2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. Oct 13-14, 2022

2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) | 978-1-6654-7433-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICRITO56286.2022.9964649

Interpretation of machine learning models using XAI - A study on health insurance dataset

Abhyudaya Bora
School of Computing
DIT University
Dehradun, India
boraabhyudaya@gmail.com

Ritika Sah
School of Computing
DIT University
Dehradun, India
sahritika17@gmail.com

Alabhya Singh
School of Computing
DIT University
Dehradun, India
singhalabhya32@gmail.com

Deepak Sharma
School of Computing
DIT University
Dehradun, India
d.sharmaxxx05@gmail.com

Ranjeet Kumar Ranjan
School of Computing
DIT University
Dehradun, India
ranjeetghitm@gmail.com

***Abstract*—This study aims to dive into the complex Machine Learning algorithms using XAI and provide explanations for outcomes that we received from these algorithms. In this paper, we have proposed to predict the cost of health insurance. The proposed work is composed of two Machine Learning algorithms, namely Multiple Linear Regression and Random Forest, followed by an explanation of predicted results using XAI. Here, we first provide a simple explanation with the help of model-specific approaches based on Microsoft's InterpretML library. Then the predicted insurance premium cost is further explained by model-agnostic techniques, LIME, and SHAP. The significance of the study is to provide a better user experience and build trust between the user and the machine. These techniques can help to check the correctness of the prediction models, as the domain experts can analyze the features that affect the outcome the most and provide their expertise.**

Keywords— Explainable Artificial Intelligence (XAI), Interpretable Machine Learning, LIME, SHAP, Health Insurance Price Prediction.

I. Introduction

For everyone in the industry, the year 2020 has proven to be a challenge. During COVID-19, the data science sector emerged to be the most successful, active, and effective. One of the biggest sectors that data science and machine learning would leverage is the health sector, retails, and finance. It has been found that the use of many manual processes in medicine is excessive. For technology to help improve the current state of healthcare, the electronically generated information provided to healthcare workers needs to be enhanced by the power of Artificial Intelligence and Data Science [1]. Through research, the medical industry can determine whether the treatments are safe and effective. It is a long process of trial, error, and evidence-based decisions. When it comes to machine learning, we must follow the same method to ensure that it is both safe and successful. We need to fully understand the ethics involved in transferring part of what we do to the machine. The advancement in healthcare systems and the high risk of getting various diseases have also posed challenges to people around the world. This leads to the high demand for medical insurance. The accurate prediction of insurance premiums can solve lots of challenges for the insurance company as well as customers. In the last several years, machine learning (ML) has been employed in the field of insurance [2]. The major issue of the machine-based prediction system is that the prediction is a black-box process. Where reason behind the prediction is unknown [3]. If the insurance company and the customers know the reasons behind high or low premium values, then a strong trust can be developed between customers and the company. In general, ML models and their predictions are difficult to interpret, this limits their practical relevance, credibility, and reliability in the healthcare premium price prediction system. Therefore, there is a need for approaches to elucidate the ML models.

In recent years, Explainable Artificial Intelligence (XAI) has appeared as an option for explaining the reasons behind the predictions [4]. This paper proposes an insurance price prediction using machine learning techniques and explained the results using various XAI models. The proposed work aims to predict the cost of insurance based on different attributes and explain the predicted premium price for a customer. Generally, various factors influence the premium of insurance like Cost & Utilization, Clinical disease, etc. An insurance policy reduces or eliminates the expenses of losses incurred by various risks. These factors can influence the creation of insurance policies. The proposed XAl-enabled machine learning models will help to understand the reasons behind the prediction to build a strong trust between customers and an organization. Explanations based on the clinician's knowledge can be analyzed and will help us validate the predictions made by the model.

