HSR
HOCHSCHULE FÜR TECHNIK
RAPPERSWIL
FHO Fachhochschule Ostschweiz

# NBA Game Prediction

Projekt Arbeit

B. Kühnis

Januar 1, 2018

Advisors: Prof. Dr. Farhad D. Mehta
Abteilung Informatik, IFS, HSR

**Zusammenfassung**

This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes.

# Inhaltsverzeichnis

Kapitel 1

# Einleitung

Das Projekt beschäftigt sich mit der Frage, wie sehenswürdig ein Basketballspiel ist. Dabei werden Spiele, die bereits beendet wurden, bewertet und deren Sehenswürdigkeit in einer Skala ausgegeben. Die Schwierigkeit der Fragestellung ist die Definition des Begriffs Sehenswürdigkeit. Dieser ist subjektiv und somit gibt es für diese Frage keine allgemeingültige Antwort.

Die Arbeit beschäftigt sich neben der Analyse, wie sehenswürdig ein Basketball ist, auch mit der Frage, wie der Begriff Sehenswürdigkeit in Bezug auf ein Basketballspiel überhaupt definiert werden kann.

## 1.1 Ausgangslage

Das Projekt wird von Grund auf neu entwickelt. Die Idee entstand durch das Überangebot an Spielen, welche an einem Tag durchgeführt werden. Zusätzlich kommt noch die geographische Komponente dazu. Diese hat zur Folge, dass wegen der Zeitverschiebung Spiele oft nicht live gesehen werden können. Zu einem späteren Zeitpunkt gibt eine Auswahl an Spielen[1]. Dabei ist es schwierig, das sehenswürdigste Spiel auszuwählen. Die Auswahl kann einerseits durch die Vorlieben für ein oder mehrere Teams getroffen werden oder es werden die Resultate angeschaut, wobei dies die Spannung des Spieles mindert. Was also fehlt ist ein Richtwert, welcher die Sehenswürdigkeit des Spieles bestimmt[2].

[1] Für gewisse Personen sicherlich nicht verständlich, warum ein Spiel, welches in der Vergangenheit liegt, noch angeschaut wird. Als Vergleich kann ein Basketballspiel mit einem Film verglichen werden: auch wenn der Film schon ein paar Tage alt ist, solange die Handlung nicht bekannt ist, kann dieser trotzdem spannend sein.

[2] Wenn das Basketballspiel wieder mit einem Film verglichen wird, so wäre der Richtwert wie die Kritik des Filmes.

## 1.2 Ziel der Arbeit

Das Ziel der Arbeit ist es, der Person, welche ein Basketballspiel schauen möchte, einen Richtwert zu geben, wie spannend dieses Spiel war, ohne dabei das Resultat zu verraten.

Kapitel 2

# Problembeschreibung

NBA[1] Spiele werden in Nordamerika ausgetragen[2]. Personen aus anderen Regionen können wegen der Zeitverschiebung die Spiele oft nicht live mitverfolgen. Wenn eine Person zu einem späteren Zeitpunkt ein Spiel sehen möchte, hat sie oft mehrere Spiele zur Auswahl. Die Frage ist nun, welches das interessanteste Spiel war, ohne das Resultat vorher anzuschauen (Abschnitt 2.2). Eine Lösung dazu wäre, das Spiel des Lieblingsteams auszuwählen. Dabei ist jedoch nicht klar, ob dies das spannendste Spiel ist. Es muss also ein Richtwert erstellt werden, welcher die Sehenswürdigkeit des Spieles angibt, um bei der Auswahl des Spieles den Benutzer zu unterstützen. Die Frage ist nun, wie dieser Richtwert definiert wird.

## 2.1 Sehenswürdigkeit

Um den Begriff Sehenswürdigkeit im Bezug auf ein Basketballspiel zu definieren, müssen die Daten zu einem Spiel analysiert werden. Am Ende ist die Sehenswürdigkeit jedoch subjektiv. Wenn Daten zum Benutzer vorhanden sind, soll eine genauere Sehenswürdigkeit berechnet werden (Abschnitt 3.2), wie für den Fall, in dem keine Daten zum Benutzer (Abschnitt 3.1) vorhanden sind. Zusätzlich müssen Kriterien gefunden werden, um die Sehenswürdigkeit zu quantifizieren.

