CONTENTS

Introduction

This report includes the analysis including data preprocessing, model training, and performance evaluation performed on the NCI_SEER_CRC.csv. Both a default XGBClassifier model and an optimized version using hyperparameter tuning are considered. The objective is to assess the models based on various metrics such as accuracy, ROC curve threshold, p-value for model acceptance, and confidence intervals. The feature importance is also interpreted using SHAP plots.

Data Preprocessing

To ensure data quality and consistency, the dataset was loaded from a CSV file and went through a variety of prior stages. Firstly, several of columns were renamed to improve clarity: "Year of diagnosis" became "YDD", "Race recode (W, B, AI, API)" became "Race", "Age recode with single ages and 100+" became "Age", "Grade (through 2017)" became "Grade", "SEER registry (with CA and GA as whole states)" became "Location", "Behavior recode for analysis" became "Diagnosis", and "Marital status at diagnosis" became "MaritalStatus". After that, the categories of variables were listed: the "Diagnosis" column was mapped to 1 for "Malignant" and 0 for "Not Malignant" (or other situations), and the "MaritalStatus" column was mapped to 1 for "Married," 2 for "Single," and 3 for all other scenarios (Separated, Widowed, Divorced, unknown , etc.).

After then, the dataset was cleared of a number of unnecessary columns, such as "Origin recode NHIA (Hispanic, Non-Hispanic)", "Age recode with <1 year olds", "Primary Site – labeled", "Combined Summary Stage (2004+)", "Total number in situ/malignant tumors for patient", and "COD to site recode". The numbers in the "Age" column have been eliminated through the removal of the trailing "years." A binary column called "SurvivalRecode" was also created from the "Survival months" column; values in this column were set to 1 if the number of survival months exceeded or equaled 60 and 0 otherwise. The dataset was guaranteed to be adequately prepared for the following model development and evaluation.

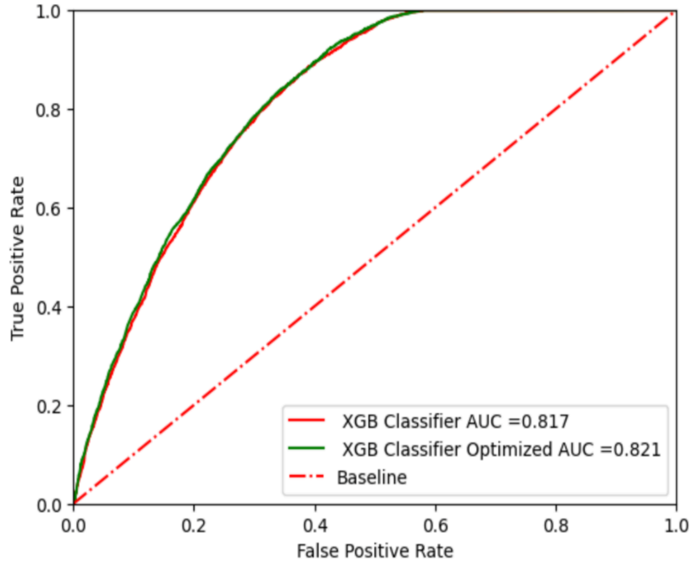
Model Training and Evaluation

The dataset was preprocessed and then divided into training and testing sets. In order, the target (y) and features (X) were defined. The initial research dealt with XGBClassifier default model training. With a precision of 0.80 for the non-survival class and 0.63 for the survival class, the default model's accuracy was 73.32%. F1-scores were 0.79 and 0.64 for the non-survival class and 0.66 and 0.77 for the survival class, respectively, based on the recall values. For the non-survival class, the confusion matrix for the default model displayed 6685 true positives and 1971 false negatives, whereas for the survival class, it displayed 3292 true positives and 1660 false negatives.

