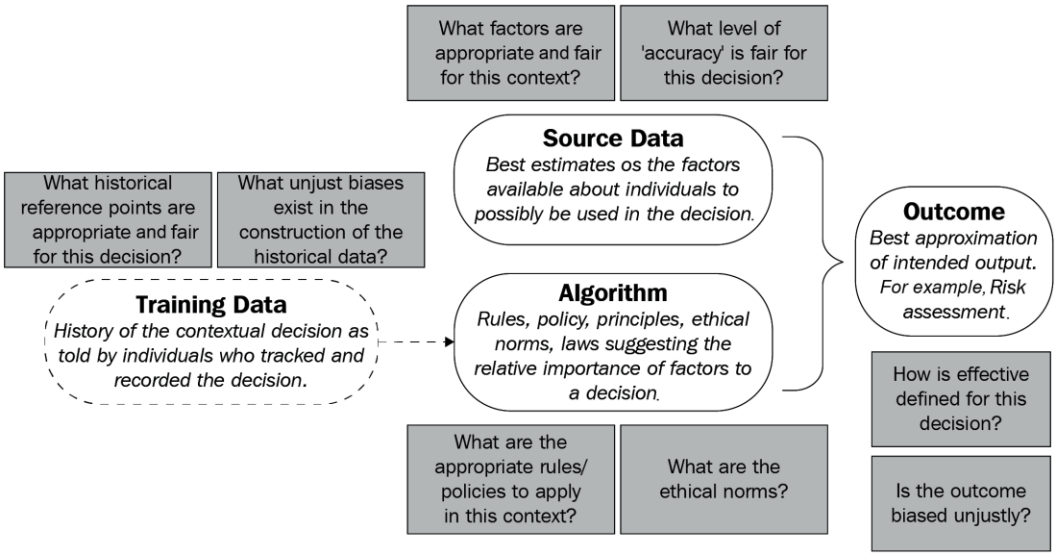
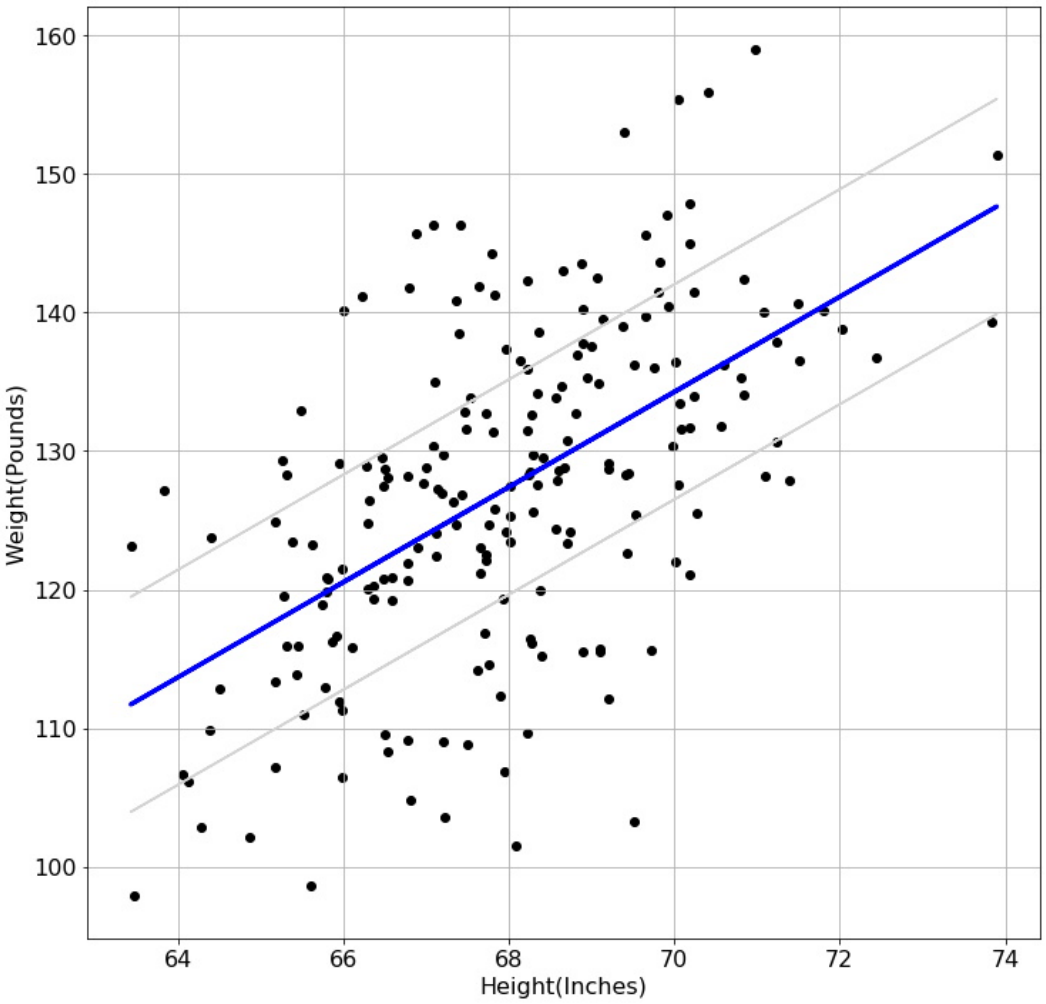
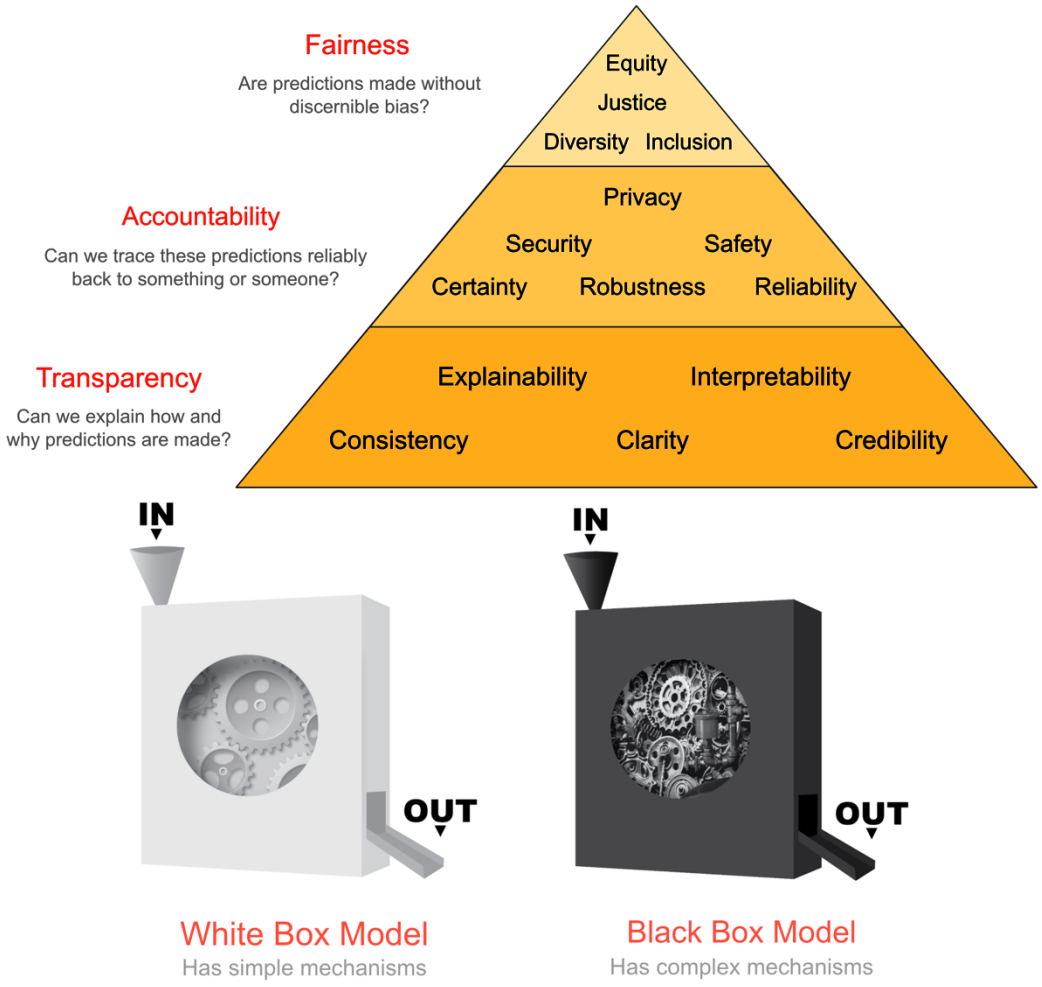
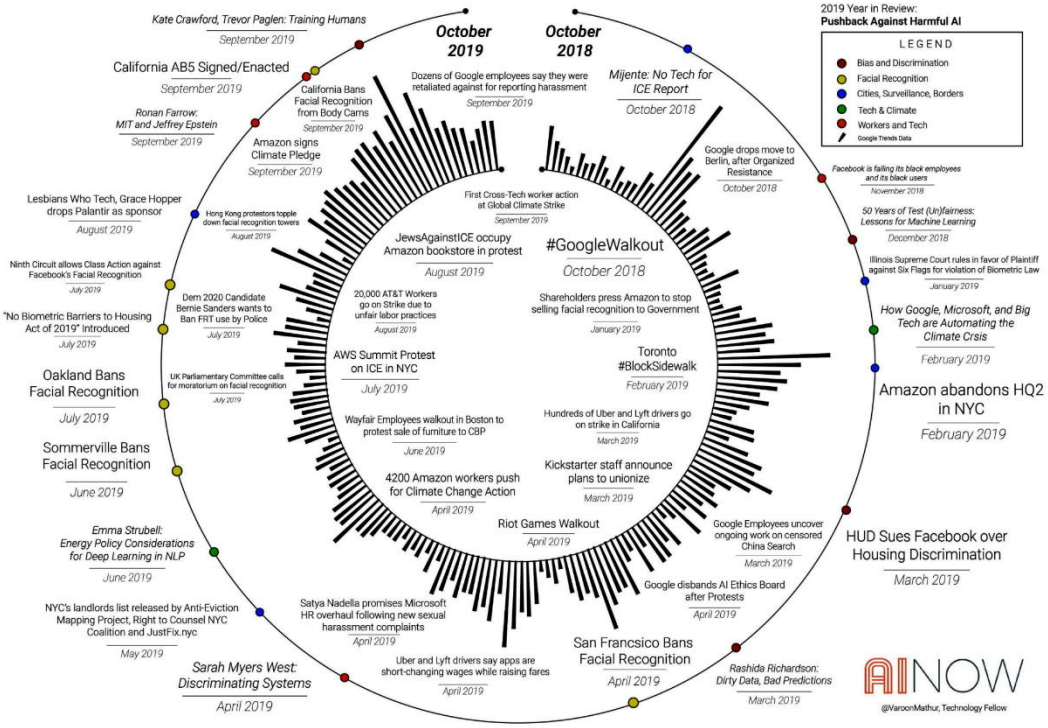
Chapter 1: Interpretation, Interpretability and Explainability; and why does it all matter?

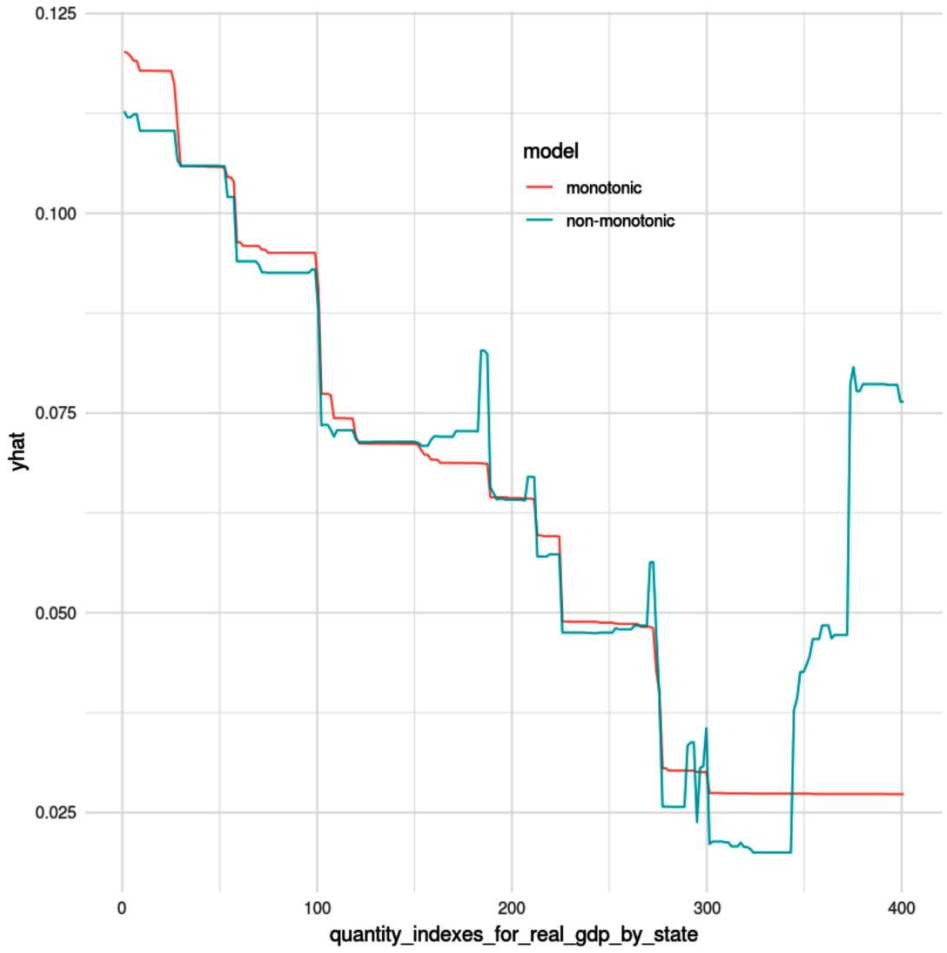



160
150
140
130
120
110
100
Weight(Pounds)
64
66
68
70
72
74
Height(Inches)

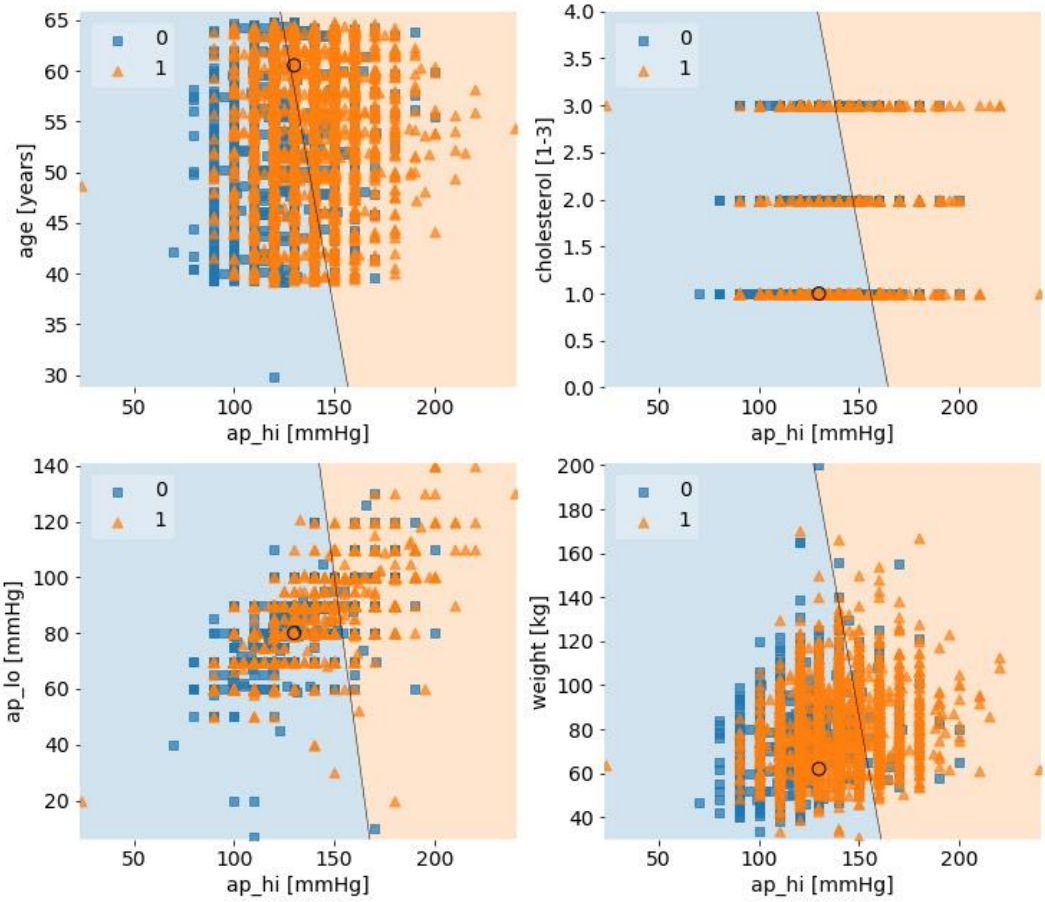

Fairness
Are predictions made without discernible bias?
Accountability
Can we trace these predictions reliably back to something or someone?
Transparency
Can we explain how and why predictions are made?
Equity
Justice
Diversity
Inclusion
Privacy
Security
Safety
Certainty
Robustness
Reliability
Explainability
Interpretability
Consistency
Clarity
Credibility
IN
OUT
IN
OUT
White Box Model
Has simple mechanisms
Black Box Model
Has complex mechanisms

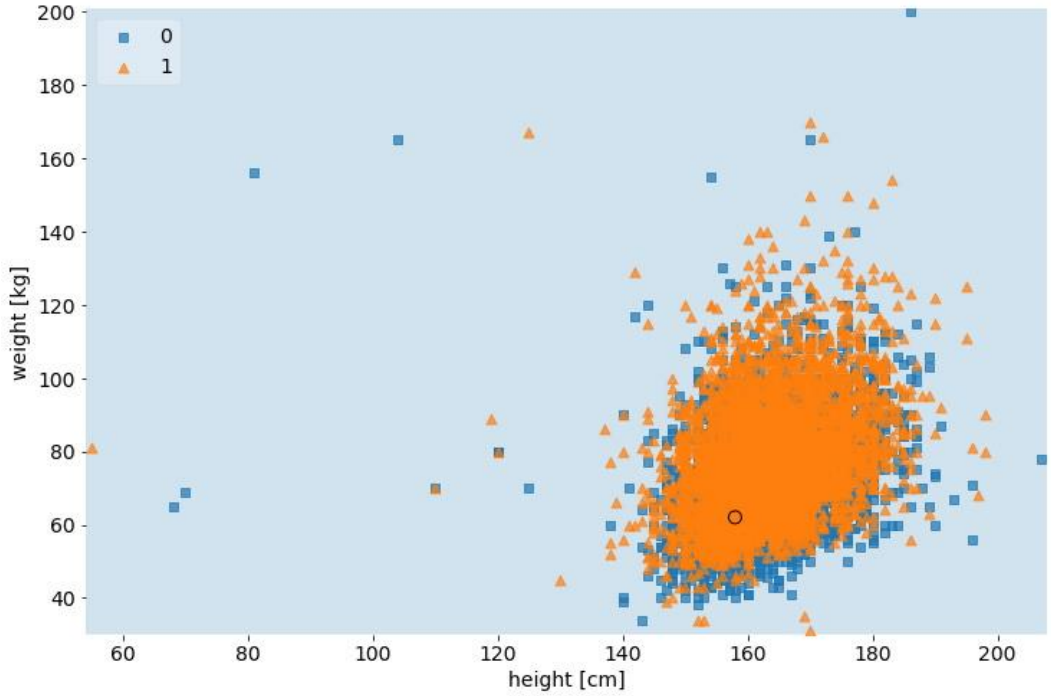

2019 Year in Review:
Pushback Against Harmful AI
LEGEND
Bias and Discrimination
Facial Recognition
Cities, Surveillance, Borders
Tech & Climate
Workers and Tech
Google Trends Data
October 2019
October 2018
Kate Crawford, Trevor Paglen: Training Humans
September 2019
California AB5 Signed/Enacted
September 2019
Ronan Farrow: MIT and Jeffrey Epstein
September 2019
Amazon signs Climate Pledge
September 2019
California Bans Facial Recognition from Body Cams
September 2019
Dozens of Google employees say they were retaliated against for reporting harassment
September 2019
Mijente: No Tech for ICE Report
October 2018
Google drops move to Berlin, after Organized Resistance
October 2018
Facebook is failing its black employees and its black users
November 2018
50 Years of Test (Un)fairness: Lessons for Machine Learning
December 2018
Illinois Supreme Court rules in favor of Plaintiff against Six Flags for violation of Biometric Law
January 2019
How Google, Microsoft, and Big Tech are Automating the Climate Crsis
February 2019
Amazon abandons HQ2 in NYC
February 2019
HUD Sues Facebook over Housing Discrimination
March 2019
Rashida Richardson: Dirty Data, Bad Predictions
March 2019
San Francsico Bans Facial Recognition
April 2019
Google disbands AI Ethics Board after Protests
April 2019
Google Employees uncover ongoing work on censored China Search
March 2019
Uber and Lyft drivers say apps are short-changing wages while raising fares
April 2019
Satya Nadella promises Microsoft HR overhaul following new sexual harassment complaints
April 2019
Sarah Myers West: Discriminating Systems
April 2019
NYC's landlords list released by Anti-Eviction Mapping Project, Right to Counsel NYC Coalition and JustFix.nyc
May 2019
Emma Strubell: Energy Policy Considerations for Deep Learning in NLP
June 2019
Sommerville Bans Facial Recognition
June 2019
Oakland Bans Facial Recognition
July 2019
"No Biometric Barriers to Housing Act of 2019" introduced
July 2019
Ninth Circuit allows Class Action against Facebook's Facial Recognition
July 2019
Lesbians Who Tech, Grace Hopper drops Palantir as sponsor
August 2019
Hong Kong protestors topple down facial recognition towers
August 2019
Dem 2020 Candidate Bernie Sanders wants to Ban FRT use by Police
July 2019
UK Parliamentary Committee calls for moratorium on facial recognition
July 2019
First Cross-Tech worker action at Global Climate Strike
September 2019
JewsAgainstICE occupy Amazon bookstore in protest
August 2019
#GoogleWalkout
October 2018
20,000 AT&T Workers go on Strike due to unfair labor practices
August 2019
Shareholders press Amazon to stop selling facial recognition to Government
January 2019
AWS Summit Protest on ICE in NYC
July 2019
Toronto #BlockSidewalk
February 2019
Wayfair Employees walkout in Boston to protest sale of furniture to CBP
June 2019
Hundreds of Uber and Lyft drivers go on strike in California
March 2019
4200 Amazon workers push for Climate Change Action
April 2019
Kickstarter staff announce plans to unionize
March 2019
Riot Games Walkout
April 2019
AINOW
@VaroonMathur, Technology Fellow

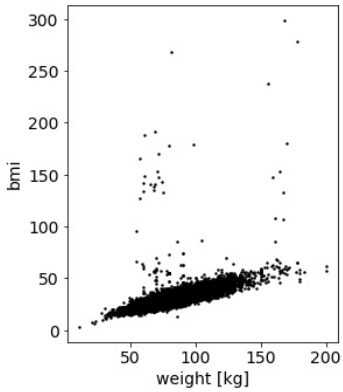
Chapter 2: Key Concepts of Interpretability

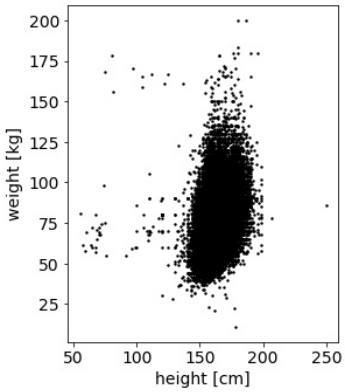


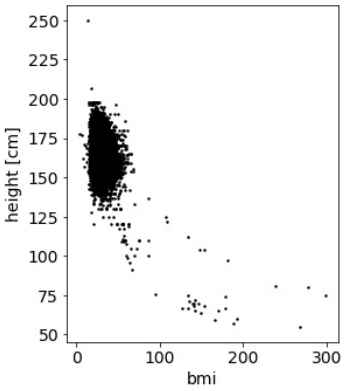
	count	mean	std	min	25%	50%	75%	max
age	70000.0	53.304309	6.755152	29.564122	48.36272	53.945351	58.391742	64.924433
gender	70000.0	1.349571	0.476838	1.000000	1.00000	1.000000	2.000000	2.000000
height	70000.0	164.359229	8.210126	55.000000	159.00000	165.000000	170.000000	250.000000
weight	70000.0	74.205690	14.395757	10.000000	65.00000	72.000000	82.000000	200.000000
ap_hi	70000.0	128.817286	154.011419	-150.000000	120.00000	120.000000	140.000000	16020.000000
ap_lo	70000.0	96.630414	188.472530	-70.000000	80.00000	80.000000	90.000000	11000.000000
cholesterol	70000.0	1.366871	0.680250	1.000000	1.00000	1.000000	2.000000	3.000000
gluc	70000.0	1.226457	0.572270	1.000000	1.00000	1.000000	1.000000	3.000000
smoke	70000.0	0.088129	0.283484	0.000000	0.00000	0.000000	0.000000	1.000000
alco	70000.0	0.053771	0.225568	0.000000	0.00000	0.000000	0.000000	1.000000
active	70000.0	0.803729	0.397179	0.000000	1.00000	1.000000	1.000000	1.000000
cardio	70000.0	0.499700	0.500003	0.000000	0.00000	0.000000	1.000000	1.000000

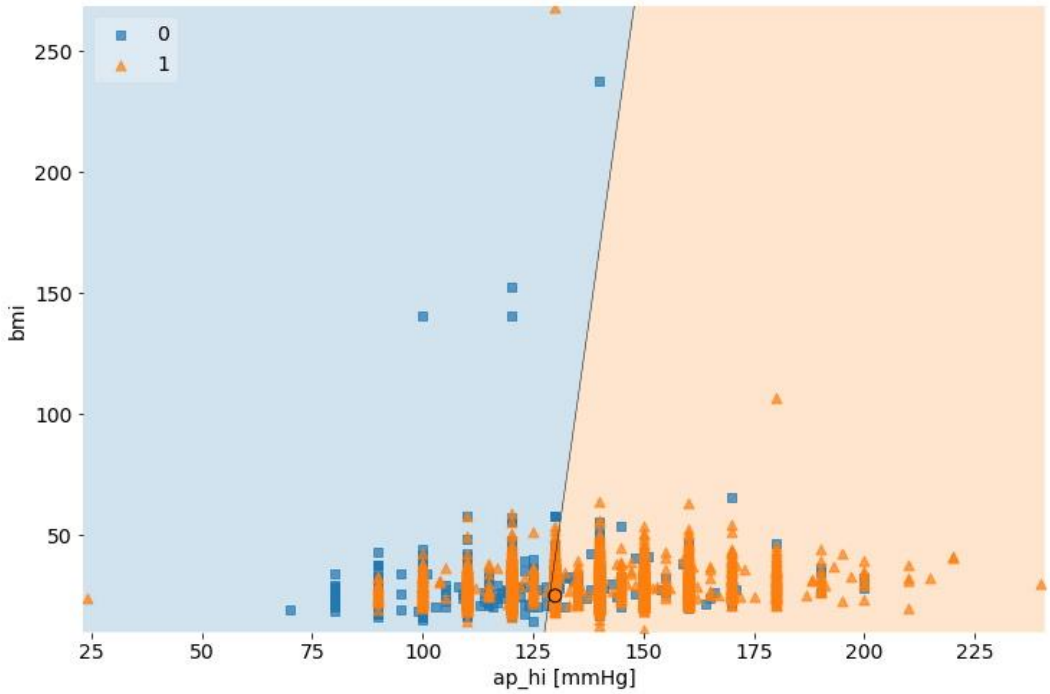



0
1
weight [kg]
height [cm]

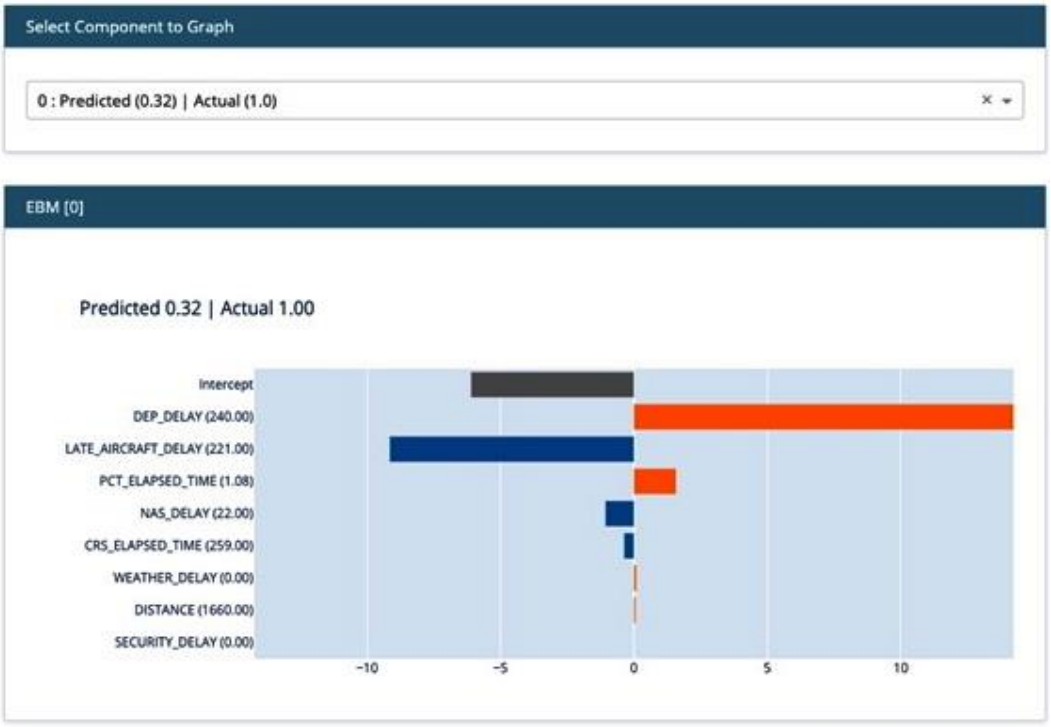

bmi
weight [kg]


weight [kg]
height [cm]


height [cm]
bmi


0
1
250
200
150
100
50
bmi
25
50
75
100
125
150
175
200
225
ap_hi [mmHg]

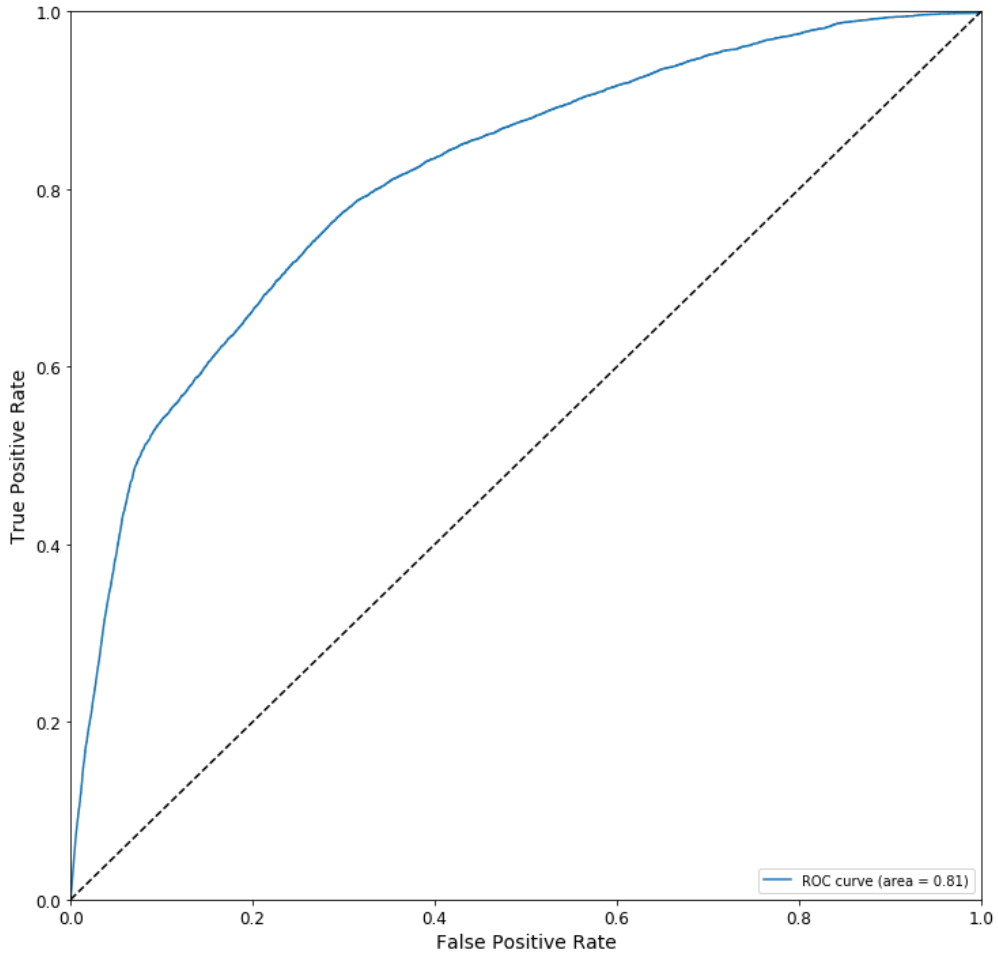
Chapter 3: Interpretation Challenges

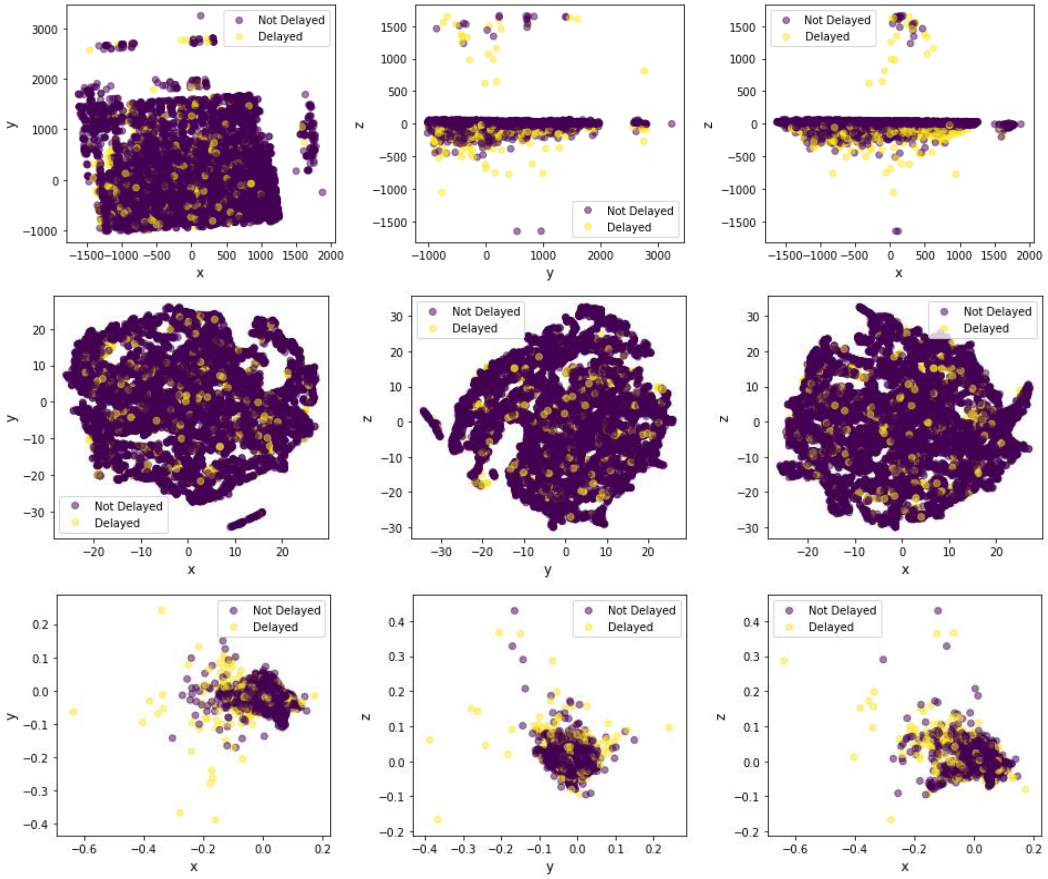


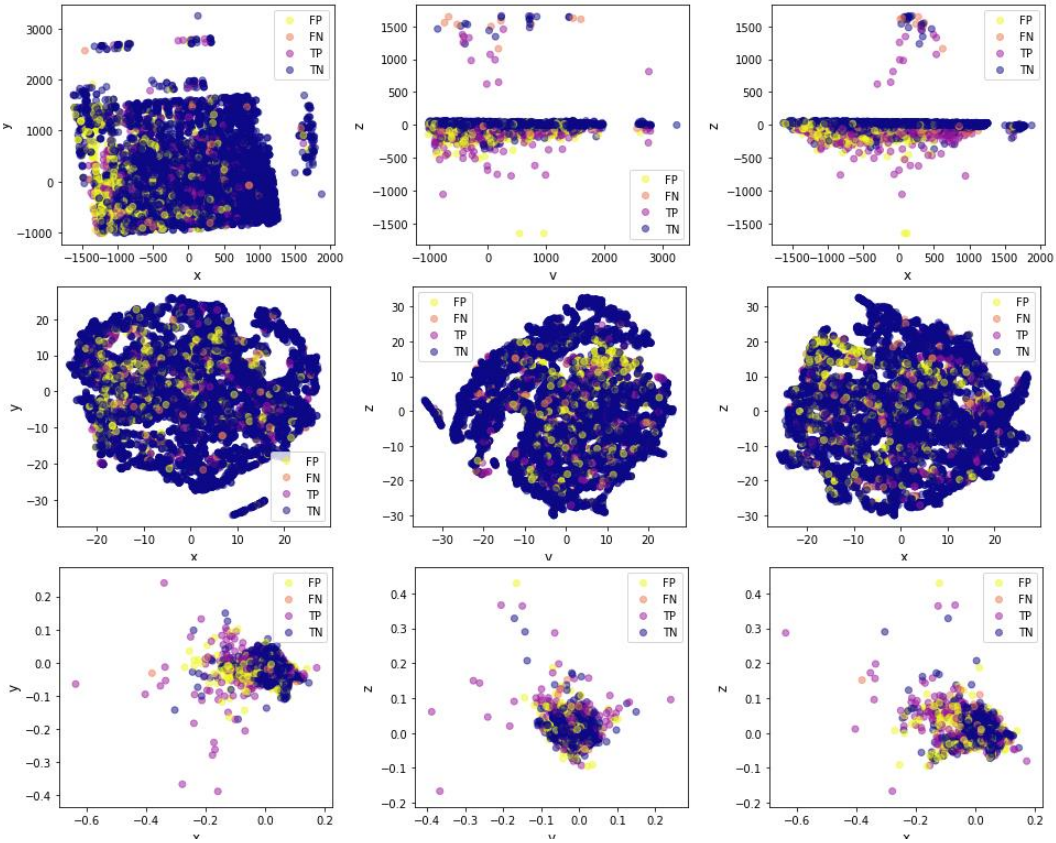
	ARR_DELAY	CARRIER_DELAY
8	168	136
16	20	5
18	242	242
19	62	62
22	19	19
26	26	0
29	77	77
32	19	19
33	18	1
40	36	16

	RMSE_train	RMSE_test	R2_test
mlp	3.18388	3.23262	0.987614
linear_poly	6.21339	6.33494	0.952433
random_forest	5.37769	6.35627	0.952112
linear_interact	6.45271	6.55931	0.949004
decision_tree	6.54292	7.53014	0.932791
linear	7.81963	7.88287	0.926347
ridge	7.92769	7.98758	0.924377
rulefit	8.8205	9.015	0.903672
knn	7.36014	9.26012	0.898362

	Accuracy_train	Accuracy_test	Recall_train	Recall_test	ROC_AUC_test	F1_test	MCC_test
mlp	0.998364	0.99851	0.984826	0.986444	0.999909	0.987819	0.987027
gradient_boosting	0.991725	0.991662	0.89293	0.893851	0.998885	0.929223	0.925619
decision_tree	0.983297	0.982895	0.856969	0.852215	0.994932	0.859182	0.85011
random_forest	0.938783	0.937879	0.997546	0.990559	0.992844	0.661333	0.677145
logistic	0.9786	0.978381	0.743923	0.742677	0.971935	0.807953	0.800067
knn	0.97289	0.965123	0.680667	0.607722	0.948387	0.680906	0.668176
naive_bayes	0.925115	0.925561	0.279126	0.274268	0.811872	0.310922	0.275073
ridge	0.890447	0.891255	0.777002	0.77802	0	0.466998	0.463706




Not Delayed
Delayed
x
y
z


FP
FN
TP
TN
x
y
z
v

	feature	coef
0	CRS_DEP_TIME	0.00454956
1	DEP_TIME	-0.00525032
2	DEP_DELAY	0.894124
3	TAXI_OUT	0.125274
4	WHEELS_OFF	-0.0006468
5	CRS_ARR_TIME	-0.000369914
6	CRS_ELAPSED_TIME	-0.0126273
7	DISTANCE	0.000676793
8	WEATHER_DELAY	-0.906354
9	NAS_DELAY	-0.674053
10	SECURITY_DELAY	-0.917398
11	LATE_AIRCRAFT_DELAY	-0.929841
12	DEP_AFPH	-0.0152963
13	ARR_AFPH	0.000548174
14	DEP_MONTH	-0.039835
15	DEP_DOW	-0.0182132
16	DEP_RFPH	-0.469474
17	ARR_RFPH	0.373844
18	ORIGIN_HUB	-1.02909
19	DEST_HUB	-0.394899
20	PCT_ELAPSED_TIME	45.0116

OLS Regression Results

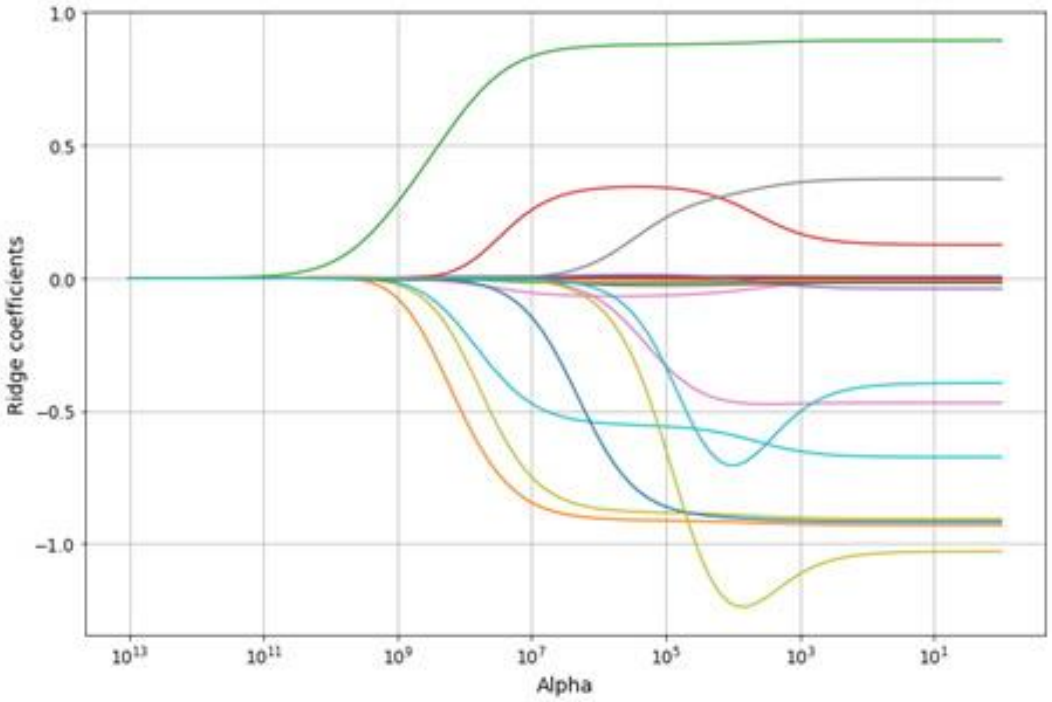
Dep. Variable:	CARRIER_DELAY	**R-squared:**	0.921
Model:	OLS	**Adj. R-squared:**	0.921
Method:	Least Squares	**F-statistic:**	4.251e+05
Date:	Wed, 02 Sep 2020	**Prob (F-statistic):**	0.00
Time:	13:32:20	**Log-Likelihood:**	-2.6574e+06
No. Observations:	764597	**AIC:**	5.315e+06
Df Residuals:	764575	**BIC:**	5.315e+06
Df Model:	21		
Covariance Type:	nonrobust		

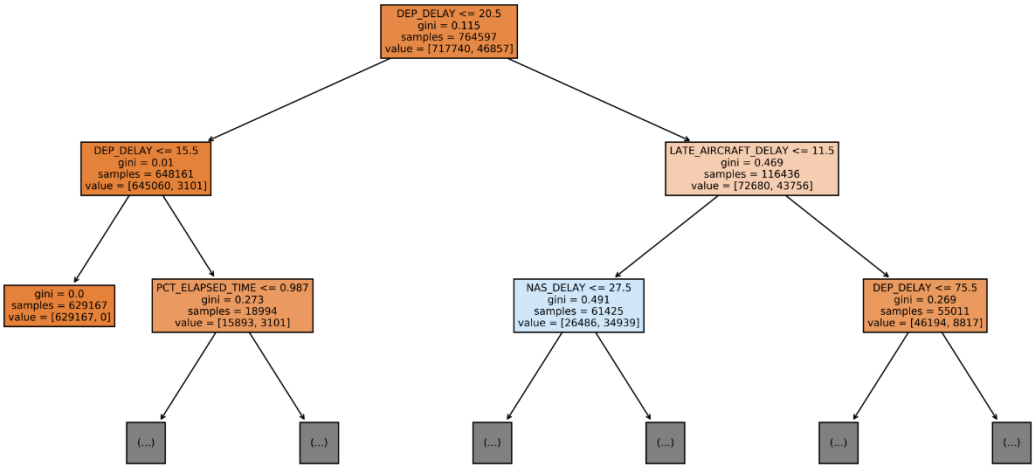
	coef	std err	t	P>|t|	[0.025	0.975]
const	-37.8618	0.125	-301.763	0.000	-38.108	-37.616
CRS_DEP_TIME	0.0045	7.24e-05	62.872	0.000	0.004	0.005
DEP_TIME	-0.0053	9.19e-05	-57.116	0.000	-0.005	-0.005
DEP_DELAY	0.8941	0.000	2951.056	0.000	0.894	0.895
DEP_AFPH	-0.0153	0.000	-47.725	0.000	-0.016	-0.015
DEP_RFPH	-0.4696	0.017	-27.353	0.000	-0.503	-0.436
TAXI_OUT	0.1253	0.001	104.120	0.000	0.123	0.128
WHEELS_OFF	-0.0006	6.7e-05	-9.646	0.000	-0.001	-0.001
CRS_ELAPSED_TIME	-0.0126	0.001	-19.132	0.000	-0.014	-0.011
PCT_ELAPSED_TIME	45.0113	0.117	384.073	0.000	44.782	45.241
DISTANCE	0.0007	8.02e-05	8.429	0.000	0.001	0.001
CRS_ARR_TIME	-0.0004	2.18e-05	-16.939	0.000	-0.000	-0.000
ARR_AFPH	0.0005	0.000	1.651	0.099	-0.000	0.001
ARR_RFPH	0.3739	0.013	28.386	0.000	0.348	0.400
WEATHER_DELAY	-0.9064	0.001	-995.366	0.000	-0.908	-0.905
NAS_DELAY	-0.6741	0.001	-829.129	0.000	-0.676	-0.672
SECURITY_DELAY	-0.9174	0.005	-167.857	0.000	-0.928	-0.907
LATE_AIRCRAFT_DELAY	-0.9298	0.001	-1827.018	0.000	-0.931	-0.929
DEP_MONTH	-0.0397	0.003	-15.019	0.000	-0.045	-0.034
DEP_DOW	-0.0180	0.004	-4.005	0.000	-0.027	-0.009
ORIGIN_HUB	-1.0291	0.027	-38.589	0.000	-1.081	-0.977
DEST_HUB	-0.3949	0.026	-15.041	0.000	-0.446	-0.343

Omnibus:	211121.387	**Durbin-Watson:**	2.001
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	24359701.834
Skew:	0.098	**Prob(JB):**	0.00
Kurtosis:	30.651	**Cond. No.**	5.69e+04

	feature	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]	t_abs
2	DEP_DELAY	0.894124	0.000302981	2951.09	0	0.89353	0.894717	2951.09
11	LATE_AIRCRAFT_DELAY	-0.929841	0.000508937	-1827.03	0	-0.930839	-0.928844	1827.03
8	WEATHER_DELAY	-0.906354	0.000910567	-995.373	0	-0.908138	-0.904569	995.373
9	NAS_DELAY	-0.674053	0.000812964	-829.13	0	-0.675646	-0.67246	829.13
20	PCT_ELAPSED_TIME	45.0116	0.117195	384.076	0	44.7819	45.2413	384.076
10	SECURITY_DELAY	-0.917398	0.00546544	-167.855	0	-0.928111	-0.906686	167.855
3	TAXI_OUT	0.125274	0.00120321	104.117	0	0.122916	0.127633	104.117
0	CRS_DEP_TIME	0.00454956	7.23674e-05	62.8675	0	0.00440772	0.00469139	62.8675
1	DEP_TIME	-0.00525032	9.19302e-05	-57.1121	0	-0.0054305	-0.00507014	57.1121
12	DEP_AFPH	-0.0152963	0.000320506	-47.7256	0	-0.0159245	-0.0146681	47.7256
18	ORIGIN_HUB	-1.02909	0.0266686	-38.5879	0	-1.08136	-0.976818	38.5879
17	ARR_RFPH	0.373844	0.0131708	28.3844	3.89612e-177	0.34803	0.399658	28.3844
16	DEP_RFPH	-0.469474	0.0171688	-27.3446	1.50325e-164	-0.503124	-0.435824	27.3446
6	CRS_ELAPSED_TIME	-0.0126273	0.000659852	-19.1366	1.3093e-81	-0.0139206	-0.011334	19.1366
5	CRS_ARR_TIME	-0.000369914	2.18388e-05	-16.9384	2.4083e-64	-0.000412717	-0.00032711	16.9384
14	DEP_MONTH	-0.039835	0.00264082	-15.0844	2.08773e-51	-0.045011	-0.0346591	15.0844
19	DEST_HUB	-0.394899	0.0262564	-15.0401	4.07781e-51	-0.44636	-0.343437	15.0401

	feature	coef_linear	coef_ridge
0	CRS_DEP_TIME	0.00454956	0.00501961
1	DEP_TIME	-0.00525032	-0.00441738
2	DEP_DELAY	0.894124	0.894292
3	TAXI_OUT	0.125274	0.125165
4	WHEELS_OFF	-0.0006468	0.000232365
5	CRS_ARR_TIME	-0.000369914	-0.00189765
6	CRS_ELAPSED_TIME	-0.0126273	-0.0125826
7	DISTANCE	0.000676793	0.0021406
8	WEATHER_DELAY	-0.906354	-0.906168
9	NAS_DELAY	-0.674053	-0.67396
10	SECURITY_DELAY	-0.917398	-0.917398
11	LATE_AIRCRAFT_DELAY	-0.929841	-0.929537
12	DEP_AFPH	-0.0152963	-0.0154111
13	ARR_AFPH	0.000548174	0.000532269
14	DEP_MONTH	-0.039835	-0.0398301
15	DEP_DOW	-0.0182132	-0.018213
16	DEP_RFPH	-0.469474	-0.469473
17	ARR_RFPH	0.373844	0.373847
18	ORIGIN_HUB	-1.02909	-1.02909
19	DEST_HUB	-0.394899	-0.394898
20	PCT_ELAPSED_TIME	45.0116	45.0116


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]
LATE_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 = 0.273
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]
(...)
(...)
(...)
(...)
(...)
(...)

