Data & Policy (2021), 3: e2
doi:10.1017/dap.2020.23

RESEARCH ARTICLE

A recommendation and risk classification system for connecting rough sleepers to essential outreach services

Harrison Wilde[1], Lucia L. Chen[2], Austin Nguyen[3], Zoe Kimpel[4], Joshua Sidgwick[5], Adolfo De Unanue[6], Davide Veronese[7], Bilal Mateen[5], Rayid Ghani[8] and Sebastian Vollmer[1,5,*]

[1]Department of Statistics, University of Warwick, Coventry, United Kingdom
[2]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom
[3]Data Science, Tripadvisor, Needham, Massachusetts, USA
[4]Master in Data Science, Northwestern University, Chicago, Illinois, USA
[5]The Alan Turing Institute, London, United Kingdom
[6]Departamento de Matemáticas, Instituto Tecnologico Autonomo de Mexico, Mexico City, Mexico
[7]Master in Public Policy Candidate, Harvard Kennedy School, Cambridge, Massachusetts, USA
[8]Machine Learning Department and Heinz College of Information Systems and Public Policy, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
*Corresponding author. E-mail: svollmer@turing.ac.uk

The four first authors (Harrison Wilde, Lucia Lushi Chen, Austin Nguyen, and Zoe Kimpel) contributed equally to this work.

Received: 28 August 2020; **Revised:** 26 December 2020; **Accepted:** 06 November 2020

Key words: decision support; homelessness; prioritization; risk classification; social good

Abstract

Rough sleeping is a chronic experience faced by some of the most disadvantaged people in modern society. This paper describes work carried out in partnership with Homeless Link (HL), a UK-based charity, in developing a data-driven approach to better connect people sleeping rough on the streets with outreach service providers. HL's platform has grown exponentially in recent years, leading to thousands of alerts per day during extreme weather events; this overwhelms the volunteer-based system they currently rely upon for the processing of alerts. In order to solve this problem, we propose a human-centered machine learning system to augment the volunteers' efforts by prioritizing alerts based on the likelihood of making a successful connection with a rough sleeper. This addresses capacity and resource limitations whilst allowing HL to quickly, effectively, and equitably process all of the alerts that they receive. Initial evaluation using historical data shows that our approach increases the rate at which rough sleepers are found following a referral by at least 15% based on labeled data, implying a greater overall increase when the alerts with unknown outcomes are considered, and suggesting the benefit in a trial taking place over a longer period to assess the models in practice. The discussion and modeling process is done with careful considerations of ethics, transparency, and explainability due to the sensitive nature of the data involved and the vulnerability of the people that are affected.

Policy Significance Statement

This study not only empowers the UK's largest homelessness charity organization in their mission to reach rough sleepers, but also offers opportunities for immediate and long-term policy-level impact. The previously untapped data generated by the StreetLink platform is now a robust source of information for gaining insight into homelessness across the UK, in terms of both its volume and its distribution. Our contributions empower

https://doi.org/10.1017/dap.2020.23 Published online by Cambridge University Press

Homeless Link to audit their processes and work on exposing systemic issues at the root of homelessness as they augment their outreach referral platform with our transparent and equitable recommendation system. The resulting data provide compelling quantitative arguments for their partnership with the government to incite positive change for the country's homeless population.

1. Introduction

Homelessness and rough sleeping comprise a pressing and worsening global issue that negatively affects a population through a host of societal and health-related pressures spanning poverty, illness, and abuse (Fetzer et al., 2019). The United Nations Human Settlements Program estimates that 1.1 billion people are living in inadequate housing, and the available data suggest that more than 100 million people have no housing at all (Institute of Global Homelessness, 2019). Homelessness affects people in every region of the world, developed and developing, and in the absence of government-level coordinated action it is likely to continue growing (Ortiz-Ospina and Roser, 2019).

Rough sleeping is defined by the UK government[1] as a category of homelessness referring to the act of sleeping in the open air or other places not designed for human habitation; carried out by people who do not have access to permanent, consistent shelter (Ministry of Housing, Communities and Local Government, 2019). A rough sleeper is vulnerable even relative to the homeless population as a whole; they are more likely to experience violence, health-related issues, sexual exploitation, and substance abuse (Meinbresse et al., 2014). Female rough sleepers are especially disadvantaged as they tend to be younger in age, require more mental health support than men, and are more likely to be victims of domestic violence. Because of this, female rough sleepers are anecdotally known to hide themselves for safety reasons, something that is shown to be true in Homeless Link's (HL) data. This behavior decreases the known number of female rough sleepers in the UK and means it is often harder for them to find support (Pleace and Bretherton, 2018). On any given night in England, there are an estimated 4,700 people sleeping rough on the streets; this represents a 169% increase in the rough sleeper population from 2010 to 2019 (White and Maguire, 2018; Ortiz-Ospina and Roser, 2019). Additionally, when censused, 60% of those sleeping rough in London were new to the streets that night, further evidencing the transient and often spontaneous nature of homelessness.

Data-based approaches aimed at tackling and properly quantifying these issues are severely lacking; it has been shown that existing statistics disproportionately affect minorities and exhibit gender biases (Toro, 2007; Toro et al., 2007). HL[2] is a UK membership charity organization working to end homelessness in England and Wales. One way HL seeks to achieve this mission is by operating a platform called StreetLink[3] that serves as a conduit for communication between members of the public and over 300 local service providers (LSPs) spread across the UK. The platform allows for members of the public, as well as rough sleepers themselves, to submit an alert via phone, web, or mobile app regarding someone who is potentially sleeping rough.

The alerts are passed on to a team of volunteers at StreetLink who manually review them for quality before dispatching them as referrals to LSPs, who then attempt outreach and report on the outcome. This review process is based on the information contained in the alerts such as descriptions of the person and their location. Due to the resource constraints seen at the LSP-level, only alerts with sufficient information to locate the reported individual are turned into referrals. This review process takes a significant amount of time, and it's timeliness and quality can negatively impact the scale and speed of service that charities are able to provide. This is because much of the information in an alert will not necessarily be valid more than

[1] https://www.gov.uk/guidance/homelessness-data-notes-and-definitions

[2] https://www.homeless.org.uk

[3] https://www.streetlink.org.uk

https://doi.org/10.1017/dap.2020.23 Published online by Cambridge University Press

a day after it was submitted as rough sleepers move around or have their situations change at short notice. Moreover, the volume of alerts can overwhelm the system during periods of extreme weather and at the peak of winter. The high volume of alerts leads to a bottleneck, meaning not all alerts can be effectively processed despite the urgency of rough sleepers' needs, especially during these potentially more dangerous and difficult times.

