Predicting FBS College Football Wins

Craig Johnson

Overview

- Project Description
- Introduction
- Data Acquisition
- Dataset Creation
- Exploratory Analyses
- Machine Learning Model
- Results
- Discussion
- Limitations and Future Directions

College Sports

- High school players identified, ranked, tracked 247sports, Rivals, ESPN
  - Millions of dollars spent before students hit campus
- 2015 report: Colleges sports generates more than 9.15 billion in revenue
- Increasing emphasis on data/data science in the college ranks
- With all the data collected in sports, can we predict wins?

Data Acquisition

- Common layman data sources
  - TV
  - Magazines
  - Newspapers
- Problem: Incomplete datasets
- Common large scale data sources
  - Databases
  - APIs
- Problem: Difficult and expensive to gain access
- Solution: Scrape NCAA’s website

Web scraping

- Framework used: Scrapy
  - Python based
- Build cycle
  - Identify
  - Pilot
  - Scale
  - Scrape

Identify

- Big question: How does this website work and how do I get what I need?
- Identify
  - Required pages
  - Required fields
  - Optimal “flow” to pages/fields

Pilot

- Use Jupyter Notebook and scrapy to build pilot programs capable of extracting data

Spider	Purpose
Teamlinks_spider	Generate team links by year
PeopleHistoryRosterStats_spider	Generate links to coach, team history, roster, stats
Coach_spider	Extract coaching history
Roster_spider	Extract team roster by year
History_spider	Extract team history
Teamstats_spider	Extract aggregate team stats by year
GagmeByGame_spider	Extract team stats for each game for each year
GameBygGameTeamName_spider	Extract the opponent team name

Scale

- Standardize code
- Migrate out of Jupyter Notebooks
  - Large scrape log will crash browser
- Run on small subsets of data to maximize performance

Scrape

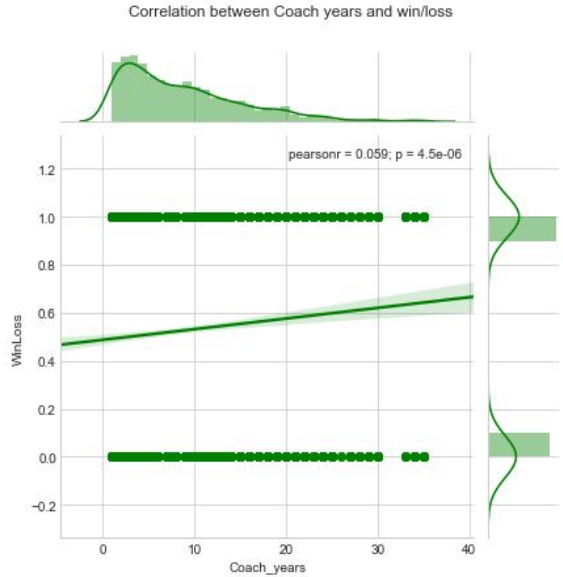
- Run the 8 scrapers to obtain data
  - Run time anywhere from minutes to over 24 hrs
  - Calls every 1.5 seconds
- Final counts
  - 61 files
  - 93 mbs of data

Dataset Creation

- End goal:
  - One line of data for every game
  - Each line of data should include the offense and defensive averages for the home and away team
- Important notes
  - Each line of data represents the AVERAGE of the preceding games for the year.
    - E.G: Game 8 is made of up the previous 7 games
  - Game 1 of 2013 dropped: NO previous data available to predict!
  - Game 1-3 of every year were predicted by the previous year plus any previous game data.
    - E.G.: Game 2 is predicted by the average statistics of the previous year + game 1

