Comparative Analysis of Linear Classification and Logistic Regression on Breast Cancer Dataset

Josh Kenn A. Viray

Department of Computer Science, College of Information and Computing Sciences, University of Santo Tomas
España Blvd, Sampaloc, Manila, 1008 Metro Manila, Philippines
joshkenn.viray.cics@ust.edu.ph

Abstract— **This study explores the comparative effectiveness of logistic and thresholded linear regression in binary classification tasks, specifically distinguishing between benign and malignant tumors in the Breast Cancer Wisconsin (Diagnostic) dataset. Feature standardization was applied to enhance model performance, and an 80:20 stratified train-test split ensured a balanced representation of classes. While linear regression, typically used for continuous prediction, was adapted for classification, its reliance on mean squared error led to suboptimal decision boundaries. Conversely, logistic regression, designed for classification tasks, employed the sigmoid function to map outputs between 0 and 1, optimizing the decision boundary through cross-entropy loss minimization. The study evaluates model behavior, feature importance, scalability, robustness, interpretability, and key performance metrics. Findings reveal that logistic regression provides superior classification accuracy (97%) and a higher F1-score (0.98), indicating a more balanced trade-off between precision and recall. While linear regression achieved higher recall, it suffered from increased false positives, making it less reliable for high-stakes medical diagnostics. The results highlight the importance of model selection in classification tasks, emphasizing that logistic regression is better suited for probabilistic classification. In contrast, linear regression remains an approximate alternative with inherent limitations.**

Keywords— **Machine Learning, Linear Regression, Logistic Regression, Feature Importance, Model Evaluation, Test Metrics**

I. Introduction

Breast cancer remains one of the most prevalent and life-threatening diseases worldwide, necessitating accurate and efficient diagnostic models. Machine learning techniques, particularly classification algorithms, have enhanced predictive accuracy in medical diagnostics. This study conducts a comparative analysis of two widely used classification methods—Linear Classification and Logistic Regression—on a breast cancer dataset provided under the sci-kit-learn import package that originated from Wolberg, Mangasarian, Street, and Street (1993), where they introduced the Breast Cancer Wisconsin (Diagnostic) dataset, which has been widely used in machine learning research. By evaluating their performance based on key metrics such as accuracy, precision, recall, and F1-score, this research aims to determine which method offers superior predictive capabilities. This comparative analysis is written as part of the partial requirements of the course Machine Learning.

II. Methodology

This formative task follows a structured approach to compare logistic regression and thresholded linear regression for binary classification in breast cancer diagnosis. The methodology consists of several key steps, including dataset description, preprocessing, exploratory data analysis, model implementation, evaluation, and comparative analysis.

A. Dataset

The WDBC dataset comprises multivariate, real-valued features that describe various morphological properties of cell nuclei in FNA images. These features include mean radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, along with their standard error and worst-case values.

TABLE I
Dataset Features and Description

Variable Name	*Role*	*Type*
ID	ID	Categorical
Diagnosis	Target	Categorical
radius (mean)	Feature	Continuous
texture (mean)	Feature	Continuous
perimeter (mean)	Feature	Continuous
smoothness (mean)	Feature	Continuous
compactness (mean)	Feature	Continuous
concavity (mean)	Feature	Continuous
concave points (mean)	Feature	Continuous
symmetry (mean)	Feature	Continuous
fractal dimension	Feature	Continuous

(mean)		
radius (standard error)	Feature	Continuous
texture (standard error)	Feature	Continuous
perimeter (standard error)	Feature	Continuous
smoothness (standard error)	Feature	Continuous
compactness (standard error)	Feature	Continuous
concavity (standard error)	Feature	Continuous
concave points (standard error)	Feature	Continuous
symmetry (standard error)	Feature	Continuous
fractal dimension (standard error)	Feature	Continuous
radius (worse)	Feature	Continuous
texture (worse)	Feature	Continuous
perimeter (worse)	Feature	Continuous
smoothness (worse)	Feature	Continuous
compactness (worse)	Feature	Continuous
concavity (worse)	Feature	Continuous
concave points (worse)	Feature	Continuous
symmetry (worse)	Feature	Continuous
fractal dimension (worse)	Feature	Continuous

B. *Exploratory Data Analysis*

To begin our analysis, we will initially utilize a heatmap to visualize the correlation among the features in the Breast Cancer Wisconsin (Diagnostic) dataset. A heatmap is a powerful tool for identifying relationships between variables, particularly in high-dimensional datasets, as it provides an intuitive way to observe patterns of association. This visualization is essential for feature selection, dimensionality reduction, and understanding multicollinearity within the dataset.

