Predicting Injury Risk for NBA Draft Prospects Using Deep Learning and NLP

Christopher Song
Department of Computer Science
University of Maryland
College Park, MD 20742
cs6643@umd.edu

Viraj Boreda
Department of Computer Science
University of Maryland
College Park, MD 20742
vboreda@terpmail.umd.edu

Abstract

In this paper, we propose a novel approach to predicting career injury risk for NBA draft prospects utilizing deep learning techniques, specifically FT-Transformers and the BERT language model. Our methodology uses FT-Transformers for tabular data from the NBA Draft Combine measurements and BERT for text data from player scouting reports to find a holistic representation of how injury prone a player will be. By constructing a custom injury risk score that factors in percentage of maximum possible career games missed, average length of each injury, number of injuries per game, and minutes played per game, we aim to develop a model that aids both amateur draft analysts and professional scouts in assessing draft prospect health and longevity once they enter the league.

1 Introduction

The NBA draft presents a pivotal moment for franchises every year, requiring them to evaluate prospects not only based on skill and potential but also on injury risk. There is no easy way to predict this as some players happen to be inherently injury prone but this may not be revealed in their college/pre-NBA games. Current mainstream approaches to determining injury risk mostly involve past injury data and aren't given much thought (i.e. a player who has been injured a lot in college may be seen as injury prone without any objective risk value), so we believe there is a lot of room to improve the effectiveness of draft scouting. Recent advancements in deep learning techniques and natural language processing models have opened up avenues for building an advanced injury likelihood prediction model. Our approach aims to integrate these methodologies by employing FT-Transformers for tabular data analysis of NBA draft combine measurements and the BERT language model for processing player scouting reports to form a holistic injury risk score based on current NBA player injuries.

1.1 Other Approaches

There have been a few past attempts to predict injuries, however none of these attempts have been very thorough for draft prospects. Sports Illustrated has an example of how some writers and draft scouts analyze draft prospect injuries in a subjective way, making guesses about injury likelihood in the NBA based on past injuries the prospects faced. This approach has been the most common method of assessing injury risk, but it hasn't yielded great results as many highly drafted NBA players tend to have injury-laden careers and end up washing out of the league.

Farghaly and Deshpande tried a different approach, using machine learning techniques like random forests and K-nearest neighbors to determine whether NBA players are similar to other players based on injury likelihood, using measurements similar to the combine measurements our project gathered.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

While they achieved some success, their paper was used for players already in the NBA to help teams prevent injuries among their players, instead of being used as a predictive tool for the NBA Draft.

Therefore, our project explored new ground both by using novel deep learning techniques for tabular data and scouting reports, as well as using these techniques on data for players before they entered the NBA. We believe this approach will make our model more useful for NBA scouts and draft analysts, not just medical teams.

2 Methodology

2.1 Data Collection

We gathered the following NBA Draft Combine measurements for all players drafted between 2013 and 2022: position, height, weight, body mass index, body fat percentage, wingspan, standing reach, hand length & width, wingspan/height ratio, and hand size from Marcus Fern's dataset on Kaggle.

For the scouting reports, we scraped the NBA Draft Room website's mock drafts from 2013 to 2022, as they have writeups for almost every player drafted in that timeframe. Although not every player in our final dataset ended up with a scouting report, the model was built to account for this and the majority of players did have at least a small writeup. Additional difficulties were encountered as many players, but not all, had detailed scouting reports on separate pages from the mock draft pages. When possible, we gathered these more detailed reports, but used the mock draft page's smaller paragraphs as a fallback when necessary.

Finally, we collected data on every injury occurring between the 2013 and 2022 seasons, which we used to calculate the injury risk score, the variable our model was attempting to predict. This process began by taking Elap733's webscraping code for Pro Sports Transactions and Basketball Reference, which required heavy modifications due to both websites updating their underlying structure. Next, we modified Elap733's cleaning and processing code to fix issues and account for the separate timeframe we used. The processing involved taking the raw dates of injury list placements and activations and converting them into separate injury events with durations, as the duration data could not be found. Next, we used ICLiu30's dataset to augment the injury dataset.

