Top Predictors of Homelessness

Research

May 2026

TABLE OF CONTENTS

1 BACKGROUND

Homelessness has been extensively studied worldwide, with research indicating that it is influenced by a variety of factors, ranging from social structures (Tessler, Rosenheck, and Gamache 2001) to underlying patterns of causality (Barile, Pruitt, and Parker 2018). Despite the efforts to address this issue from multiple perspectives (including attempts to identify root causes and prevent homelessness), research specific to Canada remains relatively limited.

Recent advancements in methodologies have opened new avenues for studying homelessness; particularly, through the use of machine learning techniques. In Canada, these approaches have been implemented to predict homelessness and inform preventive strategies. For example, VanBerlo et al. (2021) employed the Canadian Homelessness Management Information System, utilizing shelter usage patterns of frequent users to train a predictive model aimed at forecasting homelessness. Additionally, Messier, John, and Malik (2022) compared various algorithms for classifying individuals at risk of homelessness, with the goal of improving prevention efforts.

Gao, Das, and Fowler (2017) compared various machine learning techniques, including random forest, decision trees, and logistic regression, to predict whether households would re-enter the homelessness service provision system in a Midwestern US community between 2007 and 2014. Using demographic data and service-use data from publicly funded services, they evaluated how well each model distinguished between households that exited homelessness and those that later re-entered the system. The Random Forest model performed best based on the Area Under the Curve (AUC) metric. However, the authors found that its performance declined as the time gap between the training and testing samples increased. This suggests that while machine learning models can identify useful patterns in homelessness data, their predictive strength may weaken when underlying conditions change over time.

Wang (2024) aims to address the limitations of traditional homelessness models in the United States by incorporating big data sources, including housing characteristics and shelter inventories. This study makes a significant contribution to the homelessness literature by proposing multiple predictive models that focus on specific demographic groups. Wang (2024) concludes that each demographic group represents a unique set of challenges, with distinct risk factors that vary across different subpopulations.

In this context, we sought to contribute to the research implementing machine learning to provide data-driven insights for policymakers. Specifically, we trained a Random Forest algorithm using data from Canada's 2019 General Social Survey: Canadians' Safety (Victimization). Our goal was to explore the role of various life outcomes in determining the likelihood of homelessness, identifying potential factors that could inform effective prevention policies.

The Random Forest algorithm was chosen over other well-known models (like logistic regression) based on some studies concluding that the former outperforms the latter in terms of accuracy and Area Under the Curve (Couronné, Probst, and Boulesteix 2018), while also considering the increasing noise present in survey data (Kirasich, Smith, and Sadler 2018).

Our results are limited by the quality and availability of data reported in the survey and the limitations of the chosen algorithm. The relationships identified by the results may have changed in the intervening years, as the most recent survey is from 2019. Therefore, caution should be exercised when applying these findings to current policy decisions or population estimates.

2 DATA SUMMARY

To select individuals for participation, Statistics Canada's *General Social Survey: Canadians' Safety (Victimization)* survey employes a probability sampling method. Each observation is assigned a statistical weight that reflects the number of similar individuals in the population who were not surveyed (Statistics Canada 2019). These weights ensure that the sample accurately represents the broader Canadian population.

In this analysis, theses statistical weights were used to calculated the frequency of responses, as the survey captures a representative population rather than raw counts. This procedure allows us to make inferences about the entire population, enhancing the validity and generalizing our findings.

The first data exploration showed that our output question (Have you ever been homeless; that is, having to live in a homeless shelter, on the street or in parks, in a makeshift shelter, or in an abandoned building?) was highly imbalanced, with the Yes answers (364,367.1) being only 1.26% of the No (28,861,314) answers.

This imbalance affects the results, as it is expected that the algorithm learns too little from the category with less responses to classify them correctly. As a solution, we took a sample from the population with No answers that matches the weighted count of the Yes answers. **Table 1** shows the weighted frequency for each of the concepts included in the model.

Table 1: Variables and number of responses, General Social Survey data

Concept	Answer	Weighted Count	Percentage (%)
Have you ever been homeless; that is, having to live in a homeless shelter, on the street or in parks, in a makeshift shelter, or in an abandoned building?	Yes	364,367	50.45
	No	357,837	49.55
Age group of respondent	15 to 34	230,165	31.87
	35 to 54	251,393	34.81
	55 to 74	204,452	28.31
	75 years and older	36,194	5.01
Dwelling Type of the respondent	Single detached house	385,760	53.41
	Low-rise apartment (less than 5 stories)	111,708	15.47
	High-rise apartment (5 or more stories)	47,603	6.59
	Other	177,133	24.53
Education - Highest degree	Less than high school diploma or its equivalent	141,937	19.65

Concept	Answer	Weighted Count	Percentage (%)
	High school diploma or a high school equivalency certificate	214,763	29.74
	Trade certificate or diploma	65,288	9.04
	College, CEGEP, other non-univ. cert./diploma (not trades)	144,924	20.07
	University certificate or diploma below the bachelor's level	19,514	2.70
	Bachelor's degree	88,886	12.31
	University certificate, diploma, degree above bachelor's	46,894	6.49
Gender	Male	368,401	51.01
	Female	353,803	48.99
Household size of respondent	One member in respondent's household	175,638	24.32
	Two members in respondent's household	183,485	25.41
	Three members in respondent's household	114,820	15.90
	Four members in respondent's household	105,577	14.62
	Five members in respondent's household	83,944	11.62
	Six or more members in respondent's household	58,741	8.13
Income - Total (before tax)	Less than $25,000	348,439	48.25
	$25,000 to $49,999	171,607	23.76
	$50,000 to $74,999	84,667	11.72
	$75,000 to $99,999	79,416	11.00
	$100,000 to $124,999	22,259	3.08
	$125,000 and over	15,817	2.19
Marital status of the respondent	Married	288,117	39.89
	Living common-law	51,212	7.09

