Cancer Model Univariate Analysis Report

2024-02-22

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 providea 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.
Mean Area	88.6%	88.2%	94.4%	86.7%	91.2%	75.5%	87.4%
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%
log1p Mean Area	77.1%	86.5%	72.6%	78.2%	78.9%	55.4%	74.8%
log1p Mean Perimeter	76.2%	88.6%	67.2%	77.8%	76.5%	55.5%	73.6%
log1p Mean Radius	75.0%	84.8%	68.4%	76.1%	75.7%	51.8%	72.0%
Mean Texture	71.9%	81.0%	71.8%	72.0%	76.1%	42.8%	69.3%
log1p Mean Texture	65.3%	75.6%	56.4%	66.6%	64.6%	33.3%	60.3%

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 8 of the variables passed to the Cancer Model.

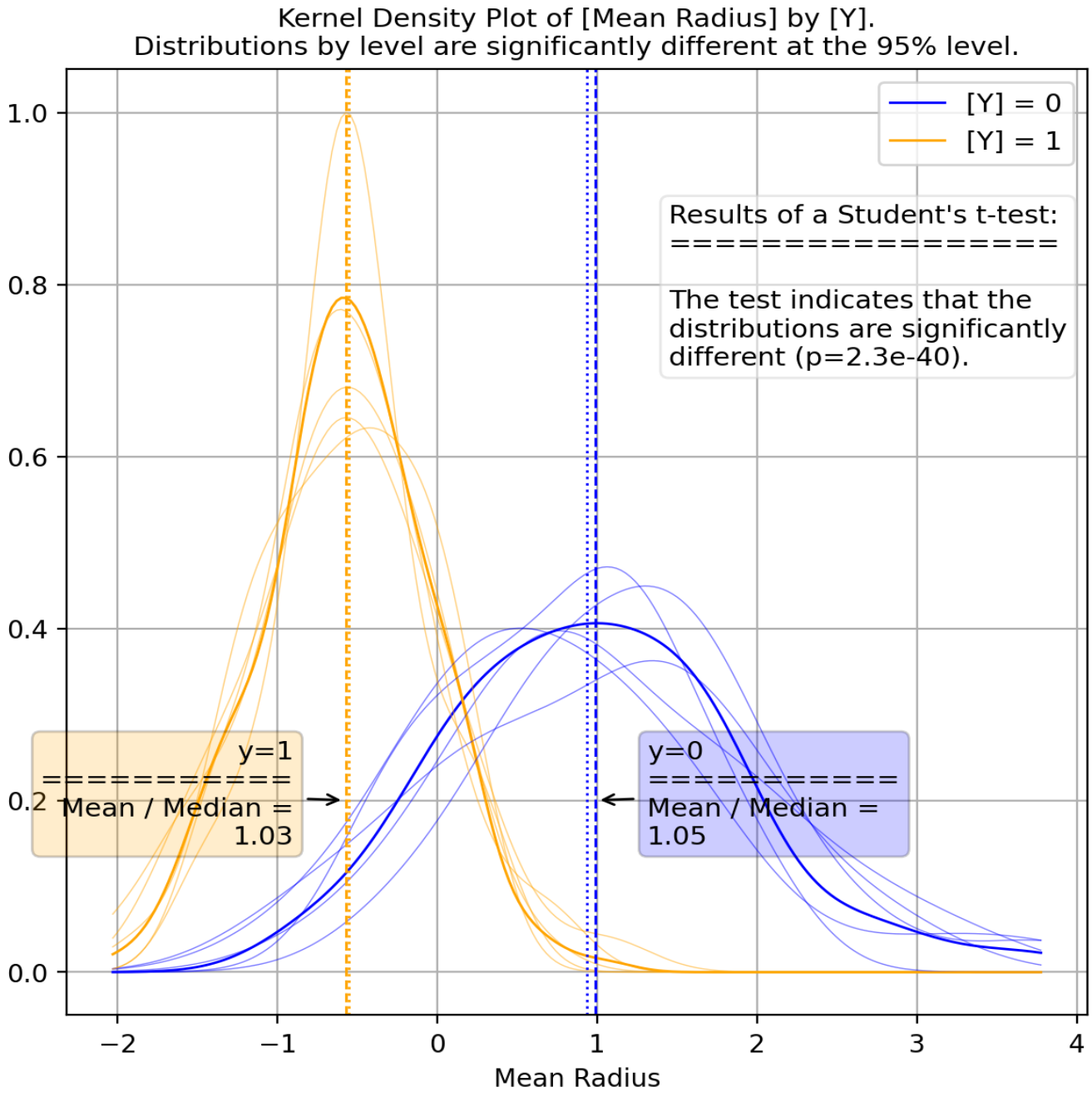
Univariate Report

Mean Radius - Results

	Coef	Pvalues	Se	Lower Ci	Upper Ci	Acc Test	Auc Test	F1 Test	Precision Test	Recall Test	Mcc Test
Fold-1	-3.98	3.4e-16	0.487	-4.93	-3.02	84.6%	84.7%	86.5%	88.9%	84.2%	68.8%
Fold-2	-3.57	3.5e-16	0.438	-4.43	-2.71	86.2%	85.8%	88.6%	89.7%	87.5%	71.0%
Fold-3	-3.52	8.6e-17	0.423	-4.35	-2.69	86.6%	86.8%	89.2%	92.5%	86.0%	71.9%
Fold-4	-3.82	3.1e-16	0.467	-4.73	-2.90	85.9%	85.6%	88.6%	90.7%	86.7%	70.3%
Fold-5	-3.59	5.7e-16	0.444	-4.47	-2.72	89.0%	89.8%	91.3%	95.5%	87.5%	77.1%
mean	-3.69	5.6e-20	0.403	-4.48	-2.90	86.8%	86.2%	89.4%	90.0%	88.7%	72.1%
std	0.19	1.7e-16	0.025	0.24	0.14	1.6%	1.9%	1.7%	2.6%	1.4%	3.2%

