Univariate Analysis Report

Cancer Model

2024-02-02

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 sortied results. This is a limitation of the univariate analysis, and it is important to keep this in mind when interpreting the results. It is also important to consider further that depending on the purpose of the model, the most appropriate features may not be the ones with the highest average test scores, if a different metric is more important.

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

	Accuracy	Precision	Recall	AUC	F1	MCC	Ave.
mean_texture	27.2%	40.0%	33.8%	25.0%	36.6%	-4.85e-01	19.0%
mean_radius	12.3%	17.8%	11.3%	12.6%	13.8%	-7.42e-01	-1.07e-02
mean_perimeter	12.3%	17.8%	11.3%	12.6%	13.8%	-7.42e-01	-1.07e-02

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 3 of the variables passed to the Cancer Model.

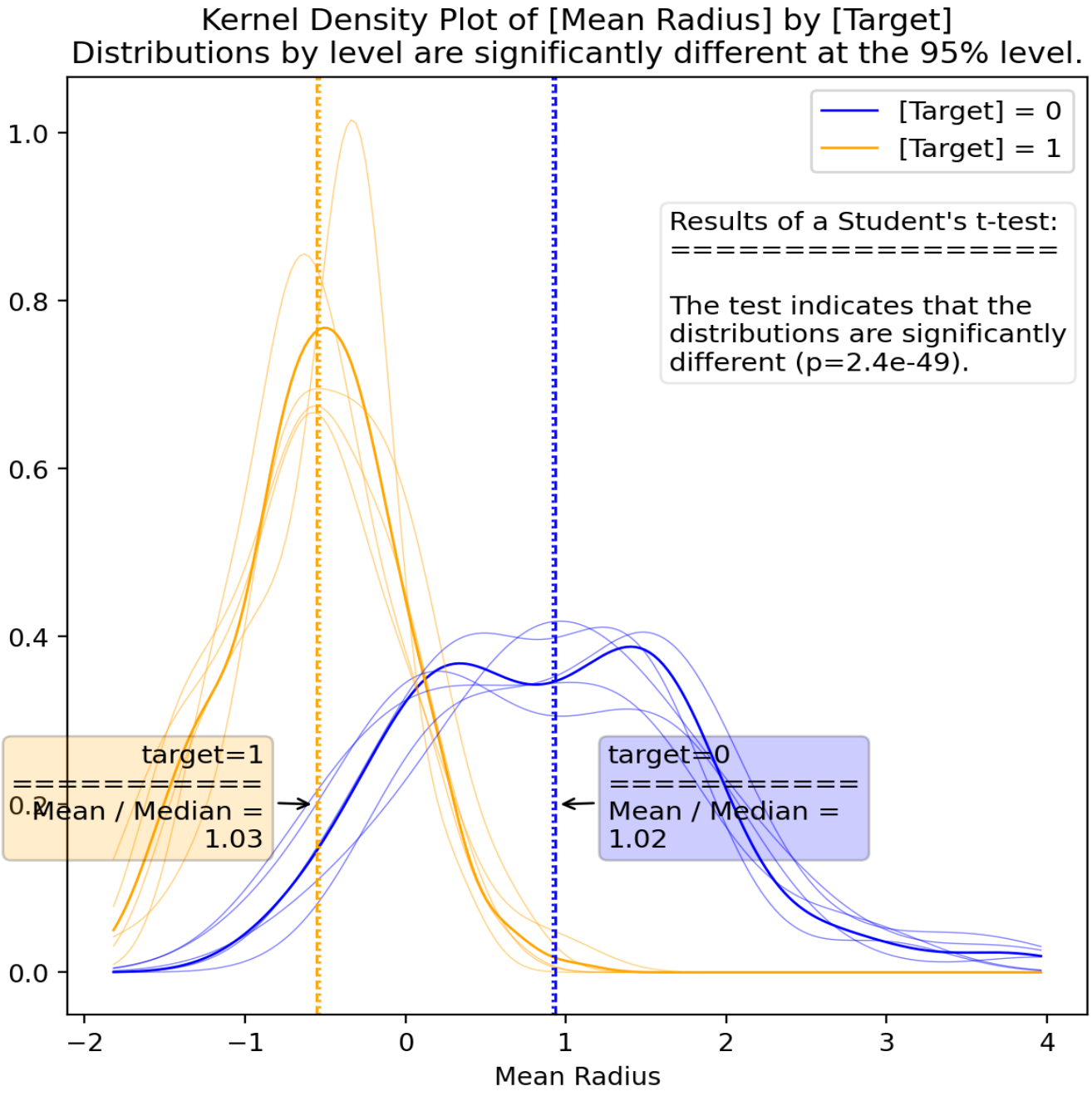
Univariate Report

Mean Radius - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Agg. Mean	Agg. SD
Fitted Coef.	-3.3e+00	-3.6e+00	-3.1e+00	-3.3e+00	-3.1e+00	1.1e-02	2.1e-01
Fitted p-Value	1.2e-21	2.6e-20	5.2e-22	2.1e-20	6.1e-21	9.1e-02	1.2e-20
Fitted Std. Err.	0.349	0.392	0.318	0.353	0.334	0.006	0.028
Conf. Int. Lower	-4.0e+00	-4.4e+00	-3.7e+00	-4.0e+00	-3.8e+00	-1.7e-03	2.7e-01
Conf. Int. Upper	-2.6e+00	-2.9e+00	-2.4e+00	-2.6e+00	-2.5e+00	2.4e-02	1.6e-01
Train Accuracy	85.9%	87.4%	84.8%	85.7%	85.4%	14.1%	1.0%
Val Accuracy	84.8%	78.3%	91.1%	86.8%	88.2%	12.3%	4.8%
Train AUC	85.7%	87.4%	84.8%	85.1%	85.1%	14.3%	1.1%
Val AUC	83.7%	78.4%	89.7%	88.0%	88.5%	12.6%	4.7%
Train F1	88.5%	90.1%	87.5%	88.3%	88.0%	16.3%	1.0%
Test F1	88.0%	79.3%	93.3%	89.4%	90.9%	13.8%	5.4%
Train Precision	90.7%	93.0%	90.0%	89.2%	89.6%	21.0%	1.5%
Val Precision	88.0%	81.5%	92.5%	95.2%	94.3%	17.8%	5.6%
Train Recall	86.4%	87.3%	85.0%	87.5%	86.5%	13.3%	1.0%
Val Recall	88.0%	77.2%	94.2%	84.3%	87.7%	11.3%	6.2%
Train MCC	70.5%	73.1%	68.5%	69.8%	69.5%	-70.4%	1.7%
Val MCC	67.3%	56.6%	80.2%	73.0%	74.7%	-74.2%	8.9%
Train Log-Loss	5.08	4.54	5.46	5.16	5.26	30.97	0.34
Val Log-Loss	5.47	7.82	3.19	4.76	4.24	31.62	1.73

