model Univariate Analysis Report

2024-02-08

Overview

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_radius	13.7%	20.4%	12.9%	14.0%	15.8%	-7.12e-01	0.9%
mean_perimeter	13.0%	19.5%	12.3%	13.2%	15.1%	-7.26e-01	0.1%
worst_radius	9.3%	13.0%	7.8%	9.8%	9.8%	-8.01e-01	-5.06e-02
worst_perimeter	8.3%	12.0%	7.3%	8.6%	9.1%	-8.24e-01	-6.21e-02

This table shows an overview of the results for the variables in this file, representing those whose average test score are ranked between 27 and 30 of the variables passed to the model.

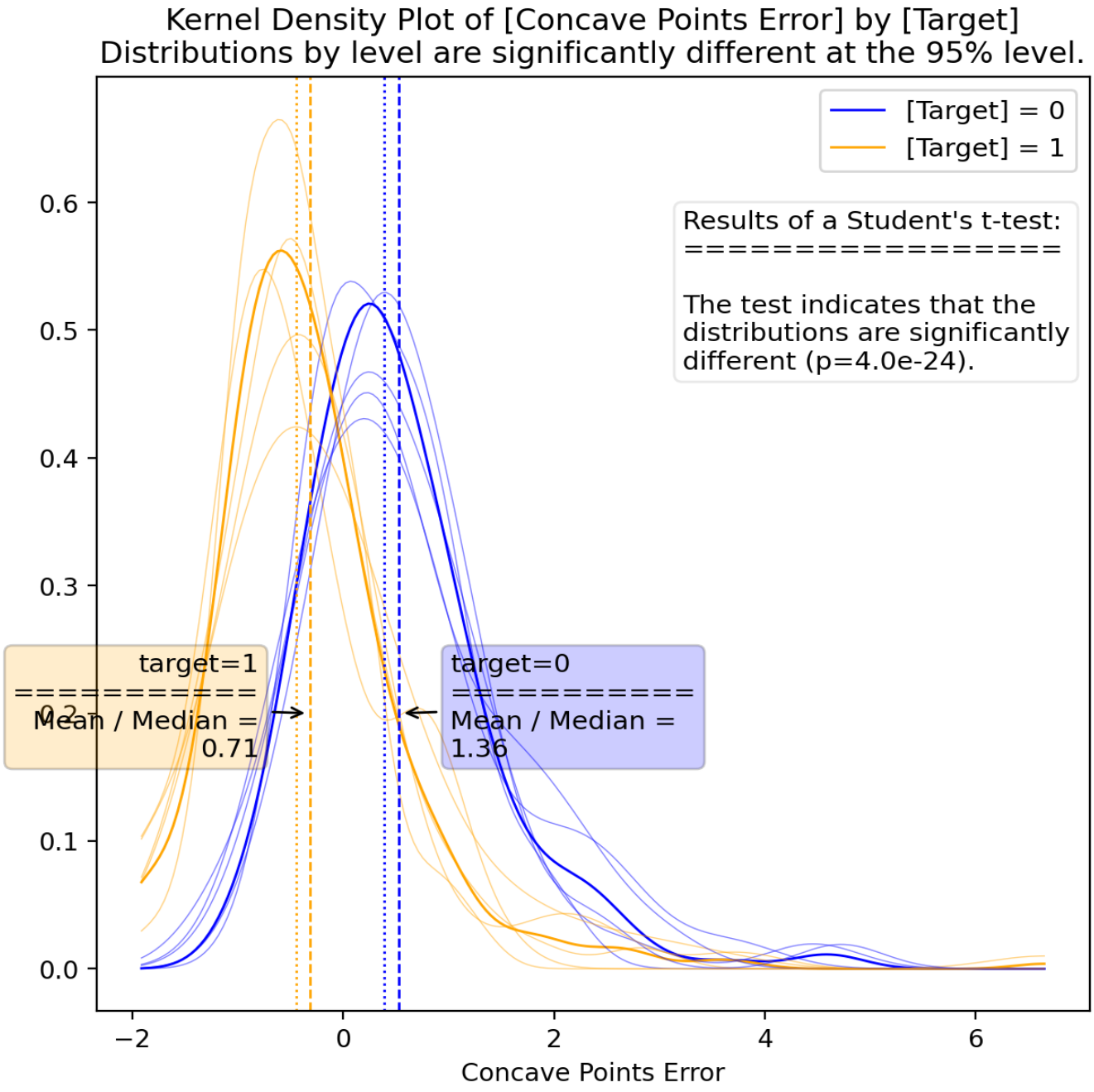
Univariate Report

Concave Points Error - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Agg. Mean	Agg. SD
Fitted Coef.	-9.9e-01	-1.4e+00	-1.0e+00	-8.5e-01	-1.1e+00	6.4e+00	2.1e-01
Fitted p-Value	6.5e-14	1.9e-18	8.6e-14	1.8e-11	3.3e-15	3.1e-01	7.8e-12
Fitted Std. Err.	0.132	0.160	0.136	0.127	0.144	6.313	0.013
Conf. Int. Lower	-1.2e+00	-1.7e+00	-1.3e+00	-1.1e+00	-1.4e+00	-6.0e+00	2.3e-01
Conf. Int. Upper	-7.3e-01	-1.1e+00	-7.5e-01	-6.0e-01	-8.5e-01	1.9e+01	1.8e-01
Train Accuracy	73.4%	74.0%	73.7%	72.1%	73.3%	26.9%	0.7%
Val Accuracy	73.9%	66.4%	73.4%	78.6%	73.0%	26.9%	4.4%
Train AUC	73.1%	74.2%	73.3%	71.6%	73.3%	27.1%	1.0%
Val AUC	72.1%	67.9%	73.3%	78.3%	72.5%	27.1%	3.7%
Train F1	77.8%	78.1%	78.0%	77.1%	77.5%	31.1%	0.4%
Test F1	78.9%	69.6%	78.5%	81.8%	78.0%	31.1%	4.6%
Train Precision	82.0%	83.1%	81.1%	80.7%	81.9%	38.1%	0.9%
Val Precision	77.6%	79.7%	84.1%	84.0%	82.1%	38.1%	2.8%
Train Recall	74.1%	73.7%	75.1%	73.7%	73.5%	26.3%	0.6%
Val Recall	80.4%	61.8%	73.6%	79.7%	74.3%	26.3%	7.4%
Train MCC	45.1%	47.0%	45.7%	42.1%	45.5%	-44.7%	1.8%
Val MCC	44.7%	34.7%	44.7%	56.0%	43.8%	-44.7%	7.6%
Train Log-Loss	9.60	9.35	9.48	10.04	9.61	26.35	0.26
Val Log-Loss	9.40	12.11	9.59	7.70	9.72	26.35	1.57

Univariate Report

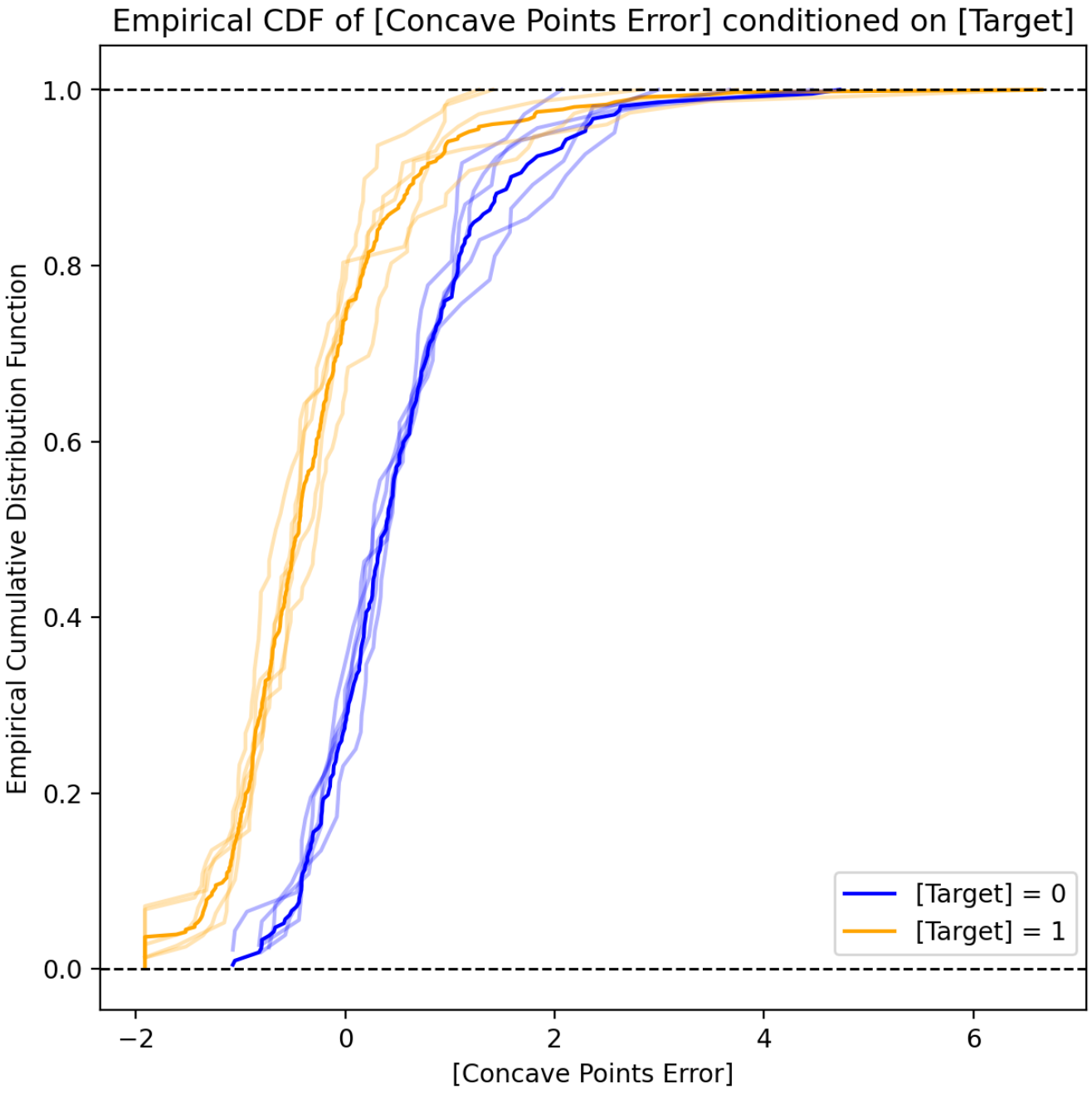
Concave Points Error - Kernel Density Plot



