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_area	37.7%	0.00e+00	0.00e+00	50.0%	0.00e+00	0.00e+00	14.6%
mean_compactness	37.7%	0.00e+00	0.00e+00	50.0%	0.00e+00	0.00e+00	14.6%
mean_concavity	37.7%	0.00e+00	0.00e+00	50.0%	0.00e+00	0.00e+00	14.6%
mean_concave_points	37.7%	0.00e+00	0.00e+00	50.0%	0.00e+00	0.00e+00	14.6%

This table shows an overview of the results for the variables in this file, representing those whose average test score are ranked between 7 and 10 of the variables passed to the Cancer Model.

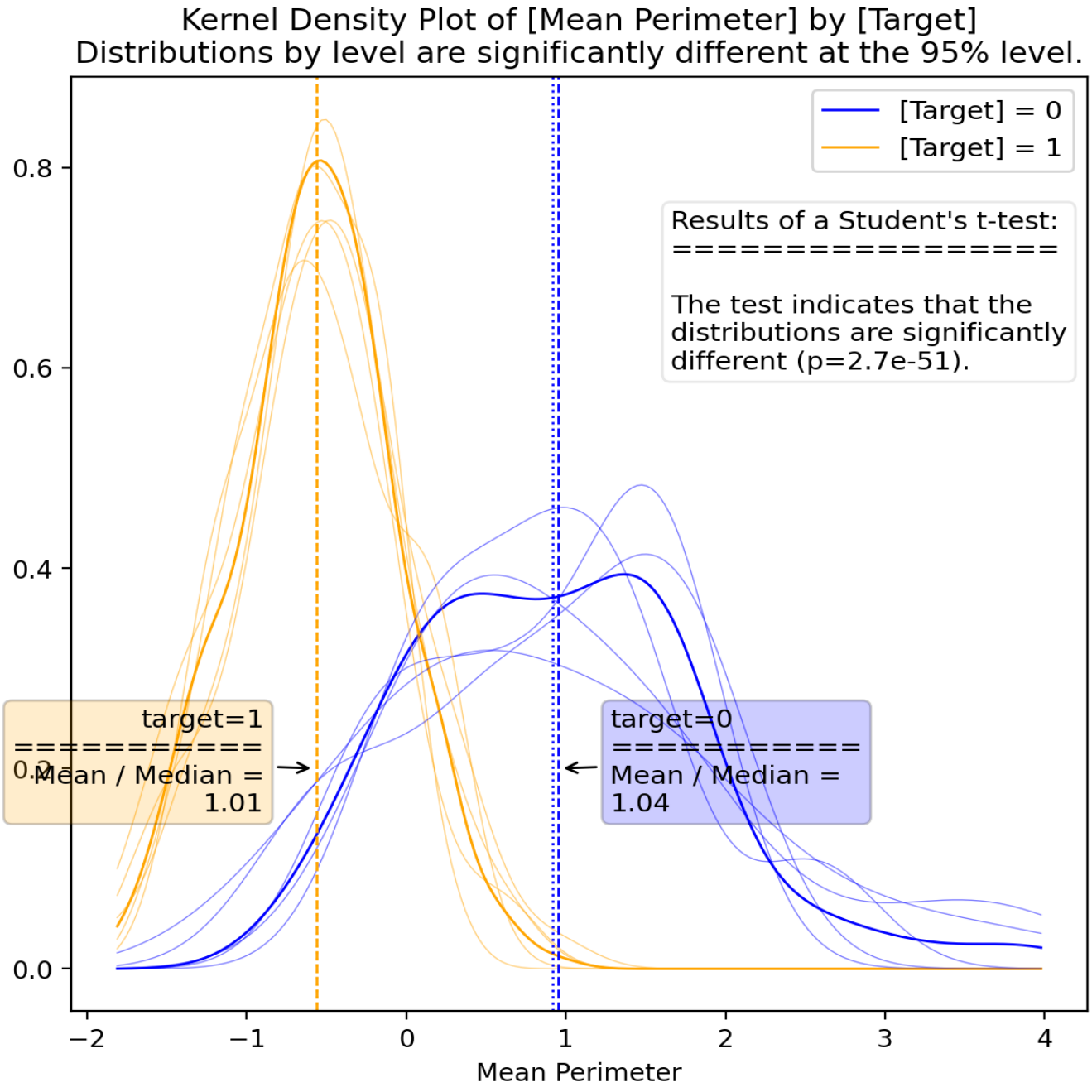
Univariate Report

Mean Perimeter - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Agg. Mean	Agg. SD
Fitted Coef.	-3.8e+00	-3.2e+00	-3.6e+00	-3.4e+00	-3.7e+00	-3.5e+00	2.5e-01
Fitted p-Value	4.8e-21	1.0e-21	1.7e-21	1.3e-21	4.3e-21	2.0e-26	1.8e-21
Fitted Std. Err.	0.405	0.334	0.374	0.354	0.394	0.331	0.029
Conf. Int. Lower	-4.6e+00	-3.9e+00	-4.3e+00	-4.1e+00	-4.5e+00	-4.2e+00	3.1e-01
Conf. Int. Upper	-3.0e+00	-2.5e+00	-2.8e+00	-2.7e+00	-2.9e+00	-2.9e+00	1.9e-01
Train Accuracy	87.7%	84.7%	88.4%	85.7%	87.5%	86.8%	1.6%
Val Accuracy	83.3%	94.8%	80.4%	91.2%	83.7%	37.7%	6.0%
Train AUC	87.5%	84.6%	88.1%	85.3%	87.5%	86.6%	1.6%
Val AUC	82.6%	94.3%	80.6%	91.3%	82.6%	50.0%	6.1%
Train F1	89.7%	87.5%	90.8%	88.6%	89.7%	89.3%	1.2%
Test F1	87.8%	95.8%	82.7%	92.2%	87.5%	0.0%	5.0%
Train Precision	91.2%	90.2%	92.4%	90.6%	91.8%	91.2%	0.9%
Val Precision	91.5%	95.0%	86.0%	94.0%	89.1%	0.0%	3.7%
Train Recall	88.3%	85.0%	89.2%	86.8%	87.8%	87.4%	1.6%
Val Recall	84.4%	96.6%	79.6%	90.4%	86.0%	0.0%	6.4%
Train MCC	74.4%	67.9%	75.3%	69.5%	74.0%	72.3%	3.3%
Val MCC	62.1%	89.0%	60.5%	82.2%	64.3%	0.0%	13.1%
Train Log-Loss	4.44	5.52	4.17	5.15	4.49	4.75	0.56
Val Log-Loss	6.01	1.88	7.05	3.17	5.87	22.45	2.17

