Predicting Breast Cancer Using Tree-Based Classifier Models

Kunal Pathak
University of North Carolina at Chapel Hill
kunal.pathak@unc.edu

Kunj Patel
University of North Carolina at Chapel Hill
kunj@ad.unc.edu

Aadit Mehta
University of North Carolina at Chapel Hill
amehta1@unc.edu

Yutika Aggarwal
University of North Carolina at Chapel Hill
yaggarwal@unc.edu

May 1, 2024

Abstract

Classification is one of the fields at the forefront of machine learning research. Our daily lives already use advanced data processing models, whether for identifying spam emails or predicting defaulted loans. This project utilizes a dataset of patient breast imaging classification containing both qualitative and quantitative aspects. In this paper, we construct and compare the accuracy, precision, and recall to classify whether a patient has a malignant or benign breast mass across various tree-based models.

I. Introduction

i. The Data

The data set used for our research comes from Kaggle and contains 6775 studies on patient mammograms. For our approach, we decided to implement various tree-based algorithms and compare the results. The dataset came with various features that were refined to using the following: 'breast density', 'left or right breast', 'image view', 'abnormality id', 'mass shape', 'mass margins', 'assessment', and 'subtlety'.

The data is a curated breast imaging subset of DDSM. Each entry relates to a new image of a breast that is classified as either benign, malignant, or normal. Some patients had various image entries listed in the dataset, so it was important to separate the patient identification data from their corresponding pathology.

ii. Implications

Our research can be applied to automating work performed by healthcare professionals like Radiologists. The dataset can also be used with image classification software to ensure patients receive accurate diagnoses to provide life-saving medicine faster.

II. Framework

i. Preparing and Exploring the Data

The steps that we took are listed below, separated into three primary categories: Preparing & Exploring the Data, Building the Classifier Models, and Evaluating Model Results.

1. We downloaded the mass_case_description_test_set.csv and mass_case_description_train_set.csv file from Kaggle and read it into our notebook combining the two into a single pandas dataframe.

2. We established a new column binarily classifying whether a patient has a malignant tumor or benign tumor based on the pathology column. During this process, we removed any patient images that did not have a tumor present.

3. We then performed a correlation matrix on the dataset using Cramers V enumerate the correlations between all the features.

   (a) The Cramers V statistic is a measure of association between two categorical variables defined by the following calculation:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}. \tag{1}$$

4. Next, we removed irrelevant fields that did not impart any information regarding the diagnosis including file paths, patient ids, and abnormality type based on the results of the correlation matrix. In doing so, we aim to reduce the chances of over-fitting the data.

5. Then, we encoded all categorical features as numeric using one-hot encoding

6. Finally, we split the data into training and testing sets using a 75/25 split. It is industry standard and is a good split to avoid over-fitting

ii. Building the Classifier Models

These general steps were applied to each of the three classification models we compared:

1. Create classification model object (Decision Tree, Random Forest, or XGBoosting)

2. Fit the models using the X and y training data

3. Compare the models' abilities to make accurate diagnoses of malignant tumors in the breast.

iii. Evaluating Model Results

These general steps were applied to each of the three classification models we compared:

1. Make predictions using each of the trained models

2. Calculate accuracy of predictions using the Jaccard similarity index for in and out of sample data

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

3. Model predictions into confusion matrices to better understand the strengths and weaknesses of our models' predictive power

4. Compare the efficacy of the predictions made across models using the previous confusion matrices

III. Building the Models

Once our Kaggle dataset is loaded, we utilize several classifier models to predict whether a tumor is malignant or benign in the breast. We will test three different models to compare the accuracy of each and find a better solution for our prediction.

After training each of our models with the full dataset containing all the features, we realized that this would cause our final prediction to over-fit our training data. This is because some variables were highly irrelevant in diagnosing whether the patient has a malignant or benign tumor. Therefore, we set up a covariance matrix using Cramers V and removed features that did not impart any valuable information on performing the classification. No additional calculations were needed here since the confusion matrix made determining unnecessary columns easy. If this dataset had additional features, then a feature selection algorithm may have been necessary.

i. Decision Tree

The purpose of using the Decision Tree model is to have a baseline classification model that is simple to implement and an effective way to check whether other models exceed the performance and accuracy of this model.

A decision tree is a flowchart-like structure where each internal node represents a "decision" based on an input feature, each branch represents the outcome of that decision, and each leaf node represents the final decision or prediction.

The main calculations involved in building a decision tree include determining the best feature to split the data at each node (based on criteria such as information gain or Gini impurity), recursively splitting the data into subsets based on the selected feature, and stopping criteria to prevent over-fitting (e.g., maximum tree depth or minimum number of samples per leaf node).

ii. Random Forrest Classifier

Random Forest is one of the most popular and powerful classification algorithms because of its similarity to human-like thinking which leads to high intuition and interpretability.

