INTERPRETABLE MACHINE LEARNING APPROACHES TO PREDICTION OF CHRONIC HOMELESSNESS

A PREPRINT

Blake VanBerlo
VanBerlo Consulting
London, Canada
blake@vanberloconsulting.com

Matthew A. S. Ross
Artificial Intelligence Research and Development Lab
Information and Technology Services
The Corporation of the City of London
London, Canada
maross@london.ca

Jonathan Rivard
Homeless Prevention
The Corporation of the City of London
London, Canada
jrivard@london.ca

Ryan Booker
Information and Technology Services
The Corporation of the City of London
London, Canada
rbooker@london.ca

September 10, 2020

ABSTRACT

We introduce a machine learning approach to predict chronic homelessness from de-identified client shelter records drawn from a commonly used Canadian homelessness management information system. Using a 30-day time step, a dataset for 6521 individuals was generated. Our model, HIFIS-RNN-MLP, incorporates both static and dynamic features of a client's history to forecast chronic homelessness 6 months into the client's future. The training method was fine-tuned to achieve a high F1-score, giving a desired balance between high recall and precision. Mean recall and precision across 10-fold cross validation were 0.921 and 0.651 respectively. An interpretability method was applied to explain individual predictions and gain insight into the overall factors contributing to chronic homelessness among the population studied. The model achieves state-of-the-art performance and improved stakeholder trust of what is usually a "black box" neural network model through interpretable AI.

***K*eywords** Machine learning · Interpretability · Forecasting · Homeless prevention

1 Introduction

1.1 Problem

Homelessness in Canada has been changing over recent years. A 2016 report claims that annually upwards of 235 000 Canadians endure periods of homelessness, with approximately 35 000 individuals lacking a place to stay each night [1]. Between 2005 and 2014, there was a downward trend in the total number of Canadians using shelters; however, the occupancy rates of shelters has been increasing [1]. One factor accounting for this ongoing decrease in the number of homeless individuals paired with an increase in shelter occupancy is an increase in chronic homelessness. London's Homeless Prevention division identifies an individual as chronically homelessness if they have spent 6 or more months ($\geq$ 180 days) of the last year in a shelter, which was based on the definition of chronic homelessness outlined by the Canadian government's homelessness strategy directives [2]. In addition to this trend, the demographics of homelessness are changing in Canada. In preceding decades, older, single males are over-represented in the homeless population; in contrast, the homeless population of today is increasingly diverse, with families, women, and youth comprising a greater fraction [1].

Given the diverse and evolving makeup of the modern homeless population in Canada, it would be advantageous to elucidate factors contributing to chronic homelessness to enable the predictive identification of individuals at risk of becoming chronically homeless. Shelters and municipal social services are faced with the task of preventing individuals from entering states of chronic homelessness, while acting as stewards of public resources. Machine learning models can help improve the efficiency and transparency of this process by identifying and triaging individuals at high risk of chronic homelessness. Proactive screening can inform targeted intervention before at-risk individuals suffer greater trauma and have their chronic homelessness further encumber an already overburdened shelter system [3]. This study aims to explore the efficacy of employing machine learning to predict the risk of chronic homelessness using data from the shelter system of London, Canada.

In consultation with Homeless Prevention and the London Homeless Prevention Network, it was speculated that early identification of individuals at risk of chronic homelessness may enable London's Centralized Intake system to provide more resources to divert them from experiencing homelessness altogether. As shelters continue to adopt a housing-focused model of care, those at risk of chronic homelessness may be rapidly rehoused to further reduce over-occupancy in the shelter system. Preventive and diversionary resources are less costly overall than the reactive consumption of shelter resources by someone who has become chronically homeless. The conservation of resources via prevention of chronic homelessness would enable the shelter system to serve a greater number of individuals.

1.2 Goals

The primary aim of this project was to develop a machine learning model to predict whether an individual would be in a state of chronic homelessness at a point 6 months in the future. The team considered false negatives to be more harmful than false positives and therefore throughout the study, the goal was to train a proficient model that primarily minimized false negatives, while balancing a desired decrease in false positives.

A secondary goal of the study was to gain insight into the factors that contribute to chronic homelessness in London. As a result, it was imperative that the model's predictions were interpretable. We pursued a system that would produce accurate predictions and accompanying explanations that relate client features to their predicted class. This approach of utilizing interpretable artificial intelligence (AI) has the added benefit of enabling the reduction of unintended bias and increased transparency in government-deployed automated decision systems.

A secondary goal was to release the source code and accompanying documentation under an open source license to enable other homeless services agencies across Canada to quickly train and deploy their own machine learning models for predicting chronic homelessness in their jurisdictions.

1.3 Precedent Research

In recent years, numerous studies have applied statistical and/or machine learning concepts to model homelessness scenarios. There exist several studies and decision support tools that have been developed to serve different regions and subsets of the homeless population.

A well-known example of a decision aid that applies homelessness modelling is the Service Prioritization Decision Assistance Tool (SPDAT) [4]. With multiple versions available, the SPDAT is a screening tool that assists communities with the prioritization of homelessness prevention resources by triaging clients based on a questionnaire. As of 2015, the SPDAT was being used in communities across multiple countries, including Canada [4]. The City of London has utilized the SPDAT for over 5 years. In essence, the SPDAT is a linear model whose features are answers to specific questions. Despite its widespread adoption, the SPDAT has its shortcomings. For instance, an American study concluded that previous versions of the Vulnerability Index-SPDAT (VI-SPDAT) struggle to hold valid and exhibit reliability, claiming that the model fell short particularly in the "Socialization and Daily Functions" and "Wellness - Health" areas [5].

A multitude of studies have applied mathematical modelling and/or machine learning to the prediction of homelessness. A 2013 study applied Cox regression to predict whether individuals in New York City would enter shelters, and presented a screening questionnaire derived from their results [6]. After appropriate thresholding, the study reported an increase in recall of 26% from its baseline [6]. Also in New York, a study investigated the use of logistic regression to predict chances of readmission to shelters and length of stay [7]. These models, trained on a database of 6000 homeless families, attained an area under the receiver operating characteristic curve (AUC) of 0.70 and a pseudo-R^2 value of 0.069 respectively [7]. Citizenship, age, medical history, and childhood foster care or shelter stays were reported to be the most influential features. Further, the study adapted K-means clustering to sort clients into 3 clusters, which the authors analyzed as representative of 3 conventionally described subtypes of homelessness: chronic, episodic, and

transitional [7]. In 2016, Greer *et al.* applied Cox regression to develop models predicting individual and familial entry into New York City shelters over 2-8 years, which achieved AUCs of 0.90 and 0.73 respectively [8].

The Economic Roundtable (ER) has undertaken several initiatives to predict homelessness. In 2011, this group developed a tool based on logistic regression that identifies the top tenth of homeless individuals ordered by expense, reporting a recall of 0.833 and precision of 0.8 for the task of predicting whether a person will fall in this top decile [9]. In 2017, ER released a report on their Silicon Valley Triage Tool that predicts which homeless individuals will be the most costly users of public resources [10]. Having also investigated decision tree modelling and least-angle regression, the authors selected their logistic regression model, which achieved an AUC of 0.83 [10]. Recent work by ER focused on creating separate predictive models for identifying recently unemployed workers and for young adults at risk of becoming persistently homeless, which achieved AUCs of 0.89 and 0.88 respectively [11]. Both models were fitted using logistic regression and their coefficients were presented to infer the relative importance of the input features, which also exemplifies the growing importance of model interpretability [11].

