Modeling

Presented to you by

Dipl. Inform.(FH) Jony Sugianto, M. Comp. Sc.
WA:0812-130-86659
Github: https://github.com/jonysugianto

Use Case 1: Profiling User

Use Case 1: Profiling User

- User activness

  -derajat keaktivan user

  -can be used as bot or spam detection
- User interest based on reading behaviour
- Personal Profile

  -for all user (login and non login)

  -estimate age, gender, marital status for non login user

Source data

- Log Data(all user)
  - Anonymous ID from Cookie Data
  - LoginID (if exist)
  - ArticleId
  - Kanal / Category
  - Browser
  - IP
  - Etc...
- Registered User Data(10 %)
  - Login ID
  - Name
  - Age
  - Gender
  - Education
  - Etc...

Compute

Click Per User Per Kanal		
Per Minggu		
User	Kanal	Click
1	News	10
1	Sport	100
1	Lifestyle	50
2	News	200
2	Sport	20
2	Lifestyle	30

Click Per User	
Per Minggu	
User	Total Click
1	160
2	250

Total Click
Per Minggu
410

Rata-rata Click Per User
Per Minggu
205

Compute

Click Per User Per Kanal			
Per Minggu			
User	Kanal	% Click	% Click
1	News	10/160	0.0625
1	Sport	100/160	0.6250
1	Lifestyle	50/160	0.3125
2	News	200/250	0.8000
2	Sport	20/250	0.0800
2	Lifestyle	30/250	0.1200

User Activness (UA)1		
Per Minggu		
User	% UA	% UA
1	160/410	0.3902
2	250/410	0.6098

User Activness (UA)2		
Per Minggu		
User	% UA	% UA
1	160/205	0.7805
2	250/205	1.2195

Why UA 2 is better than UA 1 in media online?

- Higher value of UA2 tend to be a Robot or Spammer in media online user.
- The Normal value of UA2 is always around 1
- More interpretable

User Activness and User Interest

User	News	Sport	Lifestyle	User Activness
1	0.0500	0.6500	0.3000	0.8000
2	0.8000	0.0800	0.1200	1.3000

How to update the User activness?

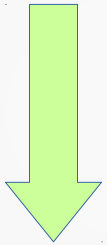
User behaviour change slower than the actual reading activity

*New-UA=w_history * UA-sofar + w_current * UA-Per-Minggu*

w_history=0.75
w_current=0.25

Assigning personal profile

User	Loginid	News	Sport	Lifestyle	User Activness
1	11	0.0500	0.6500	0.3000	0.8000
2		0.7000	0.1800	0.1200	1.3000
3		0.0500	0.5500	0.2000	0.9000
4	22	0.8000	0.0800	0.1200	1.4000
5		0.0400	0.6600	0.3000	0.7000



User	Loginid	E Loginid	News	Sport	Lifestyle	User Activness
1	11		0.0500	0.6500	0.3000	0.8000
2		22	0.7000	0.1800	0.1200	1.3000
3		11	0.0500	0.5500	0.2000	0.9000
4	22		0.8000	0.0800	0.1200	1.4000
5		11	0.0400	0.6600	0.3000	0.7000

User similarity

We Use this:

$$similarity(A,B) = (A_{news} - B_{news})^2 + (A_{sport} - B_{sport})^2 + (A_{lifestyle} - B_{lifestyle})^2 + (A_{useractivness} - B_{useractivness})^2$$

Why not this one:

$$similarity(A,B) = |A_{news} - B_{news}| + |A_{sport} - B_{sport}| + |A_{lifestyle} - B_{lifestyle}| + |A_{useractivness} - B_{useractivness}|$$

Improvement?

