The American Journal of Engineering and Technology

ISSN 2689-0984 | Open Access

TYPE Original Research
PAGE NO. 202-210
DOI 10.37547/tajet/Volume07Issue05-20

Check for updates

Analysis and Reduction of Errors in AI Models

Rinat Sharipov
Founder of UpLook AI

OPEN ACCESS

SUBMITED 24 March 2025
ACCEPTED 26 April 2025
PUBLISHED 31 May 2025
VOLUME Vol.07 Issue 05 2025

CITATION
Rinat Sharipov. (2025). Analysis and Reduction of Errors in AI Models. The American Journal of Engineering and Technology, 7(05), 202–210. https://doi.org/10.37547/tajet/Volume07Issue05-20

Abstract: The issue of errors in artificial intelligence (AI) models is a critical aspect that requires systematic analysis and the application of effective methods for their reduction. Errors in AI models can occur at various stages of development and deployment, including data collection, model training, and operation phases. The key tasks in this field involve identifying error sources and applying approaches aimed at eliminating them. Methods such as cross-validation, regularization, and the use of ensemble models play a significant role in reducing errors and improving prediction accuracy. Therefore, for the successful use of AI technologies in various domains, continuous attention to model monitoring, parameter adjustment, and the implementation of innovative methods to minimize risks is necessary.

Keywords: artificial intelligence, AI model errors, cross-validation, regularization, error reduction, data analysis.

Introduction: Artificial intelligence (AI) models have become an integral part of modern technologies and are widely used in various industries, including healthcare, finance, transportation, and education. However, despite significant progress in the development and application of these technologies, AI models still face the challenge of errors, which can significantly impact their effectiveness and reliability. Errors can occur both during the data collection phase and throughout the model training process, ultimately leading to a reduction in prediction quality and the generalization ability of models. The relevance of this issue is heightened by the widespread adoption of AI in critical areas where prediction accuracy and reliability are key factors in decision-making.

There are many sources of errors in AI models, including incorrectly represented data, overfitting, unbalanced datasets, and insufficient model interpretability. Errors not only reduce prediction accuracy but can also lead to user mistrust in AI technologies. Therefore, the analysis and reduction of errors are crucial for improving model performance and ensuring the safe use of AI. Recently, active research has been conducted to develop methods aimed at identifying and addressing various types of errors in AI models.

The goal of this work is to analyze the sources of errors in AI models and explore existing methods for their reduction.

1. CLASSIFICATION OF ERRORS IN AI MODELS

Artificial intelligence represents a set of algorithms and systems designed to train computers to perform complex tasks in data processing and analysis, mimicking human perception processes. Through these models, machines can extract information from data, identify patterns, and make decisions with minimal human involvement.

Automation and digitalization are made possible through the development of artificial intelligence. The essence of AI is to enable computers and other devices to act based on principles similar to human thinking. By programming machines to imitate human cognitive processes, significant improvements in productivity and accuracy can be achieved in many tasks. While this may seem intimidating to some, as it evokes images of robots from science fiction, AI helps optimize workflows and enhance their efficiency.

There are many different approaches to creating artificial intelligence models. AI models are a combination of methods and algorithms designed to teach machines to work with data in a way similar to how humans do. Machine learning is an important branch of AI in which computers can independently create new algorithms by analyzing large volumes of data. Some AI models require programming of specific algorithms, after which they can adjust their actions based on the experience gained. There are also models that lack the ability to learn independently and operate solely on pre-programmed algorithms, requiring regular human intervention.

An example of an AI model application is Google Maps and other navigation systems, which help users plot routes. These systems use AI algorithms to process data, such as information on traffic and road changes, improving the accuracy of predictions and route suggestions. Each new user experience is integrated into the system, making it more reliable and precise.

However, the question remains: does artificial intelligence contribute to the improvement of society, or does it pose a threat by rendering human labor unnecessary? There are differing viewpoints on this issue.

