Propeller Health (COPD)

Meet the team

Anthony Melendez
amelendez@sandiego.edu

Patrick Kelly
pkelly@sandiego.edu

Cianna Platt
cplatt@sandiego.edu

Danielle Henry
dhenry@sandiego.edu

Sebastian Perez
sebastianperez@sandiego.edu

Project Plan Overview

Main Question(s):

- What are ***early indicators*** of ***exacerbation*** within COPD patients?
- What ***effect*** does ***adherence*** have on the ***likelihood*** of ***exacerbation***?

Primary Metric/Variable:

- Number of ***rescue events***

Problem Statement:

- How can Propeller ***better predict*** the needs of COPD patients to ***better improve*** their quality of life?

Data Preparation

Programs used:

- SQL
  - Import data into SQL to ***create*** a ***database***
  - ***Aggregate*** into daily values for every patient
  - ***Merge***: Daily patient kpi, events, and user tables
- Python
  - Perform all ***supervised*** and ***unsupervised*** learning models/techniques

Qualitative Analysis:

- Brainstorm what could be ***key indicators*** of exacerbation

Data Preparation - Total Indicator

Prepared on SQL:

- Baseline:
  - Average # of puffs in the last 29 days (if applicable)
- Standard deviation:
  - Calculated from the last 29 days and lifetime # of puffs
- Risk Indicator 1:
  - 10 rescue events a day where the baseline is ***less*** than 10
- Risk Indicator 2 (BOTH criteria):
  - 2 days of steady ***increase*** in rescue puffs
  - Number (#) of rescue events on a day is ***greater*** than the sum of the baseline and 1.64 times the standard deviation

of rescue puffs 1 day prior
of rescue puffs 2 days prior
Baseline
X_bar, X_i, N, N-1
Difference Squared
Standard Deviation

Some of the variables used to calculate Total Indicator

We derived “Total Indicator” from the ***existing*** data in order to quantify an exacerbation event

Data Exploration

What didn't work:

- Morning, afternoon, evening variables
- Number of medicines prescribed per user
- Day of the week dummies (Monday - Sunday)
- Exacerbation from one day, two days, and a week before (lagged variables)
- Cluster groups:
  - Weather clusters
  - Air quality clusters

What worked:

- Weekly & monthly adherence
- Number of triggers entered
- Number of symptoms entered
- Air quality metrics
- Weather metrics
- Symptoms
- Triggers
- Age

Exploratory Analysis

Logistic Regression

Coefficients

	Estimate	Standard Error	Odds Ratio	z	Wald Test		
					Wald Statistic	df	p
(Intercept)	−1.322	0.073	0.267	−18.098	327.536	1	< .001
trigger_entered_count	0.008	0.003	1.008	2.255	5.084	1	0.024
symptom_entered_count	0.041	0.005	1.042	7.758	60.186	1	< .001
weekly_adherence_followed	−0.366	0.022	0.694	−16.278	264.963	1	< .001
aq_breezometer_avg	−0.005	0.001	0.995	−5.649	31.914	1	< 001
weather_humidity_avg	−0.001	0.000	0.999	−3.323	11.039	1	< 001
weather_temp_avg	−0.002	0.000	0.998	−6.074	36.893	1	< 001

Note. total indicator level '1' coded as class 1.

- Logistic regression is used to predict a ***categorical*** outcome (i.e. do they exacerbate or not)
- This is ***not*** a great predictive model
- Helpful to find important individual predictors
- ***Weekly adherence average*** has the ***largest*** impact on our exacerbation indicator

Logistic Regression - Monthly Adherence

Coefficients

	Estimate	Standard Error	Odds Ratio	z	Wald Test		
					Wald Statistic	df	p
(Intercept)	−0.416	0.087	0.659	−4.738	22.735	1	< .001
trigger_entered_count	0.011	0.004	1.011	2.994	8.964	1	0.003
symptom_entered_count	0.040	0.006	1.041	7.234	52.324	1	< .001
monthly_adherence_followed	−0.391	0.025	0.676	−15.919	253.422	1	< .001
aq_breezometer_avg	−0.005	0.001	0.995	−5.056	25.566	1	< .001
weather_humidity_avg	−0.002	0.000	0.998	−3.886	15.105	1	< .001
weather_temp_avg	−0.002	0.000	0.998	−4.954	24.543	1	< .001
age	−0.015	0.001	0.985	−18.424	339.436	1	< .001

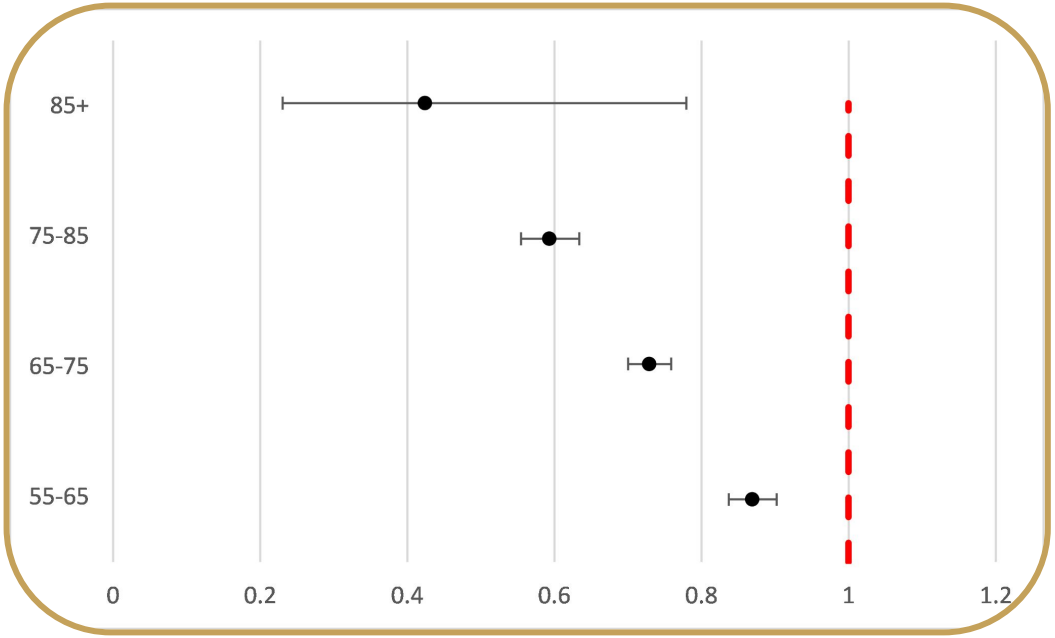
Note. total_indicator level '1' coded as class 1.