Then, in order to improve performance, an optimized XGBClassifier model was developed via parameter tuning with the following parameters learning_rate=0.02, n_estimators=600, objective='binary: logistic', and nthread=1. With an accuracy of 73.69%, the optimized model showed a minor performance gain. While it increased to 0.63 for the survival class, the precision for the non-survival class stayed at 0.80. In the survival class, the recall increased to 0.67, whereas in the non-survival class, it increased to 0.78. F1-scores of 0.79 for the non-survival class and 0.65 for the survival class were the result of this. For the non-survival class, the optimized model's confusion matrix showed 6709 true positives and 1947 false negatives, while for the survival class, 3319 true positives and 1633 false negatives.

Receiver Operating Characteristic (ROC) Curve and Threshold

The plotting of the ROC curves for both models enabled us to observe how well they performed in terms of true positive rate (TPR) against false positive rate (FPR) at different threshold values. The optimized model had an AUC of 0.821, which was slightly greater than the 0.817 of the default model. These results reveal that both models differentiate positive and negative classes well, with the optimized model showing a slight improvement. With a G-Mean of 0.742, the optimal threshold for the default model was discovered to be 0.4026. With a G-Mean of 0.743, the optimal threshold for the optimized model was 0.4280.

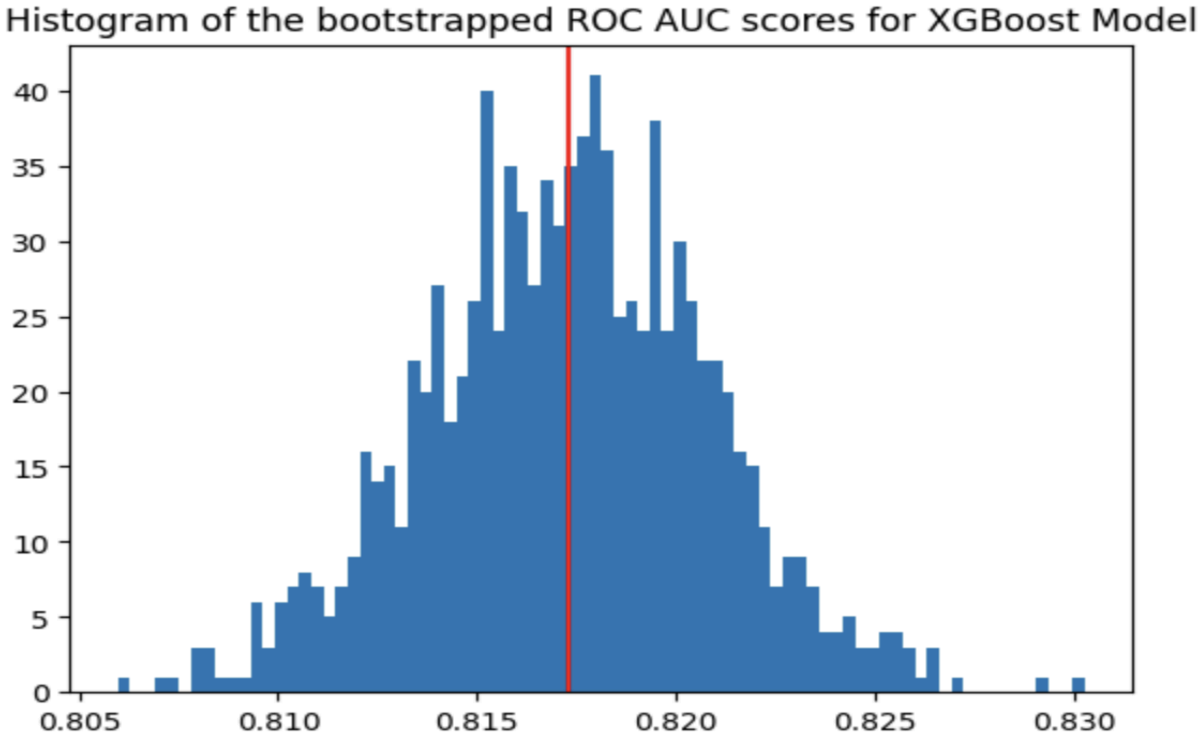


Bootstrap Analysis

The optimized model outperforms the default model, as shown by the optimized model's better mean AUROC score of 0.821, which is less than the default model's 0.817. This improvement indicates that the optimized model has slightly better discriminatory power. Furthermore, the optimized model's 95% confidence interval ([0.814, 0.828]) is both narrower and higher than the default model's ([0.810, 0.824]). This narrower interval suggests reduced variability and more consistent performance, which is further supported by the histogram of bootstrapped AUROC scores showing a large number of scores clustered near the mean. Overall, the optimized model's higher and more stable AUROC scores make it a more reliable choice for predictive accuracy.

