Technoeconomics. 2025, Vol. 4, No. 4 (15). Pp. 36–43.
Техноэкономика. 2025, Том 4, № 4 (15). С. 36–43.

Scientific article
UDC 330.47
DOI: https://doi.org/10.57809/2025.4.4.15.2

PREDICTING CLAIMS IN AUTO INSURANCE USING DEEP NEURAL NETWORKS

Andrei Klimentov ✉

Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

✉ andreiklimentov361@gmail.com

Abstract. In the modern world, the insurance market is subject to significant changes, including under the influence of the use of digital technologies and the introduction of machine learning methods in insurance scoring. The object of the study is a data set with records of insurance policies. The study uses a deep nonlinear neural network to predict the occurrence of claim loss on auto insurance policies. Before using a multilayer neural network, data is pre-processed, and possible data leakage is eliminated. At the output of the neural network model, the resulting loss probability value is converted to a binary value. The model is evaluated using the ROC-AUC metric, with a graph of the ROC curve. The results show that the obtained model has predictive accuracy, but not high enough accuracy for industrial applications of the chosen model. The findings indicate the need for further research on ways to solve this problem using other machine learning methods.

Keywords: machine learning, neural networks, insurance scoring, prediction of insurance events, auto insurance, ROC-AUC, classification, scoring

Citation: Klimentov A. Predicting claims in auto insurance using deep neural networks. Technoeconomics. 2025. 4. 4 (15). 36–43. DOI: https://doi.org/10.57809/2025.4.4.15.2

Научная статья
УДК 330.47
DOI: https://doi.org/10.57809/2025.4.4.15.2

ПРОГНОЗИРВОАНИЕ УБЫТКА В АВТОСТРАХОВАНИИ С ИСПОЛЬЗОВАНИЕМ МНОГОСЛОЙНЫХ НЕЙРОННЫХ СЕТЕЙ

Андрей Климентов ✉

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Россия

✉ andreiklimentov361@gmail.com

Аннотация. В современном мире страховой рынок подвержен значительным изменениям в том числе под влиянием применения цифровых технологий и внедрения методов машинного обучения в стразовой скоринг. Объектом исследования является набор данных с записями о страховых полисах. В исследовании используется многослойная нелинейная нейронная сеть для предсказания наступления убытка по полисам автострахования. Перед использованием многослойной нейронной сети проводится предварительная обработка данных, устранение возможны утечек данных. На выходе модели нейронной сети получаемое значение вероятности убытка преобразуется в бинарное значение. Оценка модели проводится по метрике ROC-AUC, с построением графика ROC кривой. Результаты показывают, что полученная модель имеет предсказательную, но недостаточно высокую точность для промышленного применения выбранной модели. Выводы указывают на необходимость дальнейшего исследования способов решения поставленной задачи при помощи других методов машинного обучения.

Ключевые слова: машинное обучение, нейронные сети, страховой скоринг, прогнозирование страховых событий, автострахование, ROC-AUC, классификация, тарификация

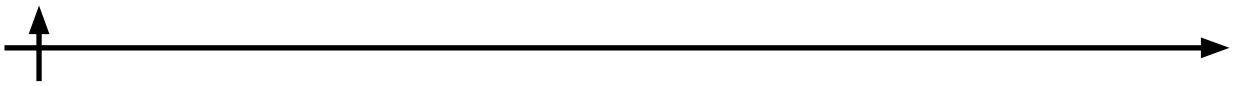
Для цитирования: Климентов А. Прогнозирование убытка в автостраховании с использованием многослойных нейронных сетей // Техноэкономика. 2025. Т. 4, № 4 (15). С. 36–43. DOI: https://doi.org/10.57809/2025.4.4.15.2

Introduction

In the modern world, the insurance market is subject to significant changes, including under the influence of the use of digital technologies. In particular, machine learning methods are actively used in insurance to solve various problems. The main tasks are to calculate the cost of insurance policies based on the assessment of the client's riskiness and forecasting claim losses. These methods are also actively used to counter fraud (Zabavin, 2009; Vorobyev, 2024; Ignatiev, Levina, 2024).

