In [1]: import os os.chdir('C:\\Users\\LENOVO\\Downloads') In [2]: import pandas as pd #Load the data file\_path = 'null.csv' data = pd.read\_csv(file\_path) #Display the first few rows of the dataframe data.head() Out[2]: Month\_Year Area Type Borough\_SNT Area name Area code Offence Group Offence Subgroup Measure Financial Year FY\_FYIndex Count Arson and Criminal Damage fy23-24 07 **0** 2023-04-01 Borough Aviation Security(SO18) Aviation Security(SO18) SO18 Criminal Damage Offences fy23-24 3 **1** 2023-04-01 Borough Aviation Security(SO18) Aviation Security(SO18) SO18 Offences **Drug Offences** Drug Trafficking fy23-24 fy23-24\_07 1 Borough Aviation Security(SO18) Aviation Security(SO18) Possession of Drugs **2** 2023-04-01 SO18 **Drug Offences** Offences fy23-24\_07 fy23-24 Borough Aviation Security(SO18) Aviation Security(SO18) SO18 **Drug Offences** Possession of Drugs Outcomes **3** 2023-04-01 fy23-24 fy23-24\_07 3 Borough Aviation Security(SO18) Aviation Security(SO18) **4** 2023-04-01 SO18 Miscellaneous Crimes Against Society Making, Supplying or Possessing Articles for u... fy23-24\_07 2 fy23-24 In [3]: #Convert 'Month\_Year' to datetime format data['Month\_Year'] = pd.to\_datetime(data['Month\_Year']) #Aggregate the data by 'Month\_Year', summing up the 'Count' column time\_series\_data = data.groupby('Month\_Year')['Count'].sum().reset\_index() #Sort the data based on 'Month\_Year' time\_series\_data = time\_series\_data.sort\_values('Month\_Year') #Display the aggregated time series time\_series\_data.head() Month\_Year Count Out[3]: 0 2023-04-01 169034 **1** 2023-05-01 185152 2 2023-06-01 191128 **3** 2023-07-01 188698 In [4]: #Check the range of dates to confirm if any are missing date\_range = pd.date\_range(start=time\_series\_data['Month\_Year'].min(), end=time series data['Month Year'].max(), freq='MS') # 'MS' stands for Month Start frequency #Identify if there are any missing months in the data missing\_dates = date\_range.difference(time\_series\_data['Month\_Year']) missing\_dates DatetimeIndex([], dtype='datetime64[ns]', freq='MS') Out[4]: In [5]: from sklearn.model\_selection import train\_test\_split from sklearn.preprocessing import MinMaxScaler #Splitting the data: 75% for training and 25% for testing train\_data, test\_data = train\_test\_split(time\_series\_data, test\_size=0.25, shuffle=False) #Normalizing the data scaler = MinMaxScaler(feature\_range=(0, 1)) train\_scaled = scaler.fit\_transform(train\_data['Count'].values.reshape(-1, 1)) test\_scaled = scaler.transform(test\_data['Count'].values.reshape(-1, 1)) # Reshaping the data to fit the RNN input requirements # RNNs require input shape of the form [samples, time steps, features] # Here, each sample is one month, and we have one feature - the count #We have only one feature, so we reshape the data to [samples, time steps=1, features=1] train\_scaled = train\_scaled.reshape((train\_scaled.shape[0], 1, train\_scaled.shape[1])) test\_scaled = test\_scaled.reshape((test\_scaled.shape[0], 1, test\_scaled.shape[1])) #Check the shapes of the processed data (train\_scaled.shape, test\_scaled.shape) ((3, 1, 1), (1, 1, 1))from tensorflow.keras.models import Sequential from tensorflow.keras.layers import LSTM, Dense from tensorflow.keras.callbacks import EarlyStopping #Define the LSTM model model = Sequential() model.add(LSTM(units=8, input\_shape=(train\_scaled.shape[1], train\_scaled.shape[2]))) model.add(Dense(1)) model.compile(optimizer='adam', loss='mean\_squared\_error', metrics=['mae']) #Early stopping callback to prevent overfitting early\_stopping = EarlyStopping(monitor='val\_loss', patience=10, restore\_best\_weights=True) #Train the model with validation split history = model.fit( train\_scaled, train\_scaled[:, :, 0], # The target is the same as the input in this case epochs=200, batch\_size=1, verbose=1, validation\_split=0.2, # Use part of the training data for validation callbacks=[early\_stopping] #Evaluate the model on training data train\_mae = model.evaluate(train\_scaled, train\_scaled[:, :, 0], verbose=0)[1] #Evaluate