Other studies have explored different subsets of the homeless population, addressing various formulations of the problem of predicting homelessness. One study applied undisclosed predictive analytical methods to forecast first-time homelessness and return to homelessness within 12 months, among a dataset of 1.9 million single adults in Los Angeles County [12]. Chan *et al.* investigated the use of logistic regression and decision trees to train interpretable models intended to function as decision aids for housing prioritization among homeless youth [13]. Finally, a 2020 Canadian study applied a custom variant of Q-learning to simulate transitions of individuals between states of homelessness, including states such as staying in a shelter, in the hospital, on the streets, and being housed [14]. The researchers' model computes transition probability matrices on a weekly basis to generate a simulation for the population. In comparison to actual outcomes from a population dataset extending over 3 years, the simulated population had a relative difference of 12.5% [14].

1.4 Our Contribution

The problem addressed in this paper is similar to some of the aforementioned studies. This study's focus was the development of a model that accurately predicts whether an individual will become chronically homeless in the next 6 months and the identification of factors that influence their chronic homelessness, as well as the general drivers of chronic homelessness in London, Canada. This study is among the first to apply an artificial neural network to chronic homelessness forecasting. Further, the trained model was specifically designed to capture time-series service usage sequences in conjunction with static demographic features. Despite the inherent opaqueness of a neural network, a post-hoc interpretability method was applied that enhanced model transparency. As most precedent research pursued inherently interpretable models such as logistic regression, our results indicate that modern interpretability algorithms may be suitable to obtain stakeholders' trust in "black box" models intended to be decision aids in public services. Further, this interpretable strategy enables utilization of machine learning approaches to mathematical modelling and prediction in future homeless prevention and public service machine learning research and development.

Perhaps of even greater importance is the replicability of our approach in other municipal jurisdictions. The source of data for our model was the City of London's Homeless Individuals and Families Information System[1] (HIFIS) application which joins the service usage information for over a dozen shelters and related homeless services. The Canadian government's homelessness strategy directive mandates that municipalities are to adopt HIFIS if they lack a preexisting homelessness information management system, and are entitled to funding to assist with its deployment [2]. Care was taken to promote readability and modularity when writing our source code for all experiments to enable other HIFIS users to quickly train and deploy their own models. It is our belief that the results described in this paper could be reproduced for other jurisdictions' homeless populations given sufficient local training data.

2 Methods

2.1 Data

The raw data for this study was extracted from the database connected to the City of London's HIFIS application (version 4.0.57.30). A SQL query was constructed to pull all records for all clients from the HIFIS database, which were subsequently saved in CSV format. The resultant raw data encapsulates interactions with social services, personal events/attributes, and demographic information. Client anonymity was preserved, as names and other identifiable information was not fetched by the query. Rather, clients were identified by a unique *ClientID*. At the time of writing, London's HIFIS database contains approximately 4 years of 6521 clients' records.

[1] `https://www.canada.ca/en/employment-social-development/programs/homelessness/hifis.html`

The model was to be trained to predict if clients were at risk of becoming or continuing to be chronically homelessness 6 months in the future. A client was considered chronically homeless if they had at least 180 stays over the most recent 365.25 days. A stay was defined as 1 or more shelter visits, occurring on the same day, that were each at least 15 minutes in duration. Multiple visits on the same day were treated as 1 stay. An example was therefore a $(\vec{x}, y)$ tuple, where $\vec{x}$ is a vector representing a client's state on a particular date, and y is the example's corresponding ground truth. For any example, if the client met the criterion for chronic homelessness 6 months after the example's date, then the ground truth is positive (i.e. $y = 1$); otherwise, the ground truth is negative (i.e. $y = 0$).

Prior to all training experiments, data was cleansed to remove any features considered to be either inconsequential or at risk of introducing unintended bias. Examples of features eliminated at the outset include height, eye colour, and hair colour. The corresponding columns in the raw data were dropped before any other pre-processing was applied.

Data preprocessing was conducted to transform the raw data to a dataset of examples which were then fed to the model. Each example included both dynamic and static features. The dynamic features were composed of numerical features describing density of usage of specific social services over the most recent 6 time steps. Services included features such as: number of shelter stays, number of days of case management, number of days receiving a housing subsidy, number of days in supportive housing, number of times an individual was refused service at a shelter, and number of SPDAT assessments conducted (see Appendix A for a complete listing). Time steps were 30 days long, and each dynamic feature represented the number of times a service was accessed during that time step. The time series input sequence length (T_x) was 6, meaning that each dynamic service feature from the raw data resulted in 6 features of preprocessed data – 1 for each of the last 6 time steps. In contrast, static features consisted of any feature not intended to capture recent time dependant service usage. Examples of static features include: total number of times services were accessed (since the beginning of a client's history in the HIFIS database), total monthly income, total monthly expenses, medical diagnoses, shelters they stayed at, as well as demographic information, such as age, citizenship and gender. See Appendix A for a described list of all features derived from raw data.

Preprocessing of numerical features and categorical features differed in the construction of an example vector $\vec{x}$. To speed up training convergence, each numerical feature of an example was normalized by applying the operation described in Equation 1, where x_i is the i^{th} feature of an example, and $\overline{x_i}$ and σ_i are respectively the mean and standard deviation of the i^{th} feature in the training set [15]. The same transformation was applied to numerical examples in the validation and test sets, using the mean and standard deviation of the training set.

$$x_i \leftarrow \frac{x_i - \overline{x_i}}{\sigma_i} \tag{1}$$

Categorical features, were represented as one-hot encoded bit arrays. *Single-valued categorical features* (SVCFs), defined as features for which an example takes on a single value (e.g. citizenship), were one-hot encoded. *Multi-valued categorical features* (MVCFs) defined as features for which an example may take on any number of values were first split into a sparse array with a new feature for every possible value of the original MVCF. For instance, health issues are a MVCF because a client may have 0 or more health issues. MVCFs were transformed into sparse bit arrays, where each element was a Boolean flag indicating the presence or lack thereof of each value in the MVCF's domain. For example, the feature *IncomeType* is split into the following binary features: *IncomeType_Pension*, *IncomeType_Student Loans(s)*, *IncomeType_Old Age Security*, etc.

In cases of missing data, assumptions were made to impute the blank fields. If a client did not have service records during a particular time step, their service usage was set to 0. Any other numerical features were also set to 0 if the client was missing records for that feature. Exceptions to this rule were made for client weight (*ClientWeightKG*) and recent SPDAT score (*TotalScore*), which were set to -1 if their values were nonexistent. Any absent values for SVCFs were set to "Unknown". If a record had no values for a MVCF, the binary features corresponding to each possible value were set to 0.

See Figure 1 for a visual breakdown of an example that has been preprocessed. The preprocessed dataset was a table, indexed by *ClientID* and *Date*, where *Date* is the final date of the current time step. The dataset contained records for each client dating back to their first records of service in the HIFIS database. Figure 2 summarizes the entire preprocessing procedure that transforms raw client records into $(\vec{x}, y)$ examples. At the time of writing, the preprocessed dataset contained 109 575 examples, 6.56% of which had a positive ground truth.

Prior to training, the data was partitioned into training, validation and test sets. As is customary for time series scenarios, validation data comprised the end segment of the dataset [16]. In keeping with this paradigm, the validation and test sets were taken to be the second-most and most recent partitions of data respectively, where each such partition was composed of all clients' preprocessed records from 1-2 time steps.

