

# **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 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.
<b>Mean Area</b>	88.6%	88.2%	94.4%	86.7%	91.2%	75.5%	87.4%
<b>Mean Radius</b>	86.8%	90.0%	88.7%	86.2%	89.4%	72.1%	85.5%
<b>Mean Perimeter</b>	86.8%	90.0%	88.7%	86.2%	89.4%	72.1%	85.5%
<b>Log1p Mean_area</b>	77.1%	86.5%	72.6%	78.2%	78.9%	55.4%	74.8%
<b>Log1p Mean_perimeter</b>	76.2%	88.6%	67.2%	77.8%	76.5%	55.5%	73.6%
<b>Log1p Mean_radius</b>	75.0%	84.8%	68.4%	76.1%	75.7%	51.8%	72.0%
<b>Mean Texture</b>	71.9%	81.0%	71.8%	72.0%	76.1%	42.8%	69.3%
<b>Log1p Mean_texture</b>	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

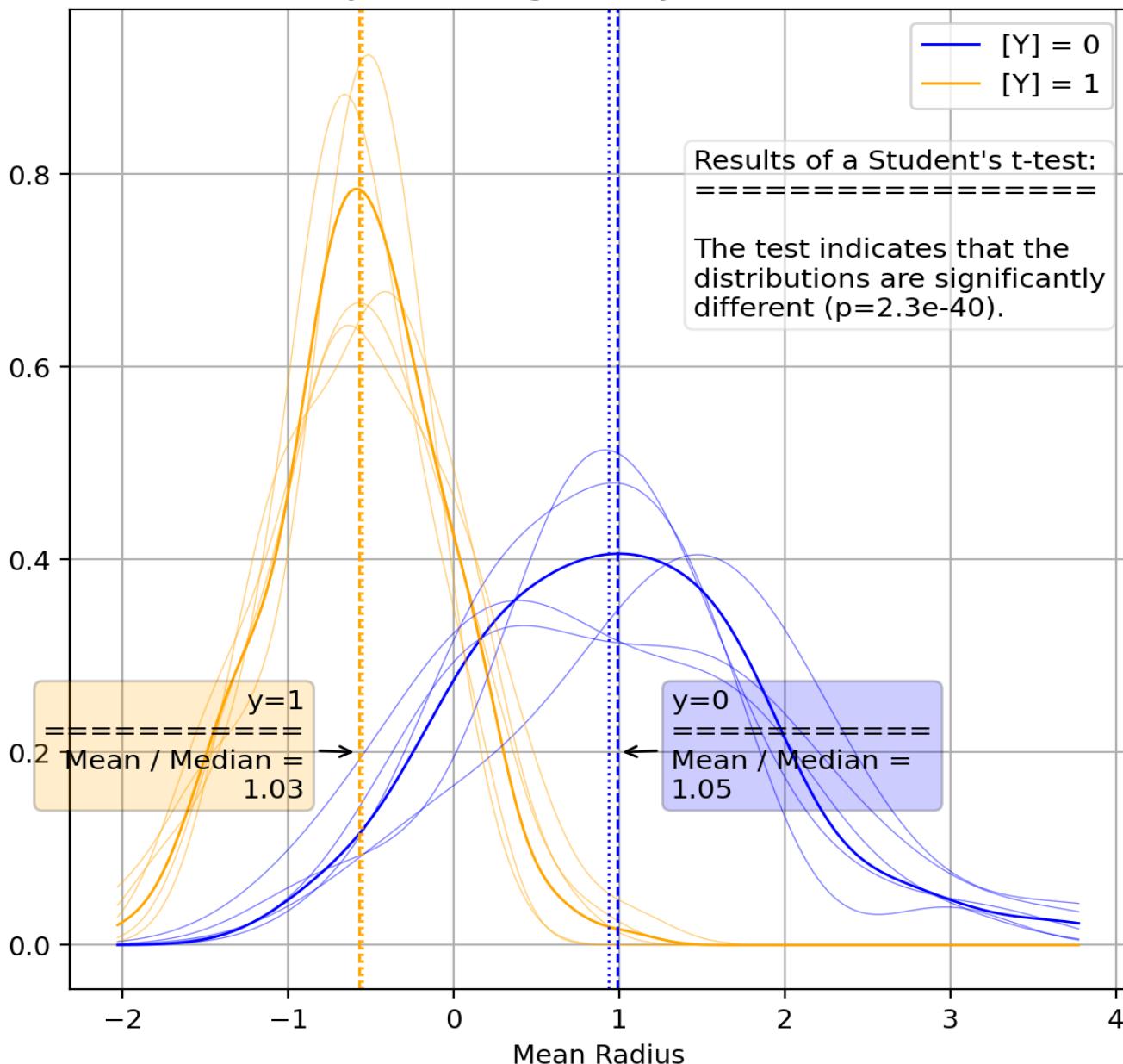
### Mean Radius - Results

	Fitted Coef.	Fitted p-Value	Fitted Std. Err.	Conf. Int. Lower	Conf. Int. Upper	Test Accuracy	Test AUC	Test F1	Test Precision	Test Recall	Test MCC
<b>Fold-1</b>	-3.72	1.3e-16	0.450	-4.60	-2.84	85.0%	83.7%	87.7%	84.2%	91.4%	69.0%
<b>Fold-2</b>	-3.64	2.0e-16	0.443	-4.51	-2.77	84.9%	85.7%	87.6%	92.9%	83.0%	69.2%
<b>Fold-3</b>	-3.66	3.1e-17	0.434	-4.51	-2.81	86.0%	88.2%	88.6%	96.9%	81.6%	72.5%
<b>Fold-4</b>	-3.48	1.1e-15	0.435	-4.33	-2.63	93.4%	92.9%	94.5%	93.5%	95.6%	86.3%
<b>Fold-5</b>	-3.98	1.8e-15	0.500	-4.96	-3.00	81.3%	82.1%	84.8%	90.7%	79.6%	61.8%
<b>mean</b>	-3.69	5.6e-20	0.403	-4.48	-2.90	86.8%	86.2%	89.4%	90.0%	88.7%	72.1%
<b>std</b>	0.18	7.8e-16	0.027	0.23	0.13	4.4%	4.2%	3.6%	4.7%	6.9%	9.0%

# Univariate Report

## Mean Radius - Kernel Density Plot

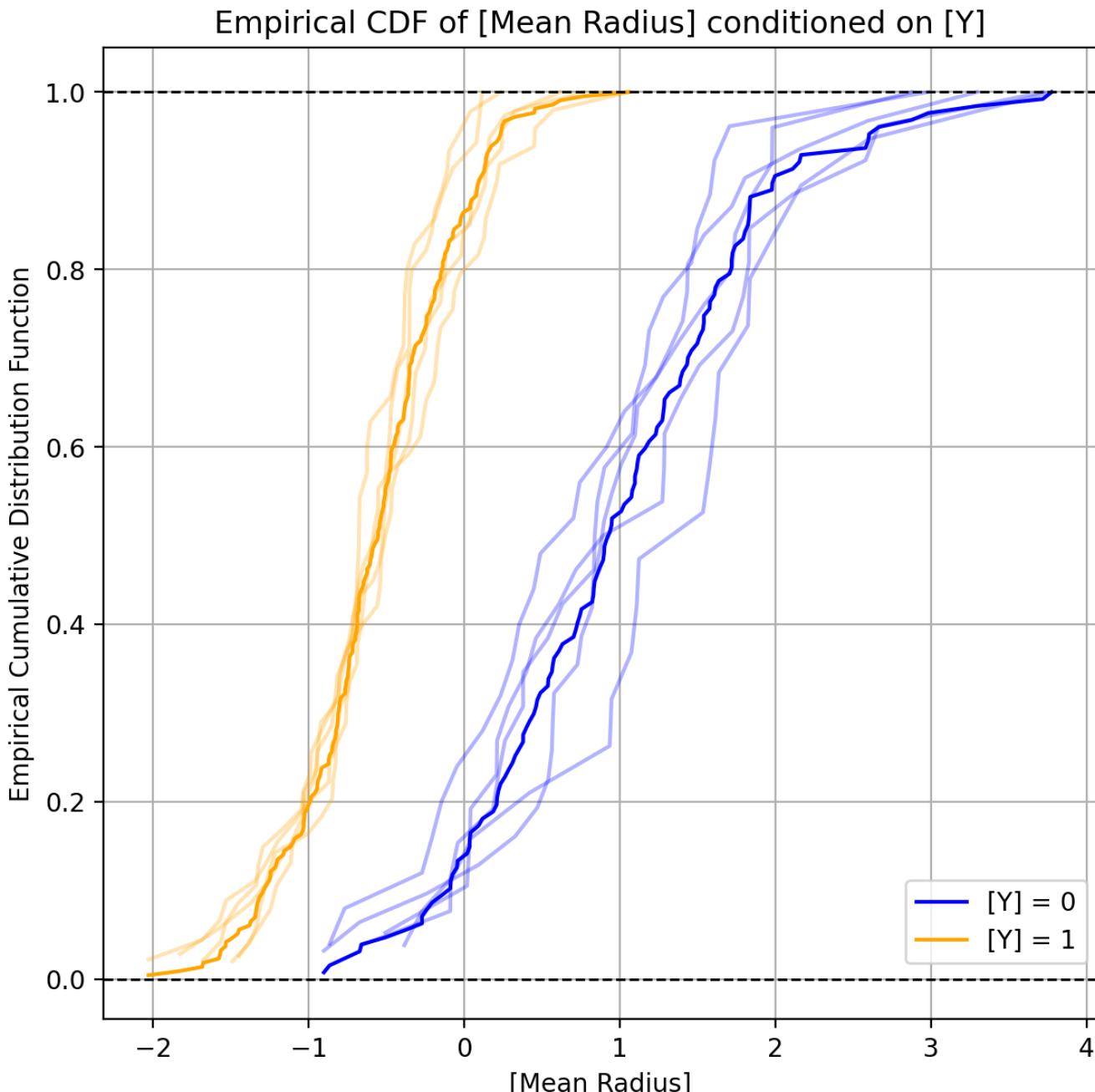
Kernel Density Plot of [Mean Radius] by [Y].  
Distributions by level are significantly different at the 95% level.



