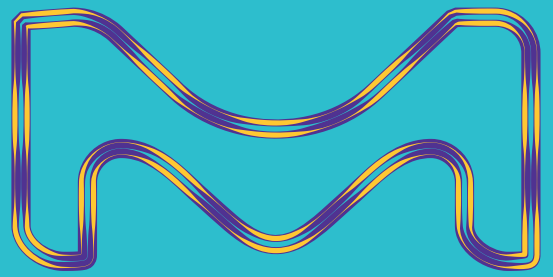
EXPLORE ADVERSE EVENTS ASSOCIATED WITH CONCOMITANT MEDICATIONS USING PREDICTION MODEL ESTABLISHED ON GRADIENT BOOSTED DECISION TREE (GBDT) METHOD

Yang Chong
Senior Program and Application Developer
Merck Serono



MERCK

Data Mining

Data Mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

Data Mining is the analysis step of the "knowledge discovery in databases" process, or KDD.

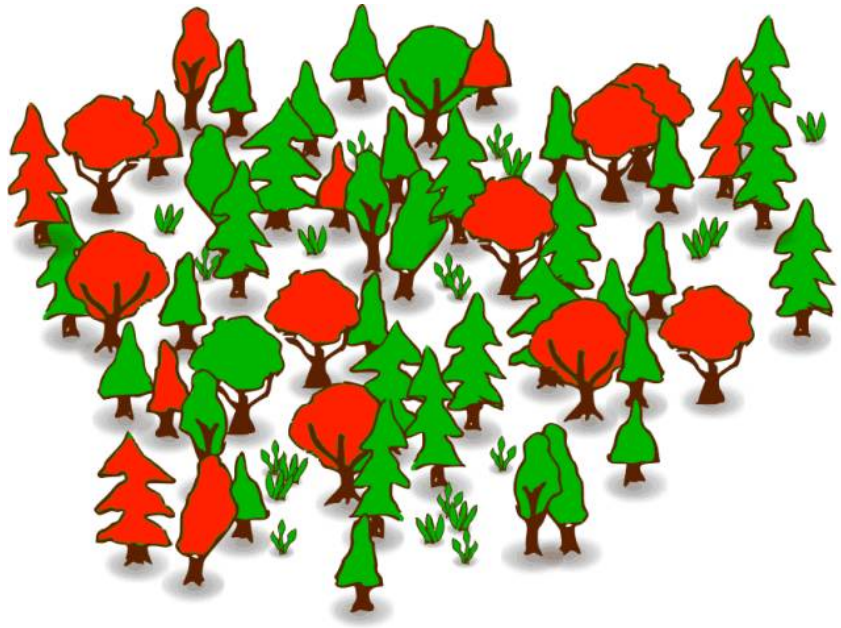
"We're drowning in information, but starving for knowledge."

MERCK

Machine Learning

Machine Learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed.

- Machine Learning explores the study and construction of algorithms that can learn from and make predictions on data.
- Machine Learning is the foundation of Artificial Intelligence.



MERCK

Gradient Boosting Decision Tree (GBDT)

Input: training set $\{(x_i, y_i)\}_{i=1}^{n}$, a differentiable loss function $L(y, F(x))$, number of iterations M.

Algorithm:

1. Initialize model with a constant value:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma).$$

2. For m = 1 to M:
    1. Compute so-called *pseudo-residuals*:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n.$$

    2. Fit a base learner (e.g. tree) $h_m(x)$ to pseudo-residuals, i.e. train it using the training set $\{(x_i, r_{im})\}_{i=1}^{n}$.
    3. Compute multiplier γ_m by solving the following one-dimensional optimization problem:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right).$$

    4. Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

3. Output $F_M(x)$.

MERCK

Data Mining in Clinical Trial

Data Sources

- Historical Trials
- Electronic Health Record
- Life Style, Vital Signs
- Drug database
- Hospital Data
- Patent/Scientific Publication
- Social Media

Machine Learning

- Association Rules
- Anomaly Detection
- Clustering
- Classification
- Prediction
- Natural Language Processing

Objectives

- Drug Discovering
- Time & Cost Reducing
- Site & Subject Selection
- Quality Control
- Adverse Event alert
- Query Prediction
- Cycle Time Risk
- Efficacy & Safety Assessment
- Visualization

