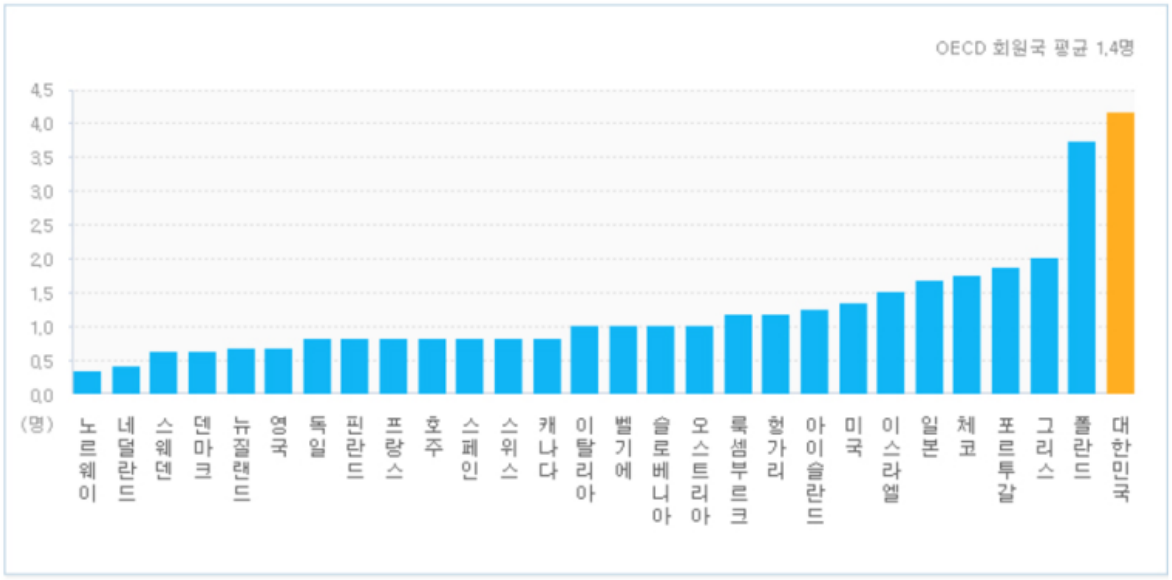
데이터마이닝 프로젝트 보고서

주제: 교통사고 요인 분석
학번: 201520852
이름: 유승우

1. 서론
-주제 선정 이유
아래의 그래프에서 보이는 바와 같이 OECD 회원국과 비교했을 때 상대적으로 높은 수치의 교통사고 사망자 수를 갖는 것을 볼 수 있습니다.



출처 : 도로교통공단. 2010 OECD 회원국 교통사고 비교 (2013년 판)

[그림1] OECD 인구 10만명 당 보행 중 교통사고 사망자수

물론 우리나라의 대인 교통사고 사망자 수는 계속해서 감소하고는 있으나 아직 개선할 여지가 많이 남아있지 않나 생각하게 되었습니다. 따라서 어떤 요인에 의해 대인 교통사고가 발생하는지 규명해보고 이에 따른 개선책을 구상해보고자 해당 주제를 선택하게 되었습니다.

2. 본론
Taas에서 2019년 한 해 동안 서울특별시 양천구에서 발생한 대인 교통사고 데이터에서 총 354개의 instances에서 다음 5개의 속성을 추출하였습니다.

-Attributes
① 면허 취득 경과 년도: 운전자가 운전 면허를 취득하고 지난 기간을 연 단위로 나타낸 것. 5년 미만, 5년 이상 10년 미만, 10년 이상 15년 미만, 15년 이상으로 분류.
② 주야: 사고 발생 시간이 낮 혹은 밤이었는지 나타냄.
③ 성별: 운전자의 성별
④ 차종: 승용차, 승합차, 화물차, 이륜차, 기타(건설기계, 농기계 등이 포함되나 비중이 너무

낮아 기타로 분류)

⑤ 도로형태: 사고 발생 지역의 도로형태. 단일로, 교차로, 터널, 기타(고가도로, 교량 위 등이 포함되나 비중이 낮아 기타로 분류)

-Class

5개의 속성을 통해 사고의 심각성 정도를 파악해보기 위해 피해자의 상태를 경상, 중상, 사망으로 나누었습니다. 다만 사망 사고의 인스턴스가 3건밖에 되지 않아 중상 이상의 피해가 발생한 사고를 중점적으로 요인을 추적하는 것을 목표로 하였습니다.

