Using Machine Learning to Generate Meaningful Vacation Rental Property Recommendations
(Technical project)

The Socio-Political Implications of the Use of Facial Recognition Technology by the US Government
(STS project)

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Vaidic Naik

December 9, 2022

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.

ADVISORS

MC Forelle, Department of Engineering and Society

Brianna Morrison, Department of Computer Science

Intro

In the field of Computer Science, there has been a growing interest in machine learning and its many applications. At a rudimentary level, machine learning takes in data and works to find trends in it to make meaningful predictions. There are different types of machine learning models that are best suited for dealing with different types of data. They also make use of different algorithms to help create these meaningful predictions and convert input data into an output that is interpretable by a user. Part of the job of a data scientist is choosing the correct machine learning model that is best suited for the problem they are trying to solve. A large phase of developing a machine learning algorithm that will make meaningful predictions is training the model with labeled data. This process involves feeding the model inputs that have prelabeled outputs which allow the model to get better at recognizing such patterns. Like any technology, there are moral and political implications that come with the use of the technology (Winner,). With federal agencies making use of machine learning in various ways it is important to understand if there are any underlying biases that may negatively impact a certain subgroup. This is especially important at a time when there is a growing sentiment of distrust towards the government from the public (Silva & Kenny, 2018). Part of the issue at hand deals with the lack of information on machine learning. Not only is there a gap of knowledge in the government when trying to use machine learning to help automate different processes but also in the public when trying to understand why there may be a bias.

In this paper, I will be discussing two separate applications of machine learning. The technical portion deals with reflecting on my previous internship experience and how it allowed me to gain insight into the use of machine learning to help generate revenue for a company. Specifically, how I worked on a team to help make recommendations for similar vacation rental

properties given an input of an Airbnb property. The STS portion of this paper will discuss the socio-political implications of the use of facial recognition technology by federal agencies and law enforcement. With both topics, we see that the impact of the use of machine learning models differs drastically based on the context they are being used. Throughout this paper, we will discuss how it is important to understand how your machine-learning model is trained and how it is making predictions.

Technical Topic

For the technical portion of this paper, I will be reflecting on a previous internship experience. For my internship I worked at Capital One as a part of their Capital One Shopping Department. Within the Capital One Shopping department I worked on the travel team which developed a price matching tool to help users searching for hotels find better deals for the same hotels through Capital One's partners. For our project, the Travel team wanted to expand this product into the space of vacation rental properties specifically Airbnb properties. The problem that we were tasked with solving was finding vacation rental properties from Capital One's partners that will be deemed similar enough to the Airbnb property that the user is viewing so they will choose our recommendation. The problem of figuring out what is deemed "similar" to an Airbnb property required us to employ machine learning models to compare a variety of features.

The team I worked on comprised of another intern and two other software engineers who acted as mentors for the project. The initial code-base was very minimal which allowed us to decide which features we would compare and which machine learning models we would use to rate their similarities. The original pipeline for our project comprised of a simple SQL query that traversed through the scraped data to only match properties that were within a certain radius of

Airbnb properties. We found that this original method of filtering properties wasn't very good at meaningful recommendations. The features we decided to compare were size (rooms, bathrooms, max occupancy), amenities, name similarity, description similarity, and distance. The data we had to work with for our project was acquired through extensive web scraping of Airbnb properties and properties from Capital One's partners which we labeled as canonical properties.

To compare the size and amenities we decided to use the same machine learning model k-nearest neighbors. K-nearest is a form of unsupervised learning that calculates the distance between two points given a list of features. We decided to generate two separate scores for these features since we wanted to weigh them differently. One of the challenges we found when working with our data for amenities was that some of our data entries didn't include data on all the amenities we had listed. So, when preprocessing this data, we decided to assume that the property probably didn't have the amenities in question, so we filled the null data.

The next set of features we compared were the names and descriptions of the properties. For the description we used Doc2Vec which is a Natural Language Processing library that works by vectorizing words based of nuance and context. Using Doc2Vec we extensively trained our model using all the descriptions we had scraped this way we were able to match for similarities in the descriptions that may include information of lodge type, if the property was family friendly etc. For comparing the names of the properties, we realized that we could get away with using a pretrained open source NLP model called spacy. It was important that we did proper research into how the model was trained and how it worked before implementing into our pipeline to ensure that we would optimize performance.

