Master's Programme in Computer, Communication and Information Sciences

Feasibility of profitable betting on FIFA World Cup tournaments

Ville Toiviainen

Aalto University

MASTER'S
THESIS

Aalto University

MASTER'S THESIS 2018

Feasibility of profitable betting on FIFA World Cup tournaments

Ville Toiviainen

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.
Otaniemi, 19 Nov 2018

Supervisor: professor Aristides Gionis

Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

Aalto University

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Abstract

Author
Ville Toiviainen

Title
Feasibility of profitable betting on FIFA World Cup tournaments

School School of Science

Master's programme Computer, Communication and Information Sciences

Major Machine Learning and Data Mining **Code** SCI3044

Supervisor professor Aristides Gionis

Level Master's thesis **Date** 19 Nov 2018 **Pages** 46 **Language** English

Abstract

This thesis examines the feasibility of building a profitable betting system that bets on FIFA World Cups. Existing literature has researched the opportunity to bet on football games, but not in the context of FIFA World Cups. Most of the existing literature related to FIFA World Cups has examined how to predict the most probable winner or what the most probable course of the tournament might be.

In this thesis, tournaments are predicted using tree-based models and a linear model. All of the models are trained using a dataset on historical matches and a player attribute dataset obtained from the EA Sport's video game series FIFA. A more in-depth analysis of the engineered features is performed using different combinations of the features and with recursive feature elimination process. Models' output is used with two betting strategies. The first strategy uses Kelly's criterion and the second bets one unit for the most probable winner.

Two of the models are profitable in all of the tested tournaments, and most of the models can be profitable at least in two out of the three tested tournaments. However, guaranteed profits cannot be stated with full confidence since the number of games in total is relatively small. Data limitations limit the possibility to test the model with more tournaments. Using video game data with historical matches increased the accuracy in probability predictions. Models predict the probability of a home win and an away win very similarly, but differ with the prediction of a draw. Models do not prefer a draw as the most probable outcome in almost all of the matches.

Keywords football prediction, random forest, Kelly's criterion classification, recursive feature elimination, betting

urn http://urn.fi/URN:NBN:fi:aalto-201712188201

Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Ville Toiviainen

Työn nimi
Feasibility of profitable betting on FIFA World Cup tournaments

Korkeakoulu Perustieteiden korkeakoulu

Maisteriohjelma Computer, Communication and Information Sciences

Pääaine Machine Learning and Data Mining **Koodi** SCI3044

Valvoja professori Aristides Gionis

Työn laji Diplomityö **Päiväys** 19.11.2018 **Sivuja** 46 **Kieli** englanti

Tiivistelmä

Tämä diplomityö tutkii, onko mahdollista toteuttaa voittoatekevä vedonlyöntijärjestelmä jalkapallon maailmanmestaruuskilpailuja varten. Mahdollisuutta tienata vedonlyönnillä jalkapallossa on aiemmin tutkittu. Tämä tutkimus ei ole kuitenkaan ulottunut koskemaan vedonlyöntiä jalkapallon maailmanmestaruuskisoissa. Jalkapallon maailmanmestaruuskisoja koskeva aikaisempi tutkimus on keskittynyt ennustamaan potentiaalista voittajaa tai simuloimaan kuinka kisoissa todennäköisesti käy.

Tässä diplomityössä jalkapallon maailmanmestaruuskisoja ennustetaan puupohjaisilla malleilla ja lineaarisella mallilla. Kaikki käytetyt mallit koulutetaan syötteellä, joka on generoitu kahdesta eri datalähteestä: historiallisista otteludatasta sekä videopelin “EA Sport’s video game series FIFA” pelaaja-attribuuttidatasta. Syötteen selittäjien merkittävyyttä tutkitaan tarkemmin hyödyntäen erilaisia selittäjäkombinaatioita sekä toteuttamalla rekursiivinen selittäjien eliminointiprosessi. Mallien tulostetta käytetään kahdessa vedonlyöntistrategiassa. Ensimmäinen strategia lyö vetoa voittajan puolesta ja toinen strategia hyödyntää Kellyn kriteeriä vedonlyöntikohteiden valinnassa.

Kaksi mallia on voitollisia kaikissa kokeiluissa turnauksissa ja suurin osa malleista on voitollisia ainakin kahdessa kokeiluista turnauksista. Tuloksien luotettavuutta heikentää rajattu turnauksien määrä mitä testauksessa voidaan käyttää. Videopeleissä käytetyn datan hyödyntäminen ennustamisessa osoittautui hyödylliseksi. Kaikki mallit ennustavat koti- ja vierasjoukkueen voiton hyvin samankaltaisesti. Eniten eroavaisuuksia mallien välillä on tasapelin ennustamisessa. Mallit eivät pidä kovinkaan usein tasapelia kaikista todennäköisempänä lopputuloksena.

Avainsanat football prediction, random forest, Kelly’s criterion classification, recursive feature elimination, betting

urn http://urn.fi/URN:NBN:fi:aalto-201712188201

Contents

1. Introduction

The motivation behind this Master's thesis is to experiment on different ways to predict and bet a FIFA World Cup tournament. After reading this thesis, the reader should understand how football matches can be predicted and what makes predicting challenging.

Football, being one of the most popular sports in the world, attracts a lot of global attention especially during the sport's main tournaments like the FIFA World Cup. This global interest has attracted big investment banks like Goldman Sachs and UBS to predict the tournament outcome. Also, the research community has been active around this topic; mostly predicting the probabilities for winning the tournament or predicting the most probable course of the tournament [1, 2, 3]. As far as we know, the opportunity to profit from betting on FIFA World Cups has not been researched.

For many, betting the outcome of a match is an essential part of the game experience. As a result, the size of the betting industry has grown to massive proportions[1]. This must be one of the reasons why the field of economics has researched this industry, especially the efficiency of the betting market. When markets are efficient, all the available information is reflected into the prices, and there is no chance to 'beat the market'. The efficient market hypothesis in betting has been discarded by Vlastakis et al. [4] and Kuypers [5]. Nevertheless, the general assumption is that the probabilities given by the betting industry are close to the true probabilities. For this reason, models are often benchmarked against the odds given by the betting industry or the odds are used as features. [3] However, during major tournaments heavy one-sided bets, marketing purposes, and increased competition can force bookmakers' odds further from the actual probabilities. These side effects might create fruitful opportunities which

[1] `https://www.caughtoffside.com/2018/07/05/world-cup-gambling-is-the-highest-its-ever-been/`

can be harvested with a right betting strategy.

From the investor's point of view, profiting from the FIFA World Cup can be a risky opportunity. Some extreme patience is needed if after the first tournament investor is left with negative profits and a promise that the strategy will be profitable in the long run. Waiting for four years for the next tournament is a long time if you've already lost a part of your savings. Also, there is no guarantee that the strategy can keep up with the development that happens during the time between the tournaments. That's is why the betting strategy needs to be profitable in every tournament, not just in the long run.

Nowadays, many games that simulate football have a vast collection of player attribute data. This data is freely available and to our knowledge, has been successfully used once [6]. Could this player attribute data be aggregated and used to improve the prediction accuracy together with the data on historical matches?

Many different methods have been used to predict football matches. In the early work parametric models, like the Poisson distribution or the negative binomial distribution, were used to estimate the probability for a football match's score [7, 8]. In football, the favorite team has a relatively low chance to win compared to many other sports [9]. These early methods often required a single parameter which reflected the team's ability as well as possible. When skill levels were similar more features were needed to differentiate the teams' abilities. Calculating an ability value out of multiple features was difficult. More complex models were required. Logistic regression models and random forest models have performed well with multiple features in comparison to ranking methods [1, 10]. Therefore in this thesis, both linear models and random forest-based models are used in the prediction.

This thesis will answer two questions: "Is it possible to bet on the FIFA World Cup matches profitably?" and "Is it possible to predict the tournaments more accurately than the betting market predicts with freely accessible data?". The idea is not to only monetize the abnormal odds but to use the average odds from the markets in a profitable way. This way 'beating the bookie' in the long run is possible. The main contributions of this thesis are the following:

1. we develop a methodology to predict a FIFA World Cup tournament profitably;

2. we demonstrate the value of football video-game data; and

3. we demonstrate the difficulty to predict a draw by observing that the model's largest error is attained on the 'draw' class.

The rest of the thesis is structured as follows: in Section 2, we go through the existing research on the topic. Next, in Section 3, we describe the dataset and how player attributes are aggregated to team-level attributes. Then, in Section 4, we describe the prediction models, betting strategies and the tournament simulation process. In Section 5, results from the simulation experiments are stated. Finally, we conclude and discuss future work in Section 6.

2. Literature Review

Modeling football scores has gained some interest in scientific communities. One of the first to mention that football score distribution resembles a Poisson distribution was Maroney [7] in his book 'Facts from figures'. Although the Poisson distribution fits the scores well, Maroney came into a conclusion that the negative binomial distribution fits the scores better. But was this enough? In 1997 Dixon stated [8] "It is not difficult to predict fairly accurately which teams are likely to be successful, but the development of models that have a sufficiently high resolution to exploit this long run predictive capability for individual matches is substantially more difficult." Chance dominates the game in football [11]. However, Maher's [12] idea to include the team's quality into the model was promising. He used an independent Poisson distribution to model the score based on the team's previous performance. Ten years later Dixon used a similar approach in his model and was able to achieve positive returns from betting [8].

To differentiate the teams the key is to describe the team's abilities as well as possible. Elo rating, initially developed for assessing the strength of chess players [13], has gained popularity in football prediction. Leitner et al.[3] simulated UEFA Euro 2008 football tournament using ratings of abilities such as the Elo rating and the FIFA/Coca-Cola World ranking with bookmaker's odds. They showed that Elo and FIFA/Coca-Cola World ranking are suitable attributes for match outcome prediction, but not as good as the bookmaker's odds are. The FIFA/Coca-Cola World[1] rating had a higher Spearman correlation with the tournament outcome than the Elo rating. This means that it would be a better metric if the corresponding winning probabilities could be computed. [3] The results of Lasek et al. [15] are contrary to this finding. When they compared rating methods,

[1] FIFA has changed the ranking algorithm three times since it was introduced. The ranking algorithm will be modified after FIFA World Cup 2018 to resemble Elo rating [14].

the Elo rating described the team's ability better than the FIFA/Coca-Cola rating. Hvattum et al. [16] concluded that the single rating difference is a highly significant predictor of the match outcomes, which justifies the increasing interest in using Elo rating to describe the team's ability. In their study, they used logistic regression but were not able to achieve as good results with Elo rating as they were with market odds. More data is needed with Elo rating to match the market's accuracy.

When the number of used features has grown, tree-based models have gained popularity in football prediction. Groll et al.[1] predicted the FIFA World Cup 2018 match outcomes using three models: a random forest model, a regression model, and a ranking model. From these models, the random forest model performed the best in classification and even outperformed the bookmakers. For features, they used economic factors such as GDP per capita, sportive factors like FIFA/Coca-Cola rating, factors that described the team's structure such as average age and factors that described the team's coach. In another study, where the outcome of a match in the Turkish super league was predicted, a random forest model was the best performing model. It outperformed a support vector machine model and a bagging REP tree model in prediction accuracy. In this study feature set selection was performed and a limited feature set performed better than the full feature set. [10]

One clear motivation behind the match outcome prediction is the possibility to earn profit from betting. If the markets are efficient, an investor cannot consistently "beat the market" [17]. This assumption is known as the *Efficient market hypothesis* (EMH) and it comes from the field of economics. Since American football's betting market resembles Wall Street, Pankoff [18] reasoned that American football betting should satisfy the EMH as well. He concluded that the systematic market errors are not large enough to be profitable to bettors. Later the efficient market hypothesis has been revoked in finance [19], and this might also be the case in betting [20, 17]. However, Goddard et al. [20] state that there is some evidence that the inefficiencies in the bookmakers' prices have diminished over time. Kuypers [5] analysis on how bookmakers calculate their odds supports this claim of price inefficiency. If bookmakers want to increase their profits while keeping their over-roundness competitive, they need to set the odds further from the market efficient odds. Also, marketing purposes and heavy one-sided bets can lead to inefficient odds. In the case of heavy one-sided bets, bookmakers expose themselves to a higher risk exposure if the match's outcome is the one that is betted very heavily. With a proper betting strategy and a model that can beat bookmakers generating profit from football betting should be possible.