Univariate Report

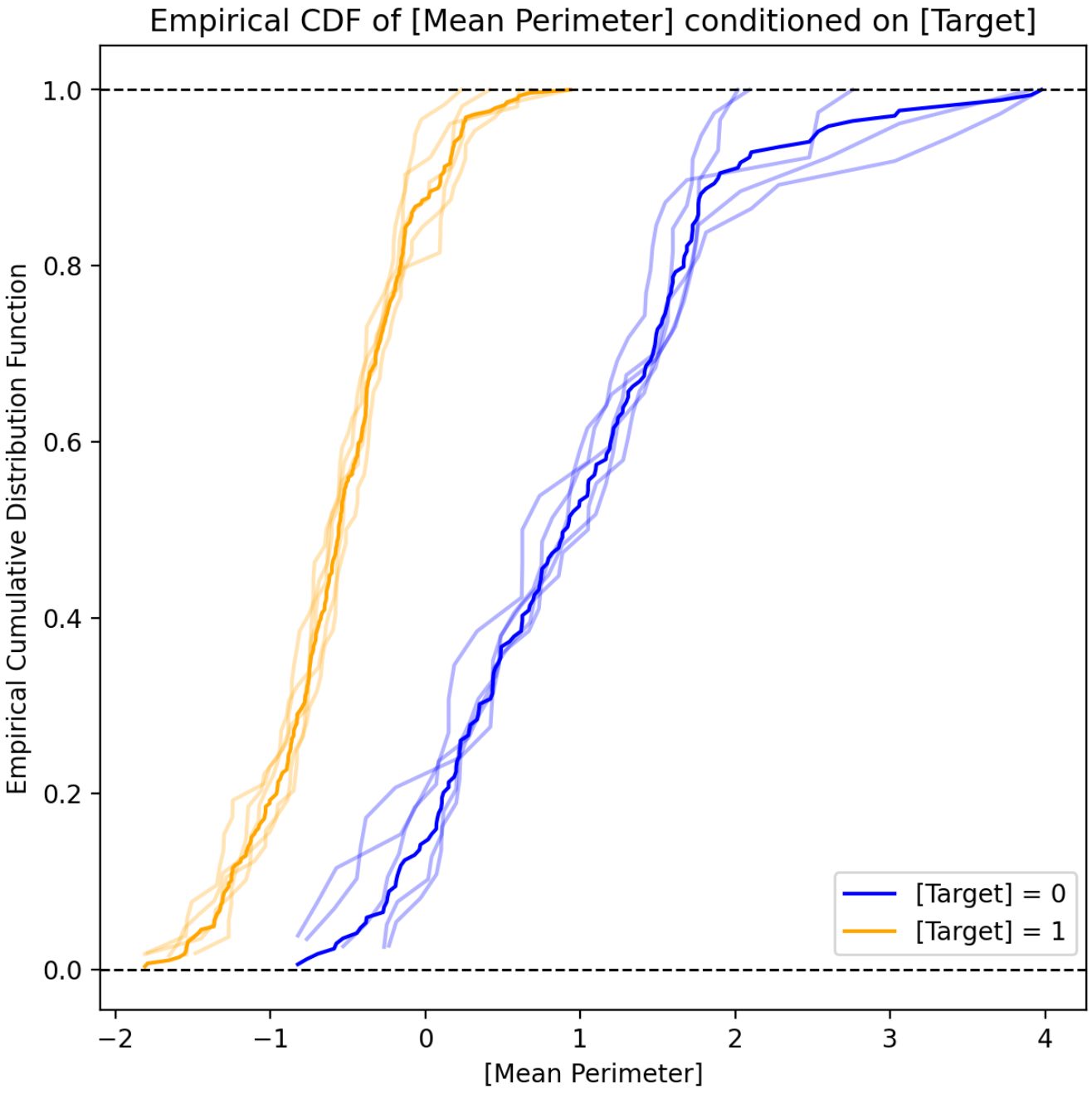
Mean Perimeter - Kernel Density Plot



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

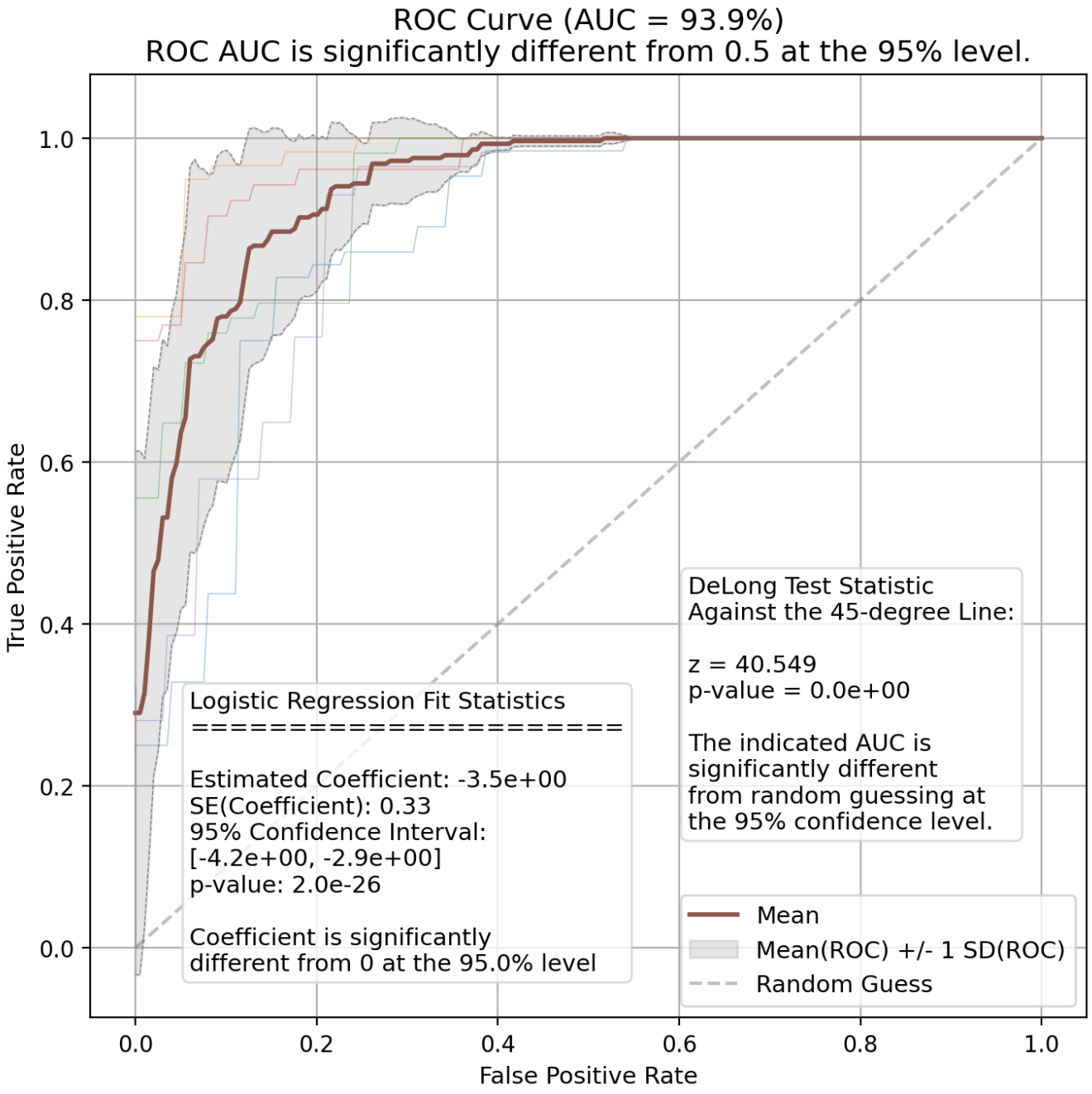
Mean Perimeter - Empirical CDF Plot



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

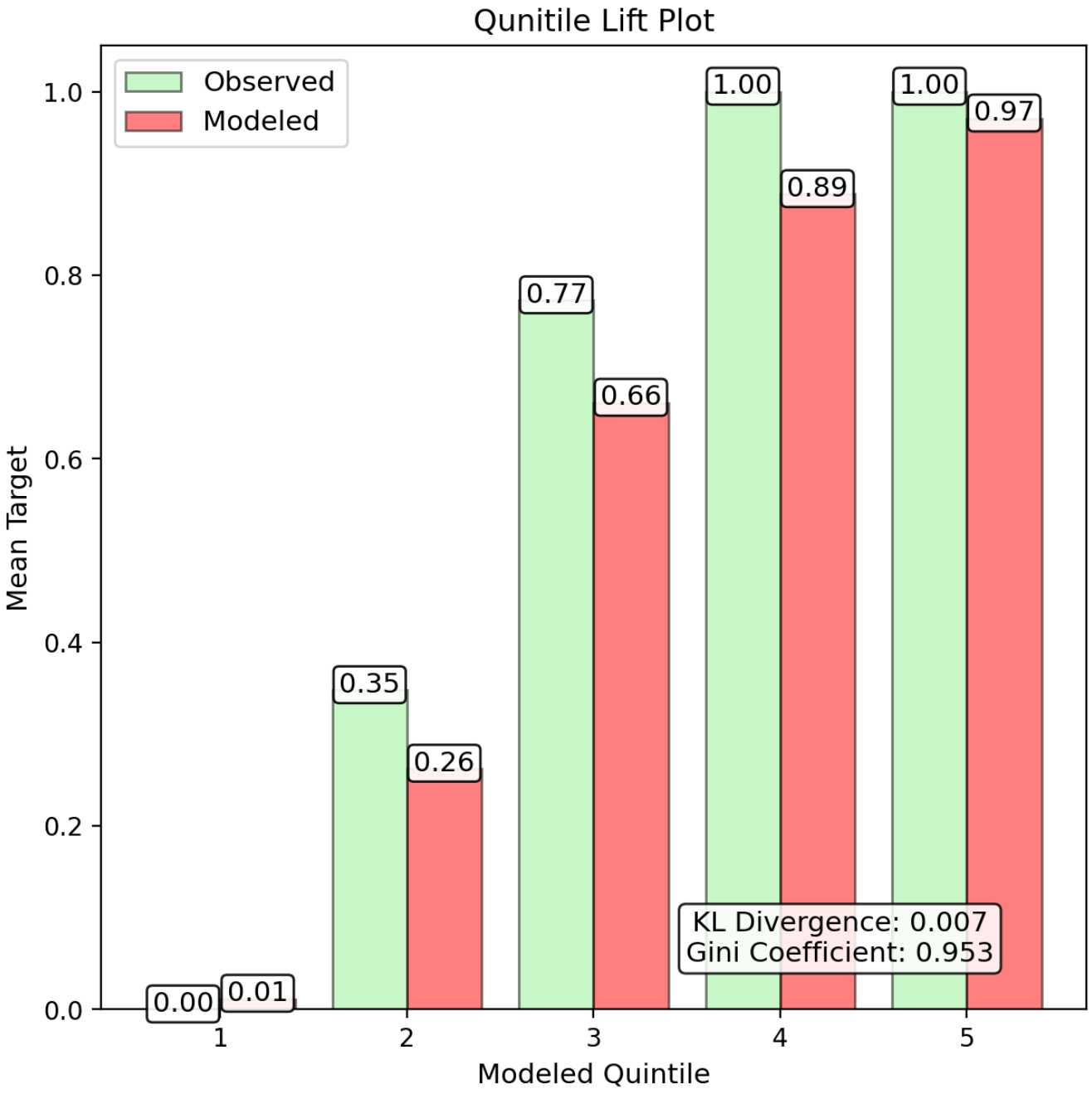


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

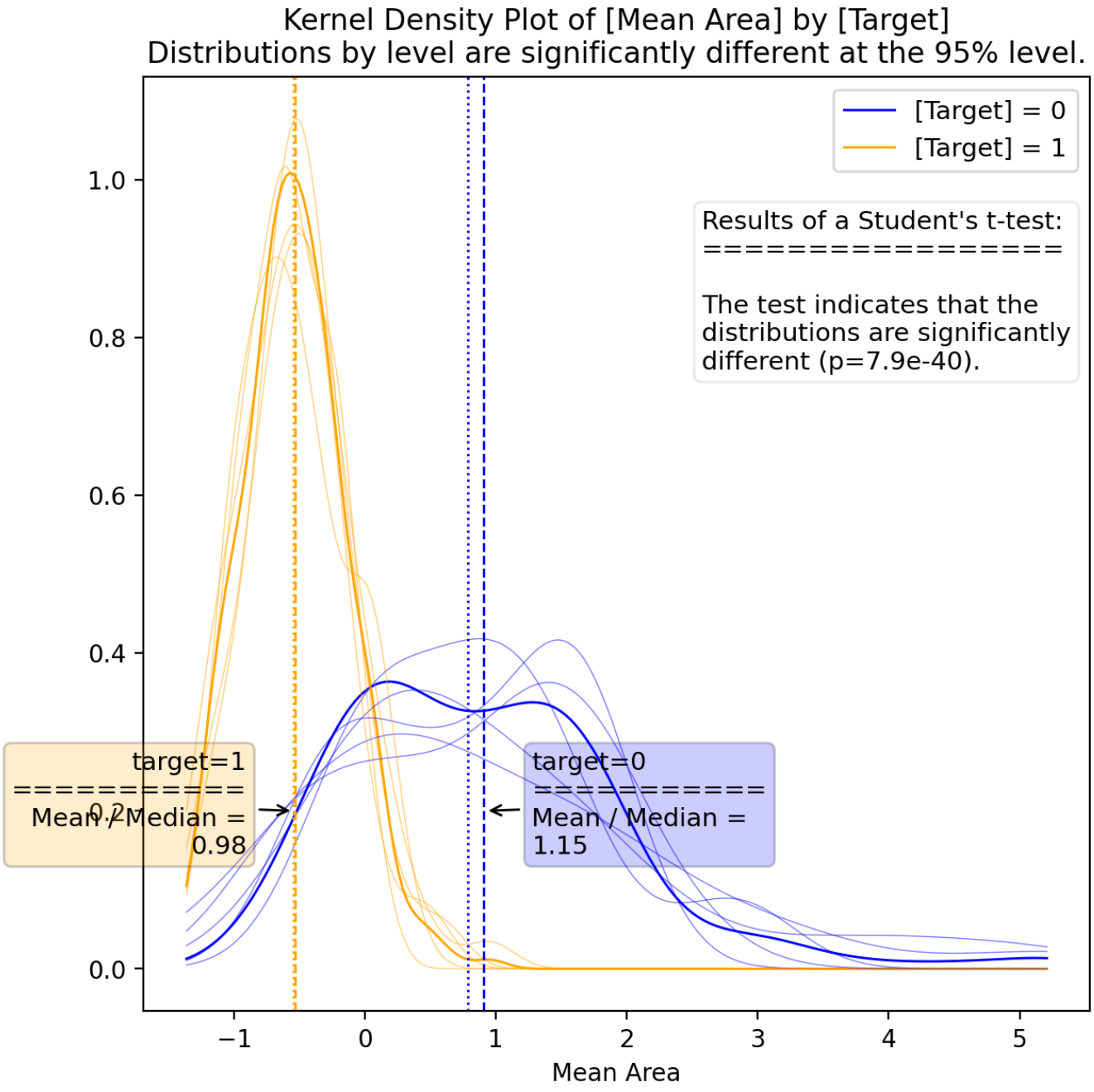
Univariate Report

Mean Area - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Agg. Mean	Agg. SD
Fitted Coef.	-3.9e+00	-3.4e+00	-3.8e+00	-3.6e+00	-3.9e+00	-3.7e+00	2.2e-01
Fitted p-Value	7.8e-22	3.0e-22	4.9e-22	4.5e-22	8.9e-22	3.8e-27	2.4e-22
Fitted Std. Err.	0.408	0.353	0.394	0.372	0.411	0.345	0.025
Conf. Int. Lower	-4.7e+00	-4.1e+00	-4.6e+00	-4.3e+00	-4.7e+00	-4.4e+00	2.7e-01
Conf. Int. Upper	-3.1e+00	-2.7e+00	-3.0e+00	-2.9e+00	-3.1e+00	-3.0e+00	1.7e-01
Train Accuracy	87.4%	85.2%	87.6%	85.4%	86.2%	86.4%	1.1%
Val Accuracy	82.2%	90.6%	81.5%	90.1%	87.2%	37.7%	4.3%
Train AUC	86.3%	84.0%	86.0%	83.9%	85.0%	85.0%	1.1%
Val AUC	79.5%	88.8%	81.1%	89.4%	85.3%	50.0%	4.5%
Train F1	89.8%	88.4%	90.4%	88.7%	89.0%	89.3%	0.8%
Test F1	87.3%	92.7%	84.1%	91.6%	90.4%	0.0%	3.5%
Train Precision	88.3%	88.2%	89.1%	88.2%	88.0%	88.4%	0.4%
Val Precision	88.7%	89.1%	84.9%	89.1%	89.7%	0.0%	1.9%
Train Recall	91.4%	88.5%	91.8%	89.3%	90.0%	90.2%	1.4%
Val Recall	85.9%	96.6%	83.3%	94.2%	91.2%	0.0%	5.6%
Train MCC	73.4%	68.2%	72.9%	68.1%	70.5%	70.6%	2.5%
Val MCC	57.8%	80.2%	62.1%	79.8%	71.2%	0.0%	10.2%
Train Log-Loss	4.54	5.32	4.47	5.25	4.98	4.91	0.39
Val Log-Loss	6.41	3.38	6.66	3.56	4.61	22.45	1.55

