Research in Collective Intelligence through Horse Racing in Hong Kong

LYU1604

Lam Cheuk Hei (1155049403)

Chan Lok Wai (1155050024)

Supervised By Prof. Michael R. Lyu

Acknowledgements

Looking back in this research, we have been doing it for one year. In this year, we acquired a lot of knowledge through this research, and the most important thing is that we learned a lot of soft skills too, such as how to cooperate with partner, a better time management and find out what we interested to discover and try to study it.

The completion of this undertaking could not have been possible without the participation and assistance of so many people. Their contribution are sincerely appreciated and gratefully acknowledged. We would like to express our deep appreciation and indebtedness to the following:

Dr. Michael R. Lyu, thank you for guiding us to complete the research, and provide us professional advice, so that we can clearly understand what should we do during the time.

Mr. Edward Yau, thank you for your endless support, kind and leading us during the project, when we were facing technical problems, you taught us patiently.

LYU other teams, thank you for supporting each other, you all are kind and positive, so that we could have the smooth and meaningful regular meeting.

Abstract

Horse racing is a popular sport in Hong Kong, there are a group of people enjoying betting on it. We found that there are many experts who give tips to newspaper reader, we want to know if collective intelligence, one type of crowdsourcing, is better than individual experts.

In this study, we have two collective intelligence methods: (1)voting system, (2) TensorFlow and (3) combine of voting system and TensorFlow.

The purpose of this study possessed several folds: (a) to investigate how accurate are the experts on betting horse racing;(b) to investigate how good are the experts on betting horse racing in terms of money returns;(c) to investigate how accurate is collective intelligence on betting horse racing;(d) to investigate how good is collective intelligence on betting horse racing in terms of money returns;(e) how can we make a better strategy to bet on horse racing.

Conclusively, we want to know if collective intelligence better than individuals. We carried it out under mainly quantitative research method on over thousand races and over 50 experts. Through this study, we want to find out the best strategy to bet on horse racing to maximize the money returns.

Key words: horse racing, crowdsourcing, collective intelligence, TensorFlow

Contents

1 Introduction

1.1 Motivation

Crowdsourcing is a hot topic in recent years. A crowdsourcing system helps people to solve problem by the data provided by user. However, will people input some wrong data that hinder the task? Is there any way to filter the inappropriate data? Collective intelligence is one of the application in crowdsourcing. It involves consensus, social capital and formalisms such as voting systems, social media and other means of quantifying mass activity (ABRAHAMS, 2016). Under quantifying mass activity, we can measure the data provided by people easily. Since we desire to know the effectiveness of crowdsourcing, we are going to conduct an experiment, to test whether the wisdom of crowd is better than decisions made by individual expert.

We choose horse racing as our research base because it provides discrete result, unlike stock market, which is wide variety. And also, many experts believe that horse racing is predictable(Ziemba, 2013).

1.2 Objective

As we have mentioned, we are trying to use horse racing as an example to investigate the power of crowdsourcing, we choose over 50 experts on horse racing, and see how they perform in the past races, and we compare them to our voting system, TensorFlow and the combine of voting system and TensorFlow.

After that, we hope to figure out the best strategy to bet on horse racing, so that we can earn money.

1.3 Research Problem

According to our research objective, we have the following problem:

1. How accurate are the experts on betting horse racing?
2. How good are the experts on betting horse racing in terms of money returns?
3. How accurate is collective intelligence on betting horse racing?
4. How good is collective intelligence on betting horse racing in terms of money returns?
5. Is crowdsourcing better than individual experts?
6. What is the best strategy to bet on horse racing?
7. Can we actually earn money using our chosen best method?

1.4 Definition

Definition of collective intelligence

Two or more people independently acquire information and this information is combined and processed through social interaction, which provides a solution to a specific problem such that it cannot be implemented by isolated individuals .[7] It is a form of universally distributed intelligence, constantly enhanced, coordinated in real time, and resulting in the effective mobilization of skills. [23] We will add the following indispensable characteristic to this definition: The basis and goal of collective intelligence is mutual recognition and enrichment of individuals rather than the cult of fetishized or hypostatized communities. [24]

Definition of crowdsourcing

Although it is hard to define crowdsourcing exactly, we can define it as a general-purpose problem-solving method: a crowdsourcing system invites a crowd of people to help solving a problem.[10] Estellés-Arolas and Gonzáles-Ladrón- de-Guevara (2012) gave us another detailed definition:

Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of

individual skills, while the crowdsourcer will obtain and utilize to their advantage what the user has brought to the venture, whose form will depend on the type of activity undertaken.[12]

Definition of machine learning

Machine learning is the subfield of computer science that, according to Arthur Samuel in 1959, gives "computers the ability to learn without being explicitly programmed."[1] It is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. Machine learning is sometimes conflated with data mining,[18] where the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning. [19]

Definition of TensorFlow

TensorFlow is Google Brain's second generation machine learning system, released as open source software on November 9, 2015.[14] It is developed by Google to meet their needs for systems capable of building and training neural networks to detect and decipher patterns and correlations, analogous to the learning and reasoning which humans use. It is originally developed for the purposes of conducting machine learning and deep neural networks research, but the system is general enough to be applicable in a wide variety of other domains as well.[19]

TensorFlow is an interface for expressing machine learning algorithms, and an implementation for executing such algorithms. A computation expressed using TensorFlow can be executed with little or no change on a wide variety of heterogeneous systems, ranging from mobile devices such as phones and tablets up to large-scale distributed systems of

hundreds of machines and thousands of computational devices such as GPU cards. The system is flexible and can be used to express a wide variety of algorithms, including training and inference algorithms for deep neural network models, and it has been used for conducting research and for deploying machine learning systems into production across more than a dozen areas of computer science and other fields, including speech recognition, computer vision, robotics, information retrieval, natural language processing, geographic information extraction, and computational drug discovery. The TensorFlow API and a reference implementation were released as an open-source package under the Apache 2.0 license in November, 2015 and are available at www.tensorflow.org. [30]

Definition of neural network

A neural network is a series of algorithms that attempts to identify underlying relationships in a set of data by using a process that mimics the way the human brain operates. [9] Neural networks have the ability to adapt to changing input so the network produces the best possible result without the need to redesign the output criteria. The concept of neural networks is rapidly increasing in popularity in the area of developing trading systems. [15]

A neural network operates similar to the brain's neural network. A "neuron" in a neural network is a simple mathematical function capturing and organizing information according to an architecture. The network closely resembles statistical methods such as curve fitting and regression analysis. [22]

In general, there are three fundamentally different classes of network architectures in Artificial Neural Networks- Single-Laye Feedforward Networks (SLFF) which have an input layer of source nodes that projects onto an output layer of neurons; Multilayer Feedforward Networks (MLFF) which have some hidden layers between input layer and output layer; and

the last class of network architecture is Recurrent Networks (RN) which have at least one feedback loop that feeding some neuron's output signal back to inputs [9].

There are two types of adaptive algorithms that help us to determine which structure is better for this application: network pruning: start from a large network and successively remove some neurons and links until network performance degraded, network growing: begin with a small network and introduce new neurons until performance is satisfactory [9].

Multilayer feedforward neural network is the best architecture that minimizes mean square error (MSE) by network growing method. This network consists of four layers: an input layer that each neuron in this laye corresponds to one input signal; two hidden layers of neurons that adjust in order to represent a relationship; and an output layer that each neuron in this layer corresponds to one output signal. Also, the network is fully connected in the sense that every node in each layer of the network is connected to every other node in the adjacent forward layer.

Definition of win

Win is one of the basic bet type of horse racing. Bettors win the bet if the chosen horse finish first. They receive the dividend determined by the win odds. The chance of winning is average.

Definition of place

Place is another basic bet type of horse racing. Bettors win the bet if the chosen horse finish first, second or third. Winner receive the dividend determined by the place odds. The chance of winning is good so place odds are usually lower than win odds.

1.4 Other related study

Horse racing is predictable, and the results on every racing day have a pattern that it's not easy to find out, but with computer we can find the pattern. [4,11]

Refer to ELNAZ's study, he did a research on horse racing prediction using artificial neural networks. Artificial Neural Networks (ANNs) have been applied to predict many complex problems. The paper ANNs were applied to horse racing prediction. He employed Back-Propagation, Back-Propagation with Momentum, Quasi-Newton, Levenberg-Marquardt and Conjugate Gradient Descent learning algorithms for real horse racing data and the performances of five supervised NN algorithms were analyzed. Data collected from AQUEDUCT Race Track in NY, include 100 actual races from 1 January to 29 January 2010. The experimental results demonstrate that NNs are appropriate methods in horse racing prediction context. All algorithms can produce acceptable prediction with and accuracy of 77% by average. The performance of each model has been influenced by the Neural Network structure and learning algorithm. [9]

As a result, it was shown that BP algorithm slightly performs better than other algorithms but need a longer training time. Furthermore, LM is the fastest.[9]

We were inspired by this study a lot, as this study was conducted in 2010, we want to apply TensorFlow in our research, and his research used 'Horse weight, type of race, horse's trainer, horse's jockey, number of horses in the race, race distance, track condition and weather' as input features, so that we also use similar features in our research.

2 Research Methods

To study the performance of experts and collective intelligence, we use quantitative research method to achieve this goal. We mainly focus on two fields: accuracy and profit percentage. As there are different types of betting method, such as win and place, so we will analysis the best betting strategy to maximize the accuracy and profit percentage.

2.1 Research Design

From term 1 to term 2

In term 1, we have developed a website which aimed to attract people to use our system, and we want them to provide their predictions to us, and our system would gather all the predictions to provide a tips. However, there were a lot of problems, such as not enough users, they may give random predictions, they may not provide predictions to us regularly. And the biggest problem was that we didn't know the algorithm is work or not. We reflected and discussed a lot, we finally decided to change our direction, to solve the problem of 'users', we choose our target audience to horse racing experts, they provide tips on newspaper regularly, and we assume they provide tips with their own professional knowledge. Another part is the evaluation of the tips, we need to know our new design is work or not, so that we to collect all horse racing data, and see which betting method is the best. For the algorithm part, it is the most important part, when we have the tips from experts, how can we choose the best method to predict horse racing? We developed voting system, and also we choose to use TensorFlow as a tool to predict horse racing, because TensorFlow is a hot machine learning method in the recent year, we want to know if this machine learning method can really help us, so we decided to use TensorFlow in our research in term 2.

Term 2 design

In this research, we need to have the past horse racing information, without them we cannot determine which one is better. The following are the information of each race we collected: place, horse number, horse, jockey, trainer, actual weight, declared horse weight(horse weight on two days before the race day), draw, running position, finish time and win odds. In total, we collected the information of 4235 races, which included the latest racing day. With those ground truth information, we can use them in the later part of our research.

As we are doing research on collective intelligence, we need to collect tips from horse racing experts, in total, we found 54 experts who provide tips on newspaper, and we got their tips in the past 569 races(up to the latest race). The format of each tips record is as follow: race year, race month, race day, race number, predicted winner horse number, predicted no.2 horse number, predicted no.3 horse number and predicted no.4 horse number.

At this point, we already have enough data for us to do our research. The next step is deciding how to analysis them. As we mentioned before, we chose to use TensorFlow and voting system as representatives of collective intelligence, to compare with individual experts, so we will design these two methods to make predictions, the details of TensorFlow and voting system will be elaborated in part 3 of this research report.

It is important to verify our designed method to check if they work, so we chose 3 types of betting method: (1)choose three horses in each race, and buy ‘win’ for all of them. (2)choose one horse in each race, and buy ‘win’ for it. (3) choose three horses in each race, and buy ‘place’ for all of them. We will do a comparison of all predictions on the latest 117 races.

The result of the comparison will divide into two parts, we first focus on accuracy, we want to know who’s tips is the most accurate, and see the difference between them, is

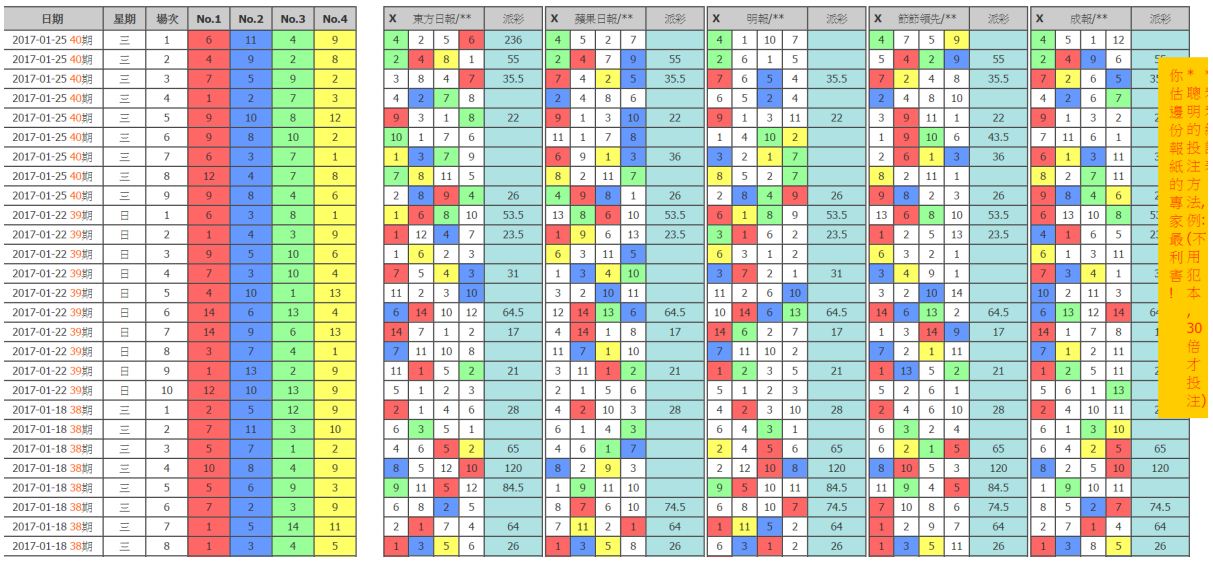
collective intelligence better? or worse? And the other part is money returns, we want to know who's tips can earn the most money, or maybe loss money?

After all, we will have the results for all experts and collective intelligences as the following:

1) Accuracy
    a) choose three horses in each race, and buy 'win' for all of them.
    b) choose one horse in each race, and buy 'win' for it.
    c) choose three horses in each race, and buy 'place' for all of them.
2) Profit/Loss Percentage
    a) choose three horses in each race, and buy 'win' for all of them.
    b) choose one horse in each race, and buy 'win' for it.
    c) choose three horses in each race, and buy 'place' for all of them.

2.2 Research Participants

The research participants are the core of our project, because we are doing research on collective intelligence, the whole project is built on the tips from horse racing expert. After a long time of data collection, we found 54 experts who provide tips on newspaper from September 2016 to recent. Those experts are all professional and experienced in predicting horse racing result.



日期	星期	場次	No.1	No.2	No.3	No.4
2017-01-25 40期	三	1	6	11	4	9
2017-01-25 40期	三	2	4	9	2	8
2017-01-25 40期	三	3	7	5	9	2
2017-01-25 40期	三	4	1	2	7	3
2017-01-25 40期	三	5	9	10	8	12
2017-01-25 40期	三	6	9	8	10	2
2017-01-25 40期	三	7	6	3	7	1
2017-01-25 40期	三	8	12	4	7	8
2017-01-25 40期	三	9	9	8	4	6
2017-01-22 39期	日	1	6	3	8	1
2017-01-22 39期	日	2	1	4	3	9
2017-01-22 39期	日	3	9	5	10	6
2017-01-22 39期	日	4	7	3	10	4
2017-01-22 39期	日	5	4	10	1	13
2017-01-22 39期	日	6	14	6	13	4
2017-01-22 39期	日	7	14	9	6	13
2017-01-22 39期	日	8	3	7	4	1
2017-01-22 39期	日	9	1	13	2	9
2017-01-22 39期	日	10	12	10	13	9
2017-01-18 38期	三	1	2	5	12	9
2017-01-18 38期	三	2	7	11	3	10
2017-01-18 38期	三	3	5	7	1	2
2017-01-18 38期	三	4	10	8	4	9
2017-01-18 38期	三	5	5	6	9	3
2017-01-18 38期	三	6	7	2	3	9
2017-01-18 38期	三	7	1	5	14	11
2017-01-18 38期	三	8	1	3	4	5

X	東方日報/**				派彩
4	2	5	6		236
2	4	8	1		55
3	8	4	7		35.5
4	2	7	8		
9	3	1	8		22
10	1	7	6		
1	3	7	9		
7	8	11	5		
2	8	9	4		26
1	6	8	10		53.5
1	12	4	7		23.5
1	6	2	3		
7	5	4	3		31
11	2	3	10		
6	14	10	12		64.5
14	7	1	2		17
7	11	10	8		
11	1	5	2		21
5	1	2	3		
2	1	4	6		28
6	3	5	1		
4	6	5	2		65
8	5	12	10		120
9	11	5	12		84.5
6	8	2	5		
2	1	7	4		64
1	3	5	6		26

X	蘋果日報/**				派彩
4	5	2	7		
2	4	7	9		55
7	4	2	5		35.5
2	4	8	6		
9	1	3	10		22
11	1	7	8		
6	9	1	3		36
8	2	11	7		
4	9	8	1		26
13	8	6	10		53.5
1	9	6	13		23.5
6	3	11	5		
1	3	4	10		31
3	2	10	11		
12	14	13	6		64.5
4	14	1	8		17
11	7	1	10		
3	11	1	2		21
2	1	5	6		
4	2	10	3		28
6	1	4	3		
4	6	1	7		
8	2	9	3		
1	9	11	10		
8	7	6	10		74.5
7	11	2	1		64
1	3	5	8		26

X	明報/**				派彩
4	1	10	7		
2	6	1	5		
7	6	5	4		35.5
6	5	2	4		
9	1	3	11		22
1	4	10	2		
3	2	1	7		
8	5	2	7		
2	8	4	9		26
6	1	8	9		53.5
3	1	6	2		23.5
6	3	1	2		
3	7	2	1		31
11	2	6	10		
10	14	6	13		64.5
14	6	2	7		17
7	11	10	2		
1	2	3	5		21
5	1	2	3		
4	2	3	10		28
6	4	3	1		
2	4	5	6		65
2	12	10	8		120
9	5	10	11		84.5
6	8	10	7		74.5
1	11	5	2		64
6	3	1	2		26

X	節節領先/**				派彩
4	7	5	9		
5	4	2	9		55
7	2	4	8		35.5
2	4	8	10		
3	9	11	1		22
1	9	10	6		43.5
2	6	1	3		36
8	2	11	1		
9	8	2	3		26
13	6	8	10		53.5
1	2	5	13		23.5
6	3	2	1		
3	4	9	1		
3	2	10	14		
14	6	13	2		64.5
1	3	14	9		17
7	2	1	11		
1	13	5	2		21
5	2	6	1		
2	4	6	10		28
6	3	2	4		
6	2	1	5		65
8	10	5	3		120
11	9	4	5		84.5
7	10	8	6		74.5
1	2	9	7		64
1	3	5	11		26

X	成報/**				派彩
4	5	1	12		
2	4	9	6		
7	2	6	5		
4	2	6	7		
9	1	3	2		
7	11	6	1		
6	1	3	11		
8	2	7	11		
9	8	4	6		
6	13	10	8		
4	1	6	5		
6	1	3	11		
7	3	4	1		
10	2	11	3		
6	13	12	14		
14	1	7	8		
7	1	2	11		
1	2	5	11		
5	6	1	13		
2	4	10	11		
6	1	3	10		
6	4	2	5	65	
8	2	5	10	120	
1	9	10	11		
8	5	2	7	74.5	
2	7	1	4	64	
1	3	8	5	26	

A part of raw data collected from the internet

The above figure shows a part of raw data, provided by 5 experts from different newspapers from 18th to 25th January 2017, each expert provides 4 horses for each race. Since our data is started from September 2016, with 54 different opinions, the amount of data is too huge. If we use all experts as our research participants, it will spend a long time for calculating, no matter for TensorFlow or voting system. Although using all data can raise the meaning of our project, we have limited time for sorting data and calculating result. So, we decide to use the tips from 4 experts with highest win rate, to ensure the accuracy and observability of our predicted result.