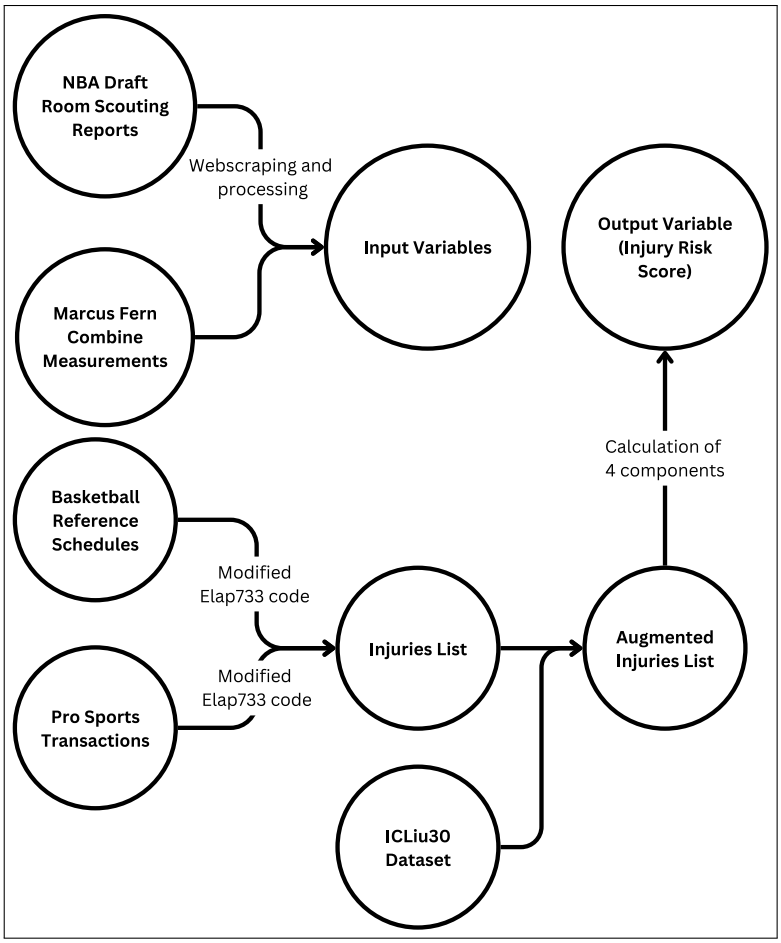

Figure 1: Data collection process.

2.2 Injury Risk Scoring

After collecting all of the data, we decided on how to map the inputs to the outputs. The injury data was in a complex format which didn't make sense to predict, so we developed a custom injury risk score (IRS), a float value from 0 to 100 that holistically captures how injury prone an NBA player is.

To set up the calculations for the IRS, we went through the following steps. First, ICLiu30's data was used to calculate the maximum number of games a player could have played in their career based on what seasons they were active in. Next, we removed all injury listings relating to personal reasons and illnesses as they are not indicative of injuries occurring during a basketball game. This meant that the only injuries used to calculate the IRS were fractures, sprains, concussions, tears, etc. The next step was to calculate the average length of each injury event in terms of games missed, the fraction of a player's maximum possible career games they missed, and the number of injury events per game. The latter two values (fraction of max career games missed and number of injury events per game) were then scaled by the player's minutes played per game, as playing heavier minutes increases injury likelihood. The final step was to weight these values, add them up, then scale them so the least injured player had an IRS of essentially zero and the most injured player had an IRS of 100. Specifically, we selected a weighting of 0.5 for the fraction of maximum career games missed, 0.3 for the average length of each injury, and 0.2 for the number of injuries per game. This weighting has no objective reasoning but was chosen based on how important we believe each factor to be.

Although it's not possible to objectively distill a concept as vague as injury likelihood to a single number, we believe the process described above is a lot more accurate than any singular stat like games missed or number of injuries. The end result is a value that's easy for a model to output that puts more of an impact on long injuries, repeated injuries, and careers shortened by injuries, while

accounting for the fact that starters are on the court risking injury for a longer time period than bench players.

