Doug Marcum

DSC 630 – Predictive Analytics

Final Report – Breast Cancer

Executive Summary

Cancer is a disease that attacks tens of thousands of people annually. The ability to quickly and accurately diagnosis this disease greatly increases one's ability to overcome and survive this affliction. The purpose of this project was to determine if machines can accurately assist doctors, particularly breast cancer specialists, in making an accurate predictive diagnosis relating to breast cancer.

Currently breast cancer is diagnosed by completing one or more the following:

- **Breast ultrasound.** A machine that uses sound waves to make detailed pictures, called *sonograms,* of areas inside the breast.
- **Diagnostic mammogram.** If you have a problem in your breast, such as lumps, or if an area of the breast looks abnormal on a screening mammogram, doctors may have you get a diagnostic mammogram. This is a more detailed X-ray of the breast.
- **Magnetic resonance imaging (MRI).** A kind of body scan that uses a magnet linked to a computer. The MRI scan will make detailed pictures of areas inside the breast.
- **Biopsy.** This is a test that removes tissue or fluid from the breast to be looked at under a microscope and do more testing. There are different kinds of biopsies (for example, fine-needle aspiration, core biopsy, or open biopsy).[1]

This project examined over 500 fine-needle aspiration biopsies of breast tissue samples. In order to accurately determine if the predictive model could make predictions, multiple characteristics were evaluated to determine their value in our formula. Additionally, numerous models were processed to see which could return the best measurable metrics. After various rounds of testing, each model was able to accurately predict a correct diagnosis of a cancerous tumor over 90% of the time. These results are exciting and will be valuable to the medical community.

[1] Centers for Disease Control and Prevention. (2020, September 14). How is Breast Cancer Diagnosed?. Retrieved from https://www.cdc.gov/cancer/breast/basic_info/diagnosis.htm.

Abstract

Breast cancer is the second leading cause of death from cancer in women, but outcomes have been shown to improve if caught and treated early. Mammogram screenings are an excellent tool in identifying a potential tumorous mass, even before symptoms may appear. While imaging technology has increased resolution and clarity, interpretation of the data and the image as still exposed to potential false positive or false negative findings by breast imaging radiologists. Can predictive analytics be used to hone and improve cancerous tumor diagnosis? Data from the "*Diagnostic Wisconsin Breast Cancer Database*" was analyzed and ran through four predictive classifier models. The findings of this analysis show that through machine learning, machines can diagnosis breast cancer at a highly effective rate. In doing so, it will be another tool in the arsenal of early detection and minimizing the damaging effects of the disease.

Introduction

Background

Breast cancer is a disease that impacts the lives of thousands annually. It is estimated that 1 in 8 U.S. women (approximately 12%) will develop invasive breast cancer over the course of their life. An estimated 276,480 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S., along with 48,530 new cases of non-invasive (in situ) breast cancer over the course of 2020. While less common, occurrences of breast cancer in men are becoming more common. Men have a 1 in 883 risk of developing during their lifetime.

Unfortunately, even with advances in the field, incidence rates have increased slightly (by 0.3% annually) in recent years. [2]

Many studies have been conducted in this field, and the "*Diagnostic Wisconsin Breast Cancer Database*" is a publicly available data set from the UCI machine learning repository that has gained a recognition as one of the first milestones of artificial intelligence. The dataset shares information about tumor features, that were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. For each observation there are ten features, which detail tumor size, density, texture, symmetry, and other characteristics of the cell nuclei present in the image. The mean, standard error and "worst" mean (mean of the three largest values) of these features were computed for each image, resulting in 30 features. The categorical target feature indicates the type of the tumor, malignant or benign.

While the approach and technology are dated by today's standards, the data is still quite valuable for extracting meaningful insights. The goal of this project is to examine the results of multiple features being utilized in providing predictive value between benign and malignant breast tumors.

