CHULA ENGINEERING
Foundation toward Innovation

COMPUTER

Chula Big Data and IoT
Center of Excellence
(CUBIC)

Data Mining Group
Machine Intelligence and Knowledge Discovery Lab
Chulalongkorn University

Lab: Employee Resignation Prediction

Python for Data Analytics

Peerapon Vateekul, Ph.D.
Department of Computer Engineering,
Faculty of Engineering, Chulalongkorn University
Peerapon.v@chula.ac.th

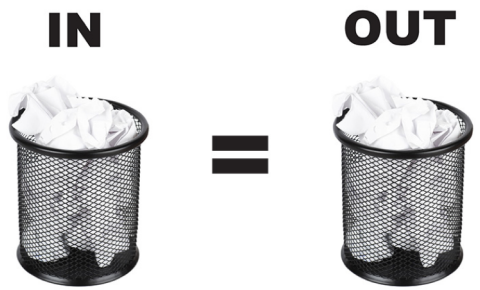
Course Objectives (Recap)

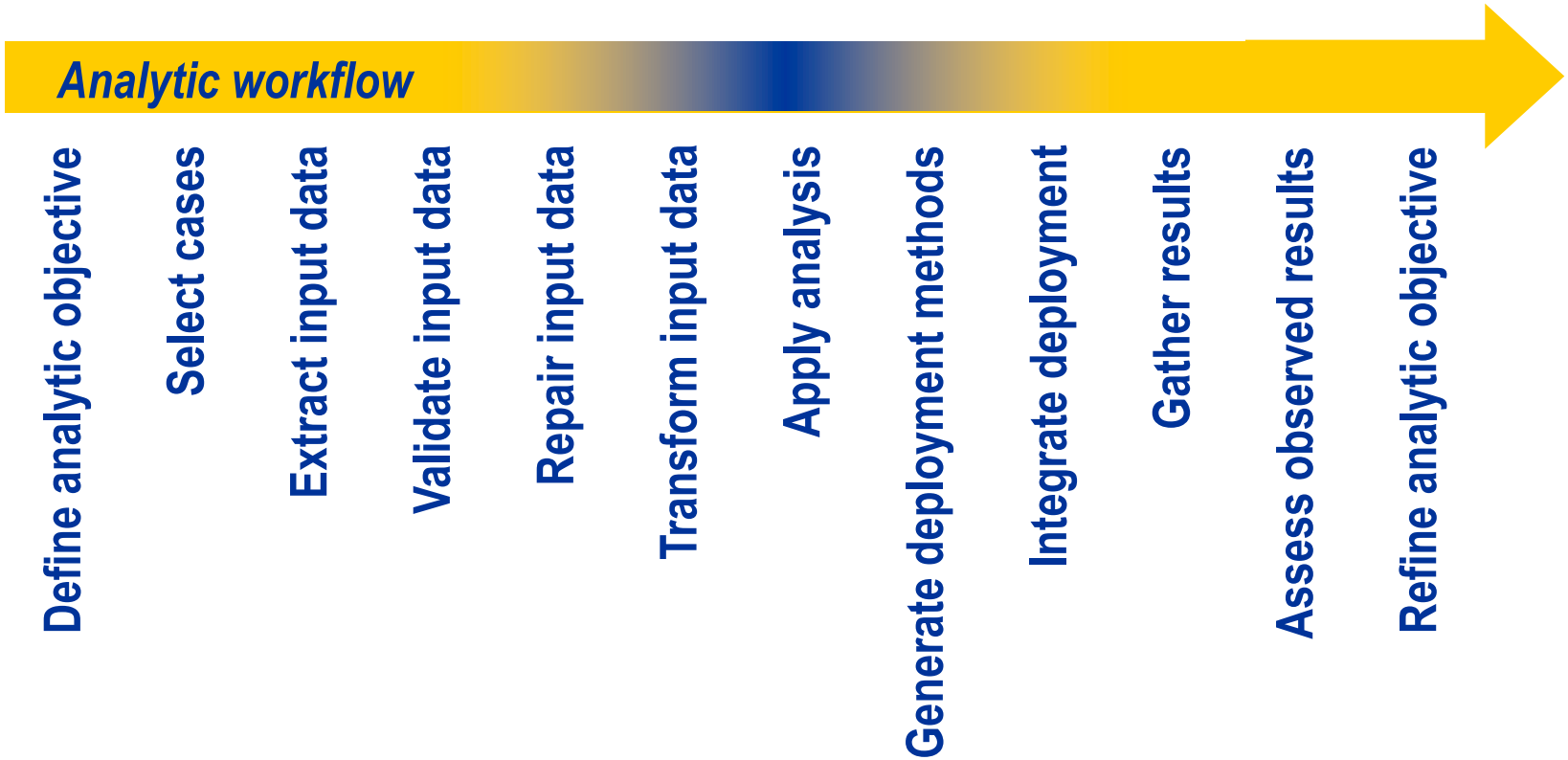
- Understand data analytics process
- Be able to conduct data analytics project
- Be able to conduct proper data processing

