FE cheat sheet

Robust ML:

1. Preview the data in order to discard useless features or ones with too many NaN values.
2. Separate features on categorical and numerical
3. Create a pipeline with the appropriate steps to preprocess the numerical features
    a. Imputing
    b. Scaling
    c. Normalizing
    d. More
4. Create a pipeline with the appropriate steps to preprocess the categorical features
    a. Imputing
    b. Encoding
    c. Handle unknown
    d. More
5. Create column transformer combining both pipelines
6. Create a dictionary with the Classifiers/Regressors that are going to be tested
7. Combine each classifier with the column transformer
8. Train and test all together using either cross validation or stratification.

Data generation:

1. Create a copy of the dataset
2. Select the modifying component
    a. A division of the each columns STD is recommended
3. Randomly add of subtract the value row by row
4. Select a random sample of the generated data to add to the training set.

Feature generation:

Approach 1:

1. Decompose a complex column into smaller individual components

Approach 2:

1. Select 2 or more columns with a real life correlation
2. Create a linear combination between the selected columns

The new feature should never include the target and should appear in both than and test.

Images:

1. Iterate through the folders of data
    a. If the class is in the name check for the file name using OS
    b. If the class is in the folder name go folder by folder
2. Create placeholder np.arrays with an extra space for the class
3. Load the image with PIL.Image
4. Convert the image either to RGB or to Black and White ‘L’
5. Resize or modify the data at will
6. Save the images with the target into np.arrays
7. Expand the dataset in each iteration

Text:

1. Train/Test Split
2. Take out words that are too specific
3. Take out words that appear in excess
4. Generate TF-IDF vectors after splitting
5. Use at least N-Grams of size 2 at least to keep some context
6. Classify the data using tree models

Time series:

Classification:

1. Select a sequence length
2. Iterate through the data in blocks of size ‘seq_len’
3. Extract statistical data from the sequence
    a. Mean
    b. Var
    c. STD
    d. MAX
    e. MIN
    f. |MAX - MIN|
    g. Col A / Col B
    h. More

4. Use the new rows as the data to perform the classification task.

Regression:

1. Select a sequence length
2. Iterate through the data in blocks of size ‘seq_len’
3. The target is part of the data
4. Use the sequence to predict one or more features that appear in the next row of data
5. Only include the target row if the train set is extremely limited.