Driverless AI Experiment: `cifeneca`

Generated on: Wed Jun 6 04:37:06 2018

Generated by: lauren

Experiment Overview

Driverless AI built 1 GBMModel to predict `y` given 376 original features from the input dataset `Mercedes-train.csv`. This regression experiment completed in 6 minutes and 27 seconds (0:06:27), using 8 of the 376 original features, and 76 of the 2,279 engineered derived features.

Performance

Dataset	R2
Internal Validation	0.567
Test	Test Data did not have Target Column

Driverless Settings

Dial Settings	Description	Setting Value	Range of Possible Values
Accuracy	Controls sophistication of the model	5	1-10
Time	Controls duration of the experiment	5	1-10
Interpretability	Controls complexity of the features	5	1-10

System Specs

System	GPU Memory	CPUs	GPU
Linux	120	16	0

Versions

Driverless AI Version
1.2.0+local_6cc49e3

Data Overview

This section provides information on the datasets used for the experiment.

data	file path	number of rows	number of columns
training	/home/ubuntu/data/Kaggle/MercedesBenz/Mercedes-train.csv	4,209	378
validation	Not provided		
testing	/home/ubuntu/data/Kaggle/MercedesBenz/Mercedes-test.csv	4,209	377

Training Data

The training data consists of both numeric and non numeric columns.

The summary of the columns is shown below:

Numeric Columns

name	min	mean	max	std	unique	freq of mode
ID	0.000	4,205.961	8,417.000	2,437.609	4,209	1
y	72.110	100.669	265.320	12.679	2,545	7
X10	0.000	0.013	1.000	0.115	2	56
X11	0.000	0.000	0.000	0.000	1	4,209
X12	0.000	0.075	1.000	0.264	2	316
X13	0.000	0.058	1.000	0.234	2	244
X14	0.000	0.428	1.000	0.495	2	1,802
X15	0.000	0.000	1.000	0.022	2	2
X16	0.000	0.003	1.000	0.051	2	11
X17	0.000	0.008	1.000	0.087	2	32
X18	0.000	0.008	1.000	0.088	2	33
X19	0.000	0.100	1.000	0.299	2	419
X20	0.000	0.143	1.000	0.350	2	601

name	min	mean	max	std	unique	freq of mode
X21	0.000	0.003	1.000	0.051	2	11
X22	0.000	0.087	1.000	0.282	2	366
X23	0.000	0.021	1.000	0.142	2	87
X24	0.000	0.002	1.000	0.044	2	8
X26	0.000	0.005	1.000	0.070	2	21
X27	0.000	0.683	1.000	0.466	2	2,873
X28	0.000	0.033	1.000	0.177	2	137
X29	0.000	0.043	1.000	0.203	2	181
X30	0.000	0.005	1.000	0.067	2	19
X31	0.000	0.232	1.000	0.422	2	978
X32	0.000	0.011	1.000	0.105	2	47
X33	0.000	0.000	1.000	0.015	2	1
X34	0.000	0.005	1.000	0.074	2	23
X35	0.000	0.232	1.000	0.422	2	978
X36	0.000	0.005	1.000	0.067	2	19
X37	0.000	0.232	1.000	0.422	2	978
X38	0.000	0.033	1.000	0.179	2	140
X39	0.000	0.000	1.000	0.015	2	1
X40	0.000	0.001	1.000	0.027	2	3
X41	0.000	0.011	1.000	0.106	2	48
X42	0.000	0.000	1.000	0.015	2	1
X43	0.000	0.072	1.000	0.259	2	304
X44	0.000	0.011	1.000	0.106	2	48
X45	0.000	0.253	1.000	0.435	2	1,066
X46	0.000	0.597	1.000	0.491	2	2,514
X47	0.000	0.013	1.000	0.113	2	54
X48	0.000	0.022	1.000	0.148	2	94
X49	0.000	0.122	1.000	0.327	2	514

name	min	mean	max	std	unique	freq of mode
X50	0.000	0.214	1.000	0.410	2	901
X51	0.000	0.722	1.000	0.448	2	3,038
X52	0.000	0.042	1.000	0.201	2	178
X53	0.000	0.007	1.000	0.083	2	29
X54	0.000	0.043	1.000	0.204	2	183
X55	0.000	0.005	1.000	0.072	2	22
X56	0.000	0.021	1.000	0.144	2	89
X57	0.000	0.013	1.000	0.115	2	56
X58	0.000	0.575	1.000	0.494	2	2,420
X59	0.000	0.001	1.000	0.027	2	3
X60	0.000	0.001	1.000	0.038	2	6
X61	0.000	0.954	1.000	0.210	2	4,015
X62	0.000	0.006	1.000	0.077	2	25
X63	0.000	0.011	1.000	0.106	2	48
X64	0.000	0.375	1.000	0.484	2	1,579
X65	0.000	0.002	1.000	0.046	2	9
X66	0.000	0.027	1.000	0.162	2	114
X67	0.000	0.002	1.000	0.044	2	8
X68	0.000	0.073	1.000	0.261	2	309
X69	0.000	0.030	1.000	0.170	2	126
X70	0.000	0.920	1.000	0.271	2	3,872
X71	0.000	0.104	1.000	0.305	2	436
X73	0.000	0.020	1.000	0.140	2	84
X74	0.000	0.999	1.000	0.027	2	4,206
X75	0.000	0.036	1.000	0.187	2	152
X76	0.000	0.043	1.000	0.204	2	183
X77	0.000	0.013	1.000	0.112	2	53
X78	0.000	0.006	1.000	0.075	2	24

