Intellectualizing analysis and assessment of statistical data on field pipeline failures due to internal corrosion

Denis P. Karmachev[1,2,*]

[1]Tomsk Polytechnic University, 30, Lenina ave., Tomsk, 634050, Russia
[2]TomskNIPIneft JSC, 72, Mira ave., Tomsk, 634027, Russia

Abstract. The paper presents some results of a study meant to create an expert system used to select a material design and method of internal corrosion protection for field pipelines at the design stage. The author has performed an intellectual analysis of operating statistical data on field pipeline failures. The first part of the paper describes the initial sample and the exploratory analysis performed. The second and the third parts describe the processes of creating and assessing a classifier based on the Random Forest algorithm. To assess the quality of the classifiers, the author has calculated the shares of correct answers in the algorithm (accuracy), precision and recall, as well as the F1-score. The author makes a conclusion about satisfactory values of quality metrics and outlines areas for further research.

1 Introduction

Oil and gas companies use field pipelines with various design specifications operated under the influence of various internal and external factors during oil and gas production, primary treatment and transportation. These internal and external factors directly influence the emergence and development of internal and external corrosion processes in some sections of field pipelines.

This work is connected with the development of a software prototype for an expert decision-making support system (hereinafter referred to as the "System") to be used to choose a material design and method of internal corrosion protection for field pipelines. The goal of this study is to conduct a sequential intellectual data analysis to create an example of predictive model that can be used as the main component of the System.

2 Data mining of operational statistics

Current work is concerned with statistical data on the failures and operating conditions of various field pipelines. The initial sample comprises 641 fields in the Russian Federation and contains 109,989 examples of field pipeline failures (data lines). These failures occurred in 2000-2019 at pipelines commissioned in 1938-2018. This analysis is done in the Python software environment.

2.1 Exploratory data analysis

Basic information on the causes and nature of failures in the original sample is presented in table 1. Current work is carried out on failures due to internal corrosion. Most of failures were due to internal corrosion. Also the causes are unknown (NaN) for 14356 failures. These failures are also accepted in the current study as we can assume that most of these failures are due to internal corrosion. The training of the predictive model will be performed based on the causes of failures that have a different failure nature.

The result of the predictive model will be the mean time between failures caused by internal corrosion. The feature "Nature of failure" cannot be considered in the model, since it is unknown until the moment of failure.

Table 1. Causes and nature of the considered failures

Column name	Category	Quantity
Cause of failure	Internal corrosion	67777
	External corrosion	22881
	NaN	14356
	Other – factory and construction defects, delamination, mechanical damage	4975
Nature of failure (cracks, holes, breaks, loss of tightness)	Body	80527
	NaN	19398
	Other	6257
	Welded joint	3807

A preliminary assessment used to involve or reject specific features is based on the following rules:
– The feature is specific at the design stage and known before actual failure occurs.
– The feature characterizes the operating reliability of the inner wall of a test target (field pipeline section).
– The feature characterizes the internal operating factors of the test target: hydrodynamic parameters, physical and chemical properties of pumped media.

For comparison, the author presents information on some features not involved in the analysis:

* Corresponding author: karmachevd@gmail.com

– "Failure coordinate," "Manufacturing plant," "Construction contractor", as these features are unknown at the design stage.
– Organizational and economic data, as these features are not connected with operating reliability.

It should be noted that failures caused by factors not connected with field pipeline inner wall reliability are excluded from the initial table. At the same time, the study comprises the following features: "Type of outer coating," "Type of heat insulation coating," "Laying method," "Type of soil," and "Laying depth." These features influence the temperature of pumped media and, consequently, lead to the emergence and development of corrosion processes [1].

The preliminary analysis restores the values of features that can be determined definitely or with some assumptions, for example:
– Water content (in percentage) and zero values of oil flow rate for water and gas pipelines meant to transport dry gas
– Zero values of gas flow rate for pipelines pumping gas-free oil or water.

Mean time to failure [2] for each unique field pipeline section is taken as a target value.

