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.
comp_1	90.7%	89.4%	96.6%	88.6%	92.9%	80.0%	89.7%
comp_2	51.3%	64.1%	51.0%	51.4%	56.8%	2.8%	46.2%

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 2 of the variables passed to the model.

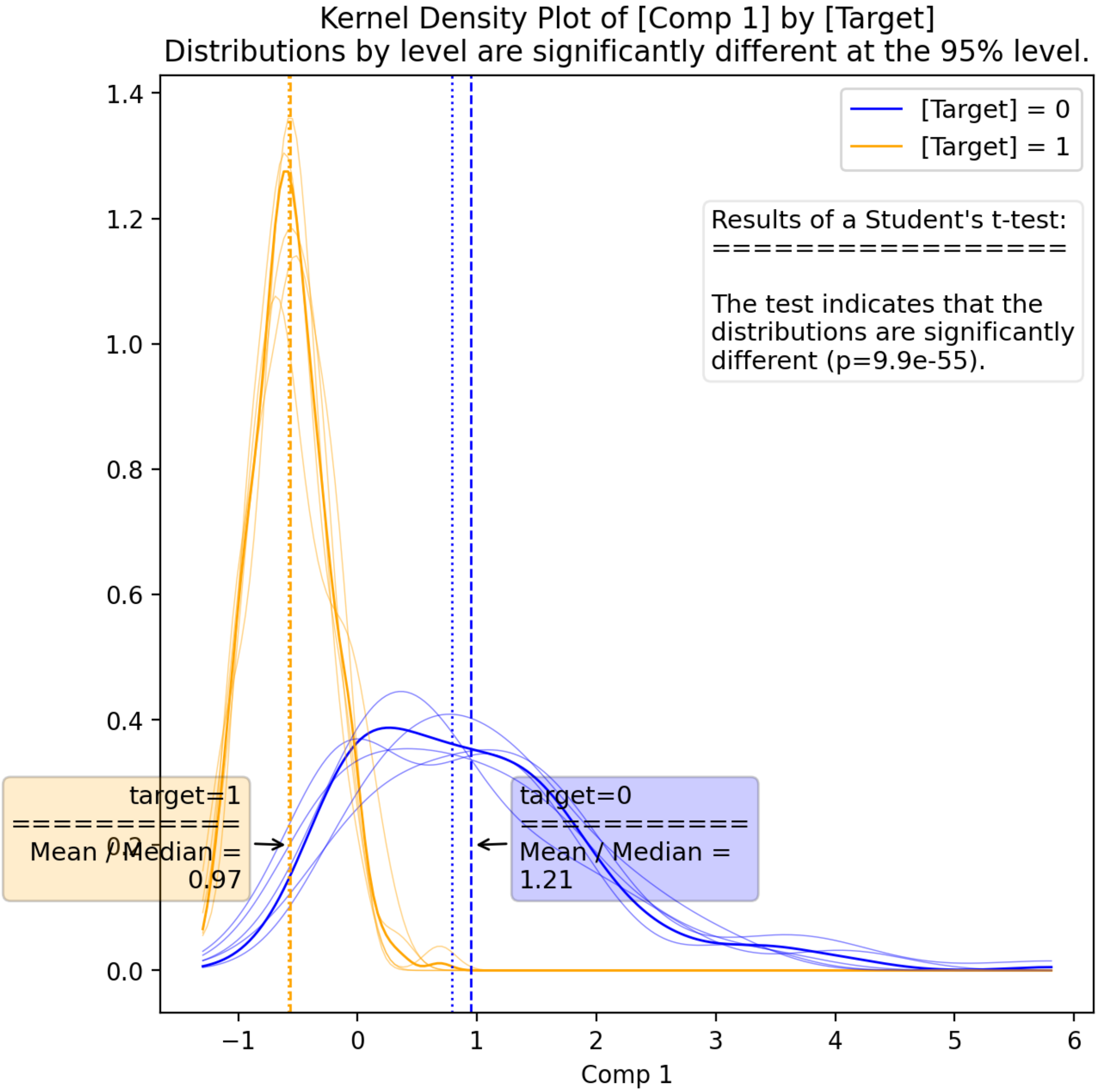
Univariate Report

Comp 1 - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Agg. Mean	Agg. SD
Fitted Coef.	-4.8e+00	-4.7e+00	-4.9e+00	-5.3e+00	-4.6e+00	-7.3e-03	2.8e-01
Fitted p-Value	2.3e-27	2.8e-27	1.6e-25	1.3e-24	3.5e-27	1.2e-32	5.7e-25
Fitted Std. Err.	0.440	0.435	0.471	0.519	0.424	0.001	0.038
Conf. Int. Lower	-5.6e+00	-5.6e+00	-5.8e+00	-6.3e+00	-5.4e+00	-8.5e-03	3.6e-01
Conf. Int. Upper	-3.9e+00	-3.9e+00	-4.0e+00	-4.3e+00	-3.8e+00	-6.1e-03	2.1e-01
Train Accuracy	90.5%	90.0%	91.2%	91.9%	89.4%	90.7%	1.0%
Val Accuracy	90.6%	92.5%	89.6%	86.9%	96.3%	90.7%	3.5%
Train AUC	88.3%	87.8%	89.5%	90.3%	87.3%	88.6%	1.2%
Val AUC	90.7%	92.0%	84.5%	83.0%	95.1%	88.6%	5.1%
Train F1	92.8%	92.5%	93.1%	93.8%	91.8%	92.9%	0.7%
Test F1	92.1%	93.4%	92.9%	90.4%	97.3%	92.9%	2.5%
Train Precision	89.3%	89.9%	89.4%	91.0%	87.8%	89.4%	1.1%
Val Precision	94.1%	89.1%	88.6%	83.3%	95.9%	89.4%	5.0%
Train Recall	96.5%	95.3%	97.1%	96.8%	96.1%	96.6%	0.7%
Val Recall	90.1%	98.3%	97.5%	98.7%	98.6%	96.6%	3.7%
Train MCC	79.4%	77.9%	81.5%	82.7%	77.4%	80.0%	2.3%
Val MCC	80.6%	85.4%	74.8%	72.8%	91.6%	80.0%	7.7%
Train Log-Loss	3.43	3.59	3.18	2.90	3.83	3.36	0.36
Val Log-Loss	3.39	2.69	3.76	4.73	1.33	3.36	1.27

