DOI: 10.25728/mlsd.2023.0950

РЕГРЕССИОННЫЕ МОДЕЛИ ПРОГНОЗА РАЗМЕРОВ ОТСТУПЛЕНИЙ ЖЕЛЕЗНОДОРОЖНОГО ПОЛОТНА

Владова А.Ю.
Институт проблем управления им. В.А. Трапезникова РАН, Москва, Россия
Финансовый университет при Правительстве РФ, Москва, Россия
avladova@ipu.ru

Аннотация. Прогноз параметров отклонений железнодорожных путей обычно опирается на три основных типа моделей: авторегрессионные модели, регрессионные модели и их гибриды. Предложенный метод прогнозирования размеров отклонений железнодорожного пути на двумерной сетке по модели линейной регуляризации включает три этапа: подготовку данных, обучение и оценка качества прогноза, визуализация результатов.

Ключевые слова: регрессия, факторы, рельс, большие данные.

Введение

Железные дороги функционируют под влиянием природных факторов и нагрузок проходящих поездов. Диагностические лаборатории измеряют геометрию дорог и выявляют отклонения параметров рельсов от нормативных значений. Эти отклонения можно классифицировать на внутренние дефекты головки рельса, дефекты рельсового полотна/стопы, приповерхностные дефекты, дефекты поверхности (рифления рельсов, проседания, поверхностные трещины), дефекты колес и подшипников (горячие колеса и горячие подшипники), неровности профиля рельса (износ рельсов и усталость контакта качения) и неровности геометрии пути (ширина колеи, скручивание, поперечный уровень, продольный уровень и выравнивание). Датчики для обнаружения отступлений классифицируют на датчики неразрушающего контроля, камеры, оптические лазеры, преобразователи акселерометра, механические датчики и дополнительные датчики [1]. При многоканальной записи данные согласуются по времени с помощью Глобальной системы позиционирования (GPS). Сигналы, полученные со всех каналов, передискретизируют до 100 Гц, чтобы обеспечить постоянную частоту дискретизации для всех наборов данных [2]. По измеренным амплитудным значениям сигналов акустической эмиссии судят о наличии и длине локальных дефектов поверхности катания железнодорожных рельсов, которая пропорциональна линейной скорости движения колеса и обратно пропорциональна длительности акустико-эмиссионного сигнала [3]. Решение о техническом обслуживании может быть принято в режиме онлайн на основе тенденции, присутствующей в данных. Однако невозможно каждый раз принимать решение, основываясь на тренде, так как данные, полученные с датчиков, зашумлены. Для обработки данных и извлечения полезной информации используют передовые алгоритмы обработки сигналов [4]. Эти алгоримы реализованы в библиотеках прогнозирования временных рядов Silverkite, Prophet и Fedot [5]. Прогноз состояния железнодорожных путей обычно опирается на три типа моделей: авторегрессионные [6] регрессионные [7] и их гибриды.

Выход размеров отклонений за нормативные значения сигнализирует о необходимости снижения скорости движения поездов. Несмотря на значительные успехи в предиктивном обслуживании протяженных объектов, адаптация известных методов прогнозирования временных рядов к оценке размеров отклонений остается ограниченной. Предлагаемое исследование направлено на преодоление этого разрыва за счет проведения статистического анализа пространства признаков; формирования входных, выходных и управляющих параметров; создания равномерной двумерной сетки, и прогнозирования размеров отклонений в узлах сетки.

1. Постановка задачи

В качестве исходных данных имеем многомерный датасет, характеризующий отступления железнодорожного полотна по набору признаков: амплитуды, нормативные значения амплитуды, длины, степени опасности, местоположения, коды и время фиксации отступлений. Поскольку кроме времени мы можем выделить местоположение отступлений, то оставшиеся признаки можно привязать к двум измерениям: по времени t и по пространству L. Дополнительно можно использовать код отступлений, разделяя двумерный датасет на количество наборов данных K, равное количеству кодов отступлений. Тогда для каждого k = 1,…, K мы можем определить подмножество объектов Xk, которые имеют код отступления k. Таким образом, каждый набор данных, характеризующий определенный тип отступлений, имеет две оси: время и пространство (рис. 1).