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

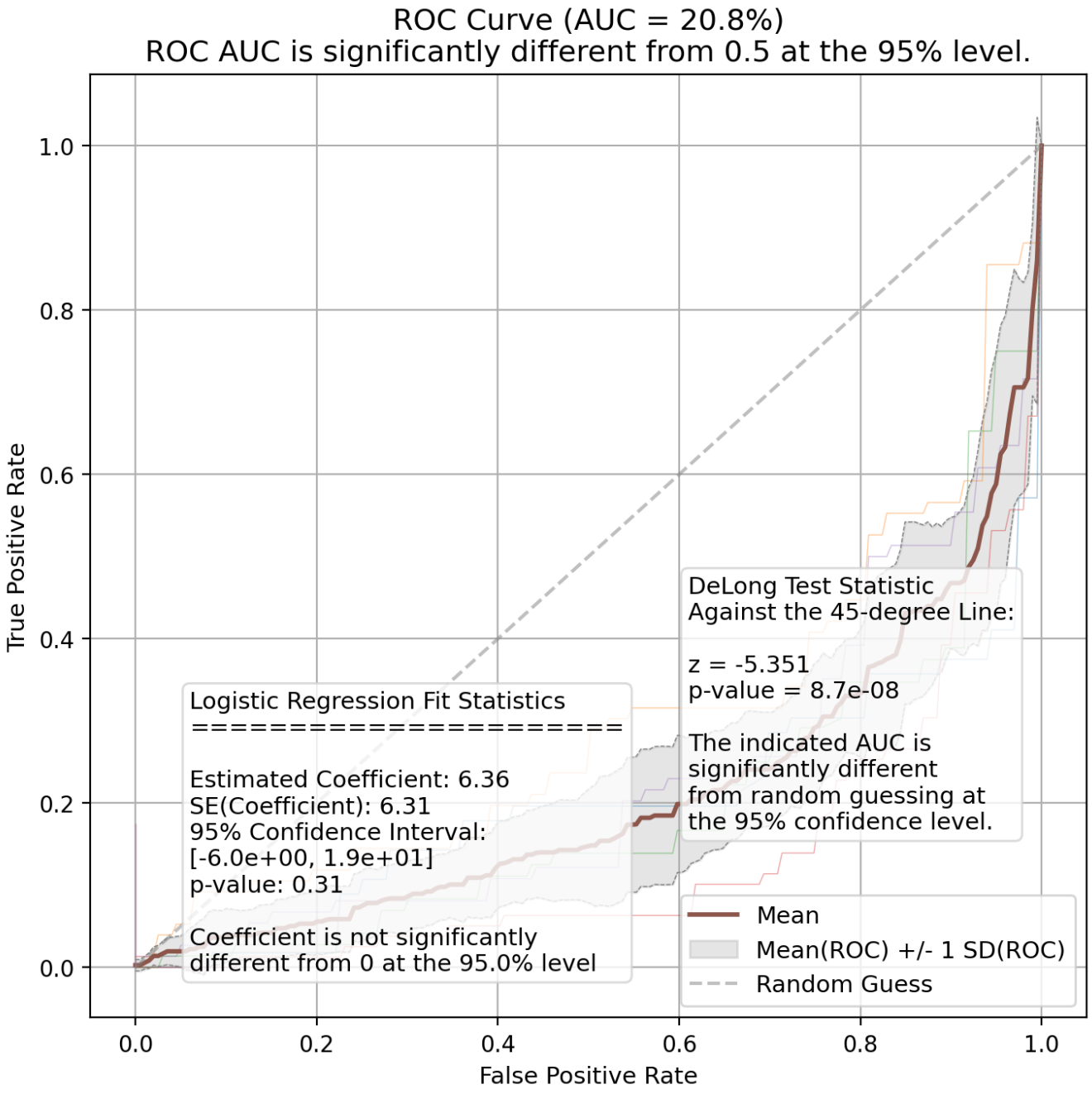
Concave Points Error - Empirical CDF Plot



This plot shows the empirical cumulative distribution function for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the cumulative distribution of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data, and whether or not it is reasonable to assume that the data is drawn from different distributions.

Univariate Report

Concave Points Error - ROC Curve

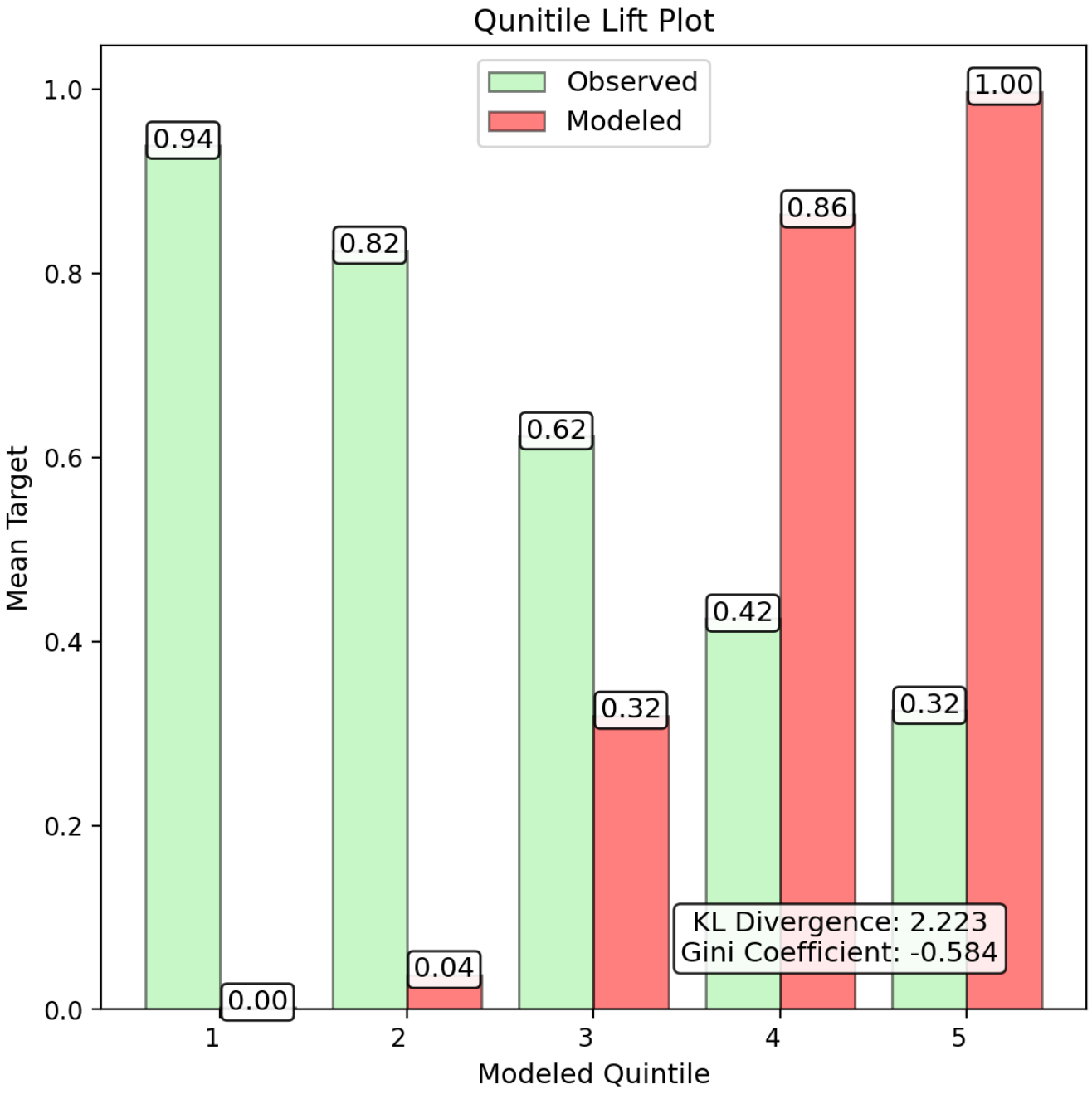


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, and no other features. Annotations are on the plot to help understand the results of the model, including the coefficient, standard error, and p-value for the feature variable. The cross-validation folds are used to create the grey region around the mean ROC curve to help understand the variability of the data. Significance of the ROC curve is determined based on 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

Concave Points Error - Quintile Lift



The quintile lift plot is meant to show the power of the single feature to discriminate between the highest and lowest quintiles of the target variable.

