香港中文大學

計算機科學及工程學系

Department of Computer Science and Engineering,

The Chinese University of Hong Kong

FYP Report (Final)

Exploiting Betting Odds using Machine Learning

Written By

WONG Wing Keung (1155093416)

WONG Ching Yeung Wallace (1155093534)

Supervised By

Prof. Michael R. Lyu

Contents

Tables

Figures

Abstract

In this project, we would apply machine learning to forecast sport events and evaluate the performances by simulating betting against bookmakers. Unlike most research projects that use performance metrics of players, teams etc. for prediction, we make use of the betting odds to forecast sport events.

In the first term, we developed a machine learning model that predicts probabilities of sport events in sport betting markets just before closing. The model was shown to be profitable when betting against bookmakers on Hong Kong horse racing.

In this term, we made improvements to the model in last semester. The improved model was showing better results in Hong Kong horse racing. Besides, we also developed models that support continuous probability prediction until closing. The continuous models also show positive results when testing in Soccer and horse racing betting markets.

Acknowledgements

We would like to thank our supervisor Prof. Michael R. Lyu and advisor Mr. Edward Yau for their guidance and feedback.

In addition, we appreciate the Department of Computer Science and Engineering, The Chinese University of Hong Kong for offering the required computing resources.

Disclaimer

According to the Gambling Ordinance, Chapter 148, Laws of Hong Kong, all gambling activities are illegal except those authorized by the Government.

Results included in this report are simulations. No participation in any forms of gambling is involved. We have no intention of promoting or facilitating illegal betting or bookmaking.

Glossary of Terms

Cantonese translation is included for terms that are difficult to be defined precisely.

Term	Definition
betting market	Known as 賭盤. A betting market is a specific type of bet. Usually, betting markets are defined by sport events. Bookmakers will offer different markets (known as 開盤) for a sport game. For example, in a soccer game, markets offered could be "Match Winner", "Over/Under" (入球大細), "Handicap" (讓球盤) etc.
bookmaker	A company that offer betting markets and accepts bets
betting exchange	A betting exchange is a platform for customers to lay (sell) and back (buy) on the outcomes of events. This is different from the traditional bookmakers where customers can only "buy" for outcomes.
closing	The time that a bookmaker won't accept bets any more. Usually, it is just before the kick-off time of the game. Different bookmakers may have different closing time for the same market. Opposite to opening.
closing odds	Odds offered by a bookmaker at "closing" Opposite to opening odds.
ensemble member	A model in an ensemble model.
line	Known as 盤口. A value set by bookmakers to create a 2-way betting for an event. For example, the lines in the "Handicap" market (讓球盤) are used to adjust the scores, say -1.5 to the Home Team. The

	line eliminates the chance of draw. Bettors need to predict which teams to win after the score is adjusted.
margin	A deduction to odds made by a bookmaker **/ betting exchange** in a market to make profits (known as 抽水). If a bookmaker / betting exchange has lower margins, its odds are higher. Margin in a market with *n* exclusive outcomes is calculated as: $$\text{margin} = \sum_{i}^{n} \frac{1}{\text{odds of outcome } i} - 1$$
payout	A measure opposite to margin. Lower the margin, higher the payout. $$\text{payout} = 1 / \sum_{i}^{n} \frac{1}{\text{odds of outcome } i}$$
odds	A number that represents the payout in a betting. There are many formats of odds. In this report, odds are referring to decimal odds. In decimal odds, the payout is calculated as: $$\text{bet} \times \text{odds}$$ And profit is calculated as: $$\text{bet} \times (\text{odds} - 1)$$
opening	The time that a bookmaker started to accept bets. Opposite to closing.
opening odds	Odds offered by a bookmaker at "opening" Opposite to closing odds.
outcome	One of the possible outcomes in a betting market. For "Match Winner" market of in soccer games, possible outcomes are "Home Team", "Away Team", "Draw".
sport event	An event that occurs in a sport game or a match.

1 Introduction

1.1. Motivation

Sport betting markets are getting more popular nowadays. There is an increasing number of online bookmakers offering betting markets for uncertain events that occur in sports games. From 2009 to 2016, the market size of the global online gambling market doubled gradually from 20 billion USD to 40 billion USD [1]. At the same time, machine learning has been shown to be successful in applying to multiple fields and industries in recent years. We want to explore if machine learning can beat the bookmakers in sport betting.

1.2. Background

1.2.1. Types of Betting

There are 2 types of betting system in general – "pari-mutuel betting" and "fixed-odds betting".

In "pari-mutuel betting", bets are placed into a "pool", which is operated by a bookmaker. The bookmaker will deduct a portion of bets from the pool as commission fees. After that, winners will share the remaining amount of money in the pool in proportion to their winning stakes.

In "fixed-odds betting", bettors will bet for the odds which are offered by bookmakers. Although odds may be adjusted from time to time until closing, the payout is based on the odds at the time that the bet is accepted. Odd changes may due to the bettors' betting activities. Pinnacle, an online bookmaker which offers almost the highest average odds among all major bookmakers [2], claimed that they will make use of the betting activities of their "sharp" bettors to correct their odds [3].

The Hong Kong Jockey Club, the only legal bookmaker in Hong Kong, accepts bets for local horse racing and soccer matches. "pari-mutuel betting" and "fixed-odds betting" systems are used for horse racing and soccer matches respectively.

1.2.2. Limitation on “pari-mutuel betting”

Due to the nature of “pari-mutuel betting”, bettors are unable to know their payouts exactly until the pool is closed. In horse racing, although the Hong Kong Jockey Club provide “odds” while accepting bets, the “displaying odds” are calculated based on the pool at that moment. It is subject to change when others’ bets are going into the pool afterwards.

In order to approximate the final payouts before placing bets, William Benter, a well-known bettor in Hong Kong horse racing market, suggested placing bets as late as possible [4]. The idea is that the sooner you place your bets, the “displaying odds” at that time will be closer to the final one.

However, from our observation in local horse racing, the last “displaying odds” that bettors can see just before the pool is closed, are still very different from the final one. Therefore, in this project, we will only focus on “fixed-odds betting”, which allows bettors to know their payouts before placing the bets.

1.3. Objectives

The overall goal is to develop betting-oriented methodologies that use machine learning to exploit the fixed-odds betting markets. Methodologies will be evaluated on horse racing and soccer betting markets.

First Term:

- Develop a profitable method that use machine learning to forecast probabilities of sport events just before closing.
- Test the proposed method in Hong Kong horse racing.

Second Term:

- Improve the machine learning model in the first term by different techniques.
- Develop a profitable method that use machine learning to forecast probabilities of sport events continuously until closing.
- Test the proposed method in Hong Kong horse racing and Soccer betting markets.

3 Methodology

3.1 Overview

After a careful study, we decided to develop a method that produces Odds-Based Forecast. This is because performance metrics may not be available in every sport. For example, for a soccer game, there is no way to determine the number of attacks, the number of defences etc. unless the game is ended. Some previous studies would use performance metrics from the past few games for prediction. This may result in inaccuracy, as past performances depend on performances of the opponents and are no guarantee of future results. Building a rating system that tracks the performance of the participants can be a solution but it is not trivial. In contrasts, betting odds are widely accessible for every match before the kick-off time.

We are going to build ensemble models to predict the winning probabilities based on the betting odds and utilize some existing betting strategies.

3.2 Betting Strategy

After having the predicted probabilities from our models, we can easily compute the expectations of each outcome in a market. In probability theory, betting for outcomes with negative expectations will result in bankruptcy in the long run. Therefore, our strategy should only bet for those with positive expectations. Besides, a wagering strategy that can produce the maximum return is needed. In gambling theory, there is a well-known formula that relates betting odds and probabilities – Kelly Formula [5].

3.2.1 Kelly Formula

Kelly Formula is used to calculate the optimal fraction of current capital that should be placed, such that the expected geometric growth rate can be maximized, given the odds and probability of winning are known in a game.

The most common version of Kelly Formula K is as follows:

$$K(\sigma,\ p) = \frac{p\sigma - 1}{\sigma - 1}$$

, where p is the probability of winning and σ is the odds offered.

