Breast Cancer

Prediction Model

Table of Contents

Introduction 1

What is Breast Cancer?

What is Breast Cancer?

Breast cancer is the most prevalent form of cancer among women in the US, excluding skin cancer. It ranks as the second-leading cause of cancer-related deaths in women, with lung cancer being the first. Timely detection and prompt treatment of breast cancer are vital in enhancing survival rates and improving outcomes. Women can play an active role in preventing breast cancer by focusing on early detection and adopting preventive measures.

Intro to Breast Cancer

- Breast cancer is a type of cancer that develops in the cells of the breast tissue. Breast cancer can occur in both men and women, although it more common in women.

- Breast cancer usually begins in the mild-producing glands (lobules) or the ducts that transport milk from the lobules to the nipple. It can also occur in the other breast tissues.

- Breast cancer can be invasive or non-invasive. Invasive breast cancer means that the cancer cells have spread to other tissues in the breast or beyond. Non-invasive breast cancer, also called in situ breast cancer, means that the cancer cells have not yet spread beyond the original site of the tumor.

Intro to Breast Cancer

- Symptoms of breast cancer may include a lump or thickening in the breast or underarm area, changes in the size or shape of the breast, nipple discharge, or changes in the skin of the breast. However, not all breast cancers cause noticeable symptoms, which is why screening for breast cancer is important for early detection.

- Breast cancer treatment depends on the type and stage of cancer, but can include surgery, radiation therapy, chemotherapy, targeted therapy, and hormone therapy.

Brife Breast Cancer History

1895

X-Ray were discovered, which led to development of mammography

1950s-60s

Radiation therapy became a common treatment. Chemotheapy beganto be used for treatment.

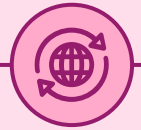
1980s

1st breast-conserving surgery (lumpectomy) was developed. (Removed the tumor and small margin of surrounding tissue.

1990s

Screening mammogram became widely available

2000s



Advances in imaging technology and genetic testing have improved the detection and treatment of breast cancer, and targeted therapies have been developed.

Breast Cancer Stats

1 in 8

Of U.S. women are going to delevop invasive breast cancer

287,850

In 2022, new cases of invasive breast cancer and 51,400 non-invasive

30%

Of newly diagnosed cancers are breast cancers & a common diagnosed cancer among American women

85%

Breast cancers occurs in women who have no family history

1 in 833

In 2022, 2,710 of Men with invasive breast cancer

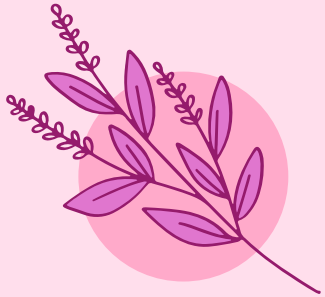
62

The median age of women diagnosed with breast cancer

3,800,000

As of January 2022,
Women with a history of
breast cancer in the U.S.

Key Takeaways



Stakeholders

Healthcare professionals
researchers
patients

Importance

By developing reliable models to predict tumor size can have significant implications for improving patient outcomes and advancing cancer research.

What is the importance of building a model and what are its accomplished?

~ Predicting tumor size accurately can be crucial in cancer diagnosis and treatment.

~ Accurate predictions of tumor size can help physicians determine the appropriate treatment options and create a personalized treatment plan for each patient.

~ It can also help monitor the progression of the tumor and the effectiveness of the treatment.

~ Accurate tumor size predictions can aid in clinical trials by identifying patients who are most likely to benefit from a particular treatment.

2 Research Questions

Can we build a model that predicts tumor sizes for patients?