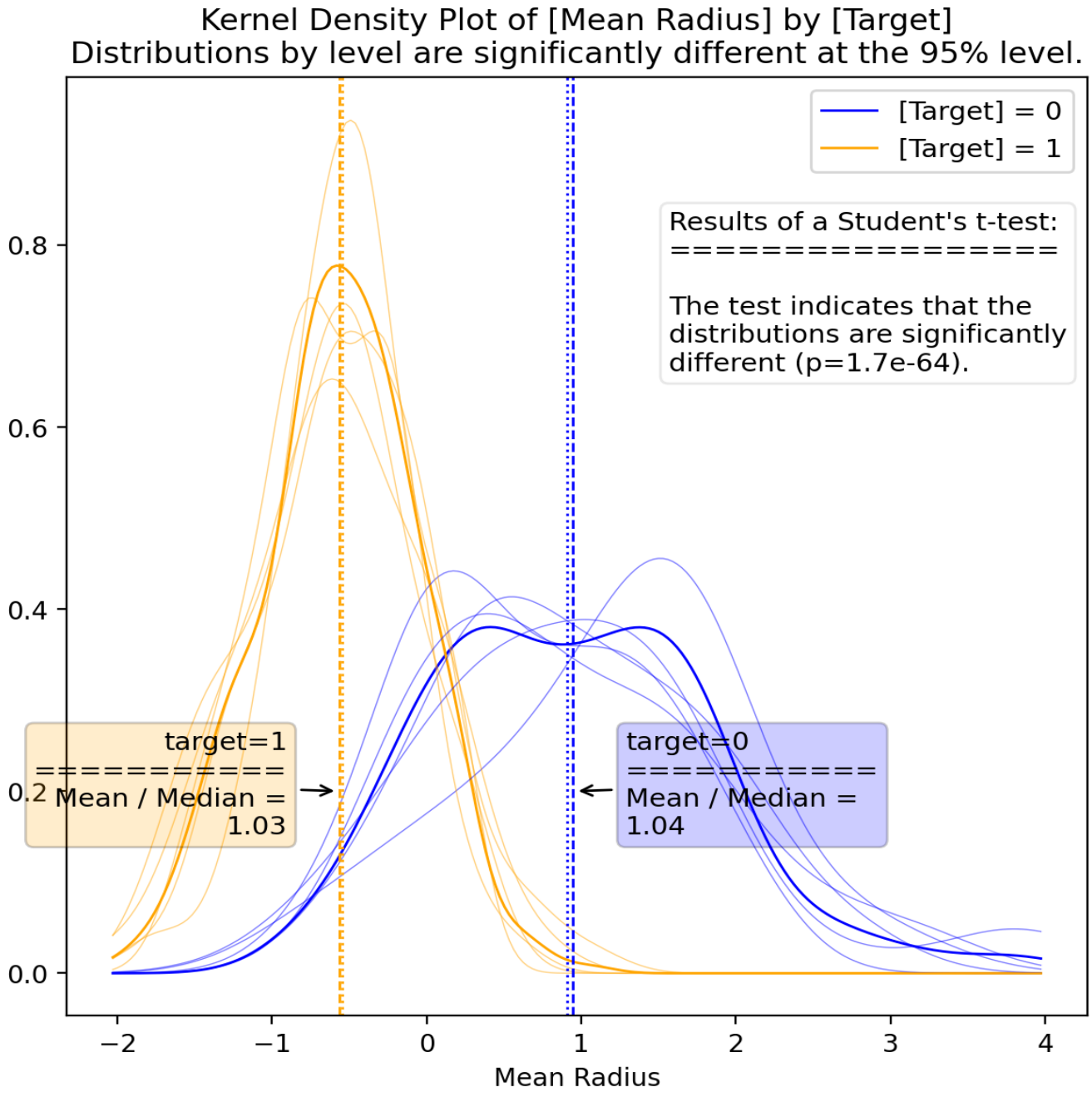
Univariate Report

Mean Radius - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Agg. Mean	Agg. SD
Fitted Coef.	-3.5e+00	-3.6e+00	-3.4e+00	-3.4e+00	-3.5e+00	1.0e-02	8.0e-02
Fitted p-Value	3.6e-27	2.8e-25	3.2e-26	7.4e-25	5.0e-26	7.4e-02	3.1e-25
Fitted Std. Err.	0.320	0.347	0.324	0.330	0.331	0.006	0.010
Conf. Int. Lower	-4.1e+00	-4.3e+00	-4.1e+00	-4.0e+00	-4.1e+00	-1.0e-03	9.8e-02
Conf. Int. Upper	-2.8e+00	-2.9e+00	-2.8e+00	-2.7e+00	-2.8e+00	2.2e-02	6.5e-02
Train Accuracy	86.6%	86.4%	85.2%	85.6%	87.2%	13.7%	0.8%
Val Accuracy	84.8%	85.2%	90.8%	85.5%	82.6%	13.7%	3.0%
Train AUC	86.4%	85.9%	85.1%	85.4%	86.9%	14.0%	0.7%
Val AUC	83.5%	85.6%	89.8%	85.7%	82.7%	14.0%	2.7%
Train F1	89.1%	89.0%	87.8%	88.4%	89.6%	15.8%	0.7%
Test F1	87.7%	87.7%	93.1%	87.6%	85.9%	15.8%	2.7%
Train Precision	91.3%	90.4%	90.0%	90.6%	90.9%	20.4%	0.5%
Val Precision	86.2%	91.4%	93.1%	90.5%	89.7%	20.4%	2.5%
Train Recall	87.0%	87.5%	85.6%	86.3%	88.3%	12.9%	1.1%
Val Recall	89.3%	84.2%	93.1%	84.8%	82.4%	12.9%	4.3%
Train MCC	71.8%	71.2%	69.3%	69.6%	73.1%	-71.2%	1.6%
Val MCC	67.8%	69.7%	79.5%	70.4%	63.7%	-71.2%	5.8%
Train Log-Loss	4.84	4.92	5.33	5.18	4.60	31.10	0.29
Val Log-Loss	5.48	5.32	3.31	5.23	6.27	31.10	1.09

