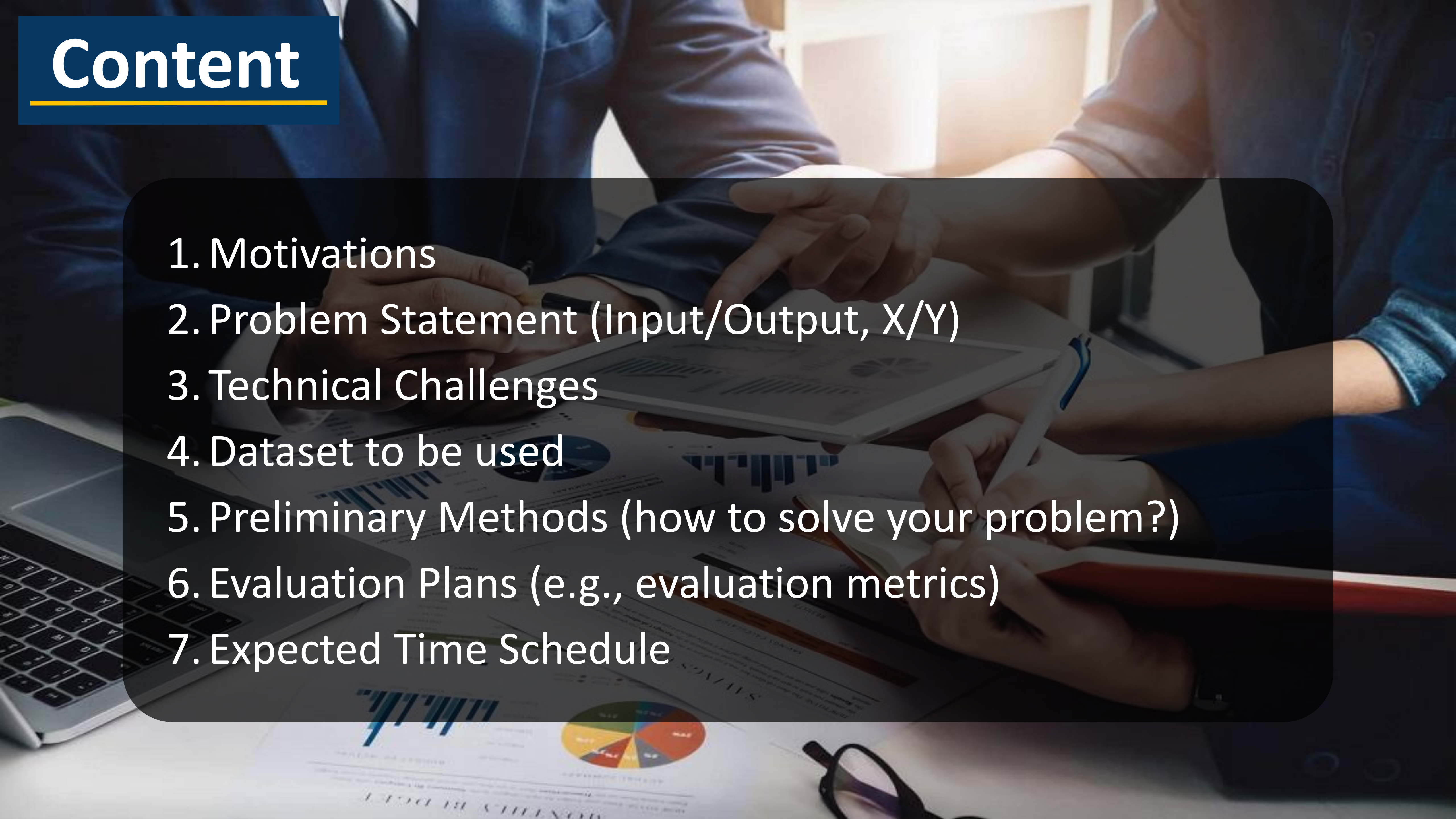
Final Project Proposal

Predict Future Sales

組員:工資112林昕霏、工資112徐圓媛



Content

1. Motivations
2. Problem Statement (Input/Output, X/Y)
3. Technical Challenges
4. Dataset to be used
5. Preliminary Methods (how to solve your problem?)
6. Evaluation Plans (e.g., evaluation metrics)
7. Expected Time Schedule

Motivations

The task for this project is to **predict the total amount of sales** next month for each product in store.