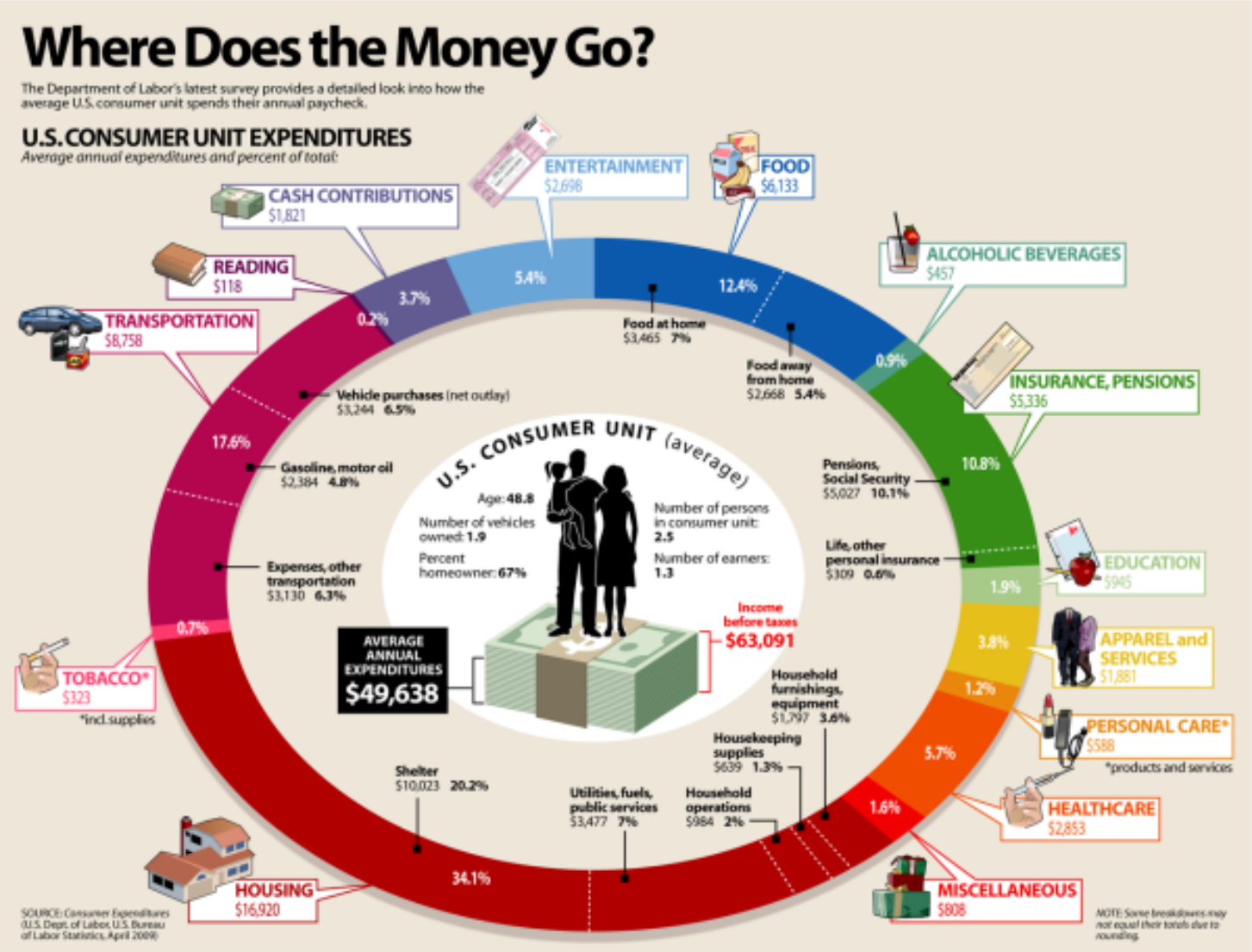
Final data

User	Loginid	E Loginid	News	Sport	Lifestyle	User Activness	Age	Sex
1	11		0.0500	0.6500	0.3000	0.8000	30	M
2		22	0.7000	0.1800	0.1200	1.3000	45	L
3		11	0.0500	0.5500	0.2000	0.9000	30	M
4	22		0.8000	0.0800	0.1200	1.4000	45	L
5		11	0.0400	0.6600	0.3000	0.7000	30	M

Complexity Analysis

- All User : 10.000.000
- Loginuser: 1.000.000
- Comparison per second per CPU: 1.000.000
- Total Comparison: 9.000.000.000.000
- Total Time: 9.000.000 second=104 hari

Use Case 2: Predicting customer spending



Use Case 2: Predicting customer spending

- Goal:

  to predict customer spending in the next 3 months

- Available Data

  transactional data form bankcards(creditscards, debitcards, etc.)

