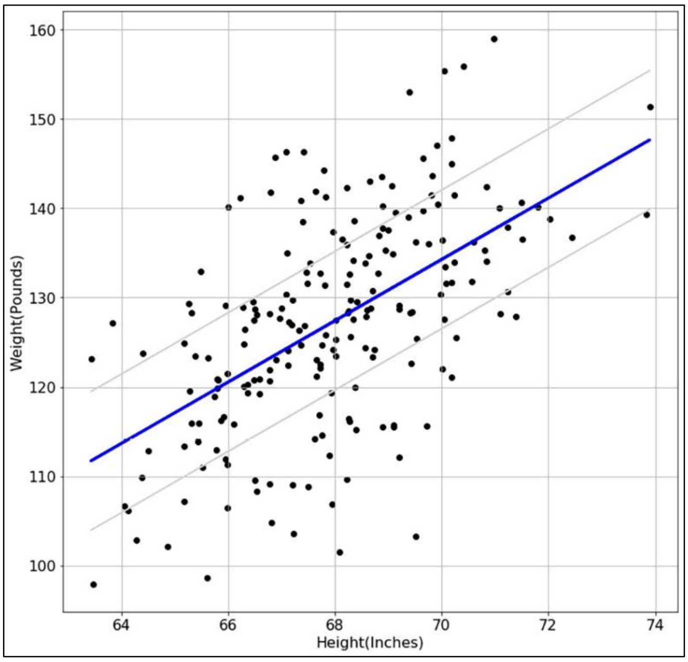
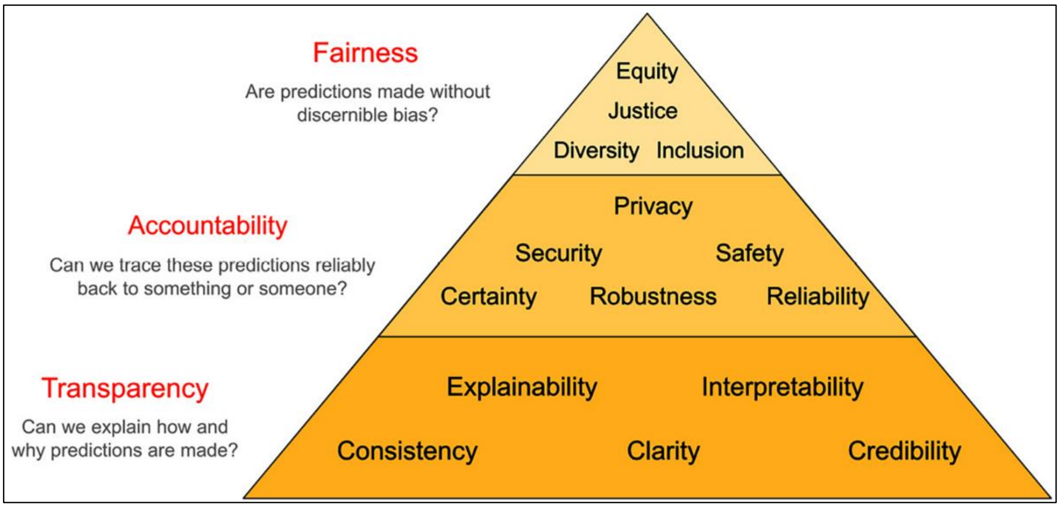
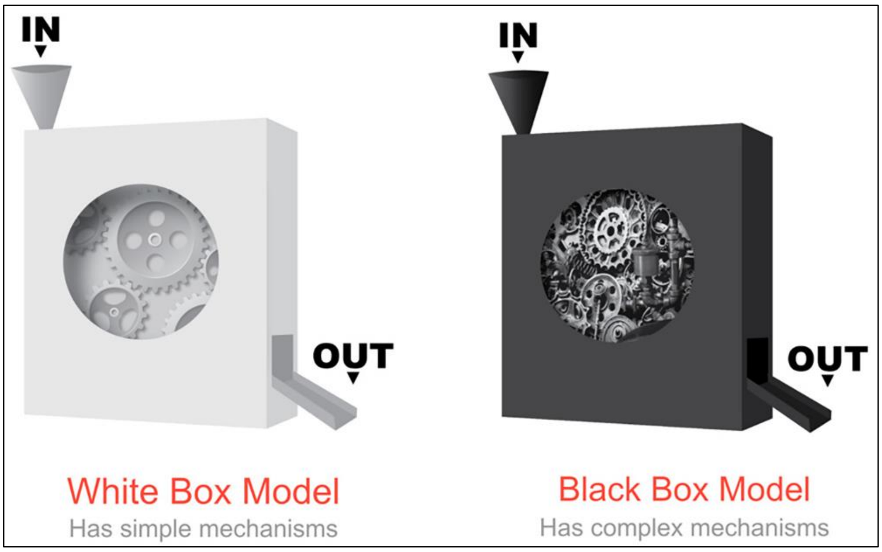
# Chapter 1: Interpretation, Interpretability, and Explainability; and Why Does It All Matter?

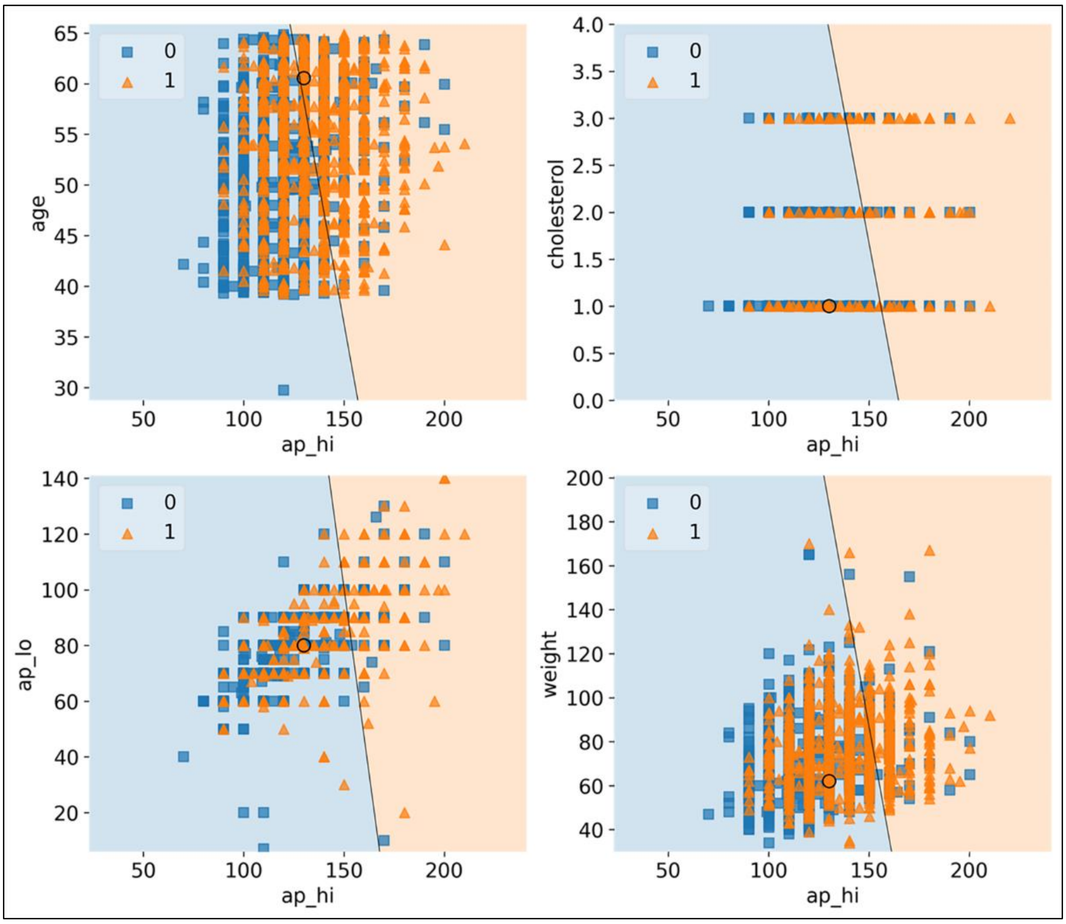


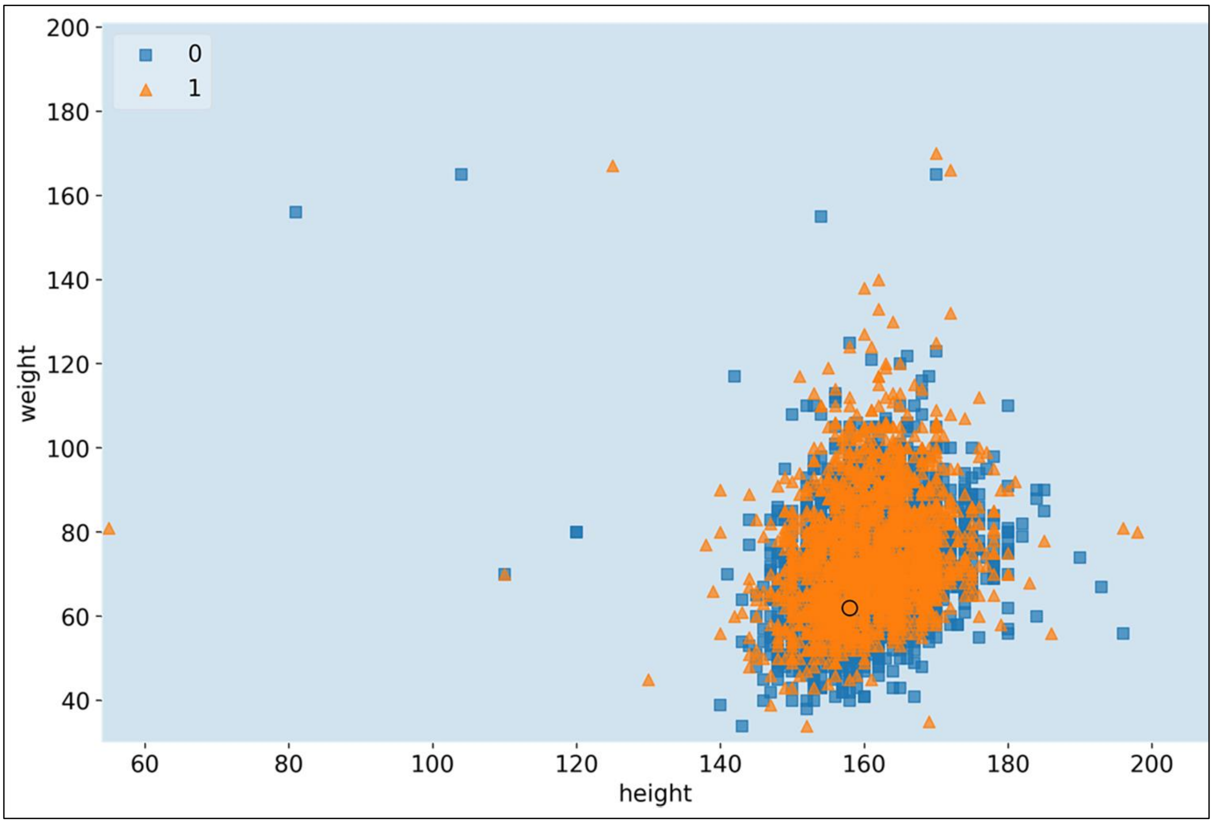



# Chapter 2: Key Concepts of Interpretability

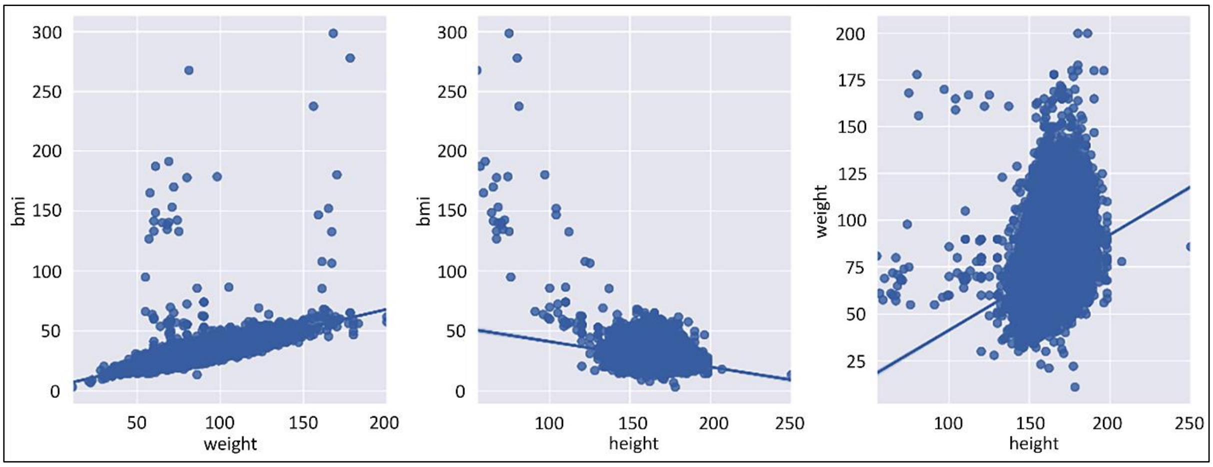
| | count | mean | std | min | 1% | 50% | 99% | max |
|---|---|---|---|---|---|---|---|---|
| **age** | 70000.00 | 53.30 | 6.76 | 29.56 | 39.61 | 53.95 | 64.31 | 64.92 |
| **gender** | 70000.00 | 1.35 | 0.48 | 1.00 | 1.00 | 1.00 | 2.00 | 2.00 |
| **height** | 70000.00 | 164.36 | 8.21 | 55.00 | 147.00 | 165.00 | 184.00 | 250.00 |
| **weight** | 70000.00 | 74.21 | 14.40 | 10.00 | 48.00 | 72.00 | 117.00 | 200.00 |
| **ap_hi** | 70000.00 | 128.82 | 154.01 | -150.00 | 90.00 | 120.00 | 180.00 | 16020.00 |
| **ap_lo** | 70000.00 | 96.63 | 188.47 | -70.00 | 60.00 | 80.00 | 1000.00 | 11000.00 |
| **cholesterol** | 70000.00 | 1.37 | 0.68 | 1.00 | 1.00 | 1.00 | 3.00 | 3.00 |
| **gluc** | 70000.00 | 1.23 | 0.57 | 1.00 | 1.00 | 1.00 | 3.00 | 3.00 |
| **smoke** | 70000.00 | 0.09 | 0.28 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| **alco** | 70000.00 | 0.05 | 0.23 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| **active** | 70000.00 | 0.80 | 0.40 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| **cardio** | 70000.00 | 0.50 | 0.50 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |

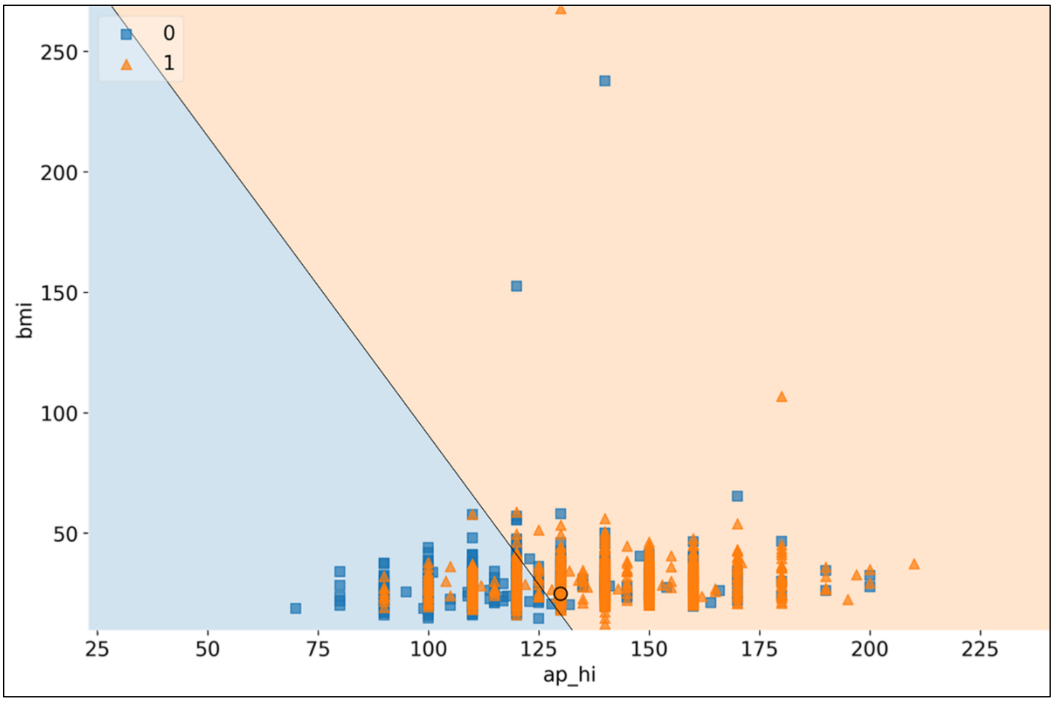


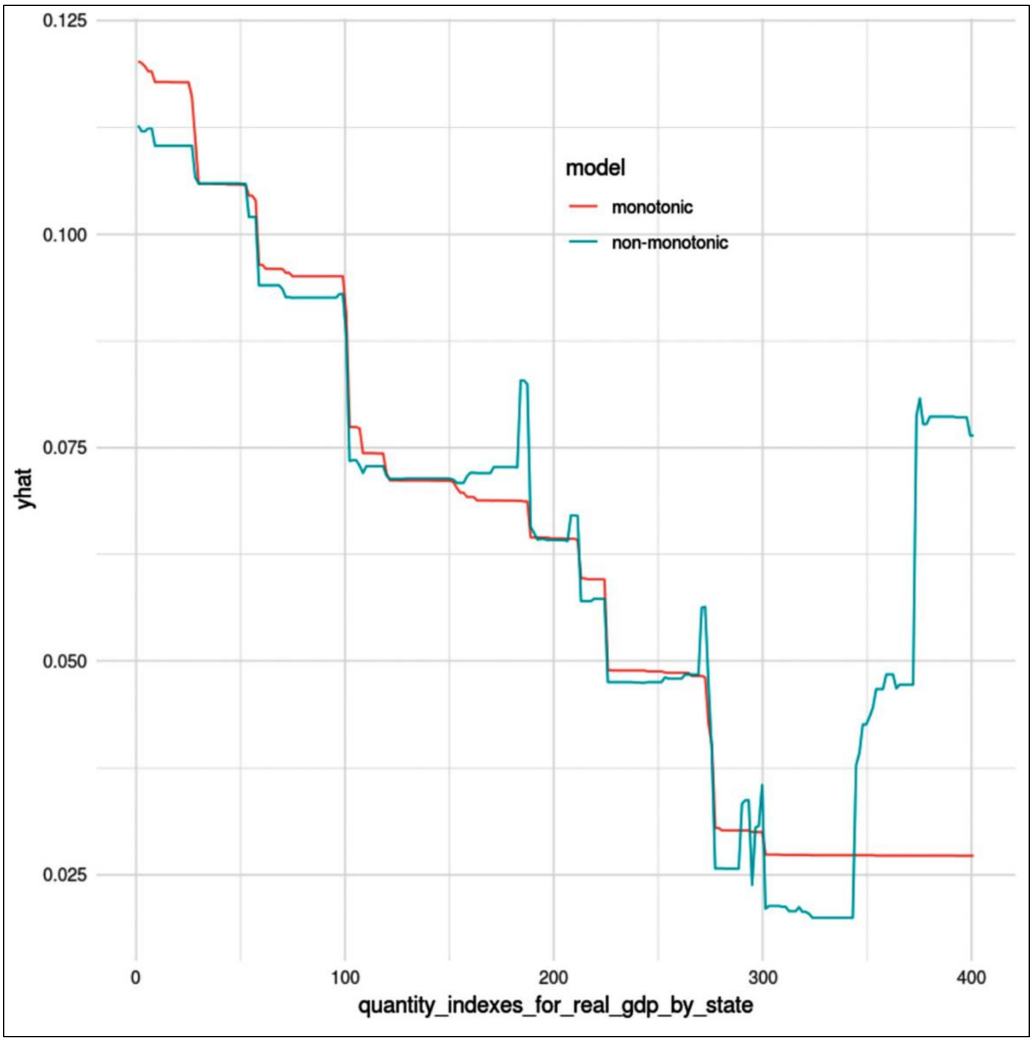
0
1
age
ap_hi
cholesterol
ap_hi
ap_lo
ap_hi
weight
ap_hi

0
1
weight
height


bmi
weight
bmi
height
weight
height


0
1
bmi
ap_hi

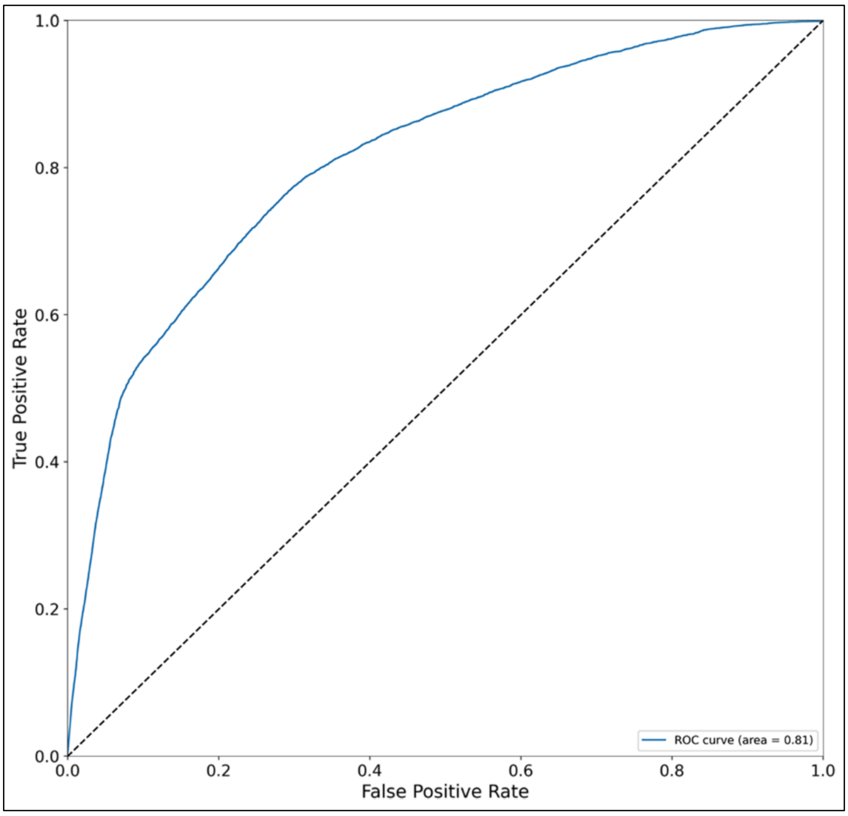


0.125
0.100
0.075
0.050
0.025
yhat
0
100
200
300
400
quantity_indexes_for_real_gdp_by_state
model
monotonic
non-monotonic

| | ARR_DELAY | CARRIER_DELAY |
|---|---|---|
| **8** | 168.000000 | 136.000000 |
| **16** | 20.000000 | 5.000000 |
| **18** | 242.000000 | 242.000000 |
| **19** | 62.000000 | 62.000000 |
| **22** | 19.000000 | 19.000000 |
| **26** | 26.000000 | 0.000000 |
| **29** | 77.000000 | 77.000000 |
| **32** | 19.000000 | 19.000000 |
| **33** | 18.000000 | 1.000000 |
| **40** | 36.000000 | 16.000000 |

| | RMSE_train | RMSE_test | R2_test |
|---|---|---|---|
| mlp | 3.24 | 3.31 | 0.987 |
| random_forest | 5.14 | 6.09 | 0.956 |
| linear_poly | 6.21 | 6.34 | 0.952 |
| linear_interact | 6.45 | 6.56 | 0.949 |
| decision_tree | 6.54 | 7.46 | 0.934 |
| linear | 7.82 | 7.88 | 0.926 |
| ridge | 7.83 | 7.90 | 0.926 |
| knn | 7.36 | 9.26 | 0.898 |
| rulefit | 9.17 | 9.31 | 0.897 |

| | Accuracy_train | Accuracy_test | Recall_train | Recall_test | ROC_AUC_test | F1_test | MCC_test |
|---|---|---|---|---|---|---|---|
| mlp | 0.998 | 0.999 | 0.987 | 0.989 | 1.000 | 0.988 | 0.987 |
| gradient_boosting | 0.992 | 0.992 | 0.893 | 0.894 | 0.999 | 0.929 | 0.926 |
| random_forest | 0.943 | 0.942 | 1.000 | 0.993 | 0.995 | 0.677 | 0.691 |
| decision_tree | 0.983 | 0.983 | 0.857 | 0.852 | 0.995 | 0.859 | 0.850 |
| logistic | 0.975 | 0.975 | 0.687 | 0.686 | 0.962 | 0.769 | 0.762 |
| knn | 0.973 | 0.965 | 0.681 | 0.608 | 0.948 | 0.681 | 0.668 |
| naive_bayes | 0.925 | 0.926 | 0.279 | 0.274 | 0.812 | 0.311 | 0.275 |
| ridge | 0.890 | 0.891 | 0.777 | 0.778 | nan | 0.467 | 0.464 |

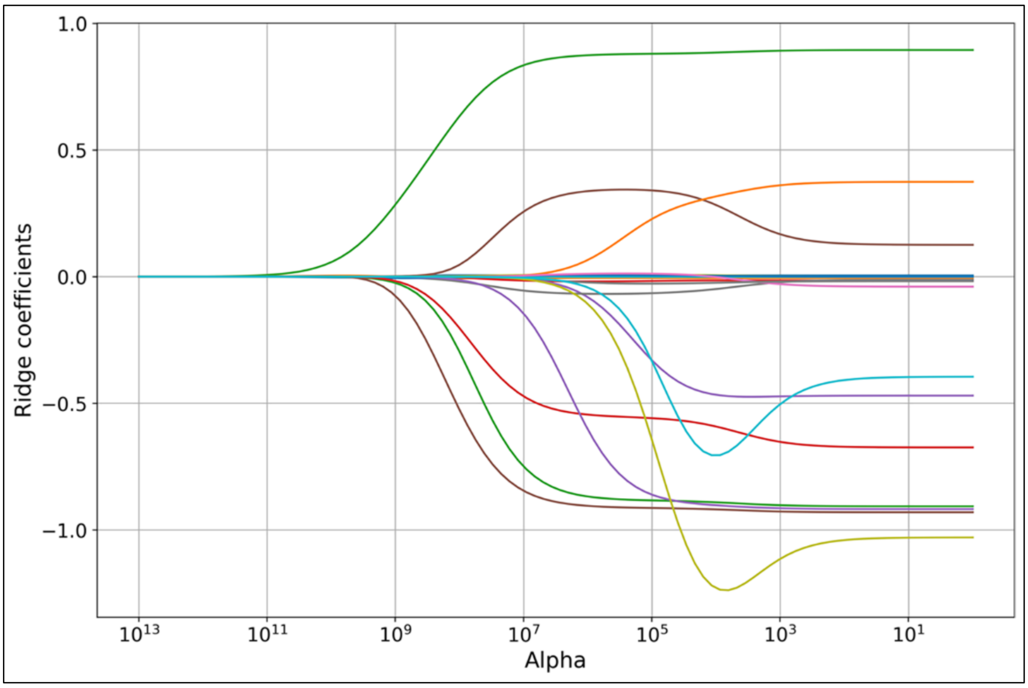


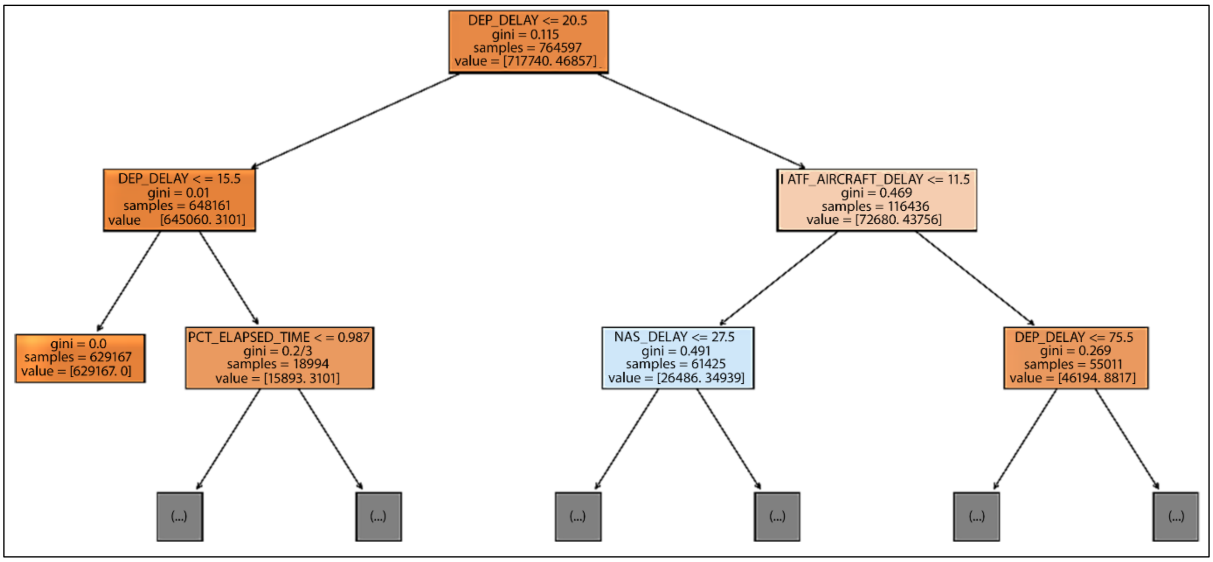
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 0.81)

| | feature | coef |
|---|---|---|
| 0 | CRS_DEP_TIME | 0.004550 |
| 1 | DEP_TIME | -0.005251 |
| 2 | DEP_DELAY | 0.894126 |
| 3 | DEP_AFPH | -0.015296 |
| 4 | DEP_RFPH | -0.469623 |
| 5 | TAXI_OUT | 0.125278 |
| 6 | WHEELS_OFF | -0.000647 |
| 7 | CRS_ELAPSED_TIME | -0.012624 |
| 8 | PCT_ELAPSED_TIME | 45.011289 |
| 9 | DISTANCE | 0.000676 |
| 10 | CRS_ARR_TIME | -0.000370 |
| 11 | ARR_AFPH | 0.000548 |
| 12 | ARR_RFPH | 0.373867 |
| 13 | WEATHER_DELAY | -0.906364 |
| 14 | NAS_DELAY | -0.674053 |
| 15 | SECURITY_DELAY | -0.917411 |
| 16 | LATE_AIRCRAFT_DELAY | -0.929844 |
| 17 | DEP_MONTH | -0.039662 |
| 18 | DEP_DOW | -0.017967 |
| 19 | ORIGIN_HUB | -1.029129 |
| 20 | DEST_HUB | -0.394935 |

| | feature | Coef. | Std.Err. | t | P>\|t\| | [0.025 | 0.975] | t_abs |
|---|---|---|---|---|---|---|---|---|
| 2 | DEP_DELAY | 0.8941 | 0.0003 | 2951.0560 | 0.0000 | 0.8935 | 0.8947 | 2951.0560 |
| 16 | LATE_AIRCRAFT_DELAY | -0.9298 | 0.0005 | -1827.0181 | 0.0000 | -0.9308 | -0.9288 | 1827.0181 |
| 13 | WEATHER_DELAY | -0.9064 | 0.0009 | -995.3664 | 0.0000 | -0.9081 | -0.9046 | 995.3664 |
| 14 | NAS_DELAY | -0.6741 | 0.0008 | -829.1287 | 0.0000 | -0.6756 | -0.6725 | 829.1287 |
| 8 | PCT_ELAPSED_TIME | 45.0113 | 0.1172 | 384.0726 | 0.0000 | 44.7816 | 45.2410 | 384.0726 |
| 15 | SECURITY_DELAY | -0.9174 | 0.0055 | -167.8571 | 0.0000 | -0.9281 | -0.9067 | 167.8571 |
| 5 | TAXI_OUT | 0.1253 | 0.0012 | 104.1196 | 0.0000 | 0.1229 | 0.1276 | 104.1196 |
| 0 | CRS_DEP_TIME | 0.0045 | 0.0001 | 62.8717 | 0.0000 | 0.0044 | 0.0047 | 62.8717 |
| 1 | DEP_TIME | -0.0053 | 0.0001 | -57.1159 | 0.0000 | -0.0054 | -0.0051 | 57.1159 |
| 3 | DEP_AFPH | -0.0153 | 0.0003 | -47.7245 | 0.0000 | -0.0159 | -0.0147 | 47.7245 |
| 19 | ORIGIN_HUB | -1.0291 | 0.0267 | -38.5894 | 0.0000 | -1.0814 | -0.9769 | 38.5894 |
| 12 | ARR_RFPH | 0.3739 | 0.0132 | 28.3860 | 0.0000 | 0.3481 | 0.3997 | 28.3860 |
| 4 | DEP_RFPH | -0.4696 | 0.0172 | -27.3532 | 0.0000 | -0.5033 | -0.4360 | 27.3532 |
| 7 | CRS_ELAPSED_TIME | -0.0126 | 0.0007 | -19.1315 | 0.0000 | -0.0139 | -0.0113 | 19.1315 |
| 10 | CRS_ARR_TIME | -0.0004 | 0.0000 | -16.9387 | 0.0000 | -0.0004 | -0.0003 | 16.9387 |
| 20 | DEST_HUB | -0.3949 | 0.0263 | -15.0415 | 0.0000 | -0.4464 | -0.3435 | 15.0415 |
| 17 | DEP_MONTH | -0.0397 | 0.0026 | -15.0188 | 0.0000 | -0.0448 | -0.0345 | 15.0188 |
| 6 | WHEELS_OFF | -0.0006 | 0.0001 | -9.6461 | 0.0000 | -0.0008 | -0.0005 | 9.6461 |
| 9 | DISTANCE | 0.0007 | 0.0001 | 8.4288 | 0.0000 | 0.0005 | 0.0008 | 8.4288 |
| 18 | DEP_DOW | -0.0180 | 0.0045 | -4.0046 | 0.0001 | -0.0268 | -0.0092 | 4.0046 |
| 11 | ARR_AFPH | 0.0005 | 0.0003 | 1.6508 | 0.0988 | -0.0001 | 0.0012 | 1.6508 |

| | feature | coef_linear | coef_ridge | coef_regularization |
|---|---|---|---|---|
| **0** | CRS_DEP_TIME | 0.004550 | 0.004496 | 0.000054 |
| **1** | DEP_TIME | -0.005251 | -0.004820 | -0.000431 |
| **2** | DEP_DELAY | 0.894126 | 0.892334 | 0.001792 |
| **3** | DEP_AFPH | -0.015296 | -0.015189 | -0.000107 |
| **4** | DEP_RFPH | -0.469623 | -0.469629 | 0.000006 |
| **5** | TAXI_OUT | 0.125278 | 0.125164 | 0.000114 |
| **6** | WHEELS_OFF | -0.000647 | 0.000013 | -0.000660 |
| **7** | CRS_ELAPSED_TIME | -0.012624 | -0.012624 | -0.000000 |
| | : | : | : | : |
| **15** | SECURITY_DELAY | -0.917411 | -0.917412 | 0.000001 |
| **16** | LATE_AIRCRAFT_DELAY | -0.929844 | -0.930708 | 0.000865 |
| **17** | DEP_MONTH | -0.039662 | -0.039664 | 0.000002 |
| **18** | DEP_DOW | -0.017967 | -0.017965 | -0.000001 |
| **19** | ORIGIN_HUB | -1.029129 | -1.029129 | -0.000000 |
| **20** | DEST_HUB | -0.394935 | -0.394935 | 0.000001 |


1.0
0.5
0.0
−0.5
−1.0
Ridge coefficients
10^13
10^11
10^9
10^7
10^5
10^3
10^1
Alpha


DEP_DELAY <= 20.5
gini = 0.115
samples = 764597
value = [717740. 46857]
DEP_DELAY < = 15.5
gini = 0.01
samples = 648161
value [645060. 3101]
I ATF_AIRCRAFT_DELAY <= 11.5
gini = 0.469
samples = 116436
value = [72680. 43756]
gini = 0.0
samples = 629167
value = [629167. 0]
PCT_ELAPSED_TIME < = 0.987
gini = 0.2/3
samples = 18994
value = [15893. 3101]
NAS_DELAY <= 27.5
gini = 0.491
samples = 61425
value = [26486. 34939]
DEP_DELAY <= 75.5
gini = 0.269
samples = 55011
value = [46194. 8817]
(...)
(...)
(...)
(...)
(...)
(...)

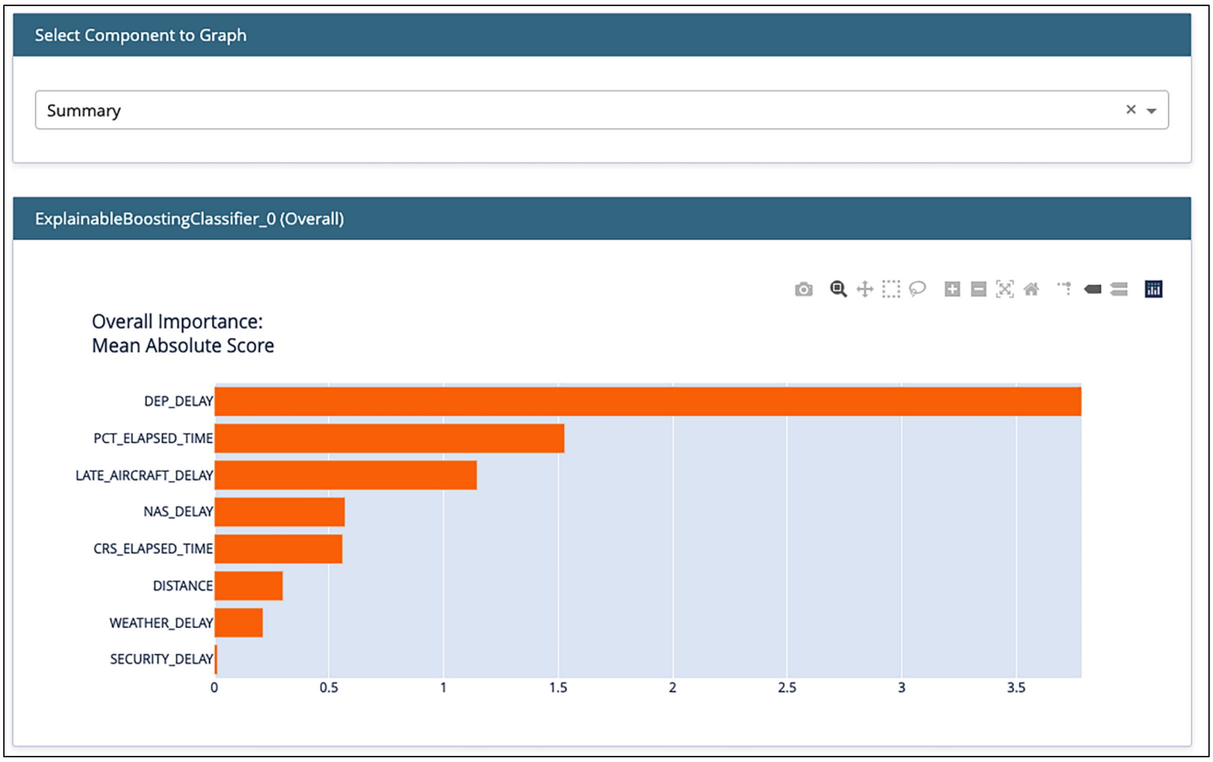
```
|--- DEP_DELAY <= 20.50
|   |--- DEP_DELAY <= 15.50
|   |   |--- class: 0
|   |--- DEP_DELAY >  15.50
|   |   |--- PCT_ELAPSED_TIME <= 0.99
|   |   |   |--- PCT_ELAPSED_TIME <= 0.98
|   |   |   |   |--- PCT_ELAPSED_TIME <= 0.96
|   |   |   |   |   |--- CRS_ELAPSED_TIME <= 65.50
|   |   |   |   |   |   |--- PCT_ELAPSED_TIME <= 0.94
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- PCT_ELAPSED_TIME >  0.94
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |--- CRS_ELAPSED_TIME >  65.50
|   |   |   |   |   |   |--- PCT_ELAPSED_TIME <= 0.95
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- PCT_ELAPSED_TIME >  0.95
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |--- PCT_ELAPSED_TIME >  0.96
|   |   |   |   |   |--- CRS_ELAPSED_TIME <= 140.50
|   |   |   |   |   |   |--- DEP_DELAY <= 18.50
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- DEP_DELAY >  18.50
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |--- CRS_ELAPSED_TIME >  140.50
|   |   |   |   |   |   |--- DEP_DELAY <= 19.50
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- DEP_DELAY >  19.50
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |--- PCT_ELAPSED_TIME >  0.98
|   |   |   |   |--- DEP_DELAY <= 18.50
|   |   |   |   |   |--- DISTANCE <= 326.50
|   |   |   |   |   |   |--- LATE_AIRCRAFT_DELAY <= 0.50
|   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |--- LATE_AIRCRAFT_DELAY >  0.50
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |--- … (goes on for 6 more pages!)
```

| | feature | importance |
|---|---|---|
| **2** | DEP_DELAY | 0.527482 |
| **16** | LATE_AIRCRAFT_DELAY | 0.199153 |
| **8** | PCT_ELAPSED_TIME | 0.105381 |
| **13** | WEATHER_DELAY | 0.101649 |
| **14** | NAS_DELAY | 0.062732 |
| **15** | SECURITY_DELAY | 0.001998 |
| **9** | DISTANCE | 0.001019 |
| **7** | CRS_ELAPSED_TIME | 0.000281 |
| | : | : |
| **4** | DEP_RFPH | 0.000000 |
| **20** | DEST_HUB | 0.000000 |

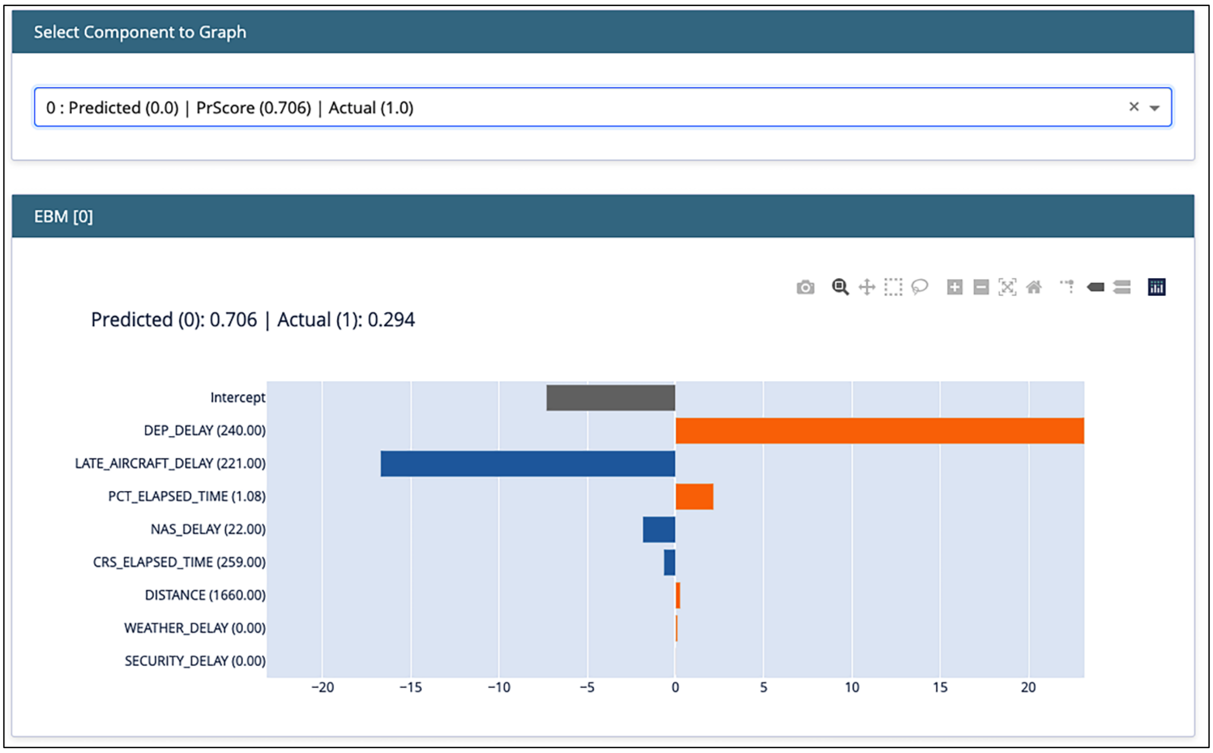
| | rule | type | coef | support | importance |
|---|---|---|---|---|---|
| 101 | LATE_AIRCRAFT_DELAY <= 222.5 & DEP_DELAY > 344.0 & WEATHER_DELAY <= 166.0 | rule | 222.024721 | 0.001684 | 9.102113 |
| 42 | LATE_AIRCRAFT_DELAY <= 333.5 & DEP_DELAY > 477.5 | rule | 172.103034 | 0.001122 | 5.762432 |
| 16 | LATE_AIRCRAFT_DELAY | linear | -0.386073 | 1.000000 | 4.523663 |
| 2 | DEP_DELAY | linear | 0.163704 | 1.000000 | 4.282909 |
| 64 | DEP_DELAY > 1206.0 | rule | 278.817372 | 0.000187 | 3.812982 |
| 142 | LATE_AIRCRAFT_DELAY <= 198.0 & DEP_DELAY > 341.5 & DEP_DELAY <= 788.0 | rule | -92.790467 | 0.001496 | 3.586813 |
| 134 | DEP_DELAY > 300.0 & LATE_AIRCRAFT_DELAY <= 158.5 & DEP_DELAY > 576.5 | rule | 115.440190 | 0.000748 | 3.156531 |
| 23 | DEP_DELAY > 66.5 & NAS_DELAY > 43.5 & LATE_AIRCRAFT_DELAY <= 19.5 & DEP_DELAY <= 849.0 | rule | -41.899504 | 0.004302 | 2.742345 |
| | : | : | : | : | : |
| 18 | DEP_DOW | linear | 0.009907 | 1.000000 | 0.019798 |
| 45 | DEP_DELAY <= 66.5 & DEP_DELAY <= 20.5 & DEP_DELAY <= 849.0 | rule | -0.042437 | 0.847924 | 0.015239 |
| 170 | DEP_DELAY <= 880.5 | rule | -0.269029 | 0.999252 | 0.007356 |

| White Box? | Model Class | Properties that Increase Interpretability | | | | | Task | | Performance Rank | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Expl. | Linear | Monotone | Non-Interactive | Regul. | Regr. | Classif. | Regr. | Classif. |
| ✓ | Linear Regression | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | 6 | |
| ✓ | Regularized Regression | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 7 | 8 |
| ✓ | Logistic Regression | ✓ | ! | ✓ | ✓ | ✓ | ✗ | ✓ | | 5 |
| ✓ | Gaussian Naïve Bayes | ✓ | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ | | 7 |
| ✓ | Polynomial Regression | ! | ! | ✓ | ! | ✓ | ✓ | ✓ | 2 | |
| ✓ | RuleFit | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | 8 | |
| ✓ | Decision Tree | ✓ | ✗ | ! | ✗ | ✓ | ✓ | ✓ | 5 | 3 |
| ✓ | k-Nearest Neighbors | ! | ✗ | ✗ | ✓ | ✗ | ✓ | ✓ | 9 | 6 |
| ✗ | Random Forest | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | 3 | 4 |
| ✗ | Gradient Boosted Trees | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | | 2 |
| ✗ | Multi-layer Perceptron | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | 1 | 1 |

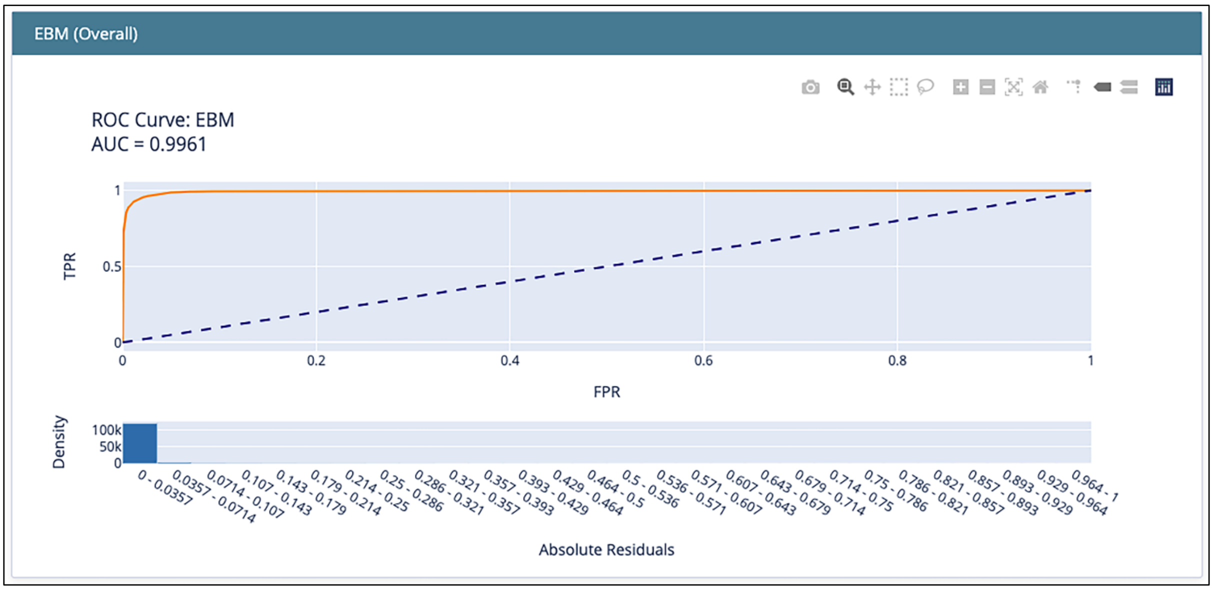
| | White Box | Glass Box | Black Box |
|---|---|---|---|
| **Inherent Interpretability** | High | Mid-High | Low |
| **Predictive Performance** | Mid | High | High |
| **Execution Speed Performance** | High | Low | Mid |

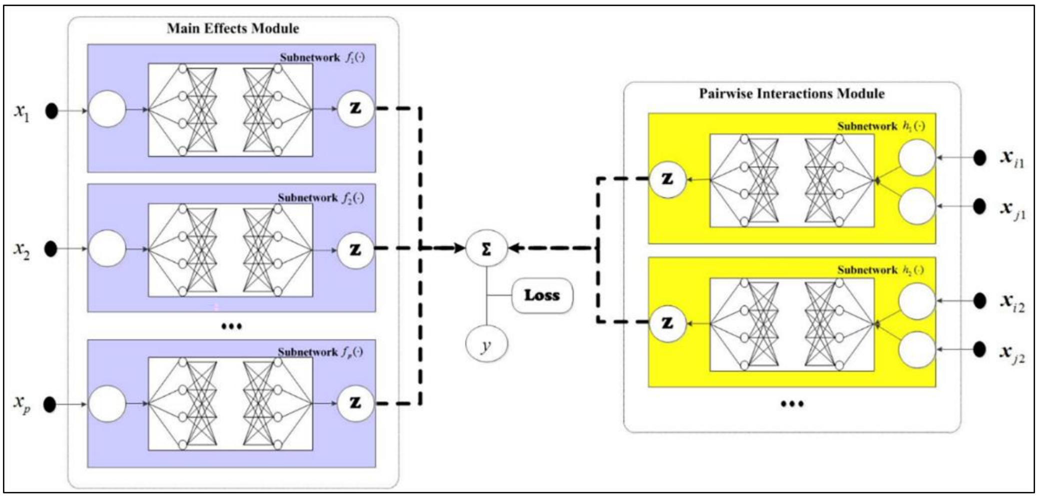


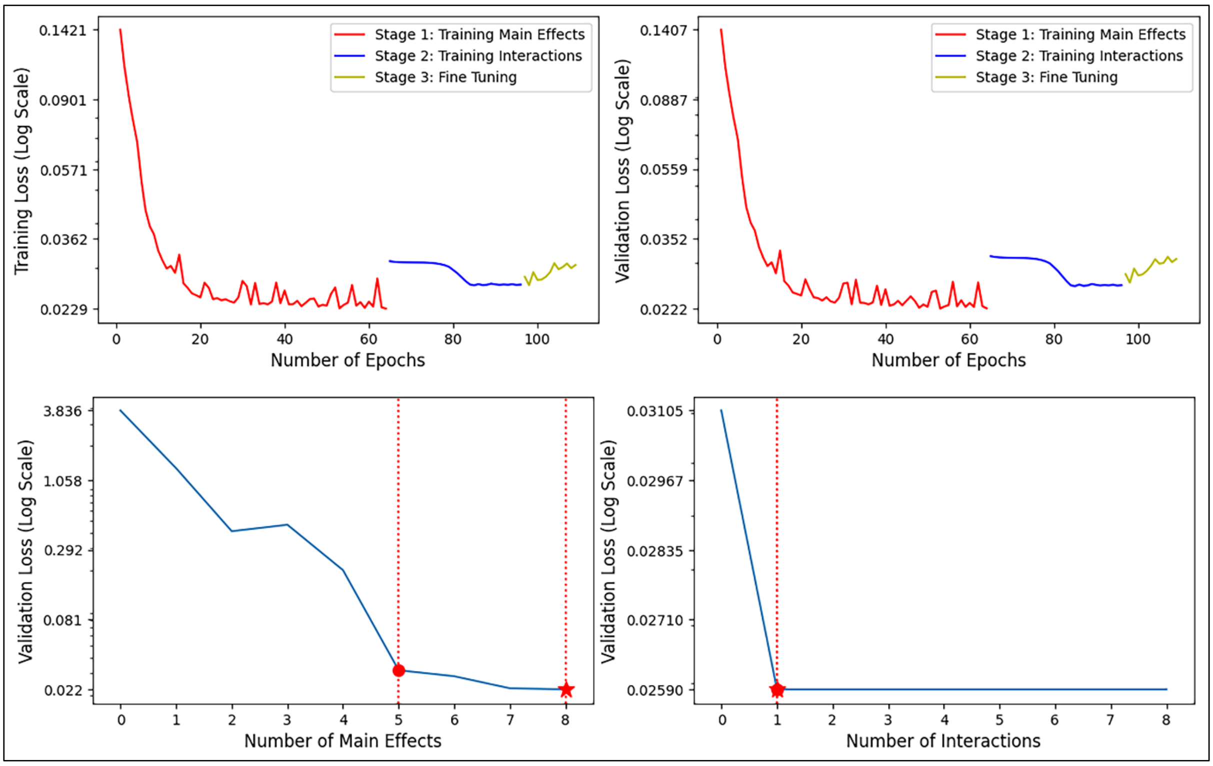
Select Component to Graph
Summary
ExplainableBoostingClassifier_0 (Overall)
Overall Importance:
Mean Absolute Score
DEP_DELAY
PCT_ELAPSED_TIME
LATE_AIRCRAFT_DELAY
NAS_DELAY
CRS_ELAPSED_TIME
DISTANCE
WEATHER_DELAY
SECURITY_DELAY
0
0.5
1
1.5
2
2.5
3
3.5

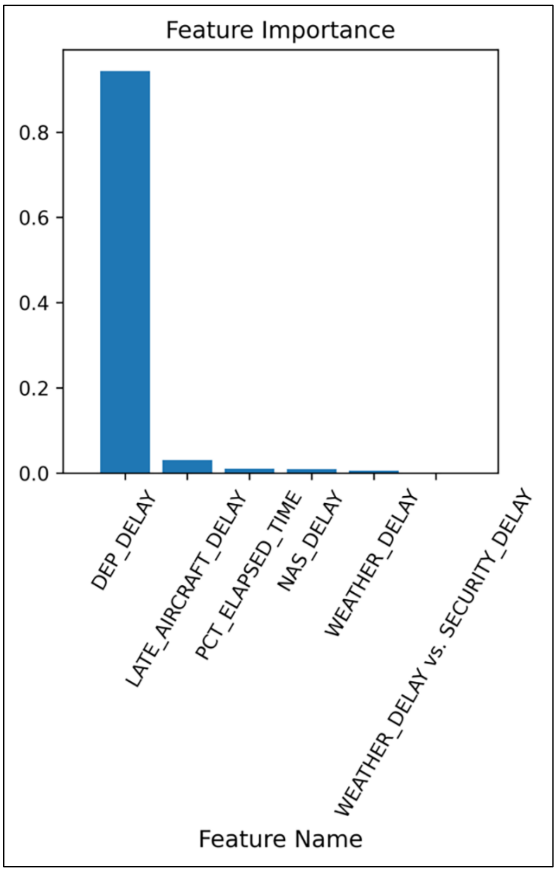
Select Component to Graph
0 : Predicted (0.0) | PrScore (0.706) | Actual (1.0)
EBM [0]
Predicted (0): 0.706 | Actual (1): 0.294
Intercept
DEP_DELAY (240.00)
LATE_AIRCRAFT_DELAY (221.00)
PCT_ELAPSED_TIME (1.08)
NAS_DELAY (22.00)
CRS_ELAPSED_TIME (259.00)
DISTANCE (1660.00)
WEATHER_DELAY (0.00)
SECURITY_DELAY (0.00)
−20
−15
−10
−5
0
5
10
15
20


EBM (Overall)
ROC Curve: EBM
AUC = 0.9961
TPR
FPR
Density
Absolute Residuals

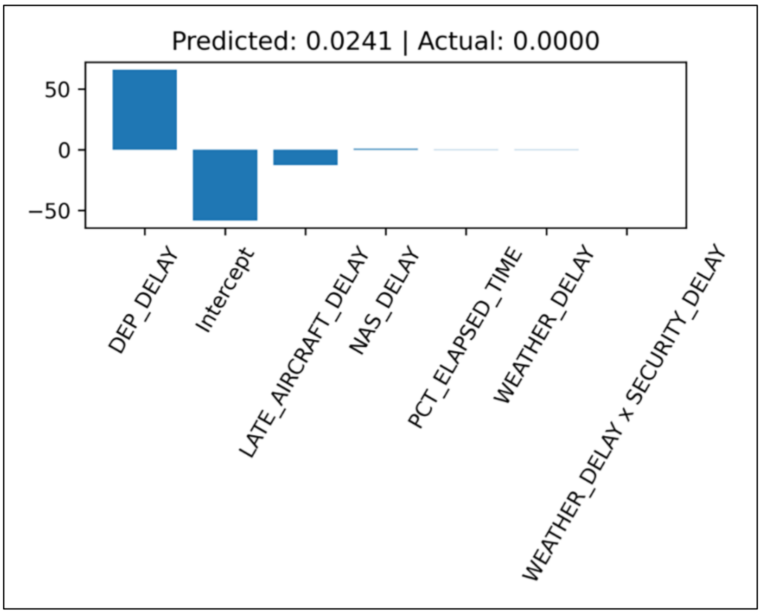

Main Effects Module
Subnetwork $f_1(\cdot)$
Subnetwork $f_2(\cdot)$
Subnetwork $f_p(\cdot)$
Pairwise Interactions Module
Subnetwork $h_1(\cdot)$
Subnetwork $h_2(\cdot)$
Loss
$x_1$
$x_2$
$x_p$
$x_{i1}$
$x_{j1}$
$x_{i2}$
$x_{j2}$
$y$



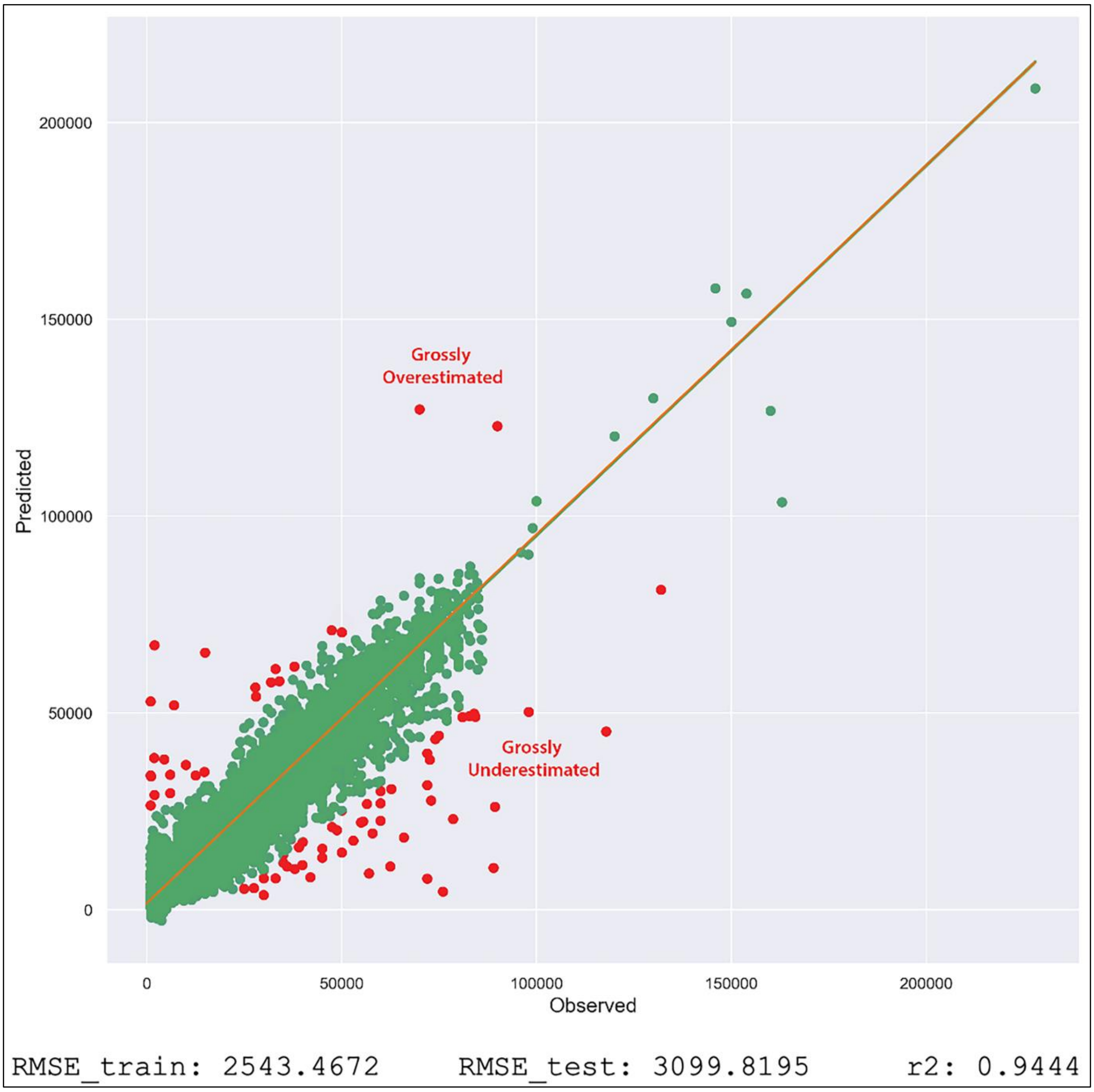
Stage 1: Training Main Effects
Stage 2: Training Interactions
Stage 3: Fine Tuning
Training Loss (Log Scale)
Validation Loss (Log Scale)
Number of Epochs
Number of Main Effects
Number of Interactions

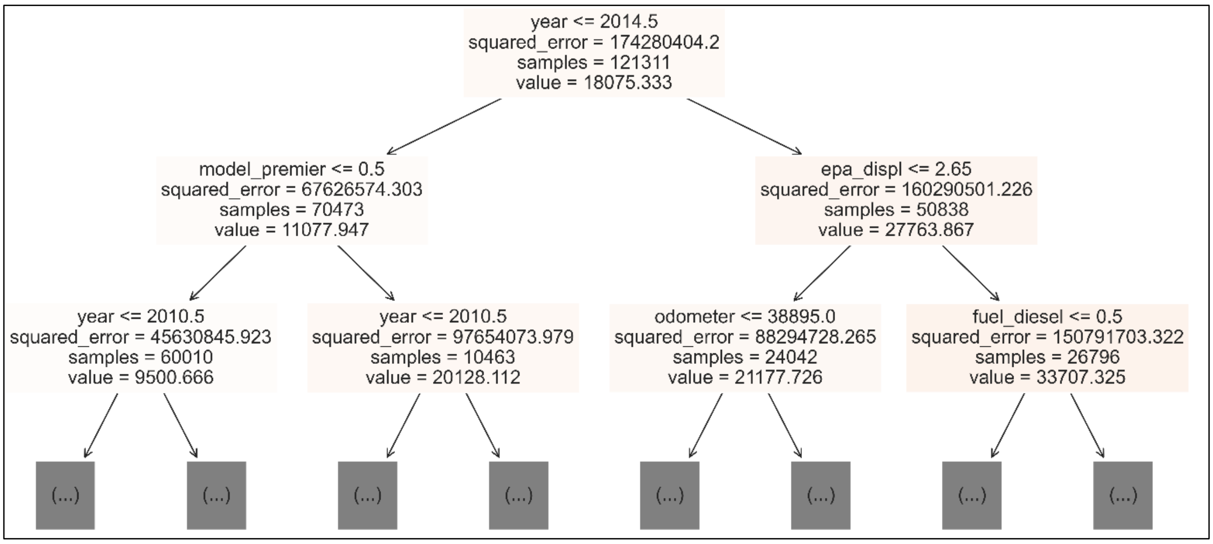



Feature Importance
DEP_DELAY
LATE_AIRCRAFT_DELAY
PCT_ELAPSED_TIME
NAS_DELAY
WEATHER_DELAY
WEATHER_DELAY vs. SECURITY_DELAY
Feature Name


Predicted: 0.0241 | Actual: 0.0000
50
0
−50
DEP_DELAY
Intercept
LATE_AIRCRAFT_DELAY
NAS_DELAY
PCT_ELAPSED_TIME
WEATHER_DELAY
WEATHER_DELAY x SECURITY_DELAY

