Cancer Model Univariate Analysis Report

2024-02-21

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.

	Feature	Accuracy	Precision	Recall	AUC	F1	MCC	Ave.
0	Pc1	92.1%	94.3%	93.0%	91.8%	93.6%	83.3%	91.3%
1	Mean Area	88.6%	88.2%	94.4%	86.7%	91.2%	75.5%	87.4%
2	Mean Radius	86.8%	90.0%	88.7%	86.2%	89.4%	72.1%	85.5%
3	Mean Perimeter	86.8%	90.0%	88.7%	86.2%	89.4%	72.1%	85.5%
4	Mean Texture	71.9%	81.0%	71.8%	72.0%	76.1%	42.8%	69.3%
5	Mean Smoothness	67.5%	80.4%	63.4%	68.9%	70.9%	36.6%	64.6%
6	Pc2	57.0%	67.7%	59.2%	56.3%	63.2%	12.3%	52.6%

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 7 of the variables passed to the Cancer Model.

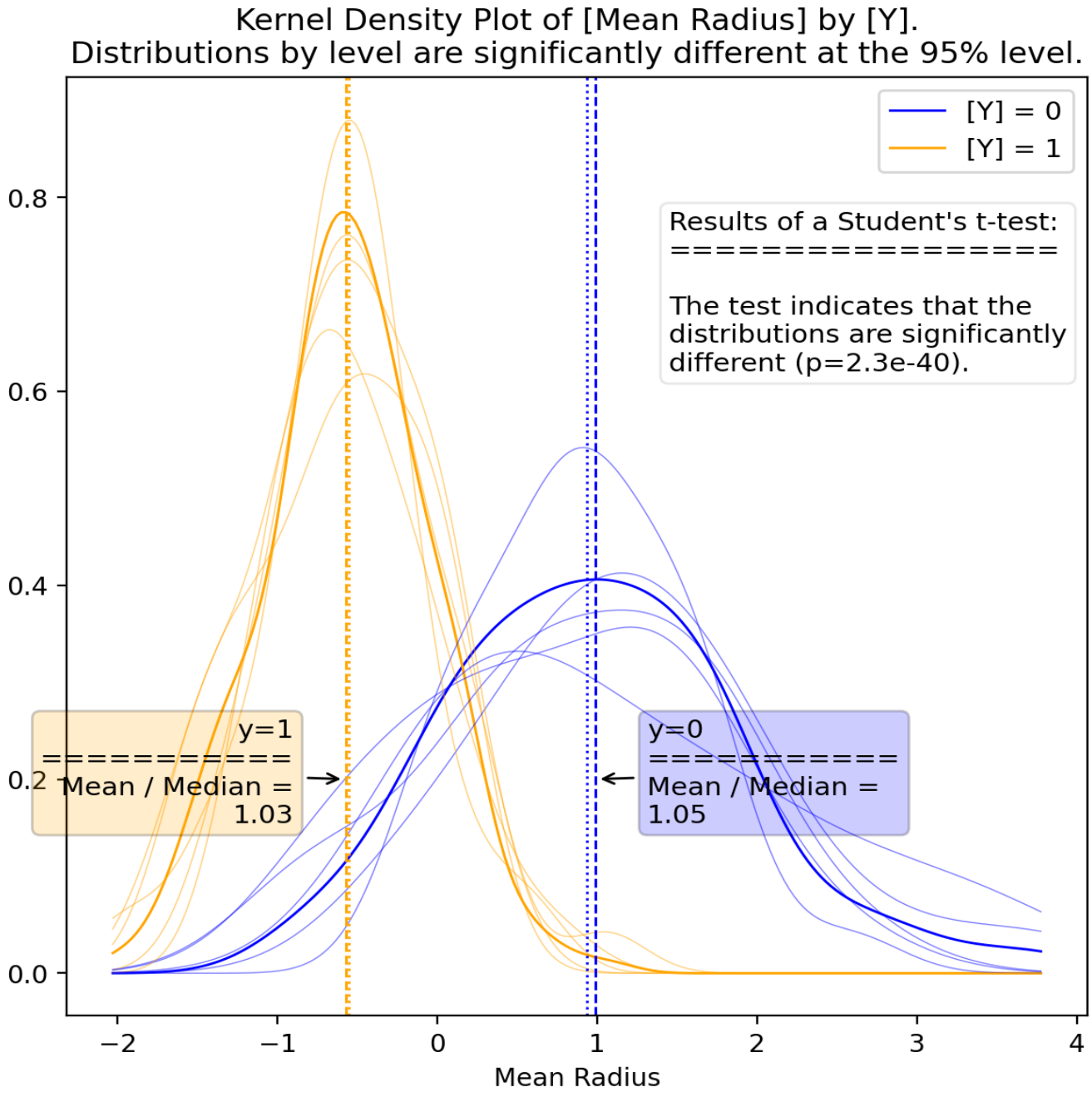
Univariate Report

Mean Radius - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-4.08	-3.92	-3.55	-3.52	-3.44	-3.69	0.28
Fitted p-Value	6.7e-16	1.8e-16	3.9e-17	5.5e-16	8.7e-16	5.6e-20	3.4e-16
Fitted Std. Err.	0.505	0.476	0.422	0.435	0.428	0.403	0.036
Conf. Int. Lower	-5.07	-4.85	-4.38	-4.38	-4.28	-4.48	0.35
Conf. Int. Upper	-3.09	-2.99	-2.72	-2.67	-2.60	-2.90	0.21
Train Accuracy	87.1%	87.7%	86.2%	85.3%	85.4%	86.5%	1.1%
Val Accuracy	83.6%	80.0%	84.5%	90.8%	88.9%	86.8%	4.3%
Train AUC	87.0%	88.0%	86.1%	85.0%	85.5%	86.5%	1.2%
Val AUC	84.2%	78.8%	85.7%	91.2%	87.5%	86.2%	4.5%
Train F1	89.5%	89.9%	88.8%	88.2%	87.6%	88.9%	0.9%
Test F1	86.5%	83.5%	86.6%	91.8%	91.9%	89.4%	3.7%
Train Precision	91.6%	93.3%	91.1%	90.7%	90.5%	91.6%	1.1%
Val Precision	91.4%	82.5%	93.5%	95.1%	92.7%	90.0%	5.0%
Train Recall	87.4%	86.9%	86.5%	85.9%	84.8%	86.4%	1.0%
Val Recall	82.1%	84.6%	80.6%	88.6%	91.1%	88.7%	4.4%
Train MCC	73.1%	74.4%	71.1%	68.8%	70.1%	71.9%	2.3%
Val MCC	66.4%	58.1%	69.5%	81.6%	74.3%	72.1%	8.8%
Train Log-Loss	4.63	4.44	4.97	5.30	5.27	4.86	0.38
Val Log-Loss	5.91	7.21	5.59	3.32	4.00	4.74	1.55