Thus, according to the data of the Central Bank of Russia for 2024 and the first quarter of 2025, the growth rate of insurance premiums year-on-year exceeds 100% (Bank of Russia, 2025). Such an increase may be due to the widespread introduction of machine learning technologies for assessing insurance policy risks (Makarenko, 2020), as well as calculating insurance premiums.

Scoring is commonly referred to as an automated mathematical scoring system that can be used to assess a client's solvency, for example in the banking sector. In insurance, scoring models can be used to determine the degree of risk in insurance based on multifactor models. For example, in auto insurance, scoring models often use the age of persons allowed to drive a vehicle (TS), as well as the power of the vehicle (Southwell, 2008).



Modern scoring systems based on ML algorithms make it possible to take into account complex nonlinear dependencies, use a wide range of data, including behavioral and external sources, and dynamically adapt to changing market conditions.

The relevance of using machine learning methods in calculating insurance premiums and predicting risks lies in their adaptability and the ability to detect nonlinear dependencies in a large amount of data.

Linear regression models (Varghese and Dash, 2012) and decision tree models (Breima, 2001; Salzberg, 1994) have become the most widespread in insurance scoring systems. However, these groups of methods have limitations that reduce their effectiveness. Thus, logistic regressions often have insufficient accuracy, especially when it comes to nonlinear dependencies or regression parameters are subject to multicollinearity (Kuznetsova, 2015). In turn, a group of machine learning methods based on decision trees are susceptible to overfitting (Karamazin, 2024; Salzberg 1994), and also have a low ability to scale, since when the system or data changes, the model must be completely rebuilt.

Ensemble models, especially those based on boosting, are also widely used. Boosting allows you to create models that consist of simple models combined sequentially, which reduces the errors of each model (Chen and Guestrin, 2016; Diana et al., 2019).

At the same time, boosting models are subject to the problem of class imbalance, complexity of configuration, and poor adaptability to sudden changes (Averro et al., 2023; Coskun and Turanli, 2023).

In addition, neural networks are used, but currently their use in scoring models is limited. Thus, it is of interest to conduct a study aimed at exploring the possibility of using neural networks to predict an insurance event, as well as to evaluate the quality of such models.

In this paper several research questions will be observed. First of all, it will be researched whether multilayer neural networks with nonlinear activation functions will be effective method for insurance events (claims) prediction. Secondly, the level of predictive accuracy by ROC-AUC metric, which can be achieved with this method, will also be question of research. Moreover, advantages and disadvantages of neural network approach to the vehicle insurance will be observed in this work. Furthermore, the possibility of practical applications of deep neural networks will be considered.

Materials and Methods

There is great variability in machine learning methods which are taken into account for dealing with insurance claim prediction. Firstly, models based on decision trees are wide spread and were used in works

The research in this paper will be conducted on the basis of a dataset (Segura-Gisbert et al., 2023) from the Kaggle repository, which characterizes the database of an auto insurance company.

The advantage of the set (Segura-Gisbert et al., 2023) is a large set of attributes, which will allow you to select only the most significant ones due to exploratory data analysis. In addition, a large number of records in the dataset will allow it to be divided into training and verification samples.

For further work, preliminary data processing was carried out. First of all, the parameters were extracted from the time attributes, so, for example, the length of service attribute in numeric format can be obtained from the "date of receipt of the driver's license" attribute. Further, categorical features were also processed, as this is a prerequisite for their inclusion in the neural network model (Valiullin, 2017; Barkov and Senotova, 2021). For encoding, the method of encoding by the name of the feature class (Label Encoding) was applied. Other temporary

attributes were also removed after that.

Also, one of the most important aspects in the pre-processing of the data was the removal of the features 'N_claims_year', 'Cost_claims_year', 'N_claims_history', which determine the number of claims under the policy, and the amount of claims. The removal of these features is necessary, as they will create a data leakage when training a neural network model.

