidak berlangganan.

Mengingat Kembali

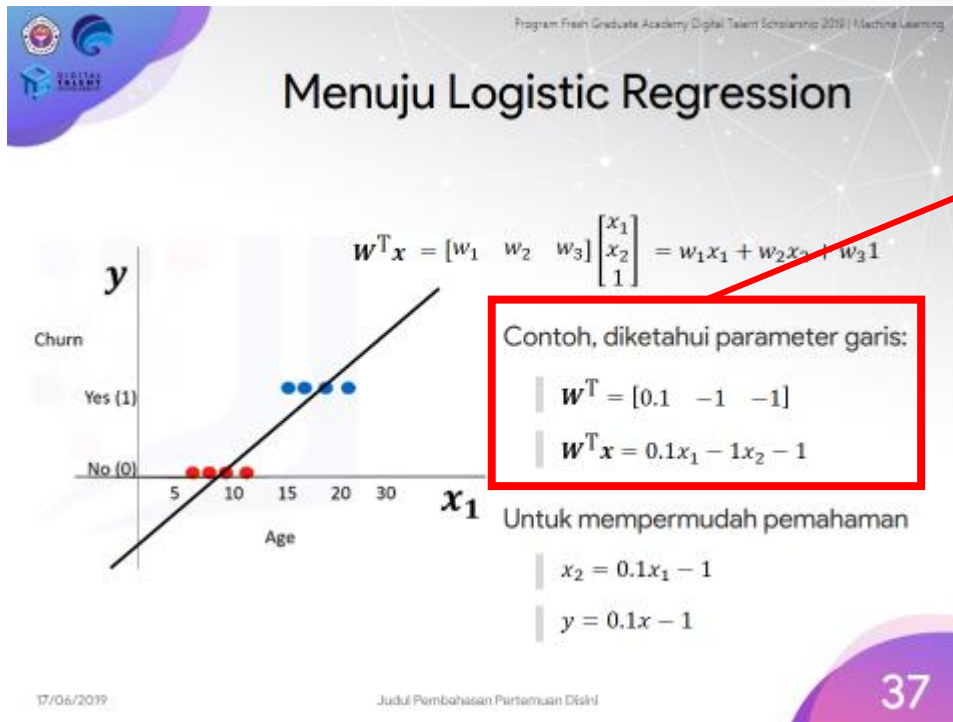


21



Alasan Dibaliknya

- Persamaan garis yang kita miliki hanya berdasarkan asumsi



Pemilihan persamaan garis hanyalah contoh, educated guess yang bukan merepresentasikan kategori data yang sebenarnya.

Lalu bagaimana caranya kita menemukan persamaan garis yang merepresentasikan kategori data pelanggan terhadap *churn* yang lebih akurat?



