Asia Pacific Journal of Information Systems
Vol. 34 No. 2 (June 2024), 565-584

ISSN 2288-5404 (Print) / ISSN 2288-6818 (Online)
https://doi.org/10.14329/apjis.2024.34.2.565

Improving Automobile Insurance Repair Claims Prediction Using Gradient Decent and Location-based Association Rules

Seongsu Jeong[a], Jong Woo Kim[b,*]

[a] *Ph.D. Candidate, The Business Informatics, Hanyang University, Korea*
[b] *Professor, School of Business, Hanyang University, Korea*

ABSTRACT

More than 1 million automobile insurance repairs occur per year globally, and the related repair costs add up to astronomical amounts. Insurance companies and repair shops are spending a great deal of money on manpower every year to claim reasonable insurance repair costs. For this reason, promptly predicting insurance claims for vehicles in accidents can help reduce social costs related to auto insurance. Several recent studies have been conducted in auto insurance repair prediction using variables such as photos of vehicle damage. We propose a new model that reflects auto insurance repair characteristics to predict auto insurance repair claims through an association rule method that combines gradient descent and location information. This method searches for the appropriate number of rules by applying the gradient descent method to results generated by association rules and eventually extracting main rules with a distance filter that reflects automobile part location information to find items suitable for insurance repair claims. According to our results, predictive performance could be improved by applying the rule set extracted by the proposed method. Therefore, a model combining the gradient descent method and a location-based association rule method is suitable for predicting auto insurance repair claims.

Keywords: Big Data, Automobile Insurance, Insurance Claim Prediction, Association Rules, Gradient Descent

Ⅰ. Introduction

Automobile insurance is a contract to prepare for an accident that occurs while using an automobile and is designed to compensate for damage to the body or property due to an accident. For this reason, auto insurance not only reduces the burden on insurance subscribers due to car accidents but also reduces the burden of medical expenses on patients. To secure a minimum safety net to protect the lives and property of others, governments in many countries actively encourage or require drivers to purchase

*Corresponding Author. E-mail: kjw@hanyang.ac.kr

automobile insurance. Globally, the auto insurance market continues to grow every year. According to IBISWorld, the scale of auto insurance premiums in the United States is expected to be $330.4 billion in 2023, having grown by 0.8% in 2022. China also predicts that auto insurance premiums will grow from about $212.1 billion in 2021 to $304.4 billion by 2026. In Korea, the size of the auto insurance market in 2021 was $15.86 billion, an increase of 11.6% in comparison to 2019. Insurers pursue a variety of strategies to secure significant returns in ever-growing auto insurance markets, and competition continues to intensify. Therefore, it will be vital for insurance companies to reduce related costs and maintain cost advantages in comparison to other companies to earn profits in the auto insurance industry.

From the point of view of an insurance company trying to maintain its auto insurance business, calculating repair costs in a reasonable and efficient way is a critical factor in company management because most insurance companies not only employ many experts to calculate auto insurance repair costs but also have huge computer systems to process claims and assess damages quickly. Therefore, being able to efficiently calculate auto insurance repair costs can greatly impact pricing strategies in the long run. In addition, by efficiently calculating repair costs, insurance companies can gain a price advantage by lowering premiums and reducing costs for business operations. Along with reducing insurance premiums for consumers, these strategies have meaningful social and economic benefits.

Although insurance companies are making steady efforts to predict insurance repair costs, they are experiencing difficulties for practical reasons such as manpower and cost. In general, insurance companies hire professionals to calculate insurance repair costs. However, millions of claims are filed every year from repair shops to insurance companies. Because specialists must handle many claims at once, time pressure means the claims are not always carefully managed. In addition, in cases where it is difficult to accurately review a claim due to insufficient experience on the part of professional personnel, the appropriateness of insurance repair costs cannot be properly judged. Even when professional personnel have sufficient experience, any judgment of whether insurance repair costs are appropriate is subjective. A single claim can lead to different judgments depending on individual employees.

To tackle the issues associated with calculating the cost of insurance repairs, experts in both the insurance industry and academia have explored prediction systems that relate to auto insurance. In order to detect insurance fraud, various techniques have been employed such as general linear models, classical complex-Poisson models, and Bayesian approaches; all of which adopt a predictive approach. Numerical data relating to policyholder characteristics, accident location, and car brand are analyzed in auto insurance-related data. Nevertheless, in-depth analysis of auto insurance repair costs involves the use of machine learning methods to predict the value of insurance claims. In this study, we propose an insurance repair item prediction model employing the association rule technique that has not been previously explored in existing methods. The auto insurance repair data used herein demonstrate a strong correlation between the damaged and repaired vehicle parts. In this way, this study creates a model using a location-based filter that considers the positional relationship between parts and enhances the efficiency of the association rule while factoring in collision and damage-related insurance properties. We assessed the prediction model's performance by con-

trasting it with actual claim items, employing diverse measures such as F1, based on gradient descent and location-based association rules. Additionally, we used a gradient descent method to determine the appropriate number of association rules. This study offers practical and academic value as it contributes to predicting auto insurance repair costs and advances a suitable method for auto insurance repair through association rules and gradient descent.

The structure of the paper is as follows. In section 2, related works are described including automobile insurance prediction models and association rules. In section 3, the proposed approach explains the current insurance claim prediction model, the framework to improve insurance claim prediction and auto insurance repair data, and the proposed approach that generates association rules with the gradient descent method and a location-based filter. In section 4, experimental design and results describe the experimental design for implementing our proposed model and the results of our experiments. Finally, in section 5, Discussion and conclusion explain the contributions and limitations of our research.

Ⅱ. Related Work

2.1. Automobile Insurance Claim Prediction Models

From the standpoint of an auto insurer, effective claims prediction is vital because it is directly related to the loss ratio in business operations. Accordingly, existing research focuses on insurance prediction models. Past models can be broadly classified into three types. The first type uses a classic regression model and a multidimensional classification method. The second type is developed from classical regression and classification methods such as double generalization and Bayesian models, and the third type is the ensemble models that improve prediction performance by integrating various methods such as regression and classification.