Univariate Report

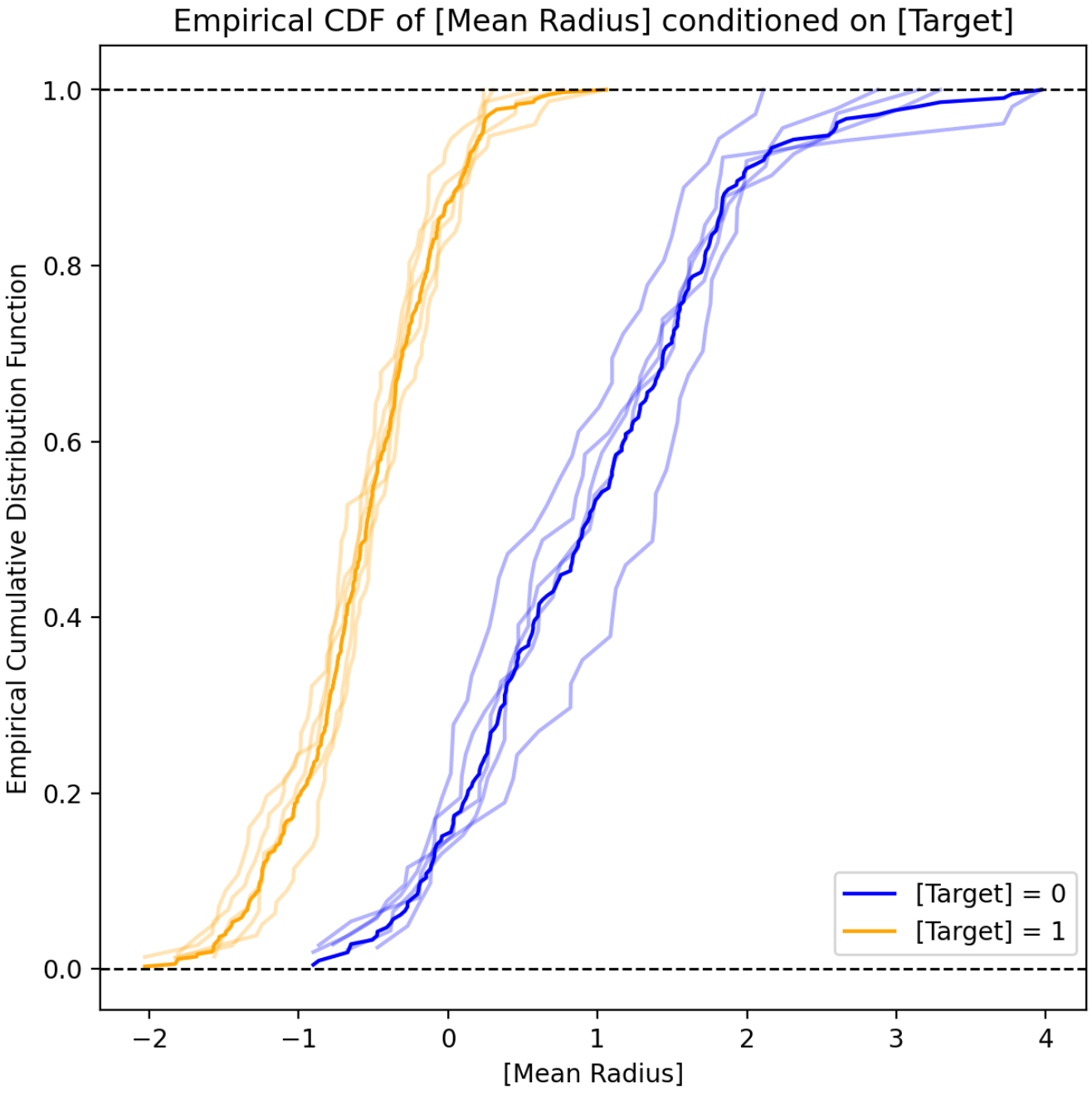
Mean Radius - Kernel Density Plot



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

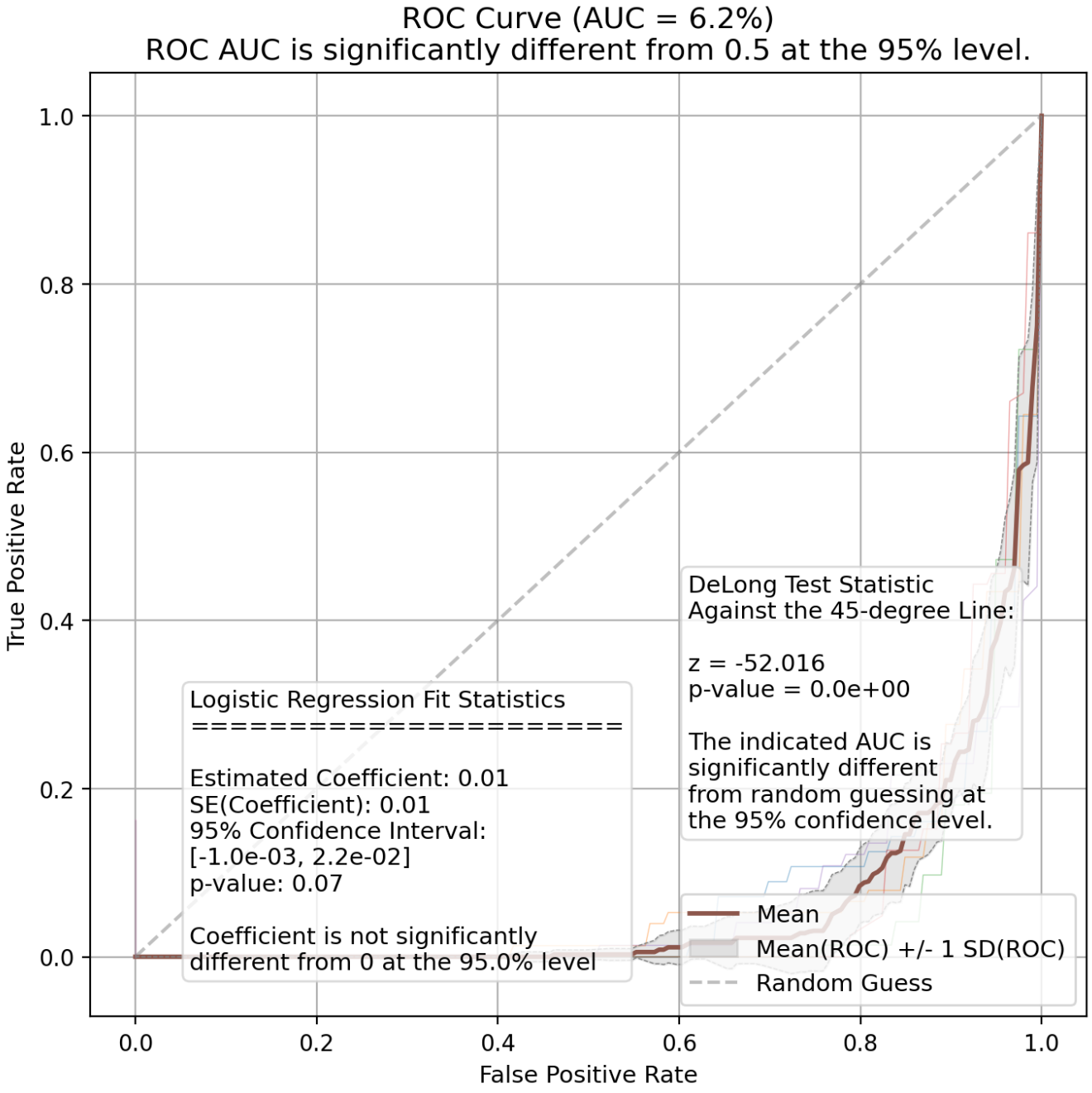
Mean Radius - Empirical CDF Plot



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

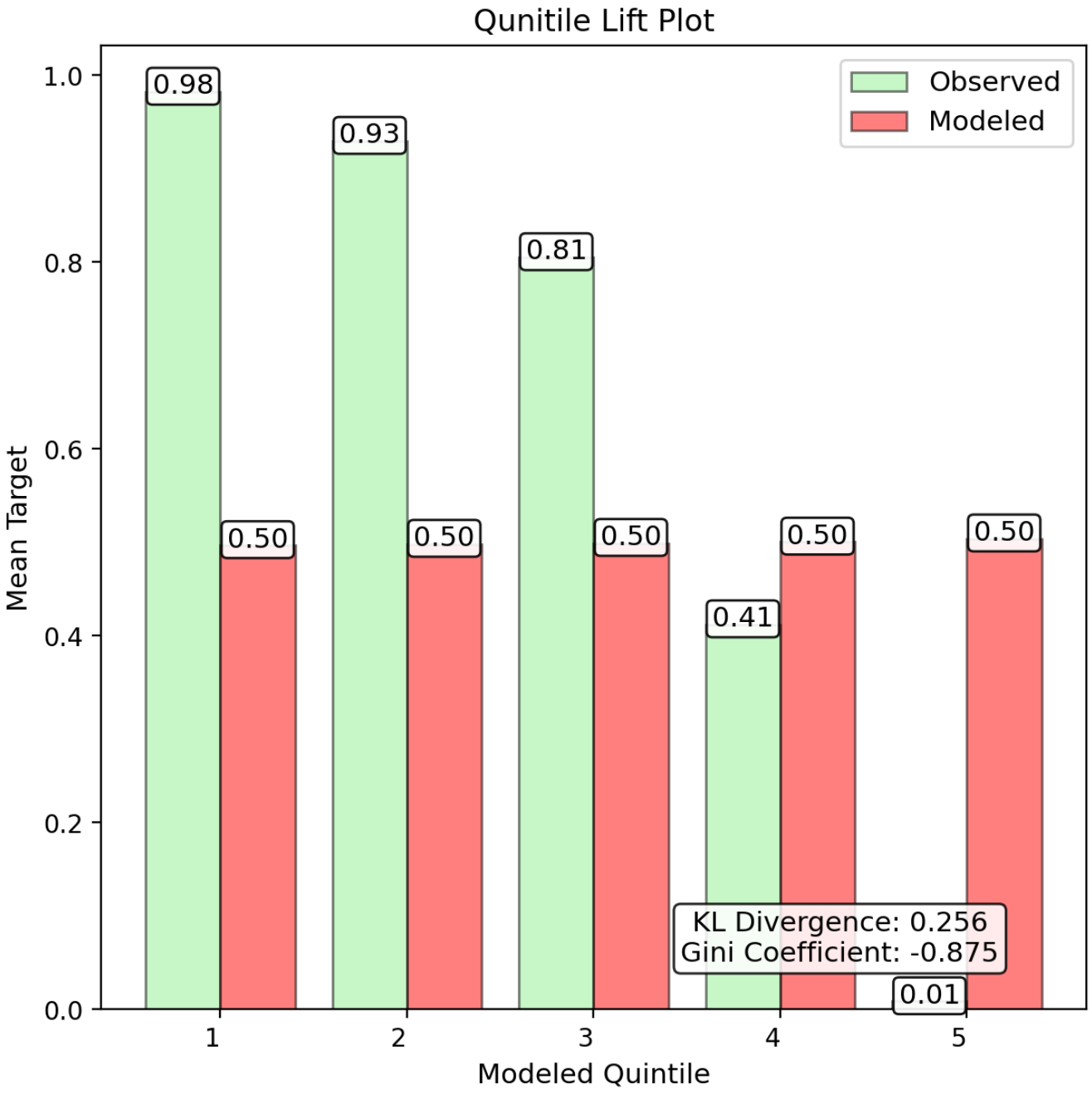


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

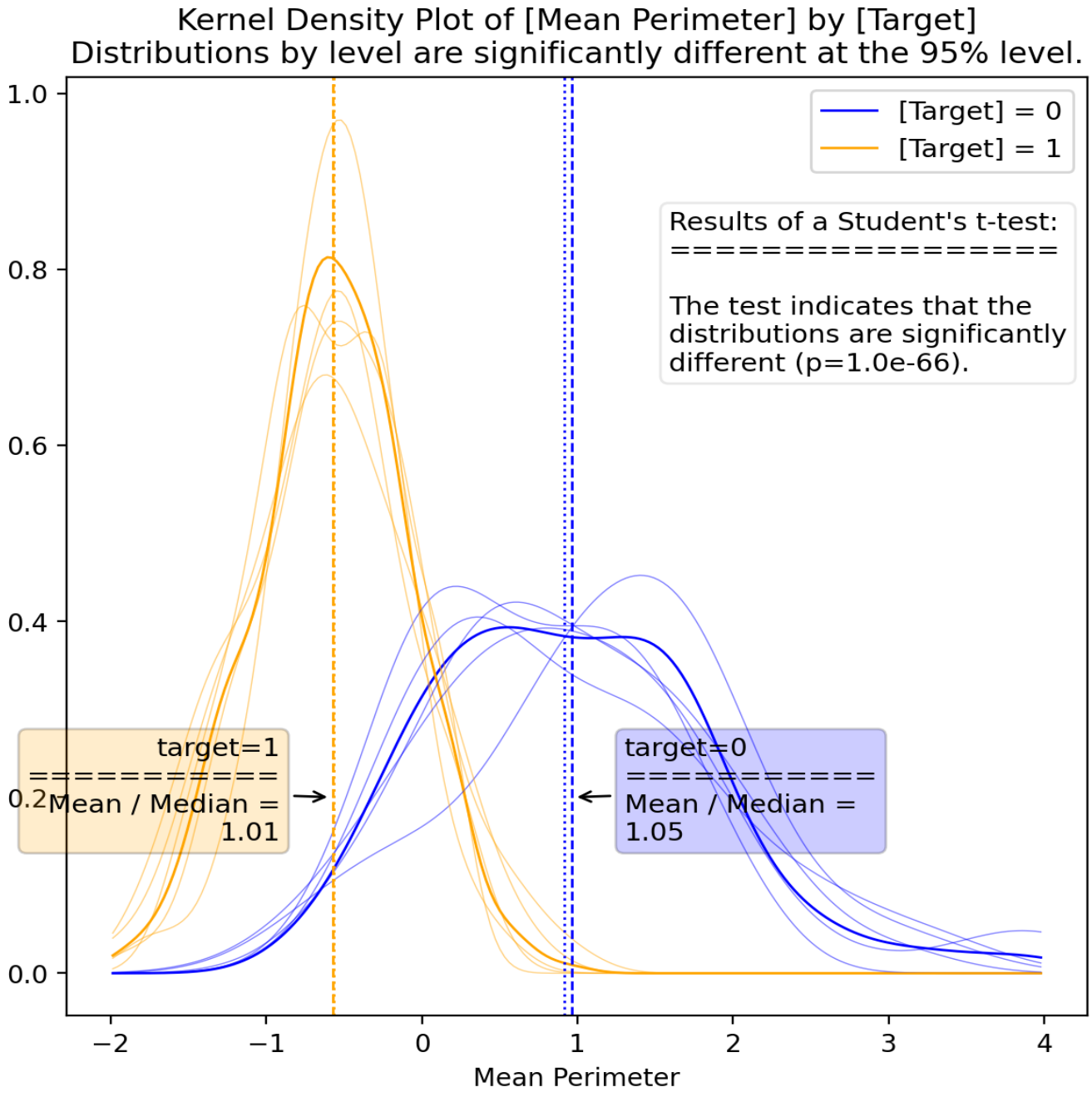
Univariate Report

Mean Perimeter - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Agg. Mean	Agg. SD
Fitted Coef.	-3.7e+00	-3.9e+00	-3.7e+00	-3.6e+00	-3.8e+00	1.3e-03	1.1e-01
Fitted p-Value	2.4e-27	4.4e-25	2.3e-26	4.7e-25	2.7e-26	1.4e-01	2.4e-25
Fitted Std. Err.	0.342	0.381	0.351	0.352	0.355	0.001	0.015
Conf. Int. Lower	-4.4e+00	-4.7e+00	-4.4e+00	-4.3e+00	-4.5e+00	-4.3e-04	1.4e-01
Conf. Int. Upper	-3.0e+00	-3.2e+00	-3.0e+00	-3.0e+00	-3.1e+00	3.0e-03	8.9e-02
Train Accuracy	87.8%	87.2%	86.1%	87.4%	87.4%	13.0%	0.7%
Val Accuracy	85.9%	86.1%	90.8%	87.0%	85.2%	13.0%	2.2%
Train AUC	87.5%	86.9%	86.0%	87.3%	87.0%	13.2%	0.6%
Val AUC	84.9%	86.2%	89.8%	86.9%	85.8%	13.2%	1.8%
Train F1	90.2%	89.7%	88.5%	89.9%	89.8%	15.1%	0.7%
Test F1	88.5%	88.4%	93.1%	89.0%	87.9%	15.1%	2.1%
Train Precision	91.8%	91.2%	90.8%	92.1%	90.9%	19.5%	0.6%
Val Precision	87.7%	91.5%	93.1%	90.8%	92.5%	19.5%	2.1%
Train Recall	88.7%	88.3%	86.3%	87.8%	88.7%	12.3%	1.0%
Val Recall	89.3%	85.5%	93.1%	87.3%	83.8%	12.3%	3.6%
Train MCC	74.3%	73.1%	71.1%	73.5%	73.5%	-72.6%	1.2%
Val MCC	70.2%	71.2%	79.5%	73.2%	69.5%	-72.6%	4.0%
Train Log-Loss	4.38	4.60	5.01	4.53	4.53	31.36	0.24
Val Log-Loss	5.09	5.02	3.31	4.68	5.33	31.36	0.81

