Задание на лабораторную работу №3 “Прогнозирование количества подвижных единиц грузового поезда в сходе с рельсов на основе квантильной регрессии”

по курсу “Управление рисками на железнодорожном транспорте”

1. Ознакомиться со структурой датасета, в котором хранятся сведения о результатах сходов/крушений подвижного состава вне стрелочных переводов по причине неисправности подвижного состава.

length	commonlength	maxder	dcar	speed	weight	load	curve	profile
81	84	82	26	54	2080	0,038826	0,002463	-0,008
…	…	…	…	…	…	…	…	…

length – количество вагонов в поезде;

commonlength – количество вагонов и секций локомотива в поезде;

maxder – количество вагонов и секций локомотива в поезде минус номер первой сошедшей с рельсов подвижной единицы (от головы поезда) + 1;

dcar – количество вагонов и секций локомотива в сходе;

speed – скорость поезда в месте схода, км/ч;

weight – вес поезда, т.;

load – показатель, характеризующий степень загруженности поезда полезной нагрузкой, чем меньше порожних вагонов, тем больше этот показатель;

curve – кривизна пути в месте схода;

profile – величина профиля пути в месте схода в тысячных.

2. Загрузить указанный датасет в любой математический пакет по выбору студента.

3. Построить различные зависимости вида dcar=f(length, …, profile) на основе квантильной регрессии, полагая, что

$$\mathrm{dcar}_k = \mathrm{f}\left(\mathrm{length}_k, \dots, \mathrm{profile}_k\right) + \varepsilon_k,$$

где $k = 1, .., n$, n – количество наблюдений, $\varepsilon_1, \dots, \varepsilon_n$ – независимые одинаково распределенные ошибки. Уровень квантильной регрессии выбрать равным 0.25, 0.5, 0.75.

Построить не менее 12 таких зависимостей (по 4 на каждый уровень). Проводить построение зависимостей на выборках с вычеркиванием строк, в которых имеется хотя бы одно пропущенное наблюдение чего-либо; а также необходимо проводить построение

зависимостей с вычеркиванием только тех строк, в которых пропущено необходимое для построения зависимости наблюдение.

Для каждой построенной зависимости привести значения скорректированного коэффициента детерминации, средней абсолютной погрешности, средней относительной погрешности, оценку дисперсии ошибок.

Результаты должны быть представлены в виде следующей таблицы

Зависимость	Уровень	R^2_{adj}	Δ (средняя абсолютная погрешность)	δ (средняя относительная погрешность)	$\hat{\sigma}^2$ (оценка дисперсии ошибок)
dcar=0.1speed+0.2	…	0.95	1.17	0.01	2
…	…	…	…	…	…

5. Определить наилучшую из построенных зависимостей и объяснить, почему она, на взгляд студента, является наилучшей.

6. Определив вид наилучшей зависимости, разбить выборку на две части в пропорции: 70 на 30. Выбор разбиения предоставляется студенту: например, можно выбрать эти 70 процентов наблюдений случайным образом, можно же взять первые 70 процентов наблюдений по порядку их нахождения в датасете. На 70 процентах наблюдений провести обучение модели (то есть подбор неизвестных коэффициентов). На этих 70 процентах наблюдений вычислить значения

- скорректированного коэффициента детерминации;
- средней абсолютной погрешности;
- средней относительной погрешности;
- оценки дисперсии ошибок.

А также для полученной модели на 30 процентах наблюдений, на которых модель не обучалась, привести значения

- скорректированного коэффициента детерминации;
- средней абсолютной погрешности;
- средней относительной погрешности;
- оценки дисперсии ошибок.

Сделать вывод из полученных результатов.

Литература

1. Горяинова Е.Р., Панков А.Р., Платонов Е.Н. Прикладные методы анализа статистических данных. М.: Изд. дом Высшей школы экономики, 2012.

2. https://www.dependability.ru/jour/article/view/255/438

3. https://lms.mai.ru/mod/resource/view.php?id=265109 (датасет)