The Wisdom of the Crowd:

Predicting Domestic Football Matches Using Crowd Estimated Market Values

Abstract

This study tries to determine whether football matches from the English Premier League can be predicted using crowd estimated market values from the website Transfermarkt. It investigates whether the wisdom of the crowd effect can be found with the market values. Using multinomial logistic regression different models are estimated. The predictions of the market value models are compared to models based on the ELO ratings and the predictions of bookmakers. The results show that it is likely that the wisdom of the crowd effect is present and that market values can be used to predict match outcomes, but that it does not outperform ELO based models. The prediction accuracy is equivalent to that of the bookmakers.

Frank Peschier – 359484

Master thesis Behavioural Economics

Supervisor: G.D. Granic

Erasmus University of Economics

1. Preface

You are reading the master thesis that serves as the completion of my master Behaviour Economics at the Erasmus University.

I foremost want to thank my thesis supervisor Dr. Georg Granic, who helped me immensely during the process of writing this thesis. He helped me with a multitude of things, from helping me lay out the structure of this thesis to advising on how to correct unusual results. I'm sure that without his guidance this thesis would not be the same quality as it is today.

Secondly I want to thank Dr. Thomas Peeters for providing the subject of this thesis and for providing his extensive database. Although I haven't used this database in my research, it proved to be a good example of possible variables that I would need.

Lastly I want to thank my family and friends for supporting me throughout the process of this thesis. I especially want to thank Leonie de Vries for her help with correcting the text and serving as a sounding board for all my thoughts and ideas.

I hope you will enjoy reading this thesis.

Frank Peschier Leiden - August 11, 2016

Contents

2. Synopsis

The topic of this thesis is the prediction of football matches in the English Premier League. I try to predict the outcomes using market values that are collectively estimated by the members of Transfermarkt, an online community where members can propose market valuations of players (Transfermarkt, 2016a). Use of crowd estimated data ranges back as far as the 18th century and the added benefit of using this data has been called the wisdom of the crowds in more recent times. Groups are remarkably intelligent, and are often smarter than the smartest people in them (Surowiecki, 2004).

I try to see whether the wisdom of the crowd effect can be found with these market values and whether these values can be used to predict football matches. The research question of this research is as follows:

Can crowd estimated market values help predict domestic football match outcomes?

More accurate predictions of match outcomes can be very beneficial for people like bookmakers and the results of this research can provide a better understanding of the benefits of using the wisdom of the crowd.

Earlier research has been done on the subject of predicting football matches. However, these studies either use different methods like ELO ratings (Hvattum & Arntzen, 2010), or focus on international football, not domestic football (Peeters, 2016).

I use four hypotheses that help answer the research question:

- **H1**: The market values from Transfermarkt do not satisfy the requirements for the wisdom of the crowd effect
- **H2**: Market values that are used as a covariate in a regression cannot help predict match outcomes
- **H3**: A model based on market values does not predict match outcomes better than a model based on ELO ratings
- **H4**: A model based on market values does not predict match outcomes better than the bookmakers

By comparing the data of Transfermarkt with the four requirements of the wisdom of the crowd devised by Surowiecki (2004), I find that it is very likely that the wisdom of the crowd effect can be found with the market values from Transfermarkt. The null of hypothesis 1 is rejected.

To help answer the other hypotheses I estimate 4 multinomial logistic models. Model 1 only has the market values as an explanatory variable while model 2 also has added control variables. ELO ratings are calculated for all teams using the goal based ELO method and 3 seasons are used to calibrate the ratings (Hvattum & Arntzen, 2010). Model 3 only has the ELO ratings as an explanatory variable while model 4 also has added control variables. The models show that the market values have a significant effect on the match outcome. The null of hypothesis 2 is rejected, market values can be used to predict match outcomes.

The models are compared by means of likelihood tests and Brier scores. The results show that the models that have added control variables are preferred over the restricted models. Model 4 has a lower Akaike Information Criterion than model 2, but this cannot be tested for significance. Tests are done to make sure t-tests can be used to compare the Brier scores. The Brier scores show that model 4 is more accurate than model 2. Hypothesis 3 cannot be rejected.

I use odds of seven bookmakers to calculate match outcome probabilities. I compare the Brier scores of these probabilities with models 2 and 4. The Brier scores show that model 4 outperforms the bookmakers, but I find no evidence that model 2 predicts match outcomes more accurate than the bookmakers.

To check the monetary value, I use the models to bet on value bets.

Table 1: The average maximum returns per model when only value bets are played.

Betting results

	Average Maximum Return	# Matches	# Value Bets
Model 1	-0.87%	340	567
Model 2	7.45%	340	519
Model 3	6.05%	340	570
Model 4	10.26%	340	501
Bet	8.00%	340	533

Betting according all models, except model 1 yields a positive average maximum return. Only when betting according to model 4, higher average maximum returns than betting according to the bookmakers are achieved. Hypothesis 4 cannot be rejected.

I conclude that crowd estimated market values can indeed help predict domestic football match outcomes, though it does not outperform a model based on ELO ratings or the bookmakers. After answering the research question I discuss a few limitations and make suggestions for further research. Lastly I make some policy recommendations.

3. Introduction

On May 2nd 2016, the English football club Leicester City FC became the champions of the English Premier League. At the beginning of the season in August, odds of 5000/1 were offered for betting on Leicester City becoming the champions, amusingly the same odds were offered for Elvis Presley being found alive in 2016. This gives an indication on how low the probability of Leicester City becoming champions was deemed by the English bookmakers. But Leicester City did win the title, which meant that English bookmakers had to pay out £25 million, the biggest loss in British history on a single sporting market (Rayner & Brown, 2016).

The question that arises is whether the bookmakers could have foreseen Leicester City becoming the champions and if they could have adjusted their odds more adequately? By predicting match outcomes, bookmakers can calculate exactly how much money they will win by setting the odds at certain levels. By this they have control over the bets that are placed. If the bookmakers end up being wrong with their predictions and for example set the odds for a certain outcome too high, they can lose a large amount of money. It is clear that bookmakers could benefit from more accurate prediction models.

Sport outcome prediction has been studied extensively in the literature. In a study by Hvattum and Arntzen, ELO ratings, which are originally used to estimate chess players' strengths, are used to predict football outcomes (2010). The authors found that their methods based on ELO ratings can be used to predict football matches, but are not able to outperform the methods using market odds. In a different study, linear regressions were estimated that allow for predictions of matches played in the English Premier League (Louzada, Suzuki, & Salasar, 2014).

The aforementioned studies all base their models on factual objective data. But it is also possible to make predictions using subjective data. A means of doing that is by making use of the wisdom of the crowd. The idea of the wisdom of the crowd is that a group of people can collectively decide on the right solution, while being more accurate than a few experts. The wisdom of the crowd is not a new concept; it has been around as early as the 18th century (Sunstein, 2006). A somewhat more recent example is a publication by Sir Francis Galton (1907). He asked a crowd of people at an ox auction to estimate the weight of a certain ox. Using all these estimations, he was able to construct a precise estimation of the ox's weight. He called this concept 'vox populi' or the wisdom of the crowd. The crowd's average judgement converged to the right solution. Since then four requirements for the wisdom of the crowd to emerge have been formulated by Surowiecki (2004).

The wisdom of the crowd has been used to help predict sport outcomes. It has been found that the mere recognition of tennis player names has an added benefit to predicting models (Herzog & Hertwig, 2011). A second example is that crowd estimated market values have been used to predict international football matches (Peeters, 2016).

Besides sport outcome prediction, the concept of crowd wisdom has been studied in many other different fields. It has for example been applied to political science to forecast US presidential elections (Murr, 2015). It can also be found in studies on the rationality of group decisions versus individual decisions under risk (Baillon, Bleichrodt, Liu, & Wakker, 2016). Study shows that decisions made by groups are more rational than decisions made by individuals (Charness & Sutter, 2012). Furthermore, research has been done on crowd wisdom and prediction markets. In the study by Wolfers and Zitzewitz it was found that crowds perform well in information aggregation tasks (2004).

A good source of subjective football data is Transfermarkt (Transfermarkt, 2016a). This community finds its origin back in 2000 when its first version went online. Since then Transfermarkt has seen a large grow in the last couple of years resulting in a more extensive database and even localized versions of their website (Transfermarkt, 2016b). Transfermarkt allows its users to make estimations of each player's market value. These crowd estimations are combined into a single market value.

The question that arises is whether these crowd estimated market values can predict domestic football matches and whether the wisdom of the crowd effect takes place with these market values. As mentioned earlier, this could be very beneficial for parties using match prediction models. But it could also give us another insight on the usability of the wisdom of the crowd. This leads to the following research question:

RQ: Can crowd estimated market values help predict domestic football match outcomes?

This research focusses on the English Premier League. This league is the most watched football league in the world with a TV audience of 4.7 billion back in 2013 (Ebner, 2013). It also a league where a great amount of money is circulating as can be seen by the record British TV rights deal worth £5.14 billion (Gibson, 2015). It has also been studied in previous research on match prediction. Focusing on the same league allows me to compare findings of my research with the findings of these other studies.

4. Theoretical framework

Previous research has been done on the subject of crowd wisdom and match outcome prediction. In this chapter, the most relevant articles are mentioned and their main findings are discussed. I start with discussing the wisdom of the crowd effect and its requirements. These requirements are very important because they will allow me to see whether this effect can be found with the market values from Transfermarkt. After discussing the wisdom of the crowd I discuss the studies done on match prediction. It is important to see what kind of methods and models previous studies have used in their research. So examining those studies gives me an idea what variables are important to add to my own models.