One well-known betting strategy is Kelly's criterion which has also been used outside of sports betting by famous investors like Warren Buffet and Bill Cross[2]. This strategy aims to maximize the logarithm of wealth by placing an optimal fraction of the total bankroll for a bet based on the probabilities [21]. MacLean et al.[22] looked at the benefits of using a fraction of the optimal fraction given by Kelly criterion. They concluded that there is a tradeoff between a small decrease in the growth rate against the increased chance of doubling your fortune before it is halved. This is a crucial finding since World Cup tournaments have a very limited number of games and for a strategy to work, it needs to be successful from the very first games onwards.

Data collection around football has increased. Betting agencies require more data to improve their prediction accuracy and many professional bettors, football scouts, and coaches are also interested in this data. More expressive datasets are valuable and hard to collect which makes it challenging to find a clean and comprehensive dataset for free. One interesting source for open data is video games. To simulate the game accurately a lot of data is required. EA Sport's video game series FIFA has included a comprehensive player attribute dataset from the year 2007 onwards. Shin et al.[6] investigated this dataset and showed that it describes the players well. When they compressed the data into 3D space, they were able to see clear separations between attackers, defensive players, and goalkeepers. With this "virtual data" (data from the video game) they were able to make more accurate predictions compared to predictions that used "real data", which contained different statistics from historical matches.

[2] http://www.financial-math.org/blog/2013/10/two-tales-of-the-kelly-formula/

3. Data

In this section, we briefly describe the main datasets: the dataset of international football matches and the dataset of player attributes from the EA Sport's video game series FIFA. We use both of the datasets to generate new data points. This process will be discussed in this chapter.

3.1 International football match dataset

We use results from all international football matches from November 11th, 1872 to June 6th, 2018 as the primary source of data. This dataset is provided by Kaggle [23] and contains match scores, tournament types, dates and more attributes that are not used in this thesis.

We don not use this dataset directly in predictions. Instead, we extract useful features from this data. Generated features from this dataset are named as *general features*.

Elo rating describes the team's quality. This metric is calculated based on previous games with the following formula

$$\mathrm{R_n} = \mathrm{R_O} + \mathrm{K} \times (\mathrm{W} - \mathrm{W_e}) \tag{3.1}$$

where $\mathrm{R_n}$ is the new Elo rating and $\mathrm{R_O}$ is the old (pre-match) rating. Paramater K is the weight constant for the tournament played. Values for K are listed in Table 3.1. These values are based on the values used in the website World Football Elo rating [1]. Team qualities can vary a lot within a confederation and for that reasons a lowering coefficient of 0.85 ($50 \times 0.85 = 42.5$) is used for AFC, CONCACAF, OFC and CAF. The selected K for the tournament type is multiplied based on the score. It is multiplied by 1.5 if the game is won by two goals, by

[1] https://www.eloratings.net/about

Table 3.1. The weight constant K for the tournaments.

Tournament	K
FIFA World Cup	60
Confederations Cup, Copa America, UEFA Euro, FIFA World Cup qualification	50
AFC Asian Cup, Gold Cup, CONCACAF Championship, Oceania Nations Cup, African Cup of Nations	42.5
African Cup of Nations qualification, AFC Asian Cup qualification, UEFA Euro qualification, CONCACAF Championship qualification, Oceania Nations Cup qualification, AFC Challenge Cup, AFC Challenge Cup qualification, Gold Cup qualification	40

1.75 if the game is won by three goals, and by $1 + (3/4 + (N - 3)/8)$ if the game is won by four or more goals, where N is the goal difference. Parameter W is the outcome of the game: 1 for a win, 0.5 for a draw, and 0 for a loss. The win expectancy $\mathrm{W_e}$ is calculated as

$$W_e = 1/\left(10^{(-dr/400)} + 1\right), \tag{3.2}$$

where dr is the difference in rating.

Goal average describes how many goals a team has scored on average in the previous games within the timespan of four years. This metric is calculated for the home team and the away team. Features' names are *home_goal_mean* and *away_goal_mean*.

Goal average difference is the difference between the home team's *goal average* and the away team's *goal average*. Feature's name is *goal_diff_with_away*.

Goal average with the opponent describes how many goals the team has scored on average against the opponent. The time lag is four years, and the metric is calculated for both teams. The features' names are *home_goals_with_away* and *away_goals_with_home*.

3.2 EA Sport's video game series FIFA's player attributes

EA Sport's video game series FIFA describes every player in the game with several different attributes. These attributes are first collected by EA's data reviewers who are a group of coaches, professional scouts, and season ticket holders. EA editors give the final value based on the reviewers' answers [24]. From here onwards EA Sport's video game series FIFA's player attributes are called just player attributes.

Player attributes are available from August 30th, 2006 onwards [25]. We collected this data ourselves since it was not available as a single dataset. All of the player attributes have a value in the range of 0-99. When two players are compared, a lower value means that the player's capability regarding that attribute is not as good as the other player's. From all possible attributes, we have used 24 player attributes that were available from the beginning. These attributes are listed here with a short description that is taken from Fifplay [24].

Goalkeeper:

Diving: determines a player's ability to dive as a goalkeeper.

Handling: determines a player's ability to handle the ball and hold onto it using their hands as a goalkeeper.

Kicking: determines a player's ability to kick the ball as a goalkeeper.

Positioning: determines that how well a player is able to perform the positioning on the field as a player or on the goal line as a goalkeeper.

Reflexes: determines a player's ability and speed to react (reflex) for catching/saving the ball as a goalkeeper.

Mental:

Aggression: determines the aggression level of a player on pushing, pulling and tackling.

Heading accuracy: determines a player's accuracy when using the head to pass, shoot or clear the ball.

Marking: determines a player's capability to mark an opposition player or players to prevent them from taking control of the ball.

Physical:

Acceleration: determines the increment of a player's running speed (sprint speed) on the pitch. The acceleration rate specifies how fast a player can reach their maximum sprint speed.

Reactions: determines the acting speed of a player in response to the situations happening around them.

Shot Power: determines the strength of a player's shootings.

Sprint Speed: determines the speed rate of a player's sprinting (running).

Stamina: determines a player's ability to sustain prolonged physical or mental effort in a match.

Strength: determines the quality or state of being physically strong of a player.

Skill:

Ball control: determines the ability of a player to control the ball on the pitch.

Crossing: determines the accuracy and the quality of a player's crosses.

Dribbling: determines a player's ability to carry the ball and past an opponent while being in control.

Finishing: determines the ability of a player to score (ability for finishing - How well they can finish an opportunity with a score).

Free kick accuracy: determines a player's accuracy for taking free kicks.

Long passing: determines a player's accuracy for the long and aerial passes.

Long Shots: determines a player's accuracy for the shots taking from long distances.

Penalties: determine a player's accuracy for the shots taking from the penalty kicks.

Short passing: determines a player's accuracy for the short passes.

Standing tackle: determines the ability of a player to performing standing tackle.

3.3 Aggregating team-level attributes

In this section, we explain how player attributes are combined to team-level attributes to describe a football team based on its players' capabilities.

To describe the team in a general level, we have used the average value from the team's 23 best players. These attributes are: *overall rating, potential, age, height, weight, international reputation* and *weak foot*. As an extra attribute, we calculated the average age from the top 11 players. The idea behind this attribute is to get the average age of the presumed starting lineup.

The other attributes are described by a subsection of the team's 23 best players. We have used our knowledge and intuition on football to select the sizes of the subsections. The goalkeeper attributes are calculated based on the team's two best values for that attributes. The average value for a skill required in a set piece situation, like the attributes *free kick accuracy* and *penalties* for example, is calculated based on the top three ratings, since in most cases a small subset of players handles these situations. Also, *sprint speed* and *marking* are only calculated based on the top three values. *Strength* and *stamina* are calculated using the 10 best ratings. Other values are calculated using the five best ratings for each attribute.

In cases where the team does not have enough players to calculate the attribute's value as mentioned in the previous part we have used this formula

$$\frac{1}{N}\sum_{i=1}^{N} x_i \cdot \max\{N/K, C\}. \tag{3.3}$$

If N (the number of all available ratings for the attribute) is smaller than K (the required number of ratings), the average is multiplied with a coefficient that has an integer value in the inclusive range from C to 1. We have set the value of C to 0.9.

Feature set containing the aggregated player attributes is named as *player features*.

3.4 Historical odds

We have collected odds for the FIFA World Cup 2018, 2014 and 2010 from Odds Portal [26]. Odds Portal has a collection of odds offered by multiple betting sites. For every match, we have used the average value from the available odds. Kuypers [5] mentions that football odds are mostly fixed. Based on values that Odds Portal offer, this is not the case anymore. Many odds have changed from the initial opening odd before the match start. We have used the latest value for every odd since the data is easier to collect that way.

Table 3.2. Feature set descriptions

Feature set's name	Description
Player features	FIFA player attributes only
All Features	General features and Player features
General Features	All excluding Player features

3.5 Features used in prediction

Features listed above are calculated for each team, but all of them are not used directly in training. For most of the features, the difference between the home team's value and the away team's value is used as the final value for that feature. For example to get the Elo difference the away team's Elo is subtracted from the home team's Elo. This process is done for all of the player attributes. From *general features* difference is only used for Elo. The main advantage of this method is that it reduces the number of features and makes the link between the home team's value and the away team's value explicit for the model. One limitation of this process is that a game between weak teams and a game between strong teams can look the same if the feature vector is only inspected.

From the available features, three feature sets are created: *all features*, *general features*, and *player features*. Table 3.2 shows the description for each of the feature sets.

4. Methodology

In the context of machine learning, supervised learning is the task of learning the relationship between input features and the target value. The structure that describes this relationship is called a model. In most cases, these models are used to predict the target value based on new input features. There are two types of models: *regression models* and *classifier models*. If the target value is in a real-valued domain, the model is called a *regression model*. *Classifier models* are used to map the input features to predefined classes. [27]

4.1 Decision trees

A decision tree is one of the most popular model types used in classification problems. A decision tree is a rooted tree, which means that all of the nodes, except the *root node*, have exactly one incoming edge. Nodes that have outgoing edges are called *internal nodes* and the nodes that have only incoming edges are called *leaf nodes*. Internal nodes in a decision tree split the instance space into two or more subspaces according to a certain discrete function of the input's feature values. Usually, a split is done based on a single feature from the whole feature vector. A single class value is assigned for the *leaf nodes*. When a new input is given the tree is navigated from the *root node* to a *leaf node* which determinates the predicted class label. In regression, these target values can take continuous values. [27]

Decision trees have many benefits and are very useful "off-the-shelf" predictors. Outliers in the dataset or many irrelevant predictors are not problematic for the trees. Scaling or any other general transformation can be done to the input space since trees are invariant under transformation of the individual predictors. [28] Decision trees have good interpretability if the trees are small.

4.2 Bootstrap aggregating

One main disadvantage of decision trees is the low prediction accuracy [28]. Decision trees can express the training data well but have a high variance, which means that the prediction accuracy for unseen data is often weak.

Bootstrap aggregating, also called bagging, is a way to improve the prediction accuracy of decision trees by averaging. In bagging the average is taken over the output of multiple estimators:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b^*(x), \tag{4.1}$$

where B is the number of estimators and $\hat{f}_b^*(x)$ is a single estimator. This reduces the high variance of a single tree and makes predictions more accurate.

