Journal of Digital Convergence
Vol. 20. No. 2, pp. 241-250, 2022

ISSN 2713-6434 / eISSN 2713-6442
https://doi.org/10.14400/JDC.2022.20.2.241

균형 랜덤 포레스트를 이용한 이륜차 보험사기 적발 모형 개발

김승훈[1], 이수일[2], 김태호[3*]
[1]국토연구원 부연구위원, [2](주)쿠팡 교통안전본부 본부장, [3](주) 쿠팡 교통안전기획팀 팀장

Bike Insurance Fraud Detection Model Using Balanced Randomforest Algorithm

Kim, Seunghoon[1], Lee, Soo Il[2], Kim, Tae ho[3*]
[1]Associate Research Fellow, Korea Research Institute for Human Settlements
[2]Director, Division of Transportation Safety, Coupang Corportation
[3]Senior Manager, Division of Transportation Safety Planning, Coupang Corportation

요 약 COVID-19 여파로 인한 비대면 서비스와 가정 재정 불안정성의 증가로 이륜차 보험사기 발생이 예상되고 있다. 이와 함께 보험사기 수법도 갈수록 교묘해지고 있다. 하지만 비대면 배달 수요와 연관된 이륜차 교통사고와 보험사기 적발 모형 관련 연구는 매우 미흡한 실정이다. 이에 본 연구는 보험사기의 표본 편중문제를 해결하기 위해 균형 랜덤포레스트 알고리즘을 이용하고 보험사기 조사 전문가의 정성적인 판단 기준을 반영한 변수를 모델에 포함하여 적용성을 향상시키며 적발력 높은 이륜차 보험사기 모형을 개발하고자 한다. 보험사기 적발 모형 개발 결과, 기존의 비균형 랜덤포레스트 모형에 비해 균형 랜덤 포레스트가 보험 사기혐의자를 분류하는 데 있어 통계적으로 우수한 점을 확인할 수 있었다. 특히, 총 26개의 변수를 토대로 탐색적 변수 조합을 적용한 모형의 예측 성능이 가장 높았지만 일부 변수만을 사용한 확인적 모형의 예측 성능도 크게 떨어지지 않은 와중에, 정성적인 보험사기 전문가가 선정한 변수만을 사용한 확인적 모형은 예측력이 떨어지는 것을 확인하였다. 또한, 총 26개의 변수 중 운전자 성별, 연령, 운전자 피보험자 일치 여부, 미수선 청구금액, 대인보험금 등이 중요한 변수로 확인되어 이를 활용해 이륜차 보험사기 혐의자 선별을 위한 적극적인 대처가 필요해 보인다.

주제어 : 플랫폼 배달 서비스, 이륜차 보험사기 랜덤포레스트 알고리즘, 데이터불균형

Abstract Due to the COVID-19 pandemic, with increased 'untact' services and with unstable household economy, the bike insurance fraud is expected to surge. Moreover, the fraud methodology gets complicated. However, the fraud detection model for bike insurance is absent. we deal with the issue of skewed class distribution and reflect the criterion of fraud detection expert. We utilize a balanced random-forest algorithm to develop an efficient bike insurance fraud detection model. As a result, while the predictive performance of balanced random-forest model is superior than it of non-balanced model. There is no significant difference between the variables used by the experts and the confirmatory models. The important variables to detect frauds are turned out to be age and gender of driver, correspondence between insured and driver, the amount of self-repairing claim, and the amount of bodily injury liability.

Key Words : Platform Delivery Service, Bike insurance fraud detection Balanced random-forest algorithm, Imbalanced data

*Corresponding Author : Tae ho, Kim(takim458@coupang.com)
Received December 6, 2021 Revised January 9, 2022
Accepted February 20, 2022 Published February 28, 2022

1. 서론

1.1 연구의 배경 및 목적

COVID-19 여파로 플랫폼 기반의 배달시장이 급격히 성장하면서 이륜차 통행량이 증가하고 있다. 이러한 통행량 급증은 이륜차 관련 교통사고와 사망자 또한 증가시키고 있다. 2020년 6월 발표된 국토교통부·경찰청 보도 자료를 살펴보면, 2020년 1~4월까지 이륜차 교통사고 사망자는 148명으로 2019년 같은 기간 131명에 비해 13.0% 증가하였다. 이는 해당 기간 보행자(-13.6%), 고령자(-18.1%), 화물차(-19.0%)의 교통사고 사망자 수가 각각 감소한 것과는 대조적이다. 또한 교통사고 건수도 6,055건으로 지난해 같은 기간 5,715건에 비해 5.9% 수준으로 증가하고 있다. 이는 비교적 낮은 구매비용과 운전면허 취득의 용이성 덕분에 이륜차를 이용하는 배달산업으로 진입하는 장벽이 높지 않아, 새로운 이륜차 배달원들이 증가하고 있고. 이에 미숙한 주행, 배송 시간을 맞추기 위한 위험한 운행으로 인해 사고율이 높아지고 있는 것으로 분석된다.

이륜차 통행량 및 사고 수의 증가와 함께 코로나19로 인한 경기 침체가 계속되면서 이륜차 보험사기 또한 증가하고 있다. 국내 A보험사에 따르면 이륜차 보험사기 적발 규모는 2020년 836억원으로 전년 793억원 대비 5.4% 증가한 것으로 발표하였다. 또한, 보험사기 기법이 복잡 교묘해지고 있어 손해보험업계 차원의 적극적인 대응이 필요하다[1]. 최근 교묘해지고 있는 이륜차 보험사기의 주요 특징은 첫째, 플랫폼 정비업체 종사자와의 공모 강화로 수리비용을 과잉 청구하는 수법의 전문성이 고도화되었다. 둘째, SNS 플랫폼을 기반으로 공모자를 모집하여 사기수법을 불특정 다수에게 전파시킨다. 셋째, 10대 이륜차 사기를 시작으로 20~30대 공유차, 렌트차, 중고 외산차로 보험사기 규모를 확대해가는 것으로 요약된다. 이와 같은 보험사기 특성은 기존의 승용차 기반의 보험사기 특성과 차별성을 가진다. 그러므로 이륜차 보험사기를 적발하기 위해서는 이륜차 보험사기 만의 인적 특성, 물차 청구 특성(대인 관련 보험금이 없는 경우의 대물과 자차 관련 보험금 청구 특성), 사기 확산 형태 등을 반영한 적발 모형 개발이 시급하다. 본 연구는 기존의 보험사기 연구를 검토하고, 이륜차 보험사기 특성을 고려한 물차 특성 변수를 이용하여 이륜차 보험사기 혐의자를 선별할 수 있는 모형을 개발하고자 한다.