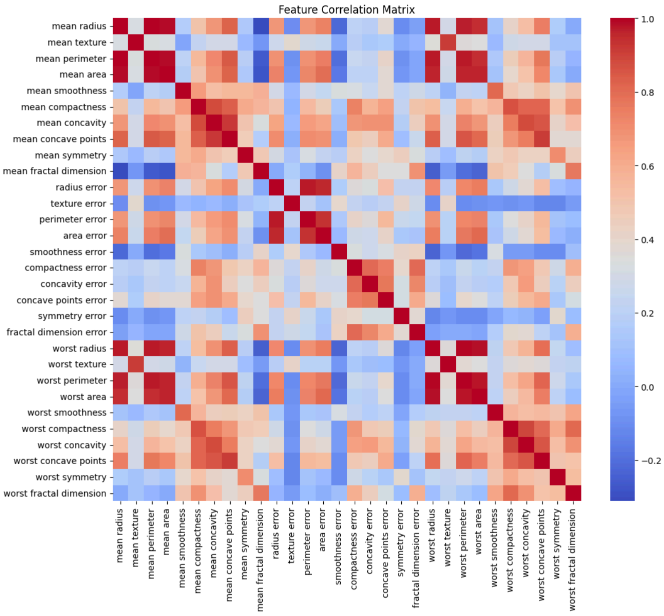

Fig. 1 Heatmap of all features' correlation from the dataset

The correlation matrices shown provide valuable insights into the relationships between different features in the dataset. The strong positive correlations observed among features like mean radius, mean perimeter, and mean area indicate that these measurements are closely related. This might suggest that these metrics increase as the tumor size increases and are greatly associated with one another.

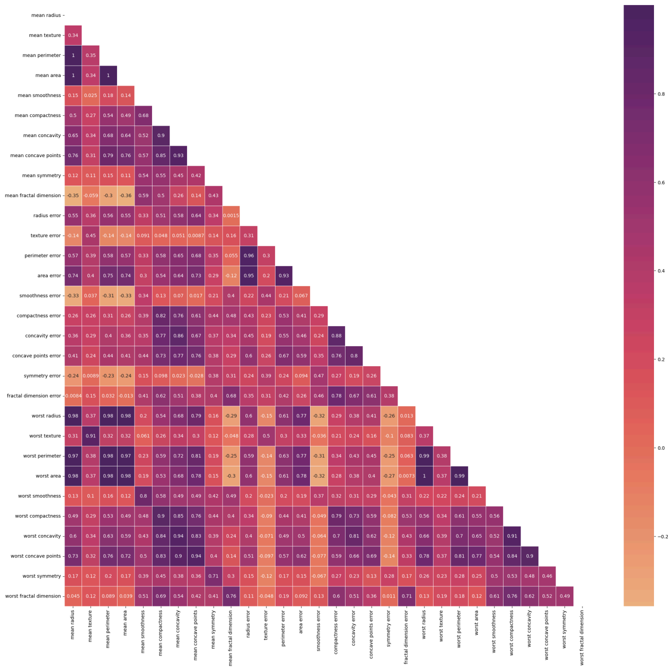

Fig. 2 Spearman heatmap of all features from the dataset

Similarly, worst radius, worst perimeter, and worst area also exhibit high correlation, reinforcing the idea that tumor size metrics are consistent across different conditions.

The heat maps also present negative correlations, particularly in features like mean fractal dimension, which show an inverse relationship between mean radius and mean area. This suggests that as tumor size increases, the complexity of its structure (as captured by the fractal dimension) tends to decrease. These insights can be important in understanding the biological behavior of tumors and selecting meaningful features for predictive modeling.

To further understand our features, we will be using violin plots to represent the distribution of numerical data per feature graphically. A violin plot serves a similar role to the traditional box plot. However, each violin is represented using a kernel density to estimate the underlying distribution.

The best features for classification in the given graphs were selected based on their ability to distinctly separate the two target classes (0 and 1). In classification problems, features that exhibit clear separation between classes contribute more significantly to the predictive performance of a model. The strip plots provided illustrate the distribution of feature values across the two classes, allowing for a visual assessment of which features provide the strongest distinction. Features with well-defined clusters or noticeable differences in distribution between the two target classes are likely to be more informative and, therefore, more useful in a classification model such as logistic regression.

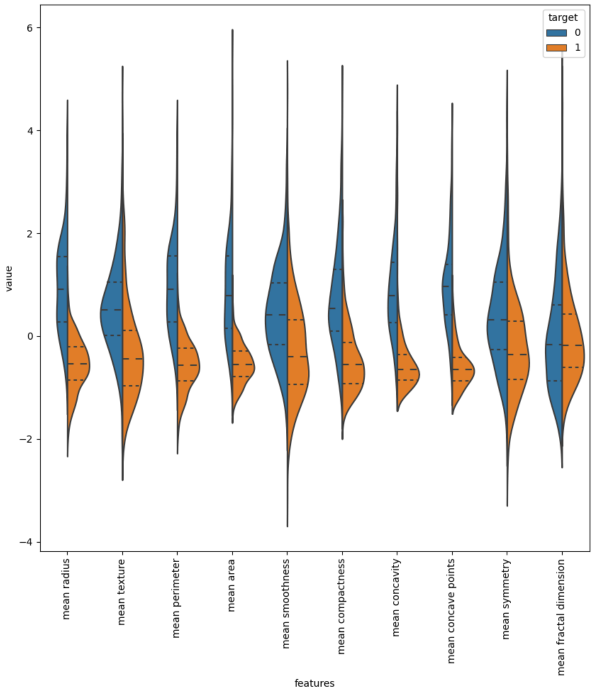

Fig. 3 Violin plot of mean values for radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and mean fractal dimension