Univariate Report

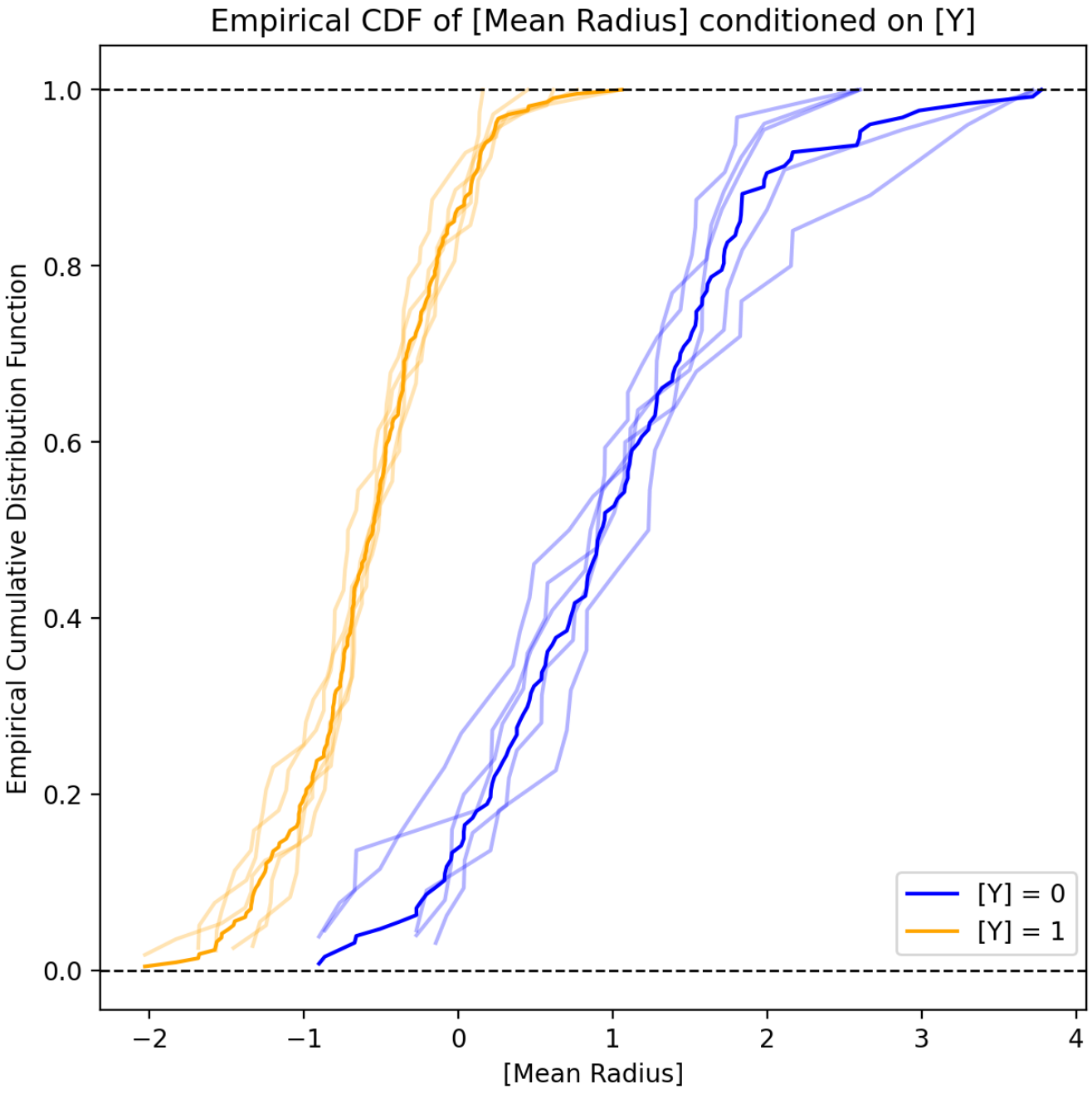
Mean Radius - Kernel Density Plot



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

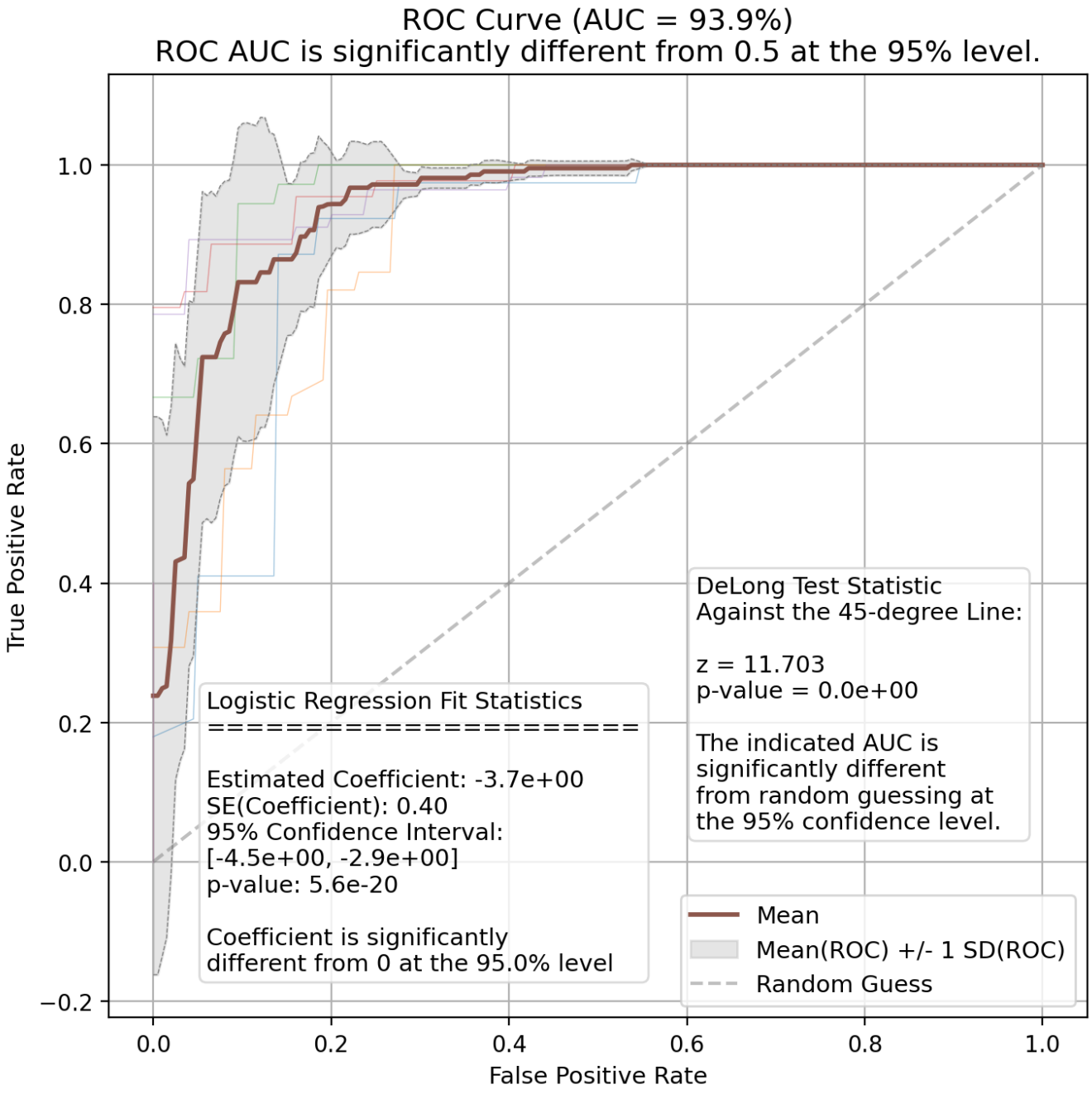
Mean Radius - Empirical CDF Plot



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

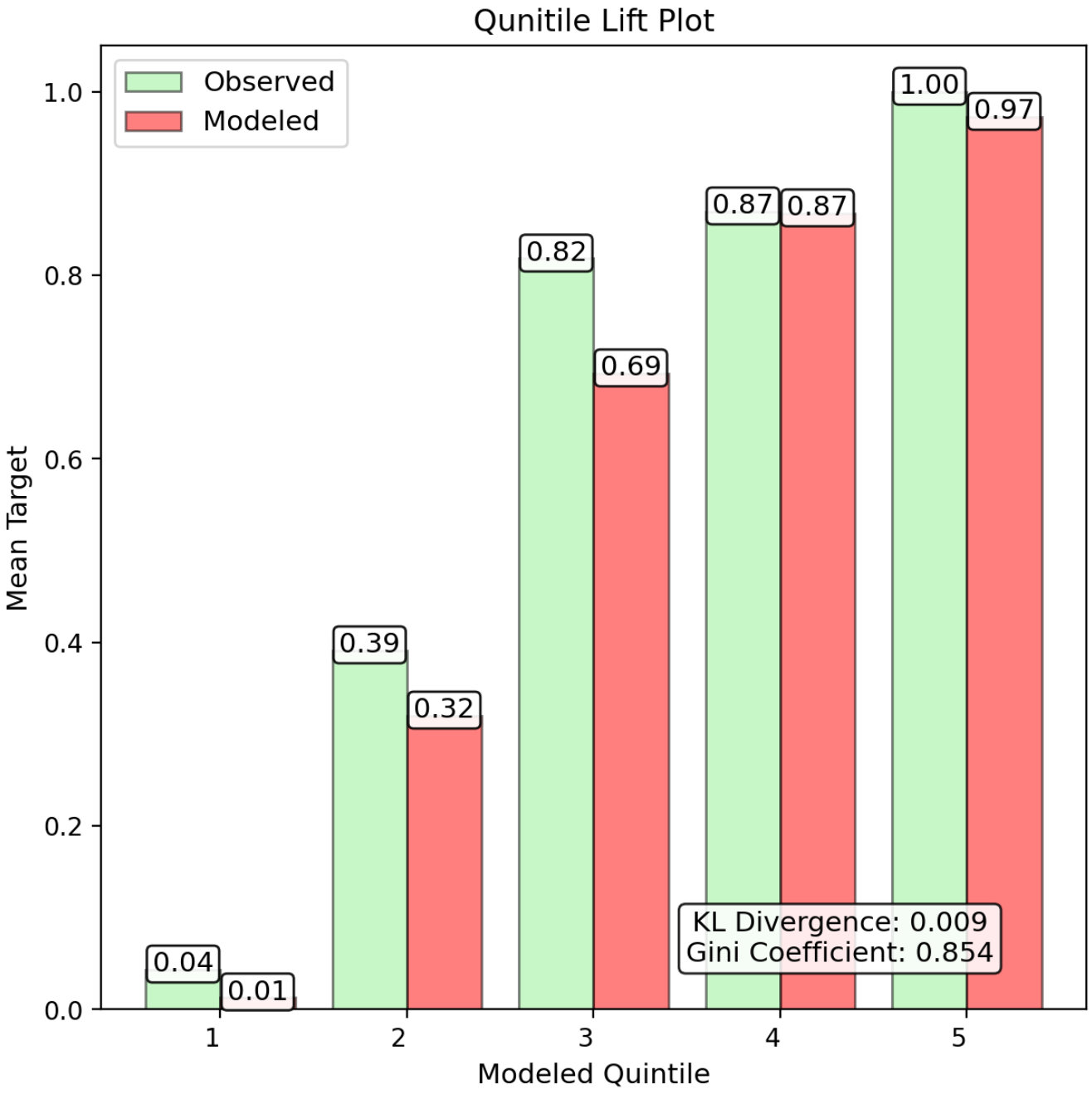


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

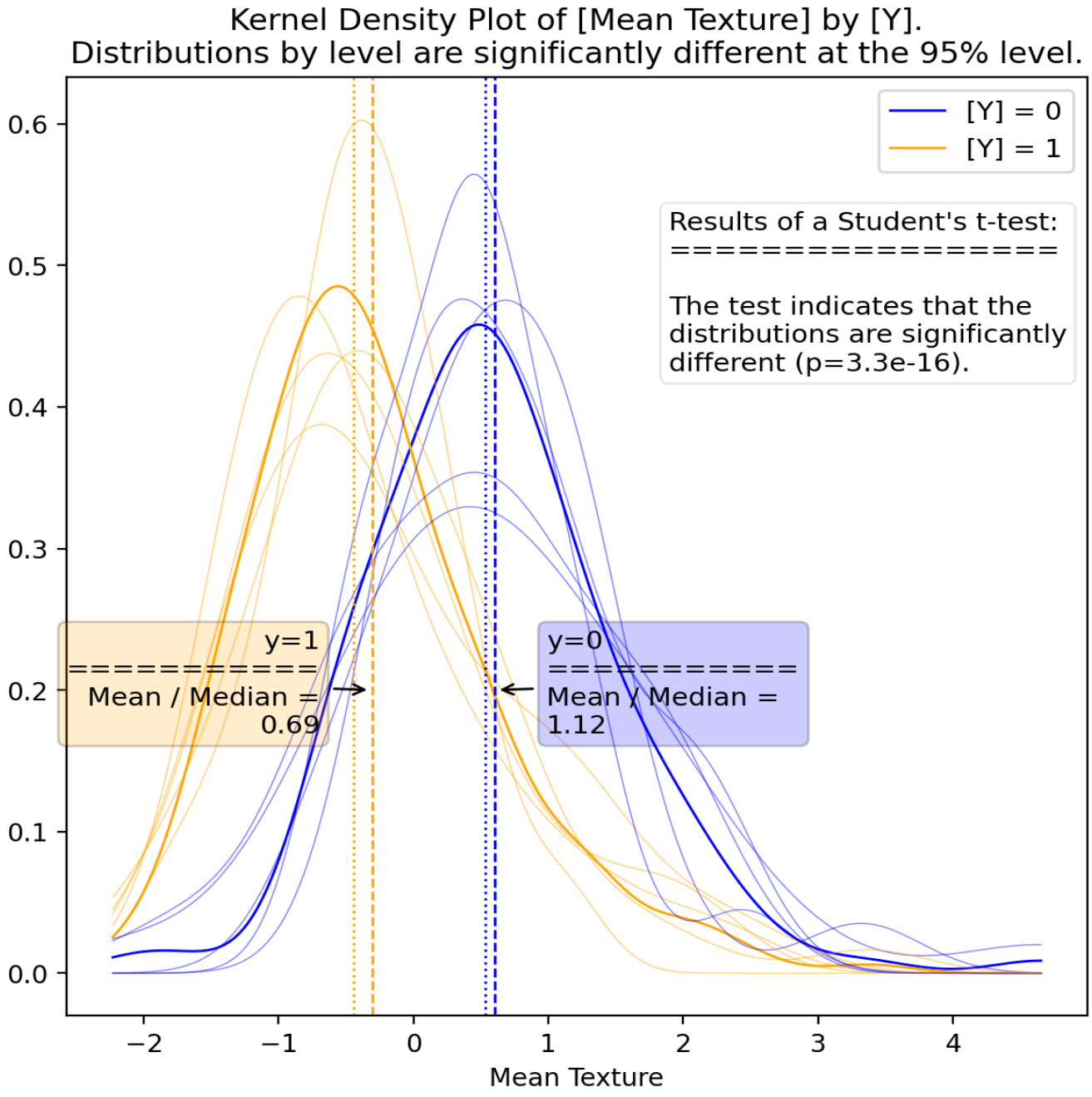
Univariate Report

Mean Texture - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-1.02	-1.07	-0.97	-0.93	-1.06	-1.01	0.06
Fitted p-Value	4.2e-10	1.5e-10	7.0e-10	3.3e-09	2.1e-10	3.5e-12	1.3e-09
Fitted Std. Err.	0.164	0.167	0.158	0.157	0.166	0.145	0.005
Conf. Int. Lower	-1.34	-1.40	-1.28	-1.24	-1.38	-1.29	0.07
Conf. Int. Upper	-0.702	-0.745	-0.663	-0.623	-0.730	-0.725	0.050
Train Accuracy	72.1%	71.7%	70.7%	71.3%	72.7%	71.8%	0.8%
Val Accuracy	70.5%	72.3%	75.9%	72.4%	69.1%	71.9%	2.5%
Train AUC	72.4%	72.1%	70.8%	72.1%	72.7%	72.1%	0.7%
Val AUC	69.0%	73.7%	77.0%	71.9%	69.9%	72.0%	3.2%
Train F1	76.2%	76.1%	75.1%	75.6%	76.4%	76.0%	0.5%
Test F1	76.3%	74.3%	78.8%	75.9%	75.2%	76.1%	1.7%
Train Precision	81.7%	82.1%	80.6%	83.1%	80.4%	81.7%	1.1%
Val Precision	78.4%	83.9%	86.7%	76.7%	84.4%	81.0%	4.2%
Train Recall	71.4%	70.9%	70.2%	69.4%	72.8%	71.0%	1.3%
Val Recall	74.4%	66.7%	72.2%	75.0%	67.9%	71.8%	3.8%
Train MCC	43.5%	42.7%	40.4%	42.5%	44.5%	43.0%	1.5%
Val MCC	37.3%	46.5%	52.5%	43.6%	37.1%	42.8%	6.5%
Train Log-Loss	10.04	10.19	10.57	10.34	9.84	10.15	0.28
Val Log-Loss	10.64	9.98	8.70	9.96	11.12	10.12	0.91