Raw data from 4 experts with highest win rate

2.3 Research Instruments

Since our project is doing a type of quantitative research, experts are acting a very important role in this case. We have to choose experts who provide the most accurate prediction. To determine whether the tips are meaningful, we need to do a lot of calculation and testing for their tips. Also, to analyze whether our methods are works, it has to predict more than hundred results of races by TensorFlow, voting system, and the combined method.

3 Research Implementation

3.1 Database Design

Since our project involves data collection and analysis, a database is needed to store the tips and horse racing result. We have chosen a MySQL database, phpMyAdmin provided by the VIEW Lab. It is a user-friendly platform for developing and managing our database. We have created some tables on phpMyAdmin to store different data, including ‘tips’ tables to store the experts’ tips, ‘race’ to store the racing information and predicted result.

Tips tables

Our tips are provided by 54 experts from 10 newspapers, so we create 10 tips tables to store the tips from each newspaper, including ‘tips_ad’, ‘tips_dn’, ‘tips_hn’, and etc. The names are the short term of each newspaper, for example ‘tips_ad’ refers to tips from Apple Daily. Figure 3.1.1 shows a part of data in ‘tips_ad’. The first six columns, ‘raceyear’, ‘racemonth’, ‘raceday’, ‘date’, ‘week’, ‘raceno’, are the basic information of each race. The next four columns, ‘no1’, ‘no2’, ‘no3’, ‘no4’, store the horse numbers of the first four horses of each race, ‘no1’ means the champion of that race and so on. The remaining columns store the tips provided by 4 experts from Apple Daily and the dividends of each race. ‘ad40’, ‘ad41’, ‘ad42’, ‘ad43’ are the id of experts, ‘ad41_1’ is the first horse predicted by ‘ad41’ and so on. A amount exists in the ‘dividend’ column only if the prediction is corrected. The structure of ‘tips_ad’ is shown in Figure 3.1.2.

raceyear	racemonth	raceday	date	week	raceno	no1	no2	no3	no4	ad41_1	ad41_2	ad41_3	ad41_4	ad41_dividend	ad42_1	ad42_2	ad42_3	ad42_4	ad42_dividend	ad43_1	ad43_2	ad43_3	ad43_4	ad43_dividend	ad40_1	ad40_2	ad40_3	ad40_4	ad40_dividend
2017	03	29	2017-03-29?58期	三	1	2	6	10	5	6	3	8	5		6	3	2	1	233.5	6	3	4	5		1	3	6	9	
2017	03	29	2017-03-29?58期	三	2	9	4	1	6	2	6	10	7		4	1	6	3		1	6	4	7		6	8	2	1	
2017	03	29	2017-03-29?58期	三	3	7	11	4	5	2	3	9	5		2	11	3	6		11	5	2	7	82.5	11	2	6	4	
2017	03	29	2017-03-29?58期	三	4	12	3	10	2	9	11	5	12	80.5	5	7	12	9	80.5	5	7	9	3		9	10	3	5	
2017	03	29	2017-03-29?58期	三	5	3	7	8	12	7	3	2	1	28	1	3	4	2	28	3	1	7	4	28	3	7	4	5	28
2017	03	29	2017-03-29?58期	三	6	9	4	8	5	3	2	9	4	88.5	3	4	12	9	88.5	4	6	8	2		3	11	4	9	88.5
2017	03	29	2017-03-29?58期	三	7	11	7	6	9	7	1	9	11	89	8	7	11	9	89	10	9	1	7		9	6	7	1	
2017	03	29	2017-03-29?58期	三	8	11	12	8	9	4	12	1	8		11	12	8	2	109	8	12	11	4	109	12	8	11	2	109
2017	03	26	2017-03-26?57期	日	1	3	4	2	8	6	7	1	8		1	3	2	8	95	1	6	8	5		1	4	8	6	
2017	03	26	2017-03-26?57期	日	2	12	7	2	9	2	4	1	8		6	2	4	12	54.5	12	1	8	14	54.5	2	14	6	12	54.5
2017	03	26	2017-03-26?57期	日	3	1	11	12	13	8	11	13	2		1	8	11	12	24.5	8	1	2	12	24.5	11	8	1	12	24.5
2017	03	26	2017-03-26?57期	日	4	1	6	2	4	4	3	1	2	40.5	2	3	1	12	40.5	1	4	6	2	40.5	3	2	4	12	
2017	03	26	2017-03-26?57期	日	5	3	11	6	1	5	7	6	11		7	11	5	2		7	9	11	5		7	2	3	5	79
2017	03	26	2017-03-26?57期	日	6	14	8	2	9	4	6	3	13		10	4	3	6		6	1	3	10		4	3	10	6	
2017	03	26	2017-03-26?57期	日	7	6	10	1	7	7	6	5	8	32	8	5	7	6	32	7	2	6	9	32	5	7	3	1	
2017	03	26	2017-03-26?57期	日	8	12	1	4	6	12	6	10	5	53	1	4	6	10		12	10	6	8	53	1	3	10	4	
2017	03	26	2017-03-26?57期	日	9	14	11	1	4	13	2	14	12	35	8	2	14	9	35	10	14	8	2	35	8	14	10	13	35
2017	03	26	2017-03-26?57期	日	10	13	3	4	11	1	11	8	13	111.5	8	11	3	13	111.5	11	1	8	5		8	3	1	4	
2017	03	22	2017-03-22?56期	三	1	6	4	1	2	4	6	5	8	69.5	2	5	6	4	69.5	1	2	8	9		2	6	9	8	69.5
2017	03	22	2017-03-22?56期	三	2	1	4	11	9	3	1	6	11	18	1	6	7	9	18	6	1	4	2	18	1	4	6	3	18
2017	03	22	2017-03-22?56期	三	3	8	2	7	12	7	6	2	1		6	2	7	1		1	2	7	5		6	11	9	7	
2017	03	22	2017-03-22?56期	三	4	7	8	1	12	5	3	1	11		2	1	8	3		6	1	3	8		1	3	8	5	
2017	03	22	2017-03-22?56期	三	5	1	12	8	2	12	2	6	8		1	8	2	12	50.5	8	12	2	1	50.5	8	6	2	1	50.5
2017	03	22	2017-03-22?56期	三	6	10	2	6	12	9	10	4	11	36.5	3	10	4	9	36.5	9	10	3	8	36.5	3	4	12	10	36.5
2017	03	22	2017-03-22?56期	三	7	9	11	2	7	9	2	7	4	28	7	2	5	8		7	1	2	9	28	9	2	1	11	28

Figure 3.1.1 Photo of tips_ad in our database

	#	Name	Type	Collation	Attributes	Null	Default	C
☐	1	**raceyear** 🔑	smallint(1)			No	*None*	
☐	2	**racemonth**	tinytext	latin1_swedish_ci		No	*None*	
☐	3	**raceday**	tinytext	latin1_swedish_ci		No	*None*	
☐	4	**date**	tinytext	big5_bin		No	*None*	
☐	5	**week**	tinytext	big5_bin		No	*None*	
☐	6	**raceno**	tinyint(1)			No	*None*	
☐	7	**no1**	tinyint(1)			No	*None*	
☐	8	**no2**	tinyint(1)			No	*None*	
☐	9	**no3**	tinyint(1)			No	*None*	
☐	10	**no4**	tinyint(1)			No	*None*	
☐	11	**ad41_1**	tinytext	latin1_swedish_ci		No	*None*	
☐	12	**ad41_2**	tinytext	latin1_swedish_ci		No	*None*	
☐	13	**ad41_3**	tinytext	latin1_swedish_ci		No	*None*	
☐	14	**ad41_4**	tinytext	latin1_swedish_ci		No	*None*	
☐	15	**ad41_dividend**	tinytext	latin1_swedish_ci		No	*None*	
☐	16	**ad42_1**	tinytext	latin1_swedish_ci		No	*None*	
☐	17	**ad42_2**	tinytext	latin1_swedish_ci		No	*None*	
☐	18	**ad42_3**	tinytext	latin1_swedish_ci		No	*None*	
☐	19	**ad42_4**	tinytext	latin1_swedish_ci		No	*None*	
☐	20	**ad42_dividend**	tinytext	latin1_swedish_ci		No	*None*	
☐	21	**ad43_1**	tinytext	latin1_swedish_ci		No	*None*	
☐	22	**ad43_2**	tinytext	latin1_swedish_ci		No	*None*	
☐	23	**ad43_3**	tinytext	latin1_swedish_ci		No	*None*	
☐	24	**ad43_4**	tinytext	latin1_swedish_ci		No	*None*	
☐	25	**ad43_dividend**	tinytext	latin1_swedish_ci		No	*None*	
☐	26	**ad40_1**	tinytext	latin1_swedish_ci		No	*None*	
☐	27	**ad40_2**	tinytext	latin1_swedish_ci		No	*None*	
☐	28	**ad40_3**	tinytext	latin1_swedish_ci		No	*None*	
☐	29	**ad40_4**	tinytext	latin1_swedish_ci		No	*None*	
☐	30	**ad40_dividend**	tinytext	latin1_swedish_ci		No	*None*	

Figure 3.1.2 Table structure of 'tips_ad'

Race

The race table store the information of races collected from The Hong Kong Jockey Club, the tips from experts, and the predicted ranking of each horse. Each row in the table represent a horse, so there will be around ten rows for every race. Since we grab the data from HKJC starting from August 2011, there are total 53554 rows in this table. Figure 3.1.3 to 3.1.5 are the photo captured from a part of 'race' table. 'raceyear' to 'place_odd' columns are the detailed information of race, horse grabbed from the website of The Hong Kong Jockey Club. 'winodds' and 'place_odd' will be used for calculating the gain and loss later on. 'jj1' to 'sp58' contain 54 columns, the values are affected by the 'tips' tables. If the corresponding expert predict the horse, the value will be changed to '1' instead of '0'. Next, the 'result' column is storing the result of each horse racing. If the horse get first, second, or the third place, the value will be '1', otherwise it will be '0'. The data pre-processing will be talked in detailed in next part. Finally, 'prediction' to 'nonten_predict' columns are the prediction part. 'prediction' stores the result of TensorFlow. 'predict_place' is the descending order of 'prediction' for each race. 'over_rank' is the descending order of 'prediction' from March 2017 to recent. 'tips_cnt' is counting the voting number of the 4 chosen experts. 'nonten_predict' is the descending rank of 'tips_cnt' for each race. The values in 'tips_cnt' and 'nonten_predict' are used for the voting system. Figure 3.1.6 to 3.1.8 shows the structure of 'race' table.

raceyear	racemonth	raceday	raceid	location	class	distance	course	going	raceno	place	horseno	horseid	jockeycode	trainercode
2017	3	26	517	ST	5	2000	TURF - "A+3" COURSE	GOOD	1	1	3	S346	RN	TYS
2017	3	26	517	ST	5	2000	TURF - "A+3" COURSE	GOOD	1	2	4	P070	PB	LAL
2017	3	26	517	ST	5	2000	TURF - "A+3" COURSE	GOOD	1	3	2	P405	YML	LAL
2017	3	26	517	ST	5	2000	TURF - "A+3" COURSE	GOOD	1	4	8	S354	TEK	FD
2017	3	26	517	ST	5	2000	TURF - "A+3" COURSE	GOOD	1	5	7	T323	WD	CCW
2017	3	26	517	ST	5	2000	TURF - "A+3" COURSE	GOOD	1	6	1	V213	PZ	OSP
2017	3	26	517	ST	5	2000	TURF - "A+3" COURSE	GOOD	1	7	5	V060	CAN	YPF
2017	3	26	517	ST	5	2000	TURF - "A+3" COURSE	GOOD	1	8	10	P168	NGA	CD
2017	3	26	517	ST	5	2000	TURF - "A+3" COURSE	GOOD	1	9	6	P293	LDE	HL

Figure 3.1.3 Race table in database(1)

actualweight	horseweight	draw	lbw	rp1	rp2	rp3	rp4	rp5	rp6	finishtime	winodds	place_odd	jj1	jj2	jj3	jj4
128	1232	6	-	3	4	4	6	1	-	125.01	9.5	30.5	0	0	0	1
126	1011	11	1-1/2	10	6	6	5	2	-	125.25	12	33.0	0	1	0	0
129	1076	7	2	11	11	9	9	3	-	125.33	10	31.0	1	1	1	0
116	1055	2	2-1/4	5	7	8	8	4	-	125.37	7.7	0.0	0	0	0	0
119	1155	1	2-3/4	6	8	7	7	5	-	125.44	11	0.0	0	0	0	0
133	1193	10	3-1/4	7	3	2	3	6	-	125.54	2.5	0.0	1	0	1	1
124	1103	8	3-1/2	2	2	3	2	7	-	125.59	7.2	0.0	0	0	0	0
108	1157	4	3-3/4	9	10	10	10	8	-	125.62	99	0.0	0	0	0	0
121	1146	9	5-1/2	4	5	5	4	9	-	125.88	9	0.0	1	1	1	1
108	1076	3	11-3/4	1	1	1	1	10	-	126.88	50	0.0	0	0	0	0

Figure 3.1.4 Race table in database(2)

sp54	sp55	sp56	sp57	sp58	result	prediction	predict_place	overall_rank	tips_cnt	nonten_predict
0	1	1	1	0	1	-0.01756	1	25	2	3
1	0	0	0	0	1	-0.280431	2	65	0	6
0	1	0	1	0	1	-0.472978	4	26	3	2
0	0	0	0	1	-1	-0.716142	7	73	0	7
0	0	0	0	1	-1	-0.737367	8	118	0	8
1	0	1	1	0	-1	-0.699956	6	9	4	1
0	0	0	0	0	-1	-0.789308	9	72	0	9
0	0	0	0	0	-1	-0.922357	11	125	0	10
1	1	1	0	1	-1	-0.457349	3	43	2	4

Figure 3.1.5 Race table in database(3)

Name	Type	Collation	Attributes	Null	Default
raceyear	int(11)			No	*None*
racemonth	int(11)			No	*None*
raceday	int(11)			No	*None*
raceid	int(11)			No	*None*
location	char(2)	utf8_general_ci		No	*None*
class	int(11)			No	*None*
distance	int(11)			No	*None*
course	char(24)	utf8_general_ci		No	*None*
going	char(24)	utf8_general_ci		No	*None*
raceno	int(11)			No	*None*
place	char(5)	utf8_general_ci		No	*None*
horseno	int(11)			No	*None*
horseid	char(5)	utf8_general_ci		No	*None*
jockeycode	char(5)	utf8_general_ci		No	*None*
trainercode	char(5)	utf8_general_ci		No	*None*
actualweight	int(11)			No	*None*
horseweight	int(11)			No	*None*

Figure 3.1.6 Table structure of race(1)

draw	char(5)	utf8_general_ci	No	*None*
lbw	char(24)	utf8_general_ci	No	*None*
rp1	char(5)	utf8_general_ci	No	*None*
rp2	char(5)	utf8_general_ci	No	*None*
rp3	char(5)	utf8_general_ci	No	*None*
rp4	char(5)	utf8_general_ci	No	*None*
rp5	char(5)	utf8_general_ci	No	*None*
rp6	char(5)	utf8_general_ci	No	*None*
finishtime	float		No	*None*
winodds	float		No	*None*
place_odd	decimal(10,1)		No	0.0
jj1	int(11)		No	0
jj2	int(11)		No	0
jj3	int(11)		No	0
jj4	int(11)		No	0

Figure 3.1.7 Table structure of race(2)

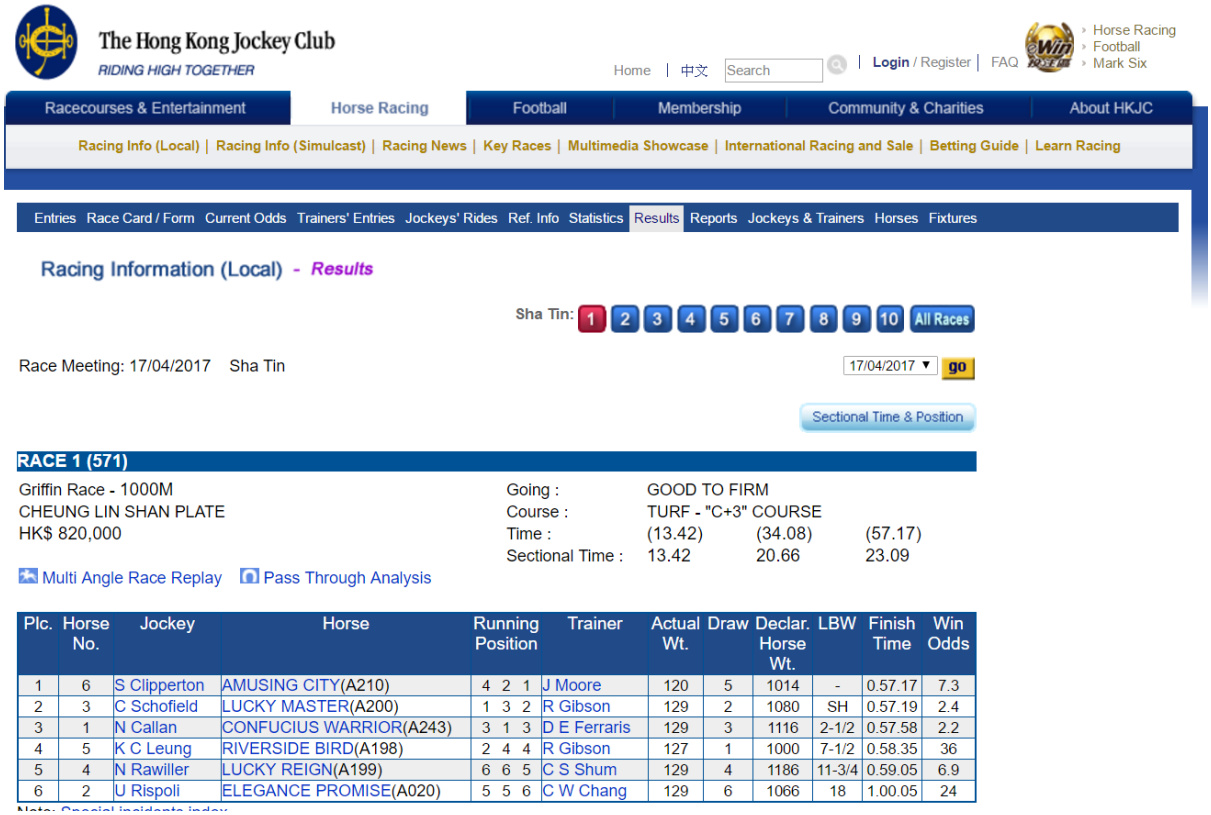
sp54	int(11)	No	0
sp55	int(11)	No	0
sp56	int(11)	No	0
sp57	int(11)	No	0
sp58	int(11)	No	0
result	int(11)	No	-1
prediction	float	No	-99
predict_place	int(11)	No	-99
overall_rank	int(11)	No	-99
tips_cnt	int(11)	No	-99
nonten_predict	int(11)	No	-99

Figure 3.1.8 Table structure of race(3)

3.2 Data Collection and Pre-processing

Before calculating and predicting the result, we need to prepare a lot of data. The data can be separated in two parts, the horse racing result and the tips. We first collect the horse racing result from the webpage of The Hong Kong Jockey Club.