Concept	Answer	Weighted Count	Percentage (%)
	Widowed	21,568	2.99
	Separated	26,387	3.65
	Divorced	47,855	6.63
	Single, never married	287,065	39.75
Medication – Help to sleep, calm down, get out of depression	Yes	191,751	26.55
	No	530,453	73.45
Number of relatives/friends respondent feels close to	None	28,021	3.88
	One or two	122,393	16.95
	Three to five	232,313	32.17
	Six to nine	131,848	18.26
	Ten or more	207,630	28.75
Region	Atlantic Region	39,384	5.45
	Quebec	160,482	22.22
	Ontario	252,467	34.96
	Prairie region	166,332	23.03
	British Columbia	103,539	14.34
Satisfaction with personal safety from crime	Very satisfied	268,304	37.15
	Satisfied	277,999	38.49
	Neither satisfied nor dissatisfied	124,541	17.24
	Dissatisfied	22,895	3.17
	Very dissatisfied	12,027	1.67
	No opinion	16,438	2.28
Self-rated general health	Excellent	105,321	14.58
	Very good	203,200	28.14
	Good	272,264	37.70
	Fair	103,536	14.34
	Poor	37,883	5.25
Self-rated mental health	Excellent	163,632	22.66
	Very good	201,294	27.87

Concept	Answer	Weighted Count	Percentage (%)
	Good	245,092	33.94
	Fair	82,832	11.47
	Poor	29,354	4.06
Using a scale of 0 to 10 where 0 means ‘Very dissatisfied’ and 10 means ‘Very satisfied’, how do you feel about your life as a whole right now?	0 – very dissatisfied	8,532	1.18
	1	1,622	0.22
	2	10,434	1.44
	3	15,512	2.15
	4	11,099	1.54
	5	70,446	9.75
	6	30,761	4.26
	7	117,606	16.28
	8	175,417	24.29
	9	124,788	17.28
	10 – very satisfied	155,986	21.60

Source: General Social Survey (Victimization) 2019.

3 MODELING AND RESULTS

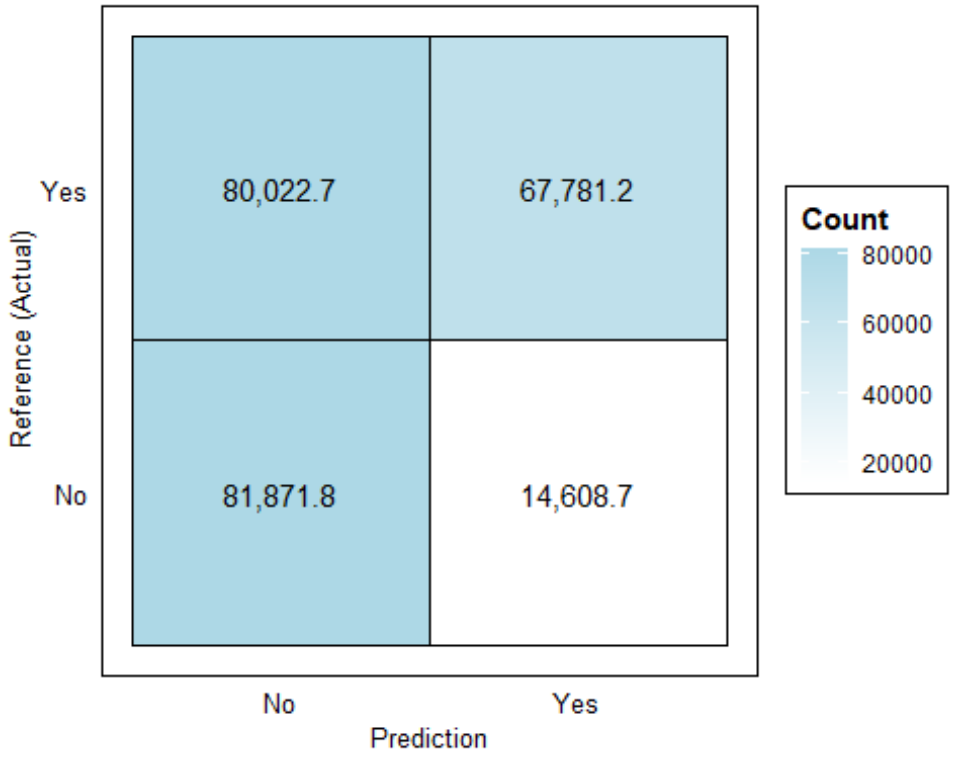
From the data described in the previous section, a random train and sample data were drawn using the 70%-30% rule. Following the classification random forest implementation suggested by Breiman (2001), 500 trees were calculated.

Given that the data is composed by categorical variables that are not balanced, the bootstrapped sampling was made without replacement, with the intention of limiting the bias across the calculated trees.

Figure 1 shows the model's results, from which the following metrics were computed:

- **Accuracy**: The model got right 61.26% of the total number of predictions.
- **Precision**: Of all homeless predictions, 82.27% were correctly classified.
- **Recall**: The model identified correctly 45.86% homeless of the total of homeless.
- **Specificity**: The model identified correctly 84.86% of non-homeless of the total non-homeless.
- **F1 score**: 58.89%. Higher values mean better overall performance.

Figure 1: Confusion matrix



Feature Importance

Although random forest models can identify which features are most useful for prediction, they are less directly interpretable than regression based approaches. Since predictions are generated from a collection of decision trees rather than a single estimated equation, the model does not produce coefficients that can be interpreted as marginal effects. As a result, random forests are

better suited to identifying predictive patterns than to explaining whether, and by how much, one variable causes another to change. Two methodologies were employed to approach the most important variables in the prediction of homelessness.

Figure 2 and **Figure 3** show feature rankings according to impurity importance and permutation importance, respectively. Impurity importance (also called Gini importance or mean decrease in impurity) ranks features by the total reduction in node impurity attributable to splits on each variable, averaged over all trees. Permutation importance measures the change in model performance (out-of-bag prediction error) when a feature's values are randomly permuted in a validation set; a larger drop indicates a more important feature. For both types of importance included, the larger the value means that the feature is more important for predicting the outcome.