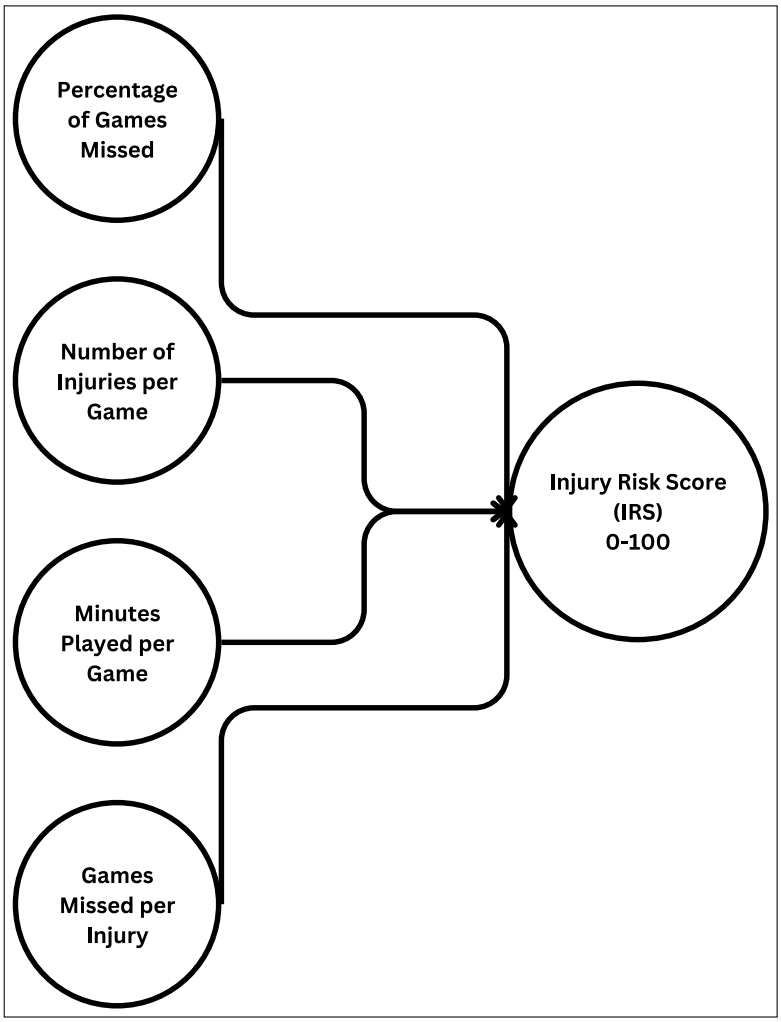

Figure 2: Injury risk score calculation.

2.3 Model Architecture

Our approach leveraged FT-Transformers, specifically designed for tabular data, which utilize self-attention mechanisms to emphasize significant measurements and capture complex interactions. Additionally, we processed the text data from the scouting reports using BERT (Bidirectional encoder representations from transformers), enabling the extraction of relevant insights related to a player's injury likelihood.

The FT-Transformers are tailored to handle tabular datasets effectively. Their self-attention mechanism dynamically weighs different features of draft combine measurements, allowing the model to prioritize the most relevant metrics, such as height and wingspan, based on their significance in predicting injury risk. Since we did not initially know which measurements are most important, the attention mechanism within the FT-Transformer determined the critical features on its own instead.

BERT is a language model known for its ability to grasp the context and nuances of language. In our approach, BERT processed predraft scouting reports, providing essential insights into each player's potential injury risk. It generated embeddings for words and sentences that capture contextual relationships, allowing the model to interpret phrases like "chronic knee issues" or "recovering from an ankle sprain" and translate them into actionable data for injury risk assessment.

After obtaining the results from both models, the feature outputs from the FT-Transformer and the embeddings from BERT were combined to create a comprehensive feature vector. This vector combined both the numerical measurements and qualitative insights from BERT, and was used to output the holistic injury risk score detailed in the previous section.

