

Data Preprocessing

First we use psycopg2 to connect to the database, and we select several attributes from our database for training the model. We select week in year, age group, client gender, case acquisition info, rain_mm, snow_cm, status ph, isresolved, is fatal as our attributes. Then we transform these data to dataframe.

We plot these data using plot function, some data is implemented as line and some data as bar. There are some Null data in rain_mm and snow_mm, we usedfillna to change these null data to 0.0. After that we found that is resolved data is imbalance, so we use no frauds data to match those fraud data. Then is resolved is chosen as label, so it's dropped from database.

"age_group", "client_gender", "case_acquisitioninfo", "status ph" are categorical data. So we use get_dummies function to do one-hot-encoding. Then we divide 80% data to training set and 20% to test set. Then we train the model and use it to test. Finally output the recall, precision and accuracy.