

Cancer Model Univariate Analysis Report

2024-02-16

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, so 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.
Worst Area	93.0%	90.9%	98.6%	91.2%	94.6%	85.2%	92.2%
Pc1	92.1%	94.3%	93.0%	91.8%	93.6%	83.3%	91.3%
Worst Radius	90.4%	91.7%	93.0%	89.5%	92.3%	79.4%	89.4%
Worst Perimeter	89.5%	91.5%	91.5%	88.8%	91.5%	77.6%	88.4%
Mean Concave Points	88.6%	93.9%	87.3%	89.0%	90.5%	76.6%	87.7%
Mean Area	88.6%	88.2%	94.4%	86.7%	91.2%	75.5%	87.4%
Mean Concavity	87.7%	91.3%	88.7%	87.4%	90.0%	74.2%	86.5%
Mean Radius	86.8%	90.0%	88.7%	86.2%	89.4%	72.1%	85.5%
Mean Perimeter	86.8%	90.0%	88.7%	86.2%	89.4%	72.1%	85.5%
Area Error	86.8%	85.9%	94.4%	84.4%	89.9%	71.7%	85.5%
Worst Concave Points	84.2%	96.5%	77.5%	86.4%	85.9%	70.6%	83.5%
Worst Concavity	84.2%	89.6%	84.5%	84.1%	87.0%	67.2%	82.8%
Perimeter Error	82.5%	83.1%	90.1%	80.0%	86.5%	62.0%	80.7%
Radius Error	80.7%	82.7%	87.3%	78.5%	84.9%	58.3%	78.7%
Mean Compactness	78.9%	86.2%	78.9%	79.0%	82.4%	56.7%	77.0%
Worst Compactness	75.4%	83.1%	76.1%	75.2%	79.4%	49.4%	73.1%
Worst Smoothness	74.6%	85.0%	71.8%	75.5%	77.9%	49.4%	72.4%
Mean Texture	71.9%	81.0%	71.8%	72.0%	76.1%	42.8%	69.3%
Concavity Error	70.2%	75.3%	77.5%	67.8%	76.4%	36.0%	67.2%
Mean Smoothness	67.5%	80.4%	63.4%	68.9%	70.9%	36.6%	64.6%

This table shows an overview of the results for the variables in this file, representing those whose average test score are ranked between 1 and 20 of the variables passed to the Cancer Model.

Univariate Report

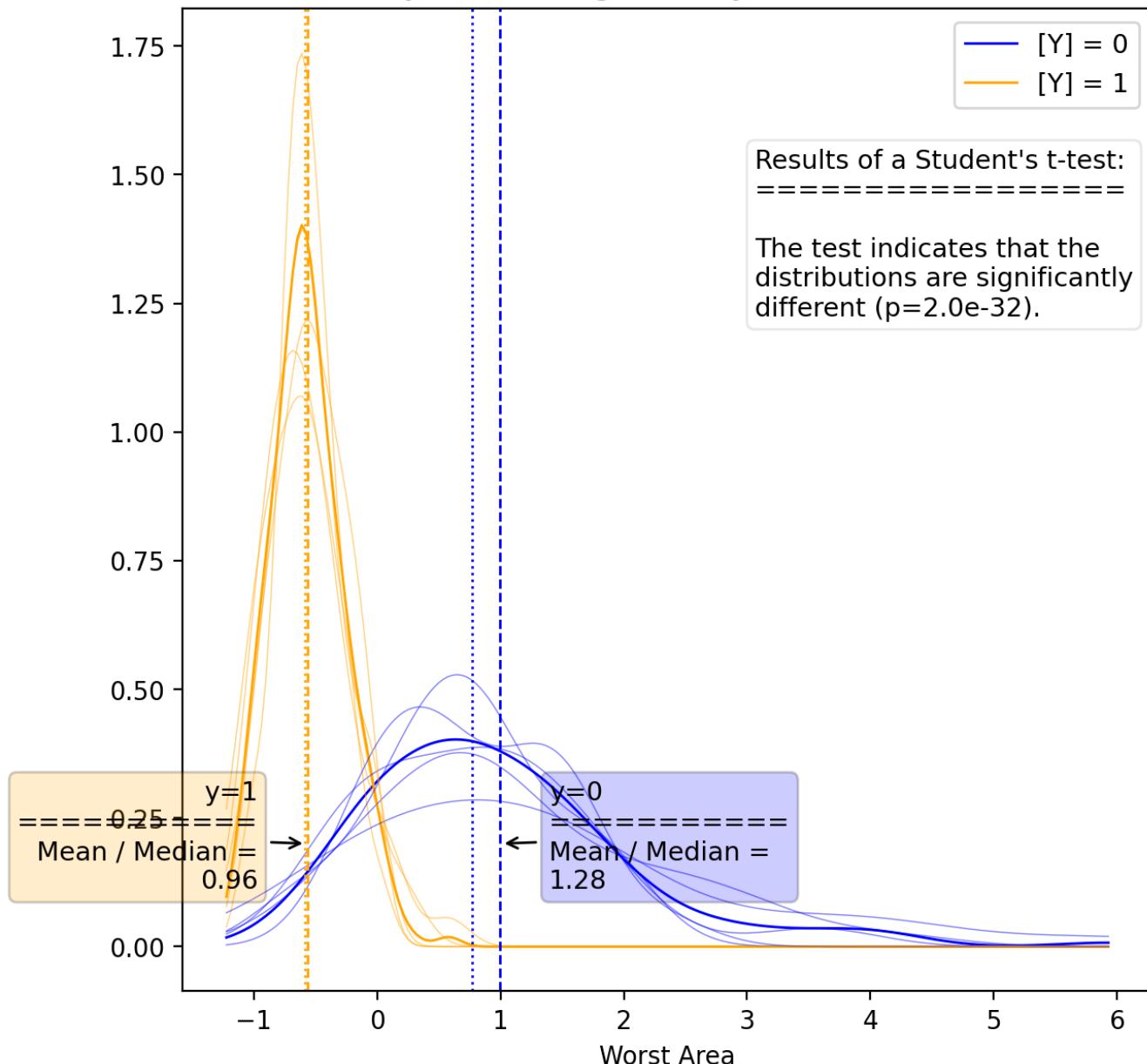
Worst Area - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-5.48	-5.84	-5.46	-6.32	-5.38	-5.67	0.39
Fitted p-Value	1.6e-15	1.4e-15	2.1e-15	2.2e-14	6.0e-17	3.9e-19	9.2e-15
Fitted Std. Err.	0.687	0.731	0.689	0.828	0.644	0.634	0.070
Conf. Int. Lower	-6.82	-7.27	-6.81	-7.95	-6.65	-6.91	0.53
Conf. Int. Upper	-4.13	-4.40	-4.11	-4.70	-4.12	-4.43	0.26
Train Accuracy	92.3%	91.5%	89.6%	92.5%	91.3%	91.5%	1.2%
Val Accuracy	90.0%	89.9%	97.3%	89.2%	92.7%	93.0%	3.3%
Train AUC	91.0%	89.8%	87.9%	90.7%	89.5%	89.7%	1.2%
Val AUC	84.8%	89.6%	95.0%	88.7%	90.9%	91.2%	3.7%
Train F1	93.9%	93.7%	91.7%	94.3%	93.3%	93.5%	1.0%
Test F1	93.1%	90.7%	98.1%	91.1%	94.3%	94.6%	3.0%
Train Precision	91.5%	91.9%	87.6%	91.2%	90.6%	90.4%	1.7%
Val Precision	87.0%	87.2%	96.4%	91.1%	89.2%	90.9%	3.8%
Train Recall	96.4%	95.5%	96.3%	97.6%	96.1%	96.7%	0.8%
Val Recall	100.0%	94.4%	100.0%	91.1%	100.0%	98.6%	4.1%
Train MCC	83.6%	81.1%	78.3%	83.9%	81.1%	81.7%	2.3%
Val MCC	77.8%	79.9%	93.1%	77.3%	85.4%	85.2%	6.6%
Train Log-Loss	2.79	3.05	3.77	2.70	3.15	3.07	0.42
Val Log-Loss	3.60	3.66	0.99	3.90	2.62	2.53	1.20

Univariate Report

Worst Area - Kernel Density Plot

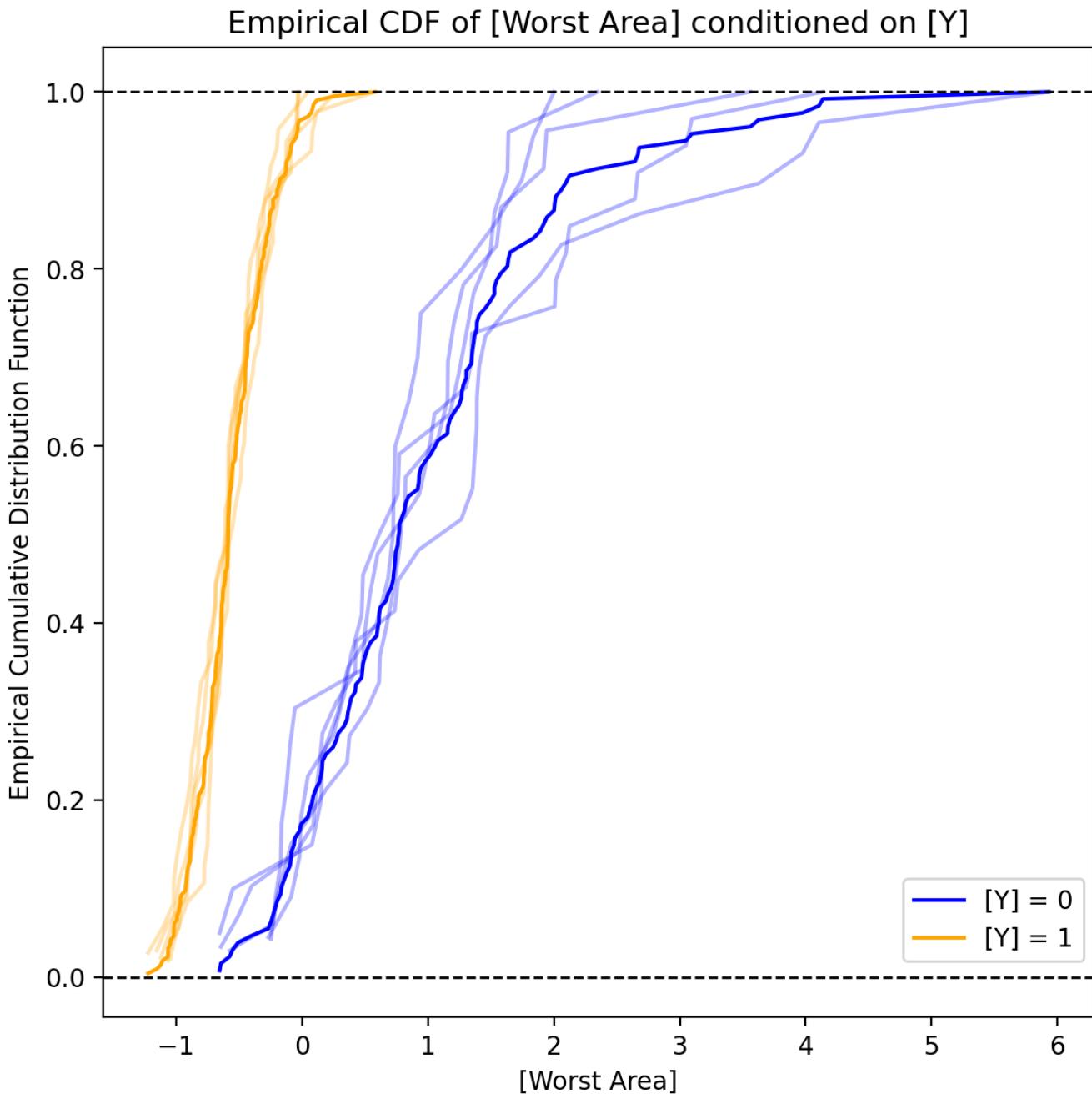
Kernel Density Plot of [Worst Area] by [Y].
Distributions by level are significantly different at the 95% level.



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

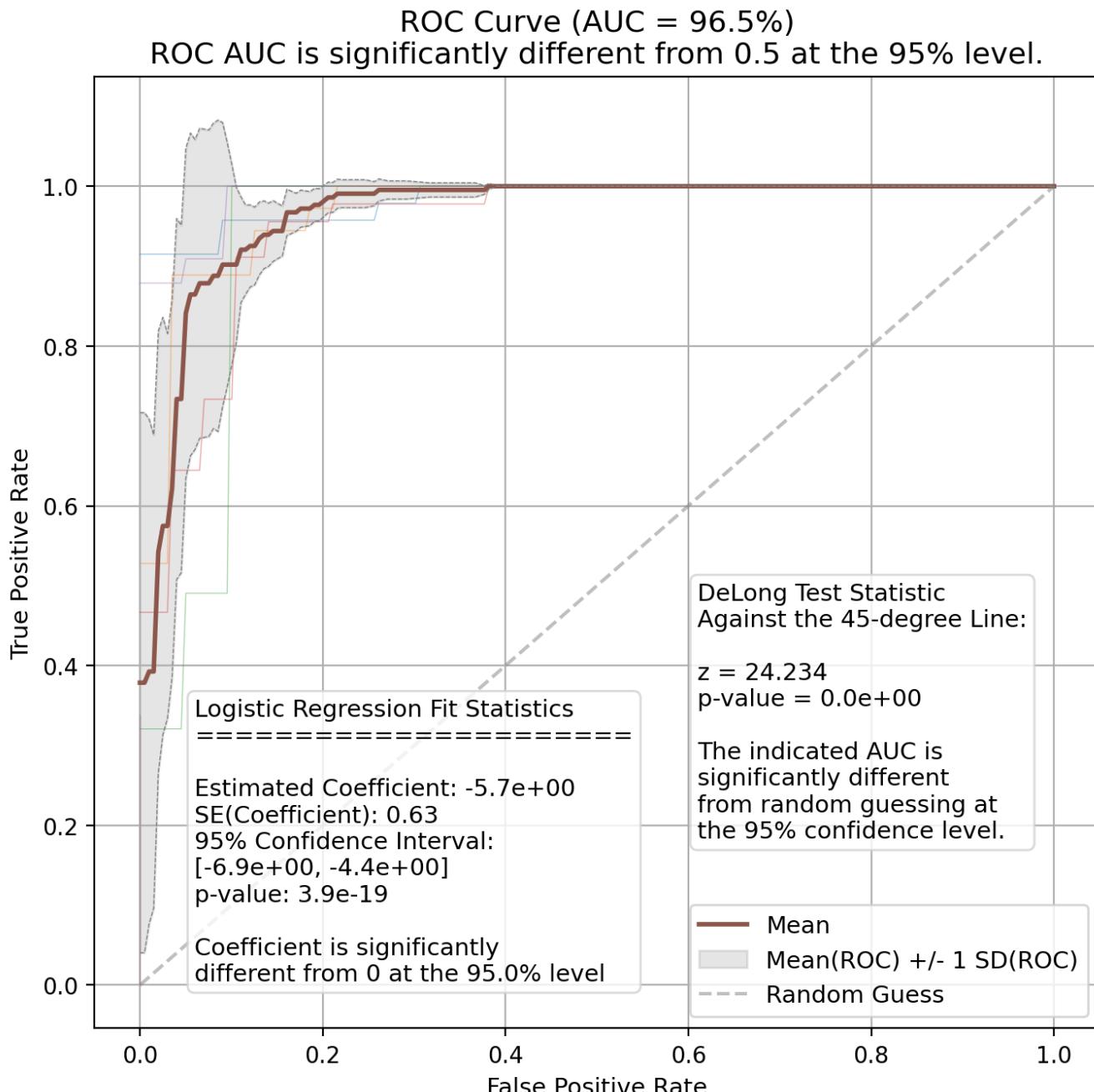
Worst Area - Empirical CDF Plot



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 Area - ROC Curve

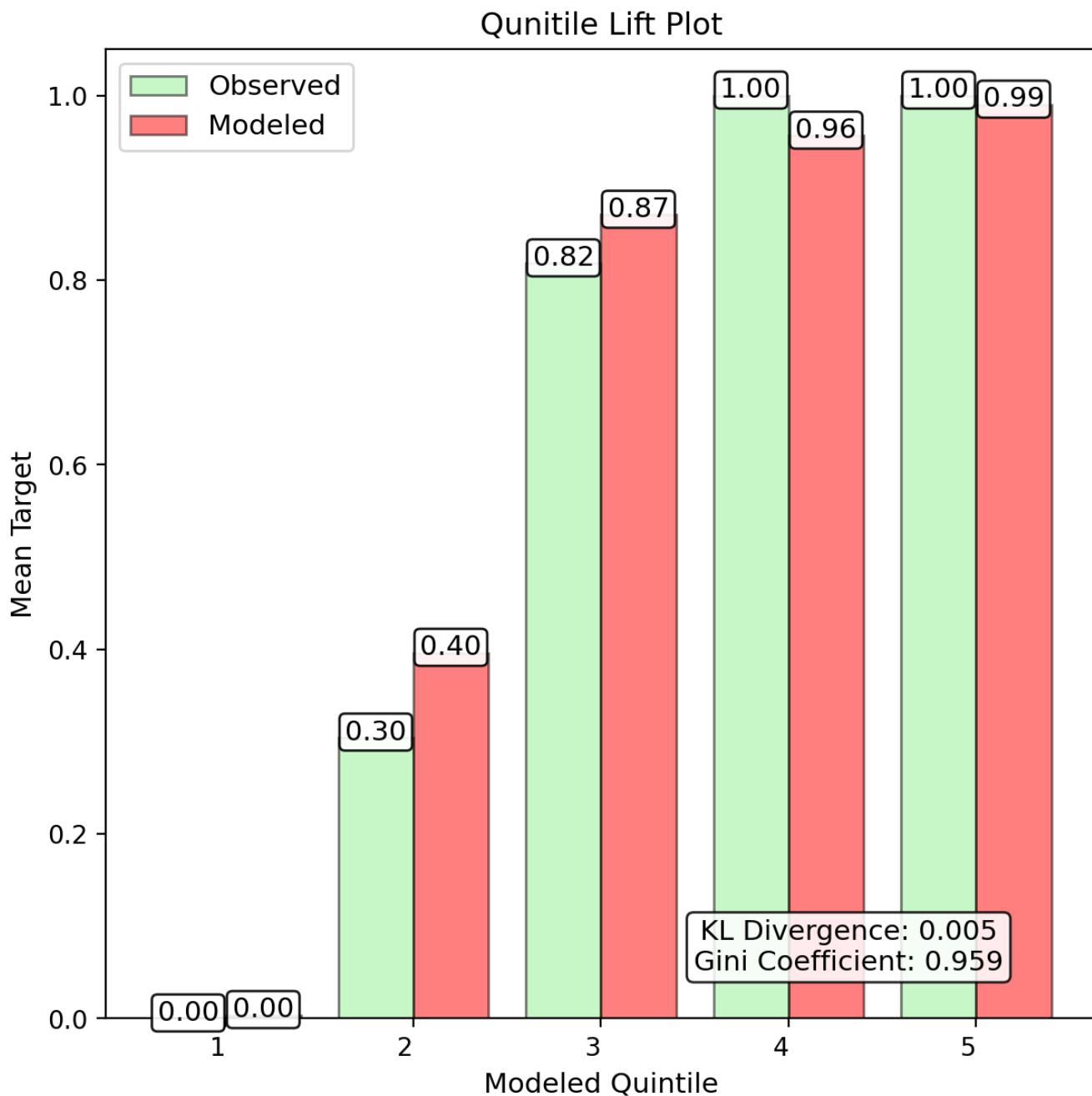


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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 Area - 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

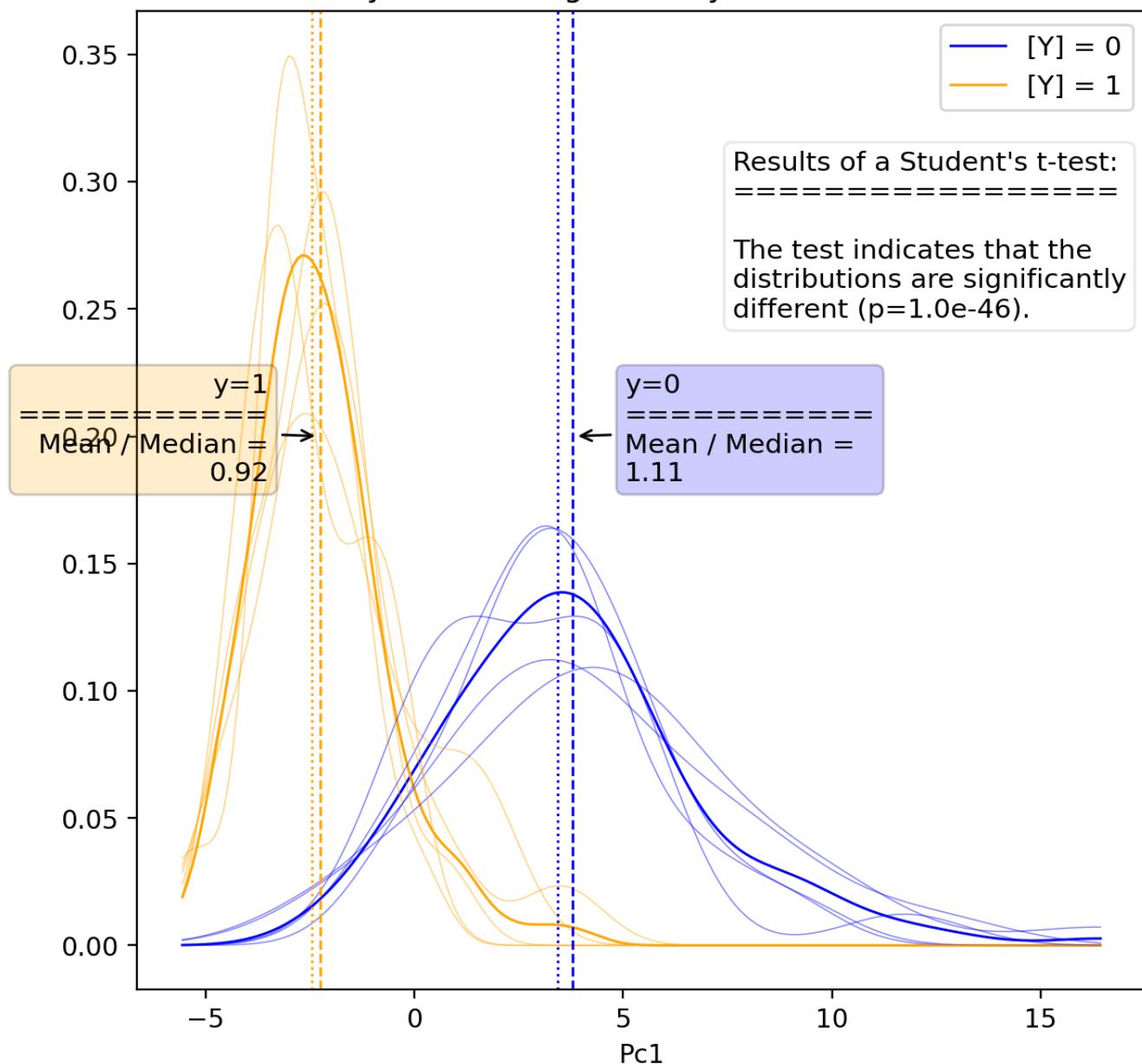
Pc1 - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-1.11	-1.38	-1.37	-1.16	-1.13	-1.21	0.14
Fitted p-Value	2.3e-17	1.9e-15	7.9e-15	1.2e-17	2.1e-18	9.2e-21	3.4e-15
Fitted Std. Err.	0.130	0.174	0.177	0.135	0.130	0.130	0.024
Conf. Int. Lower	-1.36	-1.72	-1.72	-1.42	-1.39	-1.47	0.18
Conf. Int. Upper	-0.850	-1.041	-1.026	-0.893	-0.880	-0.959	0.089
Train Accuracy	90.4%	91.9%	92.5%	90.6%	90.9%	91.5%	0.9%
Val Accuracy	95.7%	84.1%	87.7%	93.2%	94.5%	92.1%	5.0%
Train AUC	90.4%	91.8%	92.2%	90.9%	90.6%	91.3%	0.8%
Val AUC	94.6%	84.3%	88.4%	92.0%	94.7%	91.8%	4.4%
Train F1	92.1%	93.7%	93.8%	92.4%	92.7%	93.1%	0.8%
Test F1	96.8%	83.6%	91.1%	94.6%	95.4%	93.6%	5.3%
Train Precision	93.8%	95.3%	93.8%	95.0%	93.8%	94.3%	0.8%
Val Precision	95.8%	90.3%	95.8%	91.7%	96.9%	94.3%	2.9%
Train Recall	90.4%	92.1%	93.8%	89.9%	91.7%	92.1%	1.5%
Val Recall	97.9%	77.8%	86.8%	97.8%	93.9%	93.0%	8.6%
Train MCC	80.0%	82.5%	84.4%	80.4%	80.6%	82.0%	1.8%
Val MCC	90.2%	69.0%	72.2%	85.9%	88.8%	83.3%	9.9%
Train Log-Loss	3.46	2.92	2.69	3.37	3.28	3.07	0.33
Val Log-Loss	1.54	5.75	4.44	2.44	1.97	2.85	1.79

Univariate Report

Pc1 - Kernel Density Plot

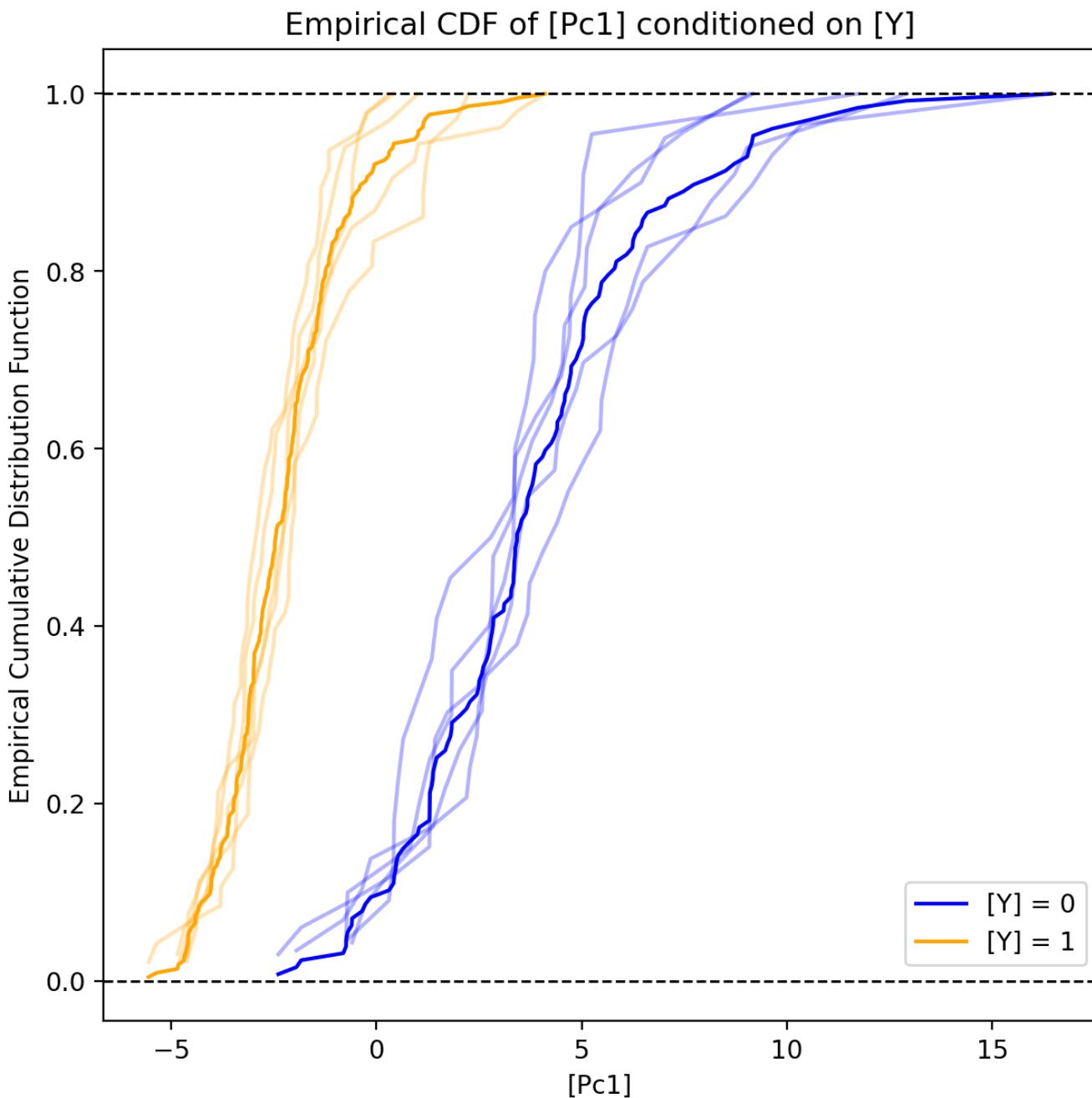
Kernel Density Plot of [Pc1] by [Y].
Distributions by level are significantly different at the 95% level.



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

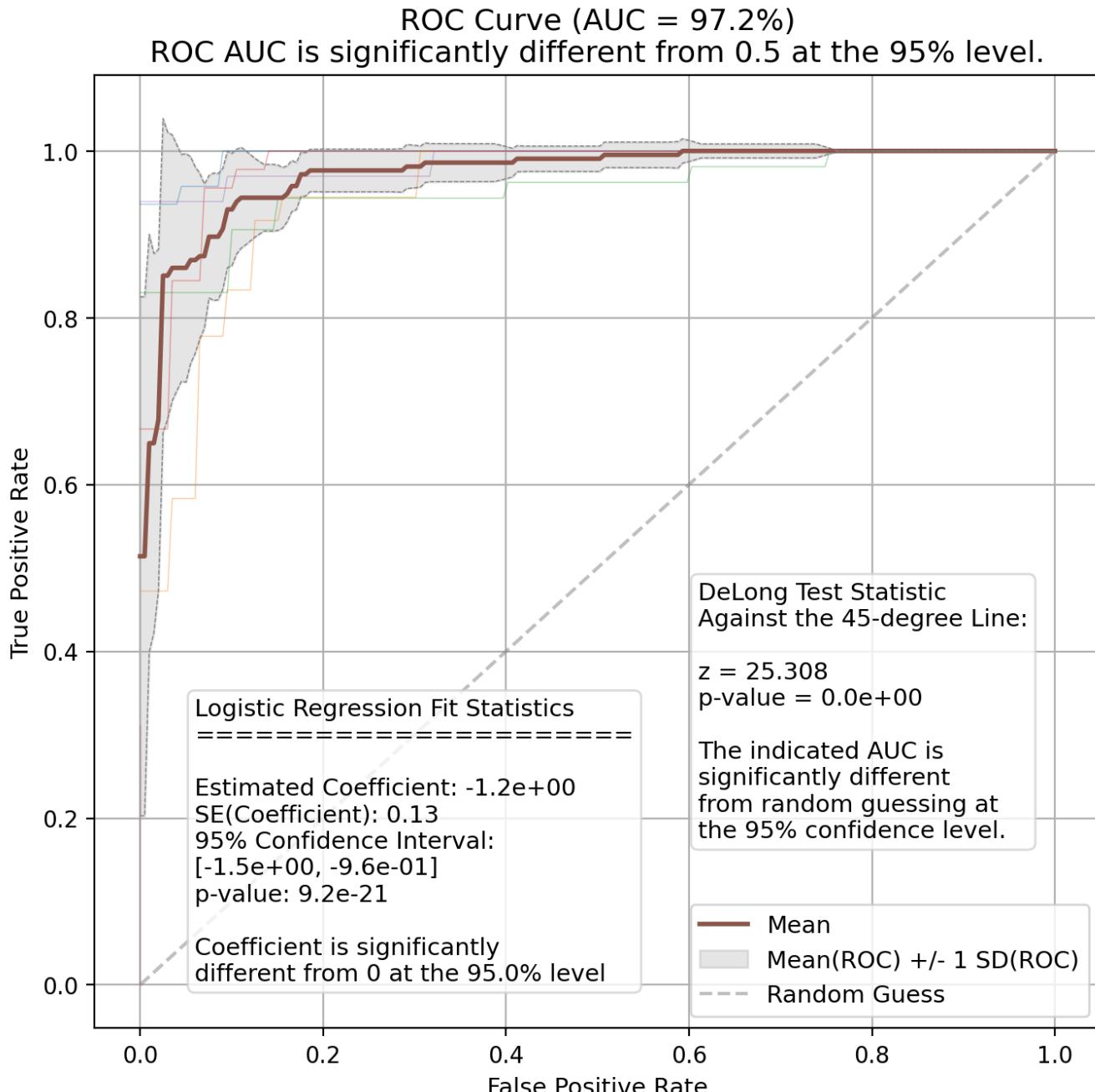
Pc1 - Empirical CDF Plot



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

Pc1 - ROC Curve

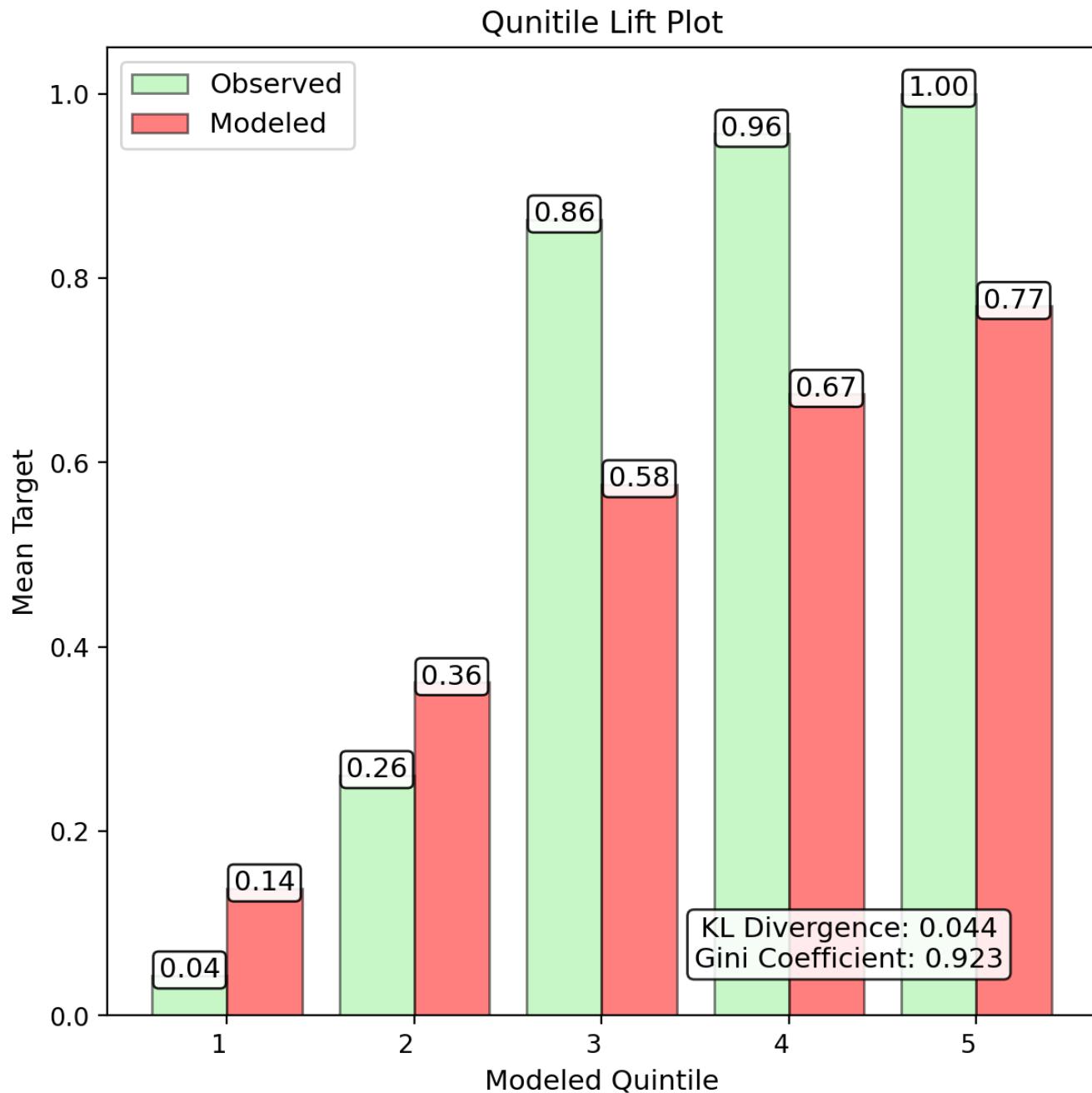


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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

Pc1 - 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

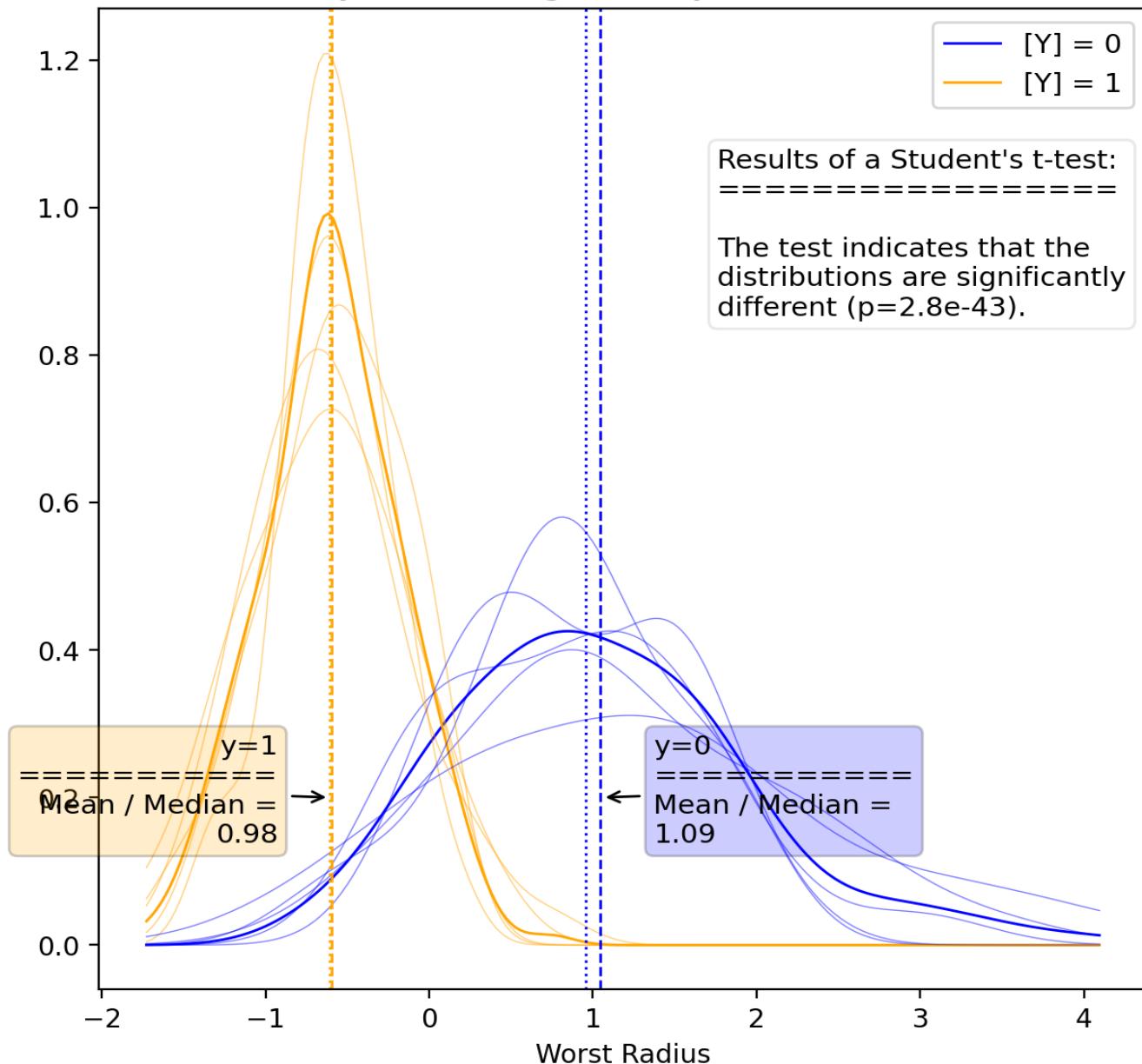
Worst Radius - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-4.85	-5.51	-5.12	-6.38	-5.00	-5.30	0.61
Fitted p-Value	7.5e-15	2.5e-14	3.9e-14	3.2e-12	1.8e-15	1.7e-17	1.4e-12
Fitted Std. Err.	0.623	0.723	0.677	0.915	0.629	0.623	0.120
Conf. Int. Lower	-6.07	-6.92	-6.45	-8.17	-6.23	-6.52	0.85
Conf. Int. Upper	-3.63	-4.09	-3.80	-4.58	-3.77	-4.08	0.38
Train Accuracy	90.8%	90.1%	90.7%	90.3%	90.6%	90.9%	0.3%
Val Accuracy	91.4%	88.4%	93.2%	89.2%	92.7%	90.4%	2.1%
Train AUC	90.2%	90.2%	89.7%	90.2%	89.7%	90.5%	0.2%
Val AUC	88.1%	88.4%	92.2%	89.3%	93.2%	89.5%	2.3%
Train F1	92.5%	92.2%	92.4%	92.2%	92.6%	92.7%	0.2%
Test F1	93.9%	88.9%	95.2%	90.9%	93.7%	92.3%	2.6%
Train Precision	92.3%	94.7%	90.5%	93.9%	92.3%	93.4%	1.6%
Val Precision	90.2%	88.9%	96.2%	93.0%	96.8%	91.7%	3.5%
Train Recall	92.8%	89.9%	94.4%	90.5%	92.8%	92.1%	1.9%
Val Recall	97.9%	88.9%	94.3%	88.9%	90.9%	93.0%	3.9%
Train MCC	80.5%	78.7%	80.5%	79.4%	79.6%	80.7%	0.7%
Val MCC	80.4%	76.8%	83.1%	77.7%	85.3%	79.4%	3.6%
Train Log-Loss	3.33	3.58	3.36	3.51	3.40	3.28	0.11
Val Log-Loss	3.09	4.18	2.47	3.90	2.62	3.48	0.76

Univariate Report

Worst Radius - Kernel Density Plot

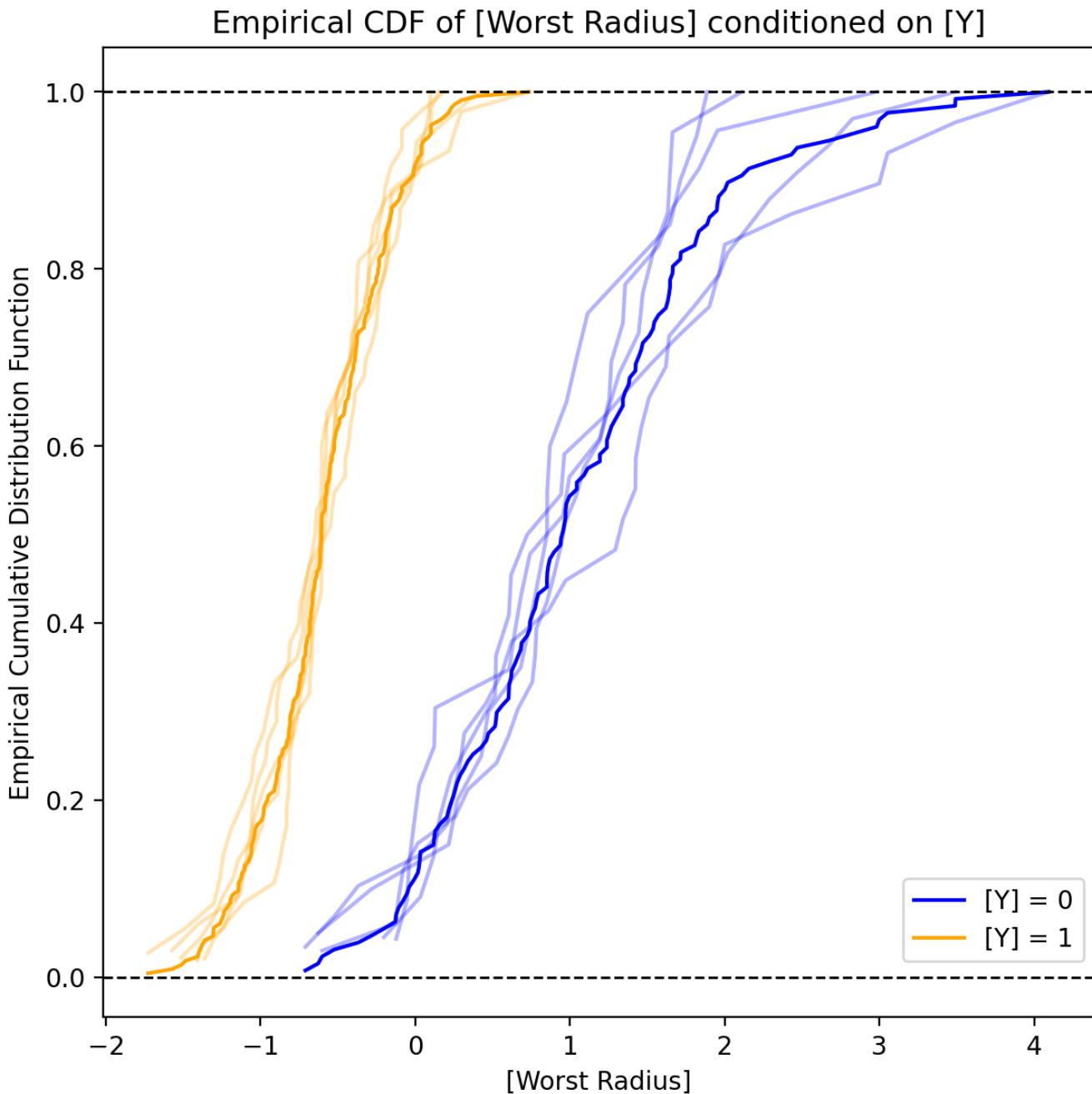
Kernel Density Plot of [Worst Radius] by [Y].
Distributions by level are significantly different at the 95% level.



