Fantasy Football Project:

Feature engineering using domain knowledge: Figure out what could be the best features to predict on using your intuition as a fantasy football player. See what other people are doing for their predictions.

Stats.stackexchange:

|  |  |
| --- | --- |
|  | I think the key is the last clause, which suggests the author means a domain expert to be someone who has internalized important data patterns enough to quickly recognize them (or their absence) in a visualization. To get to that point, the expert must understand the normal relations between relevant variables so that unusual relations stand out.  Someone who's not an expert in the problem domain can spot patterns like correlation and outliers but not readily know if such patterns are important for the given variables. |

10-19-2016

Using the nflgame/nfldb I’ve been creating dataframes in pandas to utilize it’s capabilities for statistical anaylsis. I’ve so far created two separate data frames, one for passing and one for receiving/rushing. I think it would be better for me to make one query for all of offense, and then create one data frame of offense points that I can query from to get individual positions. This will also allow to see point leaders in a give year/week.

10-23-2016

For percentage think about how likely it would be for a player to get picked, aka the Likelihood. To think Bayesian, a prior belief of ownership would be necessary.

Find distribution of past player performance in the same spot. This could be position and ownership. This is the prior.

Likelihood might be the likelihood of observing the results (points, ownership) given our prior data.

10-24-2016

Little Steps:

1. Collect data/techniques
   1. NFLDB, fantasy football R package, Draft Kings data
   2. Read up on MCMC
2. Exploratory Data Analysis/Feature Engineering
3. Investigate different distributions/models to attempt to train

Models desired: Neural Nets, K-means Clustering for player profiles (stud, mud, dud)

10-29-2016

Check and see the score of games in first quarter, second quarter, third quarter, etc… to see if that might affect the fantasy production of some players

Let’s wait on this, get more data before trying this

10-31-2016

Collecting all data

* Get defense Dataframe
  + Defense stats needed for Draft King points:
    - Sacks
    - Interception
    - Fumble Recovery
    - Kickoff Return TD
    - Punt Return TD
    - Field Goal return TD
    - Blocked punt or FG TD
    - Safety
    - Blocked Kick
    - 2 point conversion/Extra point return
    - Points Allowed
  + Other defense stats needed for ranking:
    - Rushing yds allowed
    - Passing Yds allowed
* Offense Player Stats:
  + Offense Stats for Draft Kings Points:
    - Passing TDS
    - Passing Yds
    - Passing Interceptions
    - Receiving TDs
    - Receiving Yds
    - Receptions
    - Rushing TDs
    - Rushing Yds
    - Punt/Kickoff/Field Goal Return TD
    - Fumbles Lost
    - Offense Fumble Recovery TD
* Check if agg\_play would be computationally easier to pull from
  + Would take too much time, I already have a reliable system
* Decided to make a SQL query for all relevant offensive player stats, then create a data frame based solely on calculating points by dropping irrelevant columns
  + Didn’t drop any columns, didn’t think it would be necessary. If I need to I can return the DK points column and merge it into a nicer feature matrix

GOALS FOR TOMORROW:

* More Feature engineering using common methods as starting points, research some common methods
* Scrape Vegas Lines data

11-1-2016

Features desired

* Red Zone Targets/Touches
* Receiving Yds per reception
* Rushing yds per attempt
* Pt/$/1000 value

Red zone targets is difficult. I am getting multiple players with multiple receptions during the red zone play.

Trying to find how many red zone targets each player had…

11-1-2016 afternoon:

* Moved on to web scraping. Got my feet wet by scraping a csv URL from rotogrinders. The data is for player salary and player point projections. I used StringIO to read into into pandas data frame. Will now find a way to save data to csv for weekly look ups
* Have saved for week 9 projections now would like to check out

11-2-2016

* Scrum meeting:
  + Figure out correlation matrix
  + Normal Guassian kernel density
  + Determine how to use these distributions, do I want to aim high? Or low?
    - Have seen people use 75th quantile
* Checked out multicollinearity with pearson for QBs
  + Total-td-points 100% collinear with passing\_td. Instead of total\_td\_points, do total of all points (passing\_tds, rushing\_tds, two\_ptm)
  + Other of note:
    - Team score with total\_td\_points
    - Team score with DK points
    - Total\_td\_points with DK points
    - Yds\_per\_rec/yds\_per\_rush with passing\_twoptm/passing\_twopta
    - Passing\_com with passing\_td
* Estimated simple linear regression with pymc3 for QB team\_score predicting DK points
  + Used GLM and likelihood estimation
* Have been reading PPaBMfH chapter 3. Want to incorporate these methods of find unknowns and predictions to a simple output

11-3-2016

* What to do today… Try getting a solid model from pymc3 based on prior fantasy football performance
* Continue simple feature engineering ideas
* Fit regularized linear models for certain features/Fit Random forest or boosted trees to predict fantasy points
* Want to fit a simple linear model for QB with Passing\_att, team\_score
* Realized that the dataframes display the teams that players are currently on and not the teams they were on at the time of the statistic
  + Simply changed the SQL query to get team from play\_player table instead of player table
* Trying to fit regularized linear models using SK learn. Using RMSE function from assignment to calculate this.
  + Reverted to using sklearn mean\_squared\_error in a calc\_linear function
  + Have some scores, now I need to predict the features that go with that score
    - QB: TDs, passing\_yds,
    - More feature engineering:
      * How many passing yds per attempt
      * How many yds per receptions
    - Important questions to answer:
      * How many attempts will be made
      * How many targets will player get
      * How many touches will a player get
      * How many points is the team expected to score
    - Other things to work towards:
      * Ranking players by position
  + Instead, decided to do some feature selection methods on my current data

11-4-2016

* Instead of predicting those features, I need to mess around with PyMC3 today. Get a solid starting model.
* Looking at passing data, have a chain of highly correlated data:
* Feature selection:
  + Univariate
  + Recursive
    - Plot distributions of results of the distribution of the scores

11-5-2016

* Today is a planning day:
  + Goals for the week:
    - I need to figure out how to run the GLM with many features
    - Maybe should try normalizing data
    - Feature selection based on this
    - Feature selection for all positions
    - Backtest predictions on previous data
  + Extra curricular:
    - Would like to use neural nets for predictions as well

11-6-2016

* Fit an elastic net to my data, finding a lot of data leakage, need to get feature importance
* Want to do Grid Search to find best alpha and lambda parameters for ENet
* Had to switch scoring metric of GridSearchCV to ‘neg\_mean\_squared\_error’,
  + To interpret in y-values, I found RMSE
* Okay, got sklearn ElasticNet working, now need to test predictions in a given time period
  + Redesign matrices to have training data be on previous years, then test on 2016
  + Got weekly averages for all years 2009-2015 and used them to predict average QB points per week in 2016
    - Tried to generalize to Aaron Rodgers, RMSE of 8.6
    - Was thinking about finding out how much better Aaron Rodgers is than the average quarter back and train on that
    - NOPE
  + Going to try to run some non-parametric models to see if I can get some better results

11-9-2016

* Have created a class to format data frames
* Have fit a GBR model to my data to predict points per dollar
* I would like to start clustering . This might show me where I might be able to go next with my data.
* Have fit K-means clustering to some data. Worked pretty well for points per dollar with 3 clusters
* Tried kmeans on DK points vs DK salary, but got a vertical clustering on bunches of data
* Clustered on DK points, DK salary, h/a, and team score

11-11-2016

* Elliptic Envelope
* Want to draw my own threshold line