Univariate Report

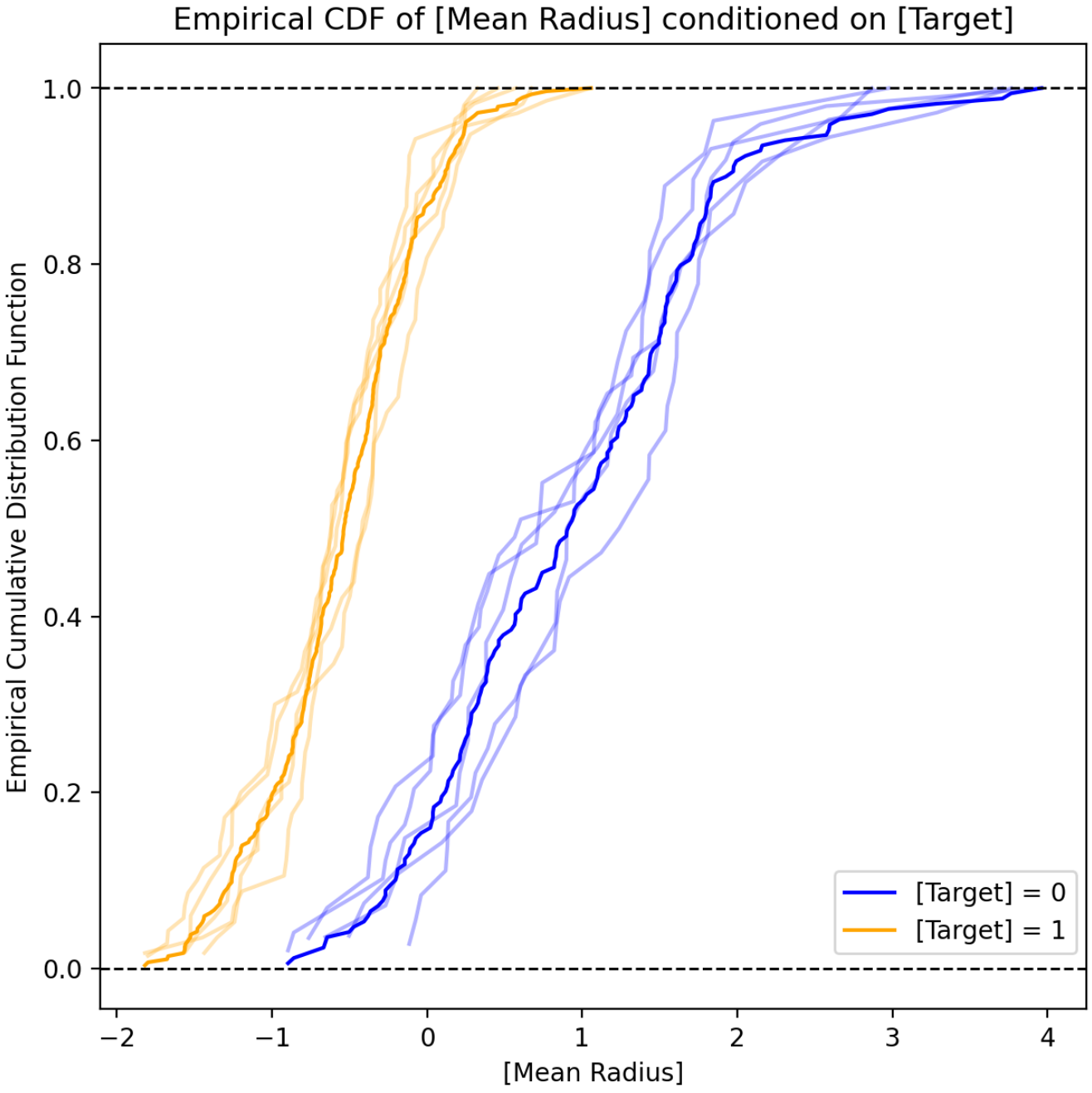
Mean Radius - Kernel Density Plot



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

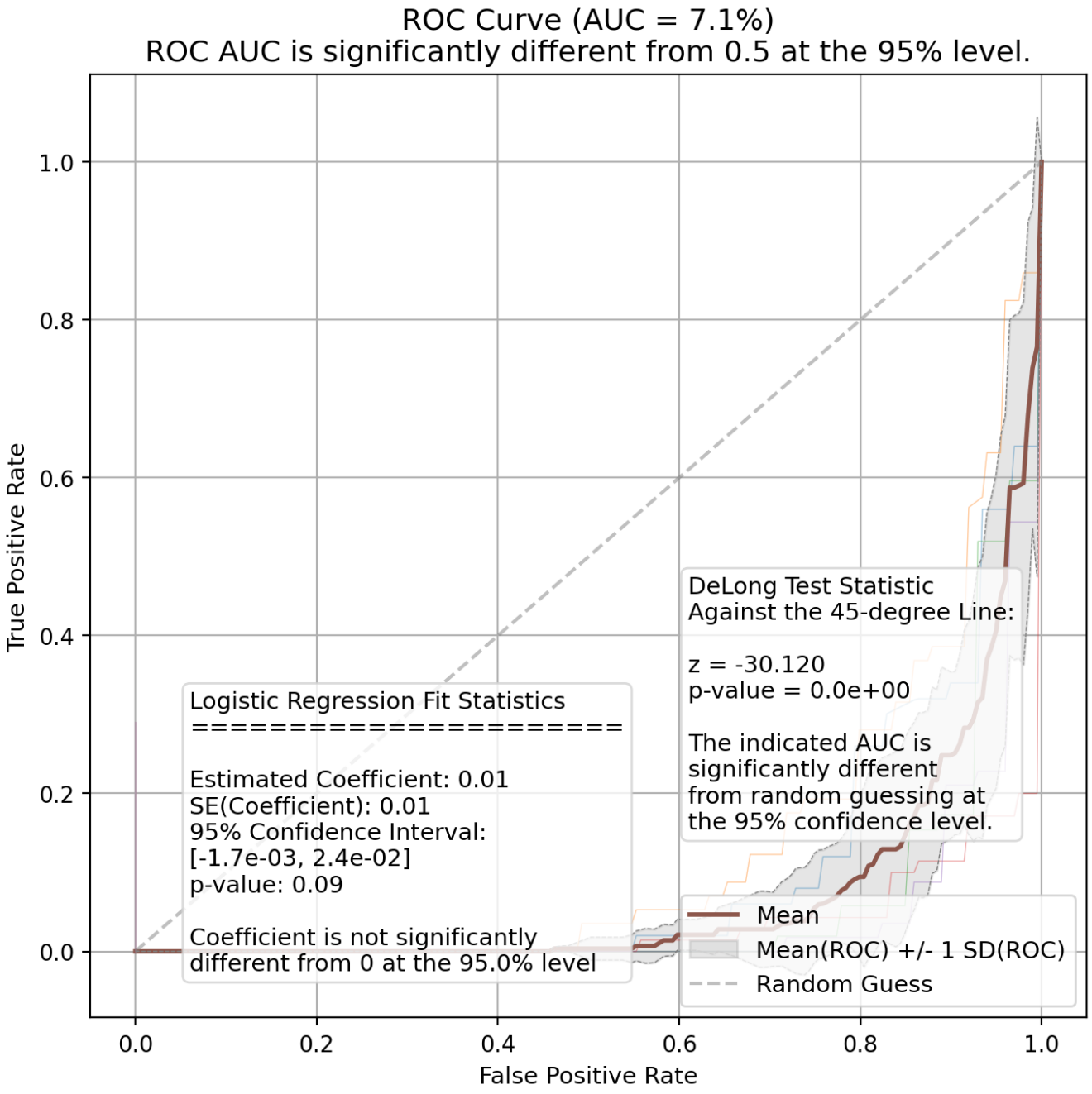
Mean Radius - Empirical CDF Plot



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

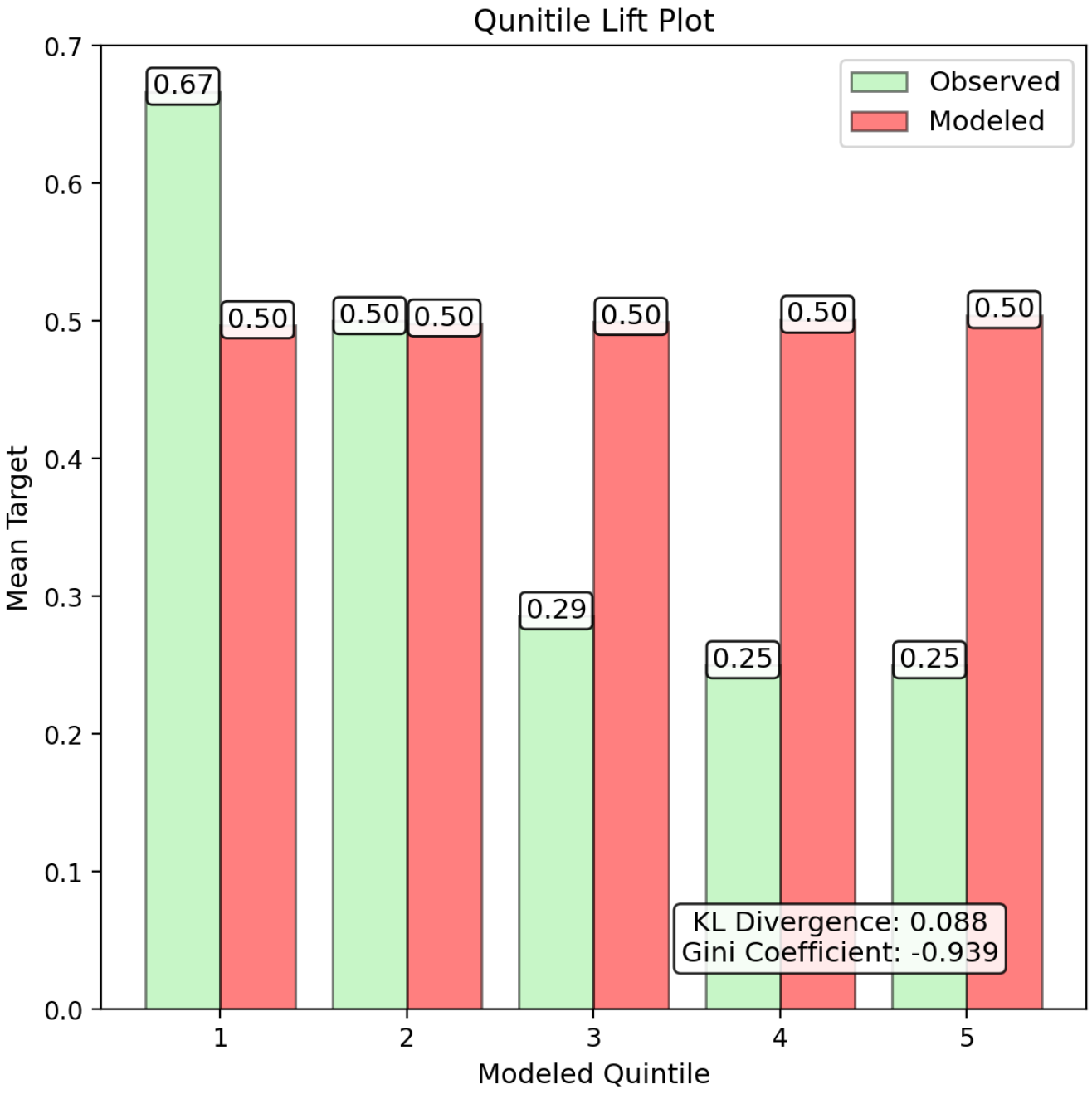


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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 the method from DeLong et al. (1988). In brief, the AUC is assumed to be normally distributed, and the z-score is calculated based on the AUC and the standard error. This z-score is compared to a +/- two standard deviations from a standard normal

distribution to get the p-value.

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.

