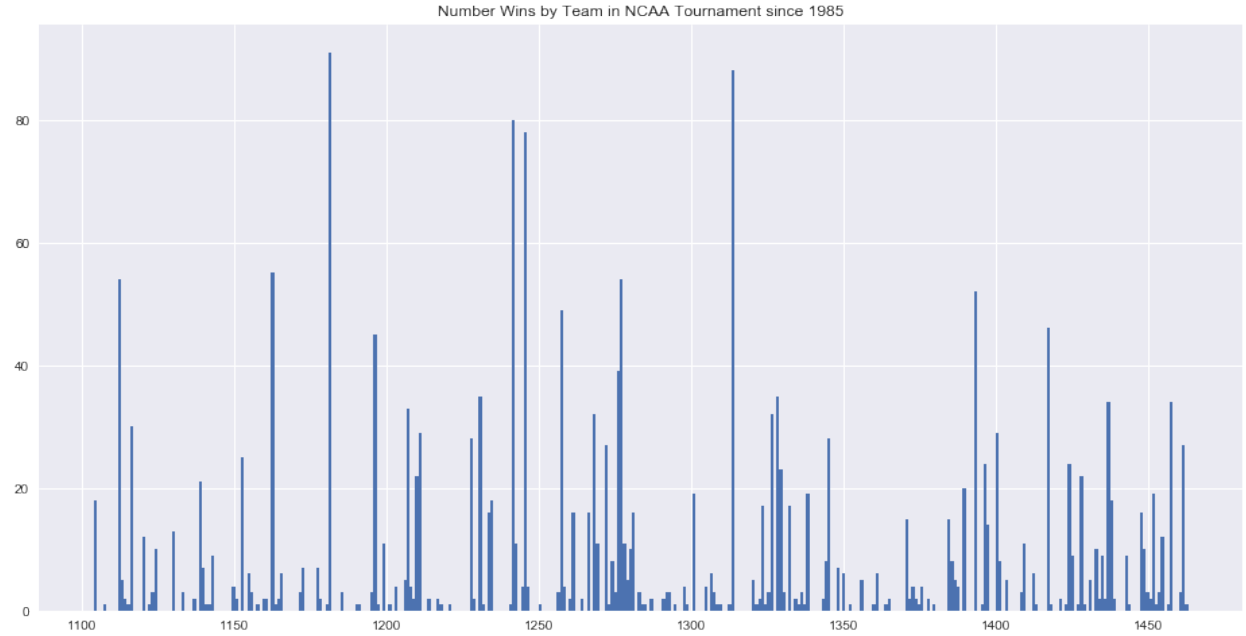
CS 401R Final Project Write-up

1. **A discussion of the dataset**
    a. Where did it come from? Who published it?
        1. The data came from a Kaggle competition from Google Cloud and the NCAA, as they teamed up to give very precise and clean data, easy for people to manipulate and use. Millions of people around the country create a bracket as they attempt to predict who will win the NCAA Men's Basketball tournament.
    b. Who cares about this data?
        1. The data is used every day by sports statisticians and athletes to understand their game and team. Each data set gives information concerning past game statistics or field goal percentage, past tournament game results and specific values relating to team and conference ID's in the data set. Supposedly, people are offering $1 million to someone who can create a perfect bracket, but so far no one known has been able to. So technically, if someone could predict a perfect bracket off this data, they could pocket $1 million and do it the year after if it is consistent and reliable.
2. **A discussion of the problem to be solved**
    a. Is this a classification problem? A regression problem?
        1. Initially, as we thought on past algorithms we discussed in class, we first thought of applying the MNIST algorithm to the data to give an average model to base our next predictions off of. But, as we discussed with Greg, he brought up logistic regression, which allows us to find specific constants for each data point we have to fit to any two-team possibility that will give us the most approximate average.
    b. Is it supervised? Unsupervised?
        1. This problem is definitely supervised because there is a definite input and output. We are inputting data concerning number of wins, all-time wins in the conference, field goal percentage, etc. and are expecting a specific probability output of winning a game.
    c. What sort of background knowledge do you have that you could bring to bear on this problem?
        1. I spend a lot of time watching basketball, especially in college. As I talked about the lab with Mark Rose, we noticed that 6 conferences in the country really dominate college basketball. There are also certain teams like Kansas, Kentucky, Duke and North Carolina that have won significantly more NCAA tournament games than other teams around the country (in the past two years, 3 of these teams have made the Final Four). So we included these weights in our logistic regression algorithm that helped it become more accurate.
    d. What other approaches have been tried? How did they fare?
        1. Looking at other techniques online, many people also used logistic regression but with k-nearest neighbors and random forest models to predict upsets that would happen between teams, especially in the first two rounds. A group tested their algorithm with the 2017 first-round data and it predicted the winner 75% of the time which is quite high especially with many upsets happening in the first round.
3. **A discussion of your exploration of the dataset**.
    a. Before you start coding, you should look at the data. What does it include? What patterns do you see?
        1. I looked a lot at the data before starting to code with it. Here is where I noticed the incredibly large difference of the NCAA 'Tigers' in 2.c. that have significantly more wins than other teams. These few teams seemed to dominate the tournament, as their win percentage is much higher than most others. Additionally,, the top twenty teams all time for number of wins in the tournament were all from the big six basketball conferences. These were the two important finds from the data, then there were also significant differences in field goal percentage, free throw percentage, average turn-overs and rebounds between high winning teams and low winning teams.
