Progress Presentation

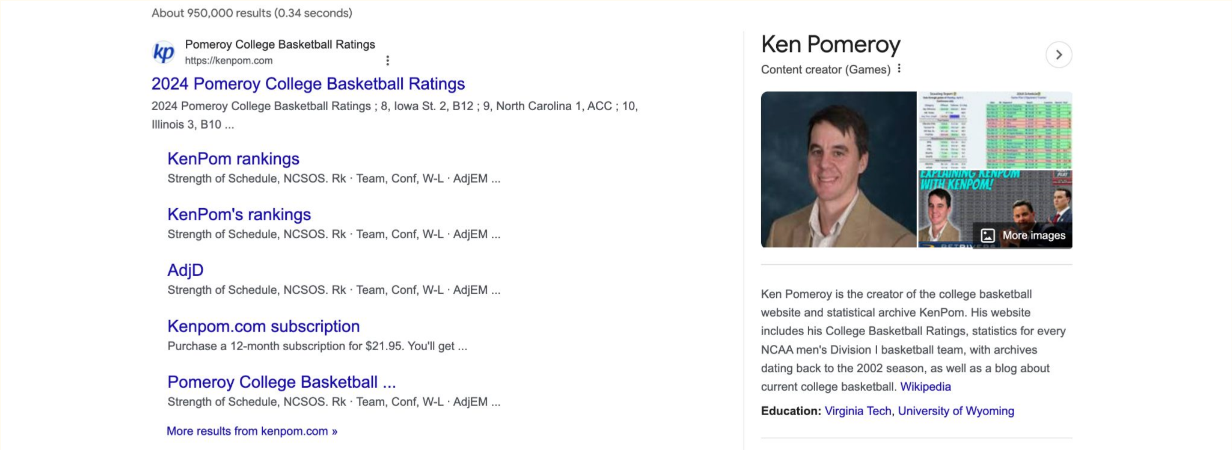
Cole Jennings and Thomas Sniezek

Tractable Data

- NBA Data
  - We added a new variable for making the playoffs (1 for making playoffs, 0 for missing playoffs)
- NCAA Data
  - We added a variable for making the NCAA Tournament (1 for making playoffs, 0 for missing playoffs)
  - We added a binary variable for what conference the team belongs to
  - Scraped Financial data from Sportico
    - OpEx and OpRev

Data Retrieval

- NBA Data was downloaded from Kaggle
  - This was scraped from stats.nba.com
- NCAA Data was downloaded from Kaggle
  - KenPom



Model Specification

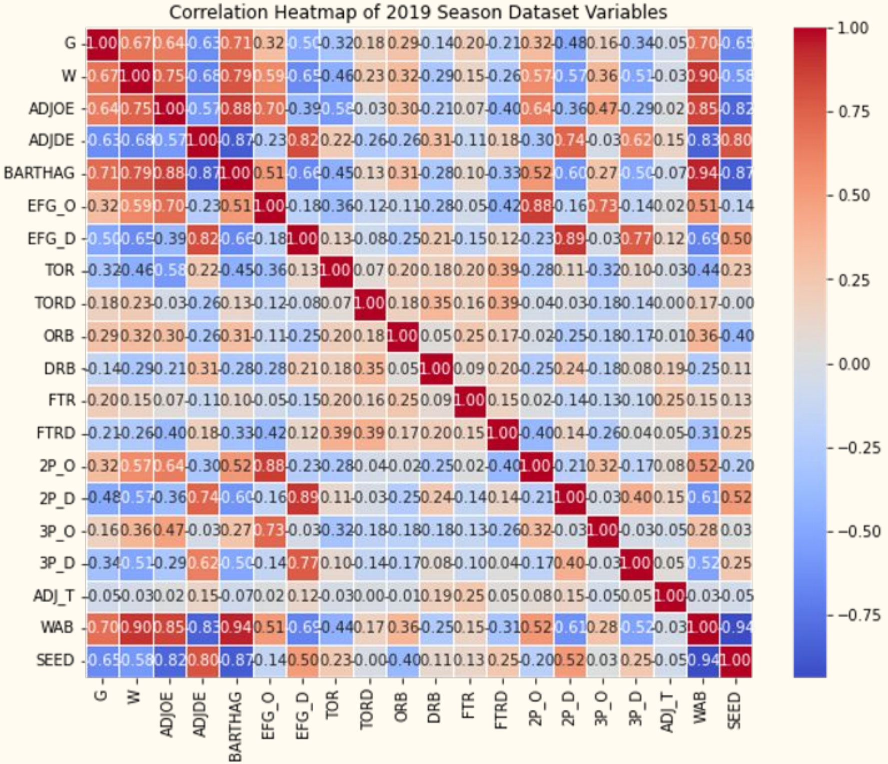
- Exploring the factors that are involved in making each league playoffs
- Create a XGBoost model or logistics regression that predicts if the team will make the playoffs based on specific variables
- Use the feature importances of the model to see what was deemed as the most important aspects to decide what teams will make the playoffs
- Compare the feature importances
- Compare the accuracies for each and tweak models until they are similar, explain what factors each model uses, show how these feature importances explain how the factors needed to make playoffs are similar/different