name	min	mean	max	std	unique	freq of mode
X79	0.000	0.025	1.000	0.157	2	106
X80	0.000	0.947	1.000	0.224	2	3,986
X81	0.000	0.229	1.000	0.420	2	964
X82	0.000	0.017	1.000	0.130	2	72
X83	0.000	0.001	1.000	0.034	2	5
X84	0.000	0.104	1.000	0.305	2	436
X85	0.000	0.408	1.000	0.492	2	1,718
X86	0.000	0.001	1.000	0.038	2	6
X87	0.000	0.001	1.000	0.031	2	4
X88	0.000	0.007	1.000	0.084	2	30
X89	0.000	0.001	1.000	0.027	2	3
X90	0.000	0.007	1.000	0.086	2	31
X91	0.000	0.002	1.000	0.041	2	7
X92	0.000	0.001	1.000	0.031	2	4
X93	0.000	0.000	0.000	0.000	1	4,209
X94	0.000	0.007	1.000	0.086	2	31
X95	0.000	0.000	1.000	0.015	2	1
X96	0.000	0.758	1.000	0.428	2	3,191
X97	0.000	0.004	1.000	0.065	2	18
X98	0.000	0.943	1.000	0.233	2	3,967
X99	0.000	0.009	1.000	0.092	2	36
X100	0.000	0.690	1.000	0.462	2	2,905
X101	0.000	0.936	1.000	0.245	2	3,938
X102	0.000	0.007	1.000	0.083	2	29
X103	0.000	0.785	1.000	0.411	2	3,302
X104	0.000	0.002	1.000	0.044	2	8
X105	0.000	0.002	1.000	0.049	2	10
X106	0.000	0.013	1.000	0.114	2	55

name	min	mean	max	std	unique	freq of mode
X107	0.000	0.000	0.000	0.000	1	4,209
X108	0.000	0.015	1.000	0.120	2	62
X109	0.000	0.040	1.000	0.197	2	170
X110	0.000	0.001	1.000	0.031	2	4
X111	0.000	0.975	1.000	0.157	2	4,103
X112	0.000	0.003	1.000	0.053	2	12
X113	0.000	0.022	1.000	0.148	2	94
X114	0.000	0.146	1.000	0.353	2	615
X115	0.000	0.286	1.000	0.452	2	1,202
X116	0.000	0.197	1.000	0.398	2	828
X117	0.000	0.049	1.000	0.216	2	207
X118	0.000	0.622	1.000	0.485	2	2,619
X119	0.000	0.622	1.000	0.485	2	2,619
X120	0.000	0.958	1.000	0.201	2	4,031
X122	0.000	0.007	1.000	0.084	2	30
X123	0.000	0.003	1.000	0.051	2	11
X124	0.000	0.000	1.000	0.022	2	2
X125	0.000	0.003	1.000	0.055	2	13
X126	0.000	0.039	1.000	0.194	2	164
X127	0.000	0.495	1.000	0.500	2	2,084
X128	0.000	0.958	1.000	0.200	2	4,034
X129	0.000	0.123	1.000	0.328	2	516
X130	0.000	0.042	1.000	0.200	2	175
X131	0.000	0.027	1.000	0.161	2	112
X132	0.000	0.689	1.000	0.463	2	2,898
X133	0.000	0.124	1.000	0.330	2	523
X134	0.000	0.022	1.000	0.148	2	94
X135	0.000	0.027	1.000	0.162	2	114

name	min	mean	max	std	unique	freq of mode
X136	0.000	0.957	1.000	0.204	2	4,026
X137	0.000	0.582	1.000	0.493	2	2,449
X138	0.000	0.041	1.000	0.198	2	172
X139	0.000	0.091	1.000	0.287	2	381
X140	0.000	0.040	1.000	0.197	2	170
X141	0.000	0.014	1.000	0.119	2	60
X142	0.000	0.770	1.000	0.421	2	3,242
X143	0.000	0.038	1.000	0.192	2	161
X144	0.000	0.808	1.000	0.394	2	3,401
X145	0.000	0.001	1.000	0.038	2	6
X146	0.000	0.041	1.000	0.198	2	172
X147	0.000	0.022	1.000	0.148	2	94
X148	0.000	0.045	1.000	0.207	2	189
X150	0.000	0.792	1.000	0.406	2	3,335
X151	0.000	0.086	1.000	0.280	2	360
X152	0.000	0.032	1.000	0.177	2	136
X153	0.000	0.001	1.000	0.027	2	3
X154	0.000	0.209	1.000	0.407	2	879
X155	0.000	0.077	1.000	0.266	2	322
X156	0.000	0.717	1.000	0.450	2	3,019
X157	0.000	0.283	1.000	0.450	2	1,190
X158	0.000	0.230	1.000	0.421	2	967
X159	0.000	0.014	1.000	0.116	2	57
X160	0.000	0.001	1.000	0.034	2	5
X161	0.000	0.198	1.000	0.398	2	832
X162	0.000	0.041	1.000	0.198	2	172
X163	0.000	0.303	1.000	0.460	2	1,277
X164	0.000	0.062	1.000	0.242	2	263

