A Study on the Improvement of Life Insurance Underwriting using the Feature Selection Method and Ensemble Classification Model

Jung-Moon Choi†, Yeong-Jin Kim††, and Je-Dong Lee†††
jmchoi@wise.co.kr, yjkim@wise.co.kr, jdlee@wise.co.kr
† Department of Research and Planning WISEiTECH, Seongnam-si, Republic of Korea
††Department of Research and Planning WISEiTECH, Seongnam-si, Republic of Korea
††† Department of Research and Planning WISEiTECH, Seongnam-si, Republic of Korea

Summary
With the changing workplace landscape evident from the recent remote working arrangement, owing to the disruption caused by COVID-19 pandemic, the pace of adopting integrated services with AI has accelerated. In the insurance industry, there has been a gradual increase in business cases and research, with the introduction of AI technology in areas such as detection of unfair claims, claims adjusting, and insurance acceptance. In this study, an insurance underwriting model for accepting/rejecting new applicants was developed to reduce these discrepancies, based on underwriters, as well as enable faster processing. The data of Prudential Life Insurance from Kaggle was utilized to develop the insurance underwriting model. Among the feature selection methods, the filter-based and embedded methods were comparatively evaluated, and a Regularized Random Forest from the embedded methods was finally selected. For the insurance underwriting model, seven classification algorithms were applied for model optimization, and using the ensemble voting, the result of models with excellent classification performance with a recall score of 0.8 or higher was finally predicted by voting to ensure derivation of reliable results.

Keywords:
Insurance Underwriting, Classification, Feature Selection, unbalanced data, Ensemble.

1. Introduction

With the recent trend in the development of artificial intelligence (AI) technology and digitalization including database construction, the business environment of various industries is undergoing rapid changes, not just the IT industry. The insurance industry is not an exception from the trend of adopting AI technology in various areas. A number of insurance companies have introduced AI technology for resolution, or improvement of problems in a range of areas, and the existing data system is digitalized to construct the database and establish a new system for use in the AI-based analyses.

The relevant fields in the insurance industry where AI technology can be implemented include insurance acceptance, insurance underwriting, policy pricing, insurance claims, and detection of insurance fraud. AI technology can be widely integrated throughout the overall processes of the insurance industry, and can be effectively adopted in various fields such as improvement of prediction accuracy, understanding of claims and customer behavior, product development, and derivation of analytic insights [1].

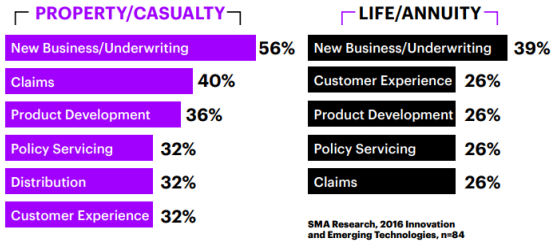

Fig. 1 Areas of insurance for utilization of machine learning technology.

The source of Figure 1 is from Accenture's 2018 report entitled “Machine Learning in Insurance”, which describes the insurance business areas where machine learning (ML) can be leveraged in the insurance industry. Among the 5 areas described above in the figure, the area with the highest potential for AI integration is the new business/underwriting. This indicates that machine learning is suitable for predicting the risk level .of insurance applicants, and for deriving useful insights.

Furthermore, it is expected that with the implementation of novel technologies, the insurance business landscape will evolve significantly from the present status, and a shift is expected from the conventional “detect and repair” approach based on statistical rules, to “predict and prevent” approach with AI application.

In the conventional process of underwriting, an underwriter reviews and evaluates the information of all insurance applicants. In this process, the underwriter gathers extensive information of the applicants, and evaluates if the application needs to be accepted, and it takes an average of 30 days to calculate premiums [2]. In addition to the problem of lengthy process, there is a challenge of human errors caused by the underwriter's mistakes or

Manuscript received December 5, 2021
Manuscript revised December 20, 2021
https://doi.org/10.22937/IJCSNS.2022.22.1.51

differences in results, depending on the work experience, knowledge, and skill levels of individual underwriters.

The adoption of AI technology can address these challenges by reducing the differences between the underwriters, and improving the business environment to enable fast and accurate processing. Underwriters can focus on important tasks such as screening insurance applicants who are considered to be high-risk groups, thereby increasing the efficiency of the overall underwriting process.