Stephen Hawking, the renowned theoretical physicist, expressed concern that the creation of fully autonomous AI could lead to the extinction of humanity, as such systems would evolve at an increasing pace, leaving humans behind due to the slow process of biological evolution. In contrast, Ginni Rometty, the CEO of IBM, believes that AI will not replace humans but rather help enhance our intellectual capabilities, suggesting that AI should be viewed not as a replacement but as a means of augmenting human intelligence [1].

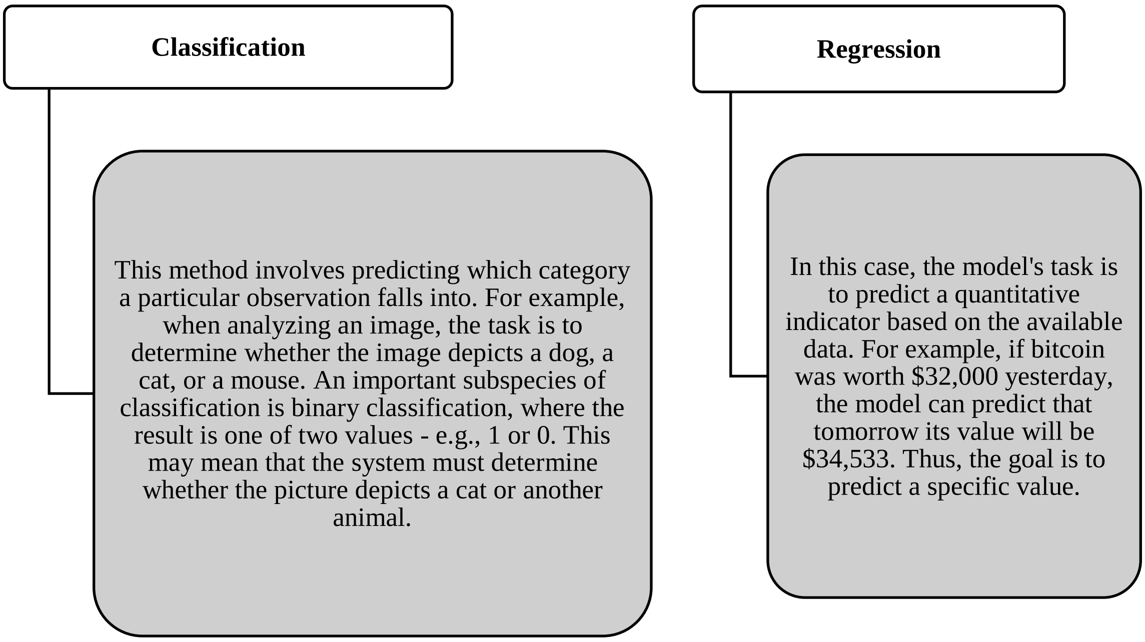
Artificial intelligence (AI) models play a crucial role across a wide range of fields, enhancing workflows, automating tasks, and addressing complex challenges. Below are several examples of key AI applications (Table 1).

Table 1. Examples of AI capabilities [2]

AI application	Description
Image analysis	AI is used for image recognition and classification, which is particularly useful in areas such as e-commerce, security, and medical diagnostics.
Natural language	AI can analyze and interpret human speech, widely applied in text analysis systems, virtual

processing (NLP)	assistants, and chatbots.
Recommendation systems	AI is actively used to develop recommendations for products, services, and content based on users' preferences and actions.
Generative models	AI can create original solutions in situations with no clear answers. These technologies are in demand for creative processes, marketing material development, and even coding.
Personalization of user experience	AI enables the creation of personalized offers for users, whether it be individualized learning plans or customized marketing messages.
Prediction	AI is used to analyze data and predict future events and trends, with applications in finance, healthcare, and marketing analytics.
Fraud detection	AI technologies can detect fraudulent activities such as data theft, insurance fraud, and credit card fraud.
Autonomous transportation systems	AI is being developed to create autonomous vehicles, which can be integrated into logistics and transportation systems.
Chatbots for user interaction	AI enables the creation of intelligent chatbots that can not only respond to standard queries but also provide customer support and facilitate online commerce.
Optimization of production processes	AI is applied in industries to automate and improve production cycles, particularly in sectors such as automotive manufacturing, electronics, and metalworking [2].