Figure 2: Feature importance - average loss function (impurity)

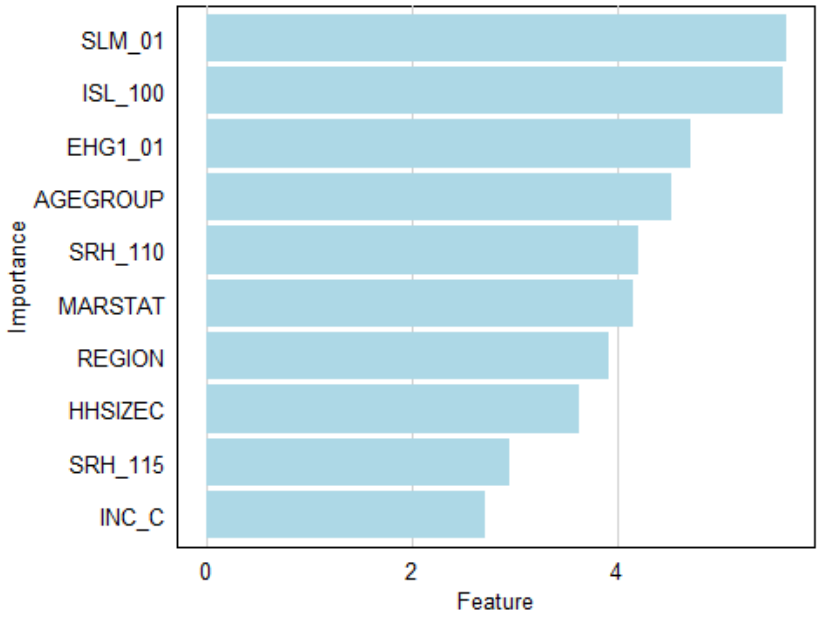


Figure 3: Feature importance - permutation

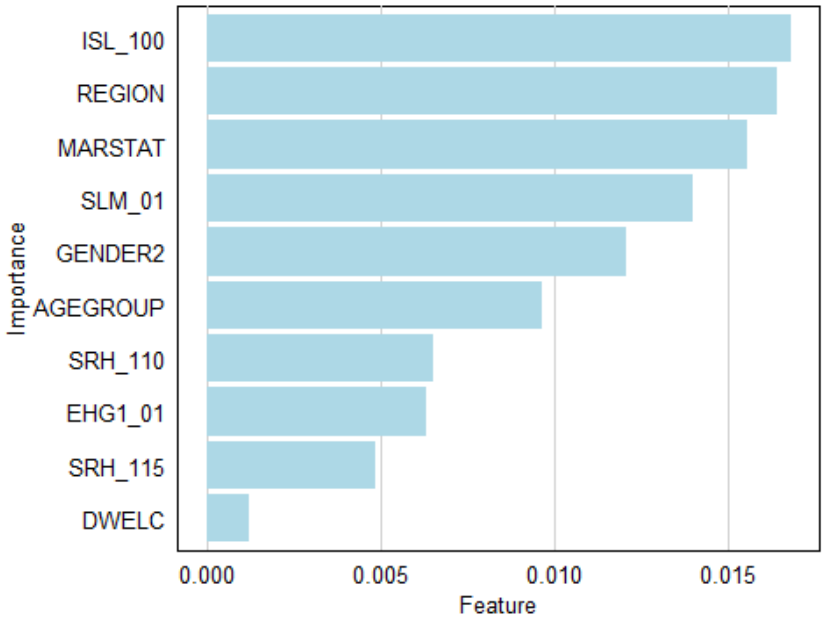


Table 2 summarizes the features that these two importance methodologies had in common and their corresponding values.

Table 2: Top feature importance

Feature	Impurity Importance	Permutation Importance
Using a scale of 0 to 10 where 0 means 'Very dissatisfied' and 10 means 'Very satisfied', how do you feel about your life as a whole right now?	5.631357	0.013997466
Number of relatives/friends respondent feels close to	5.602923	0.016807857
Education - Highest degree	4.706741	0.006321814
Age group of respondent	4.522649	0.009670417
Self-rated general health	4.198456	0.006510607
Marital status of the respondent	4.156350	0.015554679
Region	3.914519	0.016413745
Self-rated mental health	2.952666	0.004862255

The partial dependence plots below show how the model's predicted probability of homelessness changes when one predictor is varied, while all other features are averaged over. In practical terms, they show whether the model predicts a higher or lower probability of homelessness at different values of each variable, holding the rest of the data structure constant through averaging.