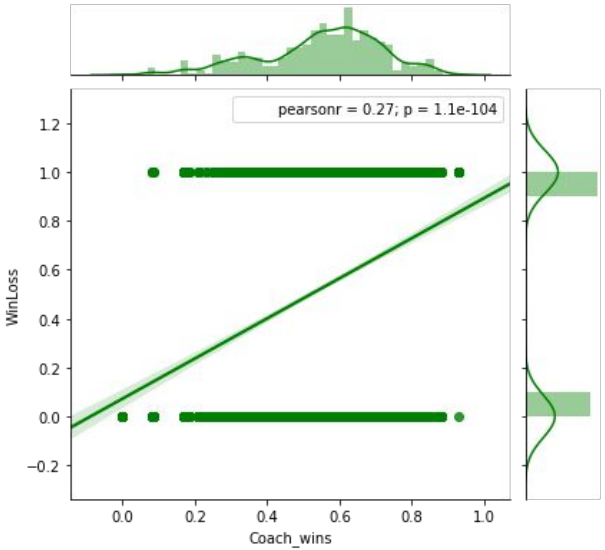
Exploratory Data Analysis

Small correlation between how long a coach has been coaching and the success in the current game.

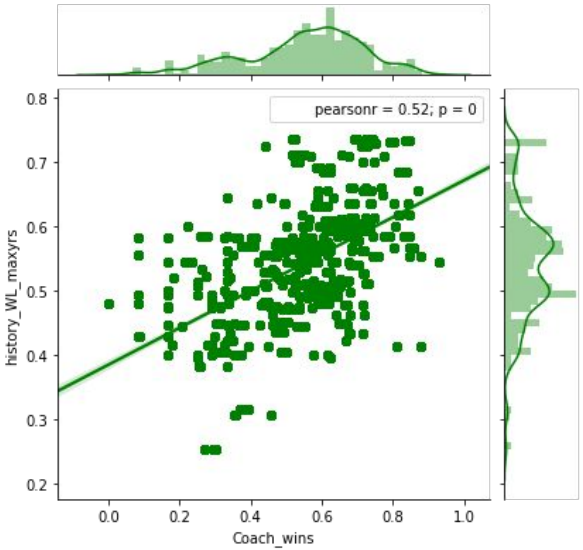


EDA - Coaching

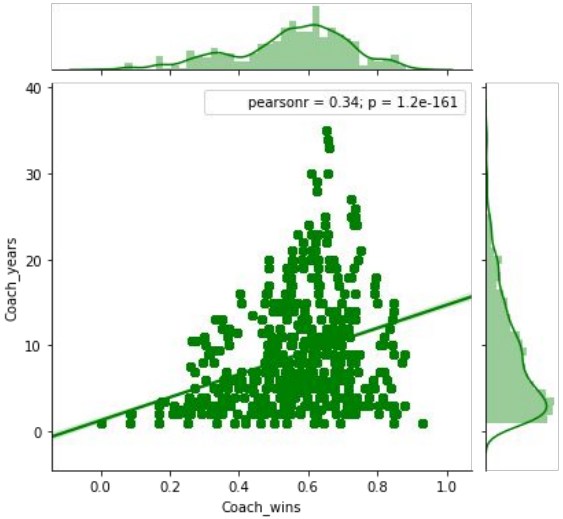
Correlation between Coach Win Percentage and Win/Loss



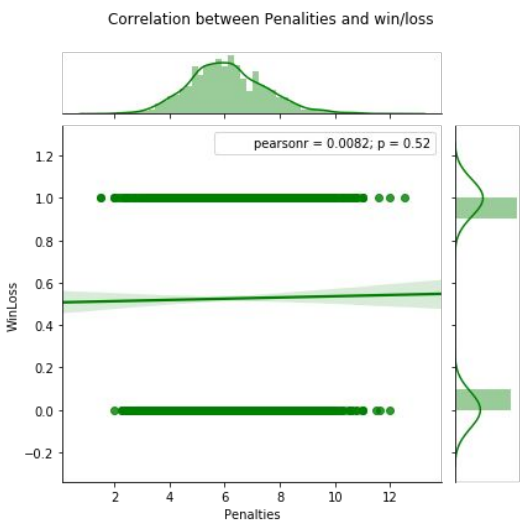
Correlation between Coach wins and school wins