Regarding classical models concerning insurance claims prediction, Bailey and Simon (1960) suggested a minimum prejudice procedure as a pricing technique for multi-dimensional classification. However, this approach does not include a statistical evaluation of the model. Subsequently, Nelder and Wedderburn (1972) employed the framework of the generalized linear model (GLM) in the analysis of insurance data. Presently, this approach stands as the conventional technique utilized in the insurance sector for assessing

<Table 1> Previous Research Table

Model Type	Authors	Techniques/Methods
Classical Regression	Bailey and Simon (1960)	Minimum bias procedure for multi-dimensional classification
	Nelder and Wedderburn (1972)	Generalized Linear Model (GLM) framework
	Jørgensen and De Souza (1994)	Complex Poisson-Gamma model
Extension of Classical Methods	Smyth and Jørgensen (2002)	Double-generalized linear model
	Gschlößl and Czado (2007)	Fully Bayesian approach
	Czado et al. (2012)	Copular models
Ensemble Methods	Liu et al. (2014)	Multi-class AdaBoost tree
	Zhang (2019)	Regression tree model with GLM using boosting algorithm
	Jain et al. (2019)	ANN and gradient boosting algorithm XGBoost

claim costs. Furthermore, Jørgensen and De Souza (1994) introduced a traditional compound Poisson-Gamma model, postulating an average claim expense and an uncorrelated count of claims conforming to Poisson and gamma distributions. Secondly, in the realm of models extending conventional methodologies for forecasting insurance claims, Gschlößl and Czado (2007) revisited the Poisson-Gamma paradigm to account for interdependence between claim counts and sizes via a fully Bayesian strategy. Additionally, Smyth and Jørgensen (2002) implemented a dual generalized linear model for situations involving solely cost outlays without considering frequency. Further research has proposed alternative frameworks such as quantile regression (Heras et al., 2018), hierarchical modeling (Frees and Valdez, 2008), machine learning (Kaš´celan et al., 2015; Yang et al., 2016), and copula models (Czado et al., 2012) for predicting insurance outcomes.

Another category of models utilized in insurance prediction involves ensemble learning techniques. In a comprehensive examination, Wolpert leveraged the predictions from the initial base learner to create meta-features for generating collective predictions in subsequent layers. Alternative methodologies proposed include stacked generalization (Wolpert, 1992). Friedman (2001) advocated for gradient-boosting machines, emphasizing their ability to produce precise predictions through the amalgamation of weighted learners. Yang (2001) introduced adaptive regression by mixing (ARM), a weighted averaging approach suitable for various analytical tasks, including regression and prediction. Furthermore, Guelman (2012) employed gradient boosting tree (GBT) ensemble learning to forecast both claim frequencies and intensities, employing techniques like under-sampling and cross-validation to mitigate issues of data imbalance. Liu et al. (2014) utilized a multi-class AdaBoost tree to predict charge intensities, comparing outcomes with those derived from GLM, two-layer back propagation (BP) neural networks, and support vector machine (SVM) methodologies. Zhang (2019) investigated prevalent claims in the Chinese auto insurance domain, integrating regression tree models with GLM via boosting algorithms, while Jain et al. (2019) scrutinized factors to propose appropriate policies based on risk assessments.

In addition to the aforementioned models, Pesantez-Narvaez (2019) introduced a novel approach to forecast auto insurance claims by leveraging telematics data, while Singh (2019) employed images depicting damaged vehicles to estimate insurance claim values. Arief Fauzan et al. (2018) investigated the application and efficacy of XGBoost in handling challenges related to data volume and missing values within claims prediction tasks. Their study encompassed a comparative analysis between XGBoost and various ensemble learning techniques, including AdaBoost, stochastic gradient boosting (GB), random forest, and online learning-based methods like neural networks. Furthermore, Kowshalya and Nandhini (2018) utilized data mining methodologies to anticipate fraudulent claims and determine premium rates tailored to individual customers' personal and financial attributes. Their research demonstrated the effectiveness of three distinct classifiers in detecting fraudulent claims.

In recent years, there has been a surge of interest in machine learning methodologies, especially artificial neural networks (ANNs), in the field of insurance pricing. Prominent research by Fialova and Folvarcna (2020), Gao and Wüthrich (2018), Sun et al. (2017), and Wuthrich (2019) has delved into the application of neural network techniques within this sector.

Furthermore, Yunos et al. (2016) utilized a back-propagation neural network (BPNN) model to examine both claim frequency and severity, highlighting the BPNN's capacity to capture nonlinear relationships inherent in loss data. Additionally, Bhowmik (2011) demonstrated the efficacy of predictive modeling techniques from machine learning in identifying instances of insurance claims fraud. However, despite the promising outcomes, the intricate nature of these models, particularly ANNs and other advanced architectures, has impeded their widespread adoption by insurance companies and regulators. ANNs, especially those with multiple hidden layers, are often regarded as black boxes, presenting challenges in interpreting the influence of input variables on output predictions (Olden and Jackson, 2002).

Previous models for predicting auto insurance claims have focused on enhancing the accuracy of detection by analyzing various numerical and categorical data. However, it is challenging to extract and employ concealed attributes such as repair relationship characteristics and parts when creating an insurance claim prediction model that only analyzes some of the total auto insurance repair data because several items in insurance repair data are based on various parts present in a car, and it is imperative to analyze the relationship between these parts to predict claims accurately. Moreover, creating a predictive model requires expertise in automobile repair because there are differences in the types of parts or fastening structures among vehicles. To address this, we studied a method of predicting auto insurance repair claims using the association rule approach that efficiently reflected the characteristics of auto parts and repair relationships.

2.2. Association Rule Analysis

The main objective of association rule analysis is to detect rules among items that co-occur in transactions by scrutinizing data. Agrawal et al. (1993) proposed a prediction model using association rules between conditional clauses and resultant clauses in data composed of categorical variables. More specifically, Agrawal et al. (1993) proposed the Apriori algorithm, which extracts frequent item sets with greater than minimum support. This method, however, has a limitation in that the time required for calculation exponentially increases as the number of items increases. According to research by Agrawal et al. (1993) and Agrawal and Srikant (1994), the Apriori method is flawed because it generates many duplicate candidates by scanning the database several times as the number of items increases. Han et al. (2000) and Liu et al. (2004) adopted pattern-growth and association rule mining methods to overcome this limitation of the Apriori algorithm. Pattern-growth and association rule mining methods make it possible to efficiently analyze even increasing numbers of items by constructing frequent pattern (FP) trees in a compressed form without generating a candidate set and mining a set of frequent items.

Goting (2004) suggested a method for detecting anomalies through association rules using the *a priori* algorithm that algorithm identifies infrequent sets, and an anomaly score is generated based on association rules between categorical variables in the data. The anomalies are then classified using a threshold-based algorithm. Otey et al. (2006) and Koufakau et al. (2010) proposed an outlier detection model that handles both categorical and numeric variables.

When analyzing auto insurance repair claims data using previous methods such as apriori and pattern-growth algorithms, performances depend on assigned minimum support values. Setting the appropriate minimum support to achieve optimal perform-

ance requires a great deal of manpower and time. Therefore, this study proposes a gradient descent and location-based association rule method that reflects vehicle characteristics. The method uses an FP-growth algorithm to find association rules between items in auto insurance repair data and applies a gradient descent and location-based filter to these rules to extract items within a certain distance. By employing this methodology, we were able to systematically identify items closely associated with insurance claims, leveraging the nuanced characteristics inherent in automobile accidents and subsequent repairs. This approach facilitated a more efficient and targeted selection process, enhancing the accuracy and relevance of the identified items in the context of insurance claims analysis.