Univariate Report

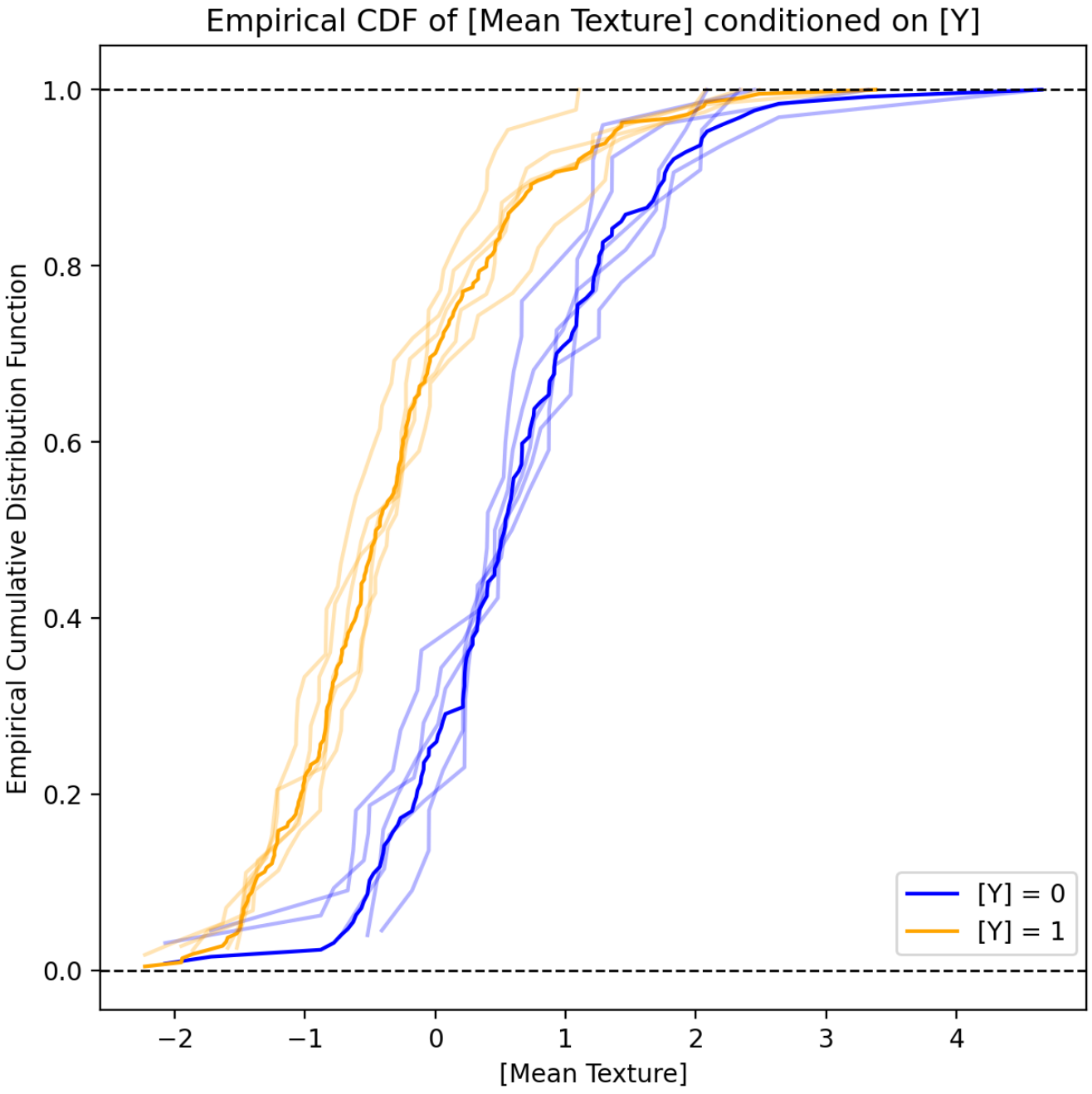
Mean Texture - Kernel Density Plot



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

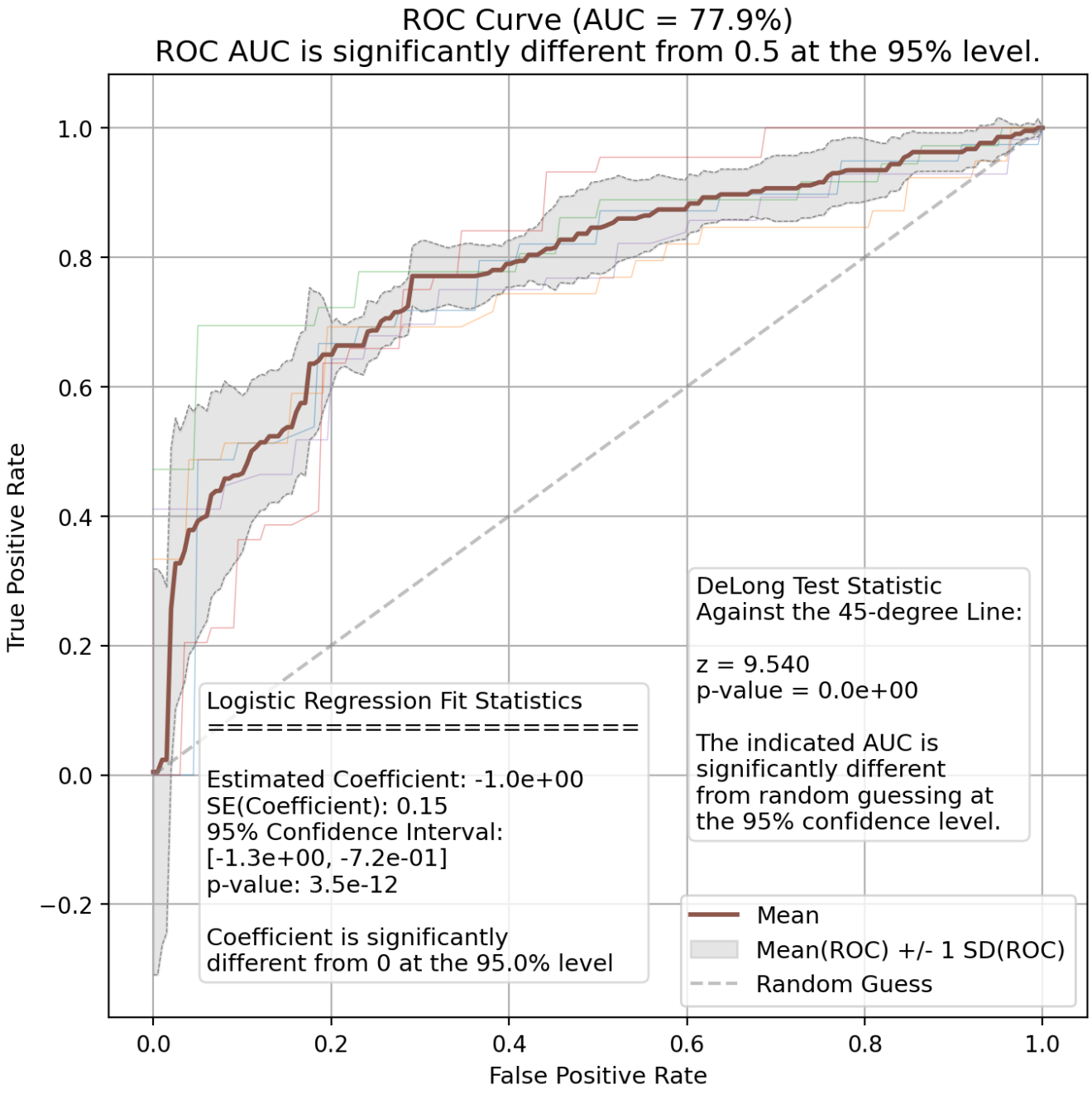
Mean Texture - Empirical CDF Plot



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

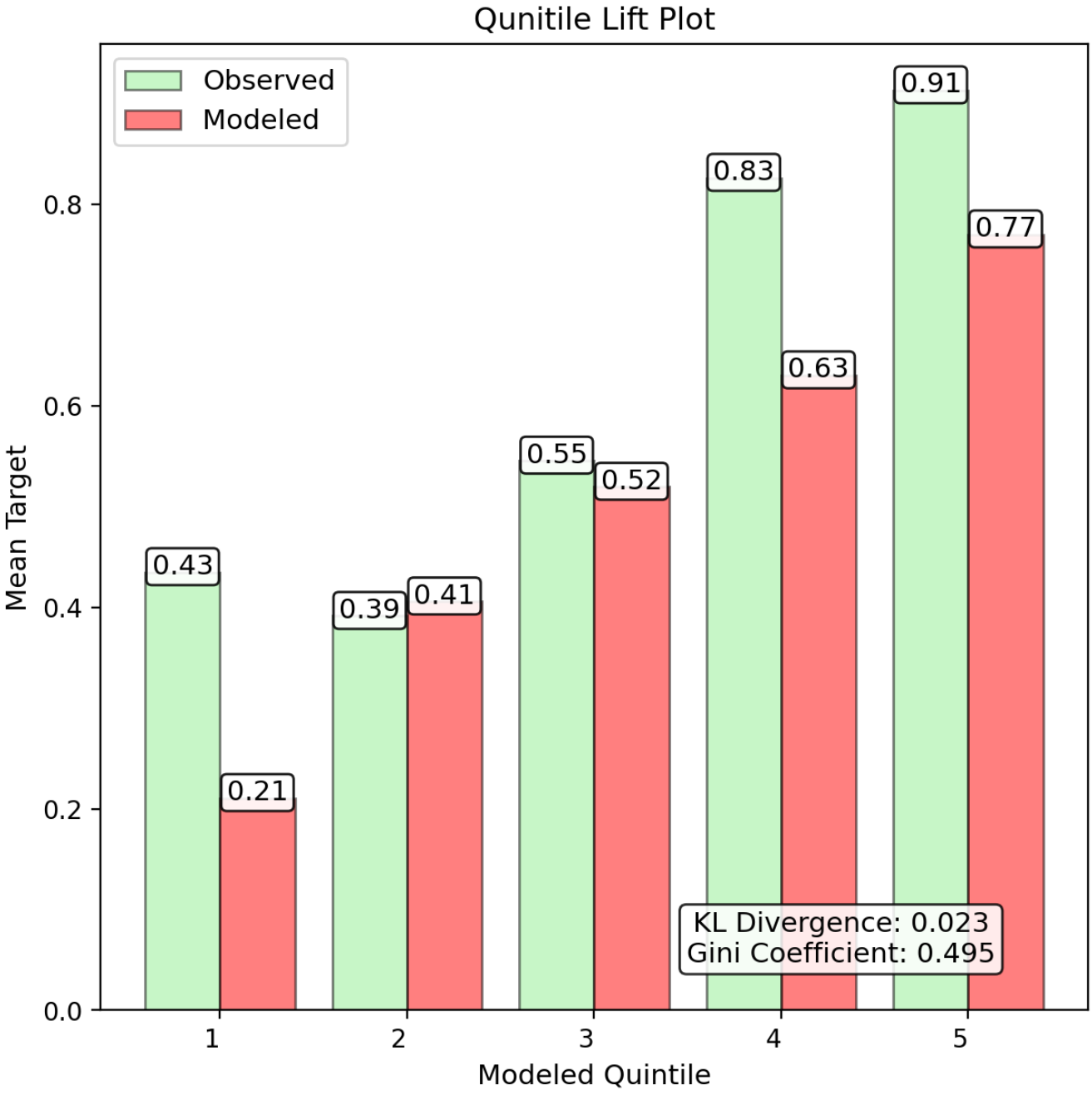


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

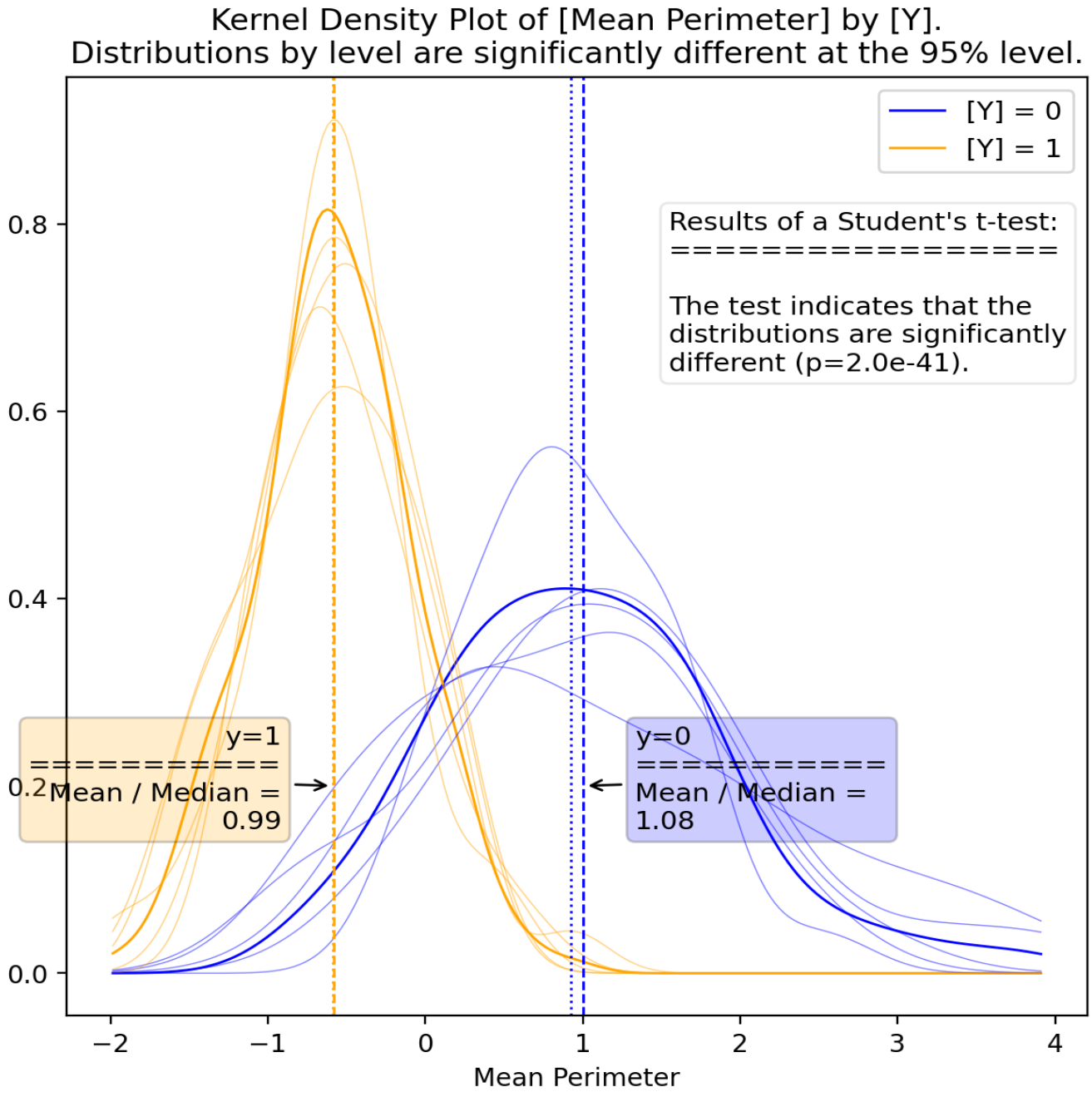
Univariate Report

Mean Perimeter - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-4.37	-4.28	-3.86	-3.82	-3.73	-4.00	0.29
Fitted p-Value	7.5e-16	3.0e-16	4.3e-17	5.7e-16	7.7e-16	6.6e-20	3.1e-16
Fitted Std. Err.	0.541	0.524	0.460	0.472	0.463	0.438	0.038
Conf. Int. Lower	-5.43	-5.31	-4.76	-4.74	-4.64	-4.86	0.36
Conf. Int. Upper	-3.30	-3.25	-2.96	-2.89	-2.82	-3.14	0.22
Train Accuracy	88.2%	88.8%	88.0%	86.4%	86.5%	87.4%	1.1%
Val Accuracy	83.6%	81.5%	87.9%	90.8%	90.1%	86.8%	4.1%
Train AUC	88.1%	88.8%	88.1%	86.2%	86.7%	87.4%	1.1%
Val AUC	84.2%	80.8%	88.5%	91.2%	88.4%	86.2%	4.1%
Train F1	90.4%	90.9%	90.2%	89.2%	88.6%	89.7%	0.9%
Test F1	86.5%	84.6%	89.9%	91.8%	92.9%	89.4%	3.5%
Train Precision	92.3%	93.4%	92.9%	91.4%	91.3%	92.1%	0.9%
Val Precision	91.4%	84.6%	93.9%	95.1%	92.9%	90.0%	4.1%
Train Recall	88.6%	88.6%	87.6%	87.1%	86.1%	87.4%	1.1%
Val Recall	82.1%	84.6%	86.1%	88.6%	92.9%	88.7%	4.1%
Train MCC	75.3%	76.4%	75.0%	71.2%	72.4%	73.7%	2.2%
Val MCC	66.4%	61.5%	75.5%	81.6%	76.9%	72.1%	8.2%
Train Log-Loss	4.25	4.05	4.33	4.90	4.85	4.55	0.38
Val Log-Loss	5.91	6.65	4.35	3.32	3.56	4.74	1.47

