Data 630: Machine Learning Summer 2023

Assignment # 2 – Statistical Data Mining

Simon A Hochmuth

Email: shochmuth3@student.umgc.edu

Professor: Ami Gates

Introduction:

Objective:

The analysis conducted on the prostate cancer dataset had the goal of conducting statistical data mining to figure out if the cancer could be predicted. Logistic Regression was used to try to predict and understand the underlying data. Using this algorithm with the dataset, the goal was to pull out relationships between variables to build predictive models for prostate cancer. Once the model is made, the overall objective is to have a model with an acceptable amount of accuracy that can predict whether a person has prostate cancer, and in turn be used to help save lives of people affected by this cancer.

Problem Domain:

Everyone has had an experience with a friend or family member fighting with cancer and has heard the cliché comment if only we had caught it sooner. Prostate cancer is the fourth most common cancer, and "Over 288000 men are diagnosed with prostate cancer, and 34000 die due to it every year" (PCF. 2022). With so many instances and a growing human population, the importance of understanding and predicting whether a person has prostate cancer has become more important than ever. The goal of this project is to understand if the relationships between the dataset's variables can be used to predict whether a person has prostate cancer with an acceptable amount of error.

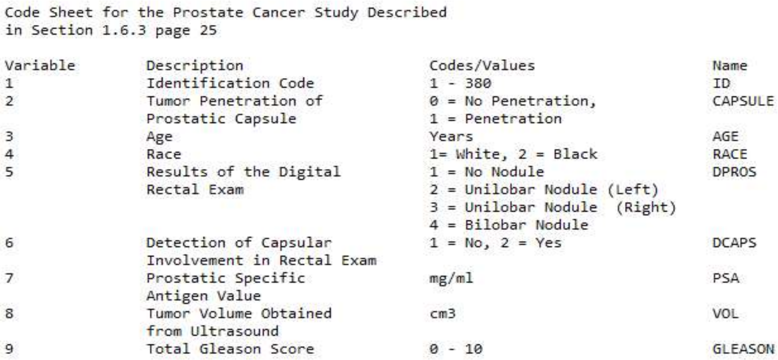
Method Rationale

After the data is preprocessed, the goal is to use it to build a Logistic Regression model that predicts whether a person has prostate cancer. To complete this, the data needs to be split into a test and training set. The reason the split occurs is to train on one dataset, then test the model on data it has not seen before, thus providing a reliable summary of the model's performance. Once the model is tested, a statistical analysis and model evaluation will be run on the model to figure out how successful the Logistic Regression model was in predicting prostate cancer.

Analysis:

Data:

The data was provided by the Data 630 UMGC team, where the description can be seen in figure 1 below. The data contains 9 variables including ID, tumor penetration, age, race, results of the rectal exam, detection of capsular, antigen values, tumor volume, and total Gleason score. Age and race were specific to the person's biological traits, while the rest of the variables were focused on lab findings during the prostate examination. There was a total of 380 observations made within this dataset between white and black people over the age of 50.



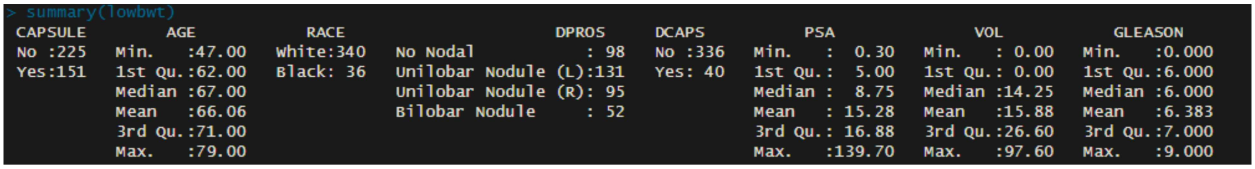
Code Sheet for the Prostate Cancer Study Described
in Section 1.6.3 page 25

Variable	Description	Codes/Values	Name
1	Identification Code	1 - 380	ID
2	Tumor Penetration of Prostatic Capsule	0 = No Penetration, 1 = Penetration	CAPSULE
3	Age	Years	AGE
4	Race	1= White, 2 = Black	RACE
5	Results of the Digital Rectal Exam	1 = No Nodule 2 = Unilobar Nodule (Left) 3 = Unilobar Nodule (Right) 4 = Bilobar Nodule	DPROS
6	Detection of Capsular Involvement in Rectal Exam	1 = No, 2 = Yes	DCAPS
7	Prostatic Specific Antigen Value	mg/ml	PSA
8	Tumor Volume Obtained from Ultrasound	cm3	VOL
9	Total Gleason Score	0 - 10	GLEASON

Figure 1: Text description provided by UMGC for the prostate cancer dataset.