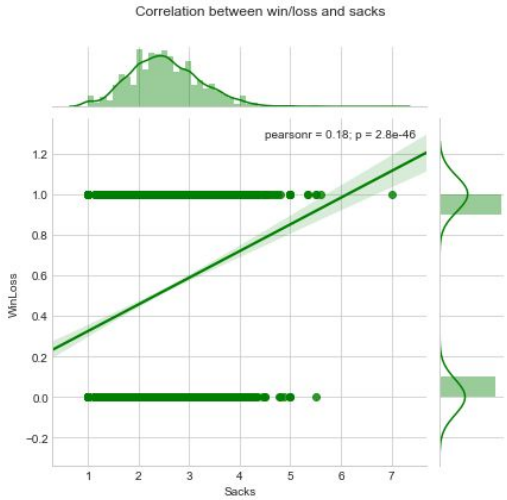


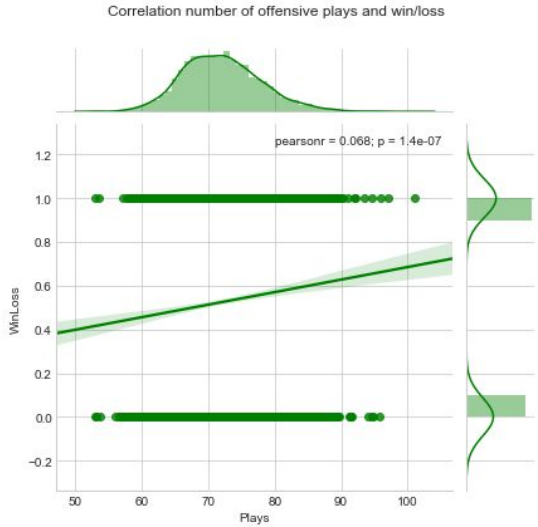
Correlation between Coach wins and number of years coaching



