

Jimmys and Joes vs X's and O's

Predicting results in college sports analyzing talent accumulation and on-field success.

Reggie Bain, Reid Harris, Tung Nguyen



Background + Motivation

Changes in College Sports Landscape

- SCOTUS O'Bannon ruling allows name-image-likeness (NIL) money for players
- Recruiting industry evaluates high school players, a difficult task
- Teams allocating resources based on player evaluations
- Sports betting legalized in many states

Goals

- Predict future on field success
- Our Targets → **regular season win percentage, individual game results**
- Our Features → **Recent on-field performance + Talent level of team**




Datasets

Data Sources

- College Football Database API
- ESPN, Sports Reference - web scraping
- 247Sports Composite Rankings - web scraping



Rank	Team	Total	5-stars	4-stars	3-stars	Avg	Points
1 1	 Georgia	28 Commits	5	19	4	93.61	317.05
2 2	 Alabama	28 Commits	5	17	6	93.12	310.74
3 3	 Oregon	27 Commits	0	22	5	92.19	293.20

Stakeholders + KPIs

Key Performance Indicators

1. Identify key features that determine on-field outcomes
2. Predict season win totals accurately
3. Highly explainable model that allows for actionable insights



Stakeholders

- University athletic departments + NIL Collectives
- College coaching staffs (for assembling rosters)
- Professional and amateur sports gamblers



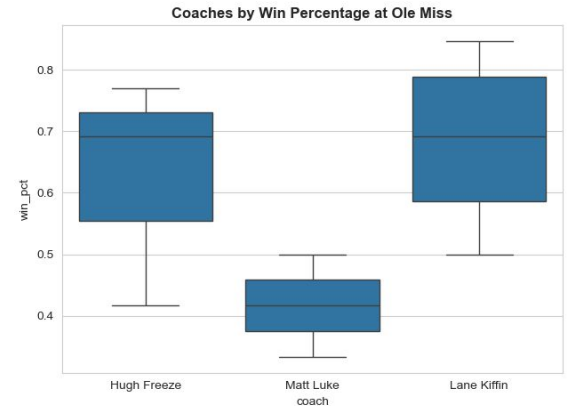
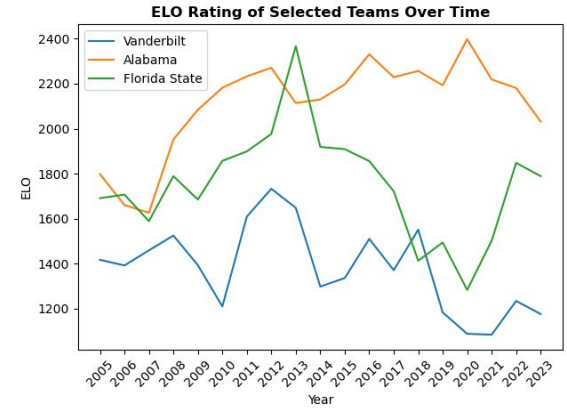
Feature Selection + Engineering

On-Field + Advanced Analytics

- ELO Rating, ESPN FPI
- Points/game, TDs/game, turnover margin, etc.
- Offensive/Defensive success rates
- Previous success of coach