Bootstrap in bagging means that in the training of a single tree a random sample with replacement is taken from the original sample. Samples used in training come from the same distribution, meaning that the trees are identically distributed (i.d.). Sampling with replacement combined with deep trees that have less bias ensures that the variance reduction achieved in bagging comes only at the expense of a small increase in bias and loss of interpretability. The loss of interpretability cannot be avoided since a single tree cannot be used anymore for reasoning. Trees in bagging are only identically distributed. The missing independent property means that the trees in the forest can have a pairwise correlation. Pairwise correlation is common in cases where input data has one strong predictor which often leads to a situation where all of the trees are split similarly. [28]

4.3 Random forest

Amit and Geman's [29] idea of random feature selection inspired Breiman to use bagging in tandem with random feature selection. With this random feature selection correlation between the trees can be reduced since the generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them [30]. Breiman was first to use the name *Random Forest* for algorithms that use bagging and random feature selection with tree predictors [30]. Step-by-step instruction from [28] for random forest algorithm are listed in Algorithm 1.

The Main usecases for random forest are *classification* and *regression*.

Algorithm 1: Random Forest for Regression or Classification.

```
1. For $b = 1$ to $B$:
    (a) Draw a bootstrap sample $\boldsymbol{Z}^*$ of size $N$ from the training data.
    (b) Grow a random-forest tree $T_b$ to the bootstrapped data, by recursively repeating the following steps for each leaf node of the tree, until the minimum node size $n_{min}$ is reached.
        i. Select $m$ variables at random from the $p$ variables.
        ii. Pick the best variable/split-point among the $m$.
        iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point $x$:
Regression: $\hat{f}_{\mathrm{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$
Classification: Let $\hat{C}_b(x)$ be the class prediction of the $b$th random forest tree. Then
$\hat{C}_{rf}^B(x)$ = majority vote $\left\{\hat{C}_b(x)\right\}_1^B$
```

4.3.1 Random forest hyperparameter selection

Many machine learning algorithms have parameters that are not optimized within the algorithm itself. These parameters are called hyperparameters. Optimizing these hyperparameters is one way to improve the model's performance since optimal parameters are often problem specific. Random forest is no exception, even though in many cases its performance is relatively decent with the default parameters [31].

In this hyperparameter selection process three important hyperparameters are optimized: *number of candidate predictors*, *minimum samples at a leaf node* and *maximum depth of a tree*.

The *Number of candidate predictors*, denoted as K, is one of the key hyperparameters to control the correlation between forest's trees [31]. In cases where there are many or only a few relevant predictor variables, choosing the value of K can have a high influence on the results. For example in the case of minuscule K with a dataset that has only a small number of important predictors most of the trees are built without the important predictor and have low prediction accuracy. [32] Often best values for K are $\sqrt{M}$ and $\log_2(M)$, where M is the number of predictor variables [32].

Segal [33] showed that increasing the number of noise variables lead to a higher optimal leaf node size. For this reason, we chose to optimize

Table 4.1. Optimized hyperparameters and the tested values.

Hyperparameter	Values
number of candidate predictors	$\sqrt{M}$, $\log_2(M)$
minimum samples at a leaf node	1, 3, 5, 10, 15
maximum depth of the tree	3, 5, 8, 12, None

the *minimum samples at a leaf node*. Reasonable default values for this hyperparameter are 1 for classification and 5 for regression [31]. Last optimized hyperparameter - *maximum depth of a tree* controls the depth of the tree. When the tree is forced to be shallow its reasoning logic is less complex. In some cases, decreasing the value of *maximum depth of a tree* might have a similar effect as increasing the value of *minimum samples at a leaf node*. This happens because in both cases the sample count in a leaf node increases.

We use two metrics to evaluate models performance. Accuracy

$$\frac{1}{N}\sum_{i=1}^{N} 1\left(\hat{y}_i = y_i\right), \tag{4.2}$$

where N is the number of observations, y is the correct class, and $\hat{y}$ the predicted class, is used to see how many observations are classified correctly. The second metric is cross entropy loss

$$-\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log\left(p_{i,j}\right), \tag{4.3}$$

where N is the number of observations, M is the number of classes, y is the binary indicator for the correct class and p is the probability for that class [34]. Cross entropy loss is used to evaluate how good model's probability estimates are.

For every model, the optimal hyperparameters are searched using the grid search algorithm. Cross-validation of 5 folds is used. Best hyperparameter combination is chosen by looking first at the highest average accuracy and then by the lowest average cross entropy loss.

4.4 Logistic Regression

Linear models like linear regression and logistic regression, which are the most well-known methods, are widely used in statistical modeling. Hosmer[35] defines the difference between these models well: "What distinguishes a logistic regression model from the linear regression model is that the outcome variable in logistic regression is *binary* or *dichotomous*. This difference between logistic and linear regression is reflected both in the choice of parametric model and in the assumptions. Once this difference is accounted for, the methods employed in an analysis using logistic regression follow the same general principles used in linear regression."

The output of a logistic regression model is a probability estimate for a class. This conditional probability for a class is denoted by $P(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$. The name of the model comes from the fact that the logistic function turns log-odds to this conditional probability. Sigmoid is the logistic function and the log-odds for the model are given by the equation

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p. \tag{4.4}$$

This combined with the logistic function makes it a logistic regression model

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}}. \tag{4.5}$$

To get the estimates for $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ maximum likelihood estimation is often used. The idea in maximum likelihood estimation is to maximize the likelihood function. Likelihood function equations are

$$\sum_{i=1}^{n} [y_i - \pi(\mathbf{x}_i)] = 0 \tag{4.6}$$

and

$$\sum_{i=1}^{n} x_{ij} [y_i - \pi(\mathbf{x}_i)] = 0 \tag{4.7}$$

for $j = 1, 2, .., p$. [35] No closed-form solution exists for logistic regression. Estimation is solved by iterative algorithms like Newton's method.

4.5 Poisson Distribution

The Poisson distribution is one of the most important distributions in statistics. It is named after the French mathematician Simèon Denis Poisson (1781–1840) who was the first to present this distribution. The Poisson distribution is given by

$$Po(r; \mu) = \frac{\mu^r e^{-\mu}}{r!}, \tag{4.8}$$

where r is the number of events and μ is the average number of events per interval. The Poisson distribution gives the probability for finding exactly r events in a given length of time if the events occur independently at a constant rate of μ. [36] The poisson distribution has been discovered to give a reasonably accurate description of football scores [12].

4.6 Prediction models

4.6.1 Outcome Model

A football match can have three different outcomes: *home win*, *draw* or *away win*. Using these three outcomes as classes, the random forest classifier can be used to predict the probabilities for each possible outcome. *Outcome model* implements this idea and its output, the outcome probabilities, are used directly as the estimated probabilities for each outcome.

4.6.2 Score Model

The *Score model* is highly influenced by Groll et al.'s [1] model that used random forest regression and a Poisson distribution to simulate each match in the World Cup 2018. They used random forest regression to get the expected number of goals for both of the teams. To simulate the tournament correctly, they needed to estimate probabilities for different results for each match. To overcome this issue they used the expected number of goals from the random forest regression as an intensity parameter μ in a Poisson distribution $Po(\mu)$ to draw a random number of goals for both of the teams. Both teams had their own intensity value which meant that both of the Poisson distributions were independent but conditional on the features.

Since we need the probabilities for each outcome, sampling just one possible result for a match is not enough. For this reason, for each team, the probabilities of scoring a number of goals between 0 to 10 is calculated from the Poisson distribution's probability mass function. As an end result both teams have their own probabilities for scoring goals between 0 and 10 in the form of a probability vector $score\_prob = (h_1, h_2, \ldots, h_{n-1}, h_n)$, where N is 10. The outer product of these two score probability vectors, called the *goal matrix*, has the probability estimates for each unique result as illustrated in Figure 4.1. Instead of probabilities, this figure has the score as cell value to clarify the matrix's structure. Probabilities for match

$$\begin{bmatrix} 0-0 & 0-1 & \cdots & 0-N \\ 1-0 & 1-1 & \cdots & 1-N \\ \vdots & \vdots & \ddots & \vdots \\ N-0 & N-1 & \cdots & N-N \end{bmatrix}$$

Figure 4.1. N by N Goal Matrix where row values are home team's score and column values are away team's scores.

outcome are simple sums from this goal matrix, since $\sum_{i=1}\sum_{j=1} p_{ij} = 1$. The sum of the lower triangular entries is the probability of the home team winning, the sum of the diagonal entries is the probability of a draw, and the sum of the upper triangular entries is the probability of the away team winning.

4.6.3 One-vs-rest model

The One-vs-rest model (OVR model) is a model that splits multiclass classifier (three or more classes) into multiple binary classifiers. Each of the classes has its own binary classifier. A multiclass classifier is then formed from the trained binary classifiers. The requirement is that the output of a binary classifier can be used with the other outputs to form the multiclass classifier. Often this means that the outputs are probabilities.

With the *OVR model* we train a single binary classifier for each outcome. For example, a binary classifier that predicts the probability for home team's win will label all true classes (the matches where the home team won) as 1 and the rest of the matches as 0. The probability of the true class $P(c_i = 1|x)$ is taken from each binary classifier i. To form the probability distribution for the match's outcome these probabilities are normalized [37]. For example, the probability of home team's win is calculated as

$$\frac{P(c_{home\_win}|x)}{P(c_{home\_win}|x) + P(c_{draw}|x) + P(c_{away\_win}|x)}. \tag{4.9}$$

One advantage of a binary classifier is that the probability estimates from the model can be calibrated. This means that model's probability estimates can be adjusted after the prediction to remove the possible bias. For example, boosted trees rarely give probability estimates close to 0 or 1. This is not the case with random forest. With random forest benefiting from calibration is more problem specific. [38] Two well-known calibration methods exists for binary classifiers. Platt scaling uses a sigmoid function to calibrate binary classifier probabilities [39]. Probability estimates from the vanilla model are passed through a fitted sigmoid function

$$P(c=1|x)=\frac{1}{1+\exp(Af(x)+B)} \tag{4.10}$$

to get the calibrated estimates. Here $f(x)$ is the output from the binary classifier and A and B are parameters that are fitted using the maximum likelihood estimation. With Platt scaling normally the assumption is that the non-calibrated probabilities tend to act like a sigmoid function. If this is not the case often the other calibration method, isotonic regression, is used. With isotonic regression the function

$$c_i = m\left(f_i\right) + \epsilon_i, \tag{4.11}$$

where m is an isotonic function, is used to calibrate the probabilities. Optimal function m is problem specific and learned by minimizing

$$\hat{m} = \arg\min_z \sum \left(c_i - z\left(f_i\right)\right)^2. \text{ [37]} \tag{4.12}$$

4.6.4 Linear Model

The linear model uses logistic regression to output the probabilities for the outcomes. Since the problem is a multiclass classification problem, logistic regression cannot be used directly. For this reason, the *Linear model* is a combination of three one-vs-rest logistic regression classifiers. Each of these classifiers gives a probability estimate for a single class, and these classifiers are fitted independently from each other. Probability estimates from the multiclass model are normalized probability estimates from the underlying models. Normalization is done the same way as it is done with the *OVR model*.

4.6.5 Bookmaker's model

The bookmaker's model is a benchmark model which predictions are implied probabilities calculated from the market odds. A single odd cannot be used as a probability directly unless the commission taken by the bookmaker is zero. To take the commission into account, the sum of inverse odds is calculated

$$k = \sum_{i=1}^{N} \frac{1}{odd_i} \tag{4.13}$$

where N is the number of odds. In the case of football it is 3, since the possible outcomes are a home win, a draw or a away win. When k is known the implied probability can be calculated as

$$\textit{implied probability} = \frac{1}{k \cdot odd_i}. \tag{4.14}$$

If there is no commission $k = 1$, but normally $k > 1$. If for example $k = 1.04$ it means that the commission is 4%.