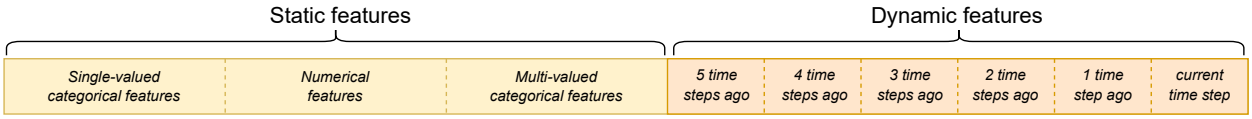

Figure 1: A breakdown of the composition of a feature vector ($\vec{x}$) for an example in the preprocessed dataset.

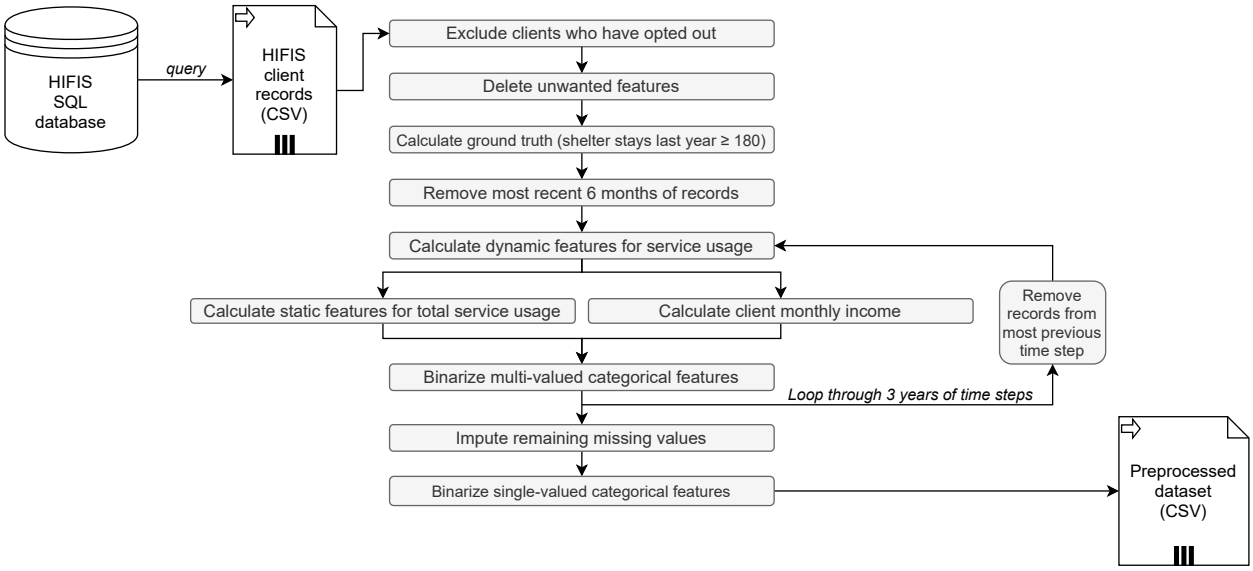

Figure 2: Major steps in data preprocessing, describing how records from the HIFIS database are transformed into the preprocessed dataset.

2.2 Model

A neural network model was designed to capture any time dependencies between dynamic features, in combination with the information contained in the static features. Our chosen model architecture, dubbed *HIFIS-RNN-MLP*, consisted of 2 components: a recurrent neural network (RNN) and a multilayer perceptron (MLP). Inspiration for this model arose from Hsu *et al.*'s application of a hybrid model that combined a RNN and random forest (RF) components [17]. To predict credit card defaults. This approach to machine learning based risk assessment can be applied to the prediction of chronic homelessness as both problems involve using dynamic and static features to predict rare undesirable events. The choice to employ a MLP as the second component made our entire model architecture differentiable and thus end-to-end trainable.

Examples were fed into our model as feature vectors. The dynamic features in the example were isolated and reshaped into a matrix, then passed to the RNN that consisted of long short-term memory (LSTM) cells. The output of the LSTM and its hidden state outputs for each time step were concatenated with the static features, then passed to the MLP. The final layer of the MLP was a single neuron with sigmoid activation, whose output represented the model's assignment of probability that the client would be chronically homeless 6 months after the date of the example. The classification threshold was set to 0.5. The output neuron's bias was initialized to the natural logarithm of the ratio of positive to negative ground truth examples in the training set. When training neural network classifiers on imbalanced data with few positives, this initialization technique accelerates convergence by coercing the model to naively predict a low probability at the start of training [18]. The model's architecture is portrayed in Figure 3.

A selection of regularization methods were applied to combat overfitting. First, the L2 regularization penalty ($\gamma = 1.78 \times 10^{-3}$) was applied to all pre-output fully connected layers in the MLP component. Additionally, dropout was applied to all fully connected layers in the MLP component at a rate of 0.44 [19]. Lastly, early stopping was employed to halt training once validation loss did not decrease for 15 epochs [20], and the model weights were frozen at the epoch corresponding to the minimum validation loss.

The model was trained using the Adam optimization method [21] at a learning rate (α) of 1×10^{-3} for a maximum of 300 epochs. Early stopping typically discontinued the training loop prior to reaching 300 epochs. Equation 2 shows the

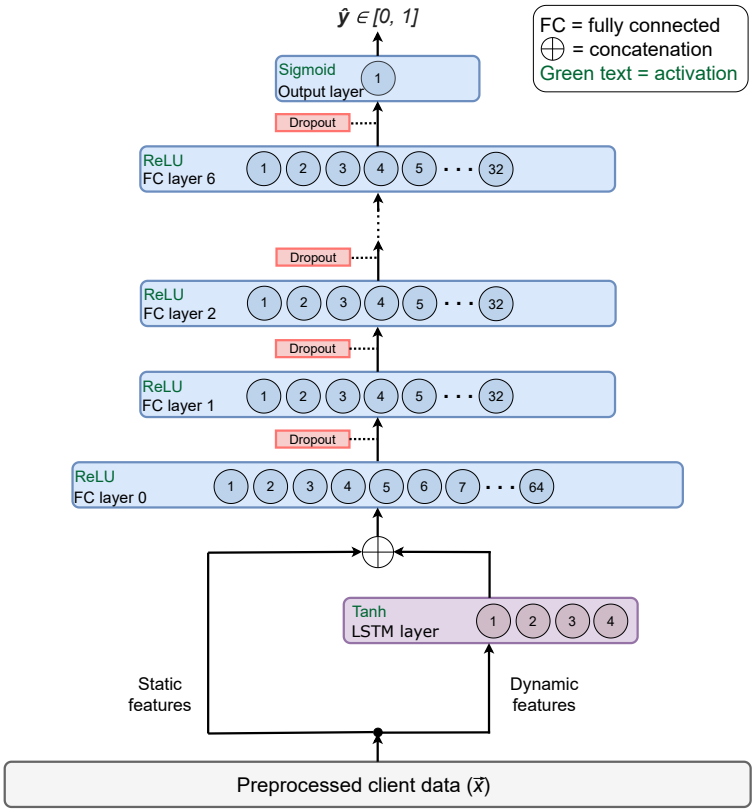

Figure 3: The HIFIS-RNN-MLP architecture. Dynamic features are passed to an LSTM layer before being concatenated with the static features to be fed to a series of fully connected layers.

