Kaggle Guideline

20215202 Jeongeun Choi

OUTLINE

INTRODUCTION

COLAB SETTING

BODY

DATA ANALYSIS AND BASIC PREPROCESSING

EVALUATION METRICS

MODEL OVERVIEW

CONCLUSION

ENTRY SUBMISSION

DATA ANALYSIS AND BASIC PREPROCESSING

1st step for the Kaggle Competition

Data should be made usable to be inserted in the model

DATA ANALYSIS AND BASIC PREPROCESSING

1st step for the Kaggle Competition

Data should be made usable to be inserted in the model

given information

id	rider_age	rider_rating	restaurant_latitude	restaurant_longitude	delivery_latitude	delivery_longitude	order_date	order_time	pickup_time	weather	traffic_density	vehicle_condition	order_type	vehicle_type	num_orders	is_festival	city	delivery_time
2D96	36	4.9	12.978453	77.643685	13.068453	77.733685	18-03-2022	22:40	22:50	clear_sky	low	1	Snack	scooter	1	No	Metropolitian	27
7F86	30	5	19.1093	72.825451	19.1193	72.835451	2022.3.1	11:25 AM	11:30	clear_sky	high	2	Drinks	scooter	1	No		17
CAB2	23	4.8	19.12663	72.829976	19.13663	72.839976	11.Mar.22	9:15	9:25	sandstorm	low	1	Snack	scooter	1	No	Metropolitian	15
860F	39	4.4	19.874733	75.353942	19.944733	75.423942	14-02-2022	7:35 PM	19:40	storm	jam	2	Drinks	scooter		No	Metropolitian	36
C0E3	27	4.8	17.429585	78.392621	17.499585	78.462621	6.Apr.22	20:40	8:45 PM	clear_sky	jam	2	Meal	scooter	0	No	Metropolitian	20
2C52	35	4.8	23.333017	85.3172	23.413017	85.3972	12.Mar.22	17:20	17:30	overcast	medium	1	Meal	motorcycle	1	No	Metropolitian	44
2E86	29	3.7	17.428294	78.404423	17.498294	78.474423	23-03-2022	19:55	20:10	high_wind	jam	2	Meal	scooter	2	Yes	Metropolitian	44
3449	40	4.9	18.55144	73.804855	18.63144	73.884855	14-03-2022	07:10PM	19:25	sandstorm	jam	1	Buffet	motorcycle	1	No	Metropolitian	27
AB43	31	4.3	17.438263	78.397865	17.568263	78.527865	2022.3.4	17:35	17:45	storm	medium	2	Meal	motorcycle	1	No	Metropolitian	33
40AE	32	3.5	12.325461	76.632278	12.395461	76.702278	8.Mar.22	05:10PM	05:15PM	storm	medium	1	Meal	motorcycle	1	No		34
A28C	21	5	18.546947	73.900626	18.586947	73.940626	24-03-2022	14:20	2:30 PM	overcast	high	2	Snack	scooter	0	No	Metropolitian	27
46C7	26	4.6	22.751857	75.866699	22.761857	75.876699	7.Mar.22	11:45	11:50 AM	foggy	high	0	Snack	motorcycle	1	No	Urban	20
A7DC	24	4.9	21.186884	72.793616	21.216884	72.823616	2022.3.15	5:15 PM	5:30 PM	foggy	medium	0	Buffet	motorcycle	1	X	Metropolitian	28
F6F1	26	4.6	-26.472001	80.354002	26.582001	80.464002	2022.2.18	17:35	17:50	overcast	medium	1	Buffet	motorcycle	0	No	Metropolitian	25
70FD			18.530963	73.828972	18.600963	73.898972	29.Mar.22		07:25PM			3	Meal	e_scooter	1	No	Metropolitian	26
5449	21	4.9	26.469003	80.316344	26.579003	80.426344	16-02-2022	18:25	18:40	storm	medium	1	Buffet	motorcycle	1	No	Metropolitian	26
4E	24	4.6	18.543626	73.905101	18.633626	73.995101	2022.3.29	22:30	10:45 PM	sandstorm	low	0	Buffet	motorcycle	1	No	Metropolitian	18
57CA	32	3.9	25.449872	81.836167	25.589872	81.976167	12-02-2022	21:55	22:05	high_wind	jam	2	Meal	motorcycle	1	No	Metropolitian	34
A38F	28	4.8	26.905287	75.794592	27.045287	75.934592	2022.3.6	21:40	21:50	overcast	jam	2	Snack	e_scooter	0	No	Metropolitian	37
3AD3	34	4.8	22.307898	73.167788	22.337898	73.197788	03-04-2022	06:40PM	18:55	high_wind	medium	1	Snack	scooter	1	No	Metropolitian	33
D4BC	21	4.8	18.51421	73.838429	18.62421	73.948429	02-04-2022	23:35	23:45	high_wind	low	2	Buffet	e_scooter	0	X	Urban	12
BC3E	32	3.5	23.351058	85.325731	23.441058	85.415731	29.Mar.22	9:30 PM	21:45	sandstorm	jam	1	Drinks	motorcycle	1	No	Metropolitian	35
BF18	26	4.7	25.449872	81.836167	25.559872	81.946167	16-02-2022	19:35	19:45	overcast	jam	0	Drinks	motorcycle	1	Yes	Semi-Urban	49
76B8	40	4.4	0	0	0.02	0.02	2022.3.21	9:20	9:35	storm	low	0	Snack	motorcycle	1	No		24

DATA ANALYSIS AND BASIC PREPROCESSING

1st step for the Kaggle Competition

Data should be made usable to be inserted in the model

you have to predic this

given information

target

id	rider_age	rider_rating	restaurant_latitude	restaurant_longitude	delivery_latitude	delivery_longitude	order_date	order_time	pickup_time	weather	traffic_density	vehicle_condition	order_type	vehicle_type	num_orders	is_festival	city	delivery_time
2D96	36	4.9	12.978453	77.643685	13.068453	77.733685	18-03-2022	22:40	22:50	clear_sky	low	1	Snack	scooter	1	No	Metropolitian	27
7F86	30	5	19.1093	72.825451	19.1193	72.835451	2022.3.1	11:25 AM	11:30	clear_sky	high	2	Drinks	scooter	1	No		17
CAB2	23	4.8	19.12663	72.829976	19.13663	72.839976	11.Mar.22	9:15	9:25	sandstorm	low	1	Snack	scooter	1	No	Metropolitian	15
860F	39	4.4	19.874733	75.353942	19.944733	75.423942	14-02-2022	7:35 PM	19:40	storm	jam	2	Drinks	scooter		No	Metropolitian	36
C0E3	27	4.8	17.429585	78.392621	17.499585	78.462621	6.Apr.22	20:40	8:45 PM	clear_sky	jam	2	Meal	scooter	0	No	Metropolitian	20
2C52	35	4.8	23.333017	85.3172	23.413017	85.3972	12.Mar.22	17:20	17:30	overcast	medium	1	Meal	motorcycle	1	No	Metropolitian	44
2E86	29	3.7	17.428294	78.404423	17.498294	78.474423	23-03-2022	19:55	20:10	high_wind	jam	2	Meal	scooter	2	Yes	Metropolitian	44
3449	40	4.9	18.55144	73.804855	18.63144	73.884855	14-03-2022	07:10PM	19:25	sandstorm	jam	1	Buffet	motorcycle	1	No	Metropolitian	27
AB43	31	4.3	17.438263	78.397865	17.568263	78.527865	2022.3.4	17:35	17:45	storm	medium	2	Meal	motorcycle	1	No	Metropolitian	33
40AE	32	3.5	12.325461	76.632278	12.395461	76.702278	8.Mar.22	05:10PM	05:15PM	storm	medium	1	Meal	motorcycle	1	No		34
A28C	21	5	18.546947	73.900626	18.586947	73.940626	24-03-2022	14:20	2:30 PM	overcast	high	2	Snack	scooter	0	No	Metropolitian	27
46C7	26	4.6	22.751857	75.866699	22.761857	75.876699	7.Mar.22	11:45	11:50 AM	foggy	high	0	Snack	motorcycle	1	No	Urban	20
A7DC	24	4.9	21.186884	72.793616	21.216884	72.823616	2022.3.15	5:15 PM	5:30 PM	foggy	medium	0	Buffet	motorcycle	1	X	Metropolitian	28
F6F1	26	4.6	-26.472001	80.354002	26.582001	80.464002	2022.2.18	17:35	17:50	overcast	medium	1	Buffet	motorcycle	0	No	Metropolitian	25
70FD			18.530963	73.828972	18.600963	73.898972	29.Mar.22		07:25PM			3	Meal	e_scooter	1	No	Metropolitian	26
5449	21	4.9	26.469003	80.316344	26.579003	80.426344	16-02-2022	18:25	18:40	storm	medium	1	Buffet	motorcycle	1	No	Metropolitian	26
4E	24	4.6	18.543626	73.905101	18.633626	73.995101	2022.3.29	22:30	10:45 PM	sandstorm	low	0	Buffet	motorcycle	1	No	Metropolitian	18
57CA	32	3.9	25.449872	81.836167	25.589872	81.976167	12-02-2022	21:55	22:05	high_wind	jam	2	Meal	motorcycle	1	No	Metropolitian	34
A38F	28	4.8	26.905287	75.794592	27.045287	75.934592	2022.3.6	21:40	21:50	overcast	jam	2	Snack	e_scooter	0	No	Metropolitian	37
3AD3	34	4.8	22.307898	73.167788	22.337898	73.197788	03-04-2022	06:40PM	18:55	high_wind	medium	1	Snack	scooter	1	No	Metropolitian	33
D4BC	21	4.8	18.51421	73.838429	18.62421	73.948429	02-04-2022	23:35	23:45	high_wind	low	2	Buffet	e_scooter	0	X	Urban	12
BC3E	32	3.5	23.351058	85.325731	23.441058	85.415731	29.Mar.22	9:30 PM	21:45	sandstorm	jam	1	Drinks	motorcycle	1	No	Metropolitian	35
BF18	26	4.7	25.449872	81.836167	25.559872	81.946167	16-02-2022	19:35	19:45	overcast	jam	0	Drinks	motorcycle	1	Yes	Semi-Urban	49
76B8	40	4.4	0	0	0.02	0.02	2022.3.21	9:20	9:35	storm	low	0	Snack	motorcycle	1	No		24

