교통사고 위험도 예측모형 개발 및 적용

SA167
임지연
이은진
이하경

목차
INDEX

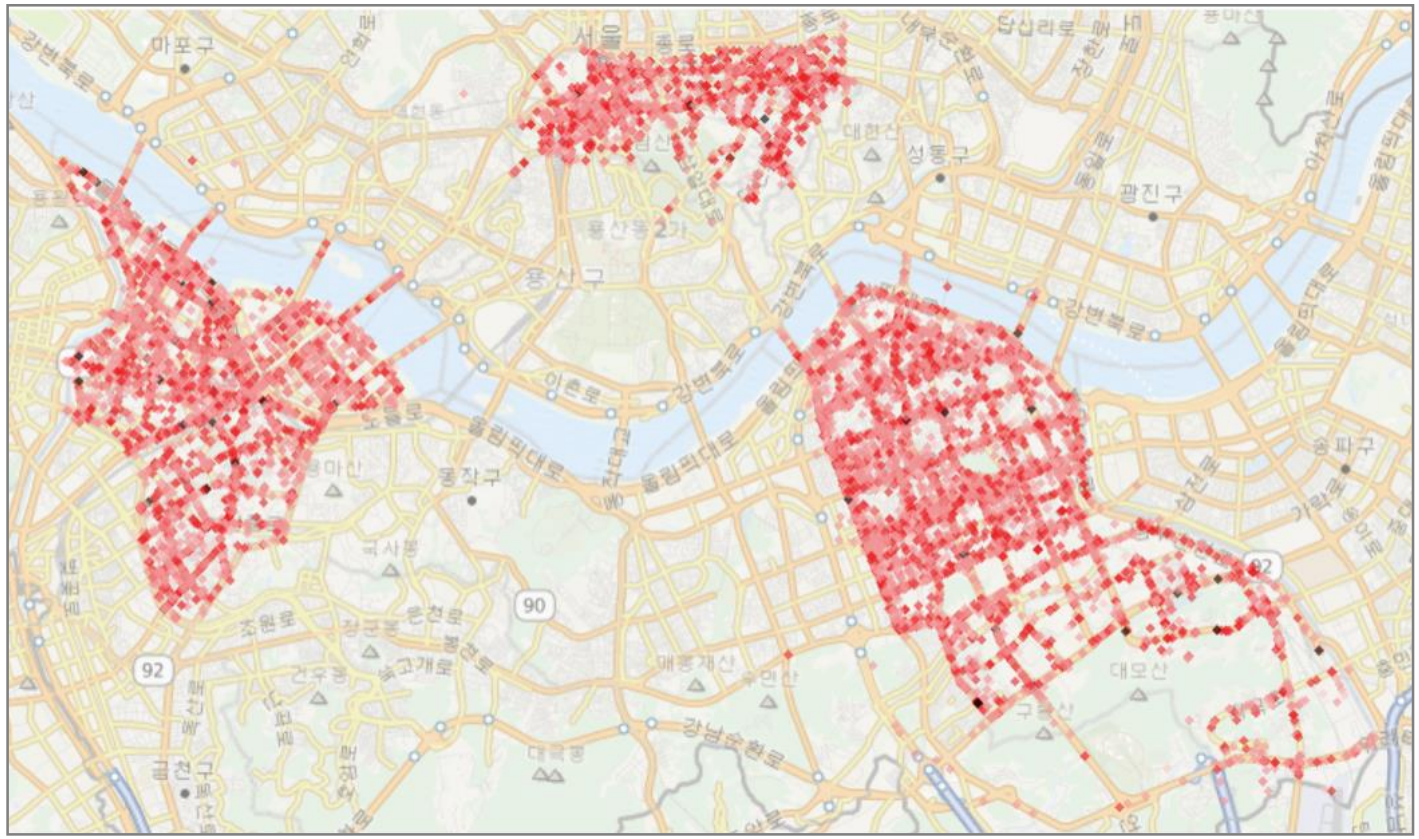
Sample
Explore
Modify
Model
Assess

Ⅰ. 조사 개요

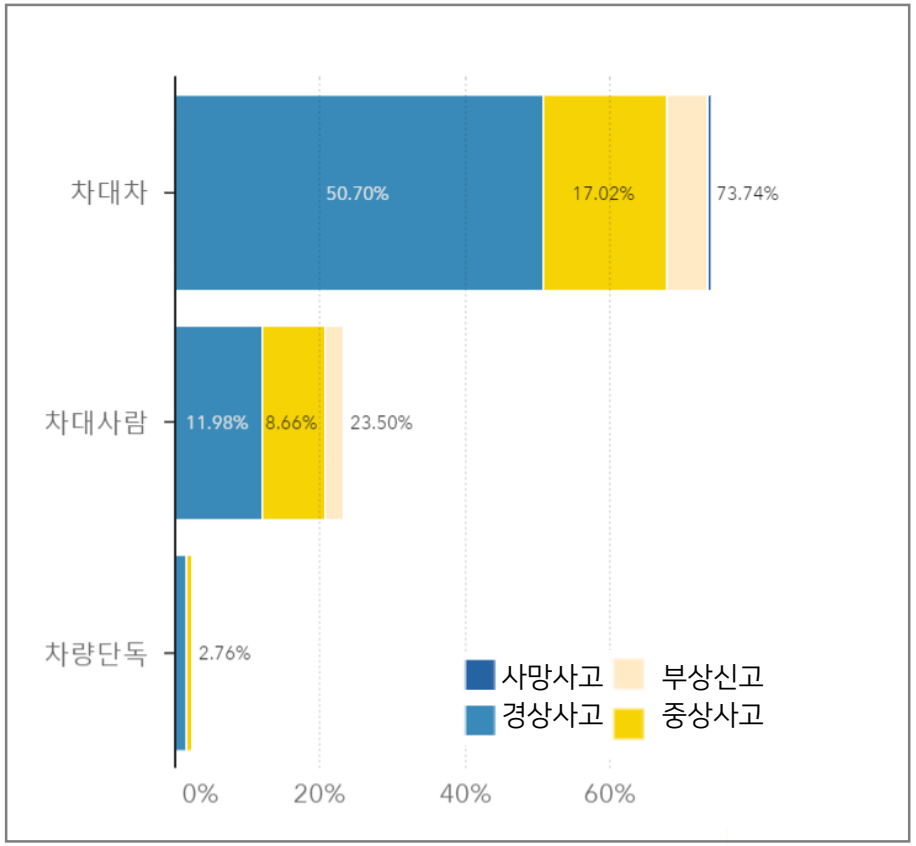
1. 거시적 관점의 시각화

- 상세한 분석을 진행하게 될 3개 구(강남구, 영등포구, 중구)의 사고내용(사망사고, 중상사고, 경상사고, 부상신고사고)에 대하여 거시적인 시각화 후 이상치 제거
- 사고유형(차대차, 차대사람, 차량단독)별 사고내용(사망사고, 중상사고, 경상사고, 부상신고사고)은 차대차 유형이 약 73.74%로 가장 높은 비율을 나타냄
  - '차대차' 유형의 사고 중 경상사고가 전체의 50%로 가장 높은 비중을 차지함
  - '차대사람' 유형의 사고의 경우는 경상사고, 중상사고가 비슷한 비율로 나타났으며 차대사람 사고가 발생할 경우 차대차에 비하여 심각도가 증가