The reason why we choose to do this project is because that it’s **an application of Data Science in industrial engineering.** From what we learn from production management, to make a shop more competitive, reducing inventory is the most important issue. Inventory significantly influences the total revenue of an organization. If we could accurately predict the amount of products sold each day, it would help to reduce the inventory cost.

Problem Statement (Input/Output, X/Y)

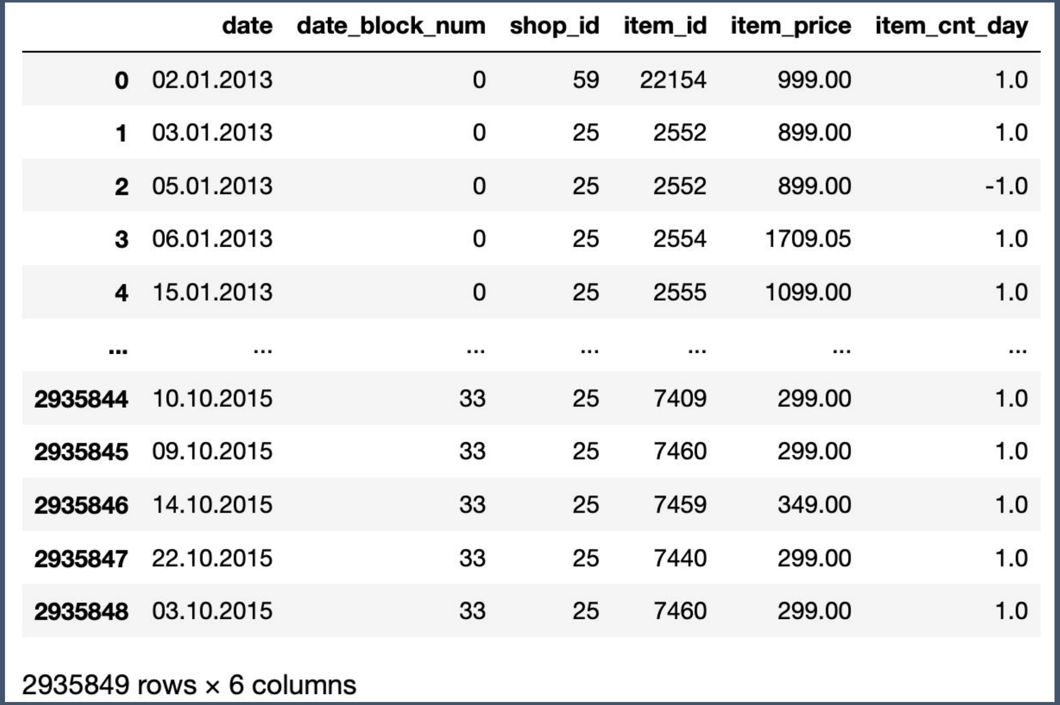
Working with a time-series dataset consisting of daily sales data, provided by one of the largest Russian software firms 1C Company. The target for this project is to **predict total sales for every product in store in the next month.**

Input : daily historical data from January 2013 to October 2015. 22.2k

Output : forecast the sales for these shops and products for November 2015.

Problem Statement (Input/Output, X/Y)

