Sports Analytics - UNCC Football

DSBA-6400 Internship Fall 2023

DSBA Akhilesh Pothuri

Mentor: Professor John Tobias

Sports Analytics Program Director

Table of Contents

Executive Summary

During the internship with the Sports Department of the UNC Charlotte football team, a good number of contributions were made to enhance the team's performance through data analysis and machine learning applications.

The primary focus of the work involved analyzing game data from the current season to extract valuable insights for informing strategic decisions. Statistical tools and techniques were leveraged to identify patterns, trends, and key performance indicators that played a crucial role in shaping the team's game plans. This data-driven approach provided a comprehensive understanding of the team's strengths and areas for improvement, contributing to more effective in-game strategies.

In addition to game data analysis, a project was undertaken to streamline and optimize the team's practice sessions. Recognizing the time-intensive nature of manual jersey number tracking during practices, a Convolutional Neural Network (CNN) model was developed for automatic jersey number detection. This innovative solution significantly reduced the time and effort required for tracking players and enhanced accuracy, ensuring a more reliable dataset for performance evaluation.

The successful implementation of the CNN model showcased the potential of integrating cutting-edge technology into sports management, emphasizing efficiency and precision. The model's accuracy and efficiency contribute not only to practice optimization but also lay the groundwork for potential applications in live game scenarios.

This internship provided valuable hands-on experience in applying data analysis and machine learning techniques to real-world sports scenarios. The outcomes of the work have the potential to revolutionize how the UNC Charlotte football team approaches game strategy and player performance evaluation. The intersection of sports and technology represents a promising avenue for continued exploration and innovation, with the insights gained from this internship expected to have a lasting impact on the team's success.

Introduction

Sports have evolved into a domain where data analysis and technology play a pivotal role in enhancing team performance and strategic decision-making. My internship with the UNC Charlotte football team aimed to leverage data science methodologies to gain insights into player performance, provide real-time statistics during games, and contribute to the optimization of coaching strategies.

The overarching objective of this project was to delve into the intricacies of player performance during football games, with a specific focus on the UNC Charlotte football team. By studying various aspects of the game, such as formations, field positions, and player actions, our aim was to provide valuable live statistics that could empower coaches to make informed decisions and build a stronger, more competitive team.

Key Project Objectives

1. **Predict 'Gain' for Defensive Strategies**: One of the primary objectives was to develop predictive models that could anticipate the defensive performance of the UNC Charlotte football team based on factors like formations, field positions, and other relevant game dynamics. This predictive insight aimed to assist coaches in devising effective defensive strategies, optimizing player positioning, and minimizing the opponent's offensive gains.

2. **Run/Pass Play Prediction**: Understanding the dynamics of offensive plays is crucial for coaches when formulating defensive strategies. The team sought to predict whether the opposing team would execute a run or pass play, providing coaches with valuable foresight to optimize defensive tactics accordingly. This predictive element aimed to enhance the team's adaptability during games and contribute to strategic decision-making on the field.

3. **CNN Model for Automated Live Tracking**: Recognizing the manual efforts involved in tracking player movements during practice sessions, the team embarked on the development of a Convolutional Neural Network (CNN) model. This technology aimed to automate the live tracking process, offering a more efficient and accurate alternative to manual data collection. By seamlessly extracting essential game data from live footage, the CNN model aimed to contribute to practice session optimization and provide coaches with reliable, real-time insights.

Statement of Business Problem

In the sports internship, challenges were encountered to enhance the team's performance. Here's the breakdown:

1. **Predict 'Gain' for Defensive Strategies**:
   - **The Problem**: Understanding the strength of the defense in advance to make quick on-field decisions.
   - **Why it Matters**: Predicting defensive performance helps coaches plan effective strategies to prevent the opposing team from scoring too many points.

2. **Run/Pass Play Prediction**:
   - **The Problem**: Determining whether the opposing team will run or pass the ball, akin to predicting moves in a game of chess.

- **Why it Matters**: Knowing the opponent's play allows coaches to position our players strategically, providing an advantage on the field.

3. **CNN Model for Automated Live Tracking**:
- **The Problem**: Tracking players during practice manually is challenging, requiring a smart tool for automated tracking.
- **Why it Matters**: Coaches want to observe players' movements and decisions during practice efficiently. An automated tracking tool saves time and provides accurate information.