The success of StreetLink and it's impact on affected individuals depends on whether the local outreach teams can successfully find rough sleepers based on the information contained in referred alerts. According to StreetLink's historical data from 2012 to 2019, only 14% of received alerts resulted in successfully connecting with a rough sleeper. Although this percentage has increased over time, it is still only around 20% in 2019.

2. Contributions

In collaboration with HL, our objective was to augment StreetLink's review process by building a recommendation system using machine learning classifiers to assist in a human-centered approach to improve the process of assessing the quality of an alert. A quality alert is one that includes sufficient information for LSPs to locate the rough sleeper in question. Without sufficient resources to review alerts in a timely manner, the quality of an alert passed on to LSPs decays as the information on the rough sleeper's location gets stale. Sixty-five percent of the alerts received by StreetLink originate in London; resources are especially constrained here due to demand combined with the urbanized distribution of population making the problem all the more critical.

If StreetLink volunteers and LSPs could focus their limited time, resources, and expertise on higher quality alerts, more rough sleepers could be connected with outreach services to alleviate their situations. Our solution ensures the quality of referrals, minimizes resource waste while taking equity into account and providing a prioritized list for StreetLink to review. To achieve this, several machine learning classification algorithms were applied to two problems: (a) identifying alerts that are likely to lead to a referral being made, and (b) of these referrals which ones are likely to result in a positive outcome for the rough sleeper. We adopt a human-centered algorithmic approach to augment the existing manual review process through the integration of knowledge and evidence gathered by StreetLink volunteers to engineer features in models that are built for explicit transparency and interpretability so that their input to decision-making is fair and ethical. The work throughout is explicitly conscious of the ethical concerns surrounding biases in such solutions and care is taken in comprehensively identifying potential model biases and group harms. This effort contributes to the ongoing mission for improved baseline expectations regarding the evaluation machine learning researchers should carry out when building similar systems to serve charities and vulnerable populations. To summarize, our main contributions are:

1. A novel approach to prioritizing incoming alerts to maximize the chance that a rough sleeper is connected with, validating the fairness of our models using a bi-model approach to ensure that referrals and suggestions are made fairly across demographics.
2. We identify a set of characteristics that could facilitate the establishment of a connection between rough sleepers and outreach teams. Previously untapped insights from HL's data could incite policy change and lead to longer term positive impact.

2.1. *Overview of the existing (manual) approach*

StreetLink manages a phone line, website, and mobile application which all feed alerts into a centralized system. The majority of these alerts are subsequently reviewed by volunteers for quality and the potential for duplication with existing alerts. Following discussions with volunteers, we can define their process as a search for the following three characteristics:

1. A location that is accessible for an outreach team and not near a known hotspot for street activity where regular outreach is done regardless.

2. Evidence that the rough sleeper has bedded down or is likely to bed down in that location.
3. Sufficiently helpful descriptions about the rough sleeper's appearance and location.

These criteria are heuristic and based on the experiences of the outreach teams that StreetLink works with and represent a baseline decision-making process for comparison. This process aims to ensure that the limited resources will go to individuals who have no choice but to sleep on the street instead of those who engage in street activities but already have a place in a shelter. In addition, StreetLink applies a simple rule-based algorithm to check for potential duplicated alerts prior to manual review (based on whether another alert was raised within 50 m in the past week). StreetLink volunteers then confirm these duplicated cases during their review.

2.2. Data-driven approach

The problem was formulated as: how can we best categorize incoming alerts based on whether they have sufficient information for outreach workers to connect with a rough sleeper, and will the eventual outcome be positive? Whilst duplicate alerts present a problem for volunteers by increasing their workload, we expect our models to use information in duplicate alerts co-operatively to gain a greater understanding of which features lead to positive outcomes. Additionally, it was necessary to build models to better understand the referral process carried out by StreetLink so as to ensure equity in the services delivered. These two components are classification problems with binary labels corresponding to whether a referral ends in a positive outcome (i.e., whether a rough sleeper will be found or not) and whether a referral was made, respectively.

Data provided by StreetLink were used to train a pair of models to be used in the review process shown in Figure 1:

1. *Positive Outcome Model*: trained on outcomes indicating whether or not a person was found within a week following a referral; alerts that did not turn into referrals were therefore excluded from the training data for this model.
2. *Referral Model*: trained on outcomes indicating whether or not an alert was turned into a referral by StreetLink.

The Referral Model's purpose lies in emulating the current process of reviewing incoming alerts at StreetLink, and can be used alongside the Positive Outcome Model to highlight alerts that should have been turned into referrals as they were likely to have led on to positive outcomes. This system can be used by StreetLink in the following way:

1. Alerts with a high enough score from the Positive Outcome Model can automatically be sent to LSPs without manual review at the discretion of StreetLink. This frees up resources for volunteers to spend more time on reviewing and following up on the more nuanced and complex alerts.
2. All other incoming alerts should be ranked based on the Positive Outcome Model's scoring so that StreetLink volunteers review the most promising alerts first in order to best use the information that they contain by quickly turning these alerts into referrals.

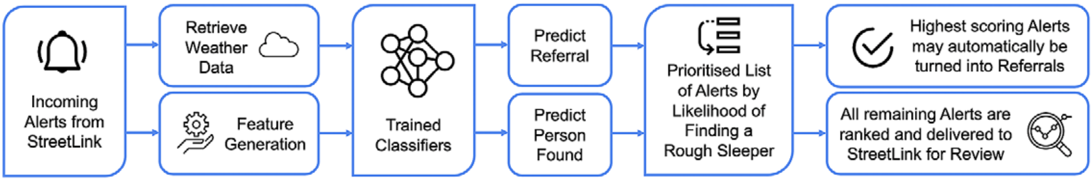
Incoming Alerts from StreetLink
Retrieve Weather Data
Feature Generation
Trained Classifiers
Predict Referral
Predict Person Found
Prioritised List of Alerts by Likelihood of Finding a Rough Sleeper
Highest scoring Alerts may automatically be turned into Referrals
All remaining Alerts are ranked and delivered to StreetLink for Review

Figure 1. *The proposed alert prioritization process.*

https://doi.org/10.1017/dap.2020.23 Published online by Cambridge University Press

3. Data

3.1. Primary data

Data were collected through StreetLink's alert reporting process from 2012 to 2019. The primary dataset contains 281,000 alerts, of which 167,000 are relevant to London and 114,000 are for non-London regions. 170,000 alerts were turned into referrals and of these 39,000 resulted in a positive outcome in which a person was found. Figure 2 shows the number of alerts, referrals, and positive outcomes per month from 2012 to 2019. There is a strong seasonal effect on the volumes due to the public's increased awareness of rough sleepers during winter and the strain placed on service providers during this time. There was a significant operational change in December 2017 that altered the way in which StreetLink collected data; for the remainder of this paper, we define our dataset to be the alerts from December 2017 to March 2019 (the end of the full dataset).