- Understand data analytical tasks: Classification & Regression
- Be able to choose suitable algorithms for each specific task

- Workshop on REAL data from Exxon's Human Resource (HR)

Analytics workflow





Data preparation challenges

- Massive data sets

- Temporal infidelity

- Transaction and event data

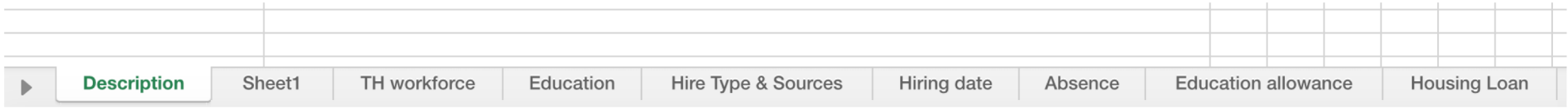
- Non-numeric data

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

- Exceptional, extreme, and missing values

- Stationarity

Data Thailand HR_(masked).xlsx : 7 sheets



- TH workforce
- Education
- Hire Type & Sources
- Hiring date
- Absence
- Education allowance
- Housing Loan

inputs

target

Serial	Income	Gender	Province	Leave
1001	25,000	Female	Bangkok	Leave
1002	50,000	Female	Nontaburi	Not Leave
1003	35,000	Male	Bangkok	Not Leave

Sheet: "TH workforce" (main)

- Refer to the last activity of each employee
- Replace the latest updated data, NOT all activities!
- Important variables are (1) Serial Number, (2) Tech Date of Entry, (3) Personnel Action Text (target)

A	B	C	D	E	F	G	H	I	J	K	L
Serial Numb	Supervisor Seria	Gender	Month of Birt	Year of Birt	Tech Date of En	Employment Sta	Chief Design	A&D Designated P	Corp Hierarc	Corp Hierarc	Corp Hierarc
1	364	F	5	1938	1/2/1955	Withdrawn	N	N			
2		M	4	1942		Withdrawn	N	N			
3		F	9	1938		Withdrawn	N	N			
4		M	12	1942	1/12/1961	Withdrawn	N	N			
5		M	6	1939	1/2/1961	Withdrawn	N	N			
6	346	M	10	1939	1/2/1962	Withdrawn	N	N			
7		M	8	1939	1/3/1962	Withdrawn	N	N			
8	27	F	6	1939	1/3/1962	Withdrawn	N	N			
9		F	8	1937		Withdrawn	N	N			
10		F	3	1943	1/6/1962	Withdrawn	N	N			
11	88	M	3	1939	1/4/1962	Withdrawn	N	N			
12		F	6	1945	1/1/1964	Withdrawn	N	N			

Sheet: "Education"

- All educations of each employees (1-to-M; employee-to-edu)
- Issue: include both degree & non-degree in the same sheet, so we need to treat them separately