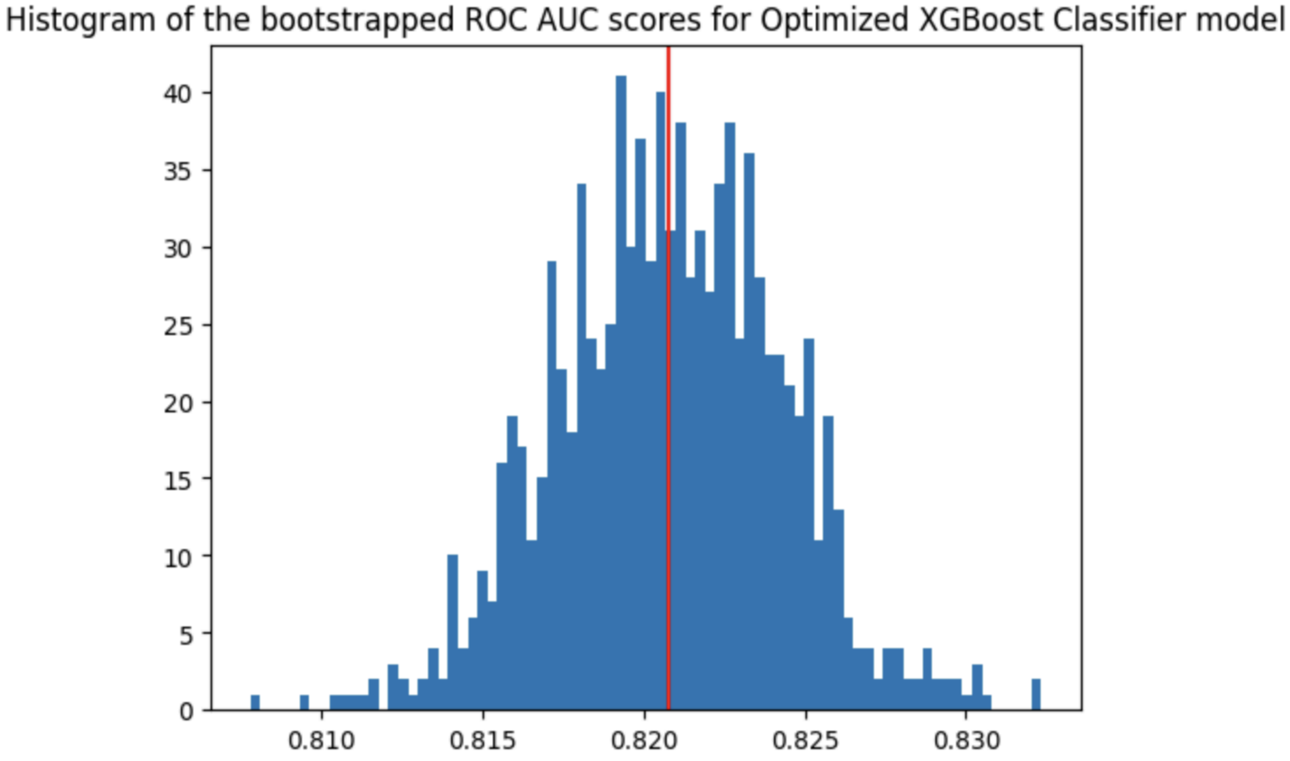
Histogram Showing Confidence Intervals

The histograms indicate that both models perform consistently, with the scores closely clustered around the mean in each case. However, the optimized model shows a slight improvement in its AUROC scores, as indicated by its higher mean and wider confidence interval. This suggests that the optimized model is marginally better at distinguishing between classes.