구별(강남구, 영등포구, 중구) 사고현황



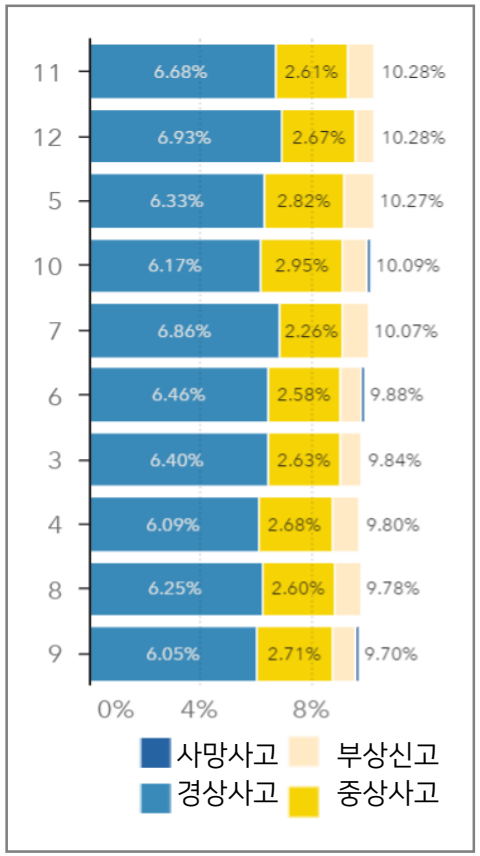
사고유형별 사고현황



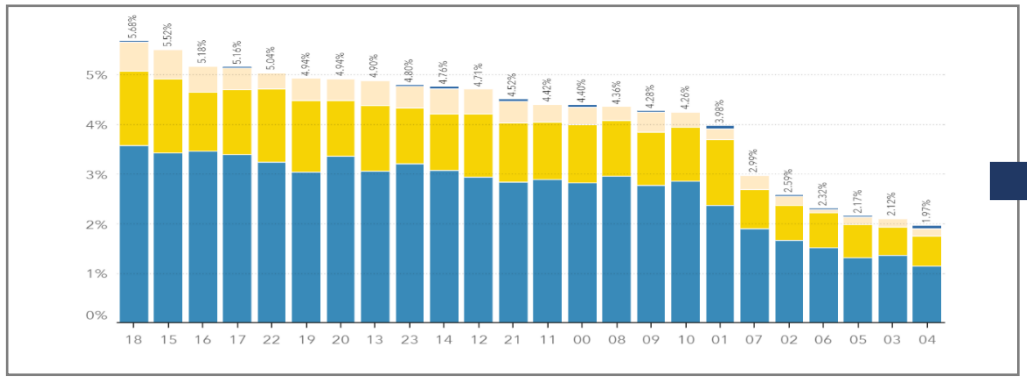
2. 원인 탐색을 위한 시각화 - 도로환경요인

- 사고 발생 월별 분포를 살펴본 결과 뚜렷한 차이가 없으나, 중상사고의 경우 10월, 5월이 각각 2.95%, 2.82%로 높은 비중을 차지함
- 사고 발생 시간은 1:00~7:00AM 까지의 빈도가 현저히 낮았으며, 시간 환경적 특성과 통행량 특성을 고려하여
  오전(7:00AM~11:59AM), 오후 (12:00PM~16:59PM), 저녁(17:00PM~19:59PM), 밤(20:00PM~00:59AM), 새벽(1:00AM~6:59AM)으로 범주화 변수 생성
- 사고 발생 요일은 금요일이 가장 높았으며, 평일/주말의 환경적 특성과 통행량 특성을 고려하여 평일을 월-목, 금,토,일로 범주화 변수 생성