Data Collection of Horse Racing Result



Plc.	Horse No.	Jockey	Horse	Running Position	Trainer	Actual Wt.	Draw	Declar. Horse Wt.	LBW	Finish Time	Win Odds
1	6	S Clipperton	AMUSING CITY(A210)	4 2 1	J Moore	120	5	1014	-	0.57.17	7.3
2	3	C Schofield	LUCKY MASTER(A200)	1 3 2	R Gibson	129	2	1080	SH	0.57.19	2.4
3	1	N Callan	CONFUCIUS WARRIOR(A243)	3 1 3	D E Ferraris	129	3	1116	2-1/2	0.57.58	2.2
4	5	K C Leung	RIVERSIDE BIRD(A198)	2 4 4	R Gibson	127	1	1000	7-1/2	0.58.35	36
5	4	N Rawiller	LUCKY REIGN(A199)	6 6 5	C S Shum	129	4	1186	11-3/4	0.59.05	6.9
6	2	U Rispoli	ELEGANCE PROMISE(A020)	5 5 6	C W Chang	129	6	1066	18	1.00.05	24

The horse racing results page in the website of HKJC

We write a parser in PHP to scrape the results from August 2011 to recent, and update the data to the database automatically. First of all, we find that we can change the URL to access every races. For example, the URL:

http://racing.hkjc.com/racing/Info/meeting/Results/english/Local/20170417/ST/1,

‘20170417’ represents the date 17th April 2017, ‘ST’ represents Sha Tin, and ‘1’ represents race number 1. So, we write some loops in PHP to get the results starting from August 2011,

```php
$baseurl = 'http://racing.hkjc.com/racing/info/Meeting/Results/English/Local/';
$locations = array("ST", "HV");
//$location = 'HV';
$racenos  = array(1, 2, 3, 4, 5, 6,7,8,9,10,11,12);
//$raceno = 4;
$trys  = array(1, 2, 3);
//$try = 1;


//set start day
$startdate = new DateTime('2011-08-15');
//get latest start day from database
$db_con=new mysqli(DB_HOST,DB_USERNAME,DB_PASSWORD,DB_NAME);
if($db_con)
{
    $result=mysqli_query($db_con, "SELECT raceyear, racemonth, raceday FROM race order by raceyear desc, racemonth desc, raceday desc");
    $row=mysqli_fetch_array($result);
    if (mysqli_num_rows($result)>=1)
    {
        $startdate=new Datetime($row['raceyear']."-".$row['racemonth']."-".$row['raceday']);

    };
    mysqli_close($db_con);
};

//set end day
//$enddate = new DateTime('2017-02-02');
$enddate = new DateTime('NOW');

$currentdate = $startdate;

while ($currentdate <= $enddate) {

    foreach( $trys as $try) {

    foreach ($locations as $location) {
        foreach ($racenos as $raceno) {...

        } // loop raceno
    } // loop location
    } // loop try

    $currentdate->add(new DateInterval('P1D'));
} // loop date
```

and skip the date without horse racing.

PHP code for looping the horse racing result

After our program can loop all combination of date, location and race number, we need to search the race information in the HTML code, for example the distance, horse number, jockey code, and etc. So, we use the PHP code to explode the webpage’s HTML by searching some keywords, then save the values to some variables. The below PHP code shows our method to get some of the information.

```php
case 1: //Place
    list($dump, $dump2) = explode('>', $text);
    list($dump, $dump2) = explode('<', $dump2);
    //list($place, $dump2) = explode(' ', $dump);
    $place = $dump;
    //echo "Place: ".$place."\n";
    break;
case 2: //HorseNo
    list($dump, $dump2) = explode('>', $text);
    list($dump, $dump2) = explode('<', $dump2);
    //list($horseno, $dump2) = explode(' ', $dump);
    $horseno = $dump;
    //echo "HorseNo: ".$horseno."\n";
    break;
case 3:
    list($dump, $dump2) = explode('(', $text);
    list($dump, $dump2) = explode(')', $dump2);
    //list($horseid, $dump2) = explode(' ', $dump);
    $horseid = $dump;
    //echo "HorseID: ".$horseid."\n";
    break;
case 4: //JockeyCode
    $as = $td->getElementsByTagName('a');
    foreach ($as as $a) {
        list($head, $tail) = explode('?', $a->getAttribute("href"));
        $args =  explode('&', $tail);
        foreach ($args as $arg) {
            if (preg_match('/jockeycode/', $arg)){
                list($head, $jockeycode) = explode('=', $arg);
            };
        };
        //echo "JockeyCode: ".$jockeycode."\n";
    }
    break;
case 5: //TrainerCode
    $as = $td->getElementsByTagName('a');
    foreach ($as as $a) {
        list($head, $tail) = explode('?', $a->getAttribute("href"));
        $args =  explode('&', $tail);
        foreach ($args as $arg) {
            if (preg_match('/trainercode/', $arg)){
                list($head, $trainercode) = explode('=', $arg);
            };
        };
        //echo "TrainerCode: ".$trainercode."\n";
    }
    break;
```

PHP code for searching the race information in HTML

Finally, after all needed information are scraped, we insert the data to ‘race’ table in

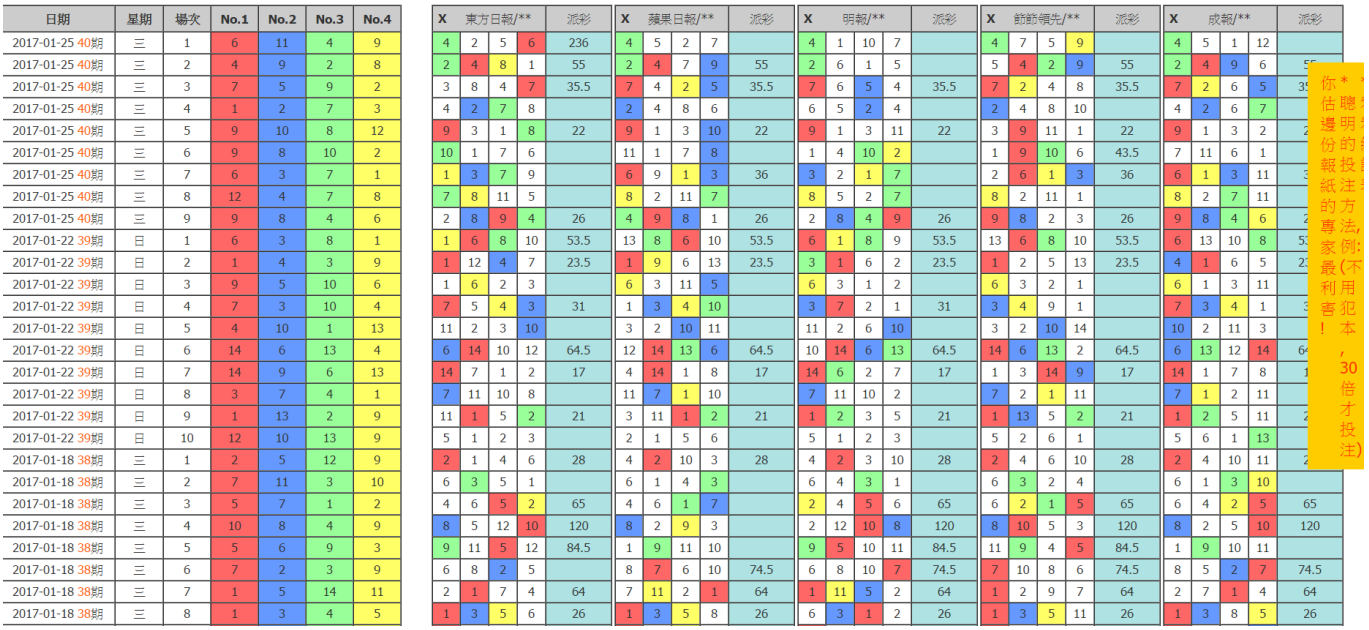
```php
$db_con=new mysqli(DB_HOST,DB_USERNAME,DB_PASSWORD,DB_NAME);
if($db_con)
{
    $result=mysqli_query($db_con, "SELECT raceyear, racemonth, raceday, raceid, horseno FROM race where raceyear =".$currentyear." and racemonth=".$currentmonth." and
        raceday=".$currentday." and raceid=".$raceid." and horseno=".$horseno);
    $row=mysqli_fetch_array($result);
    if (mysqli_num_rows($result)==0)
    {
        $stmt = $db_con->prepare("INSERT INTO race (raceyear, racemonth, raceday, raceid, location, class, distance, course, going, raceno, place, horseno, horseid,
            jockeycode, trainercode, actualweight, horseweight, draw, lbw, rp1, rp2, rp3, rp4, rp5, rp6, finishtime, winodds) values
            (?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?)");
        if($stmt)
        {
            $stmt->bind_param("ddddsddssdsdsssddddsssssssdd",$currentyear,$currentmonth,$currentday,$raceid, $location, $class, $distance, $course, $going, $raceno, $place
                , $horseno, $horseid, $jockeycode, $trainercode, $actualweight, $horseweight, $draw, $lbw, $runningposition1, $runningposition2, $runningposition3, $
                runningposition4, $runningposition5, $runningposition6, $finishtime, $winodds);
            $stmt->execute();
            $lasterr = mysqli_error($db_con);
            echo $lasterr;
            mysqli_close($db_con);
        };
    };
};
```

our database using SQL.

PHP code for inserting the data to database

Data Collection of Tips

The collection of tips is the first difficulty we have faced in this project, because we do not know where the tips can be found at the beginning. Our original plan is to release an application to collect tips. It is what we have done in the first semester. However, we find that it is so difficult to attract user to use, and the data may be meaningless. So, in this semester, we change our direction to get tips provided by experts from different newspapers. After a long period of searching, we find a webpage provides a lot of horse racing tips from 8 newspaper starting from September 2016. This part of data collection is more simple than the previous part. We save the data grouping by each newspaper in CSV format, then directly insert the data to ‘tips’ tables in our database.



日期	星期	場次	No.1	No.2	No.3	No.4
2017-01-25 40期	三	1	6	11	4	9
2017-01-25 40期	三	2	4	9	2	8
2017-01-25 40期	三	3	7	5	9	2
2017-01-25 40期	三	4	1	2	7	3
2017-01-25 40期	三	5	9	10	8	12
2017-01-25 40期	三	6	9	8	10	2
2017-01-25 40期	三	7	6	3	7	1
2017-01-25 40期	三	8	12	4	7	8
2017-01-25 40期	三	9	9	8	4	6
2017-01-22 39期	日	1	6	3	8	1
2017-01-22 39期	日	2	1	4	3	9
2017-01-22 39期	日	3	9	5	10	6
2017-01-22 39期	日	4	7	3	10	4
2017-01-22 39期	日	5	4	10	1	13
2017-01-22 39期	日	6	14	6	13	4
2017-01-22 39期	日	7	14	9	6	13
2017-01-22 39期	日	8	3	7	4	1
2017-01-22 39期	日	9	1	13	2	9
2017-01-22 39期	日	10	12	10	13	9
2017-01-18 38期	三	1	2	5	12	9
2017-01-18 38期	三	2	7	11	3	10
2017-01-18 38期	三	3	5	7	1	2
2017-01-18 38期	三	4	10	8	4	9
2017-01-18 38期	三	5	5	6	9	3
2017-01-18 38期	三	6	7	2	3	9
2017-01-18 38期	三	7	1	5	14	11
2017-01-18 38期	三	8	1	3	4	5

X	東方日報/**				派彩
4	2	5	6	236	
2	4	8	1	55	
3	8	4	7	35.5	
4	2	7	8		
9	3	1	8	22	
10	1	7	6		
1	3	7	9		
7	8	11	5		
2	8	9	4	26	
1	6	8	10	53.5	
1	12	4	7	23.5	
1	6	2	3		
7	5	4	3	31	
11	2	3	10		
6	14	10	12	64.5	
14	7	1	2	17	
7	11	10	8		
11	1	5	2	21	
5	1	2	3		
2	1	4	6	28	
6	3	5	1		
4	6	5	2	65	
8	5	12	10	120	
9	11	5	12	84.5	
6	8	2	5		
2	1	7	4	64	
1	3	5	6	26	

X	蘋果日報/**				派彩
4	5	2	7		
2	4	7	9	55	
7	4	2	5	35.5	
2	4	8	6		
9	1	3	10	22	
11	1	7	8		
6	9	1	3	36	
8	2	11	7		
4	9	8	1	26	
13	8	6	10	53.5	
1	9	6	13	23.5	
6	3	11	5		
1	3	4	10		
3	2	10	11		
12	14	13	6	64.5	
4	14	1	8	17	
11	7	1	10		
3	11	1	2	21	
2	1	5	6		
4	2	10	3	28	
6	1	4	3		
4	6	1	7		
8	2	9	3		
1	9	11	10		
8	7	6	10	74.5	
7	11	2	1	64	
1	3	5	8	26	

X	明報/**				派彩
4	1	10	7		
2	6	1	5		
7	6	5	4	35.5	
6	5	2	4		
9	1	3	11	22	
1	4	10	2		
3	2	1	7		
8	5	2	7		
2	8	4	9	26	
6	1	8	9	53.5	
3	1	6	2	23.5	
6	3	1	2		
3	7	2	1	31	
11	2	6	10		
10	14	6	13	64.5	
14	6	2	7	17	
7	11	10	2		
1	2	3	5	21	
5	1	2	3		
4	2	3	10	28	
6	4	3	1		
2	4	5	6	65	
2	12	10	8	120	
9	5	10	11	84.5	
6	8	10	7	74.5	
1	11	5	2	64	
6	3	1	2	26	

X	節節領先/**				派彩
4	7	5	9		
5	4	2	9	55	
7	2	4	8	35.5	
2	4	8	10		
3	9	11	1	22	
1	9	10	6	43.5	
2	6	1	3	36	
8	2	11	1		
9	8	2	3	26	
13	6	8	10	53.5	
1	2	5	13	23.5	
6	3	2	1		
3	4	9	1		
3	2	10	14		
14	6	13	2	64.5	
1	3	14	9	17	
7	2	1	11		
1	13	5	2	21	
5	2	6	1		
2	4	6	10	28	
6	3	2	4		
6	2	1	5	65	
8	10	5	3	120	
11	9	4	5	84.5	
7	10	8	6	74.5	
1	2	9	7	64	
1	3	5	11	26	

X	成報/**				派彩
4	5	1	12		
2	4	9	6	55	
7	2	6	5	35.5	
4	2	6	7		
9	1	3	2	22	
7	11	6	1		
6	1	3	11	36	
8	2	7	11		
9	8	4	6	26	
6	13	10	8	53.5	
4	1	6	5	23.5	
6	1	3	11		
7	3	4	1		
10	2	11	3		
6	13	12	14	64	
14	1	7	8	17	
7	1	2	11		
1	2	5	11	21	
5	6	1	13		
2	4	10	11	28	
6	1	3	10		
6	4	2	5	65	
8	2	5	10	120	
1	9	10	11		
2	7	1	4	64	
1	3	8	5	26	

A part of raw data collected from the internet

Data Pre-processing

The raw data collected cannot be used for calculation immediately. Before training and predicting the data, we need to convert the data to a appropriate format. Since the tips data collected are some horse numbers, TensorFlow may not be able to predict a meaningful result. As we mentioned before, we have collected the horse racing results. Each row in the 'race' table represent a horse. So, we add 54 columns to the race table for all experts, for example 'od28', 'ad42', and etc. All values in the 54 columns are '0' at the beginning. If the experts predict the horse will win in that race, we update the value to '1'. Besides, we decide to consider the first three tips of each experts only, since we win place when the horse get first, second or third place in that race. The fourth horse is meaningless for us. In other words, each row in 'race' contain three '1's and fifty one '0's. We use PHP and SQL to do this data pre-processing. We use the following code to update 'race' according to expert 'ad42':

```
$result=mysqli_query($db_con, "UPDATE race race,tips_ad tips_ad  SET race.ad42=1
```

```
where race.raceyear= tips_ad.raceyear and  race.racemonth= tips_ad.racemonth and
race.raceday= tips_ad.raceday and  race.raceno= tips_ad.raceno and  (race.horseno=
tips_ad.ad42_1 or race.horseno= tips_ad.ad42_2 or race.horseno= tips_ad.ad42_3)");
```

We used PHP to update the values from expert ‘jj1’ to expert ‘sp58’.

After that, we assign ‘result’ column in ‘race’ table to be our predict column for TensorFlow. The horses win the race if they get first, second or third position. So, we assign the value ‘1’ to ‘result’ column of the winning horses and ‘-1’ for the losing horses. There are total three ‘1’s for each race and the remaining horses got ‘-1’s. So, we set ‘-1’ to be the default value and use the following SQL to change the winning horses to ‘1’:

```
UPDATE race SET result=1 where place<=3
```

Now, we can train and predict the result by TensorFlow using the data. However, there is one more step of data pre-processing for doing our voting system. We have added a column ‘tips_cnt’ to ‘race’ to count the number of voting. We set ‘0’ to be the default value for the column, then use the following SQL to count the number of ‘1’ from the chosen experts:

```
UPDATE `raceold` SET `tips_cnt`=4 WHERE `st47`='1' and `ad42`='1' and `od28`='1' and
`sp51`='1'

UPDATE `raceold` SET `tips_cnt`=3 WHERE (`st47`='0' and `ad42`='1' and `od28`='1'
and `sp51`='1') or (`st47`='1' and `ad42`='0' and `od28`='1' and `sp51`='1') or (`st47`='1'
and `ad42`='1' and `od28`='0' and `sp51`='1') or (`st47`='1' and `ad42`='1' and `od28`='1'
```

```
and `sp51`='0')


UPDATE `raceold` SET `tips_cnt`=2 WHERE (`st47`='0' and `ad42`='0' and `od28`='1'
and `sp51`='1') or (`st47`='0' and `ad42`='1' and `od28`='0' and `sp51`='1') or (`st47`='0'
and `ad42`='1' and `od28`='1' and `sp51`='0') or (`st47`='1' and `ad42`='0' and `od28`='0'
and `sp51`='1') or (`st47`='1' and `ad42`='0' and `od28`='1' and `sp51`='0') or (`st47`='1'
and `ad42`='1' and `od28`='0' and `sp51`='0')


UPDATE `raceold` SET `tips_cnt`=1 WHERE (`st47`='1' and `ad42`='0' and `od28`='0'
and `sp51`='0') or (`st47`='0' and `ad42`='1' and `od28`='0' and `sp51`='0') or (`st47`='0'
and `ad42`='0' and `od28`='1' and `sp51`='0') or (`st47`='0' and `ad42`='0' and `od28`='0'
and `sp51`='1')
```

After that, we can sort the value of ‘tips_cnt’ and get the predicted place of the voting system.

