

Cancer Model Univariate Analysis Report

2024-02-08

Overview

Cancer Model Univariate Analysis Report

These sorted results for the features in this report indicate the average cross-validated test scores for each feature, if it were used as the only predictor in a simple linear model. Keep in mind that these results are based on the average, without considering the standard deviation. This means that the results are not necessarily the best predictors, but they are the best on average, and provide a fine starting point for grouping those predictors that are on average better than others. This means that nothing was done to account for possible sampling variability in the sorted results. This is a limitation of the univariate analysis, and it is important to keep this in mind when interpreting the results. It is also important to consider further that depending on the purpose of the model, the most appropriate features may not be the ones with the highest average test scores, if a different metric is more important.

In particular, this should not be taken as an opinion (actuarial or otherwise) regarding the most appropriate features to use in a model, but it rather provides a starting point for further analysis.

| | Accuracy | Precision | Recall | AUC | F1 | MCC | Ave. |
|--------------------|----------|-----------|--------|-------|-------|--------|-------|
| WorstSymmetry | 32.5% | 44.5% | 30.8% | 33.1% | 36.4% | -33.0% | 24.1% |
| CompactnessError | 30.9% | 41.7% | 25.2% | 32.9% | 31.4% | -34.1% | 21.3% |
| WorstTexture | 28.1% | 40.3% | 30.3% | 27.4% | 34.6% | -43.8% | 19.5% |
| ConcavityError | 29.0% | 39.0% | 23.2% | 31.0% | 29.1% | -38.0% | 18.9% |
| MeanTexture | 27.2% | 39.2% | 28.9% | 26.7% | 33.2% | -45.2% | 18.3% |
| ConcavePointsError | 26.9% | 38.1% | 26.3% | 27.1% | 31.1% | -44.7% | 17.5% |
| MeanRadius | 13.7% | 20.4% | 12.9% | 14.0% | 15.8% | -71.2% | 0.9% |
| MeanPerimeter | 13.0% | 19.5% | 12.3% | 13.2% | 15.1% | -72.6% | 0.1% |
| WorstRadius | 9.3% | 13.0% | 7.8% | 9.8% | 9.8% | -80.1% | -5.1% |
| WorstPerimeter | 8.3% | 12.0% | 7.3% | 8.6% | 9.1% | -82.4% | -6.2% |

This table shows an overview of the results for the variables in this file, representing those whose average test score are ranked between 21 and 30 of the variables passed to the Cancer Model.

Univariate Report

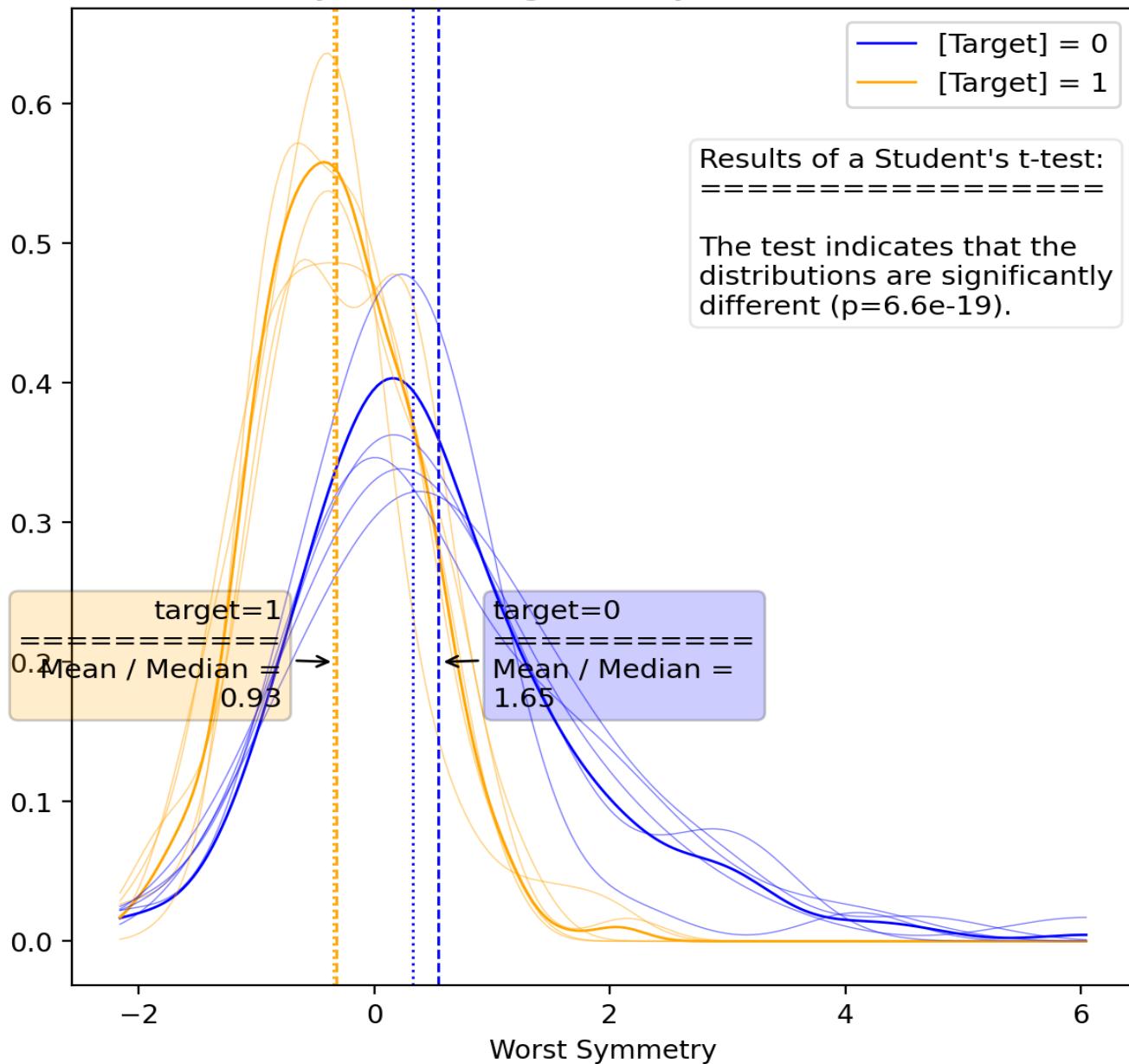
Worst Symmetry - Results

| | Fold-1 | Fold-2 | Fold-3 | Fold-4 | Fold-5 | Agg. Mean | Agg. SD |
|-------------------------|----------|----------|----------|----------|----------|-----------|---------|
| Fitted Coef. | -1.2e+00 | -1.1e+00 | -1.0e+00 | -1.1e+00 | -1.1e+00 | 1.1e+00 | 4.6e-02 |
| Fitted p-Value | 1.5e-15 | 4.6e-15 | 8.2e-14 | 8.9e-15 | 4.4e-14 | 1.2e-04 | 3.4e-14 |
| Fitted Std. Err. | 0.147 | 0.144 | 0.140 | 0.143 | 0.146 | 0.287 | 0.003 |
| Conf. Int. Lower | -1.5e+00 | -1.4e+00 | -1.3e+00 | -1.4e+00 | -1.4e+00 | 5.4e-01 | 5.0e-02 |
| Conf. Int. Upper | -8.9e-01 | -8.4e-01 | -7.7e-01 | -8.3e-01 | -8.2e-01 | 1.7e+00 | 4.1e-02 |
| Train Accuracy | 67.9% | 67.7% | 67.3% | 66.5% | 67.4% | 32.5% | 0.5% |
| Val Accuracy | 66.1% | 64.4% | 68.1% | 72.0% | 67.9% | 32.5% | 2.8% |
| Train AUC | 67.2% | 66.9% | 66.6% | 66.1% | 67.1% | 33.1% | 0.4% |
| Val AUC | 64.8% | 66.2% | 67.2% | 71.2% | 67.0% | 33.1% | 2.4% |
| Train F1 | 73.0% | 73.3% | 72.3% | 72.1% | 72.7% | 36.4% | 0.5% |
| Test F1 | 72.7% | 66.7% | 75.3% | 75.9% | 72.7% | 36.4% | 3.7% |
| Train Precision | 76.3% | 76.7% | 74.9% | 77.2% | 77.9% | 44.5% | 1.1% |
| Val Precision | 76.7% | 78.3% | 82.1% | 74.6% | 73.7% | 44.5% | 3.3% |
| Train Recall | 69.9% | 70.2% | 69.8% | 67.7% | 68.1% | 30.8% | 1.2% |
| Val Recall | 69.1% | 58.1% | 69.6% | 77.2% | 71.8% | 30.8% | 7.0% |
| Train MCC | 33.7% | 33.0% | 32.8% | 31.1% | 33.0% | -33.0% | 1.0% |
| Val MCC | 28.6% | 31.7% | 32.0% | 42.6% | 33.9% | -33.0% | 5.3% |
| Train Log-Loss | 11.58 | 11.63 | 11.78 | 12.07 | 11.77 | 24.32 | 0.19 |
| Val Log-Loss | 12.21 | 12.85 | 11.48 | 10.09 | 11.56 | 24.32 | 1.03 |

Univariate Report

Worst Symmetry - Kernel Density Plot

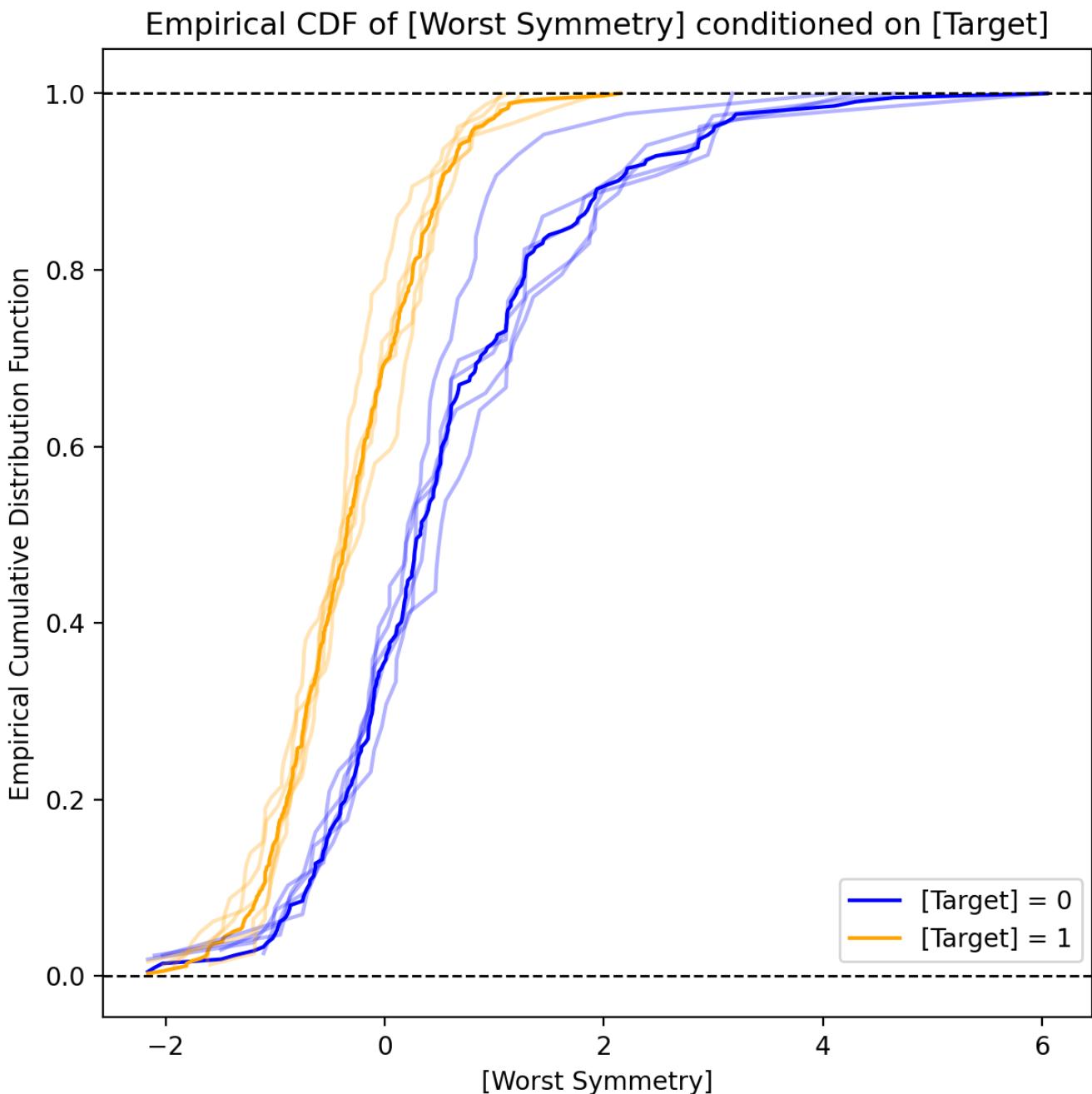
Kernel Density Plot of [Worst Symmetry] by [Target]
Distributions by level are significantly different at the 95% level.



This plot shows the Gaussian kernel density for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the density of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data. There are annotations with the results of a t-test for the difference in means between the feature variable at each level of the target variable. The annotations corresponding to the color of the target variable level show the mean/median ratio to help understand differences in skewness between the levels of the target variable.

Univariate Report

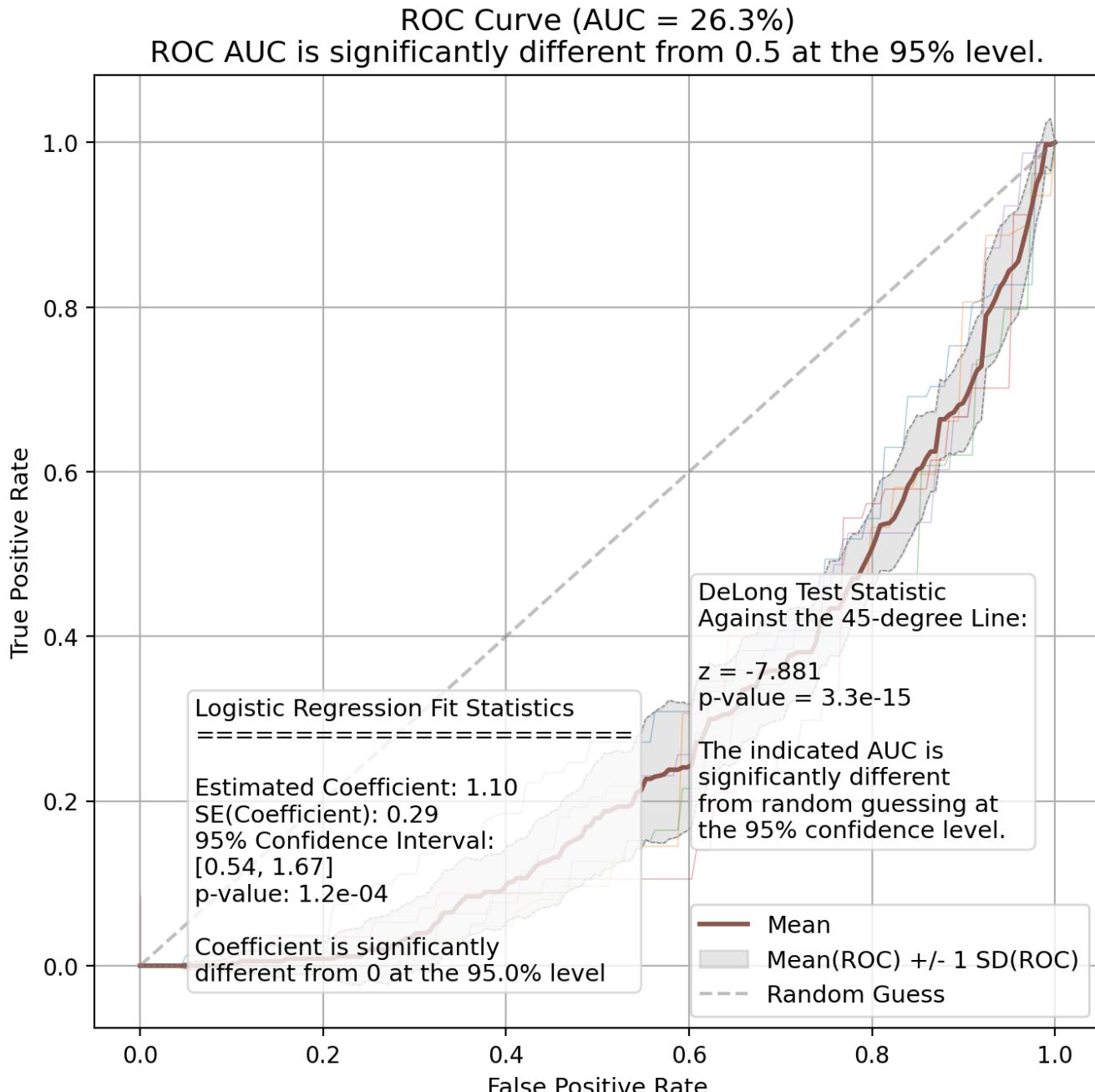
Worst Symmetry - Empirical CDF Plot



This plot shows the empirical cumulative distribution function for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the cumulative distribution of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data, and whether or not it is reasonable to assume that the data is drawn from different distributions.

Univariate Report

Worst Symmetry - ROC Curve

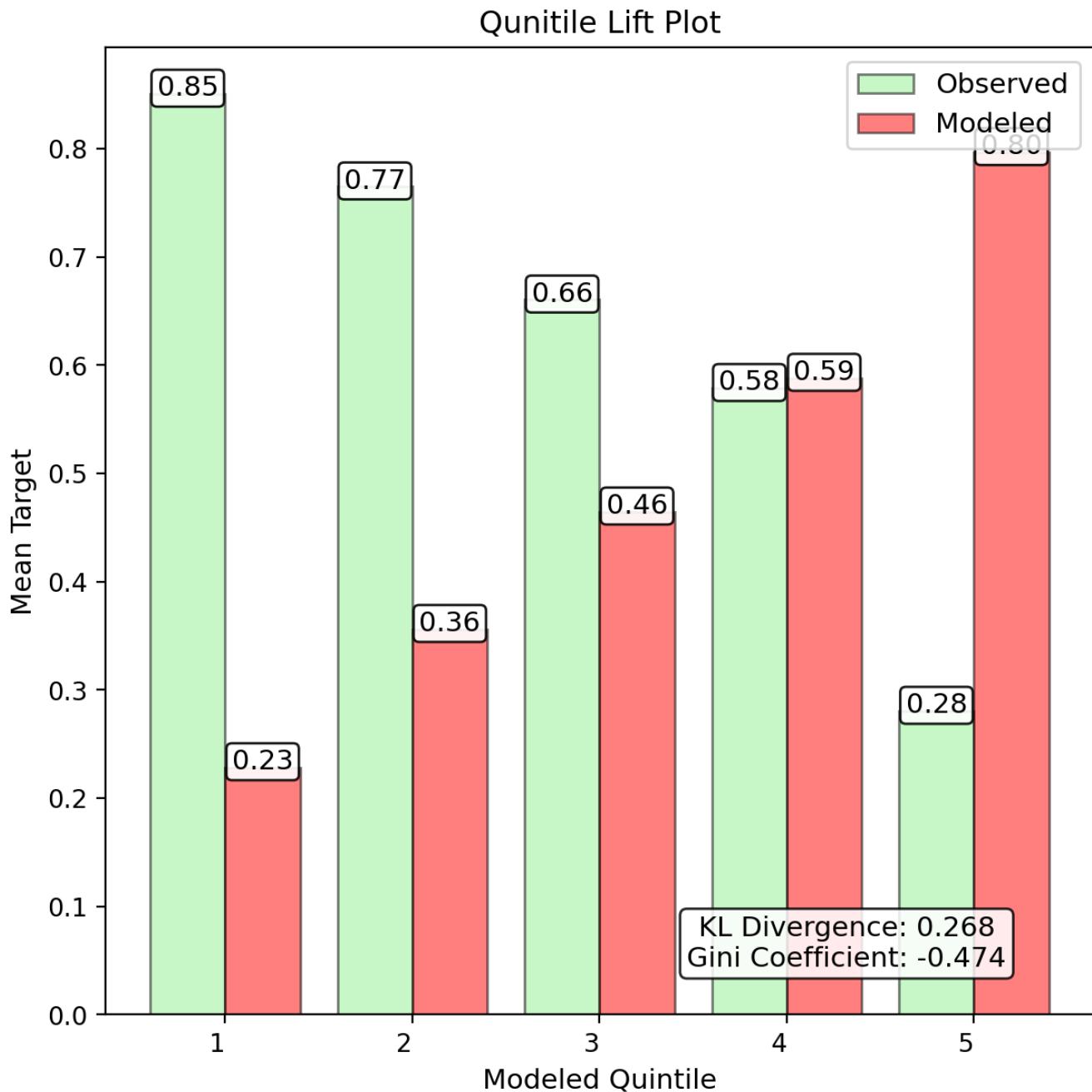


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, and no other features. Annotations are on the plot to help understand the results of the model, including the coefficient, standard error, and p-value for the feature variable. The cross-validation folds are used to create the grey region around the mean ROC curve to help understand the variability of the data.

Significance of the ROC curve is determined based on a modified version the method from DeLong et al. (1988). In brief, the AUC is assumed to be normally distributed, and I calculate the empirical standard error from the cross-validated AUC values. I then calculate a z-score for the AUC, and use the z-score to calculate a p-value. The p-value is then used to determine the significance of the AUC. This is a simple test, and should be used with caution.

Univariate Report

Worst Symmetry - Quintile Lift



The quintile lift plot is meant to show the power of the single feature to discriminate between the highest and lowest quintiles of the target variable.

