1. 연구의 목적

본 연구는 데이터 마이닝 기법 가운데 회귀분석 모형과 의사결정나무 기법을 통해 인구당 폭력 범죄율에 영향을 미치는 주요 요인을 분석하고, 지역의 범죄율을 예측하는데 도움이 되는 최적의 모형을 선택하는데 목적이 있다. 범죄는 법적으로 정해진 사회의 규범을 어기는 행위로서, 물리적 손실뿐 아니라, 시민 사회에 불안감을 조성하기도 한다. 더욱이 본 연구에서 다루게 될 폭력 범죄는 타인에게 직접적 상해를 입히는 행위로서 사회의 안녕에 큰 위협을 준다. 따라서 범죄율에 영향을 미치는 요소를 적절히 파악하고 해당 요소로 지역별 인구당 범죄율을 예측한다면, 특정한 도시에 경찰병력을 늘린다던가, 행정적 조치 등을 취하여, 보다 안전한 사회를 만드는 방법을 구체화 할 수 있다.

또한, 범죄율은 사회학적, 심리학적, 경제학적인 요소 등을 비롯한 수많은 요소들의 영향을 받는다. 다행히 현대의 통계학에서는 인구센서스 등의 조사방법과 학문 및 기술의 발전을 토대로 이들 요소에 대한 자료를 수집하고 접근하는 것이 쉽게 여겨진다. 그러나 손 쉽게 얻을 수 있는 수많은 자료들은 때때로 통계학의 주요 목적으로 여겨지는 모수 추정 및 자료의 패턴 탐사를 방해하기도 한다. 따라서 자료가 과잉 될 때 어떠한 방법의 추정을 통하여 결론을 이끌어 낼 것인지를 조금이나마 일반화하는 것 역시, 본 연구의 목적이라 할 수 있다.

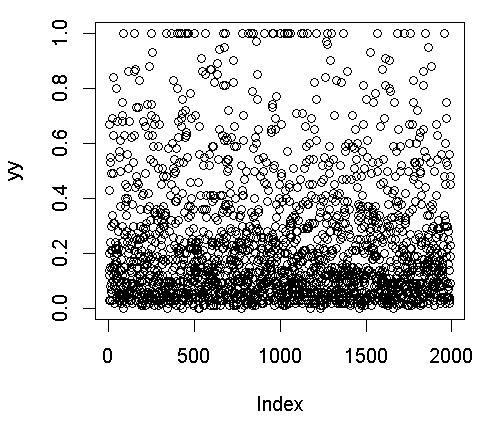
2. 사용한 자료(Data)

위와 같은 연구의 목적을 달성하기 위하여 범죄율에 영향을 미칠 것이라 예상되는 사회경제학적 요인을 한데 엮은 데이터를 이용하였다. 출처는 1990년 미국의 인구센서스, 1990년 미국 LEMAS(Law Enforcement Management and Administrative Statistic)의 설문조사, 1995 FBI UCR(Uniform Crime Reporting)이다.

(인터넷 출처: https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime)

총 1995개 community들에 대한 사회, 경제학적 요소의 관측치로 기입된 122개의 설명변수, 5개의 인덱스, 1개의 반응변수로 구성되어 있으며 설명변수와 반응변수 모두 비율이나 절대적 수치로 표현된 연속형 변수들이다. 또한 Missing data가 특정 변수들에 몰려 있었기 때문에 해당 변수를 제거하는 조치를 취했다. 이후 99개의 설명변수, 1개의 반응 변수가 남게 되었다. 자세한 변수의 내용은 해당 사이트를 참조하기 바란다. 결론에서는 유의미한 변수들만 설명하였다.

3. 방법론.

