|  |  |  |  |
| --- | --- | --- | --- |
| 모델 | 장점 | 단점 | 특징 |
| Decision Tree  (의사결정나무) | 1. 결과해석 용이 2. 직관적인 해석 가능 3. 주요 변수와의 분리기준 제시 4. 비모수 모델 5. 전통적인 통계모델에 요구되는 가정에 자유로움 6. 이상치 및 결측치에 대하여 전통적인 모델보다 강건한 성능을 보인다.   결과 해석의 직관성 + 비모수, 이상치 등에 강한 성능 | 1. 비안정성 2. 데이터 수가 적을 시 전통적인 통계 모형에 비하여 불안정 3. 과대 적합에 취약함 4. 선형적인 표현 미흡 5. 전체적인 선형관계의 파악이 어려움 6. 비연속성 7. 분리 시 연속형 범주를 구간화 처리 (비연속화) 8. 분리 경계점 근처에 오류 발생 가능   Overfitting, 소수 데이터에 취약하며 Data를 나누는 Cut-off 지점의 1종, 2종 오류 발생 | 1. Decision Tree 내 노드의 분리 기준 2. Gini-index: 지니지수란 불순도 지수(얼마나 다양한 데이터가 잘 섞여있는지에 대한 정도)  * 지니지수가 낮을수록 순도가 높은 것 (분리가 잘 됨)  1. Entropy-index: 엔트로피 또한 데이터가 섞여있는 정도를 나타냄.  * 엔트로피가 낮을수록 분리가 잘 되었음을 의미.  1. Decision Tree 내 변수별 중요도 계산 2. Information Gain: 데이터를 분리할 때 특정 노드 이전과 이후에 엔트로피 차이를 측정하는 척도 3. Information Gain이 높은 변수는 데이터 분리에 있어서 중요한 변수라는 것을 의미.   Gini-index 기준으로 노드 형성  정보획득률로 Feature importance 도출 |
| Prunning | 1. 파라미터의 개수를 조절함에 따라서 추론 속도가 빨라진다. 2. 모델 구조를 간소화(Regularization) 함으로써 일반화 성능을 높일 수 있다.   Decision Tree 내 노드, 깊이를 조절하여 빠른 계산이 가능하며 Overfitting에 강하다. | 1. 정보의 손실이 발생한다. 2. 데이터의 수가 적으면 필요 이상으로 가지치기를 진행할 가능성이 존재함.(Over Pruning)   정보 손실이 발생한다.  Underfitting에 약하다. | 1. 지나친 세분화를 막기 위하여 특정 노드 이후를 잘라내는 방법. 2. 학습의 결과를 검증하기 위한 별도의 검증 데이터를 이용한다.   깊이와 노드를 조절하여 지나친 세분화를 막아 Overfitting을 막아줄 수 있다. |
| Random Forest | 1. 앙상블 기법을 사용하기 때문에(다수의 트리를 생성한 후 투표를 실시하여 가장 많이 득표한 모델을 선정하기 때문에) 과대 적합을 예장함  * 일반화 적용  1. 파라미터 조정이 용이하다 2. 데이터 Scale 변환이 필요하지 않다.   랜덤하게 Decision Tree를 형성하고 Voting을 사용하여 최종 모델을 선택한다. | 1. 개별 트리 분석이 어렵고 트리 분리가 복잡해지는 경향이 존재함 2. 차원이 큰 텍스트와 같은 정형화 되지 않은 데이터에 대한 성능이 미흡하다. 3. 훈련 시 메모리 소모가 크다   개별 트리에 대한 분석이 어렵고 훈련시 메모리의 소모가 크다. 또한, 비정형 데이터에 약하다. | 1. 임의성: 서로 조금씩 다른 특성의 트리들로 구성 (각 트리별 무작위성 주입) 2. 비상관화: 각 트리들의 예측이 서로 연관되지 않음 3. 견고성: 오류가 전파되지 않아 노이즈에 강함 4. 일반화: 과적합 문제 극복   Random한 트리를 형성하기에 Feature가 무작위로 주입된다. 따라서 각 트리의 예측이 비상관성을 가지며 Overfitting에도 강함 |

Train, Test 방법에 따른 성능 차이 : 마지막에