O 사전 세팅

1. [워드] 단락 뒤 공백 제거
2. 캡처도구 켜두기
3. C:\windows\fonts에서 맑은 고딕 파일(malgun.ttf) Jupyter Notebook으로 옮겨두기  
   (fonts 폴더에서 ‘맑은’ 검색하면 됨)
4. 미리 코드 쳐두기(라이브러리 등)
5. 문제 파일 확인

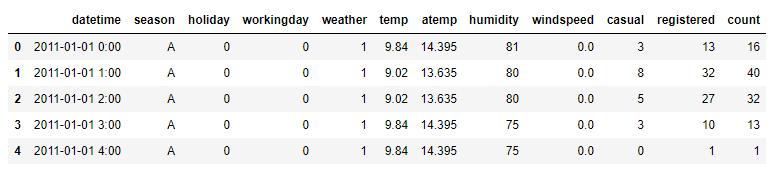
O 서식

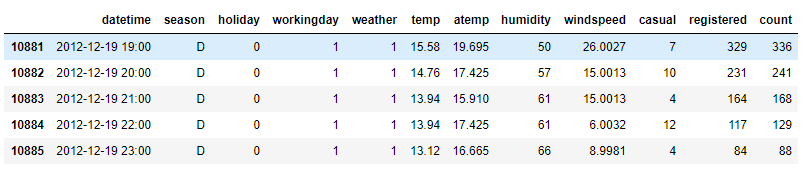
1. **대분류(Bold, 14pt)**
   1. 중분류
      1. 소분류

**2021. 3. 22**

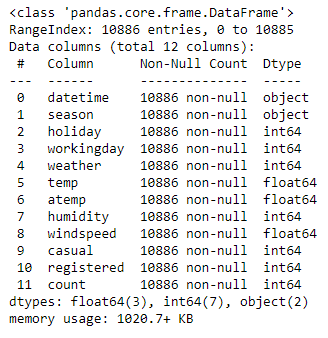
**1. 기계학습 1.1 분류**

1. **데이터 구조 확인 및 EDA**
   1. head, tail을 통해 데이터 처음과 마지막 5행 확인

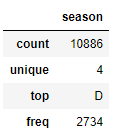
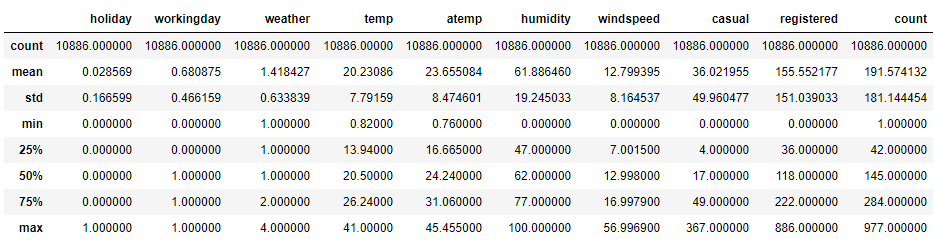




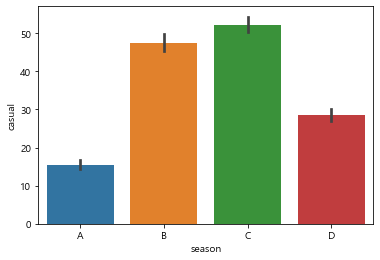
* 1. shape로 데이터 구조 확인 : 10,886행 x 12열
  2. info로 컬럼별 데이터 타입 확인



* 1. describe로 빈도와 분포 확인



* 1. 시각화를 통한 데이터 구조 확인  
     (전체 컬럼 중 몇 가지 컬럼만 선별하여 첨부함)
     1. Season class에 따라 casual 평균 값에 차이가 있음을 확인했으며, season이 C일 경우 casual의 평균은 50이 넘어 전체 season 중 가장 높았으며, 반대로 A의 경우 약 15로 가장 낮았음



