ISYE 6740 – Fall 2024
Final Report

Team Member Names: Chengxi Zhang, Lin Xu, Zhanxu Liu

Team project number: 66

Project Title: WiDS Datathon: Predicting timely treatment for metastatic cancer

Partition of roles:

Zhanxu Liu: proposal drafting, model development, final report drafting

Chengxi Zhang: proposal drafting, data preparation, EDA, final report drafting

Lin Xu: proposal drafting, NLP implementation, hyperparameter tuning, final report drafting

Table of Contents

Problem Statement

Metastatic triple-negative breast cancer (TNBC) represents an aggressive subtype of breast cancer characterized by its dissemination beyond the breast and regional lymph nodes [1]. Distinguished from other forms of invasive breast cancer, TNBC exhibits a more rapid growth and metastatic potential, alongside a limited array of treatment options and a poorer prognosis [2].

The "triple-negative" designation refers to the absence of three critical markers: estrogen receptors (ER), progesterone receptors (PR), and human epidermal growth factor receptor 2 (HER2) [1-4]. Upon confirmation of breast cancer via imaging and biopsy, these markers are routinely evaluated. If cancer cells lack ER, PR, and HER2 expression, the diagnosis is classified as TNBC. This subtype accounts for approximately 10-15% of all breast cancer cases [5].

The therapeutic management of metastatic TNBC is notably challenging due to the scarcity of targeted treatment options, leaving chemotherapy as the mainstay therapy. However, ongoing research into novel therapeutic strategies, including immunotherapy and targeted treatments, holds promise for improving outcomes in patients with metastatic disease [1]. Importantly, timely initiation of treatment following diagnosis is critical, as delays in diagnosis and subsequent therapy can have severe implications for this highly aggressive cancer. The disparities in treatment waiting times serve as a valuable indicator of inequities in healthcare access.

The aim of this study is to evaluate whether patients received a diagnosis of metastatic cancer within 90 days of their initial screening. This investigation is guided by two primary objectives:

1. To conduct a comprehensive data analysis that elucidates the relationship between patient demographics and the likelihood of receiving timely treatment.
2. To explore the secondary objective of assessing the impact of environmental hazards on proper diagnostic and treatment.

Data source

The data utilized in this study was sourced from Kaggle: The Women in Data Science (WiDS) Worldwide Datathon 2024 Challenge[6]. The dataset for this challenge consists of real-world healthcare data, spanning from 2015 to 2018, and includes detailed information on patient demographics, diagnosis, treatment options, and insurance status for individuals diagnosed with breast cancer. The data was enriched with socio economics features and air quality data that may contribute to health equity. The data contains a binary response variable, indicating if the patient was diagnosed within 90 days (1 for yes, and 0 for no).

To complement the patient zip data, we acquired the latitude and longitude data from the US Census Bureau [7] to better visualize the distribution of timely diagnosis.

Methodology

Exploratory Data Analysis

The data set contains 12906 records of patient data with 8060 records shown to have diagnosed cancer within 90 days. This is a relatively balanced dataset (response variable in 2:1 ratio) which means additional oversampling techniques could be performed to ensure equity of majority and minority classes if necessary. There are in total 82 features provided by the data

set, initial inspection of data reveals that there are some columns that are missing more than 50% of information required (Table 1). We assessed that it is necessary to drop these columns as imputation methods are unlikely to provide insights in such circumstances. Namely they include: metastatic_first_novel_treatment, metastatic_first_novel_treatment_type, bmi, patient_race.

Fields	Missing Values	Percentage Missing
metastatic_first_novel_treatment	12882	99.814040
metastatic_first_novel_treatment_type	12882	99.814040
bmi	8965	69.463815
patient_race	6385	49.473113
payer_type	1803	13.970246
Region	52	0.402913
Division	52	0.402913
patient_state	51	0.395165
N02	29	0.224702
PM25	29	0.224702

Table 1 Missing value analysis. The table ranked top 10 percentage of missing values of the variables.

In addition, as all detected cases from the training dataset are for females only. We have thus dropped the variable 'patient_gender' as it does not add any further value.