The raw data contain 43 fields that fall into the following five categories:

1. Demographics (age and gender)
2. Outcomes and labels (signifying whether a person was found or not, or if a referral was created, as well as a number of other possibilities)
3. Temporal information (time of alert creation and resolution, time a rough sleeper was seen at the location described)
4. Location data (latitude and longitude provided by the user placing a pin on Google maps, full street address)
5. Free text data (appearance description, location description, immediate concerns about the rough sleeper)

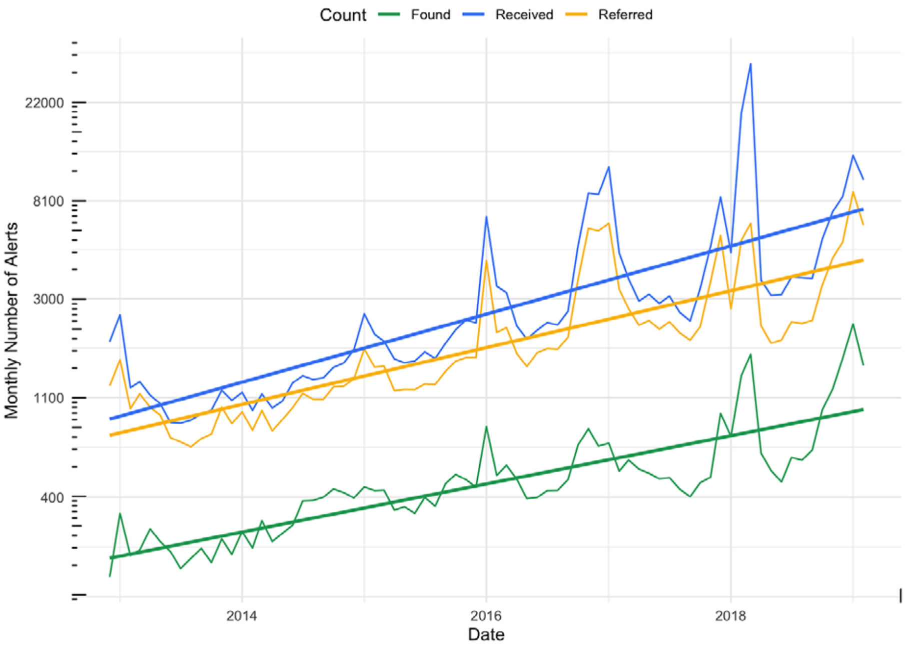

Figure 2. *StreetLink's platform is experiencing exponential growth, likely due to increased awareness of the problem and StreetLink itself. This increased level of demand has not been matched with an increased level of resources. The number of referrals made and people connected with follow a similar but slightly smaller exponential growth indicating the severity of the challenge that StreetLink faces.*

https://doi.org/10.1017/dap.2020.23 Published online by Cambridge University Press

3.2. External data

In addition to the data provided by StreetLink, we retrieved historical weather data for London (latitude: 51.50, longitude: 0.1278) through DarkSky's API.[4] The weather data range from 2012 to 2019, and include hourly information about the temperature, wind speed, and precipitation.

3.3. Data storage and integration

The data were stored in a PostgreSQL database in a Tier 3 Secure Environment hosted by The Alan Turing Institute, which helps prevent data loss and security breaches, see Arenas et al. (2019). Multiple database schema were created to store the data, features, and model results.

3.4. Data preparation

3.4.1. Cleaning the primary dataset

The historical data used throughout are an extract from StreetLink's Salesforce database. We renamed and formatted the variables for consistency, standardized text strings to lowercase, ensured empty values were recorded as NULLs, and standardized the format of temporal variables.

3.4.2. Labels

We created two levels of labels corresponding to each model discussed above. The first of which is a binary categorical variable capturing whether an alert was turned into a referral. The second label is also a binary categorical variable representing whether the outcome was positive, albeit one that can take NULL values in certain scenarios: where a referral was not made and the would-be outcome is essentially unknown. The data originally contained 24 different outcome values that were mapped appropriately in order to define these binary labels. A number of alerts did not have clear outcomes, including alerts that were not handled by local authorities due to a lack of resources or that were still open following a referral; we removed these alerts from the training dataset. Examples of some non-sensitive mappings are provided in Table 1.

3.5. Features

Two hundred and seventy features were created in total that can be split into five groups. All of the categorical variables representing demographic information were converted into dummy variables. Date-time features were created in two dimensions using sine and cosine functions to ensure consistent spacing between the last day of one month and the first day of the next, etc. Word counts of each free text field were also included.

3.6. Spatio-temporal features

3.6.1. Distances from known hotspots

Known hotspots (locations where StreetLink will not send alerts for as they know regular outreach is already carried out) were extracted using one of the outcomes in the data. Following this, a set of spatial variables was created to indicate if an alert was raised within x meters of a known hotspot, $x \in X := \{50,\ 100,\ 250,\ 500,\ 1{,}000,\ 5{,}000,\ 10{,}000\}$.

Table 1. *Examples of label mappings for the Referral and Positive Outcome Models.*

Outcome	Referral label	Positive outcome label
Person found	Yes	Yes
LSP did not respond	Yes	NULL
Person not found	Yes	No
Not enough information	No	NULL

Abbreviation: LSP, local service provider.

[4] https://darksky.net/dev

https://doi.org/10.1017/dap.2020.23 Published online by Cambridge University Press

3.6.2. Alerts, referrals, and people found by geographic location, time, and source

Similarly to the features described above, we generated spatio-temporal aggregate features to identify the number of alerts, referrals, and positive outcomes within $x \in X$ meters, within the last $y \in Y$ days and for $z \in Z$ sources of alerts, $Y{:=}\{7,\ 28,\ 60,\ 360\}$ and $Z{:=}\{\text{Phone},\ \text{Website},\ \text{MobileApp}\}$. One example of this is the count of alerts received via phone within the last 28 days and within 100 m of a given alert.

3.6.3. LSP-level features

Additionally, a set of features was created to represent the response statistics for different LSPs, including the average response time (based on the time between receipt of a referral and an outcome being reported), alert count, referral count, positive outcome count, and person found rates for each LSP. These features were again created over various $y{\in}Y$ temporal windows to capture trends in efficacy and activity at different resolutions.

