

PP19: Utilizing Machine Learning Algorithms to Detect Outliers or Inconsistencies in the Analysis Results in Datasets and Summary Reports

Linshan Yuan (LYuan7@ITS.JNJ.com), Weiwei Xu (wxu52@ITS.JNJ.com)

- Introduction

Safety related summary tables (e.g. AE, LB, VS, EG) are a core part of the Clinical Study Report (CSR) and are usually lengthy reports when we display the statistical results by visits/time points. In this poster we will explain how we applied multiple ML methods to 1) detect the outliers (e.g. extreme values in small Ns) and 2) obtain data pattern related to the derivation of Treatment Emergent Adverse Event (TEAE) flag.

- Methods

Method 1: Unsupervised Outlier Identification to identify abnormal counts across all the small Ns by visits/time points in the summary table.

ML parameters	Unsupervised ML methods			
	Isolation forest	Local outlier factor	One class SVM	Elliptic envelope Sklearn
contamination	0.1, 0.2, 0.3	0.1, 0.2	/	0.1, 0.2, 0.3, 0.4
n_neighbors	/	Less than sample size -1	/	/
Nu	/	/	0.1, 0.2	/

By adjusting the sensitivity on how many are higher or lower based on study data to concentrate the more positive alarm for outlier/abnormal values.

Case 1: extreme values of small Ns

Load extracted small Ns to model and outliers are selected based on contamination level.

Vital signs summary table’s small Ns by each time point as: Baseline, Week 1 (pre, 30 min, 1 hour), 2, 3, 4...

vsdata=[100,100,97,97,97,94,94,94,93,91,89,87,84,83,81,81,78,78,78,78,78,74,74,66,65,66,66,73,2,4,67,67,63,64,62,61,63,60,60,57,56,54]

	Unsupervised ML methods			
	Isolation forest	Local outlier factor	One class SVM	Elliptic envelope Sklearn
Abnormal	cont.=0.1:	n_ne.=10, cont.=0.1:	Nu=0.1:	cont.=01:
small N	[100, 100, 2, 4, 54]	[2, 4, 57, 56, 54]	[74, 74, 2, 4, 56, 54]	[100, 100, 2, 4, 54]
identified	cont.=0.2:	n_ne.=20, cont.=0.1:	Nu=0.2:	cont.=0.2:
	[100, 100, 2, 4, 57, 56, 54]	[100, 100, 2, 4, 54]	[100, 100, 97, 97, 97, 2, 4, 57, 56, 54]	[100, 100, 97, 97, 97, 2, 4, 56, 54]
Positive alarm	[2, 4]	[2, 4]	[2, 4]	[2, 4]

Case 2: hidden outlier using other method (Small N is either higher or lower than its neighbor)

Lab Cholesterol table’s small Ns by each time point as: Baseline, Week 4, 8, 12, 24.

choldata=[98,90,84,86,82]

	Unsupervised ML methods			
	Isolation forest	Local outlier factor	One class SVM	Elliptic envelope Sklearn
Abnormal	cont.=0.1: [98]	n_ne.=4, cont.=0.1: [90]	nu=0.1: [98, 90]	cont.=01:[98]
small N	cont.=0.2: [98]	n_ne.=4, cont.=0.2: [90]	nu=0.2: [98, 90]	cont.=0.2: [98]
identified	cont.=0.3: [98, 82]			cont.=0.3: [98, 82]
Positive alarm	[82]	[90]	[90]	[82]

Note: cont.=contamination, n_ne.=n_neighbor

- Methods

Combined with rule-based algorithms

With trial specific knowledge, rule-based algorithms can be applied as additional check to identify the outlier.

Rule-based algorithms	Case 1: vital signs data	Case 2: lab cholesterol data
Higher than neighbors	[73] (negative alarm)	[86] (positive alarm)
Lower than neighbors	none	[84] (positive alarm)

