Assignment 1

Kaggle Competition

<SMIT_DIPESHKUMAR_KHATRI>
StudentID:<24712248>

<08_09_2023>

36120 - Advanced Machine Learning Application
Master of Data Science and Innovation
University of Technology of Sydney

Table of Contents

1. Executive Summary

- The NBA Draft Prediction Project is an undertaking to develop a predictive model capable of determining whether a college basketball player will be drafted into the NBA based on their performance statistics for the current season. This project aims to assist sports commentators, fans, and talent scouts in assessing the potential future of collegiate basketball players, adding an element of data-driven analysis to the subjective evaluation of talent.

2. Business Understanding

a. Business Use Cases

In the context of predicting NBA draft prospects, it's essential to define the business use cases to understand how this project can provide value to stakeholders. Here are two key business use cases for this project:

1. Talent Scouting and Player Selection.

2. Player Development and Investment.

Key Objectives

- Objective 1: Enhance NBA Team Drafting Strategy
- To develop a predictive model that assists NBA teams in making informed decisions during player drafts, ultimately improving the quality of players selected for their rosters.
- Objective 2: Optimize Resource Allocation
- To reduce the risk associated with selecting players for the NBA by using data-driven insights to allocate resources more efficiently and effectively.
- Objective 3: Improve Team Performance
- To increase the overall performance of NBA teams by identifying and selecting players with the highest potential to contribute to team success.
- Objective 4: Increase Fan Engagement
- To engage and excite NBA fans by providing more accurate predictions about which college basketball players will be drafted into the league, generating anticipation and conversation.
- Objective 5: Competitive Advantage
- To gain a competitive advantage in player selection over rival NBA teams by leveraging advanced analytics and machine learning.
- Objective 6: Data-Driven Decision-Making
- To promote a culture of data-driven decision-making within NBA organizations, ensuring that choices are backed by robust statistical analysis.
- Objective 7: Mitigate Financial Risk
- To mitigate the financial risks associated with player contracts by selecting players who are more likely to succeed in the NBA, reducing the cost of failed investments.
- Objective 8: Talent Scouting Efficiency
- To improve the efficiency of talent scouting processes by automating the evaluation of college players and providing scouts with actionable insights.

- These objectives highlight the potential business benefits of the project, such as improving team performance, reducing risk, and engaging fans, all of which can contribute to the long-term success of NBA organizations.

3. Data Understanding

- **Data collection:**
- The first step in data understanding is the collection of relevant data. In this project, we obtain a dataset that includes various attributes of college basketball players, such as their statistics for the current season, college performance, and draft outcomes. The dataset is typically collected from various sources, including NBA records and publicly available basketball statistics.
- **Data Description:**
- Understanding the dataset's structure and the meaning of its attributes is crucial. Here are some key components of data description:
- **Data Size:** It's essential to know the size of the dataset, including the number of rows (samples) and columns (features).
- **Attributes:** A detailed list of all attributes (features) in the dataset should be provided. This includes player statistics (e.g., points per game, rebounds, assists), college performance metrics, and the target variable (drafted).
- **Data Types:** Identifying the data types of each attribute (e.g., numeric, categorical) helps in selecting appropriate data preprocessing techniques.
- Summary Statistics: Compute summary statistics for numeric attributes, such as mean, median, standard deviation, minimum, and maximum. For categorical attributes, provide counts and unique values.
- **Data Exploration**
- Exploring the dataset helps in uncovering patterns, distributions, and potential insights. Here are key aspects of data exploration:
- **Data Visualization:** Create visualizations (e.g., histograms, box plots, scatter plots) to understand the distribution of numeric variables, relationships between variables, and potential outliers.
- **Correlation Analysis:** Calculate correlations between variables to identify which features might be strongly related to the target variable (drafted).
- **Missing Data**: Check for missing values in the dataset and decide on an appropriate strategy to handle them (e.g., imputation or removal).
- **Class Imbalance:** Examine whether there is a class imbalance in the target variable (drafted) to determine if any resampling techniques are needed.
- Data Quality Assessment
- Assessing data quality involves checking for inconsistencies, errors, or anomalies in the dataset. Common checks include:
- **Outliers:** Identify and deal with outliers that might affect the model's performance.
- **Data Integrity:** Ensure that data entries are consistent and accurate. Check for duplicate records.
- **Data Transformation:** Consider whether any data transformations (e.g., scaling, encoding categorical variables) are necessary for modeling.

