Insert here your thesis' task.

Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Computer Science

Master's thesis

Exploiting betting market inefficiencies with machine learning

Bc. Ondřej Hubáček

Supervisor: Ing. Gustav Šourek

7th January 2017

Acknowledgements

I would like to express my gratitude to my supervisor Gustav Šourek for his invaluable consultations and persisting optimism.

In addition, I thank Assistant Professor Erik Strumbelj from the University of Ljubljana, for provided data.

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"

Last but not least, I would like to thank all of my family for their continued support during my studies.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular that the Czech Technical University in Prague has the right to conclude a license agreement on the utilization of this thesis as school work under the provisions of Article 60(1) of the Act.

In Prague on 7th January 2017 .....

Czech Technical University in Prague
Faculty of Electrical Engineering

Citation of this thesis

Hubáček, Ondřej. *Exploiting betting market inefficiencies with machine learning.* Master's thesis. Czech Technical University in Prague, Faculty of Electrical Engineering, 2017.

Abstrakt

Cílem naší práce bylo najít způsob, jak profitovat na sázkařském trhu pomocí strojového učení. Zatímco vědecké práce se v minulosti hojně věnovaly otázce, zda lze vytvořit model, který bude přesnější než bookmaker, ziskovosti modelů nebyla věnována dostatečná pozornost. Tvrdíme, že zisk není důsledkem pouze samotné přesnosti. Místo toho navrhujeme dekorelaci od predikcí bookmakera, stále však nezapomínaje na přesnost predikce. Dále představujeme inovativní přístup k agregaci statistik hráčů pomocí konvolučních neuronových sítí. V neposlední řadě ukazujeme využití "Modern Portfolio Theory", matematického rámce pro optimalizaci portfolia, v kontextu sázení pro porovnávání různých sázecích strategií. Námi navržené modely byly v souhrnu 15 sezón NBA ziskové. Není nám známo, že by doposud byla zveřejněna práce podobného rozsahu.

Klíčová slova prediktivní modelování, sportovní analýza, sázkařské trhy, neuronové sítě, basketbal

Abstract

The goal of our work was to find a way to profit from betting market using machine learning. While a lot of research has been devoted to determining whether a bookmaker can be beaten in terms of accuracy, surprisingly little attention was paid to evaluating the profitability of statistical models over the bookmaker. We argue that the profit is not solely affected by the models' accuracy. Instead, we encourage decorrelation of the models' output from the bookmaker's predictions, while keeping the accuracy reasonably high. Moreover, we introduce a novel approach of aggregating player-level statistics using convolutional neural networks. Last but not least, we illustrate the use of Modern Portfolio Theory, a mathematical framework for portfolio optimization, in the context of betting for comparison of diverse betting strategies. Our proposed models were able to achieve a positive profit totaling over the course of 15 NBA seasons. To the best of our knowledge, no work of similar scale has been done and made publicly available before.

Keywords predictive modeling, sports analytics, betting markets, neural networks, basketball

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Statement

Advances in machine learning allow us to tackle a variety of problems in diverse domains. I this work we focus on predictive sports analytics in the context of betting markets. Our goal is to exploit betting market inefficiencies, i.e. to profit from the market. Betting markets thus serve for validation of our findings as well as a subject of our research inquiry by themselves.

Since there are many sports events offered at the betting markets, identifying suitable sub-domain to apply our ideas on is a vital part of the thesis, too. Our approach is generally data-driven, therefore relevant historical data have to be aggregated and processed.

To assess the theoretical profits of the model, authentic simulation of betting has to be implemented. The resulting evaluation criterion of the models' is their profitability. Comparison with state-of-the-art w.r.t. the accuracy of the prediction is also desirable.

1.2 Betting Markets

Sports betting is the act of placing a wager on a subset of outcomes of random sports events, each of which is associated with a corresponding profit as predefined by a bookmaker (Sec. 1.2.1). Such an assignment of bookmaker's estimate to the random event's outcome defines a *betting opportunity*. In the case of a correct identification of an outcome, the subject wins back the wager plus the profit or losing the wager otherwise. The structure of the outcomes and associated profits are based on various types of bets as described in Sec. 1.2.3.

1.2.1 The Bookmakers

The bookmaker is a company or a person accepting wagers on various events of inherent stochastic nature. Historically the bookmakers operated betting shops, but with the

expansion of the Internet, most bookmakers these days operate online through betting sites. The goal of the bookmakers is to maximize their profits.

1.2.2 Prediction Markets

An alternative to a traditional bookmaker are the prediction markets in sports betting, also known as betting exchanges. In a betting exchange, there is not a single authority defining the odds. Instead, each participant can use the exchange to allow other participants to bet against his odds. The betting exchange is in some way similar to a financial market. Although we find betting exchanges very interesting, their complicated structure makes it harder to model their behavior and to reliably evaluate our model in a hypothetical scenario against the exchange. Therefore we evaluate our models against the bookmaker where such problems simply do not arise.

1.2.3 Types of Bets

Bookmakers these days offer a variety of betting opportunities. We describe the most frequent types in this subsection.

1.2.3.1 Moneyline Bets

To win a so called moneyline bet, predicting the winner of the game is required. For each of the outcomes, the bookmaker sets the corresponding odds. There are different formats for representing the odds, too. In Europe, the most common one is the so-called *decimal odds*. For example, if bookmaker sets the odds to 1.8 for the home team to win, a bettor place a wager of 100 Eur and the home team actually wins, the bettor's profit will be $1.8 \times 100 - 100$ Eur.

1.2.3.2 Spread Betting

Bookmakers can assign spread, also called line, to handicap a team and favor its opponent. For example, if the line is *Home team -7.5 : Visiting team +7.5* and the bettor bets on the Home team, he wins when the home team beat its opponent by at least 8 points.

In this work, we decided to use the moneyline betting format as, among the European bookmakers, the moneyline bets are more common. Also, to reflect the spread betting format, we would have to switch from classification to regression, which would make the task slightly more complicated.

1.2.3.3 Total Bets

Besides betting on the game outcome, one of the most popular bets are so called *totals* or *under/over bets*. The goal of the bettor is to predict how many points (or goals etc.) will be scored in a game in total. For example, the most common under/over bet in

soccer is under/over 2.5. Bettor has an option to bet *under* or *over*. If he bets over, and more than 2 goals are scored, the bet is won.

Again, both moneyline bets and spread bets are feasible. In spread betting, the bettor's winnings are based on how many goals or points were scored above the threshold. In moneyline betting, fixed odds are assigned for each threshold.

Had we decided to approach the task of predicting the outcome of the game as regression, we could have instantly obtained the prediction for under/over bets too. However finding historical odds or lines for under/over turned out to be very problematic.

1.2.3.4 Proposition Bets

The betting opportunities are almost limitless these days, and proposition bets are proof of that fact. Not only bettors can bet whether a particular player will score a goal in a soccer match or how many passes will the team successfully finish in the game, but also how many corners will occur in first 30 minutes or who will win the coin flip at the start of the game. Soccer match serves only for illustration here; the proposition bets are not limited to a specific domain.

While some of the subjects of the proposition bets would be interesting to model, finding freely available historical odds for these events is not possible.

1.2.3.5 In-play Bets

Some online bookmakers offer the possibility to bet during the course of a game, while updating their odds according to the current situation inside the game. Although we find this mode interesting, beating the in-play odds would require real-time data gathering and evaluation, which would make the whole task much harder, and thus we focus solely on the scenario of placing the bets before the actual start of a game.

1.2.4 Bookmaker's Edge

In the gambling market, the operator usually has the advantage over the gamblers. For example in roulette, casino wins all bets in case the ball falls onto the number 0. While poker is being mostly played among the players and not against the casino, the casino takes a percentage of the placed bets in each hand - the so-called "rake". In sports betting, bookmaker secures his edge by offering unfair odds.

A line for spread betting usually looks like this: *Home team -7.5 : Visiting team +7.5 (-110)*. We have explained, what do the decimal numbers mean. The value -110 relates to bookmaker's edge. It implies, that to win 100 Eur, one has to bet 110. In other words, the value represents $\approx 9\%$ commission.

In moneyline betting, the bookmaker is slightly more devious. His edge is not that obvious sometimes. If the odds being laid out were fair, the inverse odds could be interpreted as the probability of the outcome estimated by the bookmaker. In practice for example, when the bookmaker is indifferent which outcome is more probable, he does not set the fair odds as $2.0 - 2.0$ but offers a lower portion of profit such as

1.95−1.95. The absolute difference between the sum of true probabilities (1 by definition) and probabilities implied by inverted odds is called the *margin*. In our example, the bookmaker's margin would be $1.95^{-1} + 1.95^{-1} - 1 \approx 2.5\%$.

The presence of the margin becomes less apparent when the bookmaker is more confident about the game outcome, and the odds of one team to win are close to 1. For example, when bookmaker estimates winning chances of the team as 90 % against 10 %, the fair odds would be 1.11, 10. If the bookmaker tweaks the odds at first sight marginally in his favor, for example, 1.07 - 10, the margin would suddenly jump to $\approx 3.5\%$

The size of the margin differs among bookmakers. The bookmaker has to optimize between the size of the margin and the number of customers. Should the odds on his site be consistently worse (lower) than on some other site, the bookmaker would quickly lose his customers. The market ensures that the margin is usually between 2 to 5 percent. Also, local laws can affect the market. Enforcing additional taxes forces bookmaker to increase the margin or implement a commission from winnings. Banning foreign betting sites takes options from bettors and allows local betting sites to get away with a higher margin than in the free market.

1.2.5 Arbitrage Betting

Arbitrage betting happens when a bettor makes use of bookmakers offering odds for the same event. In rare cases, two bookmakers offer odds steered towards opposite outcomes. For example one bookmaker (A) can set odds for one game 1.4 and 3.3 while the other bookmaker (B) lays down the odds as 1.3 and 4. If the bettor posses 100 Eur, he can make a risk-free profit by betting 74 Eur on the home team at bookmaker A with odds 1.4 and the remaining 26 Eur on the visiting team at bookmaker B with 4 odds. This split ensures him profit about 3.7 Eur regardless of the result of the game.

Arbitrage betting looks very tempting on the paper. However, Franck; Verbeek; Nüesch, 2009 showed, that 98 % arbitrage bets lead to the profit of 1.2 % or lower. Moreover, to make such profit, the bettor would have to have his finances spread across multiple betting sites to be able to catch the odds. These so called "sure bets" are immediately identified and exploited by the market. Once the bettor would make such a bet, he would be committed not only to winning the money at one of the betting sites, but also losing the whole bet at the other site he placed his bet. This would lead to a constant need of moving money around, assuring that they are available at each betting site for next "sure bet" opportunities.

1.3 National Basketball Association

In order to design and evaluate forecasting model we needed to obtain both the data to train our models and the bookmakers' odds. We found out that NBA offers the most comprehensive sets of statistics, thus we decided to test our ideas on NBA.

1.3.1 Structure

NBA is a men's professional basketball league in North America founded 1946 currently played by 30 teams.

The teams are separated into the Western and the Eastern conferences. Moreover, each conference is divided into three divisions – the Eastern into Pacific, Central, Southeast and the Western into Northwest, Southwest, Pacific. Each division contains 5 teams. 82 are played in regular season. Each team plays other teams from the same division 4 times, 6 teams from the other 2 divisions 4 times and remaining 6 teams from the same conference 3 times. Teams from the second conference are faced 2 times.

A roster of each team consists of 15 players (12 active and 3 inactive). For each game, the coach can change active and inactive players. Active players are then available for the game. 5 players from each of opposing teams are on the court simultaneously. The number of substitutions is not limited. In late October rosters are locked. The trade deadline is usually in February. Teams are not allowed to trade players, they may, however, sign or release players.

At the end of regular seasons 8 teams from each conference advance into playoffs. Playoffs are played in tournament format. Teams face their opponents in best of seven series – first one to win 4 games advances to next stage, loser is eliminated.

1.3.2 Data Available

There are several types of data recorded.

Box score data provide a summary of a game. Players' and teams' statistics (number of shots, number of steals, ...) per game are recorded.

Play-by-play data represents a time-line of a game. Each event that happens in the game (shot taken, steal, ...) is recorded with exact time.

Player tracking data are recorded using video tracking system. Position of the ball and all players is recorded in short time intervals. While these data allow complete reconstruction of the game, they are not freely available.

1.4 Predictive Modeling

The goal of predictive modeling is to build models capable of predicting outcomes of events, in our case the outcomes of sports matches, given some qualitative description of the events. Machine learning is about creating algorithms that can learn such predictive models while generalizing from training data, i.e. quantitative descriptions of the historical events.

There are various machine learning techniques available that might be used for the task at hand. Artificial neural networks (ANNs) were our models of choice. The nature of ANNs allows us to tune the complexity of the models easily. This makes the ANNs very scalable – if we were in need of adding additional features, the model structure

could be tweaked in an appropriate manner. The data we are dealing with have a relational character. The architecture of the neural network allowed us to aggregate the player-level statistics using a convolutional or a locally connected layers.

In this section, we briefly introduce ANNs. Much more details can be found for example in Goodfellow et al., 2016 and Orr et al., 2003.

Artificial neural network is a model inspired by the human brain. ANN consists of a set of highly interconnected units – neurons.

Each neuron accumulates signal from its (weighted) inputs to get the so called net input (Figure 1.1). This input is passed with a threshold into an activation function. The output of the activation function is then passed to other neurons or treated as the output of the network.

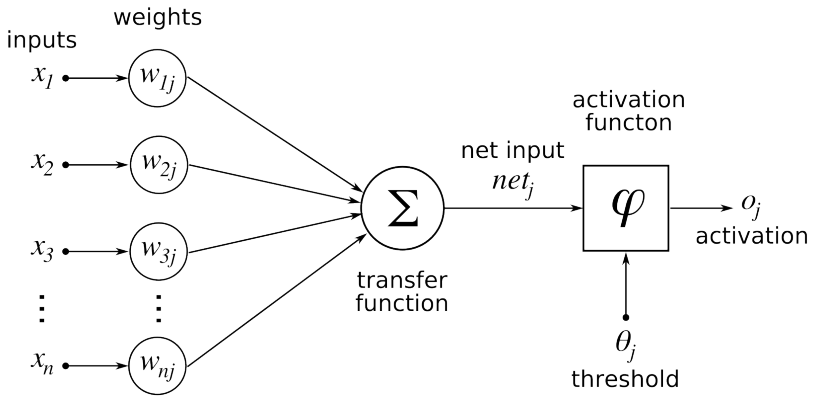

Figure 1.1: A schema of a neuron

The neurons are usually clustered into layers (Figure 1.2). The first layer of the network is called the input layer and the last one output layer. Between these layers, there can be arbitrarily many hidden layers.