II. Literature Review

First, Artificial Intelligence (AI) is a technology that can predict the development of health problems in users by capturing and analyzing their health data. And it can be integrated and deployed in the healthcare industry. Explainable AI (XAI) is a field in which methodologies are used to provide ethical justifications for all predictions made by ML models. Although research into interpretable and explainable AI is rapidly increasing, there is a lack of a comprehensive evaluation and systematic classification of these works. But there are few review papers in this field. According to Shad et al. [5], different interpretability techniques are used. The goal is to enlighten practitioners on how to comprehend and interpret explainable AI systems using a range of methodologies. Using several Neural Network Models, to predict the various stages of Alzheimer's in patients using MRI imaging data. One of the XAI

978-1-6654-7433-7/22/$31.00 ©2022 IEEE

frameworks i.e., LIME, is used for T1 weighted MRI scans from the Kaggle dataset which tells the exact classification stage of the patients. In a similar paper presented in [6], the authors have used another model-agnostic explainability approach called Doctor XAI. The proposed technique deals with multi-labelled, subsequent, ontology-linked data. In this clinical history, the patient is taken as input to predict the next visit. They also illustrate how combining the temporal dimension of the data with the domain knowledge represented in the medical ontology improves the quality of the explanations. Similarly, a review was proposed on Explainable artificial intelligence models using real-world electronic health record data, which tells deep learning is the most used ML algorithm for the prediction of various diseases like cancer and diabetes. We also discovered that XAI necessitates a greater focus on medical applications. It also discusses the opportunity and challenges faced in XAI. Finally, Adadi et al. [7] proposed a general overview of the topic. They presented a survey according to which AI-based systems are often not transparent and lack accountability. For building trust between users, it requires a natural system of black boxes that enables strong predictions and trust. Researchers use various machine learning algorithms to analyze medical insurance data. Several papers have discussed the prediction of health insurance. In [8], the author uses a Computational Intelligence Approach to demonstrate how different regression models can predict insurance prices and to compare the models' accuracy, various machine learning methods are applied. The results show that the Stochastic Gradient Boosting (SGB) model performs better than the other models. In the proposed research in [9], National Health Insurance Service-Health Screening datasets are used for cardiovascular disease prediction. After applying various ML models, results showed that the gradient boosting, and random forest algorithms had the best average prediction accuracy.

In [10], an ensemble method of supervised learning is used for medical insurance cost prediction. For feature selection, Pearson coefficient and Boruta algorithm are used. A boosting mode is created based on the regression tree and stochastic gradient descent. Boosting and stacking ensembles showed better accuracy than bagging. To combine the predictions random forest algorithm is used. Several publications have been published previously on the ML techniques used for auto claim prediction. Frempong [11] created a predictive model for auto insurance claims, CART, Entropy, and Decision Tree are the methods used in this model, around 1500 Ghana insurance data, and the age of vehicle and customers predictor variables were used, it gives the maximum claim to the policyholders from age 18 to 48 and vehicle's age is 0 to 8 years. A system was proposed in [12] to predict customers' claims in an auto insurance company. The techniques used in this model were Logistics regression, Artificial Neural network, and Decision Tree, and the result for Insurance claims are categorized as low, high, or fair. Also, the neural network has the best prediction accuracy for claims. In this research paper [13] we see how the Machine learning model can be applied to insurance using big data. On comparing various model performances such as naïve Bayes, decision trees, XGBoost, random forest, logistic regression, and KNN. The results showed that RF is better than other methods. Although in [14], the author does not cover the auto claim prediction but predicts the car insurance policies using random forest.

III. Methodology Overview

A. Machine Learning Models for Insurance Prediction

In this paper, we have used two regression models, multi-linear regression and random forest.

1) Multi-Linear Regression

Regression models are used to predict the relationship between dependent variables and independent variables. One of the prevalent examples is predicting the price of a house. It's a supervised learning technique. A regression model is an approach to finding the relationships between predictors or independent variables and response or dependent variables as defined in equation (1).

$$y = a + bx_1 + cx_2 + dx_1x_2 + e(x_1)^2 + \cdots \qquad (1)$$

where x_1 and x_2 are independent variables and y is the dependent variable, and $a, b, c, d, and\ e$ are the parameters.

In this paper, we have used multiple linear regression, which often has more than one predictor. For example, in this data, we analyzed independent variables like BMI, age, gender, smoker, etc. The dependent variable in this work is selected as insurance price.

2) Random Forest

Random forest is a bagging (also called bootstrap aggregation) ensemble technique. It works on both classification and regression models. Random forest builds tons of different decision trees on the trained dataset, as many decision trees provide high accuracy, better generalization, more stability, and prevent overfitting problems. The final prediction is done by majority voting or called aggregation for classification. In the case of regression, the final prediction mean of all output is taken. For example, we have a training dataset that contains d records and n columns. We will take a random sample of rows and features from the dataset and give them to each decision tree. This is defined as bootstrapping. The output is now generated by each decision tree. The result is estimated based on a majority of votes or average. This part is called aggregation. In random forest, we also use hyperparameters to make the model faster and improve performance. There are many decision trees combined, and the prediction is made by averaging the predictions of every component tree. The random forest model has much better predictive accuracy than a single decision tree, in general, and works well with default parameters.

