Prediction of first-time homelessness risk based on utility payment history

Colin D. Middleton[1], Andrew M. Oster[1*]

1 Department of Mathematics, Eastern Washington University, Cheney, WA, USA

*aoster@ewu.edu

Abstract

In order to reduce homeless numbers in Spokane County, the City of Spokane, Avista Utilities, Urbanova, Eastern Washington University and Washington State University formed a consortium, the Spokane Predictive Analytics Group, to collect data and develop new predictive tools to identify those at risk of homelessness. In this work, we present model performance results using monthly de-identified residential utility customer billing data. Our logistic regression model achieved an Area Under the Curve of 0.810, and a False Positive Rate of only 0.508 at a True Positive Rate of 0.899, ranking near the top of current research. This study aims at predicting first-time homelessness based on data available on the general public where nearly all precedent research focuses on predicting chronic homelessness based on much richer data.

Both the City of Spokane and Avista Utilities provided de-identified data

Introduction

Background

The United States Department of Housing and Urban Development (HUD) defines categories of homelessness with their *"Category 1: Literally Homeless"* as

> 1) Individual or family who lacks a fixed, regular, and adequate nighttime residence, meaning:
> (i) Has a primary nighttime residence that is a public or private place not meant for human habitation;
> (ii) Is living in a publicly or privately operated shelter designated to provide temporary living arrangements (including congregate shelters, transitional housing, and hotels and motels paid for by charitable organizations or by federal, state and local government programs); or
> (iii) Is exiting an institution where (s)he has resided for 90 days or less and who resided in an emergency shelter or place not meant for human habitation immediately before entering that institution ([1])

The agencies that collected the data utilized in this work used HUD's *"Category 1: Literally Homeless"* definition of homeless, so we adopt that definition here as well.

The Current State of Homelessness

Homelessness in the United States persists despite the efforts of PROGRAMS and MONEY, MONEY spent each year on assistance programs. From 2005 until 2020,

impressively there has been a decrease of 42% in homelessness in the USA according to the Department of Housing and Urban Development (HUD). HUD still reported over 442,000 people experiencing homelessness in 2020 through their Point in Time Count program [2].

In this work, we consider homelessness in Spokane County, Washington, USA. Spokane County has seen homelessness levels decrease by 32% from 2005 to 2020 with 1,244 people recorded as experiencing homelessness in 2020 [2]. These numbers include individuals sheltered in emergency shelters, sheltered in transitional housing, and unsheltered [2], however some fear PIT counts far underestimate homeless numbers****.
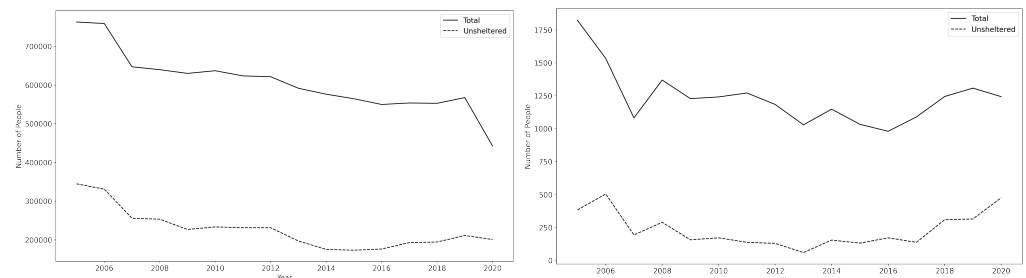
Fig 1. Annual Homelessness Counts in the USA and Spokane County [2]
font size need to be far larger on axes, perhaps thicker lines in plots

Programs focused on providing assistance to people experiencing homelessness are having an impact. HUD's January 2019 survey found that "there were 144,000 more permanent supportive housing (PSH) beds dedicated to people with chronic patterns of homelessness than there were in 2007 (a 380% increase)" [3]. Likely because of this increase in resources the rate of chronic homelessness (people who have experienced homelessness for at least 12 months in the last three years) has declined by 20% from 2007 to 2019 in the US [3]. Current assistance programs and the positive impacts they have on people in need represent a significant accomplishment, but these programs largely focus on providing support to people already experiencing homelessness. This type of assistance program can only help so much. In order to substantially reduce and eventually end homelessness altogether, homelessness prevention programs must be utilized.

We will refer to programs as homelessness prevention programs (HPPs) if they focus on preventing people from experiencing homelessness. These programs include permanent deep rental housing subsidies, eviction prevention programs, community based services such as short term financial assistance, education, and job placement. Homeless assistance programs (HAPs), we shall say, are those that are aimed at assisting people who are currently experiencing an episode of homelessness and include shelters and other emergency services. There are two aspects of HPPs that make them more attractive than HAPs. The first is cost. The services that HPPs typically offer are less expensive for taxpayers than those provided by HAPs [4]. The second reason is that if the HPP is successful it will spare people the trauma typical of experiencing homelessness.

Cost Effectiveness of Homelessness Prevention Programs

Research has shown that many types of HPPs are cost effective in practice, meaning that treating the social issue of homelessness with HPPs instead of HAPs in the same area is at least as effective at keeping people off the streets and costs less money. This is due to the high cost of the financial health and mental health rehabilitation service

HAPs must provide. Once someone experiences an episode of homelessness, they have no option but to use high-cost services such as shelters and emergency health services add source.

A study on permanent deep rental housing subsidies, one of the most promising types of HPP, found that "among the 67 percent of families who successfully used their [rent subsidy] voucher to lease housing, homelessness was prevented entirely" [4]. Another study analyzed an eviction prevention program from 2010 to 2012 in Chicago that distributed rent assistance based on at-risk individuals calling a Homelessness Prevention Call Center. Researchers found that "the volatile nature of funding availability leads to good-as-random variation in the allocation of resources to individuals seeking assistance" and that "those calling when funding is available are 76% less likely to enter a homeless shelter" [5]. The program was especially helpful for the lowest income callers and the authors claim could have been more efficient with better targeting [5]. In New York City a program called HomeBase provides community based services to at-risk individuals. Two papers have concluded that this program is cost effective at preventing homelessness [6, 7].