	feature	importance
2	DEP_DELAY	0.527482
11	LATE_AIRCRAFT_DELAY	0.199153
20	PCT_ELAPSED_TIME	0.105381
8	WEATHER_DELAY	0.101649
9	NAS_DELAY	0.0627577
10	SECURITY_DELAY	0.00199756
7	DISTANCE	0.000993382
6	CRS_ELAPSED_TIME	0.000280958
3	TAXI_OUT	0.000238682
4	WHEELS_OFF	3.46469e-05
12	DEP_AFPH	3.10537e-05
5	CRS_ARR_TIME	0
1	DEP_TIME	0
13	ARR_AFPH	0
14	DEP_MONTH	0
15	DEP_DOW	0
16	DEP_RFPH	0
17	ARR_RFPH	0
18	ORIGIN_HUB	0
19	DEST_HUB	0
0	CRS_DEP_TIME	0

	rule	type	coef	support	importance
129	LATE_AIRCRAFT_DELAY <= 222.5 & WEATHER_DELAY <= 166.0 & DEP_DELAY > 344.0	rule	207.246	0.0016835	8.49625
80	DEP_DELAY > 477.5 & LATE_AIRCRAFT_DELAY <= 333.5	rule	170.948	0.00112233	5.72377
53	WEATHER_DELAY > 255.0 & DEP_DELAY > 490.5	rule	-333.579	0.000187056	4.56188
11	LATE_AIRCRAFT_DELAY	linear	-0.383065	1	4.48841
2	DEP_DELAY	linear	0.162592	1	4.25384
46	LATE_AIRCRAFT_DELAY <= 198.0 & DEP_DELAY <= 788.0 & DEP_DELAY > 341.5	rule	-95.8115	0.00149645	3.70359
57	DEP_DELAY > 1206.0	rule	254.29	0.000187056	3.47755
84	DEP_DELAY > 300.0 & DEP_DELAY > 576.5 & LATE_AIRCRAFT_DELAY <= 158.5	rule	121.199	0.000748223	3.31401
64	DEP_DELAY > 880.5	rule	102.969	0.000748223	2.81552
147	DEP_DELAY <= 37.5 & DEP_DELAY <= 370.5	rule	-9.13357	0.898429	2.7591
52	LATE_AIRCRAFT_DELAY <= 19.5 & DEP_DELAY <= 849.0 & DEP_DELAY > 66.5 & NAS_DELAY > 43.5	rule	-41.4699	0.00430228	2.71422
63	WEATHER_DELAY <= 61.0 & DEP_DELAY <= 849.0 & LATE_AIRCRAFT_DELAY <= 19.5 & DEP_DELAY > 270.0 & NAS_DELAY <= 43.5 & DEP_DELAY > 66.5	rule	99.0067	0.000748223	2.70718
153	WEATHER_DELAY <= 61.0 & DEP_DELAY <= 849.0 & LATE_AIRCRAFT_DELAY <= 19.5 & NAS_DELAY <= 43.5 & DEP_DELAY > 109.0 & DEP_DELAY > 66.5 & DEP_DELAY <= 270.0	rule	29.733	0.00598578	2.29348
169	WEATHER_DELAY > 61.0 & DEP_DELAY <= 849.0 & LATE_AIRCRAFT_DELAY <= 19.5 & NAS_DELAY <= 43.5 & DEP_DELAY > 66.5	rule	-45.9107	0.00224467	2.17271
162	DEP_DELAY > 117.0 & WEATHER_DELAY <= 10.0 & DEP_DELAY <= 225.0 & LATE_AIRCRAFT_DELAY <= 56.5 & DEP_DELAY <= 459.0 & DEP_DELAY > 68.5 & NAS_DELAY <= 66.0	rule	28.4973	0.00467639	1.9442
38	LATE_AIRCRAFT_DELAY <= 32.5 & NAS_DELAY <= 40.5 & DEP_DELAY <= 491.5 & DEP_DELAY > 57.5 & DEP_DELAY <= 245.5 & WEATHER_DELAY <= 20.0	rule	12.1724	0.0226337	1.81044
51	DEP_DELAY <= 20.5 & DEP_DELAY <= 68.5 & DEP_DELAY <= 459.0	rule	-4.56733	0.846053	1.64834

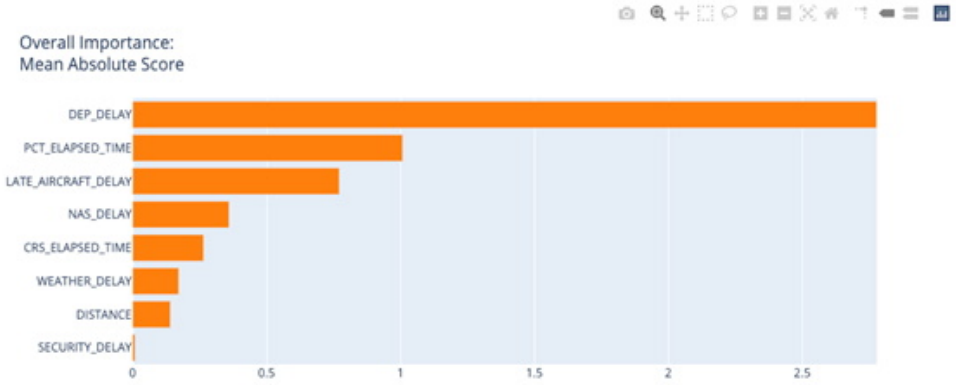
White Box?	Model Class	Properties that Increase Interpretability					Task		Performance Rank	
		Expl.	Linear	Monotone	Non-Interactive	Regul.	Regr.	Classif.	Regr.	Classif.
✓	Linear Regression	🟢	🟢	🟢	🟢	🟢	✓	✗	6	
✓	Regularized Regression	🟢	🟢	🟢	🟢	🟢	✓	✓	7	8
✓	Logistic Regression	🟢	🟡	🟢	🟢	🟢	✗	✓		5
✓	Gaussian Naïve 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
Interpretability	High	Mid-High	Low
Predictive Performance	Mid	High	High
Execution Speed Performance	High	Low	Mid

Select Component to Graph

Summary

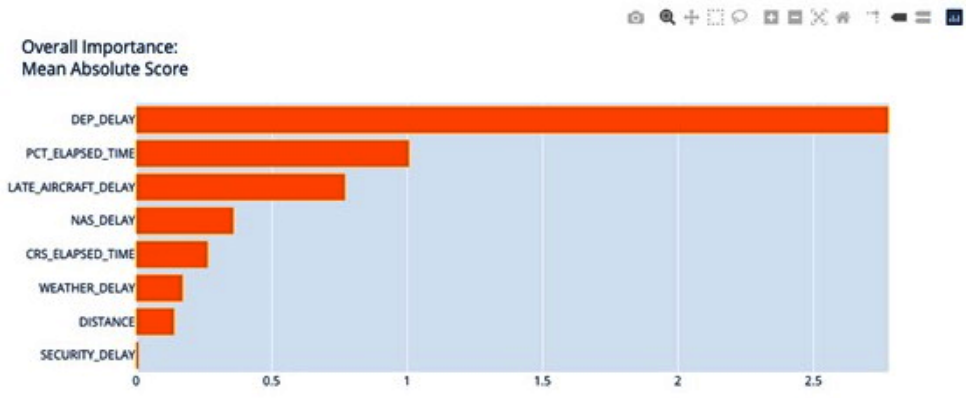
ExplainableBoostingClassifier_0 (Overall)



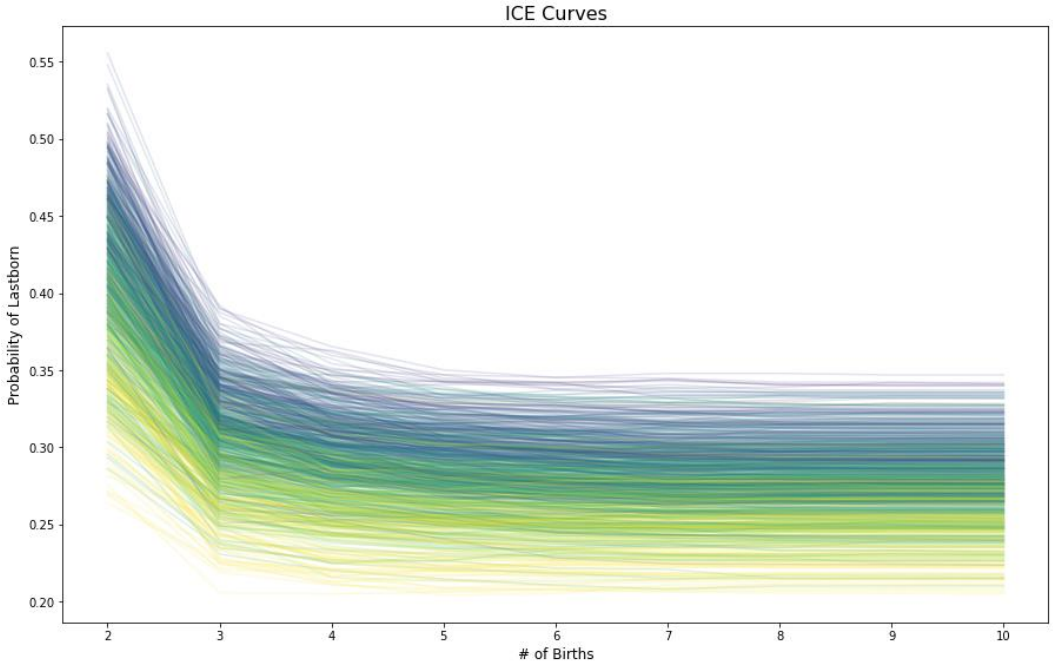
Select Component to Graph

Summary

ExplainableBoostingClassifier_0 (Overall)



Chapter 4: Fundamentals of Feature Importance and Impact



	Accuracy_train	Accuracy_test	Recall_train	Recall_test	Precision_train	Precision_test	F1_test	MCC_test
decision_tree	0.497	0.464	0.497	0.464	0.541	0.494	0.441	0.246
gradient_boosting	0.625	0.496	0.625	0.496	0.637	0.490	0.482	0.232
logistic	0.496	0.493	0.496	0.493	0.498	0.494	0.491	0.231
mlp	0.522	0.494	0.522	0.494	0.517	0.485	0.480	0.223
lda	0.501	0.492	0.501	0.492	0.500	0.489	0.477	0.201
random_forest	0.912	0.484	0.912	0.484	0.921	0.478	0.447	0.198

	name	dt_imp	dt_rank	gb_imp	gb_rank	rf_imp	rf_rank	avg_rank
28	birthn	0.851533	1	0.371305	1	0.198748	1	1
82	testelapse	0.0137081	3	0.0335579	2	0.0275725	2	2.33333
26	age	0.00667898	7	0.030532	3	0.0248301	3	4.33333
0	Q1	0.0253401	2	0.0236222	6	0.0159306	6	4.66667
81	introelapse	0.00505607	9	0.0297233	4	0.0224896	5	6
12	Q13	0.0080825	4	0.014516	7	0.0113429	8	6.33333
	:	:	:	:	:	:	:	:
90	country_GB	0	91	0.000755431	91	0.00194744	90	90.6667
92	country_NZ	0	93	0.00103713	90	0.000736748	91	91.3333
84	gender_undefined	0	87	0.000316311	94	0.000302447	94	91.6667
91	country_IE	0	92	0.000596172	92	0.000499432	93	92.3333

	name	first_coef_norm	middle_coef_norm	last_coef_norm	coef_weighted_avg
28	birthn	-0.412945	1.3538	-0.0132044	0.499431
26	age	0.0552764	-0.0265002	-0.149019	0.0804694
0	Q1	0.110523	0.0224566	-0.00631052	0.0540604
12	Q13	0.0793163	-0.0382582	-0.000743793	0.0427518
15	Q16	0.0604051	-0.0542339	-0.000581668	0.0385124
19	Q20	-0.0609848	0.0508594	0.0015853	0.0382996
39	EST1	0.0498431	-0.0622704	0.00411372	0.0371717
3	Q4	0.044028	-0.0576418	-0.000594055	0.0324218
59	CSN1	0.0316447	-0.0699186	0.00127486	0.0303448
	:	:	:	:	:
90	country_GB	-2.218e-05	-0.00138172	4.47252e-05	0.000352116
91	country_IE	-0.00014727	0.000314903	7.59173e-07	0.000136968
92	country_NZ	7.70629e-05	-0.000394417	5.52904e-06	0.000127852
87	gender_other	7.0736e-05	0.000394679	-8.67944e-06	0.000126324
84	gender_undefined	5.65254e-05	-9.94834e-05	-3.80303e-07	4.75334e-05

	name	first_coef_norm	middle_coef_norm	last_coef_norm	coef_weighted_avg
28	birthn	-0.315215	1.00305	-0.307128	0.475483
0	Q1	0.0899109	-0.0122606	-0.102456	0.0757905
12	Q13	0.0564803	-0.0337293	-0.0462968	0.0476102
51	AGR3	-0.0392475	-0.00558213	0.0523123	0.0357299
15	Q16	0.0395618	-0.0363935	-0.0235674	0.0333487
6	Q7	-0.00407858	0.0644172	-0.0396745	0.0305362
24	Q25	-0.0350918	0.0343628	0.01946	0.0295807
16	Q17	0.034915	-0.00978912	-0.036297	0.0294317
77	OPN9	-0.0326552	0.0447374	0.00925253	0.0275268
33	EXT5	-0.000968246	0.0064997	-0.00331607	0.0030811
85	gender_male	0.00272646	-0.00239968	-0.00169943	0.00229827
81	introelapse	-0.00168655	0.00484121	-0.00127851	0.00229491
36	EXT8	-0.000826513	-0.00324433	0.00327175	0.00223465
7	Q8	-0.000763955	0.0039999	-0.00183324	0.0018961

	name	first_coef_norm	middle_coef_norm	last_coef_norm	coef_weighted_avg
28	birthn	-0.315215	1.00305	-0.307128	0.475483
0	Q1	0.0899109	-0.0122606	-0.102456	0.0757905
12	Q13	0.0564803	-0.0337293	-0.0462968	0.0476102
51	AGR3	-0.0392475	-0.00558213	0.0523123	0.0357299
15	Q16	0.0395618	-0.0363935	-0.0235674	0.0333487
6	Q7	-0.00407858	0.0644172	-0.0396745	0.0305362
24	Q25	-0.0350918	0.0343628	0.01946	0.0295807
16	Q17	0.034915	-0.00978912	-0.036297	0.0294317
77	OPN9	-0.0326552	0.0447374	0.00925253	0.0275268

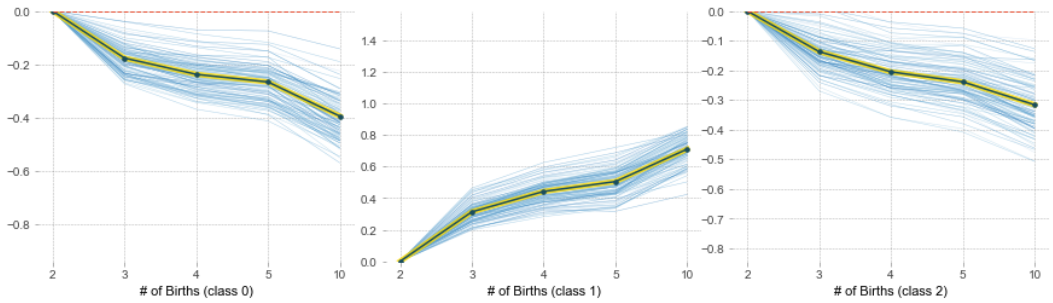
33	EXT5	-0.000968246	0.0064997	-0.00331607	0.0030811
85	gender_male	0.00272646	-0.00239968	-0.00169943	0.00229827
81	introelapse	-0.00168655	0.00484121	-0.00127851	0.00229491
36	EXT8	-0.000826513	-0.00324433	0.00327175	0.00223465
7	Q8	-0.000763955	0.0039999	-0.00183324	0.0018961

	name	dt_imp	gb_imp	rf_imp	log_imp	lda_imp	mlp_imp	avg_imp
28	birthn	0.1385	0.10735	0.07604	0.11818	0.08199	0.11172	0.10563
0	Q1	0.00832	0.00688	0.00428	0.00509	0.01103	0.0093	0.00749
26	age	0.00107	0.00327	0.00496	0.00713	-0.00122	0.00183	0.00284
12	Q13	0.00098	-0.00252	-6e-05	0.00428	0.00235	0.00499	0.00167
3	Q4	0	0.00274	0.00163	0.00178	0.0006	0.00214	0.00148
16	Q17	0.00119	-0.00201	0.00255	0.00122	0.00179	0.00273	0.00124
51	AGR3	0.00032	-7e-05	-0.00156	0.00109	0.00339	0.0039	0.00118
24	Q25	0	-6e-05	-0.00087	0.00106	0.00112	0.00465	0.00098
30	EXT2	0	0.00073	0.00161	0.00075	-0.00076	0.00348	0.00097
	:	:	:	:	:	:	:	:
69	OPN1	0.00015	-0.00035	-0.00175	-0.00062	-0.00018	0.00088	-0.00031
21	Q22	0	-0.00279	-0.00025	0.00019	-0.00207	0.00242	-0.00042
79	source	0	-0.00048	6e-05	0.00094	-0.00135	-0.0017	-0.00042
25	Q26	0.00126	-0.00144	-0.00216	0.00015	-0.00211	0.0007	-0.0006
22	Q23	0	-0.00169	-0.00012	-7e-05	-0.0017	-0.00028	-0.00064

	Accuracy_test
decision_tree	0.325639
gradient_boosting	0.388483
random_forest	0.40958
logistic	0.383053
lda	0.409265
mlp	0.37977

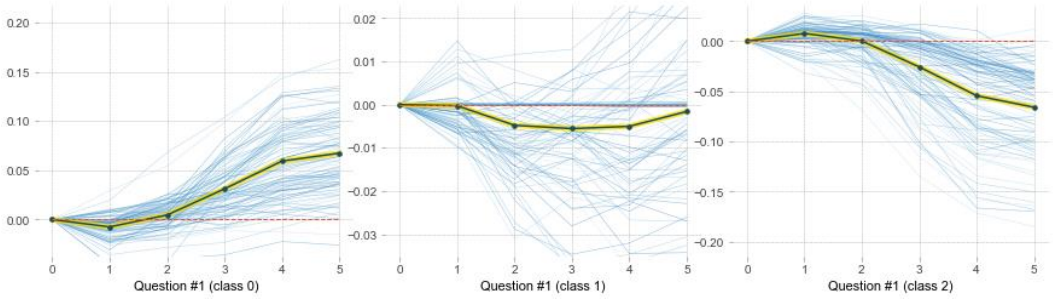
PDP for feature "# of Births"

Number of unique grid points: 5



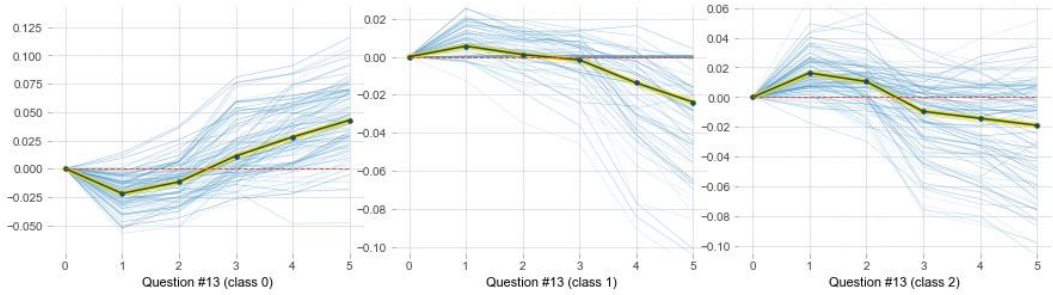
PDP for feature "Question #1"

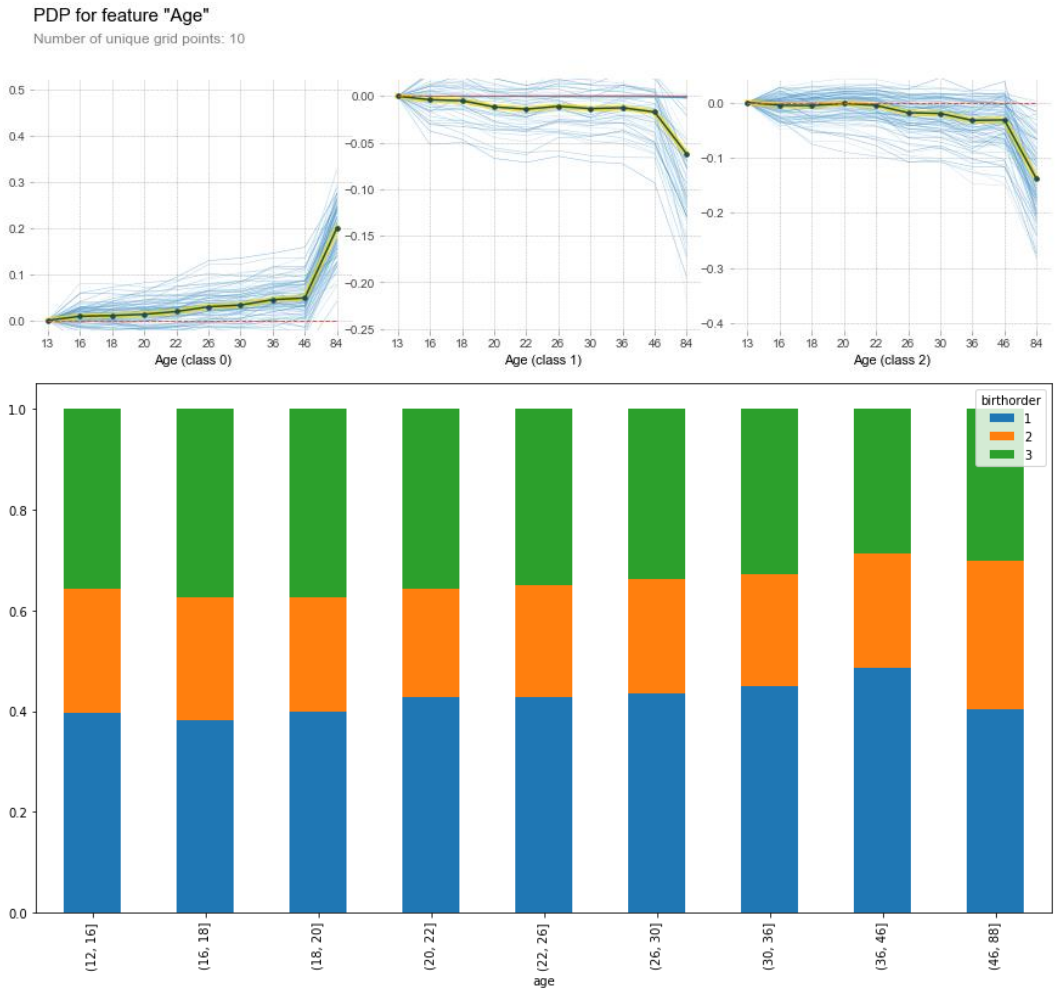
Number of unique grid points: 6



PDP for feature "Question #13"

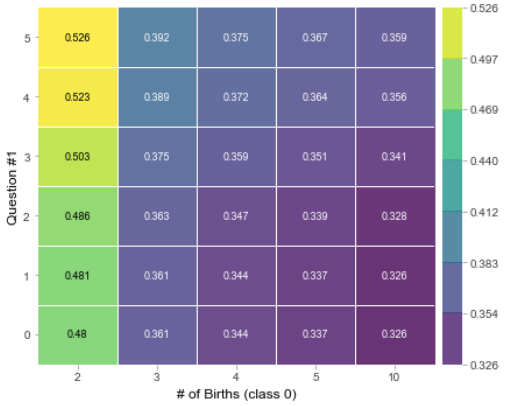
Number of unique grid points: 6

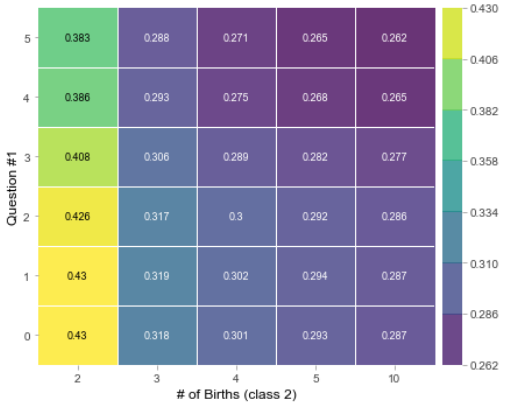



PDP for feature "Age"
Number of unique grid points: 10
Age (class 0)
Age (class 1)
Age (class 2)
birthorder
1
2
3
(12, 16]
(16, 18]
(18, 20]
(20, 22]
(22, 26]
(26, 30]
(30, 36]
(36, 46]
(46, 88]
age

PDP interact for "# of Births" and "Question #1"

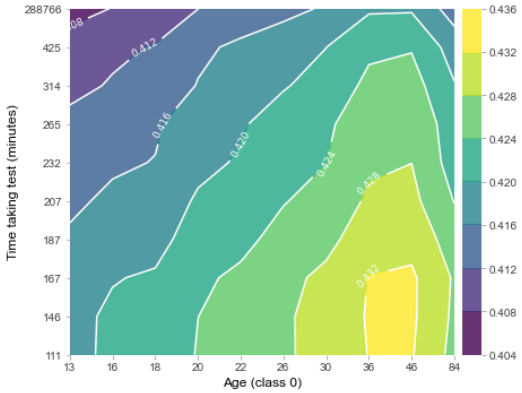
Number of unique grid points: (# of Births: 5, Question #1: 6)

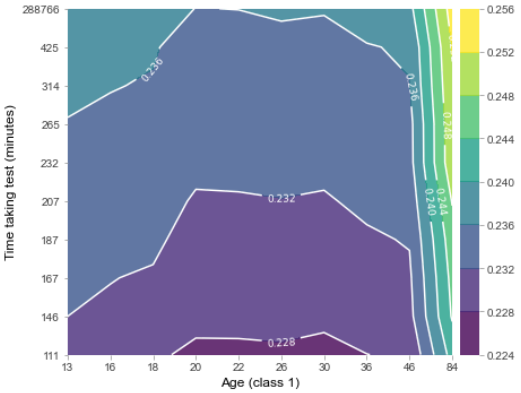


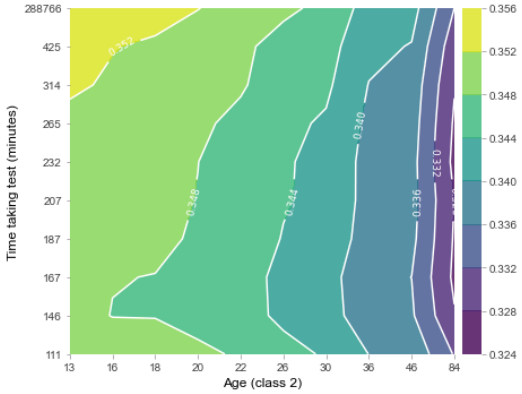


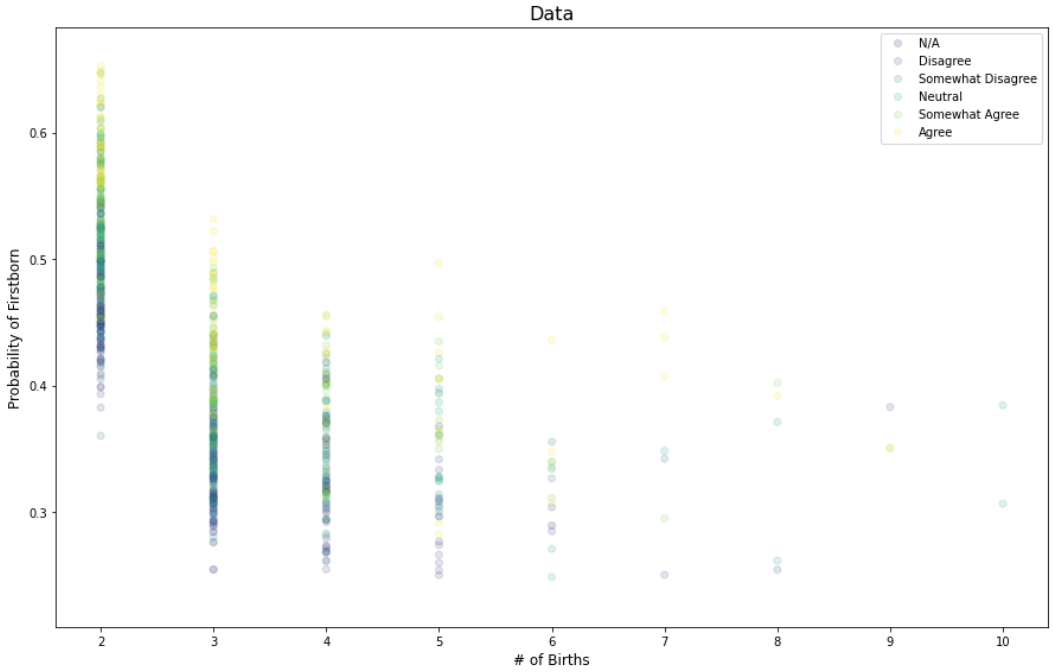
PDP interact for "Age" and "Time taking test (minutes)"

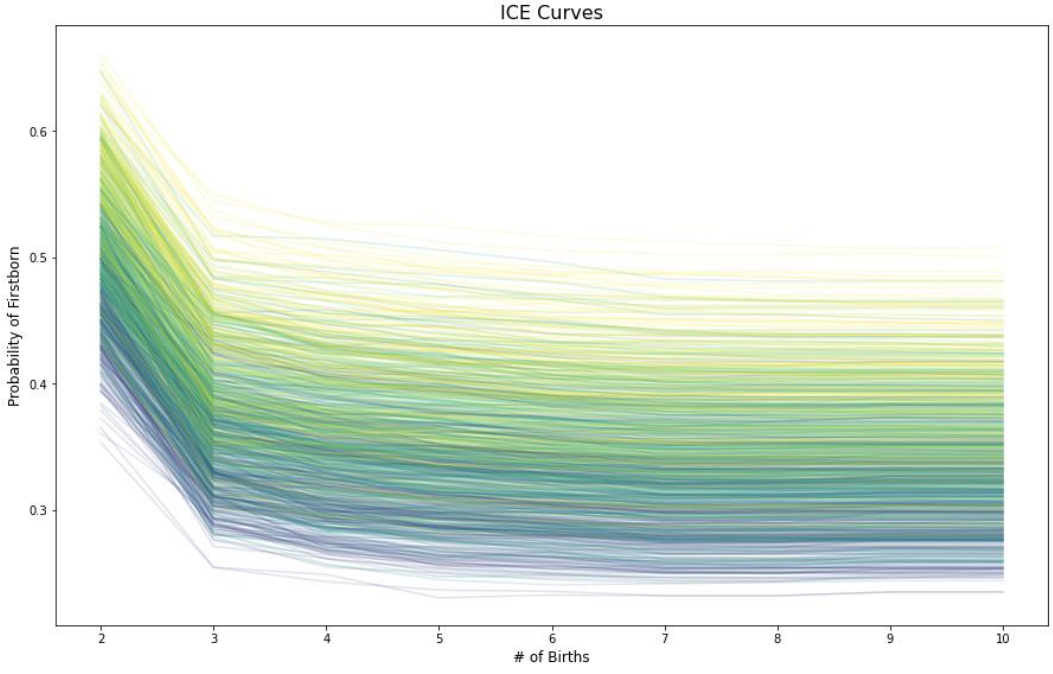
Number of unique grid points: (Age: 10, Time taking test (minutes): 10)

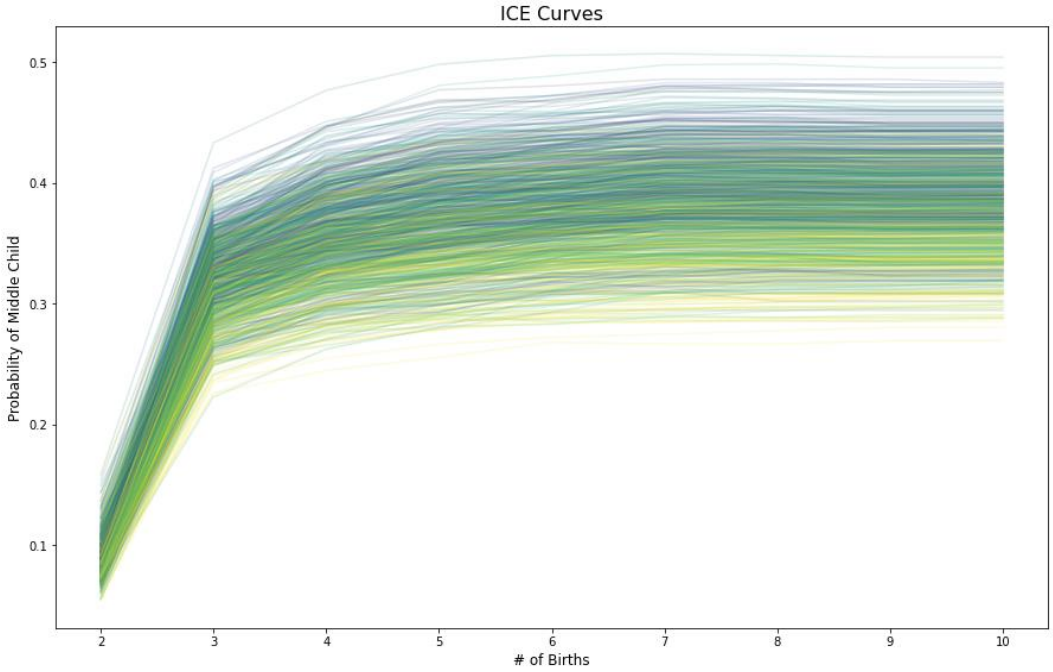






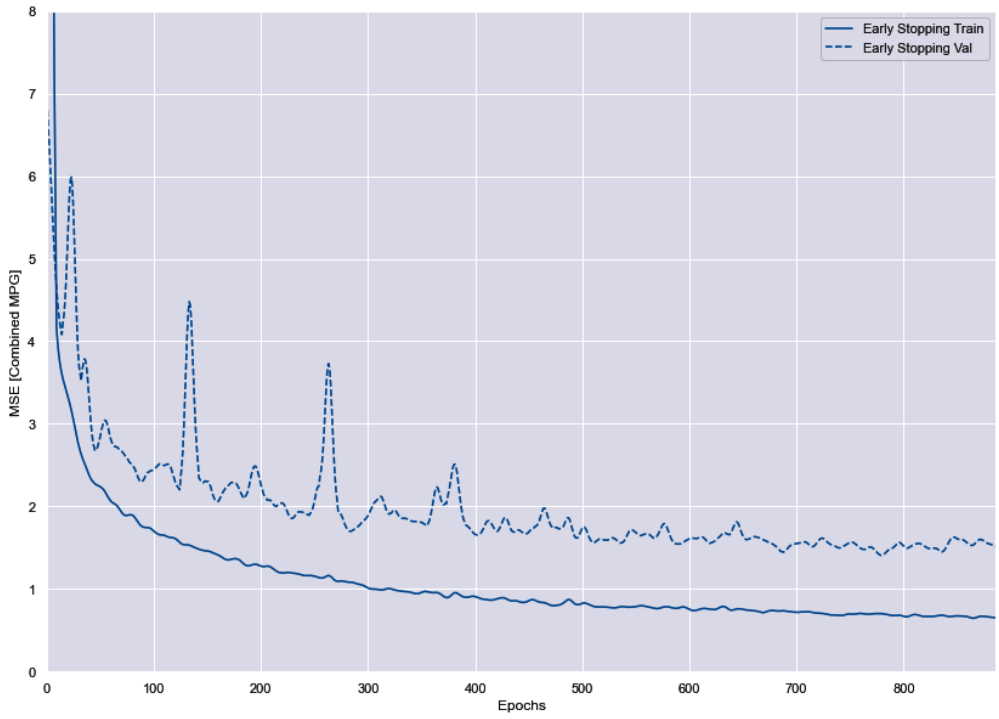

Data
N/A
Disagree
Somewhat Disagree
Neutral
Somewhat Agree
Agree
Probability of Firstborn
0.6
0.5
0.4
0.3
2
3
4
5
6
7
8
9
10
# of Births

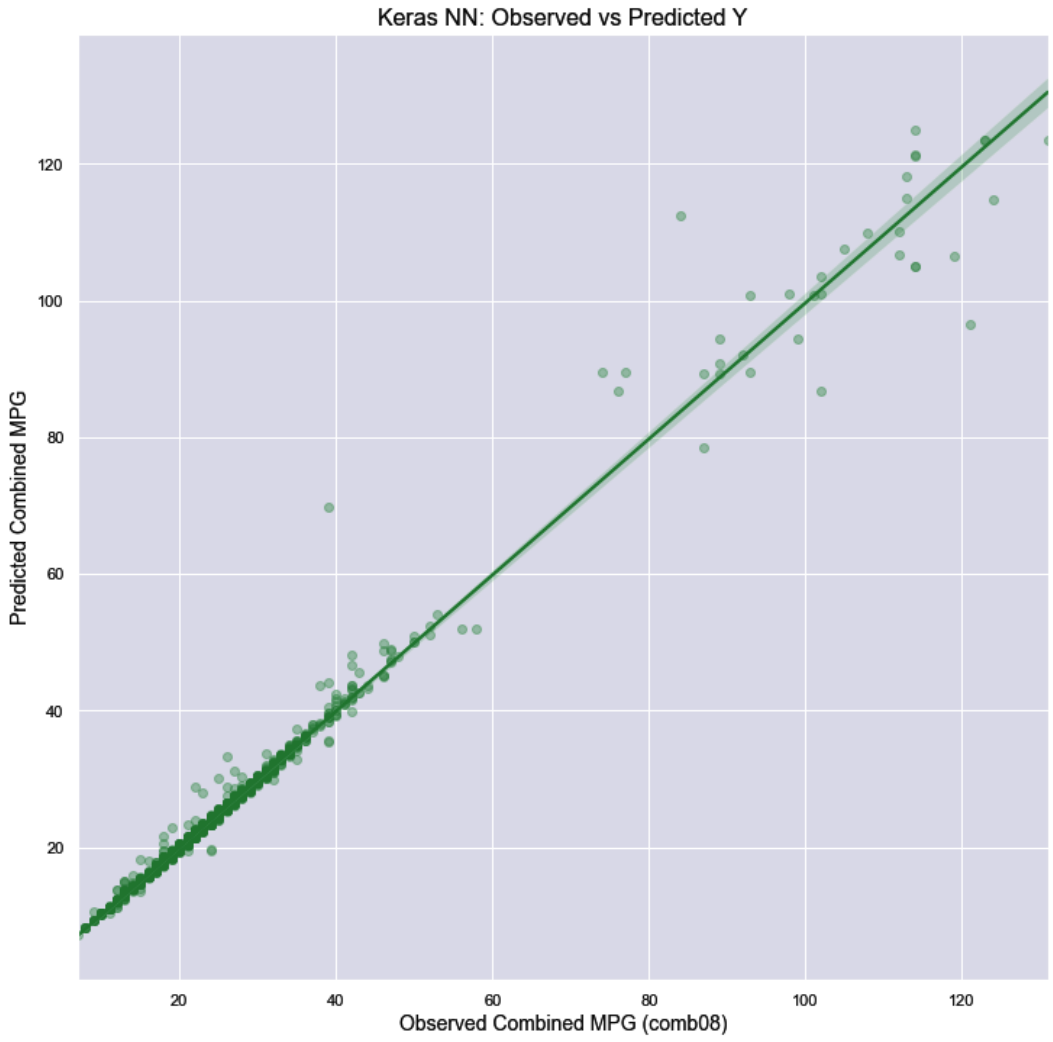

ICE Curves
Probability of Firstborn
0.6
0.5
0.4
0.3
2
3
4
5
6
7
8
9
10
# of Births

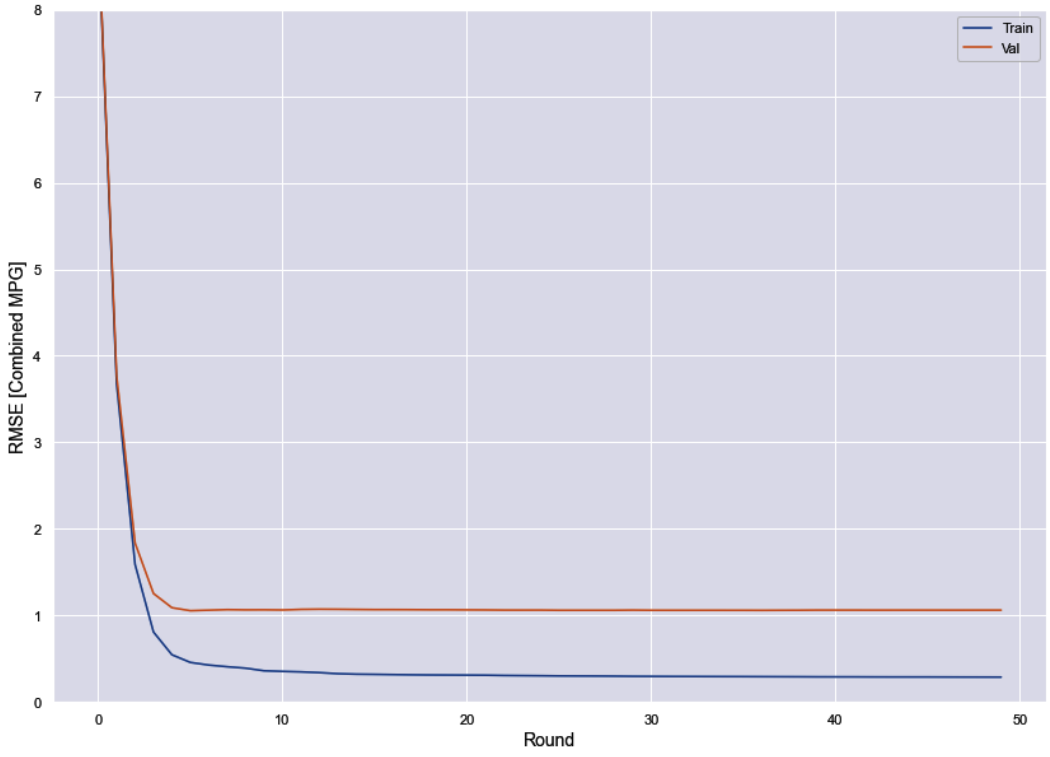

ICE Curves
Probability of Middle Child
0.5
0.4
0.3
0.2
0.1
2
3
4
5
6
7
8
9
10
# of Births

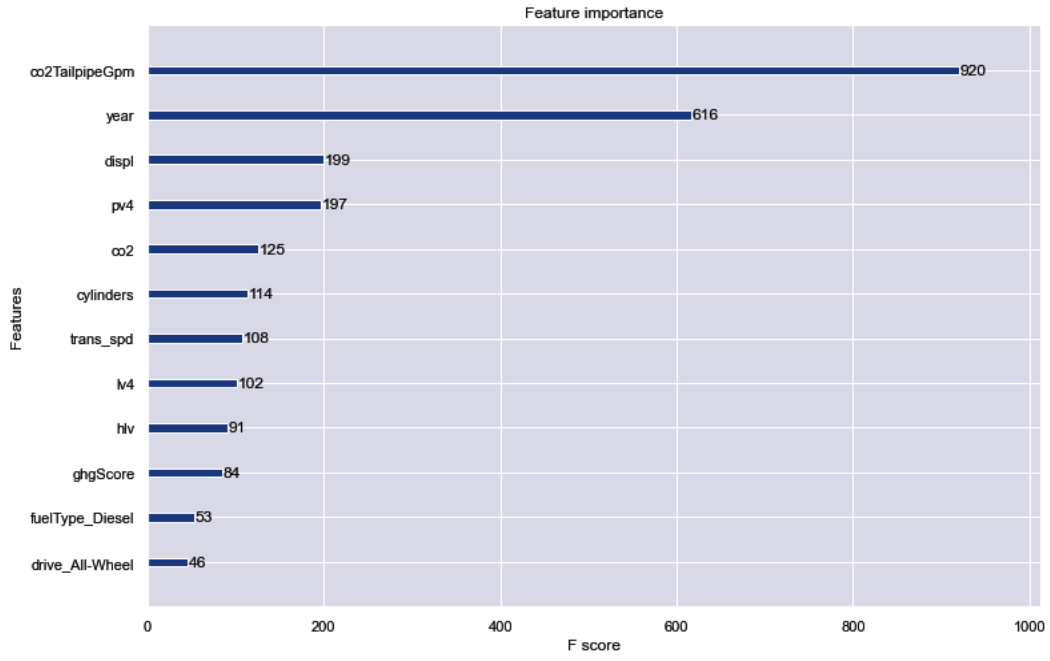
Chapter 5: Global Model-Agnostic Interpretation Methods

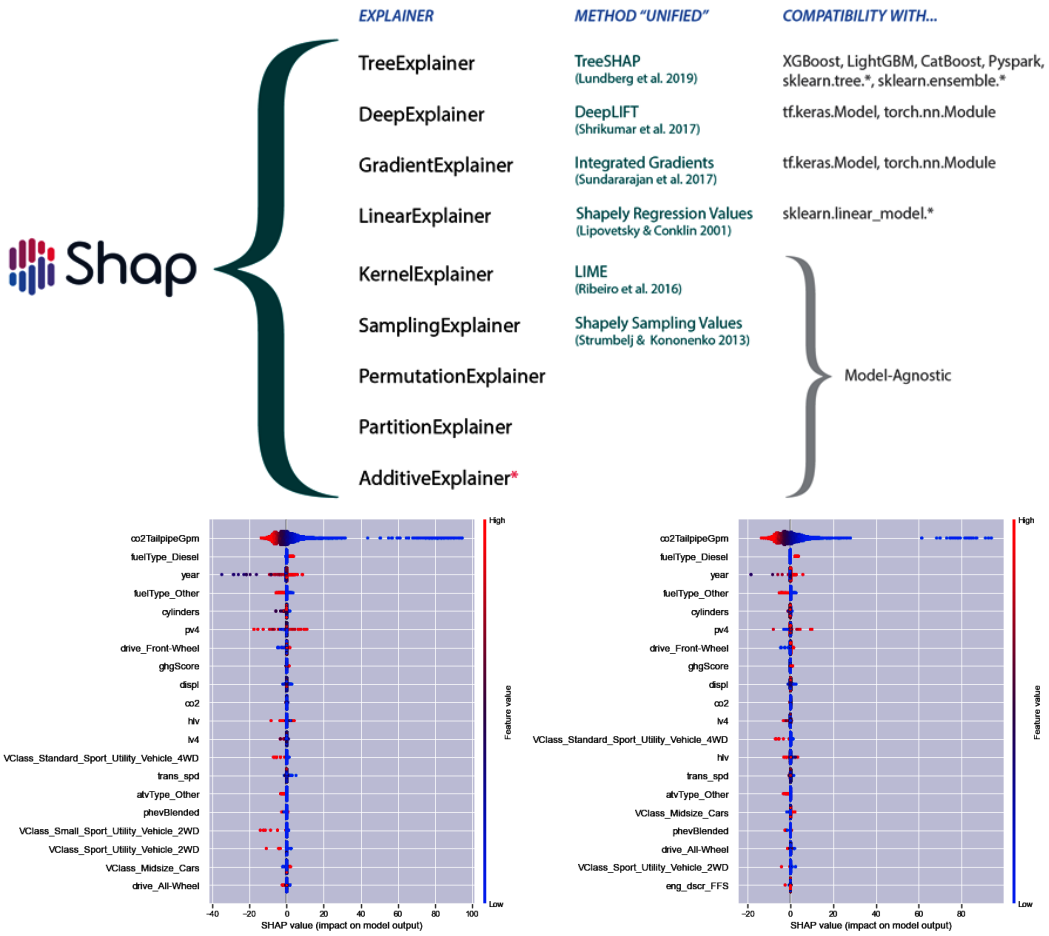
	rule	type	coef	support	importance
4	co2TailpipeGpm	linear	-0.034222	1.000000	3.865460
90	fuelType_Electricity > 0.5	rule	18.393562	0.006206	1.444470
104	co2TailpipeGpm <= 367.0 & fuelType_Electricity > 0.5 & ghgScore > 4.5	rule	20.776127	0.003546	1.235005
148	atvType_EV > 0.5 & pv4 > 42.0	rule	15.512440	0.003546	0.922113
193	eng_dscr_PFI > 0.5 & displ <= 0.30000001192092896	rule	12.607181	0.005319	0.917024
14	fuelType_Diesel	linear	2.967849	1.000000	0.488102
95	hpv > 45.0 & co2TailpipeGpm <= 28.5	rule	7.200693	0.003546	0.428034
127	co2TailpipeGpm > 124.0 & trans_spd > 0.5 & atvType_Other > 0.5 & cylinders > 1.5 & co2TailpipeGpm <= 408.5 & co2TailpipeGpm <= 302.5	rule	-7.168212	0.003546	0.426103
215	co2TailpipeGpm <= 250.53571319580078 & co2TailpipeGpm <= 320.5 & co2TailpipeGpm <= 410.5 & co2TailpipeGpm > 40.5	rule	3.457322	0.015071	0.421223
146	VClass_Small_Sport_Utility_Vehicle_2WD <= 0.5 & co2TailpipeGpm <= 45.5	rule	7.574405	0.002660	0.390100

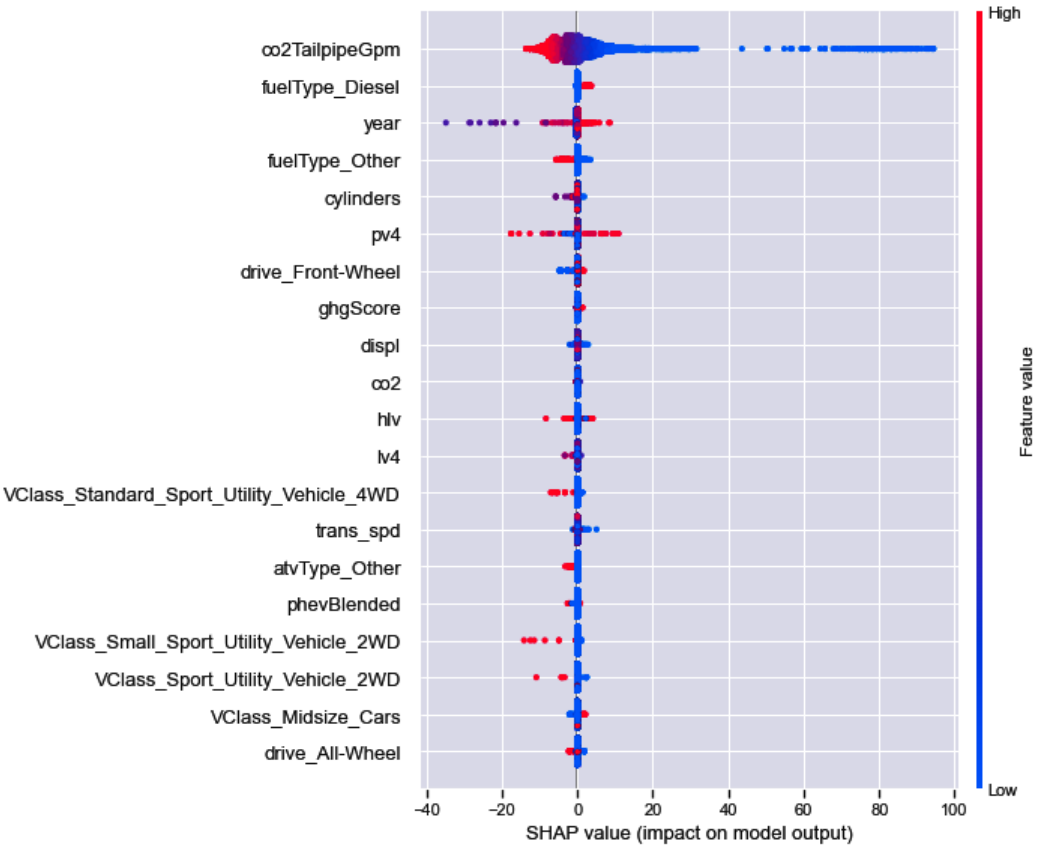


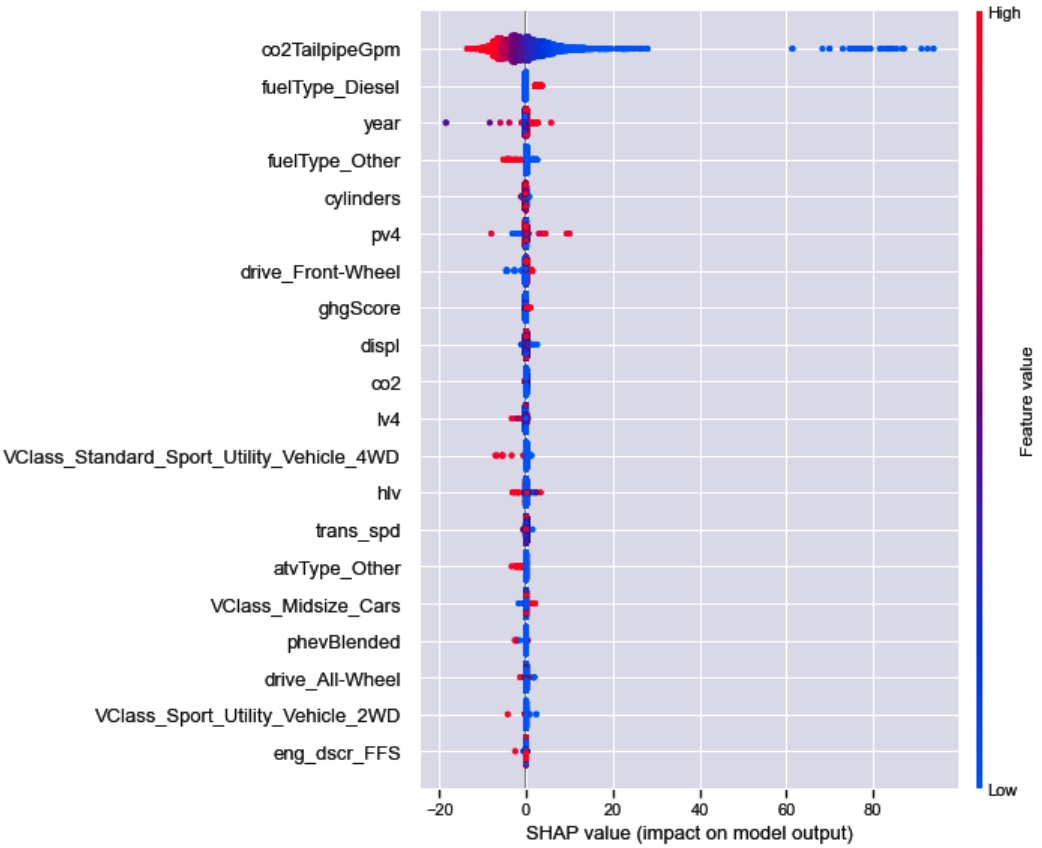

Keras NN: Observed vs Predicted Y
Predicted Combined MPG
Observed Combined MPG (comb08)
20
40
60
80
100
120

