**ИСПОЛЬЗОВАНИЕ ИНФОРМАЦИИ О ПОГРЕШНОСТИ ИЗМЕРЕНИЯ ПЕРЕМЕННЫХ ДЛЯ ДОСТОВЕРНОЙ ЭКСТРАПОЛЯЦИИ РЕГРЕССИИ**

**Постановка задачи**

Имеется набор данных dataset\_0, который содержит количественные предикторы A\_1, A\_2, A\_3, A\_4, A\_5 и количественную целевую переменную b. Набор данных dataset\_0 представляет собой точную зависимость целевой переменной b от предикторов.

Даны два набора данных dataset\_1 и dataset\_2, каждый из которых содержит количественные предикторы A\_h\_1, A\_h\_2, A\_h\_3, A\_h\_4, A\_h\_5 и количественную целевую переменную b\_d. Наборы данных dataset\_1 и dataset\_2 моделируют результаты измерений b от предикторов и сгенерированы по данным dataset\_0 путем случайных отклонений предикторов в пределах 1% и целевой переменной b в пределах 5% от соответствующих точных значений из dataset\_0.

Необходимо по данным dataset\_1 и dataset\_2 спрогнозировать значение целевой переменной b\_max при значениях предикторов (1800, 1500, 1800, 2000, 1500). Датасет dataset\_0 использовать только для валидации обученных моделей.

**РЕШЕНИЕ ЗАДАЧИ**

**1. Характеристика датасетов**

Значения предикторов в каждом наборе данных являются положительными вещественными числами, максимальное значение которых не превышает 1269. Значения целевой переменной в каждом наборе данных также являются положительными вещественными числами, максимальное значение которых не превышает 115.

Для dataset\_0 выявлена функциональная линейная зависимость целевой переменной от всех предикторов. При этом имеется практически абсолютная мультиколлинеарность всех предикторов.

Для dataset\_1 и dataset\_2 коэффициенты корреляции для всех предикторов между собой практически равны 1, а для целевой переменной с предикторами практически равны 0.99.

Таким образом, для dataset\_1 и dataset\_2 также имеются практически функциональная линейная зависимость целевой переменной от предикторов и практически абсолютная мультиколлинеарность всех предикторов.

**2. Обучение различных моделей**

Данная задача относится к классу задач экстраполяции.

Обучение моделей выполнено на языке Python 3.

Обучение проводилось как при разбиении на обучающую и тестовую выборки (70/30), так и на полных наборах данных.

Модели строились на всех предикторах. Все предикторы масштабировались с помощью MaxAbsScaler.

Для построения моделей применялись:

1) алгоритмы LinearRegression, Ridge, Lasso, ElasticNet, TheilSenRegressor из библиотеки scikit-learn с поиском оптимальных значений их гиперпараметров по сетке с кросс-валидацией.

2) алгоритм регуляризации Тихонова с поиском коэффициента регуляризации методом обобщенной невязки [1].

Отличие алгоритма регуляризации Тихонова от модели Ridge из scikit-learn состоит в том, что коэффициент регуляризации определяется по условию оптимизации нестандартной метрики – обобщенной невязки, которая учитывает информацию о погрешности измерения предикторов и целевой переменной.

Решение системы линейных алгебраических уравнений для модели на основе регуляризации Тихонова имеет следующий вид

(1)

Уравнение обобщенной невязки для определения коэффициента регуляризации имеет следующий вид

(2)

(3)

(4)

где – обозначение нормы матрицы или вектора;

– относительная погрешность измерения элементов матрицы A;

– относительная погрешность измерения элементов вектора b.

Для данной задачи = 0,01, = 0,05.

Метрики и результаты прогноза целевой переменной (b\_max) и их осредненные значения (Mean) для различных моделей регрессоров из библиотеки scikit-learn приведены в таблицах 1 – 7. Там же приведены результаты прогноза для алгоритма регуляризации Тихонова (Regularized) и точный прогноз (Accurate).