# Chapter 4: Global Model-Agnostic Interpretation Methods

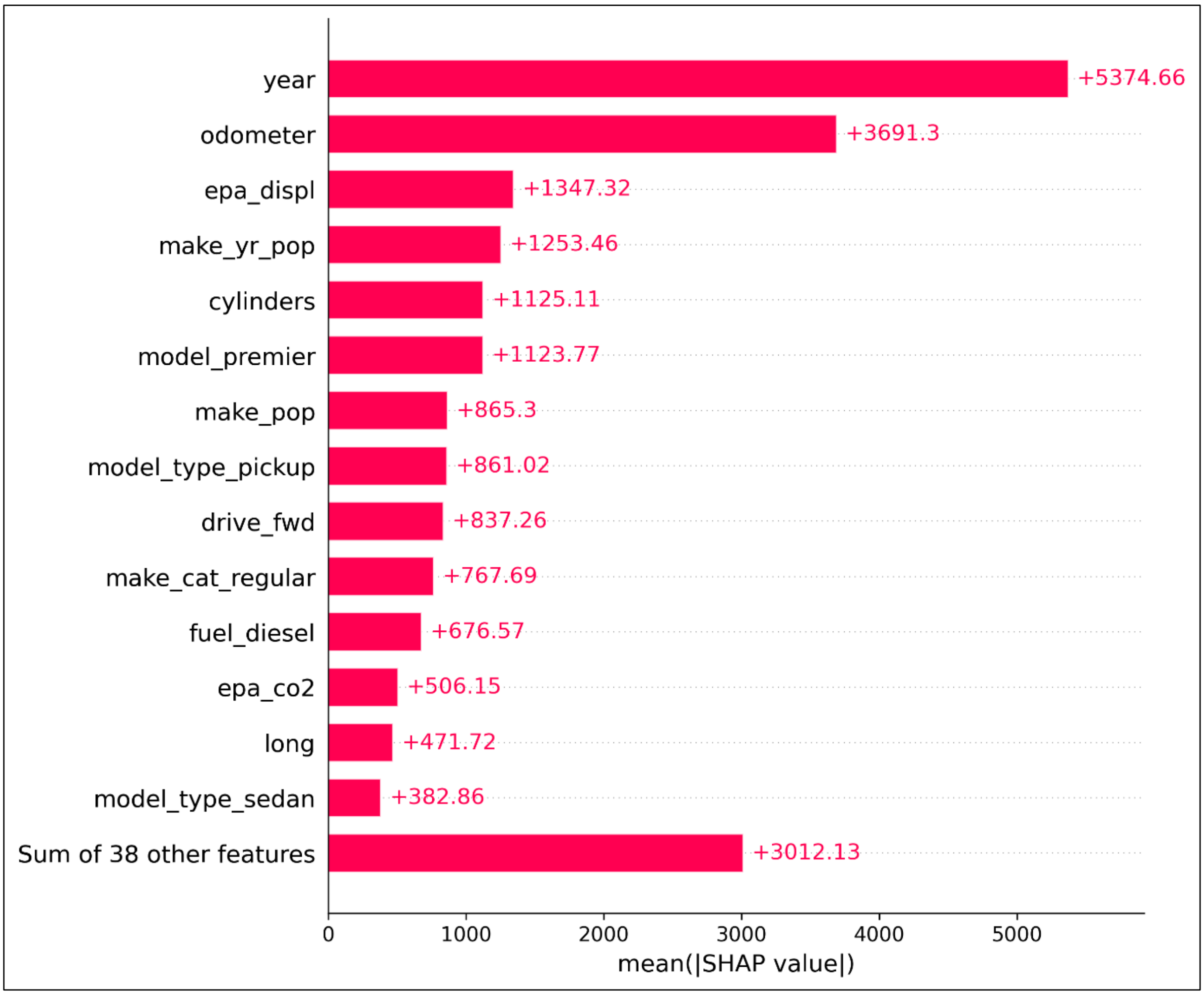



year <= 2014.5
squared_error = 174280404.2
samples = 121311
value = 18075.333
model_premier <= 0.5
squared_error = 67626574.303
samples = 70473
value = 11077.947
epa_displ <= 2.65
squared_error = 160290501.226
samples = 50838
value = 27763.867
year <= 2010.5
squared_error = 45630845.923
samples = 60010
value = 9500.666
year <= 2010.5
squared_error = 97654073.979
samples = 10463
value = 20128.112
odometer <= 38895.0
squared_error = 88294728.265
samples = 24042
value = 21177.726
fuel_diesel <= 0.5
squared_error = 150791703.322
samples = 26796
value = 33707.325
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

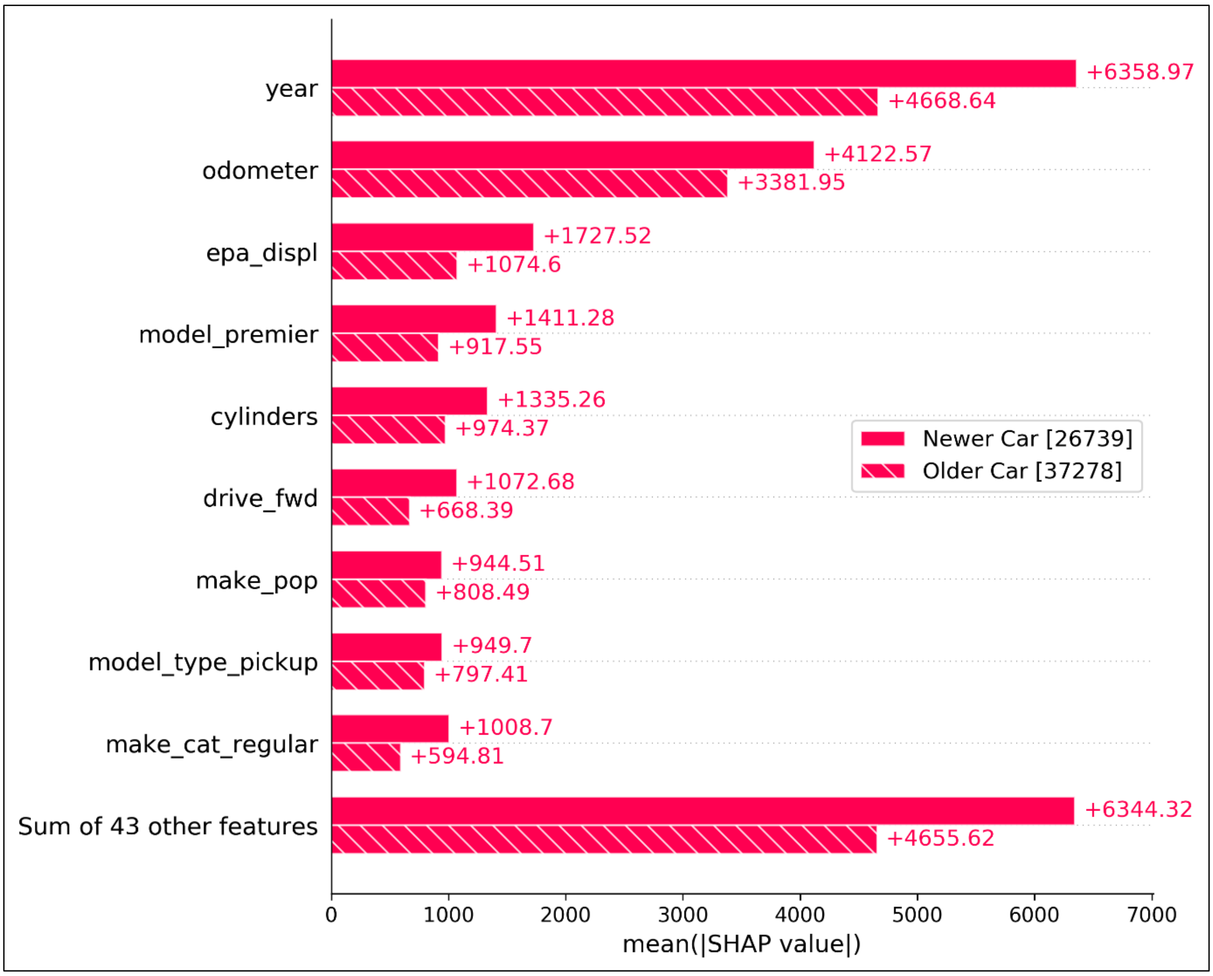
| | feature | cb_feat_imp | rf_feat_imp |
|---|---|---|---|
| 4 | year | 23.24% | 45.99% |
| 7 | odometer | 13.05% | 9.41% |
| 11 | epa_displ | 9.95% | 11.79% |
| 15 | fuel_diesel | 6.16% | 3.74% |
| 2 | make_pop | 5.11% | 2.41% |
| 3 | model_premier | 5.06% | 5.91% |
| 9 | cylinders | 5.06% | 2.45% |
| 10 | epa_co2 | 4.03% | 1.53% |
| 31 | model_type_pickup | 3.41% | 1.86% |
| 24 | make_cat_regular | 3.20% | 1.86% |
| 44 | drive_fwd | 2.53% | 0.58% |
| 5 | make_yr_pop | 2.49% | 2.36% |
| 6 | model_yr_pop | 2.34% | 1.73% |
| 38 | condition_good | 1.93% | 0.80% |
| 1 | long | 1.47% | 1.05% |
| 0 | lat | 1.23% | 0.83% |
| | : | : | : |
| | title_status_other | 0.00% | 0.00% |

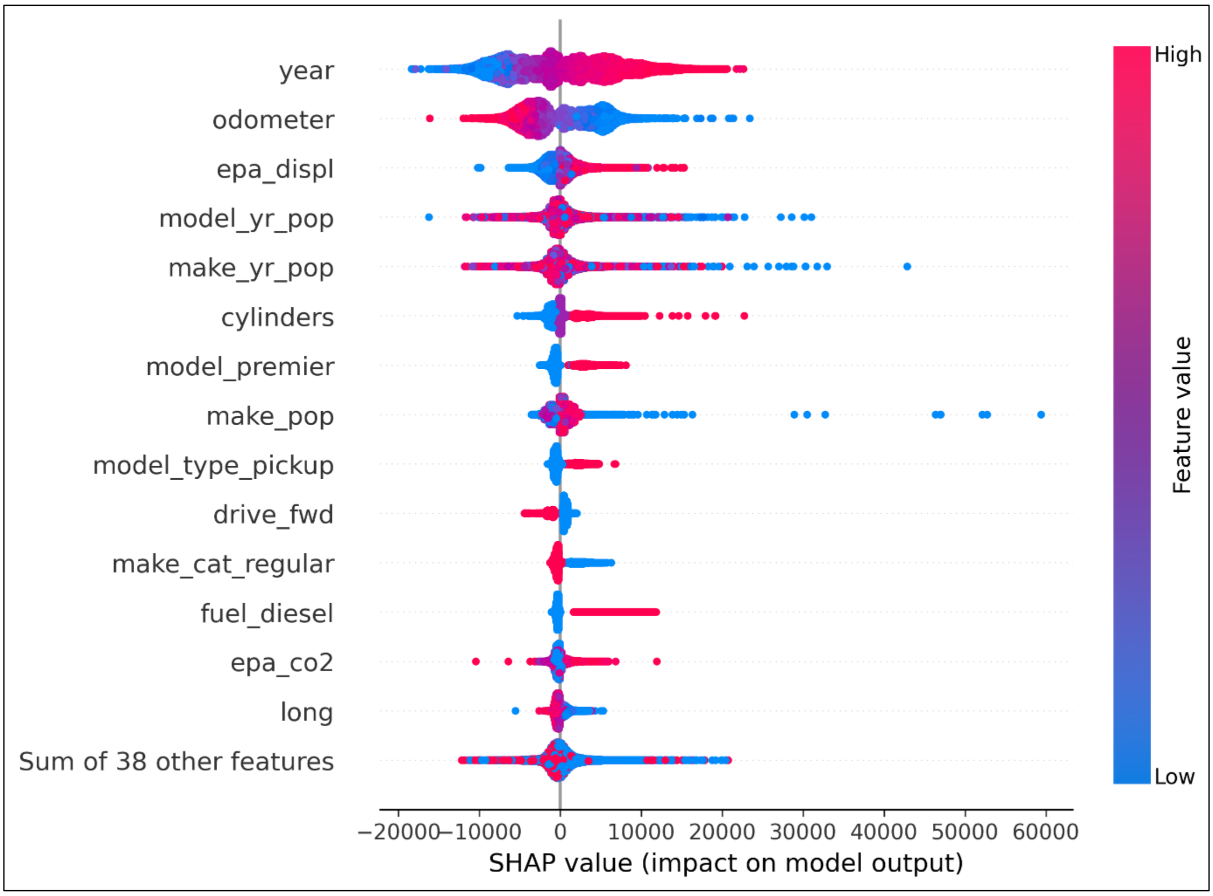
| | feature | cb_perm_mean | cb_perm_std | rf_perm_mean | rf_perm_std |
|---|---|---|---|---|---|
| 4 | year | 5424.1815 | 15.1548 | 6330.7544 | 19.9509 |
| 7 | odometer | 3014.9332 | 11.8530 | 3288.1512 | 13.0719 |
| 11 | epa_displ | 1330.8259 | 7.5498 | 2178.9189 | 10.0180 |
| 9 | cylinders | 1062.5487 | 9.9618 | 1229.9463 | 10.7112 |
| 2 | make_pop | 997.8885 | 7.0639 | 522.7714 | 4.7237 |
| 3 | model_premier | 578.3021 | 2.6501 | 1129.9980 | 7.5607 |
| 10 | epa_co2 | 555.8032 | 2.9135 | 719.1720 | 5.2183 |
| 31 | model_type_pickup | 544.0469 | 3.3891 | 533.8197 | 6.0525 |
| 5 | make_yr_pop | 532.4376 | 5.3494 | 484.2530 | 2.7652 |
| 24 | make_cat_regular | 486.2884 | 2.6164 | 709.5006 | 4.9734 |
| 15 | fuel_diesel | 349.7755 | 2.8628 | 519.0072 | 7.3040 |
| 6 | model_yr_pop | 320.0405 | 2.3302 | 531.8455 | 4.1025 |
| 38 | condition_good | 249.2664 | 2.2921 | 264.4274 | 3.9697 |
| 44 | drive_fwd | 245.5661 | 3.0388 | 304.5629 | 4.0389 |
| 1 | long | 207.2723 | 2.7572 | 285.6865 | 2.4887 |
| 32 | model_type_sedan | 165.2068 | 2.7577 | 208.8010 | 3.5156 |
| 43 | drive_4wd | 140.6390 | 1.6213 | 230.8113 | 3.9377 |
| 0 | lat | 139.9203 | 1.6919 | 147.5605 | 1.4378 |
| 20 | make_cat_luxury | 135.7519 | 1.9622 | 108.5608 | 1.5728 |
| 25 | model_type_SUV | 88.4956 | 1.4549 | 115.0064 | 2.1095 |
| | : | : | : | : | : |
| 40 | condition_new | 0.2814 | 0.0831 | 0.1344 | 0.0455 |

| feature | cb_shap_imp | rf_shap_imp |
|---:|---:|---:|
| year | 5374.6635 | 5704.8741 |
| odometer | 3691.2966 | 2837.3699 |
| epa_displ | 1347.3157 | 1935.6400 |
| cylinders | 1125.1109 | 636.3444 |
| model_premier | 1123.7732 | 534.7291 |
| make_pop | 865.3036 | 376.7030 |
| model_type_pickup | 861.0232 | 532.9261 |
| drive_fwd | 837.2566 | 313.6196 |
| make_cat_regular | 767.6872 | 348.1303 |
| fuel_diesel | 676.5713 | 332.9367 |
| epa_co2 | 506.1507 | 392.3129 |
| long | 471.7182 | 319.1998 |
| model_type_sedan | 382.8636 | 300.3917 |
| model_yr_pop | 324.9925 | 414.3527 |
| make_yr_pop | 311.5293 | 280.4556 |
| condition_good | 280.9659 | 132.0956 |
| lat | 224.7691 | 108.0771 |
| make_cat_luxury | 204.4589 | 52.7512 |
| drive_4wd | 186.8555 | 134.9422 |
| zip_density | 140.3340 | 80.8189 |
| : | : | : |
| condition_new | 1.0416 | 0.0869 |

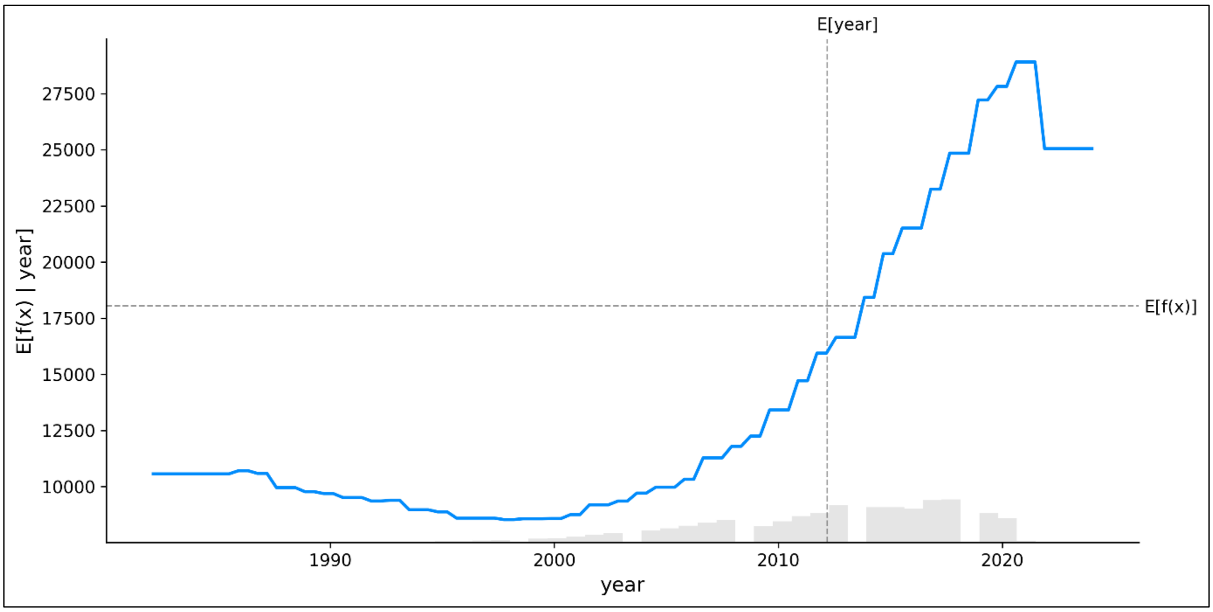


year
odometer
epa_displ
make_yr_pop
cylinders
model_premier
make_pop
model_type_pickup
drive_fwd
make_cat_regular
fuel_diesel
epa_co2
long
model_type_sedan
Sum of 38 other features
+5374.66
+3691.3
+1347.32
+1253.46
+1125.11
+1123.77
+865.3
+861.02
+837.26
+767.69
+676.57
+506.15
+471.72
+382.86
+3012.13
0
1000
2000
3000
4000
5000
mean(|SHAP value|)

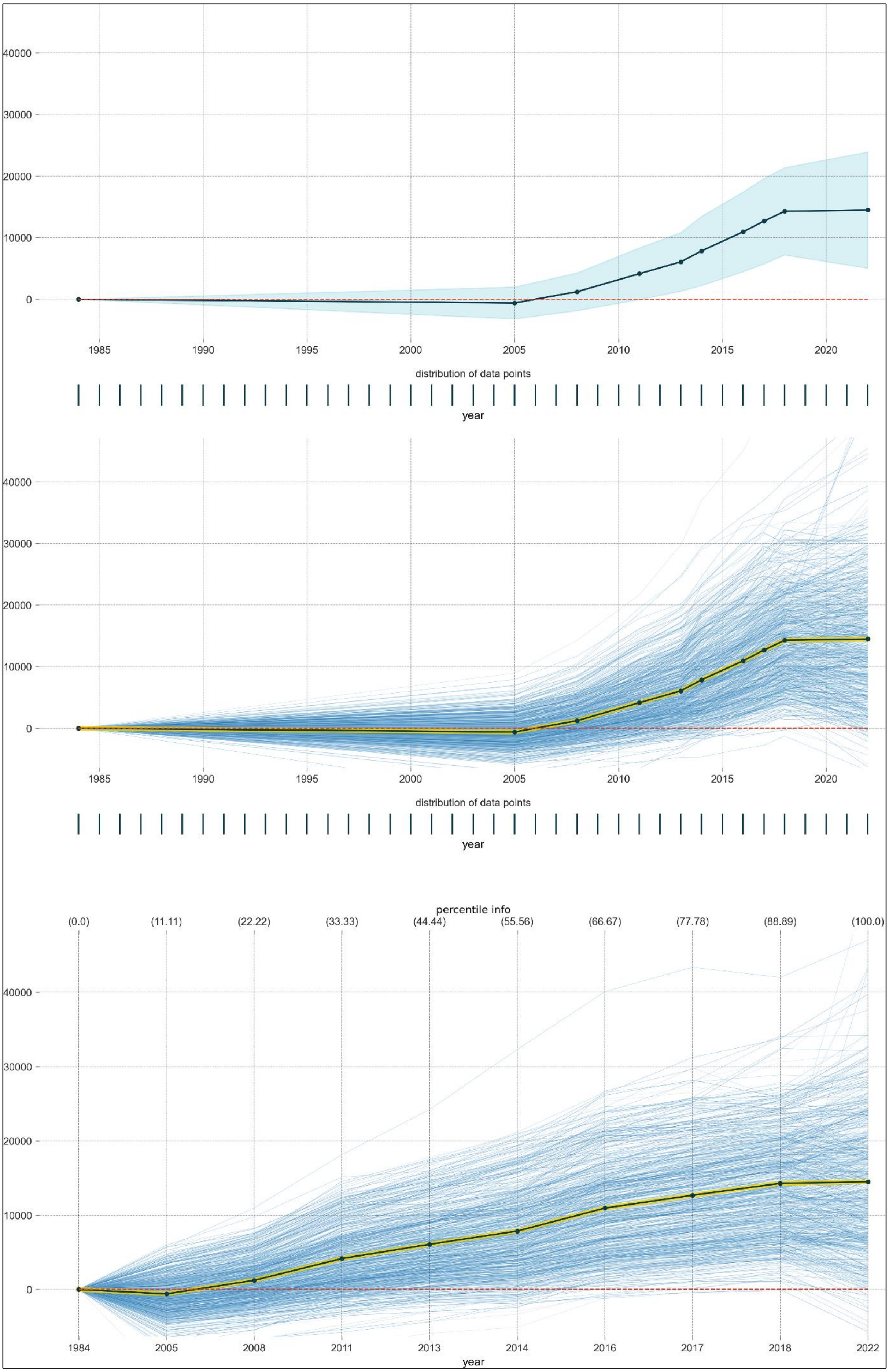

year
odometer
epa_displ
model_premier
cylinders
drive_fwd
make_pop
model_type_pickup
make_cat_regular
Sum of 43 other features
+6358.97
+4668.64
+4122.57
+3381.95
+1727.52
+1074.6
+1411.28
+917.55
+1335.26
+974.37
+1072.68
+668.39
+944.51
+808.49
+949.7
+797.41
+1008.7
+594.81
+6344.32
+4655.62
Newer Car [26739]
Older Car [37278]
0
1000
2000
3000
4000
5000
6000
7000
mean(|SHAP value|)



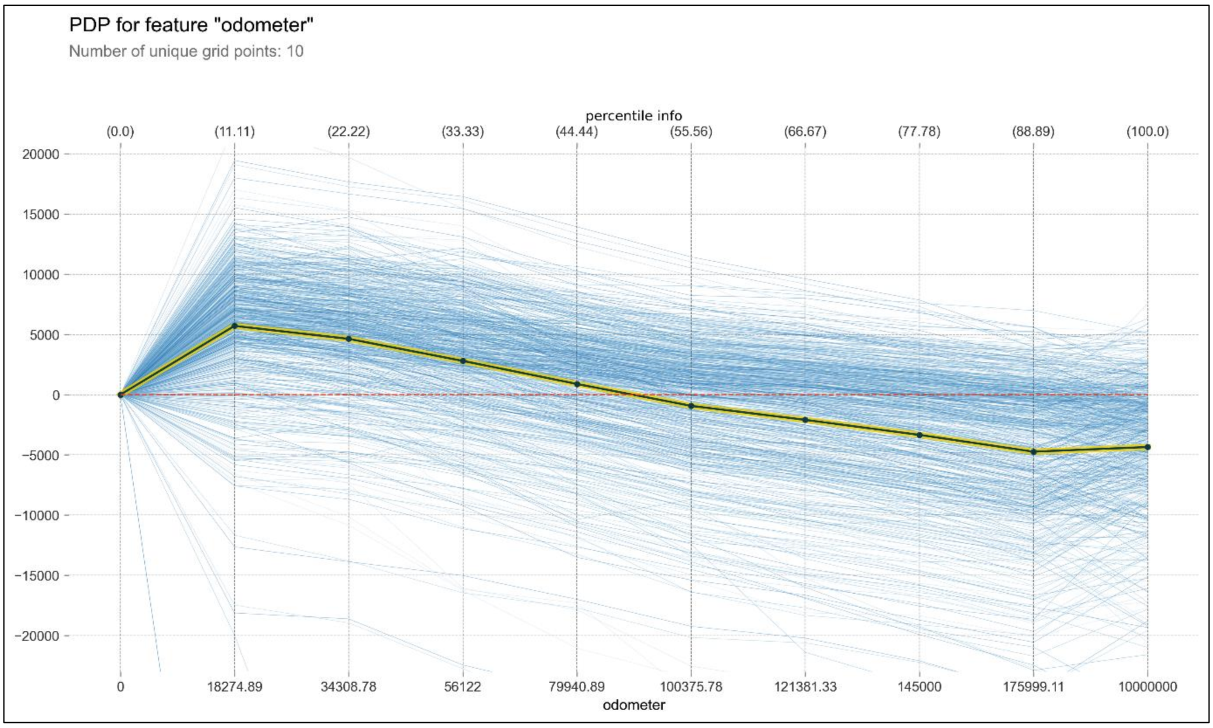
year
odometer
epa_displ
model_yr_pop
make_yr_pop
cylinders
model_premier
make_pop
model_type_pickup
drive_fwd
make_cat_regular
fuel_diesel
epa_co2
long
Sum of 38 other features
−20000 −10000 0 10000 20000 30000 40000 50000 60000
SHAP value (impact on model output)
High
Feature value
Low

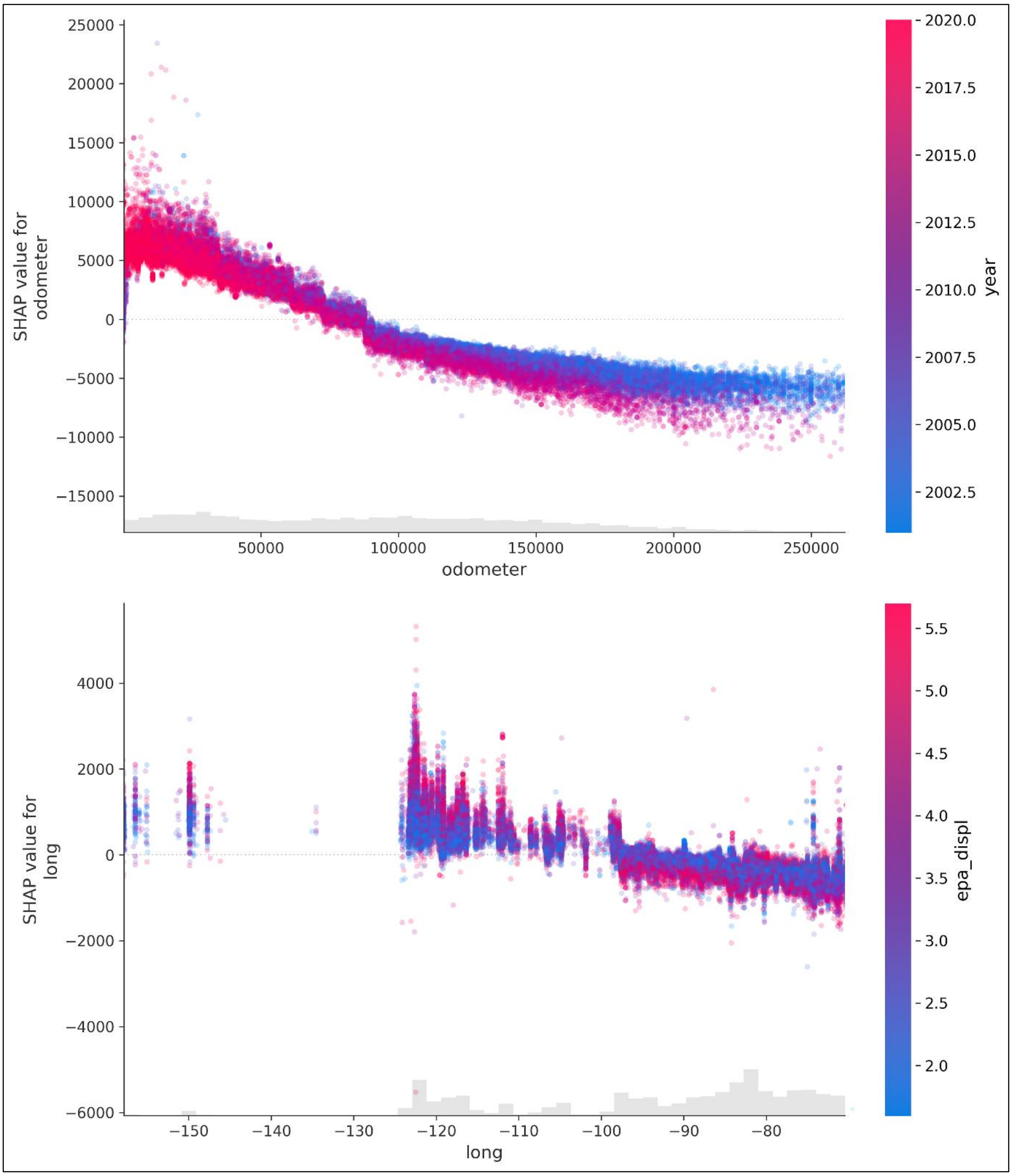
E[year]
E[f(x)]
27500
25000
22500
20000
17500
15000
12500
10000
E[f(x) | year]
1990
2000
2010
2020
year



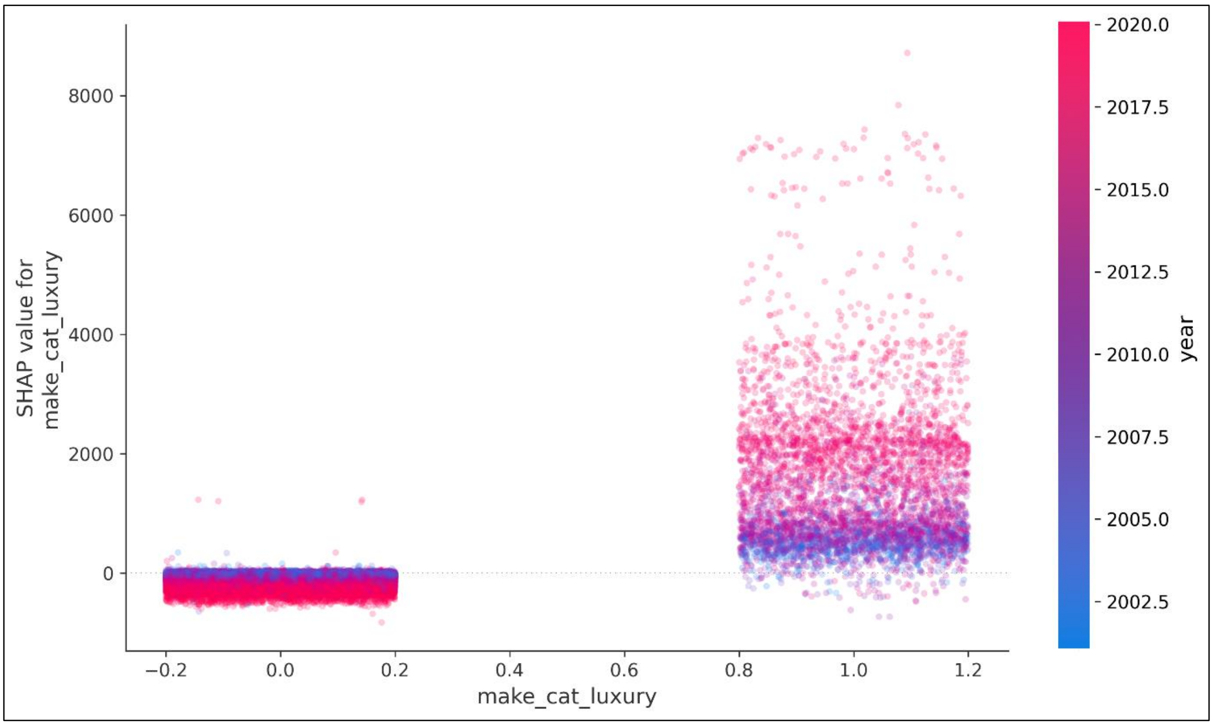
40000
30000
20000
10000
0
1985
1990
1995
2000
2005
2010
2015
2020
distribution of data points
year
40000
30000
20000
10000
0
1985
1990
1995
2000
2005
2010
2015
2020
distribution of data points
year
percentile info
(0.0)
(11.11)
(22.22)
(33.33)
(44.44)
(55.56)
(66.67)
(77.78)
(88.89)
(100.0)
40000
30000
20000
10000
0
1984
2005
2008
2011
2013
2014
2016
2017
2018
2022
year

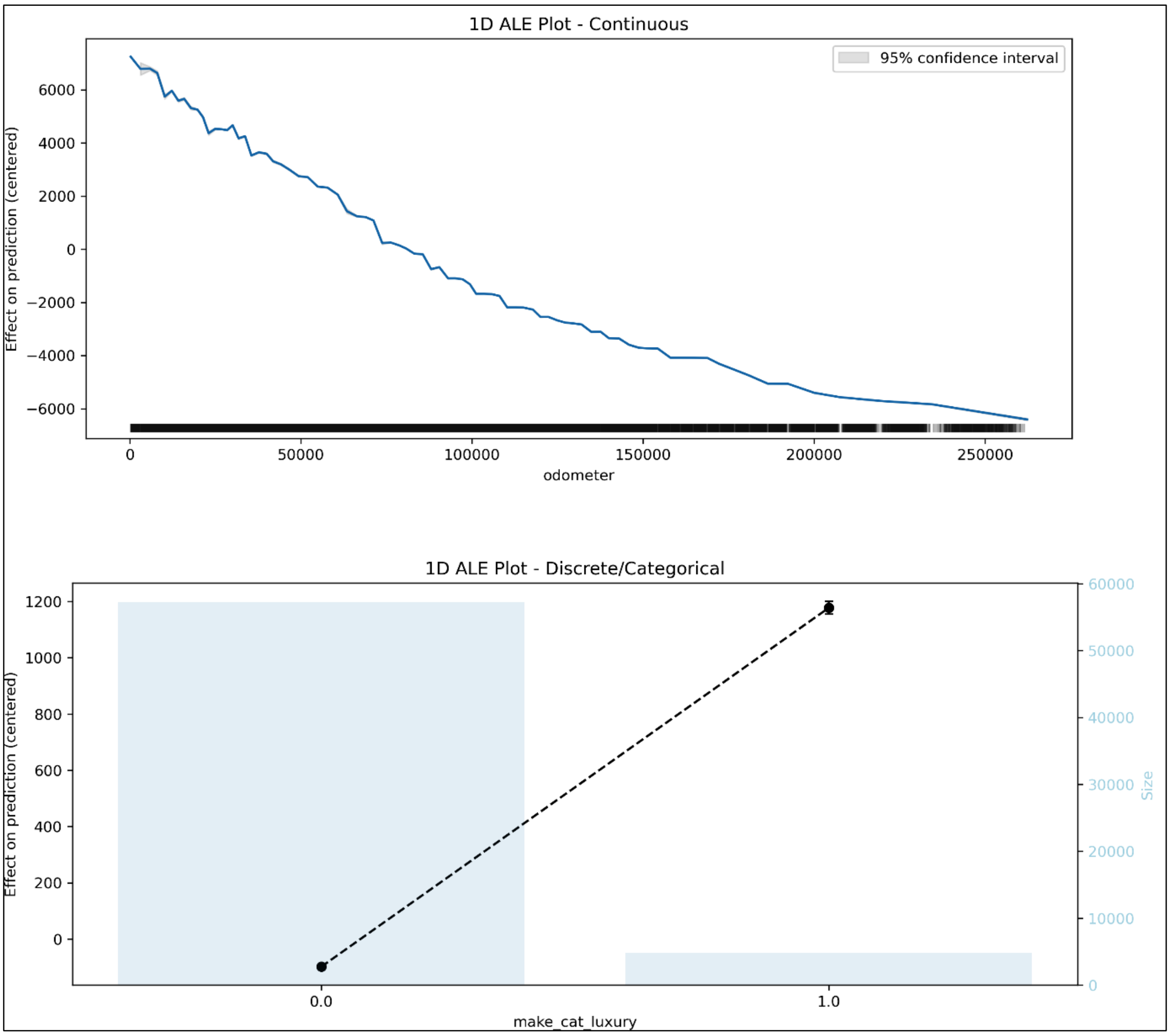

PDP for feature "odometer"
Number of unique grid points: 10
percentile info
(0.0)
(11.11)
(22.22)
(33.33)
(44.44)
(55.56)
(66.67)
(77.78)
(88.89)
(100.0)
20000
15000
10000
5000
0
−5000
−10000
−15000
−20000
0
18274.89
34308.78
56122
79940.89
100375.78
121381.33
145000
175999.11
10000000
odometer


PDP for feature "make_cat"
Number of unique grid points: 5
make_cat_obsolete
make_cat_regular
make_cat_premium
make_cat_luxury
make_cat_luxury_sports
make_cat
Actual predictions plot for make_cat
Distribution of actual prediction through different feature values.
prediction dist
4639.987
13823.496
20886.103
20304.522
29552.919
count
740
52229
5446
4874
728
make_cat_obsolete
make_cat_regular
make_cat_premium
make_cat_luxury
make_cat_luxury_sports
make_cat

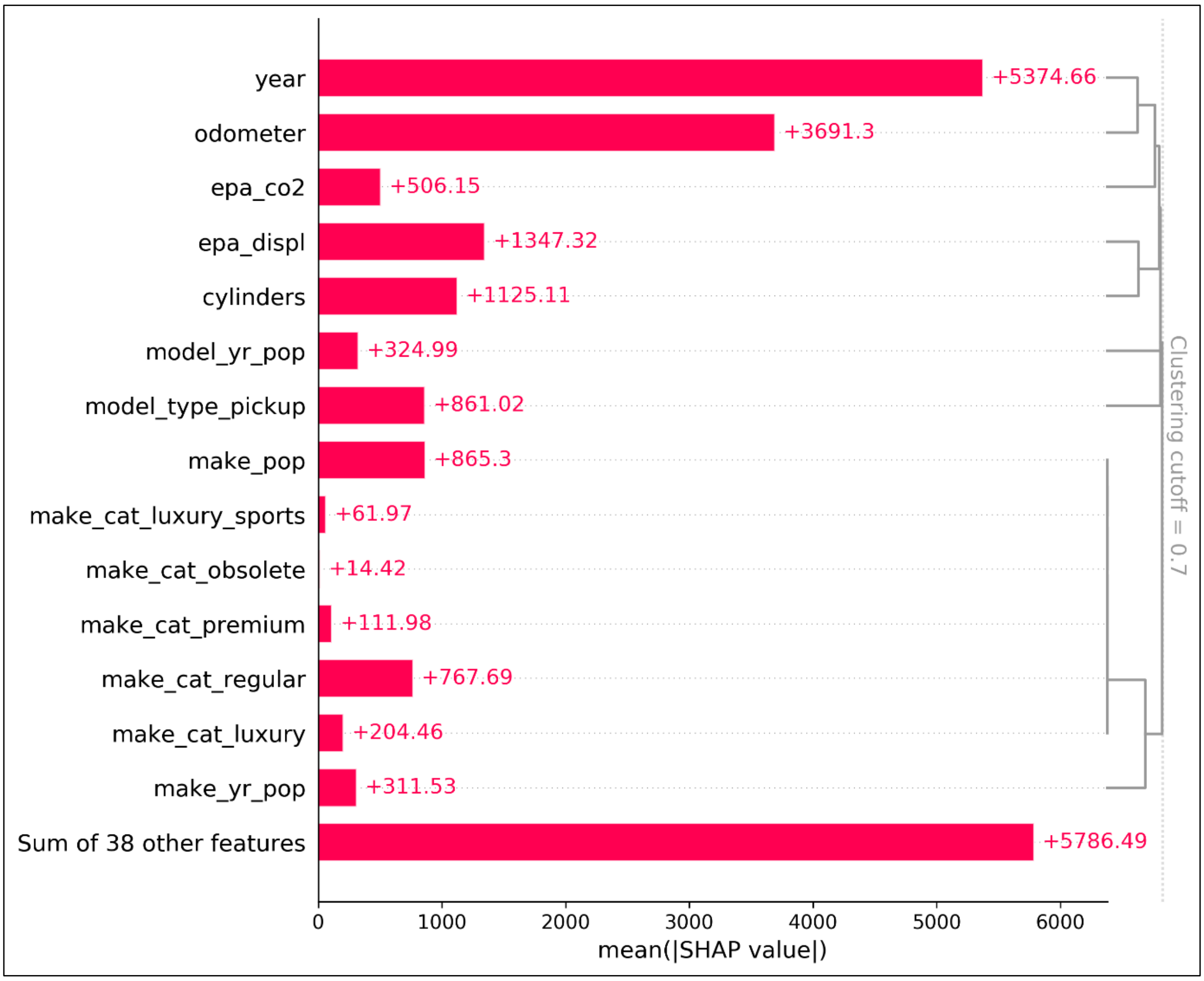


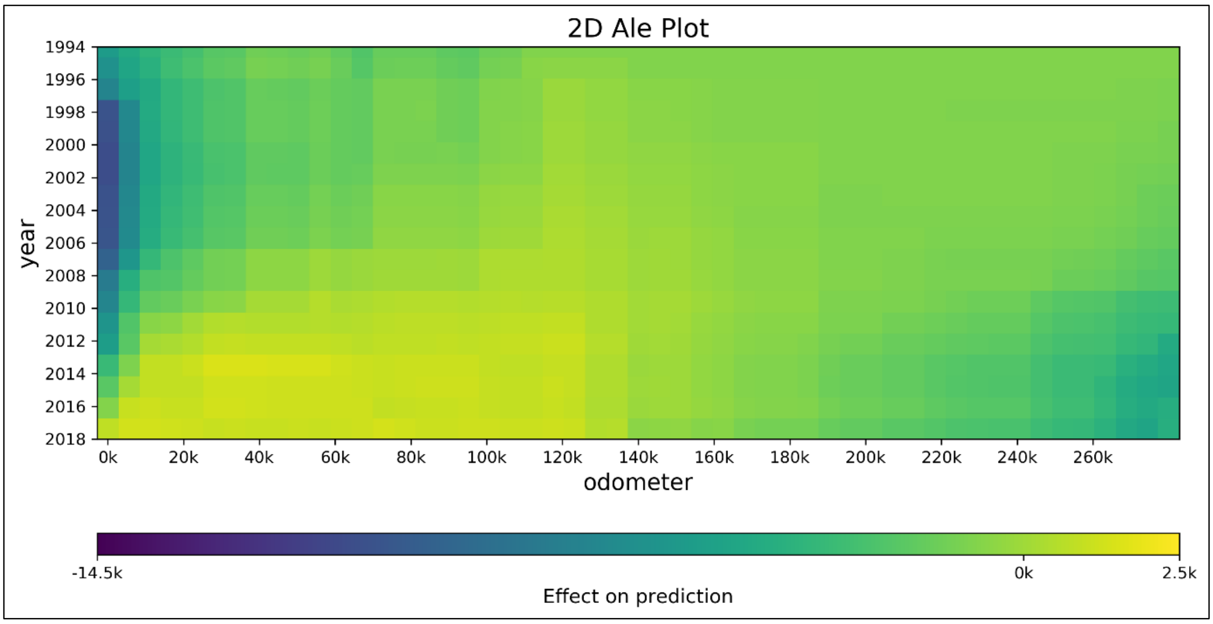
SHAP value for odometer
odometer
year
SHAP value for long
long
epa_displ

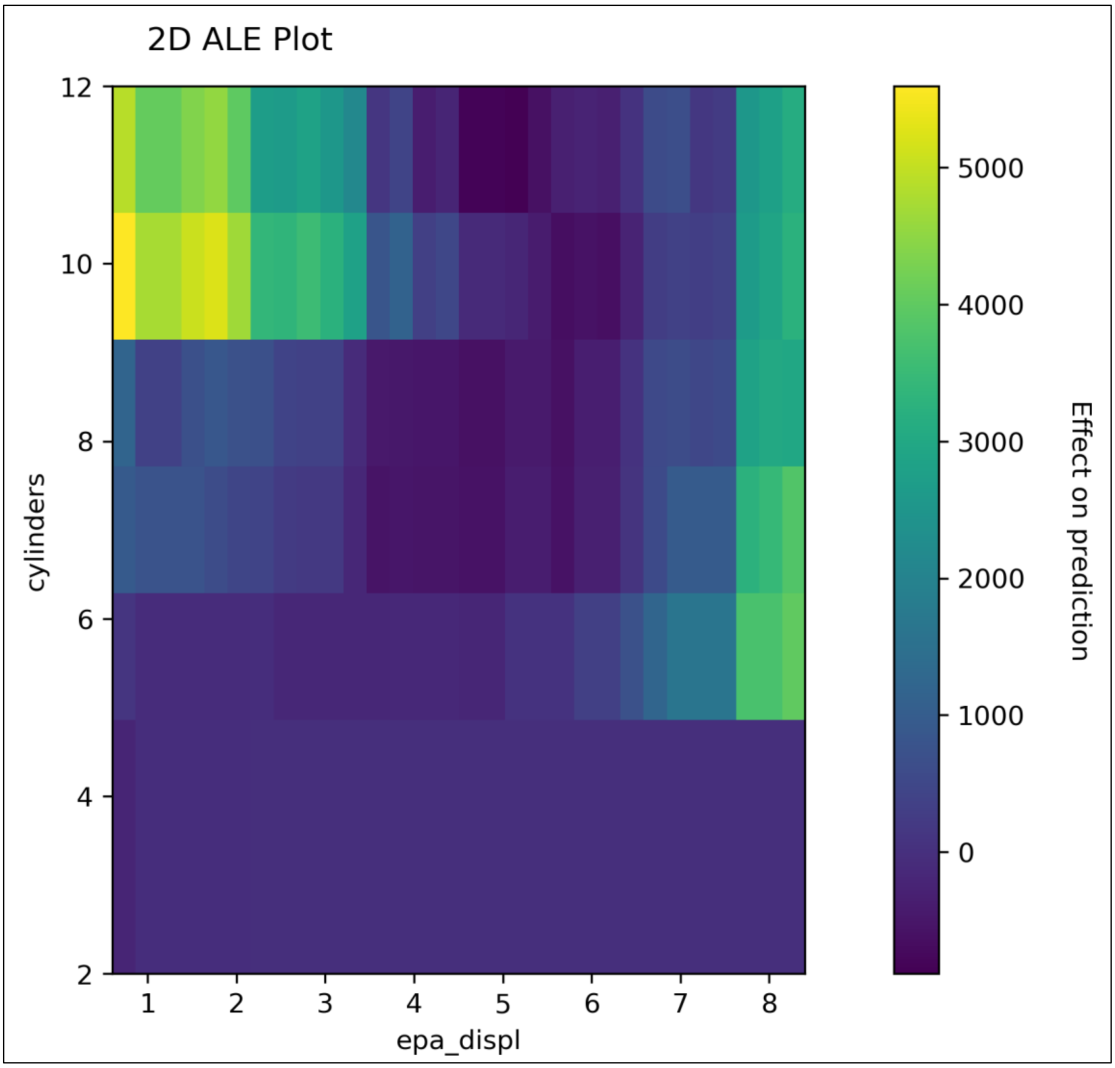

8000
6000
4000
2000
0
SHAP value for
make_cat_luxury
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
make_cat_luxury
2020.0
2017.5
2015.0
2012.5
2010.0
2007.5
2005.0
2002.5
year

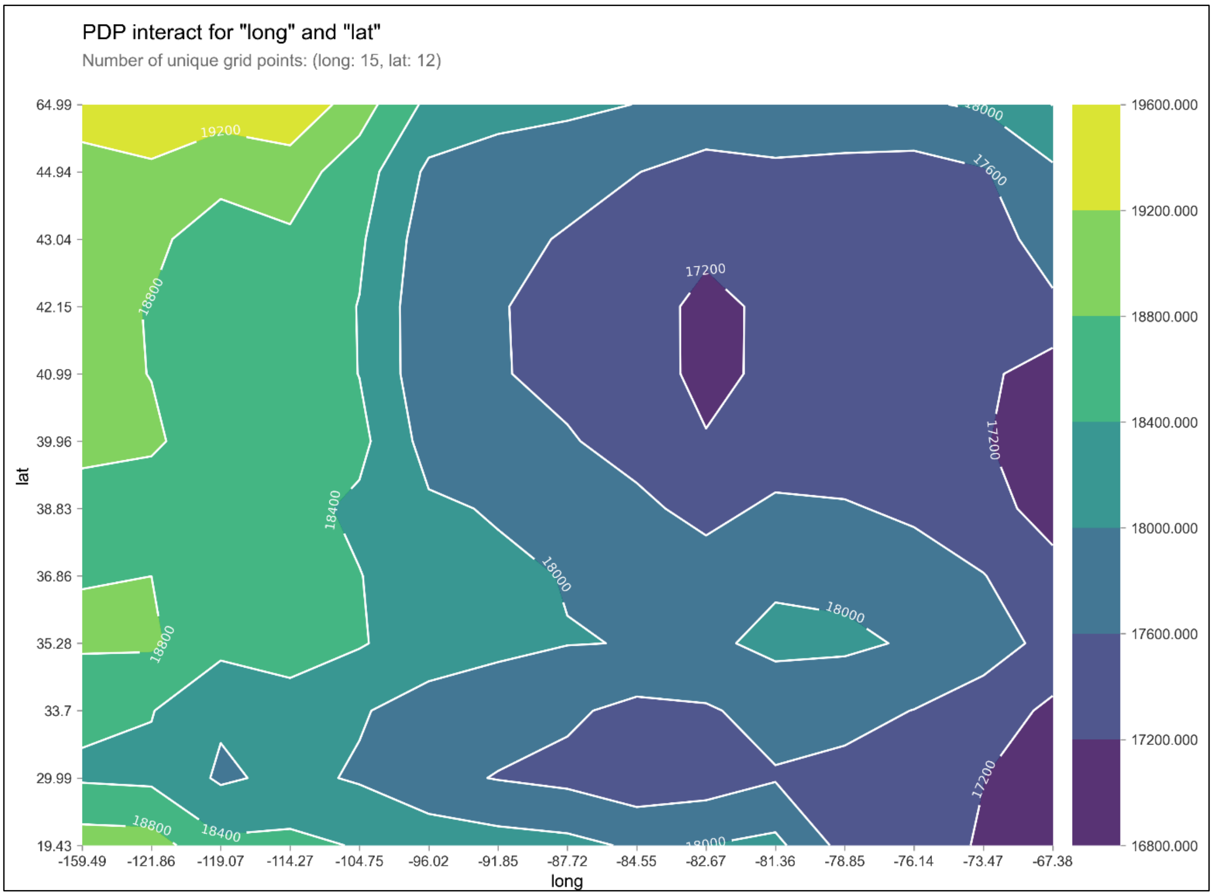


1D ALE Plot - Continuous
95% confidence interval
Effect on prediction (centered)
6000
4000
2000
0
−2000
−4000
−6000
0
50000
100000
150000
200000
250000
odometer
1D ALE Plot - Discrete/Categorical
Effect on prediction (centered)
1200
1000
800
600
400
200
0
60000
50000
40000
30000
20000
10000
0
Size
0.0
1.0
make_cat_luxury


year
odometer
epa_co2
epa_displ
cylinders
model_yr_pop
model_type_pickup
make_pop
make_cat_luxury_sports
make_cat_obsolete
make_cat_premium
make_cat_regular
make_cat_luxury
make_yr_pop
Sum of 38 other features
+5374.66
+3691.3
+506.15
+1347.32
+1125.11
+324.99
+861.02
+865.3
+61.97
+14.42
+111.98
+767.69
+204.46
+311.53
+5786.49
Clustering cutoff = 0.7
0
1000
2000
3000
4000
5000
6000
mean(|SHAP value|)

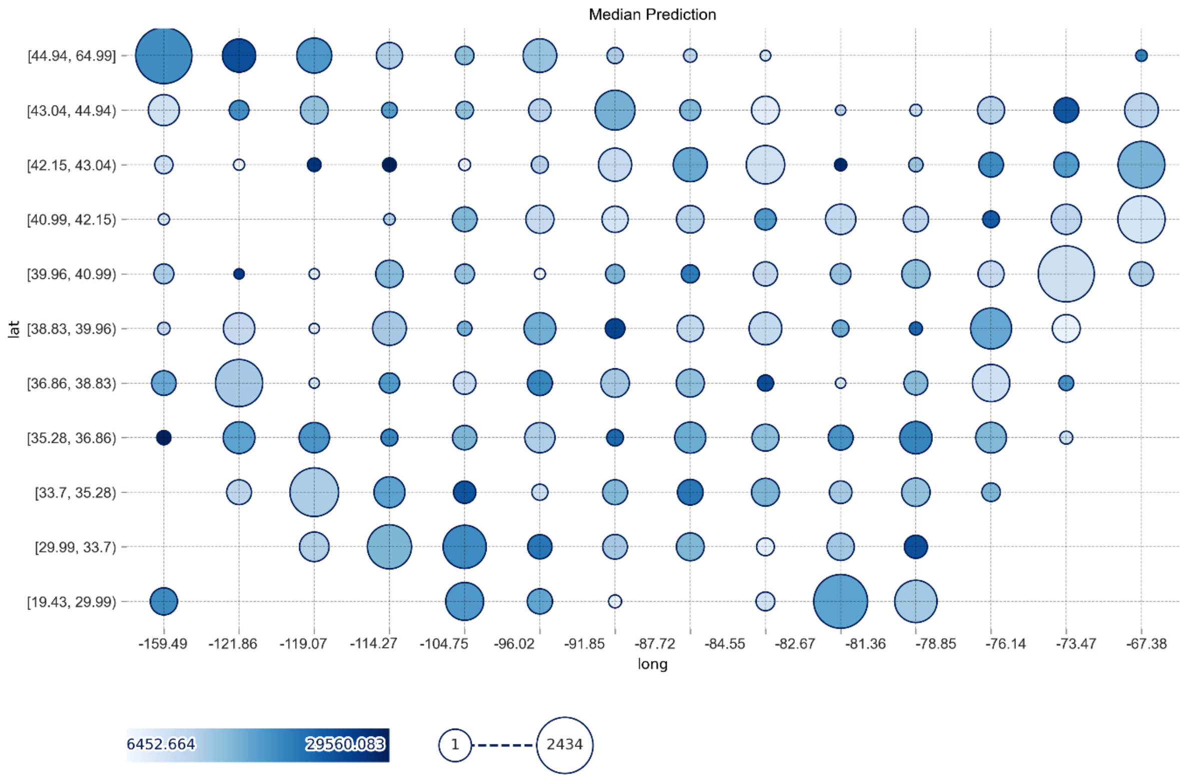

2D Ale Plot
year
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
0k
20k
40k
60k
80k
100k
120k
140k
160k
180k
200k
220k
240k
260k
odometer
-14.5k
0k
2.5k
Effect on prediction


2D ALE Plot
cylinders
epa_displ
Effect on prediction
12
10
8
6
4
2
1
2
3
4
5
6
7
8
5000
4000
3000
2000
1000
0


PDP interact for "long" and "lat"
Number of unique grid points: (long: 15, lat: 12)
lat
long
64.99
44.94
43.04
42.15
40.99
39.96
38.83
36.86
35.28
33.7
29.99
19.43
-159.49
-121.86
-119.07
-114.27
-104.75
-96.02
-91.85
-87.72
-84.55
-82.67
-81.36
-78.85
-76.14
-73.47
-67.38
19600.000
19200.000
18800.000
18400.000
18000.000
17600.000
17200.000
16800.000
19200
18800
18400
18000
17600
17200

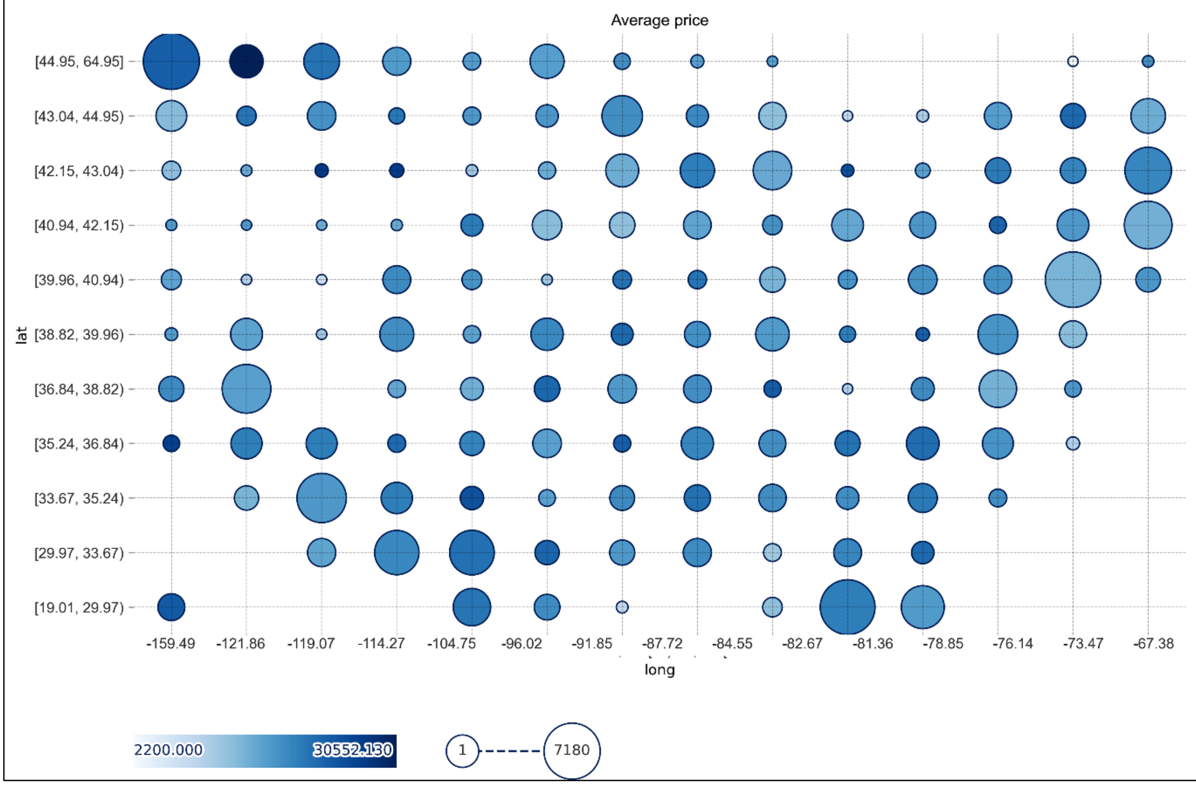
## Actual predictions plot for long & lat

Medium value of actual prediction through different feature value combinations.




## Target plot for feature "long & lat"

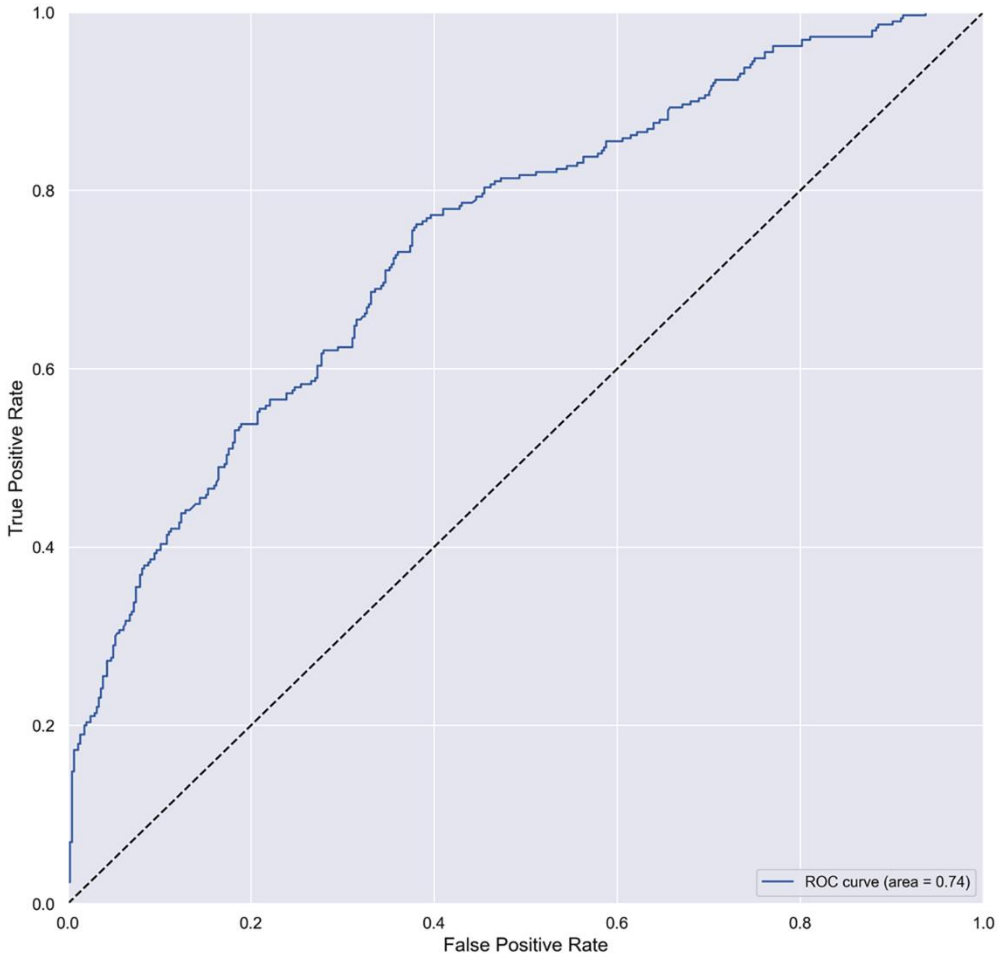
Average target value through different feature value combinations.



