**Predictions of Five-Year Survival of Breast Cancer**

Jolene Branch

Data Science Program, Bellevue University

DSC 550: Data Mining

Prof. Becky Deitenbeck

**Abstract**

90% of women diagnosed with breast cancer in the United States survive more than five years after diagnosis. (Dollinger & Dubrow). I chose to explore the individual characteristics of women who make the five-year survival mark versus those who do not. The goal was to accomplish this through a binary classification system. The K-means classifier was used for classifying five-year survival. This turned out to be unsuccessful, likely due to overfitting of data, and was confirmed by a published Codeacademy project. (Grijaiva, 2018). A late-realized confounding variable related to pursuit and type of treatment is understood to completely invalidate this work. Lack of domain knowledge and use of the complex National Cancer Institute SEER Database contributed to an abundance of data. A new question was gained from this project; “Are there financial or regulatory disincentives to providers for listing breast cancer as cause of death?”

*Keywords:* breast cancer, five-year survival, K-means classifier

**Predictions of Five-Year Survival of Breast Cancer**

The diagnosis of breast cancer is dreaded by all, and most adults know someone who has had breast cancer. Although 91% of women diagnosed with breast cancer in the United States survive more than five years after diagnosis and 84% survive more than ten years, it is still met with fear, which extends to immediate family members who have concern for their own health. (Vrinten et al., 2018). With similar family background, I chose to explore the individual characteristics of women who make the five-year survival mark versus those who do not. The goal was to accomplish this through a binary classification system.

I am interested in this topic due to increasing number of women Veterans utilizing the services of the local VA Medical Center, my mother-in-law’s unsuccessful bout with cancer the second time around, and the acknowledgement or personal risk factors – which have been exacerbated by too much sedentary time during the ‘stay-at-home’ recommendations prompted by COVID-19 infection and death rates. The question I tried to answer was “Is Stage at diagnosis a predictor of breast cancer survival?”

**Methods**

***Data Selection***

Since many data sets and completed projects related to breast cancer were available on the internet (such as Kaggle, University of Wisconsin, seaborn, UCI Machine Learning Repository, and Haberman’s Survival Data), I was worried that no matter what I did, that it could be suspected of copying one of them (see, e.g., Brownlee, 2020; Goel, 2018; Vrinten et al., 2017). I decided to go to the highest source form of the data available and requested permission to use the National Cancer Institute Surveillance, Epidemiology, and End Results Program (SEER). Also influencing this decision was the thought that if I used the entire population, that I would not need to run my model twice – as in splitting the data into test and train data sets. My plan was to look for the more highly correlated features, then use those in a classification model and compare against true five-year survival rates.

***Importing the Data***

I experienced early challenges in gaining access, understanding, selecting, and retrieving the data from SEER. After logging in the first time, I browsed through one web page link after another until I finally found that I needed to download software onto my desktop. After doing that, I was at the point of trying to figure out exactly what data I need. I mistakenly thought that the term ‘public use database’ carried with it some implication of ease of use. That was not to be the case.

I discovered that even though the site contained a data dictionary, that abbreviations and medical coding terminology even within the data dictionary required the user to have some level of domain knowledge simply to navigate the site and retrieve data. There are many decisions required to be made ‘up front’ when requesting SEER data, and there is no going back and changing answers without starting the data request over. Even seemingly simple variables all had numerous selections cascading below them, making the risk of selecting redundant features high. After exporting the data, saving to a .csv file, and then reading the data into Jupyter Notebook into a pandas dataframe, there were over seven million rows and 66 columns. The downloaded datatypes were mostly ‘objects.’ The dataset did not include any genetic factors, as I mistakenly thought that the values in the ‘Breast Subtype (2010+)’ column were the gene values.

***Data Preparation***

The first modification to the dataset was to remove all male records. This resulted in removal of 3.5 million rows! I removed all rows in which the value of the ‘Site rec ICD-0-3/WHO 2008 (individual sites only’ column was *not* ‘breast.’ This left me with 1,064,157 rows, which still seemed like way more than I needed. I decided to pull a subset based on year of diagnosis and selected 2010 because five-year survival data should be available by now for any cancer initially diagnosed in 2010. At this point I made my first copy, thinking I had already done *so* much work that I would not want to have to redo. This took several minutes to run due to the size.