MERCK

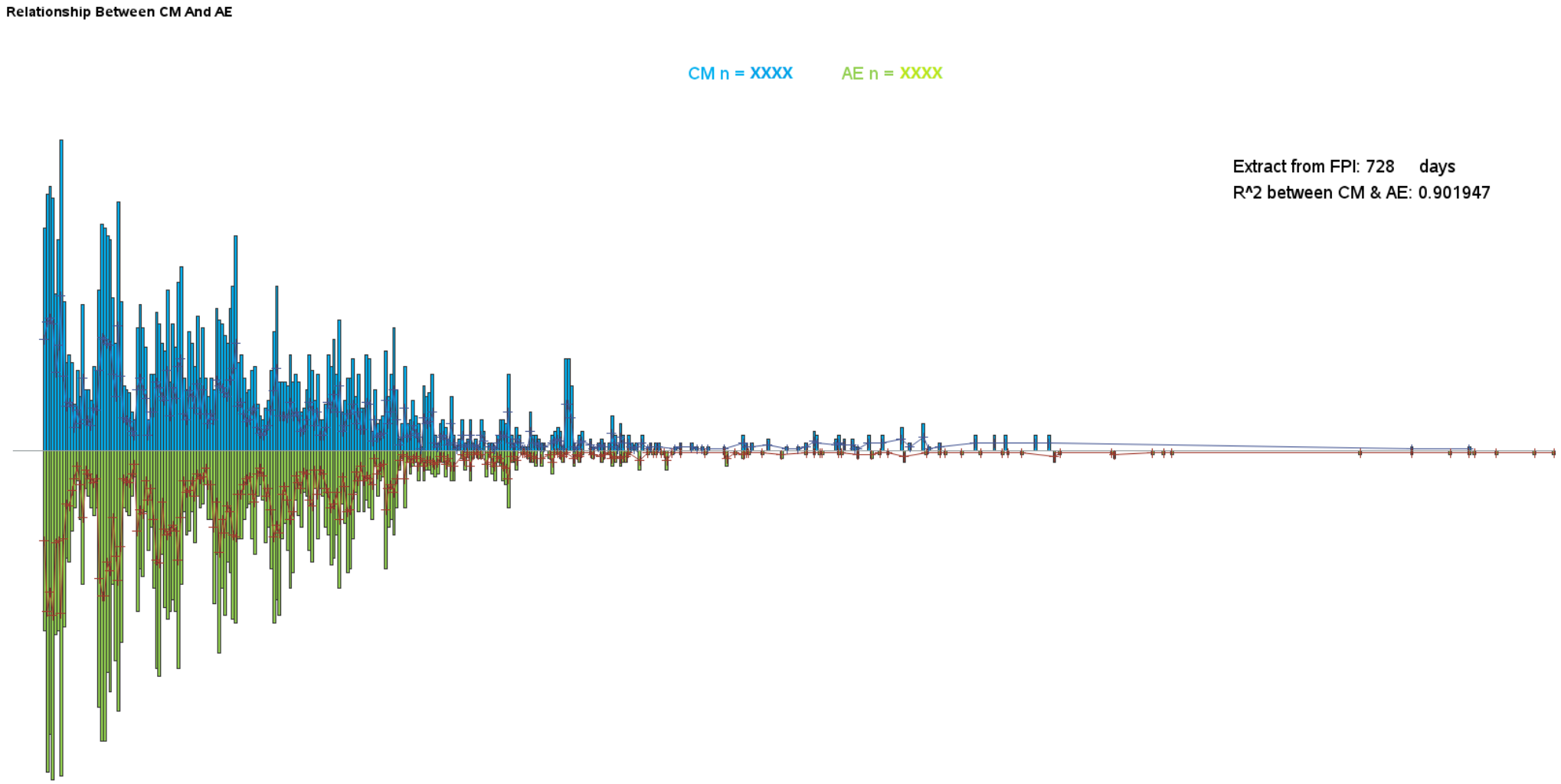
Concomitant Medications

CMTRT	CMDECOD	CMINDC	CMDOSE	CMDOSU	CMSTDTC	CMENDTC	CMSTDY	CMENDY	CMINDAE
MANNITOL		HYDRATION	125	mL	2016-01-08	2016-01-08	2	2	
PALONOSETRON	PALONOSETRON	PREVENTING V...	0.5	mg	2016-01-08	2016-01-12	2	6	
PANTOLOC		INHIBITS GAST...	40	mg	2016-01-08	2016-01-09	2	3	
LACTULOSE	LACTULOSE	RELAXING THE ...	15	mL	2016-01-09	2016-01-09	3	3	Y
CETYLPYRIDINI...		PREVENTION O...	15	mL	2016-01-10	2016-01-10	4	4	
MEGESTROL	MEGESTROL	PREVENT LOW ...	160	mg	2016-01-10	2016-01-10	4	4	
MOTILIUM	MOTILIUM ...	IMPROVE THE ...	10	mg	2016-01-10	2016-01-10	4	4	
MEGESTROLAC...	MEGESTROL ACE...	IMPROVES APP...	160	mg	2016-01-13	2016-01-14	7	8	Y
CARBOHYDRATE	CARBOHYDRATES	NUTRITION SU...	55.8	g	2016-01-14	2016-01	8	.	
CETYLPYRIDINI...		PREVENTION O...	15	mL	2016-01-14	2016-01-14	8	8	Y
KANGFUXIN LO...		PREVENTION O...	25	mL	2016-01-14	2016-01-14	8	8	Y
LACTULOSE	LACTULOSE	RELAXING THE ...	15	mL	2016-01-20	2016-01-20	14	14	Y
CETYLPYRIDINI...		PREVENT ORAL...	15	mL	2016-01-21	2016-01-21	15	15	Y
KANGFUXING L...	UNSPECIFIED HE...	PREVENT ORAL...	25	mL	2016-01-21	2016-01-21	15	15	Y

