CS216 Project Final Report: Machine Learning for Fantasy Football

Eddie Kim, Kevin Shah, Kush Gulati, Rob Haywood

Part 1: Introduction and Research Questions

Nothing quite stimulates the average sports fan like the opportunity to actively engage and feel connected to every touchdown, dunk, and home-run from their professional sports players. The numbers speak for themselves: as fantasy sports has developed into a competitive industry from a unique activity, a 2017 survey by the Fantasy Sports & Gaming Association indicated that some 59.3 million players participated in some fantasy sports league.

So what is fantasy sports? Using the Internet and access to lots of data, participants build virtual teams of professional sports players, and these teams compete based on the individual performance of these players. Points are collected and verified by computer servers that track the results and statistical performances of players in actual games. Fantasy football is the most played fantasy sport, where participants draft from the pool of all National Football League (NFL) players. Our team is particularly interested in predicting player performance given several statistical predictors of player performance, including rushing yards, touchdowns, passing yards, position, etc. One way of doing this is taking a train set of the player data from the 2021-2022 NFL season and predicting on a test set of the player data, and comparing our predictions to the actual scores. Another point of inquiry is subsetting the players by position to create individual models for quarterbacks, running backs, and wide receivers, the three most important positions for fantasy football teams. The third point of inquiry that could be particularly interesting is comparing our algorithm's predictions to the projections given by expert fantasy football sources at NFL.com, ESPN, and Yahoo Fantasy.

The research questions that we have set out to answer are: 1) how accurately can one predict the fantasy football performance of NFL players using machine learning, 2) what correlations can be drawn between variables (or interactions of variables) and the fantasy football performance of NFL players. We plan to use our machine learning model and target predictions as a backend for an interactive web app where one could search up an NFL player that they are interested in and visualize their current and projected fantasy statistics.

The topic we have chosen is substantial because it involves utilizing a large dataset to produce robust and meaningful predictions, which can help sports analytics tools understand which features are most important to influence a predictive model. Our topic is feasible because our plan to scrape data, complete exploratory data analysis, develop a machine learning model, and connect our data and predictions to an interactive web app is well possible by our team within six weeks. Finally, it is relevant due to the huge growing market size for fantasy sports, as players are looking deeper for advanced analytics tools to predict player performance. This makes machine learning a huge demand for the industry, and after doing some market research, we have found that supply for ML solutions in this industry is quite limited.

Part 2: Data Sources

The most inclusive and robust dataset we have found to pursue our research question comes from Yahoo's Fantasy Football webpage (football.fantasysports.yahoo.com). Yahoo Sports pages give access to several parameters and research tools for not only the average fantasy football player, but also those who are interested in utilizing analytics to augment their fantasy football experience. When compared to the research tools and data available at ESPN and Nfl.com, Yahoo Fantasy Football tools have more in-depth league history and record books and more unique weekly player write-ups and recaps. After doing some competitive analysis of the data available from several sources, our team decided on implementing the top 300 players in Yahoo Fantasy's 2020-2021 NFL player data, because the majority of leagues will not draft players past the top 300 players. The data that we collected is appropriate and sufficient to address our research questions because it has enough rows (players) to split into a valid train

and test set, and it also has enough columns (parameters) to create robust predictions. Parameters such as percentage of fantasy owners, pre-season rankings and actual rankings go beyond traditional on-field statistics which make this data especially beneficial.

One thing we want to note about our dataset is as of last week, the source that we are using was removed from Yahoo Fantasy's servers as the 2021-2022 fantasy football season has officially ended. Fortunately, we saved the initial dataset to our Github folder so it is still available there. Basically, we cannot run our scraper anymore because the dataset we pulled previously is wiped from Yahoo's servers. Our scraper did a majority of the work with pulling and preparing the data, which we will explain more in Part 3.

