Aalto University
School of Business

Rating National Hockey League teams: the predictive power of Elo rating models in ice hockey

Bachelor's Thesis
Santeri Tenkanen
23.05.2019
ISM Program

Approved in the Department of Information and Service Management
xx.xx.20xx and awarded the grade

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of Bachelor's thesis

Author Santeri Tenkanen

Title of thesis Rating National Hockey League teams: the predictive power of Elo rating models in ice hockey

Degree Bachelor

Degree programme Information and Service Management

Thesis advisor(s) Olga Gorskikh

Year of approval 2019 **Number of pages** 31 **Language** English

Abstract

The purpose of rating systems is to have an accurate and reliable way to describe the strength of competitors in sports and games. Elo model is the base for the various real-life applications of these systems and interest to forecast sport events has also shed academic interest towards this area. Although academic research has covered various of different sports, Elo model hasn't been researched on an academic level in ice hockey context up to this date. This thesis presents and evaluates several applications of Elo rating models in National Hockey League context in order to evaluate their predictive power to forecast full time match results.

The objective of the research was to find the most accurate application of Elo to rate NHL teams and to evaluate their predictive power against some benchmark methods. Emphasis of this paper is more on evaluating predictive power of Elo-based ratings with different error measurement, rather than using economic measures in order to determine if Elo models could be profitable in the betting market. Different applications of Elo rating models from previous academic literature are considered and evaluated to provide necessary scope. These include the choice of normal distribution against logistic distribution and using modified coefficient to better reflect the actual performance. Different error measurements are assessed to ensure that the predictive power is measured with wide enough perspective.

Publicly available NHL data from season 2005-2006 up to season 2016-2017 were used in applying different models. Data were modified and validated in Excel before importing to R, which was used to build models and analyse results. In addition to Elo models, two naïve prediction methods, market odd probabilities and an accumulative rating system, adapted from International Ice Hockey Federation, were used as benchmarks.

Results indicated that adding goal difference variable to adjust the rate of change in ratings did not provide enhanced results in NHL context when compared against the original Elo model. Normal distribution was speculated to be better in modelling performances in National Hockey League when compared against the logistic distribution. However, paired t-test on these models indicated that this difference was not statistically significant.

Comparison results gave further evidence on Elo's capability to model performances in sports. Elo models were able to outperform naïve predictions with statistical significance, which is further proof for the consensus amongst previous academic papers. However, results couldn't provide statistically significant evidence to prove Elo-based probabilities to be inferior against market odds-based models. However, odds-based model was the best performing amongst models and benchmarks, indicating similar results to previous research on other sports. Overall, results indicate that properly constructed Elo system can be used in modeling performances in ice hockey and that Elo-based probabilities can be used as a considerable benchmark for future research on match result forecasting.

Keywords Elo, rating system, performance evaluation, sport forecasting, National Hockey League

Table of Contents

1 Introduction

Sport rating systems are perhaps most notably used by many sport organizations that arrange competitions in their respective sports. A motivation for constructing proper rating system is to be able to describe the differences in strength amongst competitors. This system can then be used to various purposes. In example, the international governing body of association football, FIFA, used their ranking system to allocate teams to different pots for the purpose of group-stage draw for the 2018 FIFA World Cup (FIFA.com, 2017).

Since the World Cup draw in 2017, FIFA has moved to a new rating system (FIFA, 2018), which is an adapted version of the original Elo rating system introduced by Arpad Elo (1987). The original Elo model was developed for and used as a system to rate chess players for the World Chess Organization FIDE. Over the years it has been applied to many other sports and games, such as croquet and sumo wrestling (Stefani, 2011). Recent academic research has also applied Elo for other popular sports, including basketball (Štrumbelj & Vračar, 2012) and football (L. Hvattum & Arntzen, 2010). Research has suggested that Elo-based rating systems are able to provide significant information about the strength of competitors in various sports. Many of the papers have used Elo ratings to calculate propability estimates in order to assess system's predictive power and in most of the cases, the ability to provide profits on the betting market.

Despite the successful applications in numerous other sports, ice hockey is still a sport where Elo's possibilities haven't been researched on academic level. Previous studies and the increasing use of Elo as a rating system basis calls for research in ice hockey context. Elo models have beneficial properties being continuous, simple and reliable rating systems that allow evaluation of performances without advanced statistical methods. In the era where advanced methods such as machine learning and tracking technologies are becoming popular, simplified method for performance evaluation and result forecasting is likely to have use as a benchmark for further research in ice hockey.

1.1 Introduction to rating systems and match predictions in sports

Ratings systems are commonly used by international sport federations and other competition organizers in different sports and mind games. They are automated systems developed in the intention to rank or rate competitors based on their performances

(Stefani, 2011). These performances can be evaluated based on previous match results and/or other match statistics. While rankings merely place competitors in order based on their strength, ratings can further describe the difference in strength between competitors.

According to Stefani (2011), different types of rating systems can be categorized into three groups: subjective, accumulative and self-adjustive systems. Subjective systems are rare and only used in few combat sports where a panel of judges rate competitors. In accumulative rating systems ratings are based on the sum of the past results. These past results are often weighted and aged over time in numerous ways. Other factors such as additional statistics can also be added to accumulative models.

Adjustive rating systems, such as Elo-based rating systems, adjust the previous ratings of competitors after their encounter based on the difference between actual and predicted result. The size of rating adjustment can be modified and this adjustment varies in many sports. Based on Stefani's research, both accumulative and adjustive systems can identify strength of teams and players, when they are properly constructed. Most rating systems that were in use at the time of Stefani's research were accumulative systems. Elo rating system was the most commonly used amongst adjustive rating systems. (Stefani, 2011).

Many other groups can also benefit from the information that rating systems can provide in addition to sport competition organizers. Many have attempted to form models that would estimate probabilities for sport competitions. These estimates are then used to forecast match results and to find possible profitable winning strategies in the betting market.

In football, Hvattum and Arntzen (2010) found that Elo ratings were significant in predicting football match outcomes. Their Elo based models were outperforming other benchmarking methods in terms of observed loss, but models were found to be inferior to the information that betting market odds hold. Lasek et al. (2013) found that Elo-based models were able to outperform formerly used accumulative ranking system used by the International Federation of Association Football, FIFA.

Kovalchik (2016) found similar results when applying Elo rating model into tennis results. Model was found to be performing better than other approaches used in research, but still failing to outperform market odds in terms of accuracy. While Štrumbelj and Vračar (2012) focused their research on forecasting basketball match outcomes with their multinomial logit Markov model, they also applied and used Elo as one of their

benchmarking prediction methods. The results showed that Markov model had no statistically significant difference in predictive power compared to Elo model, which only took the absolute winning margin into account. Bookmaker odds were again proven to be better in their predictive power when compared against suggested models.

1.2 Research objectives and research questions

The objective of this research is to explore possible use of Elo rating systems in evaluating strength of National Hockey League teams. Different applications of Elo applied from previous research are explained, considered and evaluated. Objective is to assess the predictive power of those systems that can be considered to be adapted to ice hockey. Predictive power is evaluated by making probability forecasts for match results. The size of error in these forecasts are then compared against other models and benchmarks with different error measurements.

Particularly, this research attempts to answer the following research questions:

(1) What is the best application of Elo rating system to measure strength of National Hockey League teams?
(2) Can rating-based probability estimates be accurate in predicting match results and how they compare against other models and benchmarks?

1.3 Scope of research

Main focus of this research is to apply previously presented versions Elo models to a new field in sports on an academic level. In addition to the original Elo model, possible applications from past research on other sport are considered. These include the choice of using logistic distribution instead of normal distribution and the use of goal-based k-coefficient, applied from the football application of Elo (Hvattum & Arntzen, 2010). Theoretical background is presented on rating systems, drawing probability estimates from ratings and on different error measurement to evaluate these estimates.