Univariate Report

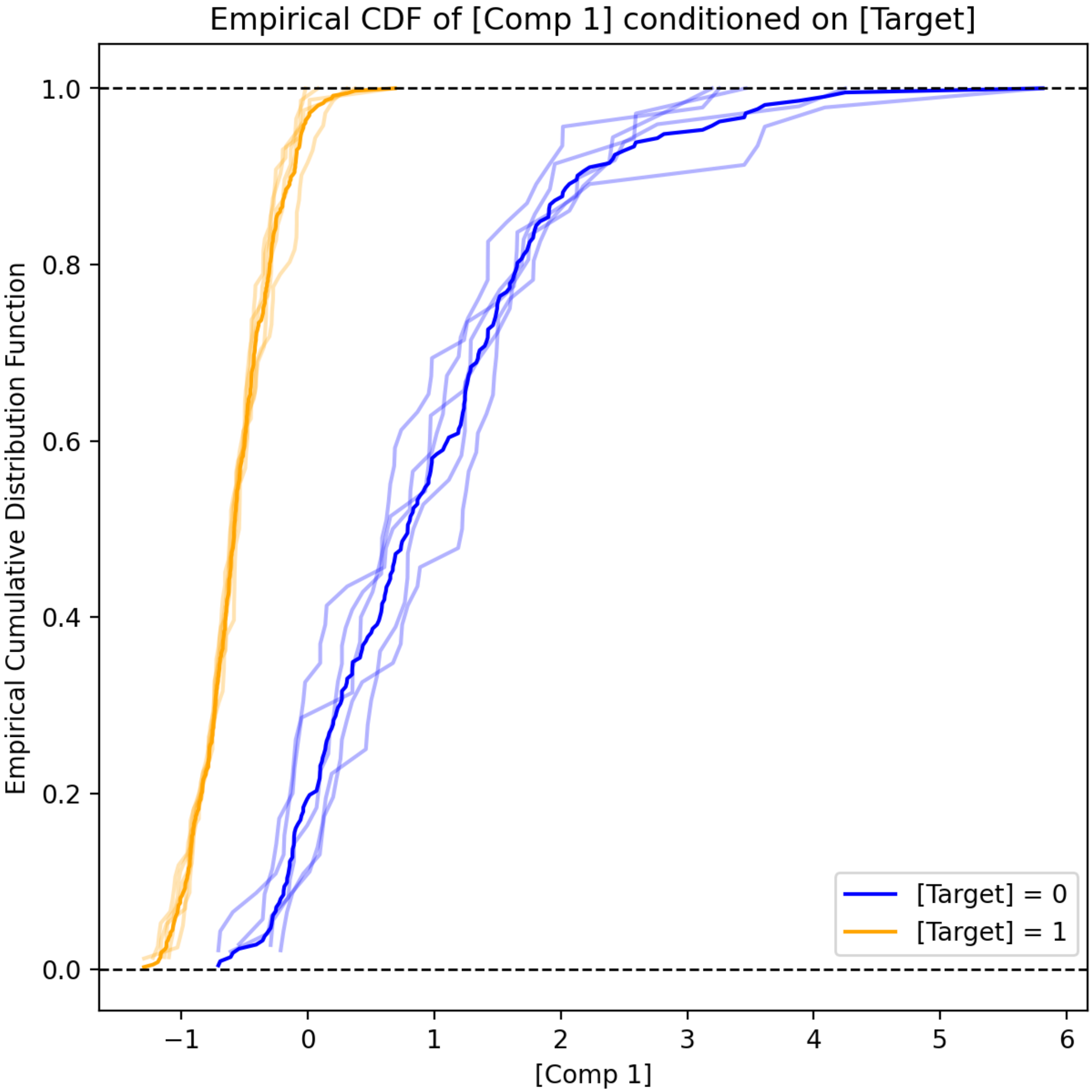
Comp 1 - Kernel Density Plot



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