Exploratory Analysis:

To start, the first step was to explore and understand the data that was available. Using functions like *dim(), attributes()*, and *str()* it was found that there were 9 variables consisting of numeric values. Looking deeper into the data it was found that the ID field was simply a unique identifier, and not a continuous numeric value. Next it was observed that the Capsule, DPROS DCAPS, Gleason, and Race variables were categorical. For example, race has only two values, 1 and 2, to show whether a person was white or black. Lastly, it was seen that the variables Age, Vol, and PSA were continuous numerical values that did not reflect any category. Figure 2 below shows the summary of the dataset, which shows a few skewed variables like race, DCAPS, and Capsule. This shows that there is a misrepresentation of black people and DCAPS equal to yes.



```
> summary(lowbwt)
 CAPSULE        AGE            RACE                       DPROS       DCAPS          PSA              VOL            GLEASON
 No :225   Min.   :47.00   White:340   No Nodal            : 98   No :336   Min.   :  0.30   Min.   : 0.00   Min.   :0.000
 Yes:151   1st Qu.:62.00   Black: 36   Unilobar Nodule (L):131   Yes: 40   1st Qu.:  5.00   1st Qu.: 0.00   1st Qu.:6.000
           Median :67.00               Unilobar Nodule (R): 95             Median :  8.75   Median :14.25   Median :6.000
           Mean   :66.06               Bilobar Nodule     : 52             Mean   : 15.28   Mean   :15.88   Mean   :6.383
           3rd Qu.:71.00                                                   3rd Qu.: 16.88   3rd Qu.:26.60   3rd Qu.:7.000
           Max.   :79.00                                                   Max.   :139.70   Max.   :97.60   Max.   :9.000
```

Figure 2: summary of dataset

Next the distributions of the less obvious variables were studied. Age and Gleason score seemed to have evenly distributed data as shown in figure 3 below, while PSA and Vol has skewed data as shown in figure 5 and 6 below. PSA and Vol being skewed makes sense because if a person has an abnormally high antigen level or tumor volume, they have a higher likelihood of dying. The point of these tests are to find volumes and levels that are above average, which is why the skew is likely to occur.

