Datasheet for NHL player statistics and Salary for the 2023-24 season as of 12 April 2024*

My subtitle if needed

Gavin Crooks

Invalid Date

The NHL Player Evaluation Dataset, compiled by Gavin Crooks, offers insights into the performance and contract details of 448 top-performing NHL skaters during the 2023-24 season. Derived from reputable sources like the NHL website and Hockey-Reference, the dataset enables analysis of player value based on statistics. While it presents opportunities for tasks such as performance prediction and salary estimation, users must exercise caution to avoid unfair treatment and protect player privacy. Transparency and ethical considerations are crucial for responsible analysis, ensuring the dataset's integrity and reliability for research in professional sports.

Extract of the questions from Gebru et al. (2021).

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
   - The dataset was created to enable analysis of NHL players. Specifically to estimate the value of players in terms of salary based of their staistics during the 2023-24 season. This dataset is indended to allow individuals to complete analysis that NHL team front offices must consider when signing players.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - Gavin Crooks, while studying at the University of Toronto.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

[*] Code and data are available at:https://github.com/Crooksyyy/NHL-Player-Evaluation

- No funding was used in the creation of the dataset. The dataset uses publicly available statistics from the NHL (https://www.nhl.com/stats/skaters) and Hockey-Reference (https://www.hockey-reference.com/friv/current_nhl_salaries.cgi).

4. *Any other comments?*

   - The dataset is not inclusive of all players but 448 of the top 500 scorers during the 2023-24 season. The datawas reduced from 500 to 448 from the availability of salary information. It does not include goalies as they require seperate analysis from skaters.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

   - The instances in the dataset represent NHL (National Hockey League) players. Each instance corresponds to a specific player in the NHL, capturing various attributes such as games played, goals, assists, points, plus/minus, salary, and cap hit for the 2023-24 season. The dataset focuses on skaters, including forwards and defensemen, but excludes goalies.

2. *How many instances are there in total (of each type, if appropriate)?*

   - The dataset comprises 448 instances of NHL players, each representing a unique player in the league. Each instance includes various attributes such as games played, goals, assists, points, plus/minus, salary, and cap hit for the 2023-24 season. There is only one type of instance in this dataset, which is NHL skaters. Each skater instance includes performance metrics and contract details for a specific player in the league.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

   - The dataset is a sample of NHL players from the 2023-24 season rather than containing all possible instances. The larger set would encompass all NHL players from that season. The sample is not random but rather comprises the top 448 skaters based on their performance metrics during the specified season. While the sample size is relatively small compared to the total number of NHL players, it is representative of the higher-performing players in terms of offensive and defensive contributions. The representativeness of the sample is validated by the selection criteria, which prioritize players with notable statistics and contributions to their teams. However, it may not capture the entire spectrum of player performance and contract dynamics, especially for players with lower statistical profiles or those in specialized roles such as goaltenders.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

    - Each instance in the dataset consists of raw data. The raw data includes various statistics and contract details for NHL players from the 2023-24 season. Specifically, each instance includes the following features:

Games Played: The number of games the player participated in during the season. Goals: The total number of goals scored by the player during the season. Assists: The total number of assists made by the player during the season. Points: The total number of points (goals + assists) accumulated by the player during the season. Plus/Minus: The plus/minus rating of the player, indicating the goal differential when the player is on the ice. Salary: The total compensation received by the player for the season, including signing bonuses. Cap Hit: The average annual value of the player's current contract, representing their salary cap hit for the season. These features provide insights into each player's performance on the ice as well as their contractual status within their respective teams.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

    - Yes, there is a target associated with each instance in the dataset. The target variable is the "Cap Hit," which represents the average annual value of the player's current contract. It is the main focus of analysis in the dataset and serves as the target variable for regression modeling. The goal is to predict the cap hit of NHL players based on their performance metrics and other relevant factors such as games played, goals, assists, and plus/minus.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

    - Yes, some complex statistics may be missing from individual instances in the dataset to keep the data simple and manageable. For example, detailed play-by-play data or advanced analytics like Corsi or Fenwick statistics may not be included. This decision could be made to streamline the dataset and focus on more commonly available performance metrics such as goals, assists, points, and plus/minus. Additionally, certain contract details or player attributes that are not readily accessible or publicly available may also be missing from individual instances. This missing information could be due to privacy concerns, contractual confidentiality, or data limitations.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