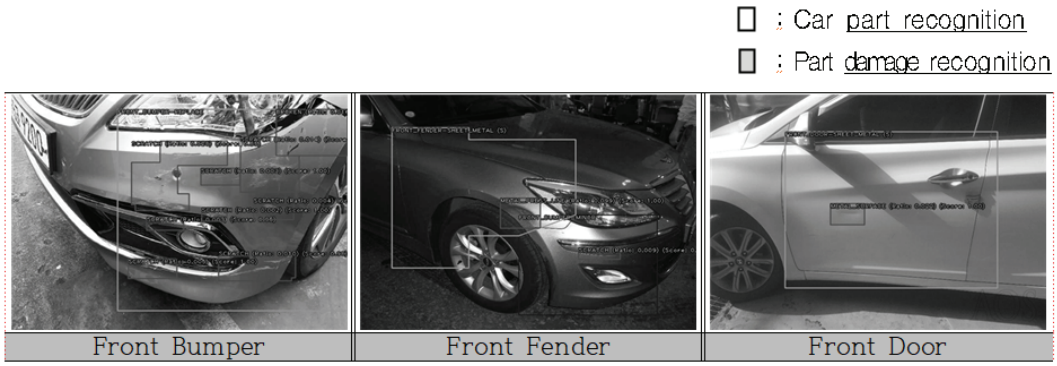
Ⅲ. Proposed Approach

3.1. Current Auto Insurance Repair Item Prediction Model

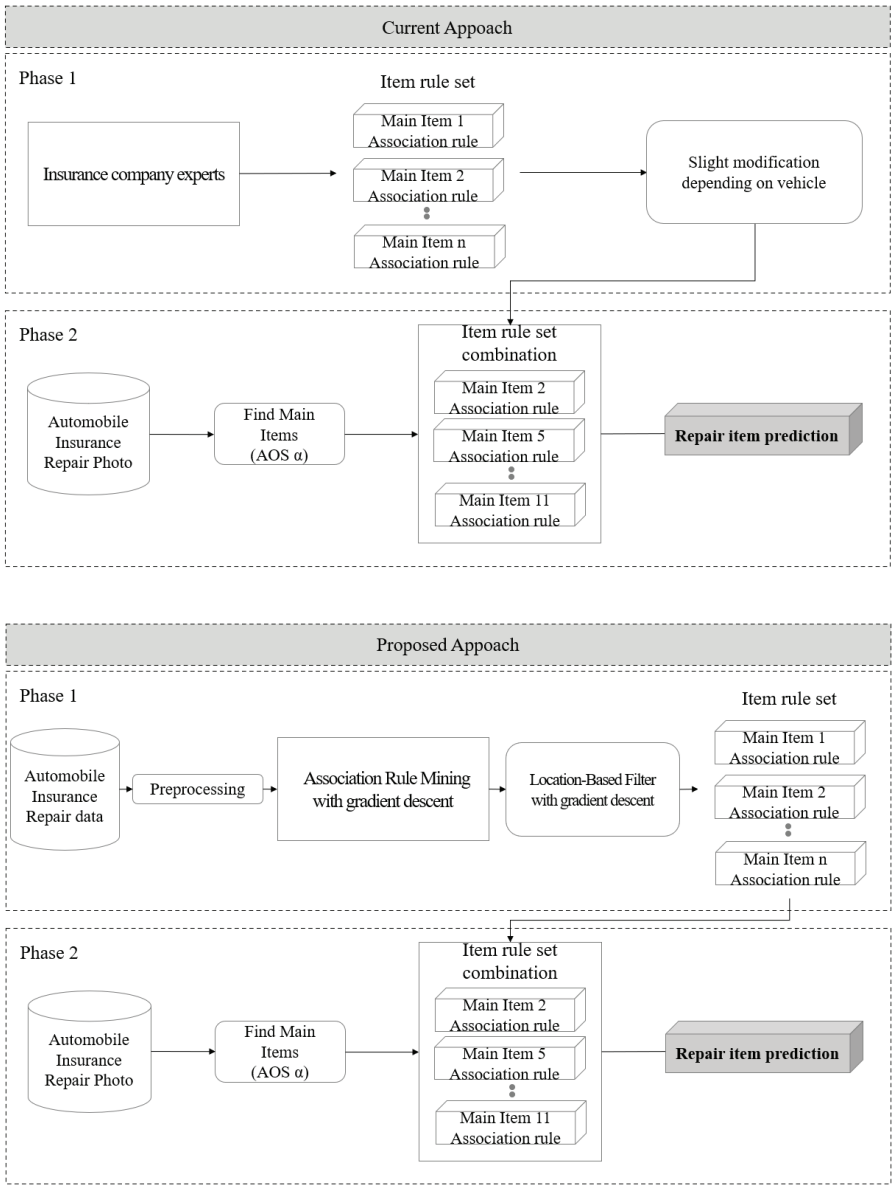
The purpose of this study is to improve the predictive performance of a program known as the automobile repair cost online service alpha (AOS-Alpha), which is used by insurance companies to predict auto insurance repair costs in Korea. AOS-Alpha is a program that recognizes damaged parts of a vehicle through photos and estimates repair costs based on a rule set connected to the damaged parts. As shown in <Figure 1>, AOS-Alpha recognizes a vehicle's exterior parts and damage. Estimated repair costs are predicted by applying a rule set based on this recognition of damaged parts. Inferring repair items from damaged parts, the rule set is the result of the long-term collection of relevant information by insurance company experts. This study aims to improve prediction performance by upgrading the rule set extraction method using gradient descent and location-based association rules.

3.2. Insurance Claim Prediction Framework

In this section, our model predicts auto insurance repair items based on association rules derived from auto insurance repair data through gradient descent and location-based filters. Our model is implemented through the process depicted in <Figure 2>. As previously noted, the rule set generated through the current approach stems from research findings by insurance company specialists during the initial phas-



<Figure 1> AOS-Alpha: Screenshots of Korea's Current Auto Insurance Repair Item Prediction Model



<Figure 2> Auto Insurance Repair Item Prediction Framework

es of AOS Alpha program development, with adjustments made by data practitioners to tailor certain items based on vehicle specifications before application. While this method ensures uniform rule application across vehicles, it faces challenges in fully capturing the repair nuances specific to individual vehicles. Our proposed approach involves predicting items through a ruleset derived from gradient descent and location-based association rules applied to vehicle-specific claims data. This method enables rapid and efficient integration of vehicle characteristics into the prediction process.

The framework of the proposed auto insurance repair item prediction model can be segmented into two main parts. The first part is the generation of gradient descent and location-based association rules for major items according to auto insurance repair data. Here, preprocessing for association rule analysis is initially performed on insurance repair item data, and association rules are generated by applying the gradient descent method to the preprocessed data. Subsequently, location-based filters developed via the gradient descent method are applied to the association rules to determine the rules for the main items. For the second part, AOS-Alpha finds the main items in need of repair in auto insurance claim photos. Then, rules for the main items are combined with the item rule set created in Phase 1. Finally, insurance claim items are predicted with the combined rule set.