Part 3: Results and Methods

To follow along with our results and methods, you can access our full implementation at our Github repository: https://github.com/kg227-dev/CS216-Project, which contains a README with a link to our demo. Below are some of the methods we implemented to pull and prepare the data for an initial exploratory data analysis and predictive model:

Pulling the Data

1. To extract the data, Kush developed a web scraper using the BeautifulSoup Python package called "Scraper.ipynb".
2. The scraper sends a URL request to the Yahoo login page, where one can authenticate with their own login and password to their Yahoo account to access the Fantasy Football data.
3. To get the parameters we were interested in, Kush used BeautifulSoup to find the HTML class holding information to the data table and iterated through the stats in each row.
4. The data was written to an outfile which we designated as "ff_data21.csv".

Preparing the Data

1. Kush manually added column headers to the data because the scraper did not automatically add those to the dataset.
2. In addition, Kush added dummy variables to the dataset for player position, which will be especially useful when developing models for individual positions.
3. After making these updates, the data was written to an outfile csv which we designated as "cleaned_ff_data21.csv".
4. After doing some initial exploratory data analysis, Kush plotted the distribution of fantasy points scored by 2021-2022 NFL players using the Plotly package, as shown in Figure 1. In addition, he plotted the correlations between the parameters in the dataset, as shown in Figure 2. These plots, along with some others that the team is working on, are available in "Data Viz.ipynb".

Building the Initial Predictive Model

1. To build the model, Kush started by splitting the data into a 90:10 train and test set.
2. He then implemented Elastic net regularization into his regression, which linearly combines the L_1 and L_2 penalties of the lasso and ridge methods of regularization to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting.
3. After running AlphaSelection to find the optimal alpha value that minimizes the loss function, ElasticNet regression was run on the train data and the parameters are shown in Figure 3.
4. Next the predictions on the test data and their actual values were plotted side by side in Figure 4. Our model's code can be located under "Initial Model.ipynb".

Building the Web Application

1. Rob began by creating a Django project and styling two pages with simple dummy data: a home page from which one can search for any player in the DB and a results page to display their data.
2. Next, he grabbed an additional command-line tool from sqlite.org which he used to import “ff_data21.csv” into a SQLite3 database. He then removed the dummy data and generated links for each player which display basic data like team and position.
3. Using RegEx and JavaScript, Rob built a search bar which uses autocomplete to determine which player a user might be searching for as they type.
4. Rob then styled the site and merged his repo into the one containing the ML model.
5. Then, he created another scraper which gets logos and background images to display on players’ pages, and a second scraper to get a headshot for each player from nfl.com.
6. After that, he migrated the database to PostgreSQL for easy deployment to Heroku. Within this database, he implemented “verbose names” which are user-friendly versions of field names that can get grabbed with the field names, and added a demo of the data we want to display with their verbose names to players’ pages.
7. Finally, he deployed the demo to this page

Visualizing the Data

1. Eddie and Kevin started by developing visualizations that show our model’s efficiency and display the significance of our results.
2. Eddie created a visualization to display our model’s performance by plotting two lines (one for predicted points and one for actual points) on a graph with player rankings on the X axis and points on the Y axis, as shown in Figure 5. This graph shows the strong accuracy of our model in predicting points across all players in the database.
3. Eddie also created a violin plot, shown in Figure 7, that shows the distribution of points per position in a visually functional way. This visualization helps convey context for different player positions and how it affects point distribution.
4. Kevin created visualizations, like in Figure 6, using matplotlib. In addition to showing the significance of our model’s results, the visualizations aid the reader in comprehending the preliminary data. This allows for people with no prior NFL knowledge to understand fantasy football trends.