Altogether, research attempts to provide a necessary scope for the reader to assess Elo's practical usage in performance evaluation and match probability predictions in ice hockey. For this purpose, Elo models are compared and evaluated against benchmark methods. Although this paper doesn't attempt to form a winning strategy for the betting

market, it is seen necessary to use betting market odds as a one of the benchmarks for possible practical betting related usage in the future.

This research includes applying Elo rating system in context of National Hockey League (NHL), which is the professional ice hockey league played in North America. The nature of NHL makes it particularly favourable to be used in this research as each team plays high number of games per season and there are no teams that are relegated or promoted after each season. It was chosen that the performance of NHL teams would be evaluated based on regular time match results. Alternative option would have been to include overtime or penalty shootout results as NHL matches are played until a winner is decided. However, usual sport applications of Elo have not included aspects of overtime and shootout situations, which could have had undesired influences on models. Possible problems include difficulties on evaluating shootout and overtime results as the situation on ice is different from regular time, possibly favouring different teams. These problems were left out of the scope and for further research.

1.4 Structure of the research

The rest of this thesis is structured as follows. Chapter two reviews the theoretical background for Elo rating systems and how probabilities can be drawn from ratings. The theoretical basis for estimating the predictive power of models with different error measurements and statistical tests is also introduced. Chapter three presents data that are used in research, its sources, modifications and validation that were performed before analysis. Then, methodologies used in applying Elo rating models are discussed, following an explanation on different benchmarking prediction methods.

The results are discussed in the fourth chapter, including results from calibrations and error measurements. Findings and conclusions from these results are discussed further in the final chapter, including implications for both practice and theory. Reader is finally reminded about the limitations of the thesis and possible future research objectives are suggested.

Terms for Elo models used in this research are as follow: Elo Normal will be used to refer to the original Elo model that follows normal distribution. Elo Logistic will be used for the Elo model version which follows logistic distribution. Elo Normal Goal and Elo Logistic Goal are used to describe the goal-based versions of Elo models that follow their corresponding distributions.

2 Theoretical background

Need for a rating system comes from the desire to evaluate strength of competitors in competitive matches. However, there is very little use for a system that can only give a conclusion that team A is better than team B. Furthermore, rating system should also be able to describe the relative strength between competitors (Elo, 1987, p.9).

Elo rating system uses interval rating scale, which means that ratings are separated into intervals that can describe differences in strength between competitors. Ratio scale is similar in a way that it also uses interval scale, with an addition of having an absolute and meaningful zero value. Between these two scales, interval scale was chosen for Elo model as it was already used in chess with reference points that had general acceptance in the game (Elo, 1987, p.19). Nominal and ordinal scales are insufficient for this purpose as they lack ability to describe relative differences.

Theoretical background on Elo rating systems is further explained in this chapter. In addition to the original model, the most suitable previous sport applications of Elo models are discussed and explained. Explanation on how to form probability predictions from Elo ratings using logit and probit regression models is introduced. Finally, measurements of error are discussed for the purpose of being able to evaluate models with a proper manner.

2.1 Elo rating system

Elo rating system was first established by Arpad E. Elo in his book “The rating of chessplayers, past and present” (1987). Its original use was to provide a better way to rate chess players. Original Elo model presents a system with numerical rating values, which can then be used to derive winning probabilities for matches. The following description of this system is slightly modified from the original by removing parts that have only use in chess context. In example, original version accounts tournament situations where players participating in different tournaments might not always be the same. However, these calculations are irrelevant for NHL model as competitors in NHL are mostly same with few exceptions discussed later in Chapter 3.2.

The variation in performances of competitors in sports and games is crucial to be recognized in Elo models. The overall rating for a competitor, which is relative to ratings of other competitors, is derived from performance ratings in a large set of matches.

Performance rating is objectively derived number for each encounter between competitors and it accounts the score of a match and the relative difference in strength. The intention of Elo rating system is to define the best possible estimate for current relative strength of competitor by combining these performance ratings from previous matches (Elo, 1987, p. 20).

Standard deviation is a common measure used to describe variation in many datasets and it is also utilized in Elo system. As Elo system is an interval rating scale, this rating scale C must be defined to get a meaningful deviation in rating values. The most appropriate rating scale C for Elo system is $C = \sigma$, which is the standard deviation of NHL teams in the Elo model over a long period of time. This is also a logical way to connect rating scale into distribution functions used in Elo, which will be discussed later in this chapter. The scale parameter C and the reference point R, which is the initial value and the average value of all ratings, can be selected arbitrarily and this selection won't affect upcoming probability estimates of the system (Elo, 1987, p. 18-19).

The profound idea of Elo rating system is to derive a continuously updated rating, which is based on a large number of encounters in head-to-head matches. The weight of earlier performances diminishes when new matches are played and newer indications of strength are added to the system. System forms a new rating for a competitor after each match with the following equation:

$$R_n = R_o + k(\alpha - e),$$

where R_n is the new rating of the team after the match, R_O is the old rating of the team before the match, k is the rating point value defining the rate of change, α is the actual match score and e is the expected match score.

Match score in ice hockey and many other sports can be either home win, draw or away win. Win equals to α value of 1, draw equals to α value of 0,5 and loss equals to α value of 0. Coefficient k determines the size of rating adjustment. High values of k give more weight on recent results as the rating adjustments are larger. Conversely, lower values of k give more weight on older results.

The expected match score e is the winning probability of a team in an encounter against another team. Expected score in Elo model is based on rating difference between these teams and probability function that is chosen. Possible probability functions to model

sport team performances and to derive expected scores are now discussed in the following chapters.

2.2 The use of normal distribution in Elo model

The key assumption in original Elo's connection from ratings to probability is that performances of competitors will be normally distributed, as long as they are evaluated on an appropriate scale, such as the standard deviation used in Elo (Elo, 1987, p.19). This normality assumption on performances is used when calculating expected match score e.

Therefore, in a large set of matches, performances of a single team are expected to be normally distributed with the standard deviation of σ. When evaluating probabilities on head-to-head encounters, the performances of two teams form a joint normal distribution with zero mean. The performances are also expected to be homoscedastic, meaning that the standard deviations of teams A and B are assumed to be the same. Then, the distribution of teams' performances in a match has a standard deviation of

$$\sqrt{\sigma_A^2 + \sigma_B^2} = \sqrt{\sigma^2 + \sigma^2} = \sqrt{2\sigma^2} = \sigma\sqrt{2},$$

where σ_A is the standard deviation of team A's performances and σ_B is the standard deviation of team B's performances. Expected score for a team A to win team B can now be calculated from joint normal distribution of the teams rating difference:

$$e_A = P_A = P(X \leq x),$$

where $X \sim N(0\,; C\sqrt{2})$, e_A is the expected score in a match for team A, P_A is the probability of team A to outscore team B, X is the joint normal distribution function, x is the difference in ratings between team A and team B and C is the scale parameter and the standard deviation σ of performances for a single team.

Expected score and new rating can be calculated for team B in the same way. Rating of teams is adjusted after each match and as the performances of teams are normally distributed, also the ratings converge into normal distribution when the number of matches in system rises. An average rating in system is the arbitrarily chosen reference point R.

2.3 The use of logistic distribution in Elo model

Although the original Elo model is based on normality assumption, problems concerning this assumption were already acknowledged during the development of the system. In fact, it was possible to arrange chess performances into groups with different standard deviations, which meant a violation of the homoskedasticity assumption. In another words, the standard deviations between teams A and B may be different, which lead to breaking the normality assumption on the joint distribution (Elo, 1987, p.141).

Logistic curve was found to better represent distribution in performances and this version of Elo system has been found to have use in many applications in sports. Therefore, expected score can't be calculated from the normal distribution as performances are assumed to be logistically distributed. Formula for expected score for an encounter of teams A and B is acquired by using the logistic function:

$$\alpha_A = P\ (Team\ A\ wins) = \frac{1}{1 + D^{-x/2C}},$$

where x is the rating difference between team A and team B, D is the rating difference parameter and C is the scale class interval and standard deviation of performances for a single team. This function is derived from the logistic distribution function and has a logarithmic interval scale.