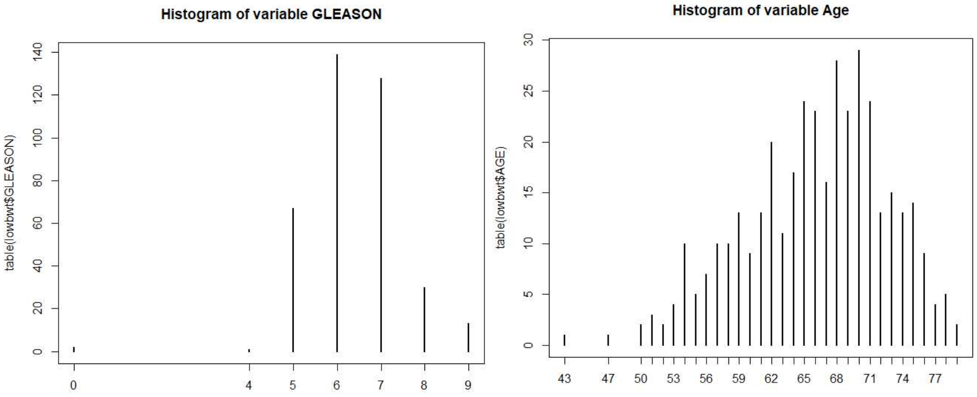


Figure 3: histograms of the variables Gleason and Age

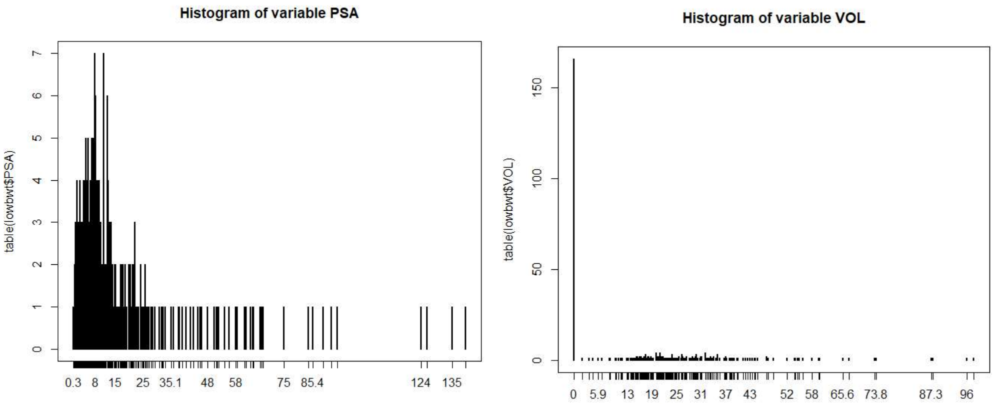


Figure 4: histograms of the variable PSA and Volume (VOL)

Preprocessing:

As discussed earlier the ID column was a unique identifier and served no purpose for the model, for that reason it was dropped from the dataset so the model wouldn't be affected. There were also a few missing values, that were dropped from the data set to provide the model with clean data. The last step was to split up the dataset into two parts for the model. The first part was the test training set, which would be used to train the two models. The second was used to test and score the models, to figure out how well they did. The reason the second test set was necessary was because it was data the model never saw while it was being trained.

Algorithm Intuition:

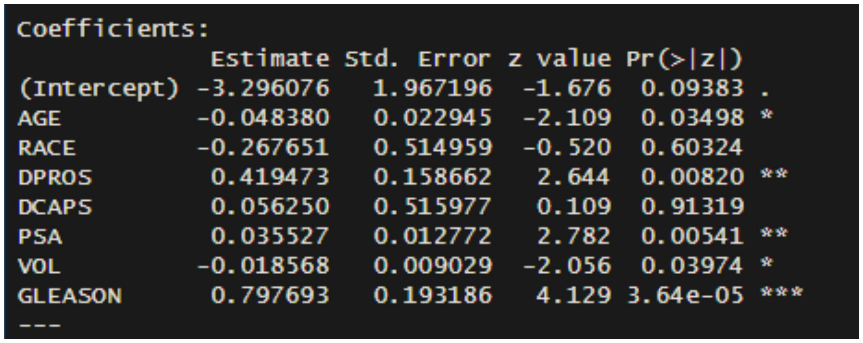
This model will be using a multi variable dataset to predict probabilities, which makes logistic regression a more suitable model than linear regression. Logistic Regression is the model used in this analysis that shows the probability of an event occurring between 0 and 1, in this case the event is prostate cancer. The logistic algorithm found, shown in figure 5 below, will calculate the probability of a person having prostate cancer based on the value for each variable in the model. Each β will be unique based on the data that the model is learning from, and together the probability of cancer will be a function based on each person observed (Firdu. 2023).

$$P(y|x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}$$

Figure 5: shows the algorithm created using logistic regression (Firdu. 2023)

Model Fitting:

After creating and testing an initial model with the original dataset, the columns race and DCAPS were decided to be dropped as well when analyzing the underlying statistical scores. After looking at the p-values in figure 6 below, it is shown that the variables Race and DCAPS had p-values well above 0.05. This shows that these two columns have very low statistical