Homeless prevention is effective on a large scale as well. In 2009 the United States government distributed \$1.5 billion through the Homeless Prevention and Rapid Rehousing Program which promoted HPPs. According to the National Alliance to End Homelessness this spending and promotion of HPPs largely caused the 1% decline in homelessness between 2009 and 2011, which is notable considering the economic recession during that time [8].

There is one important pattern in the pathology of homelessnes that is likely the cause of HPPs' effectiveness - a gradual divergence from financial stability. We hypothesise that individuals that eventually experience homelessness begin thier journey as indistinguishable from the healthy population, but, as time moves forward, these individuals begin to shows signs of financial distress and become more and more separate from the financially healthy population (and more identifiable as at-risk) as they progress towards experiencing homelessnes. As this progression continues, and especially once an individual experiences an episode of homelessness, it becomes more difficult to change their financial trajectory and help them reestablish their financial independence. This hypothesis is supported by Shinn's *Efficient Targeting of Homelessness Prevention Services for Families*, which found that "[t]he single best predictor of eventual homelessness is having previously been in a shelter" [4]. The point of HPPs is to change the trajectories of at-risk individuals early in theiri progression towards homelessness. If an intervention is performed early in the process, only a minor, inexpensive course correction is necessary. If this process is allowed to progress too long, especially if an individual experiences homelessness, the required intervention is much more costly. add source?

better transition There does not appear to be any significant portion of the population that is too high-risk for HPPs to assist. In a study on New York's HomeBase program in 2013 it was found that "[n]o level of risk was too high for families to benefit from services and, indeed, even in the highest decile of measured risk, a majority of families avoided shelter" [8]. This finding is important because it provides statistical backing to the notion that no one is beyond help. It also informs which part of the population HPPs should focus on: people who are predicted to have the highest risk of experiencing homelessness. This portion of the population is most likely to experience homelessness and is the most likely to benefit from assistance.

better transition According to a 2019 study by the California Policy Lab, HPPs should be both effective and efficient. An entirely *effective* program prevents all people from experiencing homelessness. An entirely *efficient* program provides assistance only to those who would experience homelessness if unassisted [9]. A good analogy here is

that HPPs striving to provide assistance to a certain type of person, those who will experience homelessness, is like casting a net to catch a certain type of fish. Programs that "cast a large net" by assisting many people can easily be effective because they are likely to prevent a large number of people from experiencing homelessness. However, it is difficult for such a program to be efficient because it will inevitably provide assistance to many people who would not experience homelessness even if they did not receive assistance. In a similar way a program that "casts a small net," targeting the people identified as most at-risk of experiencing homelessness, can be highly efficient by only providing assistance to people who would have experienced homelessness, but is likely less effective because some people who do not receive assistance will still experience homelessness simply because they were not identified as the most at-risk. The ideal situation is a program that assists all the people who would experience homelessness and none who would not. This ideal is almost certainly unattainable in practice due to the difficulty of identifying who will experience homelessness. The key to providing this targeted assistance is correctly identifying the most at-risk individuals and doing so as early as possible in order to give them the best chance at correcting their trajectory.

Predicting Homelessness

Identifying individuals to enroll in HPPs is a difficult task because everyone is a candidate; anyone might experience homelessness at some point in the future. The path to homelessness may be different for different individuals and people may persist in a high-risk-of-homelessness state without ever experiencing an episode of homelessness. Only when a tipping point is reached - one too many bills are due close together or an unforseen expense is incurred such as a parking ticket or a medical expense - does someone lose their ability to retain housing and experience an episode of homelessness [10]. Because of this ability for some individuals to maintain a state of financial instability but never experience homelessness, even the best prediction models will likely produce false positives, that is, predict some people as having a high risk of experiencing homelessness when they never actually experience homelessness.

Currently, screening for HPPs is almost entirely performed by healthcare workers. Using surveys, their experience, and their intuition they evaluate each individual and determine who gets which aid and how much. There are at times high volumes of individuals seeking assistance, especially in large cities, which slows down this system. Another issue is that the evaluators have biases which may affect their decisions about who gets aid and who does not [4].

The use of statistical models as a screening method shows promise in this area. Statistical models add more objectivity to the screening process; any bias they contain can be measured and corrected. One study found that the use of a screening model reduced the number of applicants to HPPs who were turned down and then eventually entered a shelter from 28.4% to 8.1% [4]. Screening systems using statistical models are easily automated which reduces the burden on healthcare workers, less expensive to operate, and fast which can help speed up the overall process of providing assistance to those in need.

Problem

In order to address the issue of homelessness in Spokane County, the City of Spokane, Avista Utilities, Urbanova, Eastern Washington University an Washington State University formed a consortium, the Spokane Predictive Analytics Group, to determine if utility customer billing data is useful for predicting first-time homelessness and to develop predictive tools to identify those at risk of homelessness. Both the City of Spokane and Avista Utilities provided de-identified residential utility customer billing

data. The City of Spokane provided de-identified homelessness data, Urbanova provided data hosting capabilities, and Eastern Washington University, through this work, analyzed the data and produced a prediction model.

Goals

The main goal of this study is to determine if ubiquitous information contained in utility customer billing data is useful in predicting first time homelessness. We accomplish this analysis by using some of this data that is locally available, but could be found anywhere across the US, to train a statistical model and predict first time homelessness, then evaluate the model's effectiveness.

Our project uses utility customer payment data to predict risk of experiencing homelessness. This data source is attractive for two key reasons. First, it contains information related to individuals' financial health such as utility bill amounts, amount owed to the utility company, and number of times a bill payment was missed. Risk of homelessness is closely linked with an individual's financial health.

