International Journal of Forecasting 35 (2019) 783–796

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Exploiting sports-betting market using machine learning

Ondřej Hubáček *, Gustav Šourek, Filip Železný

Czech Technical University in Prague, Czech Republic

ARTICLE INFO

Keywords:
Decision making
Evaluating forecasts
Neural networks
Sports forecasting
Probability forecasting

ABSTRACT

We introduce a forecasting system designed to profit from sports-betting market using machine learning. We contribute three main novel ingredients. First, previous attempts to learn models for match-outcome prediction maximized the model's predictive accuracy as the single criterion. Unlike these approaches, we also reduce the model's correlation with the bookmaker's predictions available through the published odds. We show that such an optimized model allows for better profit generation, and the approach is thus a way to 'exploit' the bookmaker. The second novelty is in the application of convolutional neural networks for match outcome prediction. The convolution layer enables to leverage a vast number of player-related statistics on its input. Thirdly, we adopt elements of the modern portfolio theory to design a strategy for bet distribution according to the odds and model predictions, trading off profit expectation and variance optimally. These three ingredients combine towards a betting method yielding positive cumulative profits in experiments with NBA data from seasons 2007–2014 systematically, as opposed to alternative methods tested.

1. Introduction

Sports betting means placing a wager on a subset of outcomes of random sports events, each of which is associated with a corresponding profit predefined by a *bookmaker*. If the outcome is guessed correctly, the bettor wins back the wager plus the profit, otherwise (s)he loses the wager to the bookmaker. Historically, bookmakers operated betting shops, but with the expansion of the Internet, most bookmakers these days operate online through betting sites. A variety of betting opportunities are offered. In our work, we focus on *moneyline bets*. To win a moneyline bet, the bettor needs to predict the winner of the game. For each of the two possible outcomes of a two-team match, the bookmaker sets the corresponding odds; the latter multiplies with the wager towards the potential profit. So if the bookmaker sets the odds to 1.8 for the home team to win, a bettor places the wager of 100 Eur on that outcome, and the home team actually wins, the bettor's profit will be $1.8 \times 100 - 100 = 80$ Eur. Naturally, both bettors and bookmakers try to maximize their profits.

If the odds were *fair*, their inverse value could be interpreted as the probability of the outcome as estimated by the bookmaker. In practice, however, this is not the case. For instance, when the bookmaker is indifferent as to which outcome is more probable, (s)he does not set the fair odds as $2.0 : 2.0$, and rather offers a lower portion of profit such as $1.95 : 1.95$. The absolute difference between 1 (the sum of true probabilities) and the probabilities implied by inverted odds is called the *margin*. In our example, the bookmaker's margin would be $1.95^{-1} + 1.95^{-1} - 1 \approx 2.5\%$. Given the margin, combined with the bookmaker's professional experience in forecasting the game outcomes, it is extremely difficult for a bettor to profit on moneyline betting systematically.

Here we design a profitable sports-betting system. Its three main properties which enable profit generation, and at the same time the most important contributions of this paper w.r.t. the state of the art, are as follows.

First, as many studies before (Section 2 provides a brief review), we use machine learning to develop an outcome

* Corresponding author.
E-mail address: hubacon2@fel.cvut.cz (O. Hubáček).

https://doi.org/10.1016/j.ijforecast.2019.01.001
0169-2070/

prediction model. However, in previous work the single emphasis has been on the predictive accuracy of such a model. Here we argue that even an accurate model is unprofitable as long as it is correlated with the bookmaker's model: if our guesses coincide with the bookmaker's, we will be losing money due to the margin. Thus we elaborate various measures to *decorrelate* the learned model from the bookmaker's model (estimable roughly from the assigned odds), while maintaining adequate prediction accuracy.

Secondly, we propose an innovative method to learn the outcome prediction model from features describing past performance statistics of individual players in both teams of the match. The approach uses *convolutional* neural network which recently achieved significant successes in visual and language data processing. Here, a convolutional network layer operates on the matrix of players and their statistics, acting as an aggregator of player-level features towards team-level features propagated further through the network towards the prediction on the output. The aggregation pattern defined by the convolution layer may be complex and is itself learned from training data.

Thirdly, we adopt the concepts of *modern portfolio theory* in the design of the betting *strategy*. Such a strategy accepts the (probabilistic) model predictions for a set of games and proposes a set of bets on these games. The portfolio theory originated in the field of economics and its application in sports-betting is novel. The proposed strategy distributes the bets under the optimal trade-off between profit expectation and profit variance. This supersedes heuristic bet spreading strategies published in previous work as well as a naive expectation–maximization strategy.

The rest of the paper is organized as follows. In the next section, we review the relevant prior work. Section 3 defines a formal framework for the problem. In Section 4, we elaborate the various kinds of predictive models employed. Section 5 formalizes several existing betting strategies and develops the novel strategy of portfolio optimization. In Section 6, we investigate selected aspects of the proposed models and strategies through simulated experiments. Section 7 then provides a thorough experimental evaluation on the US National Basketball League matches throughout seasons 2006 to 2014. In Section 8, we conclude the paper.

2. Related work

Several studies investigated the strategies of bookmakers and bettors. Focusing on the US National Football League (NFL), Levitt (2004) traced how odds are set and concluded that bookmakers rely on their ability to outperform an average bettor in outcome forecasting rather than on earning money by balancing weighted wages and profiting from the margin. This hypothesis was subjected to further scrutiny by Paul and Weinbach (2007), challenging Levitt's dataset informativeness as consisting only of bets from entry-fee betting tournaments and a limited numbers of participants. However, the conclusions essentially confirmed those of Levitt. Although the hypothesis was not confirmed in basketball (Paul & Weinbach, 2008) using National Basketball League (NBA) data, the disagreement can be explained by the smaller NBA betting market. The recent inquiry (Paul & Weinbach, 2010) into the behavior of bettors with data from NBA and NHL season 2008/09 proposes that most bettors act more like fans than investors. Combined with the conclusion of Levitt (2004), this motivates the question whether the bookmaker can be exploited with an emotionless statistical model.

The idea that a statistical model might outperform experts was first tested in Forrest and Simmons (2000). The experts were found unable to process publicly available information efficiently. Signs of using information independent of publicly available data were rare. The study deemed it unlikely that experts would outperform a regression model. Forrest, Goddard, and Simmons (2005) challenged the thesis that a statistical model has an edge over tipsters. They examined the performance of a statistical model and bookmakers on 10 000 soccer matches and concluded that bookmakers were on par with a statistical model.

Song, Boulier, and Stekler (2007) analyzed prediction accuracy of experts, statistical models and opening betting lines on two NFL seasons. There was a little difference between statistical models and experts performance, but both were outperformed by the betting line. Spann and Skiera (2009) compared prediction accuracy of prediction markets, betting odds and tipsters. Prediction markets and betting odds proved to be comparable in terms of prediction accuracy. The forecasts from prediction markets would be able to generate profit against the betting odds if there were not for the high fees. On the other hand, tipsters performed rather poorly in this comparison.

Stekler, Sendor, and Verlander (2010) focused on several topics in horse racing and team sports. Forecasts were divided into three groups by their origin – market, models, experts. Closing odds proved to be better predictors of the game outcome than opening odds. The most important conclusion was that there was no evidence that a statistical model or an expert could consistently outperform betting market.

Franck, Verbeek, and Nüesch (2010) inspired by results of prediction markets in different domains such as politics, compared performance of betting exchange against the bookmaker on 3 seasons of 5 European soccer leagues. The prediction market was superior to the bookmaker in terms of prediction accuracy. A simple strategy based of betting on the opportunities where the average odds set by the bookmakers were higher than the odds in prediction market was profitable in some cases.

Angelini and De Angelis (2018) examined effectiveness of 41 bookmakers on 11 European major leagues over a period of 11 years. Some of the markets turned out to be inefficient, since a trivial strategy of betting on opportunities with odds in certain range led to positive profit. For NBA with Pinnacle odds, however, it was shown in Hubáček (2017) that this is not possible and the bookmaker cannot be exploited that simply.

2.1. Predictive models

The review Haghighat, Rastegari, and Nourafza (2013) of machine learning techniques used in outcome predictions of sports events points out the prevailing poor results

of predictions and the small sizes of datasets used. For improving the prediction accuracy the authors suggested to include player-level statistics and more advanced machine learning techniques.

Loeffelholz, Bednar, and Bauer (2009) achieved a remarkably high accuracy of over 74% using neural network models, however their dataset consisted of only 620 games. As features, the authors used seasonal averages of 11 basic box score statistics for each team. They also tried to use average statistics of past 5 games and averages from home and away games separately but reported no benefits.

Ivanković, Racković, Markoski, Radosav, and Ivković (2010) used ANNs to predict outcomes of basketball games in the League of Serbia in seasons 2005/06–2009/10. An interesting part of the work was that effects of shots from different court areas were formalized as features. With this approach, the authors achieved the accuracy of 81%. However, their very specific dataset makes it impossible to compare the results with other research.

