Chatgpt response on why Linear regression is not good for binary classification.

Linear regression is not ideal for predicting binary class labels due to several inherent limitations:

1. Output Range

Linear regression predicts continuous values that can range from −∞-\infty to +∞+\infty, while binary classification requires outputs between 0 and 1 (or two distinct classes).

When modeling probabilities for binary outcomes, predictions outside the range [0, 1] do not make sense.

2. Poor Fit for Binary Data

Linear regression assumes a linear relationship between the input features and the output, but the relationship in binary classification problems is often non-linear.

Binary outcomes usually follow a sigmoid-like (S-shaped) curve, which linear regression cannot capture.

3. Violation of Statistical Assumptions

Linear regression assumes that residuals are normally distributed and homoscedastic (constant variance). Binary classification data violates these assumptions since the output is categorical and not continuous.

4. Suboptimal Decision Boundaries

Linear regression does not provide a natural threshold to separate classes. Arbitrarily choosing a threshold (e.g., 0.5) to classify outputs can lead to poor decision-making and misclassifications.

5. Interpretability Issues

The coefficients in linear regression do not have a clear probabilistic interpretation for binary outcomes. Logistic regression, in contrast, provides coefficients that can be interpreted as log odds, which are more meaningful in classification contexts.

Preferred Alternative: Logistic Regression

Logistic regression is specifically designed for binary classification. It uses the sigmoid function to map predicted values into a probability range [0, 1] and outputs a probability that can be thresholded to classify instances.

It also optimizes for classification-specific loss (log loss) rather than minimizing the sum of squared errors.

In summary, linear regression's assumptions and output range make it unsuitable for binary classification tasks. Models like logistic regression or decision trees are more appropriate for such problems.