Univariate Report

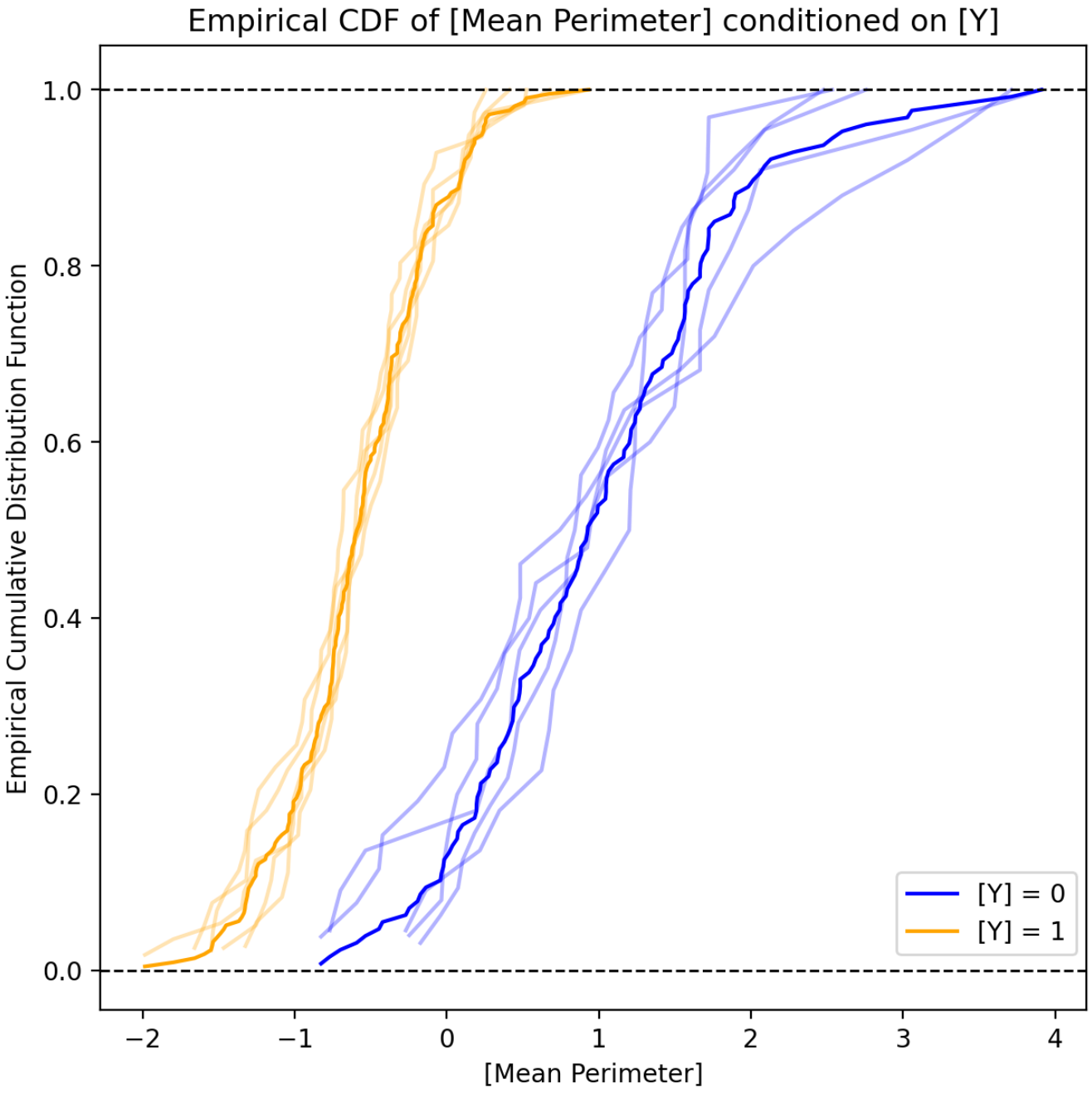
Mean Perimeter - Kernel Density Plot



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

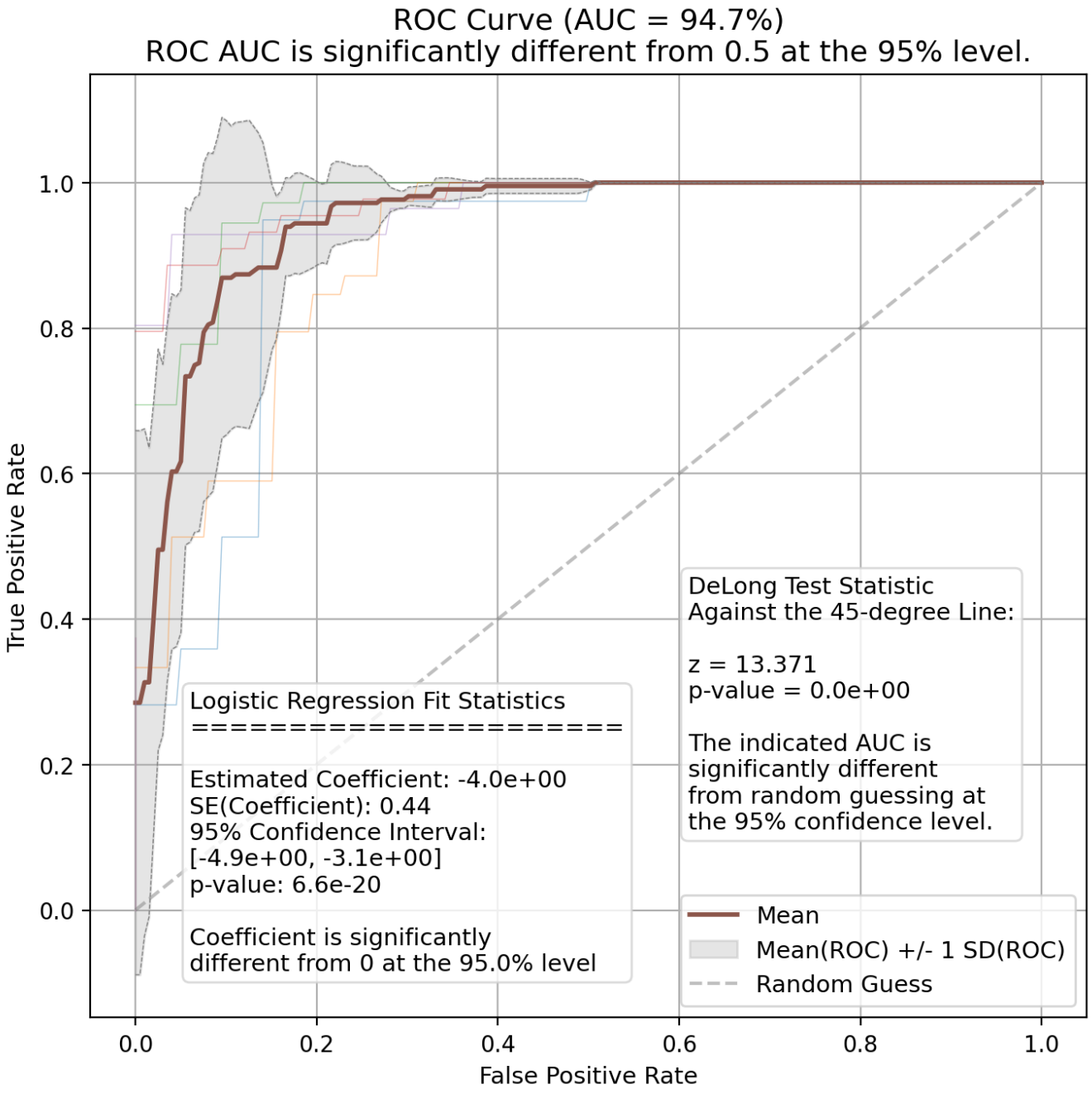
Mean Perimeter - Empirical CDF Plot



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

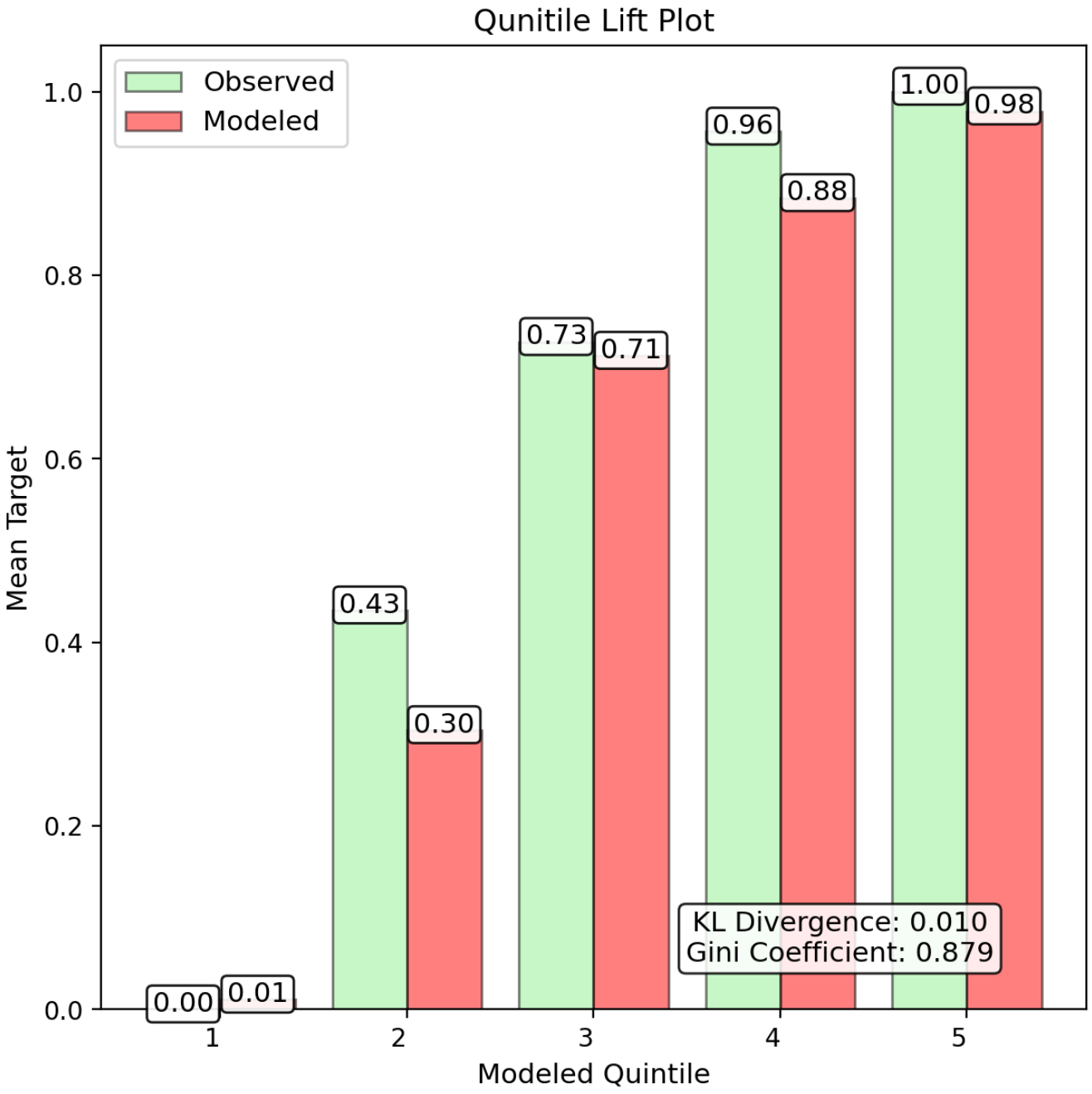


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

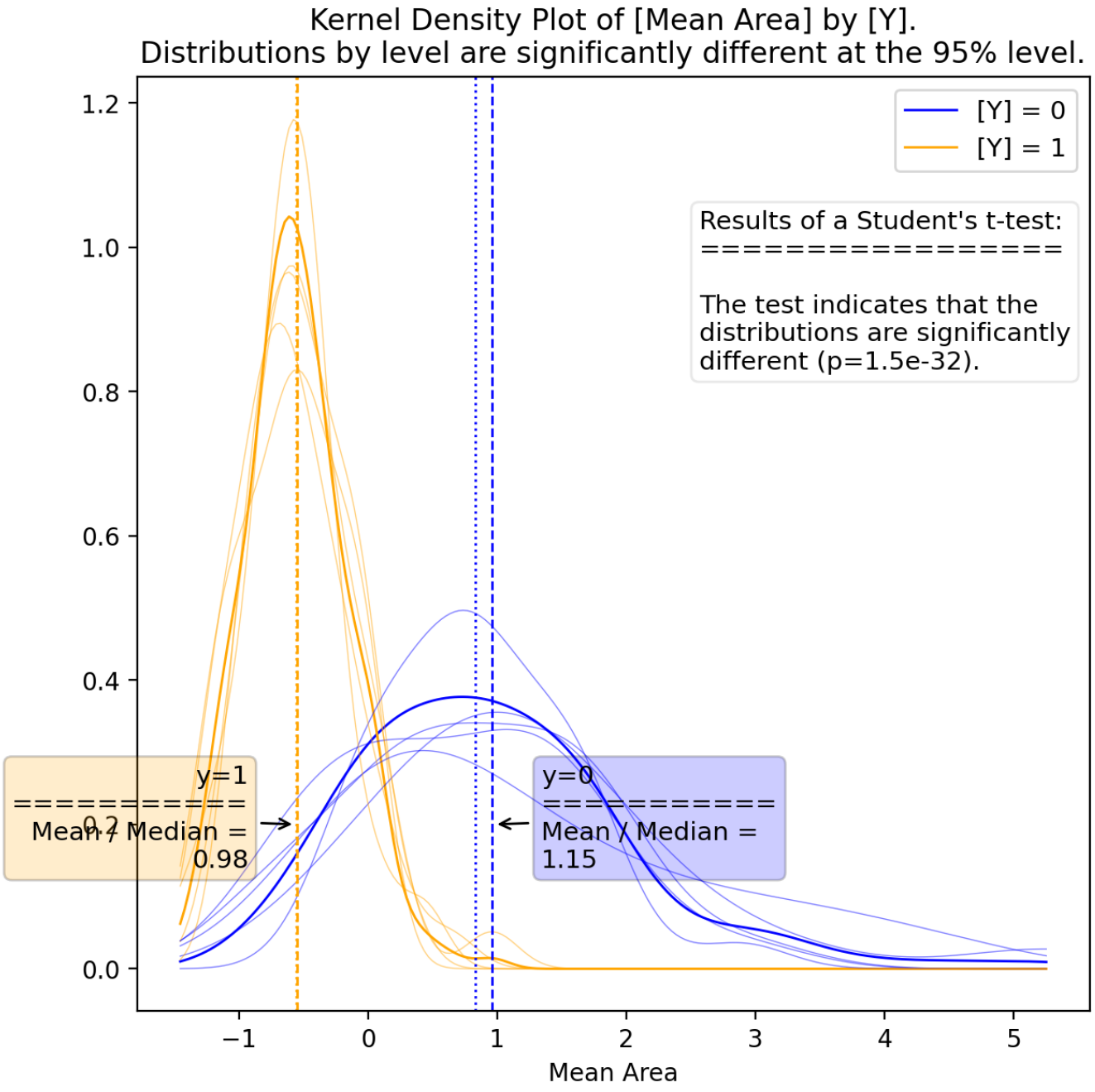
Univariate Report

Mean Area - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-4.65	-4.55	-4.04	-4.04	-3.98	-4.23	0.32
Fitted p-Value	1.8e-16	9.9e-17	1.0e-17	1.4e-16	4.2e-16	1.4e-20	1.5e-16
Fitted Std. Err.	0.564	0.548	0.471	0.489	0.489	0.455	0.041
Conf. Int. Lower	-5.75	-5.62	-4.97	-5.00	-4.94	-5.13	0.40
Conf. Int. Upper	-3.54	-3.48	-3.12	-3.08	-3.02	-3.34	0.24
Train Accuracy	88.6%	89.1%	86.2%	88.7%	88.5%	88.3%	1.1%
Val Accuracy	86.9%	89.2%	91.4%	86.8%	88.9%	88.6%	1.9%
Train AUC	87.2%	88.1%	85.1%	87.0%	87.4%	87.0%	1.1%
Val AUC	85.8%	86.5%	91.3%	86.5%	86.4%	86.7%	2.2%
Train F1	91.0%	91.5%	89.1%	91.3%	90.7%	90.8%	1.0%
Test F1	89.7%	91.8%	93.0%	88.6%	92.0%	91.2%	1.8%
Train Precision	89.5%	91.0%	88.8%	89.8%	89.0%	89.5%	0.8%
Val Precision	89.7%	84.8%	94.3%	88.6%	91.2%	88.2%	3.5%
Train Recall	92.6%	92.0%	89.3%	92.9%	92.4%	92.1%	1.5%
Val Recall	89.7%	100.0%	91.7%	88.6%	92.9%	94.4%	4.5%
Train MCC	75.4%	76.5%	70.4%	75.1%	75.6%	74.7%	2.4%
Val MCC	71.6%	78.7%	81.9%	73.0%	73.7%	75.5%	4.4%
Train Log-Loss	4.12	3.92	4.97	4.08	4.16	4.23	0.41
Val Log-Loss	4.73	3.88	3.11	4.74	4.00	4.11	0.68