- Weekly adherence and monthly adherence has somewhat the ***same*** effect
  - Monthly had .018 ***more*** effect

Triggers, Symptoms, and Age

Age Odds-Ratio

Basecase:
The **probability** of *exacerbation* for a patient under 55 is 12%



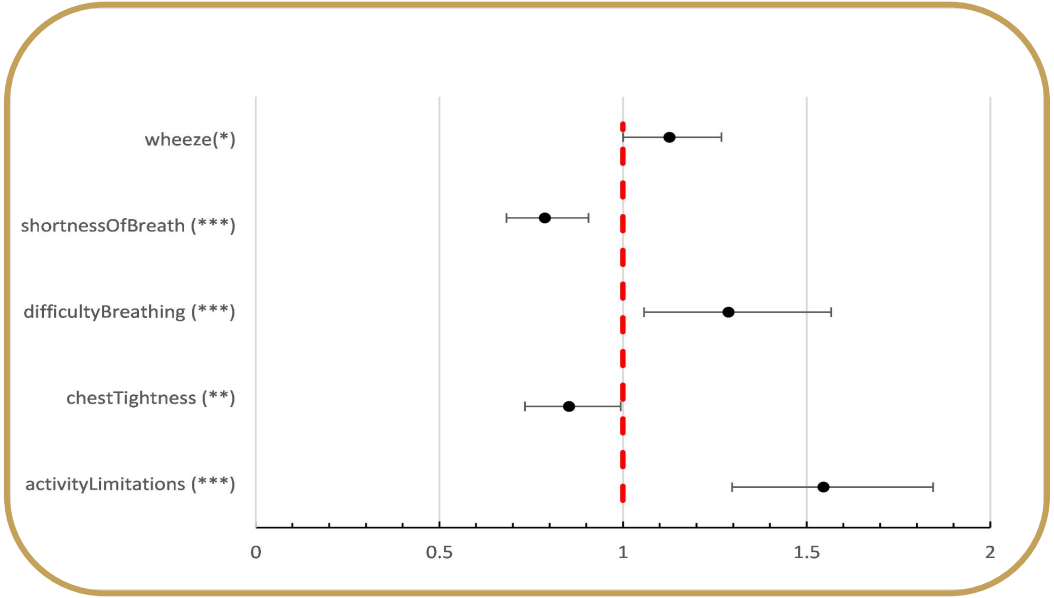
99% Confidence!

When the patient is ***55-65 years old***, they are 13% **less** likely to exacerbate than patients ***under 55***!

*See Appendix for control variables

Symptoms Odds-Ratio

Basecase:
The **probability** of *exacerbation* when a patient does not experience any of the symptoms in the model is 10%



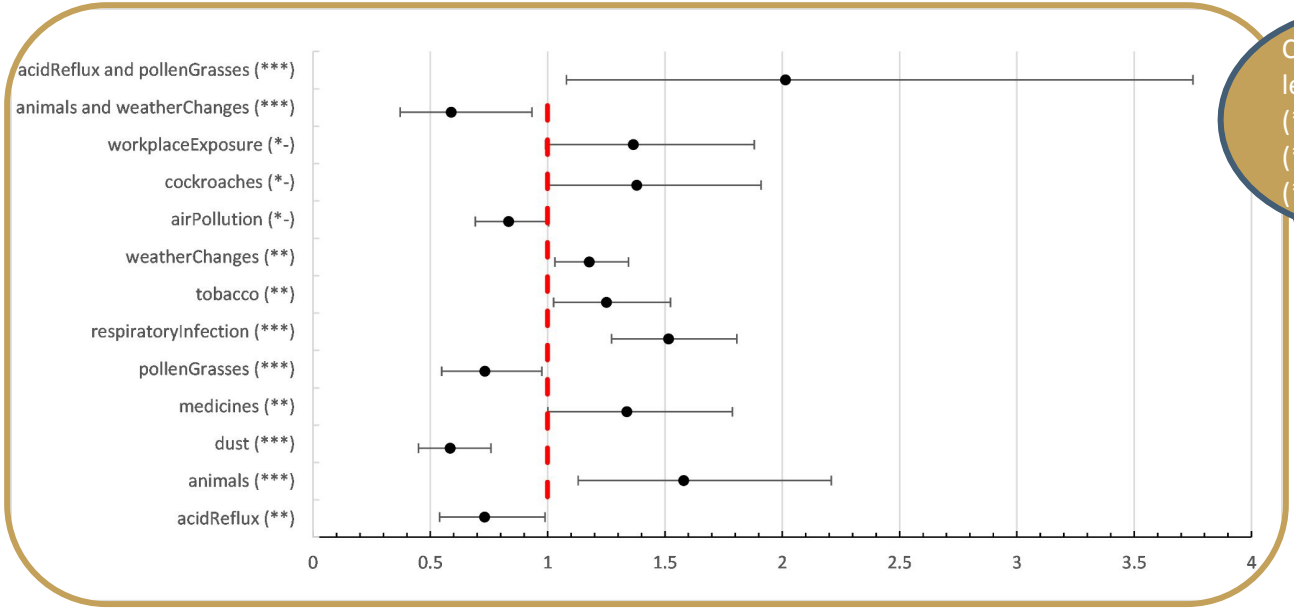
Confidence levels:
(*) : 90%
(**) : 95%
(***) : 99%

When the patient is experiencing ***activity limitations***, they are 56% **more** likely to exacerbate than when they're not experiencing it.