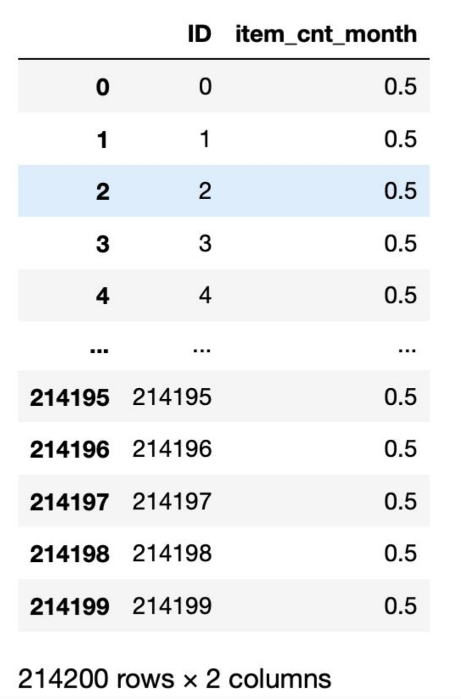
Input



	date	date_block_num	shop_id	item_id	item_price	item_cnt_day
0	02.01.2013	0	59	22154	999.00	1.0
1	03.01.2013	0	25	2552	899.00	1.0
2	05.01.2013	0	25	2552	899.00	-1.0
3	06.01.2013	0	25	2554	1709.05	1.0
4	15.01.2013	0	25	2555	1099.00	1.0
...	...	...	...	...	...	...
2935844	10.10.2015	33	25	7409	299.00	1.0
2935845	09.10.2015	33	25	7460	299.00	1.0
2935846	14.10.2015	33	25	7459	349.00	1.0
2935847	22.10.2015	33	25	7440	299.00	1.0
2935848	03.10.2015	33	25	7460	299.00	1.0

2935849 rows × 6 columns

Output



	ID	item_cnt_month
0	0	0.5
1	1	0.5
2	2	0.5
3	3	0.5
4	4	0.5
...	...	...
214195	214195	0.5
214196	214196	0.5
214197	214197	0.5
214198	214198	0.5
214199	214199	0.5

214200 rows × 2 columns

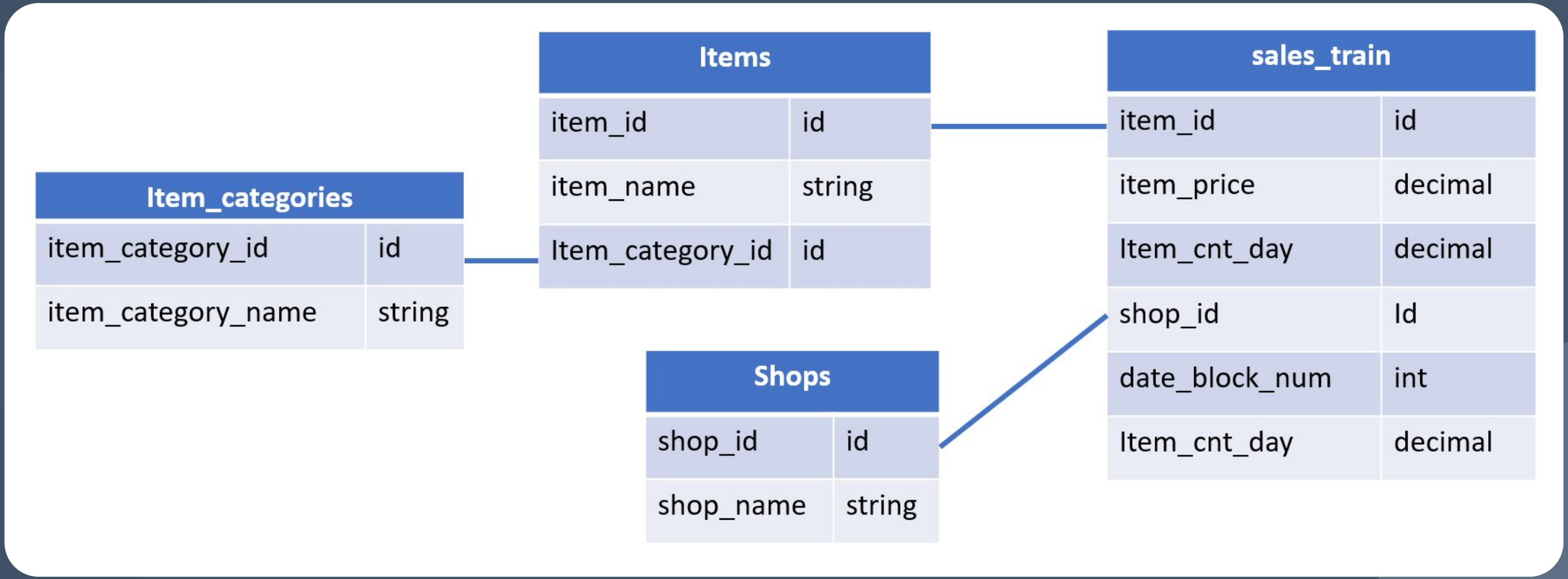
Technical Challenges

The list of shops and products **slightly changes** every month

→ a robust model is needed to handle such problem

We are not sure which kind of **model** will be best for this problem → might need to **trial and error** for several times