4.7 Betting strategies

One way to validate a model that predicts the outcome of a football match is to see if betting according to the model's predictions is profitable in the long run. Betting market odds provide a good benchmark since bookmakers have a financial interest to provide as accurate models as possible. We have used two betting strategies to validate our models' performance in FIFA World Cups.

4.7.1 Unit strategy

The first strategy, *unit strategy*, is the simplest strategy. This strategy is named as *unit strategy* since in each game one unit is placed for the predicted winner. In most of the cases, positive returns from this strategy require the model to be more accurate than the bookmaker's model. If a model and bookmaker's model predict the outcomes equally the expected return is negative since bookmakers include their commission into the odds. If the bookmaker's model predicts that the probability of a home win is 25% and bookmaker's commission is 3% the final value for the odd is calculated as $1/(0.25 + 0.03) = 3.57$. Being less accurate but profitable requires a model to predict many outcomes with low probability correctly. Unit strategy's main weakness is that it doesn't use all of the data available in betting. It doesn't use the probabilities to change the bet size or targets. For this reason, we experiment with another betting strategy.

4.7.2 Kelly strategy

The Second betting strategy is named as *Kelly strategy* by the inventor Kelly [21]. Many times it is also called Kelly's criterion. In his famous paper [21] Kelly consider how to choose the optimal bet size according

Figure 4.2. The optimal bet size formula for Kelly. *P*s are outcome probabilities, *o*s are net odds and *f*s are optimal fractions

$$\begin{aligned}\max_{f_1,f_2,f_3} p_1\log(1+o_1f_1-f_2-f_3)+p_2\log(1+o_2f_2-f_1-f_3)\\ +p_3\log(1+o_3f_3-f_1-f_2)\end{aligned} \tag{4.16}$$

$$\begin{aligned}\text{subject to: } & f_1+f_2+f_3\leq 1\\ & 0\leq f_1,f_2,f_3\leq 1\end{aligned}$$

to the available odds to maximize the logarithm of wealth. This way the typical questions of a gambler — "how much to bet" and "what are the favorable betting targets" — can be answered. In the simplest form, when calculating optimal fraction only for a single outcome, the optimal fraction of the bet is calculated as

$$f^* = \frac{p(b+1)-1}{b}, \tag{4.15}$$

where b is the net odd and p is the probability given by the model. The gambler's bankroll is multiplied with the optimal fraction f^* to get the size of the bet — if the fraction is positive. If the probabilities given by the model are close to the true probabilities, the gambler should end up exponentially increasing her wealth in the long run. To utilize the whole potential of the model a more complex function is used to calculate the optimal fraction. To get the optimal fractions for bets 1, X and 2 (home win, draw and away win), the simple formula of Kelly's criterion needs to be extended to include all the odds, probabilities and fractions as in the equation 4.2.

It is common to bet only a fraction of the optimal bet size since the short-term risk of losing a big proportion of the bankroll is high [40]. For this reason, only 30% of the suggested bet size is used.

4.8 Recursive feature elimination

The recursive feature elimination (RFE) is an algorithm that can be used to see what features matter the most. In every iteration round, RFE measures the model's performance and uses feature importance values from the model to see what feature was the least important. This feature is removed from the feature set used in the next round. The algorithm stops when the feature set is empty. [41] Random forest model includes feature importance values and can be used with RFE. To guarantee that the variance in a single fitting doesn't rule out an important feature a model is trained 100 times with a different subset of the dataset in each elimination round. The feature that performs the worst on average will be removed from the feature set. The average accuracy and the average log loss values from each elimination rounds are stored.

4.9 Tournaments simulation process

The quality of the model is measured using the results from the tournament predictions. The tournament is simulated, and the model's performance is measured for each game. The tournament simulation outputs four metrics: accuracy, log loss, unit strategy's profit, and Kelly strategy's profit. These metrics can be used to validate the model's performance. Since the same model can end up into a different local optimum between subsequent training sessions, the simulation is run for ten times per tournament to get the average performance and the standard deviation.

Simulation is done for the World Cup 2018, 2014 and 2010. In every simulation, the model uses the optimal hyperparameters that have been searched before the simulation using the whole dataset. The data used in the training span a period from the start date August 30th, 2006 to an end date, which depends on the tournament. In the World Cup 2010 the end date is June 11th, 2010, in the World Cup 2014 it is June 12th, 2014 and in the World Cup 2018 it is June 4th, 2018, which is the last date for the international matches dataset. The model is retrained before every simulation.

Each tournament is simulated according to the official tournament diagram. For each game, the most recent values for features are used. This means that the feature values are not static throughout the tournament. Elo rating is updated after the match for both teams using the real outcome of the game, not the predicted one.

This process, mentioned above, is run for every model and feature set combination.

4.10 Implementation details

In our experiments we use scikit learn's algorithm library [42] for the random forest models and for the *Linear model*. *Linear model* uses scikit learn's *newton-cg* method to optimize the weights with L2 regularization. Scikit learn's inverse of regularization strength C is set to 0.001 based on grid search results. All of the parameters that do not use the scikit learn's default values are mentioned in this thesis. Scipy's implementation of Sequential Least Squares Programming is used to calculate the optimal fractions for Kelly strategy.

5. Results

In this part of the thesis, we report the results from the experiments. Since there are no direct benchmarks, it is vital to know how to view the results. In the simplest form, successful results can be stated when betting is profitable. Unfortunately, this is a very narrow perspective, and for that reason, we will dig deeper to understand why some of the models generate profitable returns. Also, it is important to know what downsides these models and strategies might have. In short, Results section will answer to these questions 1) "Is it possible to bet on a FIFA World tournament profitably?", 2) "What are the main differences between the models?", 3) "How using different feature sets affect the results?", and 4) "How are the individual matches betted and how betting differs between tournaments?" During early experiments *OVR model* with Platt scaling outperformed the vanilla model and the isotonic scaling. In the results *OVR model* uses only Platt scaling.

5.1 Tournament simulation results

Tournament simulation results for each of the models are listed in Tables 5.2, 5.3, 5.4 and 5.5. The tables include metrics for accuracy, log loss, unit strategy's profit and Kelly strategy's profit. Listed values are averages from 10 individual simulations and include standard deviations. All metrics are listed for every feature set used in training. Abbreviations are used for feature set names: AF means *all features*, GF means *general features* and PF means *player features*. Description for each of the feature set is listed in Table 3.2.

To answer the question "Is it possible to bet on a FIFA World tournament profitably?" it is important to see if any of the models can generate profit. Preferably, a model should be profitable in each of the tournaments, since

Table 5.1. Tournament characteristics. *Underdog victory* is a case where the winning team has higher odds for winning than the other team. All of the tournaments have 64 games in total.

Characteristic	**WC 2018**	**WC 2014**	**WC 2010**
Home wins	25	28	24
Draws	14	13	16
Away wins	25	23	24
Underdog victory	14	15	14

investors prefer steady profits. Also, if the model, trained using different data periods, can continuously keep its performance it is harder to say that the model was just lucky all the time.

Unit strategy's profits for *Outcome mode*, *Score model* and *OVR model* are positive in every tournament, but not with all of the feature sets. Good profits from the unit strategy and high accuracy go hand in hand, which is no surprise. However, high accuracy is not an indicator for good Kelly profits, which is a slight of a surprise. When estimated probabilities are closer to the true ones than the market's probabilities, Kelly criterion is a working strategy. Since Kelly criterion aims to maximize the logarithm of expected wealth, it utilizes current bank more efficiently. The unit strategy uses just a static bet size of one unit per game and is not able to maximize all of the good opportunities. From the tested models only *Outcome model*, trained with player features, is profitable with both of the strategies.

Bookmaker's accuracy on the World cup 2018 and the World cup 2014 was 56.25% (36 games correctly predicted) and 51.56 % (33 games correctly predicted) on the World cup 2010. Most of the models can outperform the bookmaker's model's accuracy, which is a promising result.

But which of the feature set and model combinations is the best? This question is tricky to answer. If being profitable in every tournament with a high profit is preferred then *Score model*, trained with all features, might be the right choice. Whereas highest unit strategy's profit is achieved with *OVR model* trained with player features. Using only these metrics to validate if a model is good enough to be trusted in the next World cup 2022 is too risky. Being profitable in 192 games (64 games per a tournament) is a good start.

5.2 How models differ in game-level predictions?

To understand the total profits better, it is essential to know how the models predict individual games. Analyzing all of the tournaments at

Table 5.2. Simualtion results for *Outcome model*.

Metric		**WC 2018**	**WC 2014**	**WC 2010**	Mean
Accuracy	AF	57.34% ± 0.72	**60.94% ± 1.71**	54.84% ± 1.09	57.71
	GF	52.03% ± 1.22	56.56% ± 0.62	**55.47% ± 1.05**	54.69
	PF	**60.47% ± 1.41**	59.69% ± 0.94	54.22% ± 2.22	**58.13**
Log Loss	AF	0.9673 ± 0.0037	**0.9362 ± 0.0047**	0.9922 ± 0.0075	**0.9652**
	GF	1.0122 ± 0.0045	0.9549 ± 0.0053	**0.9676 ± 0.0052**	0.9782
	PF	**0.9406 ± 0.0026**	0.9503 ± 0.0024	1.0118 ± 0.0045	0.9676
Unit profit	AF	6.39% ± 1.96	**15.65% ± 5.21**	2.93% ± 4.43	8.32
	GF	-3.48% ± 3.4	5.17% ± 1.52	**5.04% ± 3.67**	2.24
	PF	**18.38% ± 4.26**	12.72% ± 2.22	3.32% ± 8.33	**11.47**
Kelly profit	AF	-10.58% ± 7.01	**18.55% ± 10.13**	23.26% ± 18.96	10.41
	GF	-47.32% ± 4.1	2.86% ± 8.15	**48.9% ± 16.64**	1.48
	PF	**46.63% ± 8.79**	12.26% ± 6.57	2.75% ± 9.05	**20.55**

Table 5.3. Simulation results for *Score model*.

Metric		**WC 2018**	**WC 2014**	**WC 2010**	Mean
Accuracy	AF	**59.38% ± 0.0**	58.13% ± 0.62	52.81% ± 0.62	56.77
	GF	52.03% ± 1.0	**59.38% ± 0.0**	50.94% ± 0.77	54.12
	PF	58.44% ± 1.74	61.09% ± 0.84	**53.91% ± 0.78**	**57.81**
Log Loss	AF	**0.9508 ± 0.0016**	0.935 ± 0.0005	0.9788 ± 0.0016	0.9549
	GF	0.9831 ± 0.0012	**0.9133 ± 0.0011**	**0.9522 ± 0.001**	**0.9495**
	PF	0.9566 ± 0.0013	0.9446 ± 0.0024	0.9996 ± 0.0009	0.9669
Unit profit	AF	**13.52% ± 0.0**	5.94% ± 2.11	-2.8% ± 1.29	5.55
	GF	-5.4% ± 2.57	13.66% ± 0.0	-5.08% ± 2.45	1.06
	PF	12.26% ± 5.1	**16.11% ± 2.64**	**2.23% ± 2.09**	**10.2**
Kelly profit	AF	**35.59% ± 4.45**	12.37% ± 1.02	24.2% ± 4.43	24.05
	GF	-19.06% ± 1.87	**107.61% ± 4.47**	**99.75% ± 3.63**	**62.77**
	PF	2.04% ± 2.62	10.82% ± 4.87	8.3% ± 2.1	7.05

Table 5.4. Simulation results for *OVR model*.