Univariate Report

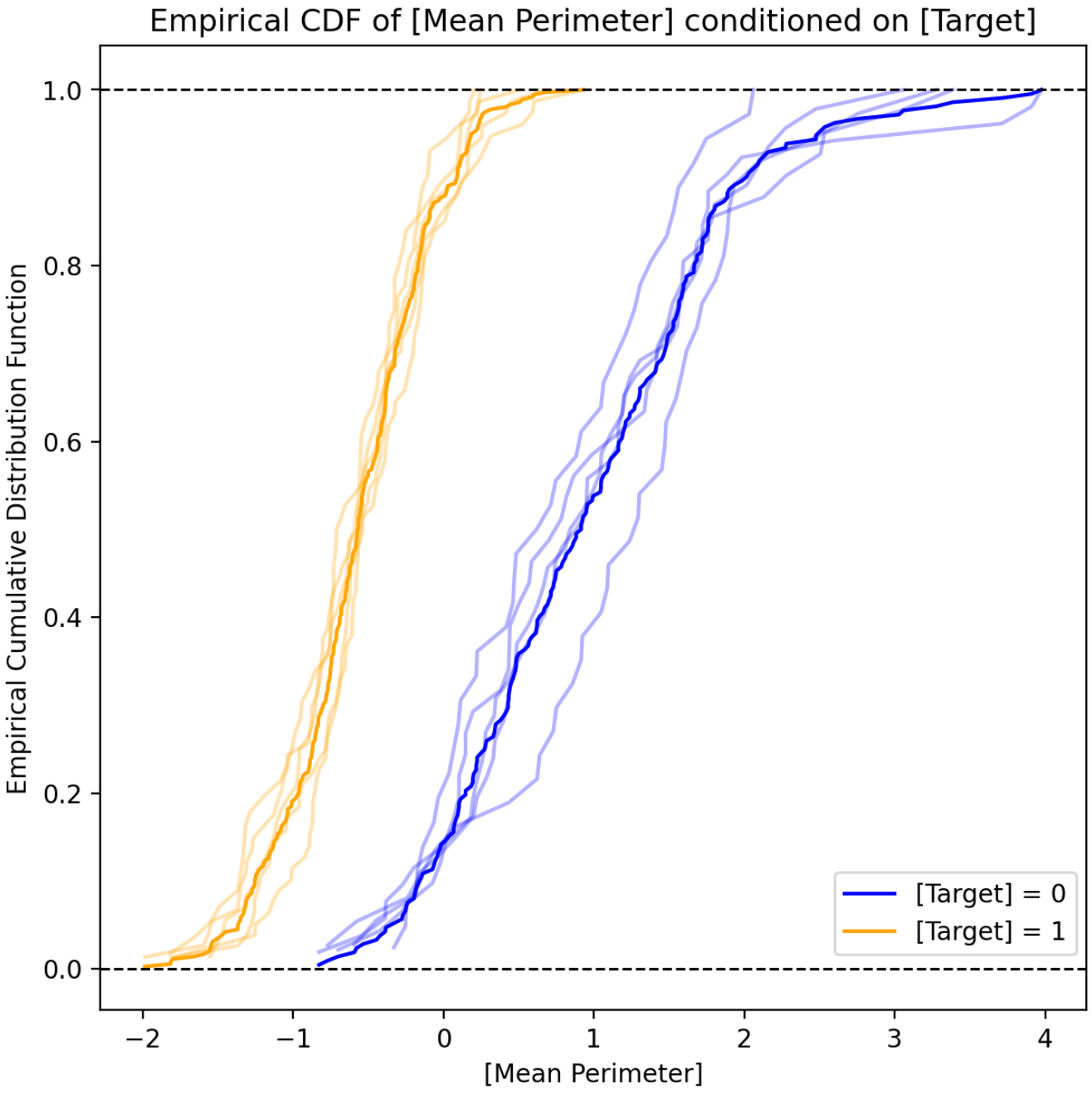
Mean Perimeter - Kernel Density Plot



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

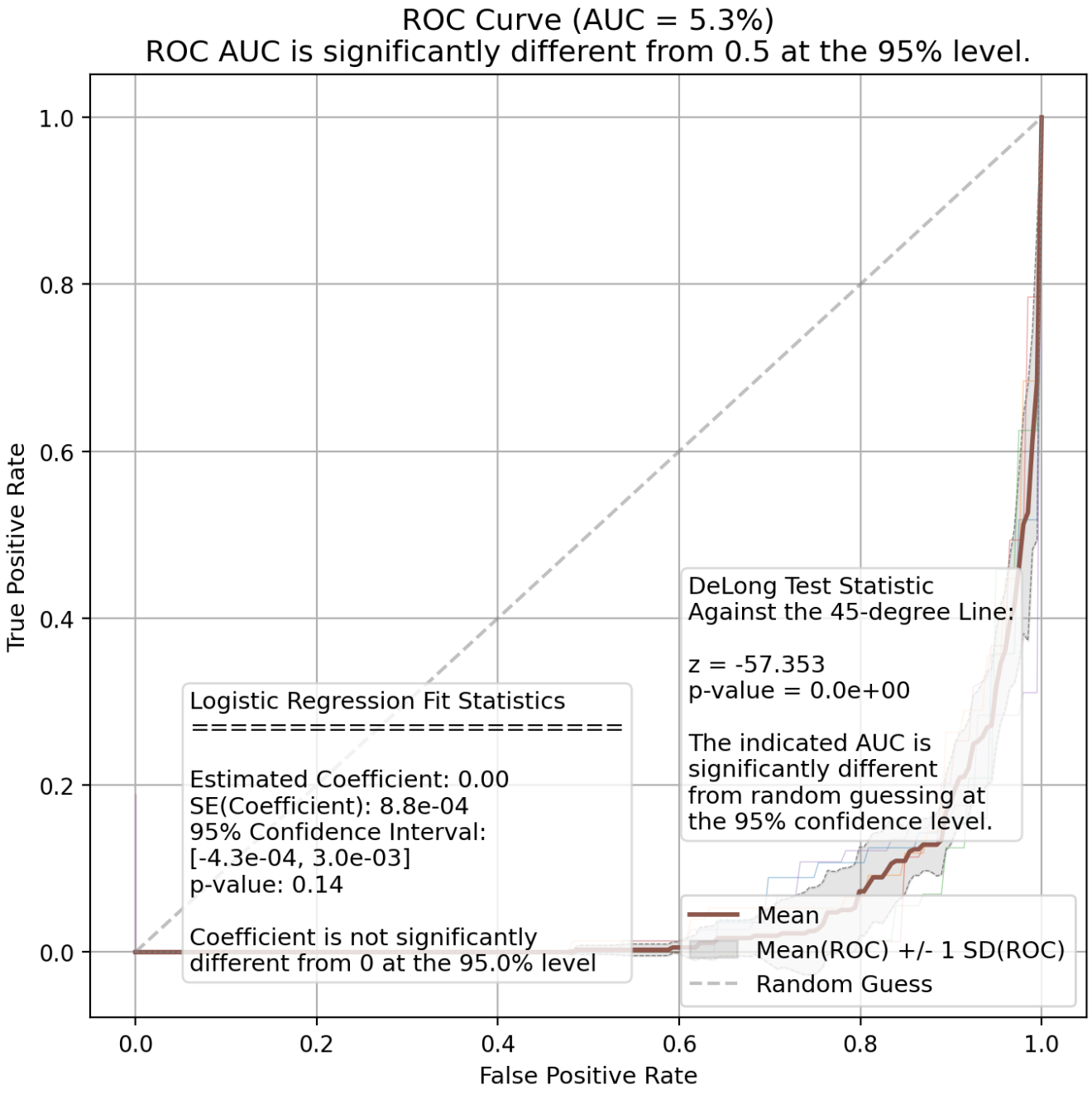
Mean Perimeter - Empirical CDF Plot



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

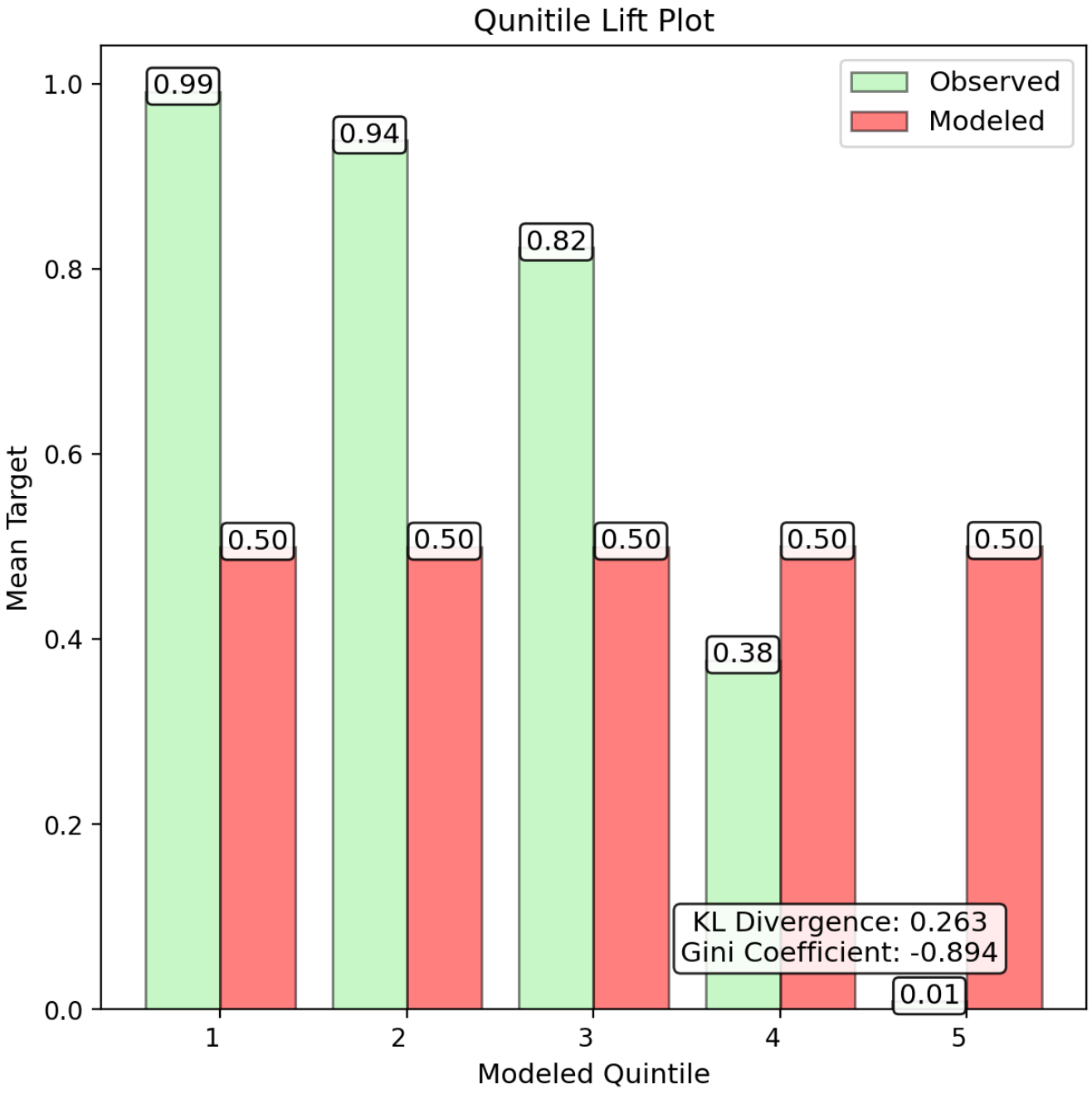


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

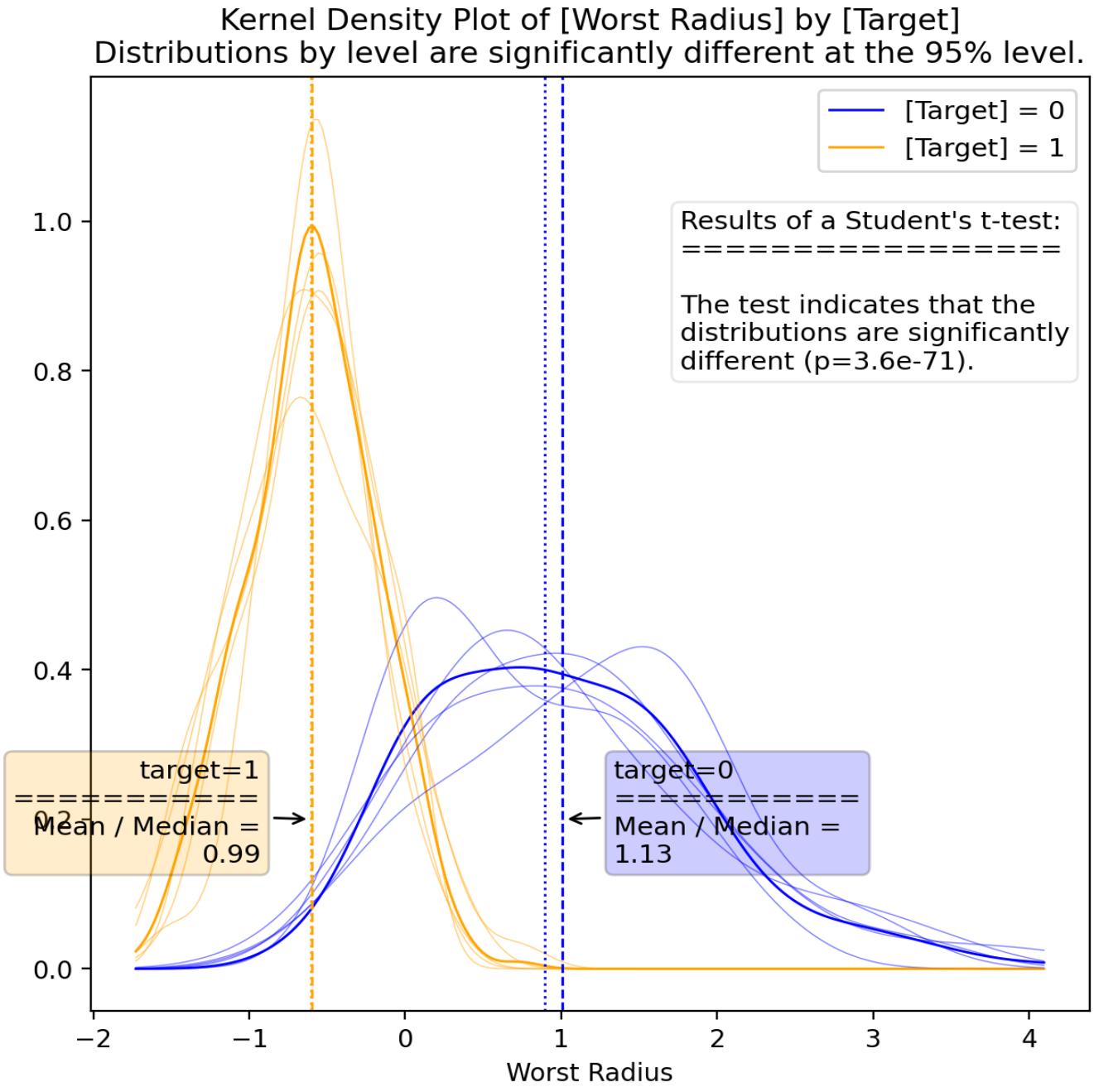
Univariate Report

Worst Radius - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Agg. Mean	Agg. SD
Fitted Coef.	-4.7e+00	-5.1e+00	-4.9e+00	-4.8e+00	-4.9e+00	3.5e-03	1.3e-01
Fitted p-Value	1.7e-25	8.5e-23	2.4e-24	4.4e-23	7.8e-24	4.7e-01	3.7e-23
Fitted Std. Err.	0.453	0.516	0.477	0.489	0.489	0.005	0.023
Conf. Int. Lower	-5.6e+00	-6.1e+00	-5.8e+00	-5.8e+00	-5.9e+00	-6.1e-03	1.7e-01
Conf. Int. Upper	-3.8e+00	-4.1e+00	-3.9e+00	-3.9e+00	-4.0e+00	1.3e-02	8.7e-02
Train Accuracy	90.6%	90.2%	90.4%	90.2%	91.2%	9.3%	0.4%
Val Accuracy	92.4%	90.2%	91.7%	90.1%	88.7%	9.3%	1.5%
Train AUC	89.9%	89.7%	90.0%	90.0%	90.5%	9.8%	0.3%
Val AUC	90.8%	90.8%	91.1%	89.8%	88.0%	9.8%	1.3%