B. Explainable AI (XAI) Description:

As the 4th industrial revolution is dawning upon us, we are still witnessing the growth and occupancy of AI in almost every field. This is making us shift towards a more algorithmic society. However, even with such unprecedented growth, we face one impediment the lack of transparency. These algorithms allow powerful predictions, but these predictions cannot be explained directly due to their black-box nature.

This issue has raised a debate about the need for explainable AI [15]. Explainable AI, also interchangeably known as Interpretable AI, is the field of study that aims at demystifying the Black-box algorithms, helping us imply responsible AI, as it can produce transparent models. This goal needs to be achieved without affecting accuracy.

In 2004, Van Lent et al coined the term for the first time [16], the term hinting at the movement and initiatives made toward transparency and trust concerns of AI. As per DARPA [17], this field aims to produce more explainable models, which will be helpful in reducing human errors as well as the cost of treatment and can correctly reflect the process that can generate the output.

Many studies have been conducted on the need and application opportunities of XAI [7], [15]. Here we will be discussing some of the different techniques used in this field to generate explanations quickly. We will then implement these techniques to regression Models (Multilinear regression and Random Forest Regressor) to see them in action. In this paper we have applied the following XAI techniques to interpret the prediction made by regression algorithms:

- *Microsoft's InterpretML*: InterpretML is an open-source python toolkit developed by Sameki et al. [18]. The focus of the toolkit is to explain black-box AI systems for a better and easier understanding of the behaviour of the systems. It provides both global and local explanations. The global explanation helps users to understand overall behaviour whereas the local explanation provides a better understanding of individual results or predictions.

- *Local interpretable model-agnostic explanations (LIME):* LIME is a visualization technique that helps in explaining individual predictions. The "Model-agnostic" phase in the name hints at the fact that it can be applied to any supervised regression or classification model. LIME was introduced in August 2016 [19]. LIME is based on the inputs and outputs of the model and can work for any black-box model and type of data (image, text, etc.). In LIME, we fit an interpretable model to the local area where the prediction is made. For LIME the only condition is that the model is locally faithful, even if it does not make sense globally. When a prediction needs to be explained, we zoom into the local area where the point under consideration is in the overall model and then we fit a surrogate model, generally an interpretable Whitebox model. This way we can easily explain what is happening in that local area.

- *Shapley Additive Explanations (SHAP):* The idea behind Shapely' Additive explanation (SHAP) comes from the corporative game theory proposed by Lord Shapely in 1953. This theory is used to find the contribution of different members in a team to the success of the team. A detailed explanation of the corporative game theory can be found in (Elkind Edith and Rothe, 2016) [20]. The way it is implemented in explaining the black-box models is rather interesting. Each feature in our input data is treated as a team member, and then each feature's contribution to the prediction is evaluated. Shapley values, based on a coalitional game theory method, inform us how much value every feature carries for the prediction.

In general, the approach requires re-training a model using a set of features S ⊆ F, where F is the feature set of a dataset. An importance value is assigned to each feature which measures the effect on the model prediction using the feature. Different combinations get executed by removing a feature and combining it with other features, in order to check how it affects the prediction and the variation that occurs in the predicted value due to that feature. Now you can see that a lot of combinations while be formed and a model for all these combinations must be fitted and compared. Therefore, this technique requires a lot of computational power and takes comparatively more time to compile.

SHAP provides a very valid global explanation by aggregating the individual predictions and the local explanation for the given point. In fact, SHAP is great for global explanations. SHAP has both, Model-Agnostic and Model-Specific, approaches [21]. Here we have used the model-agnostic (kernel SHAP)

IV. Proposed Work

In this paper, machine learning based health insurance price prediction has been proposed. The proposed approach employed two regression models, namely multi-linear regression and random forest. The limitation of these model is the complication in interpretation of predicted results since these models behave as black box. The black box concept in machine learning is unable to explain why a model made a particular decision. The black box has remained a source of concern as it reduces clinical functionality for model prediction, which reduces its utility to clinicians. To deal with this issue this study incorporates XAI models, combined with clinical knowledge, to get greater benefits from ML based systems. XAI has the potential to solve a variety of world problems. XAI does not help in better prediction but also prediction can be easily understood by the users. As described in Fig. 1, the proposed works has been carried out as follows:

- First, a dataset has been created from various resources, followed by the training of the machine learning models for prediction of insurance cost.

