Iteration #02: NBA Injury Prevention

Tea Adams, Shriniketh Mukundan, Jay Shah

1. Project Kickoff - Tea

Specific Goals & Expected Outcomes:

Our goals include:

• Analyze historical NBA injury data and player characteristics to identify key factors contributing to injuries.

• Develop a predictive model to estimate the likelihood of player injuries.

• Evaluate the model's accuracy using real-world data and refine it for practical application.

• Provide recommendations for injury prevention and load management strategies based on the model's insights

We expect to produce a reliable injury prediction model that identifies NBA players at higher risk of injury. This model will help teams take proactive measures to prevent injuries, extend player careers, and improve overall team performance. Additionally, it will offer financial benefits by reducing medical costs and minimizing the impact of injuries on player salaries.

Project Scope:

This project focuses on developing a predictive model to estimate the likelihood of injuries among NBA players using historical injury data and player characteristics such as age, position, minutes played, and performance statistics. The scope includes data collection, cleaning, analysis, feature selection, model training, and evaluation.

The project will not include real-time injury tracking, medical diagnostics, or external factors outside available performance data (e.g., mental health, off-court incidents).

The final deliverable will be a data-driven model and a report summarizing key findings, predictive accuracy, and actionable recommendations for teams to reduce injury risks.

Key Deliverables:

Data Collection - Gather historical NBA injury data and player statistics (games played, minutes, age, position). Clean and preprocess the data (Normalize formats, handle missing values) Data Analysis - Identify trends, correlations, and key factors contributing to injuries. Visualize injury patterns over time and across player types. Model Development - Build and train a predictive model using logistic regression techniques learned in class. Insights and Recommendations - Translate outputs into actionable insights for prevention and present findings visually and in a final report.

Major Miles & Deadlines:

Our milestones coincide with our deliverables:

1. Data Collection and Cleaning - ready to use dataset for analysis

2. Model Development and Testing - Validated predictive model with performance metrics

3. Final Report and Presentation - completed report and presentation

Capabilities & Gaps:

Our team has a wide variety of skills including analytical, technical and research skills to the project. We are comfortable working with data analysis concepts and have prior experience using Python and data visualization tools. Collectively, we have a strong understanding of basketball performance metrics, which will assist in interpreting the model and results.

Although our team has a range of strong capabilities, we recognize the need for continued refreshers in certain programming and data analysis libraries. We also plan to strengthen our understanding of machine learning model tuning and evaluation techniques. To address these gaps, we will collaborate closely, review class materials, utilize online resources, and hold regular progress check-ins to ensure we stay aligned and meet our project goals.

Existing Data Set:

Currently, we have two datasets available: Games Played from 2010–2020 and player injury records during the same period. We are still searching for player rosters and statistics, which will be the final dataset needed to complete our analysis.

2. Team Discussion - Shriniketh

Core Skills and Member Expertise:

Each member of our team possesses strong analytical skills that are needed for this project, and each member brings a unique contribution to this project in a cohesive manner. For example, Tea's most unique skills are Python programming, given her prior Data Analyst experience, as well as teamwork and leadership. This is crucial for the project since her role as a Data Analyst for this task involves relying on programming and how it is needed to analyze and predict models on the desired data in order to provide more insights regarding injury likelihood and predict injury trends across imperative metrics. More importantly, her ability to work as a team will provide more growth and opportunity throughout this project, regarding new insights and the presentation of our findings. Secondly, Jay's prior experience in Data Analytics, Aggregation, and Cleaning, as well as his skillsets involving Python and SQL, are very important regarding the datasets that are going to be used and whether the final datasets consist of clean and usable data that will provide more accurate injury predictions. Regarding his role as the Data Engineer, his role is comprised of handling any missing injury data, cleaning any inconsistencies within the current datasets, and providing relational schemas that are able to connect the different entities and their associated attributes. In doing so, his role is crucial for understanding the connection of different pieces of data, which will be important regarding the metrics used and how to present the final insights. Lastly, Shriniketh's prior experience in Business Intelligence, as well as his skill sets surrounding Python programming and SQL query implementation, are important for his role as the Data Visualization Lead since he has some prior experience providing data insights that were used for dashboard creation via Microsoft Power BI. Therefore, his role is important because it expands much more on the dashboard and report creation, and this is needed to visualize injury risk trends across positions and teams, and identify players who are more at risk of experiencing a severe injury. Lastly, everyone in this group possesses many soft skills, including communication, teamwork, and persistence, which are important for the success of this project because only with constant updates and collaboration can we provide meaningful insights that can have a meaningful impact regarding how coaches and their medical staff should assess their players in the future and make decisions regarding how they should play as well as other factors such as how many minutes per game they should play so that they avoid career-ending injuries that are detrimental for the players and the success of their team.

