Project 11
Predicting Purdue's Performance in the NCAA Tournament Bracket
Evaluation Report

Saul Means, Kuba Bal, Sarah Firestone, Zachary Hanson, Avi Khurana

Presentation Effort:

Saul Means 33%   Kuba Bal 33%   Sarah Firestone 33%

Report Effort:

Kuba Bal 20%   Saul Means 20%   Sarah Firestone 20%   Avi Khurana 20% Zachary Hanson 20%

1. EVALUATE RESULTS

Business objectives:

Problem: Predict how Purdue will perform in the NCAA tournament

Impact: Viewership, ticket sales, national ranking, improve Purdue's basketball program

Currently: Rely on coaching, real time analysis, and video footage

Our Solution: Create a model, predict number of games Purdue wins

Success: If within an error of 1 round

Primary – Predict how Purdue will perform in the NCAA tournament
Secondary – Identify why our model predicts Purdue performance in the way it does and use these business findings to identify areas in need of improvement for the team

Improving Purdue's performance in the NCAA tournament would lead to increases in profit in multiple areas including ticket sales, concession sales, and Purdue basketball merchandise.

Business success criteria:

Our success criteria was defined by correctly predicting the number of games Purdue will play in the tournament with an error of 1. Although the Purdue Athletics department is unlikely to use our suggestions in their strategies for the tournament, we also defined success as being able to identify particular items that our team would have been able to improve upon within the tournament.

Data mining goals:

This data mining project is a classification and prediction problem – to predict Purdue's chances of winning the 2024 NCAA Tournament, given previous teams' performances with similar regular season statistics.

Data mining success criteria:

Our model will be assessed based on the accuracy of its predictions in respect to how the Purdue Basketball team will perform during the NCAA tournament. Our model will use accuracy, precision, recall, F1-score, to evaluate its performance.

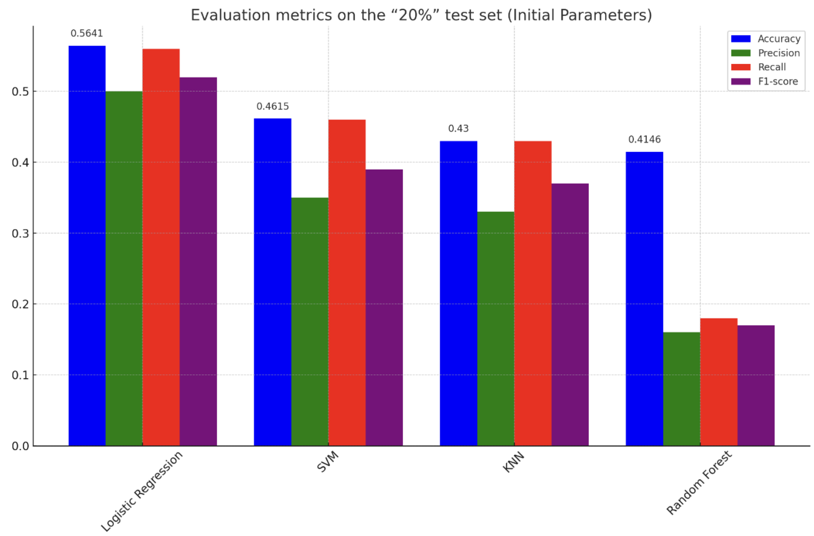
We will require that accuracy, precision, recall, and F1-score be greater than 0.80.

1.1. Assessment of data mining results with respect to business success criteria

Purdue's chances of winning the 2024 NCAA Tournament, given previous teams' performances with similar regular season statistics. So we needed to use supervised machine learning models to achieve this classification goal. We decided to use Logistic Regression because of an article from the Journal of Quantitative Analysis in Sports, stating the simplest models often perform the best. We then chose K Nearest Neighbors and Random Forest since external research told us that these models perform the best for our type of sports analytics problem. Finally, Support Vector Machines were chosen as the final model since it is a popular classification method that has been proven successful in other problems. The results are below from each model.

Below are the evaluation metrics for the initial parameters. The weighted macro-average of accuracy, precision, recall, and the F1-score are recorded in the table for each model. We then visualized the results in a plot below the table. Our accuracy is around 0.5 which is acting as almost random. So although the models met our success criteria for correctly predicting Purdue's performance within an error of 1 round, it did not meet the data mining criteria of being over 0.8.