The second reason why utility data is attractive is its ubiquity. Almost all households pay some form of utility bill, so this data is available for nearly everyone across the U.S. It may exist in different forms and in different agencies, but similar information on amounts owed and payment default likely exists in all utility-customer situations. If it were a source of useful information in predicting homelessness, any derived models could be used anywhere in the U.S. using similar data sources.

The intention is to combine the utility data, if useful, with other known sources of useful information in predicting homelessness to create the most effective prediction model possible. However, our model could be used to predict homelessness with only the utility data. A convenient model output format, and the one chosen for this project, is a ranked list of individuals and their associated risks of experiencing homelessness. With this information, the HPP using the model can assist at-risk individuals in two convenient ways: remove this?

- Option 1

    1. Assess the pool of resources available to assist at-risk individuals,
    2. determine the number of people that can receive aid based on this pool, say n,
    3. and simply choose the n individuals with highest predicted risk.

- Option 2

    1. Analyze the list of individuals and their risks of experiencing homelessness,
    2. choose a threshold risk value,
    3. then assist all individuals above that threshold. This method may be appropriate if there is a large gap in predicted risks.

Precedent Research

Data Sources

Most other studies do not predict homelessness on the entire population but instead focus on more at-risk subpopulations.

Used other data sources: health, DOJ, other assistance services

Models

logistic regression, cox regression, atrificial neural network, points-based questionaire

Performance Evaluation

Reporting on model performance varies significantly from study to study.

Our Contribution

Call for standardized model performance reporting.

Determine usefulness of ubiquitous utility customer billing data.

Present a model that performs well compared to current research that relies only on ubiquitous utility billing data.

Materials and Methods

The Data

The data used for this project was obtained from Avista Utilities and the City of Spokane. Avista Utilities provided monthly residential utility customer billing information for electricity and natural gas utilities. The City of Spokanen provided monthly combined water, sewer, and garbage utility billing information as well as outcome data on who has been recorded as experiencing homelessness each month from their Community Management Information System (CMIS), which complies with the HUD HMIS (Homeless Management Information System) regulatory requirements.

After the preprocessing steps described below, the data covers 3,674,901 rows, 96,768 Avista billing accounts, 64,728 locations (premises), 91,591 people (357 positive cases and 91,234 negative cases). The timeframe the data covers is from December, 2015, through December, 2020. Prior to December, 2015, Avista used a different account tracking system and it would have been difficult to combine data from both systems. ADD proportion of Spokane population The high degree of imbalance in the dataset, that is, the large number of negative cases, people who were never recorded as experiencing homelessness within the dataset, compared to the small number of positive cases, people who were recorded as experiencing homelessness, is a challenge to predicting homelessness. Specifically, positive cases made up 0.39% of the people represented in the dataset. This imbalance adds difficulty to the model's task of distinguishing between positive and negative cases because it is difficult to identify the few positive cases among the many negative cases, especially if there are many negative cases that are similar to the positive cases. In this study identifying the positive cases in the dataset is much more important than the negative cases.

Data Preperation

Initial data matching and de-identifying was performed by the data team at Avista Utilities. This process involved joining the utility billing data from Avista and the City of Spokane on name, address, and month. This dataset was the joined to the City of Spokane's CMIS data on homelessness where matching was based on last for of social security number. After matching, the data was de-identified by replacing any identifiable characteristics of the data such as name and address with internally generated identification numbers.

At this stage the utility billing activity was tied to utility customer accounts, not individual people, and each account had one or more people associated with it. All

accounts had at least one (main) account holder, but many had additional parties associated with them: cotenants, landlords, third party family members, and third party agencies that could assist in bill payment. The modeling task in this context is clearly people-centric, meaning the trained model should take as input a single person's billing information and output the predicted likelihood of experiencing homelessness. This required the data be reconfigured from account-centric to people centric. To acheive this only the main account holder was associated with each account's billing activity and all other people associated with the account were removed.

Only the variables that were most related to the outcome of homelessness and believed to be widely available accross the U.S. were retained for model fitting. To measure relationships between variables in the data, different association metrics were used depending on the variable types. The following is a list of variable pairing types and the metric used to measure strength of association between the variables: absolute value of Spearman Correlation for continuous-continuous, Point-Biserial Correlation for binary-continuous, Theil's Uncertainty Coefficient for categorical-categorical, and Correlation Ratio for continuous-categorical (with more than two levels). Because this modeling task focuses on prediction, intercorrelation between predictor variables was largely ignored. The chosen outcome was binary so Point-Biserial Correlation and Theil's Uncertainty Coefficient were used; the other measures of association were used for other potential outcome variables - see Discussion. The short list of predictor variables and their definitions is shown in Table 1. These are the variables used in the model. The variable BREAK_ARRANGEMENT can be interpreted as the number of times a person has established a service break arrangement with the utility company.

Feature Engineering

An effort was made to extract additional information from the data by engineering features. The amount owed by each person at the start of each month, TOTAL_CUR_BALANCE, was initially provided as separate attributes: amount owed to Avista in electricity bills, amount owed to Avista in gas bills, and amount owed to the City of Spokane in the combined bill amounts for water, sewer, and garbage. To make the variables used more closely resemble those that might be available in other utilities companies, the amounts owed were aggregated into a single attribute representing total amount owed to all utilities entities. This combined amount was more strongly correlated with the outcome of homelessness than any of the individual amounts owed.

The cumulative number of utility account holders who have lived at each location over time was calculated and recorded as NUM_PER_FOR_PREM. This variable relates to the number of people who have moved away from each location and is likely an indicator of a mismatch between the tenant's desire and ability to live at that location. The housing may be desirable to the tenant, but they are forced to move because they cannot afford to stay either because their financial situation has deteriorated or the rent has increased. The housing may also be undesirable to the tenant and their stay was intended to be temporary from the start, in which case moving away would indicate an improvement in financial health. People move for other reasons besides the housing desirability and their ability to pay rent, but this variable was intended to capture some of the information related to housing desirability and financial health.