DATA ANALYSIS AND BASIC PREPROCESSING

Data should be made usable to be inserted in the model

Missing values

id	rider_age	rider_rating	restaurant_latitude	restaurant_longitude	delivery_latitude	delivery_longitude	order_date	order_time	pickup_time	weather	traffic_density	vehicle_condition	order_type	vehicle_type	num_orders	is_festival	city	delivery_time
2D96	36	4.9	12.978453	77.643685	13.068453	77.733685	18-03-2022	22:40	22:50	clear_sky	low	1	Snack	scooter	1	No	Metropolitian	27
7F86	30	5	19.1093	72.825451	19.1193	72.835451	2022.3.1	11:25 AM	11:30	clear_sky	high	2	Drinks	scooter	1	No		17
CAB2	23	4.8	19.12663	72.829976	19.13663	72.839976	11.Mar.22	9:15	9:25	sandstorm	low	1	Snack	scooter	1	No	Metropolitian	15
860F	39	4.4	19.874733	75.353942	19.944733	75.423942	14-02-2022	7:35 PM	19:40	storm	jam	2	Drinks	scooter		No	Metropolitian	36
C0E3	27	4.8	17.429585	78.392621	17.499585	78.462621	6.Apr.22	20:40	8:45 PM	clear_sky	jam	2	Meal	scooter	0	No	Metropolitian	20
2C52	35	4.8	23.333017	85.3172	23.413017	85.3972	12.Mar.22	17:20	17:30	overcast	medium	1	Meal	motorcycle	1	No	Metropolitian	44
2E86	29	3.7	17.428294	78.404423	17.498294	78.474423	23-03-2022	19:55	20:10	high_wind	jam	2	Meal	scooter	2	Yes	Metropolitian	44
3449	40	4.9	18.55144	73.804855	18.63144	73.884855	14-03-2022	07:10PM	19:25	sandstorm	jam	1	Buffet	motorcycle	1	No	Metropolitian	27
AB43	31	4.3	17.438263	78.397865	17.568263	78.527865	2022.3.4	17:35	17:45	storm	medium	2	Meal	motorcycle	1	No	Metropolitian	33
40AE	32	3.5	12.325461	76.632278	12.395461	76.702278	8.Mar.22	05:10PM	05:15PM	storm	medium	1	Meal	motorcycle	1	No		34
A28C	21	5	18.546947	73.900626	18.586947	73.940626	24-03-2022	14:20	2:30 PM	overcast	high	2	Snack	scooter	0	No	Metropolitian	27
46C7	26	4.6	22.751857	75.866699	22.761857	75.876699	7.Mar.22	11:45	11:50 AM	foggy	high	0	Snack	motorcycle	1	No	Urban	20
A7DC	24	4.9	21.186884	72.793616	21.216884	72.823616	2022.3.15	5:15 PM	5:30 PM	foggy	medium	0	Buffet	motorcycle	1	X	Metropolitian	28
F6F1	26	4.6	-26.472001	80.354002	26.582001	80.464002	2022.2.18	17:35	17:50	overcast	medium	1	Buffet	motorcycle	0	No	Metropolitian	25
70FD			18.530963	73.828972	18.600963	73.898972	29.Mar.22		07:25PM			3	Meal	e_scooter	1	No	Metropolitian	26
5449	21	4.9	26.469003	80.316344	26.579003	80.426344	16-02-2022	18:25	18:40	storm	medium	1	Buffet	motorcycle	1	No	Metropolitian	26
4E	24	4.6	18.543626	73.905101	18.633626	73.995101	2022.3.29	22:30	10:45 PM	sandstorm	low	0	Buffet	motorcycle	1	No	Metropolitian	18
57CA	32	3.9	25.449872	81.836167	25.589872	81.976167	12-02-2022	21:55	22:05	high_wind	jam	2	Meal	motorcycle	1	No	Metropolitian	34
A38F	28	4.8	26.905287	75.794592	27.045287	75.934592	2022.3.6	21:40	21:50	overcast	jam	2	Snack	e_scooter	0	No	Metropolitian	37
3AD3	34	4.8	22.307898	73.167788	22.337898	73.197788	03-04-2022	06:40PM	18:55	high_wind	medium	1	Snack	scooter	1	No	Metropolitian	33
D4BC	21	4.8	18.51421	73.838429	18.62421	73.948429	02-04-2022	23:35	23:45	high_wind	low	2	Buffet	e_scooter	0	X	Urban	12
BC3E	32	3.5	23.351058	85.325731	23.441058	85.415731	29.Mar.22	9:30 PM	21:45	sandstorm	jam	1	Drinks	motorcycle	1	No	Metropolitian	35
BF18	26	4.7	25.449872	81.836167	25.559872	81.946167	16-02-2022	19:35	19:45	overcast	jam	0	Drinks	motorcycle	1	Yes	Semi-Urban	49
76B8	40	4.4	0	0	0.02	0.02	2022.3.21	9:20	9:35	storm	low	0	Snack	motorcycle	1	No		24

DATA ANALYSIS AND BASIC PREPROCESSING

Data should be made usable to be inserted in the model

Strings

id	rider_age	rider_rating	restaurant_latitude	restaurant_longitude	delivery_latitude	delivery_longitude	order_date	order_time	pickup_time	weather	traffic_density	vehicle_condition	order_type	vehicle_type	num_orders	is_festival	city	delivery_time
2D96	36	4.9	12.978453	77.643685	13.068453	77.733685	18-03-2022	22:40	22:50	clear_sky	low	1	Snack	scooter	1	No	Metropolitian	27
7F86	30	5	19.1093	72.825451	19.1193	72.835451	2022.3.1	11:25 AM	11:30	clear_sky	high	2	Drinks	scooter	1	No		17
CAB2	23	4.8	19.12663	72.829976	19.13663	72.839976	11.Mar.22	9:15	9:25	sandstorm	low	1	Snack	scooter	1	No	Metropolitian	15
860F	39	4.4	19.874733	75.353942	19.944733	75.423942	14-02-2022	7:35 PM	19:40	storm	jam	2	Drinks	scooter		No	Metropolitian	36
C0E3	27	4.8	17.429585	78.392621	17.499585	78.462621	6.Apr.22	20:40	8:45 PM	clear_sky	jam	2	Meal	scooter	0	No	Metropolitian	20
2C52	35	4.8	23.333017	85.3172	23.413017	85.3972	12.Mar.22	17:20	17:30	overcast	medium	1	Meal	motorcycle	1	No	Metropolitian	44
2E86	29	3.7	17.428294	78.404423	17.498294	78.474423	23-03-2022	19:55	20:10	high_wind	jam	2	Meal	scooter	2	Yes	Metropolitian	44
3449	40	4.9	18.55144	73.804855	18.63144	73.884855	14-03-2022	07:10PM	19:25	sandstorm	jam	1	Buffet	motorcycle	1	No	Metropolitian	27
AB43	31	4.3	17.438263	78.397865	17.568263	78.527865	2022.3.4	17:35	17:45	storm	medium	2	Meal	motorcycle	1	No	Metropolitian	33
40AE	32	3.5	12.325461	76.632278	12.395461	76.702278	8.Mar.22	05:10PM	05:15PM	storm	medium	1	Meal	motorcycle	1	No		34
A28C	21	5	18.546947	73.900626	18.586947	73.940626	24-03-2022	14:20	2:30 PM	overcast	high	2	Snack	scooter	0	No	Metropolitian	27
46C7	26	4.6	22.751857	75.866699	22.761857	75.876699	7.Mar.22	11:45	11:50 AM	foggy	high	0	Snack	motorcycle	1	No	Urban	20
A7DC	24	4.9	21.186884	72.793616	21.216884	72.823616	2022.3.15	5:15 PM	5:30 PM	foggy	medium	0	Buffet	motorcycle	1	X	Metropolitian	28
F6F1	26	4.6	-26.472001	80.354002	26.582001	80.464002	2022.2.18	17:35	17:50	overcast	medium	1	Buffet	motorcycle	0	No	Metropolitian	25
70FD			18.530963	73.828972	18.600963	73.898972	29.Mar.22		07:25PM			3	Meal	e_scooter	1	No	Metropolitian	26
5449	21	4.9	26.469003	80.316344	26.579003	80.426344	16-02-2022	18:25	18:40	storm	medium	1	Buffet	motorcycle	1	No	Metropolitian	26
4E	24	4.6	18.543626	73.905101	18.633626	73.995101	2022.3.29	22:30	10:45 PM	sandstorm	low	0	Buffet	motorcycle	1	No	Metropolitian	18
57CA	32	3.9	25.449872	81.836167	25.589872	81.976167	12-02-2022	21:55	22:05	high_wind	jam	2	Meal	motorcycle	1	No	Metropolitian	34
A38F	28	4.8	26.905287	75.794592	27.045287	75.934592	2022.3.6	21:40	21:50	overcast	jam	2	Snack	e_scooter	0	No	Metropolitian	37
3AD3	34	4.8	22.307898	73.167788	22.337898	73.197788	03-04-2022	06:40PM	18:55	high_wind	medium	1	Snack	scooter	1	No	Metropolitian	33
D4BC	21	4.8	18.51421	73.838429	18.62421	73.948429	02-04-2022	23:35	23:45	high_wind	low	2	Buffet	e_scooter	0	X	Urban	12
BC3E	32	3.5	23.351058	85.325731	23.441058	85.415731	29.Mar.22	9:30 PM	21:45	sandstorm	jam	1	Drinks	motorcycle	1	No	Metropolitian	35
BF18	26	4.7	25.449872	81.836167	25.559872	81.946167	16-02-2022	19:35	19:45	overcast	jam	0	Drinks	motorcycle	1	Yes	Semi-Urban	49
76B8	40	4.4	0	0	0.02	0.02	2022.3.21	9:20	9:35	storm	low	0	Snack	motorcycle	1	No		24