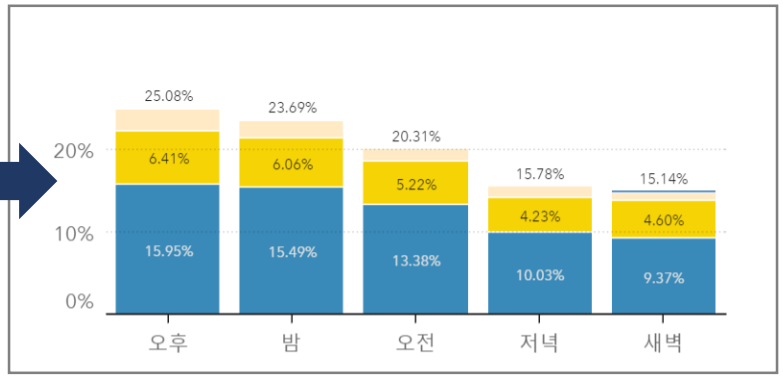
발생 월



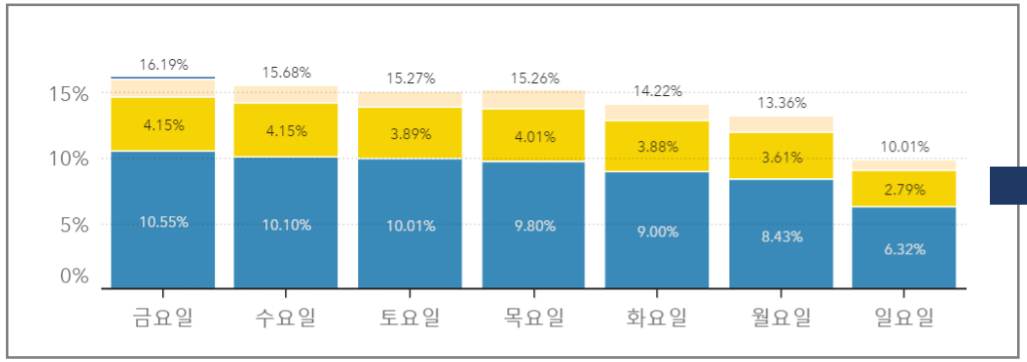
발생 시간



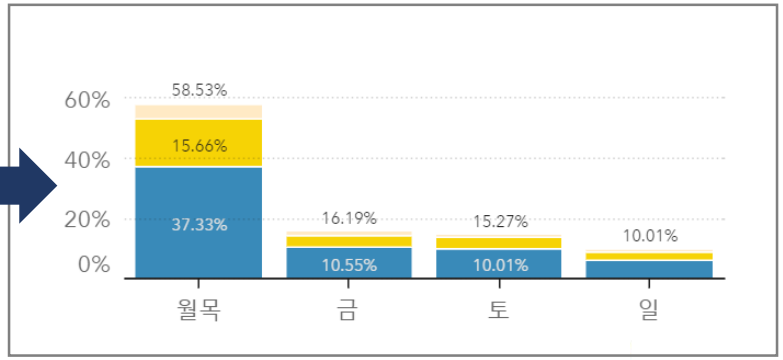
발생 시간그룹



발생 요일



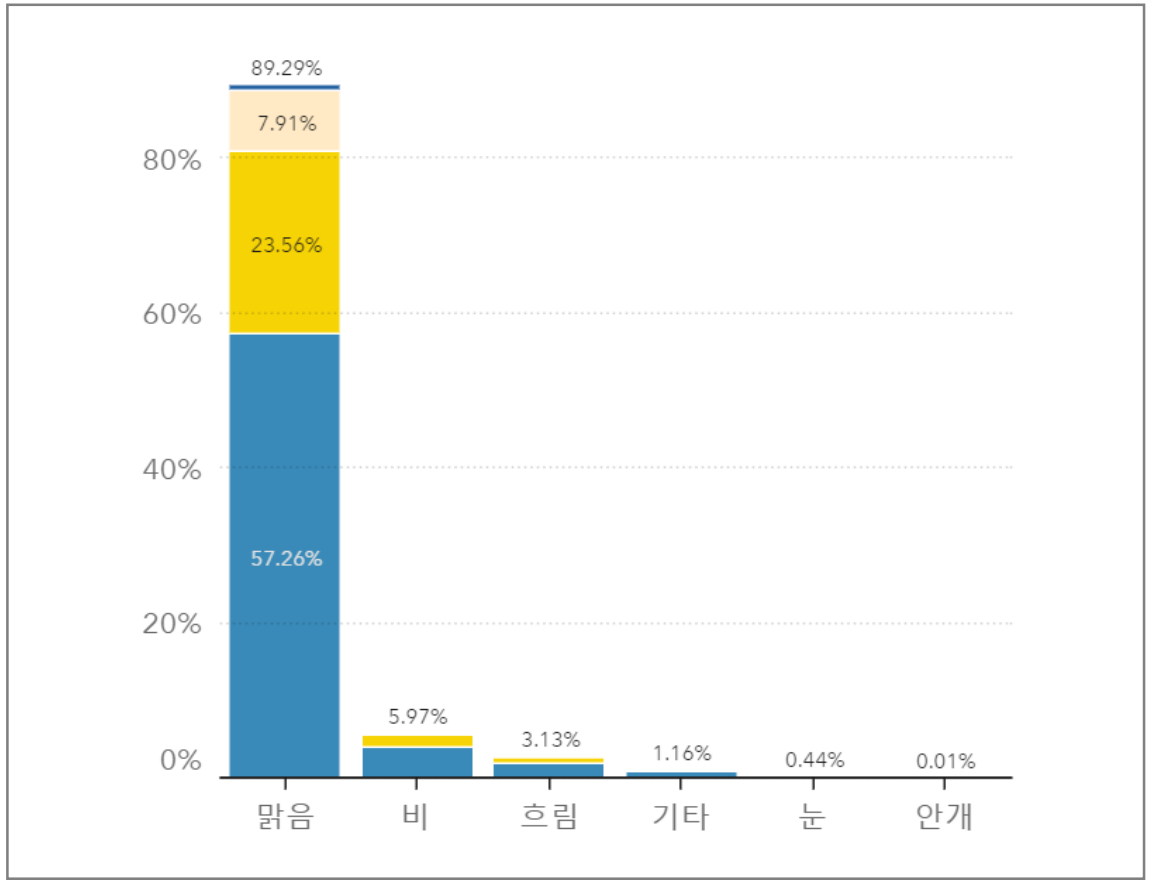
발생 요일그룹



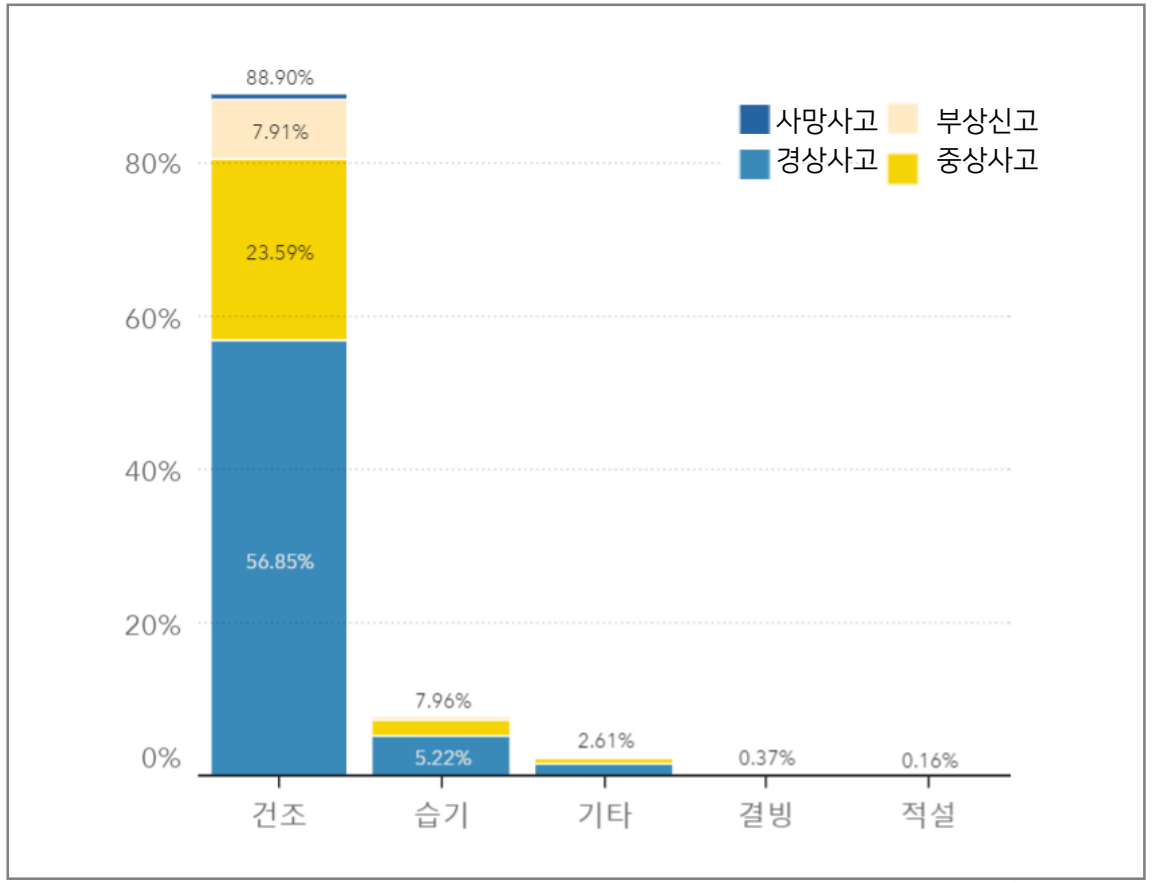
2. 원인 탐색을 위한 시각화 - 도로환경요인

- 사고 발생 기상상태의 분포를 살펴본 결과, '맑음'의 경우가 89.29%로 가장 높음
- 사고 발생 도로상태의 경우 '건조' 일 경우가 가장 높았으며, '습기', '기타', '결빙', '적설' 순으로 나타남
- '기상상태'와 '도로상태' 변수의 경우 상관관계가 높을 것으로 예상됨

기상상태



도로상태

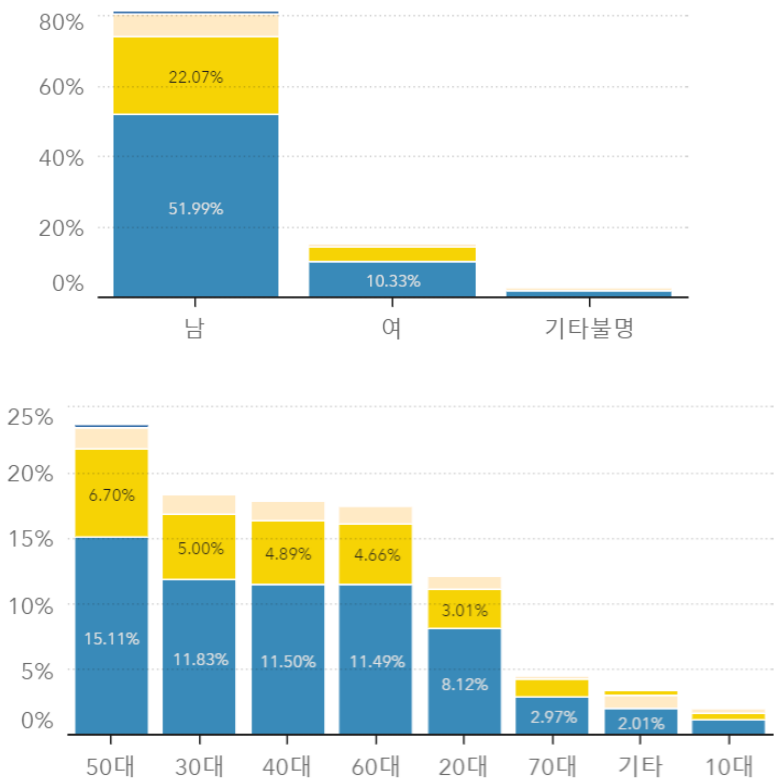


2. 원인 탐색을 위한 시각화 - 인적요인& 차량요인

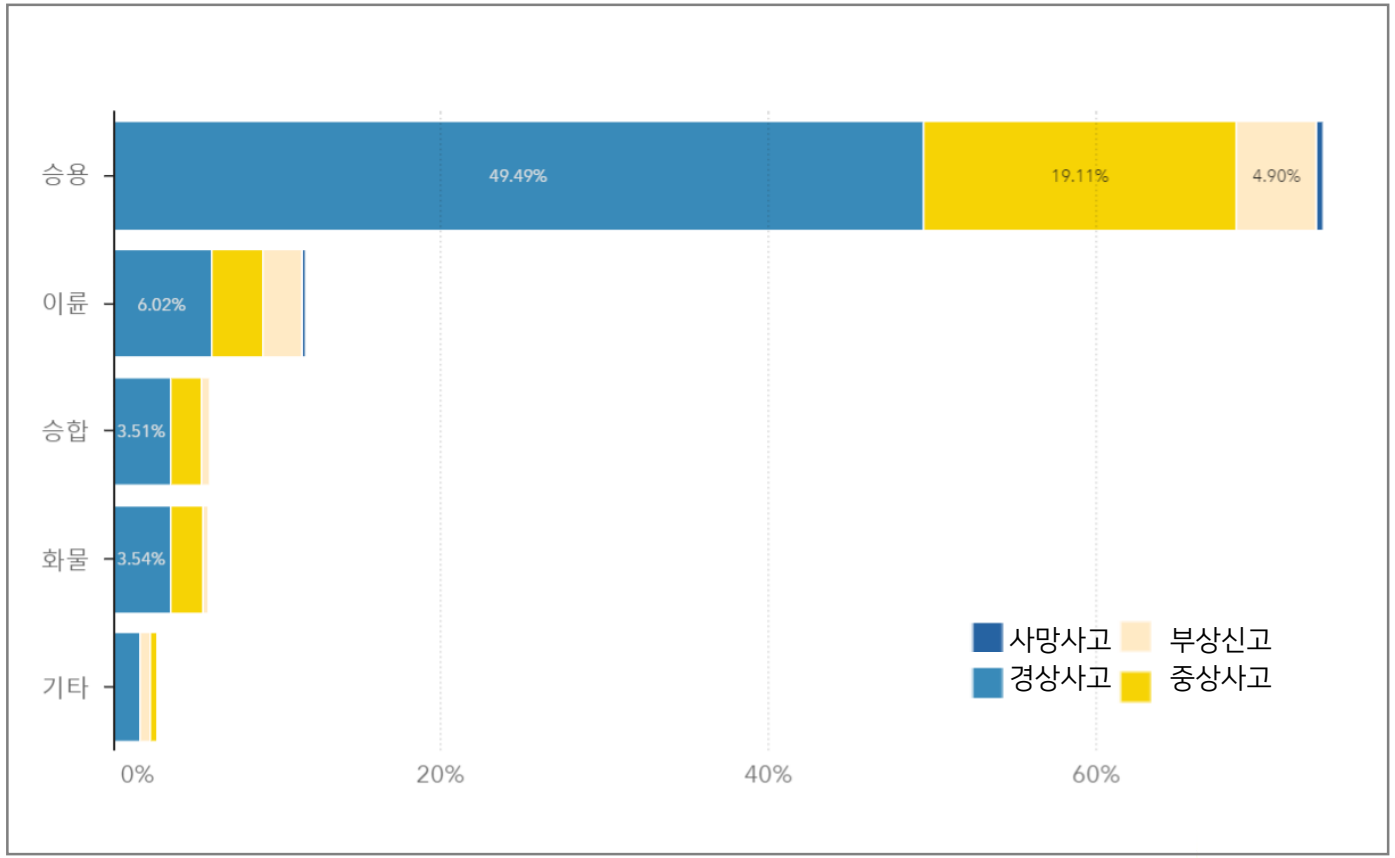
- 사고 발생 가해자의 성별은 여성인 경우에 비해 남성의 비율이 뚜렷하게 높았고, 사고 발생 가해자의 연령대는 50대, 30대, 40대 ,60대, 20대 순으로 나타남
- 사고 발생 가해자의 차종별 분석 결과, 승용의 비율이 압도적으로 높은 비율을 나타냄

-기존 데이터에서 낮은 빈도와 특성을 고려하여, '사륜오토바이,개인형이동수단,자전거, 원동기,이륜'의 1-2인용 개인형 이동수단을 '이륜'으로,
'농기계,특수,건설기계,기타불명'에 대하여 '기타'로 그룹화