- The predictions by the trained models, along with the insurance dataset, is further used by XAI methods like LIME and SHAP in the black box to generate the explanations.

- With the use of clinical knowledge, these explanations are analyzed. The analysis verifies the models' viability and provides accountability and transparency.

- If the predictions are right, explanations are used to create useful insights and recommendations and then deploy the model. If the predictions are incorrect, then re-training or improvements of the prediction models can be performed

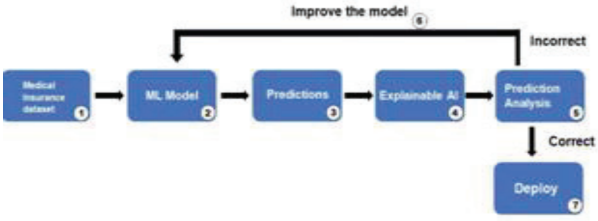


Fig. 1 A block diagram explaining the approach to the proposed work.

V. DataSet

We have used the dataset from Kaggle [22]. The dataset includes 1338 records and 7 attributes. The attributes are age, sex, BMI, children, smoker, region, and charges. While developing this model, data is divided into two sets, namely, test and train, so that the prediction of new information does not require and is dependent on previous record's memory. The training data set makes up about 80% of the total data, and the remaining 20% is for testing the model to evaluate its accuracy with help of various algorithms.

VI. Data Processing

The process of getting the data ready for our machine learning model is called data Pre-processing. Here we have to check and clean the data for any faulty or missing values and convert the string data to numerical form for our model to understand it. There are seven variables, out of which six are input variables or independent variables and one name, charges, is the output or the dependent variable. Now we run the info function to get a general overview of the dataset. We come to know we have 1338 entries/rows, indexed from 0 to 1337. as for the data type, we have two float variables, two integer variables, and three object variables. The dataset has no missing values. A statistical examination has been performed for the target variable and the result is presented in Table 1.

TABLE I. Statistic of Target Variable(charges).

Count	**Min**	**Max**	**STD**	**Mean**	**Median**	**3^{rd} Quad**
1338	1122	63770	12110	13270	9382	16640

From the analysis presented in Table 1, we can state that insurance charges are right-skewed, as we have a mean value greater than the median. The same has also been represented by visualizing the data through a histogram as shown in Fig. 2.

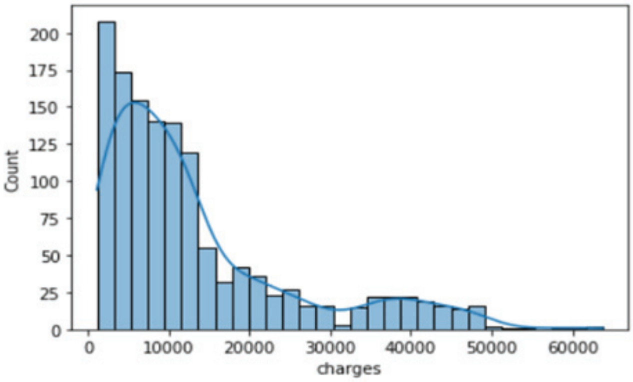


Fig. 2 Histogram for the dependent variable(charges) showing the skewed data distribution to the right

This graph represents the fluctuation in insurance charges in our data, which shows that the data is right-skewed. In regression, we can use the log transformation to normalize our highly right-skewed data set.

We then convert the data types of the features into the data types that our model can understand. We convert the categorical data into numeric or binary values. Here is the table showing the categorical data in the data set, and the numerical values that we have assigned to them. Now we divide the data set into x, the input variable, and y, the target variable. We then divide x and y into training and testing data sets in a 4:1 ratio (80% train and 20% test). The model is trained using the training data set, and the accuracy is checked using the test dataset.

