Chapter 2

Preliminary Analysis

2.1 Challenges

In the attempt to tackle the problem, various challenges arise that are worth mentioning.

2.1.1 Feature Selection

Features are pieces of information that may be useful for predictions. It is generous of Kaggle to provide a comprehensive collection of data. However, not all data is relevant to making predictions. Before a set of features can be selected, careful choices must be made on which data sets to use and which portion of that data set is relevant to solving the problem.

Regardless of the machine-learning technique used, the most obvious choice of data sets to use is the historical match records for both regular and tournament seasons. However, some data sets contain match records that date back to as early as 1985, which are no longer relevant in today's context. After all, the NCAA tournament teams consist of *college students* who can only stay with a team for a maximum of four years before graduation. As players constantly leave and join the team, it is difficult to quantify the effect brought by changes in a team's composition on the strength of that team, given only the team-level data. Moreover, there were some substantial updates on the Tournament's rules at the beginning of the 2008-2009 season [4], which also affected the strategies teams used in the subsequent tournaments. Therefore, only the most recent match records are useful for prediction. For the purpose of this project, a four-year window from 2013 to 2016 is selected and all data used in this project fall within this window. In addition, only the match records from the regular seasons are actually used, since there are only a few records from the tournament seasons which add little value to making predictions.

There are a number of potential features that can be selected from the data set **RegularSeasonDetailedResult**. The potential features are further categorised as

basic and additional features. Table 2.1 gives a summary of the basic features.

Feature	**Description**
Wteam	The id number of the team that won the game
Wscore	The number of points scored by the winning team
Lteam	The id number of the team that lost the game
Lscore	The number of points scored by the losing team
Wloc	The location of the winning team
Numot	The number of overtime periods in the game

Table 2.1: A Summary of the Basic Features

The additional features describe both *offensive* and *defensive* strengths of both winning team and losing team. Table 2.2 gives a summary of features describing offensive strengths, while Table 2.3 focuses on defensive strengths. Although the two tables list features from the winning team's perspective, another duplicated set of features also exists for the losing team.

Feature	**Description**
Wfgm	The number of field goals made by the winning team
Wfgm3	The number of three pointers made by the winning team
Wftm	The number of free throws made by the winning team
Wor	The number of offensive rebounds by the winning team
Wast	The number of assists by the winning team
Wstl	The number of steals by the winning team
Wblk	The number of blocks by the winning team

Table 2.2: A Summary of the Offensive Features

Feature	**Description**
Wfga	The number of field goals attempted by the winning team
Wfga3	The number of three pointers attempted by the winning team
Wfta	The number of free throws attempted by the winning team
Wdr	The number of defensive rebounds by the winning team

Table 2.3: A Summary of the Defensive Features

The potentially many features make model selection challenging, because they create an exponential number of possible models. If a feature has m possible values,

to select the best models based on n features, at least $\Theta(n^m)$ models have to be examined, which is impractical given that only one submission is allowed at one time. Therefore, this project uses a greedy strategy where all other features are fixed at their optimal values when one feature varies to select a suboptimal model.

2.1.2 Interrelationship

No team is in isolation. The tournament matches are interactive processes whereby complex interrelationships exist amongst all the contesting teams, which adds another layer of complication in the attempt to predict match results. For example, given the history match records that Team A beaten Team B and Team C lost to Team B, one should intuitively conclude that Team A should have a higher probability of beating Team C. But what if another record shows that Team A once lost to Team C? In that case, the relative strength levels of the three teams will be hard to determine based purely on that intuition. Moreover, the match whose result is to be predicted may be the very *first* match ever between two teams. In other words, there are no historical records that give a direct assessment on the two teams' relative strengths. Such lack of knowledge must be complemented by some forms of inference based on the interrelationships amongst the teams. So a good predictive model should not only be able to take into consideration the current game record, but also explore the interrelationships amongst all the game records and generalise on unseen matches.

2.1.3 The Curse of Model Popularity

The problem to solve can be categorised as a classification problem under supervised learning. But since the final answers to be submitted are in fact *probabilities*, non-probabilistic models like support vector machines will hardly work. Therefore, the first a few models attempted in this project include the most popular ones: logistic regression and multilayer perceptrons.

The inputs to these popular models are the seed positions of each team, since the seeds represent an official view on the relative strength of each team. However, the performances of these models are not as good as expected: the logistic model only gives a 283^{rd} position on the leaderboard. Moreover, only teams in a tournament are assigned a seed but the number of tournament matches is not large enough to support accurate predictions. Although [2] suggests a combination of logistic regression and Markov chain to predict match outcomes, these popular models alone based on simply seeds are unlikely to give a good result.