		АМПЛИТУДА	СТЕПЕНЬ	ДЛИНА
ВРЕМЯ	ДИСТАНЦИЯ			
2021-02-22	643.556	17	1	37
2021-02-24	649.982	19	1	10
	649.970	21	2	8
	647.911	11	1	20
	648.006	11	1	20

Рис. 1. Фрагмент двумерного датасета для отступлений с кодом (P)

На каждой оси мы разбиваем диапазон значений на равные интервалы [ti-1,ti] и [Lj-1,Lj] и получаем сетку из ячеек (i, j). В каждой ячейке мы строим регрессионную модель $f_k(X)$ для тех отступлений, которые попали в эту ячейку по времени и пространству: $t_{i-1} \le T < t_i$ и $L_{j-1} \le L < L_j$.

1.1. Выбор целевого признака

В качестве целевого признака y(L, t), значения которого мы должны предсказать, можно выбрать любой из привязанных признаков X(L, t): амплитуда, длина, степень опасности или их комбинацию: площадь по амплитуде, как произведение длины на амплитуду; площадь по отклонению, как произведение длины на отклонение амплитуды от нормативного значения. Возможно вычислить также не только абсолютные значения признаков, но и относительные: отношение амплитуды к нормативной амплитуде, отношение амплитуды или длины к средней амплитуде или длине по каждому отступлению (рис. 2). Это поможет учесть разный масштаб признаков и сравнить их между собой.

		ПЛОЩАДЬ_А	ОТКЛОНЕНИЕ	ПЛОЩАДЬ_О	АМПЛИТУДА_эксп	ПЛОЩАДЬ_Акорень
ВРЕМЯ	ДИСТАНЦИЯ					
2018-05-18	19.146	594	18	594	6.565997e+07	24.372115
	19.801	182	13	182	4.424134e+05	13.490738
2018-07-24	19.839	442	17	442	2.415495e+07	21.023796
2018-08-13	19.595	437	19	437	1.784823e+08	20.904545

Рис. 2. Фрагмент двумерного датасета с добавленными признаками

Кроме одиночных признаков и их агрегатов, можно использовать другие функции от признаков. Например, логарифм или экспоненту амплитуды или длины, корень квадратный из площади по амплитуде или по отклонению и т.д. Добавление признаков и их нормализация помогло приблизить распределения к нормальному виду (рис. 3), хотя тест на нормальность Шапиро-Уилка они не проходят.

