**C951**

**WGU**

**Introduction to Artificial Intelligence**

**Task 3**

**Daniel Roberts**

**Student ID: 009198202**

**Table of Contents**

**A. Project Overview 3**

* **A.1.** Organizational Need **3**
* **A.2.** Context and Background **3**
* **A.3.** Outside Works **4**
* **A.4.** Machine Learning Solution **5**
* **A.5.** Benefits **5**

**B. Machine Learning Project Design 6**

* **B.1.** Scope of Project **6**
* **B.2.** Goals, Objectives, and Deliverables **6**
* **B.3.** Methodology **7**
* **B.4.** Timeline **8**
* **B.5.** Resources **8**
* **B.6.** Criteria for Success **9**

**C. Machine Learning Solution Design 10**

* + **C.1.** Hypothesis **10**
  + **C.2.** Algorithms **10**
  + **C.2.(a)** Justification of Algorithm **10**
  + **C.3.** Tools and Environment **11**
  + **C.4.** Measurement of Performance **11**

**D. Description of Data Set(s) 12**

* + **D.1.** Sources of Data **12**
  + **D.2.** Data Collection Method **12**
  + **D.2.(a)** Advantage and Limitation of Data Collection Method **12**
  + **D.3.** Data Normalization **13**
  + **D.4.** Security of Data **13**

**E. Sources**

- **E.1.** Citations **14**

1. **Project Overview**

DBR Sports Wagers is a startup company whose mission is to simplify sports betting. They are searching for a solution that will utilize Artificial Intelligence to give people the best odds when placing sports bets. This solution will satisfy that goal for DBR Sports Wagers. This document explains the proposed AI solution that will fit the scope of DBR Sports Wagers’ objectives, goals, and requirements.

**A.1. Organizational Need**

DBR Sports Wagers has created a thorough business plan for their organization. The only thing the organization is missing is the technology to fuel that plan. The organization needs an AI system that will analyze a large and vast amount of data that is updated in near real-time to keep their customers at an advantage when placing their bets.

**A.2. Context and Background**

DBR Sports Wagers began as the idea of three college students, Daniel, Blake, and Robert. All three students were avid sports bettors with a love for technology. After completing their Introduction to Artificial Intelligence course at Western Governors University, they recognized the potential for artificial intelligence in the sports betting industry.

The three secured financial backing and created a well-planned business outline. DBR Sports Wagers needed an engineering company that specialized in artificial intelligence to complete their plan. Their goal was to create a system that analyzed a vast amount of data to predict the outcomes of sports games. They believe that data-driven insights would help bettors make better-informed decisions, ultimately increasing their chances of success.

**A.3. Outside Works**

**(1)**

In sports betting, there are many data points to keep track of including who’s playing, what time the game starts, and the location the game will be played. Machine learning and using algorithms to analyze large amounts of data can determine patterns and trends to assist in better predicting outcomes. (Grigoryan, 2022).

**Relation to Project:**

Our solution is to leverage data to make predictions. Our algorithm must identify patterns and trends in that data to accurately make those predictions and ultimately be profitable for DBR Sports Wagers. The above article shows that machine learning solutions have their place in the sports betting industry and that data can be used to identify patterns and trends to accurately predict an outcome.

**(2)**

According to intellectdata.com, “Sports analytics applies supervised machine learning algorithms like naive bayes, decision trees, linear regression, and neural networks. Sports analytics also includes unsupervised machine learning algorithms such as k-means clustering and association rules. These algorithms act on sports data collected from various sources to draw valuable conclusions concerning player performance and team efficiency. Numerous scenarios exist where machine learning may be implemented in the sporting world.” (“Machine Learning in Sports Analytics and Predictions”, 2021).

**Relation to Project:**

For our machine learning algorithm to be effective, we will need to utilize a training technique to ensure its effectiveness. The above statement shows that both supervised and unsupervised learning techniques have their place in training sports prediction algorithms to be effective.

**(3)**

According to an article published by Medium.com, a supervised learning method is one of the most common approaches when using machine learning to predict sports outcomes. (Shah, 2023).