*See Appendix for control variables

Triggers Odds-Ratio

Basecase: The **probability** of *exacerbation* when a patient does not experience any of the triggers listed is 10%



Confidence levels:
(*-) : 89-90%
(**) : 95%
(***) : 99%

- When the patient is experiencing ***acid reflux*** they are **27% less** likely to exacerbate
- When the patient is experiencing ***pollen grasses***, they are **27% less** likely to exacerbate
- However, when the patient is experiencing ***acid reflux*** and ***pollen grasses*** at the same time, they are **8% more** likely to exacerbate

Machine Learning Models

Machine Learning Model Debrief

Accuracy → The percentage of correct predictions by a model out of all predictions. Correct predictions consists of:

- Correctly predicting exacerbation (True positive)
- Correctly predicting no exacerbation (True negative)

Recall → The percentage of ***correctly*** predicting someone would exacerbate divided by the number of times they ***actually*** exacerbated, which consists of:

- Correctly predicting exacerbation (True positive)
- Incorrectly predicting no exacerbation (False negative)

Machine Learning Models Results

- While the decision tree has the highest recall, the XGBoost Classifier has the highest combination of accuracy and recall

Best Machine Learning Technique

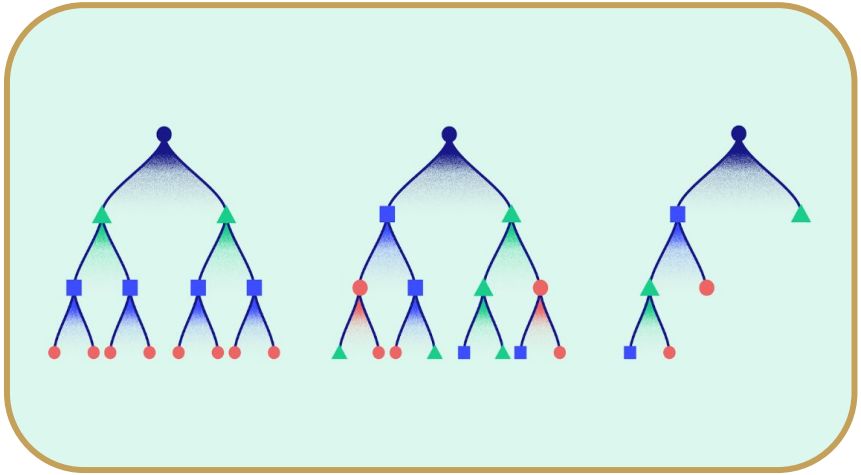
Best Recall

Machine Learning Technique	Accuracy	Recall
Gradient Boosting	90.03%	0.09%
Random Forest	89.75%	1.21%
Linear Discriminant	90.11%	0.05%
Decision Tree	43.91%	68.64%
XGBoost Classifier	70.23%	60.99%

XGBoost Model pt. 2

Inputs:

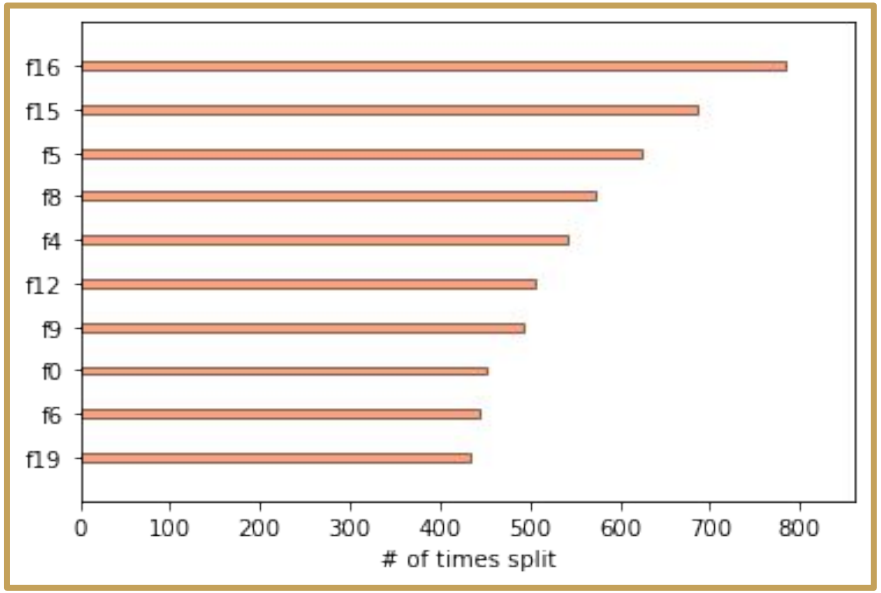
- Monthly adherence
- Weather metrics
- Air quality metrics
- Rescue medicine type
- Number of medications prescribed
- Symptoms and weekly # of symptoms
- Triggers and weekly # of triggers
- Age
- Exacerbation 1,2 days prior



Our ***final*** XGBoost model achieved an ***accuracy*** of 73% and a ***recall*** of 84%!