Train F1	92.5%	92.1%	92.3%	92.1%	93.0%	9.8%	0.4%
Test F1	94.0%	91.8%	93.7%	91.7%	91.2%	9.8%	1.3%
Train Precision	92.7%	92.8%	92.6%	93.7%	92.6%	13.0%	0.5%
Val Precision	90.2%	95.7%	94.4%	92.3%	91.8%	13.0%	2.2%
Train Recall	92.4%	91.5%	91.9%	90.6%	93.3%	7.8%	1.0%
Val Recall	98.2%	88.2%	93.1%	91.1%	90.5%	7.8%	3.8%
Train MCC	79.8%	79.0%	79.8%	79.2%	81.2%	-80.1%	0.9%
Val MCC	84.2%	80.0%	81.7%	79.4%	75.5%	-80.1%	3.2%
Train Log-Loss	3.40	3.55	3.45	3.54	3.18	32.69	0.15
Val Log-Loss	2.74	3.55	2.98	3.58	4.07	32.69	0.53

Univariate Report

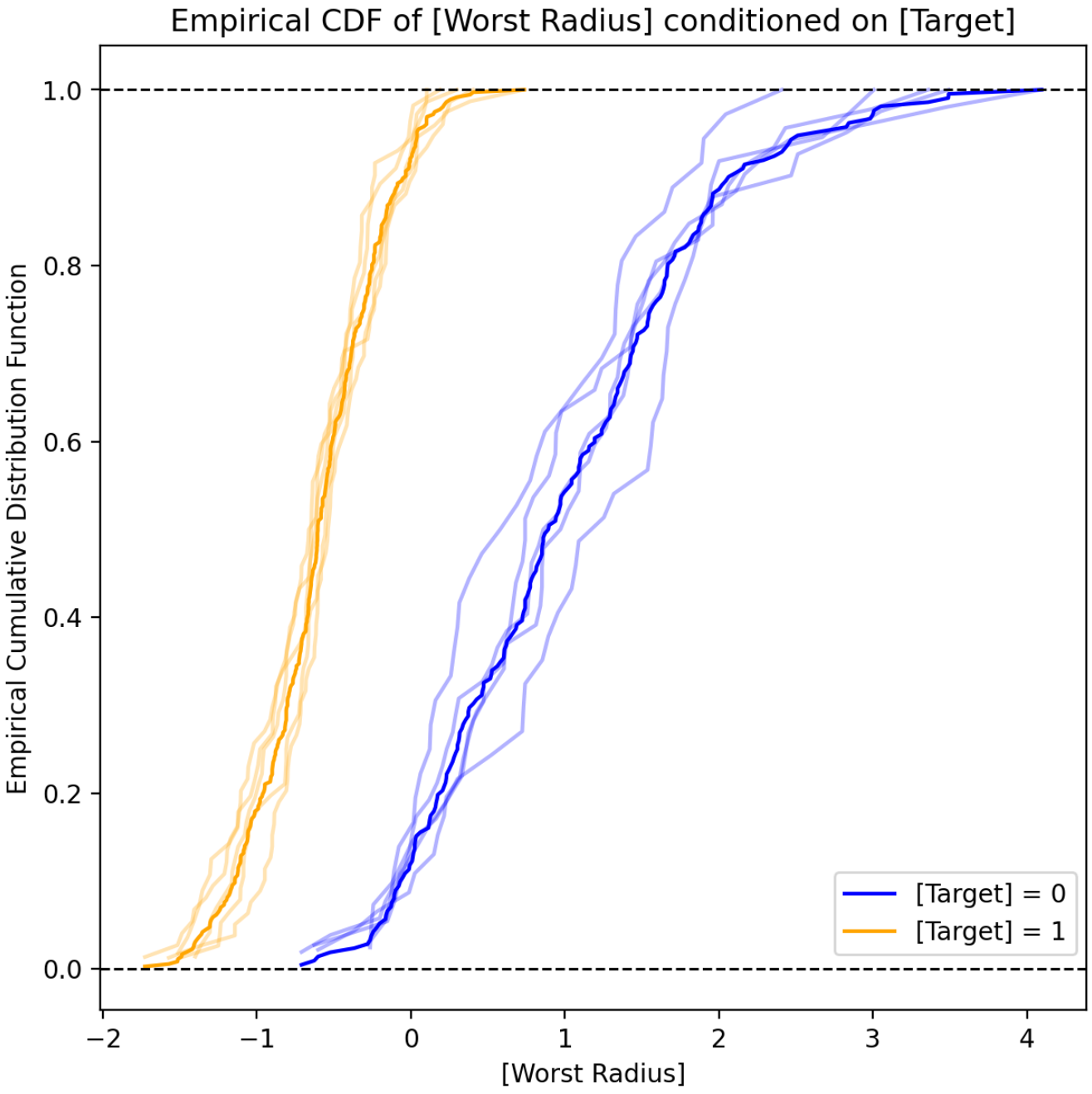
Worst Radius - Kernel Density Plot



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

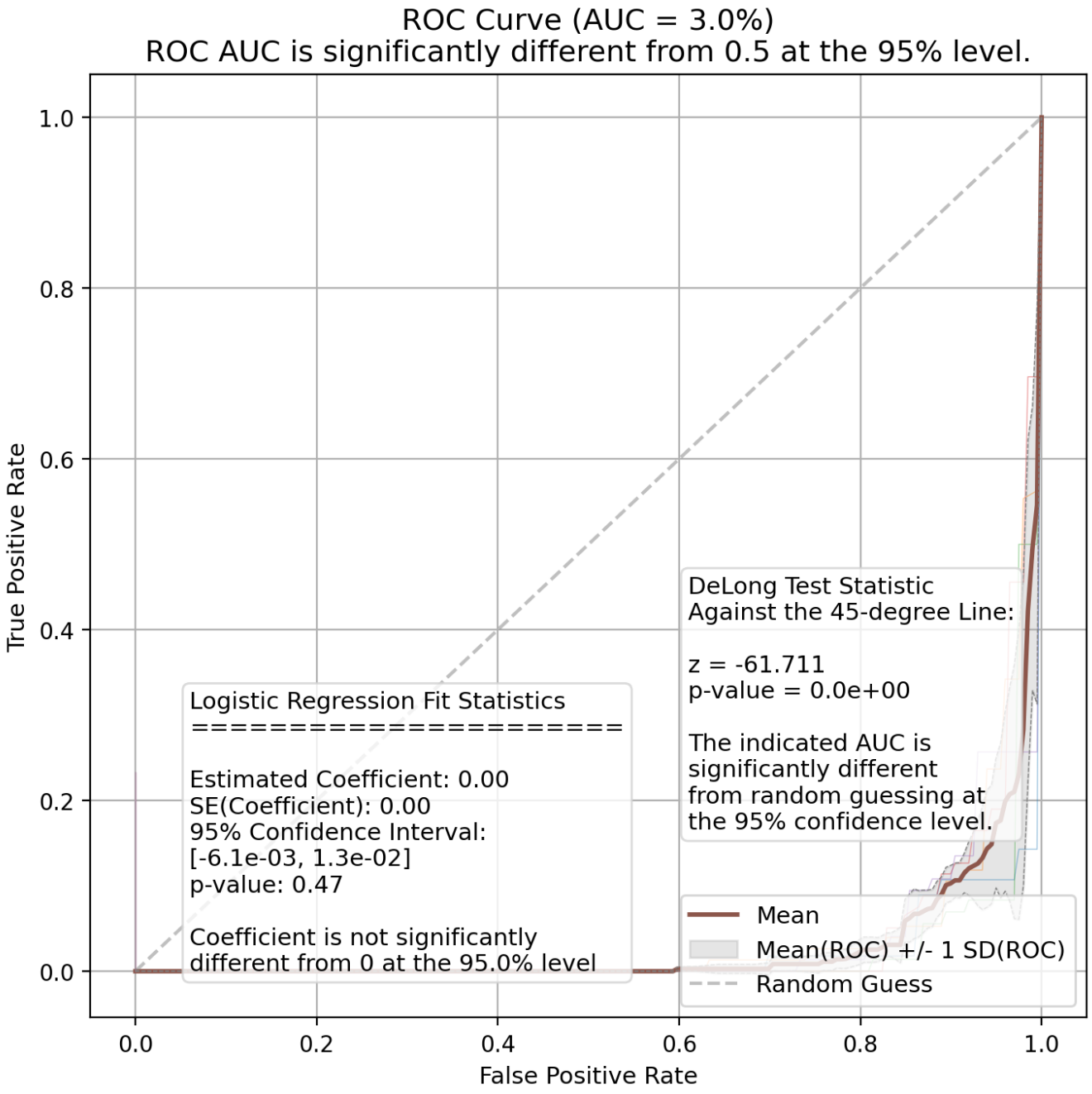
Worst Radius - Empirical CDF Plot



This plot shows the empirical cumulative distribution function for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the cumulative distribution of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data, and whether or not it is reasonable to assume that the data is drawn from different distributions.

