Samuel Kellum and Ethan Perello

Mentor: Prof. Mattei

CMPS 4010-01

Prof. Zheng

November 18, 2022

<https://drive.google.com/drive/folders/1D44chd6yC5m9Ov_EL4HK4udhtZvCzf6_>

At this point in time, we have completed our first model, which is relatively simple in terms of both the number of features we have incorporated and the supervised machine learning algorithms we implemented. For our first model, we incorporated data from 34,550 regular season games and 333 tournament games since 2016, in addition to seeding data and FiveThirtyEight ratings data.

We aggregated the data to compute the number of wins, number of losses, average margin of victory, and average margin of defeat for each team in each year. Since we wanted to compute the probability of Team A beating Team B, we engineered four features: seed difference, FiveThirtyEight ratings difference, win percentage difference, and average win margin difference.

After engineering the features, we wanted to test our data on two different machine learning algorithms: linear regression and logistic regression. The linear regression approach would compute the outcome probabilities by predicting the score gap, whereas the logistic regression approach would compute outcome probabilities by predicting the team that wins.

We used a k-fold validation method, where k is the number of seasons minus one. The reason for this is because training data consists of all of the data from previous seasons, and the validation data represents an individual season. For example, we can not validate for the year 2016 since there is no data preceding it. Below is a table showing the log loss scores of the two models:

| **Year** | **Linear Regression** | **Logistic Regression** |
| --- | --- | --- |
| **2017** | 0.580 | 0.547 |
| **2018** | 0.596 | 0.593 |
| **2019** | 0.530 | 0.509 |
| **2021** | 0.607 | 0.614 |
| **Average** | **0.578** | **0.566** |

One significant limitation our first model had was the number of years we were backtesting our model with, since we decided to use FiveThirtyEight ratings data, which only dates back to 2016. To solve this problem, we created a scraper to collect data from kenpom.com’s Pomeroy ratings data, which dates back to 2002. Additionally, if we choose to use more advanced machine learning algorithms that require a greater number of features, we will need to account for that change in our future models.

After discussion with our faculty mentor, we decided to build two new models, where each of us will individually focus on building one model. We expect to complete our new models within the next couple of weeks. The first model will use the XGBoost algorithm, and the second model will be a classification model incorporating historical sports betting odds.

XGBoost stands for Extreme Gradient Boosting, which is an ensemble model. It uses multiple decision trees to make a prediction. Models from previous years in the competition have been very effective using XGBoost so we decided to see how effective it would be for us.

With the current features we have, the XGBoost model achieved an average log-loss score of 0.578. This is not an improvement on our previous models, but it will be interesting to see if XGBoost is more effective with more and/or different features. After plotting the decision tree we saw that the FiveThirtyEight Ratings Difference was the only feature being used so adding more ratings as features will likely improve our XGBoost model.

We plan on experimenting with different parameters to potentially improve the XGBoost model as well. We have determined that the ideal max depth of the tree is two with the current features that we have, but that may change as we add more features. We plan on further experimenting with different tree structures to see if it improves our model. We will also try changing the regularization term and the learning rate. These parameters will likely have more of an impact as we add more features.

Our faculty mentor suggested building a model that incorporates sports betting odds, since the betting odds are computed through a variety of factors such as computer algorithms, rankings, injuries, and home-court advantage (when applicable). Additionally, a betting line has to attract equal action on both sides. Therefore, any betting odds that do not reflect reality will quickly change to a number that reflects reality through market forces. This is one possible way to make our model unique since we did not find any previous successful submissions that incorporated betting odds.

Our first model was completed as expected in the initial timeline. However, in our initial timeline, we wanted to have a second model built by November 7th and our third model built shortly after the end of Thanksgiving break. At that time, we intended to work together on both models. However, we decided it would be better if each of us focused on building one model, and having our models completed by our next faculty mentor meeting. The end result is the same, since we will still have three models completed shortly after the end of Thanksgiving break. However, we will be able to work more asynchronously with this schedule. Additionally, this schedule gave us more time to build scrapers to collect data we otherwise would not be able to use.

As mentioned above, in our most recent faculty meeting, we decided how to expand on the progress made in our first model with two new models. Our next faculty meeting will occur after Thanksgiving break before the end of semester presentations begin. At that meeting, we will review the new models that will be completed by then, and plan for the end of semester presentation in addition to a meeting schedule for the Spring semester.