The Big Picture:

- **Time and Smart**: In sports, time is crucial, and quick decisions are essential. The problems addressed aimed at saving time and providing coaches with intelligent information.
- **Making the Team Smarter**: Solving these challenges aimed to elevate the UNC Charlotte football team's performance, enabling them to make better moves and achieve more victories on the field.

Data Collection Process

To achieve these objectives, live football matches were actively attended, involving meticulous data collection and the calculation of essential metrics like Down and Distance. Utilizing video footage allowed for the capture of dynamic game elements, ensuring a comprehensive dataset for analysis. This hands-on approach facilitated the extraction of accurate and relevant data, immersing in the real-world context of football gameplay and enhancing the validity of findings.

Delving into the details of this internship, the subsequent sections will elaborate on the methodologies employed, challenges encountered, and the outcomes achieved in pursuit of these ambitious objectives. The fusion of sports and data science represents a burgeoning field with immense potential, and this internship seeks to contribute meaningfully to the intersection of technology and athletics.

Methods

In pursuit of the sports internship objectives, a systematic approach was employed, tailoring methodologies to each project facet. The following delineates the methods harnessed to realize the outlined objectives, accompanied by insights into potential barriers encountered during the process.

'Gain' for Defensive Strategies

In addressing the initial objective, the challenge was handling a limited dataset specific to the current season's game information. The dataset encompassed various details for each play, such as QB Comment, Down, Distance, Field Position, Gain, Off Group, Backfield, Formation, Play, Tempo, Run Concept, Pass Result, and Pass Concept.

To ensure data completeness, live games were attended, contributing additional information to enrich the dataset. Integration of data from diverse sources was facilitated through the play number of each game. Handling missing values involved imputing 'No Comment' in the QB

Comment column and dropping records with missing 'Gain' values. For features like 'Down,' 'Distance,' and 'Field Position,' missing values were imputed with mean values.

In the realm of feature engineering, checks for multicollinearity were conducted. Features like 'Play Number' and 'Series' were excluded to prevent their influence on other independent features, a decision grounded in Variance Inflation Factor (VIF) analysis. Categorical features like ['Backfield', 'Formation'] underwent transformation into numerical values using One-Hot encoding. This technique creates binary columns for each category, ensuring model compatibility and interpretability.

For categorical features like 'Backfield' and 'Formation,' One-Hot encoding is applied to convert them into a numerical format suitable for machine learning models. This involves creating binary columns for each category, assigning a 1 if the observation belongs to that category and 0 otherwise, facilitating effective interpretation and utilization of categorical information during the prediction process.

In forecasting defensive performance, four distinct regression models were employed: Random Forests, Support Vector Regression (SVR), Linear Regression, and Gradient Boosting Regressor. To optimize model configuration, GridSearchCV, a robust hyperparameter tuning method, systematically explored various combinations of parameters for each model to identify settings yielding optimal performance.

Model comparison relied on Mean Squared Error (MSE) and R-squared (R2) values, offering a quantitative assessment of how well each model captured variance in the data and minimized prediction errors. MSE quantifies the average squared difference between predicted and actual values, while R2 reflects the proportion of variance in the dependent variable explained by the model.

The choice of these regression models was deliberate, aligned with their suitability for the predictive task. Random Forests excel in handling complex relationships, Support Vector Regression captures intricate patterns in data, Linear Regression provides simplicity and interpretability, and Gradient Boosting Regressor sequentially improves model performance through ensemble learning, making it well-suited for nuanced tasks.

Post-training and fine-tuning, the evaluation process extended to the test set, rigorously comparing predictions generated by each model with the actual values. This comprehensive assessment validated the predictive capabilities of the models, ensuring reliable defensive performance predictions aligned with the distinct dynamics of each game in the current season.