Model Characteristics: Top 10

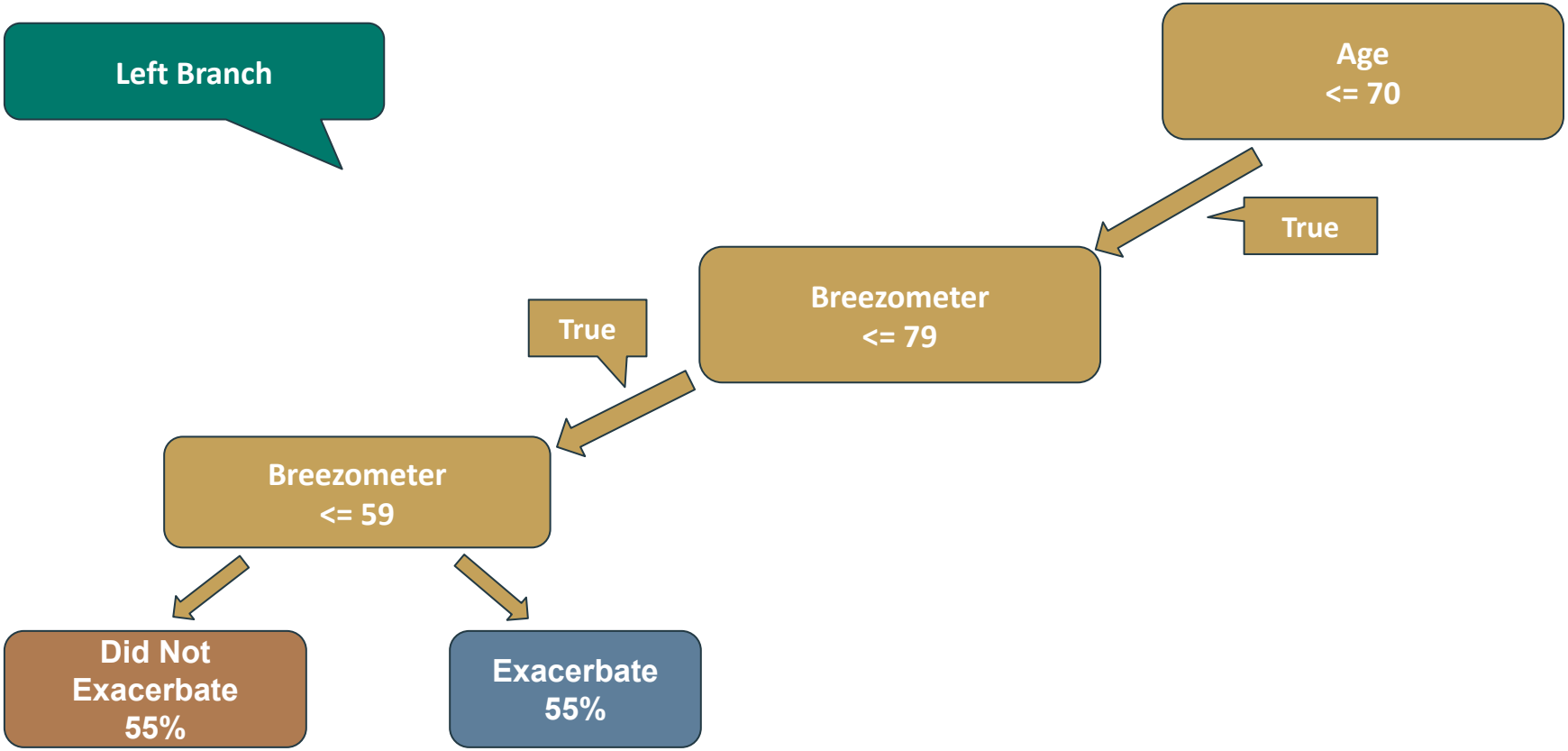
- F16: ***Wind speed*** (daily average)
- F15: ***Wind direction*** (daily average)
- F5: Daily ***NO2*** levels
- F8: Daily ***PM25*** levels
- F4: Daily ***CO*** level
- F12: ***Humidity***
- F9: Daily ***SO2*** level
- F0: Exacerbation on ***prior day***
- F6: Daily ***O3*** levels
- F19: ***Age***