A Random Forest classifier is an ensemble technique consisting of many decision trees. It builds an independent set of trees that classify inputs by voting on outcomes - an approach that tends to be more accurate than individual trees.

iii. XGBoost

XGBoost (Extreme Gradient Boosting) is a popular machine learning algorithm that is widely adapted across various applications and has proven to be a highly successful classification algorithm. The hope of this exploration is that this model will act as a gold standard and outperform both the decision tree and random forest due to its complexity.

XGBoost works by building a series of decision trees sequentially. Each tree corrects the errors made by the previous trees, leading to a more accurate prediction model. The algorithm starts with a single decision tree, which makes predictions based on the average target value of all samples in the training data. It then calculates the residuals (the differences between the actual target values and the predicted values) for each sample in the training data. Next, it builds decision trees to predict the residuals. Each tree is trained to minimize a loss function, such as mean squared error or log-loss (the model developed in this comparison utilizes a log-loss loss function). The algorithm then constructs a new tree to predict the residuals that were not explained by the previous trees. This process continues until a predefined number of trees has been built or a stopping criterion has been met. Lastly, XGBoost includes regularization techniques such as L1 and L2 regularization to prevent overfitting and improve generalization.

IV. Results

Keeping all shared variables the same, including random seed and maximum tree depth, which were set to 42 and 9 respectively, the results showed a consistent increase in accuracy, precision, and recall across the three models.

Model	Accuracy	Precision	Recall
Decision Tree	.83	.84	.81
Random Forest	.84	.86	.82
XGBoost	.86	.88	.83

Table 1: *Performance Metrics of Different Models*

The Decision Tree model was our baseline model achieving the lowest performance metrics across the three models. This met the expectations since decision tree models are the most basic and are prone to leaving out refinement opportunities addressed by the other two models. The Random Forest classifier improved on the decision tree model because of its more refined ensemble of weak learners. This supports Robert E. Schapire's idea that a collection of weak learners can achieve a more

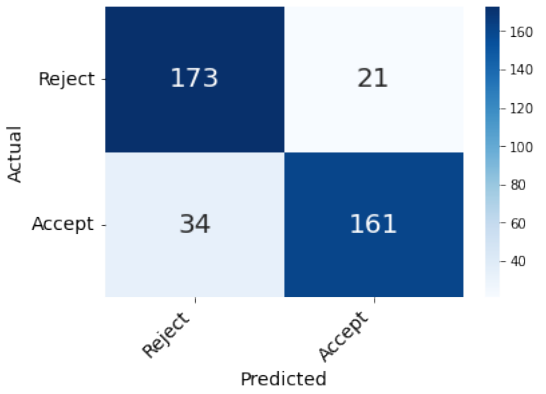

Figure 1: *Confusion matrix for XGBoost Model*

refined model. Finally, the XGBoost classifier was the most accurate machine learning model we trained, achieving an accuracy rate of 85.9%.

In the context of classifying these specific data, maximizing recall is imperative since a low recall score relates to higher false negative classifications. In the real world, this would mean a patient would believe that they do not have breast cancer when they do in reality. Fortunately, the XGBoosted solution was the most effective at minimizing false negatives and is therefore irrefutably the best model in this context.

V. Conclusion

Even though our results were fairly accurate in the XGBoosted model, this model is not refined enough to be used at an industry level. This is because the recall score must be significantly more refined to ensure that patients who have cancer are receiving the diagnosis they need to save their lives. One way these models could be improved is through the use of a larger data set or a more robust set of features. The dataset used in this exploration also included the corresponding mammogram images. An extension of this project could be to use computer vision algorithms to add quantifiable data regarding the breast mass. Some of these features could include the perimeter mean, area mean, smoothness mean, etc. Many of these features are included in the Breast Cancer Wisconsin (Diagnostic) Data Set, which has been used to train highly effective models. As a result, the model could train on exact measurements of the mass based on image inputs as opposed to mostly categorical data. This could potentially improve both the efficacy and usability of the models.

References

[1] Blockeel, H., Devos, L., Frénay, B., Nanfack, G., & Nijssen, S. (2023, July 26). Decision trees: From efficient prediction to responsible AI. Frontiers in artificial intelligence. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10411911/

[2] Natekin, A., & Knoll, A. (2013, December 4). Gradient Boosting Machines, a tutorial. Frontiers in neurorobotics. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3885826/

[3] Schapire, R.E. The strength of weak learnability. Mach Learn 5, 197–227 (1990). https://doi.org/10.1007/BF00116037

[4] Szegedi-Hallgató E, Janacsek K, Nemeth D. Different levels of statistical learning - Hidden potentials of sequence learning tasks. PLoS One. 2019 Sep 19;14(9):e0221966. doi: 10.1371/journal.pone.0221966. PMID: 31536512; PMCID: PMC6752858.