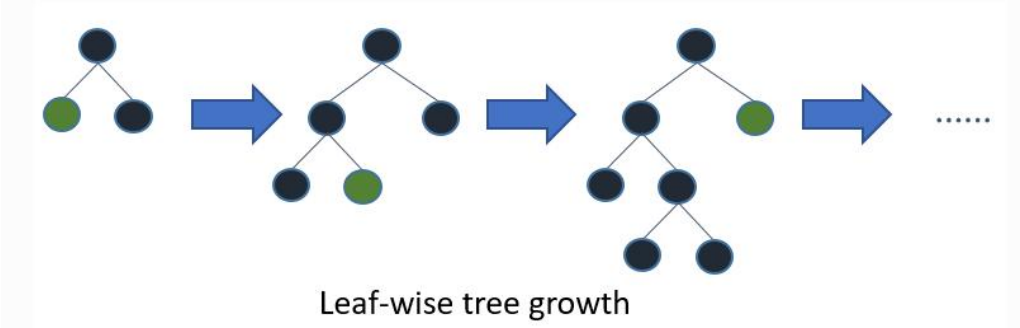
M5-forcasting-accuracy

Background: This is the guild line for kaggle contest, m5-forcasting-accuracy. It is released at mid-April in 2020. The purpose of this contest is to have a good prediction of the sales amount for Walmart in the next 28 days.

Data: The dataset includes three sheets which are history sales amount sheet, calender sheet and sale price sheet. History sales amount sheet gives sales amount in previous 5 years. Calender sheet contains information about holidays and special days. Sale price sheet holds the price data for different commodities.

Approach: This is a time-series prediction task. Many models can be useful to solve the problem. For example, we can use deep models such as lstm and neural networks or tree based models such as xgboost or light gbm. Considering the fact that there are over 60000 pieces of observations in the dataset, it could be difficult to handle it by using deep models. It seems that the most convenient model is light gbm. It is provides relatively precise predictions and a low requirement on hardware.

Lightgmb is a tree based classifier. The tree grow base on the property of leaf. At every iteration, we split the leaf with largest gain. This is a quite popular algorithm in kaggle contests.



Leaf-wise tree growth

Parameters:

More leaves may lead to overfitting. Emperically, smaller dateset should have less leaves.

Objective is the optimize function of the model. In this task ,we will use a poisson-like objective. It is natural because sales amount may follow a poisson distribution.

Boosting: we use gbdt which is the traditional gradient based decision tree.

N_estimator controls the iteration number. Higher value may lead to over-fitting. The default value is 100.

Num_leaves is the max number of leaves in a tree.

Learning_rate controls the learning speed and quality. The default value is 0.1.

Max_depth controls the maximum depth of the tree. It is related with over-fitting.

Bagging_fraction/subsample: this is used to randomly select a part of data in training.

Bagging_freq/subsample_freq controls the frequency for bagging.

Early_stopping_round: if the metric do not improve for certain number of rounds, stop the training.

Lambda_l1 and lambda_l2 are L1 regularization and L2 regularization.

Max_bin: max number of bins that feature values will be bucked in. Small number may reduce

accuracy but will increase general power.
Categorical_feature is used to specify categorical features.