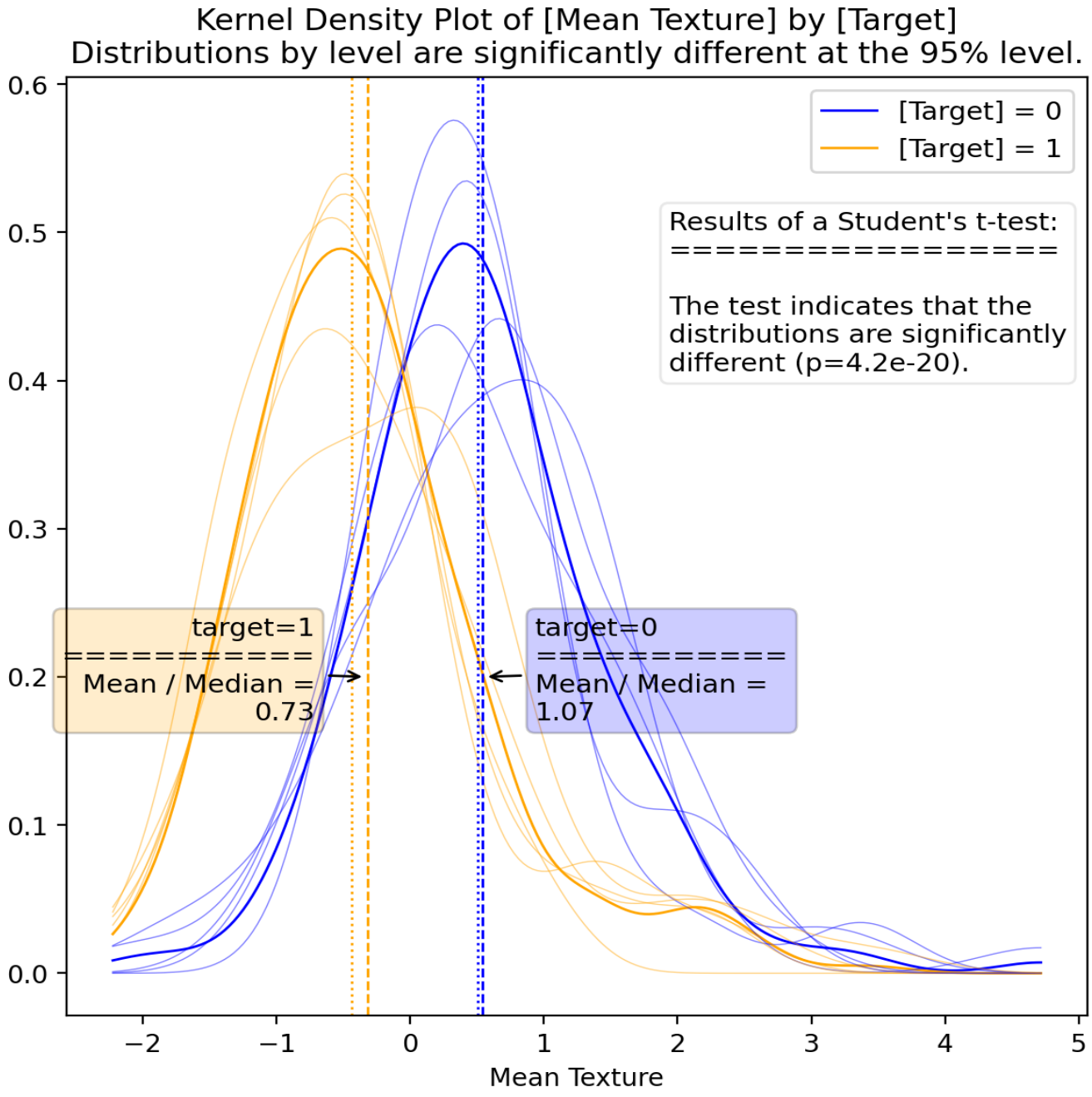
Univariate Report

Mean Texture - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Agg. Mean	Agg. SD
Fitted Coef.	-8.3e-01	-9.7e-01	-9.5e-01	-1.1e+00	-1.1e+00	1.7e-02	1.0e-01
Fitted p-Value	9.2e-11	1.1e-11	4.5e-12	8.2e-13	1.9e-13	6.6e-04	3.9e-11
Fitted Std. Err.	0.129	0.143	0.138	0.153	0.145	0.005	0.009
Conf. Int. Lower	-1.1e+00	-1.3e+00	-1.2e+00	-1.4e+00	-1.4e+00	7.0e-03	1.2e-01
Conf. Int. Upper	-5.8e-01	-6.9e-01	-6.8e-01	-7.9e-01	-7.8e-01	2.6e-02	8.6e-02
Train Accuracy	71.8%	71.6%	71.8%	74.8%	73.2%	27.3%	1.4%
Val Accuracy	78.5%	76.4%	77.2%	62.3%	70.6%	27.2%	6.7%
Train AUC	71.9%	72.4%	72.2%	74.7%	73.3%	27.0%	1.1%
Val AUC	77.2%	76.1%	76.5%	66.0%	71.7%	25.0%	4.7%
Train F1	76.1%	76.4%	75.7%	78.6%	77.1%	32.6%	1.2%
Test F1	82.8%	78.6%	82.0%	65.5%	75.7%	36.6%	7.0%
Train Precision	81.2%	84.2%	81.7%	82.7%	81.9%	39.0%	1.2%
Val Precision	83.7%	76.7%	85.4%	82.6%	84.8%	40.0%	3.5%
Train Recall	71.6%	69.9%	70.5%	75.0%	72.9%	28.0%	2.0%
Val Recall	82.0%	80.7%	78.8%	54.3%	68.4%	33.8%	11.7%
Train MCC	42.5%	42.8%	43.2%	48.4%	45.6%	-44.7%	2.5%
Val MCC	54.0%	52.4%	51.4%	30.6%	41.0%	-48.5%	10.0%
Train Log-Loss	10.16	10.22	10.16	9.09	9.64	26.22	0.49
Val Log-Loss	7.76	8.50	8.21	13.60	10.60	26.24	2.42