Here is the deviation:

Suppose p is the probability of winning, σ is the odds offered, k is the ratio of the capital to the bet size, the overall rate of return (E) after n (large enough) repeated betting will be:

$$E = (1 + k(\sigma - 1))^{np}(1 - k)^{n(1-p)}$$

$$\log E = np \log(1 + k(\sigma - 1)) + (n - np) \log(1 - k)$$

The k that maximizes $\log E$ can be found by solving:

$$\frac{d \log E}{dk} = \frac{np\sigma - n + nk - nk\sigma}{(1 - k)(k(\sigma - 1) + 1)} = 0$$

$$k = \frac{p\sigma - 1}{\sigma - 1}$$

3.2.2 Kelly Betting

The betting strategy that utilizes Kelly Formula is known as Kelly Betting. Kelly Betting requires an initial capital to start. Whenever we bet, we use Kelly Formula to compute the optimal wager:

$$\text{optimal wager} = \text{current capital} \times K(\sigma,\ p)$$

If the Kelly Formula gives a negative result, it means the expectation is negative and we should avoid placing bets on that outcome. It can be easily shown:

When the expectation is negative, the "fair odds" is larger than the one offered by the bookmaker. And thus, $\frac{1}{p} > \sigma \quad \Rightarrow p\sigma < 1 \quad \Rightarrow p\sigma - 1 < 0 \quad \Rightarrow K(\sigma,\ p) < 0$

3.2.2.1 Fractional Kelly

The above strategy that directly applying the Kelly Formula is also known as the Full Kelly. This Kelly Formula assumes that the probability of winning is deterministic and is unbiased. However, in a sport game, the true probability of an event is not known. Uncertainty is expected in the predicted probability. If the probability is being overestimated, Kelly Formula will suggest a higher bet size which may result in a negative growth rate. People often place some fraction (c) of the optimal bet size in order to reduce risks. This strategy is known as the Fractional Kelly Betting. William Benter also suggested that this strategy would be more suitable in reality [6]. The wager in Fractional Kelly Betting:

$$\text{wager} = \text{current capital} \times K(\sigma,\ p) \times c \qquad \text{where } 0 < c < 1$$

3.2.2.2 Improved Kelly

One problem of the Fractional Kelly is that the fraction chosen (c) can be critical. In general, a higher fraction will result in bigger fluctuations in the return, while a lower fraction will reduce the fluctuations in the return and the overall rate of growth. A lower or higher fraction does not necessarily produce better results. The optimal fraction found by back testing often lies between some values. Therefore, a systematic method that adjusting the fraction based on some given conditions is desirable.

Baker and McHale derived an improved version of Kelly Formula by assuming an uncertainty function exists [7]. The idea is to find the optimal fraction of Fractional Kelly under parameter uncertainty.

Suppose the probability of winning follows a probability density function b, p is the mean of the distribution and σ is the odds offered. The optimal fraction c can be found by maximizing the function below:

$$\int_0^1 b(p')\big(1 + cK(\sigma,\ p') \times (\sigma - 1)\big)^{p}\big(1 - cK(\sigma,\ p')\big)^{1-p} dp'$$

Note that this is very similar to the deviation of Kelly Formula in Section 3.2.1 but with the assumptions that a probability distribution exists and the optimal bet is in a form of cK. However, there is no direct solution exists for the calculation. In this report, the solutions are computed using SciPy's optimizer.

3.3 Optimal Loss Function for Kelly Betting

Once we have chosen Kelly Betting as the wagering strategy, models that having the same objective are required to collaborate with the strategy. In machine learning, the loss function in training decides the behavior or the objective of the model. For our experiments, the Binary Cross Entropy is chosen to be the loss function. Here we will show that it is the optimal loss function to use with Kelly Betting.

Recall that, the Kelly Formula is given by

$$K(\sigma_i,\ p_i) = \frac{p_i\sigma_i - 1}{\sigma_i - 1}$$

, where p_i is the probability of winning (predicted) and σ_i is the odds offered in an outcome *i* in an event.

Suppose we have n outcomes in total, and y_i is the label of the outcome i (whether it is the final outcome or not), σ_i is chosen to be some odds available in the market.

The rate of return (V') after performing Kelly Betting on these n outcomes:

$$V' = \prod_i^n \ (1 + \max(0, K(\sigma_i, \ p_i)) \times (\sigma_i \ - 1))^{y_i} \ (1 \ - \max(0, K(\sigma_i, \ p_i)))^{1-y_i}$$

Maximizing V' is no different from maximizing V and $\log(V)$:

$$V = \prod_i^n \ (1 + K(\sigma_i, \ p_i) \times (\sigma_i \ - 1))^{y_i} \ (1 \ - K(\sigma_i, \ p_i))^{1-y_i}$$

$$V = \prod_i^n \ (p_i\sigma_i)^{y_i} \ (\frac{\sigma_i - p_i\sigma_i}{\sigma_i - 1})^{1-y_i}$$

$$\log(V) = \ \sum_i^n \ y_i \log(p_i\sigma_i) + (1 - y_i)\log(\frac{\sigma_i - p_i\sigma_i}{\sigma_i - 1})$$

The partial derivative of $\log(V)$ with respect to p_a where $i \ \leq a \leq n$ is given by:

$$\frac{\partial \log(V)}{\partial p_a} = \frac{y_a - p_a}{p_a(1 - p_a)}$$

On the other hand, the Binary Cross Entropy (BCE) is given by

$$\text{BCE} = \frac{1}{n}\sum_i^n \ y_i \log(p_i) + (1 - y_i)\log(1 - p_i)$$

The partial derivative of BCE with respect to p_a where $i \ \leq a \leq n$ is given by:

$$\frac{\partial \ \text{BCE}}{\partial p_a} = \frac{y_a - p_a}{p_a(1 - p_a)} = \frac{\partial \log(V)}{p_a}$$

This implies that optimizing the Binary Cross Entropy is no different from optimizing $\log(V)$, which is a measure of the rate of return in Kelly Betting. Therefore, Binary Cross Entropy is the optimal loss function to use when Kelly Betting is chosen to be the betting strategy.

Note that the above conclusion is also true for Fractional Kelly Betting. It can be shown easily by multiplying a constant variable to the function K. After that, the deviated result is still the same. However, for the Improved Kelly mentioned in Section 3.2.2.2, its optimality depends on the probability density function b instead. In general, a model trained with Binary Cross Entropy will output a deterministic probability value instead of a probability density function. In the experiments, the probability density function is crafted using the predictions from the models. The details will be covered in the later sections.

3.4 Model

The technical details of the models will be mentioned in Section 4. In all experiments, the procedures for training the models are the same. In this section, we will introduce the procedures and their reasoning behind.

3.4.1 Early Stopping

Overfitting will very likely to cause bankrupt in Kelly Betting, as the bet size is highly related to the predicted probability. Overestimation should be avoided as possible. Therefore, Early Stopping will be used during training to reduce over-fitting. To train a model, we first shuffle the whole training set. The first half of data will be used in training and the second half will be used to monitor the loss continuously. Training will be stopped if the monitored loss shows no improvement in the last 50 epochs.

3.4.2 Ensemble Forecast

Each of the trained models carries its own hypothesis. A different model will be produced when we run the training again, especially under early stopping procedures mentioned above, where a different subset of the training set will be used for training every time. As a result, the performances of trained models can be different. In order to improve the robustness, ensemble forecast is used. Instead of training a single model, multiple models are trained and grouped to form an ensemble model. The output of the ensemble model will be the average of outputs from its ensemble members (ensemble mean). Figure 1 illustrates the design and the idea of the ensemble model.

It is possible that for 2 ensemble models to produce different predictions. In general, if there are enough ensemble members, the differences would be very small as the outliers will have less effect on the majority. Empirically speaking, 100-1000 members are good enough for our experiments. Note that the number of ensemble members produced depends on the actual training time. Due to time limitation, we

are not unable to produce a lot of ensemble member for some kinds of model. The exact number of ensemble member for different kinds of model will be mentioned in Section 4.

