Developing a Machine Learning model using different algorithms to predict shot success rate on the Kobe Bryant shot selection dataset

Abstract

This report presents a predictive modeling approach for determining the success or failure of shots made by Kobe Bryant during his NBA career using various machine learning algorithms. The main stages in this research include data cleaning and preprocessing, exploratory data analysis (EDA), feature engineering, selecting relevant features, and finally, implementing different algorithms on this binary experiment. The performance of each algorithm is evaluated, and as the final step, a voting ensemble approach is explored to improve model accuracy.

Introduction

During the past decades, we witnessed a surge in applying Data Science techniques in almost every aspect of our lives to embrace the potential for data to provide a competitive advantage. The sports industry is one of the significant economic, entertaining, cultural sectors in our lives that has benefited from this discipline too. Data scientists have used their analytical power to produce predictive insights for optimal/informal data-driven decision making in the sports industry. These insights can provide a range of valuable (maybe previously unknown) information that can benefit everyone in the sport sector, e.g. coaching strategies, player performance and health evaluation, team games analysis, etc. Basketball as a dynamic sport, is one the most popular sports in the world. During recent years, specifically in the National Basketball Association[1] big data analytics techniques have been used by teams to gain a competitive edge in their matches. Players' statistics are synced with on-court game data to provide a thorough data of players' performance. A valuable dataset of Kobe Bryant shots, one of the greatest players in the NBA history, was created and Kaggle[2] hosted a competition on this data to predict whether a shot will find the bottom of the net or not. This competition is a nice practice for us to test a significant part of our knowledge acquired throughout the course Fundamentals of Data Science. This report explores the use of machine learning techniques to build and

[1] https://www.nba.com/stats

[2] https://www.kaggle.com/competitions/kobe-bryant-shot-selection

test predictive models for shot success based on various features related to player actions, shot locations, game context, etc.

Related Work

Previous research has utilized machine learning techniques for sports analytics, including predicting player performance and game outcomes. However, predicting individual shot success involves unique challenges and requires careful consideration of relevant features.

Proposed Method

Our proposed method involves several key steps which are explained as follows:

1. **Data Introduction:** The dataset is loaded, and basic information about its structure is explored.
2. **Cleaning and Preprocessing:** Duplicate values are removed, and string columns are standardized to lowercase. Missing values are handled by dropping rows with null values.
3. **Data Visualization (EDA):** Exploratory data analysis is conducted to visualize the correlation between features which helped us to provide insights into the relationships between different features. Here is a summary of the results:

   We created various plots to analyze shot success rates based on different parameters. Using pairplots we figured how shot distance, latitude, and longitude relate to the success or failure of the shots. We interpreted as distance increases, the shot success rate, decreases. Closer to the end of the match, the failure rate is much more than the success rate, which may be due to the stress of the player. Shots from the center zone of the court have a higher success rate whereas the least rate belongs to the back zone shots. Dunk and Layup shots were Bryant's most successful shots.

4. **Feature Engineering:** In this section, we created new features, modified existing ones, and dropped features to improve the performance of our machine learning model. The results of the previous stage (EDA) helped us to drop columns with irrelevant and uncorrelated data. We also created new features from existing data based on our insights before (e.g. created "total_time_remaining" since we figured out that there is a high correlation between match time and the failure/success rate). Lat but not least, we converted ordinal and categorical

variables into nominal vectors using ordinal and one-hot encoding, respectively. At the end of this process, we came up with 210 features.

5. **Feature Importance:** In the context of machine learning and data analysis, feature selection is a technique in feature engineering which means to choose the most relevant/correlated features to reduce model's dimensionality as well as computational complexity. For the sake of better organization of our report, we included this technique as a separate section and implemented three feature selection techniques Chi2, Recursive Feature Elimination, and Variance Threshold to identify the most relevant features for predicting shot success. We selected the top 20 important features from chi-squared test and recursive feature elimination and chose features with variance threshold greater than 0.9. Finally, we combined the unique features from these three techniques which led us to come up with 48 features.