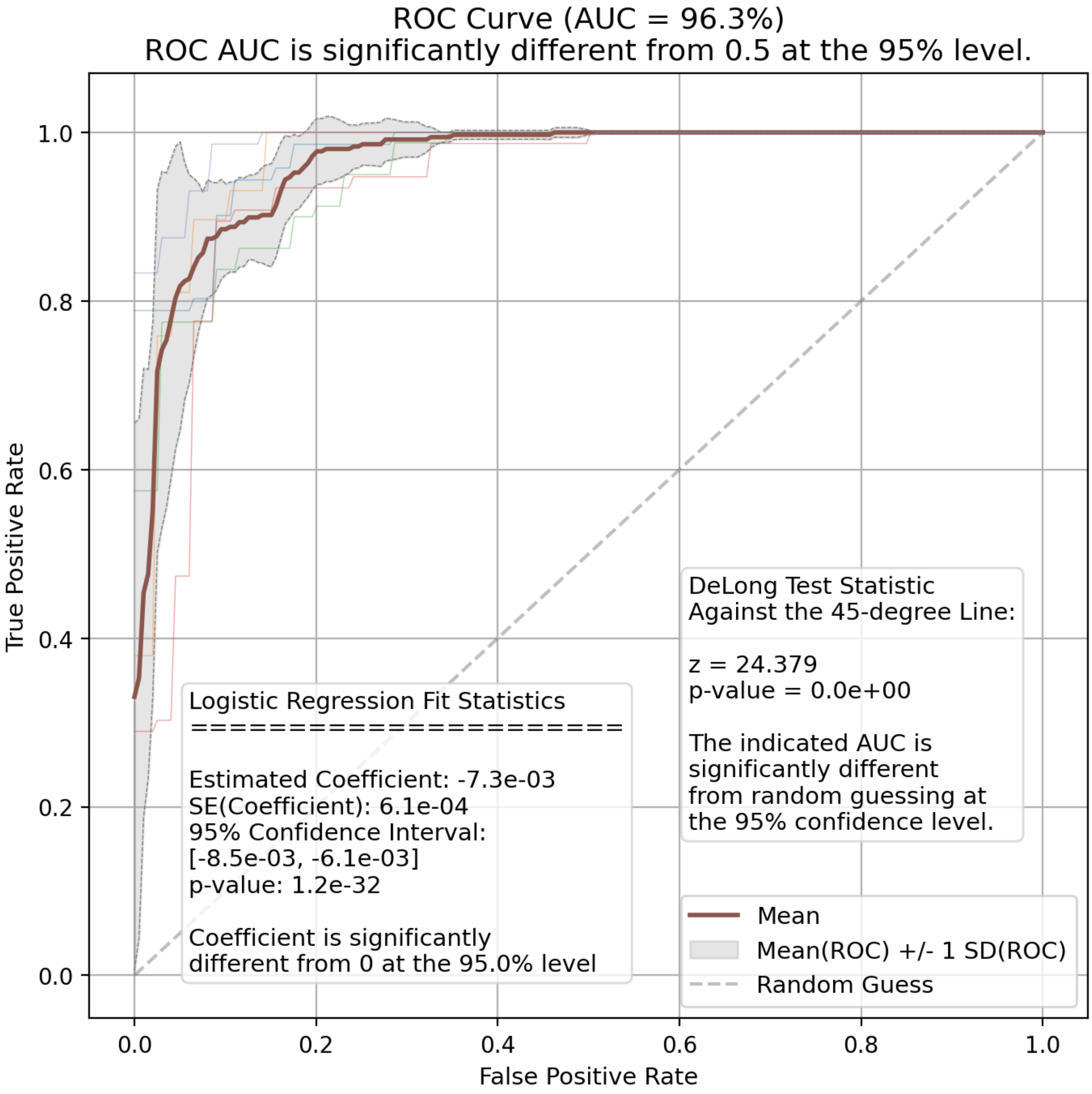
Comp 1 - Empirical CDF Plot



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

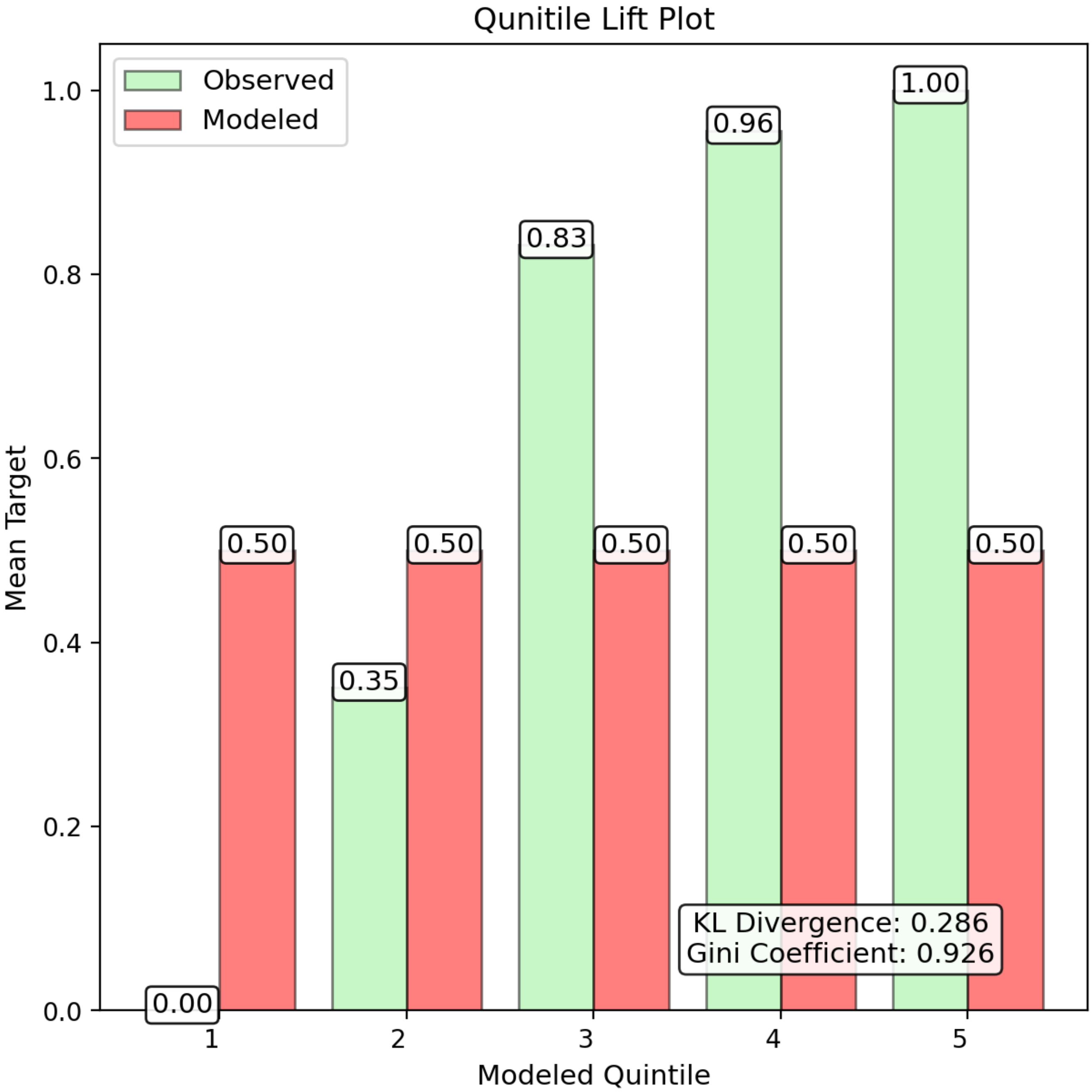
Comp 1 - ROC Curve



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

Comp 1 - 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.

