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

Jiaming Zuo, FSA, CERA, FCAA, FASHK

Any views and ideas expressed in the essays are the author's alone and may not reflect the views and ideas of the Society of Actuaries, the Society of Actuaries Research Institute, Society of Actuaries members, nor the author's employer.

The application of artificial intelligence (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.

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 $\theta_m$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.

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.

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.

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.

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.

DATA PROCESSING AND CLEANING

ELEVATING DATA QUALITY IN MORTALITY MODELING THROUGH AI INTEGRATION

The integration of 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.

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.

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.

SCENARIO: INSURANCE COMPANY USING DYNAMIC MORTALITY TABLE TO ASSESS MORTALITY RISK

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:

- 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.

DATA REPRESENTATION (SIMPLIFIED)

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.

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.

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.

* * * * *

Jiaming Zuo is a Senior Partner for EverBright Actuarial Consulting Limited. EverBright integrated AI into their digital platform for customizing and managing group health and insurance policies for clients. She can be reached at [jzuo@ebactuary.com.](mailto:jzuo@ebactuary.com)