4.1 Wisdom of the crowd

First we start with looking at research done on the wisdom of the crowd. As mentioned earlier, much attention has been devoted to the wisdom of the crowd. In the book 'The Wisdom of Crowds' authored by James Surowiecki (2004), the author states that under the right circumstances, groups are remarkably intelligent, and are often smarter than the smartest people in them. Groups do not need to be dominated by exceptionally intelligent people in order to be smart. Even if most of the people within a group are not especially well-informed or rational, it can still reach a collectively wise decision. Surowiecki states four requirements that need to be satisfied before a crowd can be deemed wise.

The first requirement is that there should be diversity of opinion. Each person should have some private information, which can also be an interpretation of known facts. Diversity among the crowd adds perspectives that would otherwise be absent. According to Surowiecki, diversity also takes, or at least weakens, some of the destructive characteristics of group decision making. Making sure a group is diverse is more important for small groups than large groups like those that can be found in markets. Due to the sheer size of the markets, a minimum level of diversity is likely to be observed. In a small group, a few biased individuals can skew the group's collective decision. Diversity makes it easier for the group to make collective decisions when they base them solely on facts rather than on for example emotional biases. A study by Hong & Page also found that the wisest crowds are the most diverse (2004). They found that a diverse group of problem-solvers made a better collective guess than the guess made by the group of best-performing solvers. So a group of diverse minds performs, when their decisions are averaged, better than a group of expert minds.

The second requirement for a group to be wise is independence of opinion. People should have relative freedom from the influence of others and should be able to form their own opinions. This is important because a group of people is far more likely to make a collectively good decision when the people in that group are independent from each other. If this is not the case, then there is a higher chance that the guesses made by the crowd are drifted towards a misplaced bias. According to Surowiecki, there are two reasons why independence is important. The first reason is that independence makes sure that mistakes made by people will not become correlated. As long as the mistakes made by the people are not systematically moving in the same direction, then mistakes by individual people won't wreck the group's collective judgement. If people's decisions are dependent on other people's decisions, then their judgements are systematically biased and a group is less likely to be wise. The second reason why independence is important is that independent individuals are more likely to possess new information instead of the same information everyone else knows. The same findings were shown by a paper in 2011 (Lorenz, Rauhut, Schweitzer, & Helbing). The authors of this paper show that even mild social influence can undermine the wisdom of crowd effect in simple estimation tasks. They found that groups are initially wise, but when they have knowledge about estimates of others, the diversity of opinions is narrowed to such an extent that it undermines the wisdom of crowd effect. This happens in three different ways. First there is the social influence effect which diminishes the diversity of the crowd without improvements of its collective error. Second there is the range reduction effect which moves the position of the truth to the outer regions of the range of estimates. This effect causes the crowd to become less reliable in providing expertise for external observers. Lastly the confidence effect boosts individuals' confidence after convergence of their estimates despite lack of improved accuracy.

The third requirement is decentralization. People should be able to specialize and draw on local knowledge. Surowiecki mentions that decentralization fosters, and in turn is fed by specialization of labour, interest and attention. Specialization tends to make people more productive and efficient. If more people in a group are specialized, the scope if diversity of the opinions and information is increased. An assumption of decentralization is that the closer a person is to the problem, the more likely that that person has a good solution for it. So decentralization both encourages independence and specialization while still allowing people to coordinate their activities and solve difficult problems.

The fourth and last requirement is aggregation; some mechanism should exist to turn private judgments into a collective decision.

These four requirements should be satisfied for the wisdom of the crowd effect to be present. Any crowd estimated data I gather should be compared to these requirements to see if they are met.

4.2 Match prediction

Now we will have a look at match prediction. Multiple studies have been done on this subject. The research by Hvattum and Arntzen examines the value of assigning ratings to teams based on their past performance in order to predict match results in association football (2010). The ELO rating system is used to derive covariates that are then used in ordered logit regression models. In order to make informed statements about the relative merit of the ELO-based predictions, the authors compare them to predictions made by a set of six benchmark prediction methods. The ELO based methods performed significantly worse than the two methods that were based on market odds. However, the ELO based methods outperformed the other methods, in terms of observed loss. The authors conclude that ELO ratings appear to be useful in encoding information on past results. In the case of association football, the single rating difference is a highly significant predictor of match outcomes. This finding can be taken as a justification of the increasingly common use of ELO ratings as a measure of team strength.

A different study published in 2014, tried to predict match outcomes in the English Premier League (Louzada, Suzuki, & Salasar). The authors of this study estimated linear models that express the sum and the difference of goals scored in terms of five covariates: the goal average in a match, the home-team advantage, the team's offensive power, the opponent team's defensive power and a crisis indicator. Their model allows them to predict multiple different things like which team will score or concede the most goals and which team will end at the top or the bottom of the table.

Another study that tries to forecast match results in the English Premier League was published in 2015 (Koopman & Lit, 2015). The authors state that the attack and defence strengths of football teams vary over time due to changes in the teams of players or their managers. To account for this, they develop a statistical model for the analysis and forecasting of football match results which assumes a bivariate Poisson distribution with intensity coefficients that change stochastically over time. They show that the model they construct can produce a significant positive return over the bookmaker's odds.

So a multitude of research has been done on football match prediction using non crowd based data. But as we are trying to predict matches using crowd estimated data, it is interesting to see what kind of studies have been done that also use such data. In a research by Herzog and Hertwig, they test whether crowd recognition of player names has any predictive power in forecasting football matches (2011). Their study is based on the collective recognition heuristic, which predicts that the better-known team or player wins a match. The names of better players should be mentioned more often in for example the media. Knowledge of a name could then be related to the quality of said player. The authors find that across three soccer and two tennis tournaments, the predictions based on the recognition of player names performed similar to predictions based on official rankings; when compared with betting odds, the heuristic fared reasonably well. Forecasts based on official rankings were improved by incorporating collective recognition information.

Besides using the crowd recognition of players, research has also been done using data from Transfermarkt. In a forthcoming paper by Peeters, a study was done on how adequately player valuations from non-expert users of an online platform predict the results of international soccer matches (2016). The online platform that is mentioned is Transfermarkt. Using a simple model with only the average crowd valuation and the number of players, the author found that this model predicts the performance of a national team better than for example the FIFA ranking and ELO ratings. The author compares the returns to different betting strategies and he finds that the gain in predictive power of the model is economically relevant. He finds no evidence of wishful thinking bias in the valuations, which means that valuations of popular players are not biased upwards.

The study by Peeters looked a little at the quality of the market values from Transfermarkt. A research that studies the quality of the market values more in depth was done in 2014 (Herm, Callsen-Bracker, & Kreis). The authors of this study mention that the market values from Transfermarkt have a good reputation in the sports industry and that the values have a high economic relevance due to the fact that the values are actually used in transfer and salary negotiations. Subsequently they look at how well a crowd is able to perform the complex task of human capital valuation by comparing the market values taken from Transfermarkt with the actual transfer fees that are being paid for these players. They find that the estimated market values are related to actual transfer fees and may serve as predictors.

Both studies are very interesting as they use data from the same source, Transfermarkt. These two studies as well as all the other studies that were discussed can be used to form the hypotheses which are described in the next chapter. The results from the discussed studies help us form expectations on the answers to these hypotheses.

5. Hypotheses

Now that we have looked at the previously done research, we can formulate the hypotheses of this research. The answers to these hypotheses can help us answer the research question. The hypotheses are based on earlier done research as well as through reasoning. Expectations of each hypothesis are made according to the findings discussed in the theoretical framework.

As we are trying to predict match outcomes using crowd estimated data, it is interesting to see whether the wisdom of the crowd effect is present. No research has been done on whether this effect can be found with market values, thus any findings on this matter would be completely new contributions to the literature.

Recall that according to Surowiecki, there are four different requirements that are needed to be satisfied for the presence of the wisdom of the crowd effect (Surowiecki, 2004). These four requirements are diversity of opinion, independence of opinion, decentralization and lastly aggregation. We need to see whether the market values from Transfermarkt satisfy these requirements. This can be done by comparing the data to each requirement and see if each of them is met. If these requirements are not met, then it is considerably less likely that these market values are close to the true values and hold any predictive power. They are then also not more accurate than valuations made by a few experts. The hypothesis that allows us to test this is formulated as follows:

- **H1**: The market values from Transfermarkt do not satisfy the requirements for the wisdom of the crowd effect
- **H1a**: The market values from Transfermarkt do not satisfy the requirements for the wisdom of the crowd effect

As mentioned earlier, Herm, Callsen-Bracker and Kreis found that the market values are good proxies for the football players' real value (2014). The fact that these market values hold some useful information leads me to believe that the wisdom of the crowd effect can indeed be found with market values and that hypothesis 1 will be rejected.

After we have determined whether it is likely that the wisdom of the crowd effect can be found with the market values, we have to see whether these market values have a significant effect when we add them to a regression. If the market values are significant, even when we control for different variables, then it is more likely that the market values have an actual causal effect on the match outcomes. The hypothesis that allows for testing this is formulated as follows:

- **H2**: Market values that are used as a covariate in a regression cannot help predict match outcomes
- **H2a**: Market values that are used as a covariate in a regression can help predict match outcomes

As we have seen earlier in the theoretical framework, it has been found that market values can be used to predict international football matches (Peeters, 2016). But no study has been done on whether this also is the case for domestic football leagues. As mentioned earlier, the market values are good proxies for the player's real value. This value is based on multiple factors including the quality of a player. It seems logical that a team with a higher average quality has a higher likelihood of winning. So the expectation is that market values can indeed help predict match outcomes and that this is a positive relation.

When we determined whether the market values have significant effect, we need to see how these models perform in comparison with other models. Otherwise we would not be able to say whether the model is good at making predictions. I compare the market value models with models based on the ELO rating (Hvattum & Arntzen, 2010). A model based on the ELO rating has proved to be able to predict match outcomes. So comparing the models based on the market values with these ELO based models gives us an idea how the market value based models perform. The hypothesis that allows me to test this is formulated as follows:

- **H3**: A model based on market values does not predict match outcomes better than a model based on ELO ratings
- **H3A**: A model based on market values predicts match outcomes better than a model based on ELO ratings

From the theoretical framework we learned that international match predictions of a simple model based on market values can outperform predictions of methods using the FIFA ranking or the ELO ratings (Peeters, 2016). The expectation is that this will also be the case for domestic football and that hypothesis 3 will be rejected.