Univariate Report

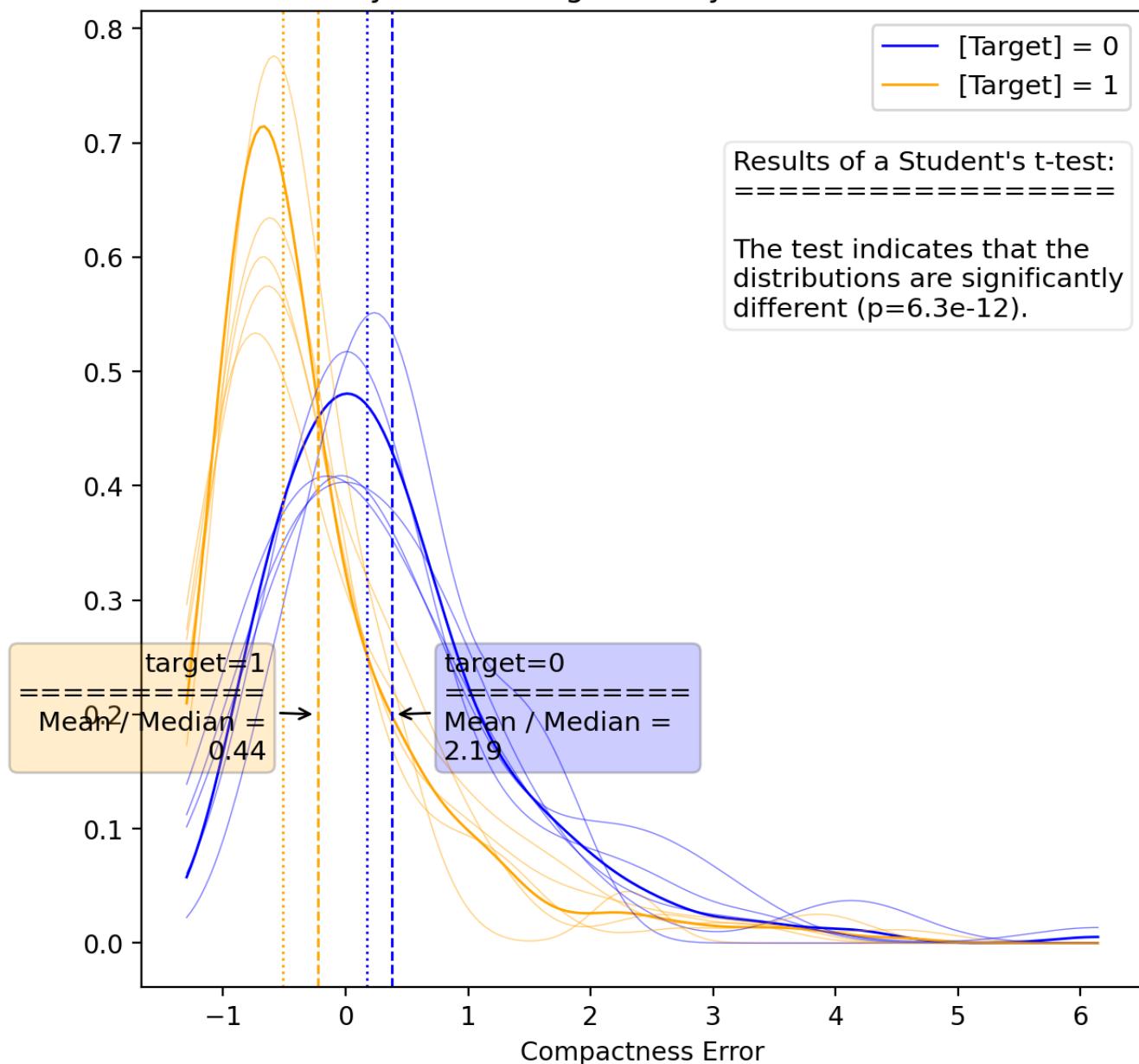
Compactness Error - Results

| | Fold-1 | Fold-2 | Fold-3 | Fold-4 | Fold-5 | Agg. Mean | Agg. SD |
|-------------------------|----------|----------|----------|----------|----------|-----------|---------|
| Fitted Coef. | -7.5e-01 | -6.0e-01 | -6.7e-01 | -6.5e-01 | -7.1e-01 | 2.9e+00 | 5.9e-02 |
| Fitted p-Value | 3.6e-09 | 1.3e-07 | 2.5e-08 | 3.2e-08 | 1.5e-08 | 2.9e-01 | 4.9e-08 |
| Fitted Std. Err. | 0.127 | 0.113 | 0.121 | 0.118 | 0.125 | 2.703 | 0.006 |
| Conf. Int. Lower | -1.0e+00 | -8.2e-01 | -9.1e-01 | -8.8e-01 | -9.5e-01 | -2.4e+00 | 7.0e-02 |
| Conf. Int. Upper | -5.0e-01 | -3.8e-01 | -4.4e-01 | -4.2e-01 | -4.6e-01 | 8.2e+00 | 4.8e-02 |
| Train Accuracy | 69.0% | 68.2% | 69.5% | 69.3% | 68.9% | 30.9% | 0.5% |
| Val Accuracy | 70.2% | 70.3% | 67.3% | 69.0% | 68.7% | 30.9% | 1.2% |
| Train AUC | 67.1% | 66.3% | 67.7% | 66.8% | 67.4% | 32.9% | 0.6% |
| Val AUC | 69.0% | 67.2% | 64.9% | 69.1% | 65.9% | 32.9% | 1.9% |
| Train F1 | 75.0% | 74.4% | 75.2% | 75.9% | 75.0% | 31.4% | 0.5% |
| Test F1 | 76.1% | 76.9% | 75.2% | 71.6% | 75.4% | 31.4% | 2.1% |
| Train Precision | 75.0% | 75.3% | 74.6% | 76.2% | 77.0% | 41.7% | 1.0% |
| Val Precision | 79.7% | 73.5% | 80.0% | 75.0% | 70.8% | 41.7% | 4.0% |
| Train Recall | 75.0% | 73.6% | 75.9% | 75.7% | 73.1% | 25.2% | 1.3% |
| Val Recall | 72.8% | 80.6% | 70.9% | 68.4% | 80.8% | 25.2% | 5.7% |
| Train MCC | 34.2% | 32.3% | 35.6% | 33.6% | 34.2% | -34.1% | 1.2% |
| Val MCC | 36.8% | 35.8% | 28.1% | 37.8% | 33.4% | -34.1% | 3.9% |
| Train Log-Loss | 11.18 | 11.48 | 10.99 | 11.07 | 11.19 | 24.89 | 0.19 |
| Val Log-Loss | 10.75 | 10.71 | 11.80 | 11.17 | 11.28 | 24.89 | 0.45 |

Univariate Report

Compactness Error - Kernel Density Plot

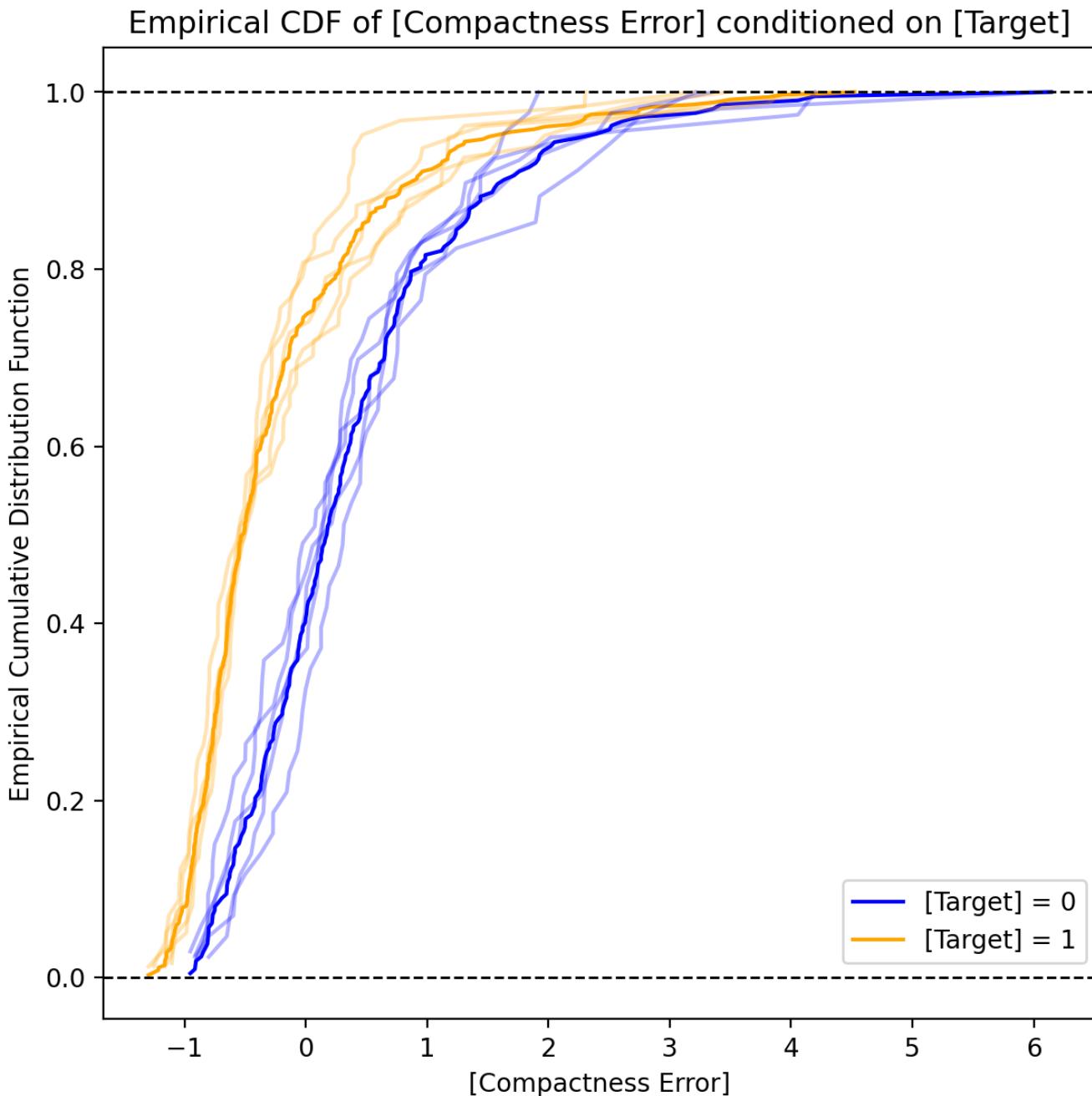
Kernel Density Plot of [Compactness Error] by [Target]
Distributions by level are significantly different at the 95% level.



This plot shows the Gaussian kernel density for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the density of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data. There are annotations with the results of a t-test for the difference in means between the feature variable at each level of the target variable. The annotations corresponding to the color of the target variable level show the mean/median ratio to help understand differences in skewness between the levels of the target variable.

Univariate Report

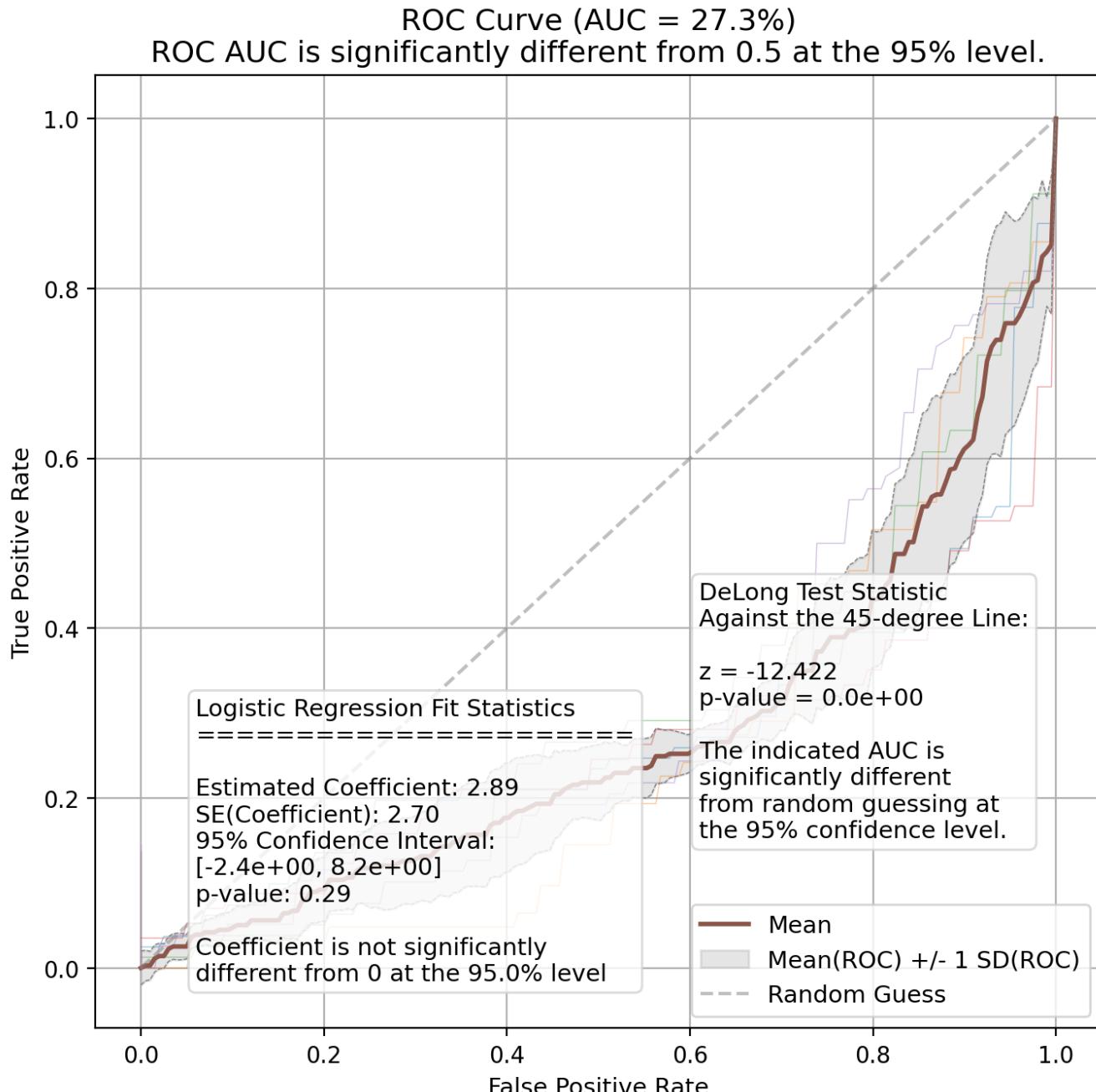
Compactness Error - Empirical CDF Plot



This plot shows the empirical cumulative distribution function for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the cumulative distribution of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data, and whether or not it is reasonable to assume that the data is drawn from different distributions.

Univariate Report

Compactness Error - ROC Curve

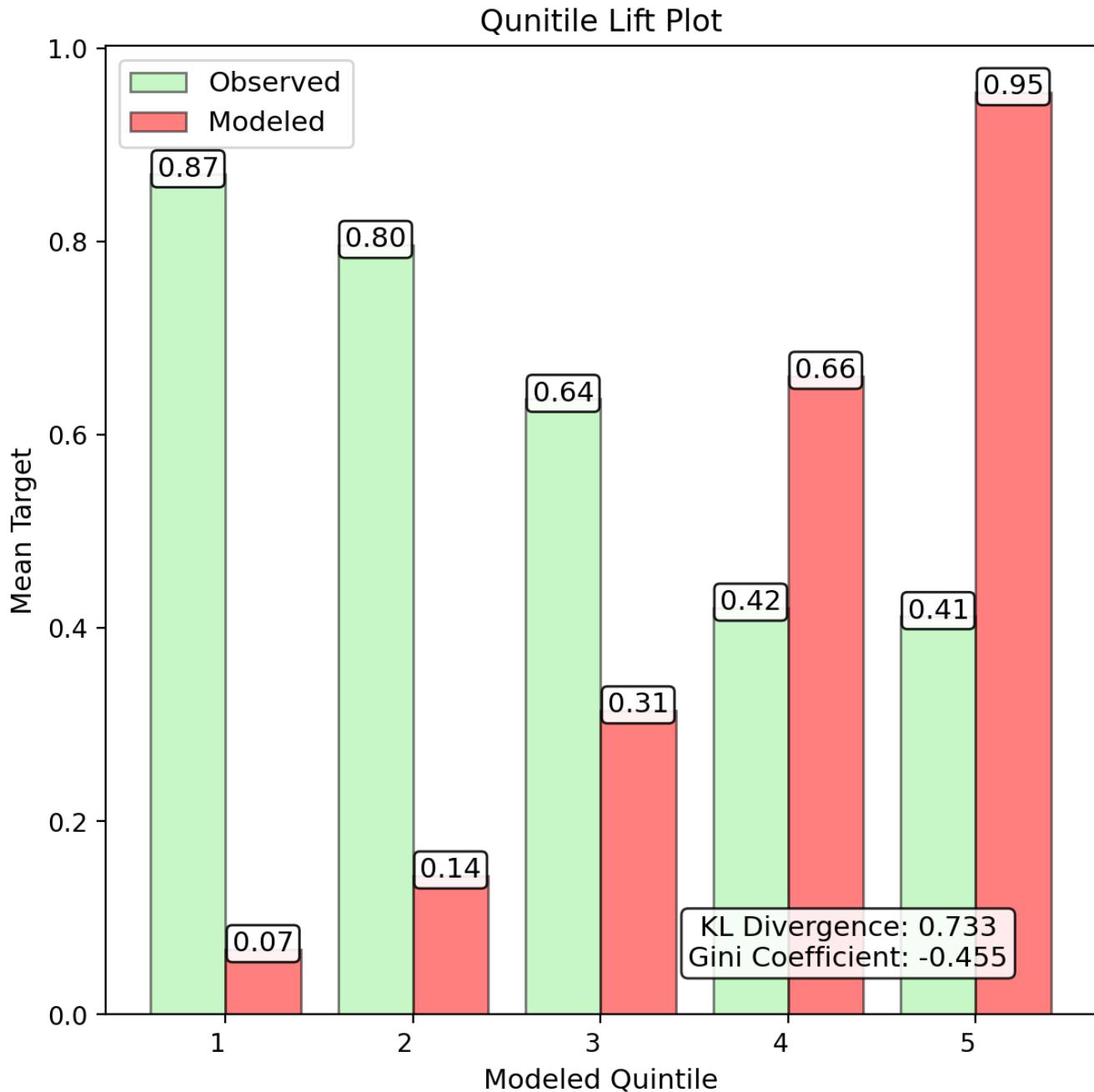


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, and no other features. Annotations are on the plot to help understand the results of the model, including the coefficient, standard error, and p-value for the feature variable. The cross-validation folds are used to create the grey region around the mean ROC curve to help understand the variability of the data.

Significance of the ROC curve is determined based on a modified version the method from DeLong et al. (1988). In brief, the AUC is assumed to be normally distributed, and I calculate the empirical standard error from the cross-validated AUC values. I then calculate a z-score for the AUC, and use the z-score to calculate a p-value. The p-value is then used to determine the significance of the AUC. This is a simple test, and should be used with caution.

Univariate Report

Compactness Error - Quintile Lift



The quintile lift plot is meant to show the power of the single feature to discriminate between the highest and lowest quintiles of the target variable.

Univariate Report

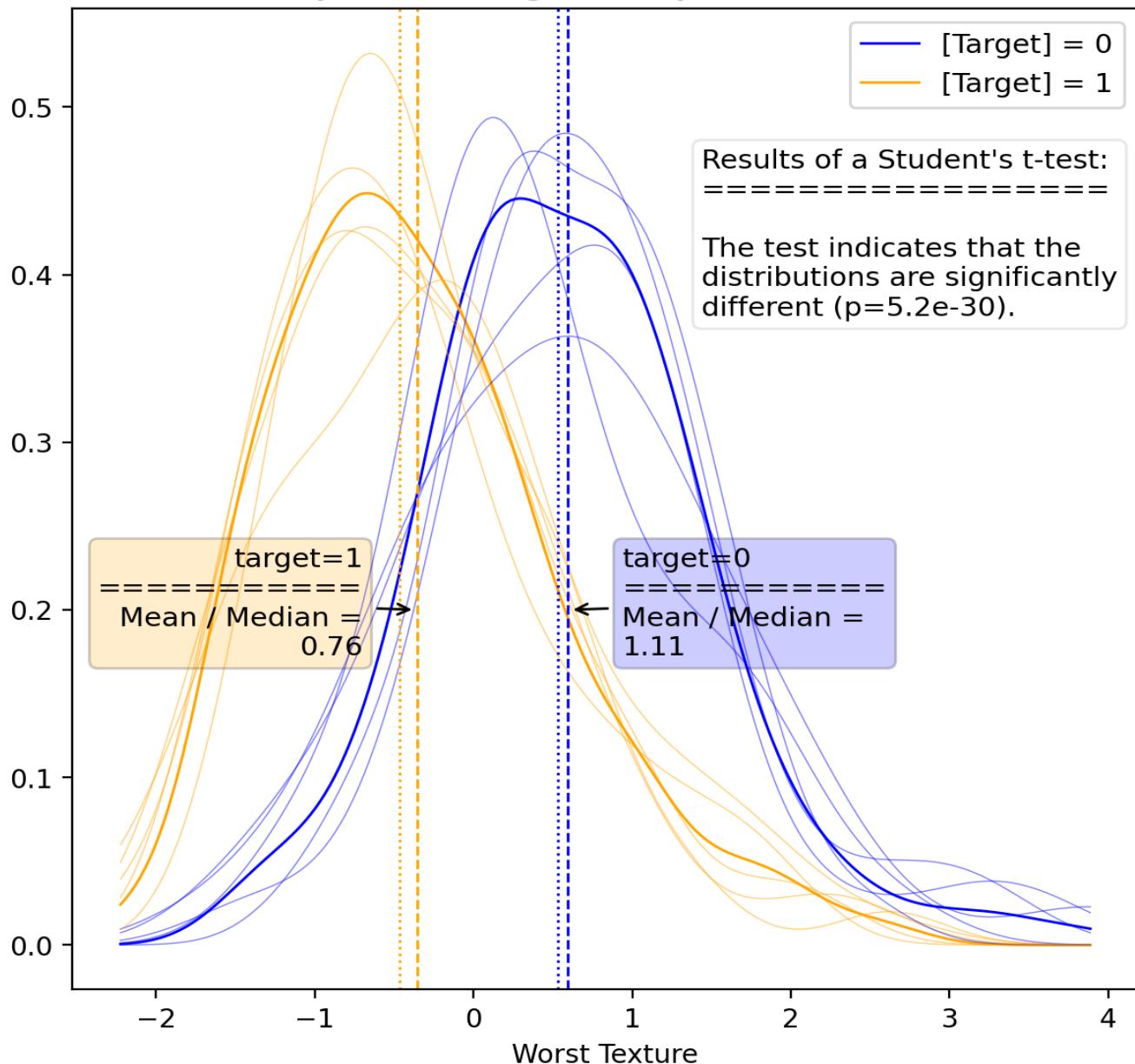
Worst Texture - Results

| | Fold-1 | Fold-2 | Fold-3 | Fold-4 | Fold-5 | Agg. Mean | Agg. SD |
|-------------------------|----------|----------|----------|----------|----------|-----------|---------|
| Fitted Coef. | -1.2e+00 | -1.0e+00 | -1.1e+00 | -1.1e+00 | -1.0e+00 | 1.1e-02 | 8.0e-02 |
| Fitted p-Value | 5.6e-19 | 7.3e-17 | 4.2e-17 | 4.2e-18 | 1.2e-15 | 7.3e-04 | 5.4e-16 |
| Fitted Std. Err. | 0.136 | 0.123 | 0.128 | 0.122 | 0.126 | 0.003 | 0.006 |
| Conf. Int. Lower | -1.5e+00 | -1.3e+00 | -1.3e+00 | -1.3e+00 | -1.3e+00 | 4.6e-03 | 9.0e-02 |
| Conf. Int. Upper | -9.4e-01 | -7.8e-01 | -8.3e-01 | -8.2e-01 | -7.6e-01 | 1.7e-02 | 7.1e-02 |
| Train Accuracy | 73.0% | 70.9% | 72.4% | 72.3% | 70.8% | 28.1% | 1.0% |
| Val Accuracy | 66.1% | 76.2% | 72.6% | 70.0% | 74.8% | 28.1% | 4.0% |
| Train AUC | 73.7% | 71.3% | 73.0% | 73.6% | 71.4% | 27.4% | 1.2% |
| Val AUC | 67.5% | 78.3% | 71.2% | 69.4% | 75.8% | 27.4% | 4.5% |
| Train F1 | 76.6% | 75.2% | 75.6% | 76.1% | 75.1% | 34.6% | 0.6% |
| Test F1 | 70.8% | 78.2% | 79.2% | 73.7% | 76.9% | 34.6% | 3.4% |
| Train Precision | 83.1% | 81.4% | 81.9% | 84.8% | 82.1% | 40.3% | 1.3% |
| Val Precision | 81.0% | 89.6% | 84.3% | 73.7% | 84.6% | 40.3% | 5.9% |
| Train Recall | 71.0% | 69.8% | 70.1% | 69.0% | 69.2% | 30.3% | 0.8% |
| Val Recall | 63.0% | 69.4% | 74.7% | 73.7% | 70.5% | 30.3% | 4.6% |
| Train MCC | 46.0% | 41.3% | 44.9% | 45.3% | 41.2% | -43.8% | 2.3% |
| Val MCC | 33.4% | 55.1% | 40.0% | 38.8% | 50.7% | -43.8% | 9.0% |
| Train Log-Loss | 9.72 | 10.47 | 9.96 | 9.99 | 10.53 | 25.91 | 0.35 |
| Val Log-Loss | 12.21 | 8.56 | 9.89 | 10.81 | 9.08 | 25.91 | 1.45 |

Univariate Report

Worst Texture - Kernel Density Plot

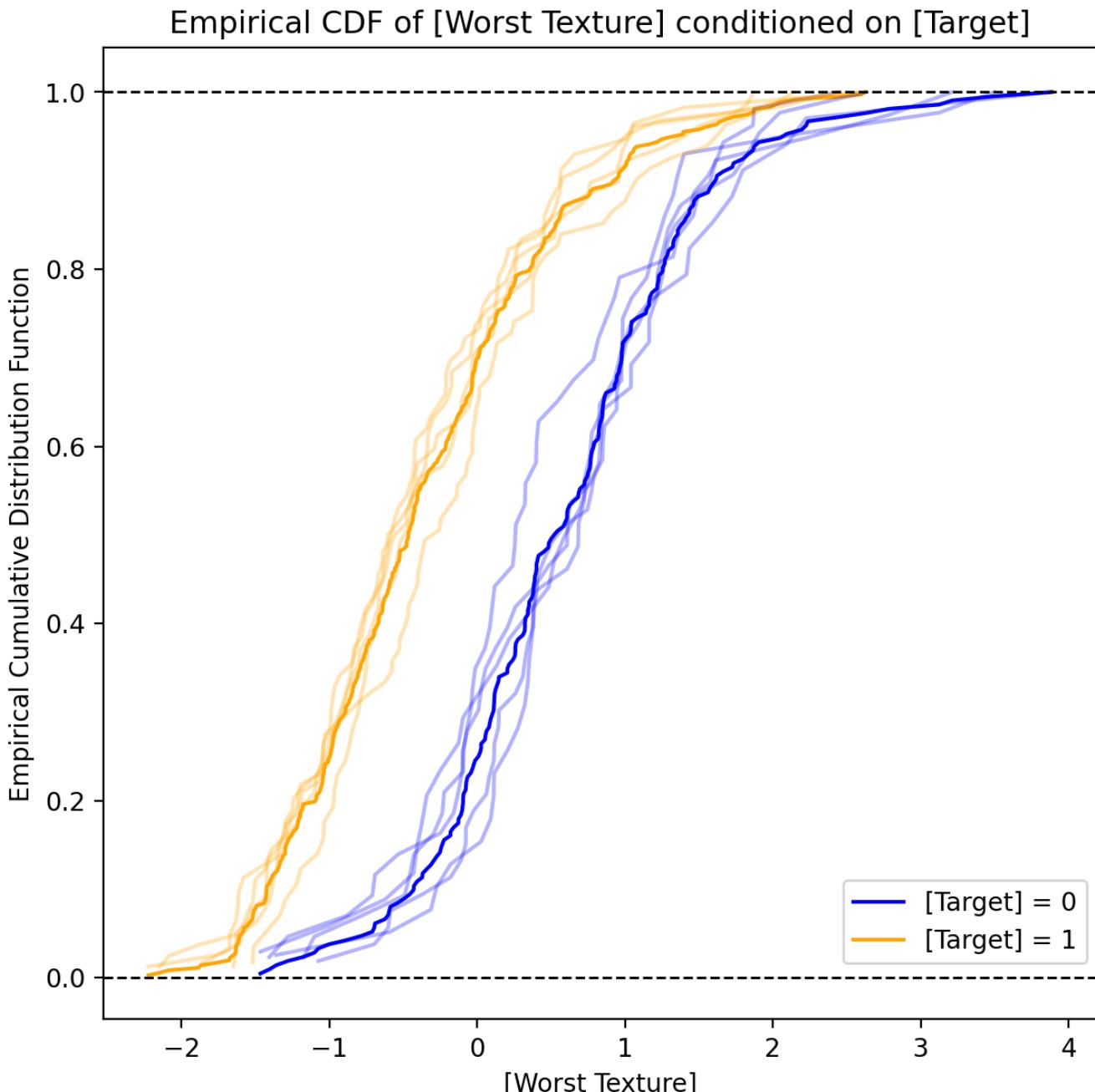
Kernel Density Plot of [Worst Texture] by [Target]
Distributions by level are significantly different at the 95% level.