R 3.2.0 프로그램을 이용하여, 다중회귀분석, Shrinkage methods, Regression tree, Bagging, Cross validation등을 시행하였다. 분석을 처음 시행하고자 했을 때, 데이터에서 먼저 반응 변수 y의 분포를 확인했다. 반응 변수 y는 10만 명당 발생하는 폭력범죄의 횟수이다. 특정 community의 값들 1에 가깝게 높고, 나머지는 0을 중심으로 적당히 분산되어 있다. 분류(classification)를 시행할 경우 표본수가 비대칭적이라 추정이 잘 되지 않는다. 따라서, 연속형 변수들의 특징을 살려 회귀식(regression)을 중심으로 예측, 해석하기로 하였으며 특히 설명 변수가 많다는 점을 고려하였다.

1) 다중회귀분석: Least Square Estimation을 기반으로 모수인 베타를 추정하는 분석으로서 잔차에 대해 정규성, 등분산성, 독립성등의 가정을 가진다. 베타의 분포를 알아내어 테스트 하는 방식으로 변수를 선택한다. 그 외에 다양한 방법으로 변수 선택 및 관측치 제거를 할 수 있다. 회귀분석을 시행할 때 가장 기본이 되며 가정만 만족하면 해석과 예측이 상대적으로 간단하다.

2) Shrinkage methods

2-1) 능형회귀분석

변수들 간의 다중공선성이 크게 존재하여 제거하기 어렵거나, 변수의 개수가 많을 때 선택한다. 다중공선성은 회귀계수의 분산을 증가시키게 되므로, 이를 상쇄시키기 위하여 불편추정량(OLS)를 포기하고 MSE를 최소화하도록 만들어진 식이다. 이 때, 식은 일반 선형회귀에서 베타를 추정하는 공식에 람다를 포함한 제약식을 집어 넣어서 의미없는 변수들의 계수 베타가 0 근처로 수렴하도록 만든다.

2-2) LASSO

이 또한 능형회귀분석과 비슷한 목적으로 존재한다. 다만, 제약식이 절대값 꼴 이여서 쓸모 없는 변수의 계수가 능형회귀보다 더 빠르게 0으로 수렴한다는 것과, 수렴 값이 0이 될 수 있다는 점에서 능형회귀와 차이가 있다. 즉, 쓸모 없는 변수의 계수가 실제로 0에 수렴하게 되므로 해석이 보다 쉬워진다.

3) Tree-based models

3-1) Regression Tree

변수를 split point S(분류자)에서 나누어 고려한다. 평균과의 오차(해당 변수 내 변동량)의 합을 가장 작게 만드는 j와 s를 생각한다. 이런 식으로 가지를 뻗어나가는 기준 점을 정하게 된다. Tree가 너무 크게 뻗으면 과적합이 일어나므로 적절한 가지치기를 위해 페널티 α를 준다. (Tree가 너무 크면 과적합, 작으면 과소적합이 일어난다.)

3-2) Bagging

n개의 sample을 B번 복원 추출하여 Boot strap sample을 생성하고, 각각에 대해 모델을 세운다. 그 과정에서 생성된 각각의 단일 분류자들에 대해 평균-연속형 변수인 경우, 분류일 경우에는 다수결-을 내어 최종적인 분류자로 선택한다. 이렇게 하면 단일 모델의 불안정성을 감소 시킬 수 있다. 즉, 모델 간 분산을 감소시켜 과적합을 예방한다는 얘기다.

3-3) Random Forest

Bagging과 유사하지만, 평균이나 다수결로 단일 분류자를 결정하지 않는 점에서 다르다. 대신 각각의 단일 분류자들을 랜덤하게 추출하여 최종적인 분류자로 삼는다. 따라서 Bagging보다 연산이 단순하고 시간이 절약되는 효과가 있다. 정확도 또한 Bagging과 유사하거나 더 높을 수 있다.

4) K-fold Cross validation

이용한 모든 모형에서 mse를 구하거나 다른 추정량을 구할 때는 K-fold CV를 이용하였다. 동일하게 나눠진 K개의 데이터 샘플 가운데, K-1개의 샘플로 추정한 모델을 이용해 나머지 1개의 test mse를 추정한다. 이 과정을 반복하여 K개의 test mse(혹은 다른 추정량)의 평균을 구하면 된다.