name	min	mean	max	std	unique	freq of mode
X165	0.000	0.005	1.000	0.067	2	19
X166	0.000	0.033	1.000	0.179	2	140
X167	0.000	0.001	1.000	0.031	2	4
X168	0.000	0.271	1.000	0.444	2	1,140
X169	0.000	0.007	1.000	0.081	2	28
X170	0.000	0.024	1.000	0.154	2	102
X171	0.000	0.657	1.000	0.475	2	2,767
X172	0.000	0.006	1.000	0.077	2	25
X173	0.000	0.010	1.000	0.098	2	41
X174	0.000	0.017	1.000	0.131	2	73
X175	0.000	0.022	1.000	0.148	2	94
X176	0.000	0.017	1.000	0.130	2	72
X177	0.000	0.050	1.000	0.218	2	211
X178	0.000	0.557	1.000	0.497	2	2,346
X179	0.000	0.048	1.000	0.214	2	202
X180	0.000	0.158	1.000	0.365	2	665
X181	0.000	0.094	1.000	0.292	2	395
X182	0.000	0.106	1.000	0.308	2	447
X183	0.000	0.004	1.000	0.063	2	17
X184	0.000	0.001	1.000	0.038	2	6
X185	0.000	0.019	1.000	0.136	2	79
X186	0.000	0.536	1.000	0.499	2	2,256
X187	0.000	0.421	1.000	0.494	2	1,770
X189	0.000	0.915	1.000	0.278	2	3,853
X190	0.000	0.000	1.000	0.015	2	1
X191	0.000	0.471	1.000	0.499	2	1,982
X192	0.000	0.002	1.000	0.049	2	10
X194	0.000	0.464	1.000	0.499	2	1,953

name	min	mean	max	std	unique	freq of mode
X195	0.000	0.012	1.000	0.107	2	49
X196	0.000	0.010	1.000	0.101	2	43
X197	0.000	0.032	1.000	0.177	2	136
X198	0.000	0.023	1.000	0.150	2	97
X199	0.000	0.003	1.000	0.053	2	12
X200	0.000	0.007	1.000	0.081	2	28
X201	0.000	0.178	1.000	0.382	2	748
X202	0.000	0.241	1.000	0.428	2	1,016
X203	0.000	0.017	1.000	0.129	2	71
X204	0.000	0.000	1.000	0.015	2	1
X205	0.000	1.000	1.000	0.015	2	4,208
X206	0.000	0.019	1.000	0.137	2	81
X207	0.000	0.000	1.000	0.015	2	1
X208	0.000	0.063	1.000	0.243	2	265
X209	0.000	0.899	1.000	0.302	2	3,782
X210	0.000	0.000	1.000	0.015	2	1
X211	0.000	0.015	1.000	0.121	2	63
X212	0.000	0.005	1.000	0.074	2	23
X213	0.000	0.002	1.000	0.044	2	8
X214	0.000	0.007	1.000	0.083	2	29
X215	0.000	0.099	1.000	0.298	2	415
X216	0.000	0.006	1.000	0.077	2	25
X217	0.000	0.007	1.000	0.086	2	31
X218	0.000	0.312	1.000	0.464	2	1,315
X219	0.000	0.067	1.000	0.251	2	284
X220	0.000	0.561	1.000	0.496	2	2,362
X221	0.000	0.008	1.000	0.090	2	34
X222	0.000	0.022	1.000	0.148	2	94

name	min	mean	max	std	unique	freq of mode
X223	0.000	0.555	1.000	0.497	2	2,337
X224	0.000	0.317	1.000	0.466	2	1,336
X225	0.000	0.097	1.000	0.296	2	408
X226	0.000	0.032	1.000	0.177	2	136
X227	0.000	0.003	1.000	0.055	2	13
X228	0.000	0.039	1.000	0.194	2	164
X229	0.000	0.960	1.000	0.196	2	4,041
X230	0.000	0.005	1.000	0.072	2	22
X231	0.000	0.016	1.000	0.126	2	68
X232	0.000	0.043	1.000	0.203	2	181
X233	0.000	0.000	0.000	0.000	1	4,209
X234	0.000	0.202	1.000	0.401	2	849
X235	0.000	0.000	0.000	0.000	1	4,209
X236	0.000	0.000	1.000	0.022	2	2
X237	0.000	0.007	1.000	0.081	2	28
X238	0.000	0.916	1.000	0.277	2	3,856
X239	0.000	0.007	1.000	0.083	2	29
X240	0.000	0.003	1.000	0.053	2	12
X241	0.000	0.097	1.000	0.296	2	409
X242	0.000	0.007	1.000	0.086	2	31
X243	0.000	0.007	1.000	0.084	2	30
X244	0.000	0.104	1.000	0.305	2	436
X245	0.000	0.001	1.000	0.027	2	3
X246	0.000	0.409	1.000	0.492	2	1,723
X247	0.000	0.241	1.000	0.428	2	1,016
X248	0.000	0.001	1.000	0.038	2	6
X249	0.000	0.008	1.000	0.087	2	32
X250	0.000	0.553	1.000	0.497	2	2,327

