IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 July 2024, accepted 12 August 2024, date of publication 19 August 2024, date of current version 30 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3446179

RESEARCH ARTICLE

Enhancing Auto Insurance Risk Evaluation With Transformer and SHAP

TIEJIANG SUN[1], (Member, IEEE), JINGYUN YANG[2], JIALE LI[3], JIAYING CHEN[4], MINGYUE LIU[5], LI FAN[6], AND XUKANG WANG[7], (Member, IEEE)

[1]Ping An Property and Casualty Insurance Company of China Ltd., Shenzhen 518043, China
[2]Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3]Tandon School of Engineering, New York University, New York, NY 10012, USA
[4]SC Johnson Graduate School of Management, Cornell University, Ithaca, NY 14853, USA
[5]Department of Computer Science, Cornell University, Ithaca, NY 14850, USA
[6]College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China
[7]Sage IT Consulting Group, Shanghai 200060, China

Corresponding author: Xukang Wang (xkw@ieee.org)

ABSTRACT The evaluation of auto insurance risks is a fundamental task for financial institutions, crucial for setting equitable premiums and managing risks effectively. Traditional machine learning algorithms often struggle to capture the intricate relationships necessary for accurate risk assessment. In contrast, deep learning methods, while capable of processing complex data structures, lack the ability to model feature interactions and provide interpretability, which are essential for transparent decision-making in the insurance industry. To address these challenges, we introduce the Actuarial Transformer (AT)–a pioneering model that leverages the self-attention mechanism of the Transformer architecture to meticulously map feature interactions. The AT integrates advanced residual models with tree-based methods, enhancing its predictive accuracy. Additionally, it incorporates the SHAP (SHapley Additive exPlanations) model, which uses Shapley values from cooperative game theory to ensure interpretability and transparency in its risk assessments. Our empirical analysis, conducted on a representative dataset of auto insurance risks, demonstrates the AT's superior performance in risk prediction. The SHAP analysis further validates the model's ability to prioritize features logically, providing clear insights into the decision-making process. The AT not only improves the precision of auto insurance risk evaluations but also enhances the interpretability of these evaluations, making it a valuable tool for both industry practitioners and clients.

INDEX TERMS Auto insurance, risk evaluation, deep learning.

I. INTRODUCTION

Property and casualty (P & C) insurance, particularly auto insurance, faces intense competition and low growth in the insurance sector industry. According to the National Association of Insurance Commissioners (NAIC) annual report [1], the overall industry profit ratio has fluctuated around -3% to 3% over the past decade. This challenging environment necessitates precise insurance pricing, considering numerous rate-making factors and their complex relationships.

Auto insurance risk evaluation is a fundamental task for financial institutions, crucial for setting equitable premiums and managing risks effectively. The process typically involves three levels: actuaries predicting pure risk premium, underwriters deciding on price strategy, and sales analysis of customer preferences [2], [3], [4]. Among these, the actuaries' task of pure risk premium prediction, also known as ratemaking, forms the foundation of pricing and significantly impacts company revenue.

Traditional machine learning algorithms, while improving upon simple mathematical models, face several challenges in practice. These include the need for periodic algorithm upgrades to match evolving data, the potential for misclassification by individual models, and the complexity of hyperparameter optimization [5]. Moreover, many advanced models lack interpretability,

The associate editor coordinating the review of this manuscript and approving it for publication was Sawyer Duane Campbell.

creating a trade-off between predictive accuracy and model explainability.

To address these challenges, we introduce the Actuarial Transformer (AT) - a pioneering model that leverages the self-attention mechanism of the Transformer architecture to meticulously map feature interactions in auto insurance risk evaluation. The AT integrates advanced residual models with tree-based methods, enhancing its predictive accuracy while maintaining interpretability through the incorporation of SHAP (SHapley Additive exPlanations) analysis.

Our contributions in this paper are summarized as follows:

- We propose the Actuarial Transformer model to conduct the auto insurance risk evaluation.
- We utilize the self-attention mechanism within the transformer architecture to model the relationships and interactions among the policyholder features.
- We design a residual modeling structure, which refines the modeling residual errors of transformer with the tree model, thus combining the advantages of them.
- We incorporate the SHAP method to analyze the feature importance for the ultimate model evaluation, enhancing the interpretability of this process.
- We perform detailed experiments on a representative auto insurance risk dataset and demonstrate the effectiveness of our proposed method.

The aim of this work is to develop a more accurate and interpretable model for auto insurance risk evaluation. We chose the Transformer architecture as the foundation of our model due to its proven ability to capture complex relationships in sequential data [6], which we hypothesize can be effectively applied to the multifaceted nature of insurance risk factors. The addition of tree-based residual modeling aims to capture any remaining patterns that the Transformer might miss, addressing the potential limitations of a single model approach [5].

The use of SHAP values addresses the critical need for interpretability in insurance applications, providing a method to explain model decisions in a way that is both mathematically sound and intuitively understandable [7]. This approach has the potential to benefit both insurance companies, by enabling more precise risk assessment and pricing, and policyholders, by ensuring fairer and more transparent premium calculations.

Our empirical analysis, conducted on a representative dataset of auto insurance risks obtained from the Casualty Actuarial Society (CAS) [8], demonstrates the AT's superior performance in risk prediction. The SHAP analysis further validates the model's ability to prioritize features logically, providing clear insights into the decision-making process.

In the following sections, we will discuss related work in the field of auto insurance risk evaluation and deep learning (Section II), detail our proposed methodology (Section III), present our experimental results (Section IV), and conclude with a discussion of our findings and future directions (Section V).

II. RELATED WORKS

A. WIDE APPLICATION OF MACHINE LEARNING APPROACHES IN AUTO INSURANCE INDUSTRY

In the past decade, artificial intelligence has revolutionized the insurance industry, particularly in auto insurance. As synthesized by Hanafy and Ruixing [9], three key areas have seen significant impact:

- Insurance process rebuilding: Companies like Lemonade have digitized entire business operations, applying AI from quote assessment to claim filing.
- Behavior detection: Approximately 10% of companies now use machine learning techniques to detect fraudulent insurance claims [9].
- Predictive analytics: Paredes's research on reducing auto insurance attrition using econometrics demonstrated an additional USD 750,000 in revenue [10].

These advancements underscore AI's pivotal role in reshaping the insurance landscape, particularly in risk evaluation and pricing.

B. TRADITIONAL MACHINE LEARNING ALGORITHMS IN RATEMAKING

Several traditional machine learning algorithms have been widely adopted in the Property and Casualty (P & C) industry for ratemaking and claim number prediction:

- Accurate Generalized Linear Model (AGLM): Based on GLM and equipped with recent data science techniques, AGLM achieves high interpretability and predictive accuracy compared to other GLMs (GLM, GAM, and GBM) [11].
- Extreme Gradient Boosting (XGBoost): Widely used in claims occurrence prediction, XGBoost has shown superior performance in car insurance claim number prediction classification compared to Neural Networks, Logistic Regression, and Naive Bayes [12], [13].
- Random Forest (RF): RF has demonstrated high accuracy in predicting claim severity, producing comparable results to using all features while only utilizing 1/3 of the overall features [14].
- Neural Networks (NN): Artificial Neural Networks (ANN) have shown promising results in auto insurance claims prediction, outperforming Decision Trees and Multinomial Logistic Regression with 61.71% overall classifier accuracy [5].