1.2 연구의 방법 및 구성

본 연구에서는 기존연구 고찰을 통해 연구의 착안점을 도출하고, 실제 자동차 보험사고 데이터를 대상으로 머신러닝 알고리즘(랜덤포레스트)을 이용해 보험사기 적발모형을 개발한다. 이를 위해 첫째, 차량 보험사기 선행연구를 변수와 방법론 측면에서 고찰한다. 둘째, 이륜차 사기 수법을 선별할 수 있는 변수 측면의 연구 착안점을 검토 제시한다. 셋째, 머신러닝 알고리즘을 이용해 방법과 변수 조합별 보험사기 적발모형을 개발한다. 넷째, 개발된 모형의 성능 비교·평가를 통해 최적의 모형을 선정한다. 다섯째, 보험사기 적발 관련 시사점을 제시한다.

2. 이론적 배경

2.1 보험사기 개념

보험사기는 “보험사고의 발생, 원인, 내용에 관하여 보험자를 기망하여 보험금을 청구하는 행위”로 규정된다(보험사기 방지법, 법률 제 14123호, 2016. 3. 29. 제정). 보험사기는 보험회사에 큰 재정적 손실과 신뢰성 저하를 야기할 수 있다[2]. 또한, 보험사기는 고의적인 사고를 유발하고, 기존에 발생한 손해를 과장·확대해서 선의의 보험 계약자의 부담을 증가시키고 보험 재원의 누수를 초래할 수 있다[3].

2.2 국내외 선행연구 검토

보험사기에 대한 연구는 1980년대에 국외에서 시작되었다[3]. 초기에는 주로 보험사기에 대한 특성과 대응(Fraud Prevention System, FPS)이 주요 대상이었지만, 최근에는 보험사기 적발을 목적으로 한 시스템(Fraud Detection System, FDS) 개발이 주요 분야로 정착되었다. 또한 FDS는 초기의 보험사기 조사원의 감시 및 적발 방법 연구에서 데이터 마이닝 기반 통계분석 연구로 전환되고 있다[2]. 보험사기 데이터는 획득하기가 어렵고 보험사기 여부가 이진 변수(0=사기 無, 1=사기 有)화 되어 있어 사용할 수 있는 통계 모형 및 머신 러닝 알고리즘에 한계가 있다. 이에 기존 연구방법으로는 Logistic Models(LM), Decision Trees(DT), Artificial neural networks(ANN), Fuzzy-Neural Network(FNN) 등이 사용되었다[2]. 예를 들면 Wen, Wang & Lan (2005)[6]는 로지스틱 모형들을 사용하여 자동차 보험의 종류와 특성, 청구 내용 등과 보험사기 여부(이진 변수)

의 상관성을 산정하였다. Artís, Ayuso, & Guillén (2002)[7]는 오분류율을 반영한 보험사기 적발 Logistic Models(LM)을 개발하였다. 한편, 국내에서는 조해균(1990, 2001)[8-9]이 보험사기의 발생원인/방안 등을 정성적으로 제시한 것이 본격적인 연구의 시초였고, 김철영(1996), 김영중(1998), Sung (2003), [10-12]의 사례연구와 김광용(1996), 김헌수(1999, 2000)[13-15]등이 보험사기 적발 모형개발을 위한 실증 연구를 수행하였다. 김정동·박종수(2006)[16]는 LM을 사용해 보험사기 모형을 개발하였다. 또한 이명진·김광용(2007)[3]이 설문 조사를 통하여 개인의 태도, 주관적 규범 지각된 행동통제 등이 보험사기 행동의도에 미치는 영향을 정성적으로 분석하였다. 최근에는 김태호·임종인(2020)[17]이 Social Network Analysis 기법을 이용한 보험사기 적발 모형의 적용 가능성을 제시하였다.

2.3 방법론 검토

2.3.1 불균형 자료 분석

Abdallah et al. (2016)[2]은 차량 보험사기 적발 연구를 고찰하면서 다양한 문제점을 제시하였다. 그 중 하나가 데이터의 불균형 문제이다. 불균형 데이터란 종속변수 계급 중 한 개 이상의 계급이 차지하는 비중이 다른 계급에 비해 월등히 높은 자료를 의미하며, 대부분의 보험사기 데이터는 여기에 해당된다[20]. 왜냐하면 사기가 적발된 케이스가 그렇지 않은 케이스에 비해 월등히 적기 때문이다. 보통의 분류 머신러닝 알고리즘들은 이러한 불균형 데이터를 분석하는 데에 있어 비효율적이다. 그 이유는 전통적인 머신러닝 알고리즘들은 보통 전체 오류율을 최소화시키는 데에 그 목적이 있기 때문이다. 곧 머신러닝 알고리즘들은 데이터 내의 주류를 이루고 있는 케이스(다수 계층, majority)를 정확도 있게 예측·분류하는 데에만 집중하게 된다[20]. 그러므로 주 목적이 비주류 케이스(소수 계층, minority)를 선별하는 보험사기 연구의 경우 불균형적인 데이터의 분포는 효율적인 모형개발에 제약조건으로 작용한다[2, 21-22]. 따라서 다수 연구들이 이러한 문제를 해결하기 위해 두 가지 해결 방법을 제안하였다. 첫째, 비용 민감성 분석(cost-sensitive learning)으로. 이 방법은 소수 계층의 오분류(misclassification) 비용을 다수 계층의 오분류 비용보다 높게 설정하는 것이다. 이는 알고리즘 내의 함수와 파라미터의 변형이 필요하며 적용 가능한 알고리즘을 선택해야 한다. 둘째, 복합표본 추출(hybrid-sampling)을 이용하여 다수와 소수 계층의 표본 수를 조정하는 것이다. 보통 저표본 추출은 분석 시간이 단축되고 과설명(overfitting)의 문제가 발생하지 않으나, 다수 계층의 표본 추출과정에서 정보가 손실될 가능성이 있다. 반면 과표본 추출은 분석 시간이 길고 소수 표본의 중첩으로 과설명의 문제가 발생하는 것으로 알려져 있다[23]. 그리하여 이와 같은 문제점을 해결하기 위해 SMOTE (Synthetic Majority Oversampling Technique)과 같은 복합표본 추출법이 개발되고 있다[20, 23]. 이러한 방법론적 문제점을 해결하기 위해 본 연구에서는 랜덤 포레스트와 복합표본 추출법을 결합한 균형 랜덤 포레스트를 활용하였다[24, 25].