Univariate Report

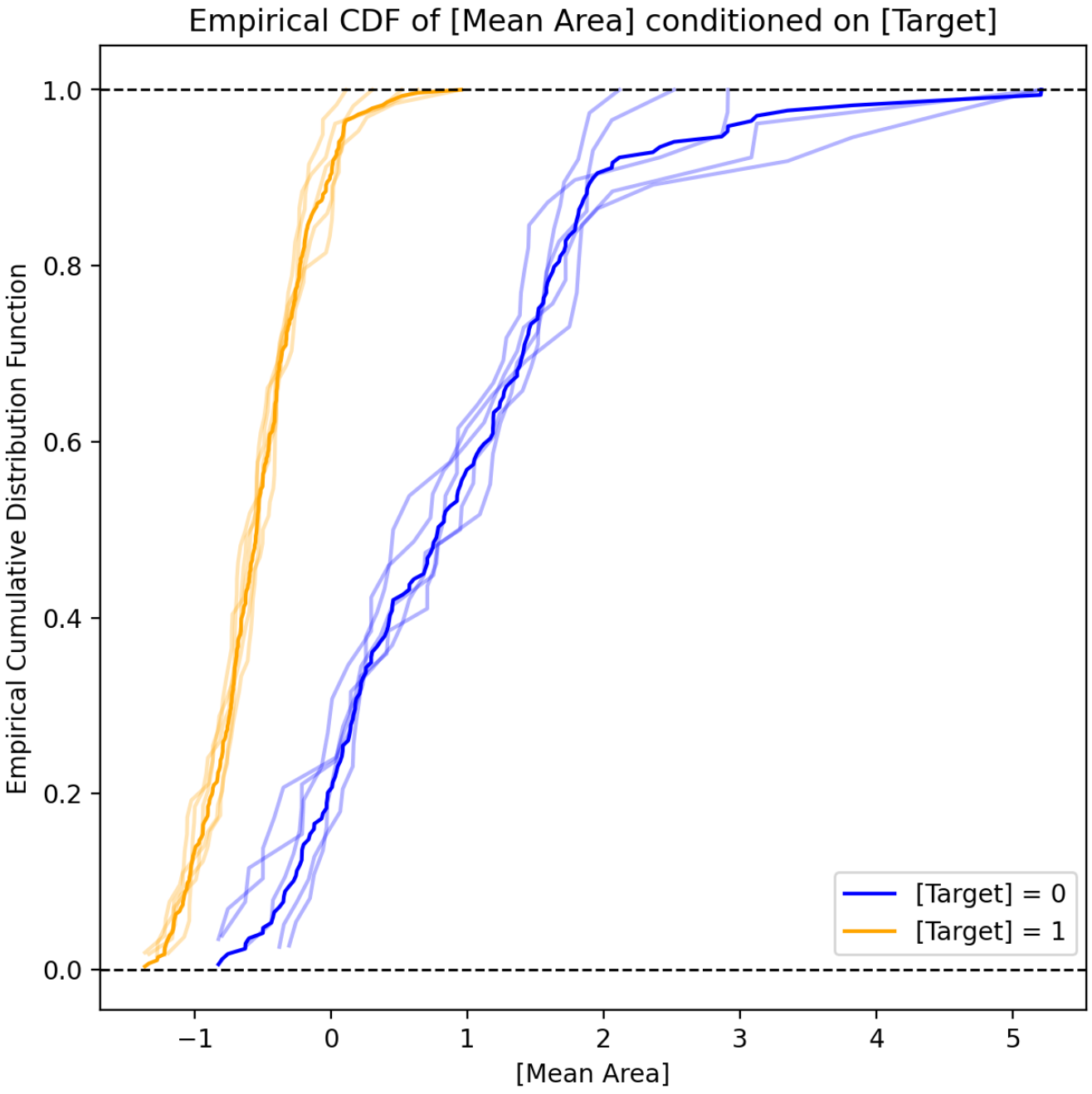
Mean Area - Kernel Density Plot



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

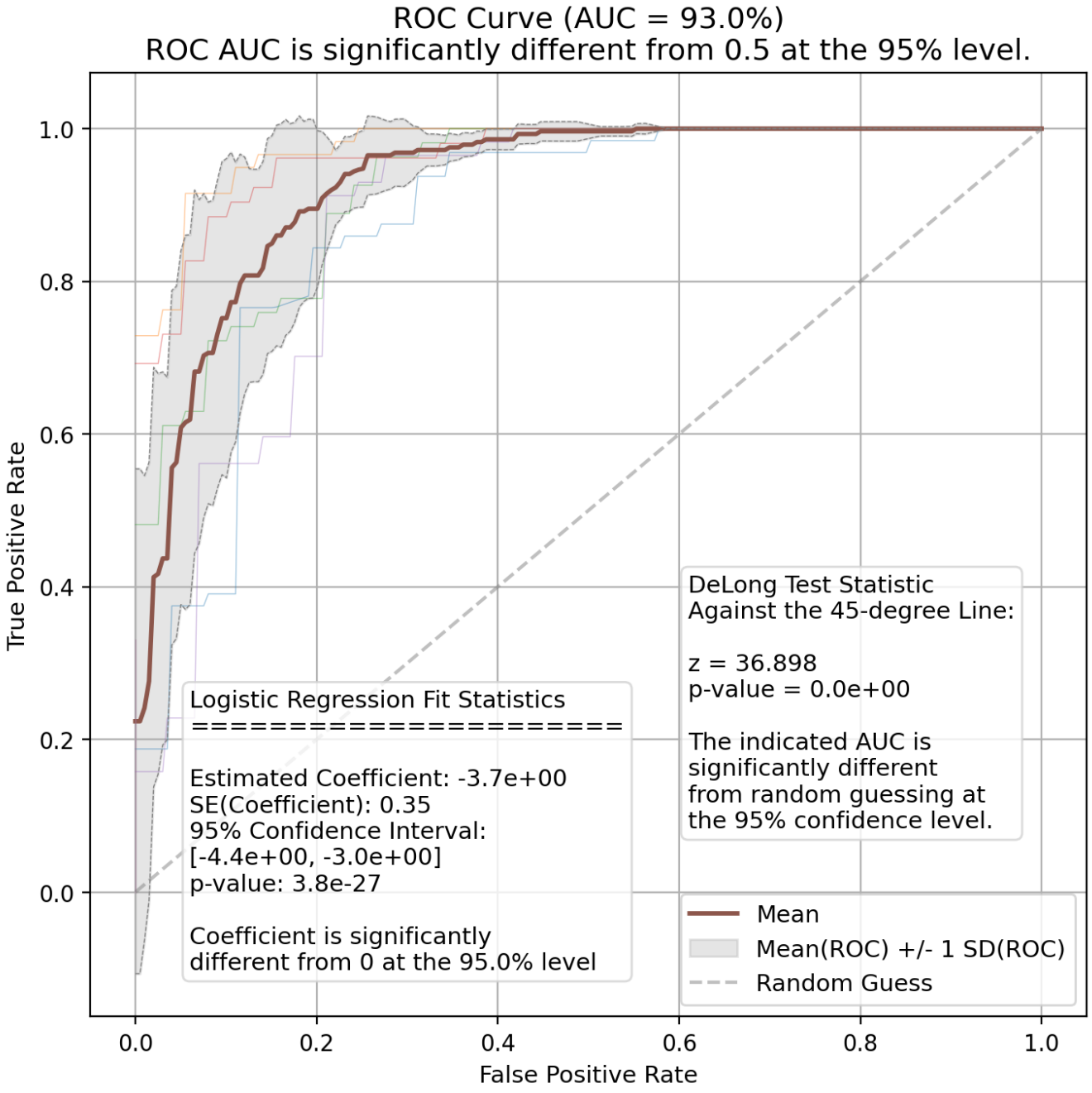
Mean Area - Empirical CDF Plot



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

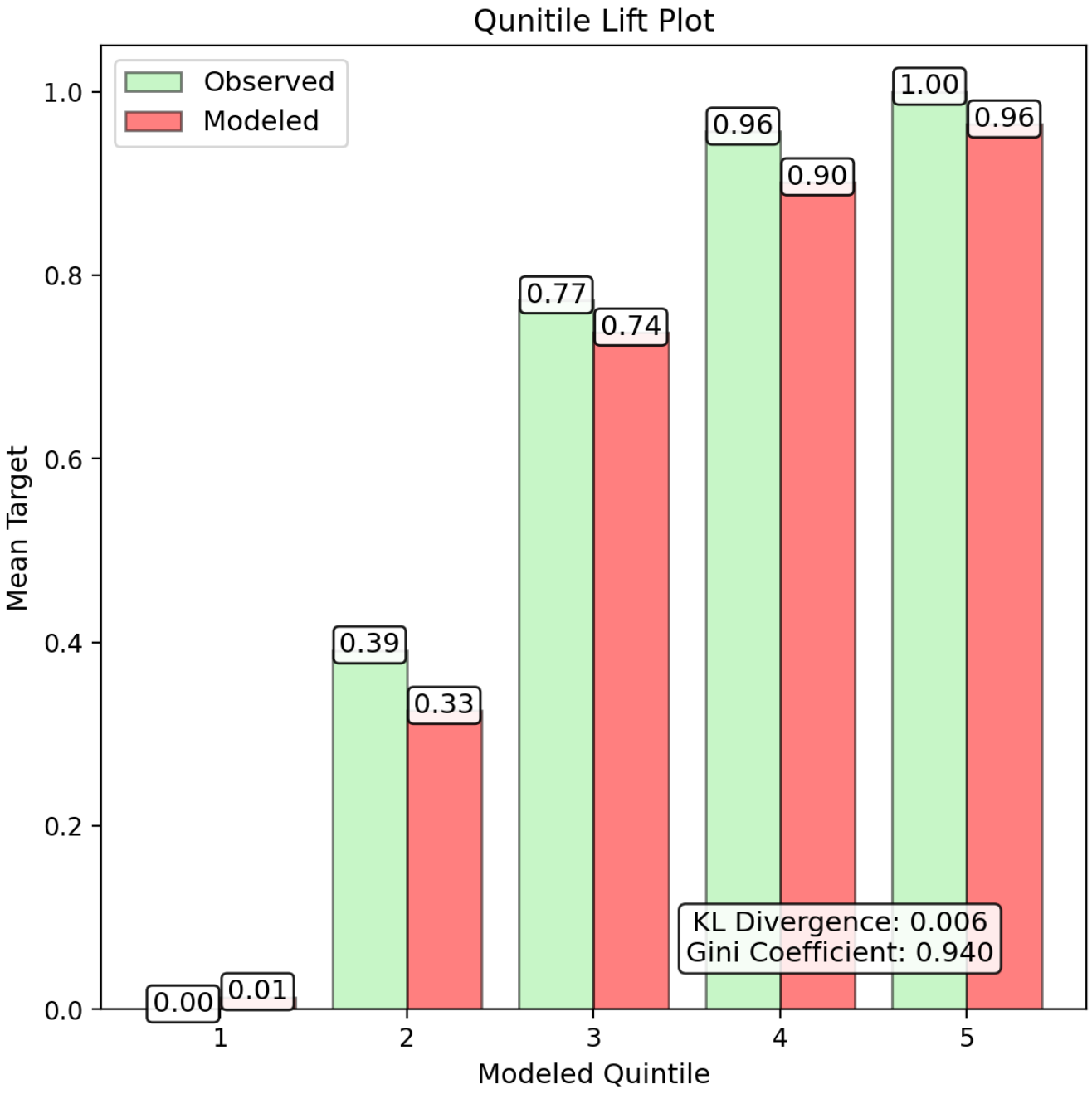


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

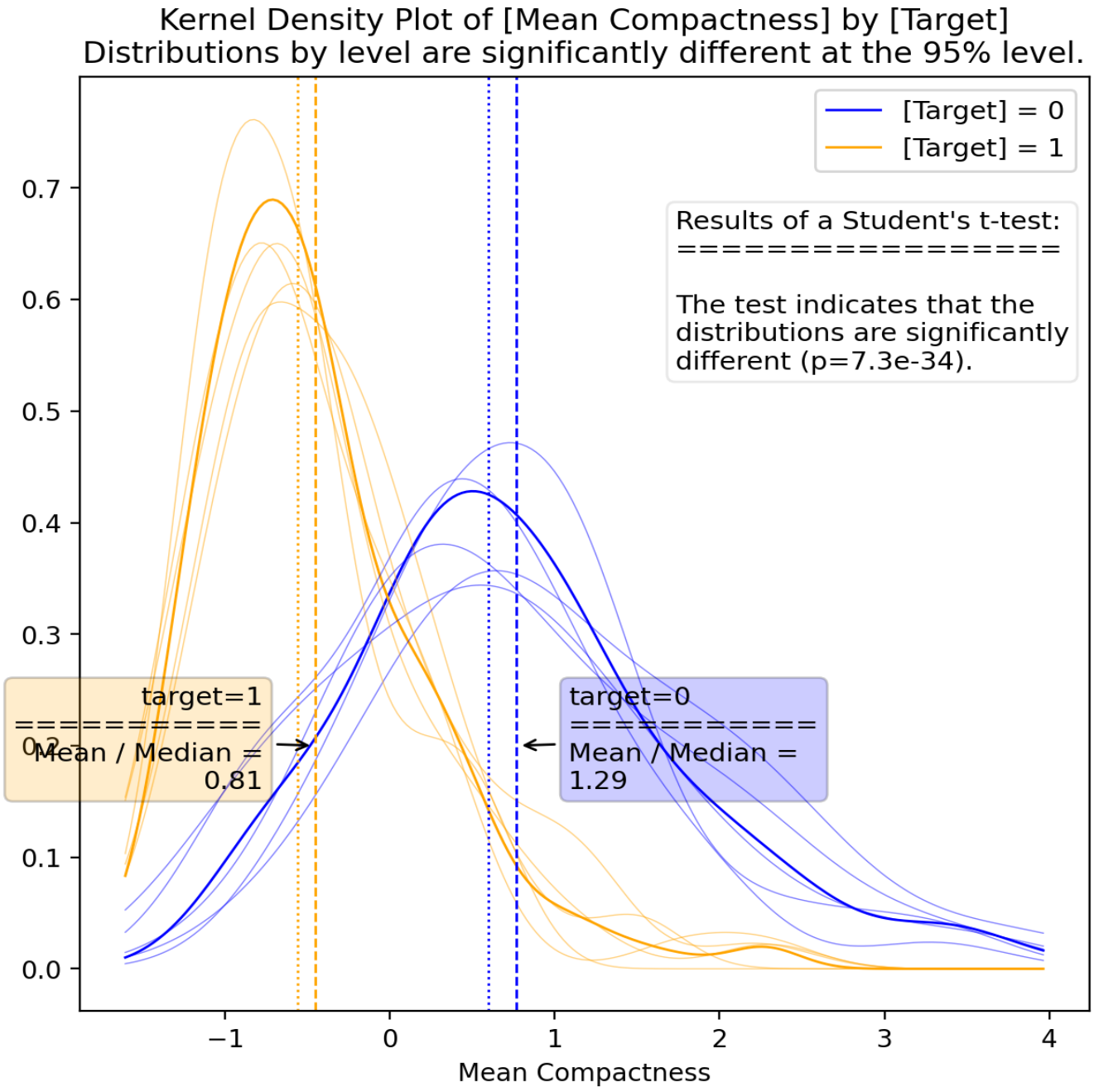
Univariate Report

Mean Compactness - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Agg. Mean	Agg. SD
Fitted Coef.	-1.9e+00	-1.7e+00	-1.9e+00	-1.7e+00	-1.8e+00	-1.8e+00	6.2e-02
Fitted p-Value	4.2e-21	1.5e-19	8.8e-21	8.6e-20	1.8e-20	4.9e-25	6.4e-20
Fitted Std. Err.	0.197	0.193	0.199	0.191	0.190	0.173	0.004
Conf. Int. Lower	-2.2e+00	-2.1e+00	-2.2e+00	-2.1e+00	-2.1e+00	-2.1e+00	6.9e-02
Conf. Int. Upper	-1.5e+00	-1.4e+00	-1.5e+00	-1.4e+00	-1.4e+00	-1.5e+00	5.5e-02
Train Accuracy	80.5%	78.3%	80.2%	79.4%	79.4%	79.6%	0.9%
Val Accuracy	75.6%	84.4%	77.2%	80.2%	80.2%	37.7%	3.4%
Train AUC	80.8%	78.2%	80.2%	79.9%	79.7%	79.7%	0.9%
Val AUC	74.8%	85.3%	77.8%	79.8%	80.0%	50.0%	3.8%
Train F1	83.3%	82.0%	83.8%	83.0%	82.6%	82.9%	0.7%
Test F1	81.7%	86.5%	79.2%	82.7%	84.4%	0.0%	2.8%
Train Precision	87.2%	86.0%	87.7%	88.4%	87.0%	87.3%	0.9%
Val Precision	87.5%	92.3%	85.1%	82.7%	88.5%	0.0%	3.6%
Train Recall	79.7%	78.4%	80.2%	78.2%	78.6%	79.0%	0.9%
Val Recall	76.6%	81.4%	74.1%	82.7%	80.7%	0.0%	3.6%
Train MCC	60.5%	55.1%	58.8%	57.8%	58.0%	58.1%	2.0%
Val MCC	46.4%	68.9%	54.8%	59.6%	58.0%	0.0%	8.1%
Train Log-Loss	7.01	7.83	7.15	7.43	7.42	7.37	0.31
Val Log-Loss	8.81	5.63	8.23	7.13	7.12	22.45	1.22

