THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

The National Collegiate Athletic Association (NCAA) Division I Men's Basketball Tournament is one of the most popular annual sport festivities in the United States. Every year, the Tournament attracts a sizeable pool of audience, with the national champion becoming one of the hottest topics throughout the whole year. In the meanwhile, interests in predicting the winning team of a particular tournament match are escalating [2] and online machine-learning communities like Kaggle are organising annual competitions to encourage creative solutions. The purpose of this project is to present a feasible machine learning-based solution to Kaggle's competition "March Machine Learning Mania 2016" [1] that accurately predicts a team's *probability* of winning a particular tournament match based on historical match data provided by Kaggle.

1.1 Background

According to Wikipedia [4], the Tournament is played during every March and April based on the rule of single-elimination. Out of the 68 participating college basketball teams, 32 *conference* match champions are automatically qualified for the Tournament, while the other 36 teams are admitted at the discretion of a NCAA selection committee based on a criterion known as Rating Percentage Index [5].

The total 68 teams are then ranked by the selection committee from 1 to 68, and distributed amongst the four regions nominally known as East, West, South and Midwest. The top four teams receiving a rank from 1 to 4 are distributed to and given a *seed* of 1 in each of the four regions, followed by the next four teams with a rank of 5-8 that receive a seed of 2 in each region. The process continues until only the last eight teams are left whereby they have to fight with one of the other teams for the 16^{th} seed position for each region, which marks the commencement of

the Tournament and is officially known as the *First Four* round.

At this point, the number of contesting teams are reduced to 64. During the next *First Round* in each region, a team with a higher seed position plays against a team with a lower seed position. For example, there are matchups between teams with the 1^{st} seed and teams with the 16^{th} seed, between the 2^{nd} seed teams and the 15^{th} seed teams and so on. Subsequently, the 32 winning teams advance to the *Second Round* where they play against one of the other teams, after which the 16 remaining teams are known as the *Sweet Sixteen*.

The number of contesting teams continues to halve until the four regional champions are determined. During the *National Semi-final*, the regional champion with the 1^{st} seed position plays against the regional champion with the 4^{th} seed position, while the other two teams play against each other. *National Final* is the last round and conducted between the two winners of the National Semi-final to determine the National Champion. However, there is no consolation game for the third place in the Tournament.

1.2 Problem & Evaluation Method

The problem of this Kaggle competition is to predict a team's probability of winning a particular tournament match. In a real tournament, there should be 68 teams and 67 matches in total because of the rule of single elimination (only the champion is not *eliminated* after the 67 matches). But how teams are paired up in these matches are not known in advance, except for the first 32 matches whose contesting pairs can be derived from the seed results. Therefore, Kaggle requires participants submit predictions for all possible matchups between any two of the 68 teams, which amount to $68 \times (68-1)/2 = 2278$ matches according to the Handshaking lemma.

According to Kaggle [1], the method for evaluating predictive models and compiling the leaderboard is based on cross entropy or log loss:

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \ln \hat{y}_i + (1-y_i)\ln(1-\hat{y}_i)] \tag{1.1}$$

where

- n is the number of games played
- $\hat{y}_i$ is the predicted probability of team 1 beating team 2
- y_i is 1 if team 1 wins, 0 if team 2 wins

The goal of any predictive models is to *minimise* the log loss. Based on the LogLoss evaluation method, a completely incorrect prediction will lead to a score of ∞, for example, a prediction of 1 when the actual outcome is 0. To avoid such an unpleasant score, Kaggle uses a threshold function to scale the submitted probability into a reasonable range as follows:

$$p' = \max(\min(p, 1 - 10^{-15}), 10^{-15}) \tag{1.2}$$

where p is the submitted probability and p' is the adjusted probability.

When using the LogLoss function to evaluate submissions, Kaggle assumes the ground truth variable y_i is discrete, namely $y_i \in \{0, 1\}$. But in general, y_i should be a continuous variable representing the *true* probability that Team 1 will beat Team 2, and $\hat{y}_i$ still denotes the *predicted* probability. In this case, the cross entropy measures how close these two probabilities are. If there exists a *perfect knowledge predictor* that *always* gives correct predictions, namely $\hat{y}_i = y_i$ always holds, then its prediction performance can be described in Figure 1-1. Moreover, if the true probability follows an uniform distribution $y_i \sim U(0, 1)$, then in the long run the perfect knowledge predictor will have an expected score of

$$-\int_0^1 [y_i \ln y_i + (1 - y_i) \ln(1 - y_i)] dy_i = 0.5 \tag{1.3}$$

which is graphically equivalent to the height of a rectangle that shares the same base of length 1 as the performance curve's.

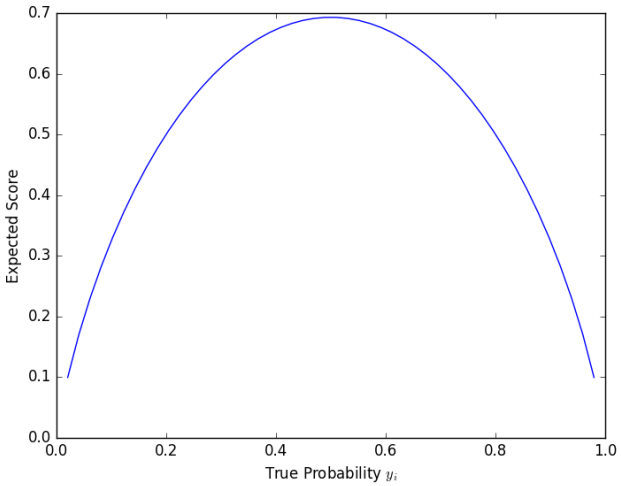

Figure 1-1: Performance of a Perfect Knowledge Predictor

1.3 Data Sets

Kaggle provides almost three decades' data about NCAA basketball matches, including matches during the regular seasons as well as the tournaments. The data sets are summarised in Table 1.1.

Data Set	**Description**
RegularSeasonCompactResults	Game-by-game results during regular seasons from 1985-2015
RegularSeasonDetailedResults	More detailed results during regular seasons from 2003-2015
Seasons	Different seasons present in the dataset
Teams	Different college teams present in the dataset
TourneyCompactResults	Game-by-game tournament results from 1985-2015
TourneyDetailedResults	More detailed tournament results from 2003-2015.
TourneySeeds	The seeds for all teams in a tournament
TourneySlots	The pair-ups between two teams based on their seeds

Table 1.1: A Summary of the Data Sets

However, not all data sets are useful. Only two of them are used in this project, namely **RegularSeasonCompactResults** and **Teams**.

1.4 Role Assignment

Table 1.2 shows the assignment of roles for each team member in this project.

Team member	**Role**	**Responsibility**
Joe Tan Chin Yong	Data Scientist	To build predictive models
Liu Zeyan	Data Analyst	To optimise selected models
Wei Yumou	Data Scientist	To build predictive models
Xie Dai	Statistician	To provide mathematical knowledge

Table 1.2: Role Assignment