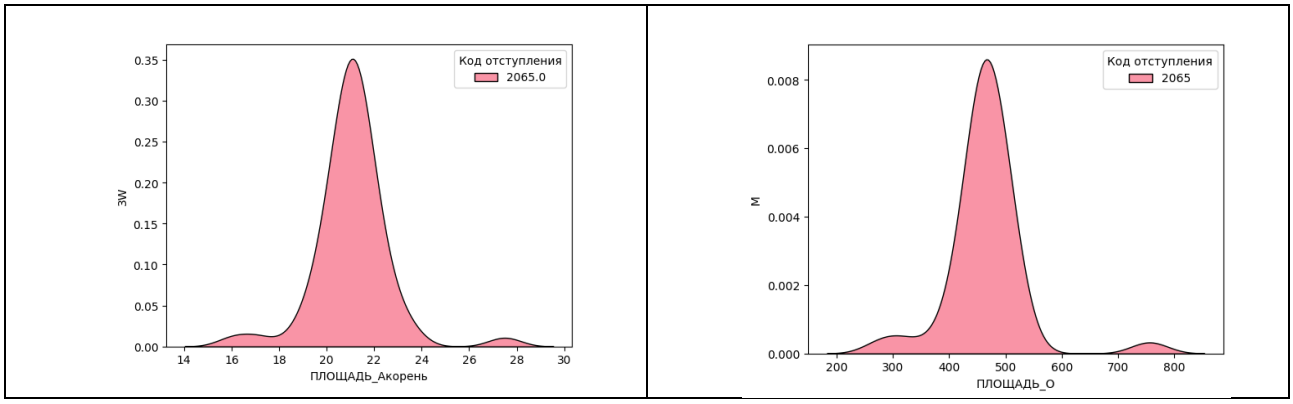


Рис. 3. Распределения добавленных признаков

При подготовке данных для регрессионного моделирования необходимо также закодировать значения признака Время, для чего использована функция OrdinalEncoder (рис. 4):

	ВРЕМЯ	ДИСТАНЦИЯ	АМПЛИТУДА	СТЕПЕНЬ	ДЛИНА
1003	111.0	26.681	25	2	34
768	72.0	26.659	16	2	38
687	174.0	650.165	42	2	34
273	46.0	611.238	12	2	17

Рис. 4. Кодирование времени

1.2. Выбор критерия минимизации и функции потерь

Для того, чтобы научиться предсказывать y(L, t) по X(L, t), используем модель вида $f_k\{X, U\}$, которая принимает на вход матрицу признаков X(L, t) и возвращает предсказание y(L, t). Модель зависит от вектора управляющих параметров U, которые определяют ее поведение. Чтобы подобрать оптимальные значения для U, использован критерий минимизации, оценивающий качество модели на обучающей выборке. Критерий минимизации — это средняя ошибка модели на всех объектах обучающей выборки. Ошибка модели на одном объекте определяется функцией потерь, которая измеряет разницу между истинным значением y_i и предсказанным значением $\hat{y}_i$. Функция потерь в теории статистических решений характеризует потери при неправильном принятии решений на основе наблюдаемых данных. Например, в линейной регрессии часто используется квадратичная функция потерь, которая равна сумме квадратов разностей между фактическими и предсказанными значениями зависимой переменной $(y_i - \hat{y}_i)^2$. Минимизация такой функции потерь приводит к нахождению оптимальных параметров модели и коэффициента детерминации.

Коэффициент детерминации R^2 — статистическая мера, показывающая, как хорошо регрессионная модель предсказывает зависимую переменную. Она равна доле дисперсии зависимой переменной, которая объясняется моделью:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \hat{y}_i)}. \quad (1)$$

R^2 может использоваться как критерий минимума эмпирического риска, который оценивает качество модели на обучающей выборке. Однако этот критерий имеет некоторые недостатки:

- всегда увеличивается при добавлении новых независимых переменных в модель, даже если они не улучшают качество предсказания. Это может привести к переобучению модели и ухудшению её обобщающей способности.
- не учитывает количество независимых переменных в модели и их взаимодействие. Две модели с одинаковым R^2могут иметь разную сложность и интерпретируемость.
- не подходит для сравнения моделей с разными зависимыми переменными или разными функциями потерь.

Поэтому для оценки качества регрессионной модели на обучающей выборке часто используются другие критерии, такие как:

Скорректированный R^2, который учитывает количество независимых переменных в модели и штрафует за их избыточность:

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}. \quad (2)$$

где R^2 — коэффициент детерминации, n — размер выборки, k — число предикторов.

Среднеквадратичная ошибка (MSE), которая измеряет среднее отклонение предсказаний модели от истинных значений зависимой переменной:

$$\text{MSE} = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}, \quad (3)$$

Средняя абсолютная ошибка (MAE), которая измеряет среднее абсолютное отклонение предсказаний модели от истинных значений зависимой переменной:

$$\text{MAE} = \sum_{i=1}^{N}|y_i - \hat{y}_i|, \quad (4)$$

Таким образом, ставится задача на обучающей выборке $\{x_i : i = 1, \ldots, h\}$ найти вектор управляющих параметров U модели $f_k(X, U)$ по критерию минимума эмпирического риска (2). Вектора X и U могут принимать значения из заданных подмножеств евклидового пространства E^n и E^r: $X \subset E^n$, $U \subset E^r$. На вектора X и U могут быть наложены ограничения: $g_i(X, U) \geq 0, i = 1, \ldots, m$.

