



Duale Hochschule Baden-Württemberg Mannheim

Projektreport

Bundesliga Match Predictions

Studiengang Wirtschaftsinformatik

Studienrichtung Data Science

| | |
|-----------------------|--|
| Verfasser: | Max Bernauer, Philipp Dingfelder, Julius Könning |
| Matrikelnummer: | 5763624, 8687786, 7305370 |
| Firma: | SAP SE, Schaeffler |
| Modul: | Data Exploration |
| Kurs: | WWI20DSB |
| Dozent: | Simon Poll |
| Bearbeitungszeitraum: | Sommersemester 2022 |

Ehrenwörtliche Erklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit mit dem Titel “*Bundesliga Match Predictions*” selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Ort, Datum

Max Bernauer, Philipp Dingfelder, Julius Könning

Inhaltsverzeichnis

| | |
|--|------------|
| Abbildungsverzeichnis | iii |
| Tabellenverzeichnis | iv |
| Abkürzungsverzeichnis | v |
| 1 Einleitung | 1 |
| 1.1 Hintergrund und Motivation | 1 |
| 1.2 Business Use Case | 1 |
| 2 Theoretische Grundlagen | 2 |
| 2.1 Related Work | 2 |
| 2.2 Verwendete Technologien und Bibliotheken | 2 |
| 3 Praktischer Teil | 3 |
| 3.1 Umsetzung | 3 |
| 3.2 Ergebnisse | 3 |
| 4 Fazit | 4 |
| Anhang | |

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

| xxx | Description |
|-----|-------------|
|-----|-------------|

1 Einleitung

1.1 Hintergrund und Motivation

Dieser Projektreport ist im Rahmen des Fachs Data Exploration entstanden. Das Ziel des Moduls ist die „Anwendung von Methoden und Verfahren des maschinellen Lernens auf eine vorgegebene Datenbasis unter Laborbedingungen“[Modulhandbuch]. Zusätzlich soll neben der informatischen Betrachtung auch der betriebswirtschaftliche Nutzen erörtert werden [vgl. Modulhandbuch].

Auf Basis dieser Vorgaben wurde das Thema des Projekts gesucht. Dabei ging es primär darum ein Themengebiet zu finden, welches sowohl breite Möglichkeiten für die informatische als auch die betriebswirtschaftliche Betrachtung bietet. Aufgrund der Interessen innerhalb der Gruppe wurde sich für das Thema **Bundesliga Match Predictions** entschieden. Wir wollten der Fragestellung auf den Grund gehen, ob es tatsächlich möglich ist, diese unzählig erscheinenden Faktoren des Fußballspiels durch Data Science-Prozesse für eine Vorhersage nutzen zu können.

1.2 Business Use Case

Wie bereits erwähnt spielt die wirtschaftliche Betrachtung dieses Projekts neben der informatischen Arbeit eine primäre Rolle. Das Ziel einer Bundesliga Match Prediction liegt hier auf der Hand. Ist es tatsächlich möglich Anbieter wie Tipico und Bwin durch ein mathematisches Modell zu schlagen?

Durch eine solche Vorhersage kann man potenziell starken Profit bei etwaigen Wettanbietern erzielen. Andersrum kann man natürlich auch diese Software an Wettanbieter verkaufen, damit diese ihre Quoten noch effizienter und genauer berechnen können.

Die Herangehensweise an dieses Projekt beginnt mit der richtigen Datenbasis. Durch diese kann man algorithmisch ein Modell erstellen, dass den genannten Business Use Case ermöglichen kann.

2 Theoretische Grundlagen

2.1 Related Work

- quellen/inspirationen"nennen und auf uns beziehen - wie gehen diese ihr problem an (quellen, algorithmen, ...)

2.2 Verwendete Technologien und Bibliotheken

Für die Datenvorbereitung werden klassische Python-Module wie Pandas, Sklearn und Matplotlib genutzt.

Bei der Modellimplementierung wurde vorerst ein Decision Tree Classifier genutzt. Dieser stammt aus dem Sklearn-Modul tree. Nach Ausprobieren dieses Classifiers wurde sich für das Ensemble-Learning entschieden. Dabei nutzt man mehrere Algorithmen um sie zeitgleich zu vergleichen und potenziell zu vereinen. Die Algorithmen wie beispielsweise Logistic Regression oder Gaussian Naive Bayes stammen auch von Sklearn. Final wurde ein Neural Network implementiert. Dazu wurde Keras genutzt. Aus Keras wurde außerdem eine Random Parameter Search zur weiteren Optimierung angewandt. Durch Matplotlib wurden die Ergebnisse visualisiert. Außerdem wurde eine Poisson-Verteilung genutzt um einen mathematischen Ansatz zu implementieren.

3 Praktischer Teil

3.1 Umsetzung

Eine der wohl wichtigsten Entscheidungen bei der Implementation eines Machine Learning Modells sind die ausgewählten Rohdaten. Nach intensiver Recherche wurde sich für einen Datensatz entschieden welcher Informationen über alle Bundesliga-Spiele seit der Saison 2005/2006 enthält. Dieser Rohdatensatz besitzt 65+ Features. Im Preprocessing wurden einige unvollständige Spieldaten gelöscht und viele unnötige Features - wie die Betting Odds der Anbieter. Nach Harmonisierung der Daten wurden die Features auf die Anzahl 21 reduziert. (Hier Grafik von Datensatz). Dadurch, dass die Zielvariablen im Datensatz bereits vorliegen handelt es sich um ein Supervised Learning-Problem. Dieses wurde wie bereits im Theorieteil erwähnt mit drei verschiedenen Methoden angegangen. Dem Ensemble Learning, einem Neural Network und einem mathematischen Ansatz.

3.2 Ergebnisse

Anfangen beim Ensemble Learning kann man verschiedene Accuracys je Algorithmus erkennen (Grafik zeigen). Hierbei sind die Logistic Regression und Linear Discriminant Analysis am Besten. Das Stacking Modell welches die anderen Modelle vereint schneidet auch gut ab. Alle drei liegen bei einer Accuracy von etwas unter 50 Prozent.

Bei Betrachtung des Neural Networks fällt auf, dass die Accuracy schon nach wenigen Epochen auf etwas unter 50 Prozent konvergiert (Grafik). Nach einer Random Parameter Search konnte diese auf 53 Prozent angehoben werden. Zuletzt haben wir mit der Poisson-Verteilung des mathematischen Ansatzes auch eine Accuracy von etwa 50 Prozent erreicht (Grafik).

4 Fazit

Abschließend lässt sich feststellen, dass man durch die Optimierung des Neural Networks die höchste Accuracy erreicht. Nach der Random Parameter Search konnten wir die Accuracy erhöhen. Wie bereits erwähnt lag diese final bei 53 Prozent. Mithilfe dieses Modells wurde das Betting bei Wettanbietern simuliert. Leider kann man sehen, dass unser Modell schnell negative Summen erreicht und einiges an Geld verliert. Der Fußball ist wohl doch ein Gebiet mit zu vielen Zufallsvariablen, sodass man keine Genauigkeit erzielen kann, die einen Wettanbieter schlägt. Die Genauigkeit von 53 Prozent ist jedoch für sich genommen recht vielversprechend. Es zeigt, dass sich der Fußball in Ansätzen algorithmisch bestimmen lässt. Für unseren Business Use Case ist dieses Ergebnis natürlich nicht hoch genug und somit muss dieser verworfen werden. Es hat sich gezeigt, dass man schnell Geld verliert bei der Nutzung unseres Modells.