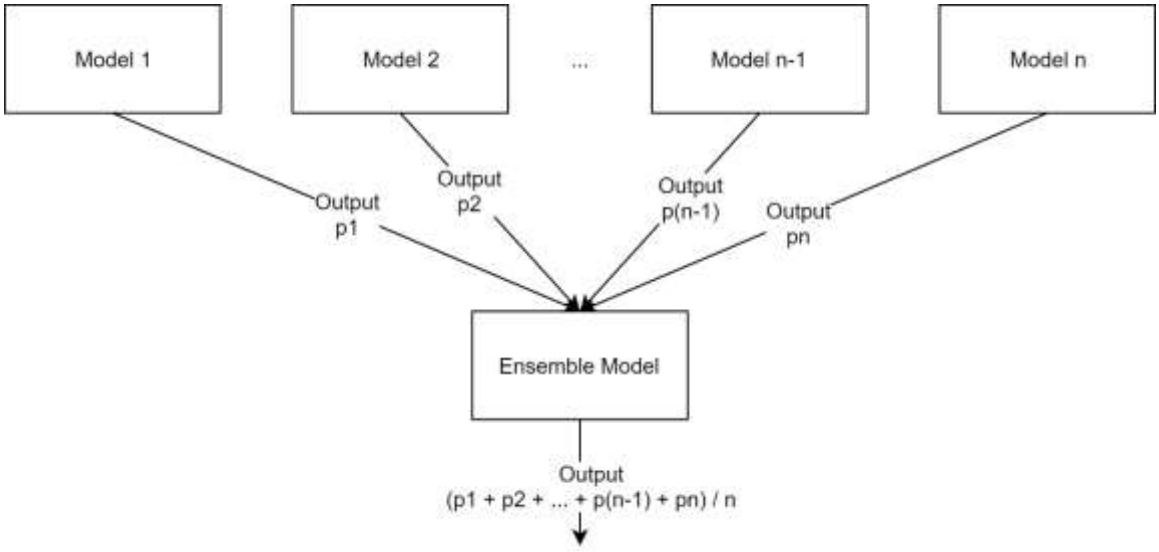

Figure 1 General Structure of the Ensemble Model

3.5 Data

We collected market data from soccer and local horse racing for training and testing purposes. The purposed models in Section 43.5 will be trained and evaluated on these data.

3.5.1 Horse Racing

In Hong Kong, there are nearly 700 horse races conducted at Sha Tin Racecourse and Happy Valley Racecourse per year. Although the Hong Kong Jockey Club is operating the pool betting, offshore bookmakers are offering fixed-odds markets for the local horse racing regularly.

In our experiments, we would focus on the Win market where bettors need to predict the race winner correctly in order to get paid. We collected the odds for the Win markets from a website, which displays the closing odds from 15 bookmakers and the average odds changes over time. The closing odds and the average odds changes of races in 2017/01/01 – 2019/12/31 were collected. However, not every bookmaker would offer markets for local horse racing. Table 1 displays a list of bookmakers which offer markets for Hong Kong horse racing by year. We split the data into training set and testing set where the set will be used in training and performance evaluation respectively.

Training Set: Data from 2017/01/01 – 2018/12/31 (1618 races / 19647 horses)

Testing Set: Data from 2019/01/01 – 2019/12/31 (805 races / 9827 horses)

Year	Count	Bookmakers
2019	12	Bet365, Bet Easy, Betstar, Bluebet, Bookmaker, Ladbrokes, Neds, Pointsbet, Sportsbet, Sportsbetting, Topbetta, Unibet
2018	13	Bet365, Bet Easy, Betstar, Bluebet, Bookmaker, Ladbrokes, Neds, Pointsbet, Sportsbet, Sportsbetting, Topbetta, Ubet, Unibet
2017	9	Bet365, Betstar, Bookmaker, Ladbrokes, Neds, Pointsbet, Sportsbet, Topbetta, Unibet

Table 1 Bookmakers offering markets for local horse racing by year

3.5.2 Soccer

Over/Under markets are our focuses on soccer games. For Over/Under, bookmakers will offer lines to each game. Bettors need to predict the total goal in the game is "over" or "under" their selected lines. We scraped the odds offered by Pinnacle and prices on Betfair exchange from a website. Games from season 2018 - 2019 in 27 different leagues are collected.

Leagues scraped: Serie A, Serie B, La Liga, LaLiga 2, Bundesliga, 2. Bundesliga, Premier League, EFL Championship, Ligue 1, Ligue 2, Eredivisie, Eerste Divisie, Scottish Premiership, Primeira Liga, Belgian First Division A, Allsvenskan, Eliteserien, J.League, J2 League, A-League, Primera División, Campeonato Brasileiro Série A, Major League Soccer, Liga MX, Chilean Primera División, K League 1, Russian Premier Liga

We split the data into training set and testing set:

Training Set: Data before 2017/07/01 (18847 lines)

Testing Set: Data from 2019/07/01 – 2020/03/08 (8567 lines)

4 Proposed Models

4.1 Overview

There are 2 types of models: Closing Model and Continuous Model. Closing Model is for betting just before closing while Continuous Model supports continuous betting until closing. In this section, we will mention their details and results. Due to time and technical limitation, some models are tested in horse racing markets only. Table 2 below shows the sports tested on different models.

	Horse Racing	Soccer
4.2.1 Closing Model: Regression-based	✓	
4.2.2 Closing Model: LSTM-based	✓	✓
4.3.1 Continuous Model: LSTM-based	✓	
4.3.2 Continuous Model: Convolution-based	✓	✓

Table 2 Sports tested on models

The models will be tested with Kelly Betting. It is possible that multiple outcomes from the same game or different games will be picked at the same time. The optimal bet that should be placed on an outcome will also depend on that of others. Modification to the original version of Kelly Formula 3.2.2 is needed to support this kind of simultaneous betting. However, for simplicity, the betting simulations in all experiments assumed that the payouts are executed immediately after the bets are placed. This makes each betting becomes independent and the original Kelly Formula can be applied.

4.2 Closing Model

4.2.1 Regression-based

Regression-based model has been introduced and discussed thoroughly in the Term 1 report, the details will not be repeated in here.

4.2.1.1 Application in Hong Kong Horse Racing

4.2.1.1.1 Results

We used the models to simulate the Kelly Betting on the testing set which includes races from 2019/01/01 – 2019/12/31. There are 805 races and 9827 horses in total. The initial capital is set to be $10,000. The bets are placed on the bookmakers offering the highest closing odds.

Note that the results in the Term 1 Report were simulated on races from 2019/01/01 – 2019/10/01. Due to time limitation, only selected models are rerun using the newer data. The rerun results will be shown below. For the old results, please refer to Section 4.5.5 of the Term 1 report.

Table 3 below shows the returns. Returns with positive gain are colored in green and red for the negatives. In order to demonstrate the positive returns from the models are not by luck, we performed Kelly Betting based on the average closing-odds-implied probability $P_{\text{avg}}(t_0)$. Figure 2 below shows the return by this strategy. Note that Fractional Kelly with fraction 1.0 is equivariant to normal Kelly betting.

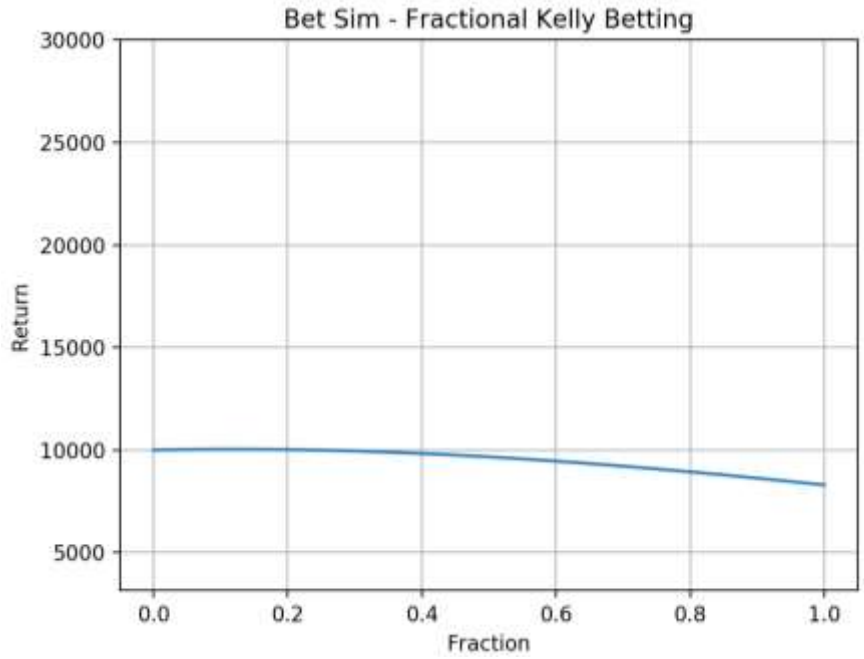

Figure 2 Return by Fractional Kelly Betting using $P_{\text{avg}}(t_0)$ as predictors in horse racing

