1. Imports
2. Define spark session
3. Load the data
   1. Extract the number of columns and rows
   2. Print schema
   3. Locate the date column
4. Make all forms of nulls into PySpark “nulls”
5. Perform any necessary transformations on the label column
   1. Define aliases for index and label
6. Handle dates if any
   1. Convert the strings format to datetime
   2. Extract features, day of week, day of month, quarter, etc
   3. Order by date
7. Check the Null values in each column (time consuming)
   1. Too many, too few, what to do?
8. Generate the list of all string columns (includes all parasitic columns)
9. How many distinct values are there for each categorical feature?
   1. Time consuming
   2. Large numbers are issues for tree methods.
   3. Drop columns with large number of distinct categories
10. Generate the list of all string columns (does not include parasitic columns)
11. Generate the list of all numerical columns
12. Convert the string columns to numerical codes
13. Do one-hot-encoding if necessary
14. Handle Null values
    1. Replace with zero
    2. Replace with median
15. Rename the dataframes to Xy and X
16. Save to feather format for later use
17. Perform any EDA’s if necessary
18. Define the train-validation sets for time series data
19. Vectorize for Spark training
20. Define the ML model
21. Define the evaluation object
22. Define the parameter grid
23. Fit the model
    1. XV
    2. Train validation
24. Compute predictions for the dev set
    1. Handled automatically for XV
    2. Due transformation and evaluation
25. Compute predictions for the test set