Artificial neural networks experienced a renaissance in the past decade. One of the reason for their rise in popularity is the available computational power. Neurons from the same layer are independent units, therefore the computations can be run in parallel.

1.4.1 Training ANN

The process of training a predictive machine learning model is carried out as a minimization of a given error function, representing some discrepancy of the output of the network and the desired output. There are three basic ways how to approach training in this learning setting as follows.

Full batch learning updates the weights after all training examples were presented to a network.

Stochastic learning algorithm updates the weight after each training example.

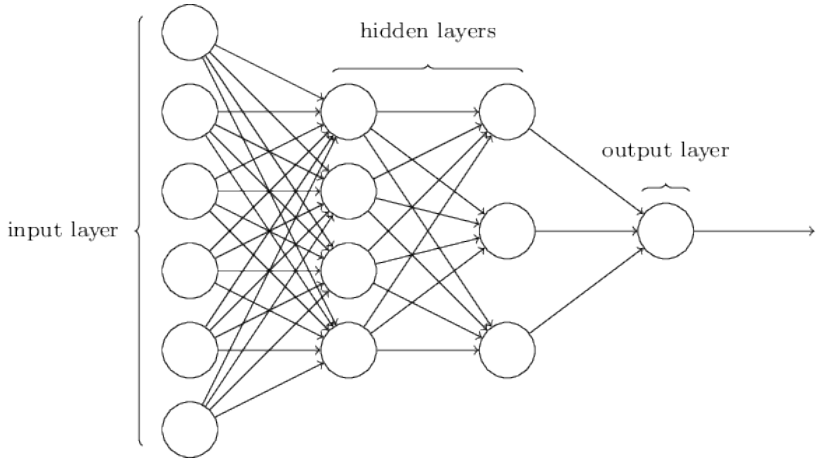

Figure 1.2: A schema of an ANN

Mini-batch learning is a compromise between previous approaches. Weights are updated after a small sample of examples is presented.

The advantages of stochastic and mini-batch learning over full batch learning are lower memory requirements and the built-in mechanism for escaping local minimum. For each example or subset of examples there are different local and global optimums, therefore ending up in bad local optimum is less likely.

1.4.2 Architecture

The architecture of the network is the most important meta-parameter. The very basic categorization of neural networks would be to feed-forward networks and recurrent networks. Since our dataset consisted of seasonal averages, we used only feed forward networks. Had we used a dataset with time series characteristics, for example sliding windows of last n games, we could have opted for recurrent neural networks (however as discussed in Sec. 3.2.2 there is no consensus if such data bear a useful information). As far as how deep the network should be and how many neurons should each layer carry, the current trend is to create large networks deep up to hundreds of layers and use regularization techniques to prevent over-fitting instead.

1.4.3 Types of Layers

The linkage pattern between the layers differs with layer type. In our work we worked with the following layers.

Fully-connected (dense) layer connects every input to each neuron.

Locally connected layer connects a different subset of inputs to each neuron.

Convolutional layer connects a different subset of inputs to each neuron. The neurons in the layer are sharing the weights.

1.4.4 Activation Functions

Activation function takes the neuron's net input and transforms it to the neuron's output. There are many possibilities when considering activation functions 1.3.

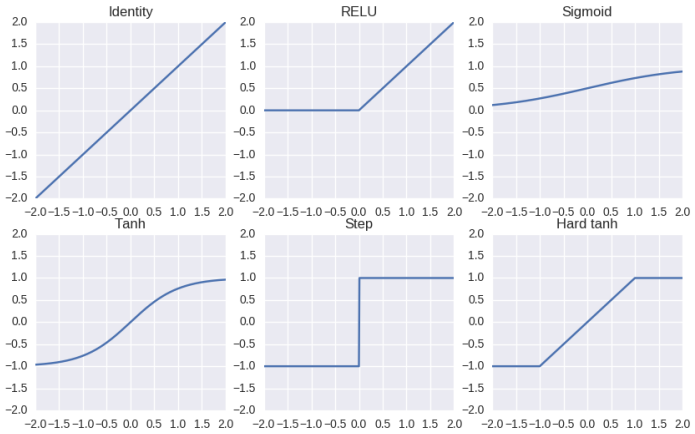

Figure 1.3: Examples of activation functions

Nowadays, the using RELU or its modifications is suggested (Goodfellow et al., 2016). In the past, sigmoidal functions were common, but their disadvantage is that they saturate quickly – when a net input is close to zero, even a small change in the net input leads to change in neuron's output, but there is very little difference when the values are far from zero.

1.4.5 Initialization

State of the network's weights before training starts can affect network's convergence. Probably the most common options for weight initialization are sampling either normal or uniform distribution. The main goal of the initialization is to make sure, that neurons with same inputs have different weights (Goodfellow et al., 2016).

1.4.6 Regularization

To prevent overfitting, regularization techniques can be used. Again, there are many techniques that are being used, we mention the methods used in this work, which also happen to be some of the most popular regularization techniques overall (Goodfellow et al., 2016), in this subsection.

1.4.6.1 Dropout

Dropout removes, with some given probability, a node together with its connections from the network. By removal of the neurons, sub-networks emerge. The sub-networks are trained in the same way as the whole network. This prevents co-adaptation of the neurons. Testing is done on the whole network with all neurons. Such network approximates averaging all the sub-networks as an ensemble model. This technique was introduced in Srivastava et al., 2014.

1.4.6.2 L1 and L2 Regularization

Another way how to regularize the network is to constrain the weights by adding a penalty to a loss function for weight far from the origin. The penalty is $\alpha||w||_1$ for L1 regularization and $\alpha\frac{1}{2}||w||_2^2$ when using L2 regularization.

1.4.7 Early Stopping

Early stopping can be looked upon as a regularization method. The point of early stopping is to separate part of training data into validation set and monitor (validation) loss during training. When the validation loss stops decreasing for some number of epochs, the training stops, and the model's weights are set to the state they were in when the validation loss was lowest.

Chapter 2

Literature Review

2.1 Betting Markets

In his work Levitt, 2004, the author explained the differences between financial and betting markets. To examine the markets' behavior, he used a dataset consisting of 20 000 wagers placed on NFL from 285 bettors. While the author had a line from only one bookmaker, he argued, that different bookmakers have similar lines probably due to outsourcing the odds from few odds setters.

Lots of authors interest had been devoted to determining how the odds are set. Two different scenarios in which bookmakers are making money were considered – one being bookmakers' capability of balancing weighted wages, second relying on bookmakers consistently outperforming gamblers in predictions.

The first hypothesis was quickly rejected by proving that in median game two-thirds of wages fall to one side. The author provided two feasible explanations. Either it is not possible for the bookmaker to balance the money wagered or it is not his objective. This rejection implies, that if there was a large enough subset of bettors outperforming the bookmaker, it would lead to the ruin of the bookmaker.

The author concluded that bookmakers are better in forecasting outcomes of games than an average bettor, which leads to higher profits than relying solely on balancing wages and winning thanks to the margin. While this strategy makes the bookmaker vulnerable against more skilled bettor than himself, the bookmaker protects himself by limiting the distortion of the odds.

The ability to hire experts that are better in forecasting game outcomes than the rest of the market was marked as the fundamental difference between betting markets and financial markets, where such solution is not possible, due to the complexity of the market and the amount of inside information.

Another interesting question answered was whether the aggregation of the bettors' beliefs carries any useful information. Data suggested that there is a correlation between the percentage of bettors choices and the game outcome such that the results chosen by the majority of bettors are more likely to happen. However, the author pointed out that

obtaining information about how the bettors acted before the game is problematic.

Rodney J Paul; Andrew P. Weinbach, 2007 raised their concerns about the generality of Levitt, 2004 findings. They pointed out that Levitt's dataset consisted of bets from betting tournament with entry fee and a limited number of participants, therefore not being representative. To test the Levitt's hypothesis, authors gathered percentage bets on opposing teams after each match day. Their findings were inline with the Levitt's hypothesis. Much more money was accepted on favorites, which led to the point spread not reflecting the market clearing price. They extended the reasons for why is not the bookmakers' strategy exploitable, stated in Levitt, 2004, with the point, that betting markets are in fact not free and betting sites can limit placed wagers or even refuse the wagers.

Same authors, such as Rodney J. Paul et al., 2008, decided to test the hypothesis on another sport – basketball. Contrary to observations from NFL, betting against public belief did not lead to profit. This lead to the idea, that bookmakers do not always try to exploit bettors biases, but rather focus on profiting on commissions in the long-run. This strategy wouldn't create chances for informed bettors. As the reason for following this policy authors mentioned smaller market for NBA betting.

Again, the same authors in Rodney J Paul; Andrew P Weinbach, 2010 explored the behavior of bettors on data from NBA and NHL season 2008/09. The key question asked was whether a typical bettor is a fan or an investor. The argument was that if the bettors are investors, it should be difficult to spot patterns affecting numbers of bets placed on different games. On the other hand, if the bettors are fans, they are expected to bet higher volumes on the most attractive games – games with television coverage, uncertain outcome, etc.

Data confirmed the hypothesis, that the bettors' behavior mimic fans' behavior to the point, that authors mention betting as a complement of watching sports events. Authors admit, that small group of bettors acting as investors might still exist; they are however the minority.

This finding follows our intuition and it is particularly interesting with the conclusion from Levitt, 2004 that the bookmaker focuses on beating the average bettor. Having confirmed that an average bettor is a fan, one might hope that with a statistical modeling approach, without any emotional attachments, it might be possible to exploit this fact towards a profit.

2.2 Human vs Machine

The idea that a statistical model might outperform experts was first tested in Simmons et al., 2000. The experts' predictions were obtained from three major magazines, and three different levels of experts' insights were tested.

- Are they outperforming a random choice?
- Do they efficiently process publicly available data?
- Do they make use of inside information?

The prediction accuracy of the experts was about 42 %. In the sample, 47 % of games were won by a home team. The authors thought of this comparison as unfair because the experts are obliged by the magazines to diversify their predictions. Moreover, when authors conducted regression analysis of the statistics in the respective magazines, the experts were found unable to process publicly available information efficiently. Only one of the experts showed signs of using information independent of publicly available data. Therefore the authors concluded that there is no reason why the experts should outperform the regression model.

Forrest et al., 2005 challenged the idea, introduced in Simmons et al., 2000, that a statistical model has an edge in comparison with tipsters. They examined the performance of a statistical model and bookmakers on 10 000 soccer matches. The authors concluded that under financial pressure, the bookmakers are on par with the statistical model. They also noticed, that bookmakers' accuracy improved through the 5-year period. Odds set by the bookmaker were suggested as a useful feature for the statistical model.

Song et al., 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 betting line. They even stated that betting lines are capable of predicting the margin of victory extremely well.

Authors stated several reasons why they think the experts performed as well as statistical models in their setup. First of all, the experts published their predictions in popular national magazines and were responsible for the magazines' creditability. Moreover, their predictions were posted close to the match day, so they could had accounted for all the information gathered or even make use of inside information since a lot of them were active players in the past and had connections to the coaches.

As a major advantage of statistical models over the experts, the authors pointed out lower variance between the models – worst models performed much better than worst experts.

Spann et al., 2009 compared prediction accuracy of prediction markets, betting odds and tipsters on three seasons of German premier soccer league. The authors tested the profitability of forecasting methods by betting the same amount of money on all the opportunities that are identified as profitable by the selected method.

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 it wasn't for the high fees. On the other hand, tipsters performed rather poorly in this comparison.

The fees make it harder to interpret the results since it is not clear if they serve as a tool for a bookmaker to increase his edge over the bettors, while offering more fair odds at first glance, or they are enforced by law as a form of tax.

Authors also examined a weighting-based and a rule-based combination of the forecasts looking for improvements in accuracy and profitability.

The weighting-based system relied on averaging the predictions made by prediction

market and bookmaker's odds. Neither the prediction accuracy nor the profits differed significantly from the standalone methods.

A rule-based system was based on a consensus of the forecasting methods. The prediction accuracy was highest when all the methods agreed. This situation, however, occurred only in about half of the games in the sample. On the contrary, the betting odds were in agreement with prediction markets in more than 90 % of the games which led to more betting opportunities and higher profits in total.

Franck; Verbeek; Nuesch, 2010, inspired by results of prediction markets in different domains such as politics, compared performance of betting exchange against bookmaker on 3 seasons of 5 European soccer leagues.

Prediction market was superior to the bookmaker in terms of prediction accuracy. As one of the reasons for this result, authors stated the bookmaker's intention to maximize the profit mentioned for example in Levitt, 2004. A simple strategy based on betting on the opportunities where the average odds set by the bookmakers were higher than odds in prediction market was profitable in some cases. Especially when predicting a win for the visiting team, the betting strategy was profitable against all bookmakers but one.

The authors also examined correlations between bookmakers' odds and prediction market. The data showed that the bookmaker and prediction market are highly correlated. Both the bookmakers and the prediction market underestimated the number of occurrences of home wins.

Kain et al., 2014 challenged the idea that betting markets are accurate predictors. For this purpose, they investigated betting markets performance not only on actual results of the games but also on under/over bets. Their dataset consisted of betting lines and under/over odds from NFL, NBA and NCAA seasons 2004-2010.

The authors found out that while betting line performs well when it comes to predicting the outcome of the game, the line for under/over fails to predict the number of points scored in the match.

2.3 Predictive Models

Besides the question, if a statistical model can outperform the human, lot of work have been dedicated to predictive modeling, not concerning comparison with the betting market. First of all, we mention two overviews, where a broader spectrum of research is described. We focus mainly on predictive models applied on basketball. However, interesting approaches tried in other domains are also mentioned.

2.3.1 Overview

Stekler et al., 2010 focused on several topics in horse racing and team sports. Forecasts were divided into three groups by their origin – market, models, experts.

Authors shared several findings that were backed by research in different sports. Closing odds proved to be better predictors of the game outcome than opening odds. Bettors belief in "hot hand" was pointed out – the bettors believe, that after a winning

game, the probability of the team to win the next game is higher, despite the fact, that there is no evidence of dependence.

As the most important result author stated that there was ho evidence that statistical model or expert could consistently outperform betting market.

Haghighat et al., 2013 provided a review of machine learning techniques used in outcome predictions of sports events. Authors more closely examined 9 papers from past decade. Common findings were that either presented results are rather poor or the used datasets are too small. For improving the prediction accuracy, it was suggested to include player-level statistics and machine learning techniques with good results in different fields.

2.3.2 Basketball

Loeffelholz et al., 2009 achieved a remarkably high accuracy of over 74% using neural networks, sadly it was on a dataset consisting of only 620 games. As features for their model, 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.

Ivankovic et al., 2010 used ANNs to predict outcomes of basketball games in the League of Serbia in seasons 2005/06–2009/10. The most interesting part of the work was using effects of shots from different court areas as features. With this approach, the authors achieved accuracy of 81 %. However, the very specific dataset makes it impossible to compare the results with other research.

