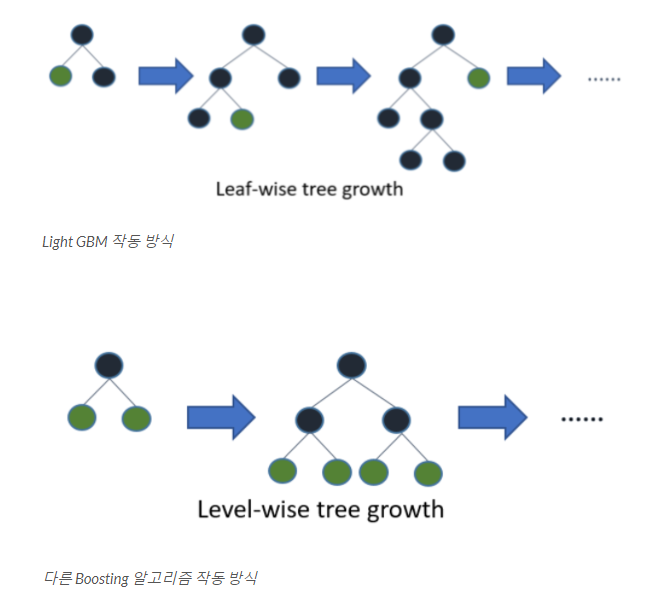
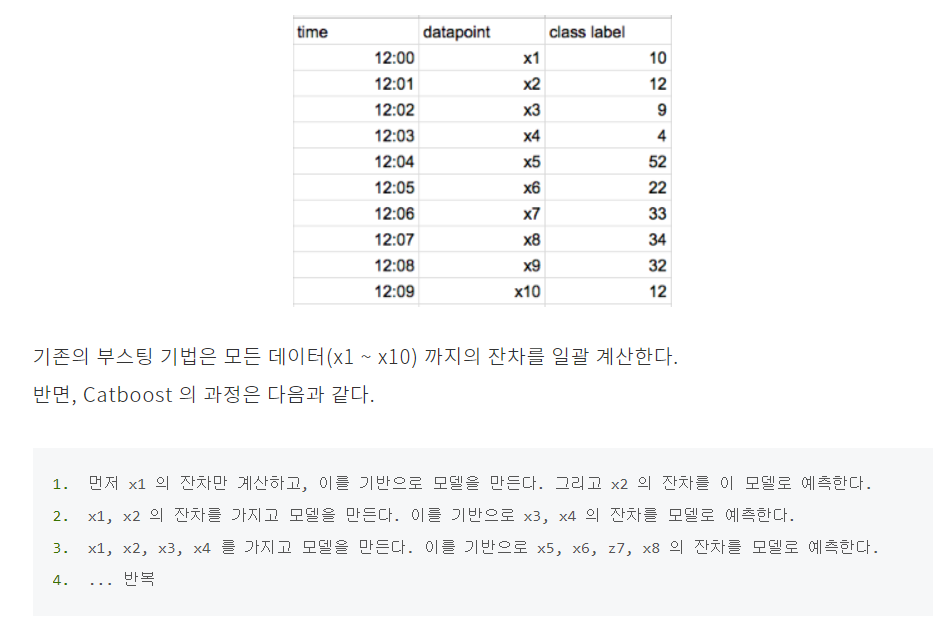
**[Ensemble 보고서 과제(7기 김형민)]**

1. voting, bagging, random forest, boosting, adaboost, gradient boost 요약
   1. **Voting**
      1. 하나의 데이터셋을 통으로 활용한다.
      2. 다양한 분류기를 활용하여 예측값을 도출한다.
      3. Hard voting은 다수의 분류기가 예측한 값을 결과물로 선정한다. (민주주의 투표 느낌)
      4. Soft voting은 분류기들이 예측한 값들의 확률 값을 계산하여 가장 확률이 높은 값을 결과물로 선정한다.
   2. **Bagging**
      1. 데이터셋을 여러 샘플로 나눈다. (복원추출)
      2. 하나의 분류기를 활용한다.
      3. Variance를 줄이는 효과를 통해 과적합을 방지하는 장점이 있다.
   3. **Random Forest**
      1. 변수를 임의추출하여 다양한 분류기를 만든다.
      2. 각 변수의 중요도를 계산할 수 있다.
      3. 변수의 임의추출로 인하여 각 분류기마다 뽑히지 않는 변수들이 존재할 수 있는데, 그에 해당하는 데이터는 test set으로서 활용될 수 있다.
   4. **Boosting**
      1. 다양한 분류기가 순차적으로 진행된다.
      2. 시간이 오래 걸리고 과적합 문제가 존재한다.
      3. 이전 분류기에서 오분류된 데이터에 가중치가 부여되는 방식이다.
      4. **Adaboost**
         1. 다양한 분류기들 중 특정 분류기(중요하다고 여겨지는 분류기)에 가중치가 부여된다.
         2. 이전 분류기에서 오분류된 데이터에 가중치를 높게 부여한 후 새로운 샘플로 iterate.
      5. **Gradient Boost**
         1. 다양한 분류기들 중 특정 분류기(중요하다고 여겨지는 분류기)에 가중치가 부여된다.
         2. 이전 분류기에서 오분류된 데이터만 남아 다음 분류기에 적용된다. (오분류된 데이터를 줄여가는 느낌)
2. Boosting의 advanced model 인 XGBoost, LightGBM, CatBoost 정리
   1. **XGBoost**
      1. 원래 boosting의 경우 분류기들이 순차적으로 연결되어 있어서 병렬 처리가 되지 않으나, XGBoost의 경우 병렬 처리가 가능하다.
      2. 기존의 방법들의 경우 예측값에 활용한 분류기들이지만 XGBoost의 경우 분류와 회귀 예측 둘 다 가능하다. (CART; Cart and Regression Tree 앙상블 모델을 사용)
      3. 즉, gbtree와 gbliner가 존재한다.
      4. 과적합 규제에 용이하며, 기존 분류기에서 사용한 것보다 훨씬 다양한 파라미터를 선정해주어야 한다.
   2. **LightGBM**
      1. 빠르다.
      2. 기존의 분류기는 수평적으로(level-wise) 확장되지만 LightGBM은 수직적으로(lead-wise) 확장된다. \* XGBoost와 Catboost는 level-wise 확장.
      3. 
      4. 분류와 회귀 둘 다 가능하고 결과물의 정확도를 최우선으로 한다. 즉, 과적합이 되기 쉽다. (max\_depth나 min\_data\_in\_leaf 등의 파라미터를 통해 과적합을 방지할 수 있긴 한다.)
   3. **CatBoost**
      1. Level-wise 확장 방식 분류기이다.
      2. Ordered Boosting을 사용한다.
         1. 기존 Gradient Boosting의 경우, 첫번째 분류기를 통해 일괄적으로 모든 데이터의 잔차를 계산하였고, 다음 분류기에 전체 데이터에서 나온 잔차를 대입하는 식이었음.
         2. CatBoost의 경우, 첫번째 분류기를 통해 일부 데이터(복원추출로 생성)의 잔차를 계산한 뒤, 이를 통해 모델을 만든다. 이 모델을 활용하여 다음 데이터의 잔차를 예측한다. (Ordered Boosting)
         3. 
         4. 대부분의 데이터가 수치형 변수가 아닌, 범주형 변수일 때 효과적인 분류기이다. 즉, 수치형 변수가 많을 경우에는 lightGBM이 더 효과적이다.