* + 1. Season별 빈도는 모두 2500 정도로 고르게 분포하고 있음



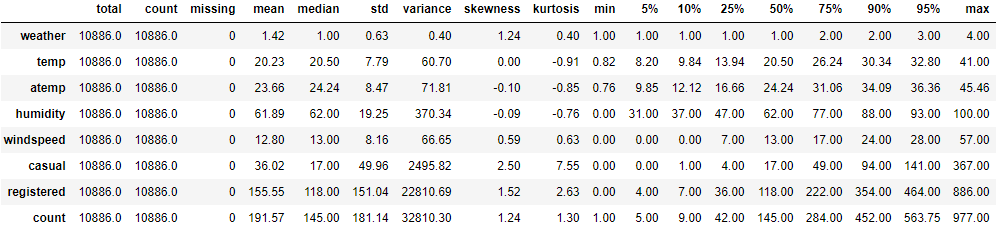
1. **데이터 마트 DQ check 및 변수 선택**

🞏 연속형 변수

⭘ 연속형 변수들의 기초통계량과 분포분석을 진행함

- 분포가 90% 이상이 동일한 변수는 모델링 시 설명력이 떨어진다고 판단하여 제외하기로 함

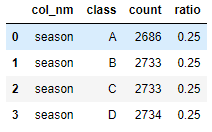
- 이 데이터에서는 모든 변수들이 고른 분포의 데이터를 가지고 있음



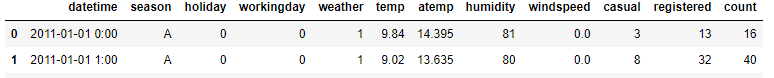
🞏 범주형 변수

⭘ 범주형 변수들의 빈도분석과 분포분석을 진행함

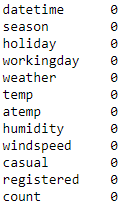
- 이 데이터에서는 모든 변수들이 고른 분포의 데이터를 가지고 있음



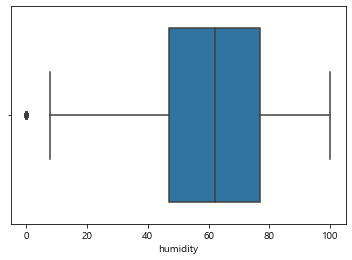
1. **데이터 전처리**
   1. 중복값 처리
      1. ~~ 사유로 중복이 발생한 것으로 생각됨

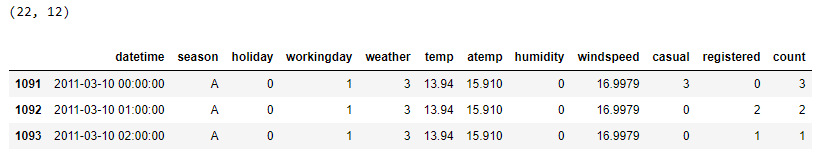


* + 1. ~~ 하여 중복값은 마지막 데이터만 남기기로 함
  1. 결측값 처리
     1. ~~와 @@에 결측값이 n개 있음을 확인함



* + 1. ~~ 사유로 결측값은 0으로 대체함 / 결측값이 있는 데이터는 제거함 / 중앙값
  1. 이상치 처리
     1. 각 컬럼별로 boxplot을 그려 IQR 방식으로 이상치를 확인함
     2. 본 보고서에는 일부 변수에 대해서만 plotting함
        1. humidity 변수에서는 총 22개의 이상치를 확인함





* + 1. 데이터를 최대한 활용하기 위해 min/max값 경계 밖의 데이터를 min/max로 보정함 / 제거함

1. **파생변수 생성**

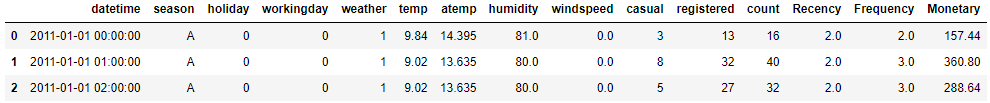
🞏 RFM의 요인을 기준으로 파생변수를 생성함 // 아니면 datetime에서 연월일, 시간대 등

⭘ Recency(최근성) : datetime이 3000일 이상이면 1, 2000일 이상면 2, 그 외는 3

⭘ Frequency(빈도) : ID별 거래 횟수(ID의 count가 10 이내면 1, 20이내면 2, 그 외는 3

⭘ Monetary(거래규모) : count와 temp의 곱(교호작용의 효과를 확인하고자 한다)