name	min	mean	max	std	unique	freq of mode
X251	0.000	0.394	1.000	0.489	2	1,659
X252	0.000	0.001	1.000	0.027	2	3
X253	0.000	0.001	1.000	0.038	2	6
X254	0.000	0.005	1.000	0.072	2	22
X255	0.000	0.019	1.000	0.138	2	82
X256	0.000	0.073	1.000	0.260	2	308
X257	0.000	0.000	1.000	0.015	2	1
X258	0.000	0.002	1.000	0.049	2	10
X259	0.000	0.000	1.000	0.015	2	1
X260	0.000	0.000	1.000	0.015	2	1
X261	0.000	0.420	1.000	0.494	2	1,766
X262	0.000	0.001	1.000	0.038	2	6
X263	0.000	0.957	1.000	0.203	2	4,028
X264	0.000	0.039	1.000	0.195	2	166
X265	0.000	0.905	1.000	0.293	2	3,811
X266	0.000	0.001	1.000	0.038	2	6
X267	0.000	0.009	1.000	0.095	2	38
X268	0.000	0.000	0.000	0.000	1	4,209
X269	0.000	0.000	1.000	0.022	2	2
X270	0.000	0.000	1.000	0.015	2	1
X271	0.000	0.002	1.000	0.046	2	9
X272	0.000	0.038	1.000	0.190	2	158
X273	0.000	0.720	1.000	0.449	2	3,031
X274	0.000	0.010	1.000	0.099	2	42
X275	0.000	0.727	1.000	0.446	2	3,059
X276	0.000	0.038	1.000	0.192	2	162
X277	0.000	0.001	1.000	0.038	2	6
X278	0.000	0.000	1.000	0.022	2	2

name	min	mean	max	std	unique	freq of mode
X279	0.000	0.043	1.000	0.203	2	181
X280	0.000	0.000	1.000	0.015	2	1
X281	0.000	0.003	1.000	0.051	2	11
X282	0.000	0.004	1.000	0.063	2	17
X283	0.000	0.141	1.000	0.348	2	592
X284	0.000	0.041	1.000	0.199	2	173
X285	0.000	0.206	1.000	0.404	2	866
X286	0.000	0.055	1.000	0.227	2	230
X287	0.000	0.016	1.000	0.125	2	67
X288	0.000	0.000	1.000	0.015	2	1
X289	0.000	0.000	0.000	0.000	1	4,209
X290	0.000	0.000	0.000	0.000	1	4,209
X291	0.000	0.010	1.000	0.102	2	44
X292	0.000	0.009	1.000	0.095	2	38
X293	0.000	0.000	0.000	0.000	1	4,209
X294	0.000	0.125	1.000	0.331	2	526
X295	0.000	0.000	1.000	0.015	2	1
X296	0.000	0.000	1.000	0.015	2	1
X297	0.000	0.000	0.000	0.000	1	4,209
X298	0.000	0.005	1.000	0.067	2	19
X299	0.000	0.005	1.000	0.067	2	19
X300	0.000	0.207	1.000	0.405	2	871
X301	0.000	0.047	1.000	0.211	2	197
X302	0.000	0.011	1.000	0.106	2	48
X304	0.000	0.924	1.000	0.265	2	3,890
X305	0.000	0.013	1.000	0.115	2	56
X306	0.000	0.044	1.000	0.204	2	184
X307	0.000	0.002	1.000	0.046	2	9

name	min	mean	max	std	unique	freq of mode
X308	0.000	0.010	1.000	0.097	2	40
X309	0.000	0.007	1.000	0.084	2	30
X310	0.000	0.003	1.000	0.051	2	11
X311	0.000	0.598	1.000	0.490	2	2,519
X312	0.000	0.004	1.000	0.065	2	18
X313	0.000	0.301	1.000	0.459	2	1,267
X314	0.000	0.432	1.000	0.495	2	1,817
X315	0.000	0.029	1.000	0.167	2	121
X316	0.000	0.196	1.000	0.397	2	823
X317	0.000	0.008	1.000	0.087	2	32
X318	0.000	0.001	1.000	0.027	2	3
X319	0.000	0.000	1.000	0.022	2	2
X320	0.000	0.007	1.000	0.084	2	30
X321	0.000	0.239	1.000	0.426	2	1,005
X322	0.000	0.022	1.000	0.146	2	92
X323	0.000	0.009	1.000	0.096	2	39
X324	0.000	0.575	1.000	0.494	2	2,420
X325	0.000	0.006	1.000	0.075	2	24
X326	0.000	0.032	1.000	0.177	2	136
X327	0.000	0.128	1.000	0.334	2	540
X328	0.000	0.040	1.000	0.196	2	169
X329	0.000	0.435	1.000	0.496	2	1,832
X330	0.000	0.000	0.000	0.000	1	4,209
X331	0.000	0.056	1.000	0.230	2	236
X332	0.000	0.001	1.000	0.027	2	3
X333	0.000	0.024	1.000	0.153	2	101
X334	0.000	0.463	1.000	0.499	2	1,947
X335	0.000	0.004	1.000	0.060	2	15