Figure 4: Weighted predicted probability of homelessness by region

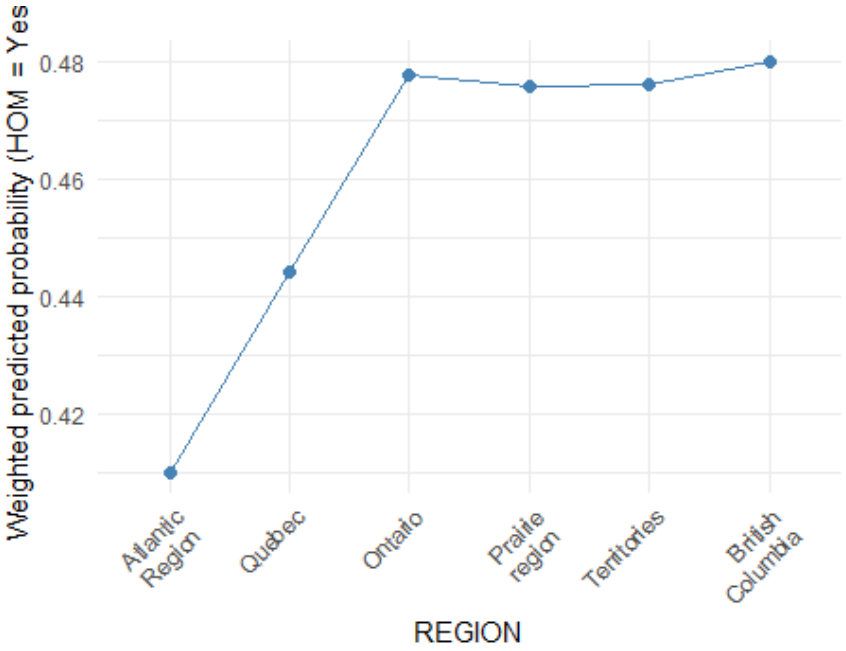


Figure 5: Weighted predicted probability of homelessness by gender

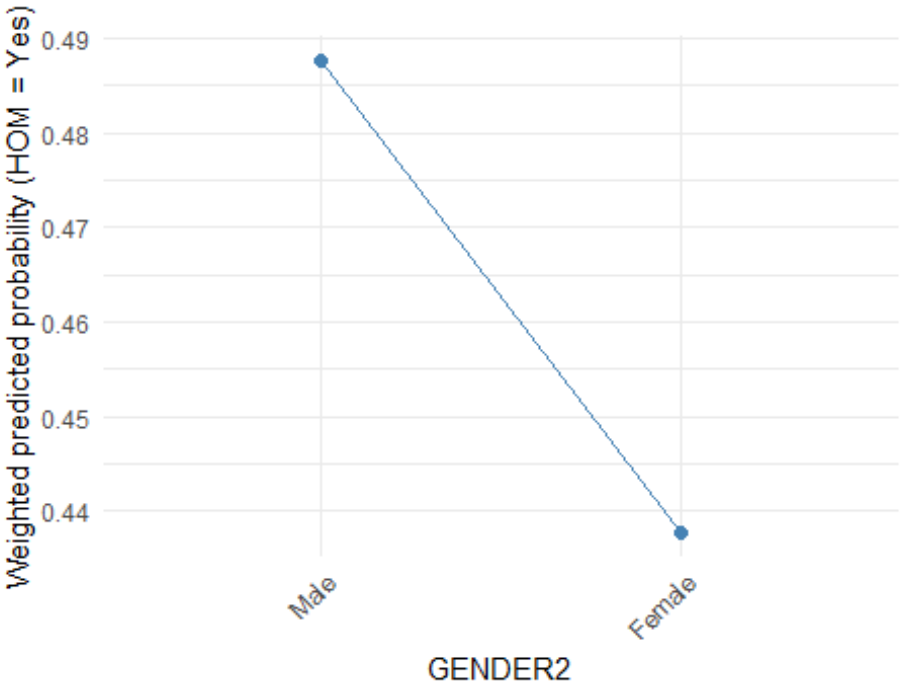


Figure 6: Weighted predicted probability of homelessness by income grouping

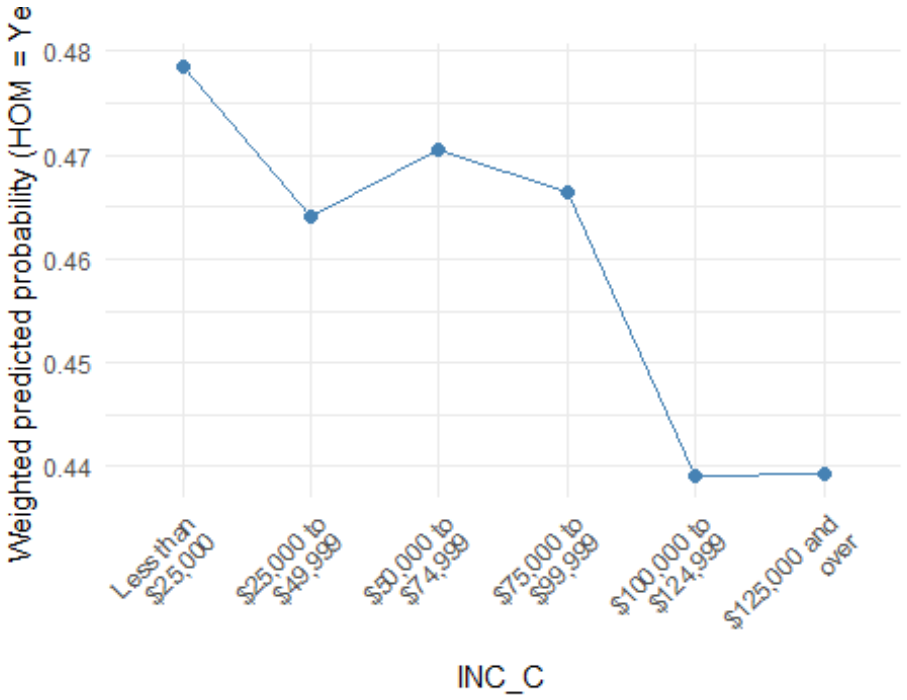


Figure 7: Weighted predicted probability of homelessness by number of close contacts

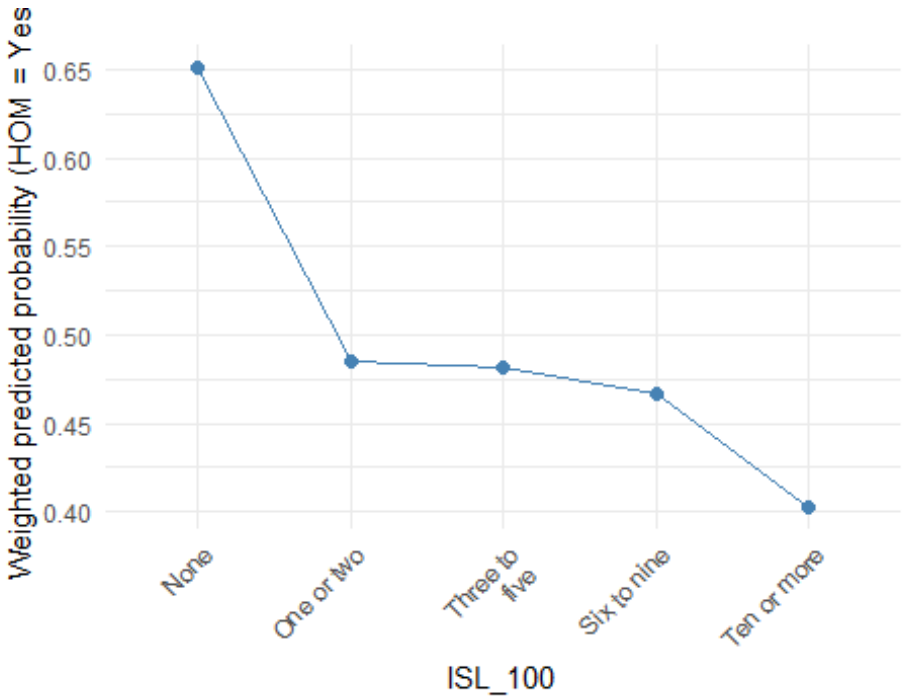


Figure 8: Weighted predicted probability of homelessness by satisfaction with personal safety