To implement the architecture described above, we built a custom model in PyTorch using the pytorch-tabular package's FT-Transformer implementation, as well as a pretrained BERT model from the HuggingFace transformers package. This combined model took the combine data, separated it into categorical and continuous columns, and fed both into the FT-Transformer architecture to generate a feature vector. Separately, we fed the scouting reports into BERT to generate the scouting report embeddings, and concatenated the embeddings with the FT-Transformer feature vector. Finally, we used a fully connected layer that took in the concatenated feature vector and outputted the injury risk score prediction.

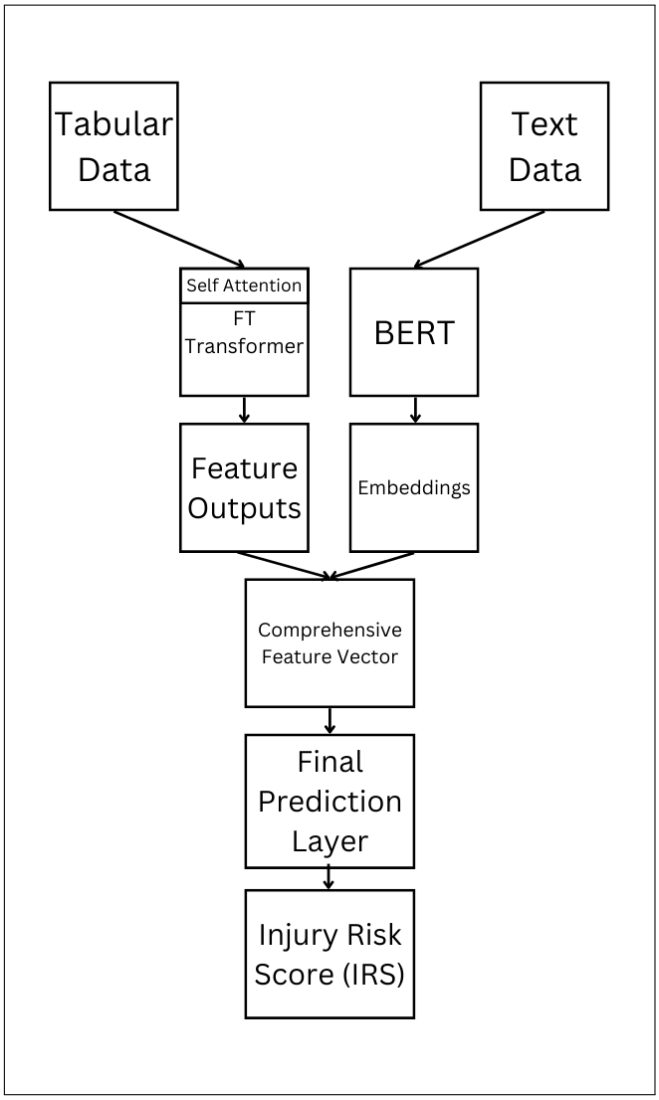

Figure 3: Diagram of the comprehensive injury risk prediction model, integrating FT-Transformers for tabular data and BERT for text analysis to produce a holistic injury risk score.

3 Experiments

To validate our model, we separated our original dataset into a training and testing set to ensure our model can predict injury proneness on players across the entire timeframe of our dataset. Then, we

trained the combined model described in the section above on the training data. After that, we fed the scouting reports and combine measurements for each player in the the test dataset to get the model's injury risk score predictions. The methods we used to analyze our model's efficiency will be explained in the next section.

4 Results and Discussion

We believe that our model has leveraged recent developments in deep learning architectures to accurately predict injury likelihood. We used Scikit-learn to run a regression between the predicted injury risk scores and actual injury risk scores for every player in the test dataset. This regression resulted in a mean absolute error of 4.47, meaning the average predicted IRS was roughly four and a half points away from the ground truth IRS. We also got a root mean squared error of 6.05, indicating that there were some outlier predictions causing the RMSE to be double the MAE.

We wanted to know how our model would perform as a classifier, so we thresholded both predicted and ground truth IRSes with a score above 20 indicating an injury prone player, and a score below 20 indicating a healthier player. This value was chosen by analyzing the data to determine a good threshold for a player being considered injury prone. When asked to classify test set players as injury prone or not, the model had an accuracy of 96%, which indicates that it will be a valuable tool for future draft prospect evaluation.

