Ensemble learning for insurance premium prediction: a comparative analysis of XGBoost, Random Forest, and SVM

Danyang Yao[1,*], Jiayu Li[1], Yiqi Shen[1]

[1]School of Mathematics and Science, XJTLU, Suzhou, Jiangsu, China

*Corresponding author: Danyang.Yao21@student.xjtlu.edu.cn

Keywords: Health Insurace, Random Forest, XGBoost, Support Vector Machines, Forecast

Abstract: The calculation of premium is an important part for insurance companies to research and introduce new types of insurance. This paper collects the data of insurance companies, and will explore which factors are related to insurance premium from the variables of gender, age, body variable index, whether smoking, number of children and region, and first explore the correlation between factors with correlation coefficient, and then use XGB, random forest and svm model to analyze them one by one. All models show that smoking, age and body mass index have greater influence on insurance premium.

1. Introduction

With the continuous development of modern society and economy, people are paying more and more attention to their own health, and as a result, the insurance industry is also receiving more and more attention from people. The insurance industry refers to a financial service industry with insurance companies as the main body, the insurance business as the core, and the purpose of providing risk protection and capital management appreciation. Insurance provides protection for customers such as families and businesses, taking on some of their risks, thereby reducing their financial pressure and providing financial support in the event of accidents. Premium is an important aspect of insurance, and an appropriate premium can maintain the profitability of an insurance company while also sharing risks for customers. It can be said that accurate insurance premiums are of great significance for social development and improving socio-economic benefits. Therefore, this article uses various machine learning models to explore the relationship between premiums and various indicators and to identify factors that affect the size of premiums. In order to delve into the relevant factors that affect premiums, we first use a correlation matrix to eliminate similar variables and obtain variables with weaker correlations that can represent independent categories. Secondly, we use Random Forest, XGBoost, and SVM models to simulate the correlation between variables and premiums. Finally, we use MAE to compare the accuracy of three models in predicting different factors.

The above is the first part of this article - Introduction. Next, we will cite many papers to conduct a review and discuss the predictive research of relevant machine learning models in premium, as well as the reasons for choosing Random Forest, XGBoost, and SVM models to simulate the correlation between variables and premiums. In the third part, we will discuss the application of the above three models and explain the relevant variables in the models. Finally, we give a summary of this paper.

2. The literature review of health insurance models

2.1 The machine learning models

Nowadays, many studies use machine learning models to predict health insurance costs. Among them, authors Aggarwal and Shruti Anmol [1] used supervised learning as a type of machine learning and tested the system using linear regression, with the best performing model being the Gradient Boosting Regression model, depending on a decision tree. Similar to the previous author, author Thejeshwar et al. [2] also compared the model's estimation with the actual premium to test its accuracy. Among Support Vector Machine Regression (SVR), Random Forest Regression, and

DOI: 10.25236/ieesasm.2023.066

Linear Regression, Random Forest Regression was thought as the highest performance. Additionally, Authors Ramya, Manigandan, and Deepa [3] also agreed that Random Forests provide higher performance and believed that Machine Learning algorithms (ML) can also be used to predict insurance claims and manage large amount of data.

With the rapid development of medical insurance, fraud in this sector is becoming a serious issue. Hence, Veena K [4] et al. used data mining and machine learning methods to automate the detection of medical fraud. Based on the pre-processed data, a machine learning model is constructed and compared with four algorithms: Logistic Regression, Random Forest, Decision Tree Classifier, and Naive Bayes. Among them, the highest accuracy was achieved through the Decision Tree Classifier algorithm, with an accuracy of 97.03%. Moreover, Venkateswarlu Nalluri [5] et al. used computational intelligence methods to predict whether a person can apply for health insurance by evaluating a number of relevant machine learning algorithms. And Saraswat, Birendra Kumar et al. [6] first extracted 19 important factors. Artificial intelligence or deep learning methods are then used to build automated auditing mechanisms that can easily detect large amounts of healthcare fraud.