After we have compared the market value models with models based on the ELO rating, it is also interesting to see how the market value models stack up when we compare them to the predictions made by bookmakers. If we are able to predict the match outcomes much better than the bookmakers, then this would allow us achieve a profit when we bet on these outcomes. The hypothesis that allows me to test this is formulated as follows:

- **H4**: A model based on market values does not predict match outcomes better than the bookmakers
- **H4A**: A model based on market values predicts match outcomes better than the bookmakers

As we have seen earlier, Hvattum and Arntzen found that an ELO rating can help predict match outcomes, but that it doesn’t outperform betting according to the probabilities based on the market odds (2010). Furthermore, we saw that a model based on market values outperforms a model based on the ELO rating (Peeters, 2016). If we find the same results as Peeters that a model based on market values can outperform a model based on ELO ratings, then it seems possible that a model based on market values can outperform the bookmakers. So the expectation is that hypothesis 4 will be rejected.

The next chapters will discuss what data and methods are used to test these hypotheses.

6. Data

Now that we looked at the previously done research and formulated the hypotheses of this research, we need to look at all the data required to test the hypotheses. I will first discuss the general match data that is used mainly for estimating the models. After that I more specifically discuss the market value data and lastly I discuss the betting data.

As mentioned earlier, this research focusses on the English Premier League. This football league is the highest football division in England. A total of twenty teams are active in the league each season. All teams play twice against all other teams, once at home and once away totalling 380 matches each season.

The match data is taken from the official Premier League website and from Football-Data (English Premier League, 2016; Football-Data, 2016).

The data that is collected includes the following:

- Match dates
- Home team
- Away team
- Goals home team
- Goals away team
- Match outcome

The data consists of a total of 1520 matches that were played in the 2012-13, 2013-14, 2014-15 and 2015-16 season of the English Premier League.

The market values of the players are taken from Transfermarkt (Transfermarkt, 2016a). As stated earlier, this community of football fans collectively estimate player market values. Members of this site come from many different countries and support many different clubs. Each member can post his valuation of a certain player. A message board allows the members to discuss the valuations and the players. All these valuations are combined to a single market value for each player. After a couple of months, the newly posted valuations are again combined to an updated market value, this to make sure players always have a representative value. For most players, two different values are used. One value for the beginning of the 2015-16 season and an updated value for all matches played after February 9, 2016.

In total, values of 683 players are collected. The average player value before February 9 is €6.75 million and €6.97 million after February 9. The lowest player value is 0 and the highest is €70 million. A player can have a value of 0 when he is just added to the system and hasn't received many valuations. This is mainly the case for young players that are added to the main squad in the course of the season.

The betting data for the 2015-16 season that is used is gathered from Football-Data. This website provides betting odds of multiple betting sites. Odds of the following seven bookmakers are used:

- Bet365
- Bet&Win
- Interwetten
- Ladbrokes
- Pinnacle Sports
- VC Bet
- William Hill

This means that for each match that is played, 7 odds are available for each outcome totalling 7980 odds.

7. Methodology

I will now discuss how the gathered data is used to answer the hypotheses. I start with explaining how the ELO ratings are calculated in section 6.1. After the ELO calculation I explain in chapter 6.2 which other variables are used in this research and look at the descriptive statistics in chapter 6.3. In chapter 6.4 I explain what kind of model will be used and what variables are added to each model. After discussing the models, I explain how the odds of the bookmakers are used and discuss the calculation of the Brier scores in chapters 6.5 and 6.6. Lastly in chapter 6.7 I explain how the monetary value of the models is examined by simulating betting on 340 different matches.

7.1 ELO rating

Earlier in the theoretical framework I discussed the study by Hvattum and Arntzen on match prediction using ELO ratings (2010). The models they constructed proved to be able to predict match outcomes, so it is interesting to compare the performance of the market value models with the performance models using ELO. But before we can do that, we need to have ELO ratings for each team.

The ELO ratings are constructed in the same way as the goal based ELO model made by Hvattum and Arntzen. The ELO rating system was originally designed to measure the strength of chess players (Elo, 1978). Based on results of a set of preceding matches, each team can be assigned an ELO rating as a measure of the team's current strength.

We start with defining ℓ_0^H and ℓ_0^A as the scores of the home and the away team at the beginning of the match. We then assume that, on average, for the match in question, the home and away teams will score γ^H and γ^A respectively where a win gives a score of 1, a tie a score of 0.5 and a loss a score of 0. The predicted scores are calculated with the formulas

$$\gamma^H = \frac{1}{1+c^{(\ell_0^A-\ell_0^H)/d}} \qquad (1)$$

and

$$\gamma^A = 1 - \gamma^H = \frac{1}{1+c^{(\ell_0^H-\ell_0^A)/d}}. \qquad (2)$$

Formulas 1 and 2 both take the scores of the home and the away team at the beginning of the match into account to calculate the predicted scores each team will have after the match.

After the match is played, we can calculate which scores that each team actually have. The score resulting from the actual match outcome for the home team is given by

$$\alpha^H = \begin{cases} 1 & \text{if the home team won,} \\ 0.5 & \text{if the match ended in a tie, or} \\ 0 & \text{otherwise.} \end{cases} \qquad (3)$$

The actual score for the away team is given by $\alpha^A = 1 - \alpha^H$. After each match the ELO scores are updated. The new rating for the home team is then

$$\ell_1^H = \ell_0^H + k(\alpha^H - \gamma^H) \qquad (4)$$

So the score before the match, ℓ_0^H or ℓ_0^A, is increased with factor k times the difference between the actual score. After that the predicted score is subtracted.

The new rating for the away team is calculated in the same way as for the home team, but now the respective away team scores are used. This new calculation is done after every match to ensure that each team has the appropriate score.

The score we have now is called the basic ELO rating by Hvattum and Arntzen (2010). They further changed the model by letting k depend on the goal difference, so a 3-0 win is rewarded more strongly than a 2-1 win.

They call this version of the model the goal based ELO rating. Comparing the basic ELO rating with the goals based ELO rating showed that the goal based ELO rating is a better proxy for a team's strength (Hvattum & Arntzen, 2010). This means that it is better to use the goals based ELO rating. We can calculate the goal based ELO ratings by making some changes to k,

$$k = k_0(1 + \delta)^\lambda \qquad (5)$$

With δ being the absolute goal difference and we take k_0 > 0 and λ > 0 as fixed parameters. Hvattum and Arntzen found that setting c = 10, d = 400, k_0 = 10 and λ = 1 gives the best results. This gives

$$k = 10(1 + \delta) \qquad (6)$$

The full goal based ELO equation for the home team is then

$$\ell_1^H = \ell_0^H + \left(10(1+\delta)\right)(\alpha^H - \gamma^H) \qquad (7)$$

Hvattum and Arntzen note that the calculation of the ELO ratings after a specific set of matches requires some initial ratings to be provided for each team. The ratings cannot be expected to be reliable indicators of strength until a sufficient number of past match results have been taken into account. To account for this, the seasons 2012-13, 2013-14 and 2014-15 are used for the initial calculations of the ELO rating.

To be able to use the ELO ratings in the models, I take the rating difference prior to the match,

$$ELO\_Diff = \ell_0^H - \ell_0^A \qquad (8)$$

When ELO_Diff is positive, then the home team is deemed stronger, while a negative value means that the away team is deemed stronger.

7.2 Variables

Now that the ELO ratings are computed, I'll give an overview of all the variables that are constructed using the data:

As we are predicting match outcomes, we need a dependent variable to predict these outcomes. For this reason, I construct the following variable:

- Match_Outcome: Can take one of three values. It takes the value 1 when the home team wins, it takes the value 0 when the match ends in a draw and it takes the value -1 when the away team wins.

Recall that we want to use the market value data from Transfermarkt. To incorporate this data into the model, I construct the following variable:

- TeamDiff: The difference in player market values between the two teams, both the starting 11 as the players on the bench are included. The difference in the team market values is calculated by subtracting the away team's market values from the home team's market values. A positive number means that the combined home team's market value is higher, while a negative number means that the combined away team's market value is higher.

Remember that Louzada, Suzuki and Salasar used a number of different variables in their model (2014). One of these variables was the goal difference in a match. This shows that the amount of goals has predictive power on the match outcome. I try to incorporate this useful information using the following four variables:

- HomeAvgScored: The average amount of goals scored by the home team.
- HomeAvgConceded: The average amount of goals conceded by the home team.
- AwayAvgScored: The average amount of goals scored by the away team.
- AwayAvgConceded: The average amount of goals conceded by the Away team.

As we base the models on the market values, changes in the team roster can have a big influence on the TeamDiff variable. Most changes in the teams take place in the summer transfer window. This is also the period when most coaches are still experimenting with their starting lineup. Combining this, it appears to be a good idea to include a dummy variable in the models for this period. I construct the following variable:

- EarlySeason: A dummy variable that takes the value 1 when a match is played before the closing of the summer transfer window (September 1) and 0 otherwise.

This variable allows me to control for any time specific effects on the match outcome.

Lastly we have the variable we constructed earlier using the ELO ratings:

- ELO_Diff: The ELO rating difference between the two teams at the start of the match. Calculated by subtracting the away team ELO rating from the home team ELO rating. A positive value means that the home team is deemed stronger, while a negative value means that the away team is deemed stronger.

This variable allows me to make models to compare the market values models with. Chapter 6.4 will discuss these models more in-depth, but first we will have a look at the descriptive statistics of the variables we just constructed.

7.3 Descriptive statistics

We start with looking at the outcome and early season dummy variables. An overview of these variables can be seen in table 2. We can see that there were more home team wins than away team wins, but that there were more away team wins than home team wins during the early season. The early season consists of 40 matches in total.

