Introduction/Background:

Fantasy football is a popular game played among friends who draft current NFL players to form the perfect team. The fantasy scores of each player on the team are calculated by the players' stats (receptions, yards, touchdowns, etc.) being accumulated. The sum of all of a member's players is then compared to the person they are up against and whoever has the higher score is the winner for the week. For our project, we want to use different machine learning techniques to predict how well a player does in the season based on their last year's fantasy points, matchups, usage, etc.

A lot of research shows what specific elements in a player's game that can affect their fantasy points. According to research [1], a player's fantasy points can be predicted using basic statistics like the number of passing yards, touchdowns, yards, and number of receptions/touches, but also takes in other factors like age, injuries, and how long the player has been on the team. To employ these methods, we can use various machine learning techniques. Some research [3] suggests that we can use the k-nearest numbers method, which essentially has us calculate the Euclidean distance between each player's statistics vs other players and the prediction is made by comparing the averages of all of the players in the position. Other research [2] shows that another method is using random forests and clustering decision trees together to predict the best player by position.

The dataset consists of all current and past NFL player statistics ranging from previous years fantasy points, toe touches and yards for specific games to age. It has every factor that would be useful to create this algorithm.

Dataset: https://www.pro-football-reference.com/

Gantt Chart: ML_Project_2024_GanttChart

Problem Definition:

Problem:

Predicting an NFL player's performance in fantasy football is hard because of the large number of factors that can affect it. A lot of prediction models now are inaccurate because they either don't take into account all the variables that are at play or they are outdated, which can lead to less optimal teams for participants.

Motivation:

We really want to create an algorithm that takes both useful statistical factors like yards, receptions/touches, touchdowns, but also external factors that may be at play like player age and injury history. By doing this, participants in the league would be able to make better decisions when drafting their fantasy team.

Methods:

For our first model, we decided to implement a random forest model for the purpose of predicting fantasy score based on a number of factors. Specifically, a player's fantasy points can be estimated using fundamental stats like passing yards, touchdowns, total yards, and the number of receptions or touches. Our model gives an even more accurate estimate, taking into account extra information as our data also considers additional factors such as age and injury history.

This model combines multiple decision trees to give one more accurate and stable prediction. Each individual tree in the random forest is built in a random subset of the features and data samples. Finally we get an output of the random forest that is determined by aggregation over all the trees. A benefit of random forests over other models is that it includes non-linearity. Random forests can capture complex, non-linear relationships between features, which a model like linear regression cannot since linear regression does assume a linear relationship between inputs and outputs.

Results and Discussion:

We have completed our models and are able to output the fantasy score of any player input by the user that is in our dataset. We have taken into account multiple years of data for maximum accuracy in the model's prediction.

Visualizations: https://docs.google.com/document/d/1cd3KoXcc0ZHYNbuG7yAc_fCRxlVIvu7Bb0ruMVY_YAQ/edit?usp=sharing

We simulated our results in the chart shown above and our results showed high accuracy in accordance to past data, validating our model's success. Random forest models are generally very accurate and robust regardless because they average the result of many trees, which significantly decreases over-fitting compared to individual decision trees. With our Random Forest model, we are able to incorporate statistics other than touchdowns and yards, especially more sophisticated variables such as the ones we mentioned above and provide the successful results.

References

- [1] Abadzic, Alan; Cheun, Jacquelyn; and Patel, Milan (2024) "Data Analysis on Predicting the Top 12 Fantasy Football Players by Position," *SMU Data Science Review*: Vol. 8: No. 2, Article 7. Available at: https://scholar.smu.edu/datasciencereview/vol8/iss2/7
- [2] Robinson, Chelsea, "The Prediction of Fantasy Football" (2020). *Mathematics Senior Capstone Papers*. 20. https://digitalcommons.latech.edu/mathematics-senior-capstone-papers/20
- [3] N. S. Parikh, "Interactive tools for fantasy football analytics and predictions using machine learning," M.Eng. thesis, Dept. Elect. Eng. and Comput. Sci., Massachusetts Institute of Technology, Cambridge, MA, 2015. [Online]. Available: http://hdl.handle.net/1721.1/100687

Contribution Table:

Name	Proposal Contributions	Codebase Contributions	
Tanay Dulipala	Report: Intro/Literature Review, Problem Definition, Gantt Chart, References	Final: KNN Model Training, Linear Regression Model Training	Model evaluation
Varshith Chekka	Report: Methods, Results and Discussion	Final: Random Forest testing, Presentation, Analysis	GUI implementation
Fawaz Sabir	Report: Methods, Results and Discussion	Final: KNN testing, Presentation, Visualizations	GUI implementation
Rithvik Pandibabu	Report: Problem, Results & Discussion	Final: Preprocessing for Both models, Linear Regression Implementation, Gantt Chart	Load and transform data, output MSE and R^2 values
Rishi Chinnapareddy	Report: Visualizations, Problem Definition	Final: Preprocessing for Both models, KNN Implementation, Model Comparisons, Quantative Metrics	Load and transform data, output MSE and R^2 values