ISYE 6740 - Summer 2024
Project Proposal

Group Number: 014

Team Member Names:

- Ty Underwood
- Ting Jennings
- Simon Liu

Project Title:

Predicting Georgia Tech Football Success Using Machine Learning Algorithms

Problem Statement:

In today's day and age, collegiate athletics are watched and enjoyed by millions of Americans. With that being said, it has become big business with product marketing, television viewership rights, and many other aspects. Most importantly, at the heart of this business are the fans. People feel a connection with their respective college teams, leading them to buy more merchandise, purchase tv channel packages, buy tickets to the games in person, and much more. A leading factor that fans follow is the performance of their team. Many fans ask themselves "I wonder how well we are going to do this next season?"

Well this is the goal we are attempting to answer. In this project we are going to use historical data from the Georgia Tech football team to try to predict the outcomes of the Yellowjackets ACC in-conference games for the upcoming 2024 season. By doing this analysis, it could incite excitement (or disappointment) for the fans for the upcoming season, and for people who work on the business side of sport it can help give an idea of what to expect for certain markets. Leading to more informed business decisions.

Data Source:

As far as the data source for this project goes, we have obtained game level data for the Georgia Tech football team spanning from the 2000 season to the most recent season (2023). This data is published by "www.sports-reference.com" and consists of 304 data points with 46 features, note each data point is one game played and the 46 features holds ID values, offensive statistics, and defensive statistics.

Methodology:

Data Cleaning: (owner: Ty Underwood)

The first thing that will need to be done is to perform some sort of "Extract, Transform, Load" (ETL) method to get the raw data into a format ready for modeling. This will entail loading in the data from its raw sources, transforming it by the means of feature analysis and standardization. Note, this feature analysis can come by statistical methods and also by expert experience on the topic. Finally, once the data is set in its final form we can load it into a usable form for the modeling. Basically, following the "Extract, Transform, Load" term pretty closely.

Classification: (owner: Simion Liu)

After cleaning the data performing some exploratory analysis, our next step will be to train and test different classification models to help aid us in our ultimate goal of predicting the 2024 in-conference games for Georgia Tech football. Some of these models include, but are not limited to: Logistic Regression, K-Nearest Neighbors, and others. Note, in this step of the process we will also explore variable selection methods to help us cut down on unnecessary features if needed. Once we have the variables we want to keep, we will follow the standard procedure of splitting the data into training and testing sets (80% train, 20% test) then evaluate our models and select one from there.

Clustering: (owner: Ting Jennings)

After the classification modeling selection process is complete, our team will apply a clustering method to help us obtain inputs for the model.

First we need to select our clustering method, for this, we are going to go with the K-Means clustering algorithm. In this case we are not wanting to use the response variable as a testing metric since we are only wanting to find games that are similar, regardless of who the opponent is or what the outcome of the game was. Overall, we are taking an unsupervised approach to this step.

Next, after getting our clusters, we want to look at the target opponent for the 2024 season and find the cluster that has that opponent the most amount of times. This will be our identified cluster for the particular target opponent. Note, for the clustering step of our project, we will be removing the outcome of the game (response variable) as well as the opponent id variable from our analysis feature set so that it does not just cluster the same teams together based on id, but focus more on the game statistics. We then will map back the opponent ids once the clustering is done, thus allowing us to find the cluster with the most instances of that particular target opponent.