The initial sample comprises more detailed information on the physical and chemical properties of pumped media: Ca^{2+} in water (mg/l), CO_2 in water (mg/l), dissolved O_2 (mg/l), H2S in water and oil (mg/l), Mg^{2+} in water (mg/l), suspended particles (mg/l), total dissolved solids (mg/l), etc. At the same time, involvement of all these features leads to a reduction in the sample size to 2,143 examples of failures, provided data lines with missing values are excluded. Therefore, practically all physical and chemical properties of pumped media are denoted through the "Group of corrosion contour (GCC)" feature. This feature takes values from 1 to 4 and testifies to the overall level of medium aggressiveness.

The importance of features (Fig. 1) is assessed using the Random Forest (RF) algorithm [3]. Category features are coded with a numerical method using the LabelEncoder function [4]. To check the adequacy of this approach, importance is assessed for standardized [4] and initial continuous features. At the same time, both resulting importance diagrams are practically identical (the deviations of the importance of each feature do not exceed 1%).

The diagram (Fig. 1) shows the features that are the initial data for the model. To predict the MTBF using the System, the user must enter the values of these attributes and additionally indicate the year of putting the pipeline section into operation.

The importance diagram shows adequate results:
– Input and output pressure of a field pipeline section (P_in, P_out), as well as the flow rate of gas and oil phases (F_gas, F_oil) are among 5 most significant features.
– Section length is also highly important, as, in general, the probability of a failure in the longest section is higher vs. that in a short section.
– Water content and temperature of pumped media are also highly important, as they directly influence the emergence and development of corrosion processes on a field pipeline inner wall.
– The importance of outer and heat insulation coatings exceeds the importance of inner coating, since over 90% of the pipelines under consideration do not have inner coating, while over 50% of the field pipelines in operation use outer and heat insulation coatings.
– The standardized value of operating pressure (P_op) provides additional support, as one should not be confident in the truth of averaged input and output pressure values.
– The influence of chemization parameters is insignificant due to the fact that about 90% of the field pipeline sections are not exposed to inhibiting.

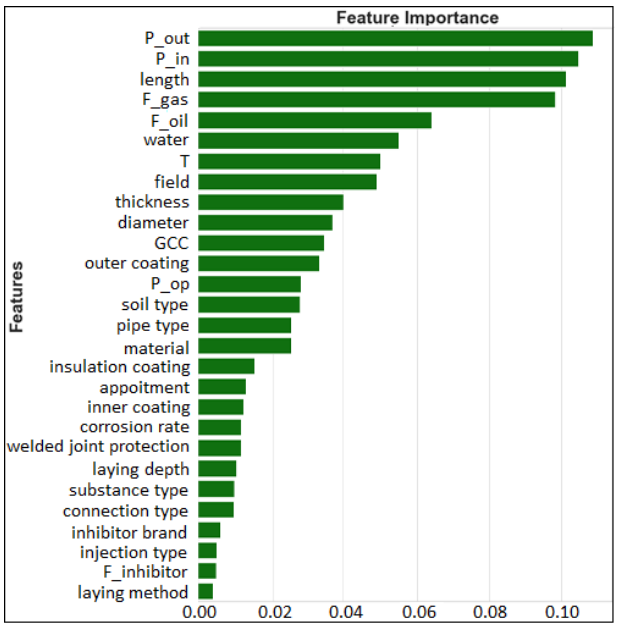

Fig. 1. Feature Importance Diagram.

With the exclusion of any features, the general dynamics of the distribution of importance is unchanged. Therefore, this diagram can also be fully considered to determine the significance of operating factors.

2.2 Model creation and history matching

The prepared sample contains 19623 examples of failures (rows) and 28 features (columns). At this stage, categorical features are coded with a binary method (one-hot encoding) [4], since numerical coding may lead to incorrect interpretation of category features by the model, which will ultimately affect the quality of the classifier performance.

The implementation of the predictive classifier is based on the RF algorithm that is easy to use, has good accuracy and is also widely used in other research studies [5]. The Gini index is used as a dispersion criterion [6].