custom weighted F1 loss function employed during training.

$$\mathscr{L} := 1 - F1_{weighted} = 1 - \frac{2PR}{(2 - \frac{2}{w_r+1})P + (\frac{2}{w_r+1})R} \tag{2}$$

$$P = \frac{true\ positives}{predicted\ positives} = \frac{\sum y \cdot \hat{y}}{\sum \hat{y}} \tag{3}$$

$$R = \frac{true\ positives}{actual\ positives} = \frac{\sum y \cdot \hat{y}}{\sum y} \tag{4}$$

In Equation 2, $P \in [0, 1]$ is precision, $R \in [0, 1]$ is recall, and $w_R \in \mathbb{R}^+$ is the recall weight. In Equations 3 and 4, model predictions and ground truths are represented by $\hat{y} \in [0, 1]$ and $y \in \{0, 1\}$ respectively. Note that precision and recall are computed here using the probabilistic predictions $\hat{y}$, guaranteeing differentiability. Setting $w_R > 1$ more harshly penalizes the model if recall is low. In training our model, we found that $w_R = 4.5$ achieved a desirable balance between precision and recall that favoured the latter. See Figure 4 for an example of training and validation curves using the weighted F1 loss function.

A handful of hyperparameters were studied to optimize model performance on validation and test sets. Bergstra and Bengio demonstrated that randomly searching the hyperparameter space is equally as effective and less computationally burdensome than grid search [22]. Several of these random hyperparameter searches were completed, narrowing down optimal ranges of the hyperparameters after each experiment. Table 1 lists the final hyperparameter values adopted for the HIFIS-RNN-MLP model.

The code used to arrive at the results presented in this paper was written in Python 3 and is publicly accessible via our GitHub repository [2]. All training experiments were conducted on a computer running Windows 10, equipped with an Intel® Core™ i7-8750 CPU at 2.2 GHz with 6 cores, and a NVIDIA® GeForce GTX ® 1050 Ti GPU with 4 GB of memory.

[2] `https://github.com/aildnont/HIFIS-model`

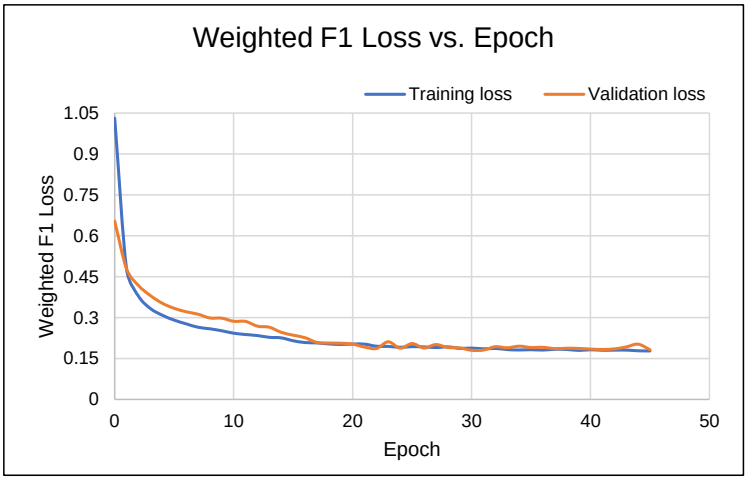

Figure 4: A sample of the training curves for the HIFIS-RNN-MLP model demonstrating convergence with the weighted F1 loss function.

2.3 Interpretability

Hyperparameter	Value
# LSTM units in RNN	4
# Fully connected layers in MLP	6
# Nodes in first fully connected layer of MLP	64
# Nodes in remaining fully connected layers of MLP	32
Dropout rate	0.44
L2 regularization parameter (γ)	1.78×10^{-3}
Learning rate (α)	1×10^{-3}
Batch size	1024

Table 1: Final hyperparameter values

The HIFIS-RNN-MLP model was a neural network and therefore it was not inherently interpretable. This is why neural networks are called "black box" models. To increase model transparency, the Local Interpretable Model-Agnostic Explanations (LIME) method was applied [23]. LIME was created to explain predictions made by any "black box" models. The basic principle of LIME is the assumption that nonlinear models may be approximated by linear models at a small scale. LIME slightly perturbs the feature values of an example, creating a set of similar examples within its neighbourhood. An exponential kernel is used to define the neighbourhood. Any inherently interpretable model trained on the black box's predictions of the generated neighbourhood of examples, may then be used to approximate the black box model's functionality within that neighbourhood. The local model's parameters are taken to represent the relative importance of the original example's feature values to the original model's prediction.

Many parameters of LIME were evaluated in pursuit of transparent and human-friendly explanations. Ridge regression was chosen as the inherently interpretable local model to explain the HIFIS-RNN-MLP predictions using LIME. To decrease the feature space of the local surrogate model, the numerical features were discretized into 4 bins. Next, the choice of kernel width is crucial because it defines the size of the neighbourhood generated around an example, which greatly influences the locality and stability of LIME explanations. Stability is a property of explanations that refers to how alike explanations for similar examples are [24]. Since we prioritized stability, we endeavoured to produce explanations that should minimally differ when produced for the same client. Too small a kernel width increased locality of explanations and compromised their stability. Whereas, increasing the kernel width too much can cause the local model to approach a global surrogate, which is counter to the goal of local explanations in the first place [25]. After investigating different values for the kernel width, we chose to use the default choice in the author's implementation of LIME, which is $0.75 \cdot \sqrt{|\vec{x}|}$. The default value produced local and reasonably stable explanations. To further enhance explanation stability, we increased the sample size, which refers to the number of slightly perturbed examples generated and used to fit the local model. According to an experimental analysis of LIME by Molnar *et al.*, "the sample size was a strictly monotonous benefactor for explanation stability and thus should not be reduced" [25]. Accordingly, we set the sample size to 40 000, which represented the ceiling of computational overhead we were willing to accept given production deployment requirements. With the aforementioned values for LIME parameters, each explanation took approximately 8.4 seconds to compute using our hardware. Overall, the application of LIME with our chosen set of parameters yielded sufficiently local and stable explanations.

3 Results

To assess the performance of the HIFIS-RNN-MLP model, various metrics were considered, including recall, precision, F1-score, AUC, and our weighted F1 loss. When conducting training experiments, it was necessary to evaluate and fine-tune towards models that were consistent with Homeless Prevention's goals. Consider that a false negative corresponds to a scenario in which a client becomes chronically homeless within the next 6 months, despite the model's prediction that they were not at risk. Missing at risk individuals is highly undesirable. Alternatively, a false positive case corresponds to a situation in which a client is predicted by the model to be at high risk of chronic homelessness in the next 6 months, but does not end up becoming chronically homeless in the future. Given a choice, the latter scenario is preferred by Homeless Prevention. The cost of a false negative is much higher than the cost of a false positive since effective preventive resources can save significant costs incurred by would-be long-term shelter users [26]. These emotional and financial costs savings can be realized to the fullest extent if our model greatly reduces the number of clients who are falsely misclassified as unlikely to be chronically homeless in the future. Models were therefore selected using recall, precision and F1-score as the primary evaluation criteria, with a preference toward recall to minimize false negatives.

