**Assignment – 3 Final Project**

Uday Bhaskar Valapadasu

11636469

**(Part -1)**

**Output Screenshots for SHAP Values, Histogram & ROC Curve**

A graph with blue bars

Description automatically generated

A graph with different colored lines

Description automatically generated

A graph of a graph

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated

**Overview Table for Part-1**

|  |  |
| --- | --- |
| **Model Performance** | - The optimized XGBoost model (72.73% accuracy) outperformed the default model (70.78% accuracy).  - AUC improved from 0.7519 (default) to 0.7728 (optimized), indicating better classification ability. |
| **Statistical Significance** | The p-value of 0.0000 suggests the model's performance is significantly better than random chance. |
| **Feature Importance** | SHAP plots were generated to visualize feature importance and their impact on predictions. |
| **Prediction Distribution** | A histogram of prediction probabilities was created to show the model's confidence across predictions. |

Overall, the optimized model shows improved performance, with SHAP plots providing insights into feature importance. The extremely low p-value indicates strong statistical significance of the model's predictive power.

**(Part -2)**

**Key Observations:**

1. **Accuracy:**
   1. Default XGBoost model accuracy: 0.7358
   2. Optimized XGBoost model accuracy: 0.7424
   3. The optimized model shows a slight improvement in accuracy over the default model.
2. **ROC Curve and AUC:**
   1. Default model AUC: 0.8201
   2. Optimized model AUC: 0.8271
   3. Both models show good discriminative ability, with the optimized model performing slightly better. The ROC curve plot visually confirms this improvement.
3. **P-value for model acceptance:**
   1. P-value: 0.0000
   2. This extremely low p-value suggests that the model's performance is statistically significant and not due to chance.
4. **Histogram of prediction probabilities:**
   1. The histogram shows the distribution of predicted probabilities, which can be interpreted as the model's confidence in its predictions. The distribution appears to be bimodal, with peaks at lower and higher probabilities, suggesting the model is often quite confident in its predictions.
5. **SHAP plots for feature selection:**
   1. Two SHAP plots were generated:
      1. A bar plot showing feature importance
      2. A summary dot plot showing the impact and distribution of feature values
   2. These plots help identify which features have the most significant impact on the model's predictions.
6. **Threshold:**
   1. While not explicitly stated, the ROC curve can be used to determine an optimal threshold for classification, balancing the trade-off between sensitivity and specificity.

In conclusion, the optimized XGBoost model shows good performance in predicting 5-year survival for colorectal cancer patients, with high accuracy and AUC scores. The model's predictions are statistically significant, and the SHAP plots provide valuable insights into which factors are most important in determining survival outcomes.

**Output Screenshots for SHAP Values, Histogram & ROC Curve**

A blue bar graph with white text

Description automatically generated

A graph of a graph with text

Description automatically generated with medium confidence

A graph with blue lines

Description automatically generatedA screen shot of a graph

Description automatically generated