Data Analysis 3

SEER BREAST CANCER DATA

https://ieee-dataport.org/open-access/seer-breast-cancer-data

Breast Cancer Dataset

4024, 16

Rows & Features

No missing values

But 1 duplicate

Split dataset into categorical & numerical features

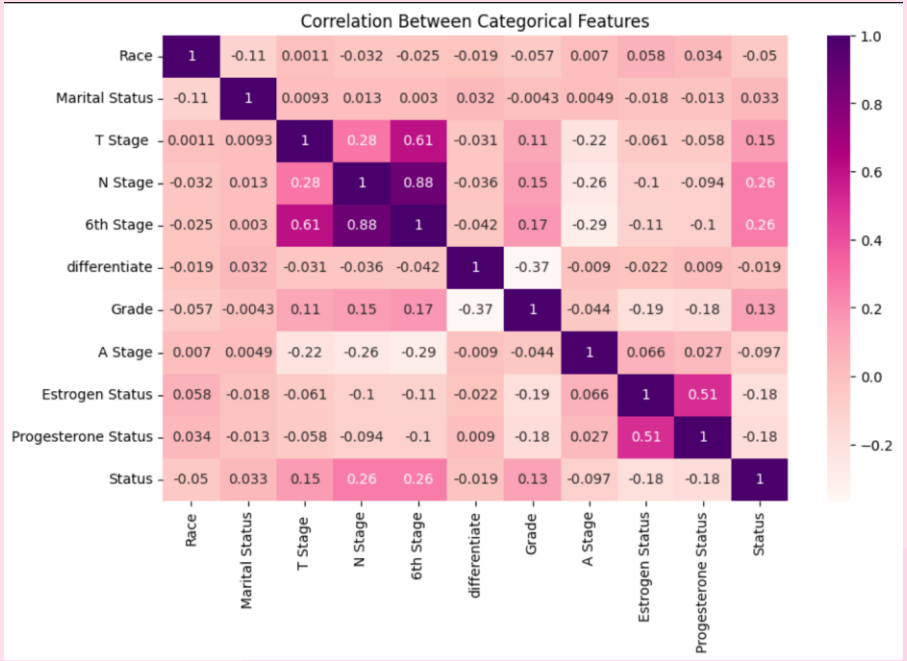
No null values

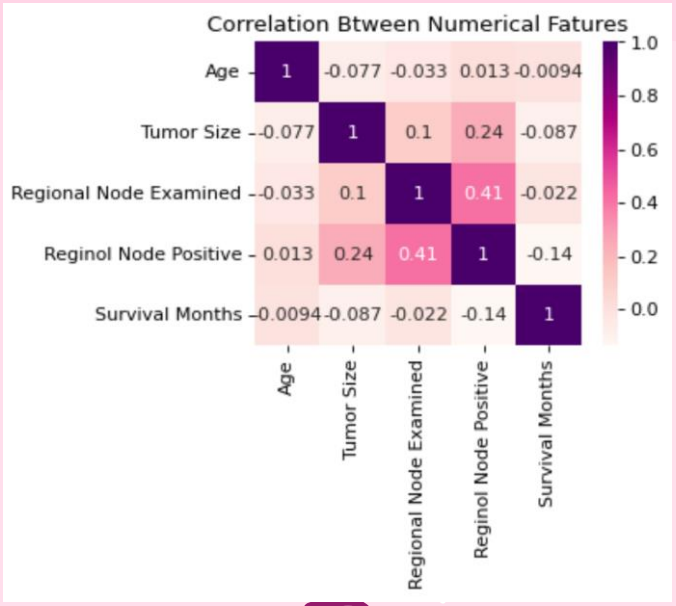
4 EDA

Exploratory Data Analysis

Include EDA visualizations

Visuals


Correlation Between Categorical Features
Race 1 -0.11 0.0011 -0.032 -0.025 -0.019 -0.057 0.007 0.058 0.034 -0.05
Marital Status -0.11 1 0.0093 0.013 0.003 0.032 -0.0043 0.0049 -0.018 -0.013 0.033
T Stage 0.0011 0.0093 1 0.28 0.61 -0.031 0.11 -0.22 -0.061 -0.058 0.15
N Stage -0.032 0.013 0.28 1 0.88 -0.036 0.15 -0.26 -0.1 -0.094 0.26
6th Stage -0.025 0.003 0.61 0.88 1 -0.042 0.17 -0.29 -0.11 -0.1 0.26
differentiate -0.019 0.032 -0.031 -0.036 -0.042 1 -0.37 -0.009 -0.022 0.009 -0.019
Grade -0.057 -0.0043 0.11 0.15 0.17 -0.37 1 -0.044 -0.19 -0.18 0.13
A Stage 0.007 0.0049 -0.22 -0.26 -0.29 -0.009 -0.044 1 0.066 0.027 -0.097
Estrogen Status 0.058 -0.018 -0.061 -0.1 -0.11 -0.022 -0.19 0.066 1 0.51 -0.18
Progesterone Status 0.034 -0.013 -0.058 -0.094 -0.1 0.009 -0.18 0.027 0.51 1 -0.18
Status -0.05 0.033 0.15 0.26 0.26 -0.019 0.13 -0.097 -0.18 -0.18 1
Race
Marital Status
T Stage
N Stage
6th Stage
differentiate
Grade
A Stage
Estrogen Status
Progesterone Status
Status
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Correlation Btween Numerical Fatures
Age 1 -0.077 -0.033 0.013 -0.0094
Tumor Size -0.077 1 0.1 0.24 -0.087
Regional Node Examined -0.033 0.1 1 0.41 -0.022
Reginol Node Positive 0.013 0.24 0.41 1 -0.14
Survival Months -0.0094 -0.087 -0.022 -0.14 1
Age
Tumor Size
Regional Node Examined
Reginol Node Positive
Survival Months
1.0
0.8
0.6
0.4
0.2
0.0

Modeling 5

Building the Linear Regression model testing the predictions of the tumor size.

