Machine learning techniques for fault-tolerance analysis and forecasting

Nataliia Kuznietsova[1], Petro Bidyuk[1], Maryna Kuznietsova[2], Jules Lepretre[3]

[1] *National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, 03056, Ukraine*
[2] *Institute for Information Recording of the National Academy of Sciences of Ukraine, Kyiv,03113, Ukraine*
[3] *Ecole Centrale de Lyon, 36 Avenue Guy de Collongue, Écully, 69134, France*

Abstract

This article is dedicated to the research of data storage reliability, fault-tolerance, and survivability. The paper performs modeling on the basis of accumulated historical data on the disks state which provided storage of information for cloud services customers. The requirements for uninterrupted access and high quality (without loss) of stored information for them are important. The main idea of our approach is as follows: if we have the SMART (Self-Monitoring, Analysis and Reporting Technology) characteristics of Hard Disk Drives (HDD) we can predict the time of disk failures and, so, restore the information on the new disks in time. To predict these moments of failure, and the performance of disks, we can use some different models such as ARMA, ARIMA, and linear regression, and forecast the daily failure rate. We took the BackBlaze datasets and showed the possibility to forecast risk exposure levels using SMART characteristics. We used a high-level analysis of risk quantification that allowed us to forecast both the failure rate and the consequent loss. Also, we used the classification models to predict the fact of failure and survival analysis to predict the moment when it could happen.

Keywords

Survivability, fault-tolerance, data mining modelling, risk analysis, neural networks, LSTM, RNN, ARMA, ARIMA, survival models, dynamic approach, loss prediction

1. Introduction

Ensuring the survivability and fault-tolerance of complex technical systems is a quite important task and now it is even more urgent due to the enormous quantity of the IT-systems, programs, and support networks of suppliers and producers. Even, for example, a logistic chain includes not only delivering goods but also such processes as support, connection, databases of goods and tracks, and roads networking. All these processes include a lot of data that should be gathered, processed and further used for solving optimization tasks.

The requirements for the operation of various online web services are associated with a heavy load on the equipment and technical systems that ensure their smooth operation. The simultaneous failures of the largest social networking services that occurred in October-November 2021 in Europe and around the world were explained by the developers as a simultaneous technical load that the technical equipment failed to cope with. Similar challenges arose in Ukraine before the application "Dija", when a large number of users were forced to use the service for Covid-certificates download, thus creating an excessive load. It led to system failure and in some cases even partial data loss. Now the requirements to data especially to their confidentiality, storage reliability and also access due to the requirements of their owners are currently the most relevant. Today, there is a growing digitalization of various areas of

XXI International Scientific and Practical Conference "Information Technologies and Security" (ITS-2021), December 9, 2021, Kyiv, Ukraine
EMAIL: natalia-kpi@ukr.net (N. Kuznietsova); pbidyuke_00@ukr.net (P. Bidyuk); marglekuz@gmail.com (M. Kuznietsova); jules.lepretre@ecl17.ec-lyon.fr (J. Lepretre)
ORCID: 0000-0002-1662-1974 (N. Kuznietsova); 0000-0002-7421-3565 (P. Bidyuk); 0000-0001-6054-418X (M. Kuznietsova); 0000-0001-8681-5195 (J. Lepretre)

activity and the transition to digital data which are regularly collected and stored in huge databases and data warehouses, updated and supplemented daily, and therefore require a lot of space for their storage.

Survivability of such systems means that they can achieve their functioning goals even in the presence of faults and troubles, ensuring required efficiency at required time. So it’s necessary to have all data and information, with all their characteristics of volume, correctness, security, links and means of processing, etc. That’s why a good storage data system is required. The well-known company in this area is BackBlaze, a cloud storage and data back-up company [1]. Its business is based on the data protection and storage of the customers’ information. BackBlaze is involved in cloud storage and data backup, providing extensive cloud storage services, helping customers develop and create their own cloud solutions, backing up existing systems and files, and providing reliable interoperability with workflows and tools. Backblaze has hundreds of pre-built integrations and partners, including Facebook. Given the above-mentioned technical failures in the services of Facebook, the purpose of our study was to test the reliability of data storage on disks from different manufacturers and determine the longest reliable models to recommend Backblaze.

Among other topics such as data privacy or data encryption, the company must provide its ability to minimize the risk of data loss. For this reason, it is needed to assess and predict its Hard Disk Drives (HDD) risk of failure, and to recommend or use some means to restore the information on the disks.

The first disk monitoring technologies of the HDD have been introduced in the 1990’s. They aimed to forecast the HDD failure. These reporting technologies were checking critical performance parameters and providing binary information {Disk OK; Disk will fail soon}. A standardization of the HDD reporting technology was introduced in 1995 and named SMART (Self-Monitoring, Analysis and Reporting Technology). Despite the importance and criticality significance of this subject only very few studies on failure characteristics of disk drives in this area were made and published. Most of the available information comes from the disk manufacturers themselves [2].

In work [3], the authors are working on BackBlaze 2011-2015 HDD reporting. They perform a K-mean clustering down sampling to manually correct the huge class imbalance and are focusing in only 4 different HDD models (2 Seagate’s and 2 Hitachi’s).