Metric		**WC 2018**	**WC 2014**	**WC 2010**	Mean
Accuracy	AF	57.66% ± 0.47	57.34% ± 1.22	56.09% ± 0.84	57.03
	GF	51.25% ± 1.95	55.16% ± 1.57	**57.03% ± 1.05**	54.48
	PF	**61.09% ± 1.3**	**60.16% ± 1.05**	54.37% ± 1.36	**58.54**
Log Loss	AF	0.9646 ± 0.0035	**0.9382 ± 0.0039**	0.9831 ± 0.005	**0.962**
	GF	1.012 ± 0.0052	0.9515 ± 0.0068	**0.9583 ± 0.0059**	0.9739
	PF	**0.9365 ± 0.0023**	0.9564 ± 0.0032	1.0083 ± 0.0052	0.9671
Unit profit	AF	7.31% ± 1.04	5.01% ± 2.93	6.52% ± 2.97	6.28
	GF	-5.03% ± 4.84	3.04% ± 4.31	**10.38% ± 3.38**	2.8
	PF	**19.83% ± 4.01**	**14.14% ± 2.82**	3.88% ± 5.11	**12.62**
Kelly profit	AF	-2.6% ± 6.57	9.34% ± 8.69	26.26% ± 10.92	11.0
	GF	-44.81% ± 3.96	**14.61% ± 13.1**	**83.44% ± 23.89**	17.75
	PF	**61.65% ± 9.88**	-3.92% ± 4.68	0.77% ± 7.02	**19.5**

Table 5.5. Simulation results for *Linear model*.

Metric		**WC 2018**	**WC 2014**	**WC 2010**	Mean
Accuracy	AF	54.84% ± 0.84	**63.59% ± 0.72**	51.41% ± 1.3	56.61
	GF	51.72% ± 1.09	63.28% ± 1.26	**54.37% ± 1.53**	56.46
	PF	**59.69% ± 0.62**	62.19% ± 0.62	50.78% ± 0.78	**57.55**
Log Loss	AF	0.9719 ± 0.0043	**0.8791 ± 0.0025**	1.0418 ± 0.0071	0.9643
	GF	0.9953 ± 0.0041	0.9112 ± 0.0035	**0.9837 ± 0.0028**	**0.9634**
	PF	**0.9334 ± 0.0017**	0.9311 ± 0.0032	1.0563 ± 0.0064	0.9736
Unit profit	AF	0.79% ± 1.51	**27.73% ± 3.09**	-5.11% ± 3.88	7.8
	GF	-3.75% ± 2.07	24.57% ± 4.51	**-0.28% ± 5.87**	6.85
	PF	**14.73% ± 1.93**	21.98% ± 2.7	-8.95% ± 2.03	**9.25**
Kelly profit	AF	-20.29% ± 5.49	**271.57% ± 25.03**	-43.87% ± 5.37	**69.14**
	GF	-36.47% ± 3.66	125.73% ± 11.94	**8.75% ± 10.44**	32.67
	PF	**55.93% ± 4.8**	52.77% ± 8.74	-49.56% ± 4.26	19.71

game-level is out of the reach of this thesis. Luckily, a single tournament can be used to see how the models differ in game-level prediction since all of the tournaments have more or less the same characteristics. These characteristics are listed in Table 5.1. The only noteworthy difference is the higher number of draws in the World cup 2010.

The reason why a tournament with few more games ending in a draw matters so much is because the models are very inaccurate when it comes to predicting the outcome of a draw. For example, the average precision of a draw for *Outcome model* with any of the feature sets is at maximum 10%. The average recall is at best 6.25%. In most of the cases, precision and recall are almost always zero. Both of those values are for the World Cup 2010, meaning that only one out of the 10 draws that are predicted is correct. Draws are not preferred by the bookmakers either. Very seldom a draw has the lowest odds (meaning the highest probability). For example, only one game in the World Cup 2010 had the lowest odds for a draw, and that game did not end in a draw.

We compare game-level results for the World cup 2018. Models are trained using all of the available features.

One interesting fact is that the predicted outcomes are very similar as Figure 5.1 shows. *Outcome model* and *Score model* are almost equal in their predictions. Based on the overall results, models should differ more since profits are not equal or even similar. What might be the reason for this? To see more details, it is wise to use the Kelly strategy's cash balance progression from Figure 5.2. Cash balances develop independently from the first game onward. The reason behind this is the different probability distributions that the models output. By observing the correlations from

Tables 5.6, 5.7 and 5.8 we can see that the models' behavior is fairly similar with the probability prediction of a home win and an away win, but differs clearly with the prediction of a draw. Even when only random forest-based solutions are compared the difference between the models is the most obvious when the probability estimates for a draw are compared. Figure 5.3 shows how differently the models estimate the probability of a game ending in a draw. This figure and the standard deviation values from Table 5.7 show that *Outcome model* and *OVR model* predict probabilities of a draw more aggressively; both of the models often predict a lower or a higher probability than the rest of the models. *Score model's* and *Linear model's* standard deviation is a half of what *Outcome model* or *OVR model* have. *Linear model* stands out from the tree-based models with low correlating probability estimates. All of the models have a similar average value for the probability estimate.

We compared *Score model's* and *Outcome model's* probability estimates to a sample probability of a draw obtained from historical data. Results are visualized in Figure 5.4. *Score model* and *Outcome model* often predict a higher probability than the sample probability from historical matches is for the games that have a relatively larger difference between teams' Elo rating. This is the case also with a sample probability of a draw obtained from historical World Cup games. With historical World Cup games, Elo difference is split into deciles so that the line would be less noisy. Each bucket contains 84 games. Based on Figure 5.3 and the metrics in Table 5.7 rest of the models have this same bias towards too high probability estimates when the difference between Elo rating increases.

Table 5.6. Means, standard deviations, and correlations of home win probability predictions for the World cup 2018.

	Measure	mean	sd	correlations			
				1.	2.	3.	4.
	Models						
1.	Score	0.3956	0.1588				
2.	Outcome	0.3854	0.1685	0.9873			
3.	OneVsRest	0.4016	0.1687	0.9870	0.9871		
4.	Linear	0.4170	0.1698	0.9654	0.9660	0.9733	

Table 5.7. Means, standard deviations, and correlations of draw probability predictions for the World cup 2018.

	Measure	mean	sd	correlations			
				1.	2.	3.	4.
	Models						
1.	Score	0.2527	0.0274				
2.	Outcome	0.2619	0.0414	0.8102			
3.	OneVsRest	0.2553	0.0414	0.7713	0.9333		
4.	Linear	0.2484	0.0197	0.2356	0.3162	0.3080	

Table 5.8. Means, standard deviations, and correlations of away win probability predictions for the World cup 2018.

	Measure	mean	sd	correlations			
				1.	2.	3.	4.
	Models						
1.	Score	0.3517	0.1504				
2.	Outcome	0.3528	0.1667	0.9824			
3.	OneVsRest	0.3431	0.1674	0.9814	0.9889		
4.	Linear	0.3345	0.1753	0.9648	0.9672	0.9680	

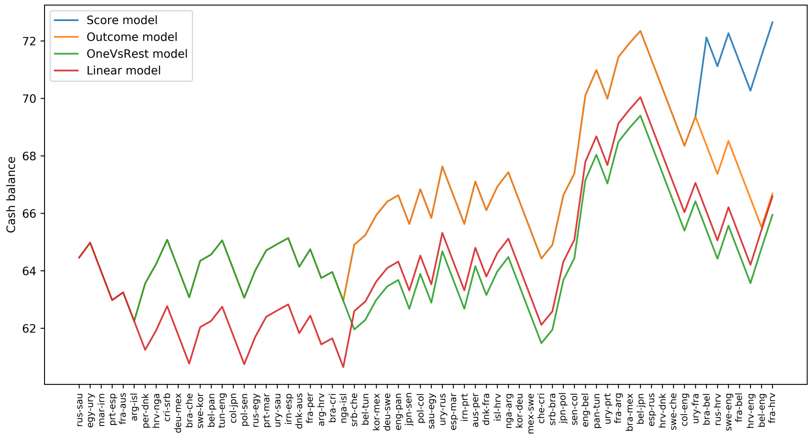

Figure 5.1. Unit strategy's cash balance progression on the World Cup 2018.

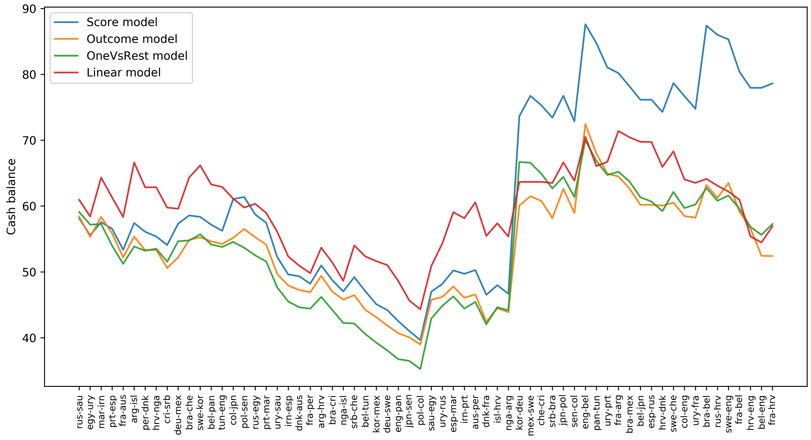

Figure 5.2. Kelly strategy's cash balance progression on the World Cup 2018.

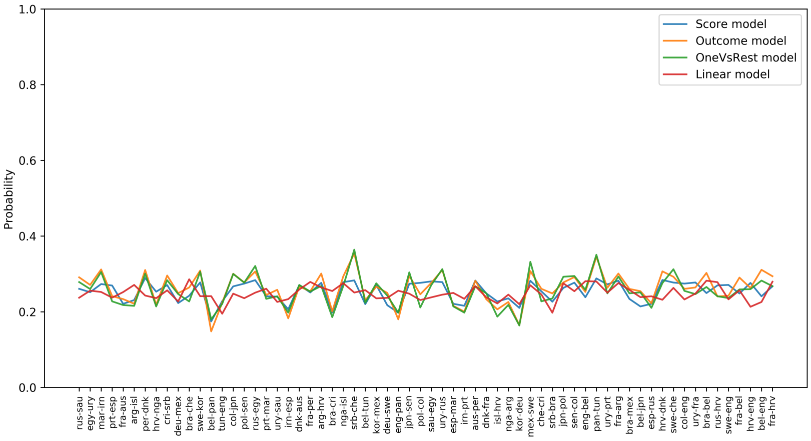

Figure 5.3. Predicted probability of draw in the World Cup 2018.

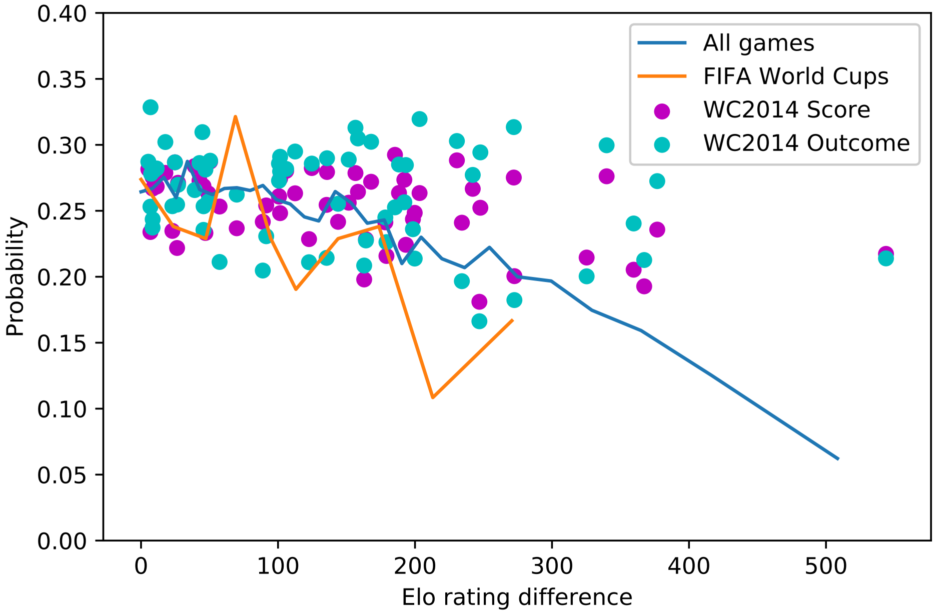

Figure 5.4. The predicted probability vs. the sample probability. The sample probability of a draw is calculated from all international football games between dates 11/30/1872 - 06/04/2018. The sample probability of a draw from World Cup games is calculated based on all World Cup games between the same period. Dots are single probability estimates from *Score model* and *Outcome model* for matches in the World Cup 2014.