VII. Experimental Setup

Hardware Requirements- A multi-core processor of i5 or higher or a processor that is similar, Minimum 8GB RAM, disc space (SSD preferable), Output display of minimum 1280x720, Nvidia Cuda support, Data connectivity up to 400 kbps of speed.

Software Requirements- OS (Linux/windows) Linux (Ubuntu) preferred, Python programming language and libraries like pandas, numpy, matplotlib, seaborn, interpret ML, Jupyter notebook/ Google Collab notebook/Visual Studio code.

VIII. Implementation and Result

In this study, we have implemented two supervised machine learning models to predict the cost of the premium for a customer. As described earlier, the main objective of the study is to explain, using XAI, that how the models have provided the price of insurance. To begin with, we implemented a quick and easy library, InterpretML developed by Microsoft. This library helped in implementing the model-specific explanation for the linear model. We were able to get the local, as well as the global explanation for the model and visualize the result as well. It provides a ready-made interactive graph with a drop-down menu. The limitation of this model is that it is model-specific, and the library has support for only three glass-box models. This library is very efficient in explaining simple models, with very good visualization. A global explanation graph using InterpretML is shown in Fig. 3.

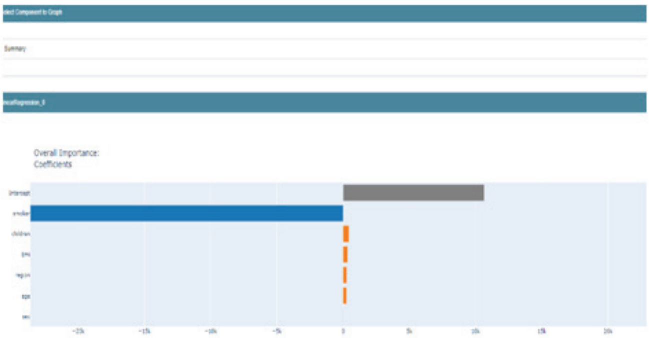


Fig. 3 Data Summary Graph produced by InterpretML, showcasing the importance of a feature globally.

Since it was a simple regression model, the accuracy of the model was very less, so we decided to go along with another, more complex model. We implemented the Random Forest Regressor as the second model. This helped us get better accuracy, but the model became a black box. To understand what was going on inside the learning models, we implemented LIME and SHAP. We used LIME to get a local explanation for a particular data point under consideration and SHAP to provide a global/overall explanation and influence on the prediction. For this, we have applied the lime package of python available publicly. Then we made a lime tabular explainer object for a regression model and passed the instance we want to predict, to the explain instance function. This provides us with graphical outcomes that help to study the features and their contribution to the prediction visually. We created this graph for random values from the data set. Table 2 presents the feature value assigned for the implementation of the LIME approach. Each Feature name has been assigned a number in the model. Table 3 maps each feature with its respective number.

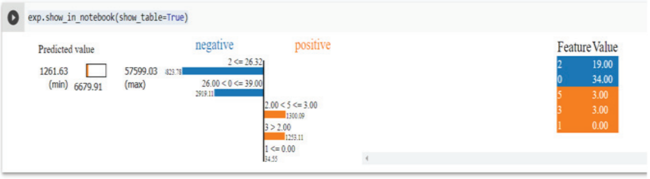
From the graph in Fig. 4, we can observe that for this client, the premium cost is affected the most by the age of the client and the BMI. These features are influencing the predicted price in a negative trend, i.e., they are lowering the price for this client. Region and Children also have some influence over the result, whereas the sex of the client does not interfere with the prediction at all. The predicted value that we get is 6,679.91, whereas without the support of the features (age and BMI) the cost could have gone up to 57,599 and if all the features showed a negative trend, the price could have gone down to 1,261.63.

TABLE II. FEATURE VALUES OF A RANDOM DATA POINT FROM THE DATA FOR THE STUDY.

Age	Sex	BMI	Children	Smoker	Region
34	0	19	3	1	3

TABLE III. NUMERICAL VALUES THAT ARE ASSIGNED TO THE FEATURES BY LIME.

Feature	Assigned Number
Age	0
Sex	1
BMI	2
Children	3
Smoker	4
Region	5



```
exp.show_in_notebook(show_table=True)
```

Feature	Value
2	19.00
0	34.00
5	3.00
3	3.00
1	0.00

Fig. 4LIME output chart for the point under consideration

To get the global explanation we used SHAP. We used the python SHAP package. Since our model is Random Forest Regressor, we used the Tree SHAP, which is a highly efficient algorithm to compute SHAP values for tree-based models. We again built an explainer (shap.TreeExplainer), and passed the entire test data set to it, to get the shapely values. We can also use SHAP to compute the local points, though it is preferred for global explanations. In this study, we have used SHAP for the global explanation only.