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

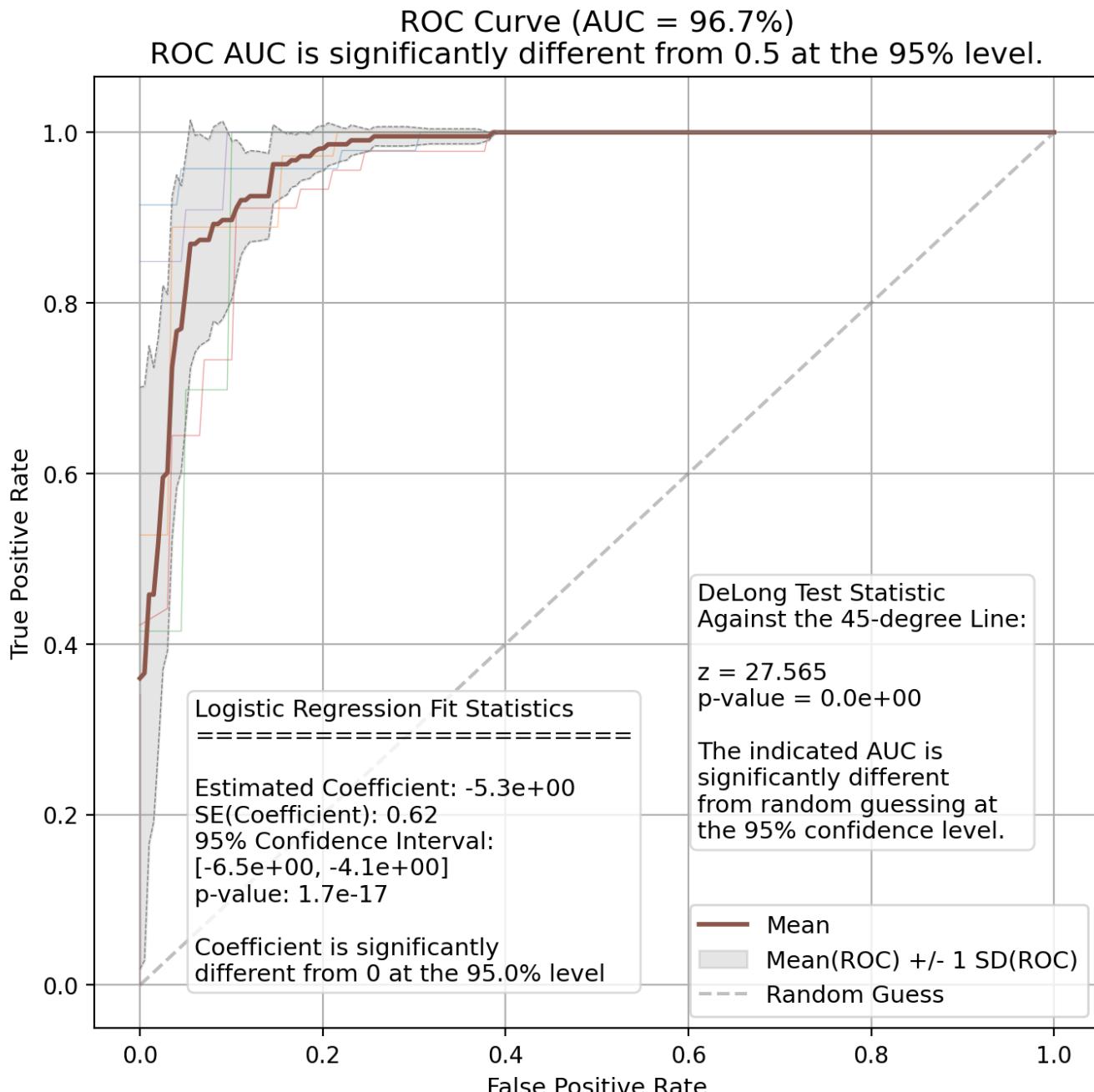
Worst Radius - Empirical CDF Plot



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

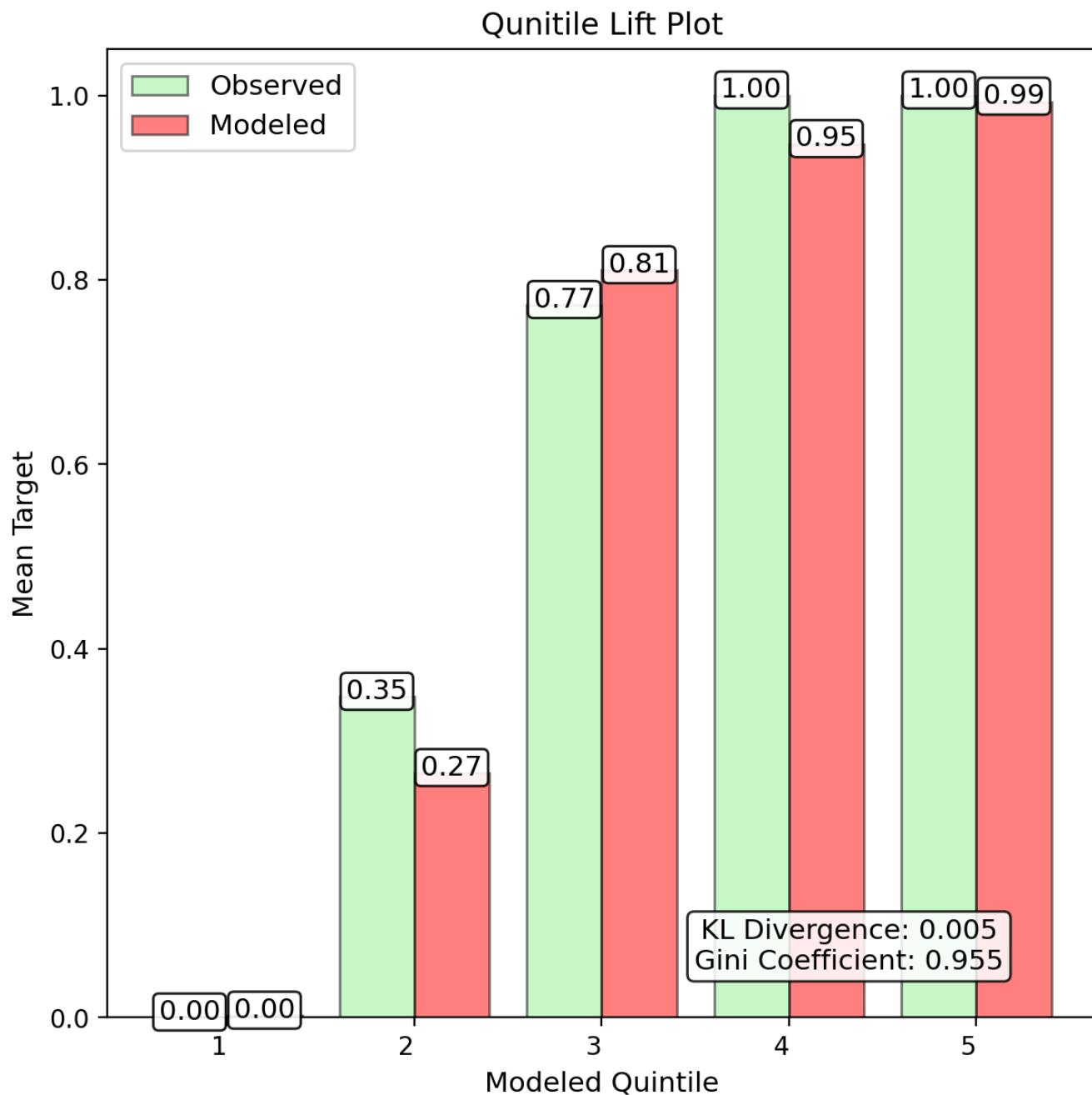


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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

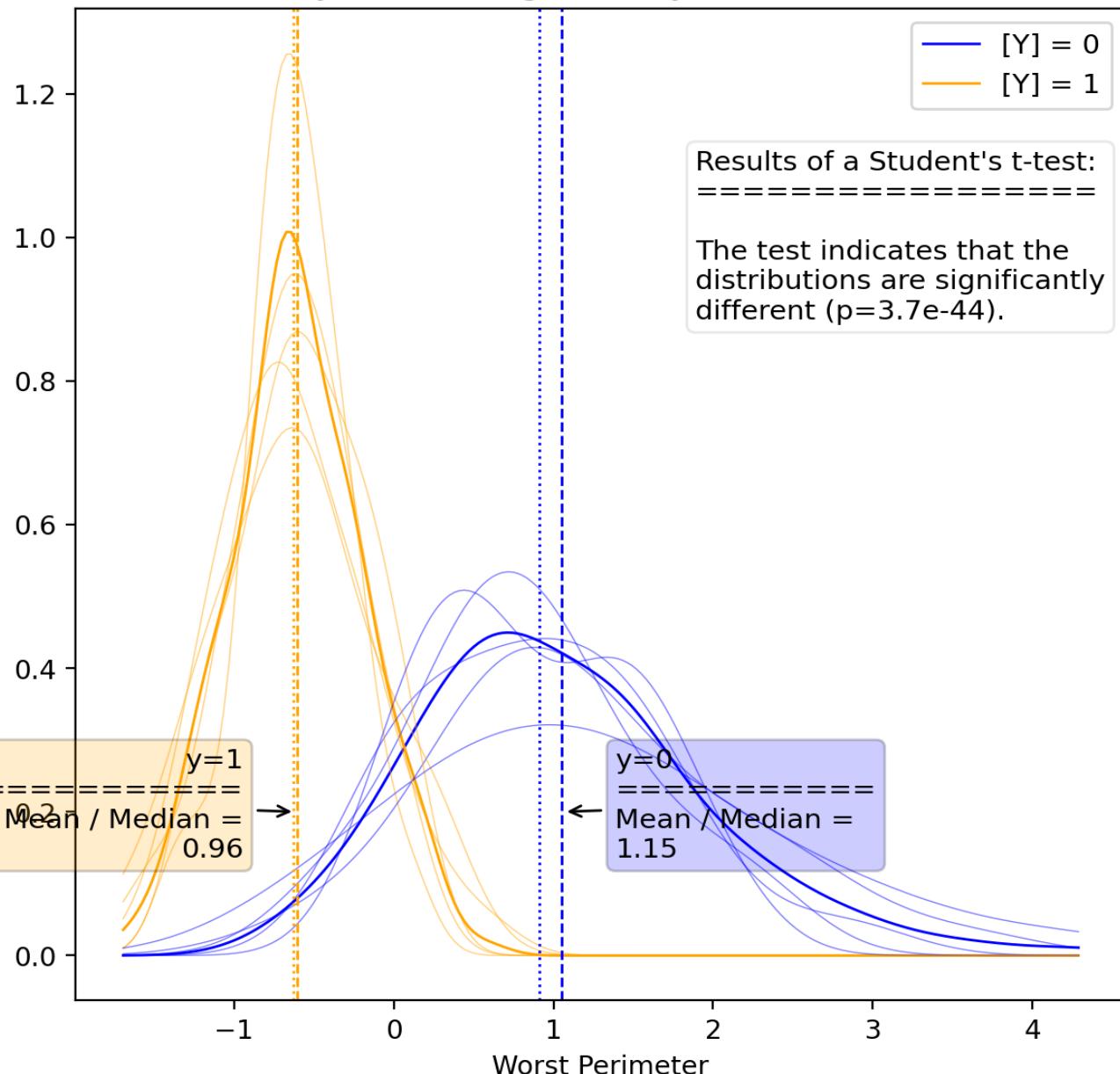
Worst Perimeter - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-5.08	-6.02	-5.35	-6.44	-5.18	-5.55	0.59
Fitted p-Value	9.7e-15	1.2e-13	4.0e-14	1.3e-12	1.4e-15	2.1e-17	5.5e-13
Fitted Std. Err.	0.656	0.812	0.708	0.907	0.648	0.654	0.111
Conf. Int. Lower	-6.37	-7.62	-6.74	-8.21	-6.45	-6.83	0.81
Conf. Int. Upper	-3.79	-4.43	-3.96	-4.66	-3.90	-4.27	0.37
Train Accuracy	91.5%	91.2%	91.8%	91.0%	91.3%	91.5%	0.3%
Val Accuracy	92.9%	87.0%	91.8%	90.5%	92.7%	89.5%	2.4%
Train AUC	91.3%	91.3%	91.3%	91.0%	90.9%	91.1%	0.2%
Val AUC	90.2%	87.1%	91.2%	90.4%	92.4%	88.8%	2.0%
Train F1	93.1%	93.1%	93.2%	92.8%	93.0%	93.2%	0.2%
Test F1	94.8%	87.0%	94.2%	92.1%	93.9%	91.5%	3.2%
Train Precision	93.9%	95.3%	92.6%	94.5%	93.8%	93.8%	1.0%
Val Precision	92.0%	90.9%	96.1%	93.2%	93.9%	91.5%	2.0%
Train Recall	92.2%	91.0%	93.8%	91.1%	92.3%	92.5%	1.1%
Val Recall	97.9%	83.3%	92.5%	91.1%	93.9%	91.5%	5.3%
Train MCC	82.2%	81.0%	82.8%	81.0%	81.3%	81.9%	0.8%
Val MCC	83.7%	74.2%	80.1%	80.3%	84.8%	77.6%	4.1%
Train Log-Loss	3.06	3.18	2.96	3.24	3.15	3.07	0.11
Val Log-Loss	2.57	4.70	2.96	3.41	2.62	3.79	0.88

Univariate Report

Worst Perimeter - Kernel Density Plot

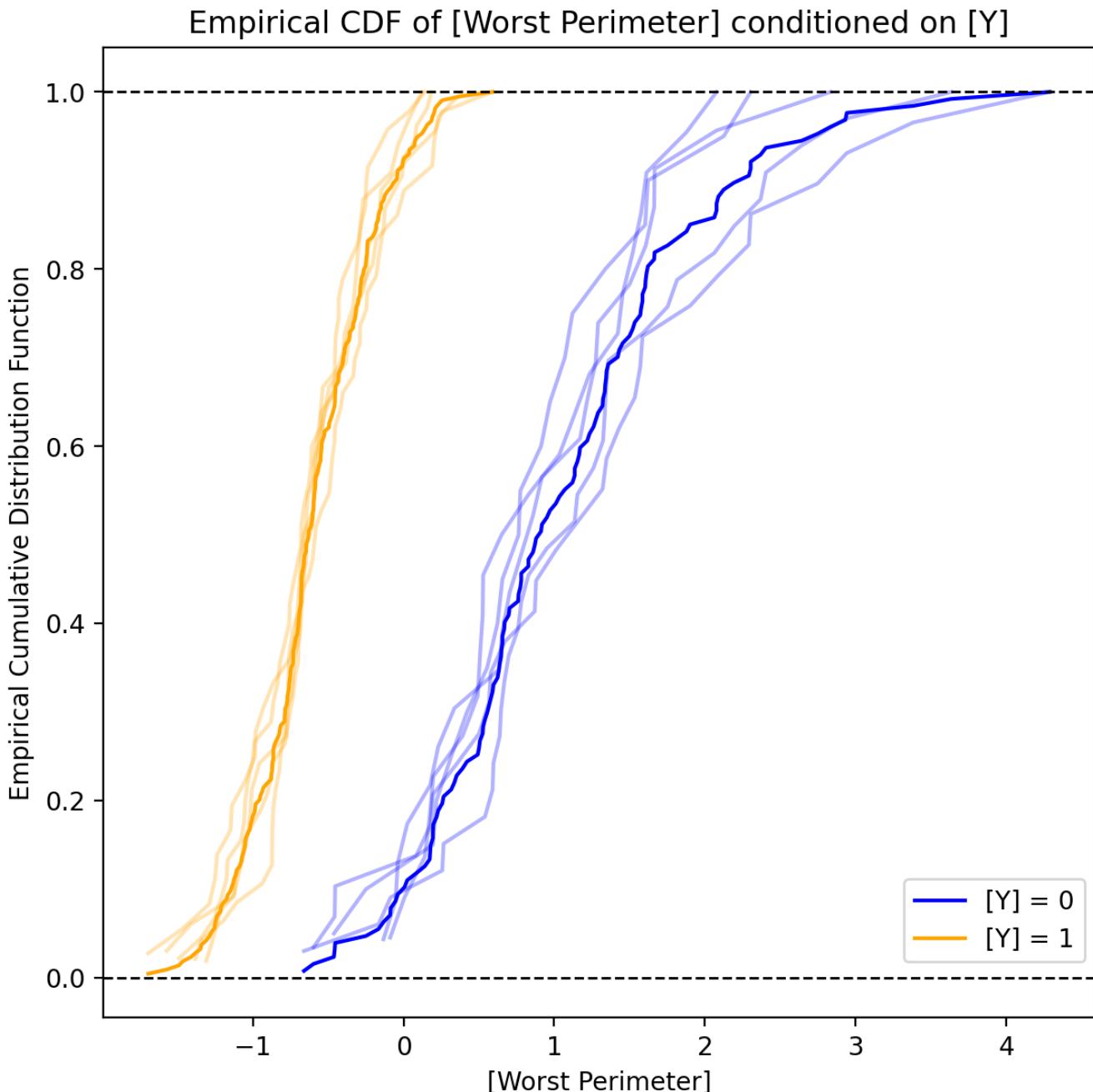
Kernel Density Plot of [Worst Perimeter] by [Y].
Distributions by level are significantly different at the 95% level.



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

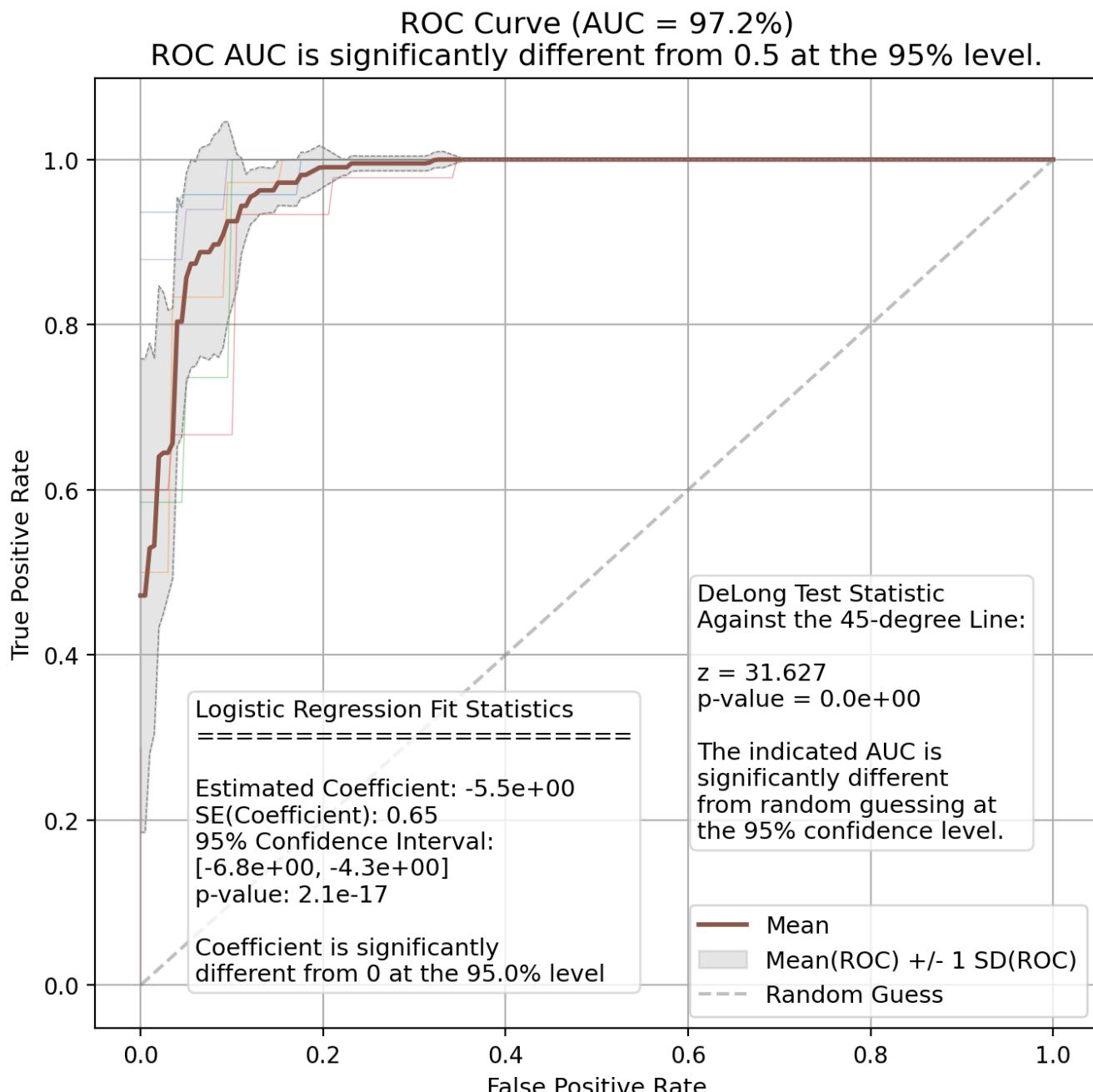
Worst Perimeter - Empirical CDF Plot



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

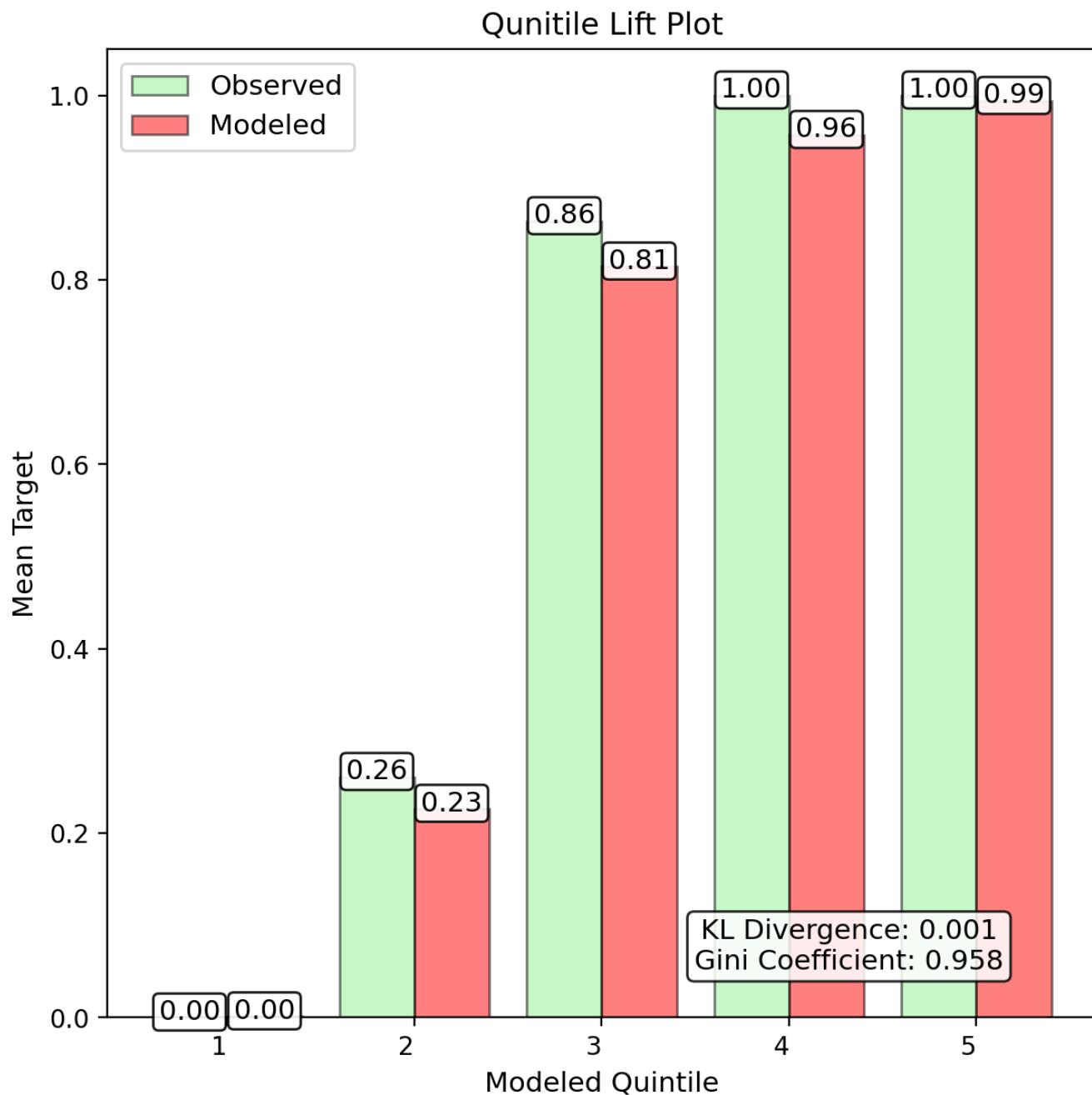


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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.

Univariate Report

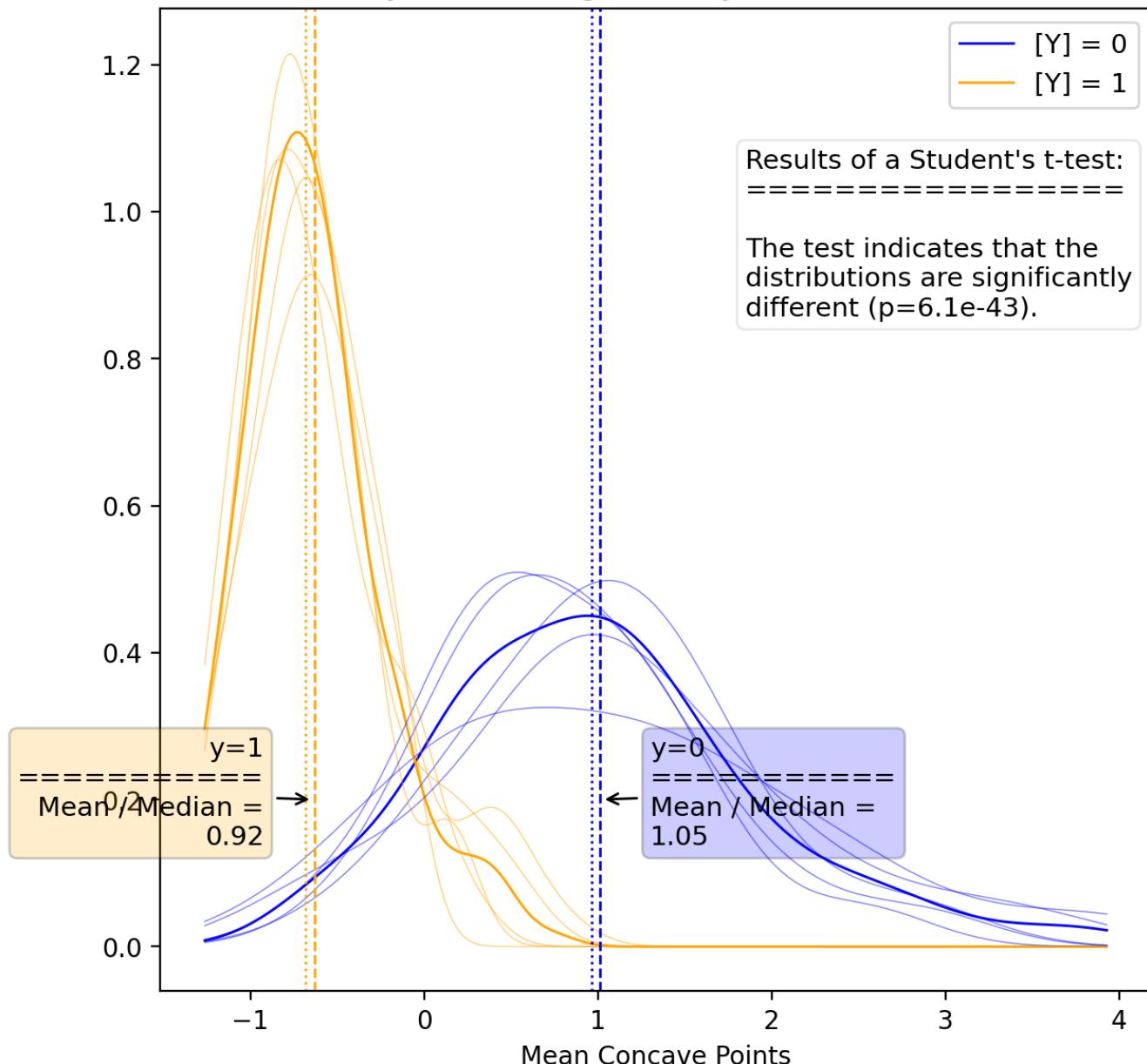
Mean Concave Points - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-4.30	-4.85	-4.88	-4.59	-4.25	-4.55	0.29
Fitted p-Value	2.0e-17	1.4e-16	1.7e-15	6.5e-17	2.0e-18	6.4e-21	7.4e-16
Fitted Std. Err.	0.506	0.586	0.613	0.549	0.486	0.485	0.053
Conf. Int. Lower	-5.29	-5.99	-6.08	-5.67	-5.21	-5.50	0.40
Conf. Int. Upper	-3.30	-3.70	-3.68	-3.51	-3.30	-3.60	0.19
Train Accuracy	91.5%	93.4%	92.9%	91.4%	91.3%	91.8%	1.0%
Val Accuracy	94.3%	87.0%	89.0%	93.2%	96.4%	88.6%	3.9%
Train AUC	91.1%	93.2%	92.5%	91.3%	91.1%	91.5%	1.0%
Val AUC	94.6%	87.0%	89.3%	92.6%	95.5%	89.0%	3.6%
Train F1	93.1%	94.9%	94.1%	93.1%	93.0%	93.4%	0.8%
Test F1	95.7%	87.3%	92.2%	94.5%	97.1%	90.5%	3.8%
Train Precision	93.4%	96.0%	93.8%	94.5%	94.3%	94.3%	1.0%
Val Precision	97.8%	88.6%	95.9%	93.5%	94.3%	93.9%	3.5%
Train Recall	92.8%	93.8%	94.4%	91.7%	91.7%	92.5%	1.2%
Val Recall	93.6%	86.1%	88.7%	95.6%	100.0%	87.3%	5.5%
Train MCC	82.1%	85.6%	85.2%	81.7%	81.4%	82.6%	2.0%
Val MCC	87.5%	73.9%	74.7%	85.8%	92.6%	76.6%	8.2%
Train Log-Loss	3.06	2.39	2.56	3.10	3.15	2.96	0.35
Val Log-Loss	2.06	4.70	3.95	2.44	1.31	4.11	1.40

Univariate Report

Mean Concave Points - Kernel Density Plot

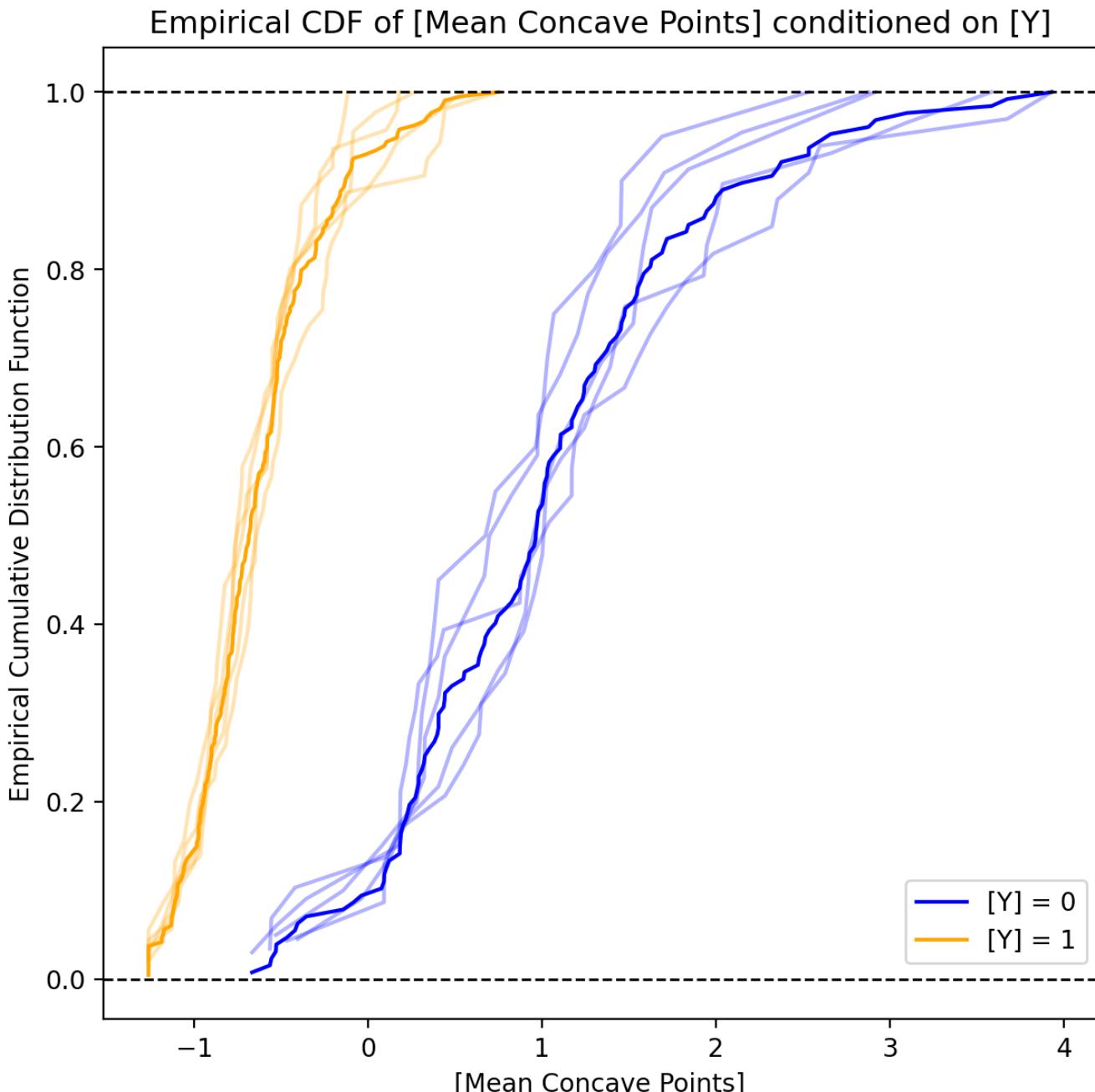
Kernel Density Plot of [Mean Concave Points] by [Y].
Distributions by level are significantly different at the 95% level.