Run/Pass Play Prediction

Similar data filtering and cleaning processes were applied to maintain relevance and data integrity. Feature selection involved identifying critical attributes, such as 'Name,' 'Play Number,' 'Down,' 'Distance,' 'Field Position,' 'Formation,' and 'Backfield.' Employing One Hot and Ordinal encoding ensured numerical compatibility for classification models, including Random Forests, SVC, and Gradient Boosting Classifier. Guided by the intrinsic capabilities of each classifier and the nuanced demands of our classification task, model selection aimed for a complementary ensemble. Random Forests addressed complexity, SVC handled subtleties, and the Gradient Boosting Classifier unraveled nuanced play distinctions through iterative learning.

The evaluation process zoomed in on accuracy, precision, recall, and F1-score metrics, ensuring a comprehensive understanding of model performance on the test set. The selection process guaranteed that our Run/Pass play predictions not only upheld accuracy but also reflected the intricate dynamics of each play, adding a layer of sophistication to our analytical endeavors.

Jersey Number Detection

In the pursuit of precise jersey number detection, reliance on the coordinates provided by helmets in the original dataset proved crucial. Instead of starting from scratch, these coordinates served as a dependable reference point for guiding our jersey number detection, eliminating the need for a new dataset.

Our strategy embraced a computer vision approach aligned with the game's dynamics, aiming for the model to accurately recognize jersey numbers using helmet coordinates as reliable markers. The focus was on maximizing the utility of existing information rather than initiating the process anew.

To tailor our model for real-world scenarios, we selected frames from video footage, prioritizing those captured from the endzone to ensure clear and readable jersey numbers.

Augmenting the images was a vital step in our methodology, introducing diversity to the training data through techniques like random rotation, zoom, and contrast adjustments. This diversity enhances the model's adaptability, preparing it to recognize jersey numbers under various conditions and player movements during live tracking.

AutoTune played a pivotal role in dynamically optimizing our processing approach, automatically adjusting parameters to enhance model training efficiency and make optimal use of computational resources.

Upon completing the model training, we tested it by extracting video frames from the test data. The helmet coordinates acted as a starting point, guiding the model to accurately detect jersey numbers in each frame. The detected jersey numbers seamlessly integrated back into new video frames, ensuring that the model's predictions translated effectively into practical and visually coherent outputs.

Throughout this process, challenges associated with the dynamic nature of football games prompted iterative fine-tuning of our model. Balancing accuracy and adaptability, we utilized the EfficientNet model as the backbone of our jersey number detection efforts, further enhancing the precision and efficiency of our approach. In essence, by leveraging existing helmet coordinates and incorporating the EfficientNet model, we optimized our strategy for effective jersey number detection in the dynamic context of live football games.

Results

'Gain' for Defensive Strategies

In evaluating the effectiveness of our gain prediction models for defensive strategies, we depend on fundamental metrics: Mean Squared Error (MSE) and Mean Absolute Error (MAE).

These metrics offer valuable insights into the consistency of our predictions in comparison to the actual values, providing a quantifiable measure of the disparity between the two.

The following tables showcase the respective scores for each model employed in our gain prediction. It's essential to note that lower MSE and MAE values signify a closer alignment between our predictions and the actual outcomes. Notably, Random Forest outperformed all other models in predicting gain.

Model	MAE	MSE
Random Forest	5.899	85.708
Gradient Boosting	5.986	87.087
SVR	5.965	88.11
Linear Regression	6.395	86.594

Table 1: Gain Model Accuracies

Moreover, GridSearchCV was utilized as a potent tool in determining the most effective model parameters. Its systematic exploration of diverse combinations within the provided set of hyperparameter options ensures the fine-tuning of our models, optimizing them for peak performance.

In terms of feature importance, our visual representation(fig 1) unveils valuable insights. The field position feature emerges as the most influential, boasting a significance of 36%. Following closely are play number, down, and distance features. These insights illuminate the aspects that carry the most weight in our predictive models, guiding our understanding of the key factors impacting defensive gain predictions.