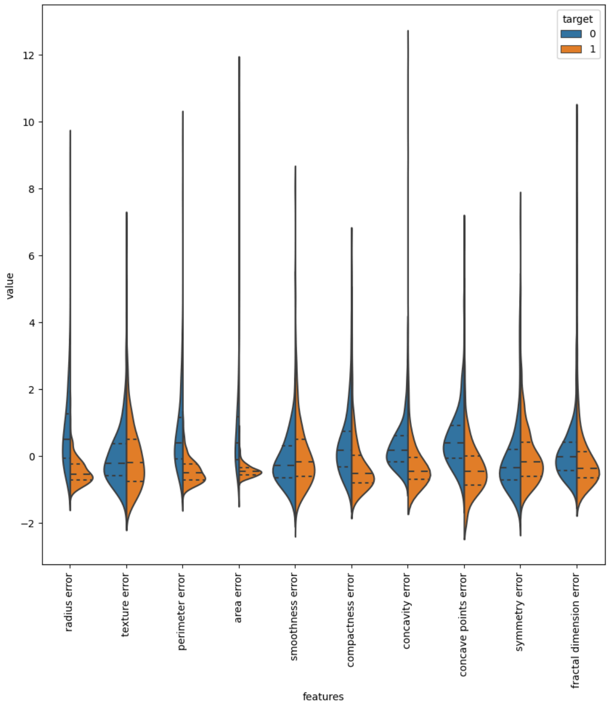

Fig. 4 Violin plot of error values for radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and mean fractal dimension

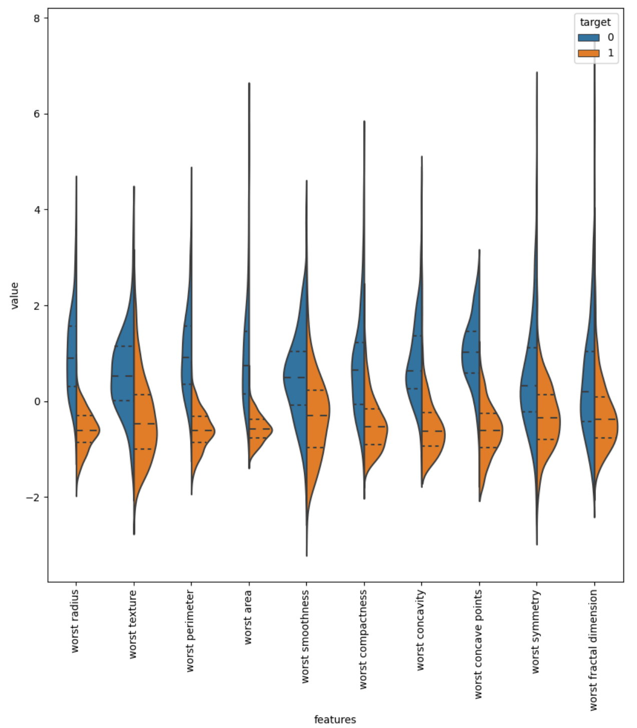

Fig. 5 Violin plot of worse values for radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and mean fractal dimension

Figures 3 to 5 show different feature names on the X-axis, while the Y-axis represents the scaled values of these features. The violin plot in each category represents the distribution of values, where wider regions indicate a higher frequency of data points, while thinner regions suggest fewer occurrences. The two different colors—blue for benign (0) and orange for malignant (1)—help differentiate the distributions for each class. The graph plot shows that malignant tumors (orange) tend to have higher feature values, particularly in features like radius, perimeter, and area. Some features clearly separate benign and malignant cases, making them more useful for classification. Meanwhile, other features have overlapping distributions, indicating they may be less effective in distinguishing between the two classes.

Further checking in Figure 5 shows that worse radius, worse perimeter, and worse area share a distinct graphic representation. Given their feature description, we can initially guess that they are greatly correlated with one another. From our initial discovery of this violin plot and heatmap, we should consider looking further into these three features by having a pair plot.

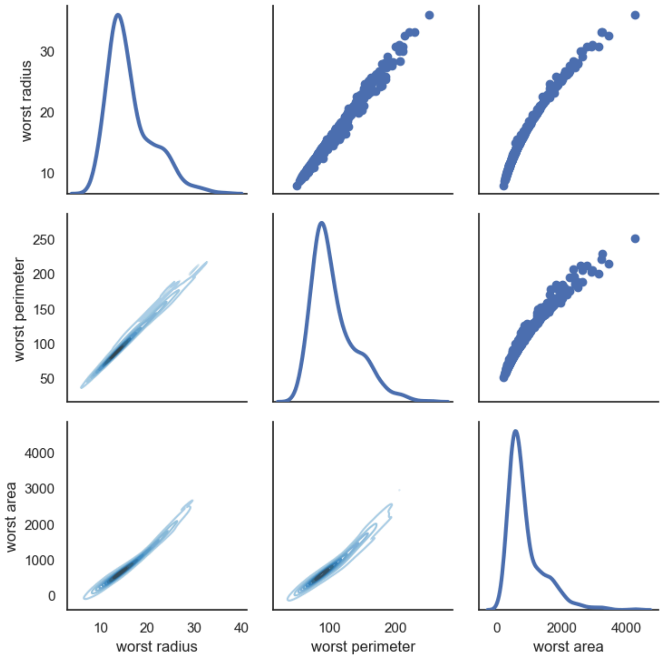

Fig. 6 Pair plot for worse radius, worse perimeter, and worse area

Given the problem statement that involves a lot of features, it would be great to visually represent the data points that do not overlap with each other. This is why we will then be applying a swarmplot to visualize the distribution of classes among all features while also having to check the appearance of outliers.