Variable Explanations - NCAA

- Adjusted defensive and offensive efficiency (ADJDE and ADJOE)
- ADJOE – Points scored per 100 offensive possessions
- ADJDE – Points allowed per 100 defensive possessions
- Possessions are not recorded officially by statisticians, so estimated using:
- FGA-OR+TO+0.475xFTA
- EFG_O – (FGM + 0.5*3PM) / FGA (on offense)
- EFG_D – (FGM + 0.5*3PM) / FGA (on defense)

NCAA Variables

- Variables of note
    - ADJOE
    - ADJDE
    - EFG_O
    - EFG_D



RESULTS AND MARGINAL EFFECTS

Logit Marginal Effects

Dep. Variable:	playoffs_binary
Method:	dydx
At:	mean

	dy/dx	std err	z	P>|z|	[0.025	0.975]
EFG_O	0.0074	0.003	2.241	0.025	0.001	0.014
EFG_D	-0.0072	0.004	-1.663	0.096	-0.016	0.001
TOR	-0.0068	0.004	-1.645	0.100	-0.015	0.001
ADJOE	0.0094	0.002	3.978	0.000	0.005	0.014
ADJDE	-0.0079	0.002	-3.544	0.000	-0.012	-0.004

Logit Regression Results

Dep. Variable:	playoffs_binary	No. Observations:	716
Model:	Logit	Df Residuals:	710
Method:	MLE	Df Model:	5
Date:	Fri, 16 Feb 2024	Pseudo R-squ.:	0.4663
Time:	19:42:01	Log-Likelihood:	-185.76
converged:	True	LL-Null:	-348.08
Covariance Type:	nonrobust	LLR p-value:	5.017e-68

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.5941	5.908	-0.608	0.543	-15.173	7.985
EFG_O	0.1520	0.070	2.187	0.029	0.016	0.288
EFG_D	-0.1475	0.086	-1.715	0.086	-0.316	0.021
TOR	-0.1401	0.085	-1.641	0.101	-0.307	0.027
ADJOE	0.1916	0.038	5.023	0.000	0.117	0.266
ADJDE	-0.1621	0.044	-3.703	0.000	-0.248	-0.076

Calculating interpretations:

(dy/dx)/mean of y variable = interpretation

A one percentage point increase in EFG_O is associated with 3.9% increase in the likelihood of a team making the NCAA tournament

Each additional turnover per 100 plays is associated with a 3.6% lower likelihood of a team making the NCAA tournament

XGBOOST MODEL

- Model makes predictions based on multiple iterations of decision trees
- Each DT builds on the last's shortcomings
- 86% Accuracy, ADJOE and ADJDE are most determinant

```
...  Accuracy on the test set: 0.8605

     Feature Importances:
     ADJOE: 0.2570
     ADJDE: 0.2337
     EFG_O: 0.0462
     EFG_D: 0.0230
     TOR: 0.0517
     TORD: 0.0444
     ORB: 0.0427
     DRB: 0.0326
     FTR: 0.0582
     FTRD: 0.0206
     2P_O: 0.0320
     2P_D: 0.0372
     3P_O: 0.0381
     3P_D: 0.0511
     ADJ_T: 0.0314
```

```python
# Split the data into features (X) and the target variable (y)
X = concatenated_df.drop(['playoffs_binary', 'TEAM', 'CONF', 'POSTSEASON', 'SEED', 'G', 'W', 'WAB', 'BARTHAG'], axis=1)
y = concatenated_df['playoffs_binary']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Create an XGBoost classifier
model = xgb.XGBClassifier(objective='binary:logistic', random_state=42)

# Train the model
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy on the test set: {accuracy:.4f}')

# Optionally, you can also analyze feature importances
feature_importances = model.feature_importances_
print('\nFeature Importances:')
for feature, importance in zip(X.columns, feature_importances):
    print(f'{feature}: {importance:.4f}')
```
✓ 0.3s

Stakeholder Implications

The stakeholders are the coaches for both the NBA and NCAA teams, the players for both the NCAA and NBA teams, the GM for the NBA teams, the fans,

College athletics generate large amounts of money for their respective universities. Those working in college athletics will be interested to see which factors are associated with making playoffs.

Making it to March Madness generates significant publicity for universities

Ethical

Bias and Fairness: We need to ensure the model is built without any bias towards certain teams or players as biased models could perpetuate inequalities or stereotypes

Privacy: We need to make sure the data that we have collected does not contain any sensitive information about individuals or teams

Accountability: We need to ensure that we are transparent on hour our model works and that we are responsible for the outcomes

Legal

Data Privacy: We need to make sure all of the data we are using are compliant with all data protection laws

Intellectual Property: We need to make sure we do not infringe on any intellectual property rights so we can avoid any and all legal disputes

Discrimination Laws: We need to ensure that the model does not discriminate against any certain groups and we do not violate any anti-discrimination laws

Social

Equity in Sports: We can promote equity and access in sports by figuring out and understanding what factors contribute to success in basketball at different levels

Talent Development: Insights from the model could inform players and coaches on how to develop talent among basketball programs and could lead to more effective training

Economic Impact: Understanding the factors that contribute to success in basketball can lead to investments in different basketball programs and new economic development through sports

Next Steps

Run XGBoost and Logistic Regression for NBA data

Draw comparisons with NCAA