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

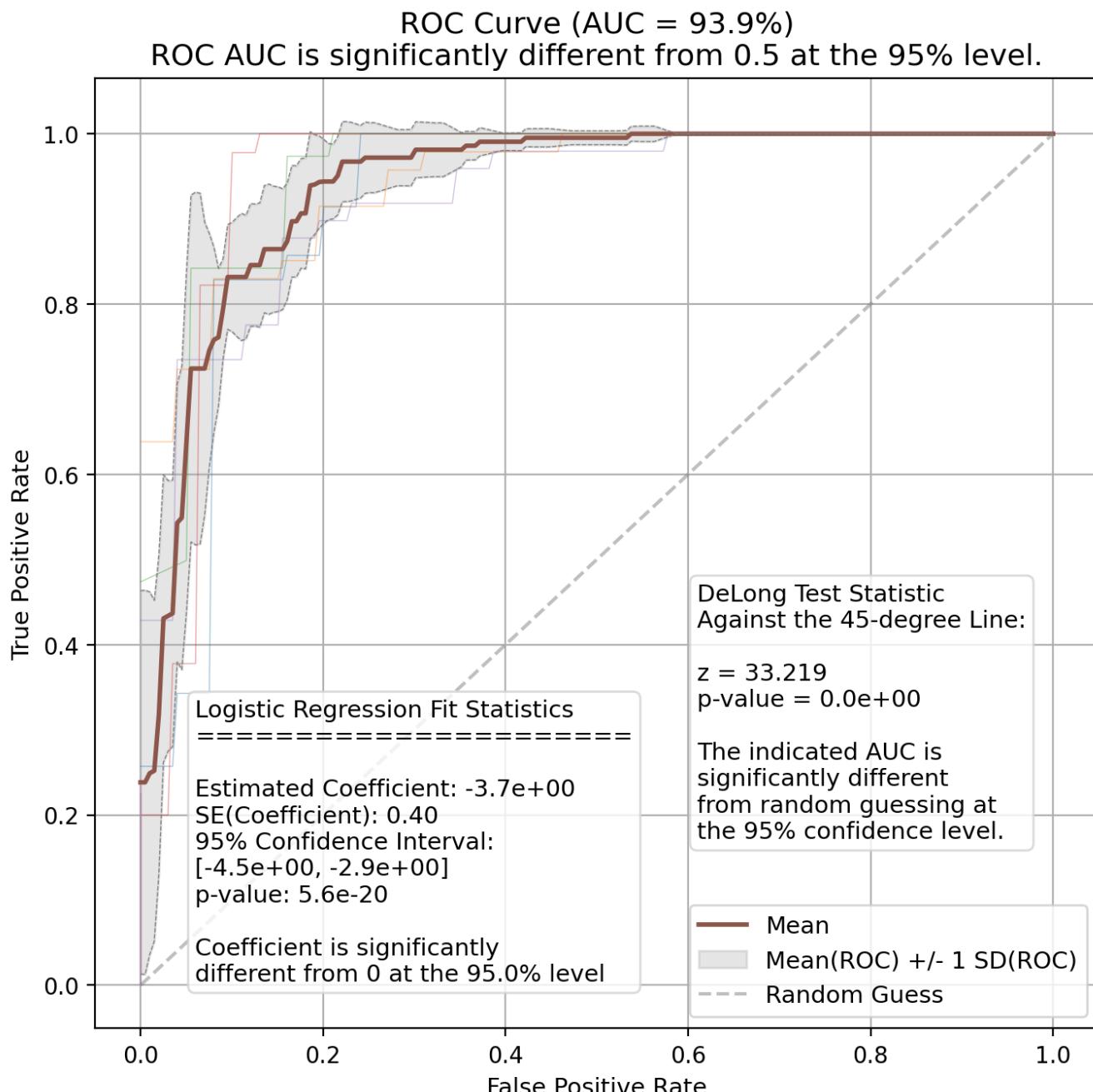
### Mean Radius - Empirical CDF Plot



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

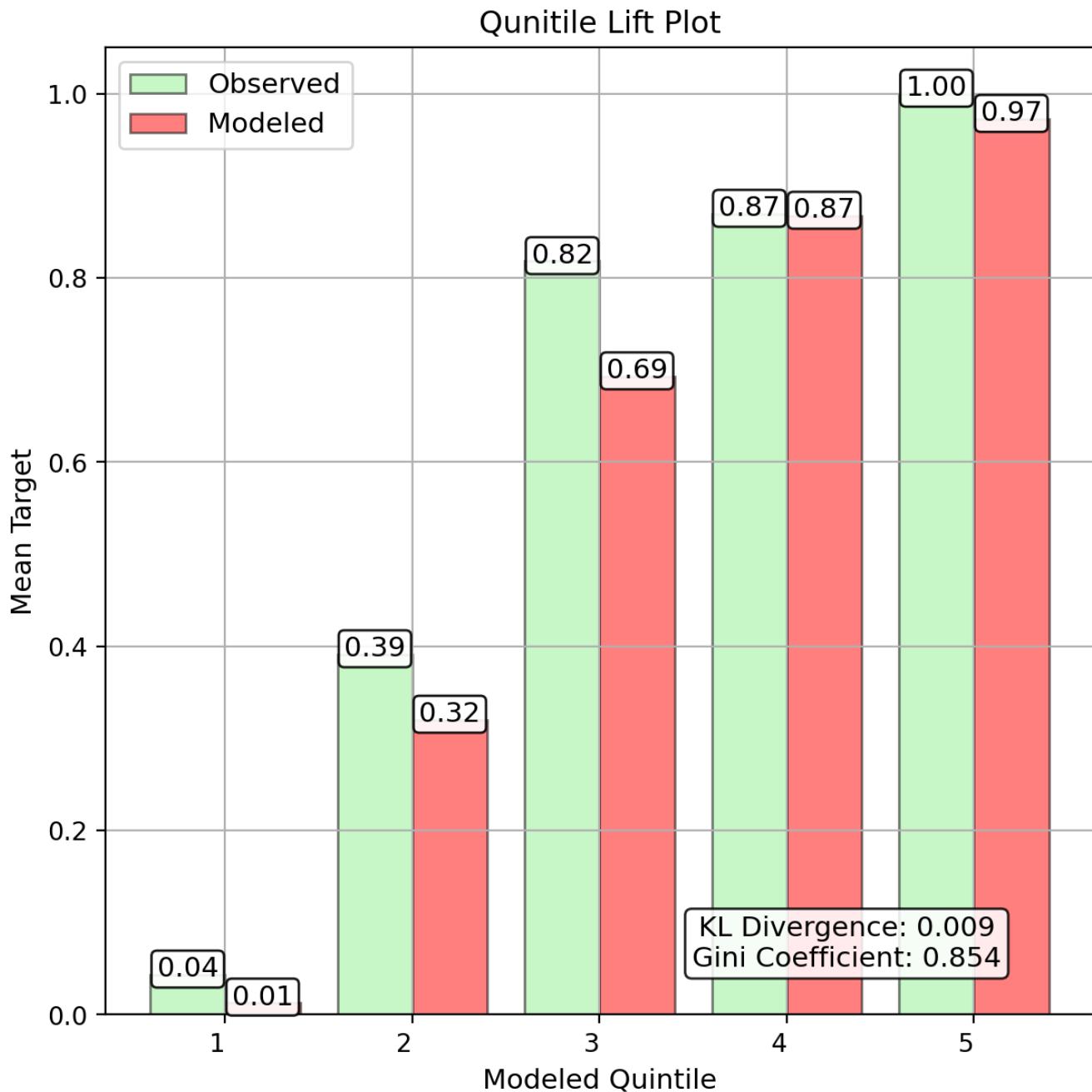


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

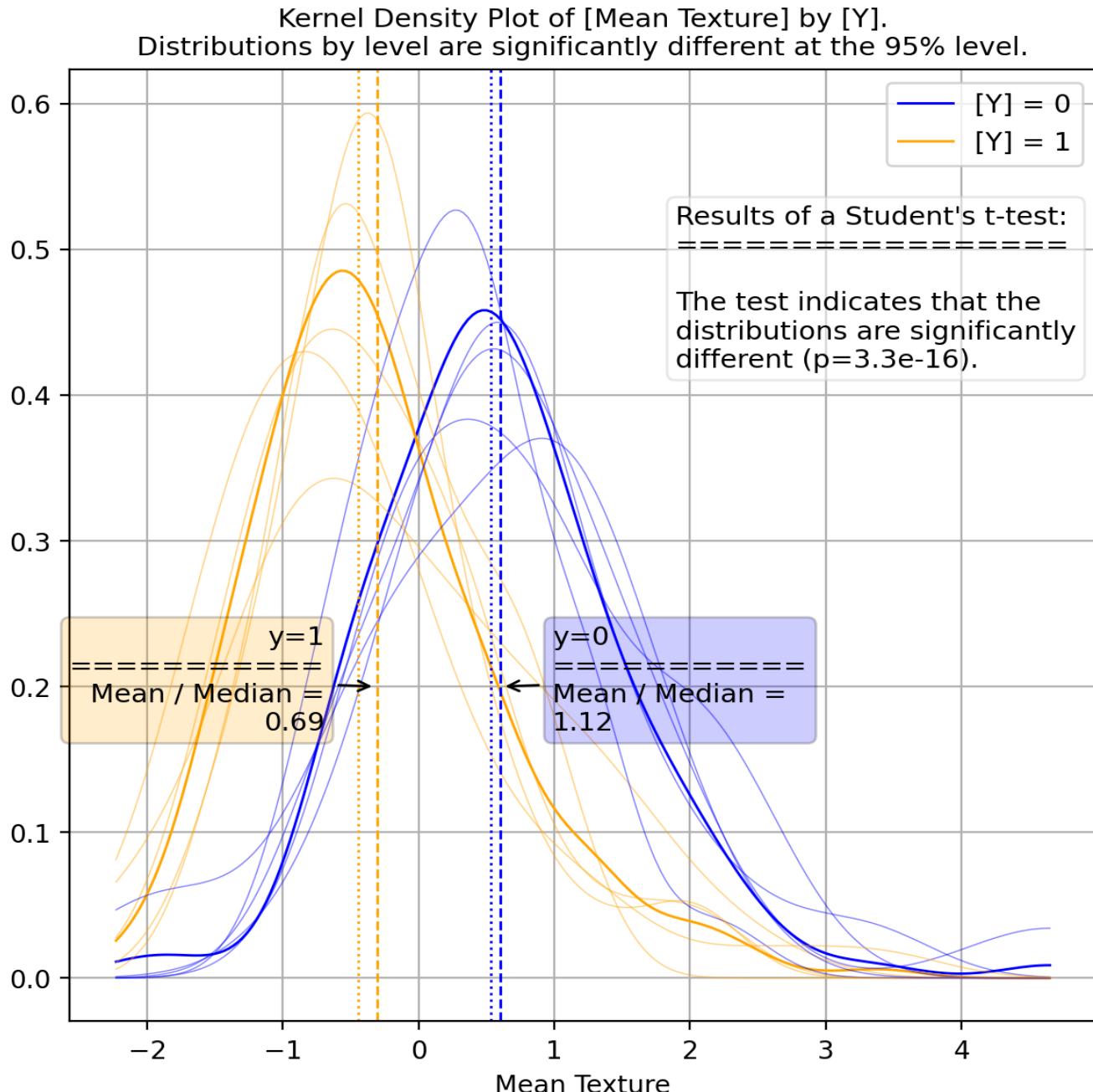
## Univariate Report

### Mean Texture - Results

	Fitted Coef.	Fitted p-Value	Fitted Std. Err.	Conf. Int. Lower	Conf. Int. Upper	Test Accuracy	Test AUC	Test F1	Test Precision	Test Recall	Test MCC
<b>Fold-1</b>	-1.09	9.8e-11	0.169	-1.42	-0.761	66.7%	68.0%	67.7%	77.8%	60.0%	35.7%
<b>Fold-2</b>	-0.97	1.8e-09	0.161	-1.28	-0.651	71.2%	72.5%	75.3%	84.2%	68.1%	43.1%
<b>Fold-3</b>	-0.94	6.3e-10	0.151	-1.23	-0.640	73.7%	77.6%	76.9%	92.6%	65.8%	52.2%
<b>Fold-4</b>	-1.00	2.3e-10	0.158	-1.31	-0.693	72.4%	71.1%	76.9%	76.1%	77.8%	42.5%
<b>Fold-5</b>	-1.07	9.4e-10	0.175	-1.41	-0.727	73.3%	70.6%	79.6%	79.6%	79.6%	41.1%
<b>mean</b>	-1.01	3.5e-12	0.145	-1.29	-0.725	71.9%	72.0%	76.1%	81.0%	71.8%	42.8%
<b>std</b>	0.07	6.9e-10	0.009	0.08	0.051	2.8%	3.6%	4.5%	6.6%	8.3%	5.9%