3.7. Other features

3.7.1. Weather

The aforementioned weather data were aggregated by day to create the following features:

1. Temperature (maximum, minimum, average)Precipitation probability (maximum, minimum)
2. Binary flag for the presence of any snow accumulation
3. Wind (average speed, maximum gust)
4. Missing values were imputed by carrying forward the most recent observations in these cases.

3.7.2. Topic features

Alerts that provide more information on the location and activity of a rough sleeper are in general more useful to LSPs. Location and activity-related entities were extracted using pre-trained entity embeddings from the Python package SpaCy (Honnibal and Johnson, 2015). These entities were then grouped via an unsupervised learning technique—Latent Dirichlet Allocation (LDA) (using the Gensim package in Python, Řehůřek and Sojka, 2010). LDA is a generative statistical model to identify groups of topics in the free text fields. In this application, each field was viewed as containing a mixture of location and verb-related topics (Blei et al., 2003). The algorithm estimates a score to represent the likelihood of an alert's fields belonging to a certain topic. These probability scores were then used as feature variables in the models.

LDA was carried out with various numbers of topics and it was found that the 10 topic solution was most representative of the extracted entities. The location topics encompassed entities including names of parks, hotels, train stations, streets, and so on. The same technique was used to extract 10 topics from the activity descriptions, hoping to separate alerts into those that described someone sleeping rough explicitly and those that described other street activity. These extracted activity-related topics included begging and sleeping, but many of the topics included the same activities. Therefore, we opted to also manually define two sets of topics to identify the activities of the rough sleepers: one containing sleep-related words ("tent," "duvet," etc.) and another including words related to begging behaviors ("beg," "small change," etc.). Word counts for all of these entities and LDA topics were used as features.

4. Modeling

Due to the temporal dependencies present in the data, a month forward-chaining temporal cross-validation approach (Varma and Simon, 2006; Roberts et al., 2017) was adopted to ensure our model error estimates were robust across the entirety of the dataset. The alerts were split by month into folds of increasing length, beginning with the first month as training/validation data and the following month as test data for evaluation. Each subsequent fold's training data is defined as the concatenation of the previous fold's test and training data; it then uses the next month in sequence as test data, and so on. Model performance was

https://doi.org/10.1017/dap.2020.23 Published online by Cambridge University Press

then calculated as summary statistics of the performance across all folds. This idea generalizes to varying periodicities but was carried out as described for a balance in robustness and model performance.

Additionally, there is a time span between an alert's creation and an outcome being provided. Therefore, for each train and test subset within a fold, we needed to define a period within which an outcome would be accepted; removing other alerts that were open as of the date defining the endpoint of the training set. In our experiment, we set this to be a week. This was necessary in order to avoid a clairvoyant model that could be trained on alerts that had outcomes which occurred during the period defined by its test set. As such, a week long buffer was maintained between the train and test sets as well as at the end of the test set, to allow for the final alerts in each subset to also have outcomes within our defined period of a week.

A grid search of parameters was carried out for a series of classification algorithms, including ensemble models (Random Forest, Extra Trees), gradient boosting models (Adaptive Boosting), Decision Trees, and dummy classifiers picking at random in a stratified manner consistent with the training data labels.

5. Evaluation

Thousands of classification models were trained and evaluated for the two classification tasks. The models generate lists of alerts ranked according to their predicted likelihood of a positive outcome, or of a referral being made. Models can then be *evaluated at varying* ***k*** by supposing that the model's top ***k*** alerts by score are its suggested referrals. Then the metrics evaluated on these alerts can be compared to the real statistics from StreetLink (when ***k*** = *the number of referrals made by HL that fold/month*) and with other models. It is impossible to calculate true precision and recall for the Person Found Model because outcomes for alerts where referrals were not made are unknown; we opt instead for the altered metrics defined in Table 2. Since each model involves training a set of nested models through the aforementioned cross-validation technique, the metric values reported in the results tables are the averaged values across all folds. Our objective is to choose a model which maximizes all of the metrics, but with a particular focus on recall during model selection due to the implications of missing a positive outcome that StreetLink did not.

5.1. Defining the baseline

For the Referral Model, there is no meaningful baseline to compare against as it is impossible to beat StreetLink's human review process whilst using the labels directly defined by their actions. However, the number of referrals made by HL on a monthly basis can still provide some insight on how well our models emulate their referral process.

For the Positive Outcome Model, we can consider metrics at ***k*** as defined in Table 2 and define two baselines:

1. StreetLink's manual review process, where we can compare our models' found rates to that of StreetLink by observing the models at ***k***'s matching the monthly referrals by StreetLink shown in Table 3. The ability to also look at our models' found rate and precision at lower values of ***k*** for each month is compeling in justifying the partial automation of referral for the alerts that the models

Table 2. *Metric definitions.*

Metric	Description
Precision at ***k***	Total number of people found in the top ***k*** alerts sorted by the model output scores and divided by the total number of people found or not found in the top ***k*** alerts (excluding NULL labeled alerts)
Found rate at ***k***	Number of people found in the top ***k*** alerts sorted by the model output scores and divided by ***k*** (including NULL labeled alerts)
Recall at ***k***	Number of people found in the top ***k*** alerts sorted by the model output and divided by the total number of people found in that period's alerts

Table 3. *Baseline homeless link statistics by fold/month.*

Fold/month	Positive outcomes	Referrals	Found rate
January 2018	741	2,707	0.2737
February 2018	1,373	5,421	0.2533
March 2018	1,706	6,442	0.2648
April 2018	625	2,279	0.2742
May 2018	524	1,909	0.2745
June 2018	467	1,972	0.2368
July 2018	599	2,373	0.2524
August 2018	582	2,332	0.2500
September 2018	646	2,407	0.2684
October 2018	970	3,448	0.2813
November 2018	1,199	4,532	0.2646
December 2018	1,639	5,334	0.3073
January 2019	2,323	8,867	0.2620
February 2019	1,521	6,331	0.2402

assign high scores to. However, it is difficult to formulate any rigorous comparisons with the baseline at lower k due to the fact that StreetLink do not currently rank the referrals they send in a meaningful way.

2. In order to compare precision and recall more clearly, a somewhat trivial baseline was formulated using a stratified dummy classifier that predicts based on the distribution of training labels. This baseline can be used for both referrals and alerts.

5.2. Model comparisons and choices

The chosen Positive Outcome Model was a Random Forest Classifier with 10,000 trees and a maximum tree depth of 5 (see Table 4). The chosen Referral Model was a Random Forest Classifier with 10,000 trees and a maximum tree depth of 10 (see Table 5). To arrive at these choices we compared models on the averaged metrics defined in Table 2 and shown in Figure 4, as well as through a number of other means described below.

