Percolata Corporation
Programming Assignment

Background: Percolata installs sensors in stores that record videos and upload to our server. Next, we train our sensors by using machine-learning algorithm to count people walking into the store (video walk-in) and walking out of the store (video walk-out). Usually, video walk-in is accurate enough to be the final walk-in result. However, sometimes there might be problems like the one below:

Something might block the camera and it would affect the accuracy of video walk-in, like the blue and yellow balloons in the above picture. In this case, directly using video walk-in as the final walk-in result is not proper.

Goal: Find a way to improve the accuracy of final walk-in using video walk-in and some other information collected by our devices.

Task: Use the information given in the attached *train_data.csv* to build a machine-learning model to get better final walk-in result. Then, test your model on the attached *test_data.csv*.

Data: Now take a quick look at the *train_data.csv*:

A	B	C	D	E	F	G	H	I	J	K	L	M	N
device_ang	distance_t	AM_or_PM	mall_or_st	average_pe	video_walk	video_walk	predict_wa	predict_wa	wifi_walki	wifi_walko	sales_in_n	sales_in_n	groundtruth
1	1	1	1	3384.5	17	21	27	23	21	20	4	2	24
3	2	0	2	1811	1	1	3	0	7	0	0	1	1
3	1	1	2	1506	9	9	10	11	35	42	1	1	4
3	1	1	1	1327.5	17	19	0	0	9	9	5	3	29
2	2	1	1	1073.25	6	6	3	0	2	5	2	4	0
3	1	1	1	1327.5	60	60	50	46	13	15	5	4	53
2	2	0	2	1224	5	9	8	8	3	5	0	0	8
2	2	0	1	4077.5	1	3	3	4	1	1	1	1	9
2	2	1	1	1252.5	26	14	21	20	4	4	3	4	34
3	1	0	2	1506	10	9	10	10	23	24	2	2	4
2	1	1	1	1161	24	24	27	28	33	38	2	3	48
3	2	1	2	1811	5	0	4	7	0	2	0	0	5
2	2	0	2	1224	2	4	6	5	3	2	0	0	1
2	2	0	1	4077.5	4	3	2	2	1	0	2	0	3
2	2	1	1	3571.5	1	3	0	0	1	0	0	0	0
2	2	1	1	890	5	4	2	2	1	2	1	2	4
2	2	1	1	3571.5	0	0	0	0	2	4	0	0	0
3	2	0	2	1811	6	7	7	9	5	5	1	3	5
2	2	1	1	1073.25	6	7	0	0	4	2	1	0	0

Here are some explanations for columns in the *train_data.csv*:

Title	Description
device angle:	Angle of camera comparing to horizontal. (1: "< 30", 2: "30 - 60", 3: "60 - 90")
distance_to_door	Direct distance to the store door. (1: "< 2m", 2: "2 - 4m", 3:"> 4m")
AM_or_PM	Moring or afternoon. (0: "AM", 1: "PM")
mall_or_street	Store located by the street or inside a mall. (1: "Mall", 2: "Street")
average person size	Average person pixel in foreground. (p.s. our video is 480*320)
video_walkin	Walk-in counted from video using computer vision algorithm.
video_walkout	Walk-out counted from video using computer vision algorithm.
predict_walkin	Walk-in predicted by traffic prediction algorithm.
predict_walkout	Walk-out predicted by traffic prediction algorithm.
wifi_walkin	Walk-in counted using Wi-Fi signature.
wifi_walkout	Walk-out counted using Wi-Fi signature.
sales_in_next_15_min	Transaction counts in next 15 minutes.
sales_in_next_15_to_30_min	Transaction counts between next 15 min and 30 min.
groundtruth_walkin	Real walk-in.

Submission:

1) **Python script:**
   It has two arguments: *train_data_file_path test_data_file_path*
   E.g. run the script: *python script.py train_data.csv test_data.csv*

2) **Test data file:**
   Append the final walk-in result generated by your model in the *test_data.csv* file.