Table 2: An overview of all the outcomes of the matches played in the 2015-16 season.

Outcome	Early season		Total
	No	Yes	
Away win	99	17	116
Draw	93	14	107
Home win	148	9	157
Total	340	40	380

Next we look at the goal related variables which can be seen in table 3. This table shows the mean amount of goals scored and conceded for both the home as the away team during the course of the season. All four variables have a mean close to 1.3 and a standard deviation of approximately 0.5.

Table 3: An overview of the four goal related variables.

Variables	Observations	Mean	Std. deviation
HomeAvgScored	380	1.293611	0.4846748
HomeAvgConceded	380	1.288257	0.5073044
AwayAvgScored	380	1.285884	0.4965712
AwayAvgConceded	380	1.290023	0.4879297

In table 4 the variables TeamDiff and ELO_Diff can be seen. The mean difference in team market value is 1.88, which means that, on average, the home team had a slightly higher team value than the away team over the course of the season. The variable ELO_Diff has a mean of -1.66, which means that, on average, the away team had a slightly higher ELO rating than the home team over the course of the season. But when we take the standard deviations of both variables into account, we can see that the reported mean differences are very minor. This means that no significant difference can be observed in the mean values for the home and away teams for both variables.

Table 4: An overview of the TeamDiff and ELO_Diff variables.

Variable	Observations	Mean	Std. deviation	Minimum	Maximum
TeamDiff	380	1.881579	159.2824	-406	439.4
ELO_Diff	380	-1.655752	171.4914	-428.2757	422.4511

Now that we have looked at the descriptive statistics of all the variables, we can start estimating the models. In the next section I discuss which kind of model and which variables the models consist of.

7.4 Multinomial logistic models

The models that are used for the analysis are multinomial logistic regressions. A multinomial logistic regression is a model that generalizes a logistic regression so it can predict probabilities of more than two different possible outcomes. The dependent variable has to categorically distributed, which means that is isn't necessary to have a clear order in the outcomes of the dependent variable. When we look at the dependent variable, Outcome_Coded, we can see that it can take on one of three different values and that there is no clear order among these values.

Existing literature is focused on predicting home wins (Hvattum & Arntzen, 2010; Peeters, 2016). In those studies, a clear order can be found in the outcomes of the dependent variable. This means that a probit or logit model is more suitable to help predict the match outcomes. These two types of models can exploit the ordered nature of the dependent variable. A part of the contribution of my research is that I try to predict the match outcomes in a different way, I predict both home as away wins. By using this different approach, the clear order in the dependent variable, as is found in other research, is not there anymore. A multinomial logistic regression model is thus used to predict the match outcomes.

A total of four different multinomial logistic regressions are modelled to help answer the hypotheses, these can all be seen below in table 5.

Table 5: An overview of the independent and dependent variables of the four different models.

	Model 1	Model 2	Model 3	Model 4
Dependent variable	Outcome_Coded	Outcome_Coded	Outcome_Coded	Outcome_Coded
Independent variables	TeamDiff	TeamDiff HomeAvgScored HomeAvgConceded AwayAvgScored AwayAvgConceded EarlySeason	ELO_Diff	ELO_Diff HomeAvgScored HomeAvgConceded AwayAvgScored AwayAvgConceded EarlySeason

Model 1 consists of just TeamDiff as an independent variable while model 2 also has the goals scored/conceded and the EarlySeason variables. Model 3 only has ELO_Diff as independent variable while model 4 also has the goals scored/conceded and EarlySeason variables.

As mentioned earlier, the EarlySeason dummy is included to control for the possible big changes in the team rosters. During the transfer window, many clubs sell and contract new players, which can lead to big changes in the team market values. Besides that, the beginning of the season also is the time where managers are often still experimenting with the starting lineups. Having a dummy for this period allows me to control for this possible effect on the match outcome. No earlier research has added a control variable for this period, so it is interesting to see if it holds any significance in predicting match outcomes.

The goal scored and conceded variables are used to control for a team's attacking or defending prowess. If a team has a high average amount of goals scored, then this team is more likely to have good attackers. If a team concedes a low amount of goals, then it is more likely that this team has a good defence. Because the amount of goals you score and concede impact the match outcome, it is good to control for these effects. Previously done study also incorporated goals in a model (Louzada, Suzuki, & Salasar, 2014).

Models 1 and 3 allow me to see whether TeamDiff and ELO_Diff have any explanatory power on their own. Models 2 and 4 allow me to see whether TeamDiff and ELO_Diff still have explanatory power when more variables have been controlled for and whether the signs stay consistent.

To compare the models likelihood ratio tests are done to compare model 1 with model 2 and model 3 with model 4. Model 2 and 4 are compared by looking at the Akaike Information Criterion (AIC). The AIC is a measure of the relative quality of a model, where a lower score means a better quality. Hypothesis 2 will be tested using the results of models 1 and 2.

The significance levels that are used in this research are 1, 5 and 10 percent significance levels. This means that if a variable has a lower p value than one of the three significance levels, then I deem the variable to be significant.

A model with both the market values as the ELO rating in a single model was also estimated. But a proper estimation of such a model proved to be impossible due to multicollinearity problems.

7.5 Odds

To be able to compare the predictions of the market value and ELO models with the bookmakers, we need to elicit the outcome probabilities of the bookmakers. For each match outcome, I compute the average odds of the seven bookmakers and I take their inverse. To calculate outcome probabilities, the inverse average odds are normalized so that the sum of probabilities over the three outcomes is equal to one. We end up with 3 probabilities, one for each match outcome, which can be used for comparison with the predicted probabilities of the four multinomial logistic models.

7.6 Brier scores

To compare the predictions of the different models with each other and with the bookmakers, I make use of Brier scores. Brier scores are commonly used to compare forecasting performance between different models and were first used for weather forecasts (Brier, 1950). They have been used for match outcomes by both Hvattum and Arntzen (2010) and Peeters (2016). The Brier scores are calculated as follows:

$$BS_{ijt} = \frac{1}{3}((f_{ijt-1}(win) - win_{ijt})^2 + ((f_{ijt-1}(draw) - draw_{ijt})^2 + ((f_{ijt-1}(loss) - loss_{ijt})^2) \quad (8)$$

f_{ijt-1} represents the forecasted probability for each outcome of a match between team i and j. This match outcome is based on all the information that is available before the start of the match. win_{ijt}, $draw_{ijt}$ and $loss_{ijt}$ take value 1 when a match ends in a win, draw or a loss for team i or 0 otherwise. Lower Brier scores mean better forecasting performance (Peeters, 2016).

Brier scores are calculated for the four multinomial logistic models and for the bookmakers. These scores are tested for univariate, bivariate and multivariate normality. This is done to decide whether it is acceptable to look at the t-tests. After these tests, the Brier scores are compared with paired t-tests to determine which models makes the most accurate predictions.

Hypotheses 3 and 4 are tested using these Brier scores.

7.7 Monetary value

Lastly I want to examine whether any money can be earned by using one of the four models. To do this I bet on every outcome for which the probabilities indicate a value bet. A value bet occurs when the expected value minus the stake is positive (Hvattum & Arntzen, 2010). The stake that is used equals 1 unit.

A value bet is calculated as follows:

$$\Sigma\left[R_{y_{ijt}}\right] = odd\left(y_{ijt}\right) * f_{ijt_{-1}}\left(y_{ijt}\right) - 1 \quad (9)$$

with the expected return being $R_{y_{ijt}}$ for a bet on an outcome y_{ijt} in the match between teams i and teams j at time t with the forecasted probability of each match outcome being equal to

$f_{ijt_{-1}}$. If the outcome of equation 9 is positive, then this bet is a value bet and a stake of 1 unit will be bet on this outcome. It is possible that multiple outcomes of a single match are deemed as a value bet. In such a case, multiple bets are placed for that specific match, one for each value bet.

The predicted probabilities of all four models are repeatedly recalculated using the data of the earlier matches. This is done to ensure that we are not making predications based on information that would not have been available at the time the match was played. If we would do this, then the results would be positively skewed and we would examine higher returns.

To prevent this, I start the prediction of the match outcomes from match 41 onwards till match 380. This is done to supply the four models with enough data so the models are calibrated. The first forty matches were chosen because these were all the matches that were played in August, the first month of the football season. If we would not calibrate the models, then the first predictions will be very inaccurate due to insufficient data to predict the outcomes with.

The predictions for match 41 are made by using the data of the first forty matches. Subsequently, the predictions for match 42 are made using the data of the earlier forty-one matches. This is done until the predictions of match 380 are made using the data of the 379 earlier played matches.

These new probabilities are used to calculate which match outcomes should be bet on. After the calculations I cross reference the bets with the actual match outcomes. If one of the match outcomes that are bet on is equal to the actual match outcome, then the highest odd for that match outcome is taken. From this odd I subtract the stake(s) that were put on the match. What we end up with is the maximum possible profit that is achieved by betting on all value bets of a maximum of 340 matches. The average maximum return is then calculated by dividing the total maximum profit by the total number of bets that have been done. We then can see what the average maximum return is for a bet of 1 unit.

Now that the methodology of this research has been discussed, we can go to the actual results of all the tests.

8. Results

Now that the methodology of this research has been discussed, we can go to the actual results of all the tests. In this section I will present all the findings of this research and answer the research questions. I start with comparing the information of Transfermarkt with the requirements of the wisdom of the crowd effect. After this the different models are discussed and compared.

8.1 Wisdom of the crowd

First up I discuss the results that are required to answer the first hypothesis. Hypothesis 1 states that the wisdom of crowd effect cannot be found with market values. As mentioned earlier, this hypothesis is answered by comparing the market values of Transfermarkt with the requirements of the wisdom of the crowd.