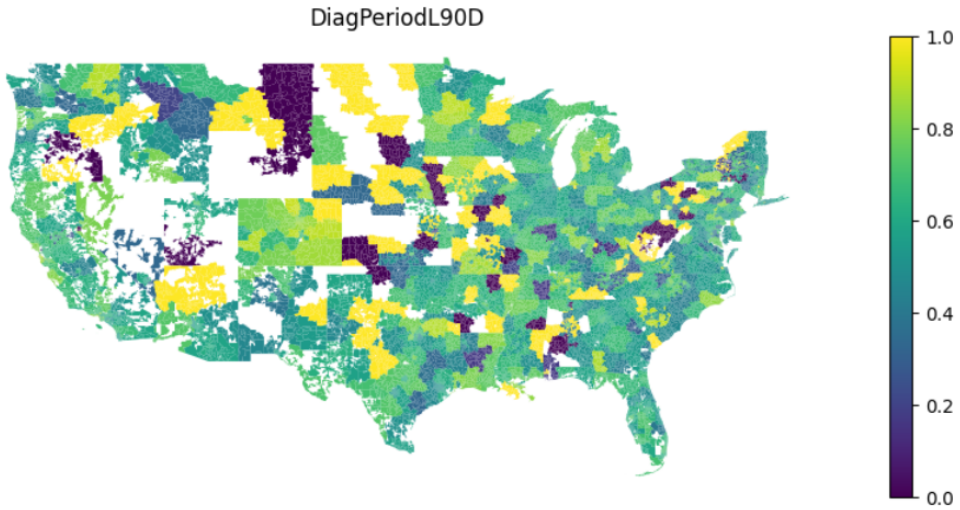
Imputation for Unknown Values

Both categorical and numerical variables with missing values were addressed using specific imputation techniques. For numerical variables, the median value of each variable was used to fill in missing data. For categorical variables, such as 'patient_state' (indicating the state of residence on the metastatic date), the most frequent state value was imputed based on grouped supplemental enrichment data derived from the corresponding ZIP codes. This approach was similarly applied to other categorical variables, such as 'region' and 'division'. For categorical variables like 'payer_type', where no clear value could be inferred from supplemental data, the missing entries were labeled as "unknown" to represent this subset. We also examined each data point (row) individually and excluded those with more than three missing entries from the analysis. These excluded rows accounted for only 0.08% of the total dataset.

Visualization of Geographical Distributions

Geographical analysis of timely triple-negative breast cancer (TNBC) diagnoses was performed by integrating patient data with latitude and longitude information from the U.S. Census Bureau. The ratio of timely diagnoses was calculated for each three-digit ZIP code region (patient_zip3) to evaluate regional patterns. Mapping these ratios revealed notable patterns, including higher rates of timely diagnoses in the northeastern U.S. and the West Coast (Figure 1A). Additionally, certain states, such as Colorado, North Dakota, and Minnesota, exhibited elevated rates of timely diagnoses compared to some regions in Central.

A



B

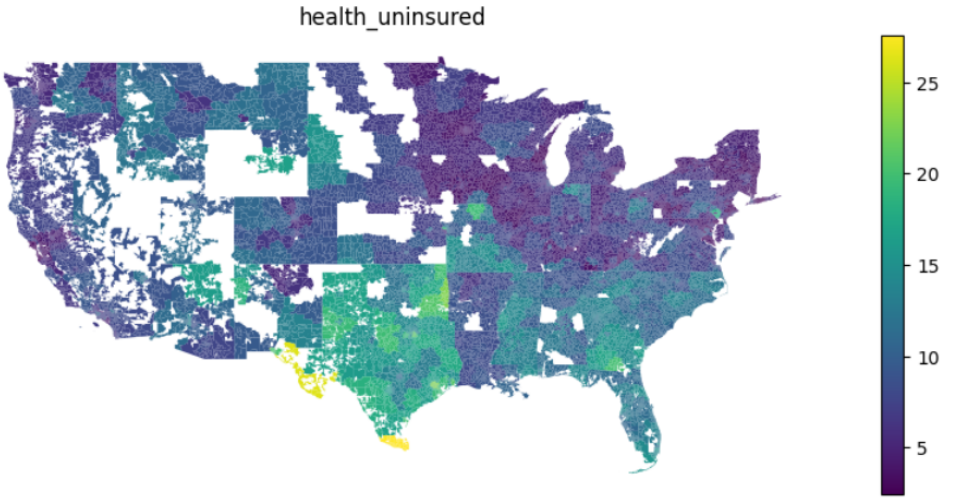


Figure 1 Geographic distribution of (A) Timely TNBC diagnosis rate by 3-digit zip prefix in the US. (B) Rate of health uninsured by 3-digit zip prefix in the US.

Using the health uninsured data based on ZIP Code, we can see that the above phenomenon does not always correspond to the availability of health insurance as only Southwestern and Central South part of US are areas where both the late diagnosis and a higher percentage of health uninsured coincide, and does not look to be conclusive for other areas of late diagnosis (Figure 1B).

Based on the observed geographic patterns, the variable ‘division’ was selected to represent the geographic locations of the reported cases. To reduce redundancy and minimize correlation between variables, 'patient_zip3', 'patient_state', and 'region' were excluded from further analysis at this stage.

Feature engineering

Natural Language Processing

One of the variables contains relevant information about diagnosis description for breast cancer. However, the entries are not all written in uniformity, with some abbreviations frequently used in the text. The dataset includes 50 unique breast cancer diagnosis codes, many of which have similar wordings. To reduce dimensionality, we applied a natural language processing technique called TF-IDF (Term Frequency-Inverse Document Frequency) as follows:

1. Split and convert phrases in each record into arrays of words and turn words used into common forms using manual stemming/lemmatization method (e.g., converting from 'malig' to 'malignant' and 'unsp' to 'unspecified'.
2. Remove any common stop words in English such as "and", "the" from the NLTK library.
3. Map arrays of words into TF-IDF scores, which quantify the relative importance of terms within the overall text corpus.