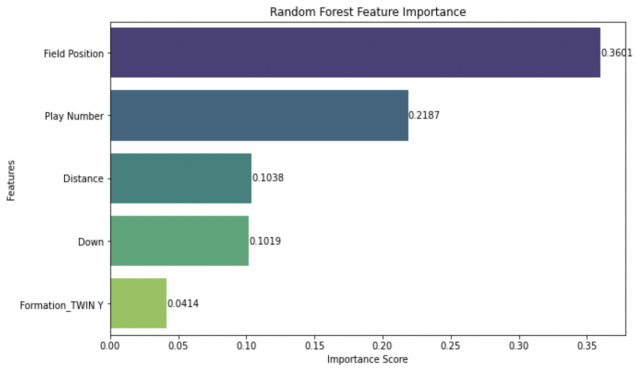

Fig 1: Regression Model Feature Importance

The second visualization(fig 2) delves into the average gain concerning down and distance. By mapping down on the y-axis and distance on the x-axis, this visualization offers a comprehensive view of how these two variables interact and contribute to the average gain. These visual tools not only enhance our comprehension of the model's performance but also provide actionable insights for refining defensive strategies based on specific game situations.

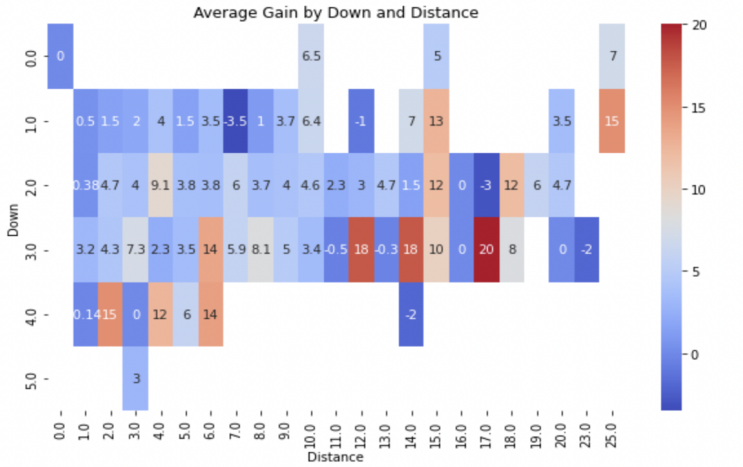

Fig 2: Average Gain by Down and Distance

Run/Pass Play Prediction

When it comes to predicting whether a play involves a run or pass, our models underwent rigorous evaluation, with Random Forest emerging as the standout performer, achieving an impressive accuracy of 70.94%. The table below provides a clear snapshot of the accuracies and confusion matrices for each model.

Model	Accuracy	Confusion Matrix
Random Forest	70.94%	[[25,26],[17,80]]
Gradient Boosting	64.86%	[[24,27],[25,72]]
Support Vector Classifier	68.24%	[[28,23],[24,73]]

Table 2: Classifier Models Results

Comprehending these accuracy values is pivotal as they represent the percentage of accurately predicted outcomes by each model. To gain deeper insights into the factors influencing these predictions, our focus shifts to the feature importance visualization (Fig 3). This graphical representation elucidates the crucial variables that significantly contribute to the decision-making process of the models. In simpler terms, it aids in identifying the aspects that play a pivotal role in determining whether a play is more likely to be a run or a pass.

In the context of our Run/Pass Play Prediction, clarity on feature importance not only enhances our understanding of the models' inner workings but also provides actionable insights for coaches and strategists aiming to optimize playcalling based on specific game scenarios.

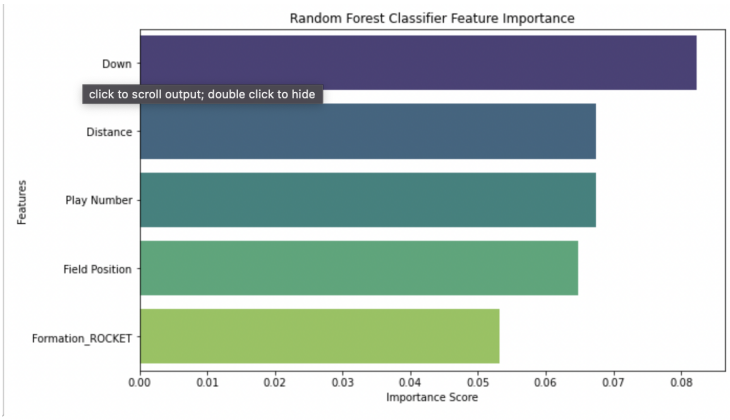

Fig 3: Classifier Feature Importance

Jersey Number Detection

The jersey number detection model underwent thorough training and validation, resulting in a training accuracy of 89.44% and a validation accuracy of 65.14%. These accuracy values provide an indication of how well the model performed in recognizing jersey numbers.