Miljkovic et al., 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/division and in home/away games. Naive Bayes in 10-fold cross-validation achieved mean accuracy of 67 %.

Puranmalka, 2013 used play-by-play data (Section 1.3.2) do develop new features. The main reason why features derived from play-by-play data are superior to box score statistics is that they include a context. The features were divided into 4 groups as follows.

The first group contained features already mentioned in the literature. The author discussed a so called “clutch performance”. The idea behind measuring clutch performance is to put emphasis on plays made in close matches. Determining which game was close based on the final score was criticized. The underlying reasoning was that even uneven games can end with a relatively low difference in points scored and on the other hand really close games can end up looking uneven because of teams taking higher risks in the final quarter. A method based on point difference at selected times of last quarter was proposed. Also measuring shot selection and efficiency was analyzed. The author explained that traditional features as points scored/allowed per possession fail to explain why is the offense/defense more or less efficient. Instead, expected points per possession were suggested. Expected points per possession are based on different types of shots having different success rates and occurring with different frequencies.

The second set of features focused on the distribution of time on the court for players of each team. Hypothesis, that in the ideal scenario, only five players for each team have to play each game is stated. Deviation from this scenario is presented as one feature, however, the author acknowledged that deviation might only signalize, that team has a strong bench. Another feature was proposed based on the variance of players time on the court between games with the idea, that if the player spends on court similar amount of time each game, his role in team is well defined and the team is effective.

The third group of features tried to catch player-to-player dependencies. The author questioned predictive models for assuming that players' performances are independent of each other. Instead, he focused on pairs of players and how their offensive efficiency changes, when they are playing together. To measure defensive efficiency author examined how offensive efficiency of opposing team changes while the player is on the court.

The fourth set of features captured team-to-team interactions, like what happens when teams strong in certain ways (rebounds, fouls drawn, etc.) plays a team weak in that way. Several statistics were measured and teams ranked in each category. Relevant ranking pairs (for example strong team in the category playing weak team in the category) were marked as the new feature.

The last group of features was considered the most important one – context added metrics. The author found out, that there is a strong correlation between points per possession and time left on shot clock. The new feature was derived from time left on shot clock, attacking team's offensive efficiency and defending team's defensive efficiency.

A genetic algorithm for feature selection was compared with forward selection. Forward selection proved to be more effective. In the model using team-level features, clutch performance defined by author proved to be one of the best predictors of game outcome. Also, team-to-team features turned out to be very valuable. In player-level model, the context added metrics were the most useful. On the contrary, classic features like distance traveled or number of rest days between matches didn't seem to have any predicting power.

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.

Even though we decided not to make use of play-by-play data, we consider this work very relevant and novel. It could serve either as possible extension of our models or for comparison of models' performance.

Zimmermann et al., 2013 leveraged multi-layer perceptrons, that hadn't been usually used in this domain. They also raised their concern, that there might be 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 features are much more important than machine learning models since even Naive Bayes performed well. To increase the predictive power of their models, features encoding experience and leadership were used. An interesting idea mentioned in the paper is to train separate model for intra-conference matches.

Yang, 2015 analyzed the relation between player statistics and its team's performance in past 20 seasons of NBA. To regress player statistics to team strength, player efficiency rating was used. Team strength (team PER) was calculated as a weighted sum of players' PER. Point was made, that basketball is flooded with overcomplicated statistics while we can use basic stats to estimate the strength of a player or a team.

Vracar et al., 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 excluding 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).

2.3.3 Other

Basketball is by no means the only sport, where various forecasting models were tested. Here we briefly mention several interesting approaches tested in other sports.

Hvattum et al., 2010 implemented ELO rating (the standard way of determining chess players' strength) to express soccer teams' strength. ELO rating method performed poorly in comparison with methods based on betting odds.

Constantinou et al., 2013 designed an ensemble of Bayesian networks to asses 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 fatigue component provided the highest performance boost. Results revealed conflicts between accuracy and profit measures. The author emphasized the importance of expert knowledge when it comes to subjective inputs. The final model was able to outperform the bookmakers. The article demonstrates predictive power of Bayesian networks when well designed. However since we do not posses domain knowledge needed to properly design such a network, we decided to stay away from this model.

Sinha et al., 2013 made use of twitter posts to predict the outcomes of NFL games. Information from twitter posts enhanced forecasting accuracy, moreover, model based solely on features extracted from tweets outperformed models based on traditional statistics.

2.4 Features

During the game of basketball, a variety of statistics is being recorded. Getting an insight into these statistics could be crucial for feature selection or later model interpretability. We conducted research in order to close the gap in our game knowledge.

Kubatko et al., 2007 had set a starting point for analyzing basketball statistics. The concept of equal possessions for opposing teams was defined as key to basketball analysis. Formulas for estimating possessions were reviewed. Statistics from non-academic sources were brought together and explained.

Sergio J. Ibáñez et al., 2008 tried to identify statistics that differentiate the successful teams from the less successful ones. All statistics were normalized according to ball possessions as was suggested in Kubatko et al., 2007. The most discriminative statistics found were assists, steals, and blocks. The classification function was correct in 82.4 % retrospectively. Another result was that pace of the games is not an indicator of the team's performance.

Sergio J Ibáñez et al., 2009 focused on variations in game-related statistics making the difference in winning and losing in three consecutive games. Authors concluded that there was no significant variance in team performance. High fitness levels of players and the unlimited number of substitutions were stated as possible causes of this outcome.

Sampaio et al., 2010 studied effects of season period, team quality on players' statistics. Their dataset consisted of more than 5000 records from nearly 200 players in Spanish professional basketball league season 2007-08. Weaker teams showed larger discrepancies between their strongest and weakest players performances. Players that spend less time on the court tend to do more errors. Authors found no variation in players' performance during the season. This result is crucial for our work, because it implies, that once we obtain the team strength, we do not need to worry about its variation in time.

Strumbelj, 2014 addressed two very important questions. First one was whether it makes a difference which bookmaker we choose and a second one concerning deriving probabilities from betting odds. To answer these questions, the authors gathered a dataset consisting of almost 50000 games from numerous team sports. In literature, underlying probabilities are usually calculated by dividing inverse odds by their sum. This work, however, showed that using model form Shin, 1991 we can obtain more accurate predictions. Betting exchange *Betfair* and betting sites *bet365* and *bwin* proved to be most accurate on average. On the other hand bookmaker *Interwetten* was pointed out as the worst source of probability forecasts. Moreover, it performed particularly poor in basketball.

Chapter 3

Analysis and Design

In order to build a system that we could test against a betting market, several steps had to be done. Firstly, we had to gather and process as much relevant data as possible from a variety of sources and decide on how to build datasets from this raw data. After that came the statistical analysis and machine learning part. These, however, were not the final step of our work, since to evaluate the learned model against odds set by the bookmaker, a betting strategy also had to be determined. Each step of this pipeline is described in the following sections.

3.1 Gathering Data

Collecting data proved to be a challenging task. There are no widely recognized datasets. Plenty of data is available online, however, it is not for free.

Obtaining historical betting odds is an even bigger problem. The betting sites do not provide such data. There are websites specializing in collecting the odds from a variety of bookmakers. It is not usually stated whether the presented odds are opening or closing odds set by the bookmaker. Sometimes these websites show the best odds available among the bookmakers for each outcome of a game. To make practical use of such information, one would have to set an account on every betting site listed and deposit money to all these accounts.

In the end, we managed to gather official box score data 1.3.2 from NBA seasons 2000 to 2014. As for betting odds, we used *Pinnacle* closing odds for seasons 2010-2014, provided by Assistant Professor Strumbelj. For previous seasons we gathered odds from different bookmakers.

3.2 Building Datasets

After the data were stored into a relational database, they had to be preprocessed and aggregated into the datasets.

The basic premise for meaningful forecasting is to use only the data obtained before the forecast could be done. In our case, that means using the data available before the start of the game. Box scores are statistics gathered during the game so they cannot be used for forecasting by themselves. On the other hand betting odds are known days before the game actually starts so they could be included as features. We discussed two approaches how to extract features from the database.

3.2.1 Teams' and Players' Seasonal Averages

Seasonal averages are widely used in literature. They serve as a good predictor of team strength. However, beginnings of the seasons are problematic because the averages change rapidly. We experimented with excluding a different number of games from season beginning from training and testing our model. The downside of this approach is that the system cannot bet in these first rounds before seasonal averages settle.

3.2.2 Teams' and Players' Last n Games

Most of the research focused on in-game momentum (see Bar-Eli et al., 2006 for review), analyzing for example if series of successful shots increase the probability that the next shot taken will also be successful. However, some research has also been done on season scale – if a streak of won games increases the probability that the team will win the next one. Vergin, 2000, Goddard et al., 2004 and Sire et al., 2009 found no evidence for such momentum while more recent work Arkes et al., 2011 claims opposite. We decided not to use the last n games of opposing teams for dataset creation, although it might be a subject of our future work.

3.3 Exploratory Data Analysis of Bookmakers' Odds

Among all the features, one type is particularly interesting - the bookmaker's odds. Bookmaker's odds can be used as an input to our model, but at the same time, the model will be evaluated using ground truth labels together with these odds through a selected betting strategy. In this section, we analyze odds set by the bookmaker to discover possible biases which could create opportunities for exploitation. Unless specified otherwise, data we are analyzing are closing odds from "Pinnacle" on game outcomes from seasons 2010/11–2014/15.

3.3.1 Odds Distribution

It is clearly visible from the Figure 3.1 that in most matches the home team is the favorite in bookmaker's eyes. This is no surprise due to the home court advantage (home teams win in about 60 % of games). Both histograms are long-tailed distributions, which is expected given that odds represent inverse probability. Thus, even a small change in the probability of winning of one team can create a relatively large difference in outsider's odds. For example, when one team is the favorite with the probability of winning 75%

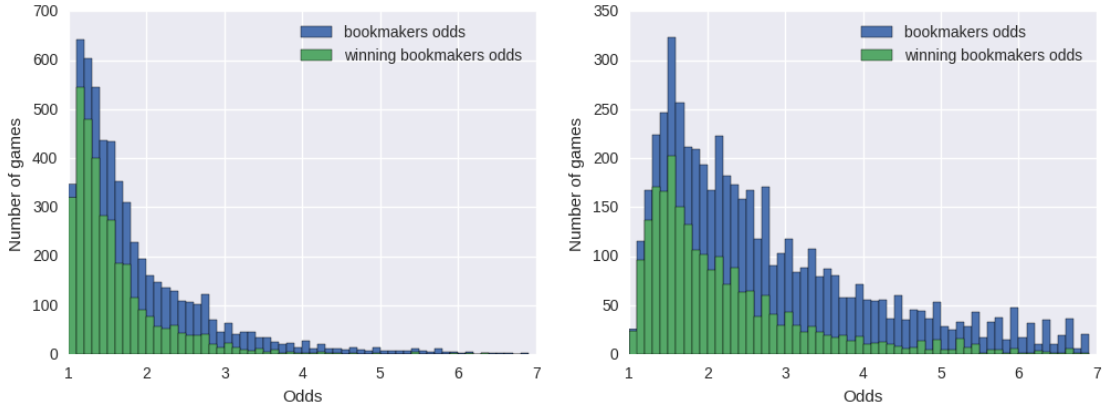

Figure 3.1: Distribution of odds set by the bookmaker for home team (left) and for visiting team (right).

and another team is also favorite, but with probability 76%, fair odds would be 1.33 and 1.32 respectively. Fair odds for their opponents would be 4 and 4.17.

3.3.2 Margin Analysis

Not having odds from the same bookmaker for all seasons is not optimal. Combining odds from various bookmakers, as we did for season's 2000-2009, can neglect the bookmaker's bias. Each bookmaker also has a different margin, therefore we computed this margin for every season.

Figure 3.2 shows, that margin in seasons 2010-2014 is more or less the same. On the other hand, in the previous seasons, margin varies a lot and is much higher. Just from this observation, we can expect higher returns from last five seasons, because the odds are simply more fair, therefore higher.

To better understand the margin, we plotted its dependency on odds. From the Figure 3.3 we can see, that the margin is larger when there is a clear favorite and its odds for the win are close to 1. This is due to asymmetry in bookmakers odds. While there are several occasion when the odds for favorite to win are around 1.1, implying probability 91%, odds above 11, representing probability 9% and lower are extremely rare. This asymmetry is increasing with favorite's odds approaching 1.

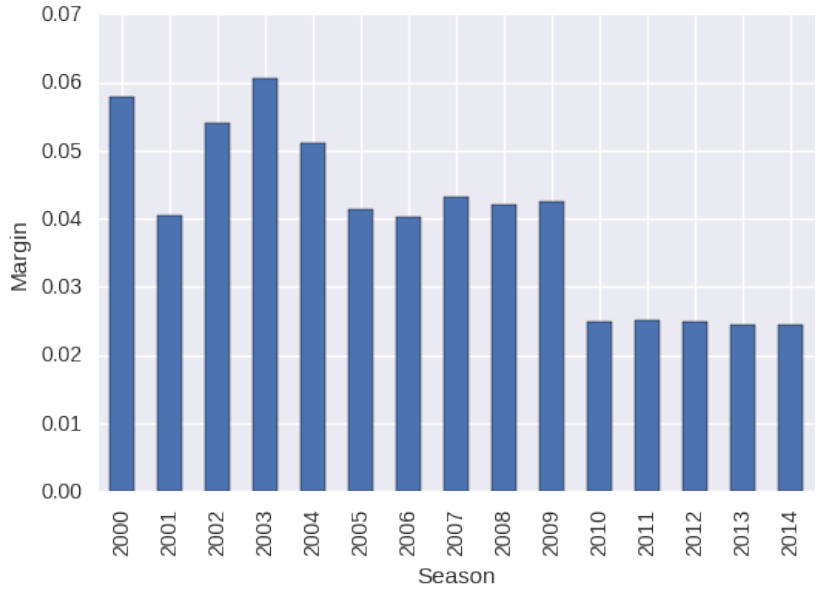

Figure 3.2: Margin distribution over the seasons

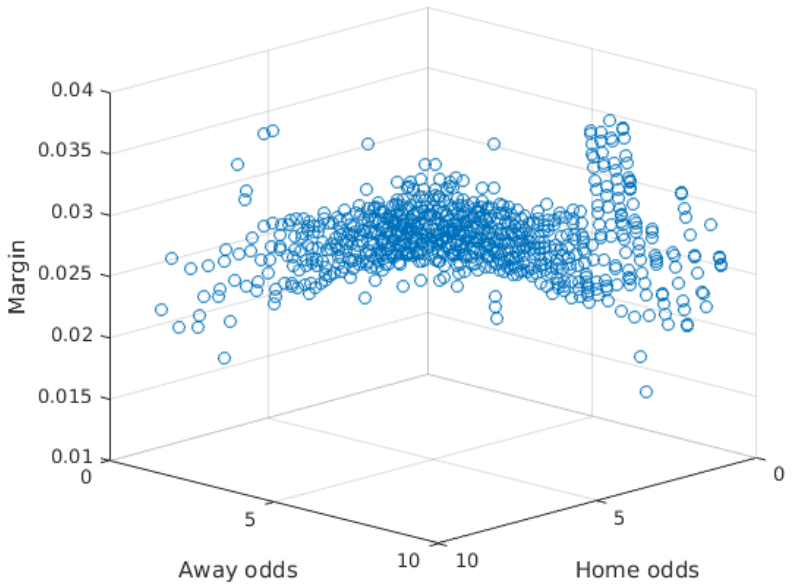

Figure 3.3: Margin distribution conditioned by the bookmaker's odds

3.4 Predictive Modeling

3.4.1 Estimates Based on Bookmaker's Odds

We used the gathered data to estimate the distribution of $P(RESULT|ODDS, COURT)$ with the assumption that this trivial model might reveal some systematic biases and provide useful information for our model.

