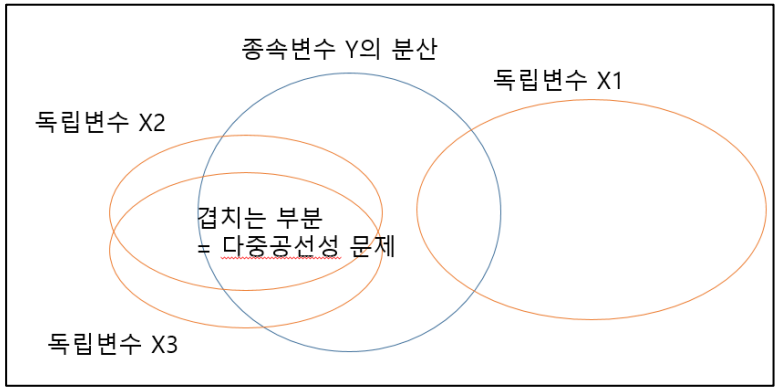
1) 전진 선택법: 모델 돌리기 전에 변수를 선택하는 방법

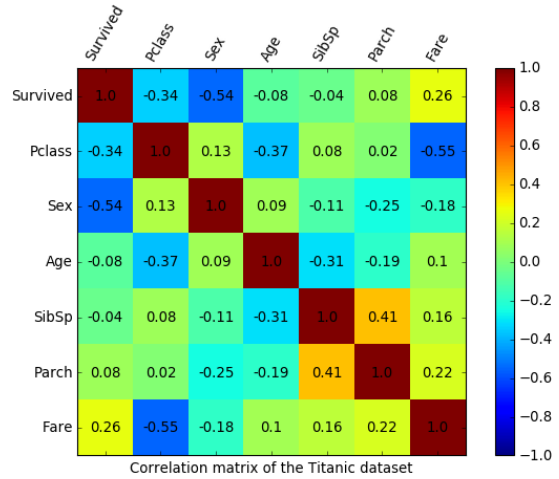
- 다중공선성 문제: 독립변수들 간에 강한 상관관계가 나타나는 심각한 회귀 방해 문제 (단 Tree 모델들은 이에 강하다.)

※ 분산팽창계수 VIF(Variance Inflation Factor)를 사용하여 다중공선성 검토 (VIF >10 제거)



- Correlation 분석 (상관계수 r ≥ abs(0.25))

※ 뒤에 추가 종류 소개할 예정



Correlation 분석 종류:

1. 연속형-연속형

1-1) Pearson correlation (노말한 연속형 vs 연속형)

예) 국어점수와 영어점수간의 상관관계

1-2) Kendall correlation (순위척도 자료형 끼리의 상관관계)

- 순위척도 자료형에 대한 상관계수 추론 방법
- 두 자료간의 상관도 정도 보다는 한개 자료의 변수가 증가할 때 다른 자료의 변수가 증가하는지 감소하는지 정도를 보는 척도

예) 국어등수와 영어등수간의 관계

1-3) Spearman correlation (순위척도 자료형 끼리의 상관관계)

- kendall과 순위척도 자료간 상관관계를 파악.
- 한 변수가 증가할때 다른 변수가 증가하는지 감소하는지 정도만 봄

예) 국어등수와 영어등수간의 관계

2. 범주형-범주형

Phi correlation (Binary 카테고리컬 vs 카테고리컬)

- 비교대상 범주 대상이 2개

예) 남/여 , O/X

Cramer's V (3개 이상의 카테고리컬 vs 카테고리컬)

- 비교대상 범주 대상이 3개이상

예) 10대/20대/30대, 단독/연립/복합/아파트

3. 연속형-범주형

Point biserial correlation

- 두개변수중 하나는 범주형 변수이고 다른 하나는 연속형 변수일 때

예) 성별과 수학점수와의 상관관계

Biserial correlation

- 두개 변수 중 하나는 명명척도이고 다른 하나는 연속변수일 때
- 명명척도의 유목은 인위적 구분하는 이분변수

예) 우열반 편성여부와 중간고사 점수와의 상관관계

Polyserial correlation

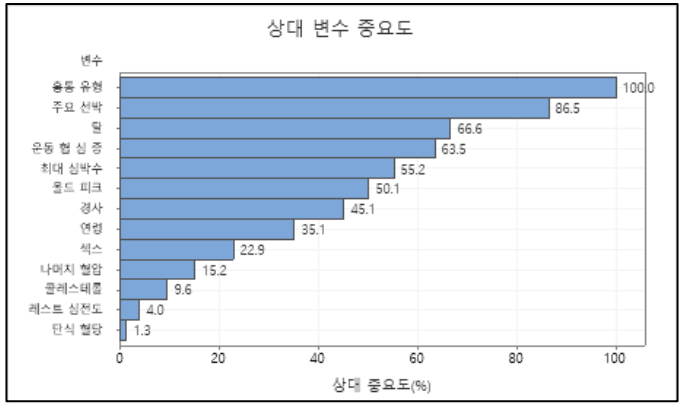
- 두개의 변수중 하나는 명명척도이고 다른 하나는 연속변수
- 명명척도의 유목은 비인위적이며 3개 이상의 유목

예) 인종과 키와의 상관관계

	Categorical	Continuous
Categorical	Lambda, Corrected Cramer's V	Point Biserial, Logistic Regression
Continuous	Point Biserial, Logistic Regression	Spearman, Kendall, Pearson

2) 후진 소거법 – 모델 돌리기 전에 변수를 선택하는 방법

- 변수중요도 (Feature importance): 모델에서의 변수 별 영향 줬던 정도



- Recursive 방법 : 모델을 돌려보면서 변수를 선별

RFE (Recursive Feature Elimination)
RFECV (Recursive Feature Elimination with Cross Validation)

- AIC (Akaike Information Criterion) 방법: 모델 품질을 검토(But, 랜덤포레스트 이후 모델을은 CV가 효과적)

3) 단계적 선택법 – 모델에 전진선택법과 후진소거법을 번갈아가면서 돌리는 것.