Recruiting + Talent Metrics

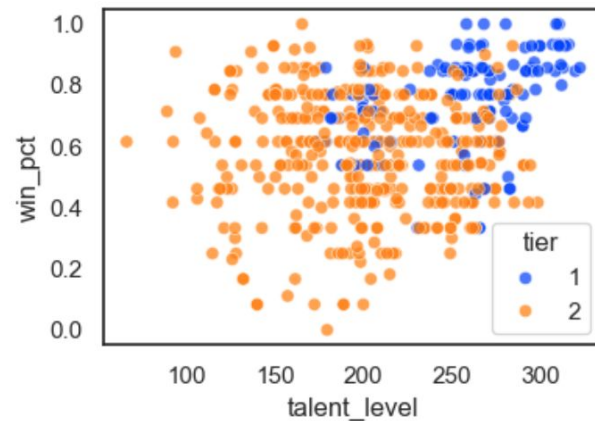
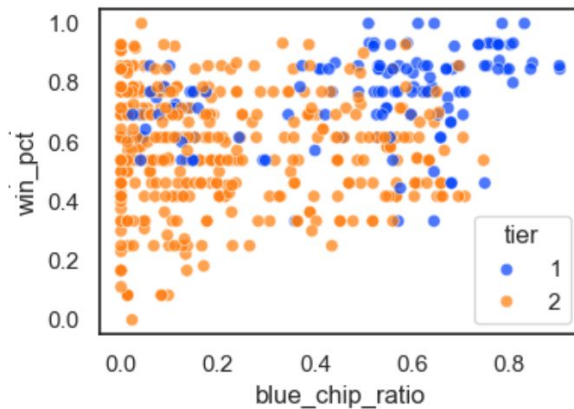
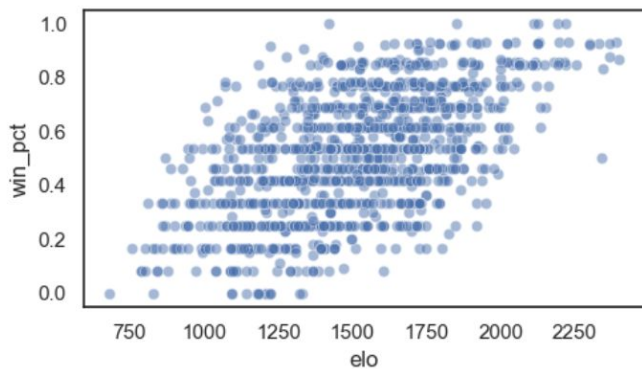
- Talent level based on recent recruiting
- Blue Chip Ratio
- Usages: %'s of returning talent from previous year



Exploratory Data Analysis

Exploring Talent & On-field Features

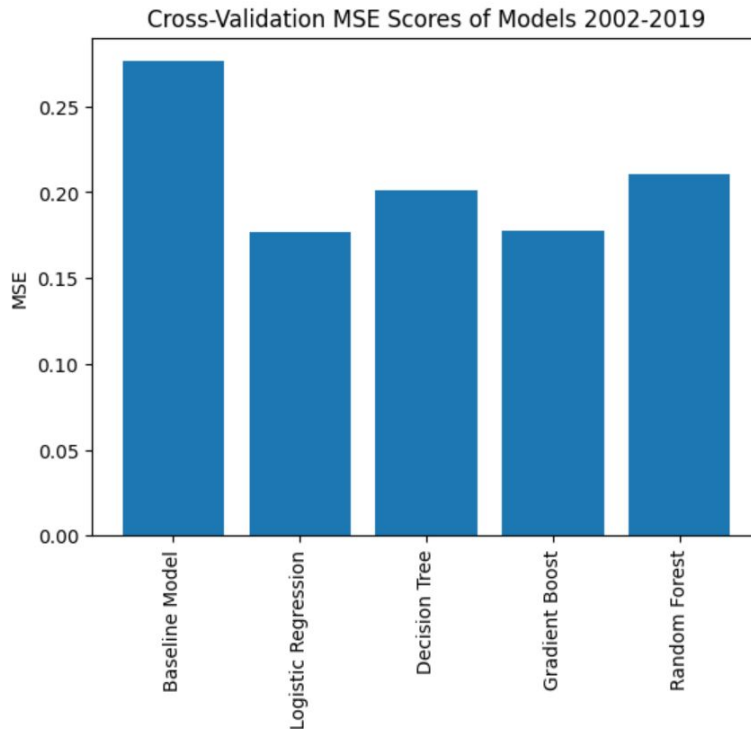
- Win Percentage vs. various features
- Explored “tiers” of teams based on recent success



Model 1 - Game by Game

Feature Importances - Game by Game

- Use in-game performance stats to predict outcomes for every matchup (win or loss)
- Prioritized recent performance, averaging performance over 4 game window
- Baseline model predicts that the team with the higher pregame ELO will win with probability 1