가해자 성별 및 연령



가해자 차종



Ⅱ. 데이터 전처리

1. 분석용 데이터(DATA_FINAL) 변수 설명

- 위도, 경도가 존재하는 중구(11140), 영등포구(11560), 강남구(11680)에 해당하는 14,579개의 관측치 중 지도 위에 나타내었을 때 해당 3개구가 아닌 15개의 지점은 관측 오류로 판단하여 분석에서 제외 후 총 14,564개의 관측치를 최종적으로 분석에 사용함
- DTG, 시설물 데이터를 사고 데이터와 매칭하기 위하여 위도,경도가 있는 자료점만을 분석에 사용함

index	column name	type	label
1	No	ID	사고번호
2	SEVERE **	Numeric	사고심각도
3	MONTH	Categoric	월
4	HOUR_GR	Categoric	발생시간대
5	DAY_GR	Categoric	발생요일군
6	Weather	Categoric	기상상태
7	Surface_A	Categoric	노면상태A
8	Surface_B	Categoric	노면상태B
9	Att_car	Categoric	가해운전자차종
10	Att_gender	Categoric	가해운전자성별
11	AGE_GR	Categoric	가해운전자연령대

index	column name	type	label
12	XCORO	Numeric	교차로 수
13	BRIDGE	Numeric	교량 수
14	CCTV	Numeric	CCTV 수
15	CROSS	Numeric	횡단보도 수
16	HUMP	Numeric	험프 수
17	d1	Numeric	과속 지수
18	d2	Numeric	급가속 지수
19	d3	Numeric	급감속 지수
20	d4	Numeric	급회전 지수
21	d5	Numeric	급진로변경 지수

2. 종속변수 설명 - SEVERE(사고 심각도)

- Y(종속변수)를 사고 건수나 사고 확률로 놓기보다는 사고 피해 정도의 의미를 담고 있는 '사고 심각도' 로 설정
- 교통사고 관련 분석에서 자주 활용되는 EPDO 지수를 변환하여 활용
- EPDO 지수에서 물피건수를 부상신고자수로 대체

*EPDO (Equivalent Property Damage Only: 대물피해환산법지수)

$$SEVERE = 12 \times (\text{사망자수}) + 3 \times (\text{중상자수} + \text{경상자수}) + 1 \times (\text{부상신고자수})$$

출처 :
-김경환, 박병호, "차량유형에 따른 교통사고심각도 분석모형 개발"(Journal of the KOSOS, Vol. 25, No. 3,2010), 132
-강승림, 박창호, "고속도로 선형조건과 GIS 기반 교통사고 위험도지수 분석"
-Lu Ma. Modeling the equivalent property damage only crash rate for road segments using the hurdle regression framework

2. 파생변수 설명 - 근방 시설물 수

- 제공된 모든 시설물 데이터에 대해 탐색 후 중요도를 판단하여 교차점, 교량, CCTV, 횡단보도, HUMP를 모형에 포함하기로 판단
- 각 교차점의 길이 및 설치 기준을 바탕으로 시설물 반경을 결정
- 위도, 경도 차이를 거리(m)로 환산 후 반경을 이용해 위도와 경도의 상하좌우 [min, max] 범위를 구하여 시설물 데이터의 위도, 경도가 범위 내에 속한 시설물 개수를 파생변수로 생성

	분류 기준	반경	분류 기준 근거
교차로 XDORO	도로폭 10m 미만	15m	왕복 2차선 도로의 경우 도로 폭이 10m 정도로 측정되었고, 교차점 부근을 포함할 수 있게 더 크게 반경을 설정하기로 결정
	10m~20m미만	30m	왕복 4차선 도로의 경우 도로 폭이 평균 20m로 측정
	20m 이상	50m	4차선 이상의 큰 도로에 대해 비례하여 설정
교량 BRIDGE	대교	750m	한강 다리 중 가장 짧은 한남대교의 길이 915m를 참조해 기준 설정
	교	250m	시설물 데이터 중 대교를 제외한 나머지 데이터들의 분포에서 평균 250m로 반경을 설정
CCTV		200m	논문 [이건학, "공공 CCTV의 공간 분포 특성과 가시 커버리지에 기반한 최적 입지" (대학지리학회지 제53권 제3호 2018), 405-425] 참조해 CCTV 영향력 범위 200m로 설정
횡단보도 CROSS		15m	가장 일반적인 왕복 4차선 도로의 폭이 20m 수준인 것을 고려하였을 때 횡단보도 부근 지점을 포함하도록 설정
과속방지턱 HUMP		5m	과속방지턱 설치 기준(도로 폭 6m 이상일 때 3m60cm, 도로 폭 6m 미만일 때 2m로 설치)을 참조해 기준 설정

2. 파생변수 설명 - d1, d2, d3, d4, d5

- 위험운전 행동분석 기준에 해당하는 데이터를 기존의 사고 데이터와 결합하기 위하여 DTG 데이터로부터 파생변수 D1,D2,D3,D4,D5를 생성하여 각 사고 발생 지점에 해당하는 위험운전 행동의 발생 정도를 반영함

Step1

'위험운전 행동분석 기준'의 정의에 따라 각 자료점이 기준에 해당하면 d1(과속유형), d2(급가속유형), d3(급감속유형), d4(급회전유형), d5(급진로변경유형) 변수의 값을 1, 해당하지 않으면 0으로 할당

Step2

위도(lat), 경도(lon) 변수를 0.0005(약 50m 차이)로 반올림함
위도그룹(lat_gr), 경도그룹(lon_gr) 변수로 그룹화한 후,
자료의 위험운전 행동 빈도수를 더하여 새로운 d1, d2, d3, d4, d5를 생성

Step3

통행 가중치를 부여하기 위하여 전체 관측치 수로 나눠준 후,
값의 범위를 조절하기 위하여 10^6 곱하여 최종 값을 증가

$$\frac{di}{n} \times \frac{n}{N} = \frac{di}{N}$$

n=그룹내 자료의 관측치수, N = 총 관측치수, di = 그룹 내 i번째 위험행동건수

Example)

lat	lon	d1	d2	d3	d4	d5
37.456146	127.04687	0	0	0	0	0
37.456151	127.04688	1	0	0	0	0
37.456198	127.0469	0	0	0	0	0
37.45615	127.0469	1	0	0	0	0
37.456215	127.0469	1	1	0	1	0
37.456216	127.04693	1	1	0	0	0
37.456248	127.04693	0	0	0	0	0
37.456119	127.04694	1	0	0	0	0
37.456135	127.04695	1	1	0	1	0
37.456213	127.04696	1	0	0	1	0

Rounding

Sum

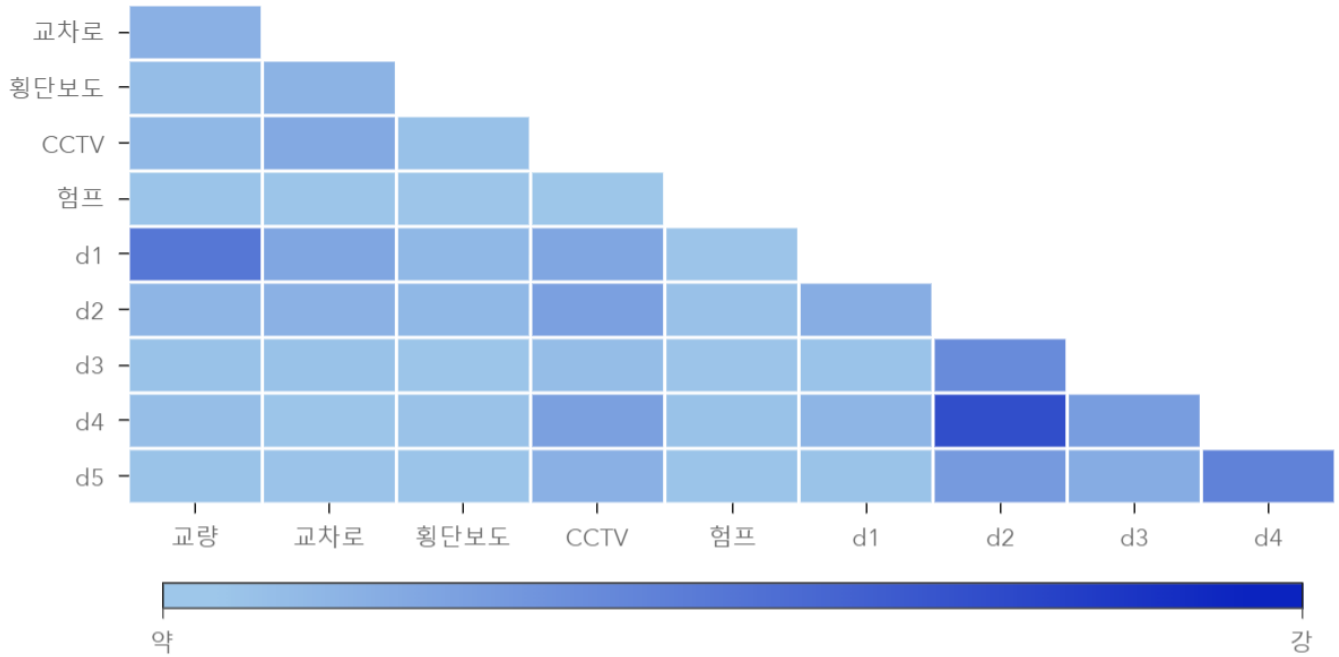
lat_gr	lon_gr	d1	d2	d3	d4	d5
37.456	127.015	0	0	0	0	0
37.456	127.0155	0	0	0	0.0135564058	0
37.456	127.0175	0	0	0.0067782029	0	0
37.456	127.018	0	0	0	0	0
37.456	127.0185	0	0	0	0	0
37.456	127.0205	0.0067782029	0	0	0	0
37.456	127.021	0.0406692174	0.0406692174	0.0406692174	0.0881166377	0
37.456	127.0465	0.0338910145	0	0	0	0
37.456	127.047	0.2236806957	0.0948948406	0.0067782029	0.0610038261	0