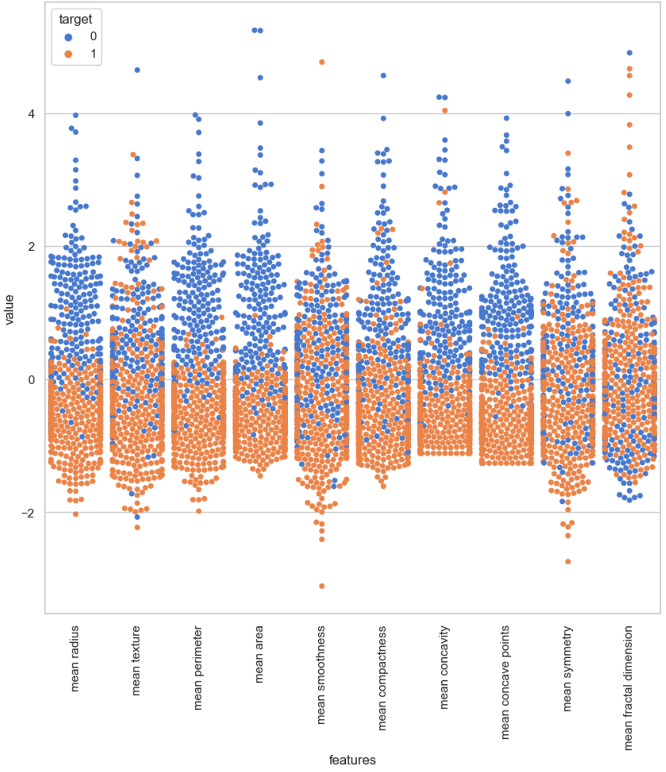

Fig. 7 Swarm plot of mean values for radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and mean fractal dimension

In Figure 7, which represents the mean values of various features, certain features such as mean radius, mean perimeter, mean area, and mean concave points demonstrate clear separability between the two classes. These features are valuable because their distributions show a distinguishable shift, meaning they have a strong correlation with the target variable. A well-separated distribution indicates that these features carry important information for distinguishing between the two classes, reducing the likelihood of overlap and misclassification.

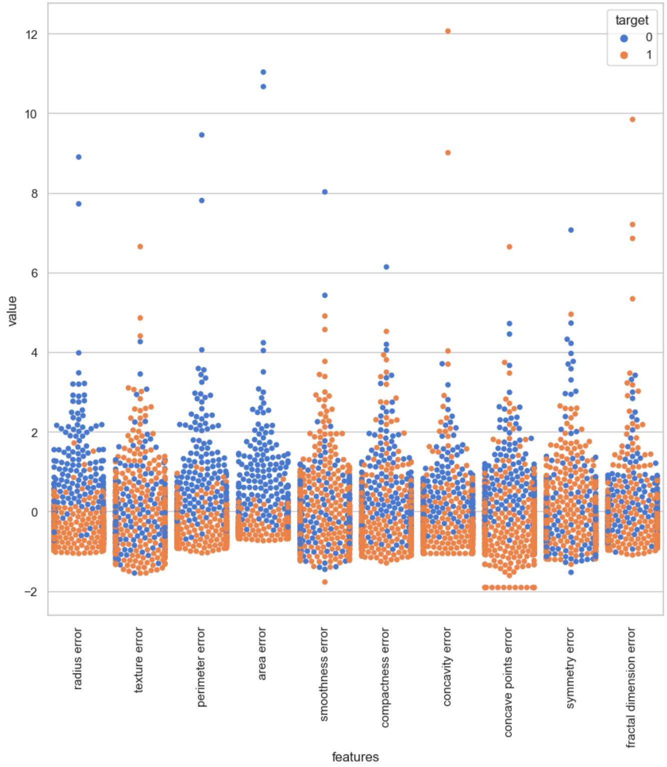

Fig. 8 Swarm plot of error values for radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and mean fractal dimension

In Figure 8, different features' standard deviation (error) displays less pronounced separation compared to the first graph. However, radius, perimeter, area, and concavity errors still exhibit moderate distinction between the two classes. Although these features may not be as strong as the mean features, they still provide additional variance-related information that could enhance classification performance. Standard deviation features may capture important variations that help refine the model's predictive ability.