5.3 Feature set comparison

Some of the models seem to work better with limited features. How can less describe more?

This phenomenon is visible if the models' accuracy is compared. Being one of the most important metrics, it is still not the most important for the sophisticated betting strategies like the Kelly strategy. Log loss is a metric that can be used to compare probability distributions' accuracy. A smaller value indicates a more accurate estimate. When log loss values are compared for each tournament, a model trained with a limited feature set can have a lower log loss value than the same model that is trained using all of the features. But when the average performance from every tournament is considered, training a model with all features seems to output the lowest log loss score with *Outcome model* and *OVR model* and second lowest (only 0.0009 difference at highest) with *Score model* and *Linear model*. This indicates that using all of the features benefits the probability distribution prediction, but not always the accuracy.

Random forest's feature importance value gives insight on model's reasoning. This data combined with optimal hyperparameters is one way to

Table 5.9. Optimal hyperparameters for the random forest models.

Model	Features	# of predictors	Min samples at a leaf	Max depth
Score	AF	$\sqrt{M}$	1	8
	GF	$\log_2 M$	10	8
	PF	$\sqrt{M}$	5	Na
Outcome	AF	$\sqrt{M}$	3	8
	GF	$\log_2 M$	15	5
	PF	$\log_2 M$	3	8
OVR	AF home win	$\sqrt{M}$	3	5
OVR	AF draw	$\sqrt{M}$	5	Na
OVR	AF away win	$\sqrt{M}$	3	5
OVR	GF home win	$\sqrt{M}$	10	12
OVR	GF draw	$\log_2 M$	15	8
OVR	GF away win	$\log_2 M$	10	5
OVR	PF home win	$\log_2 M$	3	5
OVR	PF draw	$\log_2 M$	15	5
OVR	PF away win	$\log_2 M$	5	8

interpret the models since inspecting every single tree in the forest would be too cumbersome. With a random forest, an optimal value that is higher than the default for *minimum samples at a leaf* can mean that some of the features are very noisy. Tree-based models, trained with generic features, seem to have a higher value for the hyperparameter *minimum samples at a leaf* as we can see from Table 5.9. The feature importance values for *Outcome model* trained with generic features are in Figure 5.6. Difference between Elo ratings is a valuable feature, but the rest of the features are harder to differentiate. Features *home_goal* and *away_goal* are the least important ones. When the importance of these features is listed for a model trained with all features (Figure 5.5) *home_goal* and *away_goal* seem to rank in the middle tier.

With *player features* the difference between features' importance is not as visible as it is with *general features*. Features' importance is gradually decreasing feature by feature as we can see from Figure 5.7. Same features rank high with player features as they would with all features.

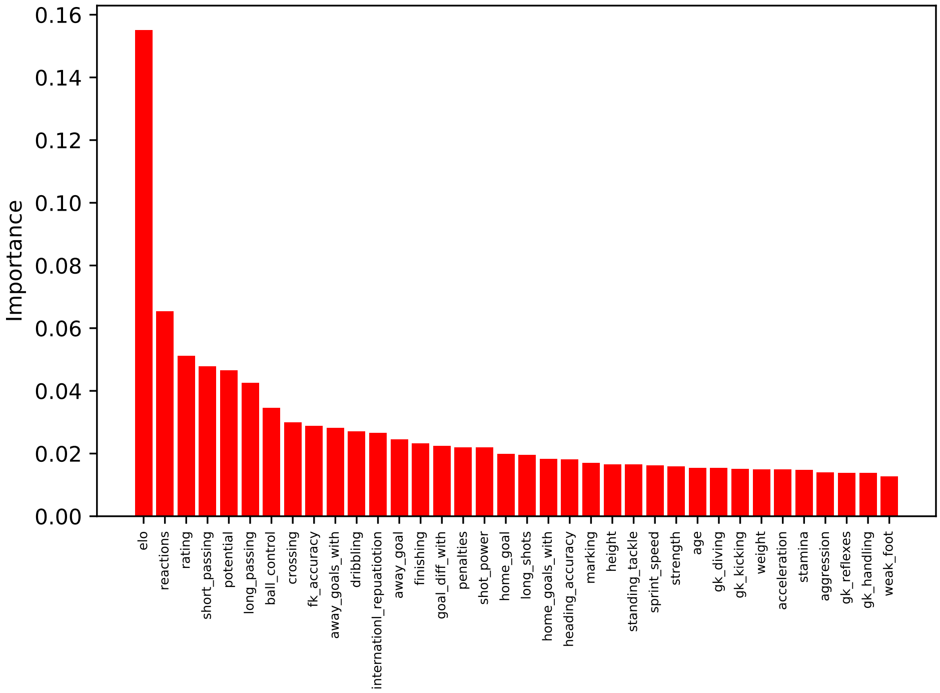

Figure 5.5. Outcome model's feature importance with all features.

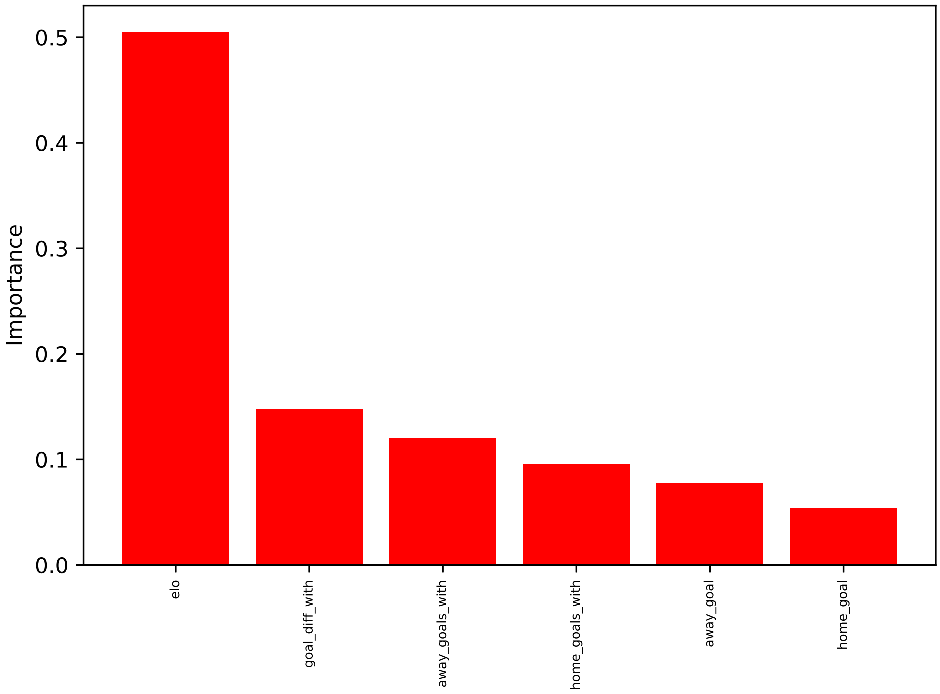

Figure 5.6. Outcome model's feature importance with generic features.

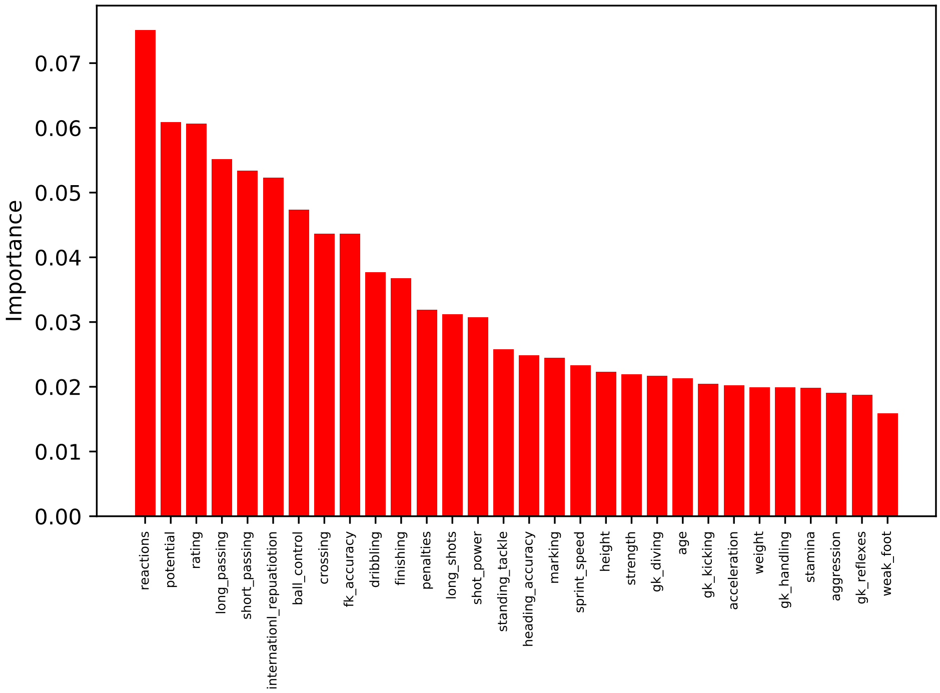

Figure 5.7. Outcome model's feature importance with player features.

5.4 Recursive feature elimination

To see if using a subset of features could improve the results, recursive feature elimination (RFE) is performed for *Outcome model*. The test is run two times: first time with default hyperparameters and the second time with optimal hyperparameters from the *all features* setup.

The first test includes the default parameters for *Outcome model*: 1000 for the number of estimators, $\sqrt{M}$ for the number of predictors, Na for the maximum depth and one for the minimum samples at a leaf node. From Figure 5.8 we can see that after the feature *gk_diving* is eliminated accuracy starts to decrease slowly, and from Figure 5.9 we can see that the log loss starts to increase. The second test's results are in Table 5.2, and these results are visualized in Figures 5.10 and 5.11. With this setup, the model performs better throughout the whole RFE process. Also, the model peaks the highest accuracy with only just four features. Improvement with this same feature set is not visible in the average log loss. The average log loss has a clear performance drop only when there is a single feature left.

To test the model's performance with a limited feature set we run the World cup simulation for *Outcome model* using features: *away_goals_with_home, finishing_diff, height_diff, crossing_diff, potential_diff, fk_accuracy_diff,* *home_goal_mean, away_goal_mean, ball_control_diff, long_passing_diff, short_passing_diff, reactions_diff, rating_diff, elo_diff*. We chose these features since they were the smallest feature set that was able to keep the performance with the model trained with the default hyperparameters. New optimal hyperparameters are grid searched for this feature set. These hyperparameters are listed in Table 5.11 and World Cup simulation results are in Table 5.10. Based on the accuracy and the log loss, model's performance does not improve with the limited feature set.

Based on the progression of the average accuracy and the average log loss, models do not perform better when features are removed from the original feature set. Eventually, performance decreases but never increases during the feature elimination. Most of the time results have natural variance, which causes small changes. It seems that using every available feature does not harm random forest's performance, even though there is no guarantee that all of the features are useful. Also, RFE's suitability for this setup can be questioned, since the results are not reliable if features are correlated. Stroble et al. concluded: "In particular, the selection of the first splitting variable involves only the marginal, univariate association between that predictor variable and the response, regardless of all other predictor variables. However, this search strategy leads to a variable selection pattern where a predictor variable that is per se only weakly or not at all associated with the response, but is highly correlated with another influential predictor variable, may appear equally well suited for splitting as the truly influential predictor variable." It is possible that correlated features are ranked to be more important than the influential feature.

