CS777 – Term Project Proposal Submission Template

**Student: Jake Giguere**

1. Data set description: Provide a detailed description of the public data set you have selected, including its source, format, and any relevant details about the data.

|  |
| --- |
| *Using a publicly available Baseball data set from* [*Baseball Reference*](https://www.baseball-reference.com/) *this site, has thousands of data sets that are formatted in tables and be easily extracted using web scraping or pulling as an excel spreadsheet. For example, data about specific player performance from multiple seasons. The size should easily fit the requirements, if we consider the last 5-10 years for each team.* |

1. Research question: Clearly define your research question and explain why studying is important. What do you want to learn from the data?

|  |
| --- |
| *Can we cluster players based on their performance data to predict the likelihood of certain expectations and game outcomes?*  *This analysis can provide insights to which player types contribute most to the teams success and help teams identify performance trends. I have been working on an architecture for the past year called PLONN it is an MLP that will be transformed into a RL agent this MLB-offseason where I can learn trends in players that are more likely to have good game performances.* |

1. Machine Learning model: Specify the type of machine learning model you plan to use, such as classification or clustering, and explain why you have chosen this model.

|  |
| --- |
| *I will start by using K-Means Clustering to group players based on their statistics, then follow up with classification using logistic regression in pySpark to predict hits. Clustering can help identify player archetypes such as contact hitters, power hitters, etc. After clustering, classification models can assess how these archetypes correlate with outcomes.* |

1. Expected outcomes: What do you expect to achieve after implementing your learning model? What do you hope to learn or discover from your data analysis?

|  |
| --- |
| *I expect to be able to determine which cluster contributes more to a specific outcome such as hitting and determine which player types are associated with standout individual performance* |

1. Evaluation plan: Explain how you plan to evaluate your project and assess the correctness of your model. What metrics or methods will you use to evaluate the effectiveness of your learning model? How well do you expect the model to work, and how will you measure its performance?

|  |
| --- |
| *The approach is the to use a within-cluster sum of squares (WCSS) for clustering quality and Accuracy score, precision recall, F1 score for classification performance. Depending on how I clean the data, I’m sure we can expect at least a reasonable accuracy, leveraging my familiarity with baseball metrics.* |