**Finding similar Kaggle solutions**

**A simple model - using the market and news data**

Competition: twosigma;

url: <https://www.kaggle.com/bguberfain/a-simple-model-using-the-market-and-news-data>;

Score: 0.46827; votes:200;

Valuable takeaways: clearly illustrated the general process of prepare data for modelling purpose, feature importance, use of lgb for modelling, result evaluation, a set of lgb parameters, use lgb module to examine feature importance

Rooms for improvement: categorical features are not used in public kernel, assetCode, assetName included in features, we can build process to examine most wrongly predicted result, PCA (dimensionality reduction)

Feature importance (top 10):month, assetName, assetCode, open, returnOpenPrevRaw10, returnOpenPrevRaw1, close, returnClosePrevMktres10, returnClosePrevRaw10, returnsOpenPrevMktres10

Feature importance (top 10, gain):returnOpenPrevRaw10, month, returnOpenPrevMktres10, returnClosePrevMktres10, returnClosePrevRaw10, returnOpenPrevRaw1, assetName, open, assetCode, returnClosePrevRaw1

|  |  |  |  |
| --- | --- | --- | --- |
| **Join Market and News Data** | **Features Selected** | **Preprocessing** | **Modelling** |
| Repeat rows in news data for each asset involved | assetCode, assetName, Urgency, takeSequence, bodySize, wordCount, sentenceCount, | Y value clip(-1, 1) | lgb |
| Preserve original index in new df | companyCount, marketCommentary, relevance, sentimentNegative | Label encoding | Train, val split of 0.8 |
| News feature aggregation perform on [time, assetCode] level | sentimentNeutral, sentimentPositive, sentimentWordCount, noveltyCount12H, noveltyCount24H, noveltyCount3D, noveltyCount5D, noveltyCount7D | Time transformation to enable common joining key for market and news data | Parameter tuning  lgb\_params = dict(  objective = 'regression\_l1',  learning\_rate = 0.1,  num\_leaves = 127,  max\_depth = -1, *# min\_data\_in\_leaf = 1000,* *# min\_sum\_hessian\_in\_leaf = 10,*  bagging\_fraction = 0.75,  bagging\_freq = 2,  feature\_fraction = 0.5,  lambda\_l1 = 0.0,  lambda\_l2 = 1.0,  metric = 'None', *# This will ignore the loss objetive and use sigma\_score instead,*  seed = 42 *# Change for better luck! :)* ) |
| Left join market\_train with news\_train\_df\_aggregated | volumeCounts12H, volumeCounts24H, volumeCounts3D, volumeCounts5D, volumeCounts7D,, | Dayofweek, month features to be created |
|  | Aggregation:  urgency: [‘min’, ‘count’], takeSequence: [‘max’], Else: [‘min’, ‘max’, ‘mean’, std’] |  |
|  | Other features:  Dayofweek, month |  |