The best model parameters are searched using the RandomizedSearchCV and GridSearchCV functions [4].

The configuration parameter diagrams are shown in figure 2.

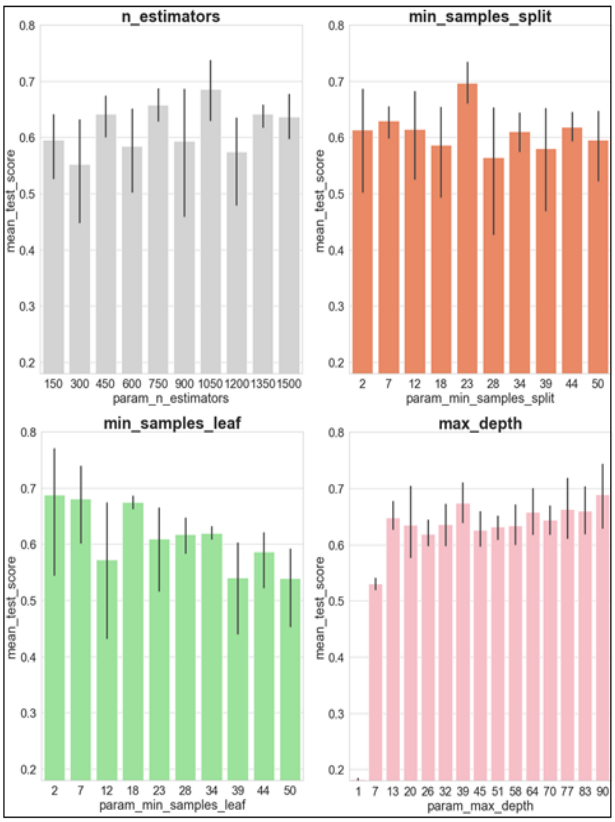

Fig. 2. Diagrams of classifier settings.

2.3 Classifier assessment

For quality assessment, the prepared data frame is divided into training and test sets with a ratio of 80% to 20%, respectively. The total number of examples of failures (19623) allows us to define this strategy.

As metrics for assessing the quality of classifier, the shares of correct answers of the algorithm (accuracy), precision and recall, as well as the F1-score [7] were determined. The values of the metrics are presented in table 1.

Table 1. Quality classifier metrics (interval classes)

Class	Precision	Recall	F1-score
1…4 years	0.88	0.86	0.87
5…7 years	0.86	0.84	0.85
7…8 years	0.84	0.89	0.86
9…10 years	0.84	0.86	0.85
11…12 years	0.86	0.86	0.86
13…14 years	0.89	0.84	0.87
15…16 years	0.88	0.84	0.86
17…19 years	0.81	0.88	0.84
20…22 years	0.85	0.87	0.86
23…25 years	0.85	0.85	0.85
26…28 years	0.91	0.89	0.90
29… years	0.94	0.91	0.92
accuracy			0.87
macro avg.	0.87	0.86	0.87
weighted avg.	0.87	0.87	0.87

The transition from interval classes to detailed classes is more convenient for perception and operation in the System, but the metrics indicators are lowered. Information is provided in table 2.

Table 2. Quality classifier metrics (detailed classes)

Class	Precision	Recall	F1-score
1 year	0.89	0.74	0.81
2 years	0.82	0.86	0.84
3 years	0.87	0.78	0.82
4 years	0.81	0.82	0.81
5 years	0.84	0.83	0.84
6 years	0.81	0.83	0.82
7 years	0.82	0.78	0.80
8 years	0.75	0.87	0.81
9 years	0.82	0.82	0.82
10 years	0.77	0.81	0.79
11 years	0.73	0.80	0.76
12 years	0.88	0.88	0.88
13 years	0.85	0.85	0.85
14 years	0.83	0.79	0.81
15 years	0.81	0.82	0.81
16 years	0.77	0.78	0.77
17 years	0.81	0.78	0.80
18 years	0.78	0.81	0.79
19 years	0.80	0.78	0.79
20 years	0.88	0.87	0.88
21 years	0.72	0.81	0.76
22 years	0.87	0.79	0.83
23 years	0.87	0.84	0.86
24 years	0.86	0.84	0.85
25 years	0.82	0.86	0.84
26 years	0.91	0.86	0.88
27 years	0.86	0.79	0.82
28 years	0.90	0.90	0.90
29 years	0.79	0.82	0.80
30 years	0.91	0.83	0.87
31 years	0.95	0.78	0.85
32 years	0.85	0.79	0.81
accuracy			0.82
macro avg.	0.83	0.82	0.82
weighted avg.	0.83	0.82	0.82