3.1 Model Performance

The model was evaluated based on average performance on held out data via cross validation. Since the model addresses a forecasting problem, traditional partitions for validation folds do not apply. A form of nested cross validation was implemented that draws inspiration from rolling-origin evaluation described by Tashman *et al.* [27] and rolling-origin-recalibration evaluation discussed by Bergmeir *et al.* [16]. For each fold, the dataset was partitioned by assigning records from the second-most and most recent time steps to the validation and test respectively, with all earlier records comprising the training set. In the first fold, the entire dataset was partitioned. For the k^{th} fold, records from the $k-1$ most recent time steps were omitted prior to partitioning. Refer to Figure 5 for a portrayal of dataset partitioning for the k^{th} fold. The model was trained on 10 folds defined by this nested cross validation method.

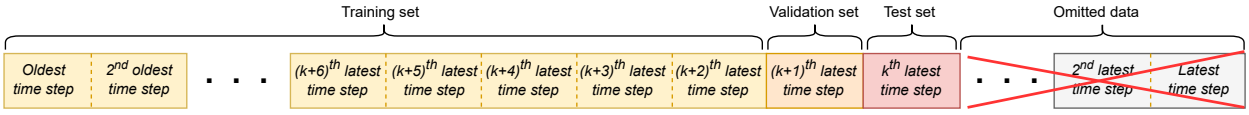

Figure 5: Dataset partitioning for the k^{th} fold in the nested cross validation. Each block represents all existing client examples for a particular time step.

To endorse the decision of framing the problem using both dynamic features and static features, we trained a MLP model that did not take into account dynamic features. Only running totals of service features were considered. In this static data modality, each example consisted of the static features for client, calculated as per their records in the HIFIS database available at the time of writing. The static dataset was indexed by *ClientID*, and could therefore be evaluated using k-fold cross validation by defining test sets composed of held out client records. Test set results from a 10-fold cross validation are reported in Table 2. As demonstrated by the results, recall was comparable for both models, but the HIFIS-RNN-MLP model achieved considerably higher precision than the MLP model using the static dataset.

Features	*Model*	*Mean metric value [standard deviation]* $\times 10^2$				
		Recall	Precision	F1-score	AUC	Accuracy
Dynamic & static	**HIFIS-RNN-MLP**	**92.1 [1.7]**	**65.1 [3.0]**	**76.3 [2.0]**	**97.6 [0.7]**	**97.1 [0.2]**
	Logistic regression	93.2 [1.8]	61.7 [1.9]	74.2 [1.8]	98.9 [0.3]	96.7 [0.2]
	Random forest	74.0 [5.1]	87.2 [1.1]	80.0 [3.1]	99.1 [0.3]	98.1 [0.1]
Static	MLP	89.7 [5.0]	36.3 [6.5]	51.3 [6.3]	96.5 [1.3]	92.2 [1.0]
	Logistic regression	80.6 [7.0]	38.9 [4.4]	52.3 [5.1]	95.0 [2.4]	93.2 [0.8]
	Random forest	17.0 [9.5]	60.8 [17.4]	25.8 [12.8]	95.6 [1.8]	95.7 [0.9]

Table 2: Holdout performance for cross validation of various models and data modalities

To further illustrate the utility of the HIFIS-RNN-MLP model, its performance was compared to classical learning algorithms, logistic regression and a 100-tree random forest. These benchmark models were trained on both the dynamic time series dataset and the static dataset and results of cross validation of all models considered is reported in Table 2. Class weighting was employed in logistic regression, random forest, and the MLP, to imbue training with our goal of

accurately identifying positives. However, HIFIS-RNN-MLP trained using the weighted F1 loss was able to achieve a balance of recall and precision closest to Homeless Prevention's goals.

To demonstrate the utility of the custom F1 loss function, in Table 3 we display the results of cross validation performance of the HIFIS-RNN-MLP model using various formulations of the loss function. As well as compared against binary cross entropy (BCE) as a loss function. Since the dataset was unbalanced, class weighting was investigated as a means to force more attention to be paid to the minority positive class examples. When conducting training experiments with weighted BCE loss, a penalty was applied based on the fraction of examples with positive ground truth (i.e. 6.56%). The results in Table 3 indicate that the custom F1 loss function is slightly more effective than class-weighted BCE at achieving the most desirable precision-recall balance.

Loss function	*Mean metric value [standard deviation]* $\times 10^2$				
	Recall	Precision	F1-score	AUC	Accuracy
Weighted F1 loss	**92.1 [1.7]**	**65.1 [3.0]**	**76.3 [2.0]**	**97.6 [0.7]**	**97.1 [0.2]**
BCE with class weighting	93.5 [1.4]	63.8 [3.8]	75.8 [2.7]	99.1 [0.2]	97.0 [0.3]
BCE	77.3 [5.2]	84.6 [2.7]	80.7 [3.0]	99.1 [0.2]	98.1 [0.2]

Table 3: Holdout set performance for cross validation of HIFIS-RNN-MLP with varying loss functions.

3.2 Interpretability

The application of LIME to predictions made by the HIFIS-RNN-MLP model resulted in explanations that were not only consistent with predictors of chronic homelessness from the literature, but provided additional insight into the chronically homeless population specific to London. Each explanation consists of a series of paired feature values and weights, listed in descending order by weight. The feature values with the highest weight magnitudes may be considered as the client's attributes that most contributed toward the model's prediction of a positive ground truth. A selection of examples of client explanations are shown in Figure 6. The local explanations for each client served three main purposes in the context of model development. First, they helped ensure that unintended bias was not present in the model's decisions. Explanations were examined to ensure that predictions were not contingent on single demographic features. Second, explanations aided in iterative feature engineering. Some categorical features take on several possible values, and thus constitute large fractions of a preprocessed example. Any categorical features that never appeared in explanations were subsequently appended to the list of features to exclude prior to training (e.g. medications). Finally, collaborative analysis of explanations with domain experts at Homeless Prevention helped validate that the model was not fixating on bizarre or unrealistic correlations.

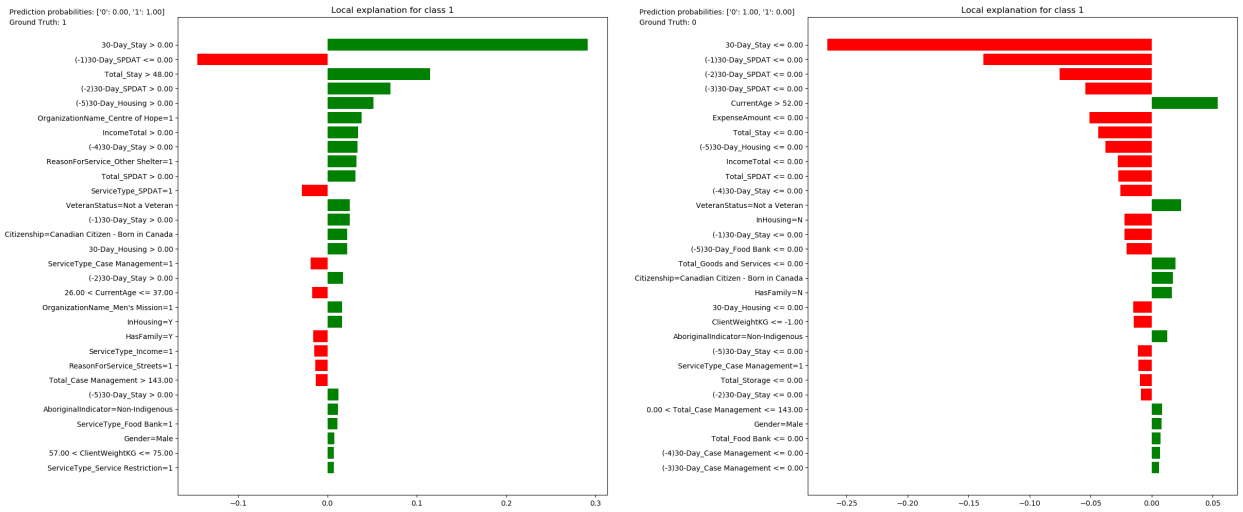

Figure 6: A selection of LIME explanations for client examples in the test set. Explanations consist of a list of feature values with weight corresponding to the bars on the graphs. Green and red bars indicate contribution towards and against a prediction of chronic homelessness respectively.