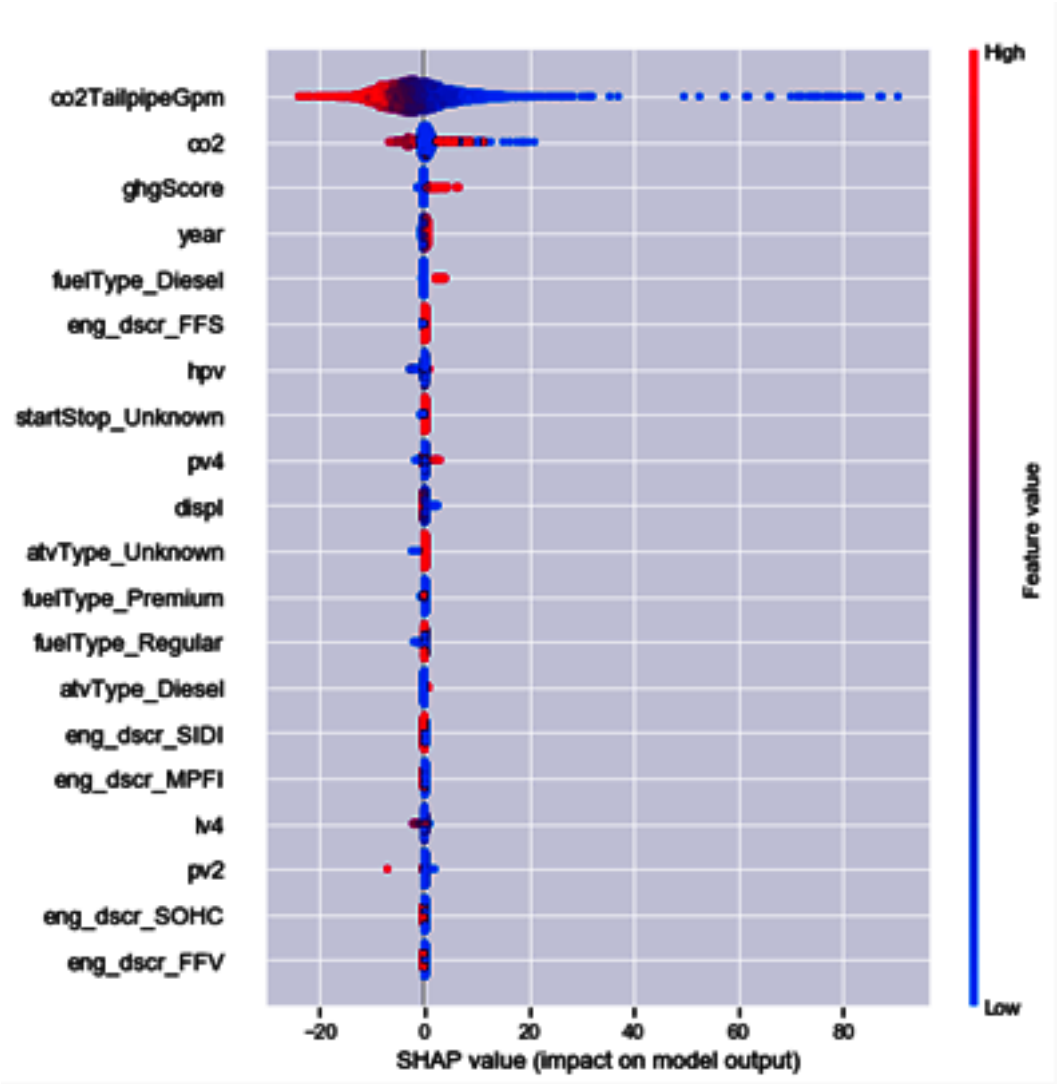

Train
Val
RMSE [Combined MPG]
Round
0
1
2
3
4
5
6
7
8
0
10
20
30
40
50

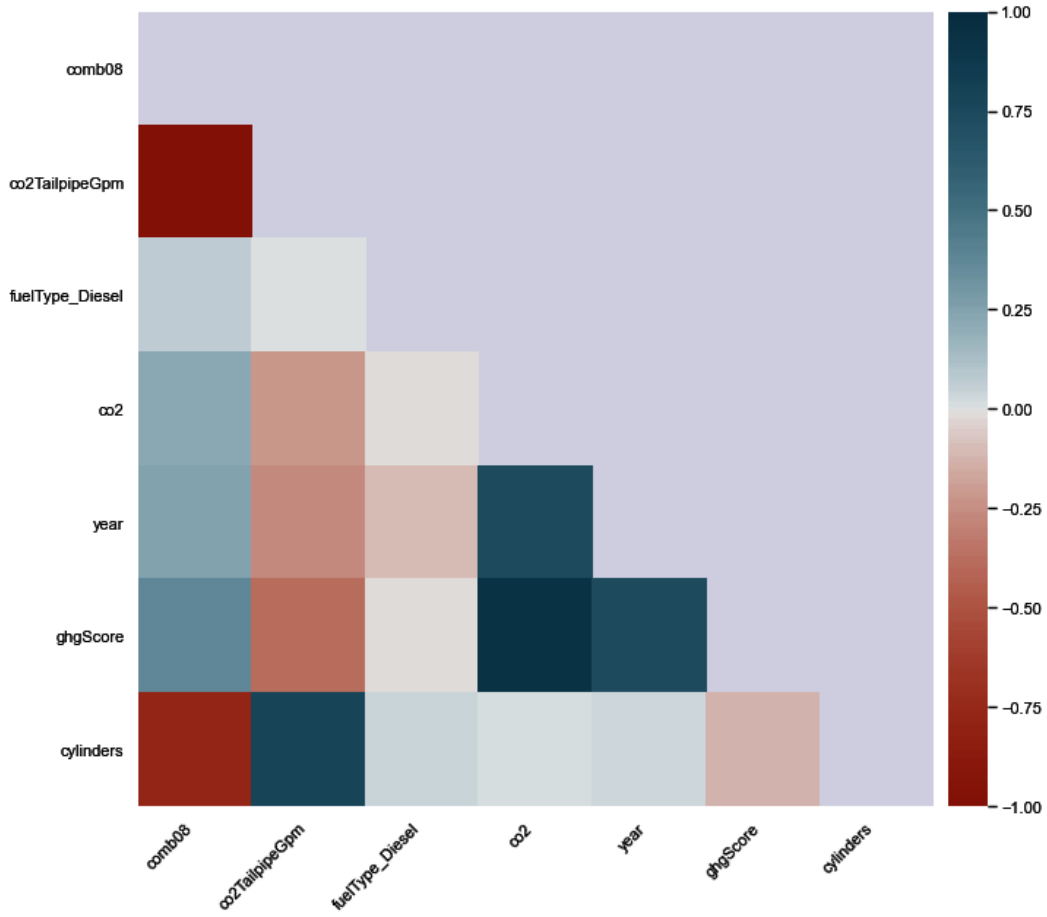

Feature importance
co2TailpipeGpm
year
displ
pv4
co2
cylinders
trans_spd
lv4
hlv
ghgScore
fuelType_Diesel
drive_All-Wheel
920
616
199
197
125
114
108
102
91
84
53
46
Features
F score
0
200
400
600
800
1000

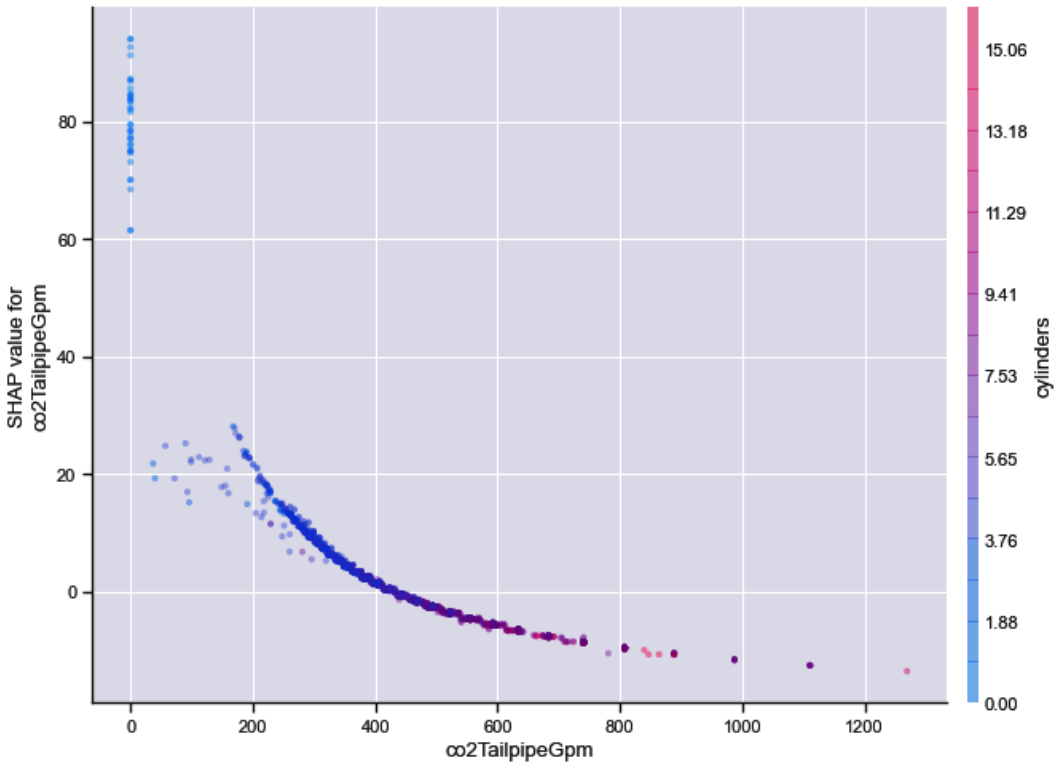

Shap
EXPLAINER
METHOD "UNIFIED"
COMPATIBILITY WITH...
TreeExplainer
TreeSHAP
(Lundberg et al. 2019)
XGBoost, LightGBM, CatBoost, Pyspark, sklearn.tree.*, sklearn.ensemble.*
DeepExplainer
DeepLIFT
(Shrikumar et al. 2017)
tf.keras.Model, torch.nn.Module
GradientExplainer
Integrated Gradients
(Sundararajan et al. 2017)
tf.keras.Model, torch.nn.Module
LinearExplainer
Shapely Regression Values
(Lipovetsky & Conklin 2001)
sklearn.linear_model.*
KernelExplainer
LIME
(Ribeiro et al. 2016)
SamplingExplainer
Shapely Sampling Values
(Strumbelj & Kononenko 2013)
PermutationExplainer
PartitionExplainer
AdditiveExplainer*
Model-Agnostic
co2TailpipeGpm
fuelType_Diesel
year
fuelType_Other
cylinders
pv4
drive_Front-Wheel
ghgScore
displ
co2
hlv
lv4
VClass_Standard_Sport_Utility_Vehicle_4WD
trans_spd
atvType_Other
phevBlended
VClass_Small_Sport_Utility_Vehicle_2WD
VClass_Sport_Utility_Vehicle_2WD
VClass_Midsize_Cars
drive_All-Wheel
eng_dscr_FFS
High
Low
Feature value
SHAP value (impact on model output)


co2TailpipeGpm
fuelType_Diesel
year
fuelType_Other
cylinders
pv4
drive_Front-Wheel
ghgScore
displ
co2
hlv
lv4
VClass_Standard_Sport_Utility_Vehicle_4WD
trans_spd
atvType_Other
phevBlended
VClass_Small_Sport_Utility_Vehicle_2WD
VClass_Sport_Utility_Vehicle_2WD
VClass_Midsize_Cars
drive_All-Wheel
−40
−20
0
20
40
60
80
100
SHAP value (impact on model output)
High
Feature value
Low

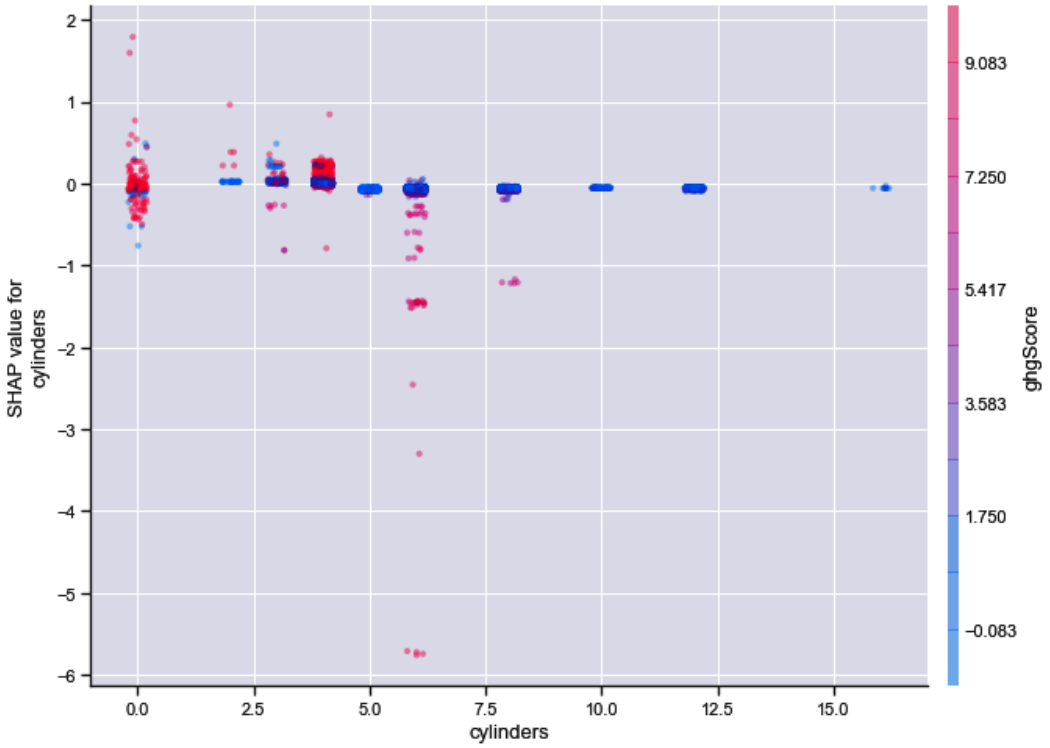

co2TailpipeGpm
fuelType_Diesel
year
fuelType_Other
cylinders
pv4
drive_Front-Wheel
ghgScore
displ
co2
lv4
VClass_Standard_Sport_Utility_Vehicle_4WD
hlv
trans_spd
atvType_Other
VClass_Midsize_Cars
phevBlended
drive_All-Wheel
VClass_Sport_Utility_Vehicle_2WD
eng_dscr_FFS
−20
0
20
40
60
80
SHAP value (impact on model output)
High
Feature value
Low


co2TailpipeGpm
co2
ghgScore
year
fuelType_Diesel
eng_dscr_FFS
hpv
startStop_Unknown
pv4
displ
atvType_Unknown
fuelType_Premium
fuelType_Regular
atvType_Diesel
eng_dscr_SIDI
eng_dscr_MPFI
lv4
pv2
eng_dscr_SOHC
eng_dscr_FFV
-20
0
20
40
60
80
SHAP value (impact on model output)
High
Feature value
Low

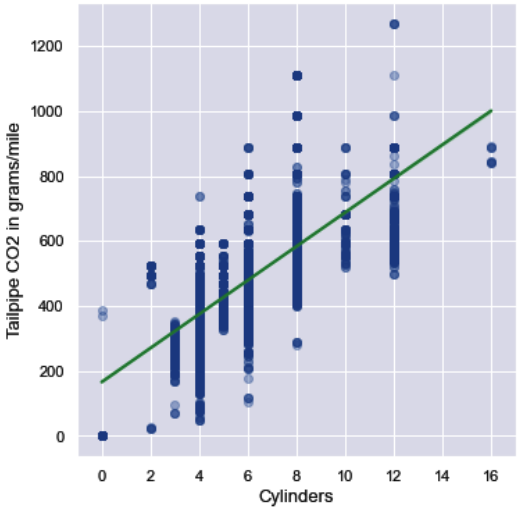

comb08
co2TailpipeGpm
fuelType_Diesel
co2
year
ghgScore
cylinders
comb08
co2TailpipeGpm
fuelType_Diesel
co2
year
ghgScore
cylinders
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

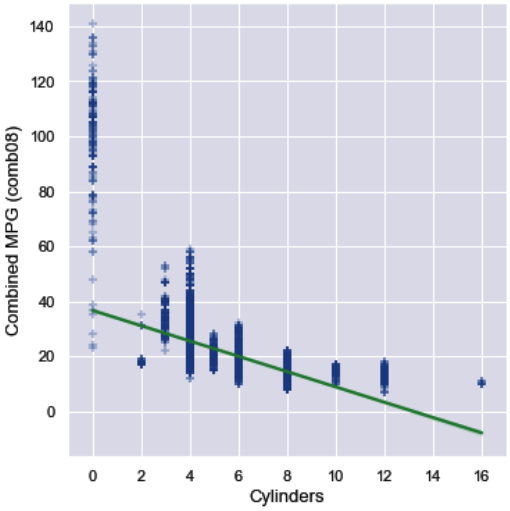


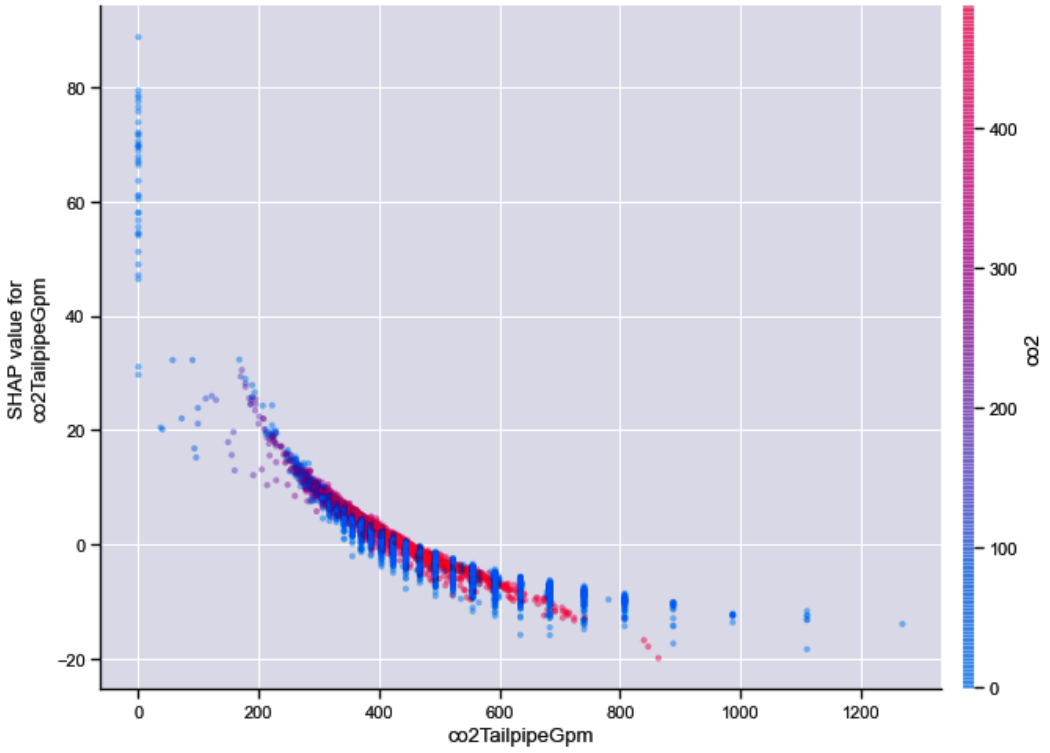
spearman cylinders→co2TailpipeGpm corr: 0.787 p-val: 0.0000



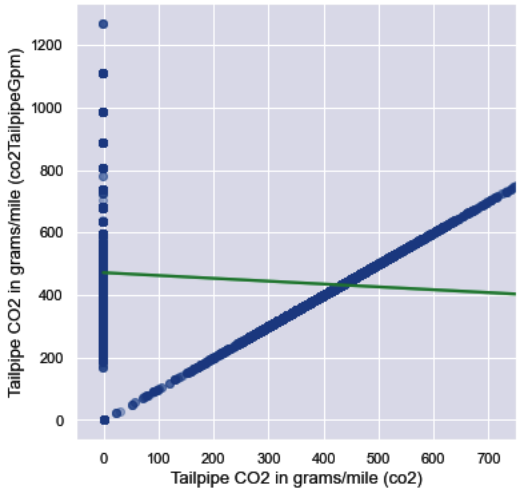

```
spearman    ghgScore→cylinders      corr: -0.117    p-val:
0.0000
```

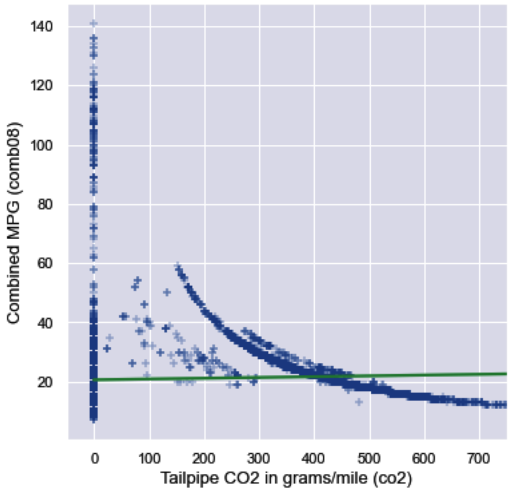


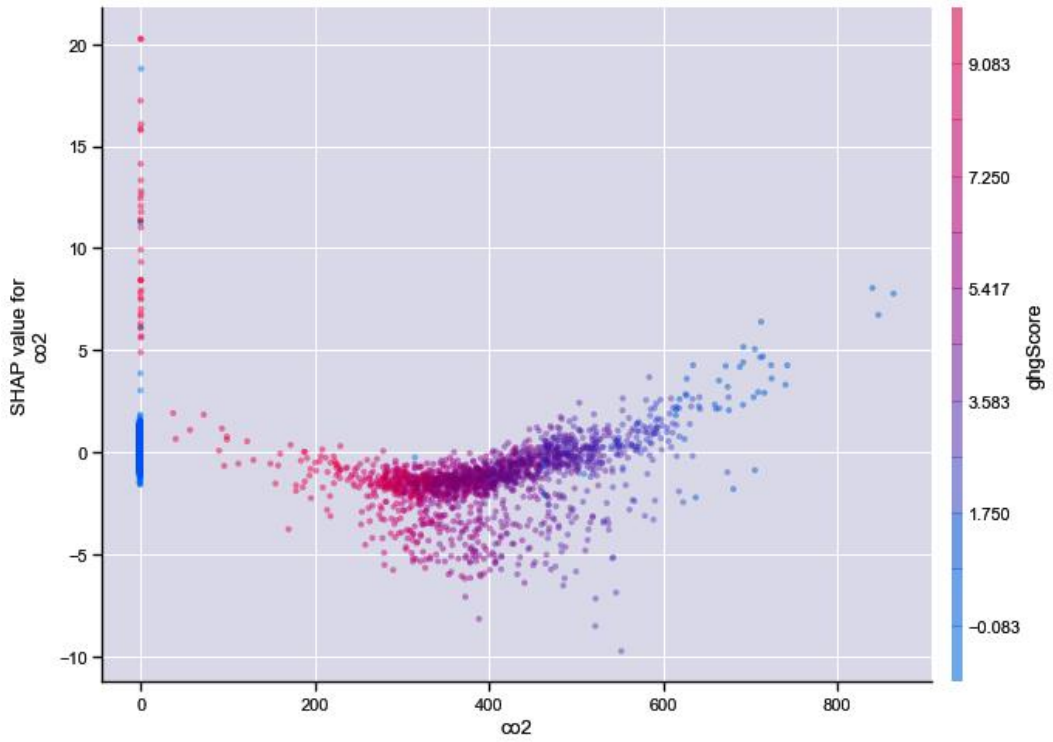




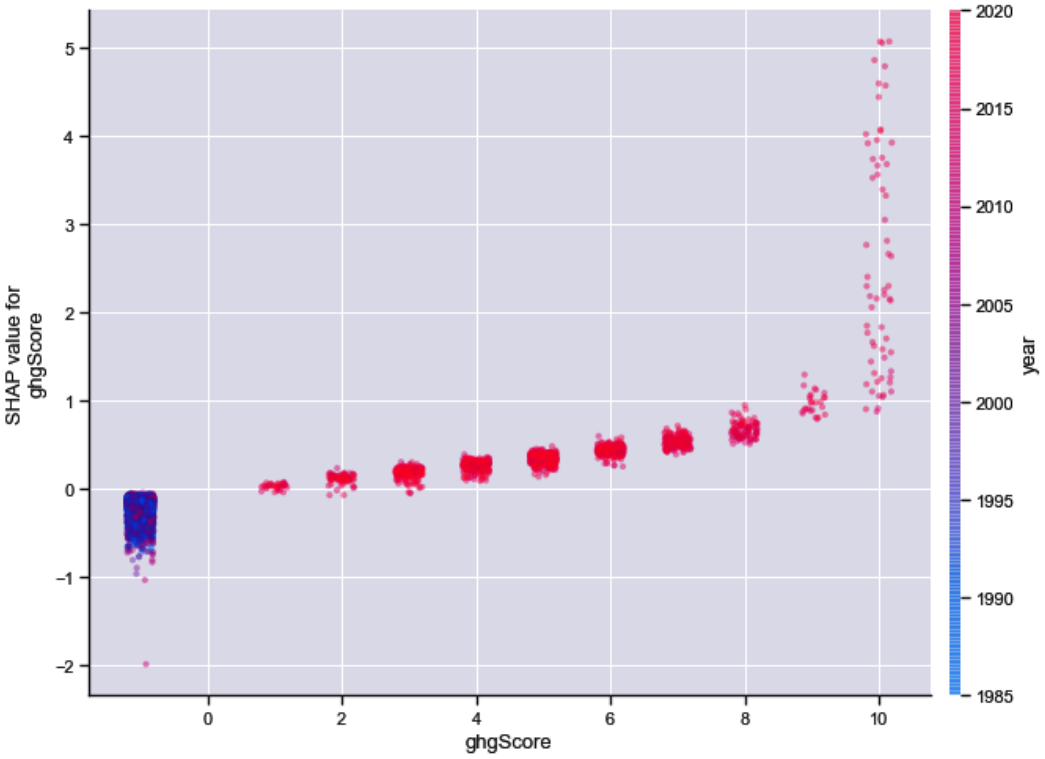
spearman co2→co2TailpipeGpm corr: -0.222 p-val: 0.0000



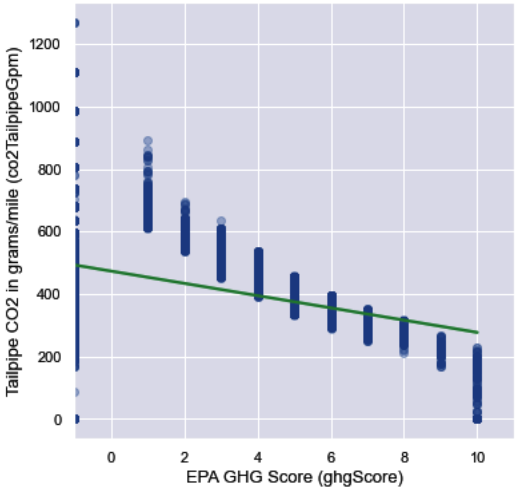


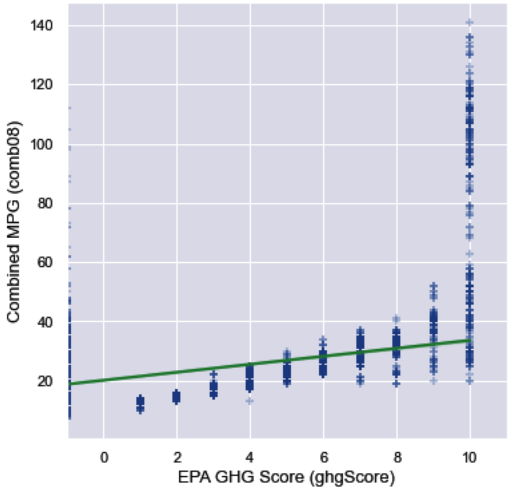


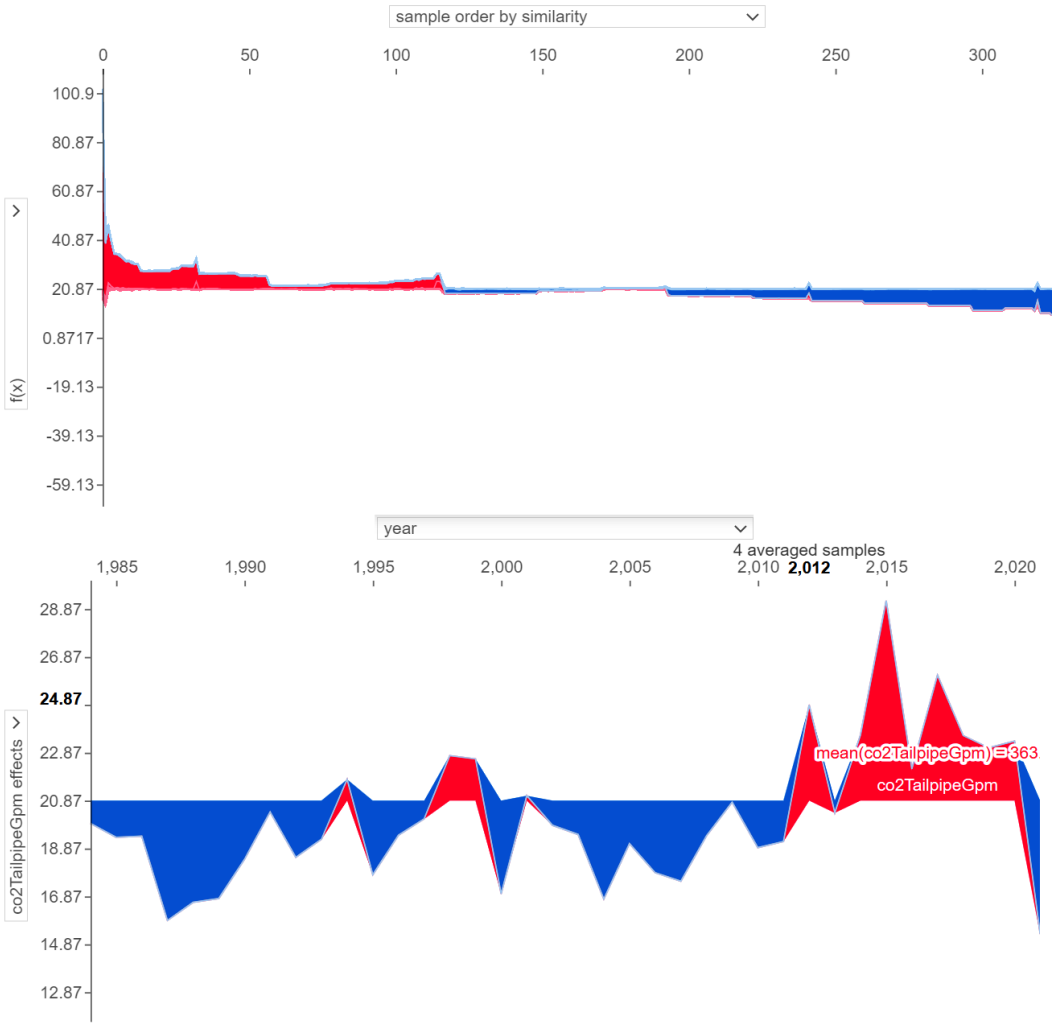
spearman ghgScore→co2 corr: 0.942 p-val: 0.0000

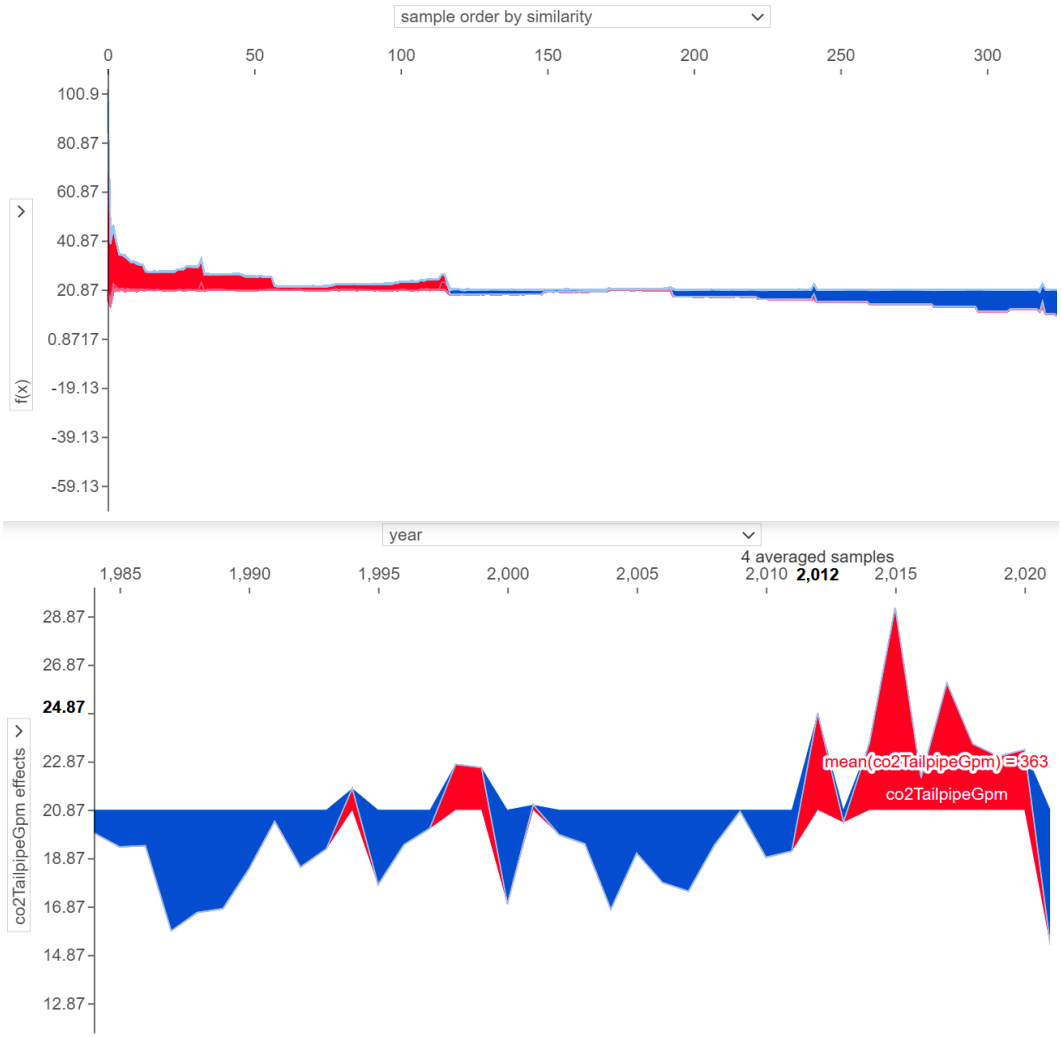


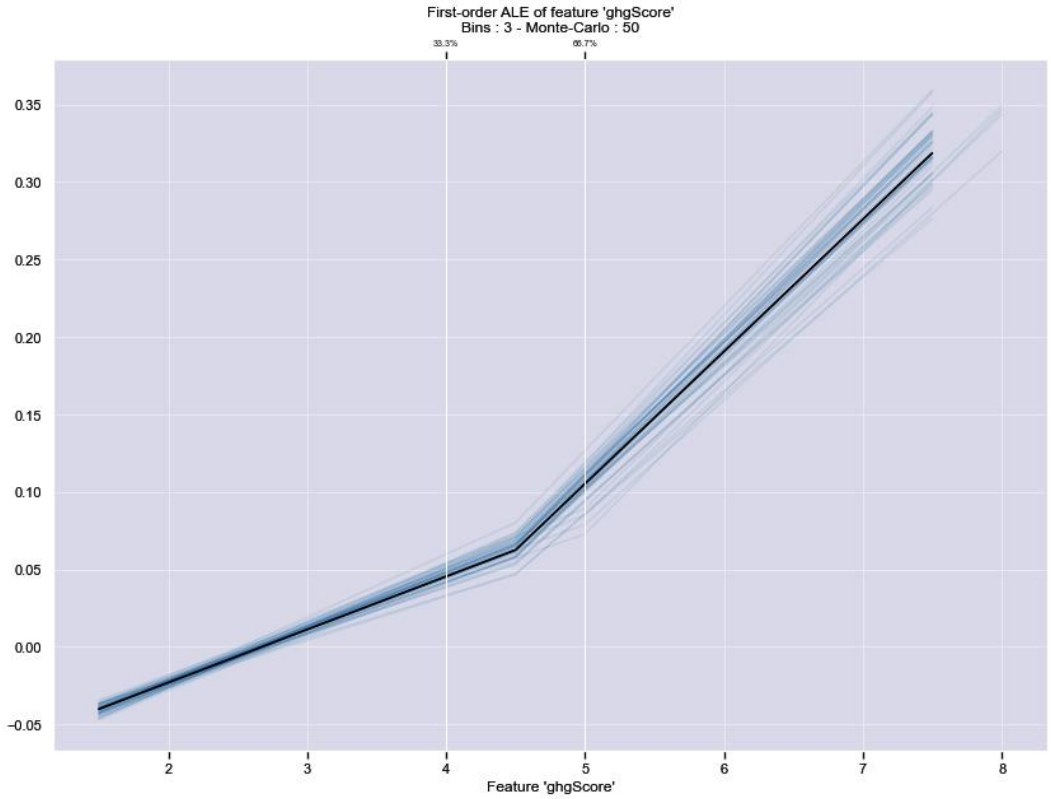
spearman ghgScore→year corr: 0.744 p-val: 0.0000

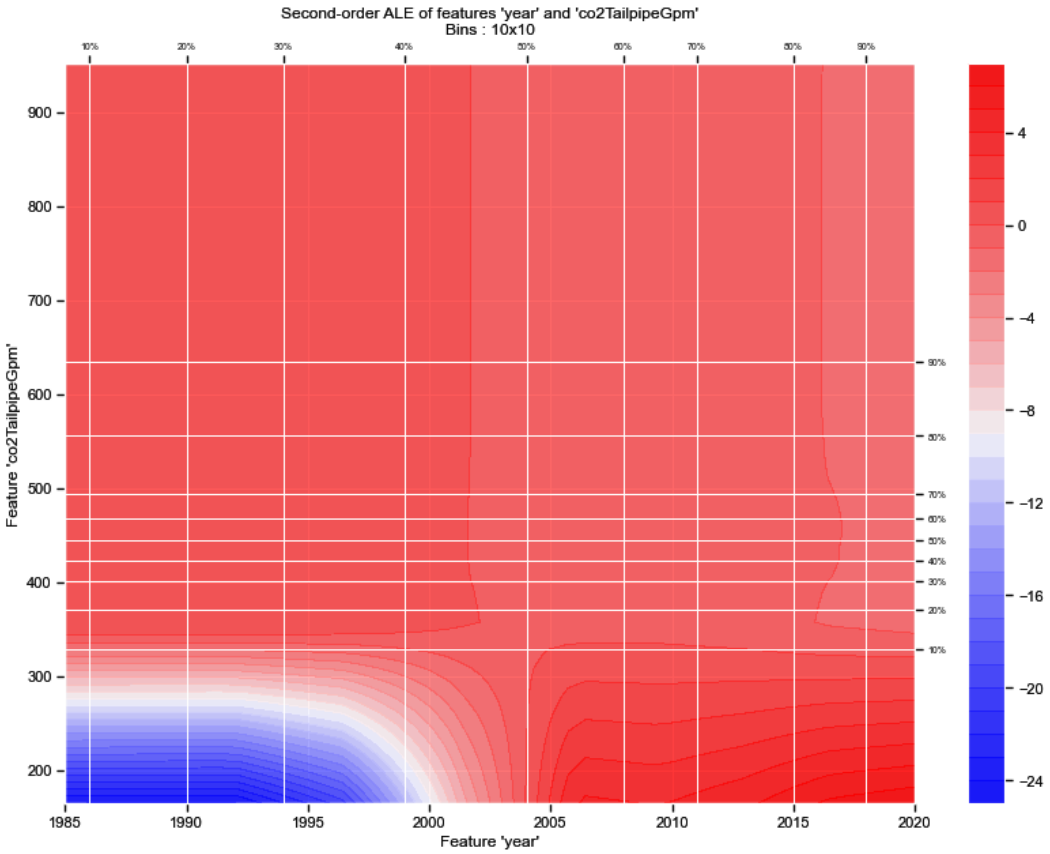


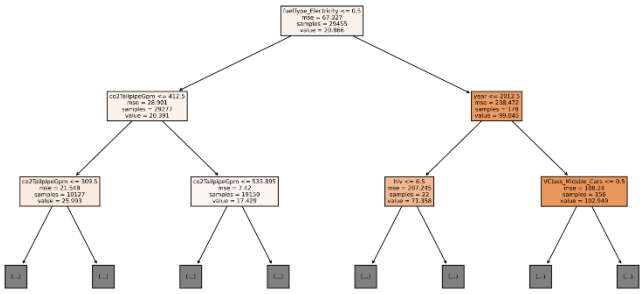



sample order by similarity
0
50
100
150
200
250
300
100.9
80.87
60.87
40.87
20.87
0.8717
-19.13
-39.13
-59.13
f(x)
year
4 averaged samples
1,985
1,990
1,995
2,000
2,005
2,010
2,012
2,015
2,020
28.87
26.87
24.87
22.87
20.87
18.87
16.87
14.87
12.87
co2TailpipeGpm effects
mean(co2TailpipeGpm) = 363
co2TailpipeGpm


sample order by similarity
0
50
100
150
200
250
300
100.9
80.87
60.87
40.87
20.87
0.8717
-19.13
-39.13
-59.13
f(x)
year
4 averaged samples
1,985
1,990
1,995
2,000
2,005
2,010
2,012
2,015
2,020
28.87
26.87
24.87
22.87
20.87
18.87
16.87
14.87
12.87
co2TailpipeGpm effects
mean(co2TailpipeGpm) = 363
co2TailpipeGpm


First-order ALE of feature 'ghgScore'
Bins : 3 - Monte-Carlo : 50
33.3%
66.7%
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
2
3
4
5
6
7
8
Feature 'ghgScore'


Second-order ALE of features 'year' and 'co2TailpipeGpm'
Bins : 10x10
Feature 'co2TailpipeGpm'
Feature 'year'
900
800
700
600
500
400
300
200
1985
1990
1995
2000
2005
2010
2015
2020
4
0
−4
−8
−12
−16
−20
−24

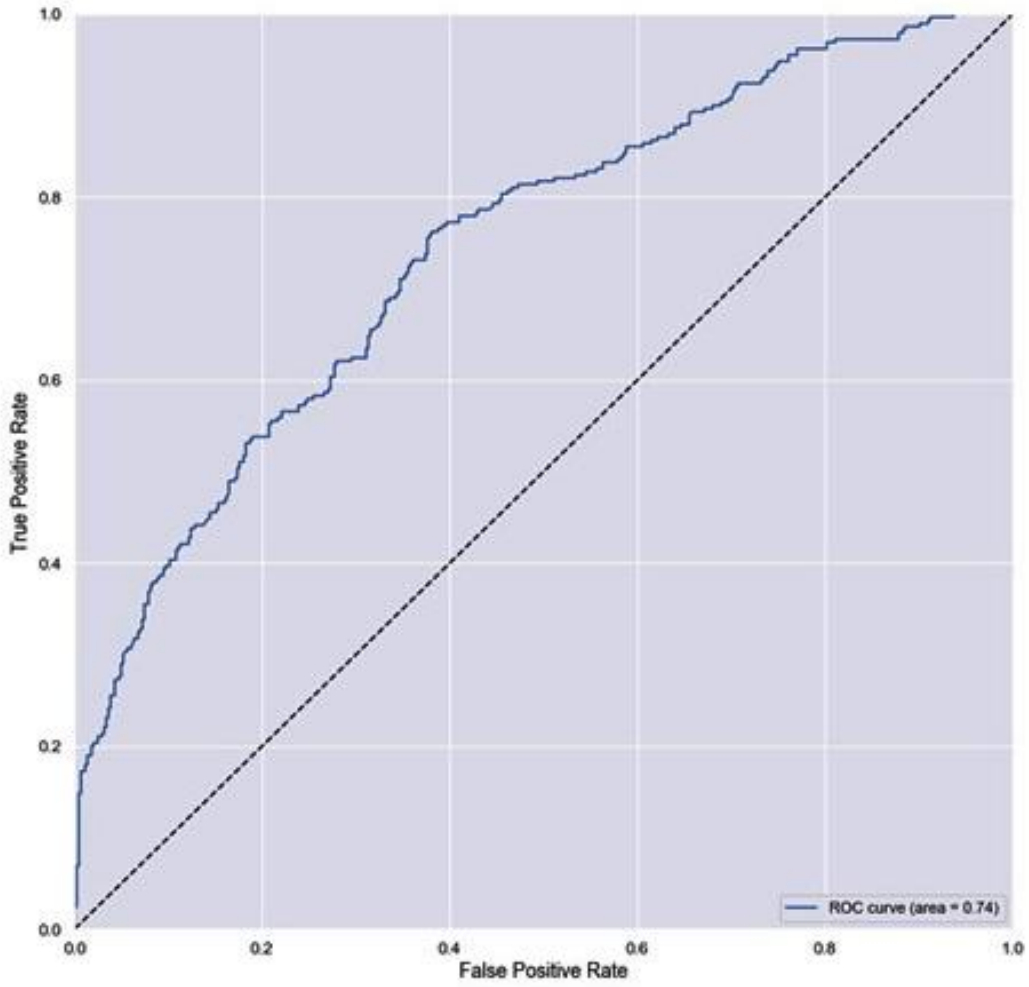


Chapter 6: Local Model-Agnostic Interpretation Methods

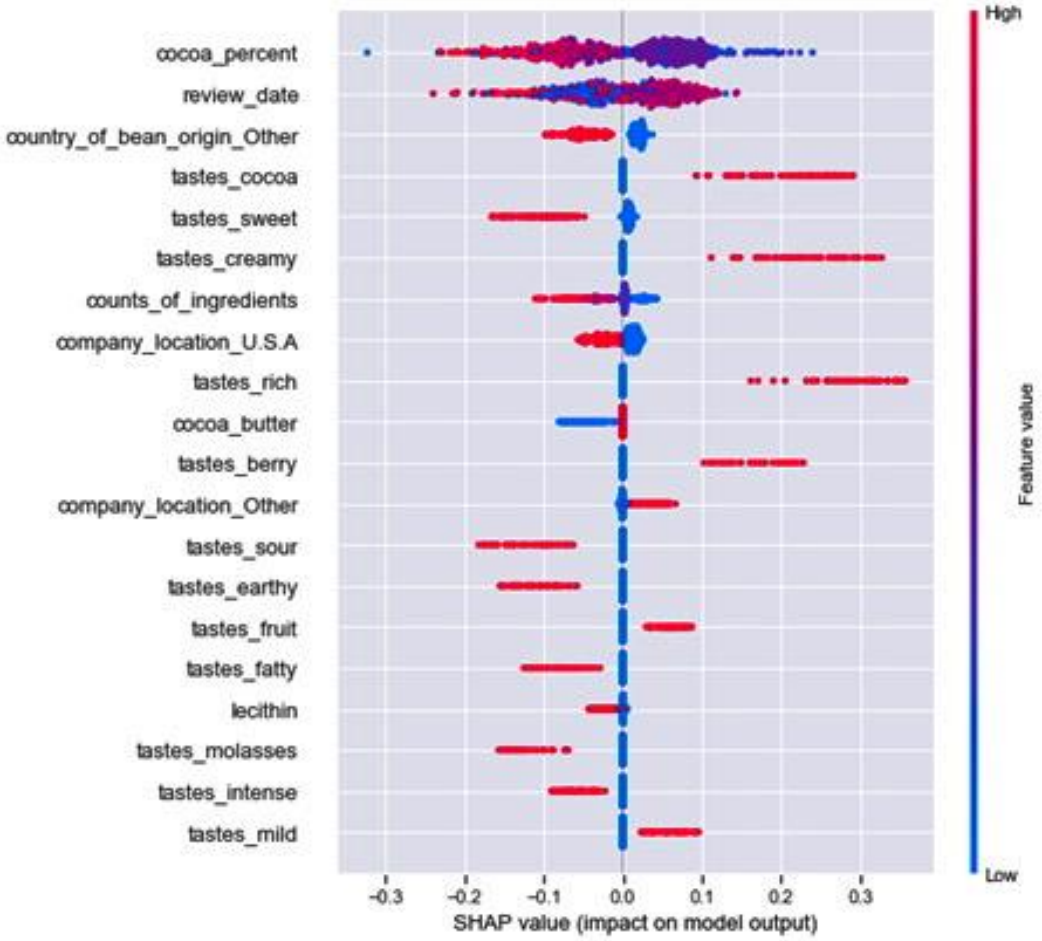
	company	company_location	review_date	country_of_bean_origin	cocoa_percent	rating	counts_of_ingredients	beans	cocoa_butter	va
0	5150	U.S.A	2019	Madagascar	76.000000	3.750000	3	have_bean	have_cocoa_butter	have_not_va
1	5150	U.S.A	2019	Dominican republic	76.000000	3.500000	3	have_bean	have_cocoa_butter	have_not_va
2	5150	U.S.A	2019	Tanzania	76.000000	3.250000	3	have_bean	have_cocoa_butter	have_not_va
3	A. Morin	France	2012	Peru	63.000000	3.750000	4	have_bean	have_cocoa_butter	have_not_va
4	A. Morin	France	2012	Bolivia	70.000000	3.500000	4	have_bean	have_cocoa_butter	have_not_va
:	:	:	:	:	:	:	:	:	:	
2219	Zotter	Austria	2014	Blend	80.000000	2.750000	4	have_bean	have_cocoa_butter	have_not_va
2220	Zotter	Austria	2017	Colombia	75.000000	3.750000	3	have_bean	have_cocoa_butter	have_not_va
2221	Zotter	Austria	2018	Belize	72.000000	3.500000	3	have_bean	have_cocoa_butter	have_not_va
2222	Zotter	Austria	2018	Congo	70.000000	3.250000	3	have_bean	have_cocoa_butter	have_not_va
2223	Zotter	Austria	2018	Blend	75.000000	3.000000	3	have_bean	have_cocoa_butter	have_not_va