We can estimate $P(RESULT|ODDS, COURT)$ directly from the histogram by calculating this probability in each bin. We can see that when the odds are low, the bookmaker is very precise on average. However in cases where odds are larger than 5 (probability of winning is less than 20%) the observed probability is oscillating around the expected value. One of the possible explanations of this phenomenon is a lack of data in the histogram for high odds.

When we estimated $P(RESULT|ODDS)$ (Figure 3.4), therefore neglecting if the odds were for home or visiting team, the distribution smoothed, but some spikes that could suggest bookmaker's bias preserved.

The expected probability, however, doesn't count in bookmaker's margin. If the bookmaker was offering fair odds and he was on average precise, he would copy the expected line. Since he provides unfair odds and the probabilities of game outcomes implied by inversed odds sums to more than 1, the bookmaker should be below the expected line, if his predictions were accurate.

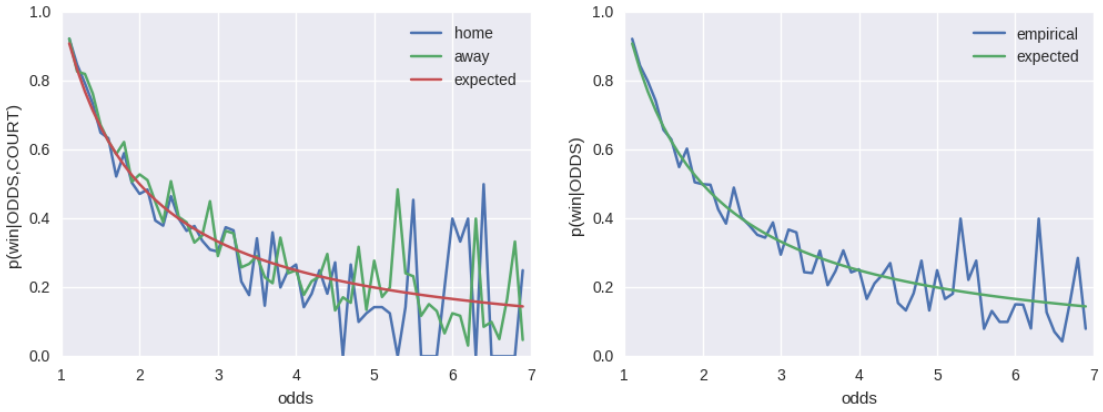

Figure 3.4: Estimates of $P(win, COURT|ODDS)$ (left) and $P(win|ODDS)$ (right).

3.4.1.1 Statistical approach

Another way to estimate $P(RESULT|ODDS, COURT)$ is to use Bayes' theorem

$$P(RESULT|ODDS, COURT) = \frac{P(ODDS, COURT|RESULT)P(RESULT)}{P(ODDS, COURT)}$$

To make a use of the formula, an assumption about $P(ODDS, COURT)$ and $P(ODDS, COURT|RESULT)$ distributions has to be made. The Beta distribution is the conjugate prior distribution for Bernoulli distribution, so it fits our case nicely. Each game can be looked upon as a Bernoulli trial. The team has a certain probability p of winning the game. The value of p is the only parameter needed for defining the Bernoulli distribution. This probability p also comes from a distribution – the distribution of probabilities. Beta distribution is defined in the interval $[0, 1]$, therefore it can be used as the distribution over probability.

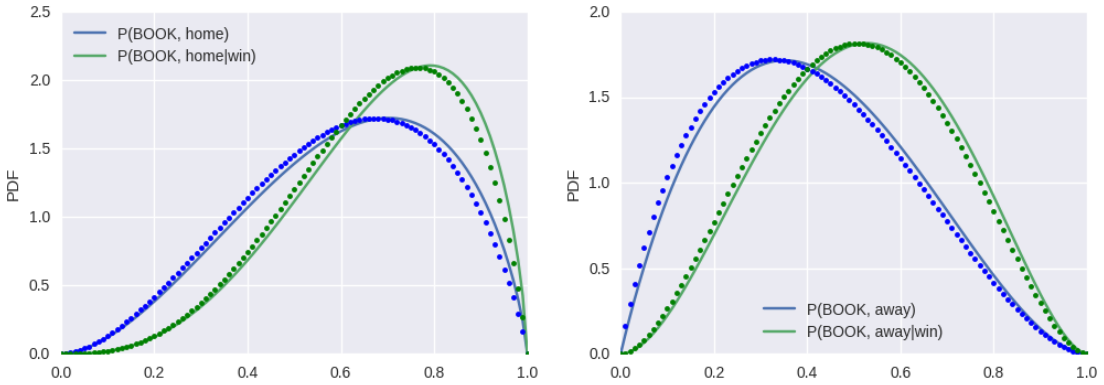

Figure 3.5: Probability density functions of $P(BOOK, COURT)$ and $P(BOOK, COURT|win)$ where $BOOK$ is the probability implied by the bookmaker's odds. Dotted lines represent the corresponding distribution fitted onto data with stripped margin.

We used probabilities implied directly by the odds, therefore including margin, as well as probabilities with stripped margin. The resulting distributions are illustrated on Figure 3.5. To get rid of the margin, we used Shin's model described in Shin, 1991. According to this model, the actual probability estimate of winning given by bookmaker's odds can be computed as:

$$p_{B_i} = \frac{\sqrt{z^2 + 4(1-z)\frac{\pi_i^2}{\Pi}} - z}{2(1-z)}$$

$$z = \frac{(\pi_+ - 1)(\pi_-^2 - \pi_+)}{\pi_+(\pi_-^2 - 1)}$$

where $\pi_1 = 1/odds_h$, $\pi_2 = 1/odds_a$, $\Pi = \pi_1 + \pi_2$, $\pi_+ = \pi_1 + \pi_2$ and $\pi_- = \pi_1 - \pi_2$

This model is known as the one providing most accurate probabilities in terms of Brier loss as was shown in Strumbelj, 2014. The Brier loss is the mean squared difference of the actual outcome and the predicted probability of that outcome.

Both distributions shifted closer to zero as expected since the implied probabilities including margin are overestimated.

Using the prior probability $p(win)$ obtained from the histogram, we estimated the probability $P(win|ODDS)$ (Figure 3.6).

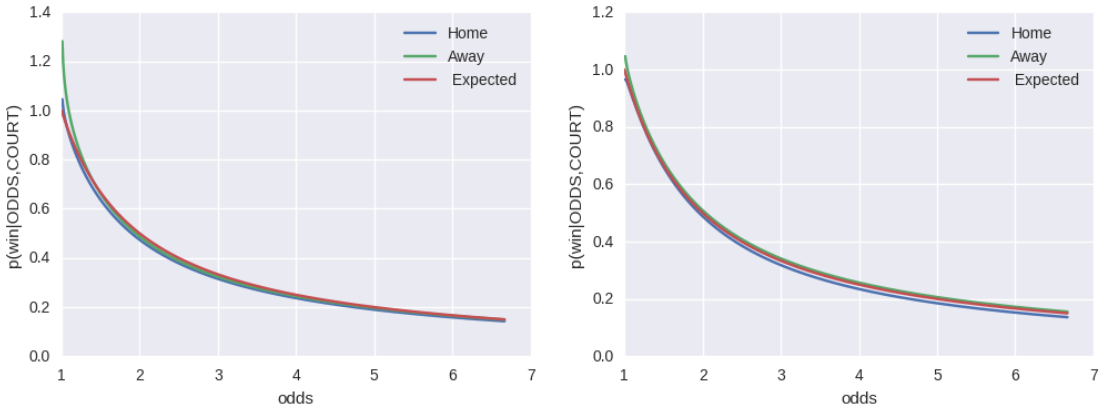

Figure 3.6: Estimate of $P(win|ODDS)$ from data with margin (left) and with stripped margin (right)

For visiting team, the estimated probability exceeded 1. That is an artifact of distribution fitting, where we fitted $P(ODDS, away)$ and $p(ODDS, away|win)$ independently. Just by looking at the graph we could say, there is no evidence of systematic bias. To confirm our suspicion we conducted a statistical test and trained a simple machine learning model.

3.4.1.2 Machine Learning Approach

Another approach how to find out whether there is a bias in bookmaker's odds is to train a model using the bookmaker's odds as the only feature. If the trained model were more accurate than the bookmaker, it would suggest, that the model learned how to exploit the bookmaker, therefore there must be a bias in bookmaker's odds. First of all, we analyzed what is the bookmaker's accuracy. In the last 5 seasons of our dataset, bookmaker's accuracy was 0.6915. Then we trained a neural network with one hidden layer. In 100-fold cross-validation, the average accuracy was 0.6904. This result suggests that the network learned a model close to the bookmaker's belief but not better than that.

3.4.2 Proposed Models

In this section, we introduce the neural models we used to obtain the reported predictions.

The differences in using different regularization techniques are only marginal. The most noticeable difference is in the models' architectures. This is in agreement with our intuition and assumption of the overall stability of our neural models.

Meta-parameter	Team-level	Player-level
Architecture	D128-D64-D32-D16-D1	c1-D64-D32-D16-D1
Activations	tanh	tanh
Dropout	0.2	0.3
L2 regularization	0.0001	0
Initialization	normal	normal

Table 3.1: Meta-parameters of the used Models

While the team-model consist solely of dense layers with 128, 64, 32, 16 and 1 neurons, we made use of the convolutional layer when dealing with the player-level statistics. The first idea was to make convolutional filters over all statistics of the player to obtain a single number for each player and filter representing some kinds of player ratings. This method proved to be ineffective. We extended the idea by using locally connected layer – different rating rules for different players. We relied on the fact that we sorted the players by minutes played in the season. Neither this approach worked particularly well. We tweaked our idea by applying the filter over each of the statistics of all players. This way we aggregated the information in a similar way as it was done in the team-level dataset, where some of the statistics were just a sum of statistics of the players. For example, the number of rebounds of the team in the season was just a sum of the rebounds of the player of the team. By using convolutional filters, we allow the model to discover more complex relations, for example in the distribution of rebounds between the players.

3.4.3 Summary

The purpose of this section was to search for a systematic bias in bookmaker's odds. Should we have found such a bias, the bookmaker would be easily exploitable without any deeper modeling of the data. However, neither statistical approach nor machine learning approach discovered any systematic bias in bookmaker's odds. Even though we conclude that the bookmaker is unbiased and thus on average precise, there is still space for opportunities for profiting with deeper modeling of the data in more complex feature spaces. Towards this goal, we propose the team-level and the player-level models.

3.5 Betting Strategies

A betting system consists of two major components. A model, estimating the probability of a becoming true and a betting strategy, combining the probabilities of associated outcomes given by the predictive model with the bookmaker's odds to determine how to split the stakes.

3.5.1 Motivation

From the odds we can derive probability of winning estimated by bookmaker p_B. Output of our prediction model is another estimation p_M of the true probability p_T. There are different scenarios that can occur as follows.

$$p_M < p_B < p_T \tag{3.1}$$
$$p_M < p_T < p_B \tag{3.2}$$
$$p_B < p_M < p_T \tag{3.3}$$
$$p_B < p_T < p_M \tag{3.4}$$
$$p_T < p_M < p_B \tag{3.5}$$
$$p_T < p_B < p_M \tag{3.6}$$

When $p_B < p_T$, the bookmaker underestimated the likelihood of the event and set the odds too high, creating an opportunity for generating profit. On the other hand, when $p_B > p_T$, then the set odds are skewed in bookmaker's favor.

When $p_B < p_M$ we assume, that bookmaker had set the odds too high, therefore we bet, otherwise we don't. However, how much we bet depends on a chosen betting strategy.

3.5.2 Definition

We can define the betting strategy for n betting opportunities as a following:

$$f : M^n \times B^n \to W^n$$

$$\begin{aligned} M_i &= \text{probability of winning given by model} \\ B_i &= \text{probability of winning given by bookmaker's odds} \\ W_i &= \text{portion of wealth staked} \end{aligned}$$

For each round of league matches consisting of $\frac{n}{2}$ matches the bookmaker sets odds for each outcome, creating n betting opportunities. Our model output its own estimates of the probabilities. A betting strategy takes the bookmakers odds and the model's estimates and outputs portion of wealth (the bets) to be waged on each betting opportunity.

3.5.3 Expected Return and Variance of Profit of a Bet

For each betting opportunity opp_i we can calculate the expected return and the variance as

$$E(opp_i) = pwin_i(odds_i - 1) + (1 - pwin_i)(-1) = pwin_i odds_i - 1 \tag{3.7}$$
$$\sigma^2 = E[X^2] - E[X]^2 = pwin_i(1 - pwin_i)odds_i^2 \tag{3.8}$$

where $pwin_i$ is the estimated probability of the opportunity coming true and $odds_i$ are odds set by the bookmaker.

The true probability of betting opportunity coming true is unknown. Therefore we use our model's estimate to calculate the expected profit instead.

3.5.4 Betting Strategies in Literature

In the literature, we have encountered several betting strategies as follows.

- Bet fixed amount on favorable odds (fixbet).
- Bet amount equal to the absolute discrepancy between probabilities predicted by the model and the bookmaker (abs disc bet).
- Bet amount equal to the relative discrepancy between probabilities predicted by the model and the bookmaker (rel disc bet).
- Bet amount equal to the estimated probability of winning (conf bet).

However these strategies have not been formally analyzed in terms of risk, therefore we question their optimality. Moreover, we can show that according to the laid out criteria, none of these strategies is optimal.

3.5.5 Betting Strategies by Tipsters

There are plenty of tipsters online, offering their free or paid services. Customer subscribes to a service and then before each league round he receives tips on which betting opportunities are profitable. The tipster usually advises staked amount in terms of units. For each match, there is a possibility to bet 0-10 units. What these units represent in the model of the tipster is unclear. However, tipsters' goals are the same - to maximize the profit and minimize the risk. Moreover, without knowing the tipster's probability estimates along with the proposed units staked, determining his betting strategy is impossible.