Scale class interval C is the standard devation of team's perfomance as in the normal distribution model and its value can be set to any desired level. This parameter is set at rating difference x, where the odds for team A to score over team B is the rating difference parameter D. The following equation illustrates this relation between parameters:

$$C \log_D(P_A/P_B) = x,$$

where P_A is the probability of team A to outscore team B and P_B is the probability of team B to outscore team A.

Usual choice of parameter D is 10 as it has beneficial relation to the normal distribution. Logarithms are therefore taken to the base $\sqrt{10}$, which also defines the odds of a rating difference of one class interval. The value of $\sqrt{10}$ is close to the odds of a rating difference of one class interval in the normal distribution system, which allows these two models to be compared relatively well later in this research (Elo, 1987, p.141).

2.4 Elo model with *k*-coefficient based on goal difference

A modified version of Elo rating system was introduced by Hvattum and Arntzen in their paper "Using ELO ratings for match result prediction in association football". They replaced the coefficient k with a formula that allows ratings react to results according to their goal difference (L. Hvattum & Arntzen, 2010).

In this version, coefficient k is replaced with the following formula:

$$k = k_0(1 + \delta)^{\lambda},$$

Where $k_0 > 0$ and $\lambda > 0$ are fixed parameters and δ is the absolute goal difference of match result. Fixed parameters k_0 and λ are calibrated to find the best fitting parameters for the model. Otherwise the model of Hvattum and Arntzen follows same formulas as represented for the logistic model. In addition to this model, this paper will also evaluate goal-based model which follows normal distribution.

Other modified versions of the Elo model have also been presented in the past. For example in tennis context, Kovalchik (2016) evaluated the predictive power of Elo system variant developed by FiveThirtyEight data journalists. Their version used a function that accounted player's career games in k-coefficient to determine the size of the rating adjustment. However, versions like this don't have a meaningful connection to ice hockey. Goal-based version of Elo is therefore the only previous adaption of Elo in other sports that will be adapted and evaluated in this research.

2.5 Ordered probit and ordered logit regression models on predicting match probabilities

As match predictions for three possible match outcomes cannot be drawn from the expected scores of teams A and B, ordered probit regression and ordered logit regression models must be adapted to make these predictions (Greene, 2012). Normal distribution is used in the probit model, giving therefore predictions for Elo models following normal distribution. Logistic distribution is used in the logit model, giving predictions for Elo models following logistic distribution.

Prediction model is formed by estimating coefficients for the following equation:

$$y^* = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon,$$

where y^* is the ordered outcome of matches to be estimated, $\boldsymbol{x}'$ is vector of the Elo rating difference between home and away team before a match, $\boldsymbol{\beta}$ is a vector of coefficients to be estimated for each outcome and ε is the error term that follows the chosen distribution.

For a model with three possible outcomes for a match, vector of coefficients $\boldsymbol{\beta}$ is used to estimate unknown parameters μ for observed outcomes:

$$y = \begin{cases} 0, & -\infty < y^* \leq \mu_1 \\ 0.5, & \mu_1 < y^* \leq \mu_2 \\ 1, & \mu_2 < y^* < \infty \end{cases}$$

Statistical software R and its function "polr" of the package MASS is used in this research to find optimal parameters μ and vector of coefficients $\boldsymbol{\beta}$.

In football context, the use of ordinal logit models as a prediction method has been questioned for its ability to estimate the draw results due to proportional odds assumption. Multinomial logit model has been considered as an alternative in a recent research (L. M. Hvattum, 2017). However, there wasn't found statistically significant differences between multinomial and ordinal model. Therefore, both of them are acceptable options in analysing related situations.

2.6 Measurements of predictive power

It is important to discuss measures which are later used in assessing and comparing the predictive accuracy of different models and benchmarks in this thesis. These different measures are sometimes called loss functions as they measure the information that is "lost" when comparing estimated probability to actual result. These functions can be used in calibration as well as in evaluation of different models against each other in attempt to find the best ones in terms of predictive accuracy.

The situation and purpose of use for error measurements should be considered when choosing which measure to use in evaluating error in models. When choosing a measurement for calibration purposes, sensitivity of error measure is a desirable property. On the other hand, when comparing forecasting methods, sensitivity is not the primary criteria and more weight should be put to reliability and outlier protection (Armstrong & Collopy, 1992).

For calibration purpose, Armstrong and Collopy (1992) recommended geometric mean relative absolute error (GMRAE) because of its high sensitivity. Hyndman and Koehler

(2006) argued against GMRAE in their paper as they found it to be adding unnecessary complexity and a level of arbitrarity in measuring. Two other options that Armstrong and Collopy consider, root mean squared error (RMSE) and mean absolute percentage error (MAPE), provide sufficient sensitivity but contain other problems. RMSE is considered unreliable as the choice of time period is likely to have significant effects on the calibration result. MAPE on the other hand is biased and favours low forecasts. In addition, MAPE uses the actual match result in the denominator, which is problematic for research context as away victories are assigned a value of zero.

Desired attributes differ when the purpose is to compare different forecast models against each other. Armstrong and Collopy (1992) suggest that primary attributes to look from error measurements in this case should be reliability, robustness to outliers, relatedness to decision making and measure's accuracy to assess what is supposed to be measured. Reliable measurements produce similar results in different time periods. This is also related to measurement's ability to be robust against outliers as single extreme values on different time periods shouldn't have undesirable effects on results. Error measurement should also be related to decision making and demonstrate properly error on which it is used to measure. Armstrong and Collopy call the latter attribute construct validity. Based on these attributes they recommended use of median relative absolute error (MdRAE) for small sets and median absolute percentage error (MdAPE) for larger sets.

However, Hyndman and Koehler (2006) argue against this recommendation. They point out that absolute percentage error (APE) measures can't be used when observed values contain zero values, as is the case with NHL match scores. In addition, these relative measures often require trimming of extreme values, which undesirably increases complexity and adds an arbitrary level to the measurement. Instead, they recommend the use of mean absolute scaled error (MASE), which scales absolute forecast error against one-step behind naïve forecast error using previous value in the data as naïve forecast.

Betting related economic measures, such as betting strategies using unit bets or Kelly criterion (Kelly, J., 1956) are particularly good for evaluating suitability of different rating-based betting strategies in the betting market. However, they evaluate Elo based models in terms of ability to produce profits in the betting market rather than directly measuring accuracy of the model. As these measures don't fully reflect the objective of this thesis, these measures are not considered further in this paper.

3 Methodology

The methodology used in applying different Elo models into National Hockey League data is discussed in this chapter. The data source and reliability of the data are discussed, including modifications and validation that were made before adapting Elo models. Calibration process in Elo models is further discussed in Chapter 3.2., but the results of calibration are left for the fourth chapter. Benchmarking methods and choices made on error measurements are also discussed in Chapters 3.3. and 3.4.

In adaption of the Elo models, data for six seasons up to 2010-2011 season were first used to get reliable ratings for teams. Then, data for seasons 2011-2012 to 2015-2016 were used to calibrate Elo models. This period is further referred to as the training period in this research. It is noteworthy to mention that the 2012-2013 season contains less games per team than usual a NHL season due to lockout. Predictions are made for season 2016-2017, which is further referred to as the testing period in this research.

3.1 Data source and modifications

NHL match data used in this research are from a public website Hockey Reference (Hockey Reference, 2019), which collects results, standing and other statistics from National Hockey League games, dated all the way to the first season 1917-1918. The main reason to use Hockey Reference as a source for data was the ability to download it on CSV format. Data were then modified and validated in Excel before importing it to R for analysing purpose. At the end, imported data contained information from seasons 2005-2006 to 2016-2017, including a total of 15 296 match results.