3.3. Auto Insurance Repair Data

This experiment was executed using auto insurance repair data obtained from the automobile repair cost online service (AOS) system of the Korean Insurance Development Institute. The accumulated data span from August to November of 2020 and contain full details of car repairs produced by repair shops that worked with insurance companies in paying for repairs. There are a total of 14 variables, as shown in <Table 2>, and approximately 4.2 million records. Among these data, the primary variables used for creating and analyzing association rules were firm, insurance claim number, vehicle model, repair method, and repair items. The remaining nine attributes (e.g., vehicle code, work item number, time, material price, and other remaining attributes) were excluded as they were irrelevant to association rules creation and analysis.

Information used the analysis that consists of work items and time as created by the Korean Insurance Development Institute based on the auto repair work

<Table 2> Data Attributes

No.	Attributes	Description
1	Firm	Insurance firm name
2	Insurance number	Insurance claim number
3	Vehicle maker	Maker of the insured vehicle
4	Vehicle model	Model of the insured vehicle
5	Vehicle code	Code of the insured vehicle
6	Registration number	License plate number of the insured vehicle
7	Repair method	Repair method of the insured vehicle
8	Repair items	Repair items of the insured vehicle
9	Items number (main)	Repair items code (main)
10	Items number (sub)	Repair items code (sub)
11	User text	User input text
12	Time	Repair time
13	Material price	Material cost
14	Total cost	Total amount

item method established by the Korean government. Specifically, work items are standardized by automobile with each vehicle's work items comprising roughly 400 items taking into account parts or work procedures. Additionally, approximately 80% of auto repair shops in Korea use AOS to bill for repairs with claims to the insurance company, so the method has universality. Furthermore, because the majority of our analyzed data is made up of insurance repair data resulting from collision accidents, it can be verified that parts installed on the exteriors of automobiles are first charged.

3.4. Association Rules Generated with the Gradient Descent Method

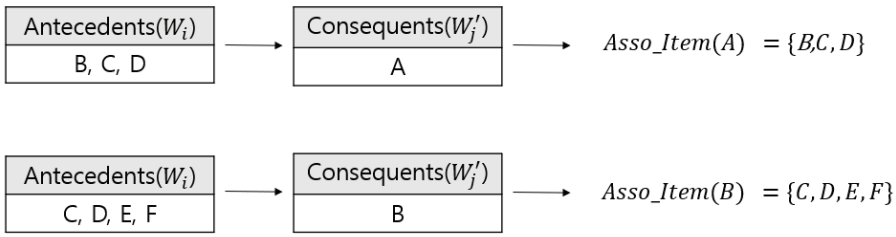
In this study, association rules were made using insurance repair item data for individual vehicles. Because work items for each vehicle are created by reflecting the unique structural characteristics of the vehicle including vehicle shape, type of engine, and options, there are differences in repair item data for each vehicle. For this reason, when a repair item is predicted based on association rules with data about different vehicles, the prediction model may predict a repair item that does not exist for the vehicle.

In creating association rules for individual vehicles, repair item variables, and repair method variables are combined to reflect insurance repair characteristics. Repair methods consist of values such as remove and refit, overhaul, renew, repair, and adjust and describe a set of potential actions in vehicle repair. When repairing a car due to a collision accident, the difference in the repair method and scope of repair depending on the severity of damage exist even for the same item. Accordingly, the repair method can be used as an appropriate variable for analyzing the association rule in detail. Thus, rules are created by combining work items and repair methods for detailed analysis to predict insurance repair items.

To create association rules for individual vehicles herein, we had to distinguish between work items and main work items most in need of repair due to damage recognized by AOS-Alpha. Work items refer to all repair items that may be required for repairing a vehicle at any time. A main work item is located outside the car and is the primary item to be damaged in a collision. In addition, any main work item has sustained damage recognized by AOS-Alpha. The reason for the distinction between work items and main work items is that most car accidents involving insurance claims are caused by collisions, so the external main work items of the car are damaged first. In this way, the relationship between the main work items and work items can be analyzed.

The aforementioned work items and main work items can be described as follows. Among the set of work items W for repairing the car, the first damaged and claimed work item is assigned to the set of main work items, $W^{'}$, corresponding to the outcomes of association rule analysis. The set of work items can be represented by $W = \{w_i | i = 1, \cdots, N\}$. The set of main work items, $W^{'}$ is a subset of W and can be represented by $W^{'} = \{w'_j | j = 1, \cdots, M\}$.

For association rule analysis, when a main work item w'_j is claimed, rules are derived by setting the work item w_i along with the main work item as a conditional clause and the main work item w'_j as a result clause. As a formula, this is expressed as $w_i \rightarrow w'_j$. At this time, to extract optimal association rules for predicting repair items, we find the highest F1 score with *Rule*=$\{w_i \rightarrow w'_j \mid$ *Minimum support= minimum support value that maximizes F1*$\}$ based on the entire claim while changing the minimum support by applying the gradient descent method.



<Figure 3> Create $Asso\_item(w'_j)$

As a result, it is possible to create a set of work items $Asso\_item(w'_j)$ with the main item w'_j as the result clause. This is expressed as formula $Asso\_item(w'_j) = \{w_i | w_i \rightarrow w'_j \in Rule\}$.

<Figure 3> shows the generation of a set of work items B, C, D, E, and F from the main items A and B through association rules with minimum support that maximize the F1 value which is obtained by applying the gradient descent method. The association rule set of main work item A can be represented by $Asso\_item(A) = \{B,C,D\}$, and the association rule set of main work item B can be represented by $Asso\_item(B) = \{C,D,E,F\}$.

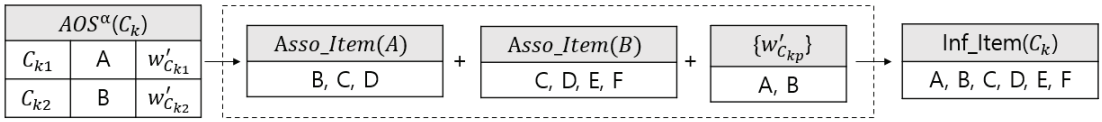
After generating a set of work items $Asso\_item(w'_j)$ based on the relationship between main work items w'_j and work items w_i, it is necessary to identify main work items w'_j among repair items in the auto insurance claim details. An individual claim C_k is a set of work items consisting of repair items extracted from photos using AOS-Alpha, and claimed items C_{kl} are work items claimed for the repair of any specific vehicle. This can be expressed as a subset of work items $C_{kl} \subset W$.

