



Predicting Outcomes of Week One College Football Games

Sam Alptekin, Jacob Beiter, Sam Berning, Ben
Shadid



Introduction / Background

- Before the season, there is a lot of uncertainty about how good a college football team will be
 - Players graduating or getting drafted
 - Coaching changes
 - Various random factors
- Make sense of the chaos



S&P+

- Not intended as a predictive tool
- Pre-season rankings based on 5 aggregate statistics
- Current method only in practice since 2014
- Pretty good at predicting team strength
 - Definitely room for improvement



Problem Statement

1. Given teams' statistics from the previous football season, how can we accurately predict the result of week 1 games for the current year?

1. Using S&P+ as a benchmark, what features can we incorporate to create a useful predictive model?



Project Overview

- Data Sources
 - S&P+ -- footballoutsiders.com
 - Recruiting scores -- 247sports.com
 - Schedule & Results data -- ESPN.com
- Scraped data using python scripts
- Integrated using Microsoft Excel
- Game-specific features defined as differences

Project Overview: Benchmark

Initial Stage (milestone):

- C4.5 Tree
- Benchmark and new model (~20 features, very small data)
- Recruitment scores

Benchmark

Result \ Predicted	Lose	Win	Total
Lose	TN = 4	FP = 19	N = 23
Win	FN = 6	TP = 45	P = 51
Totals	N' = 10	P' = 64	74

Accuracy 0.662162162162

F1 Score 0.782608695652

Precision 0.703125

Recall 0.882352941176

Proposed Model

Result \ Predicted	Lose	Win	Total
Lose	TN = 3	FP = 20	N = 23
Win	FN = 2	TP = 49	P = 51
Totals	N' = 5	P' = 69	74

Accuracy 0.702702702703

F1 Score 0.816666666667

Precision 0.710144927536

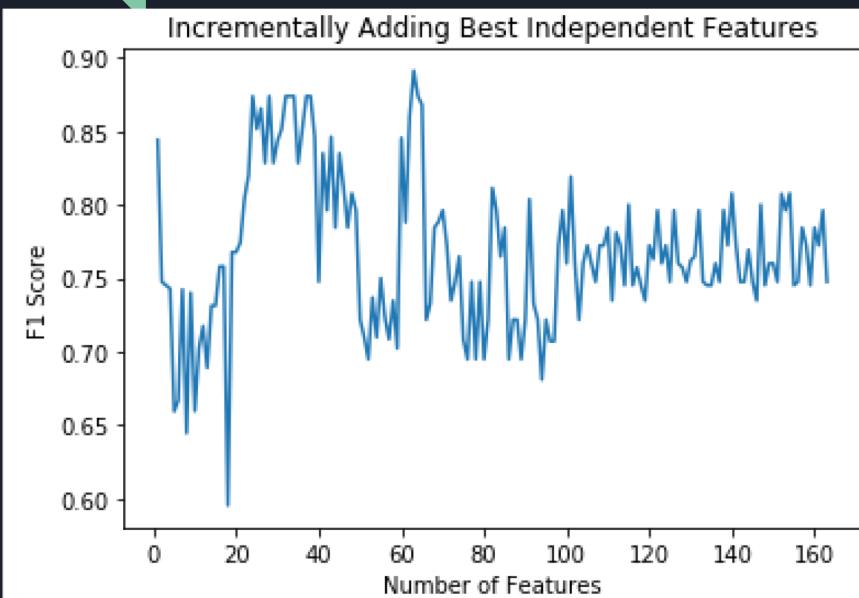
Recall 0.960784313725



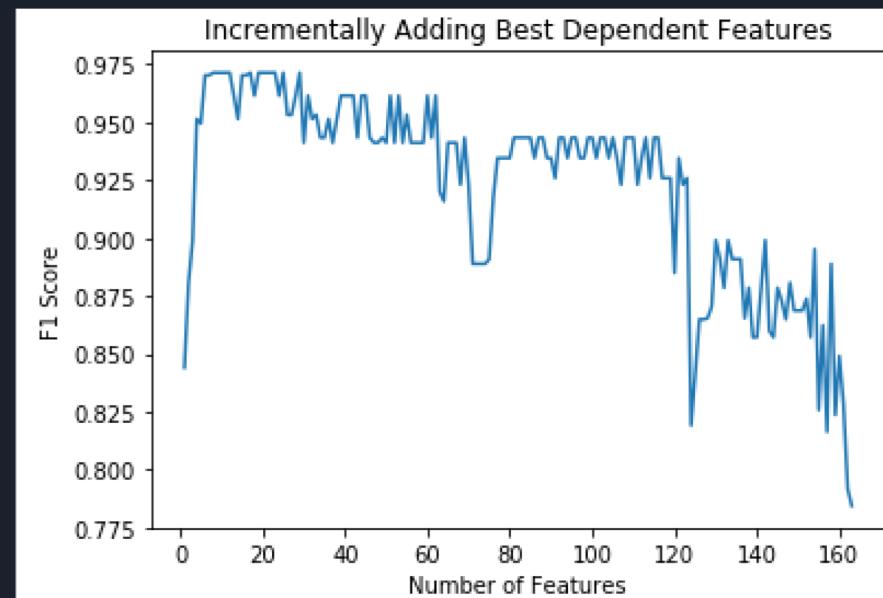
Increase dimensionality

- Get a ton of features and determine which were most likely to carry over to the next year
- Teamrankings.com
 - 144 stats for each team, each year
 - Offense, defense, special teams, penalties, turnovers, etc.
- Added ~92,160 data items!