Primary data provider for Hockey Reference is Gracenote, which is part of a large American market research company Nielsen Holdings Inc. A contributor for the historical data from past season is also Dan Diamond and Associates Inc, which has for example published NHL Official Guide and Record Book 2016. Hockey Reference has not clarified which information is from which source, but it is reasonable to assume that recent data are from Gracenote. As only match scores are needed for constructing Elo models of this research, it can be considered unlikely that data on Hockey Reference would contain errors on the easiest information that can be acquired from NHL matches.

A usual season in NHL consists of 82 regular season games and Stanley Cup playoffs, where 16 best teams of the regular season will advance. Playoffs are played with the best

out of seven format, requiring a team to win four games to move on to the next round. There is an exception to the regular format on the season 2012-2013, where there were only 48 games on the regular season because of the lockout. However, there were no modifications that had to be made to the data as the number of games in a time period doesn't affect Elo ratings.

Few modifications were made on those occasions where NHL matches had been cancelled. For example, in 7.1.2014 a match between Buffalo Sabres and Carolina Hurricanes was postponed because of a blizzard. This was recorded on the original data as null score match. This occasion and any similar ones were deleted as these matches were accounted on the days they were finished, which were already displayed in the data with remarks. Similarly, any games that were postponed after the match had started and continued on another date were accounted on the days these matches ended.

Format of NHL is desirable to rating purpose as the number of teams has been same during the whole period data period. However, one significant modification had to be done, as the organization of Atlanta Thrashers was relocated to Winnipeg after the 2010-2011 season. It was seen logical to assign the old rating of Atlanta Thrashers for Winnipeg Jets, as this was in fact a relocation of franchise with most of the staff and players moving to a different location, rather than a complete formation of new team.

Another dataset containing bookmaker average odds for test period of 2016-2017 season was acquired from Odds Portal website (Odds Portal, 2019). This data was acquired manually from the source as no proper download format was available. Odds Portal acquires its data from closing odds of several international companies that offer betting odds for NHL matches. The number of bookers that were accounted in the average odds ranged from 30 to 34, depending on the match. A large number of bookmakers contributing to the average odds of a single match diminishes the weight of a single booker's estimate and the possibility of incorrectly acquired odds for a certain game to minimal. Average odds were chosen instead of the highest possible odds for each outcome as average can better reflect probability estimates of all bookmakers. Extreme values suit better for evaluating betting strategies as they diminish the bookmaker's margin, but this was not within the scope of the research.

3.2 Applying ELO rating systems

The choice of seasons that Elo ratings would be first applied on and then tested on was done with a few criterions in mind. Firstly, enough data for ratings must be collected

before strength can be assumed to be accurate. It was also desirable to acquire a large amount of data prior to the testing period to calibrate the coefficients accordingly. Secondly, considering the choice of first and last season, it was desirable to avoid situations where a team would be added into the National Hockey League. This would cause problems as enough data would have to be collected on the performances of this team before its rating could be assumed to be accurate.

With these criterions in mind, data was acquired from seasons 2005-2006 to 2016-2017. The lockout season of 2004-2005 caused a gap of one year to data as no games were played during this lockout season, which would have caused ratings in the beginning of season 2005-2006 to be based on ratings from approximately one and a half years ago. This may have caused inaccuracies and more unexpected results as the strength of the teams could have changed during time gap. Therefore, the lockout season eventually placed a logical starting point for data.

When considering the choice of the last season, it was kept in mind that the last season would be used as testing period, and that it was desirable to use the most recent season possible for as current research result as possible. However, the most recent full season at the time of writing, 2017-2018, included introduction of Vegas Golden Knights to the league. This would have caused problems on getting accurate rating for them as was described earlier. For this reason, 2016-2017 season was chosen to be used as a testing period. Training period was set to contain data from five seasons of 2011-2012 to 2015-2016, including the short lockout season of 2012-2013.

Reference point R of the scale was chosen to be 400 and the scale parameter defining standard deviation of performances was chosen to be 200. The rating difference parameter D for the logistic version of Elo model was chosen to be 10 for the reasons discussed in chapter 2.3. A relatively high standard deviation compared to the reference point was chosen after few tests with different scale parameter values. It is necessary to notice that the choice of these parameters was completely arbitrary. However, these arbitrary choices do not affect the expected match results and probability estimates as long as the k-coefficient is appropriately calibrated. A high standard deviation is simply a way to increase rating differences in Elo models for more understandable explanation of the ratings.

Rating and expected score formulas for different Elo rating systems were adapted after selection of seasons and parameter values. Calibration of rating adjustment coefficients

was done by minimizing error in actual and expected score of the matches during testing period. Optimize-function from the "stats" package in R was used for calibration.

3.3 Benchmarking methods

Results from match predictions of Elo model must be compared against results of some benchmark methods in order to get a proper overview of the predictive power of different Elo models. Previous results of Elo model applications in other sports, such as tennis (Kovalchik, 2016) and football (Lasek et al., 2013), have concluded that Elo's predictive power is often significantly better than the naïve prediction methods but still significantly worse than the market odds. These can be considered as the two extremes of prediction methods. Therefore, they were considered as good benchmarks to evaluate Elo models against.

Two naïve prediction methods are used in comparison. The first method assumes that there is no information available prior the match, therefore assuming a uniform distribution which places 1/3 probability for each outcome. The latter naïve method takes prior matches played into account by counting the frequency of each match outcome. These frequency-based probabilities are first counted from the training period data and then assigned to each outcome of all the test period matches.

In market odds method, probabilities are simply calculated for the test period matches by transforming average market odds into probabilities. However, the sum of probabilities in odds doesn't equal to one because of bookmakers' margins. For this reason, the probability of each outcome must be normalized to make the sum of outcomes equal to one. This then allows a proper comparison of the bookmakers' predictions to other prediction models of this research.

The fourth method is an accumulative rating system that has prior usage in ice hockey. This system is adapted from the International Ice Hockey Federations' (IIHF) World Ranking system for national teams. In the benchmark accumulative rating system point scores from each match played are given in the same manner as were scores in Elo model: win equalling the score of one, draw a score of a half and lose a score of zero. A completely arbitrary decision is made to not use different weighting methods on playoff or any other type of matches.

Finally, the same uniform ageing of points is used as was in the IIHF ranking system. This ageing system accounts performance points of the past year with 100% weight,

points older than one year but not older than two years with 75%, points older than two years but not older than three years with 50% and points older than three years but not older than four years with 25% weight. Points older than four years ago will not be accounted. The performances of teams are assumed to be normally distributed and model for probabilities is constructed in a same way as for the Elo model: training period is used to estimate the coefficients of ordered probit model, which is then used to calculate probabilities for testing period.

To conclude, there are four methods overall that are used to compare the Elo models against in this thesis. These are market odds -based probabilities, naïve equal probabilities, naïve probabilities based on training period frequencies and accumulative model adapted from the IIHF national team ranking system.

3.4 Choice of measurements

Error measurements are needed in two ways in this research: to calibrate coefficients for Elo models and to estimate predictive power of models and benchmarks. Both purposes asses how much information is lost, but in different ways. In calibration, the loss is assessed as the difference between actual score and the probability distribution value of the teams rating difference. Actual scores can take either a value of one for home win, half for a draw and zero for away win. Probability distribution can take values from zero to one, where zero predicts a sure away win and one predicts a sure home win. When evaluating models against each other, estimated probabilities are compared to true probabilities of each outcome based on match result. The true probability is in this case calculated by assigning a value of one as the actual match result. This is then compared against probability estimate of a model or a benchmark to estimate the loss.

Despite the recommendation from Armstrong and Collopy (1992) to use GMRAE measurement in calibration, this method was found to produce poor results in NHL Elo models. Calibrated coefficients converged to the minimum values set to optimization even when the measurement was tested with two different benchmark error methods, one being random walk and the other frequencies from previous occurrences. Trimming of extreme values with “winsorizing” method was not used as it was seen desirable not to add an arbitrary and complex level to Elo model, which should be used as a simple model. Another likely problem and difference to Armstrong and Collopy's approach is that they used continuous ratio data, when Elo models have an interval scale data with actual

scores being discretely distributed. Therefore, despite the concerns on reliability, RMSE method was chosen to be used in calibration instead of GMRAE and MAPE.