3.5.6 Markowitz's Model

If our goal would be solely to maximize the expected profit, then the solution would be trivial – bet whole wealth on the opportunity with highest expected return of profit from presented opportunities. However, since the betting is a repeated process, we want some level of guarantee that we will have money available for the next betting opportunities. For these reasons, our criteria for selecting the betting strategy are expected profit and risk. Therefore whenever we mention the maximization of the profit, we refer to maximization of the profit while minimizing the risk.

We use a model described in Markowitz, 1952 to find the optimal strategy. The *Markowitz's model* (also known as *Modern Portfolio Theory*) is used in economics [1] for portfolio optimization, which is a problem very similar to the one we are facing. The portfolio assets can be viewed as the betting opportunities at our hand before each of the league rounds.

3.5.6.1 Expected Return and Variance of the Portfolio's Profit

The expected return of the portfolio and the variance are defined as

$$E(R_p) = \sum w_i E(R_i) \tag{3.9}$$

$$\sigma_p^2 = \sum w_i^2 \sigma_i^2 + \sum \sum w_i w_j \sigma_i \sigma_j p_{ij} \tag{3.10}$$

where R_p is the return of the portfolio, σ_p^2 is the variance of the portfolio, w_i is the wealth staked on asset i, R_i is return of asset i, σ_i^2 is the variance of return of asset i and p_{ij} is the correlation of returns of assets i and j.

In our case, we may reasonably assume that outcomes of matches are not correlated, so the variance of the portfolio simplifies to $\sigma_p^2 = \sum w_i^2 \sigma_i^2$.

3.5.6.2 Efficient Frontier

Every possible distribution of bets given opportunities can be plotted into a so called *risk-expected return space*. The efficient frontier is then another term for *Pareto-frontier*. *Pareto frontier* is a set of solutions to a problem with multiple objectives, where improving in one objective would worsen another objective. In our case, such a solution is an allocation of the bets among the betting opportunities such that different allocation would lead to higher risk or lower expected return.

3.5.6.3 Proposed Strategy

We have decided to choose the final *opt* strategy from the frontier using *Sharpe ratio* introduced in Sharpe, 1994 as

$$\frac{r_p - r_f}{\sigma_p}$$

where σ_p is standard deviation of portfolio return, r_p is expected return of the portfolio and r_f is a risk-free rate. We neglect the risk free rate because there is no risk-free method how to appreciate the money during the duration of a match.

The proposed *opt* strategy is a strategy from the frontier that maximizes the *Sharpe ratio*.

[1] The author was later on awarded Nobel Prize in Economic Sciences.
`http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1990/`

Figure 3.7: Risk-expected return space with random strategies and efficient frontier

3.5.6.4 Criticism of Markowitz's Model

There has been some criticism on Markowitz's model. Here we take a look at the main points of the criticism.

Risks and returns are based on expected values. These values change in time (they are subjects of volatility). Once the bet is placed there are only two outcomes possible. Losing the bet or winning $bet \times odds$ money. Bets can be placed up to few minutes before the match starts, so the predictive model can accommodate current information, e.g. about team injuries.

Variance is a symmetric measure therefore extremely high returns are as risky as extremely low returns This downside, unfortunately, prevails in our case, too.

3.5.7 Evaluating Betting Strategies

Now when we introduced the framework we use to evaluate the betting strategies, we can evaluate the betting strategies we have encountered in the literature in terms of risk and expected profit.

The biggest downfall of the strategies mentioned in Section 3.5.4 is that they operate with different amount of wealth making them incomparable. To compare these strategies we had to normalize the bets placed on the betting opportunities so they sum up to 1.

To demonstrate the differences between the betting strategies we randomly sampled data representing six illustrative betting opportunities with positive expected return

($\pi < p_M$). Then we applied the betting strategies and analyzed the differences in the wealth staked by the different strategies. Results are summarized in Table 3.2.

opp	p_M	π	std	return	fix_bet	abs_disc	rel_disc	conf_bet	opt
1	0.30	0.26	1.80	0.19	0.17	0.20	0.35	0.08	0.09
2	0.59	0.52	0.94	0.12	0.17	0.27	0.24	0.16	0.23
3	0.75	0.70	0.62	0.07	0.17	0.21	0.15	0.21	0.30
4	0.60	0.57	0.86	0.06	0.17	0.13	0.12	0.17	0.12
5	0.74	0.71	0.62	0.04	0.17	0.11	0.08	0.20	0.17
6	0.64	0.62	0.77	0.03	0.17	0.07	0.06	0.18	0.08

Table 3.2: Comparison of betting strategies on simulated betting opportunities. The columns p_M and π represent the probability estimates of the model and the bookmaker, respectively, std and return refer to standard deviation and expected value of the profit.

The strategies based on the absolute and relative discrepancies between the probability estimates always prefer higher expected profit of the opportunity regardless of the variance of the profit. On the other hand, the strategy based on the confidence of the probability estimate always prefers lower risk, regardless of the return. The proposed optimal strategy is looking for a compromise between these two approaches.

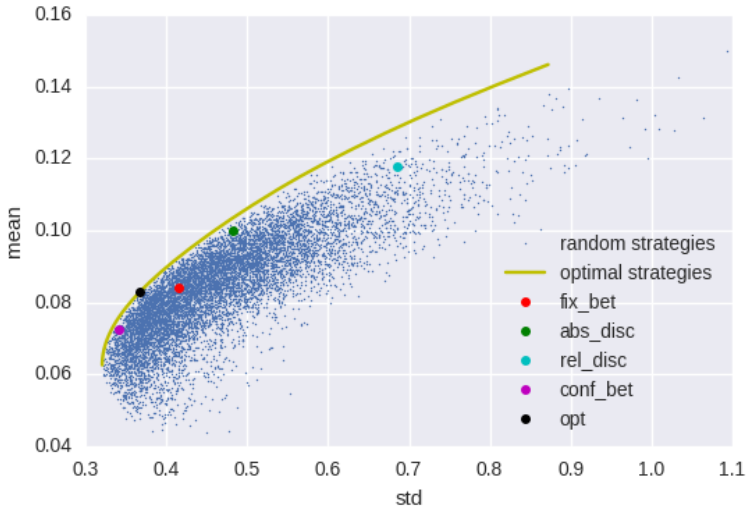

Figure 3.8: Comparison of the betting strategies in the risk-return space

Figure 3.8 shows, that none of the strategies we examined lies on the efficient frontier. In other words, none of the strategies is optimal given the introduced criteria of risk and expected profit. The distribution of the strategies we observe in the graph follows their characteristics described.

Chapter 4

Experiments

4.1 Simulation of the Betting Strategy

To get a better idea of how our selected betting strategy behaves, we generated data and observed the behavior on them. Generated data consisted of triplets representing the true probability of winning, the model's estimate and the bookmaker's estimate. For a joint analysis of all the hidden factors, the data were generated from a multivariate beta distribution parametrized with the means and the variances of the marginal distributions, representing the ground truth, model's and bookmaker's estimates, and their mutual correlations. The parameters of this joint distribution were subject to tuning and subsequent analysis. However, the marginal distribution representing the bookmaker was fitted with the real historical data obtained from a bookmaker. Since we didn't consider separate cases for home and away games, the mean of the ground truth model was set to 0.5, however, the variance was in principle unknown. The actual outcome of a game was determined by a Bernoulli trial.

Firstly we examined the effect of our model's variance on the profit. Table 4.1 shows, that model with lower variance yields higher profit regardless of the variance of the ground truth. Other metrics such as Log loss, Brier loss and accuracy are independent of the model's variance.

We inspected different scenarios that can occur depending on the bookmaker's prediction, our model's prediction and the actual outcome as follows.

If one team is a favorite in the eyes of the bookmaker, our model agrees, and this favorite wins, we call this scenario *Consensus*. If this team loses, we call it an *Upset*. If our model determined the opposite team as the favorite and was correct in that decision, we say the model *Spotted* an opportunity. If the model is not right, we conclude it as *Missed*. When we analyze the models w.r.t. these scenarios, we inspect the portion of games in each category.

We examined the relation between the profit and correlation of model's and bookmaker's estimate, as we expected this to be the core factor in profitability. In agreement with our intuition, we found out that lower correlation with bookmaker leads to higher

σ^2_{GT}	σ^2_M	MPPR	LogLoss	BrierLoss	Accuracy
0.05	0.05	0.1831	0.6078	0.2103	0.6667
	0.06	0.1270	0.6135	0.2117	0.6665
	0.07	0.0840	0.6220	0.2137	0.6673
0.06	0.05	0.1722	0.5886	0.2022	0.6829
	0.06	0.1382	0.5895	0.2023	0.6840
	0.07	0.0963	0.5955	0.2037	0.6845
0.07	0.05	0.1546	0.5705	0.1946	0.6996
	0.06	0.1372	0.5682	0.1938	0.7006
	0.07	0.1134	0.5733	0.1952	0.6991
0.08	0.05	0.1580	0.5517	0.1865	0.7162
	0.06	0.1415	0.5498	0.1861	0.7154
	0.07	0.1258	0.5512	0.1866	0.7137

Table 4.1: Effects of the ground truth and model's output distributions variances on the evaluation metrics

earnings, despite the fact that the accuracy of the model remained the same. This reflects the case, when the model correctly estimates outcomes of the games where bookmaker fails, in other words where the underdog wins (*Spotted*). Since odds for the underdogs are higher, winning these bets while wrongly estimating favorites in other games (*Missed*) is more profitable than trying to beat the bookmaker in games where there is a mutual agreement (*Consensus*).

$p_{GT,M}$	$p_{B,M}$	$MPPR_{opt}$	Accuracy	Consensus	Upsets	Missed	Spotted
0.85	0.85	0.1163	0.6763	0.5922	0.2246	0.0991	0.0842
	0.90	0.0640	0.6769	0.6077	0.2395	0.0836	0.0692
	0.95	-0.0043	0.6770	0.6274	0.2582	0.0648	0.0496
0.90	0.85	0.1781	0.6885	0.5985	0.2183	0.0932	0.0900
	0.90	0.1392	0.6884	0.6134	0.2339	0.0777	0.0750
	0.95	0.0966	0.6876	0.6323	0.2540	0.0584	0.0554
0.95	0.85	0.2388	0.6996	0.6036	0.2129	0.0875	0.0960
	0.90	0.2108	0.7003	0.6194	0.2281	0.0716	0.0809
	0.95	0.1933	0.6998	0.6384	0.2477	0.0525	0.0614

Table 4.2: Effects of different correlation levels between the true probabilities, the model and the bookmaker on the evaluation metrics.

4.2 Model Evaluation

4.2.1 Experiment Setup

In all the experiments, in order to obtain statistically conclusive results, we performed cross-validation as follows. Each season represents one fold in our setup. This way, when a model is tested on, for example, season 2012, the remaining seasons, including seasons 2013 and 2014, are in the training set. Since the datasets consist of the seasonal averages, there is no risk of corrupting the results as there is clearly no overlap between the training and test sets. Moreover, no features, that could identify the same team in different splits and therefore lead to assumptions about team strength based on matches that have yet to be played, are present.

The goal of this work was to exploit the bookmaker, hence we focus primarily on the profit measure. To measure the profit of the model, we use mean profit per round (MPPR). In NBA, the bookmakers usually set the odds for the next round of league matches in advance. After the odds are set, we can deduce the returns with our model and spread the bets according to our betting strategy. There are 30 teams in NBA, so one league round consists of 15 games. Table 4.3 summarizes the number of games for which we predicted the outcomes in each season and the bookmaker's accuracy in each season, respectively. The number of games differs between the seasons because we had to exclude games with incomplete statistics. Moreover, the first ten games of each team are not included as they serve for the initial calculation of seasonal averages.

The results presented (f.e. in Table 4.4) are averaged over the validation splits.

4.2.2 Baseline Model

Logistic Regression on team-level dataset served as our baseline model. Logistic Regression is a simple linear model that can be looked upon as ANN with only one neuron using sigmoid as the activation function. Therefore it serves well as an indicator for structural and meta-parameter setup of our extended neural models. Moreover, although being a classifier, it naturally regresses posterior probability of the target classes, and so its output can be directly interpreted as a probability, as required in our subsequent experiments with betting strategies.

In the experiments, the baseline model was losing money in the course of 15 seasons. Interestingly, the model's accuracy was lower than the bookmaker's even though the bookmaker's odds were presented in the features.

4.2.3 ANN Model

Our ANN model outperformed baseline model both in the accuracy and the profitability (Table 4.4).

Beating bookmaker in terms of accuracy proved to be a very hard task, even though the bookmaker's predictions are known to our model. This paradox signalizes, that despite the fact that we used several regularization techniques, the model was fitting

Year	Bookmaker's Acc	Games
2000	0.6955	926
2001	0.6429	983
2002	0.6697	1005
2003	0.6743	1007
2004	0.6973	839
2005	0.6753	924
2006	0.6674	953
2007	0.6832	1070
2008	0.7105	1057
2009	0.6907	1070
2010	0.7066	1043
2011	0.6707	829
2012	0.6903	1072
2013	0.6867	999
2014	0.7140	1000
Average	0.6850	985

Table 4.3: Number of games and bookmaker's accuracy in each of the seasons

some noise during the training phase. Moreover, we can assume, that bookmaker had access to the same data we have and possibly more. The bookmaker could also leverage his domain knowledge and distinguish the noise from an actually useful information.

model	$MPPR_{opt}$	$MPPR_{fixbet}$	Log Loss	Brier Loss	Accuracy
ANN	0.0057	-0.0048	0.5902	0.2028	0.6824
LR	-0.0043	-0.0151	0.5937	0.2037	0.6807

Table 4.4: Mean results over the seasons of the LR and ANN model

To further analyze the results, we plotted the models' accuracy over the course of 15 seasons. From the Figure 4.1 we can see, that the models are correlated in terms of the accuracy not only between themselves but also with the bookmaker. Only in seasons 2004/05 and 2005/06 there are some differences between the Logistic Regression and the ANN. We observed some level of decorrelation with the bookmaker in seasons 2003, 2004, 2011, and 2013. In all of these seasons, the bookmaker outperformed our models.

We took a closer look into the models' correlation. We were mostly interested in the correlation of our model with the bookmaker. We found out, that the correlation was indeed very strong.