Because the set of claims consists of K individual claims C_k, it can be expressed as $C = \{C_k | k = 1, \cdots, K\}$. Further, because each individual claim C_k consists of L claim items C_{kl}, it can be expressed as $C_{k=}\{C_{kl} | l = 1, \cdots, L\}$. In addition, the set of main work items W'_{ck} extracted through AOS-Alpha from the claim items can be expressed as $W'_{C_k} = \{w'_{C_{kp}} | p = 1, \cdots, P\}$ because the set is composed of main work items $w'_{C_{kp}}$ among P claim items. As a result, the relationship between work item w_i, individual claim C_k, claim item C_{kl}, and main work items among claim items $w'_{C_{kp}}$ can be expressed as $W = \{w_{1,} \cdots w_i\} \supset C_{k=}\{C_{k1}, \cdots C_{kl}\} \supset \{w'_{C_{k1}}, \cdots, w'_{C_{kp}}\} = W'_{C_k}$.

To predict repair claims, the inferred work item set $Inf\_item(C_k)$ is created by extracting $w'_{C_{kp}}$ (main work items obtained through AOS-Alpha among claim items) from individual claims C_k. The set $Inf\_item(C_k)$ is created by deriving $Asso\_item(w'_{C_{kp}})$ based on main work items $w'_{C_{kp}}$ among claim items and merging the work items. Because the merged work item does not include the main item $w'_{C_{kp}}$, an additional $w'_{C_{kp}}$ is added. As a result, the set of claims $Inf\_item(C_k)$ derived from claims C_k can be expressed as $Inf\_item(C_k) = \{w_i | w_i \rightarrow w'_j, Rule\ for\ w'_j \in w'_{C_k}\} \cup \{w'_{C_{kp}}\}$.

<Figure 4> expresses the formula described above as an image. First, A and B corresponding to main work items $w'_{C_{kp}}$ are extracted from claim items C_k through AOS-Alpha, and subsequently, $Inf\_item(C_k)$ is created by adding $Asso\_item(A)$, $Asso\_item(B)$, and $\{w'_{C_{kp}}\}$.

3.5. Location-based Filters with Gradient Descent



<Figure 4> Create $Inf\_item(C_k)$

This study applies the gradient descent method and location-based filters to increase the prediction rate of auto insurance repair items with association rules. As context, it is generally necessary to determine appropriate support, confidence, and lift-setting values for individual vehicles and main work items to generate rules with a high prediction rate in the case of association rules. Therefore, the rates charged for each vehicle and main work items are different, and the performance of a rule is determined according to the set value mentioned above, so it is necessary to specify various input values and repeatedly verify them. These problems cause high costs and a great deal of time in predicting insurance repair items using association rules. Accordingly, effective application of the association rule method is challenging for insurance companies. In addition, for vehicles or items with a small amount of data, there are cases wherein a rule is not inclusive even though an item may be suitable for actual repair prediction. Therefore, our model applies the gradient descent method and location-based filters for objectivity and efficiency in the association rule method. We applied gradient descent based on the F1 scale to set the most appropriate distance value when applying gradient descent and location-based filters.

The idea for location-based filters came from the Research Council for Automobile Repairs (RCAR) bumper test, which determines auto insurance ratings. The RCAR is an international organization of insurance industry-financed research centers whose explicit aim is to maintain reasonable costs for auto insurance repairs. We analyzed the RCAR low-speed crash test conducted by the Korean Insurance Development Institute. Thus, we were able to confirm that auto parts within a certain distance from the point of impact in collision accidents are repairable in both front and rear crash tests of all cars. Based on these results, location-based filters were applied to our predictive model herein.

To use location-based filters, we needed a part corresponding to each work item, its coordinates, and a distance criterion to apply to each filter. The Korean Insurance Development Institute disassembles all the parts installed on a car to calculate the work items and time required to repair all vehicles involved in automobile insurance accidents. Based on these data, we set coordinates of approximately 400 work items for each vehicle in two dimensions.

First in the process of calculating coordinates for work items and distances per item to apply to filters, coordinates for the work item w_i can be expressed as $P(w_i)$, yielding the formula $P(w_i) = (w_{i1}, w_{i2})$. In addition, the distance d between work items to which the Euclidean method is applied can be expressed as $d(w_i, w_j) = \sqrt{(w_{i1} - w_{j1})^2 + (w_{i2} - w_{j2})^2}$.

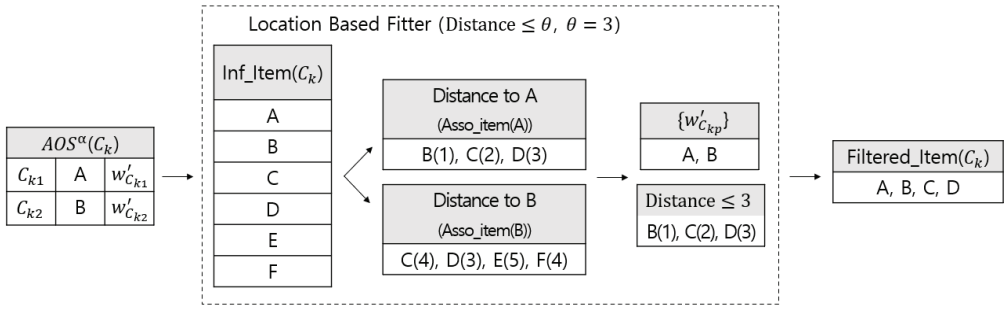
Location-based filters use this method to calculate the distance between the main item set W'_{C_k} and the work item set $Inf\_item(C_k)$ derived from the main item. At this time, the distance between the main item $w'_{C_{kp}}$ and the items belonging to the work item set $Asso\_item(w'_{C_{kp}})$ was calculated, and a value that satisfied the limit θ was extracted using a filter. Finally, the work item set $Filtered\_item_{\theta}(C_k)$ was

derived. As a formula, this process was expressed as $Filtered_{item_\theta}(C_k) = \{w_i | w_i \rightarrow w_j \; Rule \; for \; w'_j \in w'_{C_{kp}} and \; d(w'_{C_{kp}}, \{w_i | w_i \rightarrow w'_{C_{kp}}\}) \leq \theta\} \cup \{w'_{C_{kp}}\}$, $\theta = threshold$. At this time, the threshold θ to be applied to the location-based filter was set to a value that most highly satisfied the F1 scale by applying the gradient descent method.

