Project 11:

Predicting Purdue's Performance in the NCAA Tournament Bracket

Modeling

Project Lead: Professor Petros Drineas, Christos Boutsikas

Members: Zachary Hanson, Avi Khurana

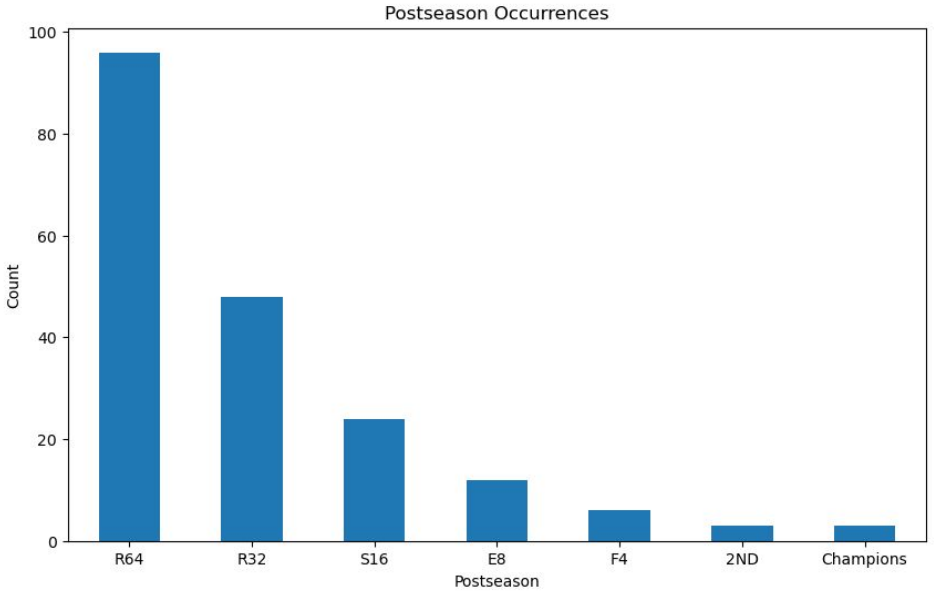
Kuba Bal, Saul Means, Sarah Firestone

PURDUE UNIVERSITY

Modeling

Background

1) Goal: Create a model to predict Purdue's performance in March Madness
2) Multiclass Classification Problem
3) Unbalanced Classes



PURDUE UNIVERSITY

Modeling

Background

4 Models: Classification

1) Support Vector Machine

2) K Nearest Neighbors

3) Multinomial Logistic Regression

4) Random Forest

PURDUE UNIVERSITY

Modeling

Modeling Assumptions

- **4 classification models:** data assumptions
  - Entirely numerical data (SVM, MLR, KNN)
  - Scaled features (MLR, SVM, RF)
  - Data separable by hyperplane (SVM)
  - Data does not contain large amounts of noise (SVM, KNN)
  - Sufficient sample size (SVM, MLR, KNN, RF)
  - Assumptions built into python library (RF)

PURDUE UNIVERSITY

Modeling

Test Design

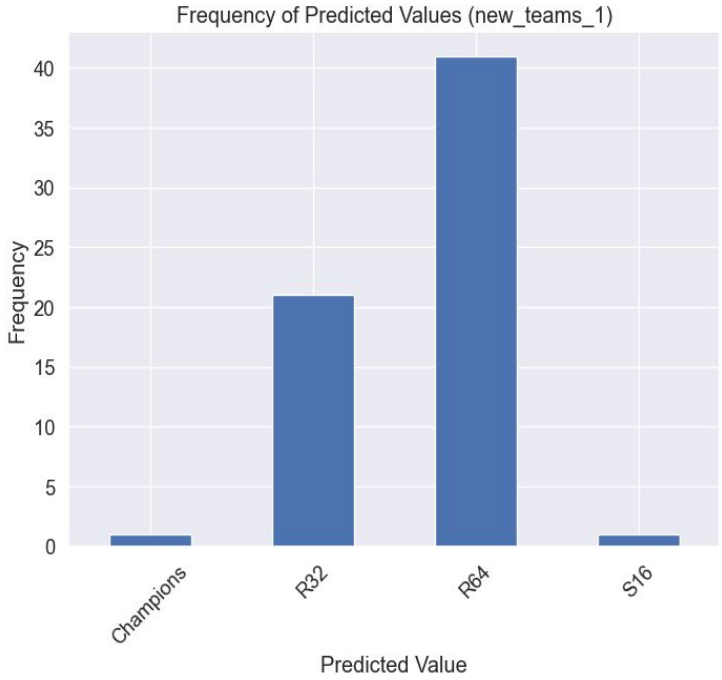
- Determine feature importance
- Split data into an 80/20 training test split
- Use a Grid search to determine hyperparameters
- Evaluate the model using accuracy, precision, recall, and F1-score
- Platt Scaling