DATA ANALYSIS AND BASIC PREPROCESSING

Data should be made usable to be inserted in the model

Strings and numbers (Mixed)

id	rider_age	rider_rating	restaurant_latitude	restaurant_longitude	delivery_latitude	delivery_longitude	order_date	order_time	pickup_time	weather	traffic_density	vehicle_condition	order_type	vehicle_type	num_orders	is_festival	city	delivery_time
2D96	36	4.9	12.978453	77.643685	13.068453	77.733685	18-03-2022	22:40	22:50	clear_sky	low	1	Snack	scooter	1	No	Metropolitian	27
7F86	30	5	19.1093	72.825451	19.1193	72.835451	2022.3.1	11:25 AM	11:30	clear_sky	high	2	Drinks	scooter	1	No		17
CAB2	23	4.8	19.12663	72.829976	19.13663	72.839976	11.Mar.22	9:15	9:25	sandstorm	low	1	Snack	scooter	1	No	Metropolitian	15
860F	39	4.4	19.874733	75.353942	19.944733	75.423942	14-02-2022	7:35 PM	19:40	storm	jam	2	Drinks	scooter		No	Metropolitian	36
C0E3	27	4.8	17.429585	78.392621	17.499585	78.462621	6.Apr.22	20:40	8:45 PM	clear_sky	jam	2	Meal	scooter	0	No	Metropolitian	20
2C52	35	4.8	23.333017	85.3172	23.413017	85.3972	12.Mar.22	17:20	17:30	overcast	medium	1	Meal	motorcycle	1	No	Metropolitian	44
2E86	29	3.7	17.428294	78.404423	17.498294	78.474423	23-03-2022	19:55	20:10	high_wind	jam	2	Meal	scooter	2	Yes	Metropolitian	44
3449	40	4.9	18.55144	73.804855	18.63144	73.884855	14-03-2022	07:10PM	19:25	sandstorm	jam	1	Buffet	motorcycle	1	No	Metropolitian	27
AB43	31	4.3	17.438263	78.397865	17.568263	78.527865	2022.3.4	17:35	17:45	storm	medium	2	Meal	motorcycle	1	No	Metropolitian	33
40AE	32	3.5	12.325461	76.632278	12.395461	76.702278	8.Mar.22	05:10PM	05:15PM	storm	medium	1	Meal	motorcycle	1	No		34
A28C	21	5	18.546947	73.900626	18.586947	73.940626	24-03-2022	14:20	2:30 PM	overcast	high	2	Snack	scooter	0	No	Metropolitian	27
46C7	26	4.6	22.751857	75.866699	22.761857	75.876699	7.Mar.22	11:45	11:50 AM	foggy	high	0	Snack	motorcycle	1	No	Urban	20
A7DC	24	4.9	21.186884	72.793616	21.216884	72.823616	2022.3.15	5:15 PM	5:30 PM	foggy	medium	0	Buffet	motorcycle	1	X	Metropolitian	28
F6F1	26	4.6	-26.472001	80.354002	26.582001	80.464002	2022.2.18	17:35	17:50	overcast	medium	1	Buffet	motorcycle	0	No	Metropolitian	25
70FD			18.530963	73.828972	18.600963	73.898972	29.Mar.22		07:25PM			3	Meal	e_scooter	1	No	Metropolitian	26
5449	21	4.9	26.469003	80.316344	26.579003	80.426344	16-02-2022	18:25	18:40	storm	medium	1	Buffet	motorcycle	1	No	Metropolitian	26
4E	24	4.6	18.543626	73.905101	18.633626	73.995101	2022.3.29	22:30	10:45 PM	sandstorm	low	0	Buffet	motorcycle	1	No	Metropolitian	18
57CA	32	3.9	25.449872	81.836167	25.589872	81.976167	12-02-2022	21:55	22:05	high_wind	jam	2	Meal	motorcycle	1	No	Metropolitian	34
A38F	28	4.8	26.905287	75.794592	27.045287	75.934592	2022.3.6	21:40	21:50	overcast	jam	2	Snack	e_scooter	0	No	Metropolitian	37
3AD3	34	4.8	22.307898	73.167788	22.337898	73.197788	03-04-2022	06:40PM	18:55	high_wind	medium	1	Snack	scooter	1	No	Metropolitian	33
D4BC	21	4.8	18.51421	73.838429	18.62421	73.948429	02-04-2022	23:35	23:45	high_wind	low	2	Buffet	e_scooter	0	X	Urban	12
BC3E	32	3.5	23.351058	85.325731	23.441058	85.415731	29.Mar.22	9:30 PM	21:45	sandstorm	jam	1	Drinks	motorcycle	1	No	Metropolitian	35
BF18	26	4.7	25.449872	81.836167	25.559872	81.946167	16-02-2022	19:35	19:45	overcast	jam	0	Drinks	motorcycle	1	Yes	Semi-Urban	49
76B8	40	4.4	0	0	0.02	0.02	2022.3.21	9:20	9:35	storm	low	0	Snack	motorcycle	1	No		24

DATA ANALYSIS AND BASIC PREPROCESSING

Missing values handling strategy

DATA ANALYSIS AND BASIC PREPROCESSING

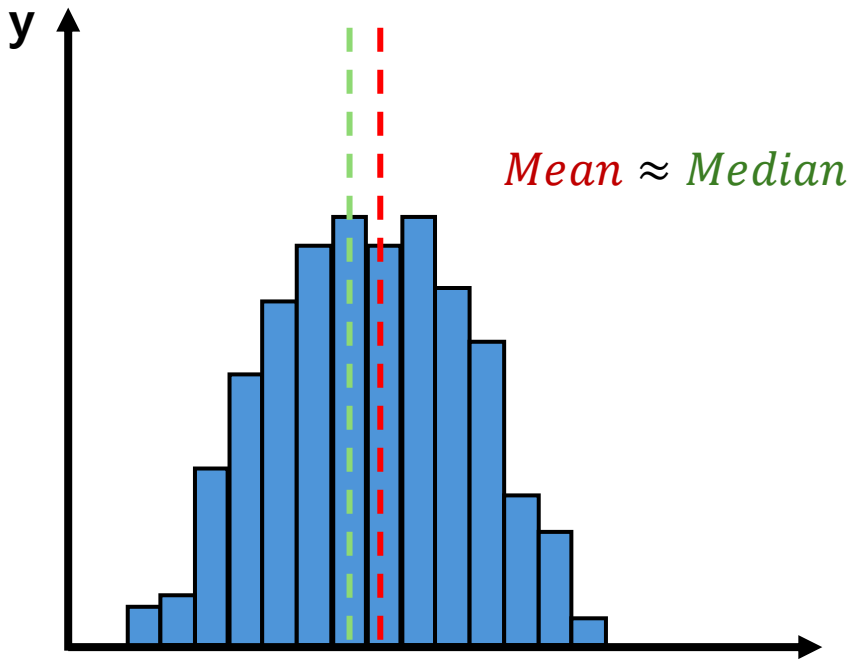
Missing values handling strategy

DATA ANALYSIS AND BASIC PREPROCESSING

Missing values handling strategy

1) Mean Imputation

Works well when the variable is roughly symmetric and has no strong outliers.

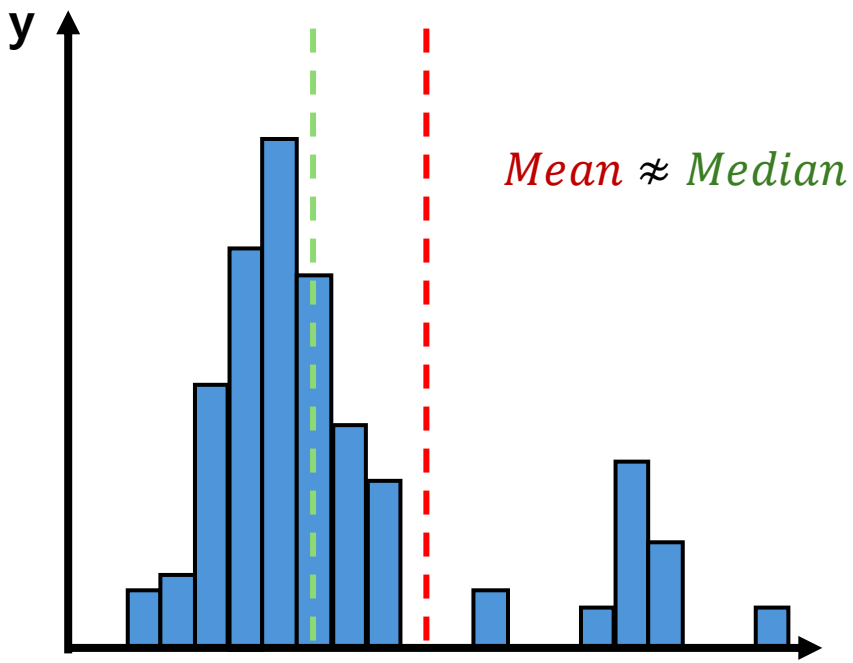


DATA ANALYSIS AND BASIC PREPROCESSING

Missing values handling strategy

2) Median Imputation

Works well when the distribution is skewed or contains outliers.