In [4] instead of forecasting the failure events, Ni, Jun, and Xiaoning Jin are providing five different decision support tools for planning maintenance operations which aim at considering both production and maintenance decisions jointly. Though providing interesting insights, the approach will be just shortly discussed for the BackBlaze application.

While utilizing and assessing the performance of different machine learning classifiers for predicting failure in industrial equipment for anode production, the work of [5] really aims at emphasizing the importance of the time window considered and its impact on the accurate measurement of a failure.

In [6] the authors considered the risk of a failed HDD over the next month based on the last 3 months’ SMART characteristics. All disks were divided into two types of classes, defined as “0” (if the disk would not fail over the next month) and “1” (if it would fail during the next month). To forecast this fact all disk information was analyzed and the most important variables were defined. Based on the five SMART characteristics it was calculated their last 3 months’ statistics. The obtained results were really accurate.

In our research, we will extend the application of this approach on new obtained data for the next years and will study in time the efficiency of SMART-characteristics time series analysis and how we can use their change for predicting disk failure.

2. Problem statement

Despite already existing and implemented tools for disk monitoring and replacement (mostly based on the average lifespan of an HDD), the preliminary study of the BackBlaze dataset highlights that the problem of disk failure still exists, and thus our research in this area will be relevant. It is still not sufficient to prevent all failures and consequent data losses. Therefore the present work aims at evaluating and forecasting the hard drive faults using the Self-Monitoring, Analysis, and Reporting Technology characteristics of each hard drive.

In this research, we focus on two different and complementary approaches. First, we will predict the level of risk exposure, faced by BackBlaze company over time based on the failure rate and the total storage capacity modeling. In addition to this high-level approach, we will use the survival models for individually forecast the failure probability of each HDD independently. This implementation is considering SMART characteristics as time-varying features. The approach is risk-oriented and therefore should be used as a part of a decision support system for predictive maintenance or early replacement.

3. Mathematical problem definition

The aim of this part is to globally assess and forecast the level of risk exposure, faced by BackBlaze company over time, and to describe the dataset at disposal. To properly assess the BackBlaze's risk exposure we propose to denote the risk of failure at the mathematical composition of the two components: the probability of failure (the failure rate) and the estimated loss which is related to the fact that because of the failure the total amount of available information on the disk is damaged or missed. Equation 1 summarizes the problem definition we will model:

$$R = P \otimes U, \tag{1}$$

where P is the probability of risk occurrence in general; U are related losses. In our task P is the probability of HDD failure in the Datacenter and is dependent in time variable and could be determined also as follows: $P(t) = f(t, P(t-1), P(t-2),..., P(t-n))$. Losses in these cases are the losses of total size storage and are increasing in time, and thus $U(t) = f(t, U(t-1), U(t-2),..., U(t-n))$.

3.1. Preliminary Data Analysis

Normalized SMART characteristics were used for time series analysis: each SMART characteristic will be considered as a time series (like a sensor).

In Figure 1 it is presented the number of reported HDD drive over time. While this number keeps increasing as BackBlaze warehouses are expanding, the figure also underlines that there are 4 days with missing a report for HDD. These missing reports shall not be considered as failed HDD, but rather as reporting error. In the following work, the missing values for these dates were replaced by a linear function between the two nearest neighbors. In Figure 2 it is shown the daily number of HDD reported as failed. This value varies over time and seems close to a stationary signal. In addition, the maximum is 24 failures in one day in 2017. This value will not be considered as an outlier, but rather as a probable important technical incident in the data warehouse.

As the number of HDD varies over time, we will consider the HDD daily failure rate as shown in Figure 3.