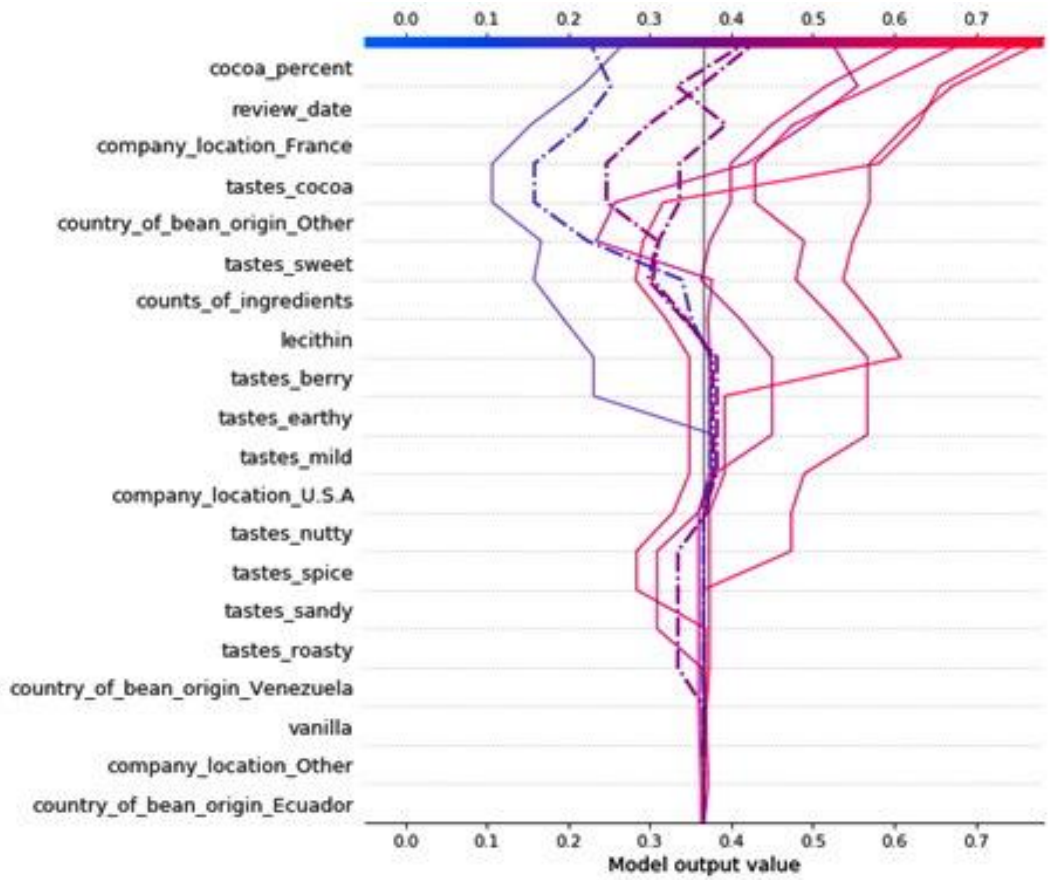
2224 rows × 14 columns

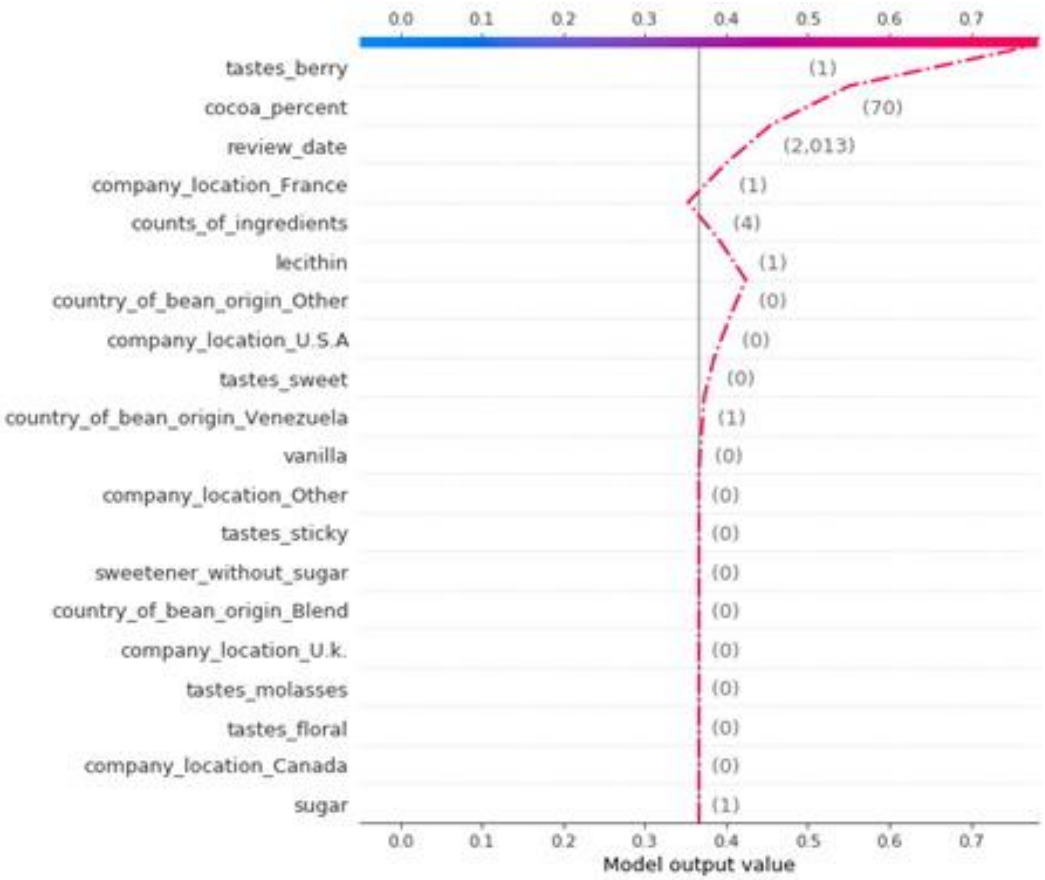
	first_taste	second_taste	third_taste	fourth_taste
80	oily	vegetal	nutty	cocoa
81	oily	vanilla	melon	cocoa
82	rich	sour	mild smoke	nan
83	fruity	sour	nan	nan
84	high roast	high astringnet	nan	nan
85	smokey	savory	nan	nan
86	sandy	roasty	nutty	nan
87	roasty	brownie	nutty	nan
88	red wine	rich cocoa	long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



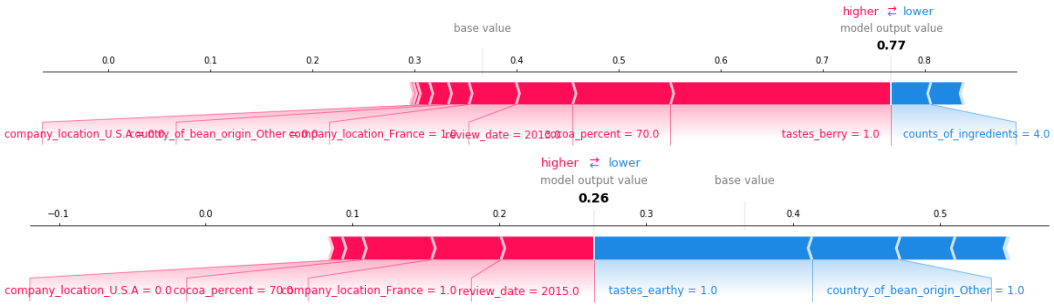

```
Accuracy_train:  0.7315    Accuracy_test:   0.6962
Precision_test:  0.6772    Recall_test:     0.4414
ROC-AUC_test:    0.7449    F1_test:         0.5344    MCC_test:
0.3399
```

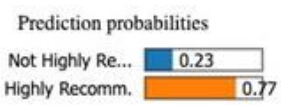

cocoa_percent
review_date
country_of_bean_origin_Other
tastes_cocoa
tastes_sweet
tastes_creamy
counts_of_ingredients
company_location_U.S.A
tastes_rich
cocoa_butter
tastes_berry
company_location_Other
tastes_sour
tastes_earthy
tastes_fruit
tastes_fatty
lecithin
tastes_molasses
tastes_intense
tastes_mild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

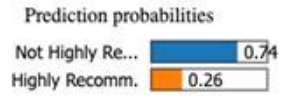

0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
cocoa_percent
review_date
company_location_France
tastes_cocoa
country_of_bean_origin_Other
tastes_sweet
counts_of_ingredients
lecithin
tastes_berry
tastes_earthy
tastes_mild
company_location_U.S.A
tastes_nutty
tastes_spice
tastes_sandy
tastes_roasty
country_of_bean_origin_Venezuela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

	5	24
rating	4.00	2.75
y	1.00	0.00
y_pred	1.00	0.00
review_date	2013.00	2015.00
cocoa_percent	70.00	70.00
counts_of_ingredients	4.00	4.00
cocoa_butter	1.00	1.00
vanilla	0.00	0.00
lecithin	1.00	1.00
salt	0.00	0.00
sugar	1.00	1.00
sweetener_without_sugar	0.00	0.00
company_location_Canada	0.00	0.00
company_location_France	1.00	1.00
:	:	:
country_of_bean_origin_Other	0.00	1.00
country_of_bean_origin_Peru	0.00	0.00
country_of_bean_origin_Venezuela	1.00	0.00
:	:	:
tastes_earthy	0.00	1.00
:	:	:
tastes_berry	1.00	0.00
tastes_vanilla	0.00	0.00
tastes_creamy	0.00	0.00

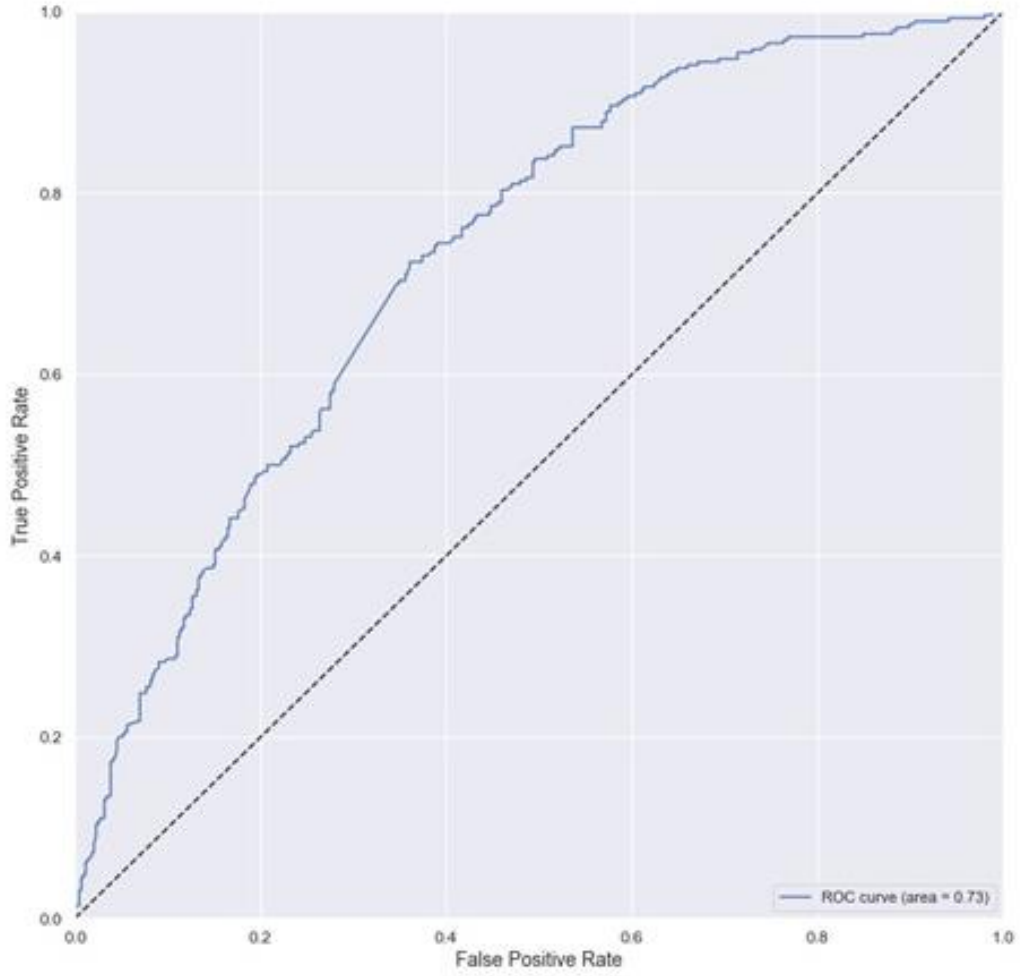




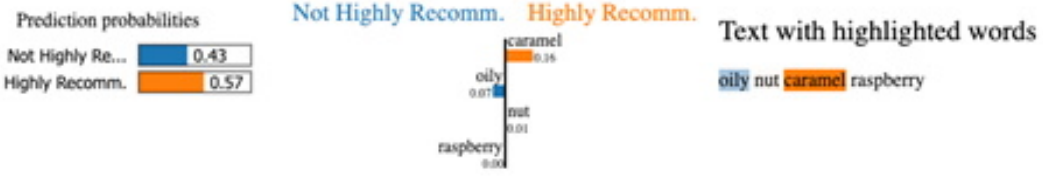


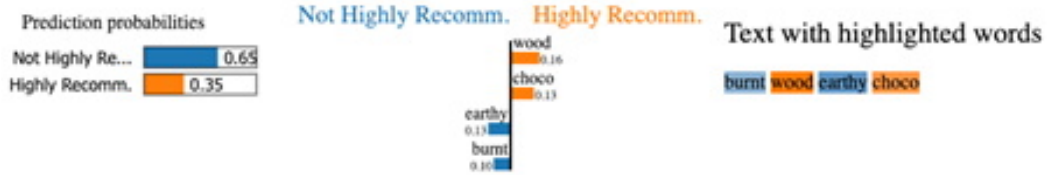
	taste	tf-idf
305	raspberry	0.59
259	nut	0.49
265	oily	0.46
64	caramel	0.45
274	papaya	0.00
:	:	:
135	edge	0.00
134	easy	0.00
133	easter	0.00
415	yogurt	0.00

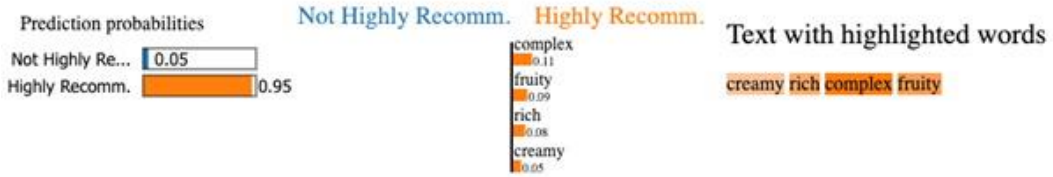
416 rows × 2 columns

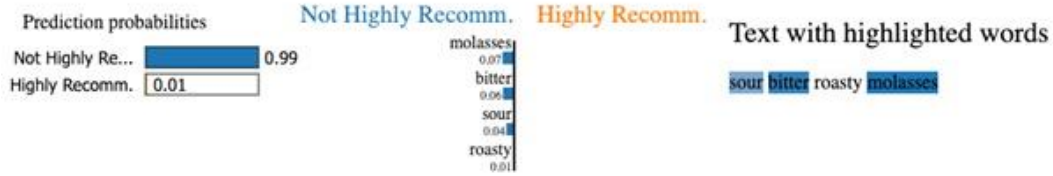


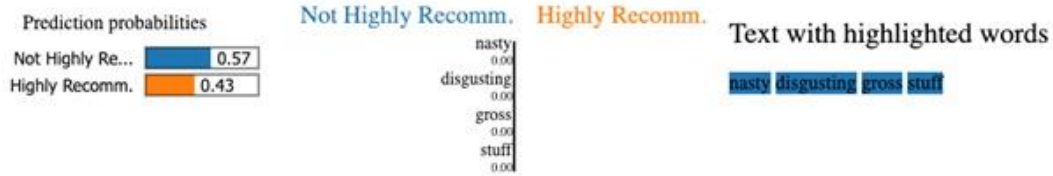

```
Accuracy_train:  0.7953    Accuracy_test:   0.6798
Precision_test:  0.6233    Recall_test:     0.4793
ROC-AUC_test:    0.7328    F1_test:         0.5419    MCC_test:
0.3084
```

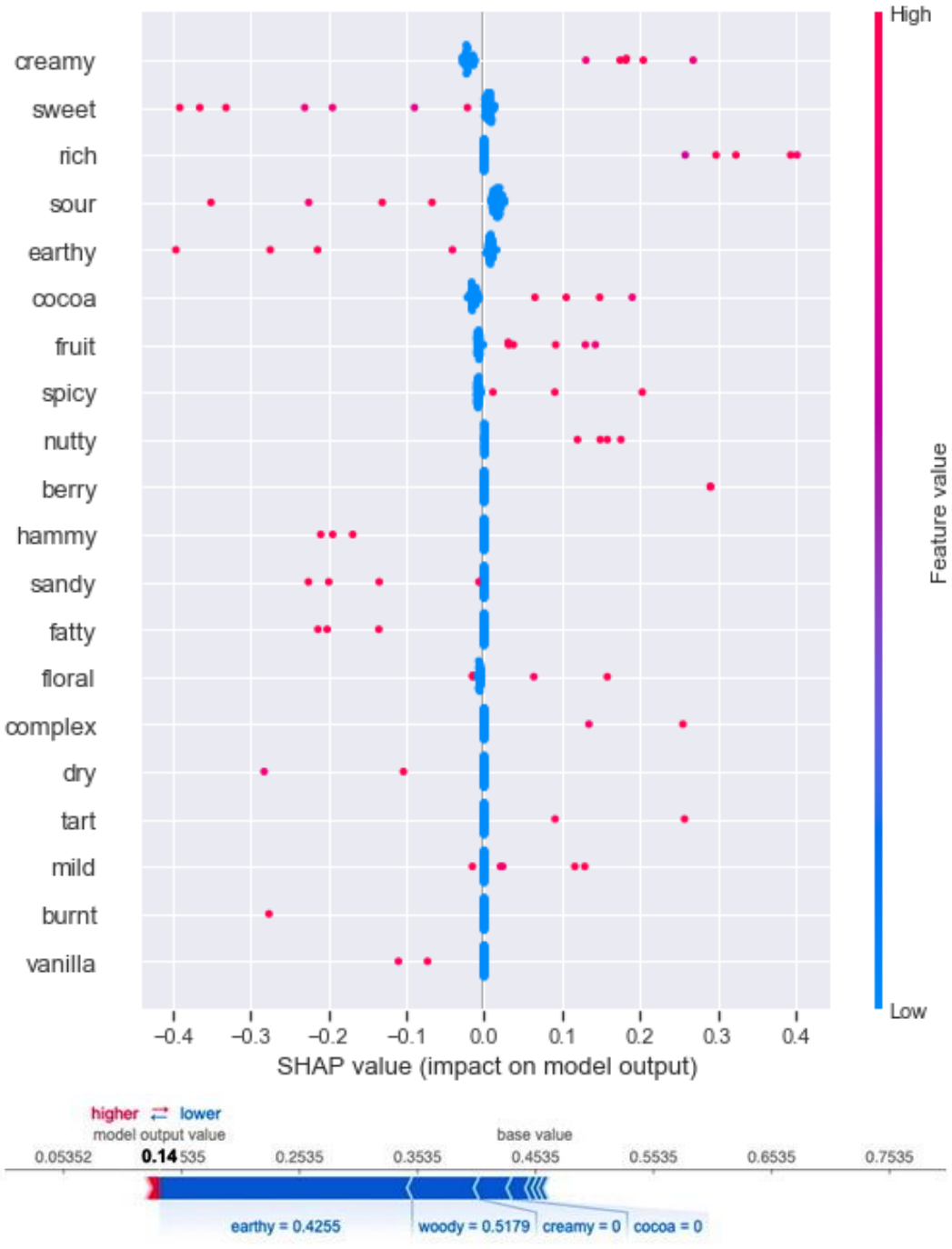



Prediction probabilities
Not Highly Re...
0.65
Highly Recomm.
0.35
Not Highly Recomm.
Highly Recomm.
wood
0.16
choco
0.13
earthy
0.13
burnt
0.10
Text with highlighted words
burnt wood earthy choco


Prediction probabilities
Not Highly Re...
0.05
Highly Recomm.
0.95
Not Highly Recomm.
Highly Recomm.
complex
0.11
fruity
0.09
rich
0.08
creamy
0.05
Text with highlighted words
creamy rich complex fruity

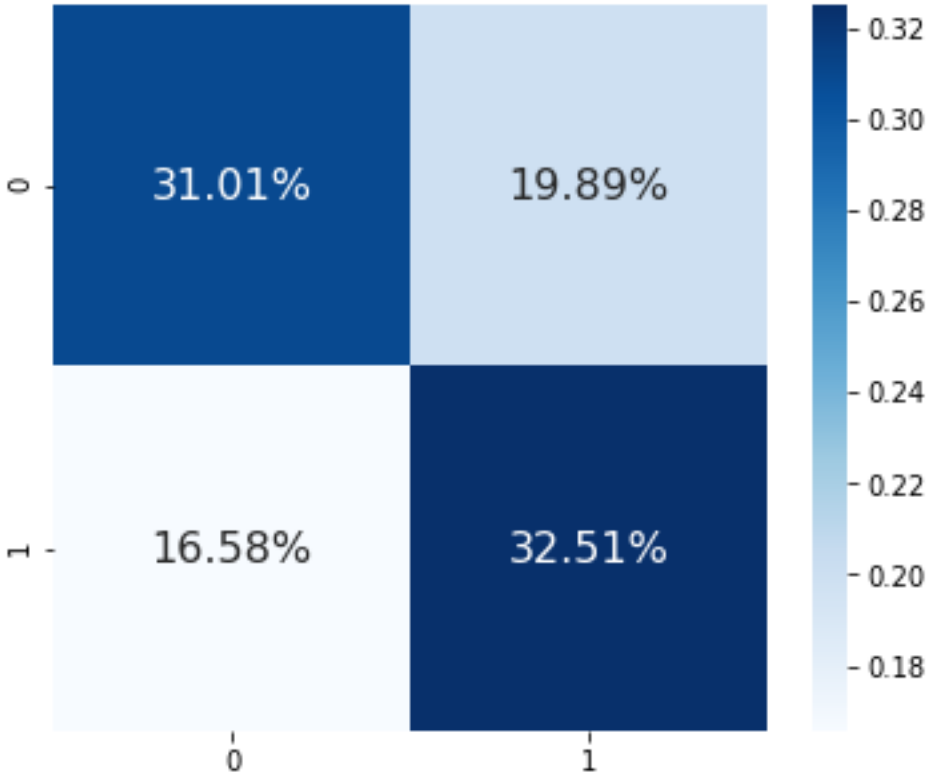

Prediction probabilities
Not Highly Re...
0.99
Highly Recomm.
0.01
Not Highly Recomm.
Highly Recomm.
molasses
0.07
bitter
0.06
sour
0.04
roasty
0.01
Text with highlighted words
sour bitter roasty molasses

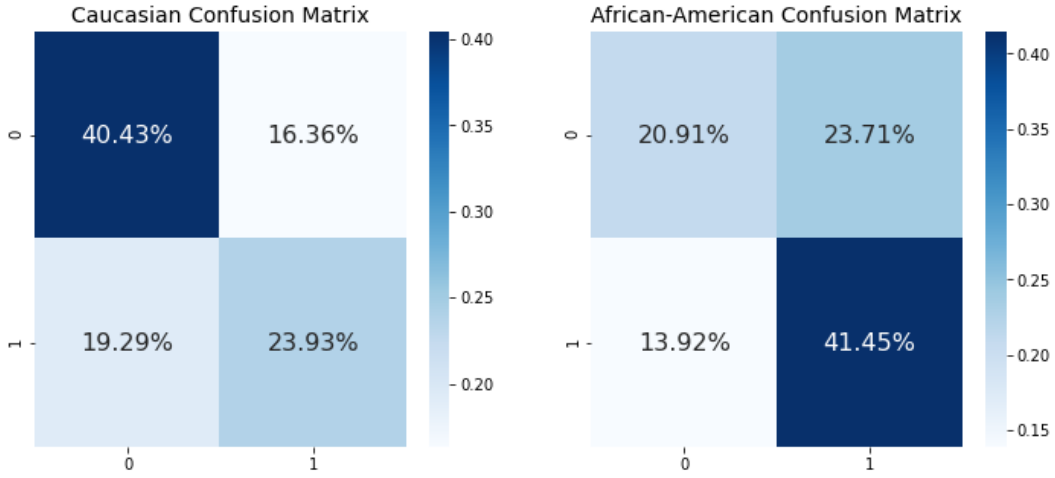

Prediction probabilities
Not Highly Re...
0.57
Highly Recomm.
0.43
Not Highly Recomm.
Highly Recomm.
nasty
0.00
disgusting
0.00
gross
0.00
stuff
0.00
Text with highlighted words
nasty disgusting gross stuff


creamy
sweet
rich
sour
earthy
cocoa
fruit
spicy
nutty
berry
hammy
sandy
fatty
floral
complex
dry
tart
mild
burnt
vanilla
−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
higher ⇄ lower
model output value
base value
0.05352
0.14535
0.2535
0.3535
0.4535
0.5535
0.6535
0.7535
earthy = 0.4255
woody = 0.5179
creamy = 0
cocoa = 0

Chapter 7: Anchor and Counterfactual Explanations

	Feature	x	PN	PN-x	PP
0	age	23	23.000000	0.000000	0.000000
4	priors_count	2	2.000000	0.000000	0.000000
5	sex_Female	0	0.397589	0.397589	0.000000
6	sex_Male	1	1.000000	0.000000	0.000000
7	race_African-American	1	0.457206	-0.542794	0.000000
16	c_charge_degree_(F7)	1	1.000000	0.000000	0.000000

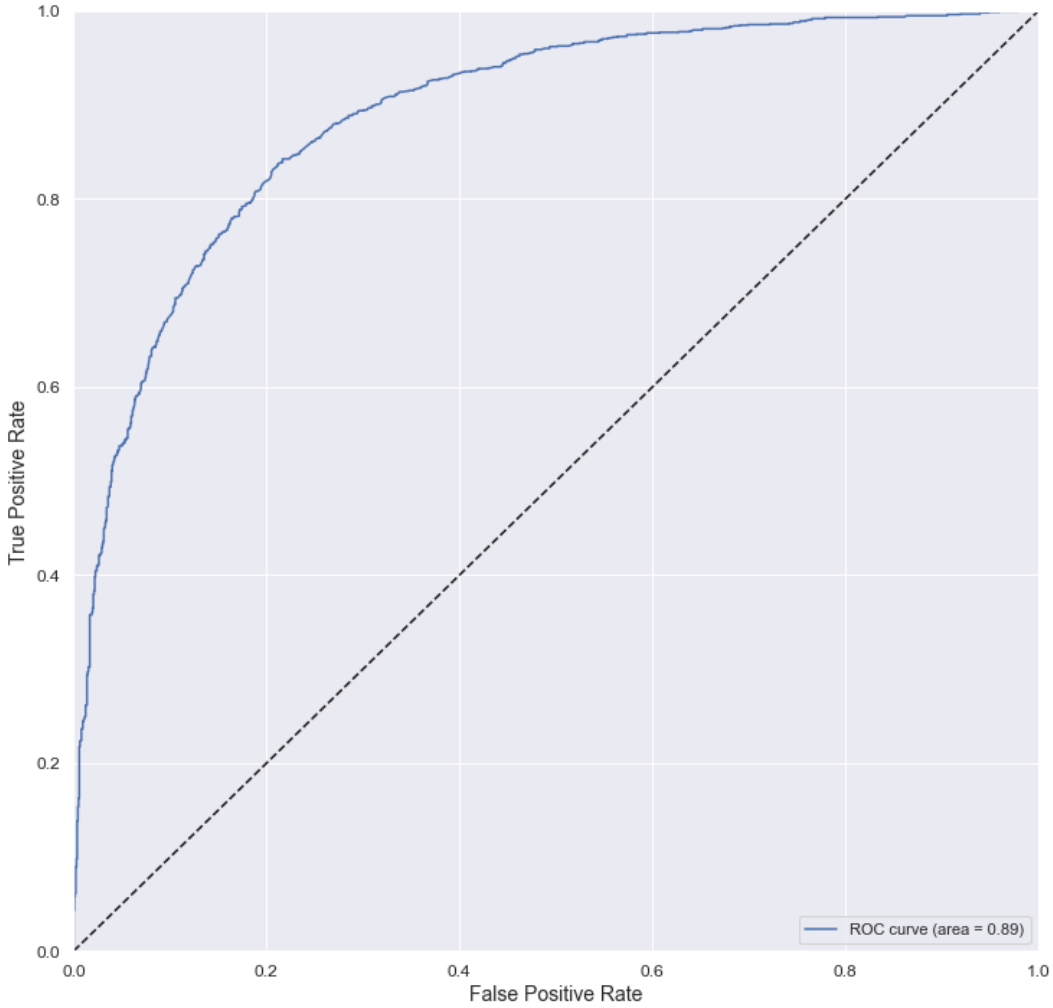




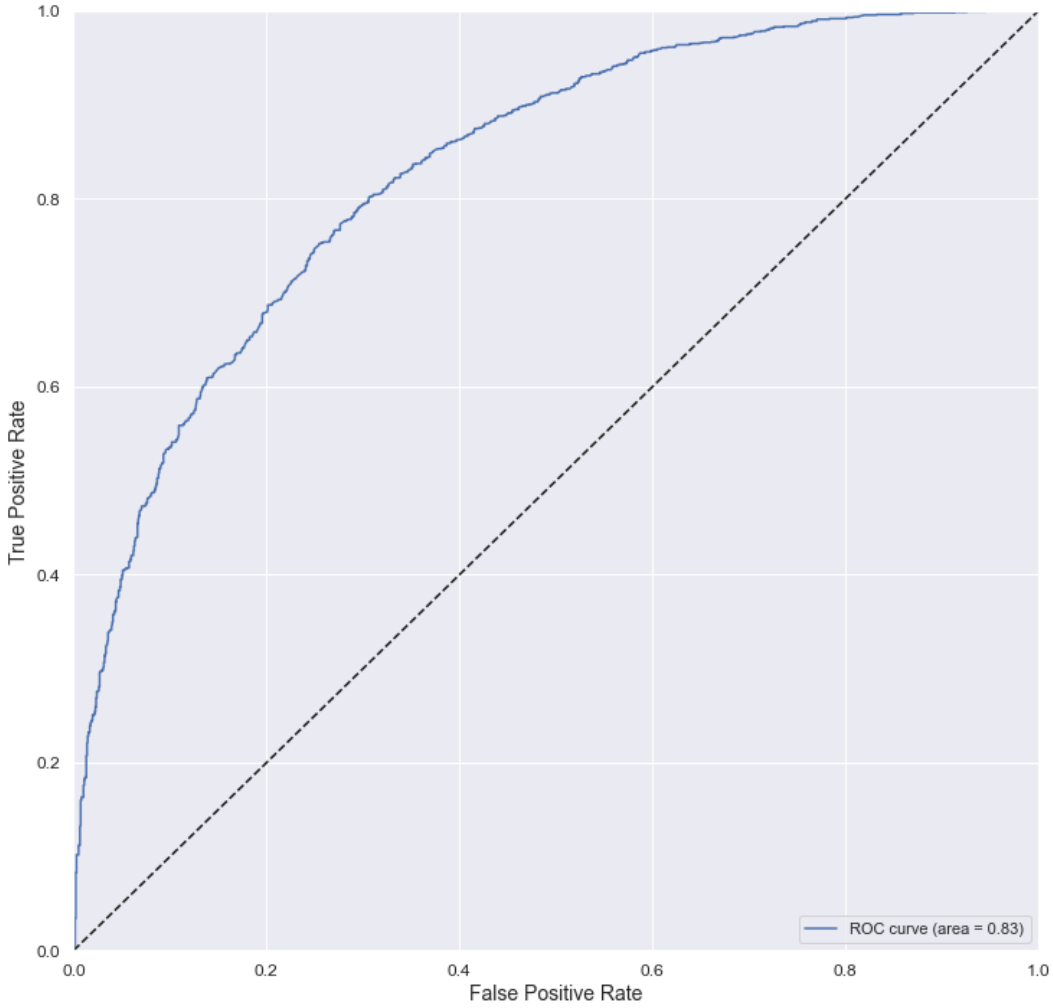
African-American FPR: 53.1%

Caucasian FPR: 28.8%

Ratio FPRs: 1.84 x

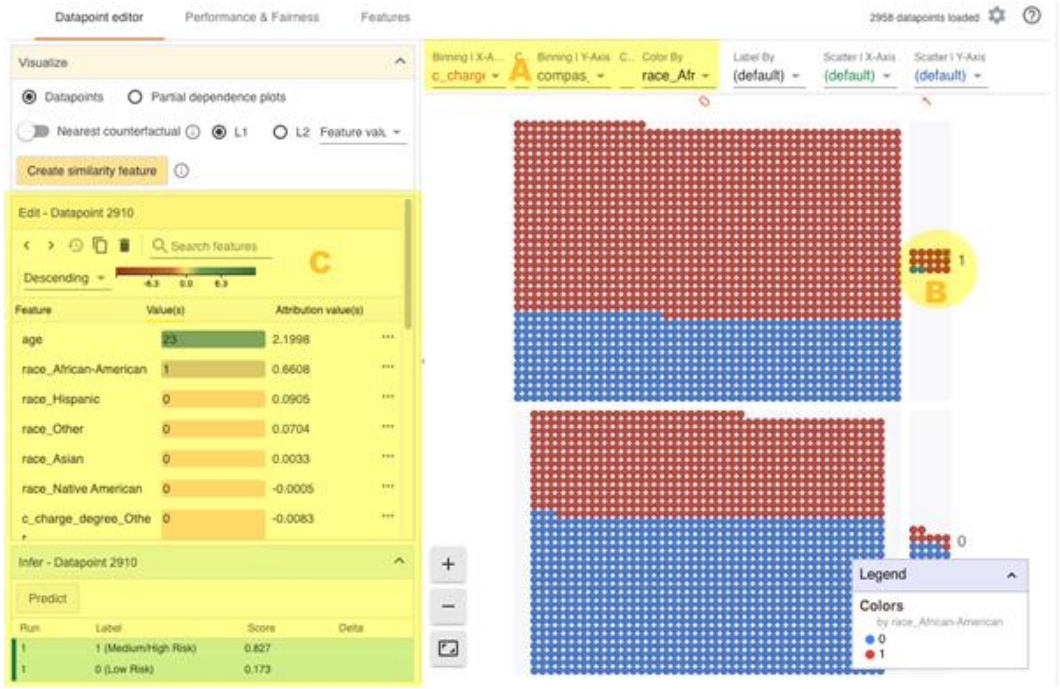


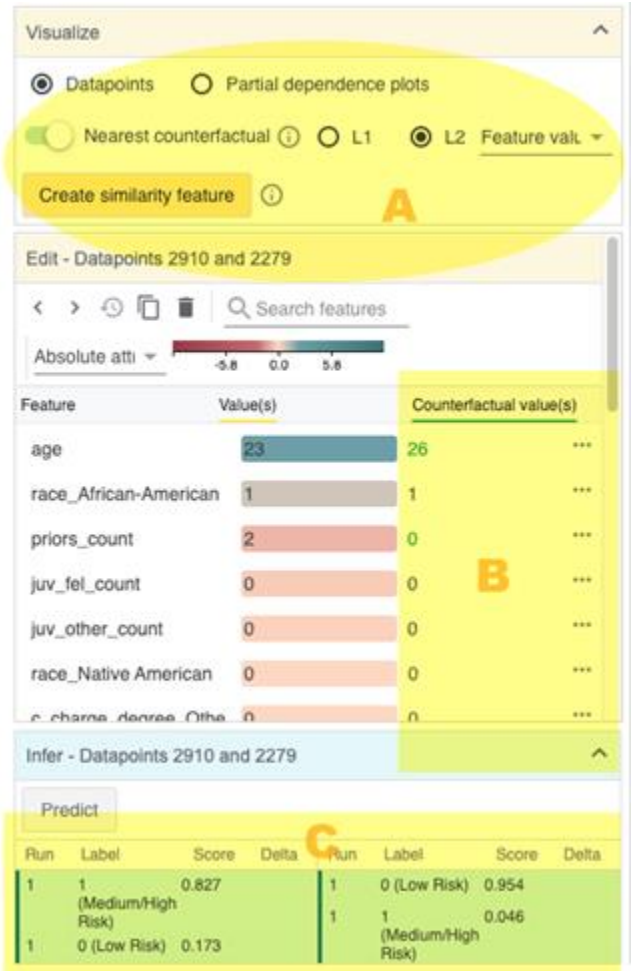

```
Accuracy_train:  0.8790    Accuracy_test:   0.8087

Precision_test:  0.8277    Recall_test:     0.8110

ROC-AUC_test:    0.8927    F1_test:         0.8193    MCC_test:
0.6162
```

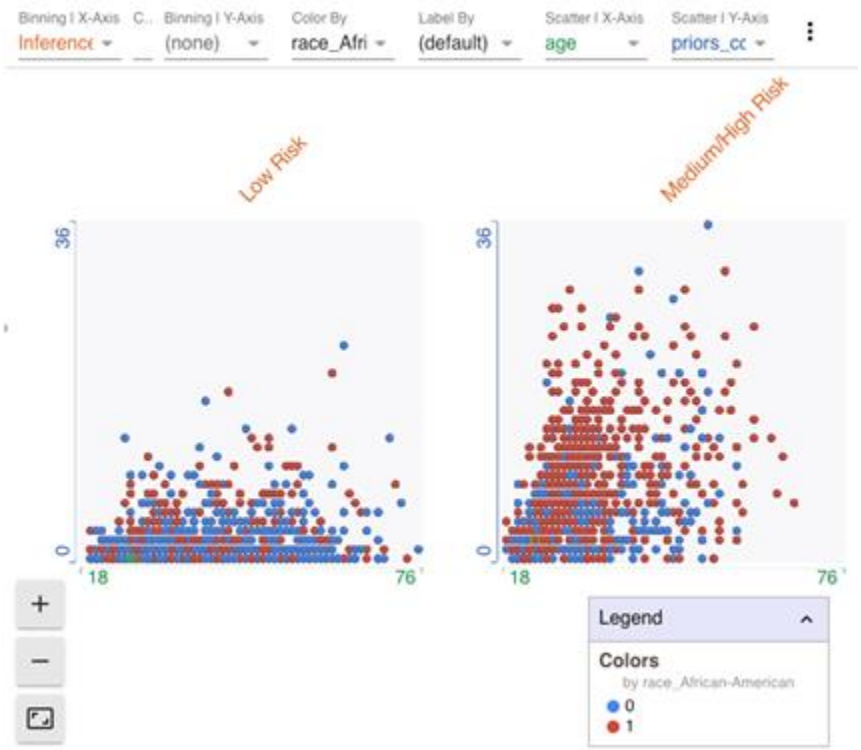


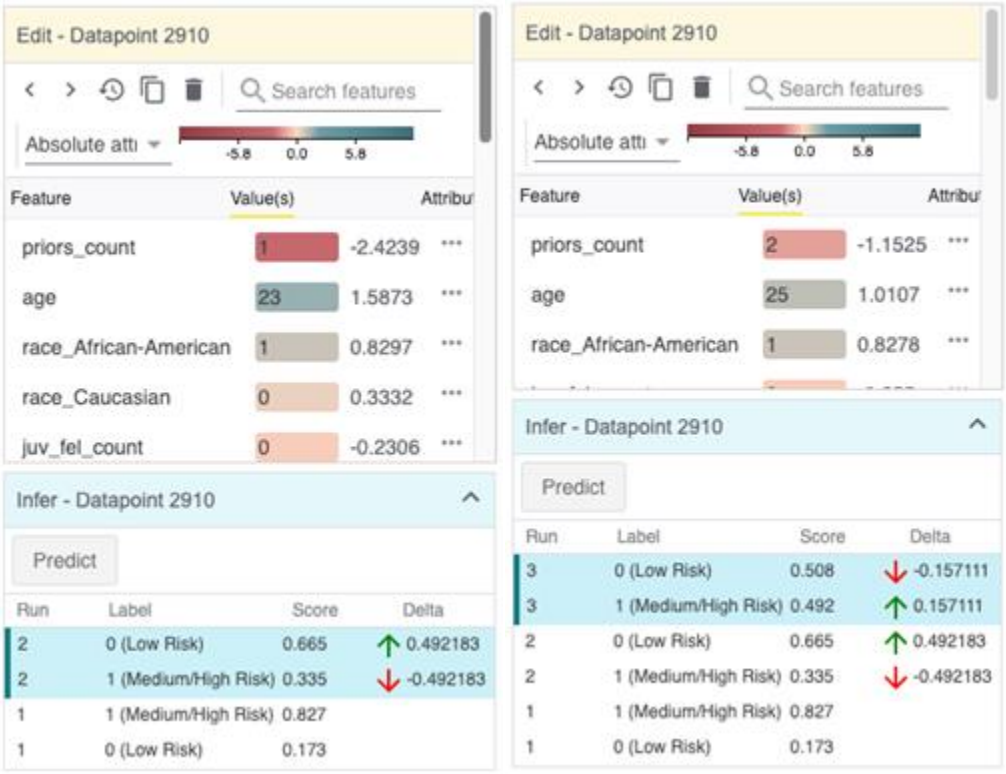
Accuracy_train:	0.7527	Accuracy_test:	0.7471	
Precision_test:	0.7653	Recall_test:	0.7604	
ROC-AUC_test:	0.8320	F1_test:	0.7628	MCC_test: 0.4920

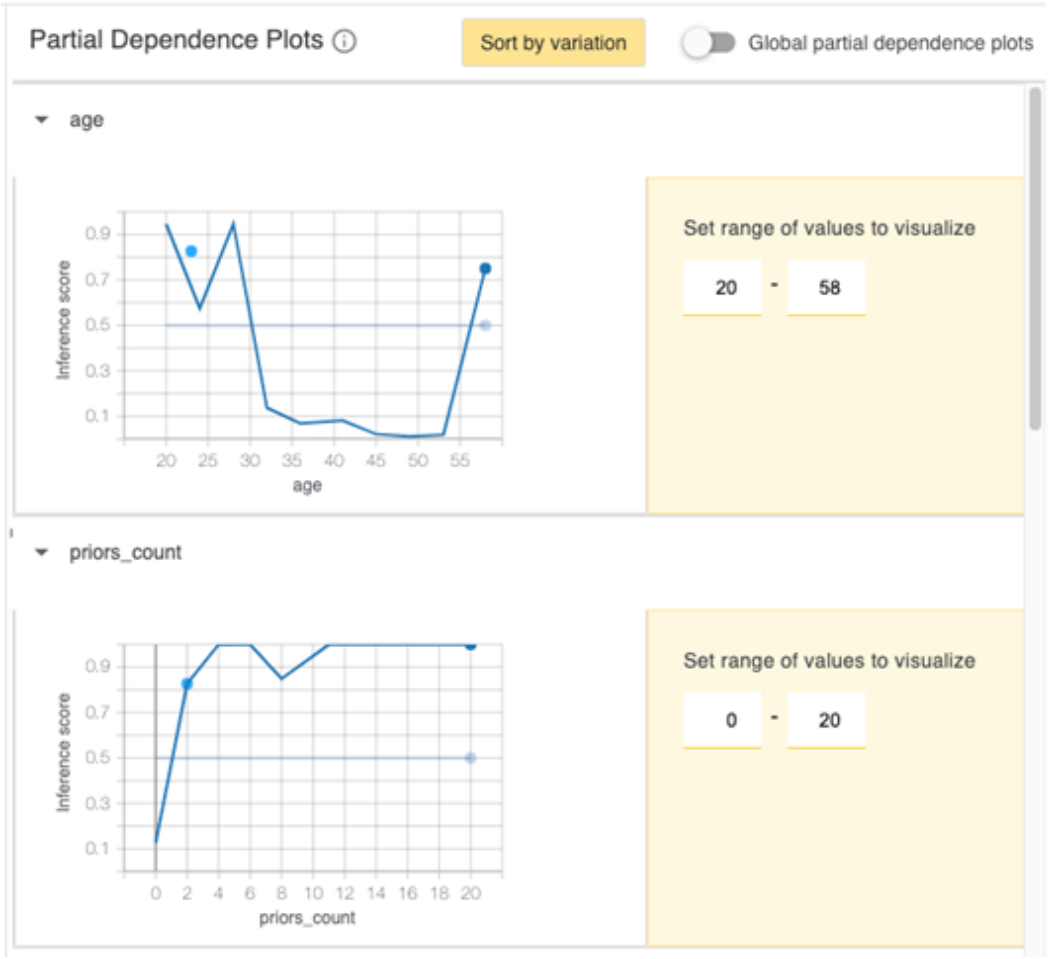
	10127	2726	5231
y	0	0	1
y_pred	0	0	1
age	24	23	23
:	:	:	:
priors_count	2	2	2
sex_Female	0	0	0
sex_Male	1	1	1
race_African-American	0	0	1
race_Asian	0	0	0
race_Caucasian	1	0	0
race_Hispanic	0	1	0
:	:	:	:
c_charge_degree_(F3)	0	1	0
c_charge_degree_(F7)	0	0	1
c_charge_degree_(M1)	1	0	0
:	:	:	:

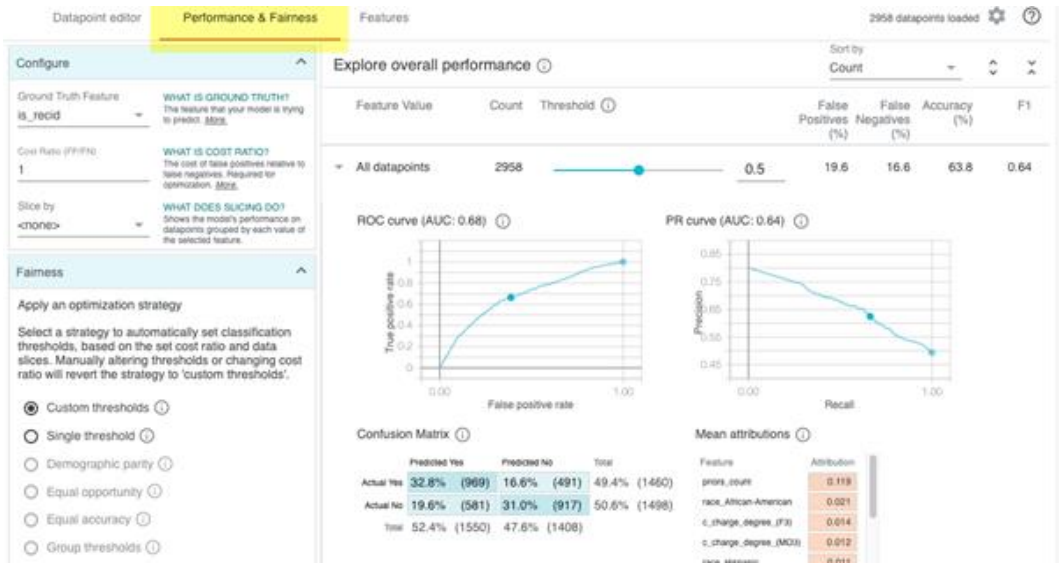

Datapoint editor
Performance & Fairness
Features
2958 datapoints loaded
Visualize
Datapoints
Partial dependence plots
Nearest counterfactual
L1
L2
Create similarity feature
Edit - Datapoint 2910
Search features
Descending
C
Feature
Value(s)
Attribution value(s)
age
23
2.1998
race_African-American
1
0.6608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

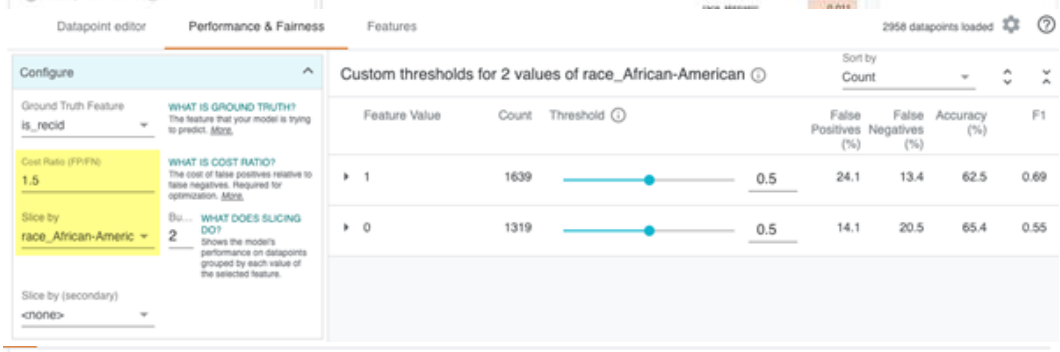

Visualize
Datapoints
Partial dependence plots
Nearest counterfactual
L1
L2
Feature valt
Create similarity feature
A
Edit - Datapoints 2910 and 2279
Search features
Absolute atti
-5.8
0.0
5.8
Feature
Value(s)
Counterfactual value(s)
age
23
26
race_African-American
1
1
priors_count
2
0
juv_fel_count
0
0
B
juv_other_count
0
0
race_Native American
0
0
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

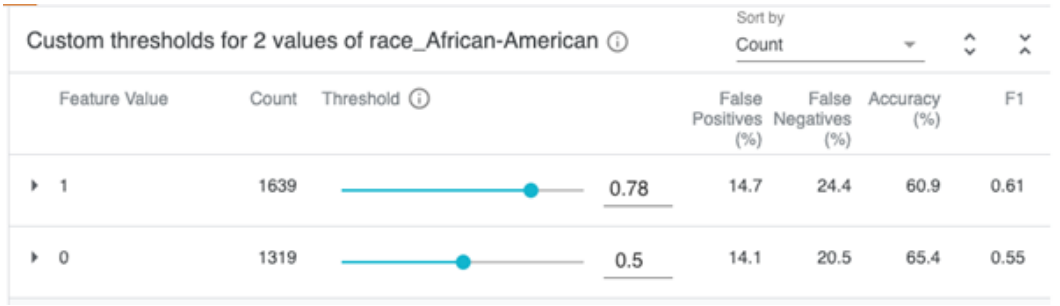

Binning | X-Axis
Inferenc
C...
Binning | Y-Axis
(none)
Color By
race_Afri
Label By
(default)
Scatter | X-Axis
age
Scatter | Y-Axis
priors_cc
Low Risk
Medium/High Risk
36
0
18
76
Legend
Colors
by race_African-American
0
1


Edit - Datapoint 2910
Search features
Absolute atti
-5.8
0.0
5.8
Feature
Value(s)
Attribu
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 atti
-5.8
0.0
5.8
Feature
Value(s)
Attribu
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

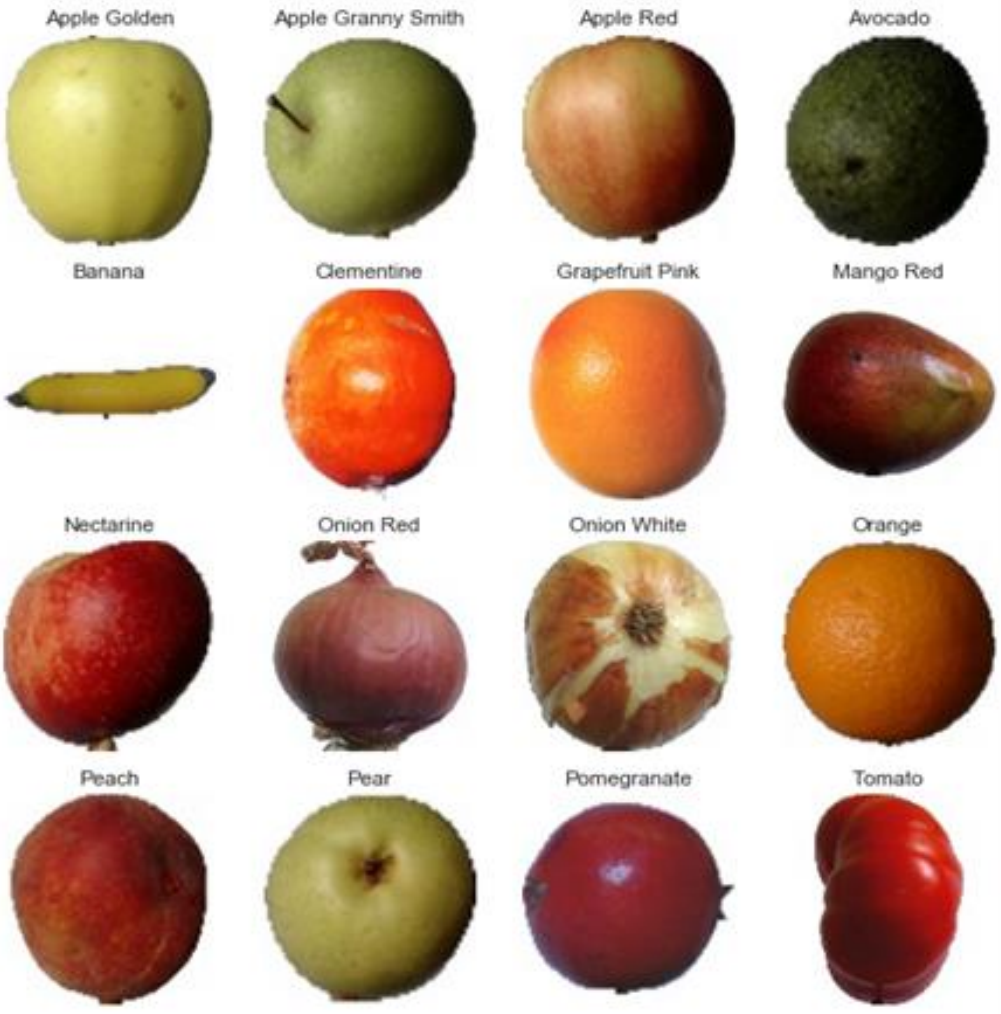

Partial Dependence Plots
Sort by variation
Global partial dependence plots
age
Inference score
0.9
0.7
0.5
0.3
0.1
20 25 30 35 40 45 50 55
age
Set range of values to visualize
20 - 58
priors_count
Inference score
0.9
0.7
0.5
0.3
0.1
0 2 4 6 8 10 12 14 16 18 20
priors_count
Set range of values to visualize
0 - 20