Univariate Report

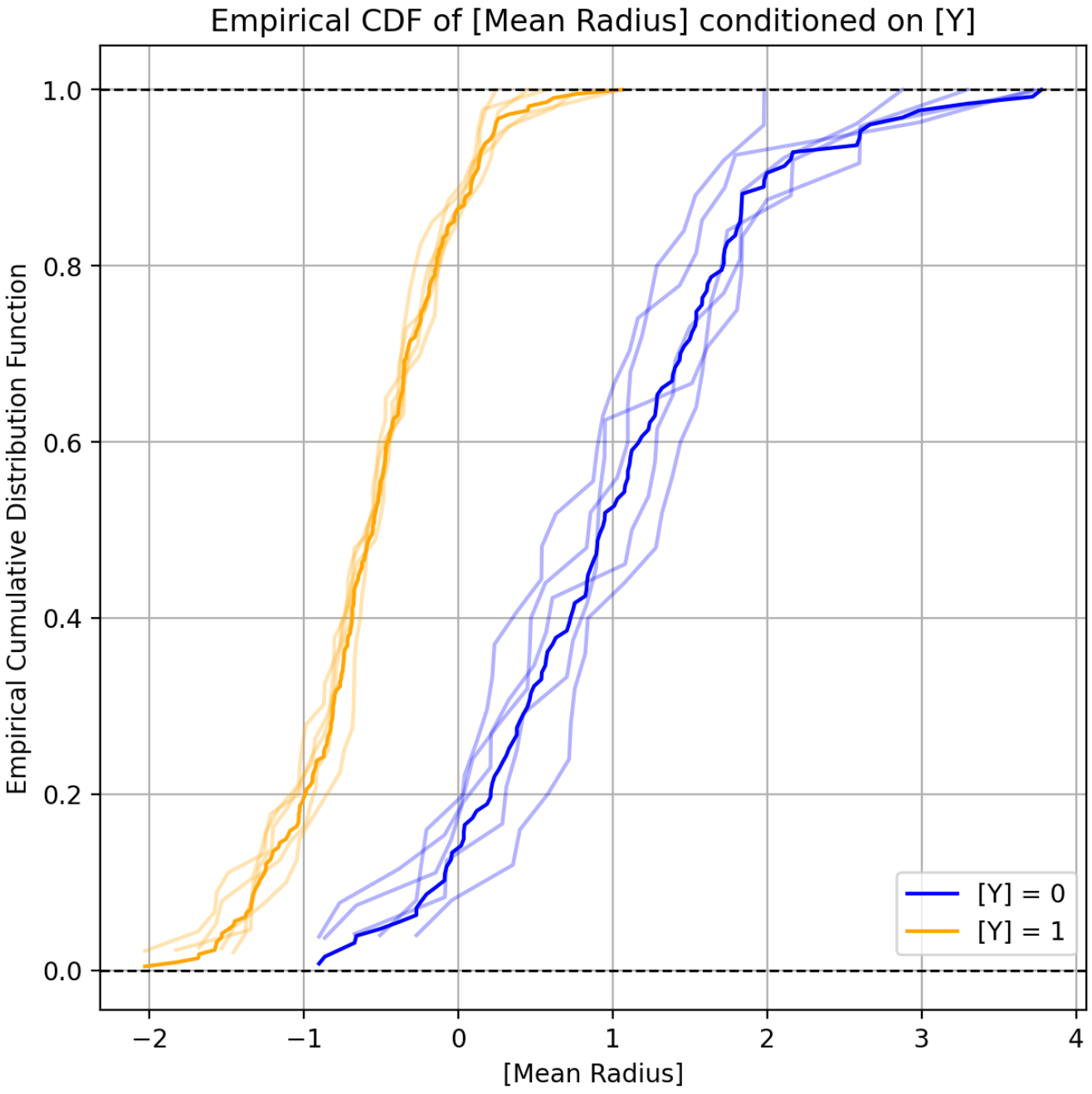
Mean Radius - Kernel Density Plot



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

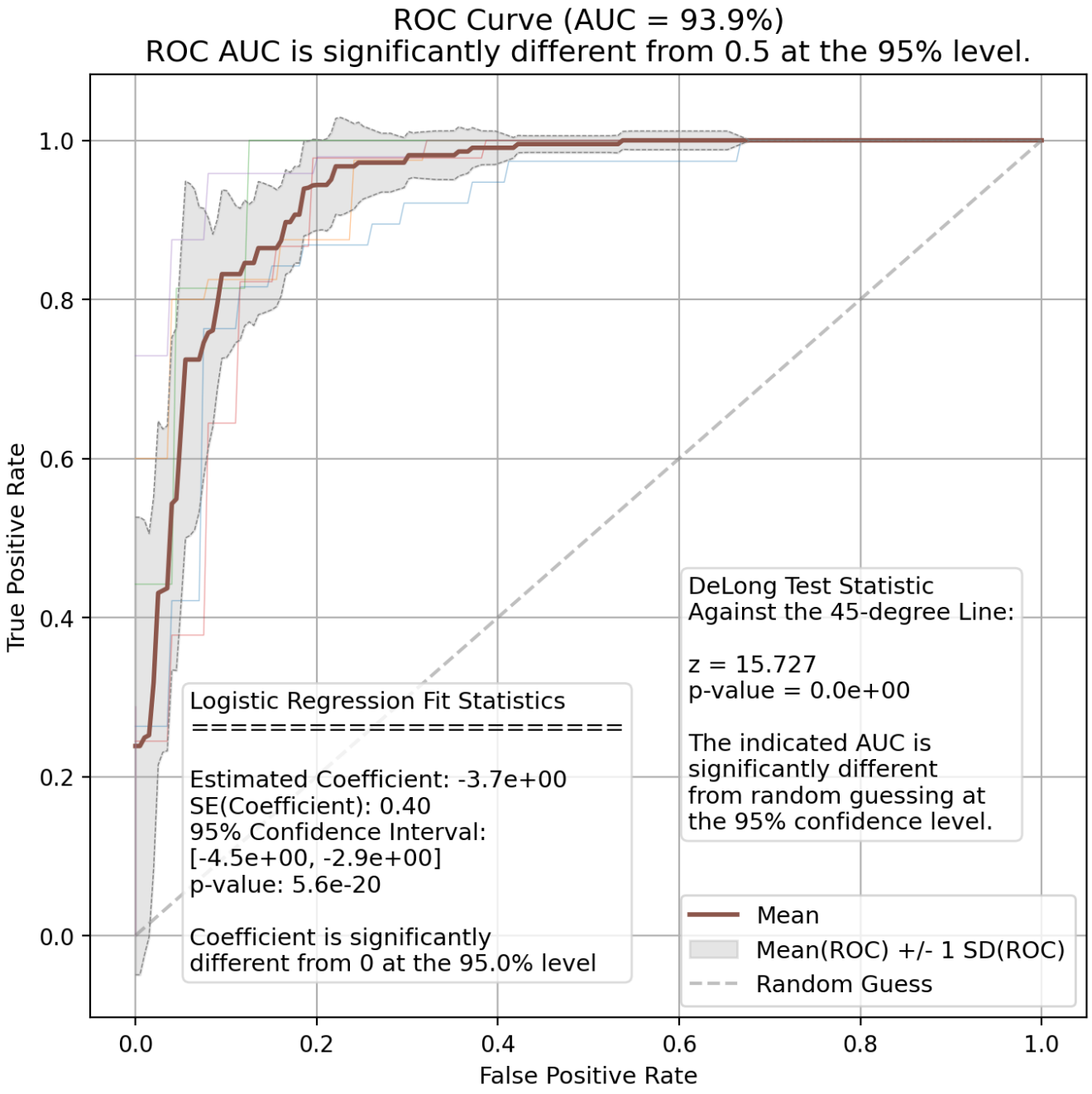
Mean Radius - Empirical CDF Plot



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

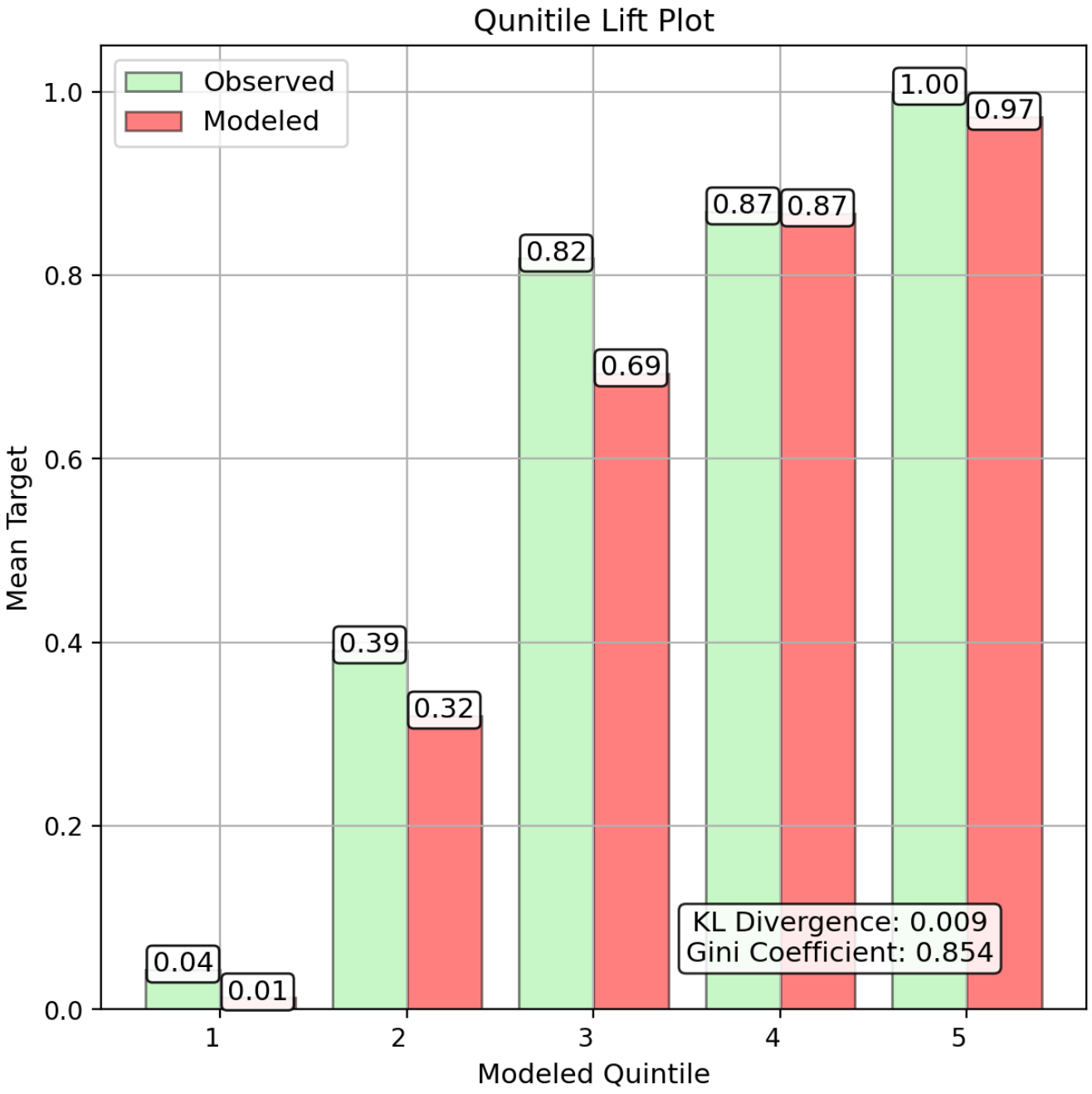


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

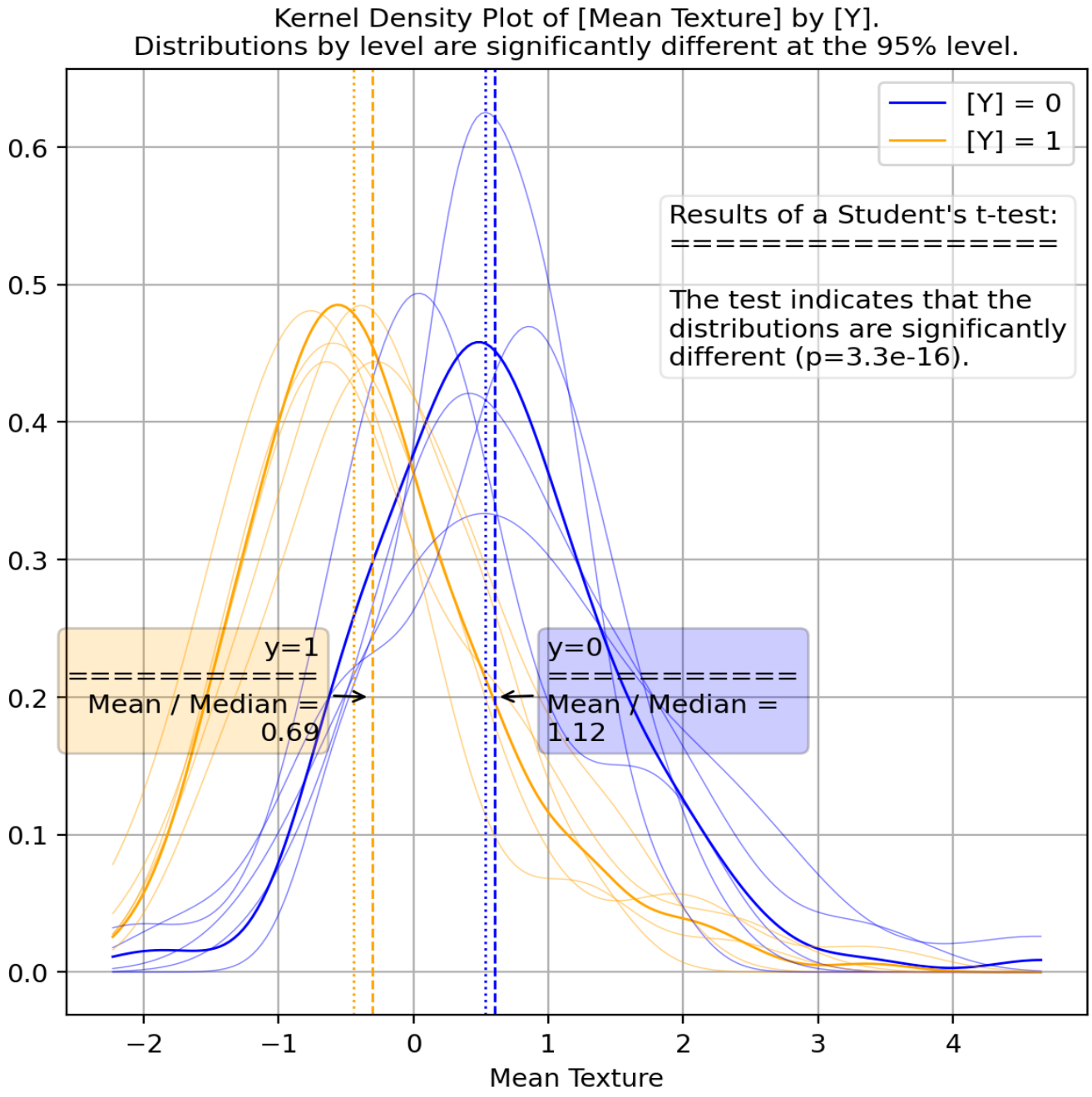
Univariate Report

Mean Texture - Results

	Coef	Pvalues	Se	Lower Ci	Upper Ci	Acc Test	Auc Test	F1 Test	Precision Test	Recall Test	Mcc Test
Fold-1	-0.97	7.8e-10	0.158	-1.28	-0.661	70.8%	70.7%	74.0%	77.1%	71.1%	40.9%
Fold-2	-1.13	2.2e-11	0.169	-1.46	-0.800	63.1%	61.8%	69.2%	71.1%	67.5%	23.2%
Fold-3	-1.06	1.4e-10	0.165	-1.38	-0.736	67.2%	70.7%	69.4%	86.2%	58.1%	40.1%
Fold-4	-0.86	2.5e-08	0.154	-1.16	-0.556	84.5%	83.7%	87.6%	88.6%	86.7%	66.9%
Fold-5	-1.04	3.9e-10	0.167	-1.37	-0.716	72.6%	74.4%	76.7%	86.8%	68.8%	46.3%
mean	-1.01	3.5e-12	0.145	-1.29	-0.725	71.9%	72.0%	76.1%	81.0%	71.8%	42.8%
std	0.10	1.1e-08	0.006	0.12	0.092	8.1%	7.9%	7.5%	7.6%	10.3%	15.7%