Feature Selection: Decision Tree (Regression)



Average Team Passer Rating



Average Team Passer Rating, S&P+, Punt
Attempts per Game, Red Zone Scoring
Percentage, Fumbles Not Lost per Game,
Opponent Turnover Margin per Game



Feature Selection: Decision Tree (Regression)

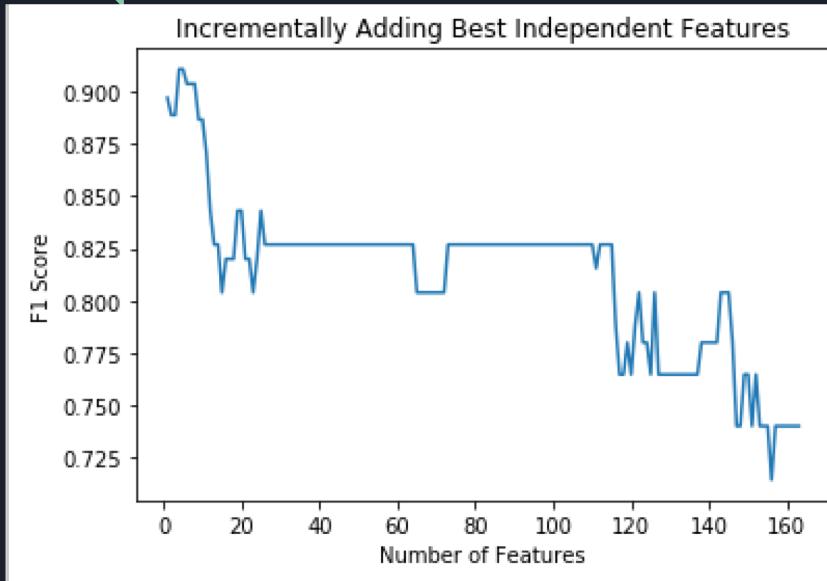
1 Feature, Independent

```
Confusion Matrix:  
[[11 12]  
 [ 5 46]]  
Accuracy: 0.77027027027  
F1 Score: 0.844036697248  
Precision: 0.793103448276  
Recall: 0.901960784314
```

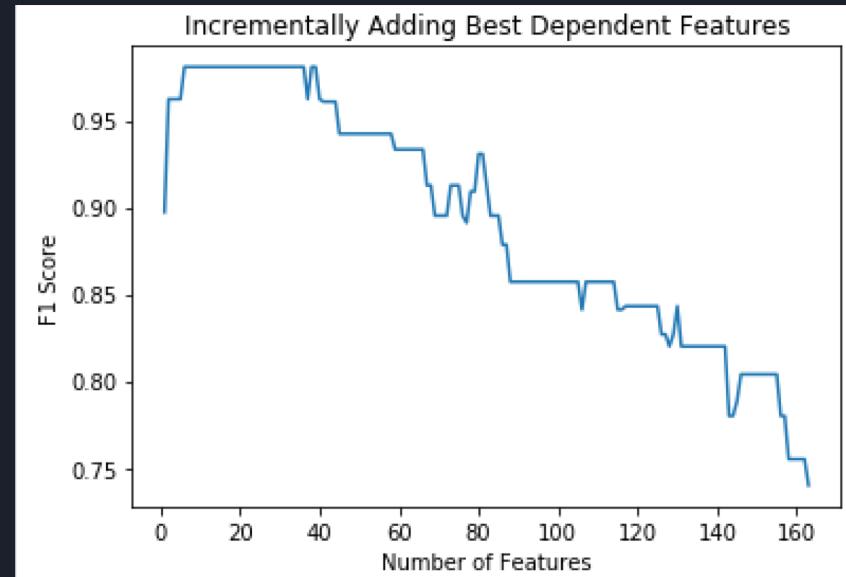
6 Feature, Dependent

```
Confusion Matrix:  
[[20  3]  
 [ 2 49]]  
Accuracy: 0.932432432432  
F1 Score: 0.95145631068  
Precision: 0.942307692308  
Recall: 0.960784313725
```

Feature Selection: Gaussian Naive Bayes



Recruiting Rank, Recruiting Points, Opponent Third Down
Conversion Percentage, Yards per Point, Defensive
S&P+, Points per Play, 1st Half points per Game,
Opponent 2nd Quarter Points per Game



Recruiting rank, Opponent Yards Per
Penalty, 2nd Quarter Possession
Percentage, 1st Half Possession Percentage,
Passing Play Percentage