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.296076   1.967196  -1.676  0.09383 .
AGE         -0.048380   0.022945  -2.109  0.03498 *
RACE        -0.267651   0.514959  -0.520  0.60324
DPROS        0.419473   0.158662   2.644  0.00820 **
DCAPS        0.056250   0.515977   0.109  0.91319
PSA          0.035527   0.012772   2.782  0.00541 **
VOL         -0.018568   0.009029  -2.056  0.03974 *
GLEASON      0.797693   0.193186   4.129 3.64e-05 ***
---
```

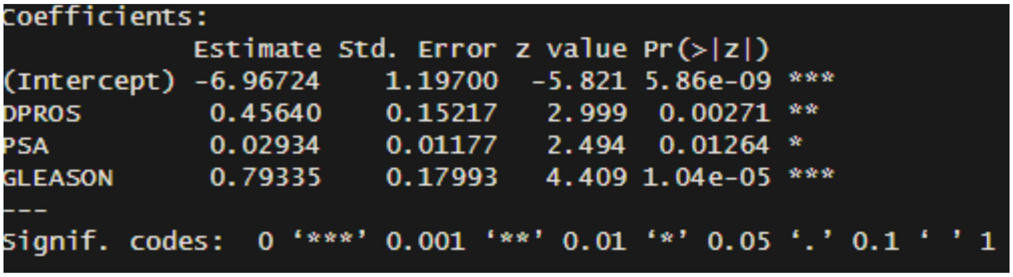
Figure 6: Shows logistics regression model on original dataset.

significance in predicting whether someone has prostate cancer. It is also shown that gleason was the most significant predicting variable with a p-value of 3.64e-05, followed by PSA and DPROS. It is also worth noting that the standard error was low for these variables, while the z-score was negative for VOL and AGE. Since z-score is a comparison against the mean, a negative score is worth flagging, because it did worse than the mean, which is why these variables were removed. After removing age, race, DPROS, and VOL the final model was created and used to predict whether a person had prostate cancer.

Result:

Output and Model Properties:

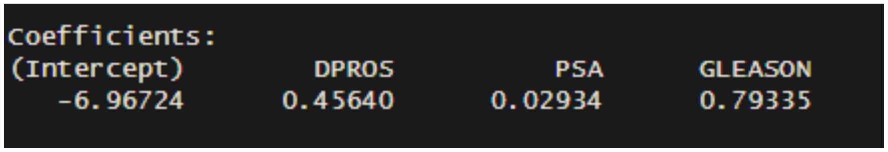
Now that a model was made with variables that produced statistically significant p-scores and positive z-scores, the results were then analyzed to see if the predictions were successful showing whether a person had prostate cancer. The final model can be seen in figure 7 below, and all variables used had low p-values of less than 0.05, proving that they are statistically significant in proving whether someone has prostate cancer. Another note is all z-values were positive showing that each variable provided value against the mean. This raised the total confidence in the model because it shows that statistically the probabilities are not due to random chance, and each variable added value to the model.



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -6.96724    1.19700  -5.821 5.86e-09 ***
DPROS        0.45640    0.15217   2.999  0.00271 ** 
PSA          0.02934    0.01177   2.494  0.01264 *  
GLEASON      0.79335    0.17993   4.409 1.04e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7: Shows logistics regression model on new cleaned dataset.

The binomial logistic regression model was created based on the variables DPROS, PSA, and Gleason, and can be seen in figure 8 below. It had a total of four β coefficients, for each of the variables plus the constant.

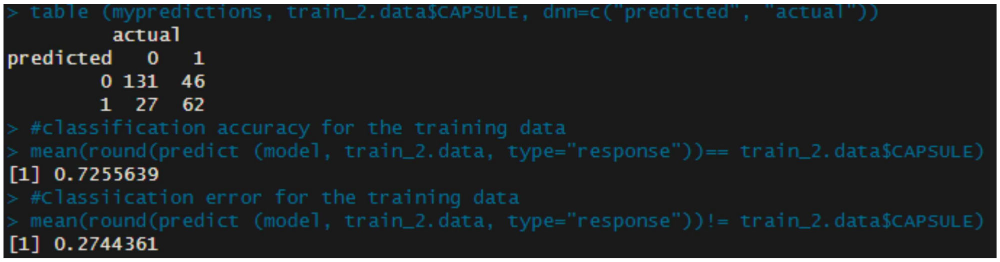