Univariate Report

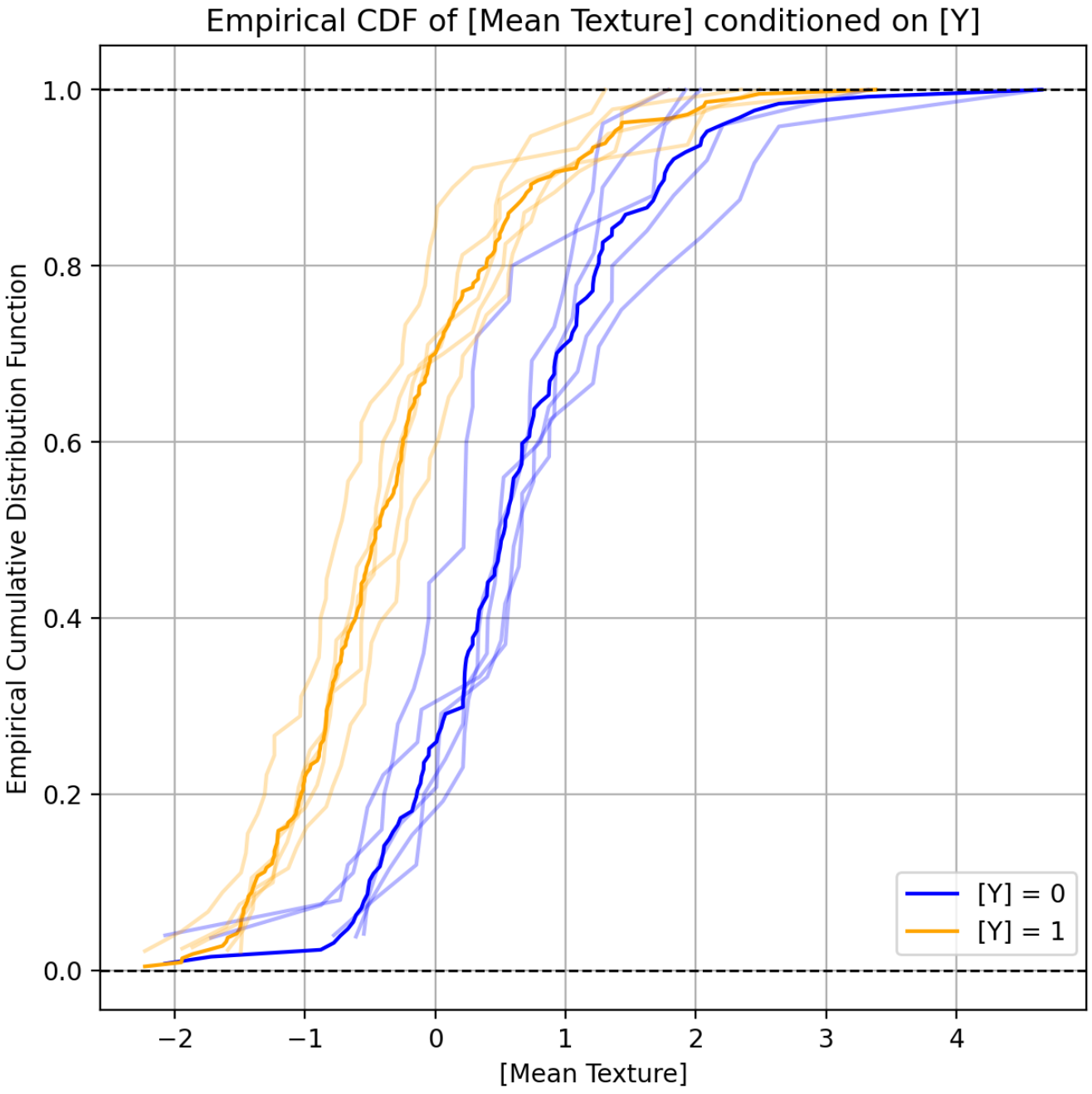
Mean Texture - Kernel Density Plot



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

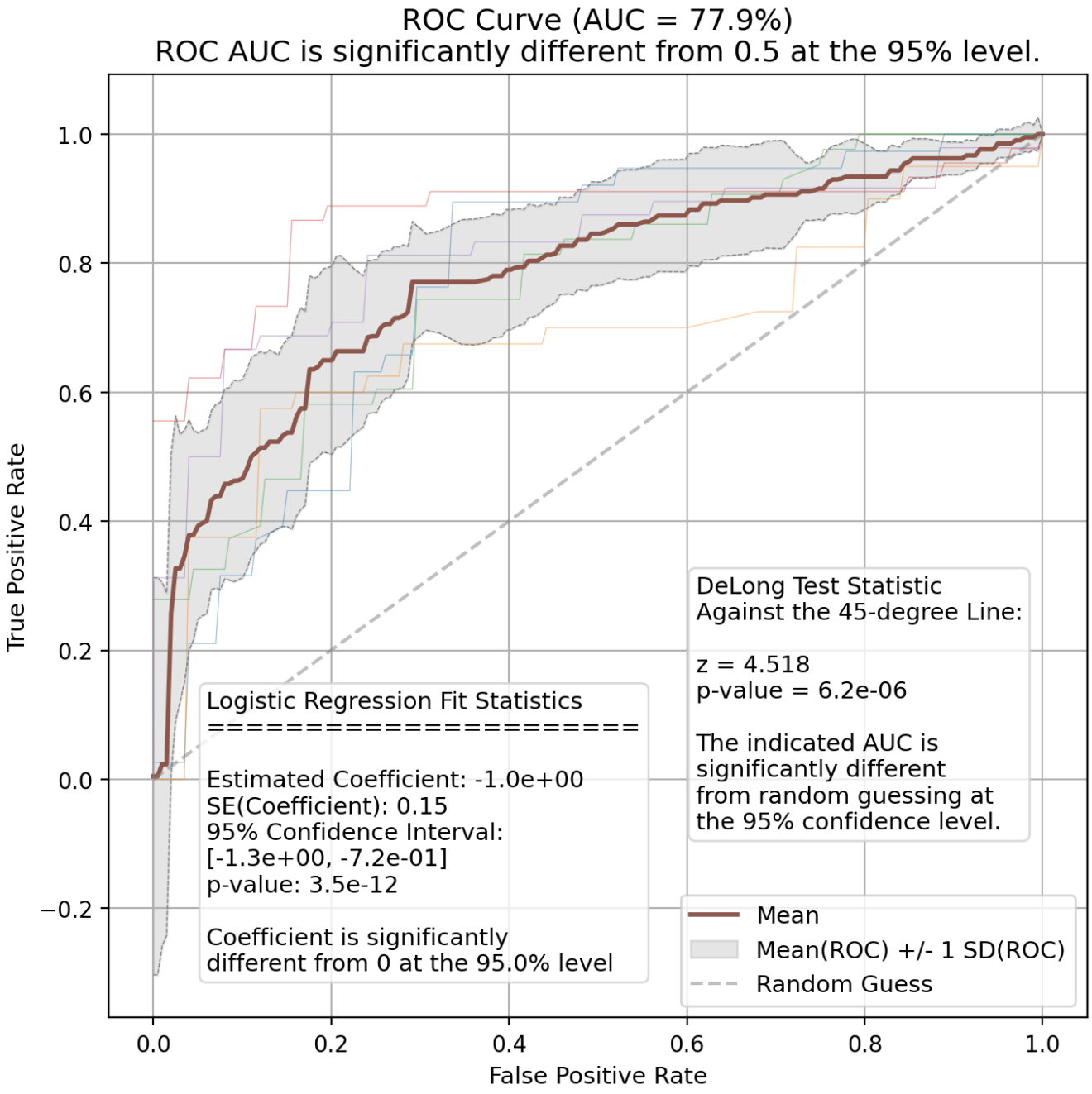
Mean Texture - Empirical CDF Plot



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

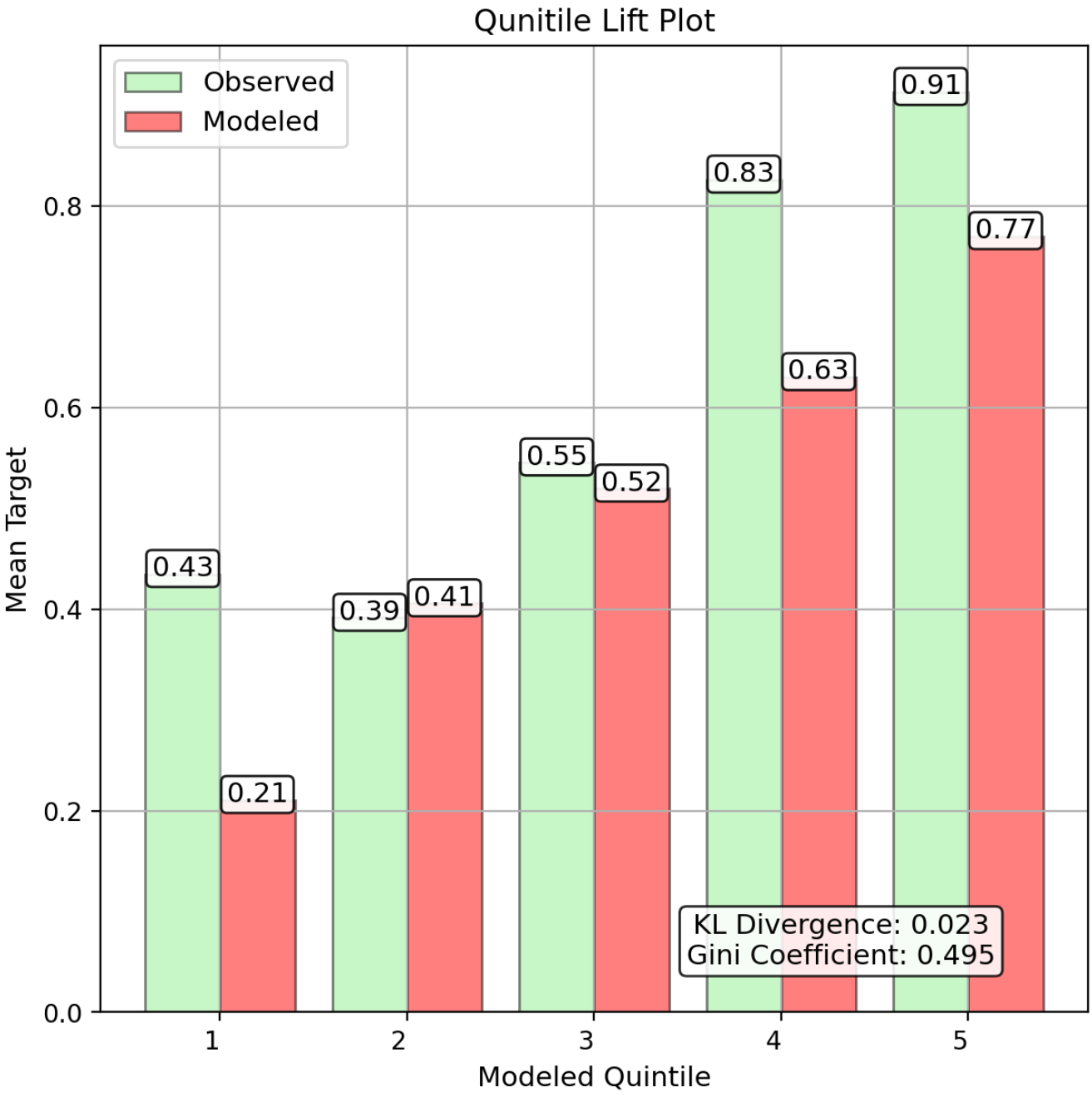


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

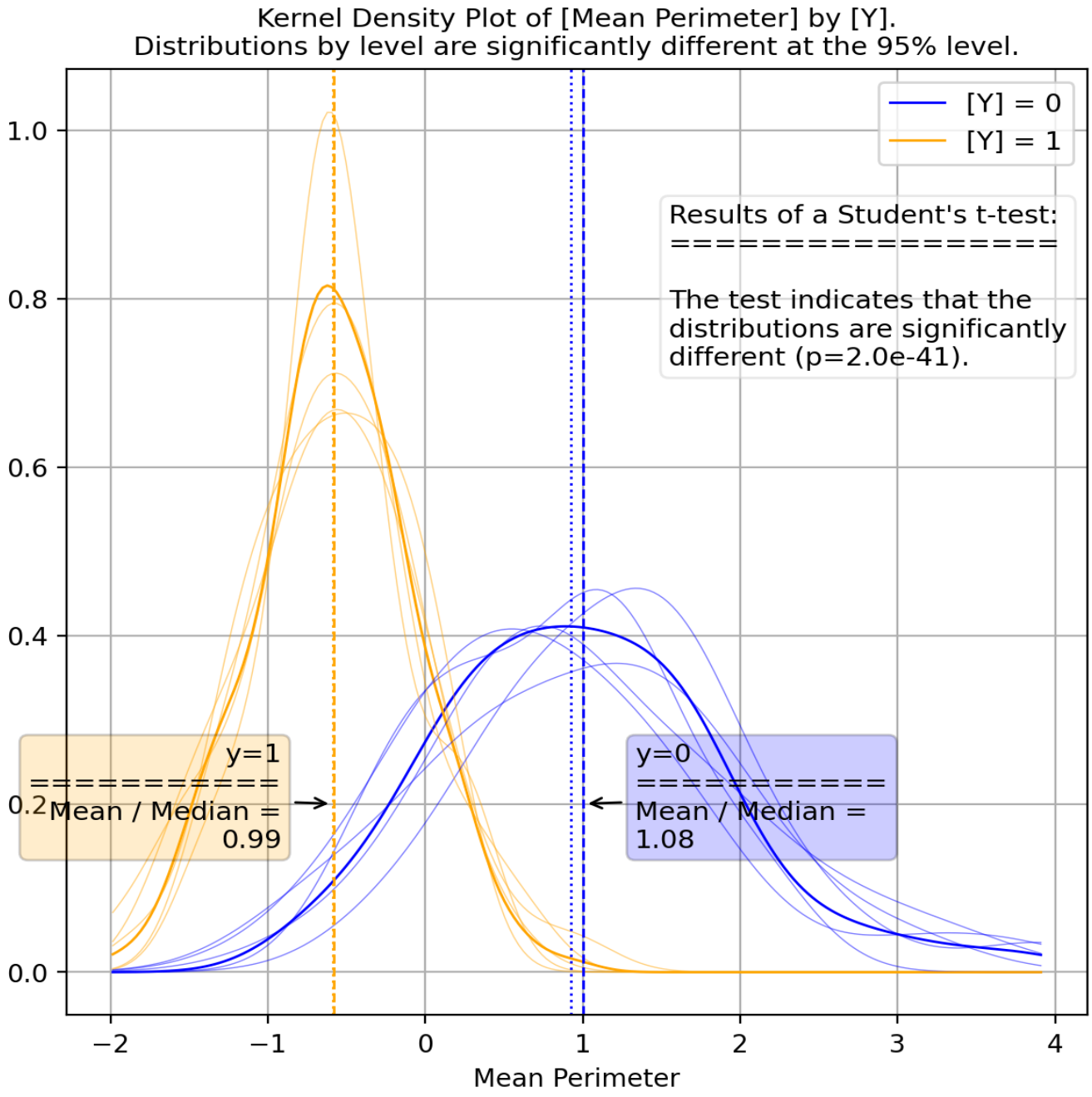
Univariate Report

Mean Perimeter - Results

	Coef	Pvalues	Se	Lower Ci	Upper Ci	Acc Test	Auc Test	F1 Test	Precision Test	Recall Test	Mcc Test
Fold-1	-4.24	3.4e-16	0.520	-5.26	-3.22	86.2%	86.0%	88.0%	89.2%	86.8%	71.7%
Fold-2	-3.92	3.7e-16	0.481	-4.87	-2.98	86.2%	85.8%	88.6%	89.7%	87.5%	71.0%
Fold-3	-3.82	8.9e-17	0.459	-4.71	-2.92	88.1%	88.9%	90.2%	94.9%	86.0%	75.5%
Fold-4	-4.11	4.3e-16	0.505	-5.10	-3.12	87.3%	86.8%	89.9%	90.9%	88.9%	72.9%
Fold-5	-3.93	6.8e-16	0.487	-4.89	-2.98	89.0%	89.8%	91.3%	95.5%	87.5%	77.1%
mean	-4.00	6.6e-20	0.438	-4.86	-3.14	86.8%	86.2%	89.4%	90.0%	88.7%	72.1%
std	0.17	2.1e-16	0.023	0.21	0.12	1.2%	1.8%	1.3%	2.9%	1.0%	2.6%