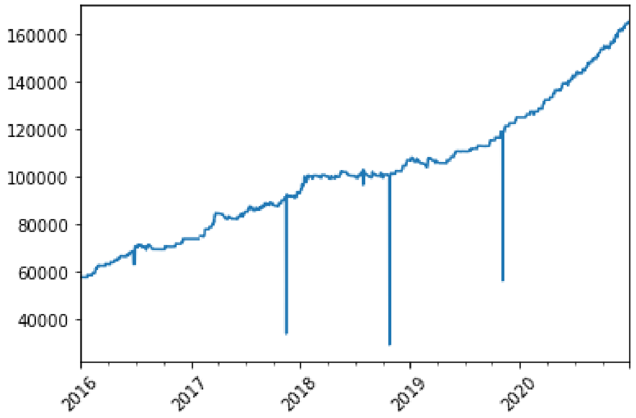

Figure 1: Number of reported HDD over time

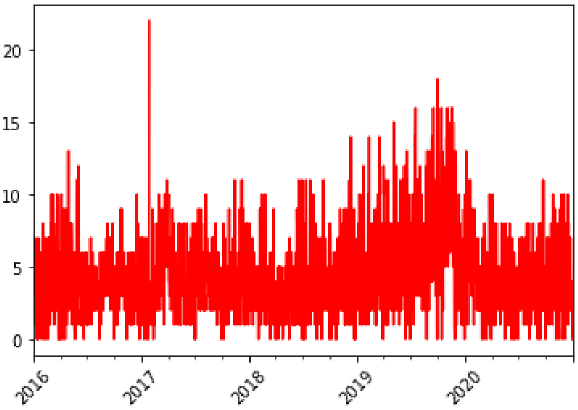

Figure 2: Number of failed HDD

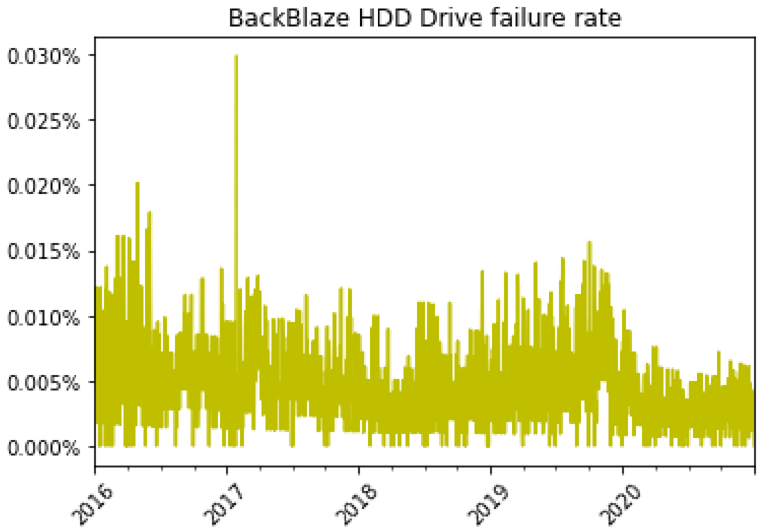

Figure 3: HDD daily failure rate

3.2. First observations

Overall, the daily failure rate is rarely exceeding 0.015%. The graph of the density function, generally, has the shape of a right-skewed Gamma distribution. But in fact, and according to the results provided by the best model, i.e. the model minimizing the chi-square value, it seems to be the beta distribution (Equation 2) with parameters (2.7, 67.4, -6.6e-06, 1.4e-3). Tested distributions and their fitting results are shown in Table 1. The fitted model on the distribution is shown in Figure 4.

$$f(x,a,b)=\frac{\Gamma(a+b)x^{a-1}(1-x)^{b-1}}{\Gamma(a)\Gamma(b)}. \qquad (2)$$

Table 1
Fitting distribution results for the daily failure rate

Distribution	Chi square
Beta	3,70
Gamma	4,08
Pearson3	4,08
Lognorm	25,87
Norm	407,65
Invgauss	473,61
Exponential	742,52
Triangle	1767,91
Weibull_min	3808,81
Uniform	4055,02
Weibull_max	9116,31

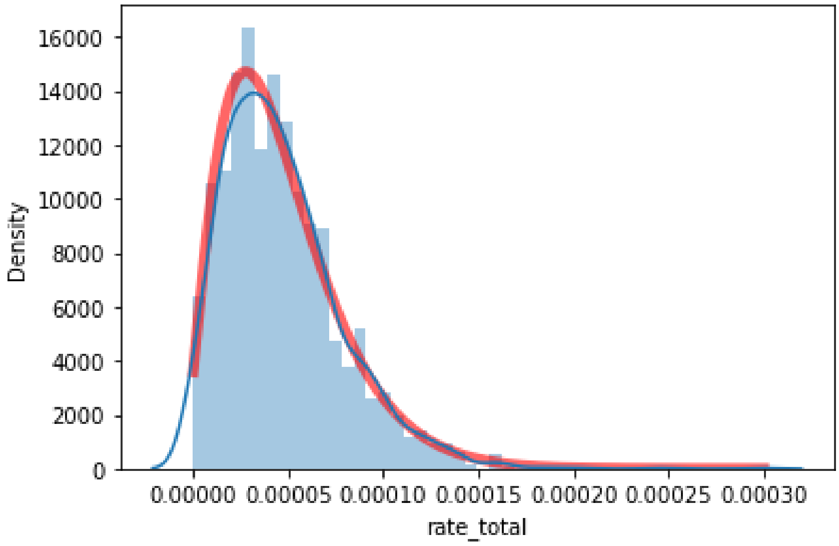

Figure 4: Fitted beta distribution for the failure rate

4. Modelling and forecasting the disks failure using time series analysis

In this part, we will try to model and predict the daily failure rate (signal shown in Figure 3) using different time series models. Based on Figure 3 we can make an assumption that the quantity of HDD failures is non-stationary and failures can have the seasonal effects. That's why we will try to remove these effects and then return to the initial time series.

4.1. ARMA preprocessing

To perform the ARMA modelling, we will implement the following signal transformations:

a. Apply a natural logarithm to stabilize the variance: the Equation 3 will be applied to the signal. Logarithm transformation applied to the signal is presented as such Equation 3:

$$\text{Failure\_Rate}_{\text{log}} = \ln(2 + 100 * \text{HDD}_{\text{rate\%}}) \tag{3}$$

b. Perform a differentiation of the values to remove the eventual trend and seasonality.

The transformed signal is represented in Figure 5. The signal seems now stationary with a stabilized variance, an ARMA model can therefore be considered for modelling the failure rate.