Adverse Events

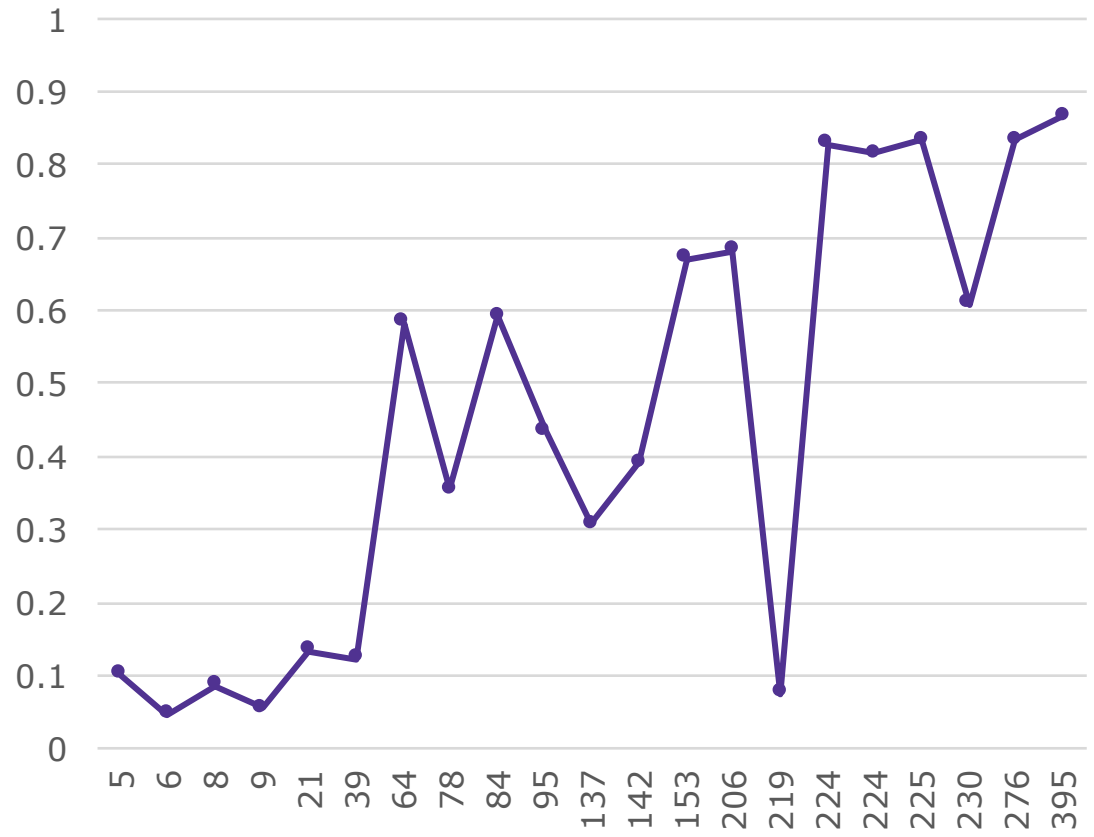
AETERM	AEDECOD	AEPTCD	AESOC	AESOCCD	AESER	AEREL	AETOXGR	AESTDTC	AEENDTC	AESTDY	AEENDY
CHEST DISTR...	Chest discomfort	10008469	General disorders...	10018065		UNRELATED	1	2016-01-04	2016-01-06	-3	-1
CONSTIPATION	Constipation	10010774	Gastrointestinal di...	10017947		UNRELATED	2	2016-01-05	2016-01-11	-2	5
CONSTIPATION	Constipation	10010774	Gastrointestinal di...	10017947		UNRELATED	1	2016-01-20	2016-01-21	14	15
CONSTIPATION	Constipation	10010774	Gastrointestinal di...	10017947		UNRELATED	1	2016-03-07	2016-03-08	61	62
DIZZY	Dizziness	10013573	Nervous system ...	10029205		UNRELATED	1	2016-02-14	2016-02-15	39	40
DIZZY	Dizziness	10013573	Nervous system ...	10029205		UNRELATED	1	2016-04-19	2016-04-20	104	105
HYPOALBUMIN...	Hypoalbuminaemia	10020942	Metabolism and n...	10027433		UNRELATED	1	2016-01-05	2016-01-27	-2	21
HYPOALBUMIN...	Hypoalbuminaemia	10020942	Metabolism and n...	10027433		UNRELATED	1	2016-02-03	2016-02-14	28	39
HYPOALBUMIN...	Hypoalbuminaemia	10020942	Metabolism and n...	10027433		UNRELATED	1	2016-02-23	2016-03-11	48	65