Feature Selection: Gaussian Naive Bayes

4 Feature, Independent

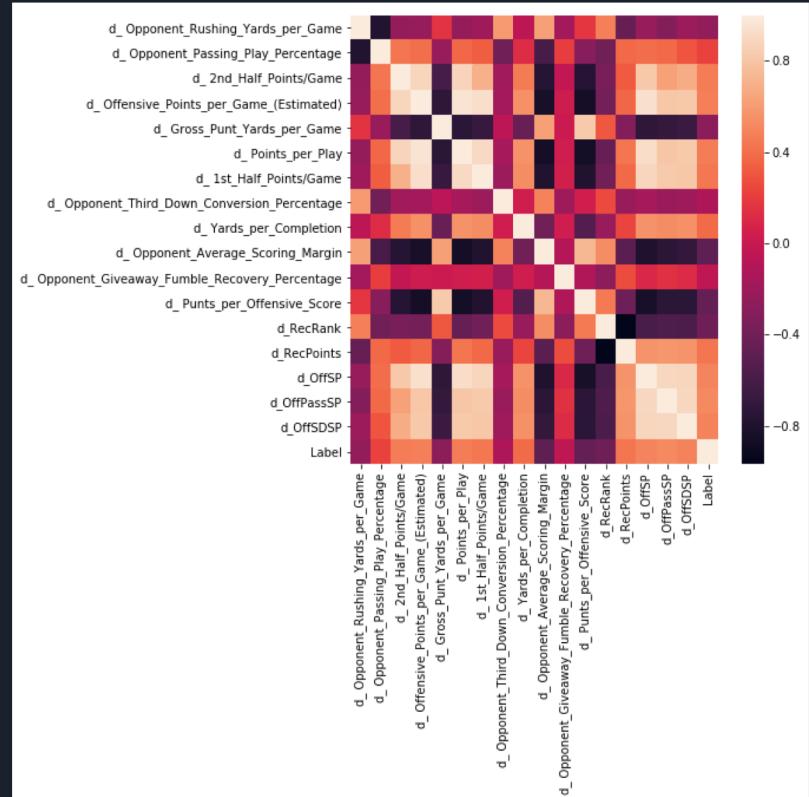
```
Confusion Matrix:  
[[19  4]  
 [ 5 46]]  
Accuracy: 0.878378378378  
F1 Score: 0.910891089109  
Precision: 0.92  
Recall: 0.901960784314
```

6 Feature, Dependent

```
Confusion Matrix:  
[[21  2]  
 [ 0 51]]  
Accuracy: 0.972972972973  
F1 Score: 0.980769230769  
Precision: 0.962264150943  
Recall: 1.0
```

Feature Selection: ETC / SVM

- Recursive Feature Elimination
 - Various feature counts tested, ExtraTree
 - Start with all features
 - Remove until threshold reached
- N = 17
 - Most consistent estimation after RFE
 - Same features for ETC and SVM
- SVM
 - RBF Kernel
 - GridSearchCV for parameter selection



Modeling and Results: Summary

S&P+ Benchmark (C4.5)

Result \ Predicted	Lose	Win	Total
Lose	TN = 4	FP = 19	N = 23
Win	FN = 6	TP = 45	P = 51
Totals	N' = 10	P' = 64	74

Accuracy: 0.662162162162
F1 Score: 0.782608695652
Precision: 0.703125
Recall: 0.882352941176

ExtraTrees

```
Confusion Matrix:  
[[16  7]  
 [ 0 51]]  
Accuracy: 0.905405405405  
F1 Score: 0.935779816514  
Precision: 0.879310344828  
Recall: 1.0
```

1-Feature Independent

```
Confusion Matrix:  
[[11 12]  
 [ 5 46]]  
Accuracy: 0.77027027027  
F1 Score: 0.844036697248  
Precision: 0.793103448276  
Recall: 0.901960784314
```

4-Feature, Independent

```
Confusion Matrix:  
[[19  4]  
 [ 5 46]]  
Accuracy: 0.878378378378  
F1 Score: 0.910891089109  
Precision: 0.92  
Recall: 0.901960784314
```

Classification SVM

```
Confusion Matrix:  
[[10 13]  
 [ 0 51]]  
Accuracy: 0.824324324324  
F1 Score: 0.886956521739  
Precision: 0.796875  
Recall: 1.0
```

6-Feature Dependent

```
Confusion Matrix:  
[[20  3]  
 [ 2 49]]  
Accuracy: 0.932432432432  
F1 Score: 0.95145631068  
Precision: 0.942307692308  
Recall: 0.960784313725
```

6-Feature, Dependent

```
Confusion Matrix:  
[[21  2]  
 [ 0 51]]  
Accuracy: 0.972972972973  
F1 Score: 0.980769230769  
Precision: 0.962264150943  
Recall: 1.0
```



Results & Analysis: ETC / SVM

Feature frequencies over 100 ETC runs:

- Offensive efficiency most important
 - Offensive Points / Game: 71
 - 2nd Half PPG : 61
 - Points / Play: 61
- S&P+ feature importance reduced
 - Second Order Wins: 1
- Defensive features considered least important
 - Defensive Success Rate (S&P+): 0
 - Sacks / Game: 0

Most Importantly...

Who will win this year?



(Obviously, Notre Dame)