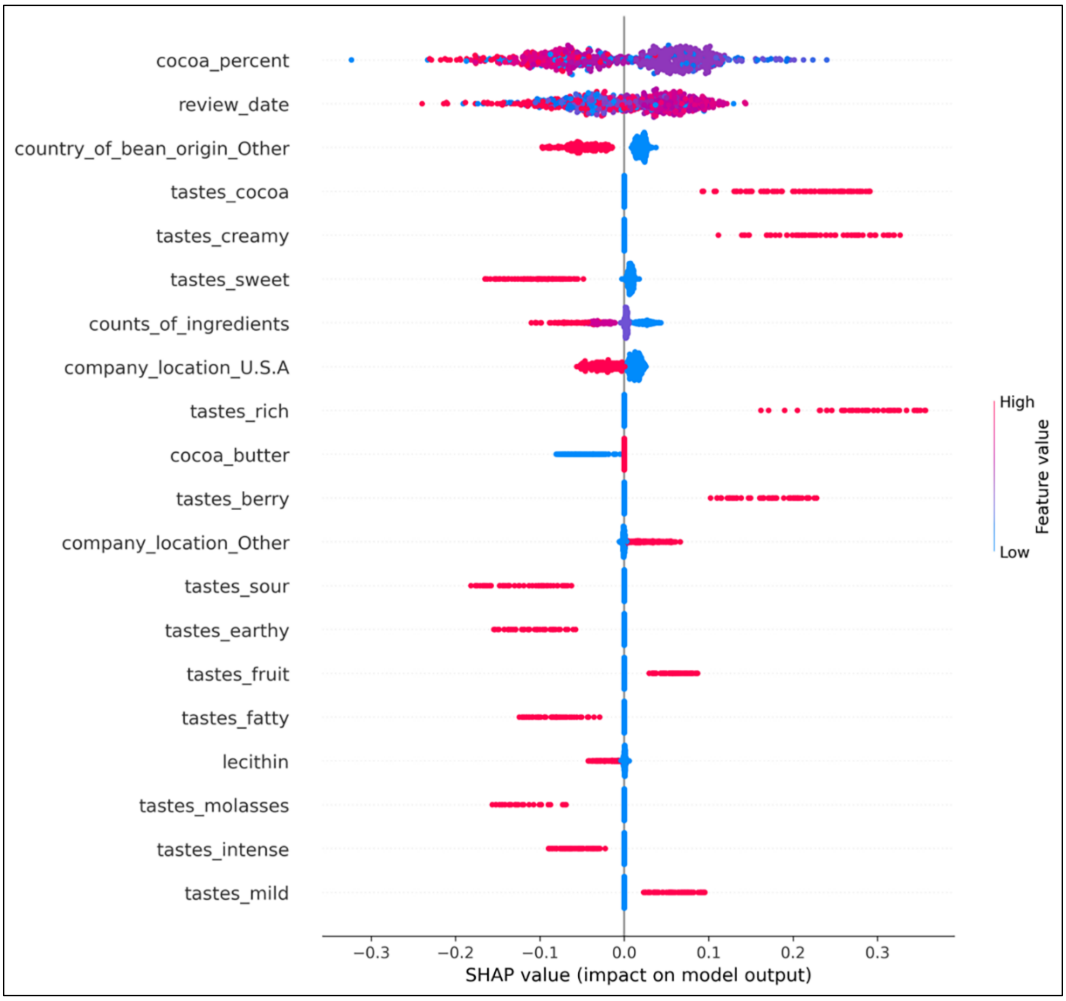
# Chapter 5: Local Model-Agnostic Interpretation Methods

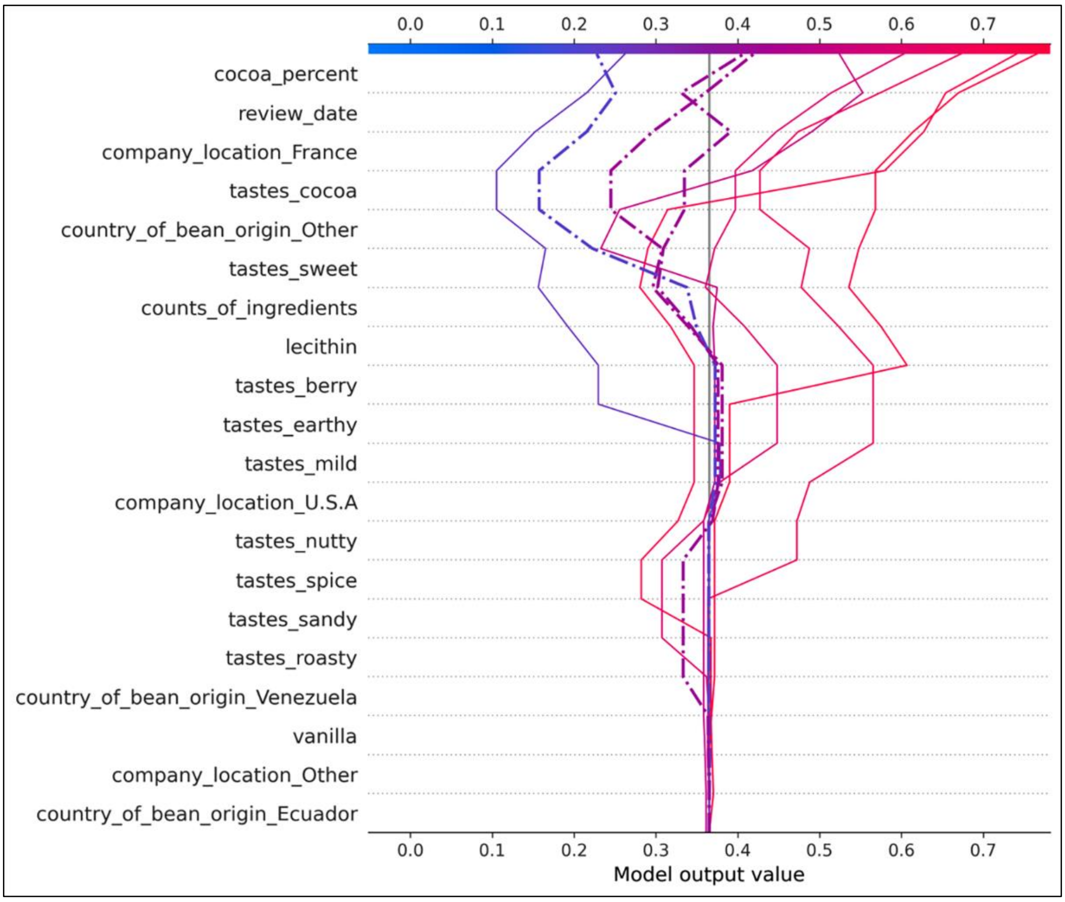
| | company | company_location | review_date | country_of_bean_origin | cocoa_percent | rating | counts_of_ingredients | cocoa_butter | vanilla |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 5150 | U.S.A | 2019 | Madagascar | 76.00 | 3.75 | 3 | 1 | 0 |
| 1 | 5150 | U.S.A | 2019 | Dominican republic | 76.00 | 3.50 | 3 | 1 | 0 |
| 2 | 5150 | U.S.A | 2019 | Tanzania | 76.00 | 3.25 | 3 | 1 | 0 |
| 3 | A. Morin | France | 2012 | Peru | 63.00 | 3.75 | 4 | 1 | 0 |
| | : | : | : | : | : | : | : | : | : |
| 2222 | Zotter | Austria | 2018 | Congo | 70.00 | 3.25 | 3 | 1 | 0 |
| 2223 | Zotter | Austria | 2018 | Blend | 75.00 | 3.00 | 3 | 1 | 0 |

| | first_taste | second_taste | third_taste | fourth_taste |
|---|---|---|---|---|
| 80 | oily | vegetal | nutty | cocoa |
| 81 | oily | vanilla | melon | cocoa |
| 82 | rich | sour | mild smoke | nan |
| 83 | fruity | sour | nan | nan |
| 84 | roast | high astringent | nan | nan |
| 85 | smokey | savory | nan | nan |
| 86 | sandy | roasty | nutty | nan |
| 87 | roasty | brownie | nutty | nan |
| 88 | red wine | rich | long | nan |
| 89 | creamy | fruit | cocoa | nan |

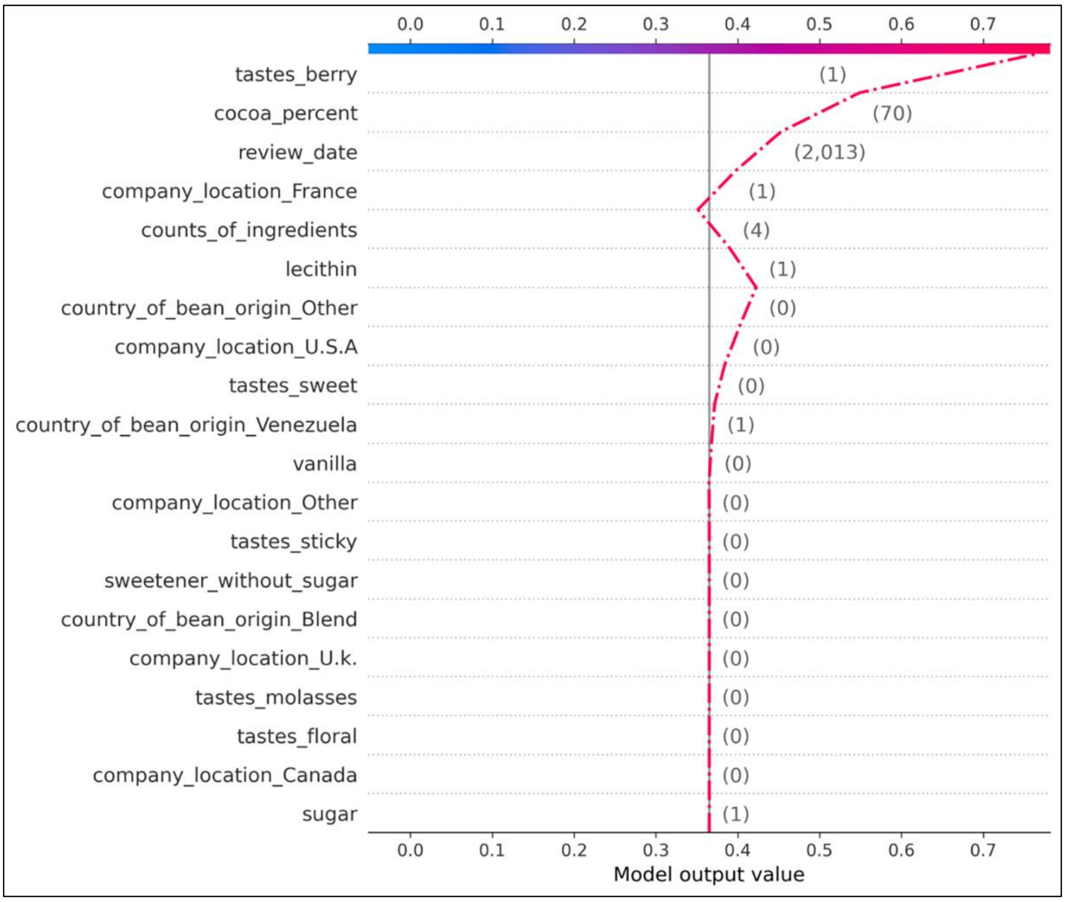



| | | | | |
|---|---|---|---|---|
| Accuracy_train: | 0.7315 | Accuracy_test: | 0.6989 | |
| Precision_test: | 0.6845 | Recall_test: | 0.4414 | |
| ROC-AUC_test: | 0.7450 | F1_test: | 0.5367 | MCC_test: 0.3461 |

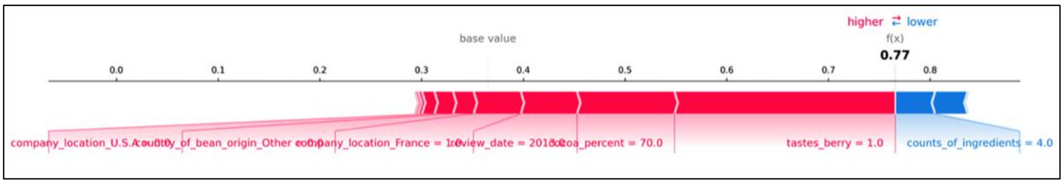

cocoa_percent
review_date
country_of_bean_origin_Other
tastes_cocoa
tastes_creamy
tastes_sweet
counts_of_ingredients
company_location_U.S.A
tastes_rich
cocoa_butter
tastes_berry
company_location_Other
tastes_sour
tastes_earthy
tastes_fruit
tastes_fatty
lecithin
tastes_molasses
tastes_intense
tastes_mild
High
Feature value
Low
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP value (impact on model output)

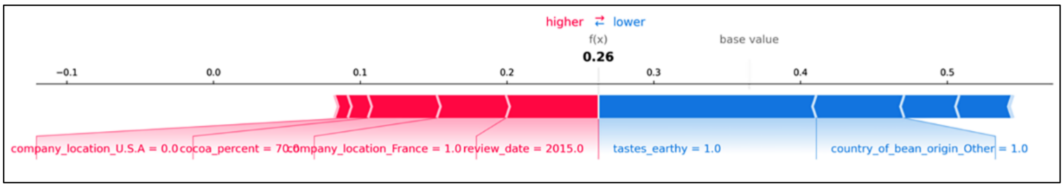


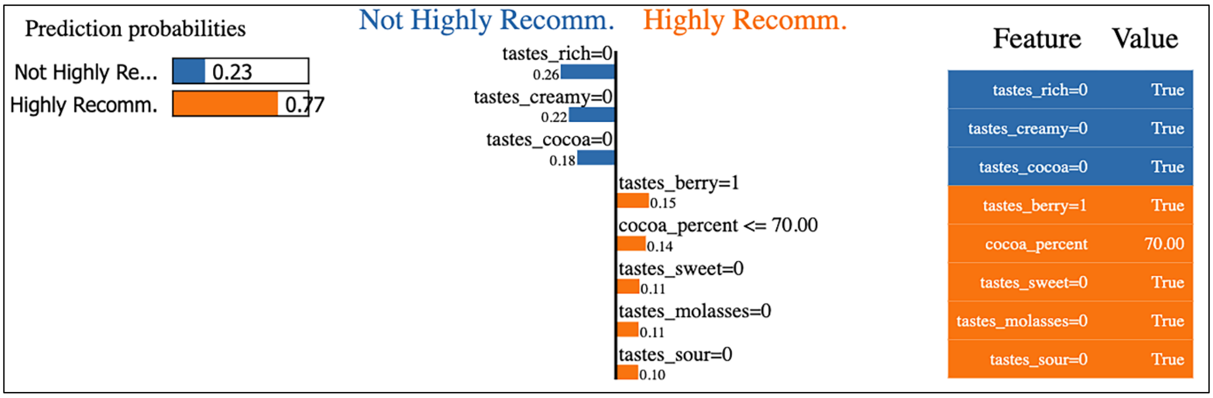
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
cocoa_percent
review_date
company_location_France
tastes_cocoa
country_of_bean_origin_Other
tastes_sweet
counts_of_ingredients
lecithin
tastes_berry
tastes_earthy
tastes_mild
company_location_U.S.A
tastes_nutty
tastes_spice
tastes_sandy
tastes_roasty
country_of_bean_origin_Venezuela
vanilla
company_location_Other
country_of_bean_origin_Ecuador
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
Model output value

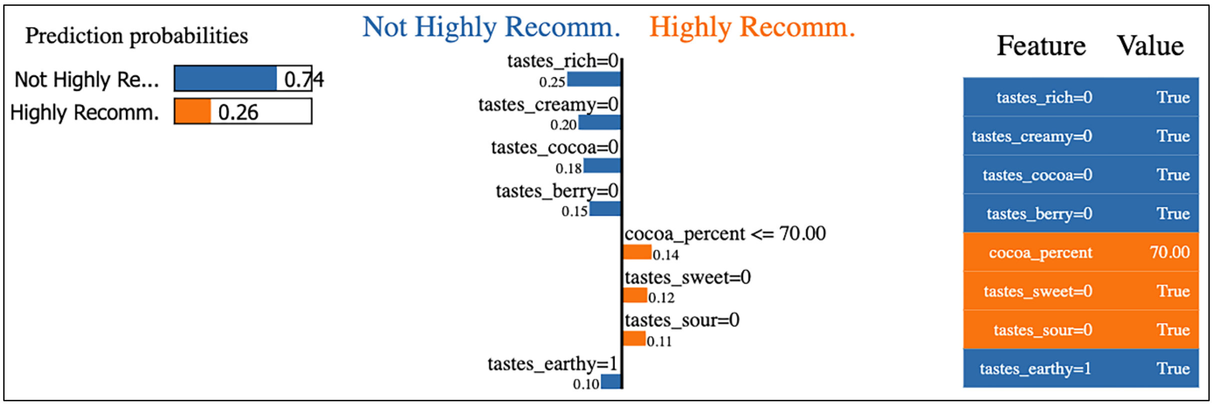

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
tastes_berry (1)
cocoa_percent (70)
review_date (2,013)
company_location_France (1)
counts_of_ingredients (4)
lecithin (1)
country_of_bean_origin_Other (0)
company_location_U.S.A (0)
tastes_sweet (0)
country_of_bean_origin_Venezuela (1)
vanilla (0)
company_location_Other (0)
tastes_sticky (0)
sweetener_without_sugar (0)
country_of_bean_origin_Blend (0)
company_location_U.k. (0)
tastes_molasses (0)
tastes_floral (0)
company_location_Canada (0)
sugar (1)
Model output value

| | 5 | 24 |
|---|---|---|
| rating | 4 | 3 |
| y | 1 | 0 |
| y_pred | 1 | 0 |
| review_date | 2013 | 2015 |
| cocoa_percent | 70 | 70 |
| counts_of_ingredients | 4 | 4 |
| cocoa_butter | 1 | 1 |
| vanilla | 0 | 0 |
| lecithin | 1 | 1 |
| salt | 0 | 0 |
| sugar | 1 | 1 |
| sweetener_without_sugar | 0 | 0 |
| company_location_Canada | 0 | 0 |
| : | : | : |
| country_of_bean_origin_Nicaragua | 0 | 0 |
| country_of_bean_origin_Other | 0 | 1 |
| country_of_bean_origin_Peru | 0 | 0 |
| country_of_bean_origin_Venezuela | 1 | 0 |
| tastes_cocoa | 0 | 0 |

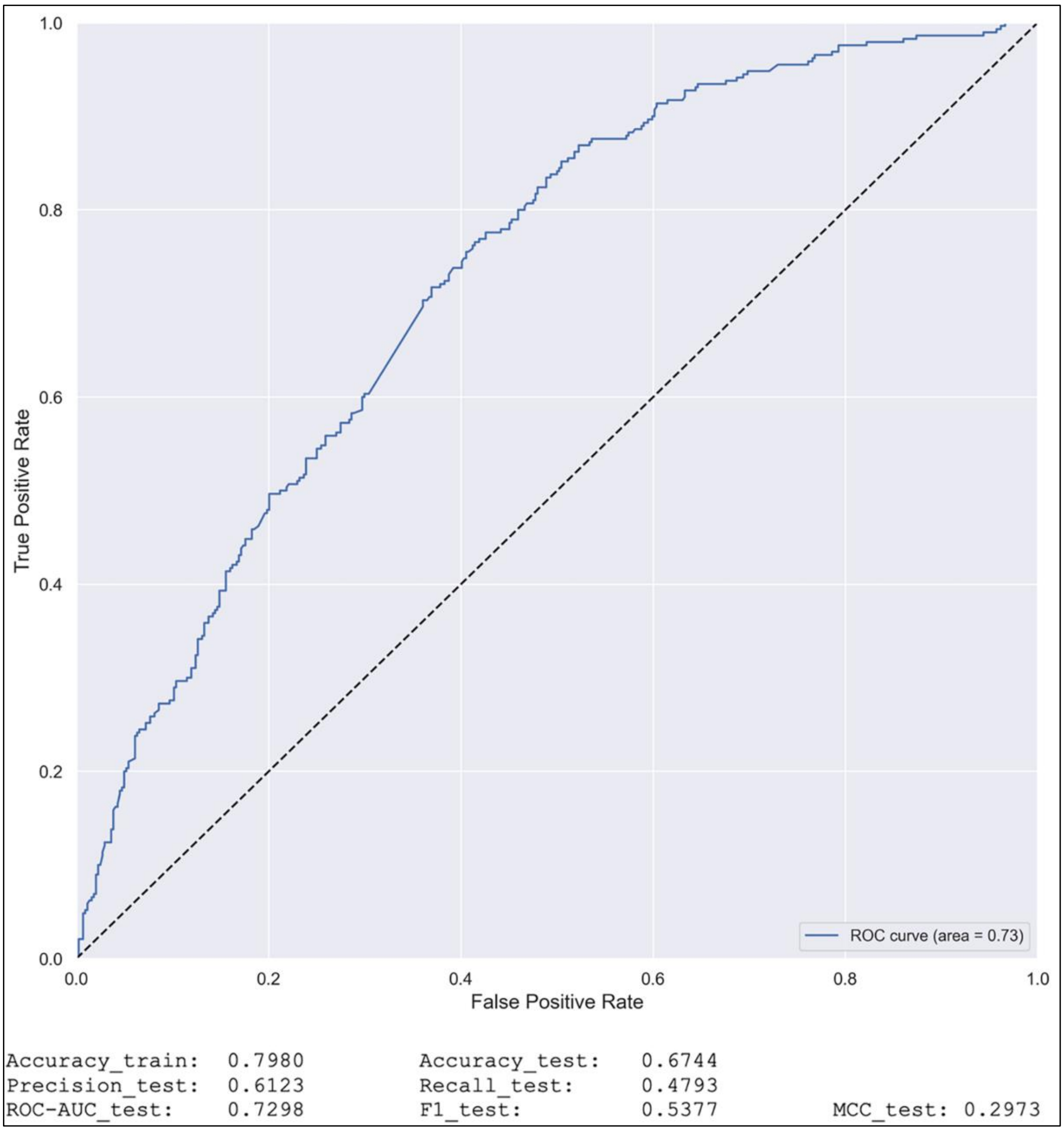

higher ⇄ lower
base value
f(x)
0.77
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
tastes_berry = 1.0
counts_of_ingredients = 4.0

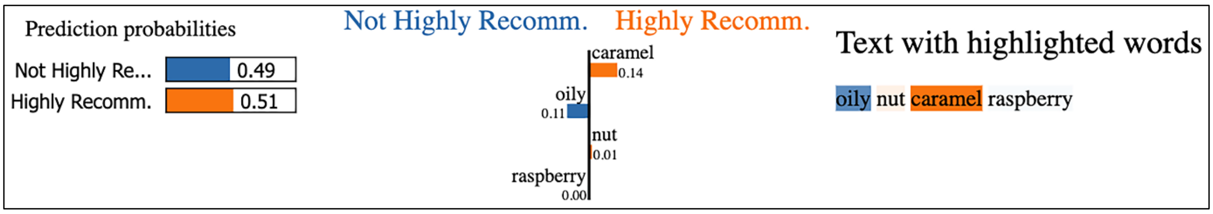

higher ⇄ lower
f(x)
0.26
base value
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
company_location_U.S.A = 0.0
review_date = 2015.0
tastes_earthy = 1.0
country_of_bean_origin_Other = 1.0

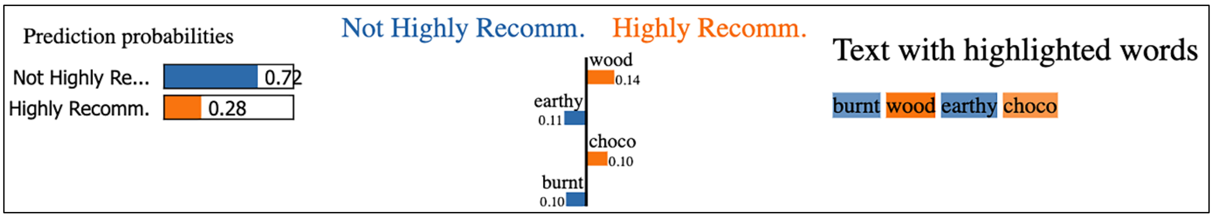

Prediction probabilities
Not Highly Re... 0.23
Highly Recomm. 0.77
Not Highly Recomm.
Highly Recomm.
tastes_rich=0 0.26
tastes_creamy=0 0.22
tastes_cocoa=0 0.18
tastes_berry=1 0.15
cocoa_percent <= 70.00 0.14
tastes_sweet=0 0.11
tastes_molasses=0 0.11
tastes_sour=0 0.10
Feature Value
tastes_rich=0 True
tastes_creamy=0 True
tastes_cocoa=0 True
tastes_berry=1 True
cocoa_percent 70.00
tastes_sweet=0 True
tastes_molasses=0 True
tastes_sour=0 True

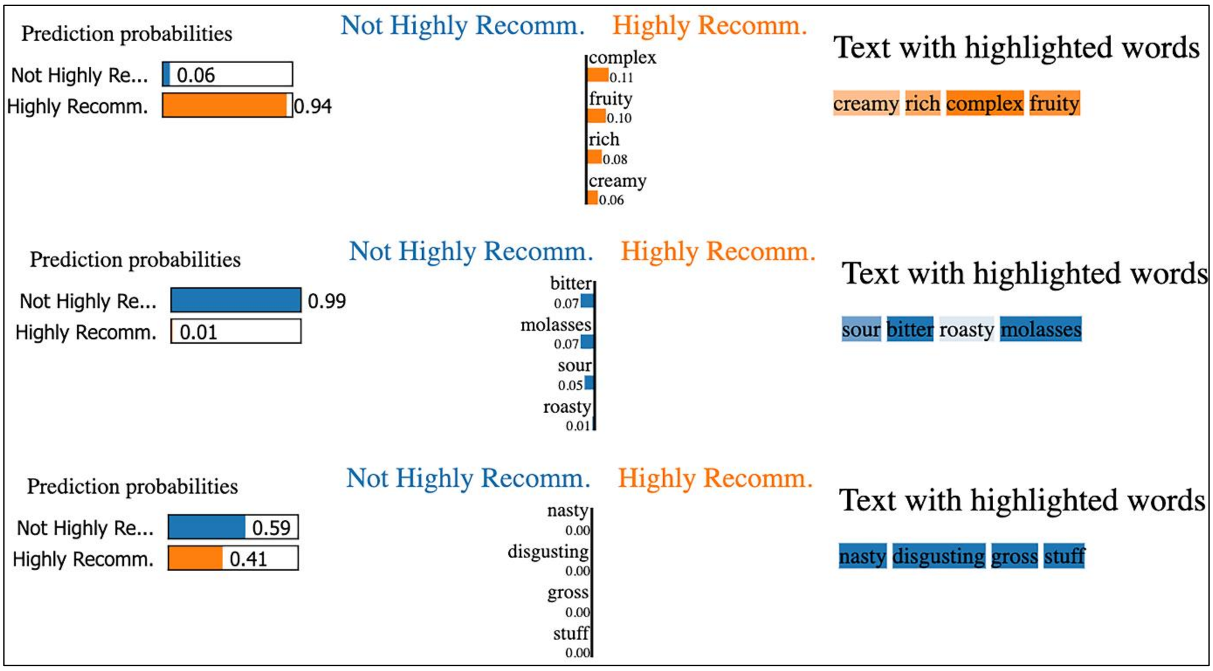

Prediction probabilities
Not Highly Re... 0.74
Highly Recomm. 0.26
Not Highly Recomm.
Highly Recomm.
tastes_rich=0 0.25
tastes_creamy=0 0.20
tastes_cocoa=0 0.18
tastes_berry=0 0.15
cocoa_percent <= 70.00 0.14
tastes_sweet=0 0.12
tastes_sour=0 0.11
tastes_earthy=1 0.10
Feature Value
tastes_rich=0 True
tastes_creamy=0 True
tastes_cocoa=0 True
tastes_berry=0 True
cocoa_percent 70.00
tastes_sweet=0 True
tastes_sour=0 True
tastes_earthy=1 True

| | taste | tf-idf |
|---|---|---|
| **305** | raspberry | 0.585538 |
| **259** | nut | 0.491542 |
| **265** | oily | 0.463973 |
| **64** | caramel | 0.447504 |
| **274** | papaya | 0.000000 |

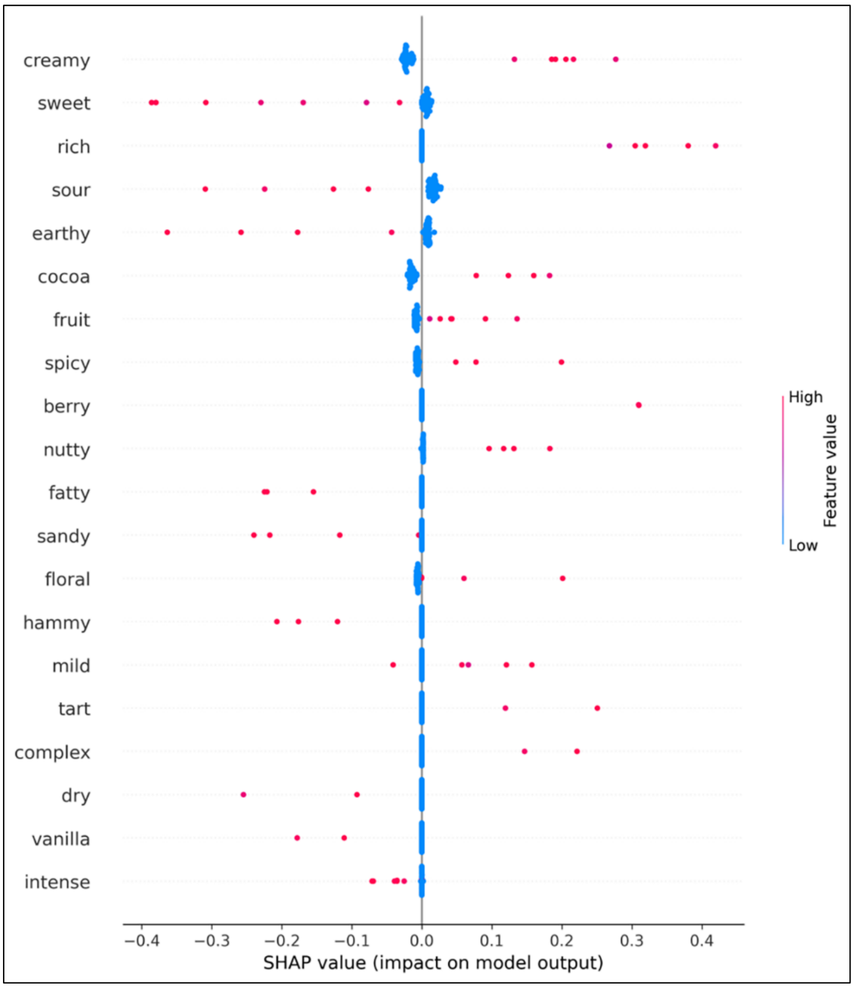

True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.73)
Accuracy_train: 0.7980
Accuracy_test: 0.6744
Precision_test: 0.6123
Recall_test: 0.4793
ROC-AUC_test: 0.7298
F1_test: 0.5377
MCC_test: 0.2973


Prediction probabilities
Not Highly Re... 0.49
Highly Recomm. 0.51
Not Highly Recomm.
Highly Recomm.
caramel 0.14
oily 0.11
nut 0.01
raspberry 0.00
Text with highlighted words
oily nut caramel raspberry

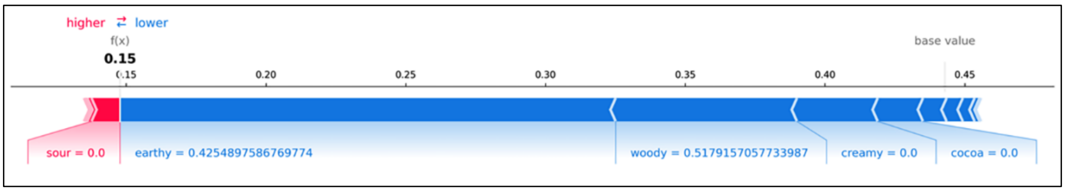

Prediction probabilities
Not Highly Re... 0.72
Highly Recomm. 0.28
Not Highly Recomm.
Highly Recomm.
wood 0.14
earthy 0.11
choco 0.10
burnt 0.10
Text with highlighted words
burnt wood earthy choco

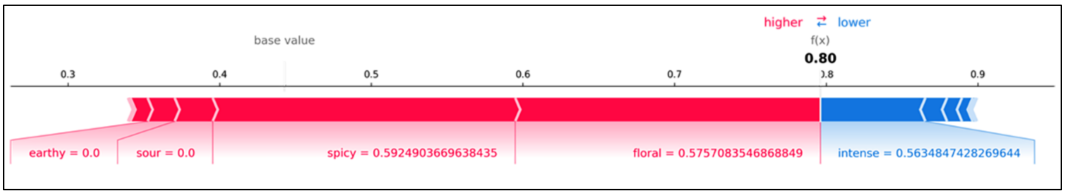


Prediction probabilities
Not Highly Re... 0.06
Highly Recomm. 0.94
Not Highly Recomm.
Highly Recomm.
complex 0.11
fruity 0.10
rich 0.08
creamy 0.06
Text with highlighted words
creamy rich complex fruity
Prediction probabilities
Not Highly Re... 0.99
Highly Recomm. 0.01
Not Highly Recomm.
Highly Recomm.
bitter 0.07
molasses 0.07
sour 0.05
roasty 0.01
Text with highlighted words
sour bitter roasty molasses
Prediction probabilities
Not Highly Re... 0.59
Highly Recomm. 0.41
Not Highly Recomm.
Highly Recomm.
nasty 0.00
disgusting 0.00
gross 0.00
stuff 0.00
Text with highlighted words
nasty disgusting gross stuff

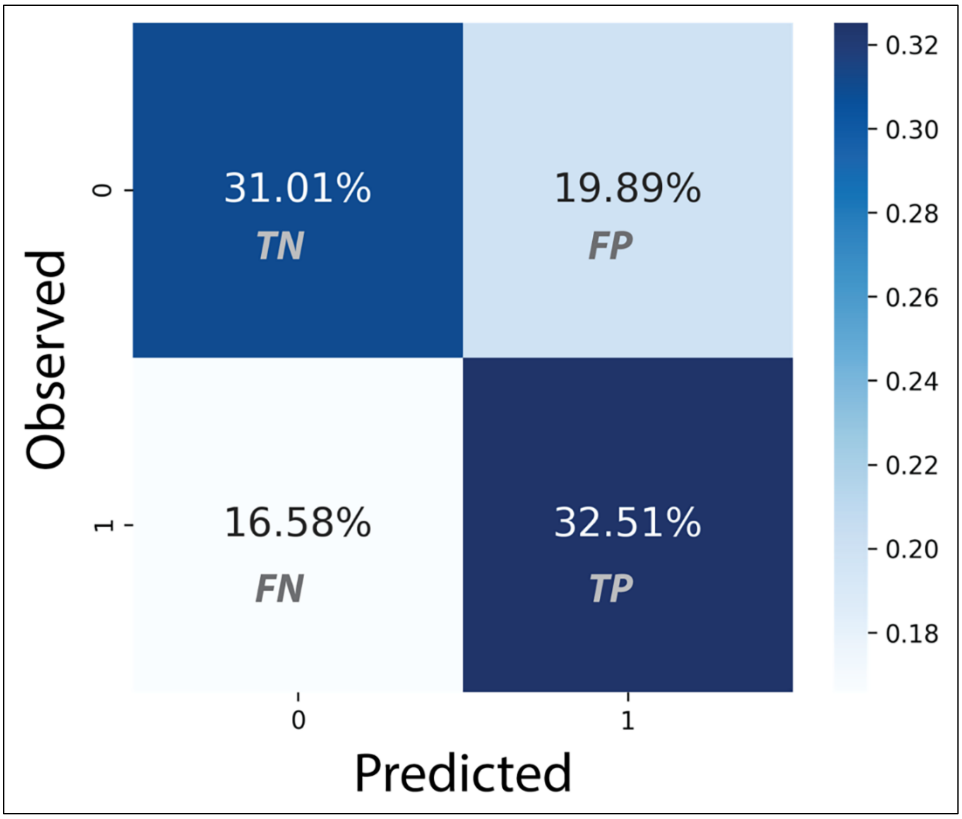



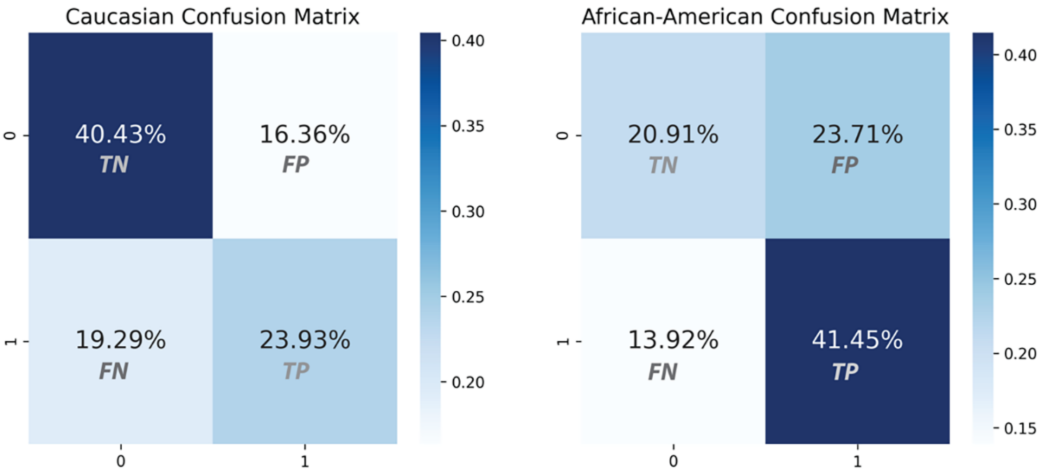
creamy
sweet
rich
sour
earthy
cocoa
fruit
spicy
berry
nutty
fatty
sandy
floral
hammy
mild
tart
complex
dry
vanilla
intense
−0.4 −0.3 −0.2 −0.1 0.0 0.1 0.2 0.3 0.4
SHAP value (impact on model output)
High
Feature value
Low



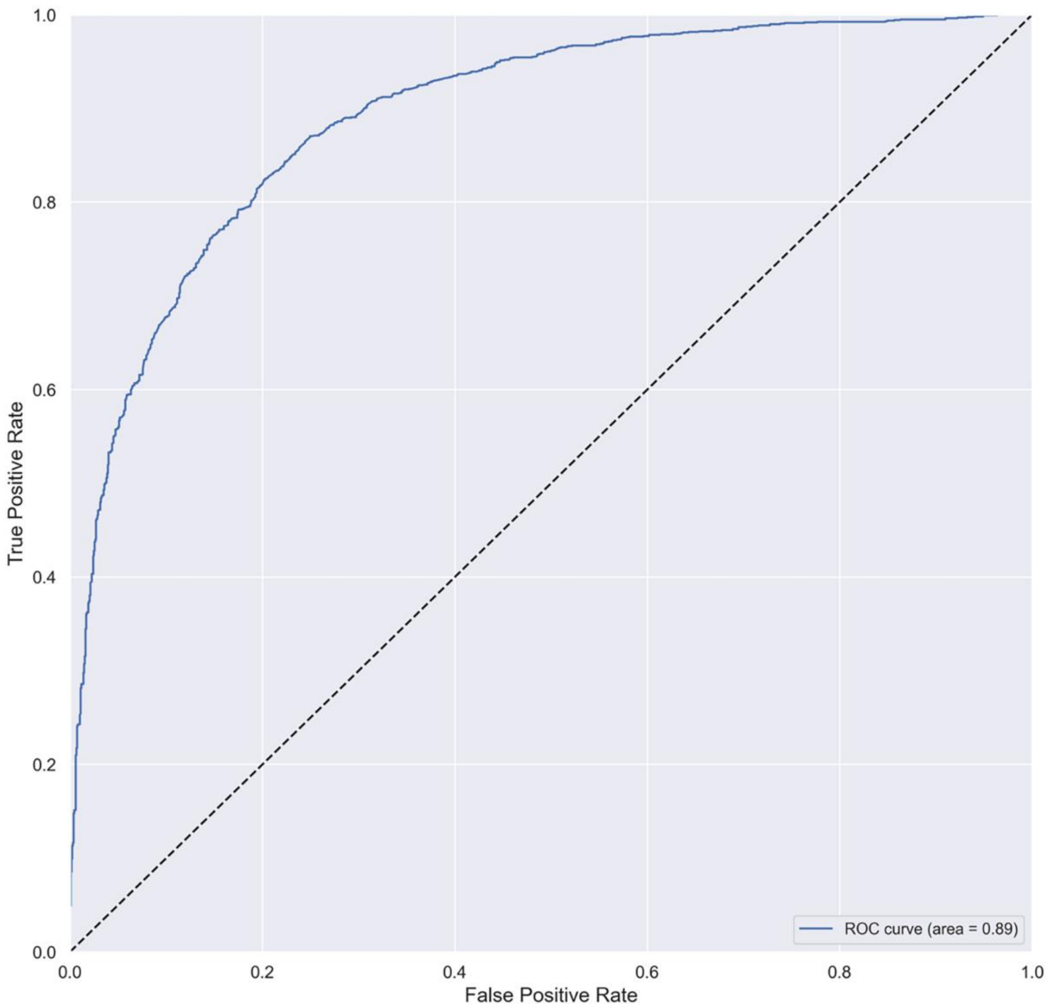




## Chapter 6: Anchors and Counterfactual Explanations



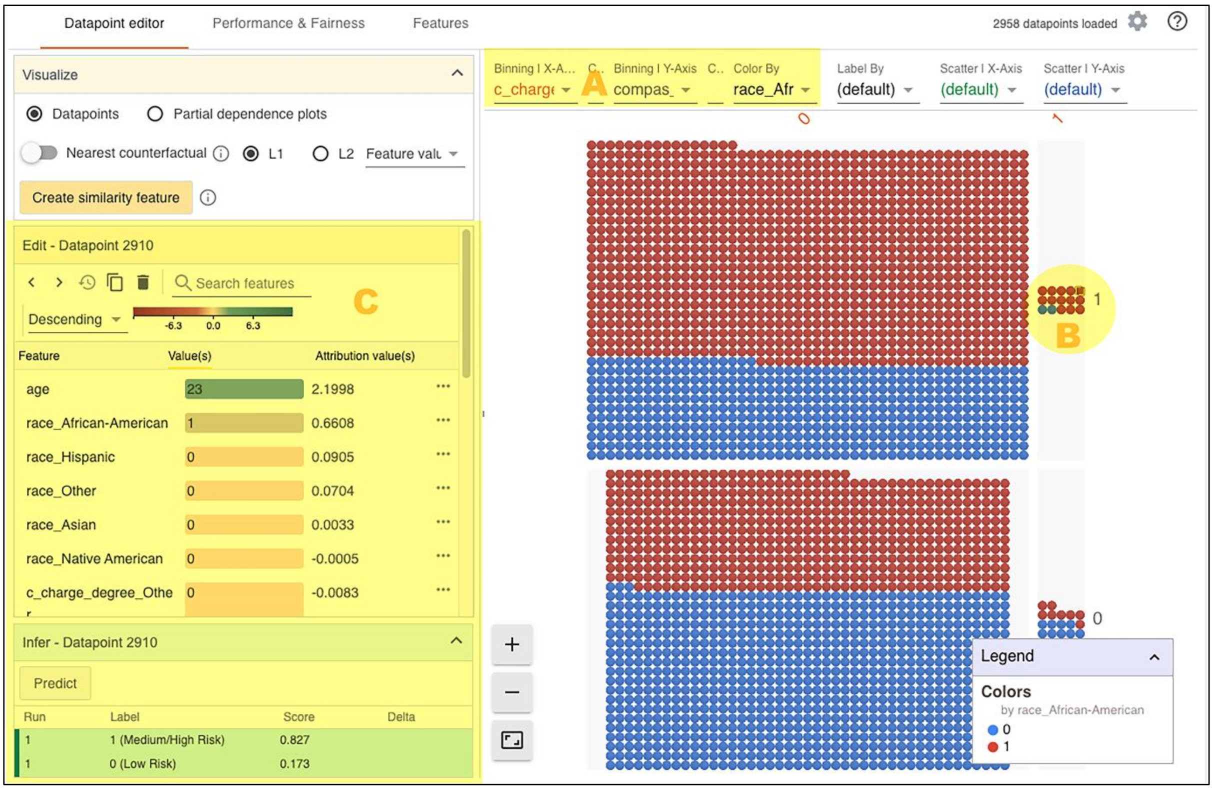



```
African-American FPR:    53.1%
Caucasian FPR:           28.8%
Ratio FPRs:              1.84 x
```

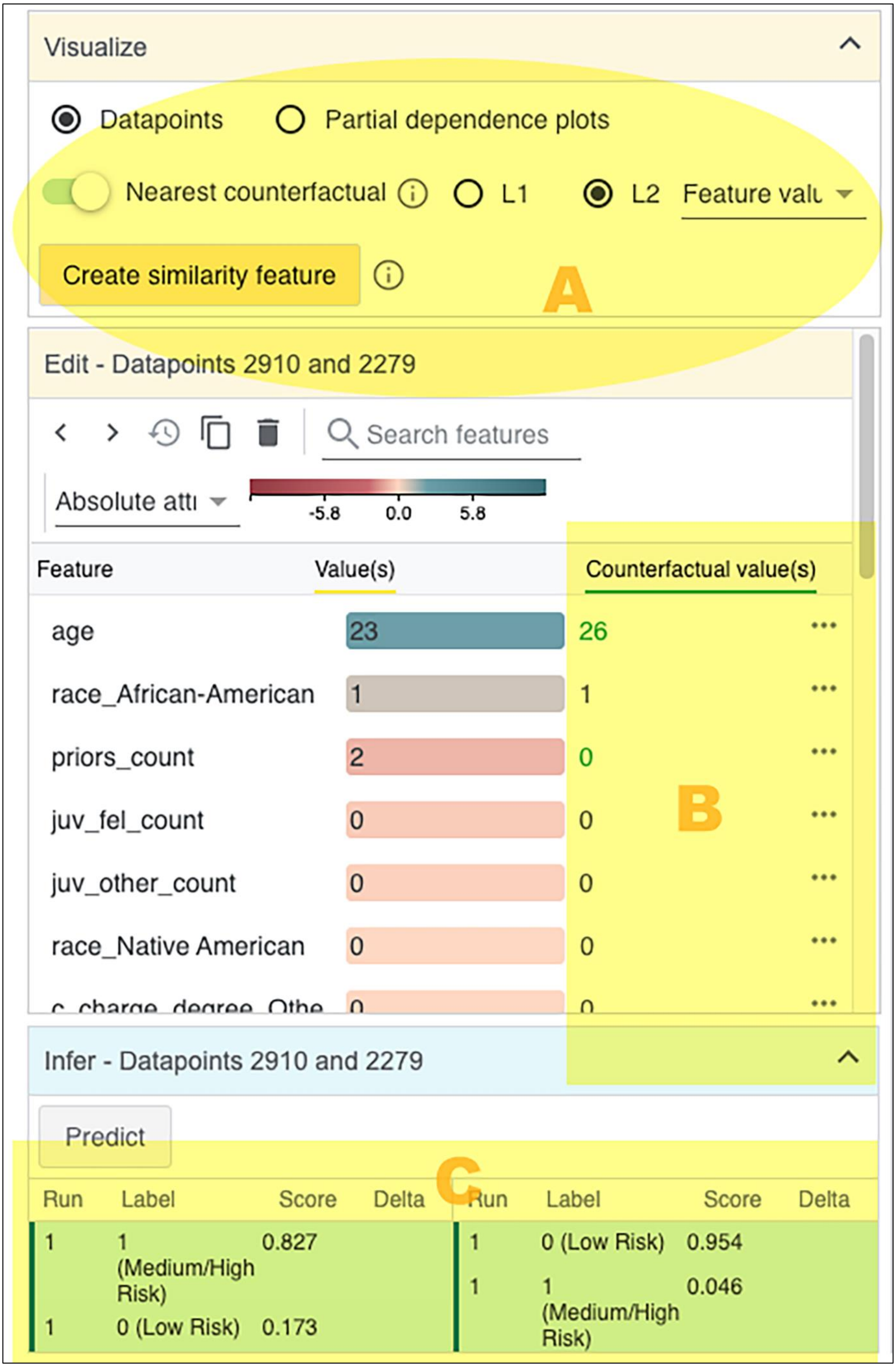



```
Accuracy_train:   0.8790        Accuracy_test:    0.8087
Precision_test:   0.8277        Recall_test:      0.8110
ROC-AUC_test:     0.8927        F1_test:          0.8193        MCC_test: 0.6162
```

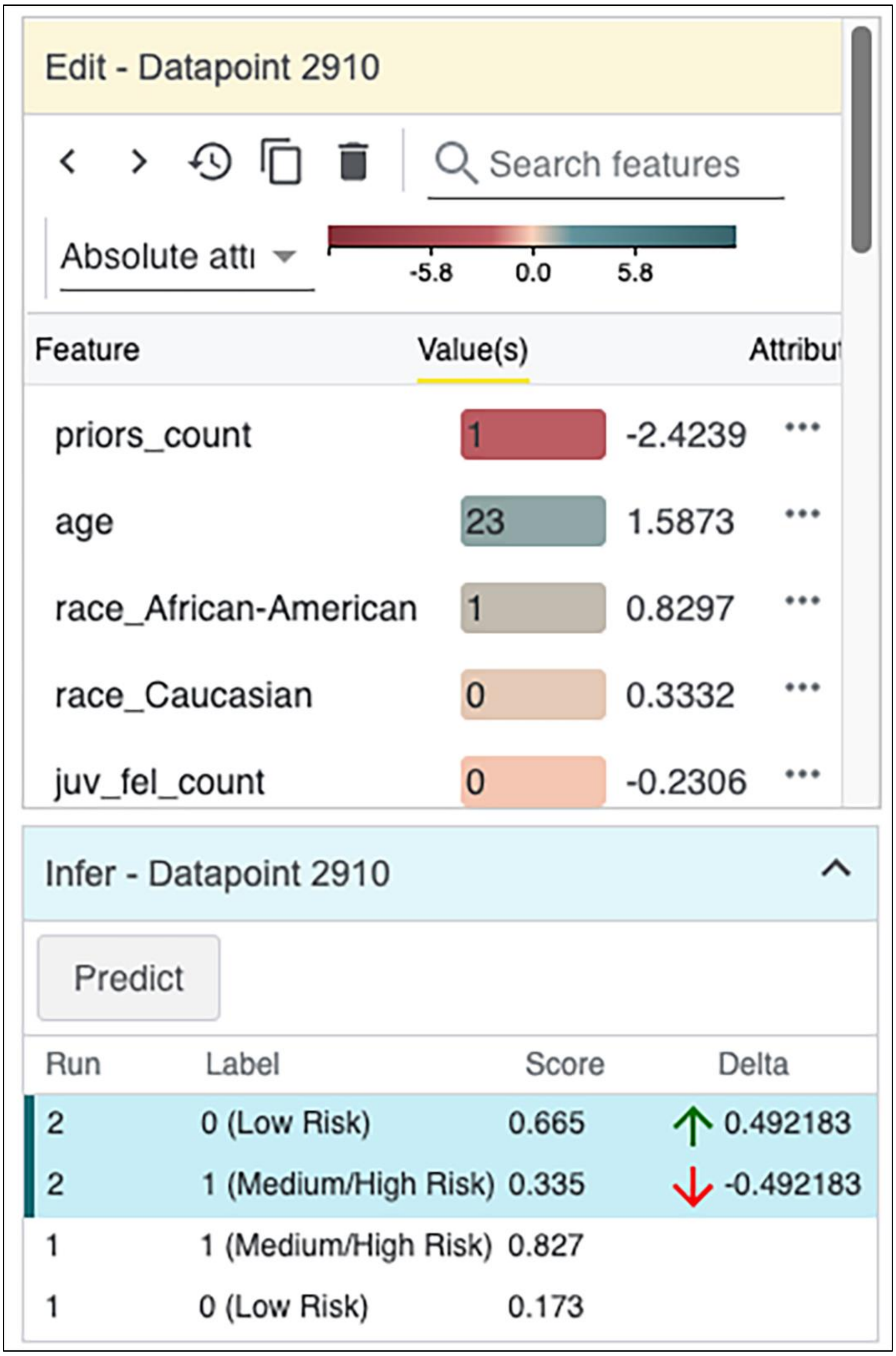
| | 10127 | 2726 | 5231 |
|---|---|---|---|
| y | 0 | 0 | 1 |
| y_pred | 0 | 0 | 1 |
| age | 24 | 23 | 23 |
| : | : | : | : |
| priors_count | 2 | 2 | 2 |
| sex_Female | 0 | 0 | 0 |
| sex_Male | 1 | 1 | 1 |
| race_African-American | 0 | 0 | 1 |
| race_Asian | 0 | 0 | 0 |
| race_Caucasian | 1 | 0 | 0 |
| race_Hispanic | 0 | 1 | 0 |
| : | : | : | : |
| c_charge_degree_(F3) | 0 | 1 | 0 |
| c_charge_degree_(F7) | 0 | 0 | 1 |
| c_charge_degree_(M1) | 1 | 0 | 0 |
| : | : | : | : |

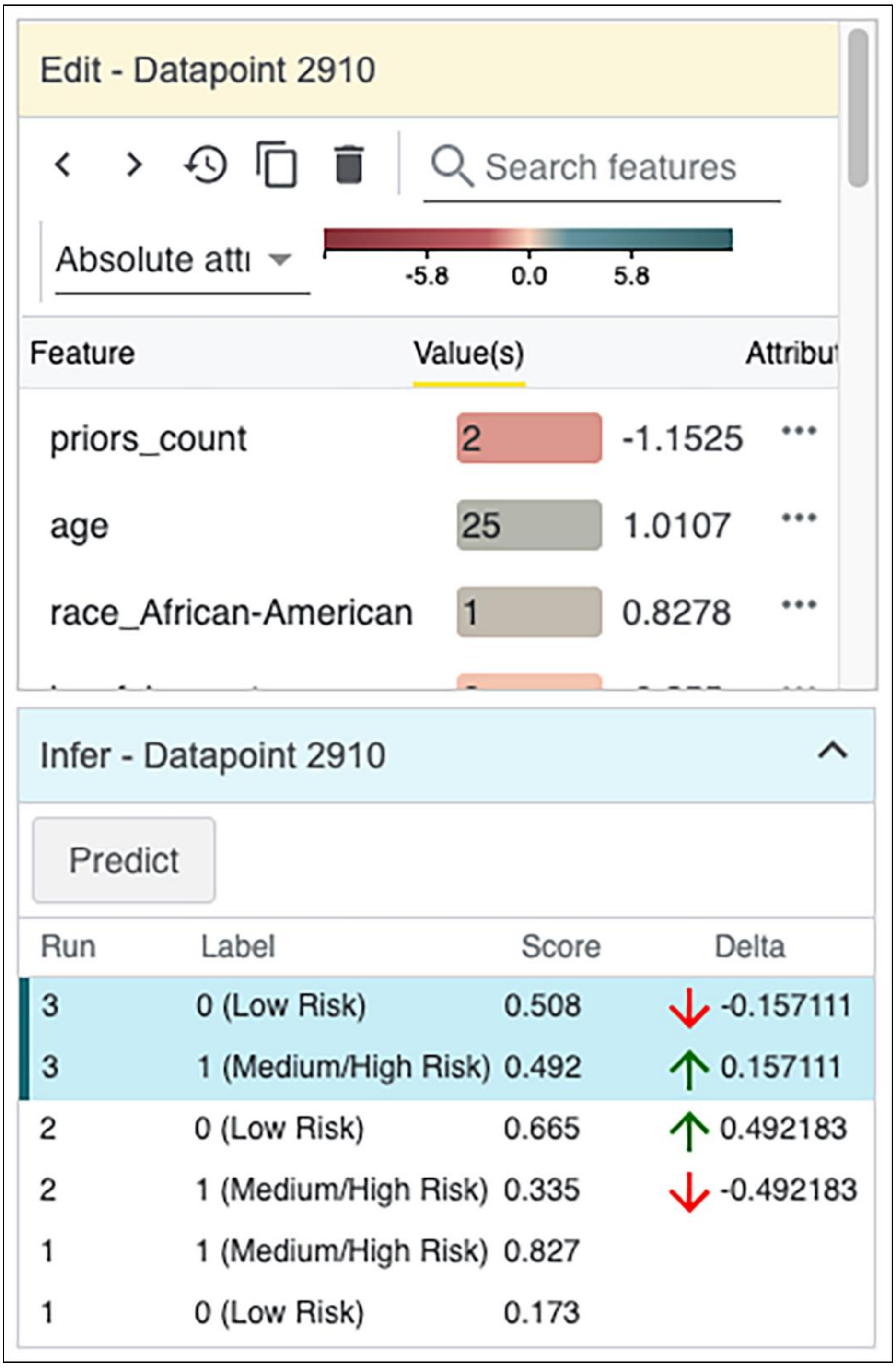


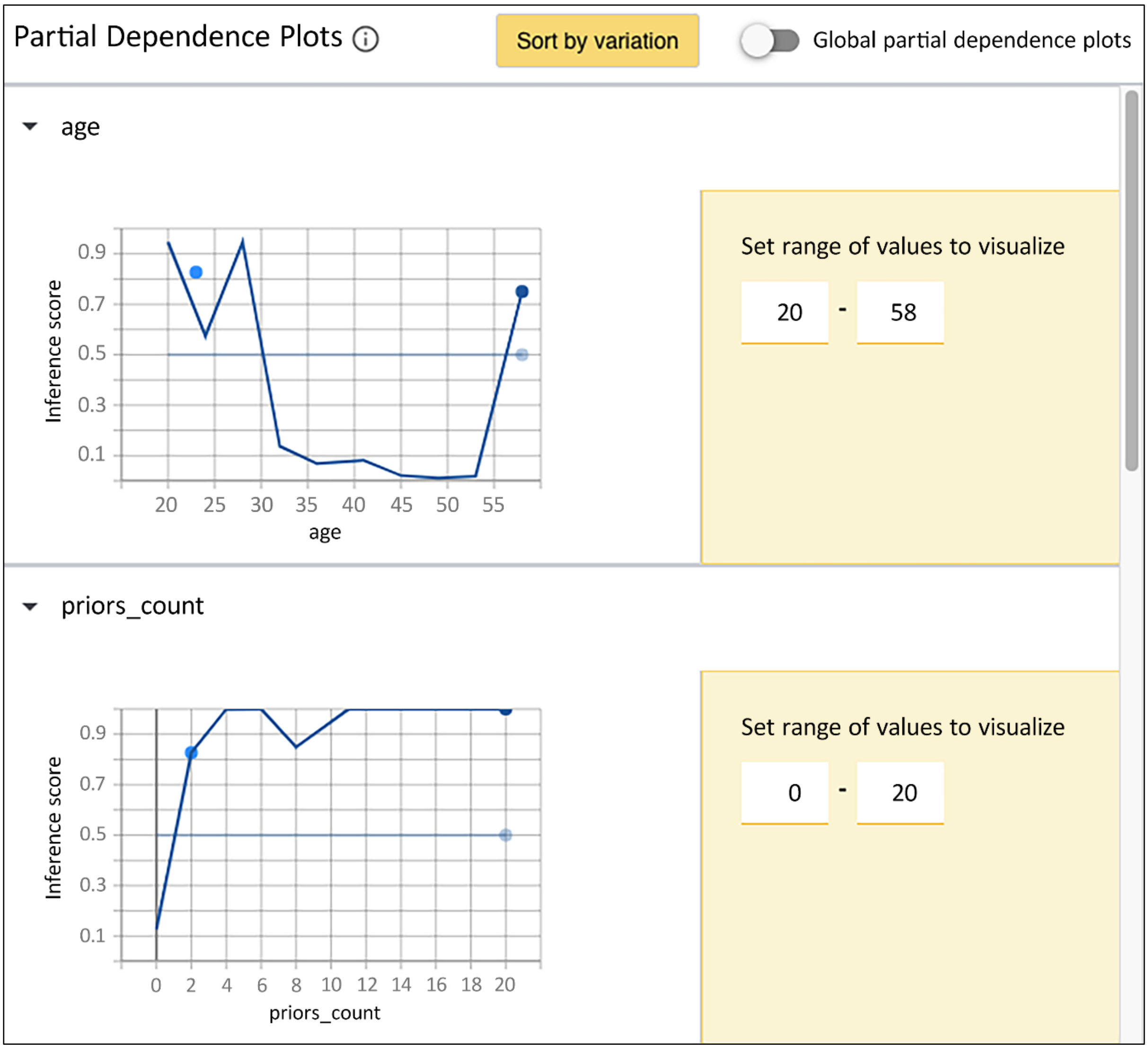
Datapoint editor
Performance & Fairness
Features
2958 datapoints loaded
Visualize
Datapoints
Partial dependence plots
Nearest counterfactual
L1
L2
Feature valu
Create similarity feature
Binning I X-A...
c_charge
A
Binning I Y-Axis
compas_
Color By
race_Afr
Label By
(default)
Scatter I X-Axis
(default)
Scatter I Y-Axis
(default)
Edit - Datapoint 2910
Search features
C
Descending
-6.3
0.0
6.3
Feature
Value(s)
Attribution value(s)
age 23 2.1998
race_African-American 1 0.6608
race_Hispanic 0 0.0905
race_Other 0 0.0704
race_Asian 0 0.0033
race_Native American 0 -0.0005
c_charge_degree_Other 0 -0.0083
Infer - Datapoint 2910
Predict
Run
Label
Score
Delta
1 1 (Medium/High Risk) 0.827
1 0 (Low Risk) 0.173
B
1
0
Legend
Colors
by race_African-American
0
1

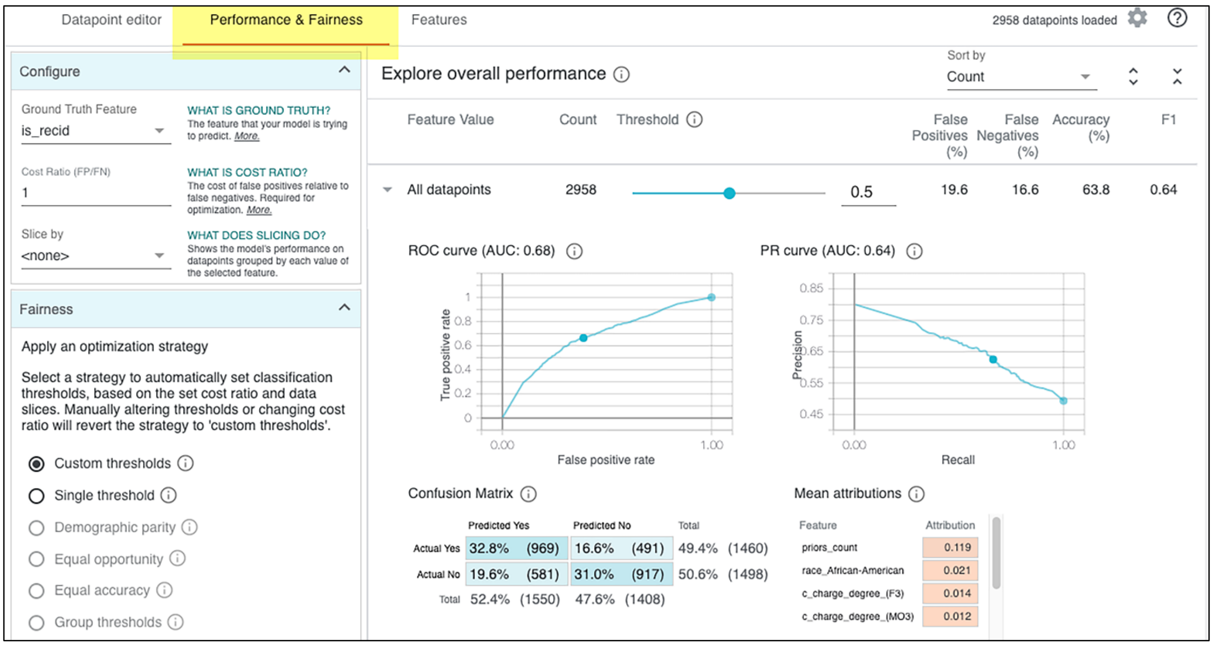

Visualize
Datapoints
Partial dependence plots
Nearest counterfactual
L1
L2
Feature valu
Create similarity feature
A
Edit - Datapoints 2910 and 2279
Search features
Absolute attı
-5.8
0.0
5.8
Feature
Value(s)
Counterfactual value(s)
age 23 26
race_African-American 1 1
priors_count 2 0
juv_fel_count 0 0
juv_other_count 0 0
race_Native American 0 0
B
Infer - Datapoints 2910 and 2279
Predict
C
Run Label Score Delta
1 1 (Medium/High Risk) 0.827
1 0 (Low Risk) 0.173
Run Label Score Delta
1 0 (Low Risk) 0.954
1 1 (Medium/High Risk) 0.046

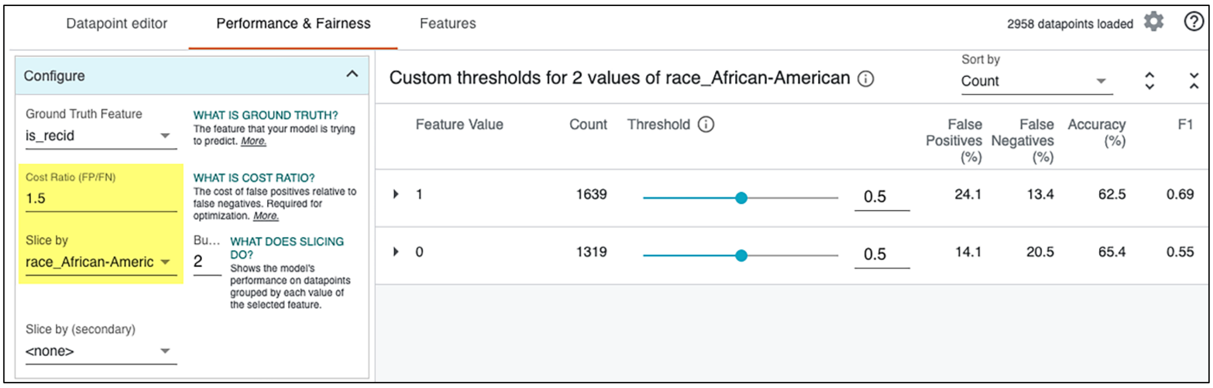

Binning | X-Axis
Inference
C..
Binning | Y-Axis
(none)
Color By
race_Afri
Label By
(default)
Scatter | X-Axis
age
Scatter | Y-Axis
priors_cc
Low Risk
Medium/High Risk
36
0
18
76
36
0
18
76
Legend
Colors
by race_African-American
0
1

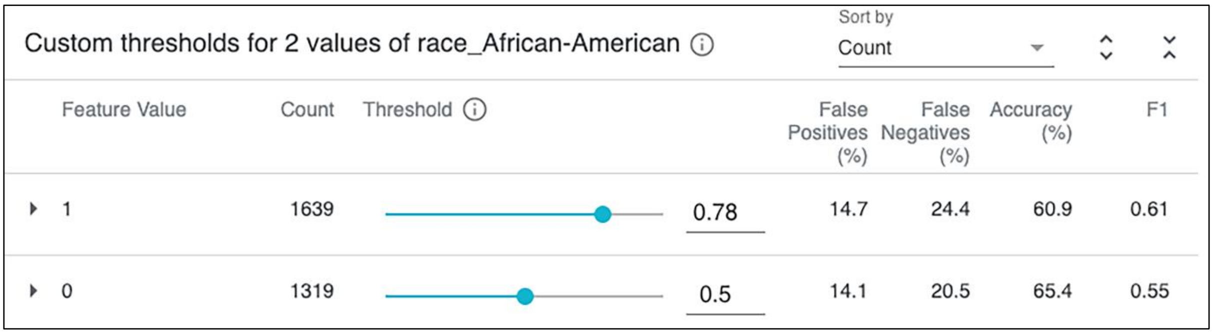

Edit - Datapoint 2910
Search features
Absolute attı
-5.8
0.0
5.8
Feature
Value(s)
Attribut
priors_count
1
-2.4239
age
23
1.5873
race_African-American
1
0.8297
race_Caucasian
0
0.3332
juv_fel_count
0
-0.2306
Infer - Datapoint 2910
Predict
Run
Label
Score
Delta
2
0 (Low Risk)
0.665
0.492183
2
1 (Medium/High Risk)
0.335
-0.492183
1
1 (Medium/High Risk)
0.827
1
0 (Low Risk)
0.173


Edit - Datapoint 2910
Search features
Absolute attı
-5.8
0.0
5.8
Feature
Value(s)
Attribut
priors_count
2
-1.1525
age
25
1.0107
race_African-American
1
0.8278
Infer - Datapoint 2910
Predict
Run
Label
Score
Delta
3 0 (Low Risk) 0.508 -0.157111
3 1 (Medium/High Risk) 0.492 0.157111
2 0 (Low Risk) 0.665 0.492183
2 1 (Medium/High Risk) 0.335 -0.492183
1 1 (Medium/High Risk) 0.827
1 0 (Low Risk) 0.173

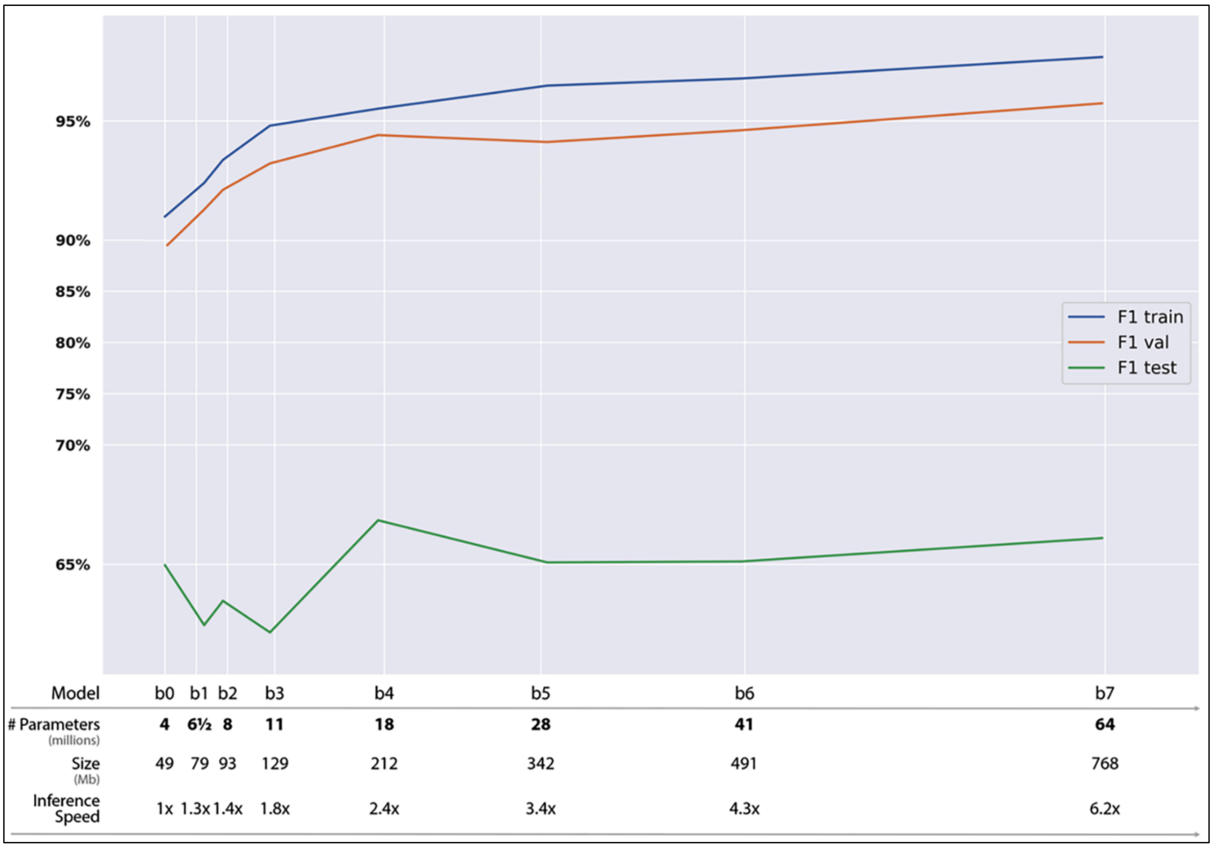

Partial Dependence Plots
Sort by variation
Global partial dependence plots
age
Inference score
age
Set range of values to visualize
20 - 58
priors_count
Inference score
priors_count
Set range of values to visualize
0 - 20