1.3. Выбор модели по критерию минимума эмпирического риска

В данной работе мы исследовали применение различных регрессионных моделей для прогнозирования размеров отступлений железнодорожного полотна (рис. 5).

Рис. 5. Выбор регрессионной модели

Сначала построили линейную регрессионную модель для прогноза нормализованных размеров каждого отступления. Однако обнаружили, что такая модель имеет низкое качество прогноза, так как коэффициент детерминации был мал и преимущественно отрицателен. Это означает, что модель не может объяснить значительную часть дисперсии зависимой переменной и имеет большую ошибку. Мы предположили, что это связано с тем, что размеры отступлений имеют большую вариабельность и зависят от множества факторов, которые не учитываются в модели.

Чтобы снизить дисперсию, мы построили линейную регрессионную модель для прогноза размеров отступлений на каждом километре пути и получили небольшое улучшение коэффициента детерминации, который стал преимущественно положителен, но все еще низок (рис. 6). Это означает, что модель все еще не может объяснить большую часть дисперсии зависимой переменной и имеет значительную ошибку.

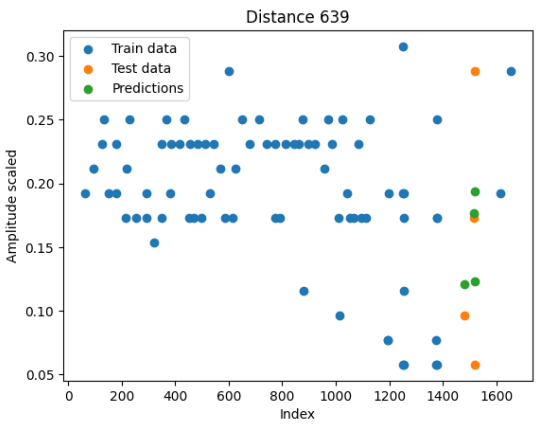
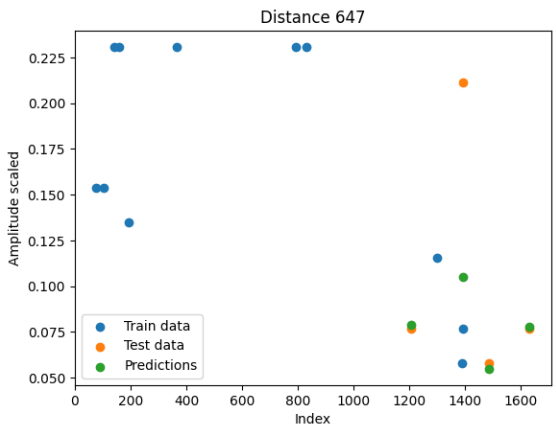
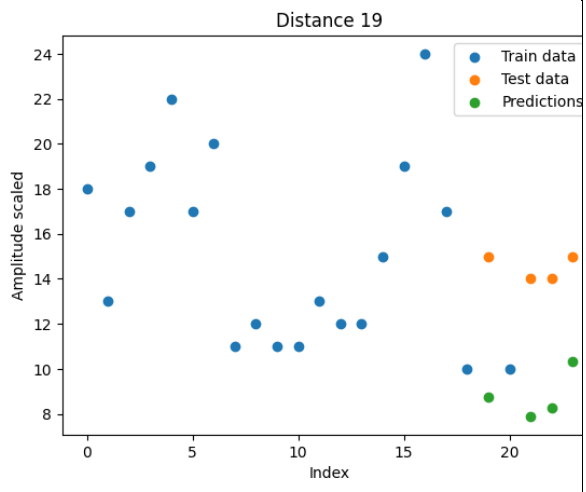
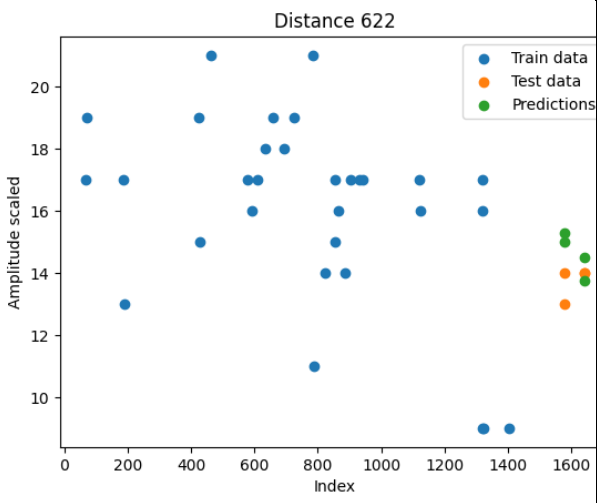
	
