금융시장의빅데이터분석

20180071 오서영

1. bias는 예측값과 실제값의 차이를 의미하고, variance는 입력에 따른 예측값의 변동성을 의미한다. 머신러닝의 결과 bias가 큰 경우, 예측값과 실제값의 차이가 크다는 것이며, underfitting이 발생하고 있음을 의미한다. 반면 variance가 큰 경우, 입력에 따른 예측값의 변동성이 크다는 것으로, overfitting이 발생하고 있음을 의미한다.

2. ACF란 AutoCorrelation Function의 약자로 특정 시차로 구분된 시계열의 현재 값과 과거값의 상관계수함수를 의미한다. PACF란 Partial AutoCorrelation Function의 약자로 ACF와 마찬가지로 시계열의 관측치간 상관관계지만, 시차가 다른 두 시계열 간의 순수한 상호 연관성을 의미한다.

3. Overfitting이란 과적합으로서 머신러닝에서 모델이 훈련 데이터에 과하게 학습이 되어 훈련 데이터에서만 좋은 성능을 보이고 테스트 데이터에서는 나쁜 성능을 보이는 것을 말한다.

4. random walk란 이전 시점 관측값에 랜덤한 노이즈 값이 추가되어 다음 시점의 관측값이 정의되는 경우를 의미한다.

5. OLS, Lasso, Ridge 모두 선형회귀모델이다. 하지만 Lasso와 Ridge는 제약조건이 추가되어있다. Lasso는 가중치의 절대값의 합을 최소화하는 것을 제약조건으로 두며, Ridge는 가중치들의 제곱합을 최소화하는 것을 제약조건으로 한다. Ridge와 Lasso 모두 OLS보다 제약조건 통제를 통해 과적합을 방지할 수 있다는 장점이 있다.

6. X, Y의 p value가 0.05 초과이므로 이는 non-stationary한 불안정 시계열이라는 것을 의미한다. 이러한 불안정 시계열의 문제점은 시간에 따라 추세나 계절성이 바뀌면 관측값의 의미도 바뀌어서 모델에 넣었을 때 정확한 예측값을 도출하기 어려워진다는 것이다. 이러한 불안정 문제는 차분이나 변화율화를 통해 안정 시계열 데이터로 변환함으로써 해결할 수 있다.

7. 횡단면 자료는 하나의 시점에 여러 변수를 관측한 자료이고, 시계열 자료는 하나의 변수를 여러 시점에서 관측한 자료이다. 횡단면 자료는 하나의 시점에 여러 변수의 자료를 랜덤 샘플링으로 구하기 때문에 변수 간 표본 선택에 문제가 생길 수 있고, 시계열 자료는 여러 시점에 관측한 자료이기 때문에 추세, 계절성, 자기상관 등의 문제가 발생할 수 있다.

8. 의사결정나무의 가장 큰 단점은 훈련데이터셋에 맞춰 과하게 복잡한 트리를 구현하는 과적합이 발생하기 쉽다는 것이다. 이러한 과적합을 해결하기 위해서는 max\_depth, min\_sample\_leafs 등의 하이퍼파라미터를 조절하여 모델에 restriction을 주는 것이다.

9. 카드사 입장에서는 정상사용자가 잘못 색출되는 것을 방지하는 것보다 부정사용자를 색출해내는 것이 중요하므로 recall이 중요하다. recall이란 실제 부정사용자 중에서 모델이 부정사용자라고 예측한 비율이기 때문이다. 반면 고객의 입장에서는 부정사용자가 색출되는 것보다 본인이 부정사용자라고 잘못 선정되지 않아야 하므로 precision이 중요하다. precision이란 모델이 부정사용자라고 분류한 것중에서 실제 부정사용자인 사람의 비율이기 때문이다.

10. 모델링시 데이터의 feature가 많은 경우 시간과 메모리상에서 cost가 크며, 유의미하지 않은 변수에 의해 overfitting되는 결과를 초래할 수 있다. 그러므로 종속변수의 예측에 유의미한 변수만을 feature selection하여 모델에 넣어주면 이러한 문제점을 해소할 수 있다.