Policy Report 2018-04

Machine Learning-Based Models for Big Data Analysis and Prediction

- Social Security Applications

Miae Oh, Hyeonsu Choi, Jaehyeon Jin, and Migyeong Cheon

KIHASA
Korea Institute for Health and Social Affairs

【Principal Researcher】
Miae Oh Research Fellow, Korea institute for Health and Social Affairs

【Publications】

A study on the Micro data linkages to promote the production and utilization of health and welfare statistics, Korea institute for Health and Social Affairs (KIHASA), 2014 (author)

Data Portal System Management and Operation 2016 for Health and Welfare Statistical Information, Korea institute for Health and Social Affairs (KIHASA), 2016 (author)

【Co-Researchers】
Hyeonsu Choi Research Fellow, Korea institute for Health and Social Affairs
Jaehyeon Jin Senior Research, Korea institute for Health and Social Affairs
Migyeong Cheon Research, Korea institute for Health and Social Affairs

Machine Learning-Based Models for Big Data Analysis and Prediction: Social Security Applic

Korea Institute for Health and Social Affairs
Building D, 370 Sicheong-daero, Sejong city 30147 KOREA
http://www.kihasa.re.kr
ISBN: 978-89-6827-510-4 93510

Contents

List of Tables

List of Figures

I Introduction

I Introduction

Amid the current era of the Fourth Industrial Revolution, it is critical to incorporate a wide range and immense volumes of data into policymaking so as to analyze and prepare for the future with policies that reflect new knowledge and values.

Artificial intelligence (AI) and big data analysis are the core technologies underlying the Fourth Industrial Revolution, and the self-sustained evolution of algorithms, based upon machine learning and big data, is key to all related progress. Machine learning, which is a part of AI, refers to the technology with which computers learn and adapt on the basis of large quantities of accumulated data. Machine learning holds the key to analytical and predictive tasks required in a variety of fields, including image processing, video and voice recognition, and Internet search.

Now that big data analysis is increasingly becoming an integral part of policymaking, there is growing hope that it will allow us to better pinpoint the blind spots of the social security system and identify children at risk. Regarding the former, a local government in South Korea has already used a machine learning-based (MLB) model of big data analysis to identify 114,000 of its citizens considered to be at high risk of deprivation and alienation in the winter.[1] This led the local govern-

ment to provide additional services for 17,000 new welfare recipients. The MLB model has proven itself to be four times more productive than the conventional method, according to which the local government surveyed all households that had experienced power disconnections in order to find citizens who could be in need of welfare services. As for identifying children at risk, the Korean government has been developing an MLB model for that purpose since July 2017. The model will make use of big data on extended absences from schools and lack of medical service records on children to identify and rescue children who could be suffering from abuse.

In this study, we examine the characteristics of big data on social security services and analyze the MLB statistical techniques, with a view to designing an MLB model capable of analyzing social security big data that can be used for evidence-based research. MLB models for prediction can make significant contributions to research and policymaking by enhancing the utility of available data and enabling diverse types of analyses. Machine learning can be applied to social policymaking with great benefit, allowing us to provide increasingly predictive, proactive, and customized social security services.

1) As of December 2015.

II

Concepts

1. Social Security Big Data
2. Machine Learning
3. Machine Learning Algorithms

Ⅱ Concepts

1. Social Security Big Data

The conceptualization of social security big data (SSBD) inevitably reflects the areas of policy services encompassed by a given definition of social security in a given political community. Big data can be categorized into structured and unstructured data. Alternatively, the concept can include behavior, image, social, and language data. Machine learning has been applied productively to both structured and unstructured types of data. As we are limiting our focus to administrative (structured) data in this study, we shall define SSBD as "administrative big data produced and accumulated in relation to social security services."

