Istanbul Technical University

Machine Learning

BLG527E

Project Preliminary Report

Anıl Öztürk
Selen Geçgel

December 20, 2018

1 General Information

	Team Member 1	**Team Member 2**
Name	Anıl ÖZTÜRK	Selen GEÇGEL
ID	504181504	504161330
Competition Name	Two Sigma: Using News to Predict Stock Movements	
Kaggle ID	nlztrk	selengecgel
Kaggle Team ID	ituml_team	
Kaggle Score	0.63629	
Kaggle Score Date	22.11.2018	
Used Methods	Cleaning the data, LightGBM with Binary Logloss	

2 Used Methods

2.1 Cleaning

Firstly, we deleted the outliers from the data. Then we set up the NaN Mktres values to Raw values. We calculated the open/close ratios and delete the rows that have abnormal o/c ratio.

2.2 LightGBM

LightGBM is a gradient boosting framework that uses tree based learning algorithms. Our main reasons for selecting LightGBM are as in below;

- Faster training speed and higher efficiency
- Lower memory usage
- Support of parallel and GPU learning
- Better accuracy

  In the beginning of study, there are used several methods such as; XGBoost etc. but using LightGBM offers higher accuracy than other approaches (especially with binary logloss feature)
- Capable of handling large-scale data

Our parameters to tune LightGbm can be seen in the table below.

Paraneter Groups by Functionality	**Parameter Name**	**Value**
Core Parameters	learning_rate	0.01
	num_leaves	60
Learning Control Parameters	max_depth	-1
	bagging_fraction	0.9
	feature_fraction	0.9
	bagging_freq	5
	bagging_seed	2018
Metric Parameters	metric	binary_logloss
IO Parameters	verbosity	-1

2.3 Future Plans

- To raise performance extraction of relationship between features and also between news data and market data should be needed certainly. Therefore to find correlation between features some approaches such as; "Pearson pairwise correlation ", "Mutual information " will be tried with LightGBM and next prediction algorithms.

- Implementation of Recurrent Neural Networks. We (ituml_team members) think that it is very available for Two Sigma competition datasets because of the success in time series data. Also we will plan to use LSTM specifically.