2.3.2 균형 랜덤 포레스트

Breiman에 의해 개발된 랜덤 포레스트는 앙상블 기법 중 하나로 bootstrapping과 bagging을 결합하여 다수의 의사결정 나무 모형(CART)을 개발한 후 다수결에 따라 관측치의 계급을 예측하는 방법이다. 국내에서 주로 사용되는 로지스틱 모형과 같은 parametric 모형은 독립변수와 종속변수 간의 선형관계를 전제로 하지만 랜덤 포레스트에 쓰인 의사결정나무 기반의 모형들은 그러한 가정에서 자유롭다. 한편 단일 나무 기반의 의사결정나무 모형은 데이터를 과설명하여 설명력이 높지만 예측력이 높지 않다. 랜덤 포레스트는 이러한 단점을 보완하기 위하여 데이터를 bootstrapping하여 매 의사결정나무 모형마다 training set과 test set을 구분한 후 각 의사 결정나무에 각기 다른 training 데이터를 사용한다. 또한 각 의사결정나무에 이용되는 변수는 무작위 선택되므로 의사결정나무 모형의 다양성이 증가한다. 마지막으로 개발된 모든 의사결정나무모형의 예측값들을 다수결로 결합하여 좀 더 예측력이 높은 모형을 만들 수 있다[24].

랜덤 포레스트는 모형의 분석 구조상 불균형 상태의 데이터 분석에 적합하게 변형될 수 있다. 그 중 한 알고리즘인 균형 랜덤 포레스트는 가중치 랜덤 포레스트와 더불어 불균형 데이터 분석이나 비용 민감성 분석에 쓰일 수 있다. 본 연구에서는 균형 랜덤포레스트를 사용하였다. 균형 랜덤포레스트는 알고리즘내의 표본 추출과정에서 다수계층의 표본 수와 소수계층의 표본 수를 일치시키는 방법을 이용한다[24]. 모형에 대한 자세한 설명은 Chen & Breiman(2004)[24]의 논문을 참조하면 된다.

2.3.3 연구의 착안점

선행 연구와 연구방법론 고찰 결과를 토대로 다음과 같은 연구의 착안점을 제안한다.

첫째, COVID-19로 인한 이륜차 보험사기 급증에 대응할 수 있는 국내 보험사기 적발모형 연구는 매우 미흡하다. 둘째, Abdallah et al.(2016)[2]이 주장한 데이터 불균형 문제를 해결하기 위해 균형 랜덤 포레스트 모형을 사용한다. 선행 연구는 대부분 표본 추출 기반의 해결책을 제시하였으나, 저표본 추출의 경우에는 데이터 손실의 문제점이, 과표본 추출의 경우에는 과설명의 문제점과 분석시간이 오래 걸리는 단점이 있다. 이와 같은 한계점들을 극복하기 위해 Hofmann은 SMOTE 기법을 사용하였다. 셋째, 보험 데이터는 방대한 변수 조합을 수반한다. “The Curse of Dimensionality”로 알려진 이 문제는 많은 수의 변수를 이용한 모형이 보다 적은 수의 변수를 이용한 모형보다 성능이 반드시 뛰어나지는 않다는 것이다. 이 문제를 해결하기 위해 일부 선행 연구는 요인분석(Principal Component Analysis, PCA)와 같은 차원 감소법(Dimensionality Reduction)을 이용하였다[26-28].

본 연구에서는 실제 적용 가능한 변수 추출을 위해 보험사기 조사팀의 정성적 의견을 반영하였다. 이는 통계적인 분석 방법 대신 실제 사기적발에 적용 가능한 주요 변수 도출에 전문가 지식과 경험을 적용하는 사례라 할 수 있다.

3. 변수추출 및 분석결과

3.1 모형의 개발

3.1.1 변수 추출

본 연구에 사용된 자료는 2017-2020년 A보험사 내부 데이터에서 사후 보험금 지급 완료 후 이륜차 사기 관련 케이스를 추출하였다. 자동차 보험사기는 두 가지로 구분할 수 있다. 경성 사기(Hard Fraud)는 개인이 사고를 고의 혹은 의도적으로 유발한 경우이다. 연성 사기(Soft Fraud)는 실제 사고가 부주의 및 비고의로 발생한 경우 보험료 및 내용 청구를 조작하는 경우이다. 연성사기는 실제 수리비보다 과다 청구하거나, 손상되지 않은 부품을 교체하는 경우까지 포함한다[2,4-5]. 단 최근에는 경성과 연성사기의 경계가 모호한 사고 또한 증가하고 있는 추세이다. 본 연구에서는 경성과 연성을 구분하지 않고 사기(Fraud)와 비사기(Non-Fraud)로 구분하였다.

Table 1. List of variables

Variable		Scale	Definition
Fraud		nominal	If an accident is bike insurance fraud or not (No=0, Yes=1)
Insured Age		ratio	The age of the Insured in 2020
Driver-Insured		nominal	If a driver is the insured (No=0, Yes=1)
Night Accident		nominal	The time that an accident occurred (before midnight = 0, midnight~4am = 1, 4 am ~ 6 am = 2)
Redemption		ratio	The amount of redemption
Delayed Time		nominal	The time difference between reporting time and accident time (0~2 hours=0, 2~24 hours=1, 24~72 hours=2)
Accident Type		nominal	Type of accident (1=car to person, 2=car to car 3=car, 4=car to others(bike, bicycle))
Foreign Brand		nominal	If an insured car is foreign branded (No=0, Yes=1)
Gross Fault		nominal	If an accident is caused by any gross fault (No=0, Yes=1)
Type of Gross Fault		nominal	Type of gross fault (1=getaway, 2=unlicensed, 3=traffic signal, 4=speed, 5=impaired, 6=center-line, 7=crosswalk, 8=overtake)
Police report		nominal	If an accident is reported to police No=0, Yes=1
Police On-site		nominal	If a police officers go out on the site (No=0, Yes=1)
Driver Gender		nominal	Gender of a driver(insured car) (0=female, 1=male)
Driver Age		ratio	Age of a driver(insured car)
Driver Age_20		nominal	If a driver's age is 20s (No=0, Yes=1)
Insurance Duration		ordinal	The duration of an insurance 1=less than 1 year , 2=2~3 years, 3=3~4 years, 4=4~5 years, 5=5~6 years, 6=6~7years, 7=more than 7 years
Vehicle Year		ordinal	Vehicle year in 2020
Rental		nominal	If an insured car is rental (No=0, Yes=1)
Corporate Vehicle_insured		nominal	If an insured car is a corporated car (No=0, Yes=1)
# of Passengers	insurded	ratio	A number of passengers in an insured car
	opponent	ratio	A number of passengers in an opponented car
Corporate Vehicle_opponent		nominal	If an opponent car is foreign branded (No=0, Yes=1)
Self-repair	insurded	nominal	If an insured car is fixed by self-repairment (No=0, Yes=1)
	opponent	nominal	If an opponent car is fixed by self-repairment (No=0, Yes=1)
	payout	ratio	The amount of payouts of self-repairment
Drunken		nominal	If an accident is caused by a drunken driver No=0, Yes=1
Payouts_Opponent passengers		ratio	The mount of payouts for opponents