4. 분석 결과

모든 모형이 inference가 가능한 만큼, 우선 낮은 Test mse와 모형의 적합성을 중심으로 가장 좋은 모델을 선정하고, 선정된 최종 모형에 대해서만 해석할 것이다.

1) 다중회귀분석

Vif로 변수간 다중공선성을 확인하고 변수선택을 실시하여 모델을 얻어냈다. 변수간 다중공선성은 분산을 팽창시키는 효과를 불러온다. 또한 유의미하지 않은 계수를 제거해야 좋은 추정이 된다.

10-Fold CV를 이용한 식:

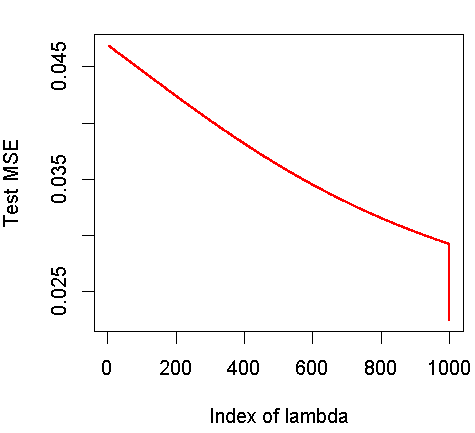
Y = 0.39713 - 0.20881\*X9 - 0.20337\*X16 - 0.14479\*X19 +0.03737\*X30 + 0.11808\*X39 -0.06033\*X40 - 0.08672\*X51 - 0.09438\*X52 + 0.28876\*X56 +0.06790\*X75 + 0.26707\*X77 + 0.09000\*X80 - 0.06631\*X81 + 0.33320\*X82 + 0.04400\*X83 + 0.06561\*X91 - 0.08862\*X94 + 0.14459\*X95 + 0.04485\*X97 + 0.06080 \*X100 + X126\*0.04249

Test MSE: 0.01965595 / Adjusted R Square: 0.6460884

R-square 값도 꽤 높고, 변수의 개수도 적절히 남았다. 다만, 등분산성 검정을 만족하지 않는 모습을 보였다. 이를 만족시키려 y에 대한 변수변환을 시행해보았다. 여러 변수변환을 시도해보았지만, 가장 단순하게는 y^(1/10)으로 추정해야 겨우 등분산성을 만족하며, 해당 모델에서는 Test mse가 급격히 상승하는 결과를 불러왔다. 따라서 다중회귀식을 사용하여 예측, 해석하기에는 어려움이 많다.

Y^(1/10)일 때의 Test MSE: 0.05400579

2) Ridge / Lasso

앞서 말했듯, 설명변수 99개 중에서 다중공선성을 가진 변수가 65개였다. 따라서 Ridge와 Lasso를 활용하는 것이 적절하다고 판단, 각각의 모델에서 람다를 추정해보았다.

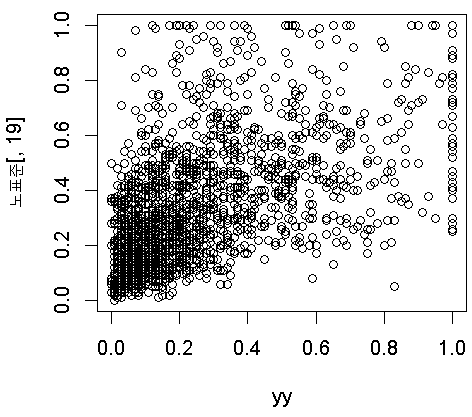
그런데, CV를 이용하고, 1000과 0, 10과 0, 0.1과 0사이, 0.01과 0사이 등을 grid로 잡아 추정했음에도 불구하고, Ridge와 Lasso 모두 Test MSE는 람다가 0일 때 가장 낮았다. 이는 ‘제약식이 없을 때’ 모델의 예측력이 가장 높다는 뜻이다.