An additional loss attribute ('claim_prob') was formed as a predicted feature, which is determined binarily (equal to 1 if there was a loss and 0 if there was no loss).

In addition, the data set is divided into training and test samples. The separation was made in the ratio of 80% of the data included in the training sample, and 20% in the test sample.

Thus, the classification task is set to predict the occurrence of a loss {0,1}. The number of regressors in this task is 14, which requires the use of a deep neural network.

Next, the architecture of the machine learning model was defined. A multilayer neural network with nonlinear activation functions will be used to predict the probability of an insurance event. The ReLU activation function will be used on the input and hidden layers of the neural network, which looks like:

$$ReLU(x) = \max(0, x) = \begin{cases} x, if\ x \geq 0 \\ 0, if\ x < 0' \end{cases} \quad (1)$$

where x – variable on the input of the neuron.

According to (Dubey et al. 2021), this activation function allows solving the problem of decaying gradients and is the standard choice for most tasks solved using neural networks.

The sigmoid activation function (2) will be used on the output layer, which allows you to project the values of the output variable to the interval [0,1], which corresponds to the problem being solved.

$$Sigmoid(x) = \frac{1}{1+e^{-x'}} \quad (2)$$

The mathematical description of the inner layers of a neural network can be represented as (3) according to (Blier-Wong et al. 2020).

$$\begin{gathered} h_j = ReLU(z_j),\ j = 1,...,J \\ z_j = \sum_{k=1}^{14} \omega_{kj} x_{ik} + b_j,\ j = 1,...,J \end{gathered} \quad (3)$$

where J is the width of the hidden layer, g is the activation function, x_ik are the input parameters of the model, i is the observation number, and p is the number of input variables.

The next part is to select the number of hidden layers and the number of neurons on them. According to (Aziz et al. 2024), there is no universal algorithm for selecting the number of hidden layers and neurons per layer. Thus, you can be guided by empirical experience, and you can vary the parameters during experiments.

The quality of the neural network model will be evaluated using a standard metric for the classification problem ROC-AUC. Thus, according to (Stern 2021), the ROC curve shows the ratio of true positive results and false positive results at different risk thresholds. In turn, the AUC area under the ROC curve can be calculated using the formula (4).

$$AUC = \int_0^1 \left(\frac{(1-r) f(r)}{1-r_{mean}} \right) \left(\int_r^1 \frac{x f(x)}{r_{mean}} dx \right) dr \quad (4)$$

where x is an artificial variable for integration, r is the probability of an object belonging to a positive class, $f(r)$ is a probability density function, and r_{mean} is the mathematical expectation of an object belonging to a class.

The advantages of using this metric are its invariance to class imbalance, as well as statistical stability (Richardson et al. 2023).

Results

During the experiments, the best classification accuracy values for the ROC-AUC metric were obtained with the following neural network configuration and represented on Fig. 1.

Fig. 1. The view of neural network architecture.

In the process of training the model on the training sample, the saturation level of the model was determined, after which an increase in the number of training epochs does not significantly affect the accuracy of the model. The training schedule for the model is shown on Fig. 2.

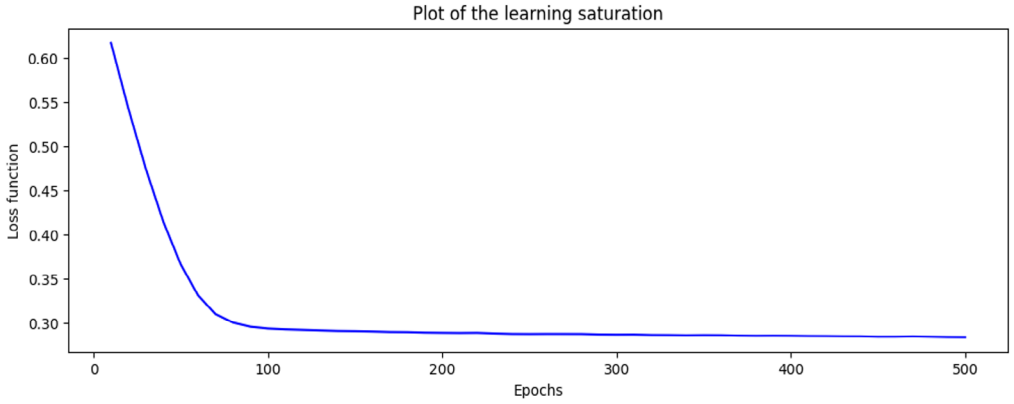


Fig. 2. The plot of the dependence of the model's loss function on the learning epoch.

