Capstone Course Project Proposal

# Project Description

In Major League Soccer (MLS) there is a large number of players available to sign all with a variety of attributes and playing styles. Combined with the variation in playing styles of the 29 teams in the league and their financial limitations, the task of identifying high quality players for a team to sign is difficult.

The objective of this project is to predict a player’s ability with a given team then find the optimal player to sign given a club’s financial and roster limitations. This project will achieve this by combining predictive modeling and combinatorial optimization to identify the best players available for a club to sign.

# Proposed Project Title

Prescriptive Modelling for Transfer Targets in Major League Soccer.

# Project Rationale

The reason for taking on this project is to aide decision making when identifying players to sign for a club. A combinatorial optimization algorithm applied to available players in the transfer market will allow an analysis department to quickly identify high-quality players within the team’s budget. This provides value to a club because it reduces the amount of time required for a club to identify key targets and it also allows clubs to apply their scouting methodology to players that may otherwise be unknown. Lastly it gives a club the ability to quantify a player’s predicted output compared to their cost.

# Project Purpose

The client for this project is soccer club Orlando City Soccer Club. They are a soccer team that competes in the top tier of American soccer, Major League Soccer (MLS). The client contact for this project is Caleb Shreve, the Head of Analytics for Orlando City Soccer Club. Caleb directs the club’s analytical efforts in support of on field team performance as well as the used of data and analytics in player identification and assessment.

The purpose of this project is to provide Orlando City Soccer Club with a tool that quantifies player quality and value and then recommends players for the club to sign based on the club’s financial and roster limitations.

# Project Objectives

1. Build a machine learning model to predict player ability.
2. Build a combinatorial optimization algorithm to optimize player ability.
3. Build an R Shiny application to support UI integration of the optimization algorithm.
4. Integrate the machine learning model’s output with the optimization algorithm.
5. Expand the R Shiny UI to pass more detailed parameters to the optimization algorithm.

# Project Data

All data used in this project will be limited to players in the MLS. The first data source used is player salary data. There are multiple datasets blended together that capture the salaries of all players in the MLS for a given season. The data is from Fbref and American Soccer Analysis. The next data source is aggregated performance data from StatsBomb. This dataset contains detailed measurements on a player’s performance aggregated for the whole season. Additionally, this project will use a third dataset to speed development as the optimization algorithm requires a metric that will not yet exist at the time of development. That dataset contains a metric called “Goals Added”. This metric will be the value the algorithm optimizes at first. Once the predictive model is complete it will produce a value that will replace Goals Added as the value in the objective function for the optimization algorithm.

# Computational Tools

Rather than using a large amount of information, this project focuses more on the intelligent use of information to inform decision making. In total there will be less than 10,000 rows of data used with some records used for different purposes than others. This will not require any special set of tools or more advanced hardware than what is available to someone’s personal laptop.

# Data Science Concepts

The core concepts are prescriptive and predictive models of which the former will depend on the latter. The prescriptive model will use a genetic algorithm to optimize an objective function. The value to be optimized by the optimization algorithm must come from a predictive model. Within the optimization algorithm there is also a need for a linear programming model to consider all the constraints on the client. Additionally these three concepts will depend on a usable application that ingests and feeds the data into the three models described above.

# Description of Final Document

1. **Introduction** 
   1. Introduction
   2. Problem Description
   3. Value Added by the Solution
2. **Methodology**
   1. Model Design
      1. Optimization Model
      2. Predictive Model
   2. Data Analysis
   3. Conclusion
3. **Analysis of Results**
   1. Findings
   2. Results
   3. Conclusion
4. **Summary and Conclusion**
   1. Summary of Findings
   2. Suggestions for Continued Research and Improvement
   3. Conclusion

|  |  |  |
| --- | --- | --- |
| Project Timeline with Activities **Bold items are deliverables required as a part of the capstone project**  Non-bold items are work that will be available at that time in the project | | |
| **Week Of** | **Summary of Work** | **Deliverables** |
| Jan 30 – Feb 5 | * Gathering data * Analysis of optimization algorithms and other approaches * Project Proposal | * **Project Proposal** |
| Feb 6 – 12 | * Working mockup of optimization algorithm with American Soccer Analysis “Goals Added” dataset (test dataset) * Data cleansing to combine Goals Added and Player Salary datasets | * .R file with optimization model |
| Feb 13 – 19 | * R Shiny UI that displays optimization algorithm output * Begin development of the predictive model to find player quality * Assessing different predictive models | * Simple R Shiny application with basic functionality and output |
| Feb 20 – 26 | * Basic working predictive model * Begin integrating the predictive model into the optimization algorithm | * **Project Update 1** |
| Feb 27 – Mar 5 | * Tuning and further evaluating the predictive model * Improve functionality of R Shiny UI | * Improved R Shiny UI * .R file with predictive model |
| Mar 6 – 12 | * Finalize predictive model * Begin analysis on linear programming model to make the optimization model more dynamic | * **Project Update 2** |
| Mar 13 – 19 | * Improve linear programming model to consider the club’s financial constraints | * .R file with linear programming model |
| Mar 20 – 26 | * Integrate the linear model with the full optimization model * Continue to improve linear model, optimization model, and predictive model based on new information, the quality of results, or necessary technical changes * Document findings from model development |  |
| Mar 27 – Apr 2 | * Improve R Shiny UI functionality to allow integration of data sources | * **Project Update 3** |
| Apr 3 – 9 | * Reevaluate the three models to determine how to improve |  |
| Apr 10 – 16 | * Improve accuracy of predictive model, and the flexibility of the optimization model and linear programming model * Document findings and functionality of the models | * **Project Update 4** |
| Apr 17 – 23 | * Improve R Shiny UI to be more flexible and user friendly * Prepare final paper and documentation |  |
| Apr 24 – 30 | * Finalize documentation | * **Completed capstone project** |