C. ACCURACY AND ROBUSTNESS PERFORMANCE OF TRADITIONAL MACHINE LEARNING

While machine learning algorithms have improved ratemaking accuracy compared to simple mathematical algorithms, they face three main challenges in practice:

- Performance dependence on data quantity and integrity, necessitating periodic algorithm updates.
- Complementary nature of different models, suggesting that a combination of models could result in better prediction performance [5].

- Sensitivity to hyperparameter settings, especially in Neural Networks, requiring expertise in AI engineering and algorithm optimization.

D. EXPLAINABILITY OF TRADITIONAL MACHINE LEARNING

Among traditional machine learning algorithms, only GLMs achieve high interpretability. XGBoost, RF, and NNs present challenges in explaining their decision-making process to stakeholders. To address this, the SHAP (SHapley Additive exPlanations) algorithm has been proposed to provide a detailed understanding of feature contributions to predictions [7], [16].

E. DEEP LEARNING IN INSURANCE RISK EVALUATION

Recent years have seen an increasing interest in applying deep learning techniques to insurance risk evaluation. Some notable works include:

- Ying [15] proposed a deep learning approach for insurance premium pricing, demonstrating improved accuracy over traditional methods.
- Wüthrich and Merz [17] explored the use of neural networks for claims reserving in non-life insurance.
- Gabrielli et al. [18] introduced a neural network approach for individual claims reserving in insurance.
- Spedicato et al. [19] reviewed machine learning methods for actuarial applications, including deep learning approaches.

F. RESEARCH GAP AND PROPOSED APPROACH

To better illustrate the research gap and justify our proposed Actuarial Transformer (AT) approach, we present a comparison of related works in Table 1. As shown in Table 1, existing methods face trade-offs between accuracy, interpretability, and the ability to model complex feature interactions. Our proposed Actuarial Transformer (AT) aims to address these limitations by combining the strengths of deep learning (specifically, the Transformer architecture) with tree-based models and SHAP analysis.

The AT leverages the self-attention mechanism of Transformers to capture complex feature interactions, which is crucial in auto insurance risk evaluation where multiple factors interact in non-linear ways. By incorporating tree-based residual modeling, we aim to capture any remaining patterns that the Transformer might miss, thereby enhancing overall prediction accuracy. The use of SHAP analysis addresses the interpretability challenge, providing clear insights into the model's decision-making process.

This approach is justified by the need for a method that can simultaneously achieve high accuracy, maintain interpretability, and effectively model complex feature interactions in auto insurance risk evaluation. The AT's design specifically addresses these requirements, filling a gap in the existing literature and potentially advancing the state-of-the-art in this field.

III. METHODOLOGY

A. PROBLEM FORMULATION

The aim of insurance risk evaluation is to predict the expected risk $\hat{y}_i$ for each policyholder i. In this work, the risk for each policyholder is quantified as the number of claims. Mathematically, our objective is to predict the conditional mean $\mathbb{E}(Y_i|X_i)$ of the risk variable Y_i, where X_i denotes the feature vector of policyholder i.

Formally, we model the functional relationship between the feature vector X_i and the resultant risk Y_i as:

$$\hat{Y}_i = f(X_i; \theta), \tag{1}$$

where $f(\cdot)$ represents the modeling function parameterized by θ, designed to capture the intricate relationship between policyholder characteristics X_i and their corresponding risk Y_i. To ease the understanding to this paper, we summarize the notations in the following Table 2.

B. DATA PREPROCESSING

In this section, we detail the preprocessing steps for the Actuarial Transformer (AT) model, which is designed to analyze and quantify complex interactions among policyholder attributes, including both continuous and categorical variables, for insurance risk evaluation.

1) FEATURE ENCODING

Consider a dataset of N policyholders, each described by M features:

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N\}, \tag{2}$$

where $\mathbf{x}_i \in \mathbb{R}^M$ represents the feature vector of the i-th policyholder, encompassing both continuous and categorical attributes. To efficiently process these heterogeneous data types, we apply:

1) An embedding layer for categorical features, translating them into dense vectors of fixed size [20]. 2) A normalization layer for continuous features, ensuring a uniform scale.

This preprocessing step transforms $\mathbf{x}_i$ into a uniform representation:

$$\mathbf{x}'_i = \text{Embedding}(\mathbf{x}_i^{cat}) \oplus \text{Normalize}(\mathbf{x}_i^{cont}), \tag{3}$$

where $\oplus$ denotes concatenation, $\mathbf{x}_i^{cat}$ and $\mathbf{x}_i^{cont}$ correspond to categorical and continuous parts of $\mathbf{x}_i$, respectively.

2) DATA NORMALIZATION

The Normalize operation in Equation 3 is achieved through:

$$\text{Normalize}(\mathbf{x}_i^{cont}) = \frac{\mathbf{x}_i^{cont} - \boldsymbol{\mu}^{cont}}{\boldsymbol{\sigma}^{cont}}, \tag{4}$$

where $\boldsymbol{\mu}^{cont}$ and $\boldsymbol{\sigma}^{cont}$ are the mean and standard deviation vectors of the continuous features across the dataset, respectively.

TABLE 1. Comparison of Related Works.

Study	Method	Accuracy	Interpretability	Feature Interaction
GLM [11]	Linear	Medium	High	Low
XGBoost [12]	Tree-based	High	Medium	Medium
RF [14]	Tree-based	High	Medium	Medium
NN [5]	Deep Learning	High	Low	Medium
Deep Learning [15]	Deep Learning	High	Low	High
Our AT	Transformer + Tree	Very High	High	Very High

TABLE 2. Summary of Notations.

Notation	Description
X_i	Feature vector of policyholder i
Y_i	Risk variable (number of claims) for policyholder i
$\hat{Y}_i$	Predicted risk for policyholder i
$f(\cdot)$	Modeling function
θ	Model parameters
N	Number of policyholders
M	Number of features
$\mathbf{X}$	Dataset of all policyholders' features
$\mathbf{x}_i$	Feature vector of the i-th policyholder
$\mathbf{x}_i^{cat}$	Categorical part of $\mathbf{x}_i$
$\mathbf{x}_i^{cont}$	Continuous part of $\mathbf{x}_i$
$\mathbf{x}_i^{'}$	Transformed feature vector
$\boldsymbol{\mu}^{cont}$	Mean vector of continuous features
$\boldsymbol{\sigma}^{cont}$	Standard deviation vector of continuous features
Q, K, V	Query, Key, and Value matrices in self-attention
W^Q, W^K, W^V	Weight matrices for Q, K, V transformations
A_{pq}	Attention score between features p and q
I	Output representation of feature interaction
$\hat{y}_i^{tree}$	Tree model prediction for policyholder i
r_i	Residual between actual and tree-predicted values
$\hat{r}_i$	Predicted residual by Transformer
$\hat{y}_i^{final}$	Final refined evaluation
$\mathcal{L}$	Loss function
L_{PD}	Poisson deviance loss function
λ	Regularization strength

C. ACTUARIAL TRANSFORMER ARCHITECTURE

The Actuarial Transformer (AT) is an adaptation of the original Transformer architecture [6] for insurance risk evaluation. It utilizes a self-attention mechanism to model the relationships and interactions between features. This approach has shown great promise in capturing complex dependencies in various domains [21], [22].