The graph shows that after 300 epochs, an increase in their number does not significantly affect the accuracy of the model. After training the model, the model was validated on a test dataset. The value of the ROC-AUC metric was also obtained (see Fig. 3).

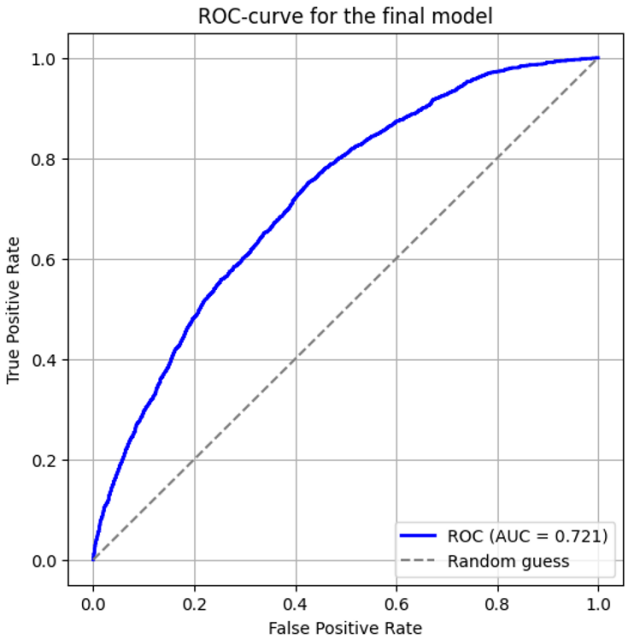


Fig. 3. The plot of ROC curve for the final model.

The exact value of the ROC-AUC metric was 0.72. A model with a similar metric value can be classified as a model that has predictive power and value, but is not optimal.

Thus, the use of multilayer neural networks with nonlinear activation functions makes it possible to solve the problem of predicting the occurrence of an insurance event. In addition, this approach allows you to use the output value of the neural network, which is the probability of an insurance event, when charging the cost of the policy.

Conclusion

In this paper, we investigated the use of a neural network with nonlinear activation functions to solve the binary classification problem in predicting the occurrence of loss.

A distinctive feature of this study is the use of deep neural networks as opposed to methods based on decision trees such as random forest or boosting methods. Moreover, approach with neural networks is not a standard practice in insurance industry due to its low interpretability. As a result, this research allows to evaluate nonstandard approach and make decision whether it is useful to apply it in industry.

The resulting neural network model has predictive power, but it is not accurate enough for industrial applications.

The advantage of using a neural network is that even in the classification task, its output value takes values in the range from 0 to 1, which can be used as the probability of a loss on the policy. This discretized value can be used more flexibly in insurance billing than a discrete value of 0 or 1. Thus, this solution has high prospects for use in the billing of auto insurance policies.

However, the accuracy of the neural network model is not high enough for it to have industrial applications, so other machine learning models should be further explored to solve this problem.

As a result, considering the research questions which are stated in introduction following conclusions could be made. The deep neural network with nonlinear activation functions is not as effective as it was expected because the value of ROC-AUC metric does not exceed 0.72. re are several Thus, the practical application of deep neural networks in auto insurance seems to be bounded due to its average evaluation results.

This is main disadvantage of this approach which was observed in this research. On the other hand, there are several advantages of this method, such as nonlinear dependencies prediction, which are also observed in this paper.

REFERENCES

Averro N., Murfi H., Ardaneswari G. 2023. The Imbalance Data Handling of XGBoost in Insurance Fraud Detection. Science and Technology Publications, 460–467.