Considering the above beneficial effects, a number of insurance companies have constructed databases and made efforts for active integration of AI technology, and the number of InsurTech cases in the domestic and international insurance industry has gradually increased. Among the InsurTech cases, there has been a gradual increase in application of AI to the underwriting process [3]. Prudential Life Insurance automated 40-50% of new insurance underwriting process, Samsung Life advanced its automatic insurance underwriting system close to AI, and Kyobo Life improved its business efficiency by utilizing substantial data for underwriting and claims adjusting. In addition, AIA Life introduced an underwriting system that can be checked in real time, and Swiss Re Insurance improved the level of standardization in the underwriting process, by introducing AI technology to support the work process of auto insurance and life insurance underwriters.

In this study, an insurance underwriting model utilizing machine learning-based insurance underwriting data, was developed in line with the trend of AI technology implementation in the insurance industry. Chapter II introduces existing works related to the insurance underwriting methods. Chapter III describes an operational definition of the data utilized in this study, and the data preprocessing process. Chapter IV discusses the development process of the insurance underwriting model developed with the machine learning-based classification algorithms, and Chapter V presents the conclusions and implications of this study.

2. Theoretical Background and Existing Works

The quality of underwriting service is one of the key factors in determining the corporate reputation in the life insurance industry, and the assurance of quality helps to maintain an advantageous position in the competitive market. Therefore, it is instrumental to improve the underwriting process to gain a competitive edge in new sales and retention of customers. There has been an active research on the improvement of underwriting processes using machine learning methods.

In [2], regarding data preprocessing, Little's Test of Missing Completely at Random was utilized , MAR(Missing At Random) was utilized as missing data structure for imputation of missing values, and data dimension was reduced using CFS(Correlation-Based Feature Selection) and PCA(Principal Components Analysis); thus, 4 different machine learning models were developed. The model performance was evaluated using the MAE (Mean Absolute Error) and RMSE (Root Mean Square Error) values. In the case of REPTree Algorithm, the lowest MAE 1.5285, and RMSE 2.027 were obtained by the CFS method.

In [4], underwriting scoring method was investigated by applying a total of 5 data mining supervised learning techniques, to determine the acceptance of the insurance applicant and claims payment approval, but the accuracy in prediction was low at 60-70%.

In [5], an accident rate prediction model using individual credit information was developed for life insurance underwriting, and the effect of business improvement was expected according to the decrease in the overall accident rate, by applying tighter conditions for acceptance of applicants in the high-risk group.

In [6], to resolve the problem of imbalance, depending on the level of risk, Synthetic Minority Over-Sampling Technique (SMOTE) was applied among combined under- and over-sampling methods, and through hyperparameter tuning in Random Forest model, which indicated the best performance among 4 ML algorithms, the accuracy of 74% was achieved.

In [7], as a prediction of the supervised learning algorithm, stacking ensemble learning was utilized, and 10-Fold cross validation was performed on the data set for each model. After applying individual models (Logistic Regression, K Nearest Neighbor, Random Forest, AdaBoost Classifier) to final ensemble models (Decision Tree, Gradient Boosting) for learning, ANN and XGBoost were added, and the prediction performance was evaluated by comparing MAE, RMSE, accuracy, and the Kappa value.

In this study, among different feature selection techniques, the filter and embedded methods were compared, and the method with the best performance was selected. Then, hyperparameter tuning was performed to optimize the hyperparameters of the machine learning classification algorithm. Based on the results obtained through this process, a model with a recall score of 0.8 or higher was selected, and the ensemble method was utilized to achieve reliable prediction results of the model.

3. Data Preprocessing

3.1 Applying Data

In this study, using the open data of Prudential Life Insurance from Kaggle, through the normalized variables that indicate the propensity of insurance applicants for Prudential Life Insurance and reprocessing, a dataset that

includes response variables that can be utilized to determine the applicants risk level was constructed.

Train data with response values were utilized as raw data for optimization, based on model performance evaluation. Raw data is composed of 59,381 cases of 128 variables, and Table 1 outlines the variable names and brief descriptions of the variables. Variables ending in _ are a set of multiple variables, and because the variables in a set have similar characteristics, they were grouped into one set.