람다가 0인 최종 모델에서 CV를 이용해 도출한 최종 mse는 다음과 같다.

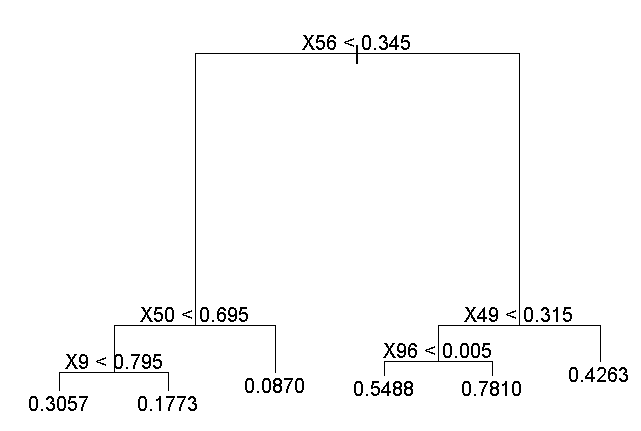
Ridge regression의 test MSE: 0.05430644

LASSO의 test MSE: 0.05430644

제약식이 없기 때문에 두 MSE모두 동일한 값이 나왔다. 또한 제약식이 없어 모든 설명변수를 고려하여 식을 짠 것이기 때문에, 100개의 설명변수 모두에 대하여 inference를 가져서 해석이 어렵다는 단점이 있다.

1과 2의 분석결과에서 다중공선성이 변수간에 많이 존재한다는 것을 알 수 있었고, 이를 해결하기 위해 제약식을 두어 설명변수의 개수를 줄이는 모델은 의외로 무의미한 것을 확인하였다. 즉, 변수들 간의 다중공선성을 고려하는 것은 모델의 분산을 줄이는데 큰 도움이 되지 않았다. 따라서 모델의 분산이 커지는 다른 이유를 찾아야 하는데, y(예측변수:폭력범죄율)과 설명변수들 간의 plot을 그려보면, 대체로 왼쪽의 그림처럼 나타났다. 즉, 각 설명변수들 자체의 분산이 크다는 것을 알 수 있다. 이는 관측치가 많을 때 보통 나타나는 현상이며, 변수 내에서의 변동성을 잘 설명해주는 flexible한 모델이 필요하다는 것을 시사한다.

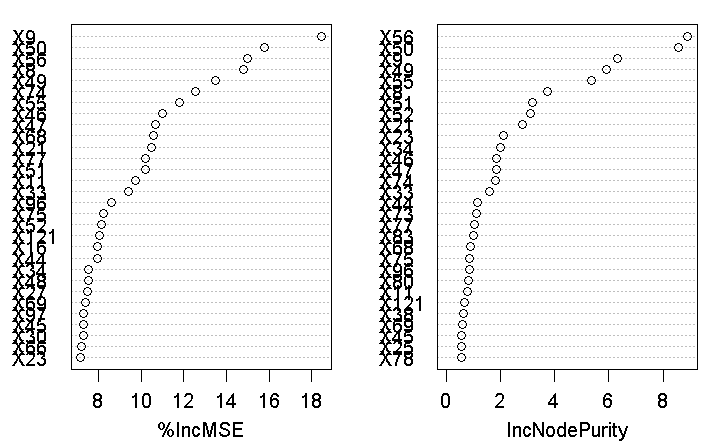
따라서 이 전에는 설명변수의 개수가 많을 때라는 점을 먼저 고려하여 inflexible하게 접근하였지만, 지금부터는 관측치의 분산이 설명변수의 개수를 압도할 만큼 크다는 점을 고려하여 flexible하게 접근할 것이다.

3) Tree-based tools

3-1) Regression Tree

우선 CV를 이용하여 적절한 Tree size와 제약식에 쓰이는 알파값을 도출해보았다. Size는 6, alpha=0.7625029 로 추정되었다. 이것을 이용하여 가지치기(pruning)을 수행한 결과에서 평균적인 Test MSE를 구했다.