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

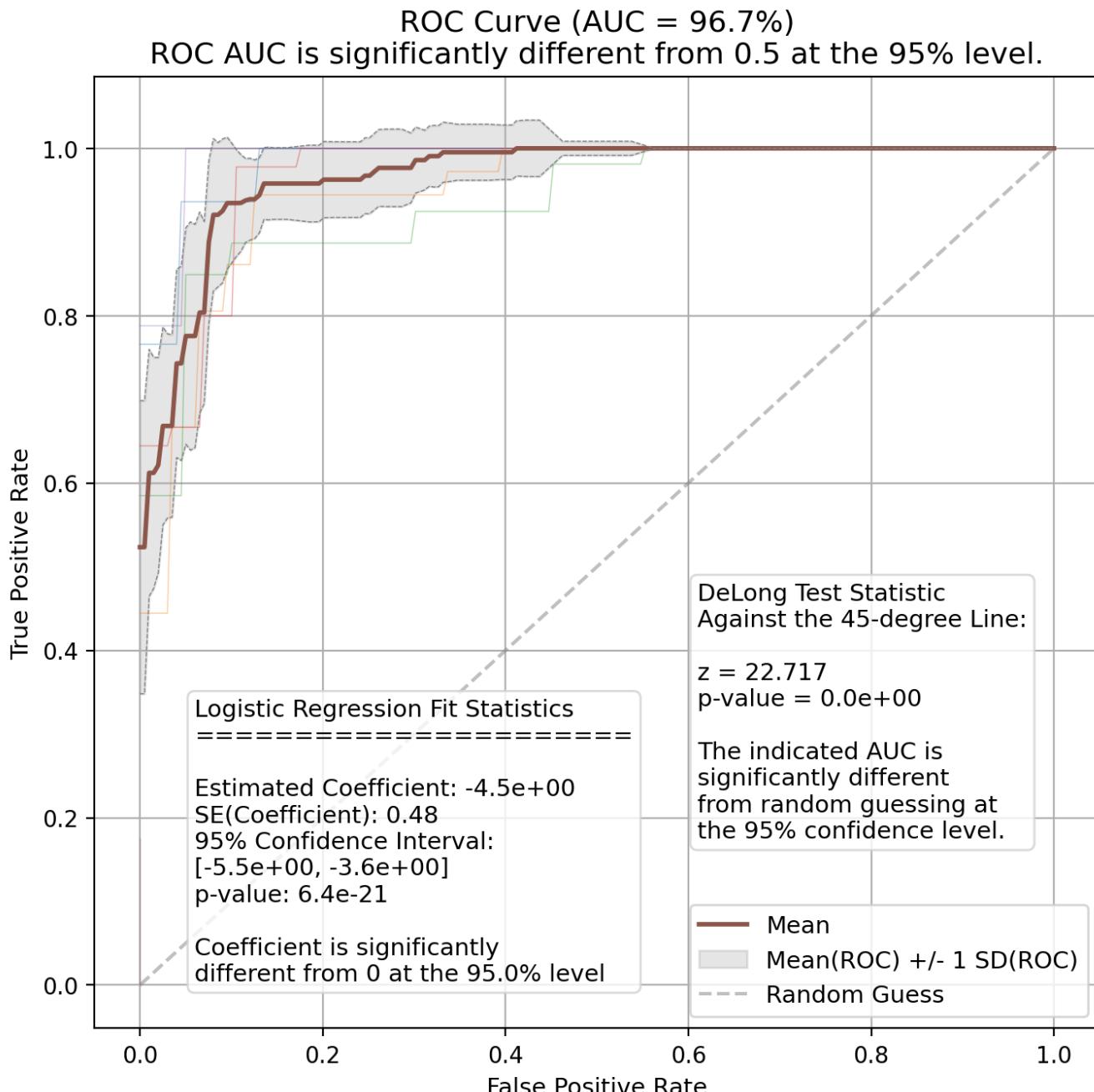
Mean Concave Points - Empirical CDF Plot



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 Concave Points - ROC Curve

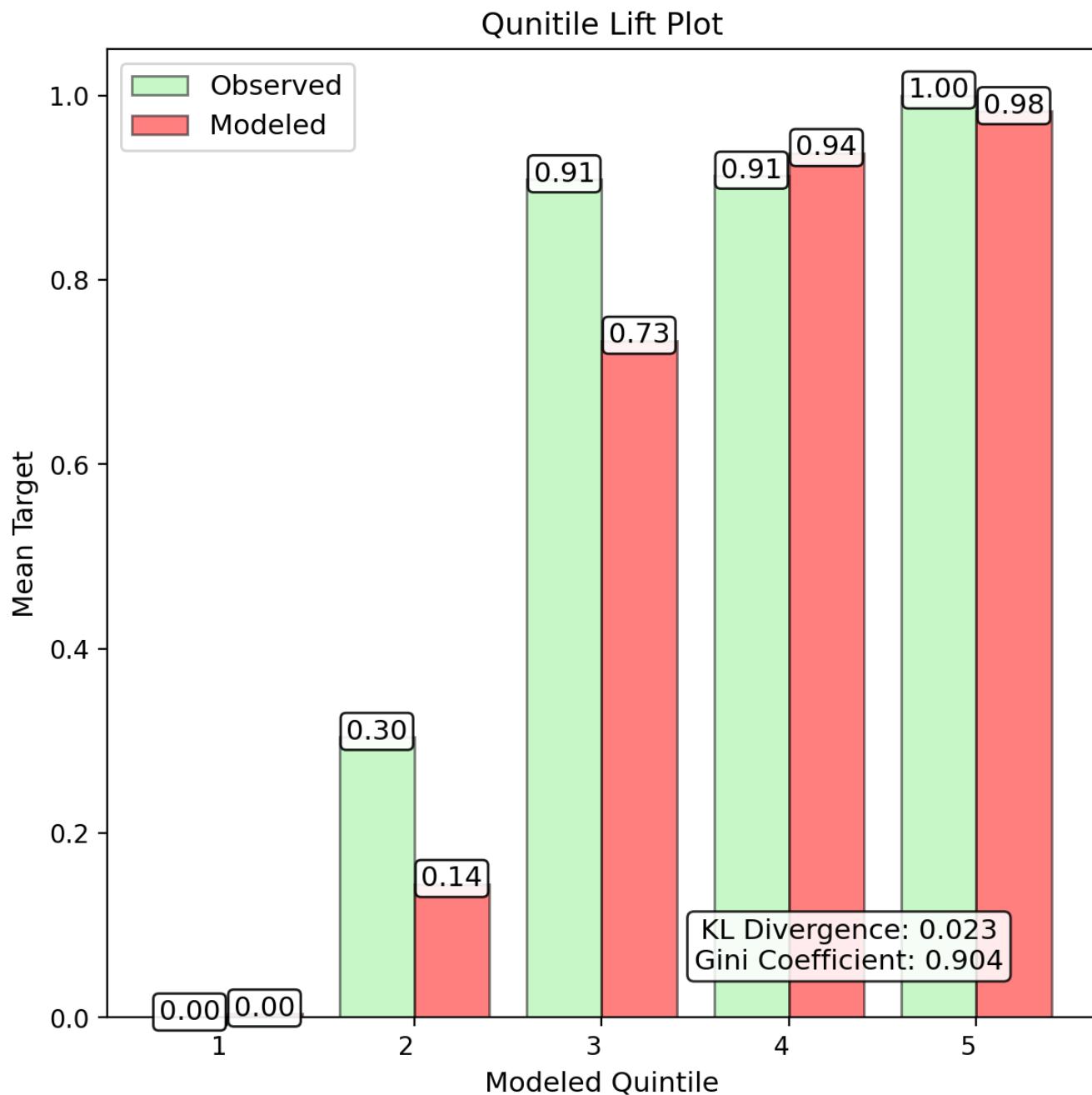


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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 Concave Points - 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

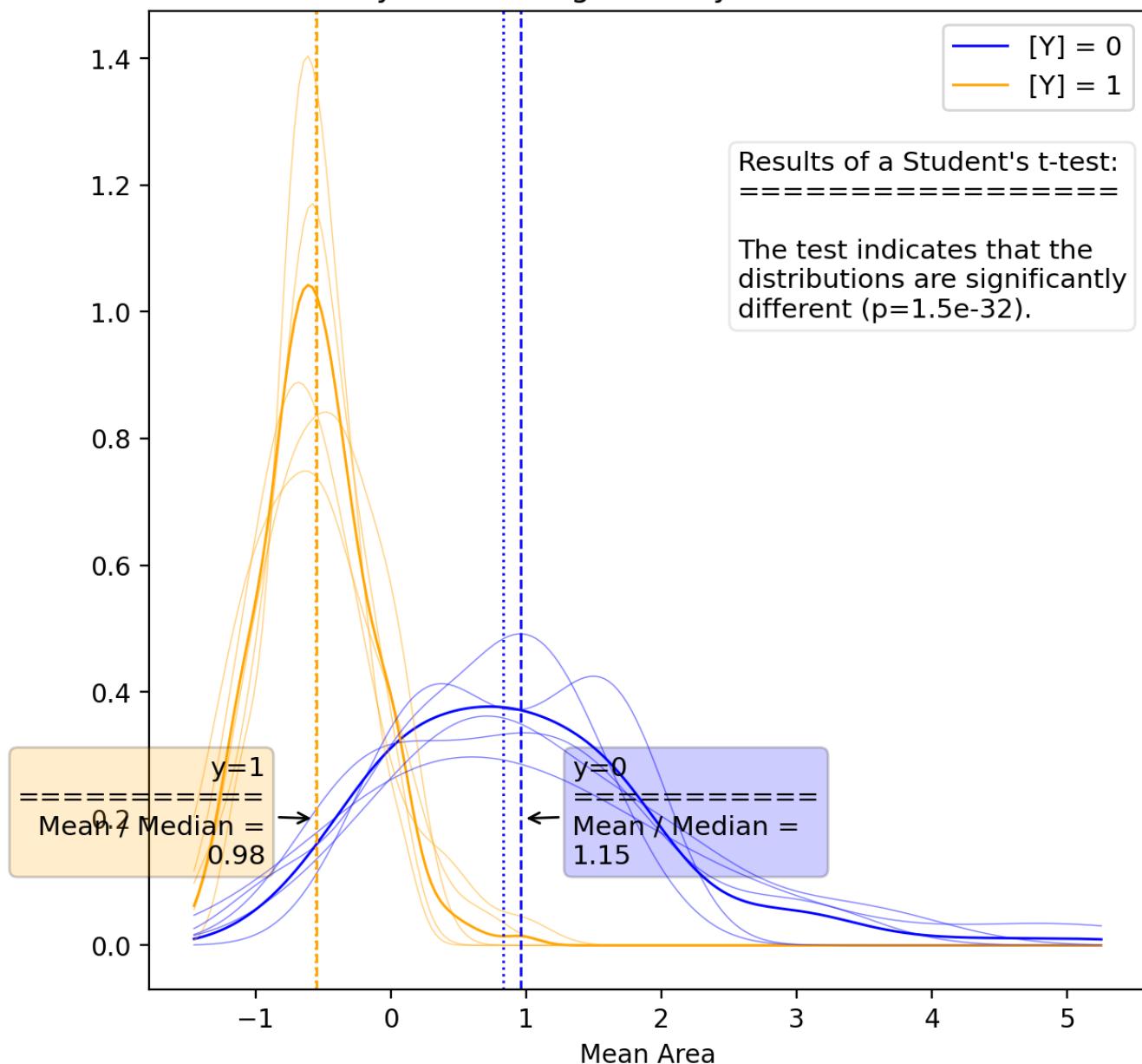
Mean Area - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-4.00	-4.34	-4.19	-4.87	-3.94	-4.23	0.37
Fitted p-Value	1.4e-16	6.1e-17	1.1e-16	2.1e-15	6.3e-18	1.4e-20	9.0e-16
Fitted Std. Err.	0.484	0.519	0.505	0.614	0.456	0.455	0.060
Conf. Int. Lower	-4.95	-5.36	-5.18	-6.07	-4.83	-5.13	0.49
Conf. Int. Upper	-3.06	-3.32	-3.20	-3.67	-3.04	-3.34	0.26
Train Accuracy	89.3%	87.1%	89.2%	87.3%	87.4%	88.3%	1.1%
Val Accuracy	88.6%	85.5%	87.7%	86.5%	92.7%	88.6%	2.8%
Train AUC	88.2%	86.2%	87.7%	86.3%	86.1%	87.0%	1.0%
Val AUC	83.7%	85.6%	83.7%	85.8%	91.7%	86.7%	3.3%
Train F1	91.4%	90.1%	91.3%	89.9%	90.2%	90.8%	0.7%
Test F1	92.0%	85.7%	91.6%	88.9%	94.1%	91.2%	3.2%
Train Precision	90.1%	90.9%	87.9%	89.9%	89.2%	89.5%	1.1%
Val Precision	86.8%	88.2%	90.7%	88.9%	91.4%	88.2%	1.9%
Train Recall	92.8%	89.3%	95.0%	89.9%	91.2%	92.1%	2.3%
Val Recall	97.9%	83.3%	92.5%	88.9%	97.0%	94.4%	6.0%
Train MCC	77.2%	71.8%	77.4%	72.6%	72.7%	74.7%	2.7%
Val MCC	73.9%	71.2%	68.6%	71.6%	84.9%	75.5%	6.4%
Train Log-Loss	3.86	4.64	3.90	4.59	4.54	4.23	0.39
Val Log-Loss	4.12	5.22	4.44	4.87	2.62	4.11	1.01

Univariate Report

Mean Area - Kernel Density Plot

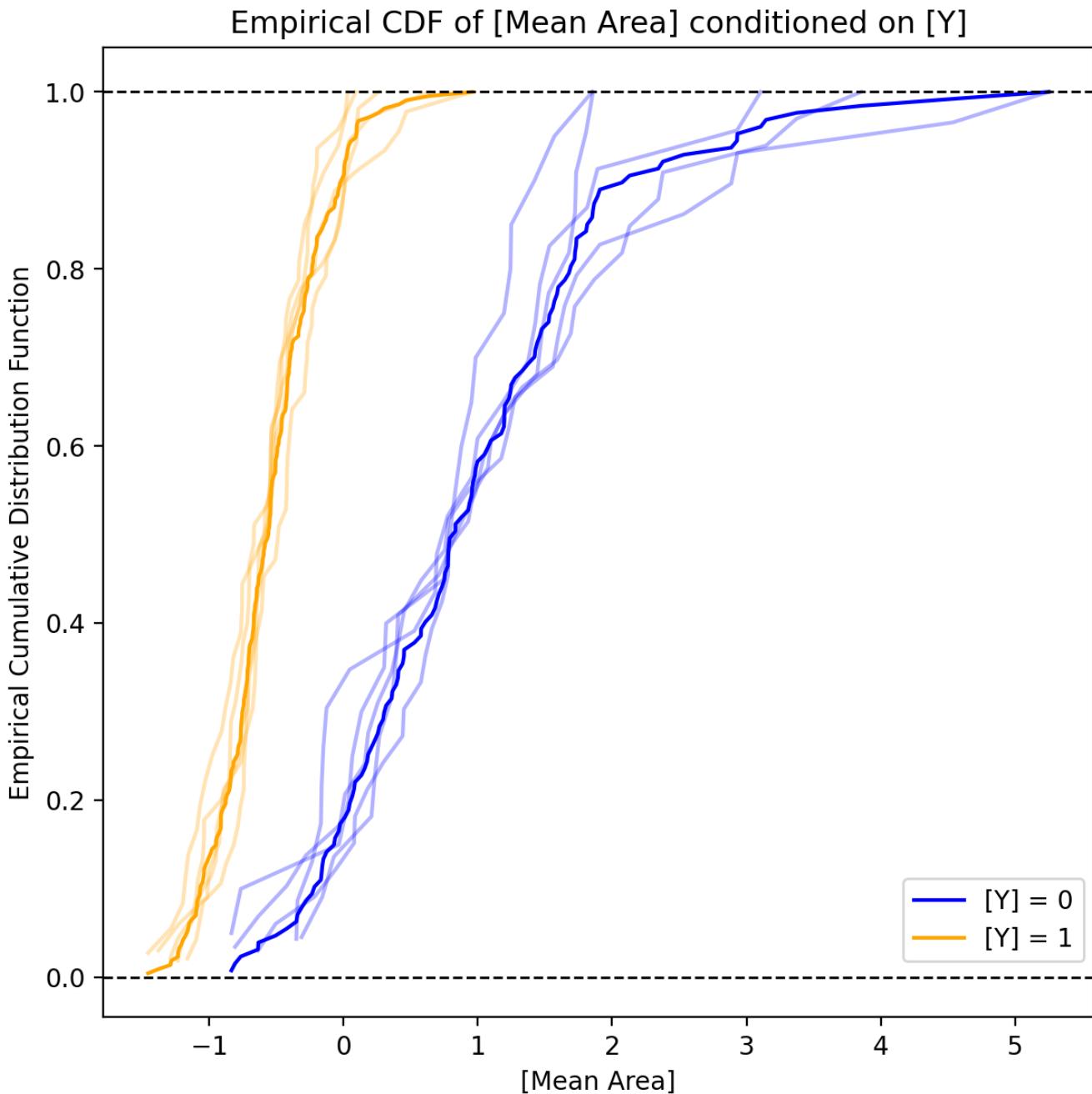
Kernel Density Plot of [Mean Area] by [Y].
Distributions by level are significantly different at the 95% level.



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

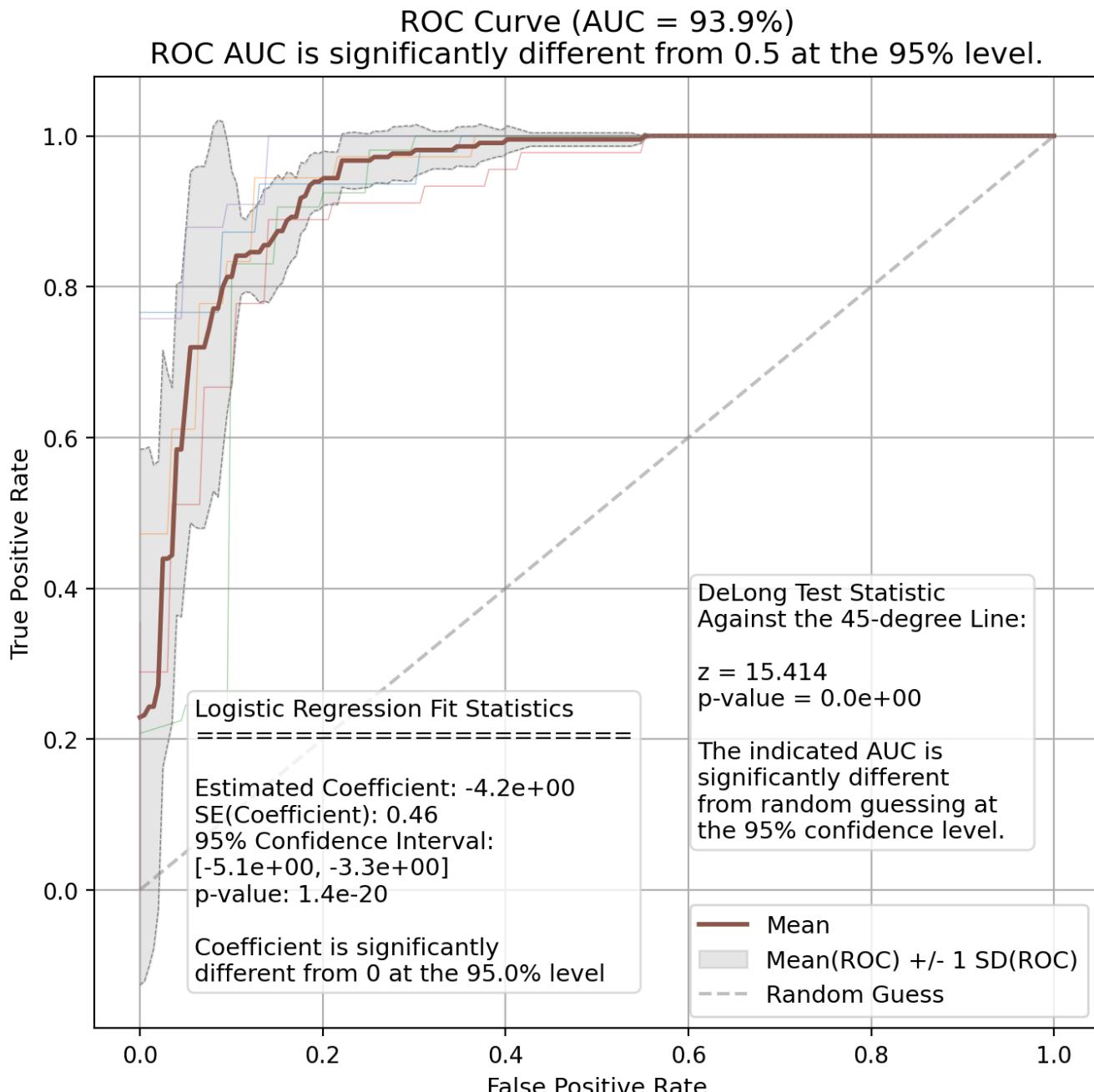
Mean Area - Empirical CDF Plot



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 Area - ROC Curve

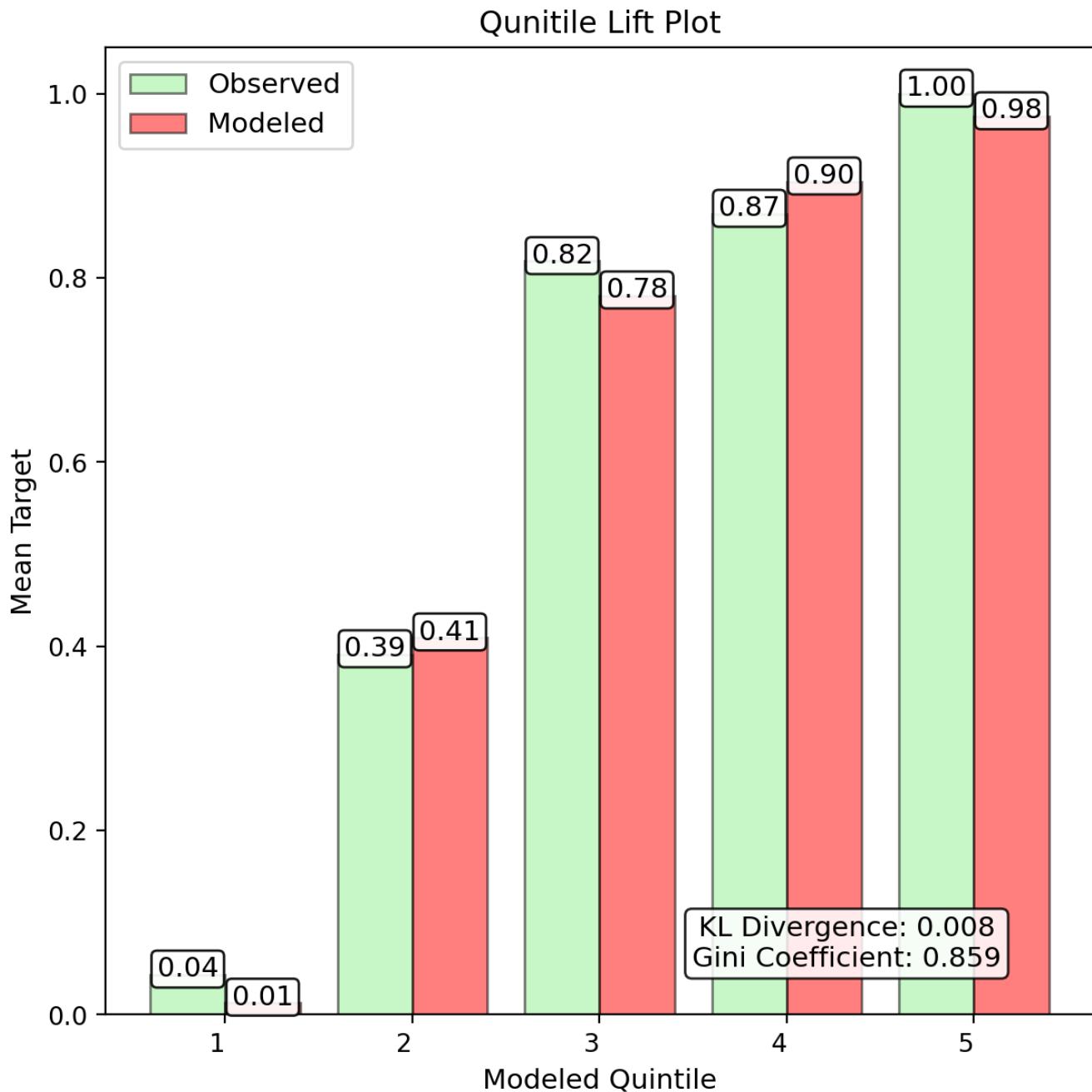


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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 Area - 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

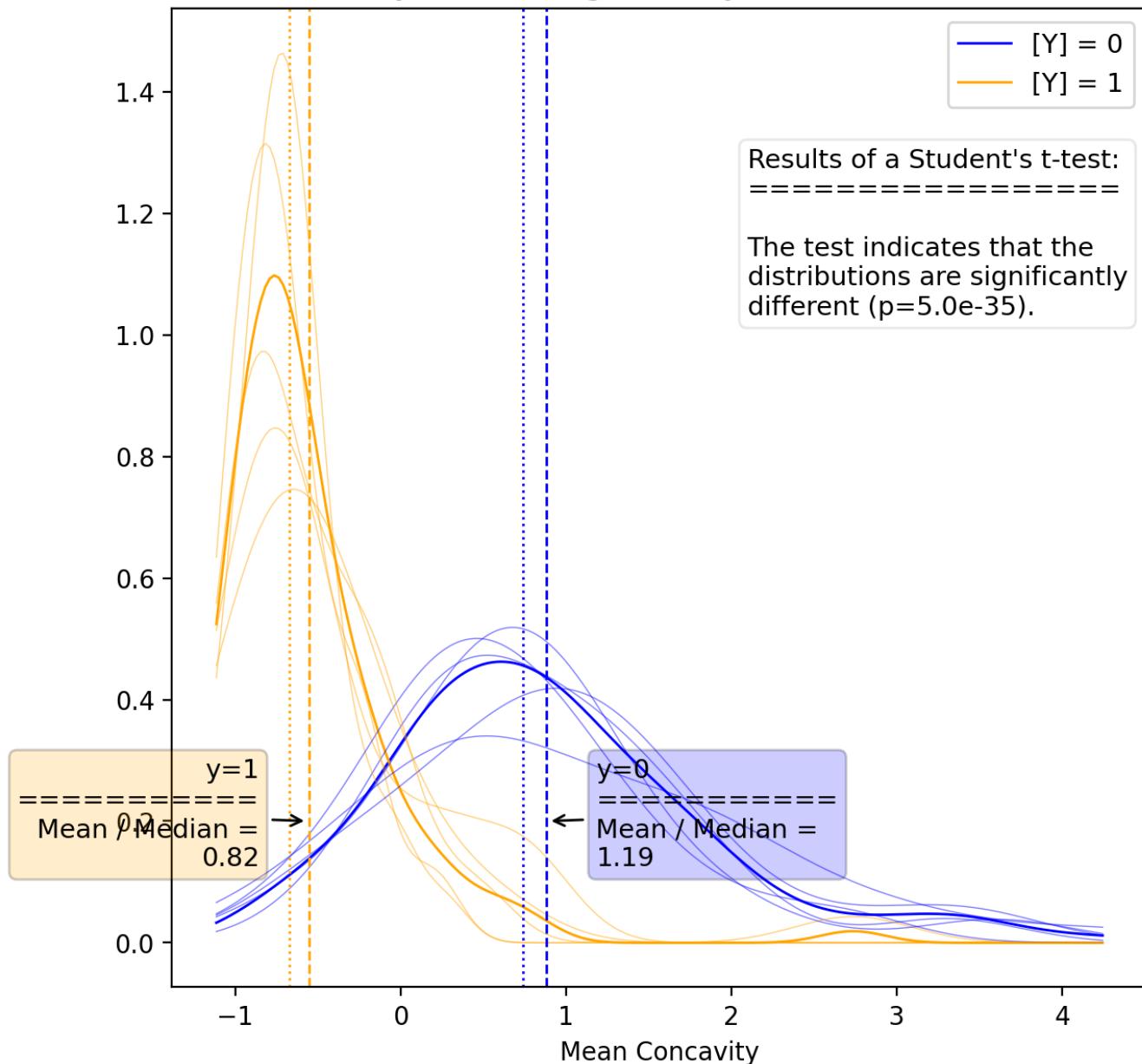
Mean Concavity - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-2.79	-3.44	-3.78	-2.91	-3.01	-3.15	0.41
Fitted p-Value	4.6e-19	1.2e-19	1.3e-18	6.9e-19	2.1e-20	4.9e-24	5.2e-19
Fitted Std. Err.	0.313	0.380	0.429	0.328	0.325	0.311	0.049
Conf. Int. Lower	-3.40	-4.18	-4.62	-3.55	-3.65	-3.76	0.51
Conf. Int. Upper	-2.18	-2.70	-2.94	-2.27	-2.37	-2.54	0.32
Train Accuracy	87.1%	90.1%	89.6%	87.6%	88.1%	88.6%	1.3%
Val Accuracy	94.3%	82.6%	84.9%	91.9%	90.9%	87.7%	4.9%
Train AUC	87.0%	90.2%	89.3%	87.7%	87.8%	88.3%	1.3%
Val AUC	93.5%	82.8%	85.0%	90.9%	90.9%	87.4%	4.5%
Train F1	89.3%	92.2%	91.3%	90.0%	90.4%	90.7%	1.1%
Test F1	95.7%	82.4%	89.1%	93.5%	92.3%	90.0%	5.2%
Train Precision	91.2%	94.7%	91.8%	92.5%	92.0%	92.3%	1.3%
Val Precision	95.7%	87.5%	93.8%	91.5%	93.8%	91.3%	3.1%
Train Recall	87.4%	89.9%	90.7%	87.6%	89.0%	89.3%	1.4%
Val Recall	95.7%	77.8%	84.9%	95.6%	90.9%	88.7%	7.7%
Train MCC	73.1%	78.7%	78.3%	74.1%	74.8%	75.9%	2.5%
Val MCC	87.0%	65.8%	65.7%	82.9%	81.3%	74.2%	10.1%
Train Log-Loss	4.66	3.58	3.77	4.45	4.28	4.12	0.46
Val Log-Loss	2.06	6.27	5.43	2.92	3.28	4.43	1.78

Univariate Report

Mean Concavity - Kernel Density Plot

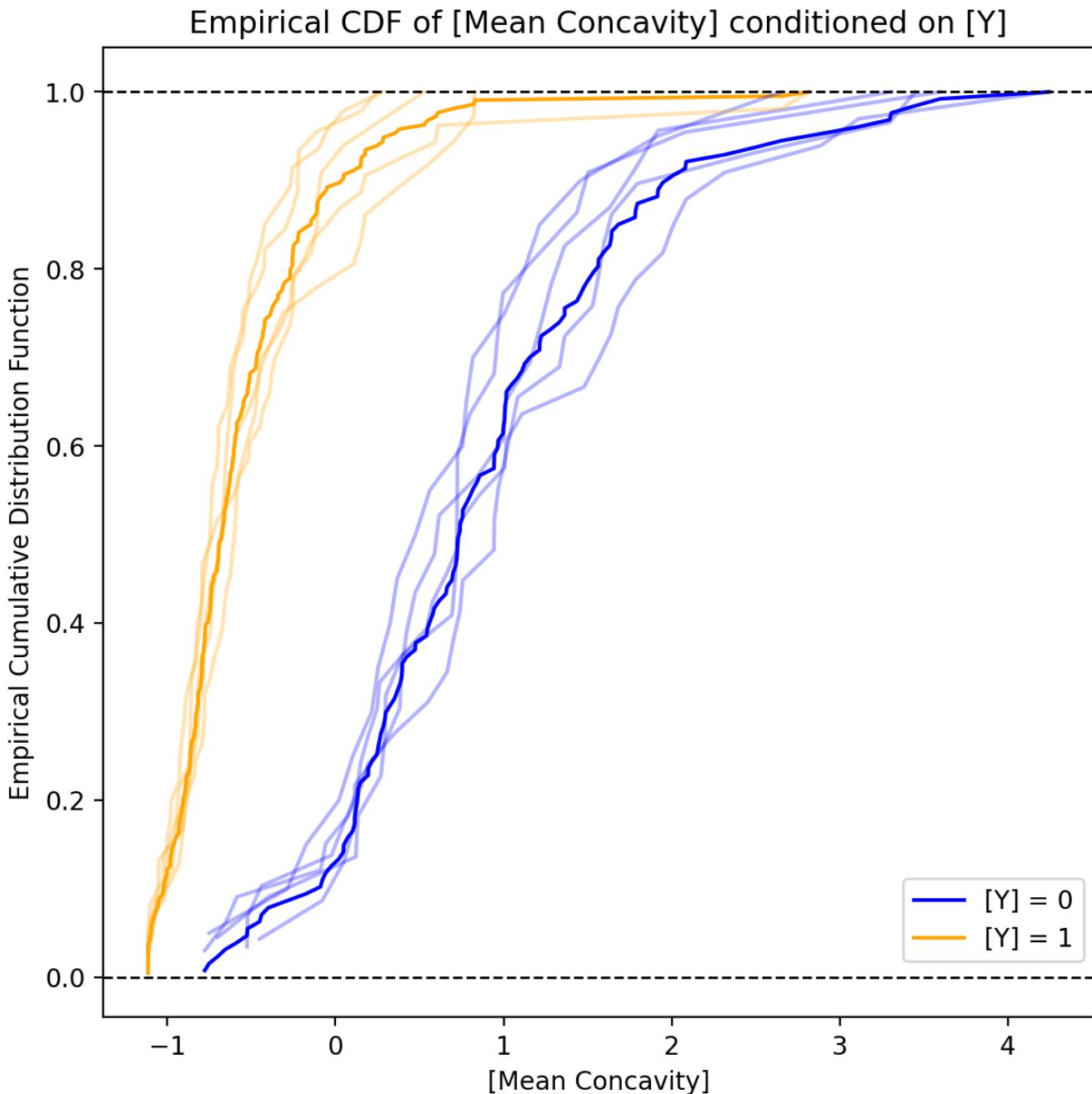
Kernel Density Plot of [Mean Concavity] by [Y].
Distributions by level are significantly different at the 95% level.



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

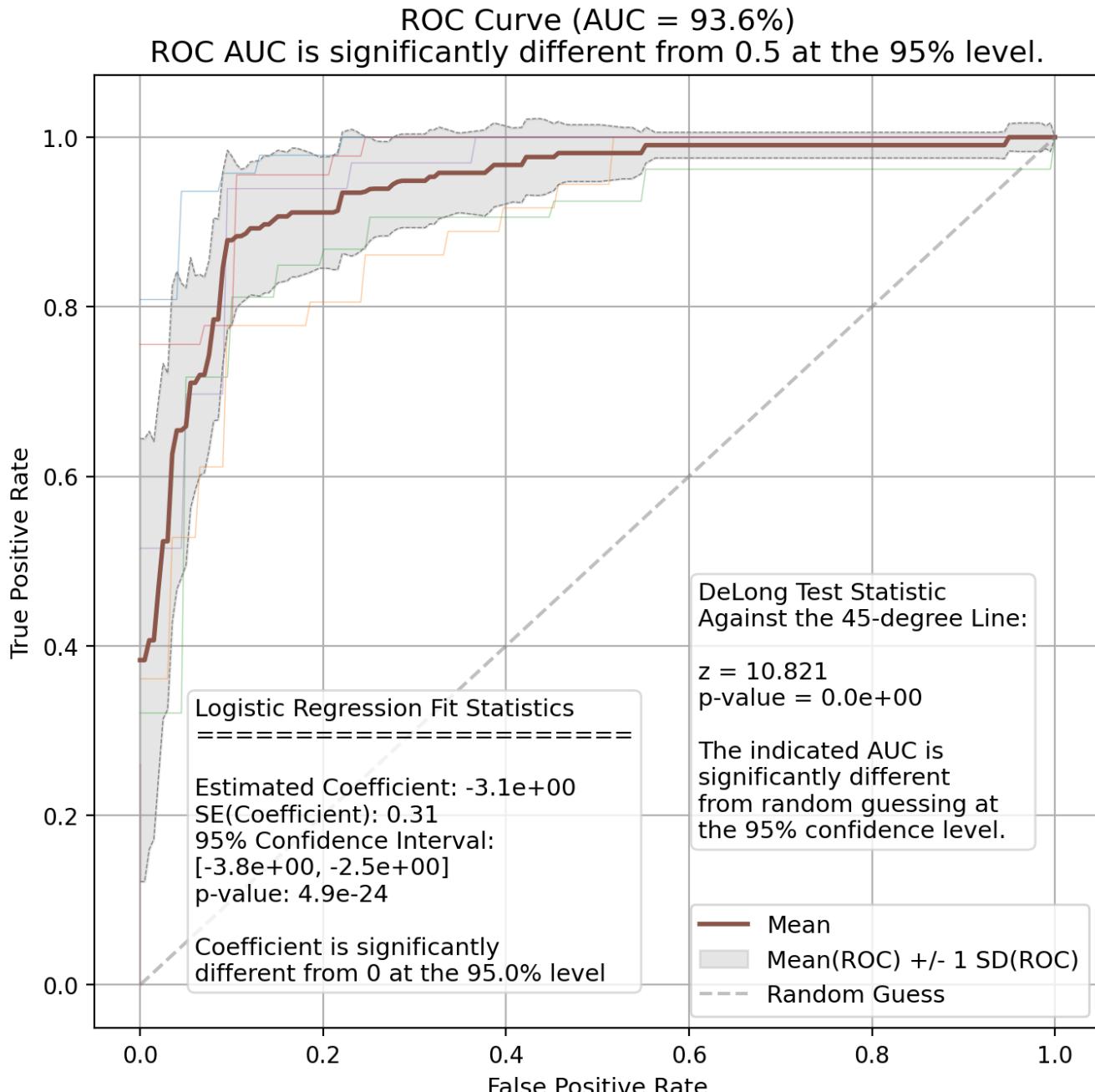
Mean Concavity - Empirical CDF Plot



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 Concavity - ROC Curve

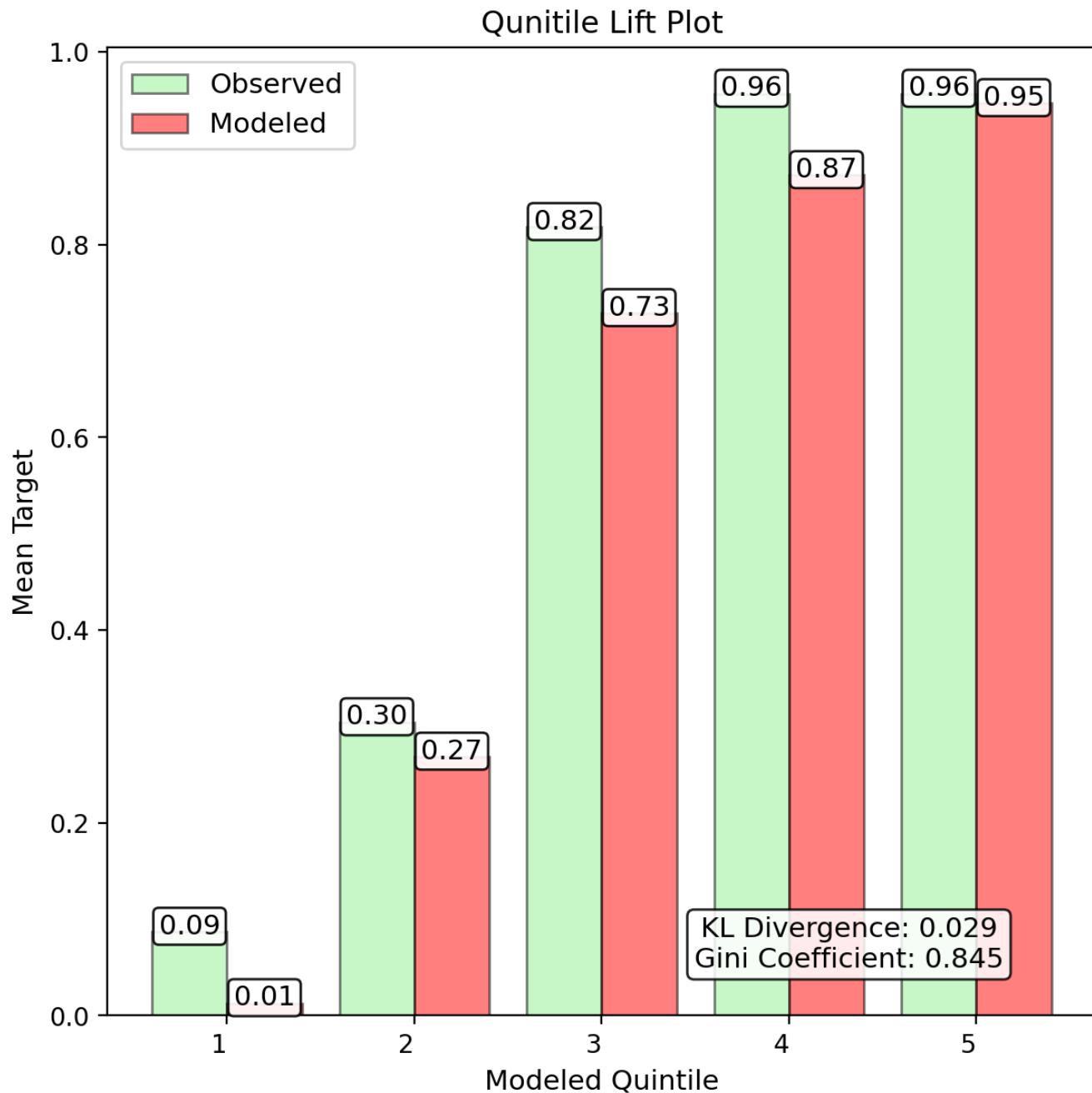


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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 Concavity - 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

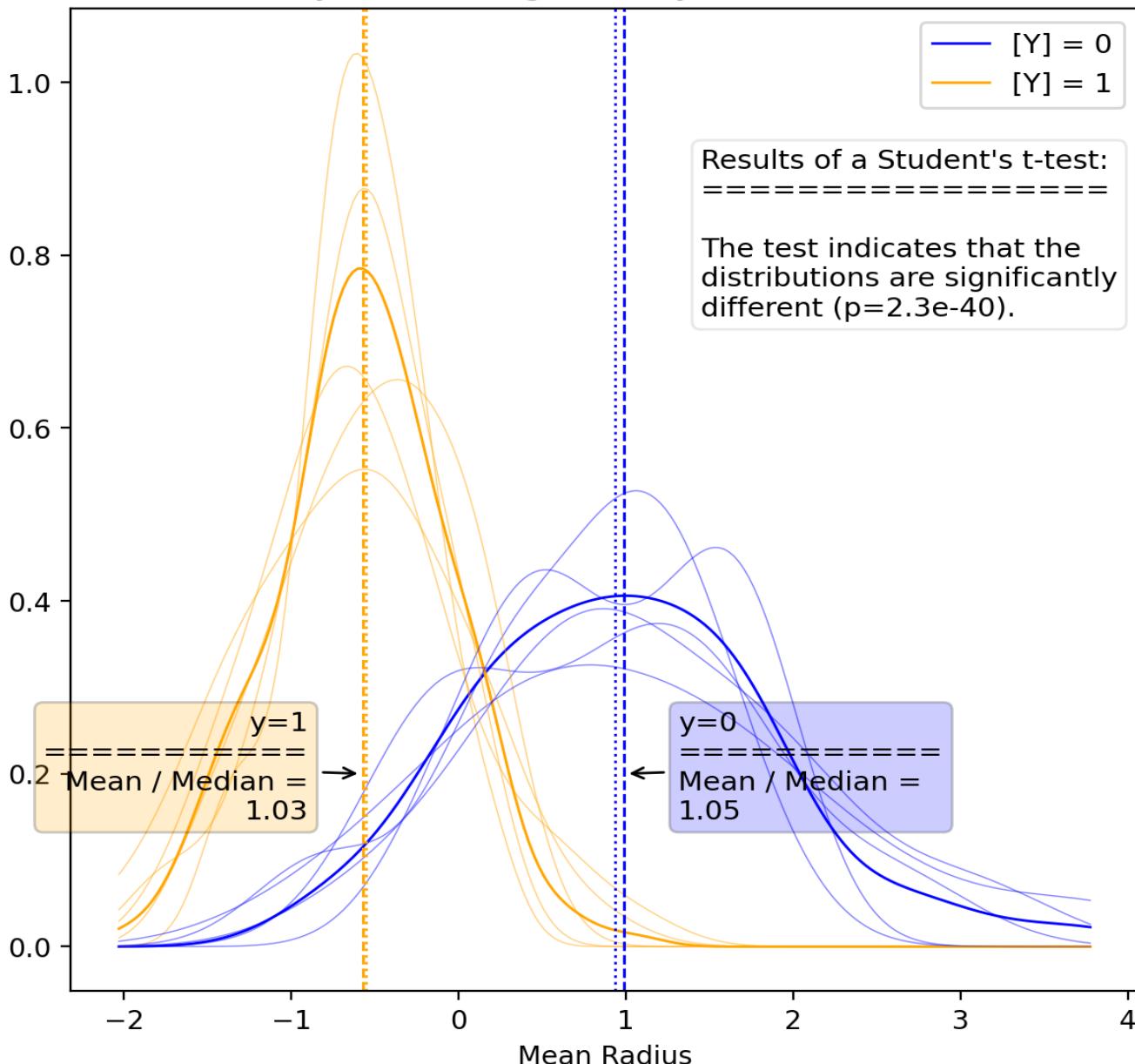
Mean Radius - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-3.38	-3.80	-3.77	-4.27	-3.40	-3.69	0.36
Fitted p-Value	1.9e-16	2.2e-16	5.6e-16	4.6e-15	2.4e-17	5.6e-20	1.9e-15
Fitted Std. Err.	0.411	0.463	0.466	0.545	0.402	0.403	0.057
Conf. Int. Lower	-4.19	-4.71	-4.68	-5.34	-4.19	-4.48	0.47
Conf. Int. Upper	-2.58	-2.89	-2.86	-3.20	-2.62	-2.90	0.25
Train Accuracy	85.6%	85.3%	86.6%	86.5%	85.7%	86.5%	0.6%
Val Accuracy	85.7%	85.5%	84.9%	85.1%	90.9%	86.8%	2.5%
Train AUC	86.0%	85.2%	85.8%	86.6%	85.7%	86.5%	0.5%
Val AUC	81.6%	85.7%	86.5%	85.3%	90.9%	86.2%	3.3%
Train F1	87.9%	88.4%	88.9%	89.0%	88.3%	88.9%	0.5%
Test F1	89.8%	85.3%	88.9%	87.4%	92.3%	89.4%	2.6%
Train Precision	91.6%	91.6%	88.3%	91.8%	91.2%	91.6%	1.4%
Val Precision	86.3%	90.6%	95.7%	90.5%	93.8%	90.0%	3.6%
Train Recall	84.4%	85.4%	89.4%	86.4%	85.6%	86.4%	1.9%
Val Recall	93.6%	80.6%	83.0%	84.4%	90.9%	88.7%	5.5%
Train MCC	70.6%	68.7%	71.9%	71.8%	70.0%	71.9%	1.3%
Val MCC	66.7%	71.6%	67.5%	69.6%	81.3%	72.1%	5.9%
Train Log-Loss	5.19	5.30	4.84	4.86	5.17	4.86	0.21
Val Log-Loss	5.15	5.22	5.43	5.36	3.28	4.74	0.91

Univariate Report

Mean Radius - Kernel Density Plot

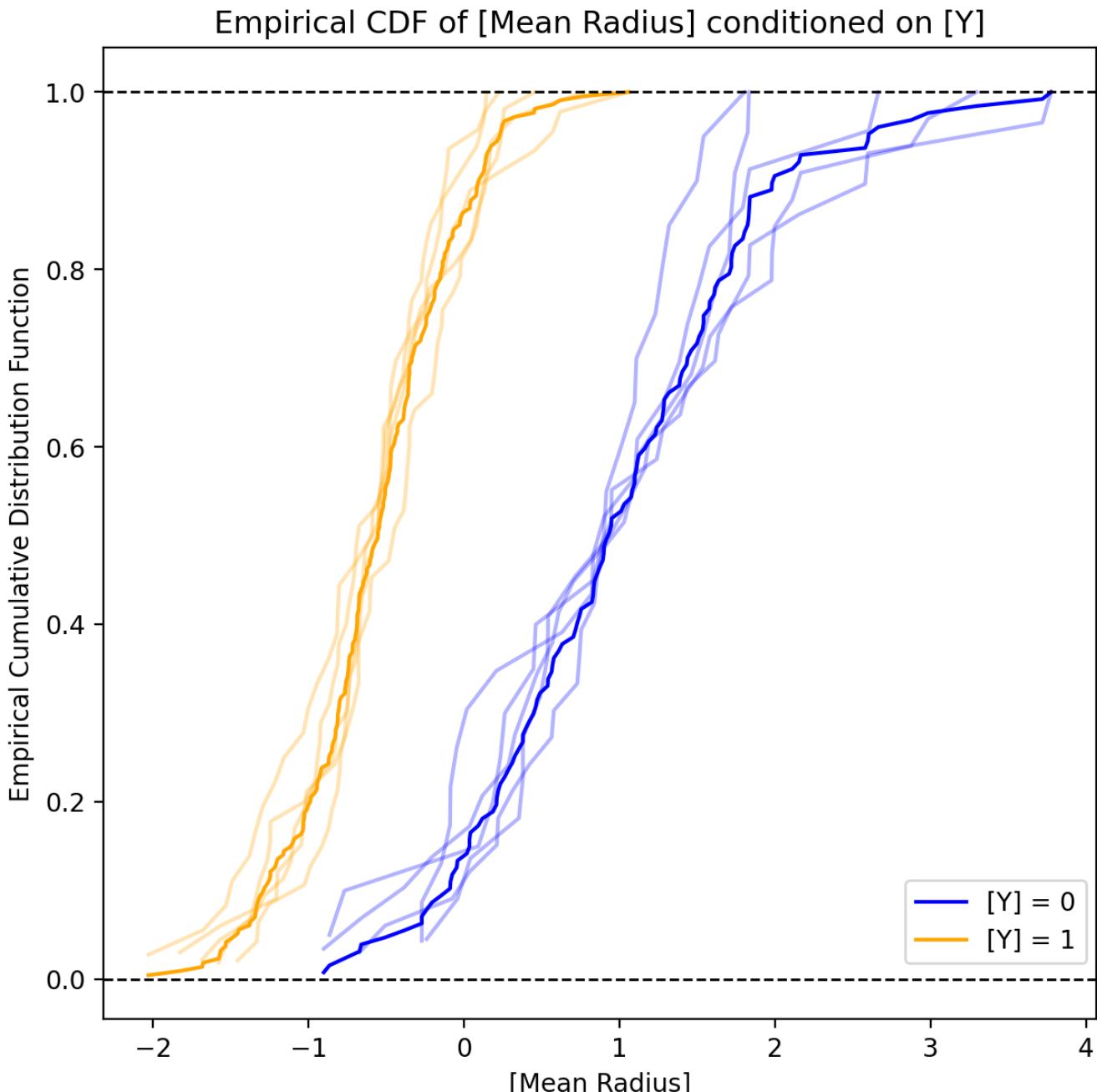
Kernel Density Plot of [Mean Radius] by [Y].
Distributions by level are significantly different at the 95% level.