Dataset to be used



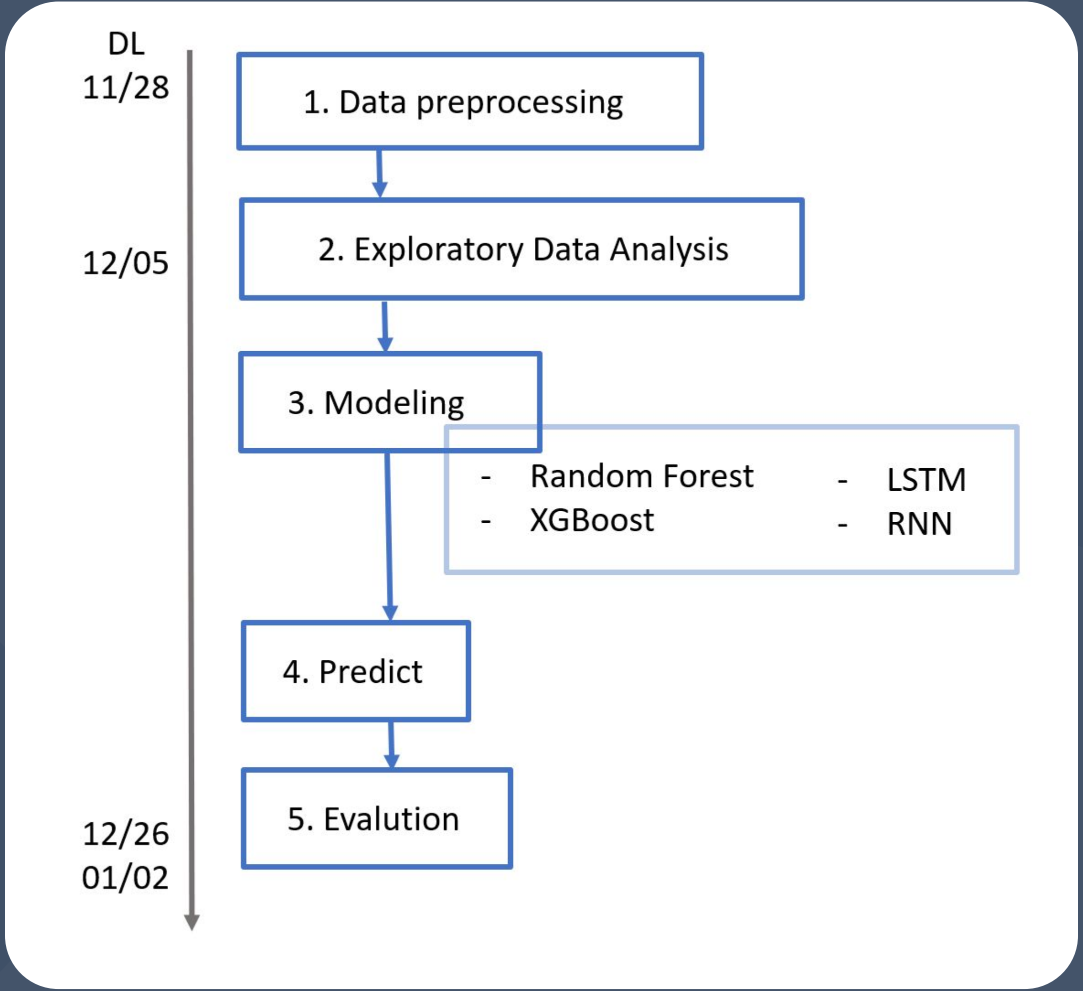
Item_categories	
item_category_id	id
item_category_name	string

Items	
item_id	id
item_name	string
Item_category_id	id

Shops	
shop_id	id
shop_name	string

sales_train	
item_id	id
item_price	decimal
Item_cnt_day	decimal
shop_id	Id
date_block_num	int
Item_cnt_day	decimal