Univariate Report

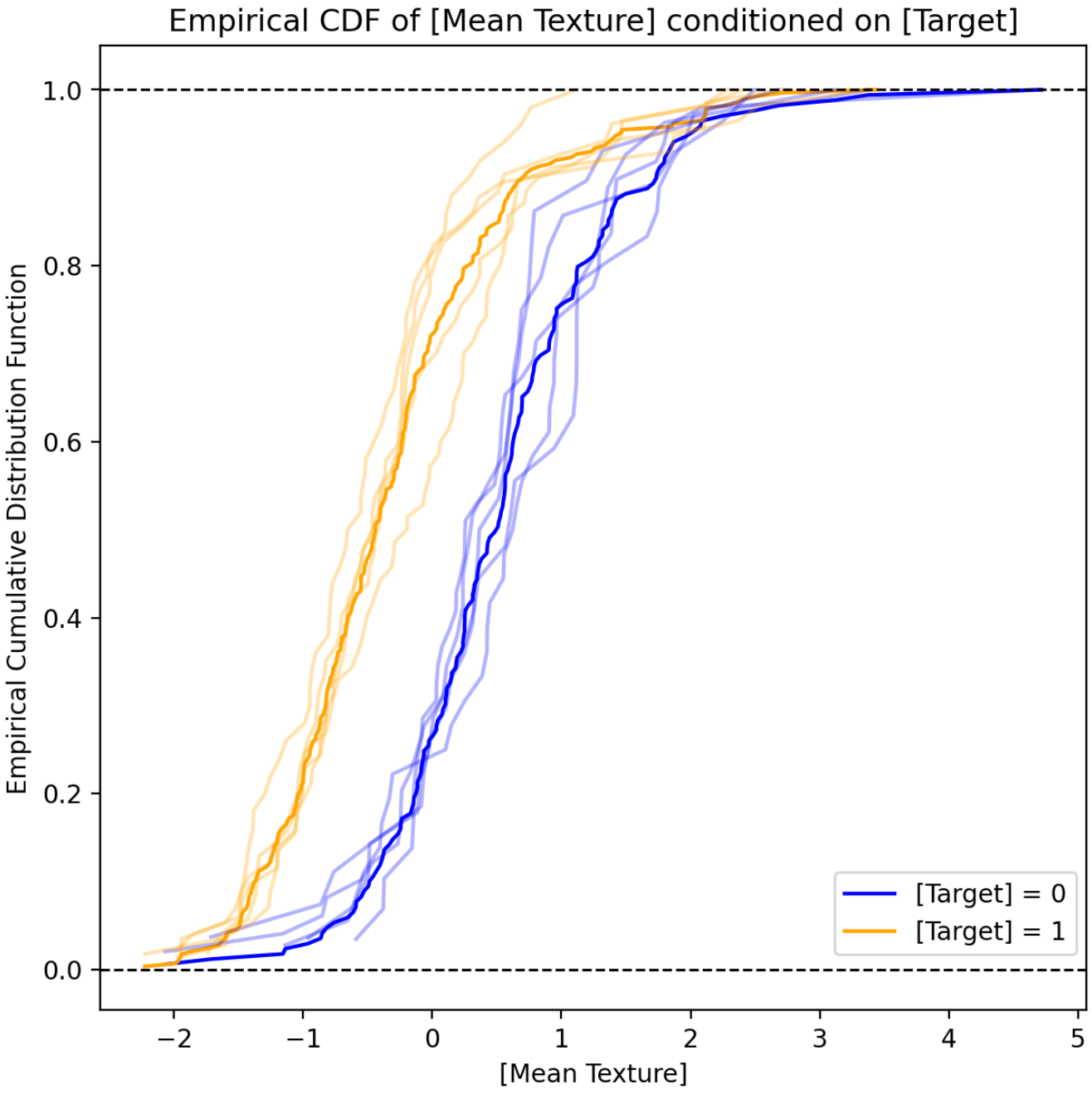
Mean Texture - Kernel Density Plot



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

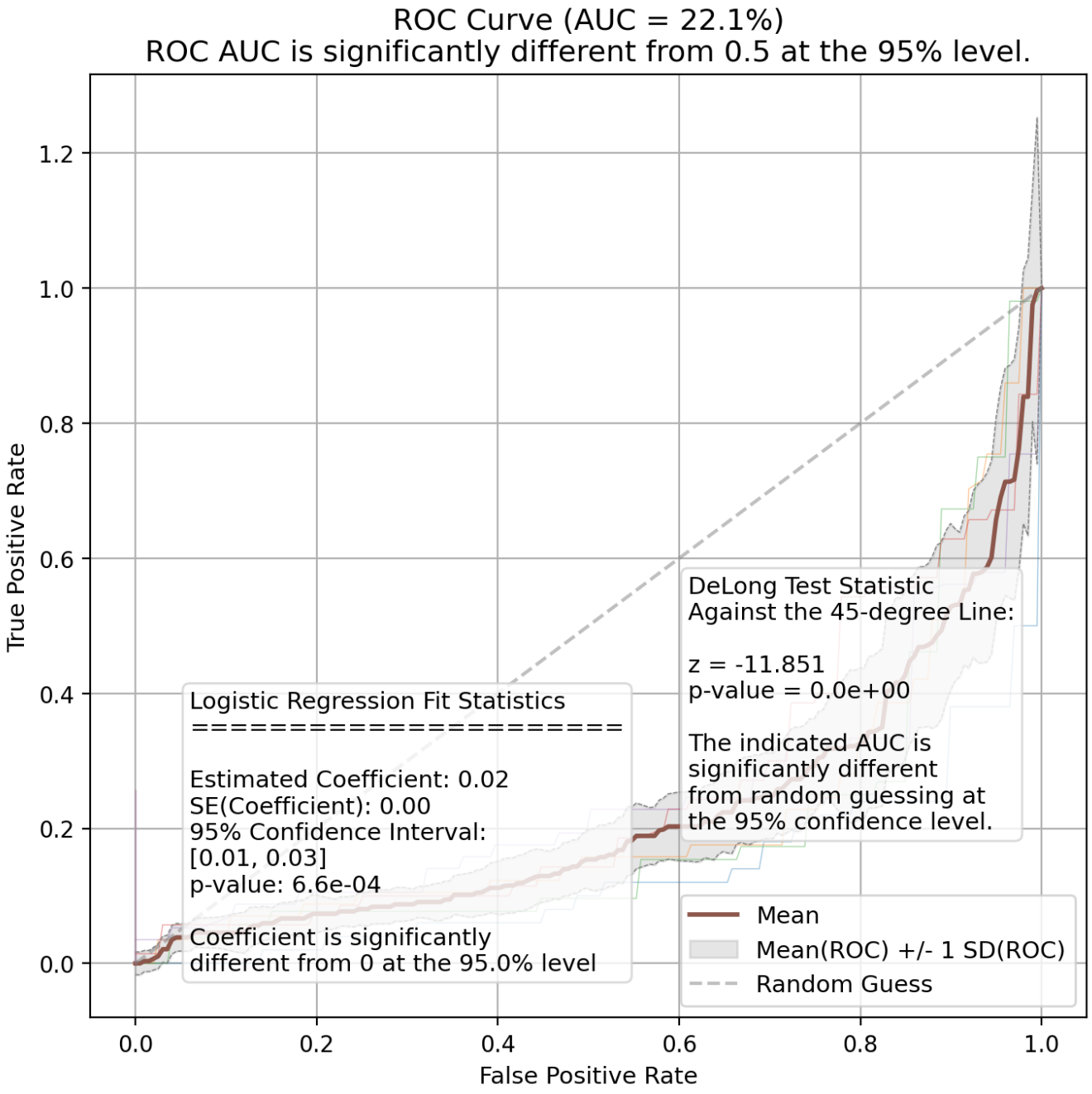
Mean Texture - Empirical CDF Plot



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

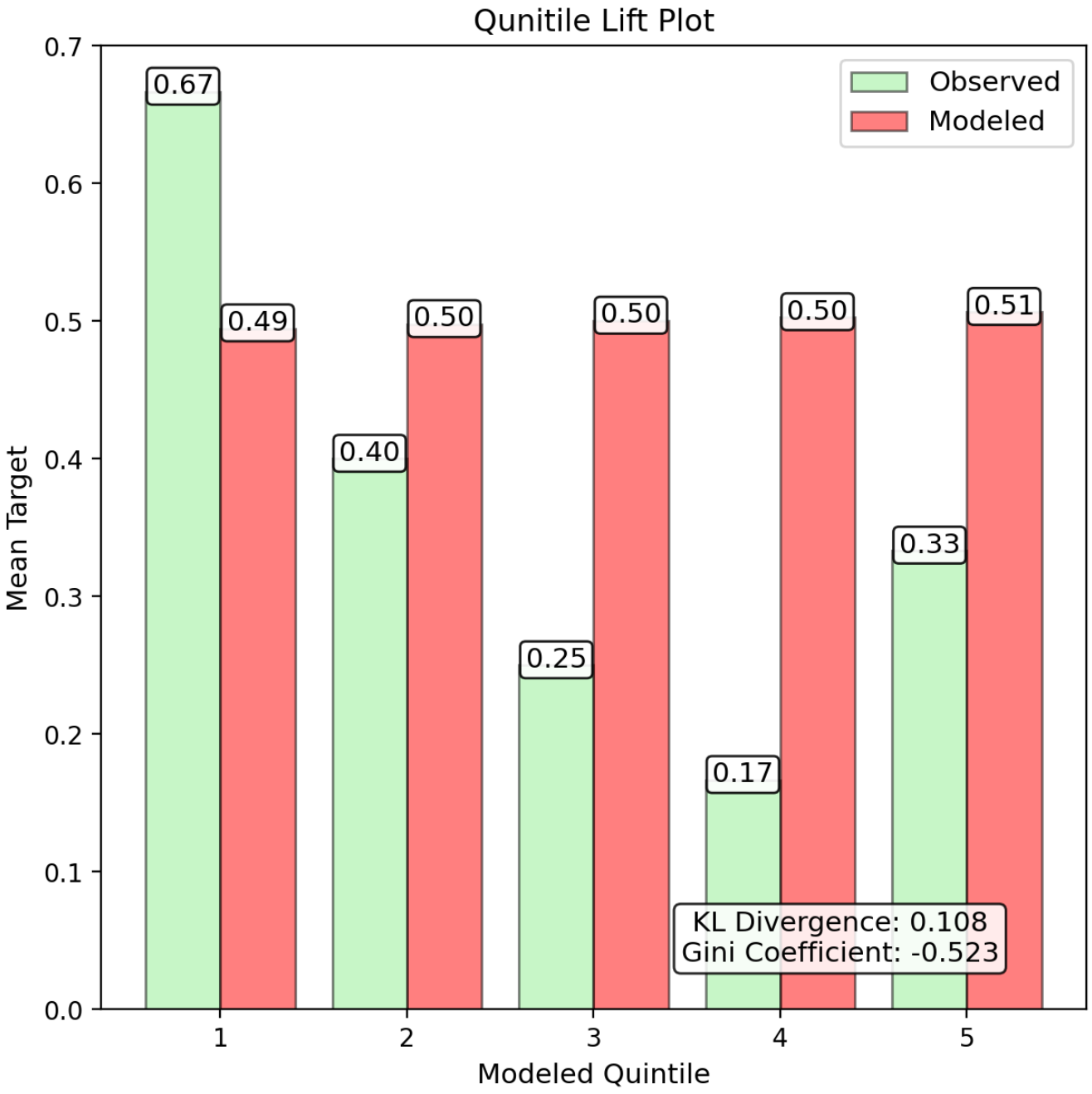


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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 the method from DeLong et al. (1988). In brief, the AUC is assumed to be normally distributed, and the z-score is calculated based on the AUC and the standard error. This z-score is compared to a +/- two standard deviations from a standard normal

distribution to get the p-value.

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.