	Evaluation metrics on the "20%" test set (Initial Parameters)			
	Accuracy	**Precision**	**Recall**	**F1-score**
Logistic Regression	.5641	.5	.56	.52
SVM	0.4615	0.35	0.46	0.39
KNN	0.43	0.33	0.43	0.37
Random Forest	0.4146	0.16	0.18	0.17

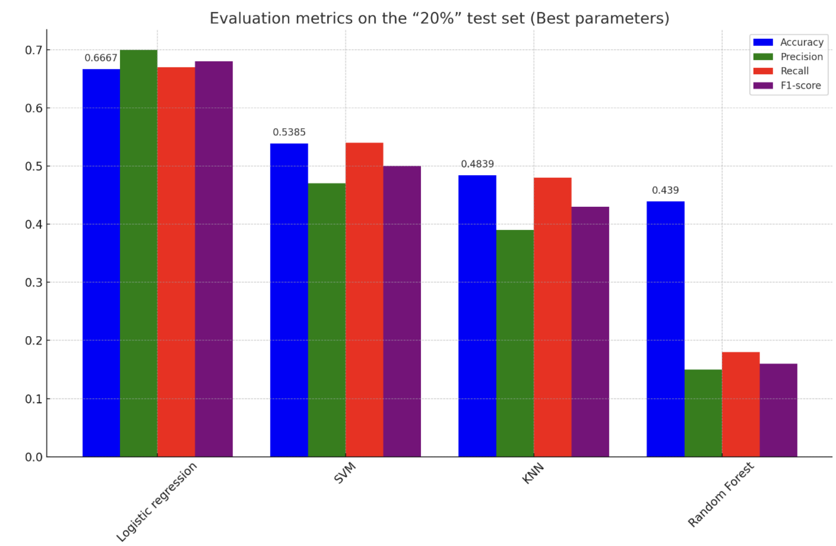


The top 5 features for the best model (Logistic Regression) are:

1) AdjDE
2) AdjEM
3) SEED
4) DE
5) W

Since our accuracy was almost equivalent to random, we fine-tuned the hyperparameters and went back to the data preparation stage of adding the Vegas data as a feature, etc. We then wanted to compare the evaluation metrics for the best parameters after fine-tuning and adjusting the data preparation. The weighted macro-average of accuracy, precision, recall, and the F1-score are recorded in the table below for each model. We then visualized the results in another plot below the table. The values were a little higher and were greater than 0.5 but still not to the 0.8 mark yet. The order of the best-performing models is still the same and consistent after adjusting for the best parameters.

	Evaluation metrics on the “20%” test set (Best parameters)			
	Accuracy	**Precision**	**Recall**	**F1-score**
Logistic regression	0.6667	0.70	0.67	0.68
SVM	0.5385	0.47	0.54	0.50
KNN	0.4839	0.39	0.48	0.43
Random Forest	0.439	0.15	0.18	0.16



The top 5 features for the best model (Multinomial Logistic Regression) are:

1) AdjEM
2) AdjDE
3) DE
4) SEED
5) W

These are the same top 5 features as before, however, they are in a different order with the hyperparameter tuning after the grid search.

As a final statement, this project meets the initial business objectives and success criteria. We wanted to predict Purdue's performance with an error of 1 round. SVM, KNN, and Logistic Regression predicted Purdue to be eliminated in the Final Four which is 1 round below the actual round they got eliminated in (the Championship round), which is a success based on our original criteria. For Random Forest, we correctly predicted the correct round Purdue will be eliminated and it correctly predicted Purdue will get 2nd place against UConn. This is really impressive that all models either predicted correctly or within an error of 1 round. It also predicted some of the other teams correctly in the Final Four, with all models predicting UConn to win which was correct. Our results agreed with the prior research stating the simplest models have the best performance as they had the highest accuracy, however, we saw Random Forest had the lowest accuracy but was the only model to predict Purdue's performance without the error. However, our accuracy was lower than our original criteria, with none of the models having an accuracy of 0.8. These results could not be used for colleges or businesses who have a lot riding on these results since we did have an accuracy close to random (even if it did predict within an error of 1 round). Since our models either predicted that Purdue would make it to the Final Four or the championships, we are still counting this a success based on our business criteria. We acknowledge the evaluation metrics are a difficult issue to fix, but we do wish our accuracy was higher than random. However, we are happy that we correctly predicted Purdue's performance which was our main objective.

In conclusion all models were successful by predicting Purdue's performance based on our criteria but none achieved our desired accuracy. Our final rankings were Logistic regression > SVM > KNN > Random Forest.