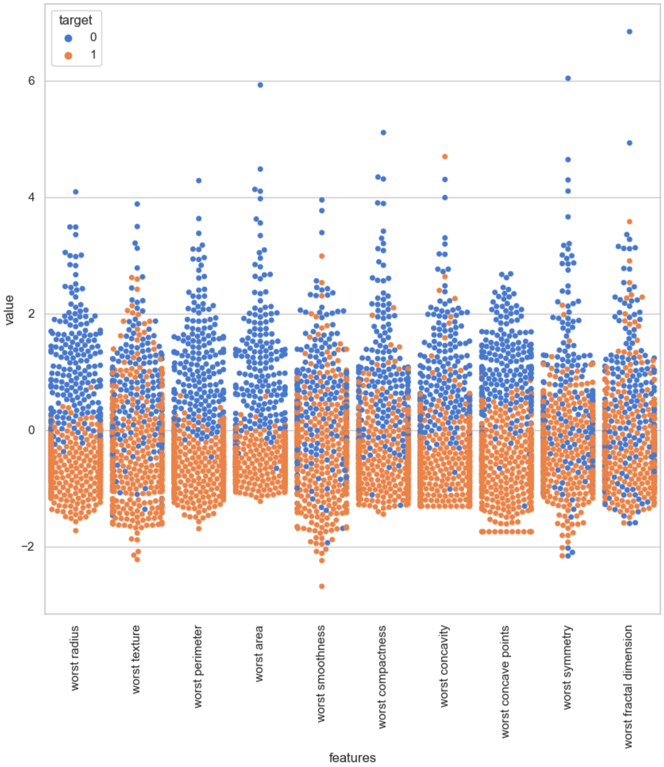

Fig. 9 Swarm plot of worse values for radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and mean fractal dimension

Figure 9, which depicts each feature's worst (maximum) values, provides another critical perspective on feature importance. Here, the worst radius, worst perimeter, worst area, and worst concave points exhibit significant class separation. These features provide some of the best discriminatory power because their distributions show a noticeable distinction between the two classes. The worst-case values of these features are particularly important as they capture extreme cases that may strongly indicate classification outcomes.

III. Methodology

Given this formative assessment task, we will provide a short overview of the method by which train and test data splits were applied and how the models were built.

A. Data Standardization

Feature standardization is a crucial preprocessing step in machine learning, particularly for models that rely on gradient-based optimization, such as logistic regression and linear regression (James et al., 2013). Standardization ensures that all features have a mean of zero and a standard deviation of one, thereby preventing features with larger magnitudes from dominating the learning process.

This step is performed using z-score normalization via StandardScaler() from the sci-kit-learn library.

This transformation helps models converge faster and improves interpretability, particularly in datasets where features have varying units or scales (Han, Kamber, & Pei, 2011).

B. Training and Testing Sets

To evaluate the generalizability of the models, the dataset was split into training and testing sets using stratified random sampling. This ensures that both classes (benign and malignant) are proportionally represented in both sets, reducing the risk of bias. The train-test split ratio was set to 80:20, a standard machine-learning practice (Hastie, Tibshirani, & Friedman, 2009). This ensures consistency in performance evaluation across models trained on both original and PCA-transformed features.

C. Linear Regression Model

While traditionally used for regression tasks, linear regression can be adapted for classification by thresholding the continuous outputs (James et al., 2013). The model estimates a linear relationship between independent variables and the target labels.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon \quad (1)$$

However, one limitation of linear regression in classification tasks is that it does not naturally constrain predictions between 0 and 1, leading to poor decision boundaries (Hosmer, Lemeshow, & Sturdivant, 2013). Despite this, it provides a baseline for comparison against more sophisticated models such as logistic regression.

D. Logistical Regression Model

Logistic regression is a widely used **classification algorithm** that models the probability of an instance belonging to a particular class. Unlike linear regression, logistic regression applies the **sigmoid function** to map outputs between **0 and 1**, allowing for probability-based classification (Menard, 2002). The logistic function is defined as:

$$P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}} \quad (2)$$

This makes logistic regression more suitable for binary classification tasks, such as distinguishing between benign and malignant tumors. Given its probabilistic nature, logistic regression can also handle imbalanced datasets better than linear regression, as it adjusts decision boundaries based on likelihood ratios (James et al., 2013).

IV. Analysis and Discussion

In this section, we will be evaluating the model–addressing their behavior, strengths and weaknesses, feature importance, scalability, robustness, interpretability, and evaluation metrics used–between linear regression and logistic regression through a series of graphic representations used in section II.b.

A. Model Behavior

When applied to high-dimensional data, logistic and linear regression (with thresholding) produce distinctly different decision boundaries due to their underlying mathematical formulations and optimization strategies. While logistic regression is inherently designed for classification tasks, linear regression requires post-processing adjustments to approximate classification, often leading to suboptimal decision boundaries.

Logistic regression establishes a non-probabilistic threshold to differentiate between classes. The decision boundary represents the point at which the model transitions from predicting one class to another, relying on the sigmoid function to map continuous input values into a probability distribution between 0 and 1. Unlike linear regression, logistic regression does not assume a linear relationship between input features and the target variable but instead focuses on modeling the log-odds of a data point belonging to a specific class.

Logistic regression ensures that probability distributions are well-aligned with the true class labels by minimizing cross-entropy loss, making it more robust in classification settings. The decision boundary in logistic regression typically appears as a sharp, well-defined separation between classes, especially in well-structured datasets. Given sufficient feature representation and minimal noise, logistic regression can achieve high classification accuracy.