1) SELF-ATTENTION MECHANISM

For an encoded and normalized input feature vector $\mathbf{x}_i'$, the self-attention mechanism operates as follows:

1) Linear Transformation: The input vector $\mathbf{x}_i'$ is linearly transformed into three different vectors known as Query (Q), Key (K), and Value (V) through learned weight matrices:

$$Q = \mathbf{x}_i' W^Q, \quad K = \mathbf{x}_i' W^K, \quad V = \mathbf{x}_i' W^V, \tag{5}$$

where $W^Q, W^K, W^V \in \mathbb{R}^{d\times d}$ are the weight matrices for queries, keys, and values, respectively, and d is the dimension of the input features.

2) Attention Scoring: The attention score between two features p and q within $\mathbf{x}_i'$ is computed by taking the dot product of their respective query and key vectors, followed by a softmax operation to ensure scores are normalized:

$$A_{pq} = \text{softmax}\left(\frac{Q_p K_q^T}{\sqrt{d_k}}\right), \tag{6}$$

where d_k is the scaling factor equal to the dimension of the key vectors, ensuring stability in the dot products [6].

3) Feature Interaction: The output representation capturing the interaction between features is obtained by a weighted summation of the value vectors V, weighted by the attention scores A:

$$I = \sum_q A_{pq} V_q. \tag{7}$$

The resultant interaction representation, denoted as I, augments the model's understanding of the intricate relationships between features, which is essential for precise risk evaluation.

D. RESIDUAL LEARNING FOR EVALUATION PRECISION ENHANCEMENT

To further enhance the precision of insurance risk evaluation, we adopt a residual learning approach [23]. This technique not only improves predictive performance but also addresses potential overfitting that might emerge when modeling complex interactions among diverse policyholder features.

Our AT model employs tree models as a robust base for initial assessments, complemented by a transformer architecture to model and rectify residual errors. The procedure is as follows:

1) Initial Tree-based Evaluation: We begin with an ensemble of tree models, which are adept at handling heterogeneous feature types. The tree models provide a set of initial evaluations $\hat{y}_i^{tree}$:

$$\hat{y}_i^{tree} = \text{Tree}(\mathbf{x}_i), \tag{8}$$

where Tree represents the function encapsulating the learning process of Gradient Boosting over decision trees [24].

2) Residual Computation: We compute the residuals–the difference between the actual target values y_i and the tree-predicted values $\hat{y}_i^{tree}$:

$$r_i = y_i - \hat{y}_i^{tree}. \tag{9}$$

3) Transformer-based Residual Modeling: The AT model processes the features $\mathbf{x}_i'$ to predict these residuals more effectively:

$$\mathbf{x}_i^{',t} = \text{Transformer}(\mathbf{x}_i'), \tag{10}$$

where Transformer represents the operations described in Section III-C. We then model the residuals using this refined feature $\mathbf{x}_i^{\prime,t}$, employing a link function f, which is learned during the training process:

$$\hat{r}_i = f(\mathbf{x}_i^{\prime,t}). \tag{11}$$

4) Final Evaluation: The final refined evaluation $\hat{y}_i^{final}$ is the sum of the initial evaluations provided by the tree models and the residual evaluations modeled by the transformer:

$$\hat{y}_i^{final} = \hat{y}_i^{tree} + \hat{r}_i. \tag{12}$$

E. MODEL TRAINING

The model is trained using an optimization objective that incorporates both the initial tree evaluations and the transformer-modeled residuals:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} L_{PD}(y_i, \hat{y}_i^{final}) + \lambda||\theta||_2^2, \tag{13}$$

where L_{PD} denotes the Poisson deviance loss function, which is suitable for count data in insurance risk evaluation [25]. θ represents the parameters of both tree and transformer models, and λ is the regularization strength.

This approach to residual modeling enables the AT model to capture fine-grained patterns in the data, offering interpretable insights into the critical factors driving risk, and representing a significant advancement toward more precise insurance risk assessments. To facilitate the comprehension and reproduction of our proposed Actuarial Transformer framework, we provide the pseudocode as follows:

Algorithm 1 Actuarial Transformer (AT) Framework

Require: Dataset $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, Target values $\mathbf{Y} = \{y_1, y_2, \ldots, y_N\}$
Ensure: Predictions $\hat{\mathbf{Y}}^{final} = \{\hat{y}_1^{final}, \hat{y}_2^{final}, \ldots, \hat{y}_N^{final}\}$
function PreprocessFeatures($\mathbf{X}$)
  for each $\mathbf{x}_i$ in $\mathbf{X}$ **do**
    $\mathbf{x}_i^{cat} \leftarrow$ CategoricalFeatures($\mathbf{x}_i$)
    $\mathbf{x}_i^{cont} \leftarrow$ ContinuousFeatures($\mathbf{x}_i$)
    $\mathbf{x}_i^{cat} \leftarrow$ Embedding($\mathbf{x}_i^{cat}$)
    $\mathbf{x}_i^{cont} \leftarrow$ Normalize($\mathbf{x}_i^{cont}$)
    $\mathbf{x}_i' \leftarrow \mathbf{x}_i^{cat} \oplus \mathbf{x}_i^{cont}$
  end for
  return $\mathbf{X}' = \{\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_N'\}$
end function
function TreeModel($\mathbf{X}$, $\mathbf{Y}$)
  Train tree ensemble on $\mathbf{X}$ and $\mathbf{Y}$
  $\hat{\mathbf{Y}}^{tree} \leftarrow$ TreeEnsemblePredictions($\mathbf{X}$)
  return $\hat{\mathbf{Y}}^{tree}$
end function
function TransformerModel($\mathbf{X}'$, $\mathbf{R}$)
  for each $\mathbf{x}_i'$ in $\mathbf{X}'$ **do**
    $Q, K, V \leftarrow$ LinearTransform($\mathbf{x}_i'$)
    $A \leftarrow$ SoftMax($\frac{QK^T}{\sqrt{d_k}}$)
    $I \leftarrow AV$
    $\mathbf{x}_i^{\prime,t} \leftarrow$ TransformerLayers(I)
    $\hat{r}_i \leftarrow f(\mathbf{x}_i^{\prime,t})$
  end for
  return $\hat{\mathbf{R}} = \{\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_N\}$
end function
$\mathbf{X}' \leftarrow$ PreprocessFeatures($\mathbf{X}$)
$\hat{\mathbf{Y}}^{tree} \leftarrow$ TreeModel($\mathbf{X}$, $\mathbf{Y}$)
$\mathbf{R} \leftarrow \mathbf{Y} - \hat{\mathbf{Y}}^{tree}$
$\hat{\mathbf{R}} \leftarrow$ TransformerModel($\mathbf{X}'$, $\mathbf{R}$)
$\hat{\mathbf{Y}}^{final} \leftarrow \hat{\mathbf{Y}}^{tree} + \hat{\mathbf{R}}$
Train model by minimizing:
$Loss = \frac{1}{N}\sum_{i=1}^{N} L_{PD}(y_i, \hat{y}_i^{final}) + \lambda||\theta||_2^2$
return $\hat{\mathbf{Y}}^{final}$

IV. EXPERIMENTS

In our study, we pose several research questions (RQs) to assess the capabilities of our Actuarial Transformer (AT) model in the context of auto insurance risk evaluation:

- **RQ1:** How does the AT model's risk evaluation accuracy compare with existing benchmarks?
- **RQ2:** Does each module within the AT model contribute to enhancing the overall performance of risk evaluation?
- **RQ3:** How consistent is the AT model's performance when subject to varying hyperparameter configurations?
- **RQ4:** Does the AT model exhibit stable and rapid convergence during the training process?
- **RQ5:** Can the feature importance rankings produced by the AT model be justified in the final evaluation?