1.2. Approved models

All the models meet the selected criteria in regards to our business success criteria by correctly predicting the number of games Purdue will play in the tournament with an error of 1. Only Random Forest predicted Purdue to get 2nd place in the Championship game, however, the other models (KNN, Logistic Regression, and SVM) predicted Purdue to get eliminated in the Final Four which is a success based on our criteria. All these models also had the best parameters.

In terms of the data mining success criteria, we required that accuracy, precision, recall, and F1-score be greater than 0.80. We did not meet that criteria for any of the models as most evaluation metric scores were a little higher than 0.50. Our results improved a little by using the best parameters however were still not to our 0.8 minimum.

We will still approve Logistic Regression, KNN, Random Forest, and SVM since they all correctly met our business criteria by correctly predicting Purdue's performance.

2. REVIEW PROCESS

The overview of the data mining process:
1) Business Understanding: First, we had to understand what the problem was and why they wanted the problem to be solved. We had to understand the topic of college basketball before we got started with touching any of the data
2) Data Preparation and Processing: Next, we had to collect all of the data. From there we removed duplicate features and features that were highly correlated. We also had to combine two different dataset to get the greatest number of features.
3) Predictive Model Training: Then, we built 4 classification models. We then trained hyperparameters and applied platt scaling to the models to get the most accurate results. After we trained the hyperparameters and applied platt scaling we would then look at where the model had placed the Purdue men's basketball team in the NCAA tournament.

All of the steps we took were integral to the process and necessary to solve the problem that we were given. We did not execute everything optimally as we were still learning how to become better data scientists. It could have been improved by running feature importance earlier rather than later when training our models on 30+ features. We ended up running feature importance and slimming our dataset down to at least 9 or less features.

We also failed to reach the target of 75% accuracy, however, our logistic regression model got very close.

Even though we did not reach the target of 75% we did achieve our goal of predicting the round that Purdue got to in this year's tournament with an error of 1 round. The majority of our models predicted that Purdue would get to the final 4.

One of the unexpected paths we faced was finding the same exact features for this year's tournament. The data we had gathered was from Kenpom and Kaggle. These sites contained data for the past years tournament however, when this year's tournament began, some of the features were missing on these sites causing us to have to manually add a few features.

3. DETERMINE NEXT STEPS

3.1. List of possible actions

Option 1: Our model can be practically deployed and used to predict the outcomes of future NCAA March Madness tournaments. Although not entirely correct, our model made fairly reasonable predictions given the highly random nature of the March Madness tournament.

Option 2: Although the model performed fairly well, there was certainly room for improvement. In particular, we could have located and assessed new data sets for possible feature additions. By the time we combined all the various sources of data we used, we only ended up using 5 features for our predictions. By locating more data sets with different features, we may have come upon features that explained larger portions of the variance in the target variable. We also could have found a way to increase the size of our training data set. For the purposes of this project, we only ended up using data from 2021 onward. We excluded previous data because we did not want to include data from COVID years and the years before contained data on a Purdue basketball team that was almost entirely different from the team we had this year. However, if we had come up with a way to use the data from those previous years, we would have had access to far more years of training data and our model would have likely been more accurate.

Option 3: The last option would have been to run a different kind of model. Something we thought of towards the end of the project was to run a game by game model instead of running a model that predicts the final outcome of the entire tournament. Our model did not take into account which specific teams played against who because our main focus was finding out where Purdue ended up. If we had instead predicted the outcome of each game separately, we would likely have a model that predicts the outcome of the tournament with much more accuracy. This, in turn, would allow us to predict with more accuracy where Purdue ended up in the end.

3.2. Decision

Of our three options, it would seem that we can rule out the last option. While it would certainly be an interesting experiment to see how our results differ, we would need to make entirely different models and essentially redo the entire project. This is not exactly feasible, and the improvement we would see from doing this would not be worth it. After all, many of our models predicted Purdue to end in the final 4 and some even predicted correctly that Purdue would get second place.

Next, it comes down to either improving the model or deploying it. While improving the model may seem like the obvious option, deploying the model, as is, makes more sense in actuality. We talked a bit about this in our final presentation, but the double edged blade of March Madness is its consistent unpredictability. While this aspect makes for an engaging tournament for fans to enjoy, it makes it very difficult to predict the outcomes of each match. For that reason, we believe that improving the model will not help us to have a better result. While we may technically be improving the model, our predictions will likely be upset with similar rates each year. Our final decision is Option 1: to deploy our current model.