Model	Return	Model	Return
0-34-4deg	17015	0-39-4deg	21287
0-34-6deg	19246	0-39-6deg	20941
0-34-8deg	18159	**0-39-8deg**	**24307 (highest)**
0-34-10deg	13210	0-39-10deg	19255
0-34-12deg	17251	0-39-12deg	17903
0-34-14deg	14996	0-39-14deg	17472
0-34-16deg	81	0-39-16deg	10185
0-34-18deg	45	0-39-18deg	213

Model	Return	Model	Return
0-59-4deg	9532	0-9-6deg	7198
0-59-6deg	9144	0-19-6deg	9705
0-59-8deg	9151	0-29-6deg	17789
0-59-10deg	8834	0-39-6deg	20941
0-59-12deg	9733	0-49-6deg	11231
0-59-14deg	11196	0-59-6deg	9144
0-59-16deg	7099	0-79-6deg	3892
0-59-18deg	8497	0-119-6deg	1355
Model	**Return**	**Model**	**Return**
0-19-10deg	14553	0-19-16deg	37
0-29-10deg	21673	0-29-16deg	3114
0-39-10deg	19255	0-39-16deg	10185
0-49-10deg	12272	0-49-16deg	10582
0-59-10deg	8834	0-59-16deg	7099
0-79-10deg	4004	0-79-16deg	4905
0-119-10deg	1754	0-119-16deg	1474

Table 3 Results of Closing Model: Regression-Based

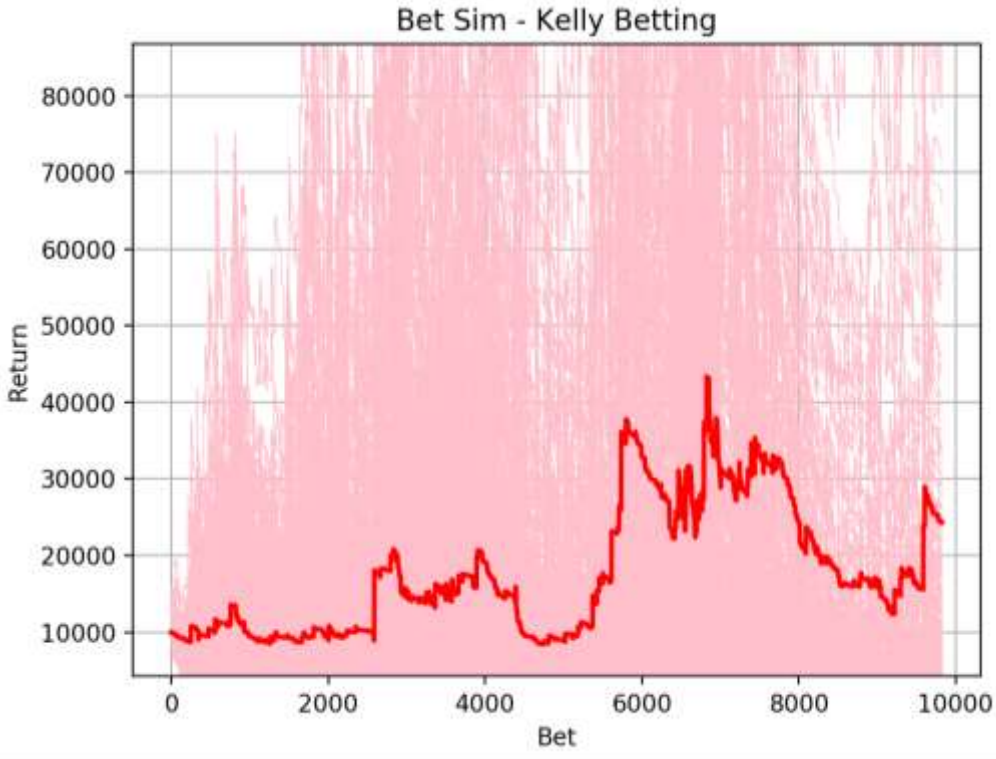

Figure 3 Return by Kelly Betting using ensemble model 0-39-8deg (red) and its members (pink)

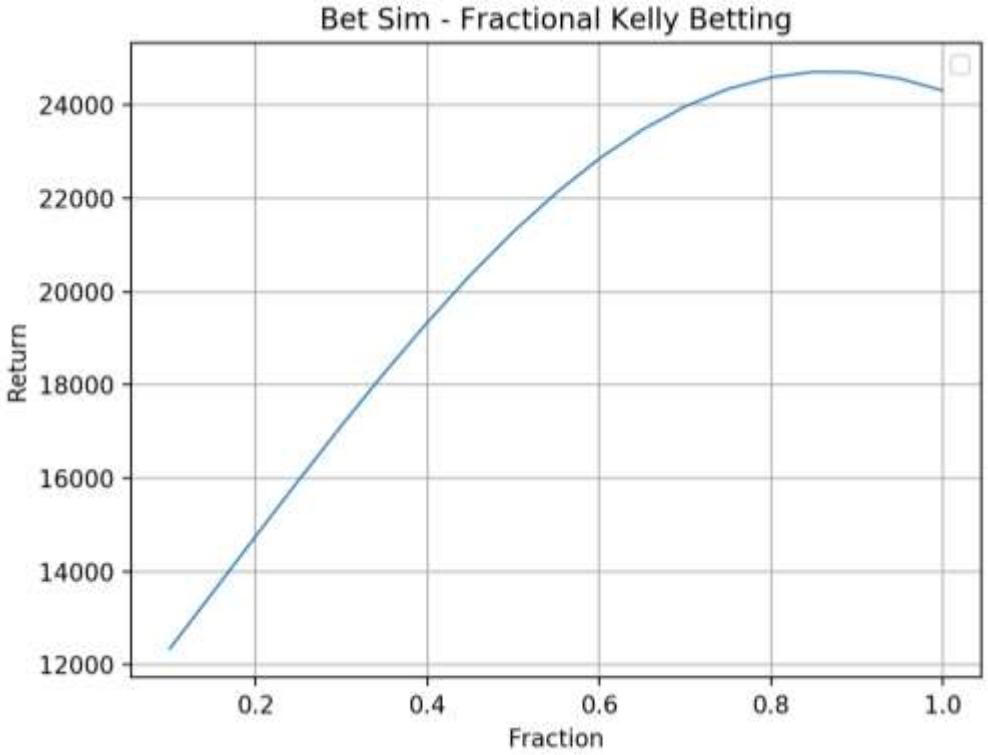

Figure 4 Return by Fractional Kelly Betting using ensemble model 0-39-8deg

4.2.2 LSTM-based

4.2.2.1 Model Structure

To improve the above model, we let the model learn the raw sequence of odds movements directly using Bidirectional LSTMs.

Let us name the block that produces extra inputs using LSTMs be “SeqExtract” for convenience. Figure 5 below shows the structure of SeqExtract that we used. From the figure, we can see that there are *k* series of blocks connected in parallel, where *k* is a parameter that we can tune for.

Figure 5 The structure of SeqExtract

4.2.2.2 Application in Hong Kong Horse Racing

We applied the above model to the same set of horse racing data. We tested the model performance for different *k*, the number of LSTM stacked. Figure 6 below shows the structure of the model used.

Figure 6 Structure of the LSTM-based Closing Model used

4.2.2.2.1 Kelly and Fractional Kelly Results

Each ensemble model includes 300 members. Again, we performed Kelly Betting on the testing set. The initial capital is set to $10000, the same we used before. We would use the best Regression-based model **0-39-8deg** as the baseline. Recall that for model **0-39-8deg**, its return in Kelly Betting is $24307 and the maximum return in Fractional Kelly Betting is $24697 archived by the fraction ≈ 90%. Figure 7 below display the betting simulation results of the LSTM-based models.