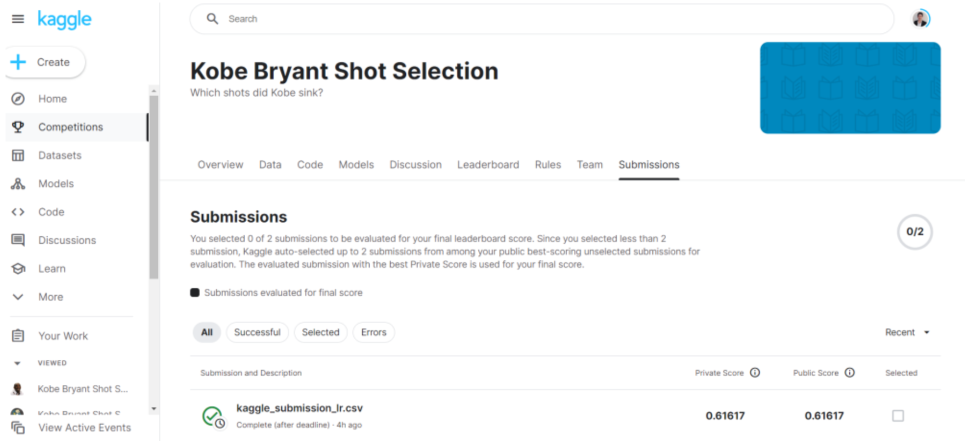
   We also realized that the failure class significantly outnumbers the success class (imbalanced data) which may prone the model to be biased towards the former class. So, we employed a popular technique for handling imbalanced data, Synthetic Minority Over-sampling Technique (SMOTE) which works by creating synthetic examples in the feature space of the minority class, thereby balancing the class distribution.

6. **Algorithms Developed:** We chose 4 different machine learning algorithms namely, Logistic Regression, Random Forest Classifier, Gradient Boosting Classifier, and Neural Networks for our classification task. Each algorithm is evaluated using standard performance metrics (Accuracy, Precision, Recall, and F1 Score). Since we may exceed the proposed report length, we do not explain the modifications and customizations suggested to each of these models here. Please refer to our submitted notebook (provided with comments) for further information. Finally, we used Voting Ensemble technique to test whether we can improve the overall accuracy of our model by using the 'collective decision of several models' developed (Gradient Boosting Classifier, Random Forest Classifier, and Ada Boost Classifier).

Benchmark, Conclusions and Future Work

The benchmark is set based on the accuracy of predictive models in determining shot success. Also, the experimental results include performance metrics (accuracy, precision, recall, F1 score) for each implemented algorithm. In the notebook, confusion matrices are provided to visualize the classification results.

In our project, we developed four different models—Random Forest, Logistic Regression, Neural Network, and XGBoost—to predict outcomes based on our dataset. Surprisingly, the simplest model, Logistic Regression, achieved the highest accuracy, reaching 70% in our test set. This outcome suggests that for our relatively small and low-dimensional dataset, where the target has only two classes, opting for a simpler model may be more effective. In our contribution to the Kaggle competition, we achieved a loss of 0.61 in the test set.



Kobe Bryant Shot Selection

Which shots did Kobe sink?

Submissions

You selected 0 of 2 submissions to be evaluated for your final leaderboard score. Since you selected less than 2 submission, Kaggle auto-selected up to 2 submissions from among your public best-scoring unselected submissions for evaluation. The evaluated submission with the best Private Score is used for your final score.

■ Submissions evaluated for final score

Submission and Description	Private Score	Public Score	Selected
kaggle_submission_lr.csv Complete (after deadline) · 4h ago	0.61617	0.61617	□

Future work could focus on specific instances where our models failed, aiming to refine and enhance overall performance. Also, further analysis of visualizations can be made to get more insights into the data.

Team Assignments

We actively did each section with the consultation and help of each other. If we need to be more specific, Arash Bakhshaee did the data cleaning and preprocessing, Ehsan Mokhtari did the EDA, Omid Ghorbani did the model development (except NN) and evaluation, and Sohrab Seyyedi Parsa did the NN model development and wrote the report.