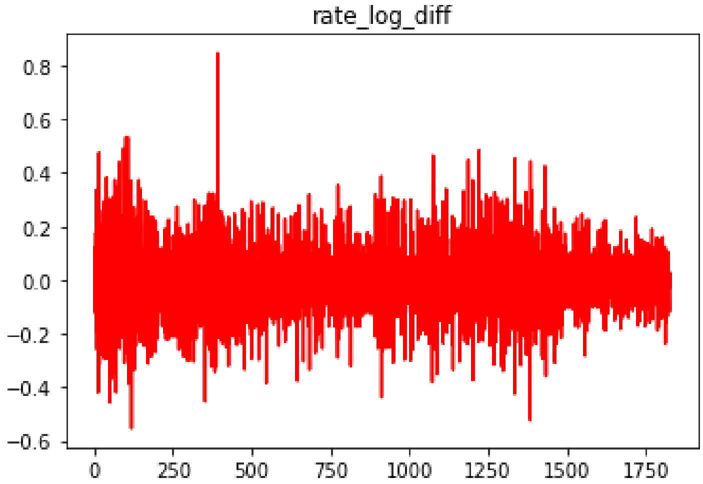

Figure 5: Transformed signal of failure rate

Based on the computed Autocorrelation and Partial Autocorrelation Functions (PACF) we can do some assumptions regarding the order of the autocorrelation or the moving average (with the ACF plot) parts of our model. ARMA (5,2) seems to be a good candidate for sufficiently capturing our signal. In addition, it is worth mentioning that the Dickey-Fuller test for stationary performed rejected the null hypothesis, suggesting the time series does not have a unit root and is stationary.

4.2. ARMA modelling and results

An ARMA model has the form depicted by the Equation 4.

$$X_t = c + \varepsilon_t + \underbrace{\sum_{i=1}^{p} \varphi_i X_{t-i}}_{AR(p)} + \underbrace{\sum_{j=1}^{q} \theta_j \varepsilon_{t-j}}_{MA(q)} . \tag{4}$$

The model implemented is a rolling-window ARMA model, where 85% of the dataset is used to fit the model, and the remaining 15% are used for testing. As ARMA models are performing better in short-term forecasting, we will forecast the next day’s failure rate value based on all past values (rolling window).

In Table 2, the performance of the model on the training set is provided with the AIC (Akaike Information Criterion), the BIC (Bayesian Information Criterion), ranked from lowest AIC to highest [7] gives the best model for the testing set and provides the R^2 score as well as the MAPE (Mean Average Percentage Error).

Table 2
Best fitted ARMA models and their performance values

ARMA(p,q)	AIC	BIC	MAPE
(3,3)	-2220,79	-2183,36	6.801
(1,2)	-2220,49	-2199,1	6.780
(2,1)	-2220,46	-2199,07	7.866
(1,1)	-2220,29	-2204,25	6.843
(0,3)	-2219,87	-2198,48	6.774
(0,2)	-2219,27	-2203,23	6.796
(2,3)	-2219,23	-2187,14	8.418
(1,3)	-2218,75	-2192,02	6.797
(1,4)	-2216,5	-2184,41	6.844

As shown in the Table 2 the most accurate model seems to be the ARMA (3,3). In the ARMA modelling it is also important to check the residuals and ensure they are close to a white noise.

5. Forecast of failure rate using linear regression

It is also possible to convert our time series analysis problem to a linear regression. Indeed, a matrix containing the n lagged values of the failure rate is built. The results considering different number of failure rate's lagged values are shown in Table 3. It appears that the consideration of the 16 past values to forecast next day's failure rate holds the best results. The feature importance of the past values is shown in Figure 6.

Table 3
Results of Linear Regression for failure rate modelling based on lagged values

# of lagged values	MAE	MAPE
10	0.0694375	8.84135
11	0.0689825	8.77823
12	0.0685862	8.72444
13	0.0679679	8.63982
14	0.0662564	8.41403
15	0.064112	8.11212
16	0.0632862	7.99901
17	0.0633625	8.00989
18	0.0636652	8.05639
19	0.063647	8.0538