MASE method, described in Chapter 2.6 and recommended by Hyndman and Koehler (2006), was chosen for its property to use previous values of data as a naïve prediction. In NHL data, this means that MASE can be considered to draw naïve random forecast from the frequency distribution, as the previous values form a discrete distribution according to the test period match result frequencies. This eliminates the possible problems of close but not identical scales used in logistic and normal Elo models. MASE method is better explained and described in chapter 2.3.

RMSE is widely used measure as it has close relation to decision making being simple and understandable measurement (Armstrong & Collopy, 1992). Despite the critic of being sensitive and unreliable, in some betting cases, sensitivity can even be desirable attribute of the model. Higher sensitivity might lead the model to recommend matches that are more profitable in some cases where the difference to odds-based probabilities increase, therefore leading to higher profits. This speaks for the use of some relatively highly sensitive and closely to betting related measure. Squared error measures have also been used previously to evaluate Elo models in academic literature, for example by Hvattum and Arntzen (2010), which speaks for the use of these measures for future comparison between models. It was also seen meaningful to include median absolute error (MAE) in comparison to better see effects of high sensitivity of RMSE measurement, which is more robust to outliers.

As the aim of this research was to evaluate the information that Elo based models hold, it was not seen appropriate to evaluate models based on economic measures or merely on correctly identifying better team and game winner of the match. In cases where models produce probability estimates, it is more natural to take these estimates into account when evaluating models against each other (Witten & Ian, 2005, 158). This is why zero-one loss function was not considered as possible evaluation measurement.

To conclude, the model calibration in this research will be done by minimizing root mean squared error (RMSE) measurement on the training period, despite the issues concerning reliability. Different models and benchmarks will then be evaluated against each other using three measures: RMSE, mean absolute error (MAE) and mean absolute scaled error (MASE). In addition to error measurements, paired t-test is used to assess whether the mean losses between two models are significantly different from each other. Results of the research will now be presented in the following chapter.

4 Results

This chapter presents the results of this research. The first part presents results from the calibration of Elo models. The assumption on which distribution should be used in modelling NHL team's performances is critically discussed and distribution options are visually compared using QQ-plots. Sensitivity of calibrated models and their coefficients will also be assessed against previous Elo-based applications in sports.

Chapter 4.2. presents results from the comparison of error measurements on the testing period, 2016-2017 season. Models will also be ranked based on different measurements and evaluated against each other with paired t-tests. Most of the conclusions and further discussion on these results is left for the fifth and final chapter of the thesis.

4.1 Calibration results

Calibrations of coefficients in Elo Normal and Elo Logistic models were executed by minimizing root mean squared error RMSE between NHL seasons 2011-2012 and 2015-2016 with the optimize-function in R, as discussed in chapter 3.3. In the same manner, calibration was then done for coefficients in Elo Goal Normal and Elo Goal Logistic models by minimizing RMSE. The results of the calibrations for Elo Normal and Elo Logistic models are presented in Table 4-1.

Model	k
Elo Normal	5.6811
Elo Logistic	7.8933

Table 4-1: Calibration results

As results show, optimal value for coefficient k in Elo Normal model was found to be lower than in Elo Logistic model. This indicates that rating adjustments in Elo Logistic model should be slightly larger compared to Elo Normal model. Calibration of goal-based Elo models, Elo Goal Normal and Elo Goal Logistic, showed that optimal values for coefficients of these two models converged to the values of their regular Elo models. Values of λ converged to the minimal value set for optimization and the k_0-coefficient values converged to the values of k in Table 4-1. Results indicate that goal-based models are unable to provide any additional information on the performance of a team in NHL

match. Therefore, it was seen evident to move forward from calibration only with the non-goal-based Elo Normal and Elo Logistic models.

As *k*-coefficient determines the rate that the ratings change in Elo models, it is rational to also have a look at some basic descriptive statistics to see the statistical differences between ratings in calibrated models. Table 4-2 presents minimum and maximum rating value, mean, median and standard deviation of New York Rangers' rating values in Elo Normal and Elo Logistic models.

Rating model	Team	MIN	MAX	Mean	Median	St. dev.
Elo Normal	NYR	381.0	480.3	416.9	411.2	21.44
Elo Logistic	NYR	373.6	511.7	423.5	415.5	29.82

Table 4-2: Statistics for different models of the New York Rangers Elo ratings

As it can be seen from Table 4-2, there is a clear difference between statistical measures in logistic and normal models. Results show that higher *k* value leads to larger differences on Elo Logistic model statistics as was expected. Minimum value, maximum value, mean and the median are all further away from the reference point than in Elo Normal model. Standard deviation in rating is also higher in Logistic model. Ratings of New York Rangers have been plotted in Figure 4-1 for the whole data period with both Elo Normal and Elo Logistic models to visually illustrate this difference between logistic and normal model.

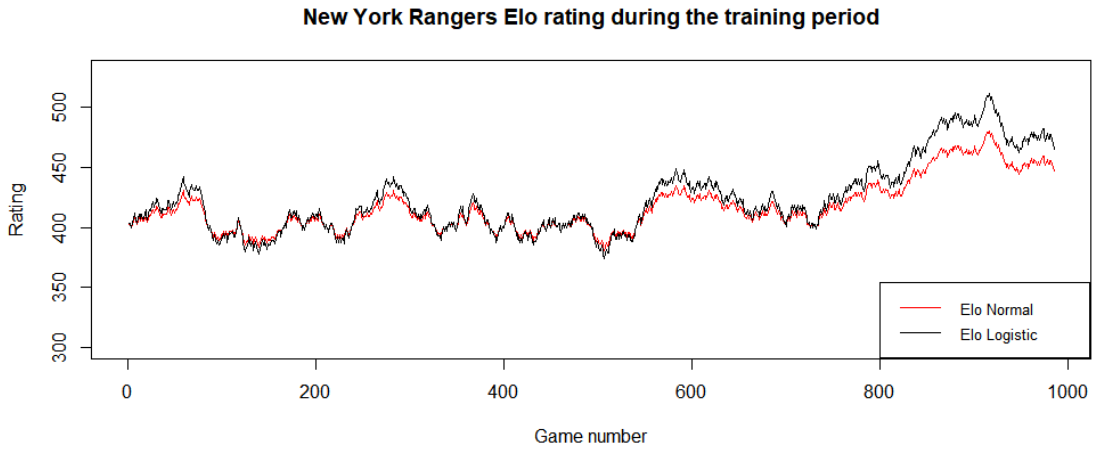

Figure 4-1: New York Rangers Elo rating between 2005-2015

Figure 4-1 shows a clear difference in ratings of New York Rangers over the training period in different models. In particular, the difference in standard deviation between Elo Normal and Elo Logistic model in Table 4-2 can be seen clearly from the plot. New York Rangers has been performing well especially during recent seasons. Median and mean ratings are above the reference point in both models but the rating in logistic model is constantly further away from the reference point when compared against normal model rating.

When probabilities are later calculated from the model, these differences aren't necessarily as large. This is because of the use of probit and logit models that calculate optimal coefficients based on rating difference. It is also worthy to mention that Elo Normal and Elo Logistic models are not directly comparable as arbitrarily chosen parameters affect expected score in a match. Therefore, the size of rating adjustment after each match is affected as well. As discussed earlier in Chapter 2.3., scale parameter value of the logistic model has been chosen in a manner which allows the comparison of logistic distribution -based model to the normal distribution -based model reasonably well, but not with certainty.