R^2 : 0.3492 Скорректированный R^2: 0.1052	R^2 : 0.2516 Скорректированный R^2: 0.0475

Рис. 6. Прогноз размеров отступлений по линейной регрессионной модели на каждом километре пути

Далее мы сменили линейную регрессионную модель на модель линейной регуляризации (рис. 7). Эта модель хороша в случаях, когда независимые признаки сильно коррелируют. Она добавляет штраф к функции потерь за сложность модели, чтобы избежать переобучения и увеличить обобщающую способность. Затем мы сравнили модели линейной регуляризации с разными значениями параметра регуляризации и выбрали те, которые минимизировали среднеквадратичную ошибку прогноза на тестовой выборке. Мы обнаружили, что модель с L1-регуляризацией имеет лучшее качество прогноза, чем модель с L2-регуляризацией, так как она отбирает наиболее значимые входные переменные и устраняет шум. Коэффициент детерминации для модели с L1-регуляризацией был преимущественно положителен и средний по величине, что свидетельствует о том, что модель объясняет значительную часть дисперсии зависимой переменной и имеет меньшую ошибку, чем линейная регрессионная модель.





R^2: 0.6889 Скорректированный R^2: 0.6686 Количество точек: 50 MSE: 2.39 MAE: 1.2529	R^2 score: 0.9036 Скорректированный R^2: 0.8747 Количество точек: 14 MSE: 0.22 MAE: 0.3963

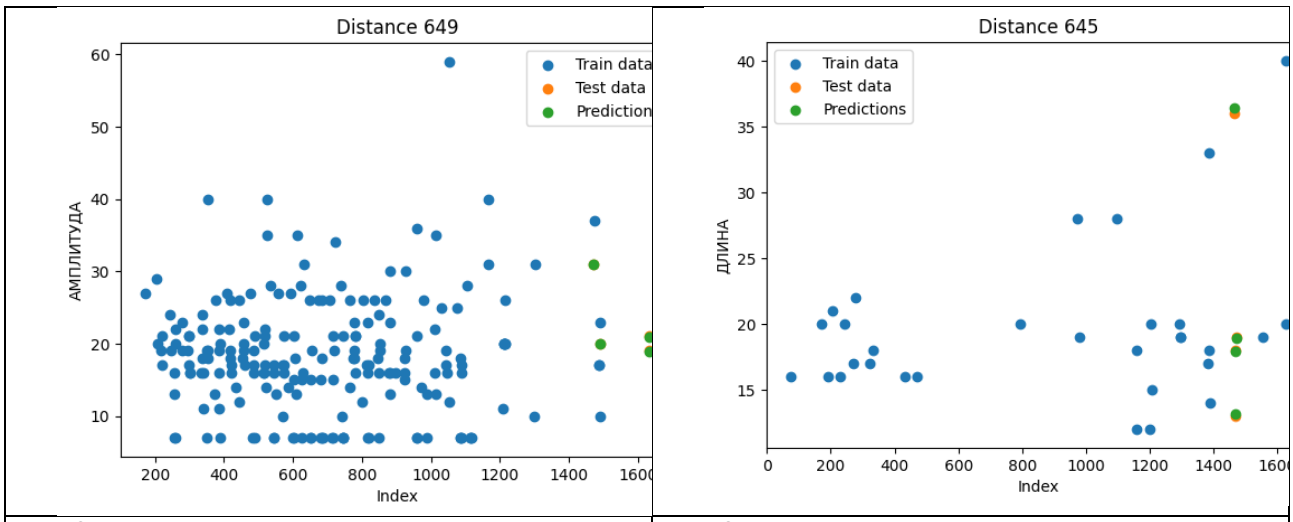
Рис. 7. Результаты прогнозирования амплитуд по модели линейной регуляризации