In our methodology, we employed an iterative technique called gradient descent to determine the optimal threshold, represented by θ, for the location-based filter. Gradient descent, a widely used optimization technique in machine learning and optimization problems, aims to minimize a function by iteratively adjusting its parameters in the direction of the steepest decrease (or ascent) in the function's value. The process begins with initialization, where the parameters or coefficients of the function are set to random values, often referred to as weights in machine learning models. Next, the gradient of the objective function is computed, representing the direction of the steepest ascent of the function at the current point. This gradient, a vector of partial derivatives, guides the adjustment of parameters in the opposite direction to minimize the function. The adjustment is determined by subtracting a fraction of the gradient, scaled by the learning rate, from the current parameter values. Iteration involves repeating the gradient calculation and parameter update steps until a stopping criterion is met, such as reaching a specified number of iterations or a threshold improvement in the function value. Throughout the process, gradient descent progressively moves towards the minimum of the function, converging to a local minimum, global minimum, or another stationary point based on the function's characteristics. This iterative process systematically adjusts the theta parameter, starting from an initial value, to enhance our model's performance. Gradient descent entails modifying the threshold value, commencing from an initial point like 0.5, and gradually reducing it in small increments, such as 0.4, 0.3, and so forth, until the F1 score reaches its zenith. Through this systematic fine-tuning, we pinpoint the θ value that strikes the optimal balance between precision and recall, thereby refining our model's effectiveness.

<Figure 5> shows a process of choosing $Filtered\_item_\theta(C_k)$ that satisfies the distance criterion θ using $Inf\_item(C_k)$, $Asso\_item(w'_{C_{kp}})$, and $w'_{C_{kp}}$. First, comparing the distance between items for A and B in {A, B, C, D, E, F} corresponding to $Inf\_item(C_k)$, the distance to itself is excluded from the calculation target (e.g., A and B). When the distance between A and the {B, C, D} items corresponding to $Asso\_item(A)$ is calculated, E is excluded from calculation because it does not belong to $Asso\_item(A)$. As in the previous process, the distance corresponding to B and $Asso\_item(B)$ is calculated and the distance to {C, D, E, F} is obtained. In the example figure below, θ is 3, so the $Filtered\_item_\theta(C_k)$ items that satisfy θ or less are {A, B, C, D}.

Our reasons for applying location-based filters after creating association rules based on claimed repair items rather than generating association rules only with location information are as follows. First, it is difficult to reflect the characteristics of fastening structures for automobile parts in rules created based on location information about work items without correlation analysis to claimed repair items. Because automotive parts are assembled on vehicle bodies in complex structures, parts can be detached due to structural characteristics even though they are not actually damaged. Because claims are made only for parts to which assembly parts are connected (even when they are close to other parts), it is difficult



<Figure 5> Process of $Filtered\_item_{\theta}(C_k)$

to reflect these precise characteristics in association rules when only distance is considered.

Second, when only distance is considered, it is difficult to reflect repair characteristics in the rules. Because repair methods are related to the depth of damage, differences in repair items claimed depend on which repair method is combined with which work item. As a result, even if the same repair item is claimed, the related and claimed items can differ depending on the repair method. It is difficult to reflect these characteristics with distance information alone. Therefore, our model combines repair methods when creating rules based on insurance repair claims data and applies location-based filters with gradient descent to increase the reliability of the rules.

Ⅳ. Experimental Design and Results

4.1. Experimental Design

To conduct this research, we excluded details of user labor, towing, and rescue that are unstructured text data difficult to apply to analysis among claimed actuarial data. We used an FP-growth algorithm suitable for the horizontal structure of auto insurance repair cost data by association rule analysis. Also, because work items for individual vehicles are different, we needed to select a vehicle for association rule analysis. We identified the Hyundai HG Grandeur because the study insurance repair claims data contain the most information about this model. Ultimately, 255,827 records were used from data spanning August to October of 2020 to create association rules. For verification, 75,723 points of data from November of the same year were used.

This study uses 33 work items recognizable by AOS-Alpha in relation to automobile repair to predict claims items. Rules for 33 items were extracted by applying the gradient descent method and location-based filters. Location-based filter no. 1 (LB1) corresponds to a distance of 200 mm, location-based filter no. 2 (LB2) corresponds to 400 mm, and location-based filter no. 3 (LB3) corresponds to 600 mm. The rules for the main items are 481 for no filter, 193 for LB1, 223 for LB2, and 252 for LB3. The current rule set of AOS-Alpha was used to compare performance according to the gradient descent method and location-based filter application. Values are shown in <Table 3>.

4.2. Results

To assess the effectiveness of our research findings,

<Table 3> Number of Rules by Minimum Support and Location-based Filter

Filter	AOS α rule set	Association Rule Set (with Gradient Descent)			
		Without LB	LB1	LB2	LB3
Rule	203	481	193	223	252

we conducted a comparative analysis between the performance of the existing rule set and the rule set obtained through our proposed research methodology against the actual billed estimate. This evaluation involved scrutinizing the predictive model's performance, formulated based on the extracted rules, in contrast to the actual billed estimates, employing a diverse range of metrics such as Accuracy, Precision, Recall, and F1 score. These performance metrics offer a comprehensive evaluation framework, offering valuable insights into the practical feasibility and efficacy of prediction models built upon the framework of these extracted rules.

Accuracy represents the proportion of predicted values that match the actual values in the evaluation dataset. Precision indicates the percentage of correctly predicted normal claim items among all predicted values. Recall measures the consistency of the model predictions with the actual values in the verification dataset. The F1-score is a composite metric that considers both precision and recall. The formula for F1-score is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

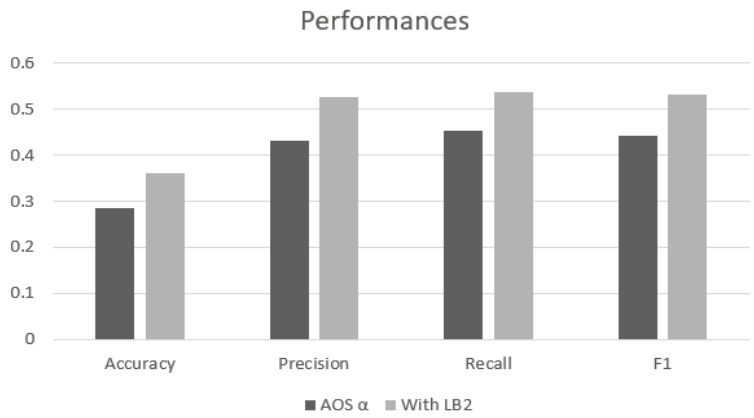
$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

Results of the linked regulations with location-based filter two are presented in <Table 4>. In this instance, there was a 7.8%p increase in accuracy, an 8.6%p increase in recall, a 9.3%p increase in precision, and an 8.9%p increase in F1-score. The findings are depicted in <Figure 6>.

Prior to assessing the effectiveness of individual filters, we conducted a paired-sample assessment of precision to establish statistical significance. Consequently, we determined that the detection rate for insurance fraud with the implementation of location-based filters differs from the rate without filters. <Table 5> displays the p-value of the paired-samples test for each filter dimension.