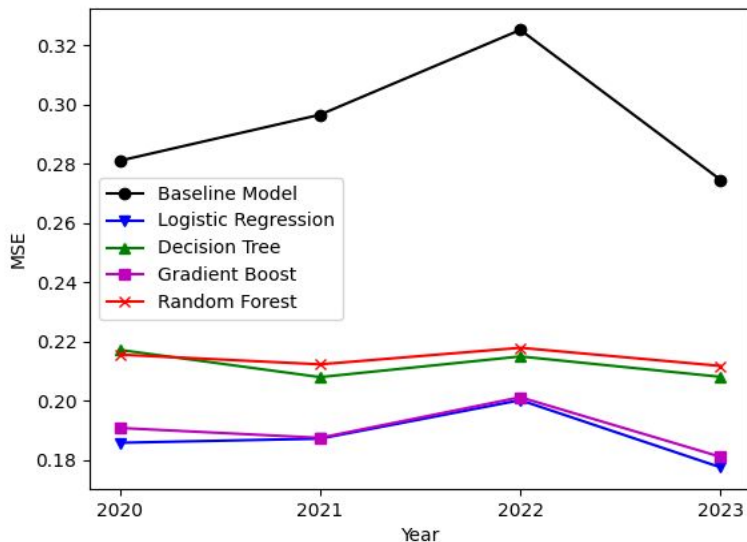


Model 1 - Results

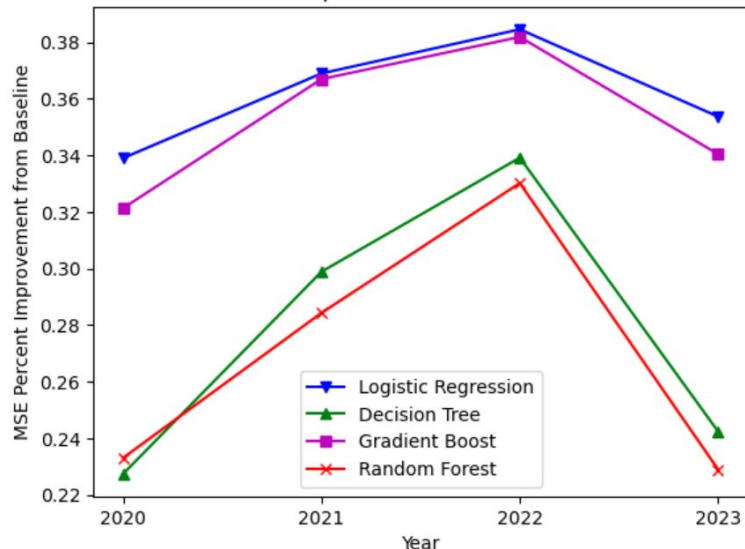
Mean Squared Error (MSE) - Game by Game Model

- Compared baseline with 4 other classification models to predict the probability of the game outcome.

MSE Scores on Games in 2020-2023



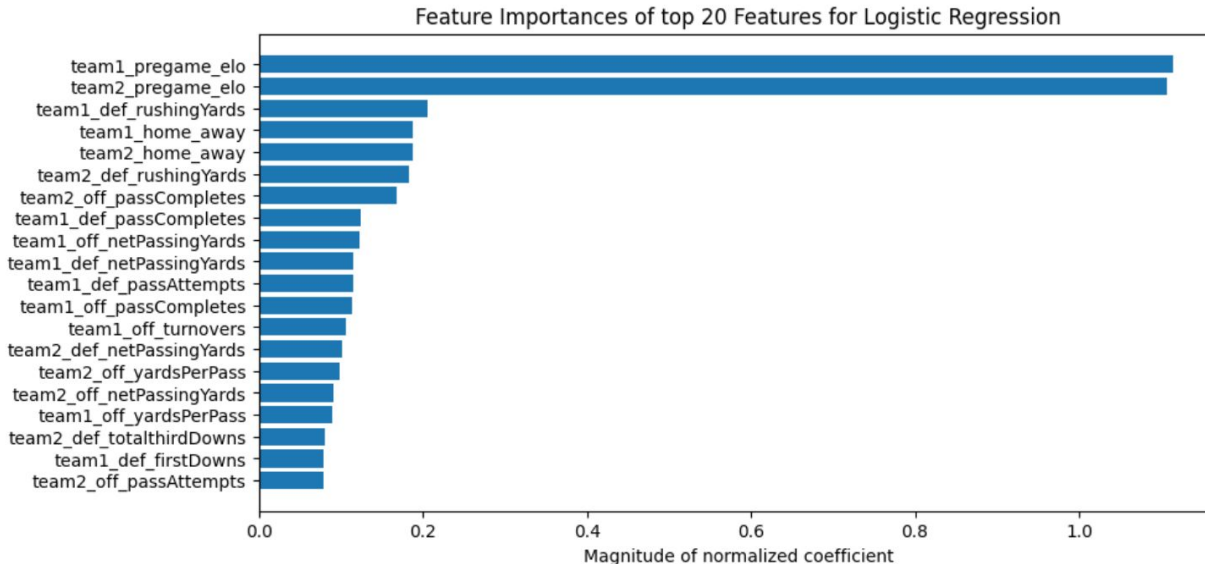
MSE Percent Improvement for Games 2020-2023