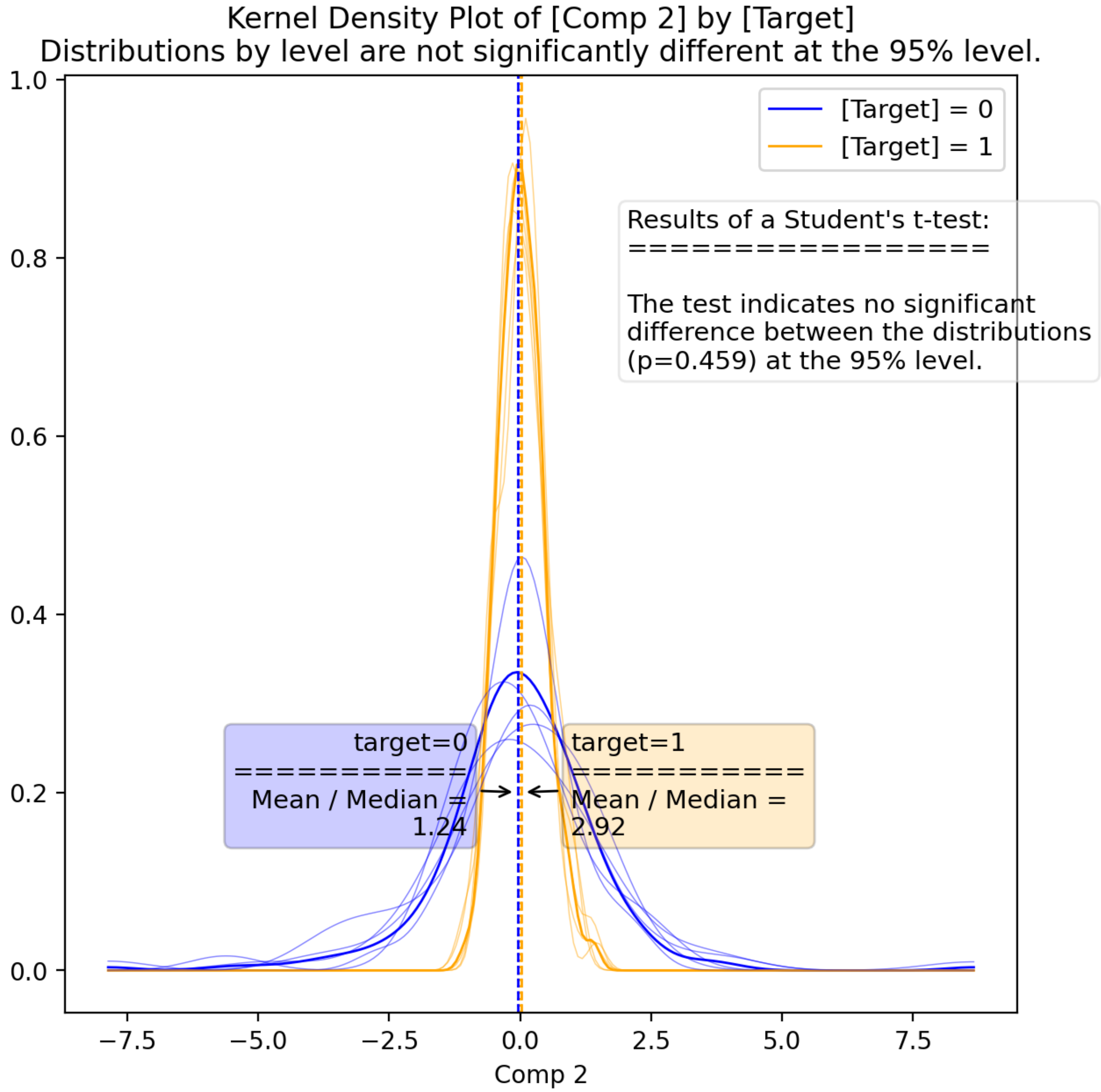
Univariate Report

Comp 2 - Results

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Agg. Mean	Agg. SD
Fitted Coef.	1.3e-01	5.1e-02	5.0e-02	8.6e-02	6.0e-02	8.7e-04	3.6e-02
Fitted p-Value	0.201	0.604	0.584	0.340	0.523	0.382	0.174
Fitted Std. Err.	1.0e-01	9.8e-02	9.2e-02	9.0e-02	9.3e-02	1.0e-03	6.0e-03
Conf. Int. Lower	-7.1e-02	-1.4e-01	-1.3e-01	-9.1e-02	-1.2e-01	-1.1e-03	2.9e-02
Conf. Int. Upper	0.340	0.244	0.230	0.263	0.242	0.003	0.044
Train Accuracy	52.0%	50.6%	51.1%	52.1%	50.1%	51.3%	0.9%
Val Accuracy	53.8%	50.5%	50.4%	48.4%	54.6%	51.3%	2.6%
Train AUC	51.7%	50.9%	51.2%	52.3%	50.5%	51.4%	0.7%
Val AUC	53.9%	50.7%	51.5%	48.3%	53.5%	51.4%	2.3%
Train F1	58.2%	56.8%	56.0%	57.5%	54.7%	56.8%	1.4%
Test F1	58.5%	51.4%	57.8%	54.0%	62.6%	56.8%	4.3%
Train Precision	64.8%	65.5%	62.1%	65.0%	62.3%	64.1%	1.6%
Val Precision	64.4%	54.9%	70.9%	60.7%	69.5%	64.1%	6.6%
Train Recall	52.8%	50.2%	50.9%	51.6%	48.8%	51.0%	1.5%
Val Recall	53.5%	48.3%	48.8%	48.7%	56.9%	51.0%	3.8%
Train MCC	3.3%	1.6%	2.3%	4.5%	1.0%	2.8%	1.4%
Val MCC	7.7%	1.3%	2.8%	-3.4%	6.6%	2.8%	4.4%
Train Log-Loss	17.30	17.79	17.62	17.26	17.98	17.55	0.31
Val Log-Loss	16.64	17.85	17.87	18.61	16.35	17.55	0.94

