# Submission description

For: Data Mining Competition @ ISMIS 2017
Participant: Quang-Vinh DANG
Email: dqvinh87@gmail.com
Date: 14 – Jan - 2017

## Introduction

The document describes our approach in Data Mining Competition organized by the conference ISMIS 2017. Further details information can be found at https://knowledgepit.fedcsis.org/contest/view.php?id=119

The taks is to predict the decision on a given stock. Three decisions are available: Buy, Hold, Sell. The training data includes stock ID, and several recommendations from experts. Each recommendation includes expert ID, the advice (Buy, Hold, Sell), expected return of this stock (numerical values, sometimes N/A) and the number of date before the real decision is made. The testing data set includes only recommendation, and we need to predict the decision.

## Approach

### Preprocessing

#### Expert Profile

We realized that not all expert have the same quality. Hence, we built a database to manage all the expert. We assigned to each expert a rating score, qualified how well he/she made recommendation in the past.
In order to assign the rating score of a particular expert, we scan the whole training dataset, count how many times this expert had give recommendation, and count how well each recommendation is, in compare to the true decision.
For each correct recommendation, the score is increased by 1.0.
For each not-so-wrong recommendation, e.g. the advice was Buy but it turned out that should be Hold, the score is increased by 0.5.
For a completely wrong recommendation, e.g. the advice was Buy but the true decision is Sell, or vice versa, the score is kept the same.
After all, we calculate the rating score by divided the total score to the number of recommendation has been made.
The rating score is a number range between 0 and 1. Higher value means better expert.

#### Accumulative recommendation

Now, for each row in the dataset, which contains a series of recommendation, we add up to see how many Buy, Sell and Hold recommendation are given. However, we take the rating

score of each expert into consideration. For instance, if there are two experts with rating score of 0.8 and 0.6 gave the advice of Buy, we will consider the case as the advice of Buy with the weight score of 1.4.

By doing so, we can build a new dataset with three features (Buy, Hold, Sell) as weighted recommendation, and one response value (true_decision). The problem now turned to be a classical multi-class classification.

## Classification

There are indeed several different ways to perform the task of classification. We validated each solution by 5-fold cross-validation in the training set. The methods we have tried are:

- Random forest
- Deep neural networks
- Gradient boosted machines

Empirical values suggested that gradient boosted machines (GBM) should be used. We optimized the model by random hyper parameter search.