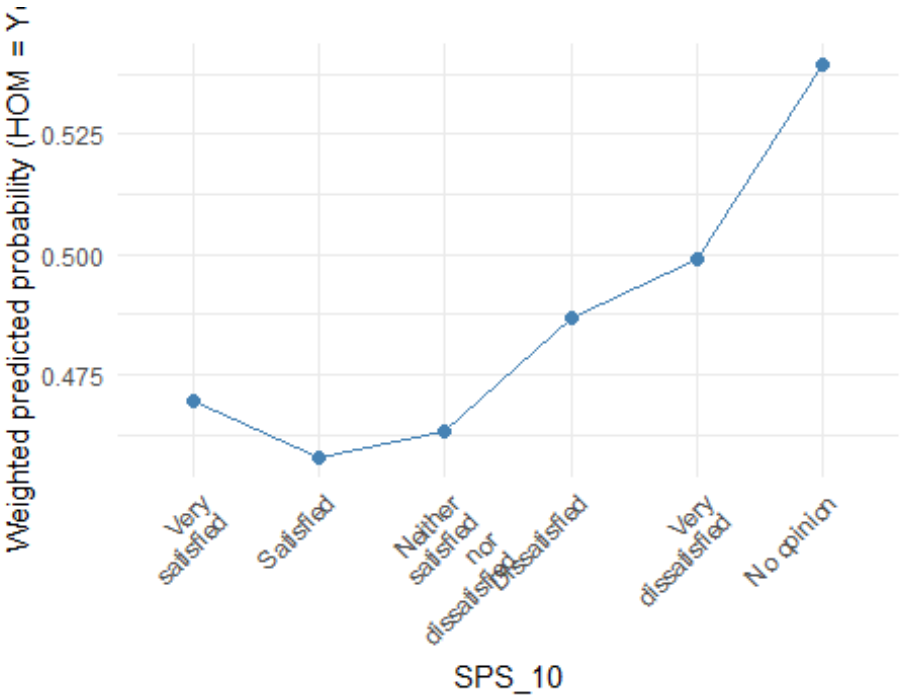


Figure 9: Weighted predicted probability of homelessness by educational attainment