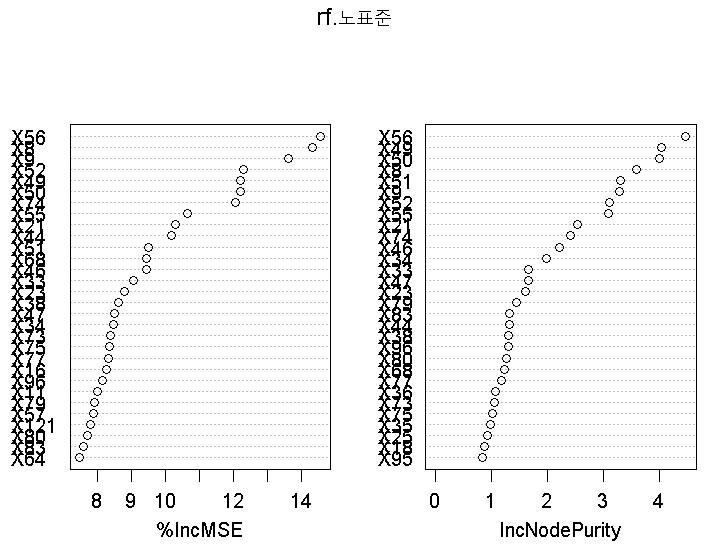
Test MSE: 0.008729434



3-2) Bagging

Regression Tree에 Bagging을 적용하여 Importance를 확인한 결과, X9, X50, X56, X49 등이 중요한 요소로 파악되었다. 또한 CV를 이용하여 Test MSE를 구하였고, 이 과정에서 Bagging은 연산이 오래 걸린다는 것을 체감하였다.

Test MSE: 0.006646991

3-3) Random Forest

Regression Tree에 Random Forest를 적용하여 Importance를 확인한 결과, X8, X9, X49, X50, X56 등이 중요한 요소로 파악되었다. 또한 CV를 이용하여 Test MSE를 구하였고, Random forest의 연산은 Bagging보다 비교적 짧게 이루어졌다.

Test MSE: 0.01845646

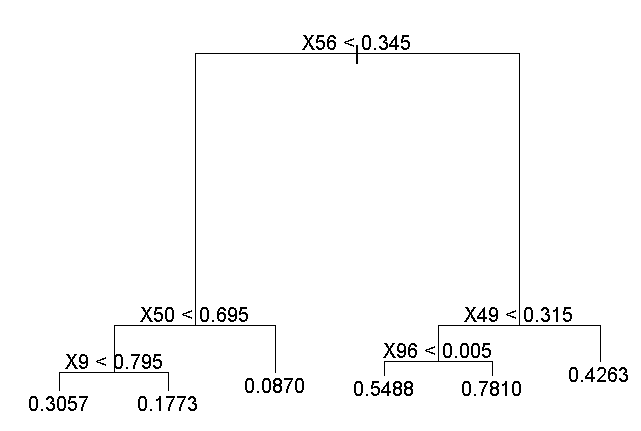
5. 결과에 대한 해석 및 결론

1) 모델 간 비교 및 최종 모형 선택

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 모델 | TEST MSE | 장점 | 단점 | 비고 |
| 다중회귀  (+변수변환) | 0.01965595  (0.05400579) | 해석과 예측을 쉽게 할 수 있다. | 잔차 가정이 까다롭다. | 변수 선택법이 다양하여 다소 자의적. |
| Ridge/LASSO | 0.05430644 | 다중공선성 제거  변수 축소(LASSO) | 람다가 충분히 크지 않으면 해석이 까다로워질 수 있다. | 제약식이 무의미할 때는, Ridge와 LASSO모델의 두 식이 같은 값을 가진다. |
| Regression Tree | 0.008729434 | 직관적 결과도출. | 일반화 하기 어렵다. Bagging은 연산이 오래 걸림. |  |
| + Bagging | 0.006646991 | 모델 안정성확보. |  |
| + Random Forest | 0.01845646 | 연산 속도향상 |  |