Next, let's take a closer look at some research on the three models used in this article. Firstly, many authors use random forest models to make predictions. For example, authors Baro, Oliveira, and de Souza Britto Junior [7] used a random forest model to predict hospitalization from health insurance data and created a hospitalization prediction model. Meanwhile, authors Reinke et al. [8] also used statistical methods such as random forest to determine accuracy and predict the risk of dementia from data from Germany's largest health insurance company. Secondly, for neural networks, both authors Kaushik et al. [9] and Kreif et al. [10]used machine learning based on regression models with various parameters to train and evaluate the artificial neural network model, achieving high accuracy and being able to better predict health insurance premiums. Finally, Author Choi et al. [5] analyzed data from the South Korean National Health Insurance Service and identified important variables related to high healthcare costs. Then, A series of models were used for supervised learning to predict high healthcare costs, with XGB having the best performance.

2.2 The reason of choosing these two models

To sum up, we can summarize the reasons of using the three models of random forest, SVM and XGB in this article. Firstly, many authors are willing to use random forest models to make predictions. It is not difficult to find that it has the advantages of extremely high accuracy and can effectively run on large data sets. Secondly, the accuracy of XGB for prediction is also relatively high, and it is currently recognized as a model with good classification effect. Thirdly, SVM model has the advantage of solving machine learning in the case of small samples. All in all, these three models have high accuracy and many advantages for prediction.

2.3 The establishment of simulation model

The Correlation Matrix is a square matrix, which is used to represent the correlation between multiple variables. Correlation measures the strength and direction of the linear relationship between two variables. Correlation matrix is usually used in statistical analysis and data mining to help identify the correlation between variables. The formula of correlation matrix is based on Pearson Correlation Coefficient, which is used to measure the linear relationship between two variables.

The calculation formula of correlation coefficient (r) is as follows:

$$r=\frac{n\sum xy-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}}$$

Among them:

(1) (r) is the correlation coefficient.

(2) (n) is the number of samples.

(3) (x) and (y) are the values of two variables of the sample respectively.

The value of correlation coefficient is between -1 and 1, which indicates the degree of correlation between two variables:

When (r = 1) means perfect positive correlation, it means that the two variables change in exactly the same proportion.

When (r = -1), it means completely negative correlation, which means that the two variables change in completely opposite proportions.

When (r = 0), it means that there is no linear relationship and there is no correlation between the two variables.

The correlation matrix is a square matrix, in which each element represents the correlation coefficient between the horizontal variable and the vertical variable. For a data set with (n) variables, the size of the correlation matrix is $(n * n)$.

Properties of correlation matrix:

The diagonal element is always 1, because the correlation coefficient between each variable and itself is 1.

The correlation matrix is symmetric, so (R{ij} = R{ji}).

Correlation matrix is widely used in data analysis, which can help identify the relationship between variables, assist feature selection, dimensionality reduction and multivariate analysis. By analyzing the correlation matrix, researchers can better understand the relationship between variables in the data set, so as to better understand the structure and characteristics of the data.

Random forest is an integrated learning method for regression and classification tasks. In the stochastic forest regression model, each tree is a regression tree, which is used to predict continuous variables. It combines the prediction results of several decision trees, and obtains the final prediction result by voting or averaging.

The training process of stochastic forest regression model is as follows:

(1) Randomly select bootstrap sample from the training set.

(2) For each self-help sample, randomly select a feature subset.

(3) Based on the selected feature subset, a regression tree is constructed by using decision tree algorithm.

(4) Repeat steps 1 to 3 times to construct multiple regression trees.

(5) For regression tasks, the final prediction result is usually the average of multiple trees.

The prediction of stochastic forest regression model can be expressed by the following formula:

$$\hat{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

Among them:

$(\hat{Y})$ is the predicted output of the model.

(N) is the number of trees in the random forest.

(Y_i) is the predicted output of the (i) tree.

The advantages of random forests include:

(1) It performs well for high-dimensional data and large-scale data sets.

(2) Be able to handle a large number of input features and evaluate the importance of features.

(3) Strong robustness to missing data.

However, random forests also have some limitations, including:

(1) For data with strong linear relationship, the performance may not be as good as that of linear regression model.

(2) Over-fitting may occur on some data sets, especially when the number of trees is large.

(3) Random forest models may encounter difficulties when dealing with sparse data.

XGBoost (Extreme Gradient Boosting) is a gradient lifting algorithm, which is used for regression and classification tasks. It is a powerful machine learning technology, which often obtains good performance in data science competitions and practical applications. XGBoost improves the prediction performance of the model by integrating multiple decision trees, and has regularization function to reduce the risk of over-fitting.