Univariate Report

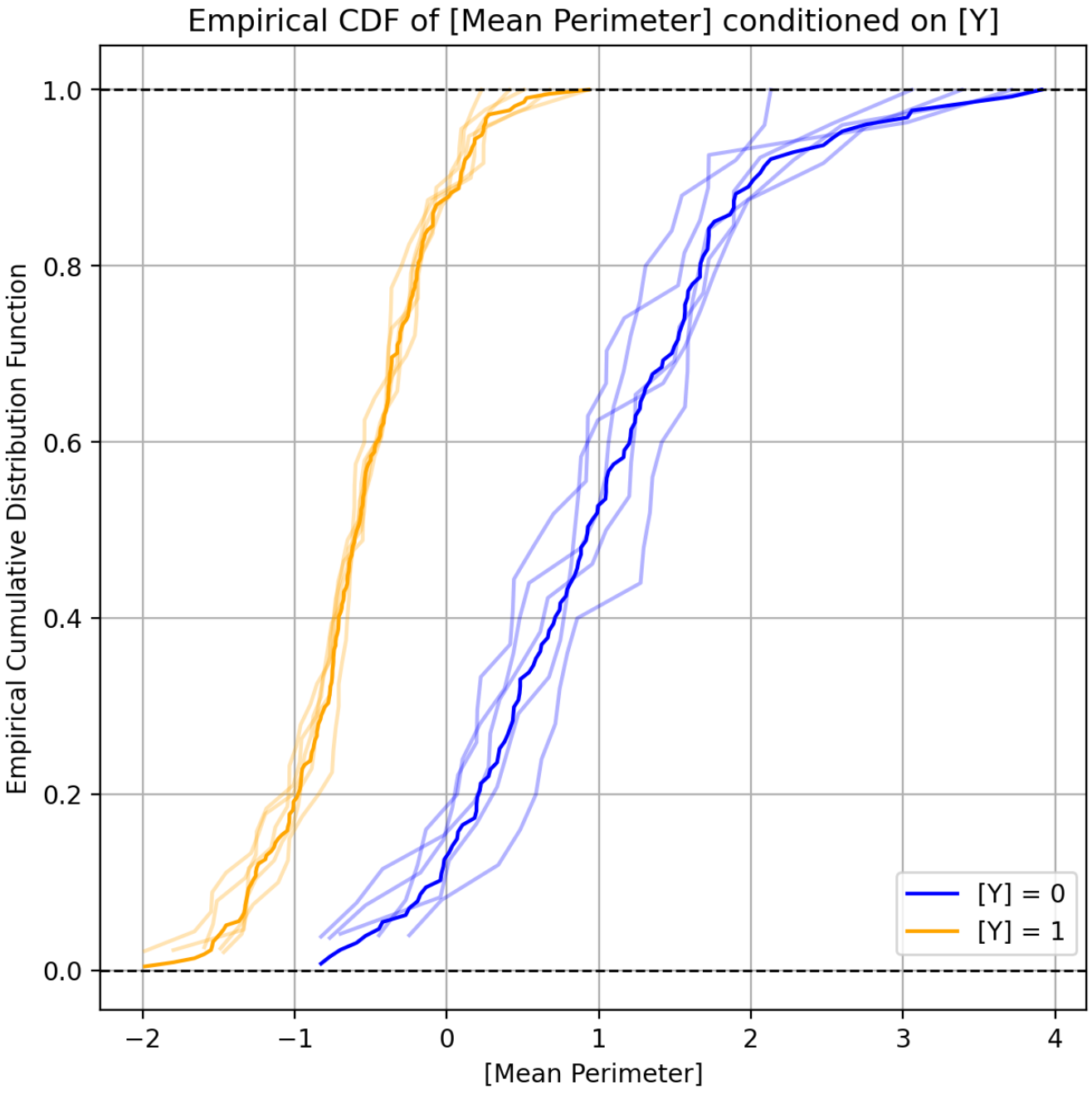
Mean Perimeter - Kernel Density Plot



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

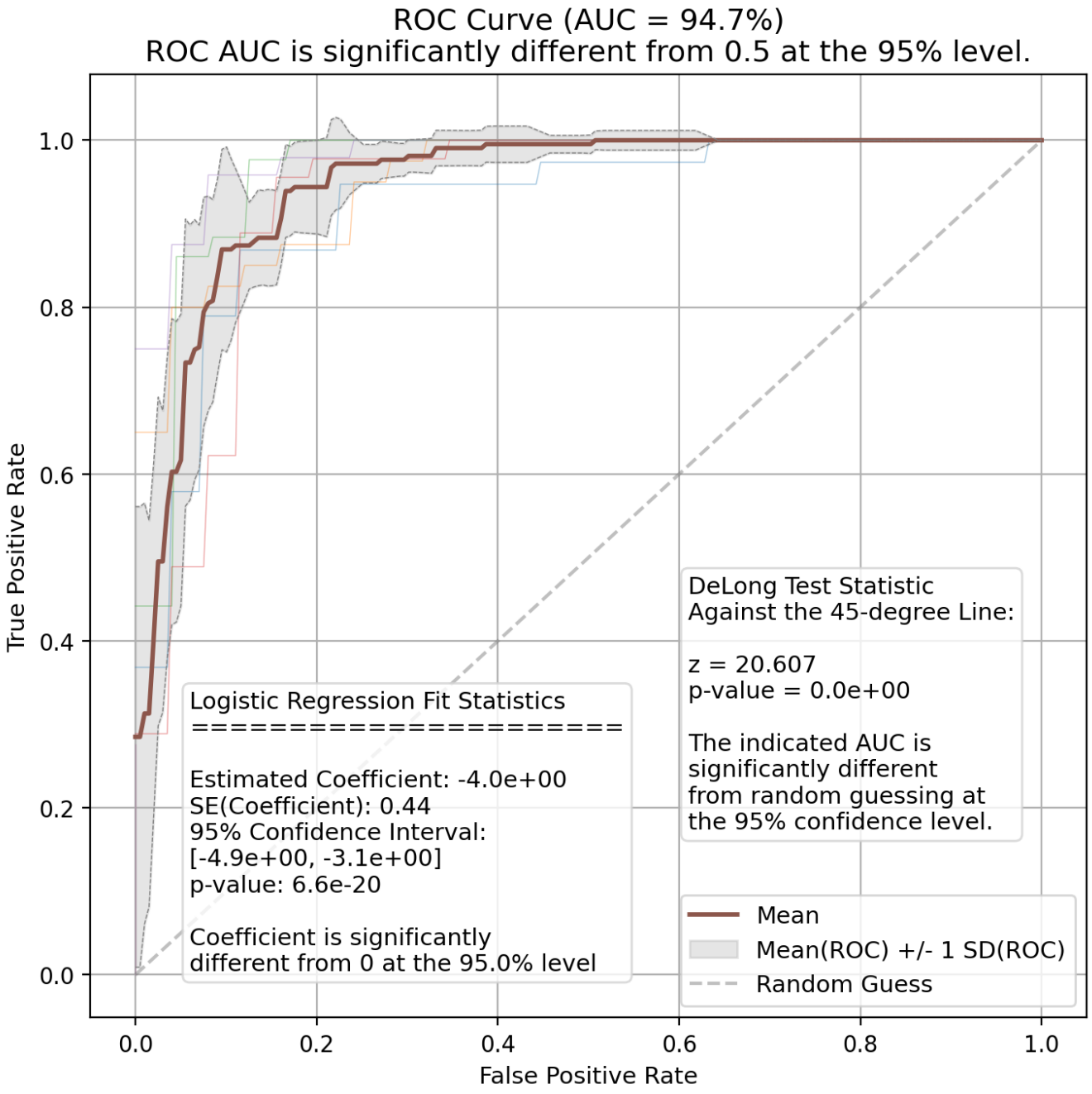
Mean Perimeter - Empirical CDF Plot



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

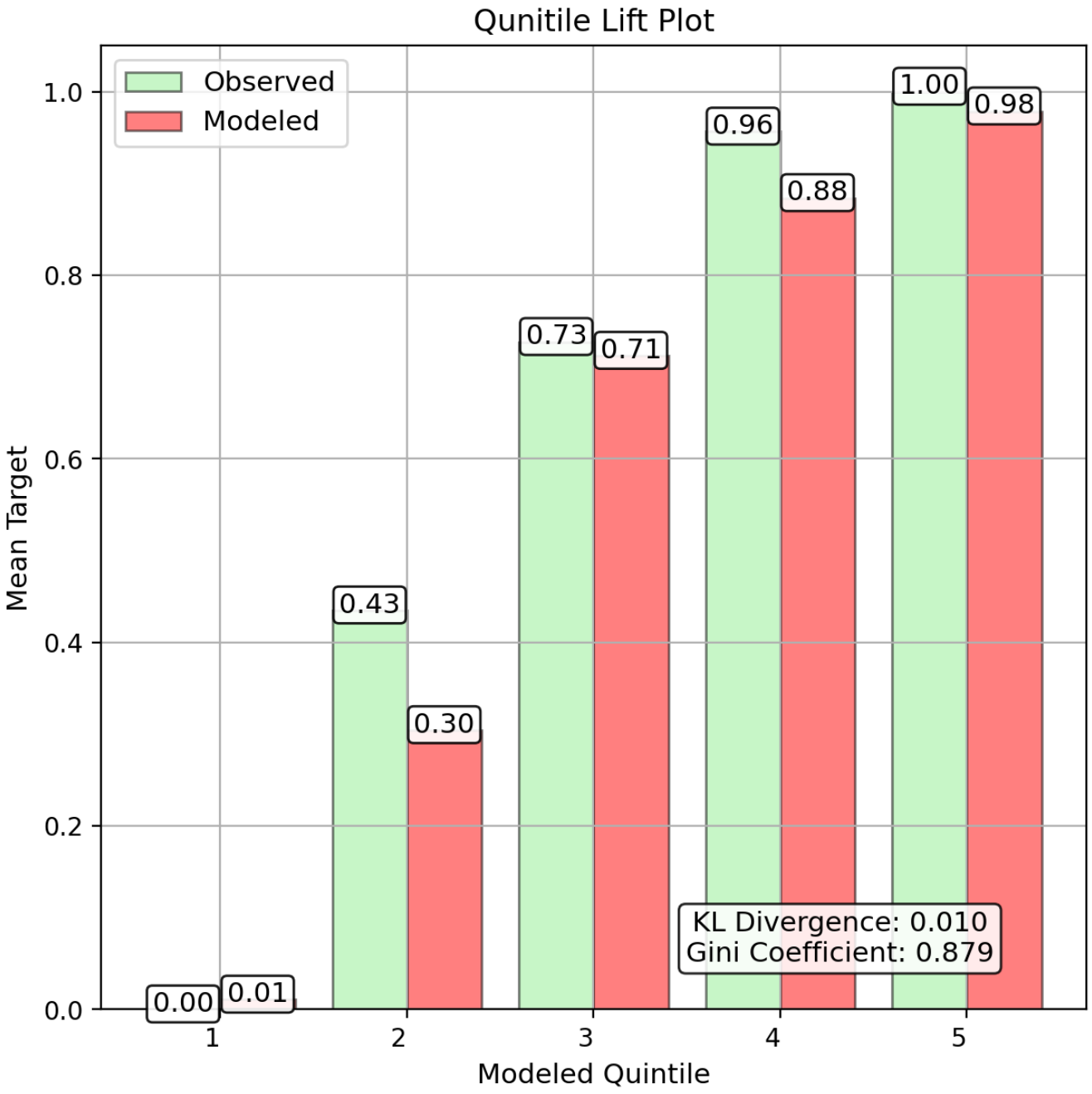


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

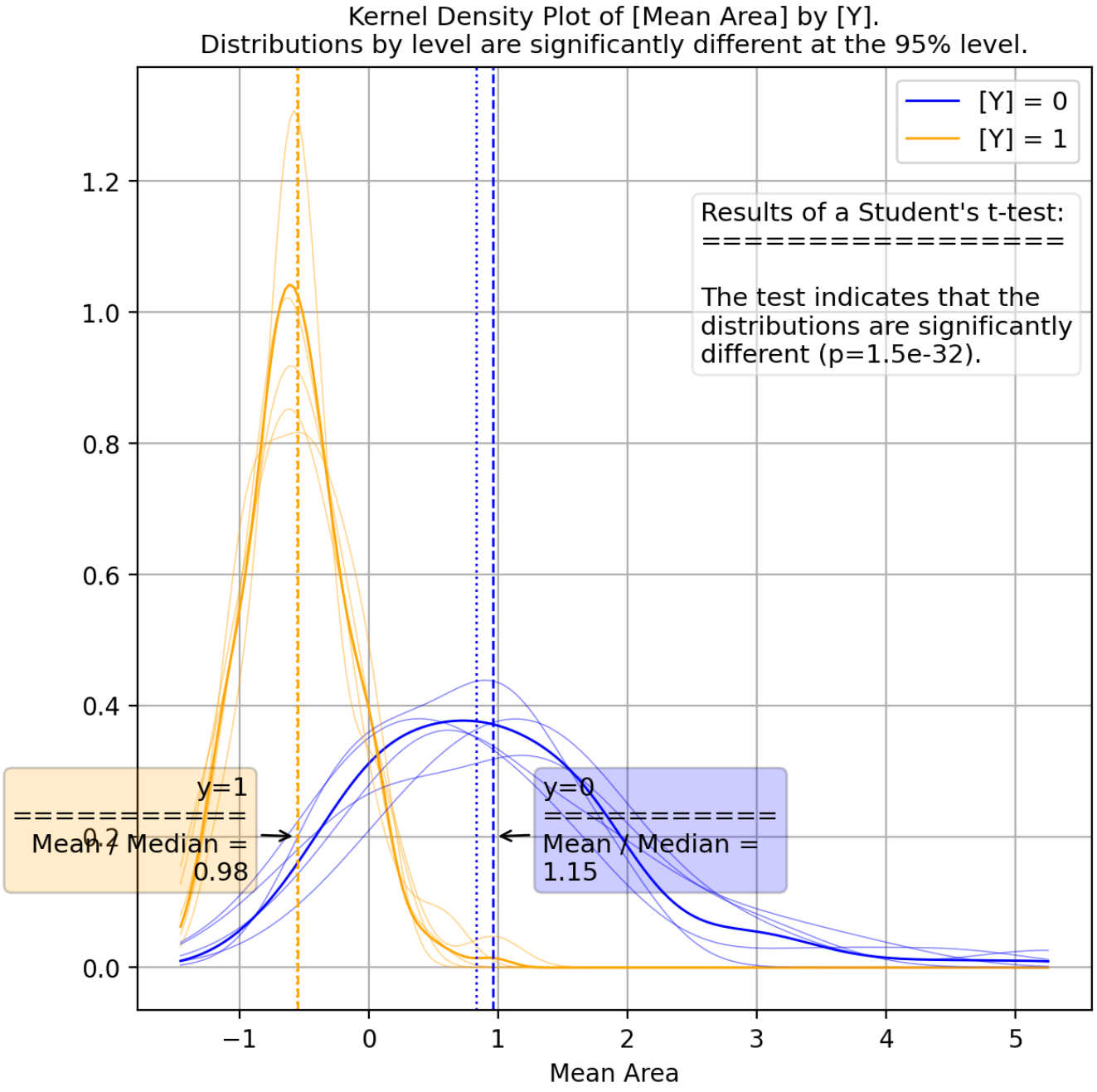
Univariate Report

Mean Area - Results

	Coef	Pvalues	Se	Lower Ci	Upper Ci	Acc Test	Auc Test	F1 Test	Precision Test	Recall Test	Mcc Test
Fold-1	-4.61	1.5e-16	0.558	-5.71	-3.52	81.5%	80.5%	84.6%	82.5%	86.8%	61.7%
Fold-2	-4.15	1.2e-16	0.501	-5.13	-3.17	86.2%	84.2%	89.2%	86.0%	92.5%	70.4%
Fold-3	-4.01	2.6e-17	0.473	-4.93	-3.08	89.6%	89.1%	91.8%	92.9%	90.7%	77.5%
Fold-4	-4.36	1.1e-16	0.525	-5.39	-3.33	91.5%	89.3%	93.6%	89.8%	97.8%	81.8%
Fold-5	-4.10	1.7e-16	0.497	-5.07	-3.12	91.8%	91.8%	93.6%	95.7%	91.7%	82.2%
mean	-4.23	1.4e-20	0.455	-5.13	-3.34	88.6%	86.7%	91.2%	88.2%	94.4%	75.5%
std	0.24	5.6e-17	0.032	0.31	0.18	4.3%	4.6%	3.8%	5.2%	3.9%	8.7%