A. EXPERIMENT SETUP

1) DATASETS

a: OVERALL INTRODUCTION

In the experiments, we utilize the **French Motor Third-Party Liability Claims (MTPL)** [26][1]**:** to evaluate the effectiveness of our proposed model AT compare with other representative baselines. In this dataset, risk features and evaluation labels (claim numbers) were collected for 677,991 motor third-party liability policies over a one-year observation period. The table contains 12 columns: *IDpol, ClaimNb, Exposure, Area, VehPower, VehAge, DrivAge, BonusMalus, VehBrand, VehGas, Density, Region*, where *ClaimNb* is the label. *Exposure* measures the exposure period. Among other columns, the *VehBrand* and *Region* are pre-processed as categorical features while the remaining columns are transformed to be continuous ones. The corresponding descriptions for each feature are provided in the Table 3. We split the overall dataset randomly into the training set and test set with a ratio of 4:1.

b: DETAILED DATA ANALYSIS

Before delving into the performance evaluation of various methods, we first conduct a detailed data analysis to the

[1] https://www.kaggle.com/datasets/floser/french-motor-claims-datasets-fremtpl2freq

TABLE 3. The features and their corresponding descriptions in the auto insurance risk dataset: MTPL.

Feature	Description
IDpol	The policy ID (used to link with the claims dataset).
ClaimNb	The number of claims during the exposure period.
Exposure	The period of exposure for a policy, in years.
Area	The area code.
VehPower	The power of the car (split into ordinal categories).
VehAge	The vehicle age, in years.
DrivAge	The driver's age, in years.
BonusMalus	Bonus/malus, <100 means bonus, >100 means malus.
VehBrand	The brand of the car, divided into the following groups: A, B.
VehGas	The car's fuel type: Diesel or Regular (petrol).
Density	The density of inhabitants, in the city that the driver lives in.
Region	The policy region in France.

dataset to better illustrate the pattern in the real-world auto insurance industry. First, we calculate the distribution for the claim number and exposure per policy. We visualize them via the histograms in Figure 1. First, we can easily notice that the claim numbers for most policies are limited to 0-2, which indicates that most policies will not trigger a claims or just trigger a small number of claims. Meanwhile, the distribution for the exposures is a little bit more balanced, while most policies are still exposed for the whole study period, i.e., corresponding to the exposure value of 1.0. Then, we discuss the exposure and claim frequency distribution according to the different Bonus-Malus groups. The related visualization has been provided in the Figure 2. From the left subfigure, we can find that low-bonus policies occupied the most exposure. Meanwhile, from the right subfigure, we can observe that as the malus value improves (<100 means bonus, >100 means malus), the claim frequency within the certain confidence interval will also increase. This is actually with the specific situations in the insurance industry, where the malus amount is directly related to the historical claim number. In the insurance industry, the term Bonus-Malus refers to a system that adjusts policyholders' premiums based on their claims history. The system is designed to incentivize safe driving behavior by offering a "bonus" or discount to those who haven't filed any claims within a specified period, while imposing a "malus" or penalty on those who have. If a policyholder has no claims over a certain period, typically a year, they receive a bonus in the form of a reduced premium rate for the following period. The longer the policyholder goes without filing a claim, the larger the bonus and, consequently, the lower their future premiums may become. The bonus encourages drivers to avoid risky behavior that could lead to accidents and claims. Conversely, if a policyholder files a claim, they can expect an increase in their premium rates–a malus, which reflects their higher perceived risk. Subsequent premiums can continue to increase with each claim filed, which discourages the making of frivolous or fraudulent claims. The malus component of the system attempts to equitably distribute the cost of insurance across all policyholders based on individual risk. The Bonus-Malus system, also known as a no-claims discount (NCD) or no-claims bonus (NCB) in some markets, acts as a merit-based pricing strategy in auto insurance, although it can also be found in other types of insurance. The exact rules, percentage discounts, and surcharges under the Bonus-Malus system can vary widely among insurers and across different regions. The system is transparently outlined in insurance policies, ensuring that policyholders are aware of the financial implications of their driving habits.

2) BASELINES

In experiments, we use following models as our baselines:

- **Generalized Linear Model(GLM) [27]:** GLM extends traditional linear regression to models with a non-normal error distribution, accommodating various types of response variables through different link functions and distribution assumptions.
- **XGBoost [28]:** XGBoost, short for Extreme Gradient Boosting, is a highly efficient and flexible machine learning library implementing the gradient boosting framework, known for its speed and performance, with built-in regularization to avoid overfitting.
- **LightGBM [29]:** LightGBM distinguishes itself through its Lightweight implementation that uses a histogram-based algorithm and gradient-based one-side sampling, optimizing both memory usage and speed for large datasets.
- **CatBoost [30]:** CatBoost is a fast, scalable, high-performance Gradient Boosting on Decision Trees library, with particular effectiveness in handling categorical features directly, hence reducing data preprocessing time. Besides, it introduces *ordered boosting* to overcome the target leakage in gradient boosting.
- **NNemb:** This model incorporates embedding layers to transform categorical inputs into continuous vectors before processing them through a neural network, enhancing the model's ability to capture deep, nonlinear relationships within the data.
- **TabNet [31]:** TabNet relies on the multi-step feature selection with sequential attention to conduct the deep insurance risk evaluation. It uses a unique instance-wise feature selection property, which helps in understanding the decision-making process by allowing inspection of which features are used for evaluations. It should be noted that TabNet is one of the most state-of-the-art (SOTA) approach in this area.

3) EVALUATION METRICS

- **Poisson Deviance:** It measures the discrepancy between risk label y_i and the risk $\hat{y}_i$ predicted by the model M for each policy i via a log-likelihood-based scoring rule:

$$PD(M) = 2\sum_{i=1}^{N} y_i log(\frac{y_i}{\hat{y}_i}) - (y_i - \hat{y}_i). \tag{14}$$

Poisson deviance is deployed as a robust assessment metric for insurance risk evaluation, where the target indicator is the count of claims. Here, Poisson Deviance is preferred over MSE or MAE for several reasons:

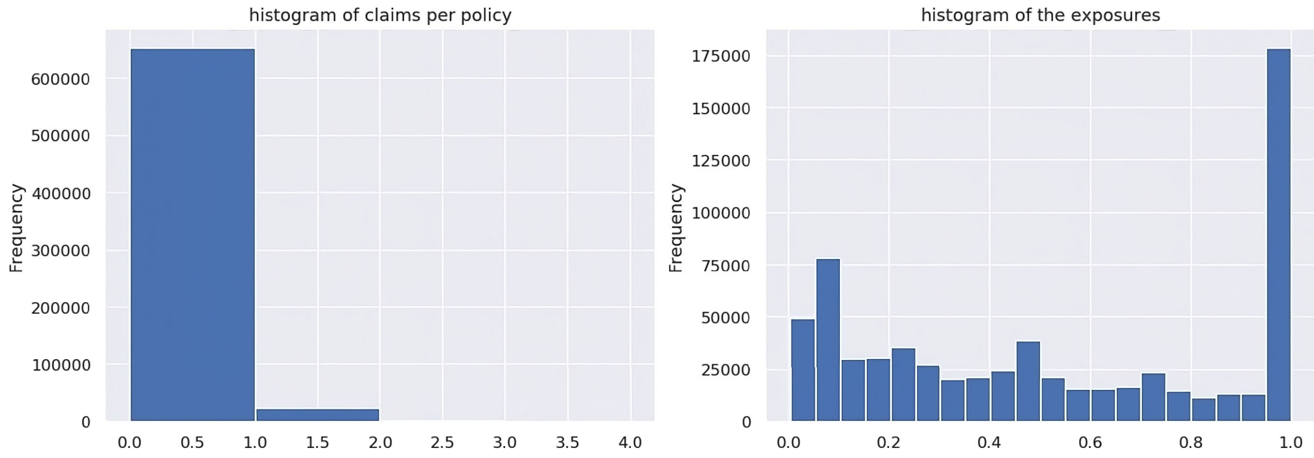

FIGURE 1. **The histograms for the number of claims and the exposure in the dataset to visualize their distribution.**