It is also important to note that the metrics for evaluating models depend on the type of task to be solved and are divided into two main categories (Fig. 1).



Classification

This method involves predicting which category a particular observation falls into. For example, when analyzing an image, the task is to determine whether the image depicts a dog, a cat, or a mouse. An important subspecies of classification is binary classification, where the result is one of two values - e.g., 1 or 0. This may mean that the system must determine whether the picture depicts a cat or another animal.

Regression

In this case, the model's task is to predict a quantitative indicator based on the available data. For example, if bitcoin was worth $32,000 yesterday, the model can predict that tomorrow its value will be $34,533. Thus, the goal is to predict a specific value.

Fig. 1. Metrics used in evaluating models depending on the tasks [3].

For each of these tasks, various metrics are used to assess the quality of the models. In this context, special attention will be given to classification.

One of the simplest and most popular metrics, often mentioned by non-experts, is accuracy. To calculate it, a formula based on the confusion matrix results is used:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} I[y_i = \hat{y}_i] = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where:

- TP: true positive (correctly classified positive cases),

- TN: true negative (correctly classified negative cases),

- FP: false positive (incorrectly classified positive cases, type I error),

- FN: false negative (incorrectly classified negative cases, type II error).

However, accuracy is not always a reliable measure of model performance, especially when classes are imbalanced. For instance, let’s assume we have 100 images of cats and only 10 images of dogs. For convenience, we’ll consider cats as class 0 and dogs as class 1. Since the number of cats is ten times higher than the number of dogs, the dataset is considered imbalanced.

Let’s assume the model correctly classified 90 out of 100 cats, meaning True Negative is 90, and False Negative is 10. Additionally, the model correctly identified 5 out of 10 dogs, so True Positive is 5, and False Positive is 5. Substituting these values into the formula, we get an accuracy of 86.4%. However, if the model simply "said" that all the images were of cats, the accuracy would be 90%, even though this result would be achieved without the model's actual involvement. This demonstrates that high accuracy does not always guarantee good model performance [3].

Therefore, the Precision metric, which characterizes the proportion of correctly predicted positive classes among all samples that the model predicted as positive, becomes an interesting alternative:

$$Precision = TPR = \frac{TP}{TP+FP} \quad (2)$$

Where:

- TP: true positive,

- FP: false positive (type I error),

- FN: false negative (type II error).

Given the wide range of available methods, selecting the optimal algorithm for solving a specific task can be challenging. It is necessary to compare different models and choose the one that best meets the task's requirements. It is important to note that a model's accuracy is not always the primary criterion for its selection, so other factors must also be considered, which will be discussed in further studies.

The application of the sklearn library provides the ability to evaluate the performance of machine learning models, helping to choose the algorithm with the best metrics for forecasting. One of the evaluation methods is calculating errors, such as mean absolute error (MAE) and mean squared error (MSE). These metrics allow for the minimization of deviations in model predictions, thereby improving its efficiency.

The man absolute error (MAE) is the average of all absolut errors, showing how much the model's predictions deviate from actual data.

The mean squared error (MSE) is the average of the squared errors, accounting for both the magnitude and frequency of deviations.