Table 1. Variable List

No	Variable Name	Variable Description
1	Ins_Age	Normalized age of applicant
2	Ht	Normalized height of applicant
3	Wt	Normalized weight of applicant
4	BMI	Normalized BMI of applicant
5	Employment_Info_	A set of normalized variables relating to the employment history of the applicant(6)
6	Insured_Info_	A set of normalized variables providing information about the applicant(6)
7	Insurance_History_	A set of normalized variables relating to the insurance history of the applicant(9)
8	Family_Hist_	A set of normalized variables relating to the family history of the applicant(5)
9	Medical_History_	A set of normalized variables relating to the medical history of the applicant(41)
10	Medical_Keword_	A set of dummy variables relating to the presence of / absence of a medical keyword being associated with the application(48)
11	Response	This is the target variable. an ordinal variable relating to the final decision associated with an application(1—8)

3.2 Data Pre-Processing

In this study, Figure 2 below is a graph comparing the data null ratio for each variable, and Figure 3 is a graph comparing the number of data cases by response. Variables not required for data analysis were deleted, and the missing values in each variable were checked to perform pre-processing. Based on this, from variables with high null ratio, whether to utilize the variable was determined, and missing data was imputed.

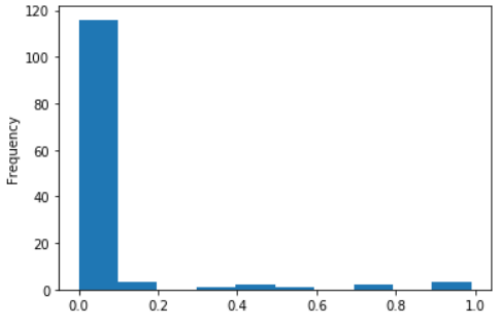

Figure 2. Data Null Ratio for each variable

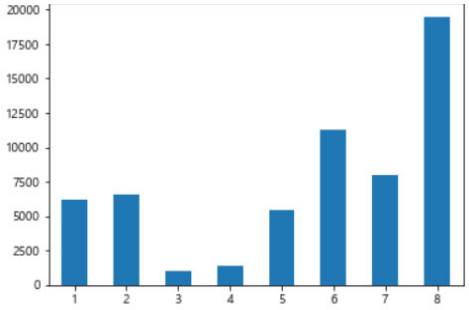

Figure 3. Number of data cases per response

Overall, although the data null ratio was low, the variables were eliminated for those whose null ratio was 50% or higher, and for variables determined to be utilized in the analysis, the missing values were replaced with -1. Through the preprocessing step, 4 out of 128 variables were eliminated, and a total of 124 variables, and additional derived variables were created to construct the data set.

The responses of the target data for analysis indicate a distribution as illustrated in Figure 3; hence, there is data imbalance depending on class. Class 3, which has the smallest number of data, has 1,013 cases, accounting for 1.7% of the total data, and Class 8, which has the most data, has 19,489 cases, accounting for 32.8% of the total data. Overall, more data is distributed in lower classes, and the number of data in Class 3 and 4 is negligible in comparison.

To address the class imbalance problem, 8 classes of the responses were grouped in two each, to produce 4 classes(0, 1, 2, 3) and a sampling method was applied. Various sampling methods were applied to reduce data imbalance, and results of the confusion matrix of the basic model were compared to select the final sampling data. The sampling methods utilized to determine the final data include Random Over Sampling, Tomek Links Sampling, SMOTE Sampling, Random Under Sampling, ENN (Edited Nearest Neighbors) Sampling, SMOTE + ENN Sampling, SMOTE + Tomek Sampling, and finally, SMOTE sampling data was utilized for analysis.

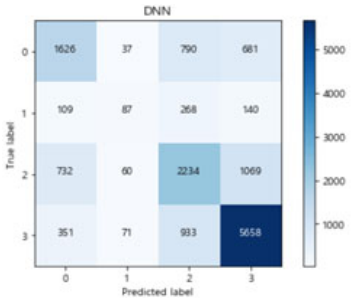

Figure 4. Result of classification performance test for raw data

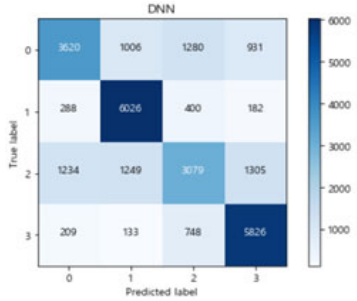

Figure 5. Result of classification performance test for SMOTE sampling data

Figures 4 and 5 are the confusion matrix applied to the basic classification model of raw data and SMOTE sampling data, and it is evident that the learning performance for the minority data labels that were not properly trained in raw data, can be improved through SMOTE sampling. The number of data cases applied with SMOTE sampling method is 110,024 in total. Data corresponding to levels 1 to 6 based on the response are automatically classified as Accept (0), and data corresponding to levels 7 to 8 are classified as Reject (1), so that the target labels for insurance underwriting model development were altered to binary format.