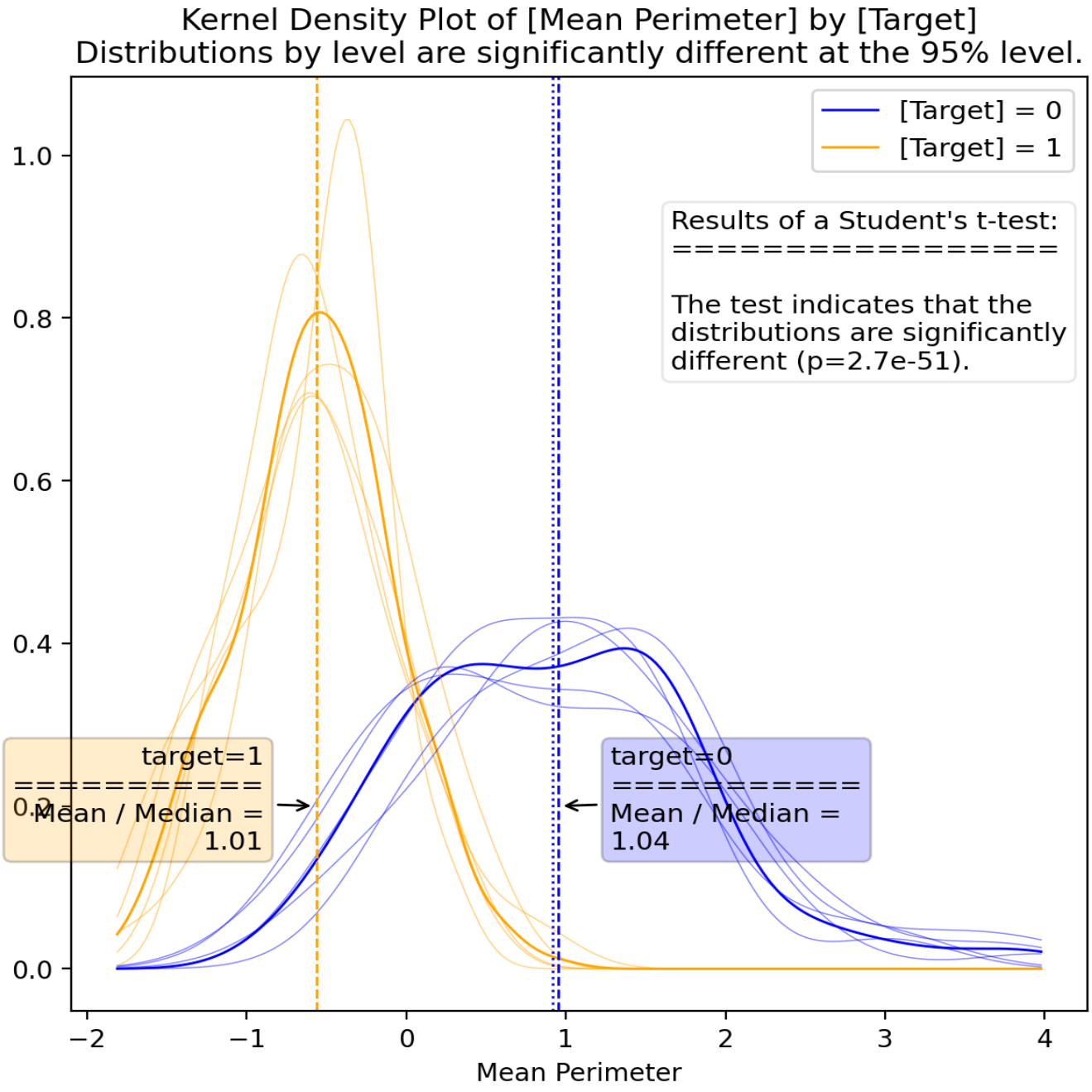
Univariate Report

Mean Perimeter - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Agg. Mean	Agg. SD
Fitted Coef.	-3.6e+00	-3.8e+00	-3.3e+00	-3.5e+00	-3.4e+00	1.4e-03	2.0e-01
Fitted p-Value	4.2e-22	1.3e-20	2.9e-22	9.1e-21	2.0e-21	1.6e-01	5.6e-21
Fitted Std. Err.	0.369	0.412	0.341	0.377	0.359	0.001	0.026
Conf. Int. Lower	-4.3e+00	-4.6e+00	-4.0e+00	-4.3e+00	-4.1e+00	-5.4e-04	2.5e-01
Conf. Int. Upper	-2.8e+00	-3.0e+00	-2.6e+00	-2.8e+00	-2.7e+00	3.3e-03	1.5e-01
Train Accuracy	87.0%	88.0%	85.9%	86.8%	86.5%	13.2%	0.8%
Val Accuracy	86.1%	83.0%	91.1%	86.8%	88.2%	12.3%	3.0%
Train AUC	86.9%	88.1%	85.9%	86.3%	86.2%	13.4%	0.8%
Val AUC	84.7%	83.1%	89.7%	88.0%	88.5%	12.6%	2.8%
Train F1	89.4%	90.5%	88.4%	89.3%	88.9%	15.4%	0.8%
Test F1	89.1%	83.9%	93.3%	89.4%	90.9%	13.8%	3.5%
Train Precision	91.6%	93.5%	91.0%	90.1%	90.5%	19.9%	1.3%
Val Precision	88.2%	85.5%	92.5%	95.2%	94.3%	17.8%	4.2%
Train Recall	87.3%	87.8%	85.9%	88.4%	87.3%	12.6%	0.9%
Val Recall	90.0%	82.5%	94.2%	84.3%	87.7%	11.3%	4.7%
Train MCC	72.7%	74.3%	70.7%	72.2%	71.7%	-72.3%	1.3%
Val MCC	69.8%	66.0%	80.2%	73.0%	74.7%	-74.2%	5.3%
Train Log-Loss	4.70	4.34	5.08	4.75	4.87	31.29	0.27
Val Log-Loss	5.02	6.12	3.19	4.76	4.24	31.62	1.07