Observations

- Many spending repeat every year:

  - birthday

  - lebaran, natal

  - school holidays, university holydays
- Most people are employee

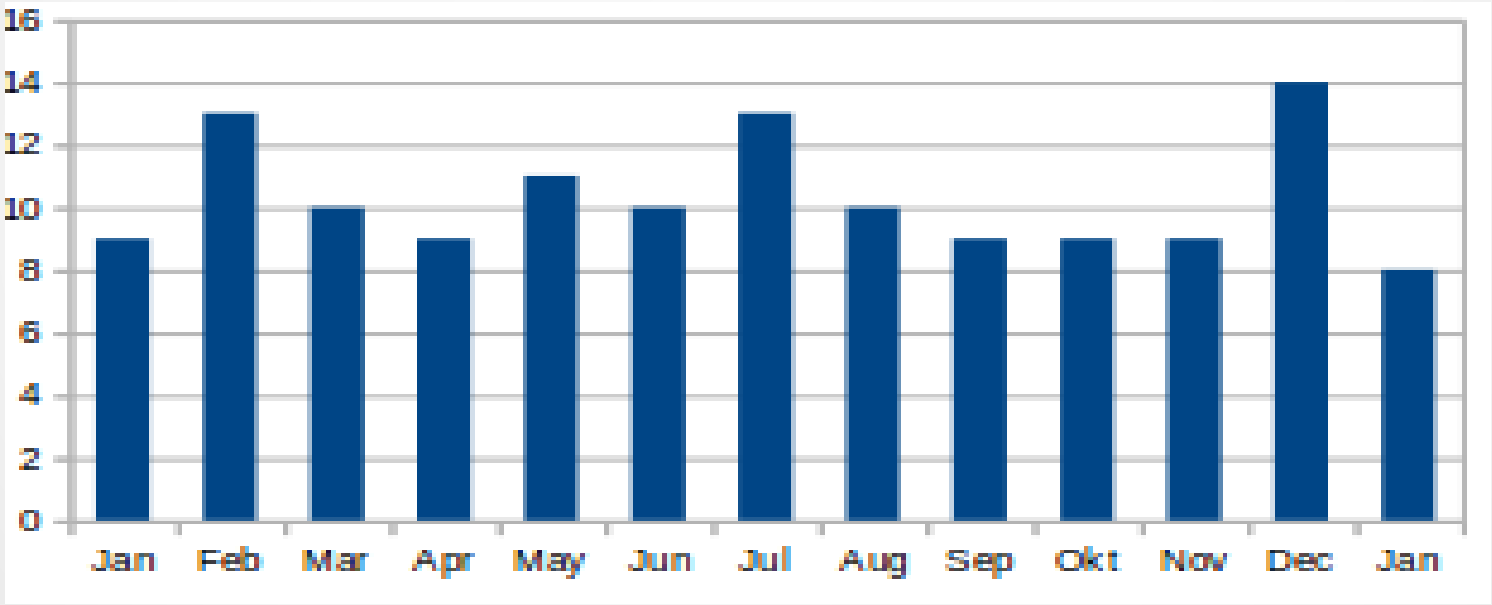
  - dramatic change in the monthly income is not expected
- Daily spending is relativ stable