Mean is pulled by outliers, but median remains stable

DATA ANALYSIS AND BASIC PREPROCESSING

Missing values handling strategy

3) Missing Indicator + Imputation

Ex 1) Income

Missing income may indicate unstable or hidden income

: These people tend to have low credit

	Income	…	credit (target)
A	300M	…	
B	NaN	…	
C	500M	…	
…	…	…	…
N	NaN	…	
M	450M	…	

Ex 2) Purchase amount

Missing purchase_amount may indicate they are new customers

: These people tend to spend low amount

	Purchase_amount	…	Amount_spent
A	2	…	
B	8	…	
C	10	…	
…	…	…	…
N	NaN	…	
M	NaN	…	

DATA ANALYSIS AND BASIC PREPROCESSING

Data should be made usable to be inserted in the model

Column rider_rating is only consisted of float and missing values

id	rider_rating
2D96	4.9
7F86	5
CAB2	4.8
860F	4.4
C0E3	4.8
2C52	4.8
2E86	3.7
3449	4.9
AB43	4.3
40AE	3.5
A28C	5
46C7	4.6
A7DC	4.9
F6F1	4.6
70FD	
5449	4.9
4E	4.6
57CA	3.9
A38F	4.8
3AD3	4.8
D4BC	4.8
BC3E	3.5
BF18	4.7
76B8	4.4

DATA ANALYSIS AND BASIC PREPROCESSING

Data should be made usable to be inserted in the model

Column weather is only consisted of string and missing values

id	weather
2D96	clear_sky
7F86	clear_sky
CAB2	sandstorm
860F	storm
C0E3	clear_sky
2C52	overcast
2E86	high_wind
3449	sandstorm
AB43	storm
40AE	storm
A28C	overcast
46C7	foggy
A7DC	foggy
F6F1	overcast
70FD	
5449	storm
4E	sandstorm
57CA	high_wind
A38F	overcast
3AD3	high_wind
D4BC	high_wind
BC3E	sandstorm
BF18	overcast
76B8	storm

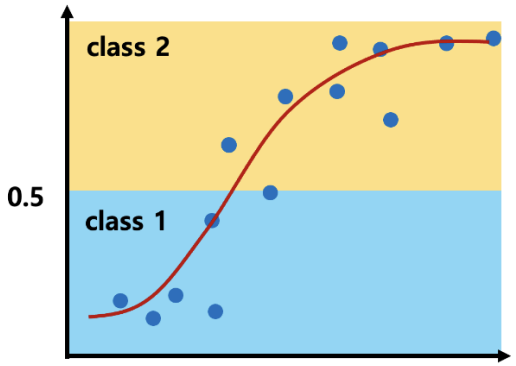
Categorical Variables

Label encoding in Regression

FAILURE OF DISCRIMINATIVE MODEL

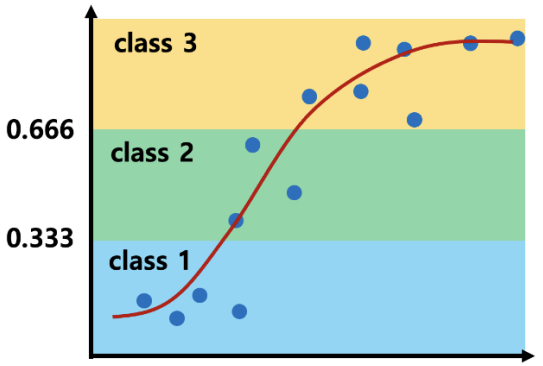
The methods in this section can be naturally extended to the case of more than two response classes.

Logistic Regression (O)



$P(Y=1|X) = 1 - P(Y=2|X)$

Logistic Regression (X)



Probable : ice (1) – water (2) – vapor (3)

Improbable : cat (1) – dog (2) – bird(3)

Automatically assigns the hierarchy between more than 2 classes

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). An Introduction to Statistical Learning with Python (ISLP). Springer.

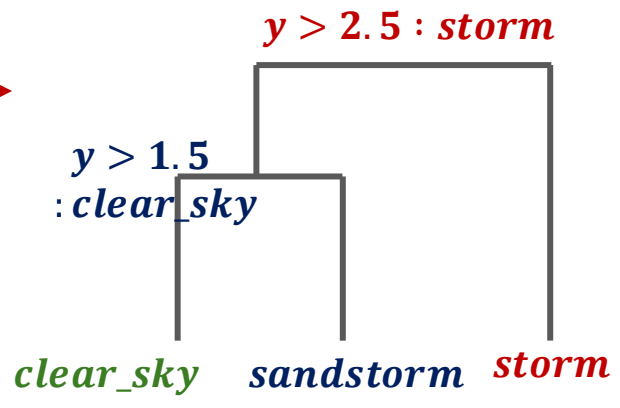
but label encoding can still create arbitrary splits. Tree models are less sensitive to ordering,

Label encoding (in Tree)

clear_sky : 1

sandstorm : 2

storm : 3



$$y = w * weather + b \quad (1D)$$

The model assumes clear_sky < sandstorm < storm: Label encoding forces a linear relationship between categories that does not exist.

DATA ANALYSIS AND BASIC PREPROCESSING

Data should be made usable to be inserted in the model

Column weather is only consisted of string and missing values

id	weather
2D96	clear_sky
7F86	clear_sky
CAB2	sandstorm
860F	storm
C0E3	clear_sky
2C52	overcast
2E86	high_wind
3449	sandstorm
AB43	storm
40AE	storm
A28C	overcast
46C7	foggy
A7DC	foggy
F6F1	overcast
70FD	
5449	storm
4E	sandstorm
57CA	high_wind
A38F	overcast
3AD3	high_wind
D4BC	high_wind
BC3E	sandstorm
BF18	overcast
76B8	storm

Categorical Variables

One Hot Encoding (OHE)

clear_sky : [0,0,1]

sandstorm : [0,1,0]

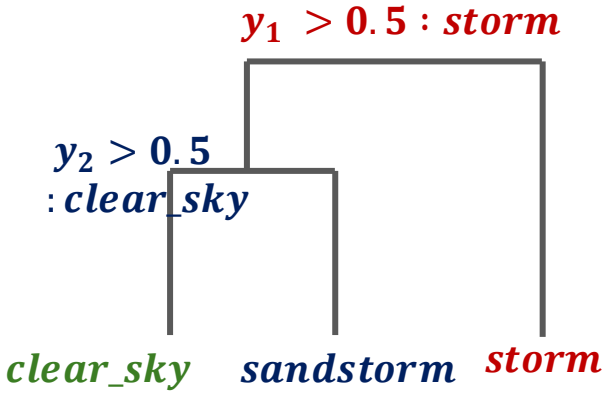
storm : [1,0,0]

$$y = w_1 * weather1 + w_2 * weather2 + w_3 * weather3 + b \quad (3D)$$

Each weather category is independently represented in the orthogonal axes.

If weather1 = 1, then all the other weathers are 0. : *No relationship between the categories in OHE.*

$y_1 = is\_storm$
$y_2 = is\_clear\_sky$



each value has closer meaning to the probability in OHE

DATA ANALYSIS AND BASIC PREPROCESSING

Data should be made usable to be inserted in the model

Column order_date is only consisted of float and missing values

id	order_date
2D96	18-03-2022
7F86	2022.3.1
CAB2	11.Mar.22
860F	14-02-2022
C0E3	6.Apr.22
2C52	12.Mar.22
2E86	23-03-2022
3449	14-03-2022
AB43	2022.3.4
40AE	8.Mar.22
A28C	24-03-2022
46C7	7.Mar.22
A7DC	2022.3.15
F6F1	2022.2.18
70FD	29.Mar.22
5449	16-02-2022
4E	2022.3.29
57CA	12-02-2022
A38F	2022.3.6
3AD3	03-04-2022
D4BC	02-04-2022
BC3E	29.Mar.22
BF18	16-02-2022
76B8	2022.3.21

order_date

18-03-2022

2022.3.1

11.Mar.22 → 2011. 03. 22 ?

11.Mar.22 → 2022. 03. 11 ?

...

How do you know ?

DATA ANALYSIS AND BASIC PREPROCESSING

Data should be made usable to be inserted in the model

How the longitude value alone can be the clue of delivery time prediction?

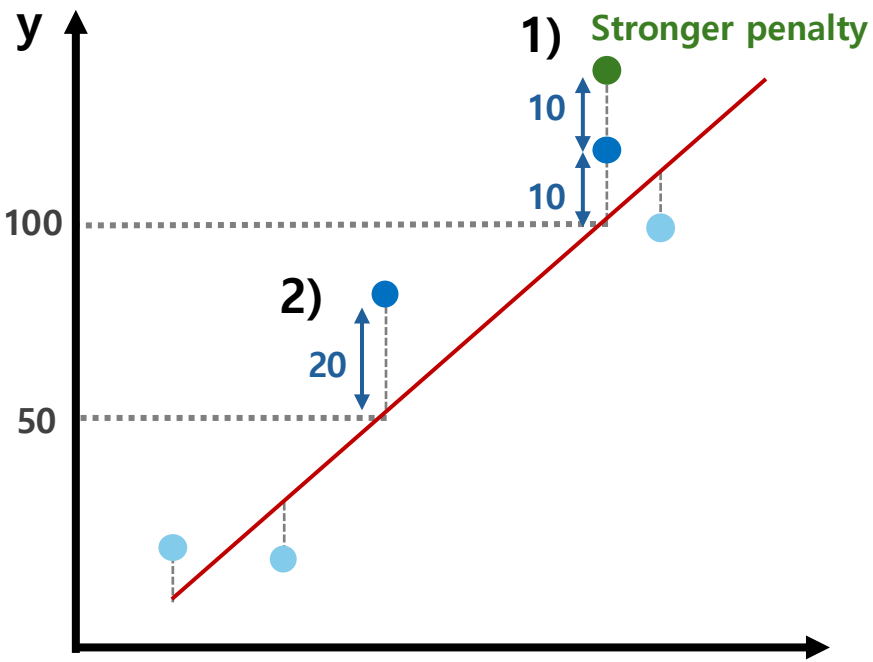
id	restaurant_latitude	restaurant_longitude	delivery_latitude	delivery_longitude
2D96	12.978453	77.643685	13.068453	77.733685
7F86	19.1093	72.825451	19.1193	72.835451
CAB2	19.12663	72.829976	19.13663	72.839976
860F	19.874733	75.353942	19.944733	75.423942
C0E3	17.429585	78.392621	17.499585	78.462621
2C52	23.333017	85.3172	23.413017	85.3972
2E86	17.428294	78.404423	17.498294	78.474423
3449	18.55144	73.804855	18.63144	73.884855
AB43	17.438263	78.397865	17.568263	78.527865
40AE	12.325461	76.632278	12.395461	76.702278
A28C	18.546947	73.900626	18.586947	73.940626
46C7	22.751857	75.866699	22.761857	75.876699
A7DC	21.186884	72.793616	21.216884	72.823616
F6F1	-26.472001	80.354002	26.582001	80.464002
70FD	18.530963	73.828972	18.600963	73.898972
5449	26.469003	80.316344	26.579003	80.426344
4E	18.543626	73.905101	18.633626	73.995101
57CA	25.449872	81.836167	25.589872	81.976167
A38F	26.905287	75.794592	27.045287	75.934592
3AD3	22.307898	73.167788	22.337898	73.197788
D4BC	18.51421	73.838429	18.62421	73.948429
BC3E	23.351058	85.325731	23.441058	85.415731
BF18	25.449872	81.836167	25.559872	81.946167
76B8	0	0	0.02	0.02