Miljković, Gajić, Kovačević, and Konjović (2010) evaluated their model on NBA season 2009/10. Basic box score statistics were used as features, as well as win percentages in league, conference or division and in home/away games. A Naive Bayes classifier in 10-fold cross-validation achieved mean accuracy of 67%.

Puranmalka (2013) used play-by-play data to develop new features. The main reason why features derived from such data are superior to box score statistics is that they include a context. Out of Naive Bayes, Logistic Regression, Bayes Net, SVM and k-nn, the SVM performed best, achieving accuracy over 71% in course of 10 NBA season from 2003/04 to 2012/13.

Zimmermann, Moorthy, and Shi (2013) leveraged multilayer perceptrons for sports outcome predictions. They proposed the existnece of a *glass ceiling* of about 75% accuracy based on results achieved by statistical models in numerous different sports. This glass ceiling could be caused by using similar features in many papers. They also argued that the choice of features is much more important than the choice of a particular machine learning model.

Vračar, Štrumbelj, and Kononenko (2016) made use of play-by-play data to simulate basketball games as Markov processes. Analysis of the results showed that a basketball game is a homogeneous process up to the very beginning and end of each quarter. Modeling these sequences of the game had a large impact on forecast performance. The author saw the application of their model not only in outcome prediction before the game but also in in-play betting on less common bets (number of rebounds/fouls in specific period of the game).

Maymin (2017) tested profitability of deep learning models trained on different datasets during the course of a single NBA season. In the paper, positive profits were only achievable with the use of detailed features extracted by experts from video recordings, while models trained using standard box-score statistics terminated with significant loss.

Constantinou, Fenton, and Neil (2013) designed an ensemble of Bayesian networks to assess soccer teams' strength. Besides objective information, they accounted for the subjective type of information such as team form, psychological impact, and fatigue. All three components showed a positive contribution to models' forecasting capabilities. Including the fatigue component provided the highest performance boost. Results revealed conflicts between accuracy and profit measures. The final model was able to outperform the bookmakers.

Sinha, Dyer, Gimpel, and Smith (2013) made use of twitter posts to predict the outcomes of NFL games. Information from twitter posts enhanced forecasting accuracy, moreover, a model based solely on features extracted from tweets outperformed models based on traditional statistics.

3. Problem definition

Each round of the league consists of n matches. Each match has two possible outcomes, *home team wins* and *home team loses*. The bookmaker assigns *odds* $o_i \in R$, $o_i > 1$ to each of the $2n$ outcomes.

We assume that the bettor places an amount $b_i \in [0, 1]$ on each of the $2n$ outcomes, wherein even two mutually exclusive outcomes (same game) may each receive a positive bet. The normalization of bets to the real unit interval is for simplicity, and as such the bets can be interpreted as *proportions* of a fixed bettor's budget, which is to be exhausted in each round. A popular alternative approach, often presented in related works (Boshnakov, Kharrat, & McHale, 2017), is reinvestment of all the previously accumulated profits (Kelly, 1956). However, we posit

$$\sum_{i=1}^{2n} b_i = 1 \tag{1}$$

The bettor retrieves $o_i b_i$ for outcome i if the latter came true in the match, and zero otherwise. Let p_i be the probability of the ith outcome. The bettor's *profit* is thus

$$P_i = \begin{cases} o_i b_i - b_i & \text{with probability } p_i \\ -b_i & \text{with probability } 1 - p_i \end{cases} \tag{2}$$

so the expected profit is

$$\mathrm{E}[P_i] = p_i(o_i b_i - b_i) - (1 - p_i)b_i = (p_i o_i - 1)b_i \tag{3}$$

and the cumulative profit

$$P = \sum_{i=1}^{2n} P_i \tag{4}$$

from all bets in the round thus has the expectation

$$\mathrm{E}[P] = \mathrm{E}\left[\sum_{i=1}^{2n} P_i\right] = \sum_{i=1}^{2n} \mathrm{E}[P_i] \tag{5}$$

Our goal is to devise a *betting strategy* which prescribes the bets given the known odds and

$$\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{2n} \tag{6}$$

which are estimates of the unknown probabilities $p_1, p_2, \ldots, p_{2n}$ of the $2n$ outcomes. The problem thus breaks down into two tasks: designing an estimator of the probabilities (6), and designing a betting strategy that uses these estimates along with the bookmaker's odds.

The former task should result in a function

$$\hat{p} : D \to [0, 1] \tag{7}$$

which estimates the probability of the home-team winning from some data $d \in D$ relevant to the match and available prior to it; D represents the domain of such background data. Assume (6) are ordered such that for each $k \in \{1, 2, \ldots, n\}$, $\hat{p}_{2k-1}$ ($\hat{p}_{2k}$, respectively) estimates the probability that the home team wins (loses, respectively) in the kth match described by data d_k. Then (6) are obtained from the function (7) as

$$\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4, \ldots = \hat{p}(d_1), 1 - \hat{p}(d_1), \hat{p}(d_2), 1 - \hat{p}(d_2), \ldots \tag{8}$$

The $\hat{p}$ function is assumed to be *learned* from data sampled from D, but also conveying the known match outcomes, i.e. from a finite sample from $D \times \{0, 1\}$. A natural requirement on $\hat{p}$ is that it estimates the true probabilities accurately. However, the precise requirements on $\hat{p}$ as well as the nature of D will be explored in the subsequent section.

The second task is to design a betting strategy, i.e., a function

$$\vec{b} : R^{2n} \times [0, 1]^{2n} \to [0, 1]^{2n}. \tag{9}$$

which for known odds $o_1, o_2, \ldots, o_{2n}$ and probability estimates (6), proposes the bets

$$\vec{b}(o_1, o_2, \ldots, o_{2n}, \hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{2n}) = b_1, b_2, \ldots, b_{2n} \tag{10}$$

subject to (1). A natural requirement on $\vec{b}$ is that the bets proposed should lead to a high expected profit (5). We will elaborate some more refined requirements in Section 5.

4. Prediction model

4.1. Data features

The information we use for predicting the outcome of a match combines data relating to the home team and those pertaining to the visiting team. For each of the two, we aggregate various quantitative measures of the team's performance in all of its preceding matches since the beginning of the season in which prediction takes place.[1] The entire range of these measures is described in the appendix. Current seasons are commonly deemed the most relevant time-windows for player and team performance prediction. The seasonal aggregation is conducted as follows. All variables depending on the match duration are divided by the duration in minutes, and for the seasonal aggregate, we consider the average of these per-minute values. Such variables are marked as "per-minute" in the appendix. For the remaining variables, the median value is considered instead.

The inputs $d \in D$ to the predictive model $\hat{p}$ (7) are tuples of real-valued features constructed out of the said season-aggregated data. Some of the variables in the latter pertain to individual players and others relate to the whole team. Consequently, we distinguish two levels of feature-based description. In the fine-grained *player level*, we collect all player-related variables as individual features, whereas the *team-level* description involves only the team-level variables as features.

Besides the historical track data considered above, the bookmaker's odds assigned to a match represent another piece of information potentially relevant to the prediction of its outcome. While the odds clearly present a very informative feature, their incorporation in a model naturally increases the undesirable correlation with the bookmaker (Section 4.4). Whether to omit or include the odds as a feature thus remains questionable and so we further consider both the options in the experiments.

[1] A few initial games of the season are thus not included among training instances and serve only to collect the statistics. We will quantify this arrangement for a particular testing data set in Section 7.1.

Table 1
The architecture and meta-parameters of the neural predictive models considered.

Meta-parameter	Standard (team-level)	Convolutional (player-level)
Architecture	D64-D32-D16-D1	C1-D64-D16-D1
Activations	tanh	tanh
Dropout	0.2	0.2
L2 regularization	0.0001	0.001

4.2. Logistic regression model

Given (7), we looked for a class of models with a continuous output in the [0, 1] interval. *Logistic Regression* is a simple such instance, which we adopt as the baseline prediction method. It can be viewed as an artificial neural network with only one neuron, using the sigmoid as the activation function.

4.3. Neural model

As an alternative prediction method, we explored two variants of a neural network. The first has a standard (deep) feed-forward architecture with 4 dense layers, while the second one uses a *convolutional layer* (LeCun, Bengio, et al., 1995) followed by 3 dense layers. Table 1 describes the architectures and the relevant meta-parameters of the two neural models.

The standard feed-forward network is intended for the team-level feature data. The convolutional network is specifically designed for the player-level data to deal with the large number of features involved. The principle of its operation, inspired by well-known applications of convolutional networks for visual data processing, is explained through Fig. 1. Intuitively, the convolutional layer may be perceived as a bridge from player-level variables to a team-level representation. However, whereas team-level variables already present in the data are simple sums or averages over all team's players, the convolution layer provides the flexibility to form a more complex aggregation pattern, which itself is learned from the training data.