Figure 1. Distribution of Fantasy Points Scored by 2021-2022 NFL Players

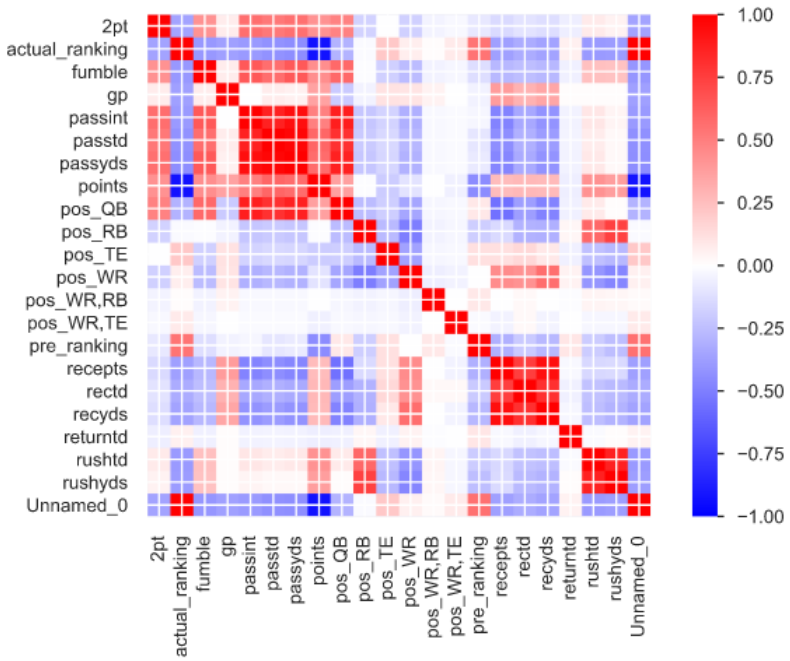

Figure 2. Correlations between Parameters

	Variables	Coefficients
0	gp	2.269916
1	pre_ranking	-0.918164
2	actual_ranking	-17.405979
3	passyds	20.692882
4	passtd	31.355824
5	passint	1.640106
6	rushyds	16.937395
7	rushtd	15.176396
8	recepts	12.181528
9	recyds	19.022774
10	rectd	14.169886
11	returntd	0.335356
12	2pt	2.376742
13	fumble	-0.651986
14	pos_QB	0.444243
15	pos_RB	-0.221098
16	pos_TE	-0.702121
17	pos_WR	0.237806
18	pos_WR,RB	0.136373
19	pos_WR,TE	-0.008266

```
Intercept: 127.23155555555554
```