2.2 Approach

In light of the challenges identified, the general approach is to quantitatively describe the *skill* of each team, from which the winning probabilities are derived. This section gives a rough idea about the approach used in this project, while the detailed mathematical principles are left to Chapter **??**.

The system used to estimate the skill of a team is known as a *rating system*. An example of a rating system is the famous Elo rating system widely applied in board games like chess and Go. In a Gaussian rating system, each team i is assumed to have a skill s_i that follows a Gaussian distribution $s_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, where μ_i is the mean skill of a team and σ_i can be considered as the system's *belief* about the team's skill level.

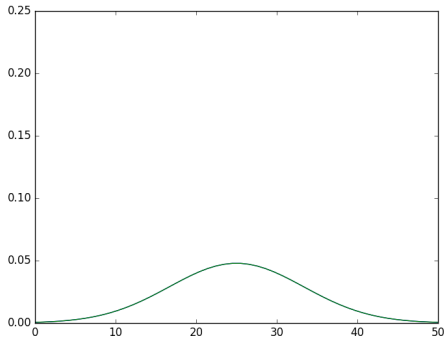
Depending on the rating system, the skill of a team is updated after a single match or a series of matches within a specific period. The amount of such a update depends on how *surprising* a match outcome is — if, prior to a match, Team A is expected to have a very high probability of beating Team B based on their *skill gap*, an outcome that Team A actually wins the game will *not* result in a significant update on either Team A's or Team B's skill; however, on the other hand, if it turns out that Team B performs superbly and beats Team A at the end of the match despite its lower skill level, which is known as an "upset" in sport term, there *will* be a significant increase in Team B's skill and a significant decrease in Team A's skill. But in either case, the belief of skills σ_i will shrink because the rating system becomes more confident about the two teams' skill level.

As an example, suppose two teams have the same initial skill estimates of s_i and s_j respectively as shown in Figure 2-1a. As the tournament begins, s_i and s_j are adjusted according to the outcomes of the matches, which is represented by the shift of skill curves as shown in Figure 2-1b and 2-1c. In the meanwhile, the skill curves become increasingly *taller*, indicating that the rating system is increasingly confident about their skill levels. Finally when the tournament ends, the rating system reports the final standings of the two teams as shown in Figure 2-1d.

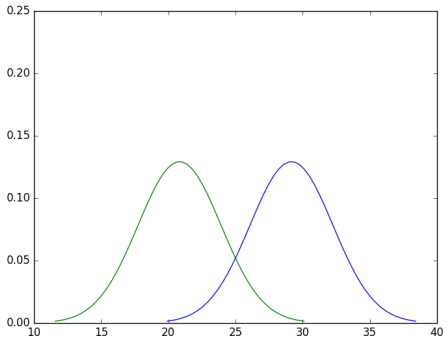
A team's probability of beating another team in an *upcoming* match can be estimated from the two teams' skill difference.

$$P(s_i > s_j) = \Phi\left(\frac{\mu_i - \mu_j}{\sqrt{\sigma_i^2 + \sigma_j^2}}\right) \tag{2.1}$$

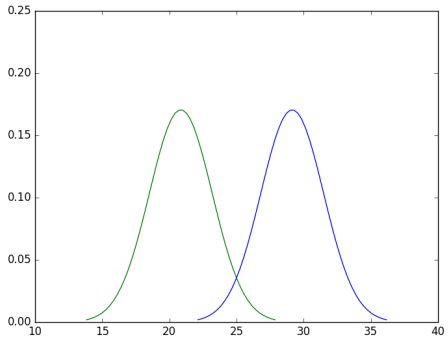
where Φ is the cumulative distribution function of a standard Gaussian distribution.



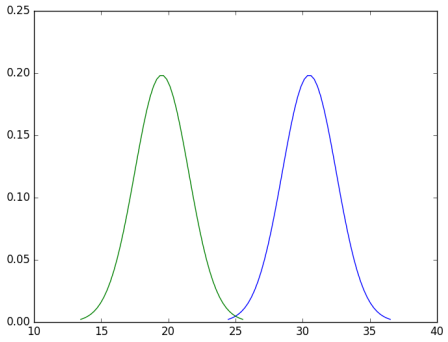
(a) Timestamp $t = 0$



(b) Timestamp $t = 1$



(c) Timestamp $t = 2$



(d) Timestamp $t = 3$

Figure 2-1: An Example of Skill Update