*International Journal of Advanced Culture Technology Vol.10 No.4 499-510 (2022)*
*DOI https://doi.org/10.17703/IJACT.2022.10.4.499*

IJACT 22-12-59

# Predicting Reports of Theft in Businesses via Machine Learning

[1]JungIn Seo, [2]JeongHyeon Chang*

*[1]Prof., Dept. of Information Statistics, Andong National., Univ., Korea*
*[2]Prof., Contents Convergence Software Research Center, Kyonggi, Univ., Korea*
*leehoo@gmail.com, crime_tiger564@kyonggi.ac.kr*

***Abstract***

*This study examines the reporting factors of crime against business in Korea and proposes a corresponding predictive model using machine learning. While many previous studies focused on the individual factors of theft victims, there is a lack of evidence on the reporting factors of crime against a business that serves the public good as opposed to those that protect private property. Therefore, we proposed a crime prevention model for the willingness factor of theft reporting in businesses. This study used data collected through the 2015 Commercial Crime Damage Survey conducted by the Korea Institute for Criminal Policy. It analyzed data from 834 businesses that had experienced theft during a 2016 crime investigation. The data showed a problem with unbalanced classes. To solve this problem, we jointly applied the Synthetic Minority Over Sampling Technique and the Tomek link techniques to the training data. Two prediction models were implemented. One was a statistical model using logistic regression and elastic net. The other involved a support vector machine model, tree-based machine learning models (e.g., random forest, extreme gradient boosting), and a stacking model. As a result, the features of theft price, invasion, and remedy, which are known to have significant effects on reporting theft offences, can be predicted as determinants of such offences in companies. Finally, we verified and compared the proposed predictive models using several popular metrics. Based on our evaluation of the importance of the features used in each model, we suggest a more accurate criterion for predicting var.*

***Keywords:*** *Evaluation Metric; Feature Importance; Machine Learning Algorithm; SMOTE and Tomek Link; Theft Report*

## 1. INTRODUCTION

To prevent and solve crimes, it is critical to report crime [1]. Contrary to popular belief that society is made safe through policing and law enforcement alone, the fact that each crime is unique makes it difficult to solve all crimes through police efforts and tasks alone [2]. In order to detect crime and maintain public safety, cooperation between members of the police force and their communities is essential. Factors that are part of this type of cooperation include willingness to report a crime, which is known to minimize the risk of important information going undetected by police. It is also an important means of responding to public order needs in cases where the police are understaffed [3]. For example, Gottfredson described individuals who report crimes as "gatekeepers" in the criminal justice context, while Take argued that crime reporting was an individual's

Manuscript received: November 26, 2022 / revised: December 4, 2022 / accepted: December 9, 2022
Corresponding Author: crime_tiger564@kyonggi.ac.kr
Tel: +82-31-249-8962, Fax: +82-31-249-8963
Research Professor, Contents Convergence Software Research Institute, Kyonggi. Univ., Korea

realistic option for gaining relief from criminal damage [4, 5]. Crime reporting is an essential component of crime prevention and victim relief. However, not all crimes are reported, a phenomenon known as "no-reporting"; this often results in secondary crimes and/or negligence. Several studies have confirmed that the severity of criminal harm impacts the relationship between crime reporting and the willingness of reporting [6]. Researchers have also identified relevant causal factors and causal models, including those pertaining to reporting costs/benefits, crime seriousness, how the police perceive procedural justice (legitimacy) and social/cultural class [7-11]. In turn, these types of complex models have enabled empirical research on the determinants of crime reporting [12]. To date, relevant studies have focused on the factors of crime reporting in the context of "individual harm," neglecting the factors of reporting in the context of a business or in other areas of public harm. From an economic perspective, reporting a crime is considered an act aimed at the victim's profit [13]. This means that an individual victim of crime decides whether to "report" a crime after calculating the benefit (gain) and the cost (loss). If the costs are low and the benefits are high, a victim is more likely to report the crime to the police [15].