Linear regression, on the other hand, is not naturally suited for classification tasks but can be adapted by applying a thresholding mechanism to its continuous numerical predictions. Instead of directly modeling probabilities, linear regression fits a linear function to the data, generating outputs that may extend beyond the [0,1] range. Using the sigmoid function, a threshold (e.g., 0.5) is applied to convert these outputs into discrete class labels, where values above the threshold correspond to one class, and values below it correspond to the other.

$$g(x) = \frac{1}{1+e^{-z}} \quad (3)$$

However, this approach introduces several challenges. Since linear regression minimizes the mean squared error (MSE) rather than cross-entropy loss, its decision boundary is often parallel contour lines rather than a well-defined separation. This makes the model more sensitive to outliers and extreme values, which can significantly distort predictions and lead to misclassification errors. This limitation becomes even more pronounced in high-dimensional datasets as linear regression fails to capture the inherent probabilistic nature of classification problems.

Ultimately, while logistic regression provides a mathematically principled and probabilistically sound approach to classification, thresholded linear regression remains a crude approximation that is prone to errors, especially in complex datasets with overlapping class distributions.

B. Feature Importance

In logistic regression, feature importance is primarily determined by examining the coefficients assigned to each feature in the model. These coefficients represent the impact of a given feature on the predicted outcome, with larger absolute values indicating stronger influence. Since logistic regression models the log odds of a classification decision, the exponentiation of these coefficients provides an odds ratio, which quantifies how much the probability of belonging to a certain class changes with a unit increase in the feature value.

Logistic regression has an inherent ability to handle irrelevant features by assigning near-zero coefficients to those that contribute minimally to the prediction. This built-in regularization effect allows the model to downweight unimportant variables, thereby reducing their influence on classification outcomes. However, if irrelevant features are numerous, they can still increase the dimensionality of the dataset, making the model more complex and harder to interpret.

As for the impact of irrelevant and highly correlated features on a model's performance, the presence of irrelevant or highly correlated features can significantly impact the interpretability and performance of both logistic and linear regression models. In the case of logistic regression, while the model can adjust the importance of certain features, an excessive number of irrelevant variables may still introduce noise, leading to unnecessary complexity and increased computation time.

The effects of multicollinearity (high correlation between independent variables) are particularly severe for linear regression. Since linear regression aims to assign precise weights to each feature, multicollinearity can lead to unstable predictions and inflated variance in the estimated coefficients. This instability arises because the model struggles to determine the individual contribution of correlated features, often leading to highly sensitive coefficient estimates that change drastically with small variations in data. This, in turn, reduces the model's generalizability and can lead to poor predictive performance on unseen data.

In order to mitigate these issues, techniques such as principal component analysis (PCA) or variance inflation factor (VIF) analysis can be used to detect and address multicollinearity. Additionally, dimensionality reduction methods and feature selection techniques help refine the model by eliminating redundant features, improving both accuracy and interpretability.

C. Scalability

When applied to significantly larger datasets with millions of samples, logistic regression and linear regression differ in their scalability and computational efficiency. Logistic regression, an iterative optimization-based method, typically relies on techniques such as gradient descent to converge to an optimal solution.

As dataset size increases, the computational cost of training logistic regression grows, mainly due to the need for multiple iterations over the dataset to optimize the loss function. In contrast, linear regression, which involves solving a system of linear equations, can be computationally expensive when using the standard equation method, especially when the feature matrix becomes large. However, stochastic gradient descent (SGD) can make both models more scalable by processing smaller batches of data at a time rather than the entire dataset.

Several strategies can be employed to optimize logistic regression for large-scale data. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), can help eliminate redundant or highly correlated features, thereby reducing the computational burden (Jolliffe & Cadima, 2016). Feature selection methods, such as L1 regularization (Lasso), can also aid in removing irrelevant features, leading to a more efficient and interpretable model (Tibshirani, 1996). Additionally, parallel and distributed computing frameworks, such as Apache Spark and TensorFlow, enable large-scale processing by distributing computations across multiple machines, significantly reducing training time (Abadi et al., 2016).

Another important approach is mini-batch gradient descent, which improves training efficiency by updating model parameters based on smaller subsets of data rather than the entire dataset at once (Bottou, 2010).

Furthermore, logistic regression benefits from online learning techniques, where the model updates dynamically as new data arrives instead of retraining from scratch. This is particularly useful in streaming data applications, such as real-time fraud detection or medical diagnostics (Bifet et al., 2010). Lastly, efficient data storage and retrieval mechanisms, such as using a compressed sparse row (CSR) format for sparse data or leveraging database indexing, can improve performance when dealing with large-scale datasets (Pedregosa et al., 2011). By implementing these strategies, logistic regression can remain a viable classification model even in high-dimensional, large-scale scenarios, ensuring both efficiency and accuracy in predictive modeling.

D. Robustness

Noise and outliers can significantly impact both logistic and linear regression, but their effects manifest differently. Linear regression is particularly vulnerable to outliers because it minimizes the mean squared error (MSE), which penalizes large deviations more heavily. As a result, extreme values can disproportionately influence the model's predictions, leading to biased coefficients and reduced generalizability. This issue is exacerbated in high-dimensional datasets, where outliers may distort the decision boundary and introduce instability in the model.