EVALUATION METRICS

MSE (Mean Square Error) / RMSE (Root Mean Square Error)

$$MSE = \frac{\sum_{i=1}^{n}(y-\hat{y})^2}{n} \qquad RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y-\hat{y})^2}{n}}$$

Linear regression



1) Large errors are penalized much more due to squaring

$$MSE \propto (y-\hat{y})^2 \qquad \begin{matrix} if\ (y-\hat{y}) = 10 \rightarrow 100 \\ if\ (y-\hat{y}) = 20 \rightarrow 400 \end{matrix}$$

2) Does not consider relative scale of the target

$$\begin{matrix} y = 50\ \&\ (y-\hat{y}) = 20 & \rightarrow 40\%\ \text{difference} \\ y = 100\ \&\ (y-\hat{y}) = 20 & \rightarrow 20\%\ \text{difference} \end{matrix} \quad : same\ penalty$$

MSE is better when Absolute Deviation is more important

Ex) car price prediction

: Even though the percentage error is larger, the actual money difference is smaller.

- **Your budget is fixed**
- **False prediction 1 : 8M to 10M (2M deviation, 25% err)**
- **False prediction 2 : 3M to 4M (1M deviation, 33.3% err)**

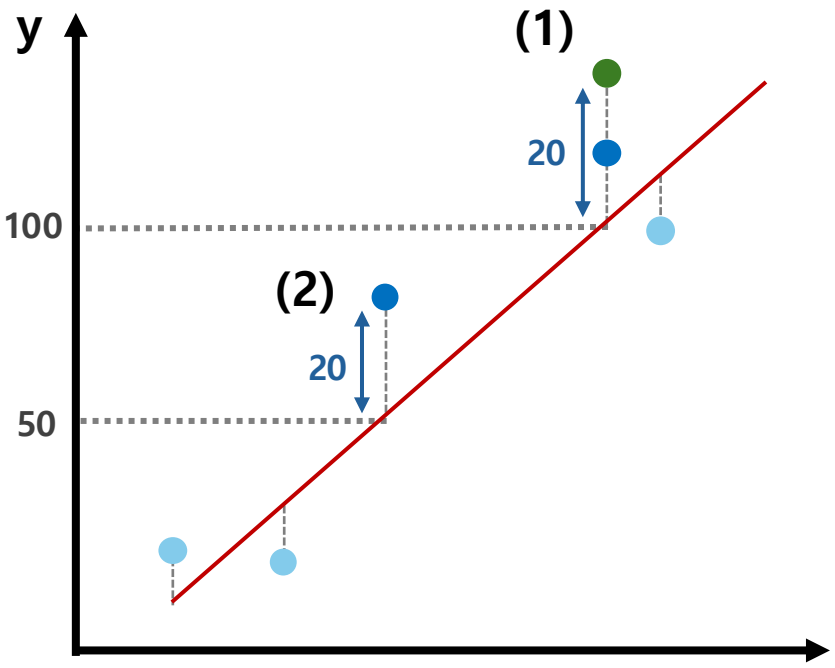
: Prediction 2 is preferred because it only deviates 1M from the true price.

EVALUATION METRICS

MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y-\hat{y}}{y}\right| \quad : \textit{penalty} \text{ is normalized by the true value y}$$

Linear regression



For the point (1) & (2) which has the same deviation value,

$$y = 50\ \&\ \frac{(y-\hat{y})}{y} = \frac{20}{50} = 0.4 \rightarrow (\sim 40\%\ err)\ \text{Larger penalty}$$

$$y = 100\ \&\ \frac{(y-\hat{y})}{y} = \frac{20}{100} = 0.2 \rightarrow (\sim 20\%\ err)\ \text{Smaller penalty}$$

Recall that)

$y = 50\ \&\ (y-\hat{y}) = 20 \rightarrow 40\%$ difference
$y = 100\ \&\ (y-\hat{y}) = 20 \rightarrow 20\%$ difference
: *For MSE*

MSE point of view **MAPE point of view**

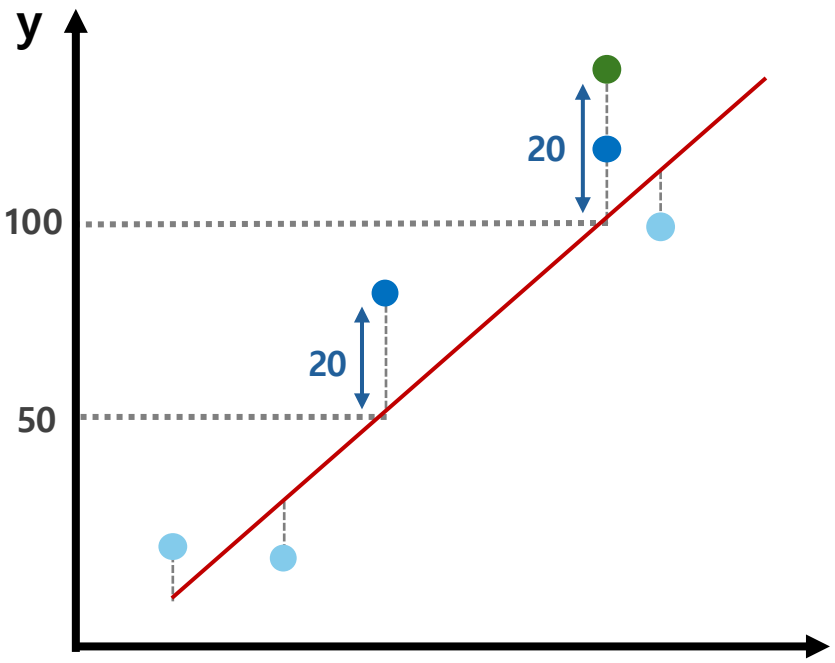
: Relative error matters more than absolute error

EVALUATION METRICS

MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y-\hat{y}}{y}\right|$$: *penalty* is normalized by the true value y

Linear regression



Recall that)

$y = 50 \;\&\; (y-\hat{y}) = 20 \rightarrow$ 40% difference

$y = 100 \;\&\; (y-\hat{y}) = 20 \rightarrow$ 20% difference

: *For MSE*

MSE point of view

MAPE point of view

: Relative error matters more than absolute error

MAPE is better when Relative Deviation is more important

Ex) delivery time prediction

- You are really hungry and impatient

- False prediction 1 : 10 min to 20 min (10 min deviation, 100% err)

- False prediction 2 : 50 min to 60 min (10 min deviation, 20% err)

: Prediction 2 is preferred because either 50 min or 60 min seems similar.

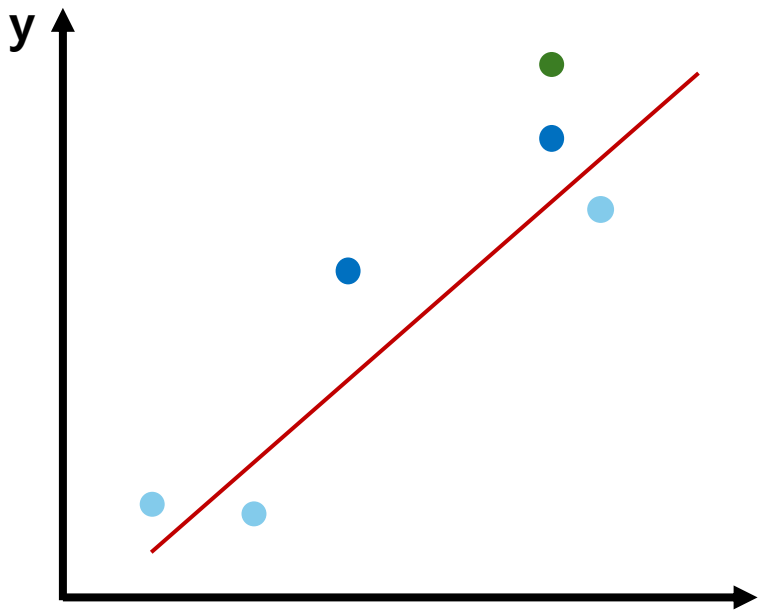
MODEL OVERVIEW

1) Linear Regression

gradual change in x

$$y = w_1 x_1 + w_2 x_2$$

gradual change in y



- Smooth change and has low variance

- Capable of extrapolation

: Can make reasonable predictions even outside the training range.

Example)

If housing prices in the training data range from 100M to 1B KRW, the model can still make a reasonable prediction for 1.2B KRW.