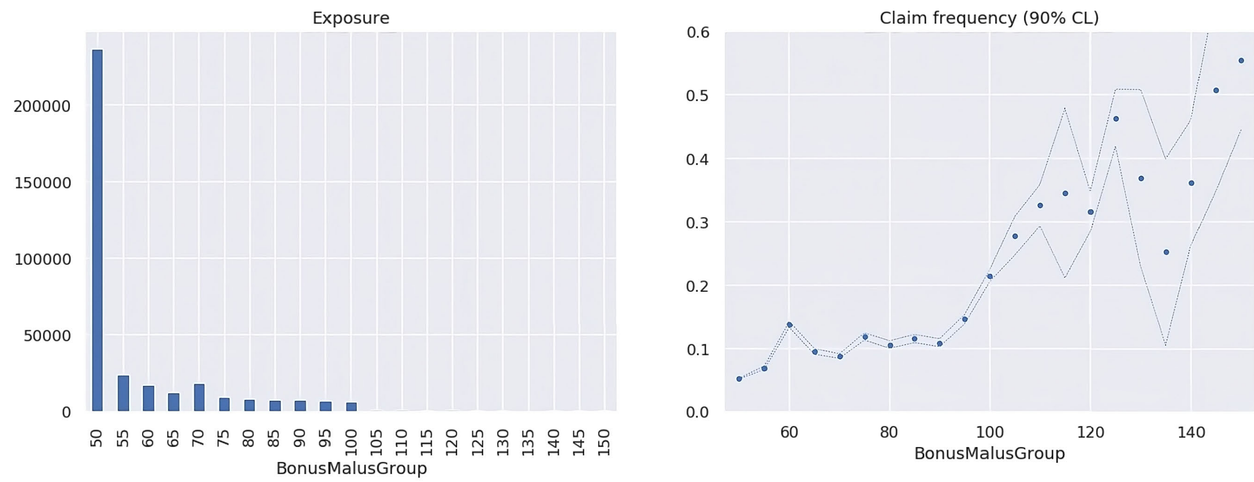

FIGURE 2. **The summation of exposure in each BonusMalusGroup and the claim frequency (sum of claim numbers / sum of exposure) in each BonusMalusGroup with 90% confidence interval (1.645 sigma).**

1) It is derived from the likelihood function of the Poisson distribution, which naturally models count data [25].
2) It can handle zero-inflated data, common in insurance claims, without the need for data transformation [32].
3) It is scale-invariant, meaning it gives equal weight to relative errors across different magnitudes of the target variable [33].
4) It is particularly adept at dealing with over-dispersion, a condition where variance exceeds the mean, which is common in insurance risk data [34].

In the Motor Third Party Liability (MTPL) dataset, different exposures for each insurance policy are considered. To achieve equitable comparison, we employ exposure-weighted Poisson deviance (E-WPD):

$$E - WPD(M) = \frac{2}{\sum_{i=1}^{N} e_i} \sum_{i=1}^{N} e_i(y_i log(\frac{y_i}{\hat{y}_i}) - (y_i - \hat{y}_i)), \tag{15}$$

where e_i indicates the exposure period of each policy i.

- **Improvement Index:** It measures the relative improvement over the baseline model GLM:

$$II(M) = \frac{PD(M) - PD(\text{INT})}{PD(\text{GLM}) - PD(\text{INT})}, \tag{16}$$

where $PD(\text{INT})$ indicates the Poisson Deviance of the intercept (i.e. no model, just mean risk value of training data).

The Improvement Index is advantageous because:

1) It provides a normalized measure of improvement, allowing for easier interpretation and comparison across different datasets or model iterations [35].
2) It accounts for the difficulty of the prediction task by considering the performance gap between the baseline GLM and the intercept-only model.
3) It is particularly useful in the insurance industry where GLMs are widely used and often serve as a benchmark [36].

4) IMPLEMENTATION DETAILS
Our method is executed using PyTorch and is operational on an NVIDIA GeForce RTX 4090 furnished with 24GB memory. In terms of implementation, we conduct each experiment three times utilizing distinct seeds and present the mean result as the ultimate performance. During each standalone experiment, the model assigned to evaluate insurance risk undergoes training for a cumulative 100 epochs. To mitigate overfitting, we have incorporated an early stopping scheme. The learning rate in the model training phase is set as 1e-4. As for the optimizer which conducts the gradient descent, we choose the Adam optimizer.

B. EVALUATION PERFORMANCE (RQ1)
In response to Q1, we juxtapose our Actuarial Transformer (AT) model's performance against linear models, tree-based models, and deep models in auto insurance risk evaluation, deployed with the Poisson Deviance (PD) and Inconsistency Index (II) metrics. As proposed in Table 4, it is evident that both tree-based and deep models (comprising NNemb, TabNet, and various AT variants) outperform the linear model, Generalized Linear Model (GLM). This suggests that the conventional linear model may fall short in navigating complex relationships and interactions among input features and risk. Yet, it's noteworthy that previous deep models such as NNemb and TabNet don't demonstrate significant advantages over tree-based models like XGBoost, LightGBM, and CatBoost. Though deep models have seen success in various scenarios, their application to auto insurance risk evaluation still demands substantial improvement and customization. Furthermore, AT versions integrated with different tree models in the residual learning module consistently exhibit significant advances compared to various traditional tree-based and deep model baselines. This substantiates that our introduced approach effectively escalates performance in auto risk evaluation on both PD and II metrics, suggesting its potential for broader application within the auto insurance sector.

TABLE 4. The evaluation performance of various baselines and our AT's different versions with different tree models on the auto insurance dataset.

Method	MTPL	
	PD	II
GLM	29.58%	0.00%
XGBoost	29.01%	127.4%
LightGBM	29.06%	126.1%
CatBoost	29.08%	122.5%
NNemb	29.12%	120.0%
TabNet	29.07%	124.4%
AT(XGBoost)	28.91%	143.5%
AT(LightGBM)	28.98%	140.9%
AT(CatBoost)	28.94%	141.8%

C. ABLATION STUDY (RQ2)
To address Research Question 2 (RQ2), we investigate the impact and contribution of each module in our proposed framework by sequentially excluding them and observing the resultant performance alteration. The findings are showcased in Table 5, where "SA" represents the self-attention module and "RM" denotes the residual modeling module. Analysis of the table reveals that for any AT variant integrated with a specific tree model, the exclusion of either SA or RM leads to an increase in Poisson Deviance (PD) and a decrease in Inconsistency Index (II), indicating a deterioration in auto risk evaluation performance. This highlights the critical role both the self-attention module, responsible for capturing feature interactions, and the residual modeling module, aimed at further reducing evaluation error, play in enhancing performance. Furthermore, the results indicate a more pronounced decline in performance upon the removal of the RM module, suggesting its substantial positive influence on final evaluation accuracy.