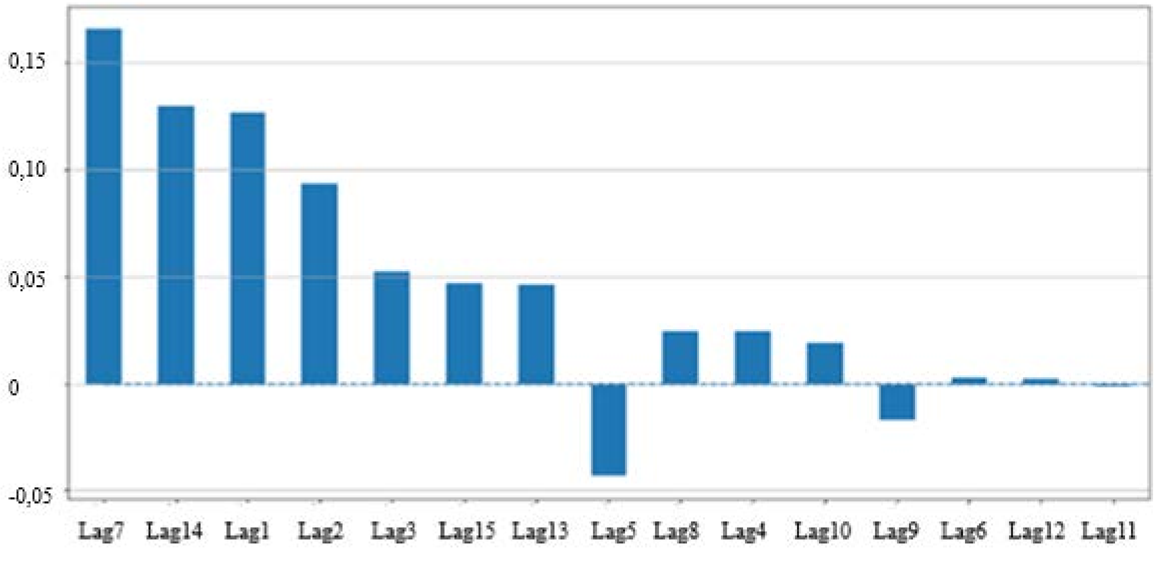

Figure 6: Feature Importance of the failure rate linear regression

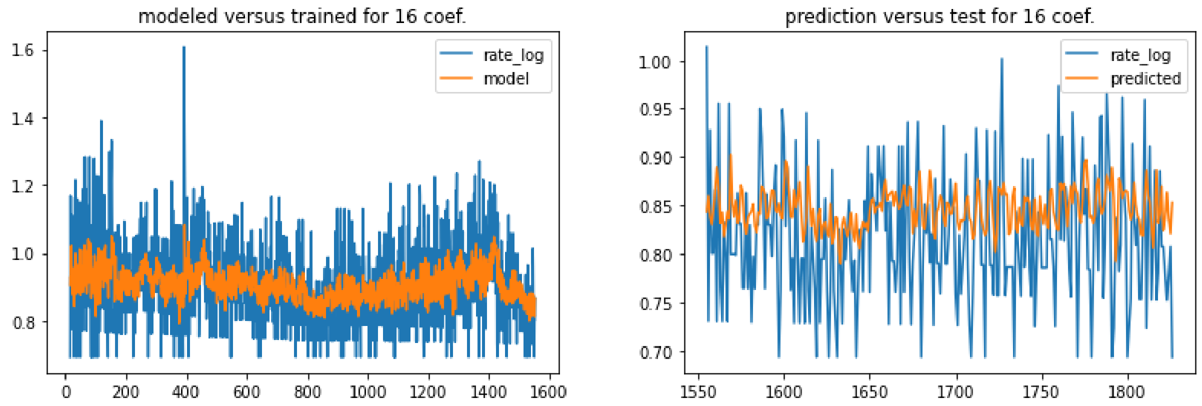

Figure 7: Linear Regression results with 16 lagged values

Figure 7 provides both the model and the predictions of our Linear Regression model. However, the ARMA (3,3) model still outperforms the Linear Regression with a MAPE of 6.8% versus 8%.

6. Forecast of total storage size

As described before, the probability of failure is varying over time. To meet customer's demand the total storage amount of BackBlaze datacentre is also increasing over time (Figure 8).

While different modelling techniques (including exponential and polynomial) have been tested, the most accurate model for predicting the total storage size is the ARIMA. The best model results are summarized in Table 4.

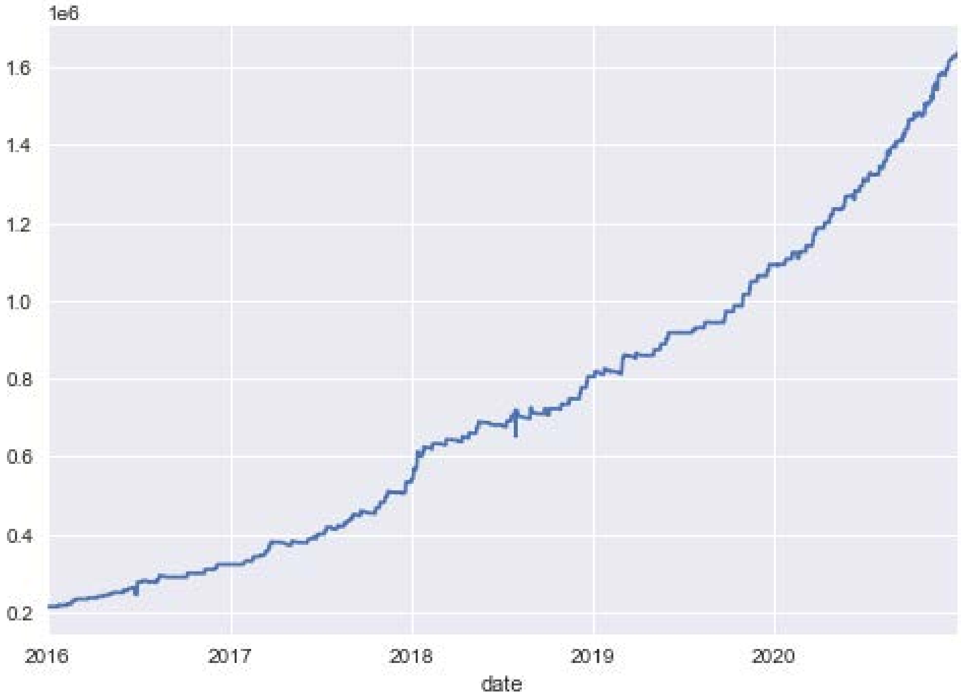

Figure 8: Total Storage size of the datacenter in terabytes

Table 4:
Modelling results for total storage size

ARIMA (p,d,q)	AIC	BIC	R2	MAPE
(1,1,1)	30227	30243	0.998	0.197
(1,1,0)	30227	30238	0.998	0.196
(0,1,2)	30228	30244	0.998	0.197

The ARIMA (1,1,1) provides a really accurate model for storage prediction with a R^2 score of more than 99%. The model and forecasted time series are provided in Figure 9.