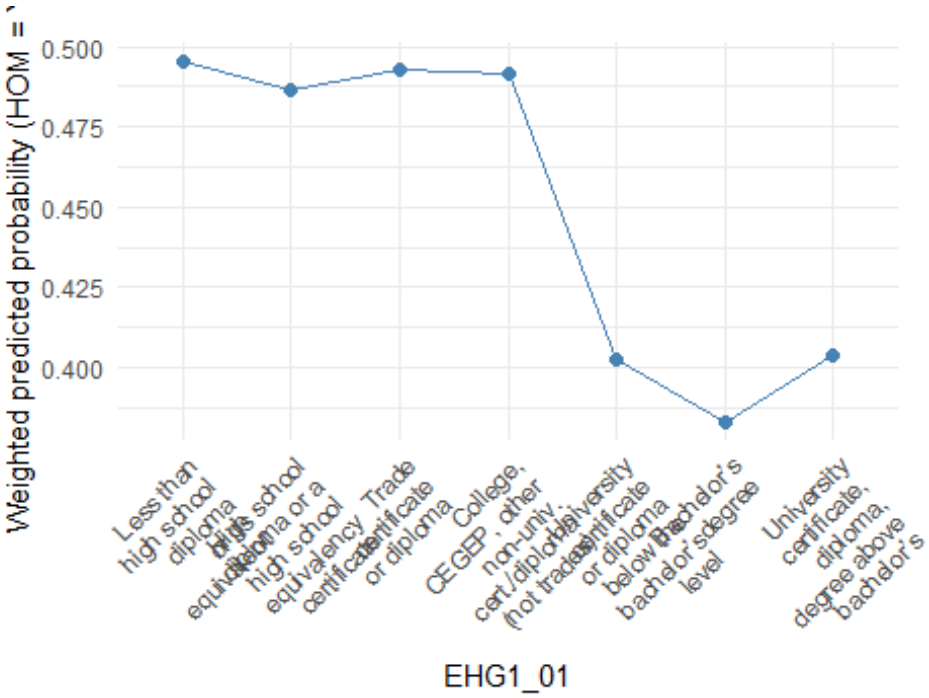


Figure 10: Weighted predicted probability of homelessness by life satisfaction

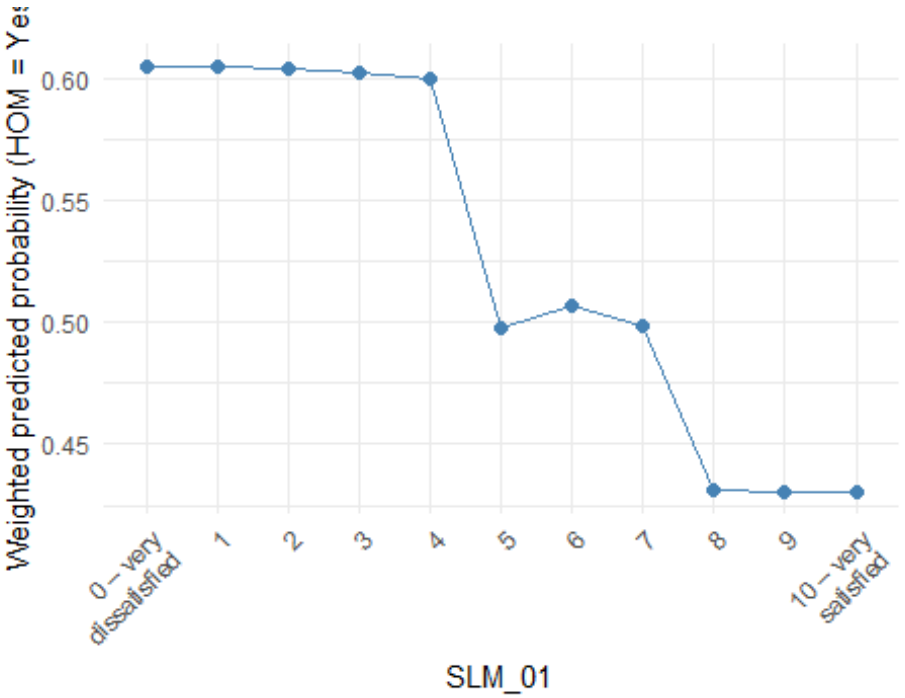


Figure 11: Weighted predicted probability of homelessness by general health

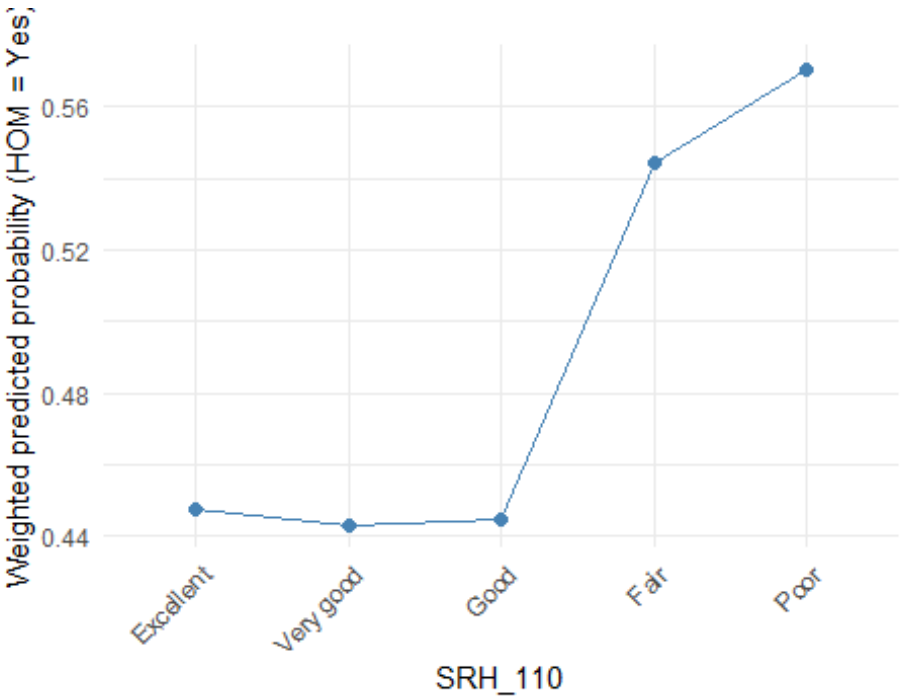


Figure 12: Weighted predicted probability of homelessness by mental health

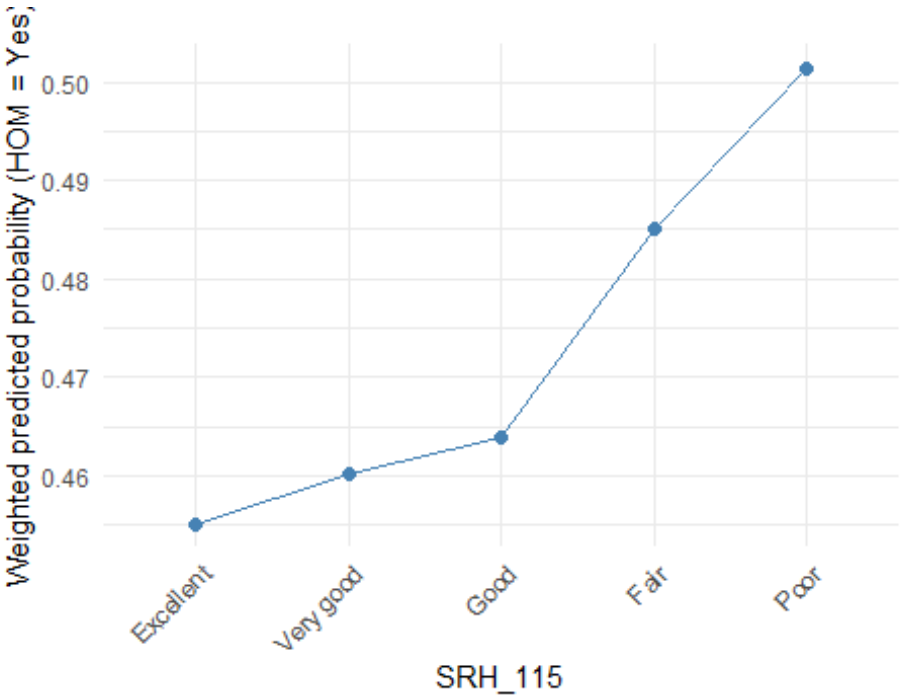


Figure 13: Weighted predicted probability of homelessness by unmet health care need