Bagian Empat

Training Logistic Regression

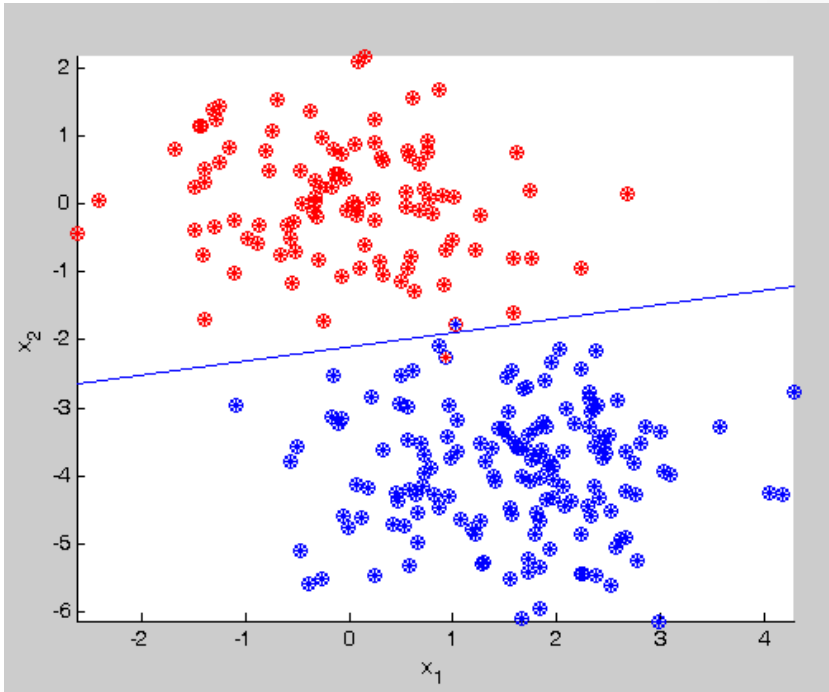
Training Logistic Regression

- Berdasarkan pengamatan yang kita lakukan, kita melihat bahwa, diperlukan persamaan garis yang tepat agar kategorisasi data menjadi lebih akurat.
- Untuk menemukan persamaan garis yang tepat, sebuah proses yang disebut training, harus dilakukan!
- Training dilakukan untuk meminimalisasi error yang dibuat oleh persamaan garis.

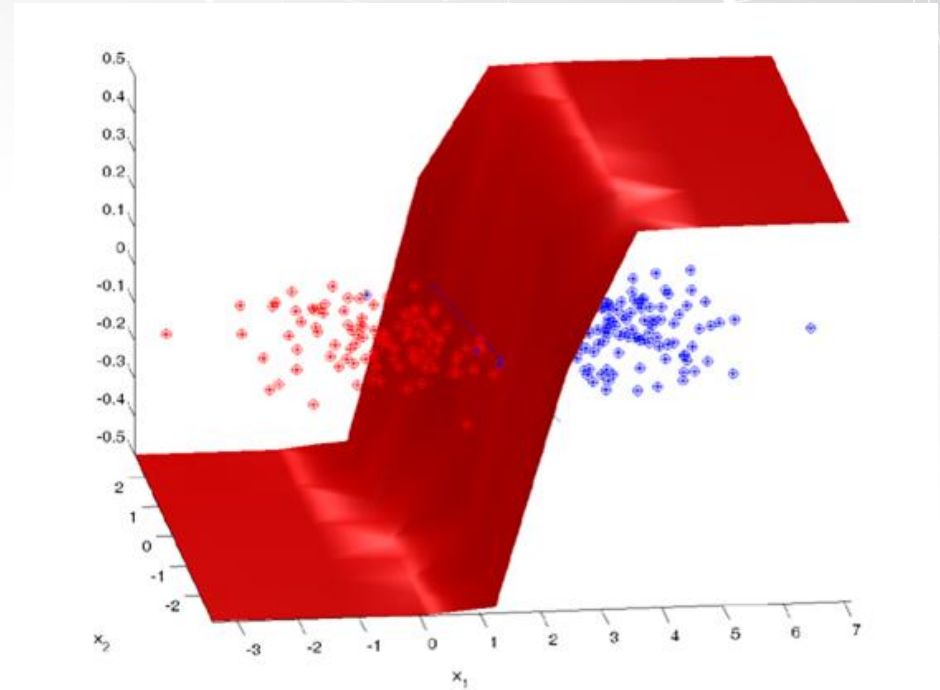
Training Logistic Regression

1. Inisialisasi \mathbf{W}^T secara acak untuk menentukan persamaan garis.
2. Hitung score, $\hat{y} = \mathbf{W}^T \mathbf{x}$
3. Bandingkan hasil \hat{y} dengan nilai label (dependent variable) yang sebenarnya, y
4. Hitung besar error (kesalahan) yang dilakukan oleh persamaan garis tersebut.
5. Ubah \mathbf{W}^T sedemikian rupa untuk mereduksi cost.
6. Kembali ke langkah 2.

Logistic Regression di 2D

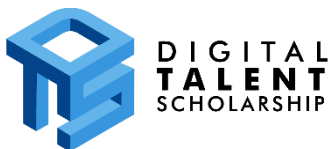






Decision Boundary



Logistic Function in Action

IKUTI KAMI



-  digitalent.kominfo
-  digitalent.kominfo
-  DTS_kominfo
-  Digital Talent Scholarship 2019

Pusat Pengembangan Profesi dan Sertifikasi
Badan Penelitian dan Pengembangan SDM
Kementerian Komunikasi dan Informatika
Jl. Medan Merdeka Barat No. 9
(Gd. Belakang Lt. 4 - 5)
Jakarta Pusat, 10110