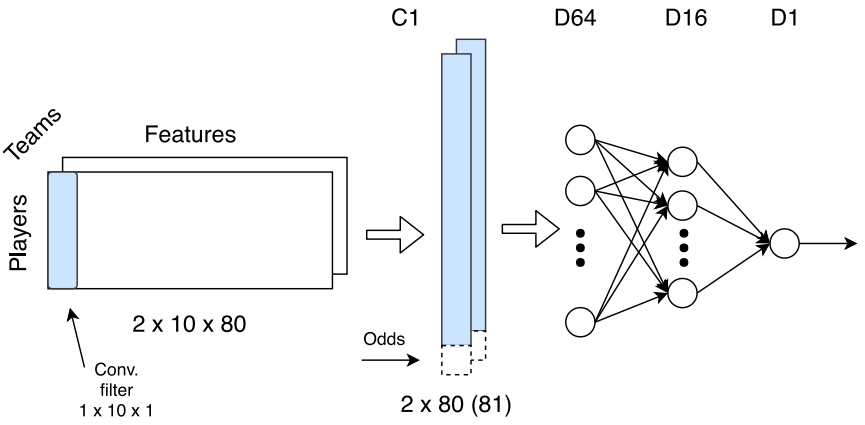

Fig. 1. The convolutional neural network for player-level data. The input to the network are two matrices (one for the home team, one for the visitors), with players in rows and all player-level features in columns. The rows are sorted by the time-in-play of the corresponding players, and only the top 10 players w.r.t. this factor are included. The convolution layer is defined by a vector of 10 tunable real weights. The output of the layer is a vector where each component is the dot product of the former vector with one of the matrix columns. The vector may be viewed as a filter sliding horizontally on the first input matrix, and then on the second.

4.4. Model decorrelation

While we deal with the construction of a bettor's predictive model, the bookmaker obviously also possesses a (tacit) model according to which the odds are set. If the probability of outcome $i = 2k - 1$ for some match k is $\bar{p}_i$ according to the bookmaker's model, the odds o_i is set to a value no greater than $1/\bar{p}_i$. So if outcome $i + 1$ is complementary to outcome i (i.e., they correspond respectively to the home and visiting team winning in the given match k), then

$$\frac{1}{o_i} + \frac{1}{o_{i+1}} = 1 + \epsilon$$

where $\epsilon \geq 0$. If $\epsilon = 0$, the odds would be called *fair*. In real life, $\epsilon > 0$ for all pairs of complementary outcomes and ϵ is called the bookmaker's *margin*. It is one of the sources of bookmaker's profit (see Figs. 3 and 4 for an analysis of odds and margin distributions in real data).

Consider a situation where the bettor's learned model coincides with the bookmakers model. Then any betting opportunity promising from the viewpoint of an estimated high outcome probability $\hat{p}_i = \bar{p}_i$ is made unfavorable by the odds set lower than $1/\bar{p}_i$. Therefore, even a highly accurate predictive model is useless as long as it coincides with the bookmaker's model. This motivates the objective to learn a predictive model under two criteria of quality: *high accuracy* on empirical data, and adequately *low correlation* with the bookmaker's model.

Recall the question from 4.1 whether to include the known bookmaker's odds as a feature when training a predictive model. Unless the bookmaker's odds are systematically biased, which they are not (Hubáček, 2017), a model considering odds as the *only* input feature would have no choice but to coincide perfectly with the bookmaker, inevitably doomed to end up with negative returns directly proportional to the *margin*. Although the odds are clearly highly accurate, given the low-correlation desideratum we just discussed, it seems reasonable considering not to provide the learning algorithm with this information. Since the bookmaker's odds represent a strongly predictive feature, the learning algorithm would likely involve it in the constructed model, which entails the undesired high correlation with the bookmaker's model. Consequently, for the models incorporating the odds as a feature, or otherwise correlated models, we propose two additional techniques to reduce the undesired correlation.

First is a simple technique in which we establish *weights* of learning samples, acting as factors in the computation of training errors. Alternatively, this can be viewed as making several copies of each example in the training multi-set so that the number of occurrences of each example in the multi-set is proportional to its weight. We explored two particular weightings. In one, we set each example's weight to the value of the corresponding odds. Thus training examples corresponding to high-odds outcomes and high potential pay-offs will contribute more to the training error. Hence the model is forced to be especially accurate on such examples, if at the price of lesser accuracy on other examples. In another variation, we set the weights to the odds only for instances where the underdog (team with odds higher than the opponent's) won, retaining unit weights for other examples. The intended outcome is that such a learned model will tend to spot mainly situations where the bookmakers underestimates the underdog's true win-probability.

Second is a more sophisticated decorrelation technique which, rather than weighting examples, directly alters the loss function which is minimized through gradient descent while fitting the neural model. The standard loss, which is just the sum of squared prediction errors, is extended towards

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2 - c \cdot (\hat{p}_i - 1/o_i)^2$$

where $\hat{p}_i$ is the model's output for the ith example, $y_i \in \{0, 1\}$ is the actual outcome of the match, and o_i are the odds set by the bookmaker, so $1/o_i$ provides a simple estimate of $\bar{p}_i$. The first term is conventional, forcing the model

Table 2
Average profits (in %) P_{opt} of the opt and P_{unif} of the unif strategies in dependence to the correlations of the (estimated) probabilities. Accuracy denotes the % of correct outcome predictions by the bettor (predict win if $\hat{p} > 1/2$). The four last columns break down the proportions (in %) of different combinations of predictions by $\hat{p}$ (bettor) and $\bar{p}$ (bookmaker): *Consensus* (both right), *Upset* (both wrong), *Missed* (bettor wrong, bookmaker right), *Spotted* (bettor right, bookmaker wrong).

$\rho(\hat{p}, p)$	$\rho(\hat{p}, \bar{p})$	$\varnothing P_{\text{opt}}$	$\varnothing P_{\text{unif}}$	Accuracy	Consensus	Upset	Missed	Spotted
0.85	0.85	11.15	3.14	70.11	61.99	20.37	9.53	8.12
	0.90	6.14	0.52	70.05	63.60	22.04	7.91	6.45
	0.95	−1.73	−5.46	70.08	65.74	24.12	5.80	4.34
0.90	0.85	18.14	8.56	71.48	62.66	19.70	8.82	8.83
	0.90	14.05	5.74	71.48	64.36	21.35	7.17	7.12
	0.95	9.60	3.38	71.45	66.39	23.50	5.05	5.06
0.95	0.85	25.30	13.42	72.91	63.34	18.98	8.11	9.57
	0.90	22.95	12.69	72.93	65.02	20.62	6.46	7.91
	0.95	20.79	11.87	72.92	67.21	22.74	4.33	5.71

to agree with the ground truth, while the second term enforces decorrelation w.r.t. the bookmaker. The constant c determines the relative significance of the decorrelation term.

5. Betting strategy

Equipped with a prediction model, we have estimates $\hat{p}_i$ of the true outcome probabilities p_i allowing us to estimate the expected profit (3) as $\widehat{E}(P) = (\hat{p}_i o_i - 1) b_i$, and the cumulative profit (5) as $\widehat{E}(P) = \sum_{i=1}^{2n} \widehat{E}(P_i)$. A straightforward betting strategy would be to place such bets $b_1, b_2, \ldots, b_{2n}$ which maximize the latter. This strategy, which we will refer to as max-ep, will obviously stipulate to put the entire budget (1) on the apparently best opportunity, i.e. $b_i = 1$ for $i = \arg\max_i \widehat{E}(P_i)$.

However, in repeated betting over several rounds, the bettor will likely prefer to spread the bets over multiple opportunities in one round to reduce the risk of losing the entire budget and being unable to continue. In other words, besides maximizing the profit's *expectation*, we also want to minimize its *variance*. In the literature reviewed in Section 2, we detected four strategies for bet spreading. We formalize them briefly as functions producing numbers B_i, while the actual bets are then prescribed by

$$b_i = \frac{B_i \, \mathbf{1}(\hat{p}_i - 1/o_i)}{\sum_j^{2n} B_j \, \mathbf{1}(\hat{p}_j - 1/o_j)}$$

where $\mathbf{1}(.)$ stands for the indicator function evaluating to 1 for positive arguments, and to 0 otherwise. So positive bets are only placed on opportunities where $\hat{p}_i > 1/o_i$, which is equivalent to $\widehat{E}(P_i) > 0$, and the denominator normalizes the bets to comply with (1). The four strategies spread the bets on such opportunities:

1. uniformly, i.e., $B_i \equiv 1$
2. according to the estimated probability of winning, i.e., $B_i = \hat{p}_i$
3. according to the absolute difference between win probabilities predicted by the model and that of the bookmaker, i.e., $B_i = \hat{p}_i - 1/o_i$
4. as above, but using relative difference, i.e., $B_i = \left(\hat{p}_i - 1/o_i\right)/\hat{p}_i$

We will refer to these strategies as unif, conf, abs-disc, rel-disc, respectively.

5.1. Portfolio-optimization strategy

The four strategies above are heuristics. Here we propose to adopt a more theoretically justified strategy for bet distribution, using the concepts of *portfolio theory* proposed in the field of economics by Markowitz (1952). As motivated earlier, a strategy should weigh in both the expectation and the variance of the profit. The portfolio optimization strategy seeks the Pareto front of $\vec{b}$'s with respect to E[P] and Var[P], i.e., the set of all bet distributions not dominated in terms of both of the factors.

