Financial Engineering and Risk Management (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/ferm.2024.070604
ISSN 2523-2576 Vol. 7 Num. 6

The Application of Artificial Intelligence in Mortality Modeling and Forecasting: GBM, Data Cleaning, and Dynamic Mortality Tables

Jiaming Zuo

Everbright Actuarial Consulting Limited, Hong Kong, China
jzuo@ebactuary.com

Keywords: AI algorithms, mortality rate modeling, mortality risk, Gradient Boosting Machine, Random Forest

Abstract: The application of AI in mortality rate modeling and prediction provides actuaries with powerful tools to more effectively identify, measure, and manage mortality risk. However, the application of this technology also brings challenges, including ensuring high-quality data, addressing model bias, complying with evolving regulatory requirements, and improving the transparency and interpretability of models. The role of actuaries is constantly evolving with the advancement of technology and the industry, requiring continuous learning of new technologies and methods to adapt to these changes. AI technologies, such as Gradient Boosting Machine (GBM) and Random Forest, are used to develop predictive models that delve into historical data and accurately predict mortality rates. AI algorithms play a key role in feature engineering, extracting features that have a significant impact on mortality outcomes, such as age, pre-existing conditions, lifestyle, and medical history. AI can also achieve personalized predictions by analyzing individual differences and segmenting populations into subgroups with similar characteristics, thereby improving model accuracy. Non-traditional data sources, such as wearable devices, electronic health records, social media, and environmental data, provide new dimensions for mortality rate prediction models, but also pose challenges in terms of privacy protection, quality control, and data source integration.

1. Gradient Boosting Machine (GBM)

GBM is a popular machine learning technique known for its effectiveness in mortality modeling and forecasting within AI applications. GBM operates by iteratively combining weak predictive models to create a strong ensemble model. In the specific context of mortality modeling, GBM can be trained on a diverse range of historical data encompassing mortality rates, demographic information, medical records, lifestyle factors, and other relevant variables to accurately predict future mortality rates.

GBM operates by iteratively adding decision trees to correct residuals, enhancing the model's predictive accuracy. The process initiates with an initial model, often a simple prediction like the mean of the target variable. A loss function is then defined to quantify the discrepancy between predicted and actual values, commonly using mean squared error (MSE) for regression tasks. Subsequently, the algorithm iterates by creating new decision trees to predict residuals, which represent the differences between observed and predicted mortality rates. A learning rate is applied to regulate the impact of each new tree on the model, with a smaller learning rate promoting better generalization. For each iteration, a decision tree is built to fit the residuals from the previous iteration, with parameters like depth and number of splits optimized to minimize the loss function. The new tree's predictions are then added to the previous iteration's predictions, adjusted by the learning rate, to update the model. The process continues until a stopping criterion is met, such as a maximum number of iterations or when the improvement in the loss function falls below a certain threshold. The final model comprises an ensemble of all trees built during the iterations, collectively making predictions about mortality rates.

One illustrative application of employing GBM in mortality modeling involves forecasting the mortality rates of a particular demographic by considering a multitude of factors, including age, gender, existing health conditions, lifestyle choices, and environmental influences. Through the aggregation of historical data on mortality rates and pertinent variables, the GBM model is trained. The mathematical expression of GBM can be summarized as:

$$F_m(x) = F_{m-1}(x) + \eta h(x;\theta_m)$$

Among them, $F_m(x)$ is the prediction function of the mth iteration, $F_{m-1}(x)$ is the prediction function of the previous iteration, η is the learning rate, $h(x;\theta_m)$ is the decision tree model for the current data set x, and θ_m is the parameter of the tree.

GBM enhances the model's prediction accuracy by progressively incorporating new decision trees to rectify the residuals from the preceding iteration. This iterative refinement process plays a pivotal role in refining mortality rate predictions and can provide valuable insights for various stakeholders. The insights derived from GBM's iterative approach hold value for healthcare providers, insurance companies, policymakers, and researchers.

1.1 Scenario: Leveraging GBM Modeling for Enhanced Preventive Care in Healthcare

In a bid to reduce mortality rates within a specific demographic group, a healthcare provider embarks on a mission to bolster preventive care efforts. Armed with historical data encompassing demographic details, health indicators, and mortality outcomes, the provider sets out to harness the power of GBM modeling to glean actionable insights that will shape their preventive care strategies.

1.1.1 Approach

The healthcare provider opts to leverage GBM modeling to delve into the data intricacies and extract pivotal insights that will steer their preventive care initiatives.

1.1.2 Data Features

- Demographic Information: Age, Gender, Location
- Health Indicators: Body Mass Index (BMI), Blood pressure readings, Presence of chronic conditions
- Historical Data: Previous healthcare interactions, Medication history, Past hospitalizations, GBM Insights
- Critical Risk Factors: The GBM model unveils that individuals surpassing a specific age threshold and those grappling with particular chronic conditions face an elevated mortality risk.