This administrative data is accumulated, managed, and applied in the form of "social security information," as defined in Article 23 of the Act on the Use and Provision of Social Security Benefits and Search for Eligible Beneficiaries ("Social Security Benefits Act," or "SSBA") and the "social security information system," as defined in Article 37 of the Framework Act on Social Security (FASS). Paragraph 1, Article 23, of the SSBA states: "The Minister of Health and Welfare may manage the following data or information[2] (hereinafter referred to as

2) "1. Data or information concerning the current status of social security

"social security information") using the social security information system so that a livelihood security agency can efficiently select persons entitled to social security benefits, manage such social security benefits, and address other relevant affairs."

The concept and scope of SSBD in our analysis shall therefore be understood as encompassing the social security information (as defined in Paragraph 1, Article 23, of the SSBA) accumulated via the social security information system (Article 37, FASS) and other information systems that support the administration of social insurances in relation to the development and implementation of policy measures for diverse areas of social security (Article 3, FASS).

---

benefits, including the statutes that provide the legal basis, the targets and details of security services, budget, etc.;

2. Data or information concerning personal information, income, property, etc. necessary for consultation, application, investigation, and eligibility management pursuant to Articles 5 through 22;
3. Data or information concerning the history of social security benefits received;
4. Data or information necessary for the Minister of Health and Welfare to conduct the duties delegated or entrusted under Article 51;
5. Data or information concerning the records of business affairs performed in accordance with the statutes related to social security information, including counseling, application (including the application filed under Article 25 (3)), investigation, determination, provision, recovery, etc.;
6. Data or information concerning the current status of the provision of social security benefits by private corporations, organizations, or facilities related to social security and the history of such private corporations, organizations, or facilities receiving subsidies;
7. Other data or information necessary for the provision and management of social security benefits and establishment and operation of the social security information system, which shall be prescribed by Presidential Decree."

2. Machine Learning

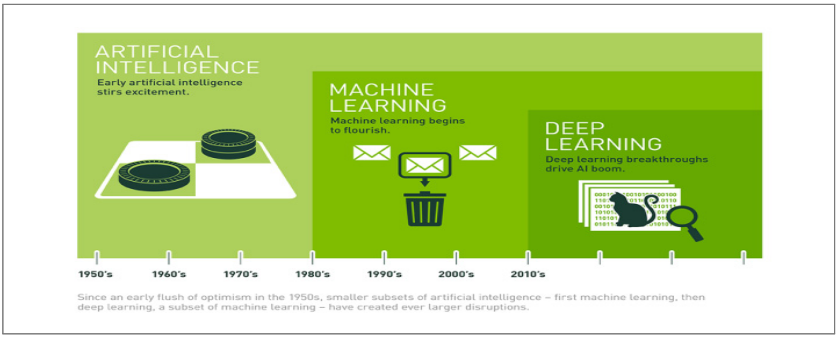
To understand machine learning, we first need to understand learning in general. From a practical perspective, learning can be defined as a combination of representation, evaluation, and optimization (Kim, 2016, p. 77). Representation refers to the process or model by which an agent, tasked with performing a certain activity, decides how to process the input so as to obtain the desired output. To train a computer to recognize cursive script, for example, the computer needs to have a model of logic according to which it can recognize and classify Arabic numerals written in cursive. Evaluation involves the use of techniques or methods by which an agent is assessed on how well he or she has performed a given task. In the field of machine learning, this can be done by measuring the probability with which the trained computer accurately recognizes Arabic numerals written in cursive. Optimization involves finding conditions that optimally satisfy the evaluation standard. After the optimization process, we can decide the weights to be used in our learning model. The learning process is complete after optimization. A learning agent, having completed such learning, can make predictions regarding new data through the process of generalization.

Machine learning is sometimes confused with data mining (Domingos, 2012, pp. 78-87). However, whereas machine

learning centrally involves a series of algorithms or processes by which a computer, without a program pre-installed, identifies patterns in given data and applies those patterns to new data, data mining is specifically performed to provide certain types of information or insights for human users. Machine learning is closely correlated to computer science, statistics, and data mining, leading some to refer to it as "statistical learning."