The expectation is given by (3) and (5). For the variance of profit on a single outcome i, we have

$$\text{Var}[P_i] = \text{E}[P_i^2] - \text{E}[P_i]^2 = (1 - p_i) p_i b_i^2 o_i^2 \tag{11}$$

and for the variance of the cumulative profit

$$\text{Var}[P] = \text{Var}\left[\sum_{i=1}^{2n} P_i\right] = \text{Var}\left[\sum_{\substack{i=1,\ldots,2n \\ b_i>0}} P_i\right]$$

where the second sum restricts the summands to those i for which a positive bet b_i was placed, since by (2), $P_i = 0$ when $b_i = 0$. We now make the assumption that the bettor never places positive bets on each of two complementary outcomes, i.e. on both the home team and the away team in a match. Under this assumption, the summands above may be safely considered pair-wise independent random variables, as no team plays in two different matches in the same round. In other words, no team influences both of outcomes i, j if $i \neq j$, $b_i > 0$, $b_j > 0$. Thus we may write

$$\text{Var}[P] = \sum_{i=1}^{2n} \text{Var}[P_i] = \sum_{i=1}^{2n} (1 - p_i) p_i b_i^2 o_i^2$$

E[P] and Var[P] are of course not known and we compute the Pareto front using the model-based estimates $\widehat{\text{E}}[P]$ and $\widehat{\text{Var}}[P]$ computed from $\hat{p}_i$'s instead of p_i's. To pick a particular bet distribution from the computed Pareto front, we rely on the *Sharpe ratio* introduced by Sharpe (1994) according to which, we pick the distribution maximizing

$$\frac{\widehat{\text{E}}[P] - R}{\hat{\sigma}_P} \tag{12}$$

where $\hat{\sigma}_P = \sqrt{\widehat{Var}[P]}$ is P's (estimated) standard deviation and R is the profit from a risk-free investment of the disposable wealth, such as through banking interests. We neglect

(i.e., set $R = 0$) this economically motivated quantity due to the short duration of the betting procedure. We use the algorithm of sequential quadratic programming (Nocedal & Wright, 2006) to identify the unique maximizer of the Sharpe ratio. The strategy just described will be referred to as opt.

5.2. Confidence thresholding

We also explored a modification applicable to each of the betting strategies, in which only high-confidence predictions are considered. More precisely, a probability estimate $\hat{p}_i$ is passed to the betting strategy if and only if

$$|\hat{p}_i - 0.5| > \phi.$$

The reasoning behind this thresholding is that we want to remove the games where the model is very indifferent about the favorite. Although being in principle valid for the strategy, our assumption is that probabilistic predictions around 0.5 are typically more imprecise than predictions of higher confidence. This is especially true for the proposed models trained with gradient descent techniques over logistic sigmoid output which is indeed most sensitive at that point.

6. Experiments on simulated data

Here we conduct two auxiliary experiments requiring simulated data, to examine the effects of correlation between $\hat{p}_i$ and $\bar{p}_i$, as discussed in Section 4.4, and to illustrate the Pareto analysis of bet distributions introduced in Section 5.1.

6.1. Decorrelation effects

Our motivation in Section 4.4 to keep the model-based estimates $\hat{p}$ of p as little correlated with $\bar{p}$ (the bookmaker's estimates) as possible stems from the hypothesis that betting profits decay if such correlation increases with other factors kept constant. To test this proposition, we simulated the ground truth p as well as both of the estimates with various levels of their correlation, and measured the profits made by the opt and unif strategies for these different levels.

More precisely, we sampled triples $p, \hat{p}, \bar{p}$ from a multivariate beta distribution. The distribution is parameterized with the marginal means and variances of the three variables and their pair-wise correlations. The mean of each of the three variables was set to 1/2, reflecting the mean probability of the binary outcome. The variance of $\bar{p}$ was determined as 0.054 from real bookmaker's data (Section 7.1), and $\hat{p}$'s variance copies this value. The variance of p was set to 0.08 reflecting the *glass ceiling* thesis.[2]

We let the correlations $\rho(\hat{p}, p)$ and $\rho(\hat{p}, \bar{p})$ and $\rho(p, \bar{p})$ range over the values $\{0.85, 0.90, 0.95\}$. The former two represent the independent variables of the analysis, acting as factors in Table 2, while the presented numbers average over the 3 values of $\rho(p, \bar{p})$.

For each setting of $\rho(\hat{p}, p)$ and $\rho(\hat{p}, \bar{p})$, we drew $p_i, \hat{p}_i, \bar{p}_i$ $(i = 1, 2, \ldots, n = 30)$ samples, to simulate one round of betting. Then we set the odds $o_i = 1/\bar{p}_i$ (the bookmaker's margin being immaterial here) for $1 \leq i \leq n$, and determined the bets $b_1, b_2, \ldots, b_n$ from $o_1, o_2, \ldots, o_n$ and $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n$ using the opt and unif strategies (Section 5.1). Finally, the match outcomes were established by a Bernoulli trial for each of $p_1, p_2, \ldots, p_n$. With these inputs, we calculated the cumulative profit $P = P_1 + P_2 + \cdots + P_n$ of one round. This procedure was repeated 10 000 times (rounds), averaging the P_{opt} and P_{unif}.[3]

Table 2 shows the average profits as well as the accuracy of the bettor's outcome predictions (call win if $\hat{p} > 1/2$), and the percentual breakdown of 4 possible combinations of bettor's and bookmaker's predictions. The accuracies, as well as the four latter proportions, are also averages over all bets in all simulated rounds.

Besides the unsurprising observation that bettor's prediction accuracy grows with $\rho(\hat{p}, p)$, the results show that profits indeed decay systematically as the bettor's and bookmaker's predictions become more correlated (increasing $\rho(\hat{p}, \bar{p})$ decreases profit). An instructive observation is that the proportion of *spotted* opportunities is in all cases higher when the bookmaker's and bettor's predictions are less correlated. Moreover, we can see that the betting strategy is another independent factor strongly influencing the profit, with the opt strategy being completely superior to the unif strategy. Clearly, our proposals for promoting decorrelation and portfolio-optimization betting are both supported by this synthetic data experiment.

Table 3
Bet distributions as dictated by 5 different strategies on 6 simulated betting opportunities.

i	$\hat{p}_i$	$\bar{p}_i$	unif	abs-disc	rel-disc	conf	opt
1	0.30	0.26	0.17	0.20	0.35	0.08	0.09
2	0.59	0.52	0.17	0.27	0.24	0.16	0.23
3	0.75	0.70	0.17	0.21	0.15	0.21	0.30
4	0.60	0.57	0.17	0.13	0.12	0.17	0.12
5	0.74	0.71	0.17	0.11	0.08	0.20	0.17
6	0.64	0.62	0.17	0.07	0.06	0.18	0.08

6.2. Pareto analysis

To provide a visual insight of the Pareto front and the placement of exemplary strategies with respect to it, we generated 6 hypothetical betting opportunities with associated bettor's estimate $\hat{p}_i$ and bookmaker's estimate $\bar{p}_i$ of the outcome probability, for $i = 1, 2, \ldots, 6$. The bet distributions assigned to the 6 cases by 5 different strategies are shown in Table 3. Fig. 2 shows the position of these distributions within the expectation-deviation diagram (also called the *risk-return* space) with respect to the Pareto front, along with other 1000 random bet distributions $\vec{b}$.

[2] Articulated by Zimmermann et al. (2013) and expressing that sports games are predictable with a maximum of 75% accuracy at best. When p is sampled with mean 1/2 and variance 0.08, then with 0.75 probability the event $(p > 1/2)$ predicts correctly the outcome of a Bernoulli trial parameterized with p.

[3] Note that this is not the same (for P_{opt}) as setting $n = 30 \cdot 10000$ without repeating the procedure, as the full unit budget is supposed to be spent in each round.

For each such random $\vec{b}$, we sampled each b_i ($1 \leq i \leq 6$) from the uniform distribution on [0,1] and then normalized them so that $b_1 + b_2 + \cdots + b_6 = 1$. As expected, the opt strategy maximizing the Sharpe ratio (12) indeed lies on the Pareto front.

7. Empirical evaluation

7.1. Data

We retrieved the official box score data from the National Basketball Association (NBA) from seasons 2000 to 2014. The gathered data provide game summaries; namely, player-level and team-level statistics such as the number of shots or number of steals per game are recorded. The detailed description of the available kinds of information can be found in the appendix. Games with incomplete statistics were removed, and thus the number of games differs slightly between seasons; on average, 985 games per year were included. 10 initial games of each team in each season were not included as training instances as they only served for the initial calculation of seasonal aggregates (c.f. Section 4.1). There are 30 teams in the NBA, so one league round consists of $n = 15$ games.

For betting odds, we used the *Pinnacle*[4] closing odds for seasons 2010–2014.[5] For earlier seasons, we had to collect odds data from multiple different bookmakers. Fig. 3 shows histograms of odds distribution for the home and away teams and their winnings, respectively. The histograms reveal that in most matches the home team is the favorite in bookmaker's eyes. This comes as no surprise due to the home court advantage (home teams win in about 60% of games). Both histograms exhibit long-tailed distributions, as expected given that odds correspond to inverse probabilities, which roughly follow the true proportions of the respective winnings.

