TERM PROJECT REPORT

Selen Gecgel, Anil Ozturk

2018/12/24

1 Introduction

In this competition, all competitors must predict a signed confidence value, $\hat{y}_{ti} \in [-1, 1]$, which is multiplied by the market-adjusted return of a given ***assetCode*** over a ten day window.

For each day in the evaluation time period, we calculate:

$$x_t = \sum_i \hat{y}_{ti} r_{ti} u_{ti}, \tag{1}$$

where r_{ti} is the 10-day market-adjusted leading return for day t for instrument i, and u_{ti} is a 0/1 ***universe*** variable.

2 DATASET DESCRIPTION

To predict confidence value, competitors have two dataset;

1. Market Data
2. News Data

2.1 Market Data

- Dataset contains financial market information such as opening price, closing price, trading volume, calculated returns, etc. relating to different companies in the time range between 2007 and present. The market data contains a variety of returns calculated over different time spans .
- Market train data size: (4072956, 16)
- Market test data size : (1823, 14)

2.2 News Data

- News data contains information about news articles/alerts published about assets, such as article details, sentiment, and other commentary.
- News train data size: (9328750, 35)
- News test data size: (2776, 35)

3 METHODS USED

Outlier data removal process was performed first. data understanding is aimed and with this goal the relationships between features are investigated. After this step, regarding the results of the data analysis, the main concept of our approach is determined. For detection of feature relevance on dataset, “pair-wise correlation matrix” is extracted.