Univariate Report

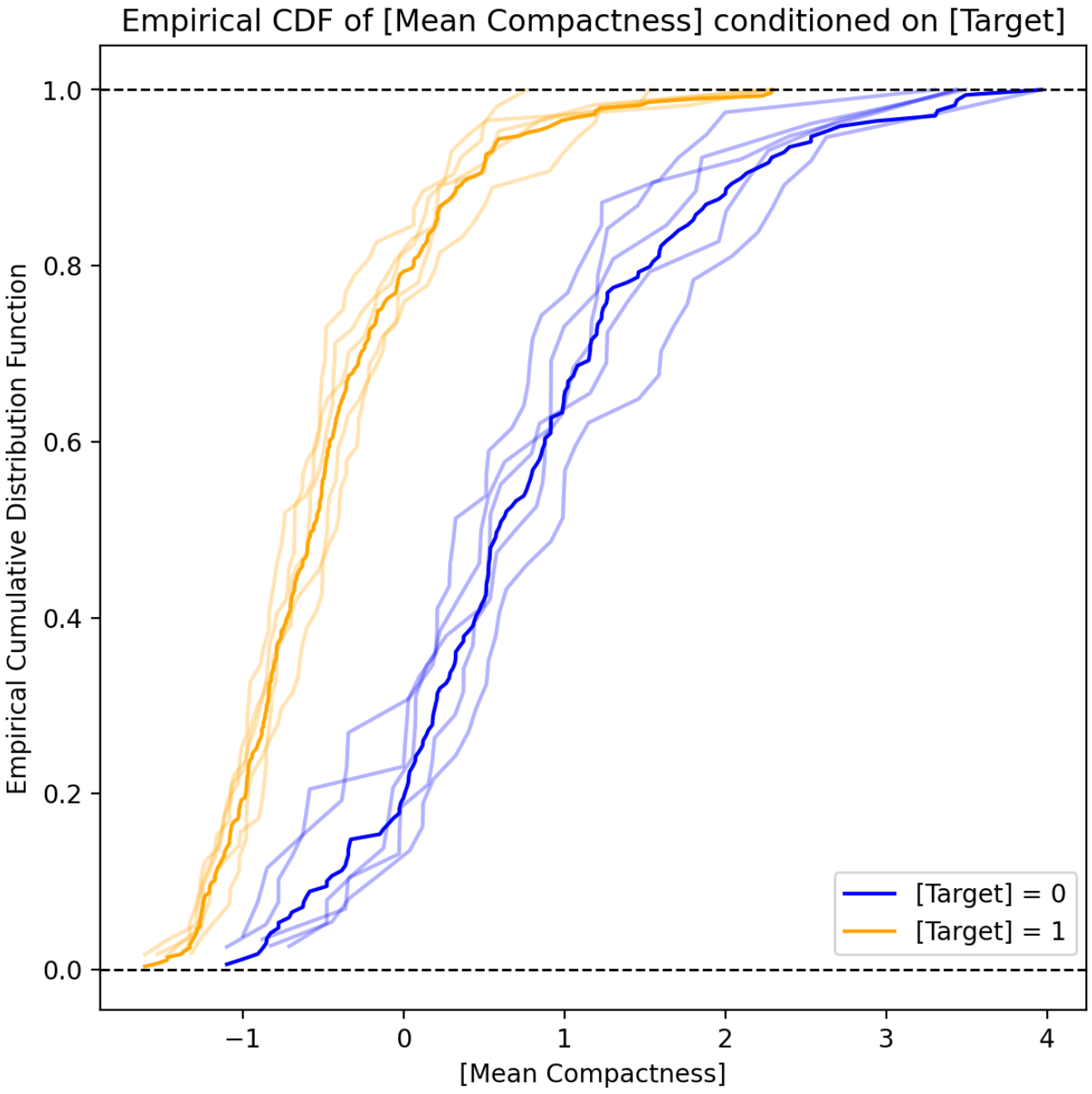
Mean Compactness - Kernel Density Plot



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

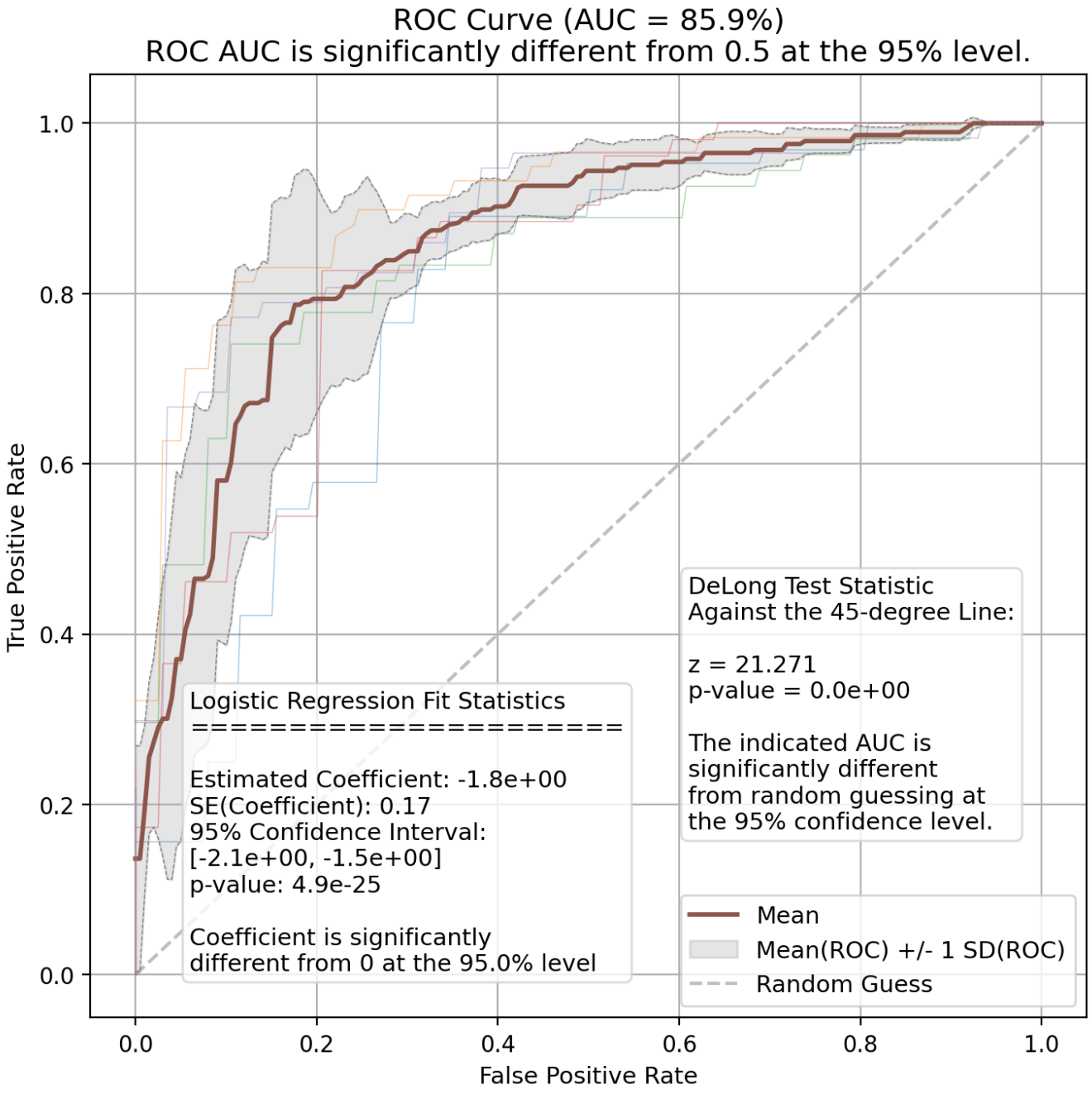
Mean Compactness - Empirical CDF Plot



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 Compactness - ROC Curve

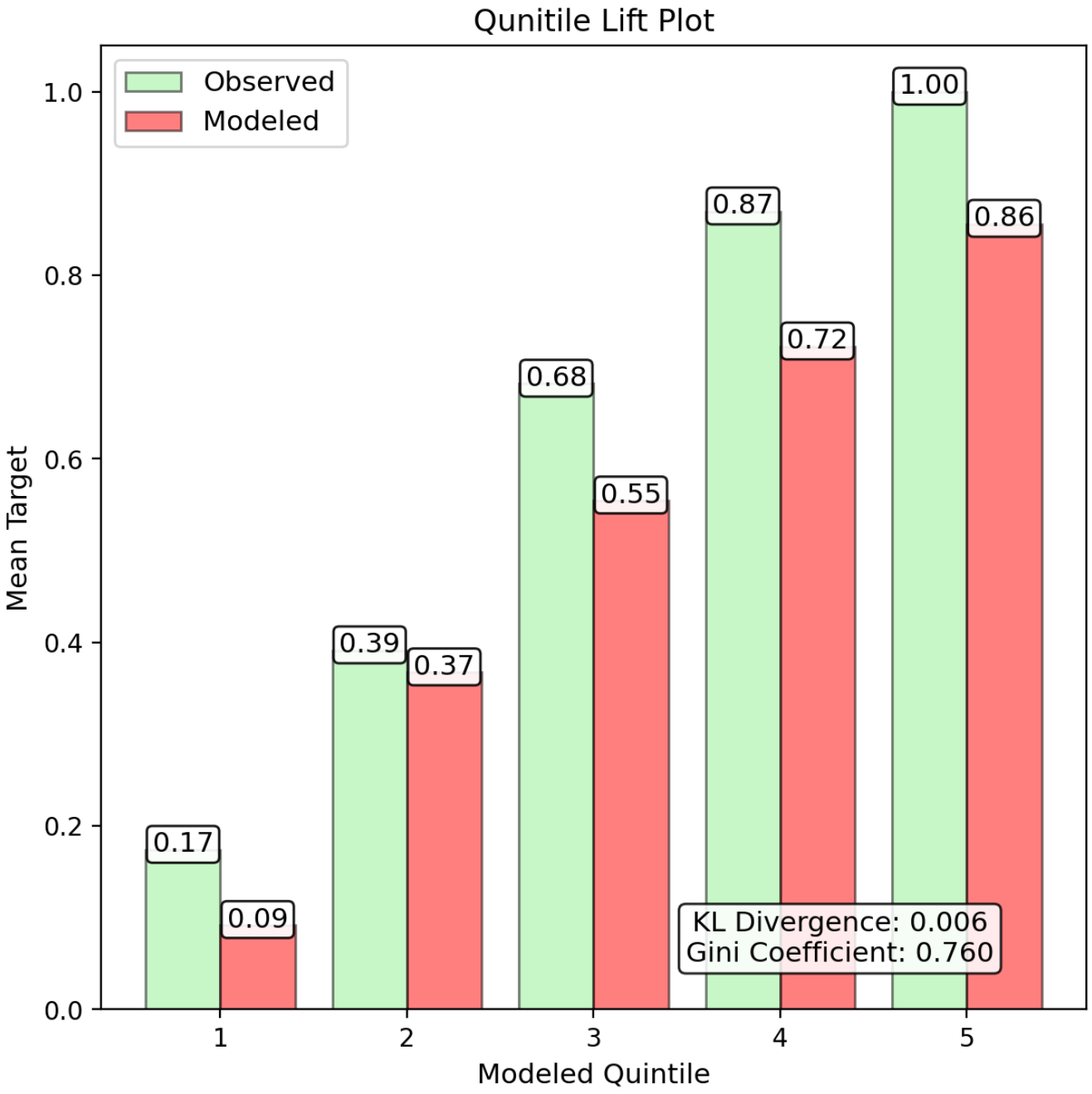


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

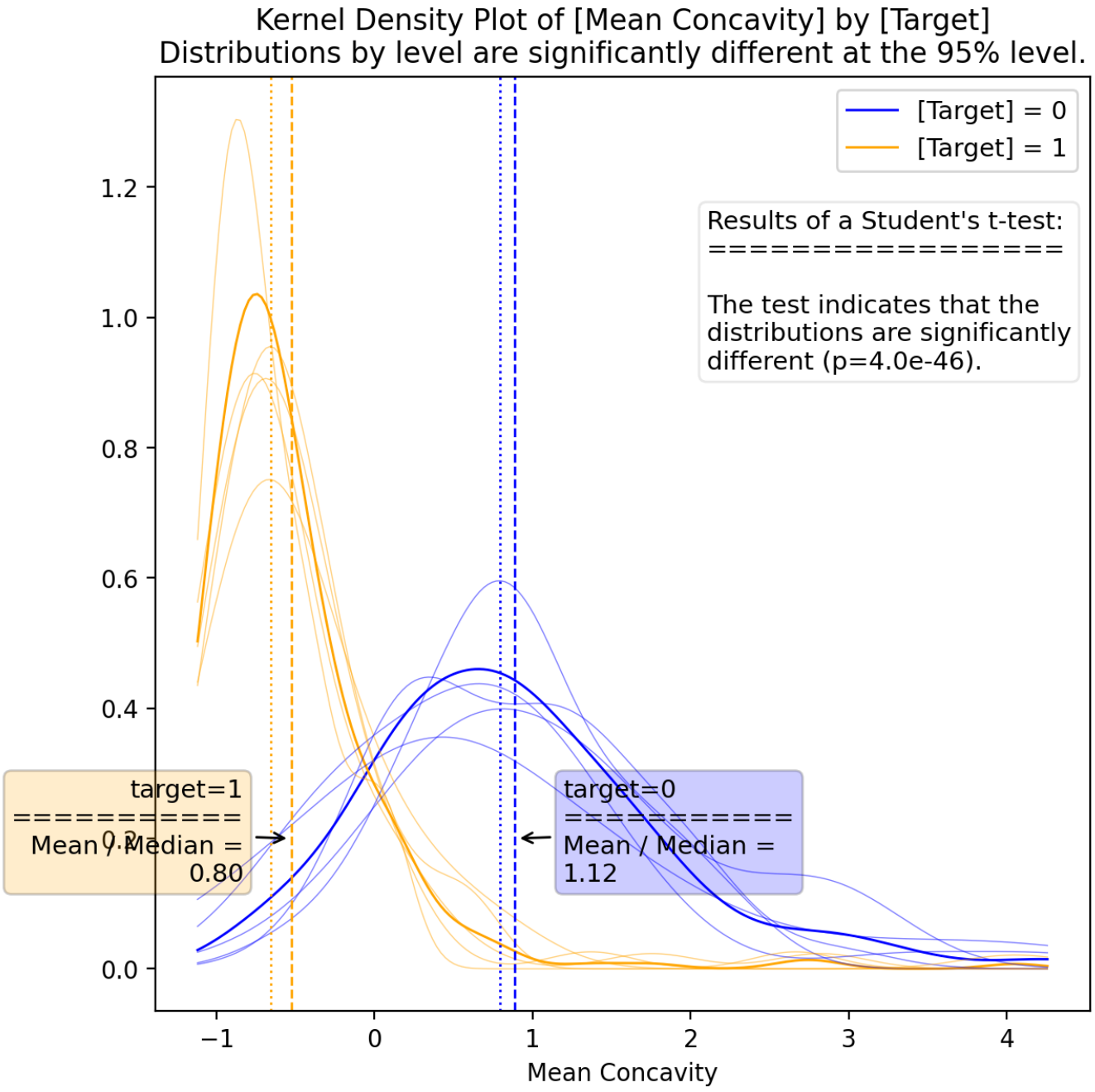
Univariate Report

Mean Concavity - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Agg. Mean	Agg. SD
Fitted Coef.	-2.7e+00	-2.8e+00	-2.9e+00	-2.5e+00	-2.7e+00	-2.7e+00	1.2e-01
Fitted p-Value	5.4e-25	1.3e-24	2.4e-25	3.1e-24	7.5e-25	1.3e-30	1.2e-24
Fitted Std. Err.	0.258	0.271	0.274	0.251	0.258	0.234	0.010
Conf. Int. Lower	-3.2e+00	-3.3e+00	-3.4e+00	-3.0e+00	-3.2e+00	-3.2e+00	1.4e-01