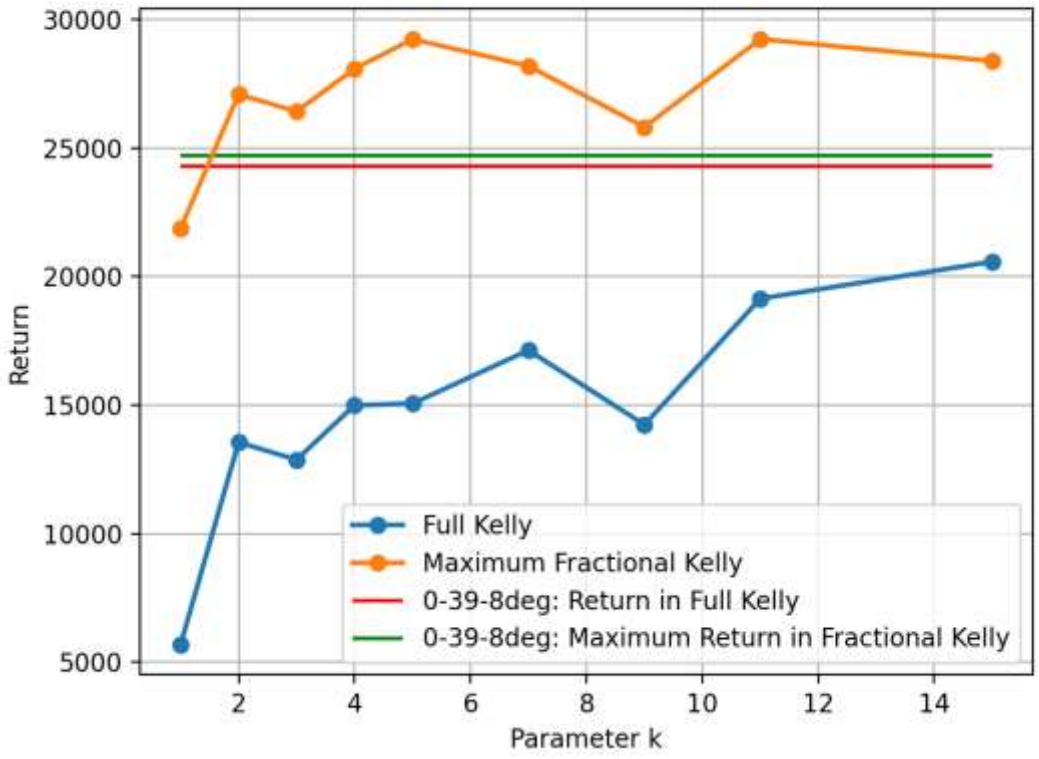

Figure 7 Returns in Kelly and Fractional Kelly Betting for LSTM-based Closing Models in horse racing

4.2.2.2.2 Improved Kelly Results

In Section 3.2.2.2, we introduced an improved version of Kelly Formula that considers the optimal fraction of the bet as well. Now, in this section, we will show a method of applying the improved Kelly. We assume that the predictions from ensemble members are drawn from the probability density function b. By further assuming b is a Beta distribution, we can obtain the b by performing the Beta fit on the predictions from the ensemble members. We evaluated the models above using this improved Kelly. Figure 8 below shows their returns.

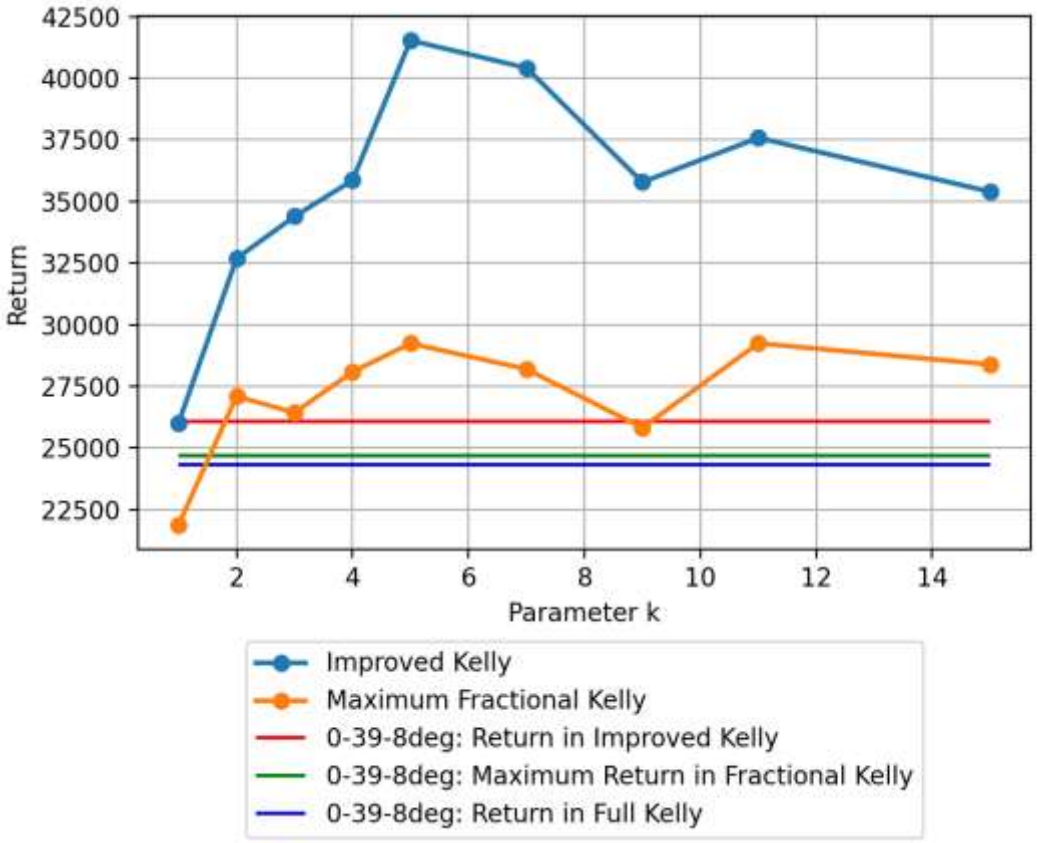

Figure 8 Returns in Improved Kelly and Fractional Kelly Betting for LSTM-based Closing Models in horse racing

4.2.2.3 Application in Soccer

Besides from horse racing, we roughly tested the method in a Soccer market as well. The models would be tested in the market Over/Under that introduced in Section 3.5.2.

4.2.2.3.1 Dataset and Model

There are total of 18847 lines in the training set. For each line, we follow the same procedures in above sections to compute the features.

4.2.2.3.2 Results

Each ensemble model includes 500 members. After training the models, we performed Kelly, Fractional Kelly and the Improved Kelly Betting on the testing set which contains 8567 lines. The highest closing odds (prices) among Pinnacle and Betfair will be chosen to bet against. The initial capital is set to $10000. Note that as Betfair charges commissions (≈5%) for winning bets [8], the Betfair's prices are multiplied by 0.95 in the following betting simulation.

Figure 10 below show the returns given by models with different parameter k. Again, in order to demonstrate the positive returns from the models are not by luck, we performed Kelly Betting based on the average closing-odds-implied probability $P_{\text{avg}}(t_0)$. Figure 9 below shows the return by this strategy. As we can see, this strategy is insufficient to produce positive returns.

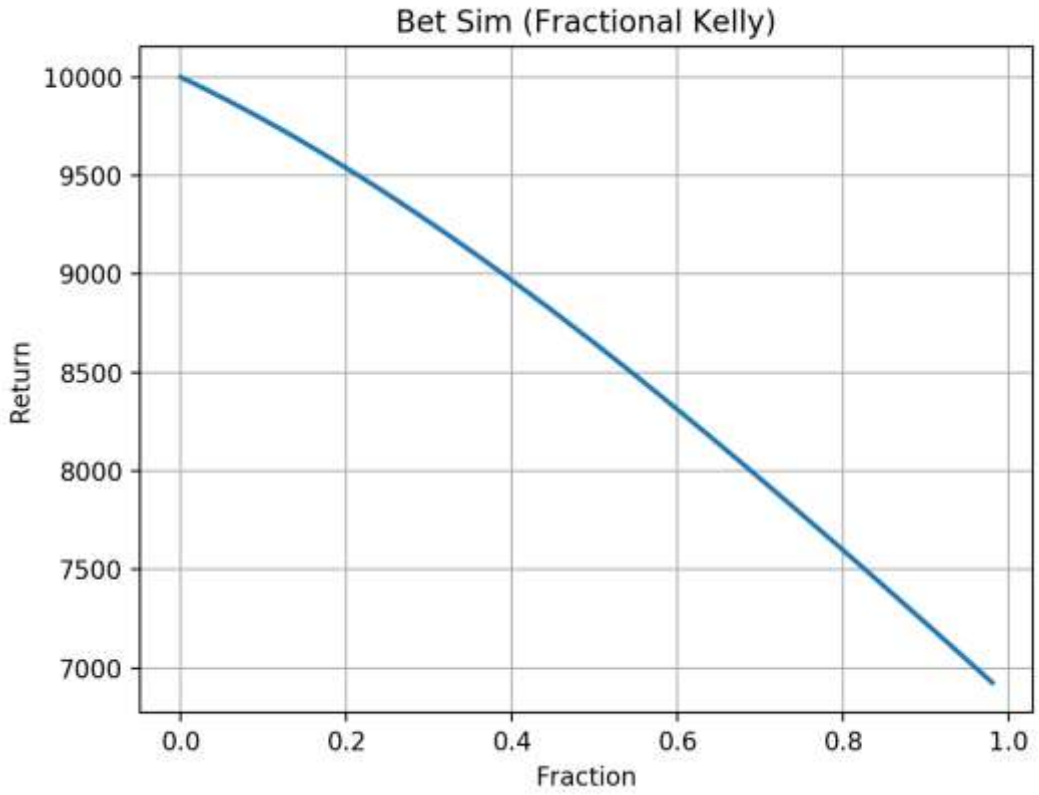

Figure 9 Return by Fractional Kelly Betting using $P_{\text{avg}}(t_0)$ as predictors in Over/Under

