- The project's **domain background** — the field of research where the project is derived;

  This project centers on evaluating rookie wide receivers entering the NFL and whether or not they are likely to become productive NFL starters for fantasy football in the ensuing years. Specifically, it relates to dynasty fantasy football where players draft incoming rookies similar to an NFL franchise and then get to keep the rights to the players in successive seasons of fantasy football until they decide to trade the rights.

  Here's a link to a similar project that investigates machine learning being applied to the NFL Draft https://mospace.umsystem.edu/xmlui/handle/10355/47027

- A **problem statement** — a problem being investigated for which a solution will be defined;

  Problem: Identifying productive young wide receivers in fantasy football is difficult yet rewarding in the context of the game for a variety of reasons. My study will use data surrounding their past success in college (college statistics), athletic profile (NFL combine data) and their position in both the NFL draft and fantasy drafts to determine if they will have a top 24 fantasy wide receiver finish within 5 years of entering the NFL. This will be a Boolean indicator that can be used as a dependent in a series of machine learning models. This is very important information to know in the dynasty fantasy football community. Since wide receivers are increasingly the core of many dynasty fantasy football teams and they have long careers, optimizing player selection can have very positive impacts on your team for years to come.

- The **datasets and inputs** — data or inputs being used for the problem;

  Datasets for the problem are difficult to compile and require various source material:

  - College Football Reference Receiving Statistics
    - Contains receiving data for college WRs which details success in college
  - Dynasty League Football (DLF) Dynasty Avg. Draft Position Data
    - Represents the average rank dynasty football participants assigned to each player. As an example, the player most consistently being drafted at the #1 overall spot is considered the player with the most promising career
  - ESPN NFL Receiving Statistics
    - This is for the dependent. This will measure their success after selection and is the measure the dynasty fantasy football community most careas about
  - Pro Football Focus Elite College Receiving Metrics
    - These are advanced statistical metrics on performance which help supplement the college statistical data
  - Pro Football Reference Combine Data
    - The NFL combine brings all incoming NFL players together to create standardized athletic tests measuring player speed, strength, agility and other measurables. In addition, this data also contains the player's

placement in the NFL draft which provides information on what professional evaluators think of the player (and differs from fantasy draft placement)

In this case they are often only available as CSV files which require heavy manipulation which I will do in the assignment after uploading them as pandas dataframes.

Dimensionality of the Problem: There will be ~30 features included in the machine learning intake after some pre-processing.

# of Datapoints: There will be ~180 data points from which to work with prior to upsampling the data.

Balance of the Dataset: Roughly 1/3 of the target variables were successes and 2/3 failures. As a result, I will be using AUC as a metric and also upsampling the successes given the unbalanced classes.

Train/test: I split the data into 60% training 40% testing sets for most model development. During the validation phase I also tested the three most promising models from the first step by taking 200 randomly generated samples with a 60% training and 40% testing split and took the mean AUC for all of the models over that timeframe. The last step during validation was splitting the data into older and more recent observations so I could test if there were severe drop offs in recent data.

- A **solution statement** — the solution proposed for the problem given;

  The solution would be whether or not a WR has a top 24 wide receiver fantasy season within five years of entering the NFL (binary). In order to increase the sample, players who haven't reached this status within three years of entering the league are assumed to be a miss until subsequent model update them as dependents.

  Corner cases due to injury: Injury is a natural part of the NFL landscape but I do not make any explicit provisions for injured players during the forward-looking window. The main reason is injury comes in many forms – some season ending and some not. Drawing a line of demarcation and what would extend the window could be difficult. Further, 5 years is a pretty forgiving runway for success and allows opportunity to produce if a player had 1-2 years of injury history.

- A **benchmark model** — some simple or historical model or result to compare the defined solution to;

  The benchmark would be to use the Dynasty League Football Avg Draft Position as feature data in a logistic regression as the initial benchmark. It is a "wisdom of the crowd" type dataset that reflects fantasy football participant opinions on how effective wide receivers will be in high stakes fantasy football leagues. The lower the rank, the more promising the player as a prospect by the community.

- A set of **evaluation metrics** — functional representations for how the solution can be measured;

  The evaluation metric in this case would be the AUC of a series of models in predicting the binary likelihood a WR hits the binary target. Given that the historical success rate is unbalanced, I will upsample the minority class and in addition use AUC. The AUC for the model solving the problem would have to be much higher than the logistic regression model trained on NFL and fantasy draft data.

- An outline of the **project design** — how the solution will be developed and results obtained.

  The project design will be as follows:

  - Data Upload & Pre-Processing
    - Upload all the CSV data
    - Create a series of scaled metrics using the raw data
    - Take that data and scale it for input into a machine learning model using a min-max scalar
    - Impute missing data
  - Create an assessment of the initial AUC using the original benchmark logistic regression model
  - Run a series of base classifier models (decision tree classifier, bagging classifier, random forest classifier, gradient boosting classifier) to assess initial predictiveness
  - Tune some of the more promising classifier models along some of the major characteristics
  - Validate the data by applying the data to other randomly sampled datasets and through time
  - Visualize a base decision tree to gather insights into the important feature variables