Many studies have examined the determinants of individual crime reporting based on an economic perspective (cost-benefit theory). Previous studies can be divided into two types. One consists of studies in which variables measuring the aforementioned costs and benefits are introduced into the model, and the other comprises studies based on the seriousness of the crime. First, since costs and benefits show great variation in both type and degree, their definitions and measurements differ in each study. Typically, an analysis of the opportunity cost of reporting is performed by comparing Internet and telephone reporting, stigma to neighbors, community safety, protection from perpetrators, and insurance coverage [16, 18] Factors measured as benefits of crime reporting increased the probability of crime reporting, while factors measured as costs of crime reporting decreased the probability of reporting. And then, some studies explain the reporting of victims based on the severity of the crime. Crime severity is discussed as the most important factor in crime reporting. This is because the nature and severity of crime play a central role in the cost-benefit decision [19]. The cost-benefit theory assumes that victims weigh benefits and costs when deciding whether to report a crime, which provides a basis for predicting the relationship between crime severity and the reporting of the crime. According to, the more serious the crime, the greater the desire for retribution and protection. In other words, the more serious the crime, the more the victim values the benefit of reporting it than the cost [20]. property crimes, for example, the higher the value of the stolen goods (damage), the greater the benefit of recovering them by reporting the theft. Moreover, for serious violent crimes involving the use of a weapon, bodily injury, or invasion of privacy, the utility of punishing the offender or securing physical protection increases. This discussion has been confirmed by several previous studies that have consistently concluded that the more serious the crime, the higher the reporting rate, both domestically and internationally [21-23]. As such, the report of a crime victim is related to the recognition of the scale of damage. And it should be noted that the damage perceived by business victims is relative to individual victims.

Therefore, in this study, we propose predictive modelling that verifies the factor of reporting a crime against a business, rather than individual willingness to report a crime or micro-level analyses. As such, we do so through predictive modelling via two approaches:(1) statistical models, including logistic regression (LR) and elastic net (EN); and (2) popular machine learning models such as support vector machine (SVM), tree-based models, and stacking. In addition, it is also a prevention model to prompt reporting to prevent corporate theft.

## 2. EXPERIMENTS

This section briefly describes statistical models and machine learning techniques that were considered for analysis.

### 2.1 Logistic Regression

Cox introduced the LR model that expands on the ideas of linear regression in order to include situations in which the target is categorical [24].

Let $X = (X_1, \ldots, X_p)$ be a p-dimensional feature vector in the LR model. Then, the logistic response function that produces the S-shaped curve is given by:

$$p(X) = \frac{e^{\beta_0 + \sum_{i=1}^{p} \beta_i X_i}}{1 + e^{\beta_0 + \sum_{i=1}^{p} \beta_i X_i}}, \tag{1}$$

where $\beta_0$ is the intercept and $\beta_i$ is the coefficient corresponding to the $i$th feature $X_i$. Note that the logistic response function(1)may be interpreted as the probability that an observation is in a specified category of the dichotomous outcome, generally called the "success probability". In addition, its logit transformation leads to the standard formulation of a logistic model given by:

$$\begin{aligned} \mathrm{g(X)} &= \log\left(\frac{p(x)}{1 - p(x)}\right), \\ &= \beta_0 + \sum_{i=1}^{p} \beta_i X_i \end{aligned} \tag{2}$$

which is a linear function of the feature vector $X$.

### 2.2 Elastic Net

Given its many features, the Ordinary Least Squares (OLS) regression model has some problems, particularly the presence of multi-collinearity and the risk of overfitting the training data. As such, several regularized regression models have been proposed as alternatives to the OLS, as constraining or regularizing estimated coefficients reduces variance and sampling errors [25]. This approach overcomes the drawbacks of the OLS regression model by constraining the total size of all coefficient estimates. To constrain the size of the coefficients, the ridge regression adds a penalty on the sum of the squared regression parameters:

$$\mathrm{SSE}_{L_2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{p} \beta_j^2 \tag{3}$$

where $L_2$ indicates that a second-order penalty is being used on the parameter estimates and $\lambda_1 (\geq 0)$ in the penalty term is a tuning parameter that controls how much of the model is regularized. In addition, $y_i$ and $\hat{y}_i$ represent the target for the $i$th sample and its predicted value, respectively.

This approach can reduce the overall mean squared error (MSE) compared to the unbiased model by reducing the variance, but unfortunately cannot conduct feature selection. To improve model performance and simultaneously conduct feature selection, the least absolute shrinkage and selection operator model (LASSO) will be considered, which uses a penalty similar to ridge regression [26]:

$$\mathrm{SSE}_{L_1} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j| \tag{4}$$

where $\lambda_2 (\geq 0)$ is a tuning parameter that plays the same role as $\lambda_1$. Another alternative is the EN model, which combines both types of penalties [27]:

$$\mathrm{SSE}_{EN} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j| \tag{5}$$

Where, the advantage of this model is that it can effectively regularize through the penalty term in the ridge model along with the feature selection quality of the penalty term in the LASSO model.