Wind Speed has the highest number of splits in the model

Decision Tree Analysis

Decision Tree Analysis



Decision Tree Analysis

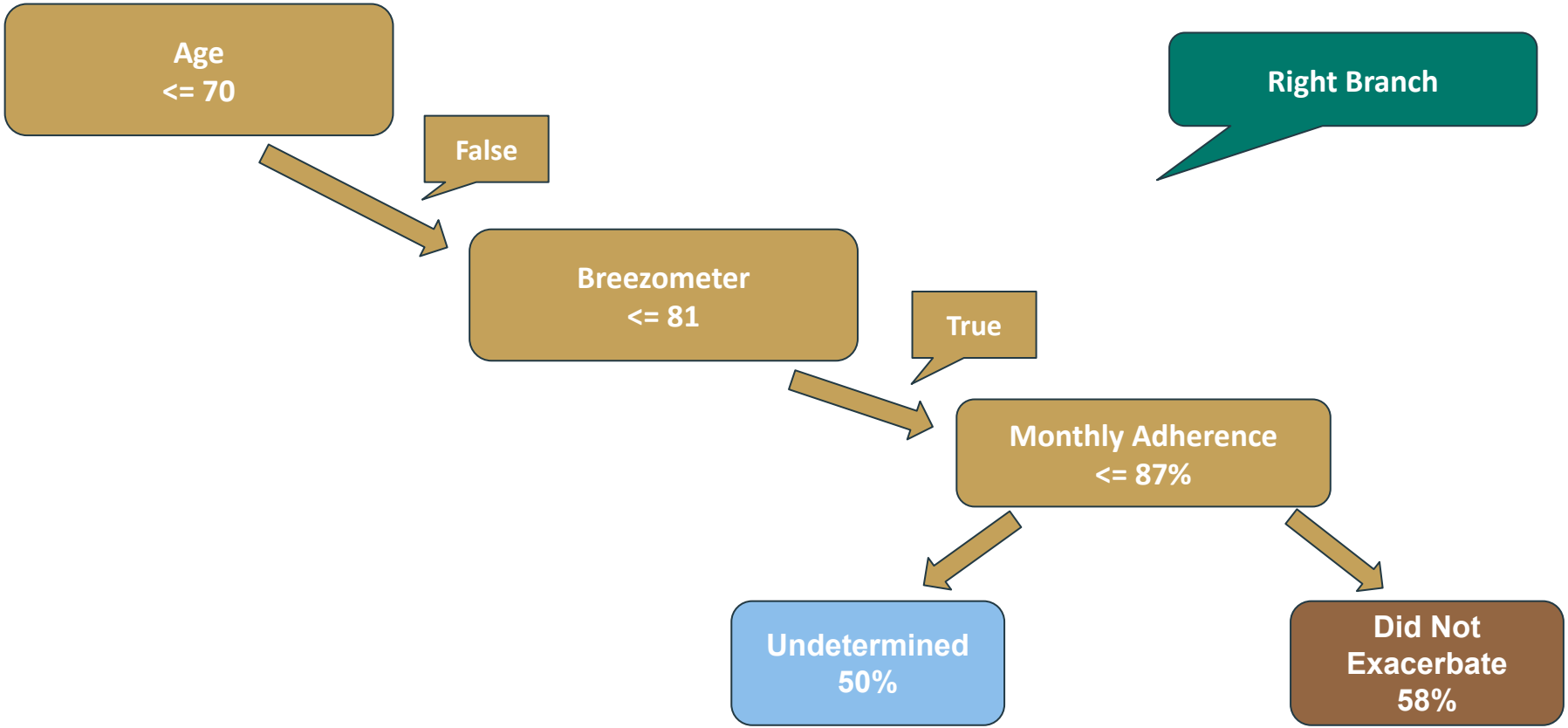
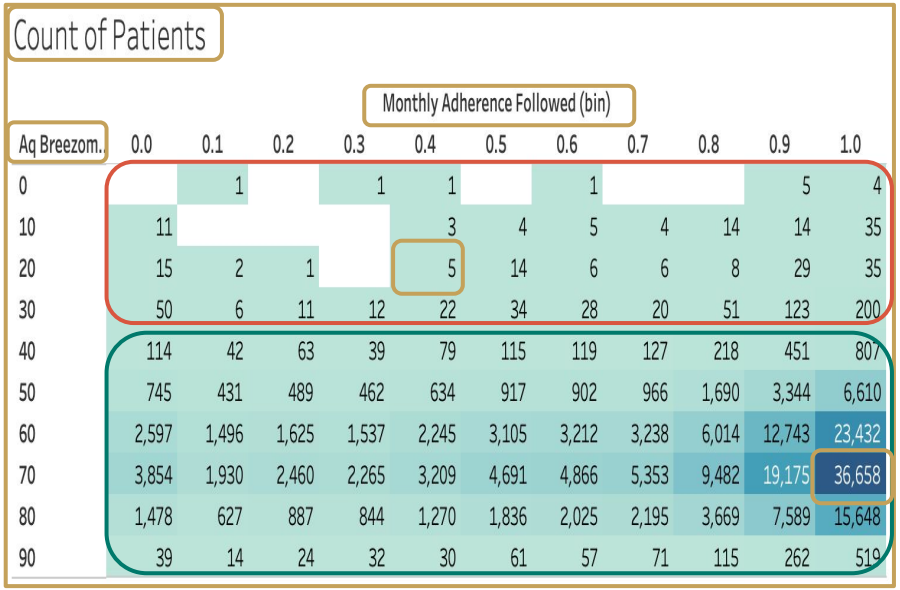
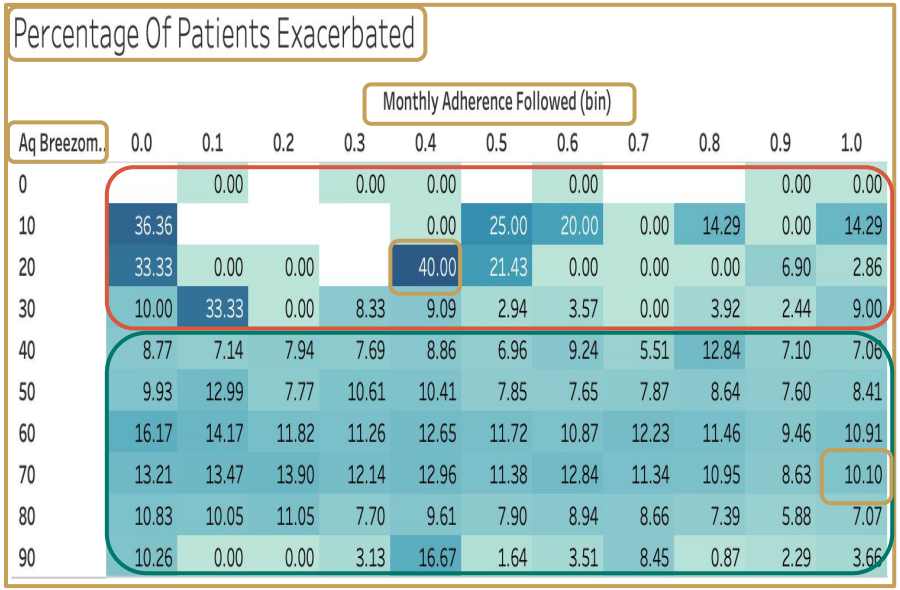