Aziz A., Khan T., Iftikhar U., Tanoli I., Qureshi A.K. 2024. Appropriate Selection for Numbers of neurons and layers in a Neural Network Architecture: A Brief Analysis. Sir Syed University Research Journal of Engineering & Technology, 2 (13).

Batten J. A., Ciner C., Lucey B. M. 2008. The Macroeconomic Determinants of Volatility in Precious Metals Markets. SSRN Electronic Journal.

Barkov D., Senotova S. 2021. Encoding of categorial features in neural networks. Scientific Papers Collection of the Angarsk State Technical University 1, 3–8.

Blier-Wong C., Cossette H., Lamontagne L., Marceau E. 2020. Machine Learning in P&C Insurance: A Review for Pricing and Reserving. Risks 1 (9), 4.

Breiman L. 2001. Random forests.

Chen T., Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System, 785–794.

Cohen G. 2022. Algorithmic Strategies for Precious Metals Price Forecasting. Mathematics 7 (10), 1134.

Coskun S. B., Turanli M. 2023. Credit risk analysis using boosting methods. Journal of Applied Mathematics, Statistics and Informatics 1 (19), 5–18.

Diana A., Griffin J. E., Oberoi J. 2019. Machine-Learning Methods for Insurance Applications. Schaumburg, Illinois: Society of Actuaries.

Dubey S. R., Singh S. K., Chaudhuri B. B. 2021. Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark.

Hansen B. E. 2005. Challenges for econometric model selection. Econometric Theory, 1 (21).

Ignatiev P., Levina A. 2024. Artificial intelligence and artificial neural networks in healthcare. Technoeconomics 3, 4 (11), 28–41. DOI: https://doi.org/10.57809/2024.3.4.11.3

Karmazin A. R. 2024. Comparative analysis of the efficiency of various scoring models in factoring.

Kuznetsova I. S. 2015. Problems of multicollinearity in regression models, 3.

Makarenko E. A. 2020. Application of computer scoring in the risk assessment system in property insurance.

Richardson E., Trevizani R., Greenbaum J.A., Carter H., Nielsen M., Peters B. 2024. The receiver operating characteristic curve accurately assesses imbalanced datasets. Patterns 5, 6, 100994.

Salzberg S. L. 1993. Review of C4.5: Programs for Machine Learning by J. Ross Quinlan. Machine Learning 3 (16), 235–240.

Segura-Gisbert J., lledo Benito J., Pavia J. 2023. Dataset of an actual motor vehicle insurance portfolio. doi: 10.21203/rs.3.rs-3631821/v1.

Stern R. H. 2021. Interpretation of the Area Under the ROC Curve for Risk Prediction Models.

Southwell W. 2008. Increasing profitability in a market with tariff competition, 1.

Valiullin A. M. 2017. Preprocessing of categorical features in machine learning problems [In Russ.] In: Robotics and Artificial Intelligence: Proceedings of the IX All-Russian Scientific and Technical Conference with International Participation, 154–157.

Varghese A. A., Dash M. 2012. A Linear Pricing Model for Life Insurance Policies. SSRN Electronic Journal.

Vorobyev I. A. 2024. ML methods for assessing the risk of fraud in auto insurance // Izvestiya of Saratov University. Mathematics. Mechanics. Informatics. 4 (24), 619–628.

Zabavin D. V. 2009. Application of scoring systems to counter crimes in the insurance sphere.

Bank of Russia. 2025. Statistics. URL: https://cbr.ru/insurance/reporting_stat/ (date of access: 13.10.2025).

СПИСОК ИСТОЧНИКОВ

Averro N., Murfi H., Ardaneswari G. 2023. The Imbalance Data Handling of XGBoost in Insurance Fraud Detection. Science and Technology Publications, 460–467.

Aziz A., Khan T., Iftikhar U., Tanoli I., Qureshi A.K. 2024. Appropriate Selection for Numbers of neurons and layers in a Neural Network Architecture: A Brief Analysis. Sir Syed University Research Journal of Engineering & Technology, 2 (13).