This plot shows the Gaussian kernel density for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the density of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data. There are annotations with the results of a t-test for the difference in means between the feature variable at each level of the target variable. The annotations corresponding to the color of the target variable level show the mean/median ratio to help understand differences in skewness between the levels of the target variable.

Univariate Report

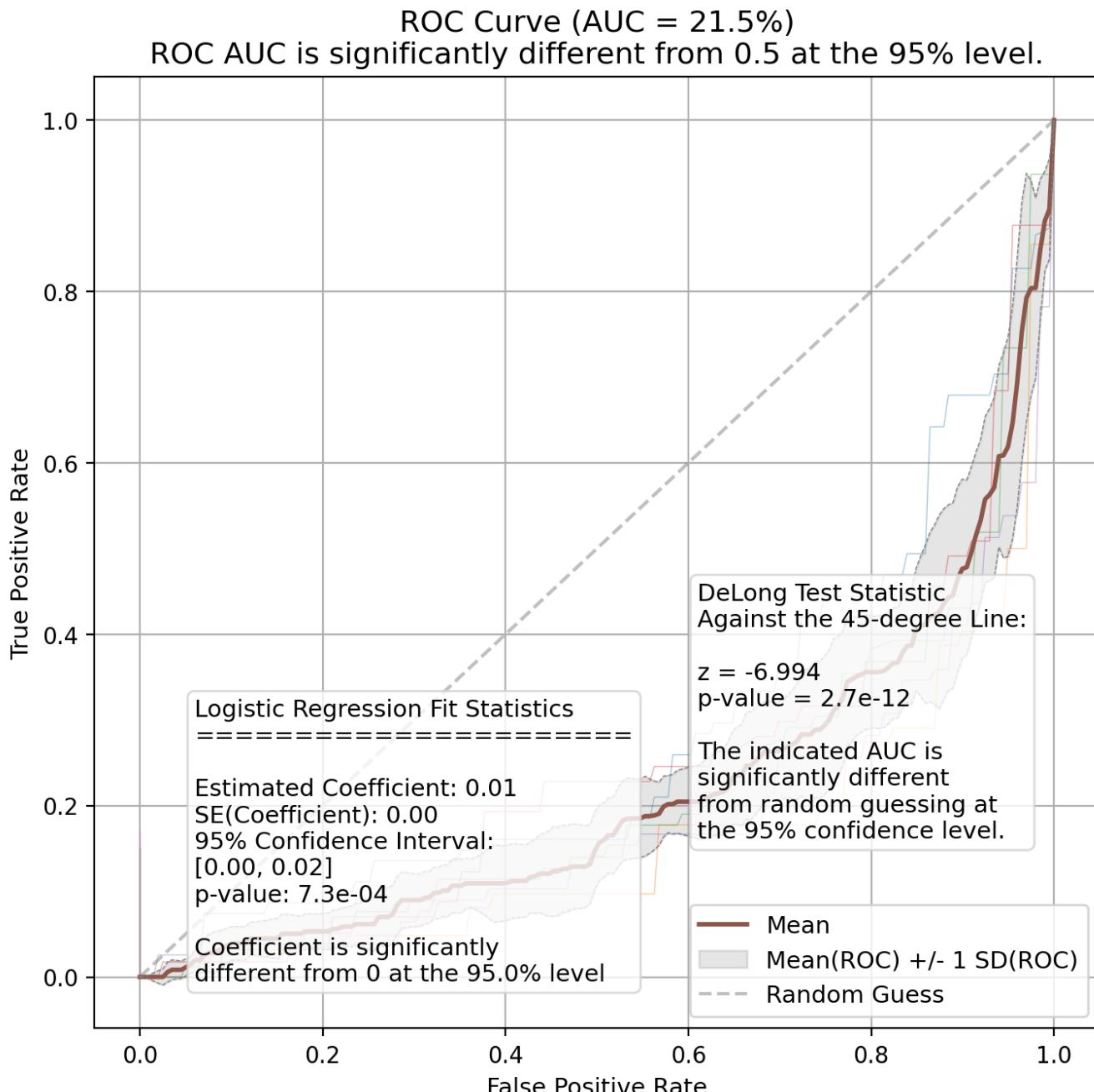
Worst Texture - Empirical CDF Plot



This plot shows the empirical cumulative distribution function for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the cumulative distribution of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data, and whether or not it is reasonable to assume that the data is drawn from different distributions.

Univariate Report

Worst Texture - ROC Curve

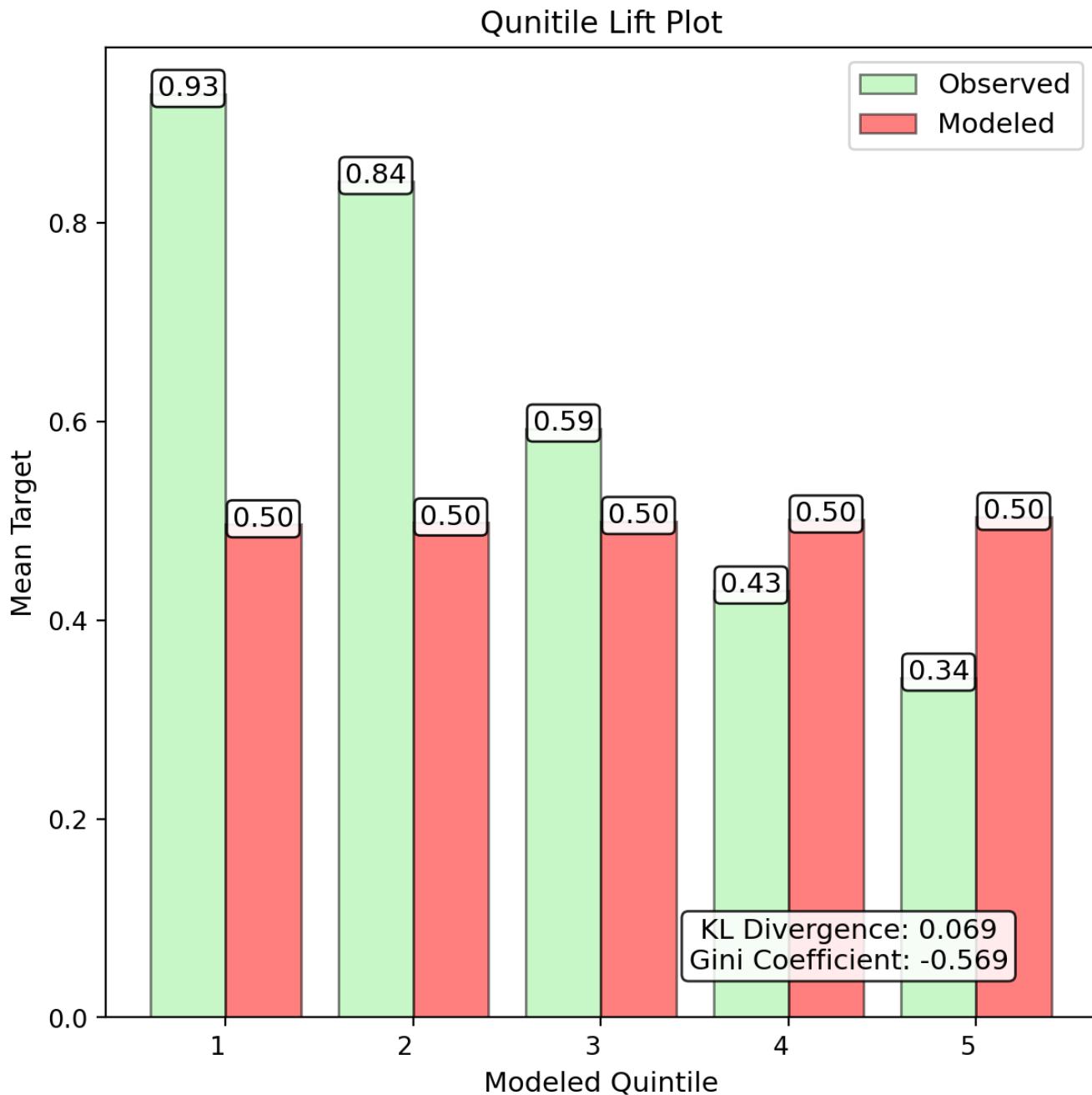


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, and no other features. Annotations are on the plot to help understand the results of the model, including the coefficient, standard error, and p-value for the feature variable. The cross-validation folds are used to create the grey region around the mean ROC curve to help understand the variability of the data.

Significance of the ROC curve is determined based on a modified version the method from DeLong et al. (1988). In brief, the AUC is assumed to be normally distributed, and I calculate the empirical standard error from the cross-validated AUC values. I then calculate a z-score for the AUC, and use the z-score to calculate a p-value. The p-value is then used to determine the significance of the AUC. This is a simple test, and should be used with caution.

Univariate Report

Worst Texture - Quintile Lift



The quintile lift plot is meant to show the power of the single feature to discriminate between the highest and lowest quintiles of the target variable.

Univariate Report

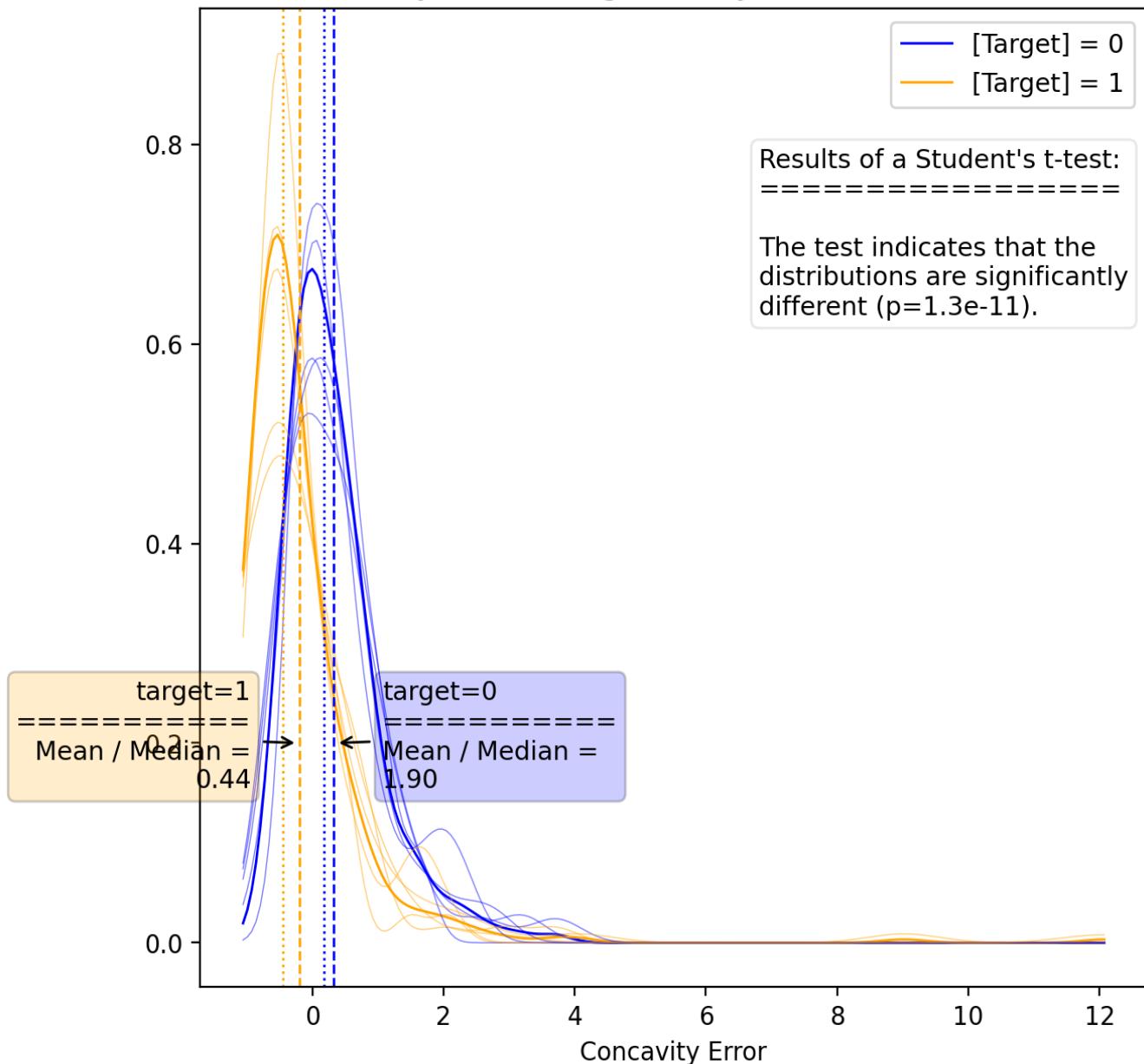
Concavity Error - Results

| | Fold-1 | Fold-2 | Fold-3 | Fold-4 | Fold-5 | Agg. Mean | Agg. SD |
|-------------------------|----------|----------|----------|----------|----------|-----------|---------|
| Fitted Coef. | -9.8e-01 | -6.5e-01 | -7.7e-01 | -7.0e-01 | -8.9e-01 | 7.4e-01 | 1.4e-01 |
| Fitted p-Value | 1.4e-09 | 2.4e-06 | 1.7e-07 | 6.1e-07 | 1.1e-08 | 7.0e-01 | 1.0e-06 |
| Fitted Std. Err. | 0.161 | 0.137 | 0.147 | 0.140 | 0.155 | 1.915 | 0.010 |
| Conf. Int. Lower | -1.3e+00 | -9.2e-01 | -1.1e+00 | -9.7e-01 | -1.2e+00 | -3.0e+00 | 1.6e-01 |
| Conf. Int. Upper | -6.6e-01 | -3.8e-01 | -4.8e-01 | -4.2e-01 | -5.8e-01 | 4.5e+00 | 1.2e-01 |
| Train Accuracy | 71.2% | 71.6% | 71.3% | 71.2% | 71.0% | 29.0% | 0.2% |
| Val Accuracy | 71.8% | 69.3% | 69.9% | 72.0% | 72.5% | 29.0% | 1.4% |
| Train AUC | 69.6% | 69.7% | 69.3% | 69.1% | 69.0% | 31.0% | 0.3% |
| Val AUC | 69.1% | 66.4% | 68.4% | 71.2% | 71.2% | 31.0% | 2.0% |
| Train F1 | 76.7% | 77.3% | 76.9% | 77.3% | 77.0% | 29.1% | 0.3% |
| Test F1 | 78.3% | 76.0% | 77.0% | 75.9% | 77.2% | 29.1% | 1.0% |
| Train Precision | 77.0% | 77.7% | 75.4% | 78.0% | 77.7% | 39.0% | 1.0% |
| Val Precision | 78.8% | 73.1% | 82.6% | 74.6% | 76.2% | 39.0% | 3.7% |
| Train Recall | 76.4% | 76.9% | 78.4% | 76.7% | 76.3% | 23.2% | 0.8% |
| Val Recall | 77.8% | 79.0% | 72.2% | 77.2% | 78.2% | 23.2% | 2.7% |
| Train MCC | 39.1% | 39.2% | 39.0% | 38.0% | 37.7% | -38.0% | 0.7% |
| Val MCC | 38.0% | 33.9% | 34.7% | 42.6% | 42.6% | -38.0% | 4.2% |
| Train Log-Loss | 10.37 | 10.24 | 10.35 | 10.38 | 10.45 | 25.59 | 0.07 |
| Val Log-Loss | 10.17 | 11.06 | 10.84 | 10.09 | 9.91 | 25.59 | 0.51 |

Univariate Report

Concavity Error - Kernel Density Plot

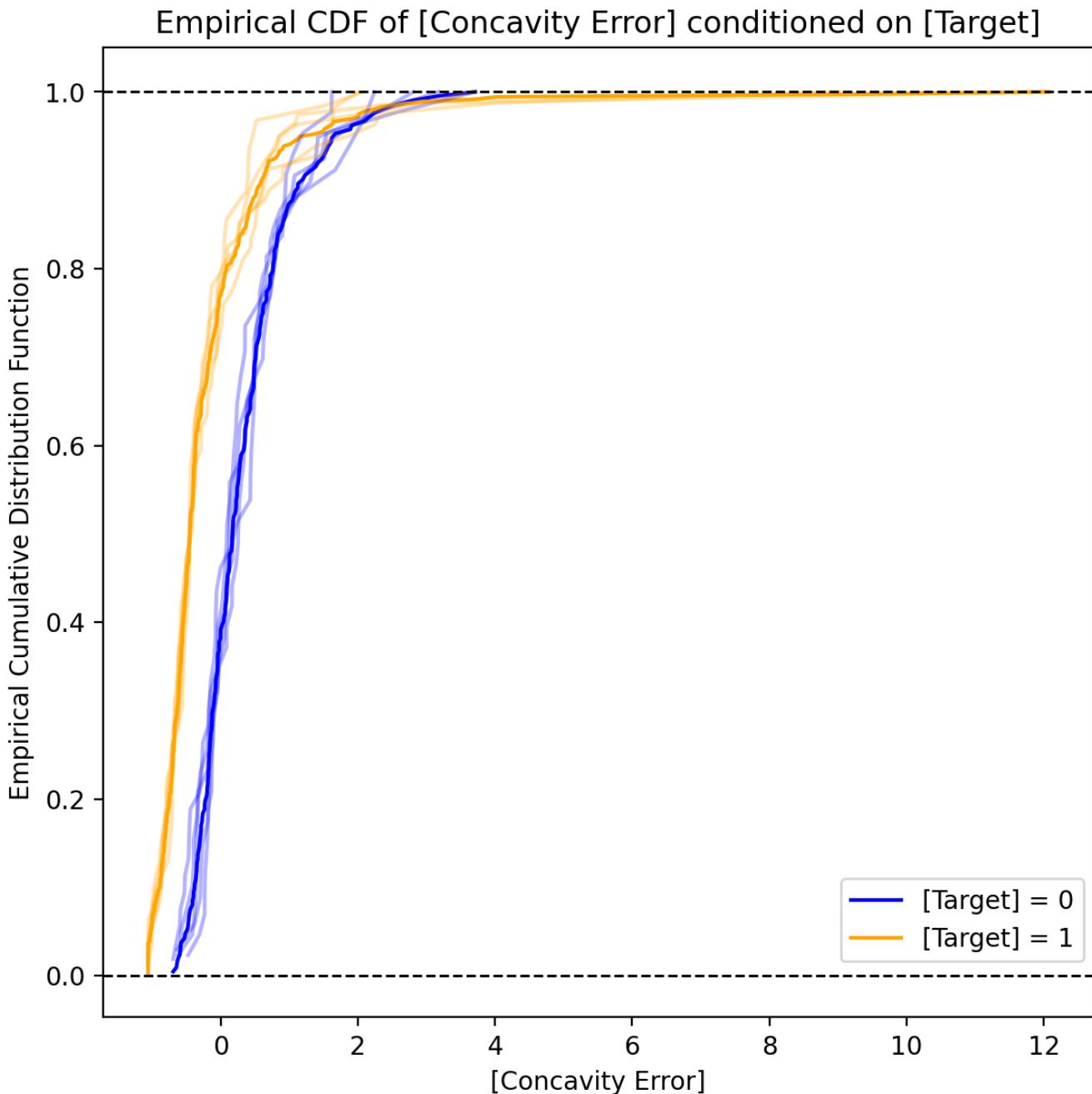
Kernel Density Plot of [Concavity Error] by [Target]
Distributions by level are significantly different at the 95% level.



This plot shows the Gaussian kernel density for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the density of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data. There are annotations with the results of a t-test for the difference in means between the feature variable at each level of the target variable. The annotations corresponding to the color of the target variable level show the mean/median ratio to help understand differences in skewness between the levels of the target variable.

Univariate Report

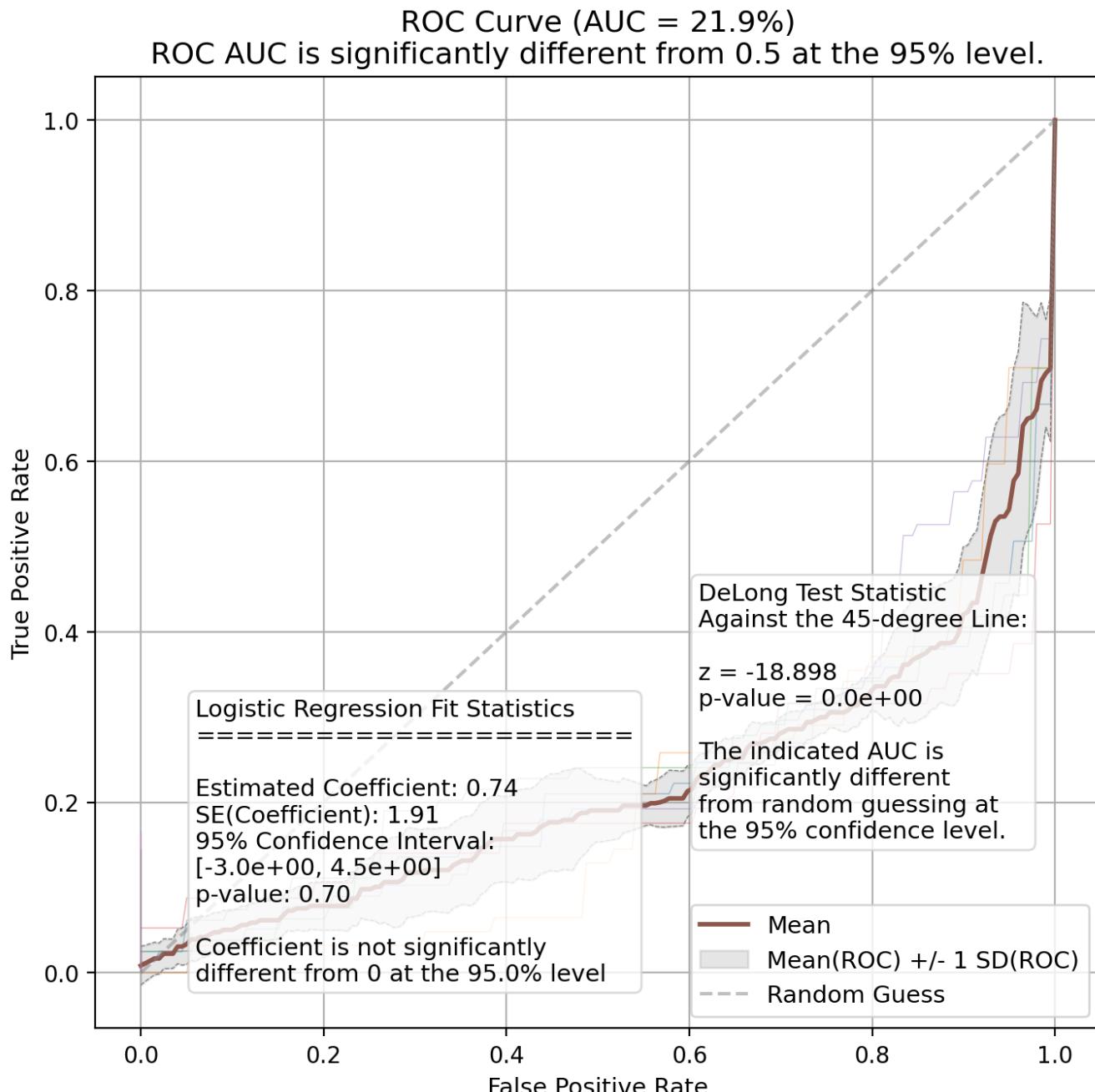
Concavity Error - Empirical CDF Plot



This plot shows the empirical cumulative distribution function for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the cumulative distribution of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data, and whether or not it is reasonable to assume that the data is drawn from different distributions.