TABLE 5. The ablation study results for different versions of AT with different tree models.

Method	MTPL	
	PD	II
AT(XGBoost)-SA	28.98%	127.2%
AT(XGBoost)-RM	29.02%	126.7%
AT(XGBoost)	28.91%	143.5%
AT(LightGBM)-SA	29.03%	126.5%
AT(LightGBM)-RM	29.06%	126.1%
AT(LightGBM)	28.98%	140.9%
AT(CatBoost)-SA	29.04%	126.3%
AT(CatBoost)-RM	29.08%	122.5%
AT(CatBoost)	28.94%	141.8%

D. HYPERPARAMETER ROBUSTNESS (RQ3)
To address Research Question 3 (RQ3), we examine the impact of three hyperparameters–embedding size, batch size, and regularization weight–on the precision of auto risk evaluation in our framework. Initially, the embedding size within our Actuarial Transformer (AT) model is varied from 2 to 30, with experiments conducted accordingly. Empirical results, as depicted in Table 6, indicate that, despite the model's version or the tree models it is combined with, changes in embedding size do not significantly alter the final performance. For instance, in the AT(XGBoost) model, the PD metric fluctuates minimally between 28.90 Similarly, Table 7 explores the effect of varying batch sizes on AT model performance, illustrating negligible variations in PD and II metrics across AT(XGBoost), AT(LightGBM), and AT(CatBoost) in comparison to their baseline improvements. Remarkably, adjustments in the batch size for AT(CatBoost), ranging from 1,000 to 20,000, consistently surpassed the 140 Moreover, we probe the AT model's robustness against alterations in the regularization weight, set between 1e-5 to 1e-3, to counteract or mitigate overfitting during training. The outcomes, presented in Table 8, affirm the model's stability against diverse regularization weight settings. However, it is essential to recognize that both excessively high and low regularization weights can diminish performance, notably for AT(LightGBM). Excessive regularization can dominate

the loss function, leading to model oversimplification and underfitting, thus impairing its predictive accuracy. Conversely, overly low regularization weights result in underregularization, risking overfitting as the model becomes too tailored to the training data, capturing noise as patterns and showing high variance, thereby affecting its reliability and performance on new data. This issue is particularly pronounced in self-attention-based transformer models like ours.

TABLE 6. The hyperparameter robustness analysis on the embedding size.

Methods		2	5	10	20	30
AT(XGBoost)	PD	28.91%	28.90%	28.90%	28.93%	28.92%
	II	143.5%	144.2%	144.7%	143.1%	142.6%
AT(LightGBM)	PD	28.98%	28.96%	28.95%	29.01%	28.99%
	II	140.9%	141.5%	141.7%	127.4%	128.8%
AT(CatBoost)	PD	28.94%	28.92%	28.89%	28.95%	28.97%
	II	141.8%	142.6%	144.7%	141.7%	141.2%

TABLE 7. The hyperparameter robustness analysis on the batch size.

Methods		1000	2000	5000	10000	20000
AT(XGBoost)	PD	28.99%	28.94%	28.90%	28.91%	28.95%
	II	135.9%	141.7%	144.5%	143.5%	141.6%
AT(LightGBM)	PD	28.97%	28.99%	28.98%	28.98%	29.00%
	II	136.5%	138.3%	141.2%	140.9%	132.8%
AT(CatBoost)	PD	28.96%	28.95%	28.93%	28.94%	28.97%
	II	140.8%	141.7%	142.2%	141.8%	141.5%

TABLE 8. The hyperparameter robustness analysis on the regularization weight.

Methods		1e-5	5e-5	1e-4	5e-4	1e-3
AT(XGBoost)	PD	28.92%	28.90%	28.91%	28.93%	28.96%
	II	142.9%	144.2%	143.5%	142.3%	140.8%
AT(LightGBM)	PD	29.00%	28.98%	28.98%	28.99%	29.01%
	II	132.7%	135.2%	140.9%	136.4%	129.8%
AT(CatBoost)	PD	28.98%	28.96%	28.94%	28.92%	28.95%
	II	138.9%	140.6%	141.8%	142.3%	141.2%

E. CONVERGENCE ANALYSIS (RQ4)

To address Research Question 4 (RQ4), we monitoring the training progression of our Actuarial Transformer (AT) model, specifically the AT(CatBoost) variant, and juxtapose its behavior against the NNemb baseline. The associated training trajectories are illustrated in Figure 3. Analysis of these curves reveals that the smoothness of our AT model's training process is somewhat inferior to that of the NNemb baseline, suggesting a lesser degree of stability. However, the training loss of AT model diminishes at a consistent rate, in contrast to the NNemb, which exhibits an initial rapid decline that gradually tapers off. Yet, both models approximately converge at the 500-epoch mark. Furthermore, upon achieving convergence, our AT model records a stable loss value of 0.1516, notably outperforming the NNemb baseline's 0.1545. This superior margin aligns with our AT model's enhanced performance on the test set, as corroborated by the data in Table 4.

F. INTERPRETABILITY ANALYSIS (RQ5)

1) INTRODUCTION TO THE SHAP TOOLBOX

The SHAP (SHapley Additive exPlanations) framework,[2] introduced by Lundberg and Lee (2017), is a robust set of tools designed to enhance the interpretability of machine learning models. Rooted in game theory, specifically the concept of Shapley values, SHAP quantifies the contribution of each feature to individual model predictions. This approach facilitates a nuanced understanding of complex model decision-making processes. SHAP conceptualizes each feature value of an instance as a "player" in a cooperative game, with the model's prediction serving as the "payout." The Shapley value methodology ensures an equitable distribution of this "payout" among the features, thereby elucidating their individual contributions to the model's output. Widely adopted across diverse domains including finance, healthcare, and marketing, SHAP provides transparent explanations for predictions generated by sophisticated machine learning algorithms. This transparency not only fosters trust among practitioners and stakeholders but also underscores the framework's robust theoretical foundations and its practical efficacy in elucidating complex models for both technical and non-technical audiences.

2) ANALYSIS OF FEATURE IMPORTANCE

We employed the SHAP framework to ascertain the relative importance of different features in our Actuarial Transformer (AT) model, as illustrated in Figure 4. The analysis reveals that the BonusMalus feature exhibits the most significant positive importance, making the largest contribution to the model's performance. Conversely, the importance of other features to the final prediction in our AT model is notably weaker, and in some cases, negative. This stark contrast can be attributed to several underlying factors that underscore the significance of BonusMalus in predicting insurance risk levels.

3) THE NATURE OF BONUSMALUS

The Bonus-Malus system [37], a cornerstone in automobile insurance risk assessment, is a merit-based rating mechanism designed to adjust policyholders' premiums based on their driving history and claim experience. This system operates on a scale where policyholders receive a "Bonus" for claim-free periods, resulting in reduced premiums, and a "Malus" for claims made, leading to increased premiums. This mechanism serves to incentivize safer driving practices and reward policyholders who maintain a clean driving record, while penalizing those with a history of claims. The Bonus-Malus system is predicated on the principle that past behavior is a strong predictor of future risk, thereby aligning premiums more accurately with individual risk profiles.