Methods

Data Preparation

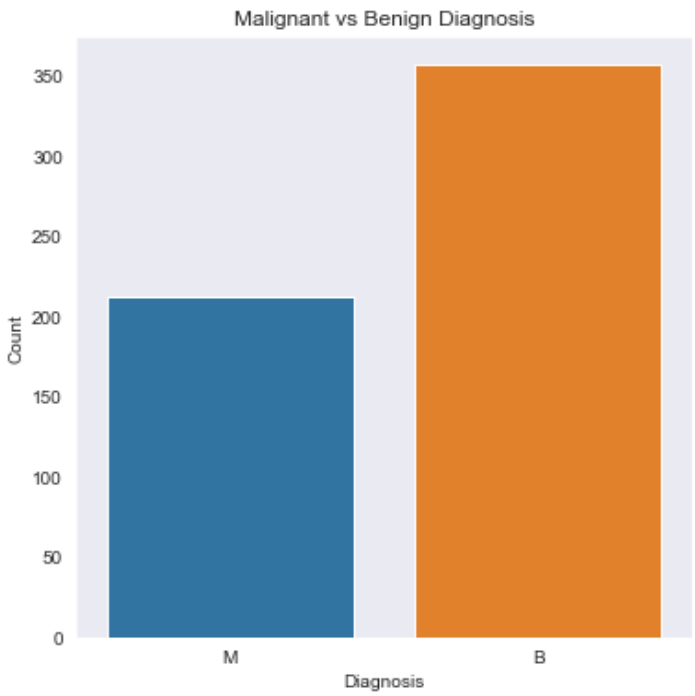
For data preparation, Python was utilized to read in and clean the initial dataset. Fortunately, not much was needed in order to have a working dataset. A generic unnamed column was created during the reading of the data, and that column was removed. Upon the

[2] American Cancer Society. (2020, January). How Common Is Breast Cancer? https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html.

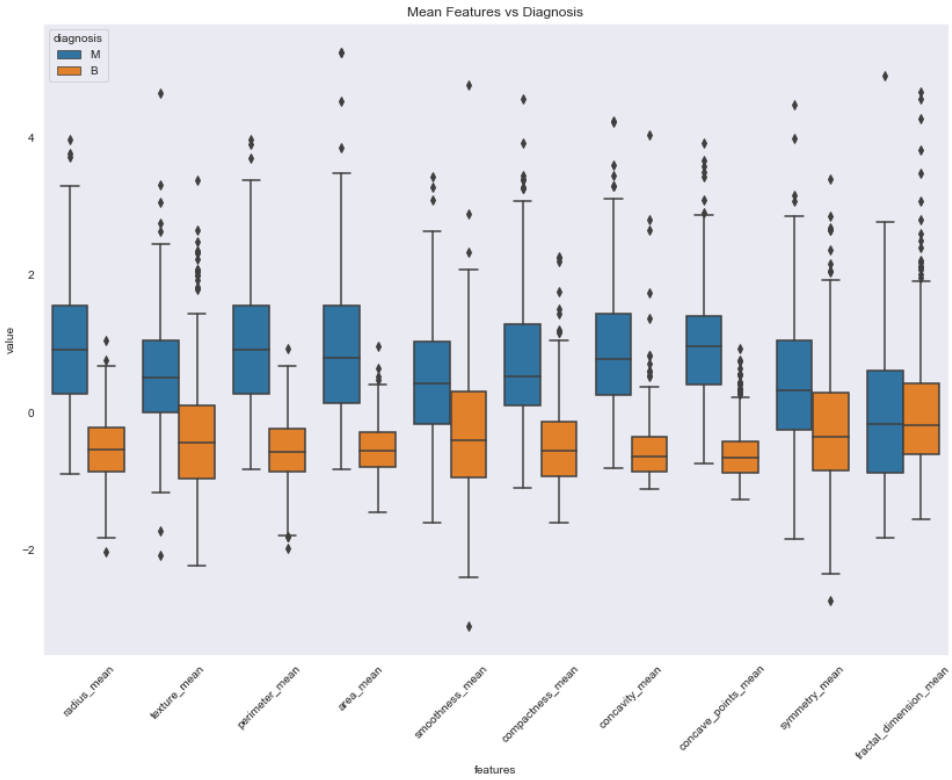
removal of the unnamed column, analysis showed that there are no NULL or empty field values. From here, no duplicate values were discovered by reviewing the “id” column for multiple entries.

Exploratory Data Analysis

Exploring the data has led to some basic and interesting discoveries. The data consists of 569 records, with 212 as malignant and 357 diagnosed as benign, 37.26% and 62.74% respectively.



Feature sets (mean, standard error, worst) were normalized and plotted in box plots and pair plots to show statistical summaries and relationships between the features and diagnosis. Two examples are shared below, both highlighting mean features of mass growths.