Univariate Report

Concavity Error - ROC Curve

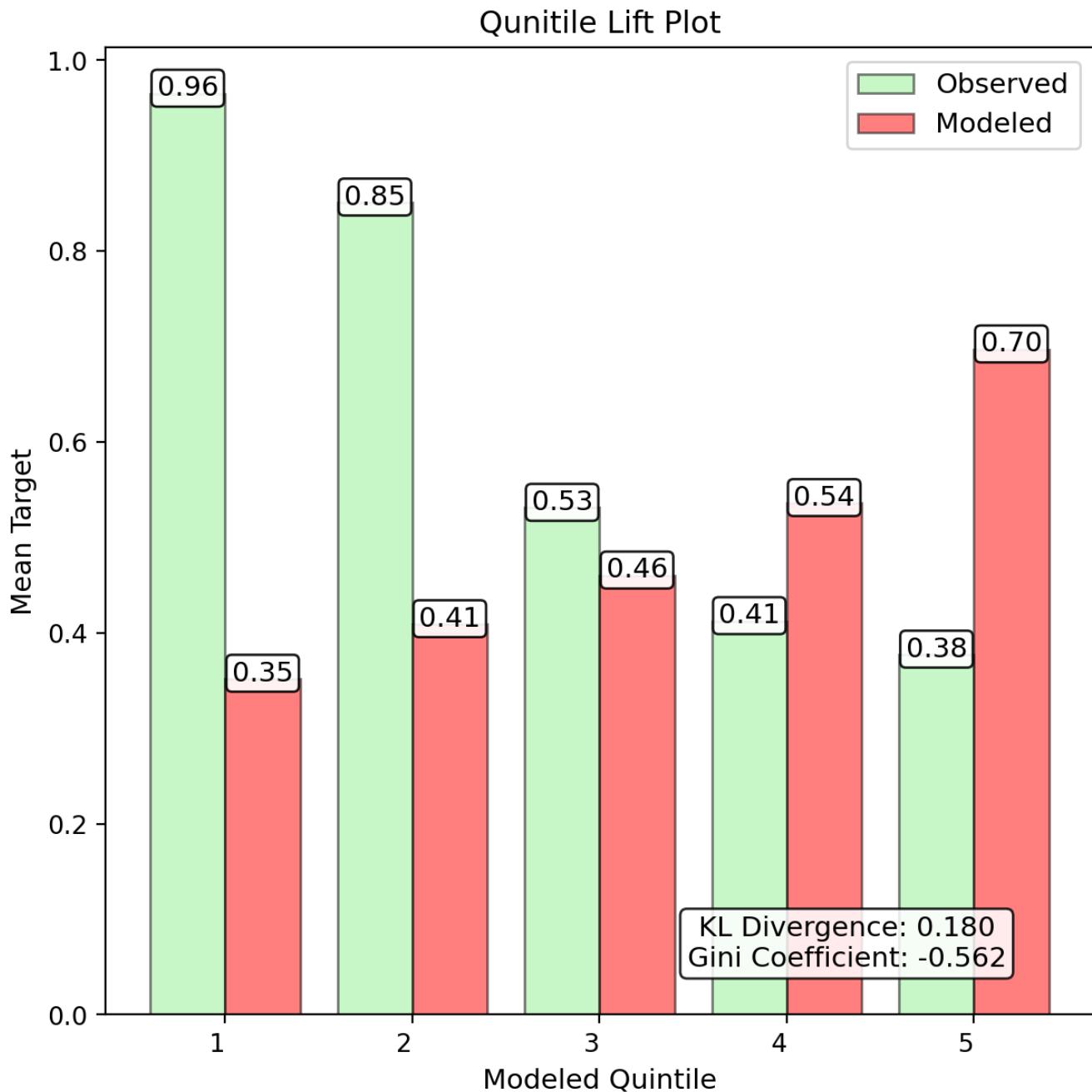


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, and no other features. Annotations are on the plot to help understand the results of the model, including the coefficient, standard error, and p-value for the feature variable. The cross-validation folds are used to create the grey region around the mean ROC curve to help understand the variability of the data.

Significance of the ROC curve is determined based on a modified version the method from DeLong et al. (1988). In brief, the AUC is assumed to be normally distributed, and I calculate the empirical standard error from the cross-validated AUC values. I then calculate a z-score for the AUC, and use the z-score to calculate a p-value. The p-value is then used to determine the significance of the AUC. This is a simple test, and should be used with caution.

Univariate Report

Concavity Error - Quintile Lift



The quintile lift plot is meant to show the power of the single feature to discriminate between the highest and lowest quintiles of the target variable.

Univariate Report

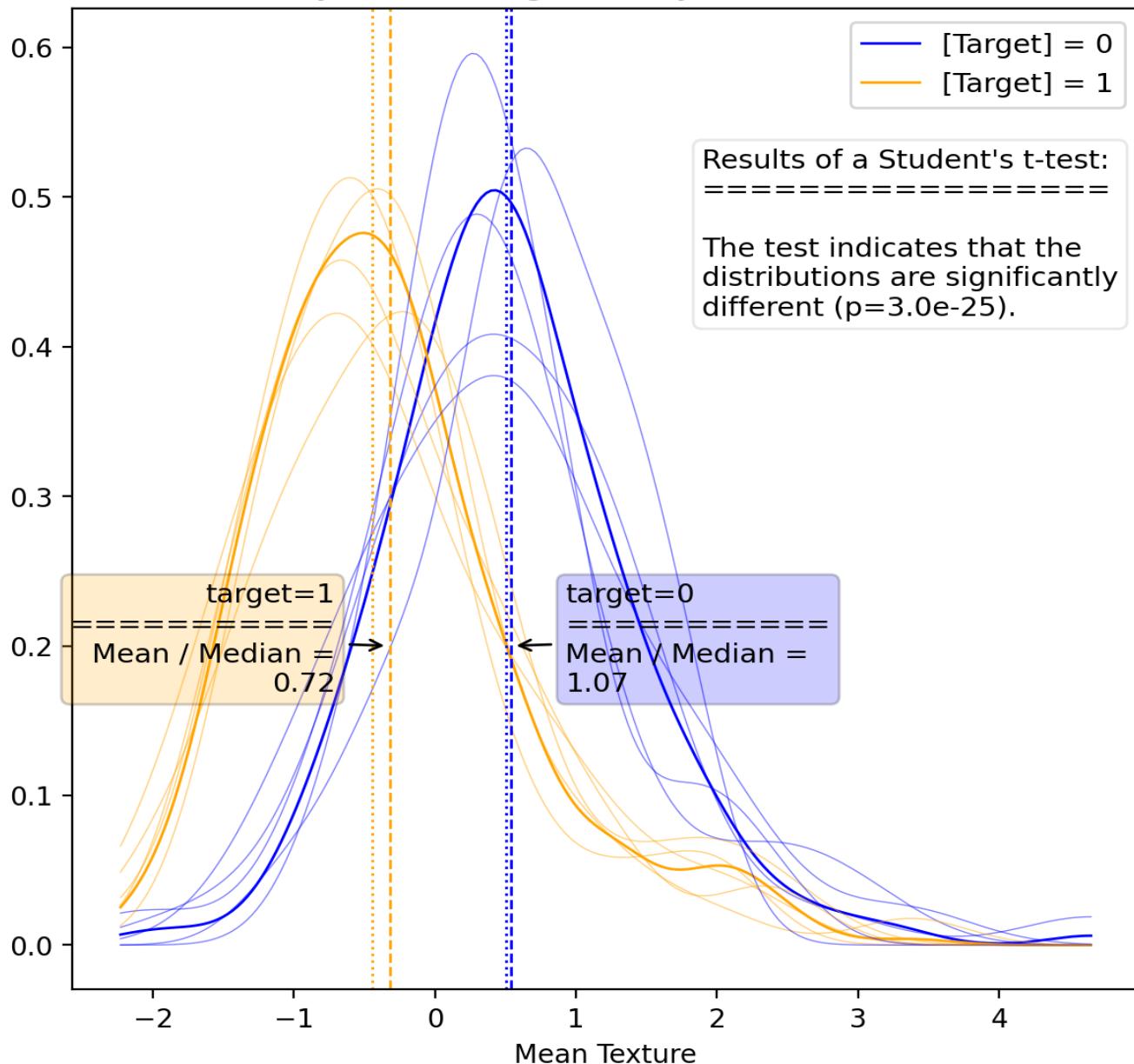
Mean Texture - Results

| | Fold-1 | Fold-2 | Fold-3 | Fold-4 | Fold-5 | Agg. Mean | Agg. SD |
|-------------------------|----------|----------|----------|----------|----------|-----------|---------|
| Fitted Coef. | -1.1e+00 | -9.7e-01 | -9.3e-01 | -9.3e-01 | -8.9e-01 | 1.6e-02 | 6.7e-02 |
| Fitted p-Value | 2.7e-16 | 2.2e-15 | 3.0e-14 | 2.6e-15 | 4.1e-13 | 1.7e-04 | 1.8e-13 |
| Fitted Std. Err. | 0.130 | 0.123 | 0.122 | 0.117 | 0.123 | 0.004 | 0.005 |
| Conf. Int. Lower | -1.3e+00 | -1.2e+00 | -1.2e+00 | -1.2e+00 | -1.1e+00 | 7.7e-03 | 7.4e-02 |
| Conf. Int. Upper | -8.1e-01 | -7.3e-01 | -6.9e-01 | -7.0e-01 | -6.5e-01 | 2.5e-02 | 6.0e-02 |
| Train Accuracy | 74.2% | 73.7% | 73.0% | 72.3% | 71.9% | 27.2% | 0.9% |
| Val Accuracy | 67.7% | 71.3% | 74.3% | 75.0% | 75.6% | 27.2% | 3.3% |
| Train AUC | 74.6% | 74.1% | 73.4% | 72.9% | 72.3% | 26.7% | 0.9% |
| Val AUC | 69.3% | 70.9% | 72.4% | 75.2% | 76.8% | 26.7% | 3.1% |
| Train F1 | 77.8% | 77.7% | 76.4% | 76.5% | 76.3% | 33.2% | 0.7% |
| Test F1 | 72.2% | 75.6% | 80.8% | 77.1% | 77.5% | 33.2% | 3.1% |
| Train Precision | 83.4% | 83.6% | 81.9% | 83.5% | 82.5% | 39.2% | 0.7% |
| Val Precision | 82.5% | 78.9% | 84.7% | 80.8% | 85.9% | 39.2% | 2.8% |
| Train Recall | 72.8% | 72.5% | 71.6% | 70.7% | 71.0% | 28.9% | 0.9% |
| Val Recall | 64.2% | 72.6% | 77.2% | 73.7% | 70.5% | 28.9% | 4.8% |
| Train MCC | 47.9% | 46.8% | 45.8% | 44.1% | 43.0% | -45.2% | 2.0% |
| Val MCC | 36.8% | 41.1% | 42.8% | 50.0% | 52.6% | -45.2% | 6.5% |
| Train Log-Loss | 9.31 | 9.47 | 9.72 | 9.99 | 10.12 | 26.23 | 0.34 |
| Val Log-Loss | 11.63 | 10.35 | 9.25 | 9.01 | 8.80 | 26.23 | 1.18 |

Univariate Report

Mean Texture - Kernel Density Plot

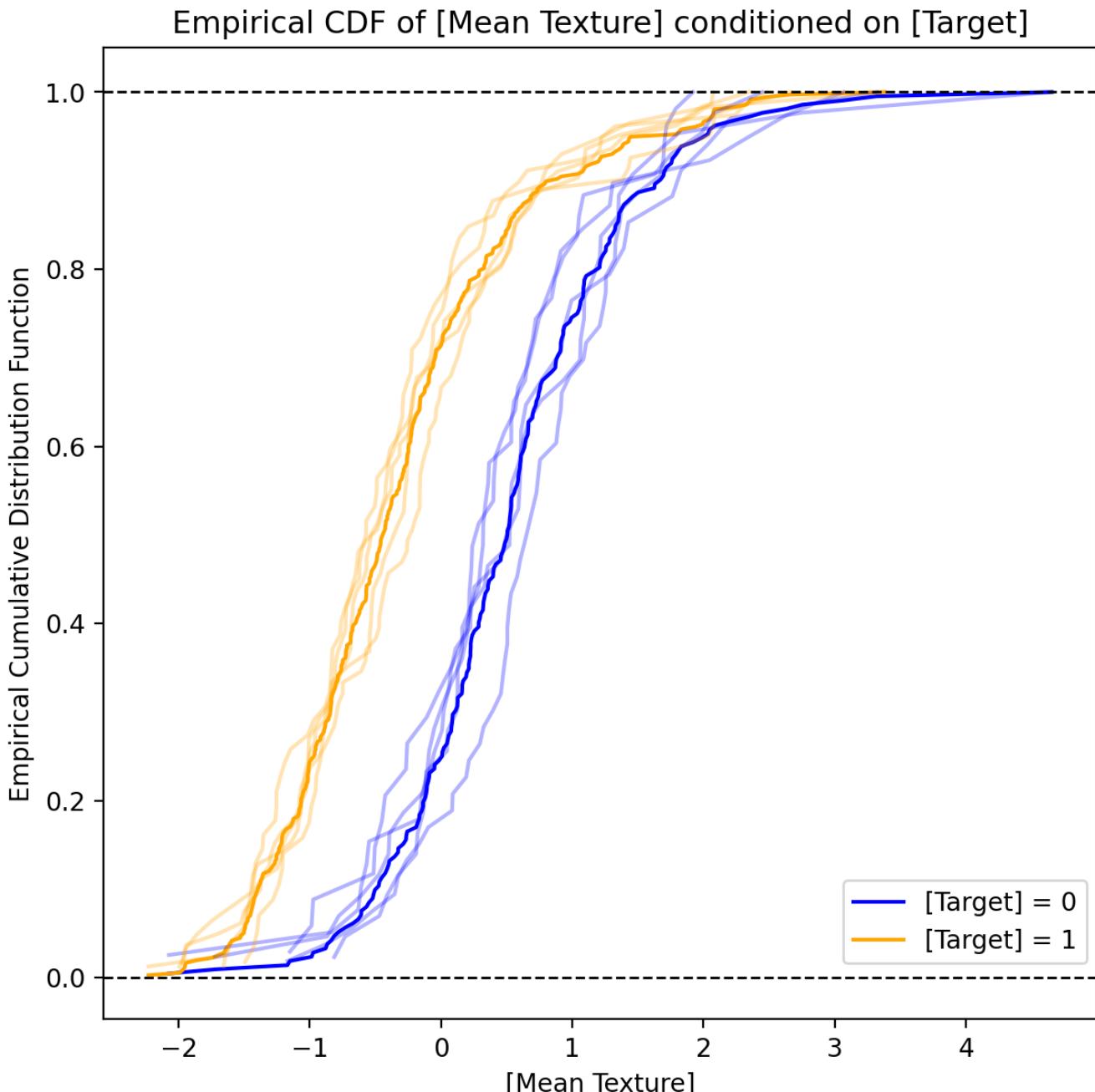
Kernel Density Plot of [Mean Texture] by [Target]
Distributions by level are significantly different at the 95% level.



This plot shows the Gaussian kernel density for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the density of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data. There are annotations with the results of a t-test for the difference in means between the feature variable at each level of the target variable. The annotations corresponding to the color of the target variable level show the mean/median ratio to help understand differences in skewness between the levels of the target variable.

Univariate Report

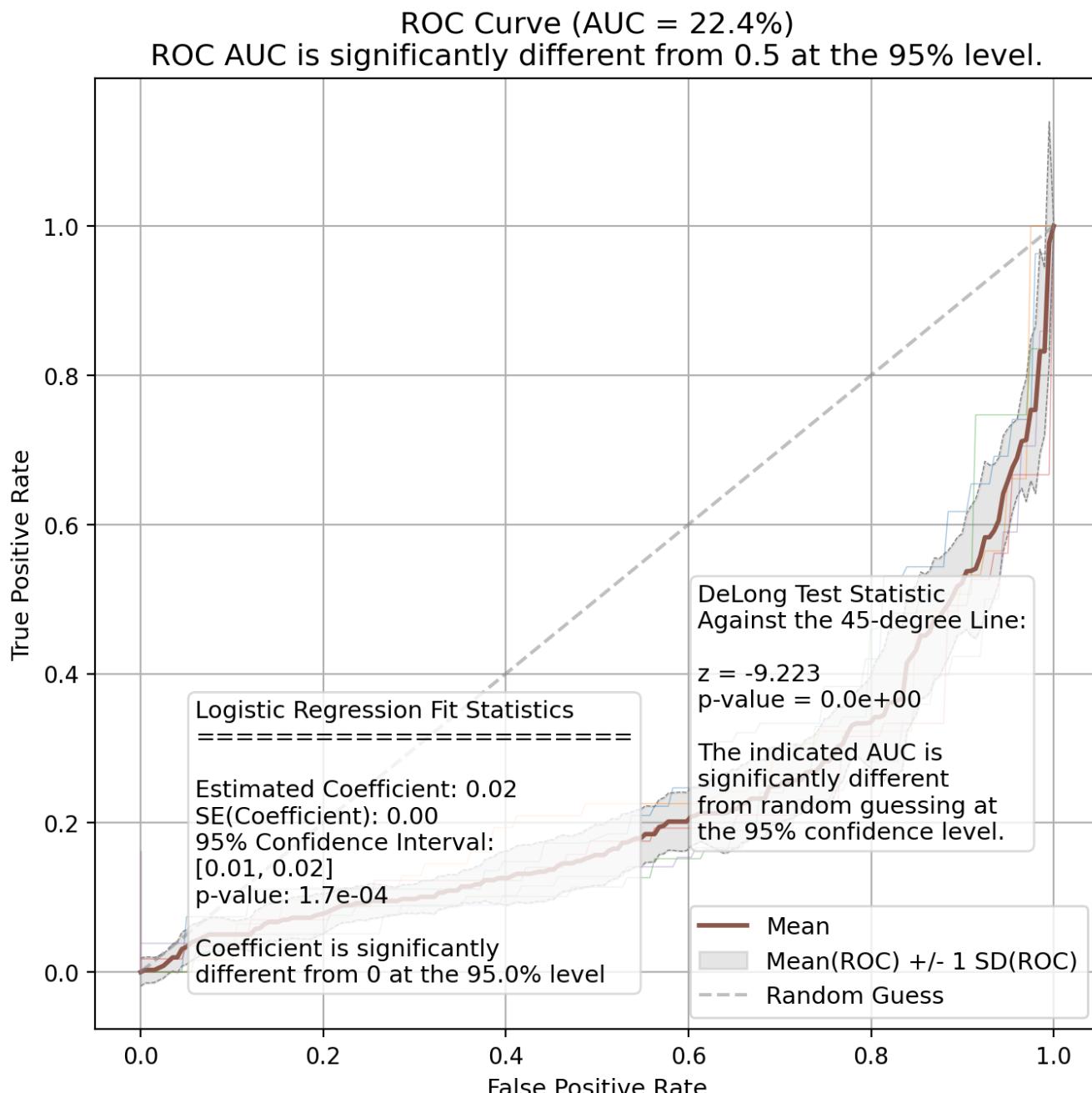
Mean Texture - Empirical CDF Plot



This plot shows the empirical cumulative distribution function for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the cumulative distribution of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data, and whether or not it is reasonable to assume that the data is drawn from different distributions.

Univariate Report

Mean Texture - ROC Curve

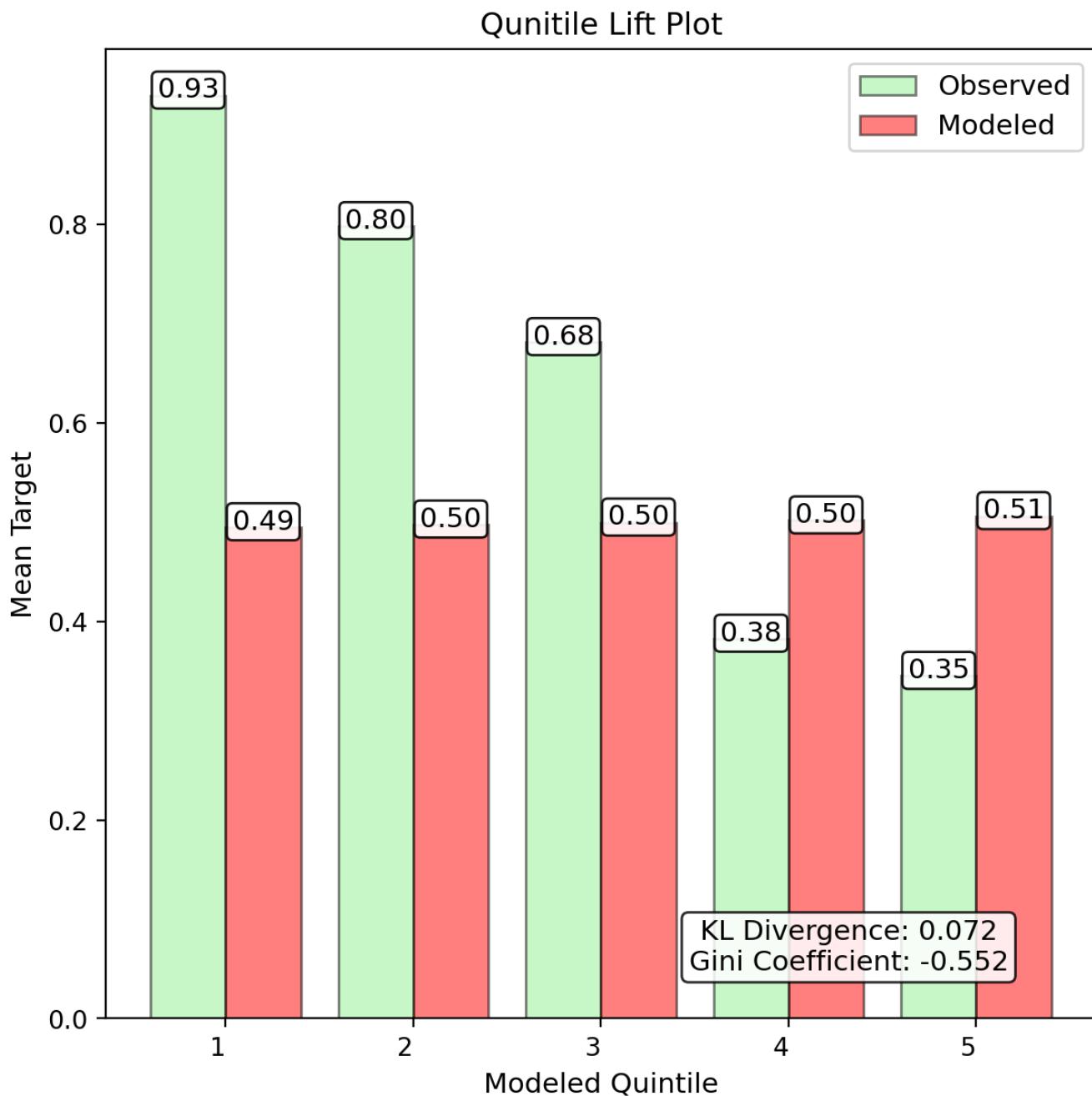


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, and no other features. Annotations are on the plot to help understand the results of the model, including the coefficient, standard error, and p-value for the feature variable. The cross-validation folds are used to create the grey region around the mean ROC curve to help understand the variability of the data.