Fig. 4 shows the seasonal averages of the bookmaker's margin ϵ, displaying the artifact caused by different sources of odds information prior and post 2010. This artifact does not confound the experimental questions below, except for causing higher profits in late seasons due to the systematically smaller margins. To get a better insight into the bookmaker's margins, we plotted their dependency on odds for the 2010–2014 period. Fig. 4 indicates a significantly larger margin in the case where there is a clear favorite with high probability of winning (odds close to 1). This is due to an asymmetry in bookmaker's odds: while there are several occasions with the favorite's odds around 1.1 implying win-probability around 91%, odds around 11 representing the complementary probability 9% are extremely rare. This asymmetry is increasing with favorite's odds approaching 1.0.

7.2. Experimental protocol

The central experimental questions are: how accurate the learned predictors of match outcomes are, how profitable the betting strategies using the predictions are, and how the profitability is related to the correlation between the bookmaker's and bettor's models.

Training and evaluation of the models and betting strategies followed the natural chronological order of the data w.r.t individual seasons, i.e. only past seasons were ever used for training a model evaluated on the upcoming season. To ensure sufficient training data, the first season to be evaluated was 2006, with a training window made up of seasons 2000–2005, iteratively expanding all the way to evaluation on 2014, trained on the whole preceding range of 2000–2013.

7.3. Results

The number of games with complete statistics available varies slightly with each individual season providing around 1000–1050 matches. The total number of 9093 games from the evaluated seasons 2006–2014 is counted towards the accuracies (% of correctly predicted outcomes) of each model, whose results are displayed in Table 4. The accuracy of the bookmakers' model, predicting the team with smaller odds to win, levels over these seasons at 69 ± 2.5. Generally in terms of accuracy, the bookmakers' model is slightly superior to the neural models, which in turn beat the logistic regression baseline (accuracy of 68.7 with odds, and 67.26 without). Overall the accuracy of all the models is considerably similar, including progress over the individual seasons.

As expected, we can observe from the results that models utilizing the highly informative odds feature achieve consistently higher accuracies. Similarly, the models that included the bookmakers' odds were anticipated to be more correlated with the bookmaker (Section 4.4). This is convincingly confirmed by measurements of Pearson coefficients which stay at 0.87 for the models trained without odds as a feature, and 0.95 for models including them, applying equally to both the player-lever and team-level models.

Table 4 also provides important insights on the profit generation. We display two selected betting strategies (opt, unif) against a range of considered variants of predictive models. Similarly to the simulation experiment, the superiority of opt over unif is still evident. Apart from accuracy, we argued for decorrelation as an important factor for profit, which we here enforce by the means of the altered loss function while varying the trade-off C between accuracy and decorrelation (Section 4.4). We can clearly see that such a trade-off is effectively possible for a wide range of $0.4 \leq C \leq 0.8$ resulting into positive returns over all the models utilizing the opt strategy.

In Fig. 5 we display insight on how the trade-off constant C influences the distribution of the four betting outcome situations. As expected, increasing the decorrelation results in a desirable increase of *spotted* opportunities, i.e., cases where the model correctly predicted the underdog's victory. If this increase is too abrupt, however, it is outweighed by the parallel increase of *missed* opportunities where the bet on the underdog was wrong. This was the case with the alternative decorrelation technique of sample weighting where we were not successful in finding the optimal trade-off between the two antagonistic factors to generate positive profit.

[4] https://www.pinnacle.com/

[5] provided kindly by prof. Strumbelj, University of Ljubjana, Slovenia.

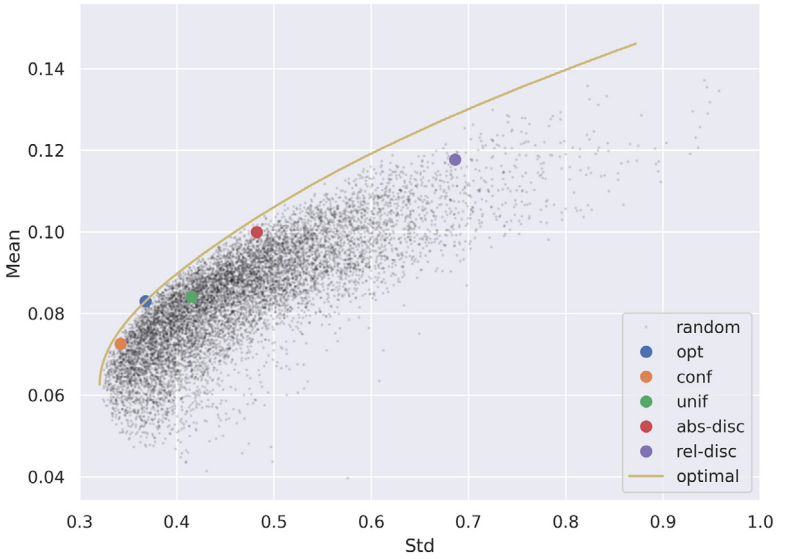

Fig. 2. Comparison of five betting strategies and 1000 random bet distributions in the risk-return space with respect to the Pareto front of optimal strategies.

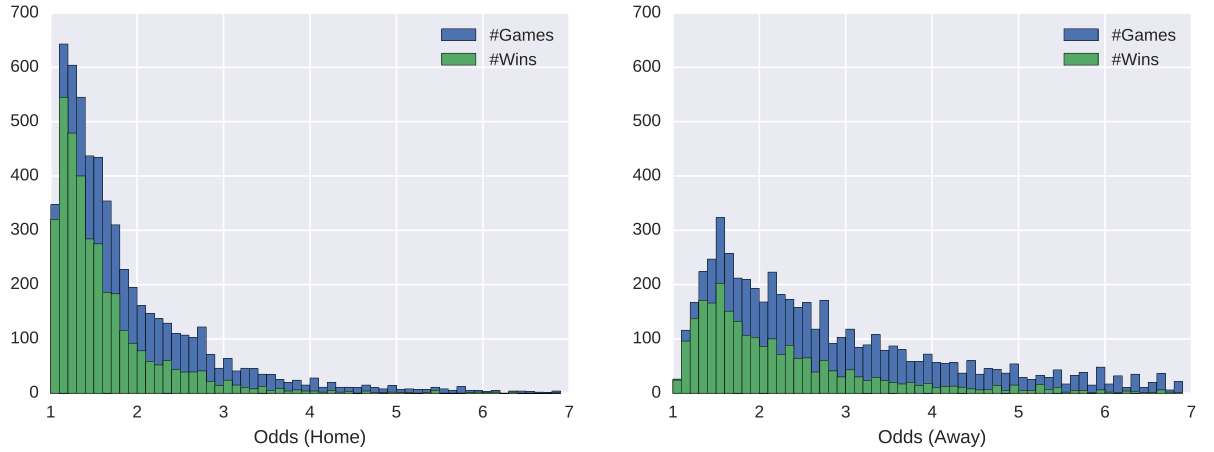

Fig. 3. Distribution of all games (blue) with respective proportions of wins (green) w.r.t odds set by the bookmaker from home (left) and away (right) team perspectives. Clearly, the home team is generally favored by the bookmaker, with the true proportions roughly following the inverse of odds. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Averages and standard errors of profits (from 10 runs over seasons 2006–2014) for the two strategies (opt, unif) with accuracies of *Player-level* and *Team-level* outcome prediction models (Section 4) across different levels of decorrelation (Section 4.4).

	C	*Team-level*			*Player-level*		
		$\varnothing P_{\text{opt}}$	$\varnothing P_{\text{unif}}$	*Accuracy*	$\varnothing P_{\text{opt}}$	$\varnothing P_{\text{unif}}$	*Accuracy*
Without odds	0.0	-0.94 ± 0.12	-4.31 ± 0.17	67.47 ± 0.05	0.38 ± 0.10	-5.12 ± 0.11	67.62 ± 0.03
	0.2	-0.58 ± 0.14	-3.60 ± 0.19	67.39 ± 0.04	1.05 ± 0.12	-3.31 ± 0.13	67.47 ± 0.03
	0.4	0.46 ± 0.15	-1.94 ± 0.20	67.30 ± 0.05	1.74 ± 0.14	-1.73 ± 0.18	67.15 ± 0.10
	0.6	0.86 ± 0.08	-1.68 ± 0.22	66.93 ± 0.06	1.32 ± 0.14	-0.61 ± 0.28	66.19 ± 0.09
	0.8	1.37 ± 0.08	-0.79 ± 0.16	65.94 ± 0.12	1.10 ± 0.29	-0.39 ± 0.22	64.93 ± 0.35
	1.0	-1.06 ± 0.35	-1.32 ± 0.31	61.38 ± 0.19	-1.92 ± 0.81	-2.59 ± 0.57	61.30 ± 0.48
With odds	0.0	0.89 ± 0.10	-2.24 ± 0.21	68.83 ± 0.05	-0.12 ± 0.24	-3.83 ± 0.22	68.80 ± 0.06
	0.2	0.92 ± 0.18	-2.10 ± 0.24	68.71 ± 0.04	0.72 ± 0.13	-2.50 ± 0.14	68.37 ± 0.04
	0.4	1.24 ± 0.12	-1.24 ± 0.22	68.42 ± 0.05	1.49 ± 0.10	-1.30 ± 0.12	67.48 ± 0.10
	0.6	1.44 ± 0.11	-0.64 ± 0.21	67.88 ± 0.06	1.02 ± 0.20	-1.15 ± 0.22	66.55 ± 0.10
	0.8	1.41 ± 0.10	-0.56 ± 0.20	66.64 ± 0.12	1.00 ± 0.35	-0.45 ± 0.28	65.19 ± 0.27
	1.0	-0.37 ± 0.16	-0.74 ± 0.13	62.49 ± 0.12	-1.22 ± 0.51	-2.25 ± 0.30	61.77 ± 0.44