## Univariate Report

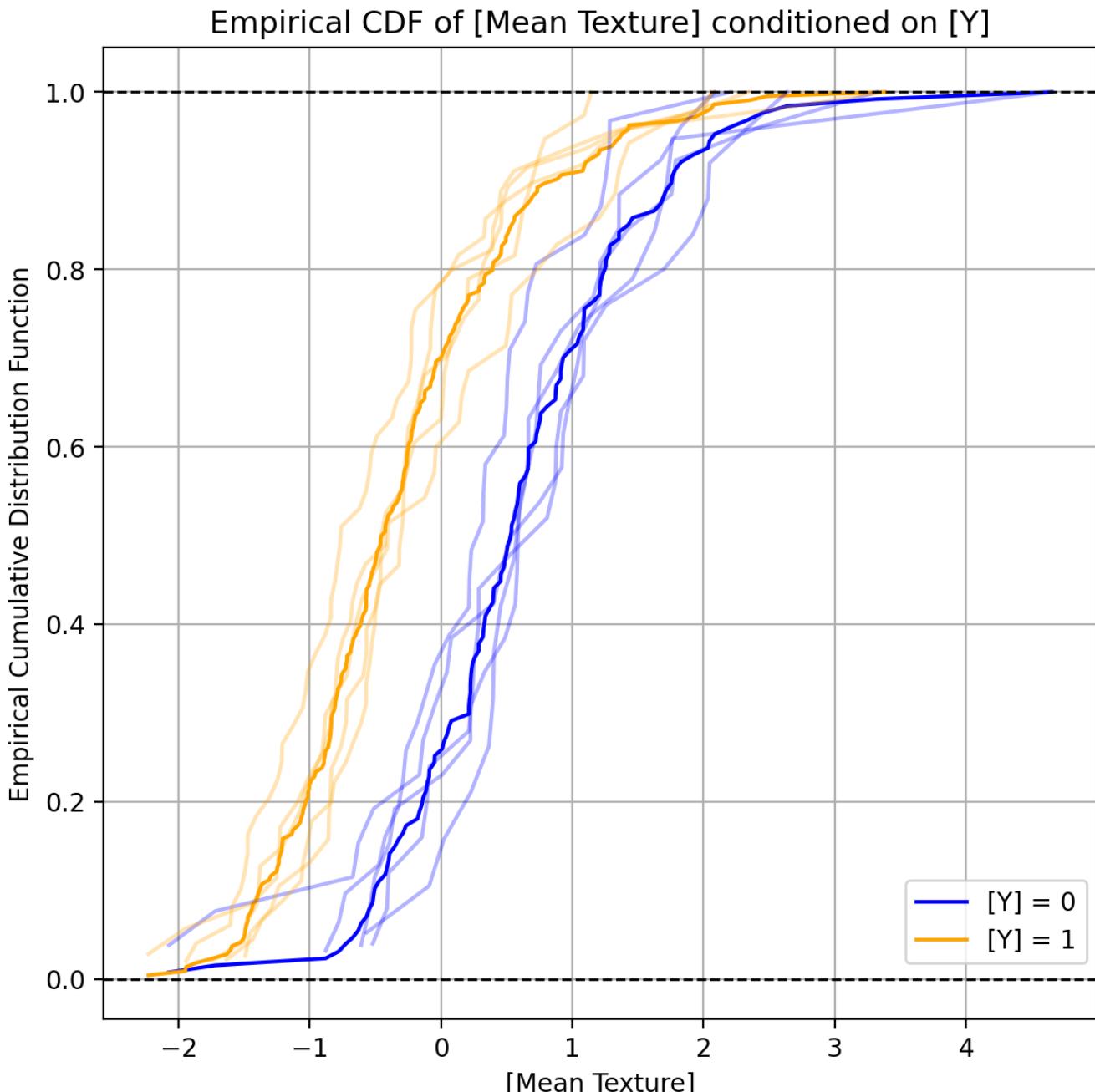
### Mean Texture - Kernel Density Plot



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

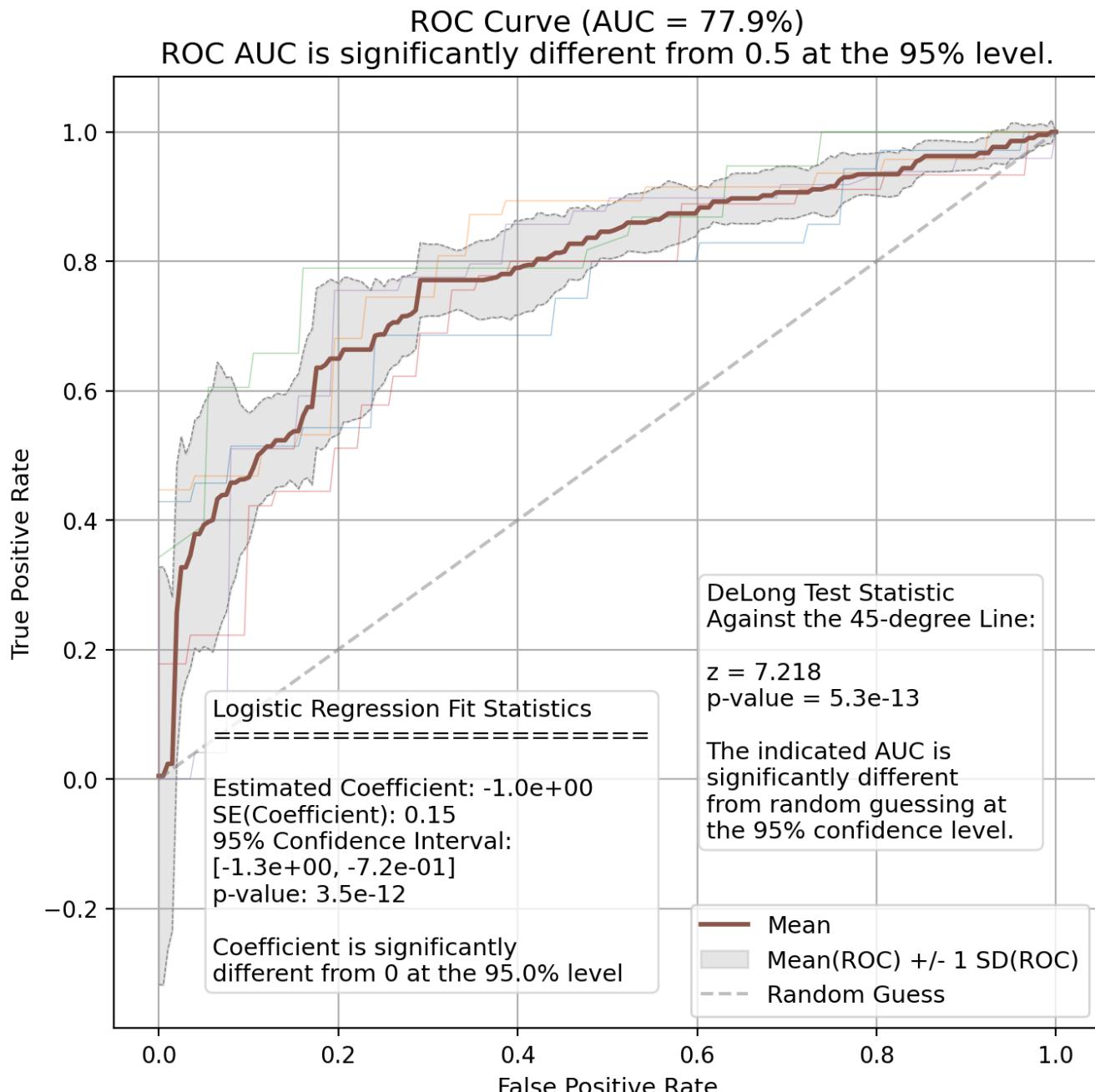
### Mean Texture - Empirical CDF Plot



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

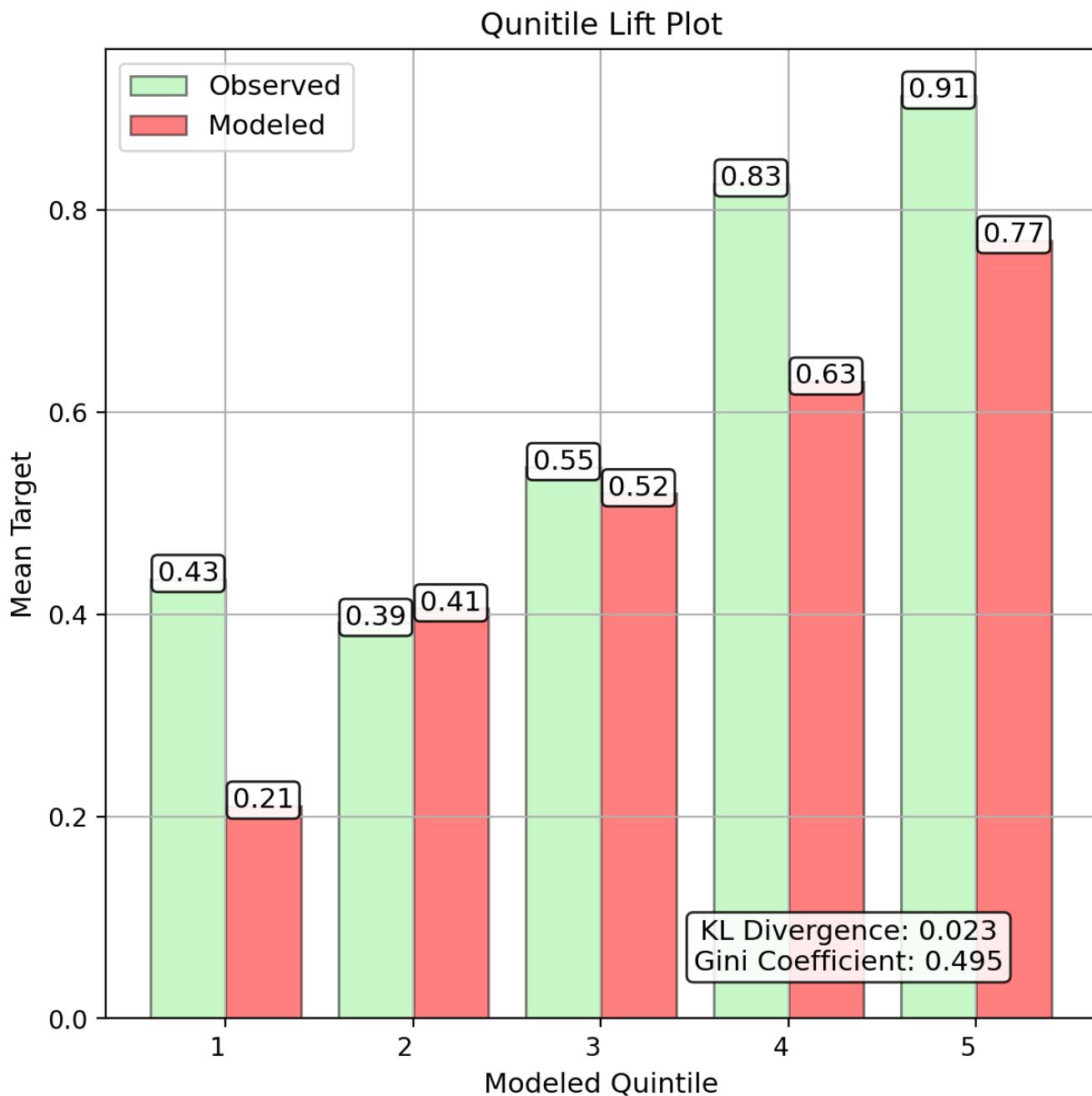


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

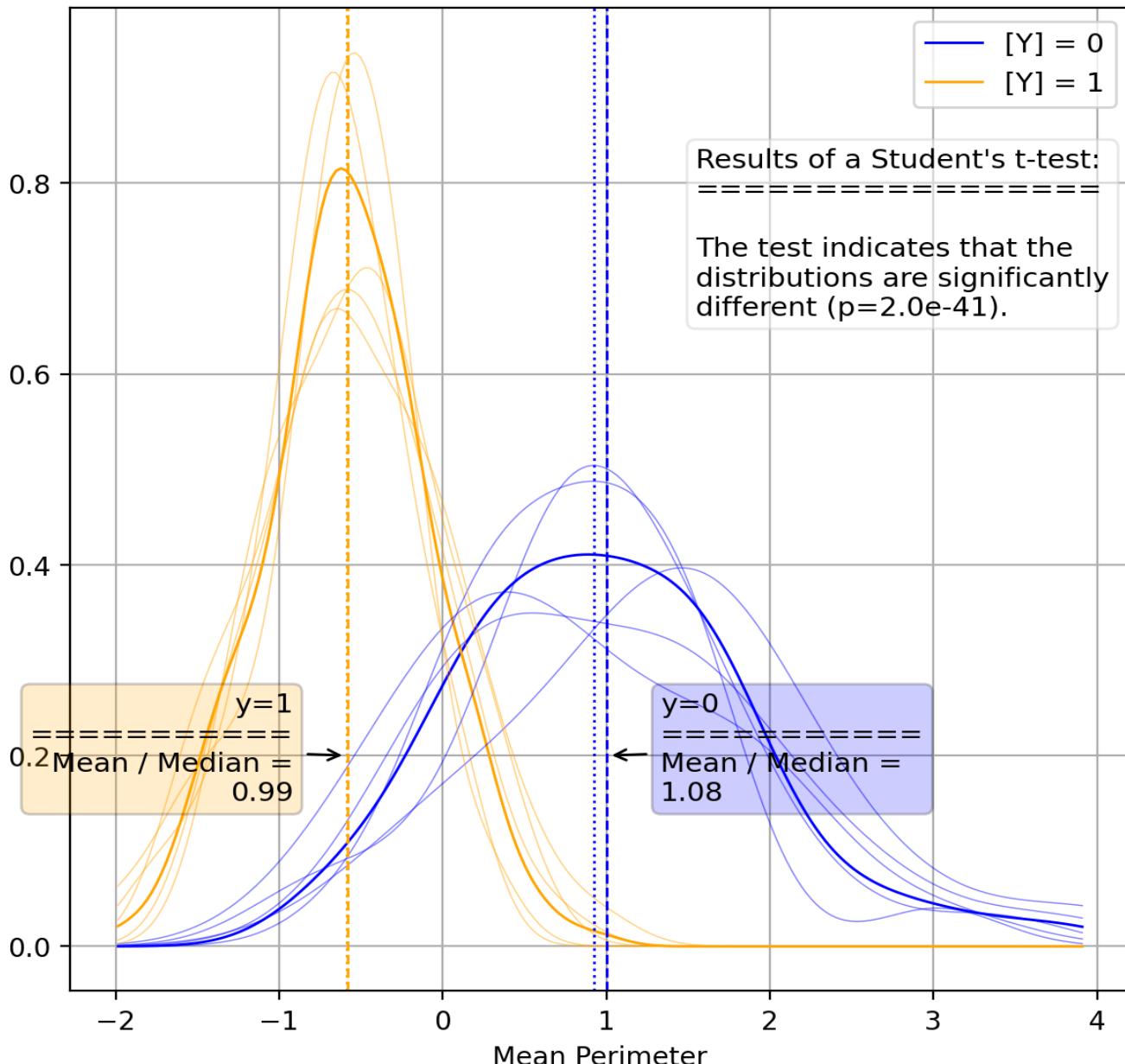
### Mean Perimeter - Results

	Fitted Coef.	Fitted p-Value	Fitted Std. Err.	Conf. Int. Lower	Conf. Int. Upper	Test Accuracy	Test AUC	Test F1	Test Precision	Test Recall	Test MCC
<b>Fold-1</b>	-4.01	1.5e-16	0.485	-4.96	-3.06	86.7%	85.1%	89.2%	84.6%	94.3%	72.6%
<b>Fold-2</b>	-3.93	2.2e-16	0.479	-4.87	-2.99	87.7%	87.8%	90.1%	93.2%	87.2%	74.1%
<b>Fold-3</b>	-3.99	4.2e-17	0.475	-4.92	-3.06	87.7%	89.5%	90.1%	97.0%	84.2%	75.4%
<b>Fold-4</b>	-3.86	1.4e-15	0.484	-4.81	-2.91	93.4%	92.9%	94.5%	93.5%	95.6%	86.3%
<b>Fold-5</b>	-4.22	1.5e-15	0.529	-5.25	-3.18	81.3%	82.1%	84.8%	90.7%	79.6%	61.8%
<b>mean</b>	-4.00	6.6e-20	0.438	-4.86	-3.14	86.8%	86.2%	89.4%	90.0%	88.7%	72.1%
<b>std</b>	0.13	7.3e-16	0.022	0.17	0.10	4.3%	4.1%	3.5%	4.6%	6.7%	8.7%

# Univariate Report

## Mean Perimeter - Kernel Density Plot

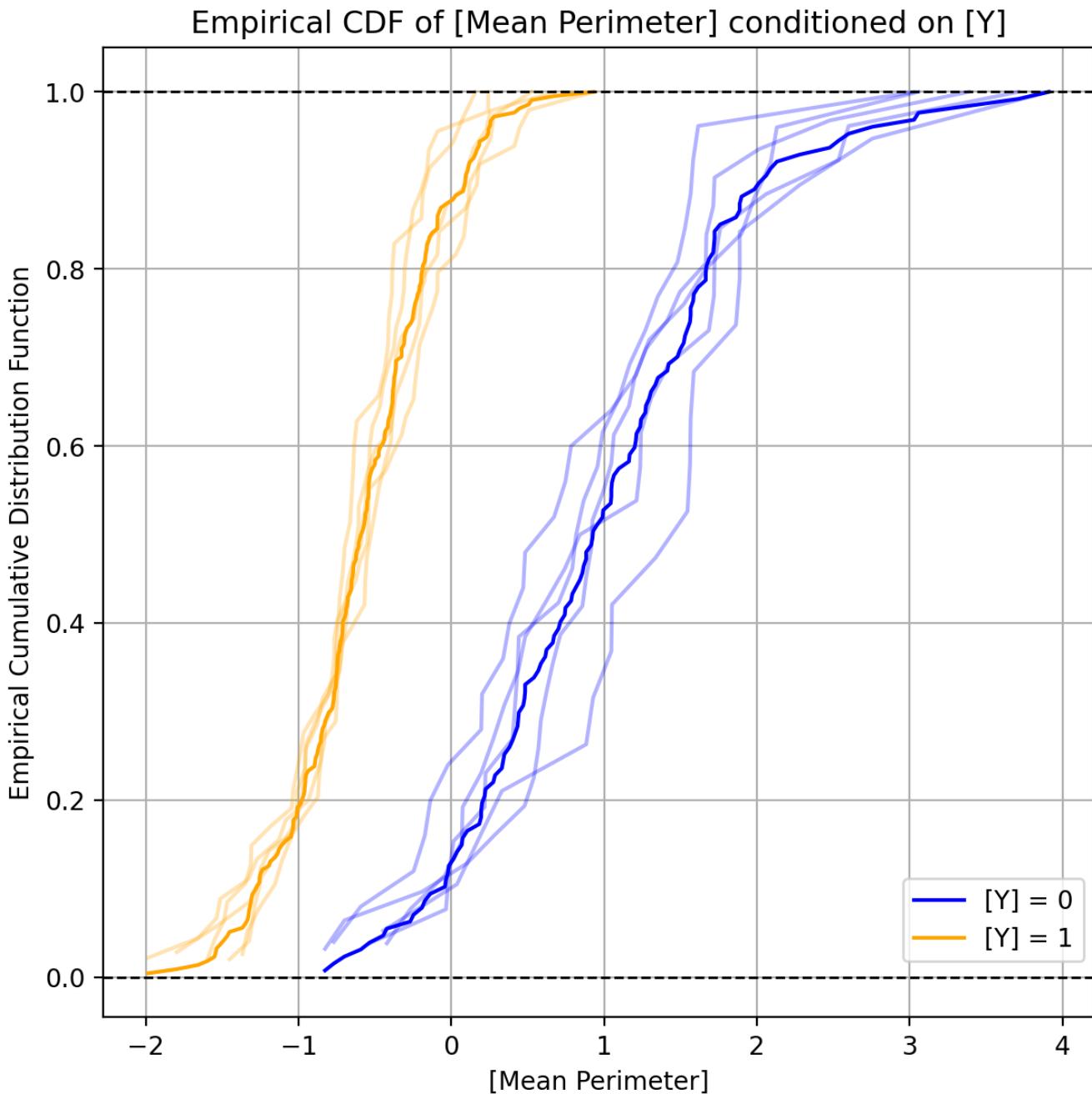
Kernel Density Plot of [Mean Perimeter] by [Y].  
Distributions by level are significantly different at the 95% level.