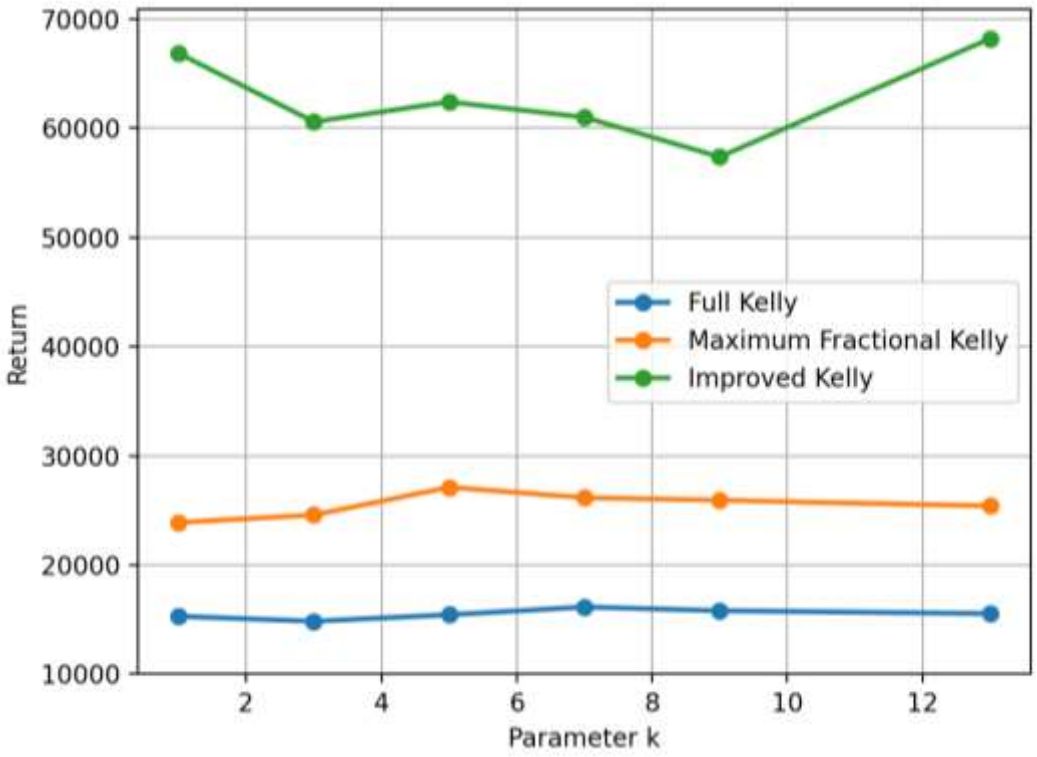

Figure 10 Returns in Improved, Fractional and Full Kelly Betting for LSTM-based Closing Models in Over/Under

4.3 Continuous Model

As betting odds are already available some time before closing, we want to explore if we can produce predictions at different timestep based on the odds at that moment. In this section, we are going to introduce some models we have tested, which support continuous prediction.

4.3.1 LSTM-based

The LSTM-based model here is similar to the one in Closed Model. Now, we would use the horse racing above as an example. Suppose we want to produce minute-by-minute predictions and up to 5 minutes before closing. Let the average odds-implied probability from bookmakers at h minutes before closing be $P_{\text{avg}}(t_h)$ and the odds considering period is from 0 minute – n minutes before closing. We will then create a total of 6 sequences of odds-implied probability P_{avg} and extra features for different timesteps and mask out the unseen features with a special value -1, which has no meaning to the probability sequence.

4.3.1.1 Application in Hong Kong Horse Racing

4.3.1.1.1 Forming the Dataset

We would use the same procedures mentioned in the example above to form the features. In this experiment, we will let the model produce minute-by-minute predictions and up to 29 minutes before closing. The odds considering period is set to be up to 60 minutes before closing, which is the time most of the bookmakers have started to offer odds.

4.3.1.1.2 Structure of Model

below shows the structure of the model. Note that the parameter k of SeqExtract is set to be a fixed value 5.

Figure 11 Structure of the LSTM-based Continuous Model for horse racing

4.3.1.1.3 Results

We trained 100 members for each ensemble model. In order to show the models' performance, we compute their Binary Cross Entropy (BCE) on the testing set. The BCE value can be understood as the loss in maximum likelihood estimation. The lower the BCE value, the model better fit the testing set. We also compute the BCE of average odds-implied probability P_{avg} for comparison. Figure 12 below shows their BCEs.

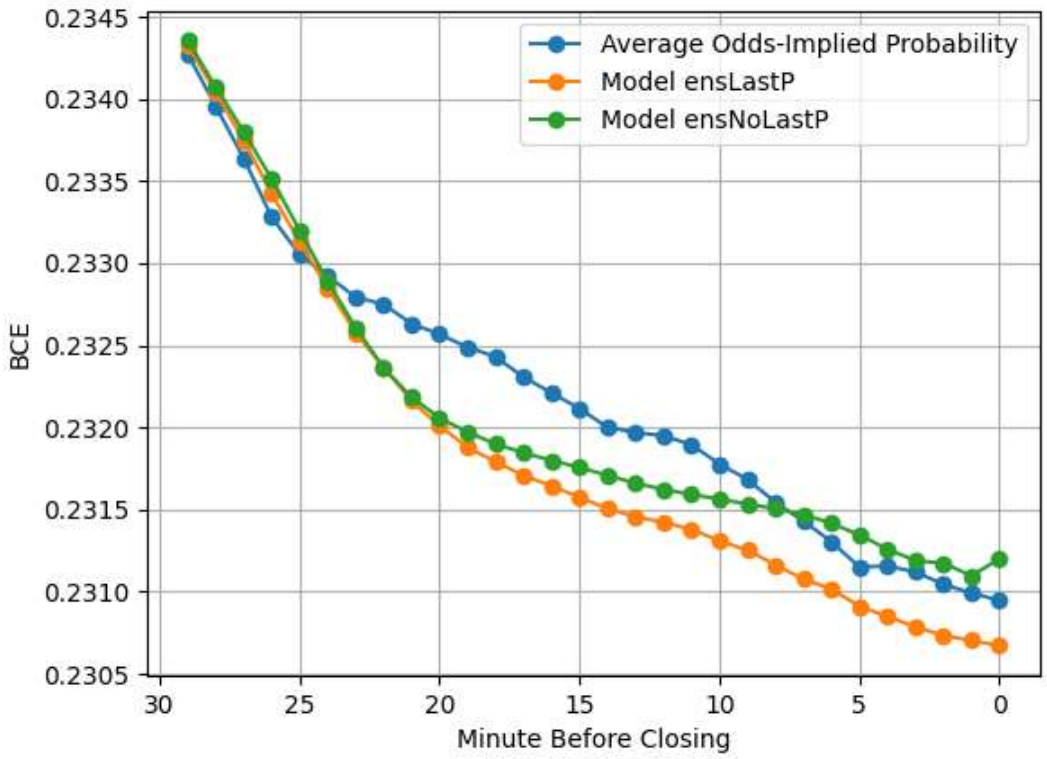

Figure 12 BCE of Model ensLastP, ensNoLastP in LSTM-based Continuous Model

4.3.1.2 Limitation

Although we show the model can outperform the odds in horse racing, there is a main limitation of forming the dataset. To archive continuous prediction, we have to create records observed at different timesteps. This step significantly increases the data size and make the method unsuitable for long period of continuous prediction. Unlike the local horse racing where the bookmakers start offering odds a few hours before the race, bookmakers usually start offering odds several days or even months in advance for Soccer games. This is the reason why we did not evaluate the method on Soccer markets.