위의 표에서 확인 할 수 있듯이, Regression Tree와 Bagging의 Test MSE가 가장 낮다. 그런데 Tree regression은 결과값 자체를 일반화하기는 어렵다. 따라서 모델 자체를 일반화 시키는 것이 더 적절하다. 다시 말해 지금의 분석과 비슷한 요소를 통해 다른 지역의 범죄율을 추정하려 할 때나, 관측치와 변수 모두 많은 상황이고 관측치 내의 분산이 클 때 등의 상황에서 Regression Tree가 효율적이라는 것을 일반화 시키는 것이 더 적절하다. 따라서 연산 속도 또한 모델 선택에 중요한 것이라 생각한다. 즉, Bagging이 일반 Regression Tree보다 Test MSE는 약간 더 좋지만, Bagging을 쓰지 않고 일반적으로 가지치기한 Regression Tree가 연산도 더 빠르고 충분히 좋은 모형이므로, 이를 최종 모형으로 선택하겠다.

물론 결과 값을 미국 내 다른 community들의 범죄율 예측에 일반화하는 것은 가능하다. 1995개의 community 샘플은 결코 적지 않으며, 미국 각 지역-최종으로 이용한 데이터에서 46개 주에서 추출된 것을 확인-에서 추출되었기 때문이다. 따라서 간략히 결과 값을 해석해 보자면,

X9 : percentage of population that is caucasian

X49 : percentage of families (with kids) that are headed by two parents   
X50 : PctKids2Par: percentage of kids in family housing with two parents

X56 : percentage of kids born to never married

X96 : number of homeless people counted in the street

우선, ‘한번도 결혼하지 않은 가정(혹은 부모)에서 태어난 아동의 백분율.’(X56)이 폭력 범죄율에 큰 상관관계를 미치는 것을 알 수 있다. 해당 지역에서 이 값이 0.345 미만이면 상대적으로 낮은 폭력 범죄율, 0.345 이상이면 상대적으로 높은 폭력 범죄율이 나타난다. 안정적인 가정의 유무와 일탈이 상관관계를 가지는 것으로 해석된다.

다음 가지로 내려오면, 두 부모들에게서 길러지는 아이들이 있는 가정의 비율(X49)과 두 부모에게서 길러지는 아이들의 비율(X50)이라는 다소 비슷한 변수가 폭력 범죄와 상관관계를 가진다는 것을 알 수 있다.

그런데 통상적인 인식과 반대로 이는 음의 상관관계를 가지는 것으로 보인다. 통상적으로 사회에는 두 부모 가정(이혼 가정 등)의 비율이 일탈의 비율과 양의 상관관계를 가지고 있다는 인식이 있다. 그러나 이번 Regression Tree의 결과에서는, X56이 0.345미만 일 때, X50이 0.695 미만이면 상대적으로 높은 폭력 범죄율 0.695 이상이면 상대적으로 낮은 폭력 범죄율을 보이며, X56이 0.345 이상일 때, X49가 0.315미만이면 상대적으로 높은 폭력 범죄율, 0.315 이상이면 상대적으로 낮은 폭력 범죄율을 보이기 때문이다. 즉, 두 부모 가정의 비율과 일탈의 비율이 반비례하는 결과가 나온 것이다.