Figure 3. Initial Model Coefficients

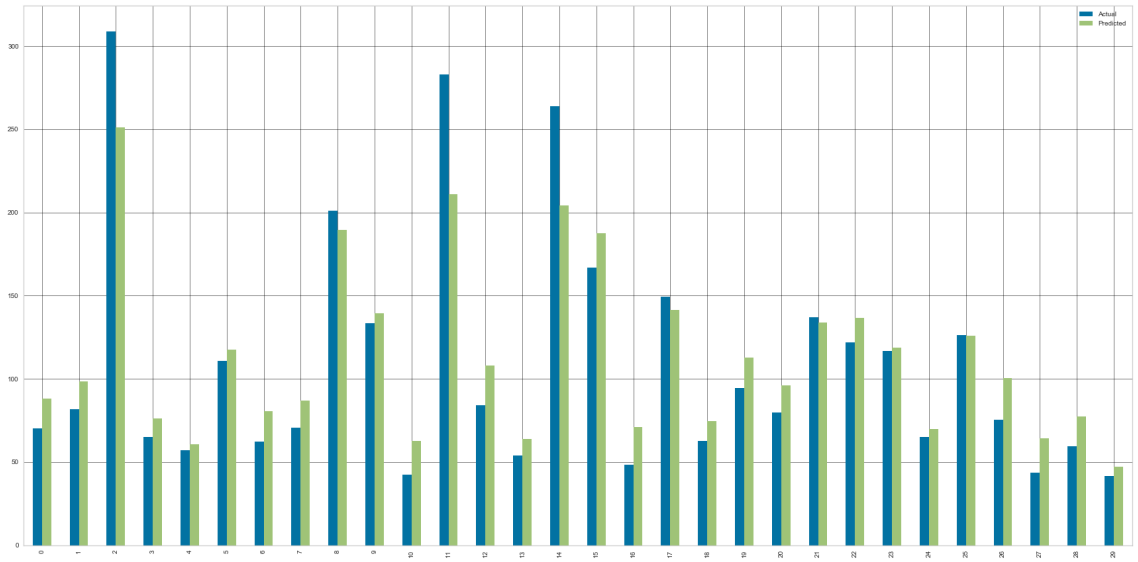

Figure 4. Actual vs Predicted Test Set Values

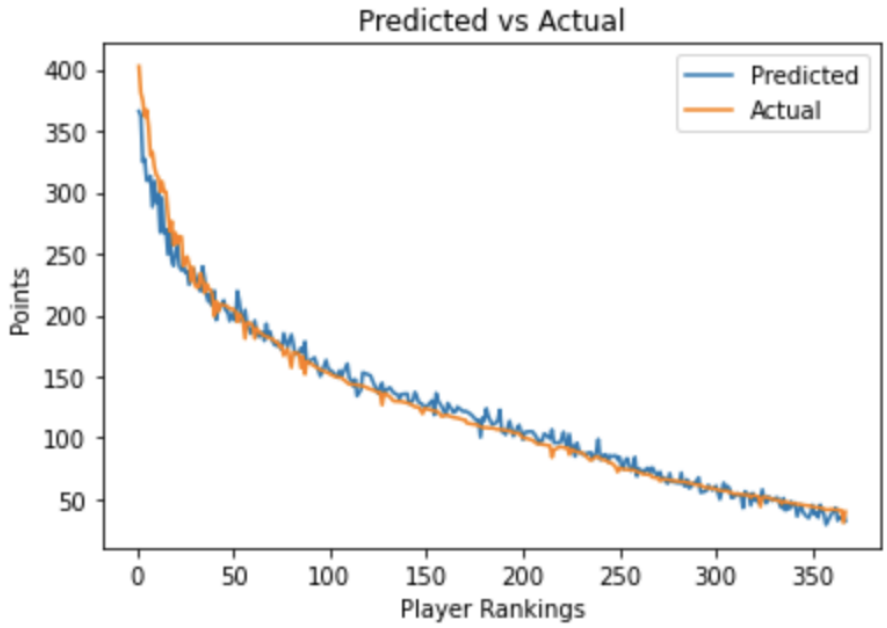

Figure 5. Actual vs Predicted Points for Player Rankings

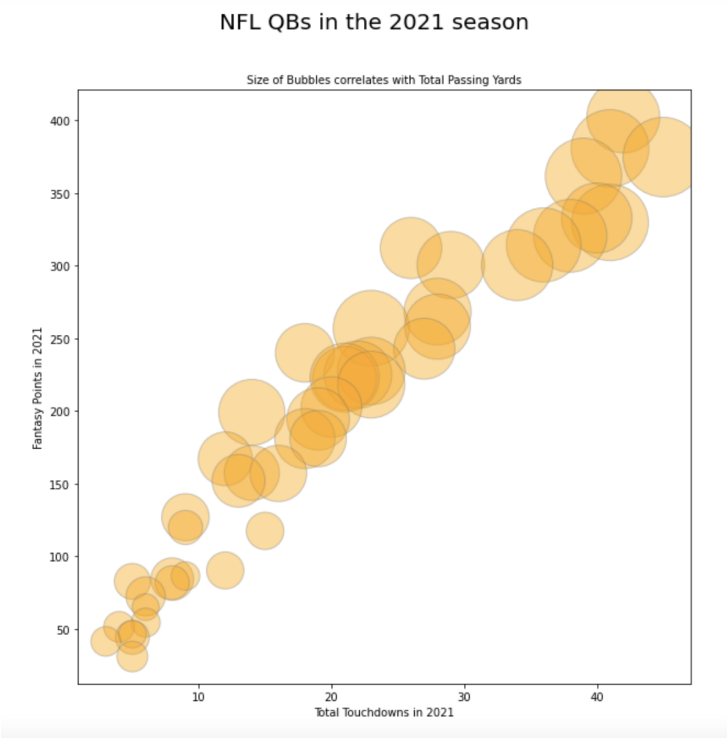

Figure 6: Total Touchdowns vs Total Fantasy Points for NFL QBs (2021 Season)