The present research optimizes filter sizes through the use of the gradient descent method and verifies performance variation with respect to filter size. The outcomes are shown in <Table 6> following the comparison of the performance of each filter. Comparative experiments T1 to T3 were conducted on location-based filters 1 to 3, with evaluation based on four performance criteria: precision, accuracy, recall, and F1-score. The accuracy of the location-based association rule was 7.5%p higher on average than that of the comparative experiment, and the maximum difference was observed in T2 at 7.8%p. The precision difference was 9.2%p on average, and the highest difference was in T1 with 11.2%p. The recall rate averaged at 8.2%p, and the maximum difference was observed in T3 with 9.3%p. The F1-score of the location-based association rule was 8.63%p higher on average, and the maximum difference was observed in T2 with 8.9%p. In sum, the model combining the location-based association rule and gra-



<Figure 6> Performance of LB2 Claim Prediction

<Table 4> Performance of LB2 Detection of Abnormal Items

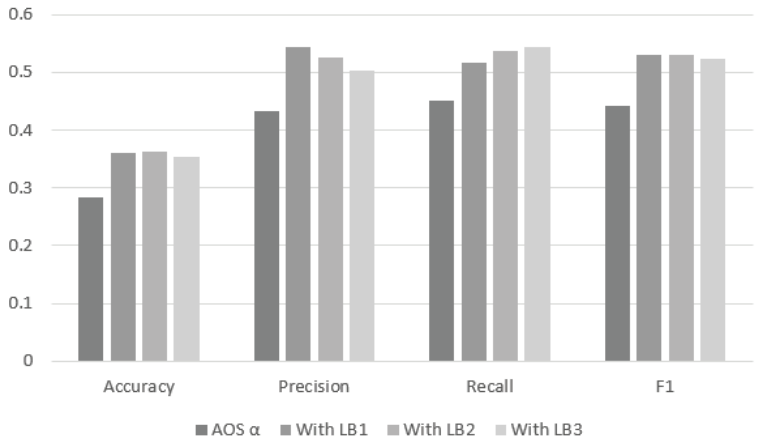
Performance	Accuracy	Precision	Recall	F1
AOS α	0.284	0.432	0.452	0.442
With LB2	0.362	0.525	0.538	0.531
Increased amount	**7.8%p**	**9.3%p**	**8.6%p**	**8.9%p**

<Table 5> P-value of Paired Samples Test with a Location-based Filter

Filter	Sig. (Two-Tailed)		
	LB1	LB2	LB3
P-value of paired samples test	0.000	0.000	0.000

<Table 6> Performance of LB1 to LB3 Detection of Abnormal Items

Performance		Accuracy	Precision	Recall	F1
AOS α		0.284	0.432	0.452	0.4420
T1	**With LB1**	0.361	0.544	0.518	0.5307
	Increased amount	7.7%p	11.2%p	6.6%p	8.9%p
T2	**With LB2**	0.362	0.525	0.538	0.5310
	Increased amount	7.8%p	9.3%p	8.6%p	8.9%p
T3	**With LB3**	0.354	0.504	0.545	0.5233
	Increased amount	7.0%p	7.2%p	9.3%p	8.1%p
Average		7.5%p	9.2%p	8.2%p	8.6%p



<Figure 7> Comparative Evaluation Criterion: With Versus Without LB

dient descent method demonstrated high performance based on our evaluation criteria while the filter producing the largest difference varied for each criterion in relation to the comparative experiment.

As evident in <Figure 7>, T1 to T3 comparative experiments show that the location-based association rule performs exceptionally well, with effective performance comparison based on filter size. The results of four previous performance evaluations indicate that T2 had an accuracy value of 36.2%; T1 had a precision value of 54.4%; T3 had a recall value of 54.5%; and T2 had the highest F1 score. Performance by the T3 location-based association rule is lower than those by T1 and T2 in terms of accuracy, precision, and F1 scores due to the prediction error rate increasing as the size of the location-based filter exceeds the required level for claims prediction.

Ⅴ. Discussion and Conclusion

This research adopts a practical approach to reduce the time and labor costs associated with calculating insurance repair expenses following automotive collisions. By analyzing authentic auto insurance repair data, the study forecasts claim details. Through the utilization of gradient descent and location-based association rules to develop a rule set and anticipate claim items, our study showcased enhanced performance compared to existing rule sets. On average, we observed a noteworthy improvement, with accuracy increasing by 7.5%p, precision by 9.2 %p, recall by 8.2%p, and F1 score by 8.6%p.

Our research findings make numerous contributions to the existing literature in auto insurance repair claims prediction. First, this study employs an association rule technique, which is not frequently used for predicting claims. Research experiments verify that the association rule approach, which reflects the relationship between automobile repair parts, is effective in predicting insurance repair cost claims. Second, the research proposes a prediction model for auto insurance repair claims that combines location-based filters derived from actual low-speed crash test data and association rule techniques using gradient descent. The experimental results demonstrate that the suggested method is superior to the current

model and that filter size affects performance. In addition, association rule techniques can serve as a training tool for insurance company staff aiding the work of both new and seasoned employees. Furthermore, if a repair shop can diagnose and prepare the necessary parts and labor required for initial repairs, all subsequent repairs can be expedited, ultimately lowering ancillary expenses linked to auto insurance, such as rental fees.

The proposed model in this research has areas that may be enhanced in various ways. First, by expanding the amount of analysis data, the prediction performance of claims may be improved. As this research only analyzes one car model, it is important to test the performance of the suggested methodology using different car models. Second, if the range of damage recognition in AOS-Alpha is increased, the accuracy of predicting claims based on association rules should also be increased. In addition, the effectiveness of association rules based on location is influenced by the dimensions of each filter, so we recommend using gradient descent techniques to identify an appropriate range for detecting auto insurance claims on a per-vehicle basis. This will have a significant impact on the proposed model's real-world application. Finally, while our suggested location-based filters were evaluated using two-dimensional coordinates, the filters have the potential to be extended to three-dimensional coordinates in the future.

<References>

[1] Agrawal, R., Imieli´nski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data* (pp. 207-216).

[2] Agrawal, R., and Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference Very Large Data Bases (VLDB)* (pp. 487-499).

[3] Bailey, R. A., and Simon, L. J. (1960). Two studies in automobile insurance ratemaking. *ASTIN Bulletin: The Journal of the IAA, 1*(4), 192-217. https://doi.org/10.1017/S0515036100009569

[4] Bhowmik, R. (2011). Detecting auto insurance fraud by data mining techniques. *Journal of Emerging Trends in Computing and Information Sciences, 2*, 156-162.

[5] Czado, C., Kastenmeier, R., Brechmann, E. C., and Min, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal, 2012*(4), 278-305. https://doi.org/10.1080/03461238.2010.546147

[6] Frees, E. W., and Valdez, E. A. (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association, 103*(484), 1457-1469. https://doi.org/10.1198/016214508000000823

[7] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

[8] Fauzan, M. A., and Murfi, H. (2018). The accuracy of XGBoost for insurance claim prediction. *International Journal of Advances in Soft Computing and its Applications, 10*(2), 159-171.