## 2.2 Spoiler-Free

Die Wertung der Sehenswürdigkeit soll keine Rückschlüsse auf den Verlauf des Spieles ermöglichen und somit spoiler-free sein. Wenn der Algorithmus nur Spiele einer bestimmten Kategorie (zB. Spiele, die in die Verlängerung

[1] National Basketball Association

[2] Es gibt eine kleine Anzahl von exhibition Spiele, welche an anderen Orten ausgetragen werden.

gehen) als sehenswürdig bewertet, mindert dies die Spannung, da der Verlauf teilweise preisgegeben wird.

Kapitel 3

# Lösungskonzept

Bei der Lösung werden zwei Use Cases beachtet. Einerseits die allgemeine Sehenswürdigkeit (Abschnitt 3.1) für anonyme Benutzer und die personalisierte Sehenswürdigkeit (Abschnitt 3.2) für Benutzer, zu welchen es zusätzliche Daten gibt. In den nächsten Kapitel werden für jeden Use Case verschiedene Lösungen vorgestellt.

## 3.1 Allgemeine Sehenswürdigkeit

Die allgemeine Sehenswürdigkeit zeigt für anonyme Benutzer auf, wie spannend ein Spiel ist. Dabei werden verschiedene Ansätze verfolgt. Die perfekte Lösung gibt es nicht, da, wie schon erwähnt, die Sehenswürdigkeit subjektiv ist.

### 3.1.1 Spielwertung durch die Benutzer

Wie oft auch auf anderen Webseiten verwendet[1] können Benutzer ein Spiel bewerten. Diese Wertung gibt dem anonymen Benutzer einen Richtwert, wie spannend das Spiel war.

**Vorteile:** Die Bewertung gibt genau den Durchschnitt der Meinungen über ein Spiel wieder.

**Nachteile:** Damit die Bewertung aussagekräftig ist, braucht es mehrere Wertungen, damit eine zuverlässige Aussage getroffen werden kann und statistische Ausbrüche vermieden werden.

### 3.1.2 Sehenswürdigkeit Modell

Mittels den Daten zu einem Spiel wird ein Modell erstellt, welches Spiele nach deren Spannung bewertet. Dabei liegt die Schwierigkeit bei der Beur-

[1] Amazon, Digitec usw..

teilung, ob das Modell das richtige Resultat liefert. Indem der Benutzer ein Spiel mit einem "+öder einem "bewerten kann (Unterabschnitt 3.1.1), ist es möglich, das Resultat des Modells zu beurteilen und das Modell mit der Zeit zu verfeinern.

**Vorteile:** Das Modell kann ohne Benutzerdaten auskommen und wenn Benutzerdaten vorhanden sind, welche das Modell bewerten, ist es möglich, dieses zu verbessern.

**Nachteile:** Die berechnete Sehenswürdigkeit ist stark vom Modell abhängig und es ist schwierig, ein perfektes Modell zu erstellen.

### 3.1.3 Supervised Learning: Regression

Superviesed Learning eignet sich sehr gut um ein Modell zu finden, welches zu den Daten passt. Dabei werden Statistiken von einem Spiel als Input-Werte verwendet. Zu jedem Spiel gibt es noch ein Resultat-Wert, der angibt, wie sehenswürdig das Spiel war. Mittels den Daten ist es möglich, eine Funktion zu approximieren, welche möglichst genau für alle Input-Werte eines Spieles die entsprechende Resultat-Werte ausgibt.

Die Schwierigkeit bei diesem Ansatz liegt darin, gute Resultat-Werte zu finden. Die folgenden Kapitel zählen einige Ideen auf, wie solche Resultat-Werte aussehen können.