3.3 Voting System Design

As we know the predictions of experts in 569 races, our voting system counts how many experts predict on a particular horse, and we choose the horse with highest vote as winner compared with other horses in the same race, and we can get a ranking from it. For example we use the best 4 experts' tips as input, in 1 Jan 2017 race 3, we can see that in total there are 4 experts predicted horse no.2, 3 experts predicted horse no.7 and 2 experts predicted horse no. 6, so their corresponding predicted rank are 1, 2 and 3 respectively.

raceyear ▲ 1	racemonth ▲ 2	raceday ▲ 3	location	raceno ▲ 4	place	horseno	horseid	finishtime	winodds	tips_cnt ▼ 5	nonten_predict
2017	1	1	ST	3	1	2	T202	83	2.2	4	1
2017	1	1	ST	3	2	7	S070	83.17	10	3	2
2017	1	1	ST	3	3	6	T339	83.25	13	2	3
2017	1	1	ST	3	4	3	S234	83.27	11	2	4
2017	1	1	ST	3	8	9	S320	83.69	15	1	5
2017	1	1	ST	3	5	8	V129	83.39	11	0	6
2017	1	1	ST	3	6	11	S051	83.49	72	0	7
2017	1	1	ST	3	7	10	T393	83.61	6.2	0	8
2017	1	1	ST	3	9	12	T263	83.75	74	0	9
2017	1	1	ST	3	10	13	S097	84.41	62	0	10
2017	1	1	ST	3	11	1	T300	84.64	21	0	11
2017	1	1	ST	3	12	5	V160	86.03	99	0	12
2017	1	1	ST	3	13	14	T067	87.59	48	0	13

photo of showing tips count and corresponding prediction in our database

In our research, we only focus on betting win and place, so we only care about the first three predictions, but there is a special case when the tips count are the same, mathematically speaking, there are more than one mode in one race, in our voting system, we choose the one with lower horse number, and it is not random but actually proved by statistic, we found those races with more than one mode, and checked the percentage of actually winning on lower horse number versus actually winning on higher horse number. There are

89% of this situation are actually winning on lower horse number, and 11% of this situation are actually winning on higher horse number.

3.4 TensorFlow Design

In recent years, machine learning has driven advances in many different fields [3 4 24 25 29]. We think that machine learning can be a tool for us to predict horse racing result since it can discover some pattern which cannot be easily found by human.

TensorFlow is a machine learning system, based on deep learning neural networks(HOPE, 2017). It operates at large scale and in heterogeneous environments. It uses dataflow graphs to represent computation, shared state, and the operations that mutate that state. We have two steps to do, first step is training, and the second is predicting.

Tips presentation on TensorFlow

We got many tips already, and each tips record contains 4 horse number(predicted winner, predicted no.2, predicted no.3 and predicted no.4), which is not match to our horse racing database, so we need to do some transformation of tips. In the following photo, it shows 4 different experts(ad41, ad42,ad43 and ad40) and their tips.

raceyear	racemonth	raceday	date	week	raceno	no1	no2	no3	no4	ad41_1	ad41_2	ad41_3	ad41_4	ad41_dividend	ad42_1	ad42_2	ad42_3	ad42_4	ad42_dividend	ad43_1	ad43_2	ad43_3	ad43_4	ad43_dividend	ad40_1	ad40_2	ad40_3	ad40_4	ad40_dividend
2017	03	29	2017-03-29?58期	三	1	2	6	10	5	6	3	8	5		6	3	2	1	233.5	6	3	4	5		1	3	6	9	
2017	03	29	2017-03-29?58期	三	2	9	4	1	6	2	6	10	7		4	1	6	3		1	6	4	7		6	8	2	1	
2017	03	29	2017-03-29?58期	三	3	7	11	4	5	2	3	9	5		2	11	3	6		11	5	2	7	82.5	11	2	6	4	
2017	03	29	2017-03-29?58期	三	4	12	3	10	2	9	11	5	12	80.5	5	7	12	9	80.5	5	7	9	3		9	10	3	5	
2017	03	29	2017-03-29?58期	三	5	3	7	8	12	7	3	2	1	28	1	3	4	2	28	3	1	7	4	28	3	7	4	5	28
2017	03	29	2017-03-29?58期	三	6	9	4	8	5	3	2	9	4	88.5	3	4	12	9	88.5	4	6	8	2		3	11	4	9	88.5
2017	03	29	2017-03-29?58期	三	7	11	7	6	9	7	1	9	11	89	8	7	11	9	89	10	9	1	7		9	6	7	1	
2017	03	29	2017-03-29?58期	三	8	11	12	6	9	4	12	1	8		11	12	8	2	109	8	12	11	4	109	12	8	11	2	109
2017	03	26	2017-03-26?57期	日	1	3	4	2	8	6	7	1	8		1	3	2	8	95	1	6	8	5		1	4	8	6	
2017	03	26	2017-03-26?57期	日	2	12	7	2	9	2	4	1	8		6	2	4	12	54.5	12	1	8	14	54.5	2	14	6	12	54.5
2017	03	26	2017-03-26?57期	日	3	1	11	12	13	8	11	13	2		1	8	11	12	24.5	8	1	2	12	24.5	11	8	1	12	24.5
2017	03	26	2017-03-26?57期	日	4	1	6	2	4	4	3	1	2	40.5	2	3	1	12	40.5	1	4	6	2	40.5	3	2	4	12	
2017	03	26	2017-03-26?57期	日	5	3	11	6	1	5	7	6	11		7	11	5	2		7	9	11	5		7	2	3	5	79
2017	03	26	2017-03-26?57期	日	6	14	8	2	9	4	6	3	13		10	4	3	6		6	1	3	10		4	3	10	6	
2017	03	26	2017-03-26?57期	日	7	6	10	1	7	7	6	5	8	32	8	5	7	6	32	7	2	6	9	32	5	7	3	1	
2017	03	26	2017-03-26?57期	日	8	12	1	4	6	12	6	10	5	53	1	4	6	10		12	10	6	8	53	1	3	10	4	
2017	03	26	2017-03-26?57期	日	9	14	11	1	4	13	2	14	12	35	8	2	14	9	35	10	14	8	2	35	8	14	10	13	35
2017	03	26	2017-03-26?57期	日	10	13	3	4	11	1	11	8	13	111.5	8	11	3	13	111.5	11	1	8	5		8	3	1	4	
2017	03	22	2017-03-22?56期	三	1	6	4	1	2	4	6	5	8	69.5	2	5	6	4	69.5	1	2	8	9		2	6	9	8	69.5
2017	03	22	2017-03-22?56期	三	2	1	4	11	9	3	1	6	11	18	1	6	7	9	18	6	1	4	2	18	1	4	6	3	18
2017	03	22	2017-03-22?56期	三	3	8	2	7	12	7	6	2	1		6	2	7	1		1	2	7	5		6	11	9	7	
2017	03	22	2017-03-22?56期	三	4	7	8	1	12	5	3	1	11		2	1	8	3		6	1	3	8		1	3	8	5	
2017	03	22	2017-03-22?56期	三	5	1	12	8	2	12	2	6	8		1	8	2	12	50.5	8	12	2	1	50.5	8	6	2	1	50.5
2017	03	22	2017-03-22?56期	三	6	10	2	6	12	9	10	4	11	36.5	3	10	4	9	36.5	9	10	3	8	36.5	3	4	12	10	36.5
2017	03	22	2017-03-22?56期	三	7	9	11	2	7	9	2	7	4	28	7	2	5	8		7	1	2	9	28	9	2	1	11	28

photo of tips from 4 experts in our database

To transform it to our horse racing database, we added all the experts' id in horse racing database as new column, if the horse number in tips database matches the particular

horse in horse database, we add value 1 on that record in the corresponding expert column. Also, we added column result, which is 1 if the place of that horse is first three, and -1 if the place of that horse is higher than 3.

One important thing is that we ignore the fourth prediction, because in our research, we mainly focus on betting on win and place, the predicted no.4 horse is meaningless to us, so that in our transformation of tips, we only check the first three horse in tips database.

In the following photo, it shows part of all experts' tips on the horse racing database.

raceyear	racemonth	raceday	raceid	location	distance	raceno	place	horseno	horseid	finishtime	winodds	jj1	jj2	hn7	hn8	hn9	hn10	hn11	ph16	ph18	ph19	sf25	sf26	od28	od39	ad40	ad41	st45	st46	st47	st48	mp49	sp56	sp58	result
2017	1	1	297	ST	2000	1	1	2	P272	123.16	9.7	1	0	1	1	1	1	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0	1
2017	1	1	297	ST	2000	1	2	9	P293	123.31	10	0	0	1	1	0	1	1	1	0	0	1	1	1	1	0	0	0	0	0	1	1	0	0	1
2017	1	1	297	ST	2000	1	3	13	P415	123.82	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2017	1	1	297	ST	2000	1	4	11	S354	123.88	8.3	0	0	0	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0	-1
2017	1	1	297	ST	2000	1	5	1	P088	124.03	15	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	-1
2017	1	1	297	ST	2000	1	6	4	S205	124.1	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1
2017	1	1	297	ST	2000	1	7	3	P405	124.13	7.9	1	1	0	1	0	0	0	0	1	0	0	0	1	1	1	1	0	1	0	0	0	1	0	-1
2017	1	1	297	ST	2000	1	8	5	T053	124.16	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1
2017	1	1	297	ST	2000	1	9	8	S007	124.45	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1
2017	1	1	297	ST	2000	1	10	12	V061	124.73	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	-1
2017	1	1	297	ST	2000	1	11	7	T323	125.66	3.7	1	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-1
2017	1	1	297	ST	2000	1	12	10	T266	126.12	92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1
2017	1	1	297	ST	2000	1	13	6	S408	134.75	8.6	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	-1
2017	1	1	298	ST	1000	2	1	2	T099	57.53	6.6	1	1	1	1	0	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	1	1	1	1
2017	1	1	298	ST	1000	2	2	11	T345	57.64	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2017	1	1	298	ST	1000	2	3	13	T073	57.8	10	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1
2017	1	1	298	ST	1000	2	4	4	V306	57.82	5.5	0	0	0	1	0	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	0	1	1	-1
2017	1	1	298	ST	1000	2	5	1	T422	57.92	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1
2017	1	1	298	ST	1000	2	6	12	T091	57.92	17	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	-1
2017	1	1	298	ST	1000	2	7	6	A031	57.98	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1
2017	1	1	298	ST	1000	2	8	5	A007	58.01	2.2	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-1
2017	1	1	298	ST	1000	2	9	8	A075	58.07	48	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	-1
2017	1	1	298	ST	1000	2	10	14	V317	58.73	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1
2017	1	1	298	ST	1000	2	11	3	V312	58.73	62	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	-1
2017	1	1	298	ST	1000	2	12	9	S011	59.03	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1

photo of transformed tips to horse racing database

Training features

To do our TensorFlow training, we need to prepare input data for training, in our database, each record represents one horse in a race as we mentioned before. There are many information such as 'race year, race month, race day, location, distance, course, going, horse id, jockey code, trainer code, actual weight, horse weight, draw, win odds, finish time etc'.

However, our purpose is want to use tips as the basic, and less rely on the race information, so we tried out many different combination of features.

At first, we use all the information and 54 experts as input, but the error is quite high, it is obviously not working, so we modify some features, and then we tried to remove race

year, race month and race day, and it showed that it is better than before, but the predictions are still really inaccurate, we thought that actually most of the experts are not really good, so that we consider to do a ranking between individual experts, and we choose only the best 4 experts as our TensorFlow input, and the result showed a big improvement.

Finally, we chose ‘distance, course, going, jockey code, trainer code, actual weight, horse weight, 4 experts’ tips and result’ as the TensorFlow input.

Number of training races and predicting races

Because we have tips from experts on 569 races, we choose to use 452 races as input to train on TensorFlow, and we use the remaining 117 as predictions.

TensorFlow model configuration

TensorFlow has neural networks, and the size of it may affect the accuracy of predictions, we tried 32, 64 and 128 neural networks, and the result showed that 128 is the best for our research.

The following table shows the improvement of TensorFlow model with different configuration:

neural networks	features	error
16	race year, race month, race day, location, distance, course, going, horseid, jockey code, trainer code, draw, 54 experts	1.279
32	race year, race month, race day, location, distance, course, going, horseid, jockey code, trainer code, draw, 54 experts	1.029
32	race year, race month, race day, location, distance, course, going, horseid, jockey code, trainer code, draw, 4 experts	0.720

64	distance, course, going, jockey code, trainer code, actual weight, horse weight, 4 experts	0.688
128	distance, course, going, jockey code, trainer code, actual weight, horse weight, 4 experts	0.624

TensorFlow prediction

At this point, we selected the features and created our TensorFlow model, how can we make predictions? We have mentioned that we added 'result' column in the horse racing database, it has 1 or -1 value, it actually our prediction column in TensorFlow, after we input the basic information of that race and the 4 experts' tips, the TensorFlow model will output some numbers from -1 to 1, we choose the highest one as winner, rank the predicted numbers as ranking, because more near 1 means it is likely to win, and near -1 means it is more likely to lose. In the following photo, we can see that the prediction column, the numbers in that

raceyear ▲ 1	racemonth ▲ 2	raceday ▲ 3	location	distance	course	going	raceno ▲ 4	place	horseno	jockeycode	trainercode	actualweight	horseweight	winodds	od28	ad42	st47	sp51	result	prediction	predict_place
2017	4	9	ST	1200	TURF - "C" COURSE	GOOD TO FIRM	4	1	8	MHT	TYS	112	1179	2.2	1	1	1	1	1	0.150422	1
2017	4	9	ST	1200	TURF - "C" COURSE	GOOD TO FIRM	4	2	6	MOJ	SJJ	124	1116	3.8	1	1	0	0	1	-0.013719	2
2017	4	9	ST	1200	TURF - "C" COURSE	GOOD TO FIRM	4	3	5	CAN	SCS	125	1154	8.1	0	1	1	1	1	-0.034013	3
2017	4	9	ST	1200	TURF - "C" COURSE	GOOD TO FIRM	4	4	7	TEK	MKL	122	1057	10	1	0	0	1	-1	-0.284753	4
2017	4	9	ST	1200	TURF - "C" COURSE	GOOD TO FIRM	4	5	10	LHW	MA	118	1048	36	0	0	0	0	-1	-0.85967	12
2017	4	9	ST	1200	TURF - "C" COURSE	GOOD TO FIRM	4	6	1	CSF	JM	133	1128	13	0	0	0	0	-1	-0.772455	10
2017	4	9	ST	1200	TURF - "C" COURSE	GOOD TO FIRM	4	7	2	CKK	YCH	127	1106	30	0	0	1	0	-1	-0.596704	6
2017	4	9	ST	1200	TURF - "C" COURSE	GOOD TO FIRM	4	8	9	DO	CCW	119	1067	12	0	0	0	0	-1	-0.652511	8
2017	4	9	ST	1200	TURF - "C" COURSE	GOOD TO FIRM	4	9	11	LDE	YPF	114	1097	99	0	0	0	0	-1	-0.608304	7
2017	4	9	ST	1200	TURF - "C" COURSE	GOOD TO FIRM	4	10	12	NGA	CAS	111	1142	40	0	0	0	0	-1	-0.672555	9
2017	4	9	ST	1200	TURF - "C" COURSE	GOOD TO FIRM	4	11	3	STH	FD	128	1039	99	0	0	0	0	-1	-0.782188	11
2017	4	9	ST	1200	TURF - "C" COURSE	GOOD TO FIRM	4	12	4	PZ	FC	128	1097	38	0	0	0	0	-1	-0.490287	5

column are predicted from TensorFlow, and predict_place column is the ranking in one race.

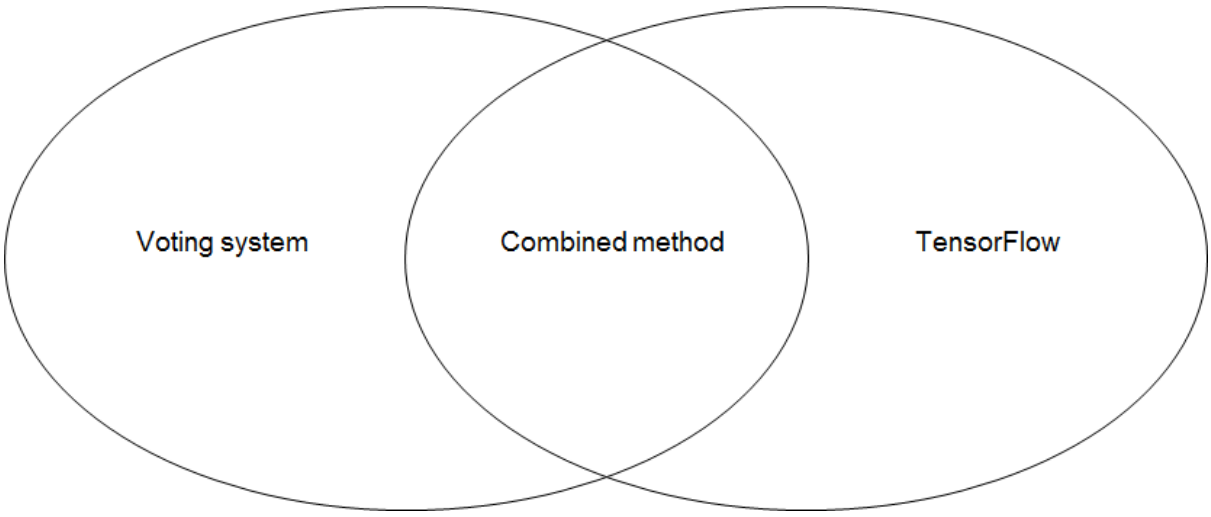
photo of horse racing database after TensorFlow prediction

3.5 Combine of TensorFlow and voting system

This combined method is a kind of filtering method, we only bet on the horse which is both agreed by TensorFlow and voting system. That means we only choose to bet when that horse is predicted by both methods. In the following photo, the predict_place(TensorFlow) column and nonten_predict(voting system) column both are 1, so we will bet on that horse to win.