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

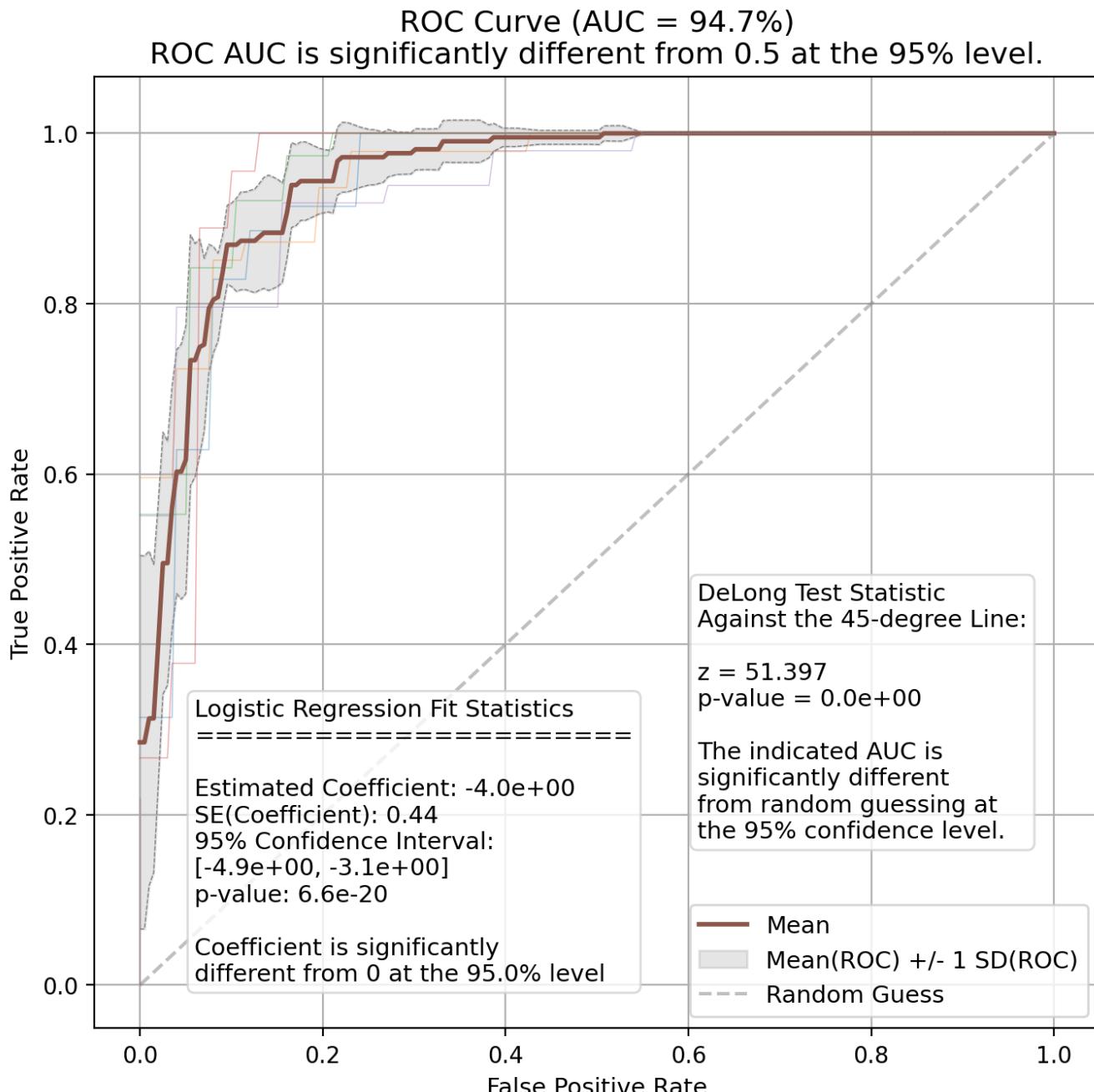
### Mean Perimeter - Empirical CDF Plot



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

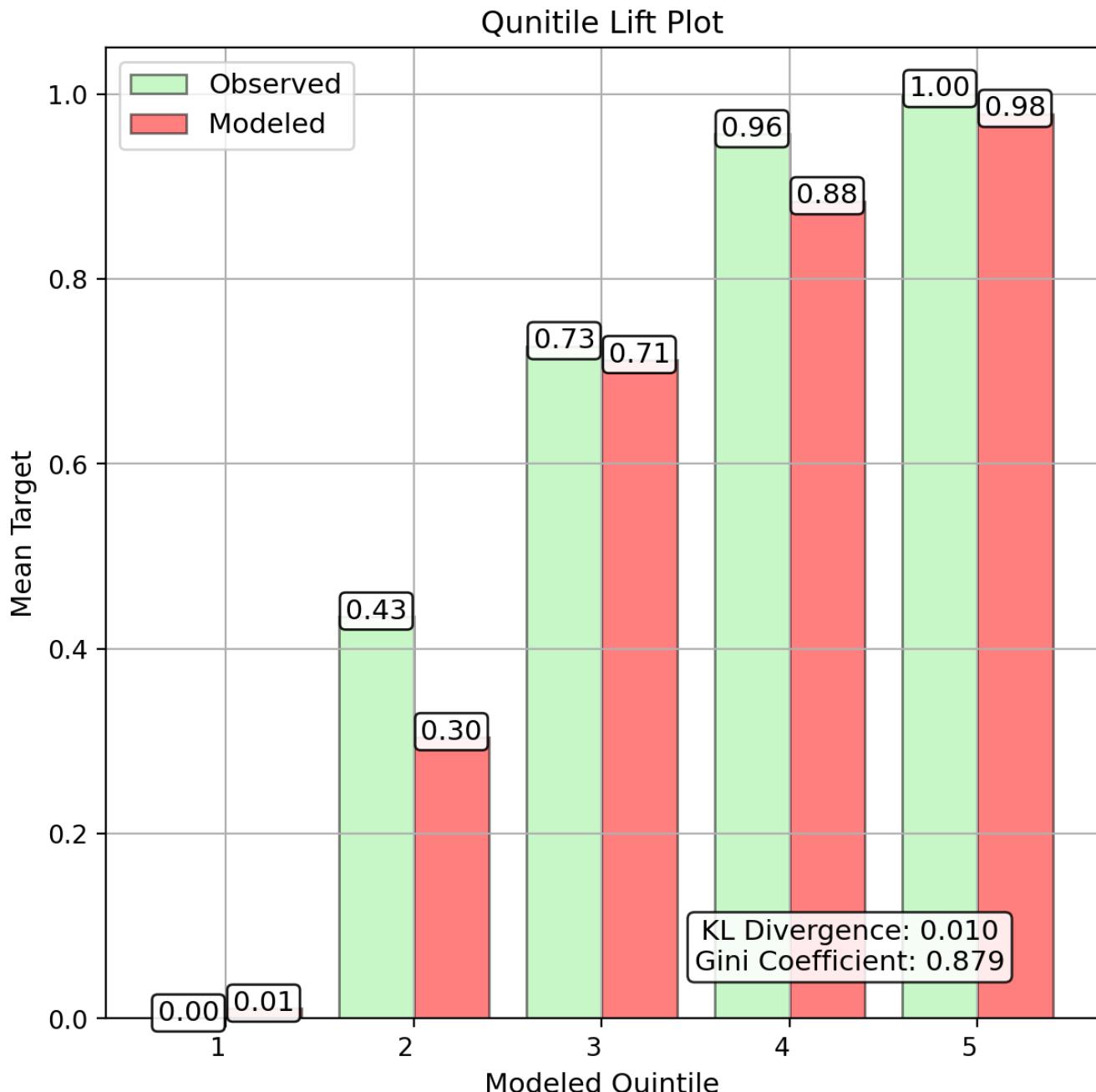


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

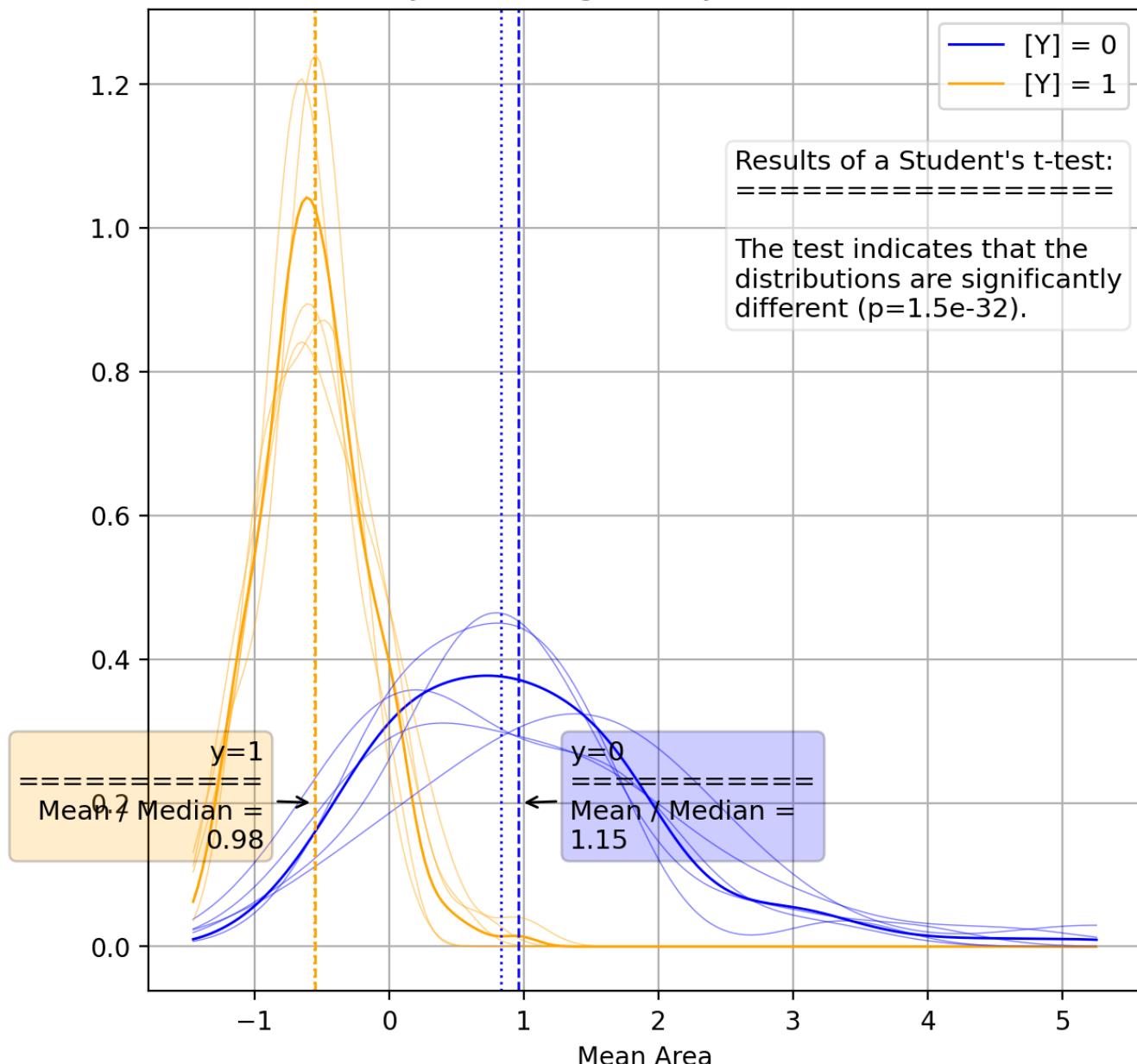
### Mean Area - Results

	Fitted Coef.	Fitted p-Value	Fitted Std. Err.	Conf. Int. Lower	Conf. Int. Upper	Test Accuracy	Test AUC	Test F1	Test Precision	Test Recall	Test MCC
<b>Fold-1</b>	-4.41	6.9e-17	0.528	-5.44	-3.37	88.3%	86.0%	90.9%	83.3%	100.0%	77.5%
<b>Fold-2</b>	-4.19	7.2e-17	0.502	-5.18	-3.21	86.3%	85.1%	89.4%	89.4%	89.4%	70.1%
<b>Fold-3</b>	-4.14	8.6e-18	0.481	-5.08	-3.19	86.0%	85.5%	89.2%	91.7%	86.8%	69.4%
<b>Fold-4</b>	-4.04	4.6e-16	0.497	-5.01	-3.06	93.4%	92.4%	94.6%	91.7%	97.8%	86.5%
<b>Fold-5</b>	-4.43	3.7e-16	0.544	-5.50	-3.36	84.0%	84.1%	87.2%	91.1%	83.7%	66.3%
<b>mean</b>	-4.23	1.4e-20	0.455	-5.13	-3.34	88.6%	86.7%	91.2%	88.2%	94.4%	75.5%
<b>std</b>	0.17	2.0e-16	0.025	0.22	0.13	3.6%	3.3%	2.8%	3.5%	7.1%	8.1%

# Univariate Report

## Mean Area - Kernel Density Plot

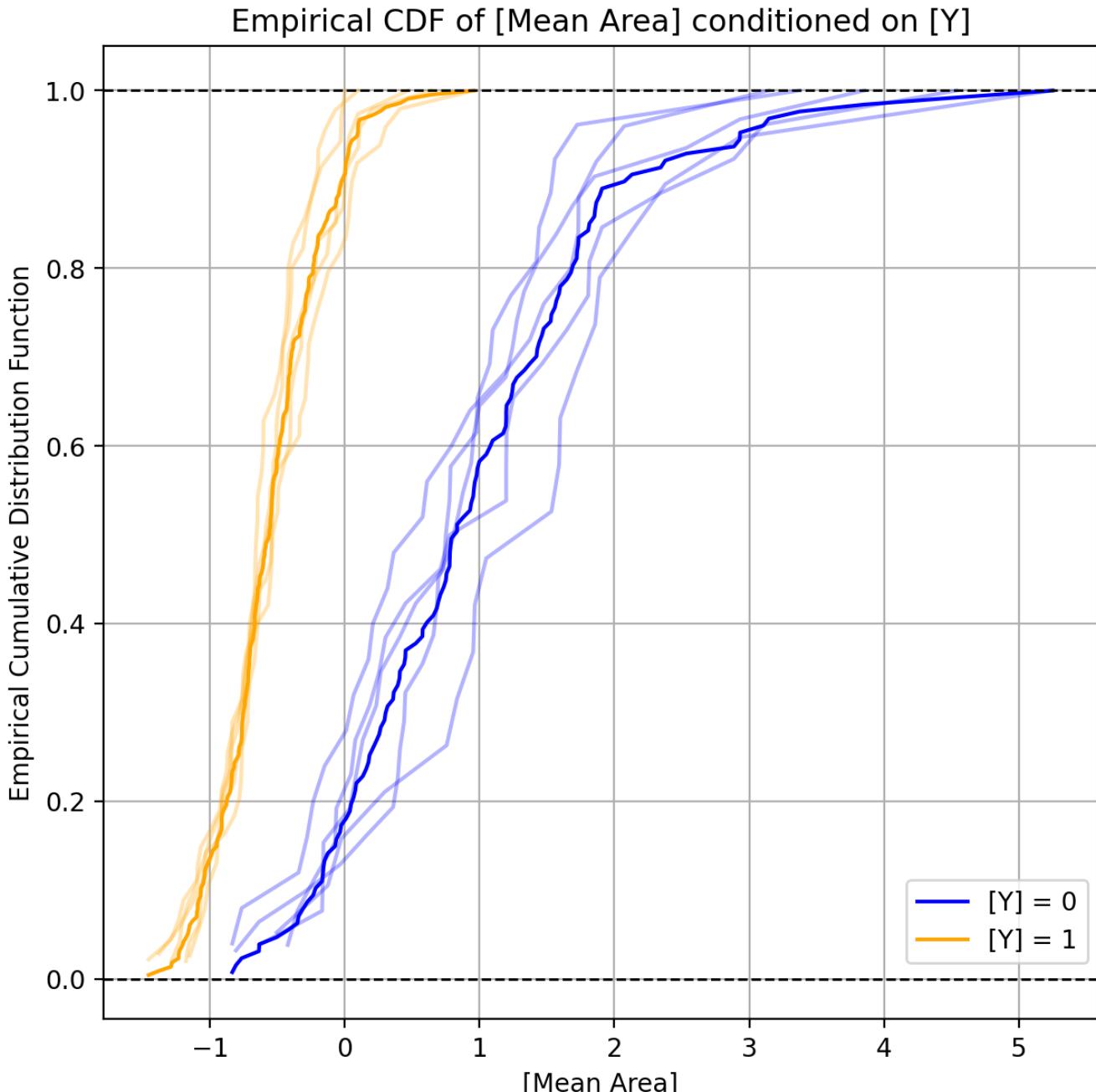
Kernel Density Plot of [Mean Area] by [Y].  
Distributions by level are significantly different at the 95% level.