The cumulative number of locations where an individual has paid utility bills was also calculated and recorded as NUM_PREM_FOR_PER. This variable is intended to capture the number of places each person has lived, though some people pay utilities at multiple premises for reasons other than moving; they may own multiple properties or pay someone else's utility bills. This variable only captures locations where an individual is a main account holder because only main account holders were retained.

Variable (Feature)	Description
SPA_PER_ID	Unique identifier for each person.
CMIS_MATCH	The binary outcome: If a person is ever recorded experiencing homelessness or not.
PAST_DUE	Bill past due notice sent to customer.
TOTAL_CUR_BALANCE	Combined amount owed by an individual in all utilities (electirc, gas, water, sewer, garbage) on the current month.
BREAK_ARRANGEMENT	Service severance begun - break arrangement established for customer.
NUM_PER_FOR_PREM	Cumulative count of unique individuals recorded at a specific premises (address).
NUM_PREM_FOR_PER	Cumulative count at each month of unique premisses (addresses) recorded for a specific person.

Table 1. All variables used in model fitting and their descriptions.

The Model

We used multivariable binary logistic regression to predict risk of experiencing first time homelessness. This type of model has been used for predicting risk of homelessness in several other studies [11–16] and fits the situation of predicting the likelihood of a binary event based on a set of continuous predictors. For each set of new person's data fed into the model, the model will produce an output of the predicted probability an event will occur given the variable values for that person. This matches the intended use-case for the model.

In the context of this project, the logistic regression model, or log model, treats each person-place-month combination as a separate entity with a separate outcome. To find a prediction for each person, the maximum is taken of all the predictions for each person over time and at all locations. This maximum risk prediction becomes the prediction for that person.

Model Evaluation

The model was evaluated on how well it could predict the outcome for each person. The performance metrics of True Positive Rate (TPR), False Positive Rate (FPR), Positive Predictive Value (PPV), Accuracy, Balanced Accuracy, and F-1 Score, and Area Under the Curve (AUC) are all reported for the model. TPR and FPR are the focus and their formulas are presented in Eq 1 and Eq 2. TPR is also referred to in the literature as Sensitivity, Recall, and Hit Rate. FPR is also referred to as Fall-Out. PPV is also called Precision.

$$\text{TPR} = \frac{\text{True Positives}}{\text{Ground Truth Positives}} \quad (1) \qquad \text{FPR} = \frac{\text{False Positives}}{\text{Ground Truth Negatives}} \quad (2)$$

The model is fit and used to predict the outcome of CMIS_MATCH, but the predictions are likelihoods which have a continuous range from 0 to 1. These must be

binned into binary class predictions before they can be compared to the true outcome values; either the model predicts the individual will or will not experience homelessness. The default binning threshold is 0.5, in which case predictions are classified as negative if they are less than 0.5 and positive if they are greather than or equal to 0.5 [17].

The default binning threshold can be adjusted to better suit model application. The model is intended to be used to inform HPPs which people to assist. Some assistance programs cast a wide net and assist many people who are predicted to experience homelessness. These are typically mass outreach programs that cost very little per person. For this type of use, the model threshold may be chosen to be relatively low, meaning many predictions are binned as positive. This may produce many False Positives, but hopefully most of the people who would experience homelessness are reached, so there are few False Negatives.

In contrast, some assistance programs are more targeted and strive to assist only the most at-risk individuals and provide them with more costly aid. For these programs the threshold may be set relatively high, meaning that only high predicted risks are binned as positive. This top tier may produce fewer False Positives, but it may also cause an increase in False Negatives, meaning not all the people that would experience homelessness are reached. In general the models can be evaluated at many different thresholds and will produce different values of the rates listed in Table UPDATE.

To best evaluate the model predictions, every possible threshold is chosen, the binned predictions are evaluated against the ground truth known outcomes, then FPR and TPR are calculated. To summarize this information the FPR and TPR are plotted as the Receiver Operating Characteristic (ROC) curve. The TPR can be interpreted as the proportion of positives correctly predicted by the model. The FPR can be interpreted as the proportion of the negative cases falsely predicted to be positive [18].

It is standard practice in data science to test prediction models by fitting them to a set of data, use them to predict on data they have not yet seen, then evaluate whether the predictions were correct on this previously unseen dataset. This is achieved in practice by selecting a random subset of the data and saving it apart from the rest. This chunk of data is referred to as the "test set". The remaining data is called the "training set" because this is the data with which the model is fit, or "trained".

To make model evaluation match how the model will be used by a HPP, the data was split over people so the model will train on one group of people, then make predictions on a different group. Unfortunately, the imbalance in the data causes the model to be sensitive to the train-test split because only a few positive cases ever end up in the test set. If the split is lucky, the positive cases in the test set will be similar to those in the training set so the model can recognize them as positive cases. If the split is unlucky, the positive cases in the test set will have different characteristics than those in the training set so the model will have difficulty recognizing them as positives. To combat this sensitivity in model evaluation, the method of K-Folds was implemented.

K-Folds is a model evaluation technique where the data is split into k disjoint groups, or folds. A loop is run where the model is trained and tested on different splits of the data; a different split for each fold. For each model run, one of the data folds is saved as the test set and the remaining folds become the training set. The model is fit and evaluated on this data split, then the whole process is repeated for the next fold. We used $k = 10$ folds, which is the default number of folds suggested in the literature [19].

The model evaluation metrics are calculated on the combined predictions of all the folds so are the average over all of these data splits. In this way the model is always evaluated on data it has not yet seen and the resulting averaged evaluation metrics are less sensitive to each individual train-test split. Also, the combined predictions from all the model runs cover the entire dataset, so every person's outcome is predicted, though some people are predicted by one model fitment and others by another.

The method of naive random oversampling the minority class (positives) was employed to combat the extreme imbalance in the dataset. Once the data split for each fold was established, the training set was balanced by oversampling the minorty class. The test set was left unaltered for each fold.