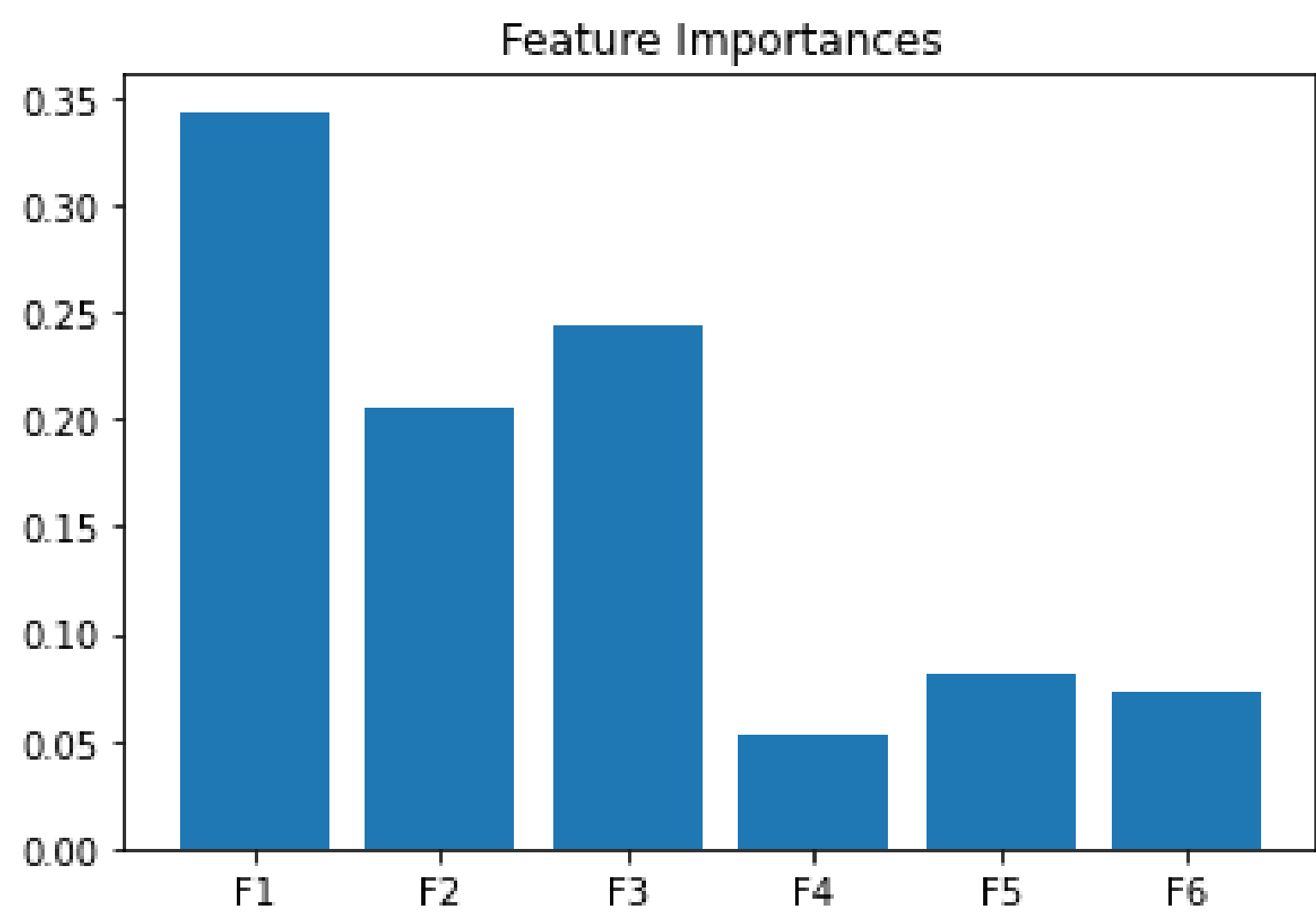
Method 2: Decision Tree to get the pattern of AE start date, AE end date, treatment start date related to the derivation of TEAE flag.

- Features from decision tree are listed in order of importance

- Performance evaluation:

Accuracy: 0.96

Precision: 0.97



- Results

- Method 1/Case 1: In this case, we found out the apparent anomaly “2” or “4” are results of mal derivation of analysis flag in ADaM. We then updated the analysis program to address the issue right away.

- Method 1/Case 2: The count at baseline differs a lot from count at last visit/time point. We can check the study data to verify whether it represents true values (e.g. xx subjects withdraw from the study) or it’s caused by incorrect data entry, derivation errors.

- Method 1: Local outlier factor and One class SVM have more abnormal elements identified from the middle of list.

- Method 2: Logic pattern of features to the outcome can help us set up the rule-based algorithm. We can harness the ML resources to facilitate our analysis in more complex scenarios without depending on subject-area knowledge.

- Conclusion

- Applying multiple ML methods along with other known rule-based algorithms can help us efficiently filter abnormal values through all of summary reports.

- A combination of multiple ML methods could provide a more comprehensive view of the outliers.