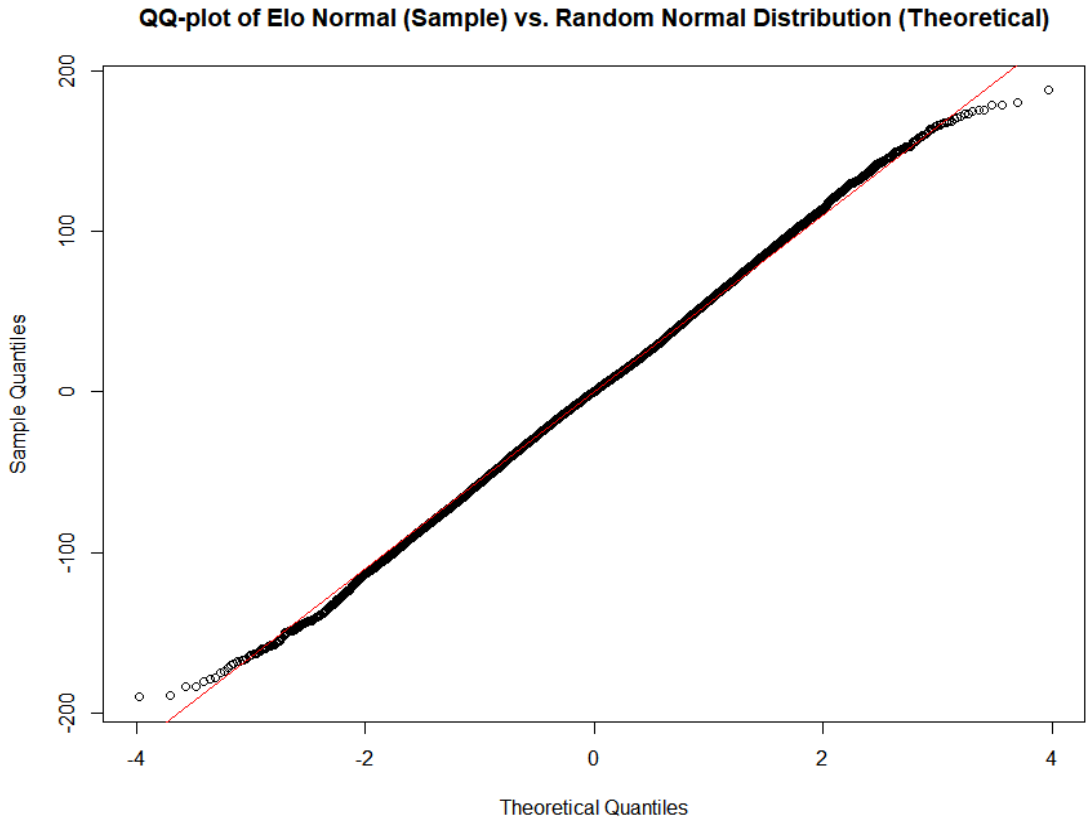

Figure 4-2: QQ-plot of Elo Normal vs. Random Normal Distribution

Another interesting comparison on calibration results is to plot the quantiles of each Elo model's rating adjustment against the quantiles of the random distribution of which the model is meant to follow. QQ-plot is used mainly to compare the size of the tails and it is not very instructive for the middle quantiles of the distribution. QQ-plot for Elo Normal model's rating adjustment against a random normal distribution of same size is plotted in Figure 4-2. Similarly, QQ-plot for Elo Logistic model's adjustments against a random logistic distribution of same size is plotted in Figure 4-3.

One can notice that some of the tail values for Elo Normal on Figure 4-2 are slightly lighter compared to the values of random normal distribution on both tails. However, the difference in tails is small, indicating that normal distribution is reasonably reliable in modelling performances of National Hockey League teams.

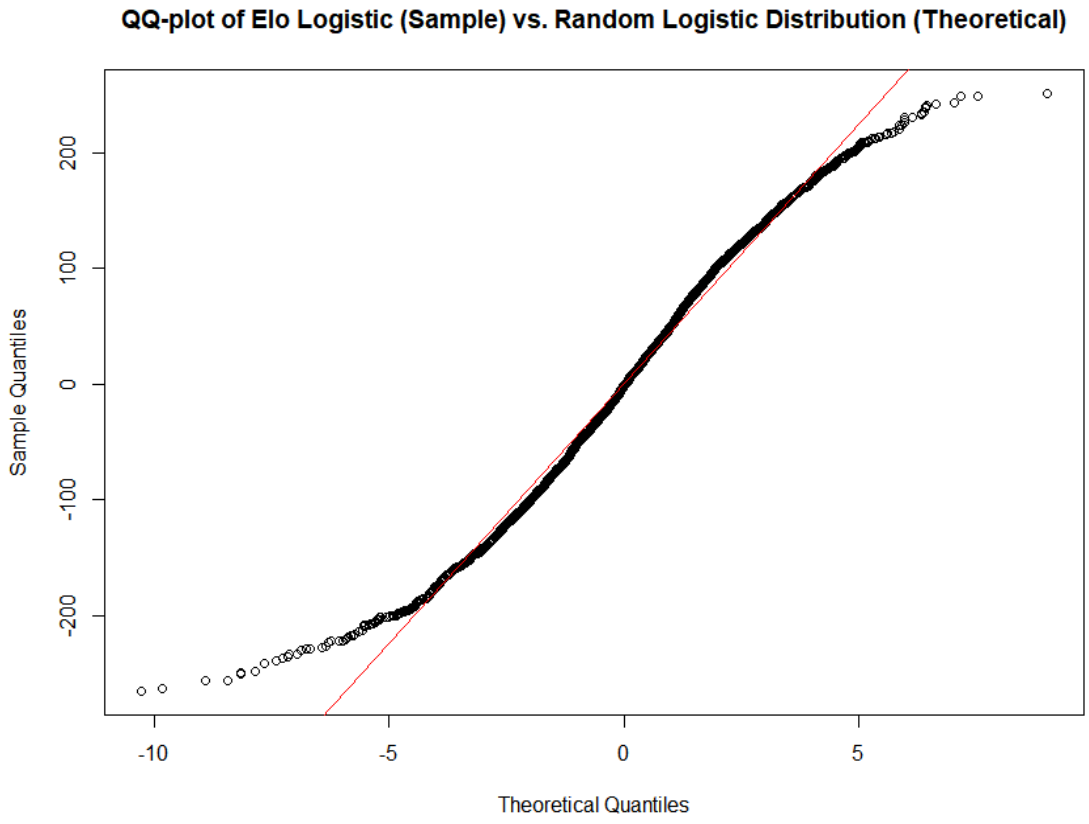

Figure 4-3: QQ-plot of Elo Logistic vs. Random Logistic distribution

In contrast, when comparing Elo Logistic model's quantiles against the quantiles of random logistic distribution in Figure 4-3, both tails of the applied Logistic model are seemingly much lighter to the random distribution. The difference is much larger than in the case of normal distribution in Figure 4-2. Therefore, results from QQ-plots

indicate that normal distribution is better in modelling the scores of NHL matches. Implications of this finding will be discussed further in Chapter 5.

It is interesting and beneficial for the use of future research to compare the models against other Elo based models in sports that have been presented in academic literature or used by sport organizations in the past. This comparison requires a unified measure which can be calculated for every model. Stefani (2011) has presented in his comprehensive survey of the different rating systems in sports a measure of sensitivity for different Elo based models. This sensitivity measure divides the maximum rating change with standard deviation of the model:

$$Sensitivity = \frac{K}{\sigma},$$

where K is the maximum rating change and σ the chosen standard deviation of the model.

Sensitivity measures of different Elo models are calculated and presented in Table 4-3. Most Elo rating models are officially used practical applications from other sports (Stefani, 2011). Hvattum and Arntzen's (2010) Elo application on football is also included in comparison. In addition to Hvattum and Arntzen's (2010) regular Elo model, they also presented goal-based application in their research. This model couldn't however be included in the comparison as the maximum goal difference from the data they used is not known.