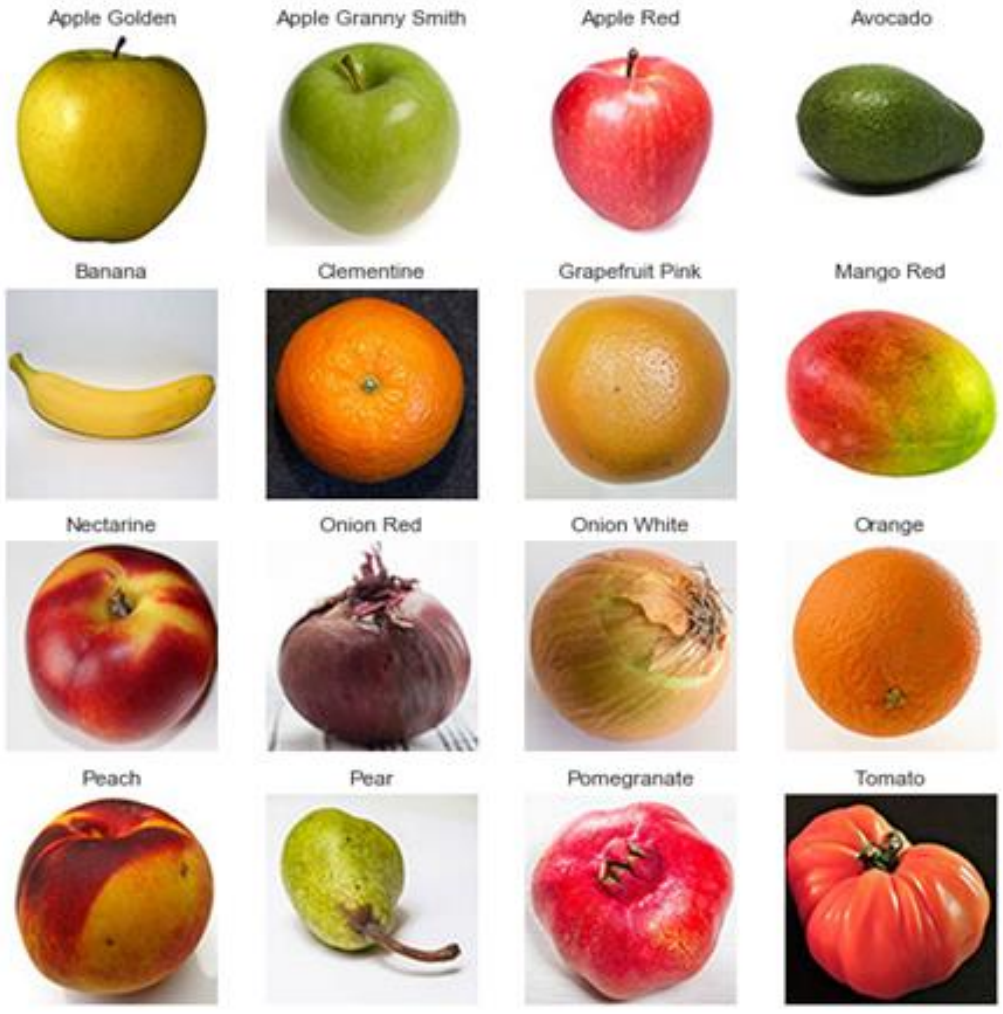

Datapoint editor
Performance & Fairness
Features
2958 datapoints loaded
Configure
Ground Truth Feature
is_recid
WHAT IS GROUND TRUTH?
The feature that your model is trying to predict. More.
Cost Ratio (FP/FN)
1
WHAT IS COST RATIO?
The cost of false positives relative to false negatives. Required for 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

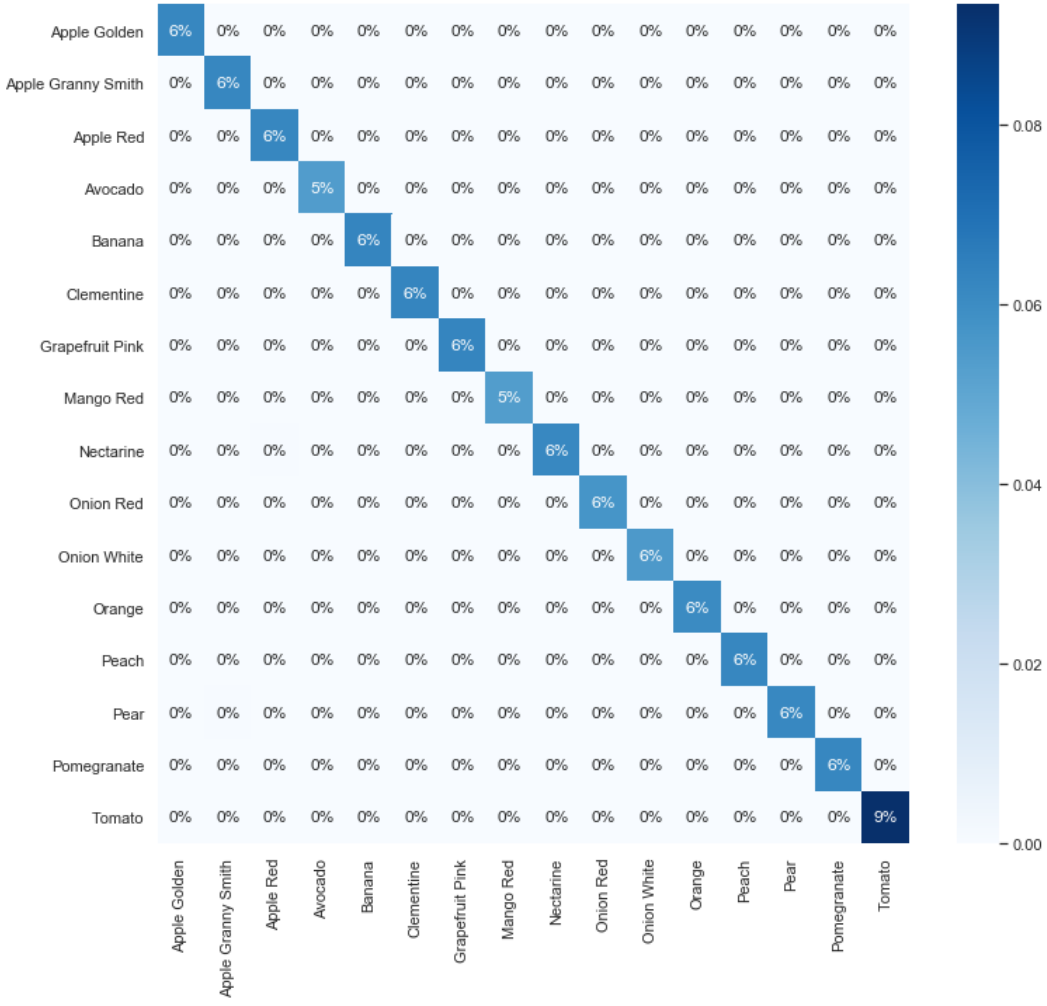

Datapoint editor
Performance & Fairness
Features
2958 datapoints loaded
Configure
Ground Truth Feature
is_recid
WHAT IS GROUND TRUTH?
The feature that your model is trying to predict. More.
Cost Ratio (FP/FN)
1.5
WHAT IS COST RATIO?
The cost of false positives relative to false negatives. Required for 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

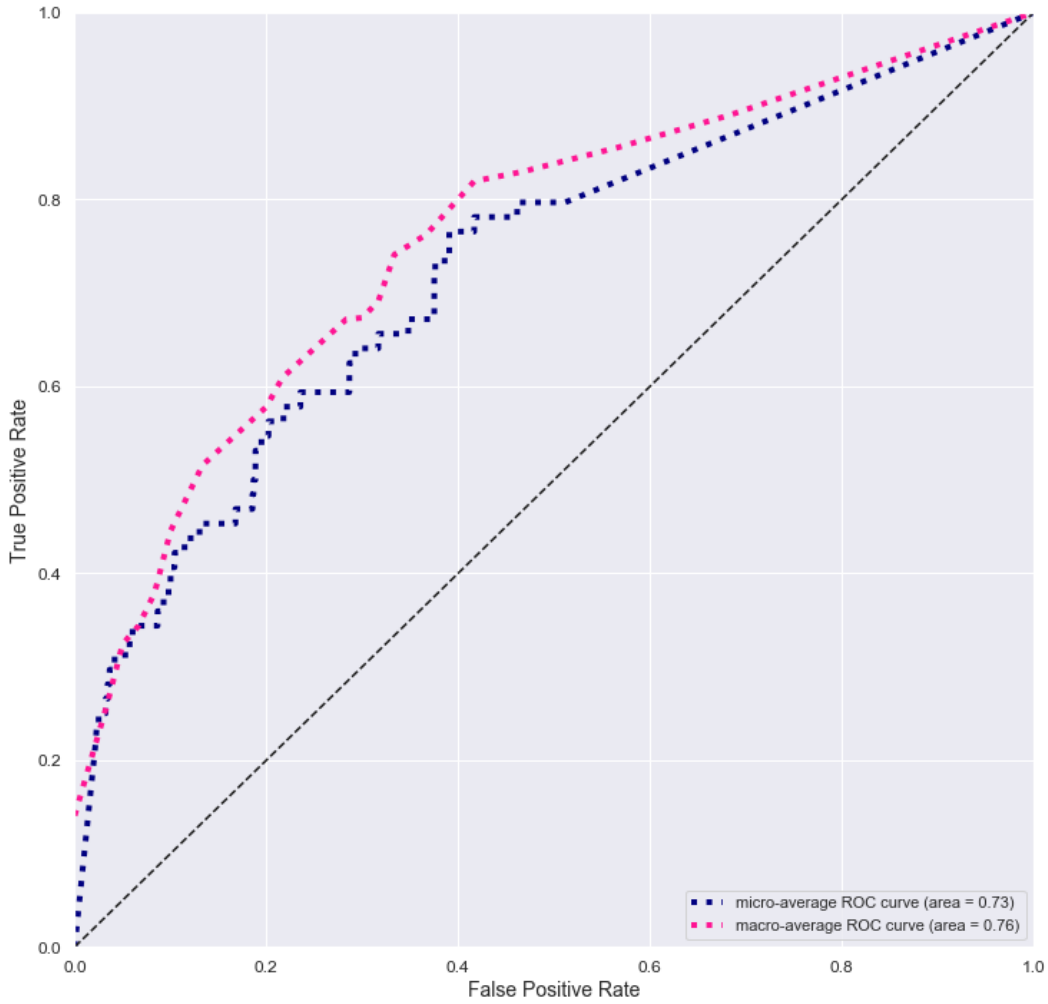
Chapter 8: Visualizing Convolutional Neural Networks

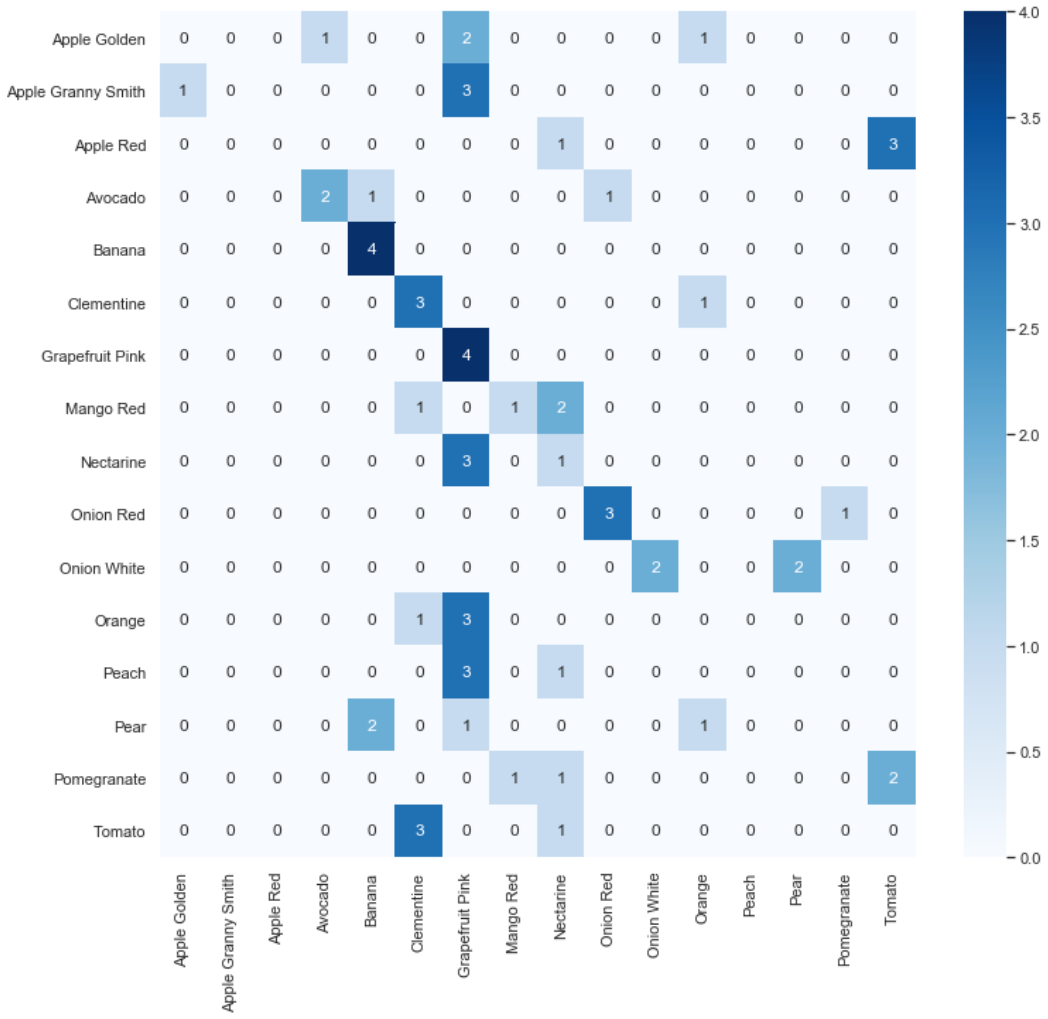

Apple Golden
Apple Granny Smith
Apple Red
Avocado
Banana
Clementine
Grapefruit Pink
Mango Red
Nectarine
Onion Red
Onion White
Orange
Peach
Pear
Pomegranate
Tomato


Apple Golden
Apple Granny Smith
Apple Red
Avocado
Banana
Clementine
Grapefruit Pink
Mango Red
Nectarine
Onion Red
Onion White
Orange
Peach
Pear
Pomegranate
Tomato


Apple Golden 6% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0%
Apple Granny Smith 0% 6% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0%
Apple Red 0% 0% 6% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0%
Avocado 0% 0% 0% 5% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0%
Banana 0% 0% 0% 0% 6% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0%
Clementine 0% 0% 0% 0% 0% 6% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0%
Grapefruit Pink 0% 0% 0% 0% 0% 0% 6% 0% 0% 0% 0% 0% 0% 0% 0% 0%
Mango Red 0% 0% 0% 0% 0% 0% 0% 5% 0% 0% 0% 0% 0% 0% 0% 0%
Nectarine 0% 0% 0% 0% 0% 0% 0% 0% 6% 0% 0% 0% 0% 0% 0% 0%
Onion Red 0% 0% 0% 0% 0% 0% 0% 0% 0% 6% 0% 0% 0% 0% 0% 0%
Onion White 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 6% 0% 0% 0% 0% 0%
Orange 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 6% 0% 0% 0% 0%
Peach 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 6% 0% 0% 0%
Pear 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 6% 0% 0%
Pomegranate 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 6% 0%
Tomato 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 9%
Apple Golden
Apple Granny Smith
Apple Red
Avocado
Banana
Clementine
Grapefruit Pink
Mango Red
Nectarine
Onion Red
Onion White
Orange
Peach
Pear
Pomegranate
Tomato
0.08
0.06
0.04
0.02
0.00

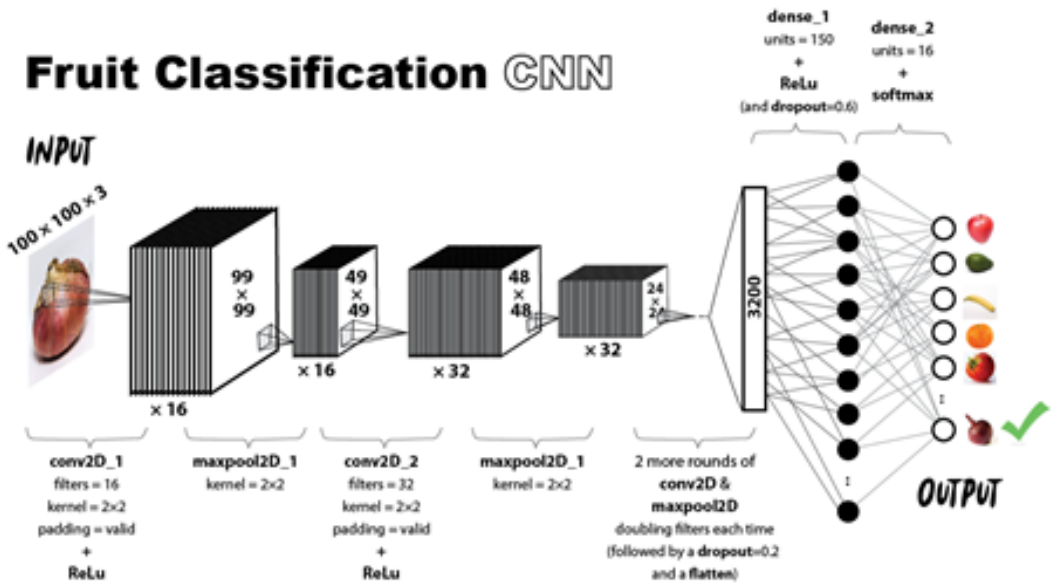
	precision	recall	f1-score	support
Apple Golden	1.000	1.000	1.000	164
Apple Granny Smith	0.994	1.000	0.997	164
Apple Red	0.994	1.000	0.997	164
Avocado	1.000	1.000	1.000	143
Banana	1.000	1.000	1.000	166
Clementine	1.000	1.000	1.000	166
Grapefruit Pink	1.000	1.000	1.000	166
Mango Red	1.000	1.000	1.000	142
Nectarine	1.000	0.994	0.997	164
Onion Red	1.000	1.000	1.000	150
Onion White	1.000	1.000	1.000	146
Orange	1.000	1.000	1.000	160
Peach	1.000	1.000	1.000	164
Pear	1.000	0.994	0.997	164
Pomegranate	1.000	1.000	1.000	164
Tomato	1.000	1.000	1.000	246
accuracy			0.999	2633
macro avg	0.999	0.999	0.999	2633
weighted avg	0.999	0.999	0.999	2633

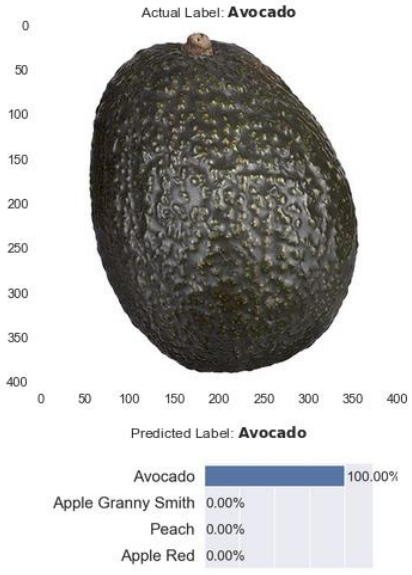

1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
micro-average ROC curve (area = 0.73)
macro-average ROC curve (area = 0.76)

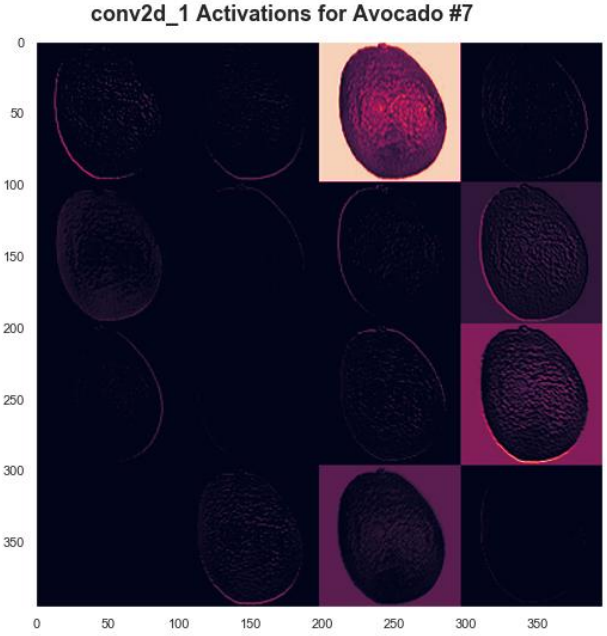

Apple Golden 0 0 0 1 0 0 2 0 0 0 0 1 0 0 0 0
Apple Granny Smith 1 0 0 0 0 0 3 0 0 0 0 0 0 0 0 0
Apple Red 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 3
Avocado 0 0 0 2 1 0 0 0 0 1 0 0 0 0 0 0
Banana 0 0 0 0 4 0 0 0 0 0 0 0 0 0 0 0
Clementine 0 0 0 0 0 3 0 0 0 0 0 1 0 0 0 0
Grapefruit Pink 0 0 0 0 0 0 4 0 0 0 0 0 0 0 0 0
Mango Red 0 0 0 0 0 1 0 1 2 0 0 0 0 0 0 0
Nectarine 0 0 0 0 0 0 3 0 1 0 0 0 0 0 0 0
Onion Red 0 0 0 0 0 0 0 0 0 3 0 0 0 0 1 0
Onion White 0 0 0 0 0 0 0 0 0 0 2 0 0 2 0 0
Orange 0 0 0 0 0 1 3 0 0 0 0 0 0 0 0 0
Peach 0 0 0 0 0 0 3 0 1 0 0 0 0 0 0 0
Pear 0 0 0 0 2 0 1 0 0 0 0 1 0 0 0 0
Pomegranate 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 2
Tomato 0 0 0 0 0 3 0 0 1 0 0 0 0 0 0 0
Apple Golden
Apple Granny Smith
Apple Red
Avocado
Banana
Clementine
Grapefruit Pink
Mango Red
Nectarine
Onion Red
Onion White
Orange
Peach
Pear
Pomegranate
Tomato
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

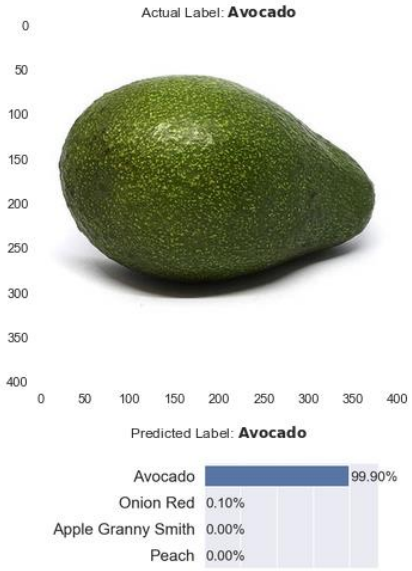
	precision	recall	f1-score	support
Apple Golden	0.000	0.000	0.000	4
Apple Granny Smith	0.000	0.000	0.000	4
Apple Red	0.000	0.000	0.000	4
Avocado	0.667	0.500	0.571	4
Banana	0.571	1.000	0.727	4
Clementine	0.375	0.750	0.500	4
Grapefruit Pink	0.211	1.000	0.348	4
Mango Red	0.500	0.250	0.333	4
Nectarine	0.143	0.250	0.182	4
Onion Red	0.750	0.750	0.750	4
Onion White	1.000	0.500	0.667	4
Orange	0.000	0.000	0.000	4
Peach	0.000	0.000	0.000	4
Pear	0.000	0.000	0.000	4
Pomegranate	0.000	0.000	0.000	4
Tomato	0.000	0.000	0.000	4
accuracy			0.312	64
macro avg	0.264	0.312	0.255	64
weighted avg	0.264	0.312	0.255	64

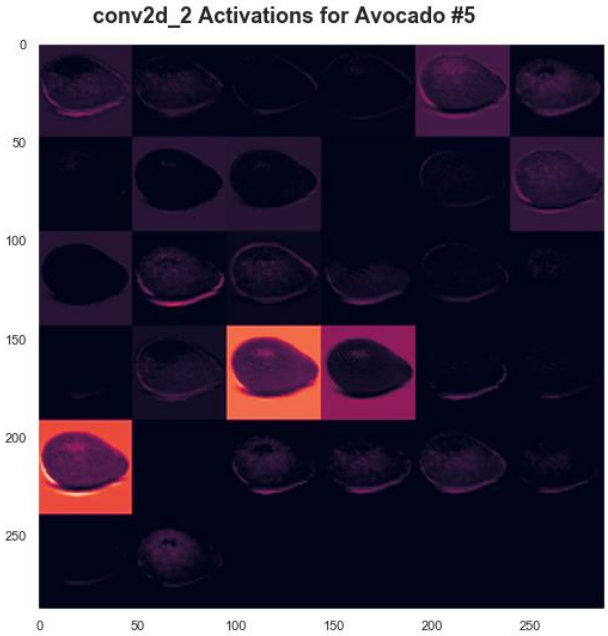
	y_true	y_pred	Grapefruit Pink	Clementine	Banana	Nectarine	Tomato	Onion Red	Avocado	Orange
0	Pear	Banana			100.0					
1	Pear	Banana			100.0					
2	Pear	Orange	5.2							94.8
3	Pear	Grapefruit Pink	96.5		3.5					
4	Avocado	Onion Red						100.0		
5	Avocado	Avocado						0.1	99.9	
6	Avocado	Banana			100.0					
7	Avocado	Avocado							100.0	
8	Pomegranate	Tomato					100.0			
:	:	:	:	:	:	:	:	:	:	:
16	Apple Golden	Avocado							99.8	0.2
17	Apple Golden	Grapefruit Pink	100.0							
18	Apple Golden	Orange								100.0
19	Apple Golden	Grapefruit Pink	100.0							
20	Nectarine	Grapefruit Pink	100.0							
21	Nectarine	Nectarine				100.0				
22	Nectarine	Grapefruit Pink	100.0							
23	Nectarine	Grapefruit Pink	100.0							
24	Clementine	Clementine		100.0						
25	Clementine	Clementine	1.9	98.1						
26	Clementine	Clementine		100.0						
27	Clementine	Orange								100.0
28	Onion White	Pear								
29	Onion White	Pear								
30	Onion White	Onion White								
31	Onion White	Onion White	8.2		0.6					
32	Apple Granny Smith	Grapefruit Pink	100.0							
33	Apple Granny Smith	Apple Golden								
34	Apple Granny Smith	Grapefruit Pink	100.0							
35	Apple Granny Smith	Grapefruit Pink	92.7		7.2				0.1	
36	Grapefruit Pink	Grapefruit Pink	100.0							
37	Grapefruit Pink	Grapefruit Pink	100.0							
:	:	:	:	:	:	:	:	:	:	:
56	Orange	Clementine	38.7	61.3						
57	Orange	Grapefruit Pink	100.0							
58	Orange	Grapefruit Pink	51.8	48.2						
59	Orange	Grapefruit Pink	91.6	8.4						
60	Peach	Grapefruit Pink	100.0							
61	Peach	Grapefruit Pink	100.0							
62	Peach	Nectarine		0.1		99.9				
63	Peach	Grapefruit Pink	87.7				11.9			

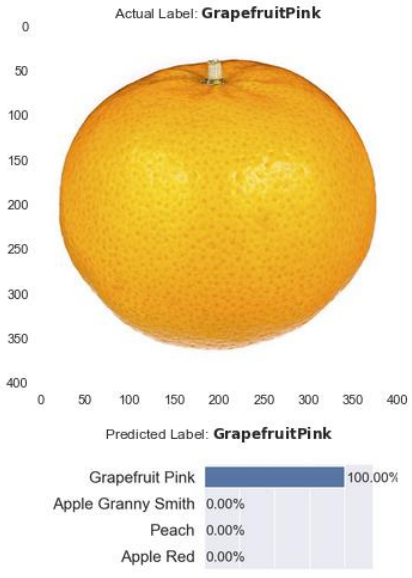

Fruit Classification CNN
INPUT
100 × 100 × 3
99 × 99
× 16
49 × 49
× 16
48 × 48
× 32
24 × 24
× 32
3200
dense_1
units = 150
+
ReLu
(and dropout=0.6)
dense_2
units = 16
+
softmax
OUTPUT
conv2D_1
filters = 16
kernel = 2×2
padding = valid
+
ReLu
maxpool2D_1
kernel = 2×2
conv2D_2
filters = 32
kernel = 2×2
padding = valid
+
ReLu
maxpool2D_1
kernel = 2×2
2 more rounds of
conv2D &
maxpool2D
doubling filters each time
(followed by a dropout=0.2
and a flatten)

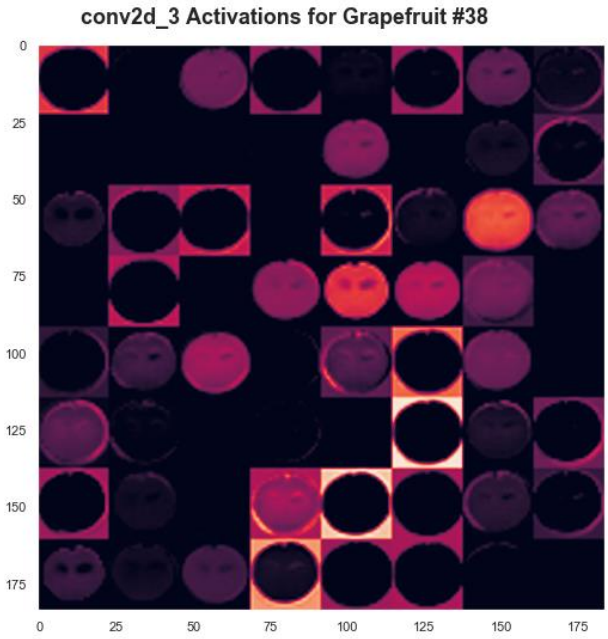

Actual Label: Avocado
Predicted Label: Avocado
Avocado 100.00%
Apple Granny Smith 0.00%
Peach 0.00%
Apple Red 0.00%

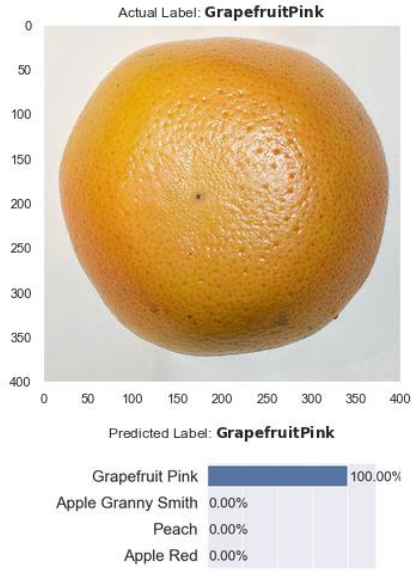

conv2d_1 Activations for Avocado #7

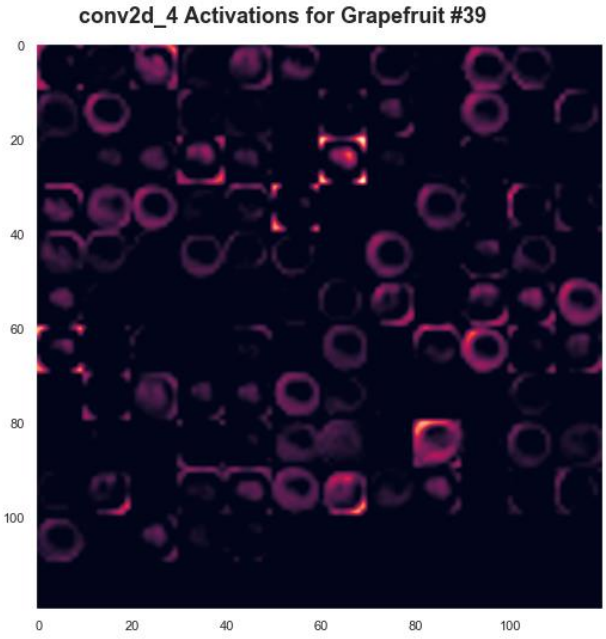

Actual Label: **Avocado**
Predicted Label: **Avocado**
Avocado 99.90%
Onion Red 0.10%
Apple Granny Smith 0.00%
Peach 0.00%


conv2d_2 Activations for Avocado #5


Actual Label: **GrapefruitPink**
Predicted Label: **GrapefruitPink**
Grapefruit Pink 100.00%
Apple Granny Smith 0.00%
Peach 0.00%
Apple Red 0.00%

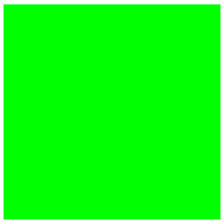

conv2d_3 Activations for Grapefruit #38


Actual Label: GrapefruitPink
0
50
100
150
200
250
300
350
400
0
50
100
150
200
250
300
350
400
Predicted Label: GrapefruitPink
Grapefruit Pink
100.00%
Apple Granny Smith
0.00%
Peach
0.00%
Apple Red
0.00%


conv2d_4 Activations for Grapefruit #39
0
20
40
60
80
100
0
20
40
60
80
100

conv2d_1 Layer

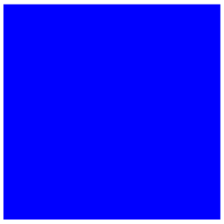
Filter #0



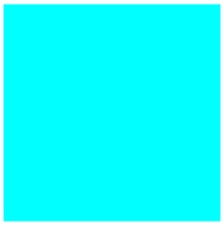
Filter #1

Filter #2

Filter #3



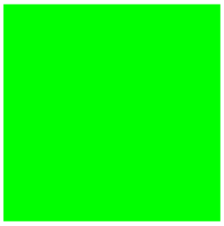
Filter #4



Filter #5

Filter #6

Filter #7

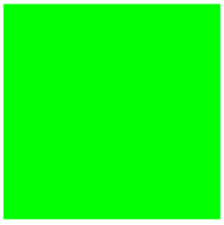


Filter #8

Filter #9

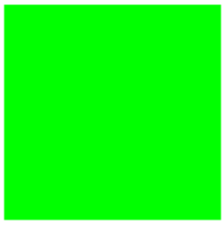
Filter #10

Filter #11

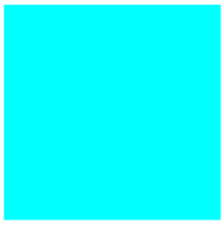


Filter #12

Filter #13

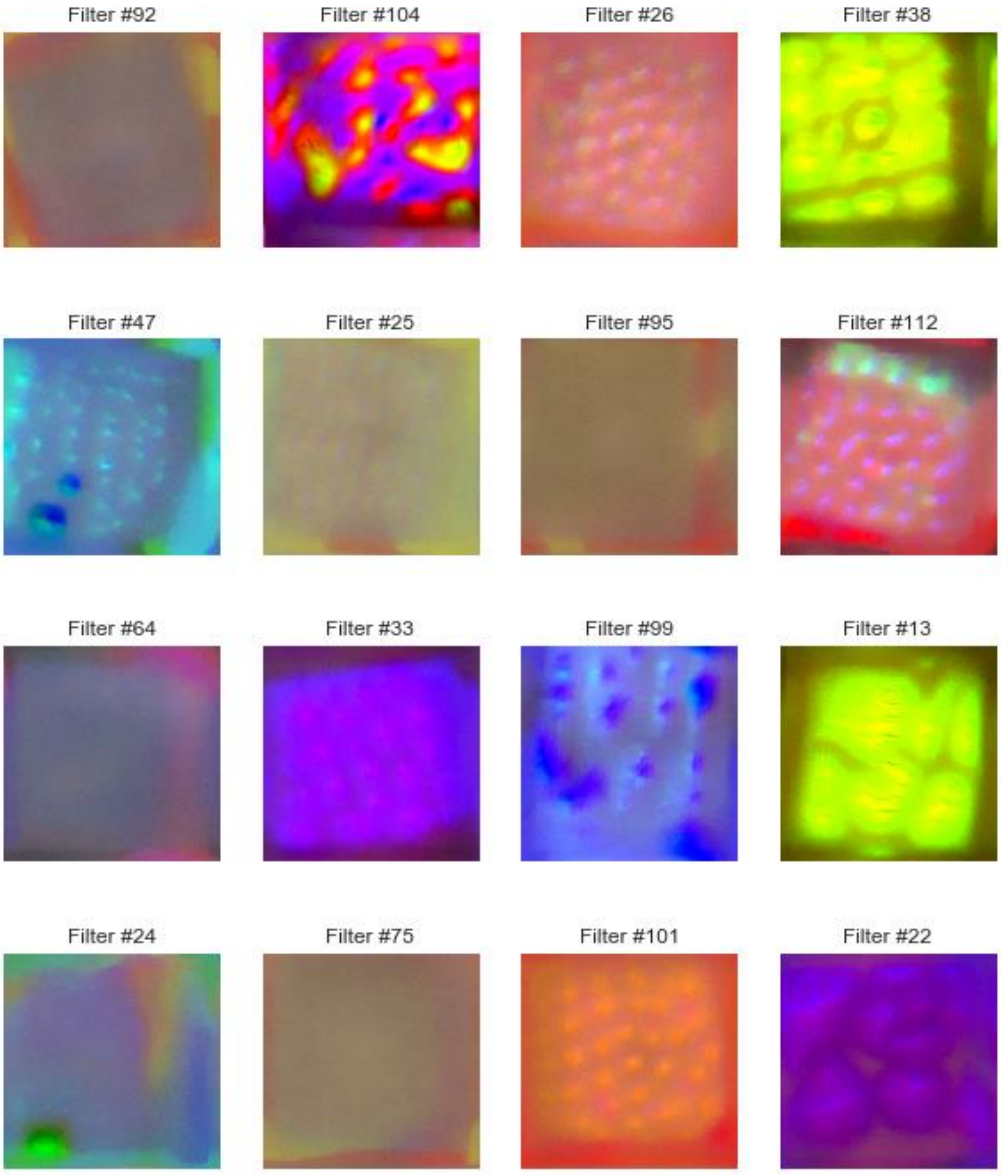


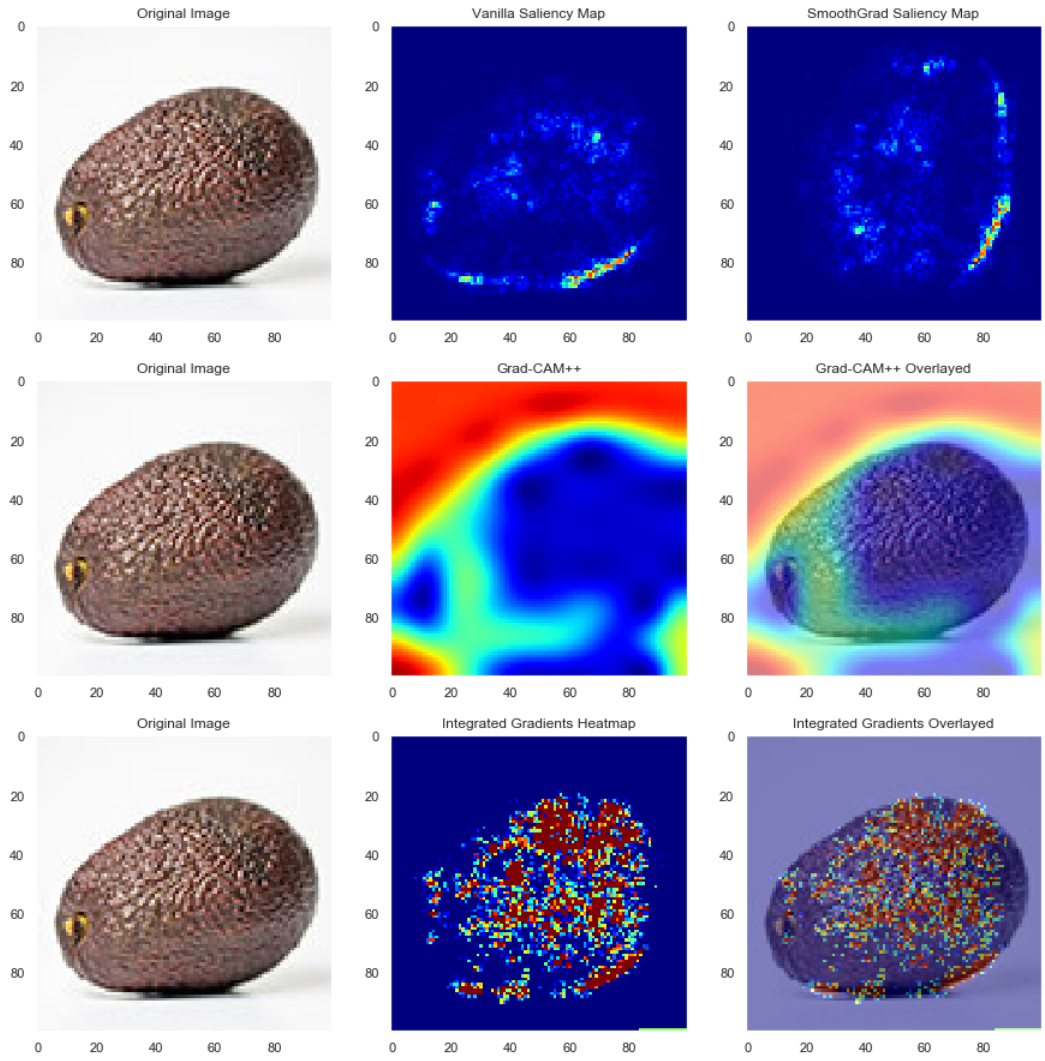
Filter #14

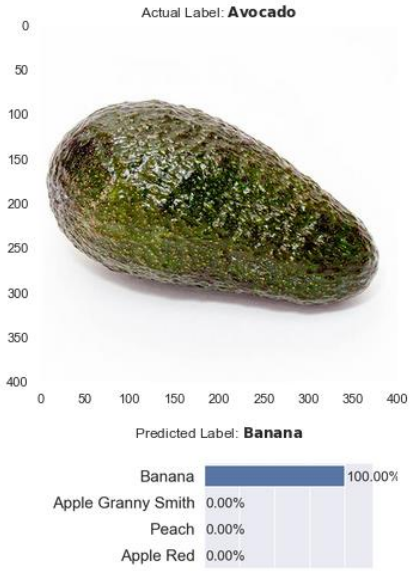


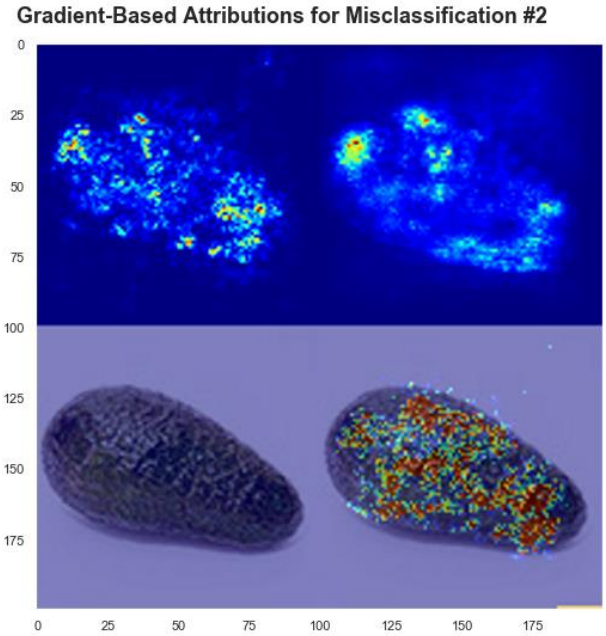
Filter #15

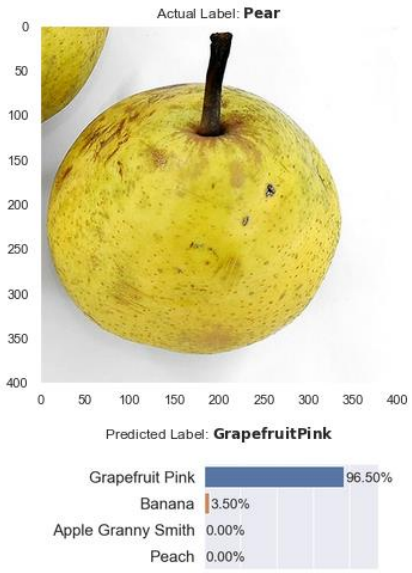
conv2d_4 Layer

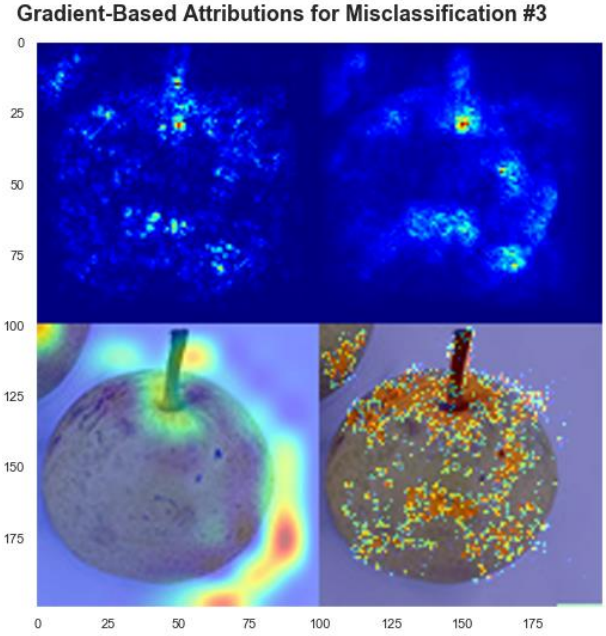


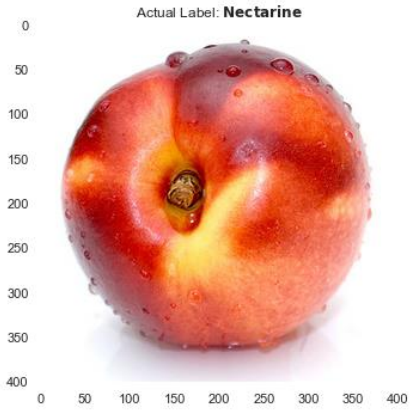

Original Image
Vanilla Saliency Map
SmoothGrad Saliency Map
Original Image
Grad-CAM++
Grad-CAM++ Overlayed
Original Image
Integrated Gradients Heatmap
Integrated Gradients Overlayed

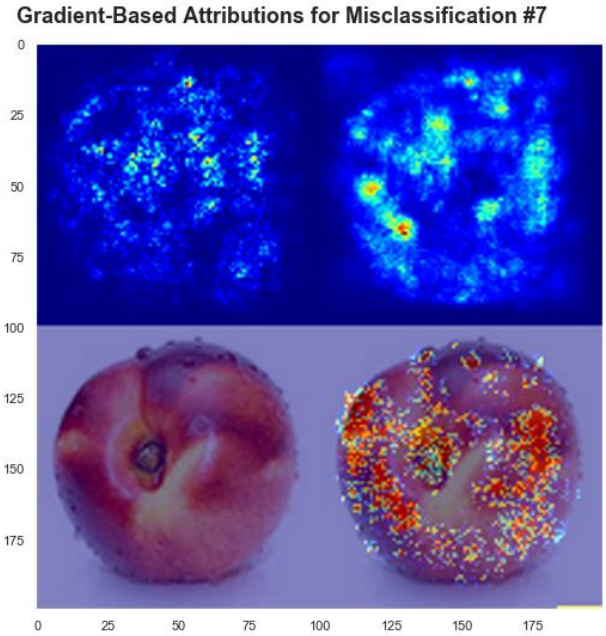

Actual Label: Avocado
Predicted Label: Banana
Banana 100.00%
Apple Granny Smith 0.00%
Peach 0.00%
Apple Red 0.00%

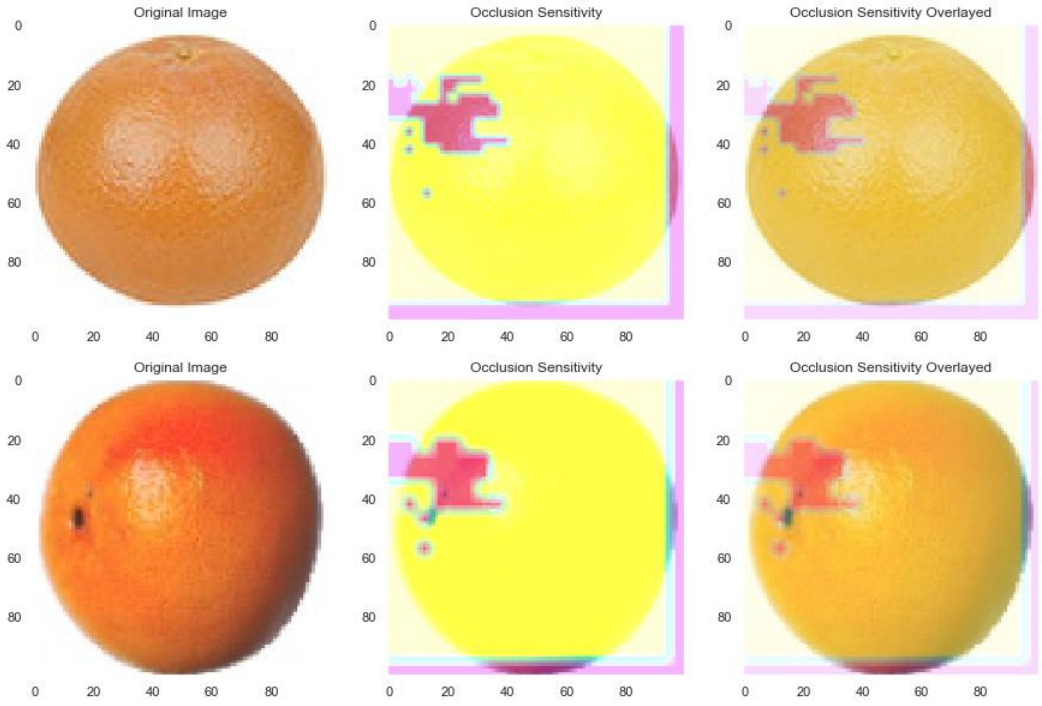

Gradient-Based Attributions for Misclassification #2

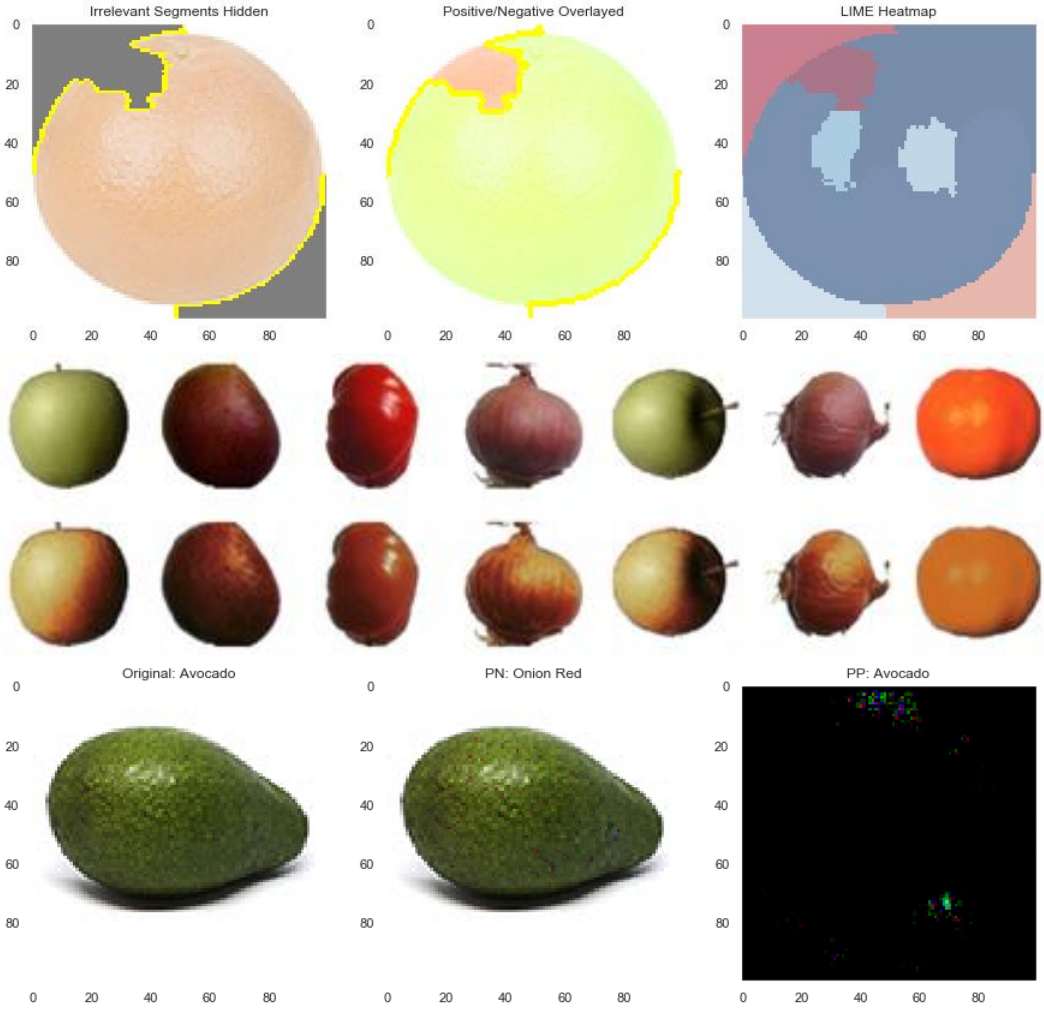

Actual Label: Pear
Predicted Label: GrapefruitPink
Grapefruit Pink 96.50%
Banana 3.50%
Apple Granny Smith 0.00%
Peach 0.00%