**Relation to Project**

Supervised learning techniques are the most used methods when training machine learning algorithms for sports predictions. We will be using a supervised learning approach for our algorithm. We will input data that is already associated with a given outcome for our algorithm to identify trends and patterns in that data.

**A.4. Machine Learning Solution**

My solution aims to utilize historical and near real-time sports data to identify patterns and make predictions on outcomes. It will need to consider a large amount of data including historical team and player statistics, weather conditions, current series statistics, injuries, etc. As stated in the article “Machine Learning in Sports Analytics and Predictions”, supervised learning algorithms are the more commonly used approaches. We will be utilizing the Random Forest algorithm approach for our project.

**A.5. Benefits**

Sports betting has grown astonishingly over the past several years due to ease of access. Bettors can now simply download an application on their phone and place whatever bet they like. According to grandviewresearch.com, “The global sports betting market accounted for USD 83.65 billion in 2022 and is expected to expand at a compound annual growth rate (CAGR) of 10.3% from 2023 to 2030”. With that expected growth, it is plausible to believe that more and more bettors will turn to sports prediction algorithms prior to placing bets. DBR Sports Wagers will benefit from this growth substantially.

**B.** **Machine Learning Project Design**

**B.1. Scope of Project**

DBR Sports Wagers’ artificial intelligence system aims to gather past and near real-time sports data to give customers a probability of sports outcomes. Those outcomes would include probabilities for which team wins, player statistics, and more. They gather to gain as much information as possible for the system to increase bettors’ chances of winning any given bet. For creating the initial algorithm, we will need a team of at least 3 including a Software Engineer, a Data Scientist, and a UI/UX Designer. For the initial project, our data and machine learning algorithm will only cover the NFL and NBA. Past sports data will only go back 2 years, to the 2021 sports season for both. DBR Sports Wagers will be responsible for any data entry to give probabilities of outcomes that are considered near real-time, however, the scope of this project does include the algorithm for that data when gathered.

**B.2. Goals, Objectives, and Deliverables**

DBR Sports Wagers aims to have a 70% proven success rate when making NFL and NBA predictions. In order to reach that success rate, we will need to gather a multitude of sports data and compare that to historic game outcomes. We will need to train our algorithm to identify patterns in the data and yield the outcome from those historic games at least 70% of the time. We will set out minimum acceptable accuracy, for deployment, of 65%. During our testing stage and prior to deployment, we will run 100 tests on historical games and 100 tests on future games. Each 100 games will include 50 NFL and 50 NBA games. Our test cases will only consider the combined overall accuracy for our target 70% (We will not differentiate the NFL games from the NBA games in reference to accuracy). With the right approach, we should have an application that can make accurate predictions and has a user-friendly interface that any legal-age sports bettor can utilize.

**B.3. Methodology**

Agile methodology is the best methodology for this project. Working in sprints will be beneficial when trying to create the algorithm to be as accurate as possible. Below, I will go through the pros and cons of using the agile methodology for this project.

There are many pros when it comes to using an Agile approach, especially for the stakeholders. Since predicting sports outcomes is a very difficult process, it is impossible to make predictions with 100% accuracy. Our goal is to be able to reach an average of 70% accuracy on most outcomes. To reach this goal, many changes may need to be made to our project. Using an Agile approach will help in this respect. We can utilize working in sprints to better perfect our algorithm and data collection which will be extremely important when trying to calculate probabilities of different outcomes.

With many advantages to something, there can come many disadvantages. Agile approaches are often more time-consuming, usually lack needed documentation, and can be very unpredictable. Time and the likelihood of changes will be the greatest disadvantage to using the Agile approach on our project. However, the main goal is to reach the highest accuracy possible when making predictions. An Agile approach will help give us more complete data and ensure our algorithm has the highest accuracy possible.