Serial Numbe	Start Date (0	End Date (00	Institute/location	Certificate Text	Branch of study 1	xt
1	01/01/1950	01/01/1953	SUANKWONG SCH.	Attended	General Studies	
2	01/01/1953	01/01/1957	VIENGPROU VITHAYA SCH.	Attended	No Specific Branch	
3	01/01/1949	01/01/1952	WADTEPSIRIN SCH.	Attended	No Specific Branch	
4	01/01/1956	01/01/1959	WATTANA VITHAYA ACADEMY.	Attended	Education	
5	06/01/1956	03/31/1960	THAMMASAT UNIVERSITY	Bachelor	Business-Economics	
6	01/01/1949	01/01/1953	WADCHAENG SCH.	Attended	No Specific Branch	
7	06/01/1959	08/30/1961	TECHNICAL TRAINING SCHOOL	Diploma	Other Degree	
8	01/01/1956	01/01/1959	BANGKOK COMMERCIAL COLLEGE	Attended	Finance/Accounting	

Sheet: "Hire Type & Sources"

- Mix between two different information:
  - Hire type
  - Source

Serial Number	Subtype (003	Statistics indicato	Exception Description (0033	Statistics Indicator Description (0033
72	9U02	03	Inexperienced Hire	Hire type
134	9U02	03	Inexperienced Hire	Hire type
479	9U02	01	Other Source	Source
479	9U02	03	Experienced Hire	Hire type
615	9U02	01	Other Source	Source
615	9U02	03	Experienced Hire	Hire type
694	9U02	01	Other Source	Source
694	9U02	03	Experienced Hire	Hire type
749	9U02	01	Employee Referral	Source
749	9U02	03	Experienced Hire	Hire type
754	9U02	01	Other Source	Source

Sheet: "Hiring date" (cannot use)

- All hiring dates of each employees. It is (1-to-M), since some employees may leave and then enter to the company again

- Issue: There is discrepancy between this sheet and the sheet "TH workforce".
  - The end date is different from the quit date. So, this sheet is not included in the analysis.

Serial Numbe	Start Date (000	End Date (000	Action type (000	Action type (0000) Text	Reason for Action text (000
1	02/01/1955	10/31/1998	Y3	Hiring Active (Migration)	Migration Upload
2	01/01/1960	05/30/1998	Y3	Hiring Active (Migration)	Migration Upload
3	10/15/1958	06/29/1998	Y3	Hiring Active (Migration)	Migration Upload
4	12/01/1961	02/29/2000	Y3	Hiring Active (Migration)	Migration Upload
5	02/10/1961	02/29/2000	Y3	Hiring Active (Migration)	Migration Upload
6	01/22/1962	12/31/1999	Y3	Hiring Active (Migration)	Migration Upload
7	03/01/1962	01/31/2000	Y3	Hiring Active (Migration)	Migration Upload
8	03/15/1962	01/21/1999	Y3	Hiring Active (Migration)	Migration Upload
9	04/01/1962	01/30/1998	Y3	Hiring Active (Migration)	Migration Upload
10	06/05/1962	01/01/1999	Y3	Hiring Active (Migration)	Migration Upload
11	04/01/1962	11/30/1999	Y3	Hiring Active (Migration)	Migration Upload
12	01/06/1964	03/31/1999	Y3	Hiring Active (Migration)	Migration Upload
13	11/23/1964	10/31/1999	Y3	Hiring Active (Migration)	Migration Upload
14	09/17/1964	12/31/2000	Y3	Hiring Active (Migration)	Migration Upload
15	05/07/1964	01/01/1999	Y3	Hiring Active (Migration)	Migration Upload
16	08/27/1964	10/31/1998	Y3	Hiring Active (Migration)	Migration Upload
17	11/24/1964	02/29/2000	Y3	Hiring Active (Migration)	Migration Upload
18	03/04/1965	02/28/1999	Y3	Hiring Active (Migration)	Migration Upload
19	06/01/1965	09/30/1999	Y3	Hiring Active (Migration)	Migration Upload
20	07/16/1965	01/31/2000	Y3	Hiring Active (Migration)	Migration Upload
21	10/01/1965	09/30/2001	Y3	Hiring Active (Migration)	Migration Upload
22	11/01/1965	04/30/1999	Y3	Hiring Active (Migration)	Migration Upload

Sheet: "Absence"

- Absence information of each employee (1-to-M)
- Since it is a transaction data, so it can capture employee's behavior.
- Issue: It can cause temporal infidelity! So, the use case must be carefully designed.