3. 파생변수 다중공선성 진단

- 설명 변수 간 다중공선성을 판단하기 위하여 산점도행렬을 시각화한 결과, d1과 교량, d2와 d4는 서로 상관관계가 높은 것으로 나타남
- 따라서 모델 적합 전 다중공선성을 해소시키기 위한 방안이 요구됨

  (이후 일반화 선형모형에서 변수간 상관관계를 제거하는 Lasso 선택 방법 채택의 근거)

index	column name	type	label
12	XCORO	Numeric	교차로 수
13	BRIDGE	Numeric	교량 수
14	CCTV	Numeric	CCTV 수
15	CROSS	Numeric	횡단보도 수
16	HUMP	Numeric	험프 수
17	d1	Numeric	과속 지수
18	d2	Numeric	급가속 지수
19	d3	Numeric	급감속 지수
20	d4	Numeric	급회전 지수
21	d5	Numeric	급진로변경 지수



Ⅲ. 모형구축

- Train Data : 70%
- Validation Data : 30%

Gamma GLM

- 일반화 선형 모형 (Gamma 분포, 로그 링크 함수)
- 반응변수(SEVERE)의 히스토그램을 확인한 결과 오른쪽 꼬리가 긴 모양을 보이고 있어 한쪽으로 치우쳐진 연속형 데이터에 대해 주로 적합이 가능한 Gamma 분포를 가정하기로 함. 또한, 관련 연구*에서 사고심도에 대해서는 주로 감마 분포를 적용한다는 점을 참고함

Gradient Boosting

Random Forest

- 연속형 종속변수에 적합한 머신 러닝 기법들 중 예측력이 높은 앙상블 기법들 중 크게 두 가지인 그래디언트 부스팅과 랜덤 포레스트 알고리즘을 사용하기로 함
- 결과해석에 있어서 GLM보다는 다소 어렵지만, 예측력 면에서는 GLM보다 더 뛰어날 것으로 예상됨

* 기승도,김대환,"일반화선형모형(GLM)을 이용한 자동차보험 요율상대도 산출방법 연구"(보험연구원, 2009), 11-12

2. 모형 적합 - 일반화 선형 모형 (Gamma 분포)

Selection Information	
Selection Method	LASSO
Stop Criterion	AIC
Choose Criterion	AIC
Effect Hierarchy Enforced	None

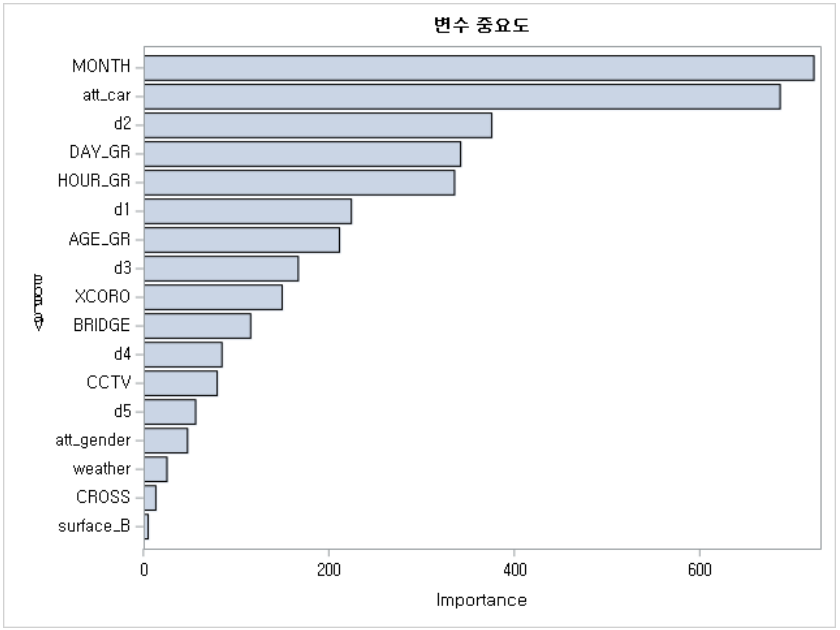
Fit Statistics		
Description	Training	Validation
-2 Log Likelihood	48757	20672
AIC (smaller is better)	48809	20724
AICC (smaller is better)	48809	20725
SBC (smaller is better)	48997	20890
Average Square Error	11.43489	7.96758

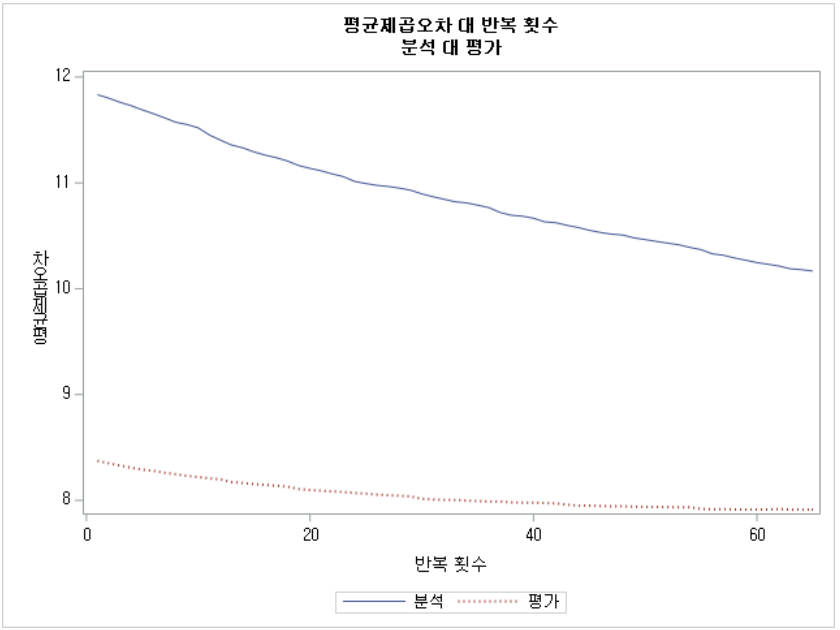
Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	1.276629
HOUR_GR 밤	1	0.023216
HOUR_GR 새벽	1	0.094767
HOUR_GR 오전	1	-0.028148
HOUR_GR 오후	1	-0.047375
HOUR_GR 저녁	1	-0.042566
DAY_GR 금	1	-0.023511
DAY_GR 월목	1	-0.053089
DAY_GR 일	1	0.039276
DAY_GR 토	1	0.037311
att_car 기타	1	-0.024888
att_car 승용	1	0.036851
att_car 승합	1	0.113324
att_car 이륜	1	-0.161842
att_car 화물	1	0.036133
att_gender 기타불명	1	-0.179878
att_gender 남	1	0.108063
att_gender 여	1	0.070840
XDORO	1	0.003857
BRIDGE	1	0.006909
CCTV	1	-0.014111
CROSS	1	-0.021003
d1	1	0.021956
d2	1	0.035409
d3	1	0.026587
d4	1	0.005465
Dispersion	0	1.000000

- AIC 기준의 LASSO 선택 방법을 이용한 결과, 최종 선택된 모형은 4개의 명목형 변수와 8개의 연속형 변수로 총 12개의 변수를 포함하고 있으며, 검증용 데이터에 대한 ASE는 7.97으로 계산됨

- 시간대(HOUR_GR) 범주 중에서는 밤과 새벽이, 요일군(DAY_GR) 범주는 토요일과 일요일이 계수가 양수이므로 사고심각도를 높이는 요인으로 작용함
- 차종(ATT_CAR)의 경우 다른 차종에 비해 승합차의 계수가 0.113324로 가장 큼
- 시설물 변수들 중 도로 주변의 교차로(XDORO)와 교량(BRIDGE)은 각각 한 단위 증가함에 따라 사고심각도는 1.0039, 1.0069배 증가하며 CCTV는1대당 0.9860배, 횡단보도(CROSS)는 0.9792배로 감소함
- 위험운전 행동지수 변수 중 급진로변경(d5)을 제외한 나머지 변수들이 모두 모형에 포함되었으며, 사고심각도를 높이는 데 영향을 줌

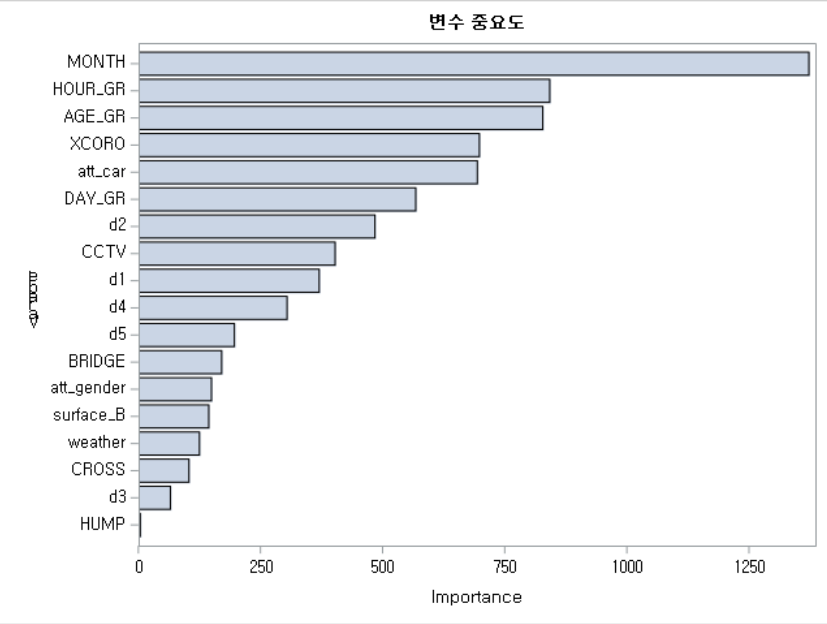
Best Configuration	
Evaluation	31
Number of Trees	74
Number of Variables to Try	7
Learning Rate	0.03487077
Sampling Rate	0.85858679
Lasso	9.70789511
Ridge	3.70120385
Average Squared Error	7.9115497756