Monthly statistics for StreetLink shown in Table 3 were used to compare found rates for the Positive Outcome Models and the real found rates across each fold. Here the Random Forest classifiers performed the best consistently, especially those with a large number of trees.

Figure 3 is an example of a type of plot used for evaluation in which scores from each of the chosen models are plotted against each other and points are colored according to the true Positive Outcome labels. The black lines are representative of the baseline in that there are a number of points to the right of the vertical line equal to the real number of people found in that month/fold. Similarly, there is an equal number of points above the horizontal line to the real number of referrals made in that month/fold. It can be seen that the majority of NULL-labeled points fall into the Positive Outcome score range of 0.45–0.5, whilst a lot of the alerts with a positive label are found in the top right quadrant of the graph. Interestingly, there are a significant number of alerts (approximately 20%) within the bottom right quadrant that were not made into referrals per our Referral Model *or* the true label, but that our Positive Outcome Model suggests would have led to positive outcomes with reasonable certainty. One of the initial objectives of this work was to increase StreetLink's efficiency so that they are able to process more referrals and potentially explore different types of referrals than they usually would; this quadrant would be a good place to start.

Further analysis by quadrant reveals that the top quadrants for Figure 3 and similar graphs for each fold contain alerts with higher word counts in the free text fields than the average. The top right quadrant tends to include alerts that fulfill the criteria initially outlined in the Overview of the Existing Manual Approach

https://doi.org/10.1017/dap.2020.23 Published online by Cambridge University Press

Table 4. *Results table for the best Positive Outcome Model.*

K	Precision	Recall	Found rate	NULL count
50	0.7978	0.02679	0.5871	13
100	0.7757	0.04932	0.5443	30
150	0.7487	0.0708	0.5162	47
200	0.7233	0.09118	0.4968	63
300	0.7004	0.1283	0.4650	102
400	0.6760	0.1591	0.4343	145
500	0.6615	0.1867	0.4137	189
750	0.6404	0.2560	0.3807	305
1,000	0.6209	0.3193	0.3606	419
1,500	0.5833	0.4406	0.3325	643
2,000	0.5548	0.5550	0.3129	870
3,000	0.5423	0.6830	0.3031	1,309
4,000	0.5683	0.6785	0.3030	1,847
5,000	0.5785	0.6615	0.3063	2,308
6,000	0.5816	0.6668	0.2911	2,950
7,000	0.5660	0.7373	0.2757	3,531

Note. All values are averages across all of the temporal folds that the model was trained on.

Table 5. *Results table for the best Referral Model.*

k	Precision	Recall
50	0.9457	0.01474
100	0.9450	0.02947
250	0.9406	0.07357
500	0.9386	0.1467
750	0.9381	0.2200
1,000	0.9348	0.2921
1,500	0.9204	0.4294
2,500	0.8688	0.6275
5,000	0.8115	0.6663
7,500	0.6975	0.7584
10,000	0.5664	0.8156

Note. All values are averages across all of the temporal folds that the model was trained on.

section, as well as more alerts that have an unknown gender and age label. This is a different property to alerts that are missing these labels, as it explicitly suggests that whoever made the alert could not determine the rough sleeper's age or gender, possibly indicating that the person is sleeping and covered up. Moreover, there are a higher proportion of alerts corresponding to females in the bottom right quadrant than in any other, which further highlights the potential of this quadrant for exploration should HL have extra resources to spare in order to tackle the biases mentioned in the introduction. The chosen models all show promise in these areas and were otherwise minimal in the biases that they exhibit.

When comparing the two chosen models for each classification task, it can be seen that the Referral Model performs a lot better than the Positive Outcome Model at comparable ***k***. This is likely due to the fact that the referral process underlying the labels for the former is much simpler than the distribution defining whether an outcome of a referral will be positive or not. This is due to the complexity and number of factors that go into determining whether a person will be found when outreach is attempted.

https://doi.org/10.1017/dap.2020.23 Published online by Cambridge University Press

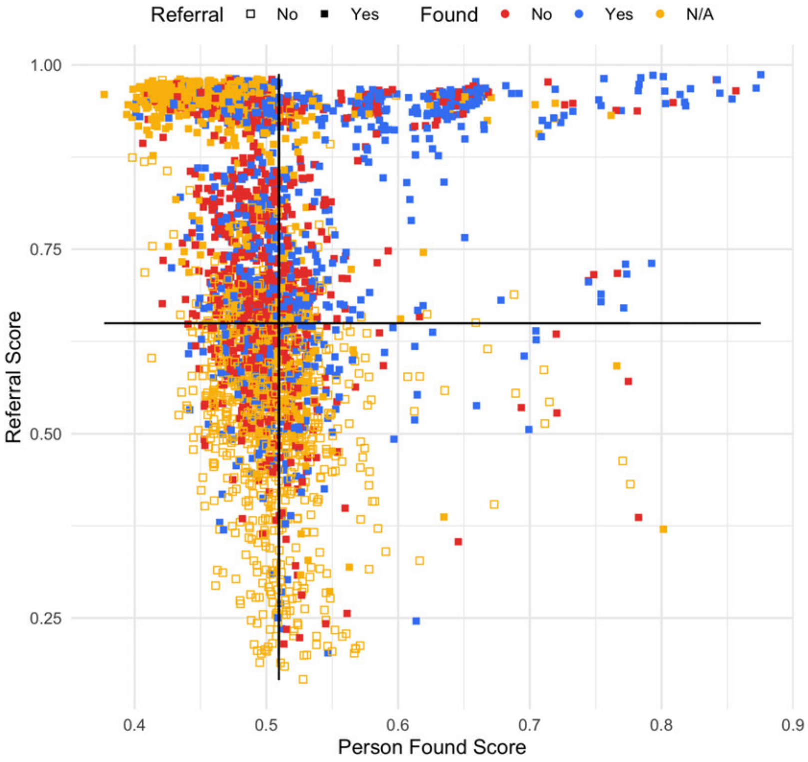

Figure 3. *Plot illustrating the distribution of scores for the chosen Positive Outcome and Referral Models. This particular plot is for February of 2019.*