name	min	mean	max	std	unique	freq of mode
X336	0.000	0.127	1.000	0.333	2	536
X337	0.000	0.517	1.000	0.500	2	2,174
X338	0.000	0.007	1.000	0.083	2	29
X339	0.000	0.000	1.000	0.015	2	1
X340	0.000	0.022	1.000	0.148	2	94
X341	0.000	0.008	1.000	0.090	2	34
X342	0.000	0.022	1.000	0.148	2	94
X343	0.000	0.078	1.000	0.269	2	330
X344	0.000	0.009	1.000	0.092	2	36
X345	0.000	0.022	1.000	0.148	2	94
X346	0.000	0.048	1.000	0.213	2	200
X347	0.000	0.000	0.000	0.000	1	4,209
X348	0.000	0.947	1.000	0.224	2	3,987
X349	0.000	0.045	1.000	0.207	2	189
X350	0.000	0.339	1.000	0.473	2	1,426
X351	0.000	0.297	1.000	0.457	2	1,251
X352	0.000	0.054	1.000	0.226	2	228
X353	0.000	0.002	1.000	0.046	2	9
X354	0.000	0.203	1.000	0.402	2	854
X355	0.000	0.380	1.000	0.486	2	1,601
X356	0.000	0.180	1.000	0.384	2	757
X357	0.000	0.001	1.000	0.034	2	5
X358	0.000	0.427	1.000	0.495	2	1,797
X359	0.000	0.032	1.000	0.176	2	134
X360	0.000	0.077	1.000	0.266	2	322
X361	0.000	0.966	1.000	0.181	2	4,066
X362	0.000	0.520	1.000	0.500	2	2,190
X363	0.000	0.754	1.000	0.431	2	3,173

name	min	mean	max	std	unique	freq of mode
X364	0.000	0.003	1.000	0.053	2	12
X365	0.000	0.003	1.000	0.053	2	12
X366	0.000	0.001	1.000	0.034	2	5
X367	0.000	0.052	1.000	0.222	2	218
X368	0.000	0.063	1.000	0.242	2	264
X369	0.000	0.000	1.000	0.022	2	2
X370	0.000	0.007	1.000	0.081	2	28
X371	0.000	0.014	1.000	0.119	2	60
X372	0.000	0.000	1.000	0.022	2	2
X373	0.000	0.019	1.000	0.137	2	81
X374	0.000	0.227	1.000	0.419	2	957
X375	0.000	0.319	1.000	0.466	2	1,342
X376	0.000	0.057	1.000	0.232	2	241
X377	0.000	0.315	1.000	0.464	2	1,325
X378	0.000	0.021	1.000	0.142	2	87
X379	0.000	0.010	1.000	0.097	2	40
X380	0.000	0.008	1.000	0.090	2	34
X382	0.000	0.008	1.000	0.087	2	32
X383	0.000	0.002	1.000	0.041	2	7
X384	0.000	0.000	1.000	0.022	2	2
X385	0.000	0.001	1.000	0.038	2	6

Non-Numeric Columns

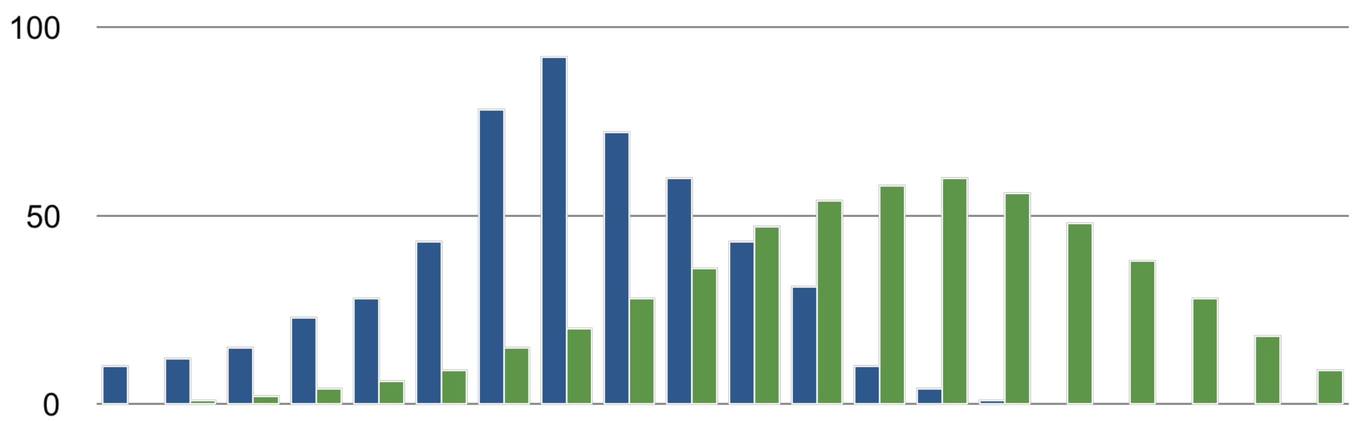
name	unique	top	freq of top value
X0	47	z	360
X1	27	aa	833
X2	44	as	1,659
X3	7	c	1,942

name	unique	top	freq of top value
X4	4	d	4,205
X5	29	v	231
X6	12	g	1,042
X8	25	j	277

Shifts Detected

Driverless AI can perform shift detection between the training, validation and testing datasets. It does this by training a binomial model to predict which dataset a record belongs to. For example, it may find that it is able to separate the training and testing data with an AUC of 0.8 using only the column: `C1` as the predictor. This indicates that there is some sort of drift in the distribution of `C1` between the training and testing data.