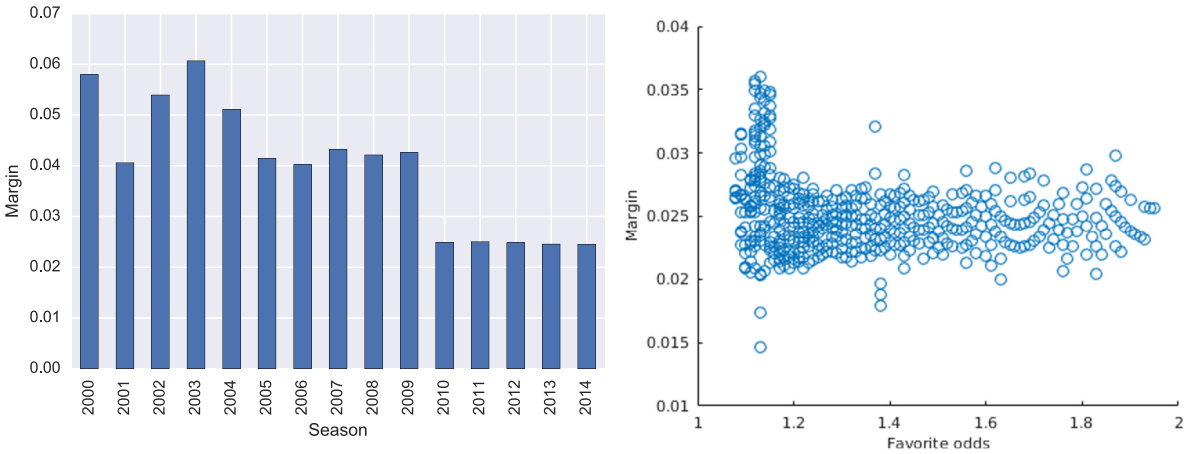

Fig. 4. Evolution of margin over the seasons (left), showing drop for seasons 2010–2014 where Pinnacle was the only source, and its dependency on bookmaker's odds for the favorite of each game (right), displaying interesting patterns with rapid growth towards the clear favorite case (minimal odds).

Revisiting the question as to whether include the odds feature or not, in terms of profit generation the results are inconclusive, with the team-level model performing slightly better with the feature and the player-level model without it.

In terms of profit measures, both the proposed models beat the baseline logistic regression, which, while operating on the same feature-set, behaves similarly to the team-level model yet yielding inferior performance (only 0.61 in the best setting with odds and opt).

Next we investigate the effects of confidence-thresholding used to filter the predictions (Section 5.2) before providing them to the betting strategy. By varying the threshold ϕ we can trade off between the confidence of the model and the number of games providing information to the strategy. Results in Table 5 are conclusive in that a reasonably low amount of thresholding bellow $\phi = 0.2$ in conjunction with the opt strategy indeed improves profits. Such a low threshold has the effect of filtering out generally those predictions that are indifferent on the winner (estimated probabilities of 0.5 ± 0.2), which was the main motivation for this technique.

7.3.1. Actual betting evaluation

We have demonstrated the effectiveness of the proposed models and techniques such as decorrelation and confidence thresholding. The ultimate question is, how a bettor could leverage these observations to exploit the betting market or, more precisely, which setting would yield the largest and most stable gains. In an actual scenario, the bettor could be continuously evaluating all the proposed models in different settings, and after each season (s)he could fix a selected setting for the upcoming season based on past performance. As a selection criterion, (s)he could once again utilize the Sharpe ratio for trading off between average historical profits per round and their standard deviation.

Following the proposed scenario, we evaluated each betting strategy under the max-Sharpe selection criterion applied over all the possible settings combining the choice of features, decorrelation, and thresholding. Fig. 6 shows the progress of their respective cumulative profits for all rounds between 2007–2014 (the first season of 2006 is used for the selection). Operating on the full range of possible settings, the resulting cumulative profits demonstrate the independent value added to profit performance separately by each of the betting strategies.

The proposed opt strategy clearly outperforms all other strategies, except the risk-ignoring max-ep. Although the expectation maximization strategy max-ep accumulates a larger ultimate profit than opt in this scenario, it suffers from great variance with abrupt drops, possibly bankrupting the bettor. On the contrary, the opt strategy maintains a steady growth of profit throughout the entire duration.

Overall, the opt strategy with the adaptive settings selection generated a positive profit of $\varnothing P = 1.63$. Interestingly, in a scenario where we deprived the strategies of decorrelation and thresholding settings, the opt achieved only $\varnothing P = 0.44$ and max-ep ended up in bankrupcy, further demonstrating usefulness of the proposed techniques.

8. Conclusions

The main hypotheses of this study were (1) that correlation of outcome predictions with the bookmaker's predictions is detrimental for the bettor, and that suppressing such correlation will result in models allowing for higher profits, (2) that convolutional neural networks are a suitable model to leverage player-level data for match outcome predictions, and (3) that a successful betting strategy should balance optimally between profit expectation and profit variance.

The first hypothesis was clearly confirmed in simulated experiments and also supported by extensive real-data experiments. In the former, for each level of constant accuracy (correlation of model and ground truth), increasing the correlation between the model and the bookmaker consistently decreased the profit in all settings. In the latter, models trained with the proposed decorrelation loss achieved higher profits despite having lower accuracies than models with higher correlation, in all settings up to a reasonable level of the decorrelation-accuracy trade-off.

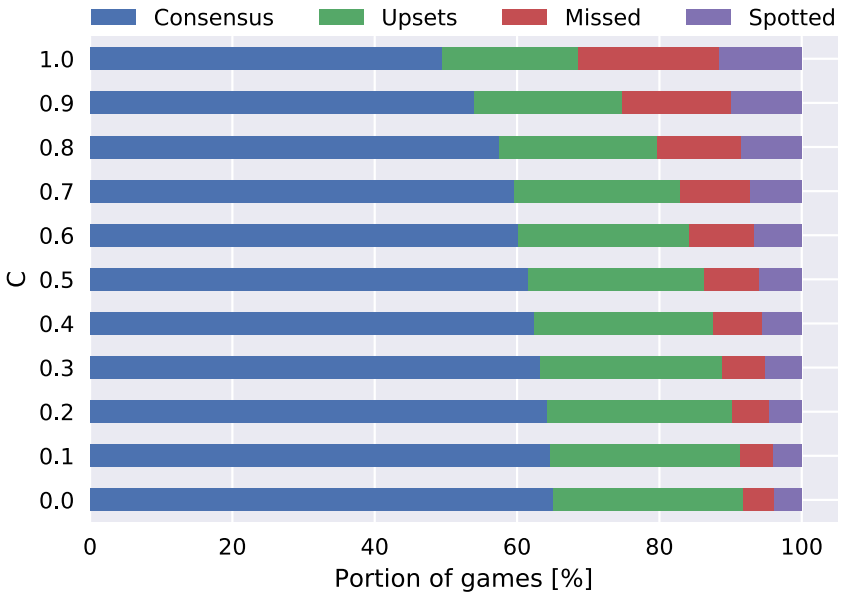

Fig. 5. The impact of loss-function-term model-decorrelation techniques, as introduced in Section 4.4 and applied on the team-level model, on the distribution of betting opportunity outcomes: *Consensus* (both right), *Upset* (both wrong), *Missed* (bettor wrong, bookmaker right), *Spotted* (bettor right, bookmaker wrong).

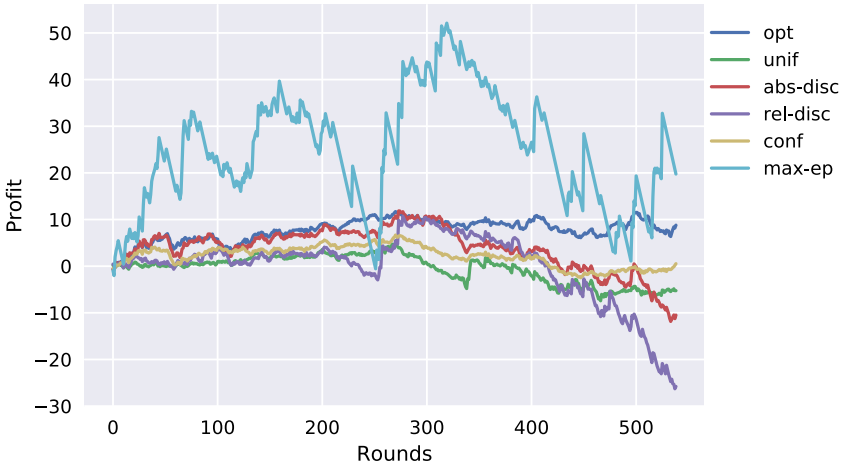

Fig. 6. Actual cumulative profits of 6 different betting strategies through seasons 2007–2014.