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

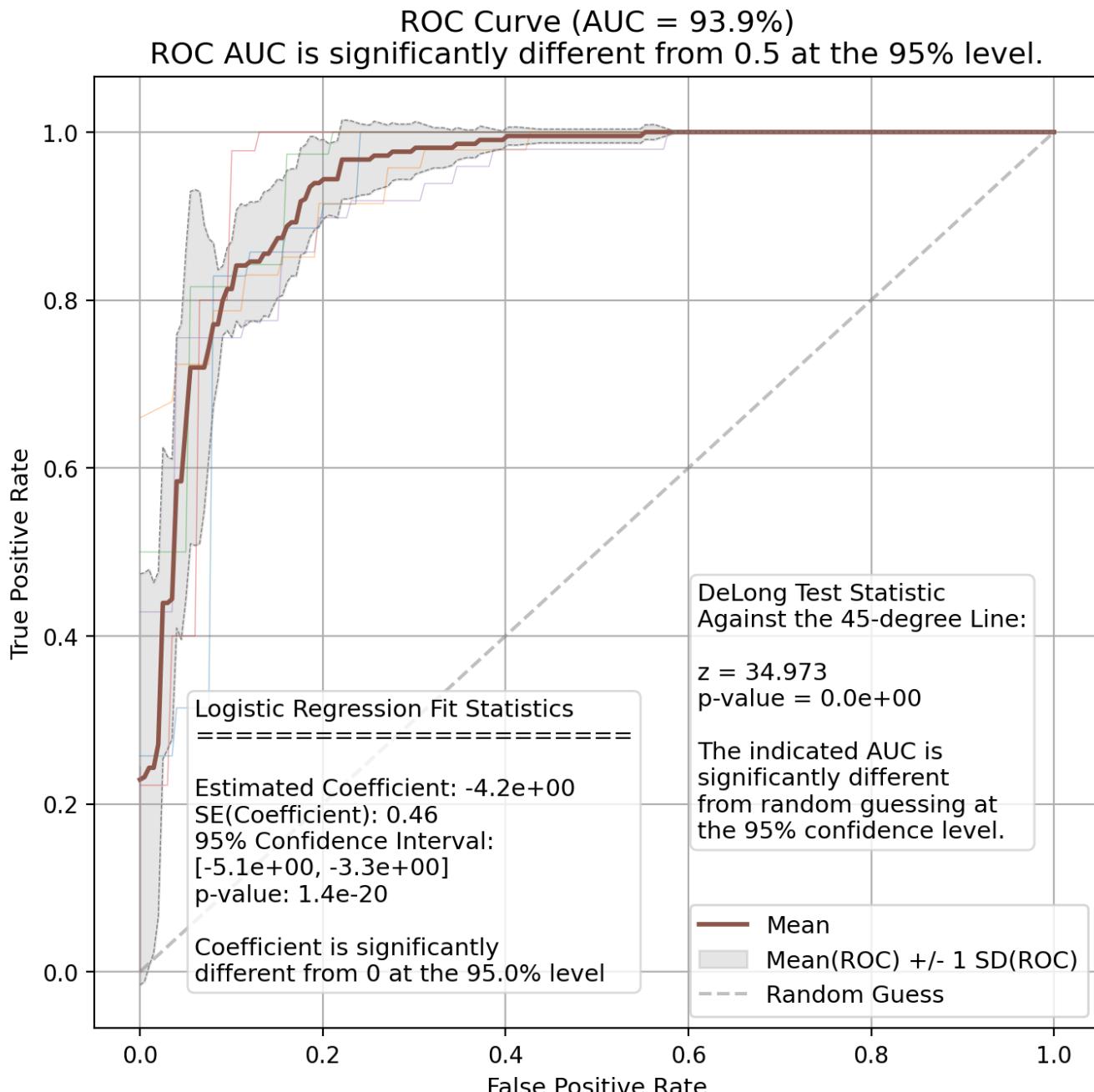
Mean Area - Empirical CDF Plot



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

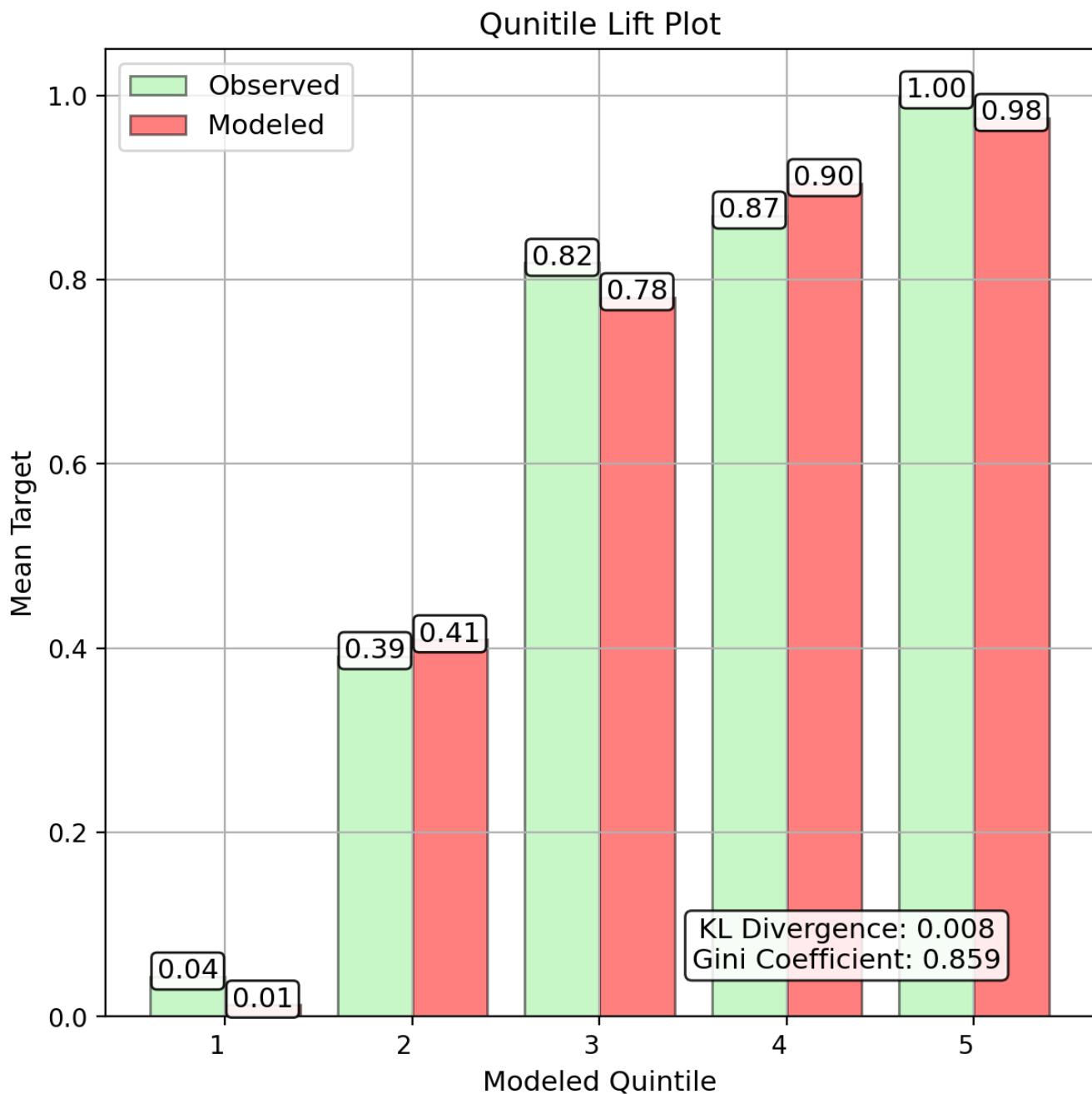


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

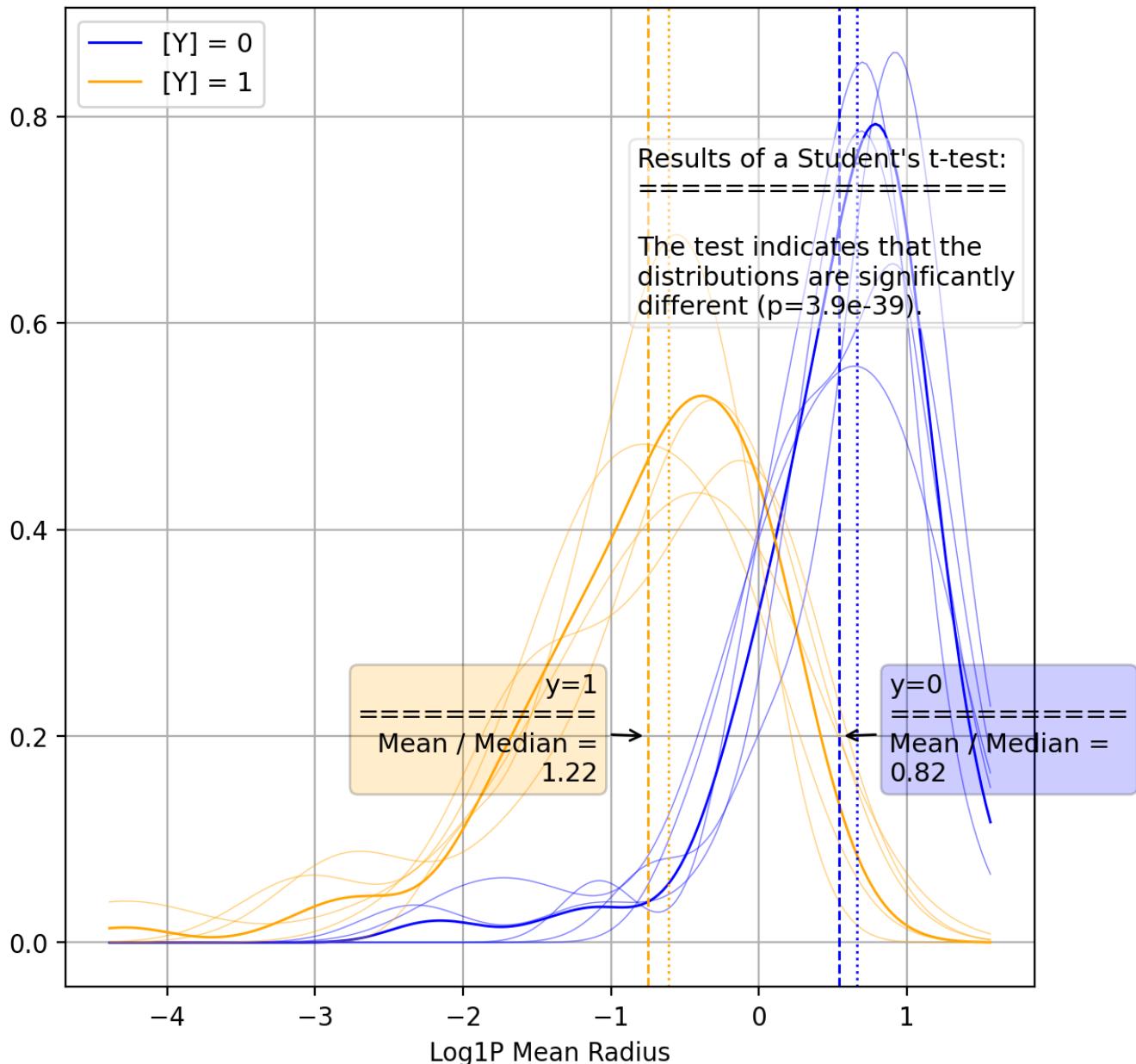
### Log1P Mean Radius - Results

	Fitted Coef.	Fitted p-Value	Fitted Std. Err.	Conf. Int. Lower	Conf. Int. Upper	Test Accuracy	Test AUC	Test F1	Test Precision	Test Recall	Test MCC
<b>Fold-1</b>	-3.23	1.8e-15	0.406	-4.03	-2.44	83.0%	83.1%	83.6%	85.2%	82.1%	66.0%
<b>Fold-2</b>	-2.84	3.4e-15	0.361	-3.55	-2.14	79.4%	81.3%	80.0%	92.9%	70.3%	62.0%
<b>Fold-3</b>	-2.88	8.8e-16	0.358	-3.58	-2.18	77.6%	80.7%	78.4%	95.2%	66.7%	60.5%
<b>Fold-4</b>	-2.95	4.3e-14	0.391	-3.72	-2.19	85.1%	85.4%	85.3%	90.6%	80.6%	70.8%
<b>Fold-5</b>	-3.02	6.3e-15	0.388	-3.78	-2.26	74.6%	78.6%	74.6%	96.2%	61.0%	57.1%
<b>mean</b>	-2.98	1.7e-18	0.340	-3.65	-2.32	75.0%	76.1%	75.7%	84.8%	68.4%	51.8%
<b>std</b>	0.15	1.8e-14	0.021	0.19	0.12	4.2%	2.6%	4.2%	4.4%	9.1%	5.3%

## Univariate Report

### Log1P Mean Radius - Kernel Density Plot

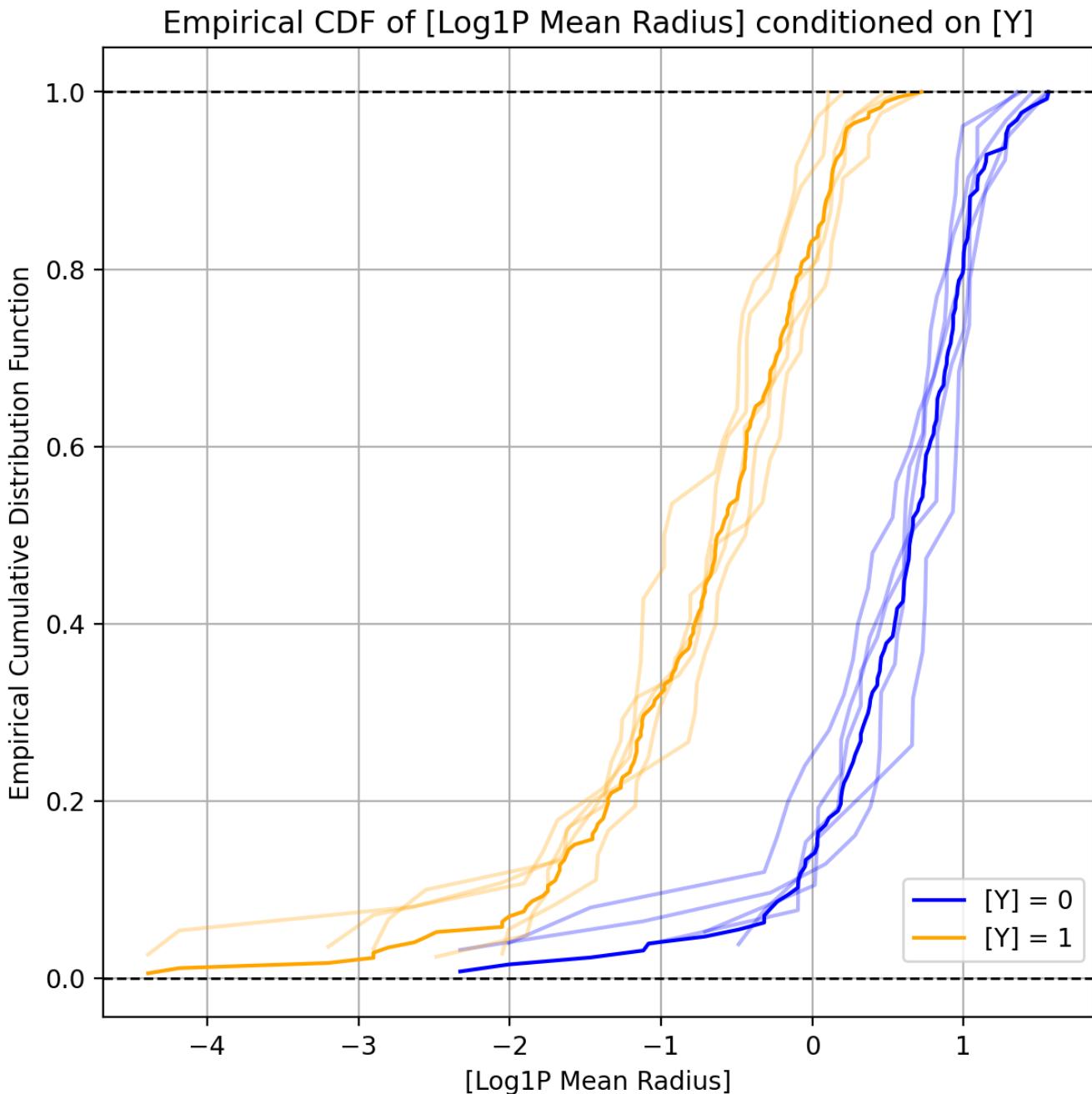
Kernel Density Plot of [Log1P Mean Radius] by [Y].  
Distributions by level are significantly different at the 95% level.