4. Development of insurance underwriting model

The insurance underwriting model was developed, with binary target labels of automatic Accept (0) and Reject (1) as dependent variables. Among the feature selection techniques, the filter embedded methods were compared, and the method that can improve the model performance was selected. Based on the preferred feature selection method, the model performance was improved through hyperparameter tuning of various machine learning algorithms, and finally, models with good performance results were selected. Then, an ensemble model was utilized to ensure reliable results in the insurance underwriting.

4.1 Comparison of feature selection method

Feature selection is a key process that must be performed to improve the performance of a model in machine learning. It can be largely divided into the filter method that selects variables based on statistical characteristics such as mutual information or correlation coefficient, the wrapper method that repeats the process of solely utilizing a few of the variables in modeling to examine the result, and the embedded method, in which feature selection is included in the model itself, as in the case of Random Forest. In this study, the filter embedded methods were compared from the feature selection methods.

Table 2. Description of Feature Selection

Feature Selection	Description
RF (Random Forest)	The method measures how much a variable contributes to mean decrease accuracy and improvement of node impurities (mean decrease Gini).
RRF (Regularized Random Forest)	The method sets similar weights for both new and existing features in each regularized tree for more robust model performance.
GRRF (Guided Regularized Random Forest)	The method sets different weights, and additionally considers the importance value so that a variable that was not present in the previous tree, but utilized frequently can be selected.
Pearson Correlation Coefficient	The method measures the linear correlation between two variables, and determines the relative importance of each predictor, when the correlation coefficient between the target value and the predictor is considered.
Spearman's Rank Correlation Coefficient	The method calculates the correlation coefficient by using the ranks of two data, instead of their actual values, and recommends the variable with a large correlation coefficient through identification of nonlinear relationship.
Chi -Squared Test	The test is performed between the classification and the variable to be predicted, and determines the adequacy of the variable for the modeling through the test of independence between the variable and the classification.

Table 2 above outlines the feature selection techniques utilized in the study. For the feature selection methods, RRF (Random Forest, Regularized Random Forest, Guided Regularized Random Forest) and CARET (Pearson, Spearman, Chi) package of R-Studio program were utilized , and for the Pearson, Spearman, and Chi-squared tests, which are classified as the filter method, variables with near zero variance, and those with high correlation coefficients ($\geq$0.9) were eliminated prior to utilizing the methods.

Table 3. Top 10 variables of importance for each feature selection method

Rank	RF	RRF	GRRF	Pearson	Spearman	Chi-Square
1	**Medical_History_4**	**Medical_History_4**	**BMI**	**Medical_History_4**	**Medical_History_4**	**Medical_History_4**
2	**BMI**	**BMI**	**Medical_History_4**	**BMI**	**BMI**	**Wt**
3	**Wt**	**Producttt_Info_4**	**BMI_Age**	**Wt**	**Wt**	InsuredInfo_6
4	**Producttt_Info_4**	**BMI_Age**	**Producttt_Info_4**	InsuredInfo_6	InsuredInfo_6	Ins_Age
5	**BMI_Age**	Medical_History_23	**Wt**	Medical_History_23	Medical_History_23	Ht
6	Ins_Age	Employment_Info_1	Employment_Info_6	**BMI_Age**	**BMI_Age**	Family_Hist_4
7	InsuredInfo_6	Family_Hist_4	Medical_History_2	InsuredInfo_1_2	InsuredInfo_1_2	**Producttt_Info_4**
8	Family_Hist_4	Medical_History_2	Medical_History_1	**Producttt_Info_4**	**Producttt_Info_4**	Employment_Info_6
9	Employment_Info_1	Medical_History_1	Producttt_Info_2	Employment_Info_5	Medical_History_33	InsuredInfo_1_2
10	Family_Hist_3	**Wt**	InsuredInfo_3	Medical_History_33	Employment_Info_5	Employment_Info_1

In Table 3, variables with high values of MDG (Mean Decrease Gini) and Attr_importance, which represents the importance of each variable, respectively, are listed in the order of rank, using six feature selection methods. Although there are differences in important variables depending on the method, variables such as **Medical_History_4, BMI, Wt, Product_History_4,** and **BMI_Age** were generally classified as the upper ranks.