As stated earlier in the theoretical framework, four requirements for the wisdom of the crowd effect to take place are mentioned by Surowiecki (2004). These are:

- Diversity of opinion
- Independence of opinion
- Decentralization
- Aggregation

Diversity of opinion

We start with the first requirement by looking whether there is any diversity among the members that give valuations to the football players. We do this by taking a look at the discussion boards of a few players of different teams.

Rihad Mahrez, Leicester City FC:

When we look at the discussion board of Mahrez, we see that there are 23 different valuations. We see that users support many different clubs from different countries. Hull City, Besiktas JK, Olympique Marseille et cetera are some of the different clubs the users support. It is very likely that there is diversity among these members. This can also be seen from the fact that they not all estimate the player at the same value (Figure 1, Appendix).

Nemanja Matic, Chelsea FC:

When we look at the discussion board of Matic, we see that there are again many members that support different clubs. There is also a considerable amount of diversity among the estimations of Matic's value, with many different opinions (Figure 2, Appendix).

Virgil van Dijk, Southampton FC:

Looking at the discussion board of van Dijk, we again see diversity among the clubs that the users support. They also don't have the same valuations (Figure 3, Appendix).

The above observations can be found for many different players. Unfortunately, no user statistics can be found on Transfermarkt[1]. So there is no recent data on the users of these message boards. The only data that is available are the most supported clubs by the users of Transfermarkt in 2013 (Peeters, 2016).

Table 6: top 15 supported teams by the users of Transfermarkt on 31/10/2013.

Rank	Team	League	Country	Number fans	% total fans
1	FC Bayern München	1.Bundesliga	Germany	5913	12.47%
2	Borussia Dortmund	1.Bundesliga	Germany	5147	10.91%
3	Hamburger SV	1.Bundesliga	Germany	2215	4.67%
4	FC Schalke 04	1.Bundesliga	Germany	2131	4.49%
5	SV Werder Bremen	1.Bundesliga	Germany	2077	4.38%
6	Borussia Mönchengladbach	1.Bundesliga	Germany	1893	3.99%
7	Galatasaray Istanbul	Süper Lig	Turkey	1414	2.89%
8	VfB Stuttgart	1.Bundesliga	Germany	1348	2.84%
9	1.FC Kaiserslautern	2.Bundesliga	Germany	1082	2.28%
10	1.FC Köln	2.Bundesliga	Germany	1006	2.12%
11	Eintracht Frankfurt	1.Bundesliga	Germany	970	2.05%
12	Fenerbache Istanbul	Süper Lig	Turkey	961	2.03%
13	1.FC Nürnberg	1.Bundesliga	Germany	822	1.73%
14	Hertha BSC	1.Bundesliga	Germany	784	1.65%
15	Hannover 96	1.Bundesliga	Germany	768	1.62%

What we can see from table 6 is that there seemed to be quite some diversity in the clubs that were supported by the users of Transfermarkt. Most of the German clubs come from different cities from all over Germany. It seems very likely that there is a difference between the

[1] I have requested more recent user data from Transfermarkt, but unfortunately this has yet to be supplied to me.

supporters of these clubs. And while table 6 shows outdated data, it still gives us an insight in the diversity of the members. Since then, Transfermarkt has only grown bigger with more users than back in 2013. One limitation of only looking at the supported teams is that we don't know which members valuate the players and specifically the players that play in the English Premier League.

Taking the top 15 supported teams in 2013 together with the finding that supporters of many clubs are commenting on the message boards, gives me reason to believe that the first requirement for the wisdom of the crowd effect is likely to be satisfied. While no recent user statistics are available, it still seems likely that the users that comment and valuate the players are diverse and thus have diverse opinions.

Independence of opinion

Next is the second requirement for the wisdom of the crowd effect, people's opinions should be independent from each other. The members of Transfermarkt are all single users, which means that it is not likely that they belong to any organizations. They are also all free to give their own valuations and opinions. As we saw earlier, the members support many different clubs. This makes it less likely that the players are overvalued by the majority of the users due to them belonging to the clubs they personally support. Study on the valuations of the members of Transfermarkt found no evidence that valuations of popular players are biased upwards (Peeters, 2016). This all leads me to the conclusion that there is indeed independence in the opinions of the members, which makes the second requirement fulfilled.

Decentralization

The third requirement of decentralization is somewhat more difficult to check than the previous two requirements. People should be able to specialize and draw on local knowledge. If we look at the discussion boards of the English Premier League, we see that there are many users that give valuations to many different players from multiple teams. It seems like these members are close followers of the English Premier League, which means they could have specialized with this specific competition. If this is the case, then these members are more likely to give good valuations of the players, as they can draw from their own specialized knowledge of the English Premier League. This means that the third requirement is also satisfied.

Aggregation

The way Transfermarkt works, is that all the estimations of the users are used to calculate a single market value for each player. The exact way this is done is not disclosed by Transfermarkt, but there is some formula they use to aggregate all the estimations. So the fourth requirement that there should be mechanism to turn all the estimations into a single estimation is satisfied.

8.2 Hypothesis 1

Summing up, it is likely that the market values from Transfermarkt satisfy all four requirements for the wisdom of the crowd effect to be present. It thus seems very likely that the wisdom of the crowd effect can be found with the market values of Transfermarkt. The null of hypothesis 1, which states that the market values from Transfermarkt do not satisfy the requirements for the wisdom of the crowd effect, is rejected. This is in accordance with the expectation that was stated in the hypotheses chapter. No other study has been done whether the market values from Transfermarkt fulfil the requirements of the wisdom of the crowd. But it has been found that the crowd estimated market values are related to actual transfer fees (Herm, Callsen-Bracker, & Kreis, 2014). Thus the results have in common that both find that there is useful information in the market values from Transfermarkt. It could be that the wisdom of the crowd is one of the reasons why Herm, Callsen-Bracker and Kreis found that the market values are related to the transfer fees. The users of Transfermarkt are able to collectively estimate the correct market values for the players.

8.3 Models

Now that we found that it is likely that the wisdom of the crowd effect can be found with the market values from Transfermarkt, we can go and estimate the models. An overview of these four models can be seen below in table 7.

Table 7: The four models, where *, ** and *** represent a ten, five and one percent significance level respectively.

		Model 1	Model 2	Model 3	Model 4
Away win	TeamDiff	-0.00197**	-0.00204**		
		(0.025)	(0.033)		
	ELO_Diff			-0.00162*	-0.00219*
				(0.054)	(0.059)
	HomeAvgScored		-0.016536		-0.02477
			(0.574)		(0.937)
	HomeAvgConceded		-0.11789		-0.28905
			(0.665)		(0.352)
	AwayAvgScored		-0.12246		-0.22830
			(0.662)		(0.448)
	AwayAvgConceded		-0.10868		0.02142
			(0.711)		(0.946)
	EarlySeason		0.04832		0.02199
			(0.907)		(0.985)
	Constant	0.03338	0.68781	0.004103	0.67183
		(0.808)	(0.256)	(0.977)	(0.272)
Draw		(base outcome)	(base outcome)	(base outcome)	(base outcome)
Home win	TeamDiff	0.00169**	0.00015		
		(0.039)	(0.873)		
	ELO_Diff			0.00338***	0.00214*
				(0.000)	(0.074)
	HomeAvgScored		0.30906		0.10458
			(0.284)		(0.739)
	HomeAvgConceded		-0.67633**		-0.31033
			(0.023)		(0.357)
	AwayAvgScored		-0.97132		-0.67167**
			(0.002)		(0.047)
	AwayAvgConceded		0.52460*		0.22279
			(0.077)		(0.505)
	EarlySeason		-1.53873***		-1.31233**
			(0.004)		(0.010)
	Constant	0.34997***	1.51223**	0.31886**	1.27862*
		(0.006)	(0.031)	(0.014)	(0.062)
	AIC	2.135	2.105	2.071	2.087

First I discuss all four models individually. After that I compare the models by means of likelihood ratio tests and by looking at the Brier scores. Hypothesis 2, 3 and 4 will be discussed and after that the monetary value of the models will be examined.

Model 1

When we look at model 1, we can see that the differences between the team values are significant for both an away as a home win relative to the base outcome, ceteris paribus. Both are significant at a five percent level. An increase of the difference in the team values by 1 unit lowers the multinomial log-odds of an away win relative to a draw with 0.00197 units, ceteris paribus. While an increase of the difference in the team values with 1 unit increases the multinomial log-odds of a home win relative to a draw with 0.00169 units, ceteris paribus.

Model 2

When we add the control variables to model 1 we end up with model 2. We can again see that the difference between the team values is significant at a five percent level for an away win relative to a draw, ceteris paribus. An increase of the difference between the teams with 1 unit decreases the multinomial log-odds of an away win relative to a draw with 0.00204 units, ceteris paribus. None of the control variables are significant for an away win.

When we look at the home win, we can see that the difference between the team values is no longer significant. Of the control variables, the home team's average conceded goals, the away team average conceded goals and the early season dummy are all significant at a five, ten and one percent level respectively relative to a draw, ceteris paribus.

An increase in the amount of goals conceded by the home team with 1 unit decreases the multinomial log-odds of a home win relative to a draw with 0.67633 units, ceteris paribus. An increase in the amount of goals conceded by the away team with 1 unit increases the multinomial log-odds of a home win relative to a draw with 0.52460 units, ceteris paribus. If the early season dummy equals 1, then the multinomial log odds of a home win is 1.53873 lower relative to a draw, ceteris paribus.

Model 3

In the model with only the difference in ELO rating as independent variable, we can see that the ELO rating difference is significant for both the away win as the home win, relative to a draw. This is significant at a ten and one percent respectively, ceteris paribus. This means that an increase in the ELO rating difference with 1 unit decreases the multinomial log-odds of an away win relative to a draw with 0.00162 units, ceteris paribus. An increase in the ELO rating difference with 1 unit increases the multinomial log-odds of a home win relative to a draw with 0.0.00338 unit, ceteris paribus.