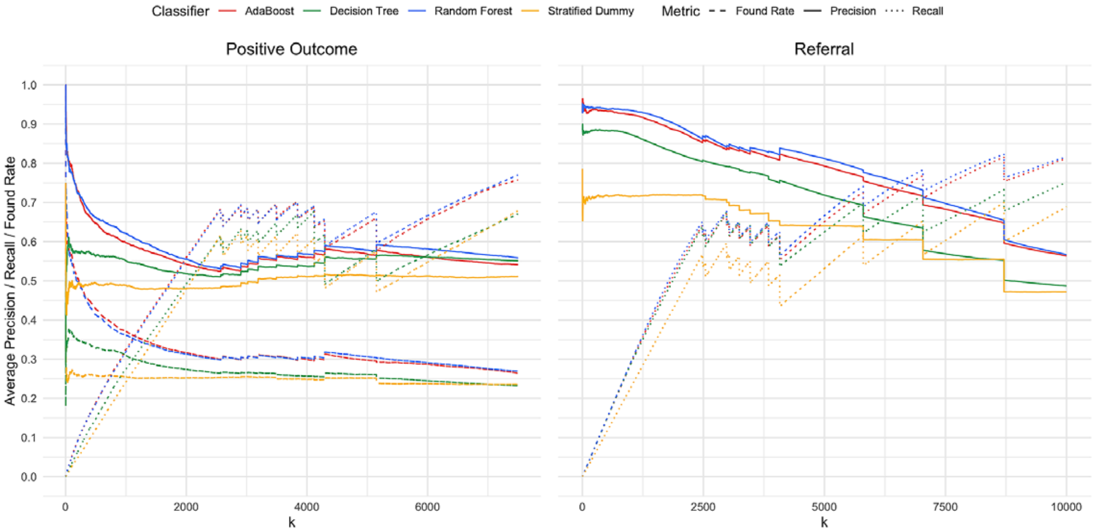

Figure 4. *The average precision, recall, and found rate across all folds for the best examples of each classifier type. Points of seeming discontinuity arise due to the nature of our temporal nested cross validation; each fold and its corresponding test set spanning a month contains a different number of alerts.*

https://doi.org/10.1017/dap.2020.23 Published online by Cambridge University Press

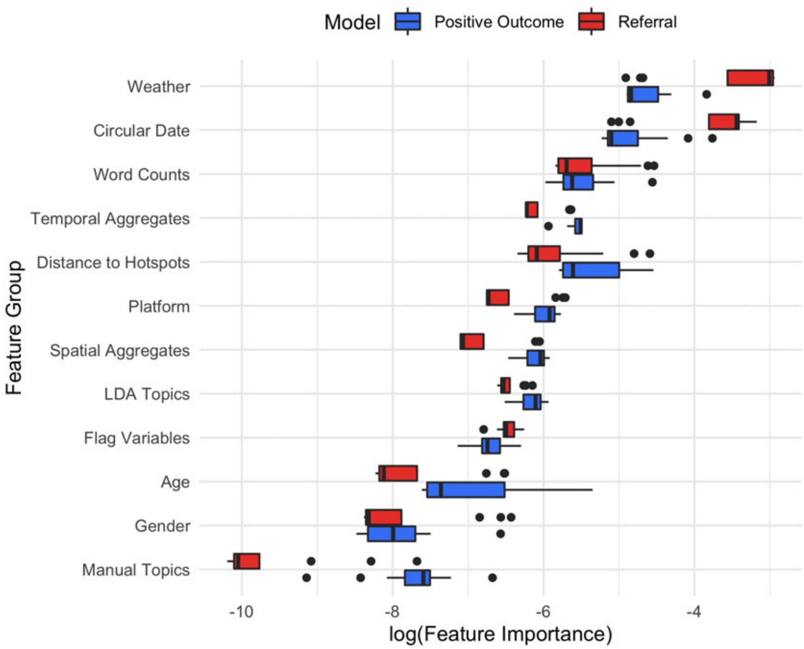

Figure 5. *Log-scaled feature group importance scores from each fold for the chosen random forest models. Temporal aggregates include all of the features generated by counting positive outcomes, referrals, and total number of alerts received in varying time windows; spatial aggregates is similar but with varying proximities; LDA topics include location and activity topics extracted from free text; manual topics are ones defined manually as important keywords to count; platform includes the features that indicate whether the alert originated from the web, mobile app, or via a phone call.*

Jaccard similarities were calculated for the label predictions made across different classifier types and configurations to assess whether different classifiers were better at predicting the outcomes of certain alerts. The results of this analysis were largely inconclusive with very few significant differences between lists of predictions; implying that some alerts are consistently more difficult to predict the outcomes of.

5.3. Feature importance

One reason for the eventual model choices over the similarly performing AdaBoost classifiers is the ease with which their feature importance scores can be extracted. Figure 5 shows the log-scaled feature importance scores for each feature group. Both of the most important two feature groups could feasibly represent seasonal effects and extreme weather, which often results in increased levels of demand and subsequently more outreach. Intuitively, word counts in the free text descriptions are likely to be indicative of the overall usefulness of an alert.

Different feature groups were included and excluded in various combinations to investigate the importance scores and interactions between features across repeated experiments. It was deemed justifiable to include all of the created feature groups in the training of the final models as many of the groups help to expose and illustrate potential biases in the recommendation system as well as carrying fairly intuitive real-world interpretations.

6. Discussion

This paper introduces a novel data-driven approach to assist StreetLink in connecting rough sleepers with LSP outreach teams. Our Positive Outcome model generates a list of prioritized alerts that can be used by

https://doi.org/10.1017/dap.2020.23 Published online by Cambridge University Press

StreetLink staff to accelerate the existing manual review process and potentially augment it through automation. We also find that weather is the most essential characteristic in identifying the outcome of an alert; weather has the potential to be used as part of a forecasting system to help assess the need for rough sleeper outreach in anticipation of extreme weather across the country.

The bi-model approach suggests that some of the alerts are currently unlikely to become a referral but have a high likelihood of having a positive outcome. This is directly in conflict with the existing review process employed by StreetLink. A partial explanation is given in that some of these alerts were identified as duplicates; the existing system does not consistently update the outcome of duplicated alerts, nor are duplicates themselves always discovered. Furthermore, high-quality alerts may not be sent as referrals during peak season due to an increased volume of alerts and associated resource constraints. Therefore, the manually assigned labels pose a significant challenge to the training and evaluation of our models due to the uncertainty present in a lot of the outcomes. The models will facilitate the collection of better data to begin filling these label gaps following deployment.

6.1. Ethical considerations

The project was given a favorable opinion by the Alan Turing Institute's Ethics Advisory group. Potential disclosure concerns were mitigated through the use of the aforementioned Tier 3 secure computational environment. Moreover, all reported results conform to the UK's Office for National Statistics (ONS) disclosure risk mitigation guidance (Office for National Statistics, 2016). For example, aggregate values are not presented for any group smaller than five distinct individuals due to risk of re-identification of specific individuals.

We find that a large number of alerts regarding female rough sleepers are also characterized by this trend of a low referral score but a high-positive outcome score. This reinforces the initial discussion on the trend of female rough sleepers being less likely to be connected with services as a result of hiding away from the public due to safety concerns. This amongst the other biases between the quadrants of Figure 3 present opportunities for HL to examine whether they reach all demographics as equitably as possible.