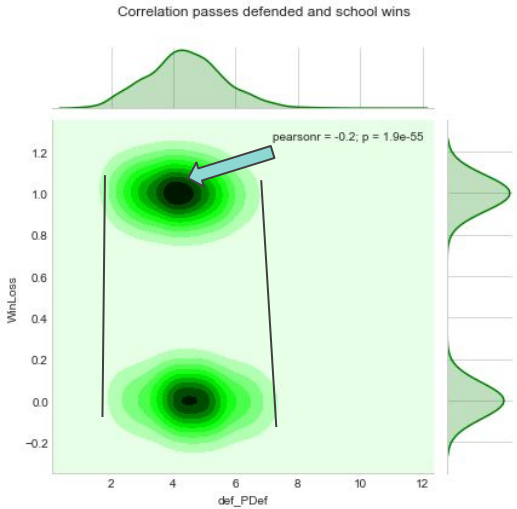
EDA - Penalties, Sacks, Offensive Plays

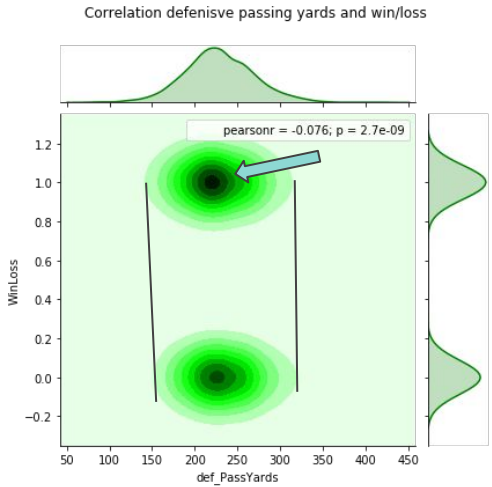


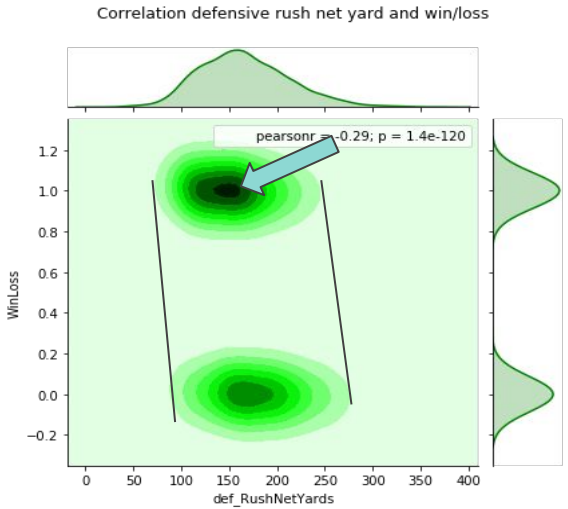




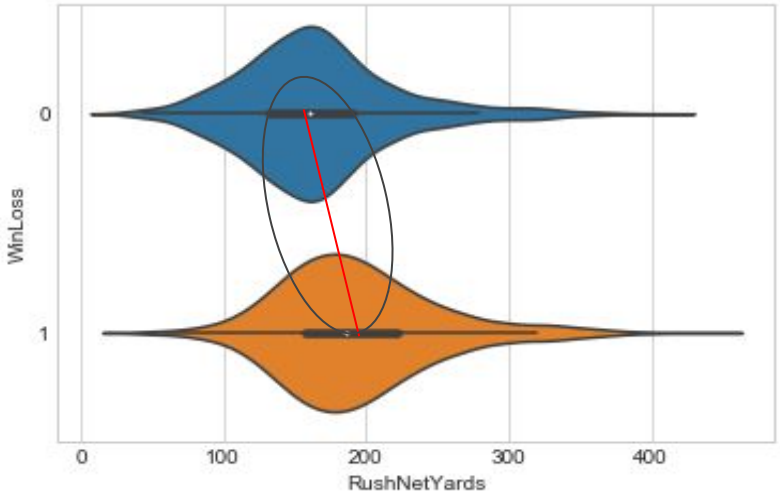
EDA - Kernel Density Plots: Defensive Rushing, Passes, Passes Defended

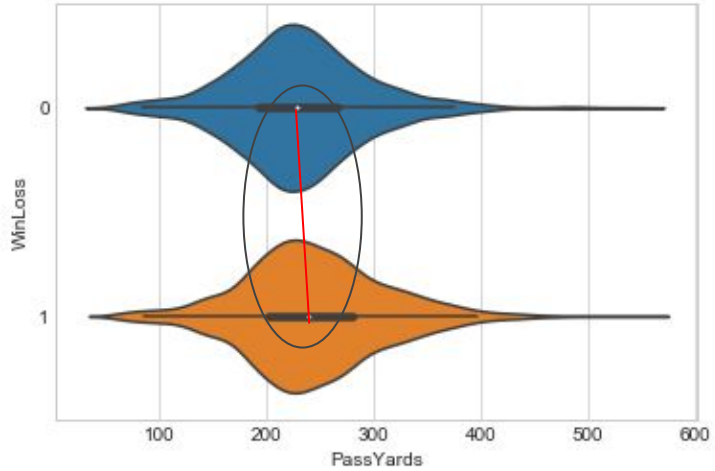




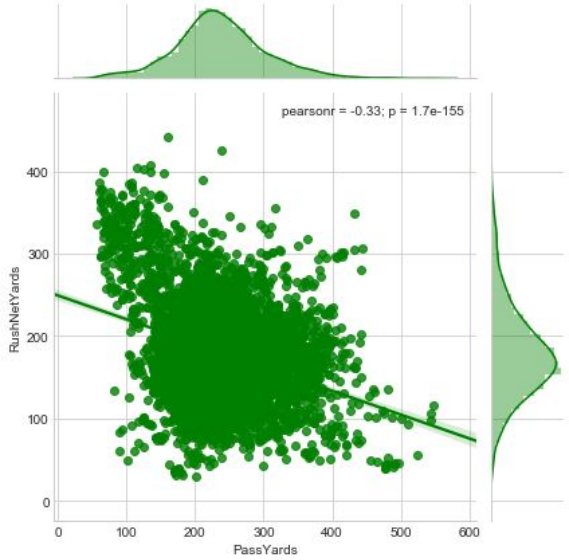


EDA - Violin Plots Rushing / Passing





EDA - Rushing Yrds & Passing Yrds



Intuitive negative correlation. If you rush a lot you probably aren't going to have a lot of passing yards as well.