[2]https://shap.readthedocs.io/en/latest/

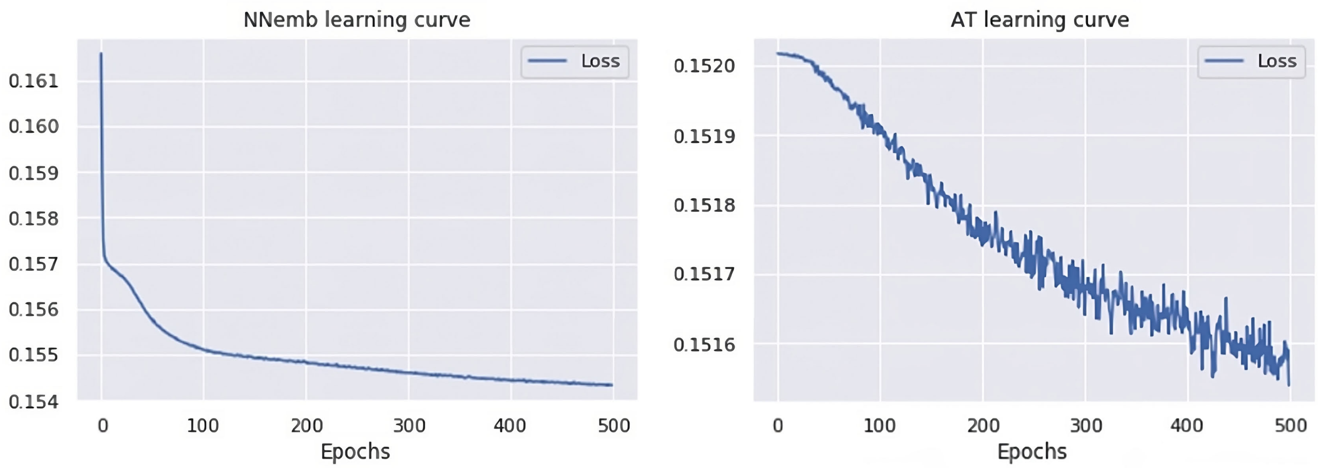

FIGURE 3. **Training loss curves of NNemb baseline and our proposed AT model.**

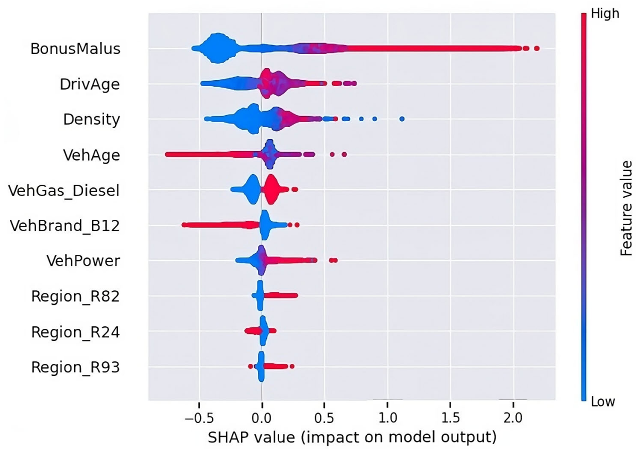

FIGURE 4. **The feature importance of different features for our AT model obtained with SHAP toolbox.**

4) IMPLICATIONS WITHIN AUTO INSURANCE RISK CONTEXT

We analyze the implications of BonusMalus within the context of auto insurance risk from the following dimensions:

- **Direct Correlation with Driver Behavior:** *Rationale:* BonusMalus directly reflects a driver's historical claim behavior, encapsulating penalties (Malus) for claims made and rewards (Bonus) for claim-free periods. *Impact:* Safe driving behavior, indicated by a low BonusMalus score, strongly correlates with a lower risk of future claims. Conversely, a high BonusMalus score signals a history of claims, suggesting elevated risk. This dichotomy enhances its predictive power.
- **Actuarial Foundation:** *Rationale:* The insurance industry extensively employs the BonusMalus system as a key actuarial tool, grounded in the principle that past behavior predicts future incidents. *Impact:* Given its widespread acceptance and actuarial significance, models incorporating the BonusMalus feature likely align closely with established insurance risk assessment practices, enhancing their predictive accuracy.
- **Quantification of Risk:** *Rationale:* BonusMalus serves as a quantifiable and dynamic metric of risk evolution over time, adjusting as new data on driver behavior becomes available. *Impact:* The dynamic nature of the BonusMalus feature allows the model to adapt to changing risk levels more accurately than static features, ensuring its Shapley value remains consistently high.
- **Comprehensive Risk Factor:** *Rationale:* While many features might indicate risk from different perspectives (e.g., age, vehicle type), BonusMalus encapsulates the cumulative effect of these factors as reflected in the driver's history. *Impact:* This comprehensive outlook renders BonusMalus a condensed measure of risk, synthesizing the effects of various underlying risk factors into a single, impactful score.
- **Enhancing Model Interpretability and Transparency:** *Rationale:* The clear, direct link between BonusMalus scores and risk perception makes this feature valuable not only for prediction but also for interpretability. *Impact:* High Shapley values for BonusMalus not only underscore its predictive power but also promote transparency in the model's assessment process, crucial for stakeholder trust and regulatory compliance.

The high Shapley value of the BonusMalus feature in our auto insurance risk evaluation model underscores its pivotal role in capturing and quantifying driver risk. This prominence derives from its direct reflection of driver behavior, its foundational basis in actuarial science, the comprehensive nature of the risk it represents, and its contribution to both model accuracy and interpretability. By leveraging the rich, historical context of drivers' claims history encoded in the BonusMalus system, our AT model is better equipped to predict future claims accurately, rendering BonusMalus an indispensable feature in risk evaluation.

V. CONCLUSION AND FUTURE WORKS

A. CONCLUSION

This study presents the Actuarial Transformer (AT), a novel model designed to improve auto insurance risk evaluation by integrating transformer architecture with tree models. The AT effectively handles the complexity of insurance datasets, which include both continuous and categorical variables, and provides an interpretable decision-making framework. Unlike traditional models that struggle with nonlinear relationships and variable interactions, the AT dynamically captures these complexities through its attention attribution mechanism, enhancing transparency and understanding for stakeholders.

Our experiments on a representative public auto insurance risk dataset confirm the AT's superior accuracy in risk assessment, outperforming existing models. The AT's use of SHAP values ensures that the model's predictions are not only accurate but also interpretable, addressing the common issue of black-box models in machine learning.

However, the study acknowledges limitations, including potential overfitting on highly specific datasets and the computational cost associated with the Transformer's self-attention mechanism. Future research should focus on optimizing these aspects to enhance the model's scalability and efficiency.

In conclusion, the AT represents an advancement in the field of auto insurance risk evaluation, offering both enhanced predictive accuracy and interpretability. Its development underscores the importance of transparent and accurate risk assessment models in the insurance industry.

B. FUTURE WORKS

The Actuarial Transformer (AT) model presents numerous opportunities for future research and development. These include expanding its application to other insurance types beyond auto insurance, advancing attention mechanisms to incorporate temporal dynamics, integrating with emerging technologies like IoT and telematics for real-time data streams, enhancing algorithmic transparency and interpretability, extending to multi-task learning frameworks for simultaneous prediction of various risk aspects, improving computational efficiency for scalability, and strengthening robustness against overfitting. These avenues for improvement underscore the ongoing journey in perfecting insurance risk evaluation models, with AT serving as a milestone in the application of sophisticated machine learning techniques to actuarial science and insurance analytics.

REFERENCES

[1] E. deHAAN, N. Li, and F. S. Zhou, "Financial reporting and employee job search," *J. Accounting Res.*, vol. 61, no. 2, pp. 571–617, May 2023.