마지막으로 세 번째 가지의 기준점 중 X9를 살펴보면 X56이 0.345미만이고 X50이 0.695 미만일때, 코카시안의 비율(X9)가 0.795 미만이면 상대적으로 높은 폭력 범죄율 0.795 이상이면 상대적으로 낮은 폭력 범죄율을 보인다. 코카시안은 인류학적으로 백인으로 분류되는 인종으로서 우리가 통상적으로 알고 있는 백인 외에 다양한 인종을 포함한다. 따라서 특정 피부색을 가진 인종이 직접적인 영향을 주었다고 해석하기는 어렵다. 다만, 코카시안 인종에 해당하는 사람들(통상적 백인, 아랍인, 인도인 등)이 미국에서 모여 사는 곳이라면 상대적으로 일탈에 있어서 보수적인 관념을 가진 곳이리라 사려된다. 또 다른 기준점인 X96을 살펴보면 X56이 0.345이상이고 X49가 0.315 미만일 때, (10만 명당) 거리 노숙자의 수(X96)가 0.05미만이면 상대적으로 낮은 폭력 범죄율, 0.05 이상이면 상대적으로 높은 폭력 범죄율이 나타난다. 즉, 거리 노숙자의 수는 폭력 범죄율과 양의 상관관계가 있다.

아주 많은 설명 변수들 가운데서 폭력 범죄율에 영향을 미치는 주요 변수를 살펴보았다. 특히 고려된 5개의 변수 가운데 4개가 가정적 요인인 것을 보아 가정에서의 교육이나 환경이 일탈과 큰 상관관계가 있음을 알 수 있다. 따라서 안정적 가정이 없는 아이들이나 사회구성원들이 일탈로부터 멀어질 수 있게끔 하는 사회적, 행정적 지원이 필요하다.

또한, 두 번째 가지들의 기준점이 되는 두 부모들에게서 길러지는 아이들이 있는 가정의 비율(X49)과 두 부모에게서 길러지는 아이들의 비율(X50)이 폭력 범죄율과 음의 상관관계를 가진다는 것은 사회적 인식과 반대의 결과를 보여준다. 이 결과를 더 자세히 알려면 두 부모에게서 길러지는 가정의 아이들이 자기검열을 통해 일탈을 적게 한다거나, 한 부모 가정에서도 가정 폭력 등에 노출될 수 있다는 점등 여러 가지 이유를 고려하여 추가 연구를 진행하는 것이 좋겠다.

2) 그 외에 알 수 있었던 것

데이터 마이닝은 최적의 모델을 찾는 과정이기에 관측치나 변수의 개수에 따라 flexibility가 높거나 낮은 모델을 사용하는 것이 굉장히 중요하다. 특히 ‘얼마나 많아야 많은 것인가’의 기준점을 잡기는 자료들마다 애매모호하다. 위의 연구에서 시행착오를 거친 결과, 이론과 동일하게 Flexibility는 Ridge/LASSO < 일반적 Regression < Tree 임을 알 수 있었고 관측치가 많을 때, 설명변수의 개수를 상대적으로 고려하여 모델의 flexibility를 정하는 것이 중요함을 알게 되었다. 설명변수가 많아서 다중공선성이 일어나더라도, ‘관측치 자체의 분산’이 ‘다중공선성이 분산을 늘리는 효과’보다 더 커다란 분산 효과를 가질 수 있기 때문이다.

3) 한계점

Bagging을 이용한 Tree에 관한 것을 자세히 알지 못한 것도 일반적인 Regression Tree 모형을 선택한 이유 중 하나여서 아쉽다. 또한, Missing Data를 처리하는 방법을 알지 못하여 일어난 정보의 손실이 있었다. 특정 변수에서만 대부분의 값이 없었던 이유로, 특정 변수들을 제거할 수 밖에 없었다.

4) 또 다른 regression model들을 사용하지 않은 이유

KNN과 PCR등의 모델이 있지만 KNN Regression은 변수가 다양한 모델에서는 Curse of Dimension의 문제를 가지고 있고, PCR regression은 설명이 매우 어렵다는 단점이 있었다. 범죄율의 예측값만을 구하여 범죄를 줄이는 것은 영화 속에서나 가능한 일이기에 inference 또한 중요하다고 생각하였다. 번외로 PCR regression을 시도해보니 Test MSE가 0.02275795로 어느 정도 낮게 나왔으나 Regression Tree보다는 높은 값을 가졌다.

<읽어주셔서 감사합니다.>