Capstone Proposal

Proposal

**Domain Background**

**In this section, provide brief details on the background information of the domain from which the project is proposed. Historical information relevant to the project should be included. It should be clear how or why a problem in the domain can or should be solved. Related academic research should be appropriately cited in this section, including why that research is relevant. Additionally, a discussion of your personal motivation for investigating a particular problem in the domain is encouraged but not required.**

Machine learning and baseball go hand-in-hand because of the volume of data that is available to teams. Volume has only gone up with the addition of Statcast. [History of Statcast]. In addition Machine Learning has been applied to baseball data both publicly among fans for pleasure and privately among professional teams in an attempt to gain an advantage in winning games. [Link to machine learning predict hits study]. Machine learning may help players improve pitch recognition. I am very interested in baseball, it got me into machine learning, it is how I learned to use the R programming language.

**Problem Statement**

**In this section, clearly describe the problem that is to be solved. The problem described should be well defined and should have at least one relevant potential solution. Additionally, describe the problem thoroughly such that it is clear that the problem is quantifiable (the problem can be expressed in mathematical or logical terms), measurable (the problem can be measured by some metric and clearly observed), and replicable (the problem can be reproduced and occurs more than once).**

Use Statcast data, consisting of perceived velocity, spin rate, release point and pitch type to classify pitches into balls and strikes before they actually cross the plate. See if there is a correlation between accuracy and pitcher quality, by using FIP or another pitching metric.

**Datasets and Inputs**

**In this section, the dataset(s) and/or input(s) being considered for the project should be thoroughly described, such as how they relate to the problem and why they should be used. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included with relevant references and citations as necessary It should be clear how the dataset(s) or input(s) will be used in the project and whether their use is appropriate given the context of the problem.**

Describe Statcast, Baseball Savant, and all metrics used. Explain balls and strikes works is conducive to unsupervised classification.

**Solution Statement**

**In this section, clearly describe a solution to the problem. The solution should be applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, describe the solution thoroughly such that it is clear that the solution is quantifiable (the solution can be expressed in mathematical or logical terms) , measurable (the solution can be measured by some metric and clearly observed), and replicable (the solution can be reproduced and occurs more than once).**

The feature we are trying to predict is classify each pitch thrown into balls and strikes, which can be identified easily by binary classification. We can use accuracy score and f-score to determine how well the algorithms were at classifying each pitch. Additionally we would want to see what (if any) relationship exists between the accuracy score and f-score of each pitcher and the effectiveness of that pitcher, as determined by FIP.

**Benchmark Model**

**In this section, provide the details for a benchmark model or result that relates to the domain, problem statement, and intended solution. Ideally, the benchmark model or result contextualizes existing methods or known information in the domain and problem given, which could then be objectively compared to the solution. Describe how the benchmark model or result is measurable (can be measured by some metric and clearly observed) with thorough detail.**

This one I’m unsure about – would benchmark model be actual results? I.e. whether pitch was classified as a ball or a strike, and the effectiveness of the pitcher.

**Evaluation Metrics**

**In this section, propose at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model. The evaluation metric(s) you propose should be appropriate given the context of the data, the problem statement, and the intended solution. Describe how the evaluation metric(s) are derived and provide an example of their mathematical representations (if applicable). Complex evaluation metrics should be clearly defined and quantifiable (can be expressed in mathematical or logical terms).**

Accuracy and F-Score, describe them. Correlation coefficient.

**Project Design**

**In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.**

Approach for solution is to download data, process/clean data, test several different algorithms, identify which works best by accuracy/f-score, pick several pitchers of different effectiveness/quality, run correlation.

**Before submitting your proposal, ask yourself. . .**

**Does the proposal you have written follow a well-organized structure similar to that of the project template?**

**Is each section (particularly Solution Statement and Project Design) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?**

**Would the intended audience of your project be able to understand your proposal?**

**Have you properly proofread your proposal to assure there are minimal grammatical and spelling mistakes?**

**Are all the resources used for this project correctly cited and referenced?**