Univariate Report

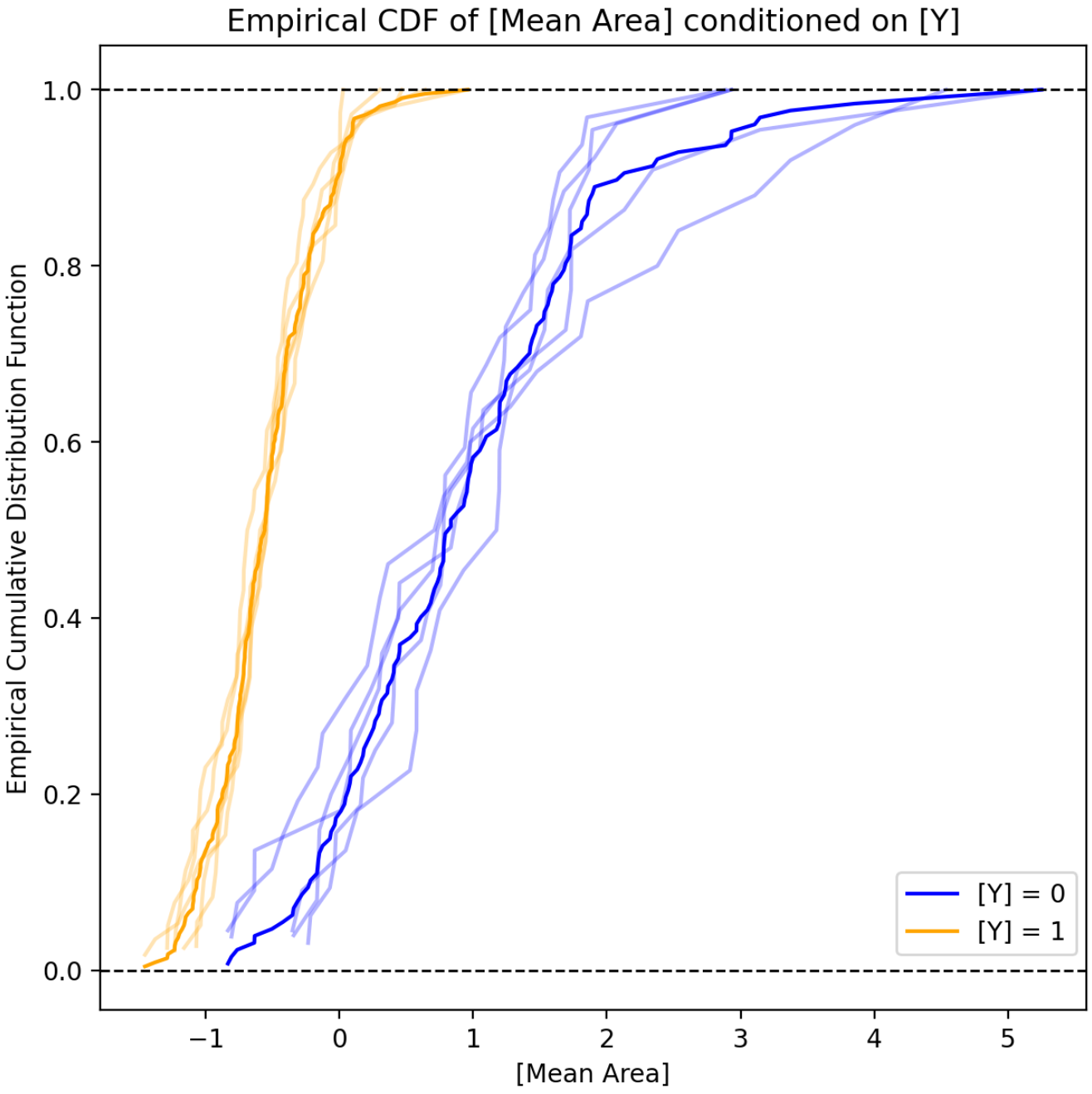
Mean Area - Kernel Density Plot



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

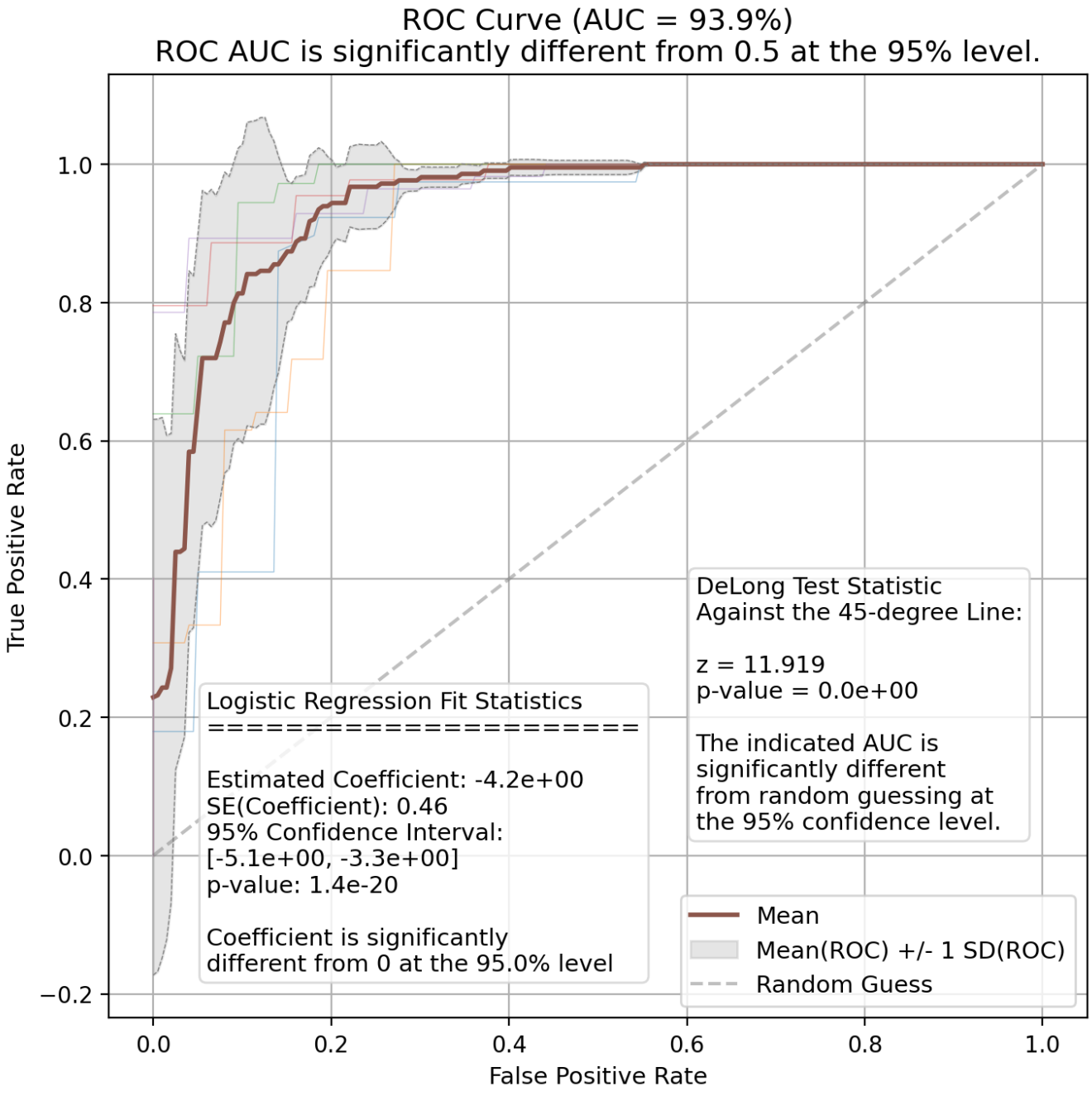
Mean Area - Empirical CDF Plot



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

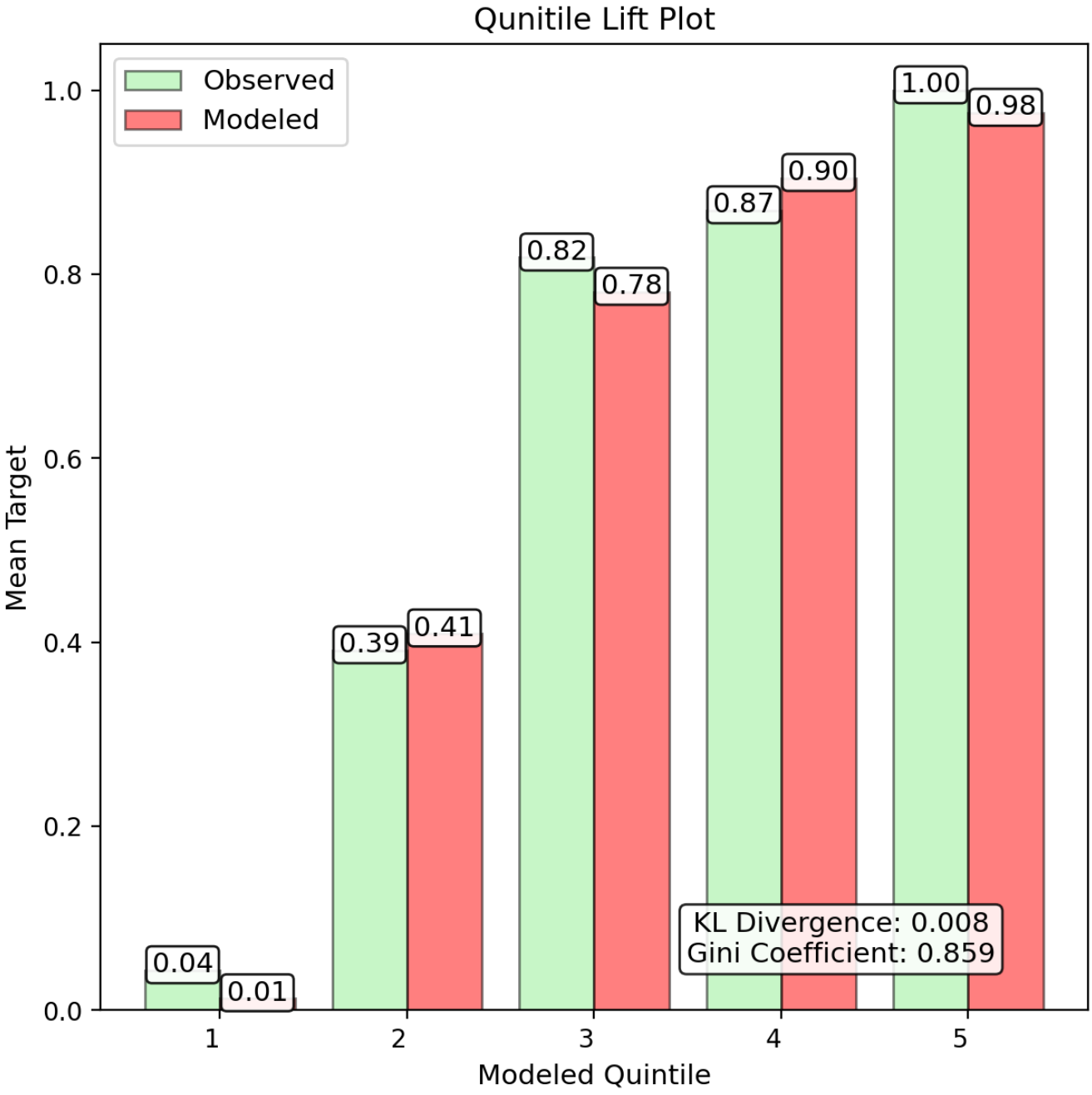


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

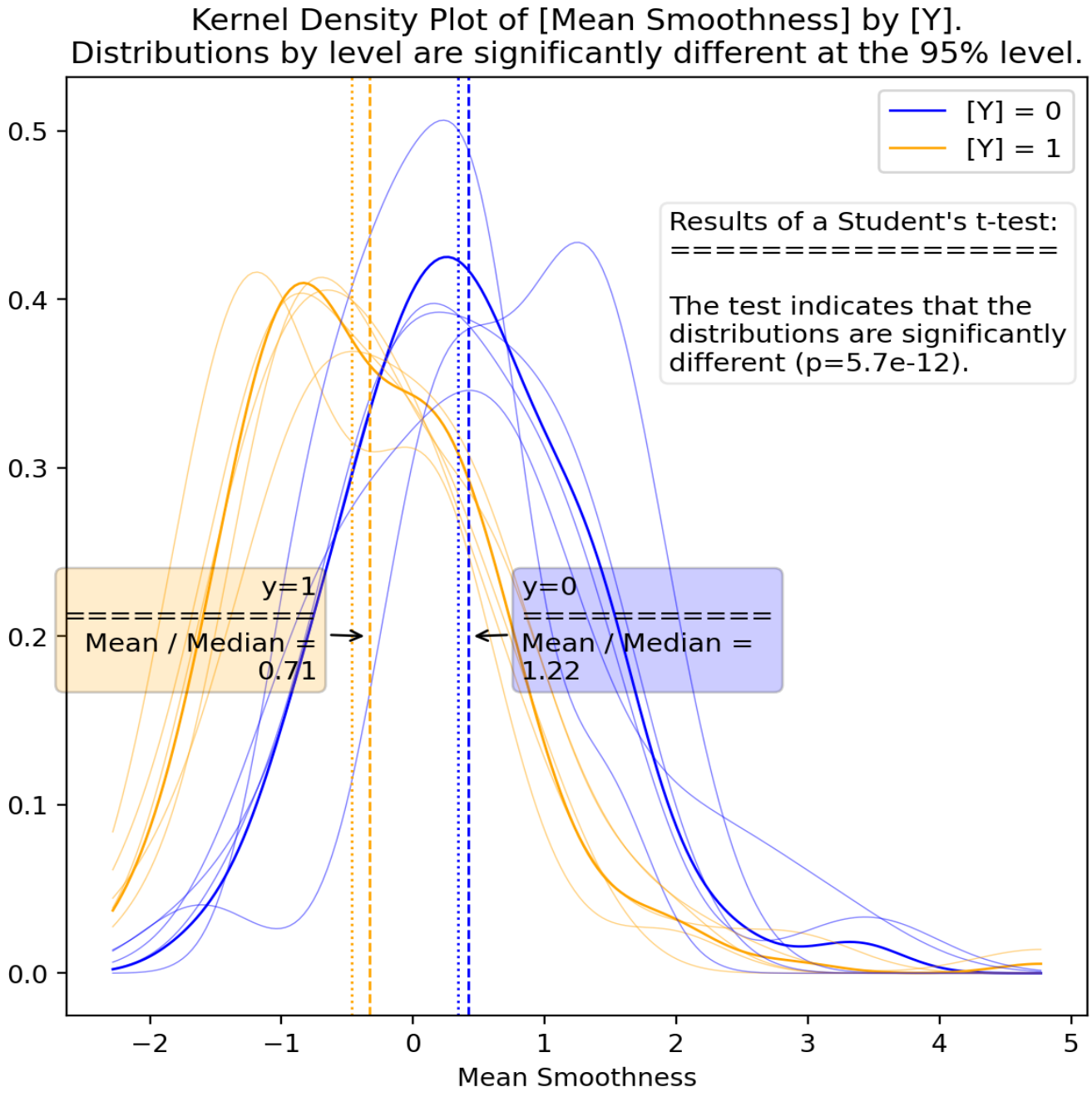
Univariate Report

Mean Smoothness - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-0.673	-0.847	-0.773	-0.898	-1.012	-0.836	0.128
Fitted p-Value	2.7e-06	1.9e-08	6.8e-08	3.1e-09	5.5e-10	4.5e-10	1.2e-06
Fitted Std. Err.	0.143	0.151	0.143	0.152	0.163	0.134	0.008
Conf. Int. Lower	-0.95	-1.14	-1.05	-1.20	-1.33	-1.10	0.14
Conf. Int. Upper	-0.392	-0.551	-0.492	-0.601	-0.692	-0.573	0.113
Train Accuracy	62.5%	64.5%	65.4%	67.5%	66.9%	64.8%	2.0%
Val Accuracy	80.3%	66.2%	69.0%	59.2%	58.0%	67.5%	9.0%
Train AUC	62.8%	65.7%	66.2%	68.9%	67.4%	65.7%	2.3%
Val AUC	83.6%	66.0%	68.8%	58.8%	60.8%	68.9%	9.8%
Train F1	67.3%	68.6%	69.6%	71.7%	70.5%	68.9%	1.7%
Test F1	82.4%	70.3%	73.5%	63.5%	63.8%	70.9%	7.8%
Train Precision	74.0%	78.1%	77.8%	81.3%	76.9%	77.3%	2.6%
Val Precision	96.6%	74.3%	78.1%	65.9%	78.9%	80.4%	11.2%
Train Recall	61.7%	61.1%	62.9%	64.1%	65.2%	62.1%	1.7%
Val Recall	71.8%	66.7%	69.4%	61.4%	53.6%	63.4%	7.3%
Train MCC	24.7%	30.3%	31.4%	36.3%	34.0%	30.4%	4.4%
Val MCC	64.7%	31.5%	36.7%	17.4%	20.0%	36.6%	18.9%
Train Log-Loss	13.52	12.80	12.48	11.70	11.92	12.68	0.72
Val Log-Loss	7.09	12.20	11.19	14.70	15.13	11.70	3.24