I removed spaces from some of the rows and shortened the names of others. The ones I was not too certain of, I left alone. I searched for features/columns with no data or have all values and same – and deleted them. I used value\_counts() to do this. That is how I found out that all values in the ‘Behavior code ICD-0-3’ feature returned as ‘malignant.’ If all the values in the column are the same, it gives no information and is not necessary. Several other columns were removed with this rationale. It occurred to me that since I had removed all males from the dataframe, that I did not need the ‘Sex’ column. The same rationale prompted the removal of the ‘Year of diagnosis’ column. A search for duplicate rows found several repeated Patient IDs. In these cases, only the first entry (for 2010) was kept for each of these women. This was accomplished using the drop\_duplicates() method with the argument of keep=’first’. My suspicion is that these are women who perhaps were initially only diagnosed with cancer in one breast, then later in 2010 received an additional location and/or diagnosis. Cancer has no respect for laterality.

Converting multiple columns into numeric format was harder than is seems like it should have been. Several columns, such as age, income, and rurality, were already ‘binned,’ and no two could be converted same way. The greatest amount of time was spent on the ‘Survival months’ column, which was an object data type with leading zeros for every value. The column needed to be changed to a series in order to use strip(), and conversion back to a dataframe column was complicated by an error statement “Unable to parse string “Unknown” at position,” which was curious to find after having already using df.isnull().sum() to check the entire dataframe. Adding the argument “errors=’coerce’ to the pd.to\_numeric() method when converting the series back to a dataframe column ultimately worked through this obstacle. Calling a Boolean for months survived greater than 60 or not proved challenging due to datatype mismatches within the conditional statement. The number 60 did not need ‘’s around it, but the numbers ‘1’ and ‘0’ did.

After attempting to use sklearn’s ‘OneHotEncoder’ (for the first time) instead of ordinal encoder to get the binned ages into a manageable numeric form, I resorted to changing them to a series and stripping the “years,” then putting them back into the dataframe and using str.slice(0,2) to strip everything from the bin values except the first two characters. That left me with 15, 20, 25… which was easily converted to numeric.

The ‘Grade’ column, which is equivalent to the Stage of cancer upon diagnosis, was manipulated using the get\_dummies() method after I renamed all values, so that each cancer grade for each event would have a binary value, which is need for the type of machine learning I was trying to accomplish. Fortunately, I did not delete the original column, as I would use that later for correlation calculations. That was the easiest column to convert to numbers, and I used the renaming technique on values in subsequent columns prior to converting them.

Throughout this data preparation process, I deleted columns as I learned more about them, their values, and their relationships with one another. I recalled reading in a text book that it is generally a good idea to keep and use ‘recode’ columns, but in the case of this dataset they presented incredible redundancy of data, so I removed almost all of them. At many times in this process, I found the output of list(df.columns) to be helpful, as it listed all the columns with apostrophes. This made it easier to copy/paste when dealing with several consecutive column names in list format within the code. The dataframe was reordered so the columns of most interest were on the left, and ‘Five-Year-Survival’ was set to be the target column.

I felt that the ‘Rural-Urban Continuum Code’ feature should be converted to a binary response, and that the additional differentiation of *size* of the urban area was not important. You either have geographic access to larger health care facilities or you do not. Another consideration for this urban vs. rural binary split is the aspect of environmental toxins. That, however, can not be split as neatly. For example, a woman who lives in a rural area but commutes to a large city would potentially have exposure to both rural and urban toxins.

When I got these binary responses into 1’s and 0’s, I started to check for correlation among select features and the target variable (of five-year-survival).

**Insights**

***Correlations***

The first insight I gained from the correlations is that the stage/grade of cancer upon diagnosis has little correlation with five-year survival. That intuitively made no sense, as cancer progresses from lower to higher stages over time and if treatment is ineffective. This data aberration did give me a hint that there was probably no sense in creating a five-year survival classifier based on the grade at diagnosis.

Next I turned to age at diagnosis, because if age was not accounted for, anyone would expect to see that a much younger person at time of diagnosis would live significantly longer after diagnosis than an elderly person, simply due to life span averages.

***Visualizations***

When grouping respondents by the grade at diagnosis, the visualization was stark; Hardly any women were diagnosed with Stage 4 breast cancer. The bar for Stage 4 in that plot is barely visible. I find this curious. The count of women diagnosed with stage 4 cancer seems like an outlier. This is where domain knowledge is important, as I don’t know the diagnostic indicators and differences between stage 3 and 4. The best I can do, without that information, is to speculate that perhaps oncologists do not want to give an initial stage 4 diagnosis for fear that the woman will be so distraught at hearing that news that she might not even attempt treatment.

When I looked at median months survived when grouped by Stage/Grade of cancer upon diagnosis, I found them to be nearly identical. The range was 84 to 87 months. The median, however, looks more intuitive. The higher the Stage/Grade of cancer, the less time you would expect to live.

***Limitations***

I’m sure any serious researcher would have examined this dataset with and without the presence of women who died with cause of death listed as something other than breast cancer. I believe this mean is essentially meaningless for two reasons; 1. It does not distinguish between women who pursued treatment and those who did not (and even which type of treatment, and 2. It does not distinguish between survival time that ended due to breast cancer or due to some other diagnosis.