4.2 Comparison of accuracy and recall according to the feature selection method

Table 4. Description of Algorithms

Feature Selection	Description
RF (Random Forest)	Algorithm that several Decision Trees compose the Forest, and each predictive result is averaged as a single result variable
MLP (Multi-layer perceptron)	Neural Network Algorithm with one or more hidden layers between imp. Networks composed of multiple layers of perceptrons
GBC (Gradient Boosting Classifier)	Boosting algorithm which creates Trees sequentially in using the Gradient descent, and compensates errors of previous Tree
LGBM (Light GBM)	Unlike the tree-based algorithms with horizontal growth, this method is based on the Gradient Boosting framework, and is a tree-based algorithm with vertical growth.
AdaBoost (Adaptive Boosting)	A representative boosting algorithm in which higher weights are assigned to misclassified instances of weak learners
ExtraBoost (Extra Boosting)	An algorithm that adds randomization, by randomly splitting each candidate feature of the forest tree
XGBoost (eXtreme Gradient Boosting)	An algorithm with enhanced speed and performance, compared to Gradient Boosting by improving the drawbacks of Gradient Boosting so that regularization terms are added. and selection of various loss functions is possible.
Soft Vote	The algorithm is also called Probability Voting, and in this algorithm, the probabilities predicted by models are added for each class, and the class with the highest probability is selected.
Hard Vote	The algorithm is also called Majority Voting, and in this algorithm, when respective models predict the results, the result with the most votes is selected.

Table 4 Outlines the algorithms utilized for the insurance underwriting model. After determining the feature selection technique that presents the best acccuracy and recall values, which are performance evaluation indices that can be calculated from the confusion matrix, hyperparameter tuning was performced for each alorithm, targeting an accuracy, and recall score above 80% each.

Table 5. Comparison of accuracy before tuning for each feature selection method

Algorithm	RF	**RRF**	GRRF	Pearson	Spearman	Chi
RF	0.871	**0.876**	0.864	0.876	0.876	0.876
MLP	0.866	**0.873**	0.862	0.862	0.869	0.869
GBC	0.871	**0.883**	0.865	0.879	0.879	0.879
LGBM	0.872	**0.883**	0.867	0.879	0.879	0.879
AdaBoost	0.867	**0.879**	0.859	0.875	0.875	0.875
ExtraBoost	0.873	**0.882**	0.866	0.876	0.877	0.877
XGBoost	0.873	**0.888**	0.870	0.883	0.883	0.883
Average	0.870	**0.881**	0.865	0.876	0.877	0.877

Table 5 above outlines the accuracy when the attribute of the algorithm is learned by default to compare the performance of feach feature selection method. The feature selection method that indicated the highest average value of accuracy was RRF, with the average at 0.881, and the algorithm that indicated a the highest accuracy was also XGBoost of RRF.

In Table 6, the average recall for 0 (Accept) was 0.929, and the average recall for 1 (reject) was 0.739 for the RRF feature selection method, which was the highest among the average values of other feature selection methods. Based on these results, the hyperparameter tuning for each classification model was performed with a focus on those with a recall of 1 (reject), which was below the target value of 0.8.

Table 6. Comparison of recall score before tuning for each feature selection method

Algorithm	Class	RF	**RRF**	GRRF	Pearson	Spearman	Chi
RF	0	0.93	0.94	0.92	0.94	0.94	0.94
	1	**0.70**	0.68	0.69	0.69	0.68	0.68
MLP	0	0.93	0.92	0.91	0.91	0.91	0.91
	1	0.66	0.73	0.73	0.72	**0.74**	**0.74**
GBC	0	0.92	0.93	0.92	0.92	0.92	0.92
	1	0.72	**0.75**	0.71	0.74	0.74	0.74
LGBM	0	0.92	0.93	0.92	0.92	0.92	0.92
	1	0.73	**0.76**	0.72	0.75	0.75	0.75
AdaBoost	0	0.91	0.92	0.91	0.92	0.92	0.92
	1	0.72	**0.75**	0.72	0.74	0.74	0.74
ExtraBoost	0	0.93	0.94	0.92	0.93	0.93	0.93
	1	**0.71**	**0.71**	0.69	0.70	0.70	0.70
XGBoost	0	0.91	0.92	0.91	0.92	0.92	0.92
	1	0.75	**0.79**	0.74	0.78	0.78	0.78
Average	0	0.921	**0.929**	0.916	0.923	0.923	0.923
	1	0.713	**0.739**	0.714	0.731	0.733	0.733