The following are the basic theories and formulas of XGBoost:

The objective function of XGBoost:

The goal of XGBoost is to minimize a loss function consisting of two parts, which respectively represent the sum of fit loss and regularization term. This objective function can be written as:

$$L(\theta) = \sum i = 1^n l(yi, \hat{y}i) + \sum k = 1^K \Omega(f_k)$$

Among them:

$(L(\theta))$ is the objective function.

(n) is the number of training samples.

(yi) is the real label of the (i) th sample.

$(\hat{y}i)$ is the model prediction value of the (i) th sample.

$(l(yi, \hat{y}i))$ is a fitting loss function, which is used to measure the deviation between the prediction of the model and the real label.

(K) is the number of trees.

$(\Omega(f_k))$ is a regularization term of the tree (k), which is used to control the complexity of the tree.

Fit Loss function (fit loss):

The fitting loss function usually adopts Mean Squared Error for regression task and Log Loss for binary classification task. For multi-classification tasks, multi-classification logarithmic loss can be used.

Regularization Term (regularization term):

XGBoost uses two regularization terms to control the complexity of the tree, namely L1 regularization of leaf node weight and L2 regularization of leaf node weight. These regularization terms help to prevent over-fitting.

Structure of the tree:

XGBoost uses CART (Classification and Regression Trees) as the basic learner, and it constructs a binary tree. Each leaf node corresponds to a segment, which represents the predicted output of a leaf node of the tree. The weight of each leaf node represents the predicted output of the leaf node.

Gradient lifting:

XGBoost uses gradient lifting method to train a series of trees iteratively. The construction of each tree will consider the prediction residual of the previous tree (that is, the difference between the real label and the prediction of the current model), so that the model can gradually approach the real label.

Importance of characteristics:

XGBoost can calculate the importance scores of features and help you understand which features play a key role in the prediction of the model. These scores can be used for feature selection and model interpretation.

In a word, XGBoost is a powerful ensemble learning algorithm, which combines multiple decision trees and improves the performance of the model by optimizing a comprehensive objective function. It performs well in many machine learning tasks, including classification, regression, ranking, recommendation system and so on.

Support Vector Machines (SVM) is a supervised learning algorithm for classification and regression. Its core idea is to find an optimal hyperplane to effectively separate heterogeneous data and maximize the interval of classification boundaries. The following are some basic theories and mathematical formulas of SVM:

(1) Support vectors: The goal of SVM is to find a hyperplane, which can divide data into two categories. Support vectors are the closest data points to this hyperplane, and they are very important to determine the position and direction of the hyperplane.

(2) Separating Hyperplane: SVM looks for a separating hyperplane, which is expressed by a linear equation: $\varpi^{Tx} + b = 0$, where (ϖ) is the normal vector, (x) is the data point and (b) is the intercept. This hyperplane divides data points into two categories, one of which is on the side of ($\varpi^{Tx} + b > 0$) and the other is on the side of ($\varpi^{Tx} + b < 0$).

(3) Maximum margin: The goal of SVM is to find a separated hyperplane, so as to maximize the distance between the two classes of support vectors and the hyperplane. This distance is called margin.

(4) Optimization problem: The goal of SVM can be expressed as a convex optimization problem, which is usually solved by Lagrange multiplier method. The objective of the optimization problem is to minimize the following loss functions:

$$[L(w,b,\alpha)=\frac{1}{2}||w||^2-\sum_{i=1}^{N}\alpha_i\big(y_i(w^Tx_i+b)-1\big)]$$

Where (n) is the number of data points, (α) is the Lagrange multiplier, and (y_i) is the label of data points.

(5) Dual Problem: By solving the dual problem, the weight of the support vector (Lagrange multiplier) can be obtained. The goal of dual problem is to maximize;

$$[W(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jy_iy_jx_i^Tx_j]$$

Where (α) is Lagrange multiplier, (x_i) and (x_j) are data points, and (y_i) and (y_j) are their labels.

(6) Kernel Function: SVM can use kernel function to deal with nonlinear problems. Common kernel functions include linear kernel, polynomial kernel and radial basis kernel (RBF kernel). Kernel function can map data from original space to higher dimensional feature space, so that nonlinear problems can also be separated by hyperplane.

Once the optimal Lagrangian multiplier (α) is found, the weight vector (ϖ) and intercept (b) can be calculated by them, and the classification hyperplane can be obtained.