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

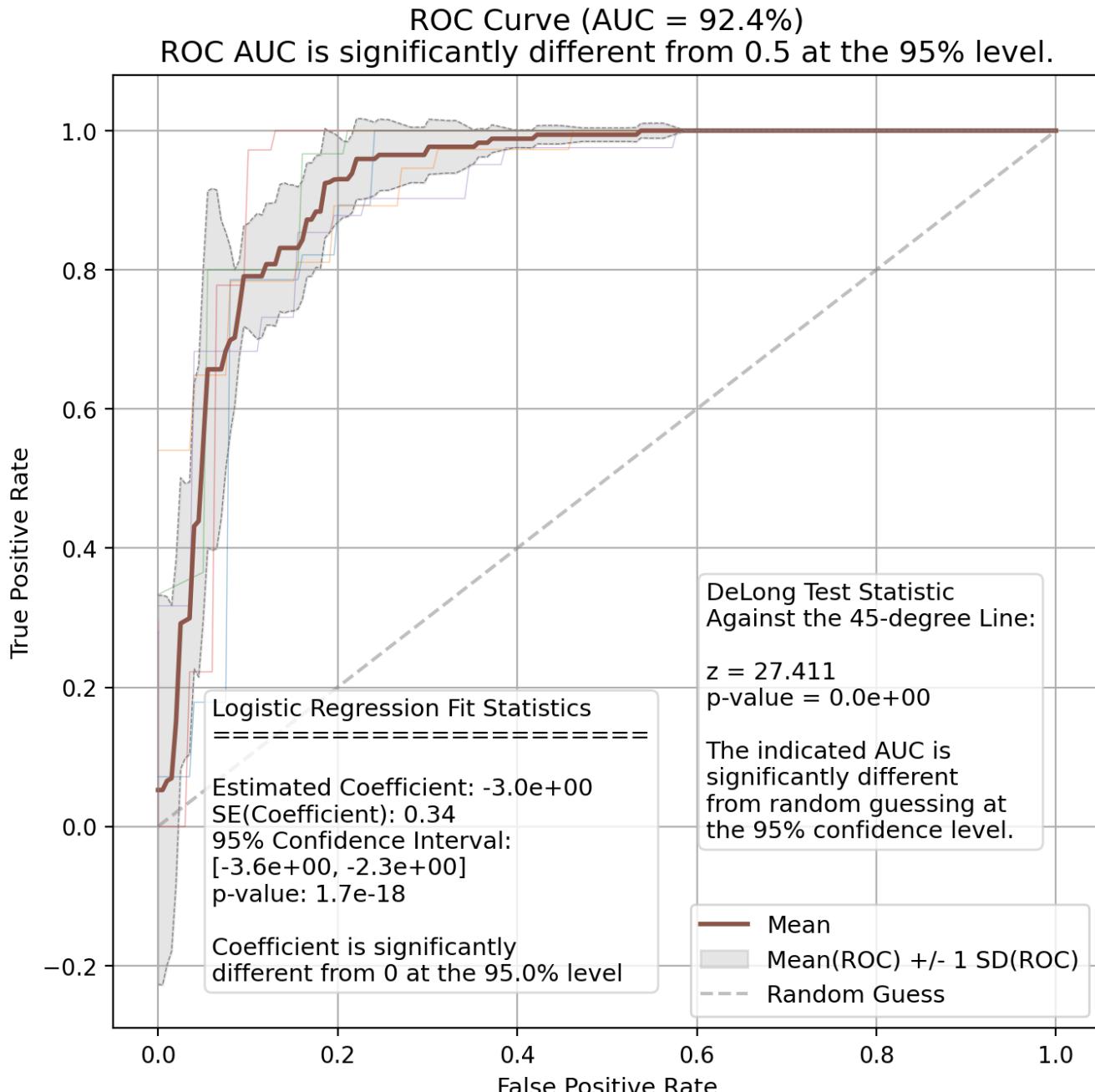
Log1P Mean Radius - Empirical CDF Plot



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

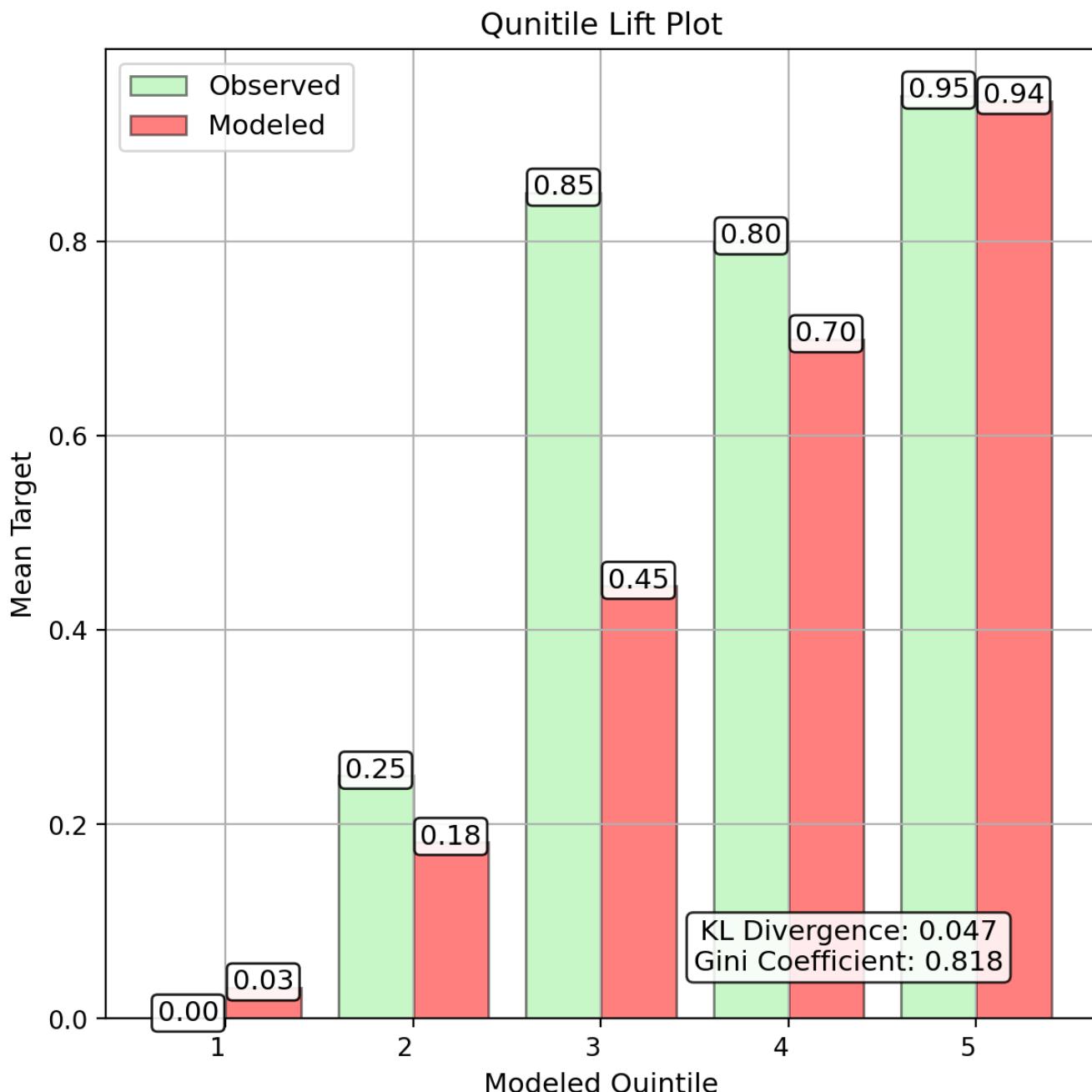


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

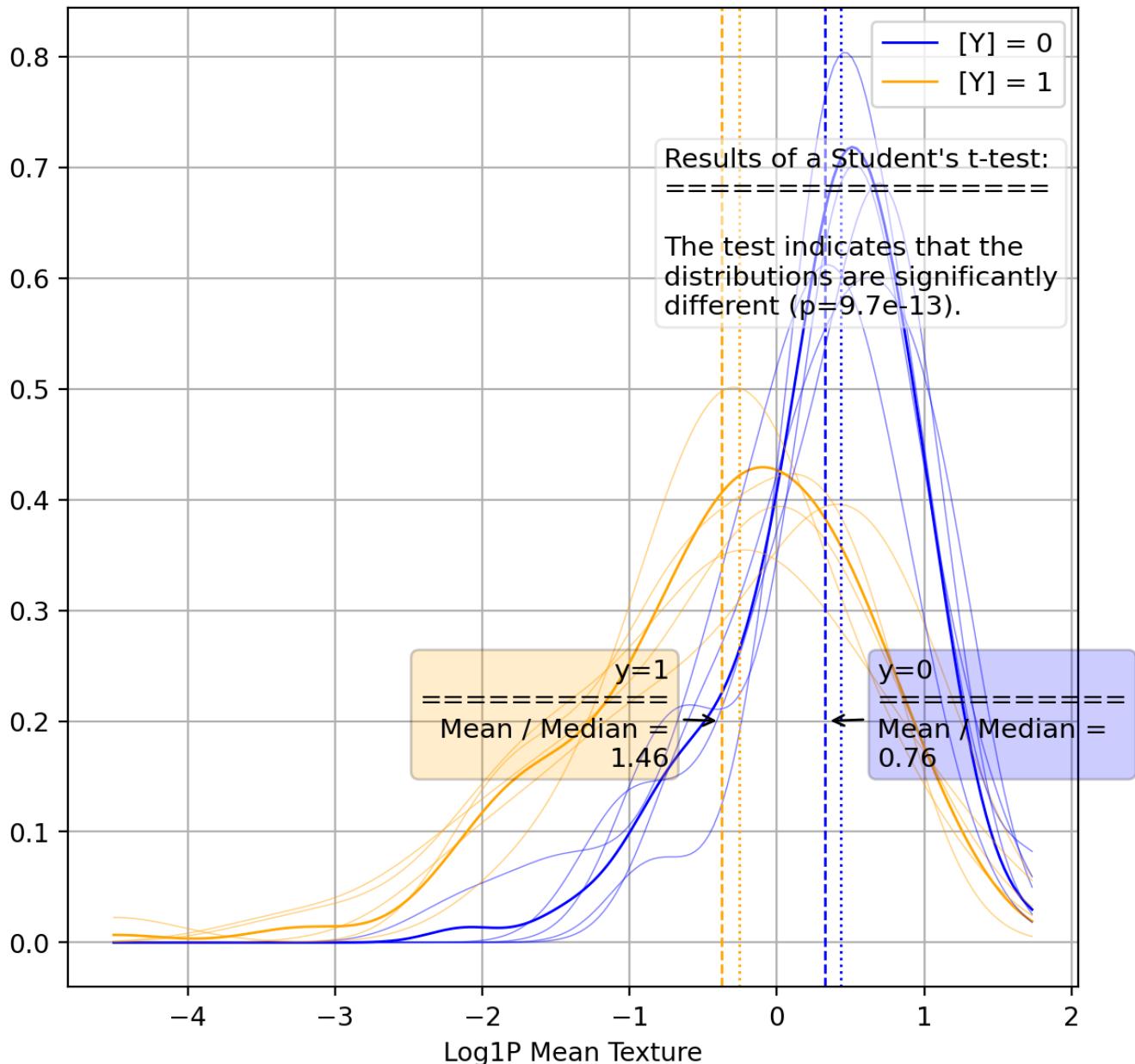
### Log1P Mean Texture - Results

	Fitted Coef.	Fitted p-Value	Fitted	Conf.	Conf.	Test Accuracy	Test AUC	Test F1	Test Precision	Test Recall	Test MCC
			Std. Err.	Int. Lower	Int. Upper						
<b>Fold-1</b>	-1.15	1.2e-08	0.202	-1.55	-0.755	61.5%	62.4%	52.4%	73.3%	40.7%	27.3%
<b>Fold-2</b>	-1.07	1.0e-07	0.201	-1.46	-0.676	68.3%	70.1%	68.8%	81.5%	59.5%	40.0%
<b>Fold-3</b>	-0.99	1.4e-07	0.189	-1.37	-0.625	67.3%	72.0%	67.9%	90.0%	54.5%	43.6%
<b>Fold-4</b>	-1.19	7.0e-09	0.205	-1.59	-0.786	69.6%	69.1%	72.7%	71.8%	73.7%	38.3%
<b>Fold-5</b>	-1.10	5.6e-08	0.202	-1.50	-0.702	69.6%	70.8%	70.2%	80.0%	62.5%	41.5%
<b>mean</b>	-1.10	7.6e-10	0.179	-1.45	-0.749	65.3%	66.6%	64.6%	75.6%	56.4%	33.3%
<b>std</b>	0.07	5.8e-08	0.006	0.09	0.064	3.3%	3.8%	8.0%	7.3%	12.0%	6.4%

## Univariate Report

### Log1P Mean Texture - Kernel Density Plot

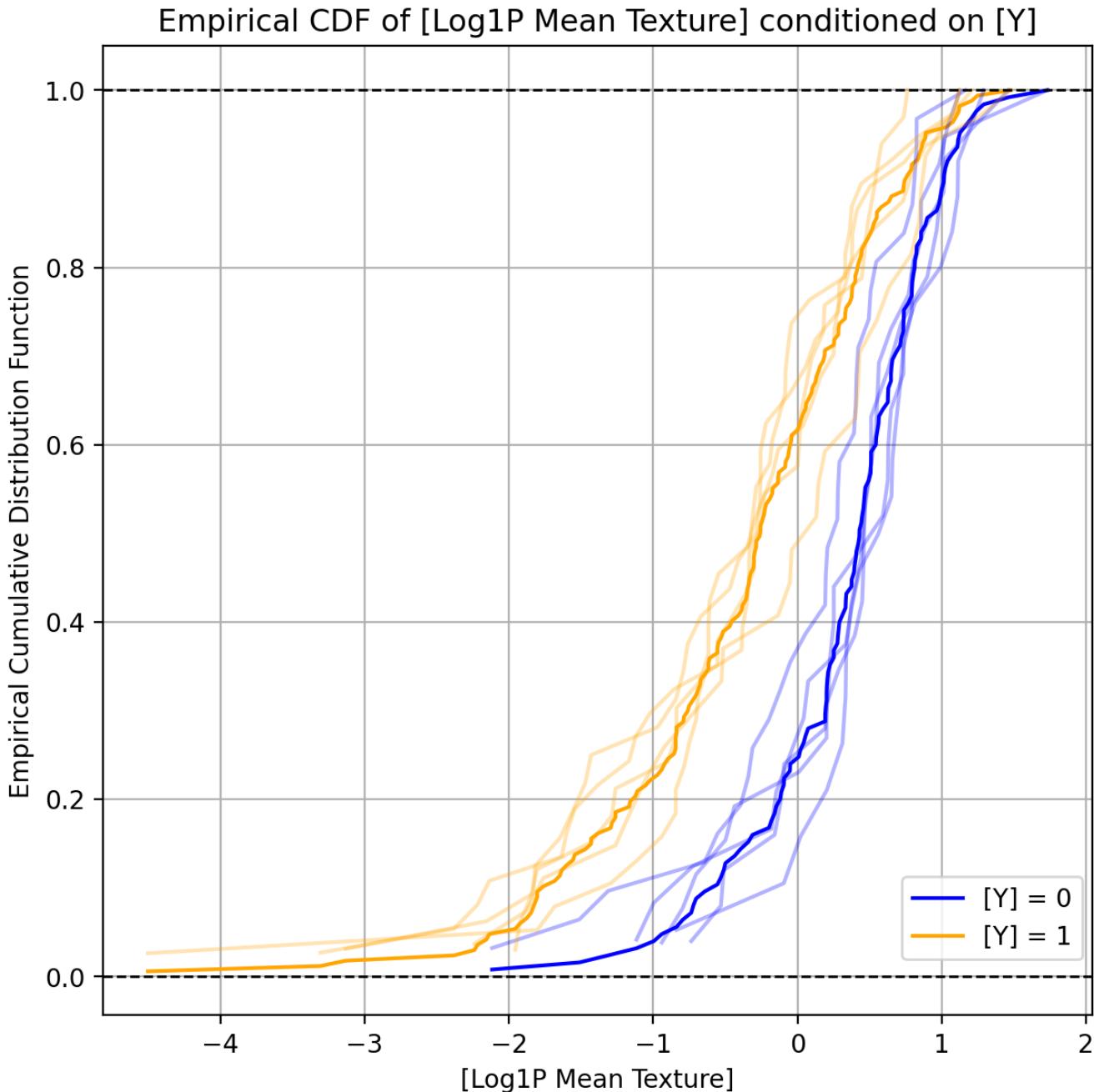
Kernel Density Plot of [Log1P Mean Texture] by [Y].  
Distributions by level are significantly different at the 95% level.



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

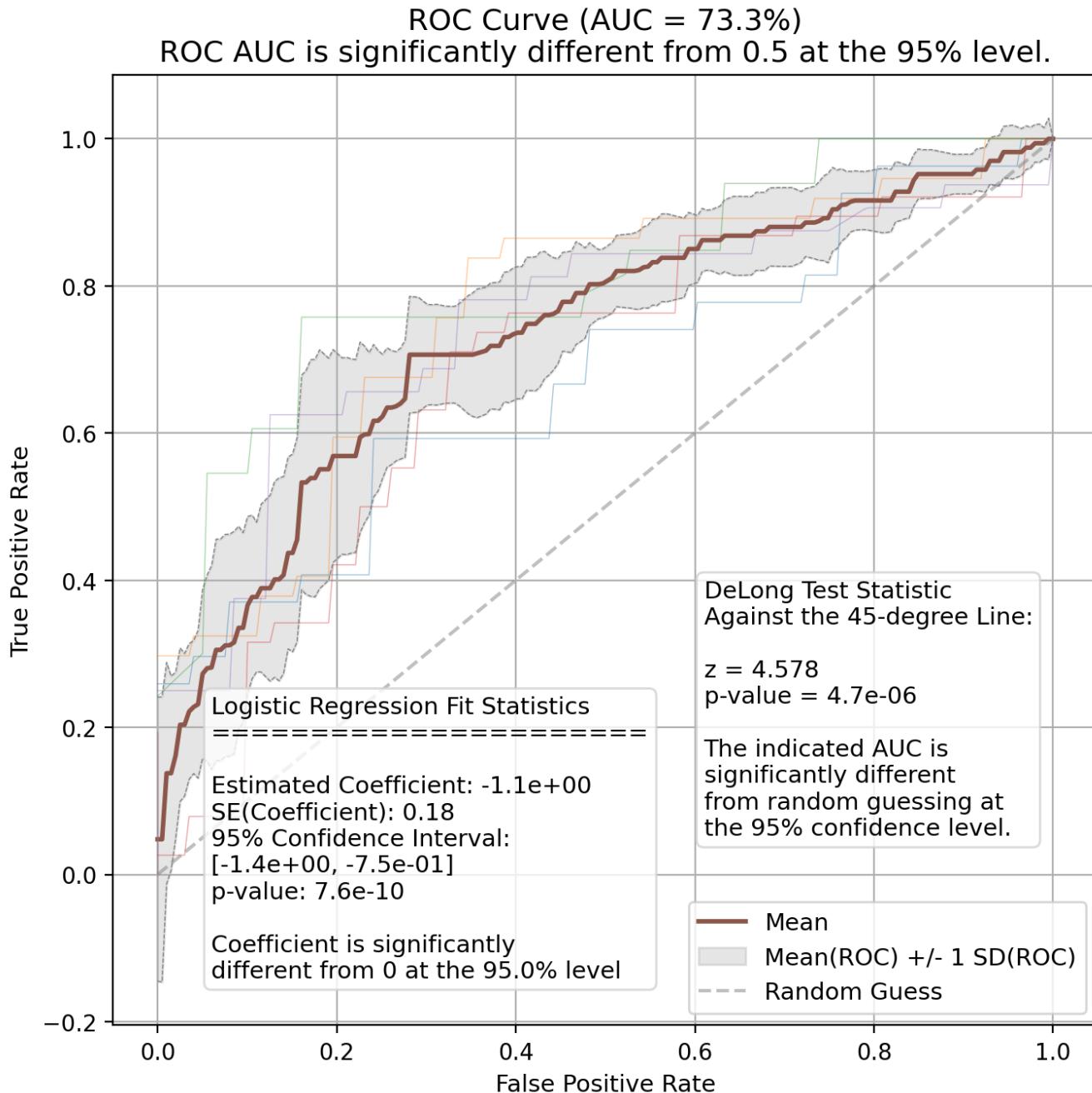
### Log1P Mean Texture - Empirical CDF Plot