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

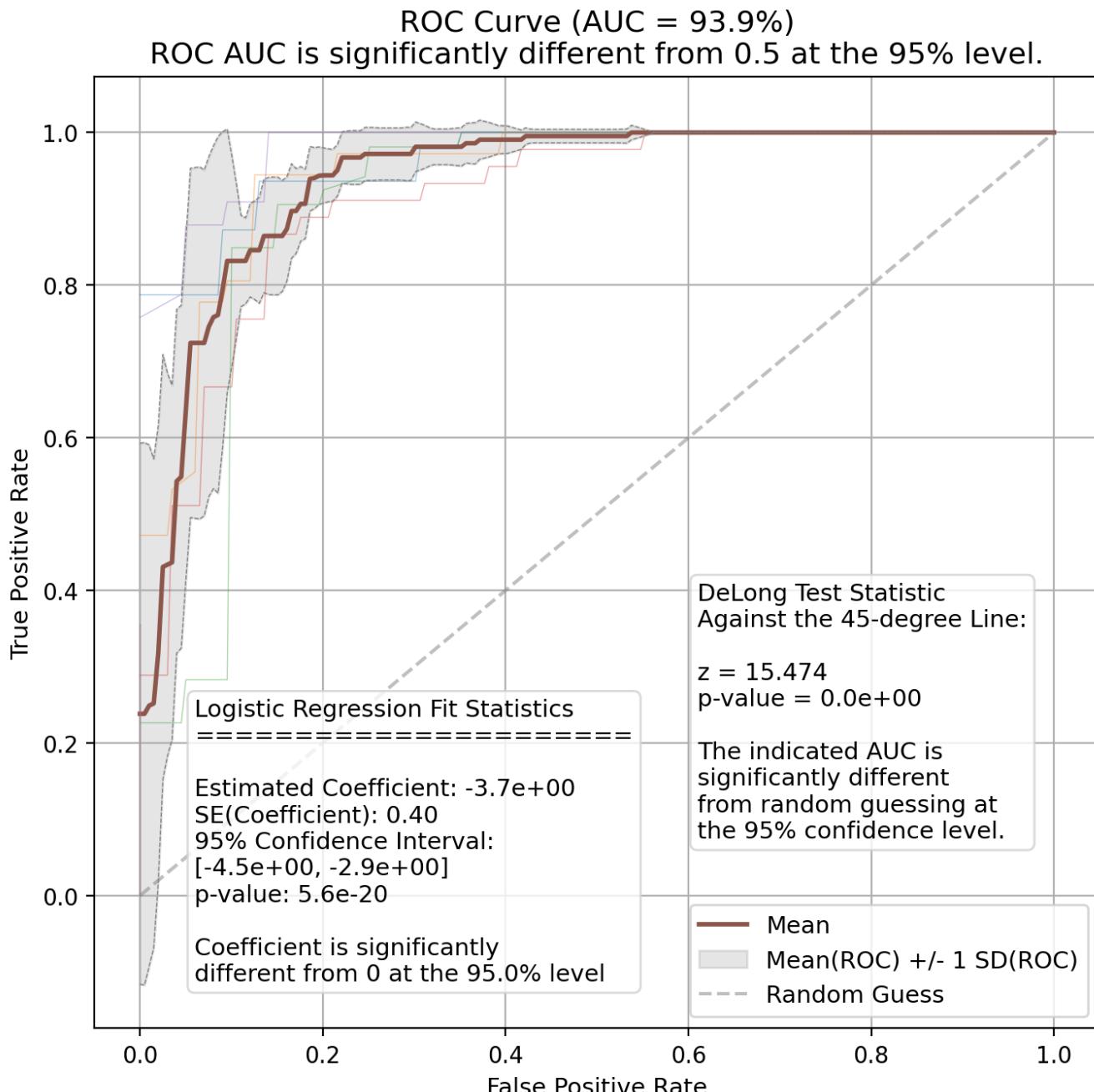
Mean Radius - Empirical CDF Plot



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

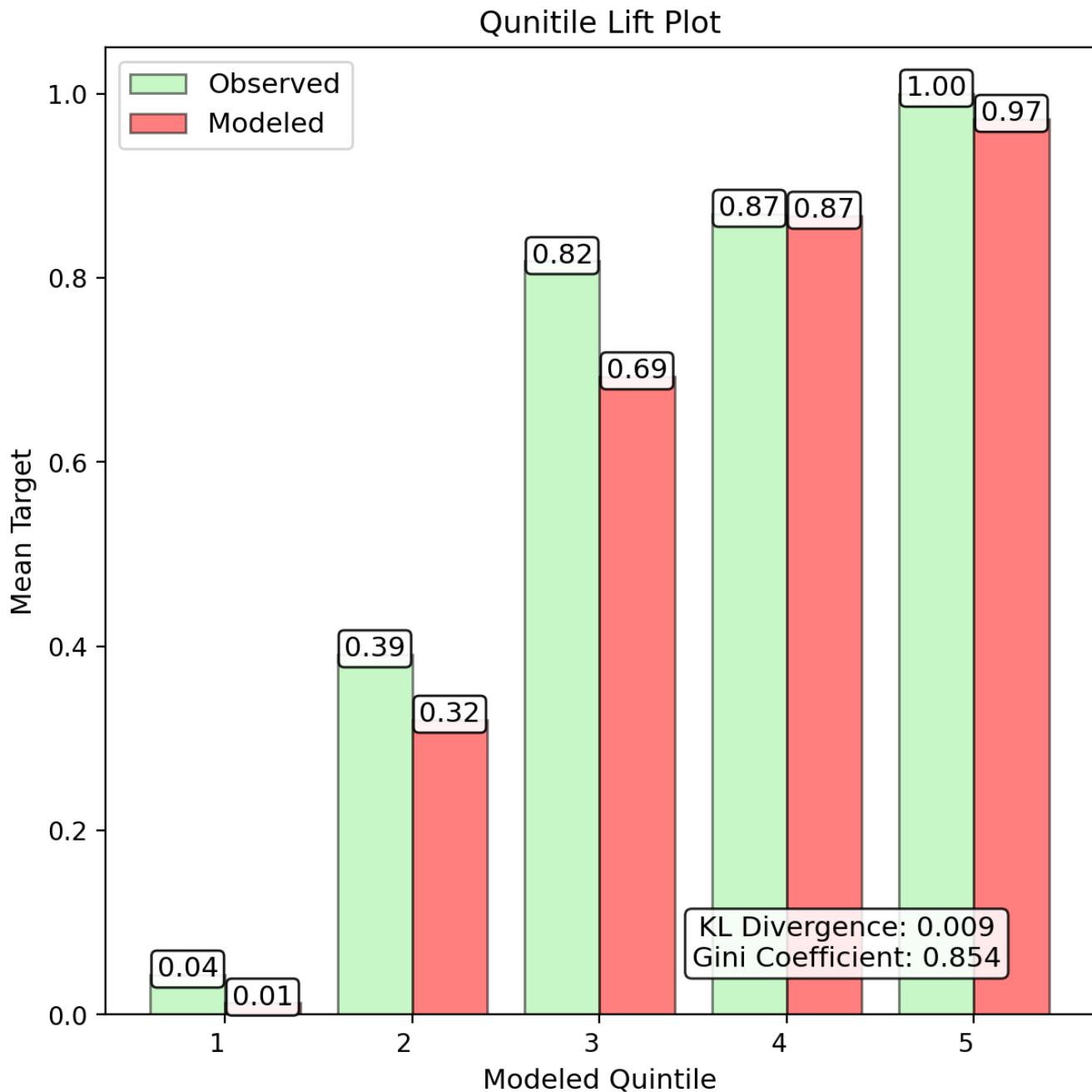


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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

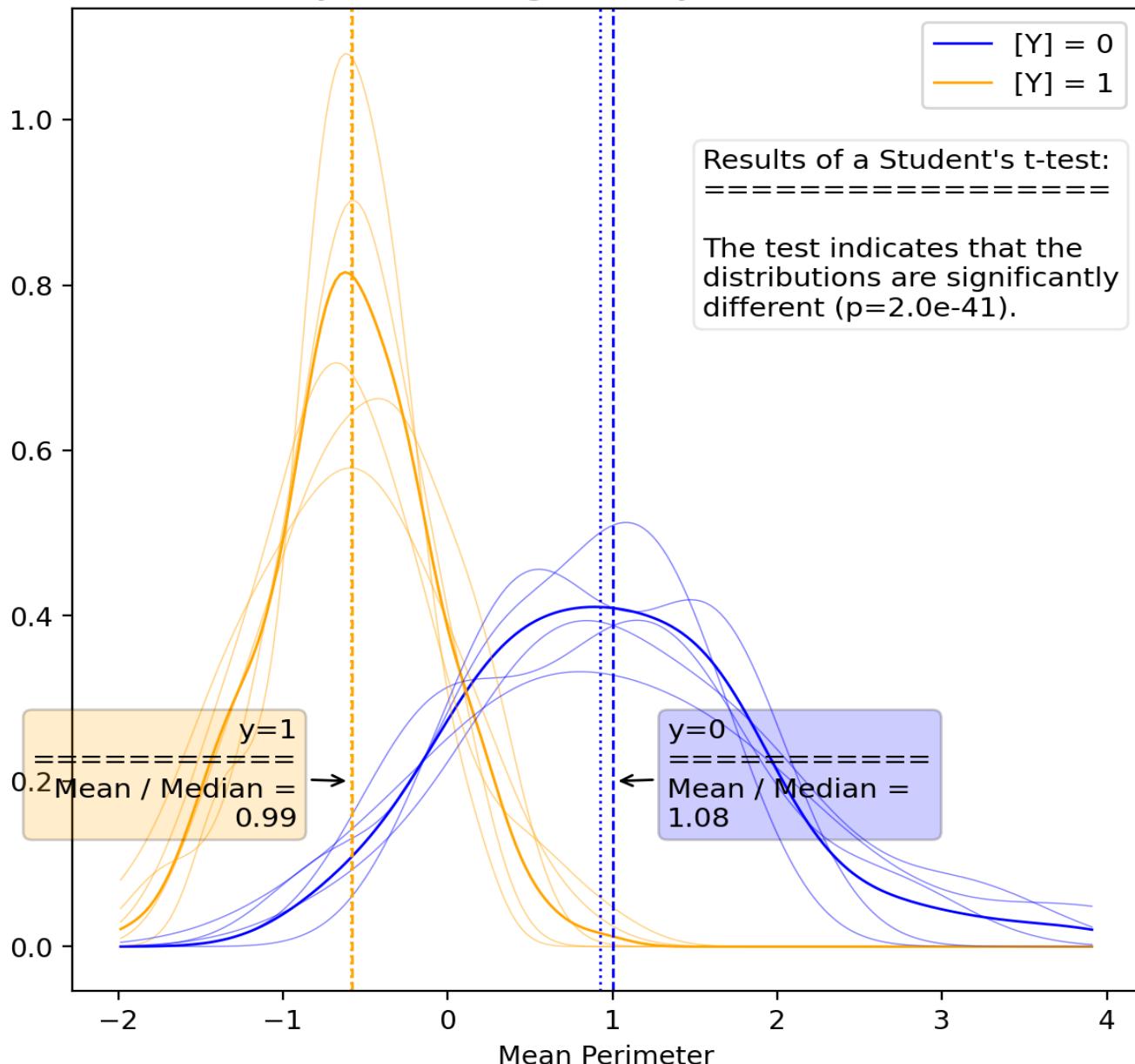
Mean Perimeter - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-3.66	-4.14	-4.13	-4.55	-3.69	-4.00	0.37
Fitted p-Value	2.0e-16	2.7e-16	8.6e-16	4.5e-15	2.3e-17	6.6e-20	1.9e-15
Fitted Std. Err.	0.445	0.505	0.513	0.581	0.435	0.438	0.059
Conf. Int. Lower	-4.53	-5.13	-5.14	-5.69	-4.54	-4.86	0.49
Conf. Int. Upper	-2.79	-3.15	-3.12	-3.42	-2.83	-3.14	0.26
Train Accuracy	87.5%	88.2%	88.4%	87.3%	86.7%	87.4%	0.7%
Val Accuracy	88.6%	88.4%	83.6%	87.8%	90.9%	86.8%	2.7%
Train AUC	87.8%	88.5%	87.9%	87.4%	86.7%	87.4%	0.7%
Val AUC	84.8%	88.6%	85.6%	87.5%	90.9%	86.2%	2.4%
Train F1	89.4%	90.7%	90.4%	89.6%	89.2%	89.7%	0.6%
Test F1	91.8%	88.2%	87.8%	89.9%	92.3%	89.4%	2.1%
Train Precision	92.9%	94.0%	90.1%	92.5%	91.8%	92.1%	1.4%
Val Precision	88.2%	93.8%	95.6%	90.9%	93.8%	90.0%	2.9%
Train Recall	86.2%	87.6%	90.7%	87.0%	86.7%	87.4%	1.8%
Val Recall	95.7%	83.3%	81.1%	88.9%	90.9%	88.7%	5.9%
Train MCC	74.4%	75.1%	75.9%	73.4%	72.2%	73.7%	1.4%
Val MCC	73.6%	77.4%	65.2%	74.7%	81.3%	72.1%	5.9%
Train Log-Loss	4.52	4.24	4.17	4.59	4.79	4.55	0.26
Val Log-Loss	4.12	4.18	5.92	4.38	3.28	4.74	0.96

Univariate Report

Mean Perimeter - Kernel Density Plot

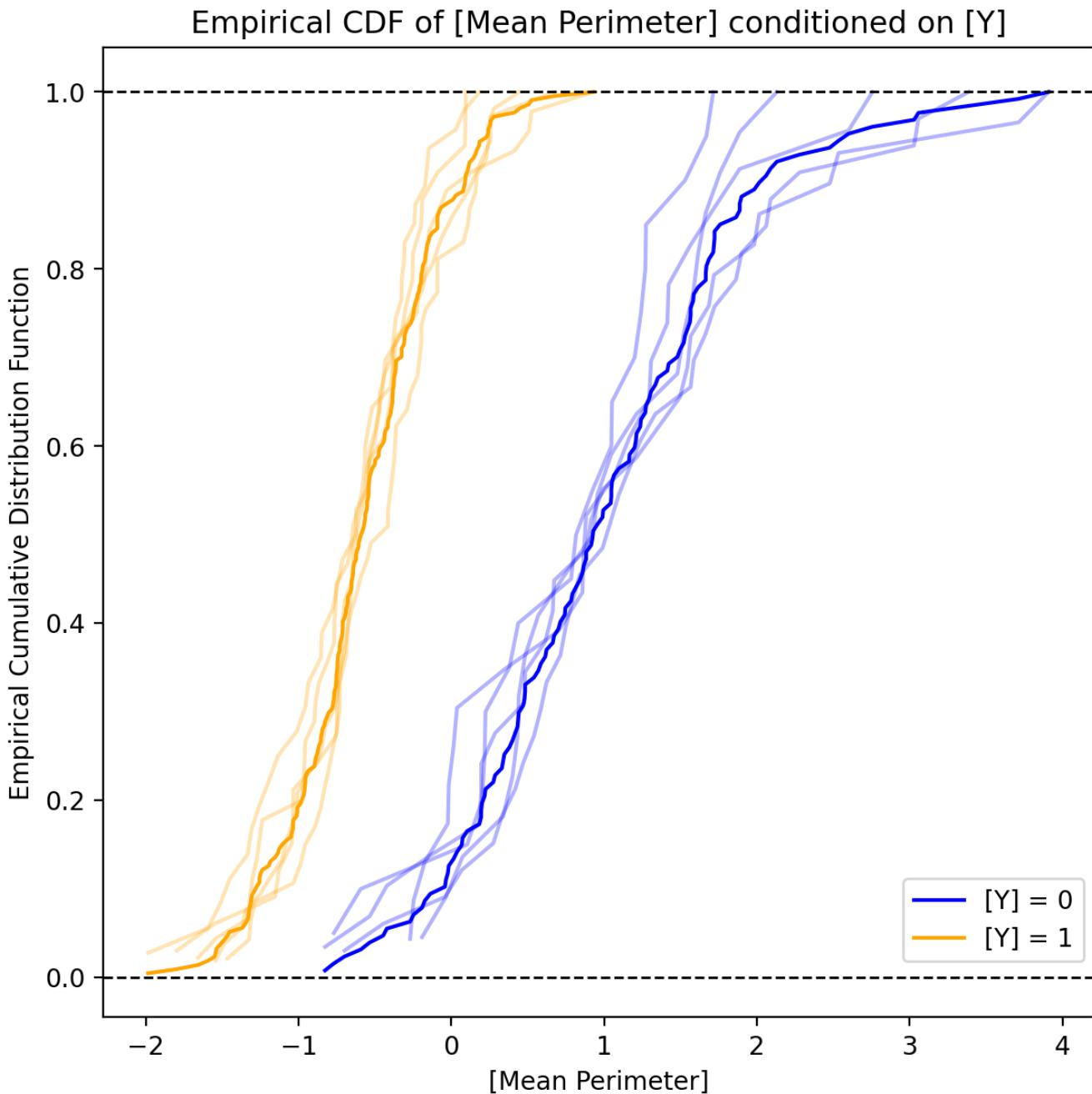
Kernel Density Plot of [Mean Perimeter] by [Y].
Distributions by level are significantly different at the 95% level.



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

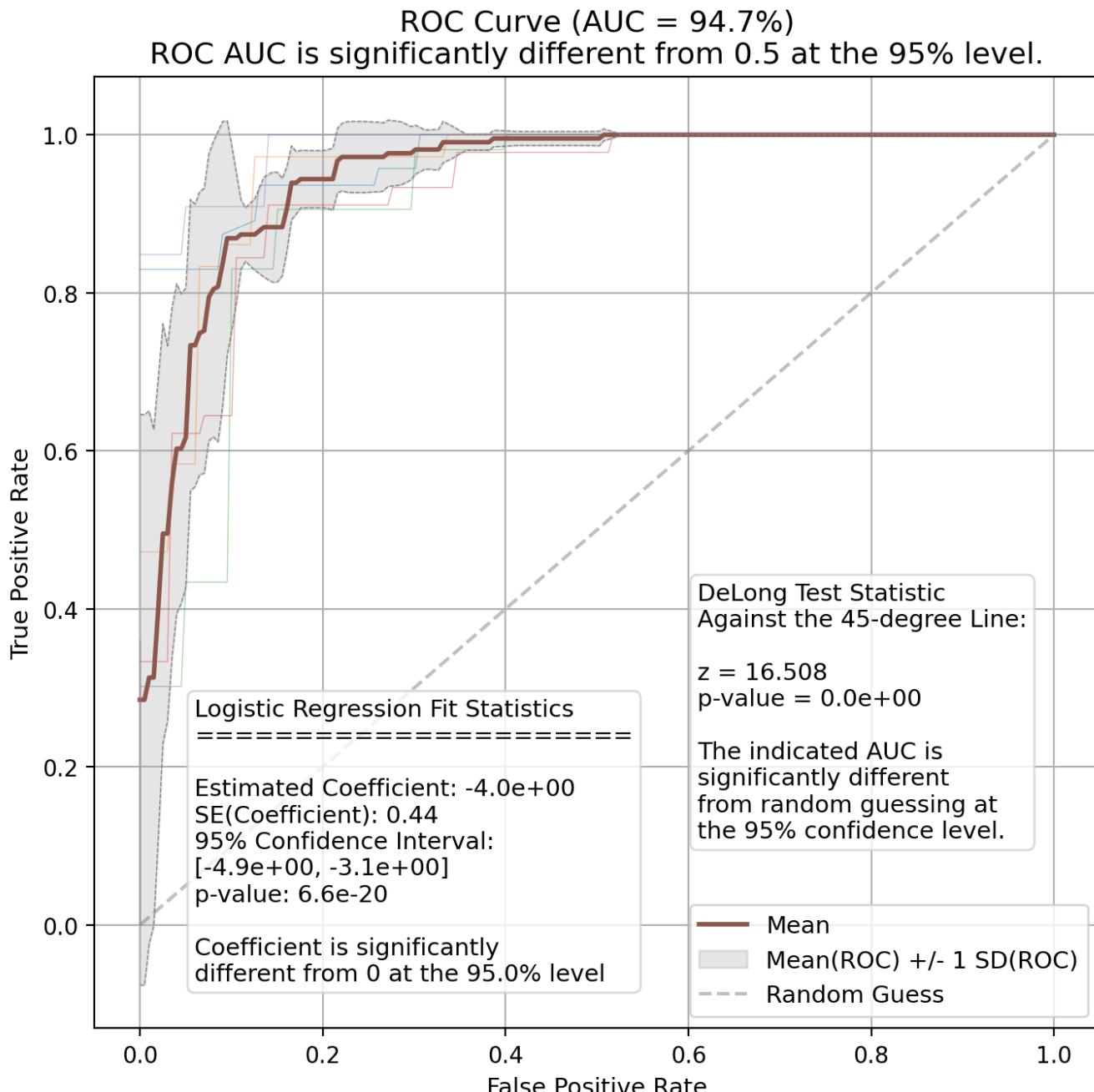
Mean Perimeter - Empirical CDF Plot



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

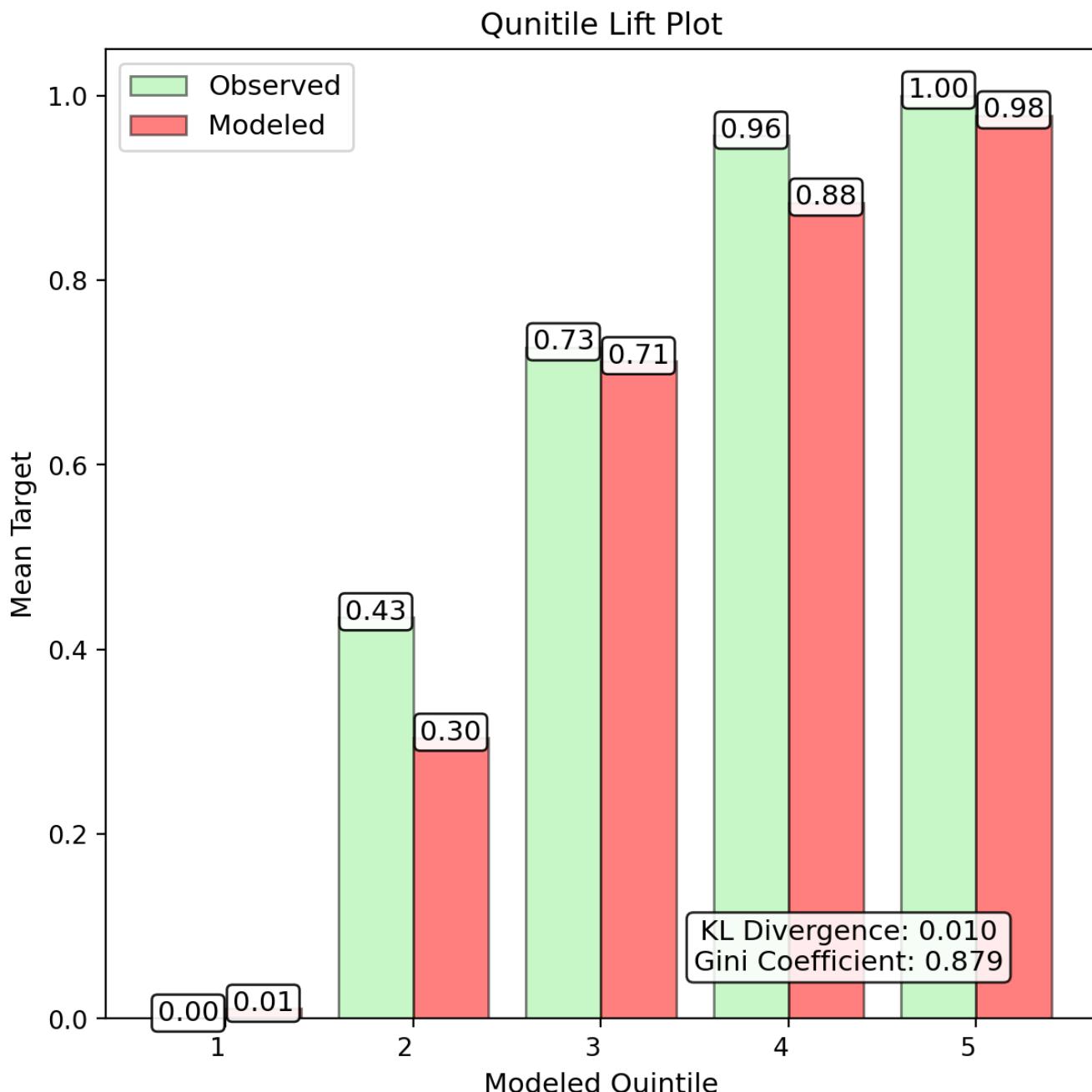


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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

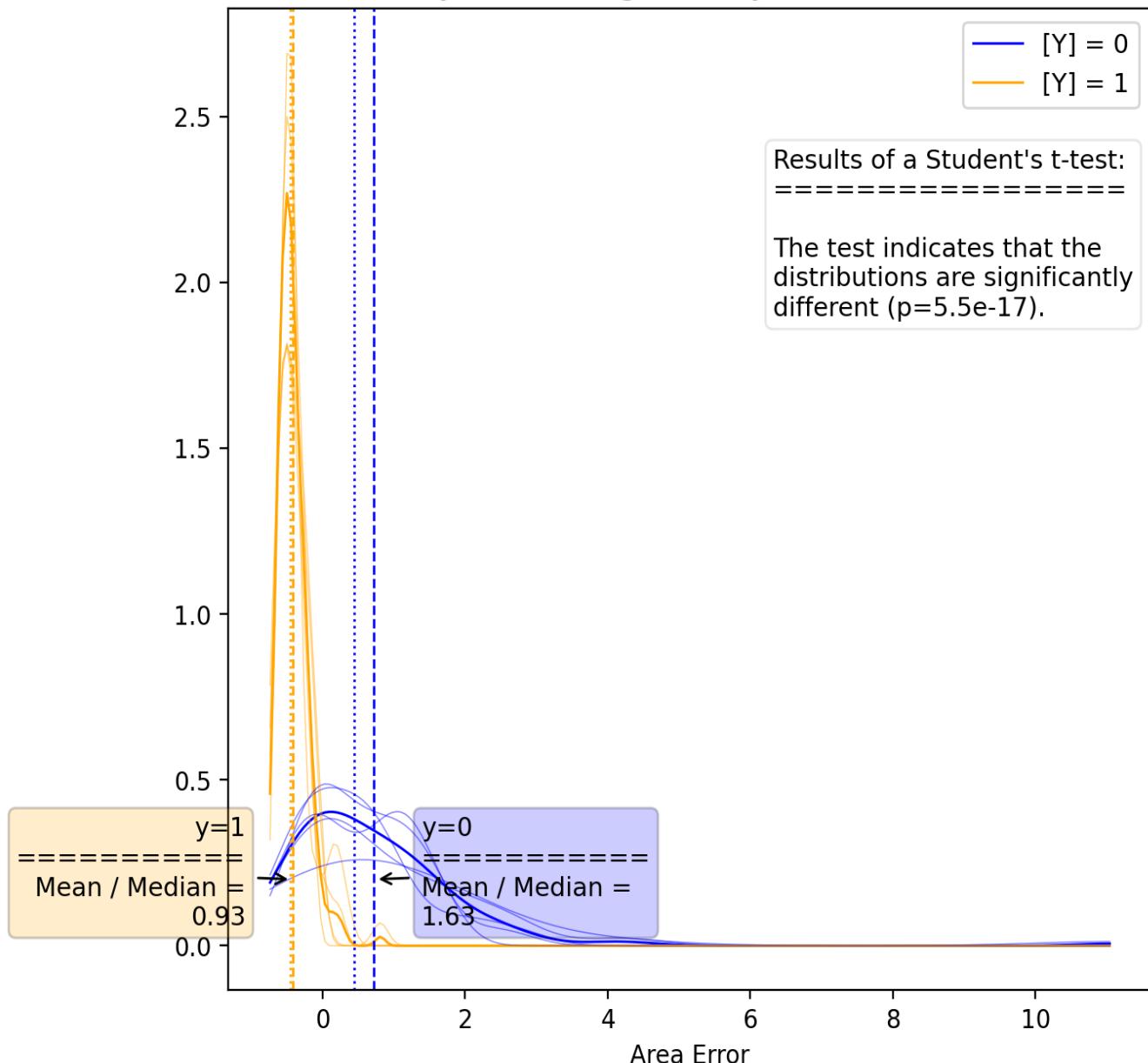
Area Error - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-4.67	-5.07	-4.76	-4.82	-4.69	-4.80	0.16
Fitted p-Value	8.3e-19	9.9e-20	3.5e-19	5.1e-19	1.7e-20	6.3e-24	3.3e-19
Fitted Std. Err.	0.527	0.557	0.532	0.541	0.506	0.476	0.019
Conf. Int. Lower	-5.70	-6.16	-5.81	-5.89	-5.68	-5.73	0.19
Conf. Int. Upper	-3.63	-3.97	-3.72	-3.76	-3.70	-3.87	0.13
Train Accuracy	85.6%	88.2%	85.4%	87.3%	86.4%	86.8%	1.2%
Val Accuracy	87.1%	82.6%	87.7%	87.8%	87.3%	86.8%	2.2%
Train AUC	82.5%	84.0%	82.7%	83.9%	82.8%	83.4%	0.7%
Val AUC	80.4%	82.3%	79.1%	85.7%	84.1%	84.4%	2.7%
Train F1	89.1%	91.6%	88.8%	90.6%	89.9%	90.2%	1.1%
Test F1	91.3%	84.2%	92.0%	90.5%	90.4%	89.9%	3.1%
Train Precision	83.3%	86.1%	82.4%	85.3%	84.5%	84.5%	1.5%
Val Precision	83.9%	80.0%	86.7%	86.0%	82.5%	85.9%	2.7%
Train Recall	95.8%	97.8%	96.3%	96.4%	96.1%	96.7%	0.7%
Val Recall	100.0%	88.9%	98.1%	95.6%	100.0%	94.4%	4.6%
Train MCC	69.6%	73.9%	70.0%	72.5%	70.5%	71.8%	1.8%
Val MCC	71.5%	65.4%	67.7%	74.5%	75.0%	71.7%	4.2%
Train Log-Loss	5.19	4.24	5.25	4.59	4.92	4.76	0.42
Val Log-Loss	4.63	6.27	4.44	4.38	4.59	4.74	0.79

Univariate Report

Area Error - Kernel Density Plot

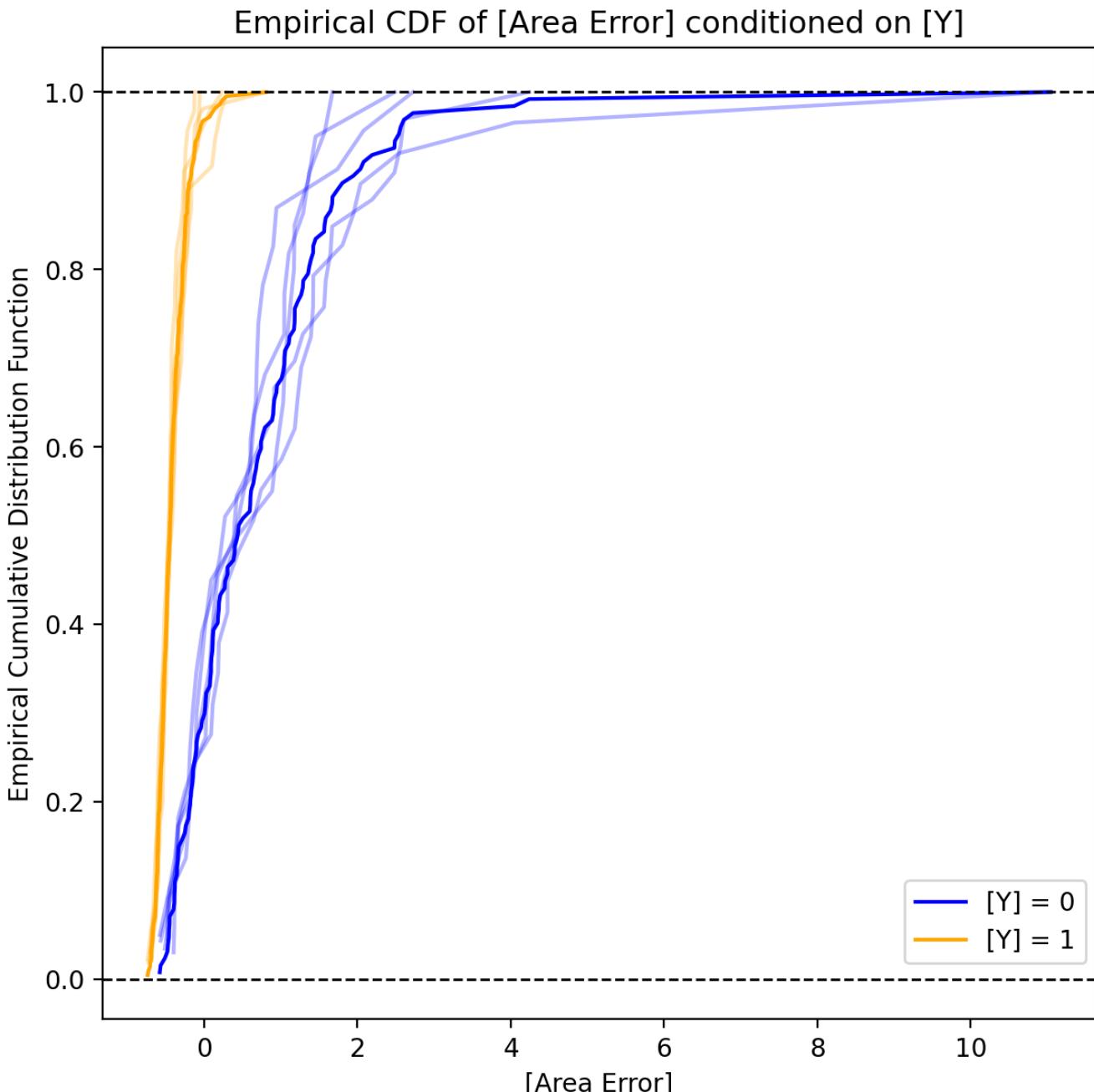
Kernel Density Plot of [Area Error] by [Y].
Distributions by level are significantly different at the 95% level.