Finding the best hyperparameters with Bayes Search CV.

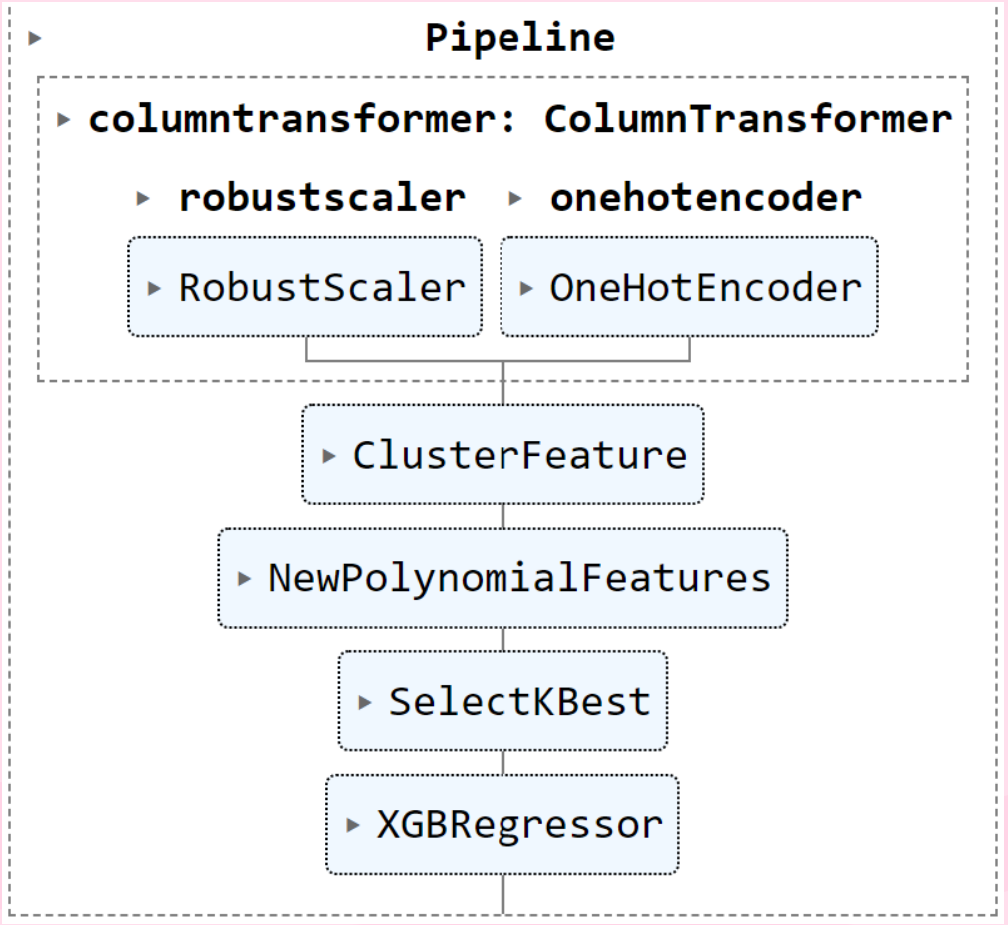
Pipeline

~ Bayes Search CV

~ XGB Pipeline

~ MLP Pipeline

This print the pipeline with the best parameters.



XGBoost Performed Best

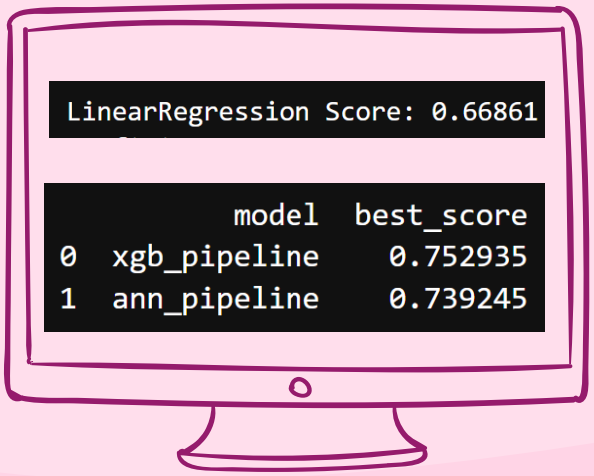
- The XGBoost model has achieved a score of 75.3%.
- The R2 score of 75.7% of the variance in the target variable.
- The root mean squared error (RMSE) average difference between the predicted tumor size and the actual tumor size is 10.3mm.
- The mean absolute error (MAE) score that the model's predictions are off by 6.88mm.