On the other hand, logistic regression is more robust to outliers than linear regression. Since logistic regression optimizes for cross-entropy loss rather than MSE, it focuses on correctly classifying data points rather than minimizing absolute prediction errors. However, if an outlier is positioned close to the decision boundary, it may still influence the classification decision. In extreme cases, highly imbalanced outliers can affect the gradient updates during training, leading to misclassification or overfitting. Preprocessing techniques such as Winsorization, trimming, robust scaling, or the use of Huber loss can be employed to mitigate the effects of outliers.

Effectively handling missing data begins with understanding the underlying mechanism that caused the data to be missing. It is essential to determine whether the missing values occur completely at random (MCAR), at random (MAR), or not at random (MNAR), as this distinction influences the appropriate strategy for managing them. Simple approaches such as imputation techniques are used when missing data is minimal. However, when a significant portion of data is missing, more sophisticated methods, such as multiple imputation or predictive modeling, may be necessary to ensure the integrity of the analysis.

Although listwise deletion may seem straightforward, it is generally discouraged unless the proportion of missing data is negligible, as it can lead to loss of valuable information and potential biases. Instead, thoughtful handling of missing data ensures that logistic and linear regression models remain robust, producing reliable insights without distorting relationships or reducing statistical power. By carefully selecting an appropriate approach, we can maintain data quality while minimizing the risk of misleading test results.

E. Interpretability

Interpretability is essential when selecting a machine learning model, especially in domains like healthcare, finance, and policy-making, where understanding the reasoning behind predictions is crucial. Both linear regression and logistic regression are considered interpretable models, but they differ in how they explain relationships between input features and outputs.

Linear regression is one of the most interpretable models because its coefficients directly represent the effect of each independent variable on the dependent variable. For example, if a coefficient is 2.5, it means that for every one-unit increase in that predictor variable, the target variable increases by 2.5, assuming all other variables remain constant. This straightforward interpretation makes linear regression a favored choice when it is essential to quantify how different factors influence an outcome.

While still interpretable, logistic regression does not allow for the same direct interpretation of coefficients. Instead of modeling direct numerical relationships, it models the log odds of an event occurring. This means that logistic regression coefficients tell us how the odds of an outcome change with a one-unit increase in a predictor. To make the results more understandable, we often exponentiate the coefficients to obtain odds ratios, which describe how much more (or less) likely an event is to occur based on a particular feature. While this transformation allows for interpretation, it is not as immediately intuitive as the coefficients in linear regression.

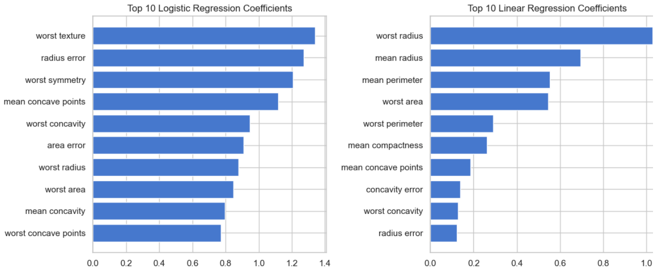

Fig. 10 Top 10 most influential features in Logistic Regression (left) and Linear Regression (right)

Ultimately, logistic regression provides a more straightforward interpretation of how features influence classification since it directly models the probability of malignancy. Linear regression is not inherently suited for classification tasks, as it assumes a continuous relationship with the target variable. However, it still identifies important features that correlate with cancer diagnosis.

The discrepancy in feature importance suggests that logistic regression is better for understanding classification, while linear regression may not fully capture the nonlinear relationships in the data.

When dealing with complex datasets, there is often a trade-off between interpretability and predictive power. More advanced models, such as deep learning and ensemble methods like random forests and gradient boosting, can capture intricate patterns in data, but they do so at the cost of transparency.

Logistic and linear regression strike a balance between simplicity and accuracy. They provide valuable insights into feature importance, allow for hypothesis testing, and are computationally efficient. However, they may struggle with nonlinear relationships or complex feature interaction datasets.

F. Metrics

The evaluation comparison between logistic regression and thresholded linear regression on the Breast Cancer Wisconsin (Diagnostic) dataset highlights critical differences in their classification effectiveness. While both models yield similar Area Under the Curve (AUC) scores, their performance across other evaluation metrics underscores their respective strengths and limitations.

1.) *Accuracy*: Logistic regression achieves an accuracy of 97%, whereas linear regression (after applying a classification threshold) attains 93% accuracy. The marginal difference suggests that both models perform well in classifying most instances correctly. However, accuracy alone is an insufficient metric, as it does not account for class imbalances or the trade-off between precision and recall, which are crucial in medical diagnostics.

2.) *Precision and Recall*: Logistic regression demonstrates a higher precision (0.96) than linear regression (0.90), producing fewer false positives. This is particularly important in medical diagnosis, where reducing false positives minimizes unnecessary psychological distress and additional medical procedures.

Conversely, linear regression exhibits a higher recall (1.00) compared to logistic regression (0.96), meaning it correctly identifies all malignant cases. However, this comes at the cost of increased false positives, potentially leading to overdiagnosis and unnecessary interventions. The trade-off between precision and recall highlights the strengths and weaknesses of each model. While logistic regression maintains a better balance, linear regression leans towards higher sensitivity at the expense of specificity.