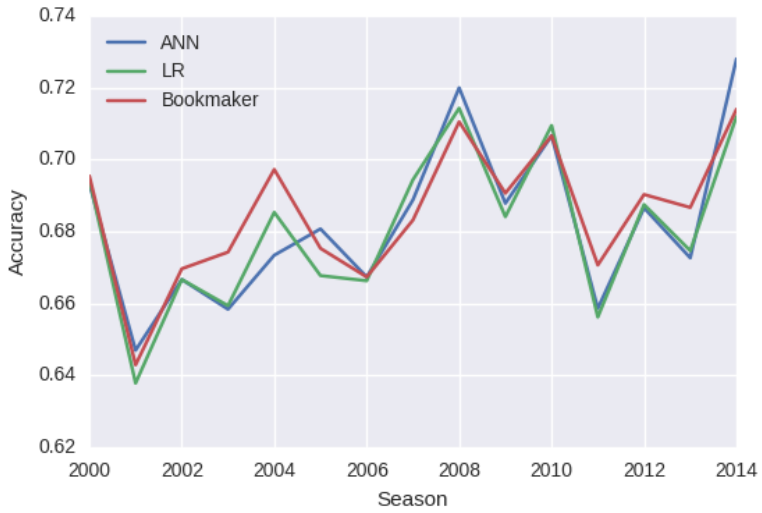

Figure 4.1: Accuracy of the ANN and LR models over seasons 2000–2014

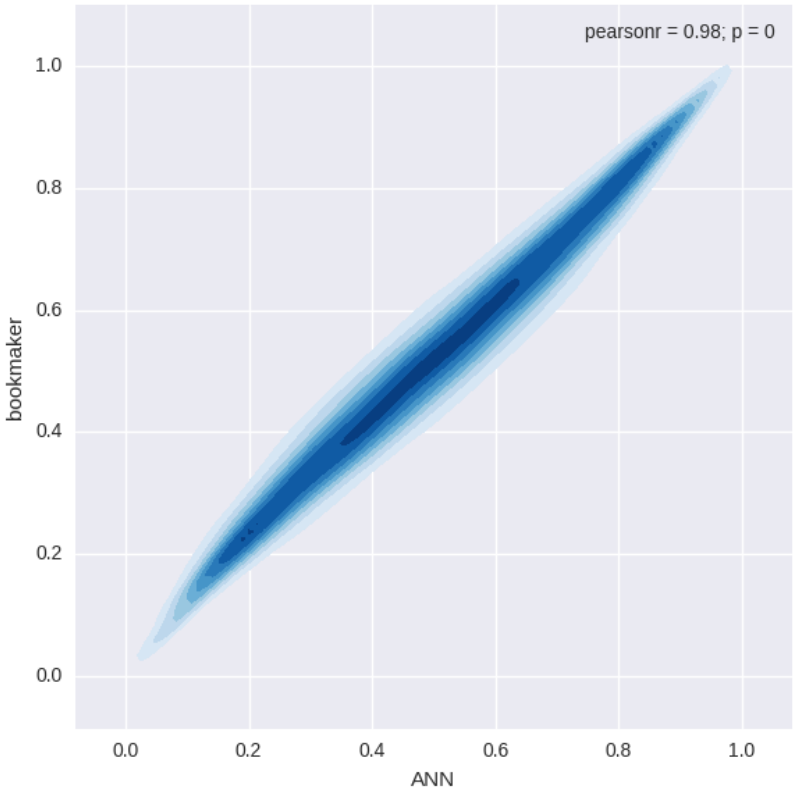

Figure 4.2: Correlation between the ANN model and bookmaker's predictions

However, we note that our goal was not to beat the bookmaker in terms of accuracy but to generate profit against his odds setup. Therefore, we examined the models' profitability over the seasons (Figure 4.3).

Figure 4.3: Comparison of models' profitability using the proposed (left) and fixbet (right) betting strategies.

Logistic Regression ended up in negative numbers when using the proposed betting strategy even though the worst seasonal result belonged to ANN in season 2013.

Both models failed to generate the profit using fixbet strategy. This result is in line with our finding that the fixbet strategy is suboptimal.

4.2.4 Early Findings

After the first series of experiments we noted several key points:

- Beating the bookmaker in terms of accuracy is very hard.
- Our model is naturally highly correlated with the bookmaker.
- Profit generated is very low.

When we say, that the generated profit is very low, it is a relative statement. Our unfounded intuition suggested profit in some low percents. However, solely the fact that the model was able to be profitable over the course of 15 seasons shows that opportunities for exploiting the bookmaker should exist.

After the preliminary analysis, we established following key options to increase the profit.

1. Increase the model's accuracy.
2. Accurately predict the games where bookmaker fails allowing for lower accuracy.
3. Bet only on the games where the model accurately predicts the winner.

We had already discovered that 1. is hard. Our simulations suggested that the level of correlation between the model and the bookmaker is crucial for the profit, overshadowing the model's accuracy. Therefore we further analyze the second approach. The third method sounds impossible since we can not place the bet after the outcome of the game is known. However, what we can do is to be more selective about the best we place and bet only if the model's confidence is above a certain level.

We further elaborate on the strategies aimed at increasing the profit, motivated by our early findings, in the following sections.

4.2.5 Model Decorrelation

If the team favored by the bookmaker wins, the payouts are smaller than in the opposite case because the favorite's odds are lower. Therefore we want to accurately predict as many underdogs' wins as possible.

We argue that being correlated with the bookmaker is not the way to exploit his model. Bookmaker's edge comes from unfair odds and, on average, superior model w.r.t. the predictions' accuracy. To decorrelate with the bookmaker, we want to increase the number of *Spotted* opportunities while keeping *Missed* games reasonably low.

In this section, we propose and evaluate several methods aimed to achieve the decorrelation with the bookmaker.

4.2.5.1 Sample Weighting

Decorrelation can be forced by telling the model which games it should focus on. This can be achieved by weighting the samples.

Since the betting opportunities with higher odds could lead to higher returns, we used winners' odds as weights of the samples. The second mode we tested was to weight only *Upsets* by the winners' odds and assigning uniform weights to the rest of the samples.

We tried two different ways to implement sample weighting – oversampling and multiplying loss of each sample by its weight. The downside of the oversampling method is a higher computational cost. Odds are given in decimal format, therefore we multiplied them by 10 and rounded the result. In the end, both implementations achieved very similar results, as expected. Slight differences were probably solely due to rounding. In Table 4.5 we present the results achieved with the loss weighting implementation.

model	$MPPR_{opt}$	$MPPR_{fixbet}$	Log Loss	Brier Loss	Accuracy
ANN	0.0057	-0.0048	0.5902	0.2028	0.6824
ANN_weighting	-0.1450	-0.1307	0.6715	0.2392	0.6514
ANN_upset_weighting	-0.1333	-0.1243	0.7009	0.2538	0.4852

Table 4.5: Results of ANN models trained with sample weighting

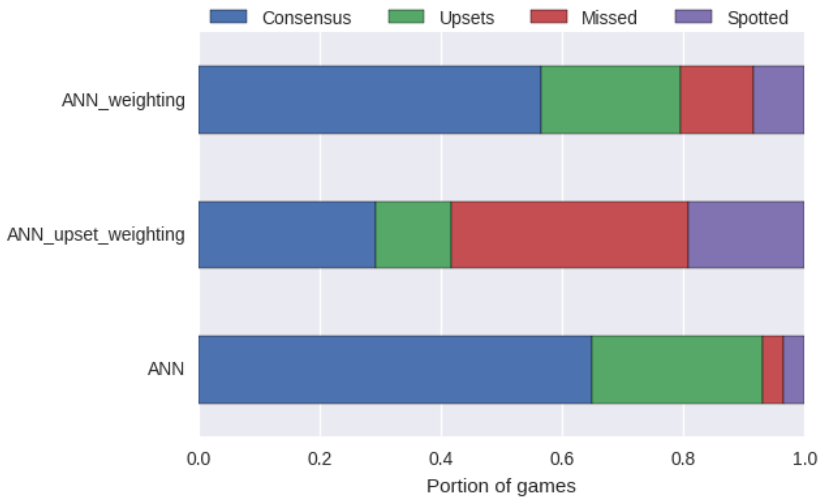

Figure 4.4: Analysis of predictions of models trained with sample weighting

However, in the experiments we found out that both weighting techniques had a negative impact on the profit. In the majority of seasons, the profit was lower than the profit of the unweighted model (Figure 4.5). The accuracy decreased as was expected, however, the number of *Spotted* opportunities wasn't high enough to compensate for this drop (Figure 4.4).

Figure 4.5: Comparison of models' profitability using proposed (left) and fixbet (right) betting strategy

4.2.5.2 Embedding Decorrelation Term into Loss Function

Another way how to ensure the decorrelation with the bookmaker is to enforce it in the loss function minimized during the training phase. Towards that purpose, we introduced a modified mean squared error as follows.

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N}((out_i - y_i)^2 - c \cdot (out_i - \pi_i)^2)$$

where out_i is the model's output for the ith example, y_i is the desired output and $\pi_i = 1/odds_i$ where $odds_i$ are the odds set by bookmaker. The first term ensures our model will be correlated with the true probability, i.e. increases accuracy, while the second term enforces decorrelation with the bookmaker. The constant c indicates relative significance of the decorrelation term. We the experiments with different values for c and analyzed the results in Table 4.6. The observed profit was generally higher for $c > 0.5$ than in the opposite case. Other metrics were worsening with increasing c.

c	$MPPR_{opt}$	$MPPR_{fixbet}$	Log Loss	Brier Loss	Accuracy
0.0	-0.0036	-0.0264	0.5910	0.2031	0.6809
0.1	0.0014	-0.0163	0.5915	0.2034	0.6818
0.2	-0.0065	-0.0239	0.5931	0.2040	0.6811
0.3	0.0012	-0.0058	0.5940	0.2044	0.6812
0.4	0.0015	-0.0123	0.5971	0.2057	0.6783
0.5	0.0086	-0.0149	0.6036	0.2080	0.6778
0.6	0.0084	-0.0119	0.6124	0.2112	0.6740
0.7	0.0119	-0.0086	0.6356	0.2183	0.6717
0.8	0.0064	-0.0048	0.6630	0.2263	0.6648
0.9	0.0035	-0.0220	0.9860	0.2630	0.6524
1.0	0.0057	-0.0023	2.3284	0.3546	0.6252

Table 4.6: Result of ANN model trained with the decorrelation term in loss function for varying values of c

The Figure 4.6 shows, that increasing c indeed leads to larger number of *Spotted* games at the expense of more *Missed* games.

We examined the distribution of our model's output depending on the c. The Figure 4.7 displays, how the variance of the output increases with larger c. When $c = 1$ the output of the model is basically binary with no additional information about its confidence.

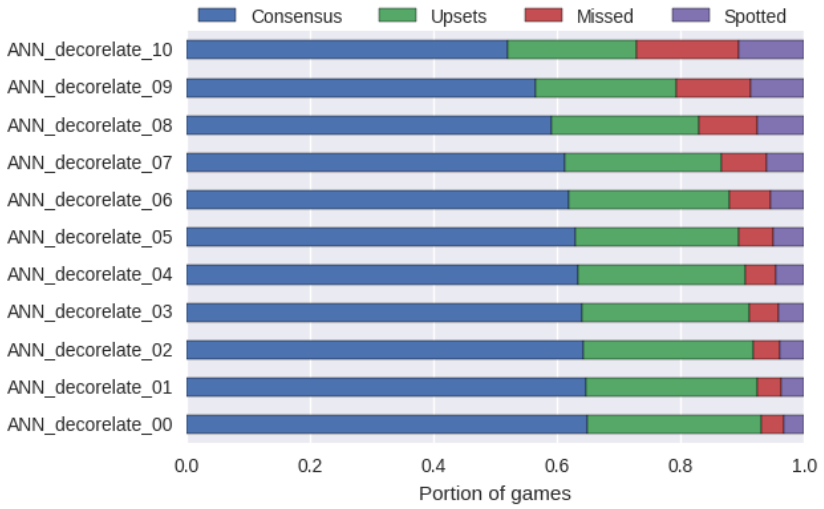

Figure 4.6: Analysis of predictions of the ANN model trained with the decorrelation term in loss function for varying values of c

4.2.6 Confidence Threshold

Another approach we discussed in Sec. 4.2.4 to increase the profit was minimizing the relative portion of mistakes our model makes, i.e. be more selective about whether we should be betting at all.

To instantiate this approach, we analyzed model's probability estimates and tried betting only when the model's confidence was above a certain threshold (Table 4.7). This way we were more conservative about the bets placed.

Despite the fact, that the number of placed bets decreased rapidly with increasing threshold, the profit increased significantly as did accuracy. The loss when using the threshold 0.9 could be due to not enough games where one team is such a strong favorite and the bookmaker's odds being very low when the outcome is so definite.

Conf	$MPPR_{opt}$	$MPPR_{fixbet}$	Accuracy	Brier Loss	Games	Bets Placed
>= 0.0	0.0057	-0.0048	0.6824	0.2028	985.1333	614.6667
>= 0.6	0.0102	0.0033	0.7385	0.1842	683.4000	403.1333
>= 0.7	0.0204	0.0104	0.8011	0.1540	406.4000	216.6667
>= 0.8	0.0315	0.0171	0.8734	0.1097	179.8667	75.8667
>= 0.9	-0.0483	0.0082	0.9651	0.0372	27.5333	8.4667

Table 4.7: Results of confidence thresholding of the team-level model

When exploring the models' predictions 4.8, we came to an interesting conclusion. When the confidence of our model is 70 % or higher, the model is always in *Consensus*