The principle behind the TF-IDF process is that it converts the frequency of a given word/phrase into term frequency-inverse document frequency values. It tries to evaluate the importance of a word in a document relative to a collection of documents (in our case, lines of texts). In doing so, words that appear too common such as stop words are ignored.

1. Term Frequency measures how often a word would appear in a document:

$$TF(t,d) = \frac{Number\ of\ times\ word\ t\ appears\ in\ document\ d}{Total\ terms\ in\ document\ d}$$

2. Inverse Document Frequency (IDF) measures how unique a word is across all given documents (lines of texts)

$$IDF(t) = \log\left(\frac{Total\ number\ of\ documents\ N}{1 + Number\ of\ documents\ containing\ word\ t}\right)$$

3. TF-IDF score combined TF and IDF to give each word/phrase a weight

$$TF - IDF(t,d) = TF(t,d) \times IDF(t)$$

We have inspected the actual diagnosis text and constructed a custom vocabulary dictionary which better represent sensible word pairs in the given diagnosis text samples [8]. They include: 'lower-outer', 'upper-outer', 'upper-inner', 'lower-inner', 'central portion', 'nipple and areola', 'unspecified', 'right', 'left', 'overlapping', 'female breast', 'axillary tail', 'secondary', 'male_breast'.

These variables are further refined by the Chi-Squared test which looks to test whether the occurrence of a specific term and the occurrence of a specific class are independent. The test itself is based on comparison of observed and expected frequencies within a contingency table. More specifically, it will be calculating the Chi-square test statistics between each feature/term and the target and select the desired number of features with best Chi-square scores. Features that exhibit high dependencies with the target variable are considered important for prediction and will be selected as a result (Table 2).

This approach is particularly suitable for TF-IDF as the feature values are frequencies of terms, because the sum will be the total frequency of that term in that class.

Features	Chi-square	P-value (< 0.05)
right	298.274332	7.83E-67
left	267.250256	4.51E-60
unspecified	221.43955	4.39E-50
overlapping	117.447953	2.29E-27
female breast	68.726398	1.13E-16
lower-inner	11.824415	5.85E-04
secondary	10.122247	1.46E-03
upper-outer	4.494086	3.40E-02

Table 2 Chi-square scores for selective features, with P-value less than 0.05.

Dummy variable/encoding

Categorical variables with multiple categories work well only with models that are part of the tree family, such as boosting models or random forest models, in their original form. However, to make it compatible with other models, we need to generalize and create separate columns with Boolean values (one-hot encoding) for each category within these variables, such as 'payer type' and 'metastatic cancer diagnosis code'.

Feature Selection

We need to assess whether each feature has a significant impact on the outcome of interest, which involves evaluating the importance of each feature and potential correlation between variables.

In our case, based on our initial EDA analysis on the underlying data, we have chosen to empirically choose variables that better represent characteristics and hope to capture the representative socio-economic and environmental factors along with the cancer specific diagnosis variable, without loss of generality.

We have chosen to use the following variables for our analysis: 'patient type', 'patient age', 'breast cancer diagnosis description (Converted into TF-IDF scores)', 'metastatic cancer diagnosis code', and zip-code based demographic and environmental variables ('division', 'unemployment rate', 'median household income', 'percentage of people with disability', 'Percentage of people with limited English', 'percentage of people with no health insurance', 'annual ozone concentration', 'annual PM2.5 concentration', 'annual nitrogen dioxide (NO2)').

Machine Learning Modeling and Results

To build the machine learning models, we have used a variety of classification algorithms to evaluate their predictive performance on the binary outcome. The models include both linear and non-linear approaches, as well as ensemble methods, which allow for a comprehensive comparison of different techniques to identify the best-performing model for the given data.

We trained the following models on the training dataset, and tuned hyperparameters for optimal performance: Logistic Regression, XGBoost, AdaBoost, LightGBM, Random Forest, Support Vector Machine, and K-Nearest Neighbors.

Logistic Regression

The logistic regression model predicts the probability of a binary outcome Y given a set of predictors $X = [x_1, x_2, \dots, x_p]$ The formula is:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p)}}$$

Where,

β_0 is the intercept term.

$\beta_1, \beta_2, \dots, \beta_p$ are the coefficients for predictors $x_1, x_2, \dots, x_p$.

AdaBoost

Adaptive Boosting is an ensemble learning method that combines the outputs of multiple weak learners to create a strong learner. It focuses on misclassified examples, adjusting their weights to improve the next model's performance. The final prediction is a weighted vote of all weak learners.