Preliminary Methods & Expected Time Schedule



DL 11/28

1. Data preprocessing

12/05

2. Exploratory Data Analysis

3. Modeling

- Random Forest
- XGBoost
- LSTM
- RNN

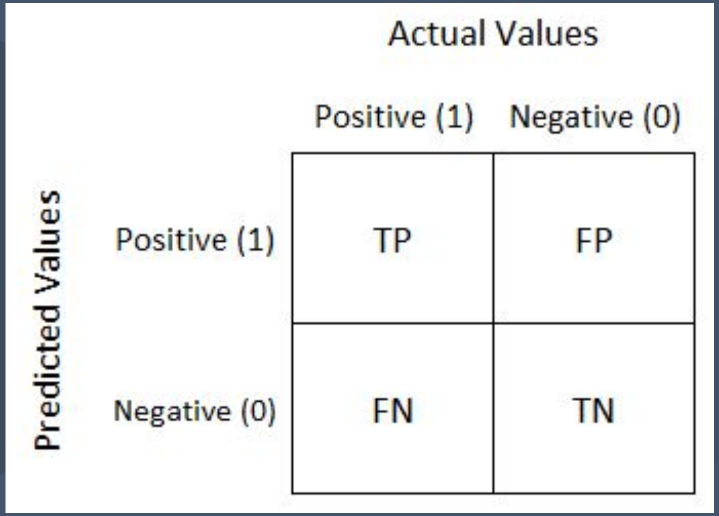
4. Predict

12/26 01/02

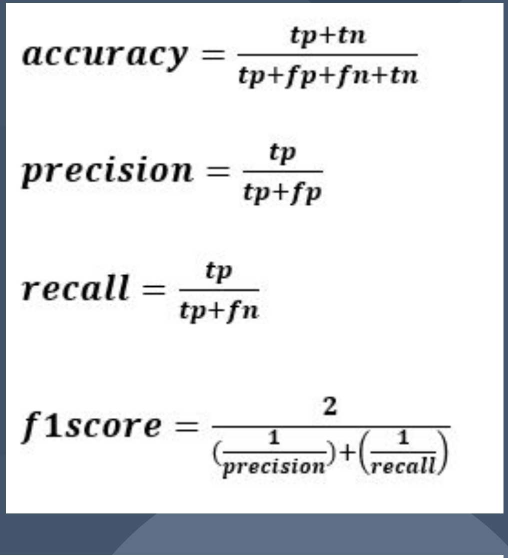
5. Evalution

Evaluation Plans

recall
precision
accuracy
F1-score
confusion matrix
RMSE



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

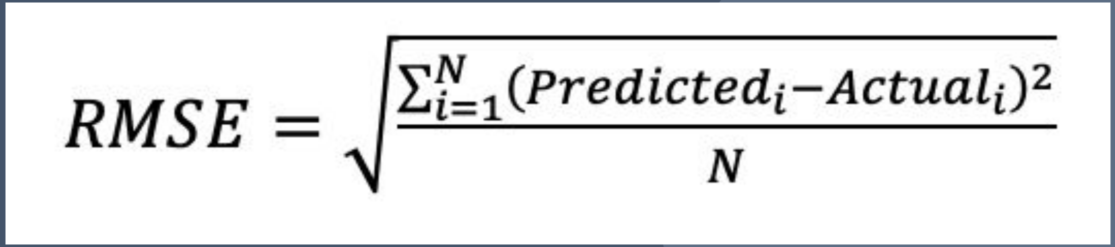


$$accuracy = \frac{tp+tn}{tp+fp+fn+tn}$$

$$precision = \frac{tp}{tp+fp}$$

$$recall = \frac{tp}{tp+fn}$$

$$f1score = \frac{2}{(\frac{1}{precision})+(\frac{1}{recall})}$$



$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

Thank you!

老師的建議

使用和時間相關的模型
每個時間點的數值或是變化特徵

使用回歸方法，自己定義新的特徵
例如：
最近這個月的銷售量
過去五個月的平均銷售量、變異數
是否逐月成長
作為特徵去做訓練，定義新的特徵