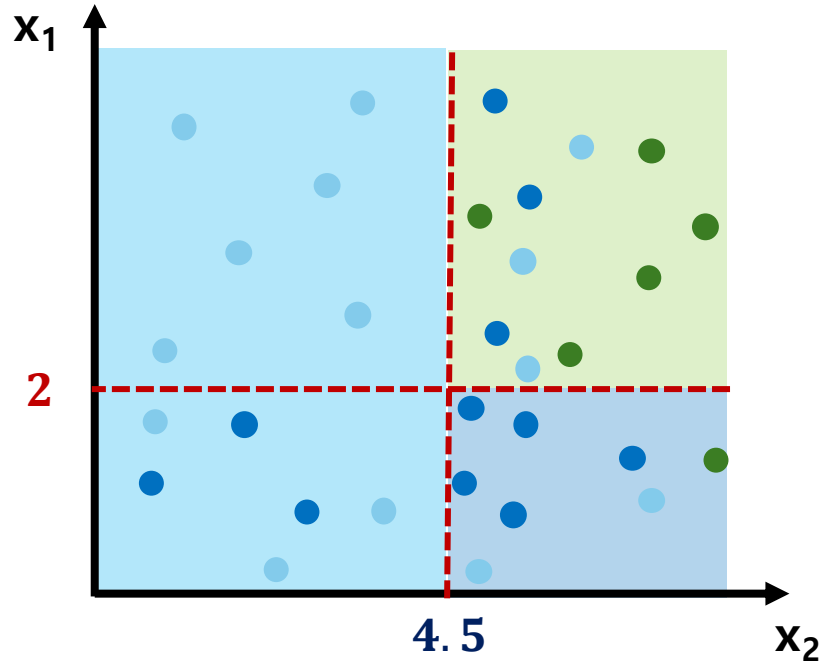
- Limited ability to model feature interactions

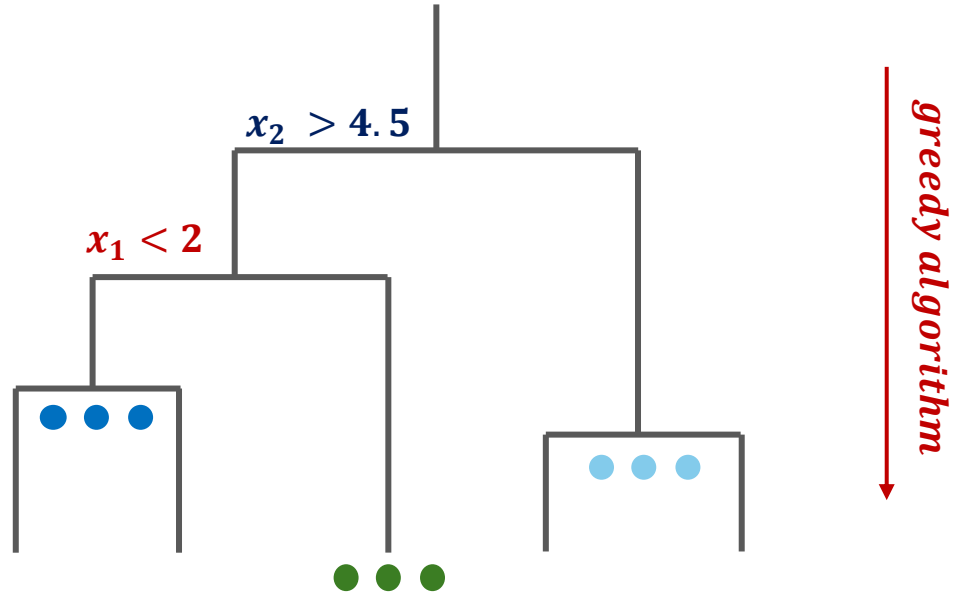
- Capture only simple relationship

MODEL OVERVIEW

2) Decision Tree

- Automatic modeling of nonlinearities and interactions
- Ability to model complex decision boundaries Ex) $x_1 > 2$ & $x_2 < 4.5$
- Discontinuous predictions (stepwise behavior)
- High variance (model instability)
- Poor extrapolation ability
- Easy to be overfitted



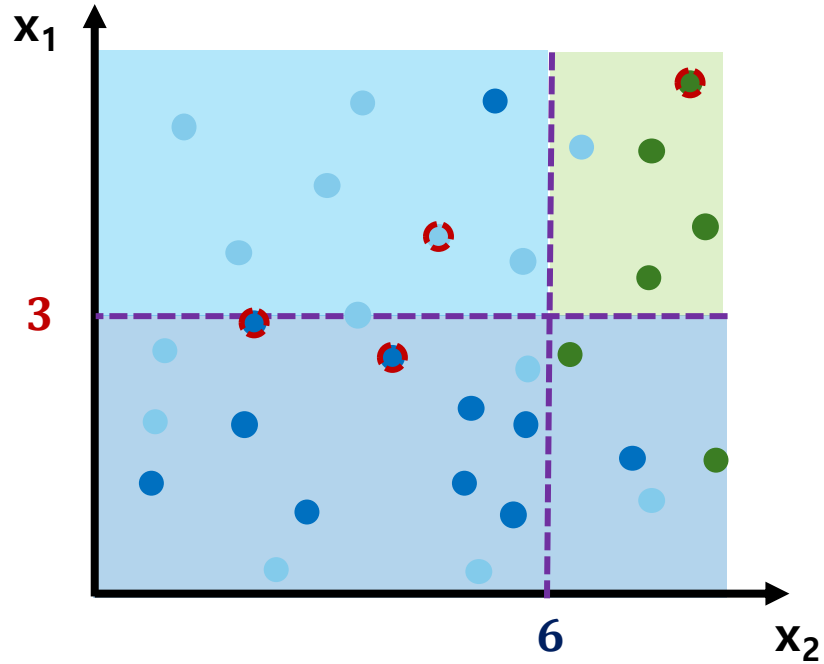


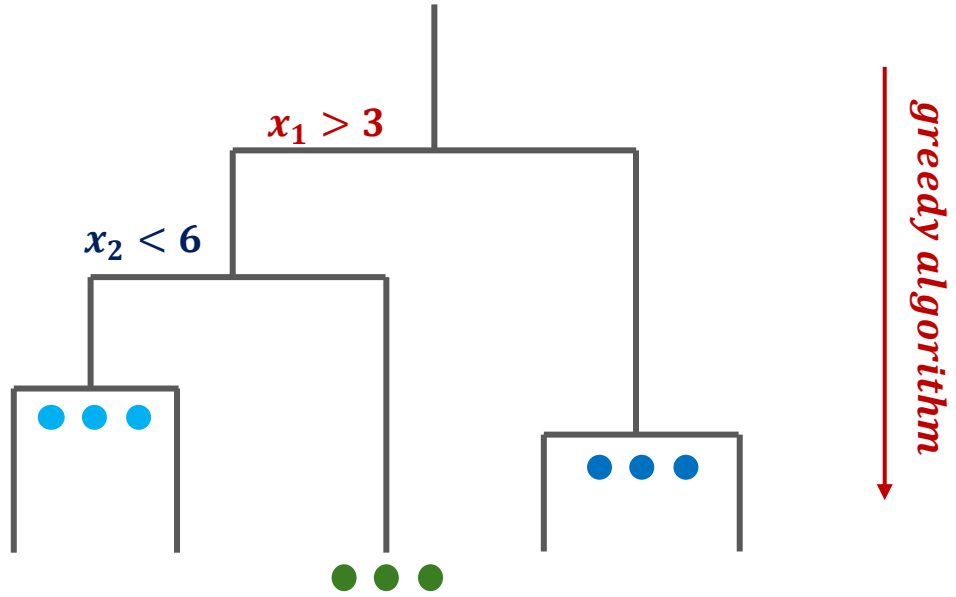
MODEL OVERVIEW

2) Decision Tree

- Automatic modeling of nonlinearities and interactions
- Ability to model complex decision boundaries Ex) $x_1 > 2$ & $x_2 < 4.5$
- Discontinuous predictions (stepwise behavior)
- High variance (model instability)
- Poor extrapolation ability
- Easy to be overfitted

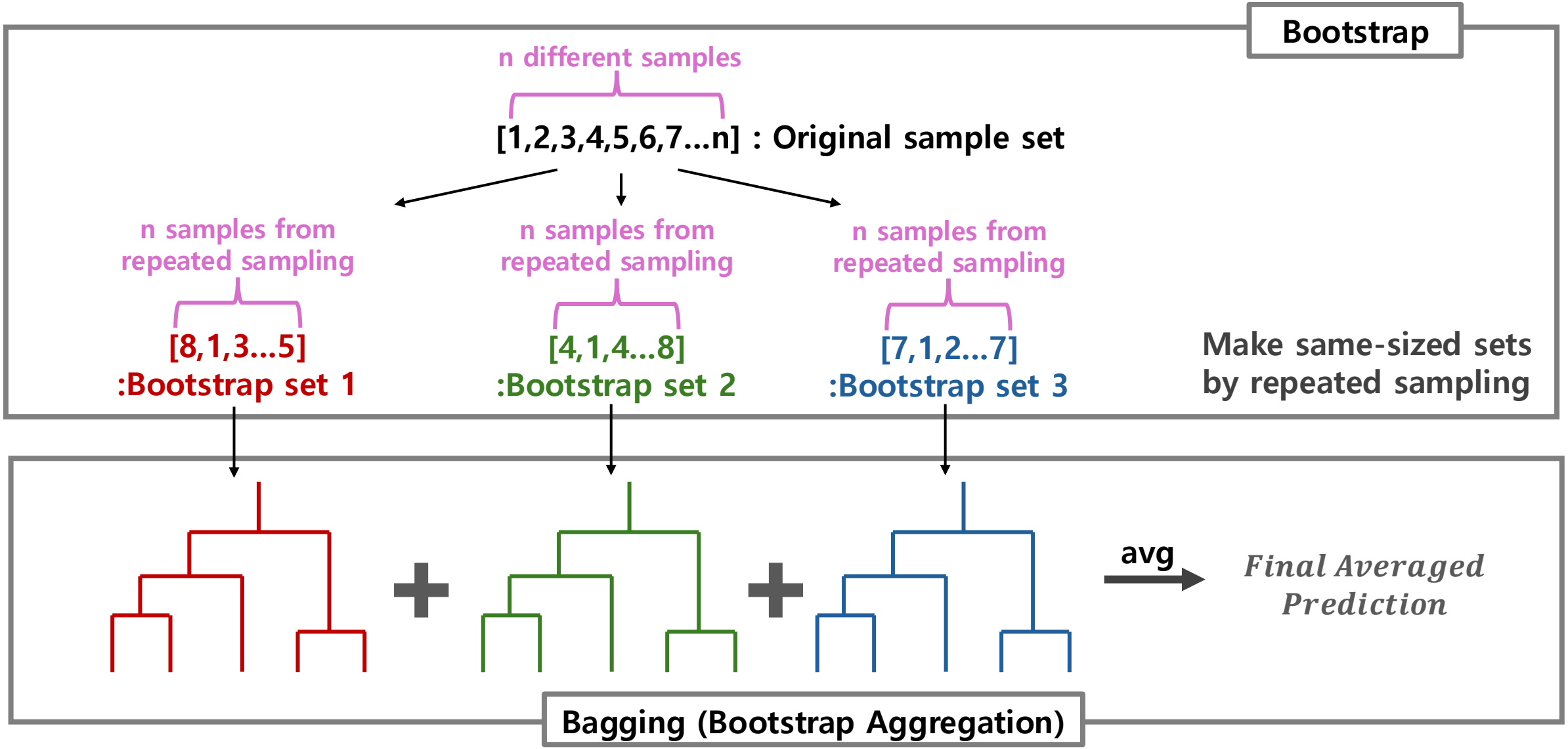
Only 4 points changed but the domain is completely altered in a discrete way