HYPOALBUMIN...	Hypoalbuminaemia	10020942	Metabolism and n...	10027433		UNRELATED	1	2016-03-21	2016-04-05	75	90
HYPOALBUMIN...	Hypoalbuminaemia	10020942	Metabolism and n...	10027433		UNRELATED	1	2016-04-13		98	.
HYPOCHLORID...	Hypochloraemia	10020955	Metabolism and n...	10027433		UNRELATED	1	2016-01-14	2016-01-21	8	15
HYPOCHLORID...	Hypochloraemia	10020955	Metabolism and n...	10027433		UNRELATED	1	2016-04-13		98	.
HYPOKALEMIA	Hypokalaemia	10021015	Metabolism and n...	10027433		UNRELATED	1	2016-03-21	2016-04-05	75	90
HYPOKALEMIA	Hypokalaemia	10021015	Metabolism and n...	10027433		UNRELATED	1	2016-04-13		98	.

The Relationship Between CM & AE



The Relationship Between R^2(CM vs AE) & Study Days

study	n	r^2
Merck001	8	0.085714
Merck002	219	0.074881
Merck003	64	0.583836
Merck004	142	0.390423
Merck005	230	0.60832
Merck006	395	0.867134
Merck007	276	0.833937
Merck008	206	0.682628
Merck009	224	0.827393
Merck010	9	0.054054
Merck011	153	0.669992
Merck012	39	0.121783
Merck013	95	0.434316
Merck014	137	0.306424
Merck015	5	0.10101
Merck016	224	0.815517
Merck017	6	0.045455
Merck018	21	0.132726
Merck019	78	0.35424
Merck020	84	0.592257
Merck021	225	0.833028