Univariate Report

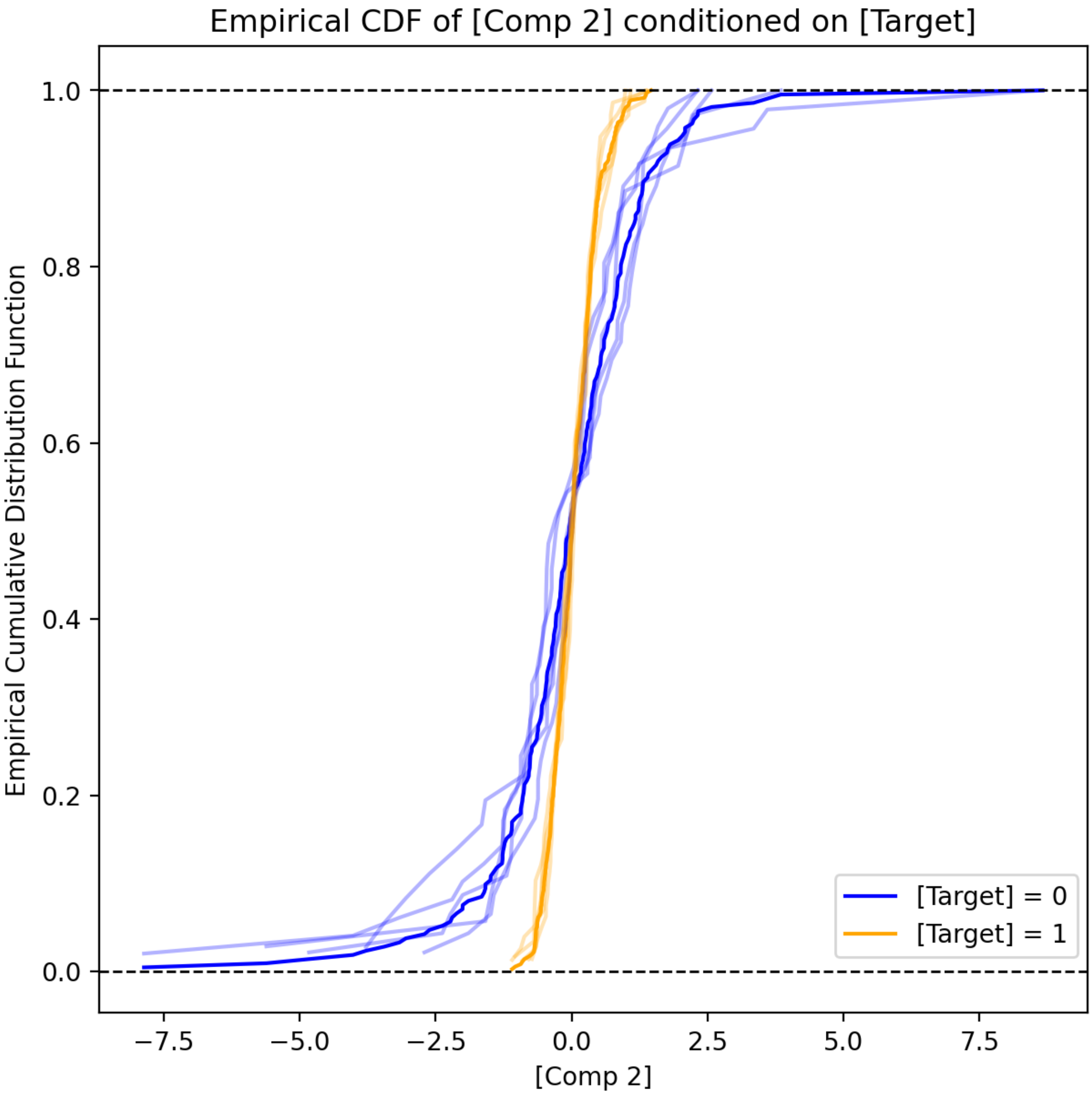
Comp 2 - Kernel Density Plot



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