```
Coefficients:
(Intercept)        DPROS          PSA      GLEASON  
   -6.96724      0.45640      0.02934      0.79335  
```

Figure 8: Model properties, specifically the coefficients

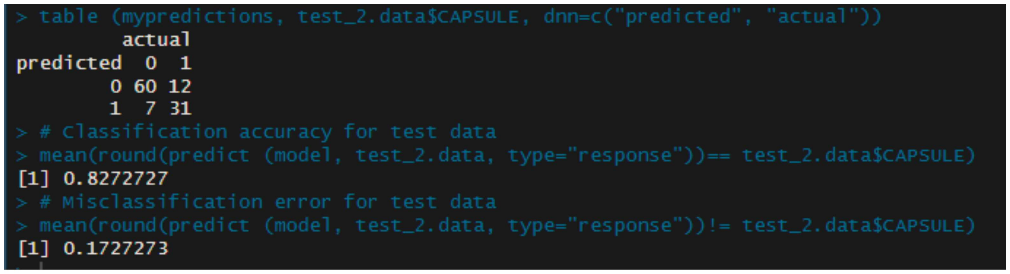
Evaluation:

Figure 8 shows that the model was 72.5% accurate in predicting prostate cancer for the training data, while figure 9 shows that the model was 82.7% accurate in predicting the data within the test dataset.



```
> table (mypredictions, train_2.data$CAPSULE, dnn=c("predicted", "actual"))
         actual
predicted   0   1
        0 131  46
        1  27  62
> #classification accuracy for the training data
> mean(round(predict (model, train_2.data, type="response"))== train_2.data$CAPSULE)
[1] 0.7255639
> #Classiication error for the training data
> mean(round(predict (model, train_2.data, type="response"))!= train_2.data$CAPSULE)
[1] 0.2744361
```

Figure 8: shows results of model against the training dataset