For the distance between properties, we decided to scale our distance based of the density of properties. Specifically, we valued the distance between two properties less if they are in a

high-density area compared to a low-density area. We computed density by looking at the number of properties in a 10-mile radius of the property and inversely scaled our score based on that number.

After calculating our five scores we made our final score by taking a weighted average of scores. We originally did this by looking at labeled data to see what features users tended to value higher. We also started to work on a way to customize our weights based of the features that the individual user was focused on. Our final pipeline was able to make five times more accurate predictions when compared to our labeled data that was scored by a variety of users.

STS Topic

On the topic of machine learning and its many applications, one of the applications employed by law enforcement and federal agencies is that of facial recognition technology. The use of this technology is widespread but the reliance on this technology has brought up several questions about its ethics. One of the main concerns with the federal governments reliance on such technologies is that there have been studies that have shown that facial recognition technologies are more likely to falsely identify certain racial groups. Specifically black women tend to be disproportionately misidentified by facial recognition technologies used by law enforcement (Bacchini & Lorusso, 2021). This ties into the technical and social portion of my topic since we see how the machine learning models behind facial recognition technologies can have some inherent racial biases. These biases can negatively impact racial groups that are already subject to racial discrimination within the law enforcement system.

I will be studying the developments of facial technology through the lens of technological citizenship. Specifically, I will be looking at how the developments of the technology has

affected the rights of the citizens. Technological citizenship deals with understanding how technologies impact the polity of citizens and how they interact with its outcomes (Andrews 2006).

It is important to understand the deep-learning models at play and see how the training of these models may be leading to this problem. There has been a long-lasting distrust in the law enforcement and judiciary system that has existed for decades, especially when talking about racial bias (Tyler & Huo, 2002). Recent events of police brutality that have heightened tensions between the public and law enforcement. This has become more severe with the rise of social media in recent years making these atrocities known around the nation and created a call to reform the law enforcement agencies. With these sentiments of distrust from the public it is important to understand the source of bias in facial recognition technologies. Facial recognition technologies use deep learning convolutional neural networks which are trained using many faces to help detect geometries in certain facial features (Gates, 2011). The network then compares new input to see if there is a match with the current set of faces that are already in a database. So, the problem of misidentification may be from the disproportionate representation of certain marginalized groups in the training data (Verma, 2018). It is also worth mentioning how heavily federal agencies rely on these technologies to facilitate their investigation. This problem is usually considered on understanding the severity of repercussions of misidentifying a suspect. Studies have found that using these technologies alongside human forensic examiners minimize the margin of error and bias (Philips et al., 2018).

Back on the issue of the bias within the facial recognition technology used by federal agencies the problem also lies in accountability. Most federal agencies will use a third-party facial recognition system without properly considering the issue of bias. These systems are

highly accurate, but it should be the government's responsibility to better ensure that they understand the shortcomings of these systems. By having a better understanding of the systems, they are using federal agencies and law enforcement can take better precautions to mitigate bias. It is also important to understand that the inaccuracies of facial recognition technology and the biases that come with its use are inevitable to a certain extent. As the technology improves the rate of inaccuracies decrease however it is still a factor that is important to be considered when dealing with racial biases.

Research and Methods

The research question I am aiming to answer is as follows: to what extent does implicit bias impact the use of facial recognition technology by the US government? This is an important question because it deals with understanding if it is justifiable for a government to use a technology that could help optimize the process of enforcing the law, even if there is a chance that there may be implicit bias in the technology itself. It is also in its relevance to the current times the issue of facial recognition is a recent one that is still being discussed to this day. The benefit of a technology that can easily identify suspects in a video with high accuracy is valuable to the government in maintaining justice. But it is important to make sure that the people who are using such technologies understand how the technology works to mitigate the biases that may come from its use. The way I will go about answering this question is primarily through literature reviews. I will analyze the different facial recognition models that are being used by multiple federal agencies and the extent to which there are inaccurate predictions. I also will look at recent legislation that was passed regarding the use of facial recognition technology by the government and see what they have sourced as the reason for or against the continued use of the

technology. The data that will be collected will be from performance reports of these facial recognition technologies. It will be important to not only consider the trends of which groups are misidentified but also which facial recognition system is being used.