At best, the mean life expectancy based on stage/grade of breast cancer from this dataset could be used as a general understanding of what is currently occurring in the field of breast cancer treatment. It cannot be used for prediction or even responsibly used for classification of outcome (in terms of five-year-survival) due to the missing feature of medical interventions (or treatments).

***Classification Modeling***

Using all the features that I was able to get into ordinal and numeric data types, I created a much smaller dataframe and performed K-means clustering. The ‘elbow’ was clear. The model hinged at two clusters. At this point I decided to look externally for validation of my model, since I'm pretty sure mine was incorrect. I believe I have a serious overfitting problem from too many features and noise within the dataset. There must be enough differentiation among the grades to support more than two clusters!! This was a big clue that there is something wrong with my model. I knew going in that there are four ‘stages’ or grades of cancer upon diagnosis. The medical community does not just pick a number. There are standards. I do not think there is a way to responsibly classify survivability based on initial presentation. Quality, method, and timeliness of treatments need to be included. A more reasonable classification project (if not including treatments) would be to use 'malignant' and 'benign' as the initial binary classifier target variables, because that decision is made before and regardless of treatment decisions.

Since I realized by now that the model I created was not valid, I saw no reason to text it further against itself. Instead, I evaluated it using an external data source. The Codecademy K-Nearest Neighbor breast cancer project gave a direct hint at where my model went wrong. I had over trained the model with lots of data, such that it probably started learning from the noise and inaccurate data entries. The model was not “able to categorize the data correctly because there’s too many details and too much noise…” (Grijaiva, 2018). It also suggested that trying to build a linear model with nonlinear data can cause this problem.

**Lessons Learned**

There were many lessons learned during the course of this project, but they are unlikely to be useful to the breast cancer research community. Some may help beginner Data Science students, though.

The target variable of five-year-survival was too far into the process. If I wanted to use that target variable, I needed to filter the data set by ‘treatment’ or ‘no treatment’ to start with, then by type of treatment. I could have used the distinction of ‘malignant’ or ‘benign’ as the target variable, because the treatment (or lack of treatment) effects would not have time to come into play.

I learned the importance of trying dimensionality reduction *first* with such a massive data set. (Sharma, 2020). This can lead to compression of features, avoidance of underfitting and noise, and increase efficiency and performance.

**Conclusions**

Even if treatment decisions are factored in, five-year survival is not a good indicator of differentiation of prognosis or prediction anymore, since nearly every woman survives at least five years after initial diagnosis.

The initial features of a person probably do not matter as much as the treatment applied. The same might be true for initial Stage of cancer.

There is no way to responsibly classify survivability based on initial presentation. Quality, method, and timeliness of treatments need to be included.

Even though this project was not successful in determining individual characteristics that can predict breast cancer survival or even that Stage/Grade can reliably make this prediction, it is important that the medical community incorporate as much machine learning in it as possible. Being human, we have (subconscious) cognitive biases that can unduly influence our ability to make accurate diagnoses, and another ‘check’ in the system is needed. (Lamb, et al., 2020).

**References**

Brownlee, J. (2020, January 14). How to Develop a Probabilistic Model of Breast Cancer Patient Survival. Retrieved from <https://machinelearningmastery.com/how-to-develop-a-probabilistic-model-of-breast-cancer-patient-survival/>

Dollinger, M., & Dubrow, B. (n.d.). Living with Mortality: Life Goes On. Retrieved from <https://med.stanford.edu/survivingcancer/cancers-existential-questions/cancer-living-mortality-fear.html>

Goel, V. (2018, October 12). Building a Simple Machine Learning Model on Breast Cancer Data. Retrieved from <https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3>

Grijaiva, C. (Director). (2018, October 23). *Machine Learning Fundamentals K Nearest Neighbor: Breast Cancer Classifier* [Video file]. Retrieved from <https://www.youtube.com/watch?v=DA0SPkZW0jA&list=PLf535ysXr-OyfLYN2Q3Hh45as8visPT0_&index=22&t=0s>

Lee, C. S., Rosvold, E., & Lamb, L. (2020, August 05). Missed breast cancer: Effects of subconscious bias and lesion characteristics. Retrieved from <https://psnet.ahrq.gov/issue/missed-breast-cancer-effects-subconscious-bias-and-lesion-characteristics>

Vrinten, C., McGregor, L., Heinrich, M., Von Wagner, C., Waller, J., Wardle, J., & Black, G. (2017, August). What do people fear about cancer? A systematic review and meta-synthesis of cancer fears in the general population. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5573953/>