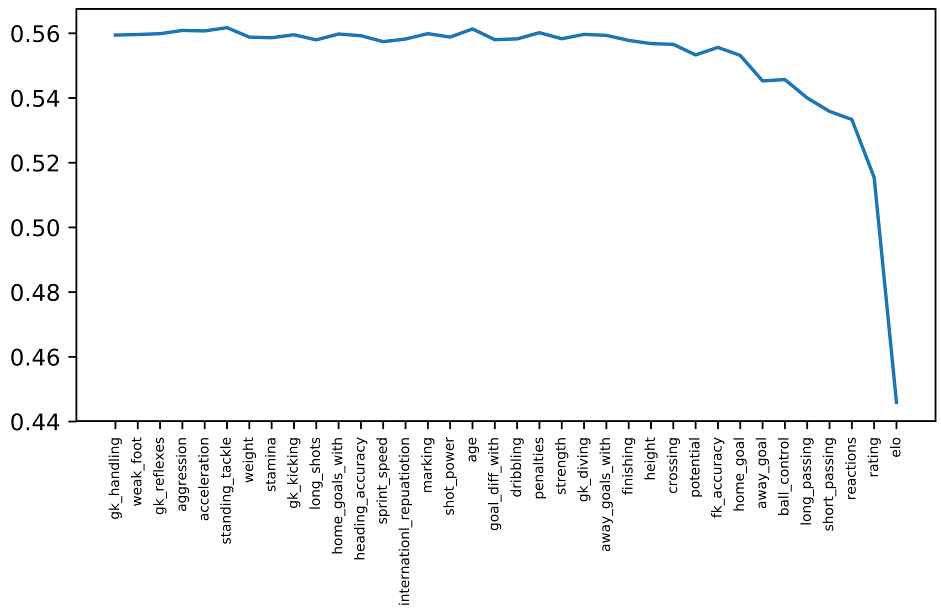

Figure 5.8. Mean accuracy for the recursive feature elimination on the *Outcome model* with the default hyperparameters. Only the features from the first round to the 30th round are list to ensure the figure's readability.

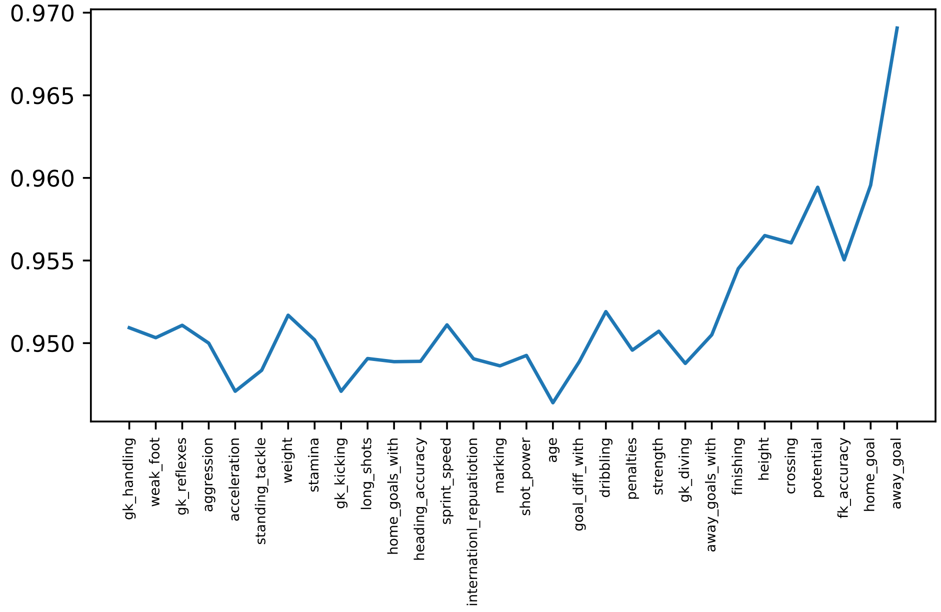

Figure 5.9. Mean log loss for the recursive feature elimination on the *Outcome model* with the default hyperparameters. Only the features from the first round to the 30th round are list to ensure the figure's readability.

Table 5.10. *Outcome model's* simulation results using the limited feature set from the RFE process.

Metric	**WC 2018**	**WC 2014**	**WC 2010**	Mean
Accuracy	52.97% ± 2.36	59.69% ± 2.3	56.09% ± 2.47	56.25
Log Loss	0.9953 ± 0.0087	0.9282 ± 0.0061	0.9743 ± 0.0069	0.9659
Unit profit	-1.79% ± 6.16	14.12% ± 6.02	8.82% ± 7.47	7.05
Kelly profit	-37.19% ± 11.43	61.45% ± 15.42	77.87% ± 12.98	34.04

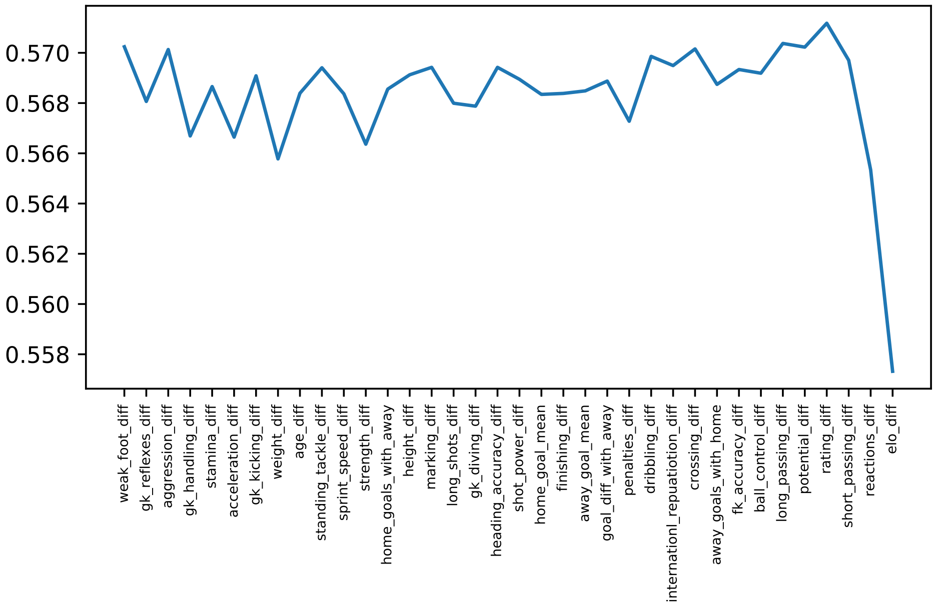

Figure 5.10. Mean accuracy for the recursive feature elimination on *Outcome model*. The optimal hyperparameters from the all features setup in Table 5.2 are used.

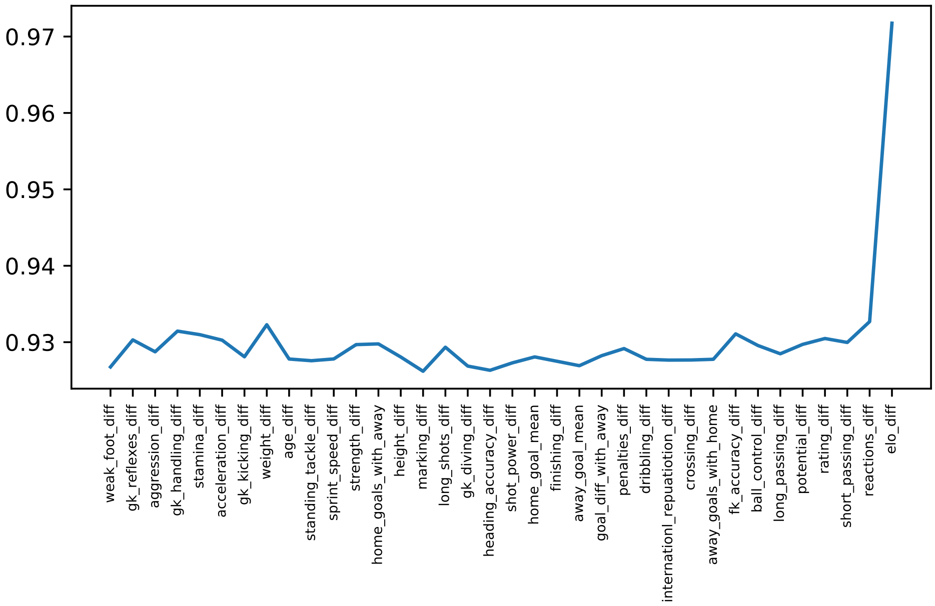

Figure 5.11. Mean log loss score for the recursive feature elimination on *Outcome model*. The optimal hyperparameters from the all features setup in Table 5.2 are used.

Table 5.11. The optimal hyperparameters for *Outcome model* trained with the limited feature set from the RFE process.

# of predictors	Min samples at leaf	Max depth
$\sqrt{M}$	10	Na

5.5 Match-level betting activity - what drives the profits?

A model can bet on the tournament in many ways. To answer the question "What drives the profits?" it is important to see how the individual games are betted. Does the model generate a small profit in most of the games or are the profits a result from a few successful opportunities? The unit strategy gains a profit when the model predicts correctly. There is no need to analyze this in more detail, since the strategy is straightforward. Kelly strategy instead is more sophisticated and deserves a more in-depth analysis. With the requirements 1) profitable in every tournament, and 2) best average profits has *Score model* trained with all features performed the best. For this reason, the following figures are plotted from *Score model's* results.

The most successful tournament for *Score model* was the World Cup 2018. It achieved on average 35.59% profit. In comparison, profit from the World Cup 2014 was only 12.37%. By the numbers performance in the World Cup 2018 was better, but what happened with match-level betting? When both of the tournaments are compared side by side in Figures 5.12 and 5.13 it is clear that a single successful bet dominates the profits in the World Cup 2018. Do not get confused by the difference in the scale of the y-axis. Total bets per match are not that different between the tournaments. Only betting frequency differs — there are more games in the World Cup 2014 without any bets. With the World Cup 2018, the winnings are more rare but bigger. Kelly strategy gains from the abnormally high odds that are available. In the match between South Korea and Germany, the odd for South Korea's win is 19.39, even though Germany had struggled the whole tournament. With the odd as high as 19.39, the implied probability for South Korea's win is close to 5%. These anomalous odds are certainly important, and Kelly strategy can utilize them properly. Without the three biggest winnings, this strategy would achieve small or negative profits. In the World Cup 2014, this strategy wins more often, but winnings are not as big as they are in the World Cup 2018. Steady wins would be ideal also for the World Cup 2018 but are more difficult to achieve. When winnings are more frequent predicted probabilities need to be closer to the true ones than the implied probabilities from the betting market's are. Based on the limited amount of evidence, smaller but more frequent winnings and sparser betting activity, it seems that the probability estimates for the World Cup 2014 are better than the implied ones based on the market. Whereas in World Cup 2018 profitability is based on the strategy's ability to gain from few favorable odds. A situation where abnormal odds are required for a profitable bet might be extremely hard to sustain in the long run unless betting markets provide opportunities like this often enough. Model's bias towards a higher probability of a draw compared to the market's when the difference between teams' Elo rating is relatively high is visible in both of the tournaments. This bias harms the strategy's performance.