An example of a shift distribution between two datasets is shown below:



For this experiment, Driverless AI checked the train and test data for any shift in distributions but found none. This indicates that all the predictors/columns in the train and test data are from the same distribution.

Methodology

This section describes the experiment methodology.

Assumptions and Limitations

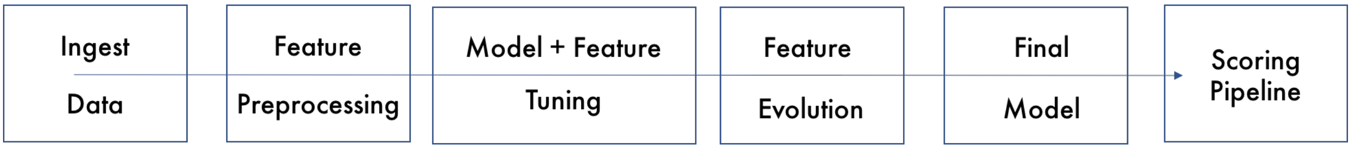
Driverless AI trains all models based on the training data provided (in this case: `Mercedes-train.csv`). It is the assumption of Driverless AI that this dataset is representative of the data that will be seen when scoring.

Driverless AI may perform shift detection between the train and test data. If a shift in distribution is detected, this may indicate that the data that will be used for scoring may have distributions not represented in the training data.

For this experiment, Driverless AI performed shift detection but found no significant changes in the distribution of the train and test data.

Experiment Pipeline

For this experiment, Driverless AI performed the following steps to find the optimal final model:



The steps in this pipeline are described in more detail below:

1. **Ingest Data**
   - detecting column types
2. **Feature Preprocessing**
   - turn raw features into numeric
3. **Model and Feature Tuning**
   - find the optimal parameter tuning settings of the xgboost model by training models with **10** different parameter combinations
   - the best parameters are those that generate the greatest **R2** on the internal validation data
   - trained and scored **18** models to evaluate features and model parameters
4. **Feature Evolution**
   - find the best representation of the data for the final model training by creating and evaluating **2,279** features over **30** iterations
   - trained and scored **22** models to further evaluate engineered features
5. **Final Model**
   - the final model is the best model from the feature engineering iterations
   - no stacked ensemble is done due to accuracy settings (consider increasing accuracy)
6. **Create Scoring Pipeline**
   - create and export the Python scoring pipeline (no MOJO Scoring Pipeline created)
   - Python Scoring Pipeline: `h2oai_experiment_cifeneca/scoring_pipeline/scorer.zip`

Experiment Settings

Below are the settings selected for the experiment by lauren:

Parameter	Value
dataset_key	tuvofoti
target_col	y
weight_col	
fold_col	
orig_time_col	
time_col	[OFF]
is_classification	False
cols_to_drop	['ID']
validset_key	
testset_key	facafego
enable_gpus	True
seed	False
accuracy	5
time	5
interpretability	5
scorer	R2
is_timeseries	False

These Accuracy, Time, and Interpretability settings map to the following internal configuration of the Driverless AI experiment:

Internal Parameter	Value
data filtered	False
tune target transform	True
number of feature engineering iterations	50
number of models trained per iteration	4
early stopping rounds	10

Internal Parameter	Value
monotonicity constraint	False
number of parameter tuning combinations	10
number of base learners in ensemble	1
time column	[OFF]