Model 1 - Logistic Regression

Feature Importances - Game by Game

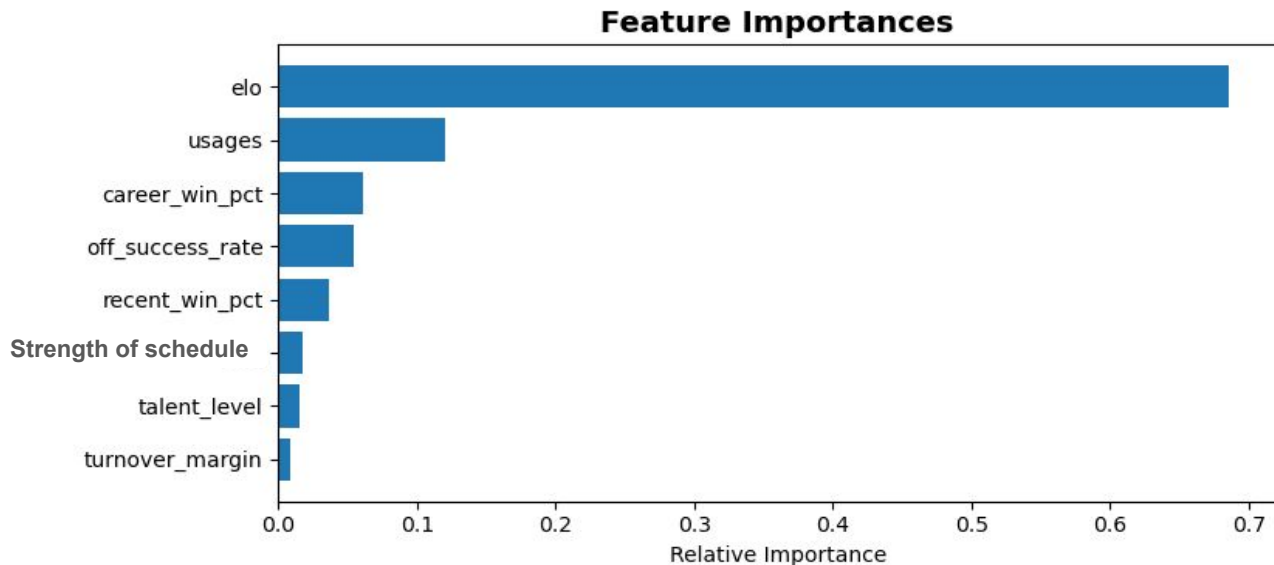
- Feature importance obtained by multiplying each coefficient by the standard deviation of that feature in the training data