### 2.3 Support Vector Machine

The SVM can be used to find a hyperplane that best separates classes by loosening what "perfect separation" means or using so-called kernel tricks to enlarge the feature space using basis functions [28]. The popular kernel functions are given by:

- $d$-th degree polynomial: $\mathrm{K}(\mathrm{x}, x') = \gamma(1 + \langle x, x' \rangle)^d$
- Radial basis function: $\mathrm{K}(\mathrm{x}, x') = \exp\,(\gamma \|x - x'\|^2)$
- Hyperbolic tangent: $\mathrm{K}(\mathrm{x}, x') = \tanh\,(k_1 \|x - x'\| + k_2)$ (6)

Where $\langle x, x' \rangle = \sum_{i=1}^{n} x_i\, x'_i$ is an inner product, and $||z||$ is the Euclidean norm. In addition, $\gamma$ is a scale parameter, $d$ is a degree term in the $d$-th degree polynomial, $\gamma$ is a parameter related to the inverse of the $\sigma$ parameter of a normal distribution in the radial basis function, and $k_i(i = 1, 2) > 0$.

### 2.4 Random Forest

The random forest (RF) is an ensemble technique that improves performance by combining several decision trees; this is based on bagging (bootstrap aggregating) and randomness [29]. Bagging reduces the impact of noisy data and overfitting. In addition, when fitting a tree model to each bootstrap sample, the RF algorithm randomly selects the proportion of features considered for that split at each node of the tree. This forces the model from each bootstrap to include different features and create individual trees that are highly uncorrelated. The RF can be implemented as shown in [Figure 1].

```
Select the number of models to builds, m
for i = 1 to m do
    Generate a bootstrap sample of the original data
    Train a tree model on this sample
    for each split do
        Randomly select l(< p) of the original features
        Select the best feature among the l features and partition the data
    end
    Use typical tree model stopping criteria to determine when a tree is complete (but do not prune)
end
```

**Figure 1. Basic RF algorithm for classification (2-class)**

### 2.5 eXtra Gradient Boost

Unlike the RF with bagging, the boosting algorithm improves errors by sequentially learning and predicting multiple weak learners, and assigning weights to incorrectly predicted data. The AdaBoost and gradient boost algorithms are most often used to implement this [30, 31]. The former performs boosting by weighting the error data, while the latter boosts by updating weights through gradient descent. However, the computation time for the boosting is often greater than for the RF, since the RF can easily be processed in parallel, given that the trees are created independently, while the boosting is an ensemble technique that, again, trains many individual models but builds them sequentially.

Among the gradient boost algorithms, the eXtra Gradient Boost (XGBoost) solves the disadvantages of the gradient boost, such as slow execution time through parallel processing [32]. It also optimizes the gradient boost algorithm by providing tree-pruning, handling missing data, and implementing regularization to avoid overfitting.

### 2.6 Stacking

The ensemble is a technique to derive a more accurate final prediction by generating and combining multiple prediction models. It includes stacking and voting, as well as the aforementioned bagging and boosting.

Bagging and boosting improve prediction results by combining multiple individual models obtained by applying the same algorithm to the same dataset, while voting and stacking improve prediction results by combining multiple individual models obtained by applying different algorithms to the same dataset. Particularly in regard to stacking, the results predicted by individual models are again used as training datasets to learn, which requires another predictive model as a final model.

Here, a stacking technique that enhances predictive performance by combining multiple individual models through a robust linear combination based on a generalized linear model (GLM) is applied [33]. The final model of this approach is given by:

$$g(\hat{Y}) = \beta_0 + \sum_{i=1}^{p} \beta_i \hat{Y}_i \tag{7}$$

where $\hat{Y}_i$ is the predicted value from the ith individual model.