Tableau Observations



Count of Patients

Aq Breezom..	Monthly Adherence Followed (bin) 0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0		1		1	1		1			5	4
10	11				3	4	5	4	14	14	35
20	15	2	1		5	14	6	6	8	29	35
30	50	6	11	12	22	34	28	20	51	123	200
40	114	42	63	39	79	115	119	127	218	451	807
50	745	431	489	462	634	917	902	966	1,690	3,344	6,610
60	2,597	1,496	1,625	1,537	2,245	3,105	3,212	3,238	6,014	12,743	23,432
70	3,854	1,930	2,460	2,265	3,209	4,691	4,866	5,353	9,482	19,175	36,658
80	1,478	627	887	844	1,270	1,836	2,025	2,195	3,669	7,589	15,648
90	39	14	24	32	30	61	57	71	115	262	519



Percentage Of Patients Exacerbated

Aq Breezom..	Monthly Adherence Followed (bin) 0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0		0.00		0.00	0.00		0.00			0.00	0.00
10	36.36				0.00	25.00	20.00	0.00	14.29	0.00	14.29
20	33.33	0.00	0.00		40.00	21.43	0.00	0.00	0.00	6.90	2.86
30	10.00	33.33	0.00	8.33	9.09	2.94	3.57	0.00	3.92	2.44	9.00
40	8.77	7.14	7.94	7.69	8.86	6.96	9.24	5.51	12.84	7.10	7.06
50	9.93	12.99	7.77	10.61	10.41	7.85	7.65	7.87	8.64	7.60	8.41
60	16.17	14.17	11.82	11.26	12.65	11.72	10.87	12.23	11.46	9.46	10.91
70	13.21	13.47	13.90	12.14	12.96	11.38	12.84	11.34	10.95	8.63	10.10
80	10.83	10.05	11.05	7.70	9.61	7.90	8.94	8.66	7.39	5.88	7.07
90	10.26	0.00	0.00	3.13	16.67	1.64	3.51	8.45	0.87	2.29	3.66

- Explorations are based off the ***outcomes*** from the ***Decision Tree***
- **Upper half** → less patients, but a higher chance they exacerbate
- **Lower half** → more patients, but a lower chance they exacerbate

Recommendations

- If the adherence, air quality, triggers passes a certain ***threshold***, then trigger a "***warning***" on the app
- Collect more information through the app for ***additional information*** on COPD patients:
  - Identifying groups of "personas"
    - ***Exercise routine***:
      - How many days?
      - Length of exercise?
      - Type of exercise?
    - ***Nutrition***:
      - Other supplements/medication taken
      - Types of food
      - Number of meals
    - How long have they been diagnosed
- Promote/incentivize more app ***usage***

Questions?

Appendix

Logistic Regression Output (1st model)

Logistic Regression

Model Summary - total_indicator

Model	Deviance	AIC	BIC	df	X^2	p	McFadden R^2	Nagelkerke R^2	Tjur R^2	Cox & Snell R^2
H_0	141586.204	141588.204	141598.500	218689						
H_1	141106.240	141120.240	141192.308	218683	479.965	< .001	0.003	0.002	0.002	0.002

Coefficients

					Wald Test		
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	−1.322	0.073	0.267	−18.098	327.536	1	< .001
trigger_entered_count	0.008	0.003	1.008	2.255	5.084	1	0.024
symptom_entered_count	0.041	0.005	1.042	7.758	60.186	1	< .001
weekly_adherence_followed	−0.366	0.022	0.694	−16.278	264.963	1	< .001
aq_breezometer_avg	−0.005	0.001	0.995	−5.649	31.914	1	< .001
weather_humidity_avg	−0.001	0.000	0.999	−3.323	11.039	1	< .001
weather_temp_avg	−0.002	0.000	0.998	−6.074	36.893	1	< .001

Note. total_indicator level '1' coded as class 1.

Multicollinearity Diagnostics

	Tolerance	VIF
trigger_entered_count	0.591	1.693
symptom_entered_count	0.592	1.690
weekly_adherence_followed	0.997	1.003
aq_breezometer_avg	0.727	1.376
weather_humidity_avg	0.778	1.285
weather_temp_avg	0.888	1.126

Logistic Regression (2nd Model)

Logistic Regression ▼

Model Summary - total_indicator

Model	Deviance	AIC	BIC	df	X^2	p	McFadden R^2	Nagelkerke R^2	Tjur R^2	Cox & Snell R^2
H_0	139205.178	139207.178	139217.454	214623						
H_1	138340.624	138356.624	138438.837	214615	864.554	< .001	0.006	0.004	0.004	0.004

Coefficients ▼

					Wald Test		
	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	−0.416	0.087	0.659	−4.758	22.735	1	< .001
trigger_entered_count	0.011	0.004	1.011	2.994	8.964	1	0.003
symptom_entered_count	0.040	0.006	1.041	7.234	52.324	1	< .001
monthly_adherence_followed	−0.391	0.025	0.676	−15.919	253.422	1	< .001
aq_breezometer_avg	−0.005	0.001	0.995	−5.056	25.566	1	< .001
weather_humidity_avg	−0.002	0.000	0.998	−3.886	15.105	1	< .001
weather_temp_avg	−0.002	0.000	0.998	−4.954	24.543	1	< .001
age	−0.015	0.001	0.985	−18.424	339.436	1	< .001

Note. total_indicator level '1' coded as class 1.