Conf. Int. Upper	-2.2e+00	-2.2e+00	-2.3e+00	-2.1e+00	-2.2e+00	-2.2e+00	9.8e-02
Train Accuracy	87.9%	86.9%	87.6%	87.9%	87.0%	87.5%	0.5%
Val Accuracy	85.6%	89.6%	87.0%	85.7%	89.5%	37.7%	2.0%
Train AUC	87.9%	86.5%	87.0%	88.4%	86.6%	87.3%	0.8%
Val AUC	84.1%	90.0%	87.7%	84.3%	90.4%	50.0%	3.0%
Train F1	89.9%	89.5%	90.2%	90.2%	89.4%	89.8%	0.4%
Test F1	89.6%	91.2%	88.2%	88.3%	91.7%	0.0%	1.6%
Train Precision	91.6%	90.9%	91.2%	94.0%	90.6%	91.6%	1.4%
Val Precision	91.8%	94.5%	93.8%	83.1%	96.2%	0.0%	5.2%
Train Recall	88.3%	88.1%	89.2%	86.8%	88.2%	88.1%	0.9%
Val Recall	87.5%	88.1%	83.3%	94.2%	87.7%	0.0%	3.9%
Train MCC	75.0%	72.2%	73.4%	74.9%	72.6%	73.6%	1.3%
Val MCC	66.2%	78.7%	74.4%	71.1%	78.1%	0.0%	5.2%
Train Log-Loss	4.34	4.72	4.47	4.36	4.69	4.52	0.18
Val Log-Loss	5.21	3.75	4.70	5.15	3.77	22.45	0.72