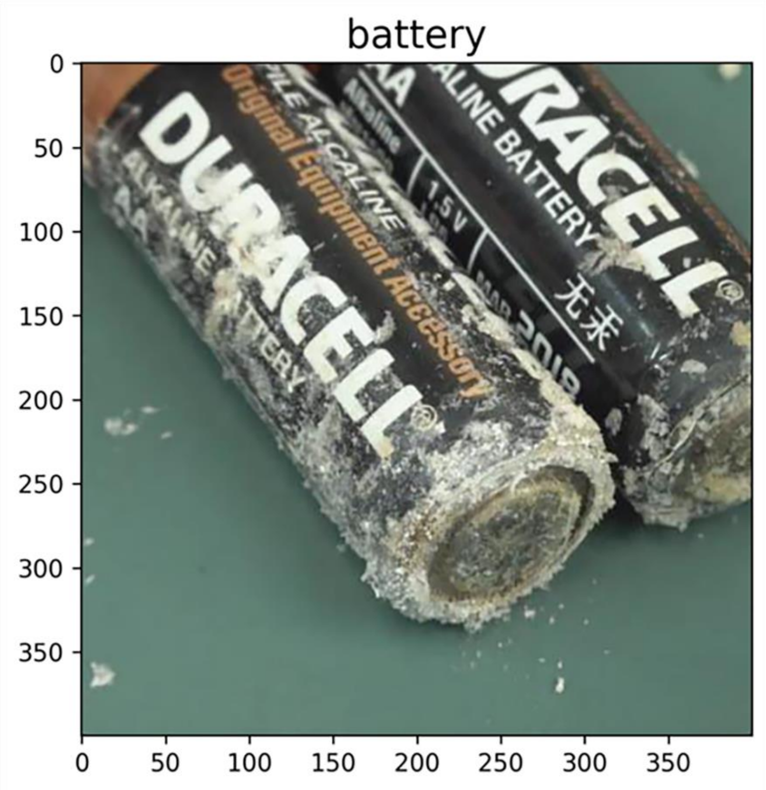

Datapoint editor
Performance & Fairness
Features
2958 datapoints loaded
Configure
Ground Truth Feature
is_recid
WHAT IS GROUND TRUTH?
The feature that your model is trying to predict. More.
Cost Ratio (FP/FN)
1
WHAT IS COST RATIO?
The cost of false positives relative to false negatives. Required for optimization. More.
Slice by
<none>
WHAT DOES SLICING DO?
Shows the model's performance on datapoints grouped by each value of the selected feature.
Fairness
Apply an optimization strategy
Select a strategy to automatically set classification thresholds, based on the set cost ratio and data slices. Manually altering thresholds or changing cost ratio will revert the strategy to 'custom thresholds'.
Custom thresholds
Single threshold
Demographic parity
Equal opportunity
Equal accuracy
Group thresholds
Explore overall performance
Sort by
Count
Feature Value
Count
Threshold
False Positives (%)
False Negatives (%)
Accuracy (%)
F1
All datapoints
2958
0.5
19.6
16.6
63.8
0.64
ROC curve (AUC: 0.68)
True positive rate
False positive rate
PR curve (AUC: 0.64)
Precision
Recall
Confusion Matrix
Predicted Yes
Predicted No
Total
Actual Yes 32.8% (969) 16.6% (491) 49.4% (1460)
Actual No 19.6% (581) 31.0% (917) 50.6% (1498)
Total 52.4% (1550) 47.6% (1408)
Mean attributions
Feature
Attribution
priors_count 0.119
race_African-American 0.021
c_charge_degree_(F3) 0.014
c_charge_degree_(MO3) 0.012

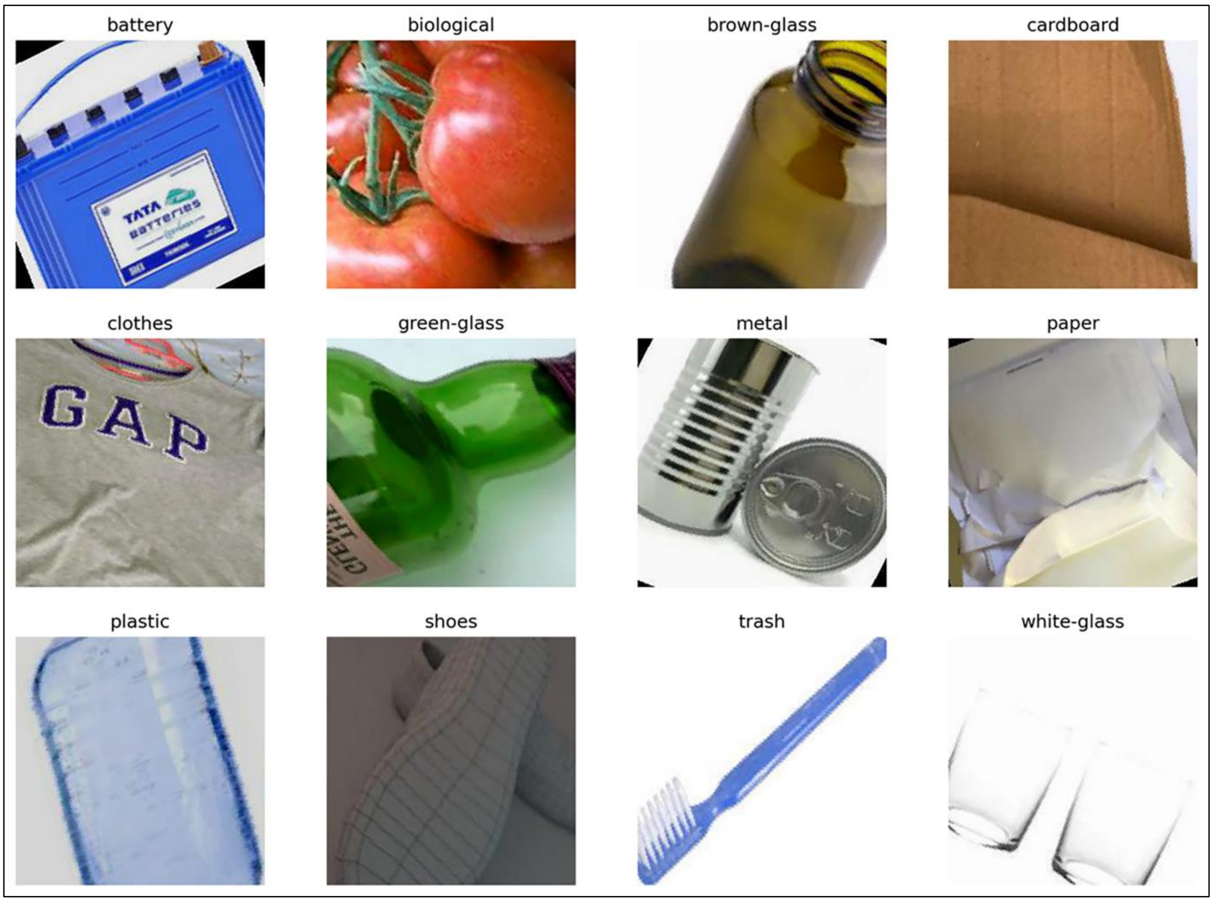

Datapoint editor
Performance & Fairness
Features
2958 datapoints loaded
Configure
Ground Truth Feature
is_recid
WHAT IS GROUND TRUTH?
The feature that your model is trying to predict. More.
Cost Ratio (FP/FN)
1.5
WHAT IS COST RATIO?
The cost of false positives relative to false negatives. Required for optimization. More.
Slice by
race_African-Americ
Bu...
2
WHAT DOES SLICING DO?
Shows the model's performance on datapoints grouped by each value of the selected feature.
Slice by (secondary)
<none>
Custom thresholds for 2 values of race_African-American
Sort by
Count
Feature Value
Count
Threshold
False Positives (%)
False Negatives (%)
Accuracy (%)
F1
1
1639
0.5
24.1
13.4
62.5
0.69
0
1319
0.5
14.1
20.5
65.4
0.55

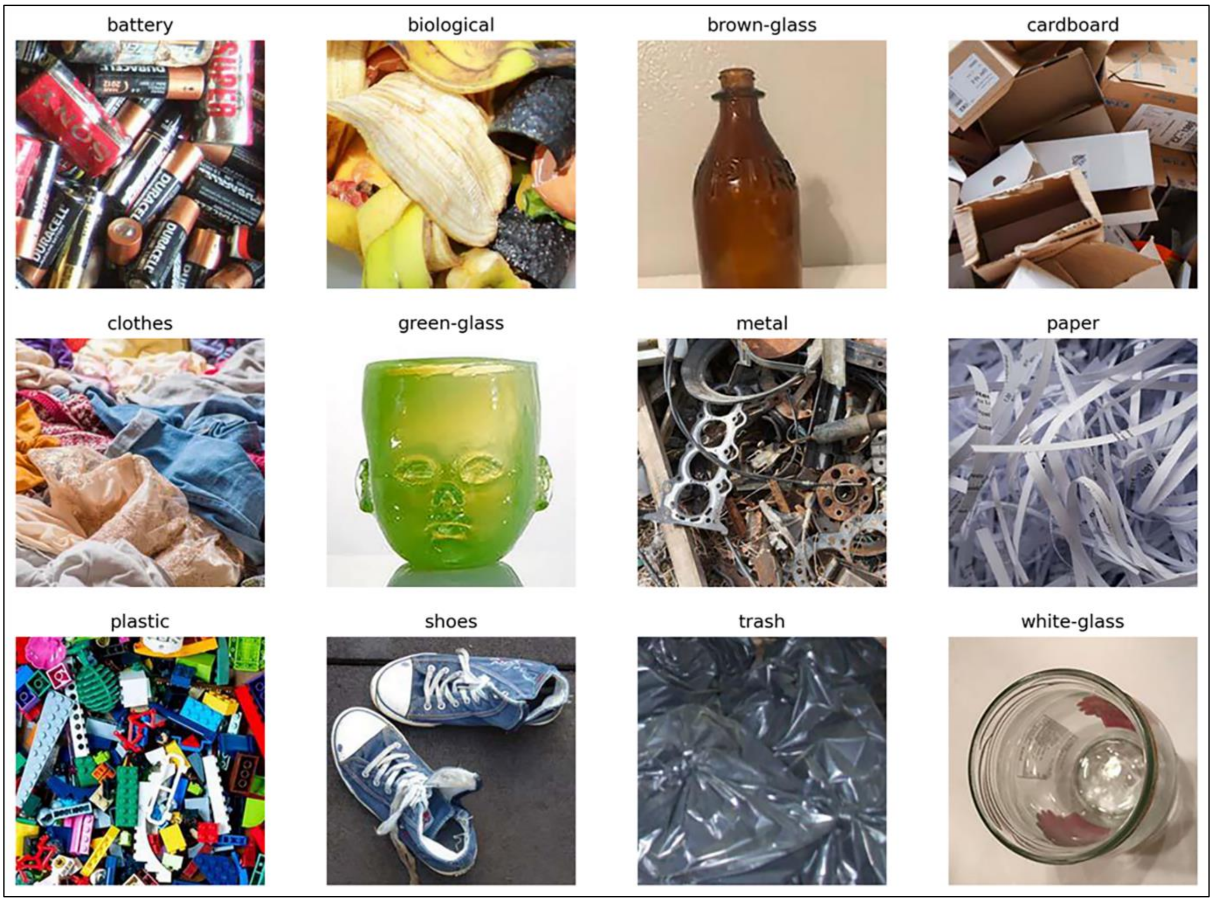

Custom thresholds for 2 values of race_African-American
Sort by
Count
Feature Value
Count
Threshold
False Positives (%)
False Negatives (%)
Accuracy (%)
F1
1
1639
0.78
14.7
24.4
60.9
0.61
0
1319
0.5
14.1
20.5
65.4
0.55

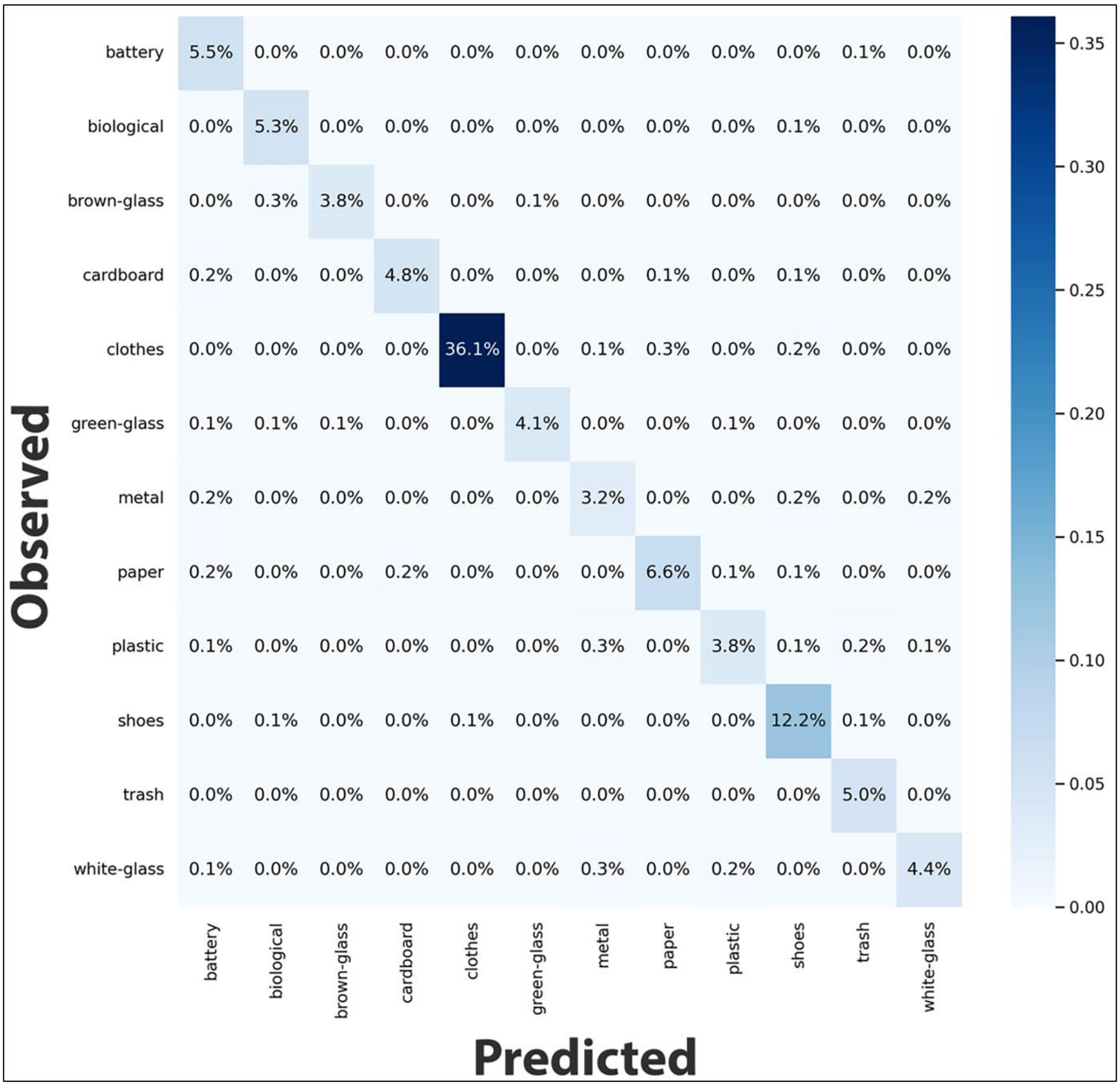
# Chapter 7: Visualizing Convolutional Neural Networks



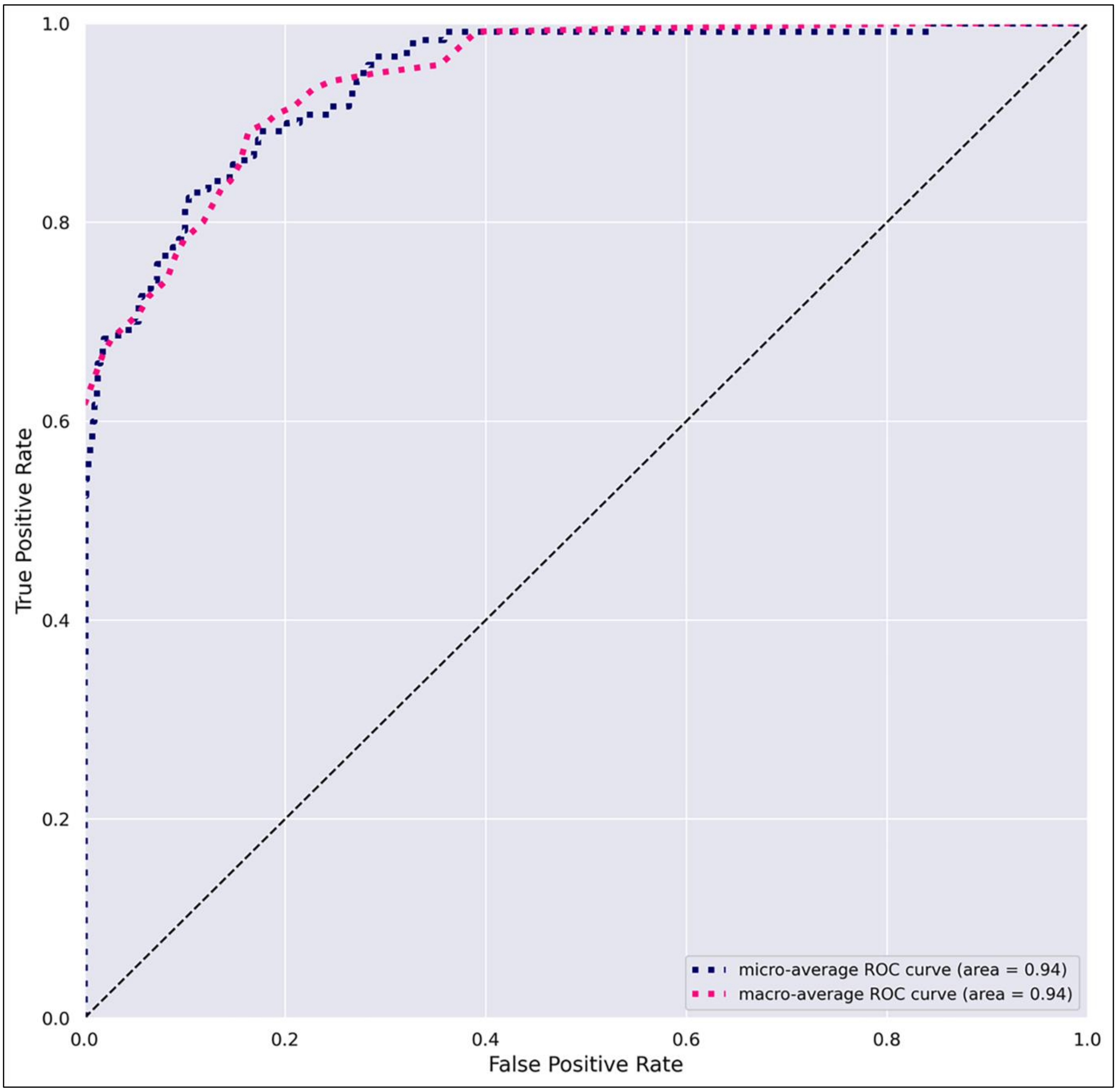

battery
biological
brown-glass
cardboard
clothes
green-glass
metal
paper
plastic
shoes
trash
white-glass


battery
biological
brown-glass
cardboard
clothes
green-glass
metal
paper
plastic
shoes
trash
white-glass

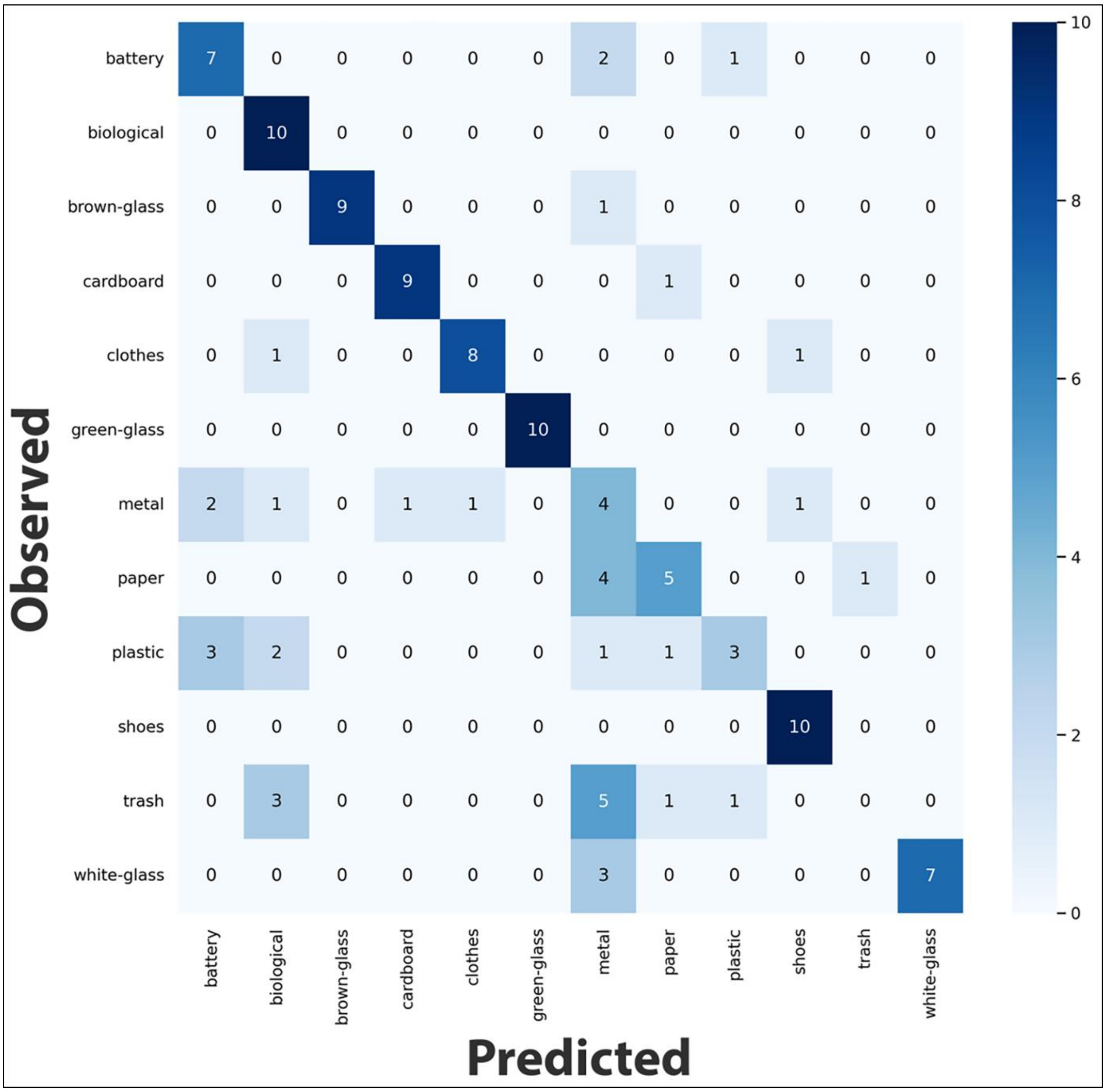


Observed
battery 5.5% 0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 0.1% 0.0%
biological 0.0% 5.3% 0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 0.1% 0.0% 0.0%
brown-glass 0.0% 0.3% 3.8% 0.0% 0.0% 0.1% 0.0% 0.0% 0.0% 0.0% 0.0% 0.0%
cardboard 0.2% 0.0% 0.0% 4.8% 0.0% 0.0% 0.0% 0.1% 0.0% 0.1% 0.0% 0.0%
clothes 0.0% 0.0% 0.0% 0.0% 36.1% 0.0% 0.1% 0.3% 0.0% 0.2% 0.0% 0.0%
green-glass 0.1% 0.1% 0.1% 0.0% 0.0% 4.1% 0.0% 0.0% 0.1% 0.0% 0.0% 0.0%
metal 0.2% 0.0% 0.0% 0.0% 0.0% 0.0% 3.2% 0.0% 0.0% 0.2% 0.0% 0.2%
paper 0.2% 0.0% 0.0% 0.2% 0.0% 0.0% 0.0% 6.6% 0.1% 0.1% 0.0% 0.0%
plastic 0.1% 0.0% 0.0% 0.0% 0.0% 0.0% 0.3% 0.0% 3.8% 0.1% 0.2% 0.1%
shoes 0.0% 0.1% 0.0% 0.0% 0.1% 0.0% 0.0% 0.0% 0.0% 12.2% 0.1% 0.0%
trash 0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 0.0% 5.0% 0.0%
white-glass 0.1% 0.0% 0.0% 0.0% 0.0% 0.0% 0.3% 0.0% 0.2% 0.0% 0.0% 4.4%
battery
biological
brown-glass
cardboard
clothes
green-glass
metal
paper
plastic
shoes
trash
white-glass
Predicted
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| battery | 0.850 | 0.981 | 0.911 | 52 |
| biological | 0.907 | 0.980 | 0.942 | 50 |
| brown-glass | 0.972 | 0.897 | 0.933 | 39 |
| cardboard | 0.957 | 0.918 | 0.938 | 49 |
| clothes | 0.997 | 0.982 | 0.990 | 342 |
| green-glass | 0.974 | 0.905 | 0.938 | 42 |
| metal | 0.811 | 0.833 | 0.822 | 36 |
| paper | 0.938 | 0.910 | 0.924 | 67 |
| plastic | 0.897 | 0.814 | 0.854 | 43 |
| shoes | 0.934 | 0.974 | 0.954 | 117 |
| trash | 0.922 | 1.000 | 0.959 | 47 |
| white-glass | 0.932 | 0.872 | 0.901 | 47 |
| accuracy | | | 0.947 | 931 |
| macro avg | 0.924 | 0.922 | 0.922 | 931 |
| weighted avg | 0.949 | 0.947 | 0.947 | 931 |



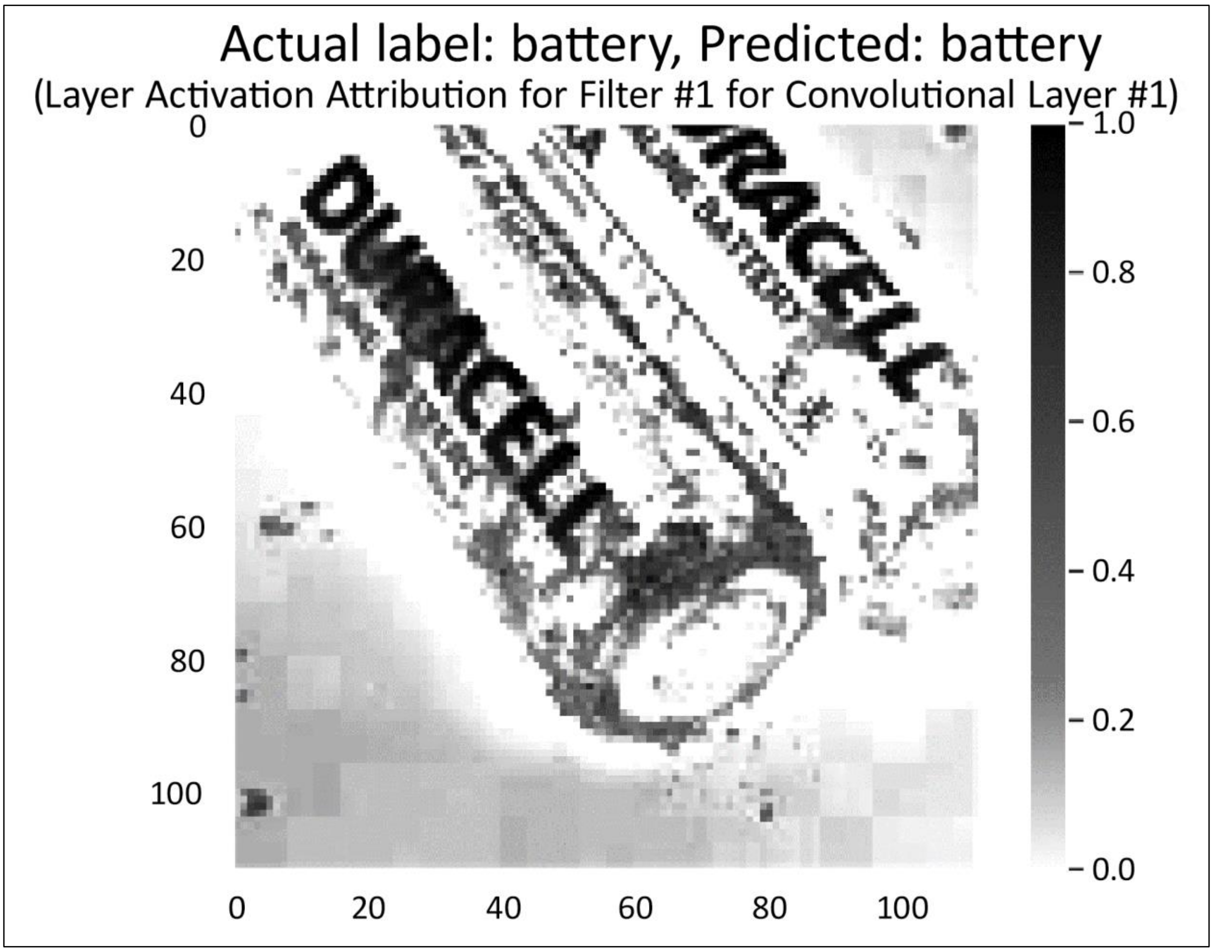
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
micro-average ROC curve (area = 0.94)
macro-average ROC curve (area = 0.94)

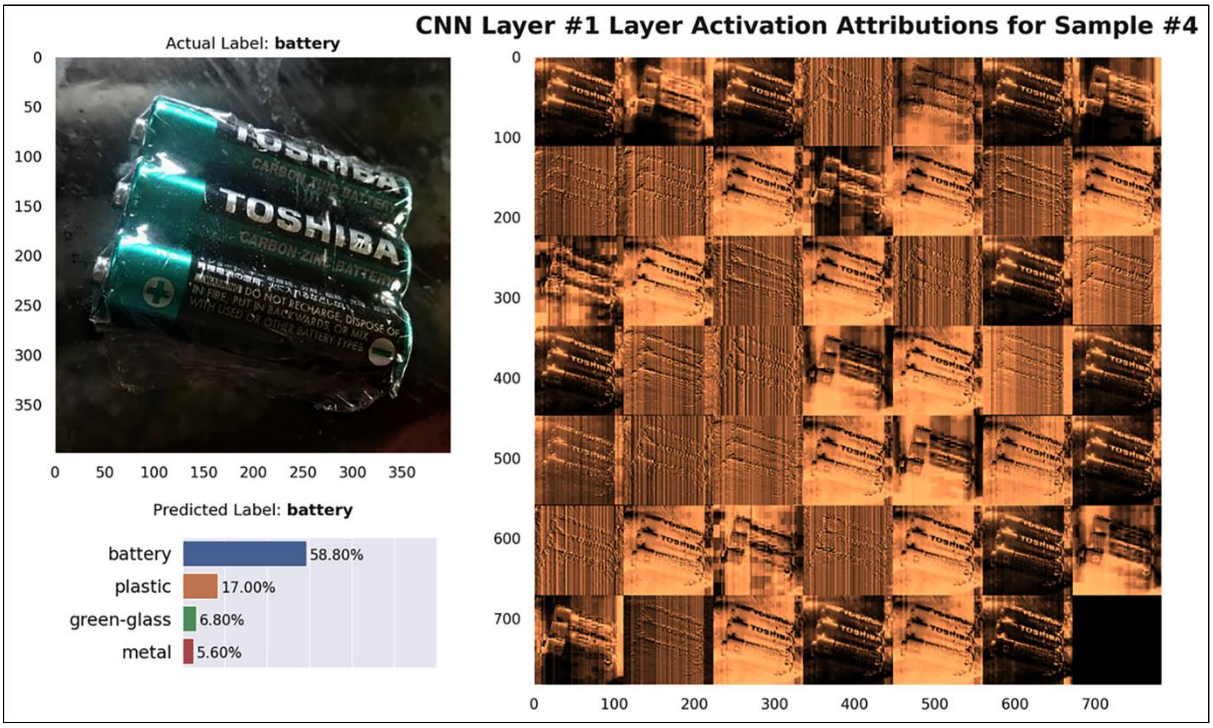


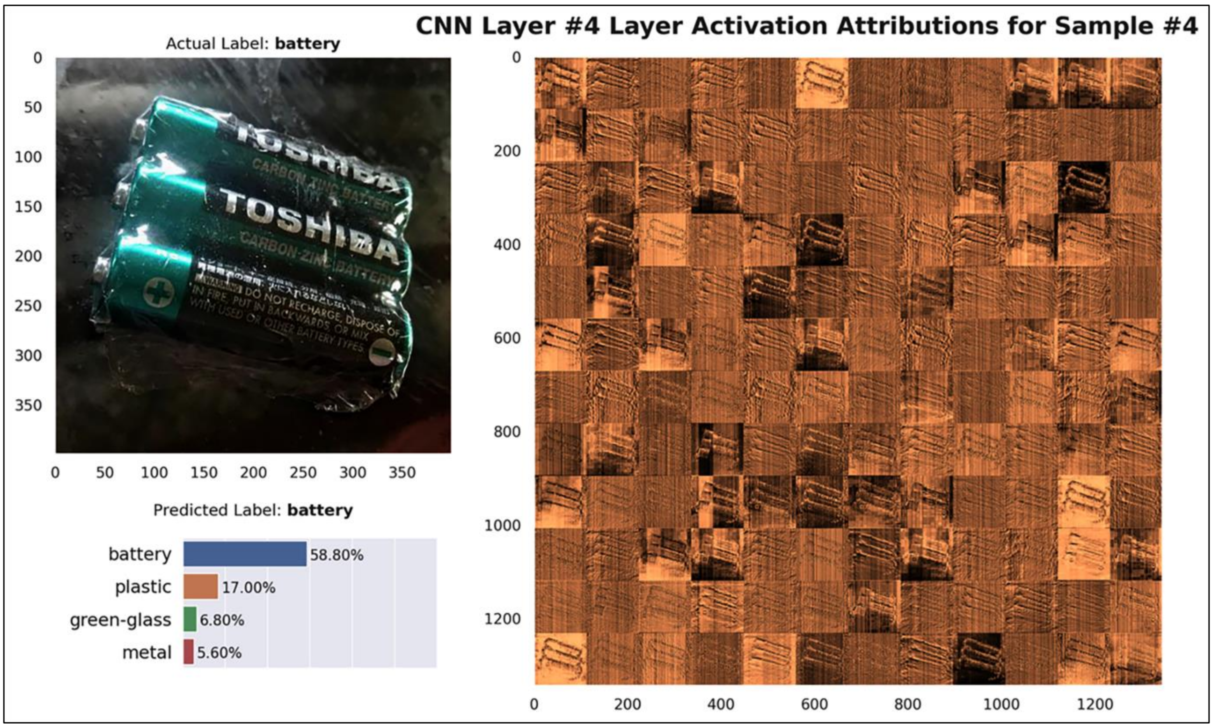
Observed
Predicted
battery
biological
brown-glass
cardboard
clothes
green-glass
metal
paper
plastic
shoes
trash
white-glass
7 0 0 0 0 0 2 0 1 0 0 0
0 10 0 0 0 0 0 0 0 0 0 0
0 0 9 0 0 0 1 0 0 0 0 0
0 0 0 9 0 0 0 1 0 0 0 0
0 1 0 0 8 0 0 0 0 1 0 0
0 0 0 0 0 10 0 0 0 0 0 0
2 1 0 1 1 0 4 0 0 1 0 0
0 0 0 0 0 0 4 5 0 0 1 0
3 2 0 0 0 0 1 1 3 0 0 0
0 0 0 0 0 0 0 0 0 10 0 0
0 3 0 0 0 0 5 1 1 0 0 0
0 0 0 0 0 0 3 0 0 0 0 7
10
8
6
4
2
0

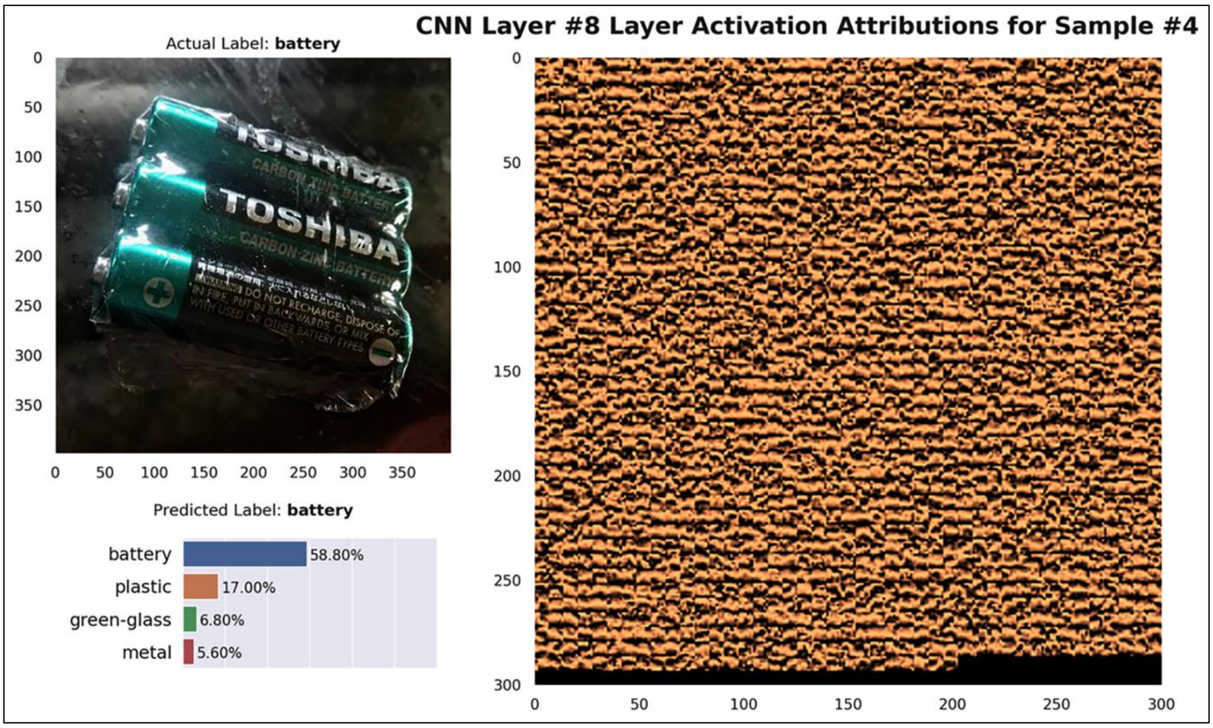
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| battery | 0.583 | 0.700 | 0.636 | 10 |
| biological | 0.588 | 1.000 | 0.741 | 10 |
| brown-glass | 1.000 | 0.900 | 0.947 | 10 |
| cardboard | 0.900 | 0.900 | 0.900 | 10 |
| clothes | 0.889 | 0.800 | 0.842 | 10 |
| green-glass | 1.000 | 1.000 | 1.000 | 10 |
| metal | 0.200 | 0.400 | 0.267 | 10 |
| paper | 0.625 | 0.500 | 0.556 | 10 |
| plastic | 0.600 | 0.300 | 0.400 | 10 |
| shoes | 0.833 | 1.000 | 0.909 | 10 |
| trash | 0.000 | 0.000 | 0.000 | 10 |
| white-glass | 1.000 | 0.700 | 0.824 | 10 |
| | | | | |
| accuracy | | | 0.683 | 120 |
| macro avg | 0.685 | 0.683 | 0.668 | 120 |
| weighted avg | 0.685 | 0.683 | 0.668 | 120 |

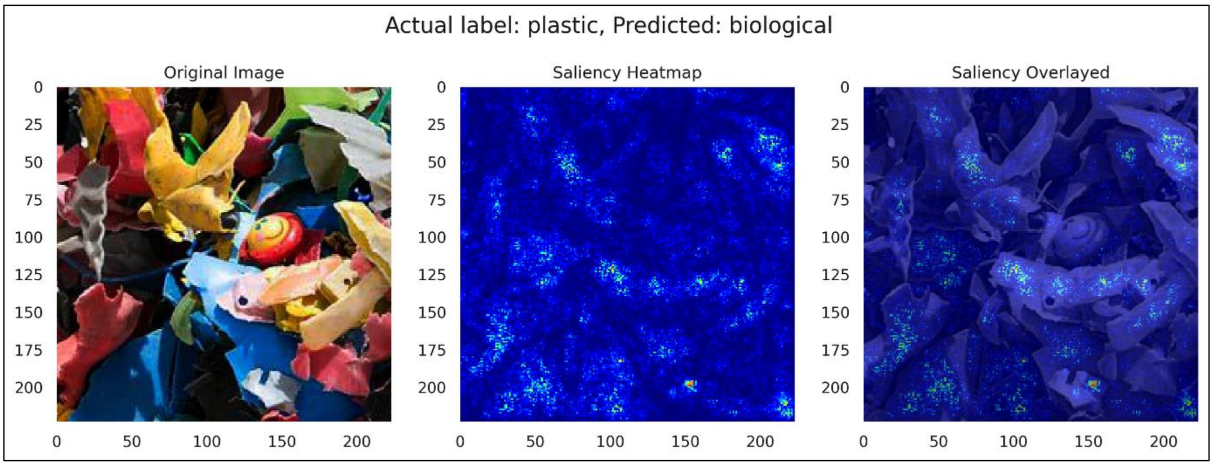
| | y_true | y_pred | biological | metal | shoes | battery | paper | clothes | white-glass |
|---|---|---|---|---|---|---|---|---|---|
| **0** | battery | battery | 0.00% | 0.10% | 0.00% | **99.70%** | 0.10% | 0.00% | 0.00% |
| **1** | battery | battery | 0.10% | 1.50% | 0.30% | **96.70%** | 0.30% | 0.00% | 0.00% |
| **2** | battery | battery | 0.00% | 0.10% | 0.00% | **99.60%** | 0.20% | 0.00% | 0.00% |
| **3** | battery | battery | 0.30% | 0.40% | 0.10% | **97.40%** | 1.00% | 0.00% | 0.00% |
| **4** | battery | battery | 1.00% | 5.60% | 0.60% | **58.80%** | 1.20% | 0.10% | 1.90% |
| **5** | battery | battery | 0.30% | 1.00% | 0.50% | **94.70%** | 0.50% | 0.00% | 0.20% |
| **6** | battery | battery | 0.00% | 0.40% | 0.00% | **99.00%** | 0.30% | 0.00% | 0.00% |
| **7** | battery | metal | 0.40% | **82.40%** | 2.10% | 1.10% | 1.50% | 1.10% | 3.60% |
| **8** | battery | metal | 0.00% | **92.60%** | 0.10% | 1.60% | 0.60% | 0.00% | 0.30% |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| **65** | metal | clothes | 2.90% | 5.20% | 2.70% | 3.90% | 4.80% | **69.60%** | 1.20% |
| **66** | metal | battery | 1.30% | 20.40% | 2.00% | **64.30%** | 2.50% | 0.20% | 0.60% |
| **67** | metal | cardboard | 10.70% | 6.20% | 3.20% | 9.60% | 13.80% | 1.00% | 1.10% |
| **68** | metal | shoes | 4.10% | 14.20% | 43.70% | 1.80% | 28.70% | 1.50% | 0.90% |
| **69** | metal | battery | 2.60% | 8.30% | 18.90% | **57.60%** | 8.00% | 0.60% | 0.20% |
| **73** | paper | metal | 1.30% | **74.80%** | 9.20% | 2.50% | 5.40% | 1.70% | 1.40% |
| **77** | paper | metal | 3.40% | 29.40% | 5.20% | 2.40% | 7.70% | 1.50% | 15.30% |
| **83** | plastic | battery | 3.90% | 5.00% | 7.80% | 46.70% | 10.00% | 1.30% | 0.70% |
| **84** | plastic | metal | 11.10% | 19.00% | 2.50% | 8.10% | 13.50% | 15.10% | 4.40% |
| **85** | plastic | battery | 4.20% | 5.20% | 5.20% | 36.30% | 27.90% | 0.30% | 0.80% |
| **86** | plastic | biological | 36.70% | 2.80% | 10.90% | 6.40% | 18.60% | 1.40% | 1.50% |
| **87** | plastic | paper | 1.80% | 1.90% | 0.90% | 5.20% | **74.10%** | 0.60% | 1.10% |
| **88** | plastic | biological | 41.10% | 2.30% | 6.00% | 3.30% | 21.90% | 1.90% | 1.90% |
| **89** | plastic | battery | 1.20% | 1.70% | 0.40% | **88.60%** | 2.00% | 0.10% | 0.20% |
| **100** | trash | biological | 49.90% | 5.00% | 4.10% | 10.00% | 5.30% | 4.60% | 1.70% |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| **109** | trash | metal | 19.30% | 24.80% | 17.50% | 7.20% | 4.30% | 1.30% | 2.40% |
| **110** | white-glass | white-glass | 0.20% | 3.40% | 0.40% | 0.20% | 0.20% | 0.10% | **90.70%** |
| **111** | white-glass | white-glass | 0.00% | 1.60% | 0.50% | 0.00% | 0.10% | 0.00% | **95.80%** |
| **112** | white-glass | white-glass | 0.00% | 0.10% | 0.00% | 0.00% | 0.00% | 0.00% | **99.30%** |
| **113** | white-glass | white-glass | 0.00% | 0.10% | 0.00% | 0.00% | 0.00% | 0.00% | **95.50%** |
| **114** | white-glass | metal | 0.10% | **82.50%** | 0.10% | 2.20% | 0.20% | 0.00% | 1.50% |
| **115** | white-glass | metal | 0.10% | **88.70%** | 0.40% | 0.50% | 0.30% | 0.00% | 5.70% |
| **116** | white-glass | metal | 3.70% | 41.90% | 3.10% | 3.30% | 4.00% | 1.00% | 10.40% |
| **117** | white-glass | white-glass | 0.10% | 1.60% | 0.00% | 0.10% | 0.10% | 0.00% | **94.90%** |
| **118** | white-glass | white-glass | 0.10% | 0.30% | 0.10% | 0.00% | 0.00% | 0.00% | **97.40%** |
| **119** | white-glass | white-glass | 0.00% | 0.40% | 0.10% | 0.00% | 0.10% | 0.00% | **95.70%** |

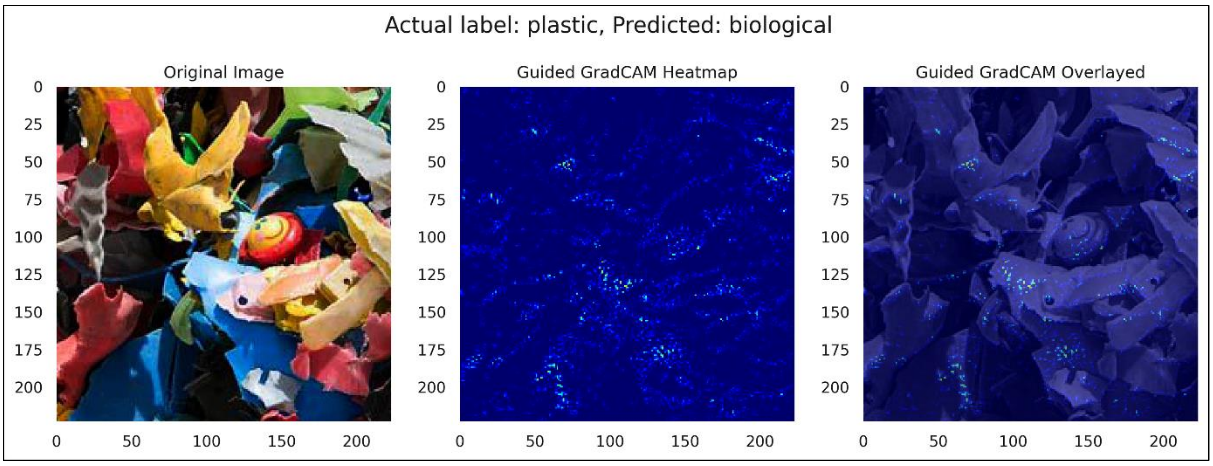

Actual label: battery, Predicted: battery
(Layer Activation Attribution for Filter #1 for Convolutional Layer #1)

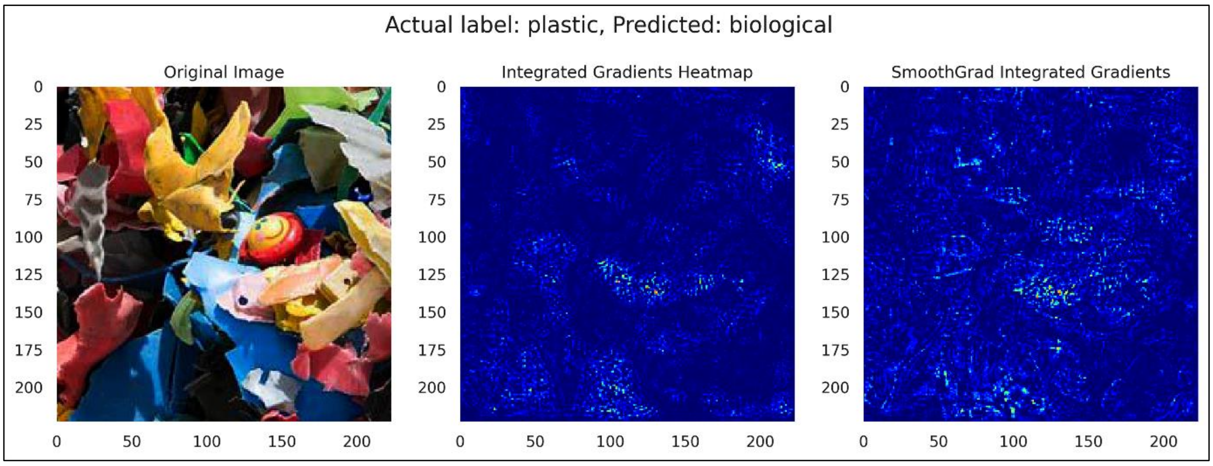

CNN Layer #1 Layer Activation Attributions for Sample #4
Actual Label: battery
Predicted Label: battery
battery 58.80%
plastic 17.00%
green-glass 6.80%
metal 5.60%

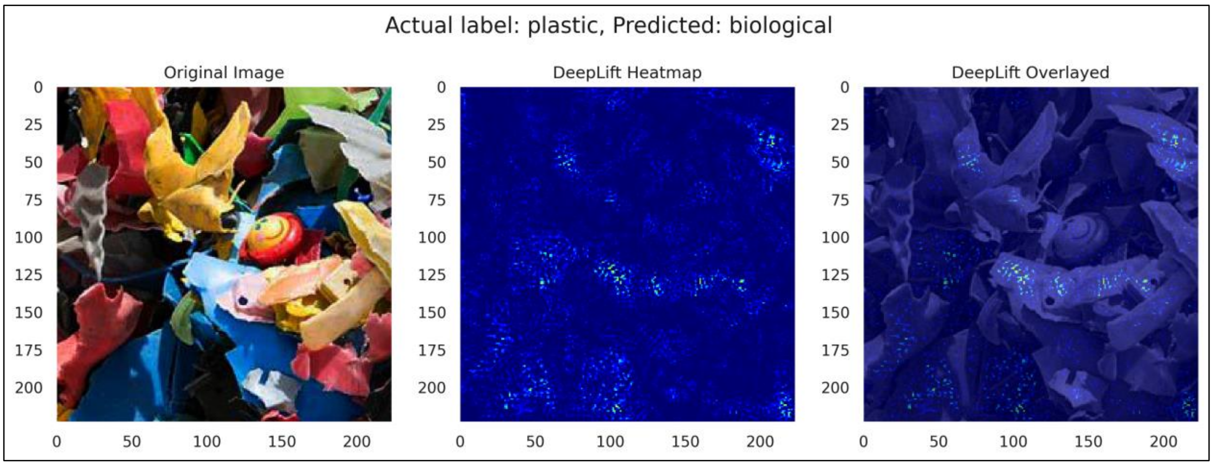

CNN Layer #4 Layer Activation Attributions for Sample #4
Actual Label: battery
Predicted Label: battery
battery 58.80%
plastic 17.00%
green-glass 6.80%
metal 5.60%

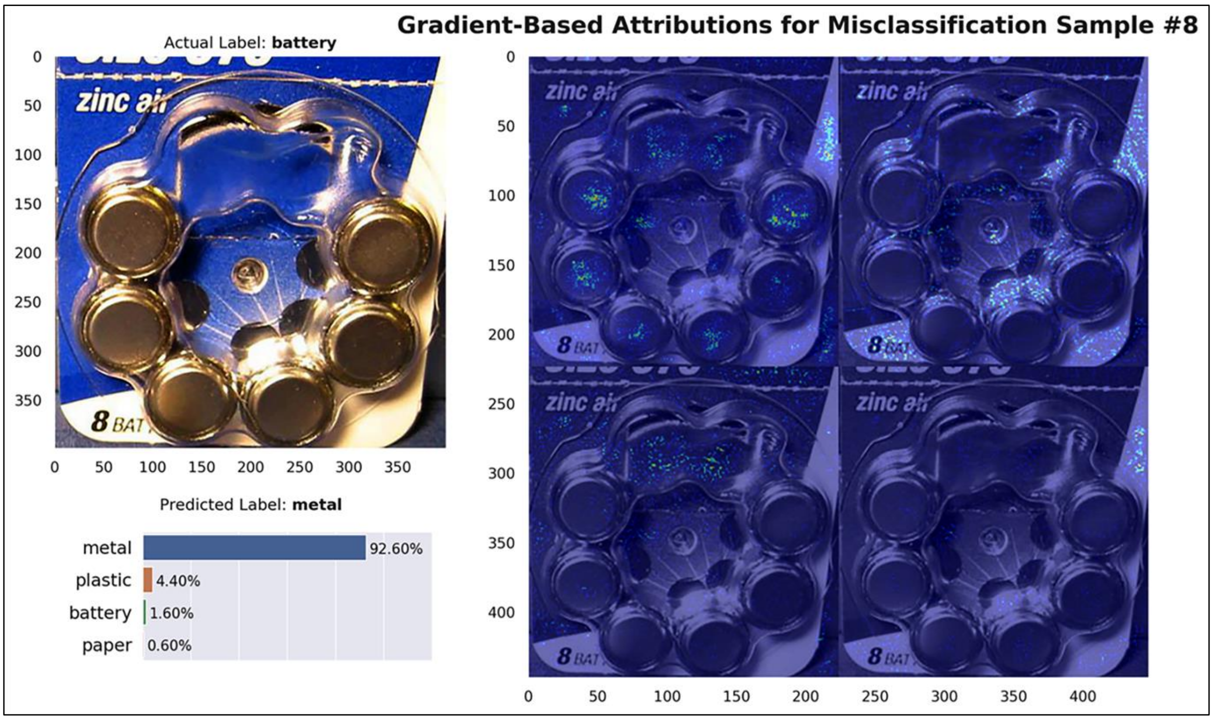

CNN Layer #8 Layer Activation Attributions for Sample #4
Actual Label: battery
Predicted Label: battery
battery 58.80%
plastic 17.00%
green-glass 6.80%
metal 5.60%

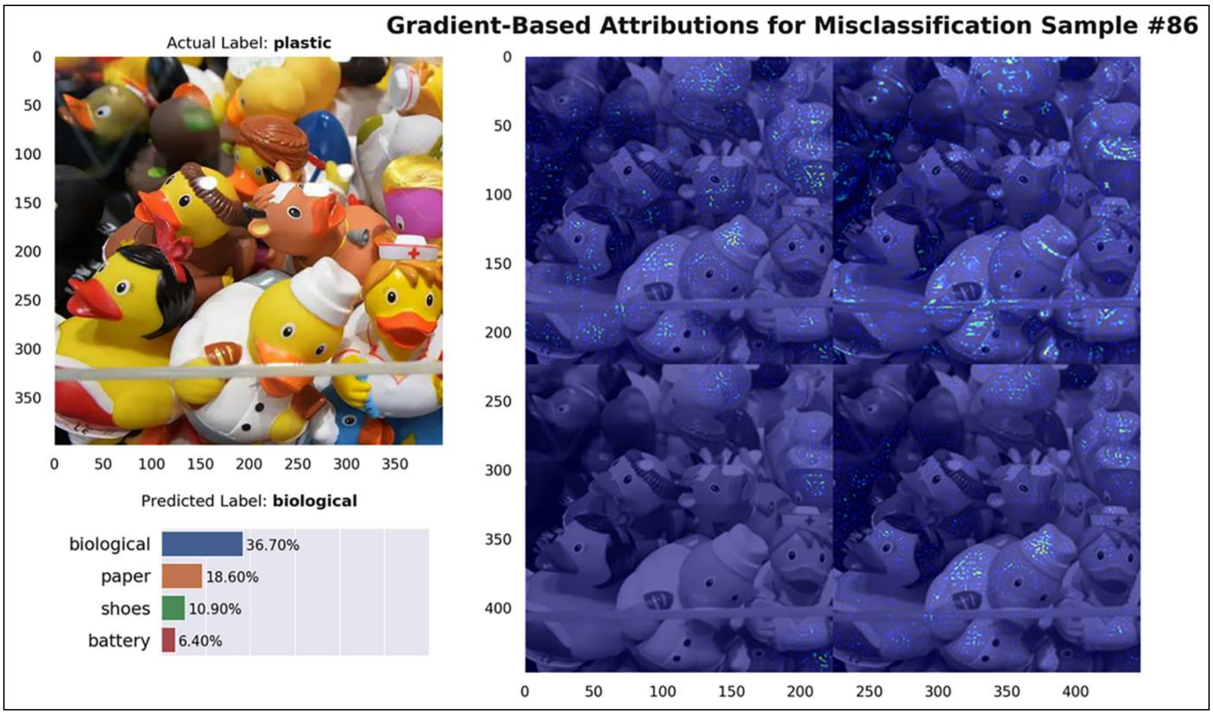

Actual label: plastic, Predicted: biological
Original Image
Saliency Heatmap
Saliency Overlayed
0
25
50
75
100
125
150
175
200
0
50
100
150
200

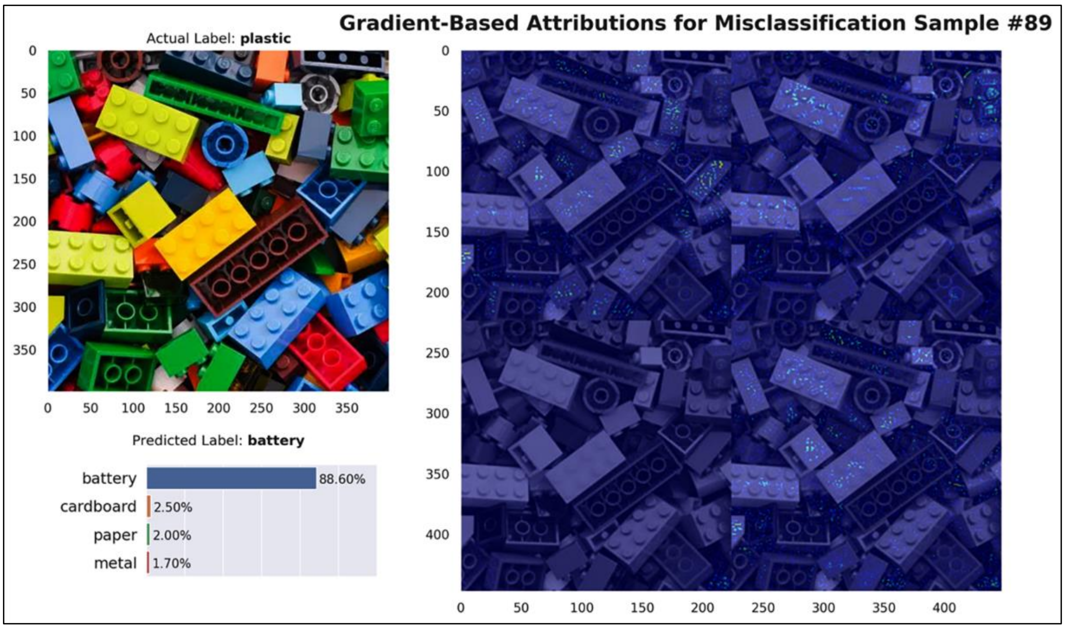

Actual label: plastic, Predicted: biological
Original Image
Guided GradCAM Heatmap
Guided GradCAM Overlayed
0
25
50
75
100
125
150
175
200
0
50
100
150
200

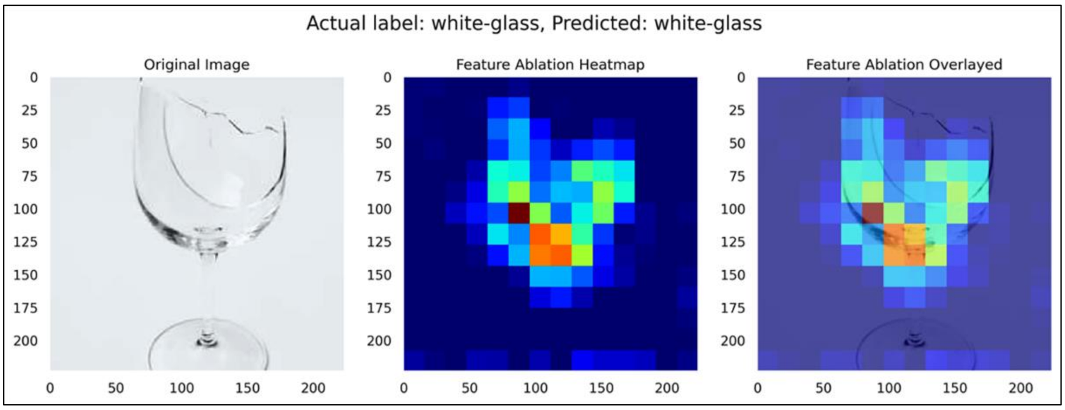

Actual label: plastic, Predicted: biological
Original Image
Integrated Gradients Heatmap
SmoothGrad Integrated Gradients
0
25
50
75
100
125
150
175
200
0
50
100
150
200

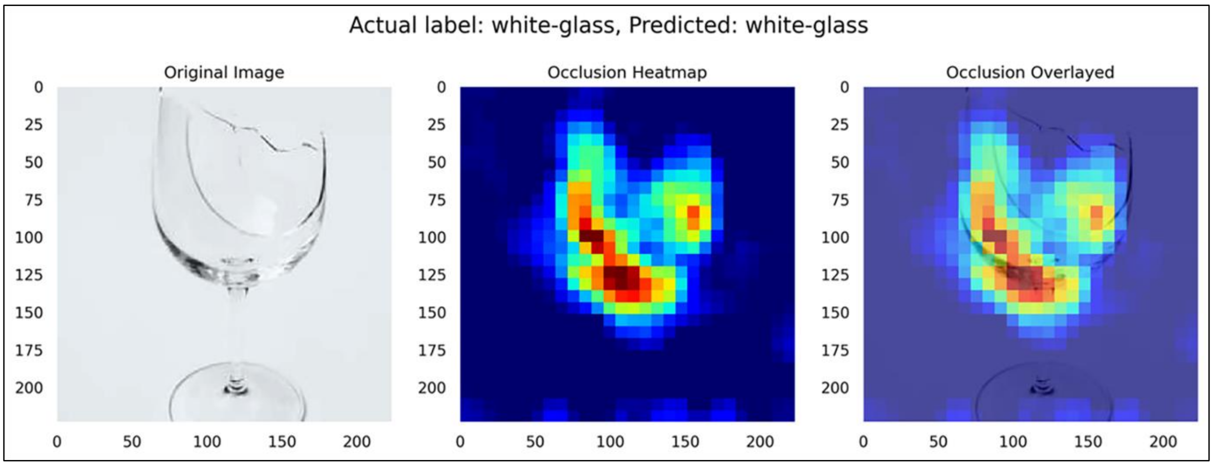

Actual label: plastic, Predicted: biological
Original Image
DeepLift Heatmap
DeepLift Overlayed

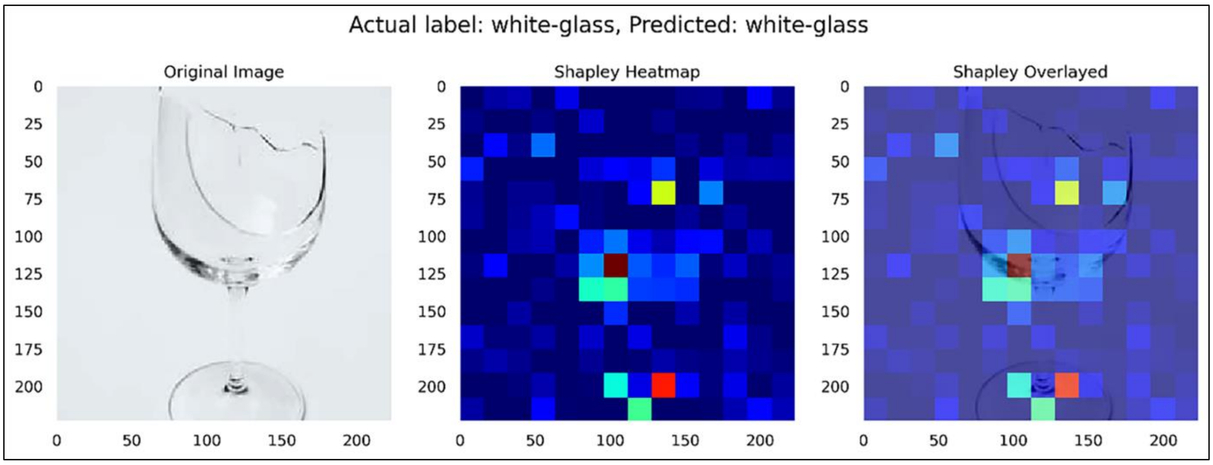

Gradient-Based Attributions for Misclassification Sample #8
Actual Label: battery
zinc air
8 BAT
Predicted Label: metal
metal 92.60%
plastic 4.40%
battery 1.60%
paper 0.60%

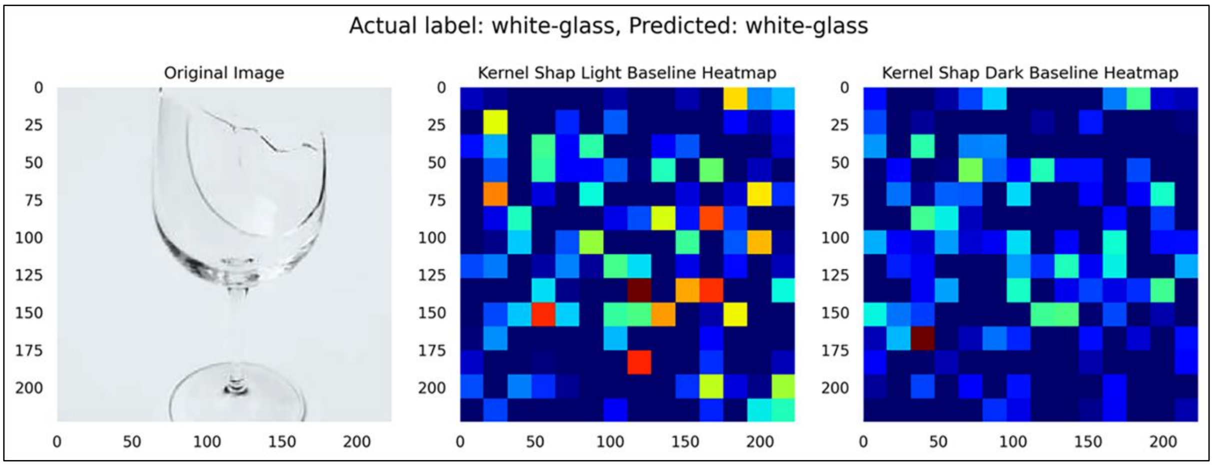

Gradient-Based Attributions for Misclassification Sample #86
Actual Label: plastic
Predicted Label: biological
biological 36.70%
paper 18.60%
shoes 10.90%
battery 6.40%