For example, let's consider a classification task using the Titanic dataset. In this case, we will build a model based on two algorithms – logistic regression and k-nearest neighbors (KNN). Other approaches could also be used for data analysis.

```
# Importing libraries
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn import metrics

# Loading data
data = pd.read_csv("gfg_data")

# Selecting features and target variable
x = data[['Pclass', 'Sex', 'Age', 'Parch', 'Embarked', 'Fare', 'Has_Cabin', 'FamilySize', 'title', 'IsAlone']]
y = data[['Survived']]

# Splitting data into training and testing sets
X_train, X_test, Y_train, Y_test = train_test_split(x, y, test_size=0.3, random_state=None)

# Logistic regression
lr = LogisticRegression()
lr.fit(X_train, Y_train)
Y_pred_lr = lr.predict(X_test)

# Logistic regression model evaluation
logreg_score = round(lr.score(X_test, Y_test), 2)
mae_lr = round(metrics.mean_absolute_error(Y_test, Y_pred_lr), 4)
mse_lr = round(metrics.mean_squared_error(Y_test, Y_pred_lr), 4)

# KNN
knn = KNeighborsClassifier(n_neighbors=2)
knn.fit(X_train, Y_train)
Y_pred_knn = knn.predict(X_test)
```

```
# KNN model evaluation
knn_score = round(knn.score(X_test, Y_test), 2)
mae_knn = metrics.mean_absolute_error(Y_test, Y_pred_knn)
mse_knn = metrics.mean_squared_error(Y_test, Y_pred_knn)

# Model comparison
compare_models = pd.DataFrame({
 'Model': ['Logistic Regression', 'KNN'],
 'Accuracy': [logreg_score, knn_score],
 'MAE': [mae_lr, mae_knn],
 'MSE': [mse_lr, mse_knn]
})
print(compare_models)
```

Thus, based on the above, it should be noted that in the scientific literature there is an opinion regarding the fact that the logistic regression model, compared to KNN, has low MAE and MSE values. What is the preferred choice for building a model [4].

2. METHODS FOR ERROR ANALYSIS AND MODEL DIAGNOSTICS

Error analysis and model diagnostics methods play a key role in the development and improvement of machine learning. These methods aim to identify the sources of inaccuracies and shortcomings in a model’s performance, as well as to enhance its overall efficiency. Several approaches exist that allow for effective diagnosis of errors and the correction of model performance.

The first method for error analysis is decomposing the error into its components. This helps to understand what contributes to the model's reduced accuracy. The main components of error include bias, variance, and data noise. For example, high bias error indicates a systematic underestimation of true values, which requires a revision of the model structure. On the other hand, high variance points to overfitting, which can be addressed by increasing the dataset size or applying regularization techniques. Below in Figure 2, the elements that make up this method will be presented.