3.) *F1 Score*: The F1 score, representing the harmonic mean of precision and recall, is 0.98 for logistic regression and 0.95 for linear regression. The slightly higher F1 score of logistic regression reinforces its reliability for classification tasks, as it effectively balances minimizing false positives and false negatives.

4.) AUC Score: Both models achieve an AUC score of 1.00, suggesting excellent performance in distinguishing between benign and malignant cases across various probability thresholds. This demonstrates their strong discriminatory power despite differences in how they model the data.

5.) Confusion Matrix: Examining the confusion matrix provides deeper insights into each model's classification patterns. Logistic regression, with its higher precision (0.96), minimizes false positives, making it a more trustworthy choice in high-risk medical settings. However, its recall of 0.96 means a small fraction of malignant cases could still be misclassified.

In contrast, thresholded linear regression achieves a recall of 1.00, meaning it correctly identifies all malignant cases but at the cost of increased false positives (lower precision of 0.90). This overclassification may lead to unnecessary follow-ups and increased anxiety for patients diagnosed with benign tumors.

While both models exhibit strong classification performance, **logistic regression emerges as the superior choice** for this task due to its well-balanced trade-off between **precision and recall**. The model's design specifically for **classification problems** allows it to provide more reliable predictions than **linear regression**, which inherently assumes a continuous target variable. The results underscore the importance of selecting appropriate models based on the application's specific needs—where precision and reliability are critical, logistic regression is preferable, whereas linear regression, when thresholded, may serve as an alternative in scenarios requiring heightened sensitivity.

While both models exhibit strong classification performance, logistic regression is the superior choice for this task due to its well-balanced trade-off between precision and recall. The model's design specifically for classification problems allows it to provide more reliable predictions than linear regression, which inherently assumes a continuous target variable. The results underscore the importance of selecting appropriate models based on the application's specific needs—where precision and reliability are critical, logistic regression is preferable, whereas linear regression, when thresholded, may serve as an alternative in scenarios requiring heightened sensitivity.

V. Conclusions

This formative assessment underscores the fundamental differences between logistic and linear regression in classification contexts, particularly in medical diagnostics, where accuracy and reliability are critical. Logistic regression's ability to model class probabilities through the sigmoid function provides a more structured approach to classification, ensuring robust decision boundaries and improved interpretability of feature importance. In contrast, thresholded linear regression, though capable of approximating classification, lacks a probabilistic framework, resulting in weaker decision boundaries and higher misclassification rates.

Additionally, the study highlights the impact of feature selection and standardization in enhancing model performance. While logistic regression demonstrated resilience to outliers and multicollinearity, linear regression's sensitivity to these factors led to more significant prediction variability. Scalability was another key consideration, with logistic regression's reliance on iterative optimization requiring additional computational resources compared to the direct equation-solving approach of linear regression. However, dimensionality reduction and mini-batch gradient descent can improve efficiency in large-scale applications.

Ultimately, the findings reaffirm that logistic regression is the preferred choice for binary classification tasks due to its well-defined probabilistic interpretation, ability to balance precision and recall, and superior handling of complex decision boundaries compared to linear regression in the case of medical diagnostics, where it is not probabilistic but definitive.

References

[1] A. Bifet, G. Holmes, B. Pfahringer, and R. Gavaldà, "Leveraging Bagging for Evolving Data Streams," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2010, pp. 135–150.

[2] Baeldung, "Gradient Descent for Logistic Regression," [Online]. Available: https://www.baeldung.com/cs/gradient-descent-logistic-regression. [Accessed: Feb. 7, 2025].

[3] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. John Wiley & Sons, 2013.

[4] Datamonje, "Gradient Descent and Variants," [Online]. Available: https://datamonje.com/gradient-descent-and-variants/. [Accessed: Feb. 7, 2025].

[5] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[6] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning with Applications in R. Springer, 2013.

[7] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2065, p. 20150202, 2016.

[8] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann, 2011.

[9] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in Proceedings of COMPSTAT, 2010, pp. 177–186.

[10] M. Abadi et al., "TensorFlow: A System for Large-Scale Machine Learning," in Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2016, pp. 265–283.

[11] R. Tibshirani, "Regression shrinkage and selection via the Lasso," Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267–288, 1996.

[12] S. Menard, Applied Logistic Regression Analysis, 2nd ed. SAGE Publications, 2002.

[13] Scikit-learn, "Breast Cancer Wisconsin (Diagnostic) Dataset," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html. [Accessed: Feb. 7, 2025].

[14] Scikit-learn, "Scikit-learn: Machine Learning in Python," [Online]. Available: https://scikit-learn.org/stable/index.html. [Accessed: Feb. 7, 2025].

[15] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, 2009.

[16] W. Wolberg, O. Mangasarian, N. Street, and W. Street, "Breast Cancer Wisconsin (Diagnostic) [Dataset]," UCI Machine Learning Repository, 1993. [Online]. Available: https://doi.org/10.24432/C5DW2B. [Accessed: Feb. 7, 2025.