Data Limitations

The data used for this project is inherently limited in its scope to some degree. Though the main benefit of using utility data is to capture as much of the population as possible, still only people who have been involved with a utility billing account are captured. The data captures 91,591 people, about 17.5% of the total Spokane County population which was 522,798 in 2019 [20].

Another major limitation of the data is incomplete outcome measurement. Positive outcomes (if someone experiences homelessness) were gathered from the City of Spokane's Community Management Information System (CMIS), an organization that collects data from various HAPs in the Spokane area. Everyone who did not appear in the CMIS database was assumed to be a negative case (did not experience homelessness). Positive cases are gathered by this method with a high degree of confidence, but some people who experience homelessness may be mis-labeled as negative cases. This could occur if a person moved out of the CMIS's data collection area before experiencing homelessness or if they simply never sought assistance.

Results

The method of K-Folds was employed with $k = 10$, so ten models were actually trained and act as an ensemble model in practice. The reported performance is an average of all the individual models. If similar performance is to be attained on new data, all models must predict the risk of experiencing homelessness for a new person, then the average prediction will be the prediction of the overall ensemble model.

In order for the model to be applied in practice, a specific binning threshold must be chosen with different thresholds yielding different model performances. The appropriate threshold will depend upon the relative costs of FPs to FNs in the specific context the prediction is being performed. For example, if a HPP uses an expensive intervention, the cost of FPs can be large so a high threshold with few FPs is preferred, but if an inexpensive intervention is used in the same context the cost of FPs is smaller so a lower threshold that correctly identifies more positives is preferred. High thresholds correspond to the lower left region of the ROC plot (Fig 2) where the TPR and the FPR are both low because few people are predicted as positive.

Feature	Mean Coeff	Mean OR	Mean [0.025	Mean 0.975]
PAST_DUE	0.406	1.501	1.499	1.502
TOTAL_CUR_BALANCE	0.001	1.001	1.001	1.001
NUM_PREM_FOR_PER	-0.992	0.371	0.369	0.372
BREAK_ARRANGEMENT	0.800	2.225	2.195	2.256
NUM_PER_FOR_PREM	0.078	1.081	1.079	1.084

Table 2. Mean logistic model coefficient (Coeff), mean odds ratio (OR), and mean 95% confidence interval for the odds ratio for each model feature over all folds. Odds ratio describes how much the predicted odds of an event will increase if a single variable is increased by one.

The odds ratio for each feature shown in Table 2 give some idea of the importance of

each feature to the logistic model. The features were not normalized to the same range so the odds ratio for one feature cannot be directly compared to that of another feature.
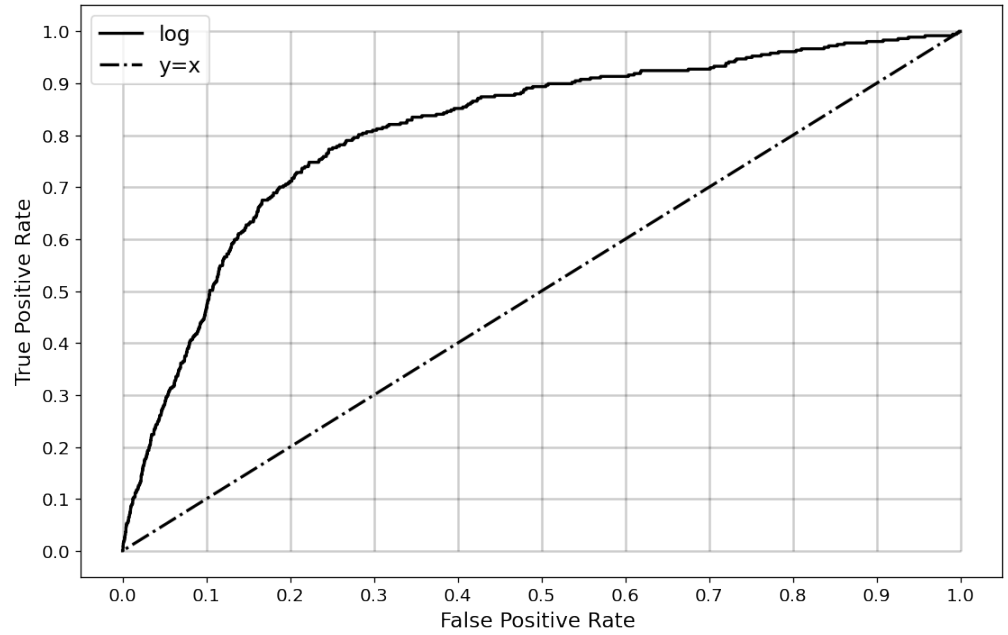
Fig 2. The ROC curve for the Logistic Regression Model. This plot is created by choosing every possible binning threshold, evaluating FPR and TPR, then moving on to the next threshold. The plot consists of only the (FPR, TPR) points. FPR (False Positive Rate) is the proportion of false positives to total ground truth negatives. TPR (True Positive Rate) is the proportion of true positives to the total ground truth positives. AUC (Area Under the Curve) is 0.810. REMOVE GRAPHICS BEFORE SUBMISSION

Threshold	TPR	FPR	PPV	Accuracy	Balanced Accuracy	F-1 Score
0.501	0.812	0.307	0.812	0.694	0.753	0.812
0.560	0.776	0.251	0.776	0.749	0.763	0.776
0.410	0.899	0.508	0.899	0.494	0.696	0.899
0.387	0.922	0.618	0.922	0.384	0.652	0.922

Table 3. Selected thresholds and their corresponding performance metrics. These (FPR, TPR) points are shown in the ROC plot shown in Fig 2. TPR (True Positive Rate) is also referred to as Sensitivity, Recall, and Hit Rate. FPR (False Positive Rate) is also referrred to as Fallout Rate. PPV (Positive Predictive Value) is also referred to as Precision.