Construct weights over T periods:

$$D_t : t = 1, \dots, T$$

Initialize weights:

$$D_1(i) = \frac{1}{m},\ i = 1, \dots, m$$

Compute the weighted error and the model weight:

$$\epsilon_t = \sum_{i=1}^{m} D_t(i)\mathbb{I}\{y^i \neq h_t(x^i)\}, \quad \alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

Update weights for data:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} e^{-\alpha_t y^i h_t(x^i)} = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } y^i = h_t(x^i) \\ e^{\alpha_t} & \text{otherwise} \end{cases}$$

where Z_t is the normalizing constant:

$$Z_t = \sum_{i=1}^{m} D_t(i) e^{-\alpha_t y^i h_t(x^i)}$$

Final classifier:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to make more accurate predictions. It can be used for both classification and regression tasks.

For b = 1, ... , B

Draw bootstrap sample Z_1 of size N from training data.

Grow a random forest tree T_1 for the boostraped data by,

- Select v variables at random from p variables
- Pick the best variable/split among the v
- Split the node into two children's nodes

Output ensemble of trees T_1 , b = 1, ... B

To make prediction for a new point x:

Regression:

$$\frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

Classification: majority vote of

$$\{T_b(x)\}, b = 1, \ldots, B$$

K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a simple and non-parametric machine learning algorithm used for classification and regression. It works by finding the k-nearest data points (neighbors) to a given query point and making predictions based on their values. In other words, it assigns x a label by taking a majority vote over the K training points x_i closest to x.

The algorithm first calculates the distance between the query point x and all points in the dataset. The most common distance metric is Euclidean distance:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{p} (x_j - x_{ij})^2}$$

x is the query point.

x_i is a data point in the training dataset.

x_i and x_{ij} are the j-th features of x and x_i , respectively.

p is the number of features.

Identify the k-nearest points to the query point x based on the calculated distances.

Assign the class label to the query point based on a majority vote among the k-neighbors:

$$\hat{y} = \text{mode}(y_{i_1}, y_{i_2}, \ldots, y_{i_k})$$

XGBoost

Extreme Gradient Boosting is a powerful gradient boosting algorithm optimized for speed and performance. It builds an ensemble of decision trees iteratively, with each tree correcting the errors of its predecessors. XGBoost minimizes a regularized loss function to prevent overfitting, making it more robust than traditional gradient boosting methods.

LightGBM

Light Gradient Boosting Machine is a gradient boosting framework optimized for speed and efficiency. It builds decision trees sequentially, using a histogram-based algorithm to reduce

memory usage and improve training speed. Instead of splitting trees level-wise, LightGBM grows tree leaf-wise, focusing on leaves that contribute the most to reducing the loss function.

Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates data points of different classes in a high-dimensional space.

Evaluation Strategies

For the evaluation of the models, we utilized the Area Under the Receiver Operating Characteristic Curve (AUROC) as the primary performance metric. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings, providing a comprehensive view of the model's classification performance across different decision boundaries.

The following two tables present the findings from running the aforementioned models in their default form on a 5-fold cross validation set and also an 80%/20% train/test set.

Using 5-fold Cross Validation, we achieved our results in Table 3, which also contains the results of a single test with data split ratio of 4:1 for training and testing.

	Test		**Cross-validation**	
	Accuracy	**ROC-AUC**	**Mean ROC-AUC**	**Std Dev**
Lasso Logistic Regression	0.735761	0.761823	0.760022	0.011323
Ridge Logistic Regression	0.735761	0.761703	0.75985	0.011276
XGBoost	0.794653	0.787719	0.789485	0.005223
AdaBoost	0.793878	0.774999	0.764353	0.011247
LightGBM	0.80279	0.79206	0.796793	0.009777
Support Vector Machine	0.783417	0.780163	0.782191	0.012719
K-Nearest Neighbors	0.721813	0.717146	0.717503	0.011107
Random Forest	0.777218	0.770349	0.782584	0.009724

Table 3 Summarization of all performance metrics across all the models tested. The Boosting algorithms perform better than other models in general.

We analyzed the key features that contribute to the performance of the selected models, using the optimal metrics representative of each model type.

The default configuration of the **random forest model** enables us to evaluate the importance of each feature in contributing to the overall outcome. As shown in Figure 2, a baseline Random Forest model identifies ‘female breast’ and ‘unspecified’ from the breast cancer diagnosis dummy variables, and ‘patient age’ as the top three important feature among all the variables. This is followed by ‘right’, indicating the right breast, and some socio-economic factors

such as the unemployment rate, disabled rate, and percentage of individuals with health insurance, and environmental factors like ozone and NO2 levels.