This is the basic theory of SVM and some mathematical formulas. SVM is a powerful classification and regression algorithm, especially when dealing with high-dimensional data and nonlinear problems.

2.4 Analysis of experimental results

2.4.1 Data source

All the data in this experiment come from American health insurance data set, including 1338 lines of insured data, in which the insurance cost is given according to the following attributes of the insured: age, gender, body mass index, number of children, smokers and region. There are no missing or undefined values in the dataset. This relatively simple data set should be an excellent starting point for EDA, statistical analysis and hypothesis testing, and training linear regression models for forecasting insurance premiums.

2.4.2 Variable explanation

Age and gender are the actual age and gender of the insured. Body mass index (BMI), referred to as body mass index, is a commonly used international standard to measure obesity and health. The formula is: body mass index = weight ÷ heigh^2. (Weight unit: kg; Height unit: meter.) Body mass index was first put forward by Belgian generalist Lambert Adolf Jacques kettler in the mid-19th century. The number of children owned refers to the post-algebra of the insured. At the same time, whether smoking or not and where the insured comes from are also variables.

2.4.3 Variable processing

For text variables, we adopt digital symbolization processing to turn non-digital indicators into numbers. For example, for gender variables, we use 1 for males and 0 for females. Among the smoking habit variables, smokers are 1 and non-smokers are 0. Analogously, the regional variables are coded as numbers: 0, 1, 2, ... 0 in Northeast China; Northwest China is 1; Southeast region is 2; The southwest region is 3. The remaining numeric variables retain their numeric attributes.

2.4.4 Experimental Process

As shown in Fig. 1, the correlation matrix shows the correlation between all variables (except the premium). In the range of 0 to 1, the stronger the correlation between variables, the closer the value is to 1, the weaker the correlation is, and the closer the value is to 0. The purpose of our move is to eliminate similar variables. According to the image, the correlation between variables is weak, and they can all represent independent categories.

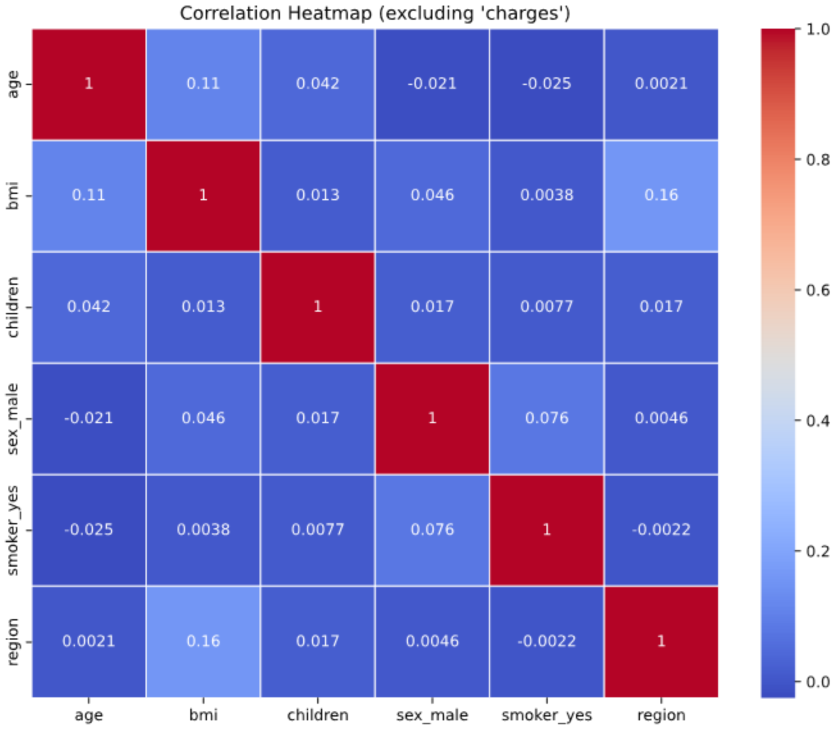

Figure 1: Correlation_Heatmap (excluding 'charges')

