SIADS 593: Milestone I
Team Project Proposal

version 2022.07.27.1.CT

Instructions: please make a ***copy*** of this template file (do not edit original).

Proposal Title: High Impact Fantasy Football Decisions

1. Team members

Please list your team members (2-3 max).

- Tristan Morgan (tristmor)
- David Brand (dbbrand)

2. Project summary

Summarize your proposed project in a few sentences.

- What is your proposed project and why are you proposing it?
- What are the question(s) you want to answer, or goal you want to achieve?

Project Title:
High Impact Fantasy Football Decisions

Proposed Project and Rationale:
This project seeks to develop a data-driven framework for optimizing high-impact decisions in Fantasy Football. Most players rely on intuitive and football knowledge, this project will attempt to achieve a quantitative approach to making Fantasy Football roster decisions. Factors such as draft strategies, waiver wire transactions, trade evaluation, and injuries will all be analyzed to see what correlates with success.

Project Goals and Questions:

- Decision Impact: Which phase of management shows the highest correlation with success? Is it the draft, trades, waiver wire pickups, or lineup decisions?
- Predictive Modeling: Can play by play data, basic Fantasy Football stats, and injury data be used to improve starting lineup

3. Datasets

Describe one primary dataset and at least one secondary dataset. If other secondary datasets will be used please describe them as well.

The proposed datasets should exhibit different features/columns and/or different access methods, e.g., *.csv file, *.json file, API retrieval, web scraping, etc. Different time periods, for example, with the same features/columns is not considered a different dataset. Remember, the focus of the project in this Milestone course is to give you the opportunity to practice your data manipulation skills, so feel free to challenge yourself.

If you're unsure if your data sets are "different enough" describe the datasets and request a review via the *#siads593_[semester]_001_project* Slack channel.

Please note: all proposed datasets ***<u>MUST</u>*** be publicly available to all members of the class (students, instructors, course support personnel, etc.). Use of proprietary datasets for this project is ***not*** permitted.

3.1 Primary dataset description

Describe your primary dataset. How is the data collected and how will you access it? Please share what features in the dataset are relevant to your topic. At a minimum, include the following information:

- Short description (i.e., 1-3 sentences) of its key features
- Estimated size (in records and/or bytes)
- Location (give the URL or other access method)
- Format (CSV, JSON, etc.)
- Access method (download, web scraping, API, etc.)

Primary Dataset:
ESPN Free API

Description:
This dataset contains Fantasy Football information and data on Fantasy Football leagues. Public leagues can be used to gather decisions that players made throughout the season. Decisions such as draft choices, Football player fantasy stats, waiver transactions, etc. can be pulled via this API

Estimated Size:
Size can vary greatly by query parameters passed into the API endpoint. Currently we have been able to acquire 501 records of league ids.

Format:
JSON Response Object

Access Method:
API

3.2 Secondary dataset(s) description

Describe your secondary dataset(s). How is the data collected and how will you access it? Please share what features in the dataset(s) are relevant to your topic and describe the data types you're expecting. At a minimum, for each secondary dataset include the following information:

- Short description (i.e., 1-3 sentences) of its key features
- Estimated size (in records and/or bytes)
- Location (give the URL or other access method)
- Format (CSV, JSON, etc.)
- Access method (download, web scraping, API, etc.)

Secondary Dataset:
Play-by-Play and Injury Data

Short Description:
Play-by-Play data contains information on each play of each game. Statistics like game data, minute, second, offensive team, down to go, yard line, description, etc. are all contained in this play-by-play data. Injury data contains data like player last name, game status, injury description, etc.

Estimated Size:
- Each year of Play-by-Play Data contains approximately 10-12MBs
- Injury Data contains 6264 records

Location:
- Play-by-Play Data
- Injury Data

Format:
Both are csv.

Access Method:
- Play-by-Play Data is accessed via web download
- Injury Data is accessed via a python package that has an object method to import injury data as a pandas dataframe which can be written as a csv.

3.3 [YES] Affirm: datasets are public.

Please write YES in the above box to confirm that your primary and secondary datasets are accessible and available to your classmates and the instructional team.

4. Cleaning and manipulation

Describe how you will need to manipulate your datasets: how will you handle missing or anomalous data? How will you join your primary and secondary datasets? What cleaning and manipulation challenges, if any, do you anticipate?

To prepare the datasets for analysis, we will clean and manipulate data from multiple sources, each of which presents distinct challenges. The primary Fantasy Football dataset is derived from public fantasy football leagues, which may include extreme outliers such as non-competitive or atypical leagues. Play-by-play data contains a wide variety of data types and semi-manual entries that may be inconsistent or noisy. Additionally, the injury dataset references players using a naming and identifier system that differs from the primary fantasy football data. Addressing these issues will require a combination of aggregation, filtering, normalization, and cleaning techniques.