Model 4

When we add the control variables to model 3 we get model 4. Looking at the away win, we can see that the ELO rating difference has a significant effect at a ten percent level relative to a draw, ceteris paribus. An increase in the ELO rating difference with 1 unit decreases the multinomial log-odds of an away win relative to a draw with 0.00219 units, ceteris paribus. The control variables are not significant.

For the home win outcome, the difference in the ELO ratings has a significant effect at a ten percent level. This means that an increase of the ELO rating difference with 1 unit increases the multinomial log-odds of a home win relative to a draw with 0.00214 units, ceteris paribus. The average goals scored by the away team and the early season dummy are both significant at a five percent level. An increase of the average amount of goals scored by the away team with one unit decreases the multinomial log-odds of a home win, relative to a draw with 0.67167 units, ceteris paribus. When the early season dummy equals 1, the multinomial log-odds of a home win decrease with 1.31233 units relative to a draw, ceteris paribus.

8.4 Likelihood ratio tests

Now that all models have been discussed separately, we can go ahead and compare the models with each other. As explained in the methodology section, the model comparison is done by means of a likelihood ratio test.

When we compare models 1 and 2, the likelihood ratio test shows that the Prob > chi2 = $p<0.001$. This means that adding control variables to model 1 results in a statistically significant improvement in model fit. This improvement is significant at a one percent level.

When we compare model 3 and 4, the likelihood ratio test reports that Prob > chi2 = $p<0.001$. This means that adding control variables to model 3 results in a statistically significant improvement in model fit. This improvement is significant at a one percent level.

For hypothesis 3, we need to compare models 2 and 4 with each other. Unfortunately, models 2 and 4 cannot be compared using a likelihood ratio test. This is due to them not being nested models of each other. To compare them we can look at the AIC. The AIC of model 2 is 2.105

while the AIC of model 4 is 2.087. Model 4 has a lower AIC than model 2 and is, according to the AIC, a better model than model 2. This difference however, cannot be tested for significance.

8.5 Hypothesis 2

Now that the models are discussed and compared with a likelihood ratio test, we can answer hypothesis 2. Recall the second hypothesis: Market values that are used as a covariate in a regression cannot help predict match outcomes.

We found that the difference between the team market values has a significant effect on the outcome in both models. In model 1, the difference is significant at a five percent level for both the home win as the away win outcome. In model 2 the difference is significant at a five percent level for the away win outcome. The likelihood ratio test shows that adding the control variables to model 1 results in a statistically significant improvement in model fit.

Taking this all together, it seems evident that the market values from Transfermarkt can help predict match outcomes, hypothesis 2 is rejected. This is in agreement with the result that was expected and it is also in agreement with the findings by Peeters (2016). Peeters found that market values can help predict match outcomes. Thus we reach the same conclusion, but there are a few differences between his and my study.

First off, the type of model that is used is different. Peeters uses both ordered probit and OLS regressions in his research. As mentioned earlier, I only make use of multinomial logistic regressions. His models are focused on predicting the outcome from the view of the home playing team, so there is a clear order to be observed in the dependent variable he uses. I on the other hand, try to predict the match outcome from no specific view. This means that the dependent variable I use is not ordered. Another difference between the two studies is which independent variables are used. I control specifically for the average amount of goals scored and conceded by both the home as the away team. Peeters incorporates the amount of goals a bit differently. He first relates the goals difference between the two teams to the explanatory variables, after which he estimates a new model to link the predicted goal difference with the match outcome. Another difference is that I control for the first month of the season with a dummy. As can be seen table 6 in the appendix, this dummy appears to have a significant effect on the match outcome. The last difference regarding the used variables is how the market values from Transfermarkt are used. Peeters adds them as logarithmic average team values, while I add the market value difference between the two teams.

Lastly his research is focused on international football while I focus on domestic football. The fact that both studies have the same findings gives me reason to believe that market values can help predict both international and domestic football matches.

8.6 Brier scores

Now that we have discussed the models and compared them with the likelihood ratio tests, we can start comparing the predictions of the models. As explained earlier in section 6.6, the predictions are compared by means of Brier scores. The computed Brier scores of each model and the bookmakers can be seen in the table below.

Table 8: An overview of the mean Brier scores of the four models and the bookmakers.

Brier Scores		
	Observations	Mean
Model 1	380	0.2124818
Model 2	380	0.2028459
Model 3	380	0.2047513
Model 4	380	0.2010306
Bookmakers	380	0.2062059

A lower Brier score means a more accurate prediction. We can see that model 1 has the highest mean Brier score and model 4 has the lowest mean Brier score. We can thus see that there are some differences between the Brier scores of the different models. We need to test the differences for significance to be sure these differences are not just by chance. Testing the differences for significance allows us to say whether there is really a difference in accuracy between the models. This can be done using t-tests, but we first need to investigate whether these t-tests are the appropriate tests to use. It is possible that we find results that are driven by chance, because we are doing multiple comparisons.

We can do this by doing an omnibus test. With this test we check for the equality of all comparisons simultaneously. If we reject the null hypothesis, then this means that some of the models are not equal to each other. If this is the case, then we can take a look at the individual t-tests. The results of the omnibus test can be found in the appendix (Table 8, 9 & 10, Appendix). The results of the omnibus test show that the null hypothesis is rejected and that some of the models are not equal to each other. We can thus use t-tests to compare the Brier scores.

A total of five paired t-tests are done on the Brier scores. Model 1 is compared with model 2, model 3 is compared with model 4, model 2 is compared with model 4, model 2 is compared with the bookmakers and lastly model 4 is also compared with the bookmakers. The t-tests comparing the different models can be seen in table 12.

Table 12: The results of the t-tests comparing the Brier scores.

Paired t-tests

	Observations	Mean	Mean(diff) < 0	Mean(diff) ! = 0	Mean(diff) > 0
Model 1	380	0.2124818	0.999	0.002	0.001
Model 2	380	0.2028459			
Model 3	380	0.2047513	0.9649	0.07002	0.0351
Model 4	380	0.2010306			
Model 2	380	0.2028459	0.9094	0.1812	0.0906
Model 4	380	0.2010306			
Model 2	380	0.2028459	0.1447	0.2894	0.8553
Bookmakers	380	0.2062059			
Model 4	380	0.2010306	0.0307	0.0615	0.9693
Bookmakers	380	0.2062059			

If we first compare model 1 with model 2, we can see that model 1 has a significantly higher brier score at a one percent level than the brier score of model 2. This means that model 2 predicts the match outcomes more accurately. This is in agreement with the result of the likelihood ratio test comparing model 1 with model 2. That likelihood ratio test found that the added control variables of model 2 result in a better model fit. So again mode 2 is showed to be preferred over model 1.

To be able to answer the last two hypotheses, we need to also compare the Brier scores of the other models.

If we compare the Brier scores of model 3 with model 4, we can see that model 3 has a significantly higher brier score at a five percent level than the brier score of model 4. This means that model 4 predicts the match outcomes more accurately.

Comparing model 2 with model 4, we can see that model 2 has a significantly higher brier score at a ten percent level than the brier score of model 4. This means that model 4 predicts the match outcomes more accurately

Comparing model 2 with the brier scores of the bookmakers, we can see that the null hypothesis that the mean difference is equal to zero is not rejected. This means that according to this test, model 2 and the bookmakers have the same accuracy predicting match outcomes.

Comparing model 4 with the brier scores of the bookmakers, we can see that the null hypothesis that the mean difference is equal to zero is rejected at a five percent level. This means that according to this test, model 4 predicts the match outcomes more accurately.

8.7 Hypothesis 3

Now that the Brier scores of all the models have been compared, we can go ahead and answer the last two hypotheses. First I will discuss hypothesis 3. This hypothesis states that a model based on market values does not predict match outcomes better than a model based on ELO ratings.

To test this hypothesis, we compare the ELO based model with the market value model. Recall that in section 7.2, we concluded that market value model 2 is preferred over model 1. The Brier score comparison between these models in the previous section reached the same conclusion. So for our comparison between the ELO and market value models we need to make use of model 2.

Now we have to find out whether model 3 or model 4 is the better ELO model. If we look back at table 7, we see that models 3 and 4 both show that the ELO rating difference has a significant effect on the match outcome. The likelihood ratio test comparing these two models found that the added control variables in model 4 result in a significant increase in model fit. The Brier scores show that model 4 is preferred over model 3. So we can conclude that model 4 is the better model of the two.

Now we know that models 2 and 4 are the best models, we can look at the comparisons between these two models. When we compared the AIC of models 2 and 4 in section 7.3, we found that model 4 has a slightly lower AIC. Which means that model 4 seems to be preferred over model 2, but this cannot be tested for significance. Thus we look at the t-test comparing the Brier scores of models 2 and 4. Model 2 has a significantly higher Brier score than model 4 at a ten percent level. This means that model 4 predicts the match outcomes more accurately than model 2.

We can conclude that hypothesis 3 is not rejected. We find that the ELO model outperforms the market value model. The result of this hypothesis is unexpected. As mentioned earlier, Peeters also researched the effect of market values on match outcomes (2016). He found that a model based on market values can outperform a model based on ELO ratings. So we reach different conclusions. There could be multiple why we find different results. First, as mentioned earlier, there is a difference in which control variables and which kind of models are used. The findings by Peeters could be biased due to not controlling for these variables. Secondly he looks at international football. It could be that a market value model is better at predicting international matches than domestic matches. If this is the case, then it could be possible that a market value model is outperformed by an ELO model. Thirdly I only use match data from the English Premier League, but there are a few other competitions in England and Europe where English teams compete in. These matches have not been taken into consideration so the results I find might be skewed. This is the case when teams for example rest their best players around UEFA Champions League matches. The team that plays in the English Premier League then has a lower combined market value than expected, which could lead to wrong predictions. Lastly, time specific effects could be influencing the results. Peeters most recent data are the WC 2014 qualifiers, while I use data on matches that were played more recently.