Data Space

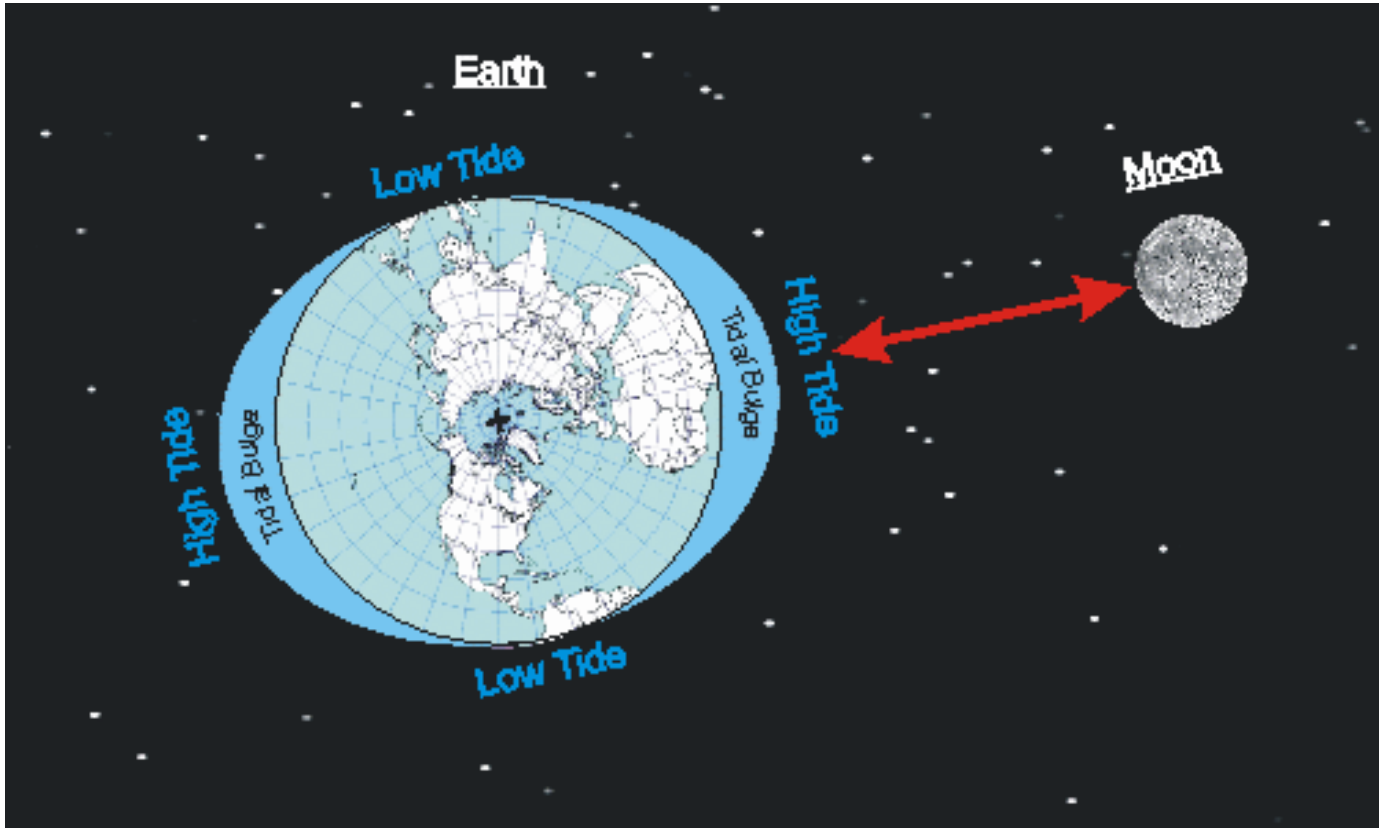
Expanding CMDECOD (n = 2245) & AEDECOD (n = 2370) by subject by event start date from Day 1（n = 7858)

stdy	CM_FUROSEMIDE	CM_CONTAC	CM_FLUTICASONE	CM_DUONASE	CM_LOPERAMIDE_HYDROCHLORIDE	CM_ONDANSETRON	CM_EPOPROSTENOL_SODIUM	CM_IBUPROFEN	CM_AZITHROMYCIN	CM_PROMETHAZINE_W_CODEINE	CM_BENZONATATE	CM_LEVOFLOXACIN	CM_PREDNISONE	CM_SODIUM_CHLORIDE	AE_Abdominal_discomfort	AE_Hypertension	AE_Pulmonary_hypertension	AE_Fatigue	AE_Diarrhoea	AE_Epistaxis	AE_Skin_ulcer	AE_Dyspnoea	AE_Dizziness	AE_Headache	AE_Nausea	AE_Vomiting
141	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
157	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
185	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
98	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
120	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
218	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
259	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
235	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
159	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
247	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
255	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
143	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
214	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
79	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