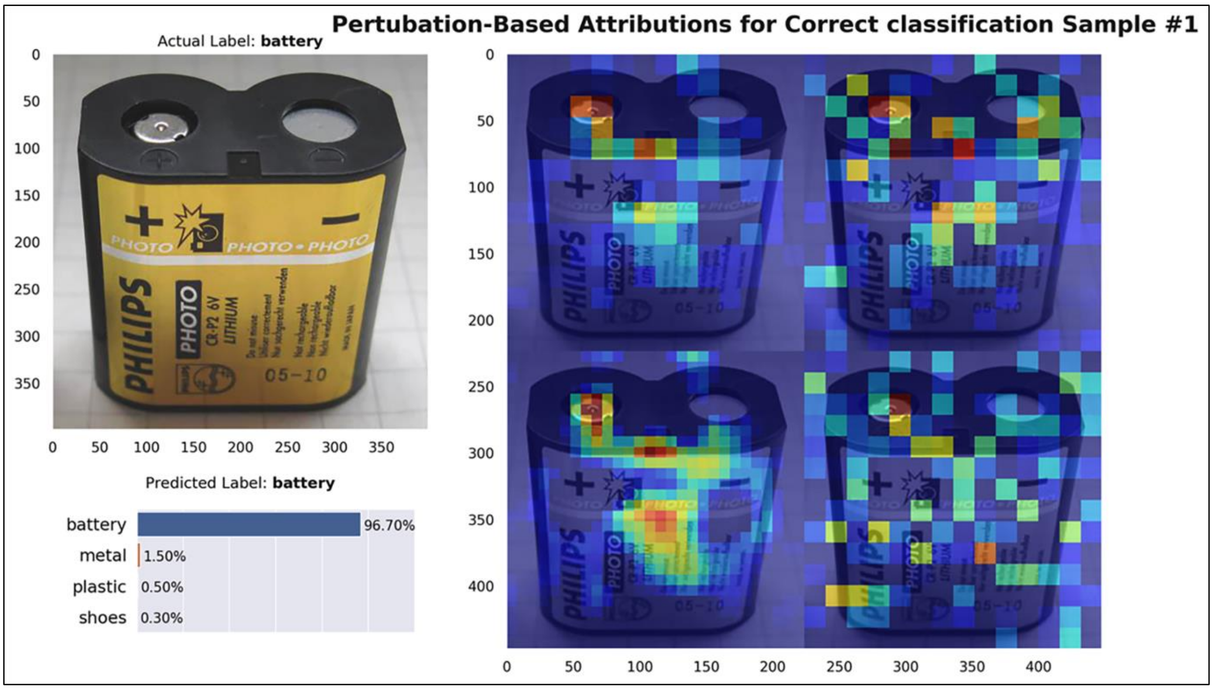

Gradient-Based Attributions for Misclassification Sample #89
Actual Label: plastic
Predicted Label: battery
battery 88.60%
cardboard 2.50%
paper 2.00%
metal 1.70%

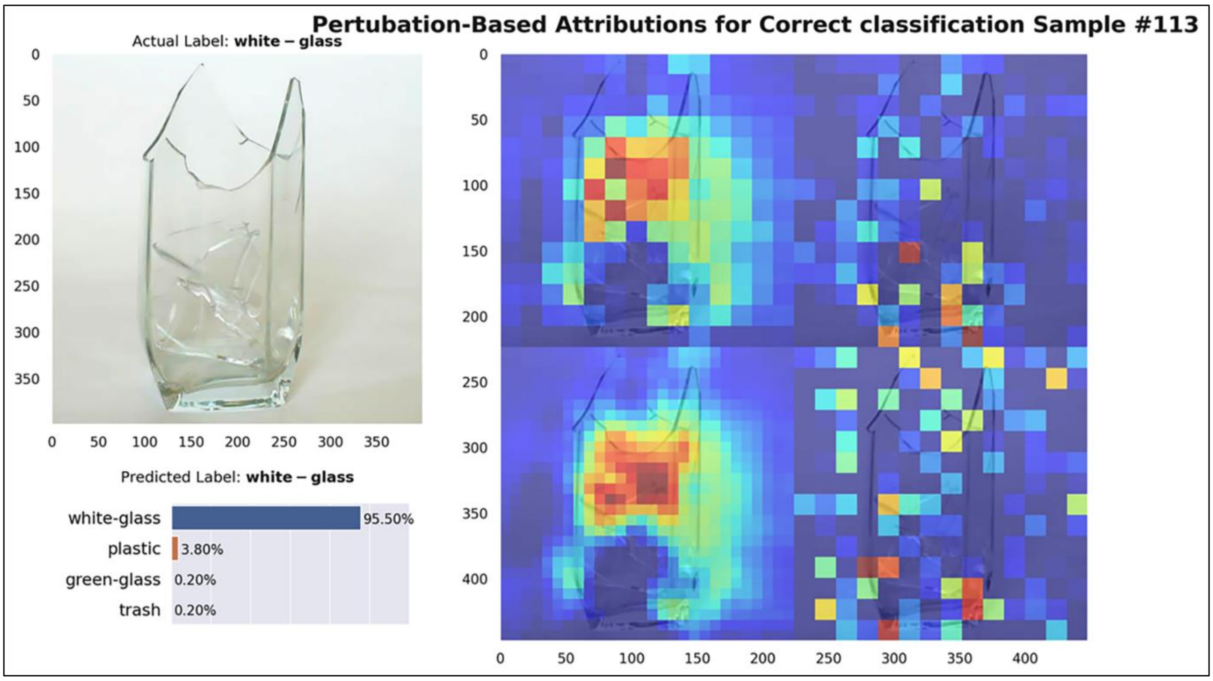

Actual label: white-glass, Predicted: white-glass
Original Image
Feature Ablation Heatmap
Feature Ablation Overlayed

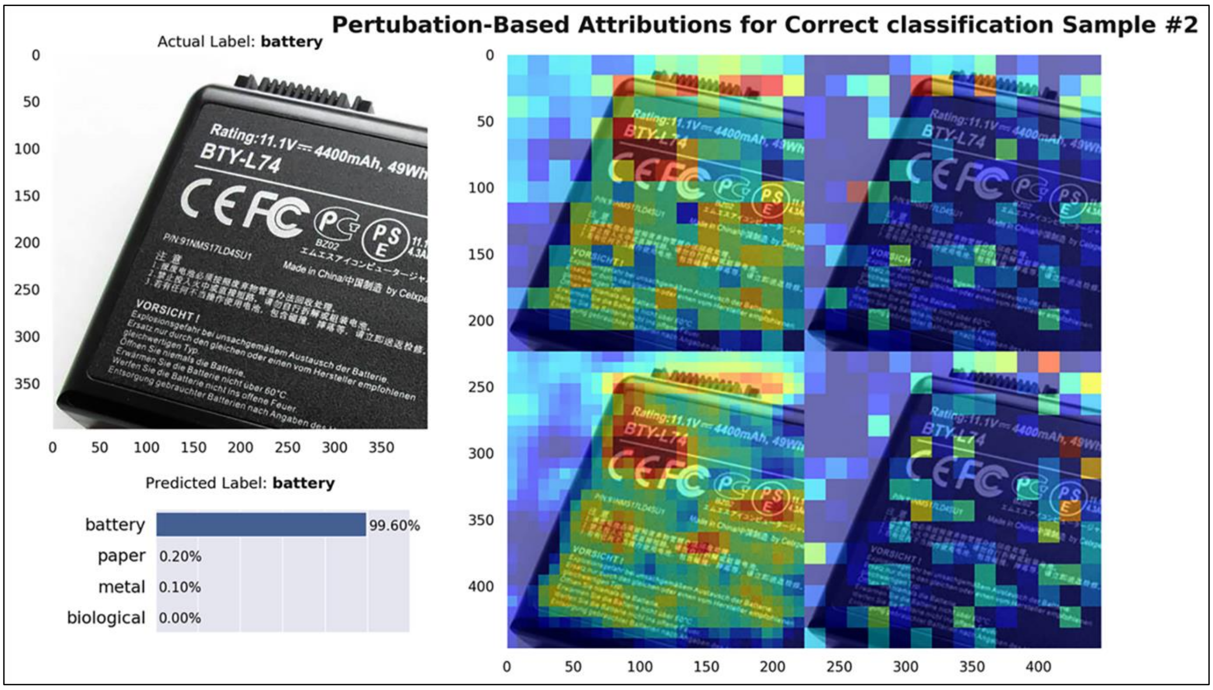

Actual label: white-glass, Predicted: white-glass
Original Image
Occlusion Heatmap
Occlusion Overlayed


Actual label: white-glass, Predicted: white-glass
Original Image
Shapley Heatmap
Shapley Overlayed

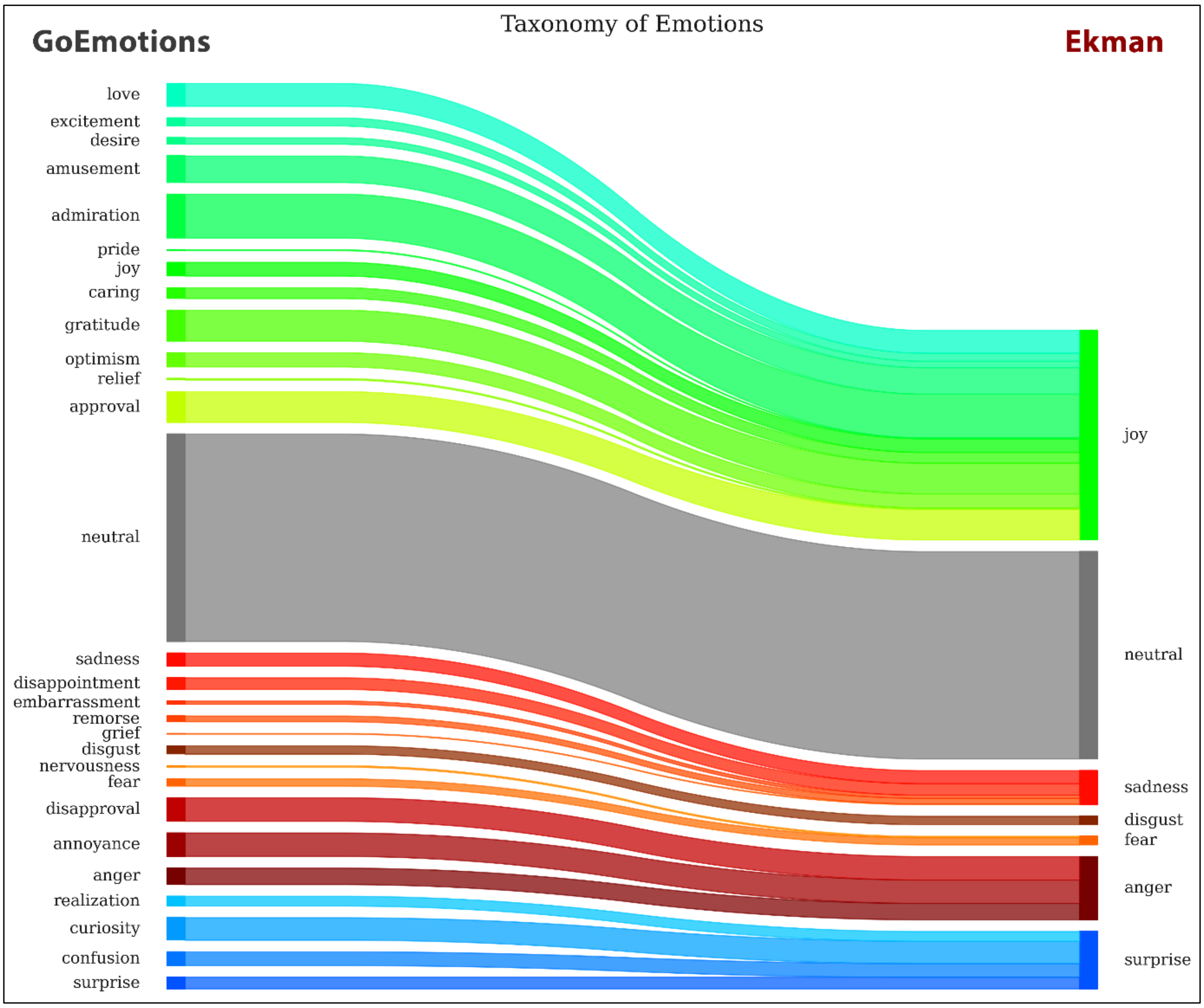

Actual label: white-glass, Predicted: white-glass
Original Image
Kernel Shap Light Baseline Heatmap
Kernel Shap Dark Baseline Heatmap
0
25
50
75
100
125
150
175
200
0
50
100
150
200


Pertubation-Based Attributions for Correct classification Sample #1
Actual Label: battery
Predicted Label: battery
battery 96.70%
metal 1.50%
plastic 0.50%
shoes 0.30%
0
50
100
150
200
250
300
350
400



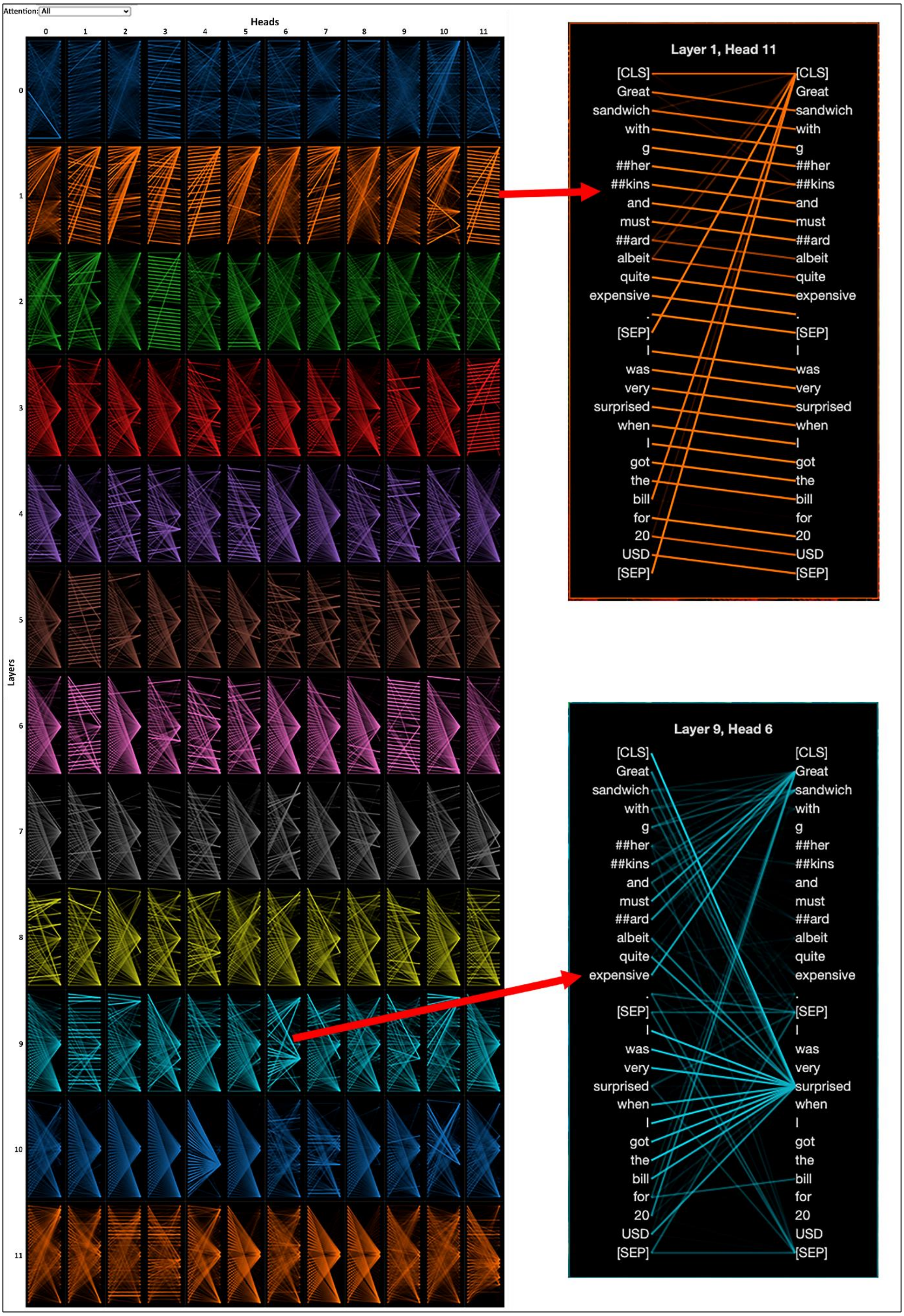

# Chapter 8: Interpreting NLP Transformers



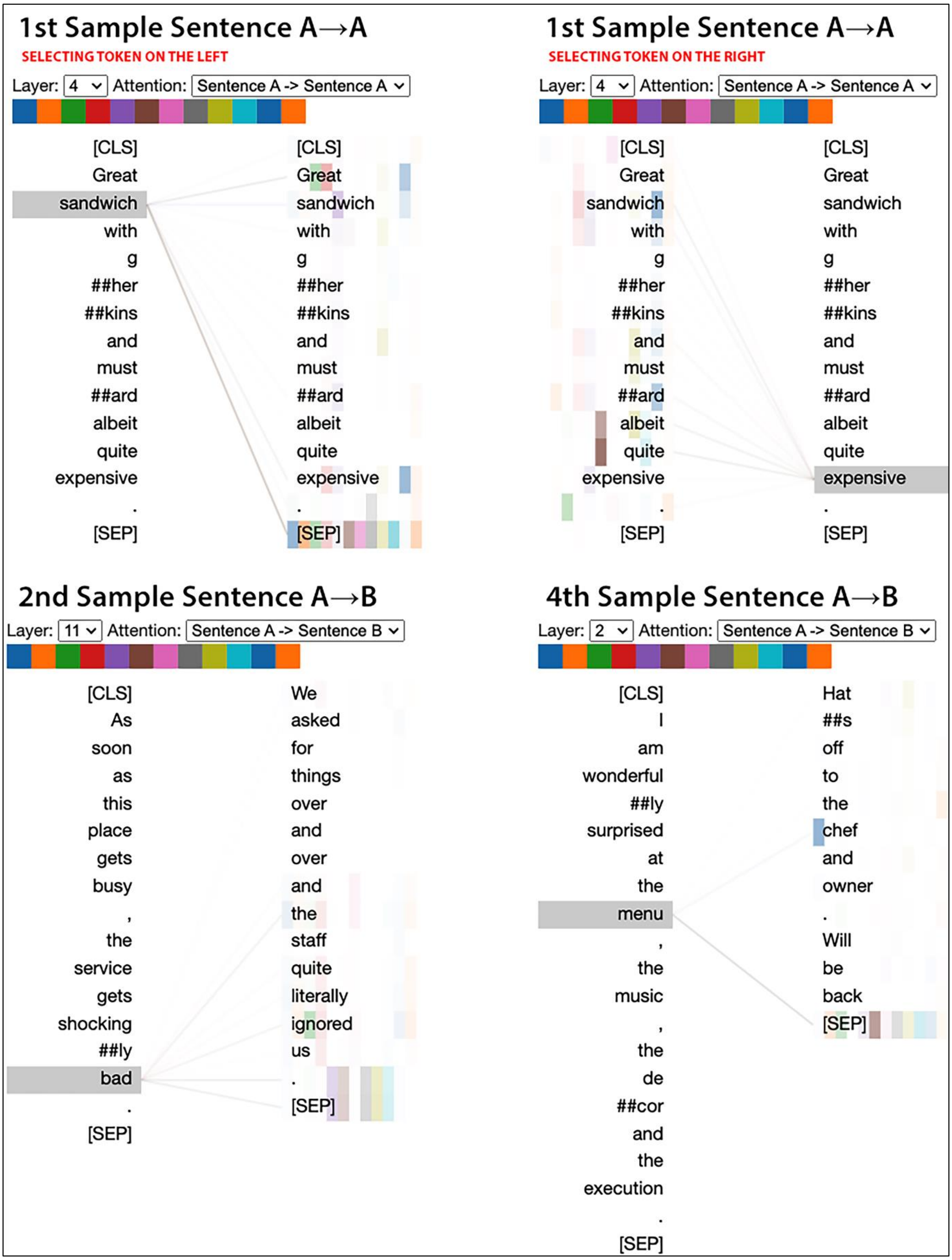

| review_title | review_full | label | score |
|---|---|---|---|
| Good neighborhood spot! | Came with family for Labor Day weekend brunch as my daughter lives nearby and it’s always been picked on previous visits. Had nice shaded and socially.. | joy | 0.987768 |
| Disappointing | Food was mediocre at best. The lamb chops are an image they feature on the websites opening page. It wasn't even listed on the menu. When I asked I wa.. | sadness | 0.504617 |
| What a find in Harlem | My co-workers were volunteering at a foodbank around the corner and we came here for lunch. What a find. Awesome Italian food with unique twists, not .. | joy | 0.999603 |

| label | count | avg. score | % positive | avg. rating |
|---|---|---|---|---|
| joy | 344,982 | 97.1% | 91.3% | 4.46 |
| surprise | 10,263 | 73.8% | 65.1% | 3.74 |
| neutral | 12,305 | 67.5% | 36.5% | 3.08 |
| sadness | 9,956 | 81.4% | 12.4% | 2.52 |
| anger | 1,398 | 56.6% | 4.2% | 1.99 |
| fear | 621 | 59.3% | 15.5% | 1.90 |
| disgust | 932 | 68.2% | 0.3% | 1.37 |

```
2nd_Avenue_Deli
        Sentiment:              Positive
        Rating:                 4
        GoEmotions Label:       surprise
        GoEmotions Score:       91.0%
        Title: Excellent salt beef sandwich
        Review:Great sandwich with gherkins and mustard albeit quite expensive.
               I was very surprised when I got the bill for 20 USD

Morning_Star_Cafe
        Sentiment:              Negative
        Rating:                 2
        GoEmotions Label:       surprise
        GoEmotions Score:       98.4%
        Title: Shocking when busy.
        Review:As soon as this place gets busy, the service gets shockingly bad.
               We asked for things over and over and the staff quite literally ignored us.

The_National_Bar_Dining_Rooms
        Sentiment:              Negative
        Rating:                 1
        GoEmotions Label:       surprise
        GoEmotions Score:       97.8%
        Title: Breakfast review
        Review:Poor service...poor omelette...poor croissant...will not try again!
               Lunch is much better but I was surprised at how bad their offering was...

Jacob_s_Pickles
        Sentiment:              Positive
        Rating:                 5
        GoEmotions Label:       surprise
        GoEmotions Score:       97.1%
        Title: All around great
        Review:I am wonderfully surprised at the menu, the music, the decor and the execution.
               Hats off to the chef and owner. Will be back
```



Attention: All
Heads
0 1 2 3 4 5 6 7 8 9 10 11
Layers
0 1 2 3 4 5 6 7 8 9 10 11
Layer 1, Head 11
[CLS] Great sandwich with g ##her ##kins and must ##ard albeit quite expensive . [SEP] I was very surprised when I got the bill for 20 USD [SEP]
Layer 9, Head 6
[CLS] Great sandwich with g ##her ##kins and must ##ard albeit quite expensive . [SEP] I was very surprised when I got the bill for 20 USD [SEP]


1st Sample Sentence A→A
SELECTING TOKEN ON THE LEFT
Layer: 4 Attention: Sentence A -> Sentence A
[CLS] Great sandwich with g ##her ##kins and must ##ard albeit quite expensive . [SEP]
[CLS] Great sandwich with g ##her ##kins and must ##ard albeit quite expensive . [SEP]
1st Sample Sentence A→A
SELECTING TOKEN ON THE RIGHT
Layer: 4 Attention: Sentence A -> Sentence A
[CLS] Great sandwich with g ##her ##kins and must ##ard albeit quite expensive . [SEP]
[CLS] Great sandwich with g ##her ##kins and must ##ard albeit quite expensive . [SEP]
2nd Sample Sentence A→B
Layer: 11 Attention: Sentence A -> Sentence B
[CLS] As soon as this place gets busy , the service gets shocking ##ly bad . [SEP]
We asked for things over and over and the staff quite literally ignored us . [SEP]
4th Sample Sentence A→B
Layer: 2 Attention: Sentence A -> Sentence B
[CLS] I am wonderful ##ly surprised at the menu , the music , the de ##cor and the execution . [SEP]
Hat ##s off to the chef and owner . Will be back [SEP]

355891: Puglia_Restaurant

**Legend:** Negative Neutral Positive

| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| Negative (1) | surprise (0.92) | surprise | 0.92 | Offensive Italian : Am simply shocked this restaurant tries to pass themselves off as authentic , traditional Italian - it is only appropriate for the university age " all you can drink " crowd that can ' t b ##oil past ##a themselves . I have been to countless Italian restaurants throughout NY and P ##ug ##lia may actually be the worst ( even inferior to $ 1 pizza places ) . Regarding the food , the best thing we ate was the garlic bread . Otherwise , simple , plain , b ##land , Italian staple ##s . If in a large group with pre fix , expect a heap ##ing amount of pen ##ne in various sauce ##s . The house wine is o ##bs ##cene ##ly und ##rin ##ka ##ble . They have music on the weekends - this actually is enjoyable momentarily , but they have a limited repertoire and will ha ##rass you for tips in between sets . |

507090: Le_Pain_Quotidien

**Legend:** Negative Neutral Positive

| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| Negative (1) | surprise (0.91) | surprise | 0.91 | 5 $ stolen ? : Food is great : no complaints . But the service at the check out counter : un ##bel ##ie ##va ##bly bad . I purchased a c ##rois ##san ##ts , tea with milk , gave cash , expecting 5 $ change . While waiting for the change , the c ##rois ##san ##t was brought in a bag , which I checked : it was smashed . The tea had no milk . Then I realized that I had never received the 5 $ change , to the best of my knowledge . . . but the cash ##ier was now ( within seconds after taking my payments ) convenient ##ly " taking a break and no longer in the building " , according to another service person ( the manager ? ) , at 3 : 04 pm , April 21st , 2019 . Nobody else , of course , could now give me my change . I know : I should have checked . |

193197: Almond

**Legend:** Negative Neutral Positive

| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| Negative (2) | surprise (0.93) | surprise | 0.93 | Don ' t bother : There is so many great places to eat in Gram ##er ##cy I am not sure how we ended up here . The staff were more interested in gossip ##ing than looking after the 8 people that had made the same mistake we had . Lamb ch ##ops as a meal means more than 1 and a half cut ##lets , how you can make a b ##urger so dry and taste ##less is a mystery , the past ##a was ok . Had to beg for a beer even . . . . to hard don ' t bother . |

174001: Katz_s_Deli

**Legend:** Negative Neutral Positive

| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| Negative (1) | surprise (0.96) | surprise | 0.96 | Shock ##ed : A long awaited trip to New York , and Katz diner was the place I wanted to visit , only a 15 ##min walk from our Hotel , perfect . It was 10 o ' clock on a sunny Sunday morning . En ##tering we were met by a very over ##weight man leaning back in a chair " here ' s you ' re ticket order there , pay on your way out " and do you pay , 3 past ##ram ##i sandwiches 1 cheese o ##mel ##ette and 3 regular coffee ##s just shy of $ 120 . People think London is expensive , considering Katz location it ' s extremely expensive , sandwiches not impressive and a g ##loom ##y and slightly g ##rea ##sy de ##cor ( I know it ' s meant to be vintage but wipe the tables down and m ##op the floor ) overall a very bad experience from the staff and the food . Don ' t waste your money going to Katz . |

387224: Cafe_Frida

**Legend:** Negative Neutral Positive

| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| Positive (5) | surprise (0.96) | surprise | 0.96 | I was a bit per ##plex ##ed when our friends invited us here , as we ' re all from Texas . And to come all the way to NYC just to eat Mexican . . . I found it odd . Then I tried the food . . . and understood . This was authentic and out of this world ! From the Tom ##ato B ##is ##que to the En ##chal ##ada ##s Mo ##le you couldn ' t find better unless you went to Mexico itself . This will be on my must stop list every trip to NYC from now on ! |

66146: Sakagura

**Legend:** Negative Neutral Positive

| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| Positive (5) | surprise (0.58) | surprise | 0.58 | Un ##believable location and authentic Japanese food for a reasonable price . It is very hard to find a place for authentic big portion Japanese dish in the city . |
| Positive (5) | surprise (0.58) | joy | 0.40 | Un ##believable location and authentic Japanese food for a reasonable price . It is very hard to find a place for authentic big portion Japanese dish in the city . |

132864: Eataly

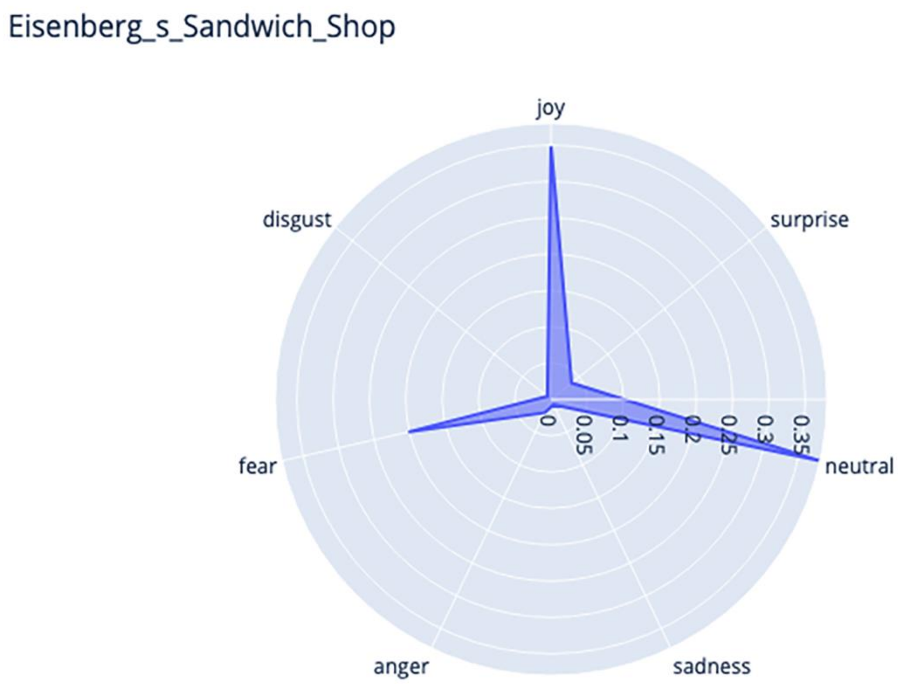
**Legend:** Negative Neutral Positive

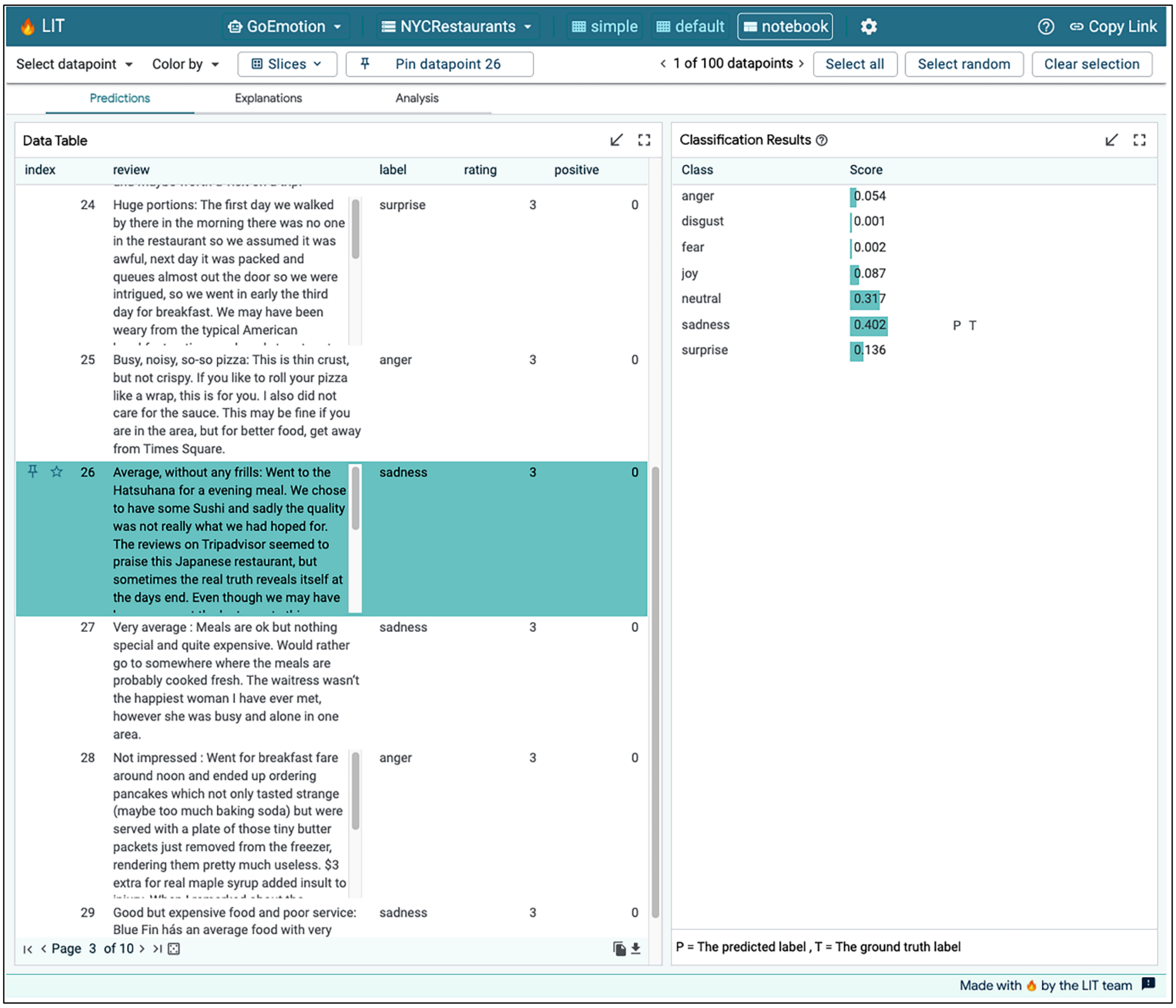
| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| Positive (5) | surprise (0.86) | surprise | 0.86 | We fell into Eat ##aly and couldn ' t believe how lucky we were to find this establishment . It was like finding ones ##elf in Italy . They have everything one could possibly want . We had wine and cheese at the wine bar , checked out the grocery store then on to the p ##iz ##zer ##ia . This place has something for everyone |

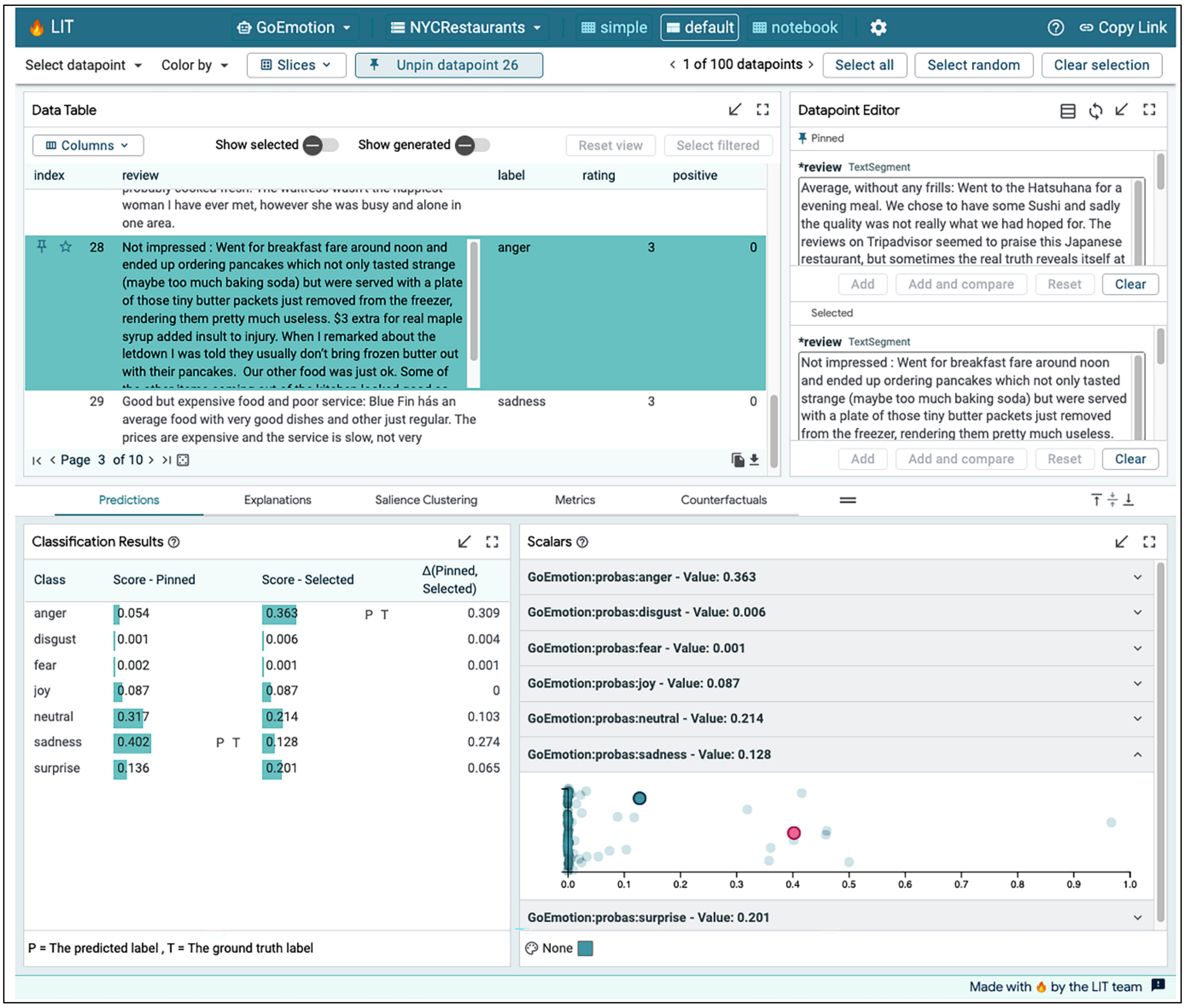
113241: Eisenberg_s_Sandwich_Shop

**Legend:** Negative Neutral Positive

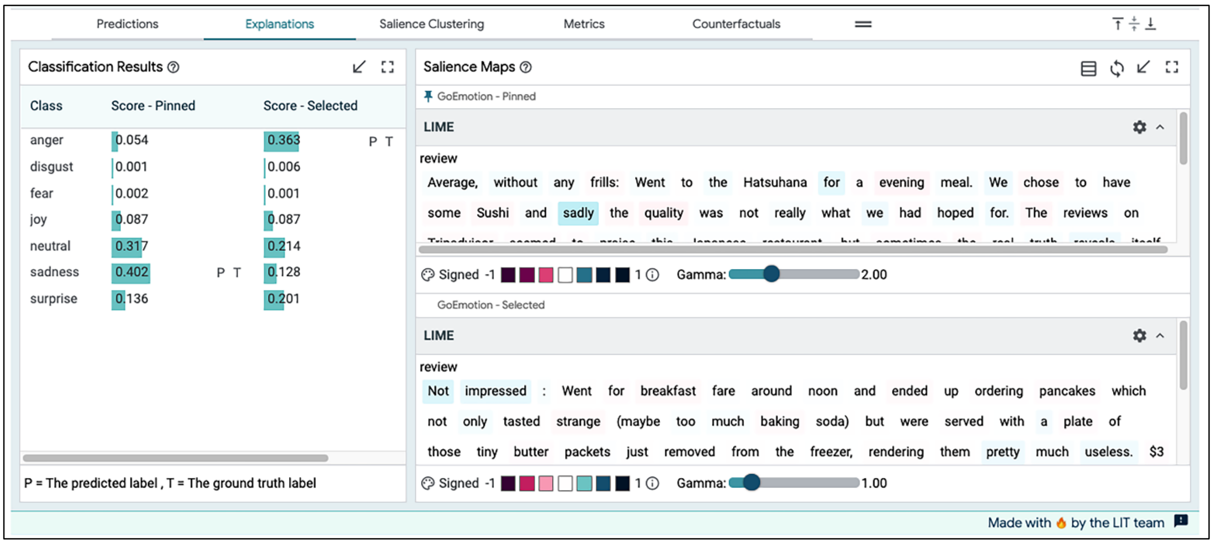
| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| Positive (4) | neutral (0.38) | neutral | 0.38 | An age worn classic del ##i complete with the extra long lunch counter and stool ##s . Has ##n ' t had a face lift in decades , but the sandwich was te ##rri ##fic and the staff was friendly . |
| Positive (4) | neutral (0.38) | joy | 0.35 | An age worn classic del ##i complete with the extra long lunch counter and stool ##s . Has ##n ' t had a face lift in decades , but the sandwich was te ##rri ##fic and the staff was friendly . |
| Positive (4) | neutral (0.38) | fear | 0.20 | An age worn classic del ##i complete with the extra long lunch counter and stool ##s . Has ##n ' t had a face lift in decades , but the sandwich was te ##rri ##fic and the staff was friendly . |



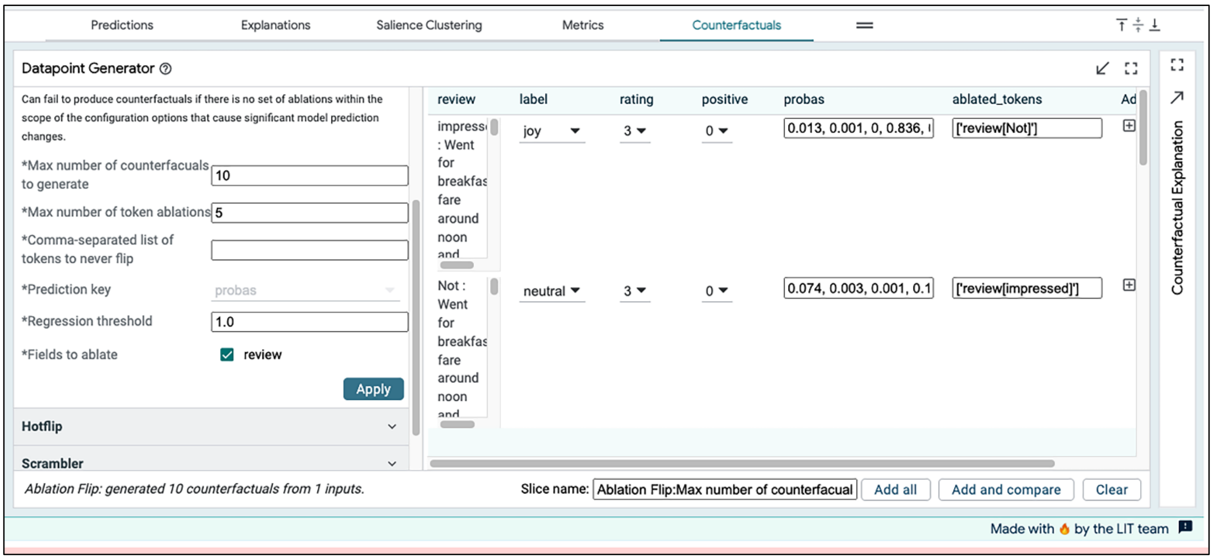

LIT
GoEmotion
NYCRestaurants
simple
default
notebook
Copy Link
Select datapoint
Color by
Slices
Pin datapoint 26
1 of 100 datapoints
Select all
Select random
Clear selection
Predictions
Explanations
Analysis
Data Table
index
review
label
rating
positive
24
Huge portions: The first day we walked by there in the morning there was no one in the restaurant so we assumed it was awful, next day it was packed and queues almost out the door so we were intrigued, so we went in early the third day for breakfast. We may have been weary from the typical American
surprise
3
0
25
Busy, noisy, so-so pizza: This is thin crust, but not crispy. If you like to roll your pizza like a wrap, this is for you. I also did not care for the sauce. This may be fine if you are in the area, but for better food, get away from Times Square.
anger
3
0
26
Average, without any frills: Went to the Hatsuhana for a evening meal. We chose to have some Sushi and sadly the quality was not really what we had hoped for. The reviews on Tripadvisor seemed to praise this Japanese restaurant, but sometimes the real truth reveals itself at the days end. Even though we may have
sadness
3
0
27
Very average : Meals are ok but nothing special and quite expensive. Would rather go to somewhere where the meals are probably cooked fresh. The waitress wasn't the happiest woman I have ever met, however she was busy and alone in one area.
sadness
3
0
28
Not impressed : Went for breakfast fare around noon and ended up ordering pancakes which not only tasted strange (maybe too much baking soda) but were served with a plate of those tiny butter packets just removed from the freezer, rendering them pretty much useless. $3 extra for real maple syrup added insult to
anger
3
0
29
Good but expensive food and poor service: Blue Fin hás an average food with very
sadness
3
0
Page 3 of 10
Classification Results
Class
Score
anger
0.054
disgust
0.001
fear
0.002
joy
0.087
neutral
0.317
sadness
0.402
P T
surprise
0.136
P = The predicted label , T = The ground truth label
Made with by the LIT team



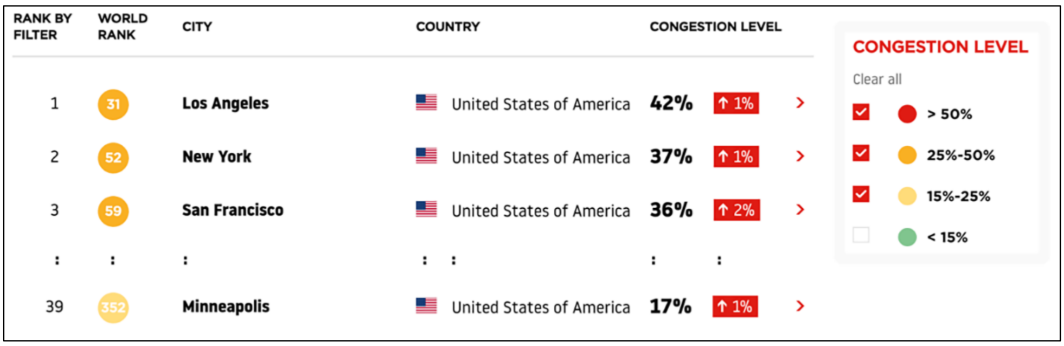
LIT
GoEmotion
NYCRestaurants
simple
default
notebook
Copy Link
Select datapoint
Color by
Slices
Unpin datapoint 26
1 of 100 datapoints
Select all
Select random
Clear selection
Data Table
Columns
Show selected
Show generated
Reset view
Select filtered
index
review
label
rating
positive
woman I have ever met, however she was busy and alone in one area.
28
Not impressed : Went for breakfast fare around noon and ended up ordering pancakes which not only tasted strange (maybe too much baking soda) but were served with a plate of those tiny butter packets just removed from the freezer, rendering them pretty much useless. $3 extra for real maple syrup added insult to injury. When I remarked about the letdown I was told they usually don't bring frozen butter out with their pancakes. Our other food was just ok. Some of
anger
3
0
29
Good but expensive food and poor service: Blue Fin hás an average food with very good dishes and other just regular. The prices are expensive and the service is slow, not very
sadness
3
0
Page 3 of 10
Datapoint Editor
Pinned
*review TextSegment
Average, without any frills: Went to the Hatsuhana for a evening meal. We chose to have some Sushi and sadly the quality was not really what we had hoped for. The reviews on Tripadvisor seemed to praise this Japanese restaurant, but sometimes the real truth reveals itself at
Add
Add and compare
Reset
Clear
Selected
*review TextSegment
Not impressed : Went for breakfast fare around noon and ended up ordering pancakes which not only tasted strange (maybe too much baking soda) but were served with a plate of those tiny butter packets just removed from the freezer, rendering them pretty much useless.
Add
Add and compare
Reset
Clear
Predictions
Explanations
Salience Clustering
Metrics
Counterfactuals
Classification Results
Class
Score - Pinned
Score - Selected
Δ(Pinned, Selected)
anger 0.054 0.363 P T 0.309
disgust 0.001 0.006 0.004
fear 0.002 0.001 0.001
joy 0.087 0.087 0
neutral 0.317 0.214 0.103
sadness 0.402 P T 0.128 0.274
surprise 0.136 0.201 0.065
P = The predicted label , T = The ground truth label
Scalars
GoEmotion:probas:anger - Value: 0.363
GoEmotion:probas:disgust - Value: 0.006
GoEmotion:probas:fear - Value: 0.001
GoEmotion:probas:joy - Value: 0.087
GoEmotion:probas:neutral - Value: 0.214
GoEmotion:probas:sadness - Value: 0.128
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
GoEmotion:probas:surprise - Value: 0.201
None
Made with by the LIT team

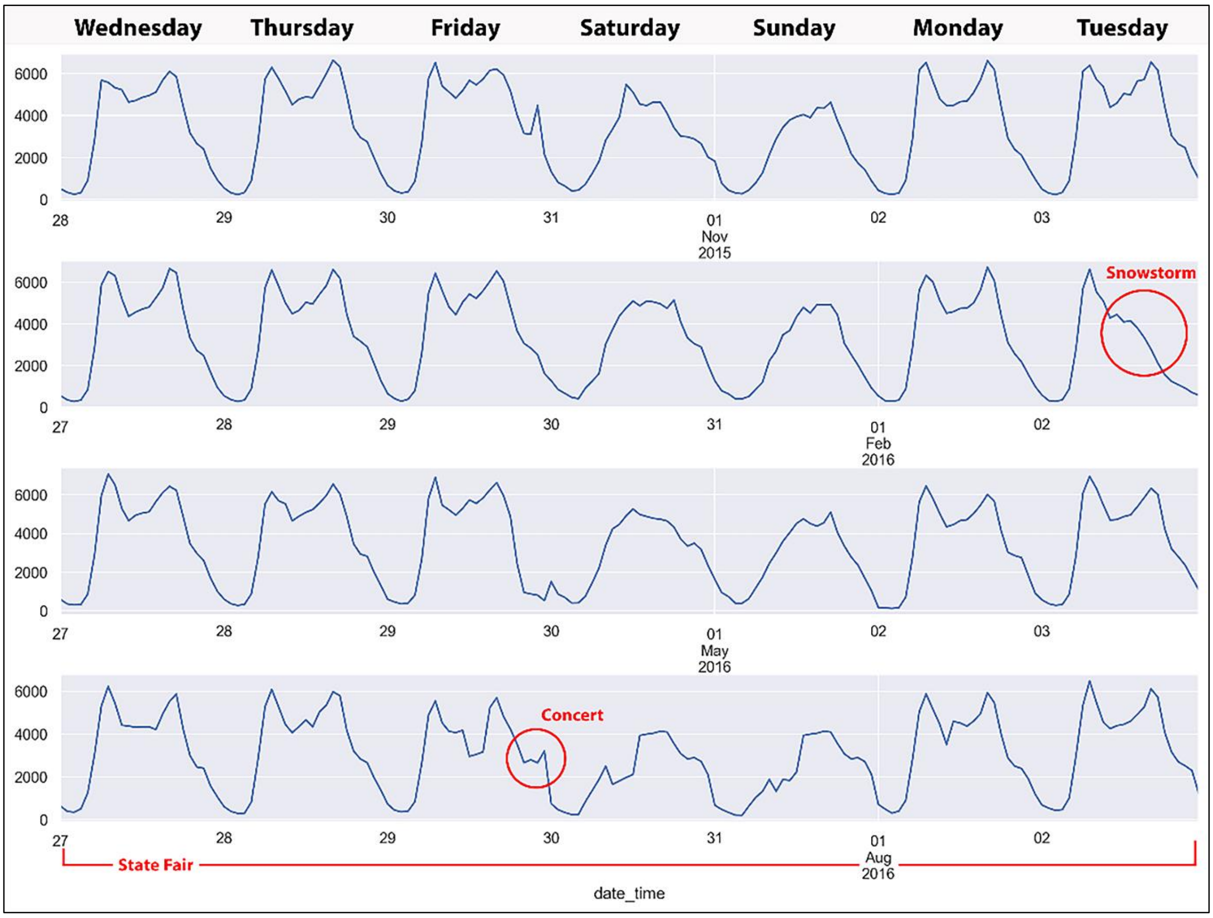



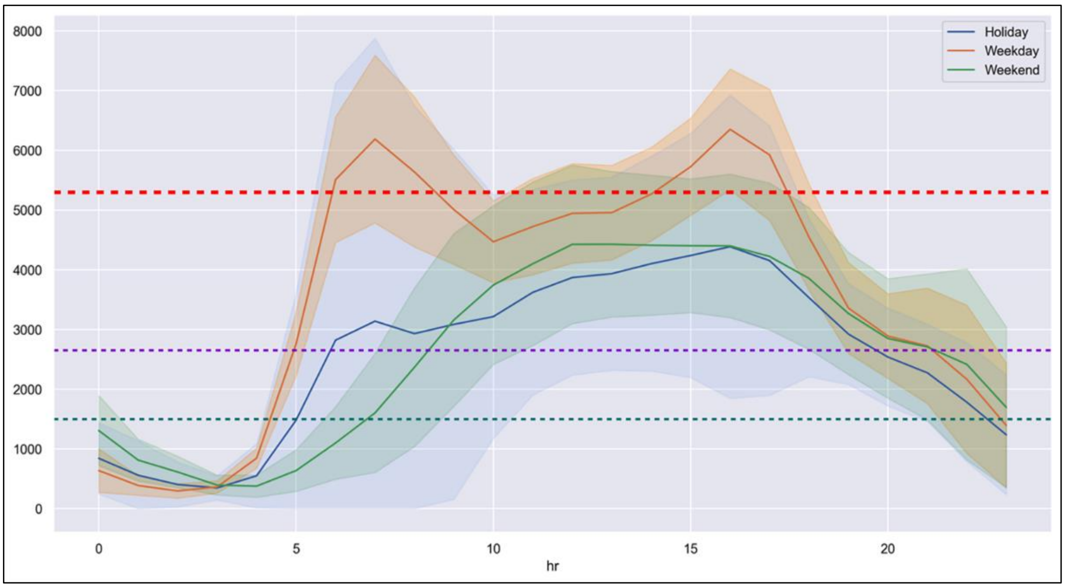
Predictions
Explanations
Salience Clustering
Metrics
Counterfactuals
Classification Results
Class
Score - Pinned
Score - Selected
anger 0.054 0.363 P T
disgust 0.001 0.006
fear 0.002 0.001
joy 0.087 0.087
neutral 0.317 0.214
sadness 0.402 P T 0.128
surprise 0.136 0.201
P = The predicted label , T = The ground truth label
Salience Maps
GoEmotion - Pinned
LIME
review
Average, without any frills: Went to the Hatsuhana for a evening meal. We chose to have some Sushi and sadly the quality was not really what we had hoped for. The reviews on
Signed -1 1
Gamma: 2.00
GoEmotion - Selected
LIME
review
Not impressed : Went for breakfast fare around noon and ended up ordering pancakes which not only tasted strange (maybe too much baking soda) but were served with a plate of those tiny butter packets just removed from the freezer, rendering them pretty much useless. $3
Signed -1 1
Gamma: 1.00
Made with by the LIT team

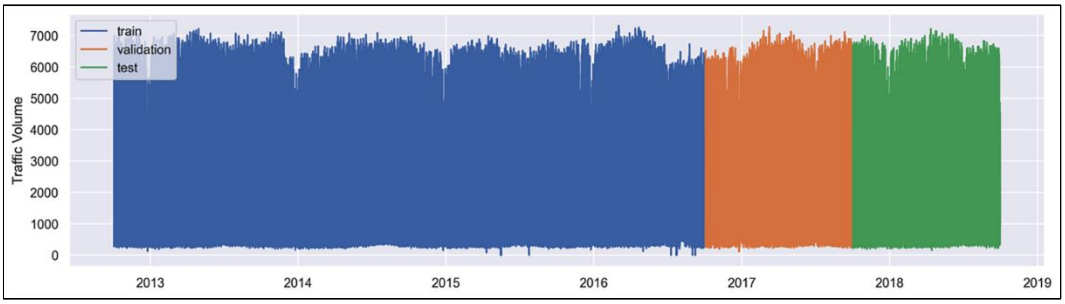



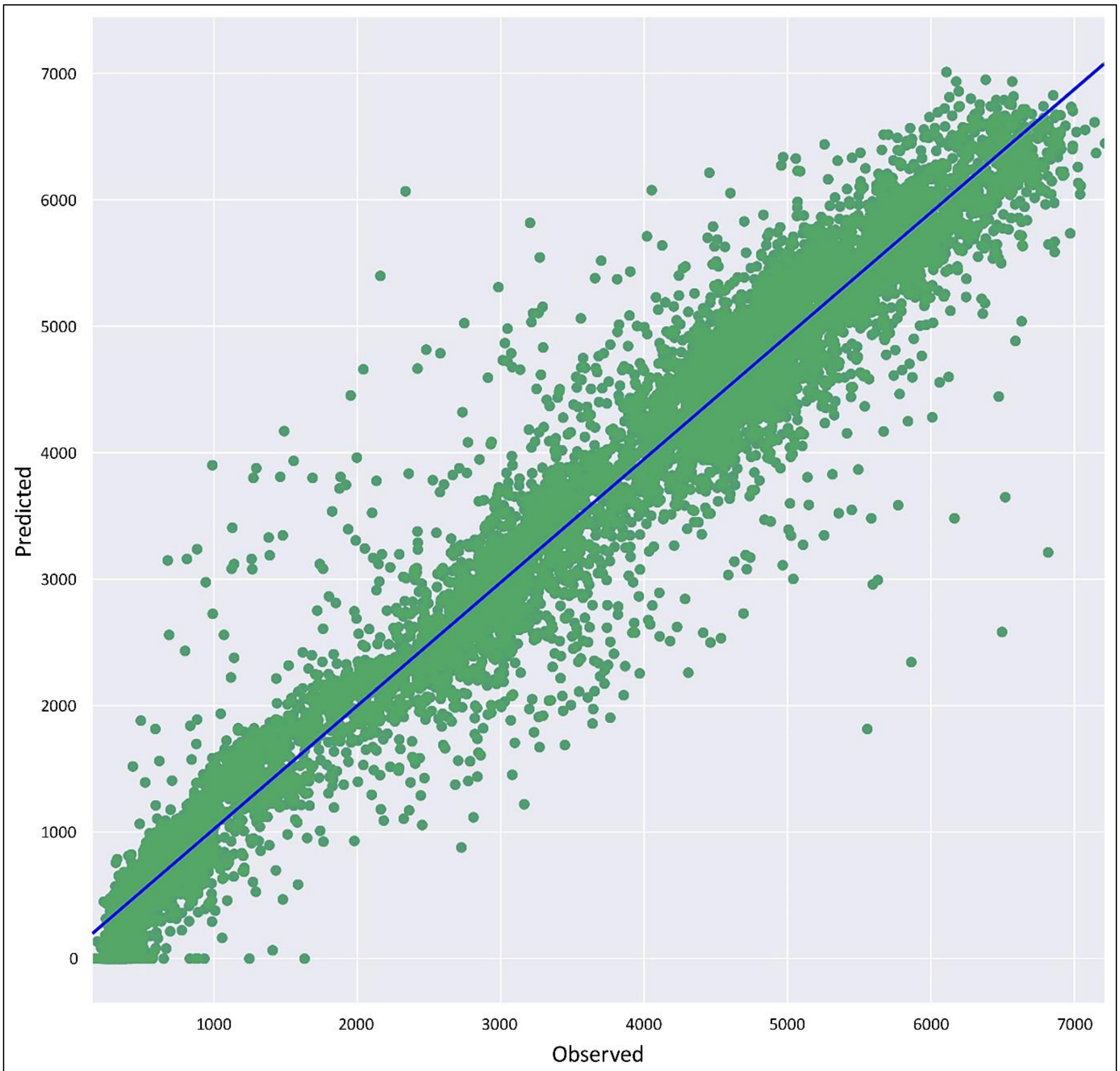

# Chapter 9: Interpretation Methods for Multivariate Forecasting and Sensitivity Analysis



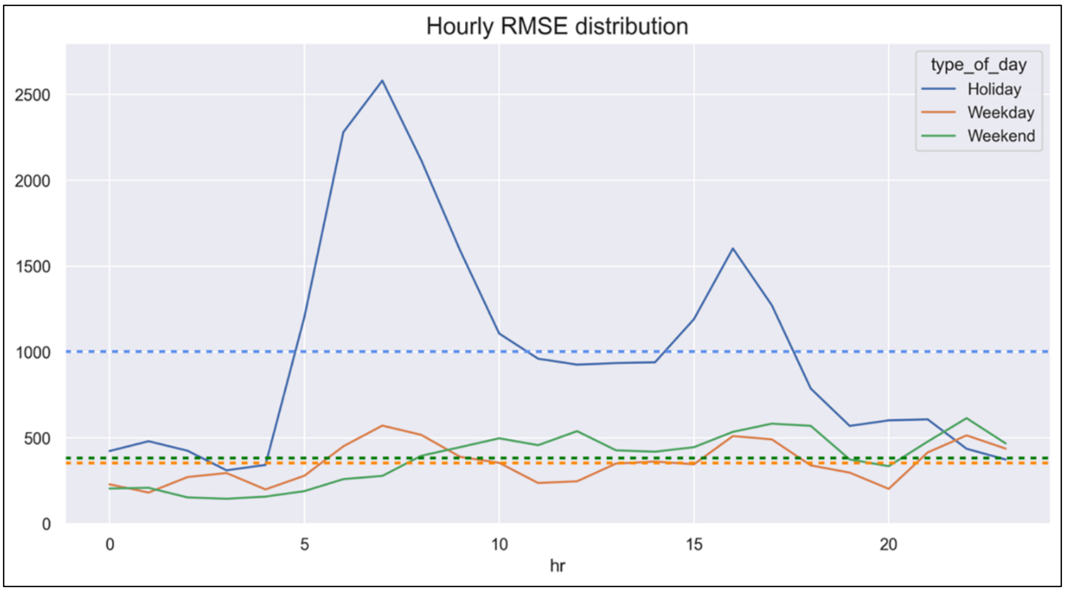

Wednesday
Thursday
Friday
Saturday
Sunday
Monday
Tuesday
Nov
2015
Snowstorm
Feb
2016
May
2016
Concert
Aug
2016
State Fair
date_time

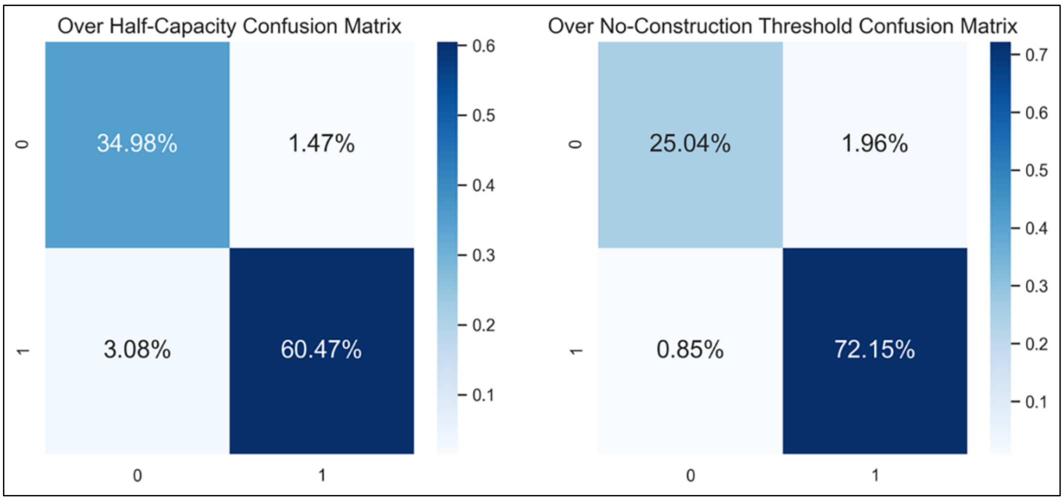

Holiday
Weekday
Weekend
hr

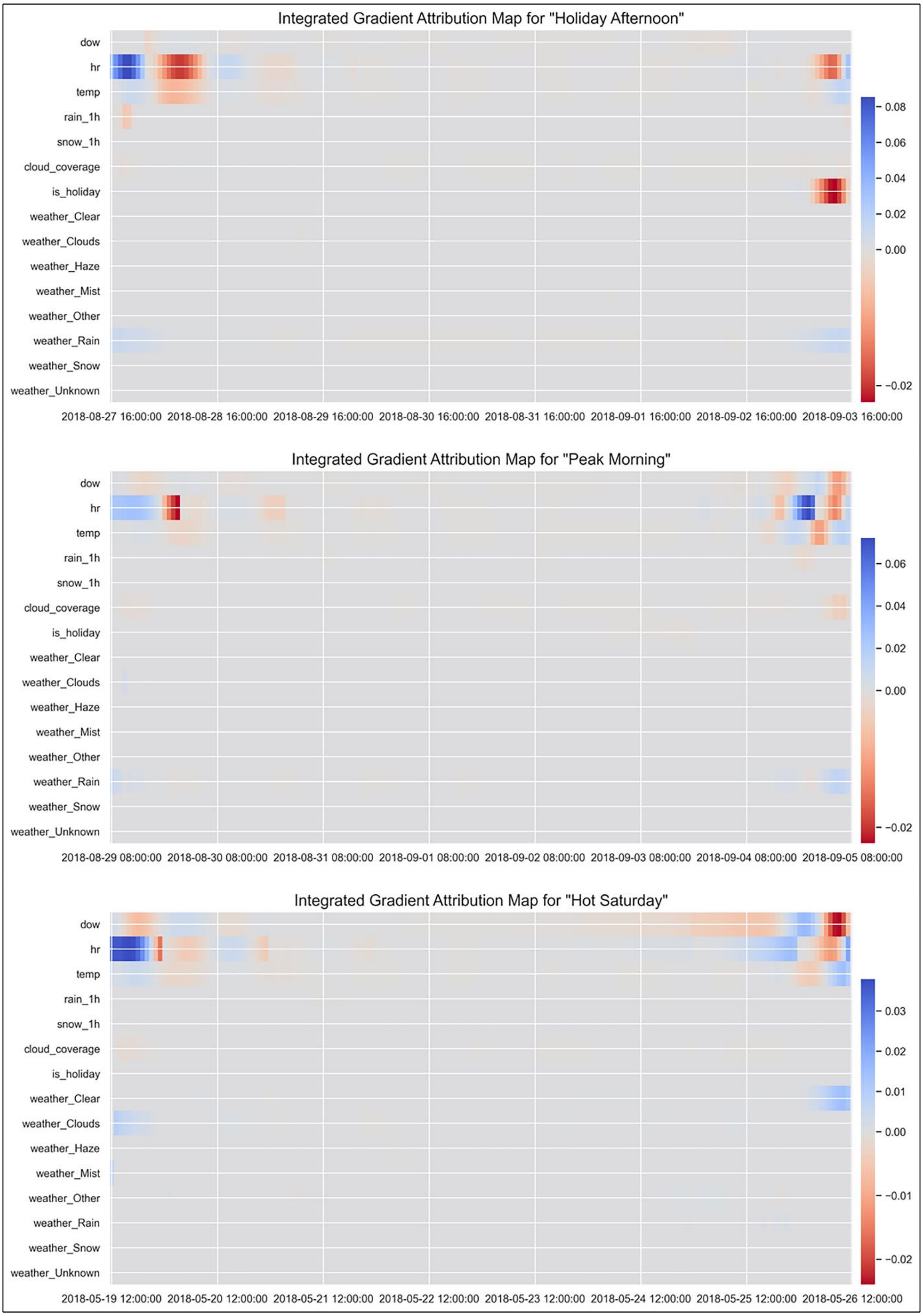


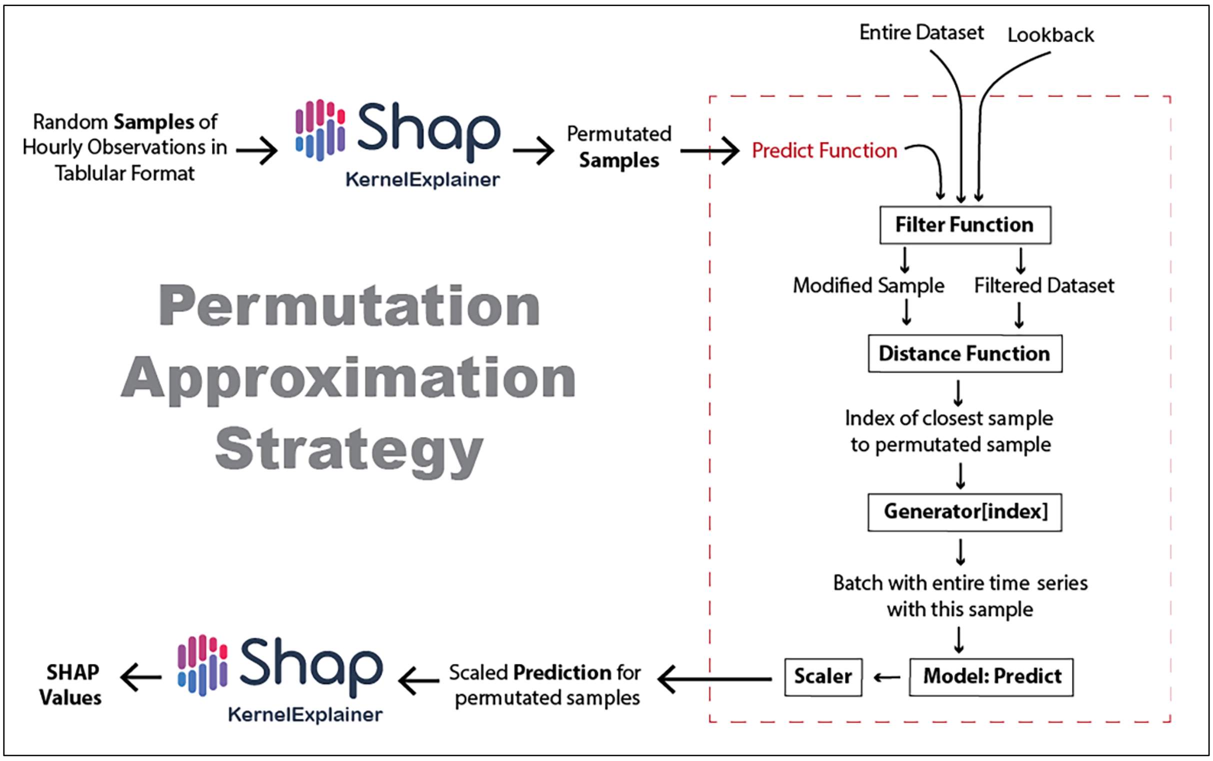

RMSE_train: 422.76 RMSE_test: 430.26 r2: 0.9528

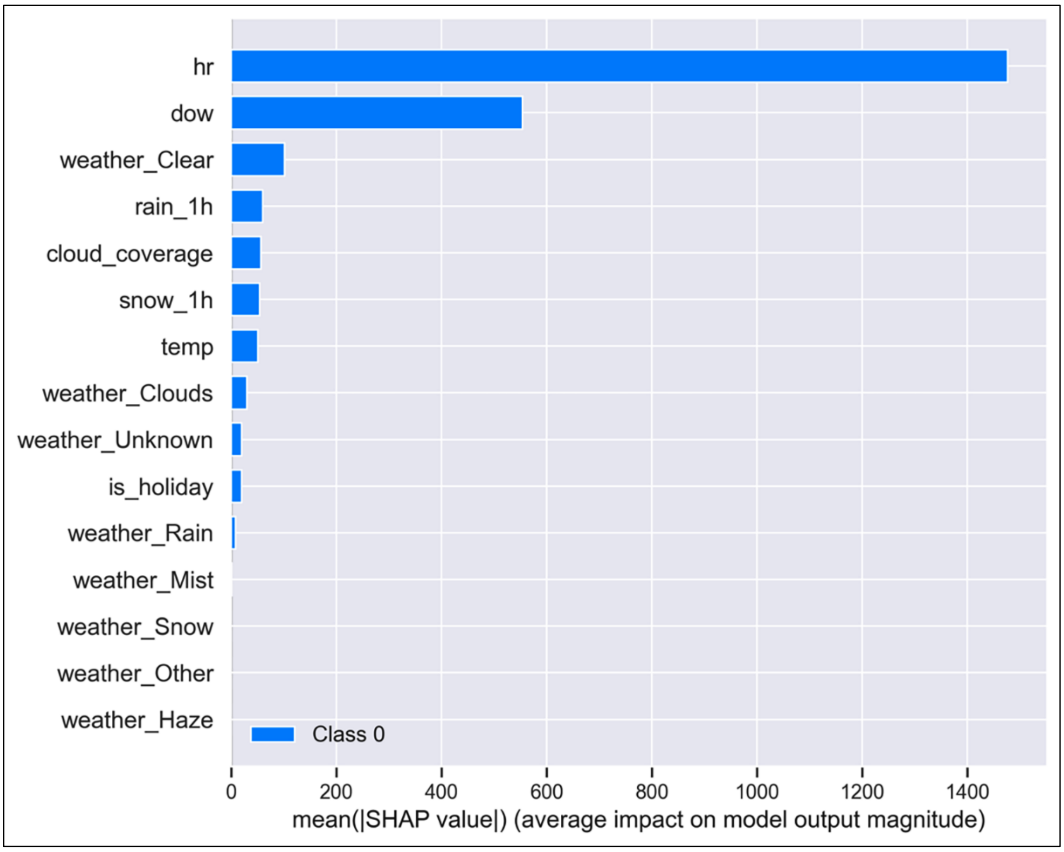

Hourly RMSE distribution
type_of_day
Holiday
Weekday
Weekend
2500
2000
1500
1000
500
0
0
5
10
15
20
hr