We will break our project into sprints. Working in sprints will help give us continuous improvement and feedback from DBR Sports Wagers. The more input from DBR Sports Wagers, the better we can mold our application to their wants. The below table shows how our sprint cycles will be conducted.

|  |  |
| --- | --- |
| **Sprint** | **Event** |
| 1 | Securing Personnel |
| 2 | Securing Resources |
| 3 | Data Collection |
| 4 | Source Comparison |
| 5 | Creation of Database |
| 6 | Final Data Cleanse |
| 7 | Coding |
| 8 | Training |
| 9 | Testing |

**B.4. Timeline**

If a team of 3 is approved by DBR Sports Wagers, the estimated time to complete the entire algorithm would be 6 months. Otherwise, if no other team members are available, the estimated time to complete the algorithm would be upwards of 1 year. Below is an estimated timeline of events leading to the completion of the project with 3 team members. It should be noted that, to reach a sustainable accuracy of up to 70%, testing and training would need to be completed for at least 1 season of NFL and NBA games. For the initial project, we will keep our goal of reaching 70% accuracy by training the system for 1 month using historic game data to compare it to our outcome probability given by the system.

|  |  |  |
| --- | --- | --- |
| **Event** | **Start Date** | **Date Completed** |
| Proposal Accepted | N/A | 06/25/2023 |
| Resources and personnel secured | 06/26/2023 | 07/25/2023 |
| Data Collection | 07/26/2023 | 08/25/2023 |
| Coding | 08/26/2023 | 10/25/2023 |
| Testing and Training | 10/26/2023 | 11/25/2023 |
| Product Review and Approval | 11/26/2023 | 12/25/2023 |

**B.5. Resources**

|  |  |
| --- | --- |
| **Resources** | **Estimated Cost** |
| Team | $200,000 (6-month contract) |
| Data | $1,000 (licensing and subscription fees for a minimum 1 year) |
| Infrastructure | $10,000 (Hardware and/or cloud-based services) |
| Software | $5,000 (Software programs for machine learning, IDEs, code editors, etc.) |
| Testing | $5,000 (Estimated cost for testing application and fixing bugs) |

**B.6. Criteria for Success**

Our criteria for success will be based on several factors including the model’s accuracy and consistency, profitability for users, its ability to handle continuous changes in data (injuries, weather, etc.), and its comparison to already existing sports prediction algorithms. Our goal for consistent and acceptable accuracy for our algorithm is 65% to 70% for the 100 historic and 100 future games. We will need to keep continuous comparison checks to other algorithms and set our target to the top 5 algorithms in terms of accuracy and profitability for its consumers.

**C. Machine Learning Solutions Design**

**C.1. Hypothesis**

If we can gather enough sports data and train our algorithm to produce the desired outcome from historic games, then we will be able to reach the desired goal of a 70% accuracy rate on future games.

**C.2. Algorithms**

We are proposing to go with a supervised learning approach for our machine learning algorithm. We will input the data from previous games and give the system the desired output for the game that it’s predicting. For our supervised machine learning technique, we will implement the Random Forest algorithm.

**C.2(a). Justification of Algorithm**

Utilizing a supervised learning approach will be the most time-appropriate way to reach our goal. We are allotting a 1-month time period for testing and training. Having our system pull the required data and giving the system the desired outcome from historic games will be our most time-cautious way to reach a 70% accuracy rate for future games. Given that a vast amount of historical game data is readily available and accessible, this will be the quickest method for our algorithm to begin to identify patterns and make predictions. A supervised learning approach has been proven to be successful in machine learning with sports predictions. According to an online article from Medium.com, “One of the most common approaches to predicting sports outcomes is supervised learning, a type of machine learning algorithm that uses labeled data to train a model. In the case of sports prediction, the labeled data might include historical records of game outcomes, along with data about the teams, players, and other relevant factors.” (Shah, 2023). For our supervised learning technique, we have chosen the Random Forest algorithm. The Random Forest algorithm combines multiple decision trees to make predictions. It can be trained to use historical data including team and player statistics, game and weather conditions, etc. to make accurate predictions on labeled outcomes. According to an article written on Algobeans.com, “Random forests are widely used because they are easy to implement and fast to compute. Unlike most other models, a random forest can be made more complex (by increasing the number of trees) to improve prediction performance without the risk of overfitting.” (Algobeans, 2021).

**C.3. Tools and Environment**

We will utilize Microsoft Visual Studio for our programming IDE. Visual Studio supports the Python programming language, which will be utilized to create our machine learning algorithm. We will need to utilize data tools such as Python’s Pandas to help process our data. We will utilize a cloud-based platform, such as Amazon Web Services, for hosting and deployment.