In order to make the results of this study comparable to other homelessness prediction studies, the results of our model evaluated at the threshold where the TPR is near the levels listed in the other studies are shown in Table 3. Each row in Table 3 represents a single (FPR, TPR) point on the ROC curve in Fig 2.

Discussion

Model Performance

Performance of models intended for a specific use, such as homelessnes prediction, should be analyzed and presented in a manner most convenient for the intended users of the models. In this case the intended users are HPPs, so the performance of this model is presented in a manner conducive to an HPP shopping for a prediction model for their particular situation.

Comparing binary classification models is difficult because their scores in performance metrics differ based on the chosen prediction binning threshold. The model performance captures how accurately the model predicts positive cases and negative cases; these are both driven by the chosen threshold, but typically the relationship between the threshold and model performance is not simple. This complicated relationship between threshold and model performance causes difficulty in model comparison. For instance, Model A may perform better than Model B at one threshold, but Model B performs better at another threshold. Additionally, at a single threshold, Model A may perform better than Model B in one metric, but worse in another metric. The most helpful way to compare models may be to compare only their performances on positive and negative cases without considering the threshold at all; the threshold can be treated as a model parameter.

To further complicate model comparison, the makeup of the population over which predictions are made has a strong effect on the performance metrics commonly used for binary classification. For example, let Model A predict an event for a population with 10% positives, while Model B predicts the same event for a population with 90% positives. If both models achieve a 0.90 TPR, then Model B has predicted nine times as many true positives, in number, than Model A. REMOVE THIS PARAGRAPH?

Call for Standardized Reporting

Even with the caveats presented above, models from different studies can and should be compared by academics and HPPs. In fact, studies should design their performance reporting to facilitate the comparison of their model with others to allow HPPs to make informed decisions about which model to pursue. This requires some form of standard performance reporting for homelessness prediction models. So far the related research has reported model performances either the default prediction binning threshold of 0.5 or choosing another threshold deemed more appropriate by the authors. The performance metrics reported in this area of research are typically: True Positive Rate (TPR), False Positive Rate (FPR), Positive Predictive Value (PPV), and Area Under the Curve (AUC).

based on the default prediction binning threshold of 0.5 with a variety of performance metrics reported Add citation. Because in this use-case it is much worse to incur a Type II Error (false negative) than a Type I Error (false positive), model performance and evaluation should focus on a threshold where the True Positive Rate (TPR) is high, meaning most of the positive cases are correctly predicted. This focus on correctly predicting positive cases has been repeatedly pointed out in the literature [12] ADD SOURCES HERE. We propose the reporting standard of choosing the prediction binning threshold that produces a TPR as near to 0.90 (90.0%) as possible, then report False Positive Rate (FPR) and other metrics at the same threshold. Reporting the FPR allows homelessness prediction programs to asses the costs and savings of the prediction model based on their service population. Reporting the Area Under the Curve for the Receiver Operator Characteristic is another helpful metric to report that provides a measure of model performance at all thresholds. Displaying the ROC plot is also very

helpful for comparing models.

Comparison with Current Research

We compare our model to those developed in previous studies by choosing the binning threshold that produces most closely the same TPR as reported by each of the previous studies, then comparing the FPR or Positive Predictive Value (PPV), when provided. For the same TPR a lower FPR and higher PPV are preferred. Several studies presented thier model performance in ways allowing for the evaluation of the model at a TPR of 0.90. At this TPR Shinn's Cox Proportional Hazards model achieved a FPR of about 0.61 [8], Toros' Logistic Regression models achieved FPRs of 0.297 (unemployed workers model) and 0.364 (youth receiving puclic assistance model) [16], and Hong's Logistic Regression model achieved a FPR of about 0.69 [15]. At a TPR of 0.889 our model achieved a FPR of 0.508. VanBerlo's HIFIS-RNN-MLP machine learning model achieved a TPR of 0.921 and PPV of 0.651 with no FPR reported [12]. At the same TPR our model achieved a PPV of 0.922. Byrne's Logistic Regression model achieved a TPR of 0.778 with a PPV of 0.117 [11]. At the same TPR our model achieved a PPV of 0.776.

In comparison with the selected studies listed above at the comparable TPR, our model achieved a lower (better) FPR than Shinn's model, a much higher (worse) FPR than either of Toros' models, a lower (better) FPR than Hong's model, a much higher (better) PPV than VanBerlo's model, and a much higher (better) PPV than Byrne's model. It is important to note that all studies where model performance was reported other than Byrne's were predicting homelessness, return to homelessnes, or chronic homelessness based on some subpopulation already seeking government assistance. There is much more relavent data available to predict homelessness on these subpopulations because they have submitted financial, health, employment, etc. information as part of applying for government assistance. Byrne's study predicted homelessnes on all "persons aged 11 years or older who had health insurance between 2011 and 2015 as reported in the Massachusetts All-Payer Claims Database (APCD)" [11]. Byrne's is the study most directly comparable to ours, though Byrne's covers "more than 98% of Massachusetts residents" [11] while our study covers about 17.5% of Spokane County residents.

Data Preparation

Including More People Per Account

The data preparation scheme associated only the main account holder with each account's billing history and removed all other people associated with the account. This step removes many people from the dataset because they are not a main account holder on any account. Only 38% of people are retained after preprocessing, though many of these are removed because their accounts did not exist in the time frame in which billing histories were provided. Ideally as many people as possible would be retained, but testing the inclusion of cotenants and all people associated with an account produced models with much worse performance. This reduced performance is likely due to the account billing history holding little or no information about the financial health of people other than the main account holder.

Data Homogeneity Over Time

For this project we investigated if the relationships between the predictor variables and the outcome changed over time. The year 2020 was expected to contain different relationships than the previous years because the COVID-19 pandemic caused people

and the utility agencies to behave differently than they had in previous years. For example, for the greater part of 2020, the utility agencies tracked in our dataset did not perform service shut-offs as they had in previous years [21].