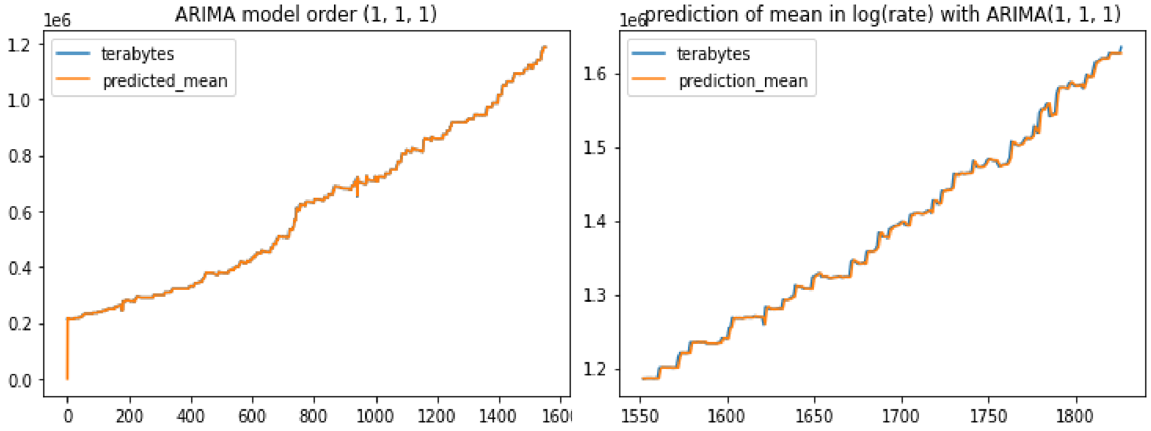

Figure 9: ARIMA model and prediction for total storage size

Thanks to the modelling of both the failure rate and the total storage capacity, BackBlaze is now able to forecast its risk exposure level. This risk level quantification enables it to properly set-up the back-up solutions (quantity of HDD that will probably need to be replaced) and to quantify the required maintenance/replacement actions. So, the disks are removed before they fail, and customers even don't know about it.

7. Classification analysis – next-day failure forecasting based on the SMART characteristics and machine-learning classification algorithms

In our research [6] we used different data mining methods for forecasting the failure of the disk. The best model on received data was the random forest. It seems for us that we need also to develop approach for finding appropriate compromise between monitoring and preventing disk failure in time. The most used statistical indicators for assessing the classification quality are general accuracy, error matrix (Confusion Matrix), first and second kind of errors, index GINI [7]. Usually, the error matrix is determined for binary classification, size 2×2. Let's give some definitions.

FN (False Negative) – first kind of errors, false negative value. For our task such error means that the disk was evaluated as damaged in next day but actually it worked.

FP (False Positive) – second kind of errors, false positive value. Here, this error means that the disk was considered as a good one but actually it failed [8].

The values of the confusion matrix are used to calculate the main metrics for estimating the classification model.

Precision shows that some of the objects which were called positive by the classifier and in fact they were indeed positive:

$$precision = \frac{TP}{TP+FP}. \quad (5)$$

Recall shows which part of the positive class objects out of all the positive class objects were found by the algorithm:

$$recall = \frac{TP}{TP+FN}. \quad (6)$$

The recall characteristic demonstrates the ability of the algorithm to find this class in general, and precision – the ability to distinguish this class from other classes [9].

F-measure (f1-score) is the average harmonic between precision and recall:

$$f_\beta = (1+\beta^2) \cdot \frac{presicion \cdot recall}{(\beta^2 \cdot precision) + recall}, \quad (7)$$

where β determines the accuracy measure in this metric, and for $\beta = 1$ it is the harmonic mean (with the factor $(1+\beta^2)$ equal to 2, so that in the case when $precision = 1$ and $recall = 1$), the f1-score reaches its maximum at completeness and accuracy equal to one. The measure of f1-score approaches zero if one of the indicators approaches zero [9].

For our task we propose to build the confusion matrix for evaluating the faults' risks as the tool for evaluating the costs of losses in a case when the disk was not changed timely and the information was lost, and if the disk was detected as required to change and in fact, it could still operate. In Table 5 the results of the classification models tested in our investigation and confusion matrixes are presented.