  -Food

  -Transportation

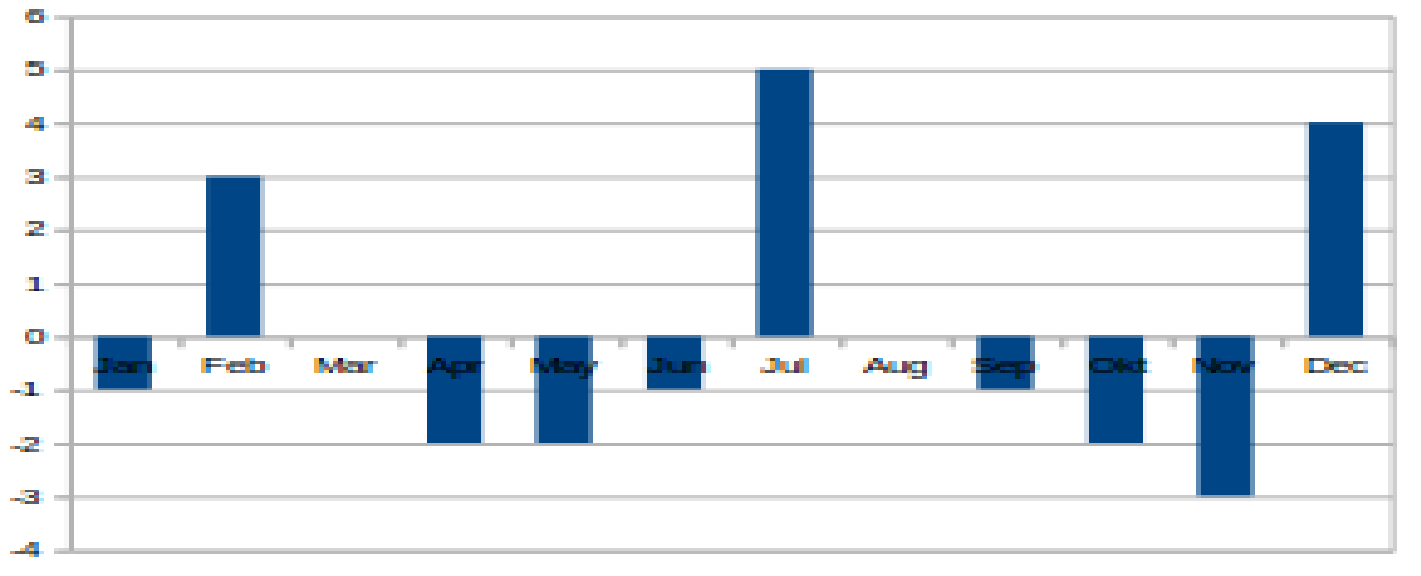
Features engineering

Aggregate monthly spending to reduce the amount of data



Features engineering

- Compute the mean of then monthly spending from the last 12 months
- Compute the different between the monthly spending and the mean from the last 12 months to make the trend in the spending clearer(careful)



Build training data

- We need minimum 12 months data as inputs for the predicting model.

  Why ?

- We need minimum 26 months data to build the training data for the model.

  Why ?

Build training data

Input variables:

- The last 12 months spending(1)
- Mean of the spending from last 12 months(2)
- The different between point 1 and point 2