Details

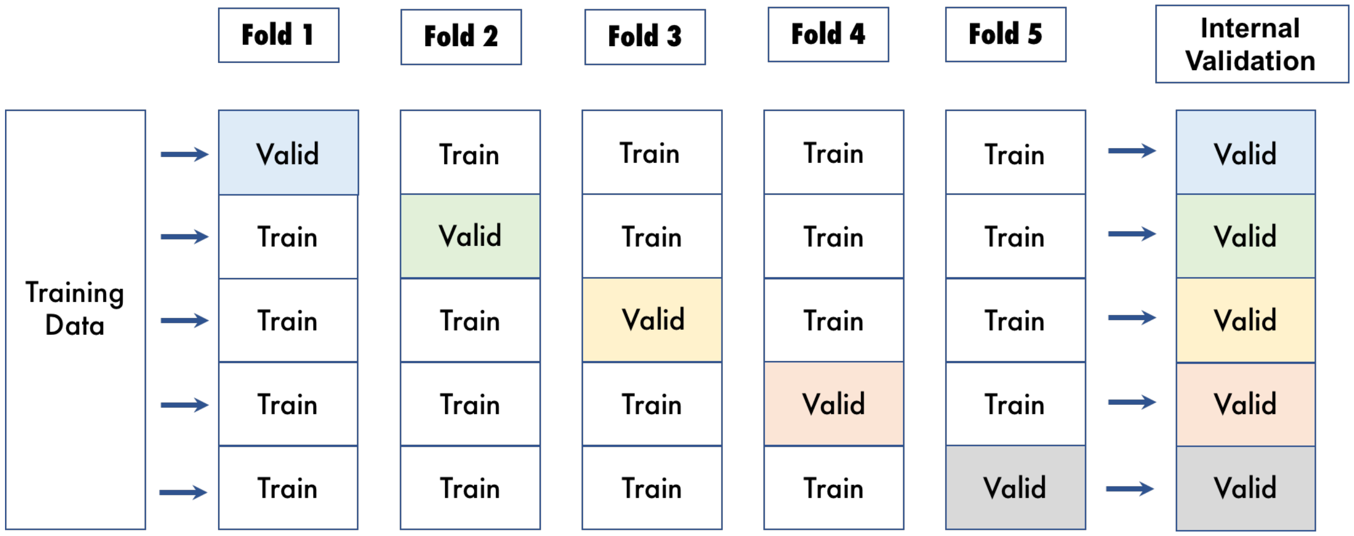
- **data filtered**: Driverless AI may filter the training data depending on the number of rows and the Accuracy setting.
    - for this experiment, the training data was not filtered.
- **tune target transform**: whether Driverless AI evaluated the model performance if the target was transformed.
    - ex: the model performance may be better by predicting the log of the target column instead of the raw target column
- **number of feature engineering iterations**: the number of iterations performed of feature engineering.
- **number of models evaluated per iteration**: for each feature engineering iteration, Driverless AI trains multiple models. Each model is trained with a different set of predictors or features. The goal of this step is to determine which types of features, lead to the greatest R2.
- **early stopping rounds**: if Driverless AI does not see any improvement after 10 iterations of feature engineering, the feature engineering step is automatically stopped.
- **monotonicity constraint**: if enabled, the xgboost model will only have monotone relationships between the predictors and target variable.
- **number of parameter tuning combinations**: the number of parameter combinations evaluated to determine the optimal parameter settings for the xgboost models
- **number of base learners in ensemble**: the number of base models used to create the final ensemble. If the number is 1, no ensemble is done and the final model is a single model.
- **time column**: the column that provides time column. If a time column is provided, feature engineering and model validation will respect the causality of time. If the time column is turned off, no time order is used for modeling and data may be shuffled randomly (any potential temporal causality will be ignored).

Validation Strategy

Driverless AI automatically split the training data to determine the performance of the model parameter tuning and feature engineering steps.

For the experiment, Driverless AI randomly split the data into 4 fold cross validation. With cross validation, the whole dataset is utilized by training 4 models where each model is trained on a different subset of the training data.

The visualization below shows how cross validation is utilized to get predictions on hold out data. The visualization shows an example of cross validation with 5 folds. For this experiment, however, 4 folds were created.



Model Tuning

The table below shows a portion of the different parameter configurations evaluated by Driverless AI for the xgboost models and their score and training time. The table is ordered based on a combination of greatest score and lowest training time.

tree method	grow policy	max depth	max leaves	colsample bytree	subsample	scores	training times
hist	lossguide	0.000	64.000	0.8	0.7	0.562	4.878
hist	depthwise	6.000	0.000			0.559	4.454
hist	lossguide	0.000	16.000			0.558	2.100
hist	lossguide	0.000	256.000			0.557	8.947
hist	depthwise	8.000	0.000			0.557	6.039
hist	depthwise	4.000	0.000			0.556	1.895
hist	depthwise	8.000	0.000			0.556	6.181
hist	depthwise	6.000	0.000			0.556	3.516
hist	lossguide	0.000	256.000			0.556	8.788

tree method	grow policy	max depth	max leaves	colsample bytree	subsample	scores	training times
hist	depthwise	8.000	0.000			0.555	6.042

Feature Transformation

The top 14 features used in the final model are shown belowed ordered by importance.If no transformer was applied, the feature is an original column.

Relative Importance	Feature	Description	Transformer
1.000	0_CVTE: X0.0	Out-of-fold mean of the response grouped by: ['X0'] using 5 folds [internal:(10, 1, None)]	Cross Validation Target Encoding
0.201	39_CVTE: X136.0	Out-of-fold mean of the response grouped by: ['X136'] using 5 folds [internal:(10, 1, None)]	Cross Validation Target Encoding
0.183	225_CVTE: X314.0	Out-of-fold mean of the response grouped by: ['X314'] using 5 folds [internal:(10, 1, None)]	Cross Validation Target Encoding
0.139	143_CVTE: X232.0	Out-of-fold mean of the response grouped by: ['X232'] using 5 folds [internal:(10, 1, None)]	Cross Validation Target Encoding
0.090	20_CVTE: X118.0	Out-of-fold mean of the response grouped by: ['X118'] using 5 folds [internal:(10, 1, None)]	Cross Validation Target Encoding
0.080	585_X314	X314 (original)	None
0.059	29_CVTE: X127.0	Out-of-fold mean of the response grouped by: ['X127'] using 5 folds [internal:(10, 1, None)]	Cross Validation Target Encoding
0.045	95_CVTE: X189.0	Out-of-fold mean of the response grouped by: ['X189'] using 5 folds [internal:(10, 1, None)]	Cross Validation Target Encoding
0.018	147_CVTE: X238.0	Out-of-fold mean of the response grouped by: ['X238'] using 5 folds [internal:(10, 1, None)]	Cross Validation Target Encoding