MODEL OVERVIEW

3) Random Forest

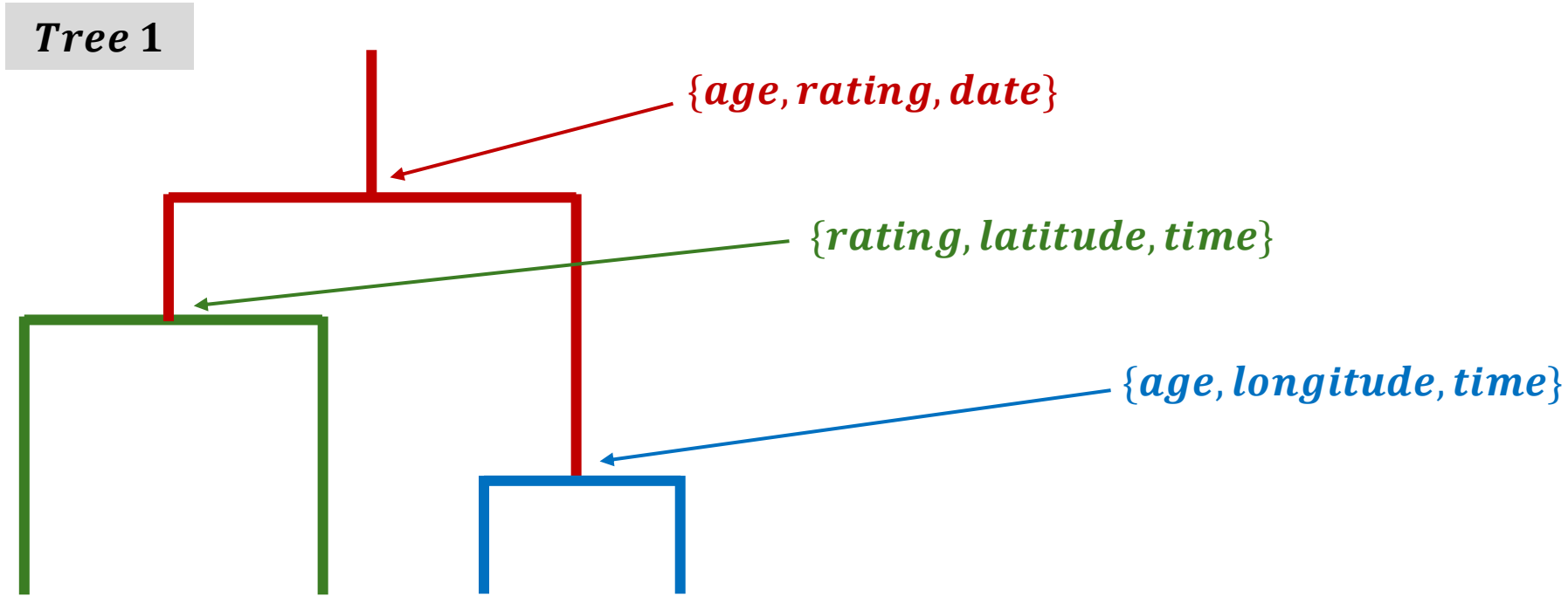


MODEL OVERVIEW

3) Random Forest

Also at each split, only a random subset of features is considered so that not all trees are very similar.

$feature\ ex) \{age, rating, latitude, longitude, date\ , time\}$



Both the training data and the candidate features are randomized in Random Forest.

MODEL OVERVIEW

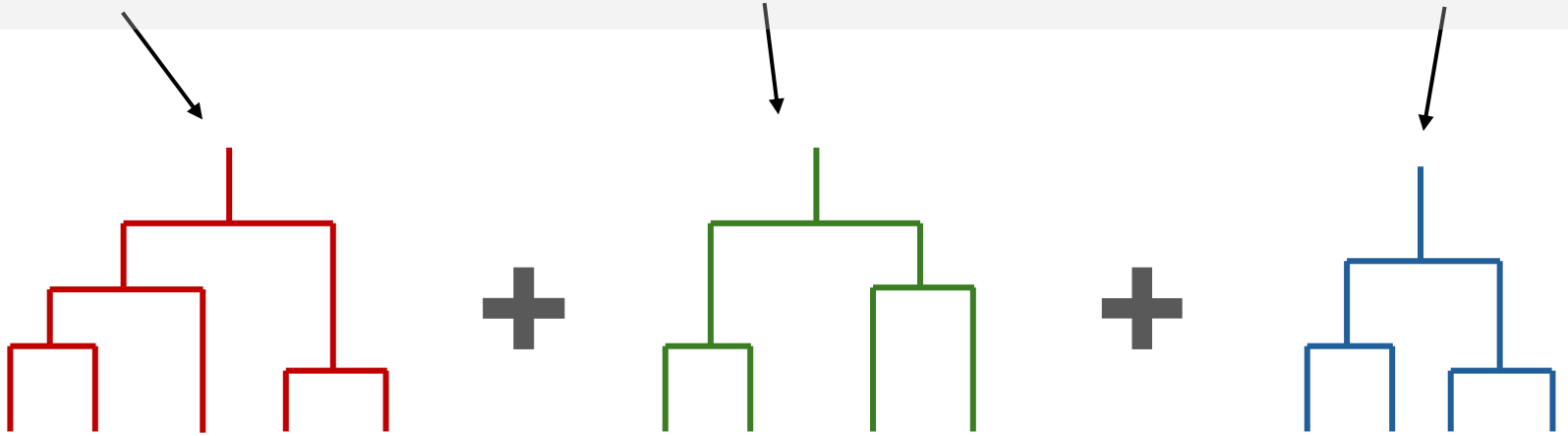
4) XGBoost : One of the Boosting Models

Starting from a weak tree, each new tree is added to fit the residual of the current model. ≡ *Boosting*

$$\hat{y}_i^{(0)} = c \qquad \hat{y}_i^{(1)} = \hat{y}_i^{(0)} + f_1(x_i) \qquad \hat{y}_i^{(2)} = \hat{y}_i^{(1)} + f_2(x_i) \qquad \hat{y}_i^{(T)} = \sum_{t=1}^{T} f_t(x_i)$$

Weak tree *Addtional tree 1* *Addtional tree 2* ≡ *Ensemble*

Serial addition of trees



- average
- weighted average
- majority vote
- weighted sum
- stacking(meta model)
- blending

Final Ensembled Prediction

Serially add trees to fit the residual

In Random Forest, the trees are parallelly made

MODEL OVERVIEW

4) XGBoost : One of the Boosting Models

Starting from a weak tree, each new tree is added to fit the residual of the current model. ≡ *Boosting*

$$\hat{y}_i^{(0)} = c \qquad \hat{y}_i^{(1)} = \hat{y}_i^{(0)} + f_1(x_i) \qquad \hat{y}_i^{(2)} = \hat{y}_i^{(1)} + f_2(x_i) \qquad \hat{y}_i^{(T)} = \sum_{t=1}^{T} f_t(x_i)$$

Weak tree — *Addtional tree 1* — *Addtional tree 2* — ≡ *Ensemble*

Serial addition of trees

Objective Function : Target function that has to be minimized ~ error, complexity etc

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t)$$

$l(y_i, \hat{y}_i)$: prediction error (loss function)

$\Omega(f_t)$: regularization term (tree complexity penalty)

For each step, additional trees are designed to minimize the objective function

Smaller error is preferred

Less complex model is preferred
(Complex model tends to be overfitted)

By balancing between the complexity and loss minimization, the next tree is determined to minimize L.