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

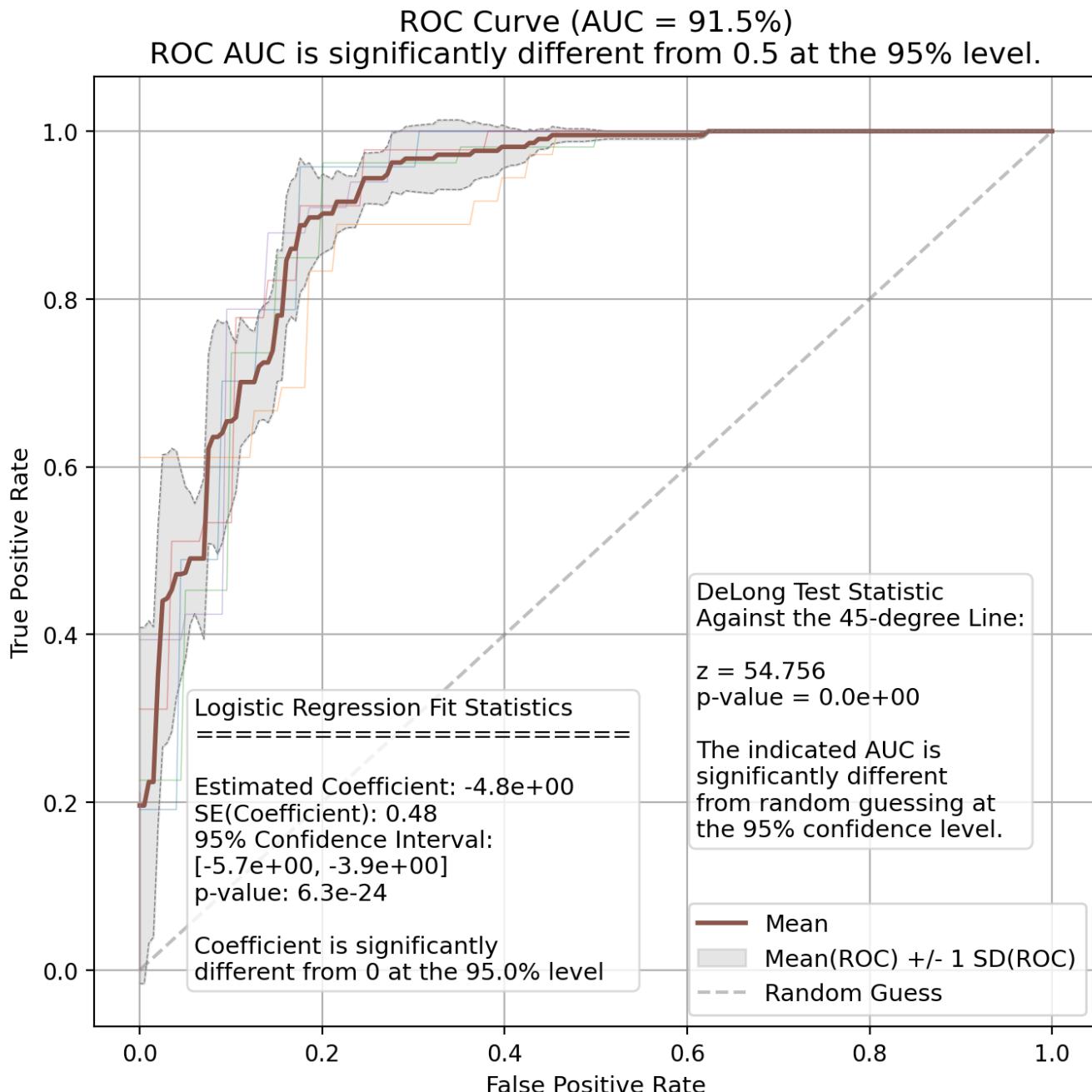
Area Error - Empirical CDF Plot



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

Area Error - ROC Curve

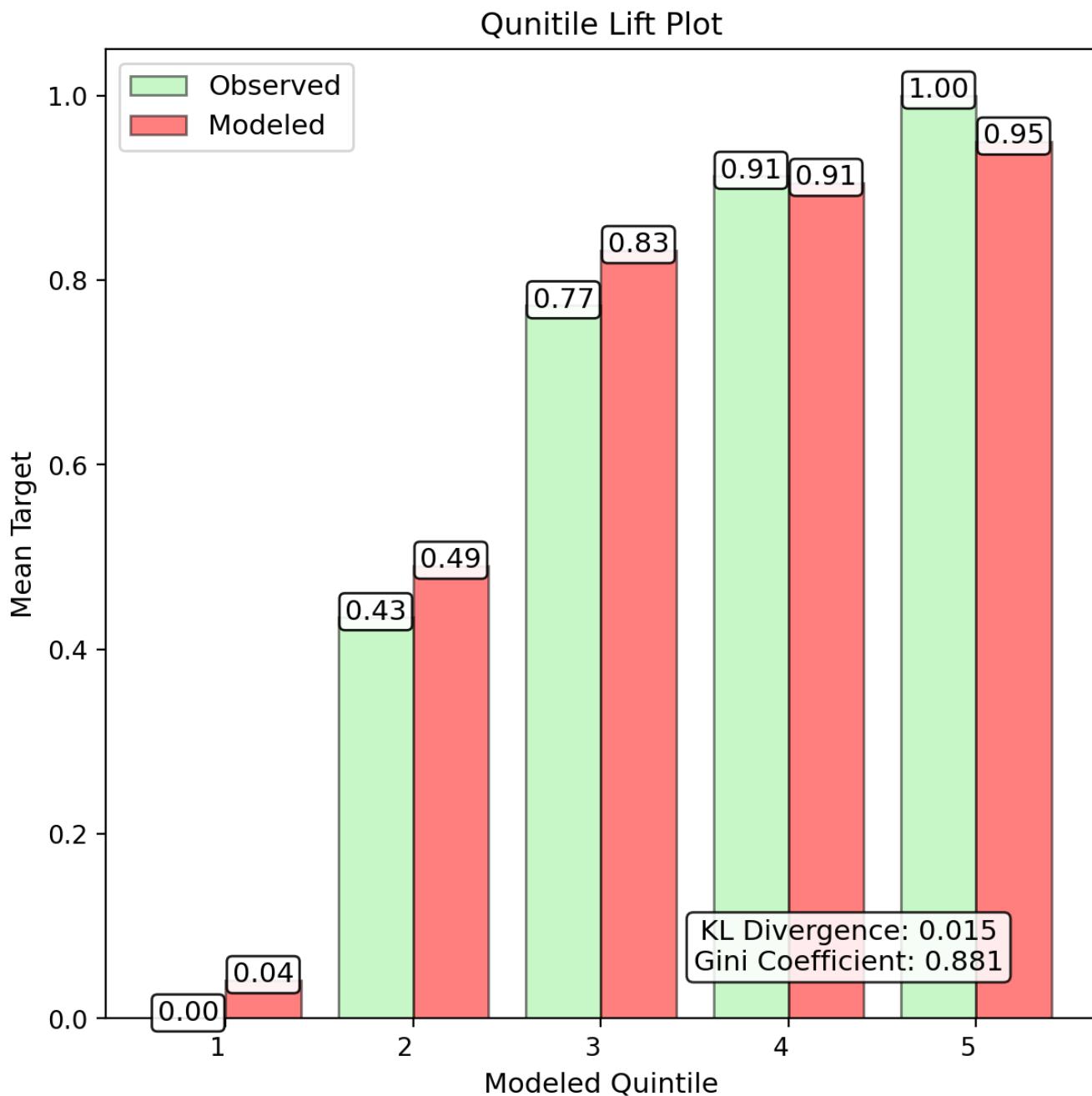


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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

Area 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

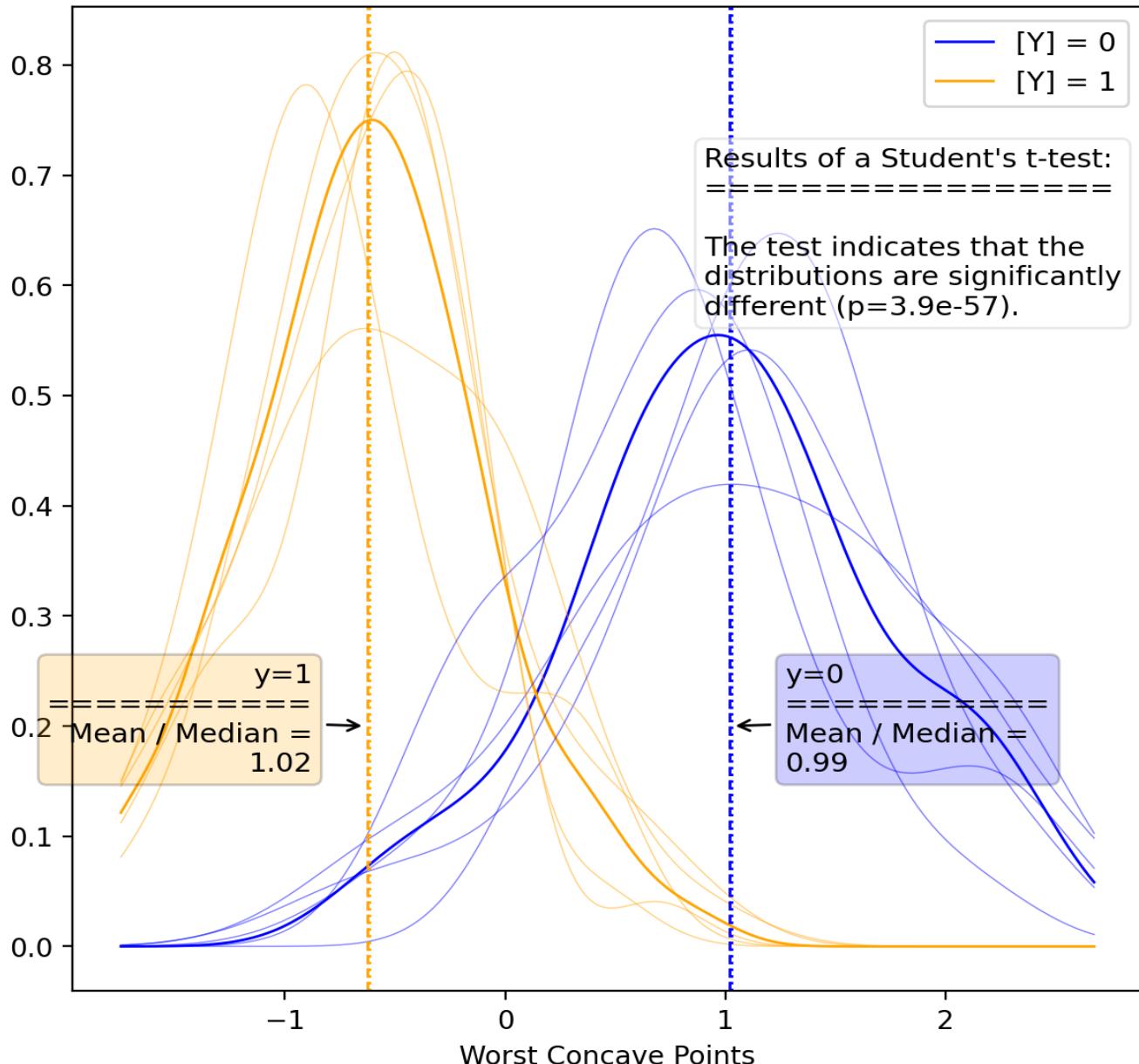
Worst Concave Points - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-3.60	-4.20	-4.06	-4.10	-3.72	-3.92	0.26
Fitted p-Value	1.8e-17	8.1e-17	1.0e-15	1.8e-16	4.8e-18	8.5e-21	4.3e-16
Fitted Std. Err.	0.423	0.504	0.506	0.498	0.429	0.419	0.042
Conf. Int. Lower	-4.43	-5.18	-5.05	-5.08	-4.56	-4.74	0.34
Conf. Int. Upper	-2.77	-3.21	-3.07	-3.13	-2.87	-3.09	0.18
Train Accuracy	87.8%	90.4%	91.0%	89.9%	89.9%	89.4%	1.2%
Val Accuracy	95.7%	84.1%	86.3%	89.2%	89.1%	84.2%	4.4%
Train AUC	87.9%	91.2%	91.0%	90.1%	90.6%	89.7%	1.3%
Val AUC	96.8%	84.3%	87.5%	89.9%	87.1%	86.4%	4.7%
Train F1	89.8%	92.4%	92.5%	91.8%	91.6%	91.3%	1.1%
Test F1	96.7%	83.6%	90.0%	90.7%	91.4%	85.9%	4.7%
Train Precision	92.4%	96.3%	93.6%	94.4%	95.8%	94.1%	1.6%
Val Precision	100.0%	90.3%	95.7%	95.1%	86.5%	96.5%	5.2%
Train Recall	87.4%	88.8%	91.3%	89.3%	87.8%	88.8%	1.5%
Val Recall	93.6%	77.8%	84.9%	86.7%	97.0%	77.5%	7.5%
Train MCC	74.8%	80.1%	81.5%	78.8%	79.3%	78.1%	2.5%
Val MCC	91.0%	69.0%	69.8%	78.3%	77.5%	70.6%	8.9%
Train Log-Loss	4.39	3.45	3.23	3.64	3.65	3.81	0.44
Val Log-Loss	1.54	5.75	4.94	3.90	3.93	5.69	1.58

Univariate Report

Worst Concave Points - Kernel Density Plot

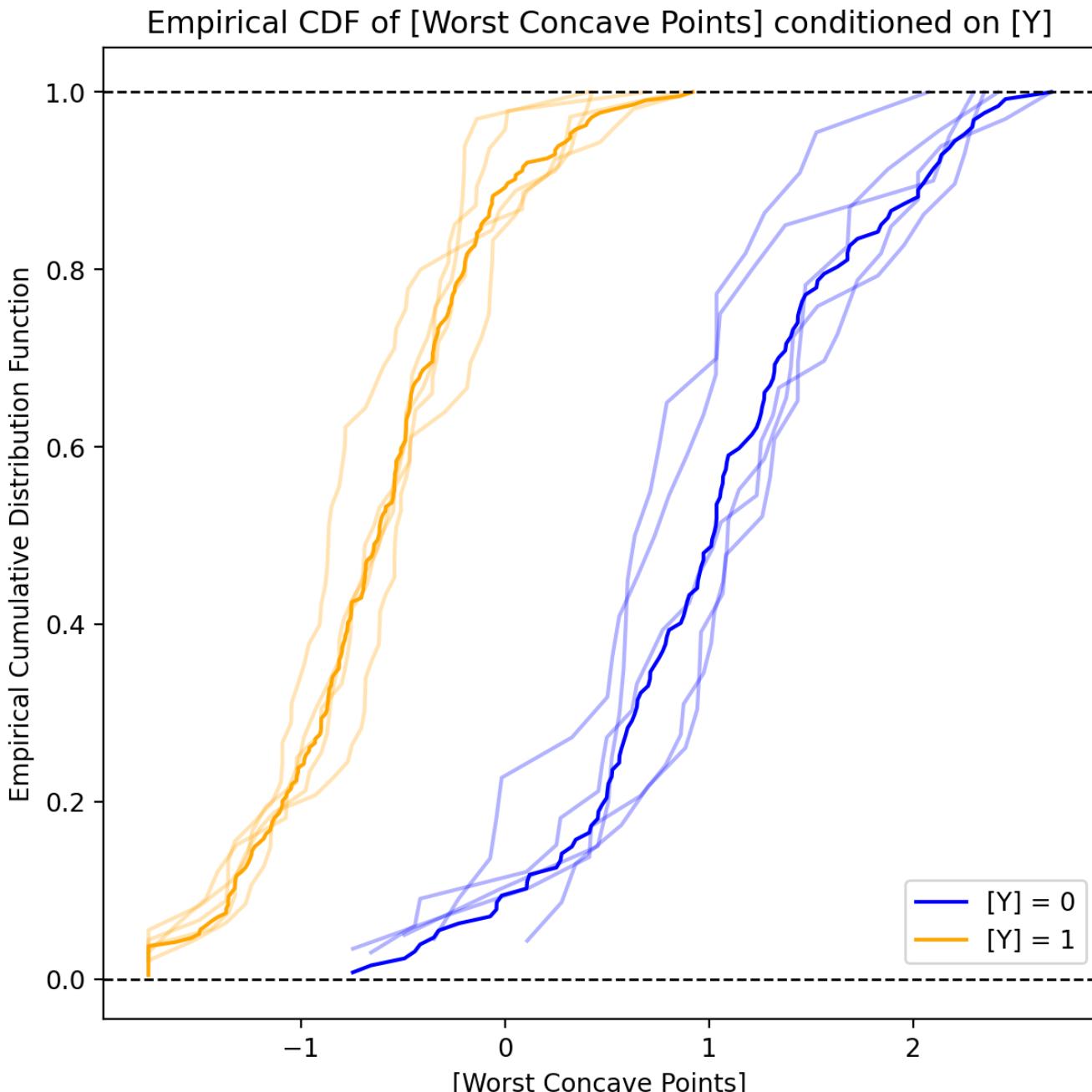
Kernel Density Plot of [Worst Concave Points] by [Y].
Distributions by level are significantly different at the 95% level.



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

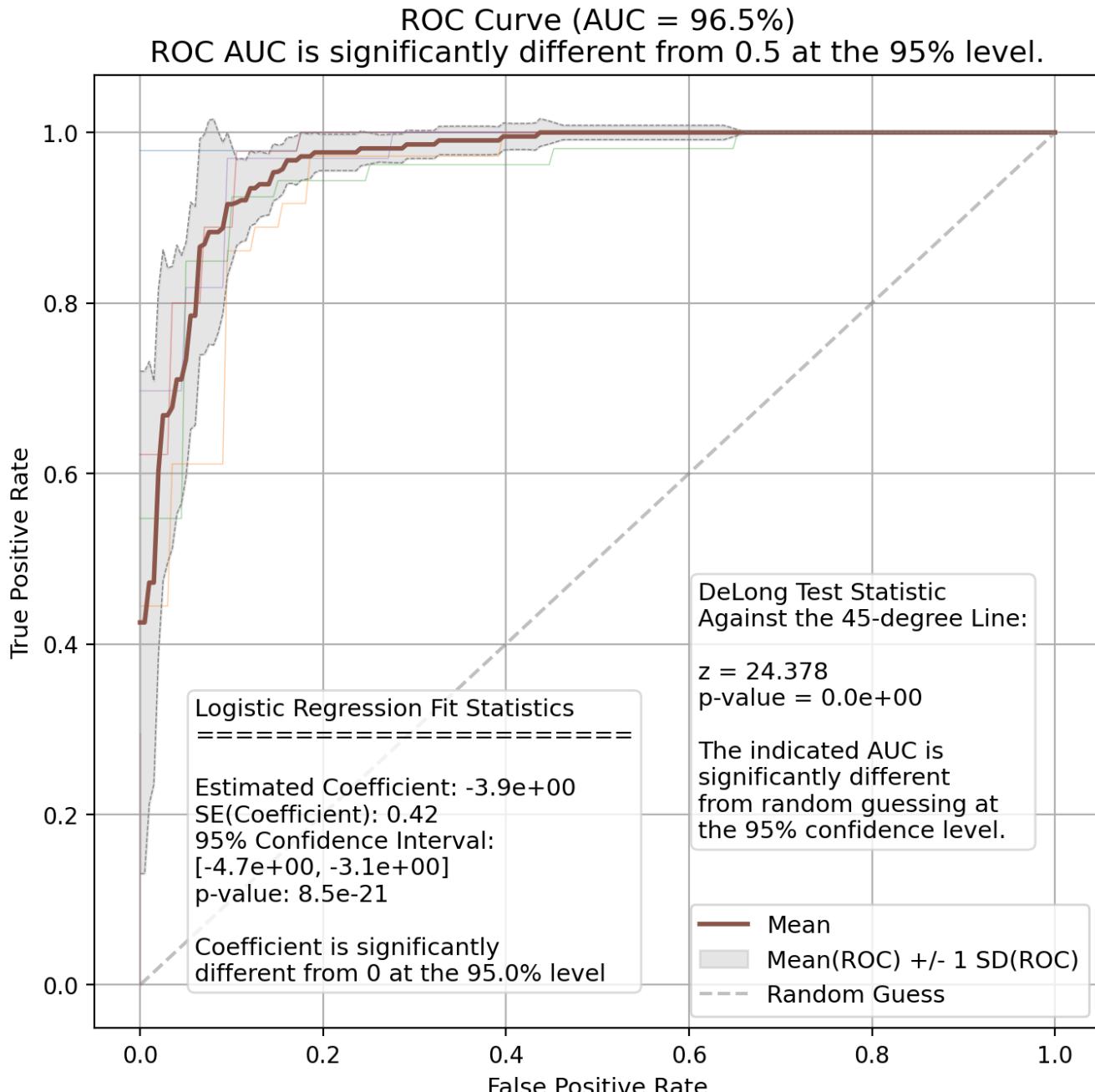
Worst Concave Points - Empirical CDF Plot



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 Concave Points - ROC Curve

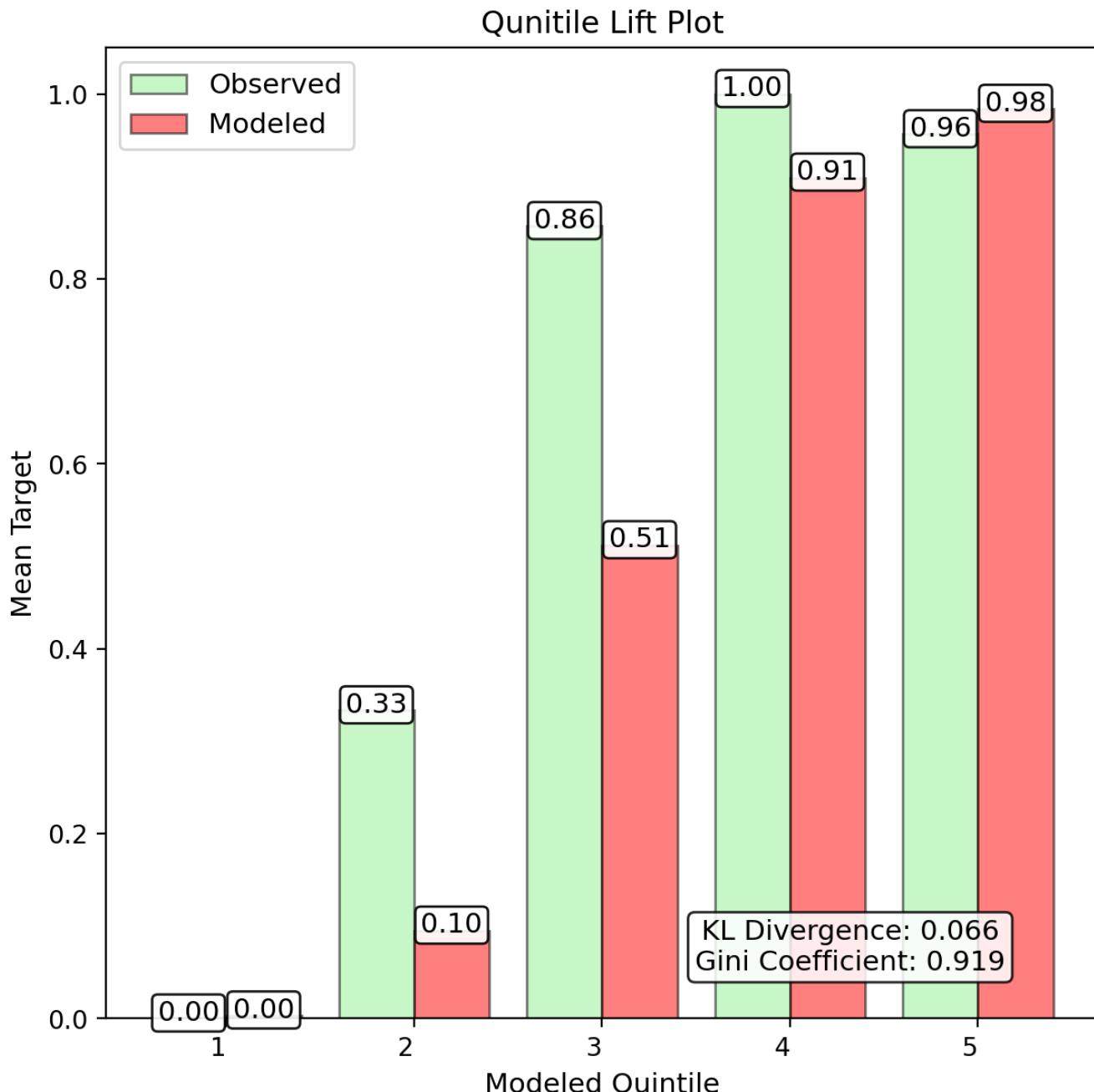


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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 Concave Points - 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

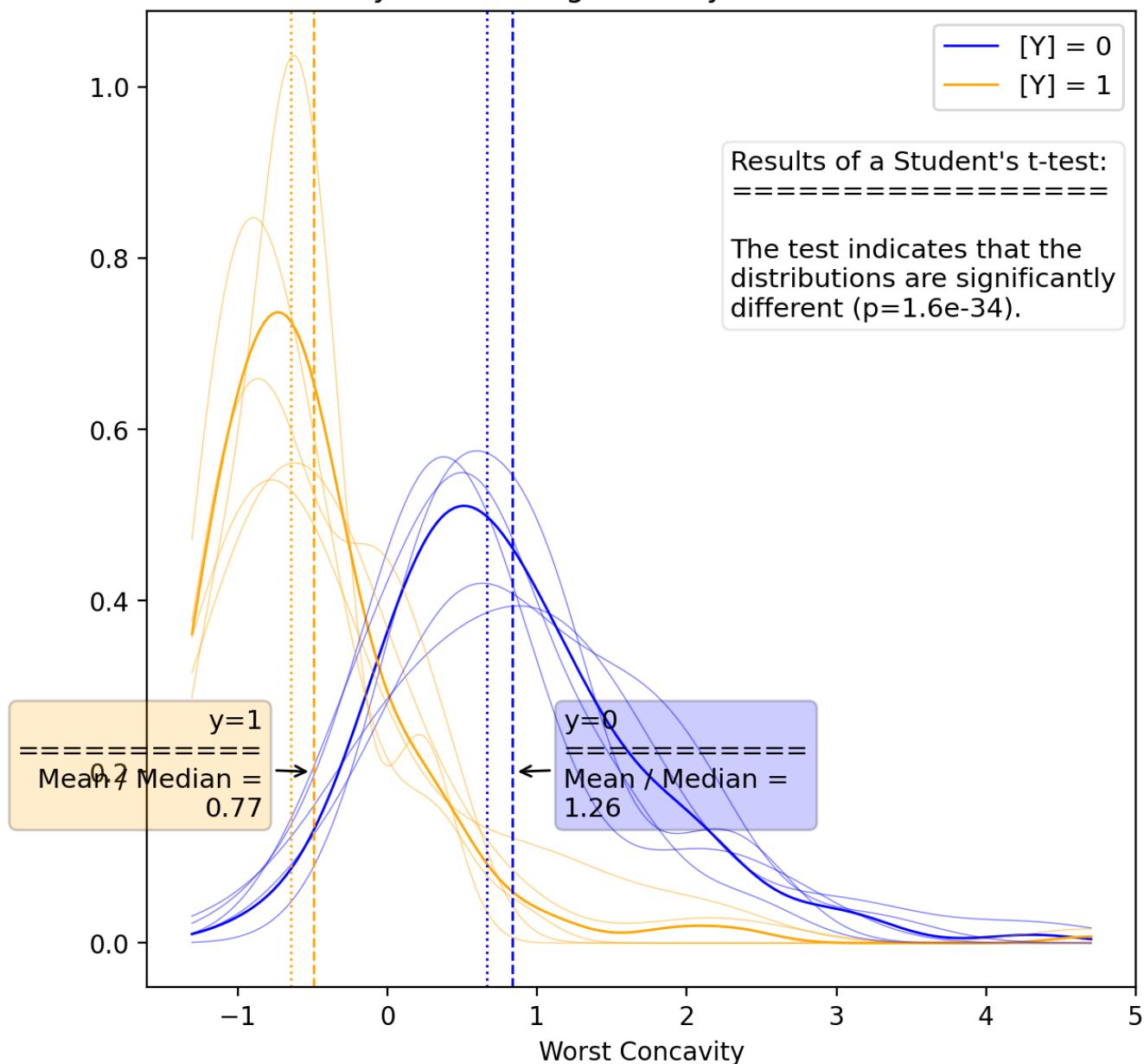
Worst Concavity - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-1.92	-2.62	-2.74	-2.00	-2.12	-2.24	0.37
Fitted p-Value	1.7e-16	3.1e-18	2.0e-18	4.4e-16	2.7e-18	1.1e-21	1.9e-16
Fitted Std. Err.	0.233	0.301	0.313	0.247	0.243	0.235	0.037
Conf. Int. Lower	-2.38	-3.21	-3.36	-2.49	-2.59	-2.70	0.45
Conf. Int. Upper	-1.47	-2.03	-2.13	-1.52	-1.64	-1.78	0.30
Train Accuracy	84.1%	84.9%	85.1%	83.5%	85.3%	84.8%	0.7%
Val Accuracy	88.6%	81.2%	84.9%	89.2%	83.6%	84.2%	3.4%
Train AUC	84.4%	86.0%	85.2%	84.0%	85.8%	85.1%	0.9%
Val AUC	89.3%	81.4%	86.5%	89.3%	83.3%	84.1%	3.5%
Train F1	86.6%	87.8%	87.2%	86.3%	87.9%	87.3%	0.7%
Test F1	91.1%	80.6%	88.9%	90.9%	86.2%	87.0%	4.4%
Train Precision	90.3%	93.6%	90.1%	90.8%	92.1%	91.3%	1.5%
Val Precision	95.3%	87.1%	95.7%	93.0%	87.5%	89.6%	4.2%
Train Recall	83.2%	82.6%	84.5%	82.2%	84.0%	83.6%	0.9%
Val Recall	87.2%	75.0%	83.0%	88.9%	84.8%	84.5%	5.4%
Train MCC	67.6%	69.3%	69.6%	66.2%	69.9%	68.7%	1.6%
Val MCC	75.8%	63.1%	67.5%	77.7%	66.2%	67.2%	6.3%
Train Log-Loss	5.72	5.43	5.38	5.94	5.29	5.50	0.27
Val Log-Loss	4.12	6.79	5.43	3.90	5.90	5.69	1.22

Univariate Report

Worst Concavity - Kernel Density Plot

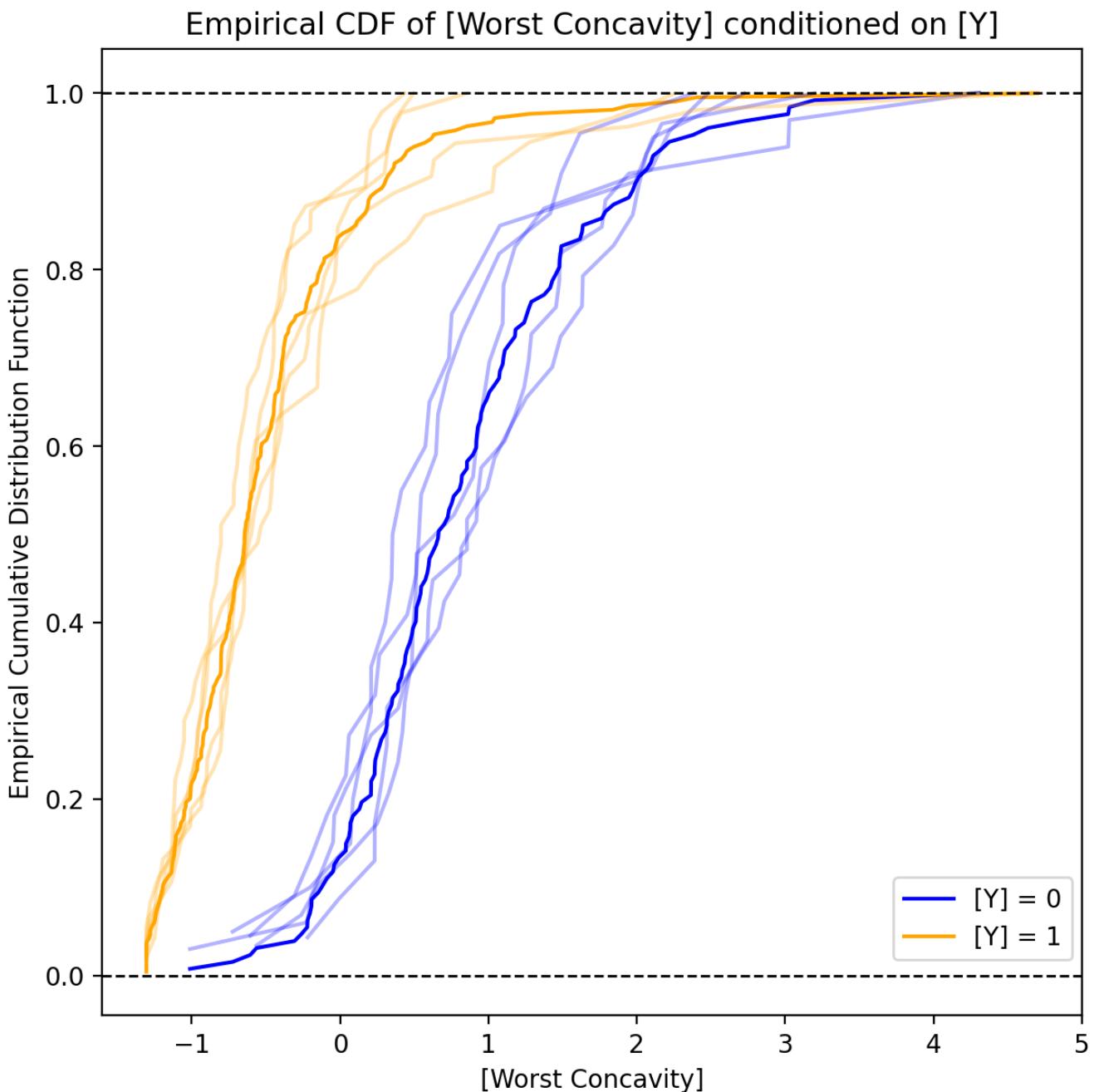
Kernel Density Plot of [Worst Concavity] by [Y].
Distributions by level are significantly different at the 95% level.



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

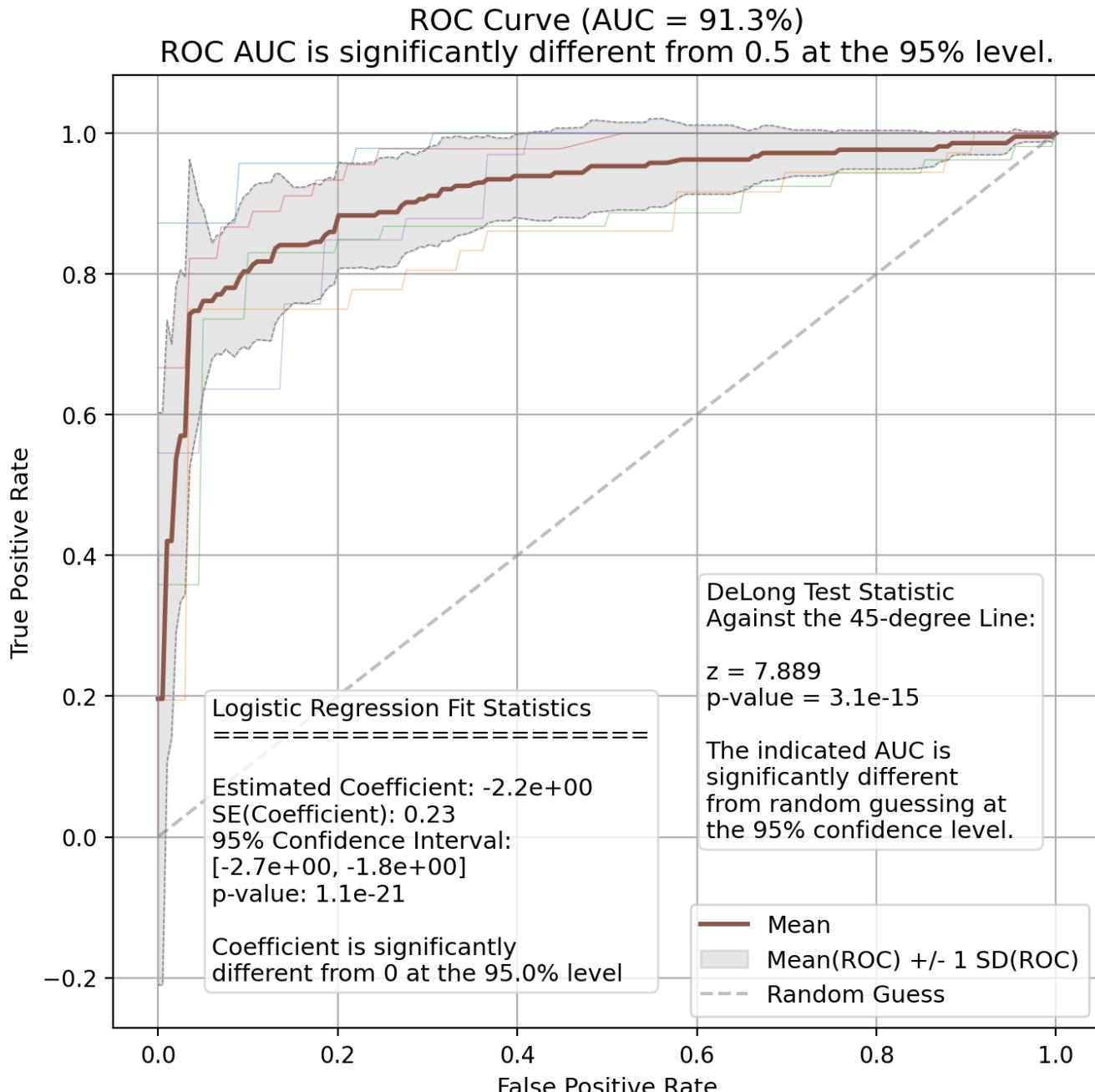
Worst Concavity - Empirical CDF Plot



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 Concavity - ROC Curve

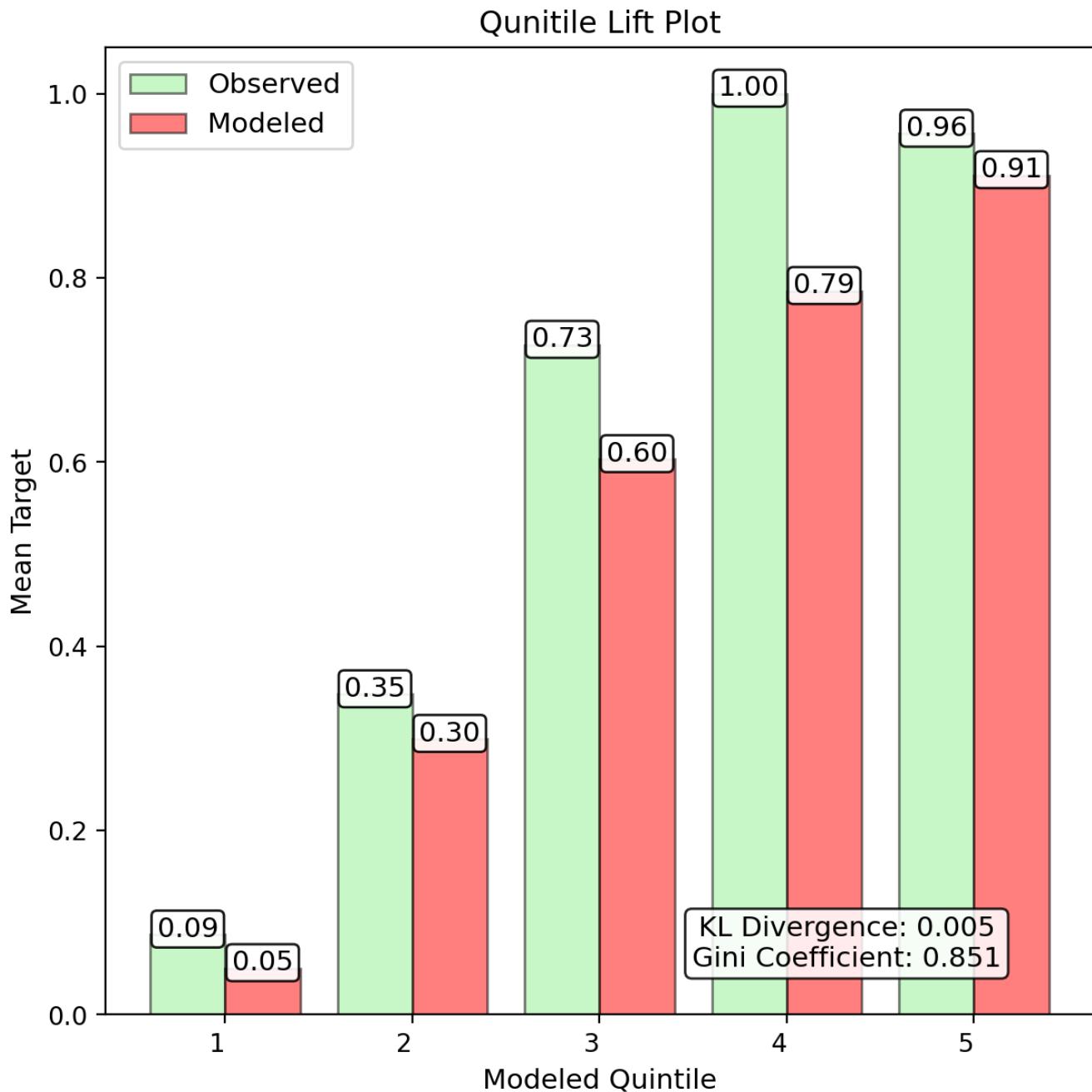


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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 Concavity - 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

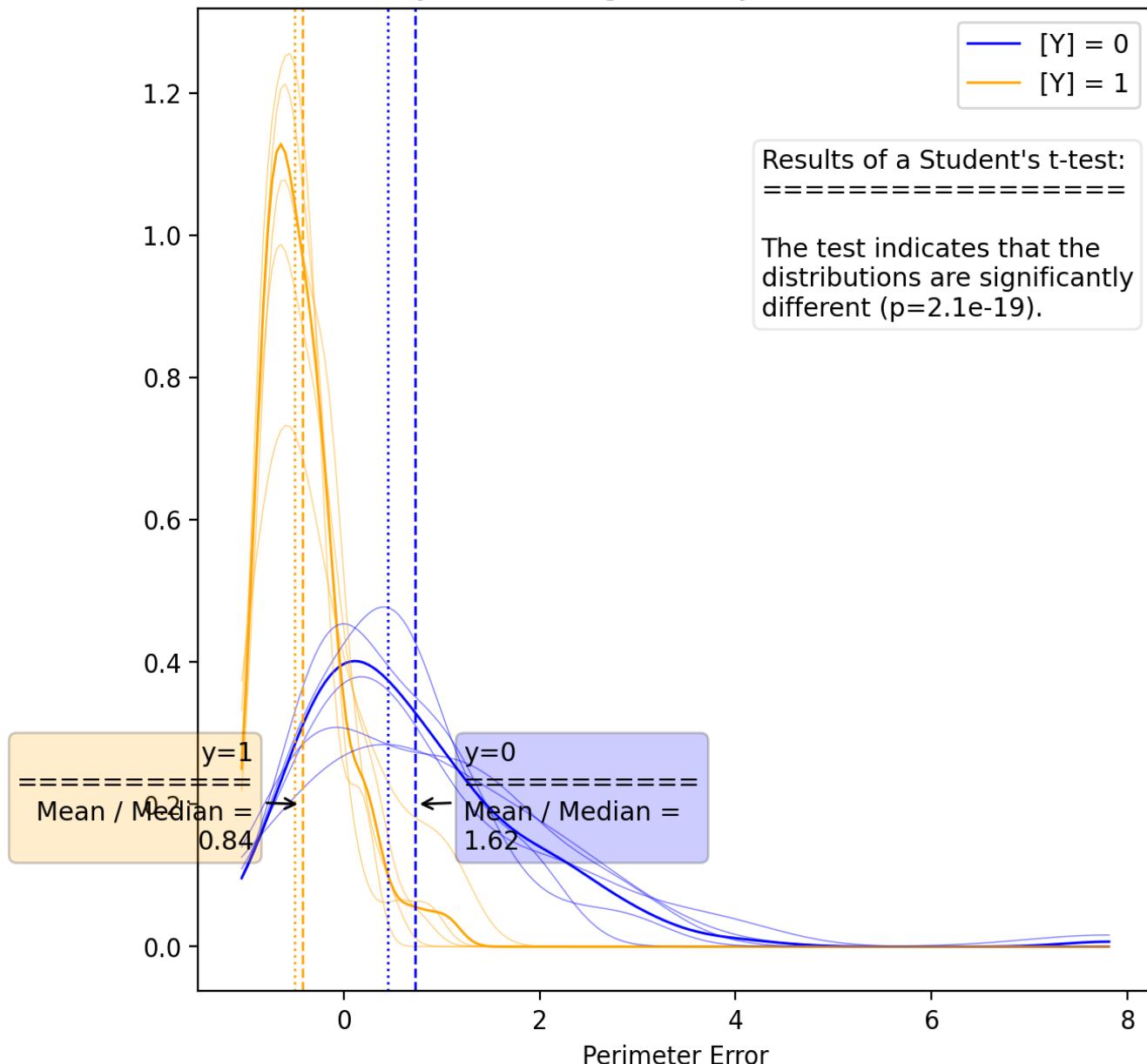
Perimeter Error - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-2.64	-3.38	-2.86	-2.78	-2.78	-2.87	0.29
Fitted p-Value	6.6e-16	3.9e-18	1.2e-16	2.1e-16	9.7e-18	7.4e-21	2.7e-16
Fitted Std. Err.	0.326	0.390	0.345	0.339	0.324	0.307	0.027
Conf. Int. Lower	-3.28	-4.15	-3.53	-3.45	-3.42	-3.48	0.34
Conf. Int. Upper	-2.00	-2.62	-2.18	-2.12	-2.15	-2.27	0.24
Train Accuracy	79.3%	84.2%	81.3%	79.8%	81.8%	80.9%	1.9%
Val Accuracy	84.3%	73.9%	79.5%	87.8%	78.2%	82.5%	5.4%
Train AUC	77.2%	81.6%	79.6%	77.6%	79.8%	78.9%	1.8%
Val AUC	79.4%	74.1%	73.4%	86.9%	75.0%	80.0%	5.6%
Train F1	83.7%	88.2%	85.0%	84.3%	85.9%	85.1%	1.7%
Test F1	88.9%	73.5%	86.0%	90.1%	83.3%	86.5%	6.6%
Train Precision	81.4%	86.5%	82.1%	82.9%	84.5%	83.4%	2.1%
Val Precision	84.6%	78.1%	85.2%	89.1%	76.9%	83.1%	5.1%
Train Recall	86.2%	89.9%	88.2%	85.8%	87.3%	86.9%	1.6%
Val Recall	93.6%	69.4%	86.8%	91.1%	90.9%	90.1%	9.8%
Train MCC	55.7%	64.5%	60.6%	56.0%	60.5%	58.7%	3.7%
Val MCC	63.2%	48.3%	47.6%	74.4%	53.9%	62.0%	11.3%
Train Log-Loss	7.45	5.70	6.72	7.29	6.55	6.87	0.69
Val Log-Loss	5.66	9.40	7.41	4.38	7.86	6.32	1.96

Univariate Report

Perimeter Error - Kernel Density Plot

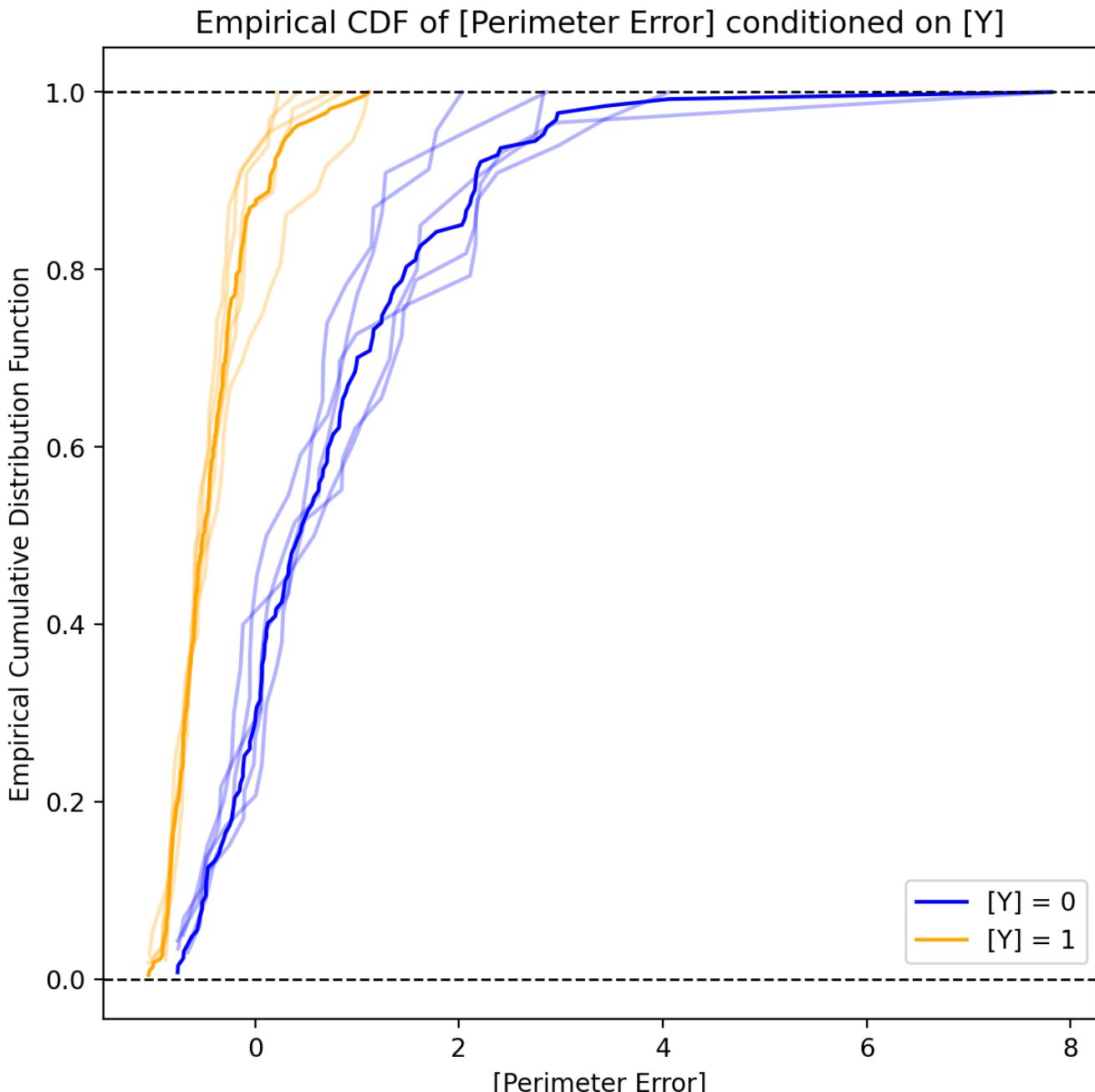
Kernel Density Plot of [Perimeter Error] by [Y].
Distributions by level are significantly different at the 95% level.