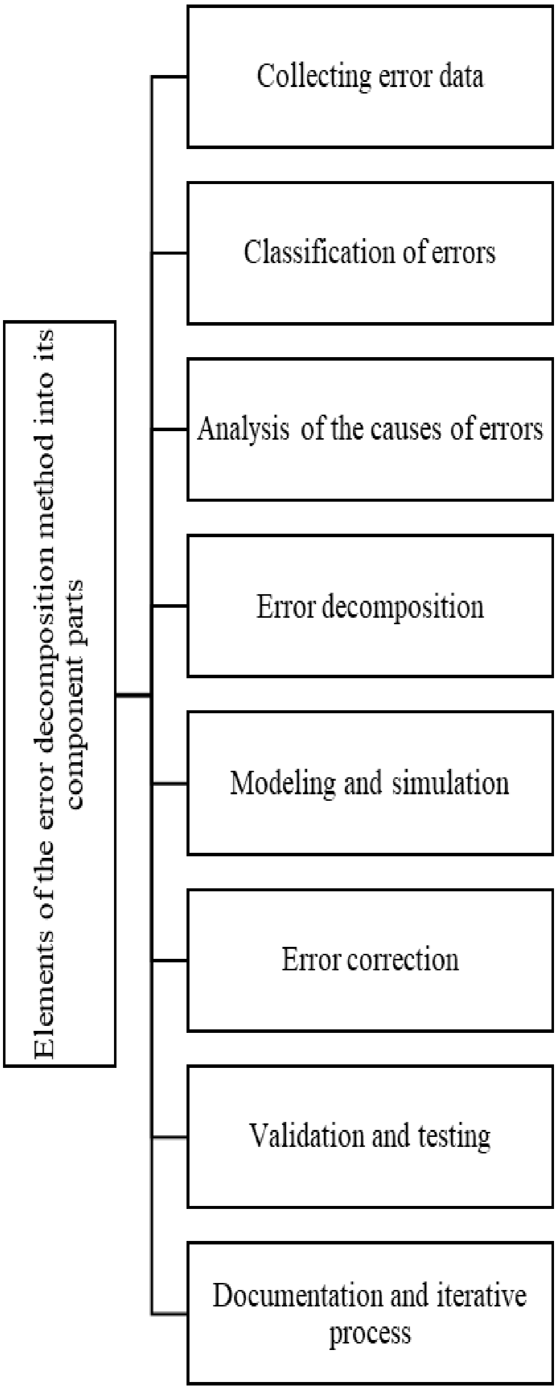

Elements of the error decomposition method into its component parts
Collecting error data
Classification of errors
Analysis of the causes of errors
Error decomposition
Modeling and simulation
Error correction
Validation and testing
Documentation and iterative process

Fig.2. The elements that make up the method of decomposing errors into their component parts (compiled independently)

The next method is residual analysis, i.e., the difference between predicted models and actual values. This method helps identify where and why the model is making errors. Visualizing residuals using graphs can help uncover systematic errors, such as underestimation or overestimation of certain classes or objects [5]. Below, Figure 3 presents the components of the residual analysis method used to analyze and reduce errors in artificial intelligence models.