6 Evaluation

Metrics:

R2, MSE, MAE

Evaluation Metric



- **The evaluation metric use on the models were Mean Squared Error (MSE), Mean Absolute Error (MAE) and R-Squared (R2).**
- **Based on our evaluation metrics, the XGBoost model performed the best out of the three pipelines we compared.**

7 Results

What performance could our model predict?

Patient Medical History

JANE DATA (Row 24)

55	3
White	Regional
Single	Postive
T2	Positive
N1	4
IIIB	1
Poorly Differentiated	78

Status of Patient

Alive

29

Patient tumor size

The Test Model on Sample Data Result

- Test the model on the sample data
- Retrain data on entire dataset
- Testing the model on sample data and print results

```
real tumor size = 29mm
predicted tumor size = 32.87976mm
```

```
real tumor size = 29mm
predicted tumor size = 32.88052mm
```

Conclusion 8

Conclusion

- **The conclusion is that the data utilized for conducting exploratory data analysis could be a valuable tool in predicting tumor size and aiding in medical decision-making.**

- **XGBoost model is a good fit for the data and can be used to predict tumor size with fairly high R^2 score. Although, the MAE seems a bit high to be off the mark by almost 7mm.**

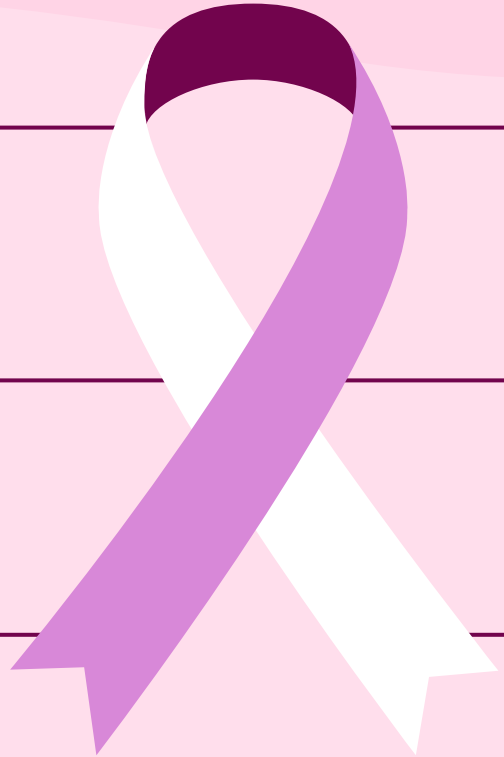
- **The part of the project I enjoyed the most was exploring and analyzing the data through EDA. This allowed me to gain a deeper understanding of the relationship between the different features and the target variable.**

Prevention

Self-exam: Women can self-exam at home to check for lumps or any changes

Regular check-ups: With a healthcare provider, perform clinical breast exams and mammograms

Healthy lifestyle: Maintain healthy lifestyle with regular exercise and healthy diets



Genetic testing: Women with history of breast cancer may benefit from genetic testing to determine the risk

Breastfeeding: can help reduce the risk of breast cancer

Early detection: key to successful treatment, be aware of changes in the breast & report to healthcare provider

“THERE IS NO LIMIT TO WHAT WE, AS WOMEN, CAN ACCOMPLISH. WE CAN DO ANYTHING AND EVERYTHING.”

—Michelle Obama

References

SEER Breast Cancer Data
https://ieee-dataport.org/open-access/seer-breast-cancer-data

Data Sets – Cancer Bioinformatics and Data Science
https://cancer.ubrite.org/data-sets/

American Cancer Society
https://www.cancer.org/cancer/types/breast-cancer/about/what-is-breast-cancer.html

Breastcancer.org
https://www.breastcancer.org/about-breast-cancer#section-breast-cancer-facts-and-statistics

Mayo Clinic
https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc20352470#dialogId55498219

Breast Cancer Foundation
https://www.komen.org/