#### Bild/Video/Ton Analyse

Die Aufnahmen des Spiel werden analysiert. Dabei werden Bild, Video und Ton nach deren Sehenswürdigkeit bewertet. Aus diesen Daten wird die Sehenswürdigkeit eines Spieles bestimmt. Dieser Ansatz ist sehr aufwändig umzusetzen. Wenn die Analyse sehr gut ist, kann es zu einem sehr guten Resultat führen, da weit mehr Daten gesammelt werden können als nur die Statistiken von einem Spiel.

**Vorteile:** Neben den Statistiken über ein Spiel werden zusätzliche Daten beachtet.

**Nachteile:** Die berechnete Sehenswürdigkeit anhand von Bild/Video/Ton ist schwierig und aufwändig.

#### Zuschauerzahlen

Indem die Anzahl der Zuschauer pro Spiel betrachtet wird, kann eventuell eruiert werden, wie spannend ein Spiel ist. Die Idee dahinter ist, dass ein Spiel mit mehr Zuschauer interessanter ist, als eines mit weniger. Die Zuschauerzahl kann entweder durch die Anzahl an Stadionbesucher oder die Fernseheinschaltquote gemessen werden.

**Vorteile:** Die Sehenswürdigkeit ist abhängig von der Anzahl der Benutzer, woraus sich schliessen lässt, dass die Sehenswürdigkeit der Meinung der meisten Benutzer entsprechen wird.

**Nachteile:** Nicht in jeder Region hat es gleich viele Menschen, somit werden gewisse Teams immer mehr Zuschauer haben als andere. Zusätzlich kommt dazu, dass die Zuschauerzahlen nach dem Spiel sich nicht mehr ändern, was dazu führt, dass Methoden wie Social Media Echo, die auch Daten nach dem Spiel betrachten, eine genauere Aussage treffen können bezüglich der Sehenswürdigkeit eines Spiels.

#### Anzahl Highlights

Pro Tag werden über alle Spiele Highlights erstellt. Es wird davon ausgegangen, dass das Spiel mit den meisten Highlights das interessanteste war.

**Vorteile:** Highlight sind einfach zu quantifizieren.

**Nachteile:** Das Problem dabei ist, dass Highlights immer erstellt werden, auch wenn kein Spiele an einem Tag sehenswürdig sind.

#### Social Media Echo

Das Social Media Echo wird darüber bestimmt, wie viel über ein Spiel in den Social Media berichtet wird. Dabei wird die Annahme getroffen, dass wenn mehr über ein Spiel berichtet wird, dieses spannender ist.

**Vorteile:** Gleich wie bei der Anzahl der Zuschauer ist beim Social Media Echo das Resultat abhängig von der Benutzeranzahl.

**Nachteile:** Es ist unklar, ob das Social Media Echo die Sehenswürdigkeit reflektiert, da auch Kommentare darüber, wie langweilig ein Spiel ist, beachtet werden.

#### Reddit Score

Auf Reddit wird pro Spiel ein Thread erstellt, wobei Benutzer den Thread Up- oder Down-voten können. Daraus resultiert ein Score, welcher als Quantifizierung von Sehenswürdigkeit betrachtet werden kann.

**Vorteile:** Wie bei der Anzahl der Zuschauer ist der Reddit Score abhängig von der Anzahl Benutzer.

**Nachteile:** Die Relation von Score zu Sehenswürdigkeit stimmt eventuell nicht, da die Anzahl Reddit-Benutzer über die Zeit steigt. Somit erhalten aktuelle Threads einen besseren Score als ältere Threads, da auf Reddit vor allem neue Threads bewertet werden.

## 3.2 Persönliche Sehenswürdigkeit

Um die persönliche Sehenswürdigkeit zu bestimmen, werden neben den Daten zu einem Spiel auch Daten vom Benutzer betrachtet. Dabei soll eine personalisierte und genauere Wertung der Sehenswürdigkeit des Spiele entstehen.