## 3. RESULTS

To evaluate and compare the performance of the considered predictive models, we consider the Receiver Operating Characteristic (ROC) curve, which plots the sensitivity (Sen) and specificity (Spe) pairs as the cutoff value (CV) descends from 1 to 0; the closer the curve is to the top-left corner, the better the performance. In a prediction problem with two classes, assuming that the two classes are "event" and "nonevent," the Sen is defined as the rate at which the event of interest is correctly predicted for all samples having the event, while the Spe is defined as the rate that nonevent samples are predicted as nonevents. The evaluation and comparison through the ROC can be accomplished by calculating its corresponding area under the ROC curve (AUC). Furthermore, the predictive power of the predictive models can be measured by comparing the overall accuracy rate (Acc), Sen, and Spe at different CVs since the ROC curve evaluates the model's performance at different CVs. The uncertainty is examined through the 95% confidence interval (CI) for the metric that is computed with 2,000 stratified bootstrap replicates.

### 3.1 Data and Preprocessing

This study used data collected through the 2015 Commercial Crime Damage Survey conducted by the Korea Institute for Criminal Policy (https://www.data.go.kr/data/15095483/fileData.do). The target population was comprised of businesses considered "wholesale and retail" and "accommodation and restaurant" under the "Korea Standard Industrial Classification." The representation group used the same definition of "National Business Survey" as a sampling frame in the relevant data as a "business entity." This is the same concept as all management units that independently perform economic activities (e.g., the production, sale, and service of goods) in certain physical places (e.g., individual stores, offices, banks, schools, hospitals, inns, restaurants, academies, churches, public institutions, and social welfare facilities, regardless of profit/non-profit status). The survey consisted of face-to-face interviews, and self-written responses in which the investigators visited the businesses in person.

Through this, data on a total of 8,140 businesses were collected, and 961 businesses that experienced theft crimes were extracted among the crimes committed against businesses. However, among 961 business data, data with no damage due to theft (no stolen goods) and missing values were excluded, and only 834 data that actually occurred were used for analysis. Table 1 describes the features considered for analysis from within this dataset.

**Table 1. Features of interest within the theft reporting data**

| Feature | Type | Description |
|---|---|---|
| crime reporting | Categorical | Report of theft<br>(1=Yes, 0=No) |
| theft_price | Numeric | Amount of damage |
| invasion | Categorical | Trespassing<br>(2=Success, 1=Attempt, 0=Not) |
| remedy_dm | Categorical | Remedy by Offender<br>(1=Yes, 0=No) |
| insurance | Categorical | Insurance subscription<br>(1=Yes, 0=No) |
| CE_dm | Categorical | Participation of Community Environment<br>(1=Postive, 0=Negative) |
| SEX | Categorical | Sex<br>(1=Male, 0=Female) |
| one_person | Categorical | Manage<br>(1=Only, 0=Not only) |
| residence | Categorical | In work place<br>(1=Inside, 0=Outside) |
| type | Categorical | Business type<br>(4=Bar, 3=Restaurant, 2=Accommodation, 1=Sales) |
| franchise | Categorical | Managing a franchise<br>(1=Yes, 0=No) |
| income | Categorical | Monthly income (Million won)<br>(6=income$\geq$ 1, 000, 5=500 $\leq$income< 1, 000,<br>(4=100 $\leq$income< 500, 3=50 $\leq$income< 100,<br>2=30 $\leq$income< 50, 1=income< 30) |

For analysis, the full dataset was partitioned into training and test sets at a ratio of 70 to 30, respectively, using a stratified train-test split that splits the data set into training and test sets in a way that preserves the proportion of each class as observed in the original data set. Note that this dataset was affected by the problem of class imbalance, which can significantly impact model predictions and performance [34]. To remedy this, we sequentially applied the Synthetic Minority Over-Sampling Technique (SMOTE) and Tomek link to the training data prior to analysis [35, 36]. The former is a method of generating new data that is slightly different from the existing data by finding the k nearest neighbor of the individual data in a small dataset, then making the difference between these data and k neighbor a constant value. However, this synthetic data generation method generates a new data sample that causes class overlaps, which results in the problem of overgeneralization. This can be overcome by the latter, which cleans up unwanted overlapping between classes by removing pairs of minimally distanced nearest neighbors of opposite classes.

While training the model using the training data, we performed 5-fold cross-validation by splitting the original training data into five partitions to both avoid overfitting and tune the model's hyper parameters. For the tree-based machine learning models, such as the RF and XGB, we applied the out-of-bag (OOB) error to evaluate model performance in addition to cross-validation. Figure 2 shows the overall framework for prediction.