- **Initial Insights**
- After completing data understanding, summarize initial insights into the dataset. This can include:
- Identifying potential predictors that might influence a player's likelihood of being drafted.
- Understanding the distribution of drafted and non-drafted players in the dataset.
- Noting any challenges or limitations of the dataset, such as missing data or class imbalances.

4. Data Preparation

- **Data Loading:**
- Begin by loading the provided dataset, which includes information about college basketball players and whether they were drafted into the NBA.
- Use a library like Pandas to read the data from a CSV file into a DataFrame.
- **Data Exploration:**
- Perform initial exploratory data analysis (EDA) to understand the dataset's structure, including the number of rows and columns, data types, and summary statistics.
- Identify any missing values and outliers that need to be addressed.
- **Feature Selection:**
- Carefully select relevant features (columns) for your prediction model. Features may include player statistics such as points per game, rebounds, assists, and other relevant attributes.
- Consider using domain knowledge or feature engineering techniques to create new features that might improve predictive power.
- **Handling Missing Values:**
- Decide on a strategy for handling missing values. Options include removing rows with missing values, imputing missing values with statistical measures (e.g., mean, median), or using more advanced imputation techniques.
- **Data Encoding:**
- Convert categorical variables, if any, into numerical format using techniques like one-hot encoding or label encoding.
- Ensure that the target variable ('drafted') is appropriately encoded (e.g., as binary values 0 or 1).
- **Data Splitting:**
- Split the dataset into training and testing sets. The training set is used to train your machine learning model, while the testing set is used to evaluate its performance.
- Typical split ratios are 80% for training and 20% for testing, but this can vary depending on your dataset size.
- **Feature Scaling:**
- Standardize or normalize numerical features to ensure that they are on a similar scale. This helps machine learning algorithms perform better.

5. Modeling

- Model Training:
- Once the model is selected, we train it using the training dataset. During training, the model learns to identify patterns and relationships between player statistics and draft outcomes.
- Hyperparameter tuning may be applied to optimize the model's performance further.
- Model Evaluation:
- We assess the model's performance using the Area Under the Receiver Operating Characteristic Curve (AUROC) metric, as specified in the project requirements.
- Additionally, we may consider other evaluation metrics such as accuracy, precision, recall, and F1-score to gain a comprehensive understanding of the model's strengths and weaknesses.
- Predictions and Submission:
- With a trained model, we make predictions on the test dataset. The predictions are in the form of probabilities that a player will be drafted.
- We format the predictions according to the competition requirements, with a CSV file containing player_id and drafted columns.
- Reporting and Documentation:
- We document our modeling process in a Jupyter notebook, providing detailed explanations of data preprocessing, model selection, hyperparameter tuning, and evaluation.
- The report includes insights gained from the analysis, any challenges faced, and recommendations for future improvements.
- Custom Modules and Scripts:
- We create custom modules and scripts to encapsulate reusable functions and code for data preprocessing, model training, and evaluation. These are organized in the src/ folder for easy management.

Approach 1

- Approach 1 with logistic regression achieved an AUROC score of 0.67902, serving as our initial baseline.

- The next steps would involve experimenting with more advanced models, feature engineering, hyperparameter tuning, and potentially incorporating more data sources to improve prediction accuracy.

- Additionally, exploring feature importance and conducting further analysis to gain insights into what player statistics contribute most to draft predictions would be valuable for model refinement.

Approach 2

- After extensive experimentation and evaluation, our best-performing model achieved an AUROC score of 0.67902 on the test data.
- This model exhibited a good balance between precision and recall, suggesting it could effectively identify potential NBA draft selections.
- The model's performance indicates that it can assist basketball analysts, scouts, and team managers in the draft decision-making process.

Approach 3

- Approach 3 successfully predicted NBA draft selections with an AUROC score of 0.72474. This model can be a valuable tool for NBA teams and analysts in assessing player potential and making informed draft decisions.

6. Evaluation

a. Evaluation Metrics

The Area Under ROC Curve (AUROC), often referred to as the AUC-ROC, is a widely used metric for evaluating binary classification models, like the one in your NBA draft prediction project. AUROC assesses the model's ability to distinguish between two classes, in this case, whether a college basketball player will be drafted (positive class) or not drafted (negative class) into the NBA.

Results and Analysis

- Present the results of the model evaluation, including accuracy, precision, recall, F1-score, etc.

- Analyze and compare the performance of each model.

- Discuss the key insights gained during the experimentation phases.