Univariate Report

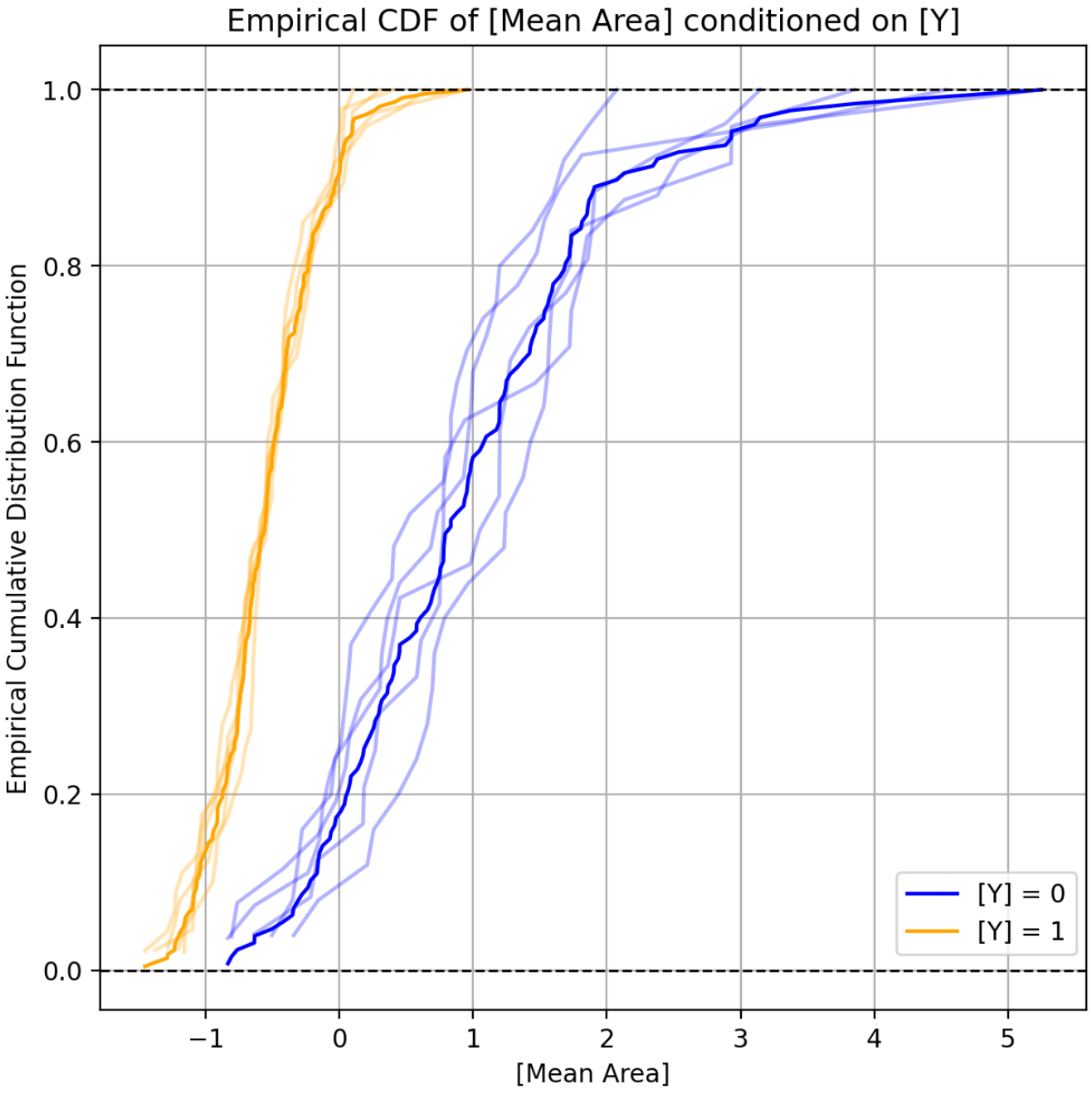
Mean Area - Kernel Density Plot



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

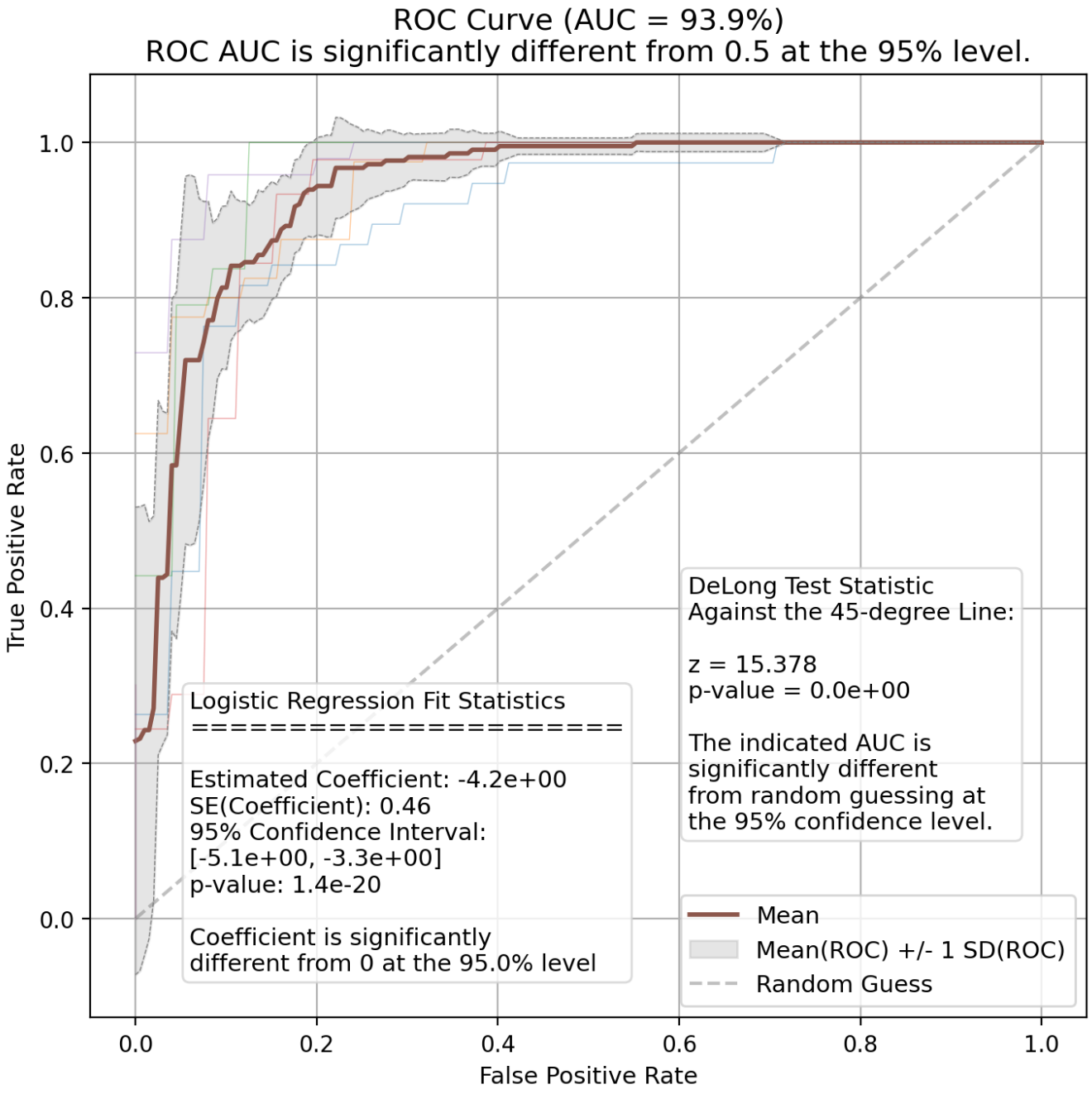
Mean Area - Empirical CDF Plot



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

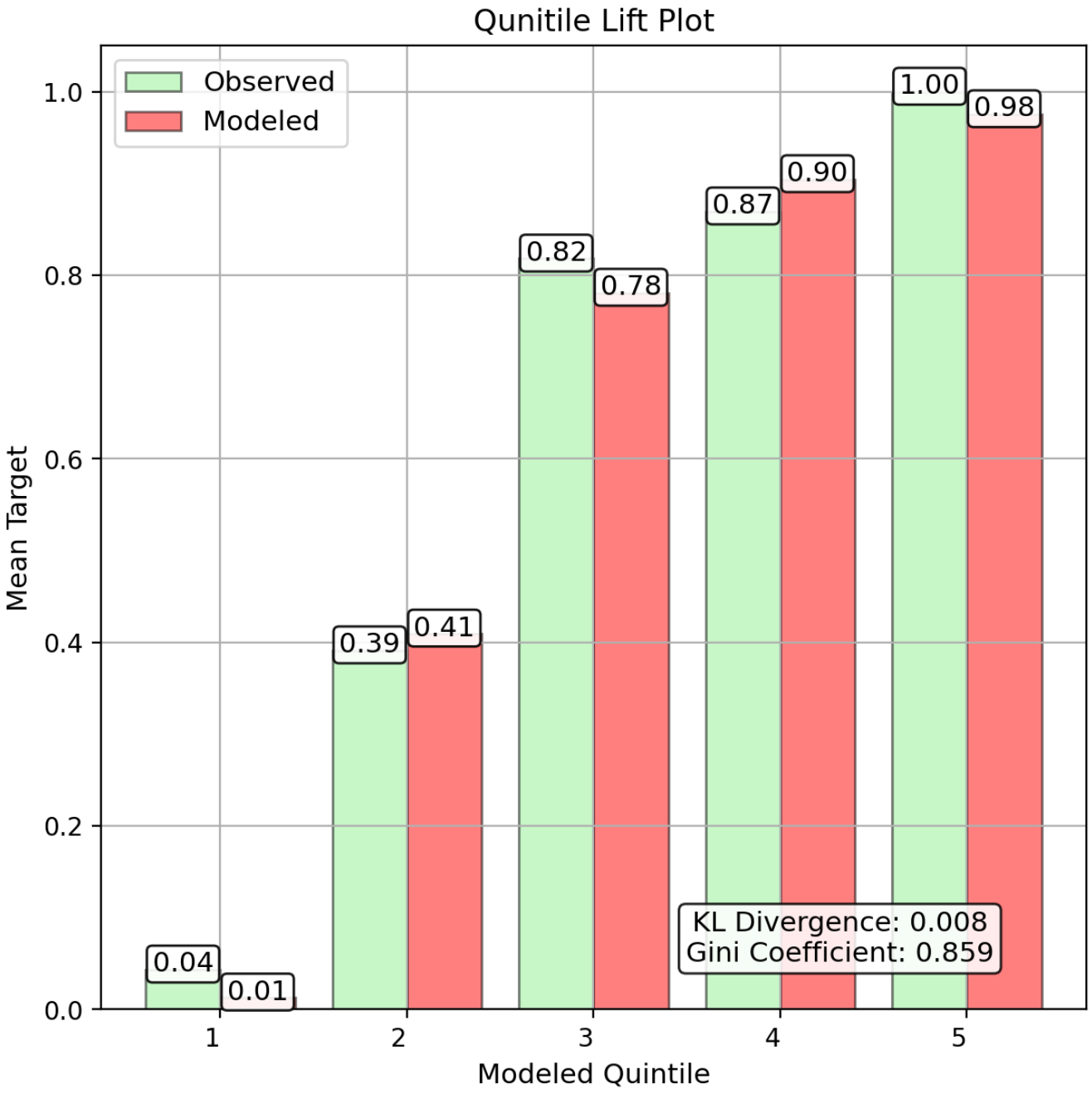


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

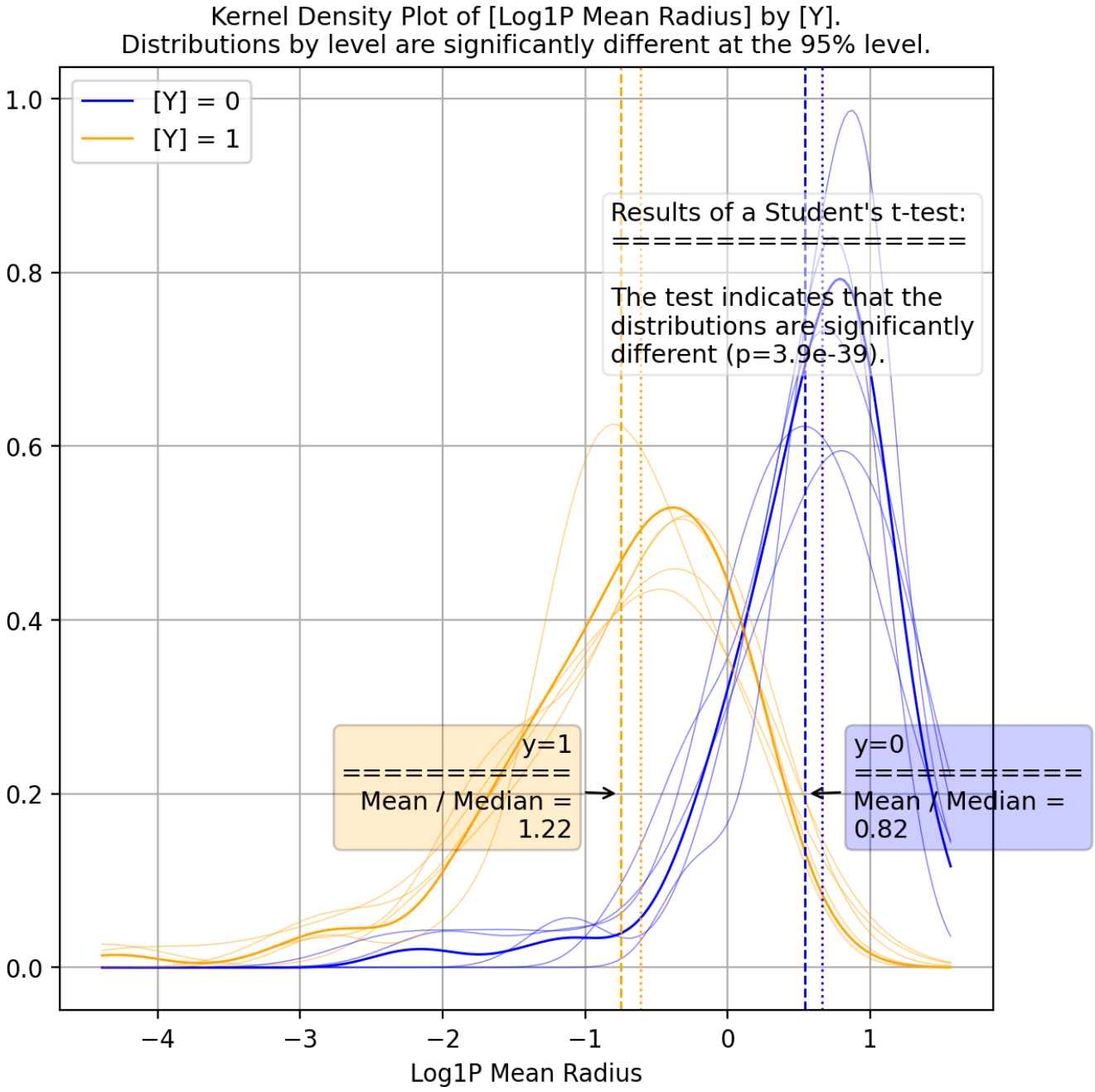
Univariate Report

Log1P Mean Radius - Results

	Coef	Pvalues	Se	Lower Ci	Upper Ci	Acc Test	Auc Test	F1 Test	Precision Test	Recall Test	Mcc Test
Fold-1	-3.26	2.3e-15	0.412	-4.07	-2.46	80.7%	81.3%	79.2%	91.3%	70.0%	63.7%
Fold-2	-2.81	9.7e-15	0.364	-3.53	-2.10	81.4%	81.7%	83.1%	87.1%	79.4%	62.8%
Fold-3	-2.84	1.5e-15	0.356	-3.53	-2.14	76.8%	79.2%	75.5%	95.2%	62.5%	59.6%
Fold-4	-3.32	4.9e-15	0.423	-4.15	-2.49	80.3%	81.4%	81.2%	89.7%	74.3%	62.1%
Fold-5	-2.75	1.3e-14	0.358	-3.46	-2.05	83.3%	85.8%	84.9%	96.9%	75.6%	69.5%
mean	-2.98	1.7e-18	0.340	-3.65	-2.32	75.0%	76.1%	75.7%	84.8%	68.4%	51.8%
std	0.27	5.0e-15	0.033	0.33	0.21	2.4%	2.4%	3.6%	4.0%	6.5%	3.7%

Univariate Report

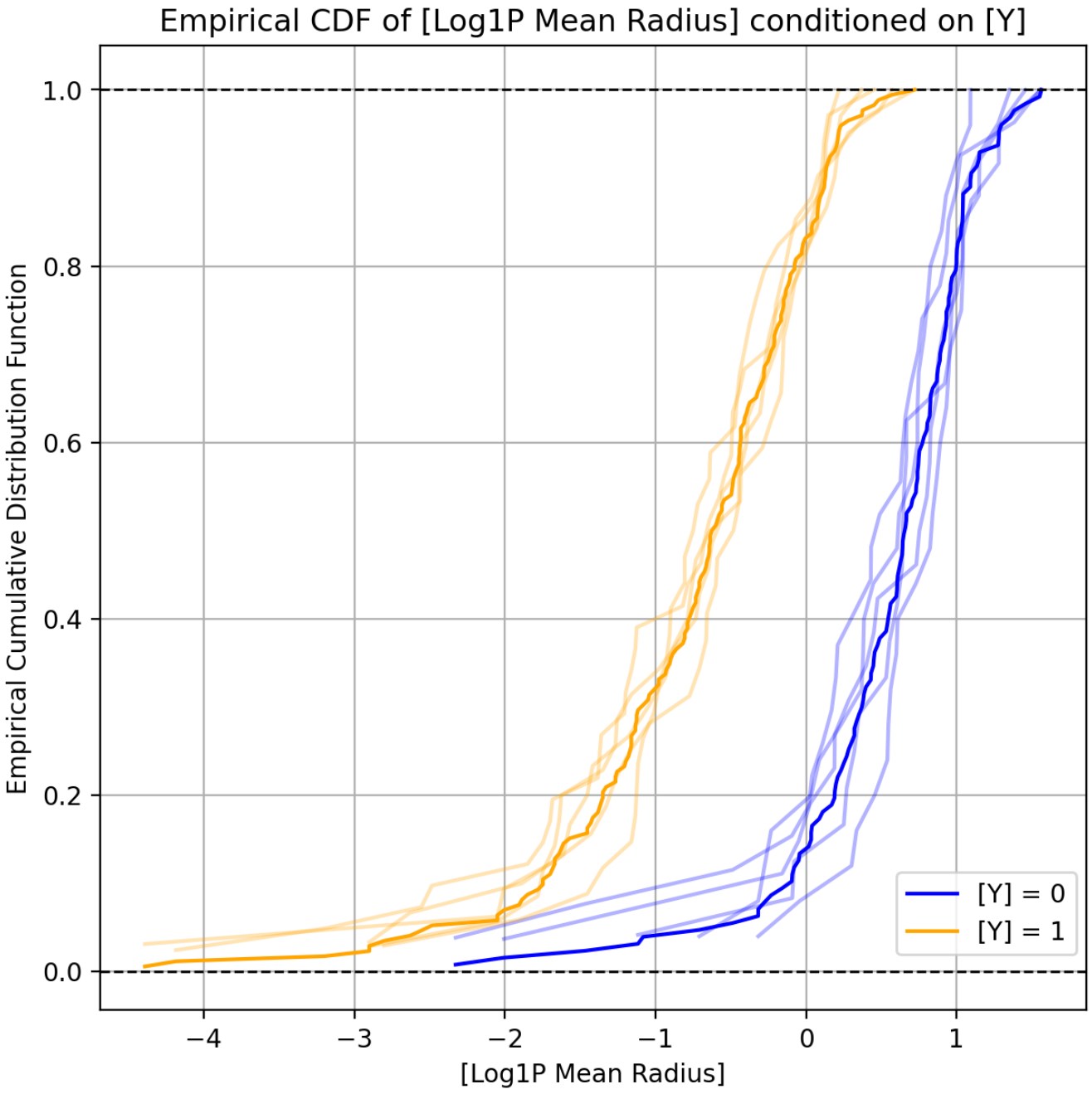
Log1P Mean Radius - Kernel Density Plot



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

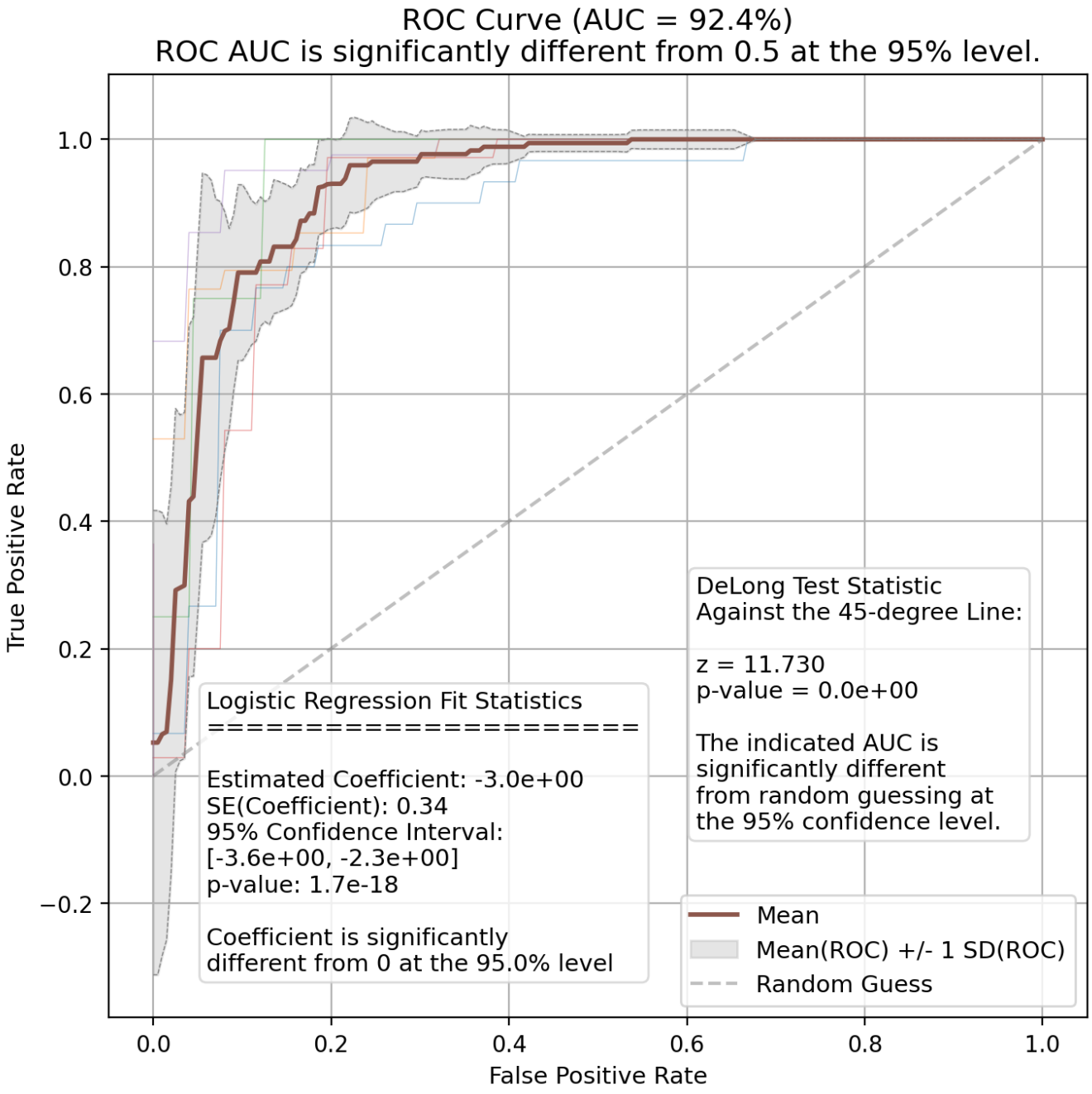
Log1P Mean Radius - Empirical CDF Plot



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

Log1P Mean Radius - ROC Curve

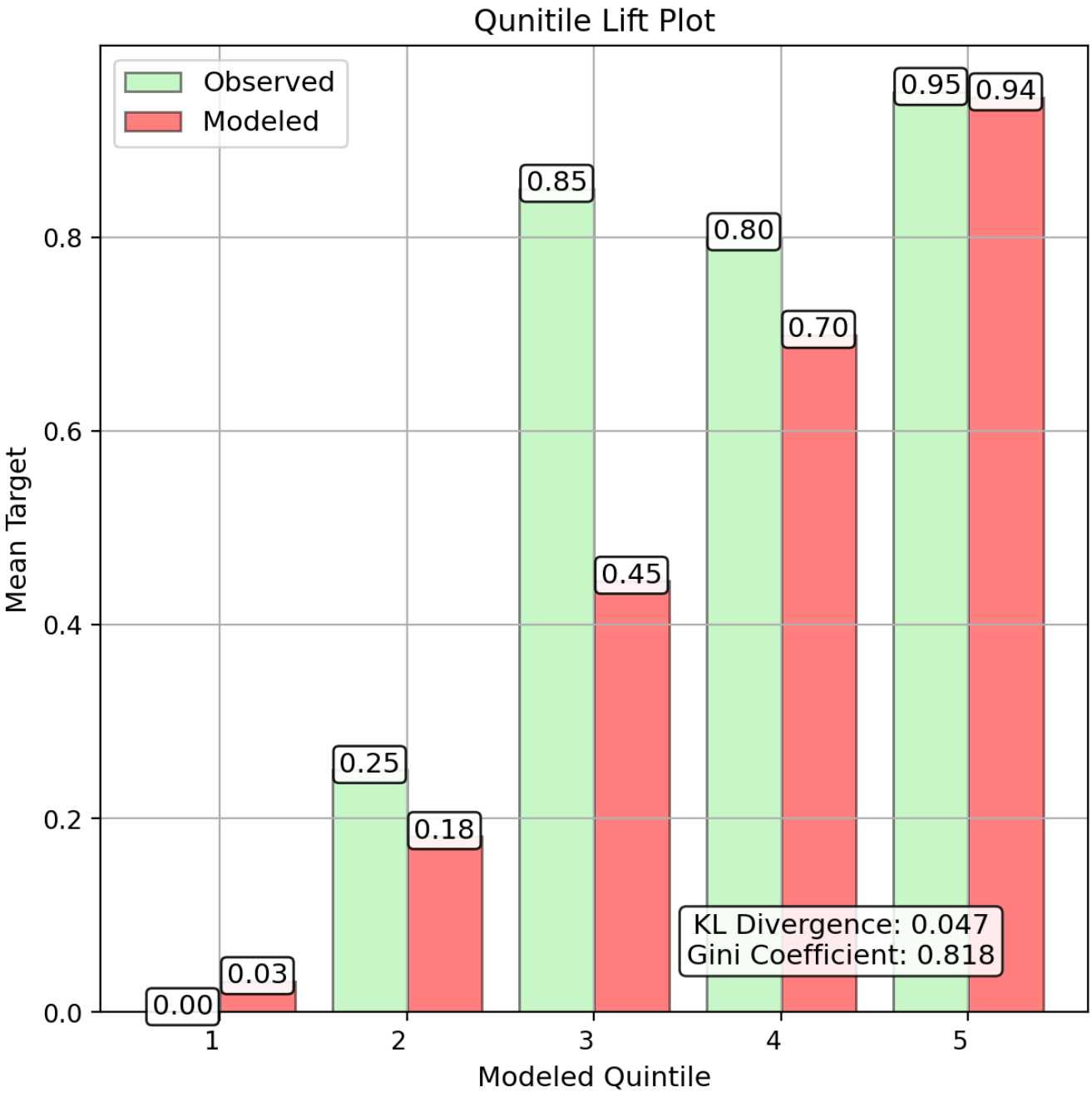


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

Log1P 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.