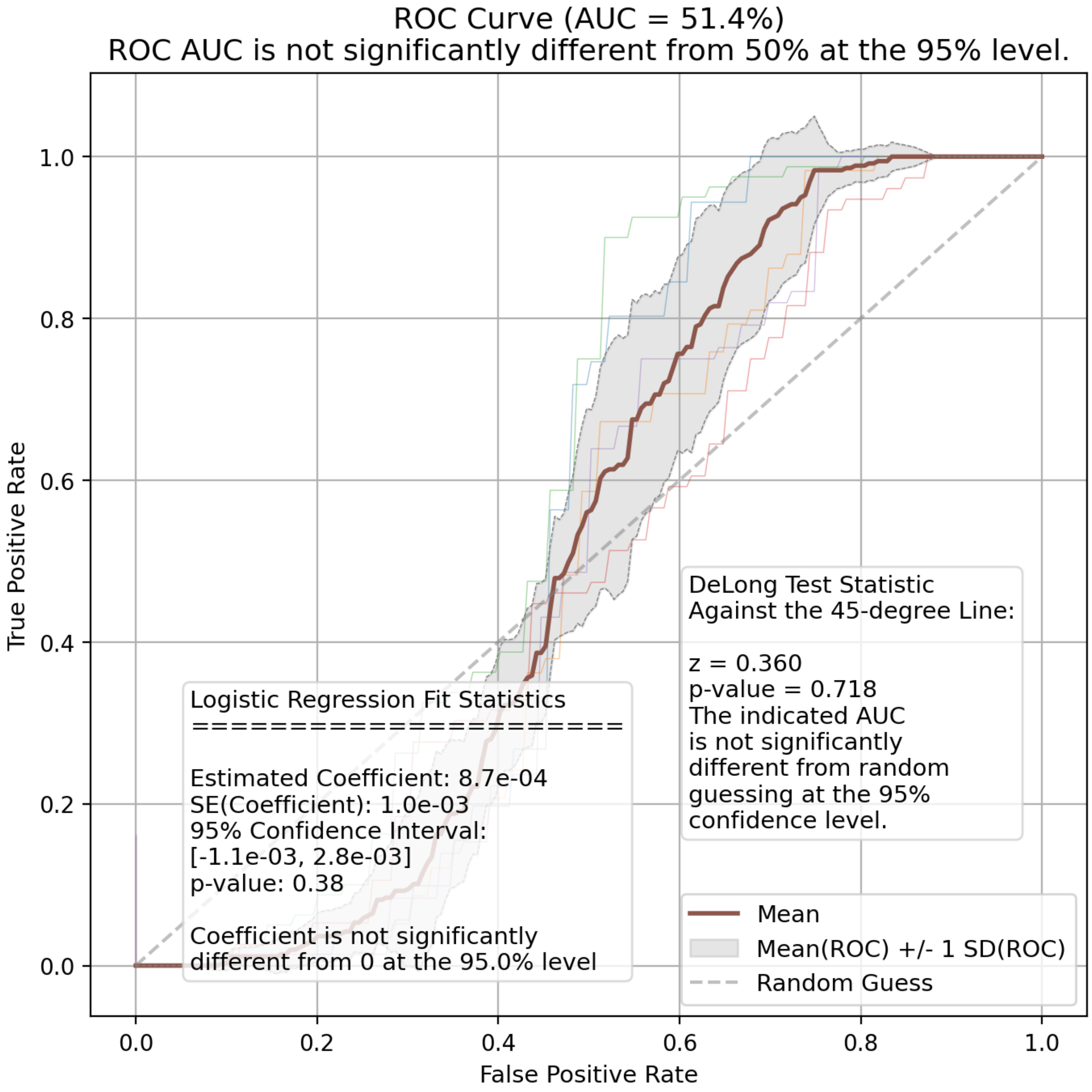
Comp 2 - Empirical CDF Plot



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

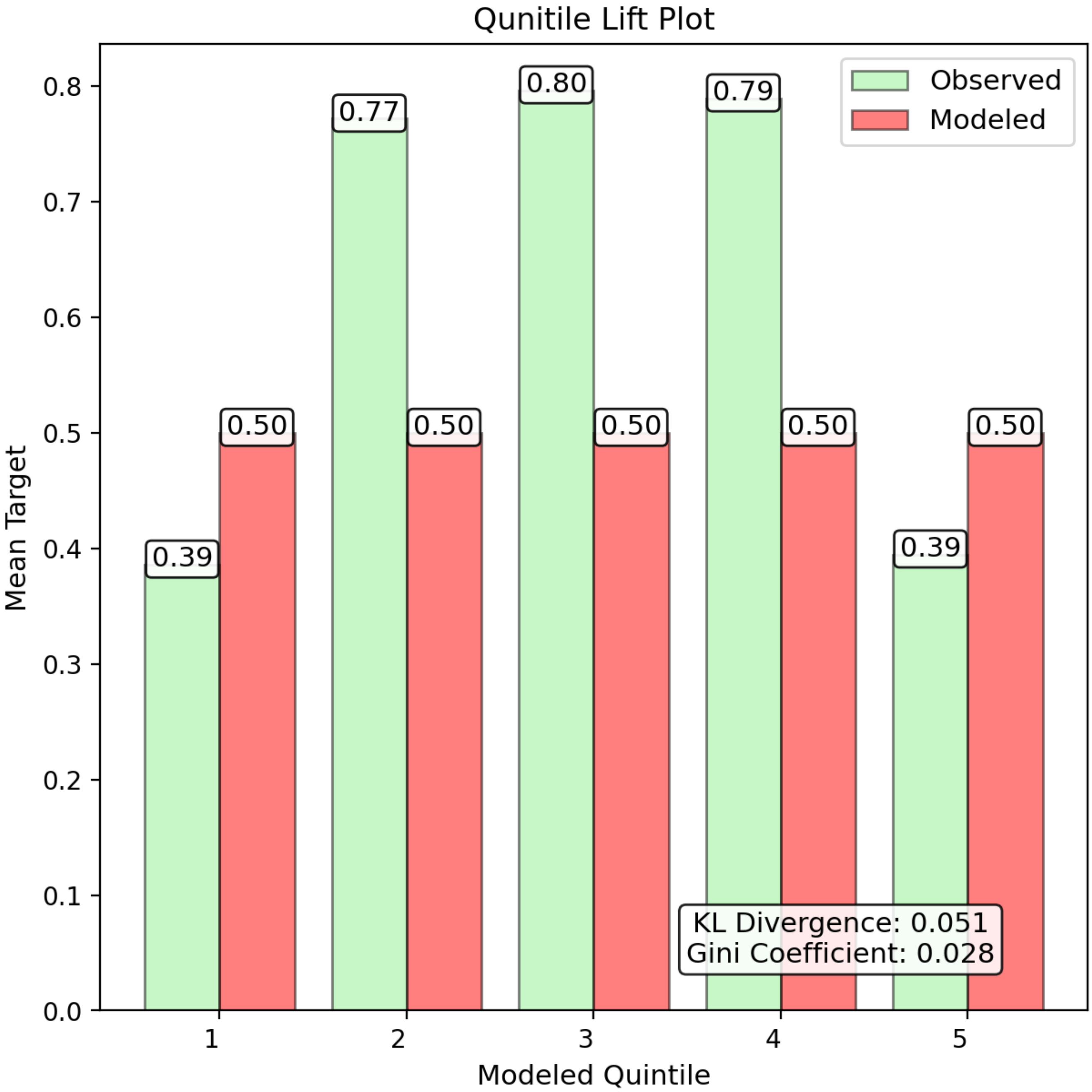
Comp 2 - ROC Curve



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

Comp 2 - 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.