Univariate Report

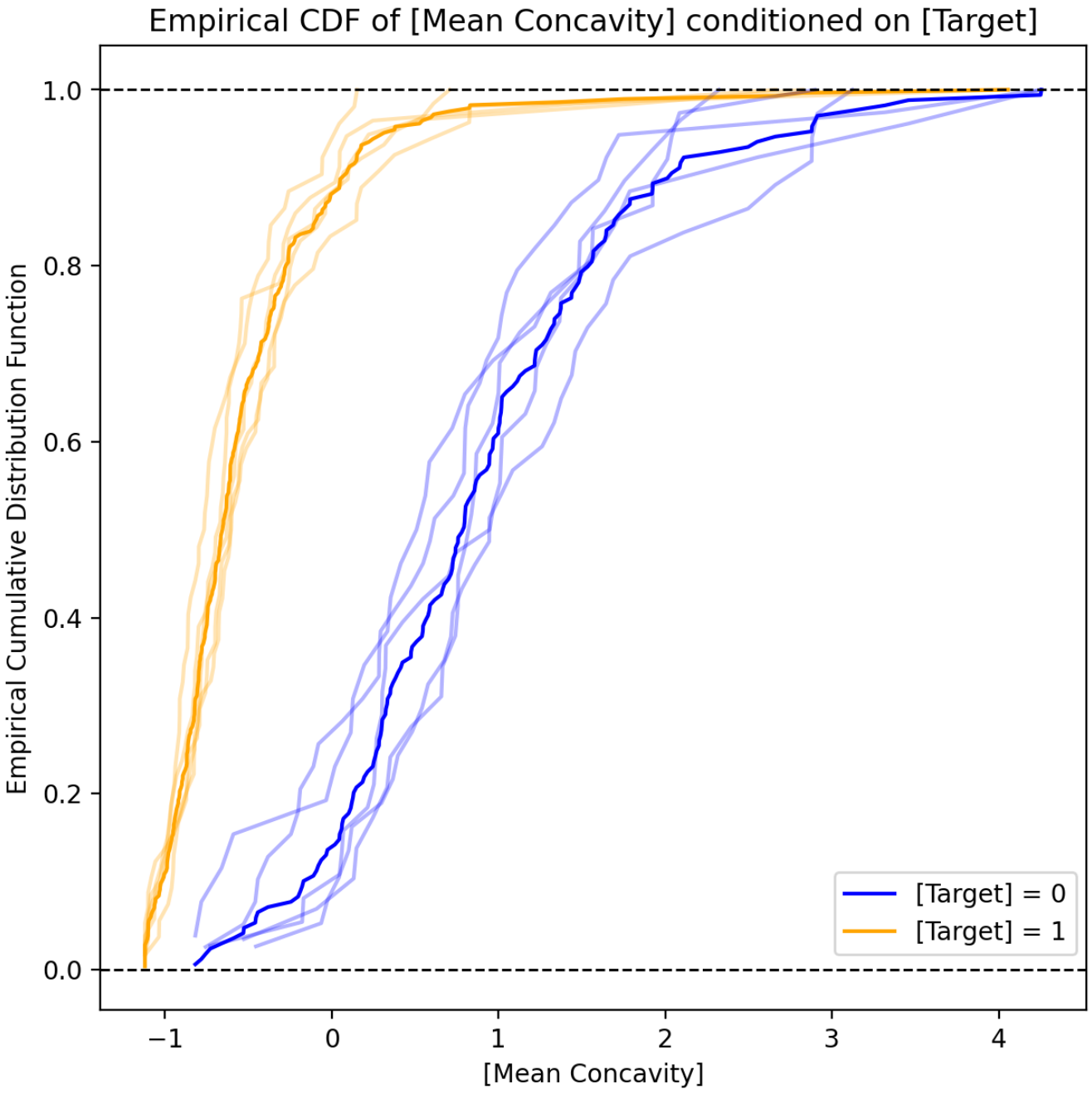
Mean Concavity - Kernel Density Plot



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

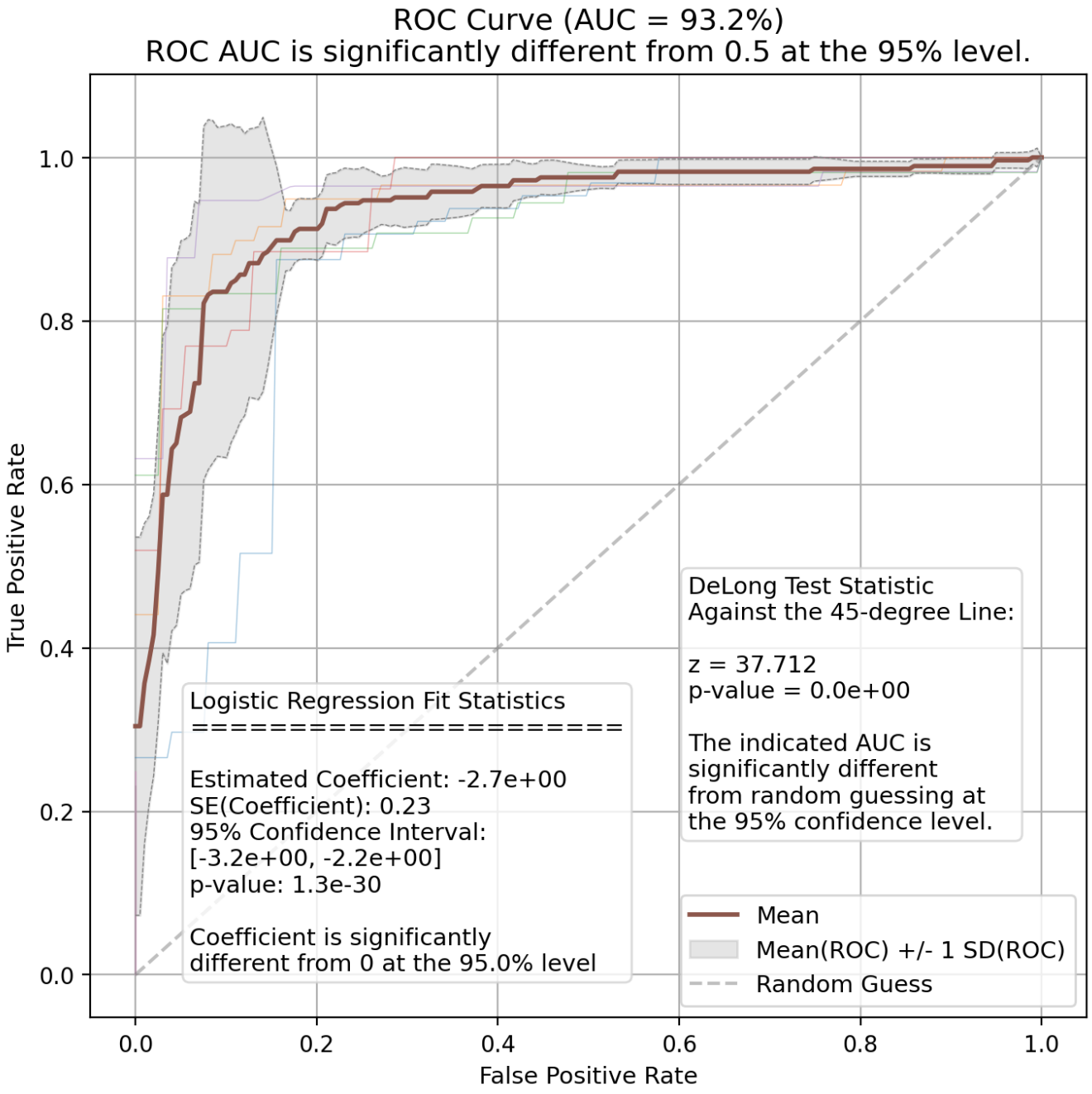
Mean Concavity - Empirical CDF Plot



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 Concavity - ROC Curve

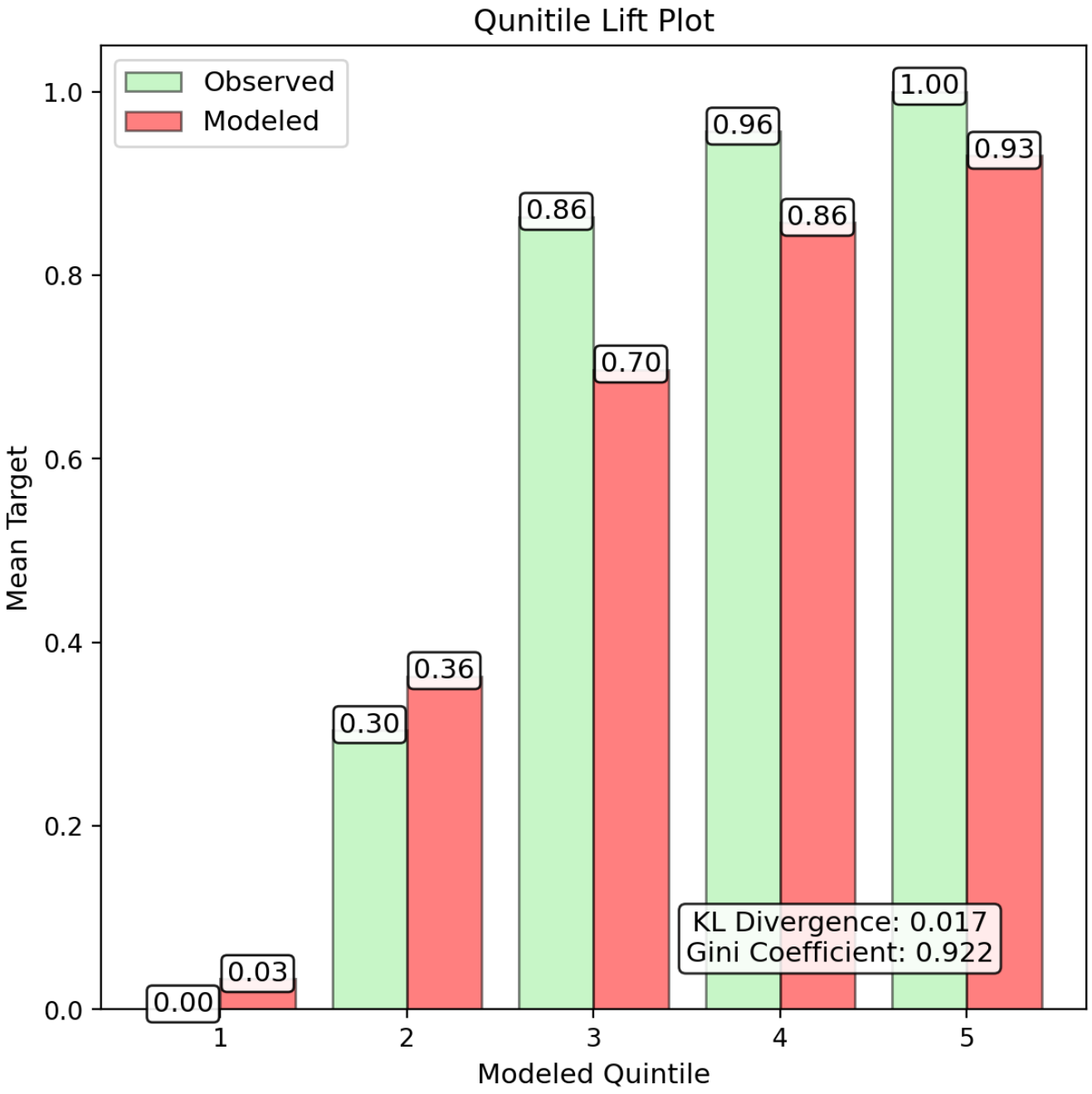


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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

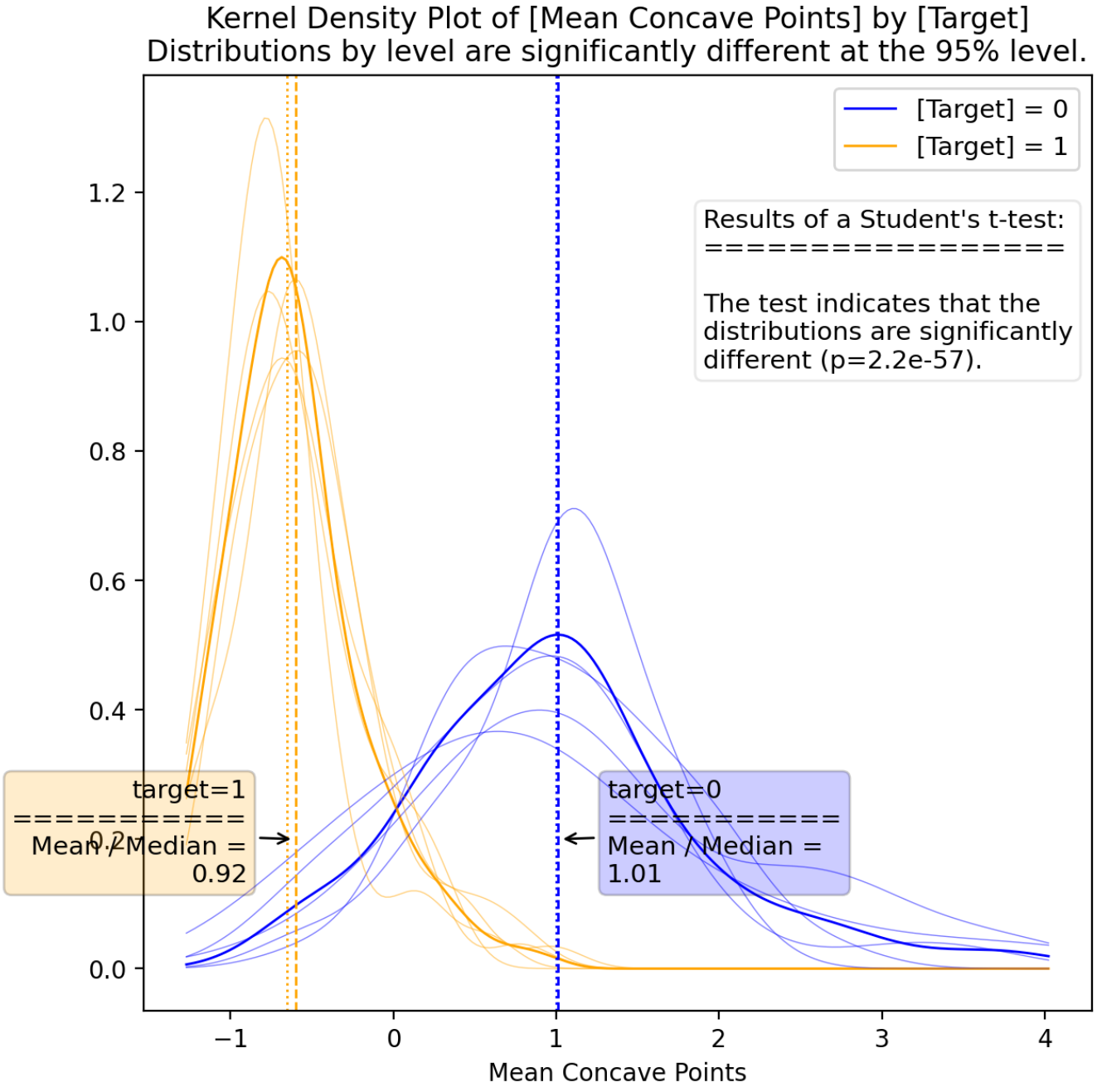
Univariate Report

Mean Concave Points - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Agg. Mean	Agg. SD
Fitted Coef.	-4.0e+00	-3.7e+00	-3.8e+00	-3.9e+00	-3.7e+00	-3.8e+00	1.1e-01
Fitted p-Value	8.3e-24	1.1e-24	1.5e-24	6.4e-24	4.5e-25	3.9e-30	3.5e-24
Fitted Std. Err.	0.393	0.363	0.367	0.385	0.360	0.334	0.015
