Undergraduate Thesis Prospectus

Analyzing the Creation of the March Madness Bracket with a Machine Learning Approach

(technical research project in Computer Science)

Methods of Phone Scam Prevention in the United States

(sociotechnical research project)

by

Andrew Cornfeld

October 27, 2023

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Andrew Cornfeld

Technical advisor: Briana B. Morrison and Rosanne Vrugtman, Department of Computer Science

STS advisor: Peter Norton, Department of Engineering and Society

General Research Problem

How can machine learning systems improve safety and other performance criteria?

Machine learning systems are quickly becoming more prevalent in society. Some may see them as harmful, but they can be used to solve problems and improve the safety of civilians. These algorithms can recognize patterns in speech and caller activity to save nearly $40 billion in 2022 by preventing phone scams. These algorithms are also designed to be preconfigured; given some inputs, they can use a learned function to predict an outcome with reasonable confidence. This is evident in the creation of the March Madness bracket, where a formula adding up wins and losses factoring in caliber of competition can result in precise seed prediction.

Analyzing the Creation of the March Madness Bracket with a Machine Learning Approach

How can machine learning be used to predict the admission and seeding of the NCAA basketball tournament?

Bracketology is the process of predicting and creating the March Madness tournament bracket, which is not well understood and complex. The March Madness bracket is selected by a committee of twelve members. To predict tournament bids and their respective seeds, I utilized Machine Learning techniques to create a model which analyzes a team's resume to determine its seed placement. The advisors are Rosanne Vrugtman and Briana Morrison. I am in the Computer Science department. The project is an independent project I worked on this past summer. The project goals are to gain understanding of and find a formula for creating the March Madness tournament bracket.

Although many have used machine learning techniques to predict outcomes of tournament games, there is less existing research into the creation of the bracket. Lunardi (2022)

is the most well-known of bracketologists, and his book on bracketology suggests there is no explicit formula for what a committee any given year favors. Some members may look heavily into box scores and predictive metrics, while others simply refer to the Associated Press rankings to make their decisions. Lunardi is often consulted by programs to evaluate their schedule. He realizes smaller programs often cannot schedule the winningest competition (Duke, Kansas, etc.) and recommends they schedule other small programs with recent success. It is well understood that a win against a strong team, especially on the road, greatly improves your resume, and a home loss against a weak team is detrimental, but there is no quantification to compare their impact.

Strack (2023) proposed another change to bracket construction, which involves minimizing travel distances to game sites while maintaining a fair bracket. He analyzed the NCAA tournament rules for seeding and created a penalty value analysis which would lead to fewer required flights. This would work to prevent situations similar to one in the 2023 NCAA Tournament where West Virginia was assigned to play in Birmingham, AL, while same-seeded Florida Atlantic was assigned to Columbus, OH. Geography plays a major role in bracket creation, with the 1st overall seed getting venue preference.

The Kaggle dataset the model is trained on is the 2023 March Madness data, which includes all games from the 2023 season and their outcomes. It also includes the seed each invited team received and their overall seed (EX: one team could be a 1-seed in the south region, but the 2nd overall seed). The ranking of each team is based on the NET (NCAA Evaluation Tool) rankings on Selection Sunday, the day the committee's bracket is released. The model uses linear regression techniques to determine the seeding impact of each type of win or loss. In this way, it creates a formula which allows the number of wins and losses in each quadrant to

calculate a seeding value, which can be stack ranked against other teams in contention. All code will be written in Python in a Jupyter Notebook file. I will train the data on the 2023 season dataset, and test the model on the 2022 season dataset.

At the end of the project, I will have an understanding of the value of each type of win or loss. Directors of scheduling for teams could use this data to understand why strong schedules are essential for an at-large bid (invitation to March Madness without winning a conference tournament), and effectively schedule to pursue a tournament bid. The model will also predict seeding midseason, allowing teams to know standings and requirements for reaching the tournament. Future work could include tournament projections in other sports, such as football, soccer, field hockey, etc.

Methods of Phone Scam Prevention in the United States

In the US, how do telecom companies, federal agencies and scammers compete to protect or subvert telecommunications security?

Truecaller, which has a spam call blocking app, estimates scammers have made nearly $40 billion in 2022 from nearly 70 million victims (Truecaller, 2022). The average reported loss has gone up since 2021, and the percentage of people reporting being scammed has increased. Several apps like Truecaller have been developed, such as T-Mobile's Scam Shield and Robokiller, but each has its own issues, including missing legitimate calls because of inaccurate spam filtering. It is not sufficient anymore to rely on intuition to discern scams. Scammers are quickly getting better at posing as real companies, coworkers, and family to steal your information and money. Americans don't want to waste time engaging with scams, which is where machine learning is used for detection.

Participants in this problem include machine learning engineers, telecom companies, federal agencies, mobile phone carriers, scam-savvy mobile phone users, scam-unsavvy mobile phone users, and scammers. Machine Learning Engineers train their models on caller data and can use them to detect activity that could be fraud (Rudolph, 2022). Telecom companies want to prevent fraudulent accounts from stealing from other customers (Apple, 2023). These companies are also motivated by profit, and collecting the customer's data is valuable to sell. Federal agencies like the FCC encourage phone companies to block any unwanted calls based on reasonable call analytics (FCC, 2023). Many mobile phone carriers like T-Mobile have apps like Scam Shield which offer scam protection and options to change your number if necessary to protect their customers (T-Mobile, 2023). These carriers are motivated to give you free devices for your recurring payment to their network, which can collect data from devices used around you.