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

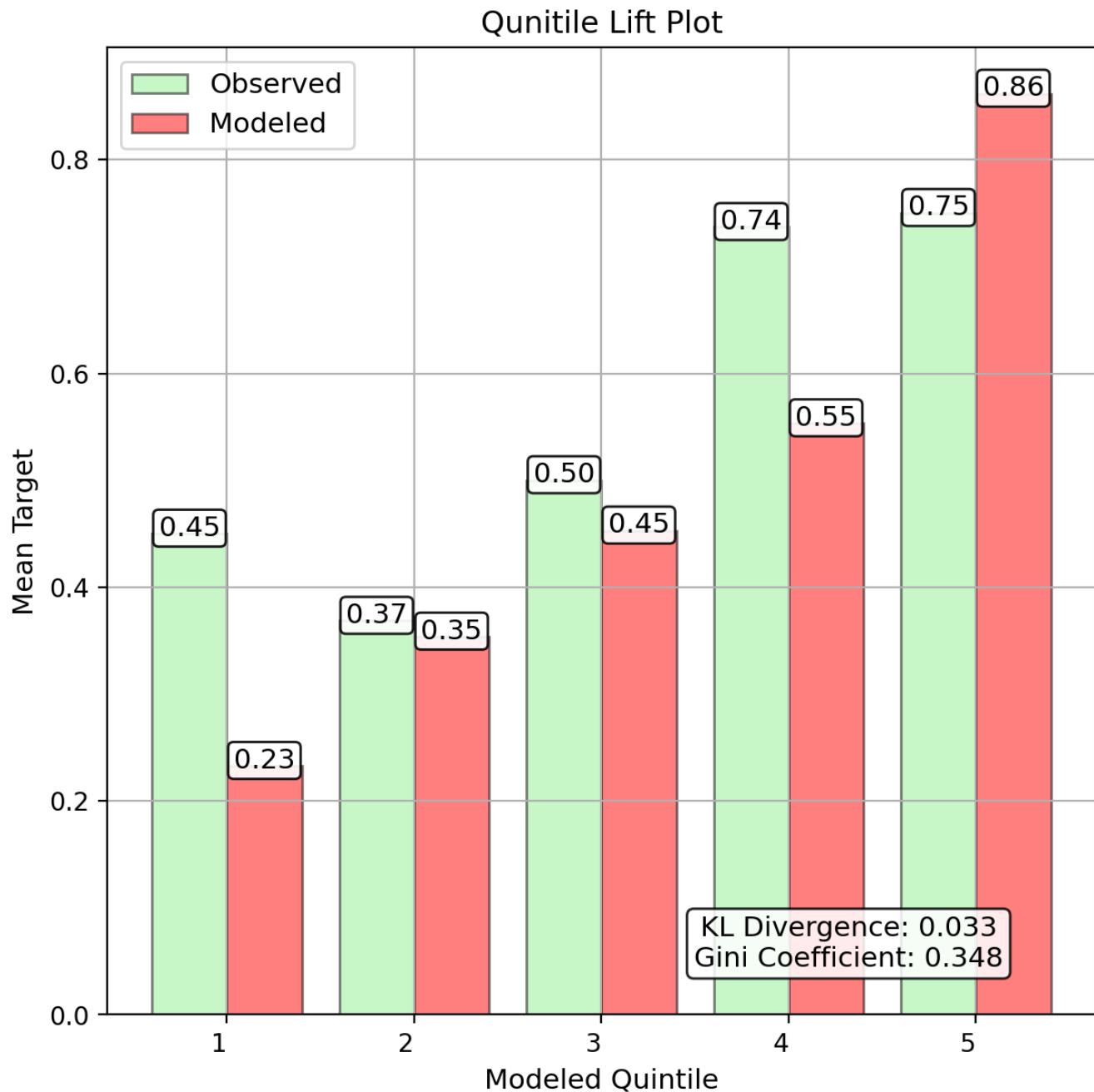


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

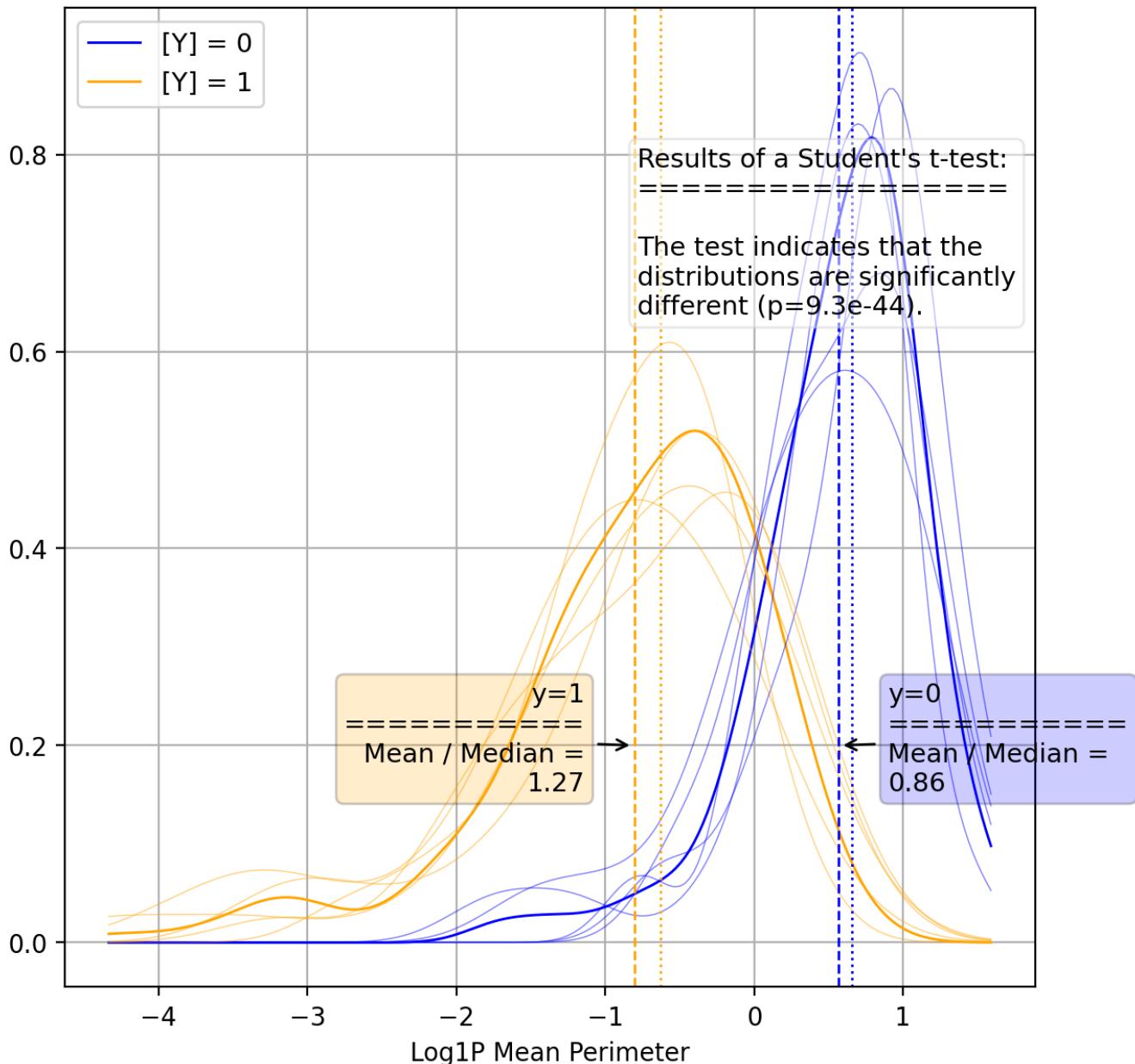
### Log1P Mean Perimeter - Results

	Fitted Coef.	Fitted p-Value	Fitted Std. Err.	Conf. Int. Lower	Conf. Int. Upper	Test Accuracy	Test AUC	Test F1	Test Precision	Test Recall	Test MCC
<b>Fold-1</b>	-3.55	2.1e-15	0.447	-4.43	-2.67	85.2%	85.1%	86.2%	86.2%	86.2%	70.2%
<b>Fold-2</b>	-3.25	2.9e-15	0.412	-4.06	-2.45	79.0%	80.9%	79.4%	92.6%	69.4%	61.5%
<b>Fold-3</b>	-3.31	9.6e-16	0.412	-4.11	-2.50	75.5%	79.0%	76.0%	95.0%	63.3%	57.6%
<b>Fold-4</b>	-3.42	5.0e-14	0.454	-4.31	-2.53	88.2%	88.7%	88.6%	93.9%	83.8%	77.1%
<b>Fold-5</b>	-3.39	6.7e-15	0.435	-4.25	-2.54	77.6%	81.0%	78.3%	96.4%	65.9%	61.3%
<b>mean</b>	-3.38	1.9e-18	0.386	-4.14	-2.62	76.2%	77.8%	76.5%	88.6%	67.2%	55.5%
<b>std</b>	0.11	2.1e-14	0.020	0.15	0.08	5.4%	3.9%	5.4%	4.0%	10.6%	8.0%

## Univariate Report

### Log1P Mean Perimeter - Kernel Density Plot

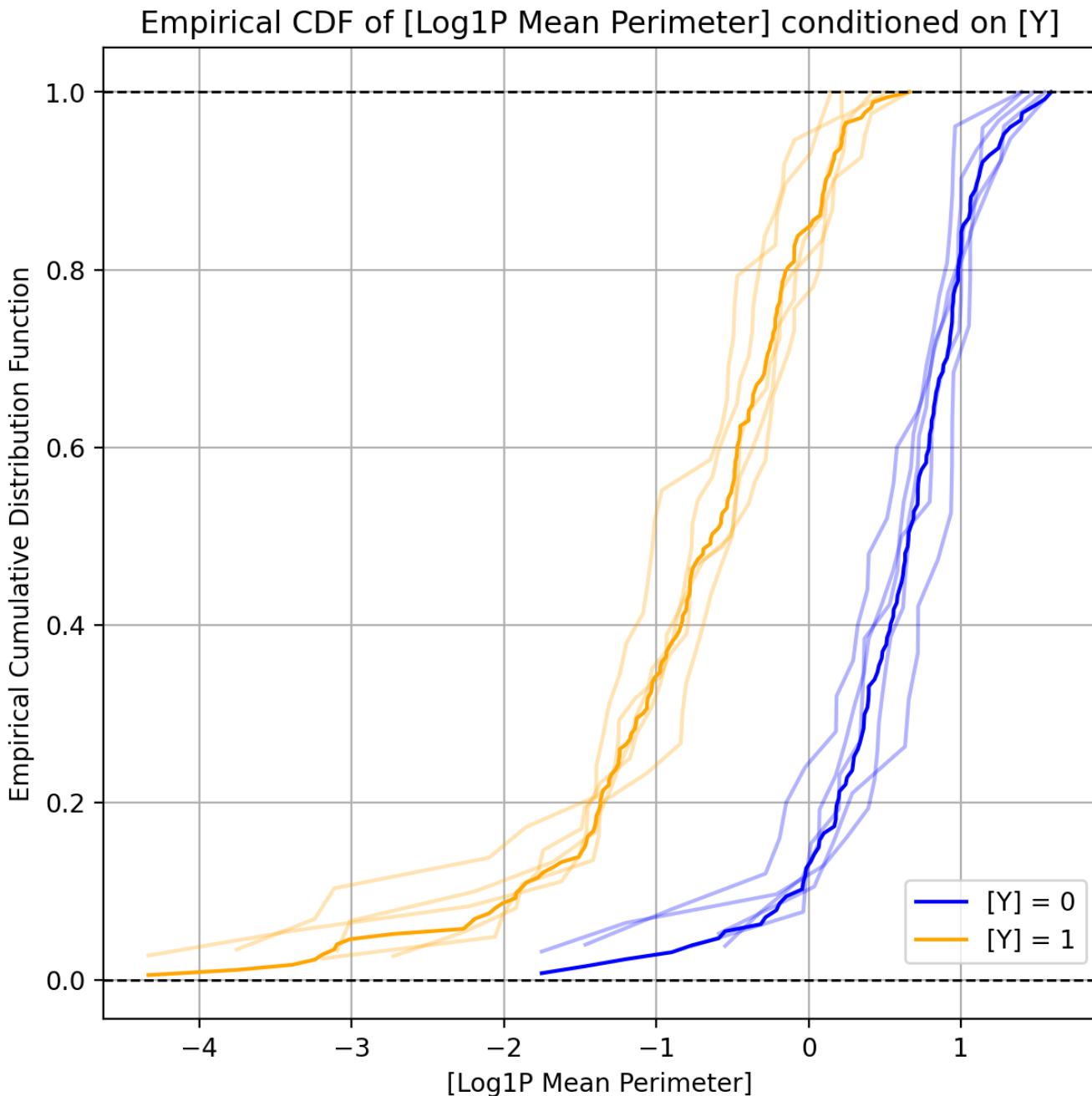
Kernel Density Plot of [Log1P Mean Perimeter] by [Y].  
Distributions by level are significantly different at the 95% level.



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

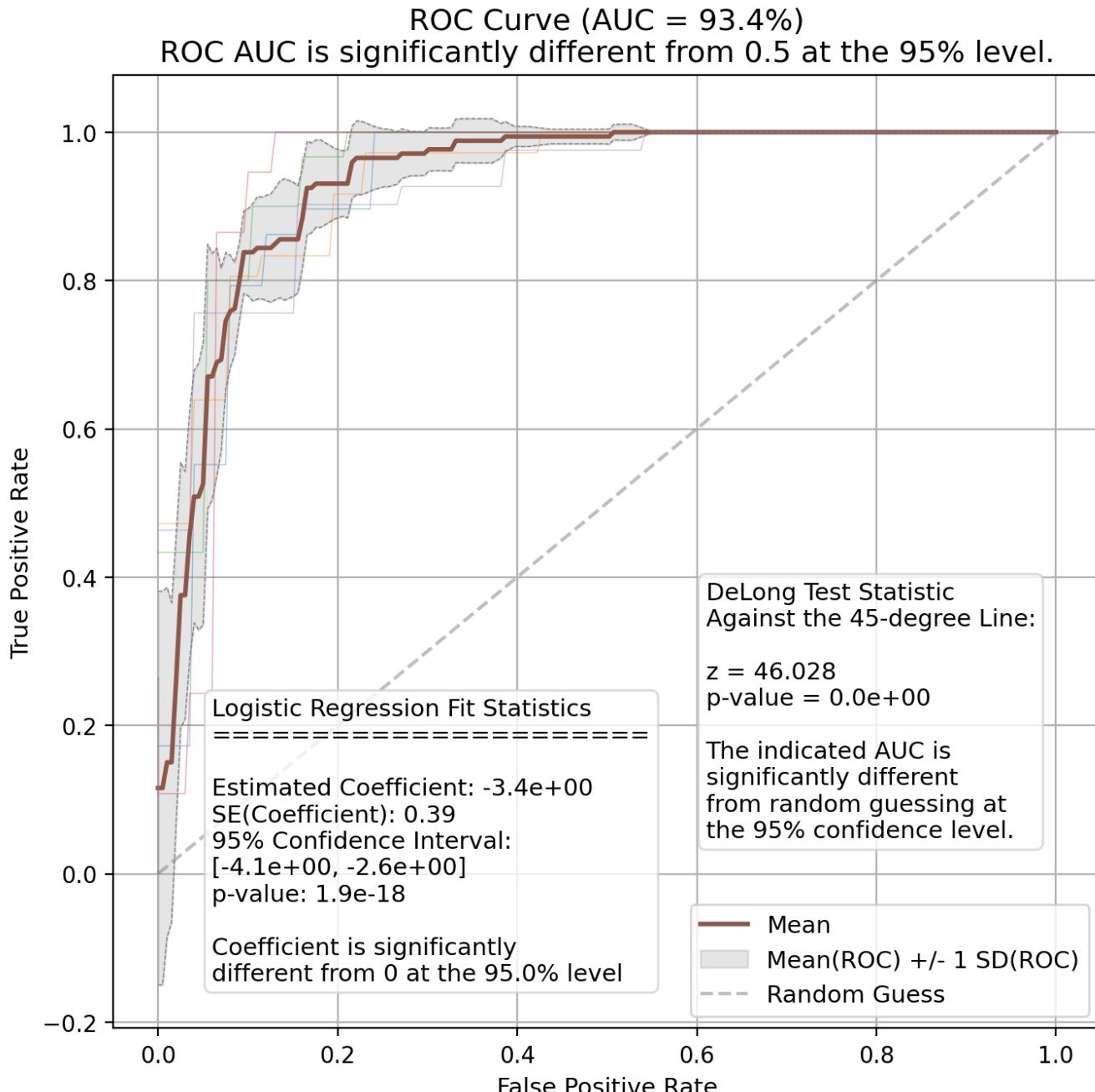
### Log1P Mean Perimeter - Empirical CDF Plot