Conclusion

Both the technical and STS portion of this paper dealt with the various use cases of machine learning. The technical aspect focused on exploring and learning different machine learning models. This helped me understand the process of training a machine learning model. This understanding helped me make a more informed discussion in my STS portion as I went into understanding why there is a bias in the facial recognition technologies being used by law enforcement. The problem with this bias is that it targets communities that are already negatively impacted by law enforcement and creates a negative light on the use of machine learning. Part of mitigating this is by increasing the understanding of how these machine learning models within the government. This way we can be better about training the model and a more informed public will understand that implicit racial biases in facial recognition technology may be unavoidable to a certain extent because of the data that the model is trained on.

References

Andrews, C. J. (2006). Practicing technological citizenship. *IEEE Technology and Society Magazine*, *25*(1), 4.

Bacchini, F., & Lorusso, L. (2019). Race, again: How face recognition technology reinforces racial discrimination. *Journal of Information Communication & Ethics in Society*, *17*(3), 321–335. https://doi.org/10.1108/JICES-05-2018-0050

Collectif, C. (2018). *Research Ethics in Machine Learning* (p. 51) [Research Report]. CERNA ; ALLISTENE. https://hal.archives-ouvertes.fr/hal-01724307

Deck, C. A., & Wilson, B. J. (2002). The Effectiveness of Low Price Matching in Mitigating the Competitive Pressure of Low Friction Electronic Markets. *Electronic Commerce Research*, *2*(4), 385–398. https://doi.org/10.1023/A:1020567515249

Gates, K. (2011). *Our biometric future: Facial recognition technology and the culture of surveillance*. New York University Press.

Hein, (2021). *Facial Recognition Technology: Federal Law Enforcement Agencies Should Better Assess Privacy and Other Risks (gao-21-518)* (Internet materials). United States. Government Accountability Office. http://RE5QY4SB7X.search.serialssolutions.com/?V=1.0&L=RE5QY4SB7X&S=JCs&C=TC_0046728234&T=marc

Liddy, E. D. (n.d.). *Natural Language Processing*. 15.

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., Chen, J.-C., Castillo, C. D., Chellappa, R., White, D., & O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(24), 6171–6176.

Racial Disparities in Law Enforcement Stops. (n.d.). Public Policy Institute of California. Retrieved October 27, 2022, from https://www.ppic.org/publication/racial-disparities-in-law-enforcement-stops/

Ray, S. (n.d.). *A Quick Review of Machine Learning Algorithms | IEEE Conference Publication | IEEE Xplore*. Retrieved October 28, 2022, from https://ieeexplore.ieee.org/abstract/document/8862451

Samet, H. (2008). K-Nearest Neighbor Finding Using MaxNearestDist. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(2), 243–252. https://doi.org/10.1109/TPAMI.2007.1182

Silva, S., & Kenney, M. (2018). Algorithms, Platforms, and Ethnic Bias: An Integrative Essay. *Phylon (1960-)*, *55*(1 & 2), 9–37.

Spencer, K. B., Charbonneau, A. K., & Glaser, J. (2016). Implicit Bias and Policing. *Social and Personality Psychology Compass*, *10*(1), 50–63. https://doi.org/10.1111/spc3.12210

T.G. (2019). Facial Recognition: Transcending Bias. *ASEE Prism*, *29*(4), 10–10.

The Social Psychology of Racially Biased Policing: Evidence-Based Policy Responses—Kimberly Barsamian Kahn, Karin D. Martin, 2020. (n.d.). Retrieved October 28, 2022, from https://journals.sagepub.com/doi/abs/10.1177/2372732220943639

Tyler, T. R., & Huo, Y. J. (2002). *Trust in the Law: Encouraging Public Cooperation With the Police and Courts* (Internet materials). Russell Sage Foundation. http://proxy01.its.virginia.edu/login?url=https://muse.jhu.edu/book/19346

United States Congress House Committee on Oversight and Reform. (2019). *Facial Recognition Technology: Hearing Before the Committee on Oversight and Reform, House of Representatives, One Hundred Sixteenth Congress* (Internet materials). U.S. Government Publishing Office. https://purl.fdlp.gov/GPO/gpo124209

Verma, N., & Dombrowski, L. (2018). Confronting Social Criticisms: Challenges when Adopting Data-Driven Policing Strategies. *Proceedings of the 2018 Chi Conference on Human Factors in Computing Systems (Chi 2018)*. https://doi.org/10.1145/3173574.3174043

Winner, L. (n.d.). *Do Artifacts Have Politics?* 17.

Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M. C., & Dafoe, A. (2021). Ethics and Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers. *Journal of Artificial Intelligence Research*, *71*, 591–666. https://doi.org/10.1613/jair.1.12895