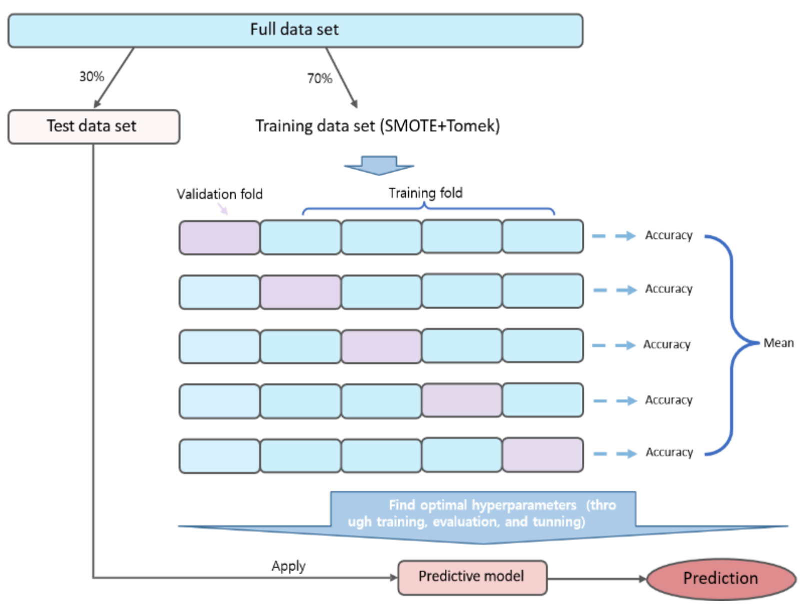


**Figure 2. Prediction process diagram**

### 3.2 Experimental Results

This section reports the results of the metrics described in Section 3 to evaluate and compare the performance of the considered predictive model by employing test data. Figure 3 shows the plotting of the ROC curve for the considered prediction models and reports the values of the corresponding AUC and its Cis . As in Figure 3, we verified that the tree-based machine learning models (RF and XGB) and the Stacking model were better in terms of AUC, as evidenced by slightly higher AUC values when compared to other predictive models. In addition, in the case of the CI for the AUC, the tree-based machine learning models (RF and XGB) and Stacking model have shorter widths than other predictive models, which reveals that they are also better predictive models in terms of uncertainty.

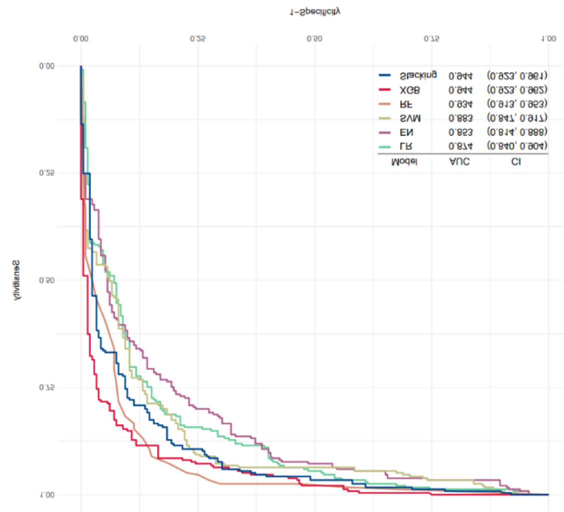

**Figure 3. ROC curve for the considered predictive models**

However, there was little difference between the tree-based models. A closer examination was performed by evaluating the predictive models not only at the default CV of 0.5 but also at the best CV, which is the point at which the sum of Sen and Spe is maximum. In doing this, the Acc that is the simplest metric and the most widely used Sen and Spe were computed at the best CV and the default CV of 0.5. These values are reported in Table 2, and the 95% CIs for the metrics in each predictive model are reported in parentheses below the predicted values. Table 2 shows the same results as Figure 3. That is, the tree-based machine learning models (RF and XGB) and the Stacking model have better overall performance on all metrics and their CIs. In addition, the predicted Sen at the best CV showed a significant improvement over the value generated by the default CV of 0.5 while worsening in other metrics. This means that the considered predictive model is better at correctly predicting the class of interest label in the theft reporting dataset at the best CV. To induce a business to report, it is thus suggested to use the predictive models at the Best CV.