Elo rating model	**Sensitivity**
Elo Normal NHL	0.03
Elo Logistic NHL	0.04
WCF Croquet	0.20
FIFA Women's Football	0.05-0.20
FIDE Chess and FMJD Draughts	0.05-0.12
IGF Go	0.14-0.58
ISF Sumo Wrestling	0.25-0.45
Elo Football (L. Hvattum & Arntzen, 2010)	0.10

Table 4-3: Sensitivities of Elo models in sports, adapted mostly from Stefani (2011)

As the k-coefficient is rarely constant in many of the Elo model applications in academic literature, maximum rating change couldn't be calculated for some of the other interesting applications of Elo. On the other hand, some Elo applications in sports use adjusted systems that weight the coefficient differently and according to competition

level or importance of the match. This means that sensitivity measure is represented as a scale in some models. Generally, in mind sports like FIDE's Chess Elo system, largest maximum rating change is for beginner level games, when in FIFA Women's football ratings, the largest weight is put on the most important matches (Stefani, 2011).

Table 4-3 shows that Elo Normal and Elo Logistic models for NHL are the two least sensitive models in this comparison. Stefani (2011) suggests the following categorization of sensitivities in Elo models: low sensitivity amongst Elo models is set at 0.15, medium sensitivity at 0.30 and high sensitivity at 0.60. With this categorization, models constructed in this research are in the low sensitivity category. Possible reasons for the low optimal sensitivity of models are further discussed in Chapter 5.

4.2 Results on the predictive power

After coefficients had been calibrated, Elo ratings were calculated for the seasons 2005-2006 to 2015-2016 according to the calibrated coefficients. Then, probit or logit model (depending on the distribution of the model and described in chapter 2.5.) was applied to training period using the polr-function of R-package MASS. This model was finally applied to make predictions during the test period of season 2016-2017. Three measures of error, root mean squared error (RMSE), mean absolute error (MAE), and mean absolute scaled error (MASE), were calculated for each model and benchmark. In addition, also the standard deviations of squared and absolute errors were calculated. Results of these measurements are seen in Table 4-4.

Model	RMSE	Sd. SE	MAE	Sd. AE	MASE
Elo Logistic	0.6466	0.1240	0.6390	0.0986	0.9439
Elo Normal	0.6465	0.1240	0.6390	0.0985	0.9439
Odds	0.6429	0.1292	0.6345	0.1038	0.9396
Accumulative	0.6471	0.1249	0.6395	0.0991	0.9319
Naïve (frequency)	0.6533	0.0962	0.6493	0.0723	0.9456
Naïve (equal)	0.8165	0.0000	0.6666	0.0000	0.9811

Table 4-4: Error measurements and standard deviations of different prediction models during the test period 2016-2017 NHL season, n=1317 games

To make comparison easier, models and benchmarks are ranked based on measurements in Table 4-4. Ranked results are then presented in Table 4-5. RMSE and MAE measurements are combined in the ranked table as these measurements result the same rank order. In error measurements, the model or benchmark ranked as first has the

smallest error according to that measurement. Models and benchmarks are also ranked based on standard deviations, where the one with the smallest standard deviation is ranked first.

Model	RMSE / MAE	MASE	Sd.
Elo Logistic	3	4	4
Elo Normal	2	3	3
Odds	1	2	6
Accumulative	4	1	5
Naïve (frequency)	5	5	2
Naïve (equal)	6	6	1

Table 4-5: Models and benchmarks ranked based on error measurements in Table 4-4, during the test period 2016-2017 NHL season, n=1317 games

RMSE and MAE measurements indicate that odds-based probabilities contain the largest amount of information amongst models and benchmarks having the smallest error in comparison. Similar to Elo applications in other sports, NHL Elo models cannot match or outperform the market odd probabilities in error measurements. Elo Normal is ranked second in RMSE/MAE and third in MASE, while Elo Logistic is ranked third in RMSE/MAE and fourth in MASE. Normal distribution model results to only slightly better results compared to the logistic distribution model. MASE measurement, recommended by Hyndman and Koehler (2006), ranks accumulative model as the best in comparison. However, RMSE and MAE measurements rank it only as the fourth best option. Accumulative model is still able to outperform naïve models even with the latter measurements.

Probabilities on both Elo models, odds and accumulative model were able to outperform the naïve benchmarks. Naturally, naïve prediction models also had the lowest standard deviation. Although odds-based model was the best performing method in RMSE/MAE and second best in MASE, it had the highest standard deviation. Three other non-naïve prediction models were close to each other not only in standard deviation, but also in RMSE and MAE measures. Their ranked order was same in each measure: Elo Normal was slightly better and had lower standard deviation than Elo Logistic, whereas Elo Logistic had lower error and standard deviation compared against accumulative model.

Table 4-6 shows paired t-test results as models were compared against each other. Low p-value indicates that the difference in mean loss of two models is significantly different,

when high p-values indicate that mean losses are not significantly different from each other.

Model	Naïve (Equal)	Naïve (Freq)	Accumulative	Elo Logistic	Elo Normal	Odds
Odds	0.0000	0.0000	0.2052	0,1889	0.2557	-
Elo Normal	0.0000	0.0021	0.8913	0.9906	-	-
Elo Logistic	0.0000	0.0022	0.9006	-	-	-
Accumulative	0.0000	0.0037	-	-	-	-
Naïve (Freq.)	0,0000	-	-	-	-	-
Naïve (Equal)	-	-	-	-	-	-

Table 4-6: Paired t-test results for absolute losses of compared models. Test results as p-values with 4-decimal accuracy

T-test results show that mean losses of odds model, two Elo models and accumulative model are significantly different than the two naïve models. As seen in Table 4-4, the MAE of all these models is lower than the one of naïve based models. This leads to conclusion that the models are significantly better than the naïve benchmarks. However, t-tests show that there isn't enough evidence to draw confident conclusions amongst non-naïve models. Odds seem to have the best possibility to stand out from other models in terms of mean loss. Especially Elo Logistic and Elo Normal models are very close to each other in mean loss.

5 Findings and conclusions

The research questions for this thesis were as follows:

(1) What is the best application of Elo Rating system to measure strength of National Hockey League teams?
(2) Can rating-based probability estimates be accurate in predicting match results and how they compare against other models and benchmarks?

Different versions and applications of Elo models were assessed to answer the first research question. Calibration results showed that goal-based rating adjustment coefficients were not able to enhance results from non-goal-based models. When normal and logistic distribution versions of original Elo rating model were compared visually to their respective distribution's theoretical quantiles with QQ-plots, it was found that Elo Normal model seemed to be a better fit when compared to Elo Logistic model. This indicates that normal distribution is the best model in capturing the performance distributions of NHL teams. However, this difference resulted to only a marginally better accuracy in error measurements during testing period with no statistical significance in t-test.

Naïve predictions, odds-based probabilities and accumulative system were used as benchmarks and compared against Elo models in different error measurements. This was the way to assess the predictive power of Elo models to answer the second research question. Similar to previous research on Elo's predictive power in other sports (i.e. L. Hvattum & Arntzen, 2010), Elo models presented in this research were able outperform naïve prediction models with statistical significance. Although no statistical significance could be drawn on the mean loss being different between Elo models and the bookmaker's odds, the latter had smaller errors in RMSE and MAE measurements. This indicates similar results to previous research in other sports where odds have been found to be superior to Elo rating-based probability models.

Some key findings in addition to these main conclusions will be discussed further in this final chapter of the thesis. Practical and theoretical implications are suggested based on past research and current trends in sports probability modelling. Finally, the limitations of this research, including some suggestions on possible future research are discussed.

5.1 Discussion and implications

As discussed in Chapter 4.1., calibration of goal-based Elo models showed that optimizing k-coefficients led these models to converge towards the non-goal-based models. To find reasons on why goal-based models were not able to provide additional information to models, the coefficients were re-optimized without the constraints requiring them to take positive values. This optimization resulted to negative optimal values for coefficient λ. This would lead to a conclusion that a bigger goal difference in a match would result a smaller rating adjustment. Similarly, this would quite illogically suggest that a small goal difference is more meaningful indicator of team performance compared to a larger goal difference.