Serial Nu	Start date	End date	Attendance c	Days
13	27/1/2005	28/1/2005	Vacation Leave	2
13	1/2/2005	1/2/2005	Vacation Leave	1
13	7/2/2005	7/2/2005	Vacation Leave	1
13	14/2/2005	14/2/2005	Vacation Leave	1
13	17/2/2005	18/2/2005	Vacation Leave	2
13	4/3/2005	4/3/2005	Vacation Leave	1
13	17/3/2005	18/3/2005	Vacation Leave	2
13	11/4/2005	12/4/2005	Vacation Leave	2
13	22/4/2005	22/4/2005	Vacation Leave	1
13	20/5/2005	20/5/2005	Vacation Leave	1

Sheet: "Education allowance"

- Although there are start and end dates, there is only one duration (just ONE DAY) for all row.
- So, each row refers to each studying day.

Serial Num	Start date	End date	Wage Type
446	30/9/2004	30/9/2004	Edu Assist'nce-Public
446	26/6/2006	26/6/2006	Edu Assist'nce-Public
446	30/6/2006	30/6/2006	Edu Assist'nce-Public
500	30/4/2005	30/4/2005	Edu Assist'nce-Public
500	30/11/2005	30/11/2005	Edu Assist'nce-Public
500	31/8/2007	31/8/2007	Edu Assist'nce-Public
540	31/7/2013	31/7/2013	Edu Assist'nce-Public
543	31/8/2013	31/8/2013	Edu Assist'nce-Public
546	28/2/2014	28/2/2014	Edu Assist'nce-Public
581	30/12/2008	30/12/2008	Edu Assist'nce-Public
581	11/5/2009	11/5/2009	Edu Assist'nce-Public
581	3/8/2009	3/8/2009	Edu Assist'nce-Public

Sheet: "Housing Loan"

- Monthly payments of employee's loan
- For one loan, there can be several monthly payments.
- For one employee, there can loan many times (many periods).

Serial Num	Start date	End date	Wage Type
34	29/11/2002	29/11/2002	Housing Int. Subsidy
34	30/12/2002	30/12/2002	Housing Int. Subsidy
34	31/1/2003	31/1/2003	Housing Int. Subsidy
34	28/2/2003	28/2/2003	Housing Int. Subsidy
34	31/3/2003	31/3/2003	Housing Int. Subsidy
34	30/4/2003	30/4/2003	Housing Int. Subsidy
34	30/5/2003	30/5/2003	Housing Int. Subsidy
34	30/6/2003	30/6/2003	Housing Int. Subsidy
34	31/7/2003	31/7/2003	Housing Int. Subsidy

Lab1: Feature Selection

- Lab1.1_FS_SelectCase
- Lab1.2_FS_HireType (consolidate load_grouping_table)
- Lab1.3_FS_Absense (consolidate load_grouping_table)
- Lab1.4_FS_HouseLoan
- Lab1.5_FS_Edu (consolidate regex; load_grouping_table)
- Lab1.6_FS_EduAllowance
- Lab1.7_FS_Workforce (not use)
- Lab1.8_FS_JoinAll by Serial

Lab2_DataCleansing

- Data exploration:
  - Checking missing
  - Histogram
- Data cleansing
  - Impute missing: numeric & nominal (object)
- Feature selection:
  - Chi-squared test

Lab3 Prediction Model

- Lab3.1_DecisionTree
- Lab3.2_LogisticRegression
- Lab3.3_LogisticRegression_WithFS
- Lab3.4_NeuralNetwork
- Lab3.5_ParamSearch

+

Any Questions?