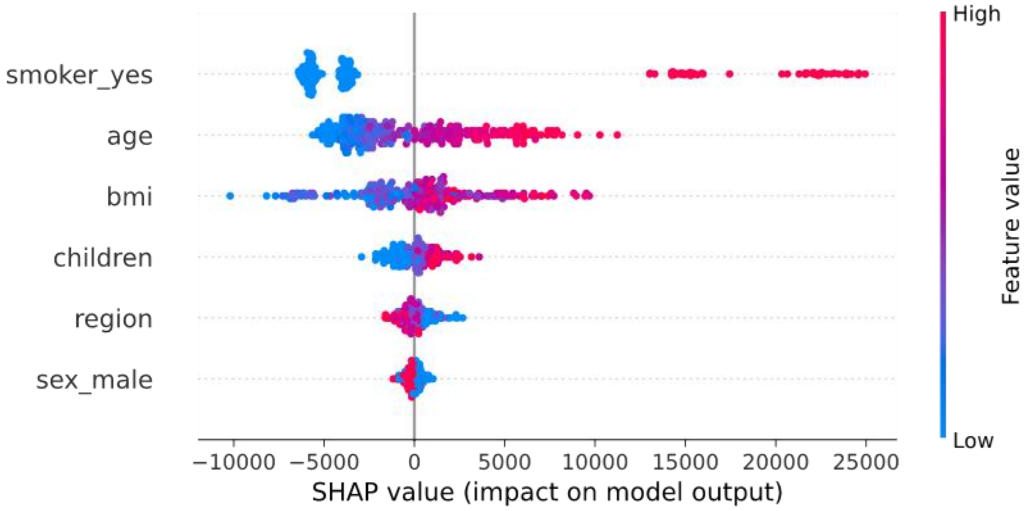

Figure 2: Using XGB model to simulate the correlation between variables and premiums.

As shown in fig. 2, the X-axis of the image represents SHAP value and the Y-axis represents variable indicators. The model ranks the correlation between variables and premiums. The stronger the correlation, the greater the absolute value of SHAP valve, which is displayed in red; the weaker the correlation, the smaller the absolute value of SHAP Value, which is displayed in blue. It can be seen that smoking, age and BMI index have great influence on the premium, while the remaining variables have little influence.

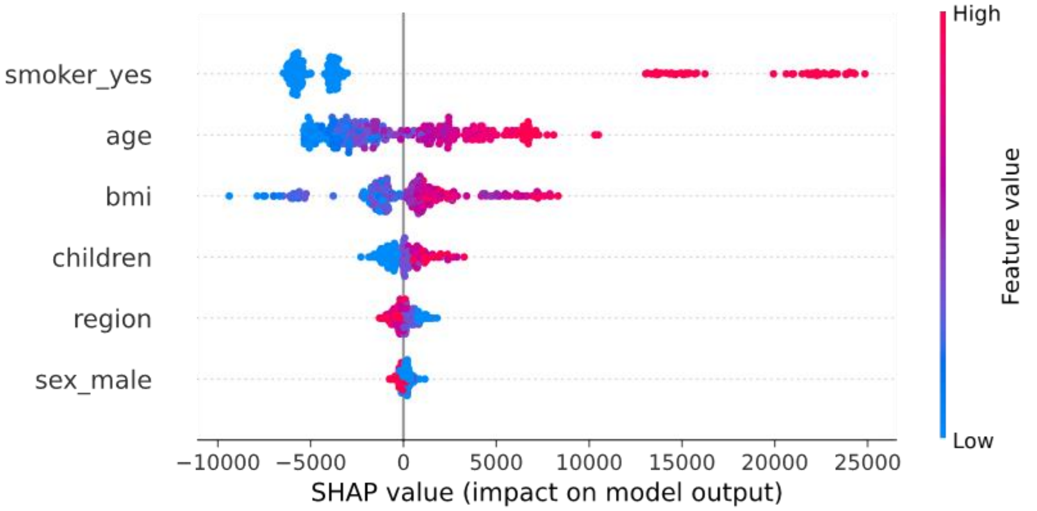

Figure 3: Using random forest model to simulate the correlation between variables and premiums.

As shown in Figure. 3, similar to XGB model, the stochastic forest model also sorts the correlation between variables and premiums, with the X-axis representing SHAP value and the Y-axis representing variable indicators. The stronger the correlation, the greater the absolute value of the SHAP value, which is displayed in red at the top; conversely, the weaker the correlation, the smaller the absolute value of the SHAP value, which is displayed in blue at the bottom. It can be seen that the results simulated by XGB model are consistent: smoking, age and BMI index have great influence on the premium, while the remaining variables have little influence.

