Thoracic Surgery Data Set

Data Wrangling

What kind of cleaning steps did you perform from the original data set?

- The original data set from the UCI machine learning repository is in the form of a Weka ARFF file. So for easy utilization, I converted the file to CSV using a github tool found at https://pulipulichen.github.io/jieba-js/weka/arff2csv/ .
- Analyzing the data set's info shows many columns as object strings for T and F values. These include PRE7, PRE8, PRE9, PRE10, PRE11, PRE17, PRE19, PRE25, PRE30, PRE32, and Risk1Yr. So, I converted the T and F object data types to 1 and 0 int data types in these columns.
- The columns DGN, PRE6, and PRE14 contains data in the form of a string with an int value attached. Reviewing the column data description, I concluded the string value was redundant and it will be more useful for analysis later on just utilizing the int value. So these three columns were adjusted to just have the int value as data type int.
- The id column was removed because it is not necessary and lacking in any useful description of each patient. The index number will suffice.
- The column names were renamed with more human readable words instead of the codes. The corresponding column descriptions found on the UCI machine learning repository site is shown below:
    - DGN: Diagnosis - specific combination of ICD-10 codes for primary and secondary as well multiple tumours if any (DGN3,DGN2,DGN4,DGN6,DGN5,DGN8,DGN1)
    - PRE4: Forced vital capacity - FVC (numeric)
    - PRE5: Volume that has been exhaled at the end of the first second of forced expiration - FEV1 (numeric)
    - PRE6: Performance status - Zubrod scale (PRZ2,PRZ1,PRZ0)
    - PRE7: Pain before surgery (T,F)
    - PRE8: Haemoptysis before surgery (T,F)
    - PRE9: Dyspnoea before surgery (T,F)
    - PRE10: Cough before surgery (T,F)
    - PRE11: Weakness before surgery (T,F)
    - PRE14: T in clinical TNM - size of the original tumour, from OC11 (smallest) to OC14 (largest) (OC11,OC14,OC12,OC13)
    - PRE17: Type 2 DM - diabetes mellitus (T,F)
    - PRE19: MI up to 6 months (T,F)
    - PRE25: PAD - peripheral arterial diseases (T,F)
    - PRE30: Smoking (T,F)
    - PRE32: Asthma (T,F)
    - AGE: Age at surgery (numeric)
    - Risk1Y: 1 year survival period - (T)rue value if died (T,F)

How did you deal with missing values, if any?

- There are no missing values in the original data set.

Were there outliers, and how did you handle them?

- The only numeric columns that could contain outliers are PRE4, PRE5, and AGE.
- Data analysis with box plots and scatter plots reveal 16 outliers.
    - PRE4: 0 outliers
    - PRE5: 15 outliers.
        - Most of the data is below 8 FEV1, except for 14 data points above 40 and 1 data point being just beyond 8.
    - AGE: 1 outlier.
        - Majority data ranges from 40 to 80 years old. The one outlier is in the early 20 range.
- Considering the data set contains a total of 470 instances, I decided to omit the 16 outliers to reduce the data set to 454 afterwards.
- This quantity seems sufficient compared to the original data set size without omitting too much of the original data.

The above procedure can be found in the **Data_Wrangling** Jupyter notebook file.

https://github.com/sychi77/CapstoneProject1/blob/master/Data_Wrangling.ipynb