Model 2 - Season Level

Feature Importances - Season Level

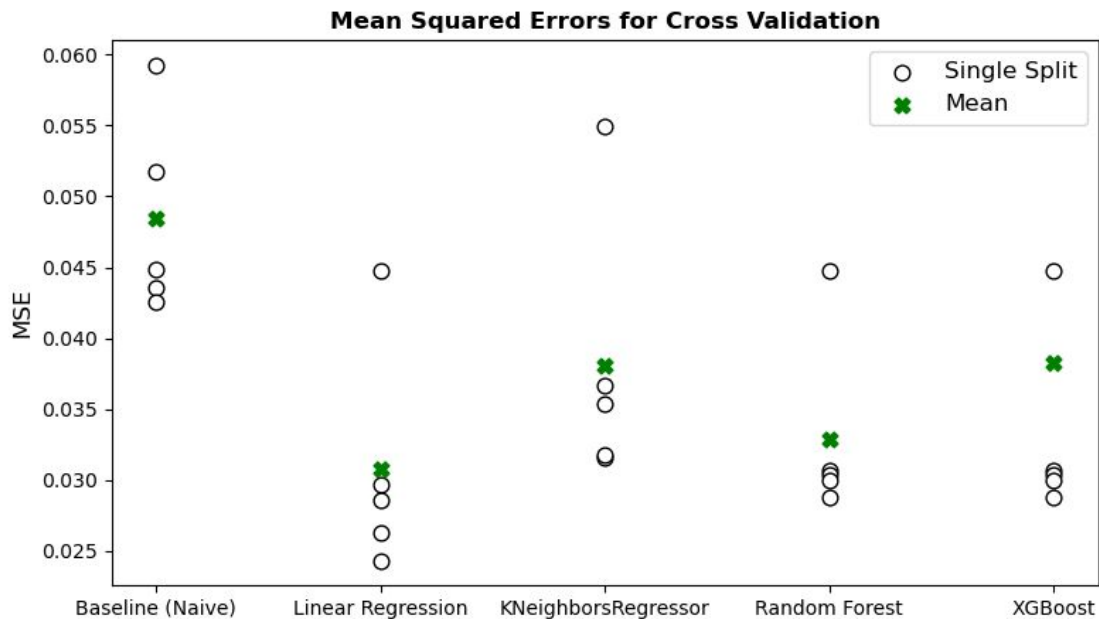
- Predict regular season win percentage for each team using previous performance + recruiting
- Used Random Forest + Lasso to measure relative feature importance



Model 2 - Results

Mean Squared Errors – Season Level Model

- Baseline Model: Naive forecast, same win % as last year



Model 2 - Results

Percent Improvements - Cross Validation

- Baseline Model: Naive forecast, same win % as last year
- Linear Regression outperformed out-of-box other models
- Used 5-fold cross validation, averaged mean-squared errors

model	avg_mse	avg_rmse	pct_improve_mse	pct_improve_rmse
Baseline Naive Forecast	0.0484019	0.220004	0	0
LinearRegression	0.0307367	0.175319	36.497	20.3112
KNeighborsRegressor	0.0380674	0.195109	21.3515	11.316
RandomForestRegressor	0.0329434	0.181503	31.9378	17.5002
XGBRegressor	0.0382383	0.195546	20.9983	11.1171
LSTM	0.0337128	0.18361	30.3482	16.5424

Model 2 - Results

Evaluating on the Test Set

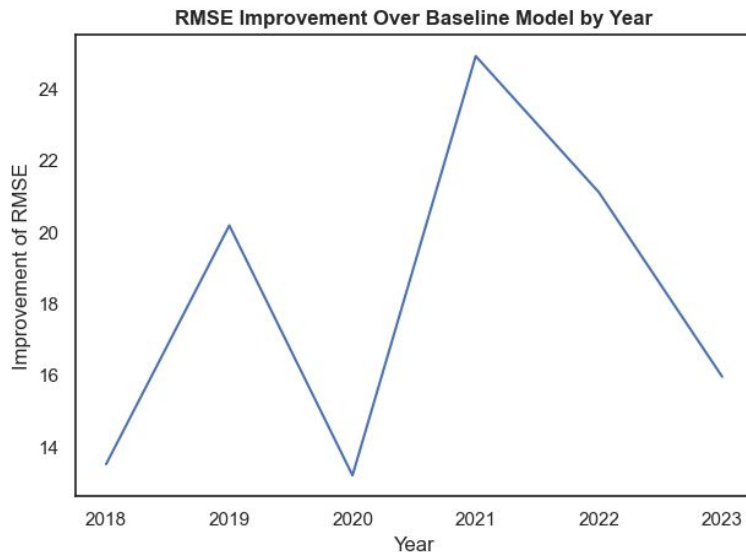
- Evaluated best model performing model (Linear Regression) on our test set (2023 data)

model	test_mse	test_rmse	pct_improve_mse	pct_improve_rmse
LinearRegression	0.0252754	0.158982	35.5216	17.9031
LSTM	0.0239591	0.154787	40.6756	20.5526