Table 5
Averages and standard errors of profits (from 10 runs over seasons 2006–2014) for the two strategies (opt, unif) with accuracies of the *Player-level* ($C = 0.4$) and *Team-level* ($C = 0.6$) prediction models (Section 4) across different levels of confidence thresholding (Section 5.2). *Games* represent numbers of games that passed the respective threshold.

	ϕ	*Team-level*				*Player-level*			
		$\varnothing P_{\text{opt}}$	$\varnothing P_{\text{unif}}$	*Accuracy*	Games	$\varnothing P_{\text{opt}}$	$\varnothing P_{\text{unif}}$	*Accuracy*	Games
Without odds	0.0	0.86 ± 0.08	-1.69 ± 0.22	66.93 ± 0.05	9093	1.74 ± 0.14	-1.73 ± 0.18	67.15 ± 0.1	9093
	0.1	1.61 ± 0.14	-1.37 ± 0.21	70.31 ± 0.07	7370	2.39 ± 0.2	-1.42 ± 0.16	72.01 ± 0.15	6686
	0.2	1.99 ± 0.25	-1.25 ± 0.21	74.08 ± 0.13	5442	3.24 ± 0.32	-1.18 ± 0.29	77.22 ± 0.19	4228
	0.3	0.51 ± 0.59	-2.56 ± 0.7	79.64 ± 0.2	2937	-4.81 ± 0.82	-6.9 ± 1.09	84.32 ± 0.48	1841
With odds	0.0	1.44 ± 0.11	-0.64 ± 0.21	67.88 ± 0.06	9093	1.49 ± 0.1	-1.3 ± 0.12	67.48 ± 0.1	9093
	0.1	2.18 ± 0.14	-0.13 ± 0.25	70.93 ± 0.06	7538	2.43 ± 0.16	-0.94 ± 0.24	72.2 ± 0.08	6749
	0.2	1.8 ± 0.24	-0.73 ± 0.29	74.47 ± 0.09	5749	3.39 ± 0.46	-0.7 ± 0.52	77.41 ± 0.12	4336
	0.3	0.75 ± 0.33	-1.61 ± 0.38	80.26 ± 0.21	3315	-4.57 ± 0.93	-7.36 ± 0.85	84.35 ± 0.29	1940

Regarding the second hypothesis, the convolutional network achieved generally higher accuracies and profits than the rest of the models in the settings excluding bookmaker's odds from features. This can evidently be ascribed to its ability of digesting the full matrix of players and their performance statistics through a flexible (learnable) pattern of aggregation, as opposed to just replicating the bookmakers estimate from input.

As for the third hypothesis, the portfolio-optimization opt strategy, which we designed as an essential contribution of this study, consistently dominated the standard unif strategy and, reassuringly, it was the only strategy securing a steady growth in profits with minimal abrupt losses in all actual betting simulations performed. Additionally, we proposed confidence-thresholding as an enhancement to the strategy when used in conjunction with models utilizing logistic sigmoid output. This technique effectively removes very uncertain predictions from the strategy, leading to additional increases in profit.

To our best knowledge, no work of similar scale evaluating sports prediction models from the viewpoint of profitability has yet been published.

8.1. Future work

There are avenues for future work stemming from each of the evaluated hypotheses. In the modelling part, we utilized neural networks operating on detailed feature sets of basketball games, however, this choice is completely independent of the remaining contributions, and we could equally employ different models on different sports and data representations. Particularly, we intend to explore models operating on pure game results data from various sports. In the betting strategy part, we assumed a scenario where a bettor is given a fixed budget to spend in each round. We plan to extend this to the more complex case where the bettor is continually reinvesting his wealth. Finally, following the decorrelation objective, we will aim to integrate the modelling and portfolio optimization parts in an end-to-end learning setting.

Acknowledgment

The authors are supported by Czech Science Foundation project 17-26999S Deep Relational Learning. FZ is also supported by OP VVV MEYS funded project CZ.02.1.01/0.0/0.0/16_019/0000765 Research Center for Informatics. Computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures". We thank the anonymous reviewers for their constructive comments.

Appendix

Below is the list of player and team performance data we used for constructing features. The grouping of variables and the acronyms shown match the source of the data http://stats.nba.com.

Basic statistics

- AST: Number of assists. An assist occurs when a player completes a pass to a teammate that directly leads to a field goal. *(per minute)*
- BLK: Number of blocks. A block occurs when an offensive player attempts a shot, and the defense player tips the ball, blocking their chance to score. *(per minute)*
- DREB: Number of rebounds a player or team has collected while they were on defense. *(per minute)*
- FG_PCT: Percentage of field goals that a player makes. The formula to determine field goal percentage is: Field Goals Made/Field Goals Attempted. *(per minute)*
- FG3_PCT: Percentage of 3 point field goals that a player or team has made. *(per minute)*
- FG3A: Number of 3 point field goals that a player or team has attempted. *(per minute)*
- FG3M: Number of 3 point field goals that a player or team has made. *(per minute)*
- FGA: Number of field goals that a player or team has attempted. This includes both 2 pointers and 3 pointers. *(per minute)*
- FGM: Number of field goals that a player or team has made. This includes both 2 pointers and 3 pointers. *(per minute)*
- FT_PCT: Percentage of free throws that a player or team has made.
- FTA : Number of free throws that a player or team has taken. *(per minute)*
- FTM: Number of free throws that a player or team has successfully made. *(per minute)*
- MIN: Number of minutes a player or team has played.
- OREB: Number of rebounds a player or team has collected while they were on offense. *(per minute)*
- PF: Number of fouls that a player or team has committed. *(per minute)*
- PLUS_MINUS: Point differential of the score for a player while on the court. For a team, it is how much they are winning or losing by. *(per minute)*
- PTS: Number of points a player or team has scored. A point is scored when a player makes a basket. *(per minute)*
- REB: Number of rebounds: a rebound occurs when a player recovers the ball after a missed shot. *(per minute)*
- STL: Number of steals: a steal occurs when a defensive player takes the ball from a player on offense, causing a turnover. *(per minute)*
- TO: Number of turnovers: a turnover occurs when the team on offense loses the ball to the defense. *(per minute)*

Advanced statistics

- AST_PCT: Assist Percentage - % of teammate's field goals that the player assisted.
- ST_RATIO: Assist Ratio - number of assists a player or team averages per 100 of their own possessions.
- AST_TOV: Number of assists a player has for every turnover that player commits.

- DEF_RATING: Number of points allowed per 100 possessions by a team. For a player, it is the number of points per 100 possessions that the team allows while that individual player is on the court.
- DREB_PCT: The percentage of defensive rebounds a player or team obtains while on the court.
- EFG_PCT: Effective Field Goal Percentage is a field goal percentage that is adjusted for made 3 pointers being 1.5 times more valuable than a 2 point shot.
- NET_RATING: Net Rating is the difference in a player or team's Offensive and Defensive Rating. The formula for this is: Offensive Rating-Defensive Rating.
- OFF_RATING: The number of points scored per 100 possessions by a team. For a player, it is the number of points per 100 possessions that the team scores while that individual player is on the court.
- OREB_PCT: The percentage of offensive rebounds a player or team obtains while on the court.
- PACE: The number of possessions per 48 min for a player or team.
- PIE: An estimate of a player's or team's contributions and impact on a game: the % of game events that the player or team achieved.
- REB_PCT: Percentage of total rebounds a player obtains while on the court.
- TM_TOV_PCT: Turnover Ratio: the number of turnovers a player or team averages per 100 of their own possessions.
- TS_PCT: A shooting percentage that is adjusted to include the value three pointers and free throws. The formula is: $\frac{\text{Points}}{2(\text{Field Goals Attempted}+0.44\text{Free Throws Attempted})}$
- USG_PCT: Percentage of a team's offensive possessions that a player uses while on the court.

Four factors, as described by Kubatko, Oliver, Pelton, and Rosenbaum (2007)

- EFG_PCT: Effective Field Goal Percentage is a field goal percentage that is adjusted for made 3 pointers being 1.5 times more valuable than a 2 point shot.
- FTA_RATE: The number of free throws a team shoots in comparison to the number of shots the team attempted. This is a team statistic, measured while the player is on the court. The formula is Free Throws Attempted/Field Goals Attempted. This statistic shows who is good at drawing fouls and getting to the line.
- OPP_EFG_PT: Opponent's Effective Field Goal Percentage is what the team's defense forces their opponent to shoot. Effective Field Goal Percentage is a field goal percentage that is adjusted for made 3 pointers being 1.5 times more valuable than a 2 point shot.
- OPP_FTA_RATE: The number of free throws an opposing player or team shoots in comparison to the number of shots that player or team shoots.
- OPP_OREB_PCT: The opponent's percentage of offensive rebounds a player or team obtains while on the court.
- OPP_TOV_PCT: Opponent's Turnover Ratio is the number of turnovers an opposing team averages per 100 of their own possessions.
- OREB_PCT: The percentage of offensive rebounds a player or team obtains while on the court.
- TM_TOV_PCT: Turnover Ratio is the number of turnovers a player or team averages per 100 of their own possessions.