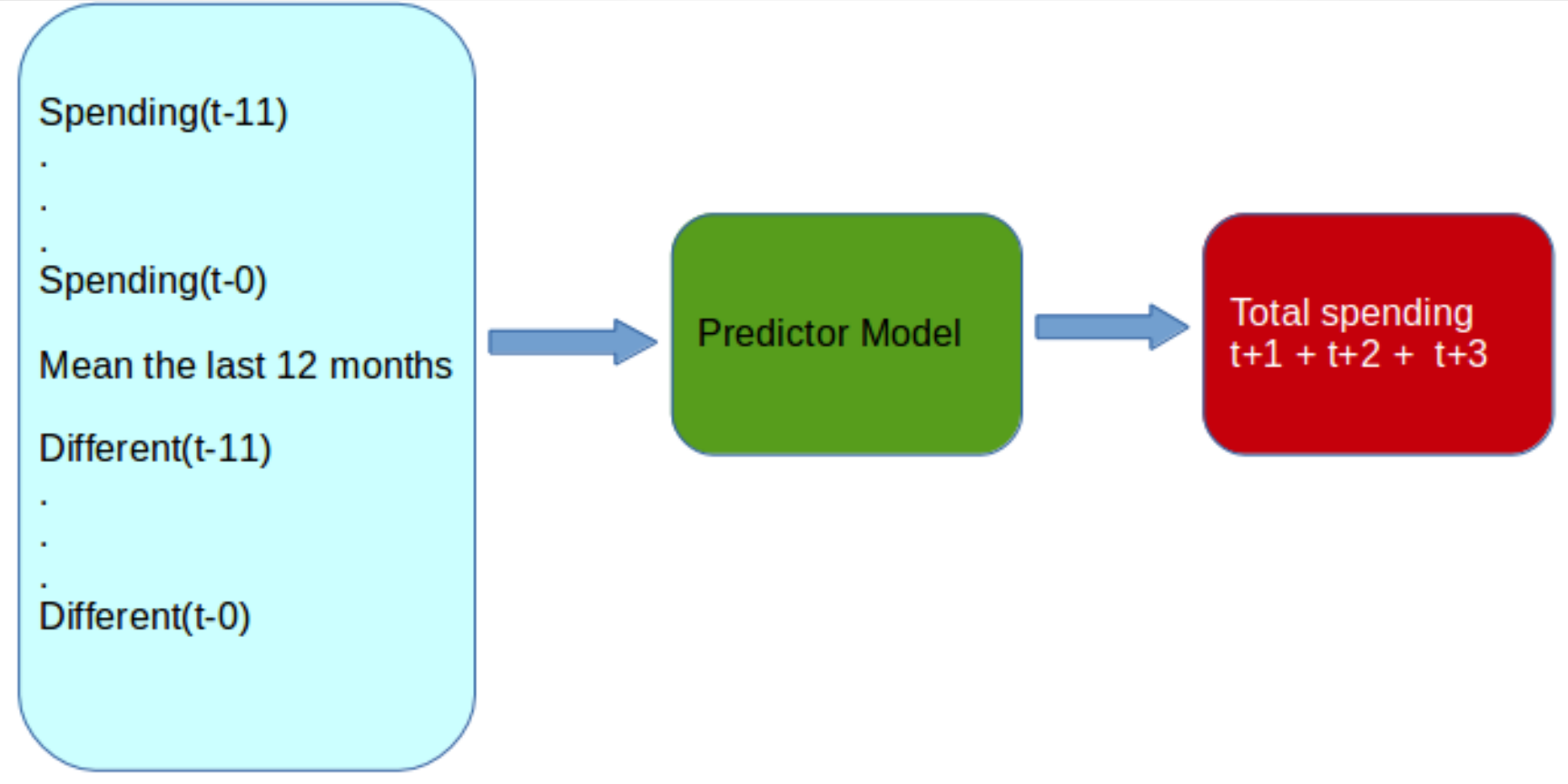
Output Variable:

- Aggregate the spending from the next 3 months

Build training data

Input Values			Target Values
Pengeluaran	Rata rata	Selisih	Jumlah Pengeluaran
0...11	0...11	0...11	Bulan(12+13+14)
1...12	1...12	1...12	Bulan(13+14+15)
2...13	2...13	2...13	Bulan(14+15+16)
3...14	3...14	3...14	Bulan(15+16+17)
4...15	4...15	4...15	Bulan(16+17+18)
5...16	5...16	5...16	Bulan(17+18+19)
6...17	6...17	6...17	Bulan(18+19+20)
7...18	7...18	7...18	Bulan(19+20+21)
8...19	8...19	8...19	Bulan(20+21+22)
9...20	9...20	9...20	Bulan(21+22+23)
10...21	10...21	10...21	Bulan(22+23+24)
11...22	11...22	11...22	Bulan(23+24+25)

Predictor


Spending(t-11)
.
.
.
Spending(t-0)
Mean the last 12 months
Different(t-11)
.
.
.
Different(t-0)
Predictor Model
Total spending
t+1 + t+2 + t+3

Questions?