Model 2 - Results

Evaluating on the Test Set

- COVID-19 altered 2020 season significantly. Player opt outs, shortened schedules, player illness, etc. Hence bad MSE in cross-validation



Conclusions

EDA + Feature Selection

1. ELO dominates importance - recent results matter most (+ home field)
2. Talent levels matter less over all teams, more for elite teams
3. Coaching career success important, even from previous coached teams

Modeling

- Game x Game - upwards of 38% improvement over baseline w/ Logistic Regression
- Season Level - upwards of 37% improvement over baseline w/ Linear Regression
- Predict number of wins to within 1.908 per season

Future Work

Improving Models

- Hyperparameter tuning, especially for XGBoost, LSTM
- Change optimizers, learning rates, number/size of hidden layers
- Gather more data over longer time frame
- Expand investigation of game by game predictions

Expanding Features + Targets

- Predict more targets:
 - Points per game, TDs per game, etc..
 - Total score over/unders
 - Other more granular predictions
- Classification on sports bets: to take bet vs. not take bet
- Expand to other sports, particularly college basketball

Web Application

Allows users to interact with data + model wins

- Explore your team's data at: <https://bain-cfb-modeling-erdos.streamlit.app/>



Explore Your Team's Data

Select Year: 2024 | Select Team: South Carolina



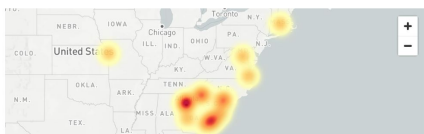
South Carolina Gamecocks

Model Predicted 2024 Record: 6-6 (± 1.908 wins)

Note: Record predicted using model trained on 2014-2023 data. See <https://github.com/reggiebain/cfb-modeling-erdos/tree/main> for more info!

Location: Columbia, SC | Conference: SEC
Stadium Capacity: 80250.0 | Current Coach: Shane Beamer

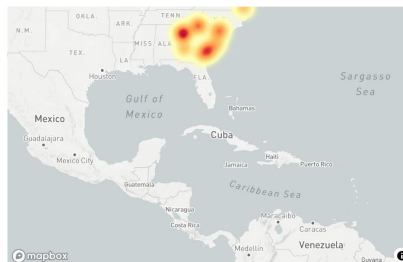
Heatmap of 2024 South Carolina Recruits



ELO Rating of South Carolina Since 2024

An ELO rating R_A sets/updates an expectation that a team will win a given game using the formula $E_A = 1/(1 + 10^{(R_B - R_A)/400})$. ELO ratings were the most important factor in our model for determining wins.

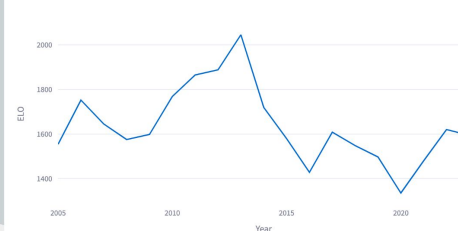
ELO Rating Over Time



Information on 2024 South Carolina Recruits

year	name	star	state	ranking	rating	position	height	weight
2024	Dylan Stewart	5	DC	15	0.9939	EDGE	77	235
2024	Josiah Thompson	5	SC	35	0.9839	OT	78	280
2024	Michael Smith	4	GA	147	0.9361	TE	76	235
2024	Wendell Gregory	4	GA	160	0.9326	LB	74	217
2024	Wendell Gregory	4	GA	179	0.9262	EDGE	74	220
2024	Kam Pringle	4	SC	180	0.9261	OT	79	338

ELO Rating Over Time



Recent Stats for South Carolina Leading Into 2024

For definitions of these terms see our writeup: <https://github.com/reggiebain/cfb-modeling-erdos>

year	recent_win_pct	talent_level	blue_chip_ratio	total_tds	totalYards	off_success_rate	sc
2014	0.6286	239.05	0.3053	57	5,880	0.4595	
2015	0.6223	235.65	0.3295	51	5,754	0.4655	
2015	0.6223	235.65	0.3295	51	5,754	0.4655	