⭘ 회당 값 : ~~을 파악하기 위해 temp를 count로 나눔



1. **종속변수 전처리(n개 클래스로 변환)**

🞏 종속변수의 클래스를 4개 클래스의 범주형으로 변환함

⭘ count <= 150 : 1

⭘ 150< count <= 300 : 2

⭘ 300< count <= 450 : 3

⭘ count > 450 : 4

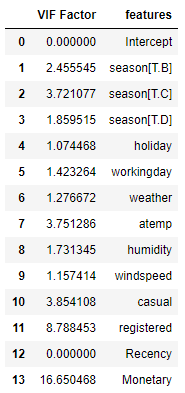
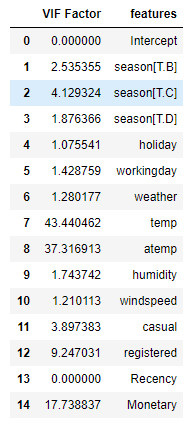
1. **변수 더미화**
   1. 특정 알고리즘을 활용한 모델링 과정에서의 에러 발생을 방지하기 위해 범주형 변수에 대해 더미화를 진행하였음
2. **다중공선성 제거**

🞏 다중공선성은 예측변수 사이에 높은 상관관계가 있을 때 발생하며, 회귀계수 추정치의 신뢰성과 안정성에 문제를 발생시키므로 VIF(분산팽장계수)를 활용하여 사전 제거함

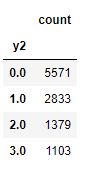
⭘ temp와 atemp의 VIF Factor값이 각각 43.4와 37.3으로 다중공선성 발생이 관찰되어 둘 중 temp 변수를 제거함

⭘ temp 제거 후 atemp의 VIF Factor값도 3.8로 안정화 된 것을 관찰함

⭘ Monetary변수의 VIF Factor값은 16.7로 10을 초과하나 데이터의 특성을 잘 반영하고 있으므로 유지하여 모델에 사용하기로 함



1. **종속변수 분포 확인**



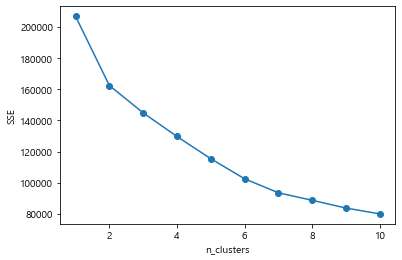
1. **Train, Test set 분리**
   1. 데이터 마트를 Train, Test set 7:3 비율로 분리함
2. **데이터 정규화**
   1. Standard Scaler로 데이터를 정규화함
3. **오버샘플링 진행**
   1. 종속변수의 class별 빈도가 불균형하여 SMOTE 방식 오버샘플링을 진행함

 🡪 

1. **군집화를 통한 파생변수 생성**

🞏 X\_train과 X\_test에 대한 군집화를 수행한 후 군집화 결과를 데이터 마트에 더하여 모델 구축에 활용하기로 함

🞏 K-means clustering을 통해 군집 수 2개~10개로 변화시켜가며 오차제곱합(SSE)의 변화와 실루엣계수를 통해 최적군집수를 찾아보았음

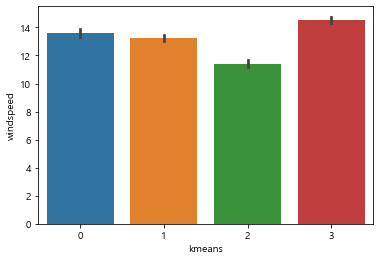
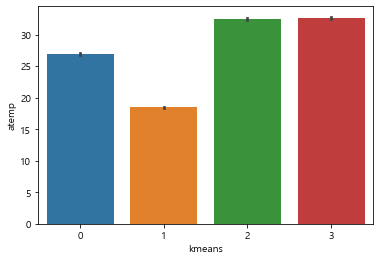
 ⭘ 군집 개수가 4개일 때가 최적인 것으로 판단됨

* 1. 군집화 결과 프로파일링

🞏 군집별 통계적 특성을 확인해 보았음(본 보고서에는 일부 그래프만 첨부)

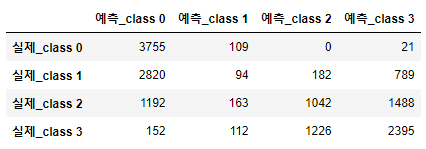
⭘ 군집2과 3의 atemp의 평균은 30을 상회하나, 군집1은 20을 밑돌았음