- Emerging Trends: Signs of a correlation between heightened BMI levels and mortality begin to surface through the analysis.

- Optimized Resource Allocation: Insights derived from the GBM model empower the healthcare provider to allocate resources judiciously. This enables targeted interventions for high-risk individuals based on the identified risk factors.

- Tailored Preventive Care: Armed with GBM insights, the healthcare provider can craft personalized preventive care plans tailored to individuals' unique risk profiles.

1.1.3 Expected Outcomes

- Tailored public health strategies tailored to the specific demographic group.
- Enhanced healthcare services honing in on high-risk individuals.
- Data-driven decision-making guiding resource allocation and intervention planning.

GBM's strength lies in its ability to handle complex, non-linear relationships in the data and capture interactions between different variables, providing valuable insights for researchers and actuaries in understanding mortality trends, risk factors, and patterns for informed decision-making in healthcare planning, insurance, and public policy.

2. Data Processing and Cleaning

2.1 Elevating Data Quality in Mortality Modeling through AI Integration

The integration of Artificial Intelligence (AI) in data processing and cleaning represents a paradigm shift, introducing advanced automated methodologies that not only enhance efficiency but also elevate the quality of data handling to a new level. This approach lays a foundation for subsequent analyses, ensuring that the insights derived are not only reliable but also actionable, driving informed decision-making in the healthcare sector [1].

2.2 AI-Driven Data Processing and Cleaning Example

By delving into the details of AI-driven data processing, cleaning, and missing value handling in mortality modeling, stakeholders in the healthcare sector can harness the power of AI to ensure data integrity, drive accurate predictions, and make informed decisions that positively impact public health outcomes. Methods for Outlier Identification:

• Z-scores and IQR Analysis: Utilizing Z-scores or the Interquartile Range (IQR) aids in pinpointing data points that deviate significantly from the mean.

• Model-Based Approaches: Decision trees, for instance, can discern outliers that exhibit substantial deviations from the rest of the data points.

• Visualization Tools: Leveraging visualization tools such as box plots and scatter plots facilitates the visual identification of outliers within the dataset.To ensure that data of different magnitudes and measurement units can be handled appropriately by the model, data standardization and normalization are necessary.

To guarantee that data with varying magnitudes and measurement units are appropriately handled by the model, the following techniques are imperative:

• Data Standardization: Standardizing data ensures that variables are on a similar scale, preventing biases due to differing magnitudes.

• Data Normalization: Normalizing data aids in adjusting the range of values, promoting consistency and enhancing model performance across diverse datasets.

By implementing these strategies, AI models can effectively identify and manage outliers, ensuring robust performance and accurate predictions in various applications.

Min-Max Normalization: Scales the data to be between 0 and 1, using the formula:

$$x_{norm} = \frac{x - min(x)}{max(x)min(x)}$$

Standardization: Converts the data to have a mean of 0 and a standard deviation of 1, using the formula:

$z = \frac{x - \mu}{\sigma}$ Where (μ) is the mean and (σ) is the standard deviation.

AI can automate the entire data processing and cleaning workflow, automatically identifying patterns in the data through machine learning algorithms and applying appropriate methods for handling them. For example, unsupervised learning algorithms can be used to identify outliers in the data, or deep learning models can be used to predict missing values.

Suppose we have mortality data for a specific population, including factors like age, pre-existing conditions, and mortality outcomes. In our data analysis endeavor aimed at predicting mortality risk and mitigating the impact of outliers on model performance, we will adopt a meticulous approach utilizing AI techniques. Our strategy involves leveraging Z-scores, visualization tools, and potentially decision trees to identify and address outliers within the mortality dataset.

The dataset comprises features such as Age (reflecting individuals' age), Pre-existing Conditions (indicating the presence of chronic illnesses like diabetes, heart disease), and Mortality Outcome (a binary variable denoting mortality as 1 or survival as 0). Our methodological steps include the process of Identifying Outliers with Z-scores or IQR, where Z-scores or IQR calculations for variables like age or pre-existing conditions will help flag data points deviating significantly from the mean. Furthermore, the Model-based Outlier Detection strategy will harness decision tree models to pinpoint data points that exert a substantial influence on mortality risk prediction, potentially uncovering outliers that could skew results. To enhance our outlier detection capabilities, we will employ Visualizing Outliers through the use of box plots or scatter plots, enabling us to visually depict the distribution of age or pre-existing conditions and effectively identify any outliers that may impact the integrity of our mortality prediction model.

The Isolation Forest model will identify outliers in the mortality data based on age and pre-existing conditions. The scatter plot will visualize the outliers, helping stakeholders in the healthcare industry to manage and understand data anomalies that could influence mortality risk predictions. This example illustrates a hypothetical use case of identifying and handling outliers in mortality data using AI techniques, emphasizing the importance of outlier detection in improving the accuracy and reliability of mortality risk predictions.