raceyear ▲ 1	racemonth ▲ 2	raceday ▲ 3	location	distance	course	going	raceno ▲ 4	place	horseno	winodds	od28	ad42	st47	sp51	result	prediction	predict_place	tips_cnt	nonten_predict
2017	3	1	ST	1200	ALL WEATHER TRACK	GOOD	5	1	11	2.7	1	1	1	1	1	0.207019	1	4	1
2017	3	5	ST	1000	TURF - "C" COURSE	GOOD TO FIRM	3	1	5	1.8	1	1	1	1	1	0.367894	1	4	1
2017	3	5	ST	1200	TURF - "C" COURSE	GOOD TO FIRM	5	1	4	4.1	1	0	1	1	1	-0.019418	1	3	1
2017	3	5	ST	1200	TURF - "C" COURSE	GOOD TO FIRM	8	1	5	1.9	1	1	1	1	1	0.211915	1	4	1
2017	3	5	ST	1800	TURF - "C" COURSE	GOOD TO FIRM	11	1	10	3.1	1	1	1	1	1	-0.039552	1	4	1
2017	3	8	HV	1000	TURF - "A" COURSE	GOOD	1	1	7	1.9	1	1	1	1	1	0.301034	1	4	1
2017	3	8	HV	1650	TURF - "A" COURSE	GOOD	4	1	4	3.3	1	1	1	1	1	0.11241	1	4	1
2017	3	8	HV	1800	TURF - "A" COURSE	GOOD	7	1	3	5.2	0	1	1	1	1	1.54593	1	3	1
2017	3	12	ST	1200	TURF - "C+3" COURSE	GOOD	2	1	1	3.7	1	1	1	1	1	0.150422	1	4	1
2017	3	15	HV	1800	TURF - "B" COURSE	GOOD	2	1	8	3.5	1	1	0	1	1	0.13523	1	3	1
2017	3	19	ST	1200	TURF - "A" COURSE	GOOD TO YIELDING	1	1	2	7.4	1	1	1	1	1	0.125857	1	4	1
2017	3	19	ST	0	TURF - "A" COURSE	GOOD TO YIELDING	8	1	1	1.8	1	1	1	1	1	0.153697	1	4	1
2017	3	19	ST	1400	TURF - "A" COURSE	GOOD TO YIELDING	10	1	1	1.8	1	1	1	1	1	0.126853	1	4	1
2017	3	22	HV	1650	TURF - "C" COURSE	GOOD	1	1	6	6.9	1	1	1	1	1	0.225531	1	4	1
2017	3	22	HV	1000	TURF - "C" COURSE	GOOD	2	1	1	1.8	1	1	1	1	1	0.348078	1	4	1
2017	3	22	HV	1200	TURF - "C" COURSE	GOOD	5	1	1	5	1	1	1	1	1	0.157864	1	4	1
2017	3	26	ST	1400	TURF - "A+3" COURSE	GOOD	9	1	14	3.5	1	1	1	1	1	0.156424	1	4	1
2017	4	2	ST	1400	TURF - "B+2" COURSE	GOOD TO FIRM	6	1	7	4.5	1	1	1	1	1	-0.044992	1	4	1
2017	4	2	ST	1200	TURF - "B+2" COURSE	GOOD TO FIRM	8	1	6	2.2	1	1	1	1	1	0.353085	1	4	1
2017	4	5	ST	1200	ALL WEATHER TRACK	GOOD	7	1	1	2.3	1	0	1	1	1	0.093077	1	3	1
2017	4	9	ST	1000	TURF - "C" COURSE	GOOD TO FIRM	1	1	8	1.8	1	1	1	1	1	0.309714	1	4	1
2017	4	9	ST	1200	TURF - "C" COURSE	GOOD TO FIRM	4	1	8	2.2	1	1	1	1	1	0.150422	1	4	1
2017	4	9	ST	1200	TURF - "C" COURSE	GOOD TO FIRM	7	1	4	3.4	1	1	1	1	1	0.367894	1	4	1
2017	4	12	HV	1200	TURF - "A" COURSE	GOOD	5	1	8	4	1	1	1	1	1	0.23404	1	4	1
2017	4	12	HV	2200	TURF - "A" COURSE	GOOD	6	1	2	5.2	1	1	1	1	1	1.54593	1	4	1

photo of showing the horses agreed by both methods

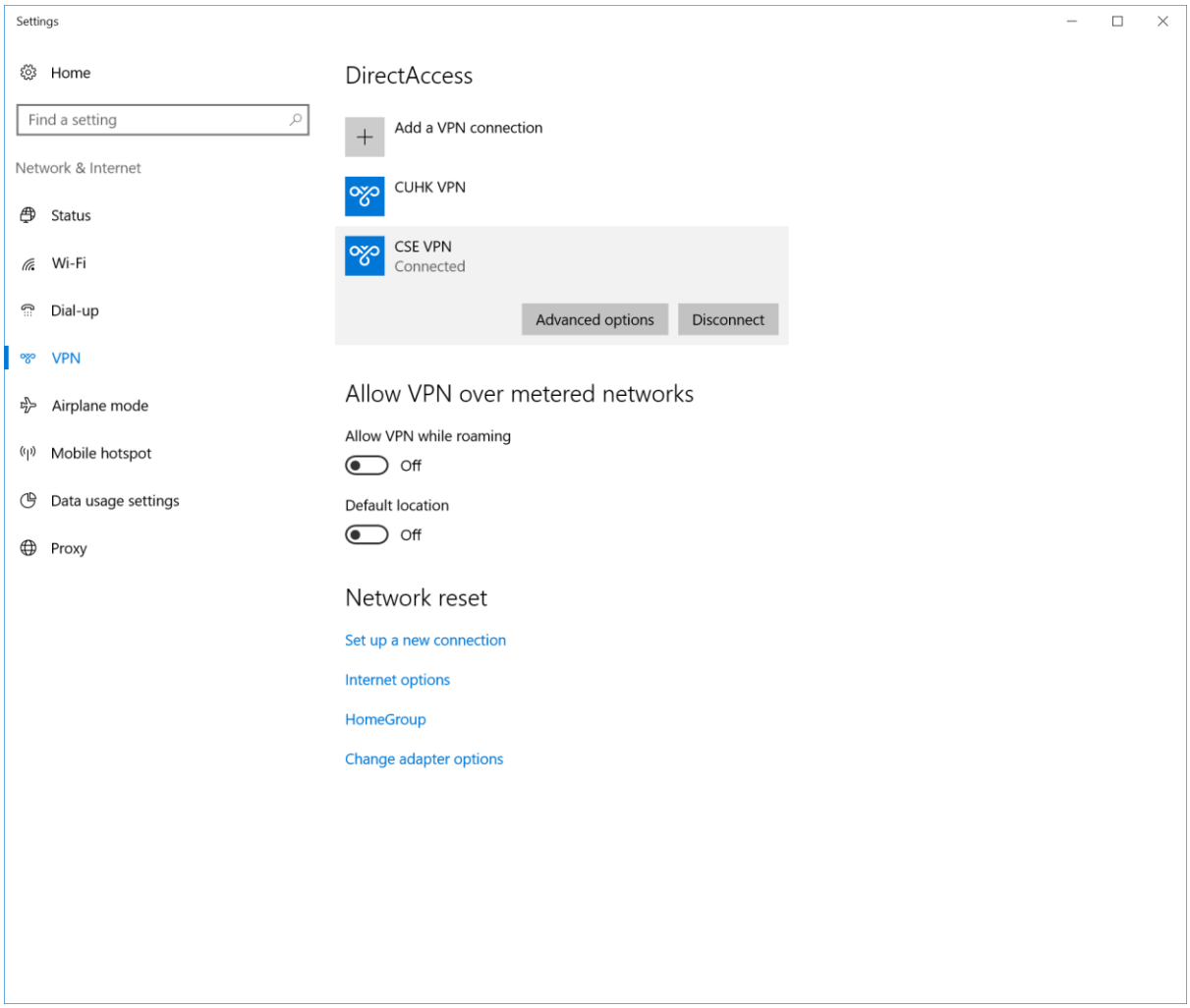


3.6 Demonstration of TensorFlow and voting system

1. Connect CUHK CSE VPN

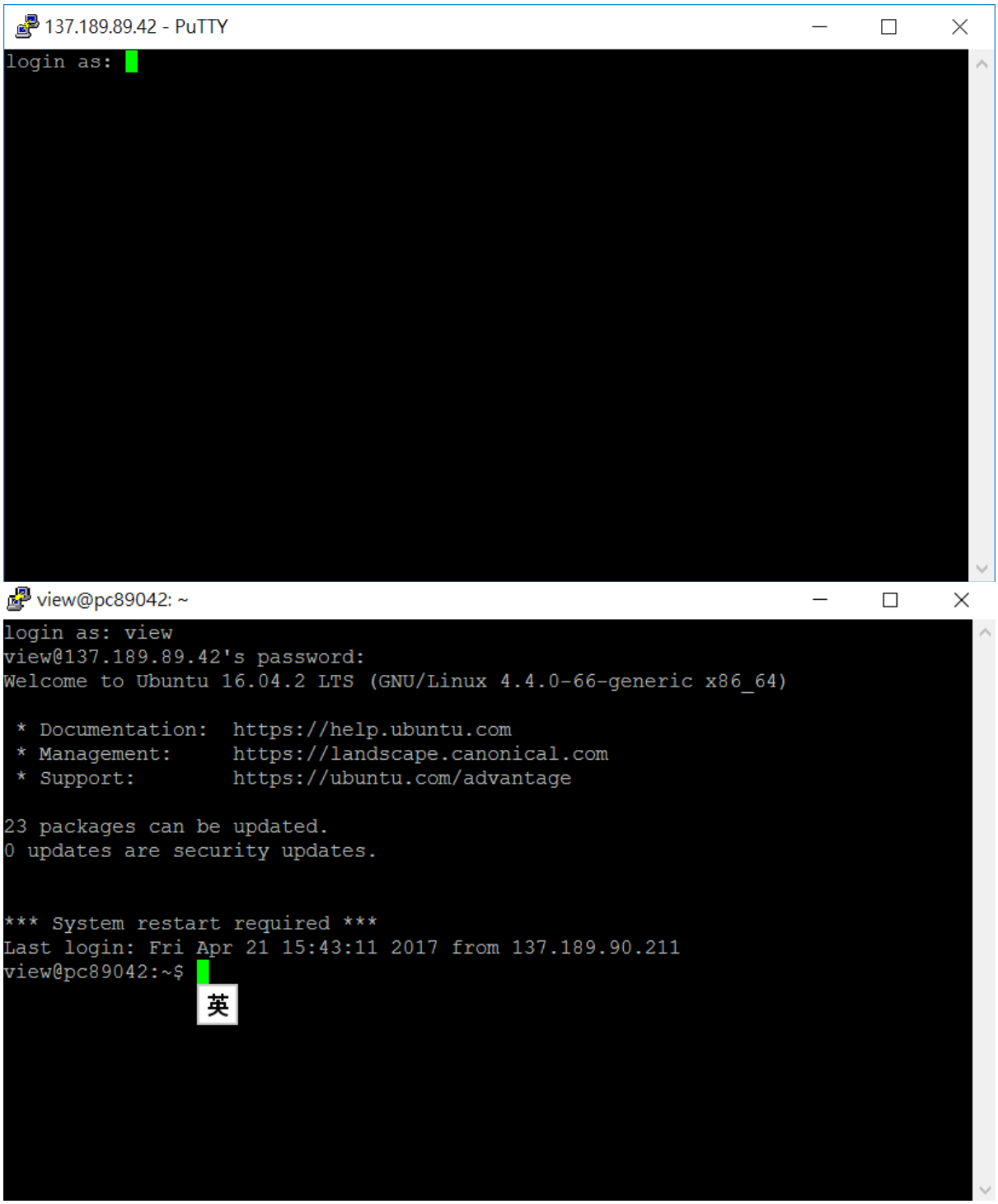
Follow the CUHK VPN Setup Procedures

(https://www.cuhk.edu.hk/itsc/network/vpn/vpn.html), but change the internet address to vpn.cse.cuhk.edu.hk.



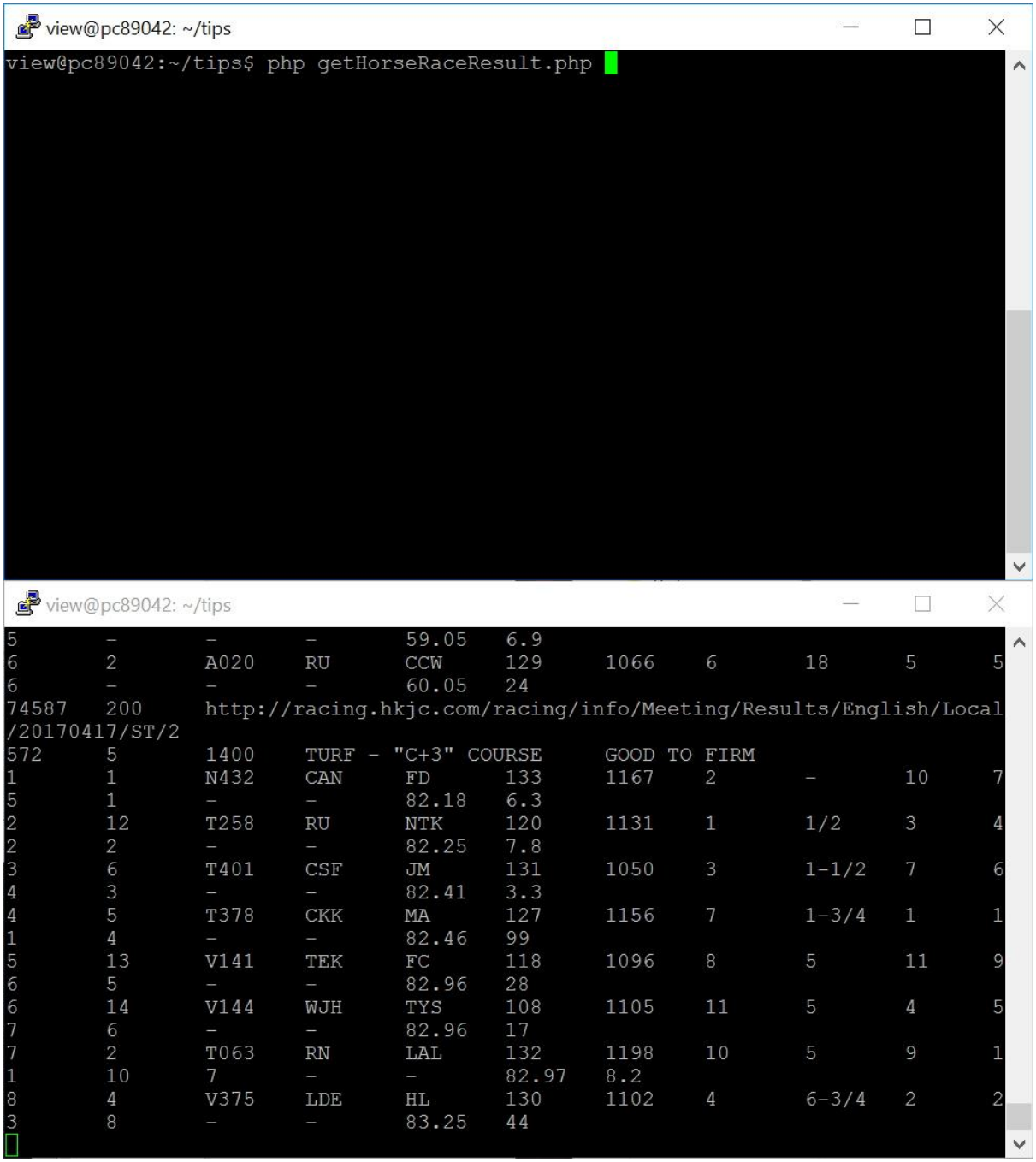
2. Connect to viewlab’s server via putty

Type in user name and password.



3. Download latest horse racing result and update database

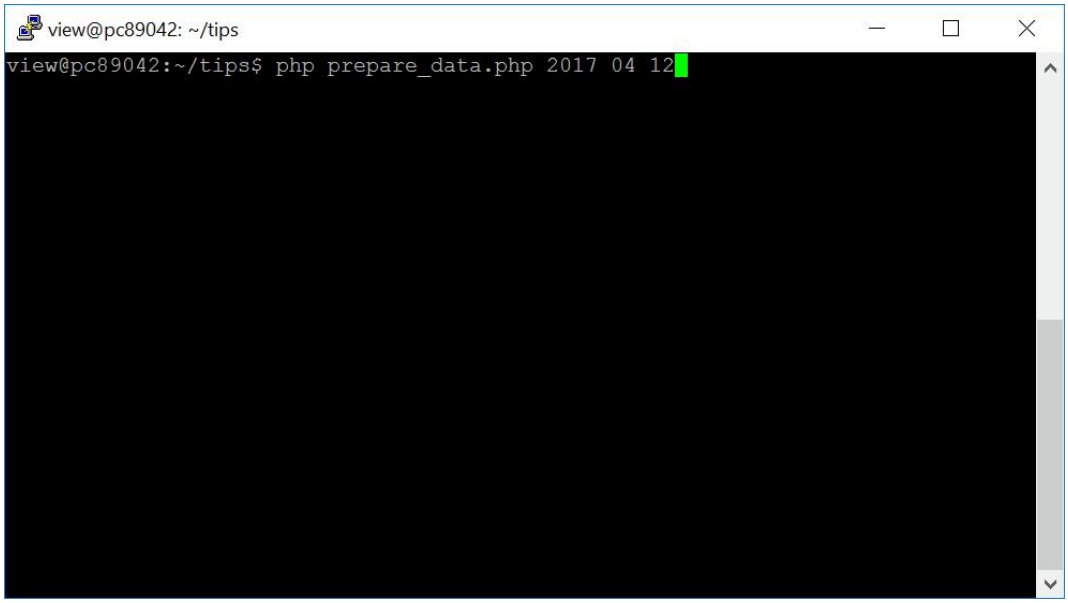
Change directory to the tips folder by command ‘cd tips’, then type ‘php getHorseRaceResult.php’ to run the web scraper for HKJC webpage.



TensorFlow method

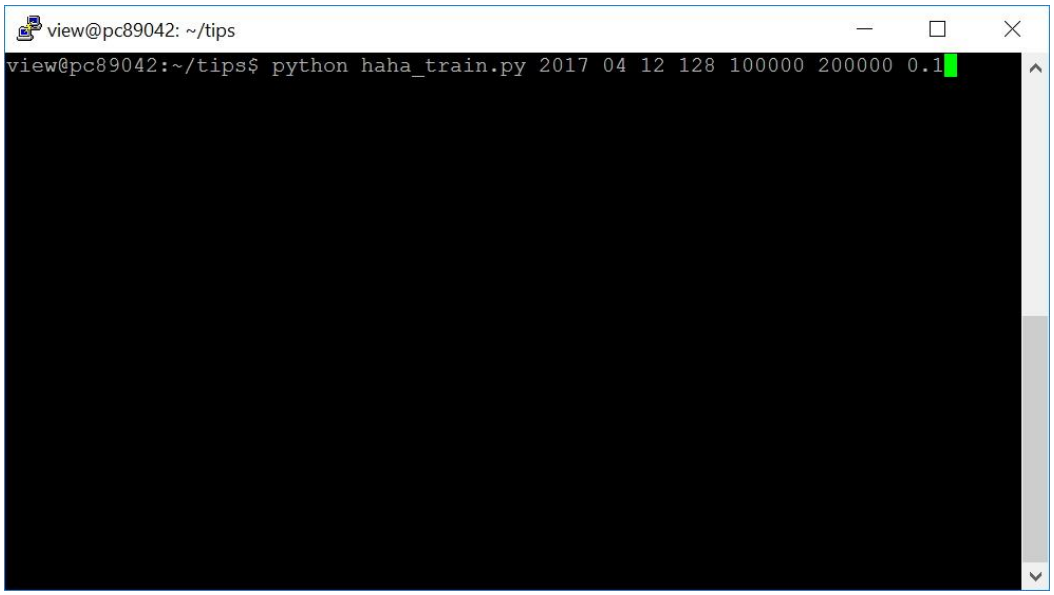
4. Prepare train, test and predict data

After typing ‘php prepare_data.php’, we need to type the date that we want to predict.