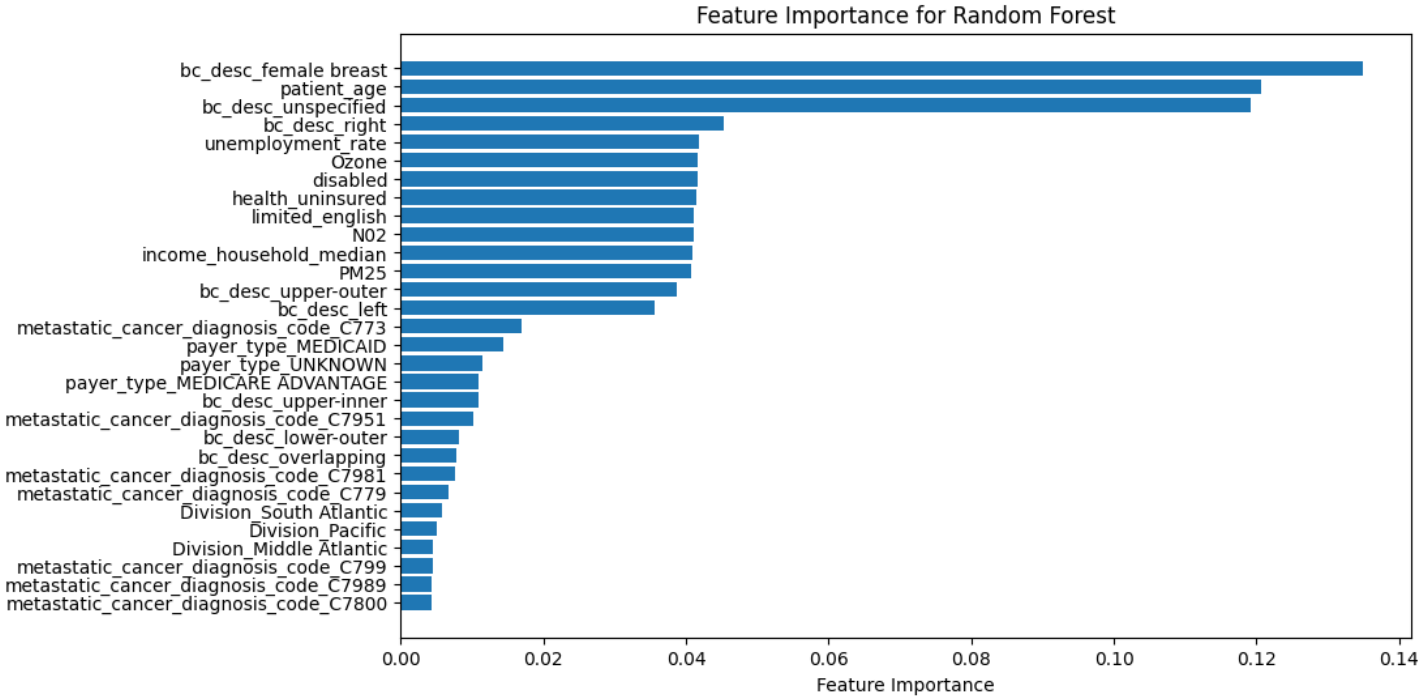


Figure 2 Feature importance for Random Forest.