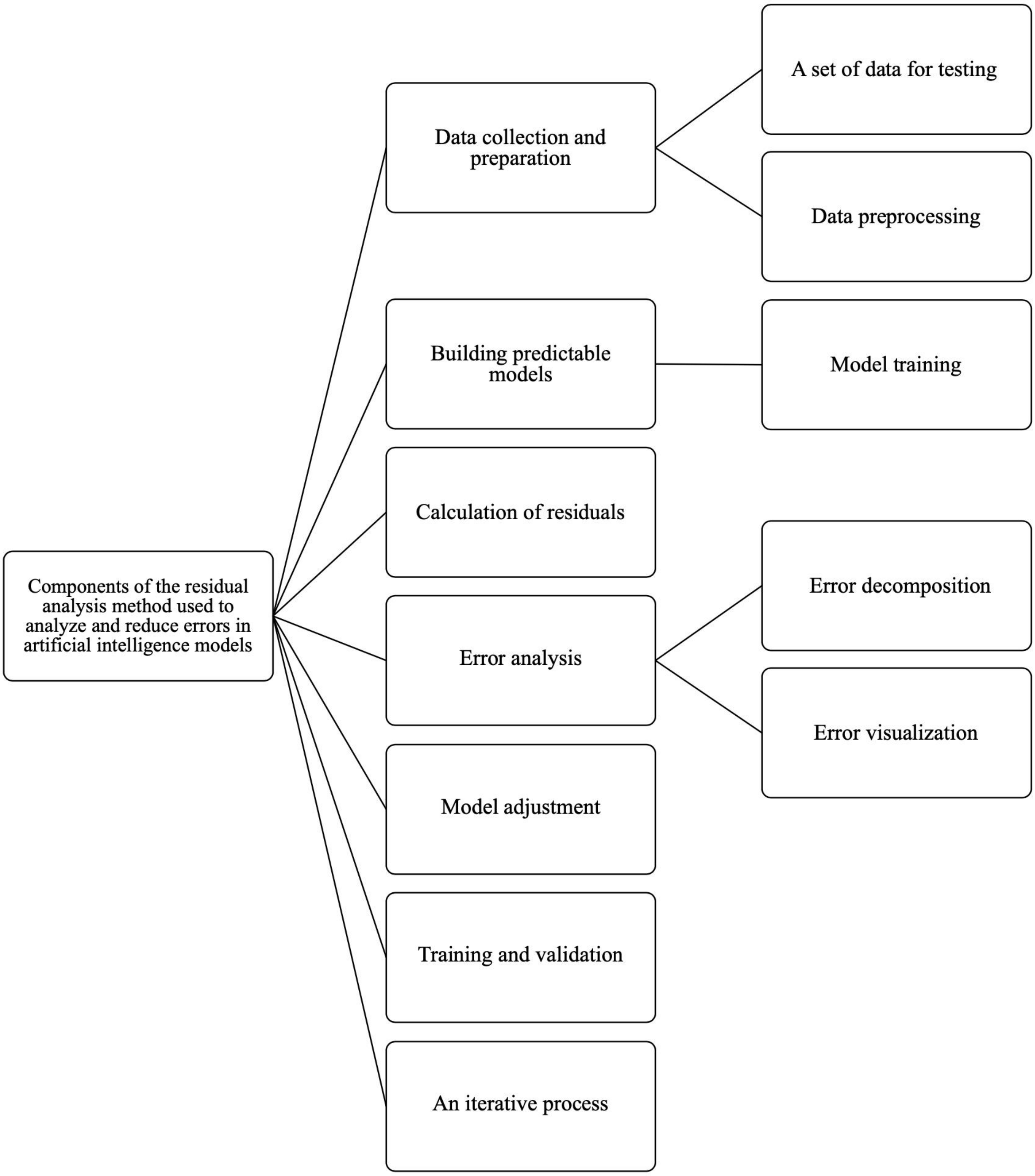

Components of the residual analysis method used to analyze and reduce errors in artificial intelligence models
Data collection and preparation
A set of data for testing
Data preprocessing
Building predictable models
Model training
Calculation of residuals
Error analysis
Error decomposition
Error visualization
Model adjustment
Training and validation
An iterative process

Fig.3. Components of the residual analysis method used to analyze and reduce errors in

Another significant approach is cross-validation, which helps assess a model's generalization ability. It prevents overfitting by splitting the data into training and testing sets. Moreover, techniques such as K-fold cross-validation provide a more accurate evaluation of model performance through repeated testing on various data subsets.

Sensitivity analysis and feature importance analysis are also essential diagnostic tools. They help determine which features in the data contribute most to the model’s predictions and which ones may be noise. Eliminating less important features can improve both the model’s performance and its interpretability.

Additionally, data visualization methods, such as heatmaps and feature importance charts, help visually represent data relationships and contribute to understanding the causes of model errors. This approach is especially useful for complex models, such as deep neural networks [6].

3. APPROACHES TO REDUCING AND PREVENTING ERRORS

Analyzing the results of McKinsey's "The State of AI in 2021" report, released in December 2021, several key areas can be identified for mitigating risks in the development and application of artificial intelligence (AI) solutions.[7].

Both groups of respondents also emphasized the importance of the interpretability of AI models. This term refers to the need to explain the decisions made by AI algorithms. Business clients demand transparency because the responsibility for decisions made by the system still lies with people, not machines. Organizations that have successfully integrated AI into their processes have already established methods and procedures to ensure model transparency.

Ways to mitigate risks when implementing AI. McKinsey identified three key areas where companies can reduce risks when working with AI:

1. Working with data for training and testing models. To mitigate risks in creating and testing AI systems, companies use several methods. One of them is checking data samples for insufficient representation (according to 47% of companies with highly effective AI implementation). The quality of models directly depends on data quality, so protecting against errors and omissions is crucial for improving AI prediction accuracy.

Another important step is checking data for bias and distortions, also supported by 47% of respondents. Despite automatic verification procedures, the involvement of data engineers is still necessary to clean and correct the data.

A further method includes protecting data from incorrect input as volumes increase (36% of companies). As data accumulates, ensuring its correctness is essential for maintaining model quality.

2. Evaluating accuracy and model adjustments. AI models are dynamic systems that require regular updates and retraining. Companies that have successfully applied AI use the following approaches: retraining when issues are detected, monitoring data deviations and model concepts, and reviewing decisions with experts. Additionally, training users to identify errors in AI operations is a key aspect.

3. Documenting processes. Documentation remains an integral part of any IT infrastructure. For AI, this includes systematically recording model performance, architecture, the data used, and the training process. It is also important to keep records of issues and trade-offs that arise during the operation of AI solutions [7].

However, there are problematic aspects. Thus, the problem of overfitting in machine learning presents a significant challenge for specialists developing intelligent systems. This phenomenon occurs when a model becomes overly adapted to the data on which it was trained, greatly reducing its ability to generalize to new, previously unseen data. While the model may show excellent results on the training data, its performance in real-world conditions can be ineffective. A deep understanding of the mechanisms behind overfitting helps identify its causes and develop strategies to mitigate this effect. Overfitting should be considered within the framework of model risk management, as it impacts the quality of decision-making.