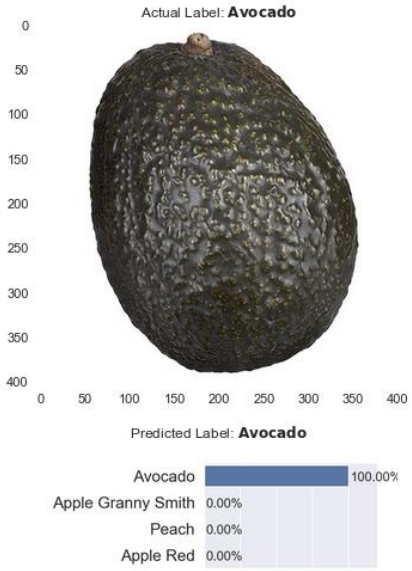

Gradient-Based Attributions for Misclassification #3

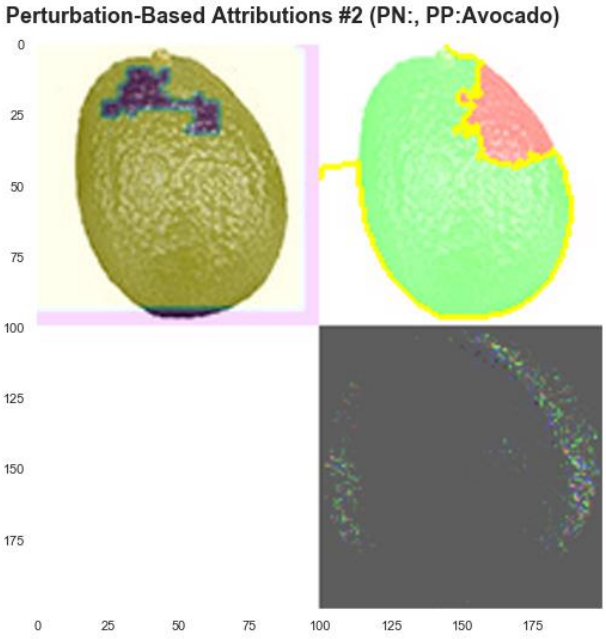

Actual Label: **Nectarine**
Predicted Label: **GrapefruitPink**
Grapefruit Pink 100.00%
Apple Granny Smith 0.00%
Peach 0.00%
Apple Red 0.00%

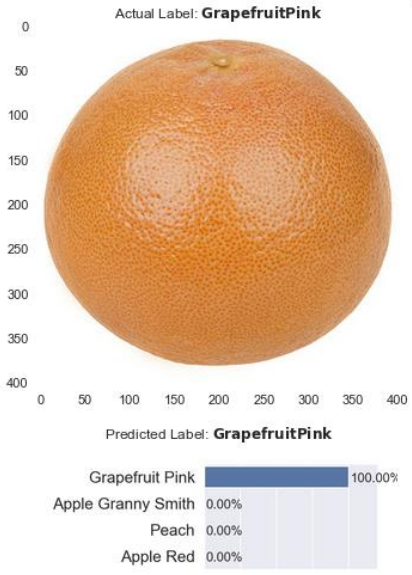

Gradient-Based Attributions for Misclassification #7

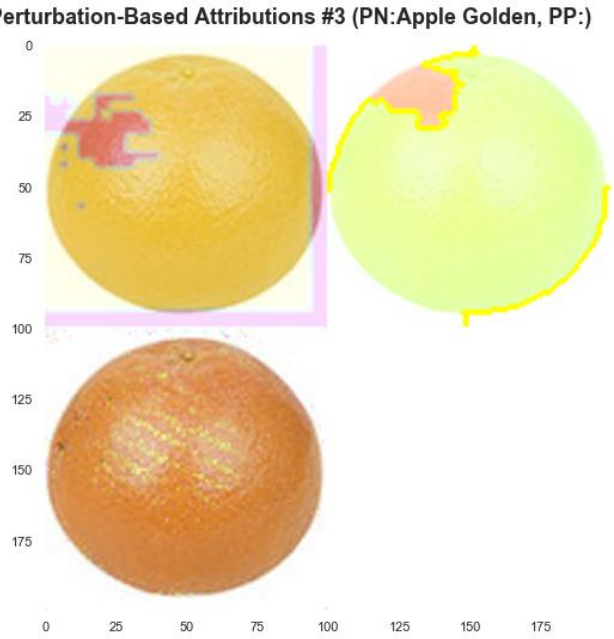

Original Image
Occlusion Sensitivity
Occlusion Sensitivity Overlayed
Original Image
Occlusion Sensitivity
Occlusion Sensitivity Overlayed

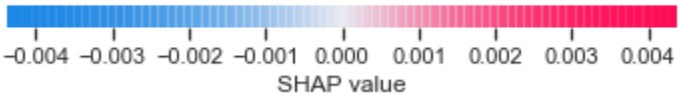

Irrelevant Segments Hidden
Positive/Negative Overlayed
LIME Heatmap
Original: Avocado
PN: Onion Red
PP: Avocado


Actual Label: Avocado
Predicted Label: Avocado
Avocado 100.00%
Apple Granny Smith 0.00%
Peach 0.00%
Apple Red 0.00%

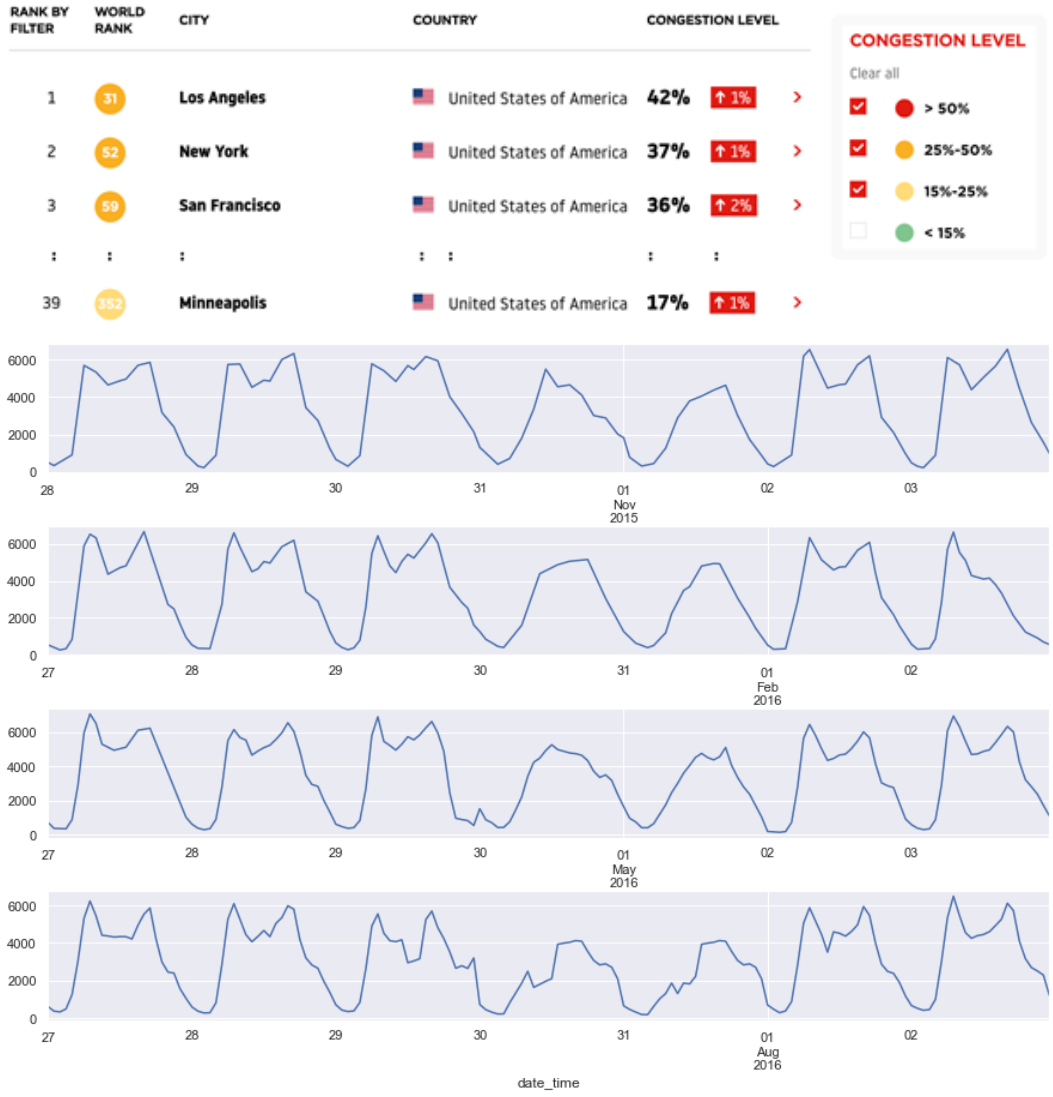

Perturbation-Based Attributions #2 (PN:, PP:Avocado)

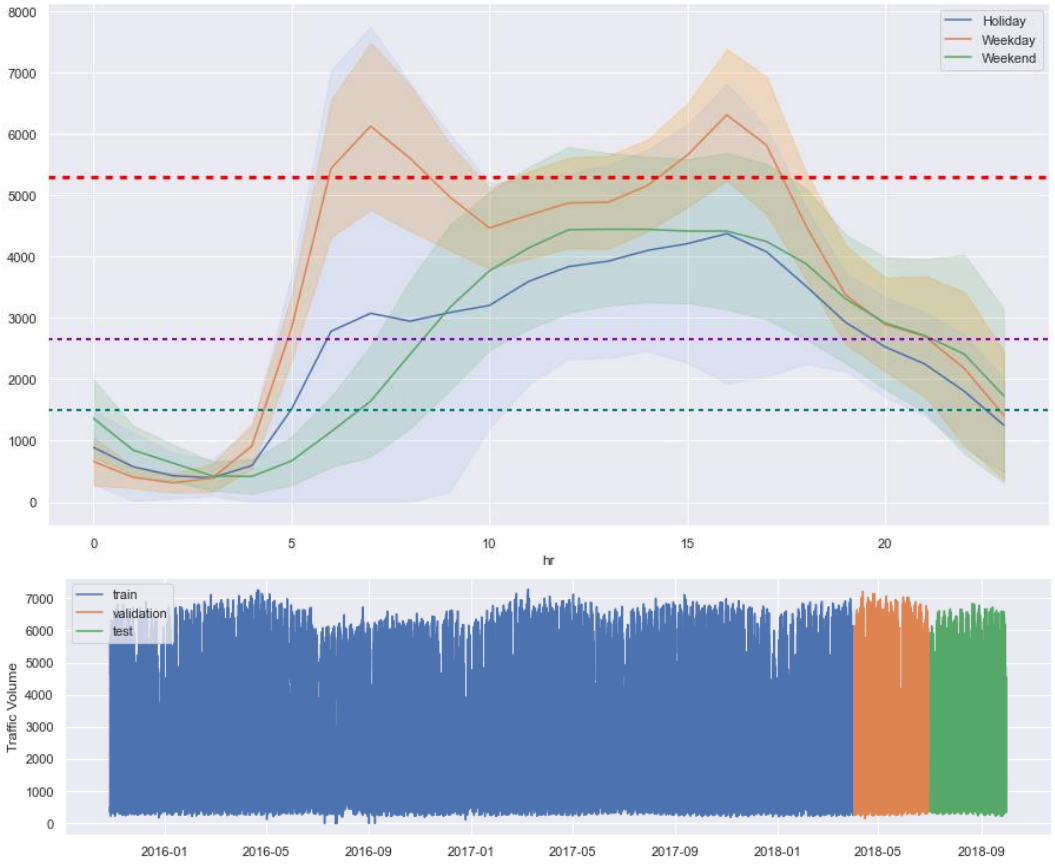

Actual Label: GrapefruitPink
Predicted Label: GrapefruitPink
Grapefruit Pink 100.00%
Apple Granny Smith 0.00%
Peach 0.00%
Apple Red 0.00%


Perturbation-Based Attributions #3 (PN:Apple Golden, PP:)


−0.004 −0.003 −0.002 −0.001 0.000 0.001 0.002 0.003 0.004
SHAP value

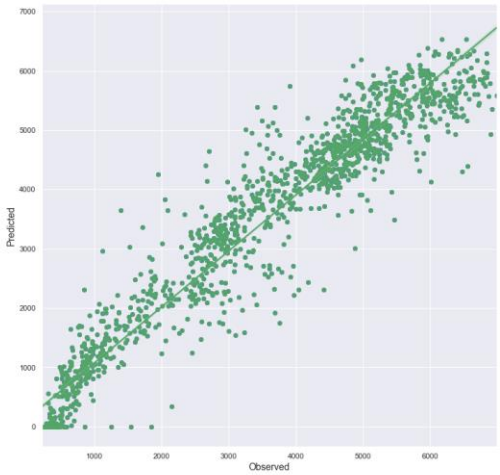
Chapter 9: Interpretation Methods for Multivariate Forecasting and Sensitivity Analysis




Holiday
Weekday
Weekend
8000
7000
6000
5000
4000
3000
2000
1000
0
0
5
10
15
20
hr

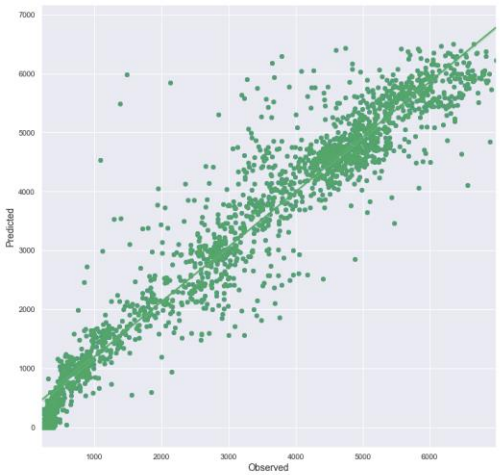
train
validation
test
Traffic Volume
7000
6000
5000
4000
3000
2000
1000
0
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05
2018-09

Traffic_Bidirectional_LSTM_672

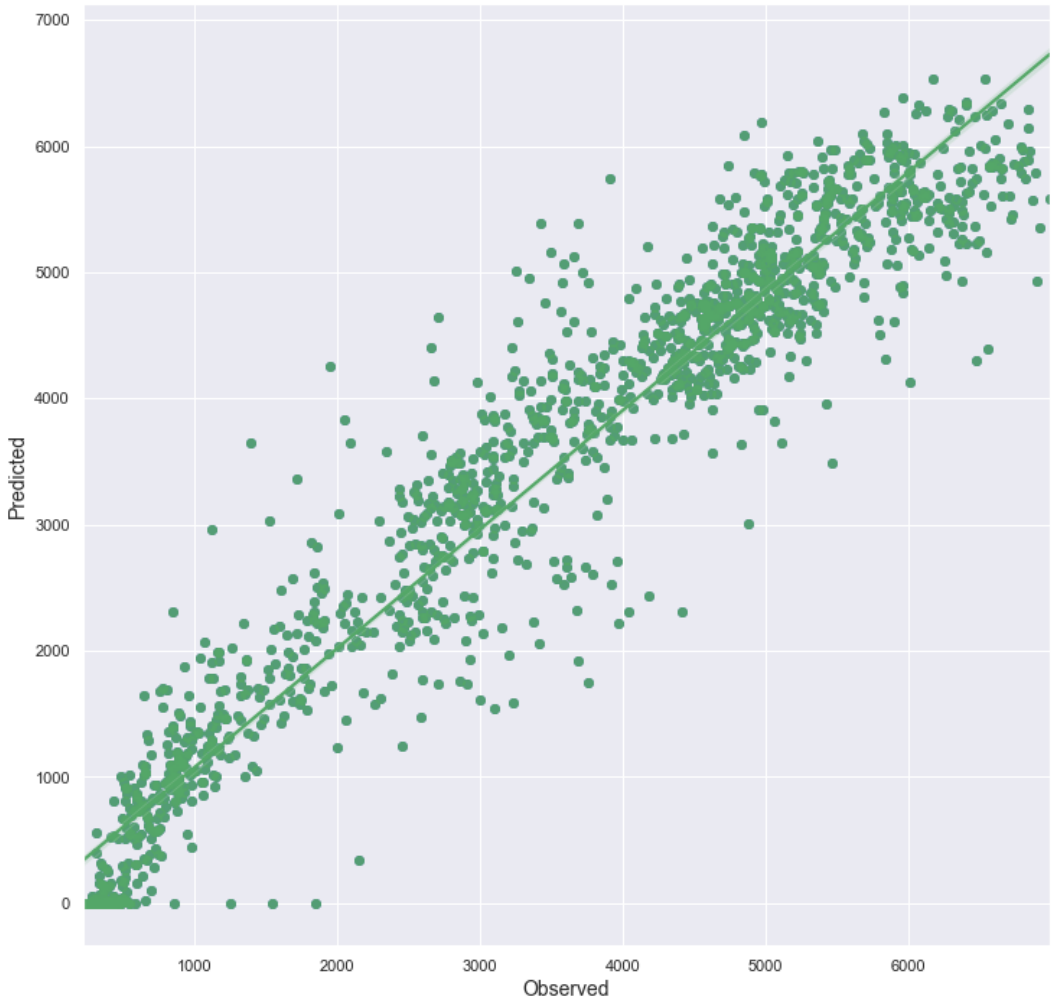


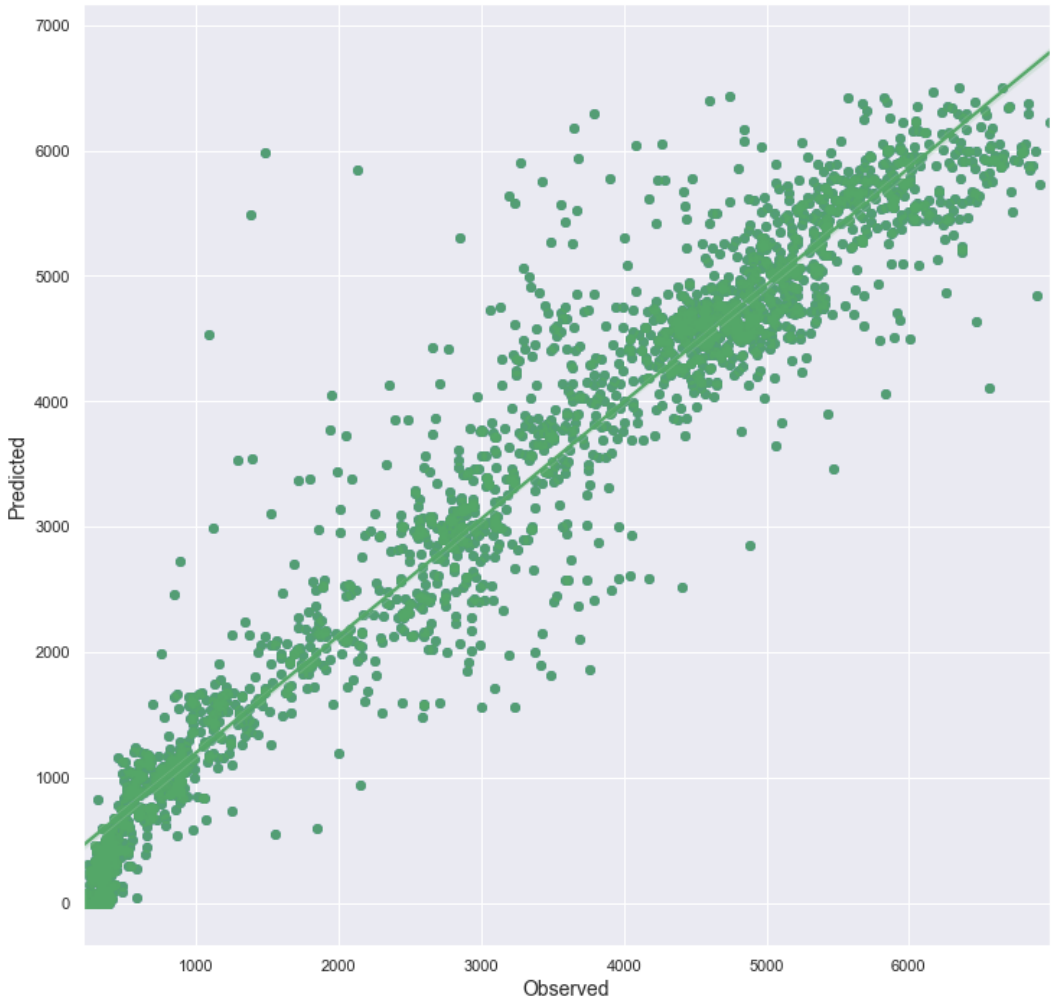
RMSE_train: 537.1589
RMSE_test: 542.6649
r2: 0.9263

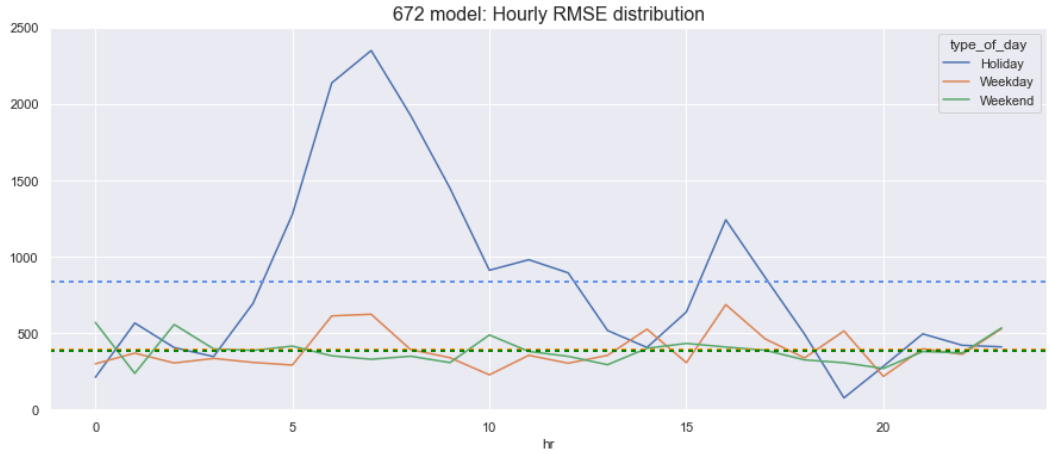
Traffic_LSTM_168

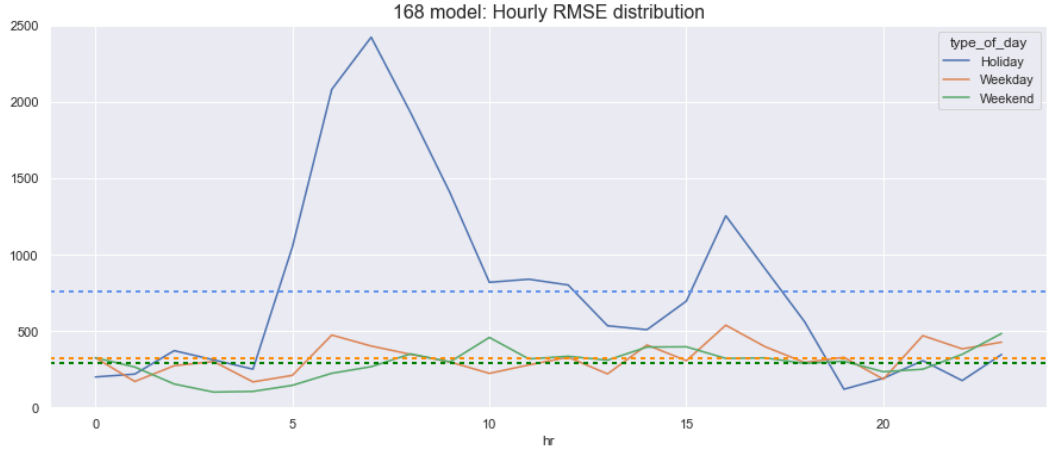


RMSE_train: 473.4274
RMSE_test: 561.8984
r2: 0.9187

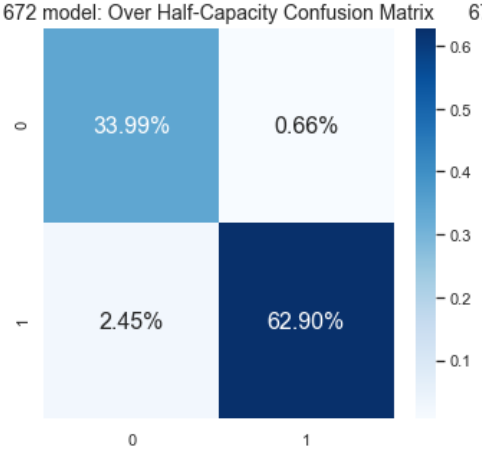

7000
6000
5000
4000
3000
2000
1000
0
Predicted
1000
2000
3000
4000
5000
6000
Observed


7000
6000
5000
4000
3000
2000
1000
0
Predicted
1000
2000
3000
4000
5000
6000
Observed

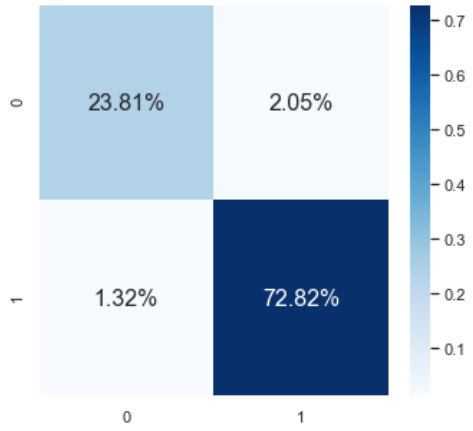

672 model: Hourly RMSE distribution
type_of_day
Holiday
Weekday
Weekend
2500
2000
1500
1000
500
0
0
5
10
15
20
hr


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

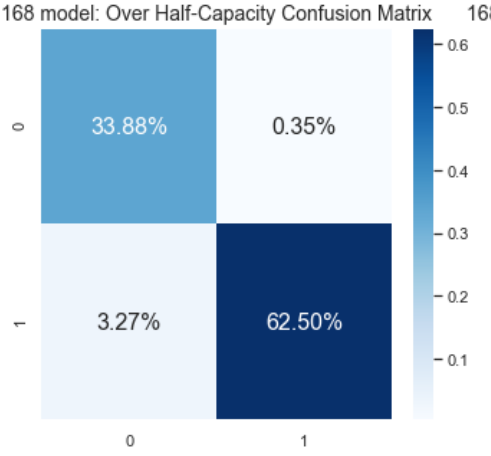
672 model: Over Half-Capacity Confusion Matrix



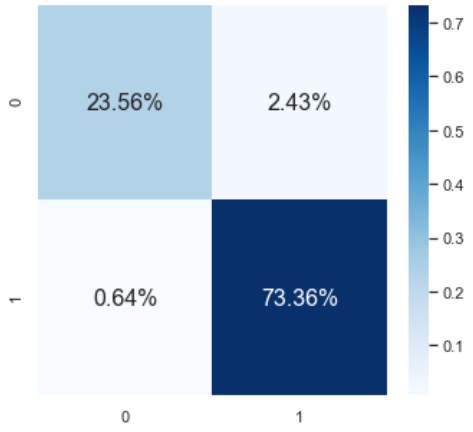
672 model: Over No-Construction Threshold Confusion Matrix



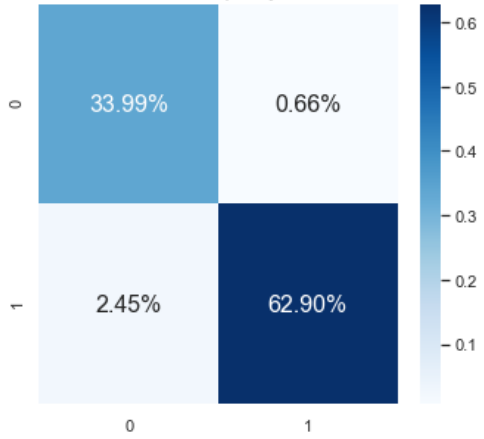
168 model: Over Half-Capacity Confusion Matrix



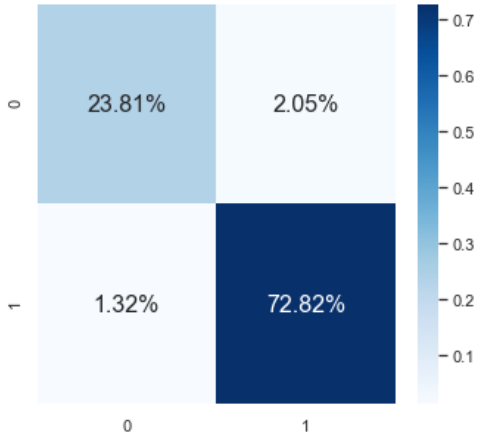
168 model: Over No-Construction Threshold Confusion Matrix



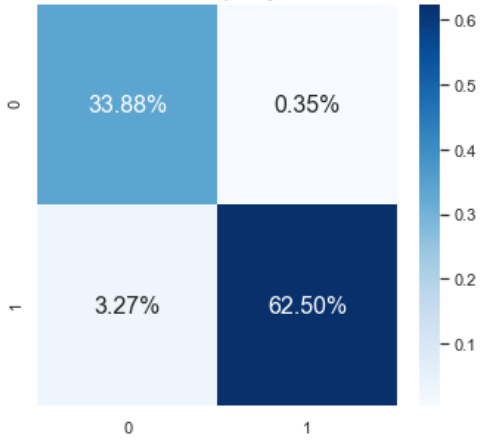
672 model: Over Half-Capacity Confusion Matrix



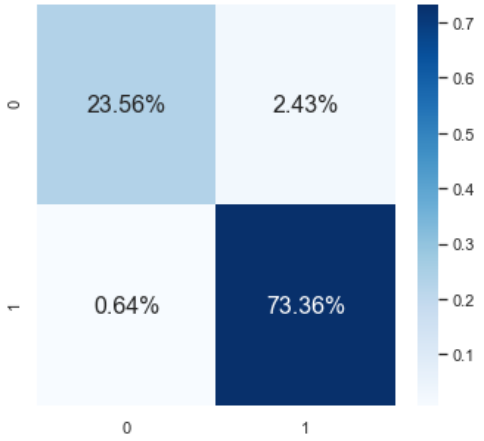
672 model: Over No-Construction Threshold Confusion Matrix

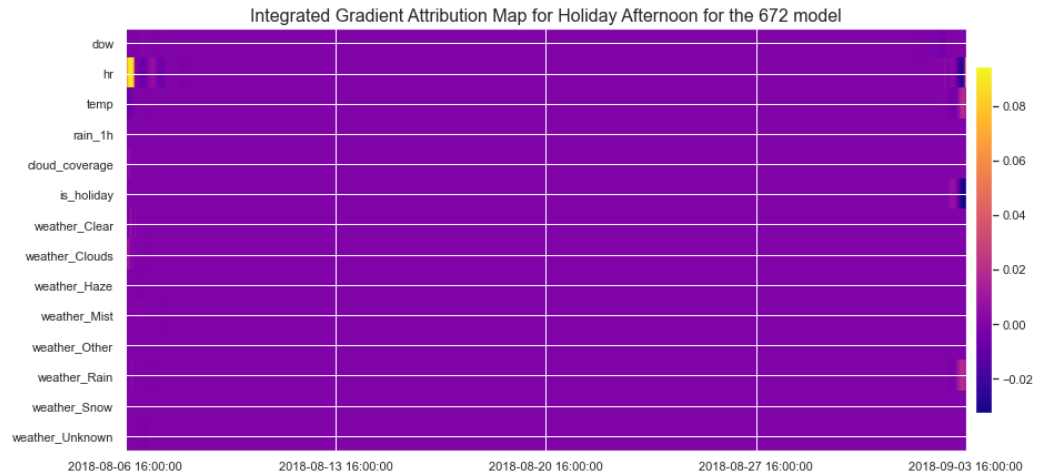


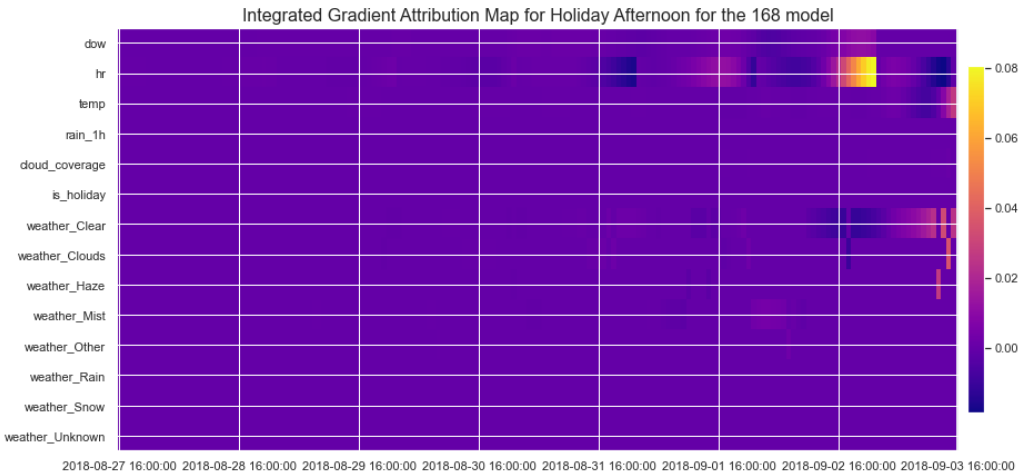
168 model: Over Half-Capacity Confusion Matrix

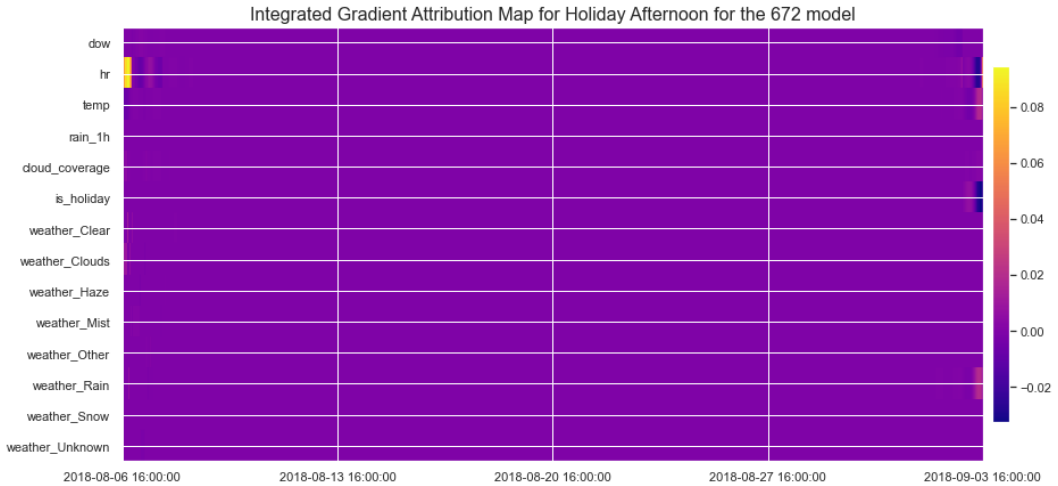


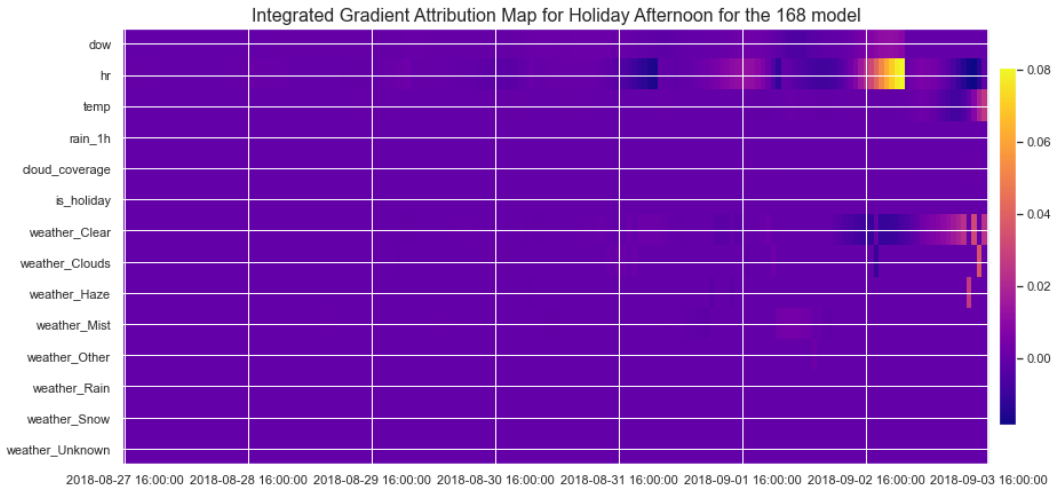
168 model: Over No-Construction Threshold Confusion Matrix

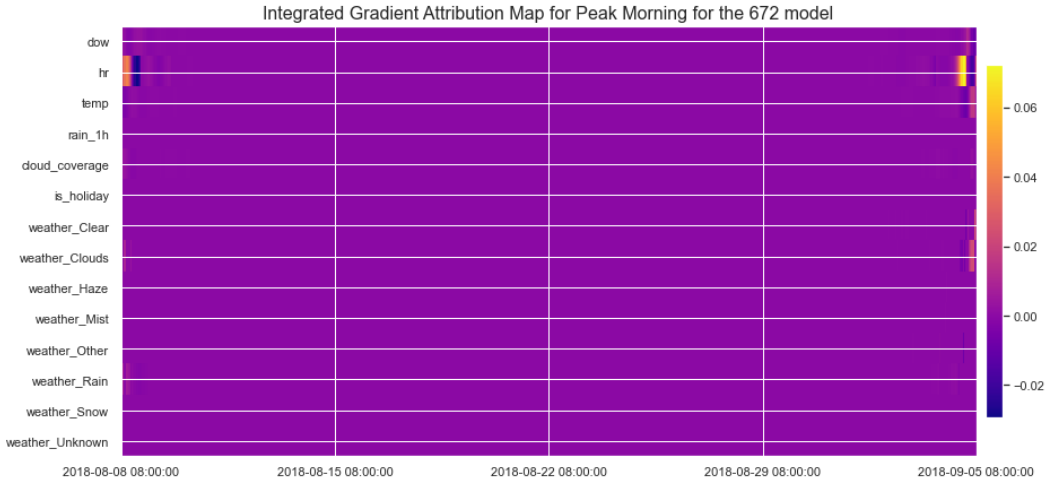


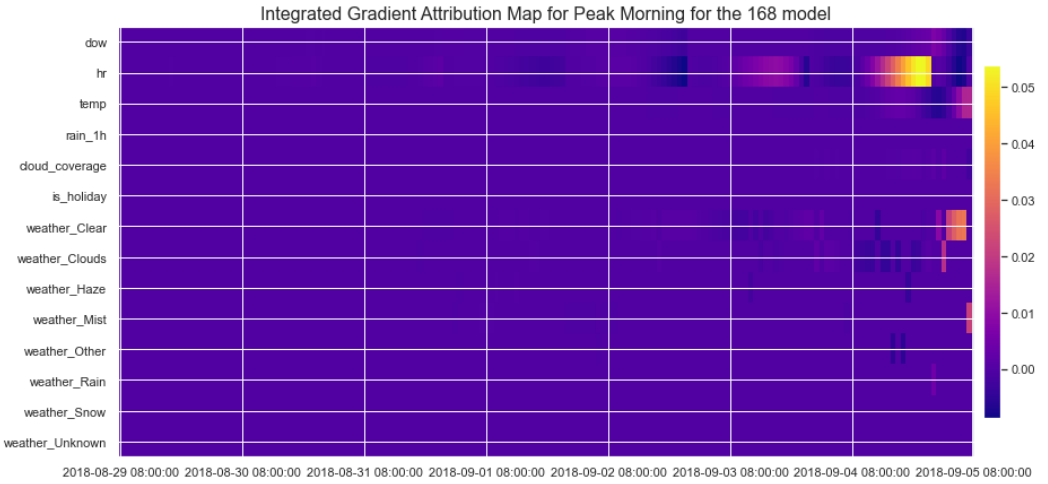

Integrated Gradient Attribution Map for Holiday Afternoon for the 672 model
dow
hr
temp
rain_1h
cloud_coverage
is_holiday
weather_Clear
weather_Clouds
weather_Haze
weather_Mist
weather_Other
weather_Rain
weather_Snow
weather_Unknown
2018-08-06 16:00:00
2018-08-13 16:00:00
2018-08-20 16:00:00
2018-08-27 16:00:00
2018-09-03 16:00:00
0.08
0.06
0.04
0.02
0.00
−0.02

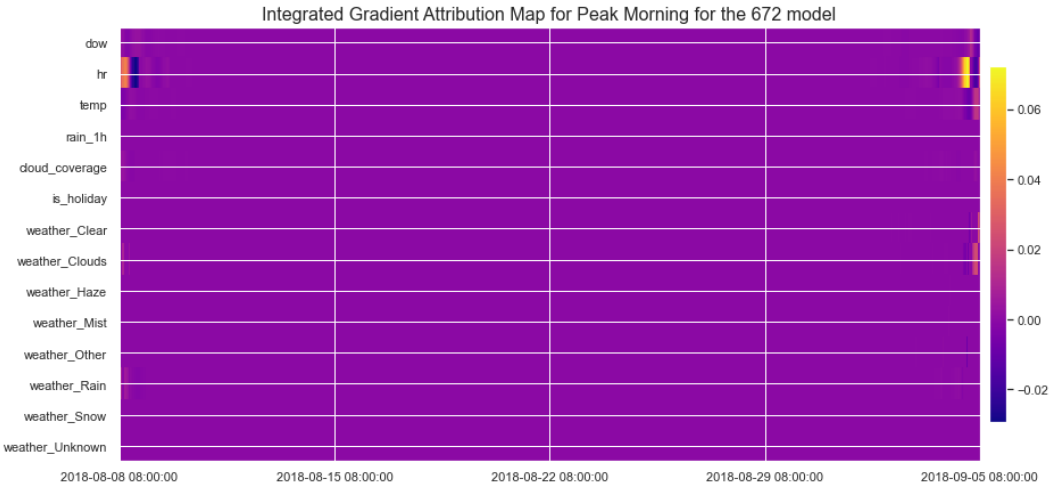

Integrated Gradient Attribution Map for Holiday Afternoon for the 168 model
dow
hr
temp
rain_1h
cloud_coverage
is_holiday
weather_Clear
weather_Clouds
weather_Haze
weather_Mist
weather_Other
weather_Rain
weather_Snow
weather_Unknown
2018-08-27 16:00:00
2018-08-28 16:00:00
2018-08-29 16:00:00
2018-08-30 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

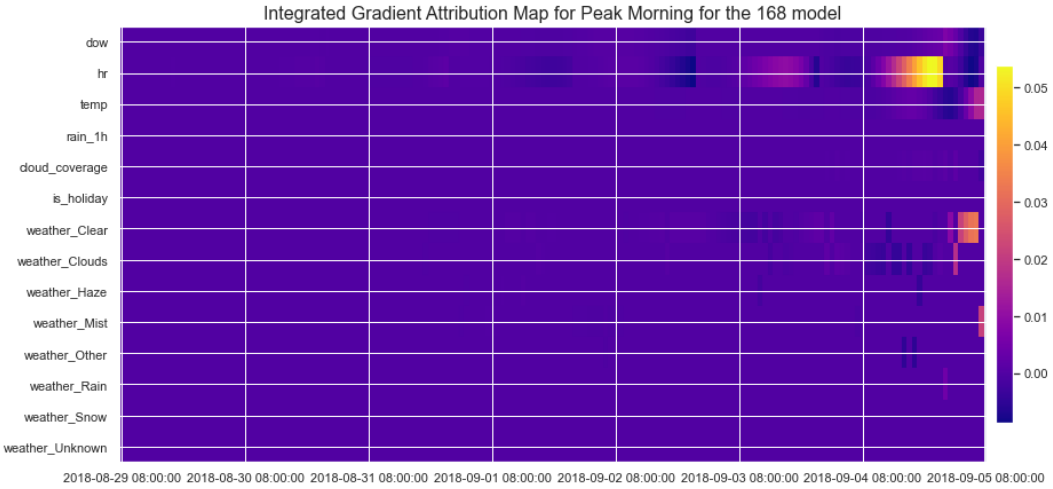

Integrated Gradient Attribution Map for Holiday Afternoon for the 672 model
dow
hr
temp
rain_1h
cloud_coverage
is_holiday
weather_Clear
weather_Clouds
weather_Haze
weather_Mist
weather_Other
weather_Rain
weather_Snow
weather_Unknown
2018-08-06 16:00:00
2018-08-13 16:00:00
2018-08-20 16:00:00
2018-08-27 16:00:00
2018-09-03 16:00:00
0.08
0.06
0.04
0.02
0.00
−0.02

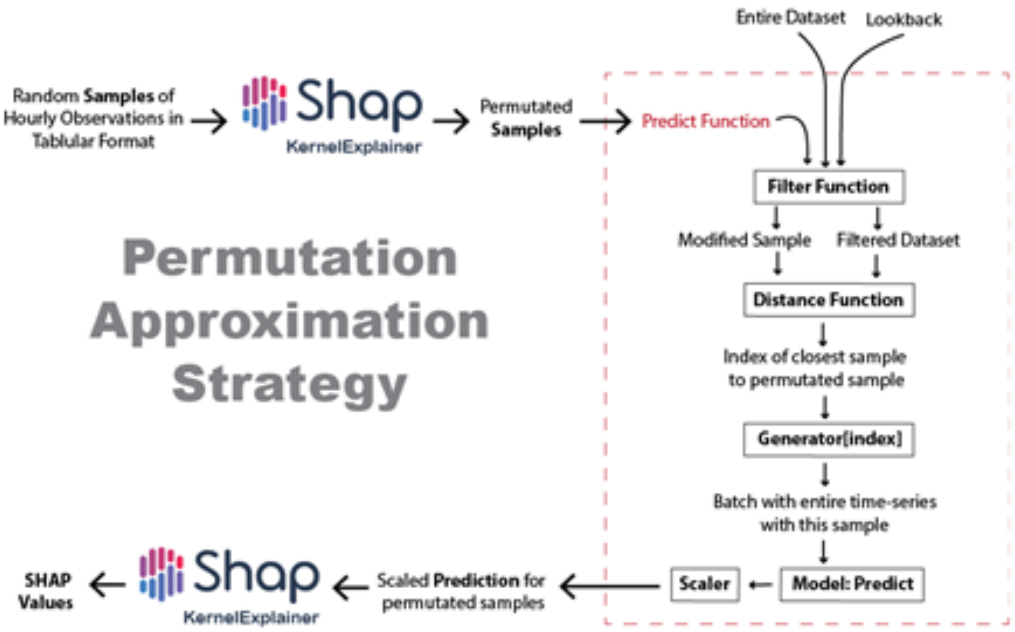

Integrated Gradient Attribution Map for Holiday Afternoon for the 168 model
dow
hr
temp
rain_1h
cloud_coverage
is_holiday
weather_Clear
weather_Clouds
weather_Haze
weather_Mist
weather_Other
weather_Rain
weather_Snow
weather_Unknown
2018-08-27 16:00:00
2018-08-28 16:00:00
2018-08-29 16:00:00
2018-08-30 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

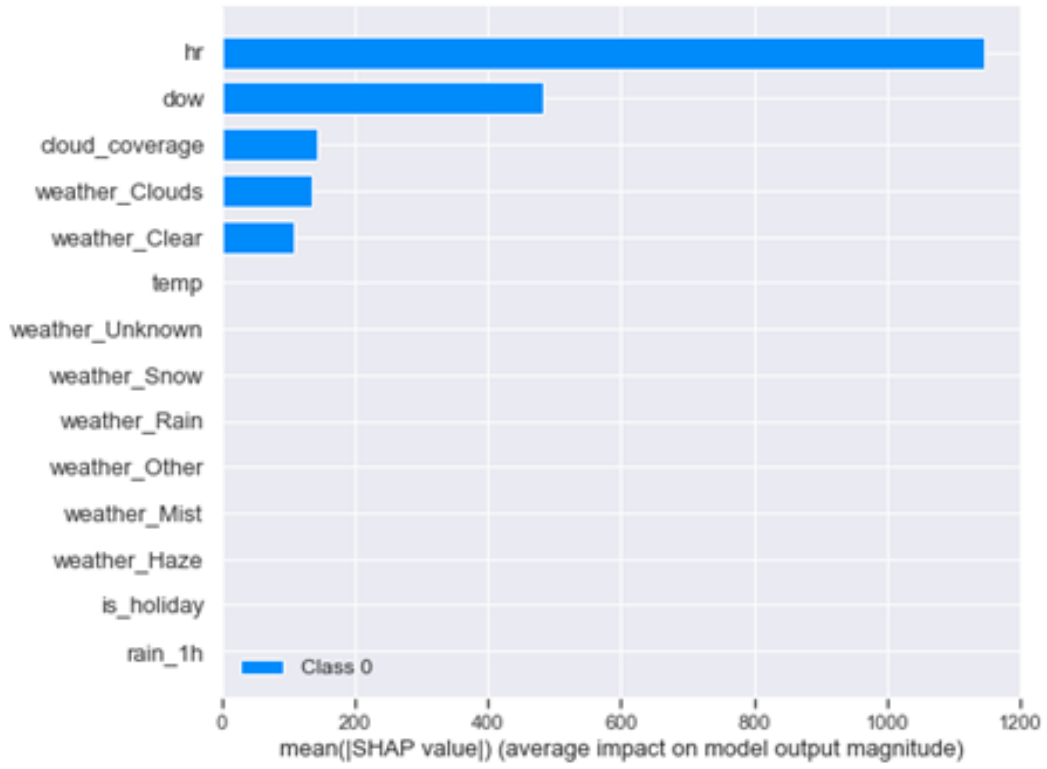

Integrated Gradient Attribution Map for Peak Morning for the 672 model
dow
hr
temp
rain_1h
cloud_coverage
is_holiday
weather_Clear
weather_Clouds
weather_Haze
weather_Mist
weather_Other
weather_Rain
weather_Snow
weather_Unknown
2018-08-08 08:00:00
2018-08-15 08:00:00
2018-08-22 08:00:00
2018-08-29 08:00:00
2018-09-05 08:00:00
0.06
0.04
0.02
0.00
−0.02


Integrated Gradient Attribution Map for Peak Morning for the 168 model
dow
hr
temp
rain_1h
cloud_coverage
is_holiday
weather_Clear
weather_Clouds
weather_Haze
weather_Mist
weather_Other
weather_Rain
weather_Snow
weather_Unknown
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
0.05
0.04
0.03
0.02
0.01
0.00