Significance of the ROC curve is determined based on a modified version the method from DeLong et al. (1988). In brief, the AUC is assumed to be normally distributed, and I calculate the empirical standard error from the cross-validated AUC values. I then calculate a z-score for the AUC, and use the z-score to calculate a p-value. The p-value is then used to determine the significance of the AUC. This is a simple test, and should be used with caution.

Univariate Report

Mean Texture - Quintile Lift



The quintile lift plot is meant to show the power of the single feature to discriminate between the highest and lowest quintiles of the target variable.

Univariate Report

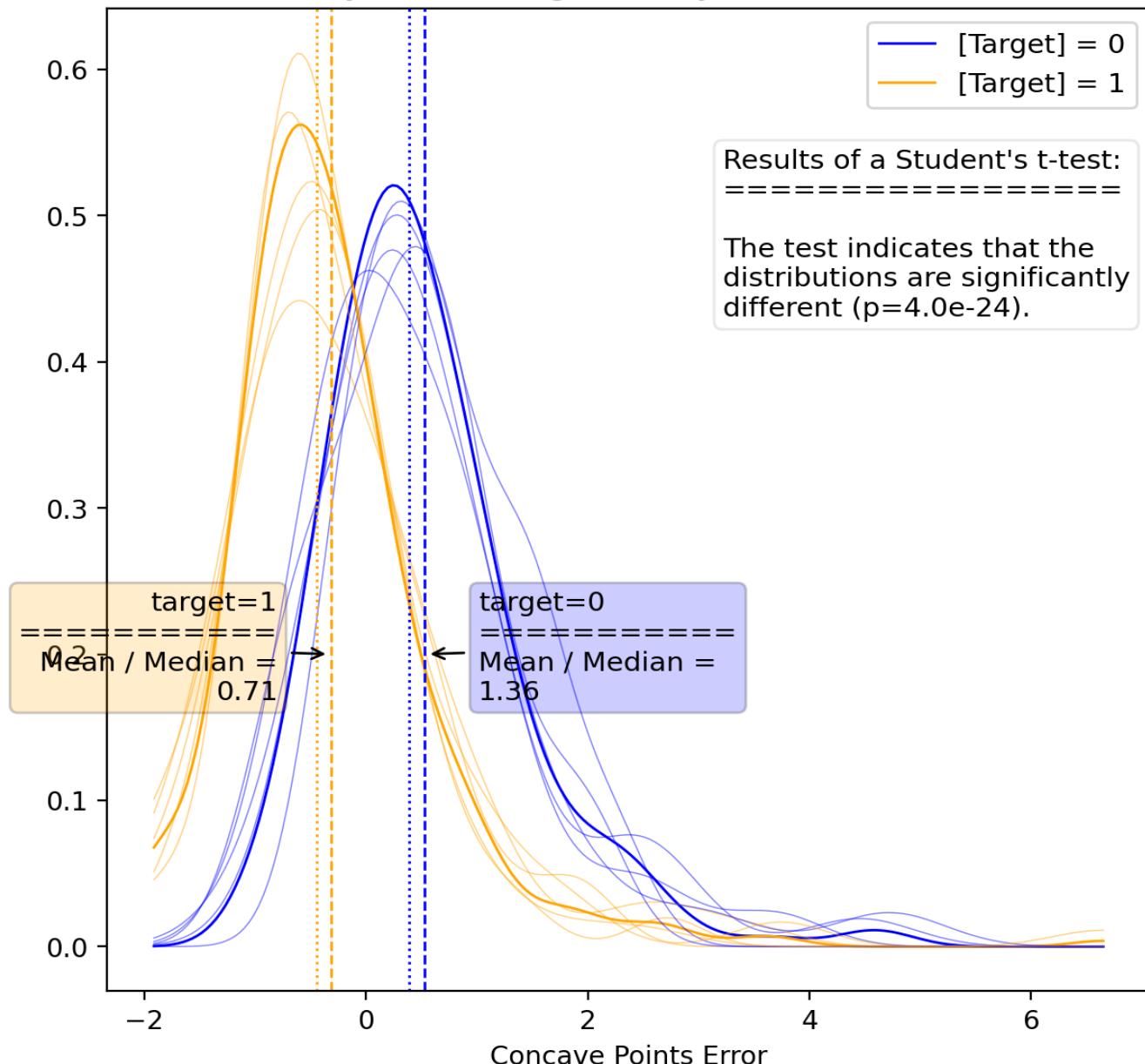
Concave Points Error - Results

| | Fold-1 | Fold-2 | Fold-3 | Fold-4 | Fold-5 | Agg. Mean | Agg. SD |
|-------------------------|----------|----------|----------|----------|----------|-----------|---------|
| Fitted Coef. | -1.2e+00 | -1.0e+00 | -1.0e+00 | -1.0e+00 | -1.1e+00 | 6.4e+00 | 6.8e-02 |
| Fitted p-Value | 2.1e-15 | 2.3e-14 | 3.9e-14 | 3.8e-14 | 2.2e-14 | 3.1e-01 | 1.5e-14 |
| Fitted Std. Err. | 0.148 | 0.133 | 0.135 | 0.135 | 0.143 | 6.313 | 0.006 |
| Conf. Int. Lower | -1.5e+00 | -1.3e+00 | -1.3e+00 | -1.3e+00 | -1.4e+00 | -6.0e+00 | 8.0e-02 |
| Conf. Int. Upper | -8.8e-01 | -7.5e-01 | -7.6e-01 | -7.6e-01 | -8.1e-01 | 1.9e+01 | 5.6e-02 |
| Train Accuracy | 75.3% | 74.4% | 73.2% | 71.4% | 73.1% | 26.9% | 1.5% |
| Val Accuracy | 66.9% | 70.3% | 74.3% | 79.0% | 75.6% | 26.9% | 4.7% |
| Train AUC | 74.9% | 74.3% | 72.8% | 71.3% | 72.9% | 27.1% | 1.4% |
| Val AUC | 68.1% | 68.2% | 74.9% | 79.6% | 74.3% | 27.1% | 4.9% |
| Train F1 | 79.3% | 78.6% | 77.3% | 76.2% | 77.7% | 31.1% | 1.2% |
| Test F1 | 71.7% | 76.2% | 80.0% | 80.4% | 79.7% | 31.1% | 3.7% |
| Train Precision | 82.4% | 83.0% | 80.0% | 81.4% | 82.3% | 38.1% | 1.2% |
| Val Precision | 81.2% | 75.0% | 87.9% | 86.0% | 78.8% | 38.1% | 5.3% |
| Train Recall | 76.4% | 74.6% | 74.8% | 71.7% | 73.5% | 26.3% | 1.8% |
| Val Recall | 64.2% | 77.4% | 73.4% | 75.4% | 80.8% | 26.3% | 6.2% |
| Train MCC | 48.9% | 47.3% | 44.9% | 41.3% | 44.5% | -44.7% | 2.9% |
| Val MCC | 34.6% | 36.8% | 46.4% | 58.6% | 49.0% | -44.7% | 9.7% |
| Train Log-Loss | 8.91 | 9.24 | 9.64 | 10.30 | 9.71 | 26.35 | 0.52 |
| Val Log-Loss | 11.92 | 10.71 | 9.25 | 7.57 | 8.80 | 26.35 | 1.69 |

Univariate Report

Concave Points Error - Kernel Density Plot

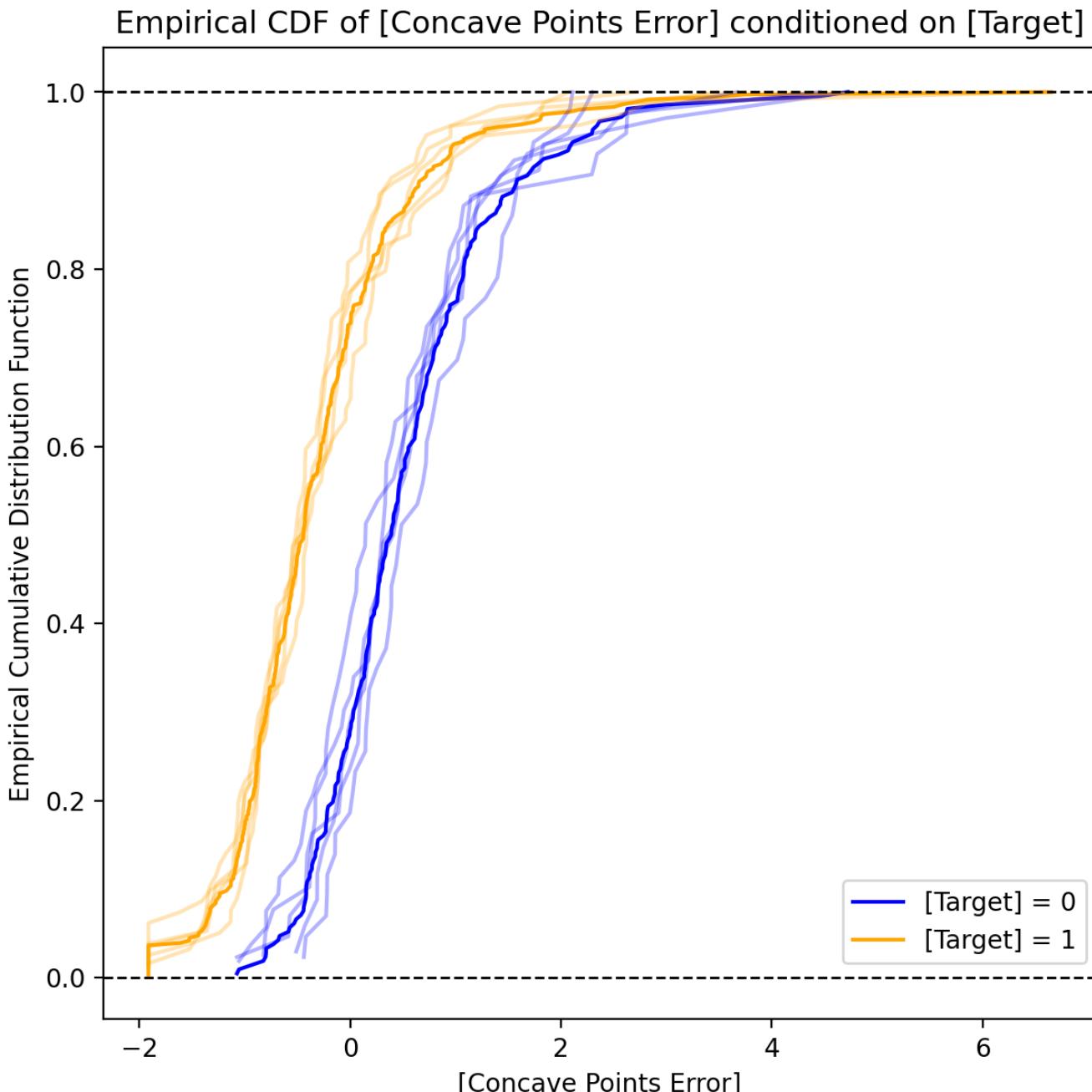
Kernel Density Plot of [Concave Points Error] by [Target]
Distributions by level are significantly different at the 95% level.



This plot shows the Gaussian kernel density for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the density of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data. There are annotations with the results of a t-test for the difference in means between the feature variable at each level of the target variable. The annotations corresponding to the color of the target variable level show the mean/median ratio to help understand differences in skewness between the levels of the target variable.

Univariate Report

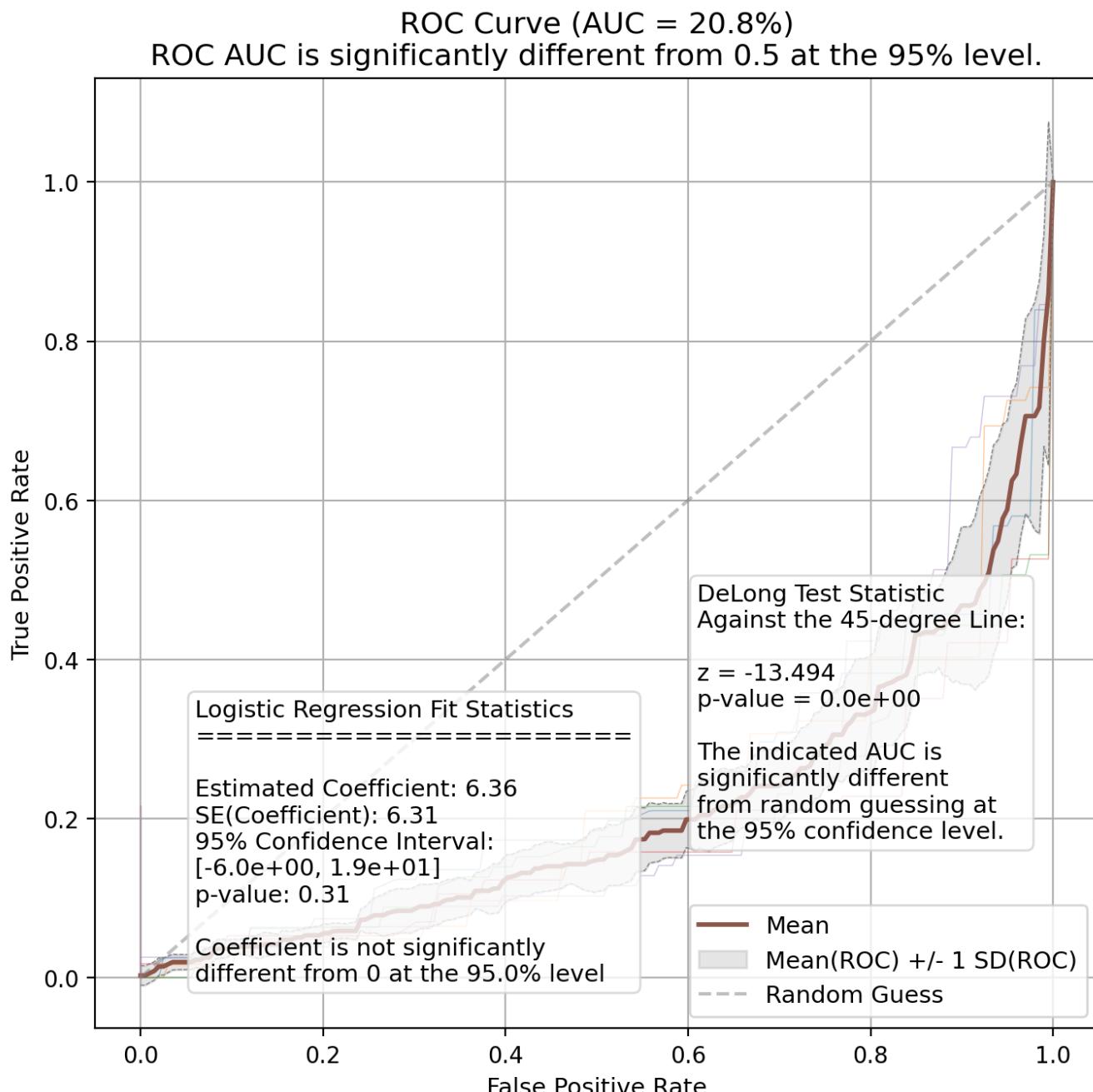
Concave Points Error - Empirical CDF Plot



This plot shows the empirical cumulative distribution function for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the cumulative distribution of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data, and whether or not it is reasonable to assume that the data is drawn from different distributions.

Univariate Report

Concave Points Error - ROC Curve

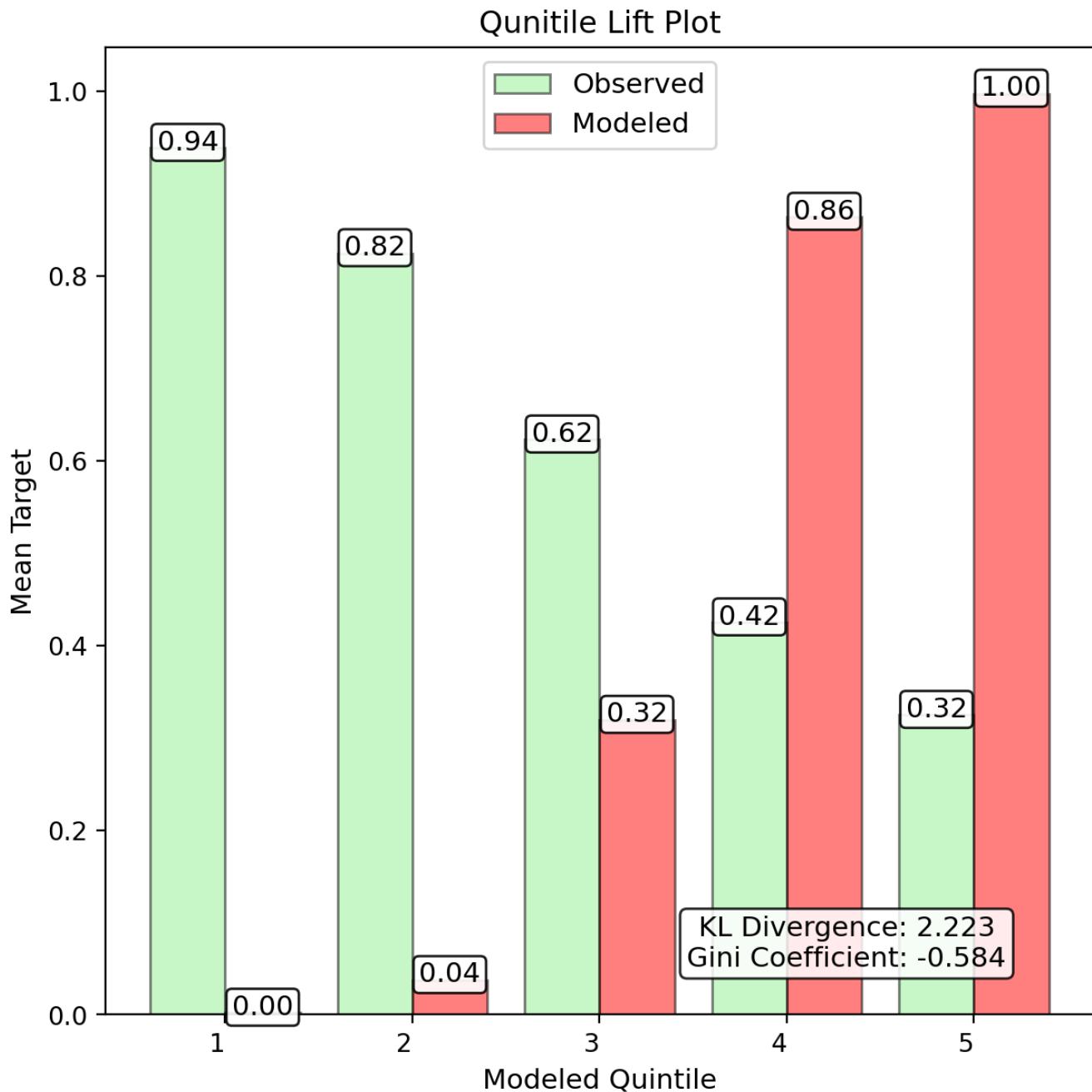


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, and no other features. Annotations are on the plot to help understand the results of the model, including the coefficient, standard error, and p-value for the feature variable. The cross-validation folds are used to create the grey region around the mean ROC curve to help understand the variability of the data.

Significance of the ROC curve is determined based on a modified version the method from DeLong et al. (1988). In brief, the AUC is assumed to be normally distributed, and I calculate the empirical standard error from the cross-validated AUC values. I then calculate a z-score for the AUC, and use the z-score to calculate a p-value. The p-value is then used to determine the significance of the AUC. This is a simple test, and should be used with caution.

Univariate Report

Concave Points Error - Quintile Lift



The quintile lift plot is meant to show the power of the single feature to discriminate between the highest and lowest quintiles of the target variable.

Univariate Report

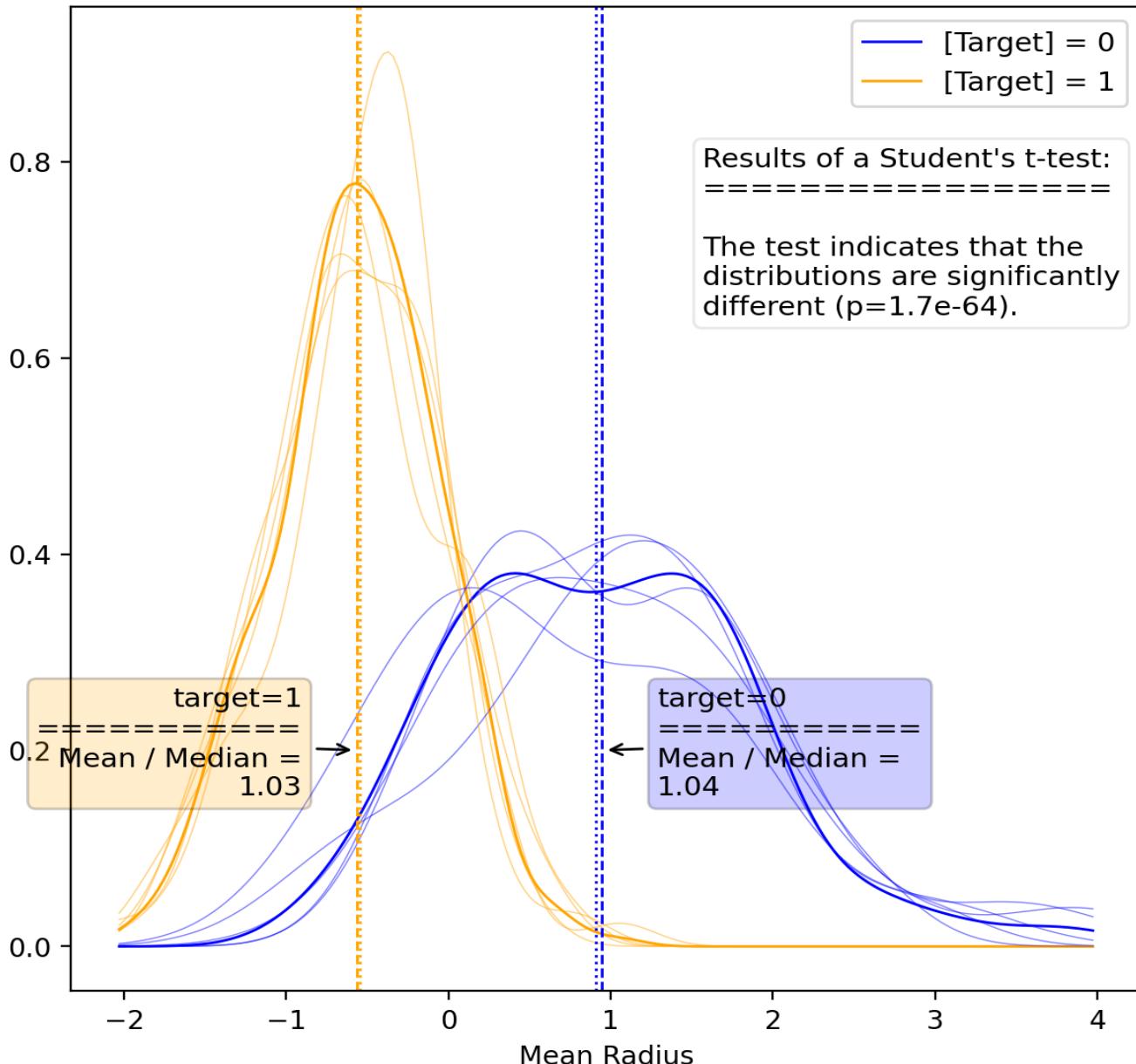
Mean Radius - Results

| | Fold-1 | Fold-2 | Fold-3 | Fold-4 | Fold-5 | Agg. Mean | Agg. SD |
|-------------------------|----------|----------|----------|----------|----------|-----------|---------|
| Fitted Coef. | -3.4e+00 | -3.6e+00 | -3.6e+00 | -3.3e+00 | -3.5e+00 | 1.0e-02 | 1.2e-01 |
| Fitted p-Value | 1.3e-25 | 1.5e-26 | 2.2e-25 | 3.8e-27 | 7.7e-25 | 7.4e-02 | 3.2e-25 |
| Fitted Std. Err. | 0.324 | 0.341 | 0.342 | 0.310 | 0.336 | 0.006 | 0.014 |
| Conf. Int. Lower | -4.0e+00 | -4.3e+00 | -4.2e+00 | -4.0e+00 | -4.1e+00 | -1.0e-03 | 1.4e-01 |
| Conf. Int. Upper | -2.8e+00 | -3.0e+00 | -2.9e+00 | -2.7e+00 | -2.8e+00 | 2.2e-02 | 9.7e-02 |
| Train Accuracy | 86.3% | 86.5% | 86.0% | 84.4% | 86.8% | 13.7% | 0.9% |
| Val Accuracy | 86.3% | 85.1% | 86.7% | 89.0% | 84.0% | 13.7% | 1.9% |
| Train AUC | 85.9% | 86.6% | 85.7% | 84.5% | 86.1% | 14.0% | 0.8% |
| Val AUC | 86.8% | 83.1% | 86.3% | 88.6% | 84.7% | 14.0% | 2.1% |
| Train F1 | 88.8% | 89.0% | 88.3% | 87.4% | 89.5% | 15.8% | 0.8% |
| Test F1 | 89.0% | 88.4% | 90.2% | 90.4% | 85.7% | 15.8% | 1.9% |
| Train Precision | 90.0% | 91.7% | 89.6% | 90.7% | 90.5% | 20.4% | 0.8% |
| Val Precision | 93.2% | 85.1% | 93.2% | 89.7% | 91.3% | 20.4% | 3.4% |
| Train Recall | 87.7% | 86.4% | 87.1% | 84.3% | 88.5% | 12.9% | 1.6% |
| Val Recall | 85.2% | 91.9% | 87.3% | 91.2% | 80.8% | 12.9% | 4.6% |
| Train MCC | 71.2% | 71.9% | 70.8% | 67.4% | 71.6% | -71.2% | 1.8% |
| Val MCC | 71.4% | 68.3% | 70.1% | 77.5% | 68.3% | -71.2% | 3.8% |
| Train Log-Loss | 4.94 | 4.85 | 5.06 | 5.61 | 4.77 | 31.10 | 0.33 |
| Val Log-Loss | 4.94 | 5.35 | 4.78 | 3.96 | 5.78 | 31.10 | 0.68 |