decod_stdtc_train_01

Parameters Setting in PYTHON

```
gbdt=GradientBoostingRegressor(
  loss='ls'
, learning_rate=0.1
, n_estimators=100
, subsample=1
, min_samples_split=2
, min_samples_leaf=1
, max_depth=3
, init=None
, random_state=None
, max_features=None
, alpha=0.9
, verbose=0
, max_leaf_nodes=None
, warm_start=False)
```



Predicted AE Values

Predicted AE values are floating-point number

A threshold is needed for 0-1 classification

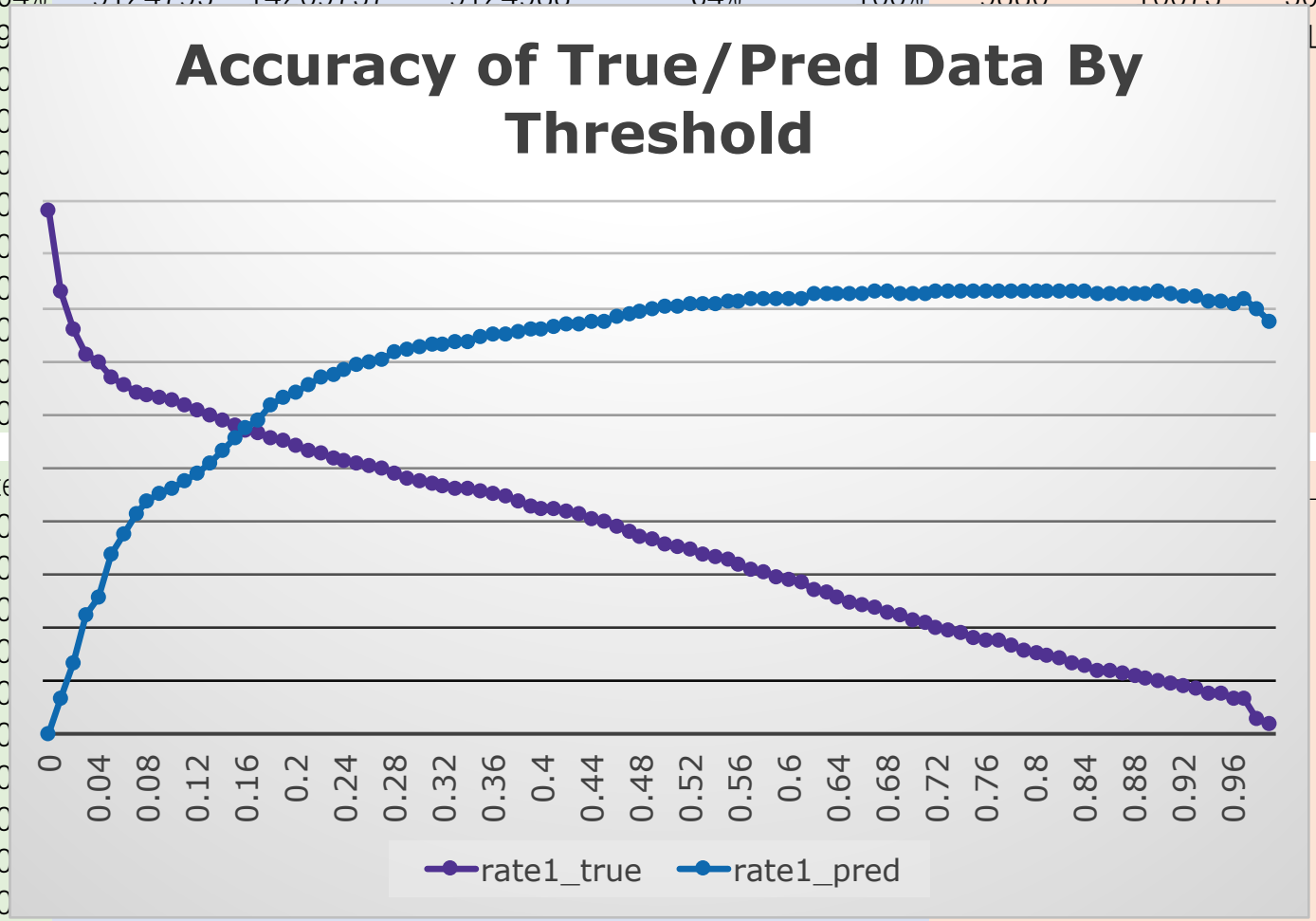
AE_Abdon	AE_Hypert	AE_Pulmor	AE_Fatigue	AE_Diarrhc	AE_Epistax	AE_Skin_ul	AE_Dyspnc	AE_Dizzine	AE_Headac	AE_Nausea	AE_Vomiti	AE_Arthral	AE_Hypoa	AE_Muscul	AE_Neck_p	AE_Produc	AE_Viral_u
-6.02E-05	0.006246	0.972191	0.016553	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.006636	0.022939	0.017067	0.005317	-0.00014	0.000823	0.001404	0.00208	0.002005
-6.02E-05	0.006246	3.54E-06	0.016553	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.006636	0.022939	0.017067	0.005317	-0.00014	0.000823	0.001404	0.841245	0.774686
-0.00017	0.006246	3.54E-06	0.016553	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.006636	0.022939	0.017067	0.005317	-0.00014	0.000823	0.001404	0.008017	0.057116
-0.00017	0.006246	3.54E-06	0.016553	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.006636	0.022939	0.017067	0.005317	-0.00014	0.000823	0.001404	0.00208	0.002005
0.000151	0.006246	3.54E-06	0.016553	0.838284	0.000298	4.45E-05	0.015821	0.0032	0.006636	0.022939	0.017067	0.005317	-0.00014	0.000823	0.001404	0.00208	0.002005
0.000109	0.006246	3.54E-06	0.016553	0.015887	0.000298	0.97737	0.015821	0.0032	0.006636	0.022939	0.017067	0.005317	-0.00014	0.000823	0.001404	0.00208	0.002005
-1.08E-06	0.006246	3.54E-06	0.02224	0.015887	0.000298	4.23E-05	0.015821	0.006379	0.006636	0.026376	0.017067	0.005317	0.000316	0.001259	0.001404	0.00208	0.002005
0.009624	0.006246	3.54E-06	0.016553	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.176195	0.022939	0.017067	0.037698	0.002828	0.000823	0.029824	0.00208	0.002005
-1.08E-06	0.006246	3.54E-06	0.016553	0.015887	0.000298	4.45E-05	0.070003	0.0032	0.006636	0.022939	0.017067	0.005317	-8.57E-05	0.000823	0.001404	0.00208	0.002005
-1.08E-06	0.006246	3.54E-06	0.016553	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.006636	0.022939	0.017067	0.005317	-8.57E-05	0.000823	0.001404	0.00208	0.002005