Artificial intelligence (AI) is a broader concept of which machine learning is only one part. Deep learning, a central topic of discourse today, can be treated as both a subset of machine learning and an independent field of research on its own.

[Figure 2-1] Relationships between AI, Machine Learning, and Deep Learning



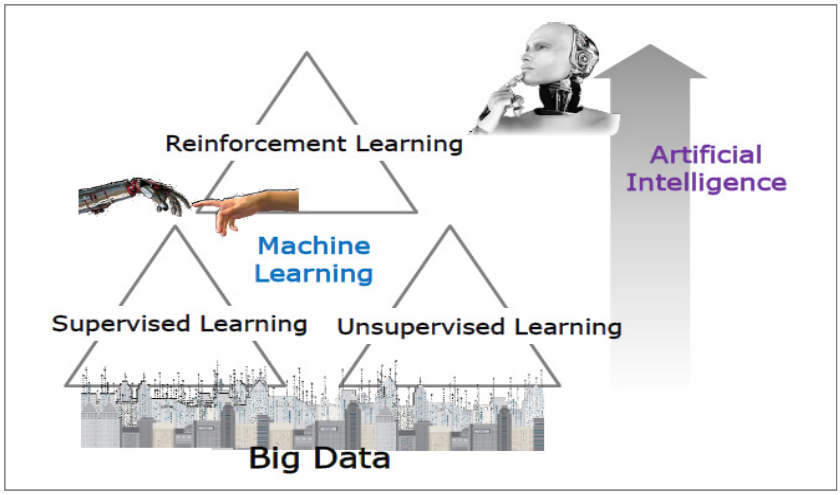
Source: Castrounis, A. (2016). Artificial Intelligence, Deep Learning, and Neural Networks, Explained. https://www.kdnuggets.com/2016/10/artificial-intelligence-deep-learning-neural-networks-explained.html (retrieved November 29, 2017).

3. Machine Learning Algorithms

Machine learning algorithms can be categorized as algorithms involving labeled training data (e.g., dependent or response variables) and those involving unlabeled training data. Alternatively, they can also be categorized as algorithms of supervised learning and algorithms of unsupervised learning. Labels define the attributes of given training data (Kim, 2016, p. 78). Supervised learning involves classification and prediction models, while unsupervised learning features clustering models. Reinforcement learning is treated as either a form of supervised learning or an independent form of machine learning separate from supervised and unsupervised forms of learning. Reinforcement learning enables the computer's task mode to evolve on its own based upon the outcomes of tasks performed according to given algorithms. In his introduction of AlphaGo 2.0, Demis Hassabis, of Google DeepMind, explained that no human manuals had been entered into the system and that the program, with only the basic rules of go having been entered into it, had learned how to win games all on its own just by playing (The Joongang Ilbo, 2017). AlphaGo Zero, as described in Nature in October 2017, features reinforcement learning algorithms, and therefore learns on its own through trial and error. This feature of reinforcement learning sets it apart from supervised and unsupervised learning. Learning

through trial and error certainly bears some resemblance to how humans learn, leading some to argue that reinforcement learning lies at the core of machine learning.

[Figure 2-2] Types of Machine Learning Algorithms



Source: Choi, D. (2017). Big Data and AI in the Age of the Fourth Industrial Revolution. special lecture at ICT Convergence Korea 2017. Seoul.

While machine learning can make powerful predictions based upon new data processed according to learned algorithms, the predictions so generated may be difficult for humans to interpret. Efforts are thus currently being made to develop various tools, including visualization tools, to interpret and explain difficult machine learning methodologies.

III

Methods

1. Statistical Techniques of Machine Learning: Pros and Cons
2. Model Evaluation

III Methods

1. Statistical Techniques of Machine Learning: Pros and Cons

There is almost an infinite variety of statistical techniques used in predictive MLB models today, including deep learning-based neural network, regression, shrinkage method, decision tree, boosting, random forest, support vector machine (SVM), bagging, and deep learning. Our discussion in this section focuses on elastic net, decision tree, random forest, boosting, and SVM, which are logistic regression and shrinkage methods applicable to models with binary response (dependent) variables.