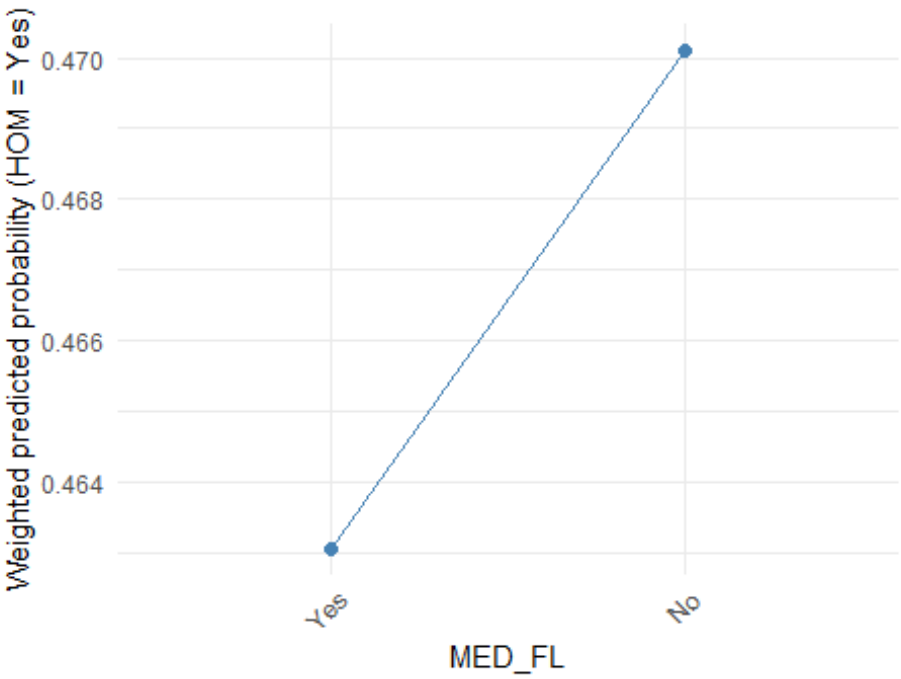


Figure 14: Weighted predicted probability of homelessness by marital status

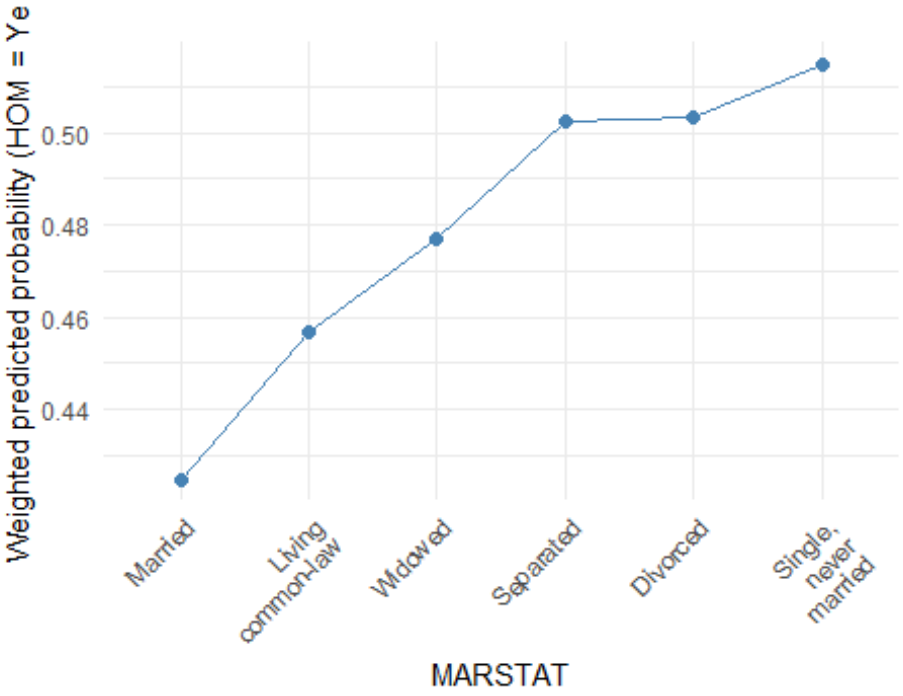


Figure 15: Weighted predicted probability of homelessness by household size

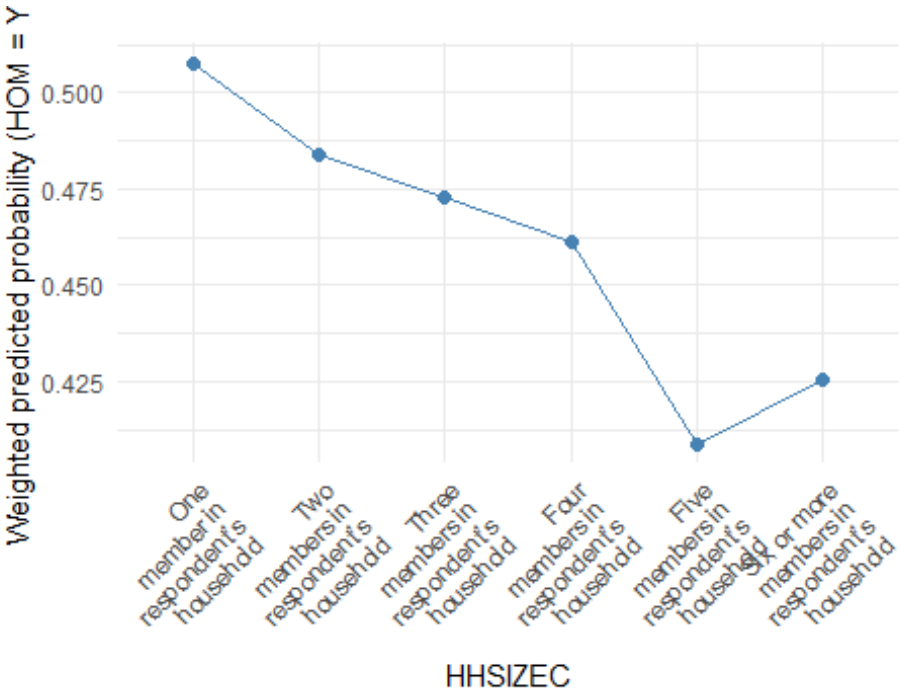


Figure 16: Weighted predicted probability of homelessness by age group

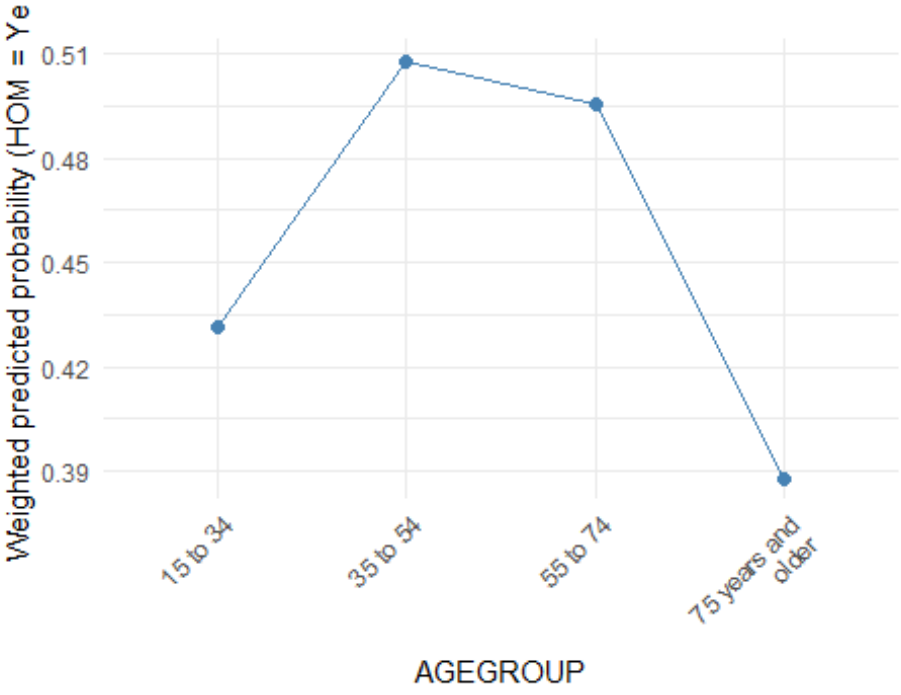
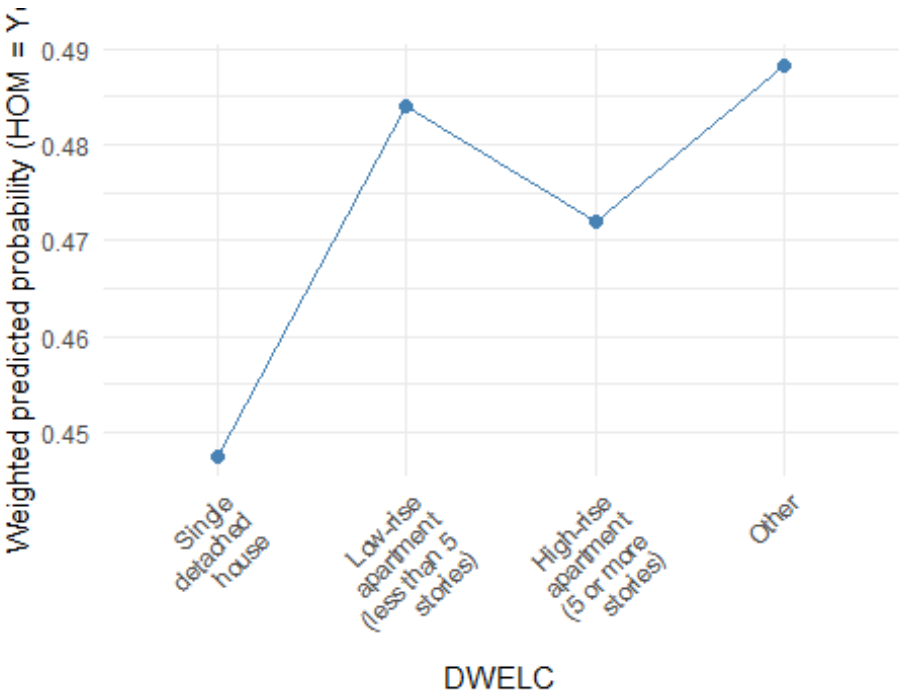


Figure 17: Weighted predicted probability of homelessness by dwelling type



4 CONCLUSION

We built a random forest model using 2019's General Social Survey to identify population-representative predictors of ever having experienced homelessness and to quantify how those predictors change the weighted probability of that outcome. The model yields modest predictive performance on a held-out sample and highlights the relative importance of socio-demographic and life-course features (e.g., housing and dwelling characteristics, income, age group, and measures of self-rated health and social supports).

These results should be interpreted with caution: the analysis is cross-sectional, depends on self-reported survey data and sampling weights, and reflects the 2019 context. Model performance is also affected by class imbalance and the choice of algorithm. For policy use, we recommend external validation on newer or administrative data, targeted exploration of the top predictors with simpler explainable models, and the design of interventions informed by the identified risk domains rather than by individual predictions alone.

5 REFERENCES