At the next step let us evaluate the value of wrong management decisions. Here we assume that the cost of disks is defined as some value D, but the cost of information in case of loss is guaranteed by the sum of the contract between BackBlaze and the client of storage. Let's denote it as L. Usually, the penalty in case of a fault is more than the cost of the disk. We can make also an assumption that $L = 10 \times D$. Now let us evaluate for each case the cost of the wrong decision not in time to change the disks. There are four possible cases based on the different data mining methods for the classification of the disks and prediction of their failure:

$Losses_1 = 329 \times L + 82 \times D = 329 \times 10 \times D + 82 \times D = 3372 \times D$ – obtained for the random forest classifier with deep = 1;

$Losses_2 = 257 \times L + 71 \times D = 257 \times 10 \times D + 71 \times D = 2641 \times D$ – obtained for the random forest classifier with deep=10;

$Losses_3 = 359 \times L + 79 \times D = 359 \times 10 \times D + 79 \times D = 3669 \times D$ – obtained for the logistic regression;

$Losses_4 = 307 \times L + 82 \times D = 307 \times 10 \times D + 82 \times D = 3152 \times D$ – obtained for the support vector machine.

Table 5.
Confusion matrixes and quality for different methods

Fact (0/1)	Forecast (0)	Forecast (1)	Precision	Recall	F1-score
		Random Forest (deep = 1)			
0	2150	82	0.87	0.96	0.91
1	329	473	0.85	0.59	0.7
		Random Forest (deep = 10)			
0	2161	71	0.89	0.97	0.93
1	257	545	0.88	0.68	0.77
		Logistic Regression			
0	2153	79	0.86	0.96	0.91
1	359	473	0.85	0.55	0.67
		Support Vector Machine			
0	2150	82	0.88	0.96	0.92
1	307	495	0.86	0.62	0.72

As we can see from the presented results, the lowest cost required for this time period is for the model of Random Forest for deep = 10 and thus this method is more effective from the costs point of view. Certainly, we can see that the most effective management decisions are really dependable on the costs of the disks and penalty for the company BackBlaze in the case of information losses. Of course, the reputation for the company is more expensive and thus the real time monitoring and prevention of risks is really significant. That's why our next proposal was to develop such dynamic approach based on the time changing of the disks statuses and their statistics which gives us the opportunity to catch the increasing risks of the failure.

8. The task of predicting the time of failure risk

8.1. Approach development

For each hard disk we have five main characteristics time series – measurements of SMART-characteristics (as well as in the previous point, the zero instances were rejected identically and the data were logarithmically transformed). Different approaches can be used to predict the time of risk. If it could be formed a sample similar to the first part, and built a regression (i.e., the vector of features is formed according to the data for three months, and as a label – the time that the disk worked after this point). If we consider the tools of survival analysis, then, since we have time-varying covariates, it is advisable to use the Cox Proportional Hazards Model with Time-Dependent Covariates [7, 10]. To do this, the data for each disk was divided into intervals of 1 day, and the table will have columns: ID – disk number; tStart and tStop – numbers of the day of measurement and the next day; Censoring – whether the disk broke that day: Smart_5, Smart_187, Smart_188, Smart_197, Smart_198.

Using the coxphfit function (from the survival library of the R language, also similar functions are available for Matlab and SAS) from the given interval measurements it is possible to obtain coefficients for the proportional Cox model. However, for its use in calculating the probability of failure after the certain number of days, the predicted values of the covariant are required (and the task of predicting SMART-characteristic itself implies additional error), so this model is generally not used for forecasting, but still allows to obtain survival function.

In order to move to a time-independent covariant and use Cox's Proportional Hazard model let's assume that the basic statistics sufficiently describe the state of the hard disk, and do not depend much from the length of the observed series. To form a training sample for each hard drive, a feature vector was compiled: maximum, minimum, average, and nonzero values of each of the five selected SMART characteristics over the entire drive observation period. The label was the lifetime of the disc and the censorship indicator. The following coefficient was obtained for the covariance: Concordance index = 0.87. To predict the time to failure, we form a vector of features from all available observations for a given disk, using the built model we can determine the length of the disk lifetime for a given probability threshold, and subtract the time that the disk has already worked.

8.2. Example of calculation

For example, for the disk 'Z3029FSB' we have the vector formed from data for 100 days:
[2.1972245773362196, 0.0, 0.06462425227459469, 0.029411764705882353, 2.1972245773362196, 0.0, 0.06462425227459469, 0.029411764705882353, 0.0, 0.0, 0.0, 0.0, 1.0986122886681098, 0.0, 0.8778107128746405, 0.8294117647058824, 1.6094379124341003, 0.0, 0.18858392848979214, 0.14411764705882352].

In figure 10 obtained probability of surviving, i.e. the probability that the disk will continue working without failing in time, is presented. We build the survival model for one year ahead. Next, the company defines the probability trigger which is the less acceptable probability that the disk will work next months. So from figure 10 we can define that for example, with a probability of 0.95, the disk will work for up to 210 days. So, confirmed by surviving models, the estimated operating time is 110 days. In fact, the disk had been working for another 240 days.

It means that if the company determines the triggers as the cut-off for the probability of disk survival, it means the moment when the disk should be removed. In our previous section, we showed that spending money on disks 30 days ahead could lead to extra losses. But on the practical and reputational side it is better to change the disk later and for this reason, the cut-off should be decreased to 0.94. So, survival models are good instruments to define the correct disk cut-off and are more flexible for risk triggers varying.