-알고리즘 학습 내용

선택한 알고리즘

One Rule, Decision Tree, Association Rules

선택 이유

프로젝트의 목적이 중상 이상의 사고에 대한 요인 파악에 있었으므로 가장 분류 능력이 높은 속성을 추출하고, 연관 규칙을 생성해보는 것이 목적에 부합한다고 생각했습니다.

결과

① One Rule

면허 취득 경과 년도가 규칙의 기준이 되는 속성으로 추출되었습니다.
5년 미만 -> 경상
5년 이상 10년 미만 -> 중상
10년 이상 15년 미만 -> 경상
15년 이상 -> 중상
총 354개의 인스턴스를 학습시켰을 때 약 76%의 정확도가 나왔습니다. 의외로 면허 취득 후 15년 이상 지난 운전자가 중상 이상의 교통사고로 이어진다는 점이 눈에 띄었습니다. 각 클래스에 따른 정밀도, 재현율은 다음과 같습니다.

| | 정밀도(Precision) | 재현율(Recall) |
|---|---|---|
| 경상 | 0.733 | 0.327 |
| 중상 | 0.764 | 0.955 |
| 사망 | ?(데이터가 너무 적음) | 0 |

경상 클래스에 대해서는 모델에서 경상으로 분류한 것은 73%정도로 실제로 경상이었지만 실제 경상인 데이터에 대해 모델이 경상으로 분류한 것은 33%정도로 낮았습니다.
Confusion Matrix에 의해서도 다시 확인되는데 경상을 중상으로 잘못 분류한 경우가 68건으로 가장 많았습니다.
중상 클래스에 대해서는 정밀도, 재현율 모두 경상 클래스보다 높은 것을 볼 수 있습니다. 이는 아무래도 중상 클래스의 학습 데이터가 가장 많았고, 이후 학습곡선에서도 나오지만 One Rule의 경우 학습 데이터가 늘어날수록 정확도가 향상되는 모습을 보였기 때

문에 학습 데이터의 비중에 의한 차이라고 생각됩니다.

② Decision Tree

One Rule에서와 마찬가지로 instances의 분류 능력이 가장 높은 속성으로 면허 취득 경과 년도가 선택되었습니다. 형성된 트리의 분류도를 표현해보자면 다음과 같습니다.

5년 미만 -> 경상(26개의 instances)
5년 이상 10년 미만 -> 낮 -> 경상(10개의 instances)
-> 밤 -> 중상(13개의 instances)
10년 이상 15년 미만 -> 경상(19개의 instances)
15년 이상 -> 차종 -> 승합차 -> 경상(10개의 instances)
-> 승용차 -> 중상(214개의 instances)
-> 화물차 -> 주야 -> 낮 -> 경상(18개의 instances)
-> 밤 -> 중상(18개의 instances)
불명(뺑소니 사고) -> 중상(24개의 instances)

면허 취득 경과 년도 외에 주야, 차종이 고려되기는 하였으나 214개나 되는 instances가 15년 이상, 승용차라는 속성 값에 의해 중상으로 분류되는 것을 볼 수 있습니다.
승용차는 시내에서 가장 흔한 차종이다 보니 instances 대부분이 승용차 사고에 해당된다는 점을 고려했을 때 15년 이상 경력의 운전자가 중상 이상의 상해를 발생시킨 사고를 주로 일으키는 것으로 나타났다고 볼 수 있습니다.
각 클래스에 대한 정밀도와 재현율은 다음과 같습니다.

| | 정밀도(Precision) | 재현율(Recall) |
|---|---|---|
| 경상 | 0.701 | 0.535 |
| 중상 | 0.812 | 0.907 |
| 사망 | 0(데이터가 너무 적음) | 0 |