8.8 Hypothesis 4

As we compared the Brier scores of all the models, we can also answer hypothesis 4. Recall that this hypothesis states that a model based on market values does not predict match outcomes better than the bookmakers.

When we compare the Brier scores of model 2 with the Brier scores of the bookmakers, we find that there is no reason to think that the mean difference is different than zero. This means that according to the Brier scores, both models have the same accuracy predicting the match outcomes. Concluding, we do not reject hypothesis 4. Model 2 has the same accuracy predicting match outcomes as the bookmakers.

No previous research has been done examining this difference, so it is not possible to compare the results. But previously done research did find that an ELO based model cannot outperform betting according to the probabilities based on the market odds (Hvattum & Arntzen, 2010). It is interesting to see if we find the same results. When we compare the Brier score of model 4 with the Brier score of the bookmakers, we find that model 4 has a significantly lower Brier score at a five percent level. This means that according to the Brier scores model 4 more accurately predicts the match outcomes.

This result is contrary to the findings by Hvattum and Arntzen. An explanation of this result could possibly

8.9 Monetary value

It is interesting to see how much model 4 outperforms the bookmakers and how much money can be made in general while betting according to the models.

In table 13 you can see the average maximum return of each model when bets are placed on all value bets. A total number of 340 matches were bet on with each model. Model 1 found 567 value bets and has a negative average maximum return of 0.87%. We can conclude that betting according to model 1 means that, on average, you will lose money. Model 2 found a total od519 value bets and achieves an average maximum return of 7.45%. Again we find that model 2 is preferred over model 1. And the positive return also shows that the market values can help predict match outcomes, further evidence that hypothesis 2 should be rejected.

Table 13: The average maximum returns per model when only value bets are played.

Betting results

	Average Maximum Return	# Matches	# Value Bets
Model 1	-0.87%	340	567
Model 2	7.45%	340	519
Model 3	6.05%	340	570
Model 4	10.26%	340	501
Bookmakers	8.00%	340	533

Model 3 identifies a total of 570 value bets and nets an average maximum return of 6.05% for each unit that is bet. Model 4 found the least amount of value bets of all the models, but it achieves the highest average maximum return, 10.26%.

We can see that model 4 has a higher average maximum return than model 2. This further supports the earlier findings, that model 4 makes better predictions than model 2 and thus more evidence that hypothesis 3 should not be rejected.

Betting according to the probabilities of the bookmakers themselves finds a total of 533 value bets and nets an average maximum return of 8%. This is return is close to the maximum average return of model 2. This result is in accordance with the finding that model 2 and the bookmakers have the same accuracy predicting match outcomes.

While these returns seem promising, some caution has to be taken with respect to the average maximum returns of the different models. Due to the way the probabilities were calculated, it has not been taken into account that some matches might have been played at the same time. There might be cases were data was used of a match that was not yet played. For example, if match 50 and 51 were both played at 3pm, then the data of match 50 is still used to calculate the probabilities of match 51. For this reason, the listed average maximum returns might be lower or higher when this would be taken into account. I expect that the accuracy of the predictions might be a bit higher due to this. The predictions are made using more data than actually should be used. Due to the higher accuracy the average maximum return is likely a bit higher than it should actually be.

8.10 Research question

Now that all hypotheses have been answered we can go ahead with answering the research question.

The research question of this research is whether crowd estimated market values can help predict domestic football match outcomes. The findings show that there is evidence that crowd estimated market values can indeed help predict football match outcomes. I find that the wisdom of the crowd effect is likely to be found with the market values from Transfermarkt. The multinomial logistic models showed that the market values have a significant effect on the match outcomes and that adding the control variables improves the model according to the likelihood ratio test and the Brier scores. This means that a positive return can be achieved when betting on value bets according to a model with market values and control variables. Although it seems like such a model can help predict match outcomes, it doesn't seem to outperform a model based on ELO ratings. It performs lower according to the AIC and the Brier scores. Betting according to an ELO rating based model also yields higher average maximum returns than betting according to a market value based model.

9. Shortcomings & research recommendations

There are a few shortcomings of the research that I have done. These are things that could be improved upon in further research.

The first shortcoming of this research is that it only focusses on the English Premier League. Different results might be found when other domestic leagues are also studied. The characteristics of the English Premier League might be very different from other leagues. Characteristics like tactics, mentality, match schedules and type of players might all affect the results. Further research could incorporate more leagues in the research and try to control for more variables that might affect the match outcome.

The second shortcoming of this research is that it only focusses on a few seasons of the English Premier League. Three seasons are used to calibrate the ELO rating, but only one season worth of market values is used in the models. Seasonal effects might be driving the results. Incorporating more years' worth of data might yield different results.

Thirdly, I do not take any other competitions beside the English Premier League into consideration. Competitions like the FA Cup and the Football League Cup are played during the football season. It could be that teams rest their best players when a match of one of these competitions is close to a match in the English Premier League. Some teams also compete in the UEFA Champions League and Europa League. So different results might be found when we would also take all these other competitions into consideration.

As mentioned earlier, the average maximum returns of the different models might be skewed. It has not been taken into account that some matches are played at the same time while predicting the match outcomes. The returns might be higher or lower when this has been taken into account, with the expectation being that the average maximum returns will be lower. Further research could try to control for the time the matches were played while predicting the match outcomes

Lastly, as in almost all research there could be variables that have an effect on the dependent variable that have not been controlled for. Variables like weather during the match and the time the match is played could all have an effect on the match outcome. Controlling for these variables could improve the predictions of the models. Further research could control for more relevant variables to see if than the findings still persists.

10. Policy recommendations

Due to the findings of this research, some policy recommendations can be made.

Foremost the findings of this research prove like many other studies that there is value in crowd estimated data. With large data becoming more and more important, the fact that a large group of people can be accurate can be very useful in many situations. Governments could let people help decide on different issues that it finds difficult to decide on its own. An example of such an issue could be a traffic problem in a certain city. Letting the citizens collectively decide on a solution can possibly be the optimal solution.

Bookmakers could benefit from incorporating crowd estimated market values to their models. While a model with market values does not seem to outperform the bookmakers, combining them could prove beneficial. If these companies have a better understanding of what the match outcome will be, they can perhaps prevent big losses due to unforeseen match outcomes.

11. References

Baillon, A., Bleichrodt, H., Liu, N., & Wakker, P. P. (2016, April). Group decision rules and group rationality under risk. *Journal of Risk and Uncertainty, 52*(2), 99-116. doi:10.1007/s11166-016-9237-8

Brier, G. W. (1950, January). Verification of Forecasts Expressed in Terms of Probability. (J. E. Caskey, Ed.) *Monthly Weather Review, 78*(1), 1-3.

Charness, G., & Sutter, M. (2012). Groups Make Better Self-Interested Decisions. *Journal of Economic Perspectives, 26*(3), 157-176.

Ebner, S. (2013, July 2). *History and time are key to power of football, says Premier League chief*. Retrieved from The Times: http://www.thetimes.co.uk/tto/public/ceo-summit/article3804923.ece

Elo, A. E. (1978). *The Rating of Chessplayers, Past and Present.* Arco Publishing.

English Premier League. (2016). *Home*. Retrieved from Premier League: http://www.premierleague.com/

Football-Data. (2016). *Historical Football Results and Betting Odds Data*. Retrieved from Football-Data: http://www.football-data.co.uk/data.php

Galton, F. (1907, March 7). Vox Populi. *Nature, 75*, 450-451.

Gibson, O. (2015, February 10). *Sky and BT retain Premier League TV rights for record £5.14bn*. Retrieved from The Guardian: https://www.theguardian.com/football/2015/feb/10/premier-league-tv-rights-sky-bt

Herm, S., Callsen-Bracker, H.-M., & Kreis, H. (2014). When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community. *Sport management Review, 17*(4), 484-492. doi:doi:10.1016/j.smr.2013.12.006

Herzog, S. M., & Hertwig, R. (2011, February). The wisdom of ignorant crowds: Predicting sport outcomes by mere recognition. *Judgement and Decision Making, 6*(1), 58-72.

Hong, L., & Page, S. E. (2004, November 16). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. (W. J. Baumol, Ed.) *Proceedings of the National Acadamy of Sciences of the United States of America, 101*(46), 16385-16389. doi:doi:10.1073/pnas.0403723101

Hvattum, L. M., & Arntzen, H. (2010, July). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting, 26*(3), 460-470.

Koopman, S. J., & Lit, R. (2015, January). A dynamic bivariate Poission model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 178*(1), 167-186. doi: doi: 10.1111/rssa.12042

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. (B. H. Singer, Ed.) *Proceedings of the National Acadamy of Sciences of the United States of America, 108*(22), 9020-9025. doi:10.1073/pnas.1008636108

Louzada, F., Suzuki, A. K., & Salasar, L. E. (2014, April). Predicting Match Outcomes in the English Premier League: Which Will Be the Final Rank? *Journal of Data Sciences, 12*(2), 235-254.

Murr, A. E. (2015, July - September). The wisdom of crowds: Applying Condorcet's jury theorem to forecasting US presidential elections. *International Journal of Forecasting, 31*(3), 916-929. doi:10.1016/j.ijforecast.2014.12.002

Peeters, T. (2016, May). Testing the Wisdom of Crowds in the Field: Transfermarkt Valuations and International Soccer. 23.

Rayner, G., & Brown, O. (2016, May 2). *Leicester City win Premier League and cost bookies biggest ever payout*. Retrieved from The Telegraph: http://www.telegraph.co.uk/news/2016/05/02/leicester-city-win-premier-league-and-cost-bookies-biggest-ever/

Simmons, J. P., Nelson, L. D., Galak, J., & Frederick, S. (2011, June). Intuitive Biases in Choice versus Estimation: Implications for the Wisdom of Crowds. *Journal of Consumer Research, 38*(1), 1-15. doi:http://doi.org/10.1086/658070

Sunstein, C. R. (2006, September). When Crowds Aren't Wise. *Harvard Business Review*. Retrieved from https://hbr.org/2006/09/when-crowds-arent-wise

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.* Doubleday.