3.1.2 분석 방법

본 연구의 최종 목적은 보험사기 가능성이 높은 사고를 효과적으로 선별할 수 있는 예측 모형을 개발하는 것이다. 보험사기 예측 모형을 개발하기 위해서 의미가 있는 변수를 선정하는 과정이 필요하다. 이는 의미가 없는 변수를 랜덤포레스트 알고리즘에 투입할 경우에 예측력을 저하시킬 수 있기 때문이다. 변수 선정을 위해 첫째, 전체 데이터와 모든 변수를 이용하여 랜덤포레스트 모형을 개발한 후, 중요 변수를 선정한다. 둘째, A 보험사의 보험사기 조사 팀을 대상으로 이륜차 보험사기 적발에 필요한 변수에 대해 설문조사를 실시한다. 다양한 변수 구성을 이용하여 랜덤포레스트와 균형 랜덤포레스트 모형을 개발한 후, 모형의 예측 성능을 비교하여 최적의 모형을 도출한다(Fig 1 참조).

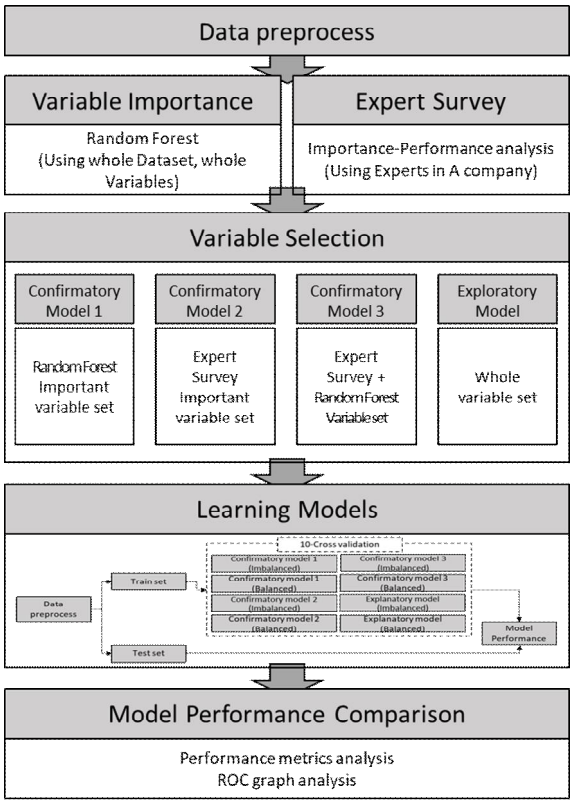

Fig. 1. Analytical framework

3.1.3 분석 과정

1) 변수 중요도 분석

모든 변수를 포함한 랜덤포레스트 모형 개발을 통해 영향력이 높은 변수를 선정하였다. 최적의 랜덤 포레스트 모형개발을 위해서 R 프로그램을 활용하였으며, 모형의 분석 결과는 다음과 같다.

```
Call:
randomForest(x = x, y = y, ntree = 500, mtry = param$mtry)
      Type of random forest: classification
      Number of trees: 500
      No. of variables tried at each split: 13
      OOB estimate of error rate: 1.12%
        Confusion matrix:
          No Yes  class. error
No   37272  24 0.0006435006
Yes    398   6 0.9851485149
```

10-fold Cross-Validation 결과, 500개의 의사 결정 나무에 12개의 변수가 최적 모형으로 나타났다. 이 모형은 약 99%의 예측력을 보이지만 대부분 사기가 아닌 사고를 성공적으로 분류한다(37284+7 / [37284+12+397+7] = 98.9%). 반면 사기 사건은 총 384 건 중 7건(1.8%) 만을 성공적으로 분류하였다. 이는 종속변수의 계급(class) 구성이 균등하지 않기 때문이다. 다시 말하면, 종속변수 내 사기자가 아닌 표본이 편중되어 있기 때문에 (총 37,700 중 37,296건) 기계 학습모형은 사기자가 아닌 계급을 예측하는 것에 더 중점을 두게 된다. 이와 같은 단점을 보완하기 위해 본 연구는 균형 랜덤포레스트를 활용하여 소수 케이스인 사기 적발 사고를 예측하는 데에 중점을 두었다. 한편, 영향력이 높은 변수를 선정하기 위하여 Mean Decrease Gini 계수에 기반을 둔 변수의 중요도를 분석하였다. 상위 10개 중요 변수는 미수선 청구 금액, 피보험자 나이, 운전자 연령, 대인 보험금, 사고 형태, 차량 연식, 접수 지연시간, 보험 가입 경력, 운전자 성별, 현장 출동 여부 순으로 확인되었다.

2) 보험사기 전문가 설문조사 분석

A 보험사의 보험사기 적발 전문가를 대상으로 보험사기 적발에 실제 활용 가능한 변수에 대해 설문조사를 실시하였다. 중요도-만족도 설문 조사 결과를 토대로 Portfolio 분석을 진행한 결과, 만족도와 중요도가 비교적 높게 나온 변수들은 접수 지연시간, 대차 미수선 청구, 심야사고, 차량 연식, 20대 여부, 현장 출동 여부, 외산차 여부, 경찰 신고, 보험 가입경력 등이다. 데이터 분석을 통해 선정된 중요변수와 실무에서 고려되는 변수를 비교해보면, 접수 지연시간, 보험가입경력, 연령(운전자, 보험 가입자), 현장 출동여부, 차량 연식 등 많은 변수들이 공통으로 나타났고 운전자 성별이나 운전자 일치 여부 등은 데이터 분석에서 중요하게 나타난 반면, 전문가 판단에서는 제외해도 되는 것으로 나타났다.