Although individual explanations are highly informative in and of themselves, their value in understanding the model as a whole is limited. The authors of LIME proposed a method called *submodular pick*, which was designed to provide a holistic understanding of the model by combining a series of explanations that maximally covers the model's input space [23]. The submodular pick algorithm, after computing a tuneable number of explanations, performs a greedy pick of explanations that maximizes the representation of the input feature space [23]. The result is a list of feature values paired with weights. An explanation's weights approximate the relative importance of feature values to the model's decision, independent of a specific example. Submodular pick was executed with 20% of the combined training and validation sets as input. The resultant feature values and weights constitute the global model explanation depicted in Figure 7. Positive and negative weights correspond to contribution towards and against a prediction of chronic homelessness respectively. The most conspicuous outcome in Figure 7 is the importance of the number of stays in the most recent 30-day time step (i.e. *"30-Day_Stay"*). Similarly, total shelter stays were a very influential feature (i.e. *"Total_Stay"*). These findings corroborate Shinn *et al.*'s result that previous shelter stays were the strongest predictive feature for familial homelessness [6]. It also appears that the administration of a SPDAT screening questionnaire 1 time step ago is highly predictive of chronic homelessness (i.e. *"(-1)30-Day_SPDAT"*), perhaps commending London case workers' ability to identify high-risk clients weeks prior to transitioning to chronic homelessness as defined here. Also of remarkable magnitude is a client's aggregate days of receipt of housing subsidies (i.e. *"Total_Housing_Subsidy"*). According to the global explanation, lack of receipt of housing subsidies steers the the model towards predicting chronic homelessness, which is in keeping with Byrne *et al.*'s finding of negative association between chronic homelessness and subsidized housing [28]. Again corroborating the work of Shinn *et al.*, advanced age appears to be a predictive factor [6]. As reported in Figure 7, being in the highest age bin *"CurrentAge > 52.00"*) increases a client's risk of future chronic homelessness; whereas, belonging to the lowest age bin *"CurrentAge <= 26.00"*) seems protective.

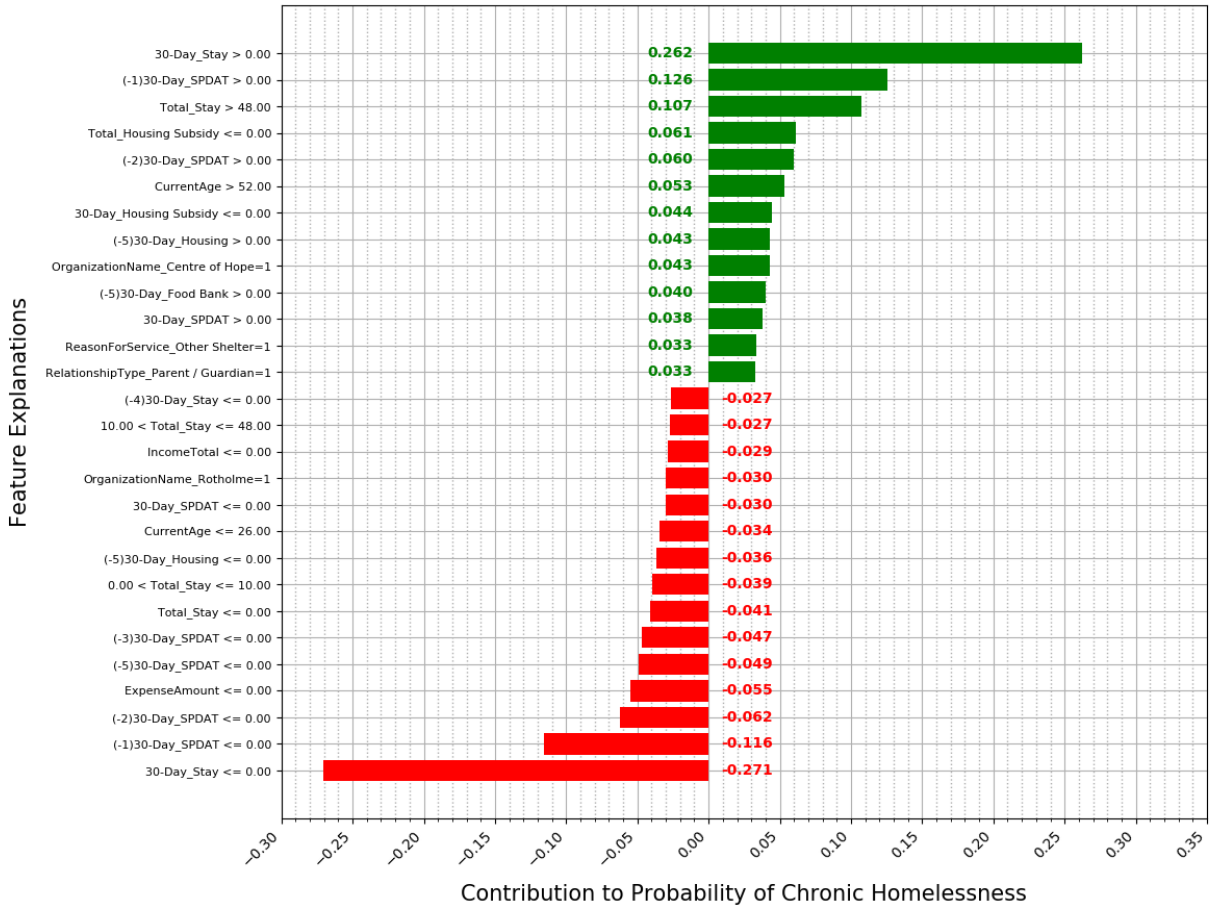

Figure 7: Results of the LIME submodular pick procedure. This graph communicates an approximation of model functionality. Each bar corresponds to the weight of a feature value or range. Green and red bars indicate contribution toward and against prediction of chronic homelessness. The magnitude of a bar indicates its relative influence in the model's decision.

4 Discussion

This project is among the first to apply an artificial neural network to model chronic homelessness. Moreover, it succeeded in illuminating the insights learned by the neural network, which are traditionally viewed as "black box" models. Performance metrics achieved by the HIFIS-RNN-MLP model exceeded Homeless Prevention's expectations. Among the work most similar to ours is the research by Toros *et al.* that developed separate models to predict persistent homelessness among adults who exited the job market recently and young adults receiving public benefits [11]. Their criterion for "persistently homeless" is having experienced more than 1 period of homelessness (defined as having no address) within 3 years. Despite not being equivalent to the "chronically homeless" state designated in this study, we consider their problem similar enough to warrant comparison. Toros *et al.* trained 2 models that attained a holdout AUC of 0.89 and 0.88 [11]. Their employment and young adult models achieved recall of 0.308 and 0.351 at classification thresholds of 0.528 and 0.471 respectively. Nested cross validation of the HIFIS-RNN-MLP model yielded an AUC of 0.976 and a recall of 0.921 at a classification threshold of 0.5.