By comparing the three models (Table 1), we find that the random forest has better fitting and is more suitable for us to study the relationship between premium and influencing factors.

Table 1: Comparison of fitting degree of models

Model	MAE
XGBoost	2789.5820
Random Forest	2535.7409
Support Vector Machines	6446.2811

3. Conclusions

This paper is to study the relationship between various variables and medical insurance premiums. Firstly, age, gender, body mass index (BMI), number of children, whether or not the insured person smokes and where they come from correspond to the independent variable x, and medical insurance premiums correspond to the dependent variable y. We then employed three models XGB, Random Forest and SVM to simulate the correlation between variables and premiums. The end result is consistent: smoking, age and BMI index have great influence on the premium, while the remaining variables have little influence. Among them, random forest model simulation is the most accurate. Because of the stronger the correlation, the greater the absolute value of SHAP valve, so we get that smoking has the biggest impact on health insurance premiums. Although this study has made some valuable findings, there are still some aspects that need further research and improvement: Firstly, broader data sample: The sample in this study has certain limitations, and future studies may consider using a broader data sample to more comprehensively understand the impact of different variables on medical insurance. Secondly, studies of long-term effects: The time span of the study is limited, and future studies can consider long-term effects to better understand how different variables affect health insurance demand over time. In this study, we also have to be honest to admit some shortcomings: Firstly, Possible data bias: There may be some bias in the data used in the experiment, because the data collection method and sample selection may lead to incomplete data accuracy. Secondly, Experimental limitations: Experimental methods were used in this study, but experiments may not capture all the complex factors in real life, which may also lead to certain limitations. In conclusion, future research needs to be further explored and improved to more fully understand this topic.

References

[1] Aggarwal, S., & Anmol. (2022). Health Insurance Amount Prediction Using Supervised Learning. 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), Technological Advancements in Computational Sciences (ICTACS), 578–581.

[2] Baro, E. F., Oliveira, L. S., & de Souza Britto Junior, A. (2022). Predicting Hospitalization from Health Insurance Data. 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Systems, Man, and Cybernetics (SMC), 2790–2795.

[3] Qin, J., Tao, Z., Huang, S., & Gupta, G. (2021, March). Stock price forecast based on ARIMA model and BP neural network model. In 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) (pp. 426-430). IEEE.

[4] Kaushik, K., Bhardwaj, A., Dwivedi, A. D., & Singh, R. (2022). Article Machine Learning-Based Regression Framework to Predict Health Insurance Premiums. International Journal of Environmental Research and Public Health, 19(13).

[5] Kreif, N., DiazOrdaz, K., Moreno-Serra, R., Mirelman, A., Hidayat, T., & Suhrcke, M. (2021). Estimating heterogeneous policy impacts using causal machine learning: a case study of health insurance reform in Indonesia. Health Services and Outcomes Research Methodology, 1-36.

[6] K, V., S, S., Deepa, D., Antony, J. C., Karpura Dheepan, G. M., Dharma Siva Pavan, A. S., & PagadalaHariprasad. (2023). Predicting health insurance claim frauds using supervised machine learning technique. 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Science Technology Engineering and Mathematics (ICONSTEM, 1–7.

[7] Nalluri, V., Chang, J.-R., Chen, L.-S., & Chen, J.-C. (2023). Building prediction models and discovering important factors of health insurance fraud using machine learning methods. Journal of Ambient Intelligence and Humanized Computing, 14(7), 9607-9619.

[8] Reinke, C., Doblhammer, G., Schmid, M., & Welchowski, T. (2023). Dementia risk predictions from German claims data using methods of machine learning. Alzheimer’s & Dementia :19(2), 477–486.

[9] Saraswat, B. K., Singhal, A., Agarwal, S., & Singh, A. (2023). Insurance Claim Analysis Using Traditional Machine Learning Algorithms. 2023 International Conference on Disruptive Technologies (ICDT), Disruptive Technologies (ICDT), 623–628.

[10] Hu, Y., Tao, Z., Xing, D., Pan, Z., Zhao, J., & Chen, X. (2020, August). Research on stock returns forecast of the four major banks based on ARMA and GARCH model. In Journal of Physics: Conference Series (Vol. 1616, No. 1, p. 012075). IOP Publishing.