**C.4. Measurement of Performance**

For our initial project goal of reaching a 70% success rate, based on our algorithm correctly choosing the outcome of a historic game, our measurement for performance will be comparing the predictions with the actual outcome of the game. The same will apply when using the predictions when comparing them to future games prior to deployment. Once our algorithm meets an average success rate of 70% on predictions made, our project will be considered a success.

**D. Description of Data Set(s)**

**D.1. Sources of Data**

We will use 3 sources to gather this information to perform an accuracy check. We will gather our data from ESPN.com, NFL.com, and NBA.com. DBR Sports Wagers will be responsible for any present or near real-time data such as weather conditions, home or away team, injuries, etc.

**D.2. Data Collection Method**

We will have to research and create our own datasets, gathering necessary information from the past 2 years of NBA and NFL games. There are a multitude of free datasets online for that, however, that information would need to be thoroughly reviewed and cleaned. Our project proposes to create our own data sets from the last 2 sports seasons. This will ensure our data is accurate, give us complete ownership of the datasets, and help us understand the information to better build our machine learning algorithm.

**D.2.(a) Advantage and Limitation of Data Collection Method**

The limitation of Data Collection will be real-time and near real-time data. Such as weather conditions, home and away team, and player and team stats on a particular drive. The responsibility for gathering that data and placing it into our prebuilt datasets will fall to DBR Sports Wagers when the time is appropriate.

Creating our own datasets has many advantages. Gathering our own data and creating our own datasets with it will give us complete ownership of that dataset. It will give us a better understanding and knowledge of the information we are gathering which will help us when building our algorithm. Most importantly, we can ensure that the data being gathered is accurate. We will be comparing the data gathered with 3 legitimate sources including ESPN, and the official NFL and NBA websites.

**D.3. Data Normalization**

Since we are creating our own data sets by gathering data from multiple sources, we will need to first compare data from all sources to ensure accuracy. We will need a method for selecting our source of data when/if two or more data sources show discrepancies. For our project, we will utilize an average approach. For example, if NFL.com shows a statistic of 90 yards for a particular game and ESPN.com shows the same statistic as being 92 yards, we will use the average between the two giving us 91 yards for our data set.

We will need to organize our data into columns and rows. Each game is a column and each statistic being used is an associated row. We will utilize a min-max scaling approach for our data normalization. This will allow our algorithm to be fairer when comparing different data points.

Since we will be creating our own data sets for use, this should strongly minimize the amount of data cleansing we have to conduct on the backend. Most of our data checks and cleansing will be conducted at the beginning whenever we are comparing the data from our sources and adding them to our data set.

**D.4. Security of Data**

Our data being gathered is not considered confidential or sensitive in any manner. All the information we will be collecting is free information provided to the public. We are proposing a cloud-based storage to host our datasets such as Amazon Web Services. Most cloud-based storage, including AWS, has many security measures already in place to secure your data including access security, encryption measures, and backup and recovery parameters.

**E. Sources**

Grigoryan, S. (2022, August 29). *How Machine Learning Works in Sports Betting.* Feedconstruct.com. <https://www.feedconstruct.com/blog/how-machine-learning-works-in-sports-betting-1>

Shah, P. (2023, May 4). *Predicting Sports Outcomes with Machine Learning.* Medium.com. <https://medium.com/nerd-for-tech/predicting-sports-outcomes-with-machine-learning-b05a8b91562f#:~:text=One%20of%20the%20most%20common%20approaches%20to%20predicting,about%20the%20teams%2C%20players%2C%20and%20other%20relevant%20factors>.

Intellectdata. (2021, October 11). *Machine Learning in Sports Analytics & Predictions.* Intellectdata.com. <https://intellectdata.com/machine-learning-in-sports-analytics-predictions/#:~:text=Sports%20analytics%20applies%20supervised%20machine,means%20clustering%20and%20association%20rules>.

Algobeans. (2021, March 29). *Random Forest Tutorial: Predicting Goals in Soccer.* Algobeans.com. <https://algobeans.com/2021/03/29/random-forest-tutorial-predicting-goals-in-football/#:~:text=Random%20forests%20are%20widely%20used,without%20the%20risk%20of%20overfitting>.