Barile, John P, Anna Smith Pruitt, and Josie L Parker. 2018. “A Latent Class Analysis of Self-Identified Reasons for Experiencing Homelessness: Opportunities for Prevention.” *Journal of Community & Applied Social Psychology* 28 (2): 94–107.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32.

Couronné, Raphael, Philipp Probst, and Anne-Laure Boulesteix. 2018. “Random Forest Versus Logistic Regression: A Large-Scale Benchmark Experiment.” *BMC Bioinformatics* 19 (1): 270.

Gao, Yuan, Sanmay Das, and Patrick J Fowler. 2017. “Homelessness Service Provision: A Data Science Perspective.” In *AAAI Workshops*.

Kirasich, Kaitlin, Trace Smith, and Bivin Sadler. 2018. “Random Forest Vs Logistic Regression: Binary Classification for Heterogeneous Datasets.” *SMU Data Science Review* 1 (3): 9.

Messier, Geoffrey, Caleb John, and Ayush Malik. 2022. “Predicting Chronic Homelessness: The Importance of Comparing Algorithms Using Client Histories.” *Journal of Technology in Human Services* 40 (2): 122–33.

Statistics Canada. 2019. “General Social Survey - Canadians’ Safety (Victimization).” https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=1235019.

Tessler, Richard, Robert Rosenheck, and Gail Gamache. 2001. “Gender Differences in Self-Reported Reasons for Homelessness.” *Journal of Social Distress and the Homeless* 10 (3): 243–54.

VanBerlo, Blake, Matthew AS Ross, Jonathan Rivard, and Ryan Booker. 2021. “Interpretable Machine Learning Approaches to Prediction of Chronic Homelessness.” *Engineering Applications of Artificial Intelligence* 102: 104243.

Wang, Sage. 2024. “Mitigating Homelessness in the United States Through Machine Learning.”