**Table 2. Prediction using the default and alternate cutoffs**

| | 0.5CV | | | Best CV | | |
|---|---|---|---|---|---|---|
| Method | Acc | Sen | Spe | Acc | Sen | Spe |
| LR | 0.837 (0.799, 0.871) | 0.736 (0.680, 0.792) | 0.868 (0.821, 0.909) | 0.815 (0.789, 0.857) | 0.813 (0.740, 0.985) | 0.816 (0.640, 0.880) |
| EN | 0.819 (0.779, 0.857) | 0.706 (0.646, 0.766) | 0.854 (0.803, 0.900) | 0.819 (0.792, 0.859) | 0.706 (0.598, 0.826) | 0.854 (0.729, 0.964) |
| SVM | 0.837 (0.801, 0.873) | 0.770 (0.710, 0.826) | 0.858 (0.817, 0.895) | 0.795 (0.769, 0.833) | 0.894 (0.731, 0.929) | 0.764 (0.732, 0.925) |
| RF | 0.877 (0.846, 0.906) | 0.847 (0.796, 0.894) | 0.887 (0.855, 0.919) | 0.839 (0.814, 0.870) | 0.928 (0.877, 1.010) | 0.811 (0.728, 0.862) |
| XGB | 0.882 (0.853, 0.909) | 0.881 (0.834, 0.920) | 0.882 (0.845, 0.914) | 0.879 (0.857, 0.908) | 0.885 (0.834, 0.932) | 0.877 (0.831, 0.928) |
| Stacking | 0.898 (0.867, 0.927) | 0.860 (0.813, 0.903) | 0.910 (0.873, 0.942) | 0.897 (0.873, 0.924) | 0.885 (0.834, 0.932) | 0.901 (0.855, 0.942) |

Figure 4 shows the feature importance score that is a score representing the "importance" of each feature for a given predictive model. A higher score means that the specific feature will have a larger effect on the predictive model. The feature the *price* represented the amount of damage and was the most influential in all predictive models in determining whether to report theft crimes. In addition, other models (except for the SVM predictive model) extracted the next most important features as *invasion2(success)* and *remedy_dm1(yes)*. In sum, the features the *price, invasion*, and *remedy*, which are known to have significant effects on the reporting of theft crimes, can be predicted as determinants of such crimes in businesses.

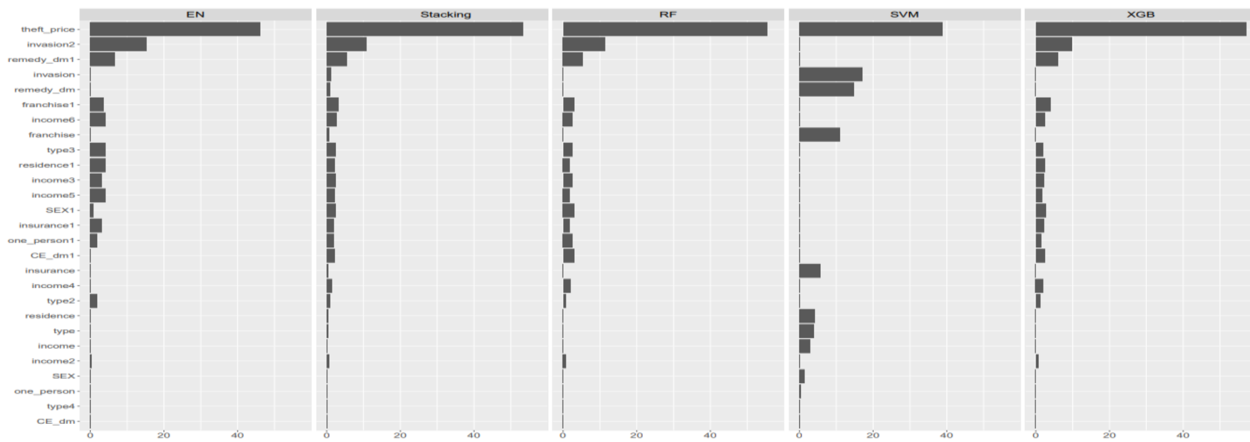


**Figure 4. Importance scores for features used in the considered predictive models**

## 4. CONCLUSION

In this study, we used machine learning techniques to predict the determinants of crime reporting in 'businesses', focusing on the individual perspective. In this regard, we used the economic cost-benefit theory as an investigative framework. Our model was shown to have sufficient goodness-of-fit in the profit calculation of reporting factors, as suggested from the economic perspective. Individual characteristics must also be taken into account to induce more frequent crime reporting in the business context. In other words, finding methods that increase the benefits of crime reporting is important. Of note, business damage is not typically recognized as harm to private property; this perception leads to a culture of tolerance for crime damage (shop thefts, credit card fraud, and property damage). As such, it is also important to induce the perception that benefits exist outside those which are derived from damage recovery.