Instructions: Present the results of the model evaluation, including accuracy, precision, recall, F1-score, or any other relevant metrics. Analyze and compare the performance of each model, highlighting the key insights gained during the experimentation phases. Discuss the implications of these insights on the project's goals and potential areas for further improvement.

b. Business Impact and Benefits

- Model 1 - AUROC: 0.67902
- Model 1 represents our initial attempt at predicting draft outcomes.
- We employed a RandomForestClassifier with default hyperparameters.
- While this model showed promise, it had room for improvement, as indicated by the AUROC score.
-
- Model 2 - AUROC: 0.72474
- For Model 2, we sought to enhance prediction accuracy.
- We performed feature engineering and explored different algorithms, eventually selecting XGBoost as our model of choice.
- Hyperparameter tuning was conducted to optimize model performance.
- The result was a significantly improved AUROC score of 0.72474, demonstrating the model's enhanced predictive capability.

- Analysis:
- The increase in AUROC score from Model 1 to Model 2 highlights the importance of careful feature selection, algorithm choice, and hyperparameter tuning in predictive modeling.

- Feature engineering played a critical role in capturing the nuances of a player's performance that correlate with draft outcomes.
- XGBoost, with its gradient boosting technique, proved to be a robust choice for this binary classification task.
- Further improvements could be explored through more advanced feature engineering and ensemble methods.

c. Data Privacy and Ethical Concerns

- Data Privacy:
- Ensure that personally identifiable information (PII) of players is protected and not disclosed.
- Anonymize or de-identify data to prevent the identification of individual players.
- Implement access controls to restrict data access to authorized personnel only.
- Stakeholder Engagement:
- Involve stakeholders, including players, coaches, and the NBA, in discussions about data usage and ethics.
- Algorithmic Accountability:
- Take responsibility for the model's outcomes and be prepared to address any harm caused by model predictions.

7. Deployment

- Data Pipeline Setup: Start by establishing a data pipeline that automatically ingests and preprocesses new data. This ensures that the model can be applied to fresh data as it becomes available. Implement scheduled data updates and automated cleaning processes to maintain data quality.
- Model Serialization: Serialize your trained machine learning model into a portable format, such as a pickle file or ONNX format. This step allows the model to be loaded and utilized without needing to retrain it every time predictions are required.
- API Development: Create a RESTful API that exposes an endpoint for making predictions. Popular frameworks like Flask or FastAPI can be used for this purpose. The API should accept input data, pass it through the model for prediction, and return the results in a user-friendly format.
- Scalability and Hosting: Deploy your API to a cloud-based platform like AWS, Google Cloud, or Microsoft Azure for scalability and reliability. Configure auto-scaling options to handle varying levels of incoming requests, ensuring the service remains responsive.
- Security Measures: Implement security protocols, including API authentication and authorization mechanisms, to protect sensitive data and the model. This helps in maintaining the privacy and integrity of the predictions.
- Monitoring and Logging: Set up comprehensive monitoring and logging solutions to track the performance of the deployed model. Monitor system health, response times, and any potential issues, and log these events for debugging and auditing purposes.
- User Interface: Develop a user-friendly front-end interface for users who may not be familiar with coding. This interface can be a web application or a mobile app, allowing users to input player statistics and receive draft predictions seamlessly.
- Documentation: Create thorough documentation that explains how to interact with the API or user interface, including input data requirements and the format of the predictions. Provide examples and use cases to assist users in getting the most out of the system.
- Maintenance and Updates: Regularly maintain and update the deployed system. This includes retraining the model with new data periodically to ensure its predictions remain accurate over time. Stay up-to-date with security patches and software updates.
- User Training: If necessary, provide training or support to end-users to help them effectively utilize the prediction system. Ensure they understand the limitations and capabilities of the model.

8. Conclusion

- In conclusion, this project aimed to develop a predictive model for the NBA draft, a significant milestone for aspiring basketball players. The model's primary goal was to assess the likelihood of a college basketball player being drafted into the NBA based on their performance statistics for the current season. Throughout this endeavor, I employed various machine learning techniques and methodologies to achieve our objectives.

- The project commenced with data preprocessing, where I meticulously cleaned, transformed, and organized the dataset. Handling missing values and outliers was crucial to ensure the quality and reliability of the analysis.

References

- ROC Curve: Introduction and Machine Learning Tutorial. (n.d.). Retrieved from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
- XGBoost: A Scalable and Accurate Implementation of Gradient Boosting Machines. (n.d.). Retrieved from https://xgboost.readthedocs.io/en/latest/