Scam-savvy mobile users have an agenda of avoiding scams, but also educating those around them. They recognize concerning things but may be unsure of how to discuss topics with those unaware, or let them know before they become victims and lose money. One such scam-savvy mobile user, Cameron Huddleston, mentioned an experience where her mother, who was suffering from Alzheimer's, was told by a scammer to wire money to claim a prize (2023). She said, "That was a wake-up call for me. If you have any cognitive decline, you don't see those red flags anymore."

Scammers often use social engineering to trick victims into revealing sensitive information. These are subdivided into three main classes: Technical, Social, and Physical (Salahdine, 2019). Examples of a technical attack include creating a fake banking website and requesting credit card information, or creating a virus pop-up which alerts the user they have a

computer virus (which doesn't exist) and they must call the scammer to fix it. A social attack could be calling as tech support to fix a computer bug, connecting to the computer to steal information. Examples of physical attacks include walking into locked buildings behind people with access, or dumpster diving through thrown away but not effectively destroyed documents.

Scams like these have existed long before telephones or email. In the 16th century, Spanish criminals used trade directories to mail contacts in richer countries (Okosun, 2022). They would send letters detailing how they knew of huge stashes of gold or money, but they had a loved one held hostage as a prisoner and needed the victim's financial help, and then the scammer would send over information about where the money was. A similar scam still exists today, known as the Nigerian prince scam. Both scams tell the victim how good of character they have to be selected to receive the money, and all scams require the victim to keep all information private.

Scammers are often competing with local governments to remain hidden or bribing them to let them stay in business. In Nigeria, 40% of people live below the poverty line, and the gap between rich and poor is so large that the only efficient way to make money is through scamming. Some scammers justify their actions by referencing the slave trade and colonial rule, and consider it "taking back what belonged to our forefathers". Scammers learn that to stay untouchable they need the right social and political connections to those in power, and they won't be held accountable.

Li et al. (2018) have developed a mobile app called TouchPal, which collects minimal amounts of data from calls and requires the user to tag the call from various options. Kubilay et al. (2023) created an experiment which analyzes the public's ability to classify scams from genuine messages, and noticed age and higher than secondary education were positively

correlated with correct identification. Mahoney (2015) analyzes the “Do Not Call List” and attributes its failure to scammers' ability to “spoof” their numbers and make calls from overseas. These scammers do not respect the Do Not Call List and aren’t afraid to get caught.

References

CBS Interactive. (2023, April 10). How to protect elderly parents from financial scams. CBS News. https://www.cbsnews.com/news/money-scams-elder-fraud-abuse/

Dema, A. (2023, August 28). App Store stopped more than $2 billion in fraudulent transactions in 2022. Apple Newsroom. https://www.apple.com/newsroom/2023/05/app-store-stopped-more-than-2-billion-in-fraudulent-transactions-in-2022/#:~:text=For%20example%2C%20with%20Apple%20Pay,714%2C000%20accounts%20from%20transacting%20again

Halladay, K. (2022, October 6). How do phones identify potential spam calls?. Built In. https://builtin.com/machine-learning/spam-calls

Help with scams, Spam, and fraud. T-Mobile. (2023). https://www.t-mobile.com/support/plans-features/help-with-scams-spam-and-fraud#:~:text=call%20is%20about!-,Scam%20Shield%20app,Store%20or%20Apple%20App%20Store.

Kubilay, E., Raiber, E., Spantig, L., Cahlíková, J., & Kaaria, L. (2023, February 2). Can you spot a scam? measuring and improving scam identification ability. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4344411

Li, H. (2018, July 26). A Machine Learning Approach To Prevent Malicious Calls Over Telephony Networks. IEEE Explore. https://ieeexplore.ieee.org/abstract/document/8418596

Lunardi, J., Smale, D., & Few, M. (2022). Bracketology: March madness, college basketball, and the creation of a national obsession. Triumph Books.

Mahoney, M. (2015, November). Dialing Back: How Phone Companies Can End Unwanted Robocalls. Consumers Union. https://advocacy.consumerreports.org/wp-content/uploads/2015/02/Dialing-Back-Complete-Report-11.16.2015.pdf

Okosun, O., & Ilo, U. (2022, October 17). The evolution of the Nigerian prince scam. Journal of Financial Crime. https://www.emerald.com/insight/content/doi/10.1108/JFC-08-2022-0185/full/pdf?title=the-evolution-of-the-nigerian-prince-scam

Salahdine, F., & Kaabouch, N. (2019). Social Engineering Attacks: A Survey. Future Internet, 11(4), 89. https://doi.org/10.3390/fi11040089

Stop unwanted robocalls and texts. Federal Communications Commission. (2023, July 7). https://www.fcc.gov/consumers/guides/stop-unwanted-robocalls-and-texts#:~:text=Under%20the%20Truth%20in%20Caller,to%20%2410%2C000%20for%20each%20violation

Strack, M. (2023, March 21). Team Assignment and Location Determination for the NCAA March Madness Tournament. NC State University Libraries. https://repository.lib.ncsu.edu/bitstream/handle/1840.20/40863/etd.pdf?sequence=1

Truecaller. (2022). Truecaller insights 2022 U.S. Spam & Scam Report. Truecaller Blog. https://www.truecaller.com/blog/insights/truecaller-insights-2022-us-spam-scam-report