    b. Any visualizations about the data you deem relevant



4. **A clear, technical description of your approach.** This section should include:
   a. Background on the approach
      1. The algorithm was first developed by David Cox in 1958, first created to estimate a binary response based off variables and changing probabilities.
   b. Description of the model you use
      1. Logistic regression is used for finding specific constants from training data that can be used to find similar probabilities for test data. This is predictive analysis to describe data and explain the relationship between variables across themselves. We used this algorithm for binary classification, as we compared two specific outcomes and took the highest.
   c. Description of the inference / training algorithm you use
      1. We used training data from all past data, with all-time wins dating back to 1985 and season wins concerning 2017. We had a plethora of data, and ran the algorithm on the 5000 games happening throughout the season, not on the tournament or conference playoff games. Here we predicted using free throw percentage, whether or not it was a major conference team, rebounds, all-time tournament wins, turnovers and wins this season.
   d. Description of how you partitioned your data into a test/training split
      1. We used data concerning matchups that would occur during the tournament. We didn't really split the data because the test data was by itself.
5. **An analysis of how your approach worked on the dataset**
   a. What was your final RMSE on your private test/training split?
      1. My final RMSE came out around 1.135, which is fairly well for this algorithm but not fantastic. The highest percentages were only around .53 because it is so difficult and borderline impossible to create a perfect bracket!
   b. Did you overfit? How do you know?
      1. I definitely think I overfit the data because of the large weight on tournament wins since 1985 and on major conferences because overtime, conferences change as do programs. Some teams from the past won large games around 20 years ago but since their program changed between coaching or conference switches, that data is quite relevant today. The teams with the most all time tournament wins including Duke, North Carolina, Kansas and Kentucky all made it to the Final Four, thus the algorithm is a little biased towards them.
   c. Was your first algorithm the one you ultimately used for your submission? Why did you (or didn't you) iterate your design?
      1. Yes, we ultimately used this algorithm because we brainstormed for quite a while and made sure this algorithm would work best for it. We spent most of our time trying different b constants to run in the algorithm that we could test each team against. In this sense, we did iterate our design, but we were satisfied with its convergence.
   d. Did you solve (or make any progress on) the problem you set out to solve?
      1. We definitely made progress on our mission for a perfect bracket! Although our quest to predict a perfect bracket wasn't entirely true, we did actually predict the winner of last year's tournament! Thankfully it was North Carolina, one of the big winners, but it also

helped me see what data is most important for a team winning the tournament. It was fascinating and quite fun to work on, and showed me what other aspects I'd like to include in the future if the data was available like injured players returning for the tournament and winningest coaches.