4.3.2 Convolution-based

Convolution-based model is designed for long period of continuous prediction. The idea is we feed the model with sequences of odds-implied probability and let it output series of predicted probability that preserves time dependency In order to control the receptive fields, we adopt Casual Convolution instead so that we can manipulate the receptive fields.

4.3.2.1 Application in Soccer

4.3.2.1.1 Model Structure

Figure 13 below shows the structure of the model used in Over/Under. As we can see, each input sequence is passed to a Dense layer and a Convolutional layer. The Convolutional layer provides the ability to lookback while the Dense layer emphasizes the latest odds as it has no ability to lookback. Besides, the Dense layer also increases the dimension of the sequence for applying the Addition layer after. After that, the Addition layer is used to merge the two output vectors from the together.

The first Addition layer merges the 4 sequences produced by Pinnacle's odds and Betfair's prices and then passes the merged vector to another set of Dense layer and Convolutional layer for exploring their interrelationships. Finally, the outputs are concatenated with the raw input sequences to output the final forecasts.

Figure 13 Structure of the Convolution-based Continuous Model used in Over/Under

4.3.2.1.2 Forming the Dataset

We would input 2 sequences of odds-implied probability which cover 0 minute - 1439 minutes before closing in minute-by-minute interval. Therefore, the model is capable to produce forecasts starting from almost 24 hours before kickoff.

Here are the features for each timestep:

1. Odds-implied probability
2. Payout
3. Percentage change of the odds-implied probability
4. Minute before kickoff

4.3.2.1.3 Results: Binary Cross Entropy Test

We trained ensemble model with different window size. Each ensemble model contains 200 members. After that, we computed their Binary Cross Entropy on the testing set for comparison. Figure 14 and Figure 15 below plot their BCEs at different timesteps.

Figure 14 BCE Test for Convolution-based Continuous Models used in Over/Under

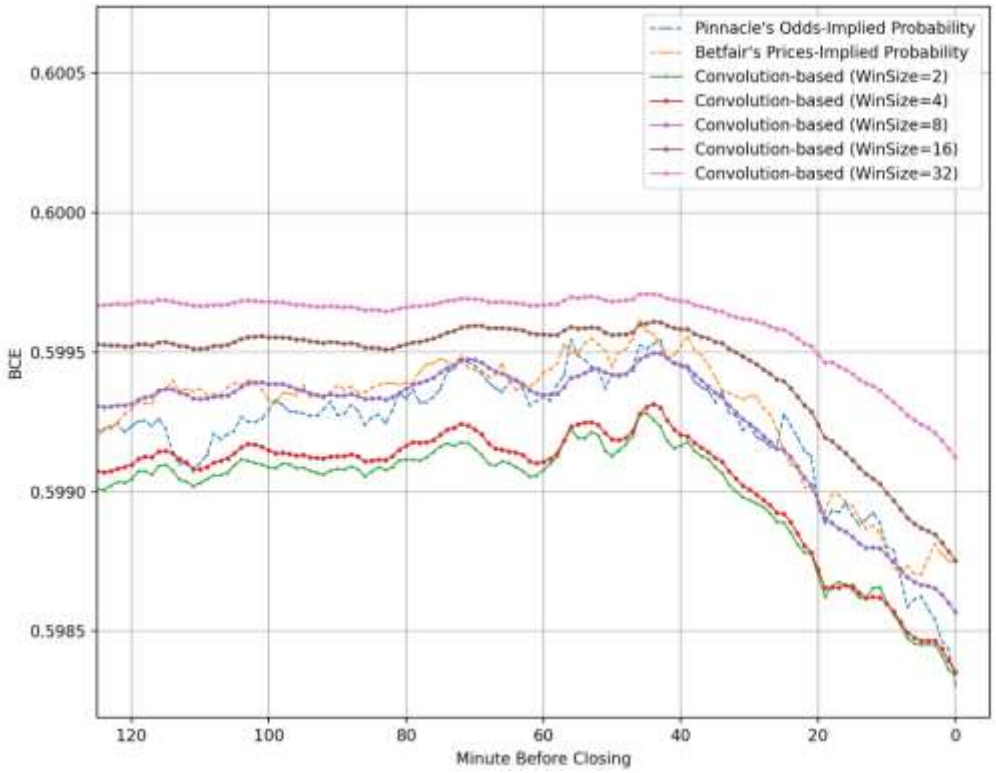

Figure 15 BCE Test for Convolution-based Continuous Models used in Over/Under (0-120 min)

Since the results show selecting smaller window size is better, we are also interested in what would happen if the window size is chosen to be 1. Figure 16 below plots the BCE of models with window size 1 and 2.

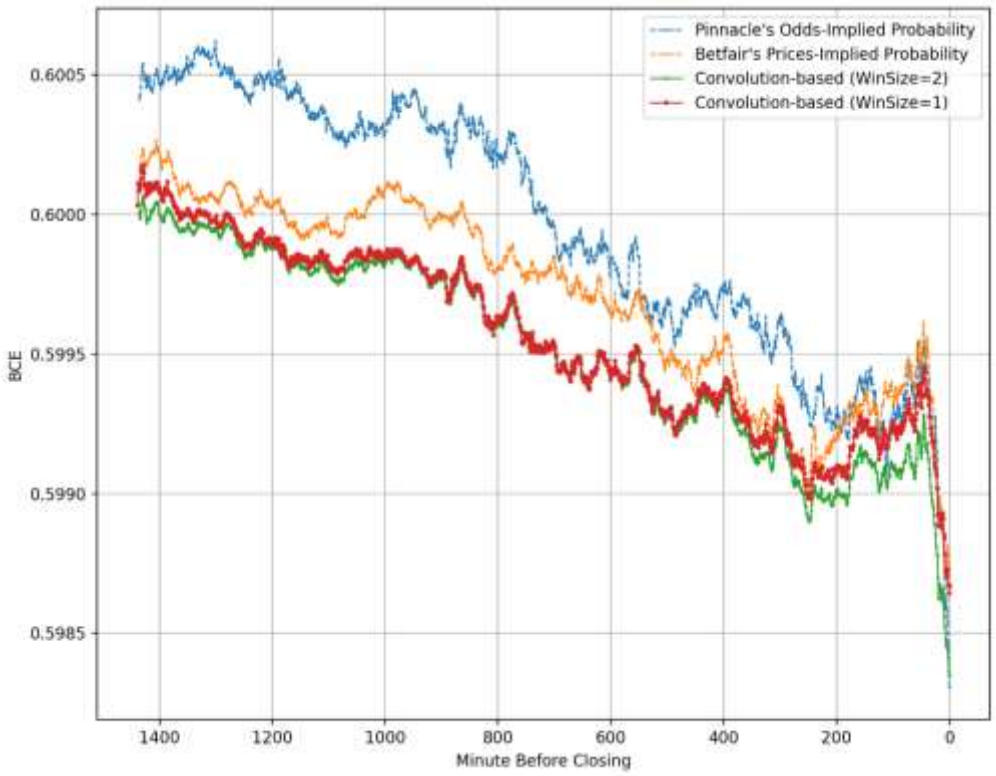

Figure 16 BCE Test for Convolution-based Continuous Models with window size 1,2 used in Over/Under

4.3.2.2 Application in Hong Kong Horse Racing

We also roughly tested the model on local horse racing. Since we only have the average odds data for horse racing, we have to slightly modify the above model which accepts 2 odds sequences. We computed the same features as the above Soccer experiment using the average odds. The window size of the convolutional layers is set to be 2, which gives the best result for Over/Under. For comparison, we use the same odds period which is the period for LSTM-based Continuous Model in Section 4.3.1.

We trained an ensemble model with 200 members and computed its BCE for comparison. Figure 17 below plots it BCE and the BCE of the LSTM-based model in Section 4.3.1. From the results, we can see that LSTM-based model is clearly better. Even so, the Convolution-based model is also capable to outperform the odds in most of the time.