The application of AI in data processing and cleaning has greatly improved the efficiency and quality of actuarial analysis. By automating the process, AI is not only capable of handling large-scale datasets but also ensures the accuracy and consistency of the data, providing a solid foundation for mortality modeling and prediction.

3. Dynamic Mortality Tables

Dynamic Mortality Tables are continuously updated with real-time data to provide the latest insights into mortality rates and trends. These tables are used in mortality prediction to reflect the most current mortality patterns and to adjust mortality assumptions accordingly. AI can assist in creating dynamic mortality tables that are continuously updated using real-time data to reflect the latest mortality trends. The construction of dynamic tables typically involves time series analysis, such as the ARIMA (Auto-Regressive Integrated Moving Average) model, which is mathematically represented as:

$$X_t = c + \phi_1 X_{t-1} + \dots + \phi_p X_{t-p} - \theta_1 \epsilon_{t-1} - \dots - \theta_q \epsilon_{t-q} + \epsilon_t$$

Where (X_t) is the observed value at time (t), (c) is a constant term, (ϕ) and (θ) are the autoregressive and moving average parameters, respectively, and (ϵ_t) is the white noise error term.

3.1 Scenario of insurance company assess mortality risk for a specific demographic group

Let's consider a hypothetical scenario where an insurance company is using a dynamic mortality table to assess mortality risk for a specific demographic group.

Components of the Dynamic Mortality Table (see Table 1):

- Initial Data: The table starts with initial data including demographic information, health indicators, and mortality outcomes for a population.
- Real-Time Updates: The table receives real-time updates on mortality data, including new mortality rates, trends, and factors influencing mortality within the population.
- Adjustable Parameters: The table has adjustable parameters that can be modified based on the latest mortality data and insights obtained from ongoing updates.
- Usage: The insurance company utilizes the dynamic mortality table to determine insurance premiums, assess risk, and make informed decisions based on the most current mortality information available.

Table 1: Mortality table by age and health group

Age Group	**Health Conditions**	**Mortality Rate (%)**
50-60	Low	1.5
60-70	Moderate	3.2
70-80	High	6.7

Updates: New mortality data reveals an increase in mortality rates for the 70-80 age group due to a recent health epidemic. This new information prompts adjustments in the mortality rates for the 70-80 age group within the dynamic mortality table to reflect the current trends accurately.

3.2 Benefits and Application

The insurance company uses the dynamic mortality table to update insurance premiums, assess risk profiles, and tailor insurance offerings based on the latest mortality insights.

In real-world applications, dynamic mortality tables provide insurers and stakeholders with up-to-date information to make informed decisions, adapt risk models, and effectively manage mortality risks based on current trends and data [2].

4. Mitigate Biased Decision-making in AI algorithms

When applying AI for mortality rate prediction, it is important to consider risks such as model overfitting, data privacy and security issues, and model interpretability. To mitigate these risks, models need to be trained on diverse and representative datasets, improve model transparency and fairness, and implement strict data protection measures.

Actuaries need to consider data legality, model transparency and interpretability, compliance with ethical and legal standards, and ongoing regulatory compliance when applying AI for mortality rate modeling. They also need to continuously update their professional knowledge and skills to keep up with the development of AI and big data technologies.

The widespread use of Artificial Intelligence (AI) in the insurance sector brings several risks and challenges that companies need to navigate to ensure the responsible and ethical deployment of AI technologies. Here are some key risks emerging from the use of AI in the insurance industry.

Addressing algorithmic bias in AI systems requires a combination of technical approaches, ethical considerations, and regulatory oversight. Techniques such as bias detection, data preprocessing, fairness constraints, and explainable AI can help mitigate bias and promote more equitable and transparent decision-making in AI systems. By understanding how bias manifests and taking proactive steps to address it, developers and users can work towards building fairer and more inclusive AI technologies.

4.1 Several factors can contribute to algorithmic bias in AI systems

Several factors can contribute to the emergence of algorithmic bias in AI systems. These factors often intersect throughout the AI development process and can influence the presence and extent of bias in the resulting algorithms. Here are some key factors that contribute to the emergence of algorithmic bias:

4.1.1 Biased Training Data:

The most common source of algorithmic bias is biased training data. Historical data often reflects societal biases, stereotypes, or systemic inequalities, which can be unintentionally encoded into AI systems during the training phase.

For example, an insurance company uses an AI algorithm to determine car insurance premiums for policyholders based on historical claims data. The training data predominantly consists of claims data from urban areas, leading to overrepresentation of claims from city drivers. The dataset lacks sufficient data from rural areas and under-represents low-income individuals.