The predictive capability of the HIFIS-RNN-MLP model introduces the possibility of early identification of clients who are at risk of chronic homelessness. In other jurisdictions, preventive strategies such as housing subsidies, diversionary efforts and community-based services are shown to be effective [3]. In their publication presenting a new framework for homelessness prevention in Canada, Gaetz and Dej present multiple arguments that preventative strategies are cost-effective for society [29]. Avoiding emotional and physical trauma for the chronically homeless population in the shelter system and conserving finite public resources makes these efforts vital. It stands to reason that an interpretable machine learning algorithm that sharpens case workers' ability to identify high risk individuals who would benefit from preventative resources would be invaluable to any municipality. Further, client predictions may be accumulated to forecast aggregate shelter demand. Aside from the immediately tangible benefits, the interpretable nature of this model may help service providers more deeply understand factors contributing to chronic homelessness in their locale.

Our work is readily replicable by other cities who use the HIFIS application. The Canadian government mandated that all municipalities have a homelessness information management system, and offered subsidized implementation of HIFIS to those lacking one [2]. Using their own HIFIS database and the open source code accompanying this paper, municipalities could apply the methods described herein. Care was taken to thoroughly document the open source repository and adhere to modular design so as to enable quick adaptation and implementation. Further the model does not require a GPU to train in an acceptable length of time which further decreases the barrier to implementation from a compute infrastructure perspective.

A key facet of this work was the application of LIME to probe the model for explanations. Interpretability is fundamental to the ethical deployment of decision-making algorithms in the public sector. Methods such as LIME promote transparency and identification of sources of unintended bias. The Canadian government in their Directive on Automated Decision Making states that some automated decision-making systems provide explanations that justify their choices [30]. As an interpretability method, LIME fit this application space. Not only is it possible to obtain explanations on granular example-wise basis, but an overall understanding of the model's behaviour may also be realized by studying the results of the submodular pick algorithm. LIME provides insight into which specific features of a client are most relevant to the decision made by the model, enabling service providers to evaluate a client's prediction in the context of their history. Although previous studies applied inherently interpretable machine learning methods to homelessness prevention (e.g. logistic regression, decision trees), the HIFIS-RNN-MLP model demonstrated performance metric superiority and was interpretable through the utilization of LIME.

Despite the encouraging results of this study, it was not free of limitations. First, the advent of HIFIS in London was relatively recent; as a result, we only had approximately 4 years of records to access. With a time step of 30 days, our dataset contained 115 515 records for 6521 clients. In contrast, Toros *et al.*'s investigation amassed records for recipients of public benefits, children and family public service usage, and homelessness information management system data spanning 10, 8, and 5 years respectively [11]. Additionally, the problem statement of this study limits the model to predicting transition into only the chronic state of homelessness. By definition, episodic and transitional states of homelessness are left out. Finally, our study's definition of chronic homelessness fails to capture the variety of states that chronically homeless people may be in when staying outside the shelter system. Sleeping rough, couch surfing, and stays in healthcare institutions were not accounted for in the calculation of the ground truth. Hence, the ground truth may fail to capture individuals who are homeless for the majority of the year, but stay in a shelter only for a minority of the year.

A possible enhancement to the methods could be the exploration of interpretability methods other than LIME, such as partial dependence plots [31] or SHAP [32]. Due to LIME meeting the interpretable requirements of service providers, no others were investigated. Further study of the model should include its evaluation in deployment. To this end, a randomized control trial could be conducted to quantify any additional preventative resources deployed to clients as a

result of the model's predictions, compared to a control group. Next, a direct comparison of the HIFIS-RNN-MLP model and the SPDAT would support the assertion that this model is a feasible decision support algorithm. In addition to comparing performance, a comparison of the SPDAT's highly weighted features to those highlighted in a LIME submodular pick of our model would be of great interest. To conduct a true comparison, all client features relating to the SPDAT could be excluded from our model. As the SPDAT's use is widespread, this proposed analysis may enhance trust in our study's methods outside of London, Canada. Finally, future investigations of modelling the homeless population in London could include locations occupied by homeless individuals other than shelters, and could incorporate episodic and transitional states of homelessness in the formulation of the ground truth.

5 Conclusion

In this project, a machine learning model was trained to effectively predict chronic homelessness among individuals receiving services in London, Canada. Dubbed HIFIS-RNN-MLP, this model connected RNN and MLP architectures to process dynamic and static features. The trained model achieved a mean recall of 0.921 and precision of 0.651 across the holdout sets of a 10-fold nested cross validation. Application of the LIME interpretability algorithm yielded local explanations that met the requirements of stakeholders. Execution of the submodular pick algorithm produced a global LIME explanation that approximates rules that the model learned from the combinations of input features. Our methods are reproducible and freely accessible. It is our hope that other municipalities may derive benefit from this work.

6 Acknowledgements

We wish to express gratitude for the support and expertise of the following individuals throughout this project: Mat Daley, Vala Gylfadottir, Craig Cooper, Trevor Fowler and Bryan Knight.

References

[1] Stephen Gaetz, Erin Dej, Tim Richter, and Melanie Redman. *The State of Homelessness in Canada 2016*. 2016.

[2] Government of Canada. Reaching Home: Canada's Homelessness Strategy Directives, 2020.

[3] Marybeth Shinn and Rebecca Cohen. Homelessness Prevention: A Review of the Literature. Technical report, 2019.

[4] Vulnerability Index - Service Prioritization Decision Assistance Tool Prescreen Triage Tool for Single Adults Welcome to the SPDAT Line of Products VI-SPDAT Series. Technical report, OrgCode Consulting Inc., 2015.

[5] Molly Brown, Camilla Cummings, Jennifer Lyons, Andrés Carrión, and Dennis P. Watson. Reliability and validity of the Vulnerability Index-Service Prioritization Decision Assistance Tool (VI-SPDAT) in real-world implementation. *Journal of Social Distress and the Homeless*, 27(2):110–117, 2018.

[6] Marybeth Shinn, Andrew L. Greer, Jay Bainbridge, Jonathan Kwon, and Sara Zuiderveen. Efficient targeting of homelessness prevention services for families. *American Journal of Public Health*, 103(SUPPL. 2):324–330, 2013.

[7] Boyeong Hong, Awais Malik, Jack Lundquist, Ira Bellach, and Constantine E. Kontokosta. Applications of Machine Learning Methods to Predict Readmission and Length-of-Stay for Homeless Families: The Case of Win Shelters in New York City. *Journal of Technology in Human Services*, 36(1):89–104, 2018.

[8] Andrew Greer, Marybeth Shinn, Jonathan Kwon, and Sara Zuiderveen. Targeting services to individuals most likely to enter shelter: Evaluating the efficiency of homelessness prevention. *Social Service Review*, 90(1):130–155, 2016.

[9] Daniel Flaming, Patrick Burns, Gerald Sumner, Manuel H. Moreno, and Halil Toros. Crisis Indicator: Triage Tool for Identifying Homeless Adults in Crisis. Technical report, Economic Roundtable, 2011.

[10] Halil Toros and Daniel Flaming. Prioritizing Which Homeless People Get Housing Using Predictive Algorithms. *SSRN Electronic Journal*, pages 1–32, 2017.