4.3 Comparative analysis between before and after tuning for each model

Table 7. Comparison of accuracy/recall before and after tuning for each model

Algorithm	**Accuracy**		**Recall**	
	Before	**Tuning**	**Before**	**Tuning**
RF	0.879	**0.886**	0.69	0.73
MLP	**0.865**	0.849	0.75	**0.81**
GBC	0.883	**0.890**	0.75	**0.80**
LGBM	0.882	**0.891**	0.75	**0.81**
AdaBoost	0.879	**0.885**	0.75	**0.80**
ExtraBoost	0.884	0.884	0.72	0.72
XGBoost	0.882	**0.891**	0.75	**0.80**
Average	0.879	**0.882**	0.737	**0.781**

With feature selected by RRF, hyperparameter tuning was performed on RF, MLP, GBC, LGBM, AdaBoost, ExtraBoost, and XGBoost. Table 7 above outlines the results of accuracy and recall, before and after hyperparameter tuning. In terms of accuracy, the values of 5 algorithms, apart from MLP whose accuracy decreased by 0.016, and ExtraBoost whose accuracy was similar at 0.884, were increased by approximately 0.007. In terms of the recall, the overall values of 5 algorithms (MLP, GBC, LGBM, AdaBoost, XGBoost), excluding RF and ExtraBoost, were increased by approximately 0.05

4.4 Model stabilization through soft and hard vote

The five algorithms (MLP, GBC, LGBM, AdaBoost, XGBoost), which were confirmed to have achieved the target values through the comparison of accuracy and recall values before and after the hyperparameter tuning were selected, and soft vote and hard vote version were developed as ensemble model to produce robust prediction results for the insurance underwriting. Table 8 below is a classification report of the two versions of the ensemble model.

Table 8. Comparison of ensemble model results

	Class	Precision	**Recall**	F1 score	Accuracy
Soft Vote	0(Accept)	0.93	**0.92**	0.93	0.89
	1(Reject)	0.77	**0.81**	0.79	
Hard Vote	0(Accept)	0.93	**0.93**	0.93	0.89
	1(Reject)	0.78	**0.79**	0.78	

The accuracy is 0.89 for both the soft and hard votes , and the recall values for 1 (reject) are 0.81 and 0.79, respectively. The recall for 1 (reject) of soft vote was higher by 0.02 than that of hard vote, and both the accuracy and recall exceeded the target value of 0.8. In addition, the values of accuracy and recall were the highest, compared to those of the algorithms (RF, MLP, GBC, LGBM, AdaBoost, ExtraBoost, XGBoost) before the voting, soft vote ensemble model was finally selected.

5. Conclusion

In this study, an AI-based insurance underwriting model was developed to reduce the differences between underwriters, and enable fast processing. The insurance underwriting model was designed as a binary classification model that classifies between accept and reject, and among the feature selection methods, the filter methods and embedded methods were compared. Training was performed with the 7 classification algorithms for the extracted variables, the accuracy and recall scores were compared, and RRF with the highest values of the accuracy and recall was selected.

Hyperparameter tuning was performed for each algorithm based on the RRF variables, and 5 algorithms (MLP, GBC,

LGBM, AdaBoost, XGBoost) were selected that exceeded the target recall score of 0.8 for the label 1(reject). The soft vote and hard vote versions of ensemble model were applied to the algorithms, and the soft vote with the highest accuracy and recall score at 0.89 and 0.81, respectively, was selected as the final classification model.

In this study, there are limitations in terms of additional collection and application of insurance data, because only the normalized data of Prudential Life Insurance were utilized for model development. Nevertheless, this study applied various feature selection methods, and an ensemble model to develop a model that can be generally applied.