One possible reason for this kind of conclusion can be found when comparing the nature of ice hockey to football, which was the sport that this goal-based model was based upon. Unlike in football, a goalie in ice hockey match is often taken to the bench on the last minutes of the game to allow a sixth player on the field to join the attack. According to MoreHockeyStats website, NHL teams allowed on average 9.87 empty net goals during the test period season 2016-2017 (MoreHockeyStats, 2019). Therefore, these situations seem to increase the goal difference in a match. Because of the changed situation, these empty net goals can't be considered to reflect true strength difference between teams.

As was seen in Figure 4-1 and discussed shortly in Chapter 4.1., rating values and statistics seemed to be further away from chosen reference point in Elo Logistic model compared against the Elo Normal model, although scale parameters were chosen to be as comparable as possible. This seemingly large difference is due to the large choice of standard deviation for the models. Large standard deviation parameter highlights the difference of normal and logistic distributions: tails of logistic distribution are slightly heavier, meaning that fewer values are therefore closer to the mean. Conversely, more values are further away from the mean.

When rating difference is small between two teams, rating adjustments after their encounter are also smaller as difference on expected and actual results cannot take high values. When rating difference is high, rating adjustments are also higher if the actual result of the match is unexpected. For logistic distribution these unexpected results on the tails of performance curve occur more often, leading to larger adjustments when compared against normal model. This explains differences between the rating values of models.

QQ-plot Figures 4-2 and 4-3 indicated that normal distribution would be better in modelling teams' performance in NHL compared against logistic distribution. Possible reason for this is the salary cap introduced to NHL for the first seasons of our data. Salary cap has been proven to strengthen the economical competitiveness in NHL, but the effects on sport performance has not yet been found to be significant (Büschemann & Deutscher, 2011). As more even competition could indicate the performances of NHL teams to be homoscedastic, logistic distribution wouldn't be required to be used to capture heteroskedasticity in performances.

Results of this research are also further proof on Elo rating systems' ability to capture statistically significant information on team's strength in sports. A consensus remains amongst academic literature on Elo's ability to measure team strengths and to outperform naïve forecasts in different sports. Results prove that Elo ratings have a practical implication for measuring team strength in NHL context.

Elo models can be utilized to practice for example in betting odds evaluation or in defining teams' performance-based success by the competition organizers. Kontinental Hockey League (KHL) is an example of an organization that uses ranking system to make overall evaluation for teams. The current sport achievement model of this system could be evaluated against Elo rating system to see whether Elo rankings would be a better measure in this category (Kontinental Hockey League, 2018).

In addition to Elo's ability to evaluate team strength, research also provided further proof on Stefani's (2011) claim that when correctly formed, both accumulative and adjustive systems are able to identify team's strengths in sports. Accumulative benchmark prediction method was significantly better than the naïve predictions and it was the best model in MASE measurement, although no statistical difference was found when compared against Elo and odds-based prediction models. It is however important to notice that weighing and ageing methods for accumulative benchmark system were entirely arbitrarily chosen. Different methods on ageing and weighing could be tested to see if accuracy of accumulative model could be further enhanced.

5.2 Limitations and future research

It must be noted that this research is limited to only National Hockey League context. Because of differences between different ice hockey leagues, for example regarding the salary cap, it is left for future research to evaluate different Elo models in other ice hockey leagues to see whether i.e. the normal model is better compared to logistic model

in other leagues or generally in ice hockey. Performance in this research was also evaluated based on the full time result of NHL match. Other indicators of performance and their ability could be tested to see if the results would differ. These other indicators include for example results after overtime or shootout competition and the use of non-score-based indicators, such as expected goals.

As discussed in Chapter 3.4., the use of RMSE as calibration measurement was undesirable in calibration for issues regarding reliability. It was nevertheless used with no other sufficient measure available with some other previous researches being also based on squared error measure calibrations. Unreliability of the calibration results may cause the *k*-values to differ significantly on different time periods, which should be accounted on evaluating the results of this research and when planning on future research on the subject.

The future possibilities on modelling probabilities in ice hockey and other sports increases constantly as more and more advanced statistics become available with richer and more accurate data. For example, National Hockey League will deploy a new tracking technology for next season, providing real-time data with inch-level accuracy (National Hockey League, 2019). These possibilities open a chance to better measure team performances, calibrate the changes in strength and to form better predictions on match probabilities, all of which can be tested and applied in order to enhance the predictive power of Elo rating systems.

References

Armstrong, J., & Collopy, F. (1992). Error Measures for Generalizing about Forecasting Methods: Empirical Comparisons. *International Journal of Forecasting*, *8*, 69–80.

Büschemann, A., & Deutscher, C. (2011). Did the 2005 collective bargaining agreement really improve team efficiency in the NHL? *International Journal of Sport Finance*, *6*(4), 298–306.

Elo, A. (1987). *The Rating of Chess Players, Past and Present*. *New York: Arco*. Retrieved from http://www.amazon.com/Rating-Chess-Players-Past-Present/dp/0668047216

FIFA.com. (2017, September 14). OC for FIFA Competitions approves procedures for the Final Draw of the 2018 FIFA World Cup. Retrieved from https://www.fifa.com/about-fifa/who-we-are/news/oc-for-fifa-competitions-approves-procedures-for-the-final-draw-of-the-2907924

FIFA. (2018). Revision of the FIFA / Coca-Cola World Ranking Confederations Cup matches. Retrieved February 19, 2019, from https://resources.fifa.com/image/upload/revision-of-the-fifa-coca-cola-world-ranking.pdf?cloudid=fzltr4s8tz3v3vy0aq01

Greene, W. H. (2012). *Econometric analysis* (7th ed). Boston : Prentice Hall/Pearson cop. 2012.

Hockey Reference. (2019). NHL Scores 2005-2017. Retrieved February 6, 2019, from https://www.hockey-reference.com/

Hvattum, L., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, *26*(3), 460–470. https://doi.org/10.1016/j.ijforecast.2009.10.002

Hvattum, L. M. (2017). Ordinal versus nominal regression models and the problem of correctly predicting draws in soccer. *International Journal of Computer Science in Sport*, *16*(1), 50–64. https://doi.org/10.1515/ijcss-2017-0004

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Kelly, J., J. (1956). A new interpretation of information rate. *IRE Transactions on Information Theory*, *2*(3), 917–926.

Kontinental Hockey League. (2018, July 2). How the KHL teams ranking changed in 2017/2018. *Kontinental Hockey League (KHL)*. Retrieved from https://en.khl.ru/news/2018/07/02/401870.html

Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports*, *12*(3), 127–138. https://doi.org/10.1515/jqas-2015-0059

Lasek, J., Szlávik, Z., & Bhulai, S. (2013). The predictive power of ranking systems in association football Zoltán Szlávik Sandjai Bhulai. *Int. J. Applied Pattern Recognition, Vol. 1, No. 1*, *1*(1), 27–46.

MoreHockeyStats. (2019). Empty Net Performance stats for NHL teams, season 2016-2017. Retrieved April 16, 2019, from https://morehockeystats.com/teams#

National Hockey League. (2019, January 26). NHL plans to deploy Puck and Player Tracking technology next season. Retrieved from https://www.nhl.com/news/nhl-plans-to-deploy-puck-and-player-tracking-technology-in-2019-2020/c-304218820

Odds Portal. (2019). NHL Bookmaker odds for season 2016-2017. Retrieved February 6, 2019, from https://www.oddsportal.com/

Stefani, R. (2011). The methodology of officially recognized international sports rating systems. *Journal of Quantitative Analysis in Sports*, *7*(4). https://doi.org/10.2202/1559-0410.1347

Štrumbelj, E., & Vračar, P. (2012). Simulating a basketball match with a homogeneous Markov model and forecasting the outcome. *International Journal of Forecasting*, *28*(2), 532–542. https://doi.org/10.1016/j.ijforecast.2011.01.004

Witten, I. H., & Ian, H. (2005). *Data mining: Practical machine learningn tools and techniques*.