Chronic Homelessness AI Model – Privacy, Consent and Ethical AI Requirements

Prepared by: Matt Ross, Manager, Artificial Intelligence, Information Technology Services, City Manager's Office, City of London

Last updated: October 26, 2020

Data Collection, Privacy and Information Security:

1) The CHAI model is built using the HIFIS database and application. HIFIS is a Homeless Information Management System (built by the federal government) which records homeless services interactions (e.g. shelter stays, other social service interactions and demographic attributes). The data collection procedures and the underlying database is an explicit consent driven database, meaning there is an explicit consent given by any individual before data collection.

2) All organizations using HIFIS have executed data sharing agreements to enable the use of this information in a de-identified format for these purposes.

3) Due to the sensitivity of the data in the HIFIS database, we took steps to ensure that all data extracted even in the earliest prototyping and design stages of the project was de-identified.

4) The risk predictions given by the model are linked to a Client ID which then are mapped back to a name. This re-identification process only happens within the HIFIS application/database enabling only those who have access to that identified data already to view the named predictions. This consistent management of access permissions is critical to protect individuals privacy and information security.

5) This system is deployed in a cloud environment utilizing our enterprise scale network authentication for facilitating access to the machine learning model.

6) The cloud machine learning system only has access to de-identified data meaning any ongoing maintenance, re-training of the model, and monitoring of performance activities don't provide any access to identified data.

7) Database connections between the HIFIS database and the cloud machine learning system (utilizing Azure Machine Learning Studios) are one-way, meaning if the machine learning system was compromised by cyberattack, access to the identified HIFIS database is denied by design. Our Information Security team has been involved at several points in this project, from pre-prototype to final production deployment to ensure information security standards were met or exceeded.

8) Before pushing the final model into production, as a secondary double-check, we had a data science support vendor review all code and deployment architecture

for machine learning, development operations and security best practices to ensure maximum information security of the production system.

Design and Model Development and Ethical AI:

1) As I mentioned above, at the outset of the project, we decided to act as if the Treasury Board of Canada's Directive on Automated Decision Making applied to our project. We conducted an Algorithmic Impact Assessment receiving a Level 2 impact assessment. We scoped the requirements of the system to include all relevant risk mitigation requirements as recommended for Level 2 automated decision-making systems. Key requirements we included as a result of this:

   a. Peer review: We both had a 3rd party data science vendor review the model, deployment architecture and security architecture; as well as submitted the article (arXiv pre-print link) to Elsevier Engineering Applications of Artificial Intelligence for formal peer-review.

   b. Notice: This is part of the Open Source strategy (GitHub link), to facilitate detailed documentation about the model. The model is currently in implementation planning stages and plain language explanations will be part of training and on-going implementation.

   c. Human-in-the-loop for decision: Though a Level 2 system by the Directive on ADS doesn't require human-in-the-loop, given the sensitive nature of this project, it was decided that this would only ever be decision aid/assistance and a human would always be in the loop.

   d. Explanation: We utilized the local-interpretable model agnostic explanations algorithm (LIME) to generate human interpretable explanations of every prediction generated by the machine learning system (see paper for examples of visual explanations output). This allows users to understand why the system made the decisions it did, to build trust in the system, and monitor the system for unintended bias.

   e. Testing: Prior to production, using the LIME algorithm, we were able to flag if unintended bias had slipped into the system and enabled us to remove or remediate any features that either unfairly impacted outcomes or otherwise led to problematic predictions of the system.

   f. Monitoring: our Support Model has a skilled data scientist monitor the performance metrics of the production system weekly, predictions are generated daily (as opposed to on-demand). Further, prior to launching, we ran analysis and found performance metrics only began being impacted when the training data of the model was older than approximately 1.5 years old. Therefore, we built into the support model that the machine learning model would undergo annual retraining with a business sign off of the performance metrics before pushing the newly trained model into production.

g. Training: Through the pre-print article, and the documentation of the model in the GitHub repo linked above, we were able to provide comprehensive documentation on the model functioning. Future implementation of standards of practice and the business level will also be documented to enable informed decisions based on the model.

h. Audit: We ensured that a record of all predictions ever made by the production system mapped to which version of the system is available to enable auditability of all decisions.

2) We also documented the data preprocessing and feature engineering process followed to ensure we eliminated any unintended bias, and any data issues throughout. This is currently an internal case study under review, which will be open-sourced once final edits are complete.

3) Another important aspect of the implementation process is seen in Homeless Prevention now facilitating conversations with social service practitioners, individuals with lived experience and other sector stakeholders to design exactly how implementation will be conducted, what standards of practice are to be developed, and to ensure this is a community led implementation decision.

4) Throughout the project development Homeless Prevention and the London Homeless Prevention Network (a body including all London shelter leaders with individuals with lived experience) provided oversight of the project.

5) We implemented a feature in the system to ensure that if a client withdraws consent, their information will no longer be processed by the machine learning system. This happens automatically if explicit consent is withdrawn at the database level, and can be done on a manual client by client basis as well for the machine learning aspect of the system as well if desired. This is more to align with machine learning regulatory legislation seen in the European Union in the GDPRs right for an individual to restrict processing through an automated decision system. This is not covered explicitly in the Treasury Board of Canada’s Directive on Automated Decision Making, but we believe it is worthwhile to have this redundancy and it is was simple enough to incorporate technically.