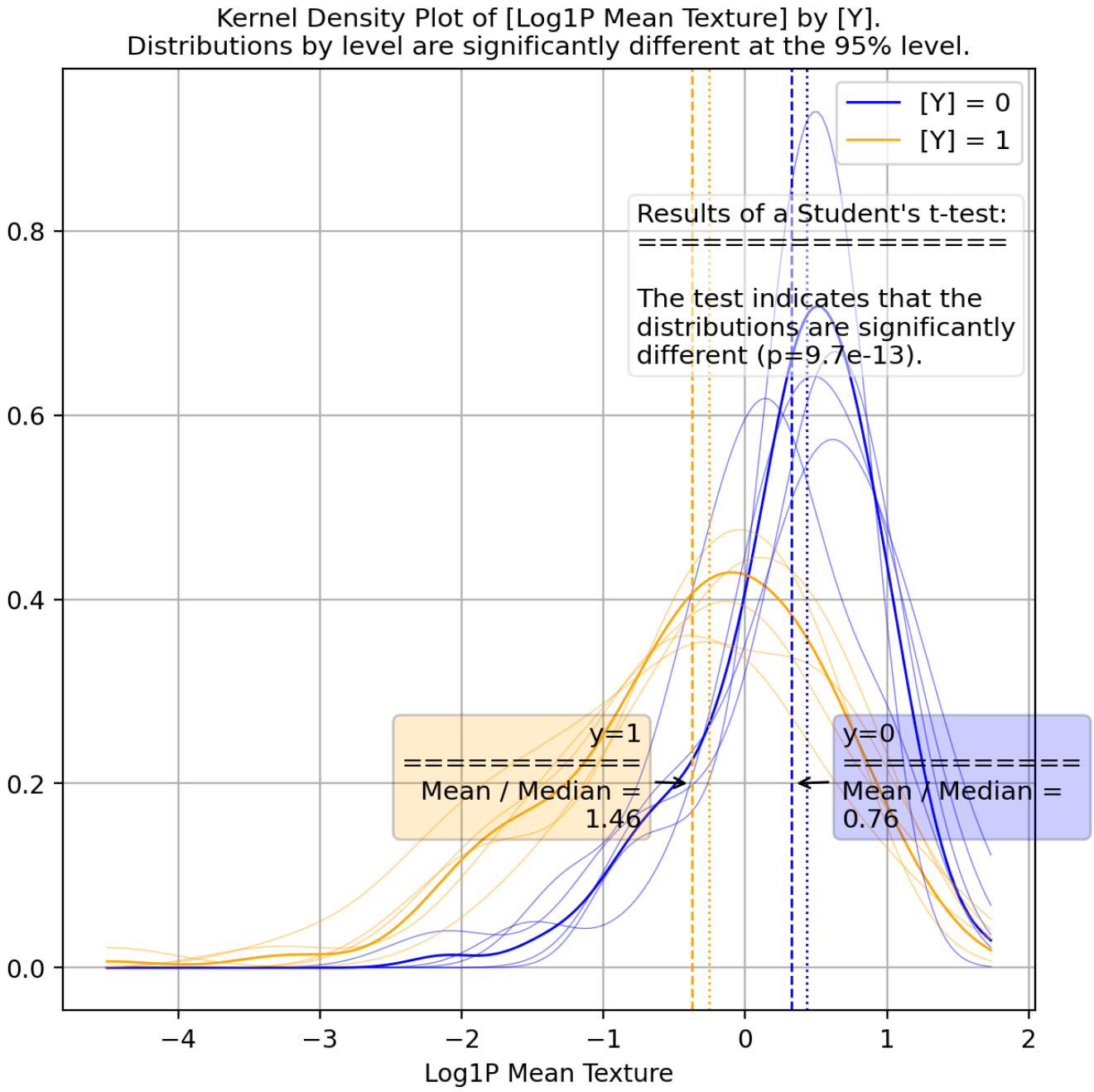
Univariate Report

Log1P Mean Texture - Results

	Coef	Pvalues	Se	Lower Ci	Upper Ci	Acc Test	Auc Test	F1 Test	Precision Test	Recall Test	Mcc Test
Fold-1	-1.11	2.5e-08	0.199	-1.50	-0.719	66.7%	67.2%	66.7%	73.1%	61.3%	34.4%
Fold-2	-1.20	5.1e-09	0.206	-1.61	-0.800	61.8%	62.4%	63.2%	69.2%	58.1%	24.6%
Fold-3	-1.12	2.2e-08	0.201	-1.52	-0.730	60.7%	64.6%	58.6%	81.0%	45.9%	30.1%
Fold-4	-0.97	4.0e-07	0.192	-1.35	-0.597	77.2%	77.8%	77.2%	84.6%	71.0%	55.6%
Fold-5	-1.09	6.7e-08	0.202	-1.49	-0.695	67.7%	69.7%	68.8%	81.5%	59.5%	39.0%
mean	-1.10	7.6e-10	0.179	-1.45	-0.749	65.3%	66.6%	64.6%	75.6%	56.4%	33.3%
std	0.08	1.7e-07	0.005	0.09	0.074	6.5%	6.0%	6.9%	6.4%	8.9%	11.8%

Univariate Report

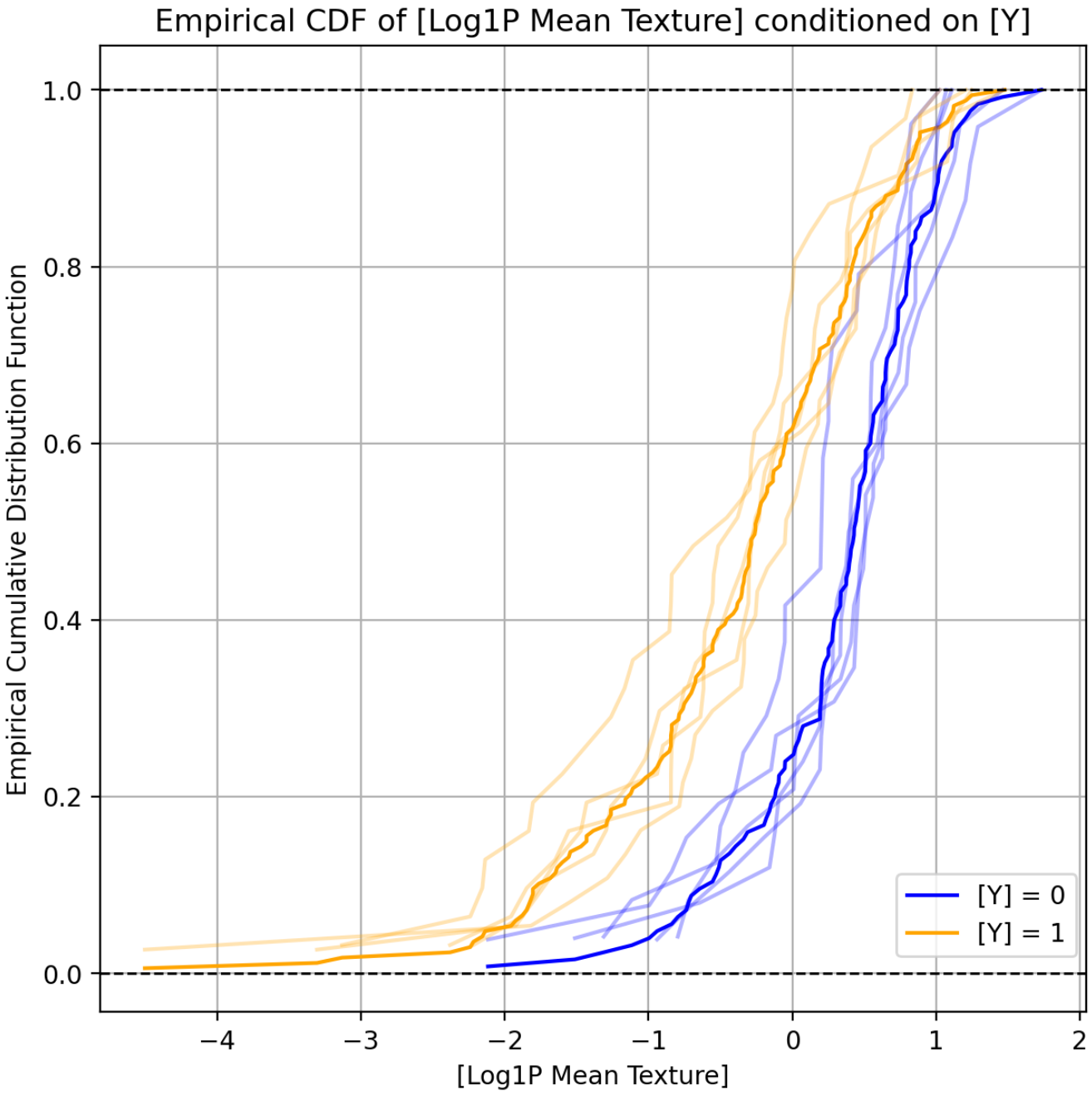
Log1P Mean Texture - Kernel Density Plot



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

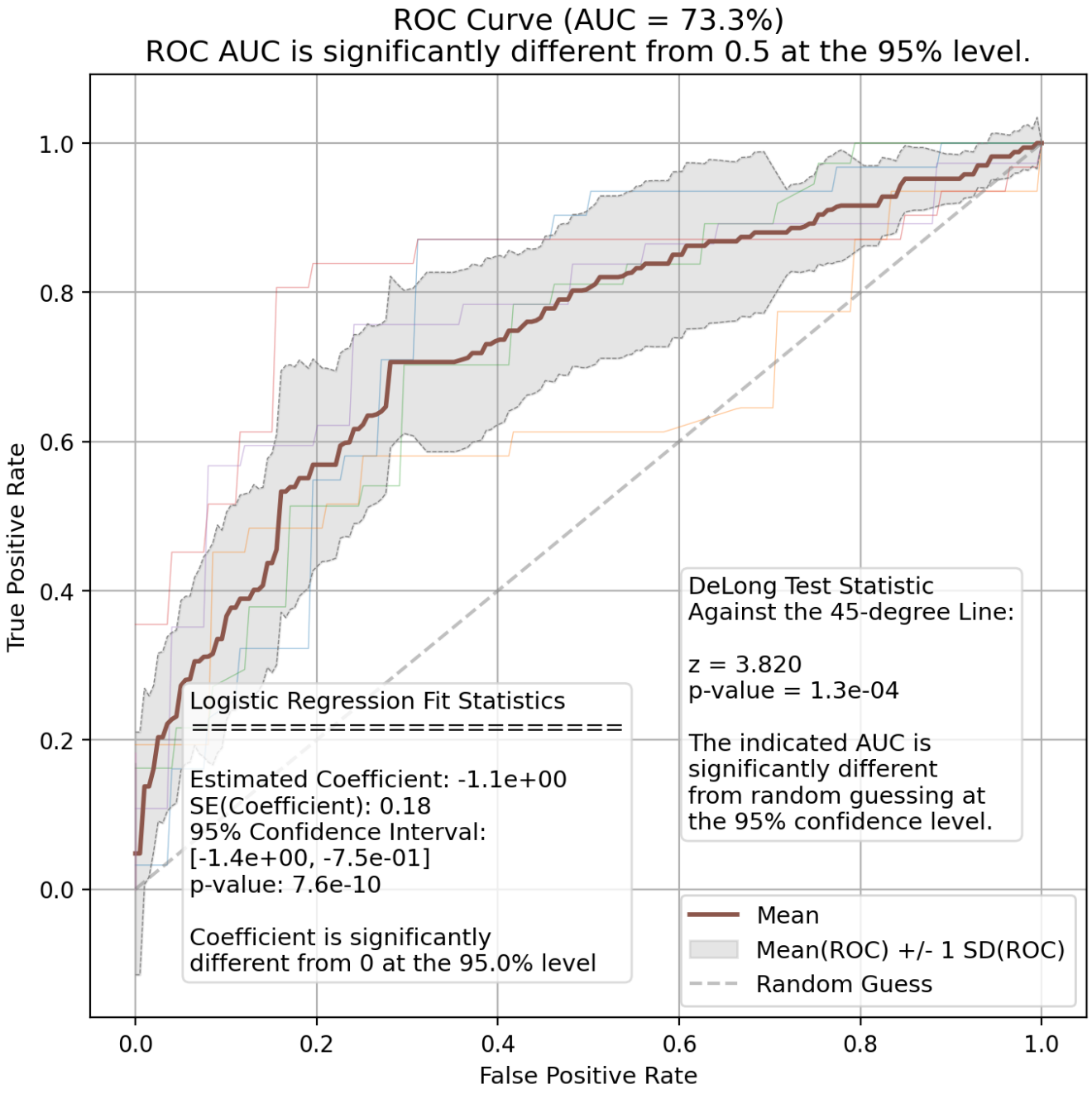
Log1P Mean Texture - Empirical CDF Plot



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

Log1P Mean Texture - ROC Curve

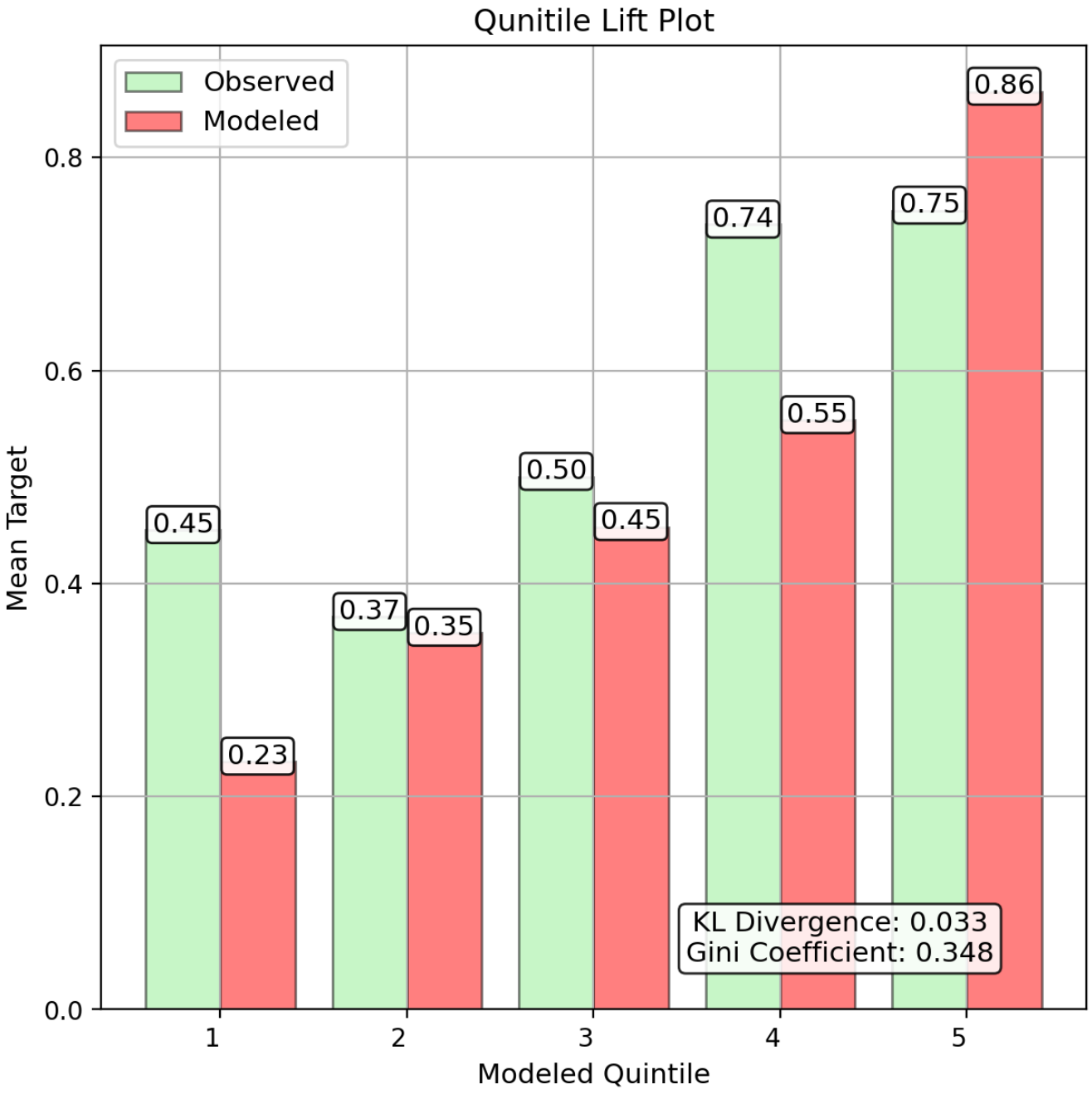


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

Log1P 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.