Из результатов видим, что модель дает более высокое качество прогнозирования для меньшего αколичества точек. Также мы видим, что параметр α модели подобран достаточно хорошо для каждого набора данных, так как значения Скорректированного R^2 близки к значениям R^2.

В дополнение к модели линейной регуляризации применили перекрестную проверку. Этот метод позволяет не только избежать переобучения и повысить обобщающую способность модели, но и определить наиболее значимые признаки, которые влияют на зависимую переменную. Мы использовали метод лассо, который добавляет к функции потерь штраф за абсолютные значения коэффициентов регрессии. Это приводит к тому, что некоторые коэффициенты становятся равными нулю, а соответствующие им признаки исключаются из модели. Для выбора наилучшей модели прогноза по разным признакам использовали кросс-валидацию, которая заключается в разбиении данных на несколько подвыборок и поочередном использовании одной из них в качестве тестовой, а остальных в качестве обучающей (рис. 8).

Прогнозные зеленые точки закрывают (или почти закрывают) тестовые оранжевые, что наглядно говорит о высоком качестве прогноза. На основании оценок качества модели выбран признак, производящий наилучший прогноз и этим признаком является ПЛОЩАДЬ_А.

Наконец, мы применили модель линейной регуляризации для прогноза размеров отступлений, объединяя несколько километров пути. Таким образом, мы увеличили количество отступлений, которые учитывались при расчете коэффициентов модели, чтобы получить более качественный прогноз. Мы получили значительное улучшение качества прогноза, так как коэффициент детерминации для модели с L1-регуляризацией положителен и высок. Это означает, что модель хорошо адаптируется к данным и улавливает общую закономерность.



R^2: 0.9989 MAE: 0.0987 Скорректированный R^2: 0.9988	R^2 : 0.8831 MAE: 0.1089 Скорректированный R^2: 0.8513

Рис. 8. Результаты прогнозирования разных признаков по модели линейной регуляризации

2. Алгоритм прогноза размеров отступлений

Алгоритм прогноза размеров отступлений состоит из трех основных шагов: подготовка данных, прогнозирование, а также визуализация результатов (рис. 9).

