Original Analysis Case Study

Case Study: Analyze data to determine factors that increase survival of breast cancer patients

Narrative:
With data science so prevalent during the ongoing pandemic, I wanted to dive into something in the health field. Unfortunately, most of the health data I found was aggregated more than I wanted. I realize a lot of this is done for protection of individuals, but it's difficult to perform analysis, without knowing all the factors that went into the outcome. I was looking for something that addresses individuals of varying demographics.

I found data from a study on breast cancer patients, originally from the Dutch Cancer Institute (NKI), which had since been cleaned and readied by Devi Ramanan. This dataset contained treatment details on individuals, along with a death indicator. Using this data, I'd like to determine if some treatments were more effective than others. I'd also like to consider the impact on various sizes and types of tumors.

Data Source: https://data.world/deviramanan2016/nki-breast-cancer-data
Data File: NKI_Breat_Cancer_Data.csv

Data Summary:
272 breast cancer patients
Meta data includes patient info, treatment, and survival.

Features:
age = Age in years (numerical)
eventdeath = Death indicator, 0 = No, 1 = Yes (categorical)
chemo = Chemotherapy indicator, 0 = No, 1 = Yes (categorical)
hormonal = Hormonal Therapy indicator, 0 = No, 1 = Yes (categorical)
amputation = Amputation indicator, 0 = No, 1 = Yes (categorical)
diam = Diameter of the primary tumor in mm (numerical)
posnodes = Number of positive lymph nodes (numerical)
grade: 1=Well Differentiated, 2=Intermediate, 3=Poorly Differentiated (categorical)
angioinv: Angioinvasion - Extent to which the cancer has invaded blood vessels or lymph vessels (categorical)
lymphinfil: Level of lymphocytic infiltration (categorical)
histtype: Histological type, Good, Average, Poor (categorical)
survival: Survival time (numerical)
timerecurrence: Recurrence time (numerical)

Part I – Graph Analysis
Step-by-step instructions for Graph Analysis:

1. Load data into a pandas dataframe.
2. Review data records to determine data types and volume of data.
3. Select features that might help determine factors that attribute to survival. Note which features are categorical and which are numerical.
    a. Think about some questions that might help you predict survival
4. Review summary statistics. Note possible outliers. What do these statistics tell you?
5. Plot histograms to review frequency distribution of numerical features. Note any outliers.

6. Create bar charts to review distribution of categorical features.
7. To see if the data is correlated, use Pearson Ranking against the numerical features with Yellowbrick's Rank2D.
8. Use Yellowbrick's Parallel Coordinates to visualize and compare the distributions of numerical variables between patients that died and those that survived.
9. Use Stack Bar Charts to compare patients who died to patients who survived based on the other categorical variables.

Part II – Dimensionality and Feature Reduction

Step-by-step instructions for Dimensionality and Feature Reduction:

10. Eliminate features.
    a. I will remove Patient, ID, and barcode, since they are irrelevant.
    b. Survival time and recurrence time are not needed, since they are results, not contributing factors.
11. Find features with missing values.
    a. Replace missing values with median or mode.
12. Adjust non-normal distributions for numerical features.
    a. Review Cumulative Distribution Function (CDF) plots to identify non-normal distributions that can be adjusted. For example, exponential, lognormal, and Pareto distributions.
    b. Perform transformations as needed.
13. Convert categorical data to numbers.
    13.1 Reduce features using Principal Components Analysis (PCA)

Part III – Model Evaluation and Selection

Step-by-step instructions for Model Evaluation and Selection:

14. Prepare data for model
    a. Standardize Data to Avoid Convergence
    b. Separate target from feature dataset
    c. Split data Training and Testing datasets
15. Run a Logistic Regression Model to predict if a passenger has survived or not
16. Evaluate the model
    a. Use Confusion Matrix to evaluate the model
    b. Use Precision, Recall & F1 score to evaluate the model
    c. Use the ROC curve and Area Under the Curve (AUC) to evaluate the model

Part IV – Improvements

a. Added to Part 2: Reduce features using Principal Components Analysis (PCA)

Part V – Self-Improvements

- Try using several machine learning techniques to try and make a prediction
- Create a pipeline and use stratified K-fold cross validation with a K of 10. Re-evaluate.

Additional References:

https://www.scirp.org/journal/paperinformation.aspx?paperid=84902

https://www.researchgate.net/publication/335211007_Prognostic_value_of_microvessel_density_in_stage_II_and_III_colon_cancer_patients_a_retrospective_cohort_study

https://www.researchgate.net/publication/340326136_Histopathologic_Assessment_of_Capsular_Invasion_in_Follicular_Thyroid_Neoplasms-an_Observer_Variation_Study