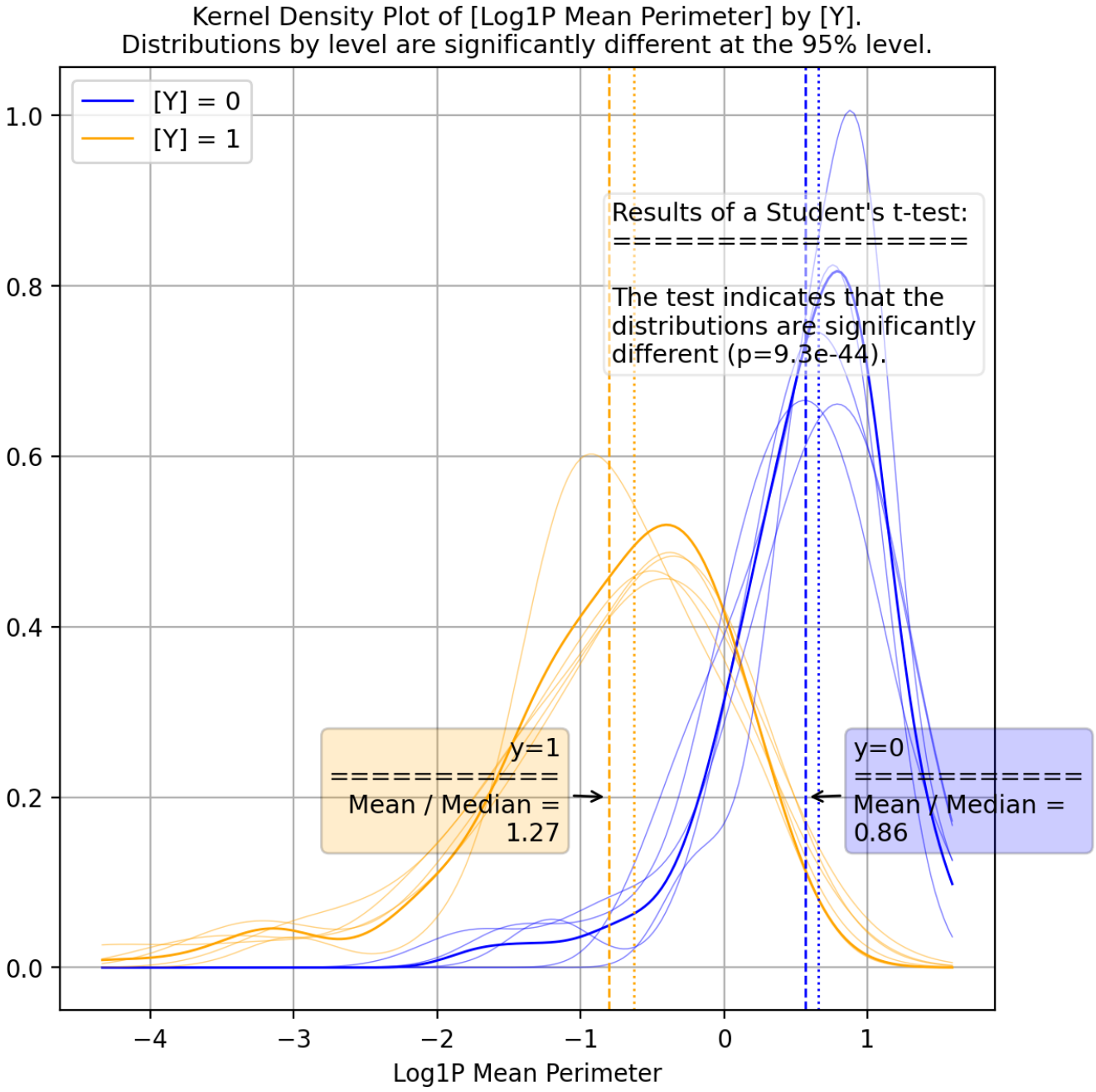
Univariate Report

Log1P Mean Perimeter - Results

	Coef	Pvalues	Se	Lower Ci	Upper Ci	Acc Test	Auc Test	F1 Test	Precision Test	Recall Test	Mcc Test
Fold-1	-3.60	3.4e-15	0.457	-4.50	-2.71	78.9%	79.6%	76.9%	90.9%	66.7%	60.8%
Fold-2	-3.27	7.2e-15	0.420	-4.09	-2.44	84.7%	85.7%	85.7%	93.1%	79.4%	70.6%
Fold-3	-3.26	1.3e-15	0.408	-4.06	-2.46	77.2%	79.7%	76.4%	95.5%	63.6%	60.3%
Fold-4	-3.64	8.1e-15	0.469	-4.56	-2.73	80.6%	81.7%	81.8%	90.0%	75.0%	62.7%
Fold-5	-3.18	9.8e-15	0.411	-3.99	-2.38	84.6%	86.8%	86.1%	96.9%	77.5%	71.5%
mean	-3.38	1.9e-18	0.386	-4.14	-2.62	76.2%	77.8%	76.5%	88.6%	67.2%	55.5%
std	0.22	3.5e-15	0.028	0.27	0.16	3.4%	3.3%	4.6%	2.9%	6.9%	5.5%

Univariate Report

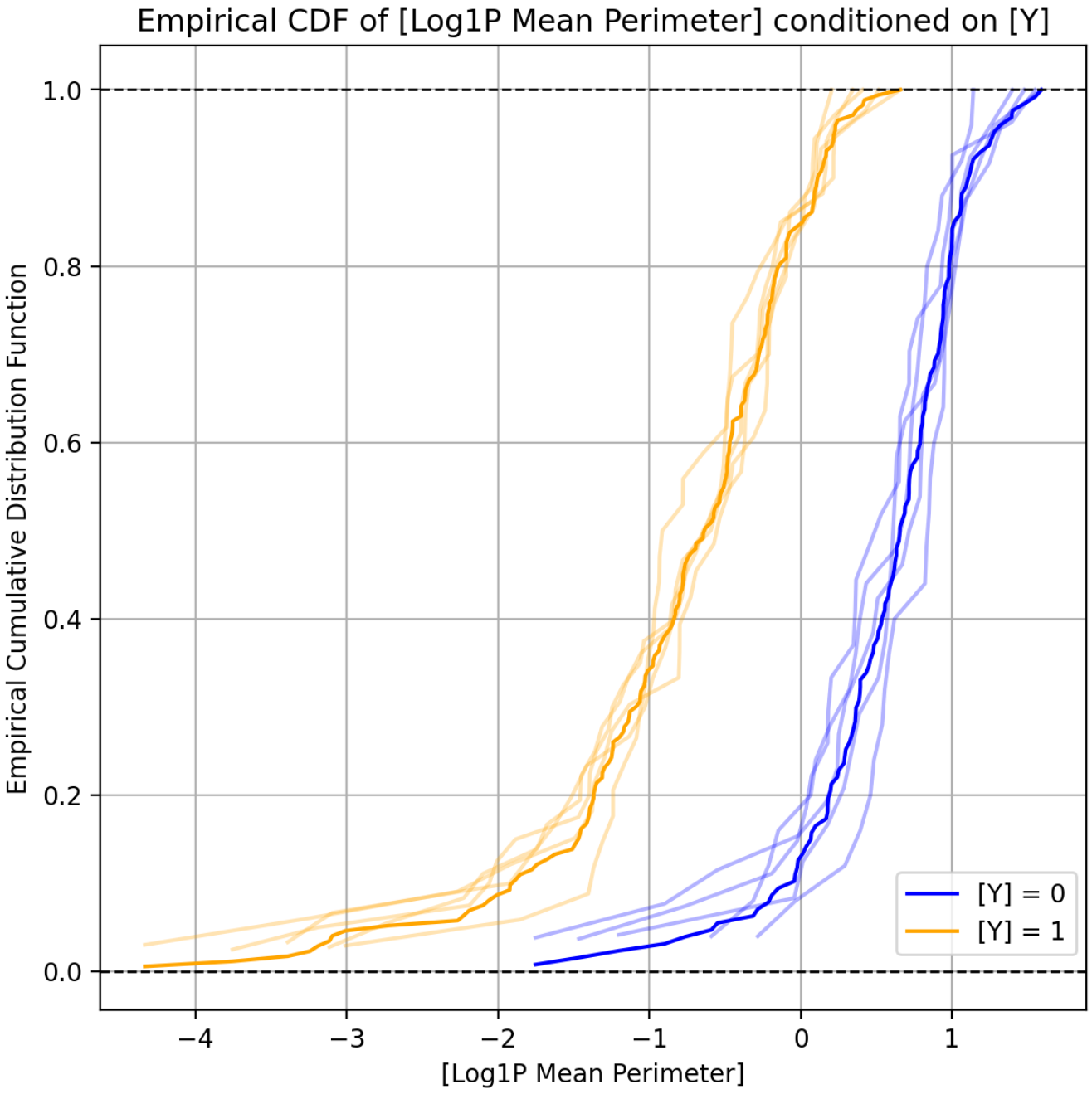
Log1P Mean Perimeter - Kernel Density Plot



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

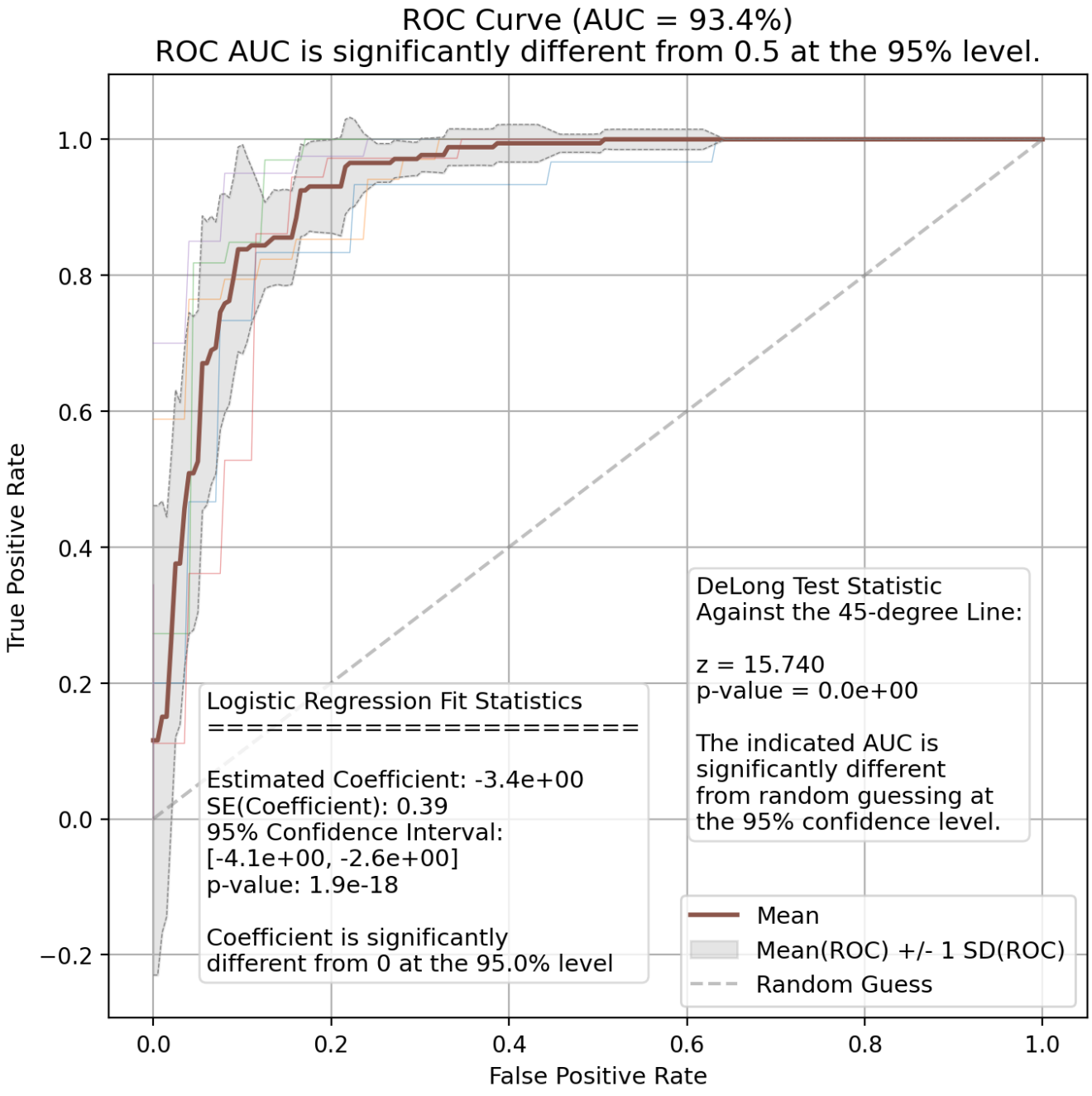
Log1P Mean Perimeter - Empirical CDF Plot



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

Log1P Mean Perimeter - ROC Curve

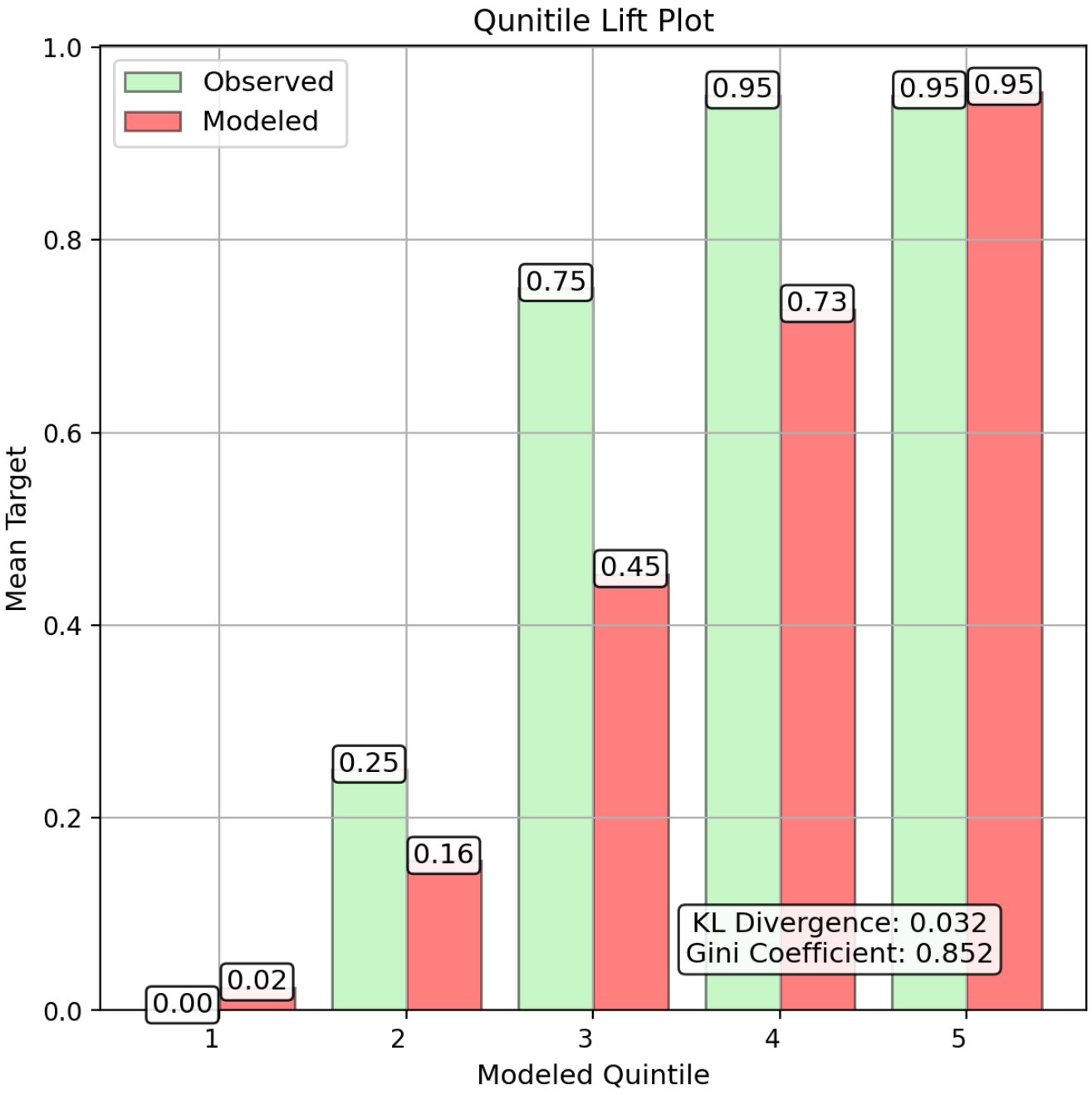


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

Log1P 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.