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

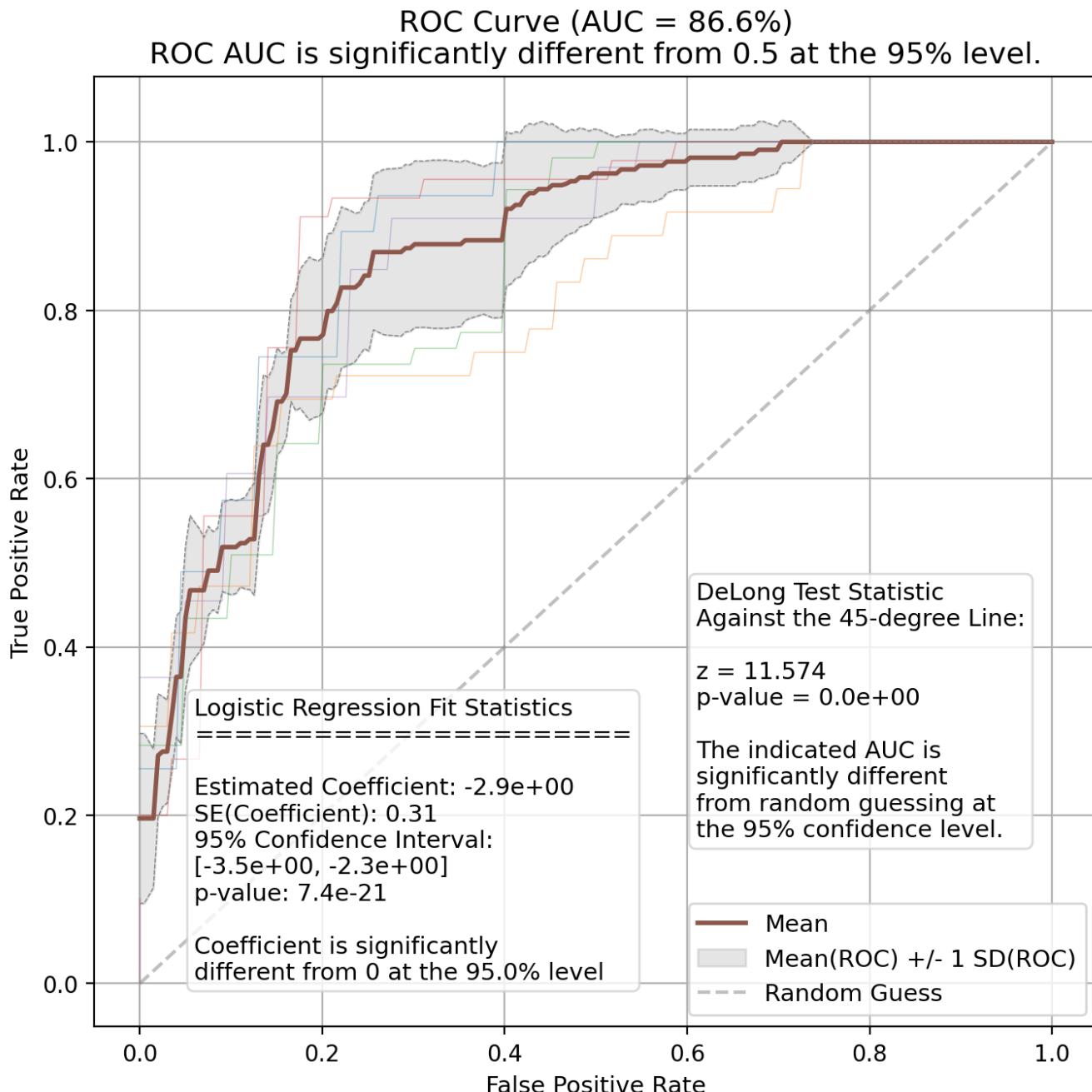
Perimeter Error - Empirical CDF Plot



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

Perimeter Error - ROC Curve

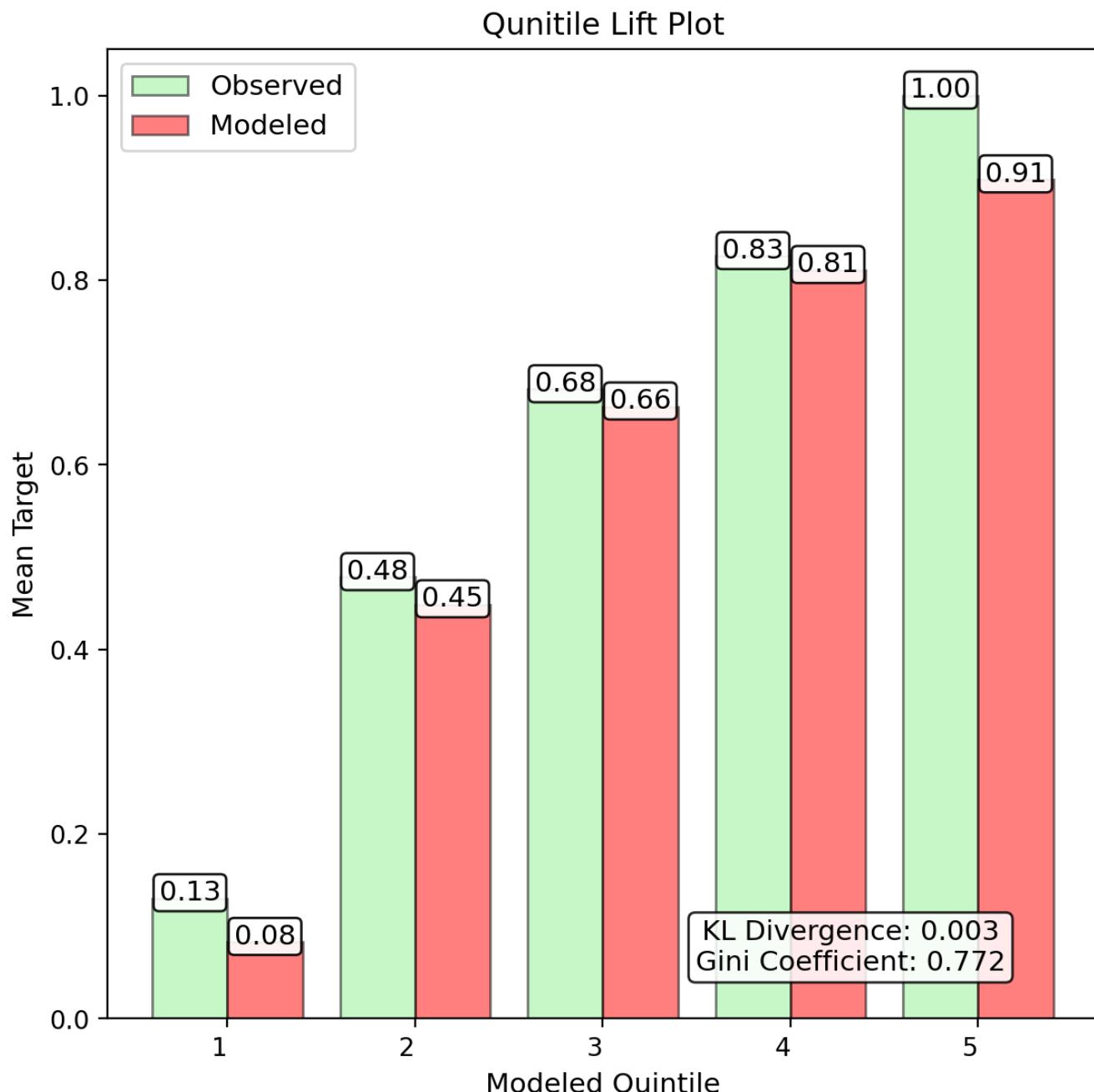


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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

Perimeter 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

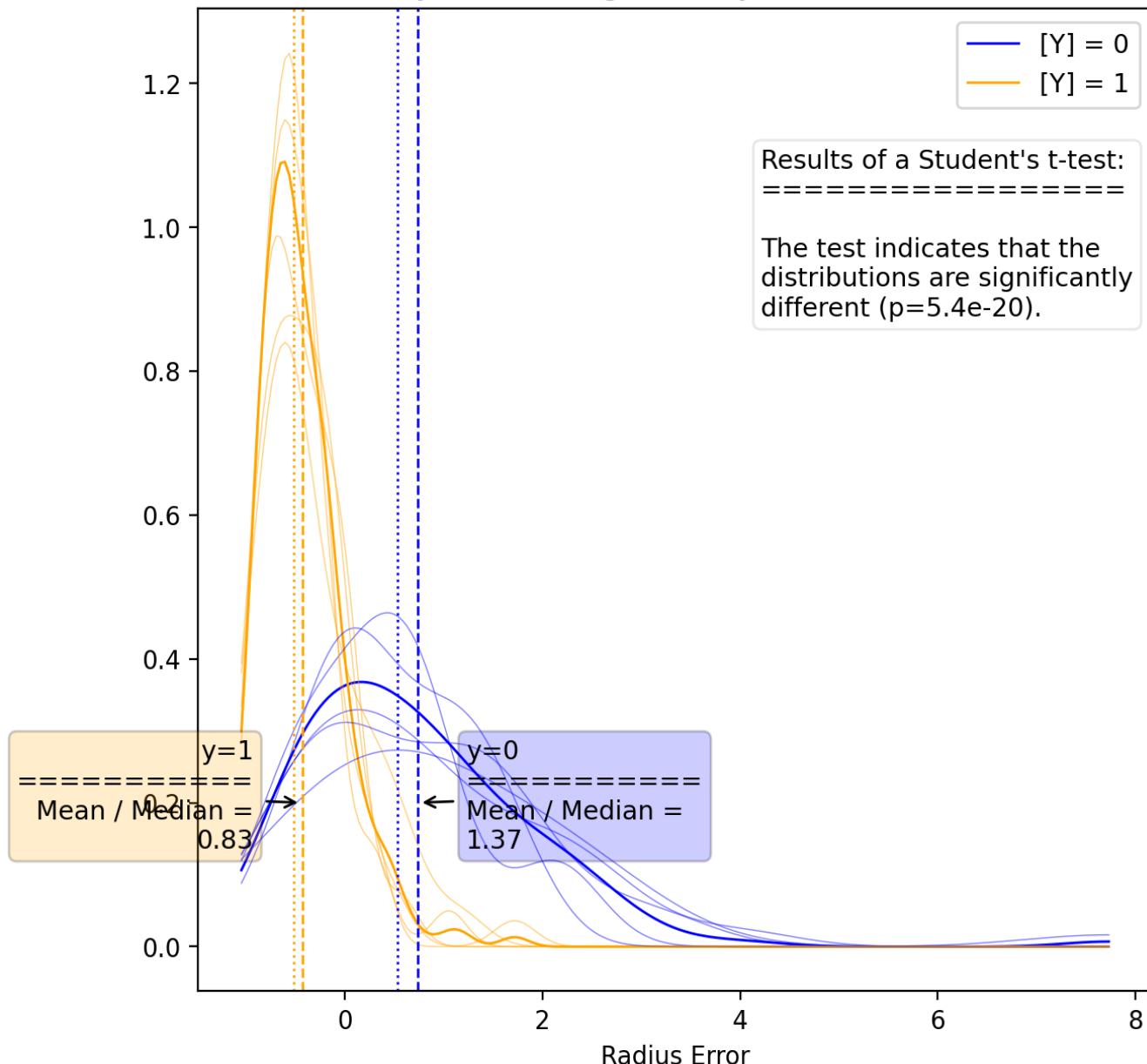
Radius Error - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-2.55	-3.09	-2.75	-2.67	-2.66	-2.73	0.20
Fitted p-Value	4.4e-16	1.3e-17	8.3e-17	2.7e-16	1.8e-17	1.1e-20	1.9e-16
Fitted Std. Err.	0.314	0.361	0.330	0.326	0.312	0.293	0.020
Conf. Int. Lower	-3.17	-3.80	-3.39	-3.31	-3.27	-3.31	0.24
Conf. Int. Upper	-1.94	-2.38	-2.10	-2.03	-2.04	-2.16	0.17
Train Accuracy	80.8%	82.4%	81.3%	80.5%	81.8%	80.9%	0.7%
Val Accuracy	85.7%	75.4%	82.2%	85.1%	80.0%	80.7%	4.2%
Train AUC	78.6%	80.0%	79.8%	78.8%	79.4%	78.7%	0.6%
Val AUC	82.7%	75.4%	75.3%	83.5%	78.0%	78.5%	3.9%
Train F1	85.0%	86.7%	84.9%	84.7%	86.0%	85.2%	0.8%
Test F1	89.6%	76.1%	88.1%	88.2%	84.1%	84.9%	5.5%
Train Precision	82.1%	85.7%	82.5%	84.2%	83.8%	83.1%	1.4%
Val Precision	87.8%	77.1%	85.7%	85.4%	80.6%	82.7%	4.3%
Train Recall	88.0%	87.6%	87.6%	85.2%	88.4%	87.4%	1.3%
Val Recall	91.5%	75.0%	90.6%	91.1%	87.9%	87.3%	7.0%
Train MCC	58.8%	60.6%	60.7%	57.9%	60.3%	58.6%	1.2%
Val MCC	67.0%	50.7%	53.4%	68.5%	57.8%	58.3%	8.0%
Train Log-Loss	6.92	6.36	6.72	7.02	6.55	6.87	0.27
Val Log-Loss	5.15	8.88	6.42	5.36	7.21	6.96	1.52

Univariate Report

Radius Error - Kernel Density Plot

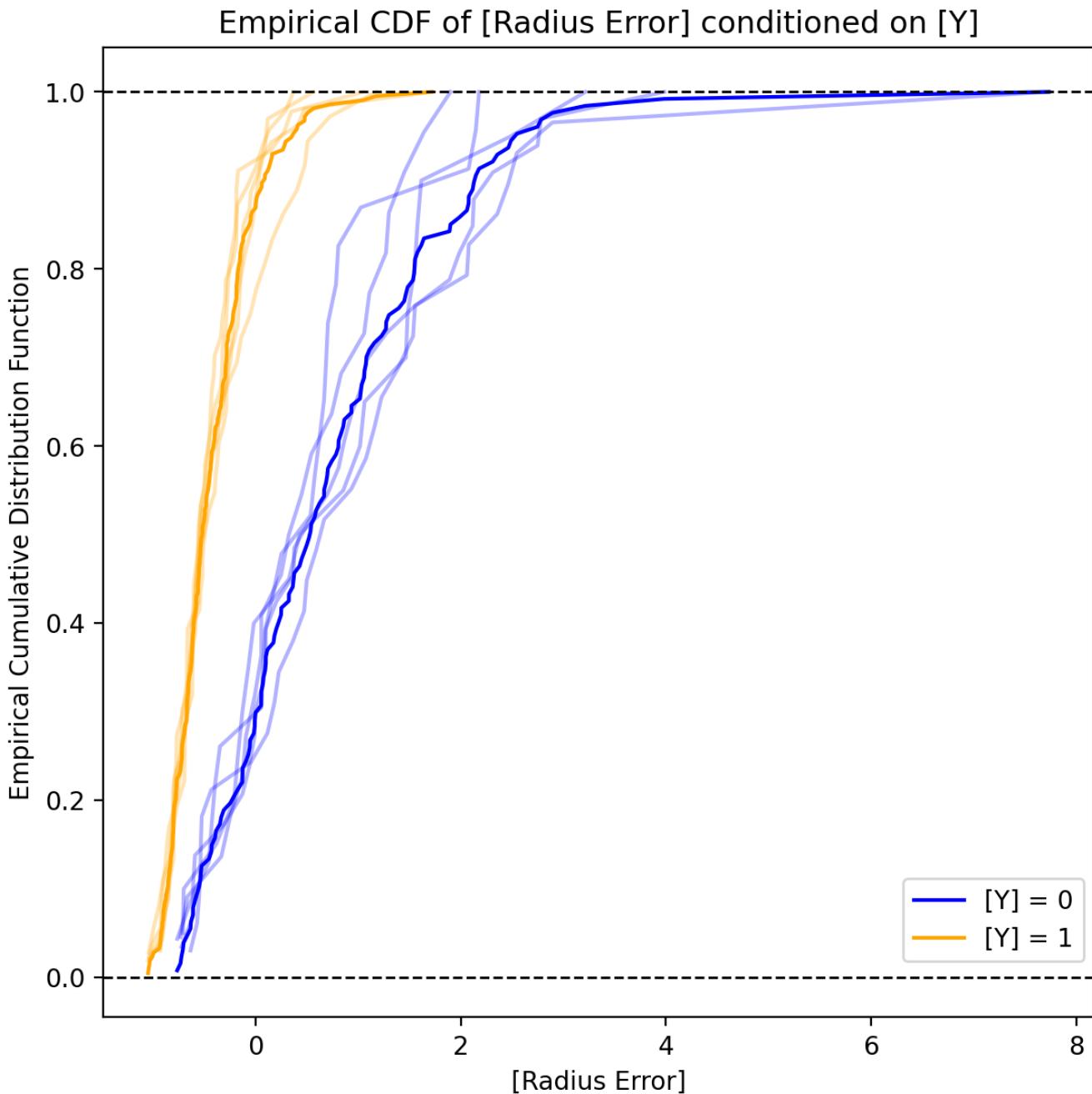
Kernel Density Plot of [Radius Error] by [Y].
Distributions by level are significantly different at the 95% level.



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

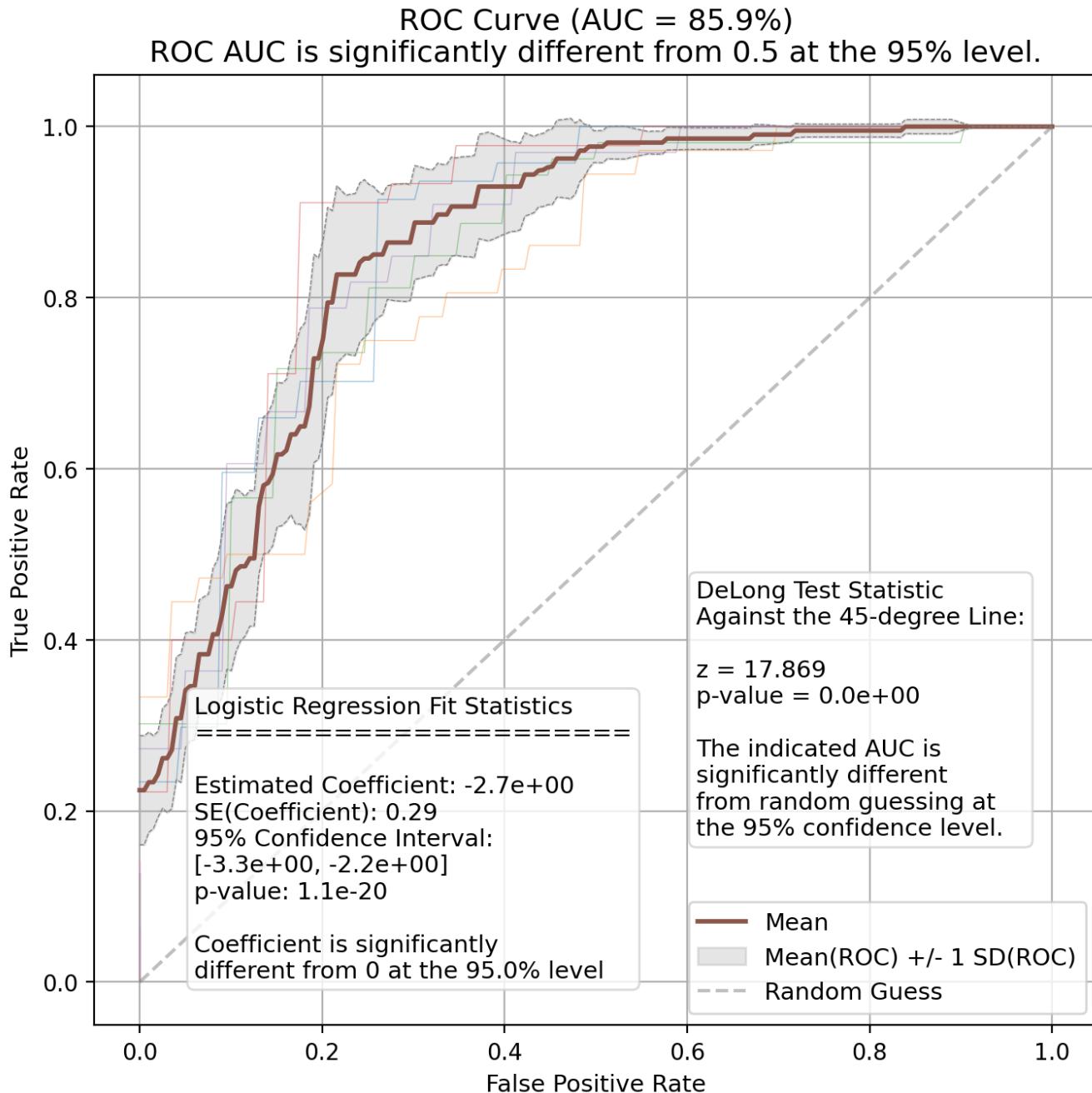
Radius Error - Empirical CDF Plot



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

Radius Error - ROC Curve

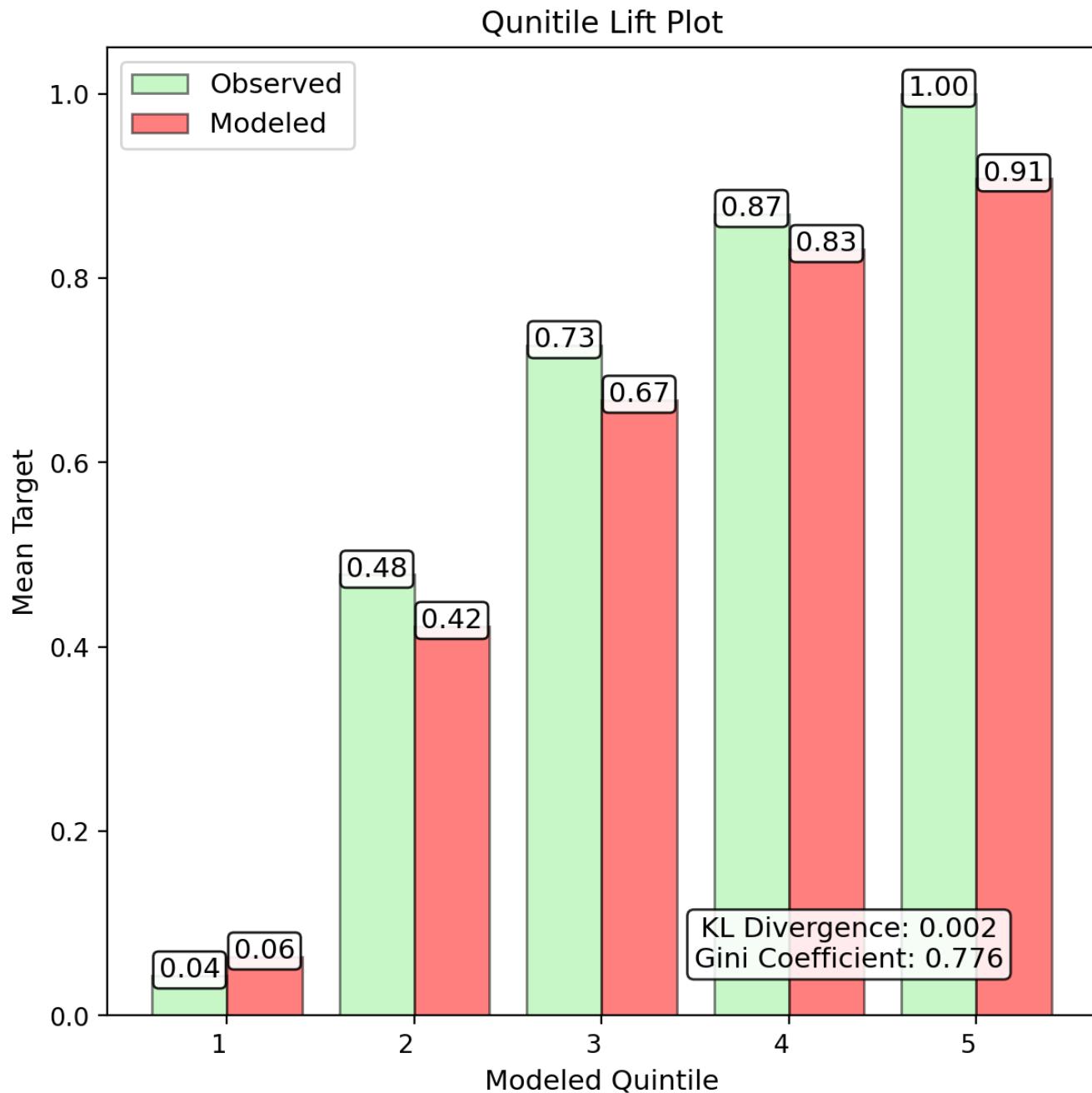


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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

Radius 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

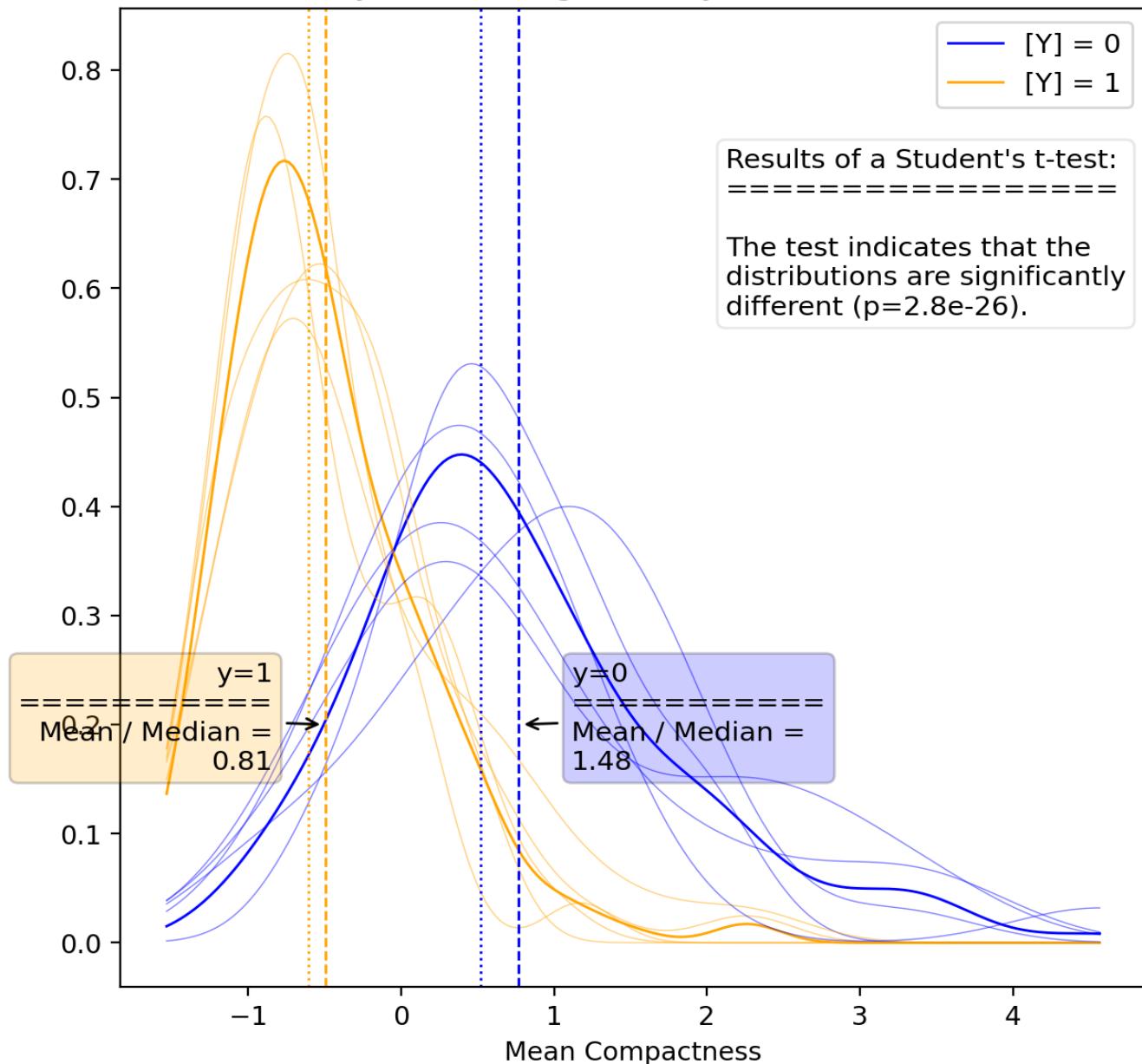
Mean Compactness - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-1.89	-2.51	-2.26	-2.10	-2.04	-2.15	0.23
Fitted p-Value	2.2e-15	9.5e-18	3.8e-17	1.2e-15	2.2e-17	1.6e-20	1.0e-15
Fitted Std. Err.	0.239	0.292	0.269	0.263	0.241	0.231	0.022
Conf. Int. Lower	-2.36	-3.08	-2.79	-2.62	-2.51	-2.60	0.28
Conf. Int. Upper	-1.42	-1.94	-1.74	-1.59	-1.57	-1.69	0.19
Train Accuracy	77.9%	81.2%	81.0%	82.0%	81.1%	81.2%	1.6%
Val Accuracy	88.6%	76.8%	80.8%	78.4%	80.0%	78.9%	4.5%
Train AUC	77.7%	82.4%	81.0%	82.2%	81.5%	81.5%	1.9%
Val AUC	87.0%	77.1%	80.6%	79.2%	78.8%	79.0%	3.8%
Train F1	81.4%	84.6%	83.6%	85.2%	84.3%	84.3%	1.5%
Test F1	91.5%	75.8%	86.0%	81.0%	83.6%	82.4%	5.8%
Train Precision	84.5%	91.5%	86.7%	89.0%	89.0%	88.7%	2.6%
Val Precision	91.5%	83.3%	91.5%	87.2%	82.4%	86.2%	4.3%
Train Recall	78.4%	78.7%	80.7%	81.7%	80.1%	80.4%	1.4%
Val Recall	91.5%	69.4%	81.1%	75.6%	84.8%	78.9%	8.5%
Train MCC	54.4%	62.1%	61.2%	62.8%	61.3%	61.6%	3.4%
Val MCC	74.1%	54.7%	56.9%	57.0%	58.1%	56.7%	7.9%
Train Log-Loss	7.98	6.76	6.86	6.48	6.81	6.76	0.58
Val Log-Loss	4.12	8.36	6.91	7.79	7.21	7.59	1.64

Univariate Report

Mean Compactness - Kernel Density Plot

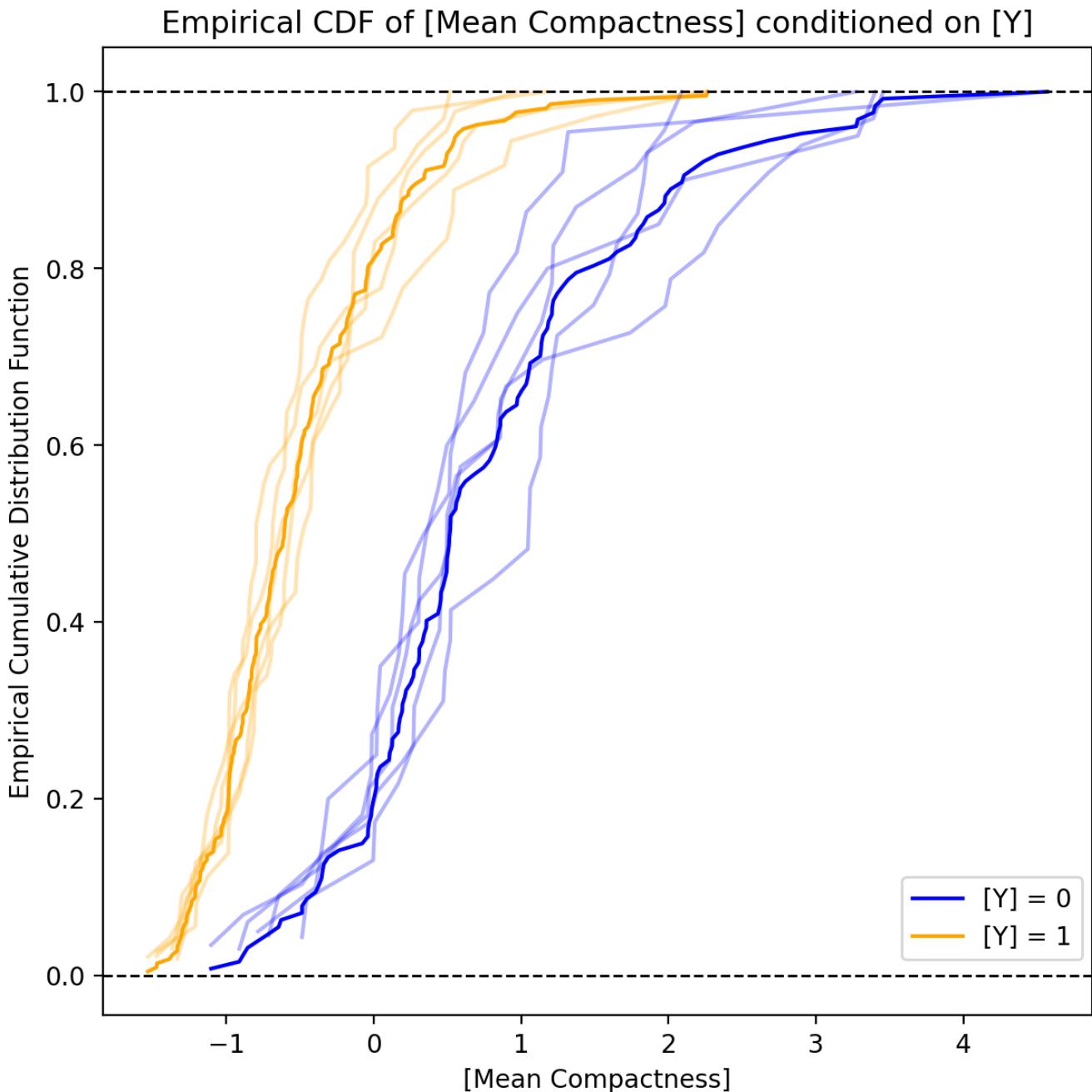
Kernel Density Plot of [Mean Compactness] by [Y].
Distributions by level are significantly different at the 95% level.



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

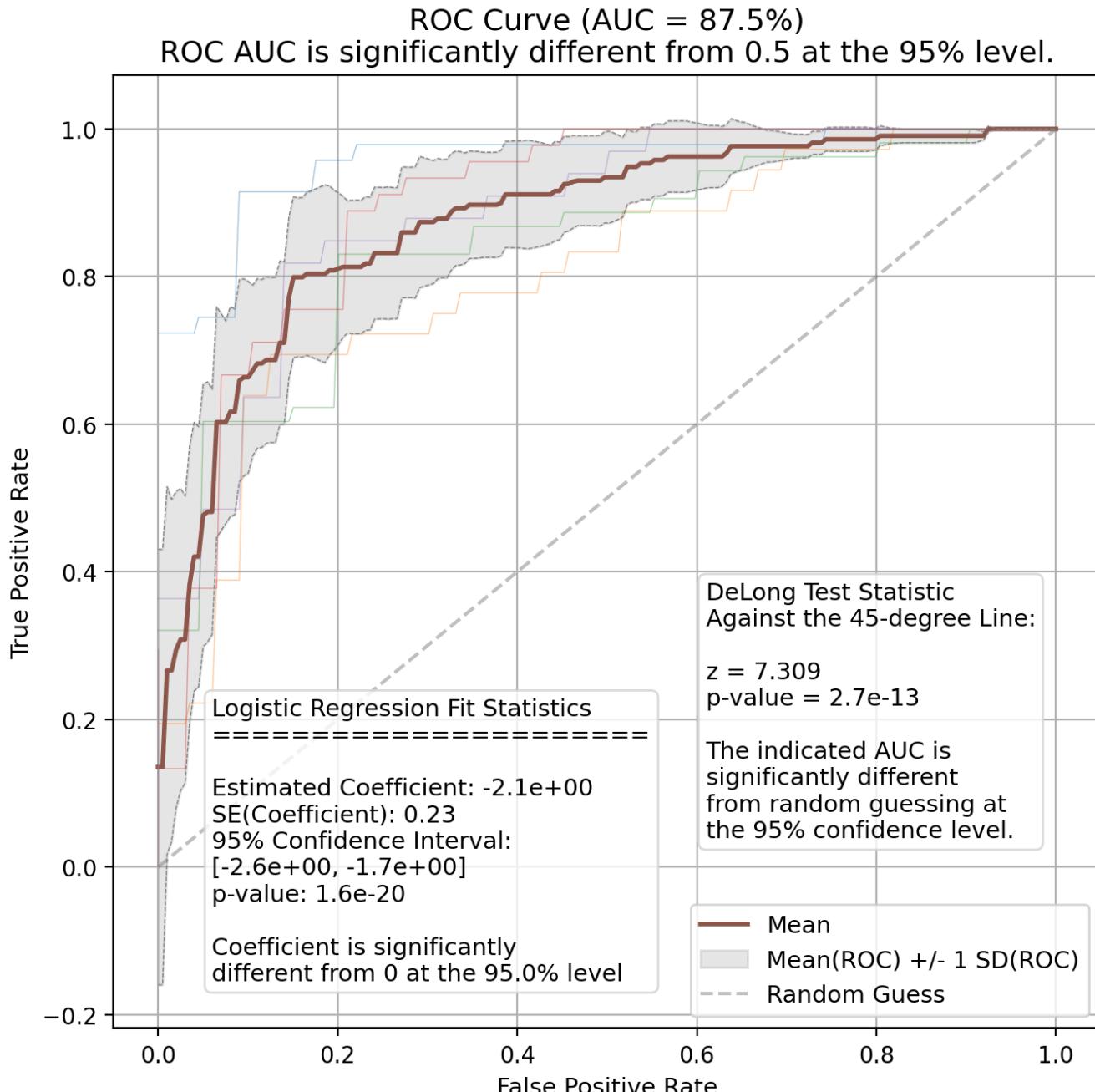
Mean Compactness - Empirical CDF Plot



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 Compactness - ROC Curve

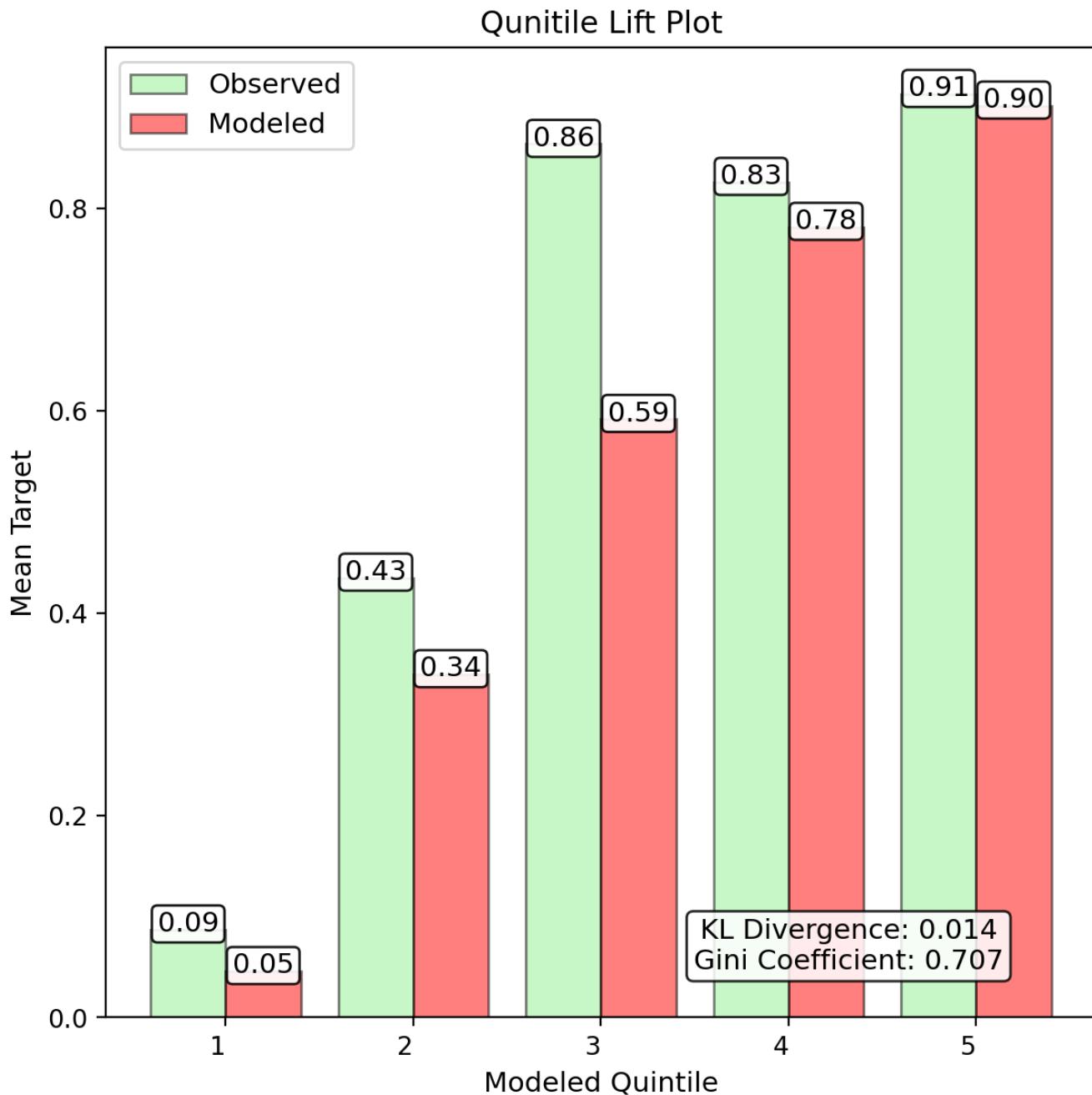


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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 Compactness - 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