We can clearly understand from Fig. 5, that whether a person is a smoker or not has a significant effect on the premium cost. The age and BMI of the person are the second most powerful features that help us come up with the predicted cost. Children and regions have very less power, whereas the sex of a person has almost no effect. Even if we were to remove the sex feature, the predictions won't show a major fluctuation. This global explanation also helps in recognizing the most valuable attributes. Here is the bar chart that represents the features that have more predictive power in descending order.

Fig. 6 shows all the features and the direction in which they moved the prediction for all the points in the dataset. The red dots represent the strong prediction stand (higher cost) for a particular prediction, and blue shows the weak stand (lower cost). From this figure, we also see how drastically smoking changes the prediction. Although a smoker may not be charged a very high price (as age is also affecting the price to the same extent.), But a non-smoker is charged significantly less for the premium. The chart provides information like:

- Feature importance: Variables are ranked as per their predictive power
- Impact: The horizontal location tells us if the value is associated with a higher or lower prediction.
- Original value: the color of the door shows us whether the value is high or low for that observation
- Correlation: from the chart, we can see that if a person is a smoker, then the premium cost for that person is usually high. Here we can say that smoking is negatively correlated two the premium cost

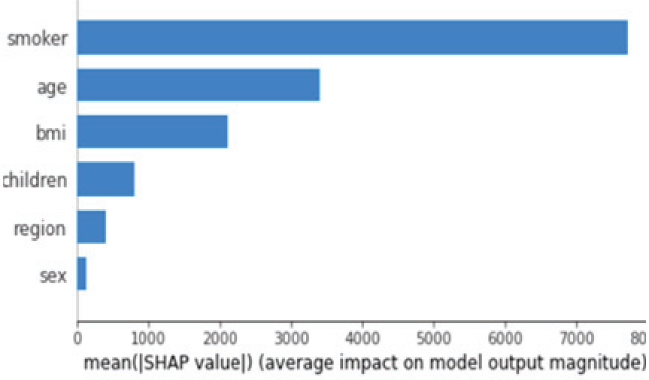


Fig. 5 Horizontal bar chart, representing the contribution of a feature globally, by using the shapely values of the whole data set

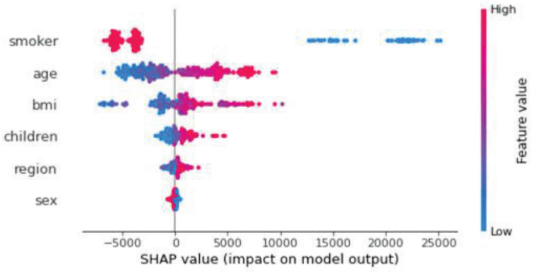


Fig. 6 Impact plot showing each value of the data set on the graph.

IX. CONCLUSION

In this paper, a study on health insurance price prediction systems based on machine learning and XAI has been performed. To predict price, multiple regression and a random forest regression model have been implemented. The predicted results have been explained using the model-specific approaches based on Microsoft's InterpretML and two model-agnostic techniques, namely LIME and SHAP have been used. The proposed study shows that using XAI, the black box prediction by traditional machine learning models can be explained inefficiently ways using graphical visualization. The interpreted results make a prediction more reliable and transparent.

REFERENCES

[1] S. N. Payrovnaziri *et al.*, “Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review,” *Journal of the American Medical Informatics Association*, vol. 27, no. 7. Oxford University Press, pp. 1173–1185, Jul. 01, 2020, doi: 10.1093/jamia/ocaa053.

[2] M. K. Severino and Y. Peng, “Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata,” *Mach. Learn. with Appl.*, vol. 5, p. 100074, Sep. 2021, doi: 10.1016/j.mlwa.2021.100074.

[3] G. Yang, Q. Ye, and J. Xia, “Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond,” *Inf. Fusion*, vol. 77, pp. 29–52, Jan. 2022, doi: 10.1016/j.inffus.2021.07.016.