Univariate Report

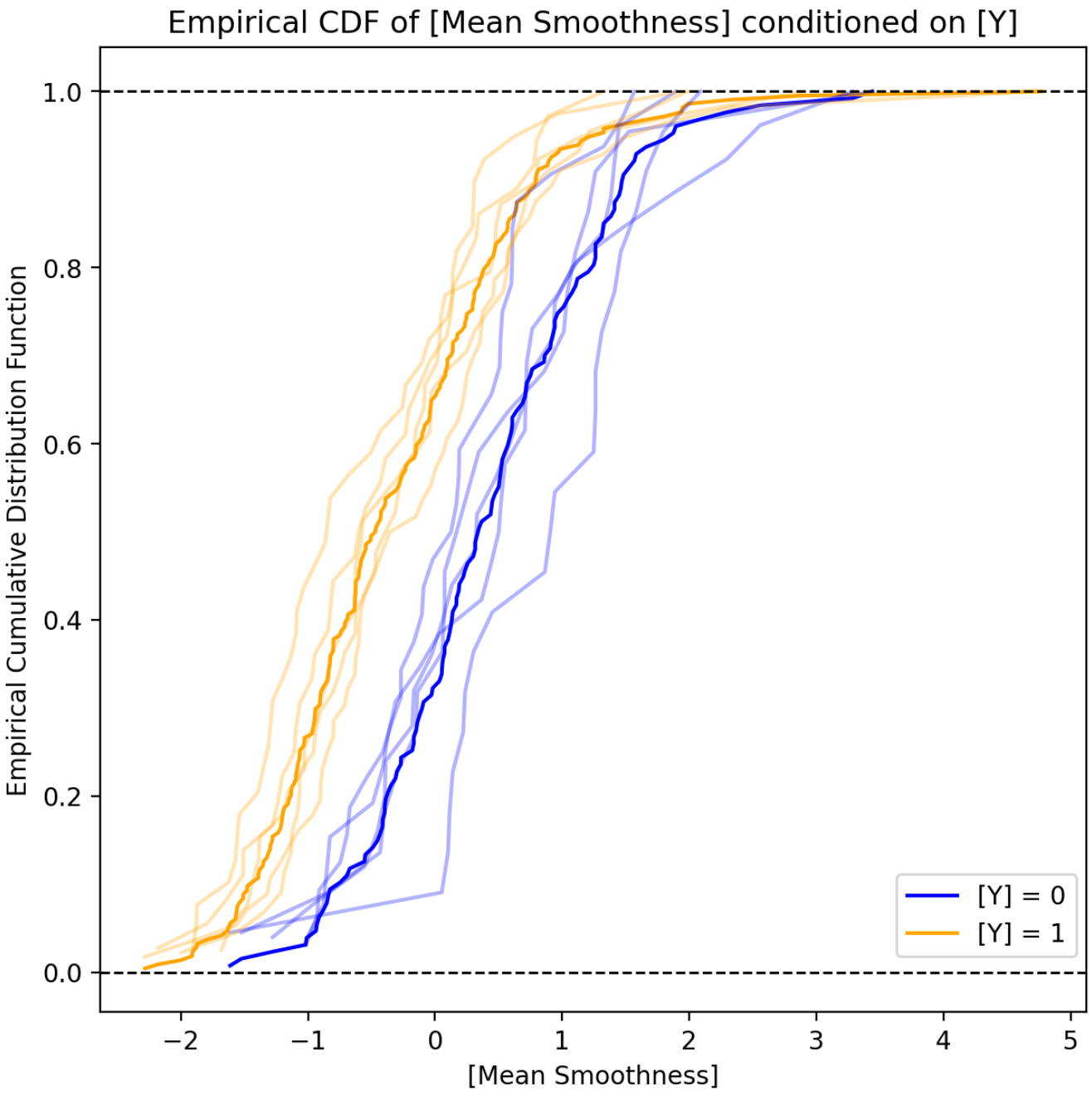
Mean Smoothness - Kernel Density Plot



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

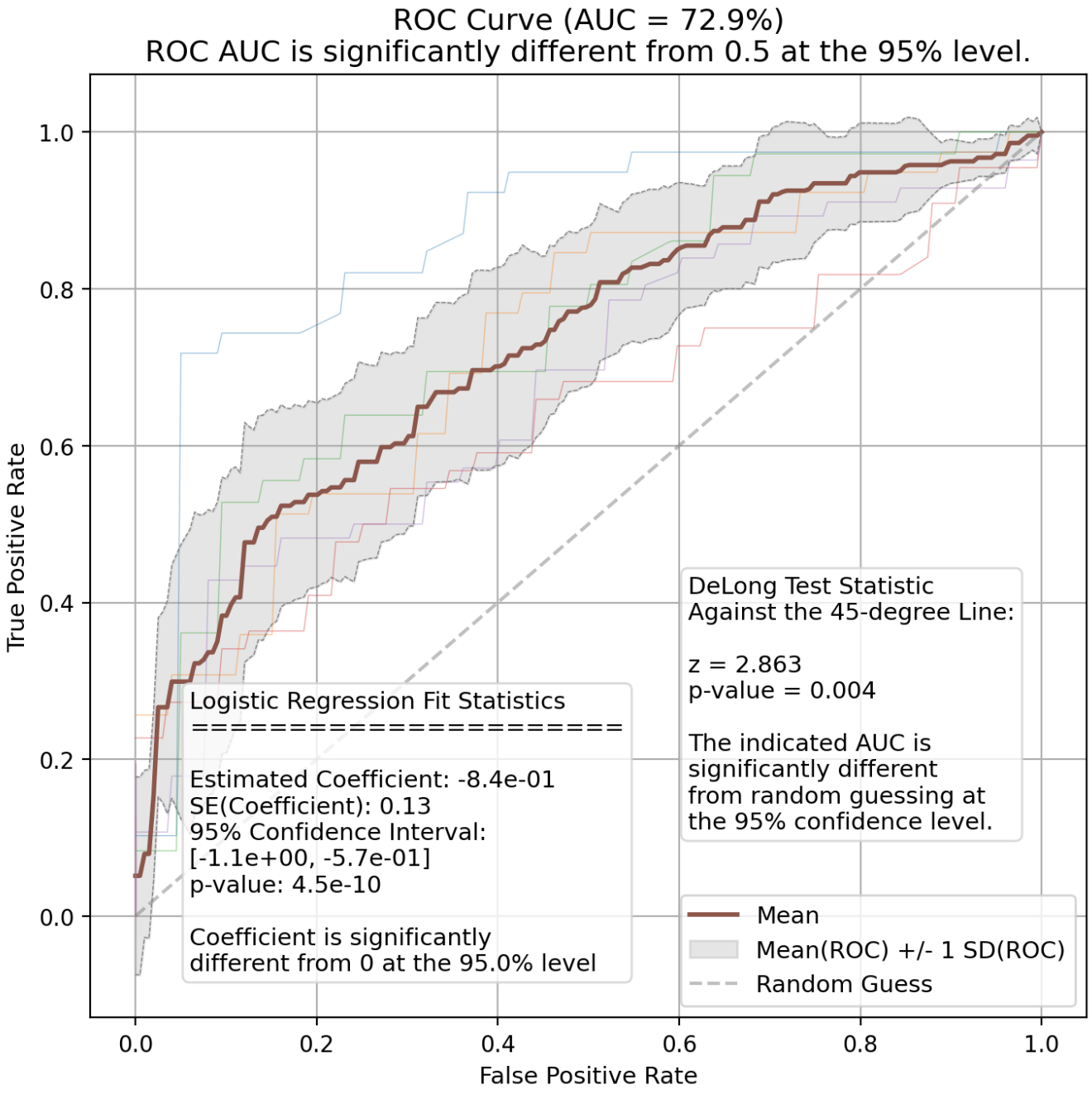
Mean Smoothness - Empirical CDF Plot



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

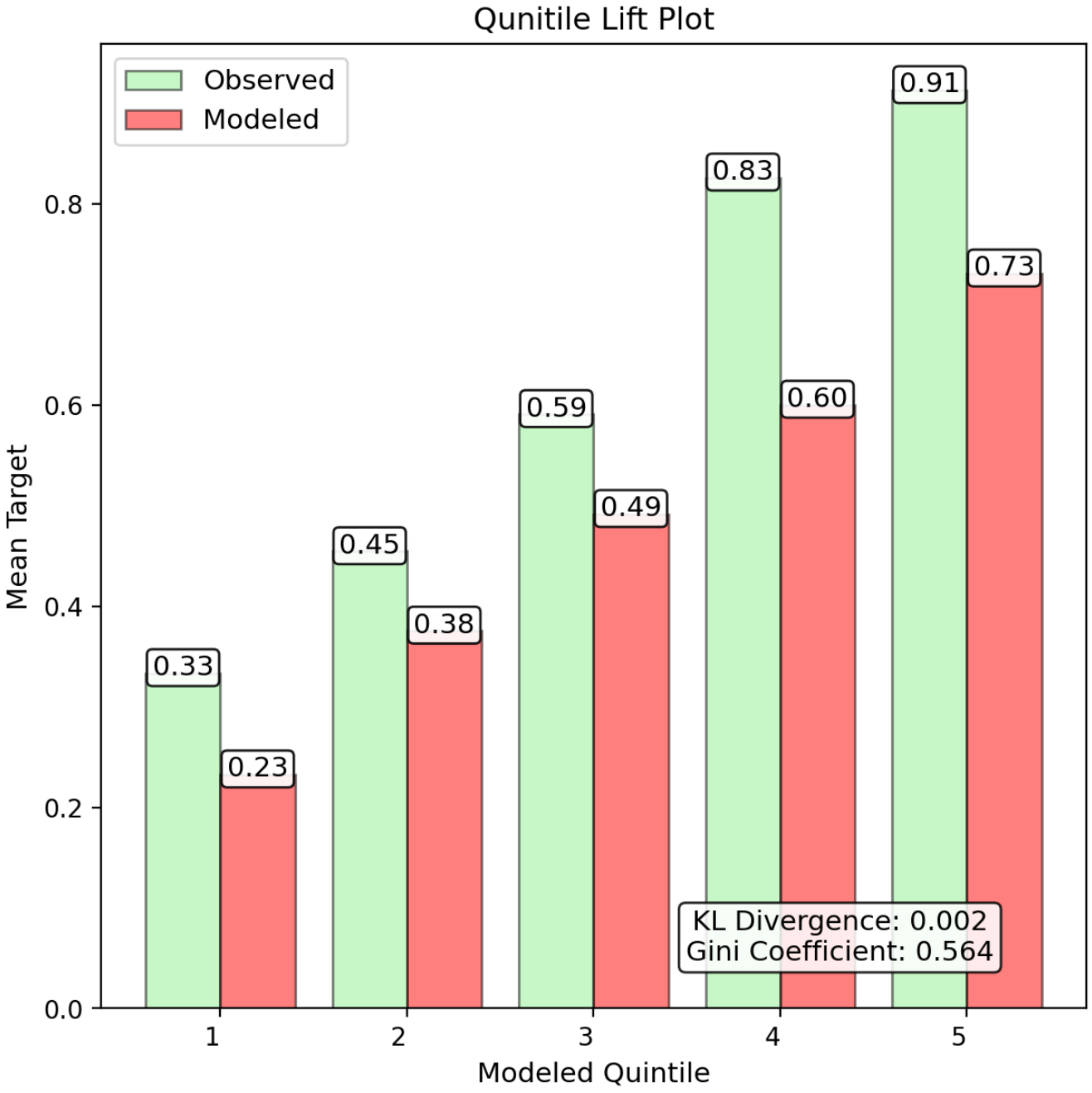


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

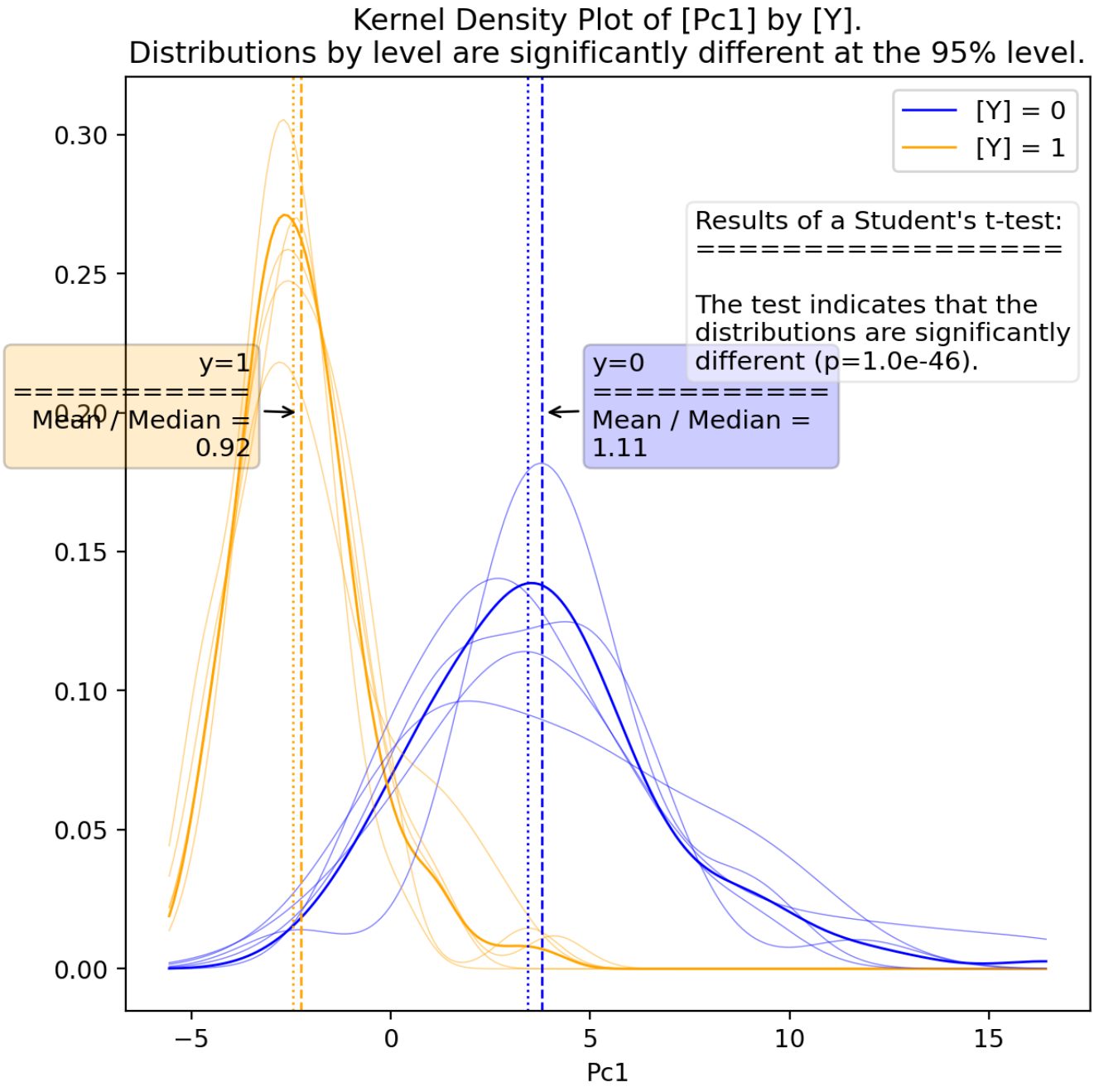
Univariate Report

Pc1 - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	-1.16	-1.36	-1.20	-1.22	-1.15	-1.21	0.08
Fitted p-Value	1.3e-17	8.5e-16	9.9e-18	2.6e-16	7.2e-17	9.2e-21	3.5e-16
Fitted Std. Err.	0.136	0.169	0.140	0.149	0.137	0.130	0.014
Conf. Int. Lower	-1.43	-1.69	-1.47	-1.52	-1.42	-1.47	0.11
Conf. Int. Upper	-0.897	-1.030	-0.925	-0.930	-0.878	-0.959	0.059
Train Accuracy	90.4%	92.4%	90.8%	92.1%	91.2%	91.5%	0.9%
Val Accuracy	96.7%	86.2%	93.1%	89.5%	92.6%	92.1%	4.0%
Train AUC	90.2%	91.9%	90.9%	92.0%	91.0%	91.3%	0.7%
Val AUC	96.4%	87.2%	91.8%	89.2%	92.4%	91.8%	3.5%
Train F1	92.2%	94.0%	92.5%	93.7%	92.7%	93.1%	0.8%
Test F1	97.4%	87.7%	94.6%	90.9%	94.5%	93.6%	3.8%
Train Precision	93.5%	94.3%	94.7%	95.2%	93.5%	94.3%	0.7%
Val Precision	97.4%	94.1%	92.1%	90.9%	96.3%	94.3%	2.7%
Train Recall	90.9%	93.7%	90.4%	92.4%	91.8%	92.1%	1.3%
Val Recall	97.4%	82.1%	97.2%	90.9%	92.9%	93.0%	6.3%
Train MCC	79.7%	83.6%	80.8%	83.0%	81.6%	82.0%	1.6%
Val MCC	92.9%	72.9%	85.3%	78.4%	83.2%	83.3%	7.5%
Train Log-Loss	3.48	2.74	3.31	2.86	3.19	3.07	0.31
Val Log-Loss	1.18	4.99	2.49	3.79	2.67	2.85	1.44