Univariate Report

Mean Radius - Kernel Density Plot

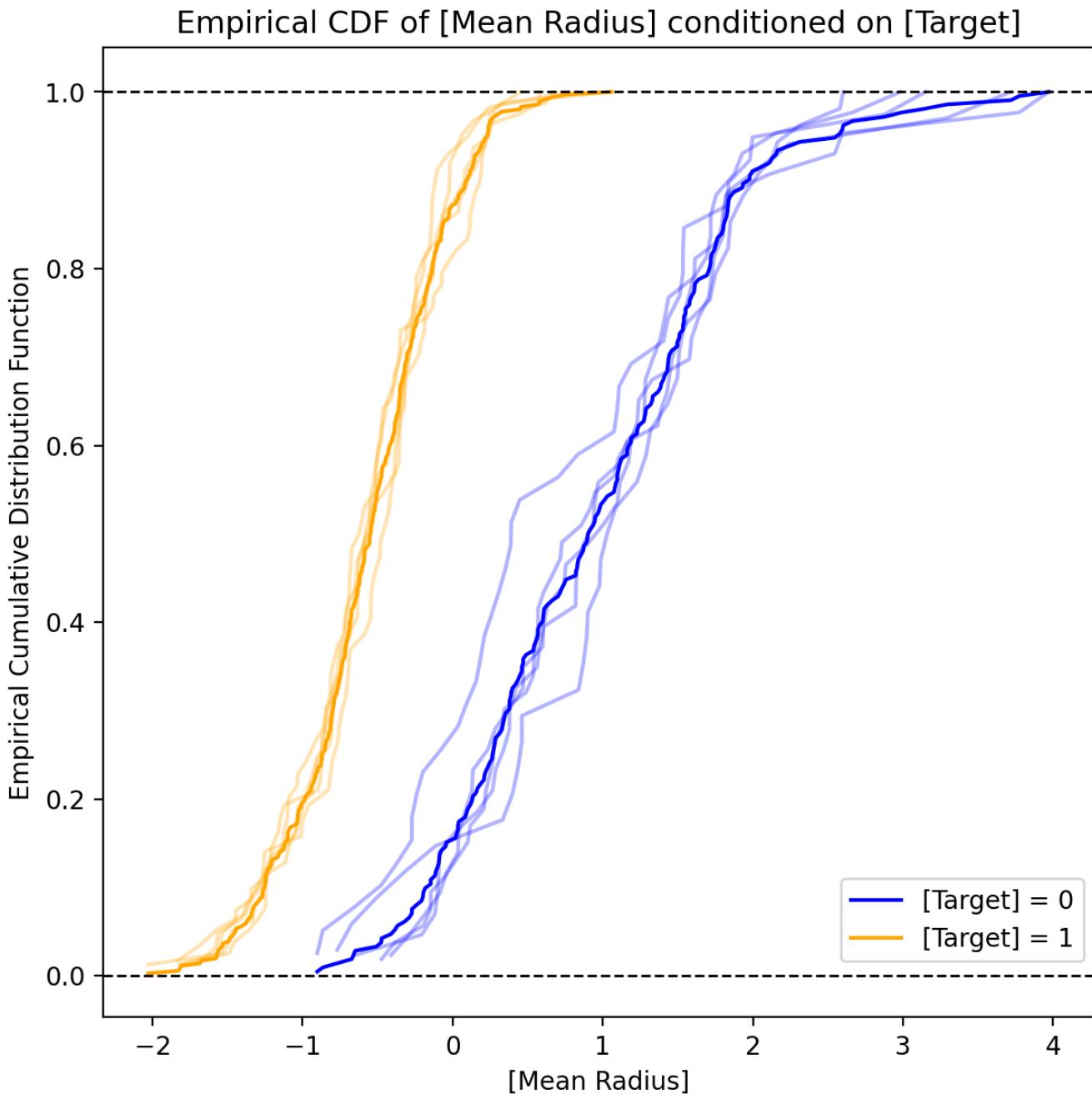
Kernel Density Plot of [Mean Radius] by [Target]
Distributions by level are significantly different at the 95% level.



This plot shows the Gaussian kernel density for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the density of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data. There are annotations with the results of a t-test for the difference in means between the feature variable at each level of the target variable. The annotations corresponding to the color of the target variable level show the mean/median ratio to help understand differences in skewness between the levels of the target variable.

Univariate Report

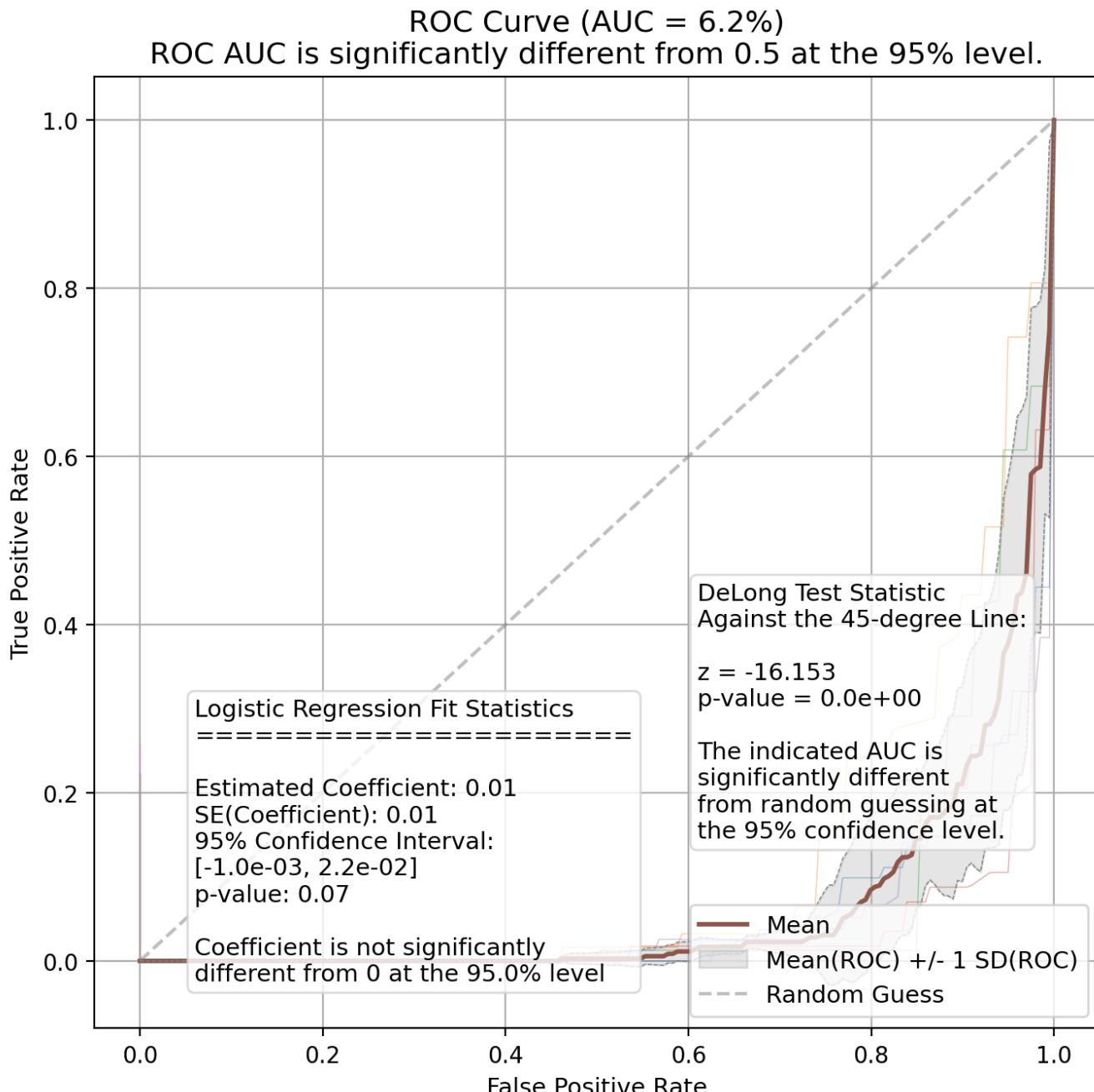
Mean Radius - Empirical CDF Plot



This plot shows the empirical cumulative distribution function for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the cumulative distribution of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data, and whether or not it is reasonable to assume that the data is drawn from different distributions.

Univariate Report

Mean Radius - ROC Curve

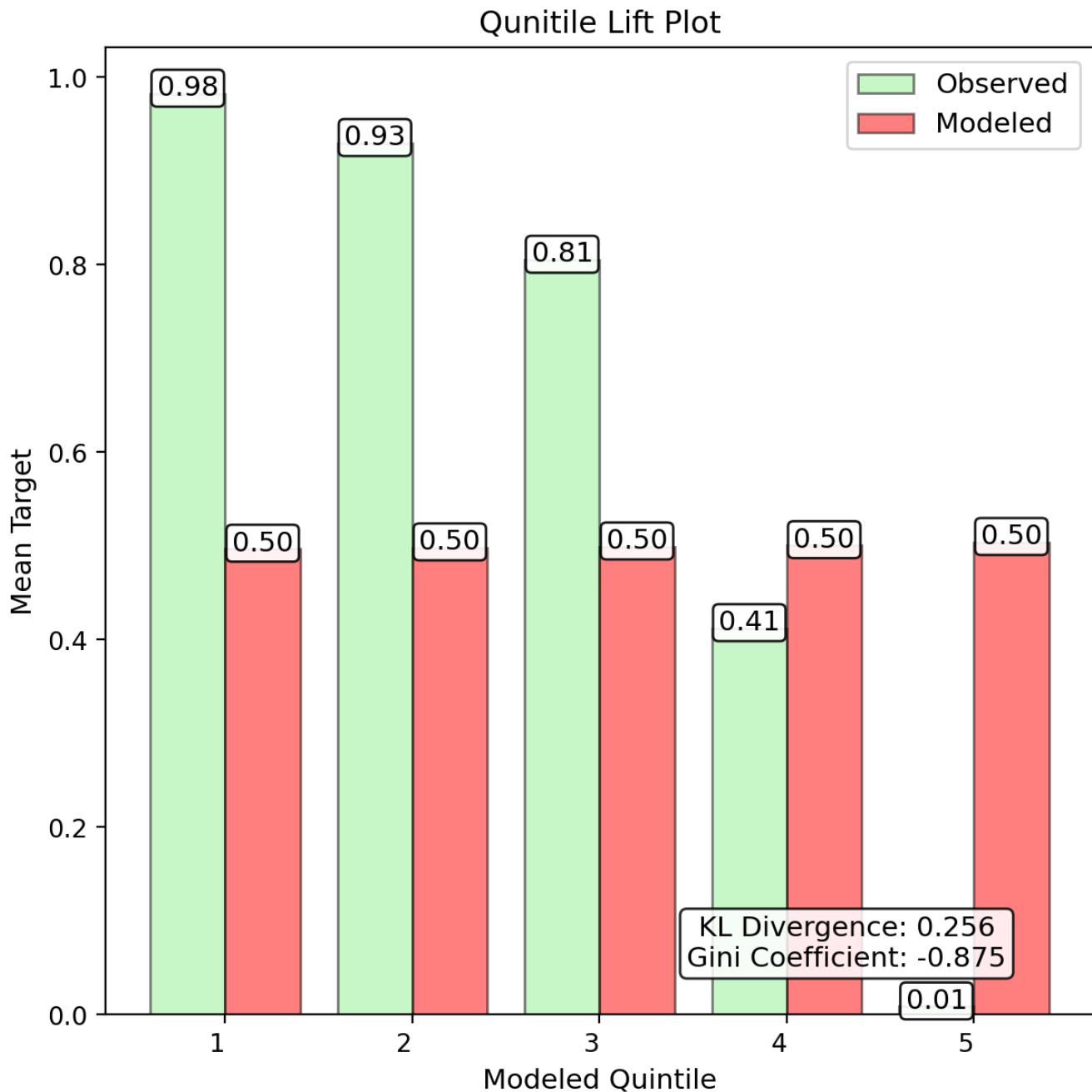


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, and no other features. Annotations are on the plot to help understand the results of the model, including the coefficient, standard error, and p-value for the feature variable. The cross-validation folds are used to create the grey region around the mean ROC curve to help understand the variability of the data.

Significance of the ROC curve is determined based on a modified version the method from DeLong et al. (1988). In brief, the AUC is assumed to be normally distributed, and I calculate the empirical standard error from the cross-validated AUC values. I then calculate a z-score for the AUC, and use the z-score to calculate a p-value. The p-value is then used to determine the significance of the AUC. This is a simple test, and should be used with caution.

Univariate Report

Mean Radius - Quintile Lift



The quintile lift plot is meant to show the power of the single feature to discriminate between the highest and lowest quintiles of the target variable.

Univariate Report

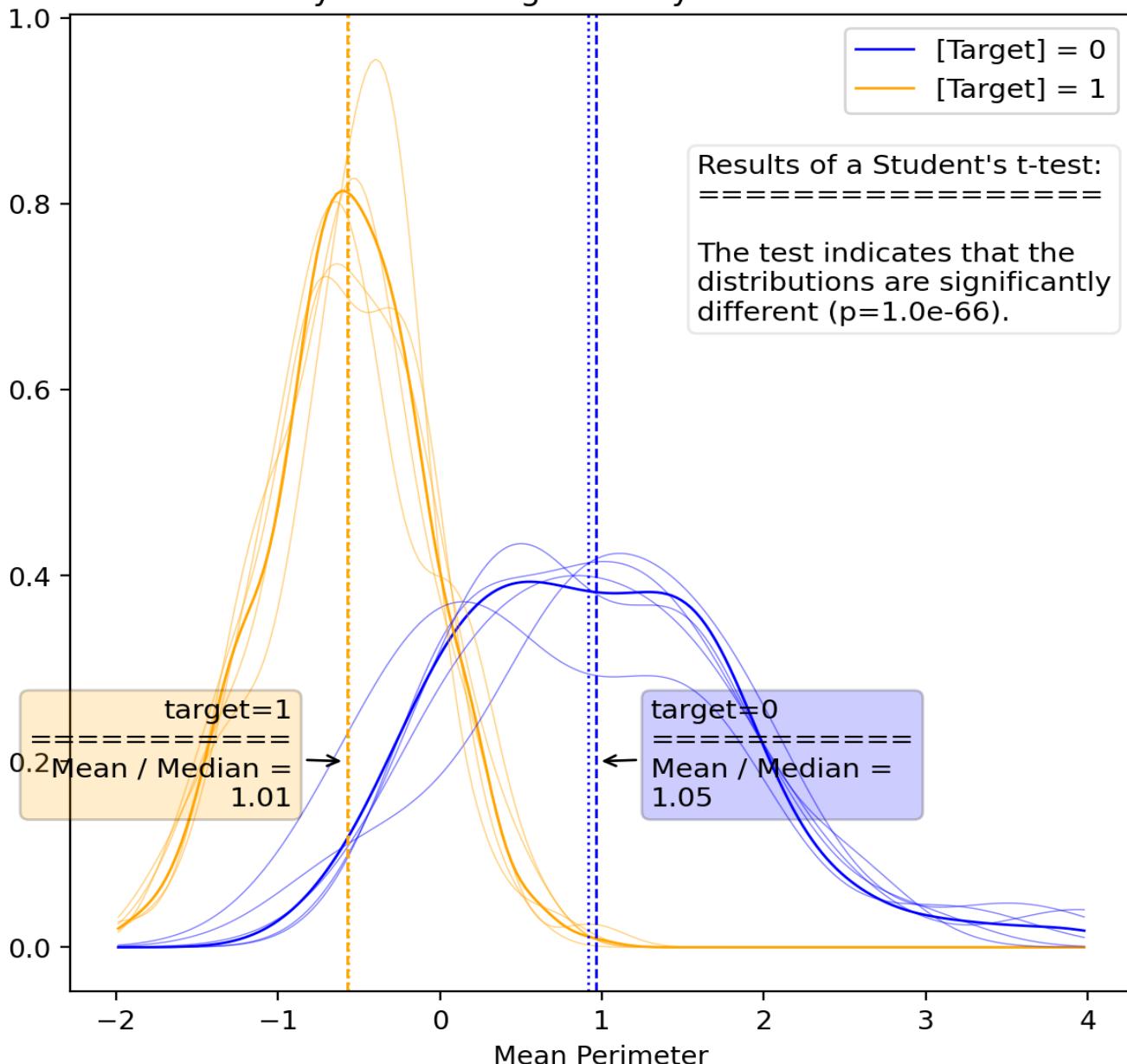
Mean Perimeter - Results

| | Fold-1 | Fold-2 | Fold-3 | Fold-4 | Fold-5 | Agg. Mean | Agg. SD |
|-------------------------|----------|----------|----------|----------|----------|-----------|---------|
| Fitted Coef. | -3.7e+00 | -3.9e+00 | -3.8e+00 | -3.6e+00 | -3.8e+00 | 1.3e-03 | 1.2e-01 |
| Fitted p-Value | 7.9e-26 | 1.0e-26 | 1.7e-25 | 3.4e-27 | 6.4e-25 | 1.4e-01 | 2.7e-25 |
| Fitted Std. Err. | 0.349 | 0.365 | 0.367 | 0.335 | 0.365 | 0.001 | 0.014 |
| Conf. Int. Lower | -4.4e+00 | -4.6e+00 | -4.5e+00 | -4.3e+00 | -4.5e+00 | -4.3e-04 | 1.4e-01 |
| Conf. Int. Upper | -3.0e+00 | -3.2e+00 | -3.1e+00 | -3.0e+00 | -3.0e+00 | 3.0e-03 | 9.5e-02 |
| Train Accuracy | 86.5% | 87.6% | 86.6% | 86.4% | 87.9% | 13.0% | 0.7% |
| Val Accuracy | 88.7% | 85.1% | 88.5% | 91.0% | 84.0% | 13.0% | 2.9% |
| Train AUC | 86.1% | 87.7% | 86.3% | 86.4% | 87.3% | 13.2% | 0.7% |
| Val AUC | 89.2% | 83.1% | 88.4% | 90.7% | 84.7% | 13.2% | 3.2% |
| Train F1 | 89.0% | 89.9% | 88.9% | 89.0% | 90.4% | 15.1% | 0.7% |
| Test F1 | 91.0% | 88.4% | 91.5% | 92.2% | 85.7% | 15.1% | 2.7% |
| Train Precision | 90.3% | 92.5% | 90.0% | 91.8% | 91.2% | 19.5% | 1.0% |
| Val Precision | 94.7% | 85.1% | 94.6% | 91.4% | 91.3% | 19.5% | 3.9% |
| Train Recall | 87.7% | 87.5% | 87.8% | 86.3% | 89.6% | 12.3% | 1.2% |
| Val Recall | 87.7% | 91.9% | 88.6% | 93.0% | 80.8% | 12.3% | 4.8% |
| Train MCC | 71.7% | 74.1% | 72.1% | 71.3% | 74.0% | -72.6% | 1.3% |
| Val MCC | 76.3% | 68.3% | 74.1% | 81.6% | 68.3% | -72.6% | 5.7% |
| Train Log-Loss | 4.86 | 4.47 | 4.82 | 4.92 | 4.36 | 31.36 | 0.25 |
| Val Log-Loss | 4.07 | 5.35 | 4.15 | 3.24 | 5.78 | 31.36 | 1.03 |

Univariate Report

Mean Perimeter - Kernel Density Plot

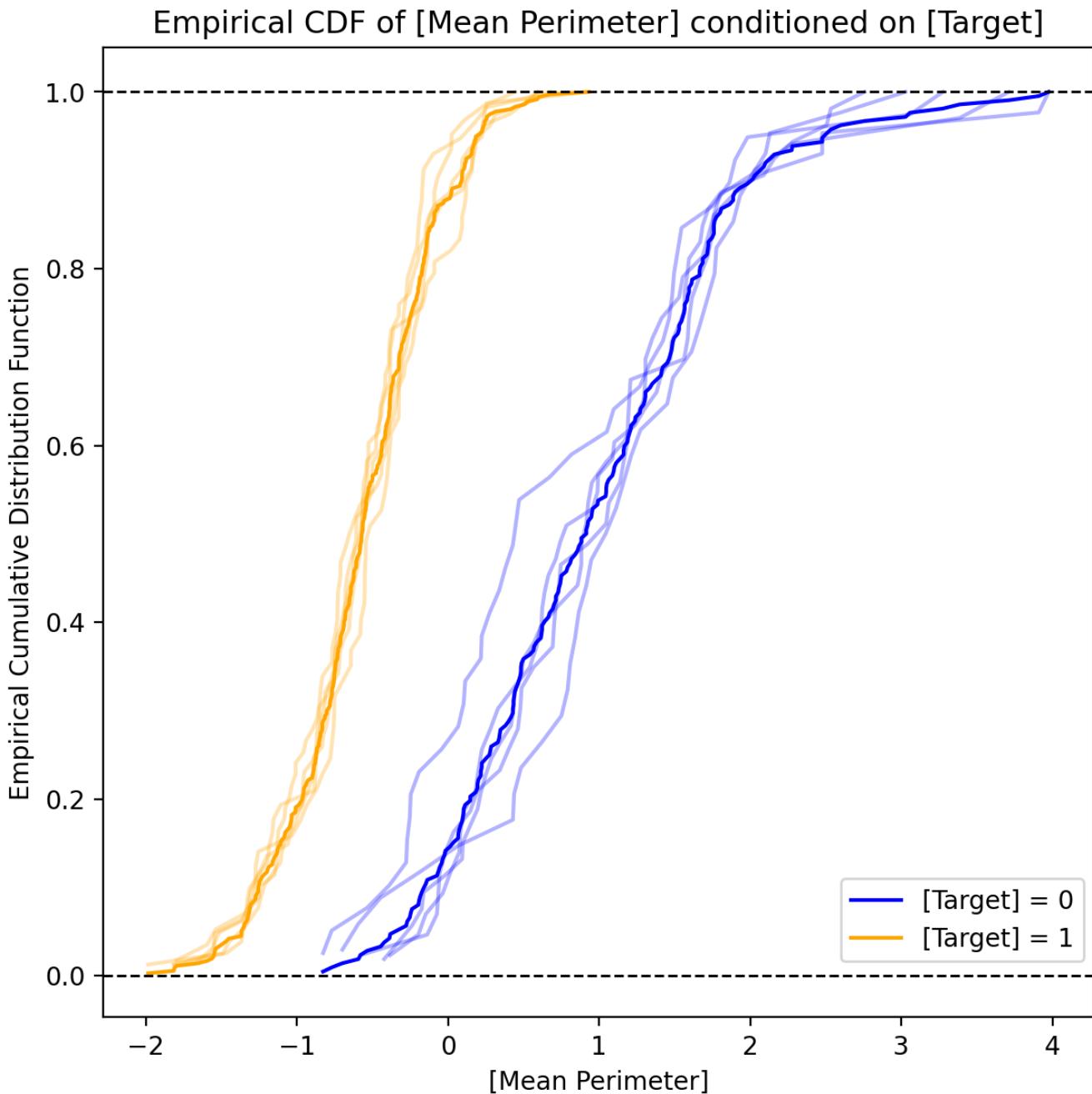
Kernel Density Plot of [Mean Perimeter] by [Target]
Distributions by level are significantly different at the 95% level.