5 Limitations

While our model shows promise, it is essential to acknowledge its limitations. The following limitations mostly revolve around inadequate data, but there are some other limitations as well.

Scouting reports can be biased and written by different people in different formats, which could nullify the effectiveness of BERT. Additionally, our data source didn't contain scouting reports for every player, which could result in the model being biased against either players with a scouting report or without, depending on the specific training examples used. Our final dataset once removing all missing values contained less than 400 players, partially due to incomplete data and partially due to not that many players entering the league in the first place. There are also other professional basketball leagues that our model and dataset do not account for at all. While we have a few international players, they are not represented as well as American basketball players.

There were also some limitations with our model and project as a whole due to time constraints. First, we weren't able to test different weightings of the injury risk score or a different method of constructing the score in the first place, which may have helped us improve our model's accuracy. Next, we were committed to our original proposal's architecture of using FT-Transformers and BERT, and we ran out of time to customize them due to difficulties getting the model working correctly. Given more time, we would have liked to experiment with different pretrained versions of BERT or experimented with other tabular architectures such as Tab Transformers, which contain embeddings for categorical data that FT-Transformers do not.

6 Conclusion

Our deep learning model presents a novel approach to predicting injury risk of NBA draft prospects by using NBA draft combine measurements and predraft scouting reports. By combining the FT-Transformer architecture for combine data and the BERT language model for scouting reports, we were able to create a model that correctly determined a player as injury prone with 96% accuracy. Additionally, using metrics such as mean absolute error and root mean squared error, our predicted injury risk scores on average were only around 4 and a half points off of the actual scores. This indicates that our model may be a superior method of predicting injury likelihood for draft prospects compared to existing subjective methods, and we hope to expand on its capabilities in the future.

To address some of the model's limitations and make it more broadly useful, there is some future work we can do. First, we can test the model on older scouting reports and combine data (i.e., players drafted before 2013) as well as international players in basketball leagues such as the EuroLeague. We also want to take a deep dive into the model's weights to see if any specific combine measurements or scouting report phrases are likely to indicate heavily injury prone or extremely healthy NBA players.

For example, we may find that tall and heavy frontcourt players who have weight issues mentioned in their scouting reports are heavily injury prone, thus indicating that these players should be drafted with caution and may require extra work with team doctors. Adding existing injury data from the NCAA or international leagues to our model would add to our multipronged approach, instead of just using scouting reports and combine data. Finally, a user-friendly interface for our model may help enable widespread use among the draft analysis community, similar to websites like Tankathon that provide other information about draft prospects.

References

[1] Gorishniy, Y & Rubachev, I & Khrulkov, V & Babenko, A (2023) Revisiting Deep Learning Models for Tabular Data

[2] Devlin, J & Lee, K & Chang, M & Toutanova, K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

[3] Marcus Fern. (December 2023). NBA Draft Combine. Retrieved December 7, 2024 from https://www.kaggle.com/datasets/marcusfern/nba-draft-combine.

[4] ICLiu30. (December 2023). NBA Player stats and injured data from 13 to 23. Retrieved December 7, 2024 from https://www.kaggle.com/datasets/icliu30/nba-player-stats-and-injured-data-from-13-to-23.

[5] Elap733. (December 2019). NBA-Injuries-Analysis. Retrieved December 7, 2024, from https://github.com/elap733/NBA-Injuries-Analysis.

[6] NBA Draft Room. (2024). NBA Mock Drafts (2013–2022). Retrieved December 7, 2024, from https://nbadraftroom.com.

[7] Pro Sports Transactions. (2024). Basketball Transactions. Retrieved December 7, 2024, from https://www.prosportstransactions.com/basketball/.

[8] Sports Illustrated. (2024). NBA Draft 2024: Top Picks Getting Injured. Retrieved December 7, 2024, from https://www.si.com/nba/draft/newsfeed/nba-draft-2024-top-picks-getting-injured.

[9] Farghaly, O & Deshpande, P (2024). Leveraging Machine Learning to Predict National Basketball Association Player Injuries. Retrieved December 7, 2024, from https://ieeexplore.ieee.org/document/10636005.