[2] S. Colella and H. Jones, "Machine learning and ratemaking: Assessing performance of four popular algorithms for modeling auto insurance pure premium," in *Proc. CAS E-Forum*, 2023, pp. 1–15.

[3] K. Sudhir, "Competitive pricing behavior in the auto market: A structural analysis," *Marketing Sci.*, vol. 20, no. 1, pp. 42–60, Feb. 2001.

[4] S. Dominique-Ferreira, H. Vasconcelos, and J. F. Proença, "Determinants of customer price sensitivity: An empirical analysis," *J. Services Marketing*, vol. 30, no. 3, pp. 327–340, May 2016.

[5] K. Weerasinghe and M. Wijegunasekara, "A comparative study of data mining algorithms in the prediction of auto insurance claims," *Eur. Int. J. Sci. Technol.*, vol. 5, no. 1, pp. 47–54, 2016.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 21–25.

[7] S. Matthews and B. Hartman, "MSHAP: SHAP values for two-part models," *Risks*, vol. 10, no. 1, p. 3, Dec. 2021.

[8] C. Dutang, A. Charpentier, and M. C. Dutang. (2020). *Package Cas-datasets*. [Online]. Available: https://www.openml.org/search

[9] M. Hanafy and R. Ming, "Machine learning and deep learning: A review for insurers," *J. Risk Financial Manage.*, vol. 14, no. 6, p. 266, 2021.

[10] M. Paredes, "A case study on reducing auto insurance attrition with econometrics, machine learning, and A/B testing," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2018, pp. 410–414.

[11] S. Fujita, T. Tanaka, K. Kondo, and H. Iwasawa, "AGLM: A hybrid modeling method of GLM and data science techniques," in *Actuarial Colloquium Paris*. Dublin, Ireland: Society of Actuaries in Ireland, 2020.

[12] S. Abdelhadi, K. Elbahnasy, and M. Abdelsalam, "A proposed model to predict auto insurance claims using machine learning techniques," *J. Theor. Appl. Inf. Technol.*, vol. 98, no. 22, pp. 3428–3437, 2020.

[13] J. Pesantez-Narvaez, M. Guillen, and M. Alcañiz, "Predicting motor insurance claims using telematics data—XGBoost versus logistic regression," *Risks*, vol. 7, no. 2, p. 70, Jun. 2019.

[14] K. C. Dewi, H. Murfi, and S. Abdullah, "Analysis accuracy of random forest model for big data—A case study of claim severity prediction in car insurance," in *Proc. 5th Int. Conf. Sci. Inf. Technol. (ICSITech)*, Oct. 2019, pp. 60–65.

[15] C. V. Colin, L. Ding, E. Ressouche, J. Robert, N. Terada, F. Gay, P. Lejay, V. Simonet, C. Darie, P. Bordet, and S. Petit, "Successive incommensurate spin orderings and excitations in multiferroic SrMnGe2O6," 2019, *arXiv:1912.10935*.

[16] S. Matthews and B. Hartman, "Machine learning in ratemaking, an application in commercial auto insurance," *Risks*, vol. 10, no. 4, p. 80, Apr. 2022.

[17] M. V. Wüthrich and M. Merz, "Yes, we CANN!" *ASTIN Bull., J. IAA*, vol. 49, no. 1, pp. 1–28, 2019.

[18] A. Gabrielli, R. Richman, and M. V. Wüthrich, "Neural network embedding of the over-dispersed Poisson reserving model," *Scandin. Actuarial J.*, vol. 2020, no. 1, pp. 1–29, Jan. 2020.

[19] G. A. Spedicato, C. Dutang, and L. Petrini, "Machine learning methods to perform pricing optimization. A comparison with standard GLMs," *Variance*, vol. 12, no. 1, pp. 69–89, 2018.

[20] T. Chen, L.-A. Tang, Y. Sun, Z. Chen, and K. Zhang, "Entity embedding-based anomaly detection for heterogeneous categorical events," 2016, *arXiv:1608.07502*.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–21.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[24] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[25] P. McCullagh and J. A. Nelder, *Generalized Linear Models*. Boca Raton, FL, USA: CRC Press, 1989.

[26] A. Charpentier, *Computational Actuarial Science With R*. Boca Raton, FL, USA: CRC Press, 2014.

[27] T. J. Hastie and D. Pregibon, "Generalized linear models," in *Statistical Models in S*. Evanston, IL, USA: Routledge, 2017, pp. 195–247.

[28] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[29] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[30] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, ‘‘CatBoost: Unbiased boosting with categorical features,’’ in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 6639–6649.

[31] K. McDonnell, F. Murphy, B. Sheehan, L. Masello, and G. Castignani, ‘‘Deep learning in insurance: Accuracy and model interpretability using TabNet,’’ *Expert Syst. Appl.*, vol. 217, May 2023, Art. no. 119543.

[32] C. Czado, T. Gneiting, and L. Held, ‘‘Insurance claim modelling with application to the collective risk model,’’ *Bull. Swiss Assoc. Actuaries*, vol. 2, pp. 129–147, Jan. 2007.

[33] B. Jørgensen, ‘‘Exponential dispersion models,’’ *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 49, no. 2, pp. 127–145, Jan. 1987.

[34] J. M. Hilbe, *Negative Binomial Regression*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[35] M. Eling, ‘‘Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models?’’ *Insurance, Math. Econ.*, vol. 51, no. 2, pp. 239–248, Sep. 2012.

[36] E. Ohlsson and B. Johansson, *Non-Life Insurance Pricing With Generalized Linear Models*. Cham, Switzerland: Springer, 2010.

[37] J. Lemaire, *Bonus-Malus Systems in Automobile Insurance*, vol. 19. Cham, Switzerland: Springer, 2012.

TIEJIANG SUN (Member, IEEE) received the degree from Chang’an University, in 2014. He is currently an experienced Software Engineer. Since then, he has been driving innovation with Ping An Property and Casualty Insurance Company of China Ltd. He is the Manager of the Analytical and Decision-Making Information Systems Department, he has successfully applied AI and big data technologies to revolutionize property insurance management.

JINGYUN YANG received the Master of Science degree in computational finance from Carnegie Mellon University, in 2024. His research interests include the application of artificial intelligence in finance and economics, covering areas, such as quantitative trading, pricing models, financial decision-making, and quantitative economics.

JIALE LI received the bachelor’s degree in statistics from Shandong University, and the master’s degree in applied urban science and informatics from NYU. She is currently a Data Analyst. She has professional experience as a Technical Analyst with Wayfair and a Data Analyst Internship. She has expertise in statistical modeling, machine learning, data visualization, and analytics using tools, such as Python, SQL, and Tableau.

JIAYING CHEN received the Master of Professional Studies degree in management from Cornell University, in 2022. Her research interests include utilizing AI and advanced data analytics to optimize financial reporting, streamline accounting processes, enhance tax compliance, and improve decision-making across various sectors.

MINGYUE LIU received the M.Eng. degree in computer science from Cornell University, in 2022. She possesses three years of experience in software engineering and artificial intelligence.

LI FAN received the Ph.D. degree in computer science from Nankai University, in 2013. He completed his postdoctoral research at the joint research laboratory of Zhejiang University and Eastcom Company Ltd. His research interests include deep learning, software and hardware codesign for AI systems, AI compilation stacks, and AI algorithms and AI applications.

XUKANG WANG (Member, IEEE) is currently a Research Scientist with the Sage IT Consulting Group, leading his team studying AI and its applications. His research interests include deep learning, blockchain, and privacy.