Predicting if a Breast Tumor is Benign or Malignant

Jonathan Daughtry

Breast Cancer is Common

Breast cancer is the second most common type of cancer in the United States (CDC)

About 1 in 8 U.S. women (about 13%) will develop invasive breast cancer over the course of her lifetime. (breastcancer.org)

About 43,600 women in the U.S. are expected to die in 2021 from breast cancer. (breastcancer.org)

For women in the U.S., breast cancer death rates are higher than those for any other cancer, besides lung cancer. (breastcancer.org)

Breast cancer became the most common cancer globally as of 2021, accounting for 12% of all new annual cancer cases worldwide, according to the World Health Organization. (breastcancer.org)

The Problem:

With the high prevalence in breast cancer, especially among women, how can doctors accurately determine early if a tumor cell is benign or malignant?

The Solution:

Create a model that can accurately predict whether a breast tumor is benign or malignant and reduce the number of false negative diagnoses

Who Might Care?

Women: About 85% of breast cancers occur in women who have no family history of breast cancer.

Doctors who perform the diagnosis

Families with a history of breast cancer or with any woman that might be at risk.

Data Information

Data collected from Breast Cancer Wisconsin (Diagnostic) Data Set
Number of records used: 469

Target Variable:
Whether a tumor was malignant(1) or benign(0)

Determining Variables:
1) radius
2) texture
3) perimeter
4) area
5) smoothness
6) compactness
7) concavity
8) concave points
9) symmetry
10) fractal dimension

For each determining variable, there was a mean value, a worst value, and a standard error

Data Cleaning:
https://github.com/dawgtree/CapstoneThreeProject/blob/main/Cancer%20Diagnosis%20Capstone%20Project%20Data%20Wrangling.ipynb

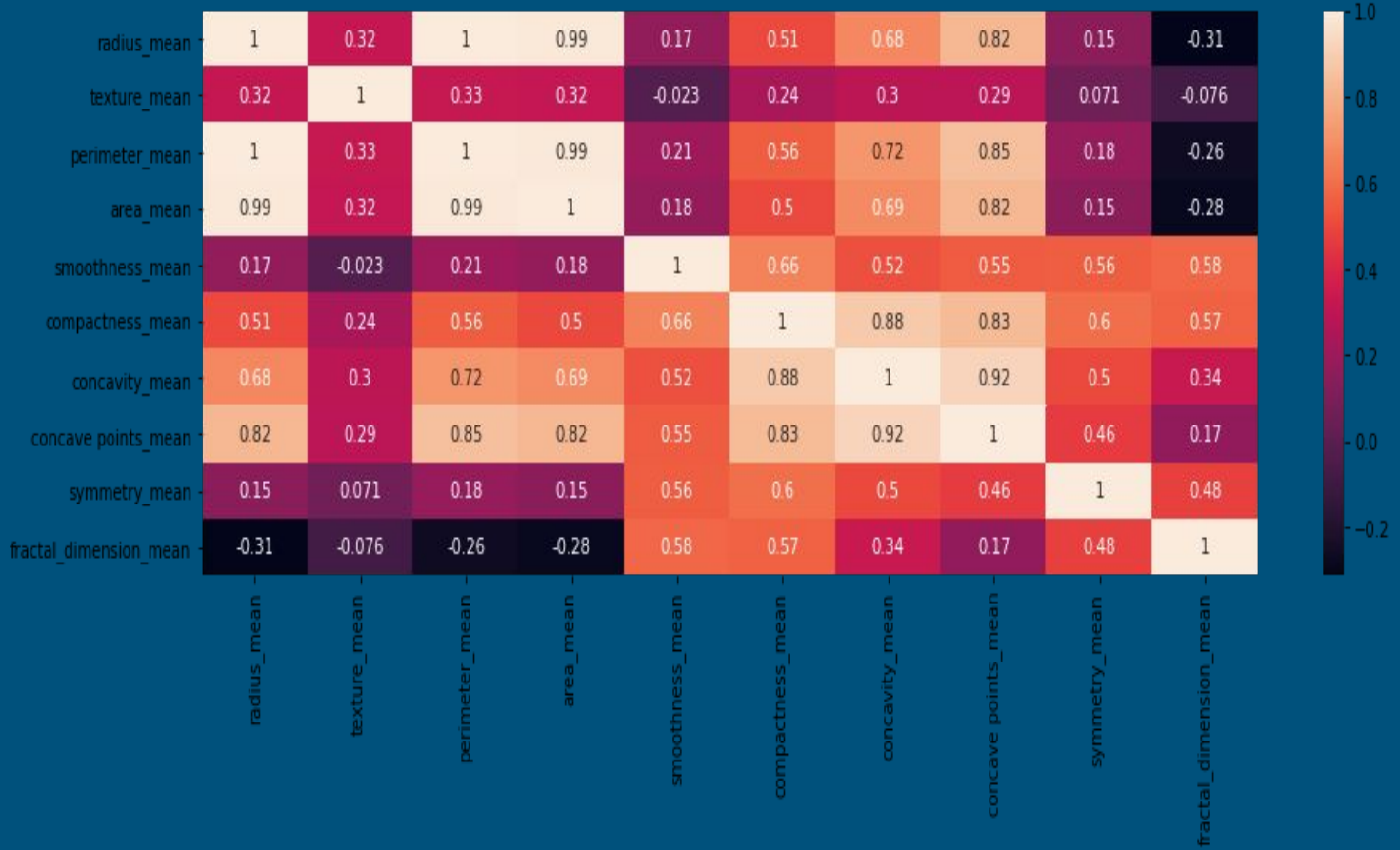
Data Source: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Data Exploration: Malignant vs Benign on Means