Missing values will be handled using informed strategies appropriate to each context. In some cases, missing data may reflect meaningful conditions, while in others it may indicate incomplete or delayed reporting. Where feasible, missing values will be investigated and supplemented using secondary data sources; otherwise, affected observations may be excluded if they compromise the analysis. We do not expect to rely heavily on data imputation, though imputation may be necessary if the data problems are systematic and impact key variables. Anomalous values will be identified using statistical summary analysis, and extreme or implausible observations will likely be removed from the dataset.

The primary Fantasy Football data will be joined with play-by-play and injury datasets using common identifiers such as player IDs, team identifiers, and game dates. Entries like player identification and datetime values will need to be carefully aligned across databases or it could invalidate the analysis. Additionally, the different datatypes from each dataset could require additional manipulation like normalization or one-hot encoding prior to analysis.

These challenges will be addressed through multiple rounds of validation, exploratory checks, and data manipulation.

5. Analysis

Describe any analyses you plan to undertake. For each, please give the technique or approach and briefly explain what you expect to learn from it.

To begin the analysis, we aim to identify which aspects of Fantasy Football decision-making present the greatest opportunity to improve performance. In Fantasy Football, there are four primary decision categories: draft decisions, trade decisions, waiver wire decisions, and start/sit decisions.

To compare the relative value of these decision types, we define a neutral baseline for each category and estimate the expected fantasy points associated with that baseline. We then analyze teams whose decisions deviate from these baselines and measure the resulting fantasy points relative to expectation.

By expressing each decision category in terms of “points above expected”, we place all decision types on a common scale. This enables direct comparison across decision categories and allows us to evaluate which types of decisions are most impactful to overall Fantasy Football success. To accomplish these goals, we will perform data cleaning, aggregation, and comparative analysis with mixed data types and data sources.

For further analysis, we will incorporate basic Fantasy Football data, play-by-play data, and injury data to better understand and improve high-impact Fantasy Football decisions. We will begin with multivariate exploratory analysis to identify strong correlations, detect outliers, and uncover meaningful statistical patterns within the data. In this phase, we will consider more complex analysis techniques including temporal analysis, categorical encoding strategies, and dimension reduction techniques, as appropriate.

This analysis will focus specifically on relationships associated with the most successful decisions, as measured by points above expected. For example, the timing of waiver wire transactions may exhibit a strong correlation with high-value waiver wire outcomes.

6. Visualizations

Describe in 1-3 sentences at least **two** data visualizations that you plan to create. Include the chart type (e.g. bar chart, scatterplot, SPLOM, etc.) as well as the variables (features) you intend to plot.

The first visualization will be overlaid probability distribution curves for each of the major decision categories. This will effectively demonstrate the average value and variance of each decision.

The second visualization will be an SPLOM that shows the pairwise relationships between a variety of basic fantasy football variables all at once. For example, the following variables presented as a SPLOM may be useful for improving Start/Sit decisions: Injury status, target share/rush share, snap share, week, player weight, yards/attempt, etc. Similarly, we will generate a correlation matrix that compares points above expected with a variety of basic fantasy football data. Given the multivariate nature of these charts, we expect to use these charts to explore many relationships quickly.

The final visualization will be a line chart that shows changes from week to week through the NFL season. This will demonstrate effective temporal analysis, and could include simple expressions of uncertainty (if appropriate). For example, for each decision category, we could plot the cumulative points added starting at the time that the decision was made in the fantasy football season. This would demonstrate the value of a timely decision in Fantasy Football.

7. Ethical considerations

Does your choice of data raise any ethical issues? If so, briefly describe the concern and how you plan to mitigate it.

Data Privacy and Anonymization:
Although we are gathering information from public Fantasy Football league's, manager and team names will be obfuscated in the analysis phase due to privacy concerns. No personally identifiable information of non-professional football players will be shared in the final report.

Responsible Use and Gambling Implications:
While this project aims to improve the decision making of Fantasy Football players, some players will be using Fantasy Football to gamble. To address this ethical concern, this report will include a disclaimer that these models are for academic and recreational purposes only. This model should not be used to aid in gambling.

8. Contributions

Indicate the contribution that each team member will make to the project.

Tristan Morgan:

- Data Collection: Download the injury dataset and aid in API calls
- Cleaning and Manipulation: Handle joining the injury dataset and API response data
- Analysis: Lead the Injury Impact Analysis and aid in key decision drivers (e.g. draft, waiver wire, etc.)
- Visualization: Create visualizations based on the analysis (e.g. visualizations to explore correlation of injury details, player position, and effect on player output)

David Brand:

- Data Collection: Download the Play-by-Play data and aid in API calls
- Cleaning and Manipulation: Aggregate Play-by-Play

- Analysis: Lead the Play-by-Play Analysis and aid in determining key decision value (e.g. draft, waiver wire, etc.)
- Visualization: Create visualizations based on the Play-by-Play analysis (e.g. touches inside redzone vs expected points added). Create visualizations to express cumulative effects of decisions over time.

Changelog

(2022.07.27.1.CT) Update for 593

(2021.07.24.1.AW) Adjust title, number sections, simplify section headings, edit text