[4] A. M. Antoniadi *et al.*, “Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review,” *Appl. Sci.*, vol. 11, no. 11, Jun. 2021, doi: 10.3390/app11115088.

[5] H. A. Shad *et al.*, “Exploring Alzheimer’s Disease Prediction with XAI in various Neural Network Models,” in *TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON)*, 2021, pp. 720–725, doi: 10.1109/TENCON54134.2021.9707468.

[6] C. Panigutti, A. Perotti, and D. Pedreschi, “Doctor XAI An ontology-based approach to black-box sequential data classification explanations,” in *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Jan. 2020, pp. 629–639, doi: 10.1145/3351095.3372855.

[7] A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” *IEEE Access*, vol. 6, pp. 52138–52160, Sep. 2018, doi: 10.1109/ACCESS.2018.2870052.

[8] C. A. ul Hassan, J. Iqbal, S. Hussain, H. AlSalman, M. A. A. Mosleh, and S. Sajid Ullah, “A Computational Intelligence Approach for Predicting Medical Insurance Cost,” *Math. Probl. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/1162553.

[9] J. O. Kim, Y. S. Jeong, J. H. Kim, J. W. Lee, D. Park, and H. S. Kim, “Machine learning-based cardiovascular disease prediction model: A cohort study on the korean national health insurance service health screening database,” *Diagnostics*, vol. 11, no. 6, Jun. 2021, doi: 10.3390/diagnostics11060943.

[10] N. Shakhovska, N. Melnykova, V. Chopiyak, and M. Gregus Ml, “An ensemble methods for medical insurance costs prediction task,” *Comput. Mater. Contin.*, vol. 70, no. 2, pp. 3969–3984, 2022, doi: 10.32604/cmc.2022.019882.

[11] N. K. Frempong, N. Nicholas, and M. A. Boateng, “Decision Tree as a Predictive Modeling Tool for Auto Insurance Claims,” *Int. J. Stat. Appl.*, vol. 7, no. 2, pp. 117–120, 2017, doi: 10.5923/j.statistics.20170702.07.

[12] K. P. M. L. P. Weerasinghe and M. C. Wijegunasekara, “A Comparative Study of Data Mining Algorithms in the Prediction of Auto Insurance Claims,” 2016. [Online]. Available: www.eijst.org.uk.

[13] M. Hanafy and R. Ming, “Machine learning approaches for auto insurance big data,” *Risks*, vol. 9, no. 2, pp. 1–23, 2021, doi: 10.3390/risks9020042.

[14] A. S. Alshamsi, “Predicting car insurance policies using random forest,” in *2014 10th International Conference on Innovations in Information Technology (IIT)*, 2014, pp. 128–132, doi: 10.1109/INNOVATIONS.2014.6987575.

[15] J. Gerlings, A. Shollo, and I. Constantiou, “Reviewing the need for explainable artificial intelligence (XAI),” in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2021, vol. 2020-January, pp. 1284–1293, doi: 10.24251/hicss.2021.156.

[16] M. Van Lent, W. Fisher, and M. Mancuso, “An Explainable Artificial Intelligence System for Small-unit Tactical Behavior,” 2004. [Online]. Available: www.aaai.org.

[17] M. Turek, “Explainable Artificial Intelligence,” *Defense Advanced Research Projects Agency*, 2019. https://www.darpa.mil/program/explainable-artificial-intelligence (accessed Sep. 14, 2022).

[18] S. Bird, M. Sameki, and K. Walker, “InterpretML : A toolkit for understanding machine learning models *,” *Microsoft*, 2020. https://www.microsoft.com/en-us/research/uploads/prod/2020/05/InterpretML-Whitepaper.pdf (accessed Sep. 14, 2022).

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should i trust you?’ Explaining the predictions of any classifier,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-August-2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[20] J. Elkind Edithand Rothe, “Cooperative Game Theory,” in *Economics and Computation: An Introduction to Algorithmic Game Theory, Computational Social Choice, and Fair Division*, J. Rothe, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, pp. 135–193.

[21] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, May 2017, vol. 2017-Decem, pp. 4766–4775, [Online]. Available: http://arxiv.org/abs/1705.07874.

[22] Miri Choi, “Medical Cost Personal Datasets,” *kaggle.com*, 2018. https://www.kaggle.com/datasets/mirichoi0218/insurance (accessed Sep. 14, 2022).