To investigate if there was a difference in the relationships between the variables and the outcome in different years, the data was split by year, the Spearman Correlation of each variable with the outcome (CMIS_MATCH) was measured, these correlations were then ranked, and the rankings were compared from year to year. Table 4 displays these rankings. Surprisingly, they are very similar from year to year indicating that the important relationships within the data are similar from year to year. Because of this finding, all available years, 2015 to 2020, were included in modeling. Note that NUM_PER_FOR_PREM and NUM_PREM_FOR_PER do not vary in 2015 because only December of 2015 was provided, so they have no meaningful correlation with the outcome. It is assumed that NUM_PER_FOR_PREM and NUM_PREM_FOR_PER would be the top two correlated variables with the outcome if more months from 2015 were provided.

Variable	2015	2016	2017	2018	2019	2020
NUM_PER_FOR_PREM	-	2	3	2	1	1
NUM_PREM_FOR_PER	-	1	1	1	2	2
TOTAL_CUR_BALANCE	3	4	4	4	4	3
BREAK_ARRANGEMENT	4	3	2	3	3	4
PAST_DUE	5	5	5	5	5	5

Table 4. Ranks of Spearman Correlation between each predictor and the outcome, by year. Variables are sorted by 2020 rankings.

Outcome Variable Selection

The eventual choice of outcome measure was CMIS_MATCH, the binary variable corresponding to if a person was ever recorded as experiencing homelessness, but several others were investigated along the the way. One was a numerical variable describing the number of months until an individual experienced an episode of homelessness. This approach fit the multidimensional time-series data well and has a clear risk interpretation (closer times are higher risk), but posed the challenge of assigning some value(s) to negative cases. Two more outcomes investigated were the binary variables corresponding to if an individual was within six months or one month of experiencing an episode of homelessness.

The outcome used for modeling, CMIS_MATCH, was chosen because out of the potential outcome measures it was the most closely associated with the predictor variables. It is surprising that the outcomes corresponding to if an individual will experience homelessness within six months and one month are associated less with the predictor variables than CMIS_MATCH. This seems to indicate that the data used for this project is better at discerning if an individual will ever experience homelessness than it is at discerning either if the individual will experience homelessness within the month or within six months.

Predictor Variable Selection

- selected those that are likely to be widely available across US

- no performance change between grouped, ungrouped billing features (interesting bc landlord likely pays some)

REMOVE THIS subsubsection?

Dependence on Time

It was hypothesized that time is an important factor in predicting homelessness. As people get closer in time to experiencing homelessness they are thought to be more distinguishable from the general population of mostly negative cases. Two approaches were considered to incorporate individual's previous characteristics in predicting the likelihood of homelessness.

The first approach was to generate simple trajectories for each person over time. This was done by fitting a linear line for each variable over time, then retaining only the slope of the line instead of the original values. This method records the long-term trend of a person in each variable over time. The performance of the models using these trajectories was worse than the methods discussed thus far, so it was abandoned.

Another method investigated for including a person's previous values as predictors was to take the difference between months for each variable. These differences capture the local changes in people's characteristics and were used as predictors in the place of the original variable values. Again, model performance was worse than the main methods described for this project.

The highest performing method tried was to treat each person-place-month combination as a separate entity, make a prediction for each entity, then take the maximum risk prediction for each person over all places and months that person appears in the dataset. This is the method used and described in the Methods section. This seems to indicate the important aspect of each person is their peak risk.

Methods more appropriate to prediction on time-series data should be tried in this context. We implemented a simple Long-Short Term Memory (LSTM) model, but its performance was much worse than the logistic regression model. This may be due to the author's inexperience with LSTM models and should be investigated further. Cox Regression was also investigated, but was deemed not to fit the situation because here the likelihood an event will occur is assumed not to be directly related to time, where Cox Regression assumes likelihood of an event never decreases over time.

Data Limitations

The first limitation is the incompleteness of the data; only a small portion of the overall population was represented. The utility billing data was structured to track accounts and ensure the people responsible for those accounts paid their bills on time. This data only contained financial information relevant to the main account holders, which turns out to be only ADD population percentage here.

There were often multiple people associated with a single utility account, but it was not recorded who paid for each bill. For this project it was assumed that the main account holder always paid since they were ultimately financially responsible for the account, but this is likely not always true. There may be useful information regarding who paid a utility bill for each month; specifically, it may be a risk indicator for the main account holder if someone other than the main account holder paid the bill for a given month. It also may be a positive financial health indicator for the bill payer.

The question of who paid each utility bill is made more difficult to answer by the fact that Avista and City utility billing information was matched using address, not person. This choice was made because matching on address and month across the billing systems of Avista and the City of Spokane had a much higher match rate than matching on the recorded account holder names. It is likely that many of the City utility bills for apartments were paid by the landlord and so would remain unaltered if the tenant was experiencing financial stress. This aspect is addressed, to some degree, by the selection of explanatory variables. It was found that the aggregated measure TOTAL_CUR_BALANCE, the combined amount owed to Avista and the City, was most

associated with the outcome measure over the amounts owed to the individual entities. This indicates that even if many landlords did pay consistently for the City utilities even for people who were at risk of homelessness, there was enough information in the amounts owed in City utilities regarding homelessness that aggregating the Avista and City amounts owed produced a variable more strongly associated with CMIS_MATCH than the individual variables.

Another important limitation of the data is our confidence in the accuracy of the outcome labels. Of all the people in the data labeled as positive, our confidence that they are actually positive is high - they matched someone who was recorded as experiencing homelessness in the Spokane area. Our confidence in the correct labeling of the negative cases, on the other hand, is lower. Some of the cases labeled as negative in the data may, in reality, be positive cases if they did experience homelessness but this fact was not recorded in the City of Spokane CMIS database. This can occur if someone experiencing homelessness does not seek assistance or if they move away from the Spokane area before experiencing homelessness. Because some of the negative cases may be positive cases in reality, the model may have predicted them as positive because they have similar characteristics to other positive cases, but that prediction was labelled as incorrect. This is an unknown factor in this data, but it may cause the model performance to appear worse than it actually is. This problem is likely common to the task of predicting homelessness.