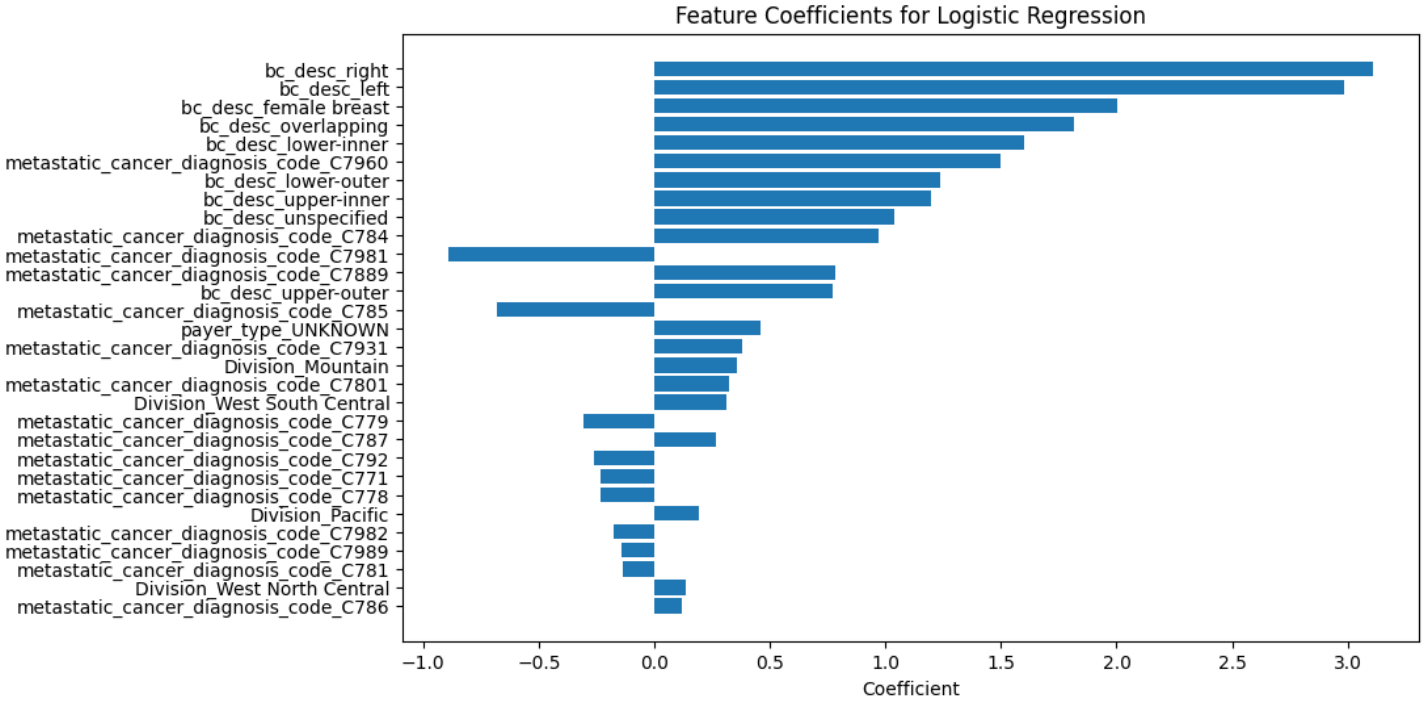


Figure 3 Feature coefficients for LASSO logistic regression.

In the **LASSO logistic regression**, based on the size of the feature coefficient, we can see that the top ranked coefficients are in fact dummy variables created from the curated list of keywords in the breast cancer diagnosis descriptions, followed by patient age, payer type and other variables (Figure 3).

Hyperparameter Tuning on Selective Models

Following the results above, we have decided to further fine-tune the LASSO logistic regression (as it deploys feature selection inherently based on its L1 regularization metric), the Random Forest and the LightGBM as it produces the best cross-validation and testing set accuracy and/or Area Under Curve.

	Cross-validation	
	Mean ROC-AUC	**Std Dev**
Lasso Logistic Regression	0.755232	0.010329
LightGBM	0.806732	0.008853
Random Forest	0.801827	0.009169

Table 4 Final performance metrics for selective models after tuning each of their hyperparameters.

LASSO logistic regression model

In the LASSO logistic regression model, the optimal parameter C was determined to be 0.0699, with a mean ROC-AUC of 0.7552 and a standard deviation of 0.0103 (Table 4).

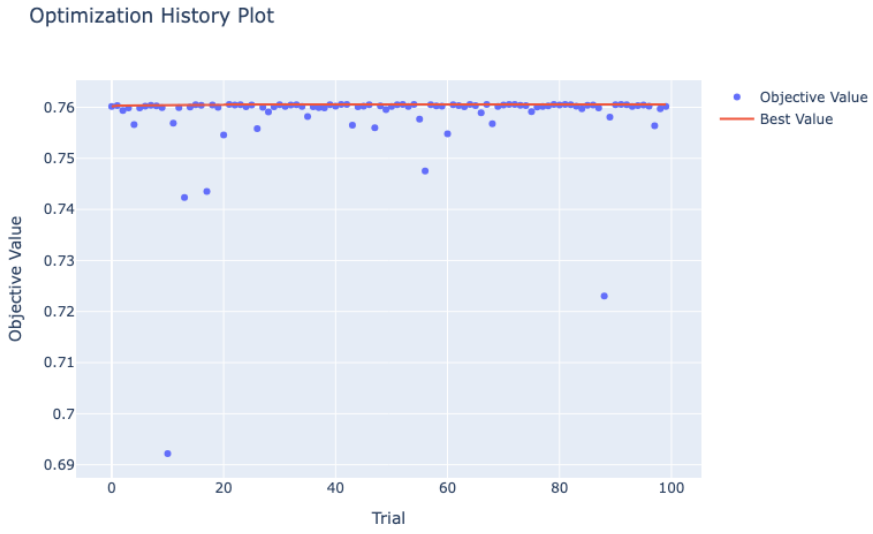


Figure 4 Hyperparameter optimization history plot of LASSO logistic regression.

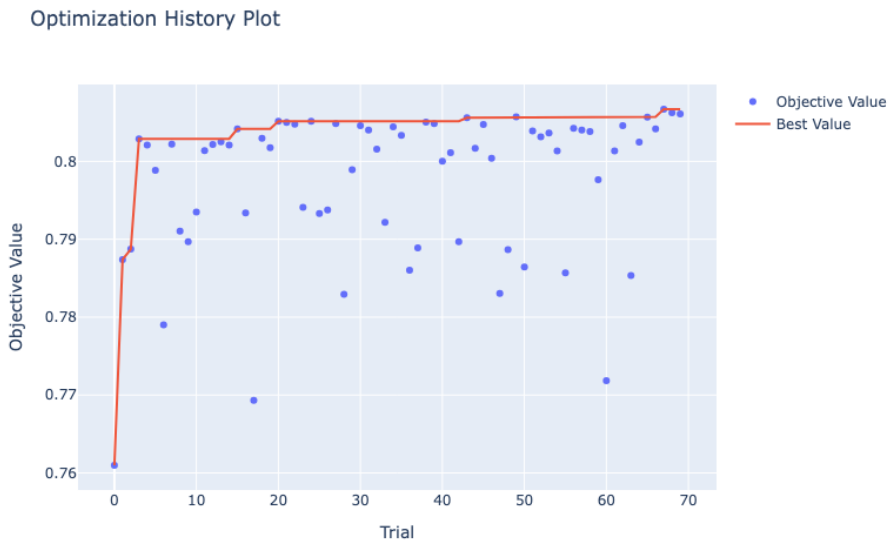
LightGBM

LightGBM is a variant of the Gradient Boosting algorithm that is meant to be optimized for speed and memory efficiency compared to other gradient boosting frameworks such as XGBoost.