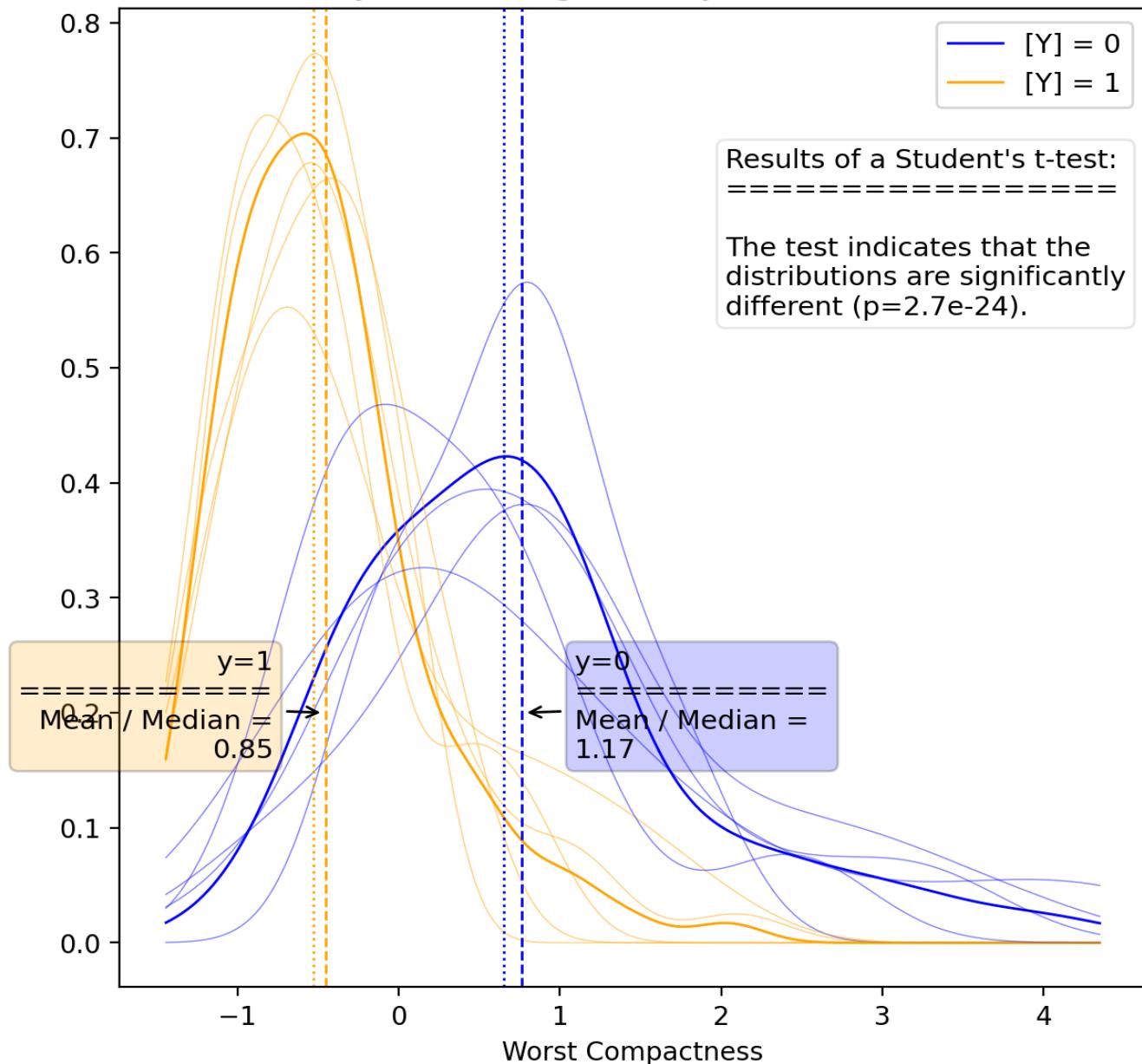
Worst Compactness - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-1.76	-2.47	-2.24	-2.00	-2.04	-2.08	0.27
Fitted p-Value	1.2e-14	2.3e-17	6.4e-17	4.3e-15	2.6e-17	5.2e-20	5.1e-15
Fitted Std. Err.	0.227	0.292	0.268	0.254	0.241	0.227	0.025
Conf. Int. Lower	-2.20	-3.04	-2.76	-2.49	-2.52	-2.53	0.32
Conf. Int. Upper	-1.31	-1.90	-1.71	-1.50	-1.57	-1.64	0.22
Train Accuracy	77.1%	79.4%	81.0%	77.2%	81.8%	80.1%	2.2%
Val Accuracy	90.0%	76.8%	76.7%	85.1%	70.9%	75.4%	7.6%
Train AUC	76.0%	78.0%	80.1%	75.7%	81.0%	78.8%	2.4%
Val AUC	88.1%	77.0%	73.1%	85.3%	68.2%	75.2%	8.3%
Train F1	81.3%	84.0%	84.2%	81.8%	85.4%	84.0%	1.7%
Test F1	92.6%	76.5%	83.5%	87.4%	77.1%	79.4%	6.9%
Train Precision	81.8%	85.5%	84.0%	82.5%	86.9%	84.4%	2.1%
Val Precision	91.7%	81.2%	86.0%	90.5%	73.0%	83.1%	7.6%
Train Recall	80.8%	82.6%	84.5%	81.1%	84.0%	83.6%	1.6%
Val Recall	93.6%	72.2%	81.1%	84.4%	81.8%	76.1%	7.7%
Train MCC	51.8%	55.2%	60.3%	51.2%	61.4%	57.5%	4.7%
Val MCC	77.1%	54.1%	44.3%	69.6%	38.0%	49.4%	16.5%
Train Log-Loss	8.25	7.42	6.86	8.23	6.55	7.19	0.78
Val Log-Loss	3.60	8.36	8.39	5.36	10.49	8.85	2.73

Univariate Report

Worst Compactness - Kernel Density Plot

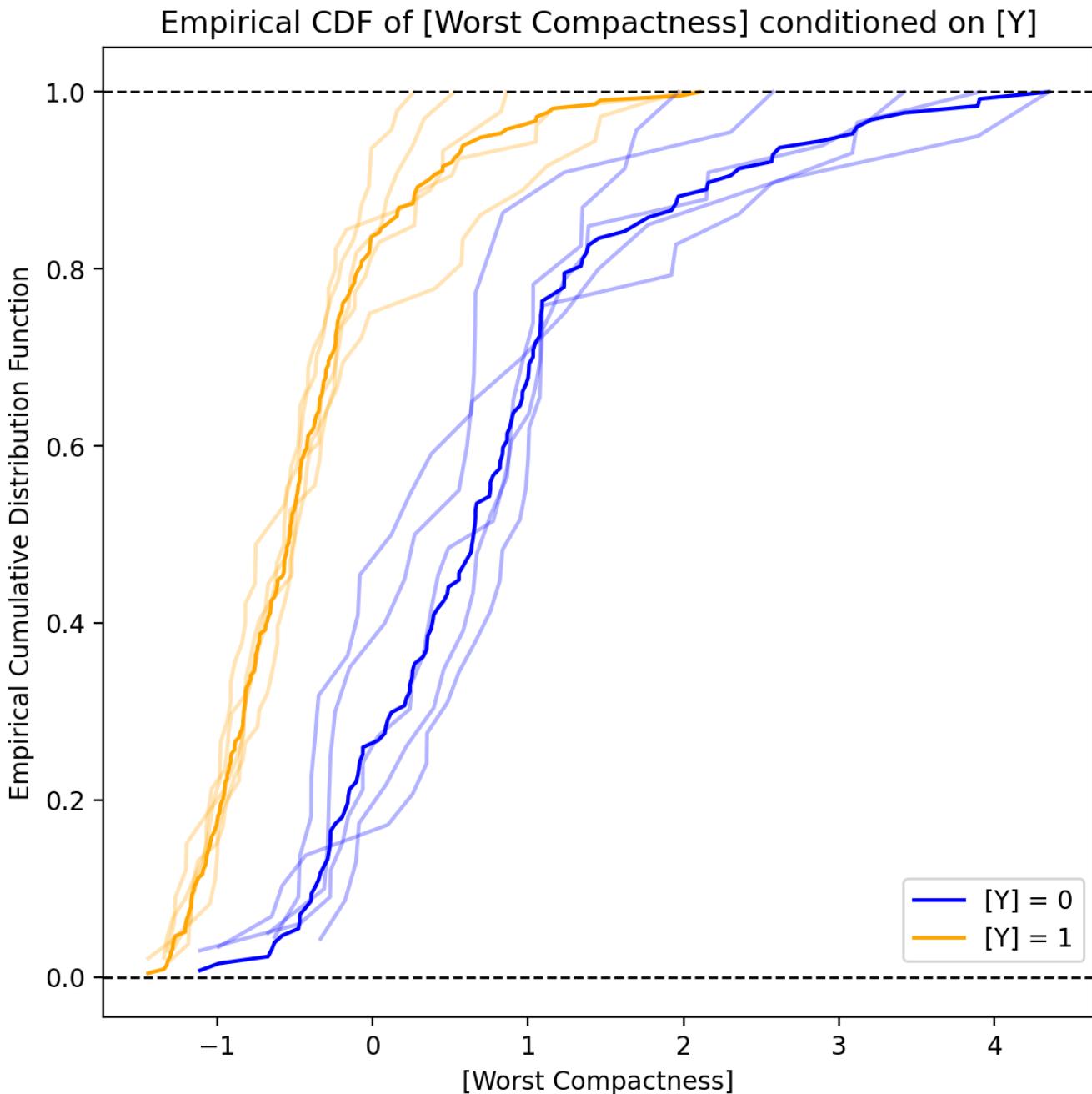
Kernel Density Plot of [Worst Compactness] by [Y].
Distributions by level are significantly different at the 95% level.



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

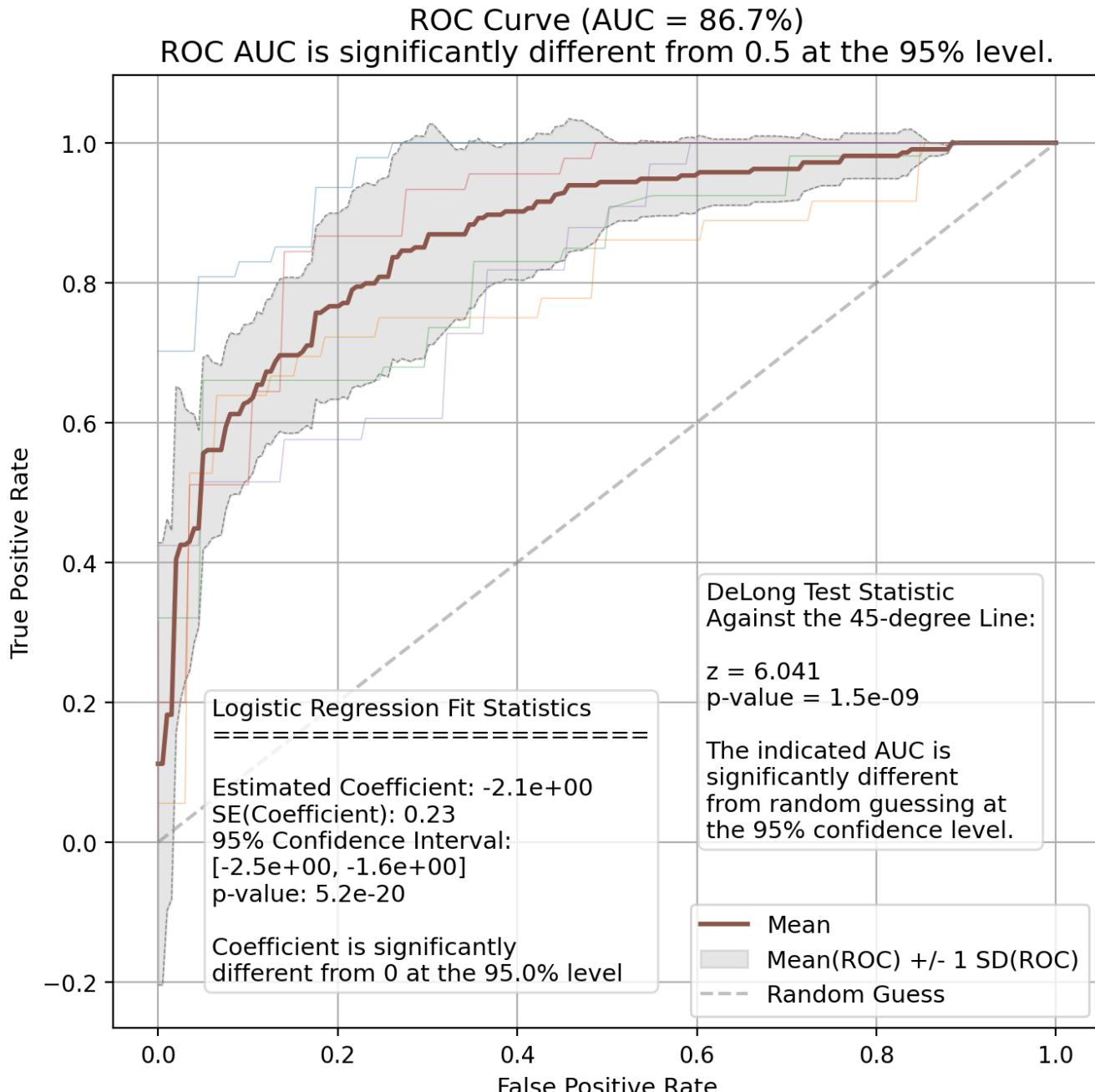
Worst Compactness - Empirical CDF Plot



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 Compactness - ROC Curve

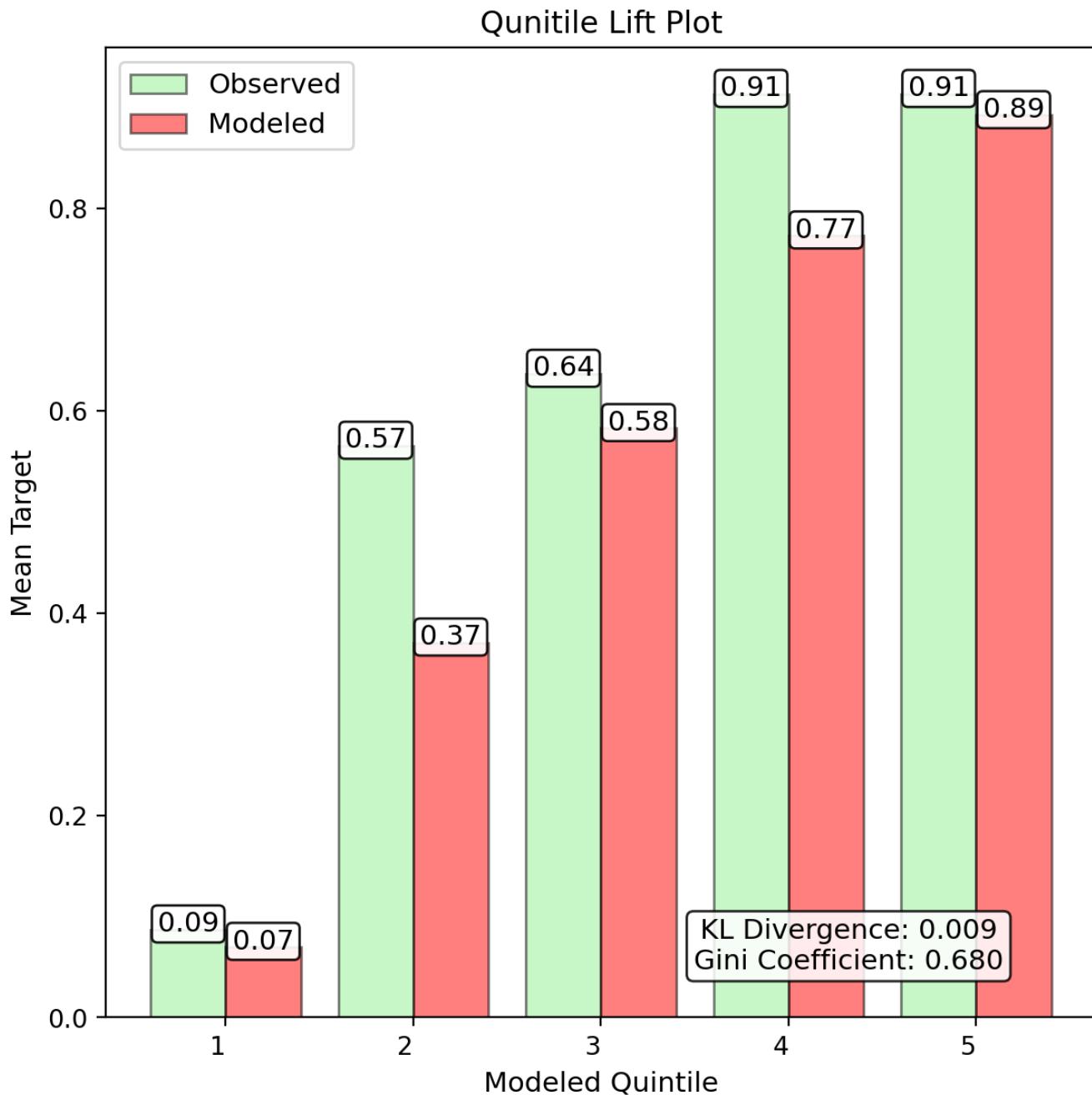


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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 Compactness - 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

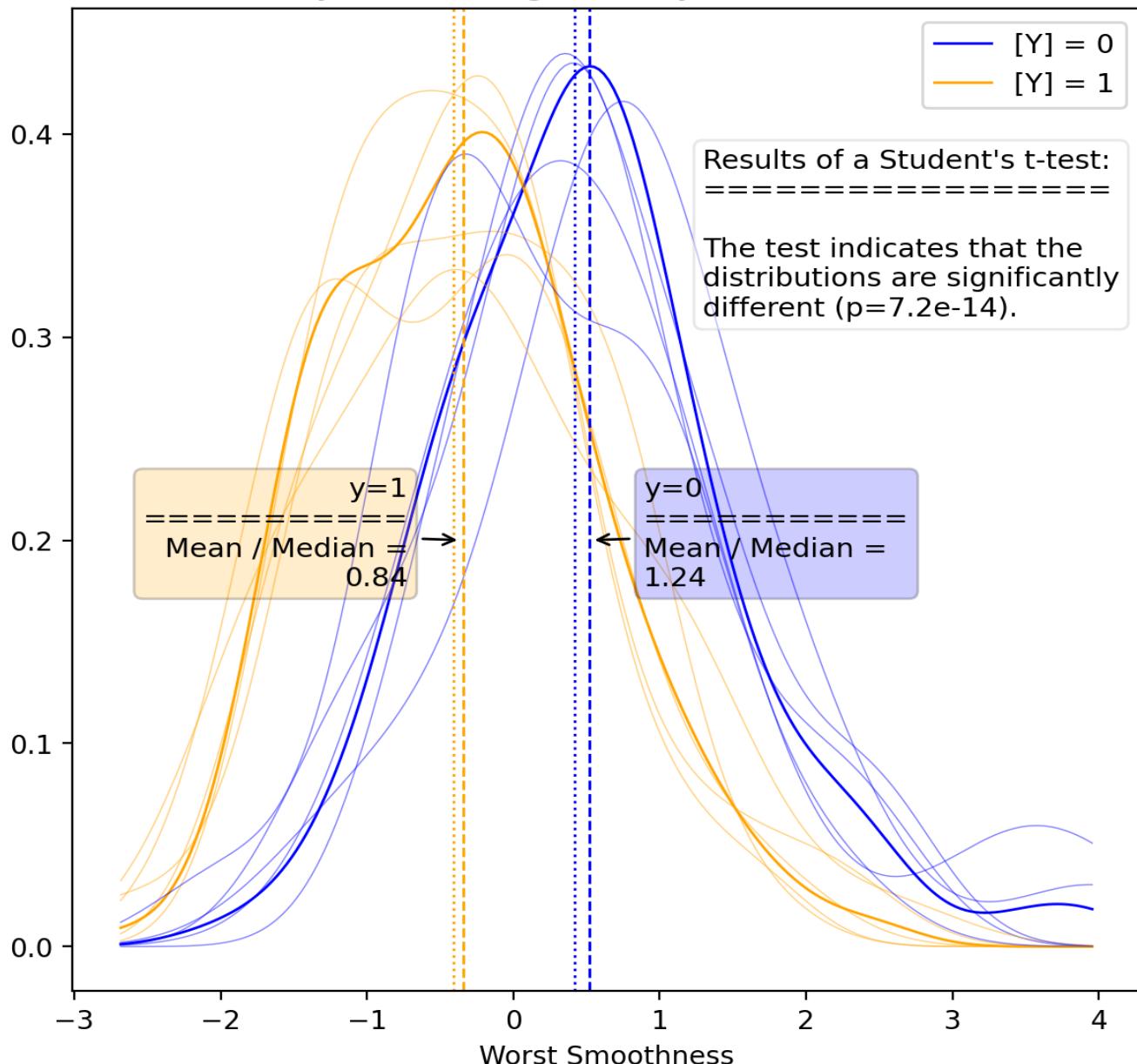
Worst Smoothness - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-0.895	-1.020	-0.847	-0.907	-0.970	-0.926	0.068
Fitted p-Value	4.5e-09	2.7e-10	1.0e-08	3.0e-09	2.4e-10	1.5e-11	4.0e-09
Fitted Std. Err.	0.153	0.162	0.148	0.153	0.153	0.137	0.005
Conf. Int. Lower	-1.19	-1.34	-1.14	-1.21	-1.27	-1.20	0.08
Conf. Int. Upper	-0.596	-0.703	-0.557	-0.607	-0.670	-0.657	0.059
Train Accuracy	66.4%	66.5%	66.8%	66.3%	68.9%	66.3%	1.1%
Val Accuracy	68.6%	68.1%	67.1%	68.9%	58.2%	74.6%	4.5%
Train AUC	66.8%	67.7%	67.2%	67.2%	70.0%	66.9%	1.3%
Val AUC	73.3%	68.3%	68.0%	70.2%	56.1%	75.5%	6.5%
Train F1	70.6%	71.5%	70.2%	70.6%	72.8%	70.6%	1.0%
Test F1	71.8%	67.6%	74.5%	71.6%	65.7%	77.9%	3.5%
Train Precision	76.8%	80.9%	76.1%	78.8%	81.5%	78.0%	2.4%
Val Precision	90.3%	71.9%	85.4%	80.6%	64.7%	85.0%	10.3%
Train Recall	65.3%	64.0%	65.2%	63.9%	65.7%	64.5%	0.8%
Val Recall	59.6%	63.9%	66.0%	64.4%	66.7%	71.8%	2.8%
Train MCC	32.7%	33.6%	33.7%	33.1%	38.6%	32.7%	2.4%
Val MCC	44.0%	36.7%	32.4%	39.4%	12.2%	49.4%	12.3%
Train Log-Loss	12.10	12.06	11.97	12.15	11.22	12.16	0.39
Val Log-Loss	11.33	11.49	11.85	11.20	15.07	9.17	1.63

Univariate Report

Worst Smoothness - Kernel Density Plot

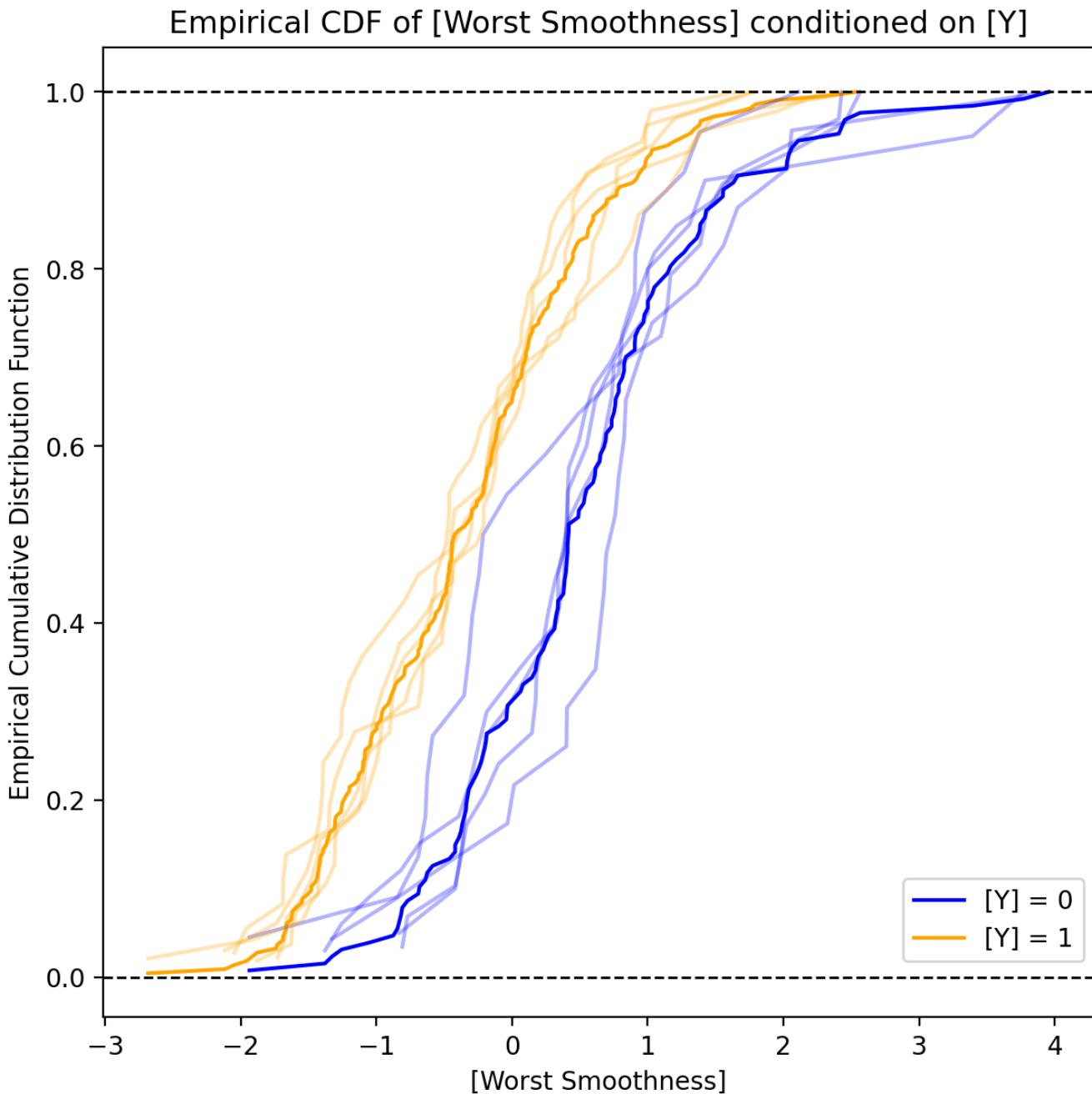
Kernel Density Plot of [Worst Smoothness] by [Y].
Distributions by level are significantly different at the 95% level.



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

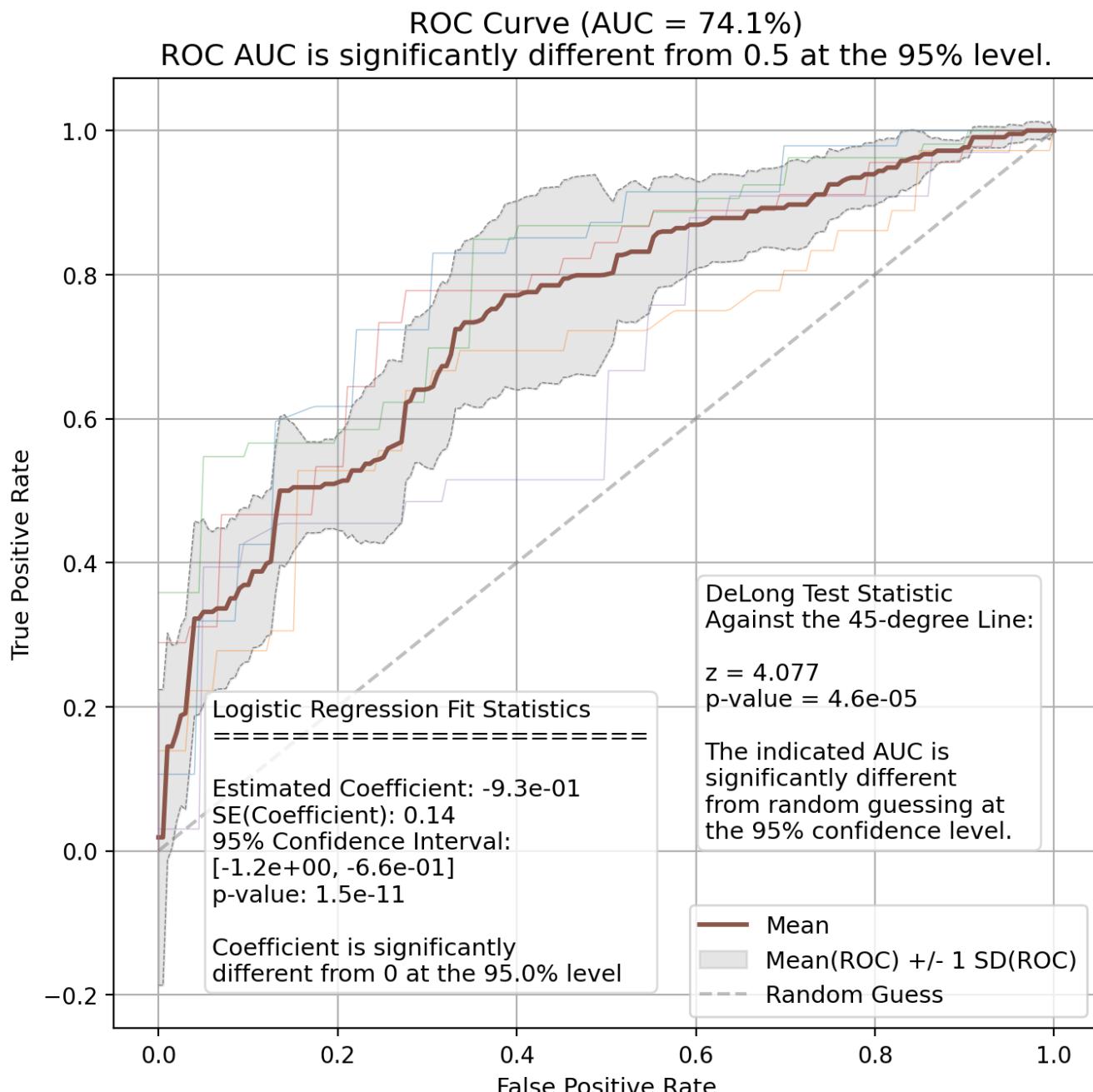
Worst Smoothness - Empirical CDF Plot



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 Smoothness - ROC Curve

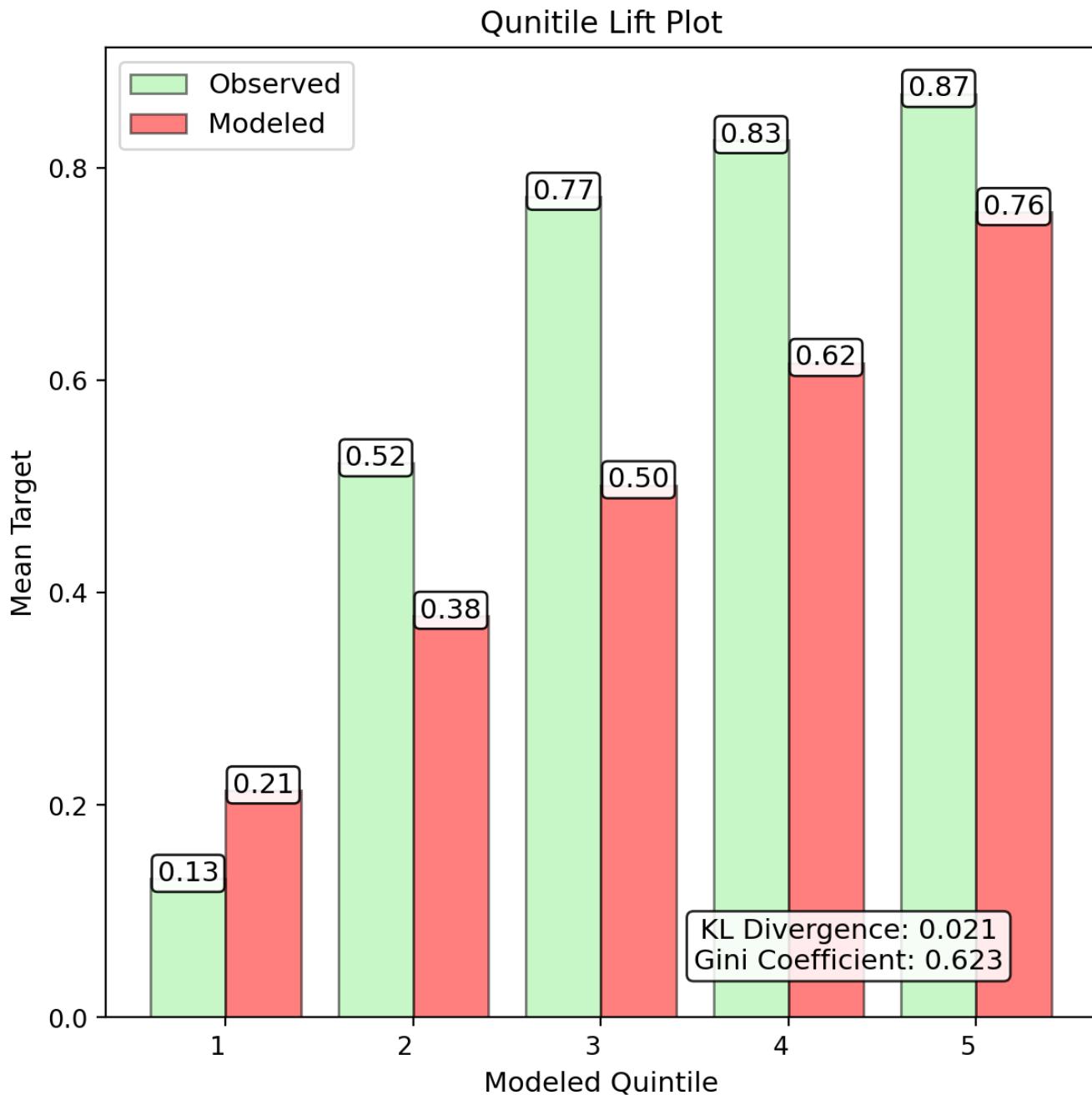


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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 Smoothness - 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

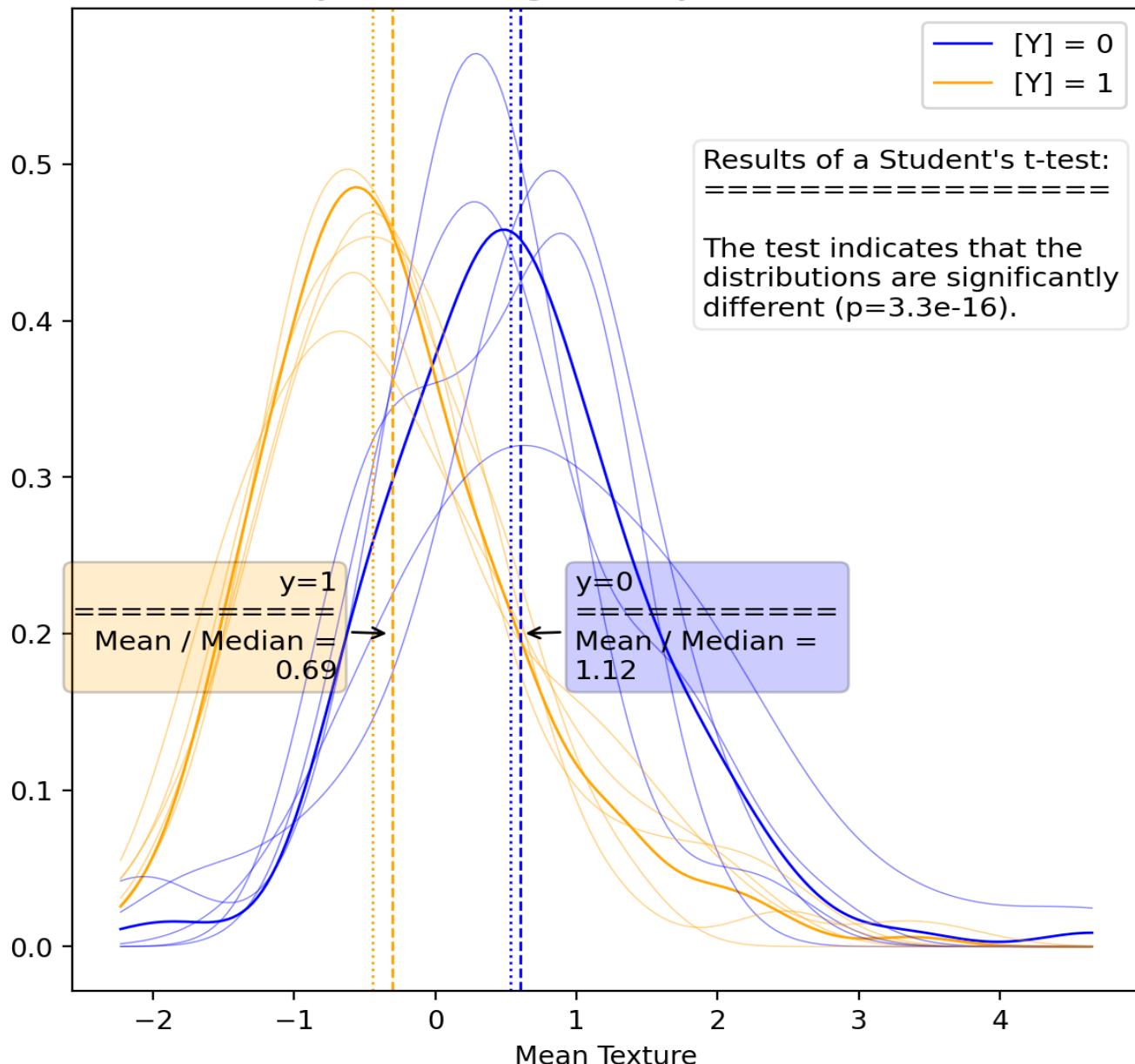
Mean Texture - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-0.928	-0.998	-1.221	-0.892	-1.032	-1.009	0.128
Fitted p-Value	1.9e-09	7.5e-10	7.2e-12	1.9e-08	1.5e-10	3.5e-12	8.1e-09
Fitted Std. Err.	0.154	0.162	0.178	0.159	0.161	0.145	0.009
Conf. Int. Lower	-1.23	-1.32	-1.57	-1.20	-1.35	-1.29	0.15
Conf. Int. Upper	-0.625	-0.680	-0.872	-0.581	-0.716	-0.725	0.112
Train Accuracy	72.7%	72.4%	73.1%	70.4%	70.6%	71.8%	1.3%
Val Accuracy	70.0%	69.6%	67.1%	75.7%	78.2%	71.9%	4.6%
Train AUC	72.9%	72.9%	73.1%	71.1%	70.8%	72.1%	1.1%
Val AUC	67.7%	69.7%	68.0%	76.3%	79.5%	72.0%	5.4%
Train F1	76.4%	77.2%	76.6%	74.6%	75.1%	76.0%	1.1%
Test F1	76.9%	69.6%	74.5%	78.6%	80.0%	76.1%	4.1%
Train Precision	81.6%	84.1%	80.3%	81.7%	80.9%	81.7%	1.5%
Val Precision	79.5%	72.7%	85.4%	84.6%	88.9%	81.0%	6.3%
Train Recall	71.9%	71.3%	73.3%	68.6%	70.2%	71.0%	1.8%
Val Recall	74.5%	66.7%	66.0%	73.3%	72.7%	71.8%	4.0%
Train MCC	44.8%	43.8%	45.5%	40.7%	40.3%	43.0%	2.4%
Val MCC	34.4%	39.4%	32.4%	51.5%	57.9%	42.8%	11.1%
Train Log-Loss	9.84	9.94	9.68	10.66	10.59	10.15	0.45
Val Log-Loss	10.81	10.97	11.85	8.77	7.86	10.12	1.67

Univariate Report

Mean Texture - Kernel Density Plot

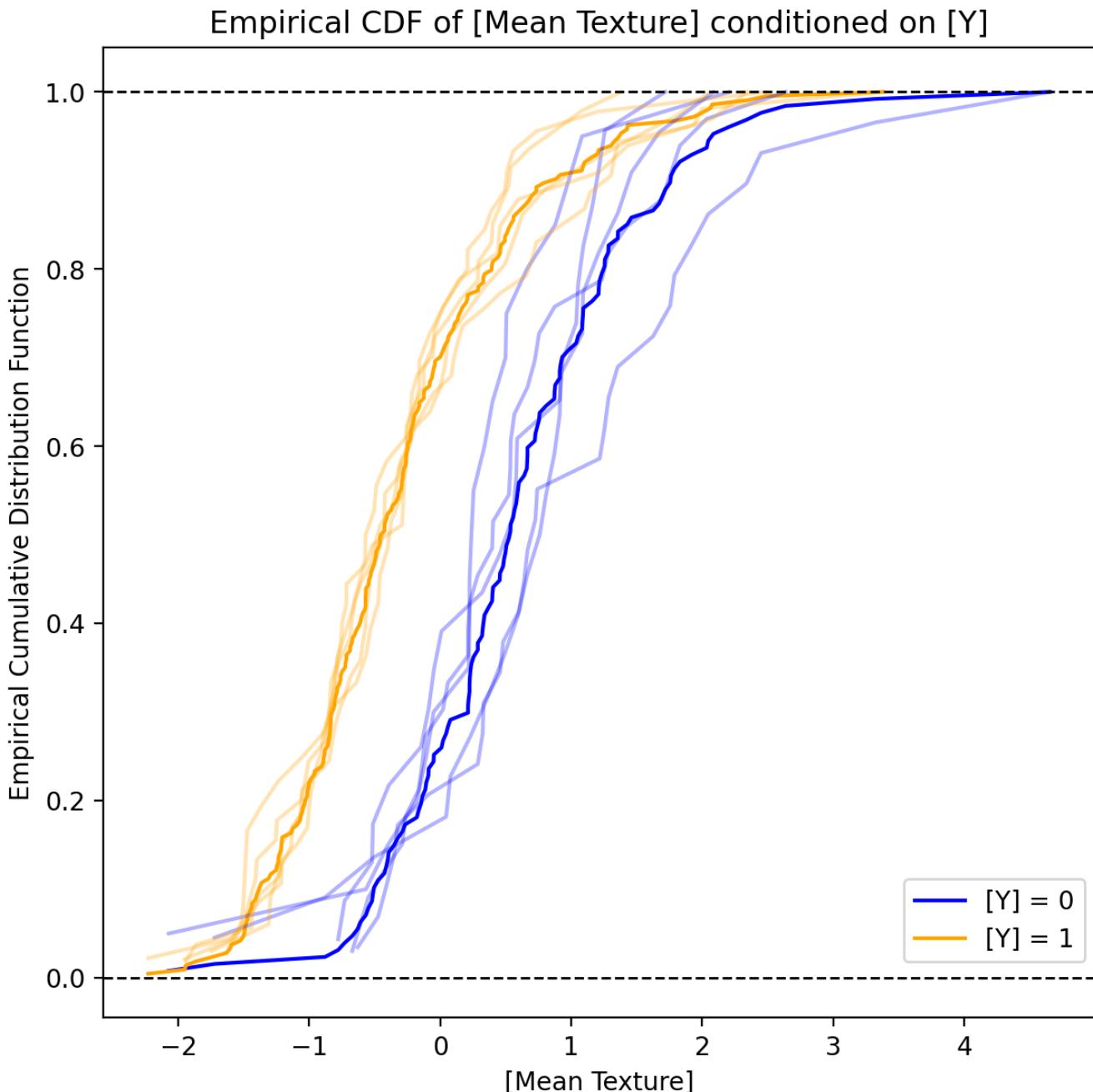
Kernel Density Plot of [Mean Texture] by [Y].
Distributions by level are significantly different at the 95% level.