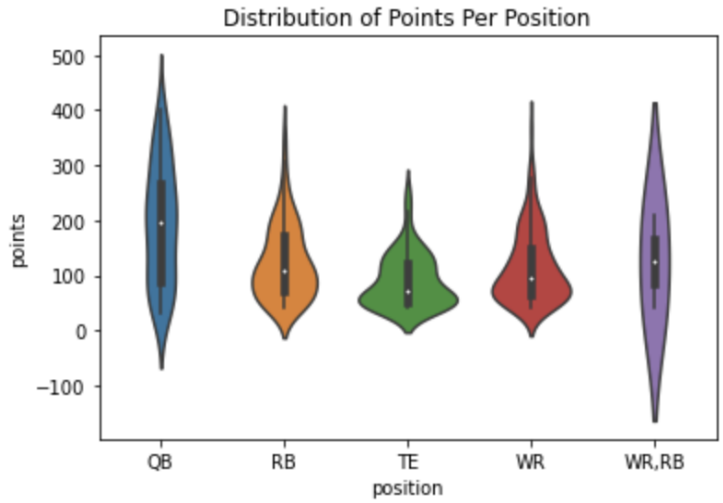

Figure 7: Distribution of Points Per Position (2021 Season)

Part 4: Limitations and Future Work

The final data analysis and the development of the culminating model has been very successful in answering our initial set of research questions. However, there are a few limitations that are important to acknowledge with respect to answering our research questions. First, our model uses the initial dataset we saved in our Github folder rather than raw data from the data source since player stats were removed from Yahoo Fantasy's servers, reducing potential flexibility. Another caveat to our results is that the model only uses data for the 2021-2022 NFL season. With access to more data from previous NFL seasons, we may have

been able to more accurately predict the fantasy football performance of NFL players using machine learning. Another limitation has been that we never did a full analysis of the statistical significance of each variable and how those variables interact with each other. Future work may include gathering further data and testing our model with a more comprehensive dataset that better captures an individual player's performance. An additional follow-up research question that would be interesting is how accurately can a machine learning model predict other player statistics such as fumbles and completed passes. Another interesting idea would be to create an algorithm to produce an overall predicted rating for each player's value during the season given their stats.

Part 5: Conclusion

In order to find how accurately one can predict the fantasy football performance of NFL players using machine learning, we gathered data for the 2021-2022 NFL season and built a culminating model that can accurately predict fantasy football performance of NFL players. Based on our model, we found substantial evidence to conclude that machine learning can accurately predict an individual NFL player's performance. Based on Figure 5, it is clear that our model computes predicted points with reasonable accuracy when compared with a player's actual points for the season. In addition to our model, we also successfully created a web application that shows each individual NFL player's current and projected fantasy statistics.

Our second research question, "What correlations can be drawn between variables and the fantasy football performance of NFL players", can be answered by Figures 2 and 3. There are some obvious correlations such as passing statistics (passing yards, interceptions, etc) with players who play QB (pos_QB variable in Figure 2). However, we also see more interesting correlations such as a slight negative correlation between playing RB or TE and the amount of fantasy points. This means that playing as a RB or TE results in fewer predicted fantasy points than playing as a QB or WR. Figure 3 shows our model's coefficients for each variable in the dataset. These model coefficients indicate how large of an impact a certain statistic has on the predicted amount of fantasy points. For instance, passing touchdowns and yards have a large impact on the amount of fantasy points a player has, while 2-point conversions have a relatively small impact on fantasy points. Figure 7 also shows that for certain NFL positions such as QB, the point distribution is much larger than WR, for example. Overall, our model and its accompanying visualizations help answer and understand certain correlations and relationships between variables and performance of NFL players.