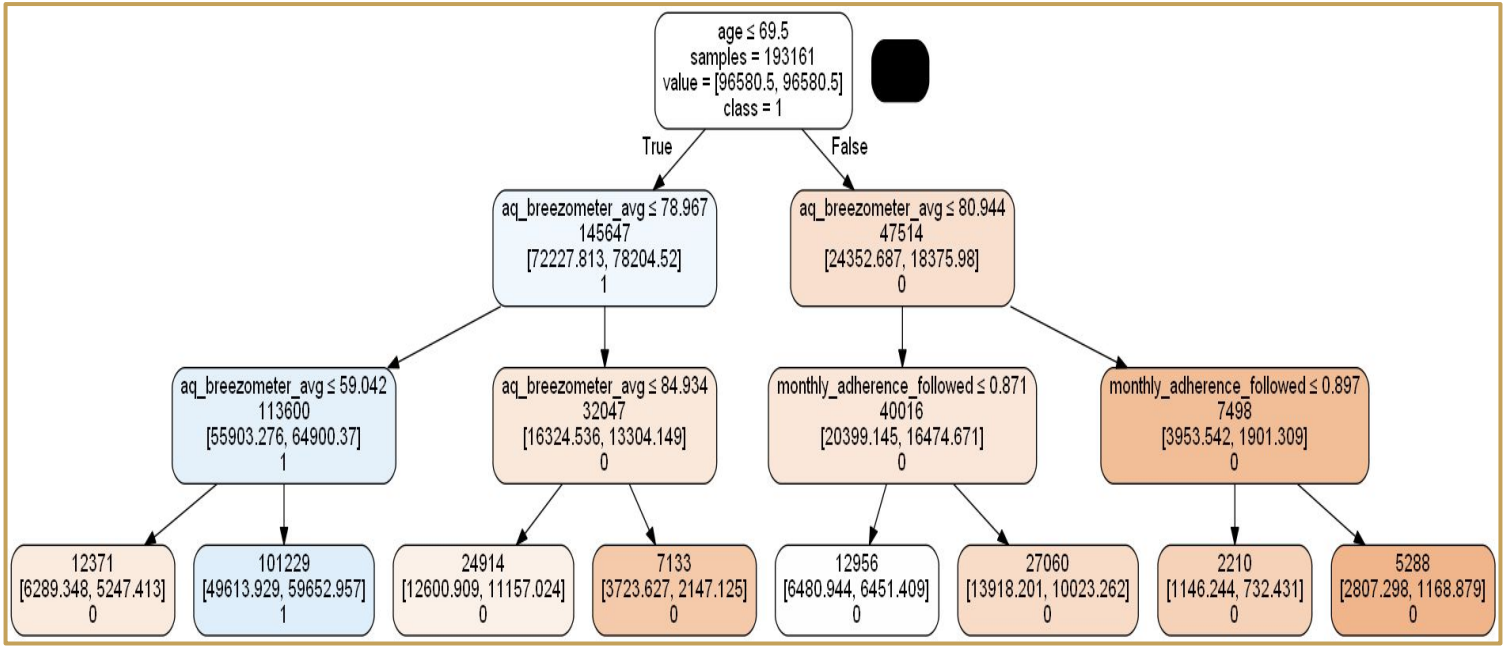
Multicollinearity Diagnostics

	Tolerance	VIF
trigger_entered_count	0.563	1.777
symptom_entered_count	0.564	1.773
monthly_adherence_followed	0.988	1.012
aq_breezometer_avg	0.727	1.375
weather_humidity_avg	0.778	1.285
weather_temp_avg	0.886	1.129
age	0.986	1.014

Full Decision Tree

Hyper parameters for this decision tree:

- Max depth: 3
- Minimum impurity decrease: 0
- Minimum sample split: 5



XGBoost Classifier

Hyper parameters used:

- scale_pos_weight = 0.89
- max_depth = 7
- objective = 'binary:logistic'
- colsample_bytree= 0.5
- gamma= 0.25
- learning_rate= 0.1
- reg_lambda= 10
- subsample= 0.8

Pre-processing steps:

- Standardized numerical
- One hot encoded categorical
- Imputer: “mean”

Logistic regression (Symptoms)

Logistic Regression

Model Summary - total_indicator

Model	Deviance	AIC	BIC	df	X^2	p	McFadden R^2	Nagelkerke R^2	Tjur R^2	Cox & Snell R^2
H_0	139205.178	139207.178	139217.454	214623						
H_1	139140.937	139156.937	139239.150	214616	64.241	< .001	4.615e−4	2.993e−4	3.038e−4	2.993e−4

Coefficients

	Estimate	Standard Error	z	Wald Test			(odds ratio scale)	
				Wald Statistic	df	p	Lower bou	Upper bound
(Intercept)	−2.202	0.007	−298.759	89257.141	1	< .001	0.109	0.113
activityLimitations (1)	0.436	0.068	6.378	40.682	1	< .001	1.297	1.844
chestTightness (1)	−0.159	0.078	−2.044	4.177	1	0.041	0.733	0.994
cough (1)	−0.092	0.072	−1.269	1.610	1	0.205	1.057	1.567
difficultyBreathing (1)	0.252	0.076	3.299	10.885	1	< .001	0.683	0.906
shortnessOfBreath (1)	−0.240	0.055	−4.377	19.161	1	< .001	1	1.268
wheeze (1)	0.119	0.072	1.650	2.724	1	0.099		
wokenAtNight (1)	−0.020	0.095	−0.206	0.042	1	0.837		

Note. total_indicator level '1' coded as class 1.