A confusion matrix (Fig. 3) is built for a detailed assessment of model performance quality (vs. interval classes). The confusion matrix shows large (over 10), medium-sized (from 5 to 10) and insignificant (under 5) groups of classification errors. The overwhelming majority (14 out of 16) of the large groups of errors falls at adjacent classes, testifies to the similarity of the features of these failures and, consequently, to the similarity of the features of the field pipeline sections under study. A similar conclusion is relevant to the two other large groups of failures falling at close but not adjacent classes: 9...10 and 13...14; 17...18 and 23...25. It should be noted that combining some adjacent and closest classes will almost completely eliminate the erroneous answers of the classifier.

		PREDICTION, YEARS											
		1-4	5-7	7-8	9-10	11-12	13-14	15-16	17-18	20-22	23-25	26-28	29...
TRUE, YEARS	1-4	226	11	8	5	3	3	2	1	2	0	0	1
	5-7	11	179	10	5	2	5	1	0	1	0	0	0
	7-8	6	7	254	7	8	1	1	2	1	0	0	0
	9-10	5	1	11	301	14	3	3	6	3	2	0	0
	11-12	3	0	9	16	320	12	4	6	1	0	0	0
	13-14	2	3	5	11	7	305	12	5	4	3	2	1
	15-16	2	5	1	3	9	9	285	18	7	2	0	0
	17-18	1	0	3	4	10	0	10	346	11	7	1	2
	20-22	0	2	1	0	0	3	2	21	335	18	2	2
	23-25	0	0	0	2	0	1	3	14	16	275	6	6
	26-28	0	0	1	0	0	0	1	5	6	11	254	8
	29...	0	0	0	3	0	0	1	2	6	6	14	319

Fig. 3. Confusion matrix (interval classes).

3 Summary

The intellectual analysis of statistical data has allowed developing a classifier meant to determine a mean time to failure for designed field pipeline sections based on historical operating data. Its satisfactory results are confirmed with the calculated quality assessment metrics. The confusion matrix indicates the similarity of features in adjacent and close classes. This problem may consist in the fact that over 80% of the data were excluded from the initial sample, as they contained missing values that could not be unambiguously identified and restored. Therefore, further research in this area will primarily be focused on the restoration of missing values. Also, a comparative analysis of other nonlinear classification algorithms will be performed in addition. It should be noted that the good classifier quality metrics show that the current model can already be applied to the System software prototype being developed now. Besides, the current results can be applied in relation to the sections of field pipelines currently in operation to optimize reconstruction and modernization schedules.

References

1. J.P. Gudas, H.P. Hack, Corrosion **35**, 2 (1980)
2. T. Nakagawa, *Advanced Reliability Models and Maintenance Polices* (London: Springer-Verlag London, 2008)
3. Leo Breiman, Machine Learning **45**, 1 (2001)
4. Sebastian Raschka, *Python Machine Learning* (Packt Pablishing, 1st edition, 2015)
5. David Richard Cutler, Thomas S Edwards, Karen H. Beard, Adele Cutler, Kyle T. Hess, Jacob Gibson, Joshua J. Lawler, Ecology **88**, 11 (2007)
6. Giovanni Maria Giorgi, Chiara Gigliarano, Journal of Economic Surveys **31**, 4 (2016)
7. Mohammad Hossin, Sulaiman M.N., International Journal of Data Mining & Knowledge Management Process **5**, 2 (2015)