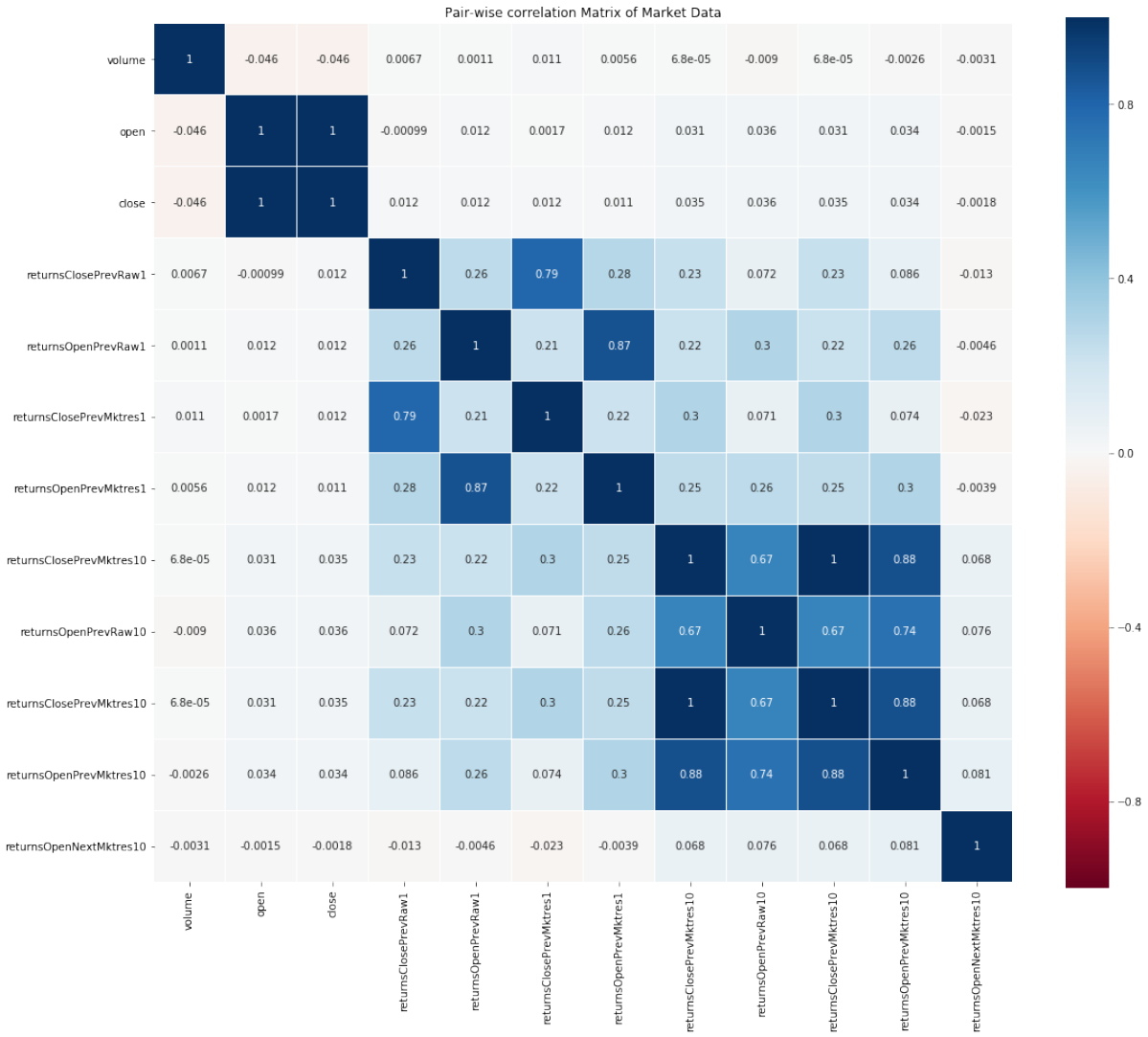

Figure 1: Pair-Wise Correlation Matrix of Market Data.

We found that decision tree based concepts were the best fit for this data-set due to the nature of the data. On account of data, it is highly noisy and is under strong effects which are changeable depends on the case at this year such as; 2010 crisis. Also, data contains a lot of outliers. At this point, using decision tree is seen logical idea because of their robustness to errors and the its ability of handling both categorical and numerical values. However, we believe that using tree based models are more applicable for news and market data. We think that a single decision tree is not strong enough to be used for best solution and can cause to over-fitting. This is why we did not prefer to use decision tree alone:instead;multiple trees are used together. Ensemble models combines multiple trees sequentially or parallel. They provide robustness over a single estimator and improve the generalizability. We prefer to implement Boosting methods with the aim of decreasing the bias. Another reason for choosing the boosting methods that they increase the predictive force by combining single models which are weighted majority vote.

Gradient boosting is a powerful machine-learning technique that achieves state-of-the-art results in a variety of practical tasks. For many years, it has remained the primary method for learning problems with heterogeneous features, noisy data, and complex dependencies. Gradient boosting is essentially a process of constructing an ensemble predictor by performing gradient descent in a functional space [1]. We implemented three state of the art gradient boosting methods; XGBoost, LightGBM and CatBoost.

Table 1: Methods are tried to get best score

Methods	**XGBoost**	**LightGBM**	**CatBoost**
Advantages	• Support for GBDT, GLM, DART • Better than GBM framework alone • It converts weak learners into strong learners	• Ability of handling with categorical features • Gradient-based One-Side Sampling(GOSS) for finding the best split • higher sampling weights to data points with larger gradients	• CatBoost provide missing value handle support • CatBoost reduces overfitting with Dynamic Boosting(compared to LightGBM and XGBoost) • Ability of handling with categorical features • Very fast inference
Disadvantages	• Pre-procesesing(label encoding, mean encoding or one-hot encoding) requirement for categorical features • Pre-sorted algorithm&Histogram-based algorithm for computing the best split • build more deep asymmetric trees	• It builds more deep asymmetric trees	• It builds large symmetric trees
Score of the ***itu_ml*** team	0,61258	0,65	**0,69**

4 CatBoost(Categorical Boosting)

All existing implementations of gradient boosting face the problem which prediction model F obtained after several steps of boosting relies on the targets of all training examples and it leads to a prediction shift of the learned model. Also, other gradient boosting methods convert categories to their target statistics. A target statistic is a simple statistical model itself, and it can also cause target leakage and a prediction shift. CatBoost uses ordering principle to solve both problems.

A gradient boosting procedure builds iteratively a sequence of approximations $F^t : R^m \to R$, $t = 0, 1, ...$ in a greedy fashion. Namely, F^t is obtained from the previous approximation F^{t-1} in an additive manner:

$$F^t = F^{t-1} + \alpha h^t, \tag{2}$$

where α is a step size and function $h^t : R^m \to R$, (a base predictor) is chosen from a family of functions H in order to minimize the expected loss:

$$h^t = \underset{h \epsilon H}{argmin} \quad \mathcal{L}\left(F^{t-1} + h\right). \tag{3}$$

CatBoost is an implementation of gradient boosting, which uses binary decision trees as base predictors. Each leaf of the tree is assigned to a value, which is an estimate of the response y in the predicted class label. A decision tree h can be written as

$$h\left(x\right) = \sum_{j=1}^{J} b_j 1_{\{x \epsilon R_j\}} \tag{4}$$

where R_j are the tree nodes corresponding to the leaves of the tree.

In CatBoost, base predictors are oblivious decision trees also called decision tables [23]. Term oblivious means that the same splitting criterion is used across an entire level of the tree. Such trees are balanced, less prone to overfitting, and allow speeding up execution at testing time significantly. , CatBoost concatenates all categorical features (and their combinations) already used for splits in the current tree with all categorical features in the dataset [1].

References

[1] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: unbiased boosting with categorical features. (Section 4), 2017.