### 3.2.1 Unsupervised Learning: Kategorisierung

Indem sich der Benutzer registriert und Spiele nach deren Sehenswürdigkeit bewertet, können mittels Unsupervised Learning die Spiele in Kategorien aufgeteilt werden. Wenn der Benutzer nun oft Spiele einer gewissen Kategorie bewertet, können dem Benutzer Spiele einer Kategorie als spannend oder langweilig vorgeschlagen werden.

**Vorteile:** Durch die Kategorisierung können persönliche Vorschläge erstellt werden und eine genauere Wertung entsteht. Indem der Benutzer weitere Spiele bewertet, werden die Vorschläge laufend verbessert

**Nachteile:** Die Kategorisierung der Spiele kann so aufgeteilt sein, dass für ein Benutzer keine Kategorie als spannend betrachtet wird. Dies führt dazu, dass die Kategorisierung verfeinert werden müsste, was mit Mehraufwand verbunden ist. Trotzdem kann nicht garantiert werden, dass eine zufriedenstellende Kategorisierung erstellt werden kann.

### 3.2.2 Persönliche Vorlieben

Eine andere Variante wäre, dass das allgemeine Modell (Unterabschnitt 3.1.2) gewichtet wird. Dabei muss der Benutzer angeben, für welche Werte des Modells, z.B. das Team, die Spieler oder eine bestimmte Statistik, er sich interessiert, damit das Modell dementsprechend gewichtet werden kann.

**Vorteile:** Das Modell kann granulös eingestellt werden und für ältere Spiele kann die Sehenswürdigkeit einfach berechnet werden.

**Nachteile:** Das Resultat hängt davon ab, wie gut das Modell ist und durch die Gewichtung können dank dem Resultat Rückschlüsse gezogen werden, wie der Verlauf eines Spieles war.

Kapitel 4

# Umsetzung

Zusammenfassung der Umsetzung:
1. prototyp extrahierung + aufbreitung Liniar regression
2. Erste Umsetzung der Webseite
3. Daten erstmal richtigtrainiert mit gültigen y
3.1 Resultat sehr schlecht
4. Neue Daten hinzugefügt
4.1 Reddit Werte hinzugefügt
4.2 Normalisierung der Resutlate
4.3 Tranierung =¿ besseres resultat
5. Asuwertung und Deutung der Resultate.
5.1 Evt Resultat Normalisiere
6. Webseite finalisierung
7. Integration zu Wikihoops (noch abzuklären)

## 4.1 Prototyp

Beschreibung vom Prototyp, Darstellung der Daten und erklären warum diese so gut waren.

## 4.2 Webseite

Webseite nur rudimentär + abklärungen mit Wikihoops

## 4.3 Sehenswürdigkeit

Nur persönliche Sehenswürdigkeit beachtet, einerseits aus Zeitlichen Gründen und anderseits, weil die Daten fehlen

### 4.3.1 Eigene Formel

Hier wurde bisher einfach die Star-Berechnung von Wikihoops verwendet.

### 4.3.2 Reddit Rating

Erklären wie Reddit Rating zusammengesetzt wird. Daten darstellen und wie die Daten Normalisiert wurden. Resultat des gelernten Präsentieren

### 4.3.3 Wikihoops

Erklären wie Wikihoops Daten zusammengesetzt sind. Daten darstellen und wie die Daten Normalisiert wurden. Erwähnen, dass Wikihoops eine eigene Formel verwendet für die Berechnung der Stars

### 4.3.4 Resultat Analyse

Vergleich von den drei Ansätze, was ähnelt sich am ehesten. Wie können die Daten normalisiert werden, gibt es ein Durchschnitt?

Kapitel 5

# Ergebnis

Umsetzung eines einfachen Protyps und Analyse wie sehenswürdigkeit definiert werden kann.

## 5.1 Offene Punkte

## 5.2 Ausblick

Anhang A

# Dummy Appendix

You can defer lengthy calculations that would otherwise only interrupt the flow of your thesis to an appendix.

# Literaturverzeichnis