Multicollinearity Diagnostics

	Tolerance	VIF
activityLimitations	0.518	1.929
chestTightness	0.532	1.880
cough	0.652	1.534
difficultyBreathing	0.513	1.951
shortnessOfBreath	0.444	2.250
wheeze	0.618	1.618
wokenAtNight	0.792	1.263

Confidence bounds adjusted per significance level!

Logistic regression (Triggers)

Coefficients

	Estimate	Standard Error	Odds Ratio	z	Wald Test		
					Wald Statistic	df	p
(Intercept)	−2.209	0.007	0.110	−298.983	89390.914	1	< .001
acidReflux (1)	−0.314	0.154	0.731	−2.030	4.123	1	0.042
airPollution (1)	−0.182	0.113	0.833	−1.614	2.603	1	0.107
animals (1)	0.457	0.130	1.580	3.517	12.371	1	< .001
anxietyOrStress (1)	−0.084	0.072	0.919	−1.176	1.382	1	0.240
cockroaches (1)	0.321	0.198	1.379	1.623	2.633	1	0.105
coldAir (1)	0.100	0.067	1.105	1.490	2.220	1	0.136
dust (1)	−0.539	0.102	0.584	−5.289	27.977	1	< .001
exercise (1)	0.031	0.061	1.032	0.514	0.264	1	0.607
flu (1)	0.221	0.161	1.247	1.375	1.891	1	0.169
foodAllergies (1)	−0.322	0.272	0.725	−1.184	1.402	1	0.236
indoorAllergens (1)	−0.051	0.102	0.950	−0.502	0.252	1	0.616
indoorMolds (1)	−0.102	0.138	0.903	−0.740	0.547	1	0.460
medicines (1)	0.291	0.148	1.337	1.964	3.857	1	0.050
other (1)	−0.029	0.076	0.971	−0.382	0.146	1	0.703
outdoorAllergens (1)	0.109	0.098	1.115	1.119	1.252	1	0.263
outdoorMolds (1)	0.120	0.117	1.128	1.031	1.063	1	0.303
pollenGrasses (1)	−0.312	0.112	0.732	−2.794	7.805	1	0.005
respiratoryInfection (1)	0.416	0.068	1.515	6.105	37.272	1	< .001
smellsOrScents (1)	0.126	0.089	1.134	1.418	2.012	1	0.156
tobacco (1)	0.223	0.101	1.250	2.215	4.908	1	0.027
unknown (1)	−0.039	0.064	0.961	−0.616	0.379	1	0.538
weatherChanges (1)	0.163	0.068	1.177	2.409	5.805	1	0.016
woodsmoke (1)	−0.098	0.144	0.906	−0.681	0.464	1	0.496
workplaceExposure (1)	0.311	0.195	1.365	1.597	2.550	1	0.110
animals (1) * weatherChanges (1)	−0.529	0.178	0.589	−2.967	8.803	1	0.003
acidReflux (1) * pollenGrasses (1)	0.700	0.242	2.013	2.894	8.378	1	0.004

Confidence bounds adjusted per significance level!

0.54	0.989
1.13	2.209
0.449	0.759
1.001	1.787
0.549	0.976
1.272	1.806
1.026	1.524
1.031	1.345
0.692	1.004
0.996	1.91
0.991	1.881
0.372	0.933
1.08	3.751
0.764	1.106
0.881	1.208
0.825	1.886
0.359	1.461
0.731	1.235
0.633	1.288
0.93	1.313
0.799	1.181
0.868	1.434
0.835	1.522
0.902	1.426
0.815	1.134
0.625	1.314

Multicollinearity Diagnostics

	Tolerance	VIF
acidReflux	0.470	2.127
airPollution	0.517	1.933
animals	0.376	2.661
anxietyOrStress	0.590	1.696
cockroaches	0.812	1.232
coldAir	0.601	1.663
dust	0.444	2.250
exercise	0.737	1.356
flu	0.759	1.317
foodAllergies	0.816	1.225
indoorAllergens	0.523	1.913
indoorMolds	0.607	1.647
medicines	0.694	1.441
other	0.686	1.458
outdoorAllergens	0.560	1.785
outdoorMolds	0.675	1.482
pollenGrasses	0.554	1.805
respiratoryInfection	0.539	1.856
smellsOrScents	0.595	1.681
tobacco	0.587	1.705
unknown	0.860	1.162
weatherChanges	0.479	2.086
woodsmoke	0.728	1.373
workplaceExposure	0.887	1.128
animals:weatherChanges	0.414	2.415
acidReflux:pollenGrasses	0.424	2.361

Logistic regression (Age)

```
Optimization terminated successfully.
         Current function value: 0.322323
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:        total_indicator   No. Observations:               219174
Model:                          Logit   Df Residuals:                   219169
Method:                           MLE   Df Model:                            4
Date:                Wed, 11 May 2022   Pseudo R-squ.:                0.002807
Time:                        17:22:05   Log-Likelihood:                -70645.
converged:                       True   LL-Null:                       -70844.
Covariance Type:            nonrobust   LLR p-value:                 8.656e-85
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                       -2.0126      0.016   -129.628      0.000      -2.043      -1.982
C(age_bins)[T.55_65]            -0.1406      0.019     -7.403      0.000      -0.178      -0.103
C(age_bins)[T.65_75]            -0.3161      0.020    -15.634      0.000      -0.356      -0.276
C(age_bins)[T.75_85]            -0.5225      0.034    -15.302      0.000      -0.589      -0.456
C(age_bins)[T.85_older]         -0.8574      0.310     -2.763      0.006      -1.466      -0.249
================================================================================================
```