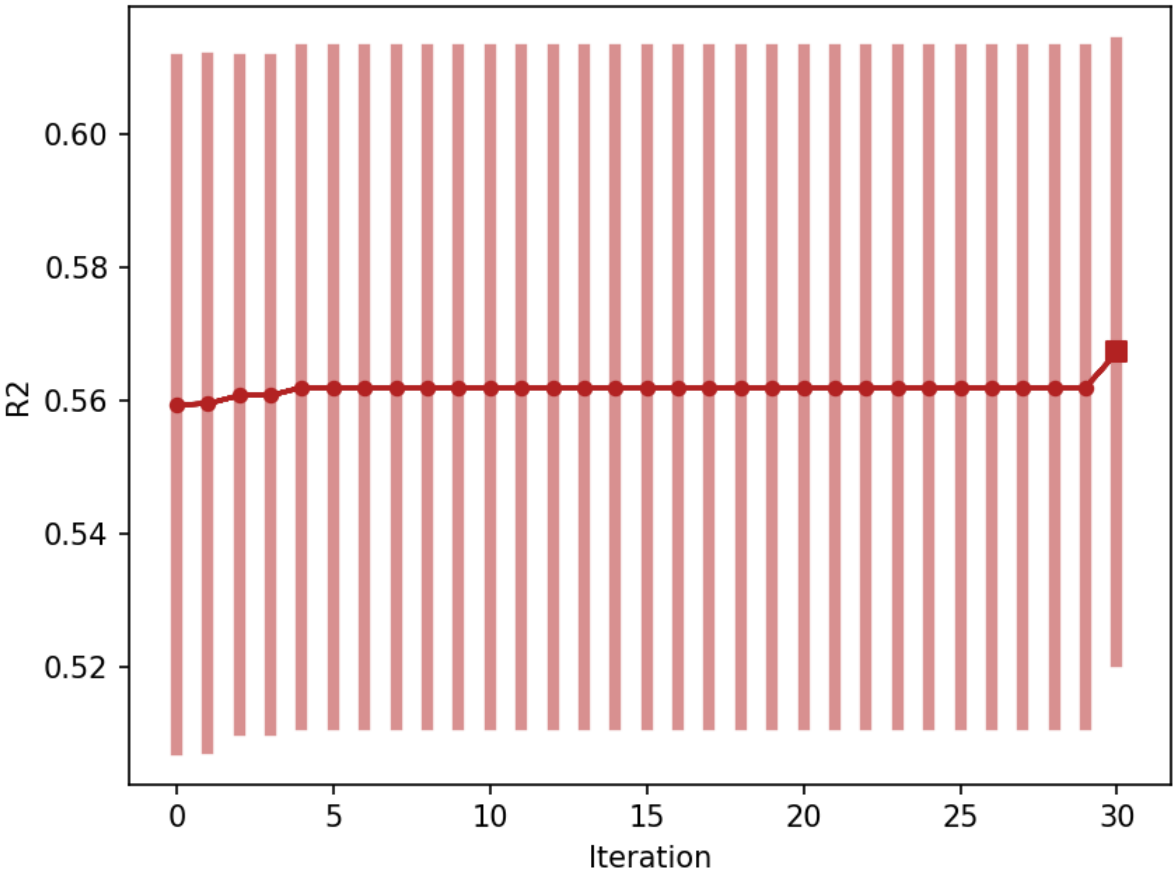
Relative Importance	Feature	Description	Transformer
0.012	226_CVTE: X315.0	Out-of-fold mean of the response grouped by: ['X315'] using 5 folds [internal:(10, 1, None)]	Cross Validation Target Encoding
0.012	313_CVTE: X5.0	Out-of-fold mean of the response grouped by: ['X5'] using 5 folds [internal:(10, 1, None)]	Cross Validation Target Encoding
0.012	21_CVTE: X119.0	Out-of-fold mean of the response grouped by: ['X119'] using 5 folds [internal:(10, 1, None)]	Cross Validation Target Encoding
0.011	174_CVTE: X263.0	Out-of-fold mean of the response grouped by: ['X263'] using 5 folds [internal:(10, 1, None)]	Cross Validation Target Encoding
0.009	382_X118	X118 (original)	None

Feature Evolution

Driverless AI trained 161 models throughout the experiment in an effort to test models with different sets of features.

Driverless AI Step	Number of Models	Number of Folds/Validation Datasets
Parameter and Feature Tuning	18	4
Feature Evolution	22	4
Final Model	1	4

The graph below shows the performance of the best model per iteration during the Driverless AI experiment. The performance of the final model is denoted by the large square marker.



The first model (our baseline model) and the final model are compared in more detail below:

Model	R2 - Internal Validation
Baseline Model	0.55936 +/- 0.0518
Final Model	0.56726 +/- 0.04657

Deployment

For this experiment, the Python Scoring Pipeline is available for productionizing the final model pipeline for a given row of data or table of data. The MOJO Scoring Pipeline is not available.

Python Scoring Pipeline

This package contains an exported model and Python 3.6 source code examples for productionizing models built using H2O Driverless AI. The Python Scoring Pipeline is located here:

- `h2oai_experiment_cifeneca/scoring_pipeline/scorer.zip`

The files in this package allow you to transform and score on new data in a couple of different ways:

- From Python 3.6, you can import a scoring module, and then use the module to transform and score on new data.
- From other languages and platforms, you can use the TCP/HTTP scoring service bundled with this package to call into the scoring pipeline module through remote procedure calls (RPC).