Univariate Report

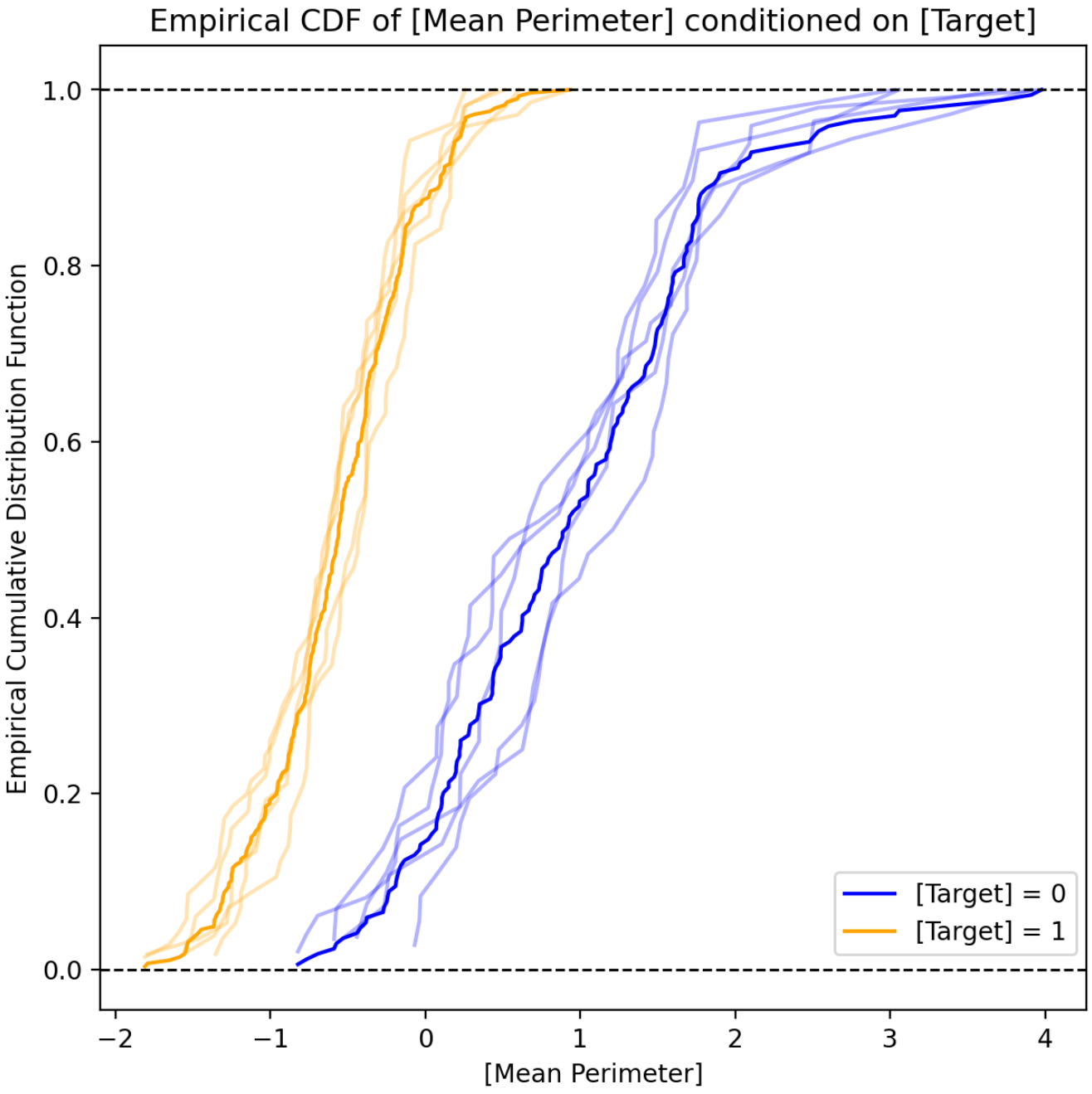
Mean Perimeter - Kernel Density Plot



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

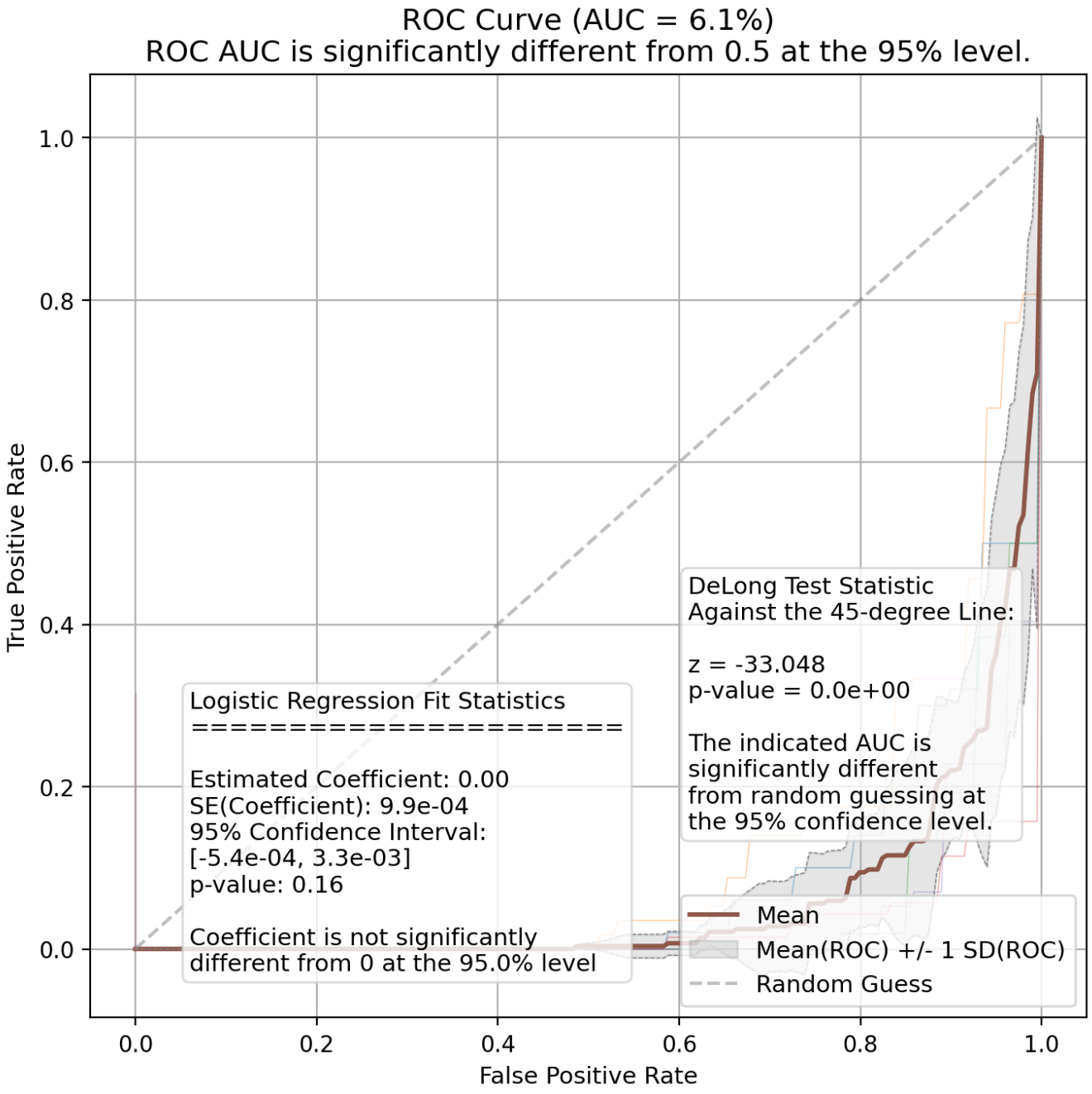
Mean Perimeter - Empirical CDF Plot



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

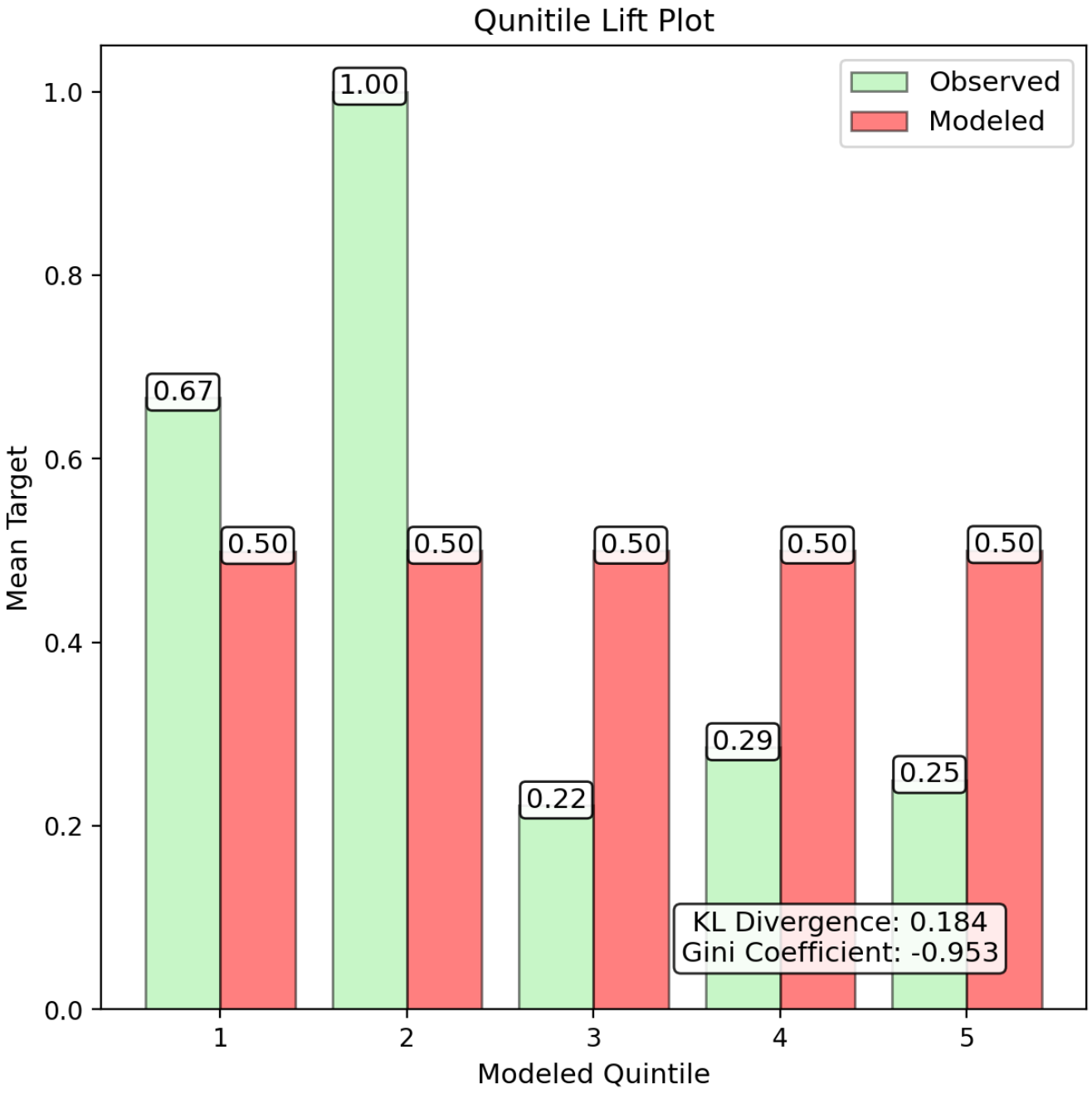


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, 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 the method from DeLong et al. (1988). In brief, the AUC is assumed to be normally distributed, and the z-score is calculated based on the AUC and the standard error. This z-score is compared to a +/- two standard deviations from a standard normal

distribution to get the p-value.

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.