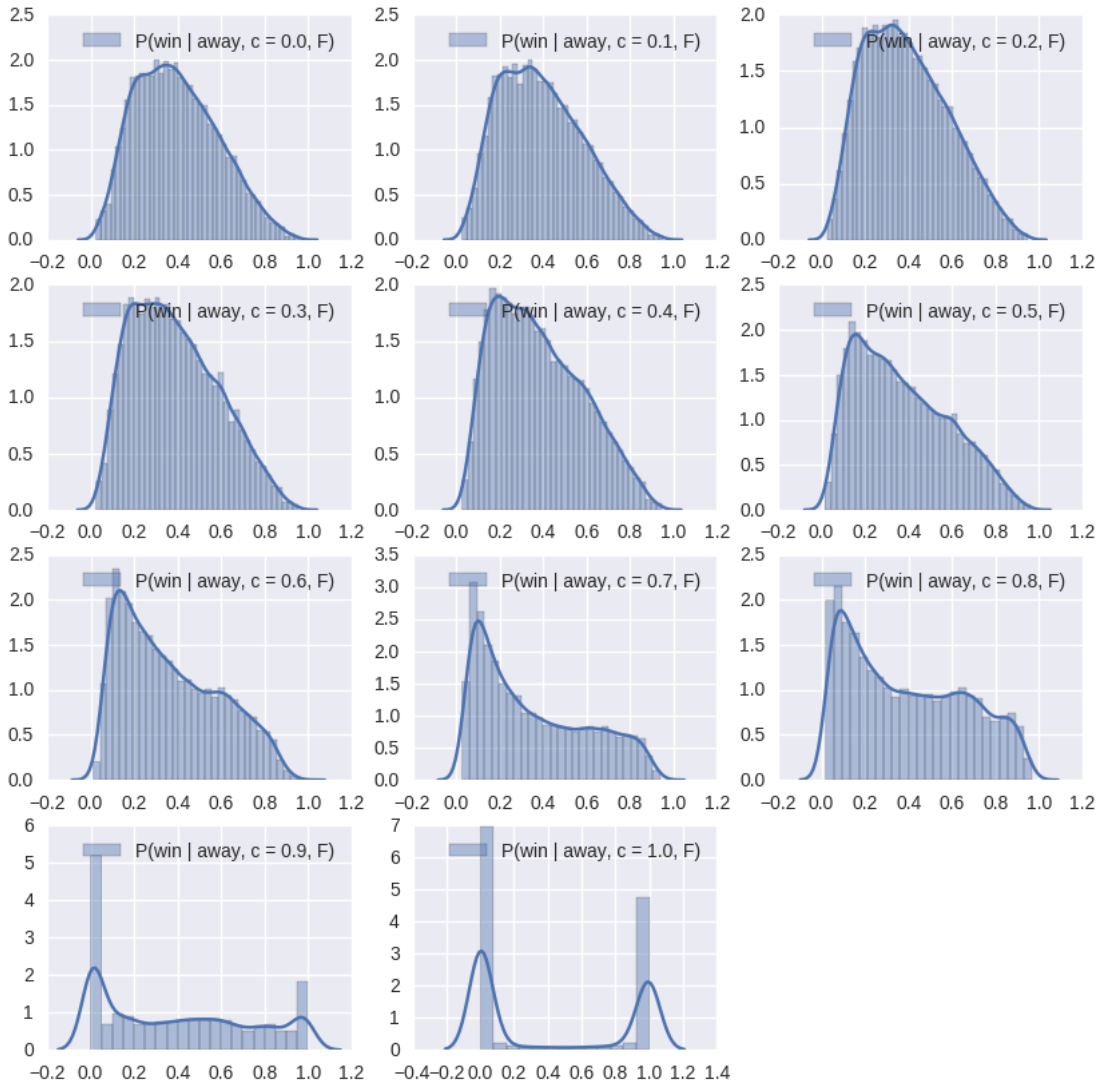

Figure 4.7: Distribution of $P(win|away, c, F)$ for different values of c, where F represents all the remaining features and parameters.

with the bookmaker. In other words, while the model and the bookmaker both correctly determined favorite of the game, the model's probability estimate was more precise, allowing profitable bet either on the favorite or on the underdog. Our hypothesis is that this is an artifact of using sigmoid activation function in the output neuron. As we stated in Section 1.4.4, the sigmoid function is very sensitive to changes in input when it's output is close to 0.5. This could result in unreliable predictions when the game has no clear favorite.

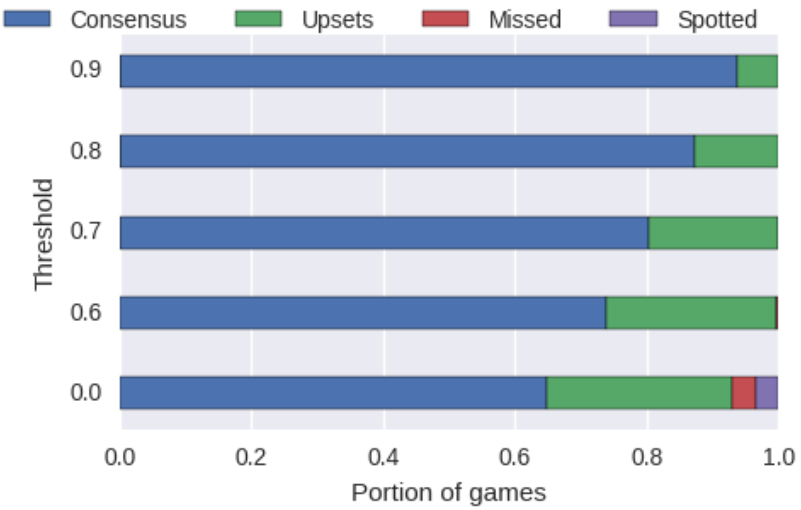

Figure 4.8: Analysis of predictions of the thresholded models

4.2.7 Player-level Model

So far we have been considering only the team-level model. The main differences between the team-level model and the player-level model are that the player-level model is trained on the dataset containing seasonal averages of each player instead of team-level which is trained using a dataset aggregating the statistics for the whole team. The second and very important difference is that the player-level dataset does not contain the odds. Even though it would be trivial to include the odds in this dataset we opted not to do so. We learned our lesson with the team-level model, where we showed that the trained model is very correlated with the bookmaker, which is an undesired behavior when we are trying to exploit the bookmaker in terms of profit.

At first, we compared the accuracy of the player-level model with the team level-model. As it can be seen on the Figure 4.9, the team-level model is superior in this objective which could be very well due to the bookmaker's odds serving as a good predictor the model can rely on.

We were particularly interested in the model's correlation with the bookmaker, now when the odds were not provided during the training. It turned out (Figure 4.10) that the level of correlation was significant, however, lower than the correlation between the team-level model and the bookmaker.

The results illustrated in Figure 4.11 were favorable. Despite the model's accuracy being lower than the team-level model's, the profit said a different story. Only in the season 2011, the lower accuracy took a toll on the model's profit. In most seasons, the model generated some profit.

The results are summarized in Table 4.8. The generated profit surpassed the team-level model by a large margin, despite achieving lower accuracy, further strengthening our point, that higher accuracy does not necessarily lead to higher profit.

Figure 4.9: Accuracy of the ANN and ANN_player models in seasons 2000–2014

model	$MPPR_{opt}$	$MPPR_{fixbet}$	LogLoss	BrierLoss	Accuracy
ANN_player	0.0239	-0.0193	0.6004	0.207	0.6748

Table 4.8: Results of ANN_player model

4.2.7.1 Player-level Model Variations

We applied the same methods for potentially increasing the profit as we did with the team-level model.

Table 4.9 summarizes the results of the model trained using loss function described in Section 4.2.5.2. In contrary to our findings from the player-level model (Table 4.6) increasing the significance of the decorrelation term decreased the profit. On the other hand, unfavorable trends such as increasing Brier Loss and decreasing Accuracy prevailed.

The story repeated itself when we tried applying the confidence threshold on the network's output. While this method dramatically increased the profit of team-level model (Table 4.7) it had a negative impact on the earnings of the player-level model, despite the fact that the accuracy was following the same increasing trend as with team-level model.

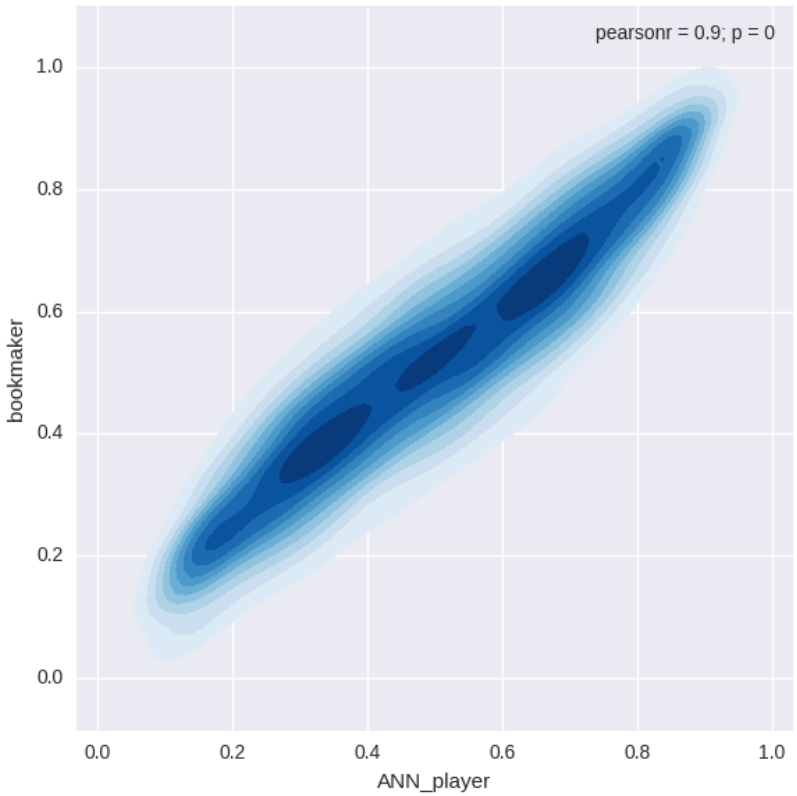

Figure 4.10: Correlation between the ANN_player model and bookmaker's predictions

Figure 4.11: Comparison of models' profitability using proposed (left) and fixbet (right) betting strategy

c	$MPPR_{opt}$	$MPPR_{fixbet}$	Log Loss	Brier Loss	Accuracy
0.0	0.0222	-0.0229	0.6000	0.2068	0.6753
0.1	0.0211	-0.0221	0.6013	0.2074	0.6739
0.2	0.0221	-0.0119	0.6020	0.2076	0.6736
0.3	0.0214	-0.0089	0.6037	0.2083	0.6719
0.4	0.0140	-0.0187	0.6094	0.2106	0.6702
0.5	0.0126	-0.0181	0.6175	0.2136	0.6682
0.6	0.0139	-0.0103	0.6292	0.2174	0.6622
0.7	0.0017	-0.0149	0.6592	0.2261	0.6578
0.8	-0.0090	-0.0268	0.7300	0.2424	0.6434
0.9	-0.0946	-0.1002	0.9064	0.2670	0.6162
1.0	-0.0435	-0.0444	2.4308	0.3503	0.6071

Table 4.9: Result of ANN_player model trained with decorrelation term in loss function with different values of c

Conf	$MPPR_{opt}$	$MPPR_{fixbet}$	Accuracy	Games	Bets Placed
>= 0.0	0.0239	-0.0193	0.6748	985.1333	795.2667
>= 0.6	0.0208	-0.0231	0.7266	704.9333	559.6000
>= 0.7	0.0226	-0.0255	0.7826	418.6000	323.1333
>= 0.8	0.0126	-0.0060	0.8608	175.8000	125.6667
>= 0.9	-0.0580	0.0162	0.9444	3.5333	2.5333

Table 4.10: Results of confidence thresholding of the player-level model

4.2.8 Joint Model

Naturally, the next step was combining the player-level model and the team-level model to ideally obtain a model combining their advantages. Unfortunately, we did not find a setting that would lead to an improvement over the independent models. We tried learning a new model, which inputs were outputs of the two models as well as merging the neural networks and both settings of pre-training the two models as well as learning the joint model end-to-end.

4.3 Comparison with State-of-the-art

In this section, we compare our results with the state-of-the-art. It is important to note, that we focused solely on the profit measures.

4.3.1 Profitability

The main goal we focused on was to generate profit against the odds set by the bookmaker.

To our best knowledge, there was no previous work of a comparable scale done. Evaluating a model, for example, on a single season is meaningless since the profit naturally varies a lot. Moreover, as we stated in Section 3.5.7, the comparison of betting strategies is usually impossible because they operate with different amount of wealth. An example of such a work can be found in Constantinou et al., 2013

Highest profits in our case were generated by the team-level model with confidence threshold and default player-level model.

4.3.2 Accuracy

Even though models' accuracies were not of our concern, we present the comparison of our models with the state-of-the-art model from Puranmalka, 2013. Table 4.11 shows, that leveraging play-by-play data and development of new features based on domain knowledge, as the author did in his Ph.D. thesis, is superior w.r.t. accuracy to using the box score data solely.

Year	ANN	ANN_player	Puranmalka
2000	0.6933	0.6911	
2001	0.6470	0.6429	
2002	0.6667	0.6597	
2003	0.6584	0.6564	0.7345
2004	0.6734	0.6806	0.7320
2005	0.6807	0.6623	0.7245
2006	0.6674	0.6569	0.7295
2007	0.6888	0.6916	0.6995
2008	0.7200	0.7077	0.7270
2009	0.6879	0.6841	0.6945
2010	0.7066	0.7057	0.7120
2011	0.6586	0.6248	0.7020
2012	0.6866	0.6782	0.6720
2013	0.6727	0.6717	
2014	0.7280	0.7080	
Average	0.6824	0.6748	0.7127

Table 4.11: Comparison of accuracy of our models with state-of-the-art model from Puranmalka, 2013.

Chapter 5

Conclusion

The goal of our work was to find a way to exploit betting market inefficiencies, i.e. to generate profit against bookmaker's odds setup, using machine learning. To the best of our knowledge, no work of similar scale, focusing on profitability, has been done and made publicly available before.

Our findings are in agreement with the previous research that the bookmaker is very hard to beat in terms of accuracy. However, we argued that accuracy is not a direct measure of profit, and we illustrated the use of Modern Portfolio Theory in the context of the betting market, allowing us to more formally analyze and compare various betting strategies, that are present in the literature, in order to optimize our betting behavior. By focusing explicitly on the profitability, our proposed models were able to generate profit over the course of the presented 15 seasons of NBA records.

We further argued that models' correlation with the bookmaker plays a crucial role when considering the profit, overshadowing the importance of accuracy. Several strategies on how to achieve decorrelation with the bookmaker were tested to confirm our assumption. Enforcing the decorrelation in the loss function used during the training phase and excluding the bookmaker's odds from the features proved to be the most effective approaches.

Another point we made is that the bettors can leverage their choice not to bet when the model's confidence in the outcome is low. Being more conservative about placing the bets turned out to be a feasible solution for generating the profit even while being correlated with the bookmaker.

Experiments with our proposed neural architectures demonstrated, that by aggregation of player-level statistics by a self-learned function, in our case a convolutional filter, we can obtain highly useful latent features, often beating the standard, expert designed statistical indicators on the team level.

5.1 Future Work

The scope of future work is naturally very broad and covers most of the parts of the described pipeline.

We have experimented with training the models with customized loss function using the decorrelation term. Although this modification proved useful, we believe that there might be more suitable loss functions with the same goal we have not yet discovered. It is also not clear, which betting strategy to use when the output of the network is affected by the decorrelation term in the loss function.

We do not claim the used Mean Portfolio Theory to be flawless. We acknowledged the drawback of using the variance of the profit as a measure of risk. Other measures for the risk could also be explored.

As we already mentioned, our dataset consisted solely of the seasonal averages. Recurrent neural networks or other techniques could be used to leverage the time-series characteristics of the seasons.

Another challenge is to find an efficient way of combining the player-level and the team-level model.

To further validate our findings, experiments on different sub-domains (sport) or different types of bets (f.e. proposition bets) should be done. Another validation can be done by evaluating the models in the upcoming seasons.