The biggest advantage of logistic regression models is that they are stable. The coefficients are also easy to interpret and not so difficult to calculate. The few shortfalls of these models include the facts that they: provide linear decision boundaries, require new variables to be generated in order to account for interaction between two existing variables, and are not invariant to changes in independent variables.

Decision tree runs on easy and simple rules stated in "if-then" form. They facilitate sorting, are capable of processing both continuous and binary variables, and support non-parametric analyses that do not require certain assump-

tions (such as homoscedasticity and linearity in linear regression models). Decision tree also begins with the most explanatory variables and are relatively less sensitive to outliers.

However, the predictive power of decision tree tends to decline in regression models with continuous variables. Complex decision tree, in general, is less predictive and more difficult to interpret and may involve significant amounts of calculation depending on the given circumstances. In addition, the outcomes may not be so reliable when the Bayes classifier borders are not rectangular. Decision tree also carries high risk of significant variance in relation to small changes in the given data.

Boosting algorithms are highly predictive, remain invariant to changes in independent variables, and support analyses of nonlinear effects with their amenability to interaction terms. However, it can be quite difficult to tune the parameters that are used in boosting models.

Random forest, too, is quite powerful tools of prediction and can generate reliable outcomes even when numerous independent variables are involved. Random forest is also comparatively less sensitive to outliers and remain invariant to changes in independent variables. These, too, support analyses of nonlinear effects using interaction terms. However, random forest lacks extensive theoretical support, and interpretation of their final outcomes can be tricky.

SVM classifies multidimensional spaces into hyperplanes and

is favored by corporations for their predictive power. SVM can be applied to classification and regression problems alike, remains indifferent to data noise, and are not given to overfitting.

However, excessive numbers of training dataset samples and dimensions may slow SVM down. It is also important to set the kernels and tuning parameters properly in order to develop optimal models. SVM-based model is also not easy to interpret.

2. Model Evaluation

In evaluating a given model of statistical analysis, it is ideal to compare it to a number of other models by applying them to the same given data. The model that optimally explains the given data is the ideal choice. In selecting such an optimal model, it is important to compare and assess multiple different models and demonstrate that the chosen model is superior to the others.

The following four factors should be considered in evaluating models.

○ Predictive power: How well do the models predict outcomes?

○ Analytical power: How well do the models explain the correlation between the input and output variables?

○ Efficiency: Do the models minimize the number of input variables required?

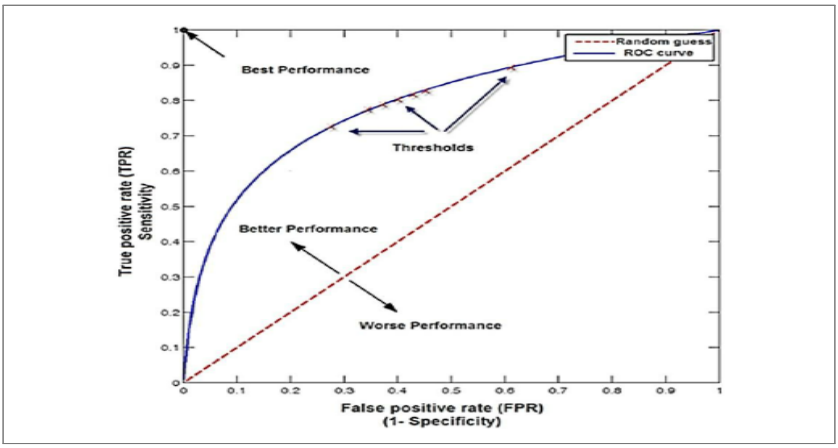
○ Stability: Do the models produce the same outcomes when applied to other data of the same target population?