Univariate Report

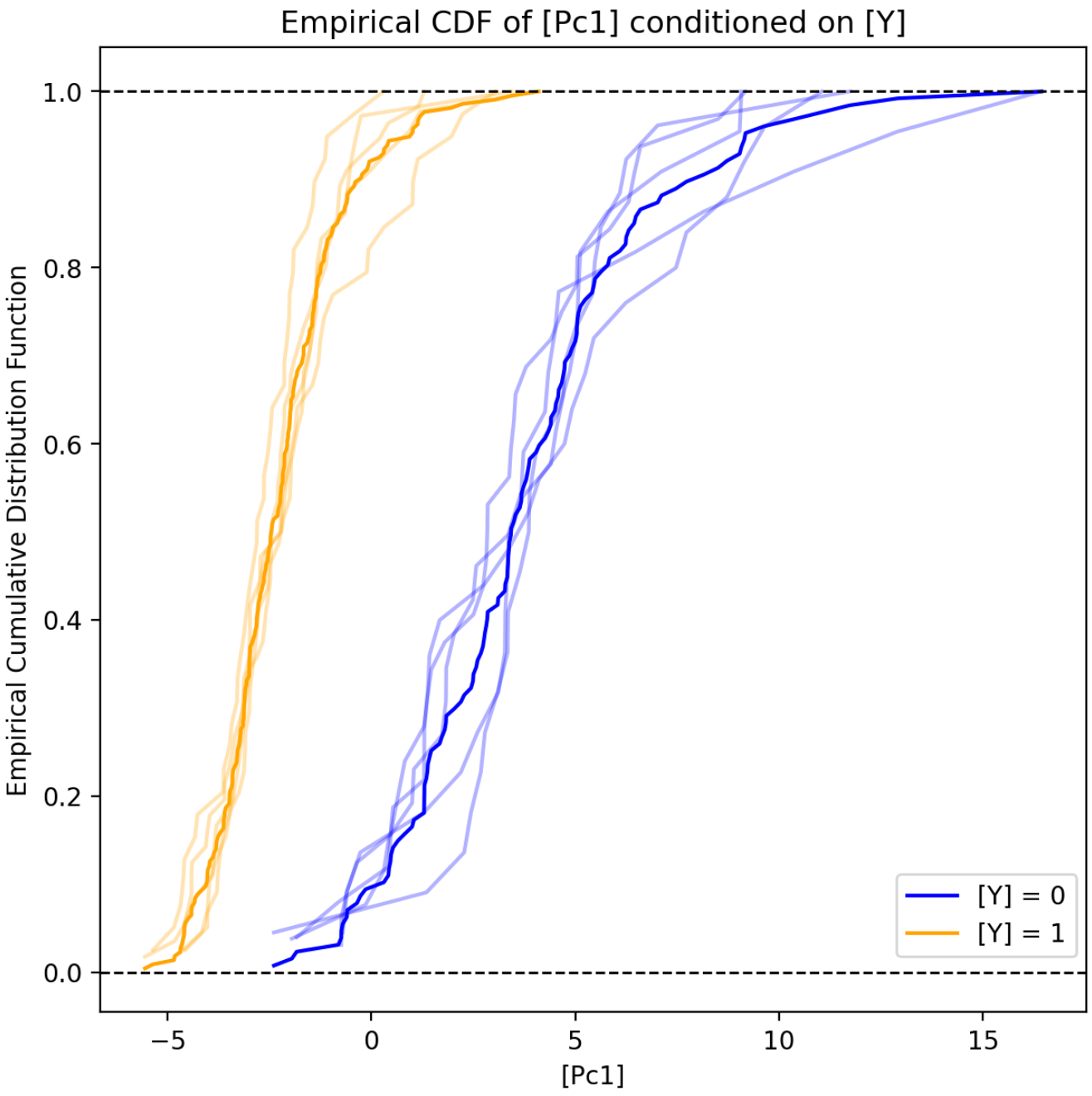
Pc1 - Kernel Density Plot



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

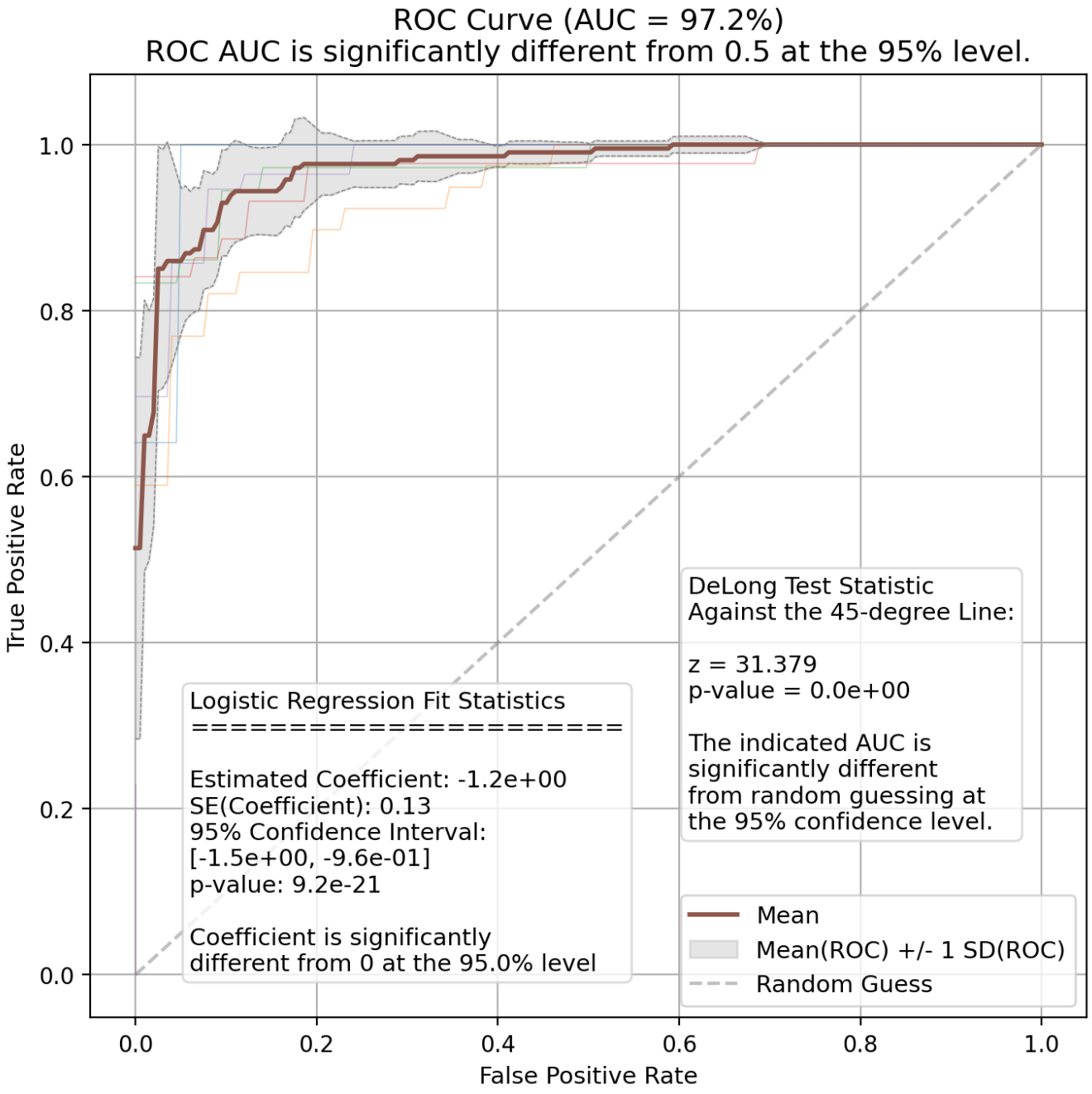
Pc1 - Empirical CDF Plot



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

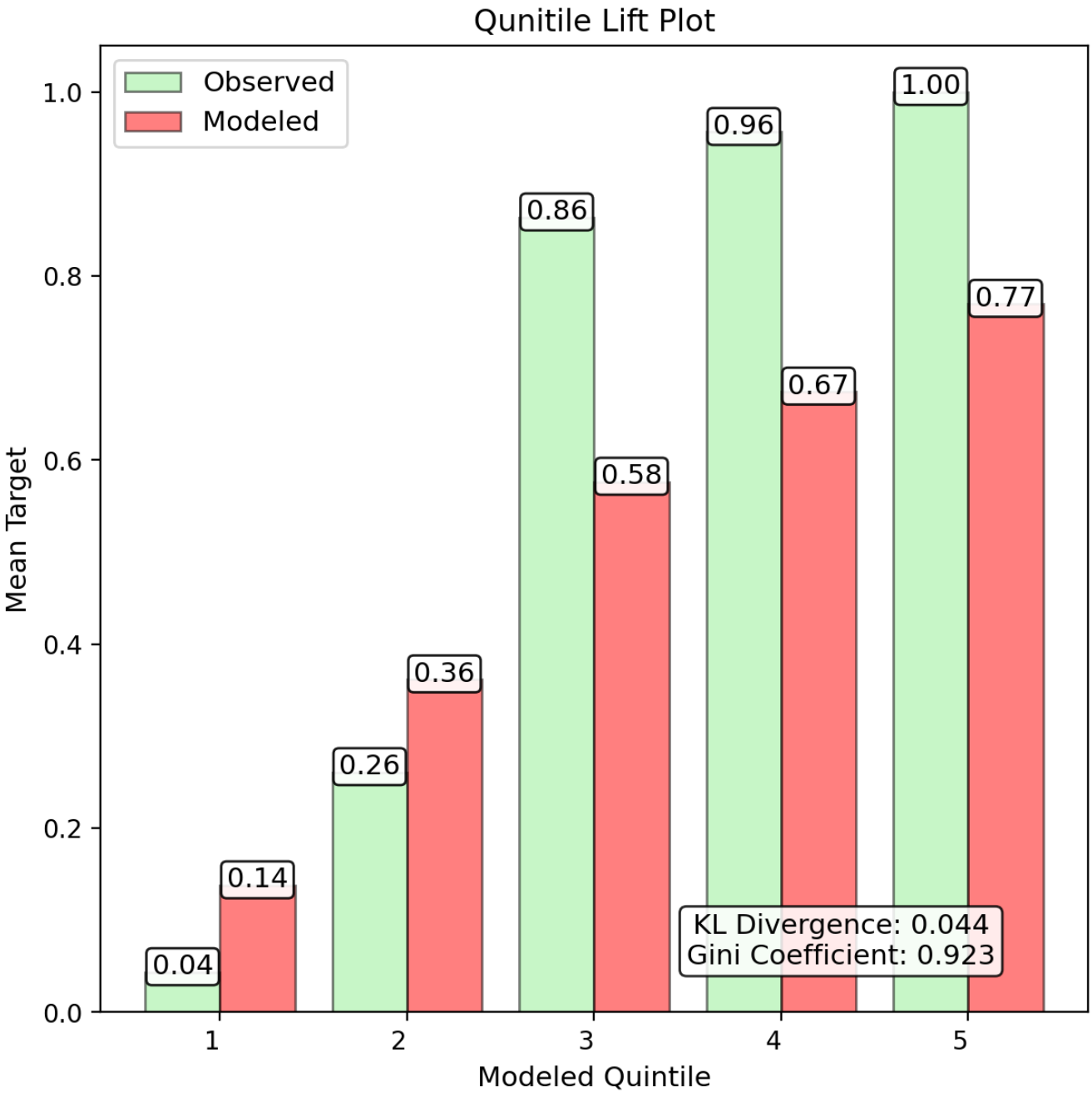


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

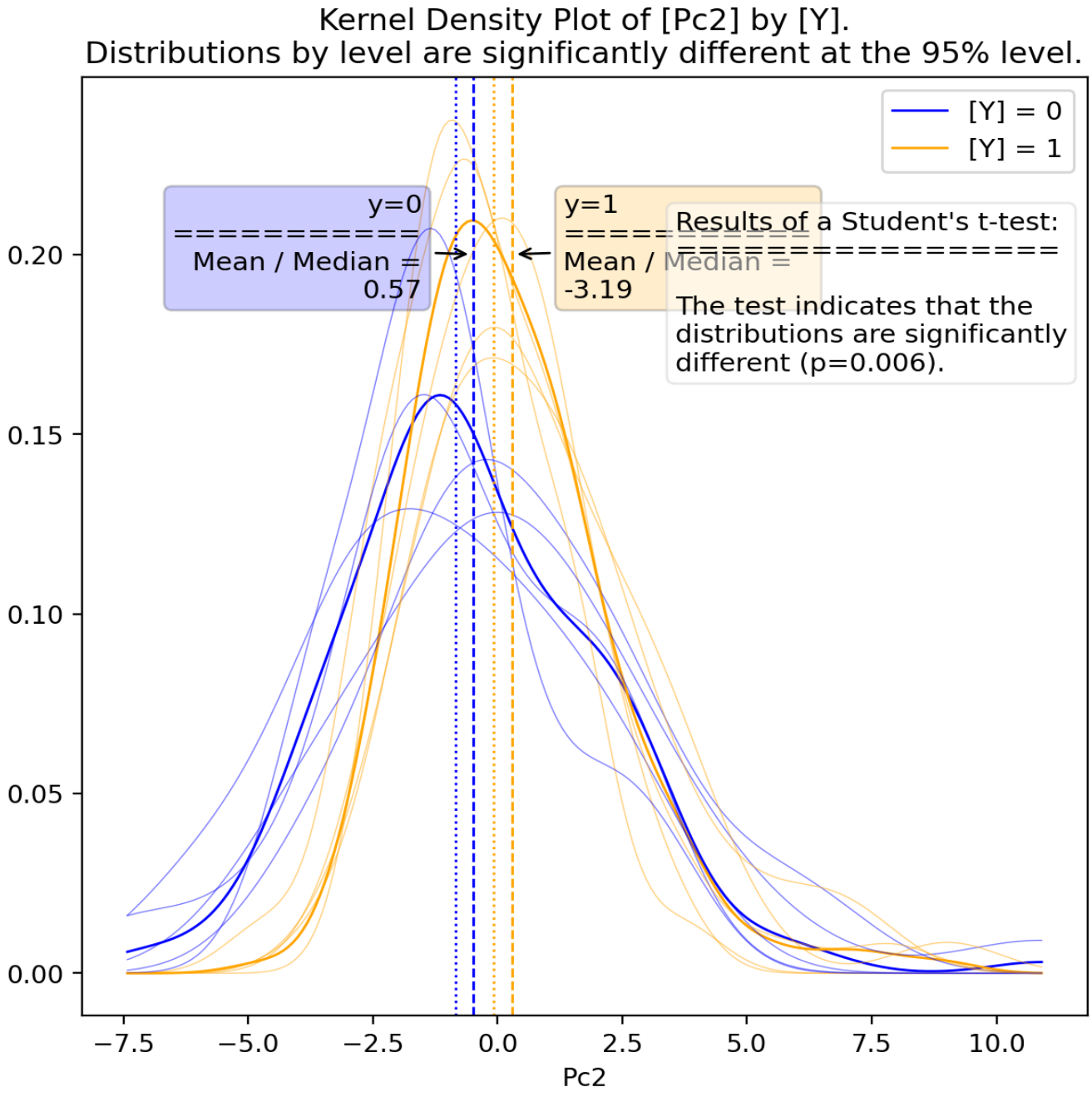
Univariate Report

Pc2 - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	mean	std
Fitted Coef.	0.221	0.153	0.138	0.081	0.109	0.140	0.053
Fitted p-Value	1.4e-04	8.5e-03	9.5e-03	1.4e-01	5.0e-02	4.9e-03	5.7e-02
Fitted Std. Err.	5.8e-02	5.8e-02	5.3e-02	5.4e-02	5.6e-02	5.0e-02	2.2e-03
Conf. Int. Lower	1.1e-01	3.9e-02	3.4e-02	-2.6e-02	-1.6e-04	4.2e-02	5.0e-02
Conf. Int. Upper	0.335	0.267	0.242	0.187	0.218	0.237	0.056
Train Accuracy	57.1%	55.1%	54.4%	50.9%	53.5%	54.0%	2.3%
Val Accuracy	39.3%	58.5%	50.0%	64.5%	55.6%	57.0%	9.5%
Train AUC	58.9%	57.0%	56.1%	52.5%	54.9%	55.6%	2.4%