PURDUE UNIVERSITY

Modeling

Found an issue

of teams predicted to be eliminated in R64 > actual # of teams allowed



PURDUE UNIVERSITY

Modeling

Solution: Create Constraints and use Platt Scaling

Sets how many teams can be eliminated each round

- **R68**: 4 teams
- **R64**: 32 teams
- **R32**: 16 teams
- **S16**: 8 teams
- **E8**: 4 teams
- **F4**: 2 teams
- **2ND**: 1 team
- **Champion**: 1 team

PURDUE UNIVERSITY

Modeling

SVM

Prior to grid search

Classification Report:

	precision	recall	f1-score	support
2ND	0.00	0.00	0.00	1
Champions	0.00	0.00	0.00	1
E8	0.00	0.00	0.00	4
F4	1.00	1.00	1.00	1
R32	0.20	0.12	0.15	8
R64	0.50	0.79	0.61	19
S16	0.33	0.20	0.25	5
accuracy			0.46	39
macro avg	0.29	0.30	0.29	39
weighted avg	0.35	0.46	0.39	39

Accuracy on the test set (top 5 features): 0.4615

Applied hyperparameters:

Classification Report:

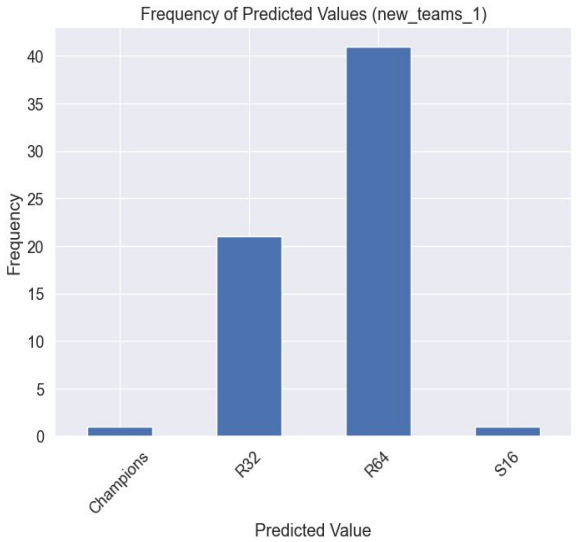
	precision	recall	f1-score	support
2ND	0.00	0.00	0.00	2
Champions	0.00	0.00	0.00	1
E8	0.00	0.00	0.00	1
F4	0.33	0.33	0.33	3
R32	0.50	0.56	0.53	9
R64	0.64	0.78	0.70	18
S16	0.25	0.20	0.22	5
accuracy			0.54	39
macro avg	0.25	0.27	0.25	39
weighted avg	0.47	0.54	0.50	39

Accuracy on the test set (top 5 features): 0.5385

PURDUE UNIVERSITY

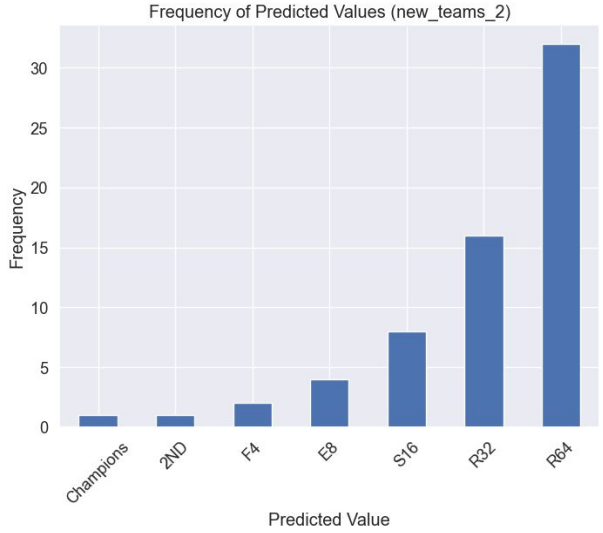
Modeling

Prior to platt scaling:



43	Purdue	R32

After platt scaling:



43	Purdue	F4

PURDUE UNIVERSITY

Modeling

KNN

Split: 0.2

k = 2: purdue S16, acc: .48

```
Model accuracy: 0.4878048780487805
```

	TeamName	PredictedStage
0	Akron	Champions
1	Alabama	2ND
2	Florida Atlantic	F4
3	Illinois	F4
4	Auburn	E8
...	...	...
59	Colorado	R64
60	Colgate	R64
61	Vermont	R64
62	Nevada	R64
63	Virginia	R64

64 rows × 2 columns

```
   TeamName PredictedStage
12  Purdue            S16
```

k =3 purdue S16, acc: 0.46

```
Model accuracy: 0.4634146341463415
```

	TeamName	PredictedStage
0	Akron	Champions
1	Alabama	2ND
2	Florida Atlantic	F4
3	Gonzaga	F4
4	Auburn	E8
...	...	...
59	Colorado	R64
60	Colgate	R64
61	Wagner	R64
62	South Dakota St.	R64
63	Yale	R64

64 rows × 2 columns

```
   TeamName PredictedStage
10  Purdue            S16
```

k=20: purdue S16, acc:0.39

```
Model accuracy: 0.3902439024390244
```

	TeamName	PredictedStage
0	Connecticut	Champions
1	Auburn	2ND
2	Houston	F4
3	James Madison	F4
4	Iowa St.	E8
...	...	...
59	McNeese St.	R64
60	Charleston	R64
61	Northwestern	R64
62	UAB	R64
63	Colorado	R64

64 rows × 2 columns

```
   TeamName PredictedStage
15  Purdue            S16
```

Split: 0.3

k=25,Purdue F4 acc = 0.48

```
Model accuracy: 0.4838709677419355
```

	TeamName	PredictedStage
0	Connecticut	Champions
1	Alabama	2ND
2	Purdue	F4
3	Baylor	F4
4	Arizona	E8
...	...	...
59	Colgate	R64
60	Akron	R64
61	Oregon	R64
62	Wagner	R64
63	Illinois	R64

64 rows × 2 columns

```
   TeamName PredictedStage
2   Purdue            F4
```

k=28,Purdue F4 acc = 0.48

```
Model accuracy: 0.4838709677419355
```

	TeamName	PredictedStage
0	Connecticut	Champions
1	Alabama	2ND
2	Auburn	F4
3	Purdue	F4
4	Arizona	E8
...	...	...
59	Colgate	R64
60	Morehead St.	R64
61	Duquesne	R64
62	Longwood	R64
63	Gonzaga	R64

64 rows × 2 columns

```
   TeamName PredictedStage
3   Purdue            F4
```

PURDUE UNIVERSITY

Modeling

Multinomial Logistic Regression

Prior to Grid Search

Classification Report:

	precision	recall	f1-score	support
2ND	0.00	0.00	0.00	1
Champions	0.00	0.00	0.00	1
E8	0.00	0.00	0.00	1
F4	0.00	0.00	0.00	0
R32	0.44	0.36	0.40	11
R64	0.67	0.89	0.76	18
S16	0.40	0.29	0.33	7
accuracy			0.56	39
macro avg	0.22	0.22	0.21	39
weighted avg	0.50	0.56	0.52	39

Applied hyperparameters:

Classification Report:

	precision	recall	f1-score	support
2ND	0.00	0.00	0.00	1
Champions	0.00	0.00	0.00	1
E8	0.00	0.00	0.00	3
F4	0.00	0.00	0.00	0
R32	0.43	0.50	0.46	6
R64	0.92	0.85	0.88	27
S16	0.00	0.00	0.00	1
accuracy			0.67	39
macro avg	0.19	0.19	0.19	39
weighted avg	0.70	0.67	0.68	39