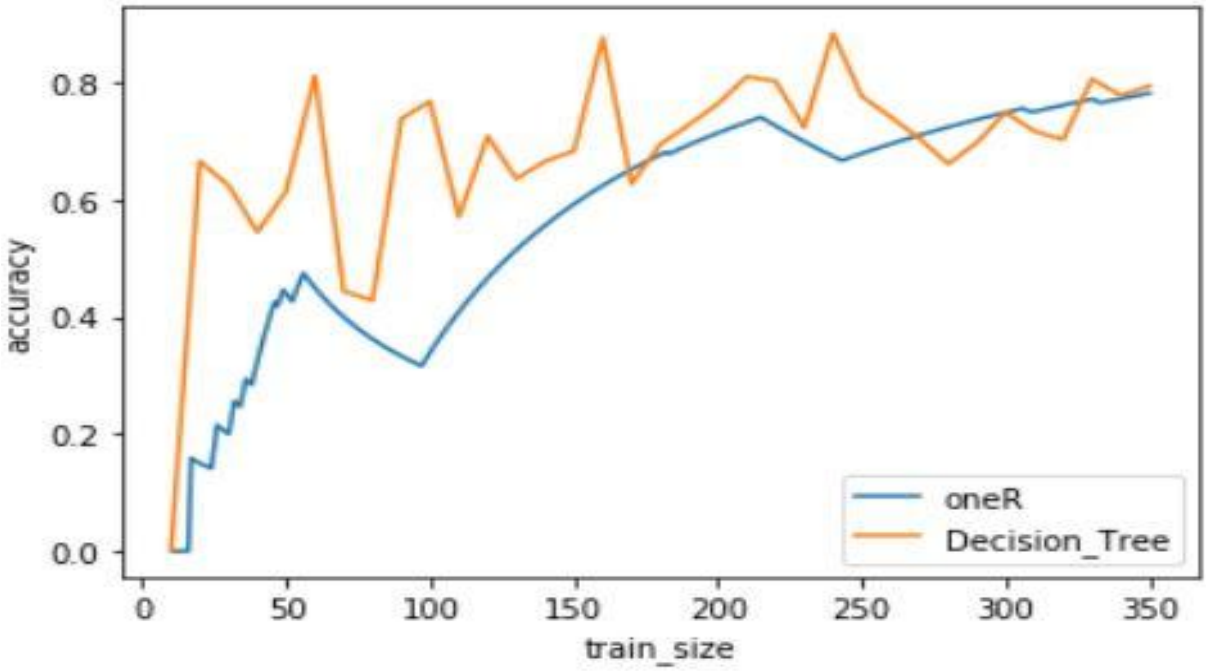
One Rule에서와 마찬가지로 중상 클래스에 비해 다른 클래스의 정밀도와 재현율이 떨어지는 모습을 보여줍니다. Confusion Matrix의 내용도 One Rule과 유사한데 경상을 중상으로 잘못 분류한 것이 47건 중상을 경상으로 잘못 분류한 것이 22건입니다.
정리하면, 실제 경상에 대해서 Decision Tree 모델이 중상으로 잘못 판단하여 경상에 대한 재현율이 떨어지고, 실제 중상 instance에 대해 모델이 경상으로 잘못 분류하여 경상 클래스의 정밀도도 함께 떨어졌다고 할 수 있습니다.
이 경우 역시 Data Set을 각 클래스 별로 균형 있게 수집하지 못해 발생한 문제라고 할 수 있겠습니다.

③ Association Rules

도출된 10개의 규칙에서 공통적으로 나타난 속성 값과 클래스 값이 있었습니다.
승용차, 중상, 15년 이상입니다.
즉, 15년 이상의 운전 경력을 가진 운전자가 모는 승용차로부터 중상 이상의 피해가 자주 발생했다는 것입니다.

도출된 규칙의 Lift 값은 모두 1이상인 1.35에서부터 1.4까지 나타났습니다.
Leverage 값의 평균을 5/345 = 0.014(속성의 개수를 instances 개수로 나눈 값)으로 잡았을 때 이보다 큰 값인 0.09에서 0.14 정도의 값이 도출되었습니다.
Convition 수치가 평균 20정도로 규칙이 성립하는 instances의 개수가 111~129 정도인 것을 감안했을 때 약 83% 확률로 규칙의 역도 성립한다는 결과가 나타났습니다..

-성능 분석(학습곡선, 베르누이 과정, ANOVA)




One Rule의 경우 354개의 instances를 학습했을 때 최고 성능 76%, Decision Tree의 경우 200개의 instances를 학습했을 때 최고 성능 90%를 보였습니다.
이를 토대로 베르누이 과정을 수행해보면 다음과 같습니다.