Transfermarkt. (2016a). *Transfermarkt homepage*. Retrieved from Transfermarkt: http://www.transfermarkt.com

Transfermarkt. (2016b). *THE TRANSFERMARKT CHRONICLE*. Retrieved from Transfermarkt: http://www.transfermarkt.com/intern/tmchronik

Wolfers, J., & Zitewitz, E. (2004). Precition Markets. *Journal of Economic Perspectives, 18*(2), 107-126.

12. Appendix

Figure 1: The message board on Transfermarkt where people discuss the market value of Rihad Mahrez.

MAHREZ, RIYAD (15 MIL. £, LEICESTER CITY) | #6 — APR 8, 2016 - 9:02 PM

KingPupka

Entries: 817
Entries worth reading: 3 / 3
IP: saved

Current estimation: **£22.95m** ~~£17.21m~~

Many journalists, managers and players think he should be player of the year, and for me, he is there in Top4 or 5 players at the moment (Kane, Vardy, Kante, Mahrez).

Alarm | Quote | Reply

MAHREZ, RIYAD (15 MIL. £, LEICESTER CITY) | #7 — APR 16, 2016 - 8:37 PM

kaanboy1

Entries: 23
Entries worth reading: 0 / 0
IP: saved

Current estimation: **£28.90m** ~~£21.68m~~

Player of the year in a small club !

Alarm | Quote | Reply

MAHREZ, RIYAD (15 MIL. £, LEICESTER CITY) | #8 — APR 23, 2016 - 11:24 PM

mirkaslan

Entries: 250
Entries worth reading: 0 / 0
IP: saved

Current estimation: **£26.35m** ~~£19.76m~~

With Kante, Vardy and Morgan, probably in team of the season of PL. His value is still low and needs to be higher.

Alarm | Quote | Reply

Figure 2: The message board on Transfermarkt where people discuss the market value of Nemanja Matic.

MATIC, NEMANJA (28,5 MIL. £, FC CHELSEA) | #5 JUN 10, 2016 - 4:49 PM

khfehome

Entries: 10
Entries worth reading: 0 / 0
IP: saved

Current estimation: **£25.50m** ~~£19.13m~~

Chelsea had a poor season, but Matic is one of the 2-3 players who looked good. And yet the price is too high for this player.

Alarm | Quote | Reply

MATIC, NEMANJA (28,5 MIL. £, FC CHELSEA) | #6 JUN 10, 2016 - 10:17 PM

KingPupka

Entries: 817
Entries worth reading: 3 / 3
IP: saved

Current estimation: **£32.30m**

Guys, just look ant the price of Busquets for example. Matic TM value fits definitely and there is no reason to change it. His best years are coming and his performances this season were good. There were other problems in Chelsea team, and Matic wasn't one of them.

Alarm | Quote | Reply

MATIC, NEMANJA (28,5 MIL. £, FC CHELSEA) | #7 JUN 11, 2016 - 12:47 PM

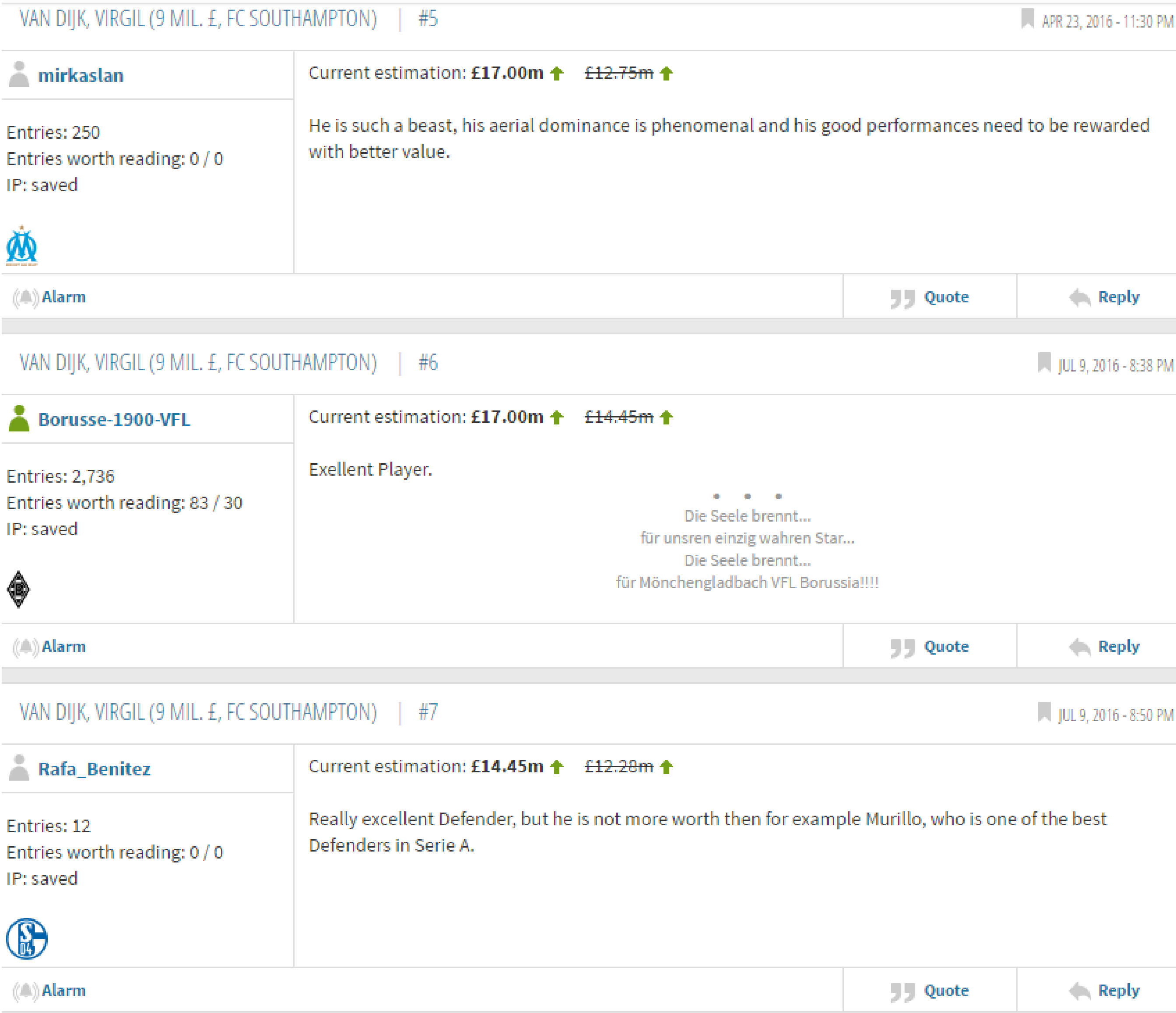
Chiki0

Entries: 14
Entries worth reading: 0 / 0
IP: saved

Current estimation: **£34.85m** ~~£26.14m~~

Its not cuz hes Serb, but hes great player, one bad season doesnt reduce his quality, i hope he moves to Juventus, he would be great boost to their team and their way of playing football suits him.

Figure 3: The message board on Transfermarkt where people discuss the market value of Virgil van Dijk.



VAN DIJK, VIRGIL (9 MIL. £, FC SOUTHAMPTON) | #5 — APR 23, 2016 - 11:30 PM

mirkaslan

Entries: 250
Entries worth reading: 0 / 0
IP: saved

Current estimation: **£17.00m** ~~£12.75m~~

He is such a beast, his aerial dominance is phenomenal and his good performances need to be rewarded with better value.

Alarm | Quote | Reply

VAN DIJK, VIRGIL (9 MIL. £, FC SOUTHAMPTON) | #6 — JUL 9, 2016 - 8:38 PM

Borusse-1900-VFL

Entries: 2,736
Entries worth reading: 83 / 30
IP: saved

Current estimation: **£17.00m** ~~£14.45m~~

Exellent Player.

• • •
Die Seele brennt...
für unsren einzig wahren Star...
Die Seele brennt...
für Mönchengladbach VFL Borussia!!!!

Alarm | Quote | Reply

VAN DIJK, VIRGIL (9 MIL. £, FC SOUTHAMPTON) | #7 — JUL 9, 2016 - 8:50 PM

Rafa_Benitez

Entries: 12
Entries worth reading: 0 / 0
IP: saved

Current estimation: **£14.45m** ~~£12.28m~~

Really excellent Defender, but he is not more worth then for example Murillo, who is one of the best Defenders in Serie A.

Alarm | Quote | Reply

Table 9: The results of the univariate normality tests of the four models and the bookmakers.

Univariate normality			
	Pr (Skewness)	Pr (Kurtosis)	Prob > chi2
Model 1	0.3226	0.0147	0.0342
Model 2	0.7988	0.0001	0.0012
Model 3	0.7684	0.0047	0.0215
Model 4	0.9586	0.0002	0.0024
Bookmakers	0.5324	0.0002	0.0017

Table 10: The results of the tests of the Doornik-Hansen test of the four models and the bookmakers.

Doornik-Hansen test for bivariate normality		
Models		Prob > chi2
Model 1	Model 2	0.0027
	Model 3	0.0604
	Model 4	0.0096
	Bookmakers	0.0008
Model 2	Model 3	0.0000
	Model 4	0.0065
	Bookmakers	0.0005
Model 3	Model 4	0.0000
	Bet	0.0323
Model 4	Bookmakers	0.0006

Table 11: The results of the multivariate normality tests.

Multivariate normality	
	Prob > chi2
Mardia mSkewness	0.0000
Mardia mKurtosis	0.0000
Henze-Zirkler	0.0000
Doornik-Hansen	0.0000