While all four of the factors above are important in the evaluation of statistical models, the most important factor, especially in relation to problems concerning prediction, is predictive power. It is possible for a model to be the most stable, efficient, and analytical of all the given options but still yield unreliable and imprecise outcomes. In evaluating predictive models, it is therefore crucial to consider whether the models, presumably designed for prediction, offer better predictions than random models, as well as which of the models has the greatest predictive power.

Regression models can be compared in terms of Mallow's Cp and/or the adjusted R^2, while classification models can be compared in terms of misclassification rates (Kim, Chu, Choi, Oh, and Kim et al., 2015).

The receiver operating characteristic (ROC) curve visualizes the performance of a given statistical model in terms of sensitivity and false positive rate (FPR). Originating from the signal decision theory, this instrument is intended to separate signals and noise.

[Figure 3-1] ROC Curve



Source: Hassouna, M., Tarhini, A., and Elyas, T. (2015). Customer Churn in Mobile Markets: A Comparison of Techniques. *International Business Research*. Vol 8(6), pp. 224-237.

The area under the curve (AUC) is the indicator of the rationality of the given classifier. The larger the AUC, the better the model.

The lift chart is another instrument used to evaluate statistical models. A lift chart is created as follows. First, the posterior probability is estimated based upon the fitness of the given model, and the data is arranged according to the order of the posterior probability. The arranged data is then divided into N equal parts, and the frequency of the target variable's certain categories is estimated at each level of the N-parts. The % captured response, % response, and lift statistics are then estimated at each level. The lift chart is then plotted, with the x-axis representing the levels of the N-parts and the y-axis

representing one of the three types of statistics so estimated.

Two of the three types of statistics, % captured response and % response, can be defined as follows.

%Captured response=frequency of the target variable's certain categories at a given level * 100 / frequency of the target variable in the same categories at all levels

%Response=frequency of the target variable in certain categories at a given level x 100 / frequency of the target variable at all levels.

% response indicates the precision of the data at the upper x-percent of the given predicted probability when the predicted probabilities are arranged according to the given order. In general, % response decreases with the levels.

If the baseline lift were to be expressed as the frequency of the target variable in certain categories multiplied by 100 and divided by all-time frequency, the lift would be defined as the % response divided by the baseline lift at the given level.

IV Results

1. Establishing a Database for Analysis Using a Predictive Model for Social Security Benefits
2. Results
3. Chapter Conclusion

IV Results

1. Establishing a Database for Analysis Using a Predictive Model for Social Security Benefits

Because a system has already been established for using the analysis of social security big data to identify the blind spots in the welfare system, the results of which are being applied to policymaking today, the database for our trial was structured similarly, drawing upon the Korean Welfare Panel Survey data. The response variable (y) represents whether households receive livelihood support from the government. Of the diverse factor variables (x), four household types were considered: single-person households; single-mother households (consisting of mothers and children under 18 years old or children under the age of 22, if enrolled in school); single-father households (consisting of fathers and children under 18 years old or children under the age of 22, if enrolled in school); and parentless households (with household heads under the age of 18 or household heads who are seniors living with their underage grandchildren). The number of household members was also taken into account, along with information on household heads (education, capability to work, work competency, etc.) and household members (chronic morbidities and disabilities, etc.). As for medical risks, experiences of defaulting on National

Health Insurance (NHI) premium payments and lengths of payment delays were considered. As for housing risks, information on whether households owned or rented the homes in which they lived was examined. Household financial situations were estimated based on disposable income, taxes, total cost of living, total debt, and total assets. As for deprivations, information on whether households experienced disconnections of water or electricity services due to their inability to pay their utility bills, failed to secure heating in winter time, and/or were burdened with credit delinquencies was examined. Finally, financial difficulties, unemployment, medical conditions, alcoholism, and housing problems were used as variables representing issues of concern to households.