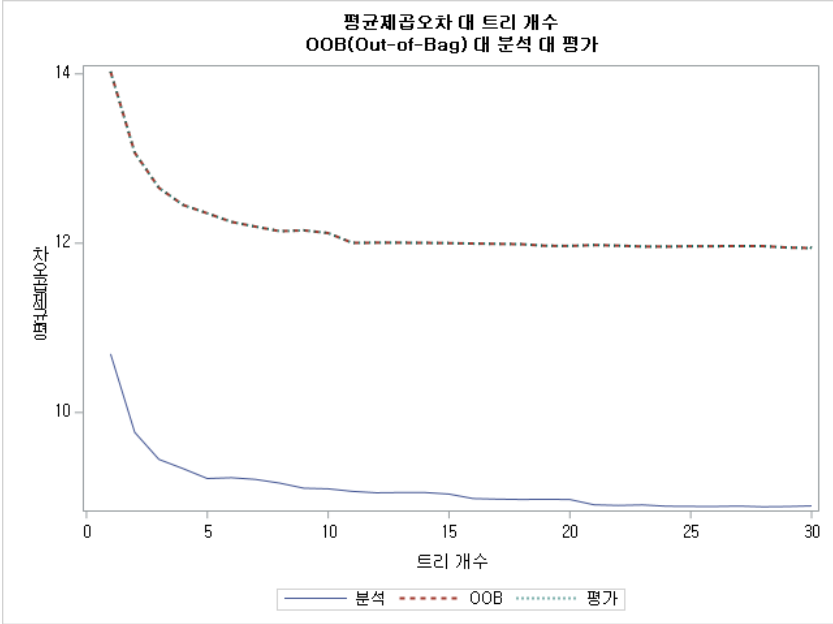




- 그래디언트 부스팅의 모든 옵션을 자동 튜닝으로 설정하여 반복 횟수, 최대 트리 깊이, 가지 수, 정규화 등을 최적으로 선택하게 함
- 최적화된 Tree의 수는 74개이며 ASE는 7.91으로 계산됨
- 변수의 중요도는 상위 7개 기준 **발생월(MONTH), 차종(ATT_CAR), 급가속지수(d2), 요일군(DAY_GR), 시간대(HOUR_GR), 과속지수(d1), 운전자연령대(AGE_GR)** 순으로 높은 것을 확인할 수 있음

Best Configuration	
Evaluation	40
Number of Trees	139
Number of Variables to Try	9
Bootstrap	0.38443454
Maximum Tree Levels	9
Average Squared Error	7.9478346407





- 랜덤 포레스트의 모든 옵션을 자동 튜닝으로 설정하여 반복 횟수, 최대 트리 깊이, 가지 수, 정규화 등을 최적으로 선택하게 함
- 최적화된 Tree의 수는 139개이며 ASE는 7.95로 계산됨
- 변수의 중요도는 상위 7개 기준 **발생월(MONTH), 시간대(HOUR_GR), 운전자연령대(AGE_GR), 교차로 수(XDORO), 차종(ATT_CAR), 요일군(DAY_GR), 급가속지수(d2)** 순으로 높은 것을 확인할 수 있음

Ⅳ. 모형평가

- 데이터 분할이 되지 않은 전체 데이터 셋을 TEST 데이터로 활용하여 스코어링함
- 각 모형의 평균 제곱 오차와 향상도 곡선을 기준으로 활용하여 모형 비교를 진행함

- 모형별 MSE 비교: 세 가지 모형의 평균제곱오차를 비교한 결과 랜덤포레스트가 0.1680으로 가장 작음

Gamma GLM

Fit Statistics							
	Squared Error			Absolute Error		Squared Logarithmic Error	
Number of Observations	Divisor of Average	Average	Root Average	Mean	Root Mean	Mean	Root Mean
14564	14564	10.39474	3.224088	1.803836	1.343070	0.179625	0.423821

Gradient Boosting

Fit Statistics							
	Squared Error			Absolute Error		Squared Logarithmic Error	
Number of Observations	Divisor of Average	Average	Root Average	Mean	Root Mean	Mean	Root Mean
14564	14564	10.22130	3.197077	1.812192	1.346177	0.178779	0.422822

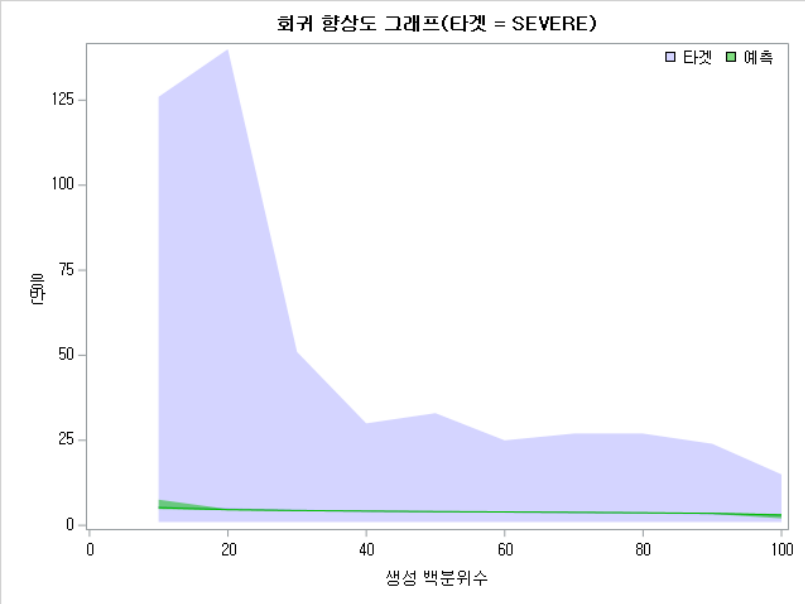
Random Forest

Fit Statistics							
	Squared Error			Absolute Error		Squared Logarithmic Error	
Number of Observations	Divisor of Average	Average	Root Average	Mean	Root Mean	Mean	Root Mean
14564	14564	9.526602	3.086519	1.751770	1.323544	0.168002	0.409880

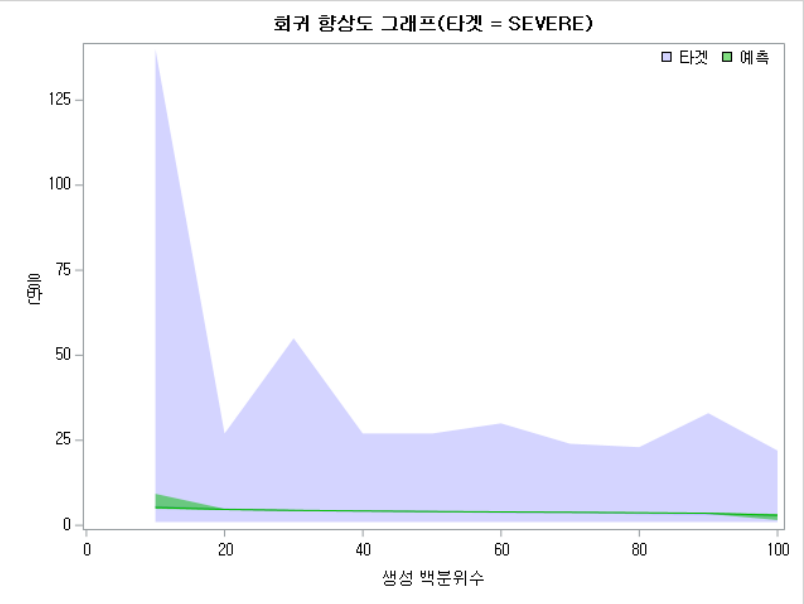
2. 최종 모형 선택

- 모형별 향상도 곡선(Lift Curve) 비교
  상위 등급에서 향상도가 제일 높고, 하위 등급으로 갈 수록 향상도가 급격하게 감소하는 경향이 랜덤 포레스트 모형에서 가장 뚜렷하므로 예측력이 좋을 것으로 판단