Overfitting is closely related to the concept of the trade-off between bias and variance in a model. A model with high bias may be too simple, leading to underfitting and an inability to capture important patterns in the data. In contrast, high variance indicates that the model is overly complex, picking up random variations in the data. The optimal solution in machine learning lies in finding a balance between these two extremes, allowing for the creation of a model with strong generalization capabilities.

One of the key methods for detecting overfitting is cross-validation. This approach splits the data into several subsets for training and testing the model, providing a more reliable assessment of its generalization ability. By using different parts of the data for validation at each stage, cross-validation helps determine how well the model will perform with new data. Techniques such as k-fold cross-validation significantly reduce the risk of overfitting, improving the model's generalization capabilities.

Regularization methods are aimed at reducing the risk of overfitting by introducing constraints on model complexity. Regularization helps prevent the model from becoming too adapted to the data, making it more robust to noise and outliers. The most popular regularization methods include L1 and L2 regularization, as well as dropout and early stopping when training neural networks. These approaches ensure the model performs more reliably on new data, avoiding overfitting.

Feature selection and engineering play a crucial role in combating overfitting. Sometimes reducing the number of features or developing new, more informative ones can significantly improve the model. Reducing unnecessary features simplifies the model, making it less prone to overfitting. This is especially important in tasks involving large datasets, such as text or image classification.

Ensemble learning methods, which involve combining multiple models, play a key role in increasing resilience to overfitting. Approaches such as random forests or gradient boosting combine the outputs of several models, helping to offset errors caused by overfitting in individual models. Averaging predictions from different models significantly reduces the risk of overfitting and improves the generalization ability of the final model [8].

In conclusion, understanding the nature of overfitting and applying various methods to prevent it are crucial aspects of developing high-quality machine learning models. Maintaining a balance between model complexity and generalization ability enables the creation of systems that will perform reliably in real-world conditions.

CONCLUSION

The analysis has shown that errors in AI models represent a significant issue that impacts the accuracy and reliability of their predictions. For the successful application of AI models, it is crucial to consider the full range of possible error sources and implement methods to reduce them at all stages of development and deployment. Cross-validation, regularization, and the use of ensemble models have proven effective in combating overfitting and other types of errors. Therefore, systematic analysis and adjustment of model parameters remain essential for the successful application of AI in real-world conditions, improving system performance and reinforcing user trust in AI technologies.

REFERENCES

Moor M. et al. Foundation models for generalist medical artificial intelligence //Nature. – 2023. – vol. 616. – No. 7956. – pp. 259-265.

Sarker I. H. AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems //SN Computer Science. – 2022. – Vol. 3. – No. 2. – p. 158.

Zinoviev V. A., Zinoviev D. V. On the generalized cascade construction of codes in modular metrics and Lie metrics //Problems of information transmission. – 2021. – vol. 57. – No. 1. – pp. 81-95.

Models Score and Error. [Electronic resource] Access mode: https://www.geeksforgeeks.org/ml-models-score-and-error / (accessed 09/06/2024).

Yablokov A. E., Blagoveshchensky I. G. Scientific and practical foundations for the creation of automated systems for technical monitoring and diagnostics of grain processing enterprises' equipment based on neural network data analysis methods //Yablokov AE, Blagoveshchenskiy IG–M., MGUPP. – 2022.

Grachev V. V. et al. The method of synthesis of neural network diagnostic models of complex technical objects //Automation in transport. - 2020. – vol. 6. – No. 4. – pp. 466-483.

Ways to reduce the risks of AI solutions. [Electronic resource] Access mode: https://datanomics.ru/artciles/sposoby-snizheniya-riskov-ai-reshenij / (accessed 06.09.2024).

Retraining: Preventing model risk by using methods to prevent overfitting. [Electronic resource] Access mode: https://fastercapital.com/ru/content/%.html (accessed 06.09.2024).