Future work could focus on investigating whether rough sleepers within these demographics are willing to be contacted by outreach teams. If they also require help from LSPs, the existing process should be improved in order to reach people who do not feel safe to exist openly in public spaces. These biases in general pose ethical concerns about excluding certain groups of people from the current outreach support system.

Due to many of the alerts received by StreetLink containing information about the description and location of vulnerable people sleeping rough, ethical concerns were raised throughout the project and we discussed this issue as a team with HL and StreetLink volunteers. It is important to note that:

1. All alerts and outcome information were created with the consent of the alert author and the LSP in the StreetLink application.
2. Individuals contacted by LSPs have the ability to refuse any service or action at any time.
3. To protect the sensitive information in the dataset, all of the data were hosted in a Tier 3 secure environment by The Alan Turing Institute with access limited to approved individuals and in line with data privacy regulations (Arenas et al., 2019).
4. The primary goal of the partnership with HL was to positively impact the lives of rough sleepers in the UK by improving HL and StreetLink's process connecting people sleeping rough with LSPs.
5. Concerns such as inappropriate use of the proposed models for harassment or tracking purposes of individual rough sleepers were discussed and are considered a strict violation. As such, use of these models is restricted to StreetLink staff and a small team of approved researchers and maintainers.
6. Concerns were also raised that the models may learn any unknown biases in the existing human-centered process. For this reason, bias metrics were evaluated for the various demographics present in the data.

https://doi.org/10.1017/dap.2020.23 Published online by Cambridge University Press

6.2. Further development

The Positive Outcome Model's precision falls off beyond a ***k*** of 1,000–2,000. This shows a potential for improvement of this model by using more complex natural language processing (NLP) and spatio-temporal modeling approaches. Additionally, future work could pursue the issue of duplicates more explicitly by investigating whether unsupervised machine learning techniques can be used to cluster incoming alerts in spatial and temporal dimensions to give a better indication as to whether they might be duplicates with previous alerts: an issue which can only become increasingly critical as the use of StreetLink grows alongside public concern for the homeless.

Future work can also focus on engineering additional features from external data sources such as census and demographic data. As shown in our evaluation, the feature importances of weather-related features ranked relatively high despite our limited integration of weather data into the pipeline. This data could be collected at a more granular level provided an appropriate data service was discovered.

More experimentation is required to fully iterate through our features and the algorithms that are available and appropriate for the task. Further tweaking may yield better results. For example, we could sacrifice explainability in the case of an XGBoost or similar model being experimented within the name of achieving stronger results. From an applied standpoint, concerns about transparency in the model's decision making would then lead to a requirement for further investigation over whether a black box model could feasibly replace the models chosen in this paper.

6.3. Deployment and further evaluation

Deployment of the project is already underway and is being carried out in a way that is conscious of HL's technical and financial capacity. The work done so far has led to various collaborations and the provision of resources to support the project in the mid to long-term, especially in overcoming some of the difficulties in ensuring that the models can work in real-time and be retrained at reasonable intervals. HL's status as a charitable organization ensures eligibility for Microsoft's NGO Programme and so all of the work is to be hosted in an Azure environment and set up so that it can survive with a minimal requirement for expertise and maintenance.

Following deployment, it is necessary to ensure that the system has a positive impact on practice. This will take the form of a series of randomized control trials and observational studies to assess the impact that the work has on StreetLink's process over a year or more. These trials will be to test for significant changes in the positive outcome rate either directly, in terms of the alerts that are referred or not, and indirectly, by assessing whether the system shortens the time taken to make a decision and whether this has a positive impact on the outcomes of alerts following the assumption that the data will remain more relevant.

7. Conclusions

We present a valuable use case for machine learning in the area of social good via an approach to recommendation and risk classification comprised of two machine learning models. The Positive Outcome Model is for predicting whether an alert will result in a positive outcome within the next week, whilst the Referral Model is used for validation of the first and to ensure biases in the recommendations of the first are minimized or at least apparent. The model outperforms both of the defined baselines significantly and the current manual process at StreetLink by at least 15% when compared to StreetLink's average person found rate for the past year. This translates to over 350 more rough sleepers being connected with per month during the busiest winter period. When smaller ***k*** is considered, the performance improvement over the current manual process is significantly greater, suggesting that the top few received alerts could immediately be referred to alleviate some of HL's resource constraints.

The Positive Outcome Model returns a ranked list of alerts allowing StreetLink staff to prioritize and augment their review process so as to more quickly and successfully deliver much needed aid to rough sleepers. Additionally, alerts sharing certain characteristics are shown to be currently underrepresented in the referrals made, but have predicted scores that suggest a positive outcome. Further work in trialing the

https://doi.org/10.1017/dap.2020.23 Published online by Cambridge University Press

solution will evaluate the impact of this prioritization system on freeing up more resources and whether this exploration of currently underrepresented alerts leads to significant increases in the person found rate. Risk classification and prioritization in this context allow organizations like HL to make better decisions on resource allocation, ensure StreetLink staff are dedicating their specialized training effectively, and maximize the overall positive impact that they can have on vulnerable people. The data-driven, evidential nature of our approach could also lead to positive policy change and longer term impact.

Funding Statement. This research was not supported directly by any grant, but the majority of the authors were working under the banner of the Data Science for Social Good fellowship funded primarily by The Alan Turing Institute and the University of Warwick. H.W. is supported by the Feuer International Scholarship in Artificial Intelligence. L.S. is funded by the University of Edinburgh School of Informatics.

Competing Interests. The authors do not perceive there to be any competing interests regarding the presentation of this work.

Data Availability Statement. The following material details what is required to verify and reproduce our results as closely as possible. The code, YAML experiment definitions, and documentation to accompany this research is all available publicly at The Alan Turing Institute's GitHub Repository[5] so that other organizations facing similar problems are able to use it as a starting point for similar projects. As previously stated, the data used by these models are sensitive in potentially identifying vulnerable individuals which means that trained models and the original historical data cannot be released. To try and mitigate the barrier this raises to furthering the research presented and to remain within the spirit of the journal, we have provided a synthetic dataset (scrubbed free-text fields due to the possibility of personally identifiable information being present, and with anonymized labels) conforming to the shape and requirements of the code, as well as to the security and privacy-related requirements of HL.

Author Contributions. Conceptualization, methodology, analysis, investigation, data curation, visualization, validation, original draft: H.W., L.S., A.N., and Z.K.; Project administration, supervision: J.S., A.D.U., R.G., and S.V.; Ethical assessment: B.M.; Partner contact: D.V., Reviewing and editing of draft: H.W., L.L.C., A.N., Z.K., J.S., A.D.U.,D.V., B.M., R.G., and S.V.