    - No, relationships between individual instances are not made explicit in this dataset. Each instance represents a standalone observation of an NHL player's performance metrics and contract details for the 2023-24 season. There are no relational attributes or links provided that connect one player's data to another player's data or establish any form of network or interaction between instances. Each player's information is treated independently, without explicit relationships or connections to other players' data.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - No, there are no recommended data splits provided with the dataset. Users are required to devise their own data splitting strategy based on their specific needs and objectives. Depending on modeling task at hand, users may choose to partition the data into training, validation, and testing sets using various techniques such as random sampling. Each subset serves a distinct purpose in the model development process, including training the model and evaluating performance, respectively. It's essential to ensure that the chosen data splitting approach adequately represents the underlying distribution of the data and avoids introducing biases that could impact model performance.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - No, there are no specific errors, sources of noise, or redundancies in the dataset.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - The dataset relies on external resources for its creation, specifically the NHL website (https://www.nhl.com/stats/skaters) for player statistics and Hockey-Reference (https://www.hockey-reference.com/friv/current_nhl_salaries.cgi) for salary information.

a) There are no guarantees that these external resources will remain constant over time.

Changes to the structure or availability of these websites could impact the accessibility and accuracy of the data.

b) There are no official archival versions of the complete dataset including the external resources as they existed at the time of creation.

c) The NHL website and Hockey-Reference may have their own terms of use and restrictions associated with accessing their data. Users should review and comply with any applicable licenses, terms of use, or restrictions when accessing and using data from these external resources.

The dataset of the 448 players in available at https://github.com/Crooksyyy/NHL-Player-Evaluation/tree/main/data/analysis_data 11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.* - No, the dataset does not contain data that might be considered confidential. It includes publicly available statistics and salary information of NHL players for the 2023-24 season, which are not protected by legal privilege or other confidentiality measures.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - No, the dataset does not contain any offensive, insulting, threatening, or otherwise anxiety-inducing content. It consists solely of statistical information related to NHL players' performance and salary details for the 2023-24 season, which are unlikely to cause any such concerns.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - Yes, the dataset identifies sub-populations based on the NHL teams to which the players belong. Each player in the dataset is associated with a specific team, allowing for the identification of subpopulations based on team affiliation. The distribution of players across teams reflects the composition of the NHL during the 2023-24 season, with varying numbers of players representing each team.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
    - While the dataset includes information about individual NHL players, such as their performance statistics and contract details, it does not contain personally identifiable information (PII) beyond what is publicly available about these players. Therefore, it is not possible to directly identify individuals from the dataset. However, if combined with other external sources containing PII, such as player biographies or

personal profiles, it may be possible to indirectly identify individuals. Nonetheless, the dataset itself does not facilitate direct identification of individuals.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - The dataset primarily consists of NHL player statistics and contract information, which do not inherently contain sensitive information such as race or ethnic origins, sexual orientations, religious beliefs, political opinions, or other personal attributes. However, it does include financial data related to players' salaries and cap hits. While this financial data is not typically considered sensitive in a broader context, it may still be subject to privacy considerations within the context of contractual negotiations and player management in professional sports. Nonetheless, the dataset does not include other types of sensitive information such as health data, genetic data, government identification numbers, or criminal history.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data associated with each instance was acquired from publicly available sources, namely the NHL website (https://www.nhl.com/stats/skaters) and Hockey-Reference (https://www.hockey-reference.com/friv/current_nhl_salaries.cgi). These sources provide comprehensive statistics and salary information for NHL players, which were then compiled into the dataset. The data was directly observable as it represents factual information such as games played, goals, assists, points, plus/minus, salary, and cap hit for each player. There was no need for validation or verification as the data was obtained from reputable sources known for providing accurate and up-to-date information about NHL player statistics and contracts.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The data collection process involved downloading information directly from the NHL website and Hockey-Reference. No complex mechanisms or procedures were used; rather, the data was obtained through manual human curation by accessing the websites and downloading the relevant statistics and salary information. Since the data was obtained from official sources known for their accuracy and reliability, there was no need for additional validation procedures beyond ensuring that the data was correctly downloaded and compiled into the dataset.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The dataset is a sample derived from a larger set of NHL players. The sampling strategy used was deterministic, specifically targeting the top 500 scorers from the 2023-24 NHL season for whom salary data was available reducing the data to 448 players. This deterministic approach ensured that the dataset included players who were statistically significant in terms of their scoring performance, while also having salary information accessible for analysis.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - The data collection process myself, Gavin Crooks, a student at the University of Toronto. No compensation was provided for the data collection process as the data was obtained from publicly available sources, namely the NHL website and Hockey-Reference. Therefore, no crowdworkers or contractors were involved in the data collection process.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - The data was collected on April 12, 2024, capturing the most up-to-date information available at that time. This timeframe aligns with the creation timeframe of the data associated with the instances, ensuring that the dataset reflects the latest statistics and contract details for NHL players during the 2023-24 season.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - No ethical review processes were conducted as part of the data collection for this project.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - The data was obtained from third-party sources, specifically the NHL website and Hockey-Reference.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