The models of statistical analysis that were compared included the logistic regression, elastic net, decision tree, boosting, random forest, SVM, and deep learning models. Because there were a significant number of missing values in the data on total assets, the variable was ultimately dropped from the analysis. The variables found to be insignificant through the t-test were also included in the analysis in order to compare the models. A 10-fold cross-validation was then conducted to develop and evaluate the models.

The models were applied to the overall data in order to identify the major variables.

Under the decision tree model, the variables found to be im-

portant in identifying households in need of social security support were the amounts of taxes paid by households, whether households lived in homes on monthly rents, and the degree of work competence of the household heads.

Under the boosting model, the important variables included the amount of tax, monthly rent, total cost of living, experiences of poor nutrition due to financial difficulties, amount of disposable income, disabilities of household members, household type, and capability of household heads to work.

2. Results

Accuracy, specificity, and sensitivity vary depending on the levels of the classifiers used in the classification models. The accuracy, specificity, and sensitivity of models with classifiers of 0.5, 0.6, 0.7, 0.8, and 0.9 are listed in the tables below.

〈Table 4-1〉 Results: Classifier of 0.5

	logistic	elastic net	decision tree	boosting	random forest	svm	deep learning
Accuracy	91.91%	91.91%	91.82%	92.38%	91.92%	91.24%	90.79%
Specificity	97.83%	97.87%	98.69%	97.75%	97.45%	98.29%	95.83%
Sensitivity	38.91%	38.61%	30.33%	44.38%	42.46%	28.10%	45.71%

〈Table 4-2〉 Results: Classifier of 0.6

	logistic	elastic net	decision tree	boosting	random forest	svm	deep learning
Accuracy	91.79%	91.80%	91.82%	92.09%	92.16%	91.16%	91.00%
Specificity	98.86%	98.86%	98.69%	98.68%	98.66%	98.71%	96.79%
Sensitivity	28.55%	28.70%	30.33%	33.14%	34.02%	23.66%	39.20%

〈Table 4-3〉 Results: Classifier of 0.7

	logistic	elastic net	decision tree	boosting	random forest	svm	deep learning
Accuracy	91.21%	91.19%	91.82%	91.39%	91.54%	90.88%	90.94%
Specificity	99.29%	99.27%	98.69%	99.22%	99.22%	98.90%	97.53%
Sensitivity	18.93%	18.93%	30.33%	21.30%	22.78%	19.08%	31.95%

When sensitivity was selected as the main criterion of evaluation, the decision tree model fared the best. This model, however, predicted 205 cases with an observed value of one when classifiers 0.5 through 0.7 were used, while predicting all cases to have an observed value of zero when classifiers 0.8 and 0.9 were used.

This suggests that more than a single criterion should be analyzed when the response variable shows significant difference between the number of cases with an observed value of zero and other cases with an observed value of one.

〈Table 4-4〉 Results: Classifier of 0.8

	logistic	elastic net	decision tree	boosting	random forest	svm	deep learning
Accuracy	90.41%	90.41%	89.94%	90.79%	90.79%	90.82%	90.86%
Specificity	99.69%	99.69%	100%	99.69%	99.77%	99.47%	98.31%
Sensitivity	7.40%	7.40%	0%	11.24%	10.50%	13.46%	24.26%

〈Table 4-5〉 Results: Classifier of 0.9

	logistic	elastic net	decision tree	boosting	random forest	svm	deep learning
Accuracy	90.06%	90.05%	89.94%	90.14%	90.08%	90.45%	90.74%
Specificity	99.95%	99.95%	100%	99.93%	99.95%	99.76%	99.02%
Sensitivity	1.63%	1.48%	0%	2.51%	1.78%	7.10%	16.71%

When accuracy, specificity, and sensitivity were all considered as the evaluation criteria, the boosting model had the best performance overall. The deep learning (CNN) model, however, performed significantly better than the other models in terms of sensitivity.