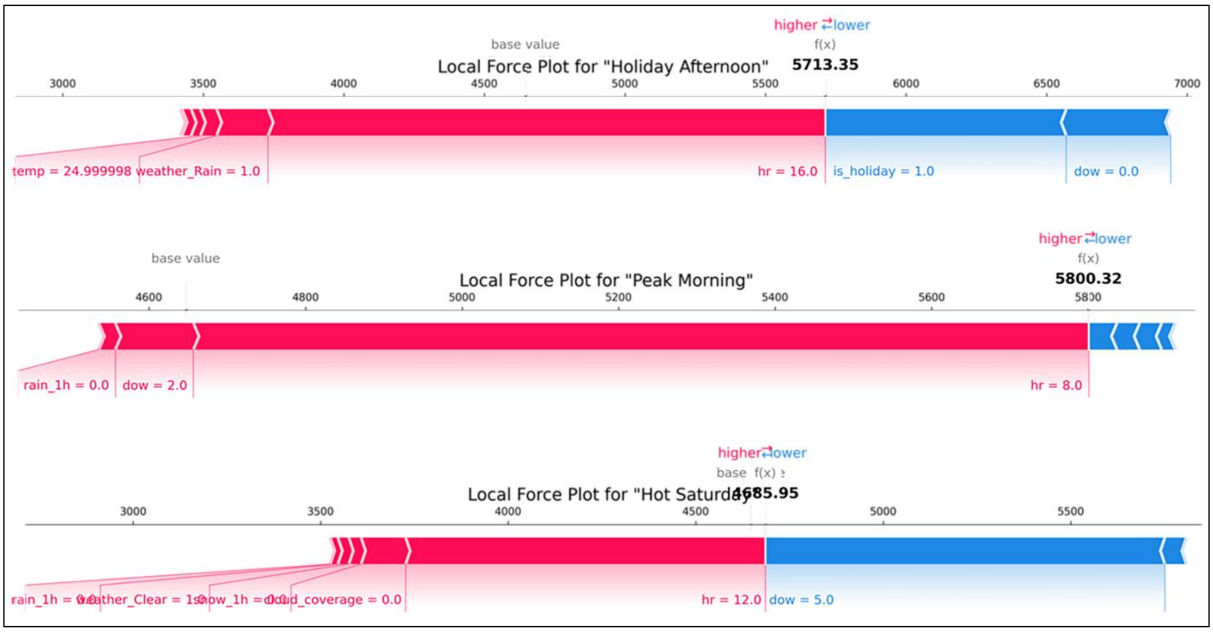

Over Half-Capacity Confusion Matrix
34.98%
1.47%
3.08%
60.47%
0
1
0
1
0.6
0.5
0.4
0.3
0.2
0.1
Over No-Construction Threshold Confusion Matrix
25.04%
1.96%
0.85%
72.15%
0
1
0
1
0.7
0.6
0.5
0.4
0.3
0.2
0.1



Integrated Gradient Attribution Map for "Holiday Afternoon"
dow
hr
temp
rain_1h
snow_1h
cloud_coverage
is_holiday
weather_Clear
weather_Clouds
weather_Haze
weather_Mist
weather_Other
weather_Rain
weather_Snow
weather_Unknown
2018-08-27 16:00:00
2018-08-28 16:00:00
2018-08-29 16:00:00
2018-08-30 16:00:00
2018-08-31 16:00:00
2018-09-01 16:00:00
2018-09-02 16:00:00
2018-09-03 16:00:00
0.08
0.06
0.04
0.02
0.00
−0.02
Integrated Gradient Attribution Map for "Peak Morning"
2018-08-29 08:00:00
2018-08-30 08:00:00
2018-08-31 08:00:00
2018-09-01 08:00:00
2018-09-02 08:00:00
2018-09-03 08:00:00
2018-09-04 08:00:00
2018-09-05 08:00:00
0.06
0.04
0.02
0.00
−0.02
Integrated Gradient Attribution Map for "Hot Saturday"
2018-05-19 12:00:00
2018-05-20 12:00:00
2018-05-21 12:00:00
2018-05-22 12:00:00
2018-05-23 12:00:00
2018-05-24 12:00:00
2018-05-25 12:00:00
2018-05-26 12:00:00
0.03
0.02
0.01
0.00
−0.01
−0.02

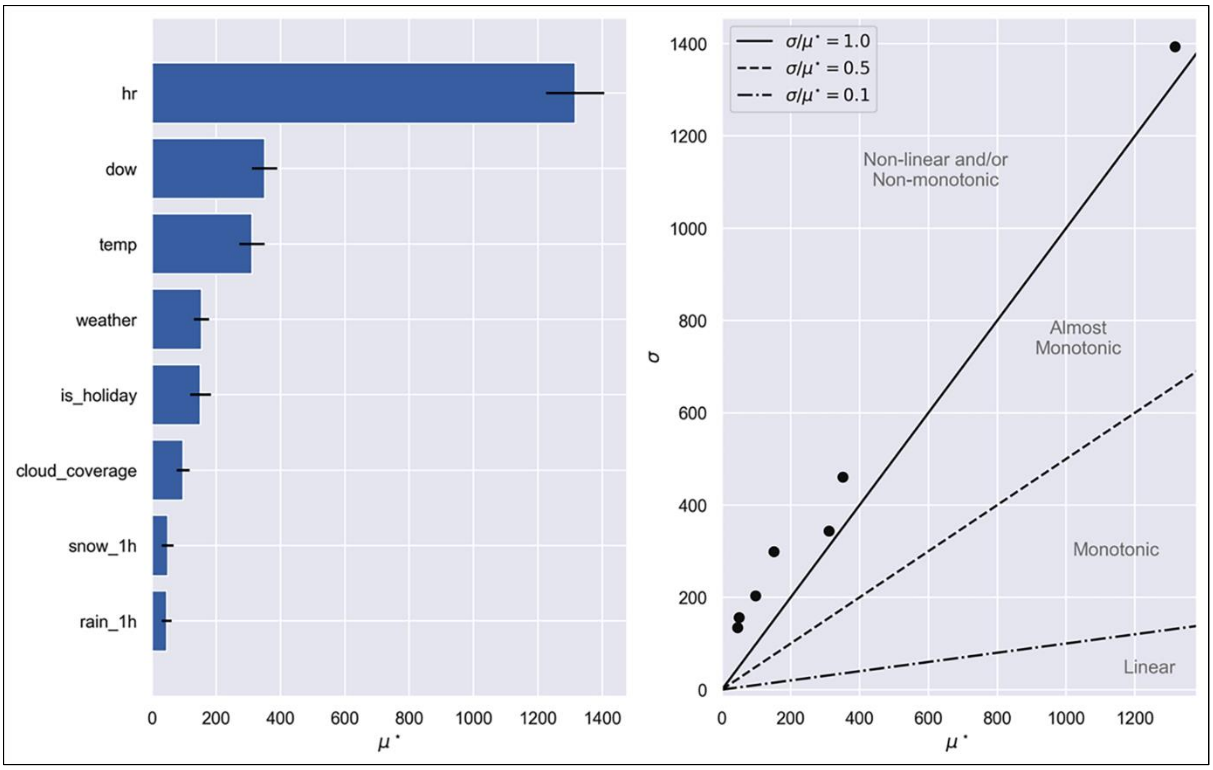

Random Samples of Hourly Observations in Tablular Format
Shap
KernelExplainer
Permutated Samples
Predict Function
Entire Dataset
Lookback
Filter Function
Modified Sample
Filtered Dataset
Distance Function
Index of closest sample to permutated sample
Generator[index]
Batch with entire time series with this sample
Model: Predict
Scaler
Scaled Prediction for permutated samples
Shap
KernelExplainer
SHAP Values
Permutation Approximation Strategy


hr
dow
weather_Clear
rain_1h
cloud_coverage
snow_1h
temp
weather_Clouds
weather_Unknown
is_holiday
weather_Rain
weather_Mist
weather_Snow
weather_Other
weather_Haze
Class 0
0
200
400
600
800
1000
1200
1400
mean(|SHAP value|) (average impact on model output magnitude)

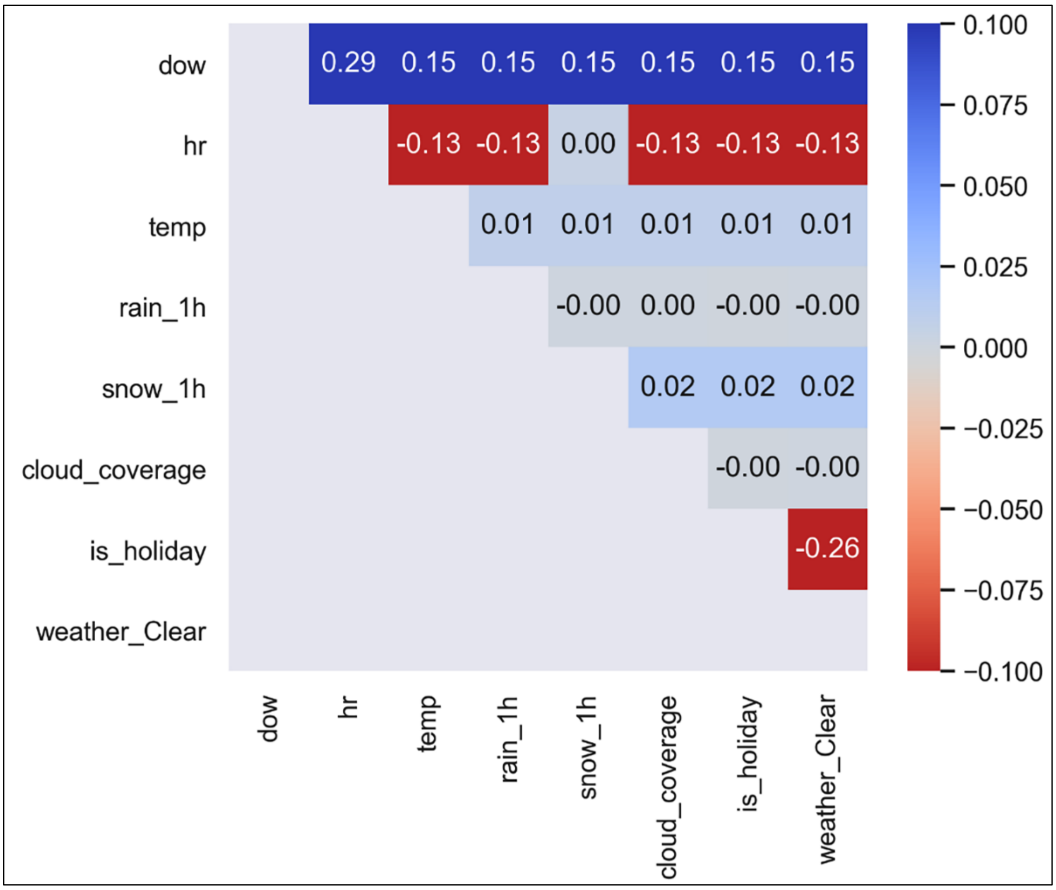



| | count | mean | std | min | 1% | 50% | 99% | max |
|---|---|---|---|---|---|---|---|---|
| **dow** | 7026.00 | 2.01 | 1.41 | 0.00 | 0.00 | 2.00 | 4.00 | 4.00 |
| **hr** | 7026.00 | 5.50 | 7.93 | 0.00 | 0.00 | 2.50 | 23.00 | 23.00 |
| **temp** | 7026.00 | 10.83 | 9.14 | -24.19 | -12.31 | 12.60 | 25.29 | 30.25 |
| **rain_1h** | 7026.00 | 0.12 | 0.67 | 0.00 | 0.00 | 0.00 | 3.10 | 20.40 |
| **snow_1h** | 7026.00 | 0.01 | 0.06 | 0.00 | 0.00 | 0.00 | 0.28 | 1.54 |
| **cloud_coverage** | 7026.00 | 38.54 | 37.77 | 0.00 | 0.00 | 27.60 | 100.00 | 100.00 |
| **is_holiday** | 7026.00 | 0.04 | 0.18 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| **weather_Clear** | 7026.00 | 0.32 | 0.46 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| : | : | : | : | : | : | : | : | : |
| **weather_Snow** | 7026.00 | 0.02 | 0.13 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| **weather_Unknown** | 7026.00 | 0.21 | 0.41 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |

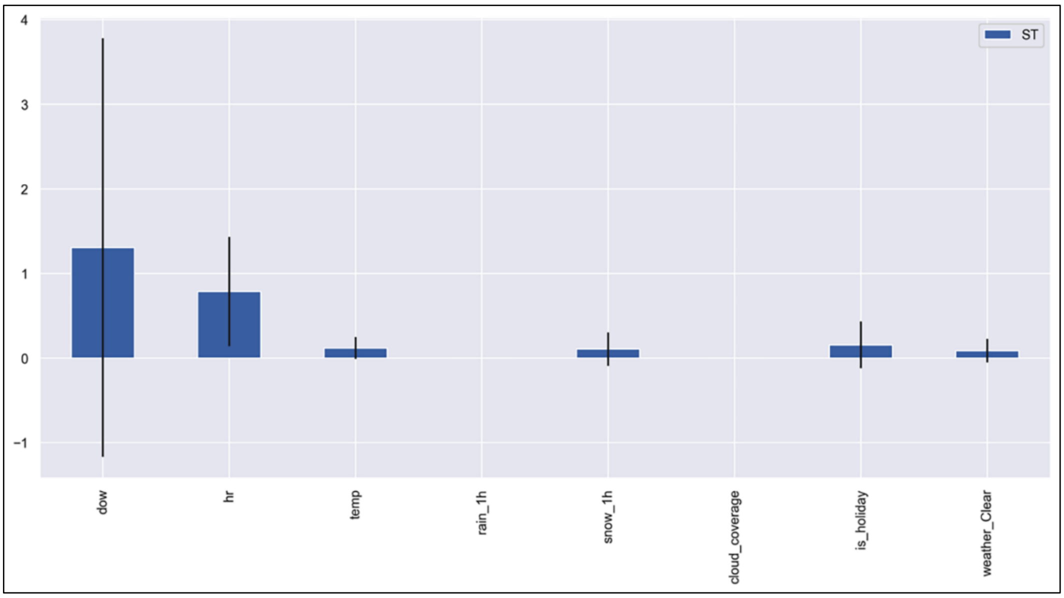
| features | μ | μ* | σ |
|---|---|---|---|
| hr | -560.18 | 1316.23 | 1393.25 |
| dow | 100.72 | 350.63 | 460.59 |
| temp | 263.15 | 311.29 | 344.00 |
| weather | nan | 154.45 | nan |
| is_holiday | -85.24 | 151.30 | 299.68 |
| cloud_coverage | -14.05 | 97.22 | 203.60 |
| snow_1h | -29.57 | 49.24 | 156.02 |
| rain_1h | 0.66 | 45.81 | 134.17 |



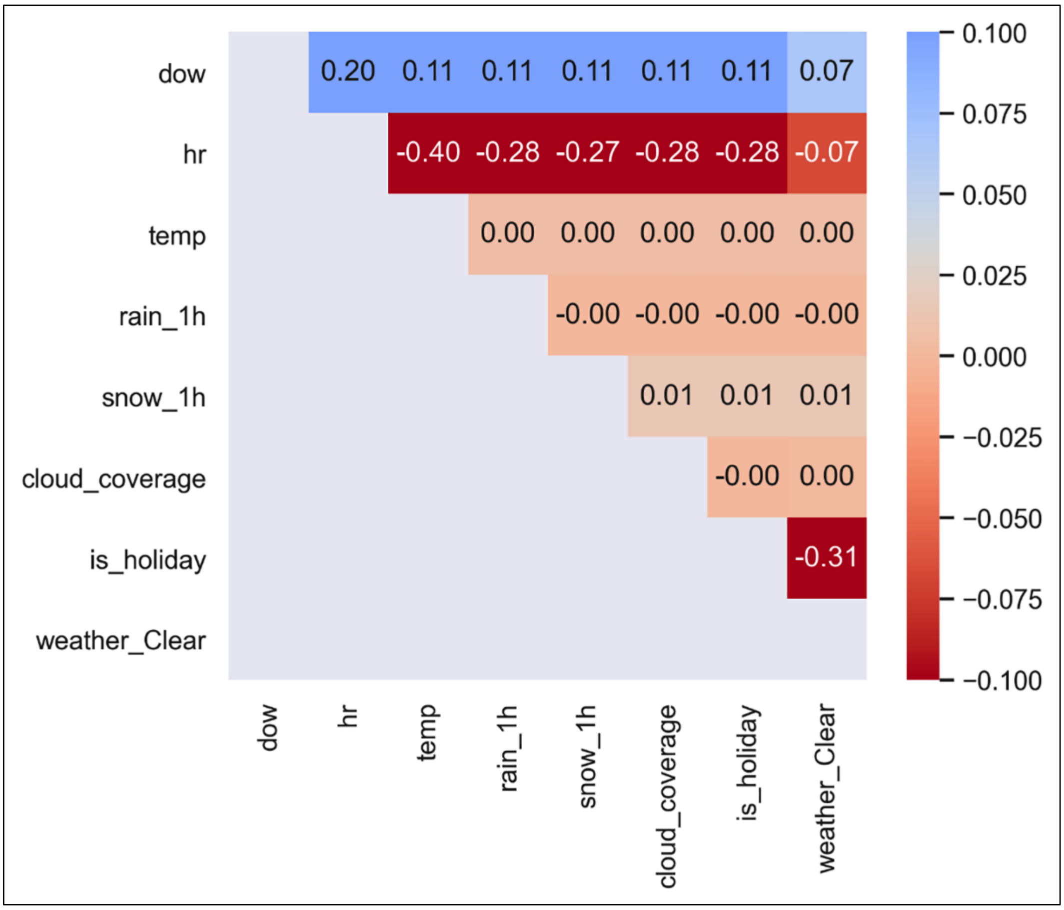

| features | 1st | Total | Total Conf | Mean of Input |
|---:|---:|---:|---:|---:|
| hr | 0.27 | 0.69 | 0.54 | 1.50 |
| dow | -0.14 | 0.62 | 0.60 | 1.99 |
| temp | -0.00 | 0.21 | 0.31 | 14.08 |
| snow_1h | -0.01 | 0.14 | 0.26 | 0.80 |
| is_holiday | 0.13 | 0.14 | 0.24 | 0.50 |
| weather_Clear | -0.00 | 0.07 | 0.17 | 0.50 |
| rain_1h | 0.00 | 0.00 | 0.00 | 10.51 |
| cloud_coverage | 0.00 | 0.00 | 0.00 | 49.69 |




| features | 1st | Total | Total Conf | Mean of Input |
|---:|---:|---:|---:|---:|
| dow | -0.03 | 1.31 | 2.47 | 1.99 |
| hr | 0.39 | 0.79 | 0.65 | 1.50 |
| is_holiday | 0.31 | 0.16 | 0.28 | 0.50 |
| temp | -0.00 | 0.12 | 0.13 | 14.08 |
| snow_1h | -0.01 | 0.11 | 0.20 | 0.80 |
| weather_Clear | -0.01 | 0.09 | 0.14 | 0.50 |
| cloud_coverage | 0.00 | 0.00 | 0.00 | 49.69 |
| rain_1h | 0.00 | 0.00 | 0.00 | 10.51 |

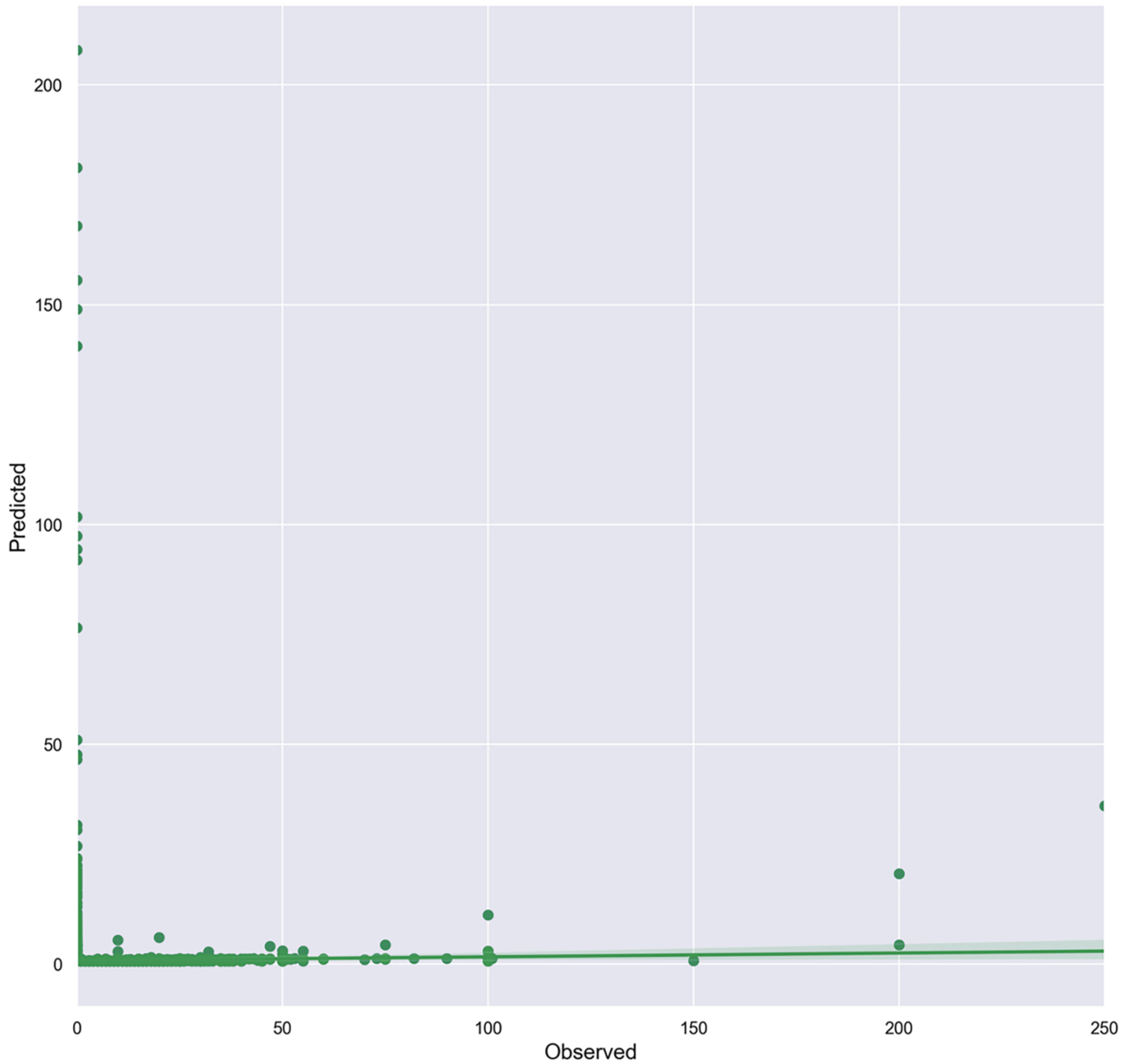


ST
4
3
2
1
0
−1
dow
hr
temp
rain_1h
snow_1h
cloud_coverage
is_holiday
weather_Clear

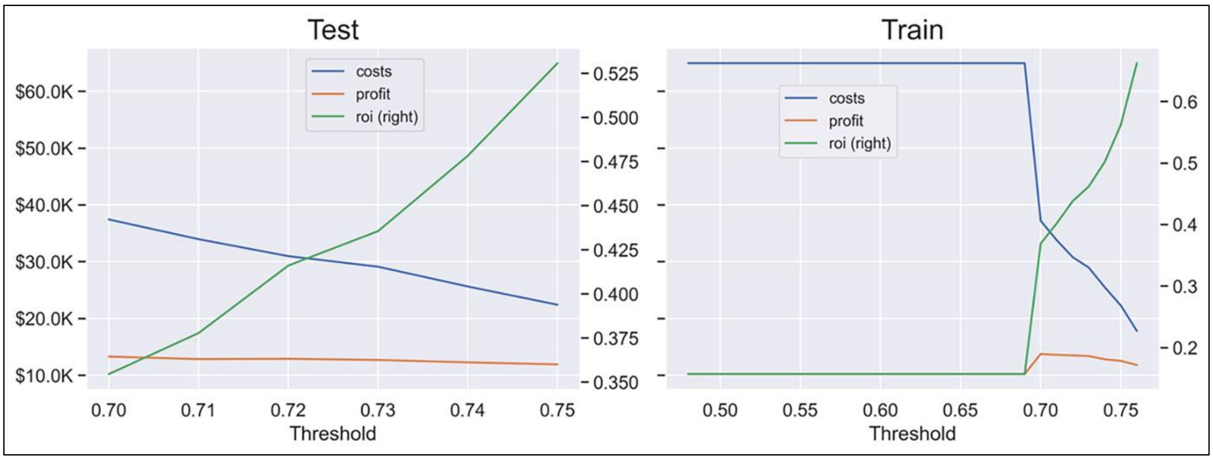



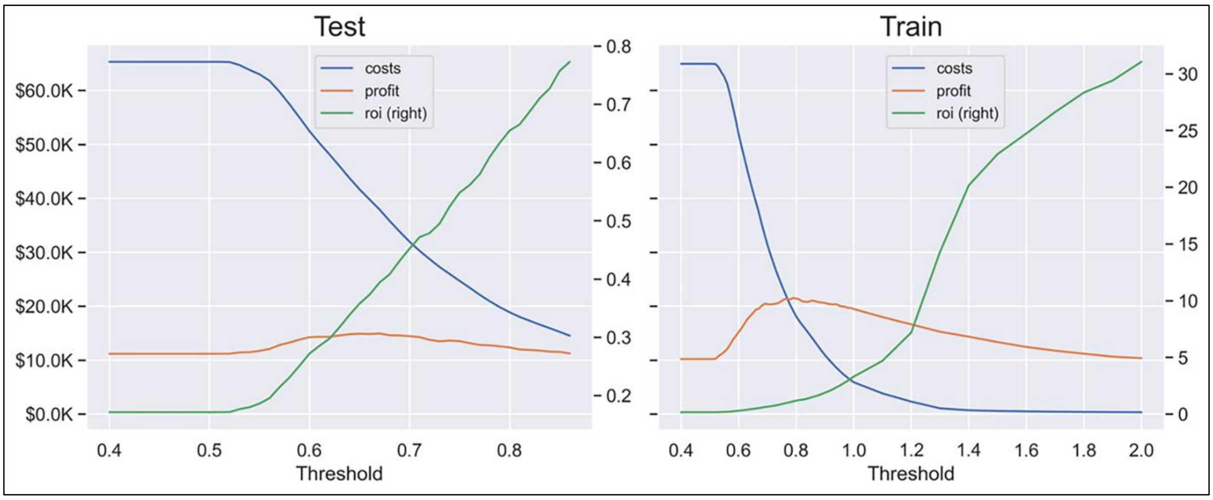
dow
hr
temp
rain_1h
snow_1h
cloud_coverage
is_holiday
weather_Clear
0.20 0.11 0.11 0.11 0.11 0.11 0.07
-0.40 -0.28 -0.27 -0.28 -0.28 -0.07
0.00 0.00 0.00 0.00 0.00
-0.00 -0.00 -0.00 -0.00
0.01 0.01 0.01
-0.00 0.00
-0.31
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100

# Chapter 10: Feature Selection and Engineering for Interpretability

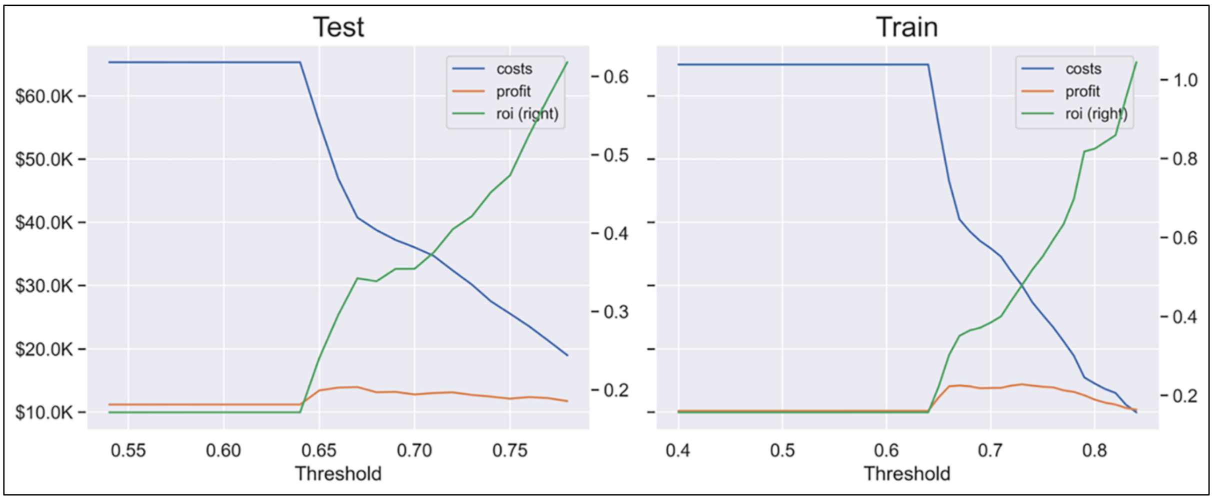



RMSE_train: 4.3210 RMSE_test: 4.6359  r2: -0.1084



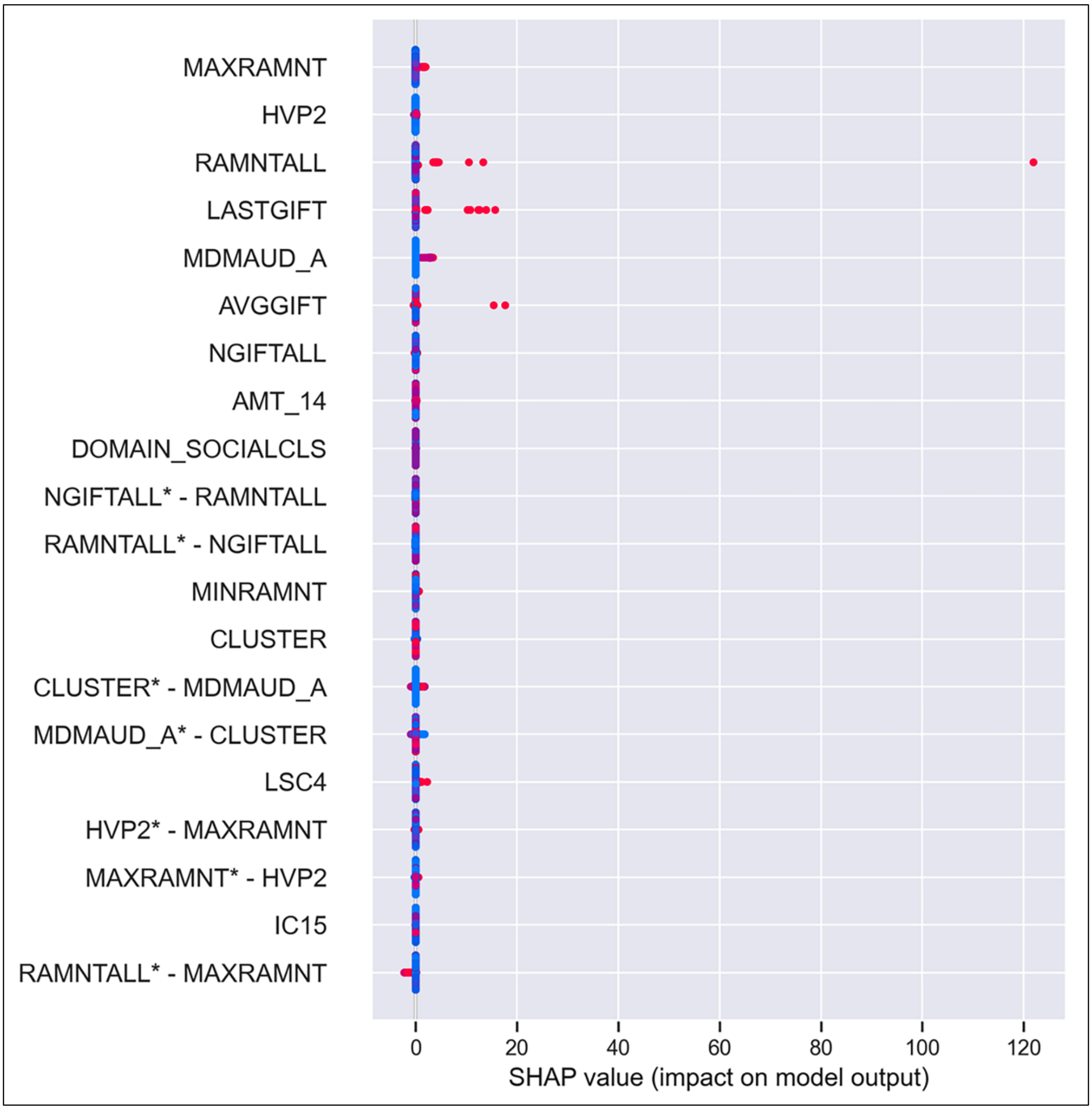

| | depth | fs | rmse_train | rmse_test | max_profit_train | max_profit_test | max_roi | min_costs | speed | num_feat |
|---|---|---|---|---|---|---|---|---|---|---|
| rf_12_all | 12 | all | 3.94 | 4.69 | $21,522 | $14,933 | 0.77 | $14,532 | 2.89 | 415 |
| rf_11_all | 11 | all | 3.99 | 4.69 | $19,904 | $15,142 | 0.76 | $14,928 | 2.74 | 398 |
| rf_10_all | 10 | all | 4.05 | 4.68 | $18,604 | $14,987 | 0.78 | $14,396 | 2.53 | 383 |
| rf_9_all | 9 | all | 4.10 | 4.68 | $17,452 | $14,778 | 0.80 | $13,997 | 2.20 | 346 |
| rf_8_all | 8 | all | 4.14 | 4.67 | $16,440 | $14,563 | 0.73 | $15,309 | 1.98 | 315 |
| rf_7_all | 7 | all | 4.18 | 4.66 | $15,435 | $14,186 | 0.66 | $17,165 | 1.71 | 277 |
| rf_6_all | 6 | all | 4.23 | 4.65 | $14,655 | $13,851 | 0.59 | $19,305 | 1.49 | 240 |
| rf_5_all | 5 | all | 4.27 | 4.64 | $14,242 | $13,752 | 0.59 | $19,199 | 1.18 | 201 |
| rf_4_all | 4 | all | 4.32 | 4.64 | $13,716 | $13,262 | 0.53 | $22,392 | 1.00 | 160 |



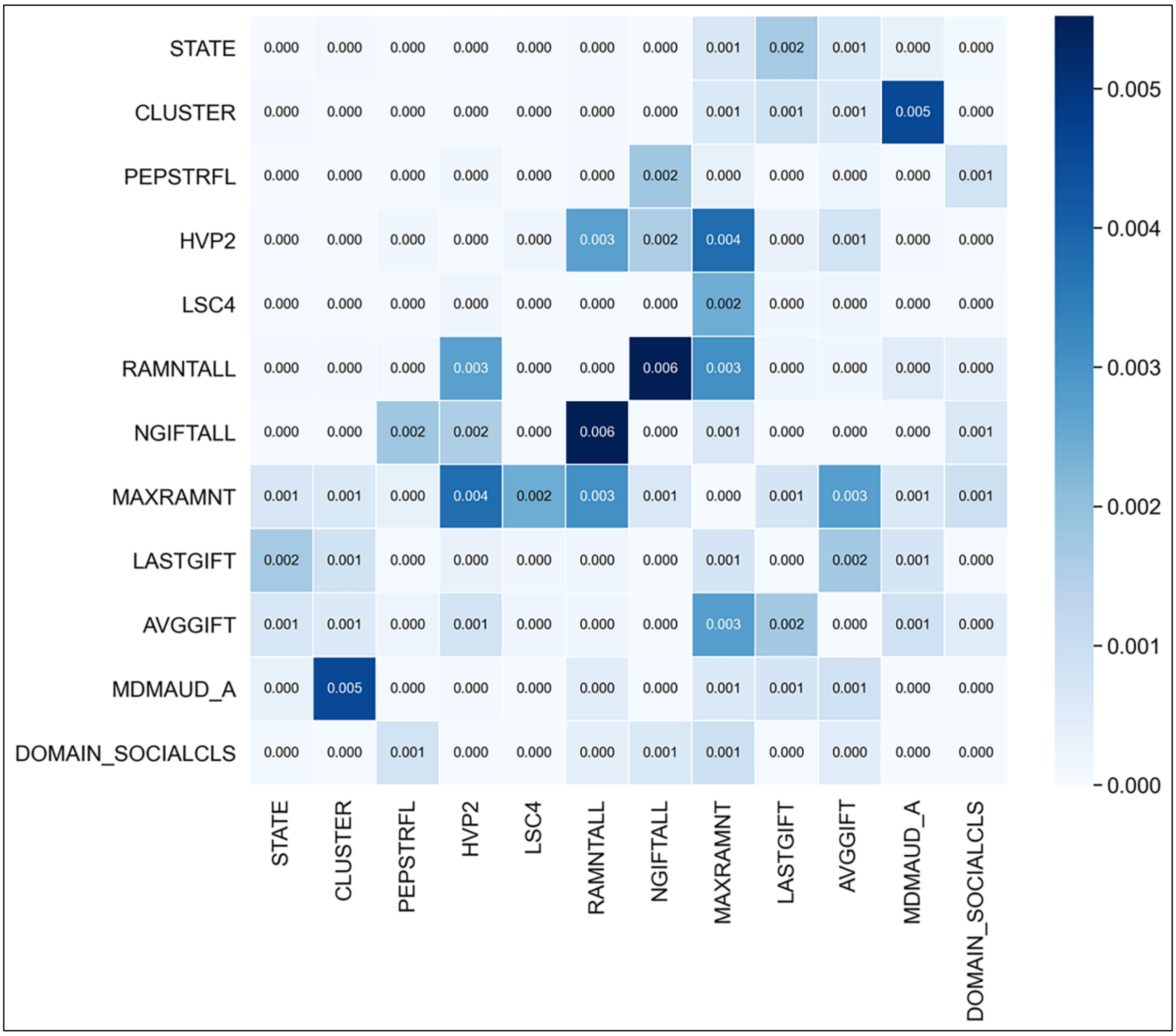

| | depth | fs | rmse_train | rmse_test | max_profit_train | max_profit_test | max_roi | min_costs | speed | total_feat | num_feat |
|---|---|---|---|---|---|---|---|---|---|---|---|
| rf_11_all | 11 | all | 3.99 | 4.69 | $19,904 | $15,142 | 0.76 | $14,928 | 2.74 | 435 | 398 |
| rf_10_all | 10 | all | 4.05 | 4.68 | $18,604 | $14,987 | 0.78 | $14,396 | 2.53 | 435 | 383 |
| rf_12_all | 12 | all | 3.94 | 4.69 | $21,522 | $14,933 | 0.77 | $14,532 | 2.89 | 435 | 415 |
| rf_11_f-corr | 11 | f-corr | 3.98 | 4.67 | $19,924 | $14,895 | 0.77 | $14,593 | 2.47 | 419 | 404 |
| rf_9_all | 9 | all | 4.10 | 4.68 | $17,452 | $14,778 | 0.80 | $13,997 | 2.20 | 435 | 346 |
| rf_8_all | 8 | all | 4.14 | 4.67 | $16,440 | $14,563 | 0.73 | $15,309 | 1.98 | 435 | 315 |
| rf_7_all | 7 | all | 4.18 | 4.66 | $15,435 | $14,186 | 0.66 | $17,165 | 1.71 | 435 | 277 |
| rf_5_f-mic | 5 | f-mic | 4.31 | 4.60 | $14,367 | $13,944 | 0.62 | $18,971 | 0.41 | 160 | 105 |
| rf_6_all | 6 | all | 4.23 | 4.65 | $14,655 | $13,851 | 0.59 | $19,305 | 1.49 | 435 | 240 |
| rf_5_all | 5 | all | 4.27 | 4.64 | $14,242 | $13,752 | 0.59 | $19,199 | 1.18 | 435 | 201 |
| rf_4_all | 4 | all | 4.32 | 4.64 | $13,716 | $13,262 | 0.53 | $22,392 | 1.00 | 435 | 160 |

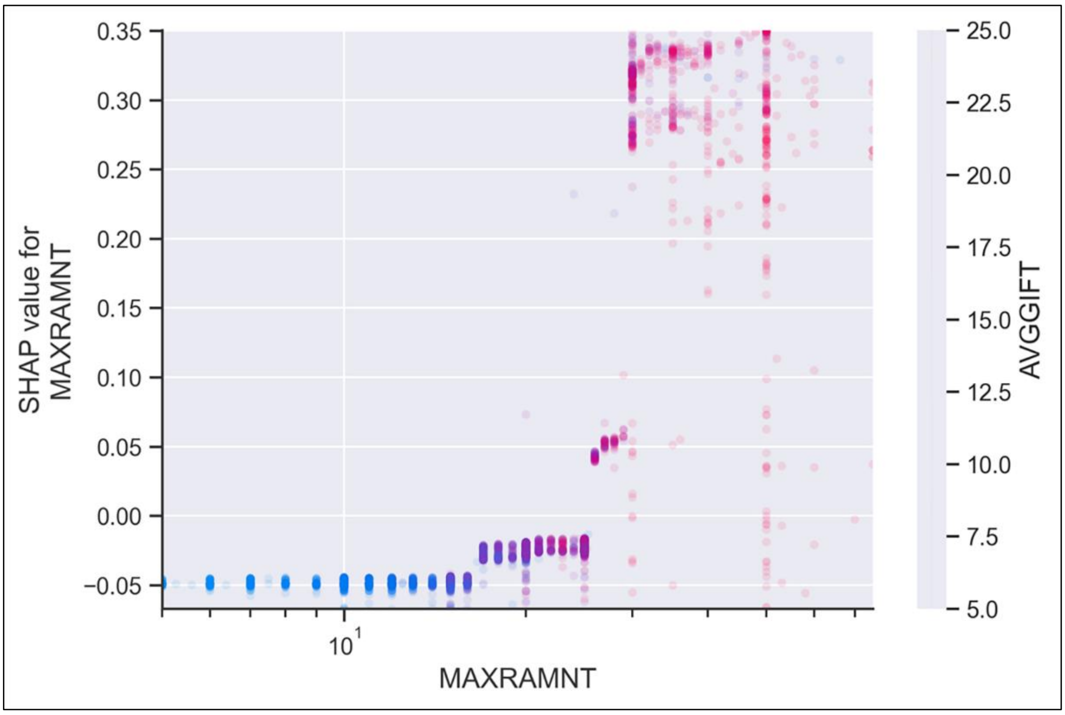
| | depth | fs | rmse_train | rmse_test | max_profit_train | max_profit_test | max_roi | min_costs | speed | total_feat | num_feat |
|---|---|---|---|---|---|---|---|---|---|---|---|
| rf_11_all | 11 | all | 3.99 | 4.69 | $19,904 | $15,142 | 0.76 | $14,928 | 2.74 | 435 | 398 |
| rf_10_all | 10 | all | 4.05 | 4.68 | $18,604 | $14,987 | 0.78 | $14,396 | 2.53 | 435 | 383 |
| rf_12_all | 12 | all | 3.94 | 4.69 | $21,522 | $14,933 | 0.77 | $14,532 | 2.89 | 435 | 415 |
| rf_11_f-corr | 11 | f-corr | 3.98 | 4.67 | $19,924 | $14,895 | 0.77 | $14,593 | 2.47 | 419 | 404 |
| rf_9_all | 9 | all | 4.10 | 4.68 | $17,452 | $14,778 | 0.80 | $13,997 | 2.20 | 435 | 346 |
| rf_5_e-llarsic | 5 | e-llarsic | 4.28 | 4.45 | $15,168 | $14,768 | 0.56 | $20,441 | 0.30 | 111 | 87 |
| rf_8_all | 8 | all | 4.14 | 4.67 | $16,440 | $14,563 | 0.73 | $15,309 | 1.98 | 435 | 315 |
| rf_6_e-logl2 | 6 | e-logl2 | 4.28 | 4.60 | $15,353 | $14,199 | 0.67 | $16,904 | 0.31 | 87 | 84 |
| rf_7_all | 7 | all | 4.18 | 4.66 | $15,435 | $14,186 | 0.66 | $17,165 | 1.71 | 435 | 277 |
| rf_5_f-mic | 5 | f-mic | 4.31 | 4.60 | $14,367 | $13,944 | 0.62 | $18,971 | 0.41 | 160 | 105 |
| rf_6_all | 6 | all | 4.23 | 4.65 | $14,655 | $13,851 | 0.59 | $19,305 | 1.49 | 435 | 240 |
| rf_5_all | 5 | all | 4.27 | 4.64 | $14,242 | $13,752 | 0.59 | $19,199 | 1.18 | 435 | 201 |
| rf_4_e-llars | 4 | e-llars | 4.36 | 4.45 | $14,014 | $13,633 | 0.52 | $22,906 | 0.04 | 8 | 8 |
| rf_4_all | 4 | all | 4.32 | 4.64 | $13,716 | $13,262 | 0.53 | $22,392 | 1.00 | 435 | 160 |
| rf_3_e-lasso | 3 | e-lasso | 4.46 | 4.49 | $14,167 | $12,930 | 0.51 | $22,249 | 0.03 | 7 | 7 |

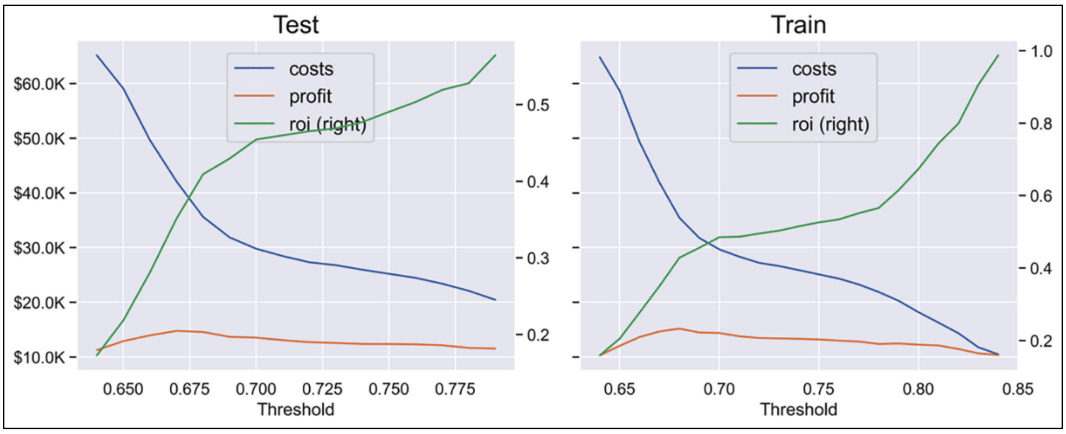
| | depth | fs | rmse_train | rmse_test | max_profit_train | max_profit_test | max_roi | min_costs | speed | total_feat | num_feat |
|---|---|---|---|---|---|---|---|---|---|---|---|
| rf_5_e-llarsic | 5 | e-llarsic | 4.28 | 4.45 | $15,168 | $14,768 | 0.56 | $20,441 | 0.30 | 111 | 87 |
| rf_6_h-rfe-lda | 6 | h-rfe-lda | 4.28 | 4.50 | $15,705 | $14,410 | 0.68 | $16,542 | 0.47 | 145 | 115 |
| rf_6_e-logl2 | 6 | e-logl2 | 4.28 | 4.60 | $15,353 | $14,199 | 0.67 | $16,904 | 0.31 | 87 | 84 |
| rf_6_a-ga-rf | 6 | a-ga-rf | 4.26 | 4.67 | $15,710 | $14,004 | 0.72 | $15,987 | 0.47 | 134 | 111 |
| rf_5_f-mic | 5 | f-mic | 4.31 | 4.60 | $14,367 | $13,944 | 0.62 | $18,971 | 0.41 | 160 | 105 |
| rf_6_all | 6 | all | 4.23 | 4.65 | $14,655 | $13,851 | 0.59 | $19,305 | 1.49 | 435 | 240 |
| rf_5_all | 5 | all | 4.27 | 4.64 | $14,242 | $13,752 | 0.59 | $19,199 | 1.18 | 435 | 201 |
| rf_4_e-llars | 4 | e-llars | 4.36 | 4.45 | $14,014 | $13,633 | 0.52 | $22,906 | 0.04 | 8 | 8 |
| rf_5_a-shap | 5 | a-shap | 4.28 | 4.51 | $14,068 | $13,350 | 0.59 | $18,935 | 0.35 | 120 | 102 |
| rf_5_w-sfs-lda | 5 | w-sfs-lda | 4.33 | 4.47 | $13,763 | $13,262 | 0.46 | $24,553 | 0.29 | 100 | 81 |
| rf_4_all | 4 | all | 4.32 | 4.64 | $13,716 | $13,262 | 0.53 | $22,392 | 1.00 | 435 | 160 |
| rf_3_e-lasso | 3 | e-lasso | 4.46 | 4.49 | $14,167 | $12,930 | 0.51 | $22,249 | 0.03 | 7 | 7 |



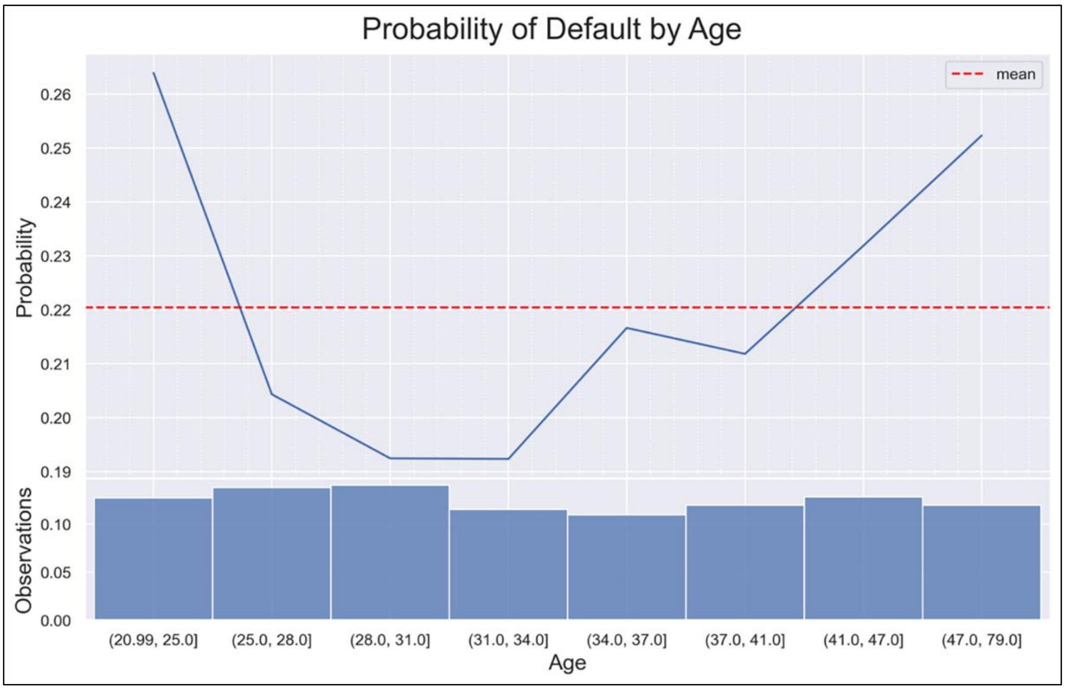
MAXRAMNT
HVP2
RAMNTALL
LASTGIFT
MDMAUD_A
AVGGIFT
NGIFTALL
AMT_14
DOMAIN_SOCIALCLS
NGIFTALL* - RAMNTALL
RAMNTALL* - NGIFTALL
MINRAMNT
CLUSTER
CLUSTER* - MDMAUD_A
MDMAUD_A* - CLUSTER
LSC4
HVP2* - MAXRAMNT
MAXRAMNT* - HVP2
IC15
RAMNTALL* - MAXRAMNT
0
20
40
60
80
100
120
SHAP value (impact on model output)

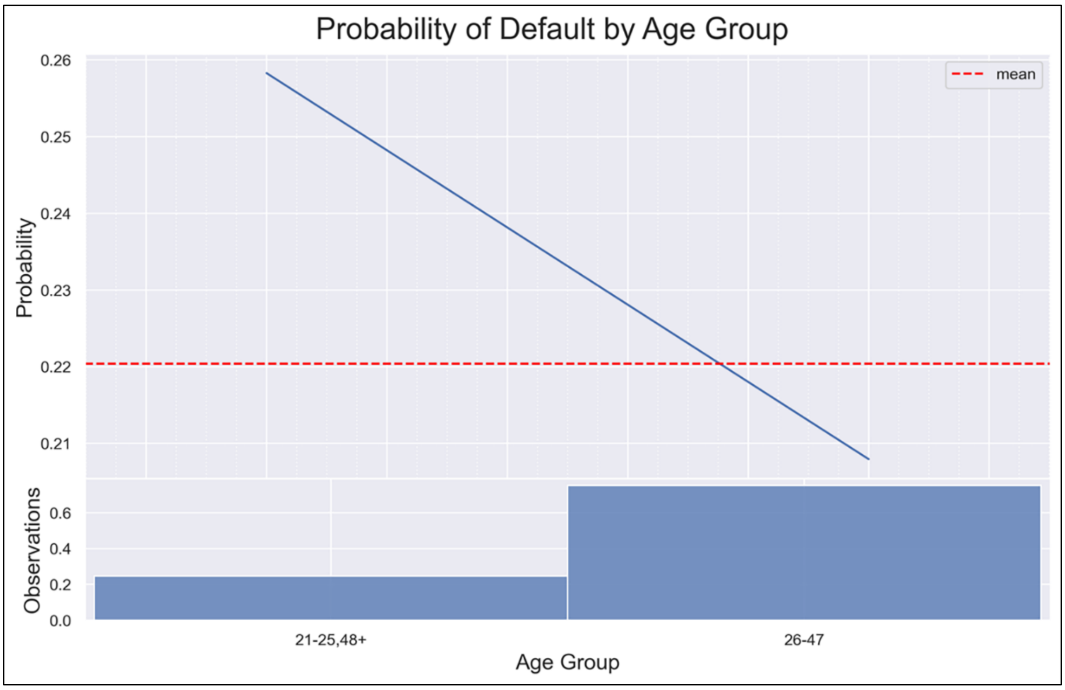

STATE 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.001 0.002 0.001 0.000 0.000
CLUSTER 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.001 0.001 0.001 0.005 0.000
PEPSTRFL 0.000 0.000 0.000 0.000 0.000 0.000 0.002 0.000 0.000 0.000 0.000 0.001
HVP2 0.000 0.000 0.000 0.000 0.000 0.003 0.002 0.004 0.000 0.001 0.000 0.000
LSC4 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.002 0.000 0.000 0.000 0.000
RAMNTALL 0.000 0.000 0.000 0.003 0.000 0.000 0.006 0.003 0.000 0.000 0.000 0.000
NGIFTALL 0.000 0.000 0.002 0.002 0.000 0.006 0.000 0.001 0.000 0.000 0.000 0.001
MAXRAMNT 0.001 0.001 0.000 0.004 0.002 0.003 0.001 0.000 0.001 0.003 0.001 0.001
LASTGIFT 0.002 0.001 0.000 0.000 0.000 0.000 0.000 0.001 0.000 0.002 0.001 0.000
AVGGIFT 0.001 0.001 0.000 0.001 0.000 0.000 0.000 0.003 0.002 0.000 0.001 0.000
MDMAUD_A 0.000 0.005 0.000 0.000 0.000 0.000 0.000 0.001 0.001 0.001 0.000 0.000
DOMAIN_SOCIALCLS 0.000 0.000 0.001 0.000 0.000 0.000 0.001 0.001 0.000 0.000 0.000 0.000
STATE
CLUSTER
PEPSTRFL
HVP2
LSC4
RAMNTALL
NGIFTALL
MAXRAMNT
LASTGIFT
AVGGIFT
MDMAUD_A
DOMAIN_SOCIALCLS
0.005
0.004
0.003
0.002
0.001
0.000


0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
SHAP value for MAXRAMNT
$10^1$
MAXRAMNT
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
AVGGIFT

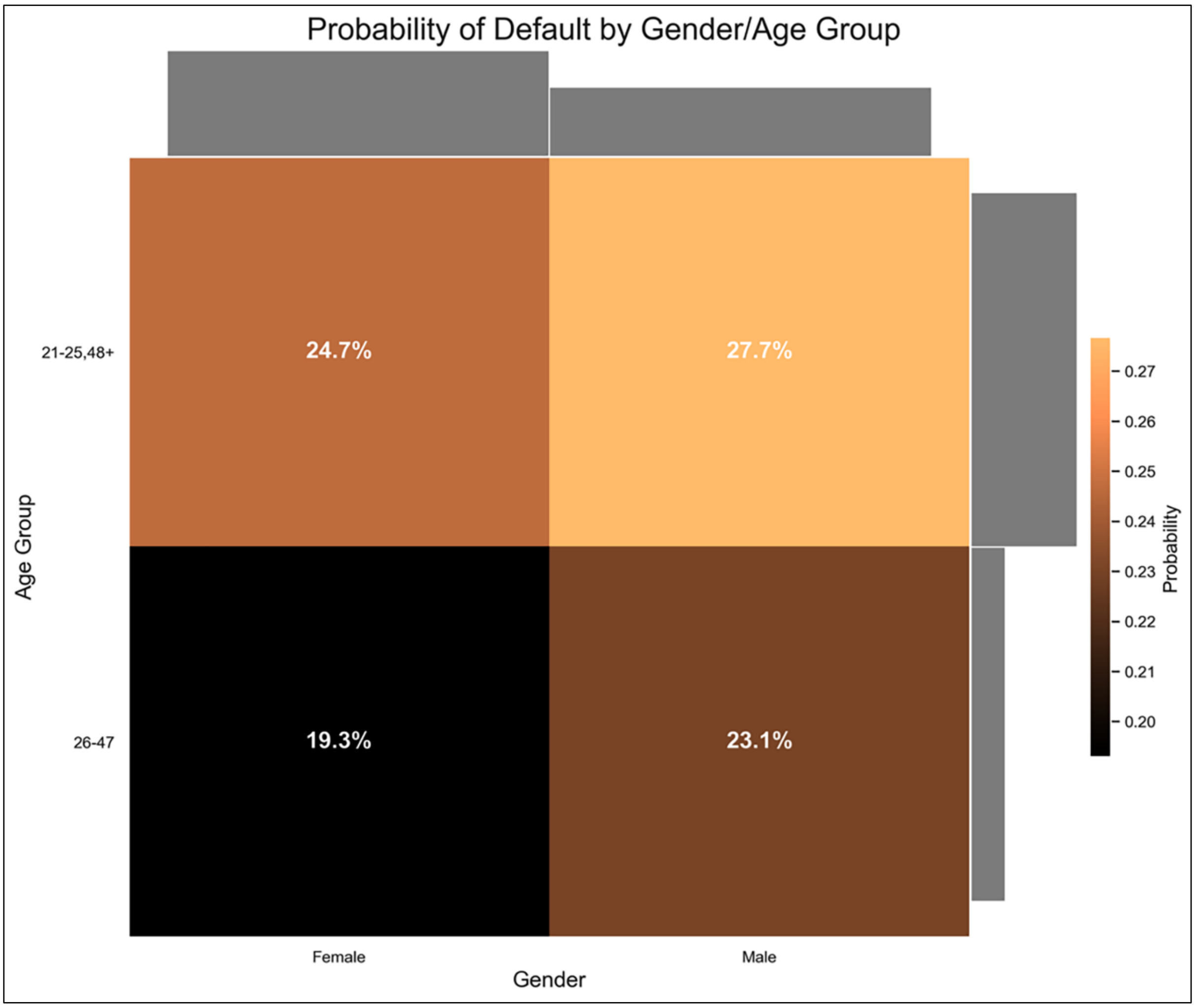



# Chapter 11: Bias Mitigation and Causal Inference Methods

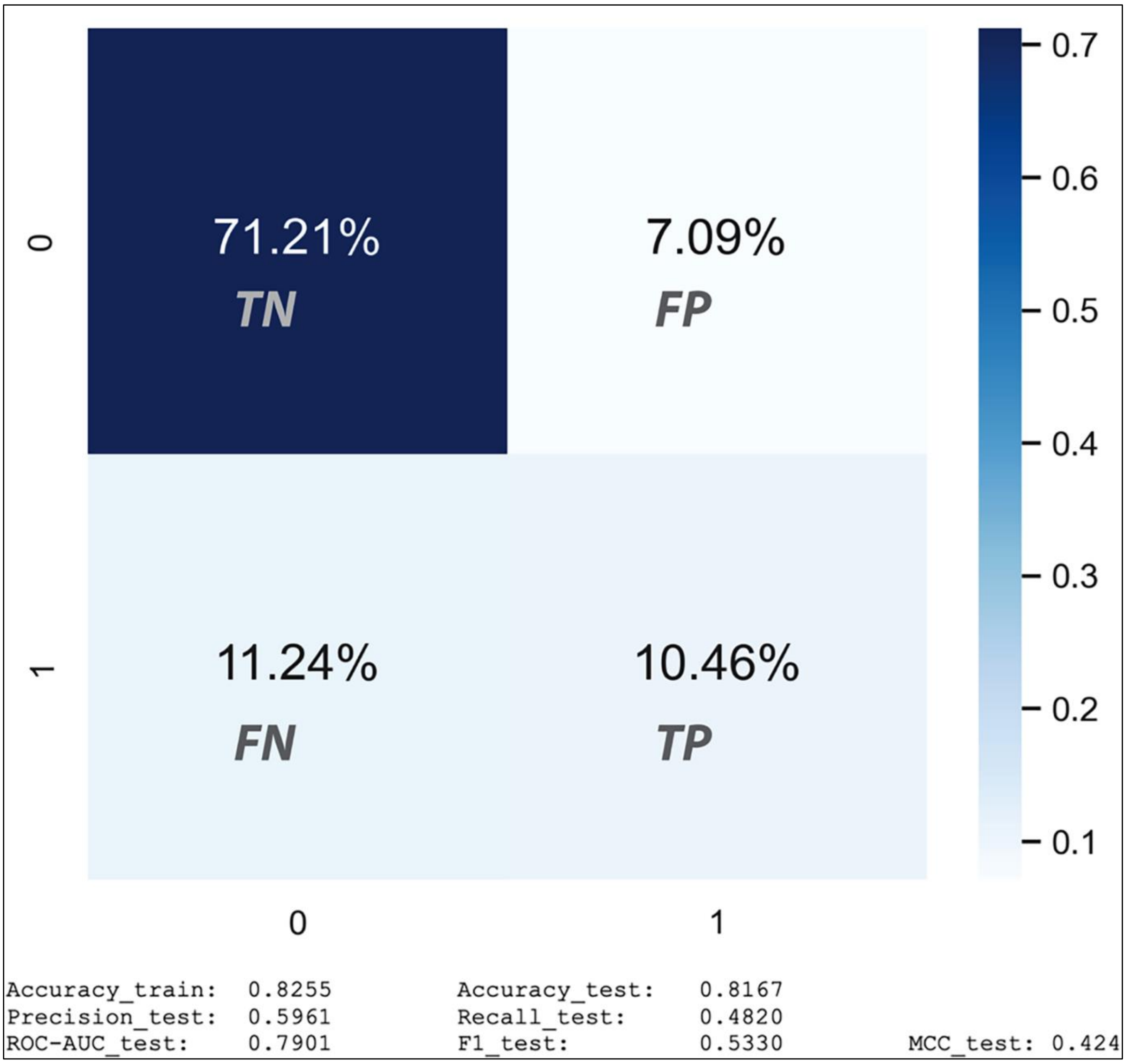




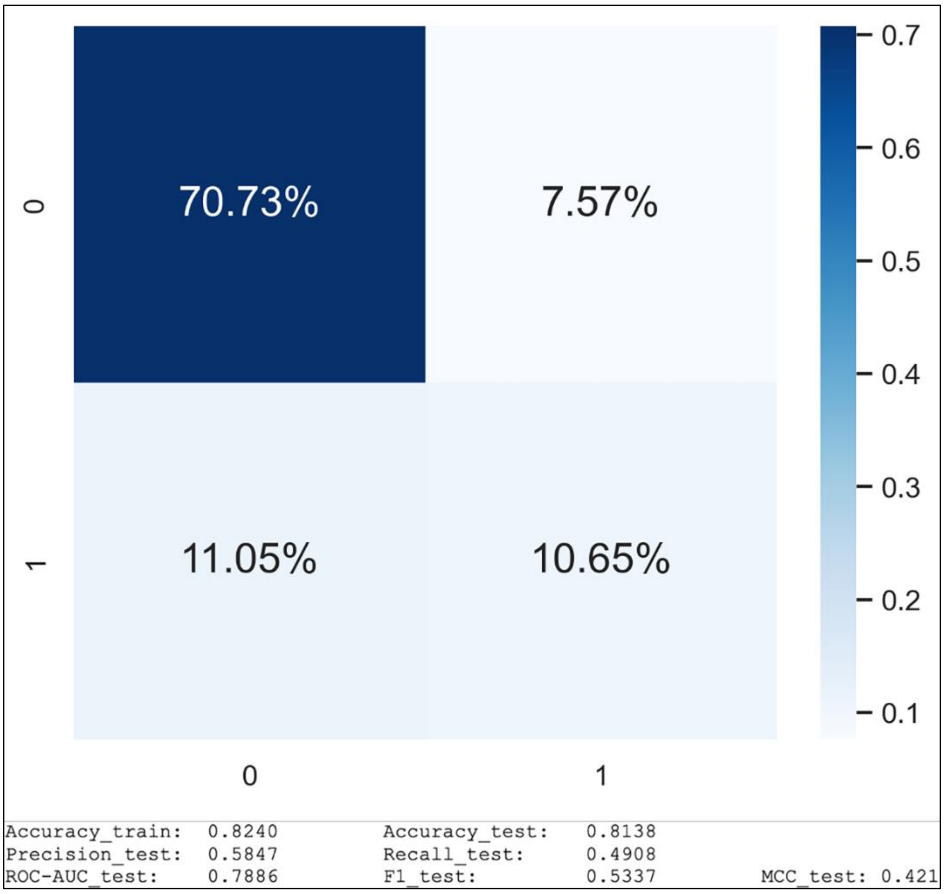
Probability of Default by Age Group
mean
Probability
0.26
0.25
0.24
0.23
0.22
0.21
Observations
0.6
0.4
0.2
0.0
21-25,48+
26-47
Age Group