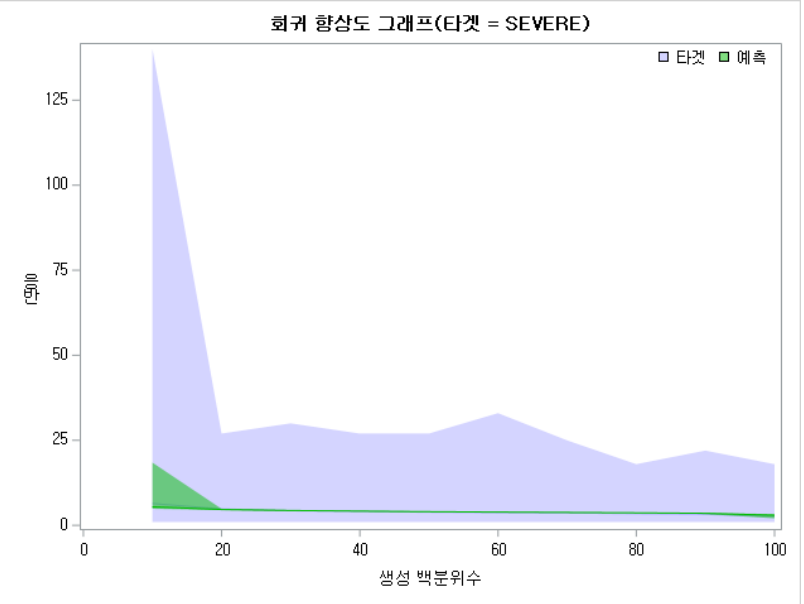
Gamma GLM



Gradient Boosting



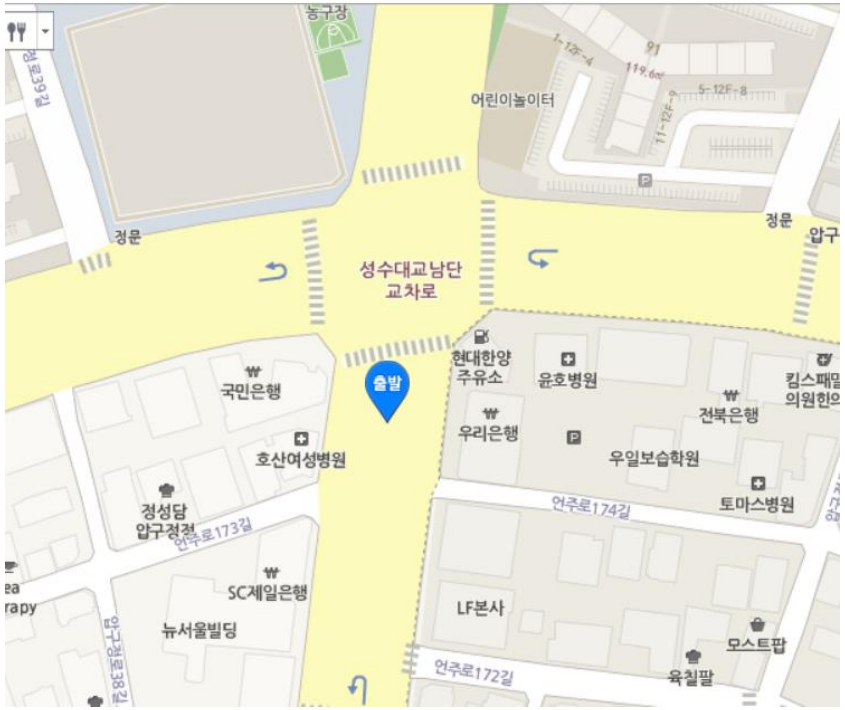
Random Forest



따라서 랜덤포레스트 모형을 최종 모형으로 선정하여 사후 분석 및 적용에 활용하기로 함

Ⅴ. 사후분석 및 활용방안

Example)



[사 고 정 보]

- 사고번호 : 20160206001004 80
- 사고유형-법규위반내용 : 차대차-안전운전불이행
- 사고내용 : 중상사고
- 가해차종 : 승용차
- 시설물 정보 : 교차로 2, 교량 0, CCTV 1, 횡단보도 1, HUMP 0
- d1 :0.560244, d2 : 3.281432, d3 : 0.040017, d4 : 3.121362, d5 : 0.16007

SEVERE(사고심각도) : 24

예측값 : 18.7396

스마트 횡단보도

보행신호 음성안내 장치

- 도로 특성에 맞는 시설물의 설치 여부 검토
- 위험운전행동지수 값이 높게 나타나는 특정구간에 대해 해당 위험행동을 집중 단속
- 설치가 추가로 필요한 횡단보도, 육교 및 어린이 보호 구역 등의 시설물 고려
- 때에 따라 스마트 횡단보도, 횡단보도 안전대기장치, 횡단보도 및 보행신호 음성안내 장치 등의 스마트 시설물 설치 및 이용
- 시간, 요일별 통행량에 따른 특성을 고려하여 위험 시간대 주의 필요

- 현 위반 벌금은 현재 승합자 기준 최대 7만원으로 책정되어 있음
- 교통사고 발생건수에서 안전운전불이행의 비율이 가장 높은 것으로 보아, 해당 위반 행동 시 벌금 및 벌칙의 강화 필요
- 졸음운전, 주행 중 핸드폰 사용, 음주운전 등 안전운전불이행 행동 신고 시 포상금 제도 도입
- 직장 워크샵, 학교 또는 구청 등에서 다양한 교육 컨텐츠 제공
- 미디어를 통해 웹드라마나 웹툰으로 안전운전의 중요성 강조
- 해외의 경우를 벤치마킹해 적립된 범칙금으로 모범 운전자에게 이벤트성 로또 지급

안전교육	공익 웹툰	스웨덴의 스피드 로또

노후차량 점검 및 폐차

타이어 체인

- 노후차량은 국가적으로 정기점검을 권유하며, 해당 차량이 위험하다고 판단되었을 때 사용 시 벌금제도 필요
- 눈이 많이 오는 날씨나 도로가 미끄러운 경우, 타이어 체인 등을 대여해주는 사업 실시
- 승합차의 경우 사고 심각도가 높아졌으므로, 승합 또는 화물차 운전자에게 교육의 필요성 증진

1) 위경도가 3개구에만 제한되어 있어 대규모의 나머지 사고 데이터를 분석시 활용하지 못한 점이 아쉬움. 3개구만을 이용한 분석을 서울시 전체에 적용할 경우 일반화에는 한계가 있을 것이라고 예상함
2) 주어진 DTG 데이터는 택시 운전자의 데이터로, 서울시 일반 운전자와 상이한 운전 패턴을 보여 실제 도로 주행 시 다른 행동을 보일 수 있음
3) 도로구간 별 길이와 폭을 고려해 각 구간을 구분한 위경도 범위 또는 폴리곤 데이터가 존재해 분석에 활용했다면 예측의 정확도를 조금 더 높일 수 있을 것이라고 예상함
4) 전체 통행량 데이터로 같은 시간, 동일한 도로 상에 사고가 일어나지 않은 데이터를 만들 수 있다면 새로운 모형을 만들 수 있을 것이라 기대
5) 공공데이터 중 자동차전용도로 위치정보에 대한 정확성이 매우 떨어져 자동차전용도로 여부를 반영하지 못함. 그로 인해 과속여부 판단 시 일반도로의 제한속도 기준을 동일하게 적용해 변수의 정확성에 한계가 있음 (ex. 자동차전용도로의 과속 기준속도는 110km/h지만, 일반도로의 기준인 80km/h으로 적용됨)
6) 시설물데이터 파생변수 생성 시 사고지점과의 거리를 측정하여 수치화한 변수를 사용했다면 기존의 모형보다 더 예측력이 높아질 것으로 기대됨

[Appendix]

Appendix. 코드 리스트

시설물 데이터 생성

```
data sa167.facility;
    set sa167.doroo3(keep=lat_min lat_max lon_min lon_max fac)
        sa167.bridge3(keep=lat_min lat_max lon_min lon_max fac)
        sa167.cctv3(keep=lat_min lat_max lon_min lon_max fac)
        sa167.cross3(keep=lat_min lat_max lon_min lon_max fac)
        sa167.hump3(keep=lat_min lat_max lon_min lon_max fac)
    ;
run;
```

각각의 3개구 시설물 데이터에서 시설물 반경의 위도와 경도 구간과 시설물 유형을 변수로 가지는 테이블 생성

사고시설물 JOIN

```
proc sql;
create table a as
select a1.*, a2.a, a2.b, a2.c, a2.d, a2.e
from sa167.accident_new3 a1
inner join (select no
          , sum(a) as a
          , sum(b) as b
          , sum(c) as c
          , sum(d) as d
          , sum(e) as e
          from (select a1.no
               , case when a2.fac  = '교차' then 1 else 0 end as a
               , case when a2.fac = '교량' then 1 else 0 end as b
               , case when a2.fac = 'CCTV' then 1 else 0 end as c
               , case when a2.fac = '횡단' then 1 else 0 end as d
               , case when a2.fac = '험프' then 1 else 0 end as e
               from (select no, lat, lon from sa167.accident_new3) a1
               left join sa167.facility a2
               on a2.lat_min <= a1.lat <= a2.lat_max
                 and a2.lon_min <= a1.lon <= a2.lon_max
               )
          group by 1) as a2
on a1.no = a2.no
;quit;
```

사고 데이터의 위경도와 시설물 구간의 위경도를 비교하여 각 시설물의 근방에 해당할 경우 a-e에 1을 할당하고, ID인 no을 기준으로 테이블 조인 시 각 사고 관측치별 a~e의 1의 개수를 더해 사고지점 근방의 해당 시설물의 개수를 각각 계산할 수 있음