This study also had some limitations. First, we did not apply all features that fit the proposed theory. However, our investigation was still novel because we designed a predictive model for crimes based on an economic theory. As for future areas of research, a given feature should be measured before and after a crime is reported in terms of the victim’s compensation. In this regard, it is necessary to consider two cases separately. This is because agreement-based compensation can hinder causal relationships between features. Thus, future studies would benefit by designing separate features. Second, we did not include community friendship or police reliability, which were both suggested in previous studies. In terms of the reporting benefits, police reliability is closely related to procedural impartiality; moreover, its influence has been verified by previous studies. Although police reliability is measured very subjectively, it should still be considered in future research.

Despite these issues, this study makes a meaningful contribution to the literature due to its employment of an interdisciplinary theory from a convergence perspective. From a practical standpoint, our findings should also be useful in predicting and preventing crimes. Finally, we expect to overcome the above limitations by further applying the proposed model to interpersonal crimes.

## ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2020R1A6A1A03040583).

## REFERENCES

[1] E. P. Baumer and J. L.Lauritsen, "Reporting crime to the police, 1973-2005: A multivariate analysis of long-term trends in the National Crime Survey (NCS) and National Crime Victimization Survey (NCVS),". *Criminology*. vol. 48, no. 1, pp. 131-185, 2010. https://doi.org/10.1111/j.1745-9125.2010.00182.x

[2] W. G. Skogan, "Concern about crime and confidence in the police: Reassurance or accountability?" *Police Quarterly*, vol. 12, no. 3, pp. 301-318, 2009. https://doi.org/10.1177/1098611109339893

[3] M. S. Kim and S. H. Kim, "An analysis of the conditions and problems of cooperative policing in Korea," *Korean Police Studies Association*, vol. 3, pp. 3-30, 2004.

[4] D. M. Gottfredson, "Prediction and classification in criminal justice decision making," *Crime and justice*, vol. 9, pp. 1-20, 1987. https://doi.org/10.1086/449130

[5] J. Y. Tak, "Reporting theft and fraud victimization in Korea," *Korean Police Studies Association*, vol. 83, pp. 53-75, 2010.

[6] W. G. Skogan, "Reporting crimes to the police: The status of world research," *Journal of research in crime and delinquency*, vol. 21, no. 2, pp. 113-137, 1984. https://doi.org/10.1177/0022427884021002003

[7] R. Bowles, M. G. Reyes, and N. Garoupa, "Crime reporting decisions and the costs of crime," *European journal on criminal policy and research*, vol. 15, no. 4, pp. 365-377, 2009. https://doi.org/10.1007/s10610-009-9109-8

[8] R. B. Felson, S. F. Messner, A.W. Hoskin, G. Deane, "Reasons for reporting and not reporting domestic violence to the police," *Criminology*, vol. 40m no. 3, pp. 617-648, 2002. https://doi.org/10.1111/j.1745-9125.2002.tb00968.x

[9] L. Zhang, S. F. Messner, J. A. Liu, "A multilevel analysis of the risk of household burglary in the city of Tianjin, China," *The British Journal of Criminology*, vol. 47, no. 6, pp. 918-937, 2007. https://doi.org/10.1093/bjc/azm026

[10] T. R. Tyler and C. J. Wakslak, "Profiling and police legitimacy: Procedural justice, attributions of motive, and acceptance of police authority," *Criminology*, vol. 42, no. 3, pp. 253-282, 2004. https://doi.org/10.1111/j.1745-9125.2004.tb00520.x

[11] F. Black, "The dividend puzzle," *J Portfolio Manag*, vol. 2, pp. 5-8, 1976. https://doi.org/10.1515/9781400829408-003

[12] C. H. Park, B.H. Ghu, "Victim-offender relationships and the reporting of crime: D. Black vs. feminist," *Journal of the Korean Criminology Assocciation*, vol. 12, pp. 121-137, 2018. https://doi.org/10.29095/jkca.12.2.6

[13] H. Goudriaan, K. Wittebrood, P. Nieuwbeerta, "Neighbourhood characteristics and reporting crime: Effects of social cohesion, confidence in police effectiveness and socio-economic disadvantage," *The British Journal of Criminology*, vol. 46, no. 4, pp. 719-742m 2006.