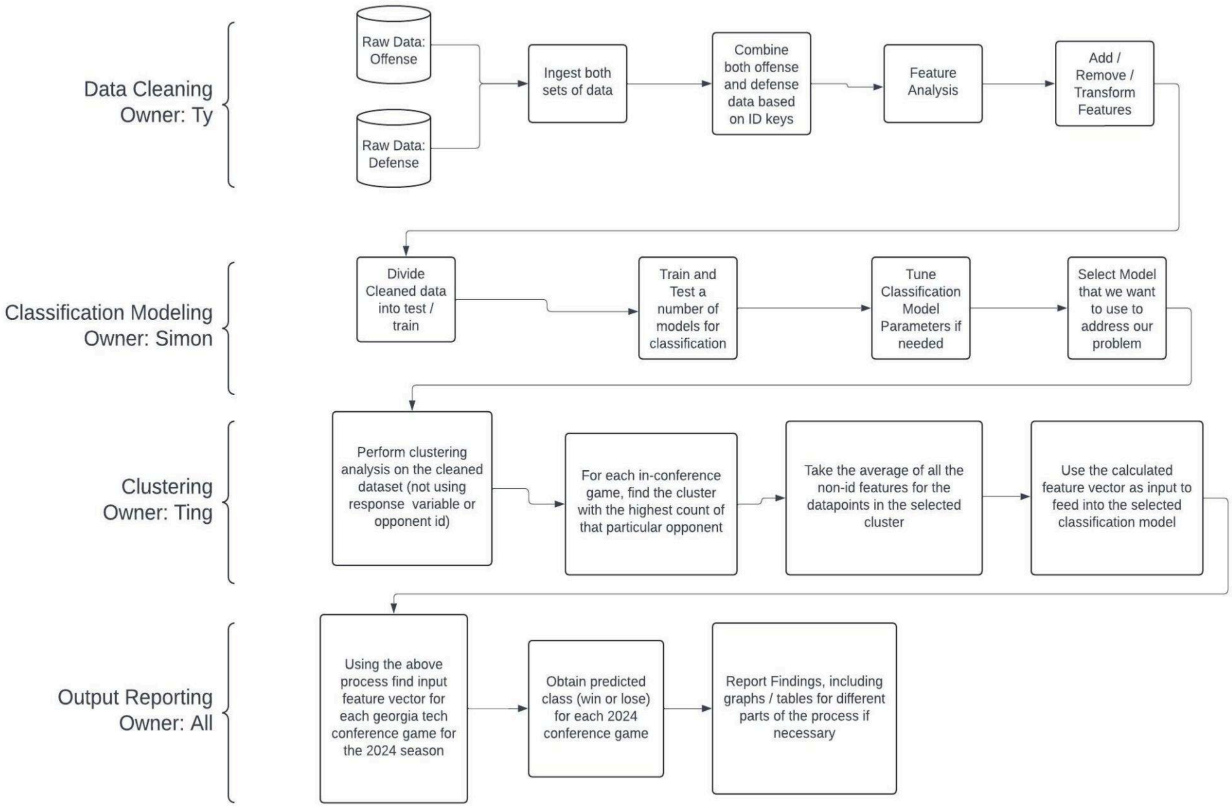
Finally, after identifying our target cluster, we will take the averages of the different features for that cluster and use those as our input data point to feed into the classification model. The idea behind this is to give us the best educated guess on the specifics of how Georgia Tech plays a particular opponent and similar teams to our target opponent. This is also because we do not have data on a game that has not been played yet.

After getting our input data point, we will repeat the identification and averaging process for each opponent on Georgia Tech’s 2024 in-conference schedule.

Evaluation and Final Results:

After going through the full process listed above, we will take the input data points then feed it into the trained classification model to obtain our final outcome predictions. It will either be a win or loss for each game in the upcoming in-conference schedule. As far as evaluation, we do not have true outcome labels for our predicted games. This makes sense because the games have not been played yet, so there is no way to test accuracy on that. However, we can evaluate and test our classification models based on the historical data we do have true labels for. The prediction evaluation for our final output can be calculated once the 2024 season is over and we can see the accuracy of our predictions.

Process Flowchart:



Data Cleaning
Owner: Ty
Raw Data: Offense
Raw Data: Defense
Ingest both sets of data
Combine both offense and defense data based on ID keys
Feature Analysis
Add / Remove / Transform Features
Classification Modeling
Owner: Simon
Divide Cleaned data into test / train
Train and Test a number of models for classification
Tune Classification Model Parameters if needed
Select Model that we want to use to address our problem
Clustering
Owner: Ting
Perform clustering analysis on the cleaned dataset (not using response variable or opponent id)
For each in-conference game, find the cluster with the highest count of that particular opponent
Take the average of all the non-id features for the datapoints in the selected cluster
Use the calculated feature vector as input to feed into the selected classification model
Output Reporting
Owner: All
Using the above process find input feature vector for each georgia tech conference game for the 2024 season
Obtain predicted class (win or lose) for each 2024 conference game
Report Findings, including graphs / tables for different parts of the process if necessary