[11] Halil Toros, Daniel Flaming, and Patrick Burns. Early Intervention to Prevent Persistent Homelessness. Technical Report March, Economic Roundtable, 2019.

[12] Till Von Wachter, Marianne Bertrand, Harold Pollack, Janey Rountree, and Brian Blackwell. Predicting and Preventing Homelessness in Los Angeles. Technical Report September, The California Policy Lab, University of Chicago Poverty Lab, 2019.

[13] Hau Chan, Eric Rice, Phebe Vayanos, Milind Tambe, and Matthew Morton. Evidence from the past: AI decision AIDS to improve housing systems for homeless youth. *AAAI Fall Symposium - Technical Report*, FS-17-01 -:149–157, 2017.

[14] Andrew Fisher, Vijay Mago, and Eric Latimer. Simulating the evolution of homeless populations in canada using modified deep Q-Learning (MDQL) and modified neural fitted Q-Iteration (MNFQ) algorithms. *IEEE Access*, 8:92954–92968, 2020.

[15] Yann LeCun, Leon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient BackProp. In Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen, editors, *Neural Networks Tricks of the Trade*, chapter 1, pages 9–50. Springer-Verlag Berlin Heidelberg, 1998.

[16] Christoph Bergmeir and José M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

[17] Te Cheng Hsu, Shing Tzuo Liou, Yun Ping Wang, Yung Shun Huang, and Che-Lin. Enhanced Recurrent Neural Network for Combining Static and Dynamic Features for Credit Card Default Prediction. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2019-May:1572–1576, 2019.

[18] Andrej Karpathy. A Recipe for Training Neural Networks, 2019.

[19] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[20] Lutz Prechelt. Early Stopping - But When? In Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen, editors, *Neural Networks Tricks of the Trade*, chapter 2, pages 55–69. Springer-Verlag Berlin Heidelberg, 1998.

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[22] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305, 2012.

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-Augu, pages 1135–1144, 2016.

[24] Christoph Molnar. *Interpretable Machine Learning*. 2019. `https://christophm.github.io/interpretable-ml-book/`.

[25] Christoph Molnar, Sebastian Gruber, and Philipp Kopper. *Limitations of Interpretable Machine Learning Methods*. 2019.

[26] Dennis P. Culhane, Jung Min Park, and Stephen Metraux. The Patterns and Costs of Services Use Among Homeless Families. *Journal of Community Psychology*, 39(7):815–825, 2011.

[27] Leonard J. Tashman. Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4):437–450, 2000.

[28] Thomas Byrne, Jamison D. Fargo, Ann Elizabeth Montgomery, Ellen Munley, and Dennis P. Culhane. The relationship between community investment in permanent supportive housing and chronic homelessness. *Social Service Review*, 88(2):234–263, 2014.

[29] Stephen Gaetz and Erin Dej. *A New Direction: A Framework for Homelessness Prevention*. Canadian Observatory on Homelessness Press, Toronto, 2017.

[30] Government of Canada. Directive on Automated Decision-Making, 2019.

[31] Jerome Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[32] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

Appendix

A Feature Descriptions

Categorizations and descriptions for features of a preprocessed client example. The dataset of preprocessed examples is indexed by *ClientID* and *Date*. Dynamic features exist for the current 30-day time step (TS) and for each previous time step $\mathbf{t} \in [1, T_x - 1]$.

Temporality	Feature type	Feature name	Description
Static	Numerical	CurrentAge	Client's age (in years)
		ClientWeightKG	Weight (in kilograms)
		ExpenseAmount	Total routine expenses (in Canadian dollars)
		TotalScore	Most recent SPDAT score, in range $[-1, 12]$. -1 indicates no SPDAT.
		Total_Stay	Total # stays in a shelter
		Total_Case Management	Total # days of case management
		Total_Housing	Total # days in supportive housing received
		Total_Housing Subsidy	Total # days receiving housing subsidy
		Total_Storage	Total # days of storage service received
		Total_Reservations	Total # shelter bed reservations
		Total_Turnaways	Total # times an individual was refused service at a shelter
		Total_Food Bank	Total # shelter meals
		Total_Goods and Services	Total # goods and services records
		Total_SPDAT	Total # times client has taken the SPDAT
Static	Single-valued categorical	Gender	Client's gender
		AboriginalIndicator	Indigenous status of client (or lack thereof)
		Citizenship	Canadian citizenship status
		VeteranStatus	Type of veteran
		InHousing	Whether client is currently housed
		ExpenseFrequency	How often routine expenses occur
		HasFamily	Whether client has a family on record
Static	Multi-valued categorical	ServiceType	Type(s) of services client has received
		OrganizationName	Name(s) of organizations that client has received services from
		ReasonForService	Reason(s) client has received service(s)
		IncomeType	Type(s) of income client receives
		ExpenseType	Type(s) of client's expense(s)
		IsEssentialYN	Whether client's expense(s) are essential
		Reason	Reason(s) for past service restrictions
		HealthIssue	Medical conditions on client's record
		DiagnosedYN	Whether client has been diagnosed for any of their medical condition(s) on record
		SelfReportedYN	Whether client self-reported any of their medical condition(s) on record
		SelfReportedYN	Whether any of client's medical condition(s) on record are suspected, but not diagnosed
		ContributingFactor	Factor(s) contributing to client's situation
		LifeEvent	Significant event(s) in client's life
		PreScreenPeriod	Periods at which client has been screened via SPDAT
		BehavioralFactor	Dangerous behaviour(s) client has exhibited
		Severity	Severity of behavioral factor(s)
		RelationshipType	Client's family role(s)
		EducationLevel	Highest reported education level(s)
Dynamic	Numerical	30-Day_Stay	# shelter stays in current TS.
		30-Day_Case Management	# days of case management in current TS.
		30-Day_Housing	# days in supportive housing in current TS.
		30-Day_Housing Subsidy	# days of housing subsidy in current TS.
		30-Day_Storage	# days of storage service in current TS
		30-Day_Reservations	# shelter bed reservations in current TS
		30-Day_Turnaways	# shelter turnaways in current TS
		30-Day_Food Bank	# shelter meals in current 30-day TS
		30-Day_Goods and Services	# goods and services in current TS
		30-Day_SPDAT	# times client took SPDAT in current TS
		(-**t**)30-Day_Stay	# shelter stays in $\mathbf{t}^{\text{th}}$ past TS
		(-**t**)30-Day_Case Management	# days of case management in $\mathbf{t}^{\text{th}}$ past TS
		(-**t**)30-Day_Housing	# days in supportive housing in $\mathbf{t}^{\text{th}}$ past TS
		(-**t**)30-Day_Housing Subsidy	# days of housing subsidy in $\mathbf{t}^{\text{th}}$ past TS
		(-**t**)30-Day_Storage	# days of storage service in $\mathbf{t}^{\text{th}}$ past TS
		(-**t**)30-Day_Reservations	# shelter bed reservations in $\mathbf{t}^{\text{th}}$ past TS
		(-**t**)30-Day_Turnaways	# shelter turnaways in $\mathbf{t}^{\text{th}}$ past TS
		(-**t**)30-Day_Food Bank	# shelter meals in $\mathbf{t}^{\text{th}}$ past TS
		(-**t**)30-Day_Goods and Services	# goods and services in $\mathbf{t}^{\text{th}}$ past TS
		(-**t**)30-Day_SPDAT	# times client took SPDAT in $\mathbf{t}^{\text{th}}$ past TS