Ex) $\min_{\theta} L(\theta)$ —by→ $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$: *θ is updated to the direction that minimizes L the most*

the opposite direction to the gradient

MODEL OVERVIEW

4) XGBoost : One of the Boosting Models

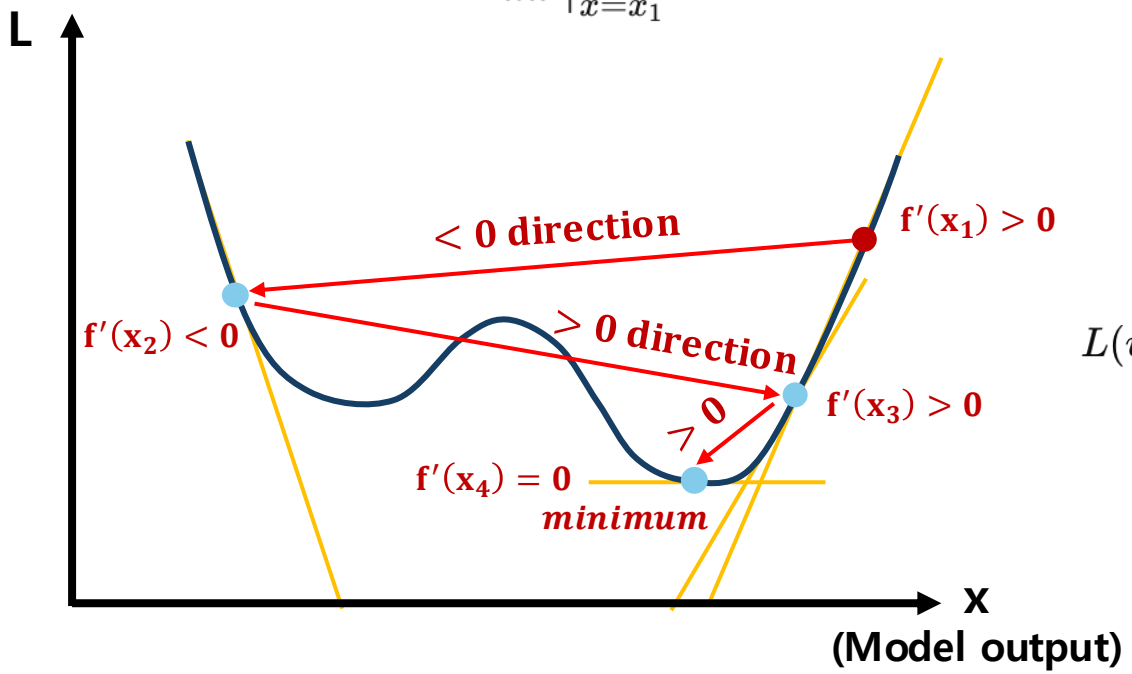
objective function (error + complexity)

Error ↓ ⇔ Accuracy ↑

Gradient : The direction of steepest increase of the loss. (the opposite direction minimizes the loss)

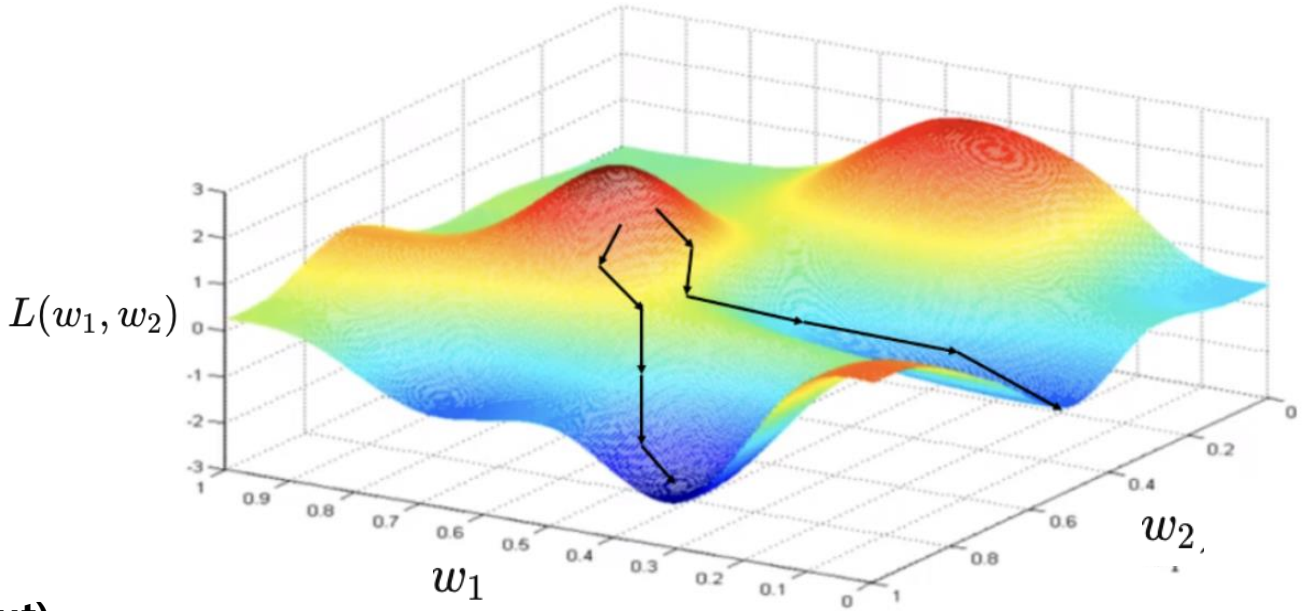
1D Function (for intuition)

$$\text{Gradient} \equiv \left.\frac{df}{dx}\right|_{x=x_1} = f'(x_1)$$



2D Function (for intuition)

$$\text{Gradient} \equiv \nabla L(w_1, w_2) = \left(\frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2}\right)$$



MODEL OVERVIEW

4) XGBoost : One of the Boosting Models

For Boosting models,

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t)$$

prediction y for n samples

$$\frac{\partial \mathcal{L}}{\partial \hat{y}_i} = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i}$$

: minimize the loss by following the negative gradient.

A new tree is trained to approximate the negative gradient.

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \eta f_t(x)$$

however, the new tree is scaled by learning rate η before being added

The learning rate controls how much each new tree updates the model.

Smaller values (e.g., 0.05–0.1) lead to more stable learning but require more trees, while larger values can speed up training but risk overfitting.

learning rate η effectively shrinks the contribution of each tree, making the boosting process more robust.

MODEL OVERVIEW

4) XGBoost : One of the Boosting Models

Tunable Parameters

eta : Learning rate - Controls how much each tree contributes to the model

$n\_estimators$: Number of Trees – Number of Boosting iterations

max_depth : Maximum depth – Maximum depth of each trees

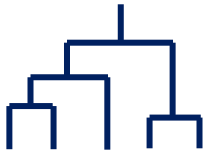
$subsample$: Subsample – Fraction of data used per tree

$colsample\_bytree$: Column Sampling – Fraction of features (predictors) used per tree

$lambda$, $alpha$: Regularization – Controls model complexity

min_child_weight : Minimum Child Weight – Minimum samples required in a leaf

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \eta f_t(x) \longrightarrow$$



id	rider_age	rider_rating	restaurant_latitude	restaurant_longitude	delivery_latitude	delivery_longitude	order_date	order_time	pickup_time
2D96	36	4.9	12.978453	77.643685	13.068453	77.733685	18-03-2022	22:40	22:50
7F86	30	5	19.1093	72.825451	19.1193	72.835451	2022.3.1	11:25 AM	11:30
CAB2	23	4.8	19.12663	72.829976	19.13663	72.839976	11.Mar.22	9:15	9:25
860F	39	4.4	19.874733	75.353942	19.944733	75.423942	14-02-2022	7:35 PM	19:40
C0E3	27	4.8	17.429585	78.392621	17.499585	78.462621	6.Apr.22	20:40	8:45 PM
2C52	35	4.8	23.333017	85.3172	23.413017	85.3972	12.Mar.22	17:20	17:30
2E86	29	3.7	17.428294	78.404423	17.498294	78.474423	23-03-2022	19:55	20:10
3449	40	4.9	18.55144	73.804855	18.63144	73.884855	14-03-2022	07:10PM	19:25
AB43	31	4.3	17.438263	78.397865	17.568263	78.527865	2022.3.4	17:35	17:45
40AE	32	3.5	12.325461	76.632278	12.395461	76.702278	8.Mar.22	05:10PM	05:15PM
A28C	21	5	18.546947	73.900626	18.586947	73.940626	24-03-2022	14:20	2:30 PM
46C7	26	4.6	22.751857	75.866699	22.761857	75.876699	7.Mar.22	11:45	11:50 AM
A7DC	24	4.9	21.186884	72.793616	21.216884	72.823616	2022.3.15	5:15 PM	5:30 PM
F6F1	26	4.6	-26.472001	80.354002	26.582001	80.464002	2022.2.18	17:35	17:50

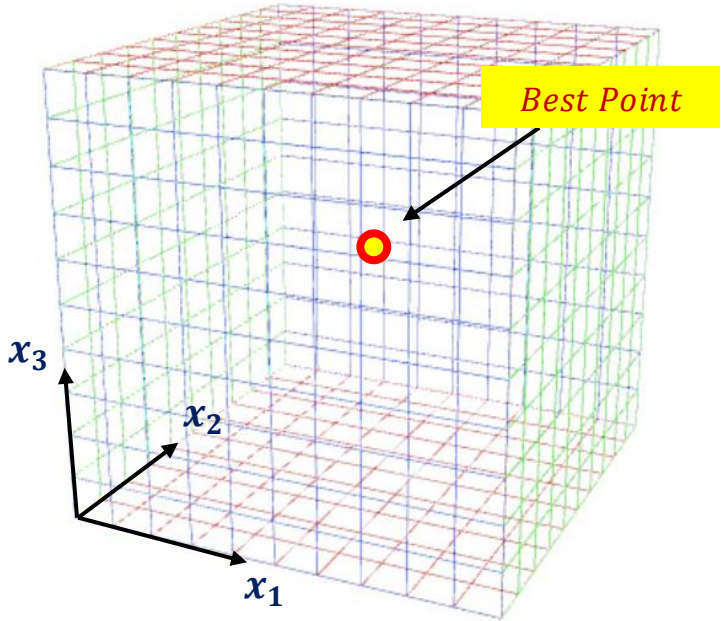
MODEL OVERVIEW

4) XGBoost : One of the Boosting Models

- *eta*
- *n_estimators*
- *max_depth*
- *subsample*
- *colsample_bytree*
- *min_child_weight*

How can we search the best parameter set?

Ex) parameter x_1, x_2, x_3 (3D)



Submit Your First Entry !