This plot shows the Gaussian kernel density for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the density of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data. There are annotations with the results of a t-test for the difference in means between the feature variable at each level of the target variable. The annotations corresponding to the color of the target variable level show the mean/median ratio to help understand differences in skewness between the levels of the target variable.

Univariate Report

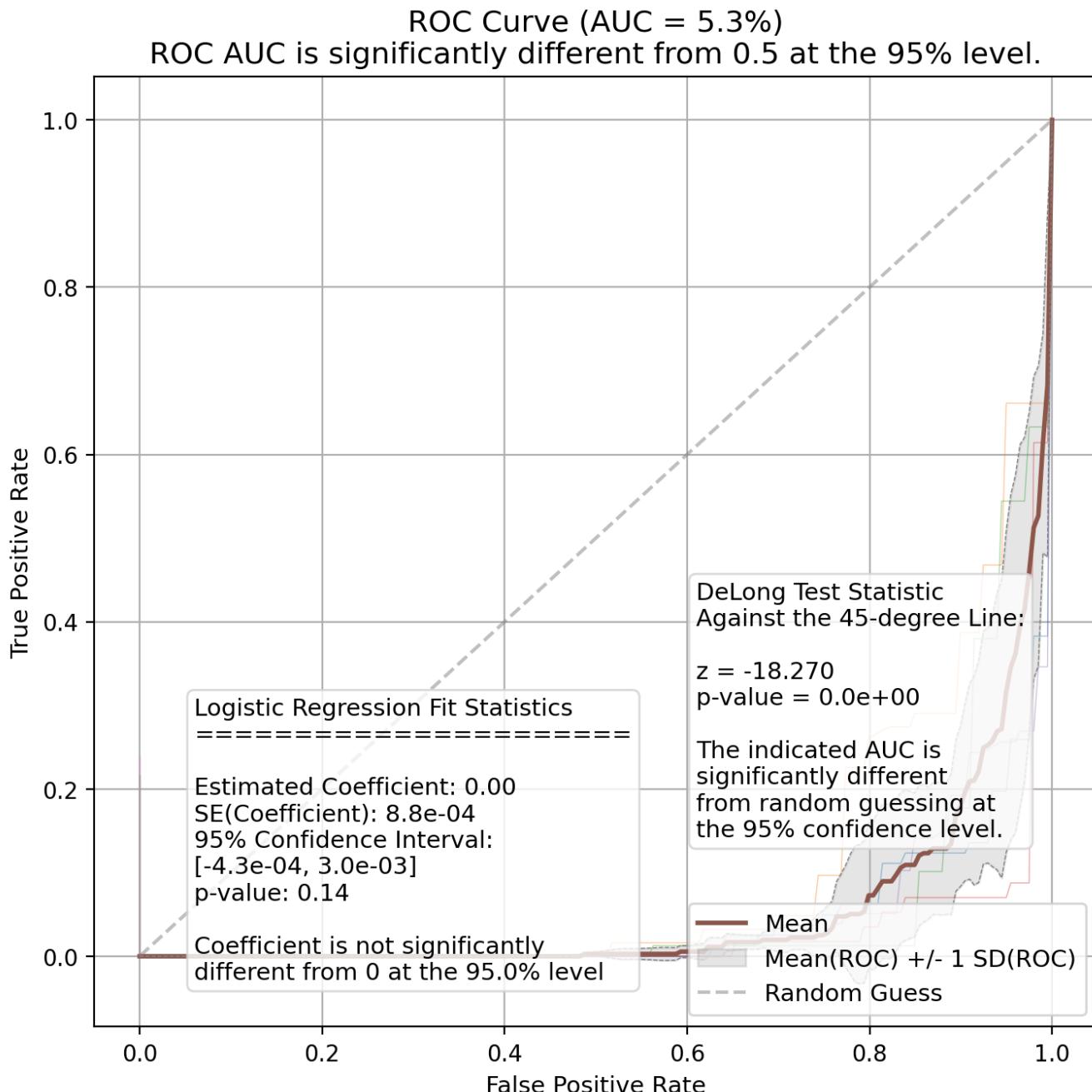
Mean Perimeter - Empirical CDF Plot



This plot shows the empirical cumulative distribution function for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the cumulative distribution of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data, and whether or not it is reasonable to assume that the data is drawn from different distributions.

Univariate Report

Mean Perimeter - ROC Curve

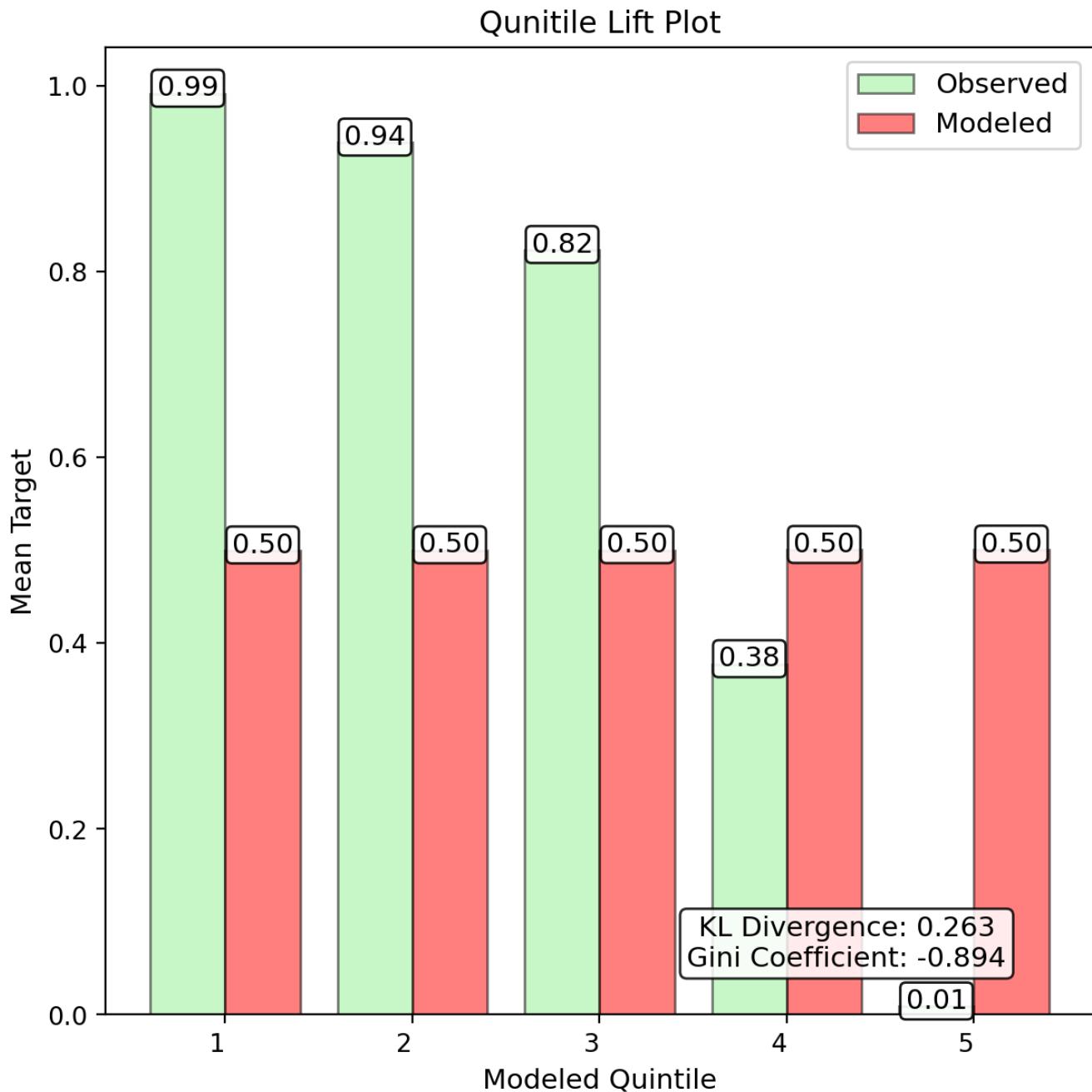


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, and no other features. Annotations are on the plot to help understand the results of the model, including the coefficient, standard error, and p-value for the feature variable. The cross-validation folds are used to create the grey region around the mean ROC curve to help understand the variability of the data.

Significance of the ROC curve is determined based on a modified version the method from DeLong et al. (1988). In brief, the AUC is assumed to be normally distributed, and I calculate the empirical standard error from the cross-validated AUC values. I then calculate a z-score for the AUC, and use the z-score to calculate a p-value. The p-value is then used to determine the significance of the AUC. This is a simple test, and should be used with caution.

Univariate Report

Mean Perimeter - Quintile Lift



The quintile lift plot is meant to show the power of the single feature to discriminate between the highest and lowest quintiles of the target variable.

Univariate Report

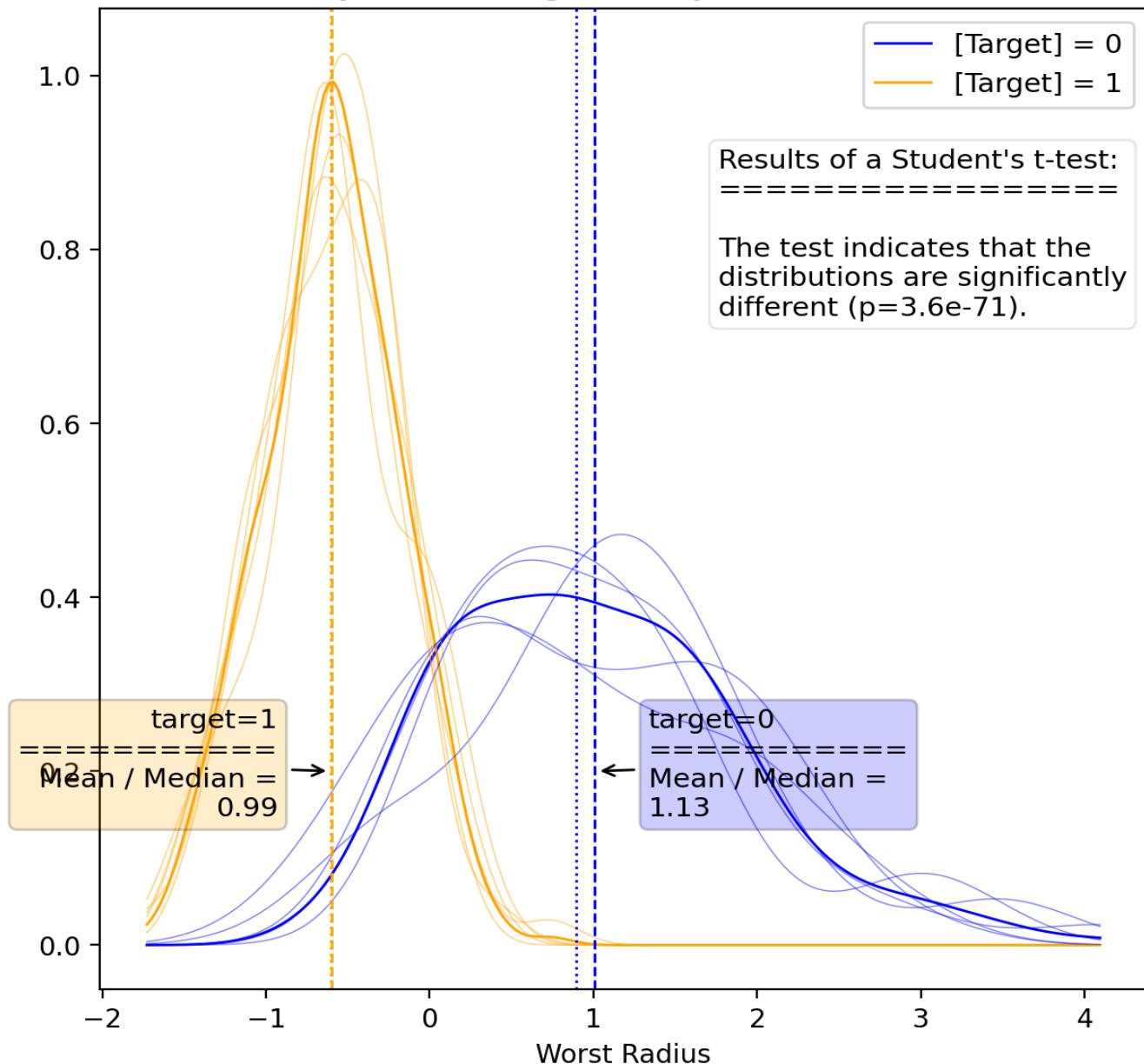
Worst Radius - Results

| | Fold-1 | Fold-2 | Fold-3 | Fold-4 | Fold-5 | Agg. Mean | Agg. SD |
|-------------------------|----------|----------|----------|----------|----------|-----------|---------|
| Fitted Coef. | -4.7e+00 | -5.1e+00 | -5.1e+00 | -4.6e+00 | -4.9e+00 | 3.5e-03 | 2.3e-01 |
| Fitted p-Value | 6.8e-24 | 6.9e-24 | 3.9e-23 | 2.7e-25 | 3.0e-23 | 4.7e-01 | 1.7e-23 |
| Fitted Std. Err. | 0.465 | 0.505 | 0.516 | 0.444 | 0.498 | 0.005 | 0.030 |
| Conf. Int. Lower | -5.6e+00 | -6.1e+00 | -6.1e+00 | -5.5e+00 | -5.9e+00 | -6.1e-03 | 2.9e-01 |
| Conf. Int. Upper | -3.8e+00 | -4.1e+00 | -4.1e+00 | -3.7e+00 | -4.0e+00 | 1.3e-02 | 1.7e-01 |
| Train Accuracy | 90.8% | 91.2% | 91.2% | 88.7% | 91.1% | 9.3% | 1.1% |
| Val Accuracy | 90.3% | 89.1% | 89.4% | 97.0% | 87.0% | 9.3% | 3.8% |
| Train AUC | 90.2% | 91.0% | 90.5% | 88.2% | 90.7% | 9.8% | 1.1% |
| Val AUC | 89.3% | 86.8% | 88.2% | 97.4% | 87.3% | 9.8% | 4.3% |
| Train F1 | 92.6% | 93.0% | 92.9% | 91.1% | 92.9% | 9.8% | 0.8% |
| Test F1 | 92.6% | 91.6% | 92.3% | 97.3% | 88.7% | 9.8% | 3.1% |
| Train Precision | 92.4% | 94.1% | 91.9% | 92.2% | 93.8% | 13.0% | 1.0% |
| Val Precision | 92.6% | 87.0% | 93.5% | 100.0% | 91.8% | 13.0% | 4.7% |
| Train Recall | 92.8% | 91.9% | 93.9% | 90.0% | 92.1% | 7.8% | 1.4% |
| Val Recall | 92.6% | 96.8% | 91.1% | 94.7% | 85.9% | 7.8% | 4.1% |
| Train MCC | 80.4% | 81.4% | 81.5% | 75.7% | 80.9% | -80.1% | 2.4% |
| Val MCC | 78.6% | 77.1% | 75.2% | 94.1% | 73.7% | -80.1% | 8.2% |
| Train Log-Loss | 3.32 | 3.16 | 3.16 | 4.07 | 3.21 | 32.69 | 0.39 |
| Val Log-Loss | 3.49 | 3.93 | 3.83 | 1.08 | 4.68 | 32.69 | 1.37 |

Univariate Report

Worst Radius - Kernel Density Plot

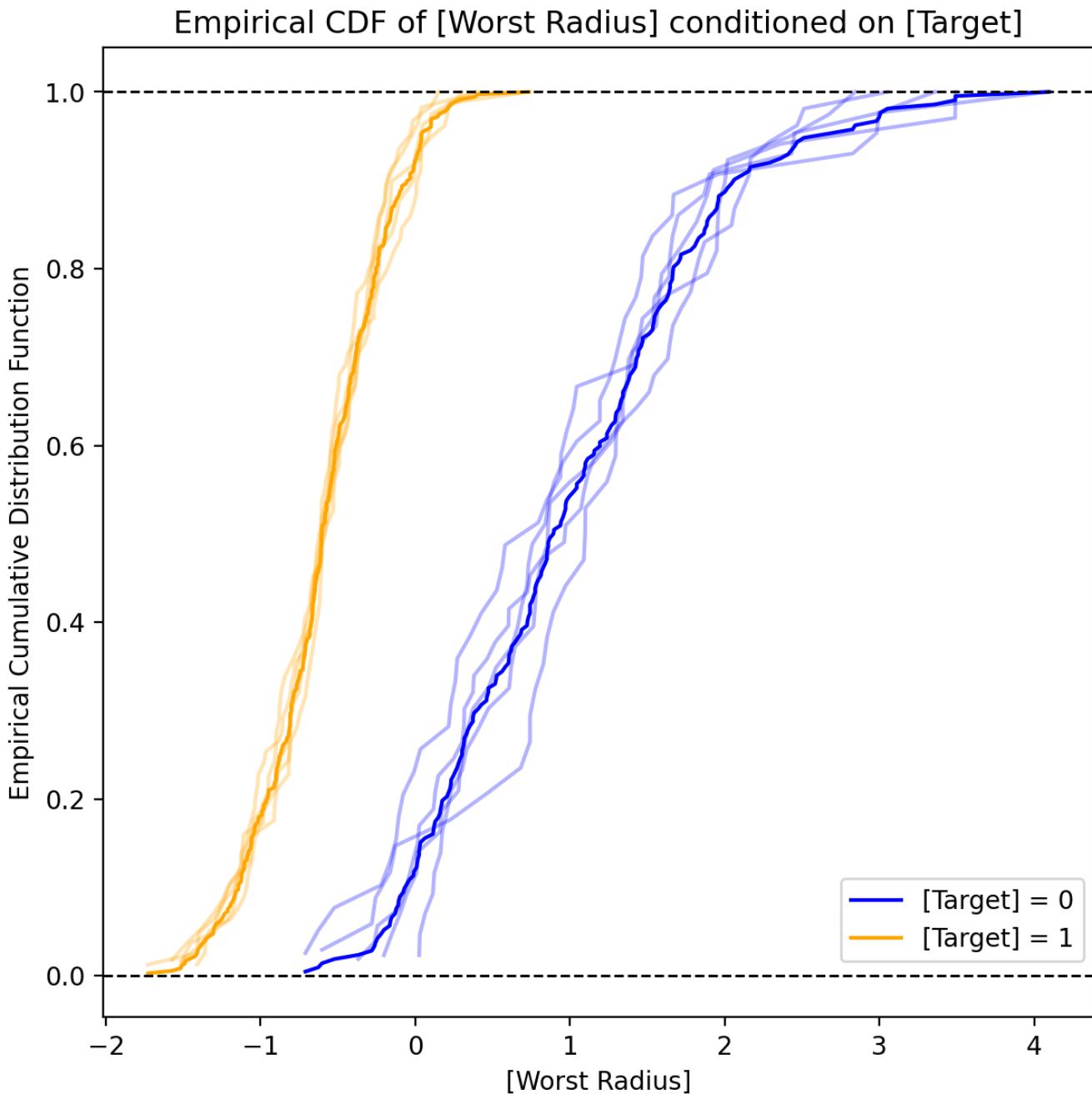
Kernel Density Plot of [Worst Radius] by [Target]
Distributions by level are significantly different at the 95% level.



This plot shows the Gaussian kernel density for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the density of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data. There are annotations with the results of a t-test for the difference in means between the feature variable at each level of the target variable. The annotations corresponding to the color of the target variable level show the mean/median ratio to help understand differences in skewness between the levels of the target variable.

Univariate Report

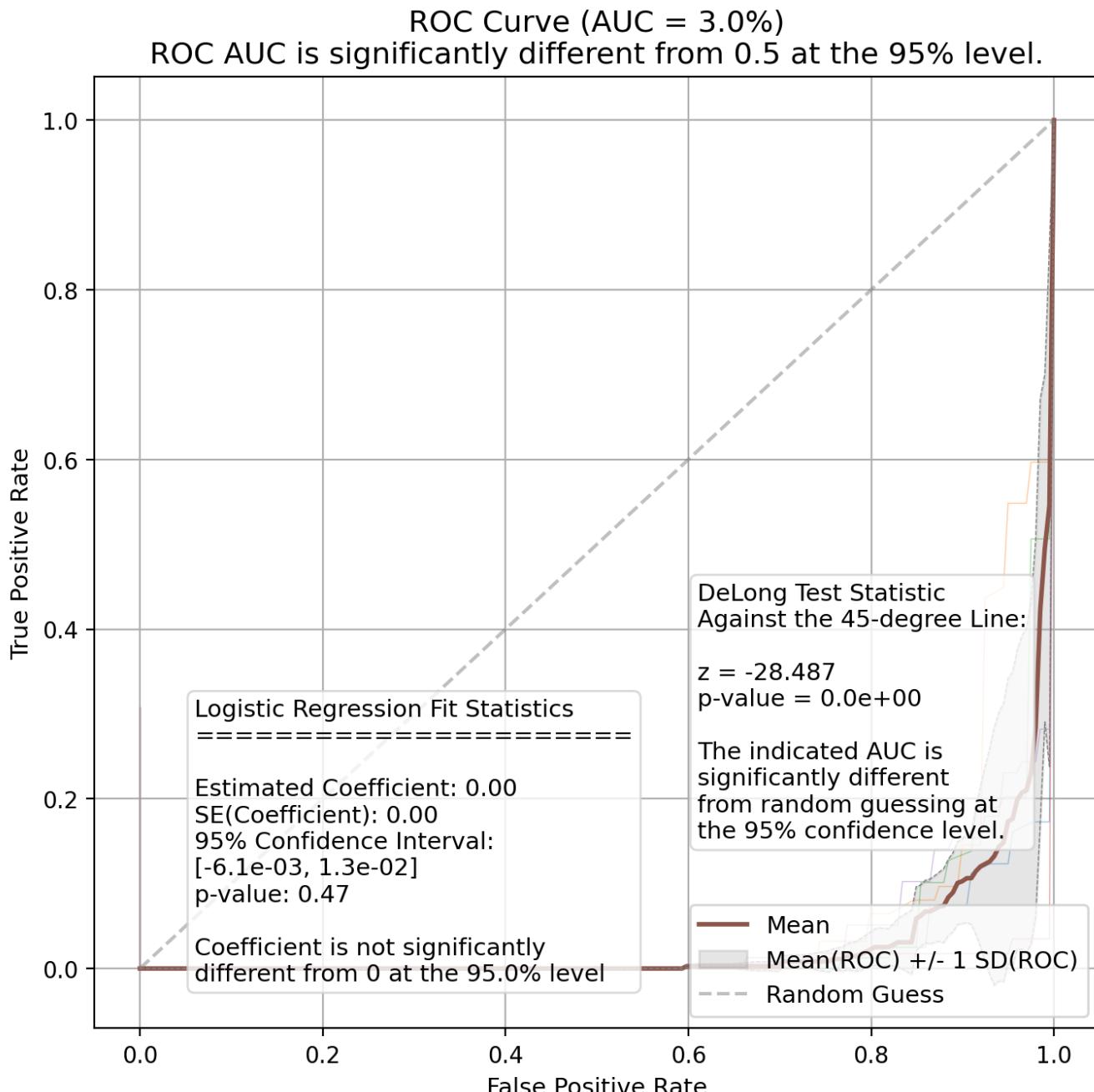
Worst Radius - Empirical CDF Plot



This plot shows the empirical cumulative distribution function for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the cumulative distribution of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data, and whether or not it is reasonable to assume that the data is drawn from different distributions.