Appendix. 코드 리스트

D1 변수생성1

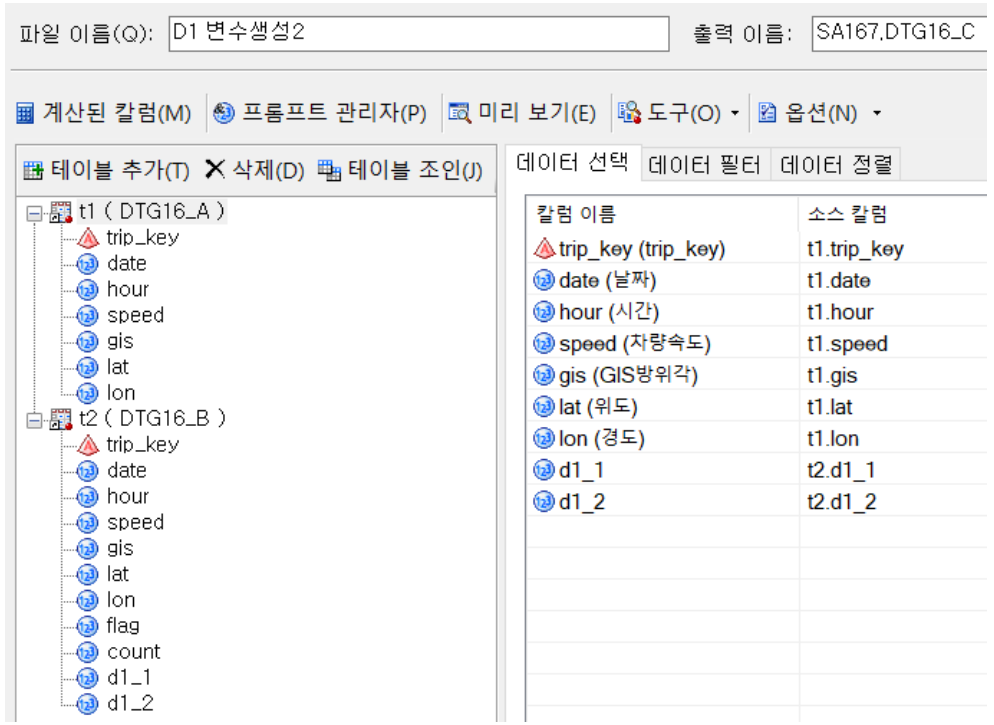
```
data sa167.dtg16_a;
    set sa167.dtg16_data;
run;

data sa167.dtg16_b;
    set sa167.dtg16_a;
    if speed>80;
run;

proc sort data=sa167.dtg16_b;
    by trip_key descending hour;
run;
data sa167.dtg16_b;
    set sa167.dtg16_b;
    by trip_key descending hour;
    if dif(hour) ne -1 then flag=2;
run;
proc sort data=sa167.dtg16_b;
    by trip_key hour;
run;
data sa167.dtg16_b;
    set sa167.dtg16_b;
    by trip_key hour;
    if dif(hour) ne 1 then flag=1;
run;
```

```
data sa167.dtg16_b;
    set sa167.dtg16_b;
    if flag=1 then count=0; count+1;

    if mod(count,3)=1 then d1_1=1;
    else d1_1=0;

    if count>=180 then
        do;
            if mod(count,3)=0 then d1_2=1;
            else d1_2=0;
        end;
    else d1_2=0;
run;
```

과속 기준을 판단하기 위해 일반 도로 기준 시속 80km을 관측치들만을 추출한 뒤 연속된 시점(초)에 대해 시작 시점에 플래그 1, 종료 시점에 플래그 2을 할당하고, 3초를 기준으로 과속(d1_1)을 1로 할당함. 장기 과속(d1_2)을 판단하기 위해 연속된 시간이 3분을 넘는 자료들에 대해서는 3분이 넘는 시점부터 3초마다 한번씩 1을 할당함. (3으로 나눈 나머지를 이용)

과속 기준에 해당하는 관측치들만 뽑아 변수를 할당하였으므로 나머지 모든 관측치들에 d1_1과 d1_2 변수를 모두 0으로 할당하기 위해 다시 테이블을 조인하고 결측치로 처리된 변수 값들에 0을 할당함

D1 변수생성2



D1 변수생성3

```
data sa167.dtg16_d1;
    set sa167.dtg16_c;
    if d1_1=. then d1_1=0;
    if d1_2=. then d1_2=0;
    d1 = max(d1_1, d1_2);
run;
```

Appendix. 코드 리스트

D2-D5 변수생성

```
proc sort data=sa167.dtg16_d1 out=work.dtg16;
    by trip_key;
run;

data sa167.dtg16_d;
    set work.dtg16;
    by trip_key;
    keep trip_key date hour speed gis lat lon d1 d2 d3 d4 d5;

    /*d2 급가속*/
    if lag(speed)>=6 and dif1(speed)>=8 then d2=1;
    else if lag(speed)<=5 and dif1(speed)>=10 then d2=1;
    else d2=0;

    /*d3 급감속*/
    if dif1(speed)<=-14 and speed>=6 then d3=1;
    else if dif1(speed)<=-14 and speed<=5 then d3=1;
    else d3=0;

    /*d4_1 급좌우회전*/
    if min(lag3(speed),lag2(speed),lag(speed),speed)>=30 then
       do;
          gisdiff=max(lag3(gis),lag2(gis),lag1(gis),gis)-min(lag3(gis),lag2(gis),lag1(gis),gis);
          if 60<=gisdiff<=120 then d4_1=1;
          else d4_1=0;
          if first.trip_key or lag(first.trip_key) or lag2(first.trip_key) then d4_1=0;
       end;
    else d4_1=0;
    /*d4_2 급유턴*/
    if min(lag6(speed),lag5(speed),lag4(speed),lag3(speed),lag2(speed),lag(speed),speed)>=25 then
       do;
          gisdiff=max(lag6(gis),lag5(gis),lag4(gis),lag3(gis),lag2(gis),lag(gis),gis)
                  -min(lag6(gis),lag5(gis),lag4(gis),lag3(gis),lag2(gis),lag(gis),gis);
          if 160<=gisdiff<=180 then d4_2=1;
          else d4_2=0;
          if first.trip_key or lag(first.trip_key) or lag2(first.trip_key)
           or lag3(first.trip_key) or lag4(first.trip_key) or lag5(first.trip_key) then d4_2=0;
       end;
    else d4_2=0;
    d4 = max(d4_1, d4_2);

    /*d5 급진로변경*/
    if min(lag6(speed),lag5(speed),lag4(speed),lag3(speed),lag2(speed),lag1(speed),speed)>=30
     and abs(lag5(dif1(gis)))>=10 and abs(dif5(gis))<=2 then
       do;
          maxdiff = max(abs(lag4(dif1(speed))),abs(lag3(dif1(speed))),abs(lag2(dif1(speed))),
                        abs(lag1(dif1(speed))),abs(dif1(speed)));
          mindiff = min(abs(lag4(dif1(speed))),abs(lag3(dif1(speed))),abs(lag2(dif1(speed))),
                        abs(lag1(dif1(speed))),abs(dif1(speed)));
          if maxdiff<=2 or mindiff>=3 then d5=1;
          else d5=0;
       end;
    else d5=0;

    if first.trip_key then do; d2=0; d3=0; end;
    if first.trip_key or lag(first.trip_key) or lag2(first.trip_key)
     or lag3(first.trip_key) or lag4(first.trip_key) or lag5(first.trip_key) then d5=0;

run;
```

d2부터 d5의 각 행동 유형 판단 기준에 맞게 조건문을 설정함. (가속과 감속 및 속도에는 speed 관측치, 회전각에는 gis 관측치 참조)
각 행동 유형에 해당할 경우 1, 그렇지 않을 경우 0을 할당
trip_key(차량코드와 시간)을 기준으로 동일 운전자의 연속된 운전을 구분하여, 서로 다른 운전의 운전 행태를 연속적으로 판단하지 않도록 함

Appendix. 코드 리스트

DTG 위험지수 생성

```
proc sql;
    create table sa167.dtg_score as
    select lat_gr, lon_gr
          , sum(d1)/count*100000 as d1
          , sum(d2)/count*100000 as d2
          , sum(d3)/count*100000 as d3
          , sum(d4)/count*100000 as d4
          , sum(d5)/count*100000 as d5
    from (select lat_gr, lon_gr, d1, d2, d3, d4, d5
                , count(*) as count
          from sa167.dtg_danger_gr)
    group by lat_gr, lon_gr, count
    ;
run;
```

반올림으로 구간화한 위경도 구간을 이용해 같은 구간의 위험운전 행동 건수끼리 묶어 더하고, 이를 전체 관측치의 수로 나누어 구간별 위험행동 지수를 계산하여 다시 d1-d5 변수로 할당

최종 데이터 생성

```
proc sql;
  create table sa167.data_final as
  select t1.*, t3.d1, t3.d2, t3.d3, t3.d4, t3.d5
  from sa167.data_a t1
  inner join (select no
                    , sum(d1) as d1
                    , sum(d2) as d2
                    , sum(d3) as d3
                    , sum(d4) as d4
                    , sum(d5) as d5
              from (select t1.no, t2.d1, t2.d2, t2.d3, t2.d4, t2.d5
                    from (select no, lat, lon from sa167.data_a) t1
                    left join sa167.dtg_final t2
                    on t2.lat_min <= t1.lat <= t2.lat_max and t2.lon_min <= t1.lon <= t2.lon_max
                    )
              group by 1) as t3
  on t1.no = t3.no
  ;
  quit;
```

DTG데이터의 위경도 지점을 기준으로 일정 구간(min, max)을 생성한 뒤 (±0.0005) 사고 데이터의 위경도의 구간 해당여부를 판단하여 각 사고지점의 위험행동 지수를 더하여 no을 기준으로 테이블을 병합함

감사합니다 ☺