Player scoring statistics

- PCT_AST_2PM: % of 2 point field goals made that are assisted by a teammate.
- PCT_AST_3PM: % of 3 point field goals made that are assisted by a teammate.
- PCT_AST_FGM: % of field goals made that are assisted by a teammate.
- PCT_FGA_2PT: % of field goals attempted by a player or team that are 2 pointers.
- PCT_FGA_3PT: % of field goals attempted by a player or team that are 3 pointers.
- PCT_PTS_2PT: % of points scored by a player or team that are 2 pointers.
- PCT_PTS_2PT_MR: % of points scored by a player or team that are 2 point mid-range jump shots. Mid-Range Jump Shots are generally jump shots that occur within the 3 point line, but not near the rim.
- PCT_PTS_3PT: % of points scored by a player or team that are 3 pointers.
- PCT_PTS_FB: % of points scored by a player or team that are scored while on a fast break.
- PCT_PTS_FT: % of points scored by a player or team that are free throws.
- PCT_PTS_OFF_TOV: % of points scored by a player or team that are scored after forcing an opponent's turnover.
- PCT_PTS_PAINT: % of points scored by a player or team that are scored in the paint.
- PCT_UAST_2PM: % of 2 point field goals that are not assisted by a teammate.
- PCT_UAST_3PM : % of 3 point field goals that are not assisted by a teammate.
- PCT_UAST_FGM: % of field goals that are not assisted by a teammate.

Usage statistics

- PCT_AST: % of team's assists a player contributed.
- PCT_BLK: % of team's blocked field goal attempts a player contributed.
- PCT_BLKA: % of team's blocked field goal attempts a player contributed.
- PCT_DREB: % of team's defensive rebounds a player contributed.
- PCT_FG3A: % of team's 3 point field goals attempted a player contributed.
- PCT_FG3M: % of team's 3 point field goals made a player contributed.
- PCT_FGA: % of team's field goals attempted a player contributed.
- PCT_FGM: % of team's field goals made a player contributed.
- PCT_FTA: % of team's free throws attempted a player contributed.
- PCT_FTM: % of team's free throws made a player contributed.

- PCT_OREB: % of team's offensive rebounds a player contributed.
- PCT_PF: % of team's personal fouls a player contributed.
- PCT_PFD: % of team's personal fouls drawn a player contributed.
- PCT_PTS: % of team's points a player contributed.
- PCT_REB: % of team's rebounds a player contributed.
- PCT_STL: % of team's steals a player contributed.
- PCT_TOV: % Percent of team's turnovers a player contributed.

Miscellaneous other statistics

- BLKA: Nnumber of field goal attempts by a player or team that was blocked by the opposing team. *(per minute)*
- OPP_PTS_2ND_CHANCE: Number of points an opposing team scores on a possession when the opposing team rebounds the ball on offense. *(per minute)*
- OPP_PTS_FB: Number of points scored by an opposing player or team while on a fast break. *(per minute)*
- OPP_PTS_OFF_TOV: Number of points scored by an opposing player or team following a turnover. *(per minute)*
- OPP_PTS_PAINT: Number of points scored by an opposing player or team in the paint.
- PFD: Number of fouls that a player or team has drawn on the other team. *(per minute)*
- PTF_FB: Number of points scored by a player or team while on a fast break. *(per minute)*
- PTS_2ND_CHANCE: Number points scored by a team on a possession that they rebound the ball on offense. *(per minute)*
- PTS_OFF_TOV: Number of points scored by a player or team following an opponent's turnover. *(per minute)*
- PTS_PAINT: Number of points scored by a player or team in the paint. *(per minute)*

References

Angelini, G., & De Angelis, L. (2018). Efficiency of online football betting markets. *International Journal of Forecasting*.

Boshnakov, G., Kharrat, T., & McHale, I. G. (2017). A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, *33*(2), 458–466.

Constantinou, A. C., Fenton, N. E., & Neil, M. (2013). Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using Bayesian networks. *Knowledge-Based Systems*, *50*, 60–86.

Forrest, D., Goddard, J., & Simmons, R. (2005). Odds-setters as forecasters: The case of English football. *International Journal of Forecasting*, *21*(3), 551–564.

Forrest, D., & Simmons, R. (2000). Forecasting sport: the behaviour and performance of football tipsters. *International Journal of Forecasting*, *16*(3), 317–331.

Franck, E., Verbeek, E., & Nüesch, S. (2010). Prediction accuracy of different market structures-bookmakers versus a betting exchange. *International Journal of Forecasting*, *26*(3), 448–459.

Haghighat, M., Rastegari, H., & Nourafza, N. (2013). A review of data mining techniques for result prediction in sports. *Advances in Computer Science: an International Journal*, *2*(5), 7–12.

Hubáček, O. (2017). *Exploiting betting market inefficiencies with machine learning* (Master's thesis), Czech Technical University in Prague.

Ivanković, Z., Racković, M., Markoski, B., Radosav, D., & Ivković, M. (2010). Analysis of basketball games using neural networks. In *Computational intelligence and informatics, 2010 11th international symposium on* (pp. 251–256). IEEE.

Kelly, J., Jr. (1956). A new interpretation of information rate. *Bell System Technical Journal*.

Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. T. (2007). A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, *3*(3).

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, *3361*(10), 1995.

Levitt, S. D. (2004). Why are gambling markets organised so differently from financial markets? *The Economic Journal*, *114*(495), 223–246.

Loeffelholz, B., Bednar, E., & Bauer, K. W. (2009). Predicting NBA games using neural networks. *Journal of Quantitative Analysis in Sports*, *5*(1).

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91.

Maymin, P. Z. (2017). Wage against the machine: A generalized deep-learning market test of dataset value. *International Journal of Forecasting*.

Miljković, D., Gajić, L., Kovačević, A., & Konjović, Z. (2010). The use of data mining for basketball matches outcomes prediction. In *Intelligent Systems and Informatics, 2010 8th International Symposium on* (pp. 309–312). IEEE.

Nocedal, J., & Wright, S. J. (2006). *Numerical Optimization*. Springer.

Paul, R. J., & Weinbach, A. P. (2007). Does sportsbook.com set pointspreads to maximize profits? *The Journal of Prediction Markets*, *1*(3), 209–218.

Paul, R. J., & Weinbach, A. P. (2008). Price setting in the NBA gambling market: Tests of the levitt model of sportsbook behavior. *International Journal of Sport Finance*, *3*(3), 137.

Paul, R. J., & Weinbach, A. P. (2010). The determinants of betting volume for sports in north america: Evidence of sports betting as consumption in the NBA and NHL. *International Journal of Sport Finance*, *5*(2), 128.

Puranmalka, K. (2013). *Modelling the NBA to make better predictions* (Master's thesis), Massachusetts Institute of Technology.

Sharpe, W. F. (1994). The sharpe ratio. *The Journal of Portfolio Management*, *21*(1), 49–58.

Sinha, S., Dyer, C., Gimpel, K., & Smith, N. A. (2013). Predicting the NFL using twitter. ArXiv preprint arXiv:1310.6998.

Song, C., Boulier, B. L., & Stekler, H. O. (2007). The comparative accuracy of judgmental and model forecasts of American football games. *International Journal of Forecasting*, *23*(3), 405–413.

Spann, M., & Skiera, B. (2009). Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, *28*(1), 55–72.

Stekler, H. O., Sendor, D., & Verlander, R. (2010). Issues in sports forecasting. *International Journal of Forecasting*, *26*(3), 606–621.

Vračar, P., Štrumbelj, E., & Kononenko, I. (2016). Modeling basketball play-by-play data. *Expert Systems with Applications*, *44*, 58–66.

Zimmermann, A., Moorthy, S., & Shi, Z. (2013). Predicting college basketball match outcomes using machine learning techniques: some results and lessons learned. ArXiv preprint arXiv:1310.3607.

Ondřej Hubáček is a Ph.D. student in Artificial Intelligence and Biocybernetics at Czech Technical University in Prague, where he graduated with a major in Artificial Intelligence. His interests lie mostly in predictive modeling in the domain of sports.

Gustav Šourek received his Masters degree in Artificial Intelligence in 2013 at Czech Technical University in Prague. Currently, he is pursuing Ph.D. in Artificial Intelligence and Biocybernetics at the same University. His research focus is mainly on Statistical Relational Machine Learning and affiliated problems of learning with logic, graphs and neural networks.

Filip Železný is professor of computer science at the Czech Technical University in Prague. Previously he was with the University of Wisconsin in Madison and the State University of New York in Binghamton. His main research interests are relational machine learning and inductive logic programming.