Potential Missing Skills & Desired Tools/Technologies Used:

While each member of the team possesses unique and extremely important skill sets, there may be some skills that cannot be overlooked or skills that need to be explored more throughout the duration of this project. For example, since Jay will be primarily using SQL for data cleaning and schema relations while Tea is using Python and predictive modeling to derive insights, it's important that EDA is conducted via scalable computing. Therefore, another skillset that would come in handy is Apache Spark, which can be used to provide a large-scale multi-language engine that can handle the data engineering, data science, and machine learning aspects of the project on single-node machines or clusters. By learning more about this analytics engine, the data cleaning, analysis, and predictions can all be conducted, making the analytics process more streamlined across the different users(project members). Lastly, expanding the advanced analysis tools of Microsoft Power BI, such as different AI capabilities that help with report creation and data patterns, can help immensely towards the success of this project and present the most meaningful insights that different teams across the NBA can use regarding how to best assess and manage their players in a way that enables them to help the team succeed while best mitigating their chances of suffering an injury.

Regarding the tools and technologies that the team already has experience with, everyone in the team already has experience using Python and SQL, which are important for cleaning and analyzing data with the use of predictive models in order to derive the most important insights that teams and players alike can utilize for their own health and safety. However, as mentioned before, some skills that should be emphasized more throughout this project are the need for Cloud-Based analytics tools implementations, such as Microsoft Azure Databricks, which is a visualization software that functions on Apache Spark as the analytical engine. Additionally, more research into the new features and capabilities of Microsoft Power BI is also needed in order to provide the most accurate and meaningful data that can be overlooked if proper time isn't devoted to this area. Lastly, more research on existing Machine Learning libraries should also be emphasized in order to understand which libraries are more suited to the data and task at hand. Additionally, given Microsoft Azure Databricks, understanding Managed ML Flow capabilities can be extremely useful regarding the machine learning implementation aspect of this project, as Databricks' Managed ML Flow delivers state-of-the-art experiment tracking, observability, performance evaluation, and model management. Therefore, based on the project's needs and the team's background, utilizing Python, SQL, and Cloud Analytics Platforms such as Microsoft Azure Databricks, and Visualization Tools, including Microsoft Power BI, is needed to best meet the goals and provide the most useful basketball insights that teams and specific players alike can utilize.

3. Skills and Tools Assessment: Jay

External Resources:

The people on our team collectively have experience in database design, data analysis, and visualization, but it is possible that we still may need help in certain areas while we work on this project. For something involving predictive modeling validation, we may need help from faculty members or TAs who are well-experienced with sports analytics and also have knowledge in machine learning to review our model and methods. For data sourcing and cleaning, we could use some help from sports data APIs like Sportradar and Basketball Reference (2010-11 NBA Schedule — Basketball-Reference.com). All of these will be helpful to find and filter reliable player and injury data. For visualizations, we will also likely use external help from videos online, peers, and people in general with experience in Tableau/other visualization tools to enhance our dashboard's clarity and meaningfulness of tools.

Tools, Frameworks, Libraries:

Project Area	**Tools/Frameworks**	**Reasoning**
Database	PostgreSQL, SQLAlchemy, Python pandas, Prefect, Azure DataBricks	Will be helpful for data mining, transformation, and modeling
Predictive Modeling	Python - NumPy, matplotlib, pandas, scikit learn, etc.	Helpful for building models, evaluating, and interpreting
Visualizations	Tableau, Power BI	Provide friendly visuals for injury trends
Collab Tools	GitHub, Google Drive	Easy ways to collaborate with one another

Ensuring Team Proficiency:

To ensure everyone is confident in the tools we will be using, we believe there are certain steps we should follow to ensure everything runs smoothly. If certain members are more experience

with a type of tool than others, they can hold some meetings where they explain and give walkthroughs on the tool. For example, Jay can lead SQL walkthroughs, Tea can show demonstrations of building models in Python, and Shriniketh can cover visualizations. Along with this, it would be helpful to have a Google Doc set up with instructions, common commands that we will need to use, and the best ways to handle certain situations. We can also do certain parts of the project in pairs, where one may be more experienced than the other, to ensure everyone is learning as best as they can as the project moves forward. Lastly, a weekly check-up meeting to make sure everyone is doing good will be a major plus.

