Секция «Теория вероятностей и математическая статистика»

Диагностика условий регрессионной модели на примере астрономических данных

Научный руководитель – Шкляев Александр Викторович

Шигин Глеб Сергеевич

Студент (специалист)

Московский государственный университет имени М.В.Ломоносова, Факультет космических исследований, Москва, Россия

E-mail: gleb.shigin@gmail.com

В работе рассматривается линейная регрессионная модель для описания фотометрических красных смещений галактик на основе их многоволновой фотометрии. Первоисточником стала работа [3], в которой применена обобщенная линейная модель с логарифмической функцией связи. Однако, в указанной работе не проводится стандартных для данного типа работ анализа остатков, диагностики модели и обоснование ее выбора.

Диагностика модели оригинальной работы позволила выявить два недостатка – чрезмерную сложность модели и вызывающий сомнения выбор класса распределений остатков. В ходе работы модель была доработана, сначала заменой GLM модели на более простую Lasso регрессию, затем добавлением кластеризации данных. Ввиду вычислительной сложности алгоритма кластеризации (в нашем случае – $O(n \log n)$ для OPTICS [1]) кластеризация была выполнена на небольшой подвыборке точек ($\sim 10\%$) с последующей классификацией всех точек с помощью метода опорных векторов (SVM [2]).

В результате нами была создана новая регрессионная модель, которая оказалась значительно более эффективной с точки зрения рассмотренной в оригинальной статье метрики.

Также, в работе обращается внимание на существенные различия в структуре данных датасета PHoto-z Accuracy Testing (PHAT, [4]), в котором целевые данные являются настоящими, а обучающие – искусственно сгенерированными.

Источники и литература

1) Ankerst M. и др. OPTICS // Proceedings of the 1999 ACM SIGMOD international conference on Management of data. 1999.
2) Cortes C., Vapnik V. Support-vector networks // Mach Learn. 1995. Т. 20. № 3. С. 273–297.
3) Elliott J. и др. The overlooked potential of Generalized Linear Models in astronomy-II: Gamma regression and photometric redshifts // Astronomy and Computing. 2015. Т. 10. С. 61–72.
4) Hildebrandt H. и др. PHAT: PHoto-zAccuracy Testing // A&A. 2010. Т. 523. С. A31.

Иллюстрации

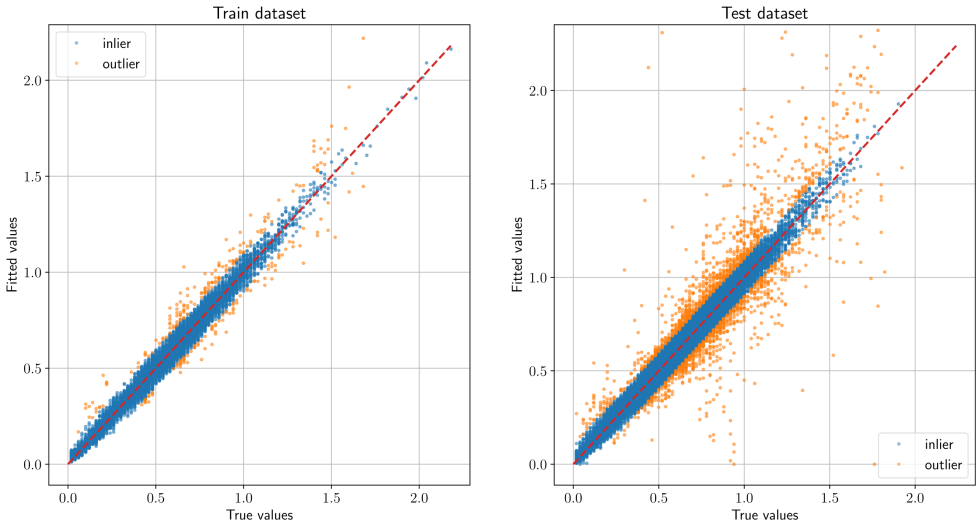

Рис. : True vs. Fitted для GLM, исходное решение из [3]

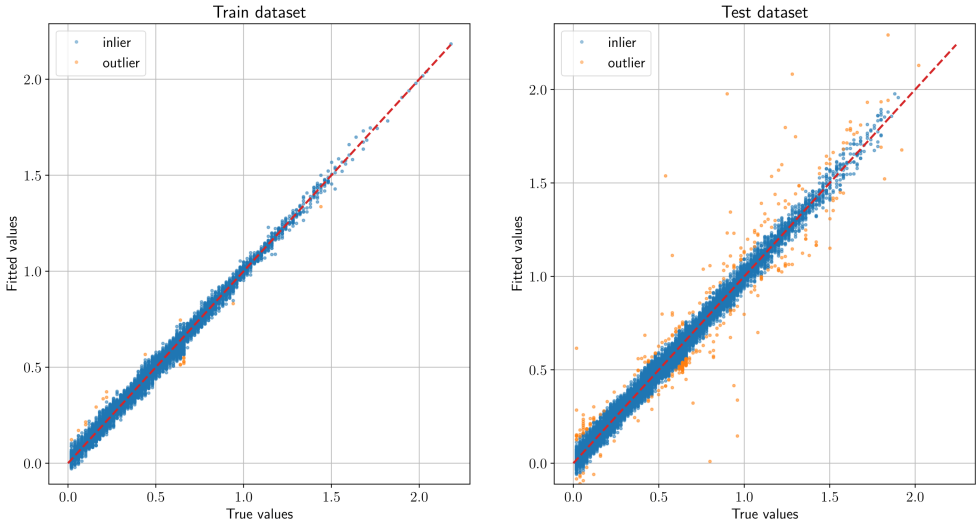

Рис. : True vs. Fitted, замена GLM на Lasso регрессию

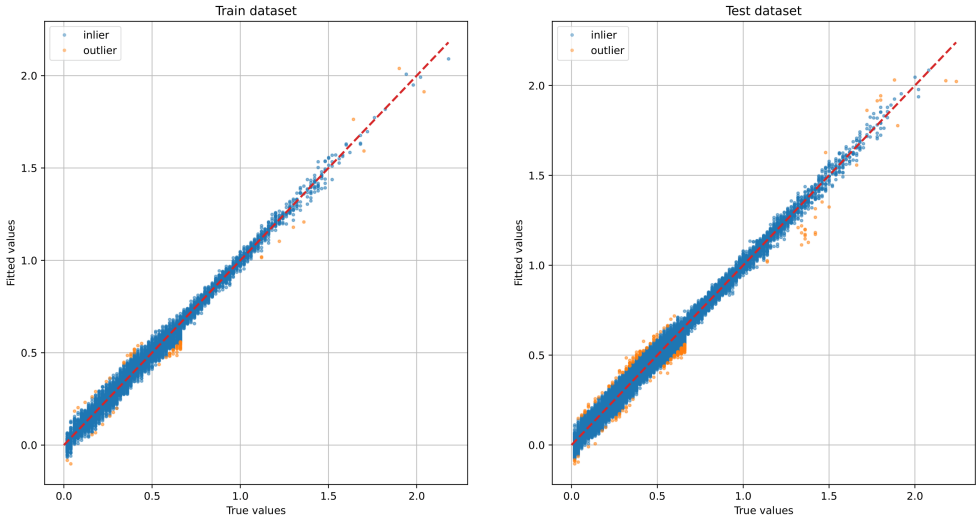

Рис. : True vs. Fitted, OLS с кластеризацией данных, итоговый результат