Univariate Report

Worst Radius - ROC Curve

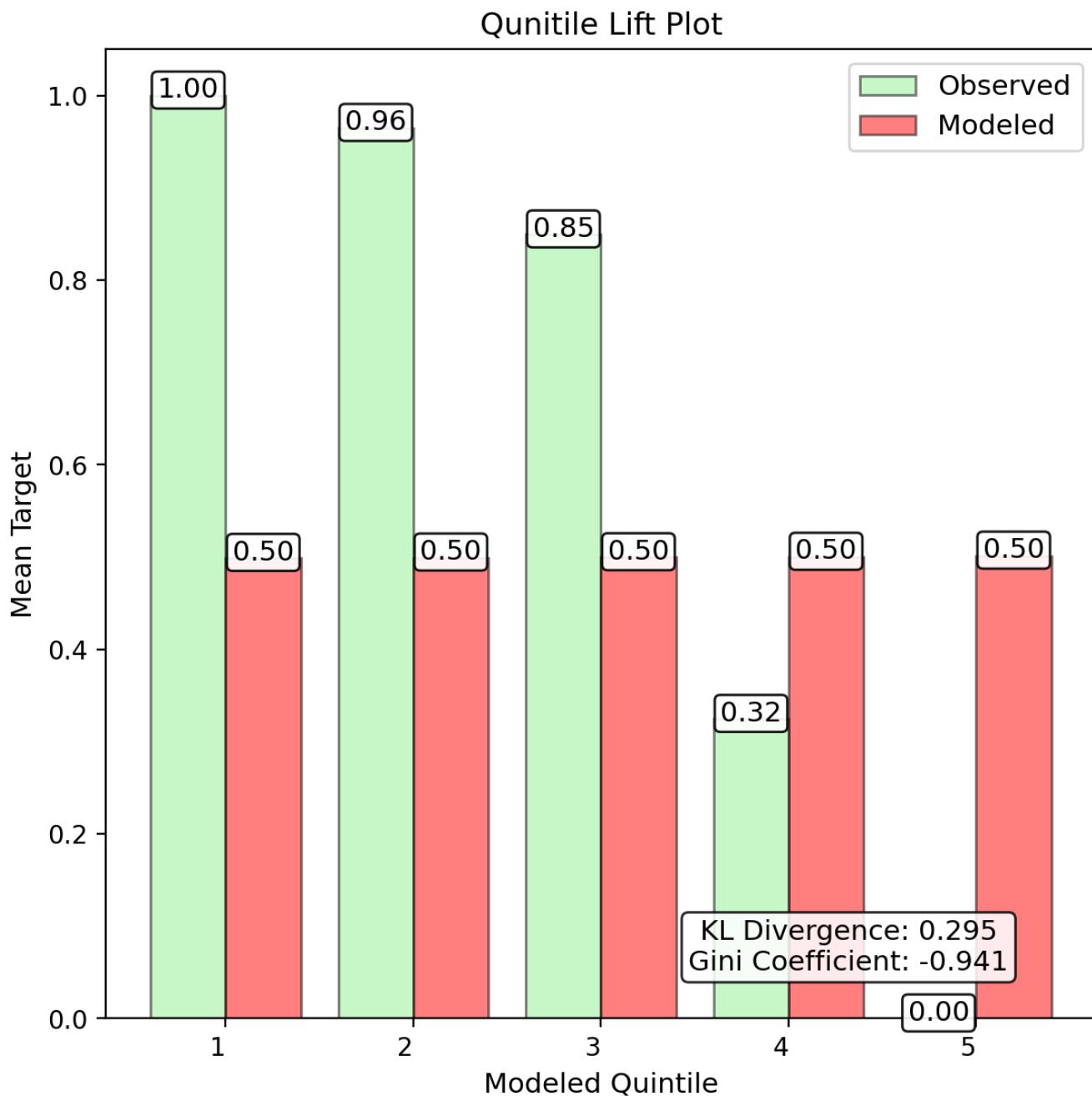


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, and no other features. Annotations are on the plot to help understand the results of the model, including the coefficient, standard error, and p-value for the feature variable. The cross-validation folds are used to create the grey region around the mean ROC curve to help understand the variability of the data.

Significance of the ROC curve is determined based on a modified version the method from DeLong et al. (1988). In brief, the AUC is assumed to be normally distributed, and I calculate the empirical standard error from the cross-validated AUC values. I then calculate a z-score for the AUC, and use the z-score to calculate a p-value. The p-value is then used to determine the significance of the AUC. This is a simple test, and should be used with caution.

Univariate Report

Worst Radius - Quintile Lift



The quintile lift plot is meant to show the power of the single feature to discriminate between the highest and lowest quintiles of the target variable.

Univariate Report

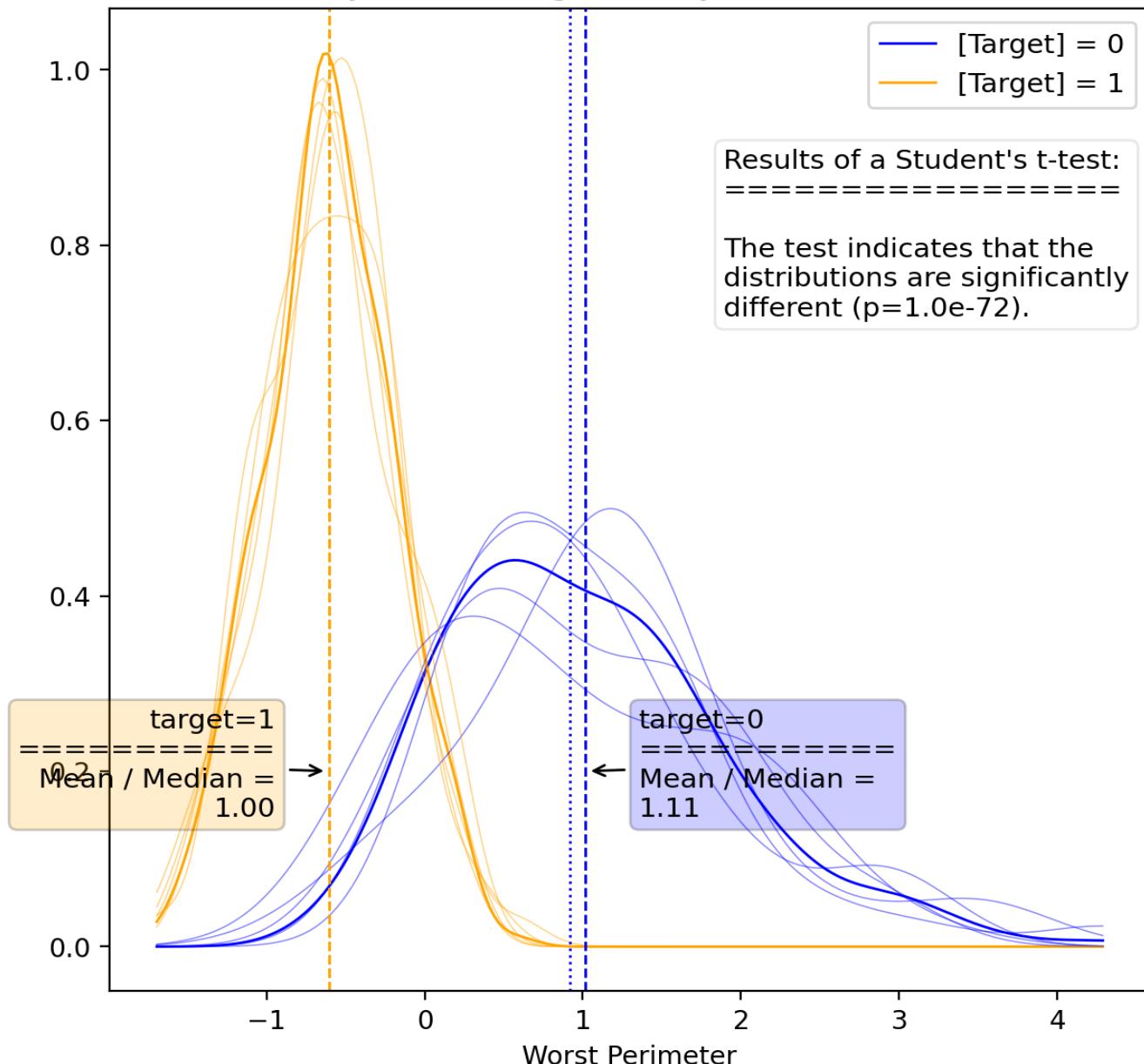
Worst Perimeter - Results

| | Fold-1 | Fold-2 | Fold-3 | Fold-4 | Fold-5 | Agg. Mean | Agg. SD |
|-------------------------|----------|----------|----------|----------|----------|-----------|---------|
| Fitted Coef. | -4.9e+00 | -5.2e+00 | -5.3e+00 | -4.8e+00 | -5.2e+00 | 3.0e-04 | 2.2e-01 |
| Fitted p-Value | 4.3e-24 | 3.0e-24 | 2.0e-23 | 3.8e-25 | 4.8e-23 | 6.9e-01 | 2.0e-23 |
| Fitted Std. Err. | 0.479 | 0.515 | 0.529 | 0.466 | 0.524 | 0.001 | 0.028 |
| Conf. Int. Lower | -5.8e+00 | -6.2e+00 | -6.3e+00 | -5.7e+00 | -6.2e+00 | -1.2e-03 | 2.7e-01 |
| Conf. Int. Upper | -3.9e+00 | -4.2e+00 | -4.2e+00 | -3.9e+00 | -4.2e+00 | 1.8e-03 | 1.6e-01 |
| Train Accuracy | 91.5% | 92.3% | 91.7% | 90.2% | 91.8% | 8.3% | 0.8% |
| Val Accuracy | 91.9% | 89.1% | 91.2% | 97.0% | 90.1% | 8.3% | 3.1% |
| Train AUC | 91.1% | 92.2% | 91.2% | 89.6% | 91.4% | 8.6% | 0.9% |
| Val AUC | 91.1% | 87.3% | 90.3% | 97.4% | 90.2% | 8.6% | 3.7% |
| Train F1 | 93.1% | 93.8% | 93.2% | 92.3% | 93.5% | 9.1% | 0.6% |
| Test F1 | 93.8% | 91.5% | 93.6% | 97.3% | 91.5% | 9.1% | 2.4% |
| Train Precision | 93.4% | 95.1% | 93.2% | 92.9% | 94.2% | 12.0% | 0.9% |
| Val Precision | 93.8% | 88.1% | 94.8% | 100.0% | 93.3% | 12.0% | 4.3% |
| Train Recall | 92.8% | 92.5% | 93.2% | 91.7% | 92.8% | 7.3% | 0.6% |
| Val Recall | 93.8% | 95.2% | 92.4% | 94.7% | 89.7% | 7.3% | 2.2% |
| Train MCC | 81.9% | 83.7% | 82.5% | 78.8% | 82.3% | -82.4% | 1.8% |
| Val MCC | 82.2% | 76.9% | 79.4% | 94.1% | 79.7% | -82.4% | 6.8% |
| Train Log-Loss | 3.08 | 2.77 | 3.00 | 3.54 | 2.96 | 33.07 | 0.28 |
| Val Log-Loss | 2.91 | 3.93 | 3.19 | 1.08 | 3.58 | 33.07 | 1.11 |

Univariate Report

Worst Perimeter - Kernel Density Plot

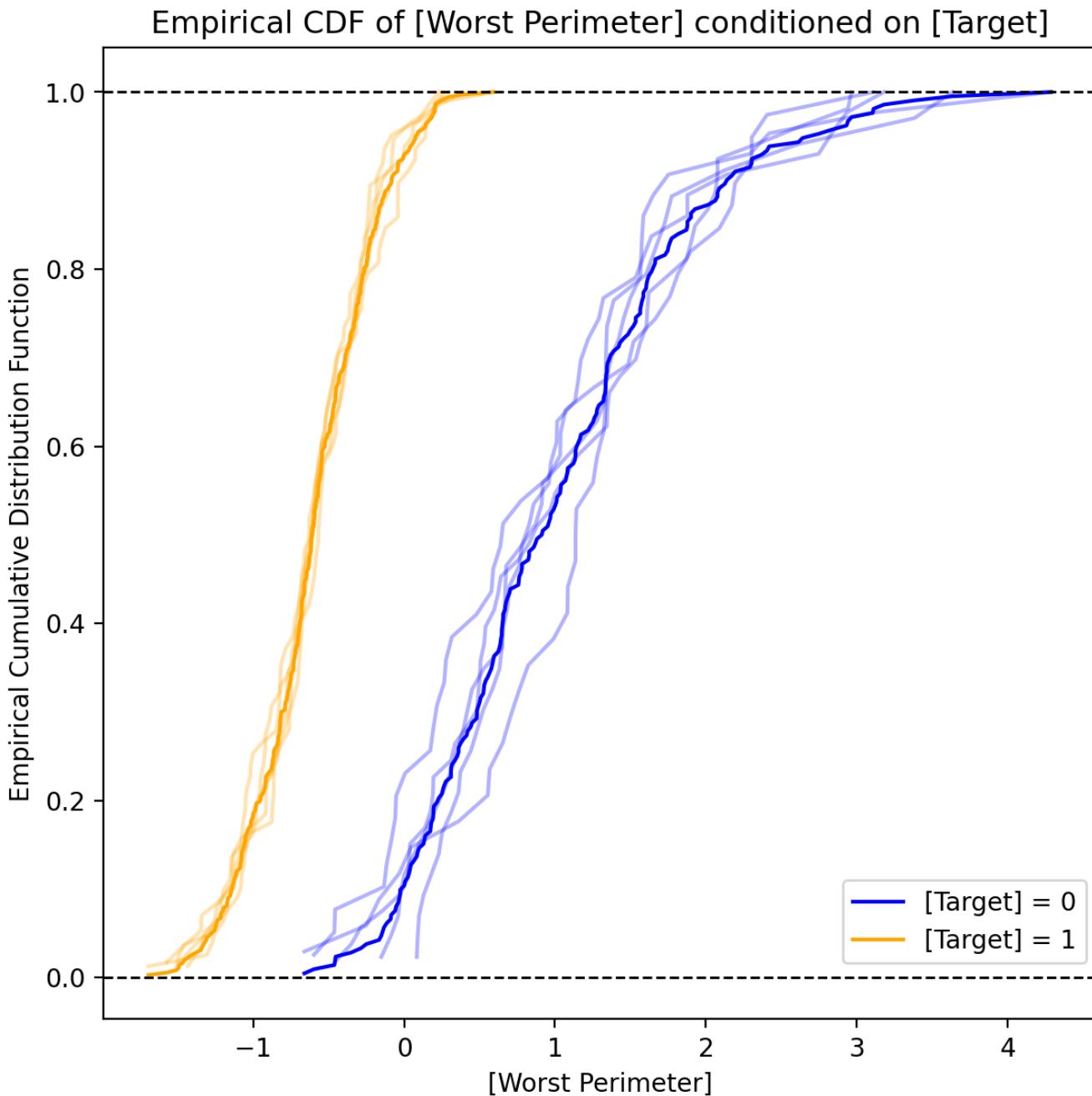
Kernel Density Plot of [Worst Perimeter] by [Target]
Distributions by level are significantly different at the 95% level.



This plot shows the Gaussian kernel density for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the density of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data. There are annotations with the results of a t-test for the difference in means between the feature variable at each level of the target variable. The annotations corresponding to the color of the target variable level show the mean/median ratio to help understand differences in skewness between the levels of the target variable.

Univariate Report

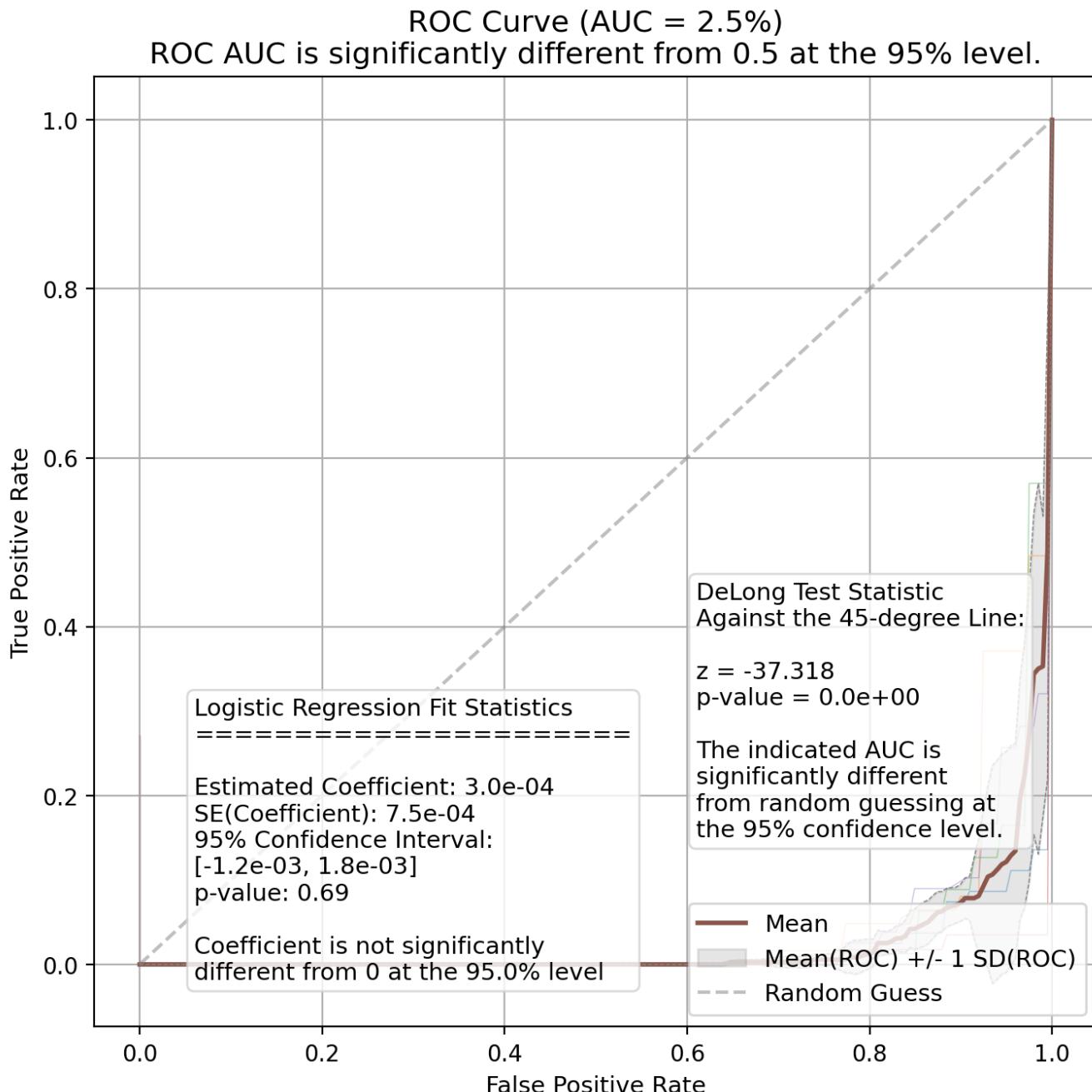
Worst Perimeter - Empirical CDF Plot



This plot shows the empirical cumulative distribution function for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the cumulative distribution of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data, and whether or not it is reasonable to assume that the data is drawn from different distributions.

Univariate Report

Worst Perimeter - ROC Curve

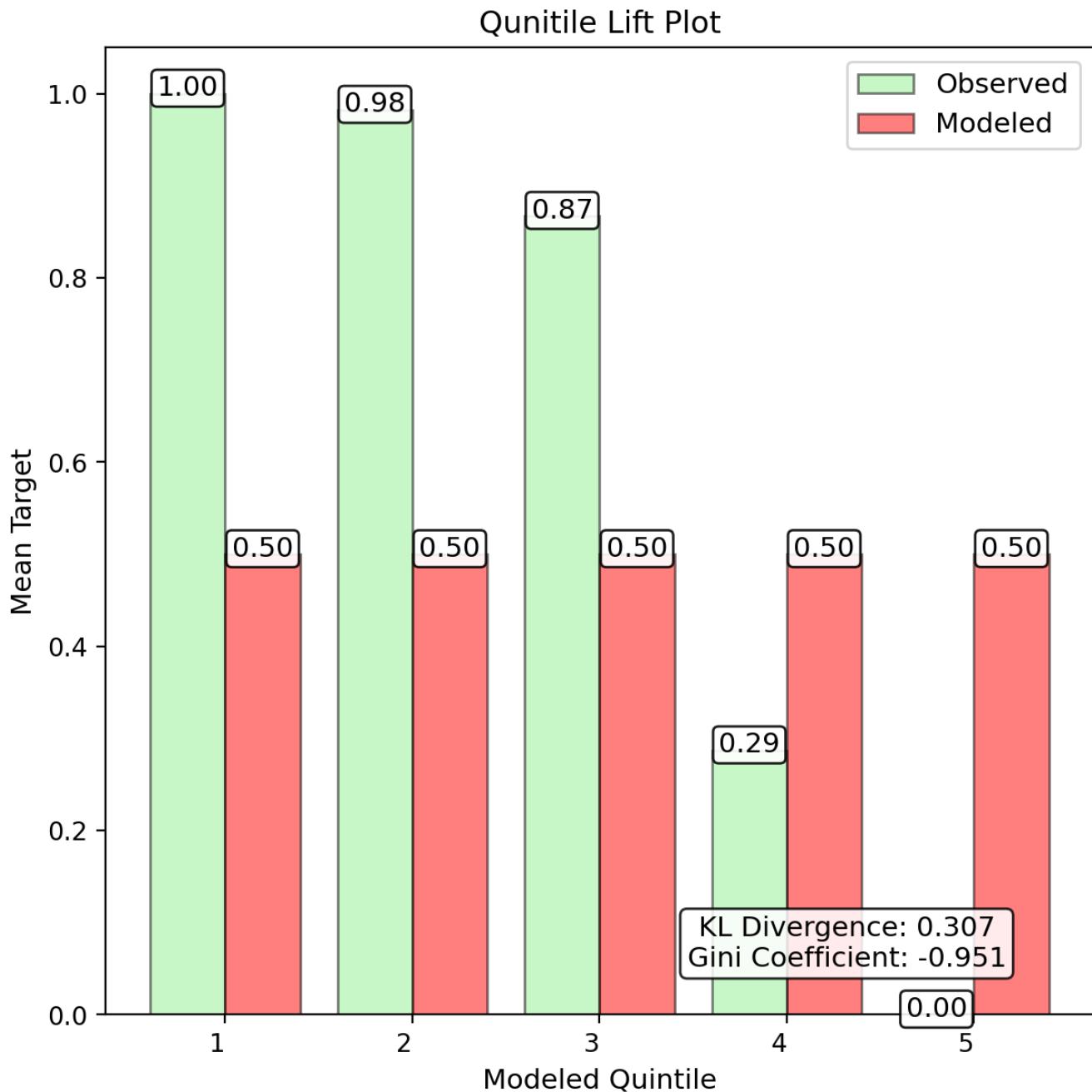


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, and no other features. Annotations are on the plot to help understand the results of the model, including the coefficient, standard error, and p-value for the feature variable. The cross-validation folds are used to create the grey region around the mean ROC curve to help understand the variability of the data.

Significance of the ROC curve is determined based on a modified version the method from DeLong et al. (1988). In brief, the AUC is assumed to be normally distributed, and I calculate the empirical standard error from the cross-validated AUC values. I then calculate a z-score for the AUC, and use the z-score to calculate a p-value. The p-value is then used to determine the significance of the AUC. This is a simple test, and should be used with caution.

Univariate Report

Worst Perimeter - Quintile Lift



The quintile lift plot is meant to show the power of the single feature to discriminate between the highest and lowest quintiles of the target variable.