3) 분석의 변수 구성

전문가 설문조사 및 데이터 분석 결과를 비교•대조해 본 결과, 본 연구에서는 다양한 변수 조합을 구성하여 효율적인 이륜차 보험사기 적발 모형을 개발하고자 한다.(Table 2)

Table 2. Variable usage for each models

Variable	Model			
	Conf 1*	Conf 2	Conf 3	Exp**
Delayed Time	o	o	o	o
Police On-site	o	o	o	o
Insurance Duration	o	o	o	o
Vehicle Year	o	o	o	o
Insured Age	o	x	o	o
Driver-Insured	o	x	o	o
Driver Gender	o	x	o	o
Driver Age	o	x	o	o
Self-repair payout	o	x	o	o
Payouts_Opponent passengers	o	x	o	o
Night Accident	x	o	o	o
Corporate Vehicle_insured	x	o	o	o
Police report	x	o	o	o
Driver Age_20	x	o	o	o
Self-repair opponent	x	o	o	o
Redemption	x	x	x	o
Accident Type	x	x	x	o
Gross Fault	x	x	x	o
Type of Gross Fault	x	x	x	o
Rental	x	x	x	o
Corporate Vehicle_insured	x	x	x	o
# of Passengers insured	x	x	x	o
# of Passengers opponent	x	x	x	o
Corporate Vehicle_opponent	x	x	x	o
Self-repair insured	x	x	x	o
Drunken	x	x	x	o

* Conf: Confirmatory model
** Exp: Exploratory model

4) 데이터의 분할

기계학습 알고리즘의 과적합을 방지하기 위하여 원본 데이터로부터 훈련데이터와 검증데이터를 비복원 랜덤 추출한다. 추출 비율은 많은 기존 연구들이 채택해온 훈련데이터 70%, 검증데이터 30%이며, 기계학습 알고리즘을 훈련데이터에 적합 시킨 후 검증데이터에 적용하여 모형을 평가한다. 훈련데이터와 검증데이터의 종속변수 구성비는 Figure 2에 나타내었다. 각 데이터 셋에서의 종속변수의 계층 분포는 동일하며 비사기 계층에 편중되어 있다는 것을 알 수 있다.

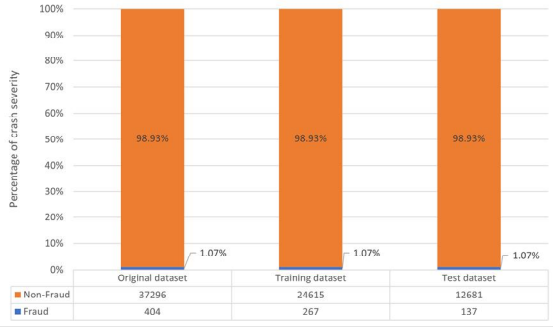

Fig. 2. Distribution of the dependent variable

3.2 분석 결과

3.2.1 학습 결과

본 연구는 R의 Caret패키지를 이용하여 8개의 랜덤 포레스트 모형을 개발하였다. 10-fold cross-validation을 통하여 500개의 의사결정나무를 이용하여 각 모형을 훈련데이터를 이용하여 학습하였다(Table 3). 학습된 모형은 검증데이터를 이용하여 모형의 성능을 분석하였다. 분석 결과, 불균형 모형의 오차율이 균형 모형의 오차율보다 적은 가운데 특이하게 확인적 모형 2 균형 모형의 오차율이 다른 모형들보다 크게 나타났다. 이는 설문조사 중요 변수들이 이륜차 보험사기를 적발하는데 효율적으로 쓰이지 않았다는 것을 시사한다.

3.2.2 모형의 성능 비교 분석

모형의 예측력 및 성능을 평가·비교하기 위하여, 본 연구는 오분류표(confusion matrix)에 기반을 둔 performance metrics와 Receiver Operation Characteristic (ROC) 그래프를 사용한다.

1) 오분류표 기반의 성능지표 분석

모형들의 오분류표와 성능지표 비교 분석결과, 첫째, 불균형 모형의 Accuracy가 균형모형보다 전반적으로 우수한 것으로 나타났다. 불균형 모형들의 Accuracy 범위는 98.92% ~ 98.93%으로 99%에 가까운 수치를 보였다. 반면 균형 모형의 Accuracy 범위는 68.88% ~ 83.65%으로 비교적 낮다(Table 3 참조). 하지만 불균형

Table 3. Performance metrics

변수	Exploratory model (Exp)		Confirmatory model 1 (Conf 1)		Confirmatory model 2 (Conf 2)		Confirmatory model 3 (Conf 3)	
	Imbalanced	Balanced	Imbalanced	Balanced	Imbalanced	Balanced	Imbalanced	Balanced
Accuracy	98.92%	83.65%	98.94%	82.74%	98.93%	68.88%	98.93%	83.45%
Precision	33.33%	3.69%	100.00%	3.29%	-	1.78%	-	3.86%
TPR	0.73%	56.93%	0.73%	53.28%	0.00%	51.82%	0.00%	60.58%
TNR	98.94%	99.45%	98.94%	83.06%	100.00%	69.06%	100.00%	83.70%
FPR	0.02%	16.06%	0.00%	16.94%	0.00%	30.94%	0.00%	16.30%
F1	1.43%	6.93%	1.45%	6.19%	-	3.44%	-	7.26%

* TPR: True Positive rate
** TNR: True Negative rate
*** FPR: False Positive rate
**** F1: F-score ($TP/(TP+(FN+FP)/2)$)

모형의 Accuracy가 높은 이유는 대부분의 보험사고를 non-Fraud로 구분하기 때문이다(Table 4 참조). 이는 다른 예측 성능 지수들과 함께 더 자세히 설명하겠다.

둘째, precision의 경우에도, 불균형 모형이 각각 33.00% 혹은 100%으로 균형모형보다 우위를 보인다. precision은 모형에 의해 예측된 보험사기 사고 중 실제 보험사기 사고인 경우의 비율로서 모형이 어떤 사고를 보험사기로 인지한 경우, 그 예측값을 얼마나 신뢰할 수 있는지를 나타낸 지표이다. 단, 확인적 불균형 모형 2와 3의 경우 TP와 FP 케이스가 0 이기 때문에(보험사기로 인지한 사고가 없기 때문에) precision이 계산될 수가 없었다.