〈Table 4-6〉 AUC

Model	AUC
Logistic	0.9058
Elastic Net	0.9059
Decision Tree	0.8335
Boodsting	0.9161
Random Forest	0.9101
SVM	0.8530
Deep Learning(CNN)	0.8726

The boosting and random forest models fared the best in terms of both the ROC curve and AUC.

The models were also evaluated in terms of the % response of lift charts by assigning the top five percent, top 10 percent, top 15 percent, top 20 percent, top 25 percent, and top 30 percent of posterior probabilities to Level 1.

The tables below list the estimates for the mean % response of the 10 test sets yielded by the 10-fold partition method.

〈Table 4-7〉 % Response of Lifts by Level

Level	logistic	elastic net	decision tree	Boosting	Ranf	svm	Deep Learning	random
5%	11.7	23.6	24.1	24.6	24.3	12.9	14.5	3.4
10%	29.5	41.1	44.4	44.4	41	32.9	32.6	8
15%	55	66.9	70.4	68	65.7	57.3	56.8	10.3
20%	81.6	94.4	98.1	96.2	94.2	84.4	82.4	14.3
25%	111.4	124.3	126.2	126.3	124.5	113.4	110.7	16.8
30%	141.6	156.1	159.4	157.6	156.2	144.2	138.9	20.7

〈Table 4-8〉 Precision of % Response of Lifts by Level

	logistic	elastic net	decision tree	Boosting	Ranf	svm	Deep Learning	random
5% acc	34.41	69.41	70.88	72.35	71.47	37.94	42.65	3.4
10% acc	44.03	61.34	66.27	66.27	61.19	49.10	48.66	8
15% acc	54.46	66.24	69.7	67.33	65.05	56.73	56.24	10.3
20% acc	60.9	70.45	73.21	71.79	70.3	62.99	61.49	14.3
25% acc	66.31	73.99	75.12	75.18	74.11	67.50	65.89	16.8
30% acc	70.1	77.28	78.91	78.02	77.33	71.39	68.76	20.7

The % response of lift charts, too, show the boosting and random forest algorithms to be the best performers, while the decision tree model was found to be rather useful here as well. The deep learning (CNN) model, on the other hand, did not fare as well according to this measure as it would have for image or language analyses.

3. Chapter Conclusion

In this chapter, a database similar to that used in an existing social security big data analysis system for policy making aid was developed. The database so developed was used to conduct an analysis of factors that influence households' need for social security benefits and a comparative analysis of multiple predictive models. The 10-fold cross-validation method was used to evaluate the predictive models, thus avoiding the problem of overfitting.

The boosting model fared the best overall, in terms of misclassification rates, AUC, ROC curve, and the % response of lifts. The latest deep learning (CNN) model performed significantly better than the other models in terms of sensitivity. This, however, is because the deep learning model tends to overestimate the number of households in need of social security benefits. Using the results of this model would reduce the probability of households in need being denied social security

benefits. However, this model could also lead households that do not need social security benefits to actually receive them, thus increasing the fiscal burden on the state. The choice of a predictive model in areas such as social security analysis should therefore be based upon not only one single criterion, such as precision, but other important factors as well. The deep learning model, in the meantime, fared poorly in comparison to other MLB models with respect to a database consisting of binary-type dependent variables. Further research is needed to determine whether deep learning models can yield reliable predictions in relation to databases that include numerous binary variables among independent variables.

The logistic regression and boosting models could be applied to binary dependent variables measured in terms of % response of lifts (top five percent of test data in terms of probabilities). Here, the boosting model performed at least twice as well as the logistic regression model.

A one-percent difference in the accuracy of the models, when applied to the discovery of blind spots in a welfare system, means that, when a local government is given a list of 10,000 potentially at-risk households, it may be able to deliver policy services and benefits to 100 additional households. This one-percent difference may not seem significant in the evaluation of analytical models, but it can make a major difference in the efficiency of public administration in the real world.