Acknowledgments. We are grateful to Gareth Thomas and Davide Veronese, both affiliated with HL, for their impassioned involvement and assistance throughout the project. Without the support of the organization and these individuals, this research would not be as close to deployment as it is now with the potential to positively impact thousands of lives in the UK. All of this work was facilitated through the Data Science for Social Good Foundation's Fellowship program running at The Alan Turing Institute and the University of Warwick in the summer of 2019. We owe our sanity to the 15 other fellows who inspired, supported, and accompanied our team to the pub and beyond on numerous occasions, something that was often much needed. Many long nights and break-throughs are owed to our mentors Adolfo De Unanue and Pablo Rosado, our project managers Joshua Sidgwick and Andrea Sipka, the organizer of the Warwick/Turing program Sebastian Vollmer and to the founder of DSSG Rayid Ghani. A pre-print of this work can be found at `arXiv:2007.15326` (Wilde et al., 2020).

Ethical Standards. The study received a favorable ethical opinion from the Alan Turing Institute Ethics Advisory Group. All analysis of identifiable data were carried out within an appropriately accredited secure computational environment, in keeping with data privacy standards (as previously described). Furthermore, all reporting of results are done at aggregate level with suitable anonymity checks, in keeping with the ONS guidance on statistical disclosure control procedures. General Data Protection Regulation(GDPR) and other EU-imposed data standards are respected throughout; there was a significant amount of discourse with HL and the government to ensure the fair treatment of the population affected by the described research.

Reproducibility Statement. Steps to reproduce the paper's main results: grid searches of the following pairings of parameter spaces and Scikit-Learn implementations of algorithms were carried out:

1. Random forest, extra trees using the Gini impurity criterion and setting the maximum number of features to be the square root of the total. Number of estimators: 100, 250, 500, 1,000, 2,500, 5,000, 7,500, 10,000; maximum tree depth: 1, 2, 3, 4, 5, 10, and none.
2. AdaBoost. Number of estimators: 100, 250, 500, 1,000, 2,500, 5,000, 7,500, 10,000; learning rate: 0.01, 0.05, 0.1, 0.25, 0.5, 1.0.
3. Decision trees. Maximum tree depth: 1, 2, 3, 4, 5, 6, 7, 8, 9, and none.

All of the generated and extracted features were used in the final experiments in order to get a complete view of feature importance in our final models as presented in 5. Month forward-chaining temporal cross validation was used in all of our final experiments with the following parameters:

Data Start Point: 2017-12-01; Data End Point: 2019-02-28; Training Data Label Span: 1 Week; Test Data Label Span: 1 Week; Training Data Span: 2 Years; Training Frequency: 1 Day; Test Data Span: 1 Month; Testing Frequency: 1 Day; Model Update Frequency: 1 Month.

[5] https://github.com/alan-turing-institute/DSSG19-HomelessLink-PUBLIC

References

Arenas D, Atkins J, Austin C, Beavan D, Cabrejas Egea A, Carlysle-Davies S, Carter I, Clarke R, Cunningham J, Doel T, Forrest O, Gabasova E, Geddes J, Hetherington J, Jersakova R, Kiraly F, Lawrence C, Manser J, O'Reilly MT, Robinson J, Sherwood-Taylor H, Tierney S, Vallejos CA, Vollmer S and Whitaker K (2019) Design Choices for Productive, Secure, Data-Intensive Research at Scale in the Cloud. *arXiv e-prints*, arXiv:1908.08737.

Blei DM, Ng AY and Jordan MI (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Fetzer T, Sen S and Souza PC (2019) Housing Insecurity, Homelessness and Populism: Evidence from the UK, p. 59.

Honnibal M and Johnson M (2015) An Improved Non-Monotonic Transition System for Dependency Parsing, pp. 1373–1378.

Institute of Global Homelessness (2019) State of Homelessness in Countries with Developed Economies. https://www.un.org/development/desa/dspd/wp-content/uploads/sites/22/2019/05/CASEY_Louise_Paper.pdf .

Meinbresse M, Brinkley-Rubinstein L, Grassette A, Benson J, Hamilton R, Malott M and Jenkins D (2014) Exploring the experiences of violence among individuals who are homeless using a consumer-led approach. *Violence and Victims 29*(3), 122–136. https://doi.org/10.1891/0886-6708.vv-d-12-00069.

Ministry of Housing, Communities and Local Government (2019) Governmental Definitions and Notes on Homelessness. https://www.gov.uk/guidance/homelessness-data-notes-and-definitions.

Office for National Statistics (2016) Working Paper 3: Risk Management. https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/healthstatistics.

Ortiz-Ospina E and Roser M (2019) Homelessness. https://ourworldindata.org/homelessness.

Pleace N and Bretherton J (2018) Women and Rough Sleeping: A Critical Review of Current Research and Methodology, p. 38.

Řehůřek R and Sojka P (2010) Software Framework for Topic Modelling with Large Corpora, pp. 45–50.

Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI, Wintle BA, Hartig F and Dormann CF (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography 40*(8), 913–929.

Toro PA (2007) Toward an international understanding of homelessness. *Journal of Social Issues 63*(3), 461–481. http://doi.wiley.com/10.1111/j.1540-4560.2007.00519.x

Toro PA, Tompsett CJ, Lombardo S, Philippot P, Nachtergael H, Galand B, Schlienz N, Stammel N, Yabar Y, Blume M, MacKay L and Harvey K (2007) Homelessness in Europe and the United States: A Comparison of Prevalence and Public Opinion (Vol. 63) (No. 3). http://doi.wiley.com/10.1111/j.1540-4560.2007.00521.x.

Varma S and Simon R (2006) Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics 7*(1), 91. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-91.

White J and Maguire E (2018) Rough Sleeping Statistics Autumn 2018, England (Revised). https://www.gov.uk/government/statistics/rough-sleeping-in-england-autumn-2018.

Wilde, H., Chen, L. L., Nguyen, A., Kimpel, Z., Sidgwick, J., De Unanue, A., Veronese D, Mateen B, Ghani R and Vollmer, S. (2020) A Recommendation and Risk Classification System for Connecting Rough Sleepers to Essential Outreach Services. *arXiv:2007*.15326.

Cite this article: Wilde H, Chen LL, Nguyen A, Kimpel Z, Sidgwick J, De Unanue A, Veronese D, Mateen B, Ghani R and Vollmer S (2021) A recommendation and risk classification system for connecting rough sleepers to essential outreach services. *Data & Policy*, 3: e2. doi:10.1017/dap.2020.23