   Before run your script, *test_data.csv* looks like:

A	B	C	D	E	F	G	H	I	J	K	L	M	N
device_ang	distance_t	AM_or_PM	mall_or_st	average_pe	video_walk	video_walk	predict_wa	predict_wa	wifi_walki	wifi_walko	sales_in_n	sales_in_next_15_to_3	
3	1	0	1	1327.5	13	11	0	0	9	5	2	2	
2	2	0	1	1073.25	4	7	32	33	0	3	1	3	
3	1	1	1	1327.5	49	48	45	39	15	19	6	3	
3	1	1	1	2299.5	17	22	21	21	7	5	1	0	
2	2	1	1	4114	2	4	3	1	1	0	0	0	
2	2	1	1	3571.5	1	2	0	1	2	6	0	1	
3	1	1	1	1327.5	14	14	0	0	7	5	2	4	
3	2	1	2	1811	1	0	7	4	1	2	0	0	
1	1	0	2	1852	1	1	1	0	3	3	1	2	
3	1	1	1	1327.5	2	3	0	0	1	2	0	0	
2	2	1	1	2064.33333	15	17	0	0	0	0	2	2	
2	2	1	2	3057.5	1	1	5	4	0	3	1	0	
1	1	1	2	1852	1	0	5	4	5	5	2	1	
1	1	0	1	3384.5	10	6	11	11	7	7	1	2	
3	1	0	2	1506	0	0	0	0	16	16	0	0	
1	1	0	2	1852	11	14	4	4	4	5	3	2	
2	2	1	1	890	3	2	2	2	2	2	0	1	
3	2	1	2	1811	4	4	2	3	1	3	0	0	
3	1	0	2	1506	6	6	8	7	51	48	1	1	

After process, column N is your model's output (column N here is fake data):

A	B	C	D	E	F	G	H	I	J	K	L	M	N
device_ang	distance_t	AM_or_PM	mall_or_st	average_pe	video_walk	video_walk	predict_wa	predict_wa	wifi_walki	wifi_walko	sales_in_n	sales_in_next_15_to_3	
3	1	0	1	1327.5	13	11	0	0	9	5	2	2	10
2	2	0	1	1073.25	4	7	32	33	0	3	1	3	5
3	1	1	1	1327.5	49	48	45	39	15	19	6	3	37
3	1	1	1	2299.5	17	22	21	21	7	5	1	0	17
2	2	1	1	4114	2	4	3	1	1	0	0	0	3
2	2	1	1	3571.5	1	2	0	1	2	6	0	1	1
3	1	1	1	1327.5	14	14	0	0	7	5	2	4	14
3	2	1	2	1811	1	0	7	4	1	2	0	0	2
1	1	0	2	1852	1	1	1	0	3	3	1	2	1
3	1	1	1	1327.5	2	3	0	0	1	2	0	0	2
2	2	1	1	2064.33333	15	17	0	0	0	0	2	2	17
2	2	1	2	3057.5	1	1	5	4	0	3	1	0	2
1	1	1	2	1852	1	0	5	4	5	5	2	1	5
1	1	0	1	3384.5	10	6	11	11	7	7	1	2	11
3	1	0	2	1506	0	0	0	0	16	16	0	0	1
1	1	0	2	1852	11	14	4	4	4	5	3	2	11
2	2	1	1	890	3	2	2	2	2	2	0	1	5
3	2	1	2	1811	4	4	2	3	1	3	0	0	4
3	1	0	2	1506	6	6	8	7	51	48	1	1	8

3) Answers to the following questions:

1. How did you preprocess the data and why?
2. How did you evaluate the model and why?
3. What are your model's pros and cons?
4. What's the other possible information may affect walk-in?

4) Any other code your wrote to complete the assignment. (Not required)

Just like the scratch paper you used when taking exams. You can put any code you wrote but make sure:

1. The code is relevant to this assignment.
2. The code readable and understandable.
3. The code has necessary comments.
4. Only one file is allowed for this part.