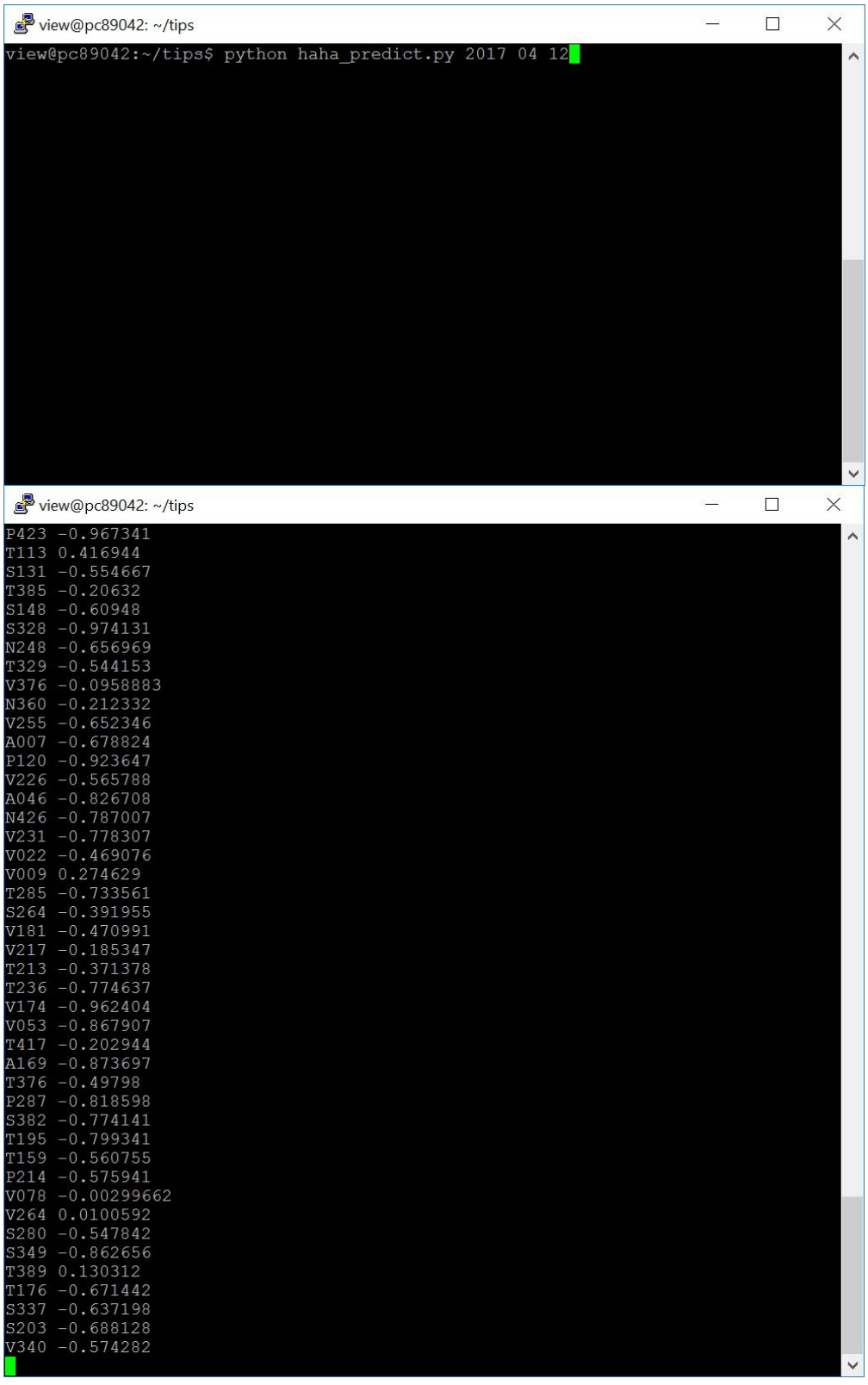
5. Train the data

Here we have 7 arguments, they are race year, race month, race day, TensorFlow model size, minimum training steps, maximum training steps, and minimum error.



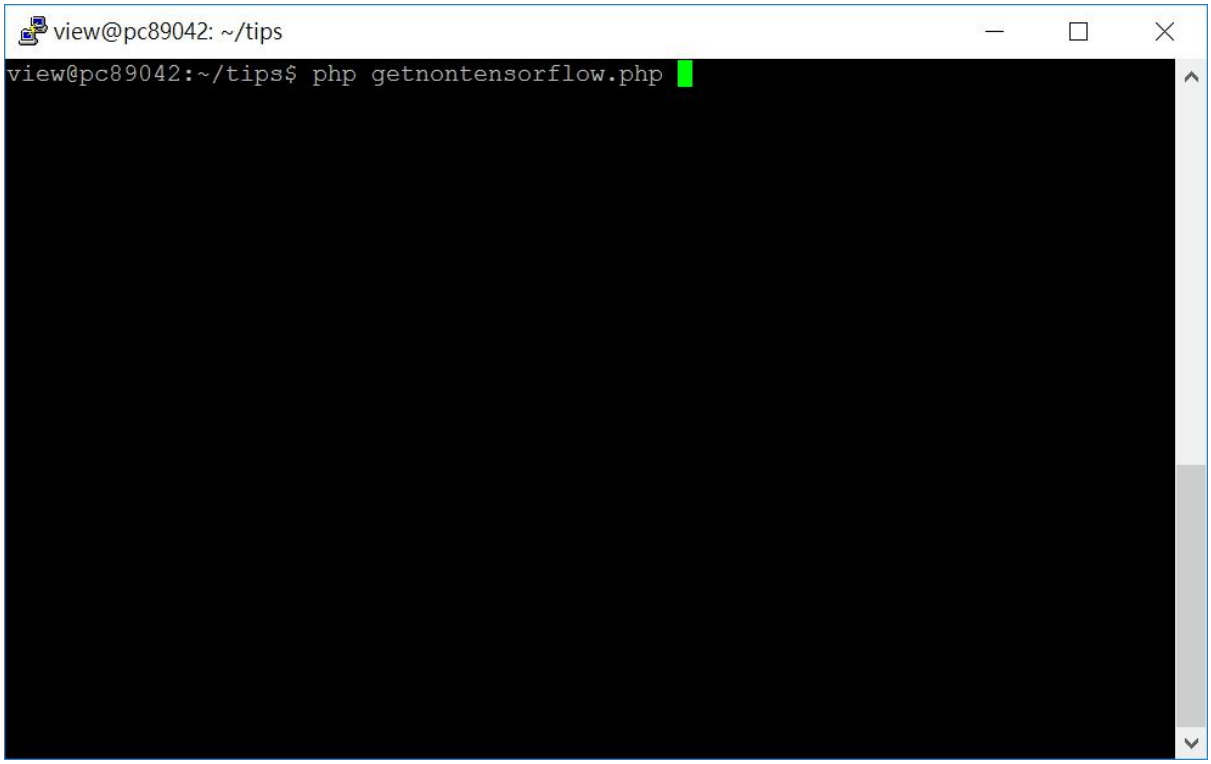
6. Get the predicted result and update the result to database

The results will automatically upload to our database and print out on putty.



Voting system method

4. Count the number of votes of horses in ‘race’, and sort the count number to get the predicted place, also, the results will upload to our database.



4 Research Results

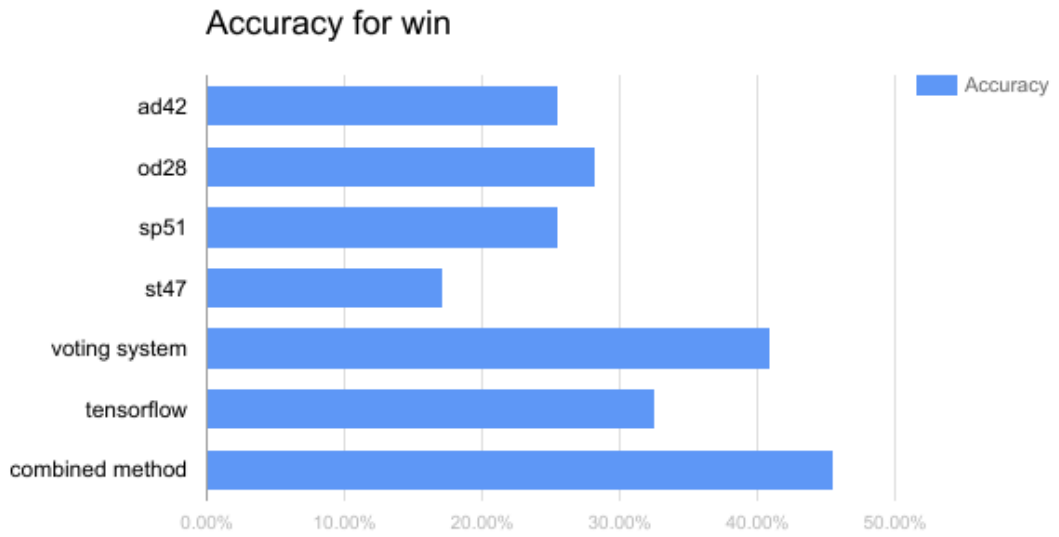
4.1 Accuracy

In this part, we choose the best 4 experts to compare with our voting system, TensorFlow and combined method. And the 4 experts are ad42, od28, sp51 and st47.

Win

In the following table, it shows the accuracy of predicting win in the latest 117 races. However, for combined method, since we are not betting on every races, after combining the TensorFlow and voting system, we only bet on 57 races.

Tips provider	Number of correct prediction	Accuracy
ad42	30	25.6%
od28	33	28.2%
sp51	30	25.6%
st47	20	17.1%
voting system	48	41.0%
TensorFlow	38	32.5%
combined method	26 (in 57 races)	45.6% (in 57 races)



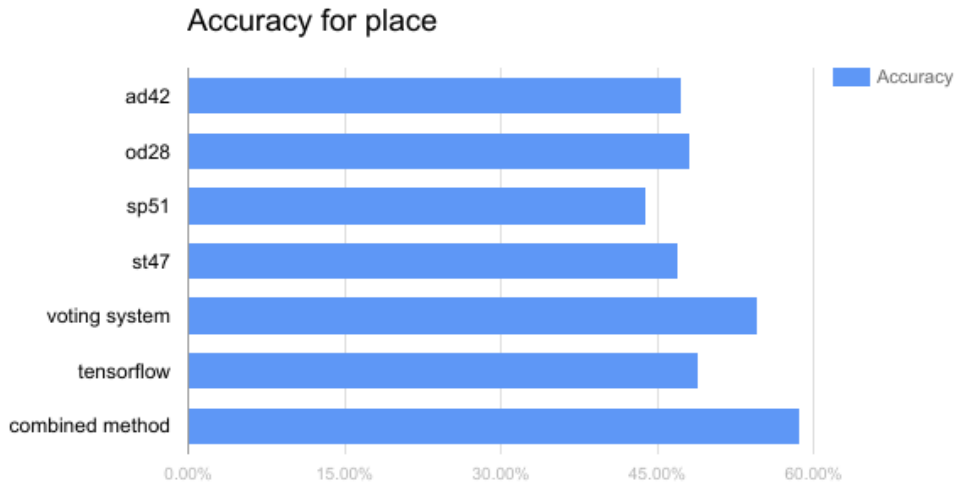
We can see that both voting system and TensorFlow are more accurate than individual experts, while voting system has a more significant different, when we do the combined method, since we filtered out many wrong predictions, the accuracy has further increased.

Place

In the following table, it shows the accuracy of predicting place in the latest 117 races, we buy 3 horses in a race, so in total we bet 351 horses over 117 races. However, for combined method, since we may not betting on every races and we may not betting on three horses in a race, after combining the TensorFlow and voting system, we only bet on 235 horses over 114 races.

Tips provider	Number of correct prediction	Accuracy
ad42	166	47.3%
od28	169	48.1%

sp51	154	43.9%
st47	165	47.0%
voting system	192	54.7%
TensorFlow	172	49.0%
combined method	138 (out of 235 horses)	58.7% (out of 235 horses)



In betting on place, the differences on accuracy between individual experts and collective intelligence are not very obvious, but we still see that voting system has a better prediction than individual experts, and the most accurate method is combined method which is same as for betting on win.

4.2 Gain and Loss Analysis

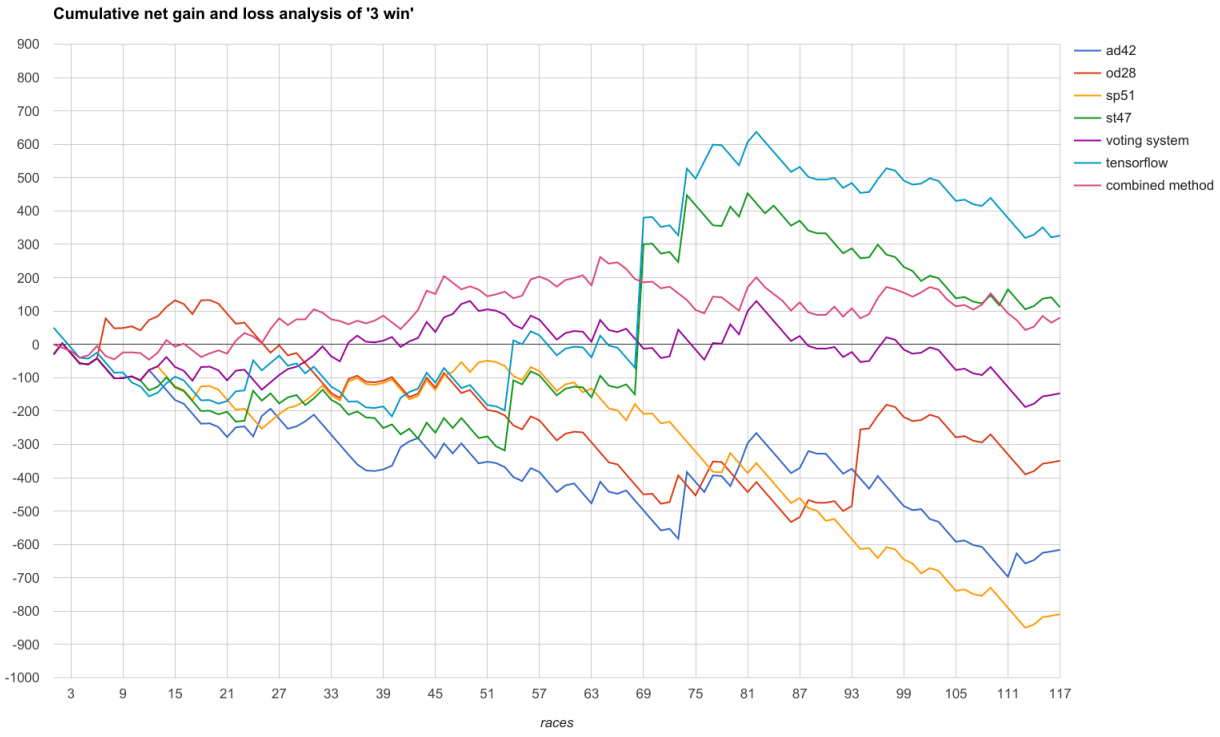
Cumulative net gain and loss analysis

In this part, we simulate betting horses according to the tips providers, voting system, TensorFlow and the combined method. Some graph will be provided for comparing different methods. We try to compare them by three betting strategies. Since the TensorFlow method can only predict racing result starting from March 2017, so our simulation is from March 2017 to recent, total 117 races.

- **3 win(choose three horses in each race, and buy ‘win’ for all of them)**

This betting strategy is to bet ‘win’ for the first three order of predicted horse. We buy $10 for each horse, so we spend $30 for each race and $3510 for the whole period. The combined method only bets the horses chosen by both TensorFlow and voting system. If no same horses are betted, combined method will skip that race.

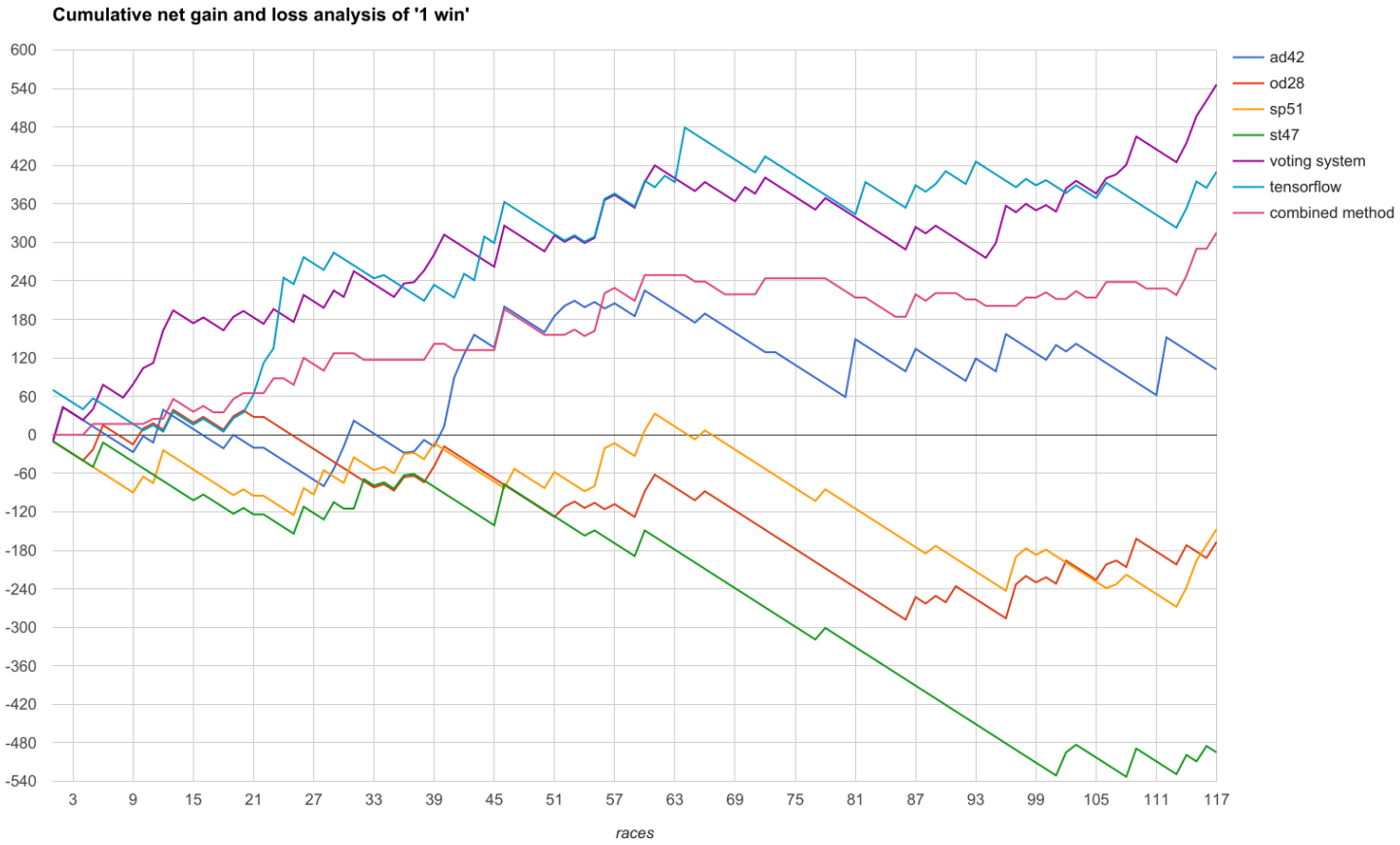
Before race 68, most of the method were losing money, instead of voting system and the combined method. At race 68, st47 predicted a unpopular horse with high win odd and it finally wins the race. The TensorFlow also agree with st47 prediction in this race, so the net gain of TensorFlow and st47 increased substantially at race 68. However, since ad42, od28 and sp51 predicted wrongly at race 68, so the voting system also predicted wrongly. Finally, TensorFlow gets the highest net gain with $326. st47 and combined method can get net gain but the other methods get net loss using this betting strategy. sp51 loss the largest amount of money with $809.



- **1 win (choose one horse in each race, and buy ‘win’ for it)**

This betting strategy is to bet ‘win’ for the first order of predicted horse. We buy $10 for each horse, so we spend $10 for each race and $1170 for the whole period. The combined method only bets the horses chosen by both TensorFlow and voting system. If the horses betted by TensorFlow and voting system are different, combined method will skip that race. The combined method finally bets 57 races over 117 races, so it spends $570 for the whole period.