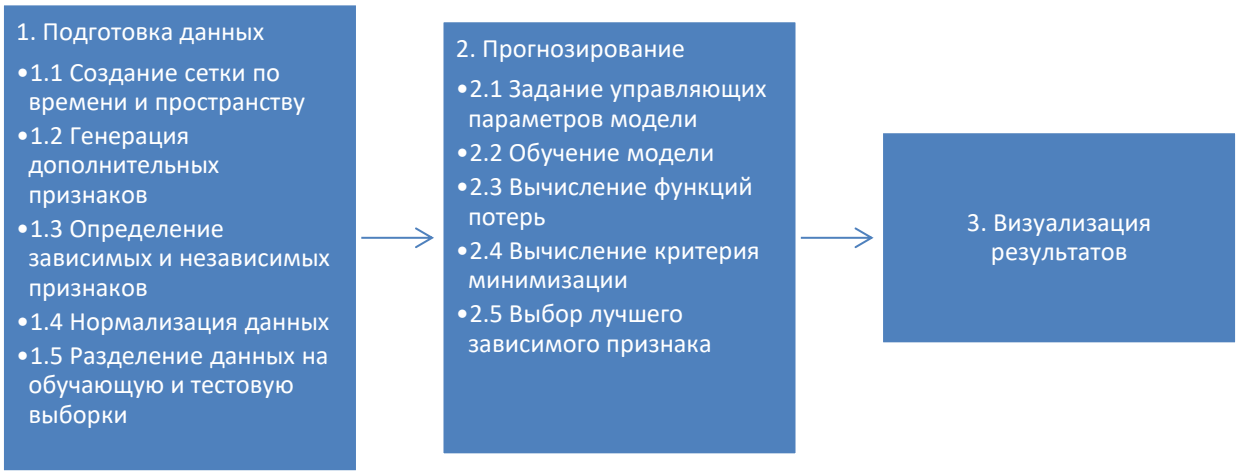


Рис. 9. Алгоритм прогноза размеров отступлений

На этапе подготовки данных мы создаем сетку по времени и пространству, в которой ячейкам сетки соответствует некоторое количество отступлений. Затем генерируем дополнительные признаки, такие как площади отступлений, которые более полно учитывают имеющиеся данные. Каждый признак при построении модели выдвигается на роль зависимого, а оставшиеся становятся независимыми. Мы нормализуем данные, чтобы устранить различия в масштабах и единицах измерения и разделяем их на обучающую и тестовую выборки, чтобы проверить качество модели. На этапе прогнозирования задаем управляющие параметры модели, такие как количество интервалов по пространству, код отступления и обучаем модель на обучающей выборке, используя метод наименьших квадратов для определения коэффициентов регрессионного уравнения. Затем вычисляем функции потерь, такие как коэффициент детерминации и на их основе критерий минимизации в виде средней квадратичной ошибки и средней абсолютной погрешности.

3. Заключение

Предложен и реализован на языке Python метод прогнозирования размеров отступлений на основе модели линейной регрессии с L1-регуляризацией. Показано, что метод позволяет выбрать целевой

признак, прогнозировать размеры отступлений по типам и дать высокое качество прогноза, так как учитывает значительное количество признаков и избегает переобучения.

Литература

1. *Falamarzi A., Moridpour S., Nazem M. A* Review on Existing Sensors and Devices for Inspecting Railway Infrastructure // Jurnal Kejuruteraan. 2019. Vol. 31, № 1. P. 1–10.
2. *Lingamanaik S.N. et al.* Using Instrumented Revenue Vehicles to Inspect Track Integrity and Rolling Stock Performance in a Passenger Network During Peak Times // Procedia Eng. No longer published by Elsevier, 2017. Vol. 188. P. 424–431.
3. *Федоров Д.В., Потапенко В.С.* Способ определения локальных дефектов поверхности катания железнодорожных рельсов: pat. RU2717683C1 USA. 2019.
4. *Kundu P. et al.* A Review on Condition Monitoring Technologies for Railway Rolling Stock // PHM Society European Conference. 2018. Vol. 4, № 1.
5. *Vladova A. Yu.* Creating feature spaces and autoregression models to forecast railway track // Control Sciences. 2023. Vol. 2. P. 54–64.
6. *Vladova A.Yu*. Identification of the Railway Track Technical State // 2021 14th International Conference Management of large-scale system development (MLSD). IEEE, 2021. P. 1–5.
7. *Дубицкий И.С., Енин А.В., Владова А.Ю.* Анализ динамики износа железнодорожных путей // Управление развитием крупномасштабных систем (MLSD’2021). 2021. P. 979–985.