[14] J. R. Lasley, B. J. Palombo, "When crime reporting goes high-tech: An experimental test of computerized citizen response to crime," *Journal of criminal justice*, vol. 23, no. 6, pp. 519-529, 1995. https://doi.org/10.

1016/0047-2352(95)00043-7
[15] D. Young, "*See no evil*," Orca Book Publishers, 2006. https://www.orcabook.com/See-No-Evil-P2781.
[16] H. Goudriaan, P. Nieuwbeerta, "Contextual determinants of juveniles' willingness to report crimes," *Journal of experimental criminology*, vol. 3, no. 2, pp. 89-111, 2007. https://doi.org/10.1007/s11292-007-9030-4
[17] R. B. Felson, P. P. Paré, "The reporting of domestic violence and sexual assault by nonstrangers to the police," *Journal of marriage and family*, vol. 67, no. 3, pp. 597-610, 2005. https://doi.org/10.1037/e53595 2006-001
[18] LEE, Soochang, KIM, Daechan, "Relationship between Change of Demographic Composition and Crime: Comparing Areas with Growth in Population to Areas with Decline," *International Journal of Advanced Culture Technology (IJACT)*, 2022, 10.3: 63-70. https://doi.org/10.17703/IJACT.2022.10.3.63
[19] W. G. Skogan, "Chicago since 1840: A time-series data handbook," *IL: Institute of Government and Public Affairs*, University of Illinois, Urbana, USA, 1976.
[20] P. Campoy-Torrente, A. A. Chelini, C. Soto-Urpina, "Evaluación de la policía de proximidad en la ciudad de Santa Fe," *Urvio, Revista Latinoamericana de Estudios de Seguridad*, vol. 19, pp. 70-89, 2016. https://doi.org/10.17141/urvio.19.2016.2392
[21] A. Blumstein, "Seriousness weights in an index of crime," *American Sociological Review*, vol. 39, pp. 854-864, 1974.
[22] R. Broadhurst, J. Bacon-Shone, B. Bouhours, T. Bouhours, L. Kingwa, "*Business and the Risk of Crime in China*," ANU Press. 2011. https://doi.org/10.22459/brcc.12.2011
[23] A. A. D. Frate, "The international crime business survey: findings from nine central-eastern European citie," *European Journal on Criminal Policy and Research*, vol. 10, no. 2, pp. 137-161, 2004. https://doi.org/10.1007/s10610-004-4122-4
[24] D. R. Cox, "The regression analysis of binary sequences," Journal of the Royal Statistical Society: Series B (Methodological), vol. 20, no. 2, pp. 215-232, 1958.
[25] Hoerl, A. "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55-67, 1970.
[26] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267-288, 1996. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
[27] H. Zou, T. Hastie, "Regularization and variable selection via the ElasticNet," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301-320, 2005. https://doi.org/10.1111/j.1467-9868.2005.00503.x
[28] C. Cortes, V. Vapnik, "Support-vector networks Mach," Machine learning, vol. 20, no. 3, pp. 273-297, 1995. https://doi.org/10.1007/bf00994018
[29] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001. https://doi.org/10.1023/a:1010933404324
[30] Y. F. R. Schapire, "Adaptive game playing using multiplicative weights," *Games and Economic Behavior*, vol. 29, no. 1-2, pp. 79-103, 1999. https://doi.org/10.1006/game.1999.0738
[31] J. Friedman, "Greedy boosting approximation: A gradient boosting machine," *Annals of statistics*, vol. 29, pp. 1189-1232, 2001. https://doi.org/10.1214/aos/1013203451
[32] T. Chen, C. Guestrin, "Xgboost: A scalable tree boosting system," *in Procc. SIG KDD*, San Francisco, CA, USA, New York: ACM, pp. 785-794, 2016. https://doi.org/10.1145/2939672.2939785
[33] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241-259, 1992. https://doi.or

g/10.1016/s0893-6080(05)80023-1

[34] M. Kuhn, K. Johnson, "*Applied Predictive Modeling*," New York, NY, USA, Springer, 2013. https://doi.org/10.1007/978-1-4614-6849-3

[35] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321-357, 2002. https://doi.org/10.1613/jair.953

[36] I. Tomek, "Two modifications of CNN," *IEEE Transactions on Systems, Man, and Cybernetics*, pp. 769-772, 1976. https://doi.org/10.1109/tsmc.1976.4309452