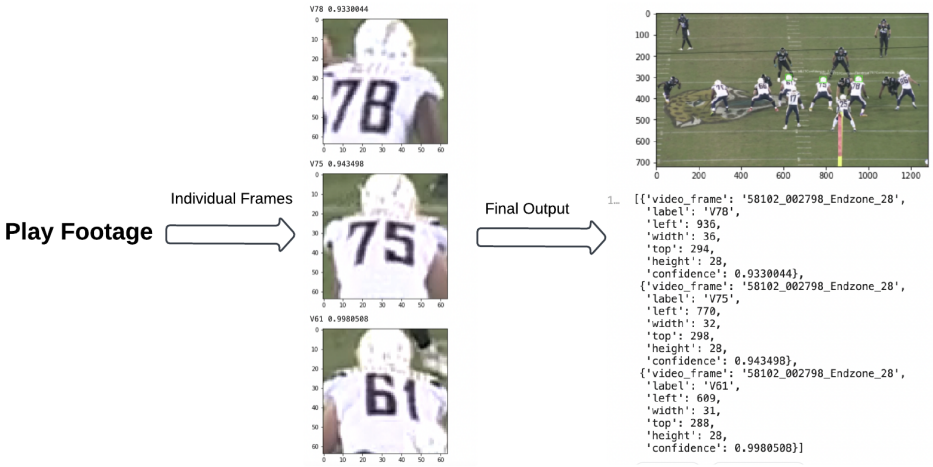

Fig 4: Jersey Number Detection

While a high training accuracy is indicative of the model's proficiency on familiar data, the validation accuracy is crucial to assess its ability to make accurate predictions in real-world scenarios. In cases where the training accuracy significantly surpasses the validation accuracy, there might be a concern of overfitting, where the model may have learned to perform well on the training data but struggles with new, diverse inputs. Achieving a balance between training and validation accuracy is essential for a robust model capable of accurately recognizing jersey numbers across various conditions and player movements during live tracking.

Discussions and Conclusions

Overall Business Impact

The sports internship experience has introduced valuable insights that can significantly impact the overall strategy of the UNC Charlotte football team. By successfully predicting defensive gains, identifying run/pass plays, and implementing efficient jersey number detection, our findings offer tangible benefits. Coaches can leverage these predictions to refine game strategies, enhance player performance, and streamline tracking processes during practice sessions. The practical application of data-driven insights in sports operations aligns with the broader industry trend, fostering a more informed and strategic approach.

New Avenues of Inquiry

The sports internship experience has sparked curiosity in exploring additional dimensions. One avenue of inquiry involves delving deeper into player-specific analyses, assessing individual strengths and weaknesses to tailor training programs. Exploring real-time applications of machine learning during live games, such as instant play recommendations based on predictive models, presents another intriguing path. Additionally, investigating the integration of player health data into the analysis could provide comprehensive insights for injury prevention and overall well-being.

Appendix

This internship experience has been a genuine roller coaster ride. Initially, we lacked clear guidance on our roles as interns and the specific tasks each of us would be undertaking. However, the journey became more enjoyable as we delved into learning about American Football, a sport entirely new to me. The team's support was instrumental in helping us grasp the intricacies of the game and its techniques.

The experience took an interesting turn when I started working on the CNN model for jersey number detection. This project held significant importance for the live tracking process, adding an element of challenge as it was my first time working on a real-time CNN project. Despite the initial uncertainties, the team's encouragement and support played a crucial role in overcoming challenges and gaining valuable hands-on experience.

The DSBA curriculum proved to be an asset during this internship, offering alternative data cleaning techniques and emphasizing the importance of storytelling in data science. The program's courses, particularly in Supply Chain and Model Risk Management, provided a unique perspective on business aspects of data science tasks. These courses expanded my understanding and encouraged me to explore different angles when approaching challenges.

Overall, this internship served as a fantastic opportunity to apply the knowledge gained in the DSBA program in real-world scenarios. It was a journey filled with learning, growth, and the application of diverse skills in various capacities.