Histogram of the bootstrapped ROC AUC scores for XGBoost Model

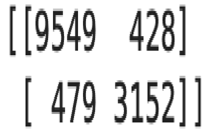
```
The Confidence interval for AUROC XGBoost Classifier model is [0.810 – 0.824]
```



Histogram of the bootstrapped ROC AUC scores for Optimized XGBoost Classifier model

```
The Confidence interval for AUROC XGBoost Classifier optimized model is [0.814 – 0.828]
```

P-Value for Model Acceptance



```
[[9549  428]
 [ 479 3152]]
```

McNemar's test results indicate that the optimized XGBoost model exhibits a slight improvement over the default XGBoost model. This is shown in the confusion matrix, where the default model correctly classified 428 instances that the optimized model did not, whereas the optimized model correctly classified 479 instances that the default model did not. The chi-squared value of 2.756, accompanied by a p-value of 0.097, exceeds the conventional significance threshold of 0.05. This suggests that although there is a difference in performance between the two models, it is not statistically significant at the 5% level. Consequently, while the optimized model demonstrates an enhancement in performance, this improvement may not be substantial enough to achieve statistical significance.

SHAP Plots for feature selection

To understand feature relevance for both models, SHAP (SHapley Additive exPlanations) plots were created. The resulting graphs highlighted important indicators for survival outcomes in the dataset and offered insightful information about how each feature affected the algorithms' predictions. The importance of characteristics including age, grade, and marital status, among others, in predicting survival outcomes has been proved by the SHAP plots.

a) Feature Importance

The default model indicates that features like "YDD," "MaritalStatus_3," and "Grade_Unknown" have a significant impact on predictions, as indicated by their SHAP values. On the other hand, the improved model may change the relevance of these features, thereby affecting the order or magnitude of SHAP values. This demonstrates a shift in how the model prioritizes distinct features as its parameters are adjusted.

b) Distribution of SHAP values

The default model's SHAP values vary widely, indicating that factors have varying influences on predictions across the dataset. In the optimized model, these values may become more concentrated or dispersed, demonstrating the model's improved sensitivity, which results in either more consistent or diverse impacts on predictions.

c) Directionality of Influence

The default model may not clearly distinguish between the direction of feature influence (e.g., increasing or decreasing predictions). The optimized model, on the other hand, demonstrates a clearer directionality, with a more visible split between features that have positive or negative effects on predictions, due to improvements in parameter tuning.