- Hypothesis testing was done to see if the differences between benign and malignant tumors were considered significant. P-values are listed below

Variable mean	p-value
Radius	8.47×10^{-96}
Texture	4.06×10^{-25}
Smoothness	1.05×10^{-18}
Compactness	3.94×10^{-56}
Concavity	9.97×10^{-84}
Concave Points	7.10×10^{-116}
Symmetry	5.73×10^{-16}
Fractal Dimensions	0.76

Data Exploration: Comparing Different Means



Determining Variables Remaining After EDA

Radius Mean
Texture Mean
Smoothness Mean
Area Standard Error
Concavity Standard Error

All other variables were removed due to very high correlation

Since a large majority of the variables were removed, I decided in my modeling I would model the data using just these five variables and model keeping all of the variables except for area and perimeter mean.

EDA: https://github.com/dawgtree/CapstoneThreeProject/blob/main/Cancer%20Diagnosis%20Capstone%20Project%20EDA.ipynb

Tableau EDA:
https://public.tableau.com/profile/jonathan.daughtry#!/vizhome/BreastCancerCapstoneProjectEDA/BreastCancerEDA

Modeling

Split into two parts: Creating baseline model and extended modeling

Baseline Model Used:	Logistic Regression
Extended Models Used:	K-Nearest Neighbors
	Random Forest Classification
	XGBoost Classification

Baseline Model: Logistic Regression

Hyperparameters used: L2 Penalty
10 for C

Effectiveness Measurements-
Accuracy: 97.37%
False Negatives: 2 of 114 (1.75%)
Recall: 95%
AUC: 99.5%

Based on these measurements, the targeted metrics to consider a model successful are:

1. Accuracy of at least 97%
2. Less than 2 % false negatives
3. A recall of at least 95%
4. AUC of at least 99%

Baseline Model:
https://github.com/dawgtree/CapstoneThreeProject/blob/main/Cancer%20Diagnosis%20Capstone%20Project%20Baseline%20Model.ipynb

Extended Models

Best Hyperparameters- KNN: 6 neighbors, a power parameter of Manhattan Distance, weight on distance.
Random Forest: max depth of 30, a log2 of max features, 300 estimators.
XGBoost were a learning rate of 1, a max depth of 2, and 10 estimators.

Model Comparison

	Accuracy	False Negatives	Recall	AUC
KNN	95.6%	4 out of 114 (3.5%)	91%	97.8%
Random Forest	96.5%	3 out of 114 (2.6%)	93%	99.6%
XGBoost	97.37%	2 out of 114 (1.75%)	95%	99.3%

XGBoost scored the best and was the only model along with the baseline model to meet the threshold. KNN scored the worst.

Modeling:
https://github.com/dawgtree/CapstoneThreeProject/blob/main/Cancer%20Diagnosis%20Capstone%20Project%20Modeling.ipynb

Takeaways

XGBoost was the only model that consistently met all of the established thresholds, along with the baseline Logistic Regression Model.
Their effectiveness could be affected by further adjustments of the hyperparameters and adding more datapoints

Reduction of false negatives was the most important factor in modeling,
Used the interactive Tableau that I created to see what would be the highest threshold for considering a tumor malignant where there would be no false negatives.
Logistic Regression: the highest threshold was 10%, and this created eight false positives, which was a difference of six more than the standard threshold of 50%.
XGBoost: at a threshold of 5%, there was still one false positive, with seven false positives. When the threshold was set to 0%, there were zero false positives, but all of the true negatives turned into false positives.
Logistic Regression performed the best.

Tableau Interaction:
https://public.tableau.com/profile/jonathan.daughtry#!/vizhome/BreastCancerModelingwithFullData/Story1

Future Research/Improvements

Adjust the features in which to model in different ways to see what would be an ideal set of features to model against that would retain the high scores that are needed and reduce correlation.

The testing set was fairly small, with only 114 sets of data to test. Ideally, these models would be best suited to test new data as it was gathered to see if it continued to meet the threshold that we want.

The models might benefit from improvement by further adjustment and expansion of their hyperparameters. I could also add hyperparameters that I did not include initially.

Recommendations

XGBoost and Logistic Regression are ready to be presented in a business setting.

Have a low threshold of an initial positive diagnosis between 0-5% to ensure that any potential malignant tumors would not be overlooked, and then do further testing on the malignant diagnoses.

This model be used as the first predictor of if a tumor is benign or malignant, but in a situation that is as serious as this, continuous monitoring should always be done.

Jonathan Daughtry
Email: daughtryje@gmail.com
https://www.linkedin.com/in/jonathan-daughtry/
https://github.com/dawgtree
Full detailed report of project:
https://github.com/dawgtree/CapstoneTwoProject/blob/main/Capstone%20Two%20Final%20Report.pdf