-1.08E-06	0.006246	3.54E-06	0.016553	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.006636	0.022939	0.017067	0.005317	-8.57E-05	0.000823	0.001404	0.00208	0.002005
0.000151	0.006246	3.54E-06	0.016553	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.133579	0.022939	0.017067	0.010352	-0.00014	0.000823	0.001404	0.00208	0.01802
0.000193	0.006246	3.54E-06	0.020278	0.015887	0.000298	4.45E-05	0.014303	0.0032	0.006636	0.028207	0.017067	0.005317	-0.00032	0.006311	0.001404	0.00208	0.002005
6.51E-05	0.61432	3.54E-06	0.020278	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.006636	0.028207	0.017067	0.005317	-0.00032	0.006311	0.001404	0.00208	0.002005
0.009624	0.006246	3.54E-06	0.020278	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.120681	0.026376	0.017067	0.037698	0.006009	0.001259	0.029824	0.00208	0.002005
0.009624	0.006246	3.54E-06	0.016553	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.120681	0.022939	0.017067	0.037698	0.002879	0.000823	0.029824	0.00208	0.002005
-1.08E-06	0.006246	3.54E-06	0.016553	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.006636	0.022939	0.017067	0.005317	-8.57E-05	0.031553	0.001404	0.00188	0.002005
-1.08E-06	0.006246	3.54E-06	0.020278	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.006636	0.026376	0.017067	0.005317	0.000316	0.0015	0.001404	0.00208	0.002005
-1.08E-06	0.006246	3.54E-06	0.016553	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.117079	0.022939	0.017067	0.012619	0.001988	0.000823	0.001404	0.00208	0.002005
-1.08E-06	0.006246	3.54E-06	0.017426	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.006636	0.022939	0.017067	0.100662	-8.57E-05	0.000823	0.001404	0.00208	0.002005
-1.08E-06	0.006246	3.54E-06	0.016553	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.006636	0.022939	0.017067	0.005317	-0.00014	0.000823	0.001404	0.00208	0.002005
-1.08E-06	0.006246	3.54E-06	0.020278	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.006636	0.026376	0.017067	0.005317	-8.57E-05	0.000823	0.001404	0.00208	0.002005
-1.08E-06	0.006246	3.54E-06	0.016553	0.015887	0.000298	4.45E-05	0.015821	0.0032	0.006636	0.022939	0.017067	0.005317	-8.57E-05	0.000823	0.001404	0.00208	0.002005

