Data Preprocessing

Data Diggers:

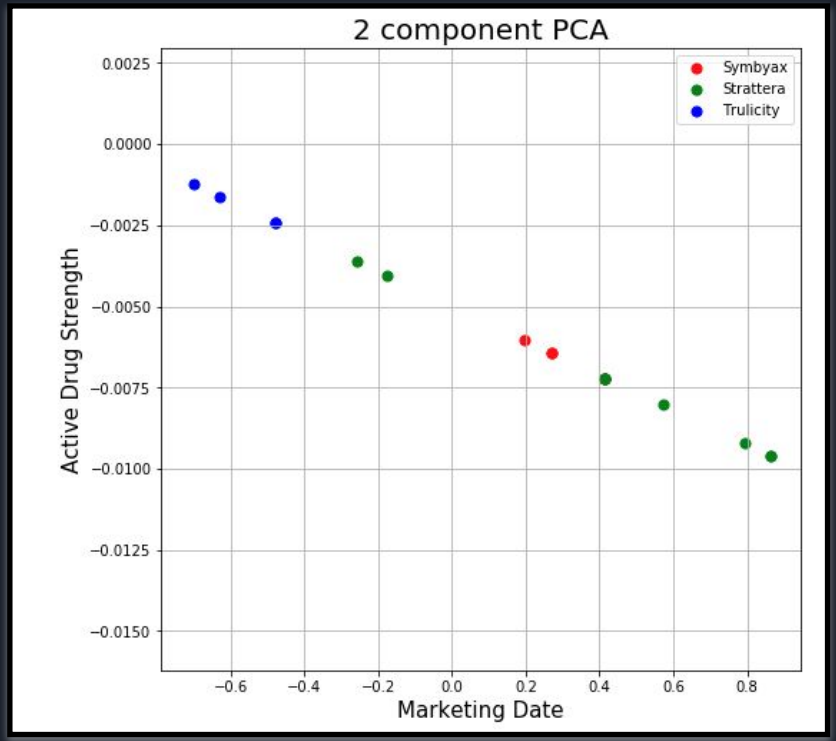
An Lam, Jimmy Le, Tom Amir,
Dianna Melendez, Amrit Singh, Talal Jawaid, Min Li

Overview

- Began by practicing each pre-processing technique
- Data selection:
  - Drugs (FDA products data)
  - London Air (Air quality levels)
- Techniques used:
  - Missing Values
  - Outliers
  - Duplicate Data
  - Shuffling/Sorting/Saving Dataframes
  - Dropping Fields
  - Calculated Fields
  - Feature Normalization
  - Concatenation
  - Principal Component Analysis (PCA)

Interesting findings

- PCA performed on drugs Symbyax, Strattera, and Trulicity
- Performed on active drug strength relative starting marketing date.
- Data suggests that as Standardized (Z-Score) Marketing Date increases, the Strattera was manufactured at a much lower strength
- When compared to Symbyax, Trulicity continued to stay unchanged.
- It was interesting to see a negative correlation in a large set of data



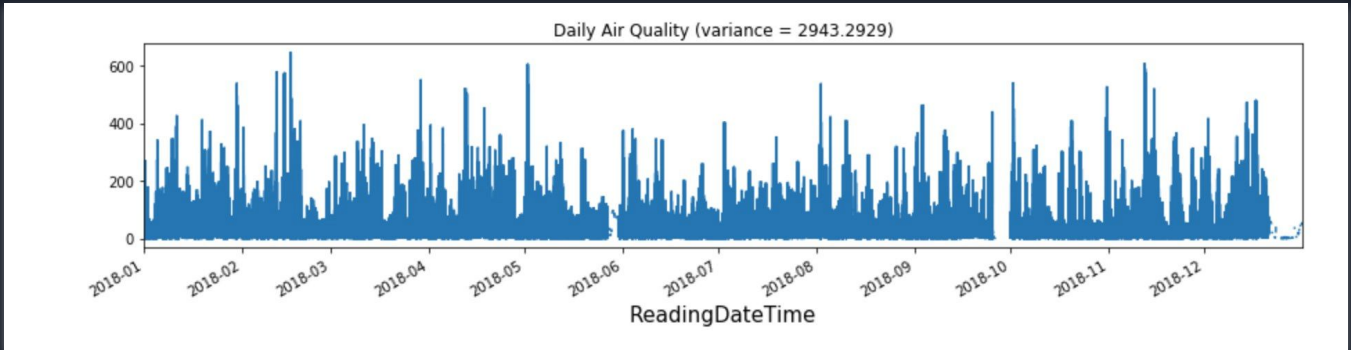
Splitting the Data

- London Air Data
- Training / Test Data Split:
  - 80% Training Data
  - 20% Test/Validation Data
  - Focusing on air quality “value” column for carbon monoxide:
- Training Data:
- Mean = 48 Standard Deviation = 54
- Test Data:
- Mean = 47 Standard Deviation = 54
- What we can draw from this:
  - Vital step in creating a machine learning model
  - The training data receives majority share of the partitions since this is the data that the model is built on
  - Remaining test set to be used to validate the training data against
  - Two sets of randomized data provide consistent results in terms of the mean and standard deviation of the “value” attribute.

Issues

Aggregation on Air Quality with respect to the Date

- Date format in our data set different than tutorial
- Macbook version versus Windows version on running Jupyter Notebook files
- to_datetime() from Pandas wouldn't accept the date column from our data set
    - to_datetime() wouldn't accept our datetime in YYYY-MM-DD HH:MM:SS
- Attempted to use Python split() to strip time from date.
- Attempted to directly modify from Excel unsuccessfully
- Difficult to debug due to no error being thrown by Jupyter
- Successfully resolved issue



Conclusion

- Pre-Processing is a vital, but tedious step in Data Mining
- Needs to be done in a specific way
- Resulted in a clean data set we are confident in using for coming projects
- Our data is clean and we are ready to proceed in the process of Data Mining.