Таблица 1 – Метрики и результаты прогноза целевой переменной для dataset\_1

(random\_state = 1937 для train\_test\_split)

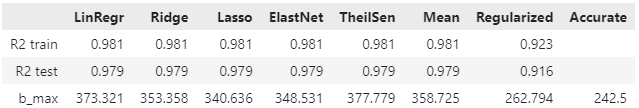


Таблица 2 – Метрики и результаты прогноза целевой переменной для dataset\_1

(random\_state = 1941 для train\_test\_split)

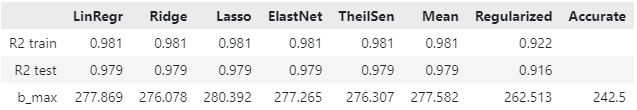


Таблица 3 – Метрики и результаты прогноза целевой переменной для dataset\_1

(обучение на полном наборе данных)

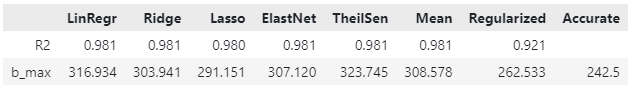


Таблица 4 – Метрики и результаты прогноза целевой переменной для dataset\_2

(random\_state = 1968 для train\_test\_split)

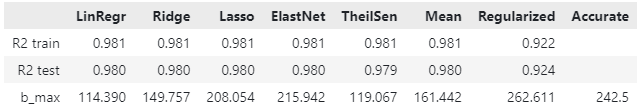


Таблица 5 – Метрики и результаты прогноза целевой переменной для dataset\_2

(random\_state = 1901 для train\_test\_split)

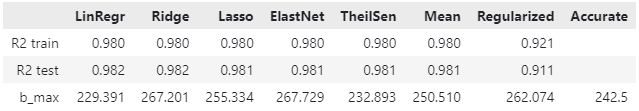


Таблица 6 – Метрики и результаты прогноза целевой переменной для dataset\_2

(обучение на полном наборе данных)

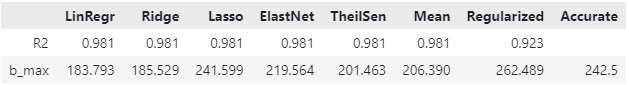
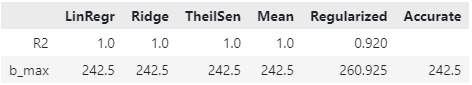


Таблица 7 – Метрики и результаты прогноза целевой переменной для dataset\_0

(обучение на полном наборе данных)



Примечание. dataset\_0 сгенерирован при коэффициентах z = (0.04, 0.035, 0.03, 0.02, 0.016)

**3. Анализ прогнозов различных моделей**

Для dataset\_1 диапазон осредненного прогноза (Mean) различных моделей регрессоров из библиотеки scikit-learn составил от 277,6 до 358,7. Диапазон прогноза регуляризованного решения с учетом информации о погрешности измерения предикторов и целевой переменной составил от 262,5 до 262,8.

Для dataset\_2 диапазон осредненного прогноза (Mean) различных моделей регрессоров из библиотеки scikit-learn составил от 161,4 до 250,5. Диапазон прогноза регуляризованного решения с учетом информации о погрешности измерения предикторов и целевой переменной составил от 262,1 до 262,6.

Для dataset\_0 точное значение прогноза составляет 242,5. Это же значение получается при вычислении прогноза по выражению , где z = (0.04, 0.035, 0.03, 0.02, 0.016) – коэффициенты регрессии для генерации dataset\_0. Значение прогноза регуляризованного решения с учетом информации о погрешности измерения предикторов и целевой переменной составило 260,9.

Таким образом, несмотря на отличные метрики всех моделей регрессоров из библиотеки scikit-learn, результаты их прогнозов не только имеют неприемлемую точность, но и сильно отличаются друг от друга, то есть являются неустойчивыми – небольшие погрешности измерений данных приводят к недопустимым погрешностям прогноза целевых переменных. При этом модель регуляризации Тихонова с учетом информации о погрешности измерения предикторов и целевой переменной выдает устойчивое решение с приемлемой точностью. Так в данной задаче погрешность осредненного прогноза различных моделей регрессоров из библиотеки scikit-learn составила от -33% до 48%, тогда как погрешность прогноза модели регуляризации Тихонова составила 8%.

**4. Выводы**

1. Игнорирование даже незначительной погрешности измерения переменных может привести к крайне большой погрешности прогноза экстраполяции.

2. Модели с мультиколлинеарными предикторами могут привести к неустойчивым прогнозам экстраполяции с недопустимой погрешностью.

3. Для моделей с мультиколлинеарными предикторами для получения устойчивого прогноза экстраполяции с допустимой погрешностью следует использовать регуляризацию с учетом информации о погрешности измерения предикторов и целевой переменной.

**Список литературы**

1. Тихонов А. Н., Гончарский А. В., Степанов В.В., Ягола А. Г. Регуляризирующие алгоритмы и априорная информация. – М.: Наука, 1983. – 200 с.