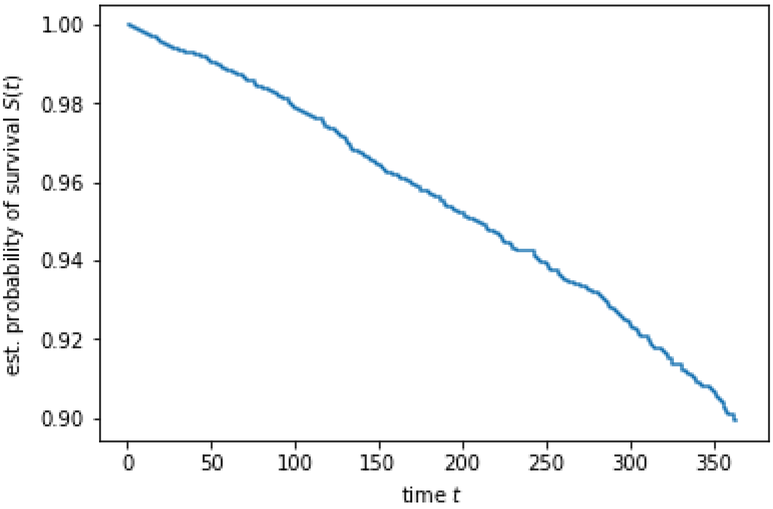

Figure 10: Survival Cox model

9. Conclusions

The research performed in the area of fault-tolerance and survivability analysis of hard disk drives confirms that the problem of fault-tolerance in the technical sphere is still relevant [11]. The methods of data mining applied for solving the forecasting problem of the fault probability give quite a high accuracy but still, there are some disks that are not forecasted as suspicious and required change. Actually, the task of searching for the compromise between changing the disks as a prevention measure is required, and defining a more precise forecasting of fault is required. That’s why the approach proposed in this paper and based on the survival models, and time SMART-characteristics monitoring is a good tool for solving this dilemma. Survival modelling gives the possibility to forecast for each disk its fault tolerance [12] and based on the experience of the technical system and production needs to make the triggers, which would allow forecasting the critical moment for disks and prevent the failure. In our next study, we are going to apply this approach in combination with other data mining methods and evaluate the accuracy of forecasting. Time-monitoring and dynamic risk assessment confirmed their effectiveness in analyzing financial systems [7], so we expect that a combination of integrated and dynamic approaches will be effective also for technical systems.

References

[1] Backblase homepage. URL: https://www.backblaze.com/b2/hard-drive-test-data.html.

[2] D. Anderson, J. Dykes, E. Riedel, More than an interface - scsi vs. ata, in Proceedings of the 2nd USENIX Conference on File and Storage Technologies (FAST’03), 2003, pp. 245 – 257.

[3] M. M. Botezatu, I. Giurgiu, J. Bogojeska, D. Wiesmann, Predicting Disk Replacement towards Reliable Data Centers, in: Proceedings of the 22nd ACM SIGKDD International Conference on

Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 2016, pp. 39–48, doi: 10.1145/2939672.2939699

[4] N. Jun, X. Jin, Decision Support Systems for Effective Maintenance Operations, CIRP Annals, vol. 61, no 1, 2012, pp. 411‑14. DOI.org, doi:10.1016/j.cirp.2012.03.065.

[5] N. Kolokas, T. Vafeiadis, D. Ioannidis and D. Tzovaras, Forecasting faults of industrial equipment using machine learning classifiers, 2018 Innovations in Intelligent Systems and Applications (INISTA), 2018, pp. 1-6, doi: 10.1109/INISTA.2018.8466309.

[6] N. Kuznietsova, M. Kuznietsova, Data mining methods application for increasing the data storage systems fault-tolerance, IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC), 2020, pp. 315–318. https://doi.org/10.1109/SAIC51296.2020.9239222

[7] N. V. Kuznietsova, P. I. Bidyuk, Theory and practice of financial risk analysis: systemic approach, Lira-K, Kyiv, 2020.

[8] F. Herrera, F. Charte, A.J. Rivera, M.J. del Jesus, Multilabel Classification Problem Analysis, Metrics and Techniques, Springer International Publishing, Switzerland, 2016. doi: 10.1007/978-3-319-41111-8.

[9] M. Hossin, M.N, Sulaiman, A Review on Evaluation Metrics for Data Classification Evaluations, International Journal of Data Mining & Knowledge Management Process 5 (2015) 1-11. doi: 10.5121/ijdkp.2015.5201.

[10] D. R. Cox, Regression Models and Life-Tables. Journal of the Royal Statistical Society, Series B, 1972, Vol. 34, No2. P. 187–220.

[11] A. Dodonov, O. Gorbachyk, M. Kuznietsova, Increasing the survivability of automated systems of organizational management as a way to security of critical infrastructures, CEUR Workshop Proceedings, 2018, Vol. 2318, pp. 261–270, http://ceur-ws.org/Vol-2318/.

[12] P. Zheng, S. Yuan, X. Wu, SAFE: A Neural Survival Analysis Model for Fraud Early Detection, 2018. URL: http://arxiv.org/abs/1812.07142.