Bibliography

ARKES, Jeremy; MARTINEZ, Jose, 2011. Finally, Evidence for a Momentum Effect in the NBA. *Journal of Quantitative Analysis in Sports*. Vol. 7, no. 3, pp. Article 13.

BAR-ELI, Michael; AVUGOS, Simcha; RAAB, Markus, 2006. Twenty years of "hot hand" research: Review and critique. *Psychology of Sport and Exercise*. Vol. 7, no. 6, pp. 525–553.

CONSTANTINOU, Anthony Costa; FENTON, Norman Elliott; NEIL, Martin, 2013. Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using Bayesian networks. *Knowledge-Based Systems*. Vol. 50, pp. 60–86.

FORREST, David; GODDARD, John; SIMMONS, Robert, 2005. Odds-setters as forecasters: The case of English football. *International Journal of Forecasting*. Vol. 21, no. 3, pp. 551–564.

FRANCK, Egon; VERBEEK, Erwin; NUESCH, Stephan, 2010. Prediction accuracy of different market structures - bookmakers versus a betting exchange. *International Journal of Forecasting*. Vol. 26, no. 3, pp. 448–459.

FRANCK, Egon; VERBEEK, Erwin; NÜESCH, Stephan, 2009. *Inter-market arbitrage in sports betting*. Technical report. National Centre for Econometric Research.

GODDARD, John; ASIMAKOPOULOS, Ioannis, 2004. Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*. Vol. 23, no. 1, pp. 51–66.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron, 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

HAGHIGHAT, Maral; RASTEGARI, Hamid; NOURAFZA, Nasim, 2013. A Review of Data Mining Techniques for Result Prediction in Sports. *Advances in Computer Science*. Vol. 2, no. 5, pp. 7–12.

HVATTUM, Lars Magnus; ARNTZEN, Halvard, 2010. Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*. Vol. 26, no. 3, pp. 460–470.

IBÁÑEZ, Sergio J; GARCÍA, Javier; FEU, Sebastian; LORENZO, Alberto; SAMPAIO, Jaime, 2009. Effects of consecutive basketball games on the game-related statistics that discriminate winner and losing teams. *Journal of Sports Science and Medicine*. Vol. 8, no. 3, pp. 458–462.

IBÁÑEZ, Sergio J.; SAMPAIO, Jaime; FEU, Sebastian; LORENZO, Alberto; GÓMEZ, Miguel a.; ORTEGA, Enrique, 2008. Basketball game-related statistics that discriminate between teams' season-long success. *European Journal of Sport Science*. Vol. 8, no. 6, pp. 369–372.

IVANKOVIC, Z; RACKOVIC, M; MARKOSKI, B; RADOSAV, D; IVKOVIC, M, 2010. Analysis of basketball games using neural networks. *Computational Intelligence and Informatics (CINTI), 2010 11th International Symposium*, pp. 251–256.

KAIN, Kyle J.; LOGAN, Trevon D., 2014. Are Sports Betting Markets Prediction Markets?: Evidence From a New Test. *Journal of Sports Economics*. Vol. 15, no. 1, pp. 45–63.

KUBATKO, Justin; OLIVER, Dean; PELTON, Kevin; ROSENBAUM, Dan T, 2007. A Starting Point for Analyzing Basketball Statistics. *Journal of Quantitative Analysis in Sports*. Vol. 3, no. 3, pp. 1–22.

LEVITT, Steven D., 2004. Why are gambling markets organised so differently from financial markets? *Economic Journal*. Vol. 114, no. 495, pp. 223–246.

LOEFFELHOLZ, Bernard; BEDNAR, Earl; BAUER, Kenneth W, 2009. Predicting NBA Games Using Neural Networks. *Journal of Quantitative Analysis in Sports*. Vol. 5, no. 1, pp. 15.

MARKOWITZ, Harry, 1952. Portfolio selection. *The journal of finance*. Vol. 7, no. 1, pp. 77–91.

MILJKOVIC, Dragan; GAJIC, Ljubisa; KOVACEVIC, Aleksandar; KONJOVIC, Zora, 2010. The use of data mining for basketball matches outcomes prediction. In: Intelligent Systems and Informatics (SISY), 2010 8. *IEEE 8th International Symposium on Intelligent Systems and Informatics*. IEEE, pp. 309–312.

ORR, Genevieve B; MÜLLER, Klaus-Robert, 2003. *Neural networks: tricks of the trade*. Springer.

PAUL, Rodney J; WEINBACH, Andrew P., 2007. Does Sportsbook.com set pointspreads to maximize profits? Tests of the Levitt model of Sportsbook behavior. *The Journal of Prediction Markets*. Vol. 1, no. 3, pp. 209–218.

PAUL, Rodney J.; WEINBACH, Andrew P., 2008. Price setting in the nba gambling market: Tests of the levitt model of sportsbook behavior. *International Journal of Sport Finance*. Vol. 3, no. 3, pp. 137–145.

PAUL, Rodney J; WEINBACH, Andrew 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*. Vol. 5, no. 2, pp. 128.

PURANMALKA, Keshav (MIT), 2013. *Modelling the NBA to Make Better Predictions*. Massachusetts Institute of Technology. PhD thesis. Massachusetts Institute of Technology.

SAMPAIO, Jaime; DRINKWATER, Eric J.; LEITE, Nuno M., 2010. Effects of season period, team quality, and playing time on basketball players' game-related statistics. *European Journal of Sport Science*. Vol. 10, no. 2, pp. 141–149.

SHARPE, William F, 1994. The sharpe ratio. *The journal of portfolio management*. Vol. 21, no. 1, pp. 49–58.

SHIN, Hyun Song, 1991. Optimal betting odds against insider traders. *The Economic Journal*. Vol. 101, no. 408, pp. 1179–1185.

SIMMONS, Robert; FORREST, David, 2000. Forecasting sport: the behaviour and performance of football tipsters. *International Journal of Forecasting*. Vol. 16, no. 3, pp. 317–331.

SINHA, Shiladitya; DYER, Chris; GIMPEL, Kevin; SMITH, Noah A., 2013. Predicting the NFL using Twitter. *ECML/PKDD 2013 Workshop on Maching Learning and Data Mining for Sports Analytics*, pp. 1–11. Available from arXiv: `1310.6998`.

SIRE, Clément; REDNER, Sidney, 2009. Understanding baseball team standings and streaks. *The European Physical Journal B*. Vol. 67, no. 3, pp. 473–481.

SONG, ChiUng; BOULIER, Bryan L.; STEKLER, Herman O., 2007. The comparative accuracy of judgmental and model forecasts of American football games. *International Journal of Forecasting*. Vol. 23, no. 3, pp. 405–413.

SPANN, Martin; SKIERA, Bernd, 2009. *Journal of Forecasting*. Vol. 28, Sports forecasting: A comparison of the forecast accuracy of prediction markets, betting odds and tipsters. John Wiley & Sons, Ltd. No. 1.

SRIVASTAVA, Nitish; HINTON, Geoffrey E; KRIZHEVSKY, Alex; SUTSKEVER, Ilya; SALAKHUTDINOV, Ruslan, 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*. Vol. 15, no. 1, pp. 1929–1958.

STEKLER, H. O.; SENDOR, David; VERLANDER, Richard, 2010. Issues in sports forecasting. *International Journal of Forecasting*. Vol. 26, no. 3, pp. 606–621.

STRUMBELJ, Erik, 2014. On determining probability forecasts from betting odds. *International Journal of Forecasting*. Vol. 30, no. 4, pp. 934–943.

VERGIN, Roger, 2000. Winning streaks in sports and the misperception of momentum. *Journal of Sport Behavior*. Vol. 23, no. 2, pp. 181.

VRACAR, Petar; STRUMBELJ, Erik; KONONENKO, Igor, 2016. Modeling basketball play-by-play data. *Expert Systems with Applications*. Vol. 44, pp. 58–66.

YANG, Yuanhao, 2015. *Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics*. Available from arXiv: `arXiv:1011.1669v3`. PhD thesis.

ZIMMERMANN, Albrecht; MOORTHY, Sruthi; SHI, Zifan, 2013. Predicting college basketball match outcomes using machine learning techniques: some results and lessons learned. Available from arXiv: 1310.3607.

Appendix A

Features' descriptions

In this section, we present all the features we used. The grouping, abbreviation and description matches the source of the data – `stats.nba.com`. The grouping of the features is following:

basic the very basic statistics such as number of assist etc.

advanced more advanced statistics such as player ratings

four factors described in Kubatko et al., 2007 provide breakdown of the ratings

scoring percentages tied to players scoring

misc other statistics

Group	Feature	Description
advanced	AST_PCT	Assist Percentage is the percent of teammate's field goals that the player assisted.
advanced	AST_RATIO	Assist Ratio is the number of assists a player or team averages per 100 of their own possessions.
advanced	AST_TOV	The number of assists a player has for every turnover that player commits.
advanced	DEF_RATING	The 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.
advanced	DREB_PCT	The percentage of defensive rebounds a player or team obtains while on the court.
advanced	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.
advanced	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.
advanced	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.
advanced	OREB_PCT	The percentage of offensive rebounds a player or team obtains while on the court.
advanced	PACE	Pace is the number of possessions per 48 minutes for a player or team.
advanced	PIE	PIE is an estimate of a player's or team's contributions and impact on a game. PIE shows what % of game events did that player or team achieve.
advanced	REB_PCT	The percentage of total rebounds a player obtains while on the court.
advanced	TM_TOV_PCT	Turnover Ratio is the number of turnovers a player or team averages per 100 of their own possessions.
advanced	TS_PCT	A shooting percentage that is adjusted to include the value three pointers and free throws. The formula is: Points/ [2*(Field Goals Attempted+0.44*Free Throws Attempted)]
advanced	USG_PCT	The percentage of a team's offensive possessions that a player uses while on the court.

basic	AST	An assist occurs when a player completes a pass to a teammate that directly leads to a made field goal.
basic	BLK	A block occurs when an offensive player attempts a shot, and the defense player tips the ball, blocking their chance to score.
basic	DREB	The number of rebounds a player or team has collected while they were on defense.
basic	FG_PCT	The percentage of field goals that a player makes. The formula to determine field goal percentage is: Field Goals Made/Field Goals Attempted.
basic	FG3_PCT	The percentage of 3 point field goals that a player or team has made.
basic	FG3A	The number of 3 point field goals that a player or team has attempted.
basic	FG3M	The number of 3 point field goals that a player or team has made.
basic	FGA	The number of field goals that a player or team has attempted. This includes both 2 pointers and 3 pointers.
basic	FGM	The number of field goals that a player or team has made. This includes both 2 pointers and 3 pointers.
basic	FT_PCT	The percentage of free throws that a player or team has made.
basic	FTA	The number of free throws that a player or team has taken.
basic	FTM	The number of free throws that a player or team has successfully made.
basic	MIN	The number of minutes a player or team has played.
basic	OREB	The number of rebounds a player or team has collected while they were on offense.
basic	PF	The total number of fouls that a player or team has committed.
basic	PLUS_MINUS	The 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.
basic	PTS	The number of points a player or team has scored. A point is scored when a player makes a basket.
basic	REB	A rebound occurs when a player recovers the ball after a missed shot.
basic	STL	A steal occurs when a defensive player takes the ball from a player on offense, causing a turnover.
basic	TO	A turnover occurs when the team on offense loses the ball to the defense.

four_factors	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.
four_factors	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.
four_factors	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.
four_factors	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.
four_factors	OPP_OREB_PCT	The opponent's percentage of offensive rebounds a player or team obtains while on the court.
four_factors	OPP_TOV_PCT	Opponent's Turnover Ratio is the number of turnovers an opposing team averages per 100 of their own possessions.
four_factors	OREB_PCT	The percentage of offensive rebounds a player or team obtains while on the court.
four_factors	TM_TOV_PCT	Turnover Ratio is the number of turnovers a player or team averages per 100 of their own possessions.
misc	BLKA	The number of field goal attempts by a player or team that was blocked by the opposing team.
misc	OPP_PTS_2ND_CHANCE	The number of points an opposing team scores on a possession when the opposing team rebounds the ball on offense.
misc	OPP_PTS_FB	The number of points scored by an opposing player or team while on a fast break.
misc	OPP_PTS_OFF_TOV	The number of points scored by an opposing player or team following a turnover.
misc	OPP_PTS_PAINT	The number of points scored by an opposing player or team in the paint.
misc	PFD	The total number of fouls that a player or team has drawn on the other team.

misc	PTF_FB	The number of points scored by a player or team while on a fast break.
misc	PTS_2ND_CHANCE	The number points scored by a team on a possession that they rebound the ball on offense.
misc	PTS_OFF_TOV	The number of points scored by a player or team following an opponent's turnover.
misc	PTS_PAINT	The number of points scored by a player or team in the paint.
scoring	PCT_AST_2PM	The percentage of 2 point field goals made that are assisted by a teammate.
scoring	PCT_AST_3PM	The percentage of 3 point field goals made that are assisted by a teammate.
scoring	PCT_AST_FGM	The percentage of field goals made that are assisted by a teammate.
scoring	PCT_FGA_2PT	The percentage of field goals attempted by a player or team that are 2 pointers.
scoring	PCT_FGA_3PT	The percentage of field goals attempted by a player or team that are 3 pointers.
scoring	PCT_PTS_2PT	The percentage of points scored by a player or team that are 2 pointers.
scoring	PCT_PTS_2PT_MR	The percentage 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.
scoring	PCT_PTS_3PT	The percentage of points scored by a player or team that are 3 pointers.
scoring	PCT_PTS_FB	The percentage of points scored by a player or team that are scored while on a fast break.
scoring	PCT_PTS_FT	The percentage of points scored by a player or team that are free throws.
scoring	PCT_PTS_OFF_TOV	The percentage of points scored by a player or team that are scored after forcing an opponent's turnover.
scoring	PCT_PTS_PAINT	The percentage of points scored by a player or team that are scored in the paint.
scoring	PCT_UAST_2PM	The percentage of 2 point field goals that are not assisted by a teammate.
scoring	PCT_UAST_3PM	The percentage of 3 point field goals that are not assisted by a teammate.
scoring	PCT_UAST_FGM	The percentage of field goals that are not assisted by a teammate.
odds	ODDS	Closing odds set by bookmaker.
rest	REST	Number of days since last game.

Appendix B

Acronyms

ANN Artificial Neural Network

MPPR Mean Profit per Round

MSE Mean Squared Error

NBA National Basketball Association

NCAA National Collegiate Athletic Association

NFL National Football League

NHL National Hockey League

Appendix C

Contents of enclosed CD

- `src` ..... implementation sources
- `text` ..... the thesis text directory
  - `thesis.pdf` ..... the thesis text in PDF format
  - `thesis.tex` ..... the thesis text in LaTeX format