Val AUC	41.7%	57.1%	52.7%	65.9%	59.0%	56.3%	9.0%
Train F1	60.3%	58.4%	57.7%	55.2%	55.7%	57.2%	2.1%
Test F1	41.3%	64.9%	50.8%	64.9%	60.9%	63.2%	10.3%
Train Precision	71.7%	70.7%	69.3%	66.7%	66.1%	68.6%	2.4%
Val Precision	54.2%	65.8%	65.2%	75.8%	77.8%	67.7%	9.5%
Train Recall	52.0%	49.7%	49.4%	47.1%	48.1%	49.1%	1.9%
Val Recall	33.3%	64.1%	41.7%	56.8%	50.0%	59.2%	12.1%
Train MCC	17.2%	13.6%	11.9%	4.8%	9.7%	11.0%	4.6%
Val MCC	-16.4%	14.0%	5.3%	31.7%	16.7%	12.3%	17.7%
Train Log-Loss	15.45	16.19	16.43	17.68	16.77	16.59	0.82
Val Log-Loss	21.86	14.97	18.02	12.80	16.02	15.49	3.43

Univariate Report

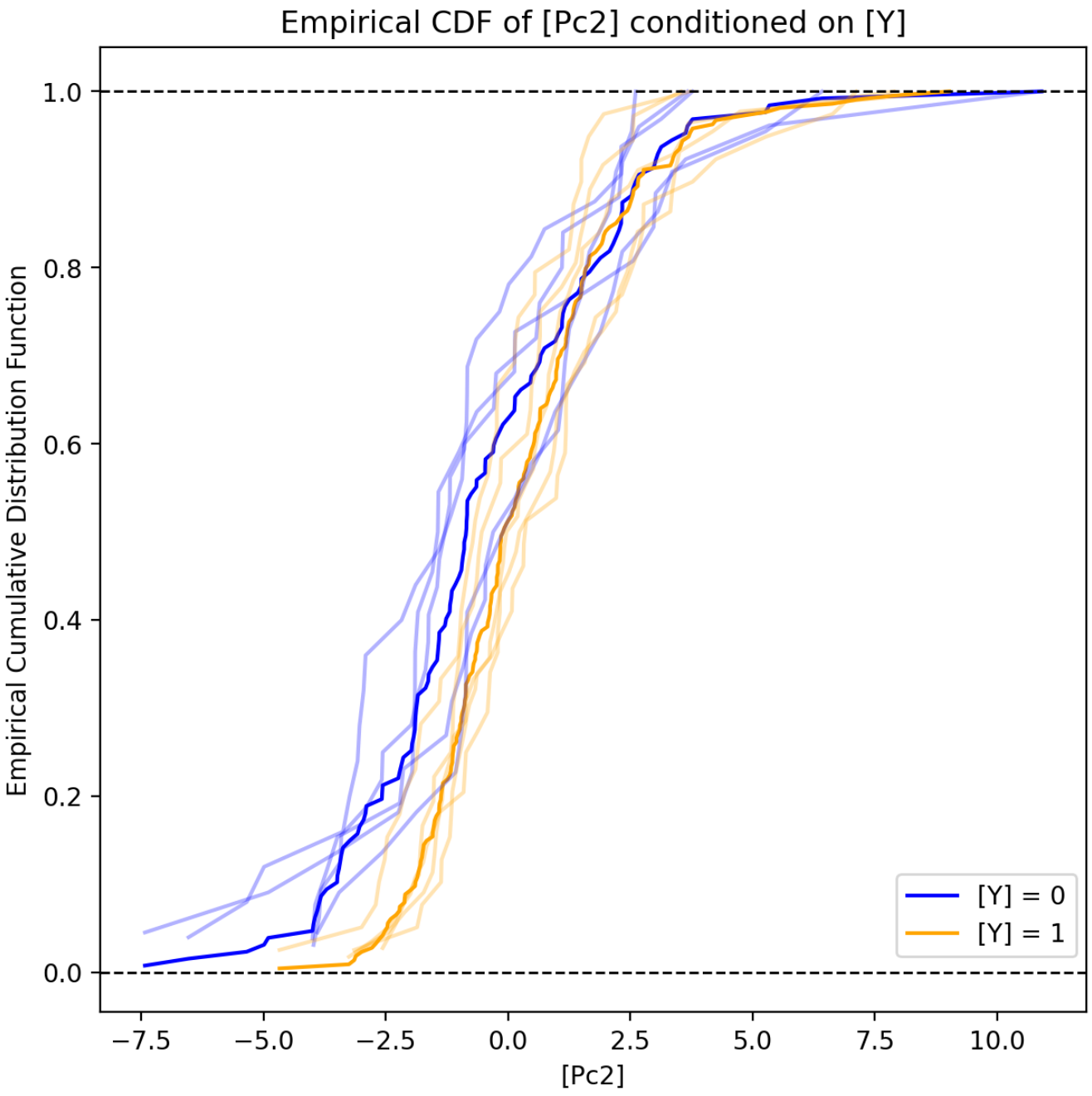
Pc2 - Kernel Density Plot



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

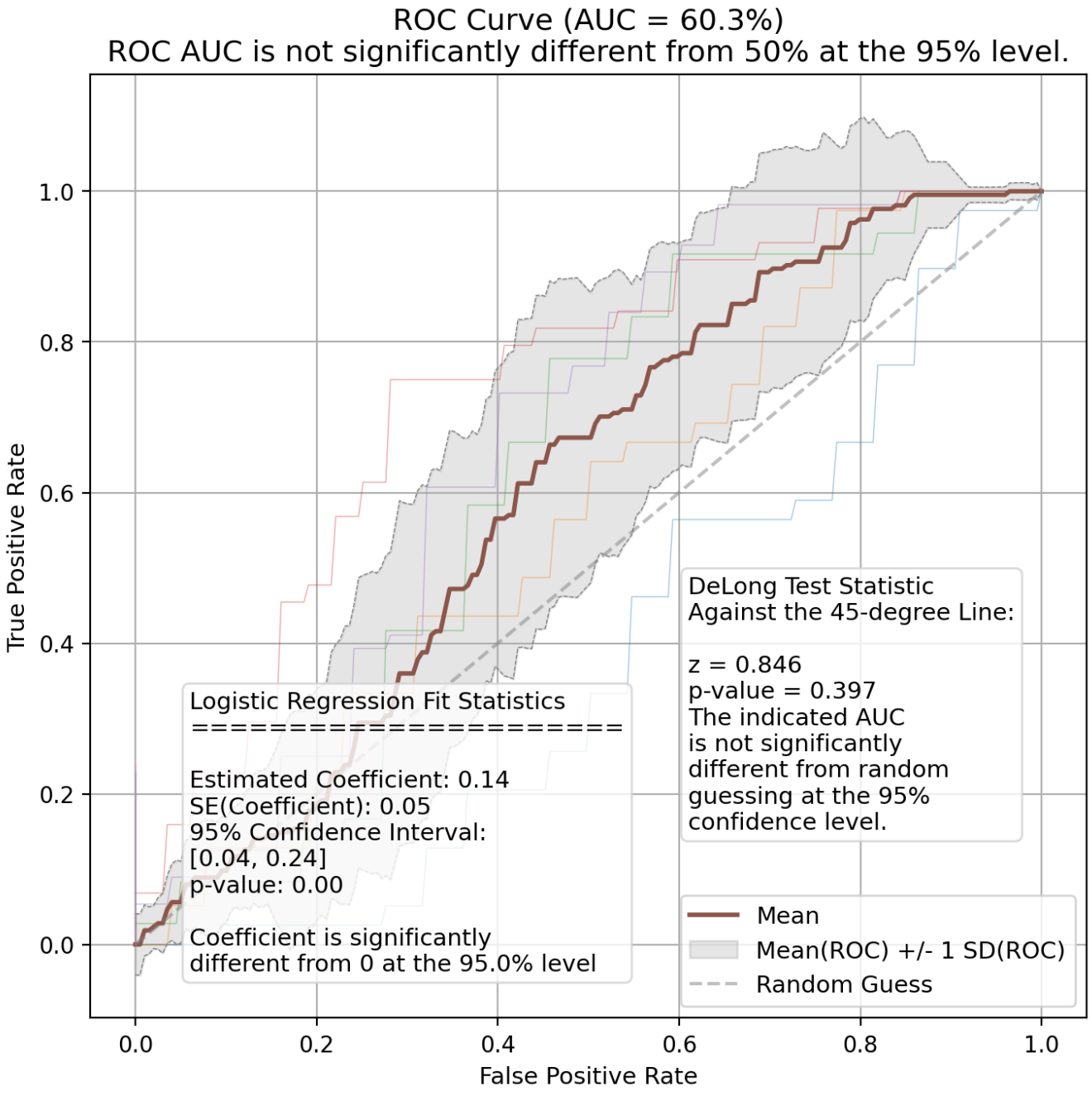
Pc2 - Empirical CDF Plot



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

Pc2 - ROC Curve

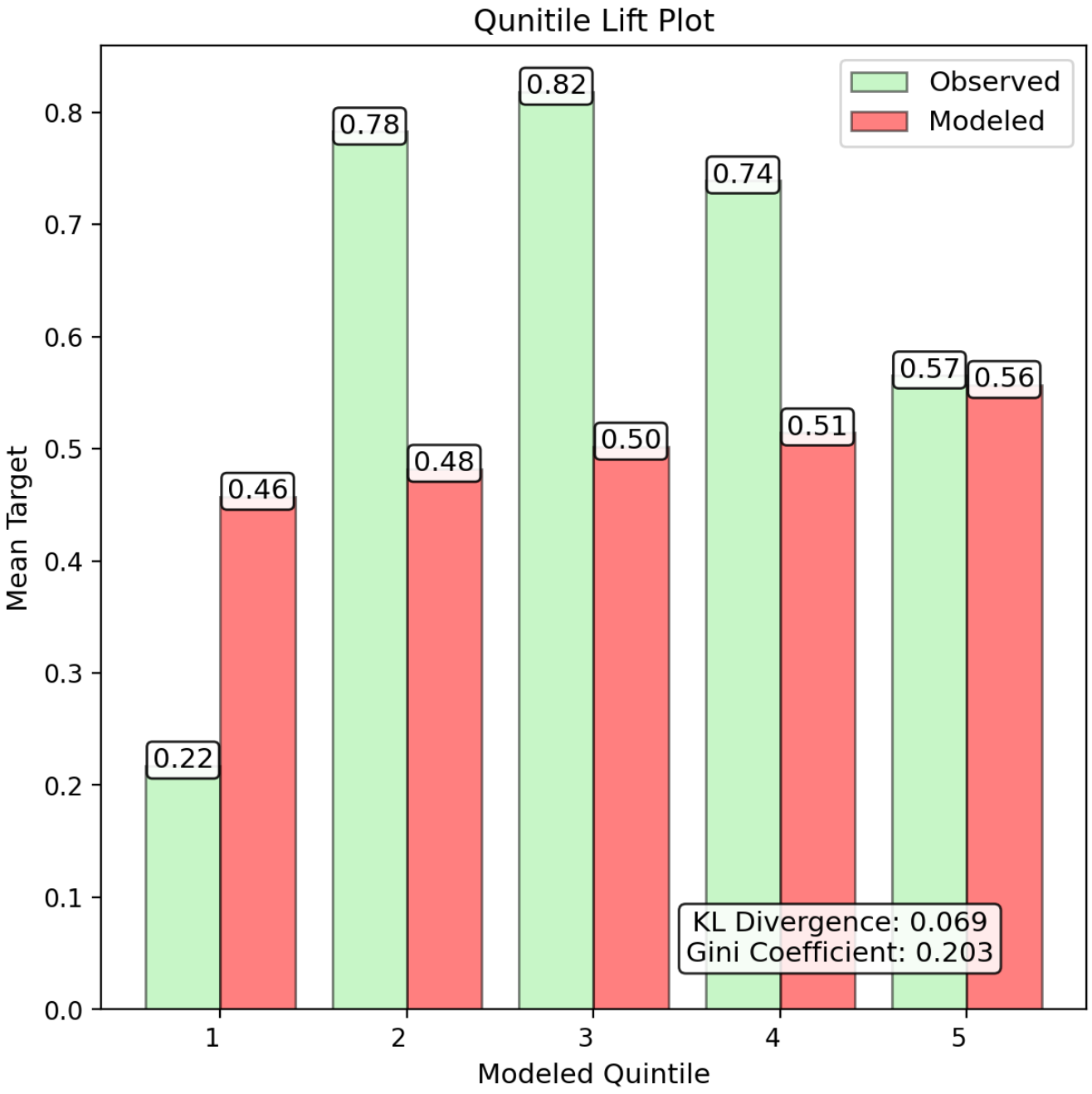


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

Pc2 - 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.