A point of interest is that the people predicted as positive by the models may be members of a more general at-risk population. The only outcomes we had data on were related to homelessness, but the model predictions and the data investigation reveal that many people labeled as negatives have similar characteristics to the positive cases. These FP's may be financially stressed or at-risk in ways other than their housing status, but we did not have data on other types of outcomes to investigate this. One study that investigated both homelessness and drug addiction found that a high predicted likelihood of homelessness was strongly correlated with drug use [11].

Variables related to housing and location were hypothesized to be a important predictors of homelessness, but were challenging to include as predictors in the models. Information such as rent, desirability of living location, and age and condition of residence should be directly related to an individual's financial health and therefore be good predictors of homelessness. The variables available that contain geographical location information were POSTAL, the person's postal code, BLOCKGROUP_GEOID, the U.S. Census block group number, and SPA_PREM_ID, the assigned identification number for a specific residence. The unit of block group is smaller than that of postal code, but block group areas often do not fall within a single postal code boundary. All of these variables are categorical with POSTAL having 17 unique levels represented in the data, BLOCKGROUP_GEOID having 279, and SPA_PREM_ID having 64,728. Because BLOCKGROUP_GEOID and SPA_PREM_ID had so many possible values, they could not be incorporated into the logistic model without causing convergence issues. POSTAL was too general to have much meaningful information regarding such a rare event as homelessness.

Other approaches to incorporating location information as a predictor should be investigated. Some prior analysis could be performed to determine a small number of at-risk areas such as low-rent and subsidized rent areas. These could be represented in a few-leveled categorical variable. Individuals in the study could be determined to reside in one or none of these areas.

Of course, there is information more directly related to homelessness than we had access to in this study. ADD useful information here

Opportunities for Future Research

This study has demonstrated the utility of utility customer billing data in the prediction of first-time homelessness. This data could be combined with other data related to homelessness to produce higher performing prediction models.

Also, as previously mentioned, further investigation into the trends of people over time as well as the incorporation of location information as a predictor could yield higher model performance.

Conclusion

Supporting information

Acknowledgements

References

1. Exchange H. Criteria and Recordkeeping Requirements for Definition of Homelessness; 2012. Available from: https://www.hudexchange.info/resource/1974/criteria-and-recordkeeping-requirements-for-definition-of-homeless/.

2. Exchange H. Point-in-Time Count and Housing Inventory Count; 2020. https://www.hudexchange.info/programs/hdx/pit-hic/.

3. Henry M, Watt R, Mahathey A, Ouellette J, Sitler A, Associates A. The 2019 Annual Homeless Assessment Report (AHAR) to Congress. The U.S. Department of Housing and Urban Development; 2020.

4. Shinn M, Cohen R. Homelessness Prevention: A Review of the literature. Center for Evidence-Based Solutions to Homelessness. 2019;.

5. Evans WN, Sullivan JX, Wallskog M. The impact of homelessness prevention programs on homelessness. Science. 2016;353(6300):694–699.

6. Rolston H, Geyer J, Locke G, Metraux S, Treglia D. Evaluation of the homebase community prevention program. Final Report, Abt Associates Inc, June. 2013;6:2013.

7. Goodman S, Messeri P, O'Flaherty B. Homelessness prevention in New York City: On average, it works. Journal of housing economics. 2016;31:14–34.

8. Shinn M, Greer AL, Bainbridge J, Kwon J, Zuiderveen S. Efficient targeting of homelessness prevention services for families. American journal of public health. 2013;103(S2):S324–S330.

9. Von Wachter T, Bertrand M, Pollack H, Rountree J, Blackwell B. Predicting and preventing homelessness in Los Angeles. California Policy Lab and University of Chicago Poverty Lab. 2019;.

10. O'Flaherty B. Wrong person and wrong place: for homelessness, the conjunction is what matters. Journal of Housing Economics. 2004;13(1):1–15.

11. Byrne T, Baggett T, Land T, Bernson D, Hood ME, Kennedy-Perez C, et al. A classification model of homelessness using integrated administrative data: Implications for targeting interventions to improve the housing status, health and well-being of a highly vulnerable population. PloS one. 2020;15(8):e0237905.

12. VanBerlo B, Ross MA, Rivard J, Booker R. Interpretable machine learning approaches to prediction of chronic homelessness. Engineering Applications of Artificial Intelligence. 2021;102:104243.

13. Van den Bree MB, Shelton K, Bonner A, Moss S, Thomas H, Taylor PJ. A longitudinal population-based study of factors in adolescence predicting homelessness in young adulthood. Journal of Adolescent Health. 2009;45(6):571–578.

14. Flaming D, Burns P. Crisis indicator: Triage tool for identifying homeless adults in crisis. Economic Roundtable Research Report; 2011.

15. Hong B, Malik A, Lundquist J, Bellach I, Kontokosta CE. 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. 2018;36(1):89–104.

16. Toros H, Flaming D, Burns P. Early intervention to prevent persistent homelessness Predictive models for identifying unemployed workers and young adults who become persistently homeless. Available at SSRN 3370634. 2019;.

17. Bewick V, Cheek L, Ball J. Statistics review 14: Logistic regression. Critical care. 2005;9(1):1–7.

18. Fawcett T. An introduction to ROC analysis. Pattern recognition letters. 2006;27(8):861–874.

19. Marcot BG, Hanea AM. What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? Computational Statistics. 2020; p. 1–23.

20. Bureau USC. QuickFacts Spokane County, Washington; 2019. `https://www.census.gov/quickfacts/fact/table/spokanecountywashington`.

21. White R, Sokol C. Utilities pause shutoffs as state prepares for financial impacts of COVID-19. The Spokane Review. 2020;.