the model on test data test\_mae = model.evaluate(test\_scaled, test\_scaled[:, :, 0], verbose=0)[1] print(f'Train MAE: {train\_mae}') print(f'Test MAE: {test\_mae}') Epoch 1/200 Epoch 2/200 Epoch 3/200 Epoch 4/200 Epoch 5/200 Epoch 6/200 Epoch 7/200 Epoch 8/200 Epoch 9/200 Epoch 10/200 Epoch 11/200 Epoch 12/200 Epoch 13/200 Epoch 14/200 Epoch 15/200 Epoch 16/200 Epoch 17/200 Epoch 18/200 Epoch 19/200 Epoch 20/200 Epoch 21/200 Epoch 22/200 ===] - Os 40ms/step - loss: 0.1346 - mae: 0.2865 - val\_loss: 0.5289 - val\_mae: 0.7272 2/2 [==== Epoch 23/200 Epoch 25/200 Epoch 26/200 Epoch 27/200 Epoch 28/200 Epoch 29/200 Epoch 30/200 Epoch 31/200 Epoch 32/200 Epoch 33/200 Epoch 36/200 Epoch 37/200 Epoch 38/200 Epoch 39/200 Epoch 40/200 Epoch 41/200 Epoch 42/200 Epoch 43/200 Epoch 44/200 Epoch 45/200 Epoch 46/200 Epoch 47/200 Epoch 48/200 Epoch 49/200 Epoch 50/200 Epoch 51/200 Epoch 52/200 Epoch 53/200 Epoch 54/200 Epoch 55/200 Epoch 56/200 Epoch 57/200 Epoch 58/200 Epoch 59/200 2/2 [====== - Os 33ms/step - loss: 0.0709 - mae: 0.2451 - val\_loss: 0.2849 - val\_mae: 0.5337 Epoch 60/200 Epoch 61/200 Epoch 62/200 Epoch 63/200 Epoch 64/200 Epoch 65/200 Epoch 68/200 Epoch 69/200 Epoch 70/200 Epoch 71/200 Epoch 72/200 Epoch 73/200 Epoch 74/200 Epoch 75/200 Epoch 76/200 Epoch 77/200 Epoch 78/200 Epoch 79/200 Epoch 80/200 Epoch 81/200 Epoch 82/200 Epoch 83/200 Epoch 84/200 Epoch 85/200 Epoch 86/200 Epoch 87/200 Epoch 88/200 Epoch 89/200 Epoch 90/200 Epoch 91/200 Epoch 92/200 Epoch 93/200 Epoch 94/200 Epoch 95/200 Epoch 96/200 Epoch 97/200 Epoch 100/200 Epoch 101/200 Epoch 102/200 Epoch 103/200 Epoch 104/200 Epoch 105/200 Epoch 106/200 Epoch 107/200 Epoch 108/200 Epoch 109/200 Epoch 110/200 Epoch 111/200 Epoch 112/200 Epoch 113/200 Epoch 114/200 Epoch 115/200 Epoch 116/200 Epoch 117/200 Epoch 118/200 Epoch 119/200 Epoch 120/200 Epoch 121/200 Epoch 122/200 Epoch 123/200 Epoch 124/200 Epoch 125/200 Epoch 126/200 Epoch 127/200 Epoch 128/200 Epoch 129/200 Epoch 130/200 Epoch 132/200 Epoch 133/200 Epoch 134/200 Epoch 135/200 Epoch 136/200 Epoch 137/200 Epoch 138/200 Epoch 139/200 Epoch 140/200 Epoch 141/200 Epoch 142/200 Epoch 143/200 Epoch 144/200 Epoch 145/200 Epoch 146/200 Epoch 147/200 Epoch 148/200 Epoch 149/200 Epoch 150/200 Epoch 151/200 Epoch 152/200 Epoch 153/200 Epoch 154/200 Epoch 155/200 Epoch 156/200 Epoch 157/200 Epoch 158/200 Epoch 159/200 Epoch 160/200 Epoch 161/200 Epoch 162/200 Epoch 164/200 Epoch 165/200 Epoch 166/200 Epoch 167/200 Epoch 168/200 Epoch 170/200 Epoch 171/200 Epoch 172/200 Epoch 173/200 Epoch 174/200 Epoch 175/200 Epoch 176/200 Epoch 177/200 Epoch 178/200 Epoch 179/200 Epoch 180/200 Epoch 181/200 Epoch 182/200 Epoch 183/200 Epoch 184/200 Epoch 185/200 Epoch 186/200 Epoch 187/200 Epoch 188/200 Epoch 189/200 Epoch 190/200 Epoch 191/200 Epoch 192/200 Epoch 193/200 Epoch 194/200 Epoch 196/200 Epoch 197/200 Epoch 198/200 Epoch 199/200 Train MAE: 0.12705300748348236 Test MAE: 0.12359803915023804 In [7]: #Summary in form of a table summary results = { 'Data Split': ['Training', 'Test'], 'MAE': [train\_mae, test\_mae] results\_table = pd.DataFrame(summary\_results) #Display the results table print(results\_table) Data Split Training 0.127053 Test 0.123598 1 The table displays the Mean Absolute Error (MAE) of the LSTM model on both the training and test data splits. The training data resulted in an MAE of approximately 0.127, while the model achieved a slightly better MAE of approximately 0.124 on the unseen test data, indicating that the model has a consistent performance on both datasets within a close margin. The LSTM model's performance, as indicated by the Mean Absolute Error (MAE), shows that it has learned to predict the monthly count of offenses with a reasonable level of accuracy given the very limited dataset. The MAE of 0.127 for the training set and 0.124 for the test set suggests that the model is not overfitting, as the test error is slightly lower than the training error. This is a positive outcome, especially considering the challenges posed by the small amount of data available. In [ ]: