**Issue**

When was doing for LSTM project, it was for the first I was working with so huge datasets. The issue caused was shape mismatch in training and testing data. During data preprocessing after splitting for training and testing, I was trying to split x and y which where not of same shape. This led to mismatches when training the model.

Lets assume the error was something like this:

**X\_sequences.shape is (14579, 30, 13)**

**y\_sequences.shape is (14579,)**

**Shape of X\_train: (11663, 15)**

**Shape of y\_train: (11663,)**

**Shape of X\_test: (2946, 15)**

**Shape of y\_test: (2946,)**

The issue was with the shape of X\_train and X\_test which might have mainly caused by data splitting.

Solution

I start print all the datasets and started check shape of split datasets including number of rows and column. Afterwards, splitting was done on X\_sequences and y\_sequences, rather than on x and y.

Code example

**X\_sequences, y\_sequences = []**

**for i in range(len(scaled\_data) - sequence\_length):**

**X\_sequences.append(scaled\_data[i: i+sequence\_length, :-1]) y\_sequences.append(scaled\_data[i+sequence\_length, -1])**

**X\_sequences = np.array(X\_sequences)**

**y\_sequences = np.array(y\_sequences)**

**train\_size = int(len(X\_sequences) \* 0.8)**

**X\_train, X\_test = X\_sequences[:train\_size], X\_sequences[train\_size:]**

**y\_train, y\_test = y\_sequences[:train\_size], y\_sequences[train\_size:]**

Let us assume the data splitting looked somewhat like the above code snippet. In order to fix this issue the I would do:

**X\_sequences, y\_sequences = []**

**for i in range(len(scaled\_data) - sequence\_length):**

**X\_sequences.append(scaled\_data[i: i+sequence\_length, :-1]) y\_sequences.append(scaled\_data[i+sequence\_length, -1])**

**X\_sequences = np.array(X\_sequences)**

**y\_sequences = np.array(y\_sequences)**

**X\_train, X\_test, y\_train, y\_test = train\_test\_split(X\_sequences, y\_sequences, test\_size=0.2, random\_state=42)**

The train-test aplit is now correctly applied to X\_sequences and y\_sequences, ensuring that the shapes remain consistent with input requirements. Now X\_train and X\_test will have shapes like (batch\_size, 30,13) which is expect by model, and y\_train and y\_test will be of shape (batch\_size,).

This was the one issue I faced while I was building my model. Although the issue might not seem that complicated for many practitioners, me begin a beginner and just starting out my learning journey it seems quite complicated for me. I learnt how much of an impact data handling and preprocess has in the entire process which many people might not notice about. After this issue, I start checking my dataset training, test and also cross validation data.