Conf. Int. Lower	-4.7e+00	-4.4e+00	-4.5e+00	-4.6e+00	-4.4e+00	-4.5e+00	1.4e-01
Conf. Int. Upper	-3.2e+00	-3.0e+00	-3.0e+00	-3.1e+00	-3.0e+00	-3.2e+00	7.9e-02
Train Accuracy	91.5%	90.8%	90.6%	91.2%	90.8%	91.0%	0.4%
Val Accuracy	88.9%	91.7%	92.4%	90.1%	91.9%	37.7%	1.4%
Train AUC	91.6%	90.7%	90.2%	91.3%	90.6%	90.9%	0.6%
Val AUC	86.5%	91.7%	93.1%	89.7%	92.2%	50.0%	2.6%
Train F1	92.9%	92.6%	92.6%	93.0%	92.5%	92.7%	0.2%
Test F1	92.2%	93.1%	93.2%	91.4%	93.7%	0.0%	0.9%
Train Precision	94.8%	94.1%	93.4%	95.1%	93.7%	94.2%	0.7%
Val Precision	92.2%	94.7%	98.0%	90.6%	96.3%	0.0%	3.0%
Train Recall	91.0%	91.2%	91.8%	91.0%	91.3%	91.3%	0.3%
Val Recall	92.2%	91.5%	88.9%	92.3%	91.2%	0.0%	1.4%
Train MCC	82.5%	80.5%	79.9%	81.3%	80.6%	81.0%	1.0%
Val MCC	73.0%	82.7%	85.1%	79.8%	82.5%	0.0%	4.7%
Train Log-Loss	3.06	3.31	3.38	3.17	3.32	3.25	0.13
Val Log-Loss	4.00	3.00	2.74	3.56	2.93	22.45	0.52