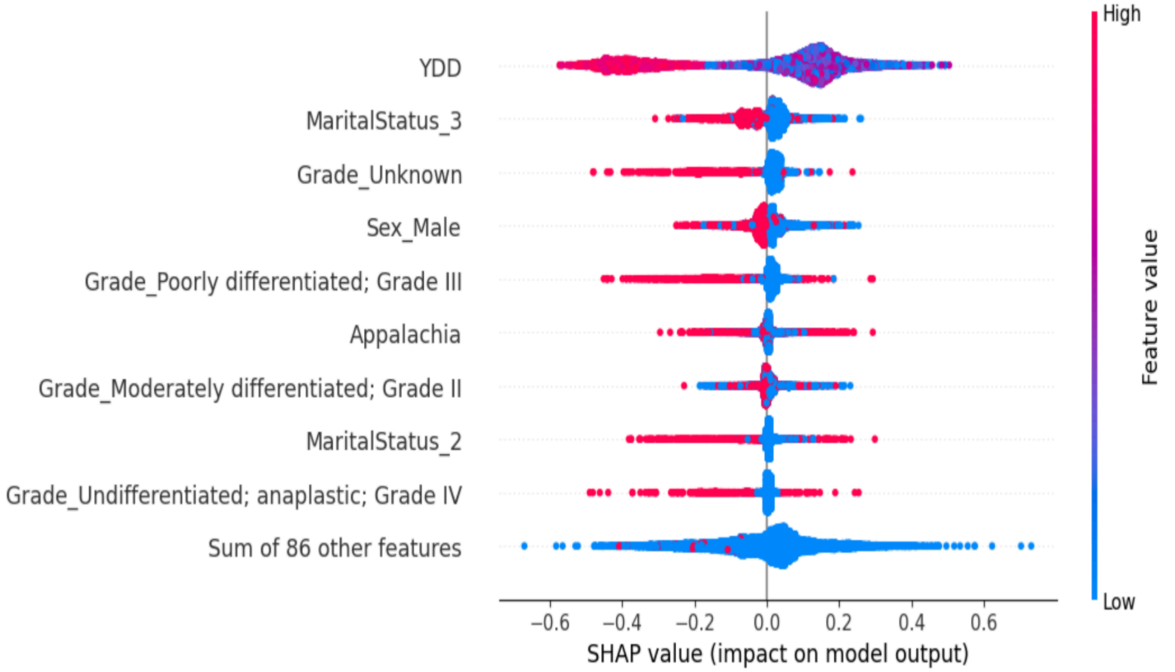
d) Potential Interaction Efforts

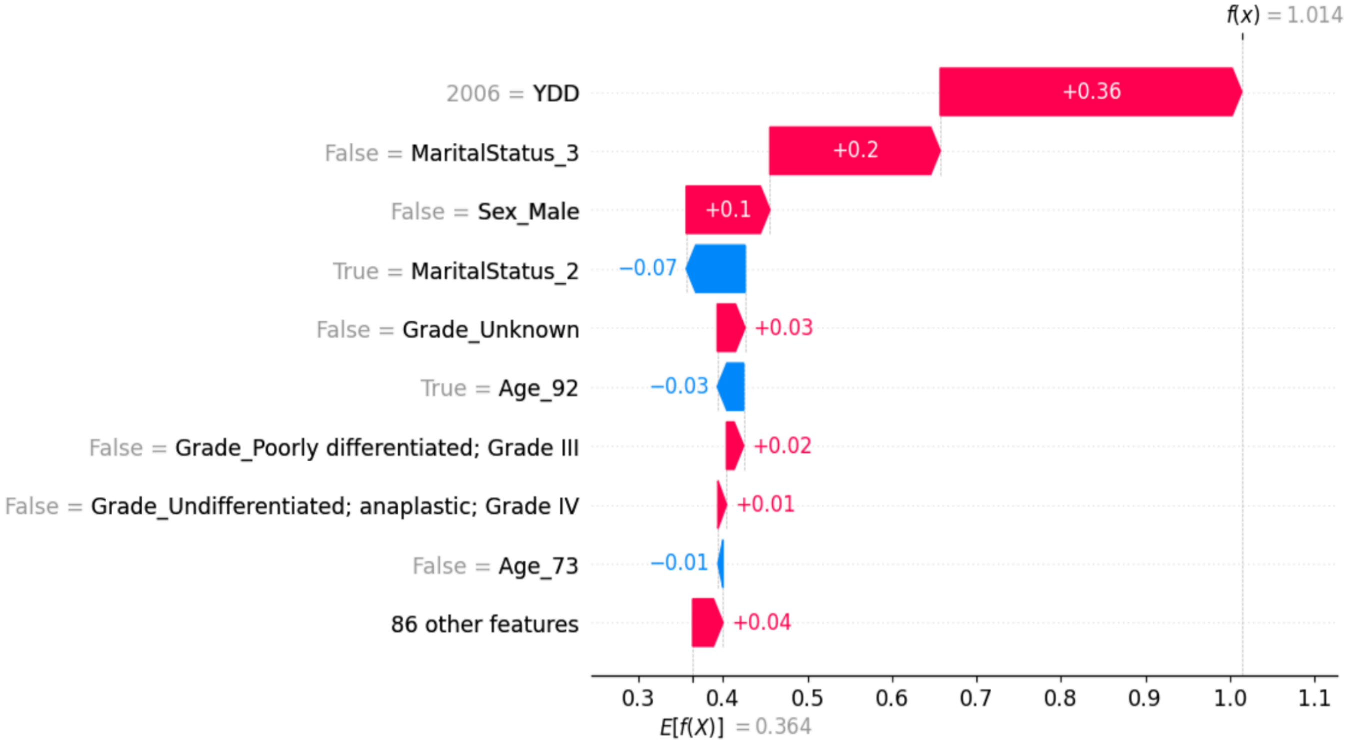
The default model suggests interaction effects between features, but lacks a clear definition, leading to uncertainty in interpretations. In contrast, the improved model highlights these interaction effects, exhibiting different patterns and demonstrating a more effective approach to feature interactions.