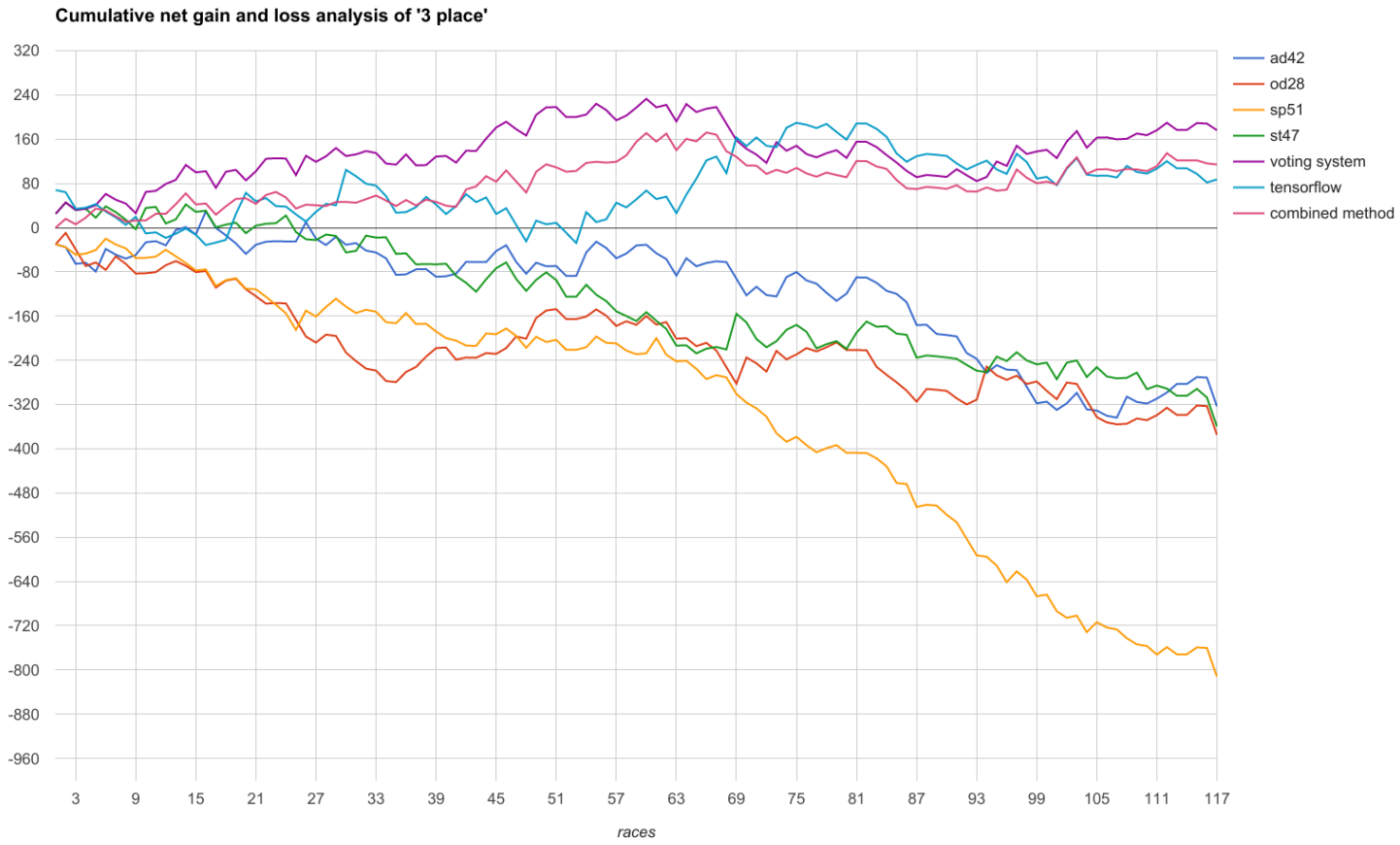
This time, the net gain of voting system, TensorFlow and combined method increase steadily all the time. Voting system gets the highest net gain with $546 among all method. Besides, ad42 also gets net gain in this period but the other tips providers get net loss using this betting strategy. st47 loss the largest amount of money with $495.



- **3 place (choose three horses in each race, and buy ‘place’ for all of them)**

This betting strategy is to bet ‘place’ for the first three order of predicted horse. We buy $10 for each horse, so we spend $30 for each race and $3510 for the whole period. The combined method only bets the horses chosen by both TensorFlow and voting system. If no same horses are chosen by TensorFlow and voting system, combined method will skip that race.

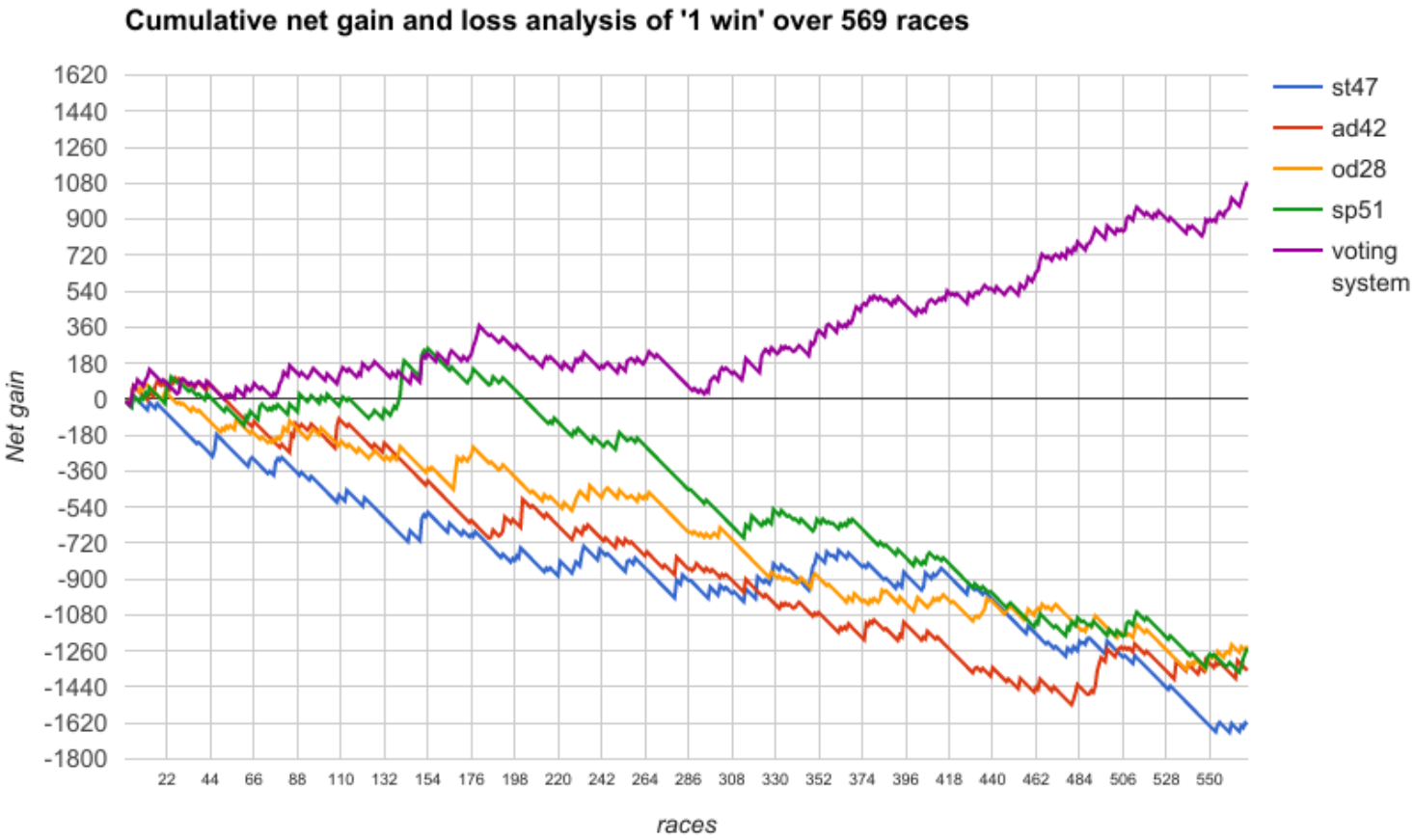
The below graph shows that the net gain of voting system, TensorFlow and combined method are always positive in the whole period. The voting system get the highest net gain with $176.2. The worst tips provider is sp51, it gets a net loss with $812.3.



- **1 win (choose one horse in each race, and buy ‘win’ for it, 569 races, without TensorFlow)**

Since we need to use over 300 races as training data in TensorFlow, the above analysis is for 117 races. However, if we remove TensorFlow, we can analyse individual experts and voting system over 569 races. So, we simulate the betting start from September 2016 to recent.

The below graph shows that the power of collective intelligence is really strong and fascinating, while all experts are losing their money over 569 races, our voting system can earn more than thousands dollar if we bet $10 on each race.

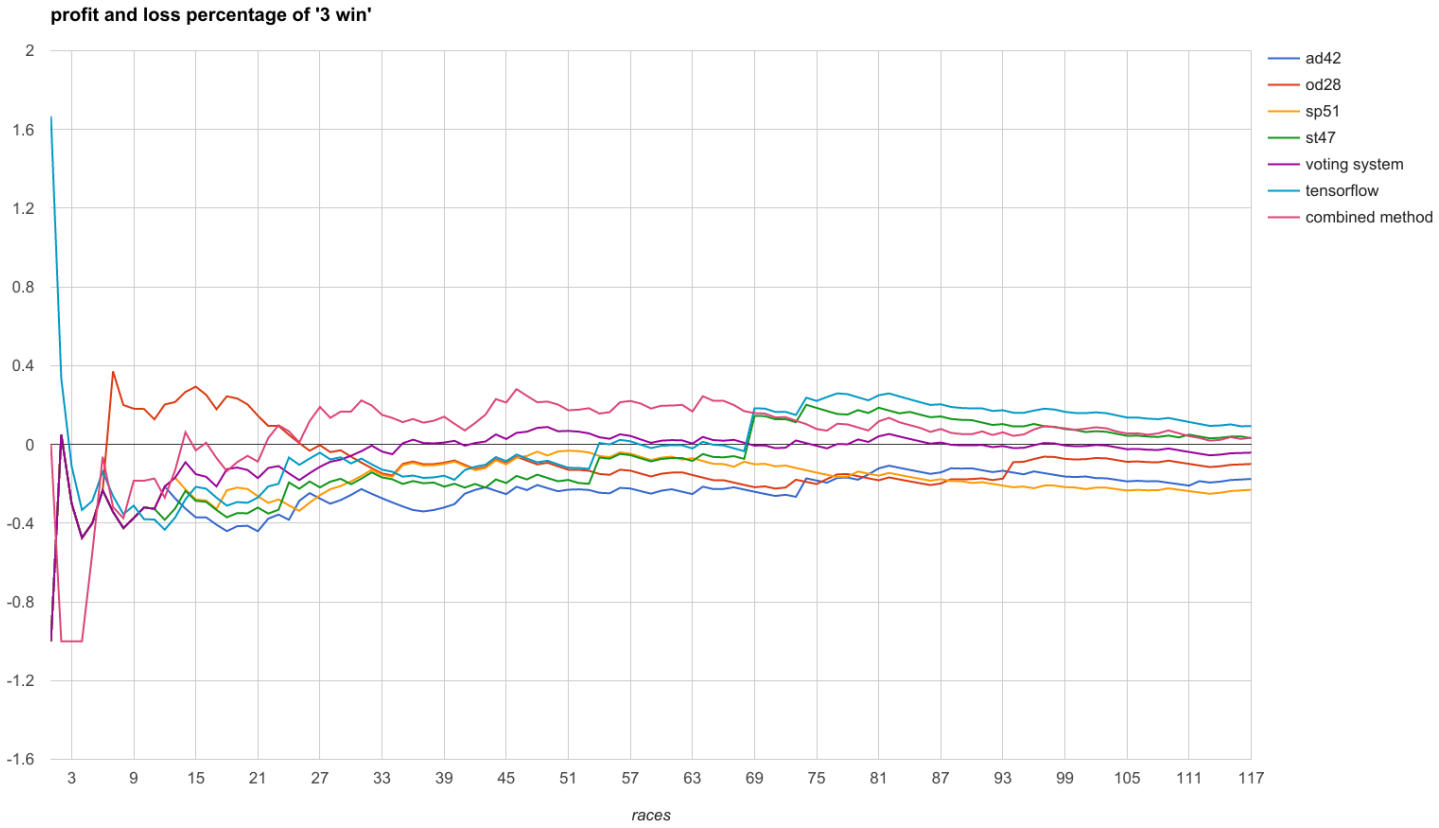


Profit and loss percentage

On the above graph, the combined method could not get the highest net gain among other method, but it does not represent it is worse than other method. Since the combined method would skip races, it should spend less money than other methods. The above comparison may not be fair for combined method. So, in this part, we are going to show the profit and loss percentage of races. The percentage are tracing the value of net gain over money spent of every races. For example, if the tips provider spent $10 and loss the betting at the first race, the percentage will be -100%. After that, the same tips provider spent $10 and win $20 at the second race, the net gain will be $10 and the total money spent is $20, so the percentage will be 50% for the second race.

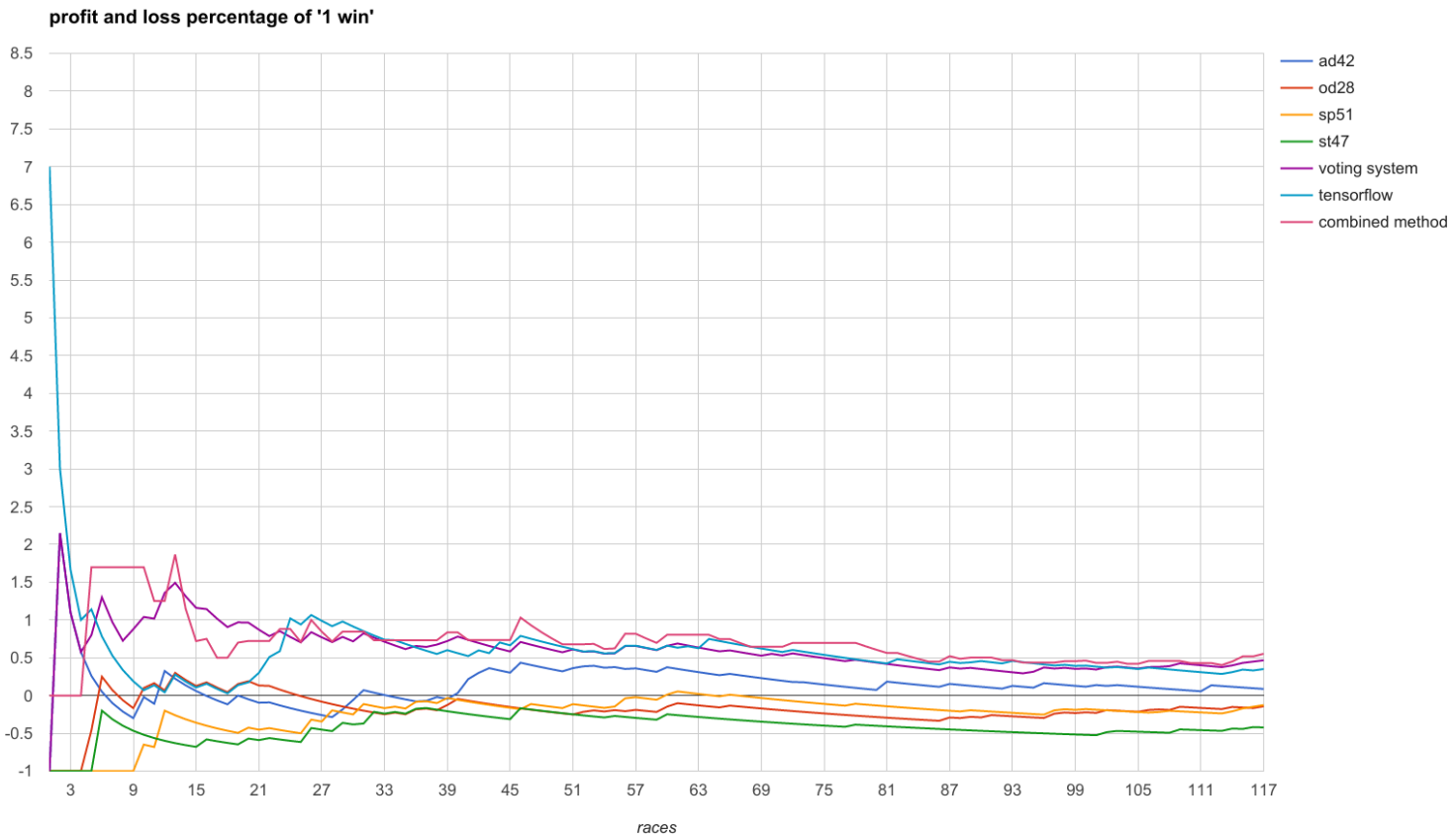
- **3 win (choose three horses in each race, and buy ‘win’ for all of them)**

In this case, ‘st47’, TensorFlow and combined method get position percentage of profit in the whole period. TensorFlow get the highest percentage with 9.29%. The other tips providers and voting system both get negative percentage.



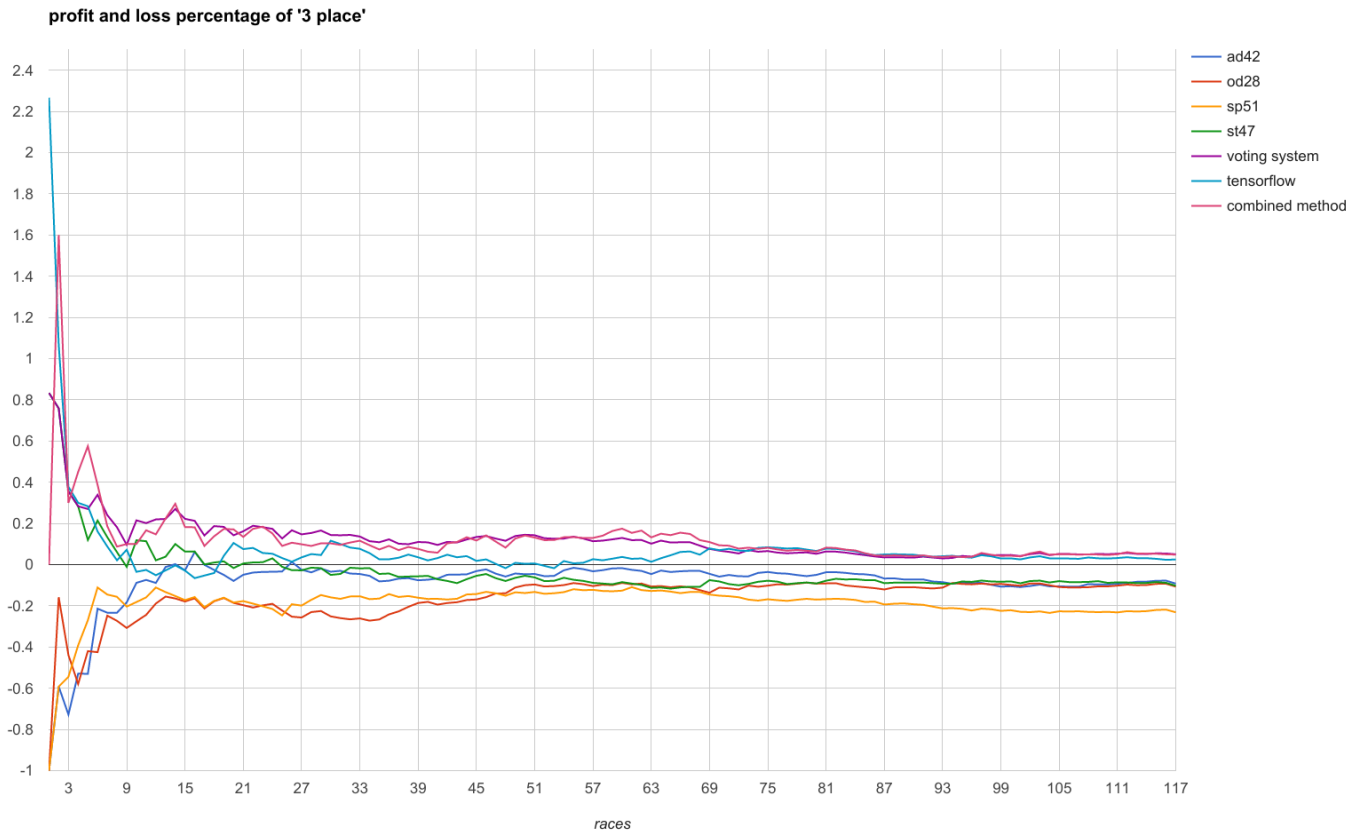
- **1 win (choose one horse in each race, and buy ‘win’ for it)**

In this case, ad42, voting system, TensorFlow and combined method get positive profit percentage. ad42 get the least profit percentage among these method. Combined method get the highest profit percentage with 55.3% which is a desirable result. As the graph show that the individual tips providers get lower profit percentage than TensorFlow, voting system and combined method, the power of collective intelligence can be seen in this case.