y_pred

Threshold Setting

threshold	count	sum	rate	count0	sum0_true	sum0_pred	rate0_true	rate0_pred	count1	sum1_true	sum1_pred	rate1_true	rate1_pred
0	9134675	14219830	64%	9124795	14209757	9124988	64%	100%	9880	10073	5094842	98%	0%
0.01	14099301	14219830	9								127214	83%	7%
0.02	14167335	14219830	10								57710	76%	13%
0.03	14192041	14219830	10								32082	71%	22%
0.04	14196191	14219830	10								27624	70%	25%
0.05	14203204	14219830	10								20043	67%	34%
0.06	14205431	14219830	10								17486	65%	38%
0.07	14207075	14219830	10								15600	64%	41%
0.08	14207866	14219830	10								14697	64%	44%
0.09	14208355	14219830	10								14114	63%	45%
0.1	14208678	14219830	10								13689	63%	46%

threshold	count	sum	rate								_pred	rate1_true	rate1_pred
0.9	14210569	14219830	10								1226	10%	83%
0.91	14210509	14219830	10								1156	9%	83%
0.92	14210470	14219830	10								1101	9%	82%
0.93	14210415	14219830	10								1028	8%	82%
0.94	14210365	14219830	10								970	8%	81%
0.95	14210330	14219830	10								915	7%	81%
0.96	14210284	14219830	10								855	7%	81%
0.97	14210261	14219830	10								800	6%	82%
0.98	14209985	14219830	10								382	3%	80%
0.99	14209882	14219830	10								229	2%	77%
1	14209779	14219830	100%	14209750	14209757	14219794	100%	100%	29	10073	36	0%	81%



Scenario: AE risk

Threshold = 1:

		True AE		
		Yes	No	Total
Pred AE	Yes	29	7	36
	No	10044	14209750	14219794
	Total	10073	14209757	14219830

Sensitivity = 29/10073 = 0.29%

Specificity = 14209750/14209757 = 99.99%

Accuracy of the model = (29 + 14209790)/14219830 = 99.93%

False positive rate = 7/14209757 = 0%

False negative rate = 10044/10073 = 99.71%

subject006 2 AE_Vomiting

CMTRT	CMDECOD	CMSTDTC	CMSTDY
20% MANNITOL	MANNITOL	2015-10-30	2
METOCLOPRAM...	METOCLOPRAMIDE DIHYDROCHLORIDE	2015-10-30	2
PIDOTIMOD OR...	PIDOTIMOD	2015-10-30	2

AETERM	AEDECOD	AESTDTC	AESTDY
NAUSEA	Nausea	2015-10-30	2

AETERM	AEDECOD	AESTDTC	AESTDY
NAUSEA	Nausea	2015-10-30	2
VOMIT	Vomiting	2015-10-31	3
VOMIT	Vomiting	2015-10-31	3

Q & A