e) Overall model Behavior

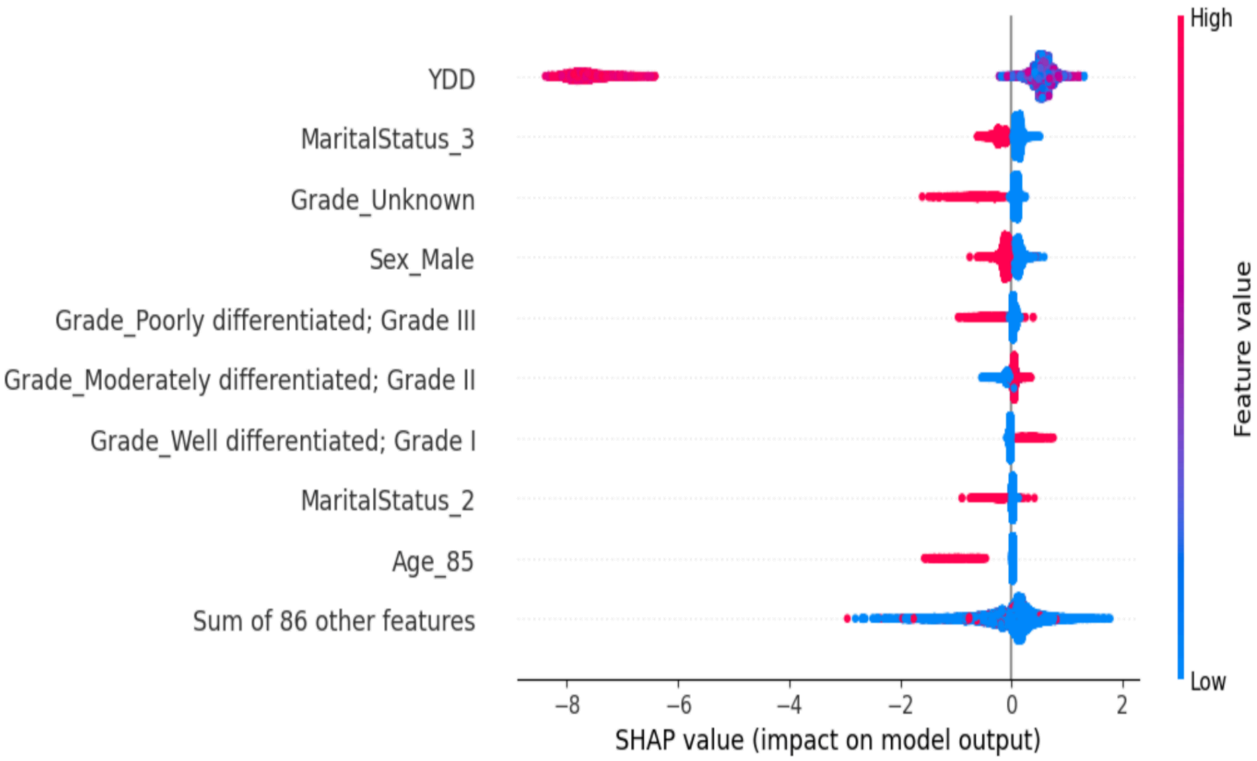
The default model performs well, however, there is a chance for improvement in terms of feature priority and interaction management. However, the optimized model exhibits improved behavior, including better-defined feature importance, clearer directionality of influence, and more stable predictions, resulting in a more interpretable and dependable outcome.

