Wichtigste Ergebnisse Gruppe 2

Maik Kebernik, Jannis Jüngert, Leon Engelhardt:
Soccer Outcome Prediction

Lineare Regression

Bei der linearen Regression versucht man, mit einer oder mehreren unabhängigen Variablen eine abhängige Variable vorherzusagen. Zur Feststellung, wie gut das erhaltene Ergebnis ist, wurden in diesem Fall sowohl R-Squared-Value, als auch Mean Sqaured Error sowie Mean Absolute Error verwendet. Des Weiteren kamen auch grafische Ausgaben zur Auswertung dazu. Gemessen an diesen Werten wurde nun die Zahl an Features zur Trainierung des Modells gewählt, die das beste Ergebnis zur Folge hatte. Parallel dazu wurde auch die Testgröße variiert. Dabei hat sich eine Featureanzahl von 5 als optimal herausgestellt und eine Testgröße von 0,1. Mit Hilfe von Recursive Feature Elimination wurden die fünf relevantesten Features des Datensatzes ermittelt, welche “right_foot_percentage_awayteam”, “average_rating_awayteam”, “average_rating_hometeam”, “B365A” und “B365H” sind. Mit Hilfe dieser Features konnte man das Modell so trainieren, dass R-Squared-Value = 0,17, MSE = 0,59 und MAE = 0,66 sind. Folglich ist die Vorhersage des Spielausgangs eher schlecht möglich.

Random Forest

Der Random Forest Algorithmus berechnet eine Vielzahl von Entscheidungsbäumen, deren Vorhersagen kombiniert werden, um bessere Ergebnisse zu erhalten. Für unser Projekt wurden verschiedene Einstellungen des Random Forest Algorithmus ausprobiert. Als bester Test-Split hat sich der 0,1er Split (10% Testdaten) erwiesen. Als ideale Hyperparameter wurden dazu diese Parameter errechnet: {'max_depth': 6, 'min_samples_leaf': 2, 'min_samples_split': 9, 'n_estimators': 493}. Dadurch ergibt sich eine Genauigkeit von 0.5314720812182742 auf den Testdaten. Als relevanteste Entscheidungskriterien haben sich die Wettquoten für Sieg, Niederlage und Unentschieden erwiesen, gefolgt vom durchschnittlichen Rating und den Ausgängen der letzten drei Spiele der Teams. Den geringsten Einfluss hatte der Anteil der Rechtsfüßer je Mannschaft. F1-Score liegt bei 0.44276195659596923, die Gesamt- AUC liegt bei 0.6697726976759606 und Recall (mit zero_division=1) liegt bei 0.5314720812182742. Damit ist eine Vorhersage des Spielausgangs möglich, jedoch nicht besonders gut. Das Modell ist deutlich besser als Raten, da Sportergebnisse aber von Faktoren abhängen, die sich schwer vorhersagen lassen, ist es weit von Perfektion entfernt.

Multinomiale logistische Regression

Im Vorhinein sollten die Features auf Multikollinearität mithilfe des Variance Inflation Factors überprüft werden. In unserem Beispiel sind hierbei vor allem bei dem durchschnittlichen Gewicht und der Größe starke Abhängigkeiten zu sehen, weshalb der höhere Wert entfernt wurde, um eine zu starke Beeinflussung der Ergebnisse zu verhindern. Beim Trainieren des Modells ist aufgefallen, je höher die Anzahl an Trainingsdaten, desto weniger wurde Unentschieden vorhergesagt. Dies ist ein Indiz für Overfitting, weshalb wir einen Split von 10% Trainingsdaten verwendet haben. Ein Versuch weniger Trainingsdaten zu verwenden führt zu Underfitting, das bedeutet, dass das Modell zwar wieder häufiger Unentschieden vorhersagt, dies aber eher durch Raten zustande kommt, als durch wertbasierte Vorhersagen. Auch wenn ein R2 von -0,5415 und MSE von 1,1050 bei dem ausgewählten Split (10% Trainingsdaten) herauskommt, so zeigen Recall, Precision und F1-Score, dass vor allem Niederlagen (F1-Score: 0,49) und Siege (F1-Score: 0,66) deutlich besser vorhergesagt werden können, als dies bei reinem Raten (33,3% Wahrscheinlichkeit einer korrekten Vorhersage) der Fall ist. Die Schwäche des Modells ist dabei jedoch Unentschieden (F1-Score: 0,02) vorherzusagen.