The policy implications of our analysis of the predictive power of various statistical models for social security demand can be summarized as follows. Statistical models can help us identify and determine individuals and households that could be in need of but have so far been denied social security benefits. The predictions made by these models can also be compared to the actual distribution of social security benefits in the real world to produce statistical estimations of the proportions of the population benefitting from not just the National Basic Livelihood Security Program but also other support programs for the poor, the elderly, and the disabled, thus providing greater empirical evidence for policymakers to enhance the reach and scope of these welfare programs. Predictive models based upon social security big data thus show great potential for broad application in the area of social policymaking.

V

Conclusion and Policy Implications

V Conclusion and Policy Implications

This study sets itself apart from the existing literature by applying diverse MLB statistical models, including deep learning, to a newly developed database that is similar to the social security big data already used in policymaking, with a view to identifying the optimal MLB model. Based on our database, boosting algorithms showed the best performance. Although deep learning is superior to other models in terms of image and language processing, boosting and random forest models outperform deep learning in relation to databases such as ours. Our findings suggest that there is not one single MLB model that performs the best with respect to all types of data. Although the boosting model was found to be the most consistently precise and predictive in our analysis, other MLB models may be better if the nature of the data changes. If social security big data, for example, were made available in the forms of images and videos, deep learning would likely be better than boosting. Therefore, before applying machine learning to policy analysis, we need to first ensure that we have a good grasp of the attributes of the given policy data. Systems for identifying social security blind spots require information on the demographic, financial, housing, and other characteristics of potential social security beneficiaries and non-beneficiaries.

Systems for identifying potentially at-risk children require information on the attributes of abused and non-abused children as well as on abusive and non-abusive parents. Video data on healthcare should be designed so that models can be used to detect abnormal data.

As with other areas of social policymaking, data innovation and technological progress are also required in healthcare policymaking. Whereas data analysis in the past was done primarily to provide simple information on occurring phenomena and/or their causes, data analysis in today's era of the Fourth Industrial Revolution should be capable of generating optimized solutions for policy problems on the basis of predictions about the future. Policymakers can help improve the lives of citizens by developing and publicly sharing databases of health and social big data and applying MLB predictive models to find optimized policy solutions.

To establish a data-centered system of predictive and preventive health and welfare policymaking, one that enables policymakers to make informed decisions in a scientific manner, it is important to transform the overall decision-making structure of the government and its health and welfare agencies with a central emphasis on data. Most importantly, it is crucial to establish infrastructure for the collection and management of health and welfare big data in order to enable the design of policy objectives, systems, and feedback according to pre-

dictions made on the basis of such data infrastructure. Such infrastructure will be pivotal in the revision and effective implementation of policies. Furthermore, government support and social trust should be fostered to establish an effective structure for data governance and data-sharing platforms so as to ensure the active application of such data.

If supported by such a data-centered policymaking system, machine learning in Korea would achieve unprecedented progress and make profound contributions to the efforts being made to improve the quality of life for Korean citizens.

References

Kim, E., B. Chu, H. Choi, M. Oh and Y. Kim et al. (2015). Identifying Welfare Blind Spots Using a Social Security Information System. Seoul: Social Security Information Service.

Kim, E. (2016). Learning Artificial Intelligence with Algorithms: Introduction to Machine Learning and Deep Learning. Seoul: Wikibooks.

The Joongang Ilbo (October 19, 2017). "AlphaGo Upgraded: No Longer Requiring Human Go Manuals." http://news.joins.com/article/22026687 (retrieved November 29, 2017).

Choi, D. (2017). Big Data and AI in the Age of the Fourth Industrial Revolution. special lecture at ICT Convergence Korea 2017. Seoul.

Castrounis, A. (2016). Artificial Intelligence, Deep Learning, and Neural Networks, Explained. https://www.kdnuggets.com/2016/ 10/artificial-intelligence-deep-learning-neural-networks-explained.html (retrieved November 29, 2017).

Domingos, P. (2012). A Few useful things to know about machine learning. New York: *Communications of the ACM*, Volume 55(10), 78-87.