Batten J. A., Ciner C., Lucey B. M. 2008. The Macroeconomic Determinants of Volatility in Precious Metals Markets. SSRN Electronic Journal. 2008.

Barkov D., Senotova S. 2021. Encoding of categorial features in neural networks. Scientific Papers Collection of the Angarsk State Technical University 1, 3–8.

Blier-Wong C., Cossette H., Lamontagne L., Marceau E. 2020. Machine Learning in P&C Insurance: A Review for Pricing and Reserving. Risks 1 (9), 4.

Breiman L. 2001. Random forests.

Chen T., Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System, 785–794.

Cohen G. 2022. Algorithmic Strategies for Precious Metals Price Forecasting. Mathematics 7 (10), 1134.

Coskun S. B., Turanli M. 2023. Credit risk analysis using boosting methods. Journal of Applied Mathematics, Statistics and Informatics 1 (19), 5–18.

Diana A., Griffin J. E., Oberoi J. 2019. Machine-Learning Methods for Insurance Applications. Schaumburg, Illinois: Society of Actuaries.

Dubey S. R., Singh S. K., Chaudhuri B. B. 2021. Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark.

Hansen B. E. 2005. Challenges for econometric model selection. Econometric Theory, 1 (21).

Ignatiev P., Levina A. 2024. Artificial intelligence and artificial neural networks in healthcare. Technoeconomics 3, 4 (11), 28–41. DOI: https://doi.org/10.57809/2024.3.4.11.3

Кармазин А. Р. 2024. Сравнительный анализ эффективности различных скоринговых моделей в факторинге.

Кузнецова И. С. 2015. Проблемы мультиколлинеарности в регрессионных моделях, 3.

Макаренко Е. А. 2020. Применение компьютерного скоринга в системе оценки рисков в страховании имущества.

Richardson E., Trevizani R., Greenbaum J.A., Carter H., Nielsen M., Peters B. 2024. The receiver operating characteristic curve accurately assesses imbalanced datasets. Patterns 5, 6, 100994.

Salzberg S. L. 1993. Review of C4.5: Programs for Machine Learning by J. Ross Quinlan. Machine Learning 3 (16), 235–240.

Segura-Gisbert J., lledo Benito J., Pavia J. 2023. Dataset of an actual motor vehicle insurance portfolio. doi: 10.21203/rs.3.rs-3631821/v1.

Stern R. H. 2021. Interpretation of the Area Under the ROC Curve for Risk Prediction Models.

Саузвел У. 2008. Увеличение прибыльности на рынке с тарифной конкуренцией, 1.

Валиуллин А. М. 2017. Предобработка категориальных признаков в задачах машинного обучения. Робототехника и искусственный интеллект : Материалы IX Всероссийской научно-технической конференции с международным участием, 154–157

Varghese A. A., Dash M. 2012. A Linear Pricing Model for Life Insurance Policies. SSRN Electronic Journal.

Vorobyev I. A. 2024. ML methods for assessing the risk of fraud in auto insurance // Izvestiya of Saratov University. Mathematics. Mechanics. Informatics. 4 (24), 619–628.

Забавин Д. В. 2009. Применение скоринговых систем в целях противодействия преступлениям в сфере страхования.

Банк России. 2025. Статистика. URL: https://cbr.ru/insurance/reporting_stat/ (дата обращения: 13.10.2025).

INFORMATION ABOUT AUTHOR / ИНФОРМАЦИЯ ОБ АВТОРЕ

KLIMENTOV Andrei R. – student.
E-mail: andreiklimentov361@gmail.com
КЛИМЕНТОВ Андрей Романович – студент.
E-mail: andreiklimentov361@gmail.com

Статья поступила в редакцию 09.11.2025; одобрена после рецензирования 20.11.2025; принята к публикации 11.12.2025.

The article was submitted 09.11.2025; approved after reviewing 20.11.2025; accepted for publication 11.12.2025.