```
> table (mypredictions, test_2.data$CAPSULE, dnn=c("predicted", "actual"))
         actual
predicted  0  1
        0 60 12
        1  7 31
> # Classification accuracy for test data
> mean(round(predict (model, test_2.data, type="response"))== test_2.data$CAPSULE)
[1] 0.8272727
> # Misclassification error for test data
> mean(round(predict (model, test_2.data, type="response"))!= test_2.data$CAPSULE)
[1] 0.1727273
```

Figure 9: shows results of model against the training dataset.

This shows that the objective of the analysis was met because the model was 82% accurate on data it had not yet seen before. Next the ROC curve, shown in figure 10, was analyzed to show that the model had good performance at various thresholds. The goal is to have a model that is above the 45 degree line, which shows that the model is better at predicting true positives and true negatives. In this case the model is significantly greater than the 45 degree line, proving it has great performance in predicting whether a person has prostate cancer.

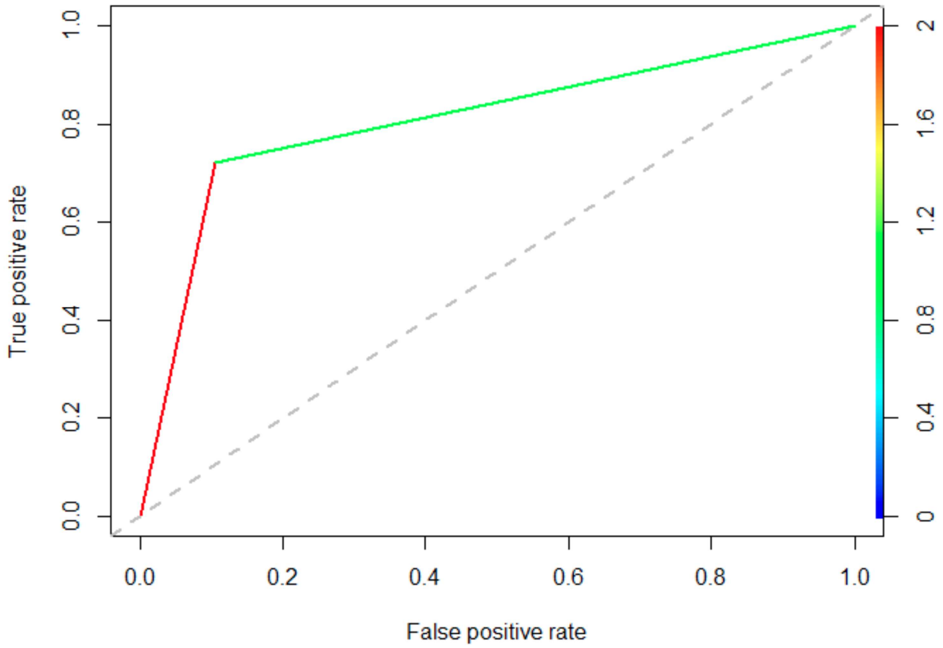


Figure 10: ROC graph for the created model

Diagnostics:

After the model was evaluated, a residual model was then used to perform diagnosis on the model, shown in figure 11 below. What is plotted are the residuals, which are "the difference between the predicted value and actual value" (Data 630. 2023) and the model predictions. There are two blue curves because a person can be predicted as having cancer (1) or not having cancer (0). As shown in the plot below, the residuals create two separate blue lines for each possible outcome. The black line shown is the predicted model, which follows the blue boundary lines set by the residuals. Based on this, the plot looks to validate the assumptions to use logistic regression because of how the residuals are set along the predicted model.

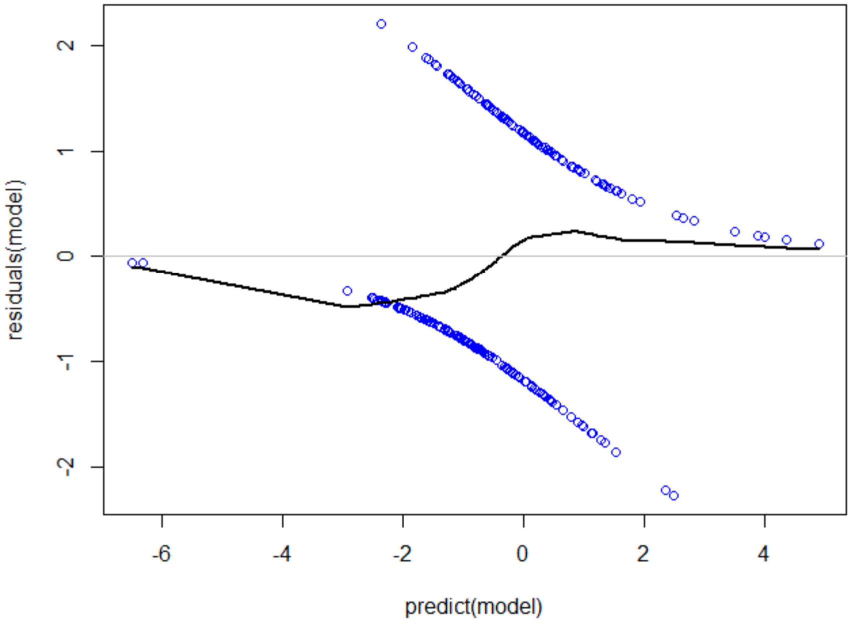


Figure 11: residual plot for the created model

Conclusion:

Summary:

The analysis was successful and showed 82% accuracy in predicting prostate cancer in the test dataset. However, when it comes to someone's life being taken, the goal would be to move the accuracy from 82% to 99% or above. The data had p-values less than 0.05 which showed that the variables provided statistical significance that was not due to random chance, and the positive z-values showed the variables added value when compared against the mean.

Limitations:

One limitation with the data was the number of observations there were. With a total of 380 observations, the data will likely be biased towards the subset and will likely not be strong enough to experience a large deviation from what it has seen. For example, there was a strong bias towards white people in the data, meaning the algorithm was worse at predicting cancer for black people. This is no fault with the model itself, but simply the scope of the data that trained it.

Improvement Areas:

As discussed in the limitation the number of observations was a weak point of this analysis. Future analysis would benefit by expanding the dataset to have more observations, but also a stronger presence of data that is of interest to the use of the model. For instance, if a hospital is in an area that has predominantly black people, then building the data to include a lot more observations that reflect its population. This in turn would allow the model to create more accuracy towards the population, because it is trained with data that is reflective of the population. By exploring models with stronger data, the next steps would test whether these provide stronger models and expanding the data to be more supportive of a population. Also, as discussed previously when it comes to medical diagnosis, it is expected people will want up to 100% accuracy when their lives are on the line. For that reason, the model will always need to be improved to push the accuracy of the predictions as close to 100% as possible.

References:

Prostate Cancer Foundation (PCF). (2022, August 24). *Top 10 things you should know about prostate cancer*. Prostate Cancer Foundation. https://www.pcf.org/c/top-10-things-you-should-know-about-prostate-cancer/

Firdu, B. (2023). *Logistic Regression* [PDF]. Data 630 - UMGC. https://learn.umgc.edu/content/enforced/770634-027339-01-2235-GO1-9040/Logistic%20Regression1.pdf

Data 630. (2023). *Logistic Regression Using R* [Word document]. UMGC.