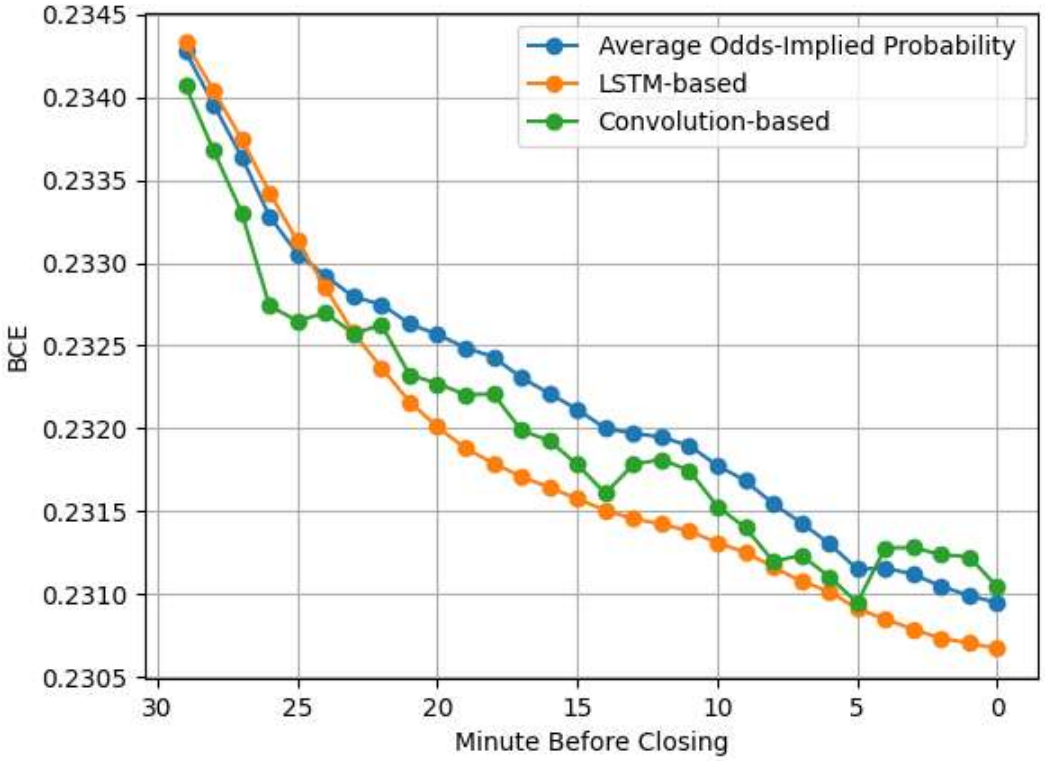

Figure 17 BCE of LSTM-based and Convolution-based Continuous Model used in horse racing

References

[1] C. Gough, "Sports Betting and Gambling Market/Industry - Statistics & Facts," Statista, 29 3 2019. [Online]. Available: https://www.statista.com/topics/1740/sports-betting/.

[2] "Odds Quality - Bookmaker Payout Ratings," OddsPortal, [Online]. Available: https://www.oddsportal.com/odds-quality/. [Accessed 29 10 2019].

[3] "Winners are Welcome - Successful players sharpen our odds," Pinnacle, [Online]. Available: https://www.pinnacle.com/en/promotions/winners-welcome.

[4] ""What Are My Odds?" - William Benter ICCM 2004," [Online]. Available: https://youtu.be/YOVrZrJ-wtc?t=2097. [Accessed 2019 10 29].

[5] J. L. Kelly, The Bell System Technical Journal, 1956.

[6] W. Benter, "Computer Based Horse Race Handicapping and Wagering Systems: A Report," 1994.

[7] Baker, Rose; I.G. Mchale, "Optimal Betting Under Parameter Uncertainty: Improving the Kelly Criterion," 2013.

[8] "Betfair Help Centre: Exchange: What is Commission and how is it calculated?," [Online]. Available: https://en-betfair.custhelp.com/app/answers/detail/a_id/413/~/exchange%3A-what-is-commission-and-how-is-it-calculated%3F.

[9] Giovanni Capobianco, Umberto Di Giacomo, Francesco Mercaldo, Vittoria Nardone, Antonella Santone, "Can Machine Learning Predict Soccer Match Results?," ICCART, 2019.

[10] Nazim Razali, Aida Mustapha, Faiz Ahmad Yatim, Ruhaya Ab Aziz, "Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL)," IOPscience, 2017.

[11] Rahul Baboota, Harleen Kaur, "Predictive analysis and modelling football results using machine learning approach for English Premier League," International Journal of Forecasting, 2018.

[12] Ali Reza Khanteymoori, Elnaz Davoodi, "Horse Racing Prediction Using Artificial Neural Networks," Recent Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing, 2010.

[13] Y. LIU, "Predicting Horse Racing Result with Machine Learning," 2018.

[14] LAU Ming Hei, CHENG Tsz Tung, "Predicting Horse Racing Result using Tensorflow," 2016.

[15] Elroy Dimson, Massoud Mussavian, "MARKET EFFICIENCY," in *THE CURRENT STATE OF BUSINESS DISCIPLINES*, SPELLBOUND PUBLICATIONS, 2000, pp. 959 - 970.

[16] Guy Elaad, J. James Reade, Carl Singleton, "Information, Prices and Efficiency in An Online Betting Market," Applied Economics Letters, 2019.

[17] Štefan Lyócsa, Tomáš Výrost, "To bet or not to bet: a reality check for tennis betting market efficiency," Applied Economics, 2017.

[18] Luca Rebeggiani, Johannes Gross, "Chance or Ability? The Efficiency of the Football Betting Market Revisited," IASE Conference, 2018.

[19] Giovanni Angelini, Luca De Angelis, "Efficiency of online football betting markets," SSRN Electronic Journal, 2017.

[20] Lisandro Kaunitz, Shenjun Zhong, Javier Kreiner, "Beating the bookies with their own numbers - and how the online sports betting market is rigged," 2017.

[21] Joseph Buchdahl, "The maths behind Pinnacle's "winners welcome" policy," 20 7 2016. [Online]. Available: https://www.pinnacle.com/en/betting-articles/educational/why-pinnacle-doesnt-close-or-limit-accounts. [Accessed 29 10 2019].

[22] John M. Gandar, William H. Dare, Craig R. Brown, Richard A. Zuber, "Informed Traders and Price Variations in the Betting Market for Professional Basketball Games," THE JOURNAL OF FINANCE, 1998.

[23] Kevin Krieger, Andy Fodor, "Price movements and the prevalence of informed traders: The case of line movement in college basketball," Journal of Economics and Business, 2013.

[24] S. R. Clarke, "ADJUSTING TRUE ODDS TO ALLOW FOR VIGORISH," Proceedings of the 13th Australasian Conference on Mathematics and Computers in Sport, 2016.

[25] S. R. Clarke, "Adjusting Bookmaker's Odds to Allow for Overround," American Journal of Sports Science, 2017.

[26] H. S. Shin, "Prices Of State Contingent Claims With Insider Traders, And The Favourite-Longshot Bias," in *The Economic Journal*, 1992, pp. 426-435.

[27] H. S. Shin, "Measuring the Incidence of Insider Trading in a Market for State-Contingent Claims," in *The Economic Journal*, 1993, pp. 1141-1153.

[28] M Viney, A Bedford, E Kondo, *Incorporating over-round into in-play markov chain models in tennis,* Las Vegas, USA: 15th International Conference on Gambling & Risk Taking, 2013.

[29] Jacek Grekow, Karol Odachowski, "Using Bookmaker Odds to Predict the Final Result of Football Matches," in *Knowledge Engineering, Machine Learning and Lattice Computing with Applications*, Springer, Berlin, Heidelberg, 2012, pp. 196-205.

[30] Engin Esme, Mustafa Servet Kiran, "Prediction of Football Match Outcomes Based on Bookmaker Odds by Using k-Nearest Neighbor Algorithm," *International Journal of Machine Learning and Computing,* vol. 8, 2018.

[31] Owens, R G, Hewson, T D, ECMWF Forecast User Guide, 2018.

[32] Jr. Kolczynski, Stauffer David, Haupt Sue & Altman Naomi, Deng Aijun, "Investigation of Ensemble Variance as a Measure of True Forecast Variance," *Monthly Weather Review,* 2011.

[33] Samuels, Peter & Gilchrist, Mollie, "Statistical Hypothesis Testing," 2014.