In the future, follow-up research is planned, in which based on the variables with universal importance, exploratory analysis is performed on the alterations in application acceptance status or level of risk, according to the values of the variables through SHAP visualization. In addition, by obtaining various additional data of insurance applicants, features will be extracted by age group, region and variables, and further analysis will be performed on the differences according to the identified trend for each type.

Acknowledgments

This work was supported by the ATC program of MOTIE/KEIT(10077293, Intelligent unfair claim detection system technology development, that improves over 40% of unfair claim detection rate by early prediction of insurance unfair claim).

References

[1] R. Balasubramanian, A. Libarikian, and D. McElhaney, “Insurance 2030—The impact of AI on the future of insurance,” McKinsey Article, (2021).

[2] N. Boodhun and M. Jayabalan, “Risk prediction in life insurance industry using supervised learning algorithms,” Complex & Intelligent Systems, vol. 4, no.2, (2018), pp.145–154. DOI: 10.1007/s40747-018-0072-1

[3] J. Kim and S. Lee, "2030 Vision and Tasks of the Korean Financial Industry: Insurance Industry - Focusing on Digitalization of Finance after the Corona Crisis-,” KIF Research Series, vol. 2021, no.1, (2021), pp.1–285.

[4] Y. Lee, J. Hur and Y. Choi, "Prediction Scoring Model for Underwriting Insurance Claims of Imbalanced Data : A Case of Life Insurance Company,” Journal of the Korean Data Analysis Society, vol.12, no.6 (B), (2010), pp. 3231–3245.

[5] J. Chung and Y. H. Yeo, "The Effects of Using Individual Credit Information on Life Insurance Underwriting,” Journal of Money & Finance, vol. 25, no. 1, (2011), pp. 25–56.

[6] S. V. Kumar and A. Gujju, “Risk segmentation of insurance data using machine learning,” Journal of Critical Reviews, vol. 7, no. 19, (2020).

[7] R. Jain, J. A. Alzubi, N. Jain, and P. Joshi, “Assessing risk in life insurance using ensemble learning,” Journal of Intelligent & Fuzzy Systems, vol.37, no.2, (2019), pp. 2969–2980. DOI: 10.3233/JIFS-190078

[8] T. Kohonen, “The Self-Organizing Map,” Proceedings of the IEEE, vol. 78, no. 9, (1990), pp.1464–1480. DOI: 10.1109/5.58325

[9] J. Wendel and B. P. Buttenfield, Formalizing Guidelines for Building Meaningful Self-Organizing Maps, GIScience Short Paper Proceedings, (2010).

[10] J. Tian, M. H. Azarian, and M. Pecht, “Anomaly detection using self-organizing maps-based k-nearest neighbor algorithm,” PHM Society European Conference, vol.2, no.1, (2014). DOI: 10.36001/phme.2014.v2i1.1554

[11] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, (2014).

[12] H. Deng and G. Runger, “Feature selection via regularized trees,” The 2012 International Joint Conference on Neural Networks (IJCNN). IEEE, (2012), pp.1–8. DOI: 10.1109/IJCNN.2012.6252640

[13] H. Deng, “Guided Random Forest in the RRF Package,” arXiv preprint arXiv:1306.0237, (2013).

[14] T. G. Dietterich, Ensemble methods in machine learning, In International workshop on multiple classifier systems, Springer, Berlin, Heidelberg, (2000), pp.1–15. DOI: 10.1007/3-540-45014-9_1

Jung-Moon Choi received the B.S. and B.L.A. degrees from Seoul Women's University in 2014. She is a senior researcher in the Department of Research and Planning at WISEiTECH. Her research has pioneered and focused on every aspect of data analytics, optimization, and predictive modeling. As a senior researcher, she has led a project to develop predictive analytics for insurance fraud detection using artificial intelligence. In addition, her research interests include machine learning. She explores neural network applications and tries to apply deep learning models based on prior knowledge obtained from different projects.

Yeong-Jin Kim received his B.A. and B.S. degrees from Hannam University in 2021. He is an assistant researcher in the Department of Research and Planning at WISEiTECH. His research interests include software engineering, ML/DL.

Je-Dong Lee received his Bachelor's degree from Seoul University in 1993, and Master's degree from Soongsil University in 2020. He is working as a Vice President at WISEiTech since 1998. His research interest include AI, Bigdata, and Data Quality.