Univariate Report

Worst Radius - ROC Curve

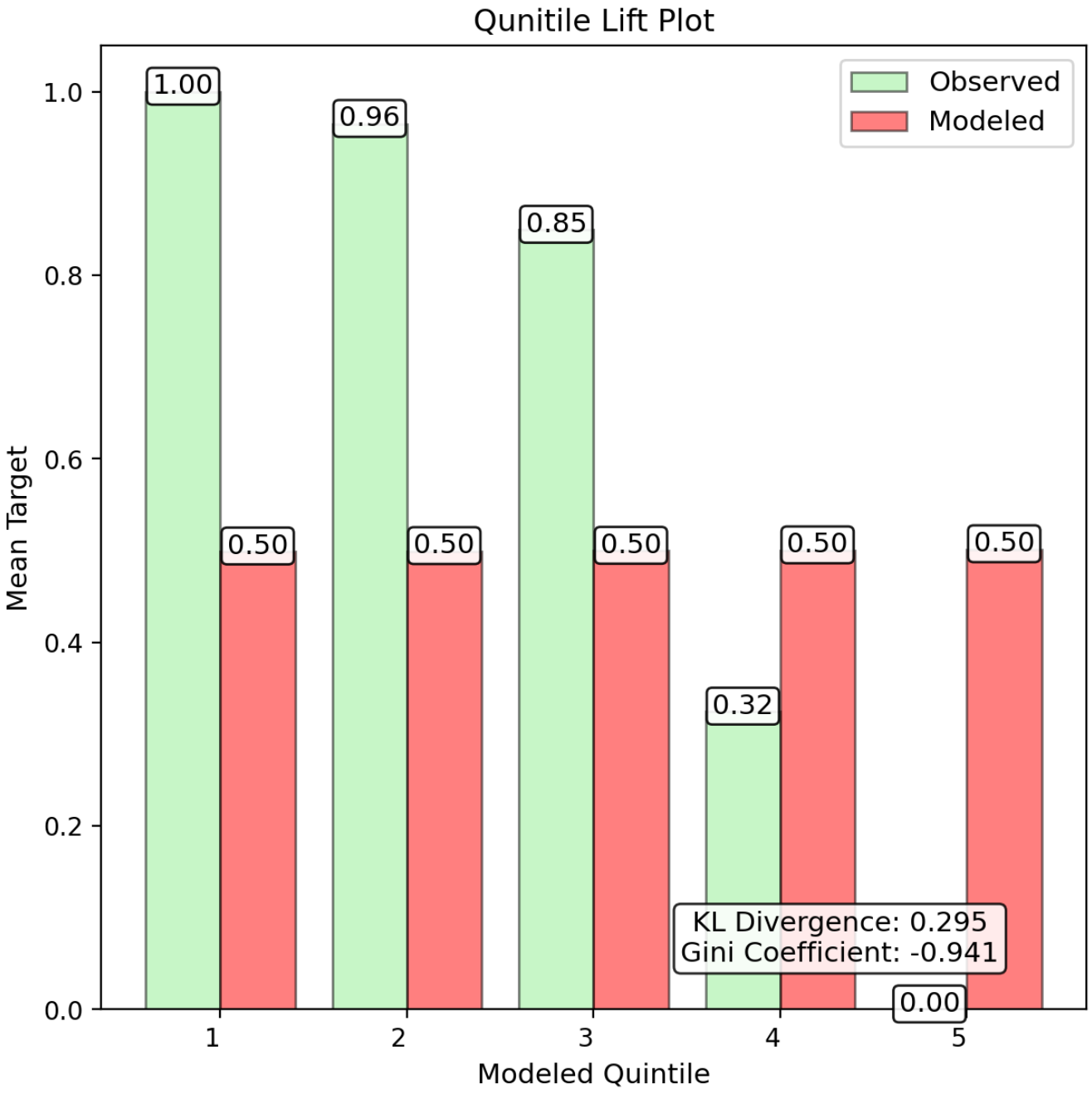


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, and no other features. Annotations are on the plot to help understand the results of the model, including the coefficient, standard error, and p-value for the feature variable. The cross-validation folds are used to create the grey region around the mean ROC curve to help understand the variability of the data. Significance of the ROC curve is determined based on 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

Worst Radius - Quintile Lift



The quintile lift plot is meant to show the power of the single feature to discriminate between the highest and lowest quintiles of the target variable.

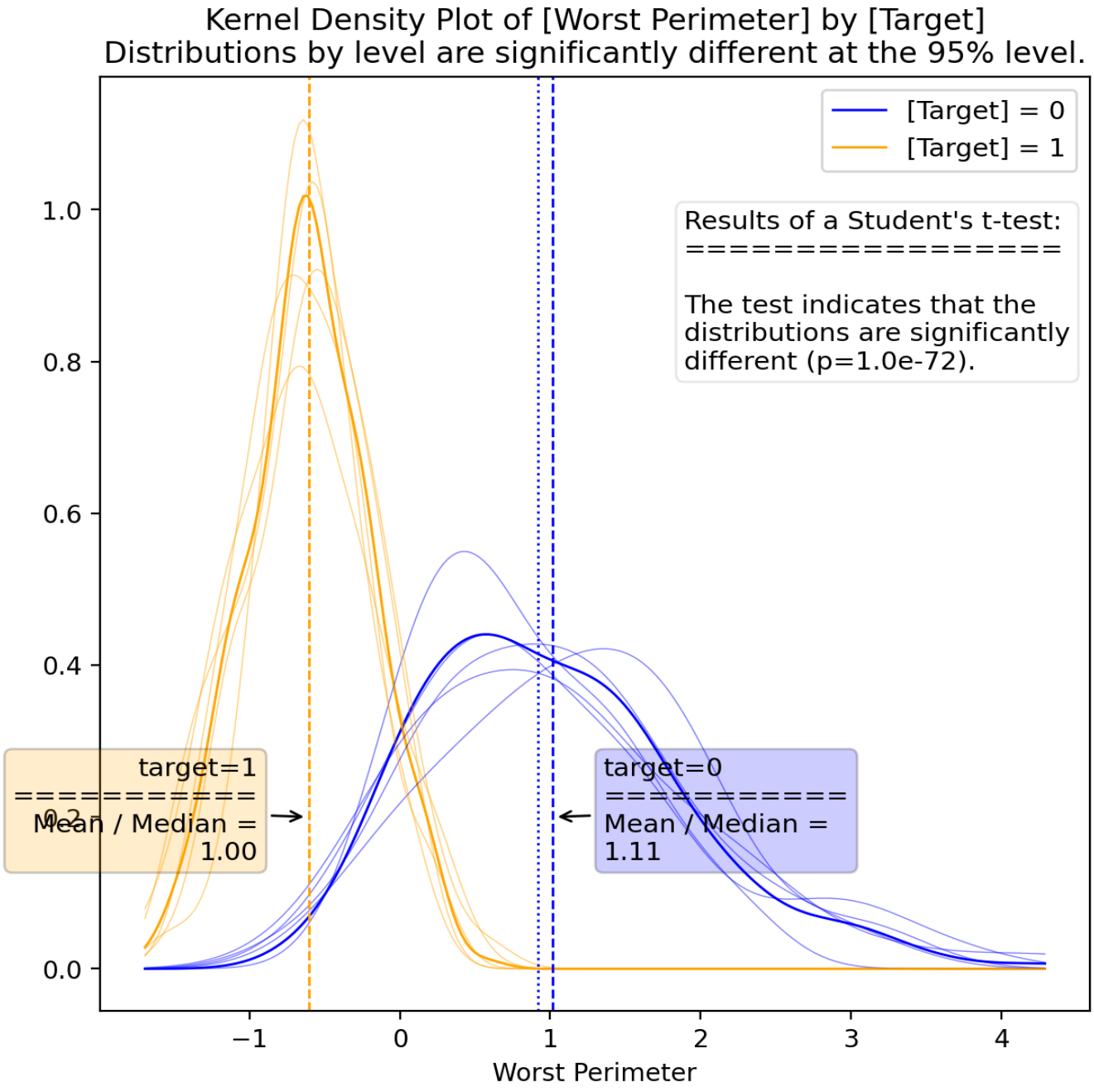
Univariate Report

Worst Perimeter - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Agg. Mean	Agg. SD
Fitted Coef.	-4.9e+00	-5.2e+00	-5.1e+00	-5.0e+00	-5.2e+00	3.0e-04	1.4e-01
Fitted p-Value	1.9e-25	3.3e-23	2.2e-24	2.9e-23	9.9e-24	6.9e-01	1.5e-23
Fitted Std. Err.	0.467	0.526	0.497	0.506	0.515	0.001	0.022
Conf. Int. Lower	-5.8e+00	-6.2e+00	-6.0e+00	-6.0e+00	-6.2e+00	-1.2e-03	1.8e-01
Conf. Int. Upper	-4.0e+00	-4.2e+00	-4.1e+00	-4.0e+00	-4.2e+00	1.8e-03	9.6e-02
Train Accuracy	91.2%	91.3%	91.7%	91.8%	91.9%	8.3%	0.3%
Val Accuracy	93.5%	91.8%	91.7%	90.1%	90.4%	8.3%	1.3%
Train AUC	90.8%	90.8%	91.3%	91.7%	91.3%	8.6%	0.4%
Val AUC	92.7%	92.1%	91.1%	89.5%	90.4%	8.6%	1.3%
Train F1	93.0%	93.0%	93.3%	93.4%	93.5%	9.1%	0.2%
Test F1	94.7%	93.2%	93.7%	91.8%	92.4%	9.1%	1.1%
Train Precision	93.6%	93.5%	93.6%	94.8%	93.3%	12.0%	0.6%
Val Precision	93.1%	95.8%	94.4%	91.2%	94.4%	12.0%	1.7%
Train Recall	92.4%	92.5%	93.0%	92.1%	93.6%	7.3%	0.6%
Val Recall	96.4%	90.8%	93.1%	92.4%	90.5%	7.3%	2.4%
Train MCC	81.2%	81.4%	82.5%	82.5%	82.6%	-82.4%	0.7%
Val MCC	86.3%	83.0%	81.7%	79.2%	79.6%	-82.4%	2.9%
Train Log-Loss	3.17	3.14	2.98	2.96	2.94	33.07	0.11
Val Log-Loss	2.35	2.95	2.98	3.58	3.45	33.07	0.48