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

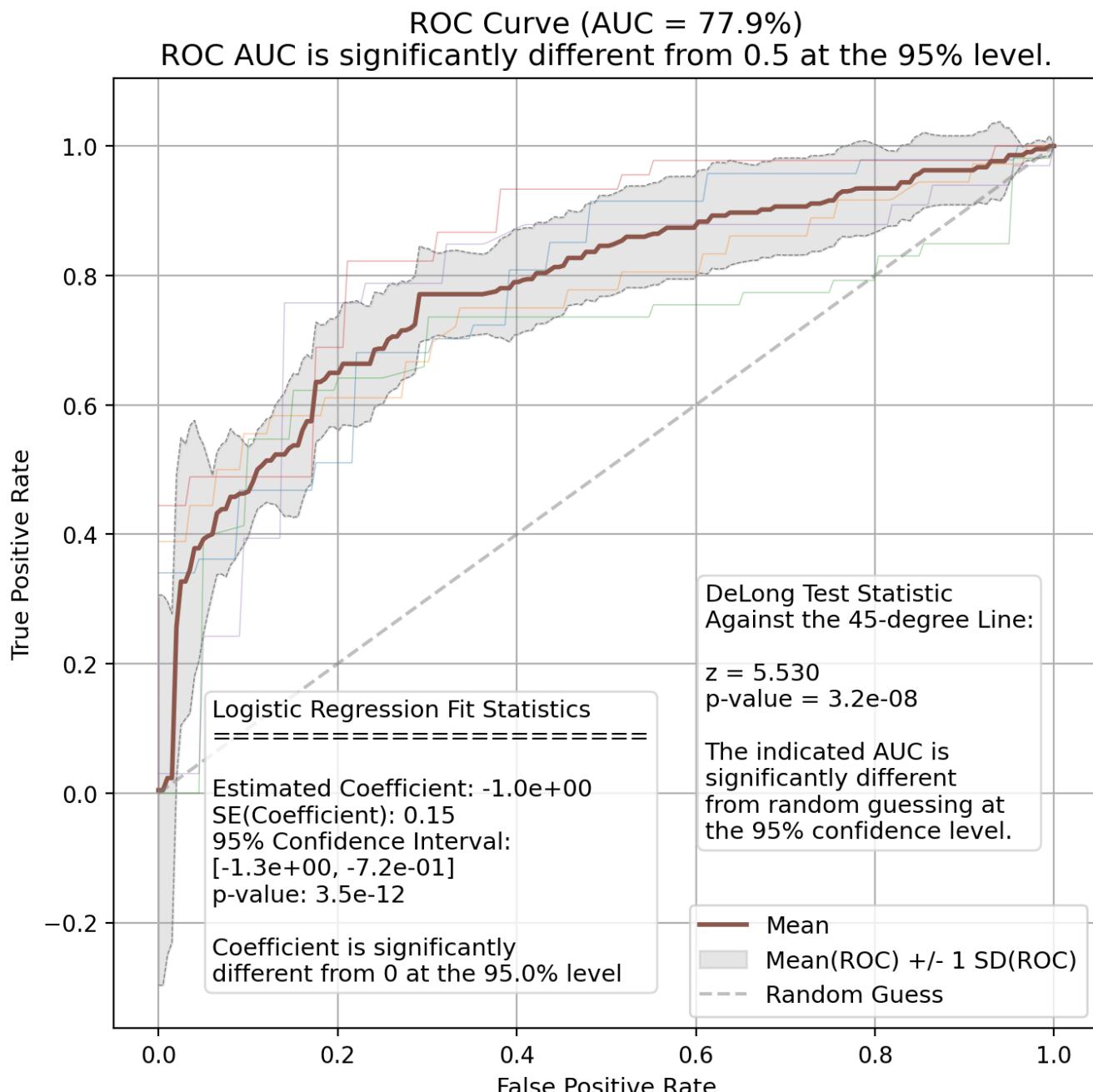
Mean Texture - Empirical CDF Plot



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

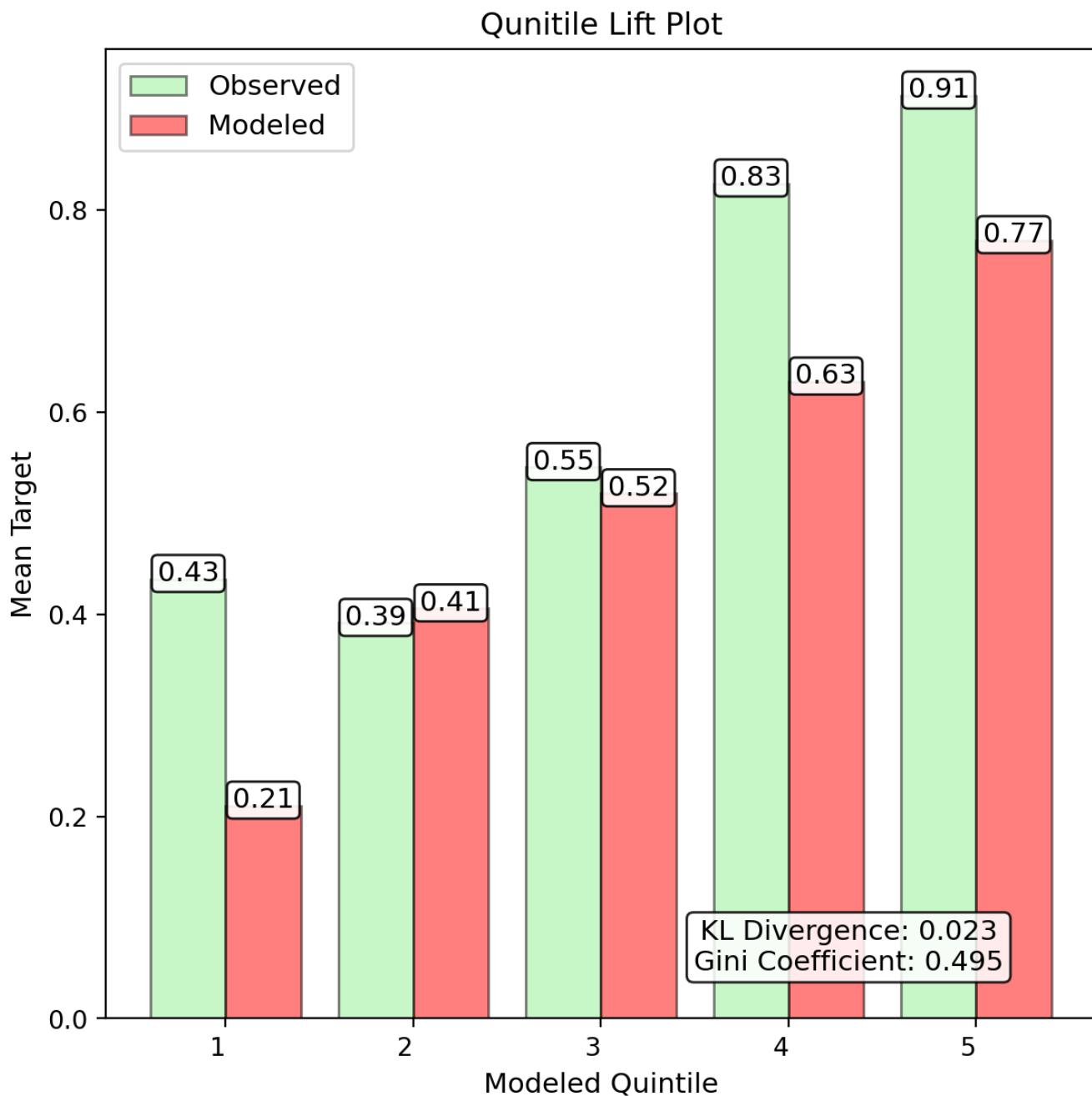


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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

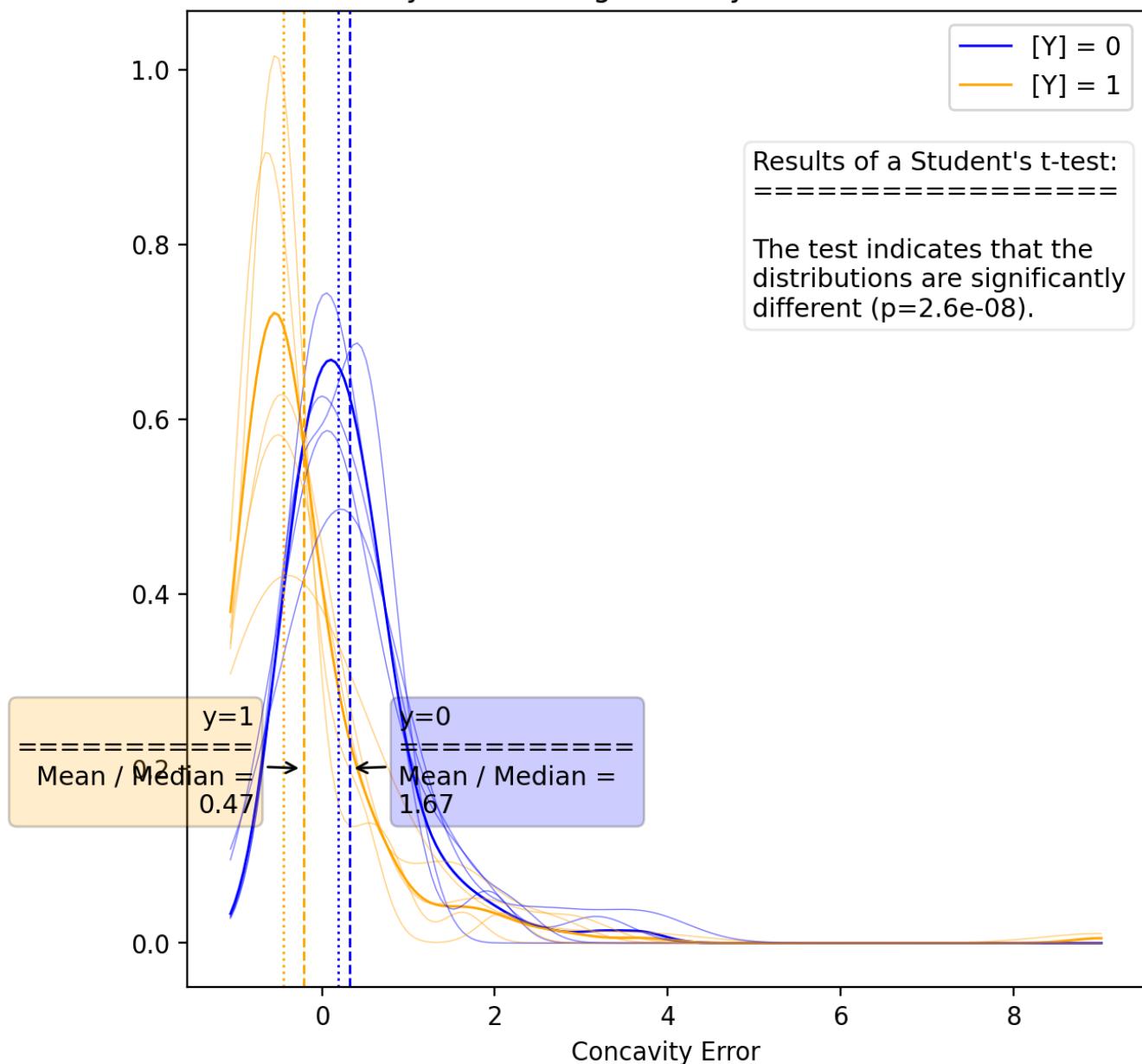
Concavity Error - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-0.663	-0.923	-1.383	-0.609	-0.875	-0.863	0.306
Fitted p-Value	1.7e-04	5.4e-06	2.7e-09	5.7e-04	5.8e-06	7.1e-07	2.4e-04
Fitted Std. Err.	0.176	0.203	0.232	0.177	0.193	0.174	0.023
Conf. Int. Lower	-1.01	-1.32	-1.84	-0.96	-1.25	-1.20	0.35
Conf. Int. Upper	-0.318	-0.525	-0.928	-0.263	-0.497	-0.522	0.261
Train Accuracy	70.5%	72.8%	74.6%	70.8%	71.3%	72.1%	1.7%
Val Accuracy	77.1%	68.1%	61.6%	74.3%	74.5%	70.2%	6.3%
Train AUC	68.8%	71.2%	73.2%	69.0%	69.8%	70.3%	1.8%
Val AUC	71.9%	67.9%	61.1%	70.9%	74.2%	67.8%	5.1%
Train F1	76.0%	78.6%	79.1%	76.6%	77.0%	77.8%	1.3%
Test F1	83.7%	70.3%	70.2%	80.4%	78.1%	76.4%	6.1%
Train Precision	76.0%	81.0%	78.2%	77.6%	78.3%	77.9%	1.8%
Val Precision	80.4%	68.4%	80.5%	75.0%	80.6%	75.3%	5.4%
Train Recall	76.0%	76.4%	80.1%	75.7%	75.7%	77.6%	1.9%
Val Recall	87.2%	72.2%	62.3%	86.7%	75.8%	77.5%	10.5%
Train MCC	37.6%	41.5%	46.8%	37.7%	39.1%	40.5%	3.8%
Val MCC	46.2%	36.0%	20.0%	44.7%	47.9%	36.0%	11.5%
Train Log-Loss	10.64	9.81	9.15	10.53	10.33	10.04	0.62
Val Log-Loss	8.24	11.49	13.82	9.25	9.17	10.75	2.26

Univariate Report

Concavity Error - Kernel Density Plot

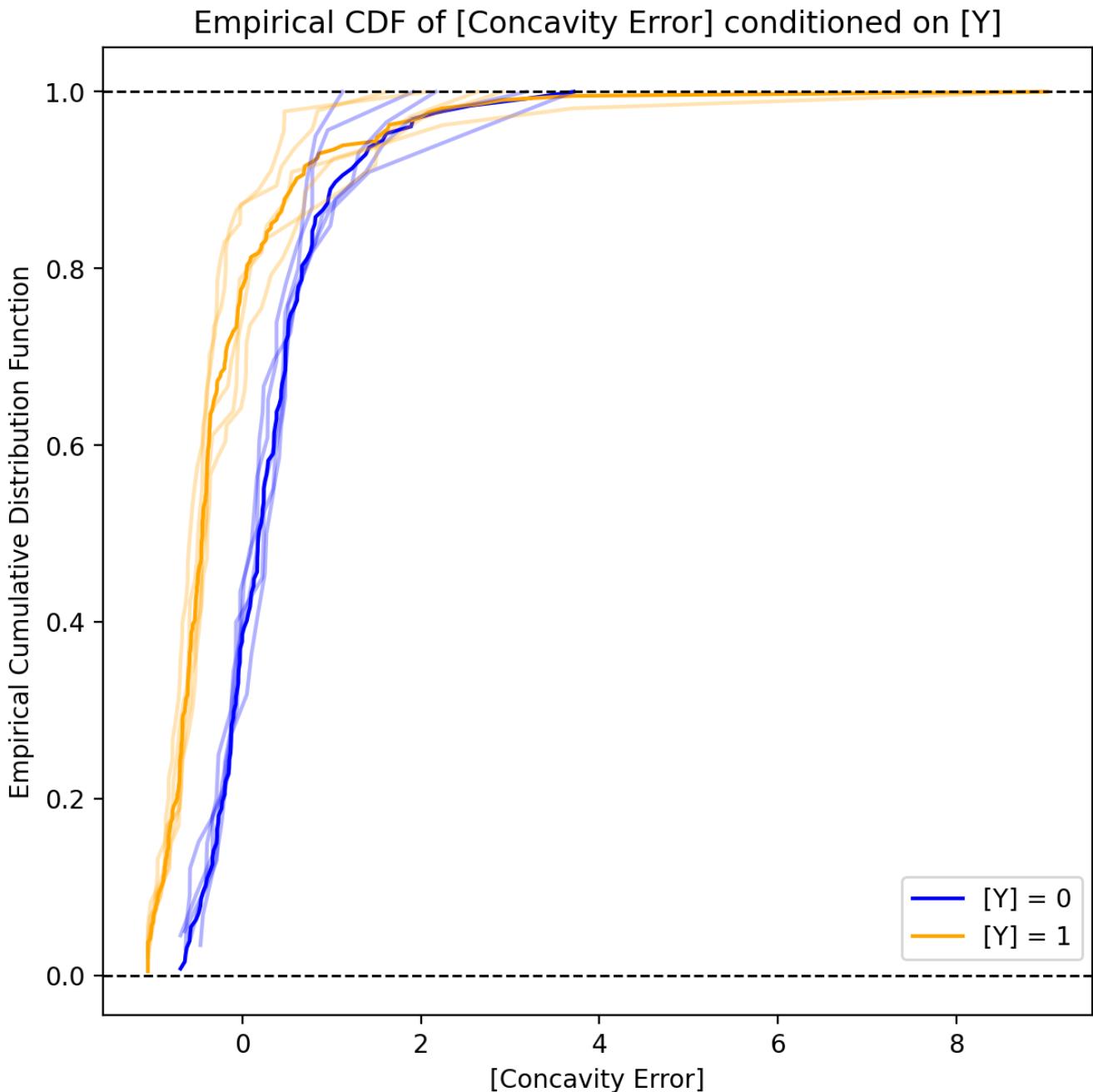
Kernel Density Plot of [Concavity Error] by [Y].
Distributions by level are significantly different at the 95% level.



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

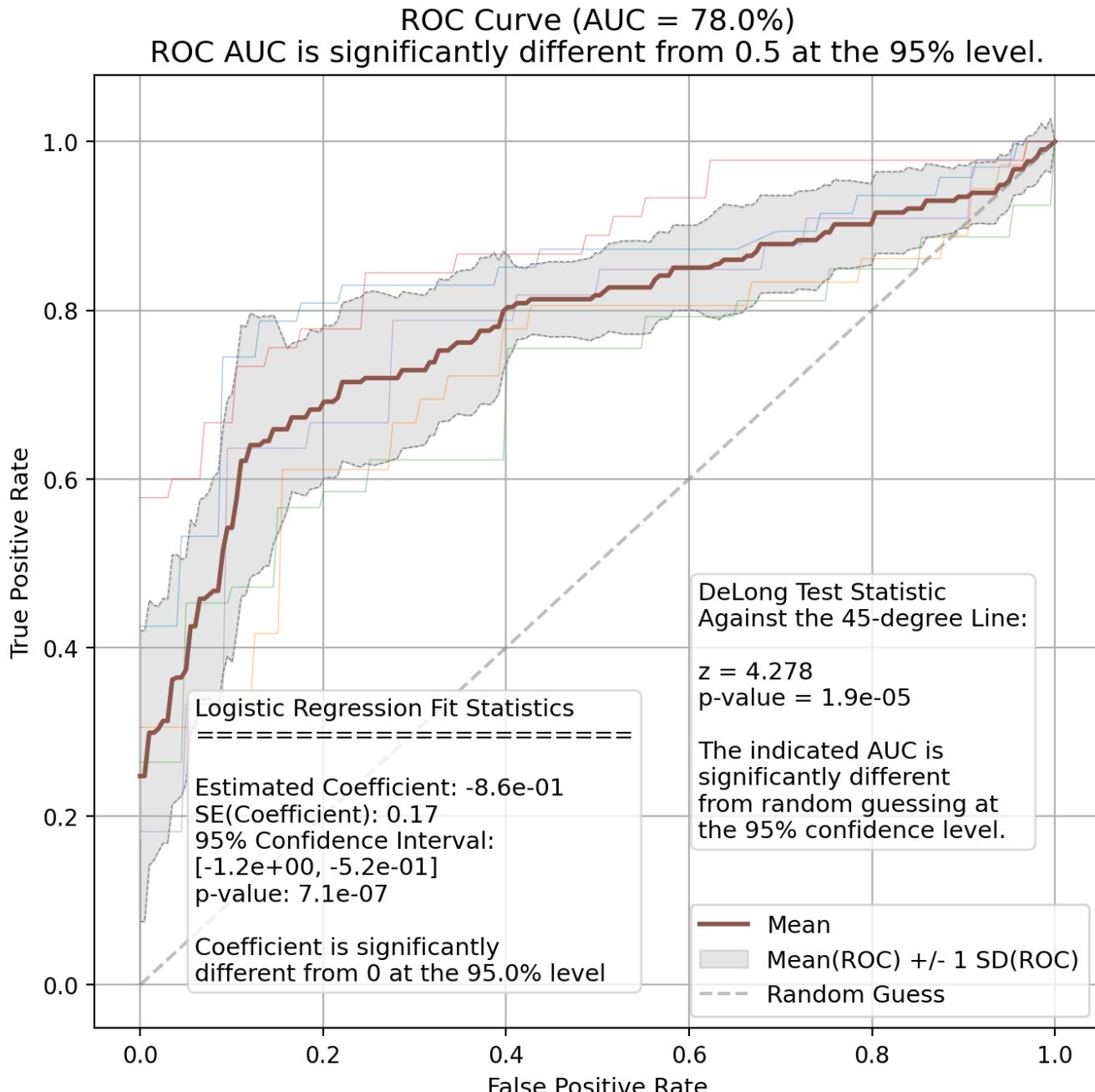
Concavity Error - Empirical CDF Plot



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

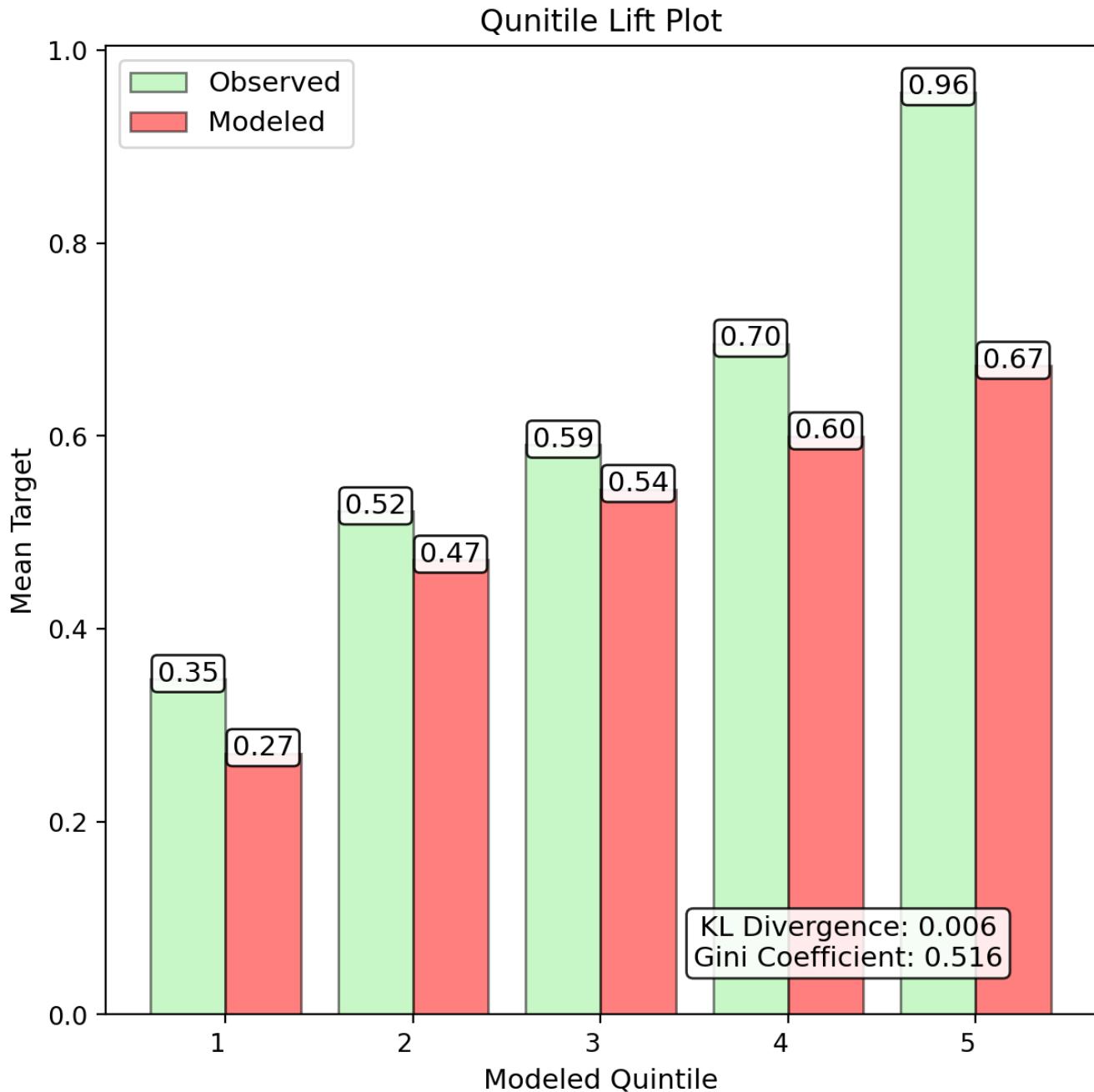


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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

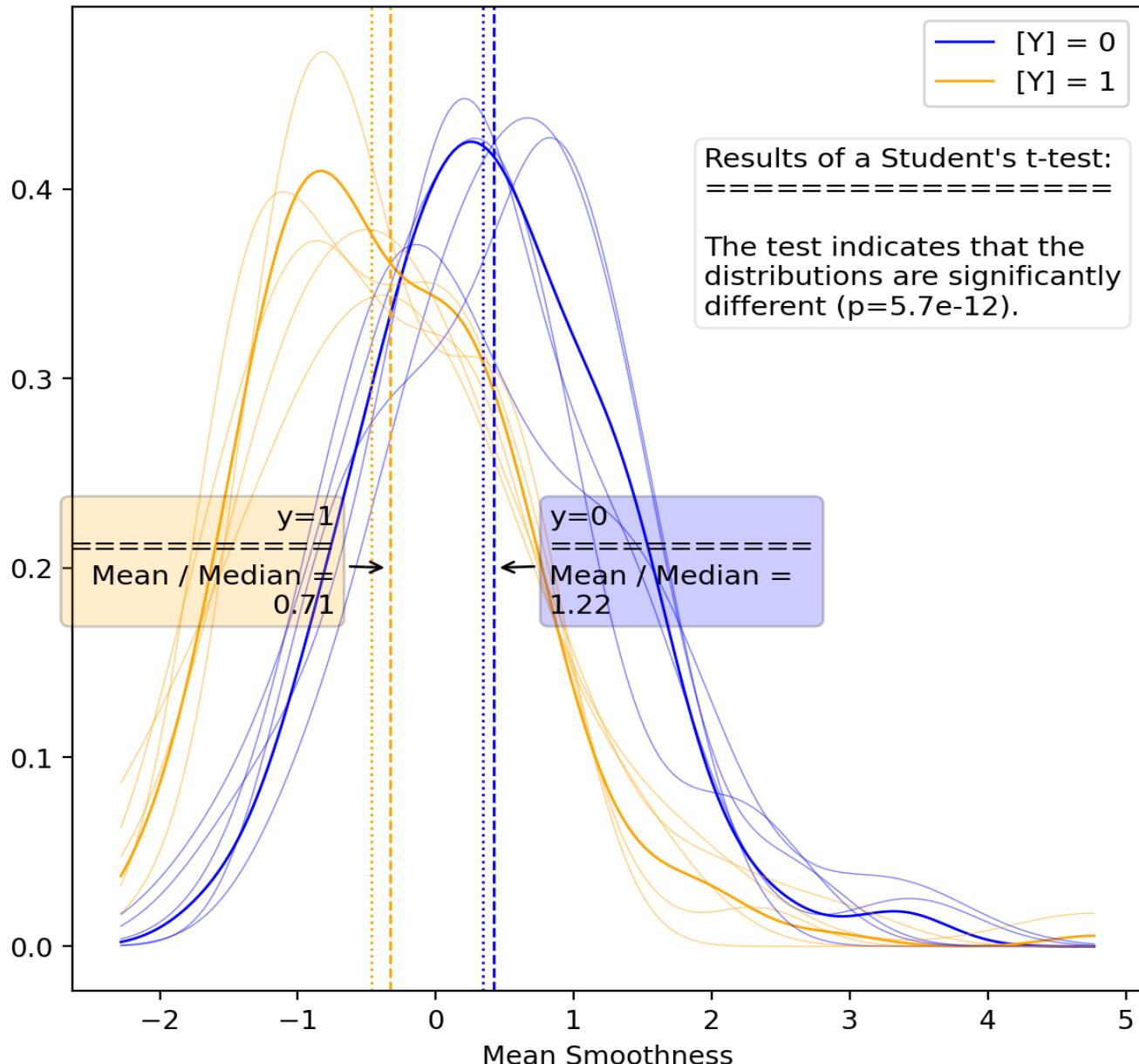
Mean Smoothness - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-0.758	-0.952	-0.803	-0.880	-0.797	-0.836	0.078
Fitted p-Value	2.5e-07	1.3e-09	5.8e-08	1.3e-08	2.6e-08	4.5e-10	1.1e-07
Fitted Std. Err.	0.147	0.157	0.148	0.155	0.143	0.134	0.006
Conf. Int. Lower	-1.05	-1.26	-1.09	-1.18	-1.08	-1.10	0.09
Conf. Int. Upper	-0.470	-0.644	-0.513	-0.577	-0.516	-0.573	0.068
Train Accuracy	64.9%	63.6%	66.4%	64.4%	67.1%	64.8%	1.4%
Val Accuracy	65.7%	68.1%	64.4%	66.2%	61.8%	67.5%	2.3%
Train AUC	65.4%	64.4%	67.2%	65.5%	67.8%	65.7%	1.4%
Val AUC	68.9%	68.4%	63.0%	66.7%	61.4%	68.9%	3.3%
Train F1	69.1%	69.0%	69.4%	68.6%	71.5%	68.9%	1.1%
Test F1	70.0%	66.7%	72.9%	69.9%	66.7%	70.9%	2.6%
Train Precision	75.7%	78.0%	76.7%	77.6%	79.2%	77.3%	1.3%
Val Precision	84.8%	73.3%	81.4%	76.3%	70.0%	80.4%	6.0%
Train Recall	63.5%	61.8%	63.4%	61.5%	65.2%	62.1%	1.5%
Val Recall	59.6%	61.1%	66.0%	64.4%	63.6%	63.4%	2.6%
Train MCC	30.0%	27.4%	33.7%	29.8%	34.4%	30.4%	2.9%
Val MCC	35.6%	37.2%	23.6%	32.6%	22.4%	36.6%	6.9%
Train Log-Loss	12.64	13.12	12.10	12.82	11.85	12.68	0.52
Val Log-Loss	12.36	11.49	12.84	12.18	13.76	11.70	0.84

Univariate Report

Mean Smoothness - Kernel Density Plot

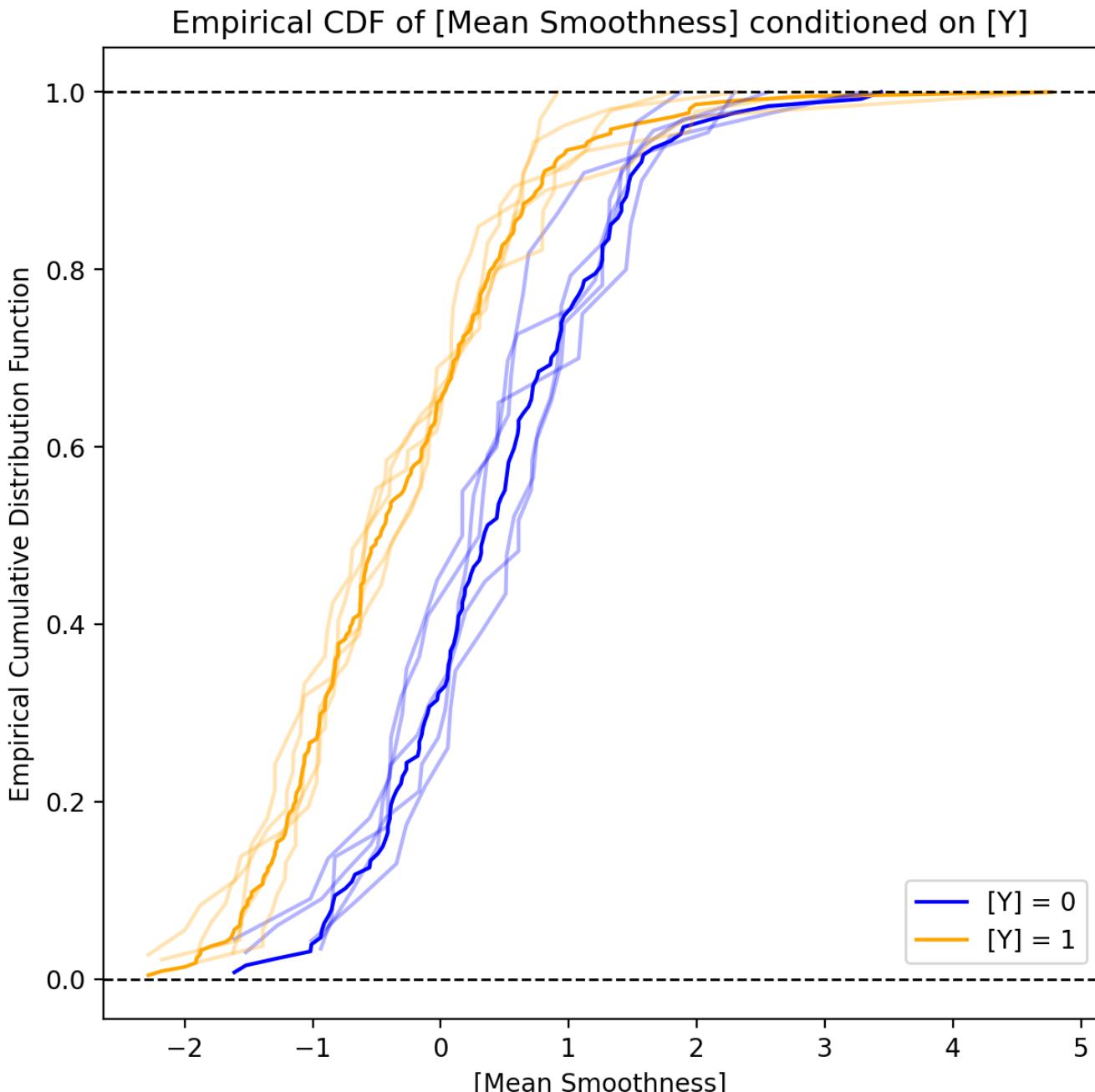
Kernel Density Plot of [Mean Smoothness] by [Y].
Distributions by level are significantly different at the 95% level.



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

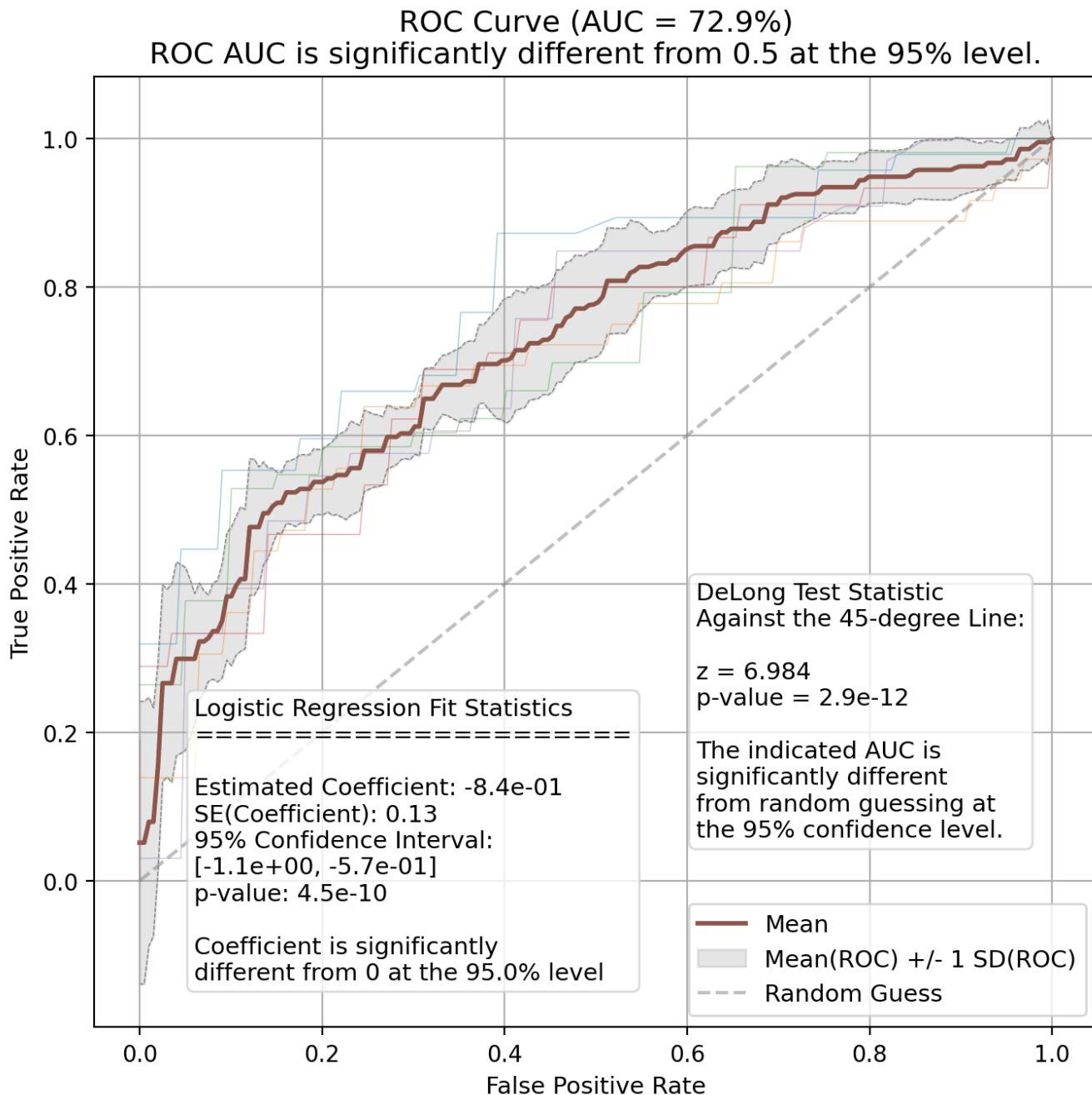
Mean Smoothness - Empirical CDF Plot



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 Smoothness - ROC Curve

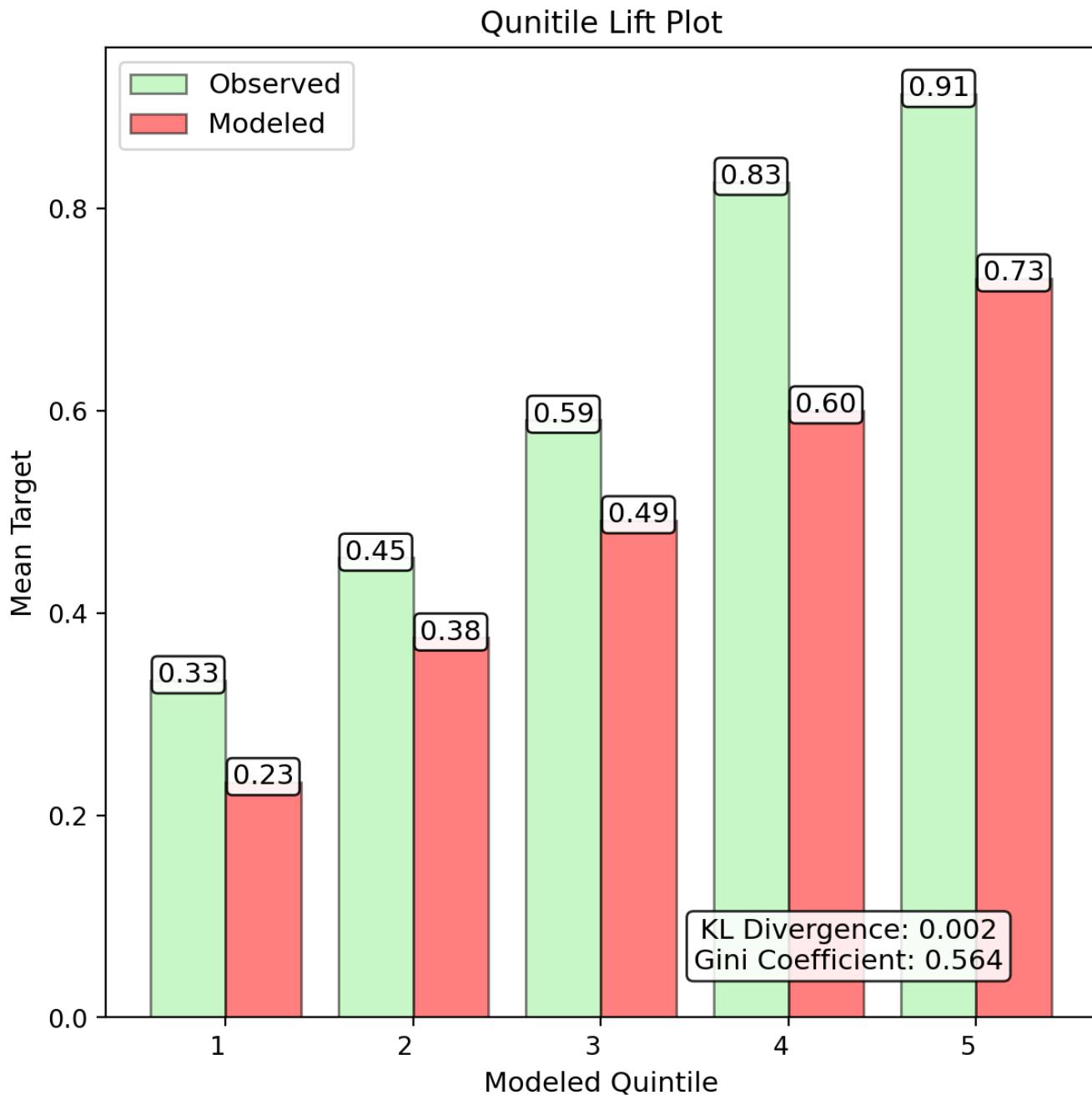


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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 Smoothness - 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.