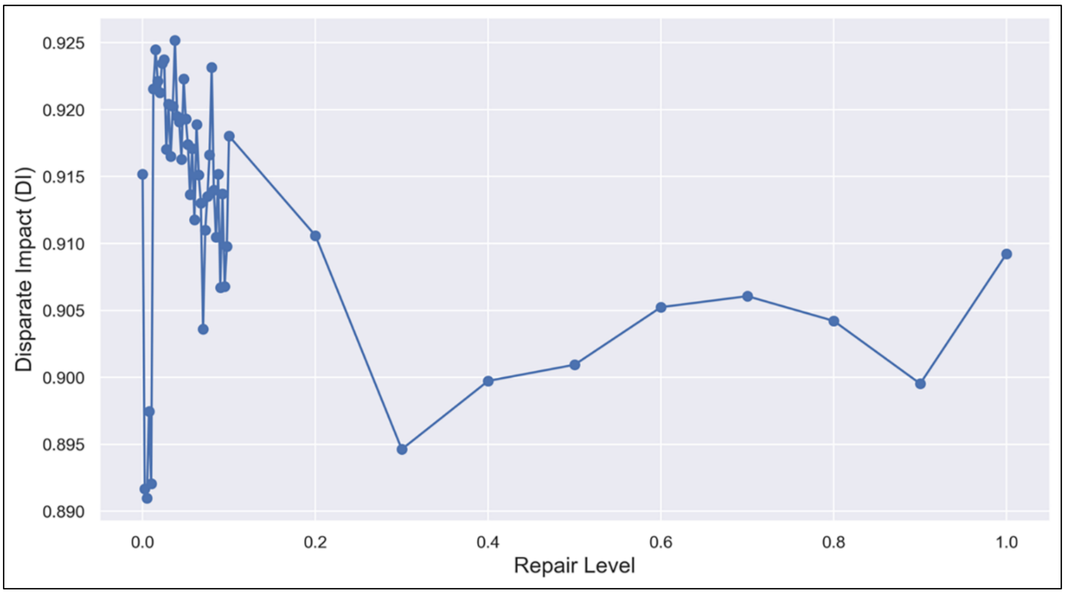



Probability of Default by Gender/Age Group
21-25,48+
26-47
24.7%
27.7%
19.3%
23.1%
Age Group
Female
Male
Gender
Probability
0.27
0.26
0.25
0.24
0.23
0.22
0.21
0.20

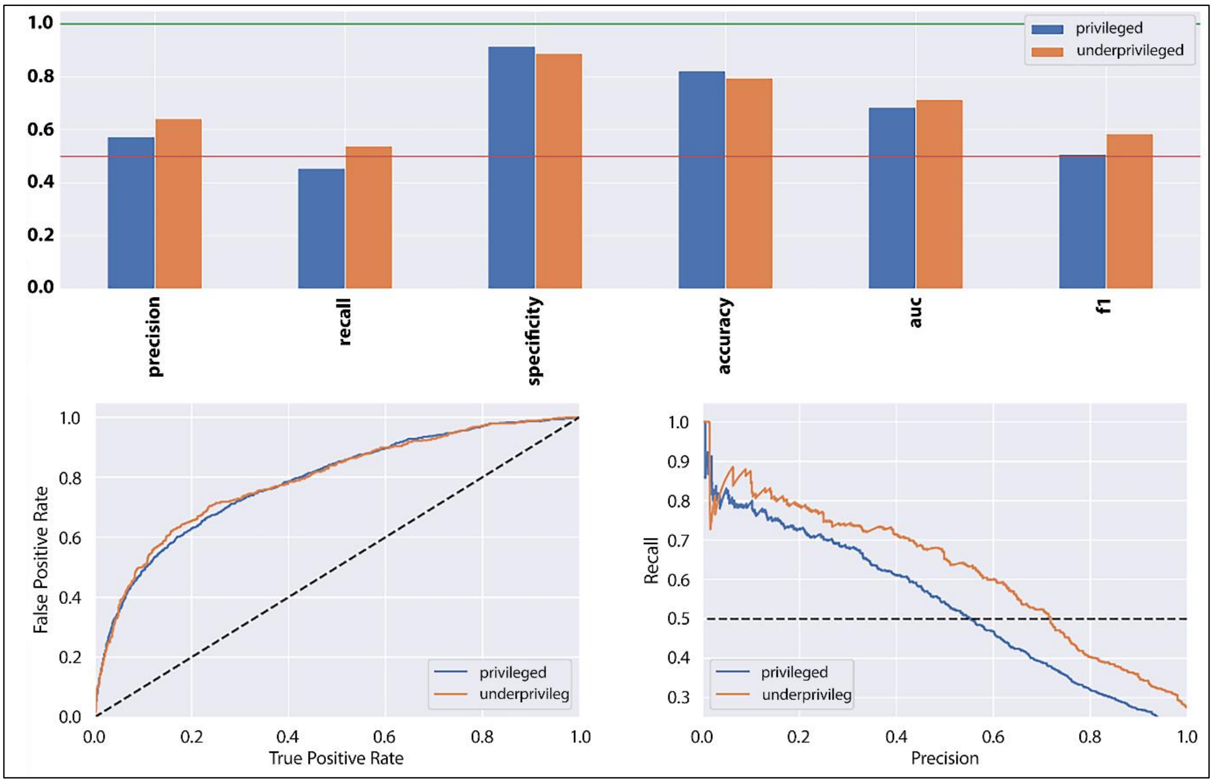


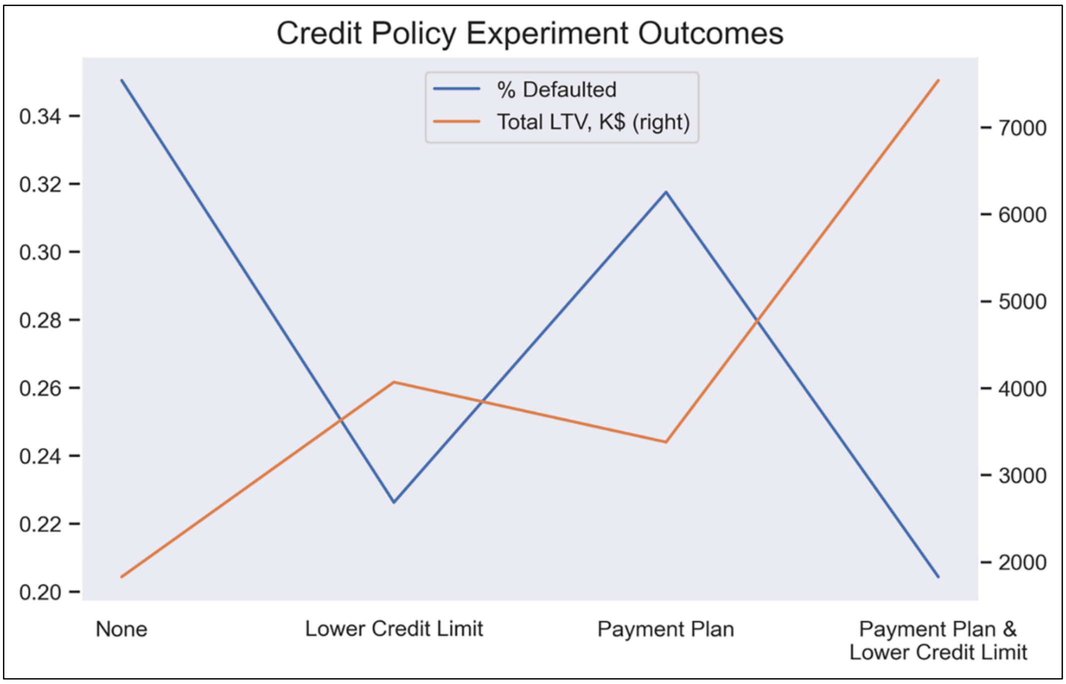
71.21%
TN
7.09%
FP
11.24%
FN
10.46%
TP
0
1
0
1
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Accuracy_train: 0.8255
Accuracy_test: 0.8167
Precision_test: 0.5961
Recall_test: 0.4820
ROC-AUC_test: 0.7901
F1_test: 0.5330
MCC_test: 0.424


70.73%
7.57%
11.05%
10.65%
0
1
0
1
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Accuracy_train: 0.8240
Accuracy_test: 0.8138
Precision_test: 0.5847
Recall_test: 0.4908
ROC-AUC_test: 0.7886
F1_test: 0.5337
MCC_test: 0.421

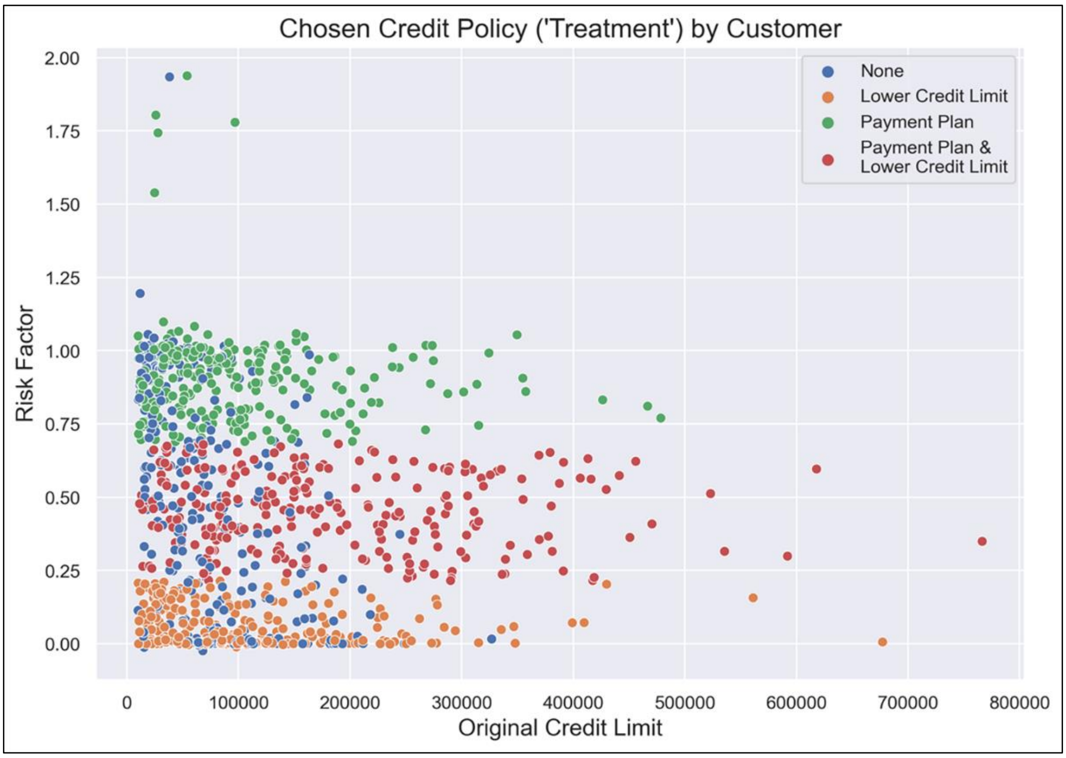

Disparate Impact (DI)
0.925
0.920
0.915
0.910
0.905
0.900
0.895
0.890
0.0
0.2
0.4
0.6
0.8
1.0
Repair Level

| | accuracy_train | accuracy_test | f1_test | mcc_test | SPD | DI | AOD | EOD | DFBA |
|---|---|---|---|---|---|---|---|---|---|
| **dt_2_gf** | 82.1% | **82.6%** | 48.1% | 0.413 | -0.055 | 0.939 | -0.043 | -0.022 | 0.252 |
| **lgb_0_base** | 82.6% | 81.7% | 53.3% | **0.424** | -0.068 | 0.919 | -0.055 | -0.026 | 0.233 |
| **lgb_1_rw** | 82.4% | 81.4% | **53.4%** | 0.421 | -0.037 | 0.955 | -0.017 | **-0.002** | 0.035 |
| **lgb_1_dir** | 82.4% | 81.3% | 53.0% | 0.417 | -0.062 | 0.925 | -0.049 | -0.021 | 0.255 |
| **lgb_2_egr** | 82.7% | 81.1% | 52.3% | 0.410 | -0.039 | 0.953 | -0.013 | -0.012 | -0.082 |
| **lgb_3_epp** | 82.6% | 81.1% | 51.8% | 0.406 | **-0.026** | **0.969** | **-0.001** | 0.002 | **-0.014** |
| **lgb_3_cpp** | 82.6% | 26.2% | 21.3% | -0.306 | -0.071 | 0.761 | -0.064 | -0.126 | 0.043 |

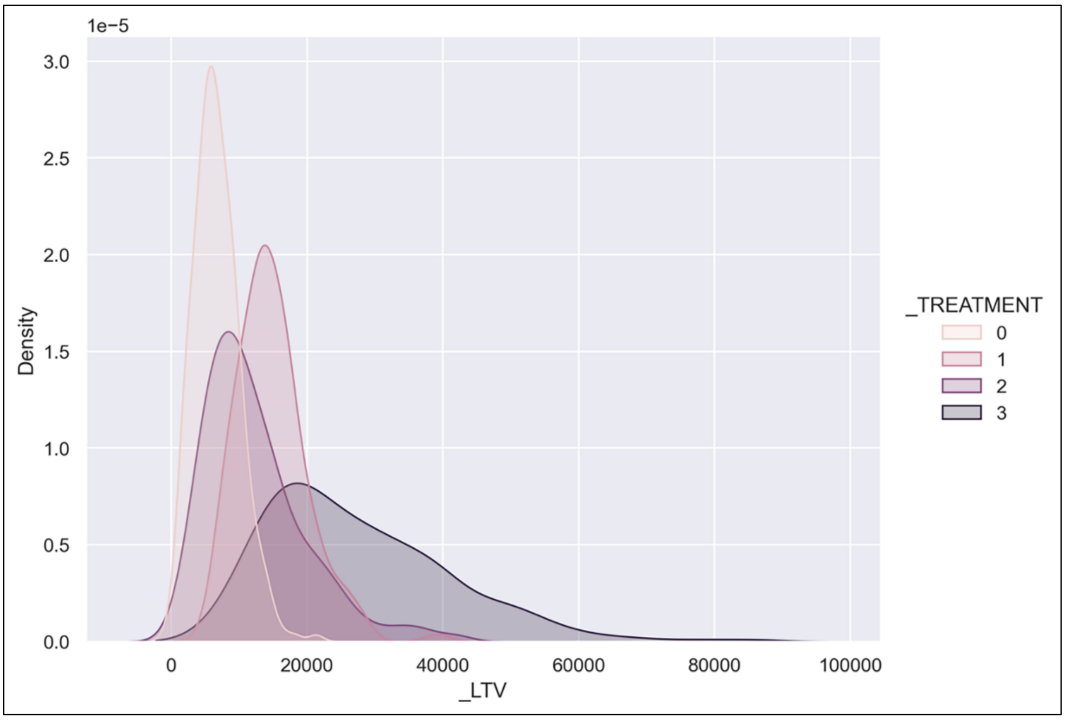


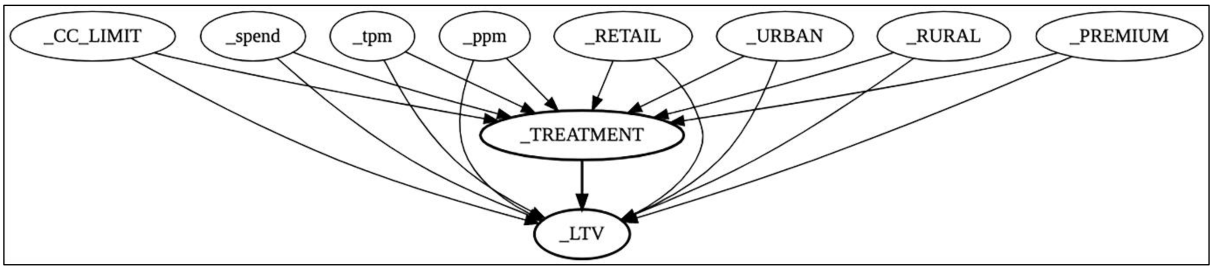


Credit Policy Experiment Outcomes
% Defaulted
Total LTV, K$ (right)
0.34
0.32
0.30
0.28
0.26
0.24
0.22
0.20
7000
6000
5000
4000
3000
2000
None
Lower Credit Limit
Payment Plan
Payment Plan &
Lower Credit Limit

Chosen Credit Policy ('Treatment') by Customer
None
Lower Credit Limit
Payment Plan
Payment Plan &
Lower Credit Limit
Risk Factor
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
100000
200000
300000
400000
500000
600000
700000
800000
Original Credit Limit




Treatment: Lower Credit Limit

**Coefficient Results**

| | point_estimate | stderr | zstat | pvalue | ci_lower | ci_upper |
|---|---|---|---|---|---|---|
| **_CC_LIMIT** | 0.006 | 0.02 | 0.322 | 0.747 | -0.032 | 0.045 |

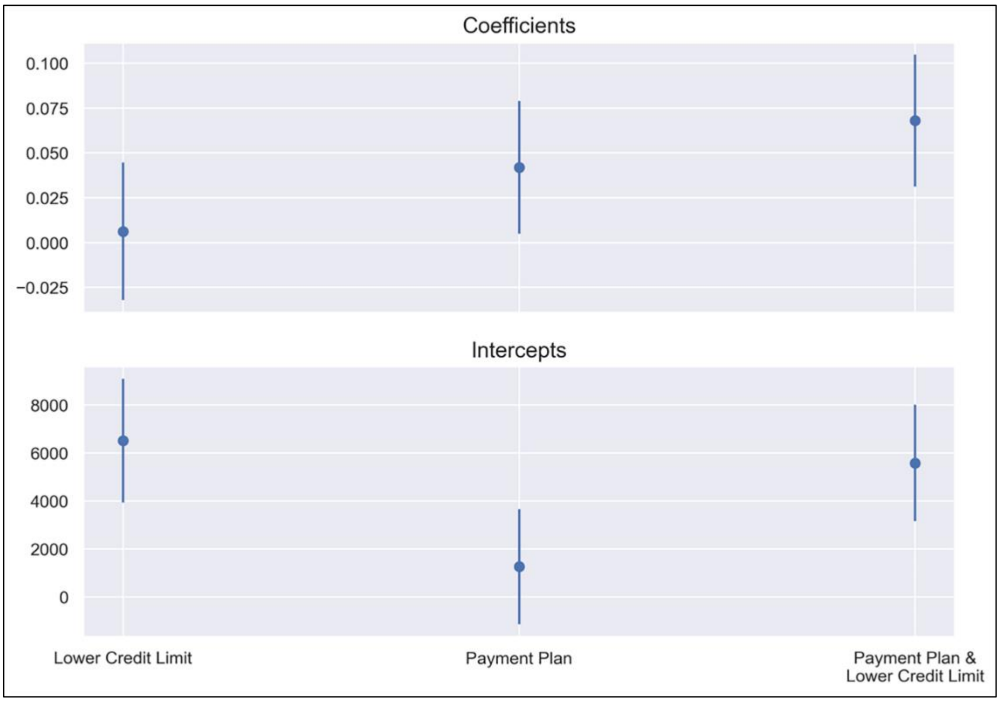
**CATE Intercept Results**

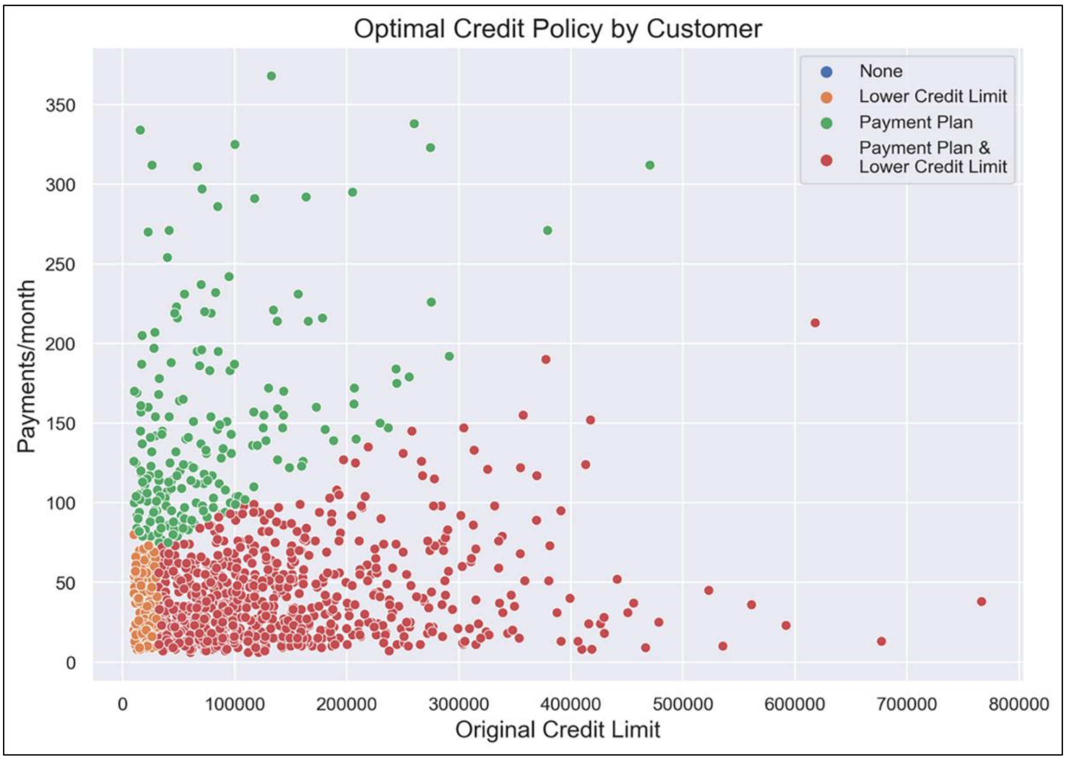
| | point_estimate | stderr | zstat | pvalue | ci_lower | ci_upper |
|---|---|---|---|---|---|---|
| **cate_intercept** | 6514.633 | 1312.662 | 4.963 | 0.0 | 3941.863 | 9087.404 |

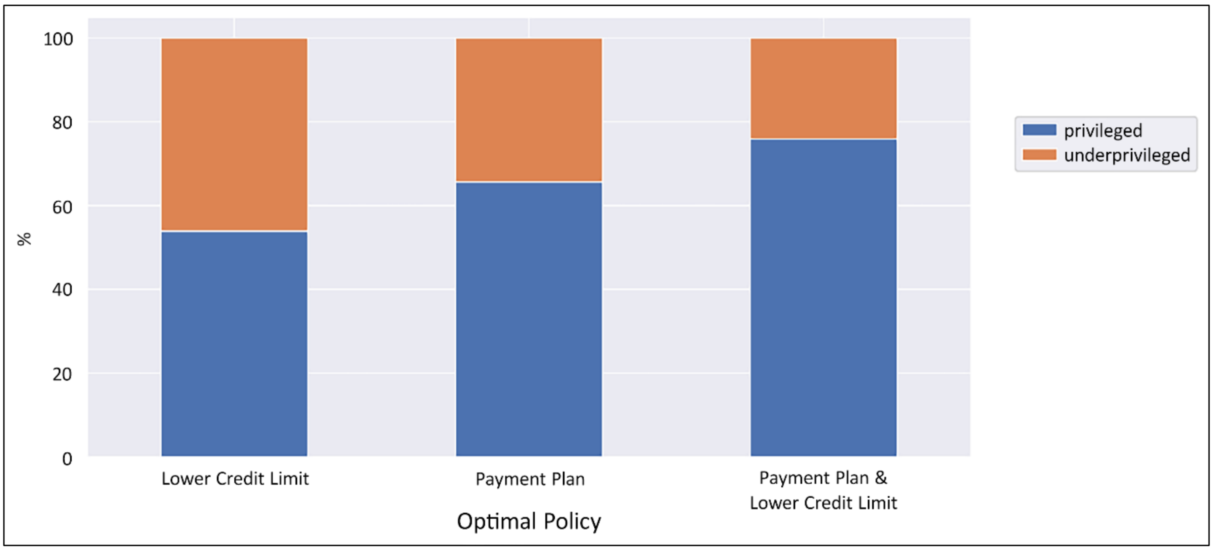
A linear parametric conditional average treatment effect (CATE) model was fitted:

$Y=\Theta(X)\cdot T+g(X,W)+\epsilon$Y=Θ(X)·T+g(X,W)+ϵ

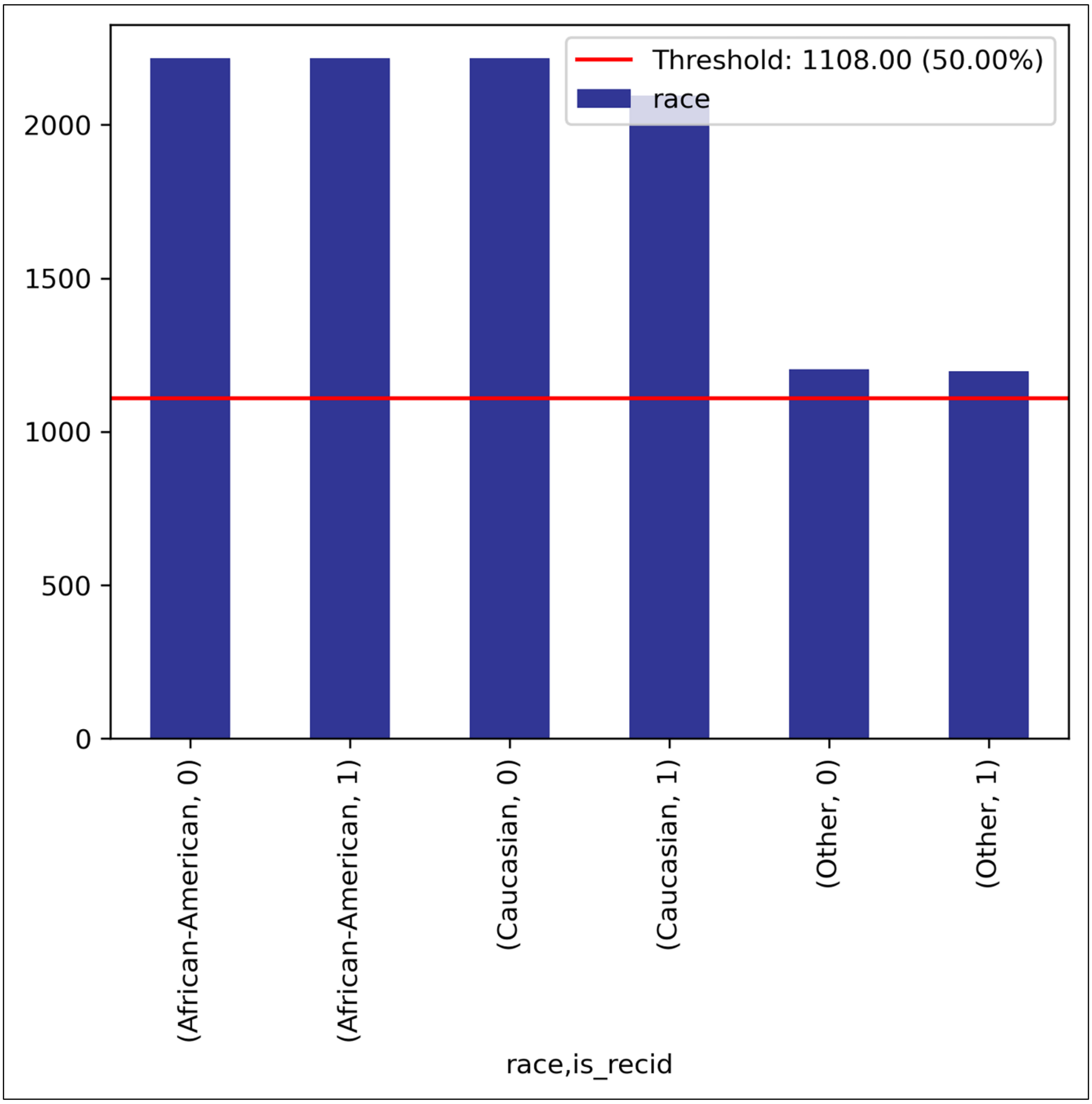
where $T$T is the one-hot-encoding of the discrete treatment and for every outcome $i$i and treatment $j$j the CATE $\Theta_{ij}(X)$Θij(X) has the form:


Coefficients
0.100
0.075
0.050
0.025
0.000
−0.025
Intercepts
8000
6000
4000
2000
0
Lower Credit Limit
Payment Plan
Payment Plan &
Lower Credit Limit

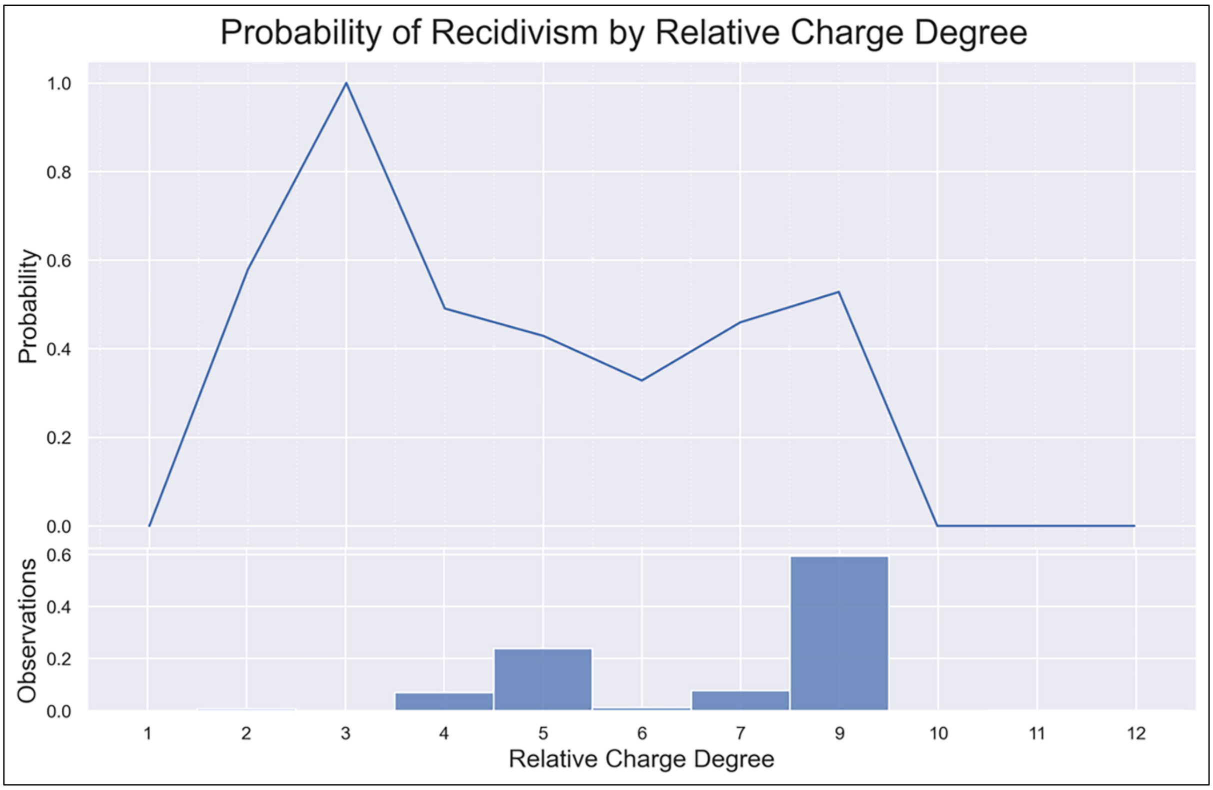

Optimal Credit Policy by Customer
None
Lower Credit Limit
Payment Plan
Payment Plan &
Lower Credit Limit
Payments/month
350
300
250
200
150
100
50
0
0
100000
200000
300000
400000
500000
600000
700000
800000
Original Credit Limit

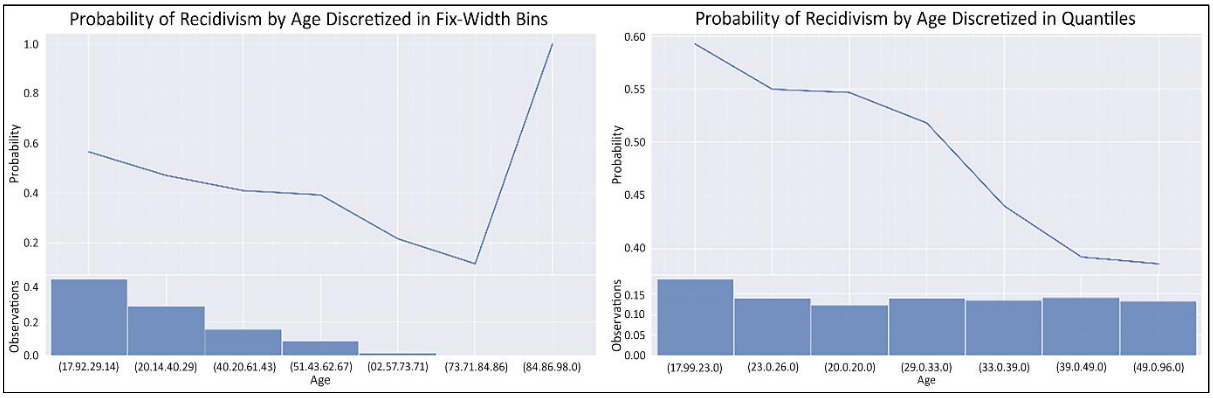



# Chapter 12: Monotonic Constraints and Model Tuning for Interpretability

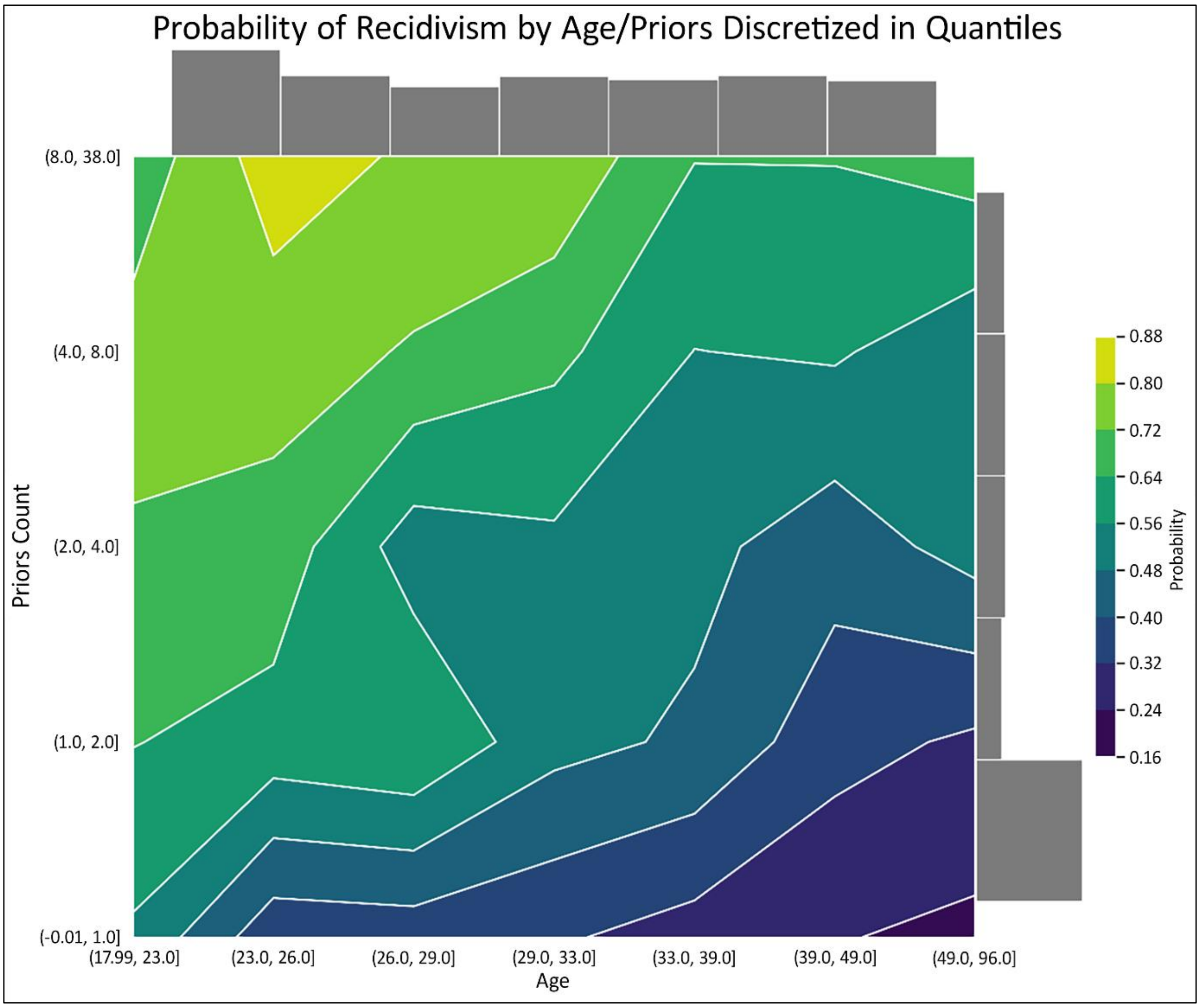

Threshold: 1108.00 (50.00%)
race
2000
1500
1000
500
0
(African-American, 0)
(African-American, 1)
(Caucasian, 0)
(Caucasian, 1)
(Other, 0)
(Other, 1)
race,is_recid

| feature | correlation_to_target |
|---|---|
| sex | 0.093255 |
| age | -0.155838 |
| race | -0.004598 |
| juv_fel_count | 0.082138 |
| juv_misd_count | 0.117976 |
| juv_other_count | 0.125797 |
| priors_count | 0.283640 |
| c_charge_degree | -0.037764 |
| days_b_screening_arrest | 0.032485 |
| length_of_stay | 0.012530 |

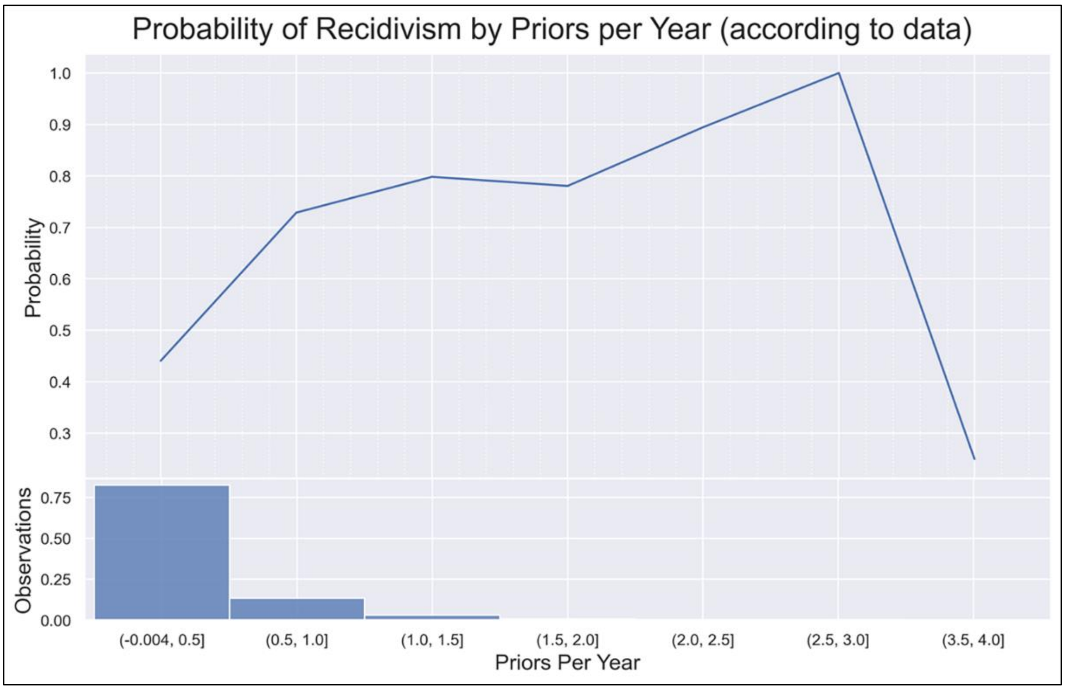




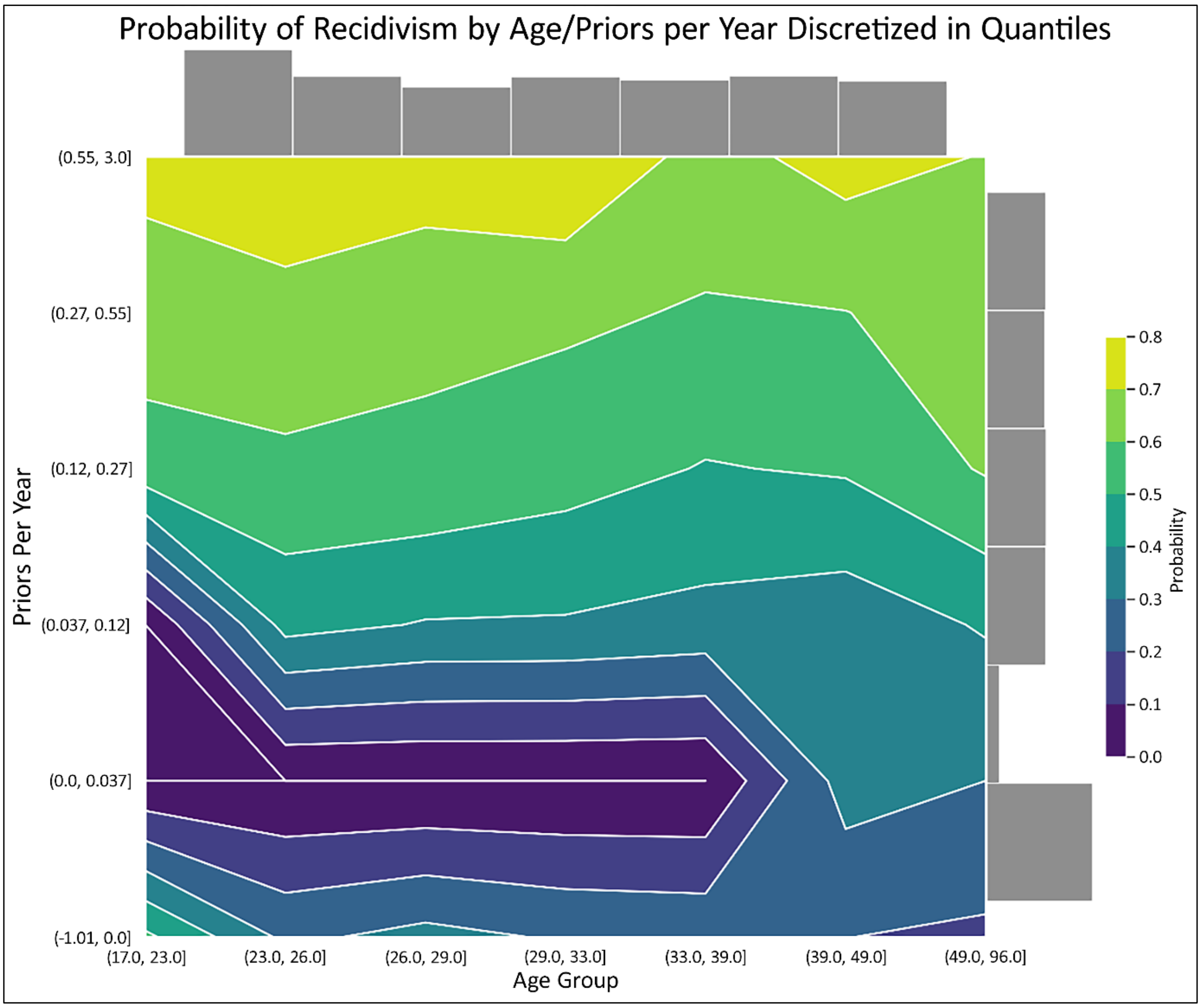
Probability of Recidivism by Age Discretized in Fix-Width Bins
Probability
Observations
(17.92.29.14)
(20.14.40.29)
(40.20.61.43)
(51.43.62.67)
(02.57.73.71)
(73.71.84.86)
(84.86.98.0)
Age
Probability of Recidivism by Age Discretized in Quantiles
Probability
Observations
(17.99.23.0)
(23.0.26.0)
(20.0.20.0)
(29.0.33.0)
(33.0.39.0)
(39.0.49.0)
(49.0.96.0)
Age

Probability of Recidivism by Age/Priors Discretized in Quantiles
Priors Count
(8.0, 38.0]
(4.0, 8.0]
(2.0, 4.0]
(1.0, 2.0]
(-0.01, 1.0]
(17.99, 23.0]
(23.0, 26.0]
(26.0, 29.0]
(29.0, 33.0]
(33.0, 39.0]
(39.0, 49.0]
(49.0, 96.0]
Age
Probability
0.88
0.80
0.72
0.64
0.56
0.48
0.40
0.32
0.24
0.16



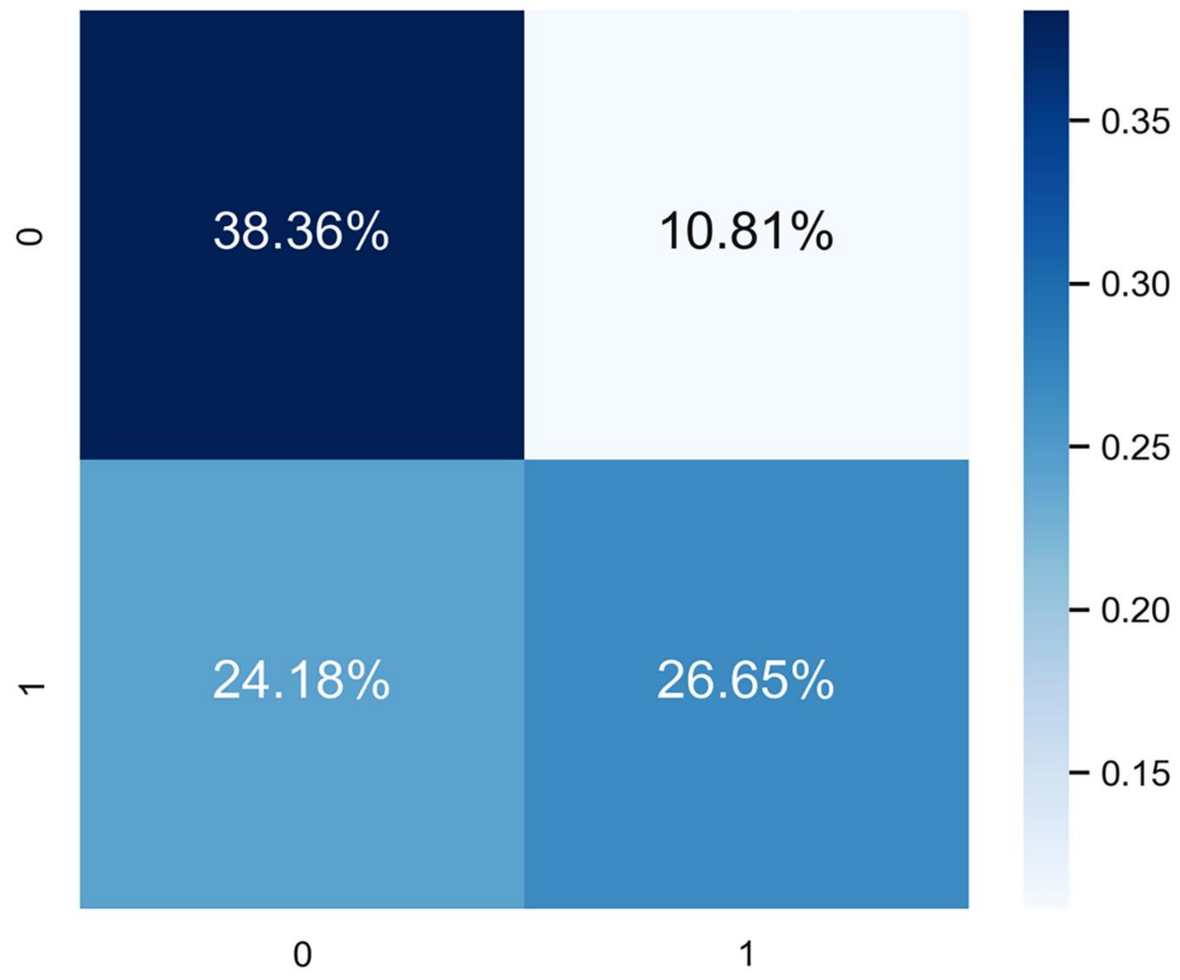
Probability of Recidivism by Priors per Year (according to data)
Probability
Observations
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.75
0.50
0.25
0.00
(-0.004, 0.5]
(0.5, 1.0]
(1.0, 1.5]
(1.5, 2.0]
(2.0, 2.5]
(2.5, 3.0]
(3.5, 4.0]
Priors Per Year

Probability of Recidivism by Age/Priors per Year Discretized in Quantiles
Priors Per Year
(0.55, 3.0]
(0.27, 0.55]
(0.12, 0.27]
(0.037, 0.12]
(0.0, 0.037]
(-1.01, 0.0]
(17.0, 23.0]
(23.0, 26.0]
(26.0, 29.0]
(29.0, 33.0]
(33.0, 39.0]
(39.0, 49.0]
(49.0, 96.0]
Age Group
Probability
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

| feature | correlation_to_target |
|---|---|
| sex | 0.093255 |
| race | -0.004598 |
| juv_fel_count | 0.082138 |
| juv_misd_count | 0.117976 |
| juv_other_count | 0.125797 |
| c_charge_degree | 0.069803 |
| days_b_screening_arrest | 0.032485 |
| length_of_stay | 0.012530 |
| age_group | -0.152131 |
| priors_per_year | 0.321885 |

| param_hidden_layer_sizes | param_l1_reg | param_l2_reg | param_dropout | mean_test_score | std_test_score | rank_test_score |
|---|---|---|---|---|---|---|
| (80,) | 0.005000 | 0.010000 | 0.050000 | 0.677700 | 0.018629 | 1 |
| (80,) | 0 | 0 | 0 | 0.670297 | 0.027577 | 2 |
| (80,) | 0.005000 | 0 | 0.050000 | 0.667625 | 0.021315 | 3 |
| (80,) | 0 | 0.010000 | 0.050000 | 0.667291 | 0.022757 | 4 |
| (80,) | 0.005000 | 0.010000 | 0 | 0.665553 | 0.017141 | 5 |
| (80,) | 0 | 0 | 0.050000 | 0.663555 | 0.011802 | 6 |
| (80,) | 0.005000 | 0 | 0 | 0.659114 | 0.026934 | 7 |
| (80,) | 0 | 0.010000 | 0 | 0.649437 | 0.019827 | 8 |



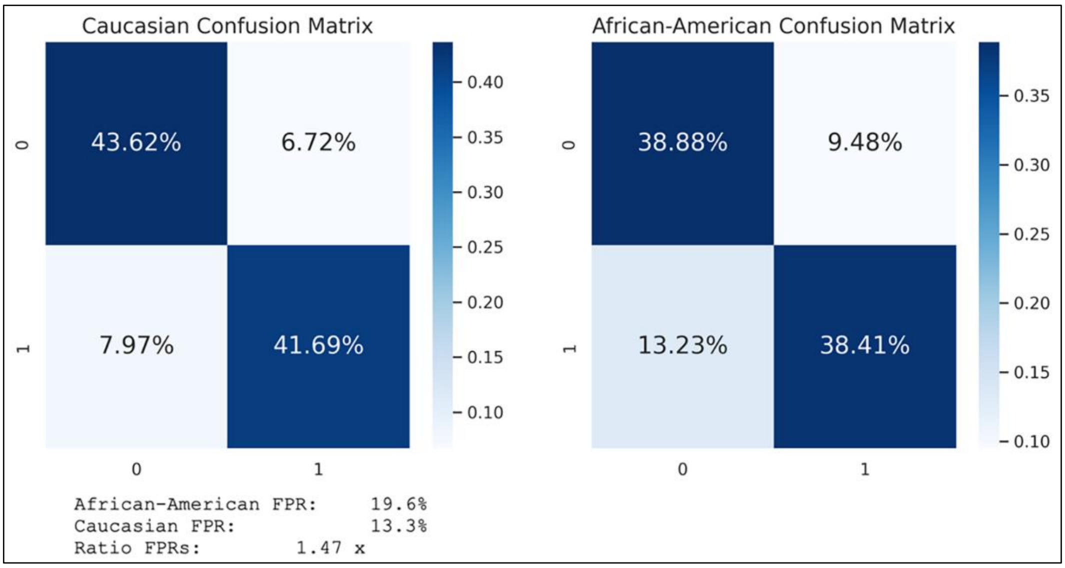

```
Accuracy_train:   0.6442        Accuracy_test:   0.6487
Precision_test:   0.7093        Recall_test:     0.5234
ROC-AUC_test:     0.6929        F1_test:         0.6023         MCC_test: 0.3115
```

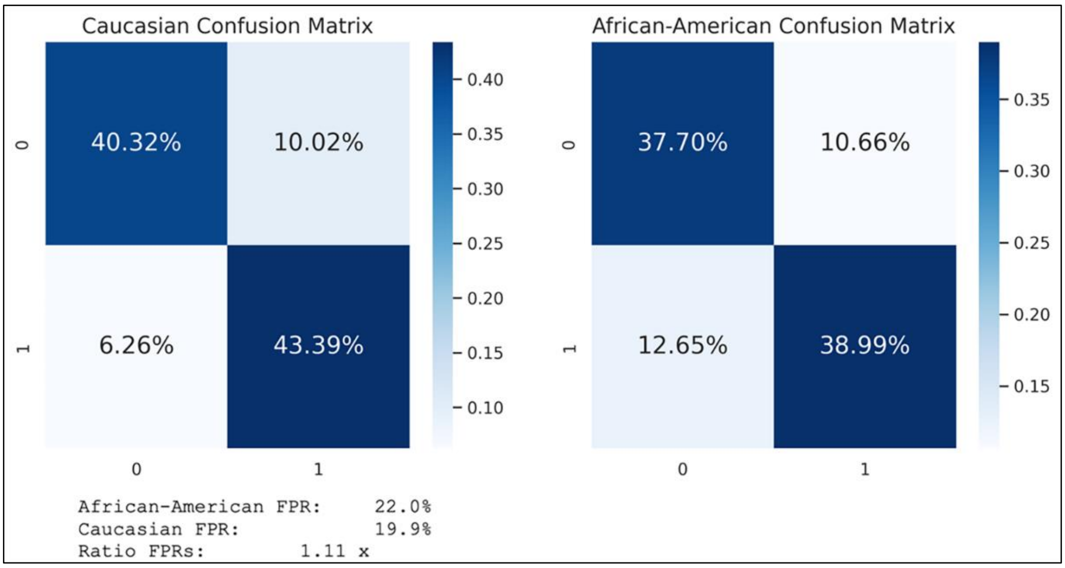
| | | **LogisticRegression** | | | **RidgeClassifier** | | | **SVC** | | | **NuSVC** | | | **MLPClassifier** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | **Ridge** | | | **SVR** | | | **NuSVR** | | | **MLPRegressor** | | |
| | **algorithm** | solver | | "lbfgs" | solver | | "auto" | kernel | | "rbf" | kernel | | "rbf" | solver | | "adam" |
| | **regularization** | penalty<br>C<br>l1_ratio | <br>+/-<br> | "l2"<br>1<br>None | alpha | +/- | 1 | C<br>gamma | +/-<br> | 1<br>"scale" | nu<br>gamma | +/-<br> | 0.5<br>"scale" | alpha | + | 0.0001 |
| OVERFITTING | **iterations** | max_iter | +/- | 100 | max_iter | + | None | max_iter | + | -1 | max_iter | + | -1 | max_iter | +/- | 200 |
| OVERFITTING | **learning rate** | | | | | | | | | | | | | learning_rate_init<br>learning_rate | | 0.001<br>"adaptive" |
| OVERFITTING | **early stopping** | tol | - | 1e-4 | tol | - | 1e-3 | tol | - | 1e-3 | tol | - | 1e-3 | tol<br>n_iter_no_change<br>early_stopping<br>validation_fraction | -<br>-<br><br> | 1e-4<br>10<br>False<br>0.1 |
| | **class imbalance** | class_weight | | None | class_weight | | None | class_weight | | None | class_weight | | None | | | |
| | **sample weight** | sample_weight* | | None | sample_weight* | | None | sample_weight* | | None | sample_weight* | | None | | | |

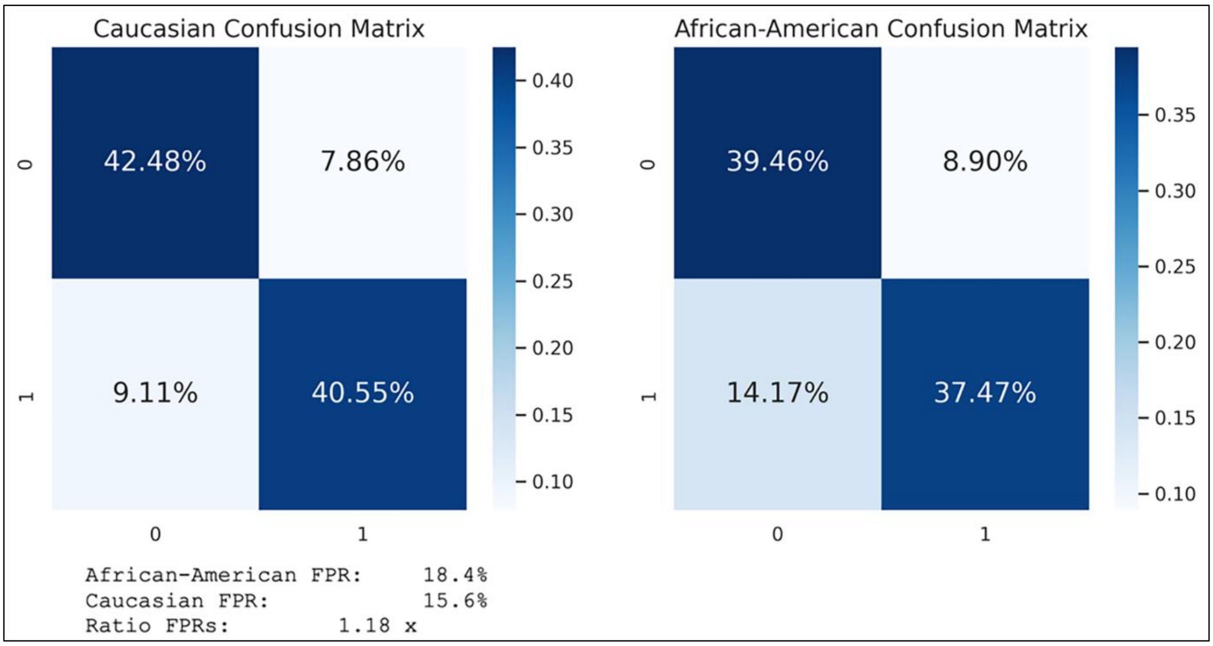
| | | LGBMClassifier<br>LGBMRegressor | | | CatboostClassifier<br>CatboostRegressor | | |
|---|---|---|---|---|---|---|---|
| | **algorithm** | boosting | | "gbdt" | | | |
| | **regularization** | lambda_l2 | + | 0 | l2_leaf_reg | + | 3 |
| | | lambda_l1 | + | 0 | | | |
| OVERFITTING | **feature sampling** | feature_fraction | - | 1 | | | |
| OVERFITTING | **learning rate** | learning_rate | +/- | 0.1 | learning_rate | +/- | 0.03 |
| OVERFITTING | **iterations / # trees** | num_iterations | + | 100 | iterations | + | 1000 |
| OVERFITTING | **early stopping** | early_stopping_rounds* | | 0 | early_stopping_rounds* | | False |
| | | eval_set* | | None | eval_set* | | None |
| | | eval_metric* | | None | eval_metric* | | None |
| OVERFITTING | **tree size** | max_depth | - | -1 | depth | - | 6 |
| | | num_leaves | - | 31 | max_leaves | - | 31 |
| | | min_data_in_leaf | + | 20 | min_data_in_leaf | + | 1 |
| | | min_sum_hessian_in_leaf | + | 1e-3 | | | |
| OVERFITTING | **splitting** | min_split_gain | + | 0 | grow_policy | | SymmetricTree |
| | | | | | random_strength | + | 1 |
| OVERFITTING | **bagging** | bagging_fraction | - | 1 | subsample | + | 0.66-1 |
| | | bagging_freq | + | 0 | | | |
| | **class imbalance** (classifiers only) | class_weight | | None | class_weights | | None |
| | | scale_pos_weight | +/- | 1 | scale_pos_weight | +/- | 1 |
| | | is_unbalance | + | False | auto_class_weights | + | False |
| | **sample weight** | sample_weight* | | None | sample_weight* | | None |
| | **constraints** | monotone_constraints | + | None | monotone_constraints | + | None |
| | | interaction_constraints | + | None | | | |

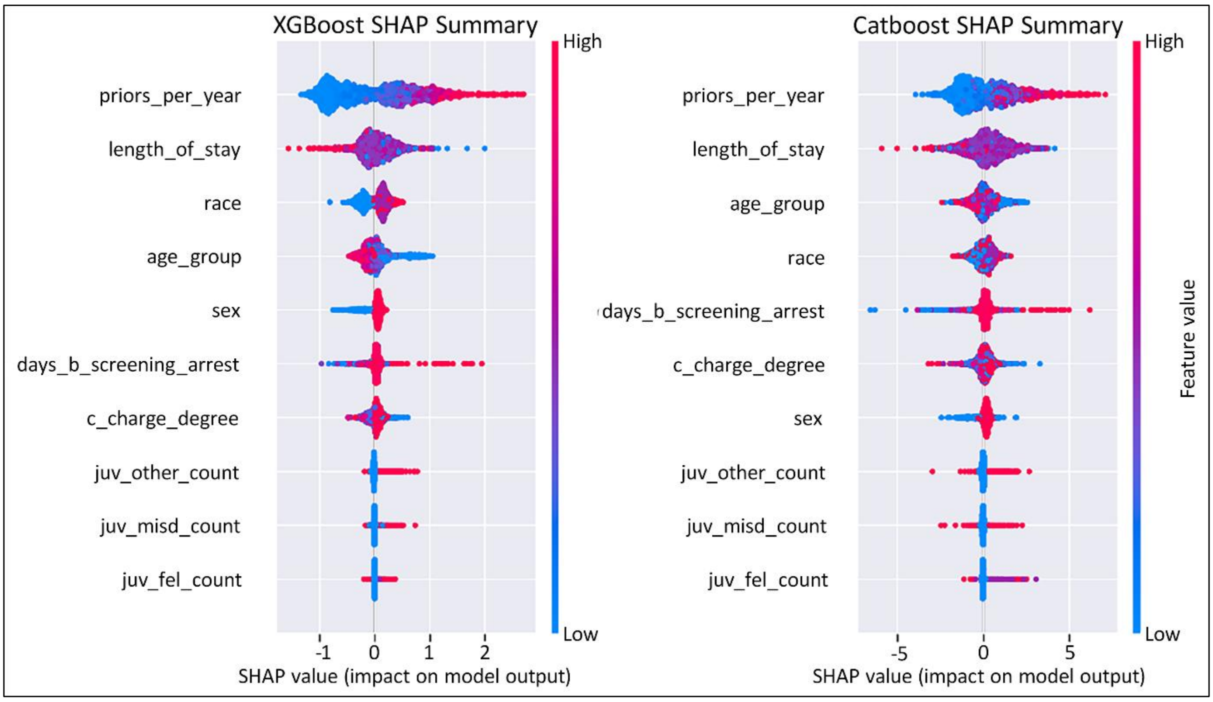
| | | RandomForestClassifier | | | XGBRFClassifier | | | XGBClassifier | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **RandomForestRegressor** | | | **XGBRFRegressor** | | | **XGBRegressor** | | |
| | **algorithm** | | | | booster | | "gbtree" | booster | | "gbtree" |
| | **regularization** | | | | reg_lambda | + | 1 | reg_lambda | + | 1 |
| | | | | | reg_alpha | + | 0 | reg_alpha | + | 0 |
| OVERFITTING | **feature sampling** | max_features | +/- | "auto" | | | | colsample_bytree | - | 1 |
| OVERFITTING | | | | | | | | colsample_bylevel | - | 1 |
| OVERFITTING | | | | | | | | colsample_bynode | - | 1 |
| OVERFITTING | **learning rate** | | | | eta | +/- | 1 | eta | +/- | 0.3 |
| OVERFITTING | **iterations / # trees** | n_estimators | +/- | 100 | n_estimators | +/- | 100 | num_round | +/- | 100 |
| OVERFITTING | **early stopping** | oob_score | + | False | early_stopping_rounds* | | None | early_stopping_rounds* | | None |
| OVERFITTING | | | | | eval_set* | | None | eval_set* | | None |
| OVERFITTING | | | | | eval_metric* | | None | eval_metric* | | None |
| OVERFITTING | **tree size** | max_depth | - | None | max_depth | - | 6 | max_depth | - | 6 |
| OVERFITTING | | max_leaf_nodes | - | None | | | | max_leaves | - | 0 |
| OVERFITTING | | min_samples_leaf | + | 1 | | | | | | |
| OVERFITTING | | min_weight_fraction_leaf | + | 0 | min_child_weight | + | 1 | min_child_weight | + | 1 |
| OVERFITTING | **splitting** | min_samples_split | + | 2 | gamma | + | 0 | gamma | + | 0 |
| OVERFITTING | | min_impurity_decrease | + | 0 | | | | | | |
| OVERFITTING | | criterion | | "gini" | | | | | | |
| OVERFITTING | **bagging** | max_samples | | None | subsample | + | 1 | subsample | + | 1 |
| OVERFITTING | | bootstrap | | True | sampling_method | | "uniform" | sampling_method | | "uniform" |
| | **class imbalance** (classifiers only) | class_weight | | None | scale_pos_weight | +/- | 1 | scale_pos_weight | +/- | 1 |
| | **sample weight** | sample_weight* | | None | sample_weight* | | None | sample_weight* | | None |
| | **constraints** | | | | monotone_constraints | + | None | monotone_constraints | + | None |
| | | | | | interaction_constraints | + | None | interaction_constraints | + | None |