[9] Fialova, V., and Folvarcna, A. (2020). Default prediction using neural networks for enterprises from the post-soviet country. *Ekonomicko-Manazerske Spektrum, 14*(1), 43-51. https://doi.org/10.26552/ems.2020.1.43-51

[10] Gao, G., and Wüthrich, M. V. (2018). Feature extraction from telematics car driving heatmaps. *European Actuarial Journal, 8*, 383-406.

[11] Gschlößl, S., and Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life

insurance. *Scandinavian Actuarial Journal, 2007*(3), 202-225. https://doi.org/10.1080/03461230701414764

[12] Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications, 39*(3), 3659-3667. https://doi.org/10.1016/j.eswa.2011.09.058

[13] Ghoting, A. Otey, M. E. and Parthasarathy, S. (2004). LOADED: link-based outlier and anomaly detection in evolving data sets. In *Fourth IEEE International Conference on Data Mining (ICDM'04)* (pp. 387-390). IEEE.

[14] Heras, A., Moreno, I., and Vilar-Zanón, J. L. (2018). An application of two-stage quantile regression to insurance ratemaking. *Scandinavian Actuarial Journal, 2018*(9), 753-769. https://doi.org/10.1080/03461238. 2018.1452786

[15] Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM Sigmod Record, 29*, 1-12. https://doi.org/10.1145/335191.335372

[16] Jørgensen, B., and De Souza, M. C. P. (1994). Fitting tweedie's compound poisson model to insurance claims data. *Scandinavian Actuarial Journal, 1994*(1), 69-93. https://doi.org/10.1080/03461238.1994.10413930

[17] Jain, R., Alzubi, J. A., Jain, N., and Joshi, P. (2019). Assessing risk in life insurance using ensemble learning. *Journal of Intelligent & Fuzzy Systems, 37*(3), 2969-2980. https://doi.org/10.3233/JIFS-190078

[18] Kaš´celan, V., Kaš´celan, L., and Buri´c, M. N. (2015). A nonparametric data mining approach for risk prediction in car insurance: A case study from the Montenegrin market. *Economic Research-Ekonomska Istraživanja, 29,* 545-558. https://doi.org/10.1080/1331677X.2016.1175729

[19] Kowshalya, G., and Nandhini, M. (2018). Predicting fraudulent claims in automobile insurance. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 1338-1343). IEEE.

[20] Koufakou, A., and Georgiopoulos, M. (2010). A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes. *Data Mining and Knowledge Discovery, 20*(2), 259-289. https://doi.org/10.1007/s10618-009-0148-z

[21] Liu, Y., Wang, B., and Lv, S. G. (2014). Using multi-class adaboost tree for prediction frequency of auto insurance. *Journal of Applied Finance Banking, 4*(5), 45-53.

[22] Liu, G., Lu, H., Lou, W., Xu, Y., and Yu, J. X. (2004). Efficient mining of frequent patterns using ascending frequency ordered prefix-tree. *Data Mining and Knowledge Discovery, 9*, 249-274. https://doi.org/10.1023/B:DAMI.0000041128.59011.53

[23] Nelder, J. A., and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General), 135,* 370-384. https://doi.org/10.2307/2344614

[24] Otey, M. E., Ghoting, A., and Parthasarathy, S. (2006). Fast distributed outlier detection in mixed-attribute data sets. *Data Mining and Knowledge Discovery, 12*(2-3), 203-228. https://doi.org/10.1007/s10618-005-0014-6

[25] Olden, J. D., and Jackson, D. A. (2002). Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling 154*, 135-150. https://doi.org/10.1016/S0304-3800(02)00064-9

[26] Pesantez-Narvaez, J., Guillen, M., and Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. *Risks, 7*(2), 70. https://doi.org/10.3390/risks7020070

[27] Singh, R., Ayyar, M. P., Pavan, T. S., Gosain, S., and Shah, R. R. (2019). Automating car insurance claims using deep learning techniques. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)* (pp. 199-207). IEEE.

[28] Sun, N., Bai, H., Geng, Y., and Shi, H. (2017). Price evaluation model in second-hand car system based on BP neural network theory. In *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and*

Parallel/ Distributed Computing (SNPD), Kanazawa, Japan, June 26-28 (pp. 431-436).

[29] Smyth, G. K., and Jørgensen, B. (2002). Fitting tweedie's compound poisson model to insurance claims data: Dispersion modelling. *ASTIN Bulletin: The Journal of the IAA, 32*(1), 143-157. https://doi.org/10.2143/AST.32.1.1020

[30] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1

[31] Wüthrich, M. V. (2019). Bias regularization in neural network models for general insurance pricing. *European Actuarial Journal, 10*, 179-202.

[32] Yang, Y., Qian, W., and Zou, H. (2016). Insurance premium prediction via gradient tree-boosted tweedie compound poisson models. *Journal of Business & Economic Statistics 43*, 1-45. https://doi.org/10.48550/arXiv.1508.06378

[33] Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association, 96,* 574-588. https://doi.org/10.1198/016214501753168262

[34] Yunos, Z. M., Ali, A., Shamsyuddin, S M., and Ismail, N. (2016). Predictive modelling for motor insurance claims using artificial neural networks. *International Journal of Advances in Soft Computing and Its Applications, 8,* 160-172. https://doi.org/10.35940/ijrte.F9873.038620

[35] Zhang, L., and Shen, Q. (2019). Improvement of the traditional auto insurance claims frequency model by boosting algorithm—Based on the traffic compulsory insurance data in five provinces of China. *Insure To Study, 7*, 67-78.

◆ About the Authors ◆

Seongsu Jeong

Seongsu Jeong is working at the Repair Research 1 Team of Korea Insurance Development Institute. He is a Ph.D. candidate in the Business Informatics at Hanyang University in Korea. He received B.S degree from Department of Mechanical Engineering at Busan National University, Busan, Korea. He received his M.S. degrees from Graduate School of Business Administration at Hanyang Cyber University, Seoul, Korea. His research interests include business analytics, data mining and machine learning applications.

Jong Woo Kim

Jong Woo Kim is a professor at the School of Business, Hanyang University, Seoul, Korea. He received B.S. degree from the Department of Mathematics at Seoul National University, Seoul, Korea. He received his M.S. and Ph.D. degrees, respectively, from the Department of Management Science, the Department of Industrial Management at Korea Institute of Science and Technology (KAIST), Korea. His current research interests include intelligent information systems, data mining applications, text mining application, sentiment and emotion analysis, social network analysis, collaborative systems, and e-commerce recommendation systems. His papers have been published in *Expert Systems with Applications, Cyberpsychology Behavior and Social Networking, Computers in Human Behavior, Information Systems Frontiers, International Journal of Electronic Commerce, Electronic Commerce Research, Mathematical and Computer Modeling, Journal of Intelligent Information Systems*, and other journals.

Submitted: December 21, 2023; 1st Revision: March 18, 2024; Accepted: April 12, 2024