Accuracy on the test set (top 9 features): 0.6667

PURDUE UNIVERSITY

Modeling

Random Forest

Prior to Grid Search

Classification Report:

	precision	recall	f1-score	support
R64	0.54	0.70	0.61	20
R32	0.09	0.14	0.11	7
S16	0.50	0.40	0.44	5
E8	0.00	0.00	0.00	5
F4	0.00	0.00	0.00	1
2ND	0.00	0.00	0.00	0
Champions	0.00	0.00	0.00	1
micro avg	0.41	0.44	0.43	39
macro avg	0.16	0.18	0.17	39
weighted avg	0.36	0.44	0.39	39

Accuracy: 41%

Applied hyperparameters:

Model Classification Report:

	precision	recall	f1-score	support
R64	0.58	0.75	0.65	20
R32	0.11	0.14	0.12	7
S16	0.33	0.40	0.36	5
E8	0.00	0.00	0.00	5
F4	0.00	0.00	0.00	1
2ND	0.00	0.00	0.00	0
Champions	0.00	0.00	0.00	1
micro avg	0.44	0.46	0.45	39
macro avg	0.15	0.18	0.16	39
weighted avg	0.36	0.46	0.40	39

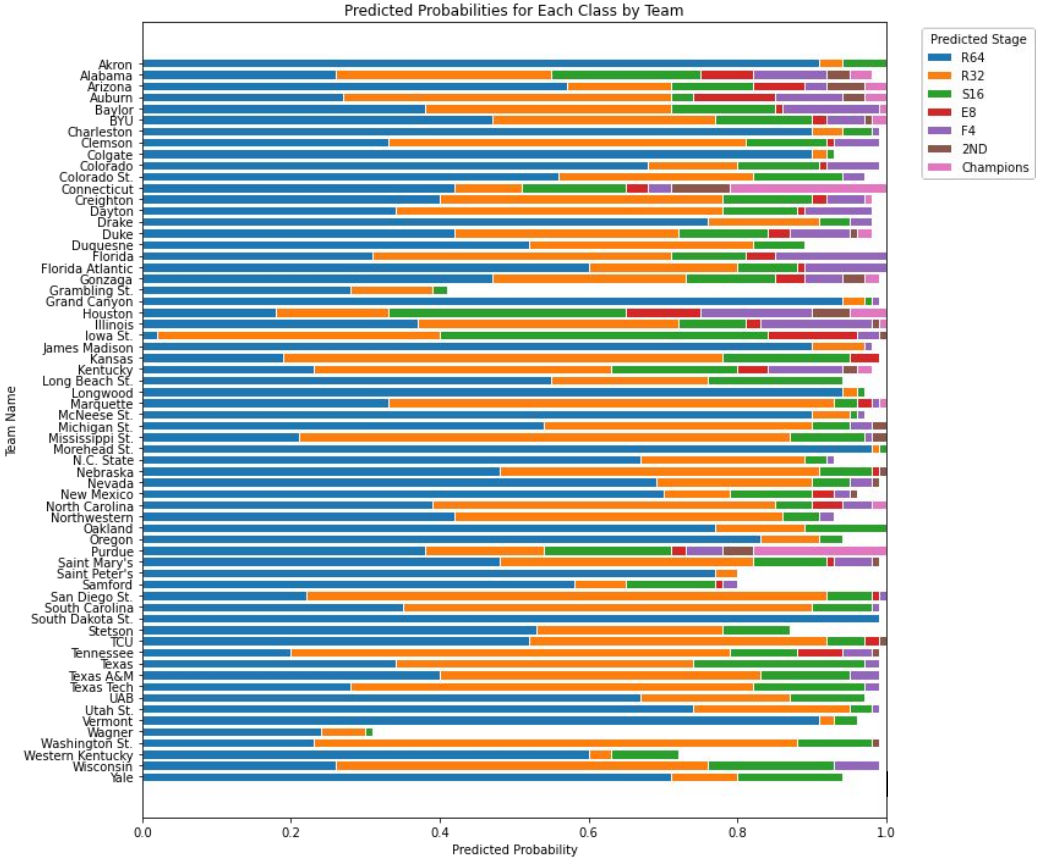
Accuracy: 44%

PURDUE UNIVERSITY

Takeaways

Unpredictability Obscures Findings

- MLR > SVM > KNN > RF
- Many teams have a decent chance to win it all
- Critical to identify highest probability candidate
- The nature of March Madness does not allow for anyone to make perfect predictions



PURDUE UNIVERSITY