Table 4. Confusion matrix

Predicted \ Observed			Fraud	Non-Fraud
Exp	IMB*	Fraud	1	2
		Non-Fraud	136	12679
	B**	Fraud	78	2037
		Non-Fraud	59	10644
Conf 1	IMB	Fraud	1	0
		Non-Fraud	136	12681
	B	Fraud	73	2148
		Non-Fraud	64	10533
Conf 2	IMB	Fraud	0	0
		Non-Fraud	137	12681
	B	Fraud	71	3923
		Non-Fraud	66	8758
Conf 3	IMB	Fraud	0	0
		Non-Fraud	137	12681
	B	Fraud	83	2067
		Non-Fraud	54	10614

* IMB: a model using original random-forest algorithm
** B: a model using balanced random-forest algorithm

셋째, 균형 모형들의 TPR이 불균형 모형들보다 현저히 높은 것으로 나타났다. TPR은 실제 보험사기 중 보험사기로 적발된 비율로써 모형의 얼마나 많은 보험사기사고를 찾아낼 수 있는지를 측정하는 중요한 지표 중 하나이다. 하지만, 불균형 데이터를 분석함에 있어 Accuracy와 Precision, TPR와 같은 성능 지표들의 올바른 해석을 위해서는 반드시 오분류표(Table 4)를 참고해야 한다.

불균형 모형들의 높은 Accuracy는 대부분의 보험사고가 Non-Fraud로 예측되었기 때문이다. 이는 오분류표에서 불균형 모형들의 예측된 Fraud 케이스 수가 0~2에 불과하다는 것을 보면 알 수 있다. 같은 맥락에서 불균형 모형의 높은 precision이 예측된 Fraud가 높은 확률로 실제 Fraud라 하더라도, 고작 3건 중 1건(탐색적 불균형 모형) 혹은 1건 중 1건(확인적 불균형 모형1)에 불과하다. 반면, 균형 모형들은 비록 Fraud로 분류해 낸 사고들 중 상당수가 Non-Fraud이더라도 (낮은 precision이더라도) 보다 많은(71~83건) 수의 보험사기 사고를 적발해낼 수 있다(높은 TPR). 요약해보면, 균형 모형이 불균형 모형보다 많은 수의 사고를 보험사기로 분류하여 절대적으로 많은 수의 보험사기 사고를 찾아 낼 수 있다.

넷째, FPR은 모형이 보험 사기가 아닌 사고를 보험사기로 예측한 비율로서, 불균형 모형들이 균형 모형보다 낮은 수치로 우월함을 보였다. 이는 피보험자가 보험사기행위를 하지 않았음에도 의심 혹은 조사를 받을 확률이다. 이 성능 지표는 두 가지 측면으로 해석 될 수 있다. 무엇보다, 피보험자 혹은 계약자 개인의 입장에서는 보험 사기행위가 없음에도 불구하고 추가 조사를 받거나 의심을 받아 피해를 입을 수 있다. 이는 보험 회사 이미지에도 손상을 입힐 수 있으며, 만약 이러한 분쟁이 소송으로 이어질 경우, 막대한 재정과 시간이 소모될 수 있다. 또한, 보험 회사 내의 보험사기 조사 업무의 효율성의 측면에서 보면(모형에서 Fraud로 분류한 사고를 조사

한다고 가정), 어떤 모형이 과도하게 많은 수의 사고를 보험사기로 분류하고 이중 상당수가 보험사기가 아닐 경우, 보험사기 조사 자원의 낭비를 가져올 수 있다[27].

2) ROC 그래프

ROC (Receiver Operating Characteristic) 그래프는 분류 모형의 예측 성능 나타내기 위한 지표이다. 만약 어떤 모형의 예측력이 전혀 없다면 대각선을 그리며 예측이 완벽하다면 (어떠한 경우에도 모든 사건을 완벽히 분류해 낸다면) (specificity, sensitivity) (1,0)(1,1)(0,1)을 지나는 그래프를 가지게 된다. 본 연구에서는 모형의 성능을 측정하기 위해서 ROC 그래프 아래의 면적 (area under the ROC curve, AUC)값을 지표로 이용하였다. AUC가 1에 가까운 모형일수록 예측 성능이 뛰어나다.

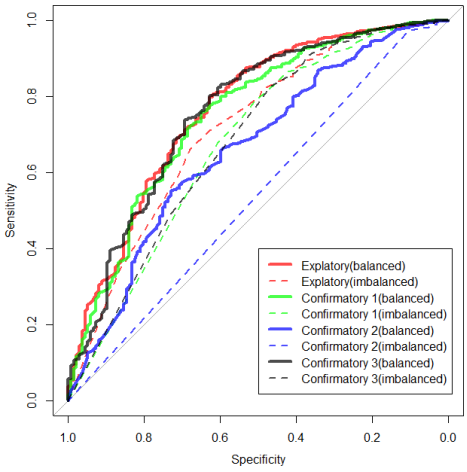

Fig. 3. ROC graph

ROC 그래프 분석 결과, 균형 모형이 불균형 모형보다 예측 성능이 뛰어났으며 확인적 모형과 탐색적 모형의 차이는 거의 없는 것으로 나타났다. 즉 탐색적 균형 모형이 가장 뛰어난 예측력을 보였으며 (AUC = 0.7649) 확인적 균형 모형 3과 1이 큰 차이 없이(0.7600, 0.7506) 그 뒤를 따랐다. 가장 예측력이 떨어지는 모형은 확인적 불균형 모형 2이었다. (0.5380)

4. 결론

본 연구는 COVID-19의 여파와 최근 성장하는 배달 시장 등의 여파로 인한 외산 이륜차의 보험사기 적발 모형을 개발하였다. 또한 기존 연구 고찰을 통해 보험사기 특성을 규명하고 균형 모형의 적용 필요성을 주장하였다. 이에 따라 본 연구에서는 랜덤포레스트 알고리즘을 이용, 확인적, 탐색적 방법론을 적용하여 불균형 모형과 균형 모형을 개발하고 예측 성능을 비교하여 발전된 보험사기 적발 모형 개발에 기여하고자 하였다.