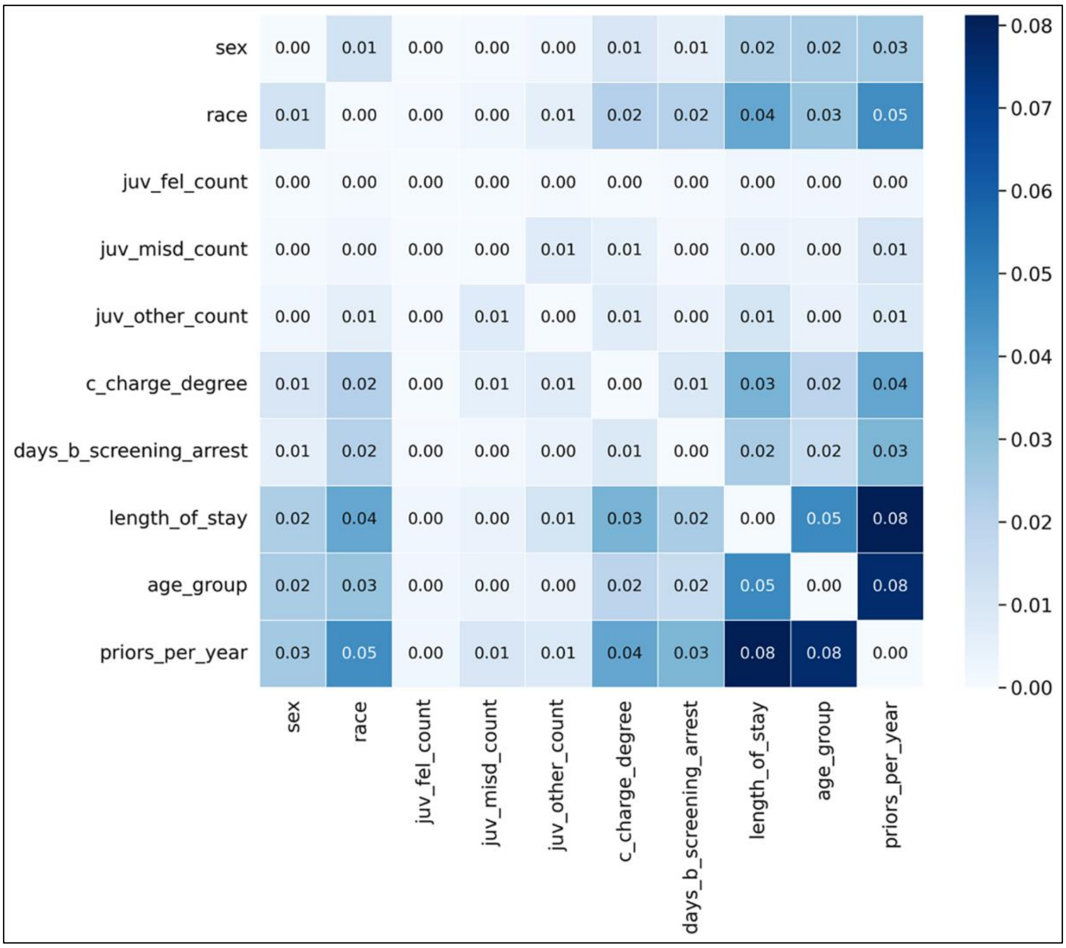
| | accuracy_train | accuracy_test | precision_train | precision_test | recall_train | recall_test | roc-auc_test | f1_test | mcc_test |
|---|---|---|---|---|---|---|---|---|---|
| **catboost_reg** | 0.968 | 0.826 | 0.991 | 0.836 | 0.943 | 0.818 | 0.885 | 0.827 | 0.652 |
| **nu-svc_reg** | 0.939 | 0.807 | 0.950 | 0.836 | 0.925 | 0.772 | 0.858 | 0.803 | 0.616 |
| **catboost_base** | 0.969 | 0.814 | 0.978 | 0.805 | 0.959 | 0.837 | 0.878 | 0.821 | 0.629 |
| **lgbm_reg** | 0.863 | 0.766 | 0.904 | 0.800 | 0.807 | 0.718 | 0.826 | 0.757 | 0.535 |
| **xgb_reg** | 0.807 | 0.727 | 0.885 | 0.800 | 0.697 | 0.618 | 0.811 | 0.697 | 0.469 |
| **lgbm_base** | 0.855 | 0.752 | 0.865 | 0.758 | 0.836 | 0.753 | 0.822 | 0.755 | 0.504 |
| **xgb_base** | 0.801 | 0.739 | 0.802 | 0.749 | 0.789 | 0.733 | 0.811 | 0.741 | 0.479 |
| **logistic_reg** | 0.643 | 0.638 | 0.721 | 0.745 | 0.445 | 0.437 | 0.701 | 0.551 | 0.309 |
| **svc_reg** | 0.623 | 0.648 | 0.741 | 0.728 | 0.356 | 0.491 | 0.716 | 0.586 | 0.317 |
| **mlp_reg** | 0.649 | 0.653 | 0.690 | 0.724 | 0.518 | 0.514 | 0.706 | 0.601 | 0.324 |
| **rf_reg** | 0.722 | 0.686 | 0.741 | 0.716 | 0.668 | 0.635 | 0.759 | 0.673 | 0.376 |
| **logistic_base** | 0.651 | 0.654 | 0.685 | 0.714 | 0.535 | 0.533 | 0.701 | 0.610 | 0.322 |
| : | : | : | : | : | : | : | : | : | : |
| **nu-svc_base** | 0.531 | 0.509 | 0.560 | 0.580 | 0.204 | 0.122 | 0.579 | 0.201 | 0.049 |

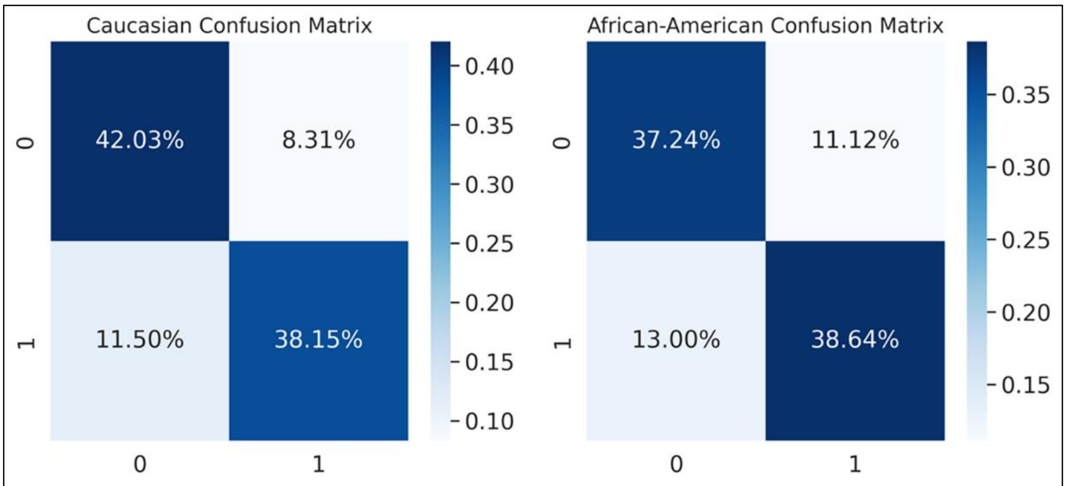

Caucasian Confusion Matrix
43.62%
6.72%
7.97%
41.69%
African-American Confusion Matrix
38.88%
9.48%
13.23%
38.41%
African-American FPR: 19.6%
Caucasian FPR: 13.3%
Ratio FPRs: 1.47 x


Caucasian Confusion Matrix
40.32%
10.02%
6.26%
43.39%
African-American Confusion Matrix
37.70%
10.66%
12.65%
38.99%
African-American FPR: 22.0%
Caucasian FPR: 19.9%
Ratio FPRs: 1.11 x


Caucasian Confusion Matrix
42.48%
7.86%
9.11%
40.55%
African-American Confusion Matrix
39.46%
8.90%
14.17%
37.47%
African-American FPR: 18.4%
Caucasian FPR: 15.6%
Ratio FPRs: 1.18 x

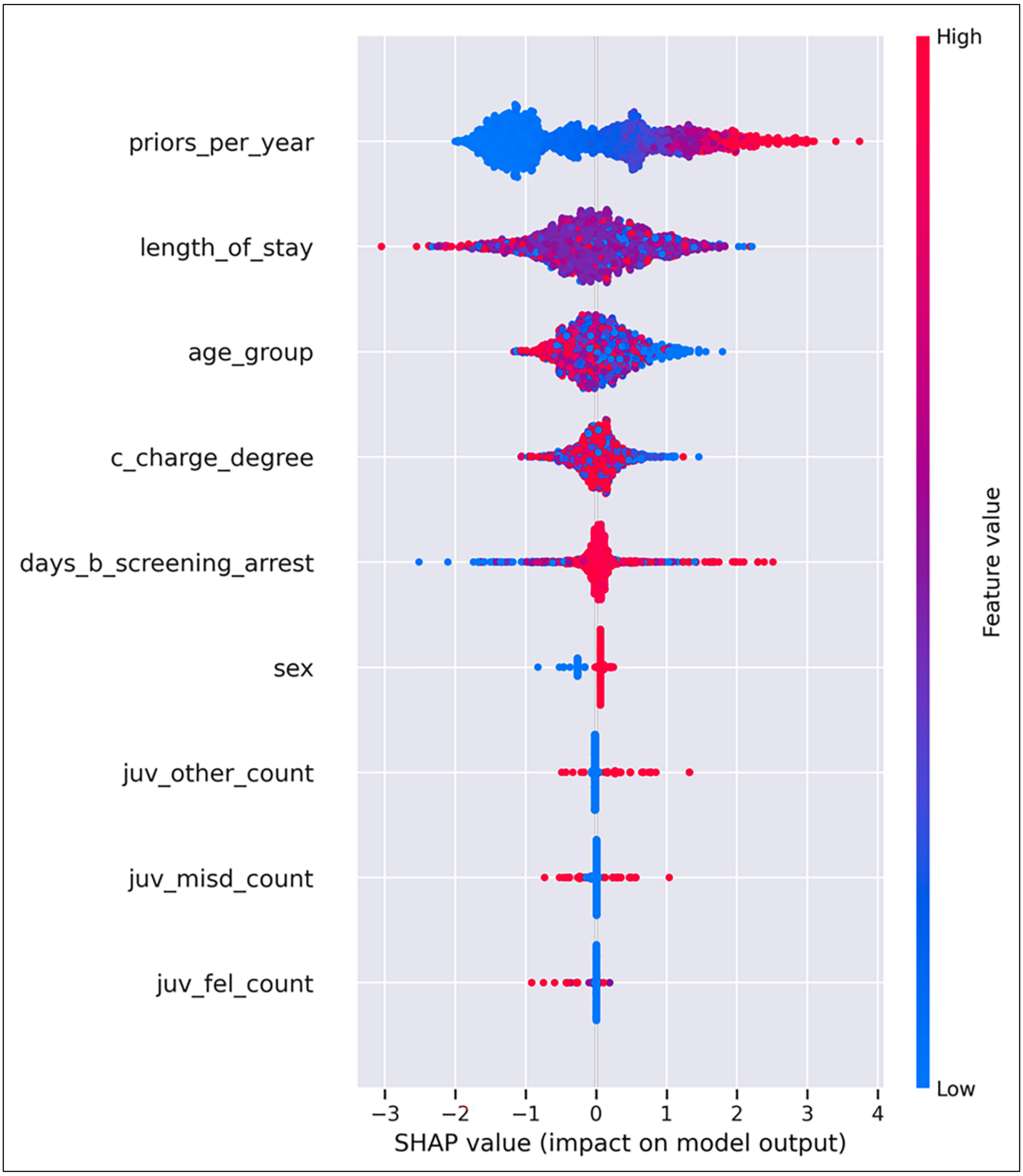

XGBoost SHAP Summary
priors_per_year
length_of_stay
race
age_group
sex
days_b_screening_arrest
c_charge_degree
juv_other_count
juv_misd_count
juv_fel_count
SHAP value (impact on model output)
Catboost SHAP Summary
priors_per_year
length_of_stay
age_group
race
days_b_screening_arrest
c_charge_degree
sex
juv_other_count
juv_misd_count
juv_fel_count
SHAP value (impact on model output)
High
Low
Feature value



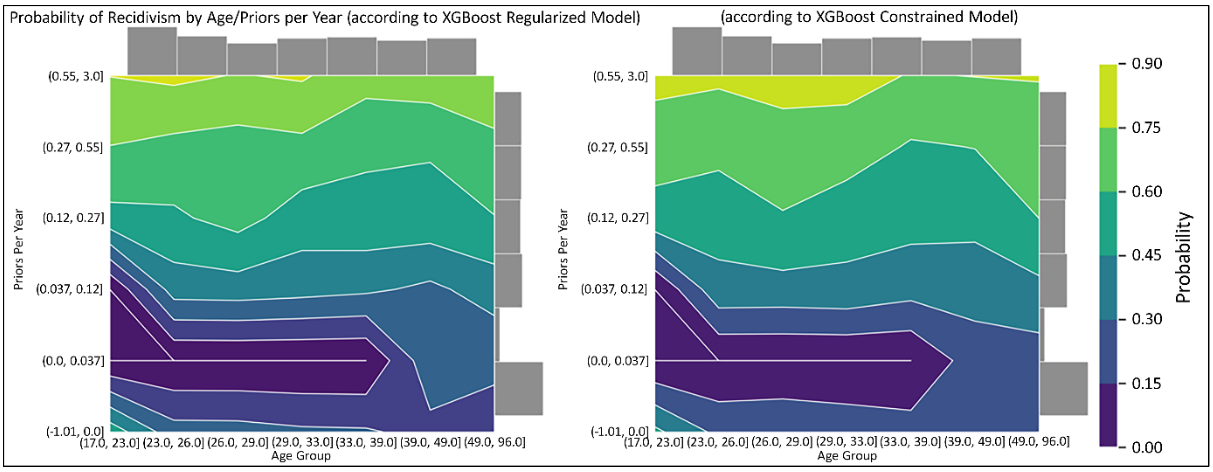



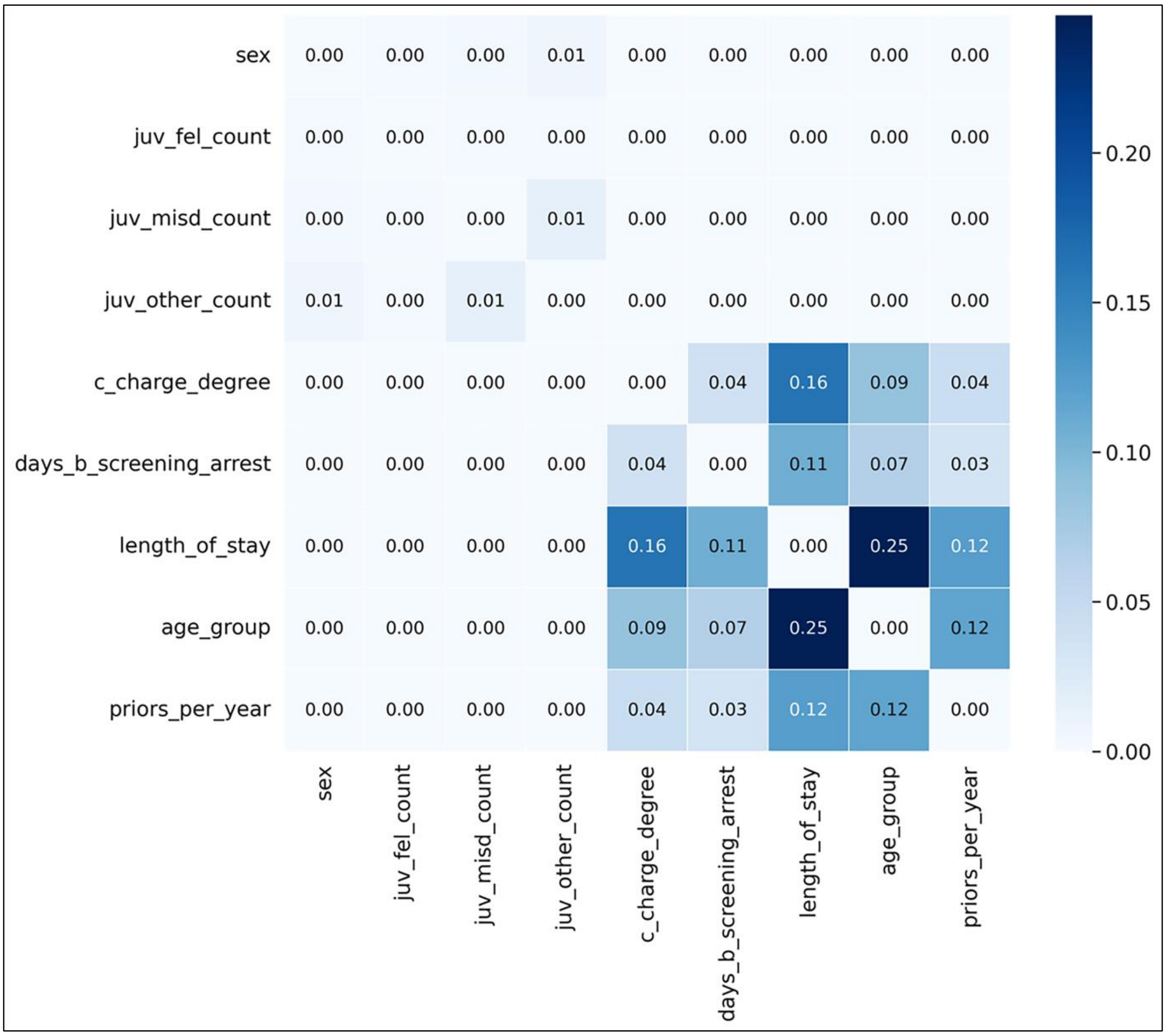

```
African-American FPR:      23.0%
Caucasian FPR:             16.5%
Ratio FPRs:         1.39 x
```

v

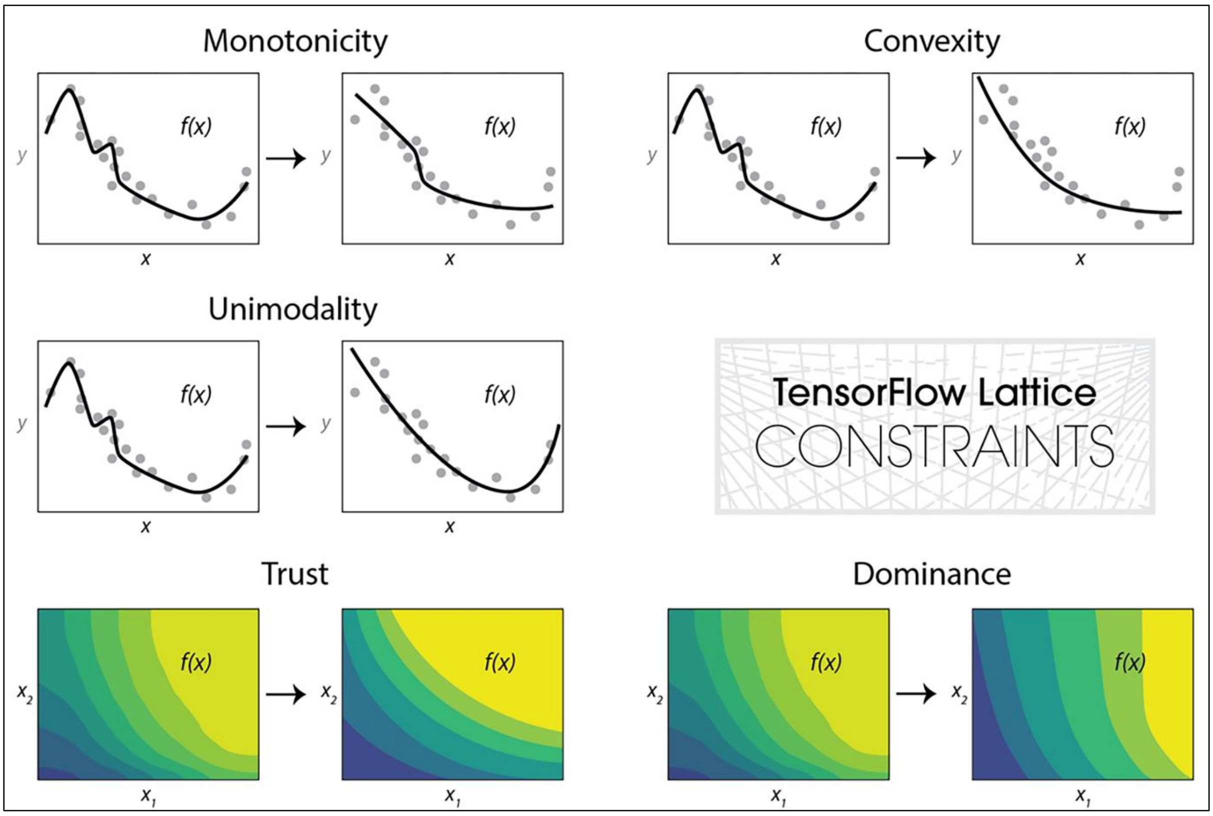


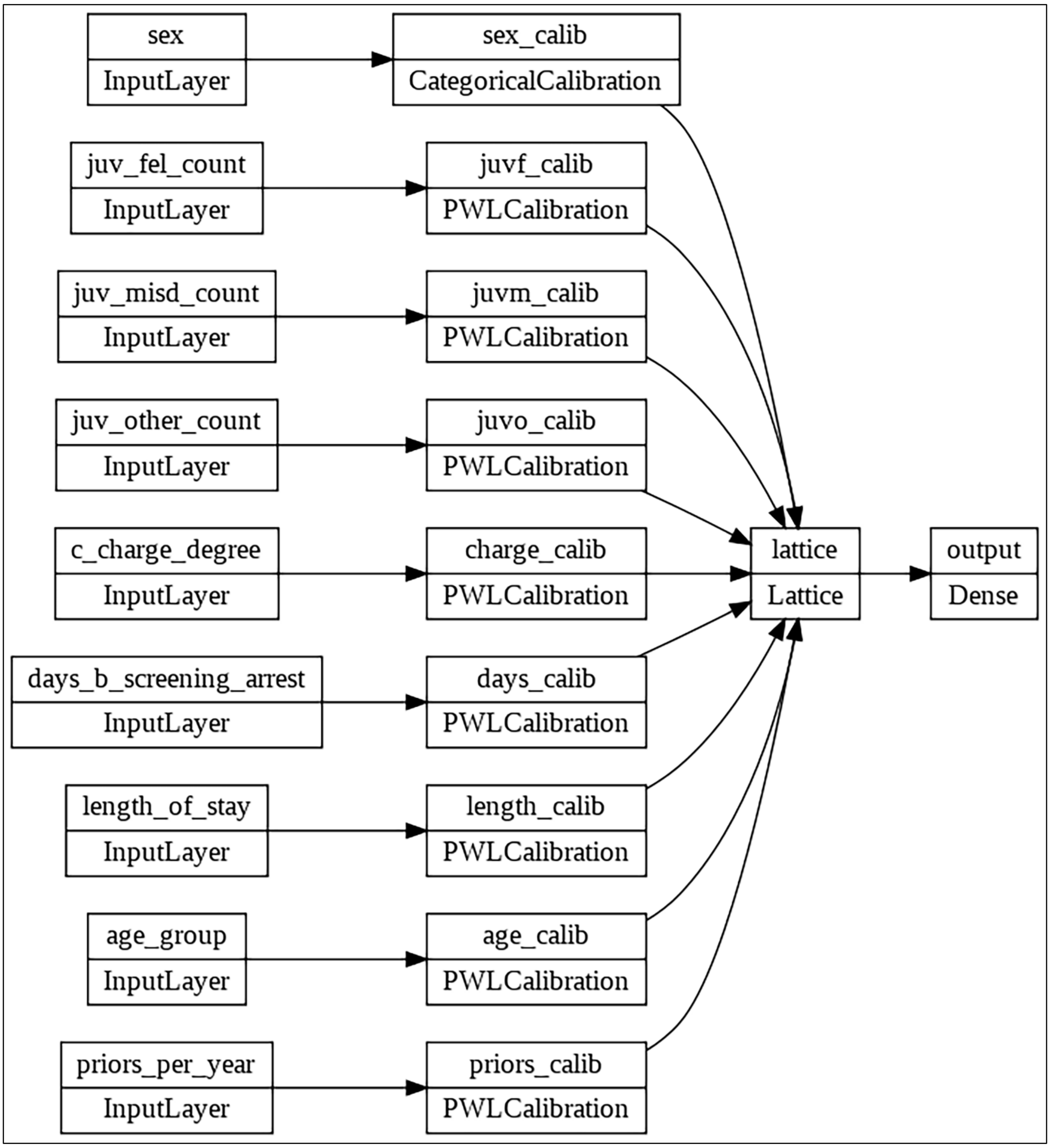
priors_per_year
length_of_stay
age_group
c_charge_degree
days_b_screening_arrest
sex
juv_other_count
juv_misd_count
juv_fel_count
−3
−2
−1
0
1
2
3
4
SHAP value (impact on model output)
High
Low
Feature value

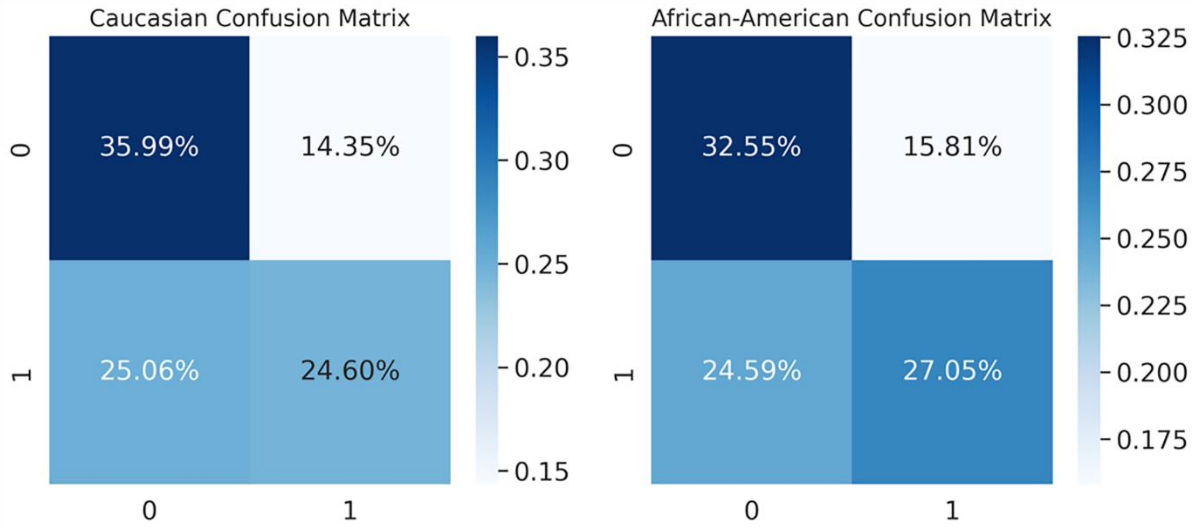



| | sex | juv_fel_count | juv_misd_count | juv_other_count | c_charge_degree | days_b_screening_arrest | length_of_stay | age_group | priors_per_year |
|---|---|---|---|---|---|---|---|---|---|
| sex | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| juv_fel_count | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| juv_misd_count | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| juv_other_count | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| c_charge_degree | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 | 0.16 | 0.09 | 0.04 |
| days_b_screening_arrest | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 | 0.00 | 0.11 | 0.07 | 0.03 |
| length_of_stay | 0.00 | 0.00 | 0.00 | 0.00 | 0.16 | 0.11 | 0.00 | 0.25 | 0.12 |
| age_group | 0.00 | 0.00 | 0.00 | 0.00 | 0.09 | 0.07 | 0.25 | 0.00 | 0.12 |
| priors_per_year | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 | 0.03 | 0.12 | 0.12 | 0.00 |


Monotonicity
Convexity
Unimodality
TensorFlow Lattice
CONSTRAINTS
Trust
Dominance
f(x)
y
x
$x_1$
$x_2$


sex
InputLayer
sex_calib
CategoricalCalibration
juv_fel_count
InputLayer
juvf_calib
PWLCalibration
juv_misd_count
InputLayer
juvm_calib
PWLCalibration
juv_other_count
InputLayer
juvo_calib
PWLCalibration
c_charge_degree
InputLayer
charge_calib
PWLCalibration
lattice
Lattice
output
Dense
days_b_screening_arrest
InputLayer
days_calib
PWLCalibration
length_of_stay
InputLayer
length_calib
PWLCalibration
age_group
InputLayer
age_calib
PWLCalibration
priors_per_year
InputLayer
priors_calib
PWLCalibration




```
African-American FPR:     32.7%
Caucasian FPR:            28.5%
Ratio FPRs:          1.15 x
```

| | precision_test | recall_test | wppra_test |
|---|---|---|---|
| catboost_reg | 0.836 | 0.818 | 0.810 |
| catboost_opt | 0.833 | 0.806 | 0.803 |
| catboost_base | 0.805 | 0.837 | 0.799 |
| nu-svc_reg | 0.836 | 0.772 | 0.791 |
| xgb_con | 0.810 | 0.777 | 0.783 |
| lgbm_reg | 0.798 | 0.709 | 0.747 |
| lgbm_base | 0.766 | 0.748 | 0.743 |
| xgb_base | 0.749 | 0.733 | 0.725 |
| xgb_reg | 0.800 | 0.618 | 0.717 |
| tfl_con | 0.646 | 0.540 | 0.591 |
| nu-svc_base | 0.580 | 0.122 | 0.406 |

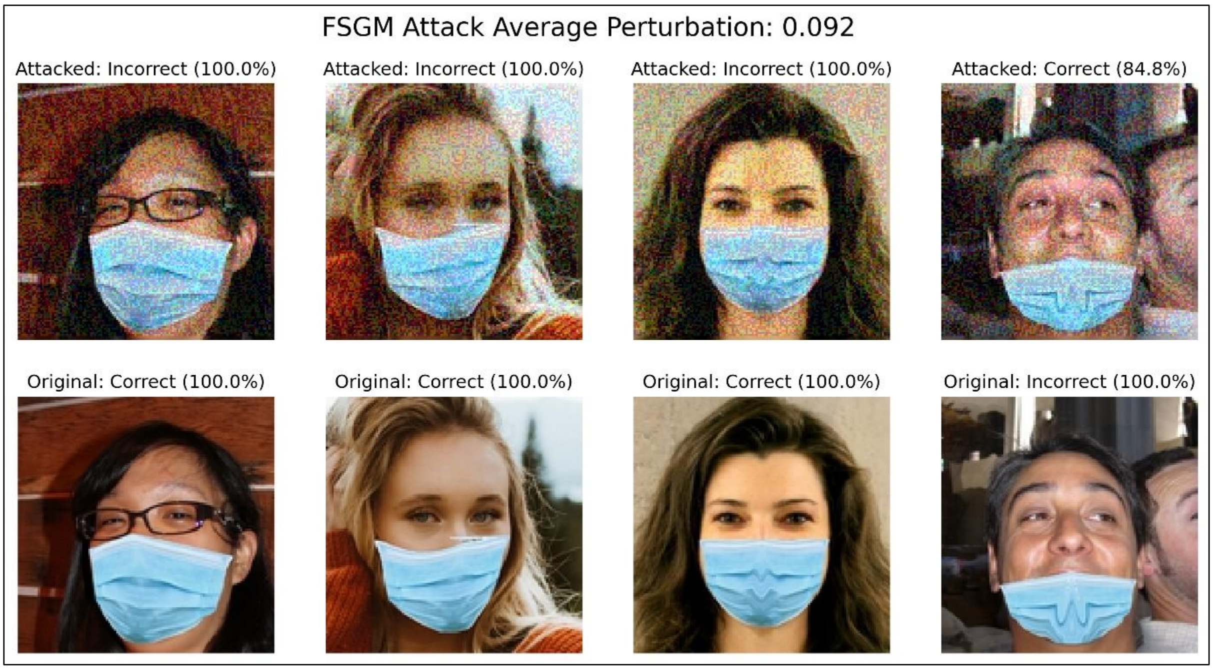
## Chapter 13: Adversarial Robustness


YOUR
SPEED
42

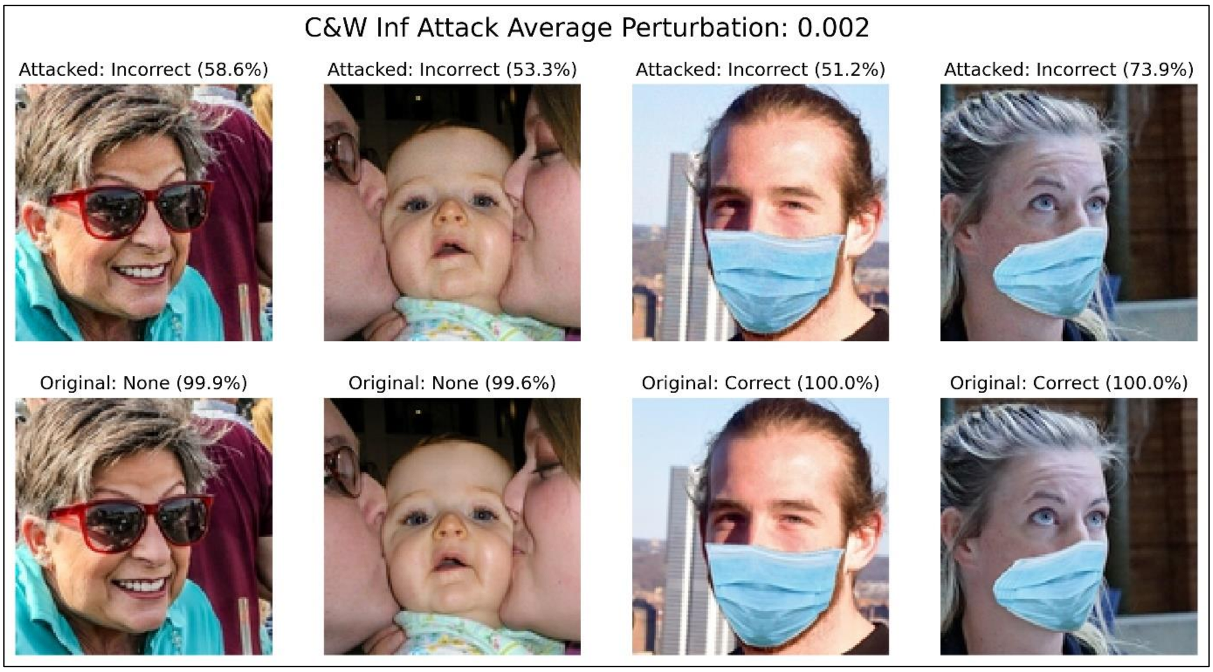

None
Correct
None
Incorrect
Correct
Correct
Correct
None
None
Correct
None
Correct
Incorrect
None
None
Incorrect
None
Incorrect
Correct
Incorrect

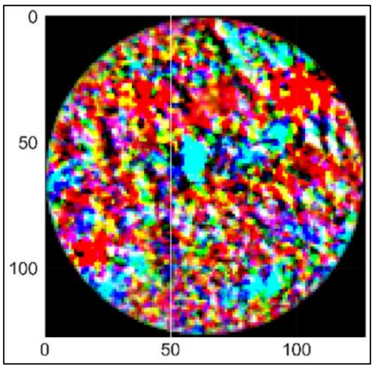

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Correct | 0.998 | 0.998 | 0.998 | 1400 |
| Incorrect | 0.999 | 0.995 | 0.997 | 1400 |
| None | 0.996 | 0.999 | 0.998 | 1400 |
| accuracy | | | 0.997 | 4200 |
| macro avg | 0.997 | 0.997 | 0.997 | 4200 |
| weighted avg | 0.997 | 0.997 | 0.997 | 4200 |

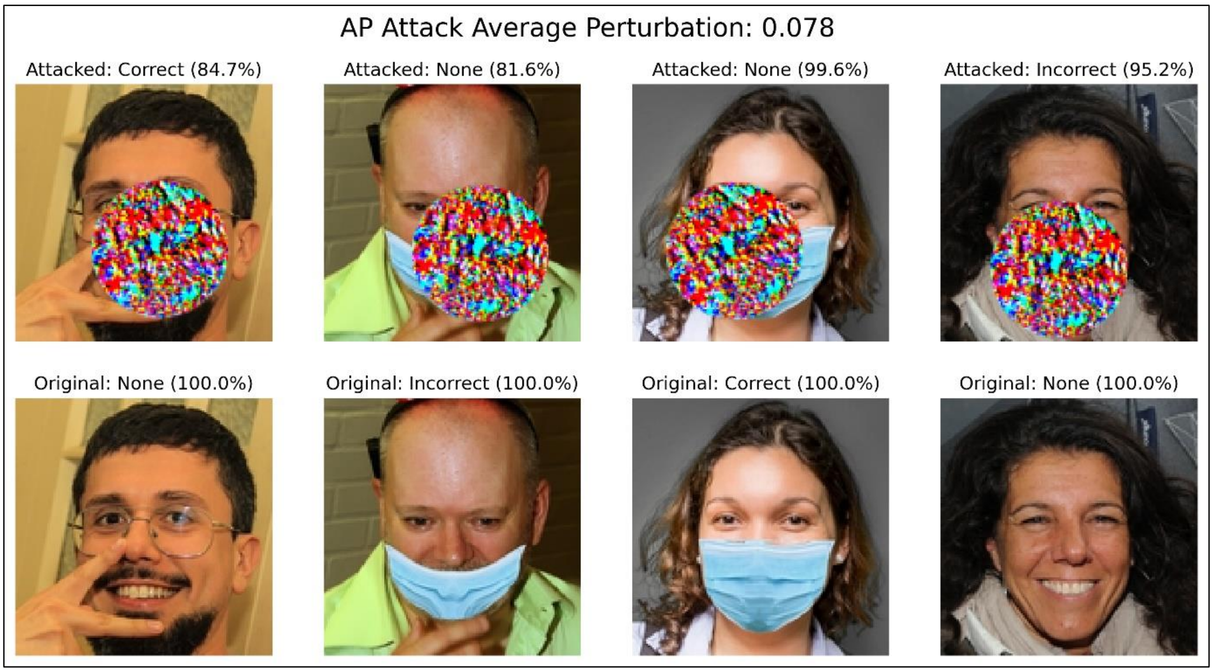
| | | Goal | | |
|---|---|---|---|---|
| | | **Espionage** | **Sabotage** | **Fraud** |
| **Stage** | **Training** | Inference (by poisoning) | Trojaning | |
| | | | Poisoning | |
| | | | Backdooring | |
| | **Production** | Inference | Reprogramming | |
| | | | Evasion | |

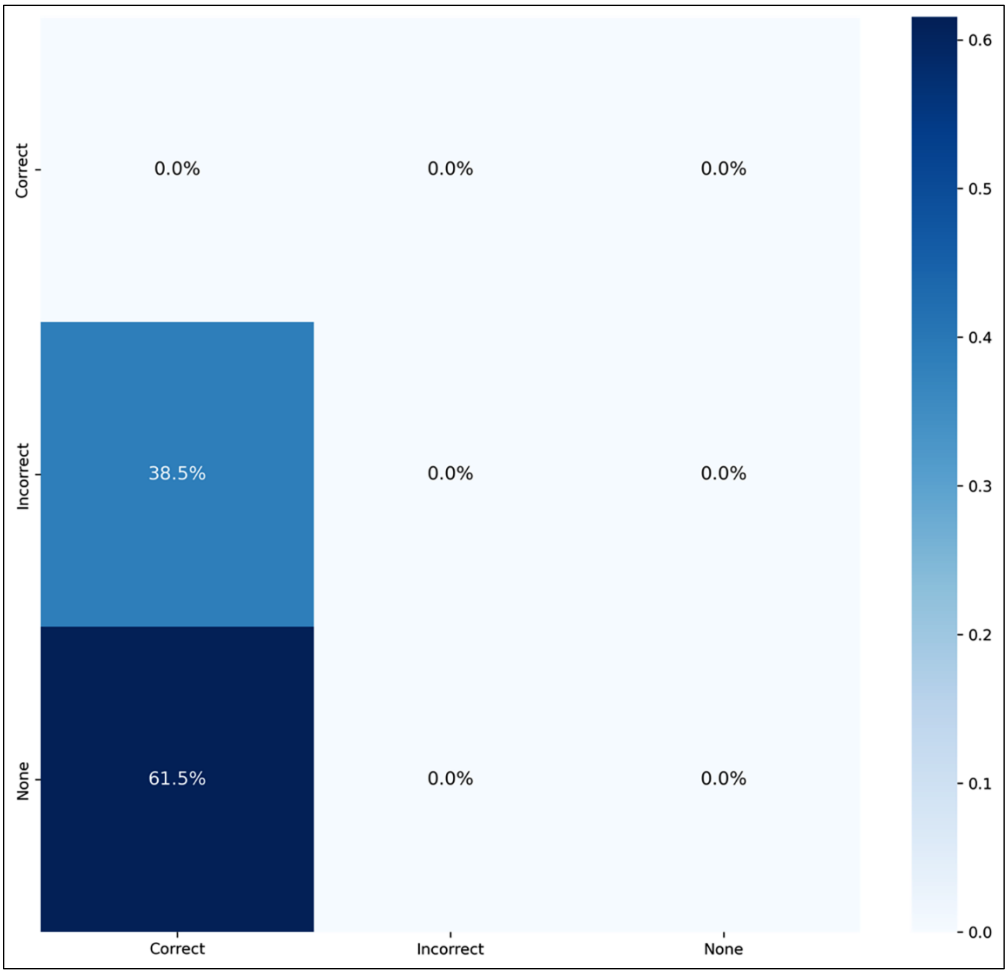


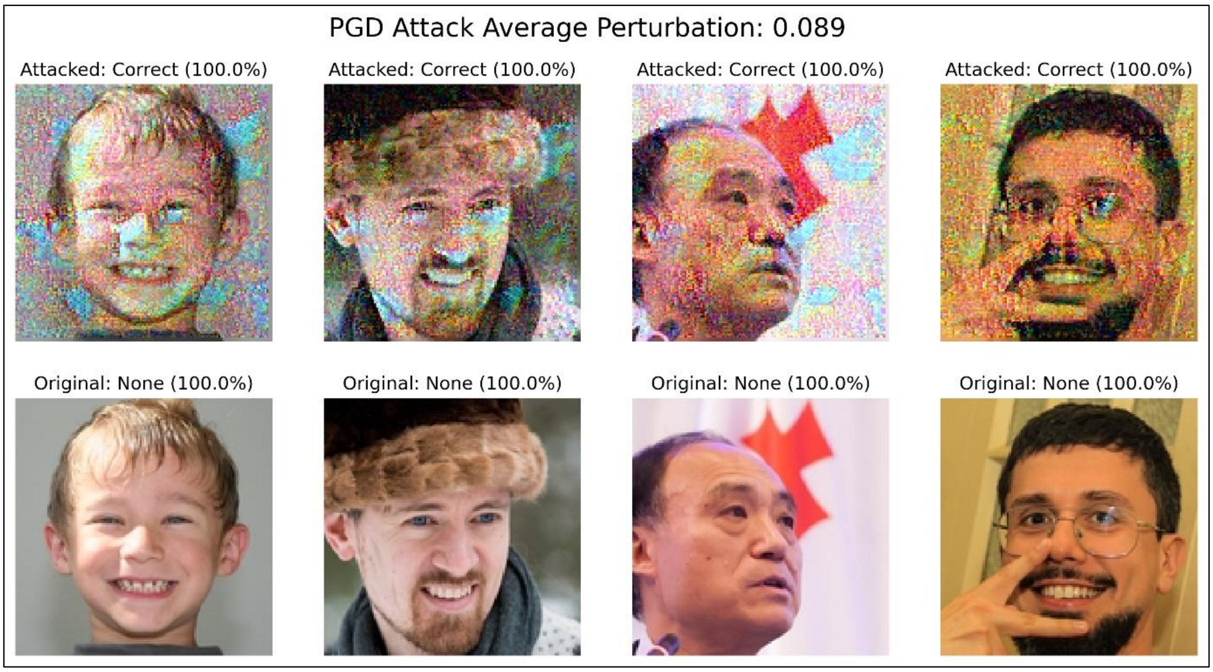
V

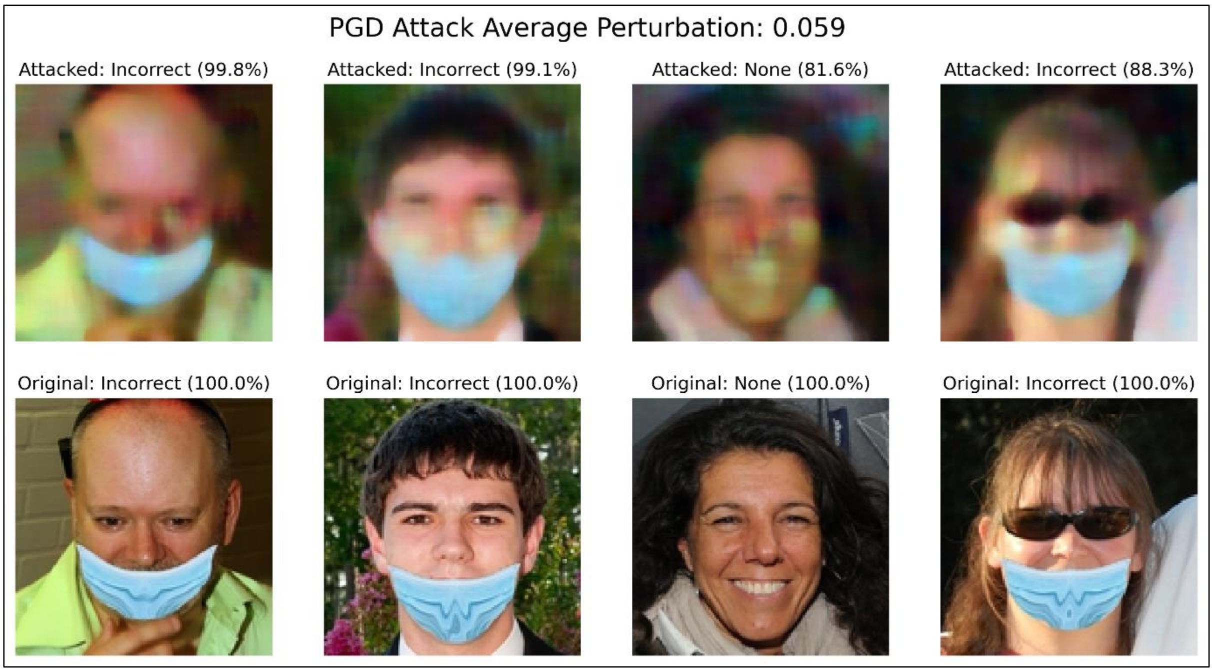

C&W Inf Attack Average Perturbation: 0.002
Attacked: Incorrect (58.6%)
Attacked: Incorrect (53.3%)
Attacked: Incorrect (51.2%)
Attacked: Incorrect (73.9%)
Original: None (99.9%)
Original: None (99.6%)
Original: Correct (100.0%)
Original: Correct (100.0%)

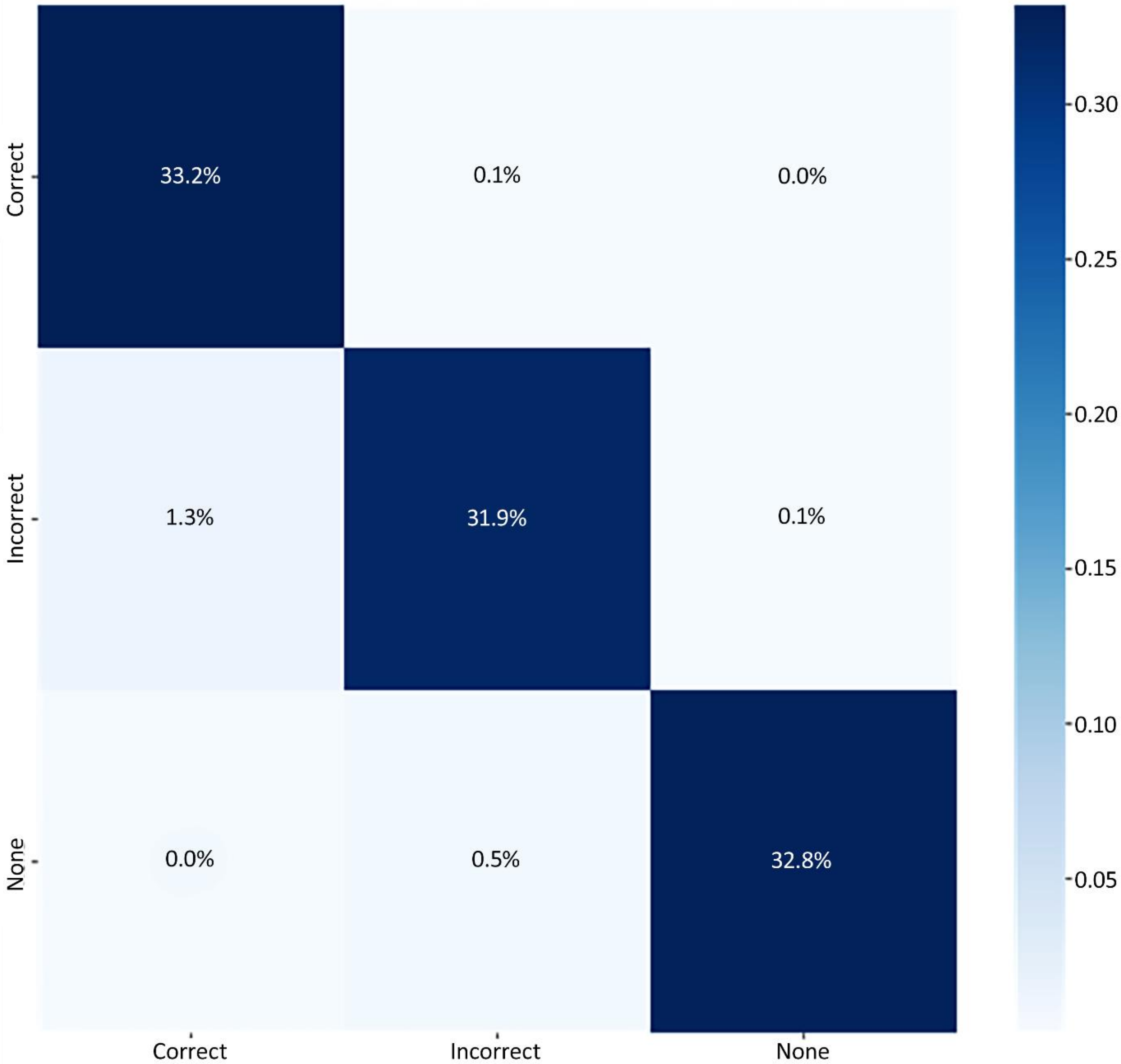

0
50
100
0
50
100


AP Attack Average Perturbation: 0.078
Attacked: Correct (84.7%)
Attacked: None (81.6%)
Attacked: None (99.6%)
Attacked: Incorrect (95.2%)
Original: None (100.0%)
Original: Incorrect (100.0%)
Original: Correct (100.0%)
Original: None (100.0%)


Correct
Incorrect
None
0.0%
0.0%
0.0%
38.5%
0.0%
0.0%
61.5%
0.0%
0.0%
Correct
Incorrect
None
0.6
0.5
0.4
0.3
0.2
0.1
0.0

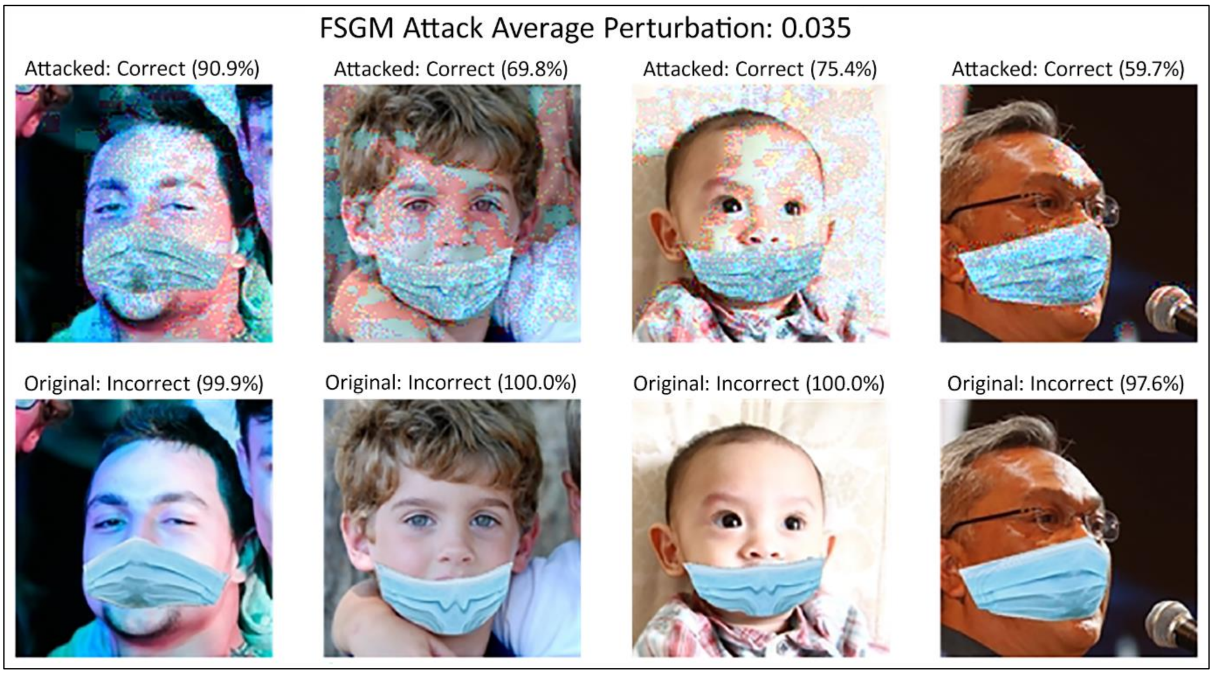

PGD Attack Average Perturbation: 0.089
Attacked: Correct (100.0%)
Attacked: Correct (100.0%)
Attacked: Correct (100.0%)
Attacked: Correct (100.0%)
Original: None (100.0%)
Original: None (100.0%)
Original: None (100.0%)
Original: None (100.0%)

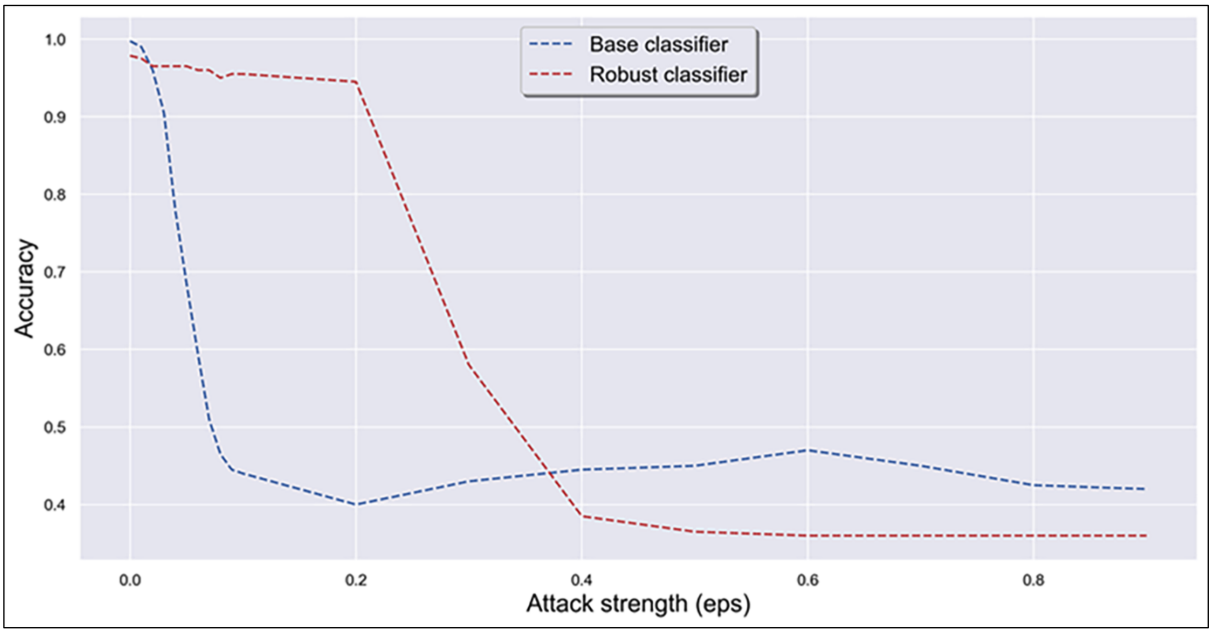

PGD Attack Average Perturbation: 0.059
Attacked: Incorrect (99.8%)
Attacked: Incorrect (99.1%)
Attacked: None (81.6%)
Attacked: Incorrect (88.3%)
Original: Incorrect (100.0%)
Original: Incorrect (100.0%)
Original: None (100.0%)
Original: Incorrect (100.0%)




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Correct | 0.962 | 0.996 | 0.979 | 1400 |
| Incorrect | 0.980 | 0.957 | 0.969 | 1400 |
| None | 0.994 | 0.983 | 0.989 | 1400 |
| accuracy | | | 0.979 | 4200 |
| macro avg | 0.979 | 0.979 | 0.979 | 4200 |
| weighted avg | 0.979 | 0.979 | 0.979 | 4200 |

v




# Chapter 14: What's Next for Machine Learning Interpretability?

## INTERPRETATION METHODS BY CHAPTER

**Evaluating models, confirming assumptions, finding problems, and certifying reliability**

### FAIRNESS

- Class Balance 3 4 6 10 11 12
- Comparing Fairness Metrics 6 11 12
  - Statistical Parity Difference
  - Disparate Impact
  - Differential Fairness Bias Amplification
  - :
- Error Analysis for Fairness 6 11 12
  - Comparing Metrics
    - FPR
    - AUC
    - :
  - Comparing Plots
    - Confusion Matrix
    - ROC Curve
    - Contour / Heatmap Plots
    - :

### ACCOUNTABILITY

- Out-of-sample Evaluations 7 8
- Sensitivity Analysis 9
- Causal Inference Methods 11
- Adversarial Robustness Evaluations 13
  - Evasion
  - *Inference, Extraction & Poisoning*

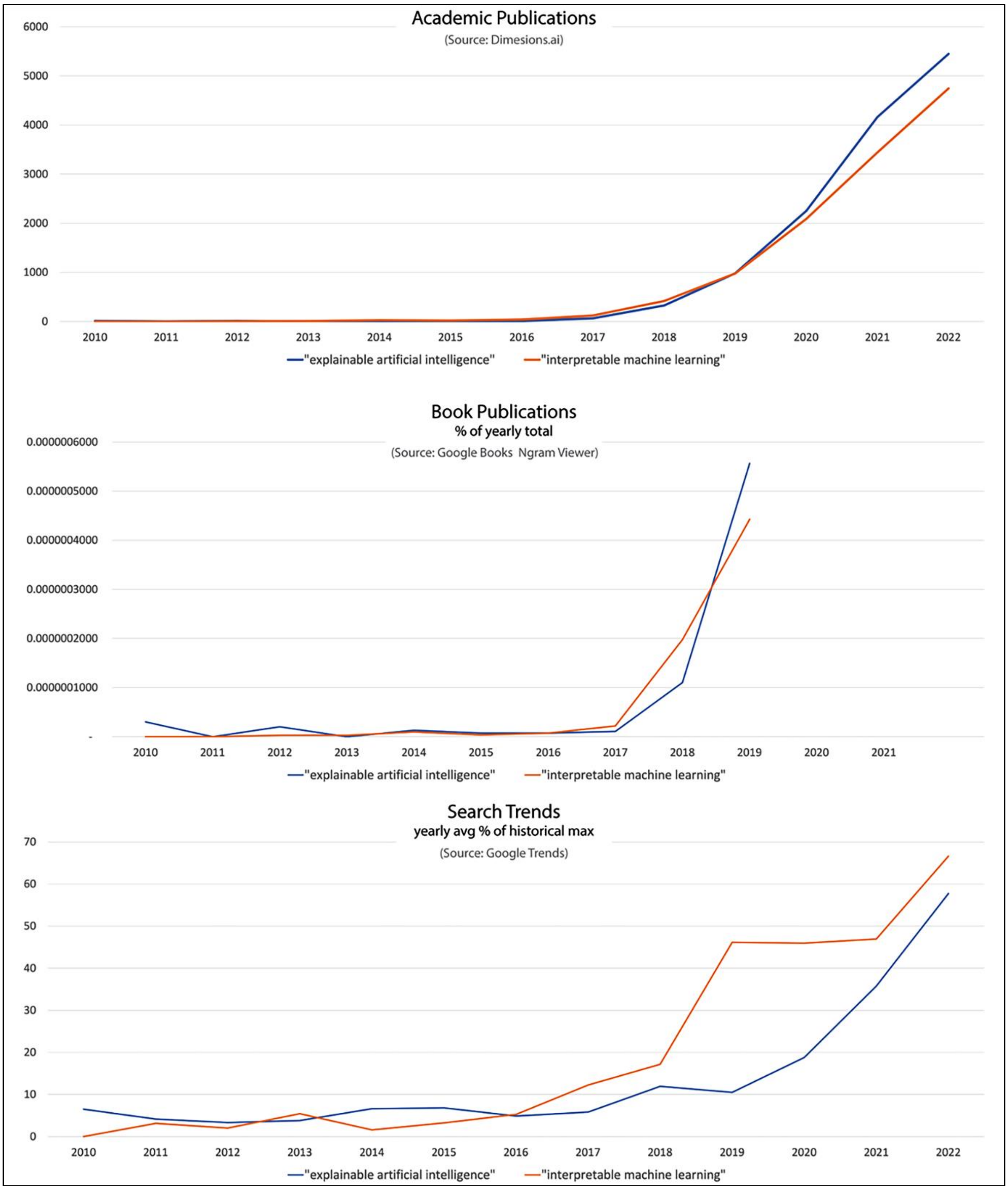
### TRANSPARENCY

- Global explanations
  - Feature Importance 1 2 3 4 5 7 8 9 10 12
    - Model-specifc (or via Global Surrogate)
    - Permutation
    - SHAP
    - :
  - Feature Summary & Interactions 4 6 9 11 12
    - Partial Dependence Plots
    - Accumulated Local Effects Plots
    - SHAP Dependence
    - :
- Local explanations
  - LIME & SHAP 5 7 9
  - Individual Conditional Expectation 4
  - Decision Boundaries/Regions 2
  - Anchors & Counterfactuals 6
  - Deep Learning Specific Explantions 7 8 9
    - Integrated Gradients
    - Saliency Maps
    - Grad-CAM
    - :

## INTERPRETABILITY SOLUTIONS BY CHAPTER

**Mitigating bias, placing guardrails, enhancing reliability, reducing complexity, and ensuring privacy**

| | DATA | MODEL | PREDICTION |
|---|---|---|---|
| **FAIRNESS** | Resampling / Reweighting 11<br>Feature Engineering 10 12<br>Data Augmentation 8 11 13<br>Feature Selection 10 (Filter, Embedded, Wrapper) | Cost-sensitive Learning 10 11 12<br>Monotonic Constraints 12<br>Adversarial Debiasing 11<br>Regularization 3 12 | Calibrating/Equalizing Odds 6 11<br>Prediction Abstention 13<br>*Fairness Model Certification* |
| **ACCOUNTABILITY** | *Feature Drift Detection*<br>Data Augmentation 8 11 13<br>Adv. Preproc. Defenses 13<br>Feature Selection 10 (Filter, Embedded, Wrapper)<br>Feature Engineering 10 12<br>*Data Anonymization*<br>*Differential Privacy* | *Uncertainty Estimation / Conformal Prediction*<br>Adversarial Robustness Certified Training & Inference 13<br>Adversarial Training 13<br>Regularization 3 12 (plus other under-fitting tuning)<br>Monotonic Constraints 12<br>*Federated Learning*<br>*Other Adversarial Defenses (for espionage attacks)* | Adv. Postprocessing Def 13<br>Calibrating/Equalizing Odds 7 11<br>*Privacy-Preserving Inference* |
| **TRANSPARENCY** | Feature Selection 10 (Filter, Embedded, Wrapper)<br>Feature Engineering 10 12 | Regularization 3 12 (plus other under-fitting tuning)<br>Model Constraints 12<br>White & Glass-Box Models 3 4 | Local Interpretation 6 7 8 9 |


Academic Publications
(Source: Dimesions.ai)
6000
5000
4000
3000
2000
1000
0
2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022
"explainable artificial intelligence"
"interpretable machine learning"
Book Publications
% of yearly total
(Source: Google Books Ngram Viewer)
0.0000006000
0.0000005000
0.0000004000
0.0000003000
0.0000002000
0.0000001000
-
2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
"explainable artificial intelligence"
"interpretable machine learning"
Search Trends
yearly avg % of historical max
(Source: Google Trends)
70
60
50
40
30
20
10
0
2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022
"explainable artificial intelligence"
"interpretable machine learning"