Univariate Report

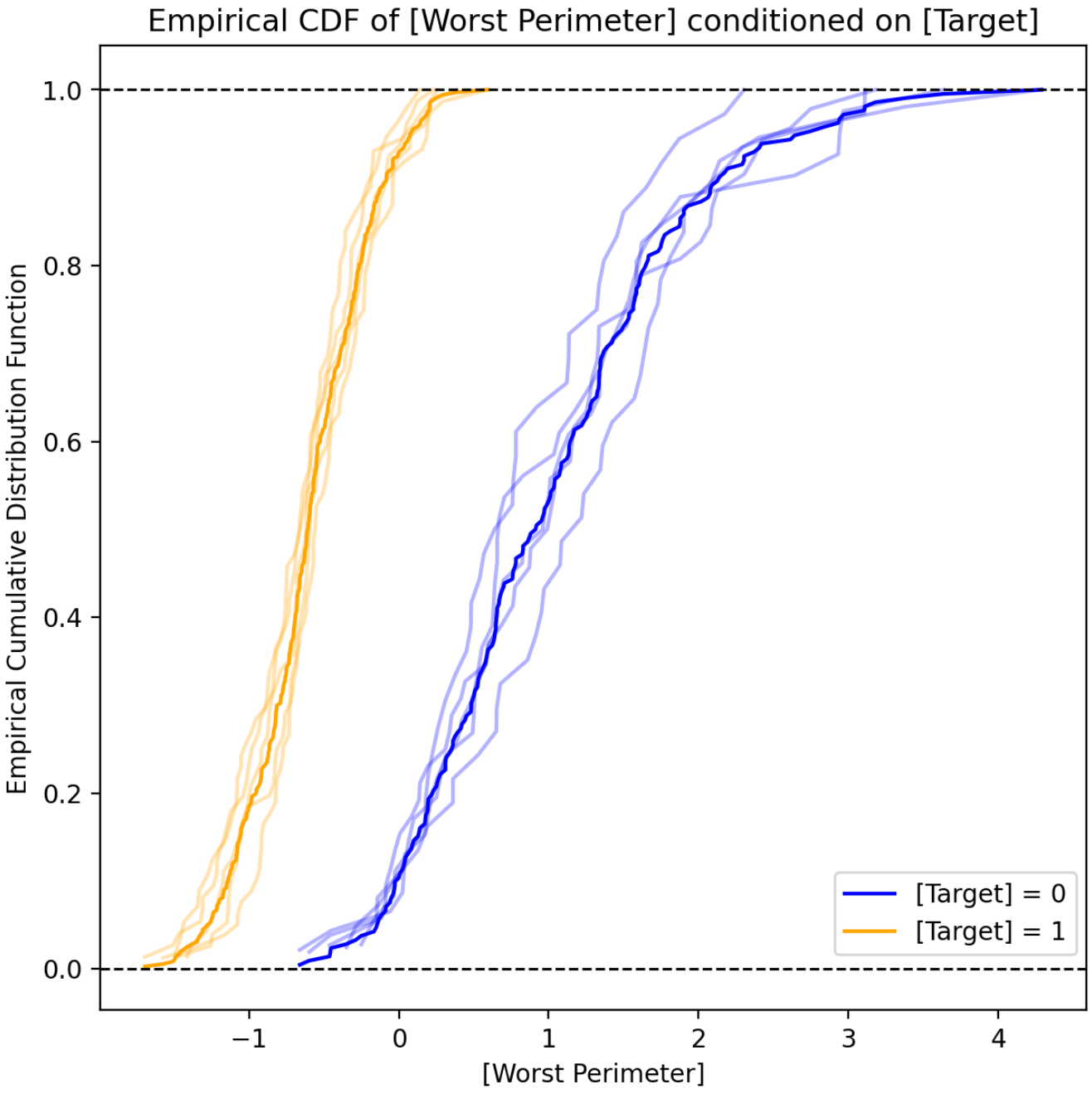
Worst Perimeter - Kernel Density Plot



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

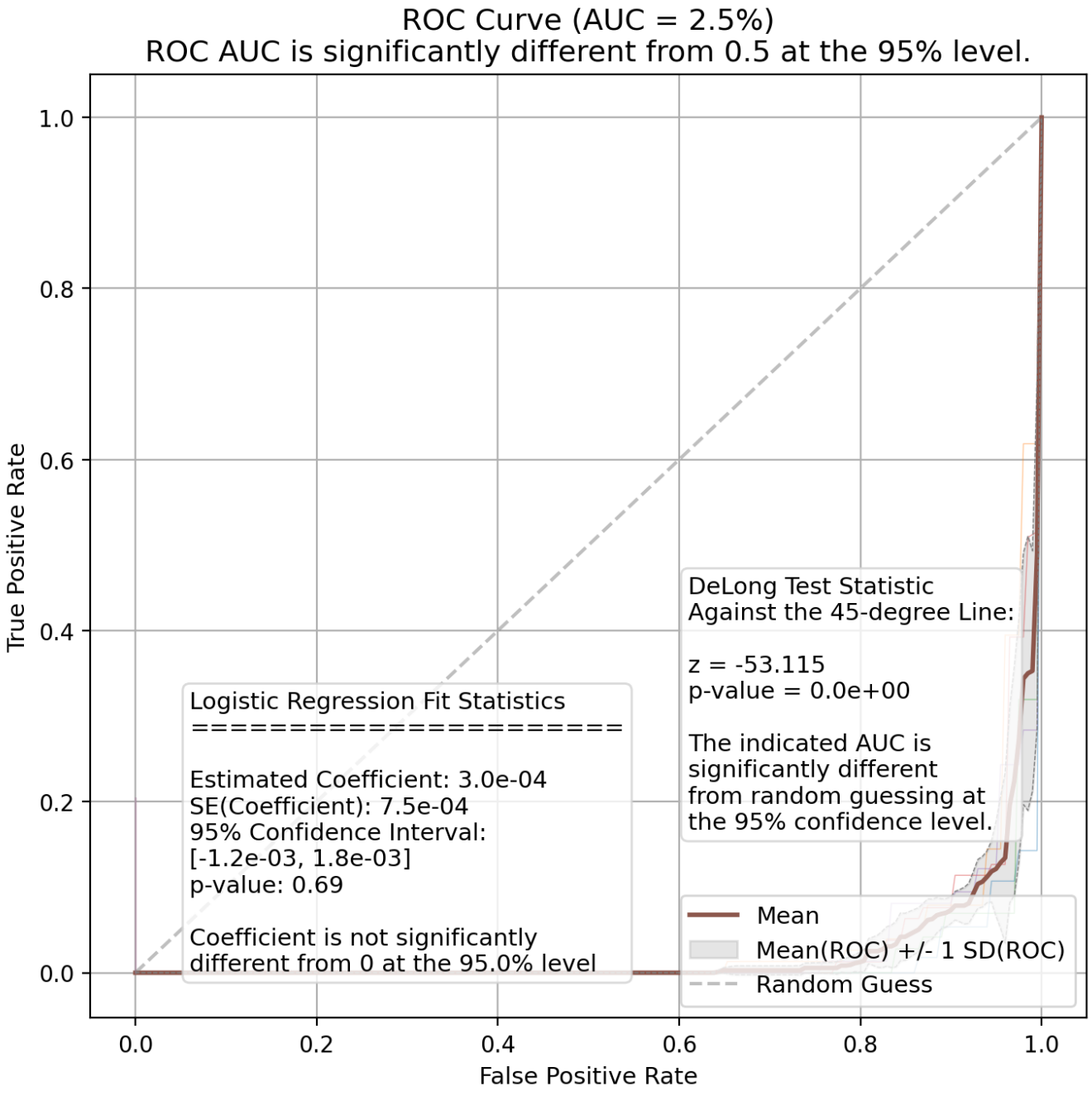
Worst Perimeter - Empirical CDF Plot



This plot shows the empirical cumulative distribution function for each level of the target variable, both in total and for each fold. The x-axis represents the feature variable, and the y-axis represents the cumulative distribution of the target variable. The cross-validation folds are included in slightly washed-out colors to help understand the variability of the data, and whether or not it is reasonable to assume that the data is drawn from different distributions.

Univariate Report

Worst Perimeter - ROC Curve

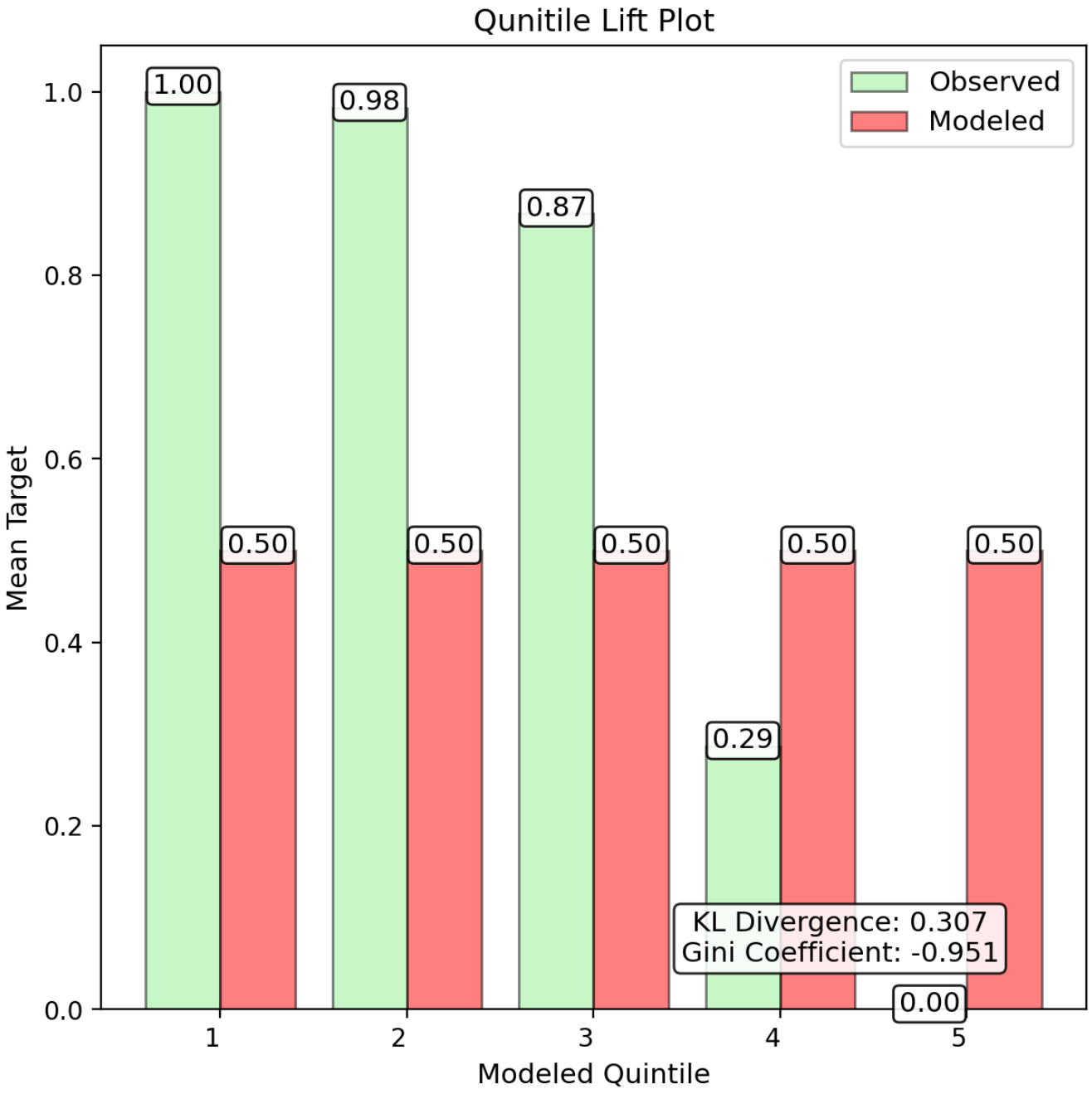


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept, and no other features. Annotations are on the plot to help understand the results of the model, including the coefficient, standard error, and p-value for the feature variable. The cross-validation folds are used to create the grey region around the mean ROC curve to help understand the variability of the data. Significance of the ROC curve is determined based on 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

Worst Perimeter - Quintile Lift



The quintile lift plot is meant to show the power of the single feature to discriminate between the highest and lowest quintiles of the target variable.