4.1.2 Data Selection Bias:

Data selection bias occurs when certain groups or perspectives are underrepresented or overrepresented in the training data, leading to skewed or incomplete datasets that do not accurately reflect the full range of real-world scenarios. Imagin an insurance company uses an AI algorithm to assess health insurance risk profiles based on historical claims data. The training data primarily consists of claims data from individuals who have regularly visited healthcare facilities and have a higher documented medical history. The dataset lacks representation of healthy individuals or those who may have had minimal healthcare needs.

4.1.3 Data Labeling Bias:

Biases can also be introduced during the data labeling process, where human annotators may unknowingly inject their biases into the training data through subjective or culturally influenced labeling decisions. An insurance company uses an AI algorithm to assess risk profiles for home insurance policies. The algorithm relies on labeled data to identify risk factors associated with properties. The data labeling process involves human annotators who unintentionally introduce biases in determining property risks based on subjective judgments or assumptions.

4.1.4 Algorithm Design Choices:

Algorithmic bias can be unintentionally introduced through design choices such as feature selection, model complexity, hyperparameter tuning, or optimization strategies. Biased assumptions embedded in the algorithm design can lead to biased outcomes.

4.1.5 Feedback Loop Effects

AI systems that interact with users and learn from feedback data can develop feedback loop bias. If the feedback data is biased, the system may reinforce or amplify existing biases over time, leading to discriminatory outcomes.

4.1.6 Contextual Biases

The context in which AI systems are deployed can also contribute to algorithmic bias. Biases may emerge from specific use cases, application domains, cultural norms, or social structures that influence the data collection, algorithm design, or decision-making processes.

4.1.7 Human Involvement

Humans involved in the AI development lifecycle, including data scientists, engineers, and designers, can introduce biases consciously or unconsciously. Their subjective judgment, prior beliefs, assumptions, or cultural influences can shape the AI system's behavior.

4.1.8 Lack of Diversity

Lack of diversity in AI development teams or insufficient representation of diverse perspectives and voices can contribute to the perpetuation of biases in AI systems. Diverse teams can bring different viewpoints and experiences to identify and address bias effectively.

Addressing algorithmic bias requires a holistic approach that involves careful data curation, transparency in algorithmic decision-making, diversity in AI teams, ongoing monitoring for bias, and the integration of fairness considerations throughout the AI development lifecycle. By understanding and mitigating the factors that contribute to bias, developers and practitioners can work towards creating AI systems that are more equitable, accountable, and inclusive [3].

4.2 The consequences of algorithmic bias on actuarial analysis

In actuarial analysis, AI algorithms trained on biased data may lead to discriminatory outcomes in insurance underwriting and pricing practices. One example of this is the use of historical claims data that may contain inherent biases related to factors like race, gender, or socioeconomic status. If AI algorithms are trained on this biased data, they may inadvertently perpetuate these biases in insurance risk assessments and pricing decisions, leading to discriminatory outcomes for certain demographic groups.

As noted above, actuarial services are data-driven; data bias, left unaddressed, can lead to incorrect conclusions, unwanted consequences, wrong policy decisions, or inadequate system performance. This section provides a few examples of actuarial services that can be impacted by data bias.

4.3 Risk Classification

Risk classification is the process of evaluating and estimating the future costs related to transferring risk. Biased data can introduce discrepancies between the actual future costs and the actuary's projections, potentially leading to overcharging or undercharging and adverse selection. Availability bias and historical bias are two significant factors that can impact actuarial decisions.

Using a machine learning model to predict life insurance policyholder mortality rates, an insurer inadvertently incorporates biased data that skews towards affluent applicants. As a result, the algorithm may underestimate risks for certain demographic groups, leading to improper risk assessments and potentially unsustainable pricing strategies. An additional instance of bias is historical bias, where differences in homeownership by race are overlooked in a personal auto rating plan. This omission can lead actuaries to base results on this bias rather than the genuine driver of future loss performance.

Imaging an insurance company uses an AI algorithm to assess risk profiles and determine premiums for auto insurance. If the algorithm is trained on biased data that correlates accidents with a specific demographic group rather than driving behavior, it may unfairly penalize individuals in that group with higher premiums, leading to discrimination and perpetuating unfair practices.

References

[1] Sonal Trivedi, M. K. Nallakaruppan, Balamurugan Balusamy, Nithya Rekha Sivakumar, 2024, Artificial Intelligence and Actuarial Applications and Case Studies from Finance and Insurance Science.

[2] Prof Joab Onyango Odhiambo, 2024, Artificial Intelligence in Actuarial Science within Sub-Saharan Africa: Practice, Processes, and Applications.

[3] Jeroen Erne, 2024, The Artificial Intelligence Handbook for Insurance Actuaries: "Future-Proof Your Skills; Save a Wealth of Time; and Secure Your Job." (AI Handbook for Finance Series)