   - There was no direct notification to the individuals whose data is included in the dataset, as the data was publicly available on the NHL website and Hockey-Reference. These websites typically do not provide individual notifications to players regarding the collection of their statistics for public access.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

   - As the data was collected from publicly available sources such as the NHL website and Hockey-Reference, there was no explicit consent obtained from the individuals regarding the collection and use of their data. These websites provide statistics and salary information for NHL players as part of their regular operations, and users accessing this data are typically bound by the terms of service and privacy policies of these websites.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - Since consent was not obtained for the collection and use of the data, there was no mechanism provided for individuals to revoke their consent.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - No, analysis of the potential impact of the dataset and its use on data subjects, such as a data protection impact analysis, has been conducted.

12. *Any other comments?*

    - Professional athletes, including NHL players, are often under scrutiny in the public eye, and various analyses, including those related to their performance and contracts, are common occurrences.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of*

instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

- Yes, preprocessing and cleaning of the data were conducted as part of the analysis. This involved merging and consolidating data from multiple sources, handling missing values, and selecting relevant variables for the analysis. For instance, missing salary information for some players resulted in the exclusion of those instances from the final dataset. Additionally, to simplify the analysis, certain complex statistics were omitted from the dataset.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

   - Yes, the raw data was saved in addition to the preprocessed, cleaned, and labeled data. The raw data can be accessed at this link https://github.com/Crooksyyy/NHL-Player-Evaluation/tree/main/data/raw_data, or directly from the websites from which it was obtained: NHL website and Hockey-Reference.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

   - The software used for preprocessing, cleaning, and labeling the data is R. You can find more information and access R via the official website: The R Project for Statistical Computing. Additionally, any scripts or code used for data preprocessing may be available in the GitHub repository associated with the project. https://github.com/Crooksyyy/NHL-Player-Evaluation/tree/main/scripts

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

   - Yes, the dataset has been used for analyzing NHL player contracts and estimating player value based on performance metrics. Specifically, regression analysis was conducted to predict players' true value against their team's cap hit. This analysis aimed to provide insights into the economics and dynamics of NHL player contracts, including identifying potential flaws or areas for improvement in current valuation methods.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

   - Yes, the repository containing the dataset also includes the paper that uses the dataset. You can access the repository at the following link: https://github.com/Crooksyyy/NHL-Player-Evaluation/tree/main

3. *What (other) tasks could the dataset be used for?*

- The dataset could be used for various tasks related to NHL player analysis and evaluation, such as:

4. Player performance prediction: Developing models to predict future performance metrics based on historical data.
5. Salary estimation: Building regression models to estimate player salaries based on performance metrics and salary cap implications.
6. Player comparison: Conducting comparative analysis between players to identify strengths, weaknesses, and overall value to their respective teams.
7. Injury risk assessment: Exploring correlations between player performance, injury history, and salary to assess injury risk and inform player management decisions.

These tasks can provide valuable insights for NHL teams, analysts, and fans to better understand player dynamics, team compositions, and overall league trends.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

   - Given that the dataset focuses on NHL players and their performance metrics, there are several considerations for future use to ensure fair treatment and mitigate potential risks or harms:

5. Avoiding unfair treatment: Dataset consumers should be cautious not to use the data in a way that unfairly stereotypes or discriminates against individual players or groups based on factors such as nationality, race, or ethnicity. Care should be taken to interpret the data within the context of sports performance and avoid extrapolating to unrelated domains.

6. Protecting player privacy: While the dataset primarily consists of publicly available statistics, consumers should respect player privacy and refrain from using the data in a manner that could compromise individual privacy rights. Avoiding the dissemination of sensitive personal information beyond what is publicly available is essential.

7. Ensuring responsible analysis: Dataset consumers should conduct analyses and draw conclusions from the data in a responsible and transparent manner, considering the limitations and biases inherent in sports statistics. Robust validation and sensitivity analyses can help mitigate potential biases and ensure the reliability of findings.