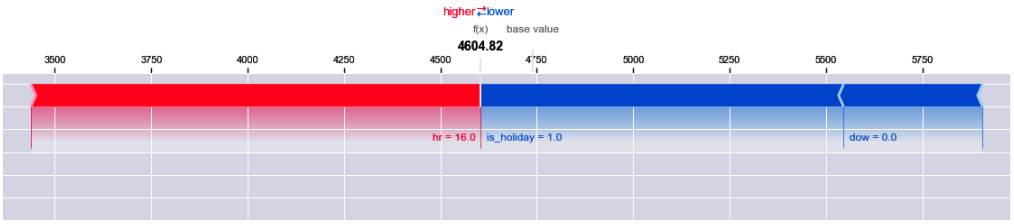

Integrated Gradient Attribution Map for Peak Morning for the 672 model
dow
hr
temp
rain_1h
cloud_coverage
is_holiday
weather_Clear
weather_Clouds
weather_Haze
weather_Mist
weather_Other
weather_Rain
weather_Snow
weather_Unknown
2018-08-08 08:00:00
2018-08-15 08:00:00
2018-08-22 08:00:00
2018-08-29 08:00:00
2018-09-05 08:00:00
0.06
0.04
0.02
0.00
−0.02


Integrated Gradient Attribution Map for Peak Morning for the 168 model
dow
hr
temp
rain_1h
cloud_coverage
is_holiday
weather_Clear
weather_Clouds
weather_Haze
weather_Mist
weather_Other
weather_Rain
weather_Snow
weather_Unknown
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
0.05
0.04
0.03
0.02
0.01
0.00

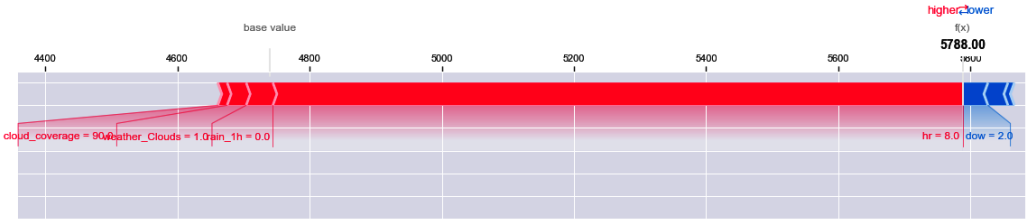

Random Samples of Hourly Observations in Tablular Format
Shap
KernelExplainer
Permutated Samples
Entire Dataset
Lookback
Predict Function
Filter Function
Modified Sample
Filtered Dataset
Distance Function
Index of closest sample to permutated sample
Generator[index]
Batch with entire time-series with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


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

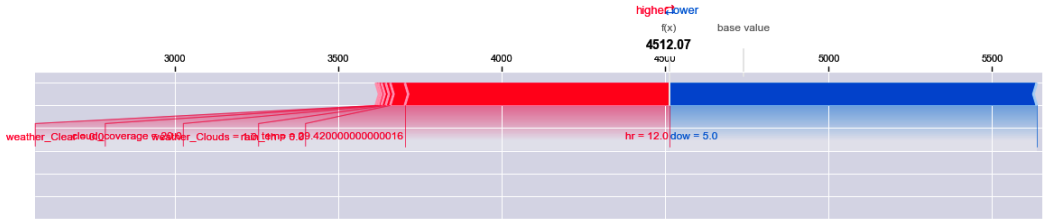
Holiday Afternoon

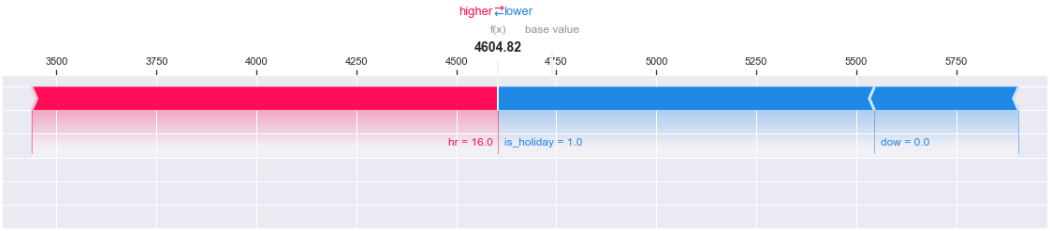


Peak Morning



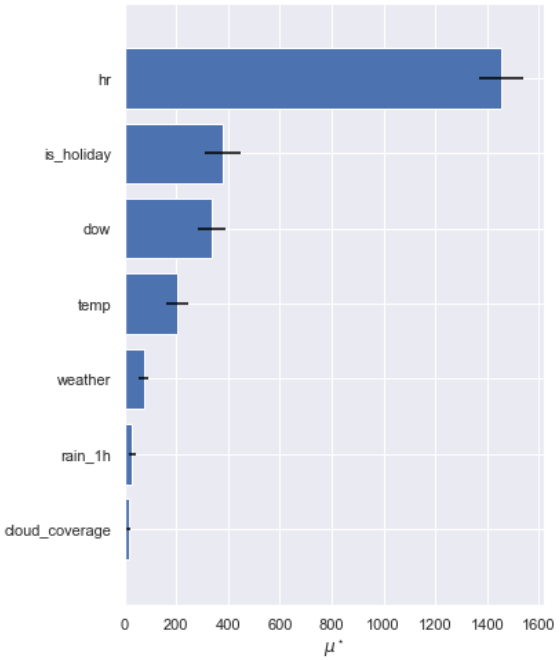
Hot Saturday

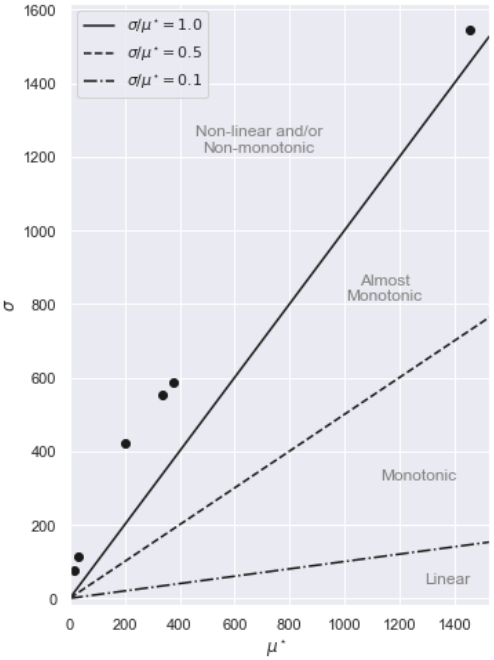




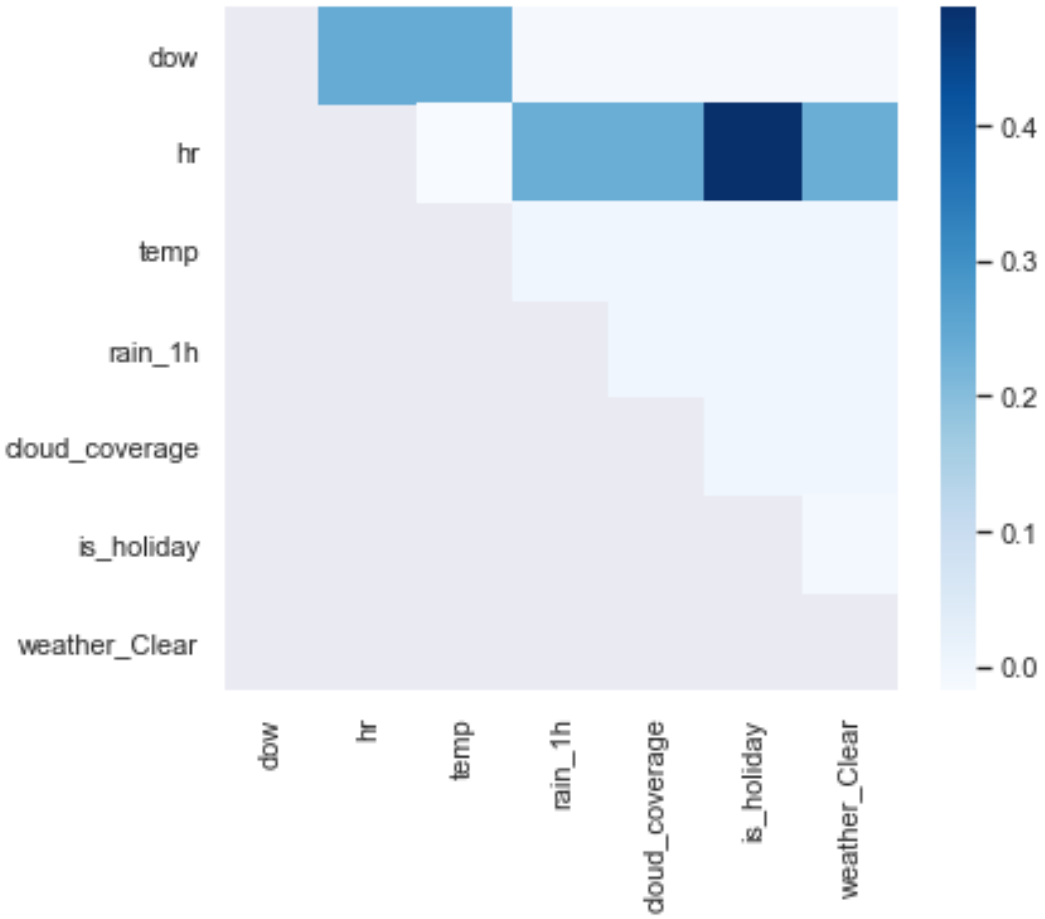
	count	mean	std	min	2.5%	50%	97.5%	max
dow	2232.000000	1.991935	1.415458	0.000000	0.000000	2.000000	4.000000	4.000000
hr	2232.000000	5.500000	7.933780	0.000000	0.000000	2.500000	23.000000	23.000000
temp	2232.000000	16.026438	5.380406	-2.570000	3.178750	16.935000	24.476750	30.458000
rain_1h	2232.000000	0.099628	0.603634	0.000000	0.000000	0.000000	1.451250	10.920000
cloud_coverage	2232.000000	29.178763	36.701417	0.000000	0.000000	1.000000	90.000000	100.000000
is_holiday	2232.000000	0.037634	0.190353	0.000000	0.000000	0.000000	1.000000	1.000000
weather_Clear	2232.000000	0.432348	0.495513	0.000000	0.000000	0.000000	1.000000	1.000000
weather_Clouds	2232.000000	0.207885	0.405885	0.000000	0.000000	0.000000	1.000000	1.000000
weather_Haze	2232.000000	0.010753	0.103159	0.000000	0.000000	0.000000	0.000000	1.000000
weather_Mist	2232.000000	0.104391	0.305835	0.000000	0.000000	0.000000	1.000000	1.000000
weather_Other	2232.000000	0.058244	0.234256	0.000000	0.000000	0.000000	1.000000	1.000000
weather_Rain	2232.000000	0.181452	0.385478	0.000000	0.000000	0.000000	1.000000	1.000000
weather_Snow	2232.000000	0.002240	0.047288	0.000000	0.000000	0.000000	0.000000	1.000000
weather_Unknown	2232.000000	0.002688	0.051789	0.000000	0.000000	0.000000	0.000000	1.000000

	features	μ	μ*	σ
1	hr	-429.300110	1455.506958	1544.544312
5	is_holiday	-345.794861	379.520477	588.769897
0	dow	130.311508	336.568451	554.439819
2	temp	62.087799	202.984299	422.309845
6	weather	nan	75.732839	nan
3	rain_1h	-2.807377	30.730101	113.262093
4	cloud_coverage	9.897467	17.152805	74.319984

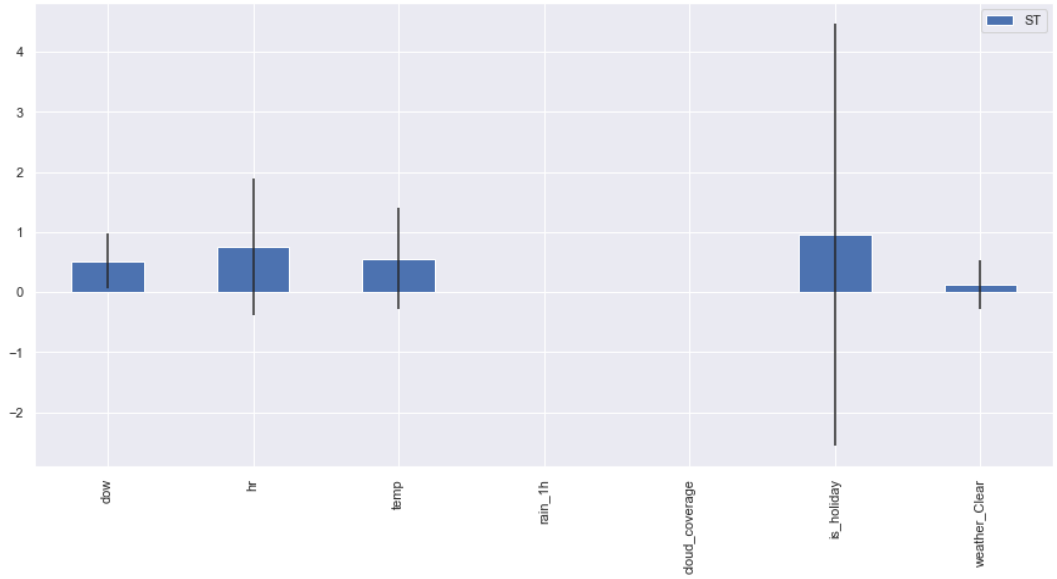


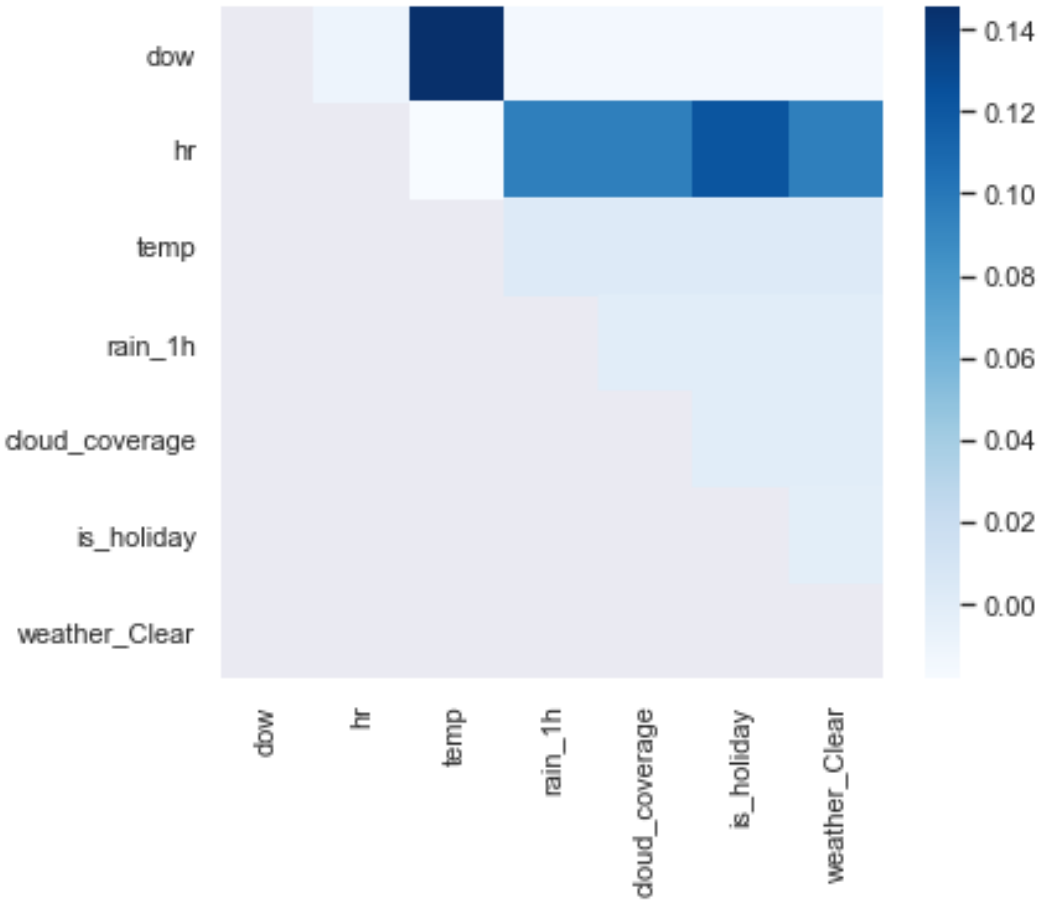


	features	1st	Total	Total Conf	Mean of Input
1	hr	0.009185	0.886824	0.912979	1.495931
2	temp	0.006123	0.506757	0.660847	14.059766
0	dow	0.009185	0.380068	0.366337	1.995599
5	is_holiday	0.003062	0.380068	0.479628	0.498047
6	weather_Clear	-0.003062	0.126689	0.314201	0.499023
3	rain_1h	0.000000	0.000000	0.000000	5.511458
4	cloud_coverage	0.000000	0.000000	0.000000	50.024740

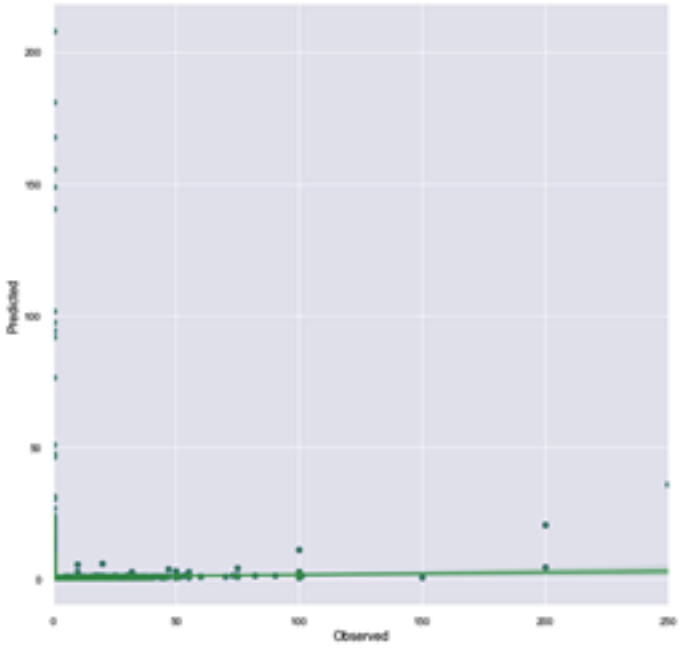


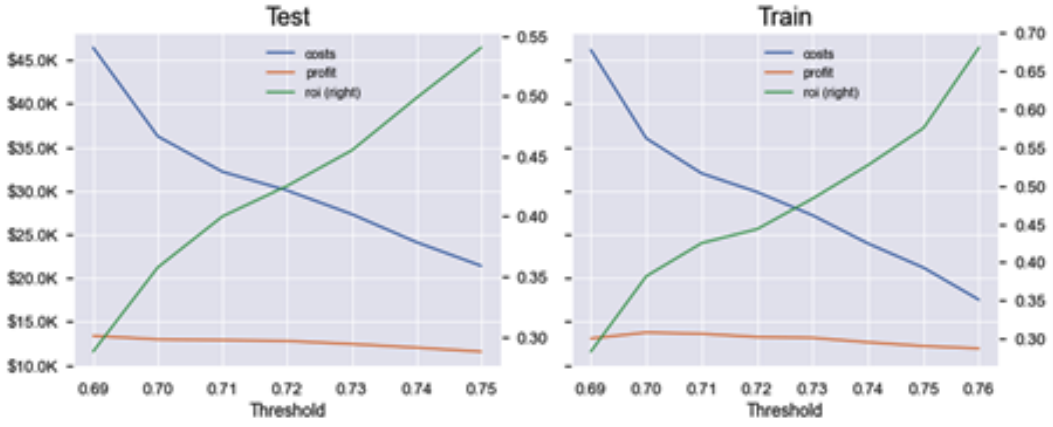
	features	1st	Total	Total Conf	Mean of Input
5	is_holiday	0.000852	0.953684	3.509326	0.498047
1	hr	0.010101	0.748595	1.132665	1.495931
2	temp	0.000677	0.552892	0.843215	14.059766
0	dow	0.009874	0.514826	0.452778	1.995599
6	weather_Clear	-0.002776	0.121222	0.404481	0.499023
4	cloud_coverage	-0.000000	0.000000	0.000000	50.024740
3	rain_1h	-0.000000	0.000000	0.000000	5.511458

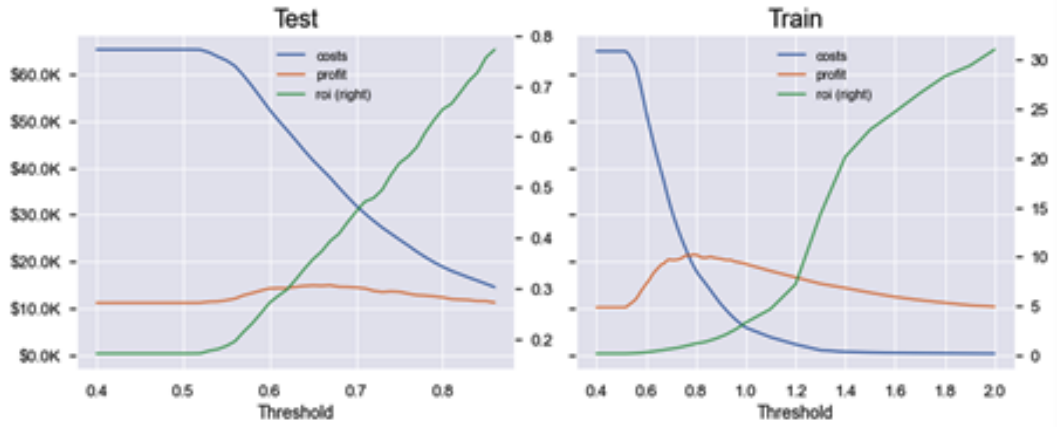

ST
4
3
2
1
0
−1
−2
dow
hr
temp
rain_1h
cloud_coverage
is_holiday
weather_Clear


dow
hr
temp
rain_1h
cloud_coverage
is_holiday
weather_Clear
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00

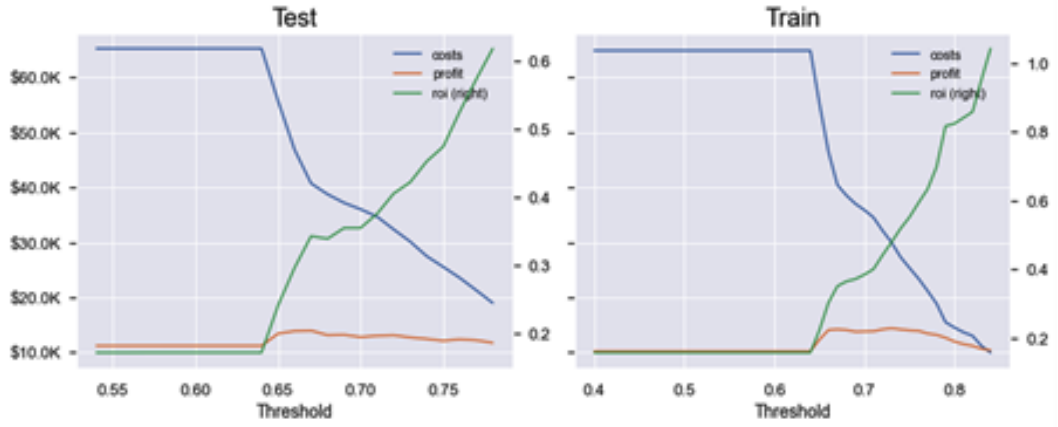
Chapter 10: Feature Selection and Engineering for Interpretability







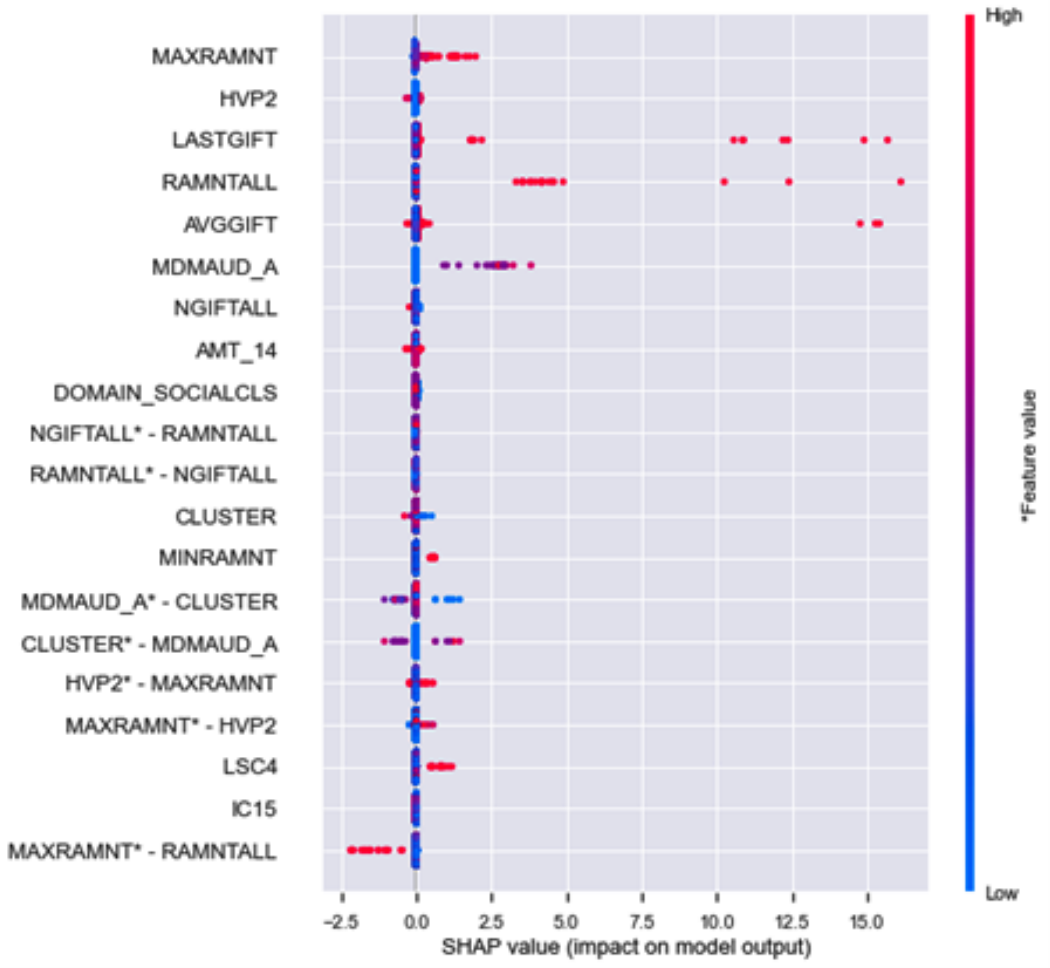
	depth	fs	rmse_train	rmse_test	max_profit_train	max_profit_test	max_roi	min_costs	speed	num_feat
rf_12_all	12	all	3.94	4.69	21521.98	14932.84	0.77	14532.28	2.89	415
rf_11_all	11	all	3.99	4.69	19904.00	15141.86	0.76	14928.04	2.73	398
rf_10_all	10	all	4.05	4.68	18603.92	14987.06	0.78	14396.28	2.43	383
rf_9_all	9	all	4.10	4.68	17453.14	14777.74	0.80	13997.12	2.19	346
rf_8_all	8	all	4.14	4.67	16439.72	14563.04	0.73	15308.84	1.94	315
rf_7_all	7	all	4.18	4.66	15435.32	14187.62	0.66	17164.56	1.71	277
rf_6_all	6	all	4.23	4.65	14651.12	13845.27	0.59	19305.20	1.41	240
rf_5_all	5	all	4.27	4.64	14242.32	13752.13	0.59	19199.12	1.22	201
rf_4_all	4	all	4.32	4.64	13715.90	13261.88	0.53	22392.40	1.00	160

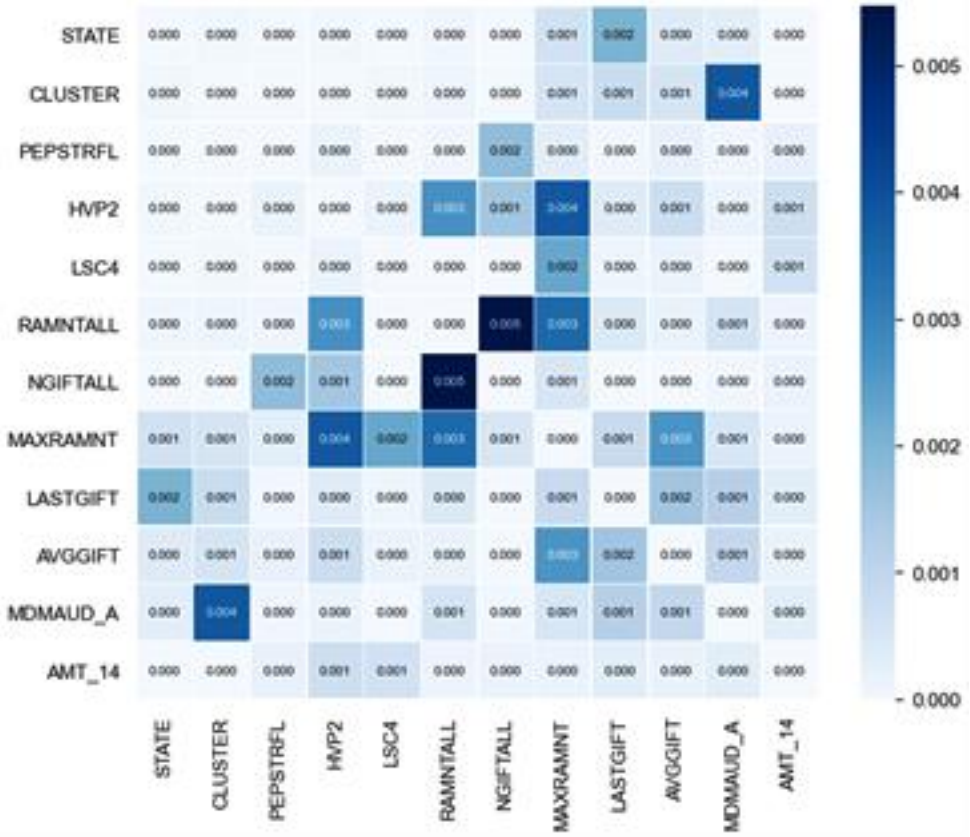


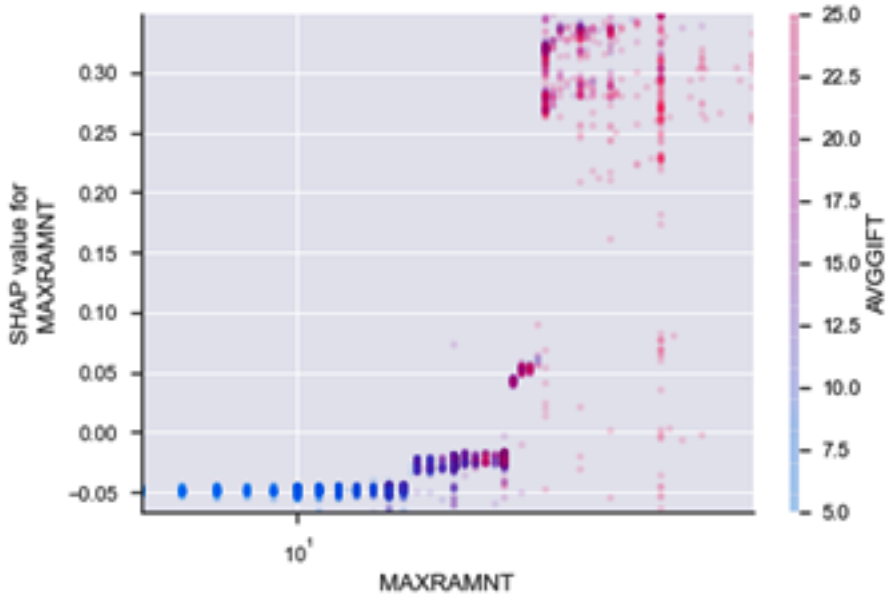
	depth	fs	rmse_train	rmse_test	max_profit_train	max_profit_test	max_roi	min_costs	speed	total_feat	num_feat
rf_11_all	11	all	3.99	4.69	19904.00	15141.86	0.76	14928.04	2.73	435	398
rf_10_all	10	all	4.05	4.68	18603.92	14987.06	0.78	14396.28	2.43	435	383
rf_12_all	12	all	3.94	4.69	21521.98	14932.84	0.77	14532.28	2.89	435	415
rf_11_f-corr	11	f-corr	3.98	4.67	19923.84	14894.94	0.77	14592.80	2.47	419	404
rf_9_all	9	all	4.10	4.68	17453.14	14777.74	0.80	13997.12	2.19	435	346
rf_8_all	8	all	4.14	4.67	16439.72	14563.04	0.73	15308.84	1.94	435	315
rf_5_f-mic	5	f-mic	4.31	4.57	14983.34	14481.39	0.62	18971.32	0.39	160	103
rf_7_all	7	all	4.18	4.66	15435.32	14187.62	0.66	17164.56	1.71	435	277
rf_6_all	6	all	4.23	4.65	14651.12	13845.27	0.59	19305.20	1.41	435	240
rf_5_all	5	all	4.27	4.64	14242.32	13752.13	0.59	19199.12	1.22	435	201
rf_4_all	4	all	4.32	4.64	13715.90	13261.88	0.53	22392.40	1.00	435	160

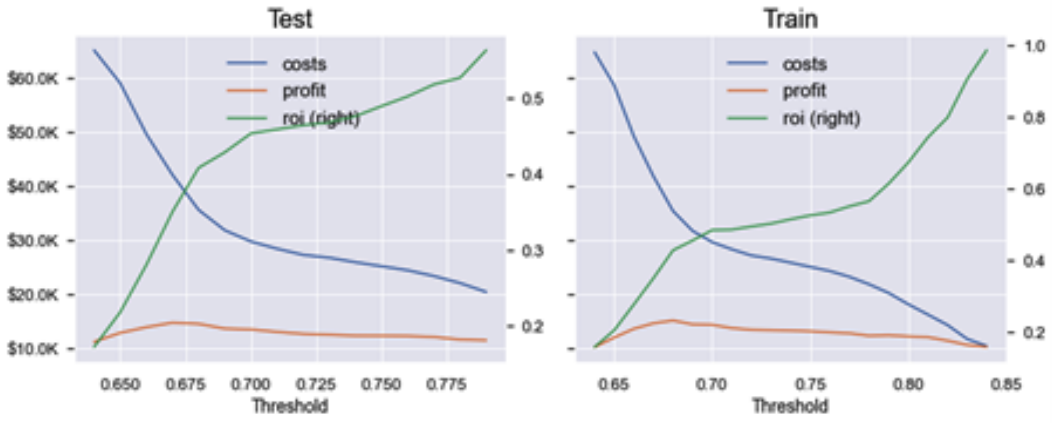
	depth	fs	rmse_train	rmse_test	max_profit_train	max_profit_test	max_roi	min_costs	speed	total_feat	num_feat
rf_11_all	11	all	3.99	4.69	19904.00	15141.86	0.76	14928.04	2.73	435	398
rf_10_all	10	all	4.05	4.68	18603.92	14987.06	0.78	14396.28	2.43	435	383
rf_12_all	12	all	3.94	4.69	21521.98	14932.84	0.77	14532.28	2.89	435	415
rf_11_f-corr	11	f-corr	3.98	4.67	19923.84	14894.94	0.77	14592.80	2.47	419	404
rf_9_all	9	all	4.10	4.68	17453.14	14777.74	0.80	13997.12	2.19	435	346
rf_5_e-llarsic	5	e-llarsic	4.28	4.45	15168.46	14768.37	0.56	20441.48	0.32	111	87
rf_8_all	8	all	4.14	4.67	16439.72	14563.04	0.73	15308.84	1.94	435	315
rf_5_f-mic	5	f-mic	4.31	4.57	14983.34	14481.39	0.62	18971.32	0.39	160	103
rf_6_h-rfe-lda	6	h-rfe-lda	4.25	4.48	15329.72	14351.74	0.71	15824.28	0.61	183	129
rf_6_e-logl2	6	e-logl2	4.28	4.60	15353.44	14199.90	0.67	16904.12	0.32	87	84
rf_7_all	7	all	4.18	4.66	15435.32	14187.62	0.66	17164.56	1.71	435	277
rf_6_all	6	all	4.23	4.65	14651.12	13845.27	0.59	19305.20	1.41	435	240
rf_5_all	5	all	4.27	4.64	14242.32	13752.13	0.59	19199.12	1.22	435	201
rf_4_e-llars	4	e-llars	4.36	4.45	14014.10	13633.19	0.52	22906.48	0.06	8	8
rf_6_h-rfe-rf	6	h-rfe-rf	4.40	4.78	13202.61	13347.15	0.41	28596.04	0.08	1	1
rf_4_all	4	all	4.32	4.64	13715.90	13261.88	0.53	22392.40	1.00	435	160
rf_3_e-lasso	3	e-lasso	4.46	4.49	14166.64	12930.30	0.51	22248.92	0.05	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	15168.46	14768.37	0.56	20441.48	0.32	111	87
rf_5_f-mic	5	f-mic	4.31	4.57	14983.34	14481.39	0.62	18971.32	0.39	160	103
rf_6_h-rfe-lda	6	h-rfe-lda	4.25	4.48	15329.72	14351.74	0.71	15824.28	0.61	183	129
rf_6_a-shap	6	a-shap	4.23	4.52	15263.60	14282.20	0.61	18767.32	0.50	150	135
rf_5_a-ga-rf	5	a-ga-rf	4.39	4.45	14274.52	14220.53	0.69	13237.56	0.07	63	63
rf_6_e-logl2	6	e-logl2	4.28	4.60	15353.44	14199.90	0.67	16904.12	0.32	87	84
rf_6_all	6	all	4.23	4.65	14651.12	13845.27	0.59	19305.20	1.41	435	240
rf_5_w-sfs-lda	5	w-sfs-lda	4.43	4.63	14377.13	13801.55	0.51	22508.95	0.11	27	27
rf_5_all	5	all	4.27	4.64	14242.32	13752.13	0.59	19199.12	1.22	435	201
rf_4_e-llars	4	e-llars	4.36	4.45	14014.10	13633.19	0.52	22906.48	0.06	8	8
rf_6_a-pca	6	a-pca	4.30	4.46	14353.54	13351.57	0.47	23901.32	0.54	150	126
rf_6_h-rfe-rf	6	h-rfe-rf	4.40	4.78	13202.61	13347.15	0.41	28596.04	0.08	1	1
rf_4_all	4	all	4.32	4.64	13715.90	13261.88	0.53	22392.40	1.00	435	160
rf_6_w-sbs-et	6	w-sbs-et	4.34	4.53	14222.49	13119.17	0.71	14711.66	0.45	135	123
rf_3_e-lasso	3	e-lasso	4.46	4.49	14166.64	12930.30	0.51	22248.92	0.05	7	7

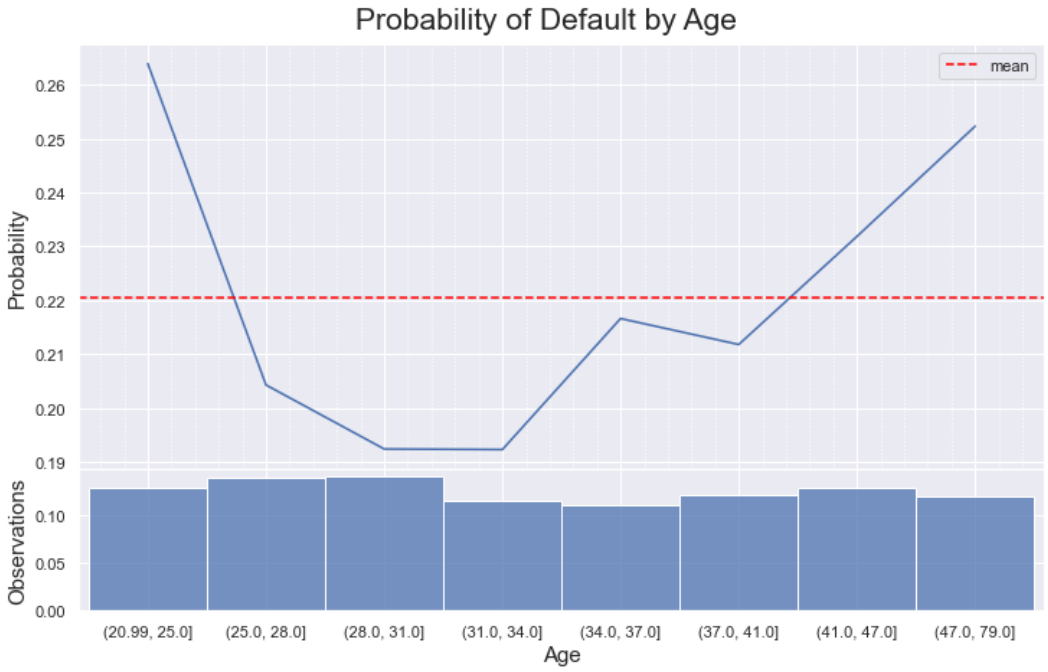

MAXRAMNT
HVP2
LASTGIFT
RAMNTALL
AVGGIFT
MDMAUD_A
NGIFTALL
AMT_14
DOMAIN_SOCIALCLS
NGIFTALL* - RAMNTALL
RAMNTALL* - NGIFTALL
CLUSTER
MINRAMNT
MDMAUD_A* - CLUSTER
CLUSTER* - MDMAUD_A
HVP2* - MAXRAMNT
MAXRAMNT* - HVP2
LSC4
IC15
MAXRAMNT* - RAMNTALL
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
SHAP value (impact on model output)
High
*Feature value
Low

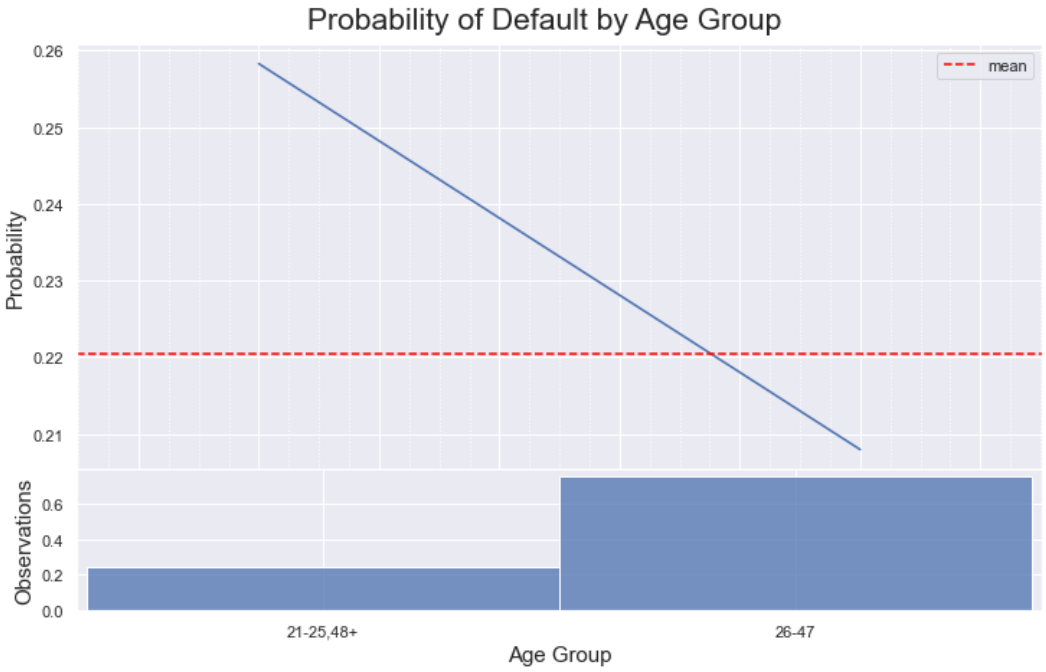

STATE
CLUSTER
PEPSTRFL
HVP2
LSC4
RAMNTALL
NGIFTALL
MAXRAMNT
LASTGIFT
AVGGIFT
MDMAUD_A
AMT_14
0.005
0.004
0.003
0.002
0.001
0.000

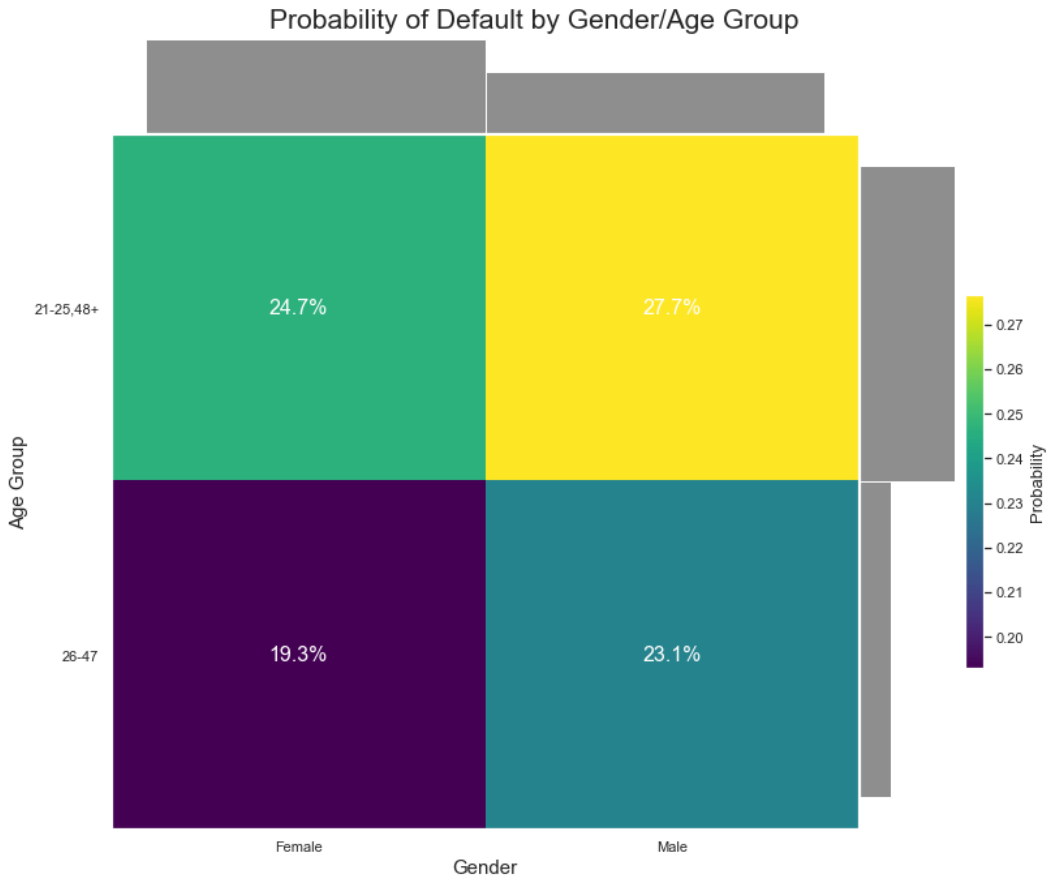

SHAP value for MAXRAMNT
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
10^1
MAXRAMNT
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
AVGGIFT

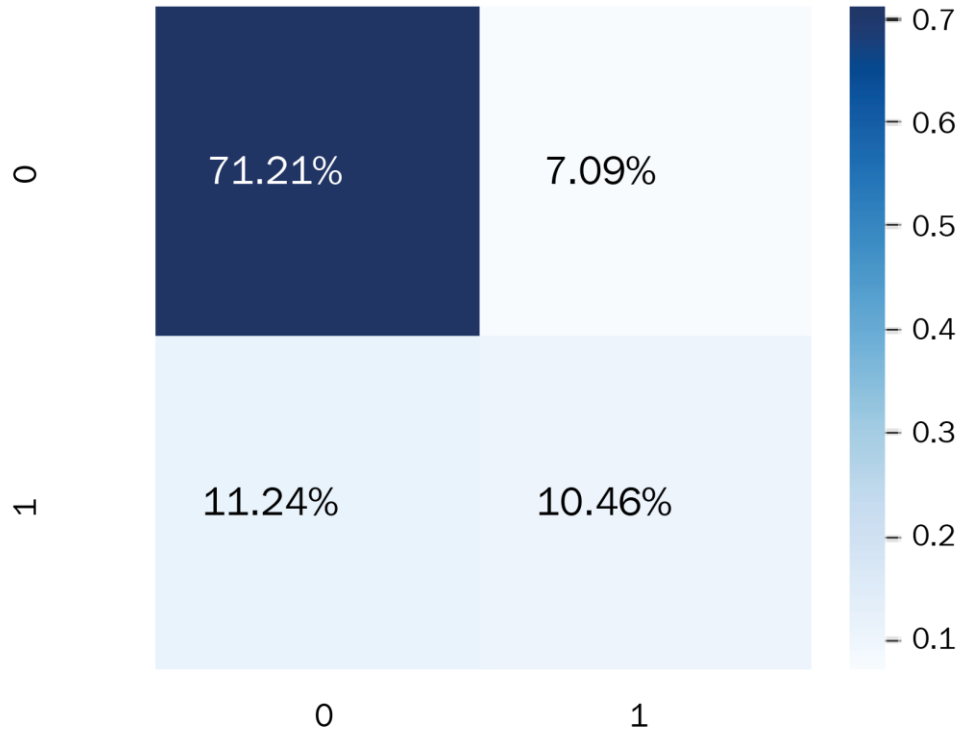

Test
costs
profit
roi (right)
$60.0K
$50.0K
$40.0K
$30.0K
$20.0K
$10.0K
0.5
0.4
0.3
0.2
0.650
0.675
0.700
0.725
0.750
0.775
Threshold
Train
costs
profit
roi (right)
1.0
0.8
0.6
0.4
0.2
0.65
0.70
0.75
0.80
0.85
Threshold

Chapter 11: Bias Mitigation and Causal Inference Methods

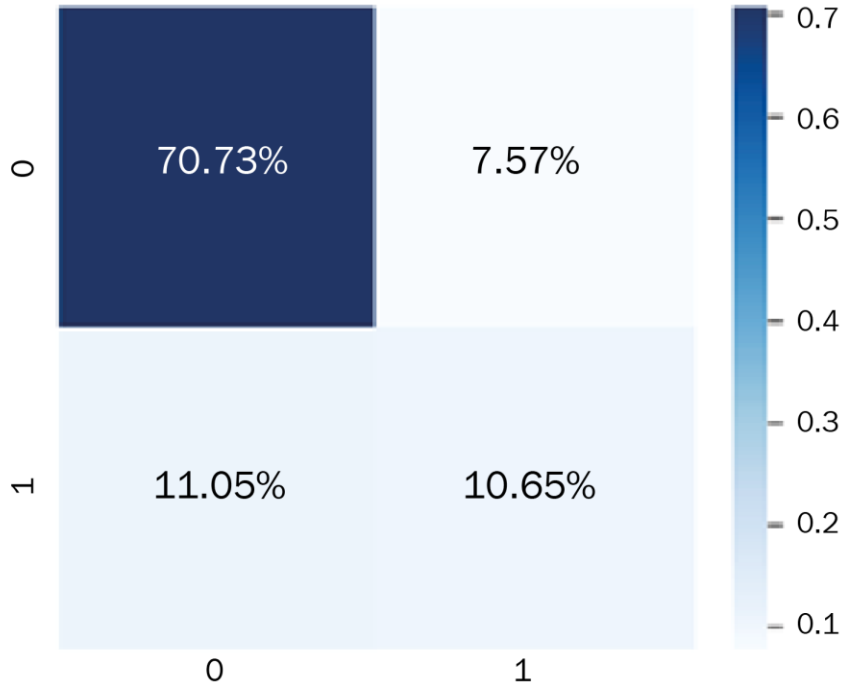



Probability of Default by Age Group
mean
Probability
0.26
0.25
0.24
0.23
0.22
0.21
Observations
0.6
0.4
0.2
0.0
21-25,48+
26-47
Age Group

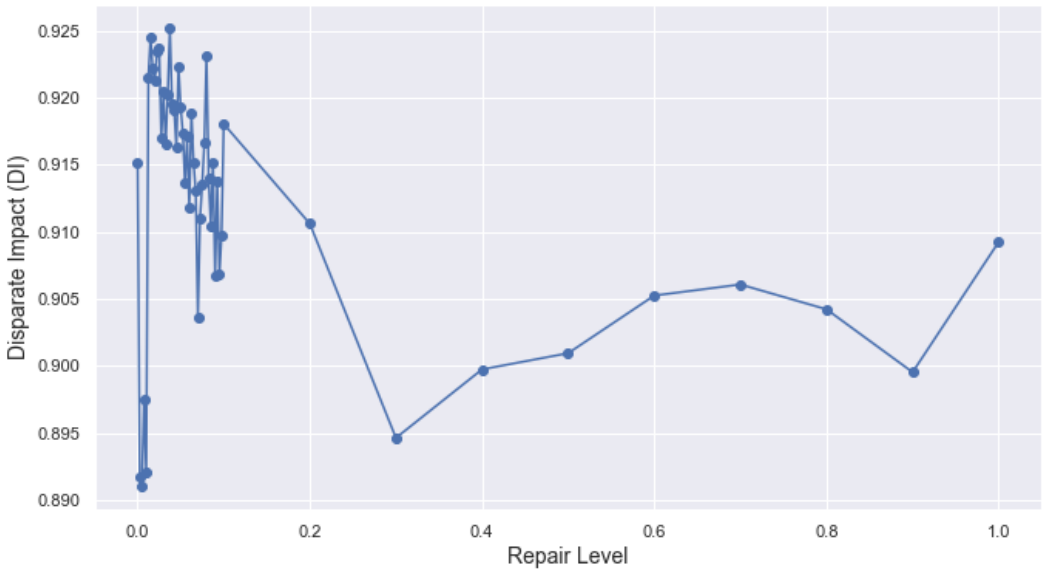

Probability of Default by Gender/Age Group
24.7%
27.7%
19.3%
23.1%
21-25,48+
26-47
Age Group
Female
Male
Gender
Probability
0.27
0.26
0.25
0.24
0.23
0.22
0.21
0.20