-Bernoulli Process

| | One Rule | Decision Tree |
|---|---|---|
| C = 80% | 0.729 < P < 0.787 | 0.872 < P < 0.922 |
| C = 90% | 0.725 P < 0.799 | 0.864 < P < 0.928 |

-ANOVA
통계적 유의성을 판단해보기 위해 ANOVA를 수행해본 결과는 다음과 같았습니다.

| | | | | | | |
|---|---|---|---|---|---|---|
| One Rule | 70% | 72% | 74% | 75% | 76% | |
| Decision T | 80% | 85% | 90% | 86% | 81% | |

분산 분석: 일원 배치법

요약표

| 인자의 수준 | 관측수 | 합 | 평균 | 분산 |
|---|---|---|---|---|
| Row 1 | 5 | 3.67 | 0.734 | 0.00058 |
| Row 2 | 5 | 4.22 | 0.844 | 0.00163 |

분산 분석

| 변동의 요인 | 제곱합 | 자유도 | 제곱 평균 | F 비 | P-값 | F 기각치 |
|---|---|---|---|---|---|---|
| 처리 | 0.03025 | 1 | 0.03025 | 27.37557 | 0.000791 | 5.317655 |
| 잔차 | 0.00884 | 8 | 0.001105 | | | |
| | | | | | | |
| 계 | 0.03909 | 9 | | | | |

최고 성능을 보인 지점 근처에서 5개의 성능 값을 추출하였고, F비의 값이 F 기각치 값보다 크므로 Null hypothesis를 기각할 수 있으므로 통계적 유의성이 있다고 판단할 수 있습니다.

3. 결론

결과적으로 면허 취득 후 15년 이상 지난 오랜 경력의 운전자가 중상 이상의 상해가 발생하는 교통사고의 주요 원인이라는 가설에 통계적 유의성이 있다고 판단되었다.
이러한 결과를 문제 개선에 적용하기 위해 생각해본 방안은 다음과 같습니다.

-운전 면허 적성 검사의 실효성을 의심해보아야 한다.
우리나라에서는 현행 상으로는 운전 면허 적성검사를 9년마다 한 번씩 받고 있습니다.
그리고 이 때 수행하는 내용은 시력, 신호등 색 구별 능력, 사지 운동능력으로 면허 소지자가 운전을 할 수 있는 신체적 요건을 확인하는데 그치고 있습니다.
우리나라의 교통사고 문제를 해결하기 위해서는 신체적 요건뿐만 아니라 정신적인 부분에 대한 개선이 필요하다고 생각합니다. 경력이 오래된 운전자들이 교통사고의 주범이 된 이유는 오랜 경력에서 나오는 안일함 때문입니다. 따라서 이번 프로젝트의 내용처럼 통계적인 근거를 들며 운전 면허 적성 검사 시에 운전자들의 경각심을 고취시키는 내용을 포함시키고 검사의 주기를 9년보다 짧게 가져가 경각심을 계속해서 상기시켜주는 것이 바람직할 것입니다.

4. 논의

-프로젝트의 한계점

고려한 속성에 과실 비율이나 위반 법규 내용 등을 포함시키지 못해 꼭 운전자만의 잘못인지 피해자의 과실도 있었는지 판단할 수가 없었습니다. 물론 15년 이상의 운전 경력자의 사고가 압도적으로 많은걸 봐서 어느 정도 유의성이 있다고는 생각됩니다만 상식적으로 생각해봤을 때 이번 프로젝트의 결론은 다소 허점이 있다고 생각합니다.
그리고 데이터의 양도 다소 통계적 유의성을 얻기에 부족하다고 사료됩니다. 데이터의 속성이 교통사고 하나에 대해서 주어지지 않고 속성별로 사고건수와 사상자가 주어지는 형태라 전처리를 프로그램의 도움 없이 해야했던 관계로 범위를 압축한 탓에 instances의 수가 너무 줄어들지 않았나 생각합니다.

-출처
Data Source: Taas(경찰청에서 제공하는 교통사고 공공데이터 플랫폼)
http://taas.koroad.or.kr/sta/acs/exs/typical.do?menuId=WEB_KMP_OVT_UAS_PDS#