4.1 최적 모형의 선정

모형 개발 및 비교 평가 결과, 모형 전체의 예측 정확도(Accuracy)에 기반을 둔다면 불균형 모형들이 균형 모형들에 비해 뛰어난 것으로 보인다. 하지만 불균형 모형의 높은 예측력은 대부분 비사기(Non-Fraud) 사고를 예측함으로써 얻어지는 것이다. 그러나 보험사기 적발 모형의 목적은 사기 (Fraud) 사고를 찾아내는 것이다. 그러므로 만약 보험 회사가 보험사기 적발에 충분한 자원을 투자할 용의가 있다면 비록 높은 수치의 FPR과 낮은 Accuracy와 Precision 값을 보이더라도 균형 모형을 선택할 것을 권장한다. 그러나 균형 모형을 이용할 시에는 많은 수의 Non-Fraud 보험사고가 Fraud로 분류되어 보험사기 조사 자원의 비효율성이 예상되므로 보험회사의 상황과 목적에 따라 적절한 모형 선정이 필요하다.

또한, 본 연구는 일부 변수만을 이용한 확인적 모형들과 탐색적 모형의 성능 비교분석을 수행하였다. 그 결과, 예측 성능에는 큰 차이가 없는 것으로 나타났다(랜덤 포레스트 모형의 장점). 이는 곧 한정적인 정보만을 바탕으로 사기 적발 모형을 이용하더라도 어느 정도 소기의 목적(보험사기 조사 대상을 선별하는 것)을 달성할 수 있다는 것을 시사한다. 한편, 확인적 모형중 실무팀이 중요하게 고려하는 변수를 이용한 모형은 다른 모형들보다 예측 성능이 확연히 떨어지는 것으로 나타났다.

4.2 중요변수 분석

최적 모형의 선정과정에서 이륜차 보험사기 적발을 위한 효과적인 변수 구성을 조사하였다. 가능한 모든 변수를 사용하는 탐색적 모형의 예측 성능이 일부 변수만 사용하는 확인적 모형에 비해 월등히 뛰어나지 않은 가운데 3가지의 확인적 모형의 변수 구성에 따른 예측 성능은 확연한 차이를 보였다. 즉, 사기조사 실무진이 주로 중요하게 생각하는 변수만을 이용하여 학습시킨 모형의 경우 낮은 예측 성능을 보인다. 확인적 분석 모형 2의 낮은 예측성능이 암시하는 바는 조사원의 고려하는 중요변수들과는 달리 실제 보험사기를 분류하기 위해서는 보다 많은 변수가 필요하다는 것이다. 반면에 데이터 중요변수와 실무팀 중요변수를 모두 사용한 확인적 분석 모형

3의 경우에는 확인적 분석 모형 2 보다 높은 예측 성능을 보여준다. 곧 확인적 분석 2에 포함되지 않았던 변수 피보험자 나이, 운전자 일치여부, 운전자 성별, 운전자 연령, 미수선 청구 금액, 대인보험금 등이 모형의 예측성능 향상에 높은 기여를 하였다. 반면에 확인적 분석 모형 3은 확인적 분석 모형 1에 비해 확연히 향상된 예측성능을 보이지는 않는다. 이는 사기조사 실무팀이 데이터 분석결과와 다르게 중요시했던 변수들 심야사고 여부, 외산차 여부, 경찰신고여부, 20대 여부, 대차 미수선 청구 여부 등이 모형의 보험사고사기 분류에 크게 공헌하지 않았다는 것을 보여준다.

이와 같은 분석 결과는 2000년대에 들어서 보험사기 조사기법이 보험사기 조사원에 의한 보험사기 조사에서 데이터 마이닝을 이용한 보험사기 분류 기법으로 전환되고 있는 이유를 보여준다. 그러나 데이터 마이닝 만을 이용한 보험사기 분류는 그 한계를 보여준다. 균형 모형의 경우, 비록 많은 사고를 비사기로 구분함으로써 높은 예측성능을 보이지만 대부분의 사기 사고를 찾아내지 못했고, 균형 모형의 경우 많은 수의 사기사고를 분류예측할 수 있지만 많은 비사기사고를 사기사고로 구분하여 문제를 야기할 수 있다. 그러므로 본 연구는 균형 모형을 이용할 경우, 보험사기 조사원이 보험사기로 분류된 사고 중 비보험사기사고를 걸러내는 과정을 추가할 것을 제안한다.

4.3 연구의 한계점 향후 연구방향

본 연구의 한계점은 다음과 같다.

첫째, 랜덤포레스트 모형의 특성 상, 모형의 정확한 사기 적발 예측 사유가 제공되지 않는다. 즉, 랜덤포레스트 모형은 여러 의사결정나무를 생성하여 다수의 의사결정나무의 분류값을 따라 최종 분류를 수행하기 때문에 단일의사결정나무를 이용한 모형과 같은 사기 적발 논리를 제공하지 않는다. 그러므로 이와 같은 한계점을 보완하기 위해서는 사기로 예측된 케이스들을 재분석(post-analysis)하는 과정이 요구된다(교차분석, 판별분석 등).

둘째, 본 연구는 과거의 보험사기 행태를 분석해 현재 혹은 미래에 있을 보험사기를 적발하고자 하는 것으로 새로운 형태나 기존의 규칙에 어긋나는 보험사기를 분류해내지 못할 가능성이 크다. 한편 이러한 특이한 사고의 형태를 인지하여 사기를 적발할 수 있는 방법론은 비지도 학습의 클러스터링, 의사나무모형의 등이 있다.

셋째, 비록 본 연구는 한 개의 알고리즘만을 사용하였지만, 차후에는 최적의 모형을 선별하기 위하여 여러 종류의 알고리즘을 여러 조건에서 시행 할 필요가 있다. 또한 neural network 기반의 딥러닝 및 AI를 이용한 분석도 요구 된다.

넷째, 현재 SNS등의 플랫폼을 이용한 보험사기가 횡행하고 있으므로 SNA(Social Network Analysis)를 이용한 보험사기 모형을 결합하여 융합된 모형을 개발한다면 보다 효과적으로 보험사기를 예방 및 적발할 수 있을 것으로 사료된다.

REFERENCES

[1] H. W. Byun, J. Y. Son. (2020). Prevention of Insurance Fraud Utilizing Data Analysis. *KIRI Report (2020.11.23.),* 1-7.

[2] Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications, 68,* 90-113.