⭘ 군집0과 1의 windspeed의 평균은 ~~~



* 1. 군집화 결과 성능평가

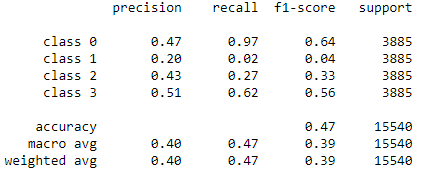
🞏 Confusion Matrix를 통해 class별 예측이 얼마나 맞았는지 확인함



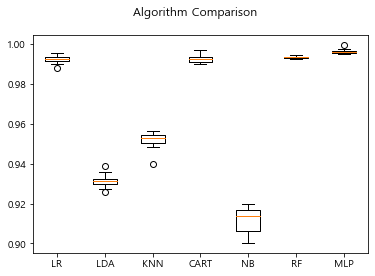
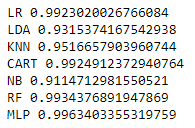
🞏 classification\_report를 통해 accuracy 및 F1 score를 확인함

⭘ acc는 0.47로 나타남

⭘ F1 score는 0.39로 나타남



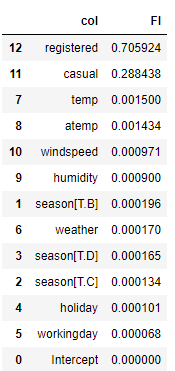
1. **모델링 및 기초 모델 선정 @@XGB는 필요시 전 버전 참고**
   1. 다양한 알고리즘을 활용하여 기본 모델을 구축함
      1. Logistic Regression
      2. Linear Discriminant Analysis
      3. K Neighbors Classifier
      4. GaussianNB
      5. RandomForest Classifier
      6. MLP Classifier
   2. Kfold=5 Cross validation을 통한 분류 정확도 기준으로 기준으로 보았을 때 MLP Classifier가 평균 0.996으로성능이 가장 높았음
   3. Training 시간을 감안하여 평균 정확도 0.993으로 두 번 째로 성능이 좋은 RandomForest Classifier를 최종 모델로 사용하기로 함



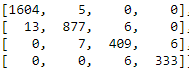
1. **파라미터 튜닝**
   1. 앞서 선택한 RandomForest Clasifier의 파라미터를 Gridsearch CV로 튜닝하여 성능을 개선함



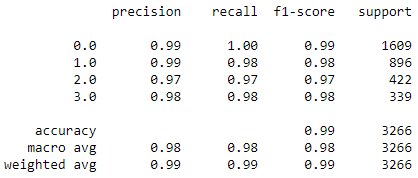
* 1. best parameter로 max\_feature : None, n\_estimators : 40이 선정됨
  2. 본 모델의 변수중요도는 registered, casual 순으로 높았음

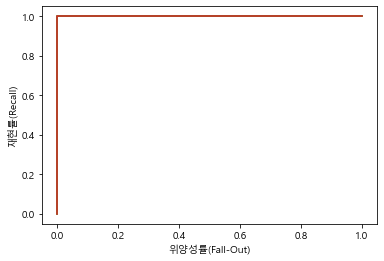
1. **Test set 활용하여 예측 수행 및 성능평가**
   1. Confusion Matrix
      1. Index 축이 실제(y\_test), col\_name 축이 예측(y\_pred)  
         엑셀로 하던가 해서 잘 그려보자



* 1. Classification Report
     1. Class별 F1-score는 class0 = 0.99, class1 = 0.98, class2 = 0.97
     2. 전체 F1-score는 0.98



* 1. ROC, AUC
     1. 4개 class에 대해 ROC를 도식화하고 ROC\_AUC를 계산함



AUC는 Class0: 1.0, Class:0.9 … 임.

//

1. **앙상블 모델 만들기**
   1. 1번 방법 : Randomforest로 그냥 제출
   2. 2번 방법 : 이항 분류 전제일 때, 모델 n개 골라서 각각 예측한 predict\_proba()의 평균을 0.5 기준으로 사용(결과 array중 왼쪽이 0일 확률, 오른쪽이 1일 확률)