It works in a similar way by minimizing the cost function using gradient descent, and each tree would improve on the residual errors from a previous tree.

The algorithm works by building trees by splitting the leaf with the largest loss reduction and is able to produce deeper trees with better accuracy with the same number of split.

A



B

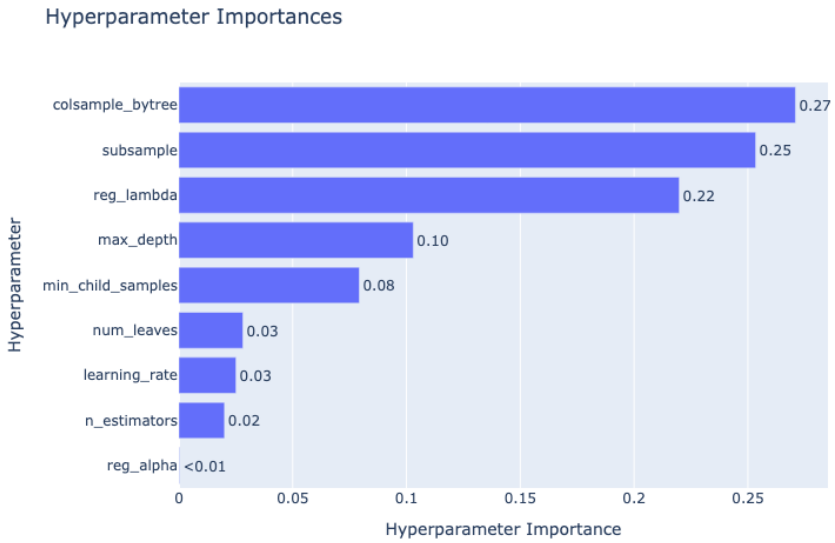


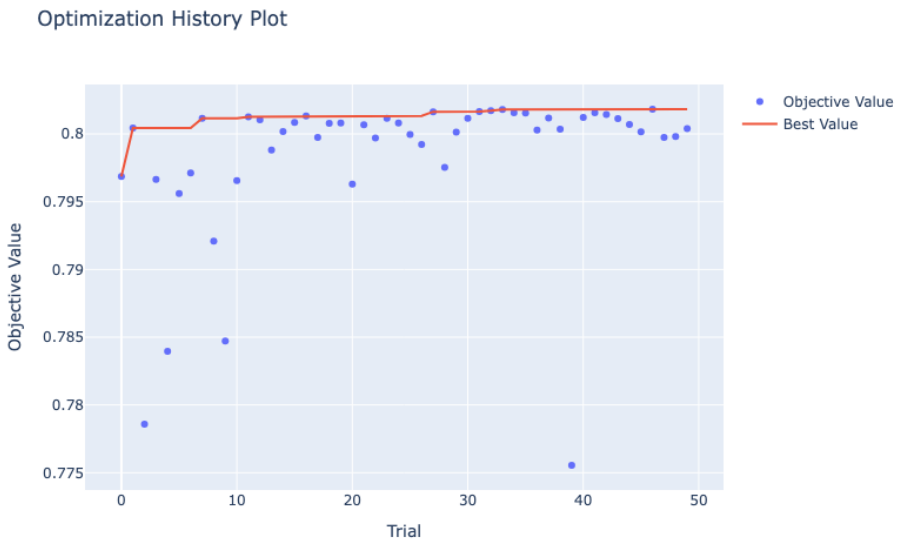
Figure 5 Hyperparameter tuning metrics for LightGBM. (A) Optimization progress chart (B) Importance of hyperparameters.

The optimized LightGBM model achieved its best performance with the following hyperparameters: 284 estimators, a learning rate of 0.0148, a maximum depth of 6, and 61

leaves. Additional parameters included a minimum of 8 child samples, a subsample ratio of 0.6823, and a column sampling ratio by tree of 0.6256. Regularization parameters were set at 1.2211 for α and 0.5027 for λ. These settings collectively contributed to the model's robust performance and efficient handling of the dataset (Figure 5).