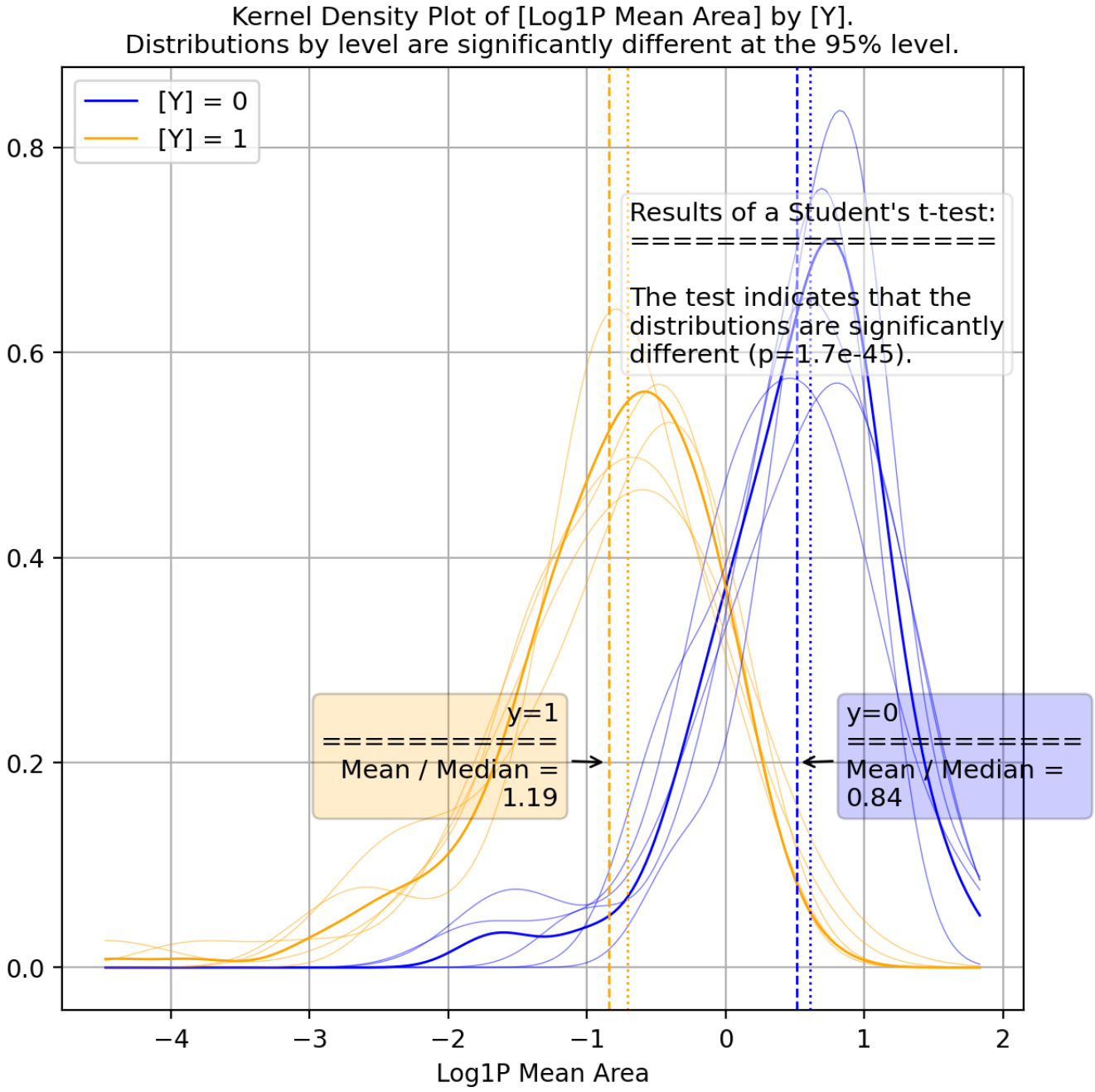
Univariate Report

Log1P Mean Area - Results

	Coef	Pvalues	Se	Lower Ci	Upper Ci	Acc Test	Auc Test	F1 Test	Precision Test	Recall Test	Mcc Test
Fold-1	-3.64	1.0e-15	0.454	-4.53	-2.75	81.4%	82.2%	80.7%	92.0%	71.9%	65.0%
Fold-2	-3.19	2.0e-15	0.401	-3.97	-2.40	83.9%	83.9%	86.1%	88.6%	83.8%	67.1%
Fold-3	-3.13	2.8e-16	0.383	-3.88	-2.38	80.0%	82.6%	80.6%	96.2%	69.4%	64.5%
Fold-4	-3.59	1.9e-15	0.451	-4.47	-2.70	81.0%	82.1%	82.4%	90.3%	75.7%	63.2%
Fold-5	-3.07	2.4e-15	0.387	-3.83	-2.31	83.8%	86.4%	85.7%	97.1%	76.7%	70.1%
mean	-3.31	3.1e-19	0.369	-4.03	-2.58	77.1%	78.2%	78.9%	86.5%	72.6%	55.4%
std	0.27	8.7e-16	0.035	0.34	0.20	1.8%	1.8%	2.7%	3.7%	5.5%	2.7%

Univariate Report

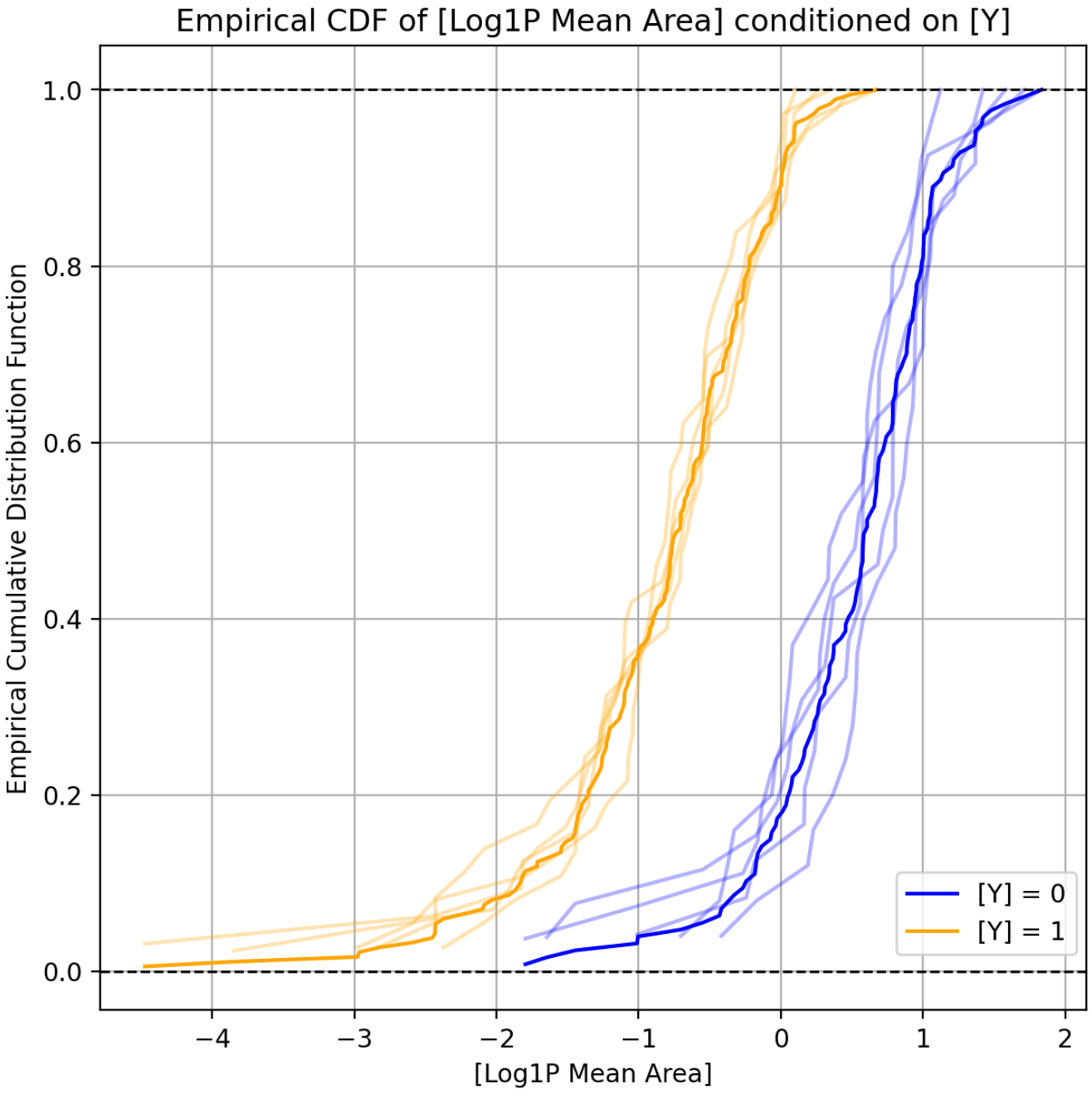
Log1P Mean Area - Kernel Density Plot



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

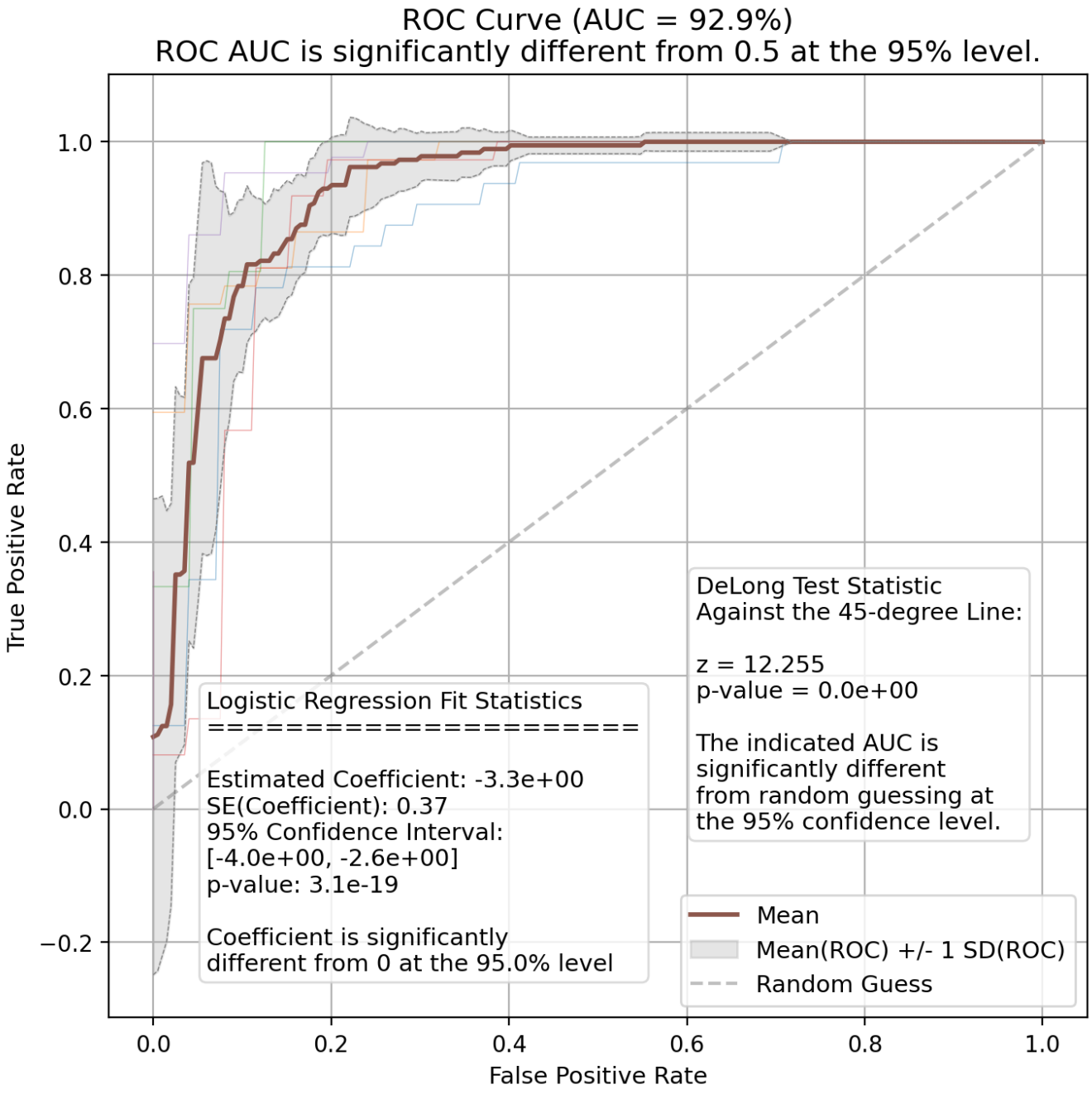
Log1P Mean Area - Empirical CDF Plot



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

Log1P Mean Area - ROC Curve

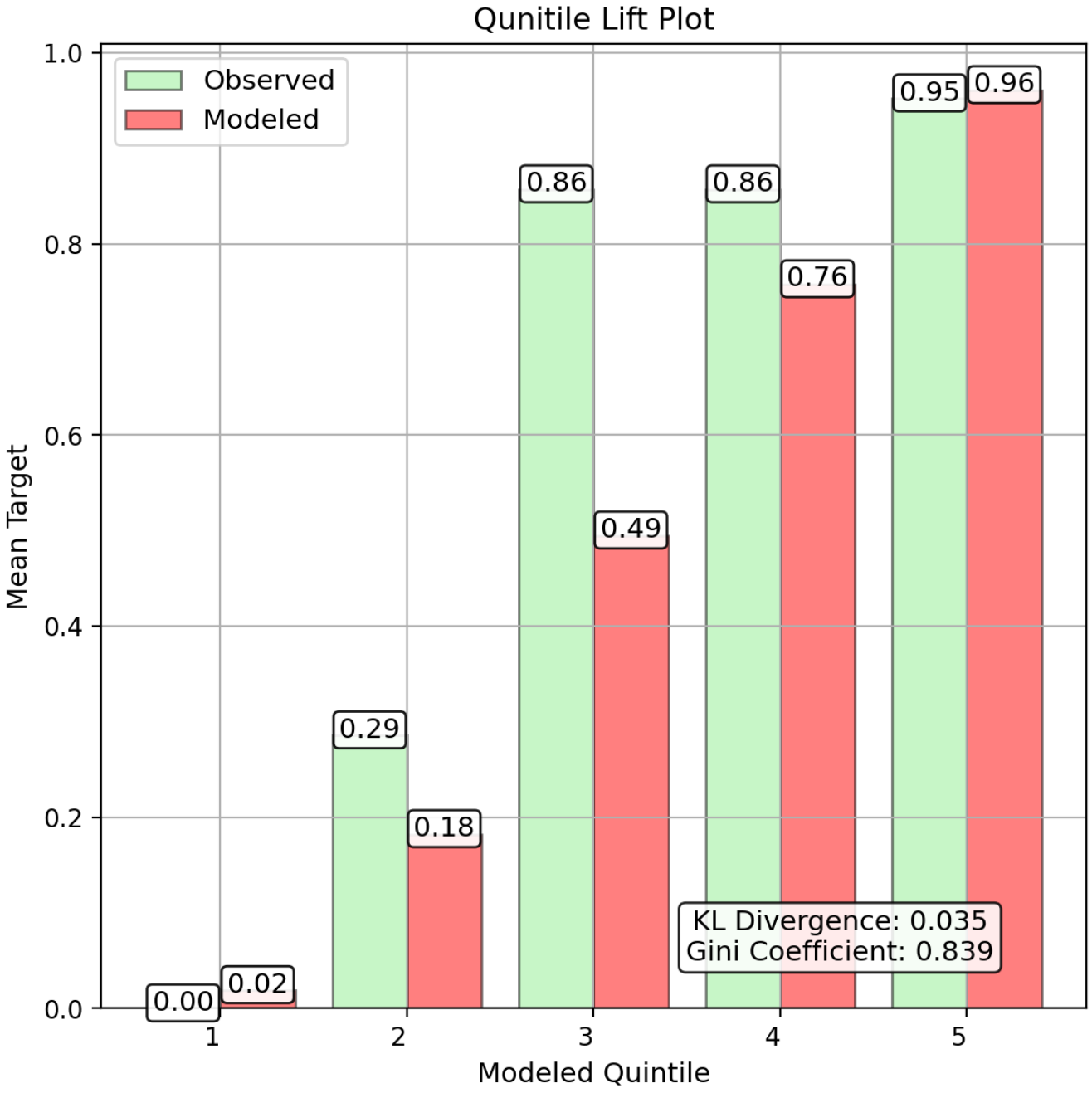


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

Log1P 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.