- **3 place (choose three horses in each race, and buy ‘place’ for all of them)**

The below graph shows that only voting system, TensorFlow and combined method get positive profit percentage. The other individual tips providers all lose their money. The voting system get the highest profit percentage with 5.02% which is a undesirable result. So, it may not be a good betting strategy. At least it shows that the combined or calculated method are all better than the individual tips.



5 Discussion, Recommendations and Conclusion

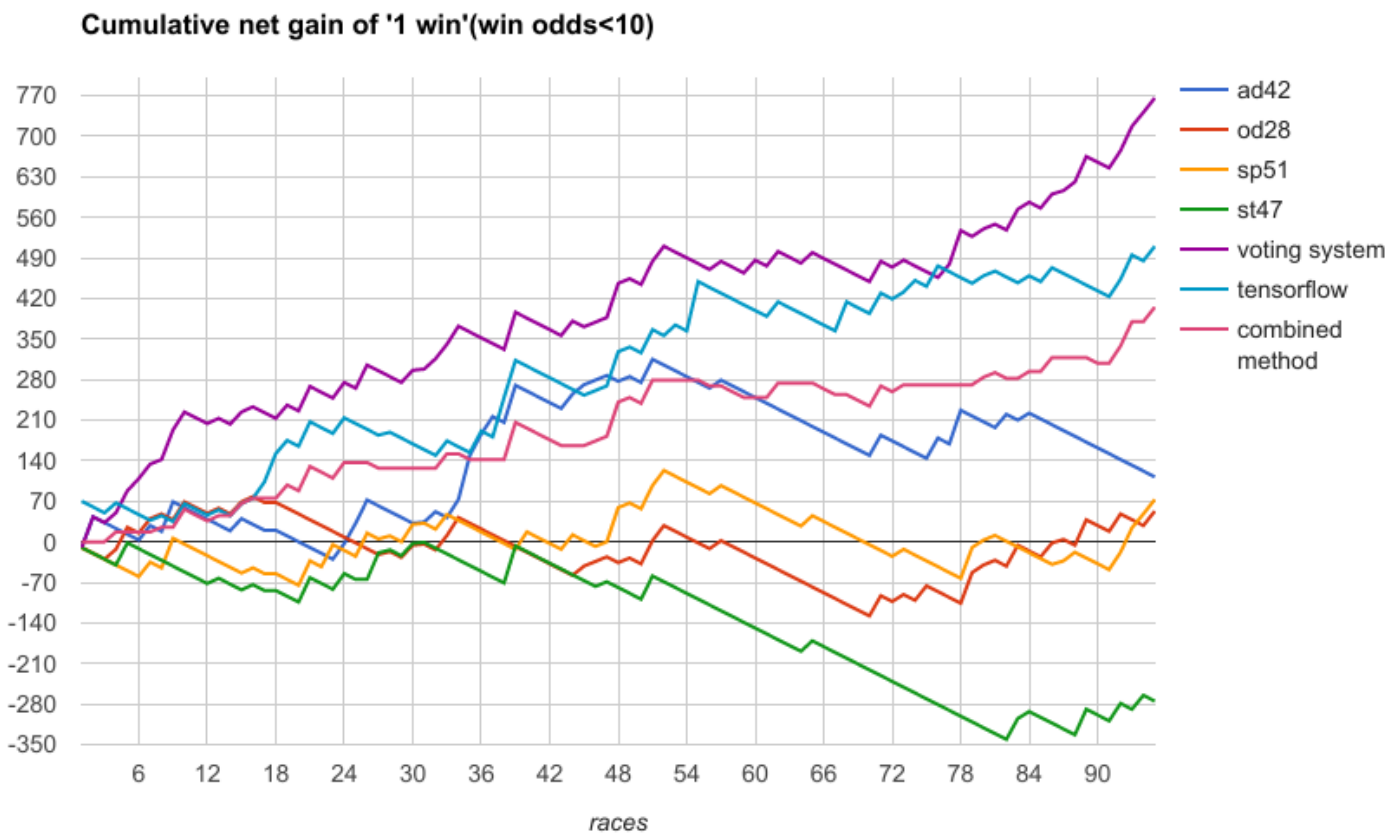
5.1 Maximization of Returns

To summarize the above results, if we want to earn more money, we should choose to use the combined method, which has the highest profit percentage(cumulative net gain divided by total cost), and the best betting method is betting win. we choose only one horse in a race, and buy it win.

- **Only bet on horses with win odds lower than 10**

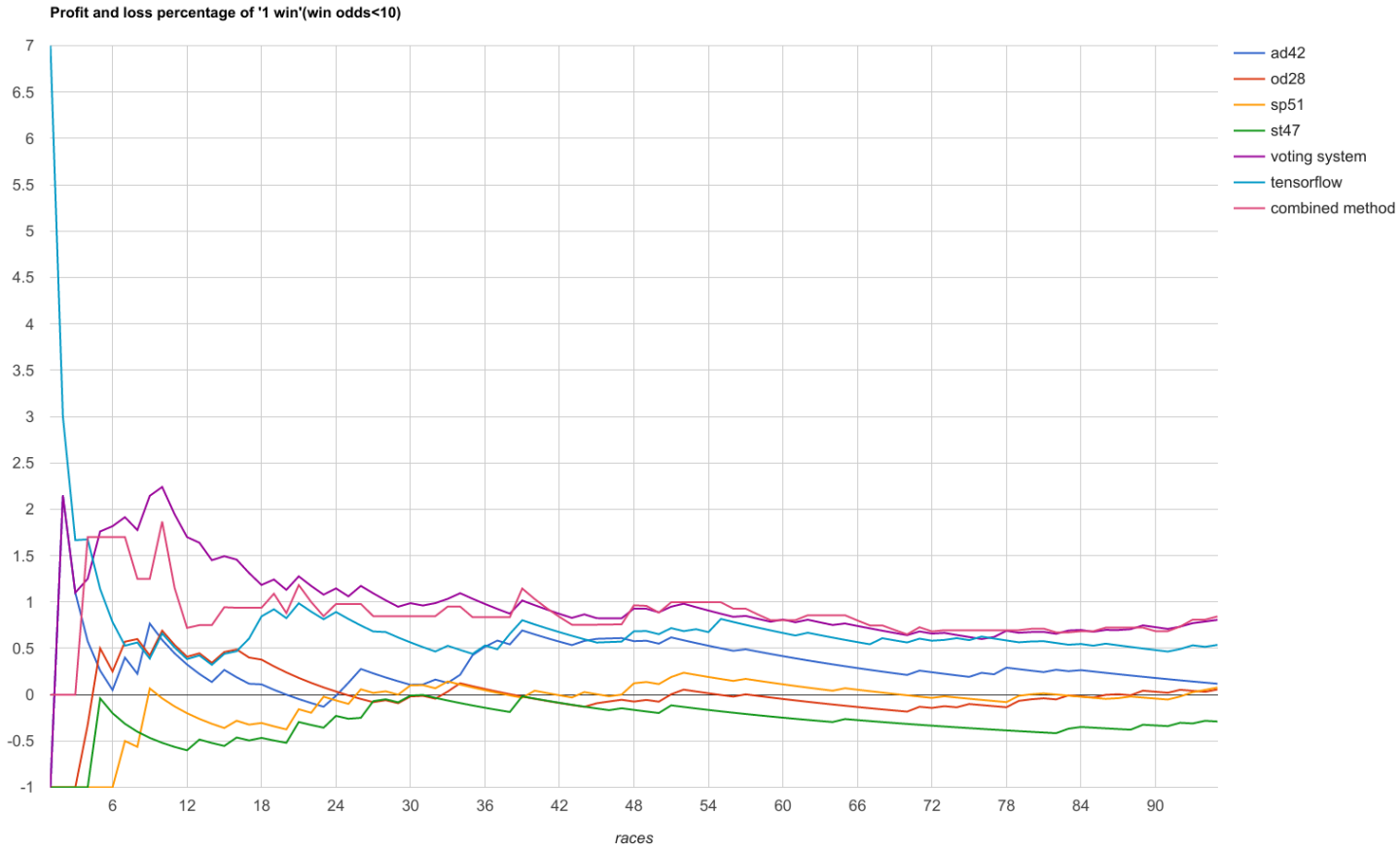
To further improve our system to earn more money, we found that most of the correct predictions, the corresponding win odds are lower than 10, so that if we got the predictions, but the current win odds is higher than 10, we don't bet on that race.

To see how good of this strategy, we choose '1 win' to analyse, after filtered out the horses with win odds higher than 10, we only bet on 95 races out of 117 races. Also, same as before, for combined method, we only bet on 48 races.



Tips provider	Final Net gain
ad42	112
od28	53
sp51	73
st47	-275
voting system	766
TensorFlow	510

combined method	405 (in 48 races)



Tips provider	Final profit and loss percentage
ad42	11.8%
od28	5.58%
sp51	7.68%
st47	-28.9%

voting system	80.6%
TensorFlow	53.7%
combined method	84.4% (in 48 races)

Tips provider	Net gain		profit and loss percentage	
	Before	After	Before	After
ad42	102	112	8.72%	11.8%
od28	-167	53	-14.3%	5.58%
sp51	-147	73	-12.6%	7.68%
st47	-495	-275	-42.3%	-28.9%
voting system	546	766	46.7%	80.6%
TensorFlow	410	510	35.0%	53.7%
Combined method	315	405 (in 48 races)	55.3%	84.4% (in 48 races)

It is obvious that the net gain and profit percentage have increased significantly, we believe that the win odds is actually meaningful, when it is too high, the horse has a very low probability to win, if we still bet on those races, we will loss money, and our filtering method showed that it is actually helpful to maximize our money return.

- **Prediction on 20-4-2017**

The following is the demonstration on 20th of April, using the prediction of voting system with the new strategy. We filtered out race 8 because the win odd was higher than 10. Besides, we bet $10 on each horse for win, that means we totally bet $70 on that day. We predicted 5 races correctly over 7 races, so our net gain is $147 in one single day which is a satisfied result for us.

race no	1st	2nd	3rd	predicted winner from voting system	Win Odds
1	5	2	10	5	2.5
2	4	2	8	3	5.7
3	7	9	4	9	4.7
4	1	7	12	1	4.6
5	2	10	5	2	4.6
6	1	8	2	1	4.7
7	8	4	2	8	5.6
8	3	12	9	7	16

5.2 Limitation and difficulties

Actually there are a lot of betting methods other than ‘win’ and ‘place’, such as ‘Quinella’(1st and 2nd in any order in a race), ‘Tierce’ (1st , 2nd and 3rd in correct order in a race), ‘Trio’ (1st, 2nd and 3rd in any order in a race) and ‘First 4’ (1st, 2nd , 3rd and 4th in any order in a race) etc. But due to the time limit, we chose to investigate our research on the two hottest betting method.

Also, we think that our betting strategy has not yet completed and it can be more perfect. Although we found out that we can actually earn a lot of money, we are now only betting on win, we believe that, some of the time, we can bet on more than one method in a race, for example, we may bet one horse win and place, and sometime we don’t bet it win, we bet it place only.

5.3 Conclusion

In conclusion, we see that collective intelligence is truly powerful, and this moment we can answer to our previous research problem.

1. **How accurate are the experts on betting horse racing?**

They are not accurate, for betting on win, their accuracy is around 20% to 30%, and for betting on place, there accuracy is around 40% to 50%.

2. **How good are the experts on betting horse racing in terms of money returns?**

It is very bad actually, no matter betting on win or place, we see that they have a net loss of money in most of the time. It is only when we choose to bet on '1 win' with win odds lower than 10, ad42, od28 and sp51 had a little net gain, which is only better than loss money, but they cannot earn money.

3. **How accurate is collective intelligence on betting horse racing?**

They are more accurate than individual experts, for betting on win, their accuracy is over 30%, and for betting on place, there accuracy is over 49%, combined method is the most accurate, it has accuracy of 58.7%.

4. **How good is collective intelligence on betting horse racing in terms of money returns?**

It is good when we bet on ‘1 win‘ and ‘3 place’, all of our three methods earned money, the highest money return is betting on ‘1 win’ with win odds lower than 10, voting system earn the most money in the same period of time. While combined method has the best profit percentage, since the total cost of it is lower than other two methods, the net gain of money is not the highest.

5. Is crowdsourcing better than individual experts?

It is absolutely true. When we look at the above graphs, voting system, TensorFlow and combined method are significantly better than individual experts, no matter in accuracy, net gain of money or profit percentage. Only our three collective intelligence can stably earn money, while individual experts are stably losing money.

6. What is the best strategy to bet on horse racing?

After seeing the results, we suggest to bet on ‘1 win’ with win odds lower than 10 using combined method. Because you can earn money with higher stability, and its profit percentage is highest. Some may argue that the graphs showed that voting system earn more money in the same period, but please remind that, the graphs assumed that we bet $10 on each race, so when we use combined method, we can increase the betting amount to increase to money return.

7. Can we actually earn money using our chosen best method?

Yes of course. For example, if we use our best strategy mentioned before, and we bet $20 on each race, we can earn $810 after 117 races.

6 Division of Labor

In this research, we work together in most of the time. For example, when we do the coding part, we discussed on the design first, and then do it at the same time, since there are a lot of parts are similar, we cooperated really well and did the thing efficiently.

Since we had regular meeting with professor and supervisor in every week, we both contributed a great deal to our research, as we know that, crowding sourcing is powerful, our cooperation must better than we do it individually.

In the report part, we also had a meeting before writing the report, we discussed the report outline, so we knew what should we write. After that, we were responsible for different parts in this report as follows:

Acknowledgements, Abstract, Contents	Lam Cheuk Hei
Motivation, Objective, Other related study, Research Problem	Lam Cheuk Hei
Definition	Chan Lok Wai
Research Design	Lam Cheuk Hei
Research Participants,Research Instruments	Chan Lok Wai
Database Design, Data Collection and Pre-processing	Chan Lok Wai
Voting System Design, TensorFlow Design, Combine of TensorFlow and voting system	Lam Cheuk Hei

Demonstration of Tensorflow and voting system	Chan Lok Wai, Lam Cheuk Hei
Accuracy	Lam Cheuk Hei
Gain and Loss Analysis	Chan Lok Wai, Lam Cheuk Hei
Maximization of Returns, Limitation and difficulties, Conclusion	Lam Cheuk Hei

7 References

[1] Andres Munoz, Machine Learning and Optimization, from https://www.cims.nyu.edu/~munoz/files/ml_optimization.pdf

[2] A. Angelova, A. Krizhevsky, and V. Vanhoucke. Pedestrian detection with a large- field-of-view deep network. In Proceedings of ICRA, pages 704–711. IEEE, 2015. www.vision.caltech.edu/anelia/publications/ Angelova15LFOV.pdf.

[3] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. DeVISE: A deep visual-semantic embedding model. In Proceedings of NIPS, pages 2121–2129, 2013. research.google.com/pubs/archive/41473.pdf.

[4]Bell, S., & Willekes, C. (2014, 08). Horse Racing and Chariot Racing. *Oxford Handbooks Online*. doi:10.1093/oxfordhb/9780199589425.013.027

[5] B. Jacob et al. gemmlowp: a small selfcontained low-precision GEMM library, 2015. github.com/google/gemmlowp.

[6] B. Jacob, G. Guennebaud, et al. Eigen library for linear algebra. eigen.tuxfamily.org.

[7]Collective Intelligence: Meme, open source intelligence, user-generated content, collaborative filtering, e-participation, swarm intelligence, collective memory, helium.com, bees algorithm, the irg solution, hierarchical incompetence and how to overcome it, smart mob. (2010). Memphis, TN: Books LLC.

[8]*Deep Learning*. (2017). Cambridge, MA: MIT Press.

[9]Deng, L., & Yu, D. (2014). *Deep learning: Methods and applications*. Boston: Now.

[10] Doan, A., Ramakhrisnan, R. and Halevy, A. (2011) Crowdsourcing Systems on the World-wide Web. Communications of the ACM, 54, 86-96.

[9] ELNAZ DAVOODI, ALI REZA KHANTEYMOORI . (2010). Horse Racing Prediction

Using Artificial Neural Networks.

[11]Encyclopedia of British Horse Racing. (2004, 12). doi:10.4324/9780203483909

[12]Estellés-Arolas, E. and González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. Journal of Information Science, 38, 189.

[13] Friedman, Jerome H.. (1998). "Data Mining and Statistics: What's the connection?".

[14]HOPE, T. (2017). *LEARNING TENSORFLOW*. S.l.: O'REILLY MEDIA, INC, USA.

[15] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean. Multilingual acoustic models using distributed deep neural networks. In Proceedings of ICASSP, pages 8619–8623, 2013. research.google.com/pubs/archive/40807.pdf.

[16] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. arXiv preprint, 1412.7755, 2014. arxiv.org/abs/1412.7755.

[17] J. Gonzalez-Dominguez, I. Lopez-Moreno, P. J. Moreno, and J. Gonzalez-Rodriguez. Frame-by-frame language identification in short utterances using deep neural networks. Neural Networks, 64:49–58, 2015. research.google.com/pubs/archive/42929.pdf.

[18] Mannila, Heikki. (1996). Data mining: machine learning, statistics, and databases. Int'l Conf. Scientific and Statistical Database Management. IEEE Computer Society.

[19] Mart´ın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Man´e, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Vi´egas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[20] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly. Dryad: distributed data-parallel programs from sequential building blocks. In Proceedings of EuroSys, pages 59–72, 2007. www.microsoft.com/en-us/research/wpcontent/uploads/2007/03/eurosys07.pdf.

[21] Neural Network. http://www.investopedia.com/terms/n/neuralnetwork.asp

Computing Science and Statistics. 29 (1): 3–9.

[22]Saxena, A. (2016, 06). Convolutional neural networks. *XRDS: Crossroads, The ACM Magazine for Students*, *22*(4), 56-58. doi:10.1145/2951024

[23]Segaran, T. (2007). *Programming collective intelligence: Building smart web 2.0 applications*. Sebastopol,: O'Reilly.

[24]Surowiecki, J. (2014, July 20). The Collective Intelligence of the Web. Retrieved April 19, 2017, from http://www.newyorker.com/tech/elements/the-collective-intelligence-of-the-web

[25] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997. deeplearning.cs.cmu.edu/pdfs/Hochreiter97 - lstm.pdf.

[26] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of ICML, pages 448–456, 2015. jmlr.org/proceedings/papers/v37/ioffe15.pdf.

[27] S. Jean, K. Cho, R. Memisevic, and Y. Bengio. On using very large target vocabulary for neural machine translation. In Proceedings of ACL-ICJNLP, pages 1–10, July 2015. www.aclweb.org/anthology/P15-1001.

[28]*The 15th annual National Institute on the Gaming Law Minefield*. (2011). Chicago, IL: American Bar Association Center for Continuing Legal Education.

[29] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of ACM Multimedia, pages 675–678, 2014. arxiv.org/abs/1408.5093.

[30]ABRAHAMS, S. (2016). *TENSORFLOW FOR MACHINE INTELLIGENCE: A hands-on introduction to learning algorithms*. S.l.: BLEEDING EDGE PRESS.