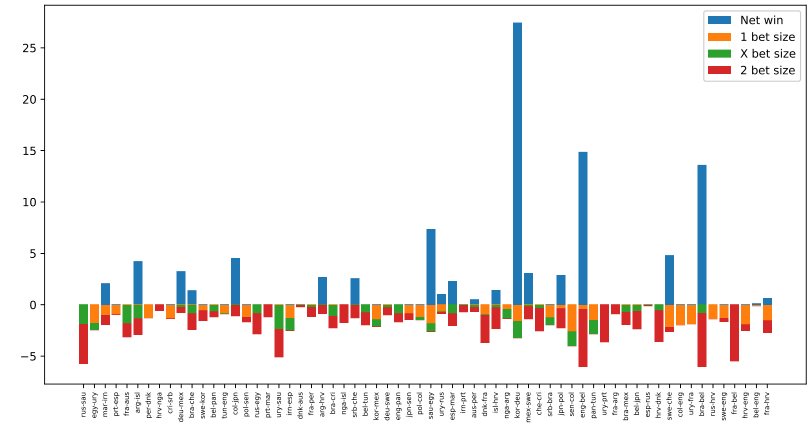

Figure 5.12. Net winnings and betting costs per label for Score model in World Cup 2018.

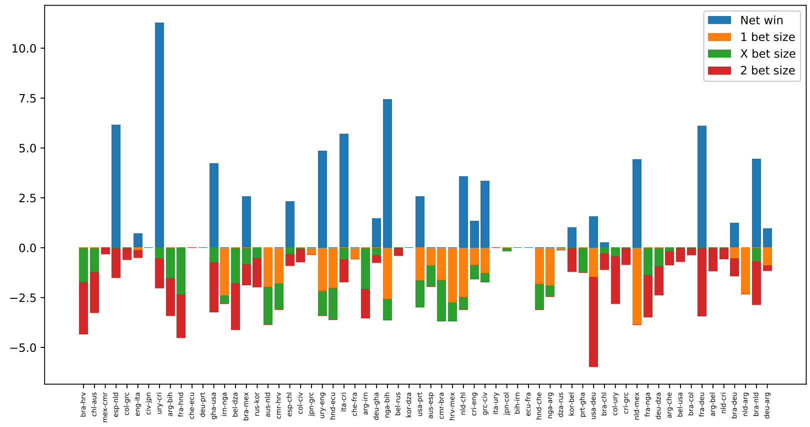

Figure 5.13. Net winnings and betting costs per label for Score model trained with all features in World Cup 2014.

5.6 Summary

In total there are 12 model and feature set combinations. In every tested World Cup tournament over half of the models perform better than the bookmaker's model. *Score model* trained with all features is the best model in performance. Its average return in all of the tournaments is 24.05%. From the models *Outcome model* trained with player features can generate profit from all of the tournaments using both of the betting strategies.

All of the models predict the probability for a home win and an away win very similarly. Probability estimates for a draw differ the most between the models. *Outcome model* and *OVR model* are more aggressive with the predictions while *Score model* and *Linear model* predict values closer to the average. All of the models miss most of the draws and very seldom give the highest probability for that class. Based on the implied probabilities bookmakers behave similarly.

Using a subset of the whole feature set improved the accuracy and the profits only for a single tournament. No subset was able to improve the results in all of the tournaments. Results from recursive feature elimination imply that the accuracy and the log loss score did not improve during the feature elimination process. With the whole feature set models can achieve the lowest log loss value or the second lowest log loss value with only a minimal difference to the best one. This constantly low log loss value indicates that using all of the available features is the best way to train any of the models.

The individual winnings differ between the World Cup 2018 and the World Cup 2014; winnings are more significant but rarer during the World Cup 2018. If this is the trend, meaning that during the upcoming tournaments finding a suitable odds to bet on is even harder, winnings become even more rare or non-existing.

6. Conclusion

In this thesis, we have investigated the possibility to "beat the bookie" on FIFA World Cup tournaments. To achieve this, the profits from betting need to be positive. In our experiments, we have used four different models with three different combinations of features. For betting, two different strategies were used. All the data used in the experiments was freely available on the internet.

Reference model, the bookmaker's model, was formed using the average odds provided for the match. This model achieved the accuracy of 56.25% for World Cup 2018 and 2014, and 51.56% for World Cup 2010. In total, 28 out of the all 36 model and feature set combinations were able to beat the bookmaker's model in prediction accuracy. "Beating the bookie" in prediction accuracy is doable.

Being profitable with both of the betting strategies in all of the tested World Cup tournaments: 2018, 2014, and 2010 is possible. More importantly, it's possible to earn on average as high as 24.05% returns using the Kelly strategy. The other strategy, the unit strategy, has less variance but the fixed bet size limits its ability to maximize the profits from favorable odds. Unit strategy's returns were more often positive, but lower in size. The answer to the question "Is it possible to 'beat the bookie'?" is *yes*. However, this result should be approached with caution. Combined results from three World Cups contains only 192 games, which means that the sample is small. It would be ideal to test the models with more tournaments if data would be available. Also, the lag of 4 years between the tournaments gives bookmakers plenty of time to improve their models. Already results from the latest World Cup indicate that excellent opportunities exist more infrequently; profits are more connected to few games, and only 7 out of the 12 model and feature set combinations were able to outperform the reference model's accuracy. There are no guarantees that the profits from the upcoming World Cup 2022 will be positive. In the future work, these methods could be used to simulate games in different football leagues. More games can be used for validation which gives more confidence in the results.

Extensive feature analysis with tree-based models indicated that no subset of features gave better results than using all of the features. Some of the tournament simulation results were in contrary to this since in some cases using a limited feature set improved the result of a single tournament simulation. However, when all tournaments were considered, using all of the features seems to be the best option. Instead of further optimizing the perfect feature set, the features themselves should be investigated. Maybe the current way of aggregate the team-level features could be improved to differentiate the teams better. Also, more data, like lineups, could be included to enhance the feature's accuracy. Using optimal hyperparameters turned out to be an essential way to improve the results. The grid search strategy for finding the optimal hyperparameters was successful.

Predicting draws turned out to be demanding. What could be done to improve this? This thesis will not provide a clear answer. Models predicted draws differently, but no model was clearly better than the rest. The reference model was no better in accuracy. Improving the prediction accuracy of a draw would be most likely very beneficial and hopefully initiates further research.

Bibliography

[1] A. Groll, C. Ley, G. Schauberger, and H. Van Eetvelde, "Prediction of the fifa world cup 2018-a random forest approach with an emphasis on estimated team ability parameters," *arXiv preprint arXiv:1806.03208*, 2018.

[2] A. Groll, G. Schauberger, and G. Tutz, "Prediction of major international soccer tournaments based on team-specific regularized poisson regression: An application to the fifa world cup 2014," *Journal of Quantitative Analysis in Sports*, vol. 11, no. 2, pp. 97–115, 2015.

[3] C. Leitner, A. Zeileis, and K. Hornik, "Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the euro 2008," *International Journal of Forecasting*, vol. 26, no. 3, pp. 471–481, 2010.

[4] N. Vlastakis, G. Dotsis, and R. N. Markellos, "How efficient is the european football betting market? evidence from arbitrage and trading strategies," *Journal of Forecasting*, vol. 28, no. 5, pp. 426–444, 2009.

[5] T. Kuypers, "Information and efficiency: an empirical study of a fixed odds betting market," *Applied Economics*, vol. 32, no. 11, pp. 1353–1363, 2000.

[6] J. Shin and R. Gasparyan, "A novel way to soccer match prediction," *Stanford University: Department of Computer Science*, 2014.

[7] M. J. Moroney, *Facts from figures*. Penguin books, 1962.

[8] D. M. J. and C. S. G., "Modelling association football scores and inefficiencies in the football betting market," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 46, no. 2, pp. 265–280, 1997.

[9] C. Anderson and D. Sally, *The numbers game: why everything you know about football is wrong*. Penguin UK, 2013.

[10] P. Tüfekci, "Prediction of football match results in turkish super league games," in *Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015*, A. Abraham, K. Wegrzyn-Wolska, A. E. Hassanien, V. Snasel, and A. M. Alimi, Eds. Cham: Springer International Publishing, 2016, pp. 515–526.

[11] E. Ben-Naim, F. Vazquez, and S. Redner, "Parity and predictability of competitions," *Journal of Quantitative Analysis in Sports*, vol. 2, no. 4, 2006.

[12] M. J. Maher, "Modelling association football scores," *Statistica Neerlandica*, vol. 36, no. 3, pp. 109–118, 1982.

[13] A. E. Elo, *The rating of chessplayers, past and present*. Arco Pub., 1978.

[14] Wikipedia contributors. (2018) Fifa world rankings — Wikipedia, the free encyclopedia. [Online; accessed 10-July-2018]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Plagiarism&oldid=5139350

[15] J. Lasek, Z. Szlávik, and S. Bhulai, "The predictive power of ranking systems in association football," *International Journal of Applied Pattern Recognition*, vol. 1, no. 1, pp. 27–46, 2013.

[16] L. M. Hvattum and H. Arntzen, "Using elo ratings for match result prediction in association football," *International Journal of forecasting*, vol. 26, no. 3, pp. 460–470, 2010.

[17] R. Badarinathi and L. Kochman, "Football betting and the efficient market hypothesis," *The American Economist*, vol. 40, no. 2, pp. 52–55, 1996.

[18] L. D. Pankoff, "Market efficiency and football betting," *The Journal of Business*, vol. 41, no. 2, pp. 203–214, 1968.

[19] N. Jegadeesh and S. Titman, "Returns to buying winners and selling losers: Implications for stock market efficiency," *The Journal of finance*, vol. 48, no. 1, pp. 65–91, 1993.

[20] J. Goddard and I. Asimakopoulos, "Modelling football match results and the efficiency of fixed-odds betting," Working Paper, Department of Economics, Swansea University, Tech. Rep., 2003.

[21] J. L. Kelly Jr, "A new interpretation of information rate," in *The Kelly Capital Growth Investment Criterion: Theory and Practice*. World Scientific, 2011, pp. 25–34.

[22] L. MacLean, W. T. Ziemba, and G. Blazenko, "Growth versus security in dynamic investment analysis," *Management Science*, vol. 38, no. 11, pp. 1562–1585, 1992.

[23] M. Jürisoo. (2018) International football results from 1872 to 2018. [Online; accessed 10-July-2018]. [Online]. Available: https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017

[24] F. Play. (2018) Fifa encyclopedia. [Online; accessed 10-July-2018]. [Online]. Available: http://www.fifplay.com/encyclopedia/

[25] Sofifa. (2018) Ea sport's video game series fifa's player attributes. [Online; accessed 10-July-2018]. [Online]. Available: https://sofifa.com

[26] Odds Portal. (2018) World cup 2018 results & historical odds. [Online; accessed 12-July-2018]. [Online]. Available: http://www.oddsportal.com/soccer/world/world-cup/results/

[27] L. Rokach and O. Maimon, "Top-down induction of decision trees classifiers-a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 35, no. 4, pp. 476–487, 2005.

[28] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, NY, USA:, 2001, vol. 1, no. 10.

[29] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," *Neural computation*, vol. 9, no. 7, pp. 1545–1588, 1997.

[30] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[31] P. Probst, M. Wright, and A.-L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *arXiv preprint arXiv:1804.03515*, 2018.

[32] S. Bernard, L. Heutte, and S. Adam, "Influence of hyperparameters on random forest accuracy," in *International Workshop on Multiple Classifier Systems*. Springer, 2009, pp. 171–180.

[33] M. R. Segal, "Machine learning benchmarks and random forest regression," 2004.

[34] N. M. Nasrabadi, "Pattern recognition and machine learning," *Journal of electronic imaging*, vol. 16, no. 4, p. 049901, 2007.

[35] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[36] C. Walck, "Hand-book on statistical distributions for experimentalists," Tech. Rep., 1996.

[37] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 694–699.

[38] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 625–632.

[39] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[40] L. C. MacLean, E. O. Thorp, Y. Zhao, and W. T. Ziemba, "Medium term simulations of the full kelly and fractional kelly investment strategies," in *The Kelly Capital Growth Investment Criterion: Theory and Practice*. World Scientific, 2011, pp. 543–561.

[41] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, "Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products," *Chemometrics and Intelligent Laboratory Systems*, vol. 83, no. 2, pp. 83–90, 2006.

[42] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," 2001–, [Online; accessed <today>]. [Online]. Available: http://www.scipy.org/