Mean Features vs Diagnosis
diagnosis
M
B
value
features
radius_mean
texture_mean
perimeter_mean
area_mean
smoothness_mean
compactness_mean
concavity_mean
concave_points_mean
symmetry_mean
fractal_dimension_mean

The box plot shows a high number of outliers for both malignant and benign diagnosed growths. Since the dataset is a smaller sample size, 529 records, the outliers remained in the analysis. Additionally, as with most abnormal growths, it is difficult to truly identify each of these as truly outliers. With the pairs plot, we are able to visualize initial patterns of how the features and diagnosis relate to one another.

Each feature's distribution was plotted, it became clear that most were skewed to the right. All features were thus transformed using log10 (with .001 being added to account for zero values in the features). This allowed most features to form normal distributions, allowing for better analysis and avoiding future overfitting.

Feature Selection

Feature selection was the most important and challenging portion of the analysis. With the data set being relatively small, 529 records, and many of the features having high correlations, it was vital to identify the most important features, but also not to eliminate those with minor significance. This was needed to preserve the integrity of the data, but also to provide a broader scope in the model's ability to accurately make consistent, valid predictions.

To begin, a Lasso linear cross validation model was executed, with irrelevant features in LassoCV having coefficients turned to zero (0). This initial feature selection was able to reduce the number of features from 30 to 23. After this reduction was completed, a correlation heatmap was constructed, and a number of features still had high correlation rates.

A principal component analysis (PCA) was completed, but this eliminated all but six (6) features. I believe this is part due to the fact that PCA underestimates space dimensionality,

which may cause over selection. This is why LassoCV was selected as the first step in feature selection.

In order to account for such high levels of multicollinearity, an Ordinary Least Squares Regression model was utilized. While maintaining an R-squared over .750, features were dropped and reevaluated if their p-value was greater than 0.05. After numerous passes, the final dataset was constructed with 11 features being selected.

Modeling

The data was randomly split into 80% training and 20% testing data sets. Since this is a classification prediction problem, the following models were selected for evaluation: Random Forest Classifier, Logistic Regression CV, K-Nearest Neighbor Classifier, and Support Vector Machines (SVM) with Support Vector Classifier. A simple function was created for displaying each model's metrics. Each model was fit and tuned using the training data set, and a prediction was made and evaluated using the training data. Models were evaluated in Python and in R.

Results

The results from each model are shared below.

Model Metrics	RF	LRCV	KNN (n=3)	SVM
Train Accuracy Mean (CV=10)	0.9273	0.9714	0.9472	0.9582
Test Accuracy Mean (CV=10)	0.9303	0.9561	0.9561	0.9644
AUC Score	0.9672	0.9415	0.9605	0.9477
Precision	0.96	0.94	0.95	0.94
Recall	0.97	0.94	0.96	0.95
F1 Score	0.96	0.94	0.95	0.94
Accuracy	0.96	0.95	0.96	0.95

After tuning we can see that each of the models performed extremely well, with high marks in each measure across the board. Logistic Regression did return a slightly lower test accuracy score than the training model, but the difference is relatively small. In future replications of these models with additional data, this should be reevaluated. In reviewing each model's performance, Random Forest Classifier did marginally better in each category than the other models, leading to its selection as the model of choice. Random forests are well designed for utilization in a classification problem because they can handle both continuous and discrete variables equally and they are not sensitive to outliers, like other models.

Conclusion

All models in this project exceeded expectations. Given the small sample size, it was questionable if the results would be as uniform as returned. While it may appear that the Random Forest Classifier or one of the other models may be ready for deployment, it is recommended that a much large sample be evaluated under the same constraints to truly validate the findings.

Overall, it is clear that predictive analytics are a tool that can be vital in the war of early detection in breast cancer and other areas of medical study. While sole reliance on the prediction of an algorithm is not wise, but rather by working in conjunction with well trained oncologists and radiologists, data scientist can continue to assist in the acceleration of diagnostic medicine.

Acknowledgements

I would like to acknowledge my family for allow me the time needed to tackle this project, my classmates for valuable insights and feedback, Professor Werner for his direction throughout the course, and to those who shared their samples to this study and others like it.

References and Data Sources

References

American Cancer Society. (2020, January). How Common Is Breast Cancer? https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html.

American Cancer Society. Cancer Facts & Figures 2020. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf.

National Cancer Institute. (2018, January). BRCA Mutations: Cancer Risk and Genetic Testing. https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet.

McKinney, Wes. *Python for Data Analysis*. O'Reilly Media. Kindle Edition.

Data Sources

Kaggle - https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.

UCI Machine Learning Repository - https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.