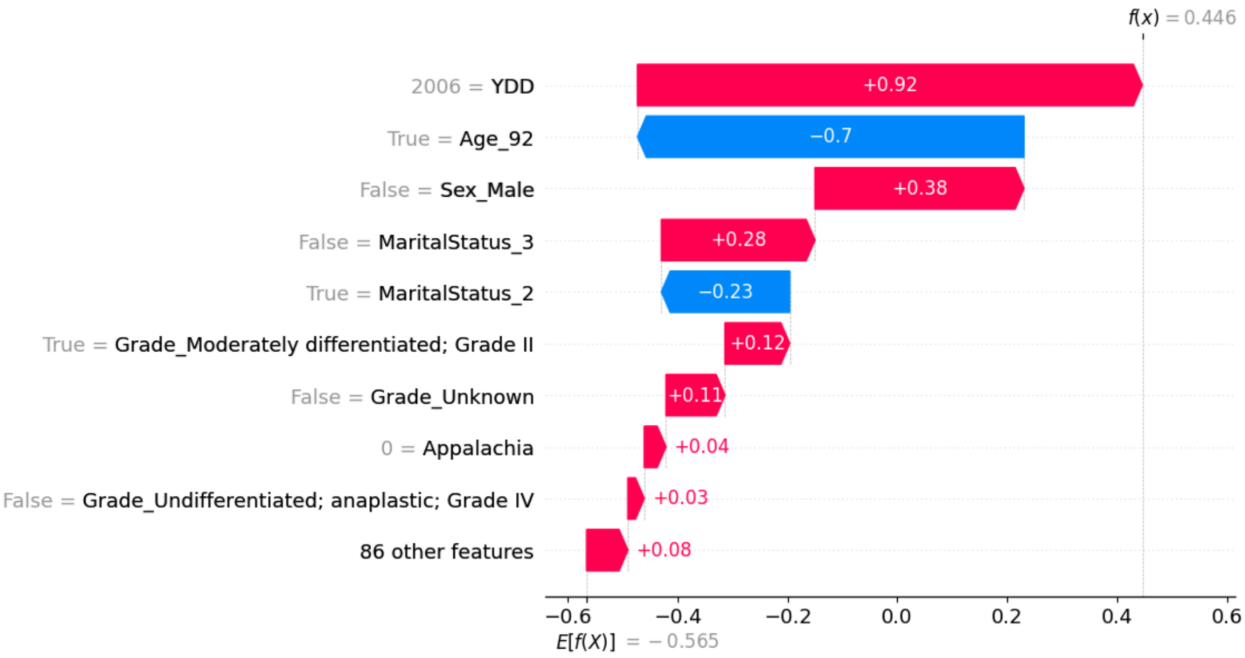
For Default XGBoost Model:





For Optimized XGBoost Model:





Conclusion

The analysis of the AUROC scores and confidence intervals for both the default and optimized XGBoost models shows that the optimized model performs slightly better. Although this improvement is visible in a slight rise in the mean AUROC score and a wider confidence interval for the optimized model, McNemar's test shows that the difference is not statistically significant at the 5% level, implying that the improvement may not be practical. Both models operate consistently and reliably, as indicated by closely clustered bootstrapped AUROC scores around their respective means, demonstrating their effectiveness in function. Furthermore, the significance of the characteristics in both models, as indicated by SHAP plots, adds relevance and confidence to the models' predictions by demonstrating how each feature contributes to the outcome. Despite the optimized model demonstrating a slight increase in accuracy and other measures, McNemar's test found no statistical significance, implying that the improvements, while useful, do not significantly affect the models' comparative performance in practical applications. Thus, while the optimized model makes significant developments they do not result in major benefits over the default model in terms of statistical assessment.