8. Acknowledging limitations: It's important for dataset consumers to recognize the limitations of the data, such as potential errors or inaccuracies in the reported statistics, as well as the exclusion of certain player attributes or performance metrics. Transparent

reporting of methodologies and assumptions can aid in understanding the scope and applicability of analyses.

9. Promoting ethical use: Encouraging ethical use of the dataset within the sports analytics community can help foster a culture of fairness, integrity, and respect for players and teams. Open dialogue and collaboration can facilitate the sharing of best practices and mitigate potential risks associated with the misuse of sports data.

By adhering to these considerations and best practices, dataset consumers can help ensure that the use of the NHL player dataset is conducted responsibly and ethically, promoting positive outcomes for players, teams, and stakeholders in the sports community.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

    - The NHL player dataset should not be used for purposes that could result in harm to individuals or groups, violate privacy rights, or perpetuate unfair treatment. Specifically, the dataset should not be used for:

6. Discriminatory practices: The dataset should not be used to discriminate against individual players or groups based on factors such as nationality, race, ethnicity, or other protected characteristics. Any analysis that could lead to unfair treatment or bias against players should be avoided.

7. Unethical profiling: Dataset should not be used to create or perpetuate stereotypes about players or teams that could lead to unfair treatment or biased perceptions. Care should be taken to analyze the data in a manner that respects the dignity and integrity of all individuals involved.

8. Privacy violations: The dataset should not be used in a way that compromises the privacy rights of players or exposes sensitive personal information beyond what is publicly available. Any analysis or dissemination of player data should adhere to applicable privacy laws and regulations.

9. Misleading conclusions: The dataset should not be used to draw misleading or inaccurate conclusions about player performance, team dynamics, or other aspects of the NHL. Any analysis should be conducted with integrity and transparency, avoiding the dissemination of false or misleading information.

10. Unauthorized commercial use: The dataset should not be used for commercial purposes without proper authorization or consent from relevant stakeholders, such as the NHL, player associations, or other rights holders. Any commercial use of the data should adhere to applicable licensing agreements and contractual obligations.

By avoiding these inappropriate uses, the dataset can be utilized responsibly and ethically to generate meaningful insights and support informed decision-making in the context of sports analytics and player evaluation.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

    - Yes, the dataset will be distributed to third parties outside of the entity on behalf of which the dataset was created. It will be made available on GitHub for public access and use. Users will be able to download the dataset and utilize it for various purposes, including research, analysis, and educational projects. The dataset will be distributed under appropriate licensing terms, ensuring that users understand their rights and obligations when accessing and using the data. Additionally, documentation and guidelines may be provided to help users understand the dataset's structure, variables, and potential applications.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - The dataset will be distributed via GitHub at this link https://github.com/Crooksyyy/NHL-Player-Evaluation/tree/main/data, where it will be hosted in a repository accessible to the public. Users will be able to download the dataset files directly from the repository. At the moment, the dataset does not have a digital object identifier (DOI), but one could potentially be assigned in the future to provide a persistent and citable reference for the dataset.

3. *When will the dataset be distributed?*

    - The dataset is currently available and can be accessed on GitHub.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

    - The dataset will be distributed without any copyright or intellectual property license, as well as without any applicable terms of use.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - No, there are no IP-based or other restrictions imposed by third parties on the data associated with the instances.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

   - No, there are no export controls or other regulatory restrictions that apply to the dataset or individual instances.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

   - The dataset will be supported, hosted, and maintained by the individual responsible for creating it, which is Gavin Crooks.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

   - Gavin can be contacted at via email gavin.crooks@mail.utoronto.ca

3. *Is there an erratum? If so, please provide a link or other access point.*

   - There is no information provided about an erratum for the dataset.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

   - There are no current plans to update the dataset.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

   - There are no applicable limits on the retention of the data associated with the instances.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

   - There are no plans to support or maintain older versions of the dataset. However, older data remains accessible at the original websites, as they provide historical data.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

- If others wish to extend, augment, build on, or contribute to the dataset, they can reach out to me directly for collaboration. Contributions will be welcomed and considered for inclusion in future versions of the dataset. However, there is currently no formal process for validating or verifying contributions. Upon receiving contributions, they will be reviewed, and if deemed appropriate, they will be communicated and distributed to dataset consumers through the same channels as the original dataset.

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. “Datasheets for Datasets.” *Communications of the ACM* 64 (12): 86–92.