[3] M. J. Lee, G. Y. Gim. (2007). An Empirical Study on the Development of Behavior Model of Insurance Fraud. *Journal of Information Technology Services, 6(2),* 1-18.

[4] Roy, R., & George, K. T. (2017). Detecting insurance claims fraud using machine learning techniques. *Proceedings of IEEE International Conference on Circuit, Power and Computing Technologies, ICCPCT 2017.*

[5] Sithic, H. L., & Balasubramanian, T. (2013). Survey of Insurance Fraud Detection Using Data Mining Techniques. *International Journal of Innovative Technology and Exploring Engineering, 2(3),* 62-65.

[6] Wen, C.-H., Wang, M.-J., & Lan, L. W. (2005). Discrete choice modeling for bundled automobile insurance policies. *Journal of the Eastern Asia Society for Transportation Studies, 6.* 1914-1928.

[7] Artís, M., Ayuso, M., & Guillén, M. (2002). Detection of Automobile Insurance Fraud With Discrete Choice Models and Misclassified Claims. *The Journal of Risk and Insurance, 69(3),* 325-340.

[8] H. G. Jo. (1990). The Cause of Insurance Fraud And Countermeasures. *Korean Journal of Insurance, 35,* 75-102

[9] H. G. Jo. (2001). Countermeasures of Insurance Fraud For Nation. *Journal of Insurance Studies, 12(2).*

[10] T. K. Sung. (2003). Detection of Insurance Fraud using Visualization Data Mining Tool. *Information System Review, 5(1),* 49-60.

[11] C. Y. Kim. (1996). Case Study of the Type of Car Insurance Frauds, *General Insurance Association of*

Korea, 328, 43-61.

[12] Y. J. Kim. (1998). Case Study of Car Insurance for Moral Hazard, *General Insurance, 359*, 60-71.

[13] G. Y. Gim. (1996). Developing Early Detecting Insurance Fraud System: Fuzzy Theory and AHP, *Insurance Development Studies, 18,* 4-28.

[14] H. S. Kim. (1999). Brief Study of The Development of Automobile insurance Fraud Early-Warning model, *General Insurance, 363,* 68-80.

[15] H. S. Kim. (2000). A Study on The Development of Automobile insurance Fraud Early-Warning model using Claim Adjusters' Expert knowledge. *The Journal of Risk management, 16*, 59-97.

[16] J. D. Kim, J. S. Park. (2006). A Fraud Detection Model for Automobile Insurance Claims. *Risk Management. 17(1),* 109-152.

[17] T. H. Kim, J. I. Lim. (2020). A Study on Conspired Insurance Fraud Detection Modeling Using Social Network Analysis, *Journal of the Korea Society of Computer and Information, 25(3),* 117-127.

[18] Martino Scheepens. (retrieved on 11.30.2021). Coronavirus, what have you done?. FRISS. https://www.friss.com/blog/coronavirus-what-have-you-done/

[19] Matthew J. Smith. (retrieved on 11.30.2021). Insurance Fraud Report (2020). https://knowledge.friss.com/hubfs/Ebooks/Insurance%20Fraud%20Report%202020-2021%20EN.pdf?utm_campaign=Fraud%20Survey&utm_medium=email&_hsmi=98996085&_hsenc=p2ANqtz-9b05tppFd4OvW5Pgn40Us4ktpp0dXzleaTZb8IQV2-j9muWaPkF6WLs3jg2XUd-udg0gUyFbZtE6ldFqd8yLfN59MVHA&utm_content=98996085&utm_source=hs_automation

[20] Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research, 61*, 863-905.

[21] Brennan, P. (2012). A comprehensive survey of methods for overcoming the class imbalance problem in fraud detection. *Thesis, (June)*, 1-107.

[22] Šubelj, L., Furlan, Š., & Bajec, M. (2011). An expert system for detecting automobile insurance fraud using social network analysis. *Expert Systems with Applications, 38(1)*, 1039-1052.

[23] Fiorentini, N., & Losa, M. (2020). Handling Imbalanced Data in Road Crash Severity Prediction by Machine Learning Algorithms. *Infrastructures, 5(7).*

[24] Chen, C., Liaw, A., & Breiman, L. (2004). Using Random Forest to Learn Imbalanced Data. *In Department of Statistics, UC berkeley.*

[25] Ai, J., Golden, L. L., & Brockett, P. L. (2009). Assessing Consumer Fraud Risk in Insurance Claims. *North American Actuarial Journal, 13(4)*, 438-458.

[26] Brockett, P. L., Derrig, R. a, Golden, L. L., & Alpert, M. (2002). Fraud Classification Using Principal Component Analysis of RIDITs. *The Journal of Risk and Insurance, 69(3),* 341-371.

[27] Viaene, S., Ayuso, M., Guillen, M., Van Gheel, D., & Dedene, G. (2007). Strategies for detecting fraudulent claims in the automobile insurance industry. *European Journal of Operational Research, 176(1)*, 565-583.

[28] Agjee, N. H., Mutanga, O., Peerbhay, K., & Ismail, R. (2018). The impact of simulated spectral noise on random forest and oblique random forest classification performance. *Journal of Spectroscopy. 2018.8*

김 승 훈(Kim, Seunghoon) [정회원]

· 2012년 2월 : 한양대학교 Urban & SOC Planning 석사 취득
· 2020년 12월 : 오하이오주립대학교 도시 및 지역계획 박사 취득
· 2021년 5월 ~ 현재 : 국토연구원 부연구위원
· 관심분야 : 교통계획, 안전
· E-Mail : sh.kim@krihs.re.kr

이 수 일(Lee, Soo Il) [정회원]

· 2006년 2월 : 한양대학교 도시공학과 박사 취득
· 2010년 6월 : 현대해상화재보험 교통기후환경연구소 연구위원
· 2021년 8월 ~ 현재 : ㈜쿠팡 교통안전본부 본부장(상무)
· 관심분야 : 교통안전, 교통공학
· E-Mail : solee94@coupang.com

김 태 호(Kim, Taeho) [정회원]

· 2008년 7월 : 한양대학교 도시대학원 SOC 교통학박사 취득
· 2012년 10월 : 현대해상화재보험 교통기후환경연구소 연구위원
· 2021년 11월 ~ 현재 : ㈜쿠팡 교통안전기획팀 팀장(부장)
· 관심분야 : 교통안전, 데이터 분석
· E-Mail : takim458@coupang.com