Task Assignment and Clarity:

So far, we have assigned tasks based on each member's strengths and career goals. Though, like said previously, we will be making sure that everyone learns a good amount of everything:

- Jay Shah - Data Engineer/Database Designer: SQL & database design, ER modeling, normalization

- Tea Adams - Data Analyst/Predictive Modeling Lead: Programming, analysis, problem-solving, and predictive modeling

- Shriniketh Mukundan - Data Visualization/Reporting Lead: Data visualization, communication, collaboration

4. Initial Setup: - Jay

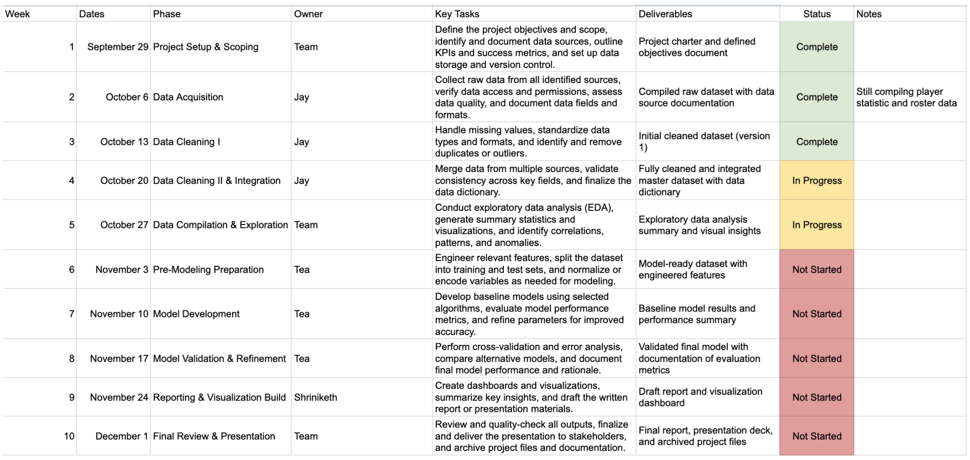
Development Environment:

Our project will be developed using both cloud and local environments to ensure collaboration, scalability, and accessibility. We will use Azure Databricks Workspace as the primary environment for data processing, model development, and experimentation. This is also very useful, as it supports collaborative notebooks for all team members. Each team member will have a local Python environment, utilizing Anaconda, for initial data exploration, SQL testing, and visualization development. For our database environment, we plan on using PostgreSQL, which can be used on Azure for the storage of data such as players, games, and injuries. There can also be connections made through Databricks for computations. For visualization tools, we will develop the Dashboard locally and deploy it through Tableau. This entire setup enables us to create an ideal cloud environment for data engineering and modeling, while also allowing for local testing and visual creation.

Version Control and Repository Access:

Version control has been configured through GitHub, and all the team members have collaborator access. The repository will store all relevant materials, including data, ETL scripts, modeling notebooks, and visualizations. Commits and documentation will help maintain consistency throughout the project. We will also link our GitHub to Databricks notebook using the built-in integration methods. This will synchronize development across all environments. Here is the link to our repository: NBA Injury Modeling GitHub Repository.

Progress Tracker: Team Progress Tracker (Google Sheets).



Week	Dates	Phase	Owner	Key Tasks	Deliverables	Status	Notes
1	September 29	Project Setup & Scoping	Team	Define the project objectives and scope, identify and document data sources, outline KPIs and success metrics, and set up data storage and version control.	Project charter and defined objectives document	Complete	
2	October 6	Data Acquisition	Jay	Collect raw data from all identified sources, verify data access and permissions, assess data quality, and document data fields and formats.	Compiled raw dataset with data source documentation	Complete	Still compilng player statistic and roster data
3	October 13	Data Cleaning I	Jay	Handle missing values, standardize data types and formats, and identify and remove duplicates or outliers.	Initial cleaned dataset (version 1)	Complete	
4	October 20	Data Cleaning II & Integration	Jay	Merge data from multiple sources, validate consistency across key fields, and finalize the data dictionary.	Fully cleaned and integrated master dataset with data dictionary	In Progress	
5	October 27	Data Compilation & Exploration	Team	Conduct exploratory data analysis (EDA), generate summary statistics and visualizations, and identify correlations, patterns, and anomalies.	Exploratory data analysis summary and visual insights	In Progress	
6	November 3	Pre-Modeling Preparation	Tea	Engineer relevant features, split the dataset into training and test sets, and normalize or encode variables as needed for modeling.	Model-ready dataset with engineered features	Not Started	
7	November 10	Model Development	Tea	Develop baseline models using selected algorithms, evaluate model performance metrics, and refine parameters for improved accuracy.	Baseline model results and performance summary	Not Started	
8	November 17	Model Validation & Refinement	Tea	Perform cross-validation and error analysis, compare alternative models, and document final model performance and rationale.	Validated final model with documentation of evaluation metrics	Not Started	
9	November 24	Reporting & Visualization Build	Shriniketh	Create dashboards and visualizations, summarize key insights, and draft the written report or presentation materials.	Draft report and visualization dashboard	Not Started	
10	December 1	Final Review & Presentation	Team	Review and quality-check all outputs, finalize and deliver the presentation to stakeholders, and archive project files and documentation.	Final report, presentation deck, and archived project files	Not Started	

Figure 1: Progress Tracker