Machine Learning Model

- Random Grid Search CV + Pipeline
  - CV = 3
- Pipeline
  - Step 1: Imputation (mean replacement)
  - Step 2: Feature Selection
    - Random Forest Classifier
    - Estimators = 100
    - Minimum samples = 20
    - Max Depth = 3
  - Step 3: Random Forest Classifier
    - Hyperparameters next slide...

Machine Learning Model continued...

Hyperparameters

- Criterion: Gini, entropy
- Max Depth: 1 to 5 (inclusive)
- Minimum Sample Size: 10 to 50 by 5
- Estimators: 100, 250, 500, 750, 100, 1250

Best Model

- Criterion: Entropy
- Max Depth: 5
- Minimum Samples: 15
- Estimators: 1,000

Results

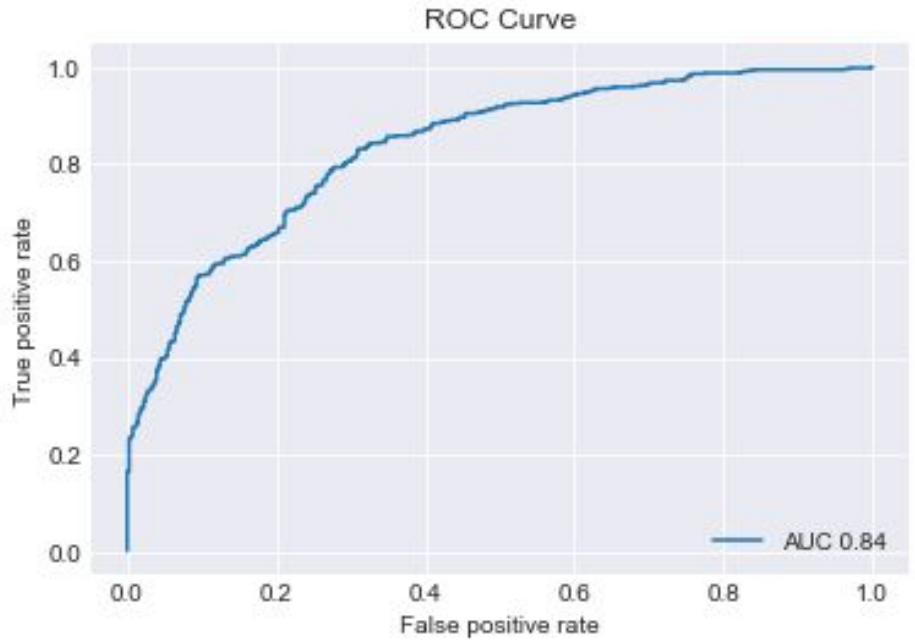
- Accuracy
  - Train: .77
  - Test: .76

Takeaways:

- Stable results between train and testing so the model isn't over fit.
- Most features are yards based, unexpected!

Feature	Importance
Rushing - Net Yards	.15
Punt Return Yards	.15
Fumble Return Yards	.14
All-Purpose Yards	.14
Receiving Yards	.12
Interception Return Yards	.11
Kickoff Return Yards	.11
Rumbles Recovered	.07

ROC Curve



Test Data

		Actual	
		Loss	Win
Predicted	Loss	246	162
	Win	78	509

Takeaway: Pretty good prediction!

Classification Report

	Precision	Recall	F1-Score	Support
Loss	.76	.60	.67	408
Win	.76	.87	.81	587
avg/total	.76	.76	.75	995

Take-away: F1-score indicates good prediction for precision and recall

Discussion

- Stable and predictive results
- Accuracy ~ 76%
- Theory != Empirical feature selection

Limitations / Future Directions

- Improve tracking and handling missing data
  - Broken links on site
  - Missing data / imputing data
- Feature Selection
  - Not all data scraped was included
  - Additional data sources exist
  - Different techniques for selecting features
- Model
  - Try different modeling techniques and grid searches