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

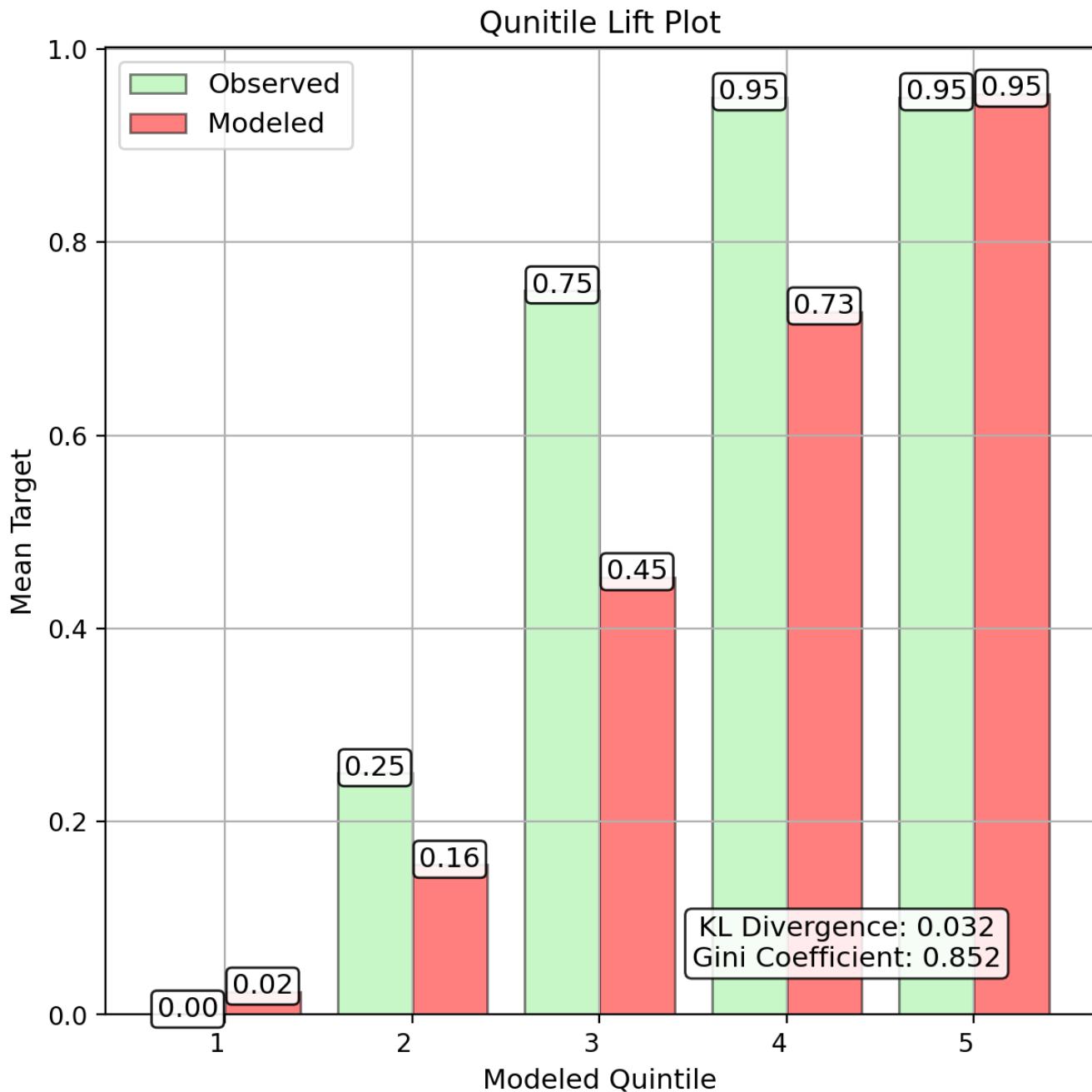


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

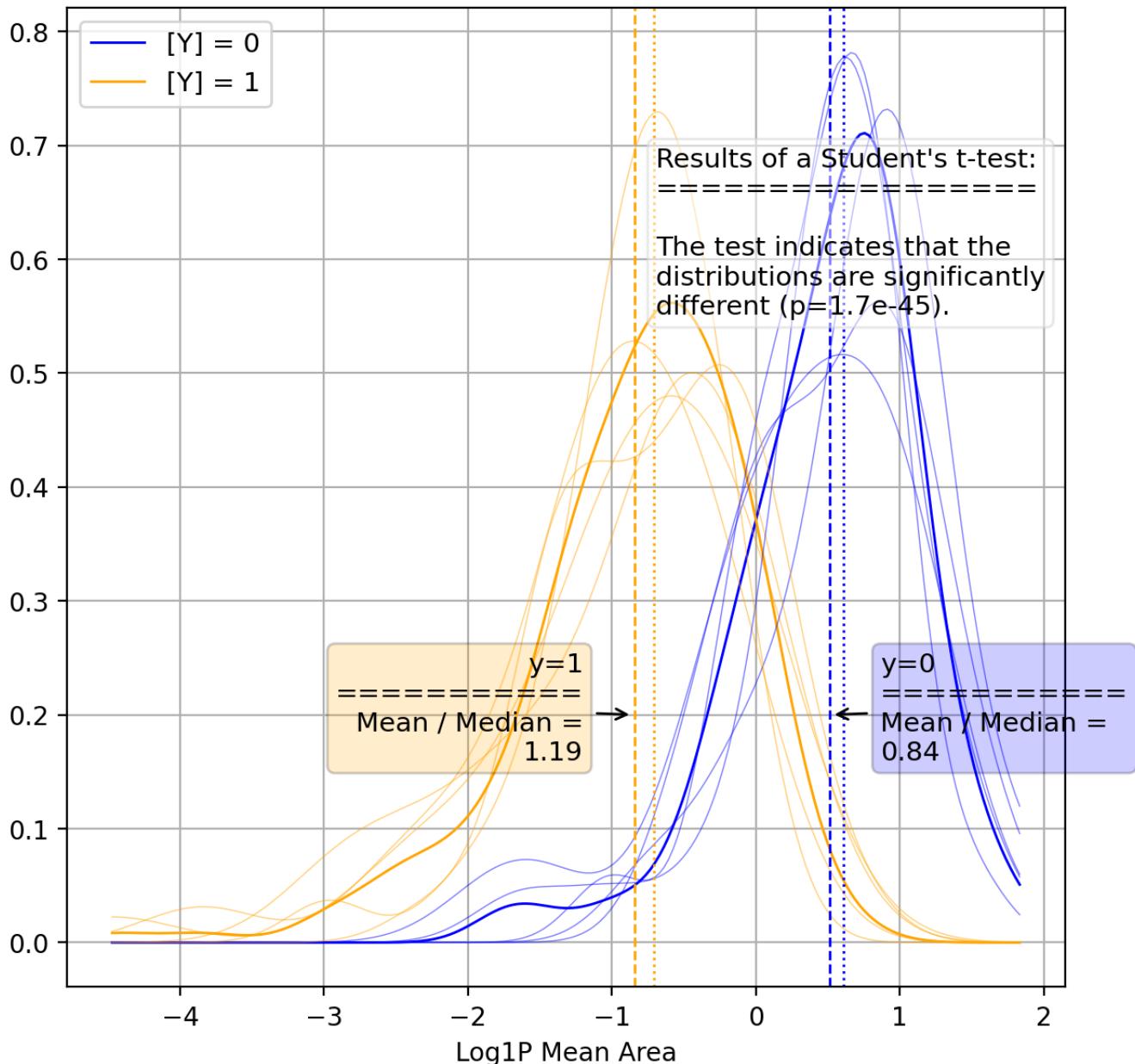
### Log1P Mean Area - Results

	Fitted Coef.	Fitted p-Value	Fitted Std. Err.	Conf. Int. Lower	Conf. Int. Upper	Test Accuracy	Test AUC	Test F1	Test Precision	Test Recall	Test MCC
<b>Fold-1</b>	-3.68	1.0e-15	0.459	-4.58	-2.79	83.9%	83.5%	85.7%	84.4%	87.1%	67.4%
<b>Fold-2</b>	-3.20	7.5e-16	0.397	-3.97	-2.42	80.3%	82.4%	81.7%	93.5%	72.5%	63.4%
<b>Fold-3</b>	-3.18	1.7e-16	0.385	-3.93	-2.42	81.1%	84.1%	83.3%	96.2%	73.5%	65.5%
<b>Fold-4</b>	-3.23	9.8e-15	0.417	-4.04	-2.41	85.5%	86.0%	86.1%	91.2%	81.6%	71.5%
<b>Fold-5</b>	-3.30	1.5e-15	0.414	-4.11	-2.49	76.5%	80.2%	77.1%	96.4%	64.3%	59.7%
<b>mean</b>	-3.31	3.1e-19	0.369	-4.03	-2.58	77.1%	78.2%	78.9%	86.5%	72.6%	55.4%
<b>std</b>	0.21	4.0e-15	0.028	0.26	0.16	3.5%	2.1%	3.6%	4.9%	8.8%	4.4%

# Univariate Report

## Log1P Mean Area - Kernel Density Plot

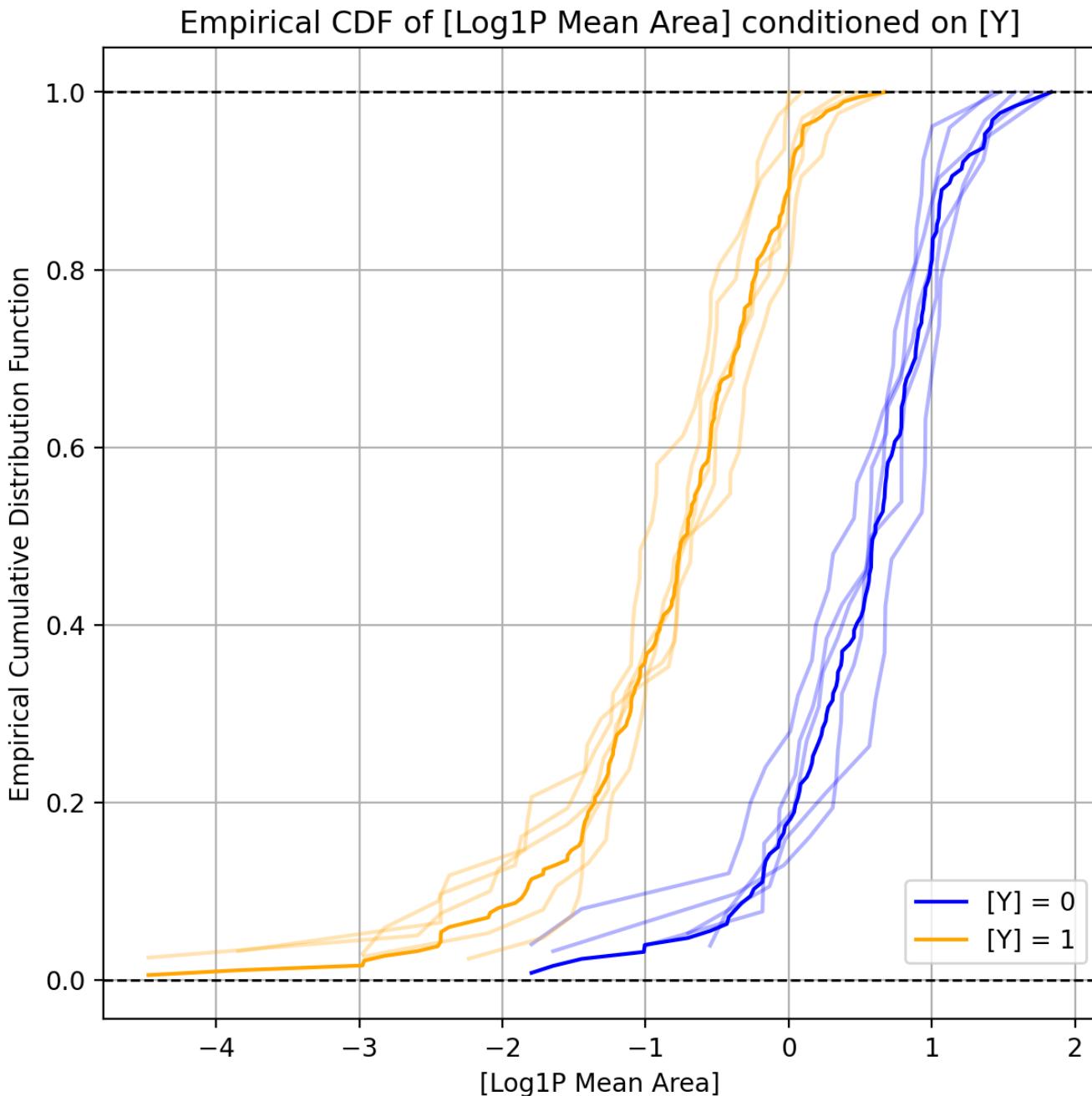
Kernel Density Plot of [Log1P Mean Area] by [Y].  
Distributions by level are significantly different at the 95% level.



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

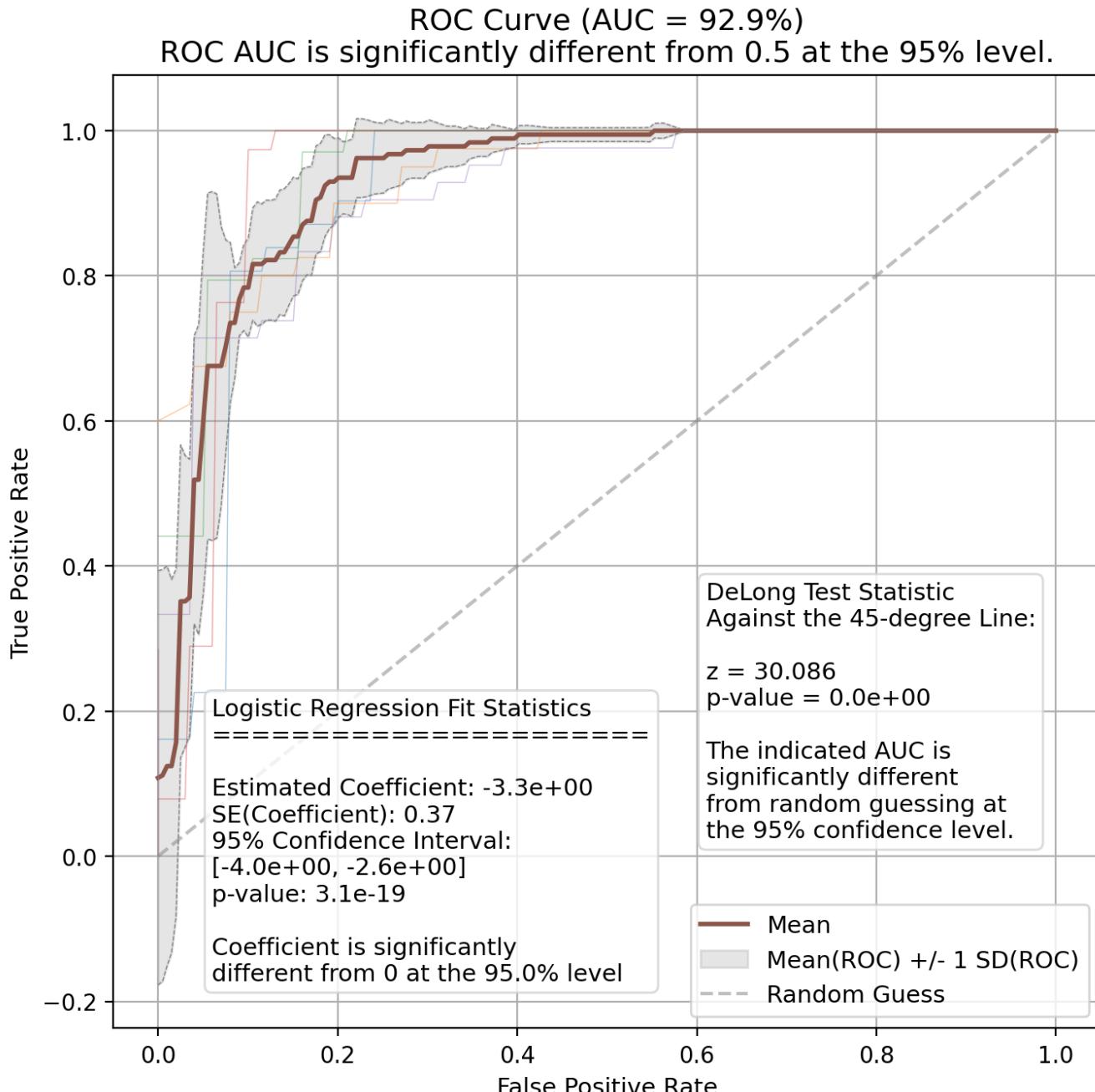
### Log1P Mean Area - Empirical CDF Plot



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

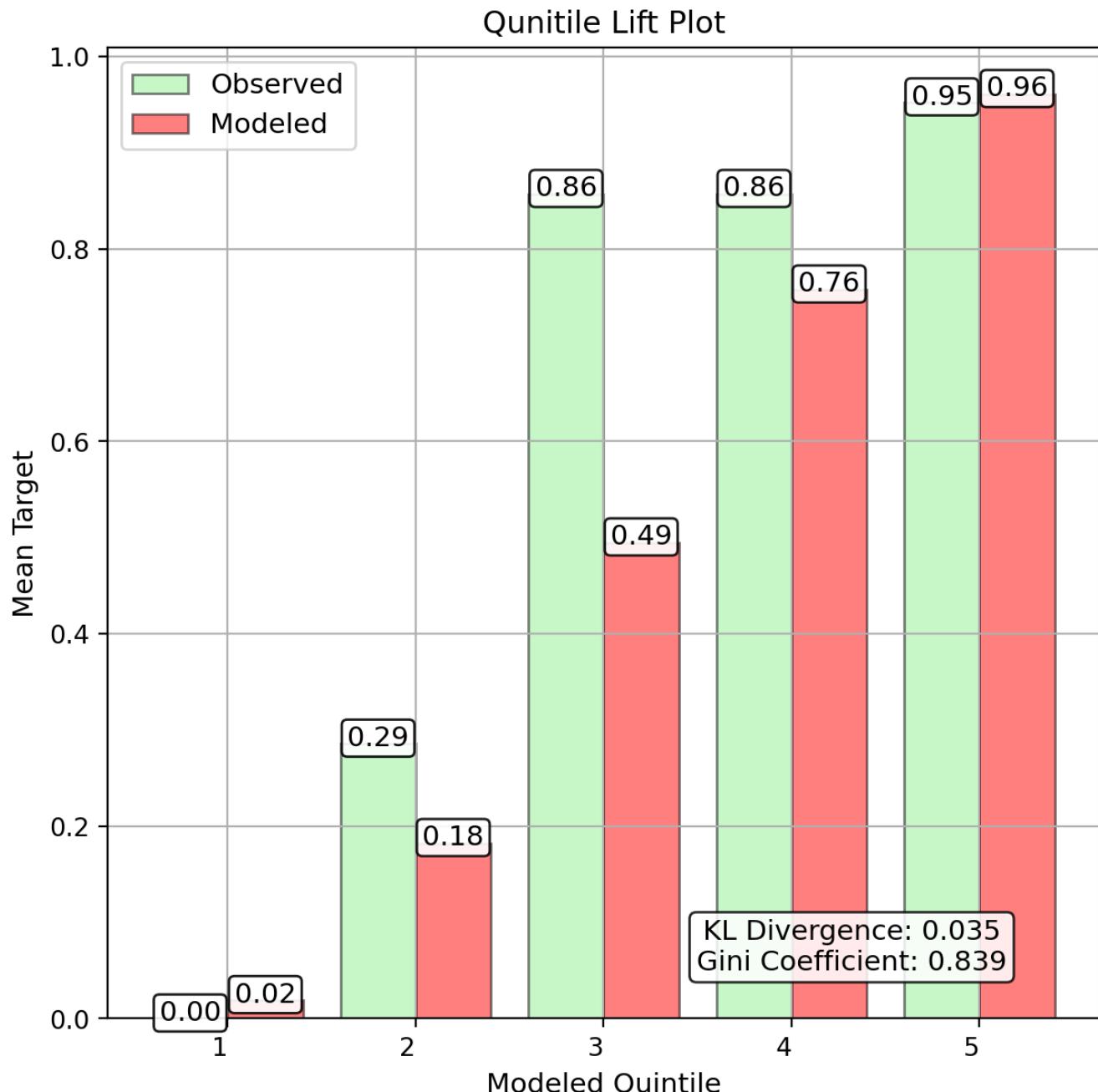


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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