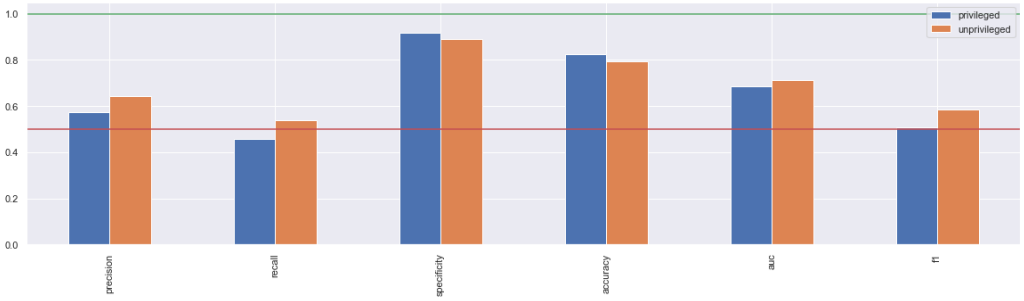
1. Accuracy_train: 0.8255 Accuracy_test: 0.8167
2. Precision_test: 0.5961 Recall_test: 0.4820
3. ROC-AUC_test: 0.7901 F1_test: 0.5330 MCC_test: 0.4243

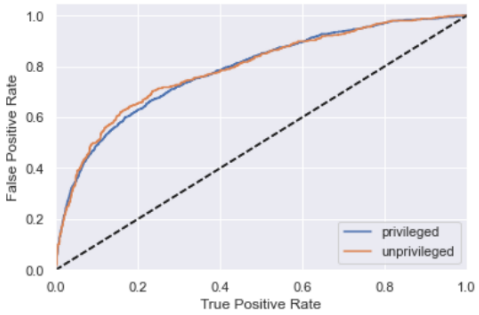


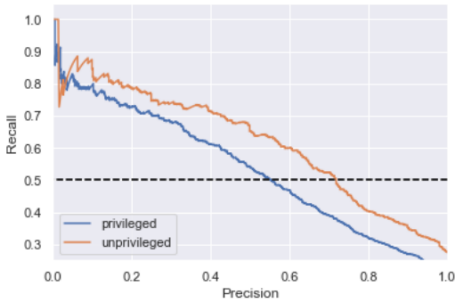
1. Accuracy_train: 0.8240 Accuracy_test: 0.8138
2. Precision_test: 0.5847 Recall_test: 0.4908
3. ROC-AUC_test: 0.7886 F1_test: 0.5337 MCC_test: 0.4210

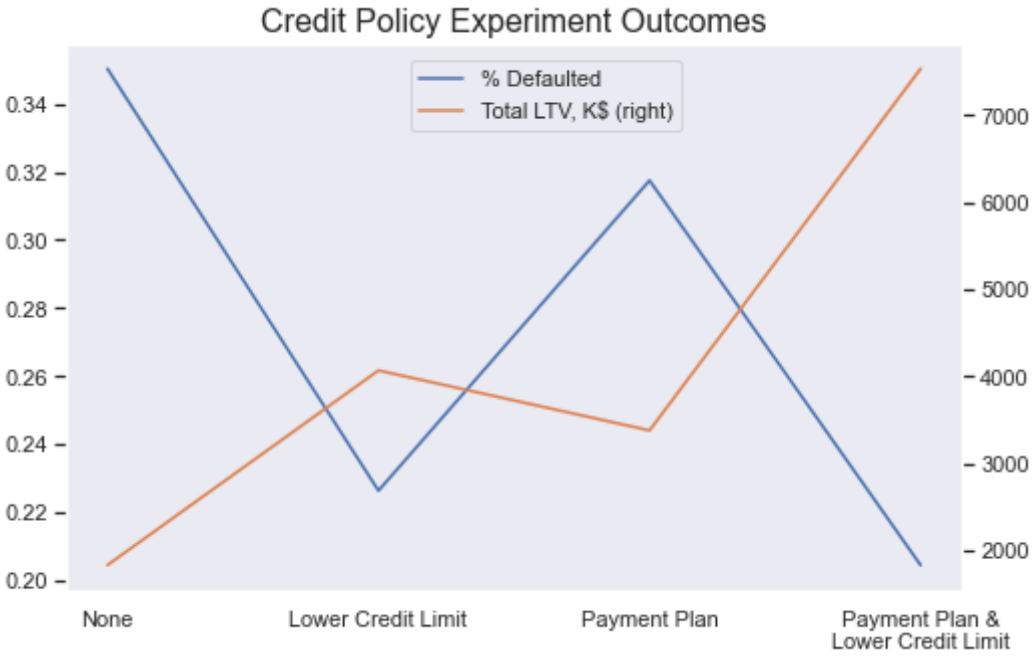


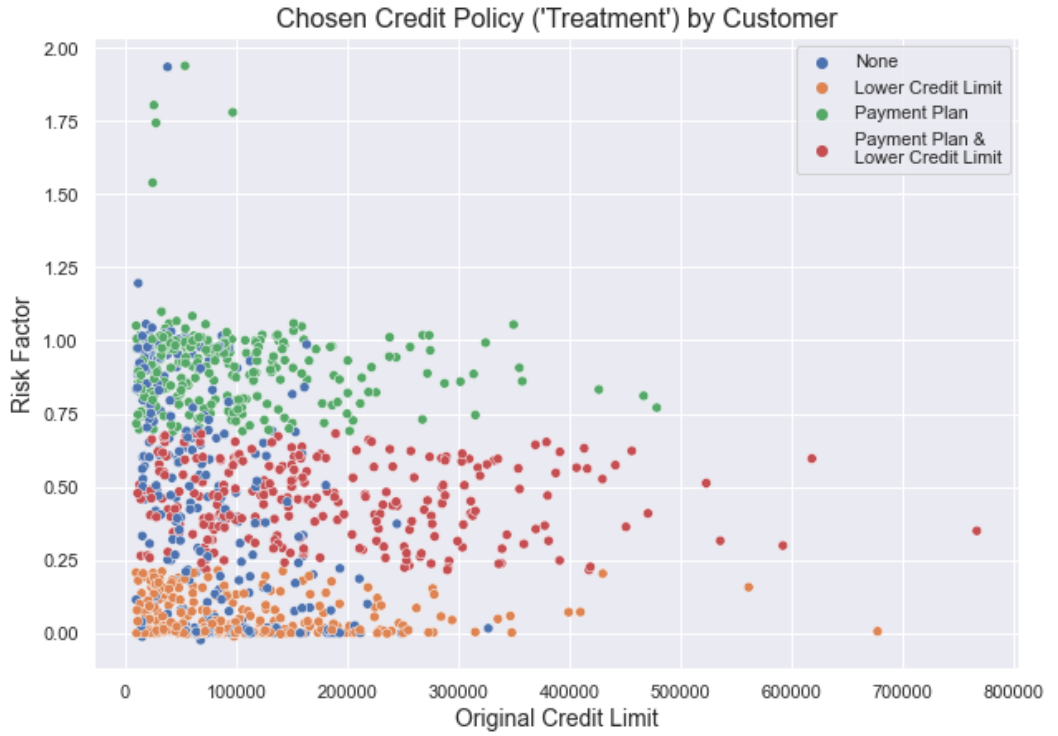
	accuracy_train	accuracy_test	f1_test	mcc_test	SPD	DI	AOD	EOD	DFBA
dt_2_gf	0.8214	0.8262	0.4812	0.4135	-0.0548	0.9388	-0.0430	-0.0216	0.2521
lgb_0_base	0.8255	0.8167	0.5330	0.4243	-0.0679	0.9193	-0.0550	-0.0265	0.2328
lgb_1_rw	0.8240	0.8138	0.5337	0.4210	-0.0371	0.9552	-0.0171	-0.0018	0.0349
lgb_1_dir	0.8237	0.8129	0.5301	0.4171	-0.0624	0.9252	-0.0493	-0.0214	0.2545
lgb_3_epp	0.8255	0.8101	0.5152	0.4025	-0.0260	0.9688	0.0022	-0.0021	0.0031
lgb_3_cpp	0.8255	0.2622	0.2129	-0.3055	-0.0711	0.7609	-0.0635	-0.1262	0.0432
log_2_pr	0.1912	0.1873	0.2844	-0.3363	0.0520	1.7627	0.0498	0.0235	0.3454

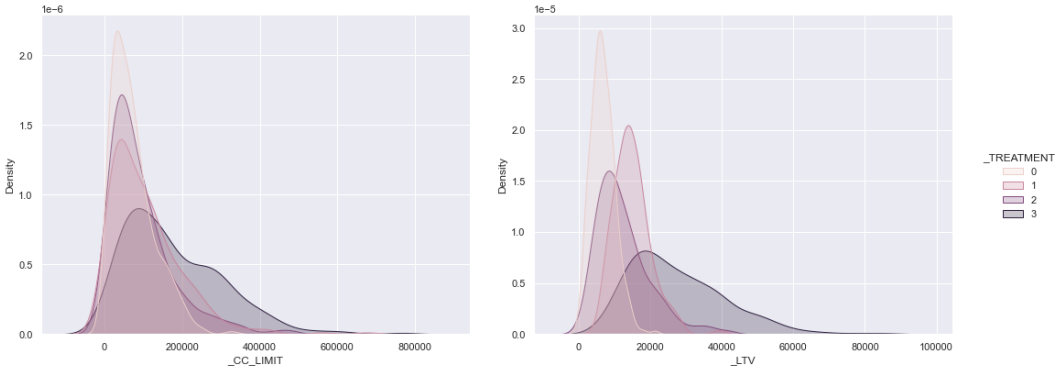

privileged
unprivileged
precision
recall
specificity
accuracy
auc
f1

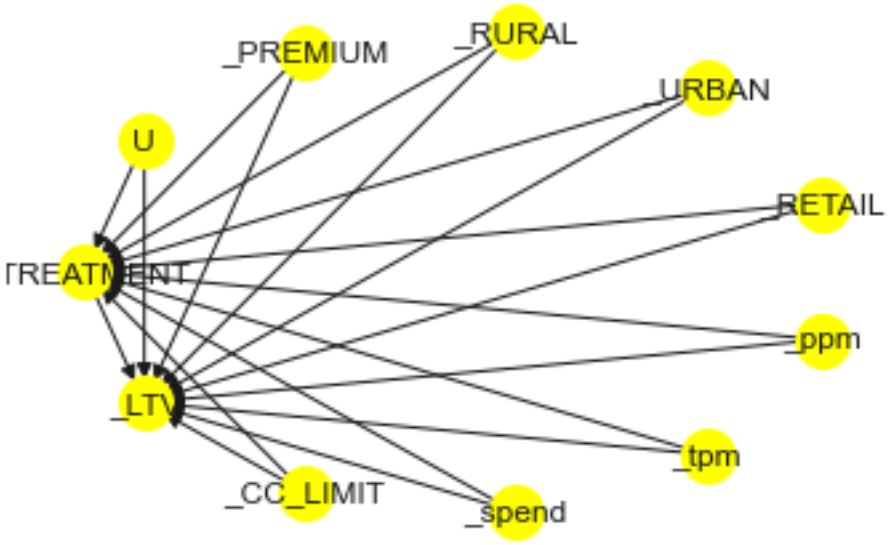

False Positive Rate
True Positive Rate
privileged
unprivileged


Recall
Precision
privileged
unprivileged


Credit Policy Experiment Outcomes
% Defaulted
Total LTV, K$ (right)
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


1e−6
Density
_CC_LIMIT
1e−5
Density
_LTV
_TREATMENT
0
1
2
3



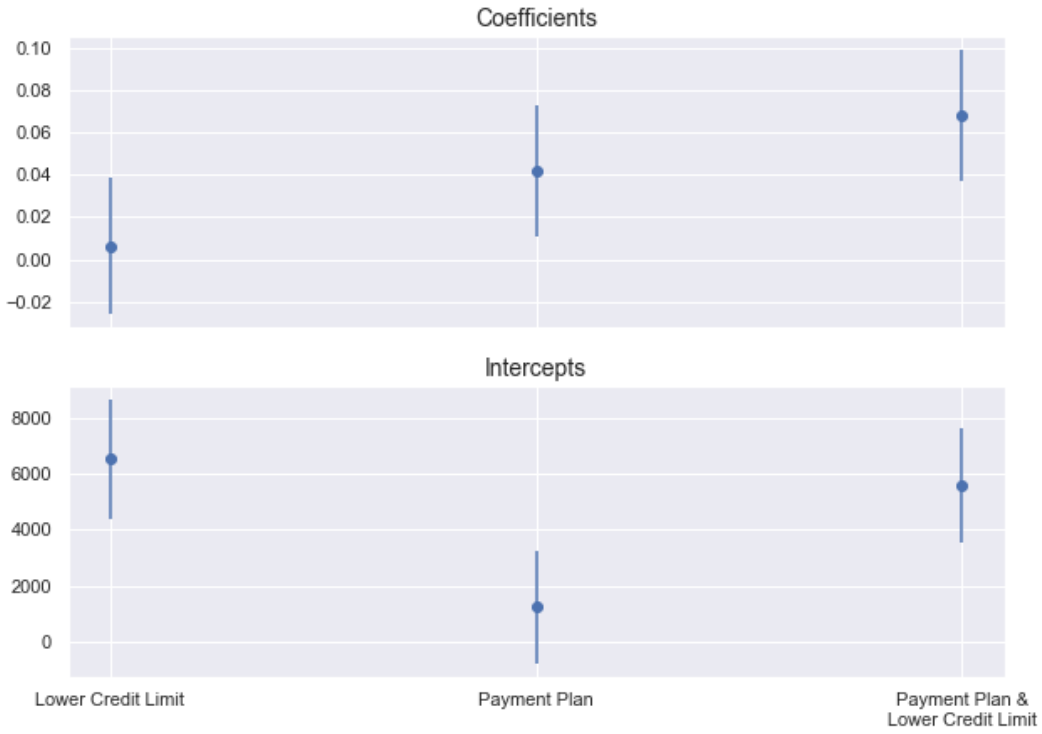
Treatment: Payment Plan &
Lower Credit Limit

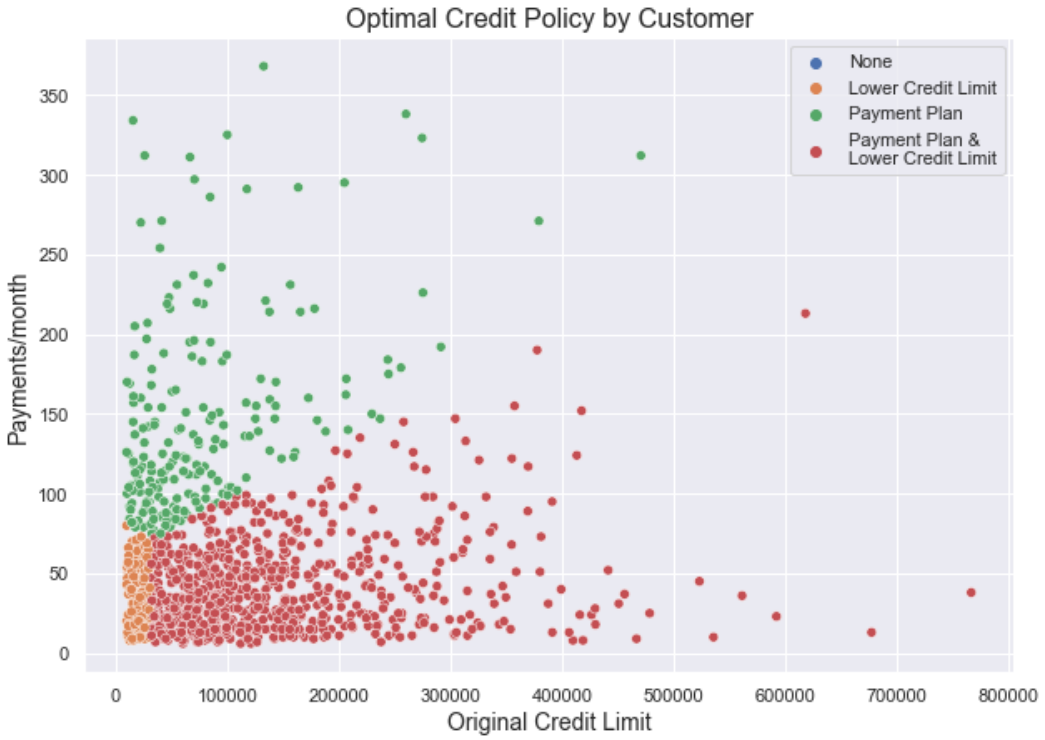
Coefficient Results

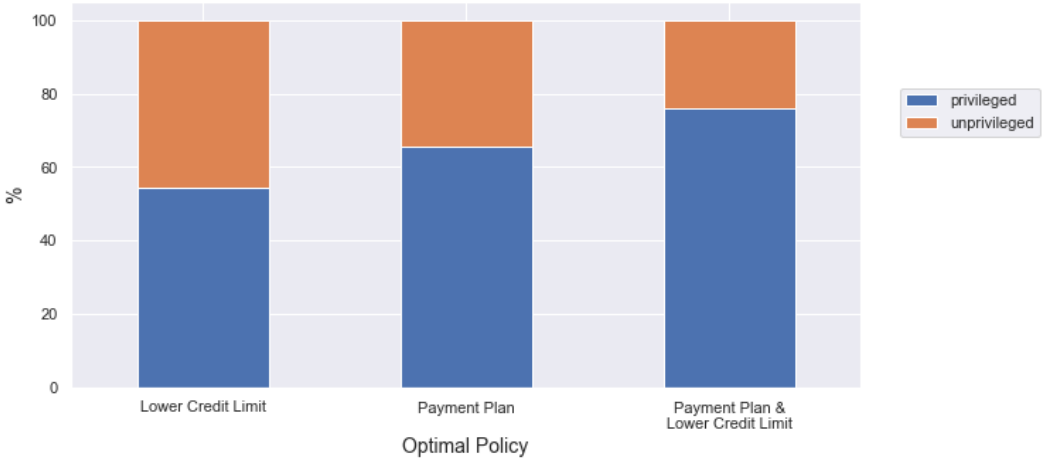
	point_estimate	stderr	zstat	pvalue	ci_lower	ci_upper
X0	0.095	0.03	3.171	0.002	0.046	0.145

CATE Intercept Results

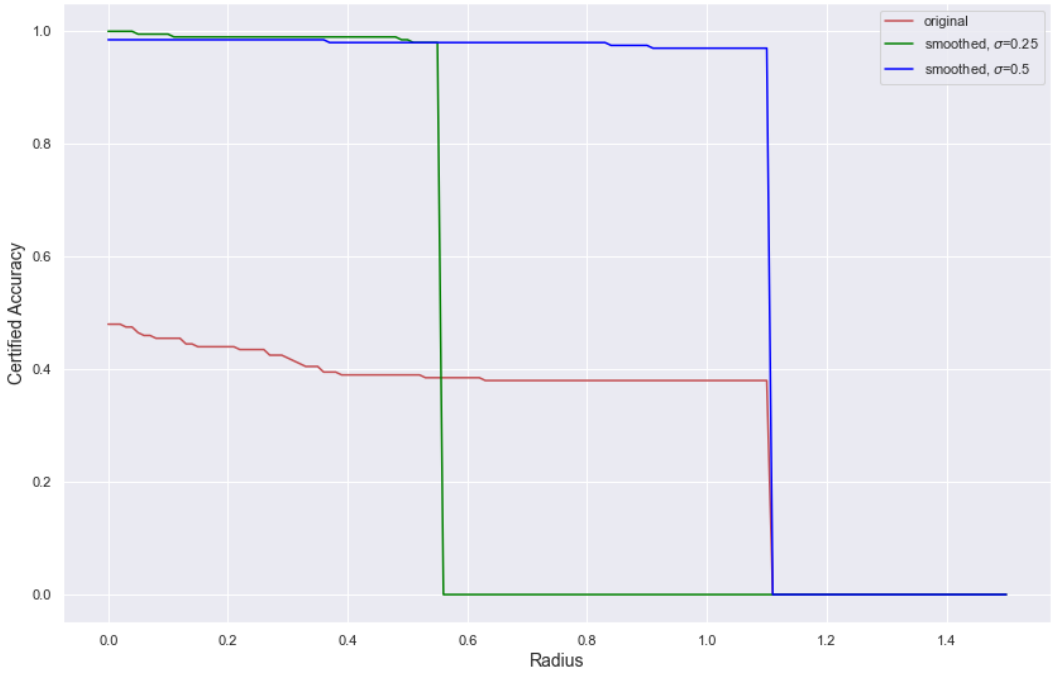
	point_estimate	stderr	zstat	pvalue	ci_lower	ci_upper
cate_intercept	4238.783	1907.747	2.222	0.026	1100.818	7376.748


Coefficients
0.10
0.08
0.06
0.04
0.02
0.00
−0.02
Intercepts
8000
6000
4000
2000
0
Lower Credit Limit
Payment Plan
Payment Plan &
Lower Credit Limit

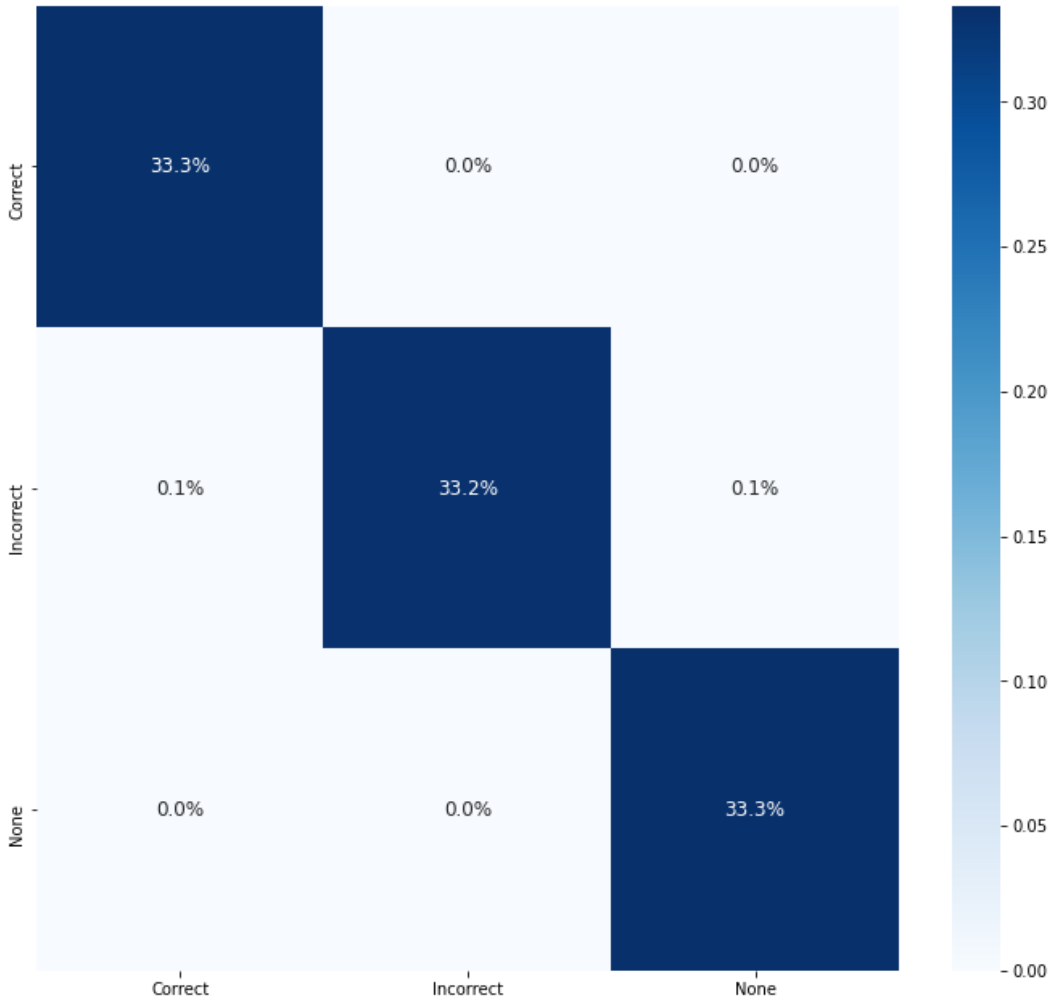

Optimal Credit Policy by Customer
None
Lower Credit Limit
Payment Plan
Payment Plan &
Lower Credit Limit
Payments/month
0
50
100
150
200
250
300
350
0
100000
200000
300000
400000
500000
600000
700000
800000
Original Credit Limit


%
0
20
40
60
80
100
privileged
unprivileged
Lower Credit Limit
Payment Plan
Payment Plan &
Lower Credit Limit
Optimal Policy

Chapter 13: Adversarial Robustness



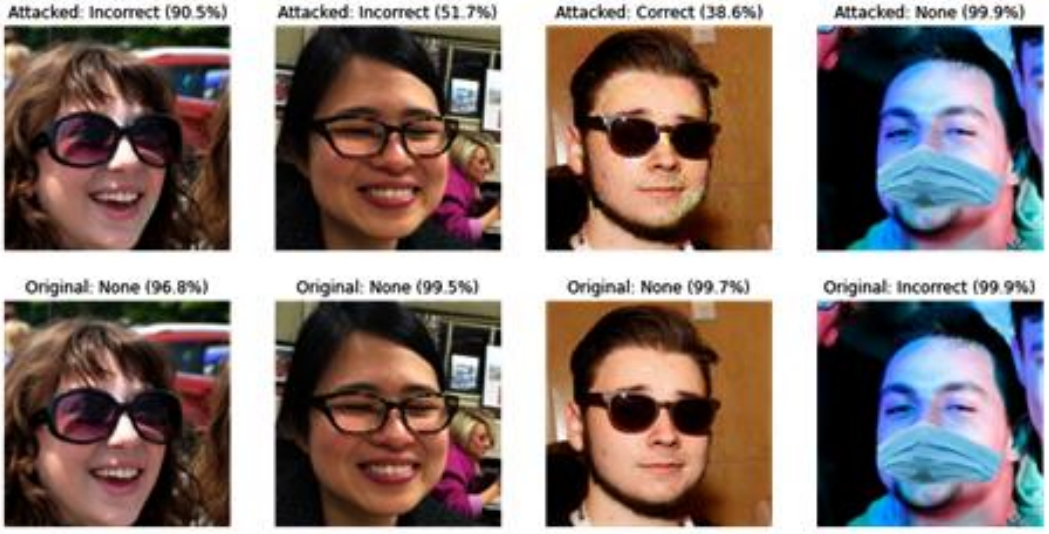

None
Correct
None
Incorrect
Correct
Correct
Correct
None
None
Correct
None
Correct
Incorrect
None
None
Incorrect
None
Incorrect
Correct
Incorrect

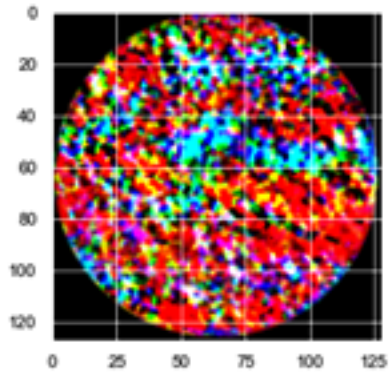


		Goal		
		Espionage	**Sabotage**	**Fraud**
Stage	**Training**	Inference (by poisoning)	Trojaning	
			Poisoning	
			Backdooring	
	Production	Inference	Reprogramming	
			Evasion	

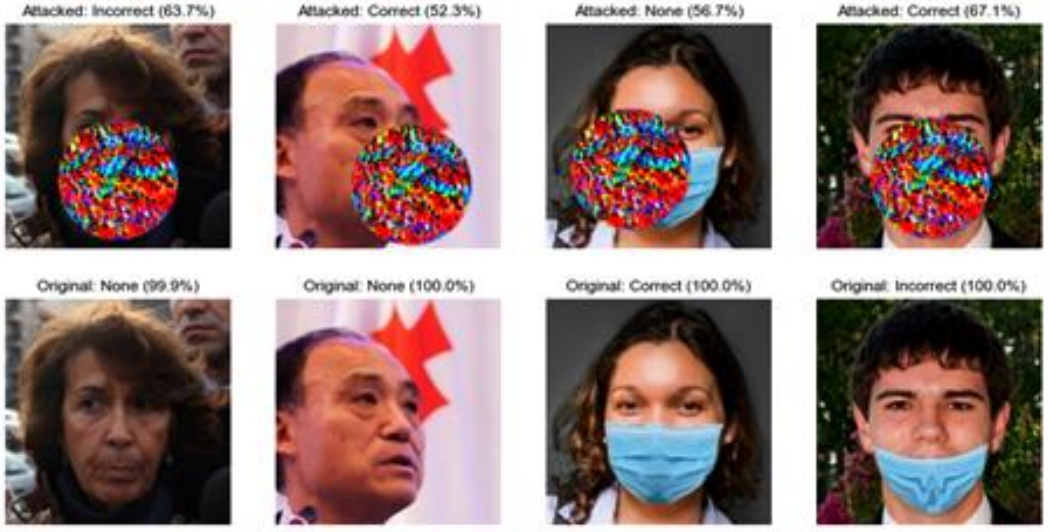
FSGM Attack Average Perturbation: 0.092

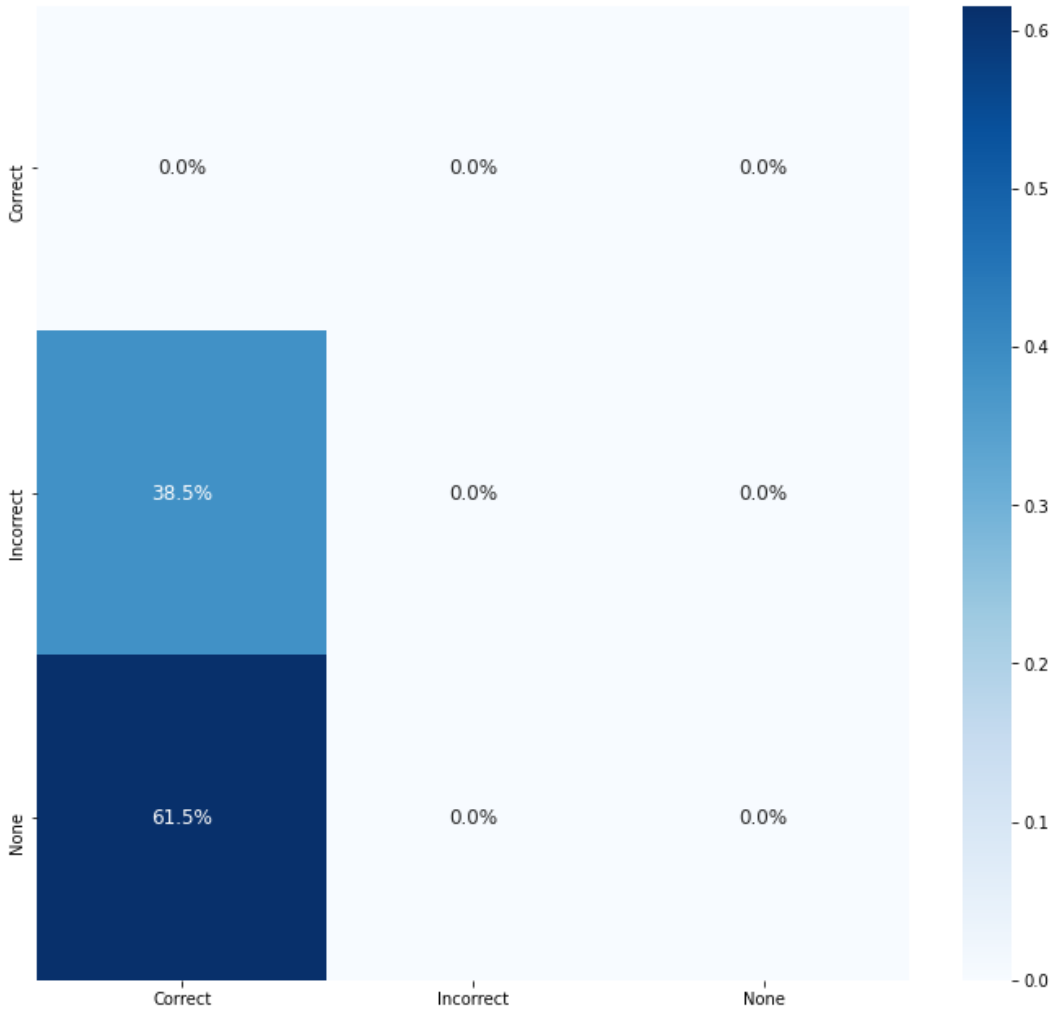
C&W Inf Attack Average Perturbation: 0.003

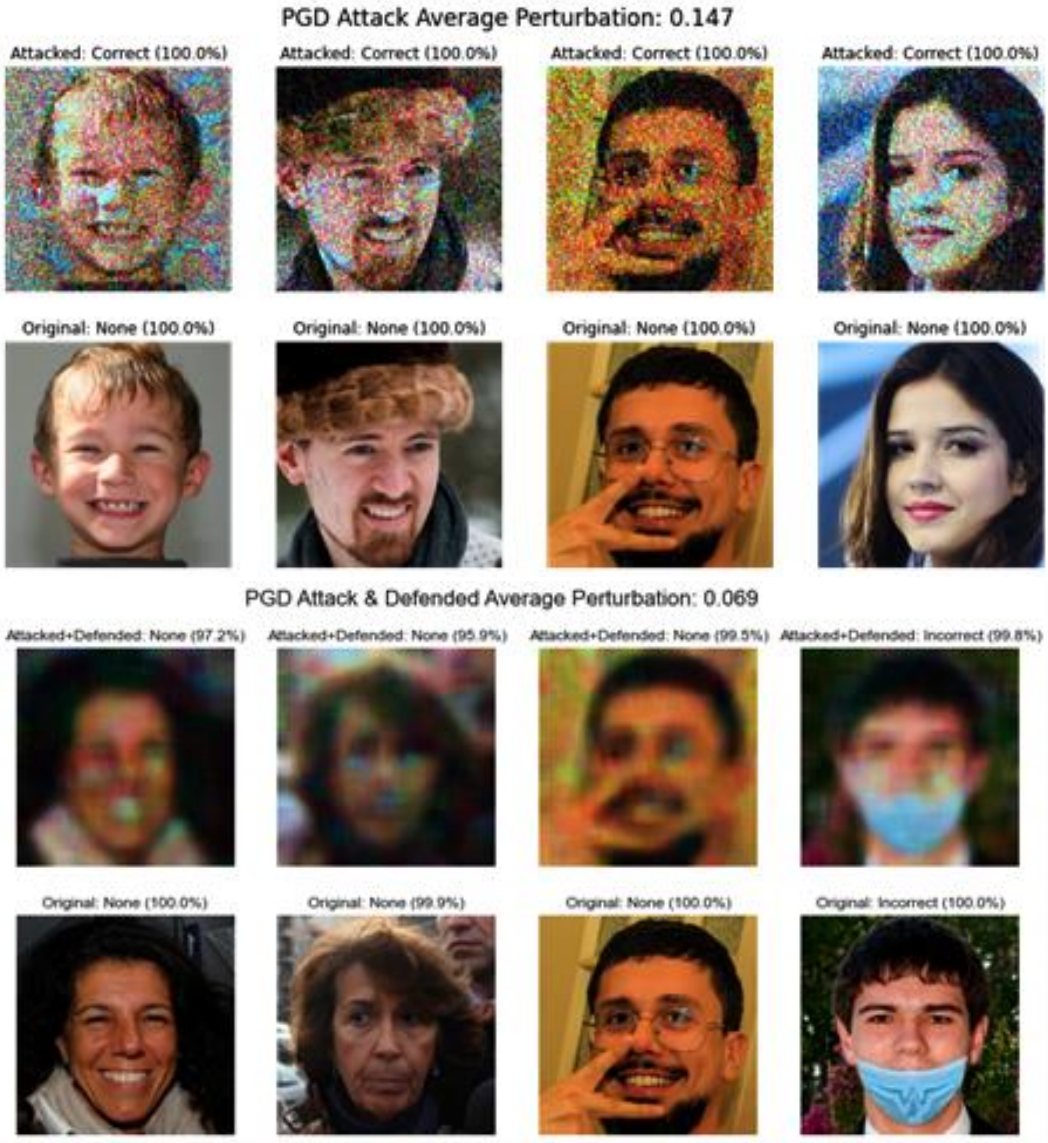


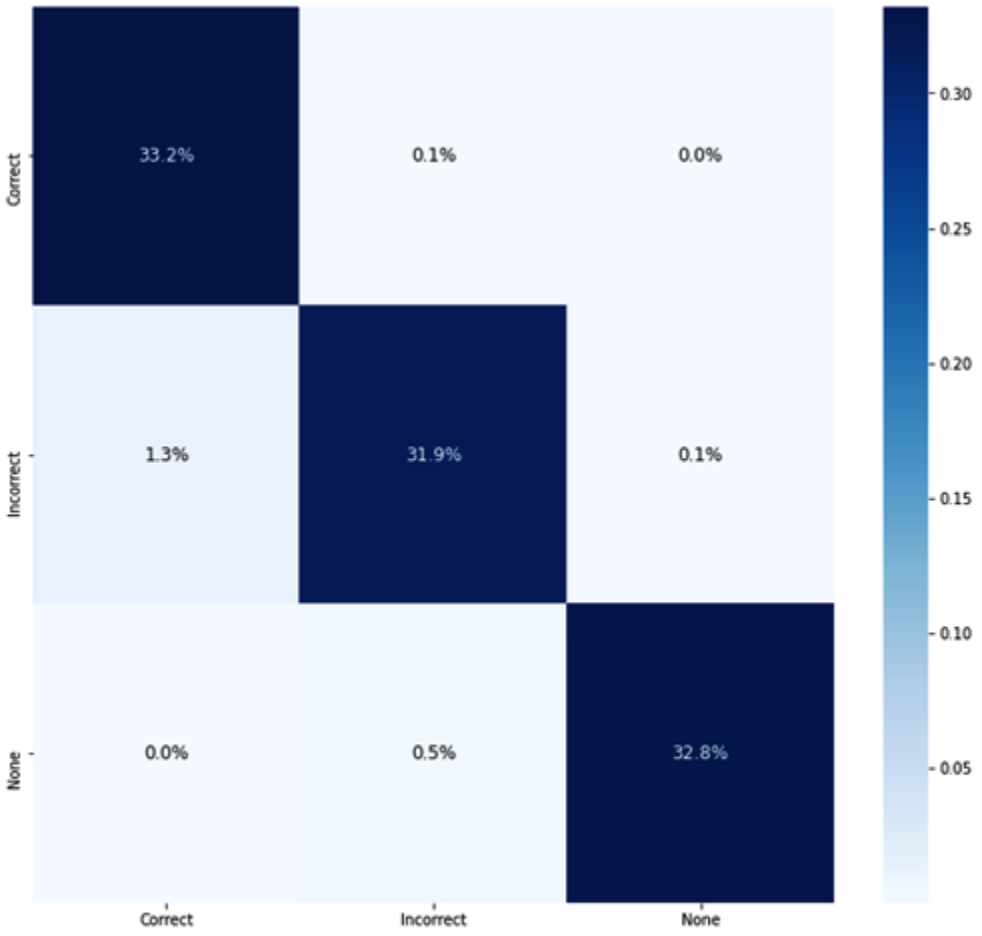


AP Attack Average Perturbation: 0.080

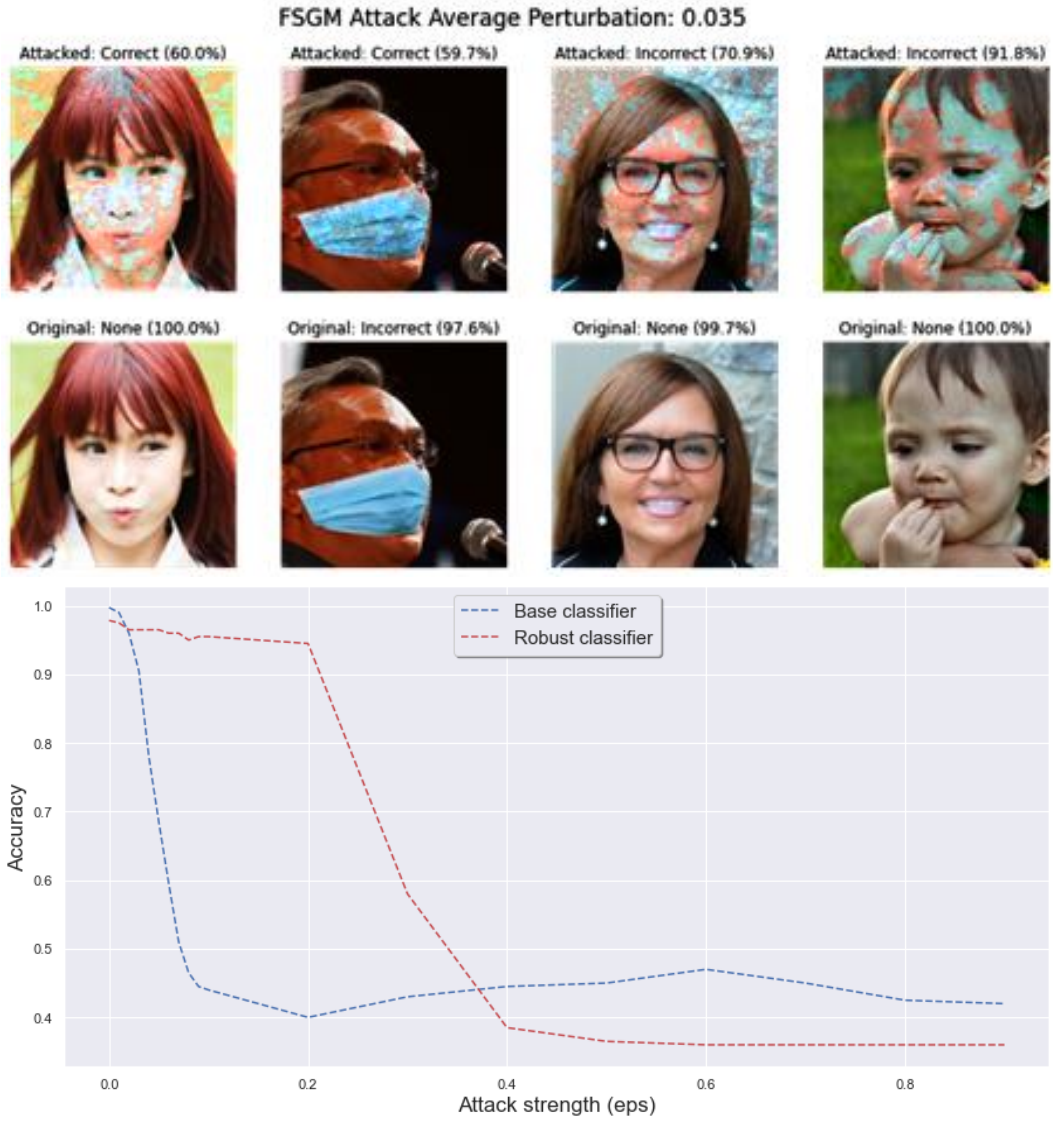



Correct
Incorrect
None
0.0%
0.0%
0.0%
38.5%
0.0%
0.0%
61.5%
0.0%
0.0%
Correct
Incorrect
None
0.6
0.5
0.4
0.3
0.2
0.1
0.0


PGD Attack Average Perturbation: 0.147
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 & Defended Average Perturbation: 0.069
Attacked+Defended: None (97.2%)
Attacked+Defended: None (95.9%)
Attacked+Defended: None (99.5%)
Attacked+Defended: Incorrect (99.8%)
Original: None (100.0%)
Original: None (99.9%)
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


FSGM Attack Average Perturbation: 0.035
Attacked: Correct (60.0%)
Attacked: Correct (59.7%)
Attacked: Incorrect (70.9%)
Attacked: Incorrect (91.8%)
Original: None (100.0%)
Original: Incorrect (97.6%)
Original: None (99.7%)
Original: None (100.0%)
Base classifier
Robust classifier
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.0
0.2
0.4
0.6
0.8
Attack strength (eps)

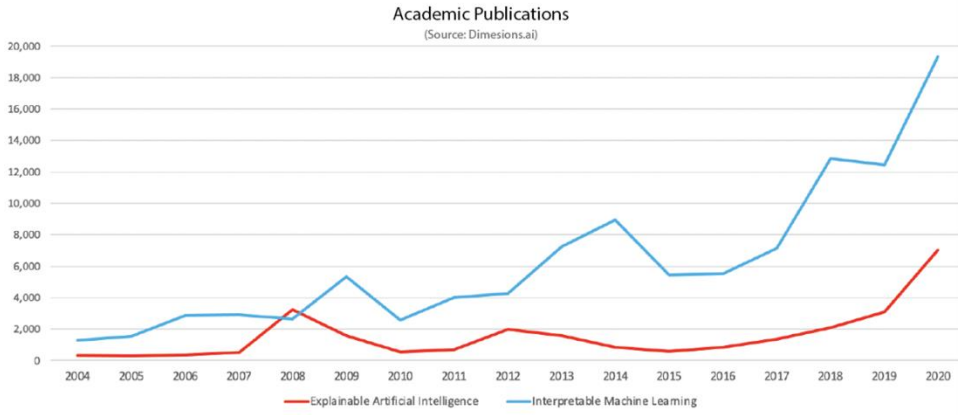
Chapter 14: What's Next for Machine Learning Interpretability?

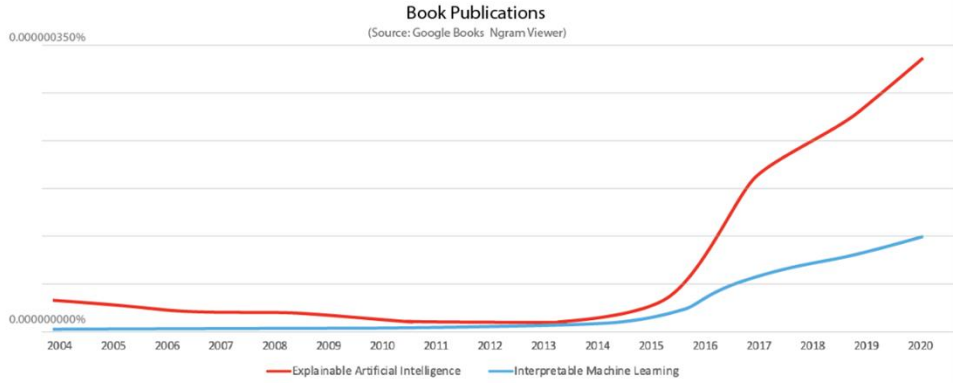
DIAGNOSTICS aim:
EVALULATING MODEL / CHECKING ASSUMPTIONS / DETECTING PROBLEMS / UNDERSTANDING DATA

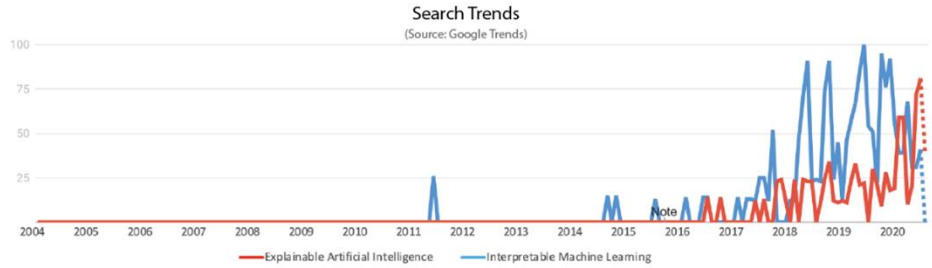
	Concerns	Interpretation Methods
FAIRNESS	Equity Justice Diversity Inclusion	• Class Balance 3 4 7 10 11 12 • Comparing Metrics 7 11 12 (FPR, FNR) • Comparing Plots 7 11 12 (Confusion Matrix, ROC Curve, PR Curve) • Group Fairness Metrics / *Individual Fairness Metrics* 11 (SPD, DI, AOD, EOD, DFBA, *CDD*) • Contour / Heat Probability Maps 12 • *Sampling Bias Evaluations*
ACCOUNTABILITY	Reliability Certainty Security Safety Robustness *Privacy*	• Out-of-sample Evaluations 8 • Sensitivity Analysis 9 (Sobol, Morris, *FAST*) • Causal Inference Methods 11 (DRL, *DML, Forest Based, Meta-Learners*) • Evasion Adversarial Robustness Evaluations 13 (FSGM, PGD, C&W, Adversarial Patches, *Boundary, PDG, B&B, DeepFool..*) • *Inference, Extraction & Poisoning Adversarial Robustness Evaluations* • *Anomaly Detection / Metrics* • *Privacy Metrics*
TRANSPARENCY	Interpretability Explainability Consistency Credibility Clarity	• Feature Importance Methods 1 2 3 4 5 8 9 10 12 (SHAP, Permutation, Model-specific) • Dimensionality Reduction Methods 3 10 (PCA, t-SNE, VAE, *DIP-VAE*) • Glass-box Models 3 (EBM, Skoped-Rules) • Partial Dependence Plots & similar 4 5 7 9 11 12 (ICE, ALE, SHAP Dependence) • White-box Surrogates 5 10 12 (Logistic Regression, Linear Regression, Rule Models, CART, KNN, *ProfWeight*) • Confirming with Statistical Tests & Correlations 5 10 12 (Spearman, Point-biserial, Cramér's V, Z-test) • Local Interpretation 6 7 9 (Decision Regions, ICE, Anchors, Counterfactuals, WIT, CEM, SHAP) • Deep Learning-specific 8 9 (IG, Saliency Maps, Grad-CAM, SmoothGrad, *Semantic Segmentation*) • *Explainability Metrics*

TREATMENT aim:
FIXING & ANTICIPATING F.A.T PROBLEMS ("TUNING FOR INTERPRETABILITY")

	APPROACH	DATA	MODEL	PREDICTION
FAIRNESS	**Mitigating Bias**	Reweighting / DIR [11] *LFR / DIR* / Unawareness [12]	Cost-sensitive Learning [10] [11] [12] Prejudice Regul. / GerryFair [11]	Calibrating/Equalizing Odds [7] [11] *Reject Option Classification*
	Placing Guardrails	Feature Engineering [10] [12]	Monotonic Constraints [12]	Prediction Abstention [11] [13]
	Enhancing Reliability	Data Augmentation [8] [11] [13]	Adversarial Debiasing [11]	*Fairness Model Certification*
	Reducing Complexity	Feature Selection [10]	Regularization [3] [12]	
ACCOUNTABILITY	**Enhancing Reliability**	*Drift Detection* Data Augmentation [8] [11] [13]	Adversarial Training [13] *Adv. Transformer Defenses* Adversarial Robustness Certified Training & Inference [13]	*Adv. Postprocessing Defenses* *Adv. Detection Defenses* *Prediction Confidence Intervals*
	Reducing Complexity	Feature Selection [10] Adv. Preprocessing Defenses [13]	Regularization [3] [12]	
	Mitigating Bias	Feature Engineering [10] [12]	Monotonic Constraints [12] (+ interaction/bi-variate constraints)	Calibrating/Equalizing Odds [7] [11]
	Ensuring Privacy	*Data Anonymization* *Differential Privacy*	*Federated Learning* *All Inference-attack Adversarial Defenses*	*Privacy-Preserving Inference*
TRANSPARENCY	**Reducing Complexity**	Feature Selection [10]	Regularization [3] [12]	
	Enhancing Reliability	Feature Engineering [10] [12]	Monotonic Constraints [12]	Local Interpretation [6] [7] [8] [9]


Academic Publications
(Source: Dimesions.ai)
20,000
18,000
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
Explainable Artificial Intelligence
Interpretable Machine Learning


Book Publications
(Source: Google Books Ngram Viewer)
0.000000350%
0.000000000%
2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
Explainable Artificial Intelligence
Interpretable Machine Learning


Search Trends
(Source: Google Trends)
100
75
50
25
Note
2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
Explainable Artificial Intelligence
Interpretable Machine Learning