Random Forest Model

A



B

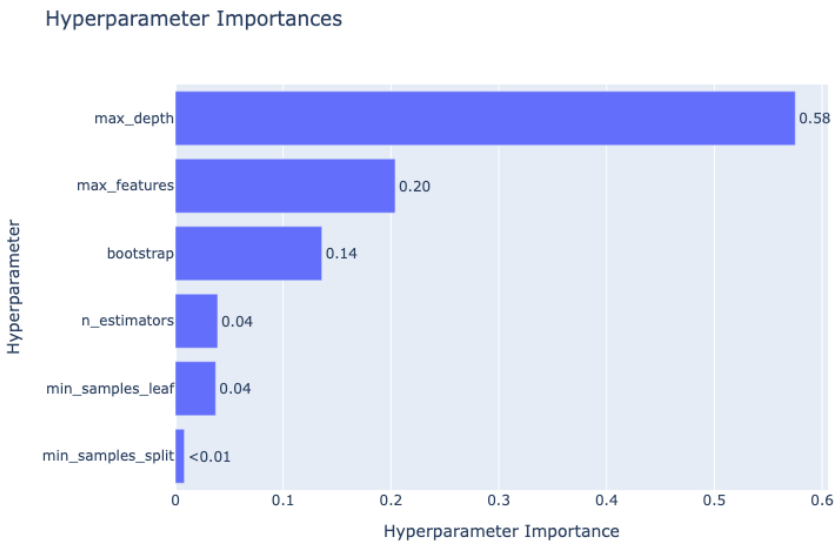


Figure 6 Hyperparameter tuning metrics for Random Forest. (A) Optimization progress chart (B) Importance of hyperparameters.

The optimized Random Forest model achieved its best performance with the following hyperparameters: 326 estimators, a maximum depth of 21, a minimum of 19 samples required to split a node, and a minimum of 14 samples required to form a leaf (Figure 6). The model used the "sqrt" method for selecting the maximum number of features and enabled bootstrapping for sampling. These hyperparameter settings ensured a balance between model complexity and generalization, resulting in strong predictive performance.

Conclusion and Discussion

This study aims to assess the timeline of Metastatic triple-negative breast cancer (TNBC) diagnosis within 90 days of the initial screening while exploring the influence of patient characteristics including demographics, socio-economic factors, and environmental variables on diagnosis outcomes. Our analysis revealed disparities in timely diagnoses, with specific diagnosis type, socio-economic and environmental factors contributing significantly to the observed patterns.

Among the tested models, the LightGBM model demonstrated the best performance, achieving the highest AUROC score, followed by the Random Forest model. The inclusion of TF-IDF-processed features and socio-environmental variables significantly enhanced the predictive capacity of these models. Key insights highlighted the importance of factors such as age, socio-economic indicators (e.g., unemployment rate), and environmental hazards (e.g., ozone and NO2 concentrations) in influencing diagnostic outcomes. These findings align with common expectations. For instance, older patients undergoing breast cancer screenings may be more likely to receive timely diagnoses due to prioritized healthcare access or awareness. Additionally, in highly polluted areas, healthcare providers might attribute certain symptoms to environmental factors rather than investigating them as potential signs of breast cancer, potentially delaying diagnosis.

Future improvements include first, addressing imbalanced outcome variables. Although the outcome variable in this study is relatively balanced with a 2:1 ratio between the two categories, future research could explore sampling techniques such as oversampling the minority class (e.g., SMOTE) or undersampling the majority class to ensure a more balanced dataset. This could make the predictive result more robust. Second, advanced feature engineering methods, such as feature interaction modeling or dimensionality reduction, could be applied to capture more complex relationships among variables. This could result in a better prediction accuracy and deeper insights into factors influencing the outcome.

Furthermore, collaboration with healthcare providers, professionals, and policymakers would ensure that the selected features are relevant and impactful, while the findings are actionable.

References

1. Derakhshan, Fatemeh, and Jorge S. Reis-Filho. "Pathogenesis of triple-negative breast cancer." Annual Review of Pathology: Mechanisms of Disease 17.1 (2022): 181-204.
2. Al-Mahmood, Sumayah, et al. "Metastatic and triple-negative breast cancer: challenges and treatment options." Drug delivery and translational research 8 (2018): 1483-1507.
3. Dass, Sylvia Annabel, et al. "Triple negative breast cancer: a review of present and future diagnostic modalities." Medicina 57.1 (2021): 62.
4. Foulkes, William D., Ian E. Smith, and Jorge S. Reis-Filho. "Triple-negative breast cancer." New England journal of medicine 363.20 (2010): 1938-1948.

5. Safran, Michal, et al. "Abstract P2-13-05: Triple Negative Breast Cancer (TNBC) patients are more likely to digitally explore clinical trial options and prior to receiving treatment for advanced disease compared to non-TNBC patients." Cancer Research 83.5_Supplement (2023): P2-13.
6. "WiDS Datathon 2024 Challenge #1", Kaggle, URL: https://www.kaggle.com/competitions/widsdatathon2024-challenge1/overview (last accessed on 10/20/2024).
7. "ZIP Code Tabulation Areas (ZCTAs)", US Census Bureau, URL: https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html (last accessed on 12/08/2024)
8. "Breast Cancer ICD-10 Code Reference Sheet", Ambry Genetics, URL: https://www.ambrygen.com/material/oncology/icd-10-code-reference-sheets/breast-cancer-icd-10-codes/630