Univariate Report

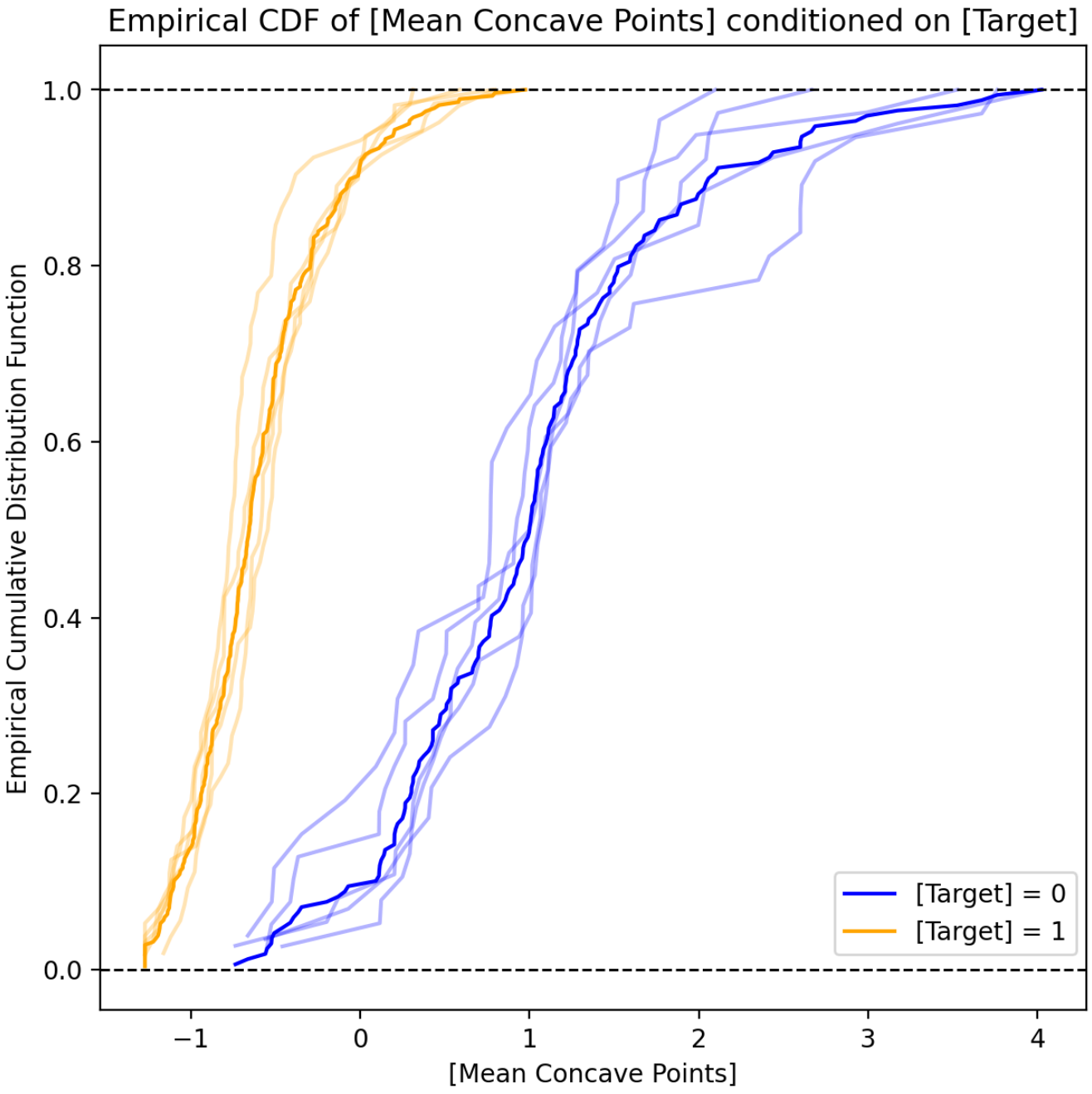
Mean Concave Points - Kernel Density Plot



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

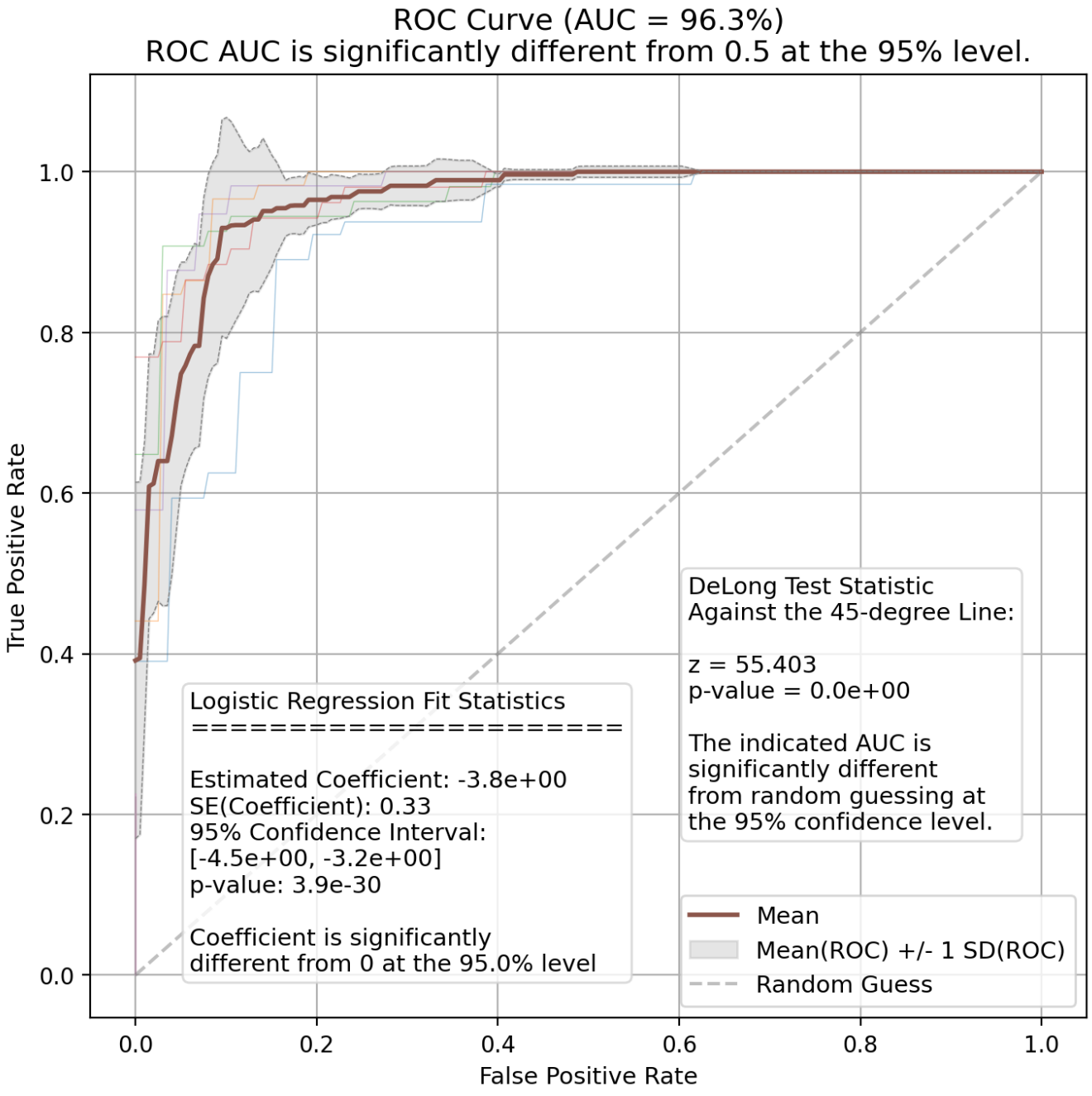
Mean Concave Points - Empirical CDF Plot



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 Concave Points - ROC Curve

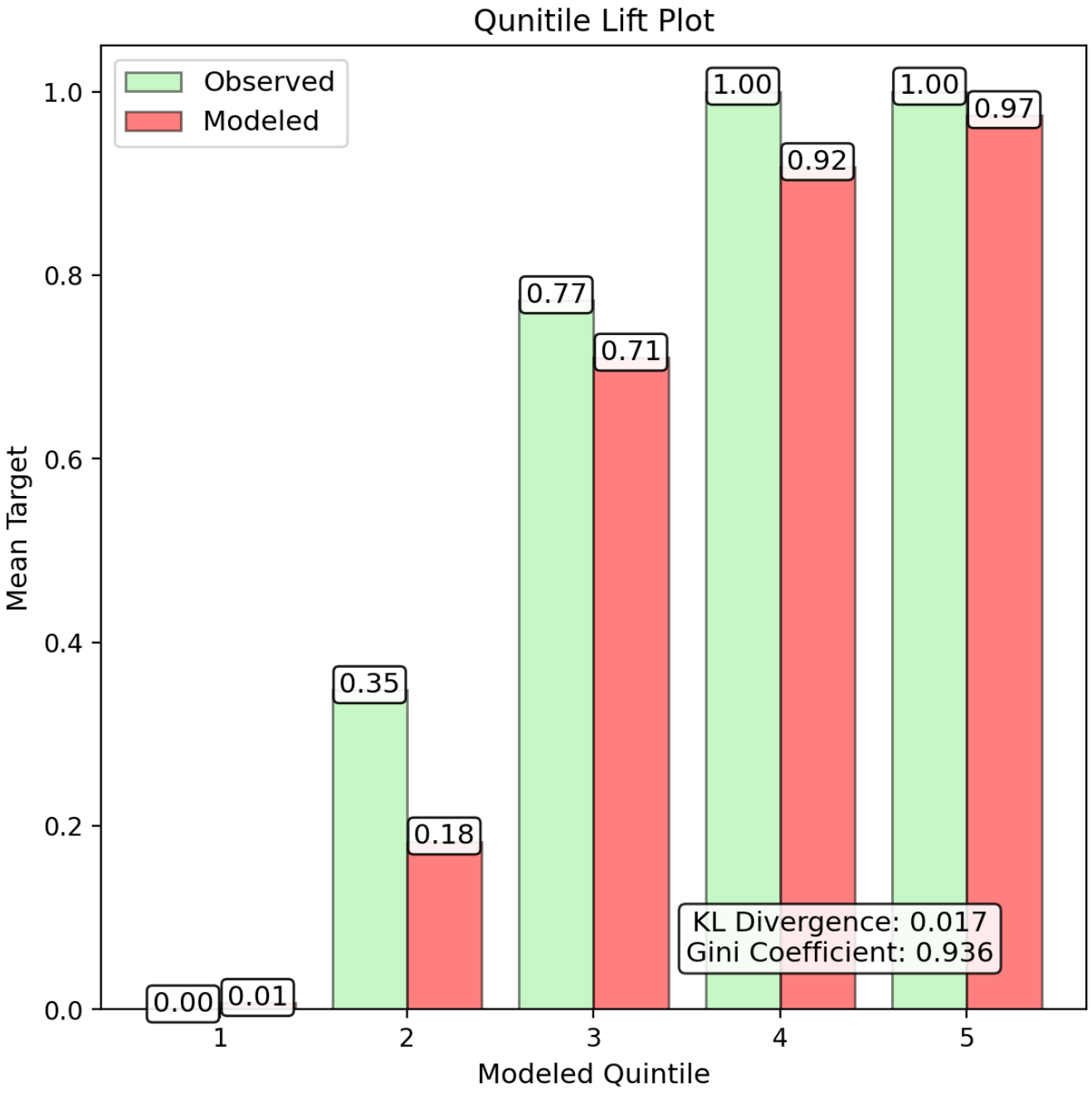


This plot shows the receiver operating characteristic (ROC) curve for the target variable in total and for each fold. The x-axis represents the false positive rate, and the y-axis represents the true positive rate. This is based on a simple Logistic Regression model with no regularization, no intercept,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